Seminar Lineare Regressionsmodelle

FS2013 – Universität Bern
Dominikus Vogl – Institut für Soziologie
e: dominikus.vogl@soz.unibe.ch
Donnerstag, 10 — 12 Uhr,
SOWI PC-Pool, 1. Stock, Unitobler, Lerchenweg 36

2.5.2013

Inhaltsverzeichnis

1 28.2. Einführung

1.1 Ablauf

1. 28.2. Einführung
    - Organisatorisches
    - Arbeiten mit Stata
    - Arbeiten mit R
    - Das ado-package *estout*
2. 7.3. Lineare Regression I
    - Bivariate Regression
    - Dummy-Variablen
    - Korrelationsanalyse
    - Multivariate Regression
    - Interaktionseffekte
    - Theoretischer Hintergrund der OLS-Schätzung
3. 14.3. Lineare Regression II
    - Regressionsdiagnostik
    - Einflussreiche Fälle
    - Transformation von Variablen
4. 21.3. Lineare Regression III
    - Regressionsdiagnostik
    - Transformation von Variablen
5. 28.3.Faktorenanalyse
6. 4.4. (Osterferien)
7. 11.4. Generalisierte Lineare Modelle (GLM) I
    - Odds-Ratios
    - Maximum Likelihood Methode
    - Logit/Probit-Modell
8. 18.4. Mulitnomiale Modelle

9. 25.4. Zähldatenmodelle / Poisson Regression
10. 2.5. Hierarchische Regressionsmodelle I
    - Variance Component Model
    - Random Intercept Model
11. 9.5. (Feiertag)
12. 16.5. Hierarchische Regressionsmodelle II
    - Random Coefficient Model
    - Hierarchische Logistische Regression
13. 23.5. Präsentation I
14. 30.5. Präsentation II

1.2 Organisatorisches

- **Sprechstunde** findet bei mir grundsätzlich auf Absprache statt, unkompliziert per mail an dominikus.vogl@soz.unibe.ch. Ich werde nach dem Seminar für Fragen, Anregungen und Wünsche da sein. Wenn der PC Pool frei ist entweder gleich dort oder sonst vor der Kaffee-Maschine im Bärengraben.
- **Leistungsnachweise**. Es soll im Laufe des Seminars eine Forschungsfrage erarbeitet und präsentiert werden. Hierfür wird am Anfang des Semesters ein Exposé verfasst, das als Leitfaden für die Seminararbeit dient. Die Seminararbeit beinhaltet die vollständige Ausarbeitung der gewählten Fragestellung, allerdings lege ich mehr Wert die statistische Datenanalyse der Hypothesen – hier bitte auch Fachliteratur angeben – als auf die vollständige Darstellung des Forschungsstandes. Der Abgabetermin wird am Ende des Semesters bekannt gegeben, wird vermutlich Ende September bzw. Anfang Oktober sein.
- **Fehlzeiten**: Ich werde für alle, die einen Leistungsnachweis möchten eine Anwesenheitsliste führen. Generell sollten alle Stunden besucht werden und ich bitte euch, sich mit einer Email abzumelden.
- **Präsentation** der Forschungsfrage. Die Präsentation dient dazu, eine Feedback zu erhalten, sie muss nicht perfekt sein, aber einen Überblick über den Arbeitsprozess liefern. Die Präsentation sollte Literaturangaben zum Forschungsstand enthalten und daraus prüfbare Hypothesen abgeleitet werden. Daten für die Datenanalyse solltet ihr zum Zeitpunkt der Präsentation bereits kennen und bearbeitet hab, da die Arbeit sonst einen zu grossen Aufwand bedeutet und unsicher ist, ob die Frage überhaupt bearbeitet werden kann. Die Methode bzw. das statistische Modell müssen explizits beschrieben werden.

1.3 Vom Sinn der Statistik

Habt ihr Ideen, warum Statistik einen Beitrag zur Erkenntnis leisten kann? Vergleiche hier anregende Gedanken von (Fox 2008: Kapitel 1).

- Kann man die Realität in Modellen darstellen?
- Gibt es systematische Faktoren, die gewissen sozialen Strukturen zugrundeliegen? Bildungsentscheidungen, Wahlpräferenzen, Berufswahl, Partnerwahl, Einstellungen, Einkommen, Status, ökonomischer Wohlstand eines Landes.
- Stastistische Modelle sind eine Vereinfachung der Realität!
- Statistik soll helfen die Welt zu beschreiben und wenn möglich auch deterministische Zusammenhänge erklären – wenn möglich. Zuerst ist es ein Mittel die Realität aufzudecken und Zusammenhänge zu enthüllen, also auch Erwartungen über Zusammenhänge zu bestätigen oder zu widerlegen.
- Was können wir mit den vorhandenen Informationen überhaupt prüfen? Kausale Zusammenhänge oder schlichtweg Beziehungsverhältnisse und die Kausalrichtung ist nicht bekannt?
- Statistik beschreibt das Ergebnis sozialer Prozesse, aber meist nicht den eigentlichen Prozess.

1.4 Forschungsfragen

Welche Fragen bewegen euch? Welche Frage wollt ihr schon immer untersuchen? Schreibe die Frage jetzt auf. Doch warte!! Erst erzähle sie deinem Nachbarn, deiner Nachbarin. Erzähle auch kurz wer du bist und was du machst. Wenn du keinen Nachbarn findest suche dir einen. Jetzt ist Raum für Phantasie und Wünsche – ja auch das gibt es in der Wissenschaft

Später werden wir an die Grenzen der Datenanalyse stossen. Aber vielleicht findet sich etwas von deinem Wunsch in deinem Ergebnis auch wieder.

1.5 Datenquellen

Wir werden in dem Seminar mit unterschiedlichen sekundär Daten arbeiten – d.h. wir greifen auf bestehende Umfragen zurück und erheben keine eigenen Daten.

- Schweizerische Gesundheitsbefragung (2007, 2002)
- Schweizerische Haushaltspanel (SHP), jährliches Panel
- Beispieldatensätze aus Lehrbüchern

- International Social Survey Programme (ISSP), jährliche Erhebung zu wechselnden Themen.
- European Values Study (EVS) (2000, 2008)
- European Social Survey (ESS) (2002, 2004, 2006, 2008, 2010)

Bitte mal nach zwei Umfragen googlen.

1.6 Transparente Arbeitsweise

Bevor man die Datenanalysen beginnt wird mit dem Befehl `cd` das Arbeitsverzeichnis bestimmt (`cd`="change working directory"). In diesem Verzeichnis sollte sich der *Do*-File befinden und ein Ordner mit dem Datensatz, den man z.B. als Ordner *data* benennt. Ein Stata-Datensatz endet auf `.dta`. Mit dem Befehl `use ../data/name.dta` wird ein Datensatz geöffnet. Natürlich kann ein Datensatz auch im working directory oder an einem beliebigen Ort gespeichert werden. Setzt man hinter den `use` Befehl die Option `clear`, so wird ein bereits geöffneter Datensatz geschlossen.

Ein Datensatz wird mit dem Befehl `save` gespeichert. Dabei sollte die Option `replace` verwendet werden, die ermöglicht, dass ein bestehender Datensatz überschrieben werden kann.

Grundsätzlich sollte ein Datensatz nach Veränderungen unter einem neuen Namen abgespeichert werden. Aus dem Datensatz `name.dta` wird dann z.B. `name1.dta`. Veränderungen sind das Umkodieren, das Erstellen und das Löschen von Variablen. Jeder Arbeitsschritt muss dabei in einem *Do*-file dokumentiert werden. Nur so ist eine Arbeitsweise möglich, die dem wissenschaftlichen Kriterium der Transparenz entspricht. Ganz nebenbei bemerkt hilft die transparente Arbeitsweise vor allem einem selbst, da Änderungen leichter gemacht werden können und Kodierungsfehler korrigiert werden können. Darüber hinaus erinnert man sich später, wie man die Berechnungen durchgeführt hat.

Do-Files enden mit `.do` und enthalten sowohl alle Befehle, die zu Änderungen des Datensatzes führen und Befehle, die den Datensatzes auswerten. Hierfür werden zwei Arten von Do-Files erstellt – die *Create*-Do-Files und die *Analyse*-Do-Files.

- Der **Create**-Do-File dokumentiert Änderungen, die den Datensatz betreffen. Zu Beginn wird der Datensatz geladen (z.B. `use name.dta, clear` und anschliessend bearbeitet. Am Ende des Do-Files werden die Veränderungen in einem neuen Datensatz gespeichert (z.B. `save name1.dta, replace`).
- item Der **Analyse**-Do-File enthält Befehle, die zur Auswertung der Datensätze führen. Zu Beginn des Do-Files wird ein (schon bearbeiteter) Datensatz geöffnet (`use name1.dta, clear`) und am Ende wird der Datensatz nicht gespeichert, sondern mit dem Befehl `clear` wieder geschlossen. In Analyse-Do-Files werden also keine Veränderungen am Datensatz vorgenommen.

In einem `Master`-Do-File werden die einzelnen Do-Files dokumentiert. Hier stehen alle `cr`- und `an`-Do-Files und werden mit dem Befehl `do crname` oder `do anname` aufgerufen. Unten seht ihr ein Beispiel für die drei Varianten von Do-Files (Es werden drei separate Do-Files gespeichert, die hier allerdings um Platz zu sparen, zusammengefasst sind.).

```
********************** crname1.do ***********************************
* Create Do-File -- crname1.do
use name, clear // Öffne Datensatz

save name1, replace // Speichere Datensatz

********************** anname1.do ***********************************
* Analyse Do-File -- anname1.do
use name1, clear // Öffne Datensatz

clear // schliesse Datensatz

********************** master.do ************************************
* master -- master.do

do crname1 // Do-File für Erstellung Datensatz name1.dta wird ausgeführt
do anname1 // Do-File für Analyse Datensatz name1.dta wird ausgeführt
```

1.7 Beispiel – Eigener Datensatz

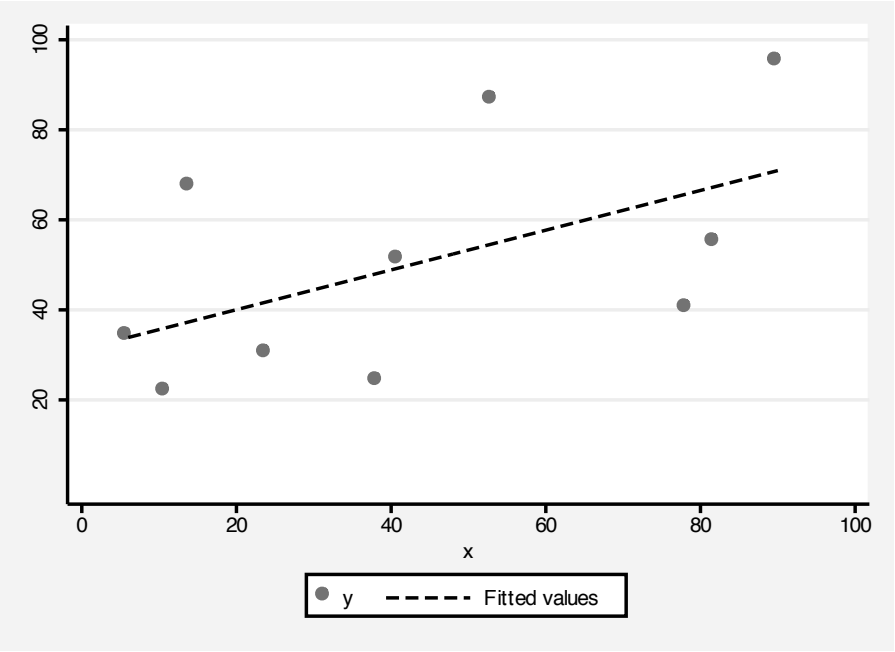
Zuerst werden aber die Variablen in einen Datensatz eingelesen. Dies geschieht mit dem Befehl `input`. Zuerst werden die Variablen `x` und `y` benannt und anschliessend die zehn Werte für `x` und `y` eingetragen. Der Befehl `end` gibt an, dass die Stichprobe zehn Beobachtungen hat. Mit `edit` wird der Dateneditor geöffnet und die Dateneingabe überprüft. Mit `sum` werden für beide Variablen statistische Masszahlen ausgegeben.

```
. clear  // offer Datensatz wird aus Arbeitsspeicher gelöscht
. input x y
             x  y
  1.         06 34
  2.         11 22
  3.         14 67
  4.         24 30
  5.         38 24
  6.         41 51
  7.         53 87
  8.         78 40
  9.         82 55
 10.         90 95
 11. end
```

```
. edit // visuelle Prüfung der Eingabe im Datensatz-Editor

. * Statistik der Variablen ausgeben lassen
. sum y x

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |        10        50.5    25.65259        22         95
           x |        10        43.7    31.04495         6         90
```

Von Interesse ist natürlich wie sich die abhängige Variabel y verändert, wenn sich die unabhängige Variable x um eine Einheit erhöht. Dies Werte werden mit der Kleisten-Quadrate Methode geschätzt. Ebenso kann man sich ein Scatterplot ausgeben lassen.

```
reg y x
estimates store ownreg // Speichere Regressionsgleichung

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    3.20
       Model |  1690.67341     1  1690.67341           Prob > F      =  0.1116
    Residual |  4231.82659     8  528.978323           R-squared     =  0.2855
-------------+------------------------------           Adj R-squared =  0.1961
       Total |      5922.5     9  658.055556           Root MSE      =      23

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .4414867   .2469487     1.79   0.112     -.127978    1.010951
       _cons |   31.20703   13.01375     2.40   0.043     1.197267    61.21679
------------------------------------------------------------------------------
```



Zur Erinnerung, und wir werden auch noch später darauf zurückkommen: die Formel einer Regressionsgleichung lautet:

$$\underbrace{y_i}_{\text{Beobachtung}} = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Schätzwert } \widehat{y}_i} + \underbrace{\varepsilon_i}_{\text{Residuum}} \tag{1}$$

Die Konstante β_0 beträgt 31.2 und der Steigungskoeffizient β_1 der Variable x ist 0,44. Die erklärte Varianz (R^2) ist mit 0.28 höher als in der eigenen Schätzung (0.21). Zusätzlich erhalten wir die Irrtumswahrscheinlichkeit von 0.11, so dass der beobachtete Zusammenhang mit hoher Wahrscheinlichkeit zufällig ist. Man sieht auch im Scatterplot, dass die Beobachtungspunkte weit um die Regressionsgerade streuen und die Ergebnisse daher nicht auf die Grundgesamtheit verallgemeinert werden sollten.

1.7.1 Speichern der Daten mit dem ado-Package `estout`

- am elegantesten geht es mit dem package *estout*, das Ben Jann (Jann 2005) (Jann 2007) geschrieben hat und das man sich über die STATA Menüleiste: Help –> SJ and user-written Programs –> Search "estout" –> packages *st0085_1* anklicken und installieren.
- mit estout kann man Tabellen erstellen, die man als Word-Tabelle abspeichert (rtf Format) und direkt in sein Word-Dokument einfügen kann.
- Man erspart sich mühsames und fehleranfälliges Kopieren per Hand.

- Also ruhig mal ausprobieren.
- man kann sich die Daten auch in einem Excel-File (.csv) oder LaTeX-file ausgeben (.tex) lassen.

```
esttab * using own_reg.rtf, replace /// Speichern der Ergebnisse
        b(a2) se(a2)  sfmt(a2) /// Format der Angaben
        r2 ar2 se label nonumber /// Angaben in Tabelle
        title({\b Table 1: This is a bold title})  ///
        mtitles("Model 1" "Model 2" "Model 3")
```

- nach `esttab` werden mit `*` alle gespeicherten Modelle verwendet. Alternativ kann man auch die Namen, der mit `estimates store` erzeugten Modelle hinschreiben: (`ownreg`).
- `using name.rtf`: in die Working Directory wird eine RTF-File geschrieben, das einen bestimmten Namen bekommt. In unserem Fall: own_reg.
- nach dem Komma kommen die Optionen
  - `replace` das rtf-file wird überschrieben, ohne dass eine Fehlermeldung kommt.
  - `b(a2) se(a2) sfmt(a2)` alle Angaben werden mit zwei Nachkommastellen ausgegeben.
  - `r2 ar2 se` die bekannten Statistiken werden ausgegeben. Setzt man `se` so erscheinen Standardfehler, ohne `se` werden t-Werte ausgegeben.
  - `label` die Variablen werden nicht mit dem Variablen*name* sondern dem Variabeln*label* angezeigt.
  - `nonumber`: die Modelle werden nicht nummeriert.
  - `title()` hier kann ein Titel eingetragen werden, der in diesem Fall noch **fett** ist.
  - `mtitles("name1name2usw...")` hier können für jedes Modell eigenen Bezeichnungen gegeben werden. Lässt man diese Option aus, so erscheint das Label der abhängigen Variabel.
- im STATA output fenster kann man nun auf den den blauen Dateiname.rtf klicken und das Dokument öffnet sich direkt.
- Alternativ könnte man auch eine Excel .csv-Datein erstellen, die man öffnet und dann mit der EXCEL-Option `Daten -> Text in Spalten -> getrennt Spalten -> Komma-getrennt -> fertig stellen` einlesen kann. Anschließend noch als gewöhnliche EXCEL-Datei .xls abspeichern (bisher ist es ja noch eine .csv Datei).

1.8 Weiterverarbeitung in Word, Excel und LaTeX

Siehe hierfür die Beispiele aus der Stunde und den Input, den es während dem Semester noch gibt.

1.9 Andere Statistikprogramme?

Ja es gibt noch andere. Wir werden die Regression versuchen in R zu lösen. Wir werden immer wieder zu R wechseln, um euch so die Option zu geben, auch dieses Programm kennen zu lernen.

```
> # in R gelten andere Regeln
> # Erstelle Variablen
> x <- c(6, 11, 14, 24, 38, 41, 53, 78, 82, 90)
> y <- c(34, 22, 67, 30, 24, 51, 87, 40, 55, 95)
> # Berechne Regression
> own.reg <- lm (y ~ x)
> # Ausgabe der Ergebnisse
> summary(own.reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-25.643 -13.650  -5.829  18.467  32.394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.2070    13.0138   2.398   0.0433 *
x             0.4415     0.2469   1.788   0.1116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23 on 8 degrees of freedom
Multiple R-squared: 0.2855,Adjusted R-squared: 0.1961
F-statistic: 3.196 on 1 and 8 DF,  p-value: 0.1116
```

1.10 Lineare Einfachregression

Das Grundprinzip der Regression ist einfach: Man versucht eine abhängige Variable durch unabhängige Variablen zu erklären. Die Erklärung erfolgt mittels einer Gleichung.

- Die abhängige Variable (AV) wird auch als Regressand, zu erklärende Variable, Prädiktand beschrieben und in einer Gleichung mit Y beschrieben.
- Die unabhängige Variable (UV) wird auch als Regressor, erklärende Variable, Prädiktor beschrieben und in einer Gleichung mit X beschrieben.

Zu Beginn jeder Forschungsfrage steht eine Überlegung zur Abhängigkeit zweier Variablen. Diese Überlegung kann als Hypothese ausformuliert werden und formal durch eine formale Gleichung dargestellt werden.

- **Hypothese:** Mit steigendem Alter (X) steigt der Body-Mass-Index (Y).
- **Gleichung:** Allgemein könnte man den Zusammenhang mit $Y = \beta_0 + \beta_1 X$ beschreiben.

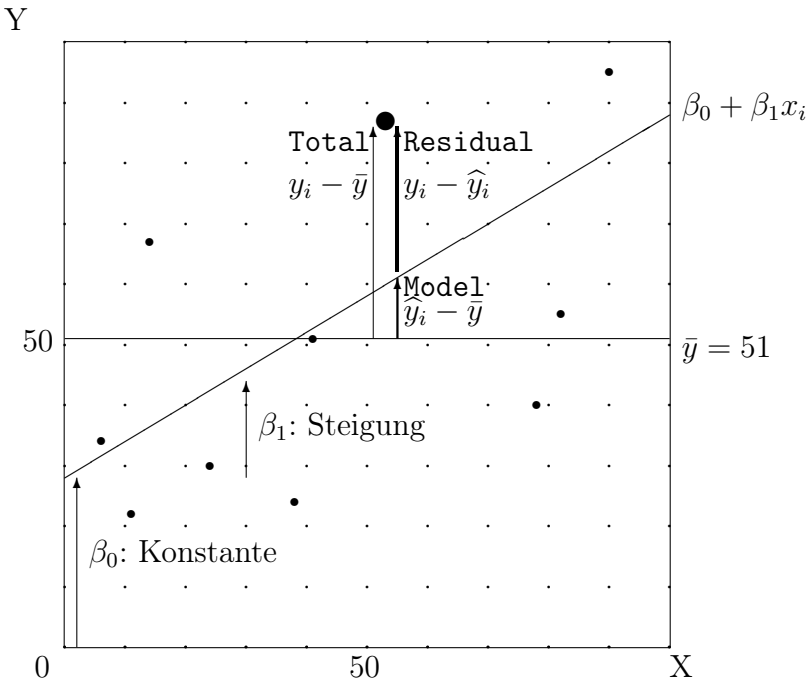
Nun können wir allgemeine Zusammenhänge nicht beobachten, sondern immer nur Ausprägungen der interessierenden Variablen (y_i und x_i) aus der Grundgesamtheit. Der Schätzwert, den man aus den Informationen der Stichprobe erhält, wird mit $\widehat{y_i}$ (gesprochen "Y-Dach") bezeichnet. Im bivariatem Fall ist der Schätzwert $\beta_0 + \beta_1 x_i$. Allerdings ist der Schätzwert $\widehat{y_i}$ einer Beobachtung nicht zugleich der tatsächlich beobachtete Wert y_i. In unserem Beispiel kann der Body-Mass-Index nicht perfekt durch die Variable Alter erklärt werden. Daher wird es einen unerklärten "Rest" geben, der als "Residuum" bezeichnet wird. Das Residuum ist die Differenz zwischen beobachtetem und geschätztem Wert.

$$\underbrace{y_i}_{\text{Beobachtung}} = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Schätzwert } \widehat{y_i}} + \underbrace{\varepsilon_i}_{\text{Residuum } (y_i - \widehat{y_i})} \tag{2}$$

Die Werte der abhängigen Variable y für alle Beobachtungen i sind somit eine *Linearkombination* der Beobachtungen der unabhängigen Variable x_i und einem unerklärten Rest ε_i. Betrachtet man die gesamte Streuung um den Mittelwert von Y, so kann ein Teil der Streuung durch das Modell erklärt werden, es gibt aber auch einen Anteil – meistens der grössere Teil – der durch das Modell nicht erklärt werden kann und unerklärt bleibt. Der Anteil der erklärten Varianz sagt etwas über die Erklärungskraft des Modells aus. Um allerdings diese Unterscheidung treffen zu können bedarf es eines Verfahrens, dass für die Gleichung 2 die Werte für β_0 und β_1 so bestimmt, dass der Unterschied zwischen dem Schätzwert und dem beobachteten Wert möglichst klein wird. Ein allgemein verwendetes Verfahren ist das OLS-Verfahren oder auch "Ordinary-Least-Squares" genannt, das in einem späteren Kapitel näher besprochen wird.

- β_0 und β_1 werden als *Regressionskoeffizienten* bezeichnet.

Der Scatterplot unten zeigt das Grundprinzip auf. Aber Achtung (!), die Schätzgerade ist nur eine Schätzung.



Insgesamt kann man die `totale` Varianz bestimmen, davon wird ein Teil durch das `Modell` erklärt und der `Rest` bleibt unerklärt. Die Regressionsgleichung wird beim OLS-Schätzverfahren so bestimmt, dass die Summe der quadrierten Residuen minimiert wird. RSS steht für "Residual Sum of Squares" und soll so klein wie möglich werden.

In einem nächsten Schritt kann jeder Studierende selbst eine Regressionsgerade in den Scatterplot zeichnen und davon ausgehend die Varianzkomponenten (RSS, TSS, MSS) berechnen. Die Werte wurden bereits in Kapitel 1.7 berechnet. Hier soll nochmals näher auf die Varianzerlegung eingegangen werden.

$$\text{RSS} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{3}$$

TSS steht für "Total Sum of Squares".

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4}$$

MSS steht für "Model Sum of Squares".

$$\text{TSS} - \text{RSS} = \text{MSS} \tag{5}$$

Die Berechnung kann man manuell in Stata vornehmen, wobei man die gewünschten Werte als Skalar (`scalar name =`) speichert, um sie für spätere Berechnungen weiterzuverwenden.

```
. * Eigenes Beispiel berechnen
. * Der Anova Block
. *RSS
. scalar rss_manual = (34-32)^2 + (22-34)^2 + (67-36)^2 + (30-45)^2 + (24-50)^2 ///
 + (51-52)^2  + (87-60)^2 + (40-76)^2 + (55-78)^2 + (95-82)^2

. display rss_manual
4734

. * RSS für die dies es ganz genau wissen wollen
. * Der Schätzwert für y ist bei jedem x: (28+3/5*x_i)
. * Die Konstante beta_0 liegt bei 28 und die Steigung beträgt 3/5
. scalar rss = (34-(28+3/5*6))^2 + (22-(28+3/5*11))^2 + (67-(28+3/5*14))^2 ///
+ (30-(28+3/5*24))^2 + (24-(28+3/5*38))^2 ///
  + (51-(28+3/5*41))^2  + (87-(28+3/5*53))^2  ///
+ (40-(28+3/5*78))^2 + (55-(28+3/5*82))^2 + (95-(28+3/5*90))^2

. display rss
4588.16

    . sum y

        Variable |       Obs        Mean    Std. Dev.       Min        Max
    -------------+--------------------------------------------------------
               y |        10        50.5    25.65259         22         95

    . scalar mean_y = 50.5

    . scalar tss = ( 34 - mean_y)^2 + ( 21 - mean_y)^2 + (67 - mean_y)^2 ///
    + ( 30- mean_y)^2 + ( 25- mean_y)^2 + ( 51- mean_y)^2  + (88 - mean_y)^2 ///
    + (40 - mean_y)^2 + (55 - mean_y)^2 + (95 - mean_y)^2

    . display tss
    6002.5


    . *MSS
    . scalar mss = tss-rss

    . display mss
    1414.34
```

Anschliessend können wir den Anteil der "totalen" Varianz berechnen, die durch das Modell erklärt wird. Dieser Anteil wird durch den *Determinationskoeffizient* R^2 ausgedrückt.

$$R^2 = \frac{\text{MSS}}{\text{TSS}} \tag{6}$$

```
. * Der Modellfit-Block
. * R^2: Erklärte Varianz
. scalar r2 = mss/tss
. display r2
.23562516
```

Neben dem R^2 kann man auch den "Root MSE" berechnen, der die durchschnittliche Abweichung der beobachteten Werte vom Schätzwert angibt. Diese Masszahl entspricht der Wurzel der durchschnittlichen Residuen des Modells. K beschreibt die Anzahl der Koeffizienten, die in dem Modell verwendet werden. In diesem Modell sind es die Koeffizienten β_0 und β_1.

$$\text{Root MSE} = \sqrt{\frac{\text{RSS}}{\text{n-K}}} \tag{7}$$

```
. * Root MSE
. scalar k = 2     // Das Model besteht aus zwei Koeffizienten beto_0 und beta_1
. scalar n = 10    // es gibt 10 Beobachtungen
. scalar rootmse = sqrt(rss/n-k)
. display rootmse
23.948278
```

Der Wert von 21,37 besagt, dass man bei der Schätzung im Schnitt um rund 21 Einheiten daneben liegen.
Schliesslich können wir mit den vorhandenen Informationen noch einen Test machen, der besagt, ob die verwendete Variable tatsächlich einen Teil der Varianz erklären kann oder ob der beobachtete Zusammenhang auch rein zufällig sein könnte. Hierfür wird ein F-Wert $F(k-1, n-k)$ bestimmt, der von der Anzahl der verwendeten Variablen $(k-1)$ und der Anzahl der Beobachtungen um die Anzahl der Koeffizienten vermindert $n-k$ abhängt.

$$F = \frac{\text{MSS}/k-1}{\text{RSS}/n-k} \tag{8}$$

```
. * F-Wert
. scalar degree = k-1     // Anzahl der Freiheitsgrade (Anzahl der UV's im Modell)
. scalar f_value = (mss/k-degree)/(rss/n-k)
. display f_value
1.5458522
```

Wir erhalten für $F_P(1,8)$ einen Wert von 1,567. Da diese Prüfgrösse einer F-Verteilung folgt, kann man für das Signifikanzniveau von $P = 0,95$ den kritischen Wert in einer Tabelle nachschlagen. Der kritische Wert beträgt 5,32 und liegt über dem beobachtetem Wert. Damit kann es sein, dass das beobachtete R^2 von $0,24$ in der Grundgesamtheit eigentlich Null ist. In `Stata` wird zusätzlich noch die Irrtumswahrscheinlichkeit "Prob > F" angegeben, die die Wahrscheinlichkeit angibt, dass das berechnete R^2 der Stichprobe in der Grundgesamtheit eigentlich Null ist.

Abschliessend werden in einem letzten Schritt die Regressionkoeffizienten β_0 und β_1 berechnet. Die Formel lautet:

$$\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{9}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \tag{10}$$

```
. quietly sum x // Ergebnisse werden nicht gezeigt

. scalar mean_x = r(mean) // verwende gespeicherten Mittelwert: 43.7

. display mean_x
43.7

* Beta_1
scalar beta_1 = ((6-mean_x)*(34-mean_y)+(11-mean_x)*(22-mean_y)+(14-mean_x)*(67-mean_y)+
 (24-mean_x)*(30-mean_y)+(38-mean_x)*(24-mean_y)+(41-mean_x)*(51-mean_y)+ ///
 (53-mean_x)*(87-mean_y)+(78-mean_x)*(40-mean_y)+(82-mean_x)*(55-mean_y)+ ///
 (90-mean_x)*(95-mean_y))/ ///
 ((6-mean_x)^2+(11-mean_x)^2+(14-mean_x)^2+(24-mean_x)^2+(38-mean_x)^2+ ///
 (41-mean_x)^2+(53-mean_x)^2+(78-mean_x)^2+(82-mean_x)^2+(90-mean_x)^2 ///
 )
display beta_1
.44148672

. * Beta_0
```

```
.
. scalar beta_0 = mean_y - beta_1*mean_x

. display beta_0
31.20703
```

Übersicht Stata-Befehle

Befehl	**Option**	**Erklärung**
`graph export`	`replace`	Graphik in Arbeitsverzeichnis speichern. Hier als pdf, aber auch jpg oder png Format sind möglich. Die Option `replace` erlaubt die Datei zu überschreiben.
`input var1 var2`		Eingabe von Werten in den Dateneditor über den do-file. Den Befehl mit `end` abschliessen.
`estimates store name`	siehe `help estimates store`	Speichern der Schätzergebnisse.
`estimates tabel`	siehe `help`	Ausgabe der gespeicherten Ergebnisse im Result-Fenster

Aufgabe

In Vorbereitung auf die Seminararbeit würde mich interessieren, welche Forschungsfelder euch interessieren. Was könnte das *Thema* der Arbeit sein? Welche *Daten* gibt es hierfür? Falls ihr schon eine Idee habt, mit welcher *statisischen Methode* ihr die Fragestellung prüfen wollt, so könnt ihr sie auch schon nennen.

Die Antwort per mail an mich bis Mittwoch 23h30 an (dominikus.vogl@soz.unibe.ch).

2 7.3. Lineare Regressionsmodelle 1

2.1 Lineare Einfachregression mit kategorialen unabhängigen Variablen

In diesem Abschnitt werden kategoriale Variablen als erklärende Variablen besprochen. Anschliessend wird die abhängige Variable gebildet, wobei mit einer `foreach`-Schleife gearbeitet wird. Abschliessend wird das Regressionsmodell gebildet, die graphische Auswertungen erläutert und mit dem Befehl `esttab` eine Tabelle der unterschiedlichen Modelle erzeugt. Wir analysieren die Daten des International Social Survey Programme (ISSP) 2005 zum Thema work-orientation (Arbeitseinstellung), die über das GESIS-Archiv ZACAT http://zacat.gesis.org/webview/velocity?v=2&mode=documentation&submode=abstract&study=http%3A%2F%2F193.175.238.79%3A80%2Fobj%2FfStudy%2FZA4350 bezogen werden kann.

2.1.1 Unabhängige Variable

Von Interesse in der Sozialwissenschaft ist der Einfluss von Geschlecht, wobei dieser Einfluss leicht zu modellieren ist, da die Variable nur zwei Ausprägungen hat. Die zwei Ausprägungen können als "Dummy-Variable" kodiert werden, die die Ausprägungen 0 und 1 besitzen.

$$\begin{aligned} \widehat{y}_0 &= \beta_0 + \beta_1 \times 0 = \beta_0 \\ \widehat{y}_1 &= \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1 \end{aligned}$$

(vergleiche: Rabe-Hesketh und Skrondal 2008: 19)

Der Koeffizient der Regressionskonstanten (β_0) gilt für diejenigen Personen, deren Ausprägung der Dummy-Variable 0 ist und unterscheidet sich um β_1, wenn die Dummy-Variable = 1 ist.

```
use ../data/issp2005, clear // oeffne Datensatz

* Unabhaengige Variable Geschlecht
tab SEX // 1= Male 2 = Female
generate female = SEX-1
label var female "Geschlecht (1=weiblich)"
label define female 1"weiblich" 0"maennlich"
label value female female
numlabel female, add
```

Sinnvoll ist es, Dummy-Variablen nach der Ausprägung zu benennen, die mit 1 kodiert ist, daher heisst die Variable nicht `sex` sondern `female`. Mit den Befehlen `label var`, `label define` und `label value` wird der Variablenname benannt, die Bezeichnung der Ausprägungen definiert und der Variable zugeordnet. Praktisch ist der Befehl `numlabel`, der die numerische Kodierung einer Variable an die Labels anfügt.

```
. tab female

  Geschlecht |
(1=weiblich) |      Freq.     Percent        Cum.
-------------+-----------------------------------
0. maennlich |     19,806       45.64       45.64
 1. weiblich |     23,586       54.36      100.00
-------------+-----------------------------------
       Total |     43,392      100.00
```

Wie verhält es sich nun mit kategorialen Variablen, die mehr als zwei Ausprägungen haben? Beispielsweise könnte neben dem Geschlecht, das Bildungsniveau der Personen von Interesse sein. Auch hier werden für jede Bildungskategorie Dummys gebildet, wobei die niedrigste Kategorie nicht in die Regressionsgleichung aufgenommen wird. Angenommen es gibt drei Kategorien (niedrig, mittel, hoch), so ist die Regressionskonstante der Wert für die Personen, die keine mittlere oder hohe Bildung besitzen, denn die Kategorie "niedrig" ist die Referenzkategorie.

$$
\begin{aligned}
\widehat{y}_{niedrig} &= \beta_0 + \beta_1 \text{edu}_{mittel} \times 0 + \beta_2 \text{edu}_{hoch} \times 0 = \beta_0 \\
\widehat{y}_{mittel} &= \beta_0 + \beta_1 \text{edu}_{mittel} \times 1 + \beta_2 \text{edu}_{hoch} \times 0 = \beta_0 + \beta_1 \text{edu}_{mittel} \\
\widehat{y}_{hoch} &= \beta_0 + \beta_1 \text{edu}_{mittel} \times 0 + \beta_2 \text{edu}_{hoch} \times 1 = \beta_0 + \beta_2 \text{edu}_{hoch}
\end{aligned}
$$

Die Bildung von kategorialen Dummy-Variablen kann in `Stata` elegant mit dem Befehl `tab` *varname*, `gen`(*newvar*) geschehen. Dabei wird automatisch für jede der Ausprägungen der Variable eine Dummy-Variable gebildet. Verwendet man die 1 bis `k` neuen Variablen in einer Regressionsgleichung, so sollte man eine Variable als Referenzkategorie nicht in das Modell aufnehmen.

```
* UV Bildung: Dummy-Variablen für mehrere Kategorien

numlabel DEGREE, add
tab DEGREE, generate(edu_) // Dummys fuer jede Kategorie erzeugt

        R: Education II-highest |
```

```
               education level |      Freq.     Percent        Cum.
-------------------------------+-----------------------------------
     0. No formal qualification |      5,643       13.12       13.12
 1. Lowest formal qualification |      6,979       16.23       29.35
  2. Above lowest qualification |      8,127       18.90       48.25
  3. Higher secondary completed |      8,839       20.55       68.80
4. Above higher secondary level |      7,051       16.40       85.19
 5. University degree completed |      6,358       14.78       99.98
         7. CH: Other education |         10        0.02      100.00
-------------------------------+-----------------------------------
                          Total |     43,007      100.00
```

Auf die gleiche Weise kann man für Variablen mit sehr vielen Ausprägungen (z.B. den 32 Ländern im Datensatz) Dummy-Variablen erstellen.

```
* Laendervariable V3
label list V3
numlabel V3, add  // fuege Labelnummern an Name an
tab V3, mis // Anzahl Befragte pro Land

tab V3, generate(countr_) // Dummys fuer jedes der 32 Laender
```

2.1.2 Abhängige Variable

Als abhängige Variable wird ein Index gebildet. Dieser besteht aus neun Variablen zum Thema "Selbsteinschätzung" (`V78` bis `V86`), mit einer fünfstufigen Likertskala als Antwortmöglichkeit. Diese Variablen sollen pro Person zu einem Index addiert werden, so dass höhere Werte eine positive Selbsteinschätzung bedeuten (was für hohes Selbstvertrauen und -bewusstsein steht).

```
label list V79
V79:
           1 Strongly agree
           2 Agree
           3 Neither agree nor disagree
           4 Disagree
           5 Strongly disagree
```

Der Befehl `label list V79` zeigt die Kodierung der Variablenausprägungen der Variable `V79`. Hohe Zustimmung ist mit einer 1 und niedrige Zustimmung ist mit einer 5 kodiert. Also müssen alle Fragen, die mit "positiv" Formuliert sind (`V79, V80, V81, V83`) rekodiert werden, so dass 1 die Ablehnung und 5 die Zustimmung der Befragten

ausdrückt. Die "negativ" formulierten Fragen (`V78, V84, V85, V86`) müssen nicht verändert werden, da Personen mit einem hohen Selbstbewusstsein, diese Fragen ablehnen sollten.

```
* AV - Selbsteinschaetzung
* richtig kodiert - Ablehnung ist hoch
tab V78
tab V84
tab V85
tab V86

* V79 V80 V81 V82 V83  // positive Wertung hat geringes Label
label list V79
* Umkodierung der positiv formulierten Variablen
generate V79r = .
replace V79r = 5 if V79==1
replace V79r = 4 if V79==2
replace V79r = 3 if V79==3
replace V79r = 2 if V79==4
replace V79r = 1 if V79==5

* Alternative
* recode V79 (1=5) (2=4) (4=2) (5=1), generate(V79r_2) // 3 bleibt identisch
```

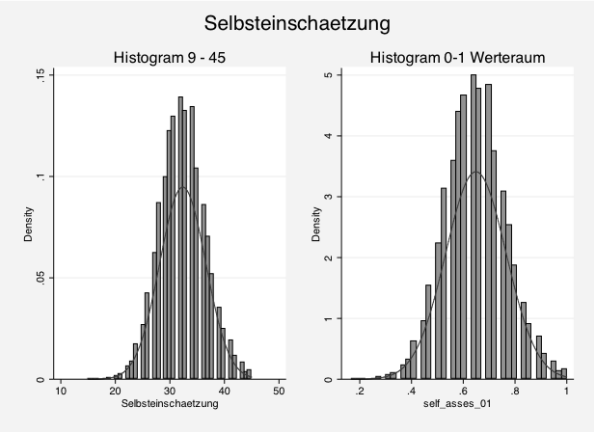
Es gibt mehrere Möglichkeiten Variablen zu rekodieren. Hier wird eine neue Variable erzeugt, die nur aus Missings besteht (`generate V79r = .`). Dann werden mit dem Befehl `replace` neue Kategorien eingefügt. Die "neue" Variable `V79r` enthält eine 5, wenn die "alte" Variable `V79` eine 1 hat, eine 4, wenn 2, usw... Ebenso kann man mit dem Befehl `recode` und der Option `generate` eine rekodierte Variable erstellen. Unabhängig von dem gewählten Befehl kann man mit einer `foreach` Schleife den gleichen Befehl für mehrere Variablen durchführen. In diesem Beispiel wird mit den übrigen positiv formulierten Variablen die gleiche Befehlsfolge durchgeführt, wie oben mit der Variable `V79r`, allerdings muss der Befehl nicht wiederholt werden, sondern wird für eine vorgegebenen Liste von Variablen (`varlist`) durchgeführt. Zu beachten sind die (!) Anführungszeichen um `‘var’`. Das vordere ist ein (‘) öffnendes Anführungszeichen, das hinter ein schliessendes (’) Anführungszeichen.

```
foreach var of varlist  V80 V81 V82 V83{
generate ‘var’r = .
replace ‘var’r = 5 if ‘var’==1
replace ‘var’r = 4 if ‘var’==2
replace ‘var’r = 3 if ‘var’==3
replace ‘var’r = 2 if ‘var’==4
```

```
replace ‘var’r = 1 if ‘var’==5
}

* Prüfe die Ergebnisse
tab V79
tab V79r

tab V80, mis
tab V80r, mis
* ok!
```

Nach der Umkodierung der Variablen kann man durch Summierung, der einzelnen Variablen eine quasi metrische Indexvariable erstellen. Der niedrigste Wert des Index ist die Anzahl der Variablen und der höchste Wert ist die Anzahl der Variablen multipliziert mit der Anzahl der Kategorien. Bei neun Variablen mit 5 Kategorien ist der Wertebereich $[min = 9; max = 45]$. Ebenso kann man mit einer relativ einfachen Rechenoperation den Wertebereich zwischen 0 und 1 normieren.

```
* Index erstellen Selbsteinschaetzung
generate self_asses = V78  + V84 + V85 + V86 + V79r + V80r + V81r + V82r + V83r
label var self_asses "Selbsteinschaetzung"
tab self_asses, mis

* Wertebereich 0-1
generate self_asses_01 = (self_asses - 9)/(9*5-9)
tab self_asses_01, mis
```



Die Verteilung kann man sich in einem Histogramm ausgeben lassen.

```
histogram self_asses, normal title("Histogram 9 - 45") name(index, replace)
histogram self_asses_01, normal title("Histogram 0-1 Werteraum") name(index_n, replace)
graph combine index index_n, title("Selbsteinschaetzung")
graph export self_asses.png, replace
```

2.1.3 Reliabilitätsanalyse (Cronbach's Alpha)

Die Reliabilitätsanalyse ist ein notwendiger Schritt bei der Indexkonstruktion. Geprüft wird, ob die ausgewählten Items eines Index tatsächlich Elemente einer gemeinsamen Skala sind. Diese wird gemessen, indem die Korrelation eines Items mit den Items der gesamten Skala ins Verhältnis gesetzt wird. Perfekte Konsistenz liegt bei einem Wert von 1 vor. In der Praxis gelten Werte ab 0.8 als ausreichend, aber auch Werte von 0.7 werden akzeptiert. Um die Reliabilität zu verbessern, können einzelne Variablen aus dem Index genommen werden.

```
* Cronbachs Alpha - Reliabilitätsanalyse
alpha V78 V84 V85 V86  V79r  V80r  V81r  V82r  V83r

Test scale = mean(unstandardized items)

Average interitem covariance:     .1116908
Number of items in the scale:            9
Scale reliability coefficient:      0.5088
```

In diesem Beispiel ist der Alpha-Wert mit 0.5088 sehr gering, wird aber in diesem Lehrbeispiel nicht weiter verändert.

2.1.4 Regressionsmodelle

Wenn alle Variablen generiert worden sind (wohl gemerkt in einem *Create*-Do-File) können nun im *Analys*-Do-File die Modelle getestet werden.

```
regress self_asses female edu_* // Stata entfernt den ersten Bildungsdummy

note: edu_1 omitted because of collinearity
      Source |       SS       df       MS              Number of obs =   23204
-------------+------------------------------           F(  7, 23196) =   37.90
       Model |  4635.44559     7  662.206512           Prob > F      =  0.0000
    Residual |  405299.998 23196  17.4728401           R-squared     =  0.0113
-------------+------------------------------           Adj R-squared =  0.0110
       Total |  409935.444 23203  17.6673466           Root MSE      =  4.1801

------------------------------------------------------------------------------
  self_asses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2502185   .0550658    -4.54   0.000    -.3581511   -.1422859
       edu_1 |  (omitted)
       edu_2 |   .5323037    .114256     4.66   0.000     .3083543    .7562531
```

```
       edu_3 |    .967245   .1053129     9.18   0.000     .7608247    1.173665
       edu_4 |   .8933325   .1031856     8.66   0.000     .6910818    1.095583
       edu_5 |   1.438627   .1058806    13.59   0.000     1.231094     1.64616
       edu_6 |   1.263674   .1101977    11.47   0.000     1.047679    1.479669
       edu_7 |   3.919645   1.324611     2.96   0.003      1.32332    6.515971
       _cons |   31.55551    .090891   347.18   0.000     31.37736    31.73366
------------------------------------------------------------------------------
```

Man sieht, dass alle Variablen signifikant sind und die Selbsteinschätzung von Frauen geringer ist und mit steigendem Bildungsniveau zunimmt. Mit 1% wird aber nur ein sehr geringer Teil der Variation erklärt. Das Programm hat einen Bildungsdummy automatisch gelöscht, da eine Variable "kollinear" ist – sie also edu_1 sein muss, wenn sie nicht edu_2 "bis" edu_7 ist. Das gleiche Ergebnis wird ebenso mit `regress self_asses female edu_2-edu_7` erzeugt. Das Minuszeichen bedeutet von Variable edu_2 "bis" edu_7.

2.1.5 Dummyregression mit Stata 11

Seit der Version Stata 11 gibt es die Möglichkeit, Dummyvariablen in der Regressionsgleichung als solche zu spezifizieren (vgl. (Cameron und Trivedi 2010: 9-11, 92-94)). Hierfür wird vor den Variablennamen ein `i.` gesetzt – der Befehl ist damit ähnlich dem in R (vgl. 2.2).

Eine Regressionsgleichung, die Dummies für Bildung (`DEGREE`) und Land aufnimmt (`V3`) sieht dann im Stata-Code folgendermassen aus:

```
regress self_asses female i.DEGREE i.V3, noheader
```

Im Output wird automatisch die Kategorie mit dem niedrigsten Label als Referenzkategorie verwendet. Im Fall von `DEGREE` ist das die Kategorie mit dem Label 0 (keine Bildung) und im Fall der Ländervariable (`V3`) ist es West-Deutschland das mit 2 gelabelt ist.

```
------------------------------------------------------------------------------
  self_asses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2515171   .0510483    -4.93   0.000    -.3515752    -.151459
             |
      DEGREE |
           1 |   .2131252   .1323271     1.61   0.107    -.0462448    .4724952
           2 |   .4256654   .1255092     3.39   0.001     .1796589    .6716718
           3 |   .4383357   .1235124     3.55   0.000     .1962431    .6804282
           4 |   .6494847   .1262768     5.14   0.000     .4019738    .8969957
           5 |     .71718   .1285521     5.58   0.000     .4652093    .9691507
           7 |   1.596446   1.234491     1.29   0.196     -.823238    4.016129
             |
          V3 |
           3 |   .0581063   .2029907     0.29   0.775    -.3397689    .4559815
           6 |   1.083605   .1611722     6.72   0.000     .7676963    1.399513
          10 |   1.832728   .1753091    10.45   0.000     1.489111    2.176346
          14 |  -2.416904    .167602   -14.42   0.000    -2.745415   -2.088393
          18 |  -1.319981   .1662446    -7.94   0.000    -1.645832   -.9941309
          19 |   .7430216   .1708071     4.35   0.000     .4082284    1.077815
          21 |   .7081417   .1659439     4.27   0.000     .3828806    1.033403
          22 |   .4962637   .1692931     2.93   0.003      .164438    .8280894
          24 |  -1.957861   .1849471   -10.59   0.000    -2.320369   -1.595352
          26 |  -.9460109   .2065142    -4.58   0.000    -1.350792   -.5412294
          28 |   .5398246   .1640838     3.29   0.001     .2182094    .8614397
          32 |   3.378251   .1650523    20.47   0.000     3.054738    3.701765
          33 |   1.606969   .1712831     9.38   0.000     1.271243    1.942696
          34 |   .0223446   .1699932     0.13   0.895    -.3108534    .3555425
          38 |  -.6157039   .1652258    -3.73   0.000    -.9395574   -.2918503
          39 |  -2.166462   .1488861   -14.55   0.000    -2.458289   -1.874635
          41 |  -2.724734    .158407   -17.20   0.000    -3.035222   -2.414245
          43 |  -.7867187    .155211    -5.07   0.000    -1.090943    -.482495
             |
       _cons |   32.27265   .1674266   192.76   0.000     31.94448    32.60081
------------------------------------------------------------------------------
```

2.1.6 Darstellung unterschiedlicher Modell mit dem Package estout

Um noch einmal den Nutzen des ado-Packages `estout` vor Augen zu führen werden drei Modelle mit unterschiedlichen Variablen berechnet und anschliessend übersichtlich in einer Tabelle gezeigt. Das Präfix `quietly` führt die Ergebnisse aus, zeigt dies aber nicht im `Stata`-Viewer. In diesem Fall werden die Ergebnisse lediglich mit `estimates store` *`name`* gespeichert und später mit dem Befehl `esttab` als Tabelle ausgegeben. Die Tabelle wird gleichzeitig als rtf-File im working-directory gespeichert und kann in ein Word-Dokument eingefügt werden.

```
* Zusammenfassung
quietly regress self_asses female
estimates store m1 // speichern der Ergebnisse als Modell 1
```

```
quietly regress self_asses female edu_2-edu_7
estimates store m2 // speichern der Ergebnisse als Modell 2

* Fuer Word
esttab m1 m2 using own_reg2.rtf, replace  r2 ar2 se label nonumber sfmt(a3)  obslast ///
        mtitles("Model 1" "Model 2" "Model 3")
```

Tabelle 1: Unterschiedliche Regressionsmodelle

	Model 1	Model 2
Geschlecht (1=weiblich)	-0.262***	-0.250***
	(0.0551)	(0.0551)
DEGREE==1. Lowest formal qualification		0.532***
		(0.114)
DEGREE==2. Above lowest qualification		0.967***
		(0.105)
DEGREE==3. Higher secondary completed		0.893***
		(0.103)
DEGREE==4. Above higher secondary level		1.439***
		(0.106)
DEGREE==5. University degree completed		1.264***
		(0.110)
DEGREE==7. CH: Other education		3.920**
		(1.325)
Constant	32.50***	31.56***
	(0.0402)	(0.0909)
R^2	0.001	0.011
Adjusted R^2	0.001	0.011
Observations	23391	23204

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.2 Regression in R

Der Abschnitt wird nur stichpunktartig auskommentiert.

- mit `library(foreign)` wird das package *foreign* geladen. Das wird benötigt, um einen Datensatz im Stata-Format zu öffnen.
- mit `read.dta` wird dann der Stata-Datensatz geöffnet und unter dem von mir gewählten Objektnamen `my.data` gespeichert.
- Die Regressionsanalyse wird mit `lm(dep ~ indep + indep )` ausgeführt, wobei vor der Tilde (∼), die im R-Output hochgestellt erscheint (˜) die abhänngige von den unabhängigen Variablen trennt. Als unabhängige Variablen werden die Dummy-Variablen `female`, `DEGREE` und `V3` aufgenommen. Spezifiziert wird auch, dass die Daten des Objekts *my.data* für die Analyse verwendet werden können. In R ist es im Gegensatz zu Stata möglich mit mehreren Datensätzen gleichzeitig zu arbeiten. Die Ergebnisse werden in dem Objekt *my.reg* gespeichert.
- mit dem Befehl `summary(objektname)` werden die Ergebnisse der Regressionsanalyse für *my.reg* ausgegeben.

```
# Package foreign zum Öffnen der Stata-Datei
library(foreign)

# Stata-Datensatz laden
my.data <- read.dta("../data/issp_2005_R.dta")
names(my.data)

# Regressionsberechnung
my.reg <- lm(self_asses ~ female + DEGREE + V3, data=my.data)
summary(my.reg)
```

```
lm(formula = self_asses ~ female + DEGREE + V3, data = my.data)
Residuals:
     Min       1Q   Median       3Q      Max
-17.3004  -2.5297  -0.0426   2.5142  14.7349
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         32.27265    0.16743 192.757  < 2e-16 ***
female1. weiblich                   -0.25152    0.05105  -4.927 8.41e-07 ***
DEGREE1. Lowest formal qualification  0.21313    0.13233   1.611 0.107282
DEGREE2. Above lowest qualification   0.42567    0.12551   3.392 0.000696 ***
DEGREE3. Higher secondary completed   0.43834    0.12351   3.549 0.000388 ***
DEGREE4. Above higher secondary level 0.64948    0.12628   5.143 2.72e-07 ***
DEGREE5. University degree completed  0.71718    0.12855   5.579 2.45e-08 ***
DEGREE7. CH: Other education          1.59645    1.23449   1.293 0.195954
V33. DE-E-Germany-East                0.05811    0.20299   0.286 0.774688
V36. US-United States                 1.08360    0.16117   6.723 1.82e-11 ***
V310. IE-Ireland                      1.83273    0.17531  10.454  < 2e-16 ***
V314. CZ-Czech Republic              -2.41690    0.16760 -14.420  < 2e-16 ***
```

```
V318. RU-Russia                    -1.31998    0.16624  -7.940 2.11e-15 ***
V319. NZ-New Zealand                0.74302    0.17081   4.350 1.37e-05 ***
V321. PH-Philippines                0.70814    0.16594   4.267 1.99e-05 ***
V322. IL-Israel                     0.49626    0.16929   2.931 0.003378 **
V324. JP-Japan                     -1.95786    0.18495 -10.586  < 2e-16 ***
V326. LV-Latvia                    -0.94601    0.20651  -4.581 4.65e-06 ***
V328. FR-France                     0.53982    0.16408   3.290 0.001004 **
V332. DK-Denmark                    3.37825    0.16505  20.468  < 2e-16 ***
V333. CH-Switzerland                1.60697    0.17128   9.382  < 2e-16 ***
V334. FLA-Flanders                  0.02234    0.16999   0.131 0.895425
V338. MX-Mexico                    -0.61570    0.16523  -3.726 0.000195 ***
V339. TW-Taiwan                    -2.16646    0.14889 -14.551  < 2e-16 ***
V341. KR-South Korea               -2.72473    0.15841 -17.201  < 2e-16 ***
V343. DO- Dominican republic       -0.78672    0.15521  -5.069 4.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.868 on 23178 degrees of freedom
  (190 observations deleted due to missingness)
Multiple R-squared: 0.154,Adjusted R-squared: 0.1531
F-statistic: 168.7 on 25 and 23178 DF,  p-value: < 2.2e-16
```

2.3 Modell der multiplen linearen Regression

Das multiple lineare Regressionsmodell mit drei Kovariablen (unabhängigen Variablen) hat folgende Form:

$$y_i \quad = \quad \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Diese Variablen können, wie im letzten Kapitel gezeigt, Dummy Variablen sein, ebenso sind auch kontinuierliche Variablen möglich und eben Interaktionen zwischen den Variablen.

2.3.1 Metrische Variablen

Bevor die Interaktioneffekte näher beschrieben werden, wollen wir metrische Variablen wie Alter, Bildungsjahre und Einkommen erstellen. Hierfür verwenden wir weiterhin den ISSP 2005 Datensatz aus Kapitel 2.

Die Variablen Alter und Bildungsjahre lassen sich schnell erstellen, hierfür wird der “Create”-Do-File verwendet.

```
use ../data/issp2005_1, clear // öffne Datensatz
```

```
describe, short

* Alter
generate age = AGE
label var age "Alter in Jahren"
tab age, mis
replace age =. if age <18 | age>80
histogram age
histogram age, normal by(V3) // Histogramm für jedes Land (V3)

* Bildungsjahre
generate yedu = EDUCYRS
label var yedu "Bildungsjahre"
tab yedu, mis

* Daten bereinigen fuer mindestens 1 und maximal 30 Bildungsjahre
replace yedu = . if EDUCYRS >30 | EDUCYRS <1
```

Mit `generate` werden die Variablen übernommen und mit `label var` benannt. Anschliessend wird mit `tabulate` eine Häufigkeitstabelle ausgegeben bzw. mit `histogram` ein Histogramm, um ggf. Extremwerte als Missing zu definieren. Im Fall der Bildungsjahre werden Personen, die mehr als 30 und weniger als 1 Bildungsjahr haben als Missing definiert. Das logische "oder"-Zeichen (|) ermöglicht Personen auszuschliessen, die entweder unter 18 Jahre oder über 80 Jahre alt sind.

Die Variable Einkommen (individuelles Nettoeinkommen pro Monat) ist ungleich schwerer zu bilden, da sie in den Teilnehmerländern des ISSP uneinheitlich erhoben wurde und in unterschiedlichen Landeswährungen vorliegt. Daher sollten in die Analysen nur Länder einfliessen, bei denen das Einkommen metrisch erhoben wurde (und nicht in Kategorien). Da es nicht entscheidend ist, für welches Land die Analysen gelten, werden hier die Schweiz, Russland und die USA berücksichtigt.
Zuerst werden die Variablen untersucht.

```
* Individualeinkommen
foreach var of varlist AU_RINC-ZA_RINC{
tab  'var', mis nolab
}
*  Kategorien 999990 kommen im Datensatz vor,
*  wenn Person nicht in Land (ist kein Missing)


* Individualeinkommen - Setze  Fälle 0, wenn nicht in Land
foreach var of varlist AU_RINC-ZA_RINC{
replace 'var'= 0 if 'var'== 999990
}
```

```
* Waehle nur drei Laender für die Analyse aus
* USA, Russland, Switzerland
* Addiere Ländervariablen zu einer Einkommensvariable

label list US_RINC
tab US_RINC
label list RU_RINC
tab RU_RINC
label list CH_RINC
tab CH_RINC

generate inc = US_RINC+RU_RINC+CH_RINC
bysort V3: tab inc, mis
tab inc, mis
* Missing, wenn Einkommen 0
replace inc=. if inc==0

* Label USA ansehen. Umkodieren
label list US_RINC // höchster Wert ist 150000
replace inc=150000 if inc==999996
```

Die Variabel Einkommen (`inc`) wird so gebildet, dass die individuellen Ländereinkommen aufaddiert werden. Allerdings stellt sich das Problem, dass die Einkommen nicht vergleichbar sind, da sie in Landeswährungen erhoben wurde, und sich der Schweizer Franken vom Russischen Rubel oder dem US-Dollar unterscheidet. Generell handelt es sich um den Vergleich von Variablen mit unterschiedlichen Masseinheiten. Die Lösung ist die Z-Standardisierung der Einkommensvariable pro Land. Dabei wird pro Land der Mittelwert und die Standardabweichung der Einkommensvariable gebildet. Anschliessend wird der Mittelwert pro Land vom beobachteten Einkommenswert subtrahiert und durch die Standardabweichung geteilt. Diese Variable hat die Eigenschaft, dass sie den Mittelwert von 0 hat und eine Standardabweichung von 1. Verwendet man die z-transformierte Variable dann als unabhängige Variable in einer Regressionsgleichung, so interpretiert man den Koeffizienten aus dem Regressionsoutput wie folgt: Die abhängige Variable ändert sich bei Erhöhung der Einkommensvariable um eine Standardabweichung um den Faktor β_{inc}.

$$\text{z-score} \quad = \quad \frac{\text{Variablenwert (score)} - \text{Mittelwert der Variable}}{\text{Standardabweichung der Variable}}$$

```
* Generiere Einkommensmittelwert pro Land
by V3, sort: egen inc_cmean = mean(inc)
* Generiere Standardabweichung für Einkommen pro Land
```

```
by V3, sort: egen inc_csd = sd(inc)

* z-Standardisierung
generate inc_z = (inc-inc_cmean)/inc_csd
```

Hilfreich ist der Befehl `egen`, wenn man Masszahlen aus metrischen Variablen erstelle möchte. Hier wird der Mittelwert (`mean(inc)`) und die Standardabweichung (`sd(inc)`) ($\sqrt{var(x_{inc})}$) der Variable Einkommen (`inc`) gebildet. Das Präfix `by V3, sort:` vor dem Befehl `egen` führt diesen Befehl separat für jedes Land (`V3`) aus.

Verwendet man den `by` Befehl nicht, so kann man auch die `egen` Funktion `std(`*`var`*`)` verwenden. In diesem Fall wurde aber pro Land die Z-Standardisierung vorgenommen, daher wird dieser Befehl hier nicht ausgeführt. Der Befehl würde dann lauten:

```
egen z_score = std(inc)
```

2.4 Multiple lineare Regression mit Interaktionseffekten

In diesem Regressionsbeispiel wird das Einkommen als abhängige Variable verwendet. Statt Bildung in Kategorien aufzunehmen, wird sie in Bildungsjahre aufgenommen. Ferner wird zum Geschlecht das Einkommen und das Alter hinzugenommen und eine Variable, die besagt, ob eine Person vollzeitbeschäftigt ist. Die Analysen geschehen im "Analyse"-Do-File.

Allerdings ist es vorher notwendig, metrische unabhängige Variablen um den Mittelwert zu zentrieren. Dies erleichtert die Interpretation der Effekte, da die Konstante sinnvoll interpretiert werden kann. Bei der Zentrierung wird der Mittelwert der Variable von jeder Beobachtung abgezogen. Die Konstante der Regressionsgleichung bezieht sich damit auf eine Person mit durchschnittlichem Alter und Bildungsjahren und nicht mehr auf eine Person mit 0 Jahren und 0 Bildungsjahren.

```
* Zentrieren der Variablen Alter und Bildungsjahre
* Alter
egen miss = rowmiss(female full age yedu inc_z)
summarize age if miss==0
generate cage =age - r(mean) if miss==0
label var cage "Alter (um Mittelwert zentriert)"


*Bildungsjahre
summarize yedu if miss==0
generate cyedu =yedu - r(mean) if miss==0
label var cyedu "Bildungsjahre (um Mittelwert zentriert)"
```

Der Befehl `egen rowmiss` bildet eine Variable (`miss`), die die Anzahl Missings einer Variablenlist (hier: `female full age yedu inc_z`) zählt. Wenn keine Missings vorliegt hat die Variable den Wert 0. Somit hat eine Person, die in der Regression berücksichtigt wird den Wert 0, wenn sie kein Missing hat. Nur für diese Personen werden die zentrierten Variablen gebildet.

Anschliessend kann die Regression durchgeführt werden.

```
regress inc_z female full cage cyedu
```

```
      Source |       SS       df       MS              Number of obs =    1437
-------------+------------------------------           F(  4,  1432) =   72.75
       Model |  43.0417256     4  10.7604314           Prob > F      =  0.0000
    Residual |  211.815571  1432  .147915902           R-squared     =  0.1689
-------------+------------------------------           Adj R-squared =  0.1666
       Total |  254.857297  1436  .177477226           Root MSE      =   .3846

------------------------------------------------------------------------------
       inc_z |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.1252553    .021219    -5.90   0.000    -.1668789   -.0836316
        full |   .2362104    .027713     8.52   0.000     .1818479    .2905728
         age |    .005648    .000891     6.34   0.000     .0039002    .0073958
        yedu |      .0295   .0026958    10.94   0.000     .0242118    .0347882
       _cons |  -.6717187   .0587893   -11.43   0.000     -.787041   -.5563963
------------------------------------------------------------------------------
```

Die Interpretation der Koeffizienten ist allerdings nicht intuitiv, da Veränderungen in den unabhängigen Variablen eine Veränderung der Standardabweichung des Einkommens bedeuten. Für Frauen verringert sich das Einkommen um 0.124 Standardabweichungen und mit jedem Bildungsjahr steigt das Einkommen um 0.028 Standardabweichungen.

Zum besseren Verständnis wird im Folgenden für die separaten Länder eine Regression in den Landeswährungen durchgeführt. Mit `estimates store` werden die Ergebnisse gespeichert und anschliessend in einer formatierten Regressionstabelle mit dem Befehl `esttab` ausgegeben.

```
label list V3 // Welches Label haben CH, RU, US
* Regressionsanalyse
regress inc female full cage cyedu if V3==33 // Schweiz
estimates store ch
regress inc female full cage cyedu if V3==18 // Russland
```

```
estimates store ru
regress inc female full cage cyedu if V3==6 // USA
estimates store us
*Fuer Word
esttab ch ru us using reg_country.rtf, replace  r2 ar2 se label nonumber ///
        title({\b Table 1: Regression pro Land})  ///
        mtitles("Schweiz" "Russland" "USA" )

*Fuer Latex
esttab ch ru us using reg_country.tex, replace  r2 ar2 se label nonumber ///
        title({Regression pro Land})  ///
        mtitles("Schweiz" "Russland" "USA")
```

Tabelle 2: Regression pro Land

	Schweiz	Russland	USA
Geschlecht (1=weiblich)	-1715.9	-3167.1***	-15155.7***
	(933.8)	(401.7)	(2020.3)
Beschaeftigungsgrad (vollzeit = 1)	2852.5**	2496.6*	19842.6***
	(1007.9)	(1004.6)	(2883.9)
Alter (um Mittelwert zentriert)	92.00**	-45.66**	513.6***
	(33.74)	(16.30)	(83.56)
Bildungsjahre (um Mittelwert zentriert)	452.1***	437.6***	4129.8***
	(116.7)	(76.11)	(336.3)
Constant	5233.9***	5795.9***	26954.7***
	(1068.0)	(1018.3)	(2911.8)
Observations	559	624	805
R^2	0.073	0.150	0.286
Adjusted R^2	0.066	0.144	0.282

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.4.1 Interaktionseffekte: Modellierung

Für die Analyse von Interaktionseffekten wollen wir uns auf die Schweiz fokussieren.

Grundsätzlich wird in einer Regression von additiven Effekten unterschiedlicher unabhängiger Variablen gesprochen.

$$y_i = \beta_0 + \beta_1 \text{female}_i + \beta_2 \text{full}_i + \beta_3 \text{cage}_i + \beta_4 \text{cyedu}_i + \varepsilon_i$$

Diese Annahme berücksichtigt nicht, dass der Effekt einer Variable vom Wert einer anderen abhängt. Beispielsweise kann man sich vorstellen, dass das Einkommen für Personen mit wenigen Bildungsjahren bei Männern wie Frauen relativ gleich ist und erst mit steigender Anzahl Ausbildungsjahre zunimmt. Es gibt demnach noch einen weiteren Effekt $\beta_5 \text{female}_i \times \text{cyedu}_i$ der Berücksichtigt werden sollte:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \text{female}_i + \beta_2 x_i + \beta_3 x_i + \beta_4 \text{cyedu}_i + \beta_5 \text{female}_i \times \text{cyedu}_i + \varepsilon_i && (11) \\ &= \beta_0 + (\beta_1 + \beta_5 \text{cyedu}_i) \text{female}_i + \beta_2 x_i + \beta_3 x_i + \beta_4 \text{cyedu}_i + \varepsilon_i && (12) \\ &= \beta_0 + \beta_1 \text{female}_i + \beta_2 x_i + \beta_3 x_i + (\beta_4 + \beta_5 \text{female}_i) \text{cyedu}_i + \varepsilon_i && (13) \end{aligned}$$

Der Interaktionseffekt verändert den Geschlechtereffekt, so dass dieser mit Anzahl Bildungsjahren im Vergleich zur Referenzkategorie ändert. Man spricht auch von einem "effect modifier" (Rabe-Hesketh und Skrondal 2008: 26).

In `Stata` werden Interaktionsterme gebildet, indem man die Variablen multipliziert.

```
* Geschlecht und Bildungsjahre
generate f_cyedu = female*cyedu
label var f_cyedu "Interaktionseffekt Geschlecht Bildungsjahre"
```

Anschliessend wird die Interaktionsvariable in das Regressionsmodell eingefügt.

```
. regress inc female full cage cyedu f_cyedu if V3==33 // Schweiz

      Source |       SS       df       MS              Number of obs =     559
-------------+------------------------------           F(  5,   553) =    9.62
       Model |  4.1027e+09     5   820541871           Prob > F      =  0.0000
    Residual |  4.7185e+10   553  85325792.4           R-squared     =  0.0800
-------------+------------------------------           Adj R-squared =  0.0717
       Total |  5.1288e+10   558  91913750.1           Root MSE      =  9237.2

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2458.843   996.2488    -2.47   0.014    -4415.738   -501.9485
        full |   2797.793    1005.17     2.78   0.006     823.3744    4772.212
        cage |   95.11549   33.66728     2.83   0.005     28.98409    161.2469
```

```
      cyedu |   657.8328   152.2269     4.32   0.000     358.8192    956.8465
   f_cyedu |  -495.1268   236.3667    -2.09   0.037    -959.4132   -30.84037
      _cons |   5565.162   1076.422     5.17   0.000     3450.786    7679.537
------------------------------------------------------------------------------
```

Tabelle 3: Interaktionseffekt

	Schweiz	Schweiz Interaktion
Geschlecht (1=weiblich)	-1715.9	-2458.8*
	(933.8)	(996.2)
Beschaeftigungsgrad (vollzeit = 1)	2852.5**	2797.8**
	(1007.9)	(1005.2)
Alter (um Mittelwert zentriert)	92.00**	95.12**
	(33.74)	(33.67)
Bildungsjahre (um Mittelwert zentriert)	452.1***	657.8***
	(116.7)	(152.2)
Interaktionseffekt Geschlecht Bildungsjahre		-495.1*
		(236.4)
Constant	5233.9***	5565.2***
	(1068.0)	(1076.4)
Observations	559	559
R^2	0.073	0.080
Adjusted R^2	0.066	0.072

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

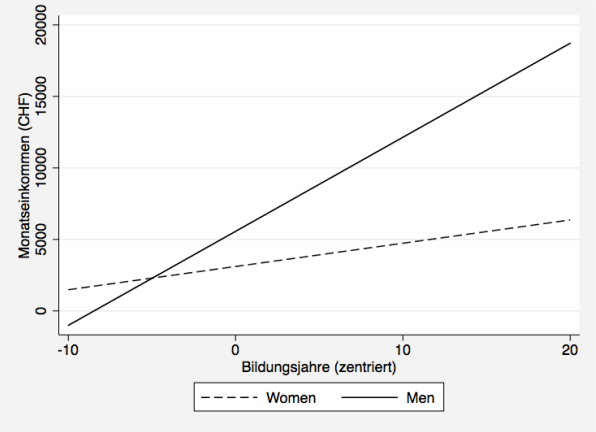
Man beachte, dass es auch seit der Version Stata 11.0 andere Möglichkeiten gibt, den Interaktionseffekt in Stata zu bilden. Hierfür wird für kategoriale Variablen ein `i.` und für kontinuierliche Variablen ein `c.` vor den Variablennamen gesetzt. Die beiden zu interagierenden Variablen werden dann mit einem ## verbunden. Dadurch werden die Haupteffekte und die Interaktioneffekte (verbunden durch ein #) gezeigt.

```
regress inc full cage i.female##c.cyedu if V3==33, noheader
------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        full |   2797.793    1005.17     2.78   0.006     823.3744    4772.212
        cage |   95.11549   33.66728     2.83   0.005     28.98409    161.2469
```

```
    1.female |  -2458.843   996.2488    -2.47   0.014    -4415.738   -501.9485
       cyedu |   657.8328   152.2269     4.32   0.000     358.8192    956.8465
             |
      female#|
    c.cyedu  |
           1 |  -495.1268   236.3667    -2.09   0.037    -959.4132   -30.84037
             |
       _cons |   5565.162   1076.422     5.17   0.000     3450.786    7679.537
------------------------------------------------------------------------------
```

Graphisch lässt sich der Interaktionseffekt gut veranschaulichen. Dabei werden allen nicht interessierenden Variablen konstant gehalten, so dass sich die Analysen auf teilzeitbeschäftigte Personen mit durchschnittlichem Alter beziehen.

```
quietly regress inc full cage i.female##c.cyedu if V3==33 // Schweiz
matrix list e(b) // Ausgabe der Variablen Labels
twoway  (function Women=_b[_cons] + _b[1.female] + (_b[cyedu]+ _b[1.female#cyedu])*x, ///
  range(-10 20) lpatt(dash)) ///
 (function Men=_b[_cons]  + _b[cyedu]*x, range(-10 20) lpatt(solid)), ///
 xtitle(Bildungsjahre (zentriert)) ytitle(Monatseinkommen (CHF))
 graph export f_cyedu.png, replace
```



2.4.2 Interaktionseffekt zwischen Dummy-Variablen

In unserem Modell kann man beispielsweise annehmen, dass für Männer ein höherer Lohn als für Frauen gezahlt wird, wenn sie vollzeitbeschäftigt sind. Die lässt sich durch einen Dummy-Interaktionseffekt testen.

```
 regress inc cage cyedu i.female##i.full if V3==33, noheader // Schweiz
------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cage |   91.52911   33.81454     2.71   0.007     25.10846    157.9498
       cyedu |    451.234   116.8217     3.86   0.000     221.7654    680.7025
    1.female |  -1305.695   1849.938    -0.71   0.481    -4939.459     2328.07
      1.full |   3232.594   1790.257     1.81   0.072    -283.9412    6749.129
             |
 female#full |
       1 1   |  -550.5675    2142.56    -0.26   0.797    -4759.119    3657.984
             |
       _cons |   4890.855   1710.141     2.86   0.004     1531.689    8250.021
------------------------------------------------------------------------------
```

Dieser Dummy-Interaktionseffekt wirkt in dem Regressionsmodell als zusätzlicher Effekt auf die Variable `full`, dann wenn es sich um eine vollzeitbeschäftigte Frau handelt: $\cdots + (\beta_{full} \times 1 + \beta_{female} \times 1)\text{full} \ldots$. Die Regressionskoeffizienten werden mit 1 multipliziert, da es sich bei `full` und `female` um Dummy-Variablen handelt. Der Effekt ist nicht signifikant, so dass es keinen Unterschied zwischen vollzeitbeschäftigten Frauen und Männern gibt.

2.4.3 Gemeinsame (Joint) Signifikanz

Die Signifikanz der Haupt- und Interaktionseffekte testet man nach der Regression mit dem postestimation Befehl `testparm` *`varlist`*.

```
. * Erweiterung des Interaktionseffekts
. regress inc cage  i.female##c.cyedu i.female##i.full if V3==33 // Schweiz

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cage |   94.62471    33.7438     2.80   0.005     28.34275    160.9067
    1.female |  -2026.647   1876.134    -1.08   0.281    -5711.883    1658.589
       cyedu |   657.0723   152.3802     4.31   0.000     357.7564    956.3882
             |
      female#|
     c.cyedu |
           1 |   -495.563   236.5703    -2.09   0.037    -960.2512   -30.87482
             |
      1.full |   3198.742   1784.871     1.79   0.074    -307.2269    6704.712
             |
 female#full |
```

```
       1 1  |  -580.9064   2136.075    -0.27   0.786    -4776.737    3614.924
            |
      _cons |   5203.516   1711.447     3.04   0.002     1841.772    8565.261
------------------------------------------------------------------------------

. * joint Effekt
. testparm  i.female##c.cyedu

 ( 1)  1.female = 0
 ( 2)  cyedu = 0
 ( 3)  1.female#c.cyedu = 0

       F(  3,   552) =    6.61
            Prob > F =    0.0002

. testparm i.female##i.full

 ( 1)  1.female = 0
 ( 2)  1.full = 0
 ( 3)  1.female#1.full = 0

       F(  3,   552) =    9.46
            Prob > F =    0.0000
```

Sowohl der Effekt für Frauen und Bildungsjahre (p-Wert $= 0.0002 < 0.05$), als auch der Effekt der Frauen und Vollzeitbeschäftigung (p-Wert $= 0.0000 < 0.05$) sind Signifikant auf dem 5 % Signifikanzniveau.

2.4.4 Interaktionseffekt zwischen kontinuierlichen Variablen

Interaktionseffekte von metrischen Variablen werden hingegen mit dem interessierenden Variablenwert multipliziert. In diesem Fall wird untersucht, ob der Effekt des Alters mit zusätzlichen Bildungsjahren zunimmt. Das kann durch unten angegebene Ergebnisse bestätigt werden. Pro Bildungsjahr erhöht sich der Alterseffekt um 22.95 CHF.

```
 regress inc female full c.cage##c.cyedu if V3==33, noheader
------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -1921.15   934.4972    -2.06   0.040    -3756.748   -85.55158
        full |   2635.119   1008.454     2.61   0.009     654.2502    4615.987
        cage |   127.6551   37.00905     3.45   0.001     54.95962    200.3506
       cyedu |   421.3597   116.9891     3.60   0.000     191.5624     651.157
```

```
             |
      c.cage#|
    c.cyedu  |   22.94648   9.974082     2.30   0.022     3.354758     42.5382
             |
       _cons |   5435.376   1067.477     5.09   0.000     3338.571    7532.181
------------------------------------------------------------------------------
```

$$\widehat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 \text{cage}_i + \beta_4 \text{cyedu}_i + \beta_5 \text{cage}_i \times \text{cyedu}_i + \varepsilon_i \quad (14)$$

$$= \beta_0 + \beta_1 x_i + \beta_2 x_i + (\beta_3 + \beta_5 \text{cyedu}_i)\text{cage}_i + \beta_4 \text{cyedu}_i + \varepsilon_i \quad (15)$$

$$= \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 \text{cage}_i + (\beta_4 + \beta_5 \text{cage}_i)\text{cyedu}_i + \varepsilon_i \quad (16)$$

Man kann auch testen, ob der Interaktionseffekt für Frauen gilt:

```
 regress inc  full c.cage##c.cyedu##i.female if V3==33 // Möglichkeit 2

      Source |       SS       df       MS              Number of obs =     559
-------------+------------------------------           F(  8,   550) =    7.70
       Model |  5.1632e+09     8   645403103           Prob > F      =  0.0000
    Residual |  4.6125e+10   550  83862995.9           R-squared     =  0.1007
-------------+------------------------------           Adj R-squared =  0.0876
       Total |  5.1288e+10   558  91913750.1           Root MSE      =  9157.7

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        full |   2429.957   1014.338     2.40   0.017     437.5072    4422.407
        cage |   205.3833   50.76093     4.05   0.000     105.6743    305.0923
       cyedu |   626.6662    151.391     4.14   0.000     329.2908    924.0416
             |
      c.cage#|
    c.cyedu  |   46.51146   13.52561     3.44   0.001     19.94328    73.07963
             |
    1.female |   -2625.13   995.1824    -2.64   0.009    -4579.954   -670.3071
             |
      female#|
      c.cage |
          1  |   -146.414   73.41646    -1.99   0.047     -290.625   -2.203069
             |
      female#|
    c.cyedu  |
          1  |  -459.6863   236.3886    -1.94   0.052    -924.0213    4.648766
```

```
              |
      female#|
       c.cage#|
      c.cyedu |
           1  |  -46.62852   19.81774    -2.35   0.019    -85.55625   -7.700792
              |
        _cons |   5920.586   1079.446     5.48   0.000     3800.246    8040.927
------------------------------------------------------------------------------
```

Der Interaktionseffekt für Frauen

```
female#|
  c.cage#|
 c.cyedu |
      1  |  -46.62852   19.81774    -2.35   0.019    -85.55625   -7.700792
```

zeigt, dass es für Frauen keinen Effekt gibt, da der Wert der Frauen 46.50 CHF niedriger als der Wert der Männer ist, dieser allerdings mit 46.50 CHF nahezu gleich ist und es für Frauen damit keinen Interaktionseffekt gibt.

Übersicht Stata-Befehle

Befehl	**Option**	**Erklärung**
`egen`	`rowmiss(`*varlist*`)`	Zählt für Variablen einer Variablenliste die Missings
`function`	`range(`*min max*`) lpatt()`	`function` erzeugt eine Linie, die aus mehreren Parametern bestehen. Mit `range()` kann die Dimension der x-Achse bestimmt werden. `lpatt()` definiert die Art der Linie (`dash solid dot`).
`testparm` *varlist*		Berechnet die gemeinsame Signifikanz der Haupt und Interaktionseffekte.
`foreach`		Eine Befehlsschleife, die einen Befehl für unterschiedliche Variablen (*varlist*) oder Zahlen (*numlist*) wiederholt.

Aufgabe

In der heutigen Aufgabe geht es darum, mir genaue Informationen zu den Daten eurer Fragestellung zu geben. Bitte recherchiert die Quelle. Meldet euch dafür auch bei

http://zacat.gesis.org/webview/ an und erstellt dort ein Benutzeraccount. Ihr werdet diese Datenquelle während eures Studiums noch öfter brauchen. Ich empfehle euch auch, die Daten der FORS in Lausanne anzusehen (http://www2.unil.ch/fors/spip.php?rubrique21&lang=de) und euch auf der Datenplattform FORS-Nestar (http://fors-nesstar.unil.ch/menu-d.jsp) anzumelden. Die Informationen sollten enthalten: Quellen der Daten (Website, Name der Umfrage), Zugang zu den Daten, wie viele Beobachtungen gibt es und Zeitpunkte. In einem nächsten Schritt müsst ihr dann prüfen, ob eine sinnvolle abhängige Variable erstellt werden kann.

3 14.3. – Regressionsdiagnostik

3.1 Kleinste-Quadrate Methode

Schauen wir uns die Regressionsgleichung für einen linearen Schätzer (Prädiktor) in einem Modell mit mehreren erklärenden Variablen nochmals genauer an (Cameron und Trivedi 2010: 82). Der Schätzer von Y ist ein bedingter Erwartungswert mehrerer Variablen X_1 bis X_k.

$$E(y|\mathbf{x}) = \mathbf{x}'\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{17}$$

Da wir es in der Realität aber nicht mit geschätzten Werten, sondern mit realen Werten zu tun haben, die nur in der Tendenz um einen imaginären "wahren Populationsmittelwert" streuen, gibt es für jede Beobachtung y_i eine Schätzgleichung, die einen Fehlerterm ε_i enthält.

$$y_i = \mathbf{x_1'}\beta + \varepsilon_i$$

Die Summer der quadrierten Residuen aller $i = 1, \ldots, N$ Beobachtungspukte wird mit der OLS-Methode (Ordinary Least Squares (OLS) = Kleineste Quadrate (KQ)) minimiert.

3.2 Annahmen des klassischen linearen Regressionsmodells

Das klassische lineare Regressionsmodell beruht auf Annahmen:

1. Es herrscht Linearität in den Parametern: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.
2. Der Mittelwert (die bedingte Erwartung) der zufälligen Störgrössen ε_i ist für alle Werte von X_i Null. Im Mittel verschwindet also der Fehler: $E(\varepsilon_i|X_i) = 0$. Was genau heisst das? Diese Annahme wird auch als "Exogenität der Regressoren" bezeichnet und bedeutet, dass der bedingte Mittelwert korrekt erzeugt wurde. Also es gibt einen linearen Zusammenhang zwischen y_i und den Einflussvariablen $\mathbf{X_i}$ und alle relevanten Variablen wurden in das Modell mit aufgenommen.
3. Die Varianz der Störgrösse ε_i ist für alle Beobachtungen von X gleich und konstant. Es existiert also Homoskedastizität, so dass $var(\varepsilon_i|X_i) = \sigma^2$ (man beachte, σ^2 ist unabhängig von i und daher konstant für alle i).

4. Für jedes gegebene Paar von Werten von X_i und X_j ist die Korrelation zwischen den zugehörigen Fehlern ε_i und ε_j stets Null. In diesem Fall gibt es keine Autokorrelation so dass die Kovarianz der Fehler Null ist: $cov(\varepsilon_i, \varepsilon_j) = 0$.
5. Regressor und Fehler sind unkorreliert, so dass es keine Kovarianz zwischen Störgrösse und Regressorwert gibt. Er darf also nicht mit unbeobachteten Variablen korrelieren und muss daher ohne Messfehler (systematischen Verzerrungen) erfasst worden sein.
6. Die Anzahl der Beobachtungen n muss grösser sein als die Zahl der zu schätzenden Parameter.
7. Bei Regressionsmodellen mit mehr als einer erklärenden Variable gibt es keine perfekte lineare Beziehung (Multikollinearität) zwischen den einzelnen Regressoren.

3.3 Gauss-Markov-Annahmen

Wenn die Annahmen des klassischen linearen Regressionsmodells gelten, dann besitzt der OLS-Schätzer bestimmt *statistische* Eigenschaften. Die statistischen Eigenschaften werden auch Gauss-Markov-Theorem genannt.

> Das **Gauss-Markov-Theorem** besagt, dass untern den Annahmen des klassischen linearen Regressionsmodells die OLS-Schätzer die minimale Varianz haben und **BLUE** (Best Linear Unbiased Estimators) sind.

Ein Schätzer $\hat{\beta}_1$ gilt als der beste lineare unverzerrte Schätzer des Populationsparamters β_1 falls gilt:

- Er ist eine lineare Funktion einer Zufallsvariable.
- Er ist unverzerrt, so dass sein Erwartungswert $E(\hat{\beta}_1)$ dem wahren Populationsparamter β_1 entspricht.
- Er ist ein effizienter Schätzer, da er die geringste Varianz aller anderen möglichen Schätzer hat. Gibt es Autokorrelation oder Heteroskedastizität, so sind ist der Schätzer nicht mehr effizient. In diesem Fall können feasible generalized least-squares (FGLS) Modelle angewendet werden (Cameron und Trivedi 2010: Kapitel 5).

Ein beliebtes Anfangsbeispiel stammt von Anscombe (1973) (vgl. auch Kohler und Kreuter 2008: 214). Dabei werden vier Regressionen ausgeführt, wobei die Koeffizienten, die Varianz der Residuen (RSS) und die erklärte Varianz (R^2) nahezu identisch sind.

```
use ../data/anscombe, clear
regress y1 x1
estimates store m1
regress y2 x2
```

```
estimates store m2
regress y3 x3
estimates store m3
regress y4 x4
estimates store m4

esttab m1 m2 m3 m4 using anscombe.tex, replace  scalars(rss) r2 se label  ///
        title({Anscombe Regressions})
```

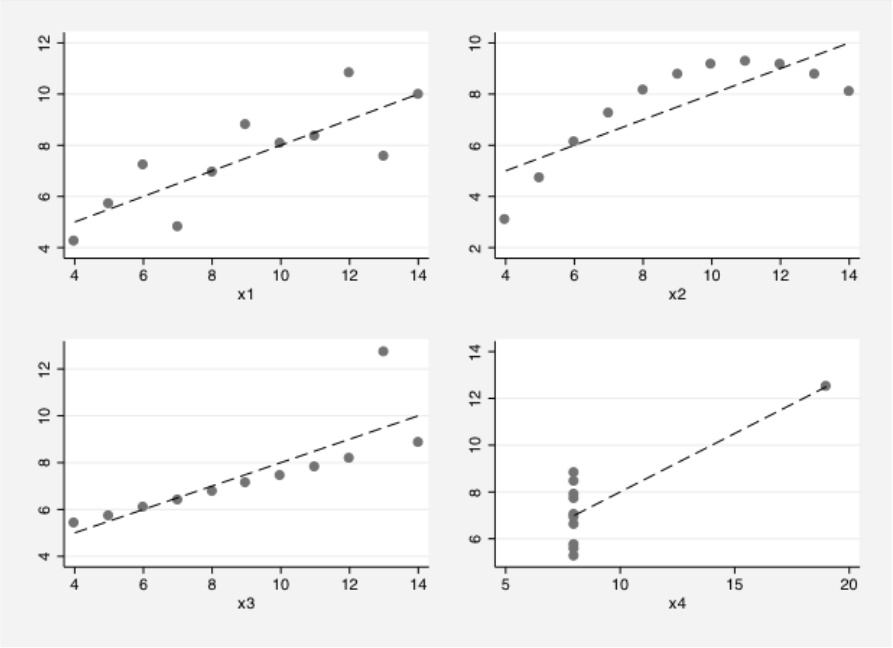
Tabelle 4: Anscombe Regressions

	(1) y1	(2) y2	(3) y3	(4) y4
x1	0.500**			
	(0.118)			
x2		0.500**		
		(0.118)		
x3			0.500**	
			(0.118)	
x4				0.500**
				(0.118)
Constant	3.000*	3.001*	3.002*	3.002*
	(1.125)	(1.125)	(1.124)	(1.124)
Observations	11	11	11	11
R^2	0.667	0.666	0.666	0.667
rss	13.76	13.78	13.76	13.74

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Allerdings liegen den Ergebnissen unterschiedliche Daten zu Grunde, wie folgende Graphik zeigt:



Welche Graphiken scheinen die Gauss-Markov-Annahmen zu verletzen?

- **x1** ist ein Beispiel in dem keine Verletzungen vorliegen,
- **x2** ist ein nicht-linearer Zusammenhang (umgekehrt U-förmig),
- **x3** hat einen einflussreichen Ausreisser und
- **x4** schliesslich hat keine Streuung der X-Werte bis auf einen Ausreisser, der alleine für die Steigung ($0.5 \times x4$) verantwortlich ist.

Aus der Regressionsstatistik werden die Verletzungen der Annahmen nicht ersichtlich, das $R^2 = 0.67$ deutet sogar auf eine hohe Modellgüte hin.

3.4 Der Fehlerterm $E(\varepsilon_i) = 0$

Der Fehlerterm ist normalverteilt mit einem Mittelwert von Null und einer konstanten Varianz σ^2 (gesprochen: "Sigma-Quadrat"). Formal heisst das: $\varepsilon_i \sim N(0, \sigma^2)$. Ob die Varianz konstant ist, werden wir am Ende des Kapitels besprechen, ebenso ob die Fehler unkorreliert sind. Zuerst wird die Annahme überprüft, ob der Mittelwert Null ist.

Verletzungen führen zu verzerrten Regressionskoeffizienten und liegen vor, wenn:

1. es einen nichtlinearen Zusammenhang zwischen der abhängigen und der unabhängigen Variable gibt,
2. Ausreisser das Regressionsmodell stark beeinflussen,
3. korrelierende Einflussfaktoren auf die übrigen unabhängigen Variablen übersehen wurden.

3.4.1 Korrelierende Einflussfaktoren

Der dritte Punkt meint Variablen, die einen Einfluss auf die abhängige Variable haben und mit mindestens einer unabhängigen Variable korrelieren. Dies ist ein theoretisches Problem! Zur Lösung könnte man weitere Einflussfaktoren gegen die abhängige Variable plotten, um mögliche Einflussfaktoren zu entdecken. Gleichzeitig sollte man aber auf "Multikollinearität" achten, also darauf ob eine Variable eine perfekte Linearkombination einer anderen Variable ist (z.B. $x_1 = 4 - x_2$).

Hat man ein Modell definiert, wird die Korrelation der Variablen untereinander getestet. Generell überprüft man die Korrelation der Variablen mit dem Befehl `correlate` oder `pwcorr` *`varlist`*`, sig`.

Das ado-package `corrtex` erzeugt wunderbare Korrelationsmatrizen für ein LaTeXSkript.[1]

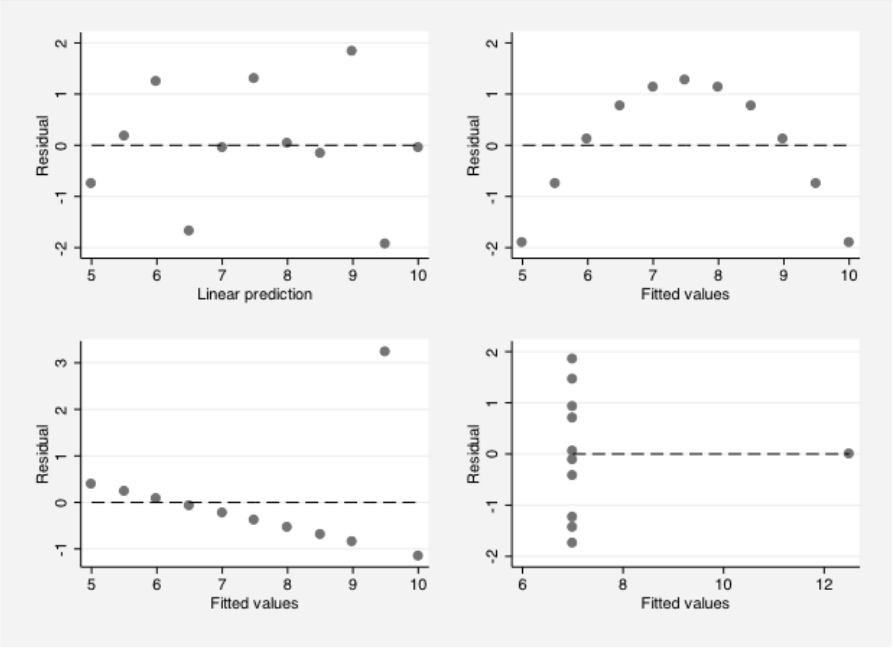
3.4.2 Graphische Prüfung der Residuen

Eine erste Möglichkeit ist die Residuen einer Regrssionsgleichung graphisch zu untersuchen. Hierfür wird nach dem Regressionsbefehl (`regress`) mit dem Befehl `predict` *`var`*`, resid` eine Variable mit den Wertend der Residuen für jede Beobachtung erzeugt. Mit `predict` *`yhat`*`, xb` können die geschätzen "fitted" Werte der abhängigen Variable erzeugt werden.

Mit beiden Werten wird ein "Residual-vs.-Fitted-Plot" gebildet, der die Streuung der Residuen um die Null-Achse zeigt. Dieser Plot stellt dar, ob es systematische Verzerrungen gibt oder ob die Residuen um den Mittelwert von Null gleichmässig streuen. Beispielhaft wird dies für das vierte Modell gezeigt. Mit `predict` werden die vorhergesagten Werte und die Residuen berechnet und in einem Scatterplot dargestellt. Alternativ kann man nach dem `regress` Befehl auch den Befehl `rvfplot` verwenden, der zu dem gleichen Ergebnis führt.

```
quietly regress y4 x4
predict yhat4
predict resid4, resid
twoway scatter  resid4 yhat4 || lfit  resid4 yhat4, ///
name(m4res, replace) legend(off) ytitle(Residual)
```

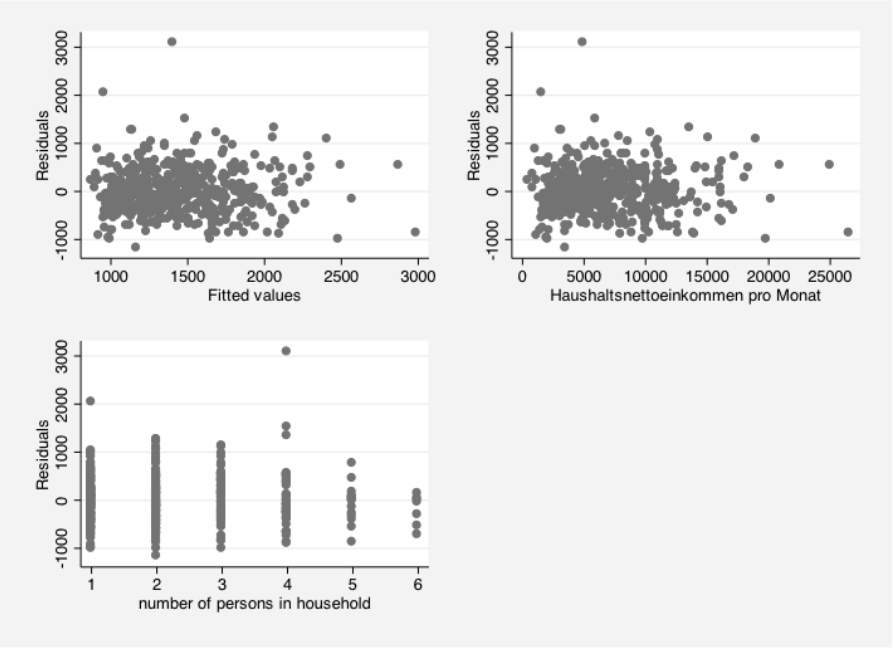
[1] Aufzurufen mit dem Befehl `findit corrtex`. Tabellen der Variablen mit weiteren Masszahlen können mit dem ado-package `sutex` erstellt werden.



Im bivariaten Fall unterscheidet sich diese Darstellung kaum vom Scatterplot. Der Vorteil des Residual-vs-Fitted-Plot ist, dass er auch für multiple Regressionsmodelle angewendet werden kann.

3.4.3 Beispiel mit dem SHP 2008 /Subsample

In diesem Abschnitt greife ich ein wenig vorweg und verwende die Daten des Schweizerischen Haushaltspanels 2008 (Subsample von 500 Fällen) um den Residual-vs.-Fittet Plot näher zu erläutern. Berechnet wird ein multiples Regressionsmodell mit drei Variablen. Gleichzeitig wird ein Resiudal-vs.Predictor-Plot gebildet, um zu sehen, ob die beobachteten Werte plausibel sind.



Es gibt wenige Personen, die trotz geringem Einkommen relativ viel Miete bezahlen. Diese Beobachtungen schauen wir uns genauer an.

```
* Residuum berechnen
predict uhat, residual
* Schaetzwert berechnen (y-Dach)
predict yhat, xb
gsort -uhat // Sortiere in absteigender Reihenfolge
list hhrent hhinc hhsize uhat yhat if uhat > 1000
```

	hhrent	hhinc	hhsize	uhat	yhat
1.	4500	4983.333	4	3091.296	1408.704
2.	3000	1566.667	1	2043.544	956.4558
3.	3000	6000	4	1513.713	1486.287
4.	3400	13633.33	4	1331.2	2068.8
5.	2400	3075	2	1264.602	1135.398
6.	2400	3200	2	1255.063	1144.937
7.	2900	10400	2	1205.619	1694.381
8.	2700	7858.333	3	1135.739	1564.261
9.	3160	15166.67	2	1101.867	2058.133
10.	3500	18950	3	1089.315	2410.685
11.	2800	11091.67	2	1052.837	1747.163
12.	2600	8566.667	2	1045.524	1554.476
13.	2300	5633.333	1	1033.21	1266.79

Tatsächlich könnte es sich bei den beiden ersten Fällen um nicht plausible Werte handeln, da die Miete im 2. Fall höher als das Einkommen liegt und im ersten Fall das Einkommen, abzüglich der Miete für einen Vierpersonenhaushalt nicht ausreichend sein dürfte. Alle anderen Fälle sind plausibel.

Wir greifen das Thema später bei der Ausreisserkontrolle noch einmal auf und untersuchen den Einfluss der Beobachtung auf den Schätzwert (leverage).

3.4.4 Linearität

Um den linearen Einfluss der unabhängigen Variablen zu prüfen, werden *nichtparametrische* Analyseverfahren verwendet. Damit ist gemeint, dass abschnittsweise der Einfluss einer unabhängigen Variable auf die abhängige Variable bestimmt wird. Die Aneinanderreihung der Abschnitte zeigt graphisch, ob es sich um einen linearen (also geraden) Einfluss handelt oder ob Abweichungen von der Geraden zu sehen sind. Zuerst wird ein Verfahren vorgestellt, wie man in einem Scatterplot die Linearität einer unabhängigen Variable testet, anschliessend wird noch gezeigt, wie man unter Kontrolle der übrigen unabhängigen Variablen die Linearität einer UV prüft.

Zuerst wird aus den Daten des Schweizerischen Haushaltspanel (SHP) 2008 eine Stichprobe von 500 Personen gezogen, die Wohnmiete (`hhrent`). Die Wohnmiete soll durch die Variablen monatliches Haushaltsnettoeinkommen (`hhinc`) und Anzahl der Haushaltsmitglieder (`hhsize`) erklärt werden. Das Beispiel orientiert sich an Kohler und Kreuter 2008: Kapitel 8.3.

Um aus einem Datensatz eine Stichprobe zu ziehen, wird der Befehl `sample` verwendet. Zuvor wird ein Startwert, ein "Seed", festgelegt (`set seed`), so dass zu einem späteren Zeitpunkt die selben 500 Fälle gezogen werden. Wird kein Seed definiert, wählt der Zufallsgenerator bei jeder neuen Berechnung unterschiedliche Fälle, was die Reproduzierbarkeit der Ergebnisse verunmöglicht.

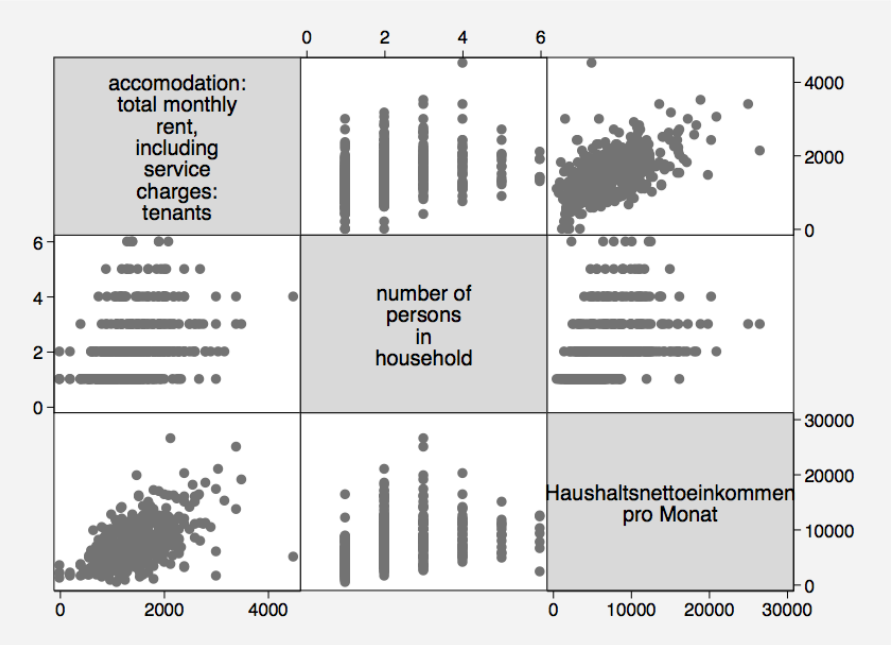
Für das Regressionsbeispiel werden die drei benötigten Variablen gebildet, wobei zum Kopieren bestehender Variablen der Befehl `clonevar` verwendet wird, der – wie der Name schon sagt – eine bestehende Variable mit allen Informationen übernimmt.

```
use ../data/SHP2008_1, clear

* HH Miete
clonevar hhrent = h08h36
* HH Groesse
clonevar hhsize = nbpers08
* HH Eikommen
generate hhinc = i08htyn/12
label var hhinc "Haushaltsnettoeinkommen pro Monat"
```

```
* Daten läschen, wenn Missing
drop if hhrent ==. | hhsize ==. | hhinc ==.
describe, short // 1900 Beobachtungen

* Setze Seed fuer Zufallszahl, um Ergebnisse zu reproduzieren
set seed 23032011  // Seed = Datum
sample 500, count  // 500 Fälle zufällig gezogen
describe, short // 500 Fälle
```

Nach der Datenvorbereitung können nun die Beziehung der Variablen untersucht werden. Einen ersten guten Überblick gibt ein Scatterplot aller verwendeter Variable, die Skatterplot-Matrix (`graph matrix`).

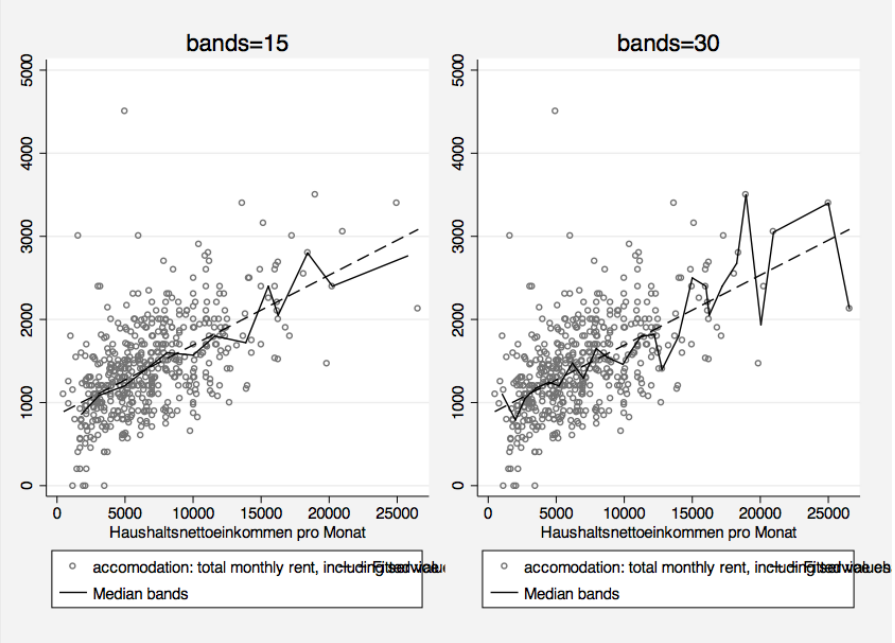
```
graph matrix hhrent hhsize hhinc
graph export scatter_matrix.png, replace // speicher Graph
```



In der ersten Zeile stehen die Scatterplots der abhängigen Variable und die beiden unabhängigen Variablen.

Um den Zusammenhang allerdings besser zu spezifizieren werden weitere statistische Hilfsmittel verwendet, insbesondere dann, wenn die Fallzahl so hoch ist, dass es optisch nicht mehr möglich ist, eindeutige Aussagen zu treffen. Hierfür werden wir einen "Scatterplot-Smoother" verwenden. Wir werden den "Median-Trace" und den "Median-Spline" kennenlernen, die jeweils den Median für definierte Abschnitte berechnen und mit einer Linie verbinden. Mit `mband` wird in der `twoway` Graphik in den Scatterplot ein Median-Trace eingefügt, mit `bands(15)` wird die x-Achse in hier 15 Abschnitte unterteilt. Je mehr Abschnitte, um so unruhiger die Linie.

```
* Bivariter Zusammenhang von Miete und HH-Einkommen
twoway  (scatter hhrent hhinc, ms(oh))  /// Scatterplot
(lfit hhrent hhinc)  /// Regressionsgerade
(mband hhrent hhinc, bands(15) clp(solid)), /// Median Trace
name(mband1, replace) title(bands=15)
```

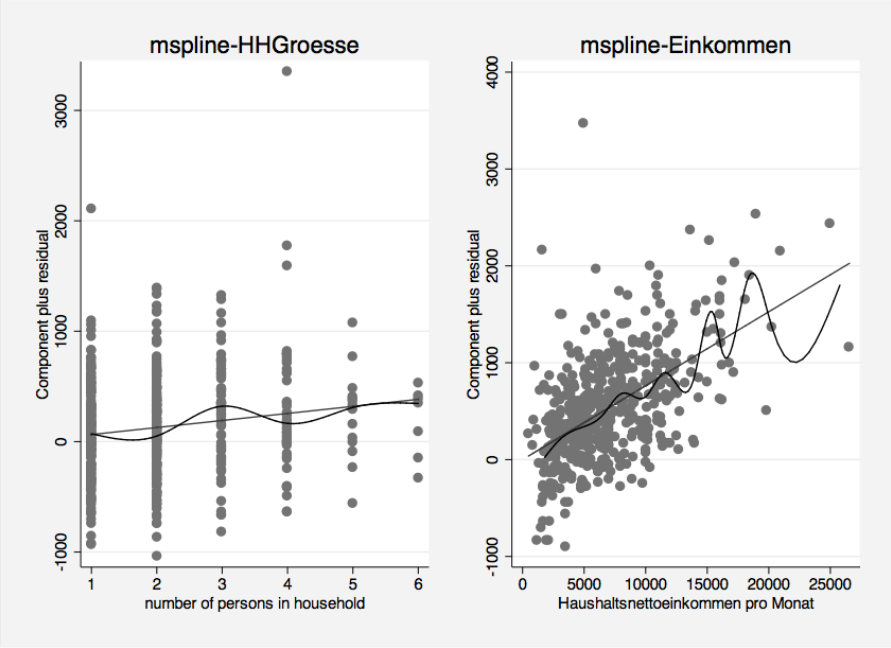


Optisch kann nun nach Ausreissern gesucht und der lineare Verlauf kontrolliert werden. Denkbar ist, dass mit steigendem Einkommen die Wohnungsmiete zwar steigt, die Steigung aber langsam abnimmt. Dieser Zusammenhang kann aber aus den Daten so nicht gelesen werden, da im nichtlinearen Median-Trace keine eindeutigen Abfall der Steigung zu erkennen ist.

Daneben muss man kontrollieren, ob der lineare Zusammenhang bestehen bleibt, wenn man die anderen Kovariablen kontrolliert. Es kann sein, dass sich unter Kontrolle der anderen Kovariablen der Zusammenhang verändert. Der "partielle Residuenplot" oder auch "Component-Plus-Residual-Plot" (`cprplot`) zeigt die funktionale Form des Zusammenhangs. Gebildet wird das Produkt aus Residuum und linearem Teil der unabhängigen Variable, welches dann gegen die restlichen unabhängigen Variablen geplottet wird. Hinter dem Befehl `cprplot` wird die uns interessierende unabhängige Variable genannt, sowie die Anzahl der Abschnitte des Median-Spline (`mspline`) definiert.

```
regress hhrent hhsize hhinc
* 1. Möglichkeit
predict e1, resid  // speichere Residuen
generate e1plus = e1 + _b[hhsize]*hhsize // addiere UV-Wert zu Residuum
twoway  (scatter e1plus hhsize, ms(oh)) ///
(mband e1plus hhsize, bands(15) clp(solid)) ///
```

```
(lfit e1plus hhsize), name(mband, replace) legend(off) title(mband)

* Alternative
cprplot hhsize, mspline msopts(bands(15))
  name(mspline_hhsize, replace) title(mspline)

graph combine mband mspline_hhsize
graph export cprplot_hhsize.png, replace
```

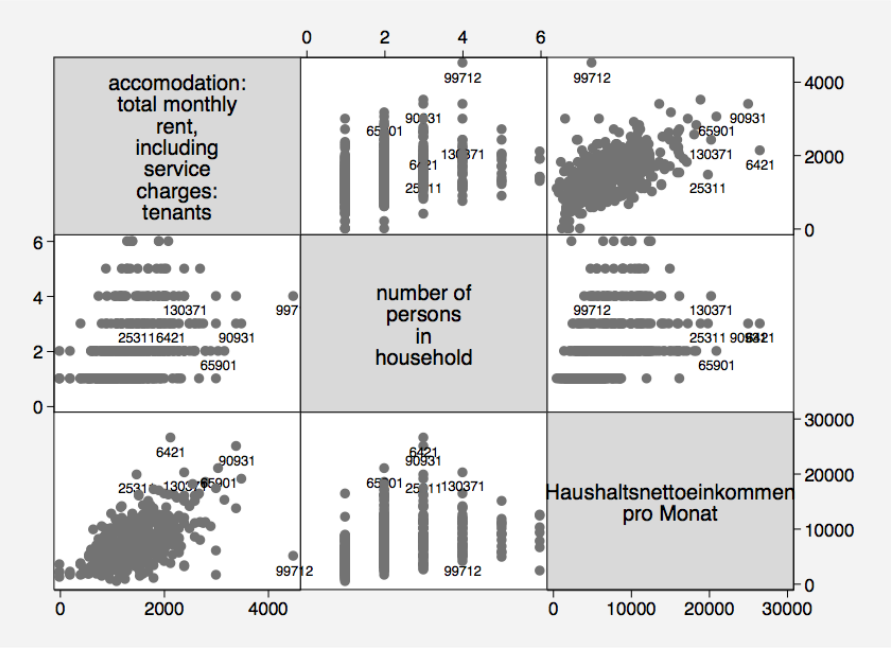


Die Variablen Haushaltseinkommen und Haushaltsgrösse scheinen tatsächlich einen linearen Einfluss zu haben und damit richtig spezifiziert zu sein.

3.4.5 Einflussreiche Beobachtungen - Aussreisser

Einflussreiche Beobachtungen lassen sich graphisch auf unterschiedliche Art feststellen. In einer Scatterplot-Matrix kann man sich die Befragtennummern ausgeben lassen. In diesem Fall für Haushalte mit eine Einkommen über 25'000 Franken oder einer Miete über 4'000 Franken.

```
generate str label= string(idhous08) if hhinc > 25000 | hhrent > 4000
graph matrix hhrent hhsize hhinc, mlab(label) mlabpos(6)
```

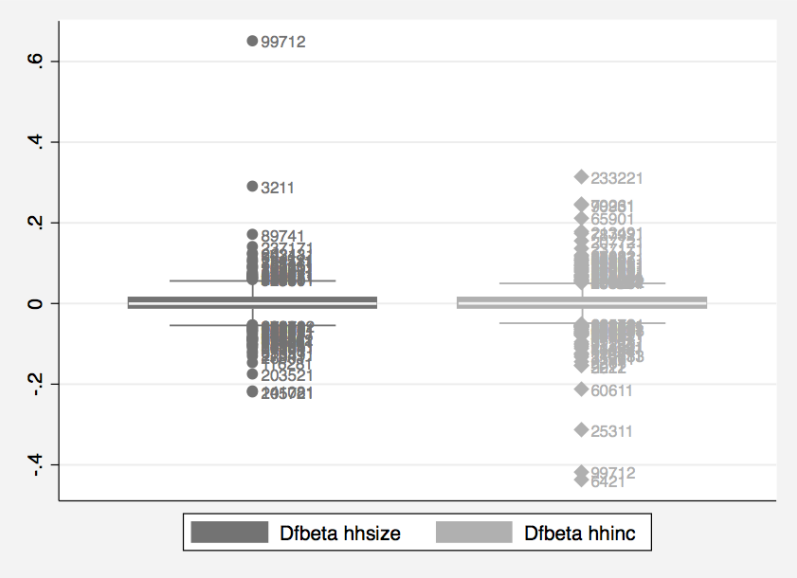


Formal lassen sich einflussreiche Fälle mit DFBETA bestimmen. Es wird ein Regressionsmodell mit allen Beobachtungen gerechnet. Dann werden Regressionsmodelle berechnet, wobei jeweils eine Beobachtung unberücksichtigt bleibt. Man beobachtete den Unterschied in den Koeffizienten und ermittelt so, ob die Beobachtung einen grossen Einfluss auf das Ergebnis hat. So erhalten wir für jede Beobachtungen einen DFBETA-Wert, der den Einfluss der Beobachtung für jeden Regressionskoeffizienten zeigt. Beobachtete wird die standardisierte Differenz zwischen dem Koeffizienten b_k und dem Koeffizienten $b_{k(i)}$ ohne der Beobachtung i. Mit dem Befehl `dfbeta` nach dem Regressionsbefehl, werden für jede Variable die DFBETA-Werte berechnet, die unter `_dfbeta_#` und einer Zahl # gespeichert werden.

```
quietly regress hhrent hhsize hhinc

  dfbeta
                        _dfbeta_1: dfbeta(hhsize)
                        _dfbeta_2: dfbeta(hhinc)
```

Ein Boxplot zeigt die DFBETA-Werte und lässt auf erste Ausreisser schliessen, die wir uns mit der Option `marker(box#, mlabel(var))` anzeigen lassen.

```
graph box _dfbeta_*, marker(1, mlabel(idhous08)) ///
 marker(2, mlabel(idhous08))
```




```
sort idhous08
foreach var of varlist _dfbeta_*{
list idhous08 ‘var’ if (abs(‘var’)> 2/sqrt(e(N))) & ‘var’<.
}
```

Die `foreach`-Schleife zeigt die Werte an, die einen absoluten DFBETA-Wert von $> 2/\sqrt{n}$ haben. Mit `abs()` werden die absoluten Werte betrachtet. In diesem Fall gibt es eine Vielzahl von einflussreichen Fällen.

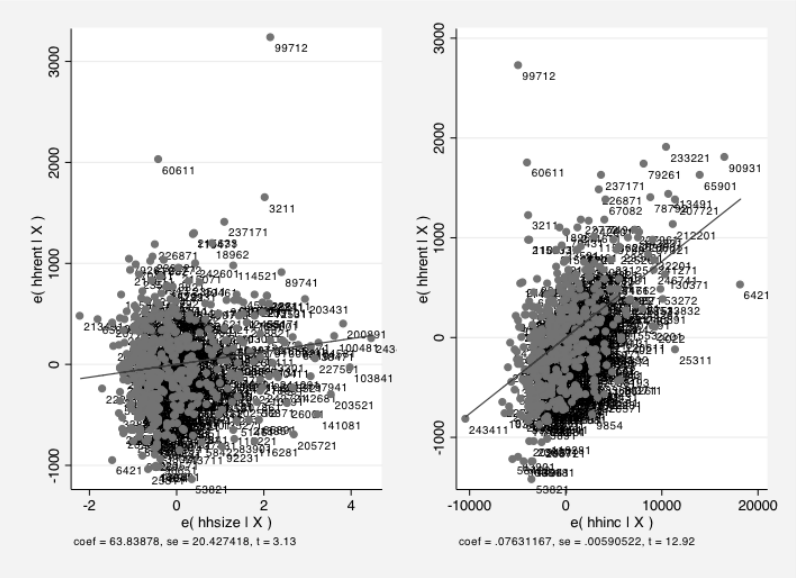
Der “Added-Variable-Plot” oder auch “Partieller-Regressionsplot” ist eine weitere Möglichkeit Ausreisser zu spezifizieren. Nach einer Regression kann dieser mit dem Befehl `avplots` für alle unabhängigen Variablen erstellt werden. Der AV-Plot besteht aus zwei Variablen. Auf der Y-Achse werden die Residuen des vollen Modells ohne einer interessierenden Variable X_i gebildet. Anschliessend wird eine zweite Regression gebildet, in der X_i aus alle weiteren unabhängigen Variablen geschätzt wird. Die Residuen dieser Schätzung ist der Anteil, der nicht durch andere Variablen erklärt wird. Die Residuen der beiden Schätzungen werden abgetragen. Diese Kombination hat einige nützliche Eigenschaften (vgl. (Fox 2008: 257)):

- Der Steigungskoeffizient der beiden Residuen entspricht dem OLS-Schätzer des vollen Modells.
- Die Residuen der bivariaten Regression entsprechen denen des vollen Modells.
- Wenn der Steigungskoeffizient signifikant ist, dann hat die Variable einen Einfluss, der über den Einfluss der anderen Variablen hinaus geht.

Die Residuen Die Steigungskoeffizient der unabhängigen Variablen entsprechen denen des vollen Modells. Damit ist ein Partieller-Regressionsplot noch immer ein Streudiagramm zwischen der abhängigen Variable und den unabhängigen Variablen, nur dass

die Variablen um den Teil, der durch die anderen Variablen erklärt wird, bereinigt wurden. Werte auf der X-Achse, die weit vom Mittelwert streuen, sind ungewöhnliche Werte.

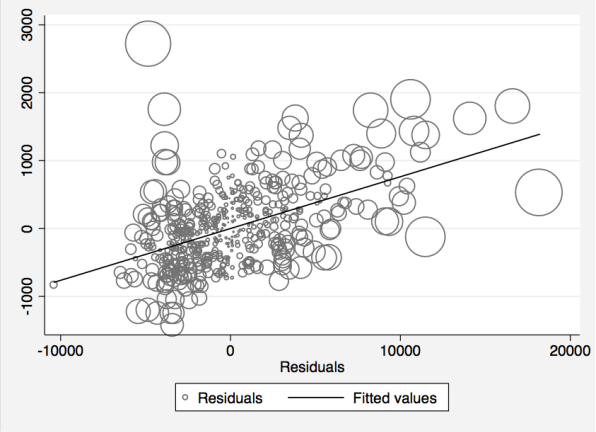
```
regress hhrent hhsize hhinc, noheader
avplots, mlabel(idhous08) mlabpos(5)
graph export avplots.png, replace
```



Weit aussen liegende Punkte haben einen grossen Einfluss und sind multivariate Ausreisser. Vergleicht man die Steigung der Schätzgeraden, so sieht man, dass unter Kontrolle der jeweils anderen Variablen die Einkommensvariable (rechts) einen stärkeren Einfluss auf die Wohnungsmiete hat, als die Variable Haushaltsgrösse (links).

Für das Haushaltseinkommen kann die Beobachtung noch mit dem DFBETA-Wert gewichtet werden. Die Gewichtung wird in der Plotsymbolgrösse dargestellt. An dieser Stelle wird auch deutlich, wie der Added-Variable-Plot erzeugt wird. Zuerst wird eine Regression ohne der interessierenden unabhängigen Variable (`hhinc`) berechnet und die Residuen gespeichert. Dann wird eine Regression von `hhinc` auf die übrigen unabhängigen Variablen – in diesem Fall nur `hhsize` – gerechnet und die Residuen gespeichert. Die Residuen der beiden Regressionen werden anschliessend geplottet.

```
regress hhrent hhsize
predict ehhrent, resid
regress hhinc hhsize
predict ehhinc, resid
generate absDF = abs(_dfbeta_2)
twoway  (sc ehhrent ehhinc [weight = absDF], msymbol(oh)) ///
  (lfit ehhrent ehhinc, clp(solid))
```



Unter Kohler und Kreuter 2008: 225ff und Jann 2009 können weitere Informationen zur Analyse von Ausreissern, insbesondere zu "Leverage" gefunden werden.

3.4.6 Leverage mit DFITS

Unter Leverage wird der Effekt verstanden, dass ein Datenpunkt eine grosse Hebelwirkung auf die Schätzgleichung hat und so die Schätzgerade an sich heranzieht (Jann 2009: 105). Hierfür wird in Stata der Befehl `dfits` (difference in fits) verwendet, der die Differenz der OLS-Vorhersage $(\widehat{y}_i - \widehat{y}_{(-i)})$ mit und ohne der i-ten Beobachtung misst. Als kritischer Wert kann $|dfits| > 2\sqrt{k/N}$ betrachtet werden (Cameron und Trivedi 2010: 97). Nach dem `regress` Befehl wird die Leverage mit dem Befehl `predict` *`newvar`*`, dfits` berechnet.

```
quietly regress hhrent hhsize hhinc, noheader
predict dfits, dfits
display e(df_m) // Konstante noch hinzurechnen (+1)
scalar schranke = 2*sqrt((e(df_m)+1)/e(N))
display "dfits Schranke = " %6.3f schranke

dfits Schranke =  0.155

tabstat dfits, stat(min p1 p5 p95 p99 max) format(%6.3f) col(stat)

    variable |       min        p1        p5       p95       p99       max
-------------+--------------------------------------------------------------
       dfits |    -0.453    -0.192    -0.121     0.120     0.276     0.737
----------------------------------------------------------------------------
```

Grob gesagt liegen über 2 % der Werte über dem kritischen Wert (aber unter 10 %), da `p1` und `p99` über dem kritischen Wert liegen, `p5` und `p95` allerdings nicht.

Nur sechs Beobachtungen liegen doppelt so hoch über dem absoluten `|dfits|` Schwellenwert. Personen die z.B. eine niedrige Miete haben aber ein hohes Einkommen beziehen oder ein geringes Einkommen haben und viel Miete bezahlen.

```
list dfits hhrent hhinc hhsize if abs(dfits) > 2*schranke, clean
```

	dfits	hhrent	hhinc	hhsize
1.	.7373407	4500	4983.333	4
2.	.3456871	3000	1566.667	1
3.	.3408907	3500	18950	3
4.	.3264144	3000	6000	4
499.	-.3383759	1473	19850	3
500.	-.4530247	2129	26550	3

Die beiden ersten Fälle sind ungewöhnliche Fälle und sollten von den Analysen ausgeschlossen werden. die anderen Beobachtungen sind plausibel und sollten in der Analyse weiterhin berücksichtigt werden.

3.4.7 Cook's D

Eine weitere Masszahl, um den Einfluss einer Beobachtung zu messen ist Cook's D. Die Masszahl bestimmt wie aussergewöhnlich die Kombination von X-Werten und Y-Werten ist. Gemessen wird im Gegensatz zu DFITS nicht mehr der Einfluss der Beobachtung auf den geschätzten Wert der abhängigen Variable, sonder der Einfluss auf die Veränderung auf die $\beta_{(-i)}$-Werte des Modells Der Einfluss einer Beobachtung ist nur dann hoch, wenn es sich um eine Kombination ungewöhnlicher Werte handelt. Eine Person mit hohem Einkommen und einer hohen Miete ist nicht ungewöhnlich. Eine Person mit hohem Einkommen, aber einer niedrigen Miete ist ungewöhnlich und umgekehrt.

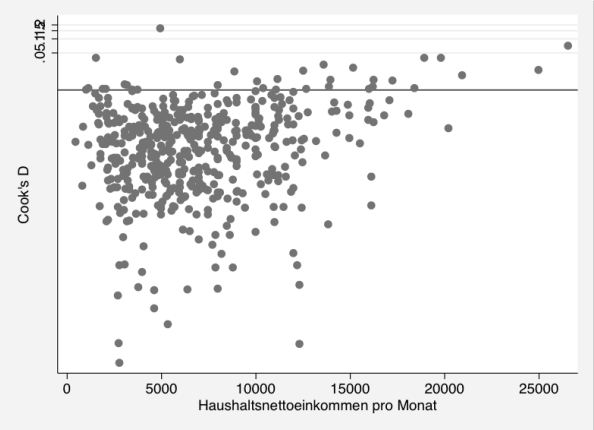
Kohler und Kreuter 2008: 226f bezeichnen den Einfluss der X-Variable als *Leverage* und den Einfluss der Y-Variable als *Diskrepanz*. Der Einfluss selber ist die Multiplikation aus Leverage und Diskrepanz. Ist einer der beiden Faktoren Null, so ist der Einfluss gering.

$$\text{Einfluss} = \text{Leverage} \times \text{Diskrepanz}$$

Cook's D-Werte messen nun den Einfluss einer Beobachtung auf das Regressionsmodell mit allen Variablen. Zur Erinnerung: DFBETA-Werte messen den Einfluss einer Beobachtung auf eine Variable. Erzeugt werden Cook's D-Werte mit dem `postestimation` Befehl `predict` *`newvar`*`, cooksd`. Werte über $4/n$ gelten als gross.

In unserem Beispiel werden Cook's D-Werte gebildet und in einem Scatterplott abgetragen.

```
quietly regress hhrent hhsize hhinc, noheader
predict cooksd, cooksd
* Schranke festlegen 4/n mit den gespeicherten Werten der Regression
* bilde ein globales Makro
global max = 4/e(N)

* Scatterplott, Y-Achse ist logarithmiert,
* die Schranke ist als horizontale Linie eingezeichnet
graph twoway scatter cooksd hhinc, yline($max) yscale(log)
graph export cooksd.png, replace
```



Es wird ersichtlich, dass tendenziell Fälle mit hohem Einkommen auch einen höheren Cook's D-Werte haben, was die Differenz der Mittelwerte für Fälle über und unter der Schranke zeigen.

```
generate bigcooksd = cooksd > $max

tabulate bigcook, summarize(hhinc)
            | Summary of Haushaltsnettoeinkommen
            |              pro Monat
  bigcooksd |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |    6704.078   3446.7829         470
          1 |     11702.5    7268.501          30
------------+------------------------------------
      Total |   7003.9833   3955.7988         500
```

Laut (Jann 2009: 109) unterscheiden sich DFITS und Cook's D nicht sonderlich, lediglich verwendet Cook's D standardisierte Residuen und DFITS studentisierte Residuen, daher sind DFITS sinnvoller.

3.4.8 Diskussion zu Ausreissern

Allgemein noch ein Wort zu den Lösungsmöglichkeiten bei einflussreichen Fällen. Generell sind einflussreiche Fälle eine Folge eines nicht korrekt spezifizierten Modells. Übersehen wird im Modell eine Variable, die den Zusammenhang beispielsweise niedriger Miete, hoher Haushaltsgrösse und geringem Einkommen erklärt – denkbar wäre dass die Variable Vermögen, Erbschaft, sonstige Einnahmequellen nicht richtig spezifiziert wurde. Rechtsschiefe Verteilungen wie häufig das Einkommen können transformiert (logarithmiert) werden. Wenn es sich um einflussreiche Fälle der abhängigen Variable handelt bietet sich eine Medianregression an.

Gerade in kleinen Datensätzen mit geringer Fallzahl sollte eine Ausreisserdiagnose gemacht werden, aber auch bei mittleren oder grösseren Datensätzen können einflussreiche Fälle einen Effekt signifikant werden lassen. Liegen Fälle über den besprochenen kritischen Werten, so heisst das aber noch nicht, dass sie auch Ausreisser sind und müssen genauer untersucht werden.

Findet man einen Ausreisser so gilt es wieder zu unterscheiden (vgl (Jann 2009: 122ff)):

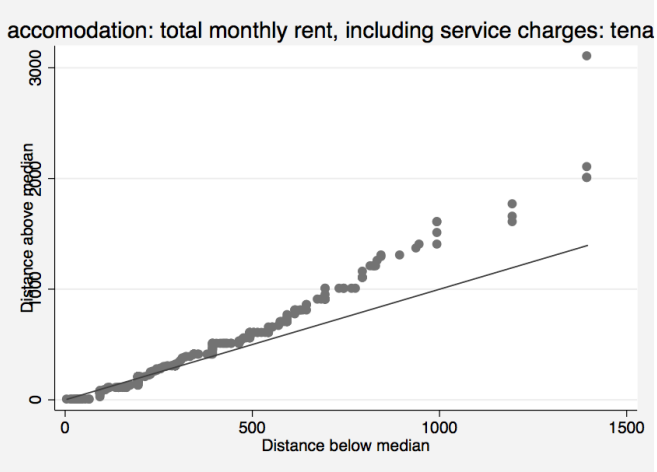
- Variablenwerte/ -kombination der Ausreisser tabellarisch ansehen,
- Messfehler entdecken und Variablen löschen,
- Modelle auf Robustheit testen. Effekte der Variablen mit und ohne Ausreisser rechnen. Instabile Modelle sollten gezeigt und diskutiert werden.
- Cluster (Gruppen) von Ausreissern identifizieren und prüfen, ob das Modell um übersehene Einflussfaktoren erweitert werden sollte. Die Erweiterung sollte allerdings nicht ad hock, sondern theoriebasiert geschehen.

3.5 Die Verletzung von VAR $(\varepsilon_i) = \sigma^2$

Die Annahme fordert, dass die Varianz der Fehler für alle Werte von X gleich sein soll. Wird diese Annahme erfüllt spricht man von "Homoskedastizität", wird sie verletzt spricht man von "Heteroskedastizität". Ist die Homoskedastizitätsannahme verletzt so sind die Schätzer nicht verzerrt, dies ist der Fall, wenn die $E(\varepsilon_i) = 0$ Annahme verletzt ist, sie sind aber nicht mehr effizient, da die Standardfehler der Koeffizienten nicht korrekt gebildet werden. Ursache sind meist nicht symmetrische Variablen, so dass man die abhängige Variablen erst noch transformieren muss.

Der "Symmetrieplot" untersucht die Symmetrie von Verteilungen. Ausgehend vom Median werden die nächst kleineren Werte gegeneinander geplottet. Werden die Abstände auf einer Seite überproportional gross, so weicht der Plot von der Geraden ab.

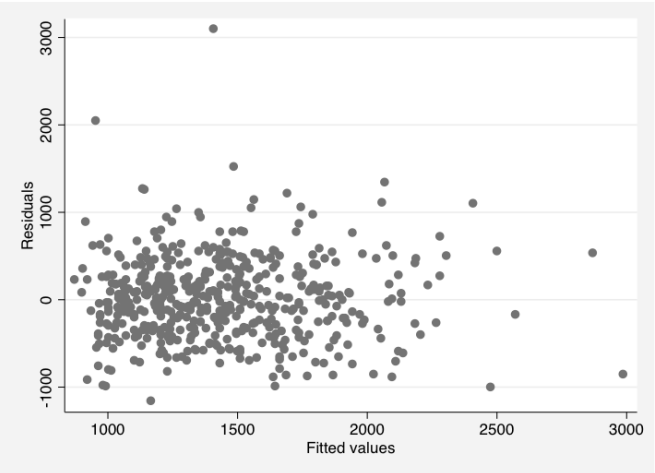
```
symplot hhrent
```



Dies deutet auf eine rechtsschiefe Verteilung hin, da die Abstände über dem Median grösser sind als unter dem Median. Im umgekehrten Fall würde es sich um eine linksschiefe Verteilung handeln.

Standardmässig wird der "Residual-vs.-Fitted-Plot" zur Prüfung der konstanten Streuung der Residuen verwendet.

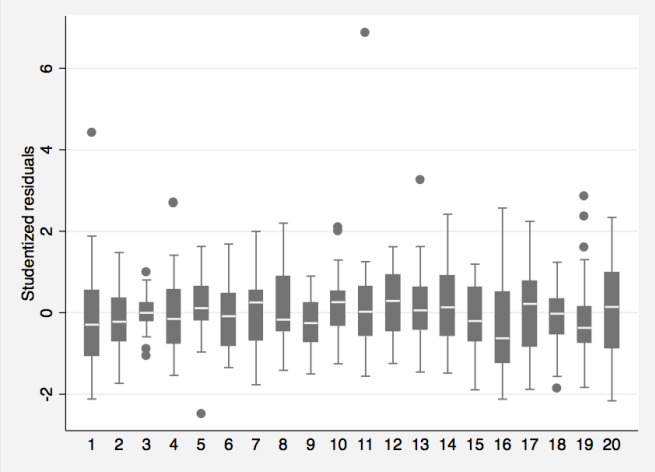
```
* Residual vs. Fitted Plot
quietly regress hhrent hhsize hhinc
rvfplot
```



Optisch lassen sich Ausreisser bestimmen, allerdings ist keine generelle Tendenz erkennbar. Daher wird eine genauere Methode verwendet, wobei hier für jeweils 25 Beobachtungen Boxplots erzeugt werden.

```
quietly regress hhrent hhsize hhinc
```

```
predict yhat_homo
predict rstud, rstud // studentisierte Residuen werden erzeugt


local groups = round(e(N)/25,1) // Runde Anzahl e(N)
// Hier Anzahl der Faelle pro Gruppe angeben (hier: 25)
xtile groups = yhat_homo, nq(‘groups’)
graph box rstud, over(groups)
```



Die Varianz der Residuen ist in den 20 Abschnitten nicht einheitlich. Allerdings gibt es auch keinen klaren Trend zu steigender oder abnehmender Varianz. Dies könnte auf Heteroskedastizität deuten.

Mit dem “Breusch-Pagan / Cook-Weisberg Test” wird statistische getestet, ob die Fehlervarianz über alle geschätzten Werte ($\widehat{y}_i$) identisch bleibt (H0= Nullhypothese) oder ob sie zu- oder abnimmt (H1 = Alternativhypothes). Hohe Chi^2-Werte deuten auf Heteroskedastizität hin. Bei einem p-Wert unter 0.05 kann die Nullhypothese abgelehnt werden und Heteroskedastizität liegt vor. Der p-Wert spricht allgemein dafür, wie glaubhaft die Nullhypothese bei gegebener Stichprobe ist. Ein kleiner p-Wert spricht demnach gegen die Nullhypothese. Bei einem p-Wert < 0.05 muss folglich die Nullhypothese der konstanten Varianz abgelehnt werden.

```
quietly regress hhrent hhsize hhinc

* postestimation hettest
estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of hhrent
```

```
chi2(1)      =     3.33
Prob > chi2  =   0.0682
```

In diesem Fall (p-Wert von $0.0682 > 0.05$) liegt keine Heterogenität vor, d.h. die Nullhypothese der Homoskedastizität kann nicht abgelehnt werden.

Wenn Heteroskedastizität vorliegt, kann man die Regression mit der Option `robust` oder `vce(robust)` berechnen, so dass robuste Standardfehler berechnet werden. Generell kann man aber immer auf Heteroskedastizität mit der Option `vce(robust)` kontrollieren.

Ein weiterer Test ist der White Test für Homoskedastizität. Der Test wird auch Omnibus Test genannt (Cameron und Trivedi 2010: 102). Er teste das gesamte Modell ob es homoskedastisch, symmetrisch und normale Wölbung (also keine heavy tails) hat. In unserem Fall kann die Symmetrieannahme (`Skewness`: p-Wert $= 0.0176 < 0.05$) abgelehnt werden. Der gesamte Test (`Total`) testet die Nullhypothese, dass es sich um normalverteilte Schätzergebnisse mit homoskedastischen Fehlern handelt. Die Nullhypothese muss bei dem P-Wert von 0.0149 abgelehnt werden. Die Zerlegung zeigt, dass die Schiefe der Verteilung nicht der einer Normalverteilung gleicht.

```
. estat imtest, white

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =      9.52
         Prob > chi2  =    0.0900

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       9.52      5    0.0900
            Skewness |       8.07      2    0.0176
            Kurtosis |       1.40      1    0.2366
---------------------+-----------------------------
               Total |      19.00      8    0.0149
---------------------------------------------------
```

3.6 Cluster-robuste Standardfehler

Wenn die Fehler unterschiedlicher Beobachtungen korreliert sind ($cov(\varepsilon_i, \varepsilon_j) = 0$), dann ist der Schätzer auch nicht mehr effizient. Beispielsweise trifft dies auf Zeitreihendaten zu, da die heutige Beobachtung, von der Beobachtung zu einem früheren Zeitpunkt abhängt. Ähnlich ist es mit Individuen bzw. Beobachtungen, die man zu Gruppen zusammenfügen kann. Personen in einem Haushalt, die Bevölkerung eines Landes etc. Die Fehler korrelieren dann innerhalb der Gruppe, sind aber zwischen den Gruppen unkorreliert. Cluser-robuste Standardfehler sollten dann angewendet werden, wenn man in seinem Regressionsmodell eine Variable aufgenommen hat, die für viele Befragten identisch ist. Beispielsweise bei einer internationalen Umfrage wie dem ISSP, das inzwischen in 45 Ländern jährlich mindestens 900 Personen interviewt, kann man pro Land Makrovariablen in die Regression aufnehmen, die Variablen sind für alle Personen eines Landes gleich, d.h. die Korrelation innerhalb des Landes ist 1 (z.B. Wirtschaftskraft, Anteil Stadt-Land Bevölkerung, Anteil von Religionsgruppen in einem Land). Die Kontrolle der Cluser führt zu unverzerrten Schätzern. In Stata erhält man mit dem Kommando `vce(cluster clustervar)` einen Cluster-robusten und Heteroskedastizität-robusten Standardfehler (Cameron und Trivedi 2010: 84f, Kapitel 5.5).

Beispiel aus eigener Forschung rechnen

Erst wird die Varianz der Variablen zwischen den Gruppen und innerhalb der Gruppen betrachtet. Ist die Varianz innerhalb der Gruppe gering (`within Std. Dev.`), sollten cluster-robuste Standardfehler verwendet werden. In diesem Fall sind die Variablen `gdppppt` und `urbanpop` für alle Gruppenmitglieder identisch. Mit `xtset` wird die Gruppenvariable definiert, anschliessend mit `xtsum` die Varianz innerhalb und zwischen den Gruppen berechnet.

```
. xtset weo_code
       panel variable:  weo_code (unbalanced)

. xtsum age1880 gdppppt urbanp,

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
age1880  overall |  43.49879   16.22768         18         80 |     N =   89543
         between |             4.158359    30.1487   49.42243 |     n =      51
         within  |             15.78151   12.07636   90.35009 | T-bar = 1755.75
                 |                                          |
gdppppt  overall |  17.27026    10.4446    .247284   38.98757 |     N =   91827
         between |             10.76034    .247284   38.98757 |     n =      51
         within  |                   0   17.27026   17.27026 | T-bar = 1800.53
                 |                                          |
urbanp   overall |  69.20337   15.57321         12        100 |     N =   91827
         between |             17.55216         12        100 |     n =      51
         within  |                   0   69.20337   69.20337 | T-bar = 1800.53
```

```
* Ohne cluster robuste Standardfehler
reg iub9 sex_female age1880   ///
  pop_dens urban_pop  gdp_1000

* Mit cluster robusten Standardfehler
reg iub9 sex_female age1880   ///
  pop_dens urban_pop  gdp_1000 , vce(cluster V4)
```

Tabelle 5: Cluster Robuste Modelle (eigenes Beispiel aus aktueller Forschungsarbeit)

	No Cluster	Cluster robust
Sex (1=female)	-0.51**	-0.51
	(0.18)	(0.34)
Age in years (18-80)	-0.13***	-0.13***
	(0.0058)	(0.022)
GDP (PPP) in 1000	-0.040***	-0.040
	(0.010)	(0.087)
Proportion urban pop.	0.0080	0.0080
	(0.0071)	(0.054)
Population density	0.29*	0.29
	(0.13)	(0.42)
Constant	56.1***	56.1***
	(0.50)	(3.56)
Observations	82354	82354

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Das Beispiel zeigt, wie sehr cluster-robuste Standardfehler die Ergebnisse beeinflussen. Die Standardfehler im unkontrollierten Modell sind viel kleiner und die Ergebnisse dadurch signifikant. Durch Kontrolle der Cluster verringern sich die Standardfehler und der Standardfehler berücksichtigt nun, grob gesagt, nur noch die Anzahl der Gruppen nicht mehr alle Individuen.

Aufgabe

- Forschungsfrage konkretisieren.
- Datenrecherche intensivieren.

Übersicht Stata-Befehle

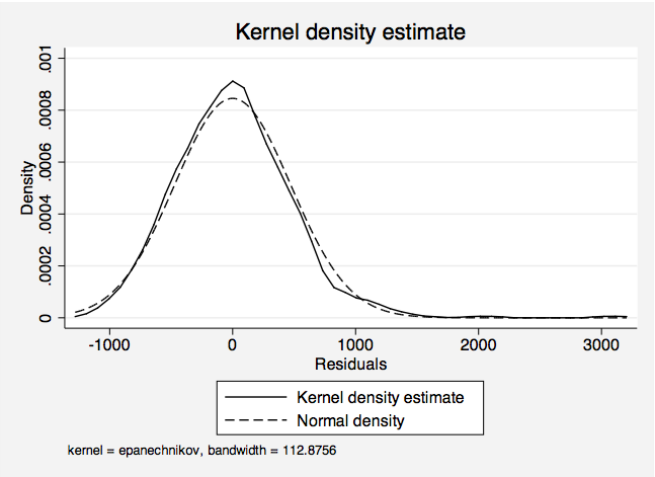
Befehl	Option	Erklärung
`clonevar` *newvar* = *oldvar*		Eine Variable mit neuem Namen wird gebildet, wobei Label und Kategorienbezeichnungen der alten Variable übernommen werden.
`set seed #`		Startwert für den Zufallsmechanismus definieren. Kann eine beliebige Zahl sein.
`sample #`	`count`	Zufallsvariabel aus Datensatz wird gezogen. Die Zahl gibt an wie viel Prozent gezogen werden. Die Option `count` ermöglicht, dass man statt Prozent, die exakte Anzahl der Fälle nennt. Man kann auch pro Kategorie die Anzahl der Fälle bestimmen, so dass man bspw. die gleiche Anzahl Männer und Frauen im Sample hat.
`graph matix` *var1 var2 ... varK*		Bivariater Scatterplot aller Variablen des Modells.
`mband` *var1 var2*	`bands(#)` `clp(solid)`	Nichtparametrischer Median-Trace. `band(#)` unterteilt die X-Achse in # Abschnitte, die mit einer durchgezogenen Linie (`clp(solid)`) verbunden werden.
`mspline`		Median-Spline, wie `mband` nur mit geglätteten Linien.
`dfbeta`		Postestimation, Ermittelt die DFBETA-Werte für alle Kovariablen.
`abs()`		Absolute Werte werden beachtet.
`symplot`		Zeigt die Schiefe einer Verteilung. Punkt über der Linie deuten auf eine rechtsschiefe Verteilung hin - unterhalb ist es ein linksschiefe Verteilung.
`regress`	`robust`	Berechnet robuste Standardfehler, wenn Heteroskedastizität vorliegt
`regress`	`vce(cluster` *clustervar*`)`	Cluster Robuste Standardfehler werden angewendet, wenn es Variablen gibt, die kaum Varianz (= Streuung) für Individuen einer Gruppe aufweisen.
`correlate`		Korrelationstabelle
`pwcorr` *varlist*	`sig`	Korrelationstabelle mit den p-Werten zum Test der Signifikanz
`predict` *newvar*	`dfits`	Berechnet den Leverage/Hebel einer Beobachtung.
`estat hettest`	`iid`	Breusch-Pagan Test für Homoskedastizität der Residuen. Die Option `iid` kann verwendet werden, wenn die Normalverteilungsannahme der Residuen nicht gewährleistet ist.

4 21.3. – Regressionsdiagnostik II

4.1 Normalverteilung der Residuen

Der Fehler eines Regressionsmodells ist normalverteilt und streut mit konstanter Varianz um den Mittelwert Null: $\varepsilon_i \sim N(0, \sigma)$. Die Normalverteilung der Fehler lässt sich einerseits graphisch veranschaulichen, wenn man den Kerndichteschätzer der Fehler mit dem einer Normalverteilung vergleicht.

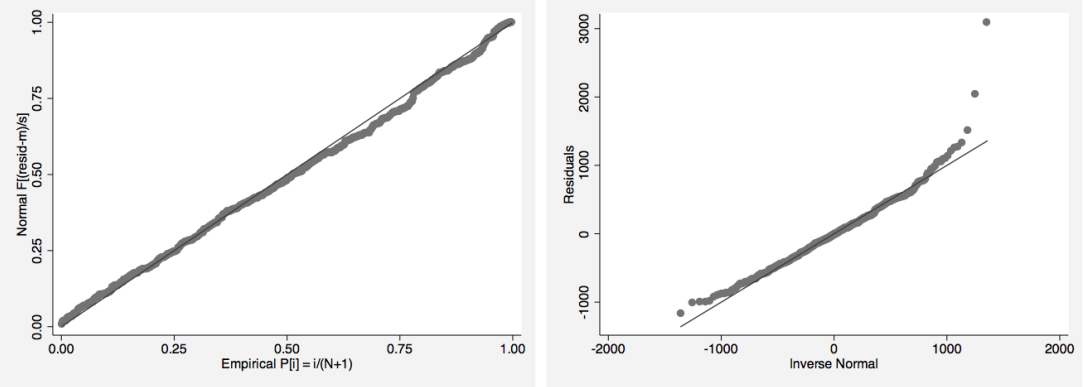
```
quietly regress hhrent hhsize hhinc
* speichern der Residuen
predict resid, resid

* Kerndichteschaetzer
kdensity resid, normal
graph export kdesity.png, replace
```



Die Kurve der Residuen folgt nicht eindeutig der Normalverteilung. Auch können wir sehen, dass die Kurve rechtsschief ist, da es grosse Streuung rechts von dem Mittelwert Null gibt.

Der `pnorm` Plot ist eine Möglichkeit die Normalverteilung der Residuen zu testen. Insbesondere wird getestet, wie die Verteilung in der Mitte einer Normalverteilung folgt. Der `qnomr` Plot hingegen testet die Normalität der Verteilung in den Enden.

```
* pnorm (Standardnormalverteilung)
pnorm resid
graph export pnorm.png, replace
*qnorm (Quantilsplot)
qnorm resid
graph export qnorm.png, replace
```



In diesem Beispiel ist die Normalverteilung in der Mitte (P-Plot links) und in den Enden (Q-Plot rechts) nicht exakt gegeben.

Numerische Tests der Normalverteilungsannahme sind zum einen der Shapiro-Wilk Test (`swilk`) und der Skewness/Kurtosis Test für Normalverteilung (`sktest`). Der Shapiro-Wilk test kann für Datensätze bist zu 2000 Fälle berechnet werden. Getestet wird jeweils die Nullhypothese, dass die Fehler normalverteilt sind. P-Werte < 0.05 lehnen die Nullhypothese ab und die Fehler sind nicht normalverteilt.

```
. swilk resid

                    Shapiro-Wilk W test for normal data

    Variable |    Obs       W           V         z       Prob>z
-------------+-------------------------------------------------
       resid |    500    0.96190     12.817     6.132    0.00000

. * Skewness and kurtois Test (Schiefe und Woelbung) fuer Normalitaet
. sktest resid

                     Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+------------------------------------------------------------------
       resid |    500      0.0000         0.0000            .          0.0000
```

Beide Tests lehnen die Nullhypothese der Normalverteilung ab, damit haben wir es mit nicht normalverteilten Fehler zu tun.

Als Lösung könnte man eine Ausreisseranalyse durchführen und einflussreiche Fälle löschen, bzw. die unabhängigen Variablen oder die abhängige Variable so transformieren, dass sie am ehesten einer Normalverteilung folgt.

4.2 Multikollinearität

Bestet zwischen zwei oder mehr Variablen perfekte Linearität, also bei Kenntnis des einen Wertes kann der Wert der zweiten Variable genau vorhergesagt werden, so lassen sich die Schätzer nicht eindeutig bilden. Die Standardfehler der Schätzer werden grösser und die Koeffizienten können sich unerwartete ändern oder das Vorzeichen wechseln.

Beispielsweise könnte Multikollinearität auftreten, wenn man mehrere Variablen in das Modell aufnimmt, die das gleiche Messen: Bildung der Eltern gemessen durch höchsten Schulabschluss, Bildungsjahre oder Abiturnote und weitere Abschlussnoten, ferner könnte das Einkommen und die Sparrate hoch korrelieren.

Mit dem postestimation Befehl `vif` (variance inflation factor) wird die Multikollinarität in einem Modell gemessen. Werte > 10 deuten darauf hin, dass die Variable eine Linearkombination der anderen unabhängigen Variablen ist. Neben dem VIF wird auch noch die Toleranz (1/vif) ausgegeben. Entsprechend deuten Werte < 0.1 auf Multikollinearität hin. Die Toleranz misst den Anteil der Varianz einer Variable, der nicht durch die anderen unabhängigen Variablen erklärt wird. In unserem Modell deutet der VIF-Wert nicht auf Multikollinearität hin.

```
quietly regress hhrent hhsize hhinc

vif

    Variable |       VIF       1/VIF
-------------+----------------------
       hhinc |      1.22    0.820933
      hhsize |      1.22    0.820933
-------------+----------------------
    Mean VIF |      1.22
```

Multikollinearität ist ein Datenproblem kein Theorieproblem. Mit höhere Fallzahl bekommt man geringere Standardfehler und genauere Schätzer. Ebenso können ähnliche Variablen zu Indizes oder neues Variablen zusammengefasst werden. Nicht alle Variablen eines multiplen Regressionsmodells sind von gleicher Bedeutung. Weisen die Kern-Variablen deiner Untersuchung keine Kollinearität auf, andere Variablen aber schon, so würde die Löschung einer Variable aus dem Modell, den Standardfehler der Kernvariablen verringern und den Effekt sogar noch signifikanter werden lassen (Baum 2006: 87). Daher ist Kollinearität nicht für alle Variablen eines Modells gleich problematisch.

4.2.1 Standardfehler eines Schätzers und Multikollinearität

Multikollinearität erhöht die Ungenauigkeit des geschätzten Koeffizienten einer Variable im multiplen Regressionsmodell. Warum? Schauen wir uns den Standardfehler eines Koeffizienten genauer an.

$$Var(B_j) = \frac{1}{1 - R_j^2} \times \frac{S_E^2}{\sum (x_{ij} - \bar{x}_j)^2}$$

mit $S_E = \sqrt{S_E^2}$ als Schätzer für die unbekannte Fehlervarianz σ_ϵ^2. Die Varianz wird geschätzt mit:

$$S_E^2 = \frac{\sum E_j^2}{n - k - 1}$$

R_j^2 ist die quadrierte multiple Korrelation der Regression von X_j auf alle anderen unbekannten Variablen in dem Modell. $1/1 - R_j^2$ ist der Variance-Inflation-Factor (VIF). Eine hohe Korrelation führt zu einem kleineren Nenner und damit einem höheren Standardabweichung ($\sqrt{Var(B)}$).

Fox (2008) schlägt folgende Interpretation des VIF vor. Er schlägt vor, die Wurzel des VIFs zu betrachten. Die Interpretation lautet dann: Die Wurzel des VIF sagt, um welchen Faktor sich die das Konfidenzintervall des Koeffizienten vergrössert in Relation zu unkorrelierten Variablen. Bei einem VIF-Wert von 4 folgt damit ein doppelt ($\sqrt{\text{VIF} = 4} = 2$) so grosses Konfidenzintervall als im unkorrelierten Fall.

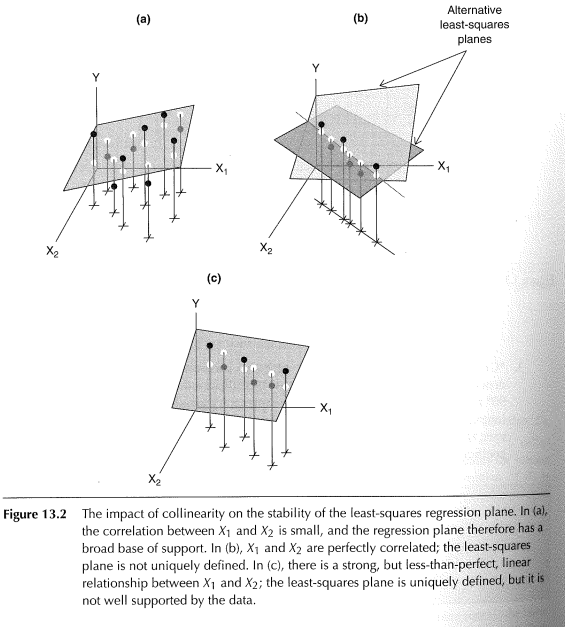


Figure 13.2 The impact of collinearity on the stability of the least-squares regression plane. In (a), the correlation between X_1 and X_2 is small, and the regression plane therefore has a broad base of support. In (b), X_1 and X_2 are perfectly correlated; the least-squares plane is not uniquely defined. In (c), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2; the least-squares plane is uniquely defined, but it is not well supported by the data.

Abbildung 1: Mulitkollinearität

Die Grafik ist entnommen aus (Fox 2008: 310) und zeigt die Ungenauigkeit in der Schätzung des Koeffizienten bei unterschiedlicher Korrelation von X_1 und X_2. (a) ist nicht korreliert und damit eine stabile Basis für die Schätzgeraden; (b) ist perfekt korreliert und damit die Punktewolke eine Gerade und ähnlich einer Wippe keine Basis für eine Fläche. (c) ist hoch korreliert und damit eine schwache Basis für eine Fläche. Stellen Sie sich die Veränderung der Residual Sum of Squares (RSS) vor, wenn Sie die Fläche leicht verändern. RSS wird sich nicht stark verändern, daher ist es schwer ein klares Minimum zu finden.

4.3 Linearität im Parameter

Eine der wichtigsten Annahmen der linearen Regressionsschätzung ist die Darstellung des Zusammenhangs von Y und X als eine linearer Zusammenhang (Linearkombination):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

So lässt sich der Zusammenhang von Y und X sehr einfach interpretieren: ändert sich X um eine Einheit, so ändert sich Y um β_1.

Problematisch sind Beziehungen zwischen zwei metrischen Variablen, die nicht linear sind. Kategoriale aber auch ordinale Variablen können durch Dummy-Variablen und damit ohne einer spezifischen Annahme über den Zusammenhang in ein Regressionsmodell integriert werden.

4.3.1 Simple monotone Transformation

Zusammenhänge, die sich nach innen oder aussen gewölbt zeigen, können durch einfache Transformation der Y oder der X Variable linear werden. Die Potenz einer Variable (die Power) wird dabei verändert. eine nicht transformierte Variabel hat die Potenz von 1:

- $X = X^1$

Höherer Power wäre dann:

- X^2
- X^3

Niedrigerer Power – also eine Verringerung der Potenz führt zu folgenden Transformationen:

- $\sqrt{X} = X^{\frac{1}{2}}$

- $log(X) = X^0$
- $\frac{1}{\sqrt{X}} = X^{-\frac{1}{2}}$
- $\frac{1}{X} = X^{-1}$
- $\frac{1}{X^2} = X^{-2}$

Je nach Wölbung kann man sowohl Y als auch X gemäss Abb 2 transformieren.

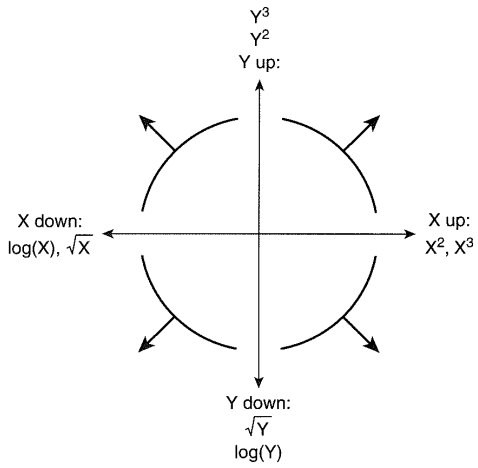


Figure 4.7 Tukey and Mosteller's bulging rule: The direction of the bulge indicates the direction of the power transformation of *Y* and/or *X* to straighten the relationship between them.

Abbildung 2: Transformation nach (Fox 2008: 60)

Beispielsweise eine Variabel mit hohen Werten für kleine X und schnell abnehmenden Werten mit steigendem X (Kindersterblichkeit je nach Wirtschaftskraft eines Landes) könnte durch Logarithmierung der Y als auch der X Variable linear transformiert werden.

4.3.2 Die Power-Leiter `ladder`

Eine weitere Möglichkeit die Variablen zu Prüfen ist der Befehl `ladder`. Zum einen wird eine Variable in unterschiedliche Power-Richtungen transformiert und dann die Schiefe und Wölbung auf Normalverteilung getestet.

```
Transformation         formula                chi2(2)       P(chi2)
-------------------------------------------------------------------
cubic                  hhrent^3                    .        0.000
```

square	hhrent^2	.	0.000
identity	hhrent	51.62	0.000
square root	sqrt(hhrent)	51.53	0.000
log	log(hhrent)	.	.
1/(square root)	1/sqrt(hhrent)	.	.
inverse	1/hhrent	.	.
1/square	1/(hhrent^2)	.	.
1/cubic	1/(hhrent^3)	.	.

Alle Modelle weisen die H0 der Normalverteilung zurück. Punkte (.) sagen, dass der Chi^2-Wert sehr hoch ist und daher nicht angezeigt wird. Warum die beiden Variablen nicht normalverteilt sind, wird durch den Test nicht ersichtlich. Daher sollte der Befehl `qladder` verwendet werden, der einen Quantil-Plot zeigt.

```
qladder hhrent
```

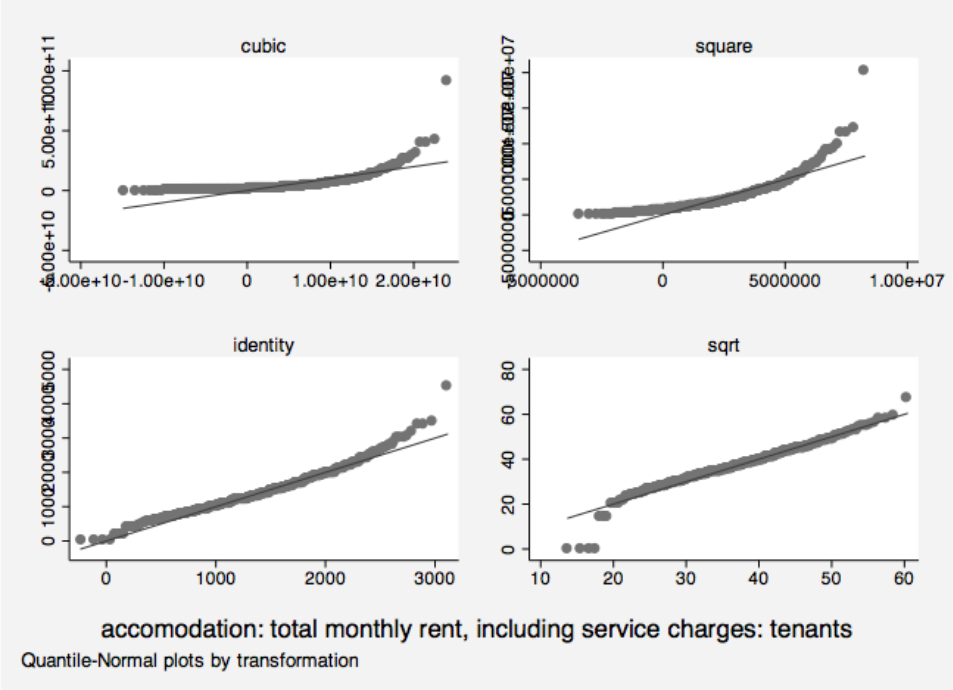


Abbildung 3: qladder: Quantilsplot der transformierten Variablen

Wir sehen, bei der `identiy` Verteilung, gibt es abweichende Beobachtungen an beiden Enden. Die Abweichungen bei der Wurzel (`sqrt`) sind dagegen nur am unteren Ende deutlich. Wo genau die Abweichungen von der Normalverteilung sind, wird durch den Test mit `ladder` nicht deutlich.

4.3.3 Box-Cox Transformation

Es gibt Schätzmethoden, die basierend auf Maximum Likelihood (ML) Schätzung, den Power einer Variable bestimmen. Die Box-Cox Transformation bestimmt die Transformation der abhängigen Variable Y so, dass Normalitätsannahme und die Annahme der konstanten Fehlervarianz verbessert wird. Dabei wird ein Transformationsparameter θ geschätzt. Für weitere Details der Schätzung siehe (Fox 2008: 292-294)

- $\theta = 1$ keine Transformation notwendig
- alle Y_i müssen > 0 sein.

Am Beispiel der Schätzung der Haushaltsmiete kann die Box-Cox Transformation getestet werden. Mit `boxcox` *`depvar indepvars`* wird das Modell spezifiziert. Die Option `lhsonly` (left-hand-side only) führt die Transformation nur für die abhängige Variabel aus. Zu beachten ist, das die Transformation nur für Werte > 0 gilt. Mit der ML-Schätzung wird der Parameter `/theta` geschätzt. Wenn der Wert signifikant ist, dann ist der Einfluss von 0 verschieden und damit die Potenz von 1 unterschiedlich. In diesem Fall wird eine Transformation von 0.49 vorgeschlagen.

```
 boxcox hhrent hhinc hhsize if hhrent>0, model(lhsonly) lrtest

                                                  Number of obs   =        496
                                                  LR chi2(2)      =     198.53
Log likelihood = -3727.1154                       Prob > chi2     =      0.000

------------------------------------------------------------------------------
      hhrent |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .4971829   .0729596     6.81   0.000     .3541847    .6401811
------------------------------------------------------------------------------

Estimates of scale-variant parameters
-------------------------------------------------------------
             |      Coef.  chi2(df)  P>chi2(df)    df of chi2
-------------+-----------------------------------------------
Notrans      |
       hhinc |   .0018625   135.192    0.000          1
      hhsize |   1.705375    10.892    0.001          1
       _cons |   54.96337
-------------+-----------------------------------------------
      /sigma |   11.88404
-------------------------------------------------------------

-----------------------------------------------------------
   Test         Restricted     LR statistic      P-value
    H0:       log likelihood       chi2       Prob > chi2
```

```
-------------------------------------------------------
theta = -1       -4004.294       554.36           0.000
theta =  0      -3752.4115        50.59           0.000
theta =  1       -3749.285        44.34           0.000
-------------------------------------------------------
```

Die transformierte Variable hängt signifikant positiv mit Haushaltsgrösse und Haushaltseinkommen zusammen.

Ein Likelihood-Ratio test (wird später noch besprochen) testet, ob ein `theta` von -1, 0 oder 1 bessere Modelle, als die Transformation von 0.49 schätzt, also das 0.49-Modell sich nicht von den anderen Modellen unterscheidet. Diese Hypothese kann abgelehnt werden, da alle P-Werte < 0.05 sind.

4.3.4 Box-Tidwell Transformation

Bei der Box-Tidwell Transformation werden die unabhängigen Variablen transformiert. Hierfür wird ein `boxcox` Modell geschätzt, nur dass nun die rechte Seite mit dem Parameter λ (Lambda) transformiert wird (`rhsonly`).

```
 boxcox hhrent hhinc hhsize if hhrent>0, model(rhsonly) lrtest

                                                Number of obs   =        496
                                                LR chi2(3)      =     200.94
Log likelihood = -3749.2609                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
      hhrent |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     /lambda |   1.041603   .1900712     5.48   0.000     .6690701    1.414135
------------------------------------------------------------------------------

Estimates of scale-variant parameters
-------------------------------------------------------------
             |      Coef.  chi2(df)  P>chi2(df)    df of chi2
-------------+-----------------------------------------------
Notrans      |
       _cons |   884.4346
-------------+-----------------------------------------------
Trans        |
       hhinc |   .0508152   124.104   0.000           1
      hhsize |   60.50575     9.478   0.002           1
-------------+-----------------------------------------------
      /sigma |   464.0792
-------------------------------------------------------------
```

```
-------------------------------------------------------------
   Test          Restricted     LR statistic      P-value
    H0:       log likelihood       chi2       Prob > chi2
-------------------------------------------------------------
lambda = -1     -3797.2709        96.02           0.000
lambda =  0     -3764.9224        31.32           0.000
lambda =  1      -3749.285         0.05           0.826
-------------------------------------------------------------
```

Vorgeschlagen wird eine Transformation der unabhängigen Variablen von 1.04 vorgeschlagen. Diese geringe Transformation unterscheidet das Modell aber nicht mehr von einem Modell mit einer Potenz von 1 für beide Variablen, wie die Likelihood-Rato-Teststatistik (`LR staistic chi2`) zeigt. Die Modelle sind nicht signifikant unterschiedlich.

Die Transformation der unabhängigen Variablen kann man auch graphisch darstellen. Hierfür eignet sich ein Component-plus-Residual plot (oder auch: Partial Residual Plot). Dieser zeigt, sowohl monotone als auch nicht-monotone Verläufe, wenn man eine nichtparametrischen Schätzer (`lowess`) zusätzlich zum linearen Schätzer des multiplen Regressionsmodells in den Graphen aufnimmt.

```
cprplot hhinc, lowess name(cprinc, replace)
cprplot hhsize, lowess name(cprsize, replace)
graph combine cprinc cprsize, col(2)
graph export cprplot.png, replace
```

Beide Variablen (vgl. Abbildugn 4) zeigen keine nennenswerten Abweichungen von einer Geraden. Für die Haushaltsgrösse gibt es für grösser Haushalte eine leicht abnehmende Tendenz, so dass man die Haushaltsgrösse durch ein Polynom zweiten Grades darstellen könnte.

4.3.5 Polynom Regression

Die simple monontone Transformation gilt nur für "eine" Wölbung (simple). Zeigt der Zusammenhang zwei oder drei Wölbungen, also einen Wendepunkt oder gar zwei, so muss pro Wendepunkt das Modell um eine höhere Potenz erweitert werden.

- 1. Grad: $Y = \beta_0 + \beta_1 X$
- 2. Grad: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$
- 3. Grad: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

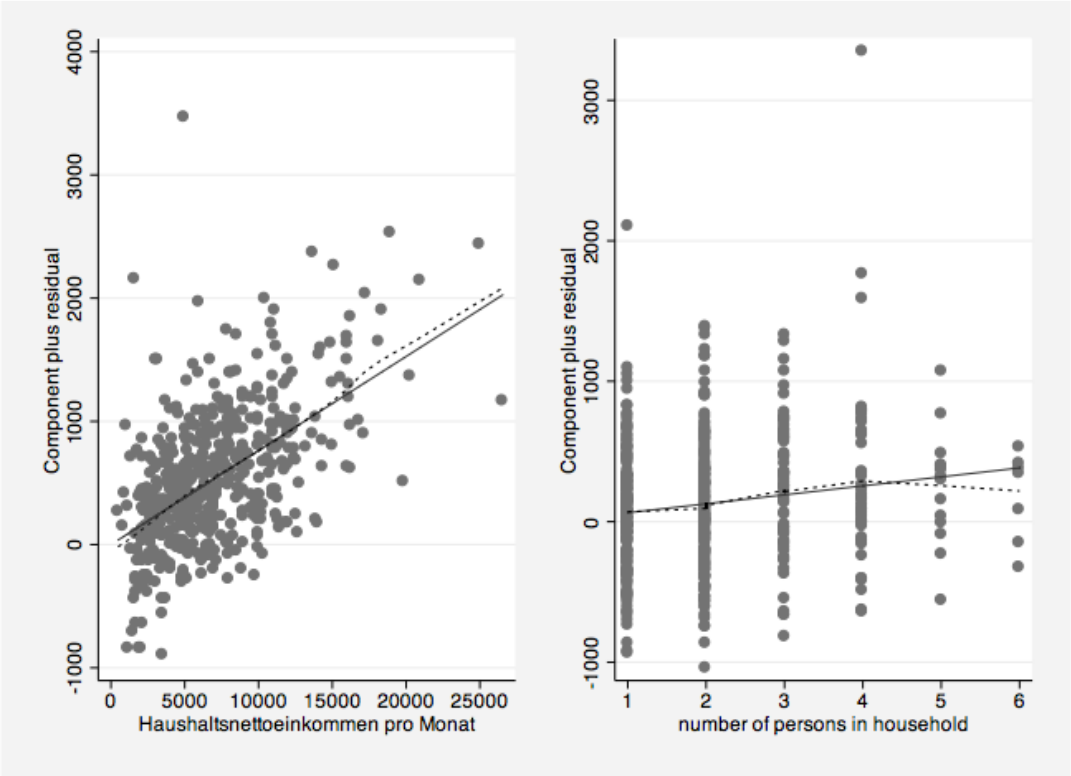


Abbildung 4: Component plus residual plot

4.3.6 Interpretation der Log-Transformation

Die Interpretation der Log-Transformation ist relativ einfach und sollte, wenn möglich, anderen Transformationen vorgezogen werden.

- $\widehat{log(y)} = \beta_0 + \beta_1 x$: wenn x um eine Einheit erhöht wird, ändert sich y im Durchschnitt um $100 \times \beta_1$ Prozent.

- $\widehat{y} = \beta_0 + \beta_1 log(x)$: wenn x um ein Prozent erhöht wird, ändert sich y im Durchschnitt um $\frac{\beta_1}{100}$ Einheiten.

- $\widehat{log(y)} = \beta_0 + \beta_1 log(x)$: wenn x um ein Prozent erhöht wird, ändert sich y im Durchschnitt um β_1 Prozent. Diese Interpretation wird auch als Elastizität bezeichnet und in den Wirtschaftswissenschaften häufig angewendet: "Wenn sich die Preise (X) eines Produktes um ein Prozent erhöhen, ändert sich die Nachfrage (Y) eines Produktes um β_1 Prozent."

Übersicht Stata-Befehle

Befehl	Option	Erklärung
`kdensity`	`normal`	Graphik des Kerndichteschätzers. `normal` fügt die Normalverteilung noch ein.
`pnorm`		Plot der Residuen mit der Normalverteilung
`qnorm`		Plot der Quantile einer Variable gegen die Quantile einer Normalverteilung.
`sktest`		Test der Schiefe und Wölbung der Verteilung der Residuen und Vergleich mit einer Normalverteilung.
`swilk`		Test auf Normalverteilung. Getestet wird H0: Fehler sind normalverteilt, H0 muss bei kleinen P-Werten abgelehnt werden.
`vif`		Variance Inflation Factor, testet für Multikollinearität im Modell.
`boxcox`	`lhsonly` `rhsonly`	je nach Option wird eine Box-Cox-Transformation für die abhängige Variable (`lhsonly`) oder unabhängigen Variablen (`rhsonly`) durchgeführt.
`ladder` *var*		Monotone Transformation. Zusätzlich wird ein `sktest` der auf Schiefe und Wölbung testet durchgeführt.
`qladder` *var*		Quantilsplot einer Normalverteilung
`gladder`		Histogramm der transformierten Variable (nicht zu empfehlen, da Graphisch nicht eindeutig zu erkennen)
`cprplot`	`lowess`	Postestimation Befehl für ein Regressionsmodell. Component plus residual plot mit Lowess-Smoother.

Aufgabe

Vorbereitung auf die Präsentation. Ausformulierung eines Exposés, das wiederum:

- die Fragestellung nennt. Literatur explizit anführt.
- Hypothesen darstellt.
- Modellgleichung zeigt (ggf. Formeleditor verwenden).
- Methode zur Hypothesenprüfung nennt (z.B. Lineare Regression, Logit-Regression).
- Wie wird die abhängige Variable gebildet? Wie die unabhängigen Variablen?
- Genaue Angaben zum Daten:
  - Hast du sie schon? Wenn ja, wie gross ist der Datensatz?
  - Hast du schon einen `master` und einen `create` Do file für das Projekt erstellt? Den Do-File bitte im Anhang anfügen.
- Das Exposé mit Anhang und allen weiteren Informationen bitte nach Ostern per Mail an mich schicken und ausgedruckt in die Stunde am Donnerstag, den 11.4.2013 mitbringen.

5 28.3. Indexkonstruktion und Faktorenanalyse

5.1 Idee eines Index

In den Sozialwissenschaften untersuchen wir oftmals nicht direkt beobachtbare Sachverhalte, wie Vertrauen, Religiosität oder in meinem Fall das Umweltbewusstsein. Diese Sachverhalte müssen erst durch ein *Konstrukt* sichtbar gemacht werden. Hilfreich bei der Operationalisierung der Konstrukte sind eine klare Definition und theoretische Annahmen über den Sachverhalt. Aus den Annahmen lassen sich messbare Fragen ableiten und formulieren. Die Fragen können dann in einen Fragebogen integriert werden. Das Umweltbewusstsein wird definiert als: "Einsicht in die Gefährdung der natürlichen Lebensgrundlage des Menschen durch diesen selbst, verbunden mit der Bereitschaft zur Abhilfe." (RSU 1978). Daraus lassen sich die theoretischen Annahmen ableiten, dass das Umweltbewusstsein aus drei Komponenten besteht (Maloney et al. 1975):

- *konative* Komponente: Bereitschaft etwas zu tun – Handlungsbereitschaft
- *affektive* Komponente: Ausmass der emotionalen Reaktion – emotionale Betroffenheit
- *kognitive* Komponente: Einsicht in die Gefährdung – Fähigkeit das Problem zu begreifen und zu erkennen

In den Datensatz des ISSP 1993 2000 und 2010 wurden jeweils neun Fragen integriert, die eine der drei Komponenten abbildet. Ganz trennscharf geht das nicht immer, aber prinzipiell entwickelt man Fragen, die den Sachverhalt messen *sollen*. Wie kann ich das *sollen* beweisen? Speziell geht es um die Frage der *Validität* der Messung, die man sich auch so stellen kann: Habe ich tatsächlich Umweltbewusstsein gemessen oder nehme ich es nur an?

Schlussendlich kann man die Annahme nie gänzlich beweisen, aber es gibt statistische Verfahren, die die Annahmen stützen. Zum einen kann die Reliabilität des Index durch *Cronbach's Alpha* (α) gemessen werden. Zum anderen prüft man mit einer Faktorenanalyse die Validität, also ob der Index auch aus den theoretisch abgeleiteten Komponenten besteht.

5.2 Reliabilitätsanalyse mit Cronbach's Alpha

Alpha ist ein Mass für die interne Konsistenz einer aus mehreren Items zusammengesetzten Skala (Index). Geprüft wird, ob die ausgewählten Items eines Index tatsächlich Elemente einer gemeinsamen Skala sind. Gemessen wird die Korrelation eines Items mit den Items der gesamten Skala, also die Interitemkorrelation aller Itempaare in dem Index. Man erhält einen Alpha-Wert, auch *Scale reliablility coefficient* genannt. Perfekte Konsistenz hat ein Index bei einem Alpha-Wert von 1 (auch negative Werte sind

möglich). Werte ab 0.8 gelten als ausreichend, allerdings wird in der Praxis auch mit werten ab 0.7 gerechnet. Um den Alpha-Wert zu verbessern, können Sie versuchsweise einzelne Variablen entfernen. In Stata erhalten Sie mit dem Befehl `alpha` *`varlist`* den Alpha-Wert. [2]

Die Formel zur Berechnung lautet:

$$\alpha = \frac{n\bar{r}}{(1+\bar{r}(n-1))}$$

Wobei n für die Zahl der Items steht und $\bar{r}$ für den Mittelwert aus allen bivariaten Korrelationen zwischen den Items. In Stata erhalten Sie die Interitem-Korrelationen mit der Option `detail`.

Beispielsweise für einen Wert mit den Daten des ISSP 2000 zum Thema Umwelt können Sie neun spezifische Variablen generieren hier `ub1` bis `ub9` genannt.

```
alpha ub*, std item
/*
mit ub* werden alle Variablen, die mit "ub" beginnen berücksichtigt
std: zeigt die z-standardisierten Korrelationen
item: zeigt die alpha-Werte an, wenn eine Variable weggelassen wird
*/

Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
ub1	27923	+	0.4527	0.2555	0.2167	0.6888
ub2	28909	+	0.5200	0.3332	0.2054	0.6741
ub3	29482	+	0.4258	0.2244	0.2232	0.6968
ub4	29513	+	0.5676	0.3906	0.1966	0.6618
ub5	29569	+	0.6554	0.5058	0.1807	0.6383
ub6	29635	+	0.6532	0.5029	0.1816	0.6396
ub7	29260	+	0.6153	0.4575	0.1879	0.6492
ub8	30060	+	0.5449	0.3624	0.2012	0.6684
ub9	28492	+	0.4699	0.2775	0.2137	0.6850
Test scale					0.2008	0.6934

[2] R gibt es Pakete, die die Funktionen zur Berechnung von Cronbachs Alpha enthalten, z.B. `psych::alpha`.

```
display (r(k)*r(rho))/(1+(r(k)-1)*r(rho)) // nur wenn std genannt ist.
.69339005
```

Mit der Option `std` werden die Items stadardisiert (Mittlewert = 0 und Varianz = 1), so erhalten wir die `interitem correlation` und können anhand der Formel den Alpha-Wert von `0.6934` nachrechnen. Dabei verwenden wir die gespeicherten Werte `r(k)`=9 und `r(rho)`=0.2008. Mit der Option `item` werden alternative Alpha-Werte gezeigt, die entstehen, wenn die `Item`-Variable der ersten Spalte weggelassen wird (z.B. sinkt der Alpha-Wert auf .65, wenn man `ub7` nicht in die Berechnungen aufnimmt). Wir sehen, dass der Index nicht deutlich verbessert werden kann, wenn man ein Variable weglässt – der Index verbessert sich nur minimal, wenn man `ub3` weglässt.

Der Wert ist noch gerade akzeptabel, da er knapp unter der kritischen Grenzen von 0.7 liegt.

5.3 Faktorenanalyse

Faktorenanalyse dient der **explorativen Datenanalyse**, d.h. es gibt prinzipiell keine Annahmen über die Zusammenhänge zwischen einzelnen Variablen meines Datensatzes. Im Gegensatz zur induktiven Statistik mit klaren, aus der Theorie abgeleiteten Annahmen, lassen wir in der explorativen Datenanalyse die Daten zu uns "sprechen", es ist ein hypothesengenerierendes Verfahren. Im Extremfall gibt es a priori überhaupt keine Annahmen über die Zahl der Faktoren als auch über die Zuordnung der Variablen. Aus der Menge der sichtbaren Variablen werden die nicht beobachtbaren, aber zugrundeliegenden Faktoren herausgebildet. Es bleiben noch viele Unklarheiten bestehen, denn weder die Zahl der Faktoren noch die genaue Zuordnung der einzelnen Variablen zu den Faktoren ist bekannt. Die Faktoren systematisieren also die beobachtbaren Variablen und durch die Zuordnung zu Faktoren werden diese gruppiert. Vergleichen Sie für eine allgemeine Zusammenfassung der Methode (Wolff und Bacher 2010).

Man unterscheidet zwischen *explorativer* Faktorenanalyse, bei der die Faktoren unbekannt sind und *konfirmatorischer* Faktorenanalyse mit vorgegebener Faktorenzahl, dann wenn Vorvermutungen bestehen anwenden. Wir betrachten eine explorative Faktorenanalyse, auch wenn wir gewissen Vermutungen haben (ganz ohne geht es nicht), aber wir geben die Anzahl der Faktoren nicht vor.

Unser Ziel ist es die Variablen für eine Skala zu verwenden, über die wir gewisse Vorannahmen haben. Also wollen wir die dimensionale Struktur der Skala testen.

Strenggenommen betrachten wir keine Faktorenanalyse, sondern nur eine Hauptkomponentenanalyse, die ein Spezialfall der Faktorenanlays ist. Für eine genaue Unterscheidung siehe (Wolff und Bacher 2010). Der Begriff Faktorenanalyse wird daher auch für eine Hauptkomponentenanalyse verwendet.

5.3.1 Hauptkomponentenanalyse

Meist wird für die Faktorenanalyse die Hauptkomponentenanalyse (PCA für Principal Component Analysis) verwendet.

Stellen Sie sich einen Raum mit drei Variablen vor. In einem ersten Schritt wird bei der PCA die erste Hauptkomponente so gebildet, dass sie den grösstmöglichen Anteil der Varianz zwischen den drei Variablen erklärt. Eine zweite Komponente soll nun grösstmöglichen Anteil der verbliebenden Varianz, der nicht durch die erste Komponente erklärten Varianz, erklären. Der Anteil der erklärten Varianz einer Variable durch den Faktor wird durch die *Komponentenladung* ausgedrückt.

In einem zweiten Schritt wird senkrecht (*orthogonal*) zum ersten Faktor, also mit einer Korrelation von 0 ein zweiter Faktor gebildet, der so viel Varianz wie möglich des verbliebenden Restanteils der Varianz erklären soll. Es werden so lange orthogonale Faktoren gebildet, maximal so viele Faktoren wie Variablen.

Wir versuchen allerdings nur die wichtigsten Komponenten zu betrachten. Die wichtigsten Faktoren sind diejenigen Komponenten, die am meisten Varianz erklären. Der Prozentanteil der erklärten Varianz kann pro Faktor angegeben werden. Ziel ist ein möglichst erklärungsreiches aber sparsames Modell. Welche Komponenten wichtig sind, das bleibt eine individuelle Entscheidung. Die Faktorenanalyse liefert lediglich Anhaltspunkte für die Entscheidung.

In einem dritten Schritt werden die Komponenten noch rotiert. Das Ziel ist, dass eine Variable möglichst hoch auf eine Komponente lädt und möglichst wenig auf die anderen. Die Rotation stellt demnach die latente Struktur in den Daten noch klarer hervor. Dadurch verändern sich aber die Faktorenladungen. Es wird zwischen orthogonaler und schiefwinkliger Rotation unterschieden. Die schiefwinklige Rotation erlaubt eine bessere Anpassung, weil man davon ausgeht, dass die Faktoren auch mit einander korrelieren dürfen.

Die Grundlage für die Entscheidung für oder gegen einen Faktor bieten die **Eigenwerte**. Ein Eigenwert gibt an, wie viele Varianz ein Eigenwert erklären kann. Ein Eigenwert von 2.6 erklärt so viel wie 2.6 Variablen. Ein Eigenwert von 0.5 zeigt das die Erklärungskraft des Faktors geringer als der einer bestehenden Variable ist. Die Summe der Eigenwerte ist die Anzahl der verwendeten Variablen, daher ist der Eigenwert auch der Anteil, der durch den Faktor erklärten Varianz. Bei 10 Variablen ist ein Eigenwert $\frac{2.6}{10} = 26\,\%$, ein Eigenwert von 0.5 erklärt nur $\frac{0.5}{10} = 5\,\%$.

Das gängigste Kriterium zur Auswahl der Faktoren ist das Kaiser Kriterium, nachdem Wert > 1 verwendet werden, da die Erklärungskraft dann grösser als bei einer Variable ist. Die zusammengefassten Variablen bedeuten so eine Datenreduktion. Weitere Verfahren sind der Screeplot und die Parallelanalyse. Der Screeplot ist ein grafisches Verfahren, bei der Parallelanalyse werden die beobachten Eigenwerte mit Eigenwerten einer Zufallsstichprobe verglichen und diejenigen Faktoren extrahiert, deren Eigenwerte deutlich über den Zufallseigenwerten liegen. Wir verwenden das Kaiser Kriterium, wobei

bei diesem Vorgehen zu kritisieren ist, dass zu viel Komponenten mit einen Wert > 1 extrahiert werden können.

Beispiel: mit den neun Variablen wird nun eine Faktorenanalyse durchgeführt. In Stata wird hierfür der Befehl `factor` *`varlist`* und die Option `pca` für die Hauptkomponentenanalyse bestimmt.[3]

```
 factor ub1  ub2  ub3  ub4  ub5  ub6 ub7 ub8 ub9, pcf
(obs=23390)

Factor analysis/correlation                       Number of obs    =    23390
    Method: principal-component factors           Retained factors =        2
    Rotation: (unrotated)                         Number of params =       17

    --------------------------------------------------------------------------
        Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    ------------+-------------------------------------------------------------
        Factor1 |      2.65152      0.89716            0.2946       0.2946
        Factor2 |      1.75436      0.85463            0.1949       0.4895
        Factor3 |      0.89973      0.06939            0.1000       0.5895
        Factor4 |      0.83034      0.07429            0.0923       0.6818
        Factor5 |      0.75604      0.02773            0.0840       0.7658
        Factor6 |      0.72831      0.15491            0.0809       0.8467
        Factor7 |      0.57339      0.08280            0.0637       0.9104
        Factor8 |      0.49060      0.17488            0.0545       0.9649
        Factor9 |      0.31571            .            0.0351       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) = 4.1e+04 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
             ub1 |   0.3397    0.5133 |      0.6212
             ub2 |   0.4539    0.5233 |      0.5202
             ub3 |   0.3668   -0.3362 |      0.7524
             ub4 |   0.5274    0.4631 |      0.5074
             ub5 |   0.7241   -0.3210 |      0.3727
             ub6 |   0.7386   -0.4298 |      0.2698
             ub7 |   0.7098   -0.4707 |      0.2747
             ub8 |   0.4583    0.3763 |      0.6484
             ub9 |   0.3658    0.4886 |      0.6275
    -------------------------------------------------
```

[3] In R wird Faktorenanalyse mit dem Befehl `factanal` durchgeführt.

Es haben zwei Faktoren einen Eigenwert > 1. Der Anteil der erklärten Varianz für Faktor 1 beträgt $\frac{2.65}{9} = 29\,\%$.

Die Summe der quadrierten Ladungen aller neun Variablen auf Faktor 1 entspricht der maximalen Varianz, die durch den ersten Faktor erklärt werden kann. Alle Variablen haben mit dem 1 Faktor etwas gemeinsam, da alle Variablen positiv laden. Der zweite Faktor hat auch negative Ladungen, so dass es auch Unterschiede gibt. Die positiven Ladungen auf Faktor 1 deuten darauf hin, dass alle Variablen inhaltlich etwas gemeinsam haben.

Zur genauen Analyse, gerade wenn mehrere Faktoren bestimmt werden, sollten die Daten allerdings rotiert werden. Hierfür in Stata nach der Faktorenanalyse den Befehl `rotate` ausführen. Standardmässig wird in Stata eine (orthogonale) *Varimax-Rotation* durchgeführt.

```
rotate

Factor analysis/correlation                          Number of obs    =    23390
    Method: principal-component factors              Retained factors =        2
    Rotation: orthogonal varimax (Kaiser off)        Number of params =       17

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.35760      0.30932            0.2620       0.2620
        Factor2  |      2.04828            .            0.2276       0.4895
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) = 4.1e+04 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
             ub1 |  -0.0153    0.6153 |      0.6212
             ub2 |   0.0726    0.6889 |      0.5202
             ub3 |   0.4932   -0.0657 |      0.7524
             ub4 |   0.1674    0.6816 |      0.5074
             ub5 |   0.7775    0.1512 |      0.3727
             ub6 |   0.8516    0.0704 |      0.2698
             ub7 |   0.8514    0.0203 |      0.2747
             ub8 |   0.1605    0.5708 |      0.6484
             ub9 |   0.0203    0.6100 |      0.6275
    -------------------------------------------------

Factor rotation matrix
```

```
------------------------------
             | Factor1  Factor2
-------------+----------------
     Factor1 |  0.8200   0.5724
     Factor2 | -0.5724   0.8200
------------------------------
```

Durch die Rotation ändern sich die Ladungen erheblich. Ausgewählt werden standardmässig nur jene Faktoren die einen Eigenwert > 1 haben (Kaiser Kriterium). Auf den Faktor 1 laden nun `ub5, ub6, ub7` deutlich und `ub3` knapp unter 0.5. Auf den Faktor 2 laden `ub1, ub2, ub4, ub8, ub9` deutlich. Damit können die Variablen zwei Faktoren zugewiesen werden. Mit der Option `promax` oder `oblimin` können schiefwinklige Rotationen durchgeführt werden. Allerdings sollte die Korrelation zwischen den Faktoren nicht zu hoch sein.

Insgesamt können die beiden Faktoren mit einem kumulierten Anteil von `0.4895` 49 % der Streuung erklären.

In der Soziologie gibt es die Faustregel, dass nur Variablen die über 0.5 auf einen Faktor laden, einen Faktor zugeordnet werden sollen. Das sind Richtwerte, an die man sich prinzipiell halten sollte. Die Option `blank(0.4)` zeigt nur Werte über 0.4. So werden auch Werte, wie in unserem Beispiel bei ub3 von 0.49 angezeigt, die knapp unter der Grenze von 0.5 liegen und faktisch auf den Faktor laden. Mit dem Befehl `quietly` vor `factor` werden die unrotierten Ergebnisse der Faktorenanlyse nicht gezeigt, da diese in unsere Analysen grundsätzlich nicht berücksichtigt werden. Erst die Ergebnisse der (orthogonalen) Varimax-Rotation werden gezeigt.

```
quietly  factor ub*, pcf
rotate, blanks(.4)
```

Inhaltlich konnte, entgegen der eigentlichen Vermutung, keine dreifaktorielle Lösung gefunden werden, die Faktorenanalyse zeigt zwei Faktoren. Diese macht inhaltlich Sinn, denn (wie hier nicht gezeigt) bildet Faktor 1 die konative Komponente ab und Faktor 2 eine Komponente die die affektive Betroffenheit zeigt.

5.3.2 Vorgehen kurz und knapp

- unrotierte Hauptkomponentenanalyse durchführen und schauen, ob alle Variablen auf den ersten Faktor laden
- Rotation (Variamax) durchführen, wenn weitere Faktoren interpretiert werden sollen. Hier nur Faktoren berücksichtigen, die einen Eigenwert > 1 haben (Kaiser Kriterium).

- Anteil der erklärten Gesamtvarianz durch ausgewählte Faktoren nennen.

- Faktorladungen genauer ansehen. Variablen die > 0.5 auf einen Faktor laden können einem Faktor zugeordnet werden. Ggf. eine schiefwinklige promax oder oblimin Rotation durchführen.

- Interpretation: Welche Variablen laden auf welchen Faktor. Entsprechen die Ladungen den vorher vermuteten Annahmen, kann die Faktorenanalyse die theoretischen Annahmen bestätigen. Wie können die Faktoren benannt werden?

- Gibt es Variablen die nicht eindeutig zugeordnet werden können, beispielsweise mit einem Wert > 0.3 auf mehrere Faktoren lädt? (Ist in oben genanntem Beispiel nicht der Fall, aber durch aus möglich und sollte dann diskutiert werden.)

5.4 Probleme und Stolpersteine der Faktorenanalyse

Mehr zu den Problemen unter (Wolff und Bacher 2010: 360-363), hier nur ein Ausschnitt, der Ihnen wichtige Hinweise für das spätere Studium zeigen kann. Die Punkte sollten Sie im Hinterkopf behalten, Sie müssen aber für die Faktorenanalyse in der Hausarbeit nicht alle beachtet werden:

- Stellen Sie die Variablen für die Analyse sorgfältig zusammen

- Insbesondere sollten Sie im Vorfeld diskutieren, ob die Variablen nur einen Faktor darstellen (unidimensional) oder mehrdimensional sind.

- Verwenden Sie nicht nur die in Stata voreingestellten Methoden wie beispielsweise die Varimax-Rotation (für die Analysen unserer Arbeit reichen die Voreinstellungen aber aus), um die Robustheit Ihrer Analysen zu unterstreichen. Betrachten Sie z.B. nicht nur das Kaiser Kriterium, sondern führen Sie auch eine Parallelanalyse durch. Vermeiden Sie die Auswahl zu vieler (Überextraktion), aber auch zu weniger Faktoren (Unterextraktion).

- Die Validität der Faktorenanalyse kann mit Kreuzvaldidierung getestet werden. Dabei testen Sie die Faktoren bei einem anderen Datensatz oder teilen einen grossen Datensatz und führen getrennte Faktorenanalysen durch.

- In separaten Stichproben könnte man auch die Frageformulierung oder -reihenfolge ändern, um Frage-Effekte zu testen.

Übersicht Stata-Befehle

Befehl	Option	Erklärung
`alpha`	`detail std`	Misst die Inter-Item Korrelation aller Itempaare. Mass für die Reliabilität. `detail` gibt die Interitem Korrelationen aus, wenn die Option `std` gewählt wird.
`factor`	`pca blank(0.5)`	Faktorenanalyse mit Hauptkomponentenanalyse wenn `pca` gewählt. `blank(0.5) zeigt nur Variablen die höher als 0.5 auf einen Faktor laden.`
`rotate`	`blank(0.5)`, `oblimin`, `promax`	Rotation (Standard: Varimax), wenn `oblimin, promax`, dann nicht-orthogonale Rotation

6 11.4. – Generalisierte Lineare Regressionsmodelle 1

6.1 Einleitung

Generell geht es um die Frage, wie man eine dichotome abhängige Variable durch eine Linearkombination von $K-1$ unabhängigen Variablen vorhersagen kann.

$$\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{K-1} x_{K-1,i}$$

Beispiele sind die Vorhersage von Mitgliedschaft in einem Verein, Wahlentscheidung für eine bestimmt Partei oder Rauchverhalten einer Person. Man kann bei allen Beispielen die Frage mit ja oder nein beantworten. Damit handelt es sich jeweils um eine dichotome abhängige Variable.

Dichotome Variablen haben die Eigenschaft nur aus zwei Werten zu bestehen, die allgemein mit 0 oder 1 kodiert sind. Zwischenwerte gibt es nicht. Man kann nur Mitglied oder nicht Mitglied sein und eine Partei kann man nur vollständig gewählt haben und nicht zu einem gewissen Anteil oder Grad, der zwischen nicht-gewählt und gewählt liegt.

6.2 Lineares Wahrscheinlichkeitsmodell

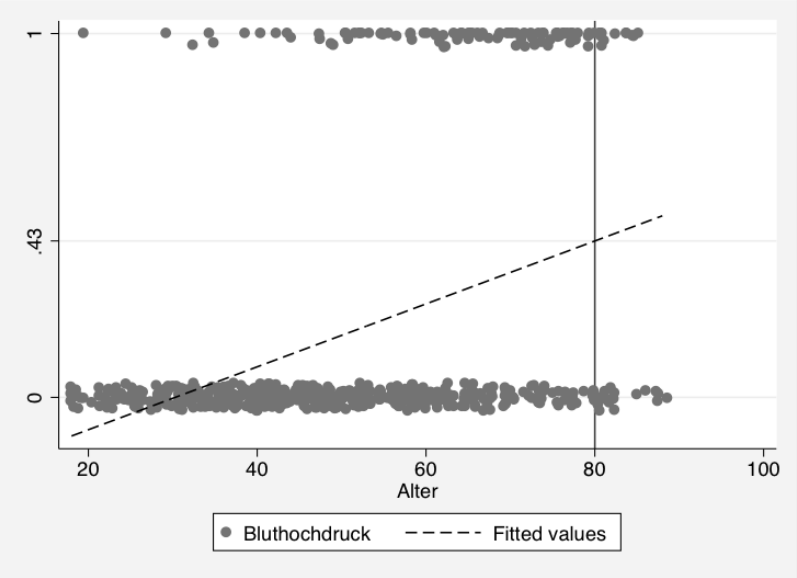
Betrachten wir die Ausprägungen der Variable Bluthochdruck einer Person.

```
 sum blut

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        blut |       597    .1758794    .3810368          0          1
```

Der Mittelwert von 0.17 besagt, dass der Anteil der Personen mit erhöhtem Blutdruck 17 % beträgt. Man kann damit sagen, dass die geschätzte Wahrscheinlichkeit, dass eine Person erhöhten Blutdruck hat 0.17 beträgt.

Den Zusammenhang von Bluthochdurck und Alter kann man durch einen Scatterplot darstellen:

```
twoway  (scatter blut alter, jitter(10)) (lfit blut alter ), ylabel(0 0.43 1) xline(80)
```



In dem linearen Regressionsmodell beträgt die Wahrscheinlichkeit unter Bluthochdruck zu leiden für Personen mit 80 Jahren 0.43. Dummerweise lassen sich negative Werte nicht sinnvoll interpretieren, da der Wertebereich einer Wahrscheinlichkeit niemals negativ sein kann.

Zwei Probleme können bei einer linearen Regression mit einer dichotomen abhängigen Variable auftreten:

1. Nicht alle vorhergesagten Werte können sinnvoll interpretiert werden.
2. Das Modell hat heteroskedastische Fehler, da die Varianz mit $\widehat{y}_i \times (1 - \widehat{y}_1)$ umso grösser ist, je näher sich die vorhergesagten Werte an 0.5 annähern. Die Folge sind falsche Standardfehler, Konfidenzintervalle und p-Werte. Es können also für das Modell keine korrekten inferenzstatistischen Aussagen gemacht werden.

Man braucht ein Modell, das ausschliesslich Wahrscheinlichkeiten zwischen 0 und 1 produziert und Annahmen hat, die das Modell erfüllen kann. Damit kommen wir in die Klasse der "generalisierten" linearen Regressionsmodelle.

6.3 Odds: Interpretation

In diesem Kapitel geht es darum, eine sinnvolle Interpretation der Wahrscheinlichkeiten zu finden. Wie untenstehende Kreuztabelle mit Reihenprozenten zeigt, ist die Wahrscheinlichkeit für unter 50-Jährige der Stichprobe an Bluthochdruck zu leiden 4.7 % . Formal: $P(Y_{<50} = 1) = 14/298 = 0.047$. Demnach ist die Wahrscheinlichkeit nicht an Bluthochdruck zu leiden $1 - P(Y_{<50} = 1) = 1 - 0.047 = 0.953$. Der Anteil der Erkrankten wird somit als Schätzung der Erkrankungswahrscheinlichkeit interpretiert. Diese erhöht sich für ab 50 Jahre auf rund 30 % und liegt damit rund 25 Prozentpunkte über dem Wert der unter 50-Jährigen.

```
 tab alter_d blut, row

               |     Bluthochdruck
Alter dichotom |      Nein         Ja |     Total
---------------+----------------------+----------
unter 50 Jahre |       284         14 |       298
               |     95.30       4.70 |    100.00
---------------+----------------------+----------
   ab 50 Jahre |       208         91 |       299
               |     69.57      30.43 |    100.00
---------------+----------------------+----------
         Total |       492        105 |       597
               |     82.41      17.59 |    100.00
```

Man kann bei der Betrachtung von Wahrscheinlichkeiten noch weiter gehen. Die geschätzten Wahrscheinlichkeiten können ins Verhältnis gesetzt werden. Dann spricht man von Chancen. "Die Chance für unter 50-Jährige an Bluthochdruck zu leiden beträgt 4.7 zu 95.3 und liegt damit bei 0.049 zu 1." Damit beträgt die geschätzte Wahrscheinlichkeit an Bluthochdruck zu leiden ein Zwanzigstel (1/0.049) der geschätzten Wahrscheinlichkeit einen normalen Blutdruck zu haben. Man kann den Gedanken auch umdrehen und nach der Chance gesund zu sein fragen. Dann beträgt die Chance zu den Gesunden zu gehören $95.3/4.7 = 20.28$ oder 20 zu 1. Damit ist die geschätzte Wahrscheinlichkeit keinen Bluthochdruck zu haben 20 mal so hoch wie die geschätzte Wahrscheinlichkeit Bluthochdruck zu haben. Als unter 50-Jähriger gehört man damit 20 mal häufiger zu den Gesunden als zu den Erkrankten. Chancen werden mit *Odds* bezeichnet, dem englischen Wort für Chance. In der Praxis spricht man von Chancen und Risiken. In unserem Fall beträgt die Chance gesund zu sein 20 zu 1 und das Risiko Bluthochdruck zu haben 1 zu 20 ($1/20.28 = 0.049$).

$$\text{Chance}_{gesund} = \frac{\text{Wahrscheinlichkeit}_{gesund}}{\text{Wahrscheinlichkeit}_{krank}}$$

Die geschätzte Wahrscheinlichkeit ab 50 Jahre keinen Bluthochdruck zu haben ist rund zwei mal so hoch ($69.6/30.4 = 2.29$) als die geschätzte Wahrscheinlichkeit Bluthochdruck zu haben.

Odds-Ratio Man kann die Chancen auch ins Verhältnis setzen. Die geschätzte Chance gesund zu sein von unter 50-Jährigen im Vergleich zu älteren Personen ist $20.28/2.29 = 8.86$ und damit rund 9 mal so hoch. Umgekehrt ist die Chance gesund zu sein für ältere Personen nur $2.29/20.28 = 0.11$ und damit rund ein Zehntel der jüngeren Personen. Die Chancenverhältnisse werden auch *Odds-Ratio* genannt. Das Alter scheint damit einen Einfluss auf das Eintreten von Bluthochdruck zu haben.

Weitere Beispiele für die Anwendung von Chancen oder Chancenverhältnissen sind z. B. die Untersuchung von Karrierechancen von Frauen oder die Bildungschancen von Kindern der Unterschicht.

6.4 Logarithmierte Odds

Wie unschwer zu erkennen ist, ist der Wertebereich der Odds von 0 bis $+\infty$. Ein Odds von 0 ist dann vorhanden, wenn es in einer Gruppe keine Erkrankungen gibt und $+\infty$ liegt vor, wenn alle Personen angeben unter Bluthochdruck zu leiden. Bei gleicher Zahl von Gesunden wie Erkrankten haben wir einen Odds von 1. Verwendet man den Logarithmus der Odds, so bekommt man einen Wertebereich, der von $-\infty$ bis $+\infty$ reicht. Der Logarithmus von 1 ist 0, so dass man mit den logarithmierten Odds oder *Logits*, wie wir sie nennen, Chancenverhältnisse erhalten. Werte kleiner Null verringern die Chance und werte grösser Null erhöhen die Chance und damit auch die Wahrscheinlichkeit, dass ein bestimmter Zustand eintrifft (Person ist krank, Person ist aufgestiegen, Person hat höhere Schulbildung etc.).

Wahrscheinlichkeit	Odds	Odds	Logits
$\widehat{P}(Y=1)$	$\widehat{Odds} =$	$\frac{\widehat{P}(Y=1)}{1-\widehat{P}(Y=1)}$	$ln(\widehat{Odds})$
0.01	$1/99 =$	0.01	-4.60
0.03	$3/97 =$	0.03	-3.48
0.05	$5/95 =$	0.05	-2.94
0.2	$20/80 =$	0.25	-1.39
0.3	$30/70 =$	0.43	-0.85
0.4	$40/60 =$	0.67	-0.41
0.5	$50/50 =$	1.00	0
0.6	$60/40 =$	1.50	0.41
0.7	$70/30 =$	2.33	0.85
0.8	$80/20 =$	4.00	1.39
0.95	$95/5 =$	19.00	2.94
0.97	$97/3 =$	32.33	3.48
0.99	$99/1 =$	99.00	4.60

Die Tabelle zeigt nochmals übersichtlich, wie die Transformation von Wahrscheinlichkeiten in Logits geschieht (vgl. Kohler und Kreuter 2008: S. 265).

Das Logit (L_i) ist nicht begrenzt und symmetrisch um den Ursprung und ist damit eine Zielgrösse unseres eingangs erwähnten Regressionsmodells:

$$L_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{K-1} x_{K-1,i}$$

Berechnet wird der Einfluss der $(K-1)$ unabhängigen Variablen, dass die logarithmierte Chance den Wert 1 erhält. Also der beobachtete Fall auftritt. Somit kann jede

unabhängige Variable die logarithmierte Chance erhöhen oder senken, was sich an den β-Koeffizienten ablesen lässt.

Logits lassen sich durch den Antilogarithmus in Odds-Ratios umwandeln.

$$e^{\beta}$$

Die Odds besagen, dass sich bei Zunahme einer unabhängigen Variable um eine Einheit, die Odds, eher die Ausprägung $y = 1$ als die Ausprägung $y = 0$ zu beobachten, um den Faktor e^{β} verändern. Die Chance verändert sich um einen bestimmten *Faktor*.

Man kann auch angeben, um wie viel *Prozent* sich die Chancen erhöhen. Hierfür wird die Formel verwendet:

$$100 \times (e^{\beta} - 1)$$

Die Interpretation lautet dann: wenn sich x um eine Einheit verändert, ändert sich die Chance um $100 \times (e^{\beta} - 1)$ Prozent.

Logits können mathematisch auch in Wahrscheinlichkeiten umgewandelt werden. Durch Auflösen des Logarithmus und Umstellen der Funktion

$$ln[P(Y = 1)/(1 - P(Y = 1))] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{K-1} x_{K-1,i} \tag{18}$$

erhält man:

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{K-1} x_{K-1,i}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{K-1} x_{K-1,i}}}$$

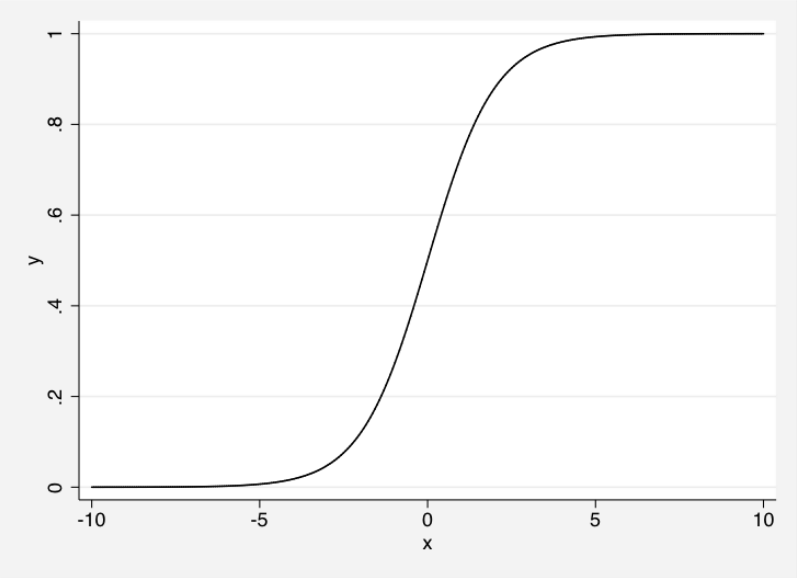
Wobei man das lineare Regressionsmodell durch L ersetzen kann:

$$P(Y = 1) = \frac{e^{L}}{1 + e^{L}}$$

e ist die Eulerische Zahl ($e \approx 2.718$) und L das Logit der Linearkombination.

Zur besseren Veranschaulichung der in Wahrscheinlichkeiten umgewandelten Logits kann folgende Funktion gewählt werden.

```
graph twoway function y=exp(x)/(1+exp(x)), range(-10 10)
```



Die aus den Logits berechneten Wahrscheinlichkeiten nähern sich asymptotisch den Werten 0 und 1 an, erreichen diese aber niemals, da sie an den Grenzen sehr stark abflachen. Zur Erinnerung: die Logits allerdings haben einen unbegrenzten Wertebereich. In der Graphik ist gut zu sehen, dass sich die Wahrscheinlichkeiten zwischen -2.5 und $+2.5$ relativ stark verändern, die Veränderung aber abnimmt, je näher man an die Grenzwerte 0 und 1 kommt.

Für uns ist wichtig festzuhalten, dass sich vorhergesagte Logits einer Linearkombination jederzeit in Wahrscheinlichkeiten zwischen 0 und 1 umrechnen lassen können. Somit ist ein Logit eine geeignete Zielgrösse der Linearkombination.

6.5 Logistische Regression

In `Stata` gibt es den Befehl `logit`, den man anstelle des gewohnten Befehls `regress` verwendet, um Logits zu berechnen. Mit der Option `or` kann man sich statt Logits die Odds-Ratios anzeigen lassen. Gleiches ist möglich, wenn man den Befehl `logistic` verwendet. Wie bei der linearen Regression folgt dem Befehl `logit` die abhängige Variable und eine Liste unabhängiger Variablen.

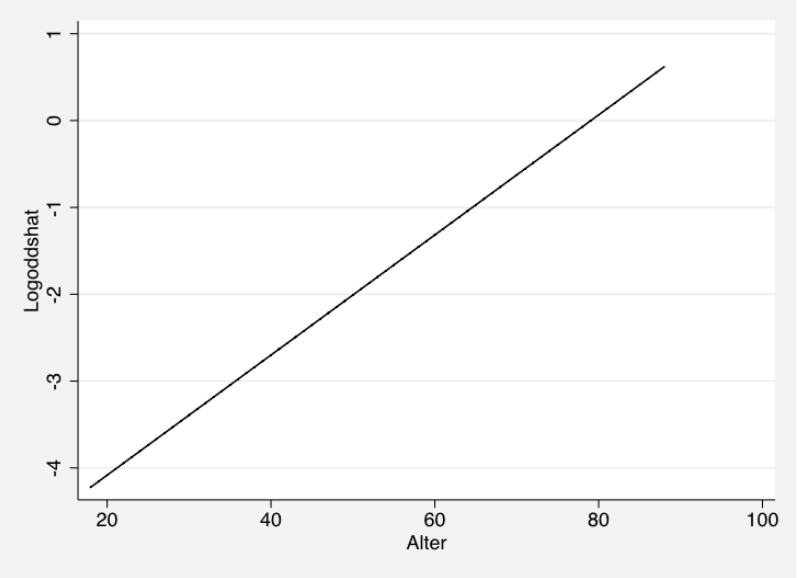
```
* Zentrierung der Variable Alter
* Interpretation der Konstanten sinnvoller
sum alter if blut <.
generate calter = alter - r(mean) if blut <.
label var calter "Alter zentriert"

* Logistische Regression
logit blut calter
```

```
Iteration 0:   log likelihood = -277.65715
Iteration 1:   log likelihood = -233.59213
Iteration 2:   log likelihood = -228.84803
Iteration 3:   log likelihood = -228.81891
Iteration 4:   log likelihood =  -228.8189

Logistic regression                                     Number of obs   =        597
                                                        LR chi2(1)      =      97.68
                                                        Prob > chi2     =     0.0000
Log likelihood =  -228.8189                             Pseudo R2       =     0.1759

------------------------------------------------------------------------------
       blut |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
     calter |   .0691486   .0079855     8.66   0.000     .0534972       .0848
      _cons |  -1.965034   .1477202   -13.30   0.000    -2.254561   -1.675508
------------------------------------------------------------------------------
```

Der Stata-Output gleicht dem der linearen Regression. Er besteht aus dem Koeffizientenblock (unten), einem Modellfit-Block (oben rechts) und einem Iterationsblock (oben links).

Wir konzentrieren uns nur auf den Koeffizientenblock und die Interpretation der Koeffizienten. Die β-Koeffizienten geben an, um wie viel sich die vorhergesagten Werte bei Anstieg der unabhängigen Variablen verändern. Die Vorzeichen zeigen schon an, ob sich die Chance erhöht (positives Vorzeichen) oder verringert (negatives Vorzeichen). Höhere Werte, weisen auf stärkere Effekte hin. Allerdings ist die Interpretation der Logits nicht intuitiv und bedarf einer Umrechnung in Odds-Ratios, Protzentwerte oder in Wahrscheinlichkeiten. Mit jedem zusätzlichen Jahr verändert sich die logarithmierte Chance kontinuierlich um 0.069.

```
generate Logoddshat = _b[_cons]+_b[calter]*calter

graph twoway line Logoddshat alter, sort
```



Die *Odds-Ratios* (die Wahrscheinlichkeitsverhältnisse) verändern sich um $e^{0.0691486} = 1.0716$. Damit steigt das Risiko erhöhten Blutdruck zu haben mit jedem Jahr um den Faktor 1.07 oder um 7 % ($= 100 \times (e^{0.0691486} - 1)$).

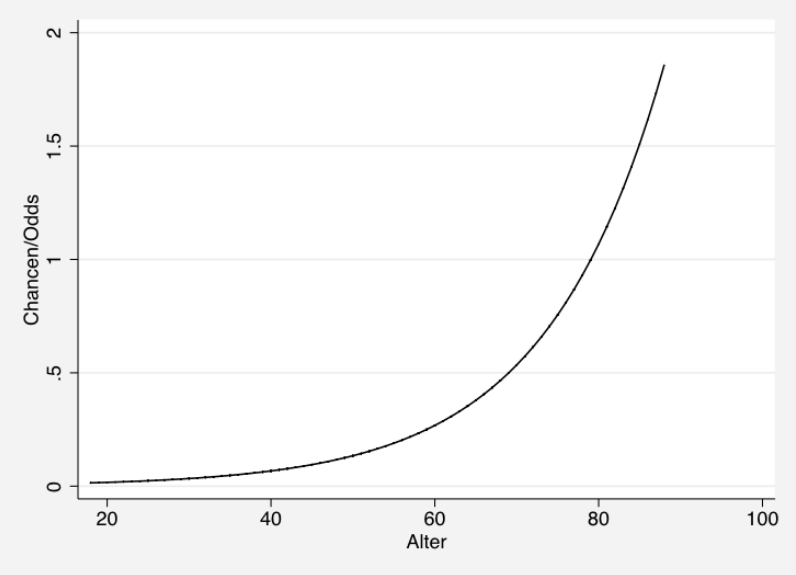
```
 dis exp(.0691486)
1.0715954

* Wahrscheinlichkeit
dis 100*(exp(0.0691486)-1)
7.1595436
```

Alternativ kann man sich die Odds-Ratios auch durch die Option `or` ausgeben lassen:

```
logit blut calter, or

Iteration 0:   log likelihood = -277.65715
Iteration 1:   log likelihood = -233.59213
Iteration 2:   log likelihood = -228.84803
Iteration 3:   log likelihood = -228.81891
Iteration 4:   log likelihood =  -228.8189

Logistic regression                                     Number of obs   =        597
                                                        LR chi2(1)      =      97.68
                                                        Prob > chi2     =     0.0000
Log likelihood =  -228.8189                             Pseudo R2       =     0.1759

------------------------------------------------------------------------------
       blut | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
      calter |   1.071595   .0085573     8.66   0.000     1.054954    1.088499
------------------------------------------------------------------------------
```

```
generate Oddshat = exp(_b[_cons]+_b[calter]*calter)

graph twoway line Oddshat alter, sort
```



Die Konstante besagt, dass die geschätzte logarithmierte Chance Bluthochdruck zu haben für eine Person mittleren Alters -1.965034 beträgt. Auch hier bietet sich die Interpretation mit Odds-Ratios an.

```
. display exp(_b[_cons])
.14015108
```

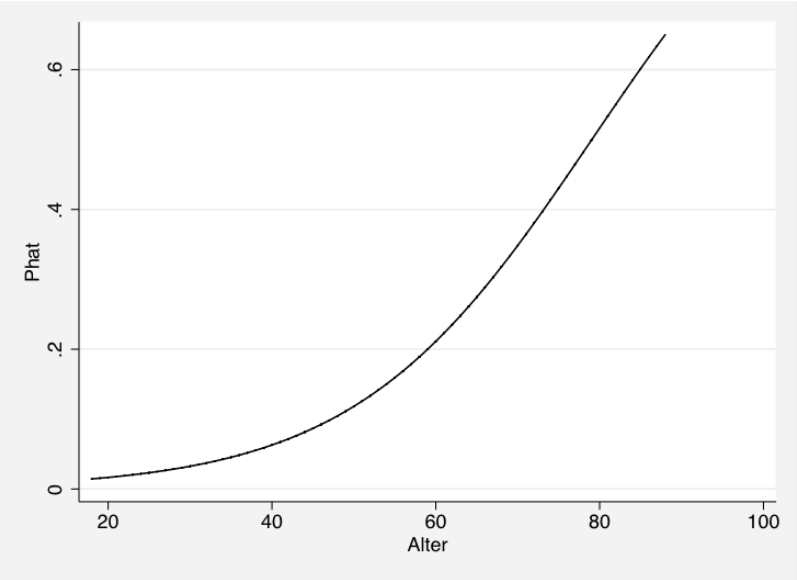
Das Risiko für eine Person mittleren Alters (50.63 Jahre) an Bluthochdruck zu leiden beträgt 1 zu 0.14. Damit ist das Verhältnis von Gesunden zu Kranken beschrieben und nicht die Wahrscheinlichkeit.

Die Wahrscheinlichkeitsinterpretation ist aussagekräftiger und lässt sich wie folgt herleiten:

```
display exp(_b[_cons])/(1+exp(_b[_cons]))
.12292326
```

Wir wenden die Formel zur Umrechnung von Logits in Wahrscheinlichkeiten an. Der geschätzte Anteil für Personen mittleren Alters Bluthochdruck zu haben beträgt ca. 12 %. Allerdings steigen die Wahrscheinlichkeiten nicht gleichmässig mit einer Erhöhung der unabhängigen Variablen an, wie folgende Graphik zeigt.

```
generate Phat = exp(_b[_cons]+_b[calter]*calter) ///
/(1+ exp(_b[_cons]+_b[calter]*calter))

graph twoway line Phat alter, sort
```



Die Wahrscheinlichkeit ist erst relativ flach und steigt immer stärker an. Eine Erhöhung des Alters führt damit nicht zu einer gleichbleibenden Veränderung der vorhergesagten Wahrscheinlichkeit. Die Wahrscheinlichkeit erhöht sich in den ersten 10 Jahren (50-60 Jahre) um 0.096, beträgt in den folgenden 10 Jahren schon 0.139 und steigt in den Jahren 70-80 um 0.169.

```
display                  exp(_b[_cons]) ///
>                        /(1+ exp(_b[_cons]))
.12292326
 * 10 Jahre über dem Durchschnitt
. display                  exp(_b[_cons]+_b[calter]*10) ///
>                          /(1+ exp(_b[_cons]+_b[calter]*10))
.21865049

. * Differenz 10 Jahre
. display .21865049 - .12292326
.09572723

. * 20 Jahre über dem Durchschnitt
. display                  exp(_b[_cons]+_b[calter]*20) ///
>                          /(1+ exp(_b[_cons]+_b[calter]*20))
.35845832

. * Differenz weitere 10 Jahre
```

```
. display .35845832 - .21865049
.13980783

. * 30 Jahre über dem Durchschnitt
. display                   exp(_b[_cons]+_b[calter]*30) ///
>                           /(1+ exp(_b[_cons]+_b[calter]*30))
.52732885

. * Differenz weitere 10 Jahre
. display .52732885 - .35845832
.16887053
```

Die Wahrscheinlichkeiten kann man sich auch mit dem Befehl `prtab` ausgeben lassen, der nach dem `logit` Befehl angewendet werden kann. Hier wird die Wahrscheinlichkeit an Bluthochdruck zu leiden für 40 bis 80 jährige Personen ausgegeben. Die Wahrscheinlichkeit steigt Anfangs nur leicht von 6 % auf 12 %. Die Differenz zwischen 70 und 80-Jährigen beträgt aber $0.52 - 0.35 = 0.17$.

```
. quietly logit blut alter

. prtab alter if alter==40  | alter==50 | alter==60 | alter==70 | alter==80

logit: Predicted probabilities of positive outcome for blut

----------------------
    Alter | Prediction
----------+-----------
       40 |     0.0629
       50 |     0.1183
       60 |     0.2112
       70 |     0.3484
       80 |     0.5164
----------------------

        alter
x=  50.454545
```

6.6 Logistische Regression: Interpretation mit durchschnittlichen marginalen Effekten

Die intuitive Interpretation der Effekte einer Logistischen Regression sind *Wahrscheinlichkeiten* und nicht Odds oder Odds-Ratios. Wünschenswert ist daher die Berechnung des Effekts der unabhängigen Variable als Wahrscheinlichkeit. Leider handelt es sich

um ein nichtlineares Modell, so dass die Steigung der Wahrscheinlichkeitskurve nicht konstant ist. Dies ist gut zu sehen durch den "Conditional-Effect-Plot". Der Conditional-Effect-Plot sollte daher immer angegeben werden.

Zusätzlich kann man eine Masszahl verwenden, die den durchschnittlichen Einfluss der unabhängigen Variable auf die Wahrscheinlichkeit, dass ein Ereignis eintritt ($P(y = 1|x)$) in einer Zahl ausdrückt. Hierfür wird ein "average marginal effect" (AME) berechnet. Zu interpretieren sind die AMEs als der durchschnittliche Effekt der Variable auf die Wahrscheinlichkeit. Das ado wird mit dem Befehl `ssc install` *`adoname`* installiert. Anschliessend kann man nach einer logistischen Regression den Befehl `margeff` ausführen.

```
* Installiere nötiges Packet margeff
ssc install margeff


quietly logit blut calter

margeff                 // Marginal effects

Average marginal effects on Prob(blut==Ja) after logit

------------------------------------------------------------------------------
        blut |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      calter |   .0084489   .0007848    10.77   0.000      .0069107     .009987
      female |   -.080584   .0287131    -2.81   0.005    -.1368607   -.0243073
------------------------------------------------------------------------------
```

Die Interpretation des *AME* ist intuitiv: Steigt x_j um eine Einheit, so steigt die Wahrscheinlichkeit, dass $y = 1$ im durchschnitt um *AME* Prozentpunkte. Allerdings ist das nur die *durchschnittliche* Veränderung, die eben den nichtlinearen Verlauf der Wahrscheinlichkeitskurve ignoriert. Demnach steigt die Wahrscheinlichkeit Bluthochdruck zu haben pro Lebensjahr um 0.8 %, wenn man auch den Effekt des Geschlechts im Modell berücksichtigt.

Der AME ist nicht zu verwechseln mit dem marginalen Effekt der Variable am Mittelwert "marginal effect at the mean" (MEM), den man (siehe Output unten `MargEfct`) durch den Postestimation-Befehl `prchange, fromto` erhält.

```
 prchange, fromto

logit: Changes in Probabilities for blut
```

```
             from:       to:      dif:     from:       to:      dif:
             x=min     x=max  min->max       x=0       x=1      0->1
calter      0.0127    0.6579    0.6452    0.1173    0.1249    0.0076
female      0.1631    0.0904   -0.0727    0.1631    0.0904   -0.0727

             from:       to:      dif:     from:       to:      dif:
             x-1/2     x+1/2     -+1/2   x-1/2sd   x+1/2sd    -+sd/2  MargEfct
calter      0.1137    0.1211    0.0074    0.0668    0.1978    0.1310    0.0074
female      0.1569    0.0867   -0.0703    0.1358    0.1011   -0.0347   -0.0698


           Nein      Ja
Pr(y|x)  0.8827  0.1173

         calter   female
   x=   2.4e-07   .567839
sd_x=   17.2893   .495792
```

Der Befehl `prchange` gibt noch zusätzliche Information über die Veränderung der Wahrscheinlichkeiten an. Für das Alter ist die Differenz zwischen minimalem und maximalem Wert (`min->max`) 0.6452 und somit 65 %. Für die Dummyvariable `female` kann man eine Veränderung von 0.1631 auf 0.0904 oder eine Veränderung von 7.3 Prozentpunkten feststellen. Mit dem Befehl `help prchange` kann man sich die Erklärung der weiteren Masszahlen angeben lassen.

Der *AME* ist nicht von unbeobachteter Heterogenität verzerrt und ist geeignet um Koeffizienten schrittweise aufgebauter Modelle zu vergleichen. Der *MEM* verändert sich, sobald in das Modell unkorrelierte Prädiktoren aufgenommen werden und ist für den Vergleich unterschiedlicher Modelle nicht geeignet.

Für eine weiterführende Diskussion siehe (Best und Wolf 2010: 840).

6.7 Maximum Likelihood Schätzung

Zur Berechnung der logarithmierten Chancen ist das "Maximum-Likelihood-Prinzip" das effizienteste. Eine OLS-Schätzung ignoriert die Heterogenität der Fehler und den nichtlinearen Verlauf der Wahrscheinlichkeiten. Die β-Koeffizienten werden so bestimmt, dass die beobachteten Anteilswerte maximal wahrscheinlich werden. Die Wahrscheinlichkeit, dass in einer Stichprobe mit n Fällen, ein Ereignis h-mal Auftritt beträgt:

$$P(h|\pi, n) = \left(\begin{array}{c} n \\ h \end{array} \right) \pi^h (1-\pi)^{n-h} \text{ mit } 0 < \pi < 1$$

π (gesprochen "Pi") ist der Anteil des dichotomen Merkmals in der Grundgesamtheit. In der Praxis ist allerdings nicht h von Interesse, sondern π, der Wert, der geschätzt wird,

da er unbekannt ist. Bei der Maximum-Likelihood-Methode wird nun gefragt, welcher Wert von π die gegebene Stichprobe (mit n und h) am wahrscheinlichsten macht. Gefragt ist nach dem Wert von π bei dem die *Likelihood* maximal wird.

$$\mathcal{L}(\pi|h,n) = \left(\begin{array}{c} n \\ h \end{array} \right) \pi^h (1-\pi)^{n-h} \text{ mit } 0 < \pi < 1 \tag{19}$$

Wir nehmen an, dass π durch eine Linearkombination mit unabhängigen Variablen dargestellt werden kann, so dass

$$\pi = \hat{P}(Y=1) = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_{K-1}x_{K-1,i}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_{K-1}x_{K-1,i}}}$$

π kann nun in die Likelihoodgleichung (19) eingefügt werden, wobei $\left(\begin{array}{c} n \\ h \end{array} \right)$ als Konstante für alle π gleich bleibt und nicht berücksichtigt werden muss.

$$\begin{aligned} \mathcal{L} &= \pi^h (1-\pi)^{n-h} \\ &= \left(\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_{K-1}x_{K-1,i}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_{K-1}x_{K-1,i}}} \right)^h \times \left(1 - \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_{K-1}x_{K-1,i}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_{K-1}x_{K-1,i}}} \right)^{n-h} \end{aligned} \tag{20}$$

Im Maximum-Likelihood-Schätzverfahren wird die Likelihood $\mathcal{L}$ nun logarithmiert und in eine Log-Likelihood umgewandelt. Die Wahl der Schätzer β_j erfolgt mit Hilfe iterativer Algorithmen, was man sich wie ein "raffiniertes Ausprobieren" (Kohler und Kreuter 2008: S. 278) vorstellen kann. Analytische Lösungen wie bei der Kleinst-Quadrate-Methode sind beim Maximum-Likelihood-Verfahren nicht mehr möglich. Ein maximaler Likelihood-Wert ist dann bestimmt, wenn sich die Änderungen der Likelihood durch weitere Iterationsschritte nicht mehr verbessern lassen.

6.8 Der Iterationsblock

Bei einer logistischen Regression mit dem Befehl `logit` erscheint links oben der Iterationsblock. Für das Beispiel des Bluthochdrucks sah dies so aus:

```
* Logistische Regression
logit blut calter
```

```
Iteration 0:   log likelihood = -277.65715
Iteration 1:   log likelihood = -233.59213
Iteration 2:   log likelihood = -228.84803
Iteration 3:   log likelihood = -228.81891
Iteration 4:   log likelihood =  -228.8189

Logistic regression                                     Number of obs   =        597
                                                        LR chi2(1)      =      97.68
                                                        Prob > chi2     =     0.0000
Log likelihood =  -228.8189                             Pseudo R2       =     0.1759

------------------------------------------------------------------------------
        blut |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      calter |   .0691486   .0079855     8.66   0.000     .0534972       .0848
       _cons |  -1.965034   .1477202   -13.30   0.000    -2.254561   -1.675508
------------------------------------------------------------------------------
```

Man sieht, dass die Iterationsschritte fortlaufen bis es praktisch keine Veränderung des Likelihood-Wertes mehr gibt. [4] Gefragt wird, wie wahrscheinlich die beobachteten Daten sind, wenn die β_j unterschiedliche Werte annehmen. Die Koeffizienten werden so lange variiert, bis die Likelihood maximal wird. Der Iterationswert 0 oder auch $\mathcal{L}_0$ entspricht dem Wert, bei dem alle β-Koeffizienten, ausser der Konstanten, keinen Einfluss haben und somit keine Steigung aufweisen – der Wert der Koeffizienten ist dann 0. Der letzte Iterationsschritt $\mathcal{L}_K$ gibt an, wie wahrscheinlich die beobachteten Daten sind, wenn die angegebenen β-Koeffizienten verwendet werden. Je stärker sich das "Nullmodell" $\mathcal{L}_0$ von dem ermittelten Modell $\mathcal{L}_K$ unterscheidet, umso grösser ist die Erklärungskraft der gewählten Variablen. Man könnte $\mathcal{L}_0$ mit dem TSS der OLS-Schätzung und $\mathcal{L}_K$ mit dem RSS vergleichen. Die Differenz $\mathcal{L}_0 - \mathcal{L}_K$ entspricht dann der durch das Modell erklärten Varianz (MSS).

Mit Hinzunahmen von unabhängigen Variablen nehmen auch die Iterationsschritte zu. Ein Modelle, das viele Iterationsschritte benötigt, deutet darauf hin, dass keine eindeutige Lösung gefunden werden kann und das Modell nicht konvergiert. Dies ist nur selten der Fall und kann durch Entfernung von unabhängigen Variablen aus dem Modell behoben werden.

[4] In Iteration 4 gibt es im Vergleich zu Iteration 3 nur noch eine Veränderung in der fünften Nachkommastelle.

6.9 Die Modellgüte

6.9.1 Modellfit-Block

Der Anteil der erklärten Varianz $R^2 = \frac{MSS}{TSS}$ gilt als Gütemass des Modells. In der logistischen Regression spricht man von einem "pseudo R^2", welches von McFadden 1973 vorgeschlagen wurde.[5]

$$R^2_{MF} = \frac{ln\mathcal{L}_0 - ln\mathcal{L}_K}{ln\mathcal{L}_0} = 1 - \frac{ln\mathcal{L}_K}{ln\mathcal{L}_0} \tag{21}$$

Der Wertebereich liegt zwischen 0 und 1 und zur inhaltlichen Interpretation kann man allerdings nur sagen, dass höhere Werte eine bessere Modellgüte bedeuten. Der Wert $R^2_{MF} = 0.18$ aus dem Beispiel ist relativ klein.

Neben dem R^2_{MF} kann man auch den Likelihood-Ratio-χ^2 Wert ($\chi^2_{\mathcal{L}}$) betrachten, der die mit -2 multiplizierte Differenz der Likelihoods ist und einer χ^2 ("Chi-Quadrat") Verteilung folgt.

$$\chi^2_{\mathcal{L}} = -2(ln\mathcal{L}_0 - ln\mathcal{L}_K) \tag{22}$$

Getestet wird die Hypothese, dass die unabhängigen Variablen keine Erklärungskraft liefern, was bedeutet, dass alle Koeffizienten ausser der Konstanten Null sind (Nullhypothese). Wie wahrscheinlich diese Hypothese ist wird unter "Prob > Chi2" ausgewiesen. Ist der Wert kleiner 0.05 kann die Hypothese zurückgewiesen werden, da die Wahrscheinlichkeit sehr gering ist.

6.9.2 Klassifikationstabelle

Gefragt wird nach der Güte der Vorhersage durch *Sensitivität*, *Spezifizität* und *Count* R^2. Mit dem Befehl `estat classification` kann man sich die Klassifikationstabelle ausgeben lassen:

```
quietly logit blut calter // output wird nicht angezeigt
estat classification

Logistic model for blut
```

[5] Es gibt noch weitere R^2-Masse, die man sich mit dem Befehl `fitstat` (Ado-Package fitstat) ausgeben lassen kann.

```
             -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        12             18  |         30
     -     |        93            474  |        567
-----------+--------------------------+-----------
   Total   |       105            492  |        597

Classified + if predicted Pr(D) >= .5
True D defined as blut != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   11.43%
Specificity                     Pr( -|~D)   96.34%
Positive predictive value       Pr( D| +)   40.00%
Negative predictive value       Pr(~D| -)   83.60%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    3.66%
False - rate for true D         Pr( -| D)   88.57%
False + rate for classified +   Pr(~D| +)   60.00%
False - rate for classified -   Pr( D| -)   16.40%
--------------------------------------------------
Correctly classified                        81.41%
--------------------------------------------------
```

Sensitivität ist der Anteil der Personen mit Bluthochdruck, die richtig vorhergesagt wurden (12/105=11.43 %), *Spezifizität* ist der Anteil der Bevölkerung, der keinen Bluthochdruck hat und auch so klassifiziert wurde (474/492=96.34 %). Insgesamt wurden $R^2_{count} = \frac{12+474}{597} = 0.8141$ oder 81.41 % Beobachtungen korrekt spezifiziert. Was nicht schwer ist, da man auch für alle Personen keinen Bluthochdruck vorhersagen könnte und damit schon $\frac{492}{597} = 0.824$ der Beobachtungen richtig spezifiziert hat. In unserem Fall verschlechtert sich die Vorhersage sogar. Dies lässt sich auch durch den "Adjustierten Count R^2" zeigen. Hierbei wird um die häufigste Kategorie der abhängigen Variable (Bluthochdruck vs. kein Bluthochdruch) korrigiert. $R^2_{count_{adj}} = \frac{(12+474)-492}{(597-492)} = -0.057$. Die Interpretation lautet somit: Unter Kenntnis der unabhängigen Variable erhöht sich der Fehler (bzw. verringert sich die Vorhersage) um 5.7 % im Vergleich zur Vorhersage allein aufgrund der Randverteilung der abhängigen Variable. Die Kenntnis der unabhängigen Variable verbessert die Vorhersage also nicht. Dies liegt daran, dass bei der Vorhersage der Personen, die Bluthochdruck haben nur 40 % und damit weniger als die Hälfte der Personen auch tatsächlich Bluthochdruck haben (12/30 = 0.4). Die Wahrscheinlichkeit, eine Person als Bluthochdruck einzuordnen, obwohl sie keinen Bluthochdruck hat (`False + rate for classified +`) beträgt über 50 Prozent (60 %), damit produziert die Vorhersage viele Fehler.

Mit dem Befehl `fitstat` des Ado-Packages von Scott Long und Jeremey Freese erhalten wir weitere – hier nicht näher erläuterte – Fit-Masse.

```
fitstat

Measures of Fit for logit of blut

Log-Lik Intercept Only:       -277.657   Log-Lik Full Model:          -228.819
D(595):                        457.638   LR(1):                        97.676
                                         Prob > LR:                     0.000
McFadden's R2:                   0.176   McFadden's Adj R2:             0.169
ML (Cox-Snell) R2:               0.151   Cragg-Uhler(Nagelkerke) R2:    0.249
McKelvey & Zavoina's R2:         0.303   Efron's R2:                    0.165
Variance of y*:                  4.719   Variance of error:             3.290
Count R2:                        0.814   Adj Count R2:                 -0.057
AIC:                             0.773   AIC*n:                       461.638
BIC:                         -3345.553   BIC':                        -91.285
BIC used by Stata:             470.422   AIC used by Stata:           461.638
```

6.9.3 Likelihood Ratio Test

Der Likelihood Ratio Test testet, ob die Hinzunahme einer Variable die Bedeutung des Modells erhöht, sich also die Likelihood verringert. Wir untersuchen, ob sich die Erklärungskraft durch Hinzunahme weiterer Variablen verbessert. In unserem Fall wird zusätzlich das Geschlecht aufgenommen.

Die Nullhypothese des Likelihood-Ration Tests lautet: Die Hinzunahme der Variable ändert das Modell nicht. Wir vergleichen daher die Likelihood eines Nullmodells (ohne der Variable) mit der Likelihood eines vollen Modells (mit der neuen Variable). Wenn die Differenz einen kritischen Wert überschreitet, kann die Nullhypothese abgelehnt werden und die Hinzunahme der Variable erhöht die Erklärungskraft des Modells.

```
logit blut calter female

Iteration 0:   log likelihood = -277.65715
Iteration 1:   log likelihood =  -230.5305
Iteration 2:   log likelihood = -224.90829
Iteration 3:   log likelihood = -224.85403
Iteration 4:   log likelihood =   -224.854
Iteration 5:   log likelihood =   -224.854

Logistic regression                               Number of obs   =        597
                                                  LR chi2(2)      =     105.61
                                                  Prob > chi2     =     0.0000
Log likelihood =   -224.854                       Pseudo R2       =     0.1902

------------------------------------------------------------------------------
        blut |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
      calter |    .071515   .0082214     8.70   0.000     .0554013    .0876287
      female |  -.6738074   .2412262    -2.79   0.005    -1.146602   -.2010127
       _cons |  -1.635452   .1824132    -8.97   0.000    -1.992976   -1.277929
------------------------------------------------------------------------------
estimates store full  // Speichere volles Modell
```

Berechnet und gespeichert (`estimates store`) wird ein Modell mit allen unabhängigen Variablen (`full`: Alter, Geschlecht) und ein Modell mit reduzierten Variablen (Alter).

```
quietly logit blut calter  // reduziertes Modell
lrtest full // Likelihood Ratio Test
```

Anschliessend wird die Differenz der zugehörigen logarithmierten Likelihoods mit -2 multipliziert, um eine Prüfgrösse zu erhalten. Dies geschieht mit dem Befehl `lrtest` *`modelname`*. Die Prüfgrösse folgt einer Chi-Quadrat-Verteilung ($\chi^2_{\mathcal{L}(\text{Diff})}$).

$$\chi^2_{\mathcal{L}(\text{Diff})} = -2(\ln\mathcal{L}_{\text{ohne Geschlecht}} - \ln\mathcal{L}_{\text{mit Geschlecht}}) \tag{23}$$

Die Anzahl der Freiheitsgrade ist die Differenz der Anzahl der Parameter zwischen den Modellen.

```
lrtest full                              // Likelihood Ratio-Test

Likelihood-ratio test                                 LR chi2(1)  =      7.93
(Assumption: . nested in full)                        Prob > chi2 =    0.0049
```

Wir erhalten einen Wert von 7.93. Die Wahrscheinlichkeit diesen Wert zu erhalten, wenn der Koeffizient des Alters in der Grundgesamtheit 0 beträgt ist 0.0049 und damit sehr klein. Wir können damit bestimmen ob der Einfluss des Alters in der Grundgesamtheit von 0 verschieden ist – über die Stärke des Einflusses können wir nichts sagen.

Der Likelihood-Ratio Test kann nur Modelle vergleichen, wenn das `full` Modell alle Variablen enthält, die auch im reduzierten Modell enthalten sind. Die Modelle müssen *hierarchisch geschichtet* sein. Ferner bezieht es sich auf die gleiche Anzahl der Personen. Gerade bei reduzierten Modellen, sollte man kontrollieren, ob sich die Fallzahlen unterscheiden. Ist dies der Fall, so zeigt `Stata` dies durch die Angabe `observation differ` an.

6.9.4 Test für nicht geschichtete Modelle BIC und AIC

Möchte man nicht hierarchisch geschichtete Modelle vergleichen kann man das "Information-Criteria" verwenden. Am häufigsten wird das "Bayesian Information Criterion" (BIC) und das "Akaike's Information Criterion" (AIC) verwendet. Zu erhalten ist dies durch `estat ic`. Das BIC und das AIC helfen ein Modell zu finden, das möglichst gut an die Daten angepasst ist und gleichzeitig sparsam ist. Daher wird die Hinzunahme von weiteren unabhängigen Variablen bestraft – ähnlich dem korrigierten R^2 der OLS-Regression. Das BIC bestraft stärker als das AIC. Je kleiner die Werte, umso besser der Fit des Modells.

```
quietly logit blut calter female
estat ic
```

```
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    597   -277.6572    -224.854      3      455.708    468.8837
-----------------------------------------------------------------------------
```

```
quietly logit blut calter
estat ic
```

```
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    597   -277.6572   -228.8189      2     461.6378    470.4216
-----------------------------------------------------------------------------
```

Das BIC und das AIC sind im Modell mit Geschlecht niedriger

Übersicht Stata-Befehle

Befehl	Option	Erklärung
`logit`	or	Logitische Regression. Mit der Option `or` kann man sich Odds Ratios statt Logits anzeigen lassen.
`logistic`		Logistische Regression wie `logit`, nur dass automatisch Odds Ratios ausgegeben werden.
`estat classification`		Ausgabe einer Klassifikationstabelle
`fitstat`		Ado-Package fitstat von Scott Long und Jeremy Freese
`lrtest`		Likelihood Ratio Test
`estat ic`		Information Criteria BIC und AIC

Aufgabe

Bitte betrachtet 3-5 dichotome Variablen, die als abhängige Variable für eine beliebige Fragestellung interessant sein könnten. Findet im Datensatz weitere metrische unabhängige Variablen und überlegt euch ein lineares Erklärungsmodell für eine oder mehrere dichotome abhängige Variablen. Zentriert die metrischen Variablen und stellt den Zusammenhang zwischen dichotomer AV und metrischer UV in einem Scatterplot (`twoway(scatter)`) dar und versucht den Zusammenhang mit weiteren graphischen Verfahren (Regressionslinie, nicht-parametrische Darstellungsformen) darzustellen. Welche Probleme könnten bei der Interpretation des Zusammenhangs auftreten? Warum ist eine logistische Regression sinnvoll?

7 18.4. – Generalisierte Lineare Regressionsmodelle 2

7.1 Regression von Zähldaten

Als einführende Literatur in die Regression mit Zähldaten bietet sich Tutz (2010) an. Eine hilfreiche Website ist auch http://data.princeton.edu/wws509/stata/overdispersion.html.

Wenn man die Häufigkeiten des Auftretens eines interessierenden Ereignisses betrachtet spricht man von Zähldaten. Es gilt die diskrete Struktur der Daten zu erfassen, was nicht mit den klassischen Verfahren der OLS-Regression mit normalverteilten Fehlern gelingt. Problematisch ist die Anwendung der klassischen Regression, da Zähldaten nur ganze Werte (0, 1, 2 ...) annehmen und die Anzahl eines Ereignisses nie negativ sein kann. Es handelt sich also nicht um eine stetige metrisch skalierte Variable mit sowohl negativen Werten, als auch Dezimalwerten.

Klassische Anwendungsbeispiele sind die Anzahl von Kindern in einer Familie, die Anzahl der Arztbesuche oder Kriminalitätsdelikte einer Person oder als Wohlstandsindikator die Urlaubstage im Ausland. Ferner werden mit Zähldatenmodellen die Anzahl von wissenschaftlichen Publikationen junger Wissenschaftler geschätzt. Ziel ist es nun die Häufigkeit des Auftretens eines Ereignisses sowohl durch metrische als auch nominal skalierte (Dummy-) Variablen zu erklären.

Da Zählvariablen generell wenige Ausprägungen haben (Anzahl Kinder, Autounfälle pro Jahr) und diskret sind, ist die Variable nicht normalverteilt. Zählvariablen mit grossen Werten (grösser als 30) hingegen sind approximativ Normalverteilt und man kann eine klassische Regression in Erwägung ziehen.

Verteilungen, die sich zur Analyse anbieten sind die Poisson-, die Quasipoisson und die negative Binomialverteilung. Darüber hinaus gibt es weitere Verteilungen wie die Gammaverteilung, die hier aber nicht behandelt wird.

7.1.1 Poissonverteilung

Die Poissonverteilung wird auch die Verteilung der seltenen Ereignisse genannt. Sie hat die Dichtefunktion:

$$f(y) = P(y = 1) = \frac{\lambda^y}{y!} e^{-\lambda} \quad \text{für} \quad y \in \{0, 1, 2, ...\} \tag{24}$$

Die Besonderheit (und die Schwäche) der Verteilung ist, dass sie nur einen Parameter λ (gesprochen "Lambda") besitzt. λ entspricht dem Erwartungswert der abhängigen Variable, d. h. $E(Y) = \lambda$ und sagt aus mit wie vielen Ereignissen im Mittel zu rechnen ist.

Da es nur einen Parameter gibt sind Erwartungswert und Varianz der Poissonverteilung identisch, $E(Y) = var(Y) = \lambda$. Diese Besonderheit der Poissonverteilung wird auch als *Equidispersion* bezeichnet. Diese Annahme lässt sich rechtfertigen, da man bei seltenen Ereignissen auch keine hohe Variabilität erwarten kann. Allerdings kann diese Annahme bei einem hohen Erwartungswert leicht verletzt sein.

Zwei Fälle können auftreten, bei denen die Annahme der Equidispersion verletzt sein kann:

- *overdispersion* liegt vor, wenn die Variabilität in den Daten grösser ist, als durch das Modell spezifiziert;
- *underdispersion* liegt zwar seltener, aber dann vor, wenn die Variabilität geringer ist.

Im Falle der Verletzung dieser Annahme kann das Poissonmodell mit einem Dispersionparameter zum Quasi-Poissonmodell erweitert werden, wobei der Dispersionparameter dann zusätzlich geschätzt wird. Mehr zum Quasi-Poissionmodell auf Seite 120.

7.1.2 Negativ Binomialverteilung

Ein flexibleres Modell ist die Anwendung der negativen Binomialverteilung, die durch zwei Einflussgrössen ν, $\mu > 0$ bestimmt wird. Der Erwartungswert und die Varianz der Verteilung sind definiert als:

$$E(Y) = \mu, \quad var(Y) = \mu + \frac{\mu^2}{\nu}$$

Damit ist eine flexiblere Modellierung der Varianz möglich und nur im Extremfall, wenn ν ("nü") sehr gross wird ist das Modell der negativen Binomialverteilung identisch mit dem der Poissonverteilung.

7.1.3 Modellierung der Zähldatenregression

Bei der Zähldatenregression muss der Erwartungswert der abhängigen Variable so modelliert werden, dass er nicht negativ sein kann. Anders als bei der logistischen Regression, muss der Wertebereich aber nicht zwischen 0 und 1 liegen, sonder kann alle Werte $\mu > 0$ annehmen. Dennoch muss bei Zähldaten spezifiziert werden, wie die gegebenen unabhängigen Variablen auf die abhängige Variable wirken. Als *Transformationsfunktion* des linearen Modells $\beta_0 + \beta_1 x_1 + ... + \beta_{k-1} x_{k-1}$ bietet sich das *loglineare Modell* an:

$$
\begin{aligned}
E(Y) = \mu &= e^{\beta_0+\beta_1 x_1+...+\beta_{k-1}x_{k-1}} \\
log(\mu) &= \beta_0 + \beta_1 x_1 + ... + \beta_{k-1}x_{k-1}
\end{aligned}
$$

Damit wirken die unabhängigen Variablen des Modells bzw. der Linearkombination auf den logarithmierten Erwartungswert. Für die Interpretation der β-Koeffizienten heisst dies, dass sich bei Erhöhung der unabhängigen Variable x_j um eine Einheit, der Erwartungswert um den Faktor e^{β_j} verändert.

- $\beta_j = 0$: keine Wirkung der unabhängigen Variable auf den Erwartungswert der auftretenden Häufigkeit,
- $\beta_j > 0$: der Erwartungswert erhöht sich, da der Faktor $e^{\beta_j} > 1$ und somit eine vergrössernde Wirkung hat,
- $\beta_j < 0$: der Erwartungswert verringert sich, da $e^{\beta_j} < 1$, der Faktor verringert den Erwartungswert.

Zusammengefasst haben die Parameter eine multiplikative Wirkung auf den Erwartungswert der abhängigen Variable. Beispielsweise hat ein $\hat{\beta}_j$-Wert von 0.08 eine Erhöhung um den Faktor $e^{0.08} = 1.083$ zur Folge. Damit erhöht sich die Häufigkeit der abhängigen Variable pro Einheit der unabhängigen Variable um 8.3 %.

Als Verteilungsannahme kann man nur bei Zählvariablen mit grossen Werten, die annähernd symmetrisch verteilt sind die Normalverteilung als Approximation verwenden. Generell muss man aber von einer nicht-normalverteilten Variable ausgehen, da die Ausprägungen diskret und nur geringe Werte mit wenigen Ausprägungen über Null annehmen. Die am Häufigsten verwendeten Verteilungsannahmen sind die:

- Poissonverteilung,
- Poissonverteilung mit Dispersionsparamtert (Quasi-Poissonmodell),
- Negative Binomialverteilung.

Die Poissonverteilung hat den Nachteil, dass sie durch einen Parameter bestimmt ist, daher sollte man lieber die negative Binomialverteilung verwenden, obwohl es auch hier Probleme bei der Konvergenz des Modells geben kann und keine eindeutige Lösung bestimmt werden kann. Alternativ kann man auch ein Poissonmodell mit robusten Standardfehlern berechnen. Ebenso kann man eine Poissonmodell mit Dispersionsparameter verwenden (Quasi-Poissonmodell).

Poissonmodell Um die Signifikanz einzelner Parameter des Modells zu testen wird ein Score Test durchgeführt. Dabei wird ein z-Wert bestimmt der angibt, ob die Nullhypothese, dass die unabhängige Variable keinen Einfluss auf die abhängige Variable hat, abgelehnt werden kann. Bei grossen und sehr kleinen Werten kann die Nullhypothese abgelehnt werden, dann, wenn der Betrag $|z|$ grösser als das $(1-\alpha/2)$-Quantil der Standardnormalverteilung ist. Der z-Wert wird bestimmt durch:

$$z = \frac{\hat{\beta}_j}{\hat{s}(\hat{\beta}_j)}$$

Quasi-Poissonmodell Im Quasi-Poissonmodell wirkt auf die Varianz noch multiplikativ ein Dispersionsparameter "Phi" (ϕ): $var(Y_i) = \phi\mu_i$. Der dazugehörige z-Wert, genauer der geschätzte Standardfehler $\hat{s}(\hat{\beta}_j)$ des Parameter β_j wird um die Wurzel des Dispersionsparameters erweitert:

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{\phi}} \times \hat{s}(\hat{\beta}_j)}$$

Negativ Binomialverteilung Im negativen Binomialmodell hängt die Varianz nicht nur von dem Erwartungswert μ ab,

$$var(Y_i) = \mu_i + \frac{\mu_i^2}{\nu},$$

sondern auch vom zu schätzenden Parameter ν (gesprochen "nü"). Er wirkt nicht multiplikativ, macht das Modell aber flexibler. Im Extremfall, dann wenn ν sehr gross wird, entspricht die negative Binomialverteilung dem Poissonmodell.

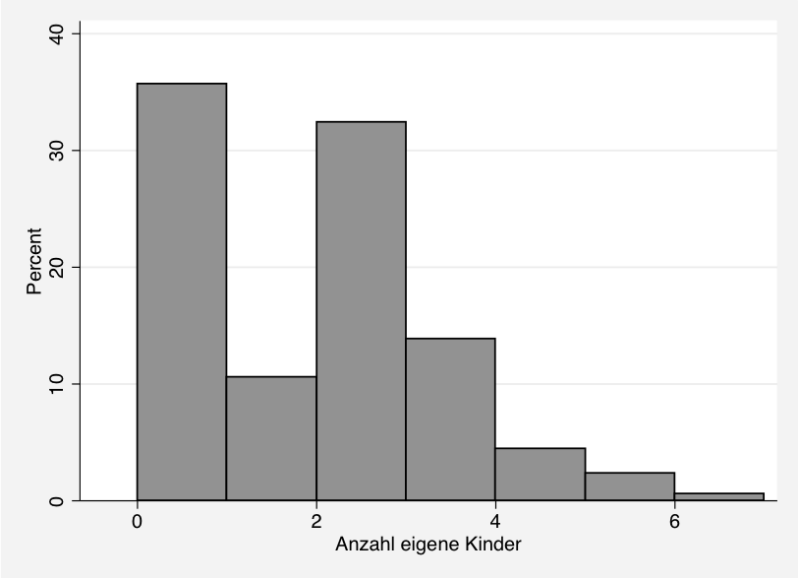
7.2 Beispiel: Anzahl Kinder in der Schweiz

Als Beispiel wird mit den Daten des Schweizer Umwelzsurvey 2007 (Reduziert auf 600 Fälle) die Anzahl der Kinder der Frauen geschätzt. Die Variable hat einen Mittelwert von 1.50 und eine Varianz von $1.39^2 = 1.93$:

```
sum kinder if female==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      kinder |       339    1.504425    1.393925          0          7
```

Die Verteilung lässt sich durch ein Histogramm darstellen, durch das auch die extreme rechtsschiefe Verteilung der Kinderhäufigkeit sichtbar wird.

```
histogram kinder
```



7.2.1 Vorbereitung für die Zähldatenregression

Als erklärende Variablen dienen:

- das Alter in Jahren (metrische Variable),
- die Nationalität (Dummy-Variable, Schweizerin = 1),
- der Bildungsgrad (quasi metrisch mit 16 ordinalen Kategorien)
- die Mitgliedschaft in der Kirche (Dummy Variable, Mitglied = 1).

Zu spezifizieren ist, wie die Variablen/Einflussgrössen wirken. In diesem Fall wird für das Alter angenommen, dass es nicht linear wirkt. Daher werden in den linearen Prädiktor `alter` noch *polynomiale Terme* aufgenommen. In diesem Fall der Ordnung vier (`alter`^{2}, `alter`^{3}, `alter`^{4}). Dadurch dass der Einfluss jetzt nicht mehr linear festgelegt ist (ein Beispiel für Linearität wäre: mit jedem zusätzlichen Jahr erhöht sich die Anzahl der Kinder um einen Faktor e^{β_j}), ist eine flexiblere Modellierung des Einflusses möglich.

```
* Alter
clonevar alter = talter

* Geschlecht
tab tsex
clonevar female = tsex

* Alter als polynomiale Terme der Ordnung vier
generate alter2 = alter^2
generate alter3 = alter^3
generate alter4 = alter^4

* AV Kinderzahl
clonevar kinder = tkinder
histogram kinder, width(1)

* Nationalität
codebook tnat1
tab tnat1
recode tnat1 (1=1) (2/4=0), gen(schweiz)
label var schweiz "Nationalität (1=Schweiz)"
label def schweiz 1 "Schweizer/in" 0 "Ausländer/in", modify
label val schweiz schweiz
tab schweiz

* Mitgliedschaft Kirche
clonevar kirche =ssoz8

* Mitgliedschaft Kirche (Dummy für Frauen)
clonevar female_kirche = kirche
replace female_kirche =. if female==0
tab female_kirche if  female_kirche !=., mis
```

Nachdem die Variablen spezifiziert sind wird der *Verteilungstyp* festgelegt. In diesem Fall kommen die Poissonverteilung, das Quasi-Poissonmodell mit Dispersionparamter und die negative Binomialverteilung in Betracht. Als *Transformationsfunktion* wird die Exponentialfunktion (e^{β_j}) festgelegt. Damit handelt es sich um loglineare Zähldatenmodelle.

7.2.2 Zähldatenmodelle in Stata

Poissonmodell Als erstes wird der Einfluss der unabhängigen Variablen durch eine Poissonverteilung spezifieziert. In `Stata` gibt es zwei Möglichkeiten die Regression zu definieren. Zum einen kann man den Befehl `poisson` verwenden. Mit `irr` wird der

antilogarithmierte Effekt (incidence-rate ratios) ausgegeben, so dass eine Wahrscheinlichkeitsinterpretation leichter möglich ist, mit `nolog` werden die Likelihood-Iterationen unterdrückt.

```
poisson kinder alter alter2 alter3 alter4 schweiz ///
            teduc kirche if female==1, irr nolog
```

```
Poisson regression                                Number of obs   =        162
                                                  LR chi2(7)      =     109.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -224.60112                       Pseudo R2       =     0.1962

------------------------------------------------------------------------------
      kinder |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alter |   57.45239   50.99601     4.56   0.000     10.08702    327.2302
      alter2 |   .8932364   .0227533    -4.43   0.000     .8497357     .938964
      alter3 |   1.001361   .0003153     4.32   0.000     1.000743    1.001979
      alter4 |    .999994   1.42e-06    -4.21   0.000     .9999913    .9999968
     schweiz |   1.107138   .2276356     0.50   0.621     .7399268     1.65659
       teduc |   .9864143   .0172842    -0.78   0.435     .9531131    1.020879
      kirche |   1.730596   .2556398     3.71   0.000     1.295562     2.31171
------------------------------------------------------------------------------
```

Besser gefällt mir der Befehl `glm` (für Generalisierte Lineare Modelle) bei dem man die Verteilungsfamilie (hier `family(poisson)`) und die Transformations- bzw. Linkfuktion (`link(log)`) definieren muss. Mit `eform` kann man sich die antilogarithmierten Koeffizienten ausgeben lassen.

```
glm kinder alter alter2 alter3 alter4 schweiz teduc kirche if female==1, ///
>         family(poisson) link(log) eform nolog
```

```
Generalized linear models                         No. of obs      =       162
Optimization     : ML                             Residual df     =       154
                                                  Scale parameter =         1
Deviance         =  177.9909913                   (1/df) Deviance =  1.155786
Pearson          =  144.1887637                   (1/df) Pearson  =  .9362907

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  2.871619
Log likelihood   = -224.6011232                   BIC             = -605.4988
```

```
------------------------------------------------------------------------------
             |                 OIM
      kinder |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alter |    57.4524   50.99602     4.56   0.000     10.08702    327.2303
      alter2 |   .8932364   .0227533    -4.43   0.000     .8497357     .938964
      alter3 |   1.001361   .0003153     4.32   0.000     1.000743    1.001979
      alter4 |    .999994   1.42e-06    -4.21   0.000     .9999913    .9999968
     schweiz |   1.107138   .2276356     0.50   0.621     .7399268     1.65659
       teduc |   .9864143   .0172842    -0.78   0.435     .9531131    1.020879
      kirche |   1.730596   .2556398     3.71   0.000     1.295562     2.31171
------------------------------------------------------------------------------
```

Diese Darstellung hat noch weitere Informationen. Die Devianz gibt die Anpassung des geschätzten Poissonmodells $\hat{\mu}_i$ an die beobachteten Werte y_i an und folgt einer χ^2("Chi²")-Verteilung. Wenn der kritische Wert grösser ist, als der beobachtete Divianz-Wert, dann wurde die Schätzung korrekt spezifiziert. Mit dem Befehl `invchi2tail(`*`Residual df,Sig. Niv.`*`)` wird der kritische Wert der χ^2-Verteilung bestimmt.

```
* Berechne kritischen Wert
. *154 = Freiheitsgrade des Modells, 0.05=5% Singnifikanzniveau
. display invchi2tail(154,0.05)
183.95861
```

Der Wert von 183.96 liegt über dem Wert des Modells von 177.99. Damit hat das Modell eine gute Anpassung und muss nicht modifiziert werden. `(1/df) Pearson = .9362907` zeigt auch, die Varianz um $0.94 - 1 = -0.06$ oder knapp 6 % unterschätzt wird. Das Modell ist also sehr passend. Ein Dispersionsparamtert, der die Standardabweichung korrigiert, liegt bei $\sqrt{0.93629067} = 0.96762114$ und damit nahe 1.

```
 quietly glm kinder alter alter2 alter3 alter4 schweiz teduc kirche if female==1, ///
         family(poisson) link(log) eform nolog
* Dispersionsparameter ‘‘Phi’’
 scalar phi = e(deviance_p)/e(df)


 * Abweichung von Varianz, Dispersionsparameter ‘‘Phi’’
di phi
.93629067


 * Multiplikativer Effekt auf Standardabweichung
 di sqrt(phi)
.96762114
```

In diesem Fall ist ein Quasi-Poissonmodell nicht notwendig. Der Vollständigkeit halber wird es aber auch noch berechnet. Dadurch dass der Dispersionsparameter < 1 ist verringern sich die Standardfehler des Modells geringfügig, die Koeffizientenwerte bleiben aber gleich.

Quasi-Poissonmodell Im Quasi-Poissonmodell ist die Varianz $var(y_i) = \phi\mu_i$. ϕ ("Phi") ist der Dispersionsparameter. Wenn $\phi = 1$, dann entspricht die Varianz der Poissonverteilung, wenn $\phi < 1$ ist die Varianz kleiner (underdispersion) und bei $\phi > 1$ ist die Varianz grösser als bei der Poissonverteilung (overdispersion). Wie oben gezeigt wurde wird die Pearson-Statistik (`Pearson = 144.1887637`)

$$\chi^2_P = \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}$$

verwendet und durch die Anzahl der Freiheitsgrade geteilt, wobei p die Anzahl der geschätzten Parameter des Modells bezeichnet (üblicherweise verwende ich in der Notation der Linearkombination für $p = (k - 1)$, also die Anzahl der geschätzten Koeffizienten abzüglich der Konstanten).

$$\chi^2_P = \frac{1}{n-p} \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}$$

In `Stata` kann man den Dispersionsparameter $\hat{\phi} = 0.93629067$ direkt unter `scale(`*`wert`*`)` einfügen oder man spezifiziert mit `x2`, dass der Pearson's-χ^2 verwendet werden soll.

```
glm kinder alter alter2 alter3 alter4 schweiz teduc kirche if female==1, ///
      family(poisson) link(log) scale(x2) eform nolog

Generalized linear models                         No. of obs      =       162
Optimization     : ML                             Residual df     =       154
                                                  Scale parameter =         1
Deviance         =  177.9909913                   (1/df) Deviance =  1.155786
Pearson          =  144.1887637                   (1/df) Pearson  =  .9362907

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  2.871619
```

```
Log likelihood   = -224.6011232                     BIC            = -605.4988

------------------------------------------------------------------------------
             |                 OIM
      kinder |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alter |    57.4524   49.34483     4.72   0.000     10.67153    309.3071
      alter2 |   .8932364   .0220166    -4.58   0.000     .8511105    .9374473
      alter3 |   1.001361   .0003051     4.46   0.000     1.000763    1.001959
      alter4 |    .999994   1.37e-06    -4.35   0.000     .9999913    .9999967
     schweiz |   1.107138    .220265     0.51   0.609     .7496447    1.635115
       teduc |   .9864143   .0167245    -0.81   0.420     .9541735    1.019744
      kirche |   1.730596   .2473625     3.84   0.000     1.307764    2.290141
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

Negative Binomialmodell Bevor die Effekt des Poissonmodells inhaltlich erklärt werden wird hier noch das negative Binomialmodell vorgestellt. Wie oben erklärt ist die Varianz:

$$E(Y) = \mu, \quad var(Y) = \mu + \frac{\mu^2}{\nu}$$

Auch hier sei angemerkt, dass im Extremfall mit $\nu \to \infty$ das negative Biomialmodell dem Poissonmodell gleicht. ν wird meistens mit Maximum-Likelihood geschätzt.

In `Stata` wird mit dem Befehl `nbreg` die negative Binomialverteilung berechnet.

```
nbreg kinder alter alter2 alter3 alter4 schweiz teduc kirche if female==1, irr nolog

Negative binomial regression                    Number of obs   =        162
                                                LR chi2(7)      =      93.04
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -224.60113                     Pseudo R2       =     0.1716

------------------------------------------------------------------------------
      kinder |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alter |   57.45849   51.00036     4.56   0.000     10.08845    327.2531
      alter2 |   .8932332   .0227526    -4.43   0.000     .8497338    .9389595
      alter3 |   1.001361   .0003153     4.32   0.000     1.000743    1.001979
      alter4 |    .999994   1.42e-06    -4.21   0.000     .9999913    .9999968
```

```
     schweiz |    1.10714   .2276333     0.50   0.621     .7399316    1.656585
       teduc |   .9864149   .0172836    -0.78   0.435     .9531147    1.020879
      kirche |   1.730692   .2556444     3.71   0.000     1.295648    2.311813
-------------+------------------------------------------------------------------
    /lnalpha |  -15.26072   614.0029                     -1218.684    1188.163
-------------+------------------------------------------------------------------
       alpha |   2.36e-07   .0001447                            0           .
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 0.0e+00 Prob>=chibar2 = 0.500
```

Von Interesse ist der `alpha` Wert. Ein Wert von 0 besagt, dass der Dispersionparameter = 1 ist und damit dem Poisson-Modell entspricht. Ebenso kann der logarithmierte `/lnalpha` Wert gegen $-\infty$ gehen. Ein signifikanter `alpha`-Wert > 1 bedeutet, dass overdispersion vorliegt und ein entsprechender Wert < 1 besagt, dass underdispersion vorliegt. Hier ist der Wert nahe bei Null und damit haben wir ein Poissonmodell.

Um die Devianz des Modells zu erhalten kann man wieder den Befehl `glm` verwenden wobei man nun bei `family(nbinomial` *`dispersionparameter`*`)` die negative Binomialverteilung mit Dispersionparameter angibt. So kann man sich die Anpassung des Modells (Devianz) ausgeben lassen und unterschiedliche Modell vergleichen. In unserem Fall können wir auch mit der Poissonverteilung weiterarbeiten, da `alpha`= 0. Siehe auch die Null im output:

```
 Variance function: V(u) = u+(0)u^2.
```

```
* bestimme alpha
quietly nbreg kinder alter alter2 alter3 alter4 schweiz teduc kirche if female==1, irr nolog

*definiere Dispersionsparameter v
local v = e(alpha)

* Setze Dispersionparameter ein
glm kinder alter alter2 alter3 alter4 schweiz teduc kirche if female==1, ///
family(nbinomial `v') link(log)  eform  nolog

* Output
Generalized linear models                         No. of obs      =       162
Optimization     : ML                             Residual df     =       154
                                                  Scale parameter =         1
Deviance         =  177.9909384                   (1/df) Deviance =  1.155785
Pearson          =  144.1887104                   (1/df) Pearson  =  .9362903

Variance function: V(u) = u+(0)u^2                [Neg. Binomial]
Link function    : g(u) = ln(u)                   [Log]
```

```
                                                AIC             =  2.871619
Log likelihood   = -224.6011261                 BIC             = -605.4989

------------------------------------------------------------------------------
             |                 OIM
      kinder |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alter |   57.45241   50.99604     4.56   0.000     10.08702    327.2304
      alter2 |   .8932364   .0227533    -4.43   0.000     .8497357     .938964
      alter3 |   1.001361   .0003153     4.32   0.000     1.000743    1.001979
      alter4 |    .999994   1.42e-06    -4.21   0.000     .9999913    .9999968
     schweiz |   1.107138   .2276356     0.50   0.621     .7399267     1.65659
       teduc |   .9864143   .0172842    -0.78   0.435     .9531131    1.020879
      kirche |   1.730596   .2556399     3.71   0.000     1.295562     2.31171
------------------------------------------------------------------------------
```

7.2.3 Auswertung des Poissonmodells

Zur Erinnerung nochmals die Werte der drei geschätzten Modelle:

```
 esttab poisson q_poisson nbreg, se mtitles("Poisson" "Quasipoisson" "Neg. Binomial" )

------------------------------------------------------------
                      (1)             (2)             (3)
                  Poisson    Quasipoisson    Neg. Binomial
------------------------------------------------------------
kinder
alter               4.051***        4.051***        4.051***
                   (0.888)         (0.859)         (0.888)

alter2             -0.113***       -0.113***       -0.113***
                  (0.0255)        (0.0246)        (0.0255)

alter3            0.00136***      0.00136***      0.00136***
                 (0.000315)      (0.000305)      (0.000315)

alter4        -0.00000597***  -0.00000597***  -0.00000597***
              (0.00000142)    (0.00000137)    (0.00000142)

schweiz              0.102           0.102           0.102
                   (0.206)         (0.199)         (0.206)

teduc              -0.0137         -0.0137         -0.0137
                  (0.0175)        (0.0170)        (0.0175)
```

```
kirche              0.548***        0.548***        0.549***
                   (0.148)         (0.143)         (0.148)

_cons              -52.56***       -52.56***       -52.57***
                   (11.26)         (10.90)         (11.26)
------------------------------------------------------------
lnalpha
_cons                                               -15.26
                                                    (614.0)
------------------------------------------------------------
N                      162             162             162
------------------------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

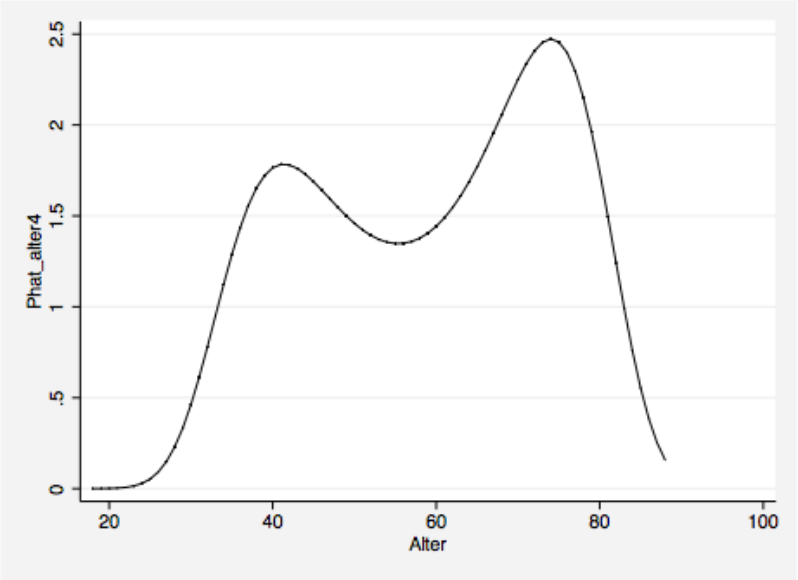
Die geschätzten Koeffizienten sind für alle drei Modelle fast identisch. Ebenso ist das Poissonmodell nahezu identisch dem negativen Binomialmodell. Die Standardfehler des Quasi-Poissonmodells sind etwas geringer, da die Varianz kleiner dem Mittelwert ist. In diesem Modell (2) sind die Koeffizienten daher eher signifikant. Nur zur Anmerkung: Bei overdispersion, also wenn die Varianz grösser als der Mittelwert ist, sind die Koeffizienten im Quasi-Modell weniger signifikant, da sie grössere Standardfehler haben.

Die Erklärung der Effekt im Einzelnen. Hier handelt es sich noch um logarithmierte Effekte, die zur sinnvollen Interpretation erst transformiert werden müssen:

- Die **Nationalität** ist nicht signifikant. Der Effekt $e^{0.102} = 1.107138$ würde bei Signifikanz sogar bedeuten, dass im Mittel bei Schweizerinnen eine um den Faktor = 1.10 höhere Kinderzahl zu erwarten wäre, im Verhältnis zu Nicht-Schweizerinnen. Der Koeffizientenwert 0.102 selber zeigt die mittlere Erhöhung des logarithmierten Erwartungswertes der Kinderzahl.
- **Bildung** hat keinen Einfluss, allerdings deutet der leicht negative Wert (-0.0137) darauf hin, dass sich mit steigender Bildung die Kinderzahl verringert (allerdings sollte auch die ungenügende Operationalisierung der Bildung in diesem Beispiel beachtete werden).
- Mitgliedschaft in der **Kirche** hat einen positiven signifikanten Effekt auf die Häufigkeit der Kinder. Im Mittel bekommen Frauen, dann eine um den Faktor $e^{0.548} = 1.730596$ erhöhte Kinderzahl.
- Alle polynomialen Terme der Variable **Alter** sind hochsignifikant und daher nicht zu vernachlässigen. Eine einfache lineare Spezifikation hätte die Datenstruktur nicht genau abgebildet. Die Berechnung des Alterseffekts mit polynomialen Termen hat den Nachteil, dass man die Wirkungsweise nicht direkt aus den Koeffizienten

ablesen kann. Eine graphische Darstellung des Effekts des Alters auf die Kinderzahl ist aber möglich.

```
generate Phat_alter4 = exp(_b[_cons]+_b[alter]*alter ///
  +_b[alter2]*alter2+_b[alter3]*alter3+_b[alter4]*alter4)

graph twoway line Phat_alter4 alter, sort
graph export Phat_alter4.png, replace
```



Man sieht, dass im Alter von 20 bis 40 die Anzahl der zu erwartenden Kinder zunimmt, dann relativ stabil bleibt und für Frauen über 65 nochmals ansteigt. Vermutlich haben wir hier noch einen Kohorteneffekt, so dass Frauen der älteren Generation noch mehr Kinder bekommen haben. Die Spezifikation mit lediglich einem linearen Term hätte die den starken Anstieg der Häufigkeit zwischen 20 und 40 Jahren nicht dargestellt.

7.2.4 Devianzanalyse

Eine Devianzanalyse untersucht die Relevanz der einzelnen Einflussgrössen des Modells. Betrachtet wird jeweils die Devianz des vollen Modell verglichen mit der Devianz des um die unabhängige(n) Variable(n) reduzierten Modells. Die Differenz der Devianz entspricht einem Likelihood-ratio-Test `(lrtest)`. Getestet wird, ob die Reduktion eines Merkmales zu einem signifikanten Unterschied führt.

Im ersten Schritt wird das volle Modell spezifiziert:

```
* Devianz
```

```
* volles Modell
quietly poisson kinder alter alter2 alter3 alter4 schweiz teduc kirche if female==1
estimates store full
* estat gof // zeigt Devianz des Modells an
```

Anschliessend wird die Variable Alter weggelassen und ein Likelihood-ratio-Test durchgeführt:

```
* Ohne Alter
quietly poisson kinder schweiz teduc kirche if female==1
* estat gof // zeigt Devianz des Modells an
lrtest full

Likelihood-ratio test                                 LR chi2(4)  =     81.99
(Assumption: . nested in full)                        Prob > chi2 =    0.0000
```

Der χ^2-Wert ist sehr hoch und die Nullhypothese, dass kein Unterschied zum vollen Modell besteht kann nicht unterstütz werden, da der p-Wert nahe Null ist. Daher ist der Einfluss des Alters signifikant.

Ebenso lässt sich der Einfluss der weiteren Variablen bestimmen:

```
* Ohne Nationalität
 quietly poisson kinder alter alter2 alter3 alter4 teduc kirche if female==1

 * estat gof // zeigt Devianz des Modells an
 lrtest full

Likelihood-ratio test                                 LR chi2(1)  =      0.25
(Assumption: . nested in full)                        Prob > chi2 =    0.6161


 * Ohne Bildung
 quietly poisson kinder alter alter2 alter3 alter4 schweiz kirche if female==1

 * estat gof // zeigt Devianz des Modells an
 lrtest full

Likelihood-ratio test                                 LR chi2(1)  =      0.62
(Assumption: . nested in full)                        Prob > chi2 =    0.4321


 * Ohne Kirchenzugehörigkeit
quietly poisson kinder alter alter2 alter3 alter4 schweiz teduc ///
         if  female_kirche ==1 | female_kirche ==0
```

```
 * estat gof // zeigt Devianz des Modells an
 lrtest full

Likelihood-ratio test                                 LR chi2(1)  =     13.11
(Assumption: . nested in full)                        Prob > chi2 =    0.0003
```

Die Reduktion um die Nationalität und die Bildung führt zu keiner signifikanten Veränderung der Modellanpassung. Lediglich die Kirchenzugehörigkeit ist signifikant. Die Devianzanalyse ist notwendig, da Variablen des Modells zwar nicht signifikante Koeffizienten haben können, eine Reduktion des Modells um die Variable aber zu einer Verschlechterung der Modellgüte führen könnte – dies ist in diesem Beispiel nicht der Fall. Eine Reduktion der Variablen ist möglich, man sollte aber testen, ob das Modell ohne den beiden Variablen Nationalität und Bildung wirklich die gleiche Modellgüte hat.

7.3 Weiterführende Modelle

Neben den vorgestellten Modellen gibt es noch weiterführende Verfahren wie das "Zero-Inflated Poisson"-Modell. Das Modell berücksichtigt, das ein Grossteil der Personen gar keine Nennung hat (z. B. keine Kinder = 0 vs. Kinder =1) und dann genauer spezifiziert, was die Häufigkeit der Anzahl Kinder beeinflusst.

```
estimates restore poisson
(results poisson are active now)

predict mup

gen kinder_pois = exp(-mup)

gen kinder0 = kinder == 0 if kinder_pois!=.

sum kinder0 kinder_pois

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     kinder0 |       310    .3612903    .4811511          0          1
 kinder_pois |       310    .3085561    .2675668   .0127109   .9998567
```

In der Population haben 36 % der Frauen keine Kinder. Das Modell schätz diesen Anteil mit 31 % recht gut. Daher ist in unserem Fall kein zero-inflated Modell notwendig. Ist der Unterschied allerdings beträchtlich höher, ist solch ein Modell in Erwägung zu ziehen. (mehr dann unter http://data.princeton.edu/wws509/stata/overdispersion.html)

Weiter Modelle sind die "Zero-Truncated" und die "Hurdl"-Modelle.

Übersicht Stata-Befehle

Befehl	Option	Erklärung
`margeff`		Ausgabe average marginal effect (AME) aus dem ado-package `margeff`
`prchange`	`fromto`	Ausgabe der Wahrscheinlichkeitsveränderung der unabhängigen Variablen
`poisson`	`irr`, `nolog`	Berechnet das Poissonmodell. Die Option `irr` steht für incidence-rate ratio und liefert die antilogarithmierten Koeffizienten. Eine Umrechnung der logarithmierten Koeffizienten ist daher nicht nötig. `nolog` gibt keine Iterationsschritte an.
`glm`	`family(`*`verteilung`*`)`, `link(log)`, `eform`	Allgemeiner Befehl für generalisierte lineare Modelle wobei man die Verteilungsfamilie und die Transformationsfuktion angeben muss. Mit `eform` kann man die antilogarithmierten Koeffizienten zeigen.
`glm`	`scale(x2)`	Quasi-Poisson-Modell mit Pearsons'sχ^2-Wert als Dispersionsparameter.
`nbreg`		Negatives Binomialmodell

Aufgabe

Untersucht ein Zähldatenmodell mit der abhängigen Variable wie Anzahl Kurzstreckenflüge oder Anzahl Langstreckenflüge. Spezifiziert geeignete unabhängige Variablen (Geschlecht, Alter, Berufstätgikeit, Einkommen) und analysiert die geeignete Verteilungsfunktion (Poisson, Quasi-Poisson, Negativ Binomial). Diskutiert die Ergebnisse.

8 25.4. – Generalisierte Lineare Regressionsmodelle 3

8.1 Ordered Logit Regression

8.1.1 Modellidee

Bisher wurden bei der logistischen Regression Modelle besprochen, die zwei Ausprägungen (0 1) haben. Dann haben wir die Wahrscheinlichkeit dass ein Ereignis eintritt $P(y = 1)$ berechnet.

Bei abhängigen Variablen die mehr als zwei Ausprägungen haben kann man unterscheiden in Variablen die eine Ordnung aufweisen (ordinalskaliert) und Variablen ohne Ordnung (nominalskaliert).

- geordnet: Modelle für ordinalskalierte Variablen
- ungeordnet: Modelle für nominalskalierte Variablen

Der Inhalt heute befasst sich mit den ordinalskalierten Variablen. Beispielsweise Variablen wie

- Sozialer Status in "tief", "mittel", "hoch"
- Lebenszufriedenheit von 0 "sehr niedrig", ... , 10 "sehr hoch"
- "Wie oft verzichten Sie der Umwelt zuliebe auf das Autofahren?": 1 "nie", 2 "manchmal", 3 "oft", 4 "immer".
- Einstellungsfragen mit vier oder fünfstufigen Likertskalen, die wir in dem Beispiel gleich sehen werden.

Prinzipiell kann man diese Variablen als *quasi*-metrisch betrachten und eine lineare OLS-Regression rechnen. Folgende Annahme wird dann aber implizit getroffen:

- Die Distanzen zwischen den Kategorien sind konstant
- Das ist unproblematisch bei einer abhängigen Variabel mit vielen Ausprägungen, beispielsweise einem Index aus mehreren Zustimmungsvariablen

Generell kann man auf logistische Verfahren zurückgreifen, die kein intervallskaliertes Niveau voraussetzen. Anschliessend kann man dann testen, ob die Abstände gleich sind und eine OLS Regression angebrachter ist.

Man geht generell von einer latent intervallskalierten abhängigen Variable (Y) aus. Die Variable wird an Schwellenwerten in unterschiedliche Kategorien geteilt, so dass wir eine *diskrete* Variable Y^* mit J Kategorien erhalten.

Das Modelle mit der diskreten Variabel Y^* (mit J Kategorien) und m unabhängigen Variablen lautet wie folgt:

$$Y^* = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon$$

Berechnet wird die Chance, eine höhere Kategorie $(J+1)$ zu erreichen. Es wird damit angenommen, dass

- die Abstände zwischen den Kategorien unterschiedlich sind (Distanzen sind nicht konstant), daher auch keine lineare Regression,
- die unabhängigen Variablen immer den gleichen Einfluss haben (Parallel Regression Assumption)

8.1.2 Interpretation

Wir berechnen in einem Beispiel mit den MOSAiCH 2011 Daten, die ich bei der FORS heruntergeladen habe (http://fors-nesstar.unil.ch/webview/index.jsp). Als ordinale Variable dient Besorgnis über den allgemeinen Umweltzustand: "Wie besorgt sind Sie im Grossen und Ganzen in Bezug auf Umweltprobleme?" mit fünf Ausprägungen von 1 "überhaupt nicht besorgt" bis 5 "sehr besorgt":

```
tab E6
          Besorgnis in Bezug auf |
                  Umweltprobleme |      Freq.     Percent        Cum.
---------------------------------+-----------------------------------
1. - 1 - überhaupt nicht besorgt |         22        1.83        1.83
                        2. - 2 - |         79        6.56        8.39
                        3. - 3 - |        251       20.85       29.24
                        4. - 4 - |        567       47.09       76.33
           5. - 5 - Sehr besorgt |        285       23.67      100.00
---------------------------------+-----------------------------------
                           Total |      1,204      100.00
```

Gemessen wird der Zusammenhang zu drei unabhängigen Variablen:

- Geschlecht `female` (1==weiblich) (bivariater Dummy)
- Alter in Jahren: `age` (metrische Variable)
- Bildungsjahre: `educyrs` (metrische Variable)

Der Befehl für ordinale Logit Regression lautet `ologit`.

```
 ologit worry female age educyrs, nolog                 // orderd Logits

Ordered logistic regression                       Number of obs   =       1191
                                                  LR chi2(3)      =      20.05
                                                  Prob > chi2     =     0.0002
Log likelihood = -1505.1619                       Pseudo R2       =     0.0066

------------------------------------------------------------------------------
       worry |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .3238702   .1082794     2.99   0.003     .1116465    .5360939
         age |   .0091481   .0030731     2.98   0.003     .0031249    .0151713
     educyrs |   .0337467   .0154231     2.19   0.029     .0035179    .0639754
-------------+----------------------------------------------------------------
       /cut1 |  -3.116485   .3316932                     -3.766592   -2.466378
       /cut2 |  -1.443886   .2659013                     -1.965043   -.9227294
       /cut3 |   .0876139   .2558284                     -.4138005    .5890284
       /cut4 |   2.166601   .2647404                       1.64772    2.685483
------------------------------------------------------------------------------
```

Wir betrachten zunächst den Effekt der drei unabhängigen Variablen:

- `female`: Für Frauen erhöht sich dich ordered Log-odds in der nächst höheren Kategorie zu sein um 0.32.
- `age`: Wenn sich das Alter um ein Jahr erhöht, erhöht sich die orderd log-odds eine Kategorie mehr zu sorgen um 0.009 – ist also fast Null.
- `educyrs`: Mit jedem zusätzlichen Bildungsjahr erhöht sich die ordered log-odds sich eine Stufe mehr zu sorgen um 0.03.

Ebenso ist eine Interpretation mit den Odds möglich:

```
 ologit worry female age educyrs, or nolog

Ordered logistic regression                       Number of obs   =       1191
                                                  LR chi2(3)      =      20.05
                                                  Prob > chi2     =     0.0002
Log likelihood = -1505.1619                       Pseudo R2       =     0.0066

------------------------------------------------------------------------------
       worry | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.382468   .1496928     2.99   0.003     1.118117    1.709317
         age |    1.00919   .0031013     2.98   0.003      1.00313    1.015287
     educyrs |   1.034323   .0159525     2.19   0.029     1.003524    1.066066
-------------+----------------------------------------------------------------
```

```
       /cut1 |  -3.116485   .3316932                      -3.766592   -2.466378
       /cut2 |  -1.443886   .2659013                      -1.965043   -.9227294
       /cut3 |   .0876139   .2558284                      -.4138005    .5890284
       /cut4 |   2.166601   .2647404                        1.64772    2.685483
------------------------------------------------------------------------------
```

Dementsprechend erhöht sich die Chance zu einer höheren Gruppe zu gehören für Frauen um 1.38 oder 38 %. Der Alterseffekt ist nahe Null und die Chance zur höheren Gruppe zu gehören steigt mit jedem Bildungsjahr um 3 %.

8.1.3 Cut-Points

Sehen wir uns die Schwellenwerte (Cutpoints) `/cut` näher an. Das sind Hilfswerte, die ungleiche Abständen haben können, daher rechnen wir auch ein orderd Logit und kein lineares Regressionsmodell. Wir betrachten uns die Differenzen und testen ob sie gleich sind. Sind sie gleich, dann kann eine lineare Regression gerechnet werden.

```
*------------------------------------------------------------------------------
*  Test der Annahme gleicher Abstände
*------------------------------------------------------------------------------
test  (([cut2]_cons - [cut1]_cons) = ([cut3]_cons - [cut2]_cons)) ///
 (([cut2]_cons - [cut1]_cons) = ([cut4]_cons - [cut3]_cons))

( 1)  - [cut1]_cons + 2*[cut2]_cons - [cut3]_cons = 0
( 2)  - [cut1]_cons + [cut2]_cons + [cut3]_cons - [cut4]_cons = 0

          chi2(  2) =   19.81
        Prob > chi2 =    0.0000
```

Der signifikante P-Wert des Chi^2-Wertes von 19.81 besagt, dass die Abstände unterschiedlich sind. Damt ist ein ordered logit-Modell gerechtfertigt.

8.1.4 Parallel Regression Annahme

Abschliessend testen wir mit dem Brant Test die Annahme, dass die Effekte der unabhängigen Variable in jeder Kategorie den gleichen Effekt haben. Für jede höhere Kategorie wird eine logistische Regression gerechnet und der Effekt der unabhängigen Variablen verglichen. Der Brant-Test (`brant`) ist im ado-package `spost9` implementiert (Installation über `findit spost9`). Führen Sie den Brant-Test als postestimation Kommando nach dem `ologit`-Befehl aus.

```
findit spost9 // installieren Sie spost9_ado
net install spost9_ado, from(http://www.indiana.edu/~jslsoc/stata)

ologit worry female age educyrs
(Ergebnisse nicht gezeigt)

. brant, detail

Estimated coefficients from j-1 binary regressions

                y>1         y>2         y>3         y>4
 female  -.63397695   .05806197   .25041557   .45301259
    age   .01785964   .01503253   .00598858   .01113019
educyrs   .02563495   .01940758   .06361162   .00898151
  _cons   3.3250134   1.4514282  -.22910464  -2.0610669

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |     18.94    0.026     9
-------------+--------------------------
      female |      5.77    0.124     3
         age |      3.94    0.268     3
     educyrs |      7.57    0.056     3
----------------------------------------
A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

Der Gesamttest `All` deutet darauf hin, dass die parallel Regression Annahme verletzt ist. Allerdings sind nur die Bildungsjare `educyrs` mit p-Wert von 0.056 fast nicht signifikant und damit unterschiedlich.

Wenn der Test die Annahme der parallelen Regression ablehnt, sollte man eher multinomiale Logit rechnen, das wir hier aber nicht mehr besprochen.

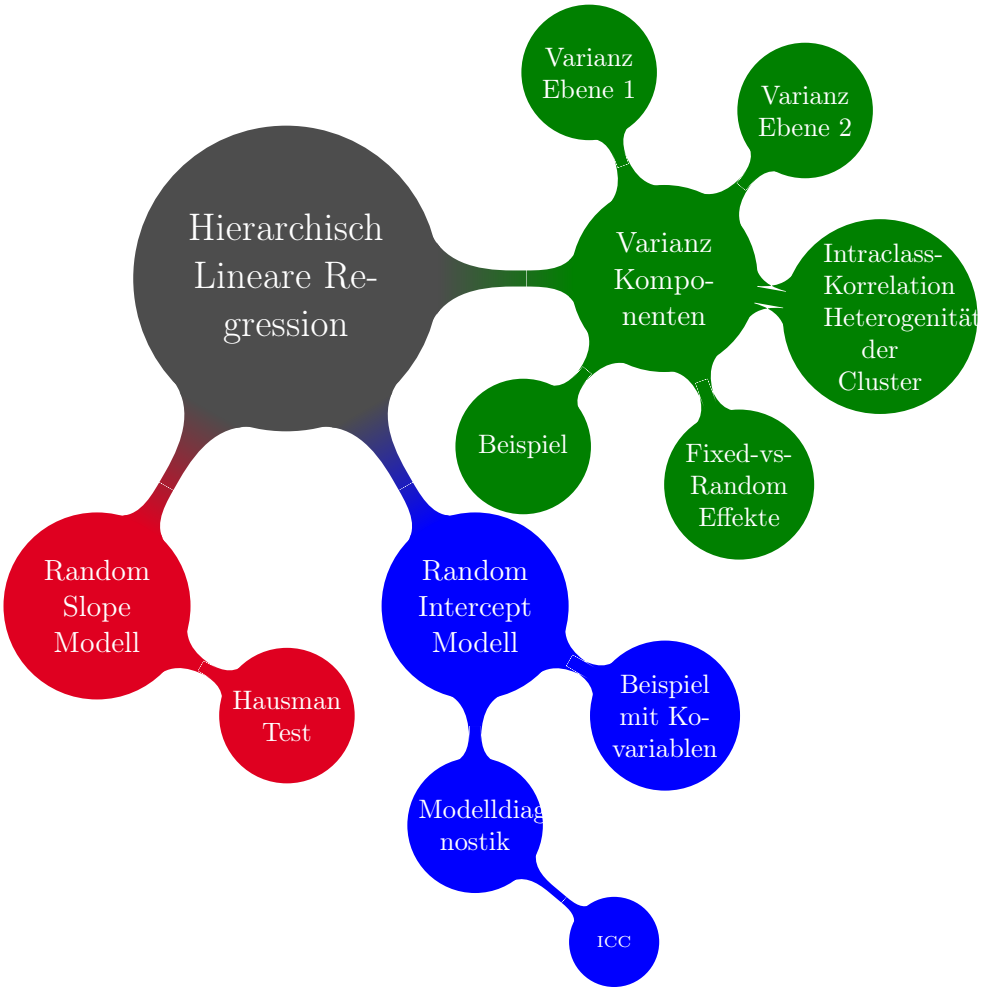
8.1.5 weiterführende Literatur

Eine vollständige und sehr umfangreiche Erklärung zur Theorie mit Beispielen findet ihr in (Long und Freese 2006: Kapitel 5). Eine knappe Einführung findet Ihr auch unter http://www.ats.ucla.edu/stat/stata/output/stata_ologit_output.htm. Für die Software R bieten (Fox und Weisberg 2010: Kapitel 5.9, Seite 269ff) eine Erklärung für das oben besprochene *proportional-odds logistic-regression model*.

Übersicht Stata-Befehle

Befehl	**Option**	**Erklärung**
`ologit`	`or`	ordered logit mit `or` Option für Odds statt log odds
`brant`	`detail`	Brant Test der Parallel Regression Annahme

9 2.5. – Hierarchisch Lineare Regression 1



9.1 Einführung

Die hierarchische lineare Regression wird auch als Mehrebeneregression oder Multi-level-Regression bezeichnet, man findet auch die Bezeichnung Mixes-Models. Es handelt sich also um *Mehrebenemodelle*. Analysiert werden also nicht nur Untersuchungseinheiten wie Individuen, sondern der Kontext ist nun von Interesse, in den das Individuum eingebettet ist. Neben den erklärenden Variablen für das individuelle Verhalten werden auch erklärende Kontextvariablen in die Analyse einbezogen. Untersucht werden also (mindestens) zwei Ebene. Zum einen die individuelle Ebene (Ebene 1 = Mikroebene) und die Kontextebene (Ebene 2 = Makroebene). Beispielsweise sind Schüler die Untersuchungseinheiten der Eben 1 und Schulen die Untersuchungseinheit der Ebene 2. Ebenso können Haushalte befragt werden, die wiederum in Kantone eingeteilt sind oder Personen befragt werden, die wiederum Ländern zugeordnet werden können. Eine Person kann auch zwei mal befragt werden. Dann ist die Person das Kontextmerkmal und die

Befragungeszeitpunkte das Ebene 1 Merkmal.

Hinsichtlich der Begrifflichkeit liest man auch, dass die Ebene 1 Variablen *genestet* sind: Kinder sind in Familien genestet. Anstelle von Ebene 2 Variablen spricht man auch von *Cluster*. Der Begriff Cluster wird vor allem bei Längsschnittanalysen verwendet. Ein Beobachtungscluster liegt dann vor, wenn für eine Person zu unterschiedlichen Zeitpunkten mehrere Beobachtungen gemacht wurden.

Die verwendete Literatur bezieht sich auf Langer (2010) und auf (Rabe-Hesketh und Skrondal 2008: Kapitel 2 und 3). Weiterführende Literatur ist Snijders und Bosker (1999).

9.2 Varianz-Komponenten Modell

Bei geclusterten Daten bietet es sich an, die Korrelation innerhalb der Cluster (within-cluster-Korrelation) zu untersuchen. Grundsätzlich sollten Ähnlichkeiten innerhalb der Cluster festzustellen sein, da die Clustermerkmale einheitlich sind. So werden Kinder in einer Familie ähnlich erzogen, Bezüglich politischer Einstellung und Bildungsniveau sollten auch Ehepaare ähnliche Muster aufweisen.

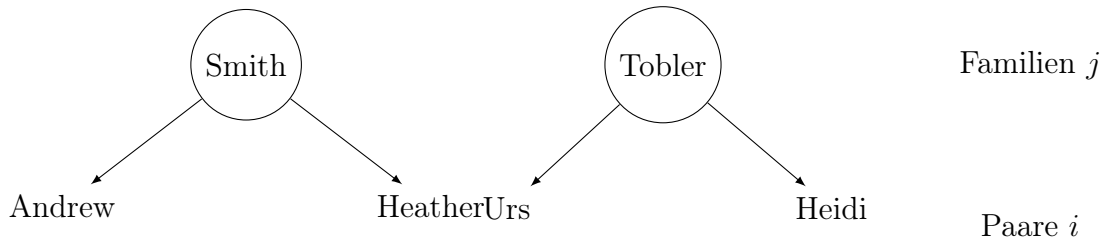


Abbildung 5: Beispiel der Ebenenstruktur

Eine Erhebung in der Schweiz könnte in mehr als zwei Ebenen aufgeteilt werden, da Haushalte in Kantonen genested sind, die wiederum in deutsche, französische und italienische Schweiz unterteilt werden könnten.

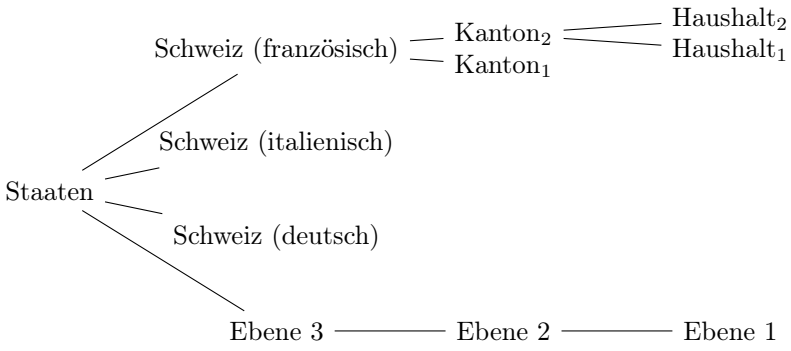


Abbildung 6: Ebenen in der Schweiz

9.2.1 Modellspezifikation

Im Mehrebenemodell spricht man von Untersuchungseinheiten y_{ij}. Die Ebene 1 wird mit i, die Ebene 2 mit j bezeichnet:

- i (Ebene 1): Eine Person y_{1j} ist die 1 Person im j-te Land.
- j (Ebene 2): Ein Land y_{i1}, in dem $i = 1 \dots n$ Personen befragt wurden.

Ein Regressionsmodell ohne unabhängigen Variablen sieht wie folgt aus:

$$y_{ij} = \beta + \xi_{ij} \tag{25}$$

In dem Modell gibt es allerdings die unrealistische Annahme, dass der Fehlerterm ξ ("xi") sowohl unabhängig von den Untersuchungseinheiten i, als auch von den Clustern j ist, in denen die Untersuchungseinheiten der Ebene 1 genestet sind. Anzunehmen ist, dass es innerhalb der Cluster grössere Ähnlichkeiten und daher weniger Varianz geben sollte als zwischen den Clustern oder Gruppen. Werden Personen z. B. in kurzem Abstand zwei mal befragt, so sollte die zweite Antwort nahe an der ersten Antwort liegen, weil die Antworten von der selben Person stammen und Personen generell ihre Meinung nicht beliebig wechseln. Zusammengefasst spricht man von einer Abhängigkeit der Subjekte innerhalb der Gruppen und eine Heterogenität zwischen den Gruppen.

Das Problem der geringen Varianz auf Ebene 1 kann dadurch gelöst werden, dass man im Regressionsmodell (25) die Fehlerkomponente ξ_{ij} in eine Ebene 2 (j) und einen Ebene 1 (i) Fehlerterm splittet.

$$y_{ij} = \beta + \underbrace{\zeta_j + \varepsilon_i}_{\xi_{ij}} \tag{26}$$

Die Gleichung 26 ist in Abbildung 7 graphisch dargestellt.

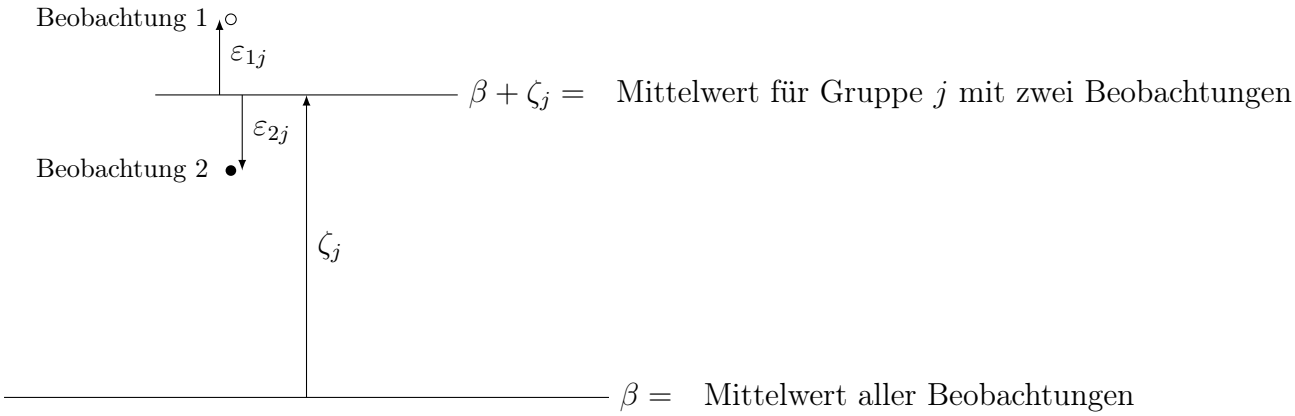


Abbildung 7: Streuung Ebene 2 um Mittelwert

Ausgangspunkt ist ein hypothetischer Mittelwert β in einer Population (z. B. Einstellung der Schweizer zu Ausländern). Jede Person wird innerhalb eines Jahres zwei Mal gefragt. Der Mittelwert dieser beiden Befragungen spiegelt die grundsätzliche Tendenz der Person gegenüber Ausländern wieder: $\beta + \zeta_j$ (ζ gesprochen "zeta"). Vermutlich sind die Messungen nicht identisch, so dass es bei einer Person Abweichungen vom Personen-Mittelwert geben wird ε_{1j} und ε_{2j}.

- ζ_j: Wird auch "random intercept" genannt und ist unabhängig zwischen den Ebene 2 Merkmalen. Der Term ist über alle Gruppen normalverteilt mit Mittelwert 0 und Varianz ψ: $\zeta_j \sim N(0, \psi)$. Die Varianz ψ ("psi") ist die Varianz zwischen den Gruppen.
- ε_{ij}: Dieser Term sind die Residuen auf Ebene 1 (innerhalb der Gruppe) und damit die Abweichungen y_{ij} vom j-ten Gruppenmittelwert. Das Residuum ist normalverteilt mit $\varepsilon_{ij} \sim N(0, \theta)$ und ist unabhängig von der Ebene 1 und Ebene 2. θ ("theta") wird die Varianz zwischen den Beobachtungen (Ebene 1) und innerhalb der Gruppen bzw. der Cluster (Ebene 2) genannt.

Abbildung 8 veranschaulicht nochmals die Verteilung der Fehler auf Ebene 1 und Ebene 2.

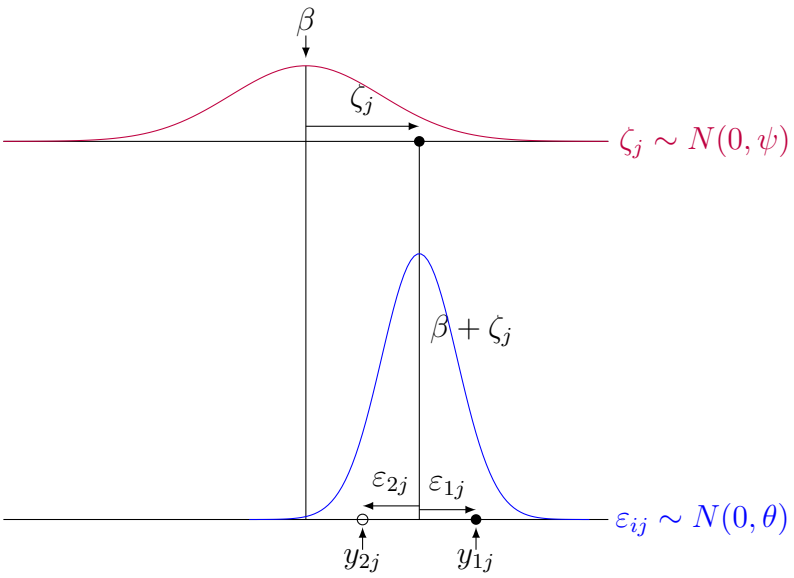


Abbildung 8: Verteilung der Fehler auf Ebene 1 und Ebene 2

Die Ebene 1 Beobachtungen sind unabhängig von einander *gegeben* die Ebene 2 Gruppe. Mathematisch heisst dies, dass die Korrelation zweier Beobachtungen 0 ist:

$$Cor(y_{ij}, y_{i'j}|\zeta_j) = 0 \tag{27}$$

Zusammengefasst lässt sich sagen, dass jede Ebene 1 Beobachtung y_{ij} vom Gesamtmittelwert β durch einen Fehler ξ_{ij} abweicht, der durch zwei Fehlerkomponenten ζ_j und ε_{ij} spezifiziert wird. ζ_j wird von allen Ebene 1 Beobachtungen der selben Gruppe (j) geteilt, während ε_{ij} für jede Ebene 1 Beobachtung (i in Ebene 2) einzigartig ist.

Die totale Varianz des Modells ist die Summer der Varianzkomponenten aus der Varianz zwischen den Gruppen und innerhalb der Gruppe:

$$Var(y_{ij}) = Var(\beta) + Var(\zeta_j + \varepsilon_{ij}) = \underbrace{\psi}_{\text{zwischen Gruppe}} + \underbrace{\theta}_{\text{innerhalb Gruppe}} \tag{28}$$

Der Anteil der Varianz, der durch die Ebene 2 erklärt wird, lässt sich durch

$$\rho = \frac{Var(\zeta_j)}{Var(\varepsilon_{ij})} = \frac{\psi}{\psi + \theta} \tag{29}$$

bestimmen. ρ (gesprochen "Rho") ist also ein Mass für die Heterogenität zwischen den Clustern – die "zwischen den Clustern Varianz". ρ drück gleichzeitig die Korrelation

innerhalb des Clusters aus und wird daher auch "intraclass correlation (ICC)" genannt. Ist die Varianz zwischen den Clustern gering und die Varianz innerhalb der Cluster gross, geht ρ gegen 0. Umgekehrt steigt ρ mit der Heterogenität der Gruppen und der Homogenität der genesteten Beobachtungen.

Fixed-Effects

Der Vollständigkeit halbe noch die Gleichung der Intraclass-Korrelation:

$$Cor(y_{ij}, y_{i'j}) = \frac{Cov(y_{ij}, y_{i'j})}{\sqrt{Var(y_{ij})}\sqrt{Var(y_{i'j})}} = \frac{\psi}{\sqrt{\psi+\theta}\sqrt{\psi+\theta}} = \frac{\psi}{\psi+\theta} = \rho \tag{30}$$

9.2.2 Fixed vs. Random Effekte

In der Mehrebenemodellierung wird neben den eben besprochenene Random-Effects noch zwischen "Fixed-Effects " unterschieden.

Das Random-Effects Modell zeigt die Streuung vom Mittelwert, die unabhängig von den anderen Gruppen ist und unabhängig von den Ebene 1 Beobachtungen.

$$y_{ij} = \beta + \zeta_j + \varepsilon_{ij}, \qquad \varepsilon_{ij}|\zeta_j \sim N(0,\theta) \quad \zeta_j \sim N(0,\psi) \tag{31}$$

Im Gegensatz dazu wird beim Fixed-Effects-Modell die Heterogenität in den Gruppen durch einen Parameter berechnet, der keiner Verteilung folgt. α ("alpha") ist ein unbekannter cluster-spezifischer Parameter.

$$y_{ij} = \beta + \alpha_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0,\theta) \quad \sum_{j=1}^{J} \alpha_j = 0 \tag{32}$$

In beiden Modellen sind die Ebene 1 Beobachtungen unabhängig voneinander. Beide Parameter ζ_j und α_j modellieren die unbeobachtete Heterogenität. Insgesamt sind sich die beiden Ansätze sehr ähnlich und es ist eine Ermessensfrage, welches Modell verwendet wird.

- **Fixed-Effects**: Prinzipiell ist es immer eine Frage, welche Aussagen man verallgemeinern möchte. Ist die *Population* unabhängig von der Gruppe von Interesse, so sollte ein Fixed-Effects-Modell gerechnet werden. Vor allem dann, wenn man keine Zufallsauswahl der Gruppen hat. Kontrolliert wird dann in α der Effekt der Gruppen und allgemein Aussagen über die Population auf Ebene 1 sind möglich. Die Fallzahl innerhalb der Gruppen sollte hoch sein.
- **Random-Effects**: Wenn sowohl die Auswahl der Gruppen als auch der Population in den Gruppen als zufällig betrachtet werden kann, bietet sich ein Random-Effects-Modell an. Von Interesse ist dann die Streuung ψ um den geschätzten Mittelwert $\hat{\beta}$

– Die Fallzahl der Ebene 2 Einheiten sollte mindestens 10 - 20 betragen. Sie sollte höher sein, wenn man die Varianz der Gruppen ψ korrekt spezifizieren möchte.

– Die Fallzahl innerhalb der Gruppen kann relativ klein sein (≥ 2).

9.2.3 Beispiel Varianz-Komponenten Modell

Das Beispiel stammt aus dem Lehrbuch von (Rabe-Hesketh und Skrondal 2008: Kapitel 2, S. 52). Bei den Daten handelt es sich um die Messung einer peak-espiratory-flow-rate (PEFR), einem Mass für die Geschwindigkeit des Ausatmens. Zwei Apparate wurden getestet, wobei 17 Personen jeweils zwei mal in jeden Apparat ausatmen mussten.

- `wp1`: Apparat 1 (Wright peak flow meter), 1. Versuch
- `wp2`: Apparat 1, 2. Versuch
- `wm1`: Apparat 2 (mini Wright peak flow meter), 1. Versuch
- `wm2`: Apparat 2, 2. Versuch

Die hierarchische Struktur der Daten ergibt sich daraus, dass hier Personen die Cluster bilden und die Versuche genestet sind, also jeder Person zugeordnet werden können.

Die Daten können direkt aus dem Internet heruntergeladen werden.

```
use http://www.stata-press.com/data/mlmus2/pefr, clear
save ../data/example_hlm, replace
* Speichere und Lade Daten wieder
use ../data/example_hlm, clear
describe
```

Untersucht werden nun die beiden Versuche mit dem neuen Apparat, dem mini Wright peak flow meter. In einem ersten Schritt wird der Mittelwert der beiden Beobachtungen gebildet. Anschliessend werden die Beobachtungen mit Mittelwert in einem Scatterplot dargestellt (Abbildung 9).

```
* Bilde Mittelwert für Apparat 2: miniWpfm:

generate wm_mean = (wm1+wm2)/2
summarize wm_mean // Mittelwert bei 453.9

* Scatterplot der zwei Testergebnisse pro Person

twoway  (scatter wm1 id, msymbol(circle)) ///
         (scatter wm2 id, msymbol(circle_hollow)), ///
xtitle(Personen ID) ///  Beschriftung der X-Achse
```

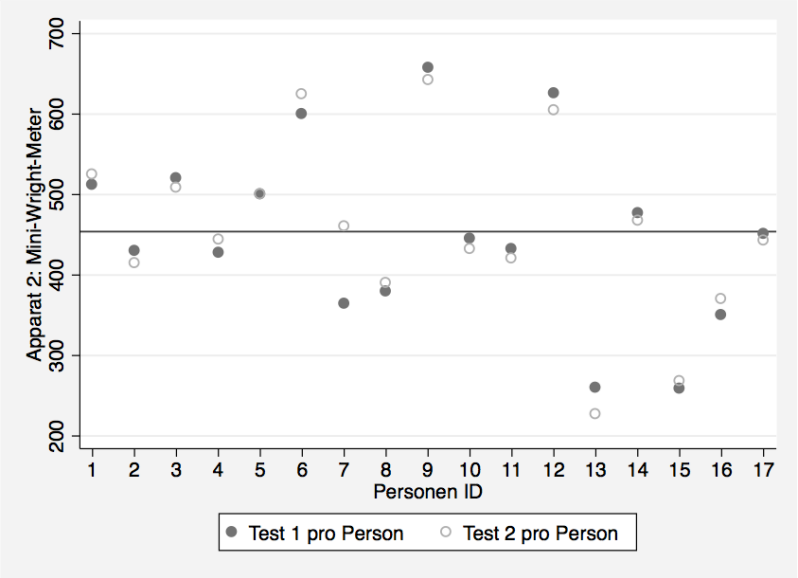


Abbildung 9: Scatterplot der genesteted Beobachtungen

```
xlabel(1/17) /// Nummerierung der X-Achse
ytitle(Apparat 2: Mini-Wright-Meter) /// Beschriftung der y-Achse
legend(order(1 "Test 1 pro Person" 2 "Test 2 pro Person")) /// Legenden Text
yline(454) // Aus summarize Tabelle kopiert und gerundet
```

In `Stata` können Mehrebenenmodelle mit drei Befehlen berechnet werden.

- `xtreg` mit der Option `mle`,
- `xtmixed` mit `mle` Option,
- `gllamm`

`xtreg` und `xtmixed` sind aus Anwendersicht zu empfehlen, da sie schnell gerechnet werden (sehr effizient sind). Die Berechnungen mit `gllamm` dauern bedeutend länger.

Generell gilt es die Datenstruktur zu beachten. Als Gruppierungsvariable dient in unserem Fall die Variable `id`. Sie kommt allerdings nur ein mal im Datensatz vor, da es sich um ein *wide*-Format handelt. Allerdings stehen die Messungen für jeden Apparat in einer gesonderten Spalte.

```
list in 1/7 // Ausgabe der ersten sieben Beobachtungen
```

```
     +-------------------------------------+
     | id   wp1   wp2   wm1   wm2   wm_mean |
     |-------------------------------------|
  1. |  1   494   490   512   525     518.5 |
```

```
  2. |  2   395   397   430   415     422.5 |
  3. |  3   516   512   520   508       514 |
  4. |  4   434   401   428   444       436 |
  5. |  5   476   470   500   500       500 |
     |--------------------------------------|
  6. |  6   557   611   600   625     612.5 |
  7. |  7   413   415   364   460       412 |
     +--------------------------------------+
```

Durch die Transformation der Daten wird erreicht, dass es pro Apparat nur noch eine Variable gibt und zusätzlich eine Variable (`test`) gebildet wird, die dokumentiert, ob es sich um den ersten oder den zweiten Versuch gehandelt hat.

```
reshape long wp wm, i(id) j(test)
(note: j = 1 2)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       17   ->      34
Number of variables                   6   ->       5
j variable (2 values)                     ->   test
xij variables:
                                 wp1 wp2   ->   wp
                                 wm1 wm2   ->   wm
-----------------------------------------------------------------------------
```

Man darf sich durch die Notation nicht verwirren lassen. In `Stata` werden die Cluster mit *i()* bezeichnet und die Einheiten innerhalb der Cluster mit *j()* bezeichnet. In der formalen Notation, die im Skript verwendet wird sind ist es umgekehrt: i sind die Ebene 1 Einheiten und j die Cluster (Ebene 2 Einheiten).

```
 list in 1/8

     +---------------------------------+
     | id   test    wp    wm   wm_mean |
     |---------------------------------|
  1. |  1      1   494   512     518.5 |
  2. |  1      2   490   525     518.5 |
  3. |  2      1   395   430     422.5 |
  4. |  2      2   397   415     422.5 |
  5. |  3      1   516   520       514 |
     |---------------------------------|
  6. |  3      2   512   508       514 |
  7. |  4      1   434   428       436 |
  8. |  4      2   401   444       436 |
     +---------------------------------+
```

xtreg Der Befehl `xtreg` wird wie der `reg` Befehl der multiplen linearen Regression verwendet. Im Mehrebenmodell wird noch die Schätzmethode (`mle`), das für die Maximum-Likelihood-Methode steht und die Gruppierungsvariable (`i()`) spezifiziert. In diesem Fall wird nur der Intercept (bzw. die Konstante) berechnet, so dass als abhängige Variable die Messungen des neuen Apparates `wm` steht und als Gruppierungsvariable die ID der Versuchspersonen `i(id)`.

```
xtreg wm, i(id) mle
```

```
Random-effects ML regression                    Number of obs      =        34
Group variable: id                              Number of groups   =        17

Random effects u_i ~ Gaussian                   Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2

                                                Wald chi2(0)       =      0.00
Log likelihood  = -184.57839                    Prob > chi2        =         .

------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   26.18616    17.33   0.000     402.5878    505.2357
-------------+----------------------------------------------------------------
    /sigma_u |   107.0464   18.67858                      76.0406    150.6949
    /sigma_e |   19.91083   3.414659                      14.2269     27.8656
         rho |   .9665602   .0159494                     .9210943    .9878545
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)=   46.27 Prob>=chibar2 = 0.000
```

Es gibt 17 Gruppen mit 34 Beobachtungen, wobei jede Gruppe zwei Beobachtungen hat. u_i steht für Random Intercept, das wir mit ζ_j bezeichnet haben. Die Konstante (`_cons`) zeigt den Mittelwert β von 453.9 an. Von Interesse ist noch die Streuung innerhalb der Gruppen (θ) und zwischen den Gruppen (ψ). In `Stata` wird nicht die Varianz, sondern die Wurzel der Varianz, also die Standardabweichung ausgegeben. `/sigma_u` ist die Standardabweichung zwischen den Gruppen (Ebene 2) $\sqrt{\psi}$, `/sigma_e` ist die Standardabweichung innerhalb der Gruppen (Ebene 1) $\sqrt{\theta}$. `rho` gibt den (geschätzte) Anteil der Varianz an, der durch Gruppenunterschiede zustande kommt. In unserem Beispiel gibt es auch die Reliabiliät des Apparates als Messinstrument an. Da die Varianz der beiden Messungen pro Person gering ist, ist das Instrument reliabel. Allerdings sind die Unterschiede zwischen den Gruppen (in diesem Fall Personen) sehr hoch, was der Intraclass Korrelationskoeffizient (ICC) von 0.97 zeigt.

$$\widehat{\rho} = \frac{\widehat{\psi}}{\widehat{\psi} + \widehat{\theta}} = \frac{107.05^2}{107.05^2 + 19.91^2} = 0.97$$

```
* Berechnung ICC (Rho) per Hand aus gespeicherten Werten
scalar rho = e(sigma_u)^2/(e(sigma_u)^2+e(sigma_e)^2)
display rho
.96656022
```

xtmixed Der Befehl `xtmixed` kommt zu den selben Ergebnissen wie `xtreg`. Nach dem die abhängige Variable `wm` genannt wird, zeigt `|| id:`, dass ein Random Intercept eingefügt werden soll, dessen Gruppierungsvariable `id` lautet.

```
xtmixed wm || id:, mle

Mixed-effects ML regression                     Number of obs      =        34
Group variable: id                              Number of groups   =        17

                                                Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2


                                                Wald chi2(0)       =         .
Log likelihood = -184.57839                     Prob > chi2        =         .

------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   26.18617    17.33   0.000     402.5878    505.2357
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   107.0464   18.67858      76.04062     150.695
-----------------------------+------------------------------------------------
                sd(Residual) |   19.91083   3.414678      14.22687    27.86564
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    46.27 Prob >= chibar2 = 0.0000
```

Mit der Option `variance` kann man sich für die Ebene 1 (`sd(Residual)` $= \sqrt{\theta}$) und Ebene 2 (`sd_cons`$= \sqrt{\psi}$) Residuen die Varianz statt der Standardverteilung angeben lassen.

gllamm Der Befehl gllamm wird hier nicht näher erläutert. Mit `which gllamm` kann man testen, ob das Package installiert ist (ggf. mit `ssc install gllamm, replace`)

9.2.4 Hypothesentest der Varianz zwischen den Gruppen

In Mehrebenenmodellen interessiert, ob die Varianz zwischen den Gruppen (between-cluster variance) von Null unterschiedlich ist. Daher wird die Hypothese

$$H_0 : \psi = 0 \quad \text{vs.} \quad H_1 : \psi > 0$$

getestet, dass es keinen Random Intercept im Modell gibt. Ist die Nullhypothese wahr, dann ist $\zeta_j = 0$. Der Fehlerterm (vgl. 26 auf S. 137) hat keine Bedeutung für das Modell und OLS-Regression kann verwendet werden. Getestet wird ein Likelihood-ratio Test mit dem Modell mit Random Intercept (l_1) gegen ein Modell ohne Random-Intercept (l_0).

$$L = 2(l_1 - l_0)$$

Im Stata-Output wird der χ^2-Testwert am Ende unter `chibar2(01)` und der p-Wert angegeben. In diesem Fall beträgt der Testwert 46.27 bei einem höchst signifikanten p-Wert < 0.0000. Die Nullhypothese kann also abgelehnt werden, ein OLS Regressionsmodell kann nicht gerechnet werden.

Alternativ kann man eine Fixed-Effects-Regression durchführen und die Nullhypothese testen, dass die Fixed-Effects $\alpha_j = 0$ Null sind. Hierfür wird `xtreg` mit der `fe` Option (für fixed effects) ausgeführt.

```
 xtreg wm, i(id) fe

Fixed-effects (within) regression               Number of obs      =        34
Group variable: id                              Number of groups   =        17

R-sq:  within  = 0.0000                         Obs per group: min =         2
       between =      .                                       avg =       2.0
       overall =      .                                       max =         2

                                                F(0,17)            =      0.00
corr(u_i, Xb)  =      .                         Prob > F           =         .
```

```
------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   3.414679   132.93   0.000     446.7074    461.1161
-------------+----------------------------------------------------------------
     sigma_u |  111.29118
     sigma_e |  19.910831
         rho |  .96898482   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(16, 17) =    62.48                 Prob > F = 0.0000
```

Der F-Test Wert (62.48) ist sehr hoch und der p-Wert höchst signifikant. Daher kann auch hier die Nullhypothese abgelehnt werden, d.h. die Fixed-Effects unterscheiden sich.

9.3 Random Intercept Modell mit Kovariablen

In Kapitel 9.2 wurde die Erweiterung des Regressionsmodells hin zu einem Mehrebenenmodell besprochen. Charakteristisch ist die Modellierung der Ebene 2 Residuen durch einen Random Intercept ζ_j. Dieses Modell wird nun um erklärende Variablen x erweitert.

In dem Beispiel aus (Rabe-Hesketh und Skrondal 2008: S. 92ff) wird untersucht, ob Rauchen der Mutter während der Schwangerschaft einen Effekt auf das Geburtsgewicht der Kinder hat. Mütter sind die Clusters und Kinder – genestet zu Müttern – die Ebene 1.

Entsprechend gibt es im Datensatz Variablen wie das Geburtsgewicht (`birwt`), die den Kindern zugeordnet werden können und daher Ebene 1 Variablen sind. Ebenso kann sich das Verhalten verändern, so dass eine Mutter zwischen zwei Kindern mit dem Rauchen aufgehört hat und das Kind ohne eine rauchende Mutter auf die Welt kam – also auch eine Ebene 1 Variable. Die Herkunft der Mutter oder die Hautfarbe der Mutter (`black`) bleiben konstant und sind daher Ebene 2 Variablen.

Die Varianz der Variablen innerhalb der Cluster und zwischen den Clustern kann man sich mit dem Befehl `xtsum` ausgeben lassen, wobei man wiederum mit `i()` die Gruppierungsvariable (`momid`) angeben muss.

```
*use http://www.stata-press.com/data/mlmus2/smoking, clear // Internet
use ../data/example_hlm_smoking, clear // Lade von Festplatte

 xtsum birwt smoke black, i(momid)

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
birwt    overall |  3469.931   527.1394        284       5642 |     N =    8604
         between |             451.1943       1361     5183.5 |     n =    3978
```

```
         within  |            276.7966    1528.431    5411.431 | T-bar =  2.1629
                 |                                             |
smoke    overall |  .1399349   .3469397           0           1 |     N =    8604
         between |             .3216459           0           1 |     n =    3978
         within  |             .1368006  -.5267318    .8066016 | T-bar =  2.1629
                 |                                             |
black    overall |  .0717108   .2580235           0           1 |     N =    8604
         between |              .257512           0           1 |     n =    3978
         within  |                    0   .0717108   .0717108 | T-bar =  2.1629
```

Für drei Variablen wird die Variation betrachtete. Insgesamt gibt es Daten zu 8604 Kinder (`N`) von 3978 Müttern (`n`). Im Schnitt wurden 2.16 Kinder pro Monat geboren (`T-bar`). Für die Variable `black` lässt sich wie erwartet keine Variation innerhalb der Gruppe feststellen, da die Hautfarbe der Mutter konstant ist. `birwt` und `smoke` variieren mehr zwischen den Gruppen als in den Gruppen, d. h. die Gruppenzusammensetzung ist relativ homogen.

Das Random Intercept Modell ist die Erweiterung des Modells ohne Kovariablen (26) auf Seite 137 um p Kovariablen.

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1ij} ... + \beta_p x_{pij} + \underbrace{\zeta_j + \varepsilon_i}_{\xi_{ij}} \\ &= (\beta_0 + \zeta_j) + \beta_1 x_{1ij} ... + \beta_p x_{pij} + \varepsilon_i \end{aligned} \tag{33}$$

Ein Beispielmodell berechnet mit `xtreg, mle` und mehreren erklärenden Variablen sieht wie folgt aus:

```
xtreg birwt smoke black, i(momid) mle

Random-effects ML regression                    Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

Random effects u_i ~ Gaussian                   Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3

                                                LR chi2(2)         =    361.17
Log likelihood  =   -65294.9                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
       smoke |  -271.7308   17.38706   -15.63   0.000    -305.8088   -237.6527
       black |  -288.0684   26.41911   -10.90   0.000    -339.8489   -236.2879
       _cons |   3527.148   7.451026   473.38   0.000     3512.544    3541.752
-------------+----------------------------------------------------------------
    /sigma_u |   341.5994    6.41113                      329.2621    354.3989
    /sigma_e |   378.5685   3.947582                      370.9099    386.3852
         rho |   .4488009   .0120053                      .4253745    .4724073
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)= 1079.83 Prob>=chibar2 = 0.000
```

```
Die Interpretation der Koeffizienten ist analog zur OLS-Regression.
```

- Wenn die Mutter raucht (`smoke`), dann reduziert sich das Geburtsgewicht um 272 Gramm.

- Wenn die Mutter schwarz ist (`black`), dann reduziert sich das Geburtsgewicht um 288 Gramm

- Die Konstante sagt, dass das geschätzte Geburtsgewicht für ein Baby einer weissen Mutter die nicht raucht bei 3527 Gramm liegt.

Übersicht Stata-Befehle

Befehl	Option	Erklärung
`reshape long`		Transformation eines Datensatzes von einem *wide* in ein *long* Format, so dass er für Mehrebenmodelle verwendet werden kann.
`xtreg`	`i()`, `mle`, `fe`	Angegeben wird die abhängige Variable und weitere unabhängige Variablen. `mle` steht für Maximum Likelihood Estimation, `fe` für fixed effects Schätzung. `i(`*`var`*`)` nennt die Clustervariable
`xtmixed` *var* `||` *groupvar*`:`	`mle variance`	Die Gruppierungsvariable wird nach `||` eingefügt und mit einem `:` abgeschlossen. `mle` steht für die Maximum Likelihood Schätzmethode. Mit der Option `variance` werden die Varianzen der Residuen auf Ebene 1 und 2 anstelle der Standardfehler angegeben.
`xtsum`	`i(`*`groupvar`*`)`	Zusammenfassung der Variabilität der Variablen auf Ebene 1 und Ebene 2.

Aufgabe optional

Verwendet die Eurpean Values Study (EVS) 2008 untersucht die Länder und such nach geeigneten Kovariablen auf Länderebene für das Jahr **2008**. Benutzt Datenbanken der Weltbank (`http://data.worldbank.org/`), der UN (`http://data.un.org/`) oder der Europäischen Union (`http://epp.eurostat.ec.europa.eu/`), um an Makrodaten aus dem Jahr 2008 zu gelangen. Als abhängige Variable bietet sich die Variable "Lebenszufriedenheit" (`v66`) an.

Mögliche Kennzahlen sind:

- BIP (Bruttoinlandsprodukt) in einheitlichen Dollar
- Masse der Ungleichheit (Gini-Koeffizient)
- Urbanisierungsrate
- Alphabetisierungsrate
- Anteil der Dienstleistungen (3. Sektor) am BIP
- Staatsverschuldung
- durchschnittliche Lebenserwartung

- ...

Erstellt eine Excel-Liste mit einer Spalte mit Ländernamen und weiteren Spalten, in denen die Variablen mit den Ausprägungen stehen.

10 16.5. – Hierarchisch Lineare Regression 2

10.1 Einführung

Im letzten Kapitel (9) wurde die Varianz einer abhängigen Variabel in eine Gruppenkomponente und eine individuelle Komponente zerlegt. Dieser Vorgang wurde als ein Varianz-Komponenten-Modell bezeichnet (Kapitel 9.2). Anschliessend wurden mehrere unabhängige Variablen zur Erklärung einer abhängigen Variable (Gerburtsgewicht von Kindern) in das Mehrebenenmodell aufgenommen und nach Gruppenvariablen und individuellen Variablen unterschieden (Kapitel 9.3). Das Modell mit mehreren Kovariablen wird Random-Intercept-Modell genannt. Besprochen wurde auch der Intraclass-Korrelationskoeffizient (ICC), der den Anteil der Varianz, der durch Gruppenunterschiede (Ebene 2) erklärt werden kann, beschreibt. In einer multiplen Regression hat das Random-Intercept Modell die Eigenschaft, dass die Effekte der unabhängigen Variablen für alle Ebene 1 Einheiten gleich sind. Die Konstante bzw. der Intercept variiert allerdings je Gruppe (vgl. Gleichung 34). Im Folgenden soll ein Modell besprochen werden, in dem sowohl die Intercepts als auch die Steigunskoeffizienten je Gruppe variieren können. Dieses Modell wird Radom-Koeffizienten-Modell genannt. Zuerst wird allerdings nochmals auf das Random-Intercept-Modell näher eingegangen.

10.2 Random Intercept Modell mit Kovariablen (Fortsetzung)

Für drei Variablen wird die Variation betrachtete. Insgesamt gibt es Daten zu 8604 Kinder (`N`) von 3978 Müttern (`n`). Im Schnitt wurden 2.16 Kinder pro Monat geboren (`T-bar`). Für die Variable `black` lässt sich, wie erwartet, keine Variation innerhalb der Gruppe feststellen, da die Hautfarbe der Mutter konstant ist. `birwt` und `smoke` variieren mehr zwischen den Gruppen als in den Gruppen, d. h. die Gruppenzusammensetzung ist relativ homogen.

Das Random Intercept Modell ist die Erweiterung des Modells ohne Kovariablen (26) auf Seite 137 um p Kovariablen.

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1ij} ... + \beta_p x_{pij} + \underbrace{\zeta_j + \varepsilon_i}_{\xi_{ij}} \\ &= (\beta_0 + \zeta_j) + \beta_1 x_{1ij} ... + \beta_p x_{pij} + \varepsilon_i \end{aligned} \tag{34}$$

Ein Beispielmodell berechnet mit `xtreg, mle` und mehreren erklärenden Variablen sieht wie folgt aus:

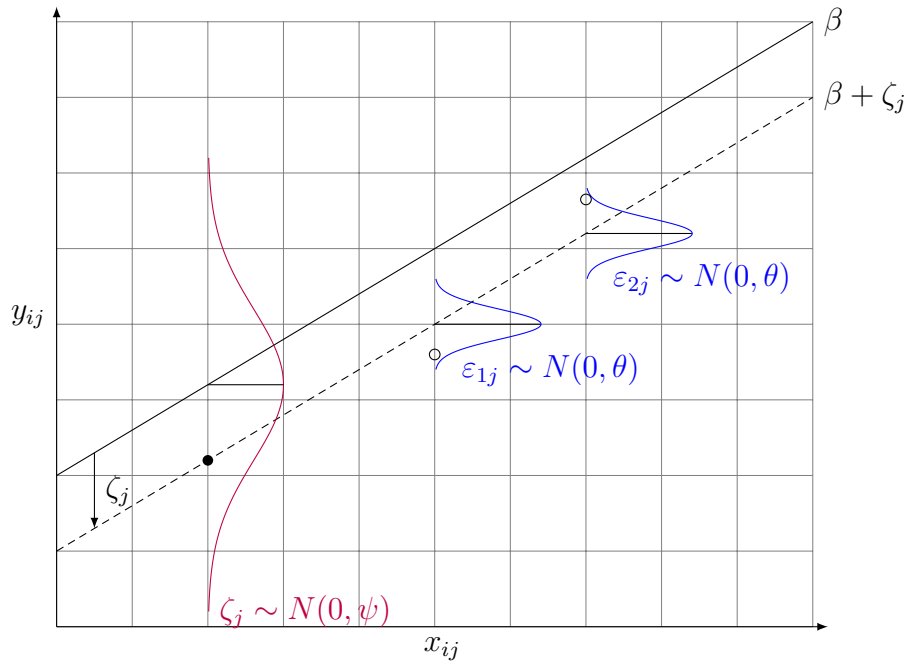


Abbildung 10: Beispiel für ein Random-Intercept Modell für eine Mutter mit zwei Kindern

```
xtreg birwt smoke black, i(momid) mle

Random-effects ML regression                    Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

Random effects u_i ~ Gaussian                   Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3

                                                LR chi2(2)         =    361.17
Log likelihood  =   -65294.9                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -271.7308   17.38706   -15.63   0.000    -305.8088   -237.6527
       black |  -288.0684   26.41911   -10.90   0.000    -339.8489   -236.2879
       _cons |   3527.148   7.451026   473.38   0.000     3512.544    3541.752
-------------+----------------------------------------------------------------
    /sigma_u |   341.5994    6.41113                      329.2621    354.3989
    /sigma_e |   378.5685   3.947582                      370.9099    386.3852
         rho |   .4488009   .0120053                      .4253745    .4724073
------------------------------------------------------------------------------
```

```
Likelihood-ratio test of sigma_u=0: chibar2(01)= 1079.83 Prob>=chibar2 = 0.000
```

Identische Ergebnisse liefert der Befehl `xtmixed`.

```
 xtmixed birwt smoke black || momid:, mle

Mixed-effects ML regression                     Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

                                                Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3

                                                Wald chi2(2)       =    380.31
Log likelihood =   -65294.9                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -271.7309   17.29763   -15.71   0.000    -305.6337   -237.8282
       black |  -288.0684   26.41897   -10.90   0.000    -339.8486   -236.2882
       _cons |   3527.148   7.446306   473.68   0.000     3512.554    3541.743
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                   sd(_cons) |   341.5987   6.411142      329.2614    354.3983
-----------------------------+------------------------------------------------
                sd(Residual) |   378.5688   3.947593      370.9102    386.3855
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  1079.83 Prob >= chibar2 = 0.0000
```

10.2.1 Varianz der Residuen

Im Mehrebenemodell mit Random-Intercept ist die Varianz der Residuen homoskedastisch. Demnach gilt für jede Beobachtung y_{ij} bei gegebenen Kovariablen $\mathbf{x}_{ij}$, dass die Varianz von ζ_j (ψ) und ε_{ij} (θ) konstant ist und nicht mit wechselnden Werten von x_{ij} variiert.

$$Var(y_{ij}|\mathbf{x}_{ij}) = \psi + \theta$$

Ebenso ist der Intraclass Korrelationskoeffizient ρ für gegenebene Kovariablen definiert als:

$$\rho = Cor(y_{ij}, y_{i'j}|\mathbf{x}_{ij}, \mathbf{x}_{i'j}) = \frac{\psi}{\psi + \theta} \tag{35}$$

Wichtig ist zwischen dem Intraclass Korrelationskoeffizienten (ICC) *mit* und *ohne* Kovariablen zu unterscheiden. Der ICC *ohne* Kovariablen wird auch *Nullmodell* oder *unkonditionaler* ICC genannt. Vergleicht man den ICC des Nullmodells mit dem des vollen Modells, so bedeutet ein geringerer ICC, dass die gewählten Ebene 2 Variablen unterschiede zwischen den Gruppen erklären können.

10.2.2 Erklärte Varianz im Modell mit Kovariablen

Im Regressionsmodell beschreibt allgemein der Determinationskoeffizient R^2 den Anteil, der durch das Modell erklärten Varianz (vgl. 1.10). Im linearen Mehrebenenmodell erklärt der Determinationskoeffizient (vgl. Snijders und Bosker (1999)) den Anteil der Varianz, der durch die Kovariablen erklärt wird. Hierfür wird die (Residuen-)Varianz des Modells ohne Kovariablen (Nullmodell) mit der Varianz des Modells mit Kovariablen (volles Modell) verglichen.

Die totale Varianz der Residuen im Random-Intercept Modell ist definiert durch: $Var(\zeta_j + \varepsilon_{ij}) = \psi + \theta$. Diese lässt sich für das Nullmodell $(\widehat{\psi}_0 + \widehat{\theta}_0)$ und das volle Modell $(\widehat{\psi}_1 + \widehat{\theta}_1)$ berechnen und ins Verhältnis setzen.

$$R^2 = \frac{\widehat{\psi}_0 + \widehat{\theta}_0 - \overbrace{(\widehat{\psi}_1 + \widehat{\theta}_1)}^{\text{mit Kovariablen}}}{\widehat{\psi}_0 + \widehat{\theta}_0} \tag{36}$$

Wenn die (geschätzte) Varianz in dem Modell mit Kovariablen ähnlich dem ohne Kovariablen ist, dann geht $R^2 \to 0$ und das Modell hat eine geringe Erklärungskraft.

```
* Nullmodell
quietly xtreg birwt, i(momid) mle
scalar null_u = e(sigma_u)^2 // zeta, Gruppe, Ebene 2
scalar null_e = e(sigma_e)^2 // theta, Individuum, Ebene 1

* Volles Modell
quietly xtreg birwt smoke black, i(momid) mle
```

```
scalar voll_u = e(sigma_u)^2 // zeta
scalar voll_e = e(sigma_e)^2 // theta

* R2 des Modells
scalar r2_smoke = (null_u + null_e - (voll_u + voll_e))/(null_u + null_e)
display r2_smoke
.0656082
```

Der Determinationskoeffizient beträgt 0.0656, d. h. 6.6 % der Varianz wird durch die Kovariablen erklärt. Mit dem Befehl `e(sigma_u)^2` wird die Standardabweichung des Nullmodells und des vollen Modells `e(sigma_e)^2` verwendet und Quadriert. Die Werte werden als Zahl (Skalar) gespeichert. Anschliessend wird mit `scalar r2_smoke` der Determinationskoeffizient nach der Formel (36) berechnet.

Ebenso kann man den Anteil der erklärten Varianz für die Varianzkomponenten (Ebene 1 und Ebene 2) separate berechnen. Ausgedrückt wird dann der Anteil der Ebene 2 Varianz und Ebene 1 Varianz, der durch die Kovariablen erklärt wird.

$$R^2_{Ebene_2} = \frac{\widehat{\psi}_0 - \widehat{\psi}_1}{\widehat{\psi}_0}$$

$$R^2_{Ebene_1} = \frac{\widehat{\theta}_0 - \widehat{\theta}_1}{\widehat{\theta}_0}$$

Die Berechnung der jeweiligen R^2 für Ebene 2 und Ebene 1 geschieht wiederum mit den gespeicherten Werten.

```
scalar r2_smoke_2 = (null_u  - voll_u )/null_u

display r2_smoke_2
.13967529

 * Ebene 1 - Individuen
 scalar r2_smoke_1 = (null_e  - voll_e )/null_e

 display r2_smoke_1
-.00482868
```

Demanch werden 14 % der Varianz auf Ebene 2 durch die Variablen erklärt. Der Anteil der erklärten Varianz durch die individuellen Variablen ist Null, bzw. in diesem Fall sogar negativ. Vergleicht man das Nullmodell mit dem vollen Modell, so erhöht sich die Varianz auf Ebene 1 durch Hinzunahme der Variable `smoke`. Ebenso ist es möglich, dass sich die Varianz auf Ebene 2 erhöhen kann, wenn Ebene 1 Variablen hinzugenommen werden.

Der Intraclass Korrelationskoeffizient ρ beträgt im Nullmodell `.4874391` und reduziert sich im Modell mit Kovariablen auf `.4488009`. Der ICC des vollen Modells kann auch grösser sein als der ICC des Nullmodells, wenn durch Hinzunahme einer Variable in der ICC-Gleichung ($\frac{\psi}{\psi+\theta}$) die Ebene 1 Varianz (θ) stärker abnimmt als die Ebene 2 Varianz (ψ).

10.2.3 Hypothesentest der Koeffizienten

Die Signifikanz des Gesamtmodells, dass eine der Kovariablen einen Einfluss hat wird generell mit einem Likelihood-Ratio Test getestet. Die Signifikanz einer Variable wird mittels einer z-Statistik ermittelt, die einer Standardnormalverteilung folgt.

$$z = \frac{\widehat{\beta}_1}{\widehat{SE}(\widehat{\beta}_1)}$$

Die z-Werte werden im `Stata`-Output angegeben und Werte $< .05$ können als signifikant interpretiert werden.

Möchte man testen, ob mehrere Variablen eines Modells eine Einfluss haben, gegeben der anderen Variablen, so macht man eine Likelihood-Ratio-Test. Zuerst wird das volle Modell mit allen Variablen gerechnet und anschliessend ein Test mit dem reduzierten Modell gerechnet. Mit dem `lrtest` Befehl werden die beiden Modelle verglichen.

```
* volles Modell
quietly xtreg birwt smoke black kessner2 kessner3, i(momid) mle
estimates store full
* reduziertes Modell
quietly xtreg birwt smoke black, i(momid) mle
lrtest full

Likelihood-ratio test                                 LR chi2(2)  =     18.32
(Assumption: . nested in full)                        Prob > chi2 =    0.0001
```

Getestet wird die Hypothese, dass der Kessner-Index (eine Masszahl der Geburtsvorsorge) *keinen* Einfluss hat. Diese Hypothese kann abgelehnt werden ($\chi^2(2)$ – Chi-Quadrat mit zwei Freiheitsgraden – 18.32, $p-$Wert < 0.0001).

Der Test, dass die Varianz zwischen den Gruppen 0 ist ($H_0 : \psi = 0$), wird auch mit einem Likelihood-Ratio Test getestet, der am Ende des `Stata`-Outputs steht ($L = 1079.83$ und $p < 0.001$).

10.3 Hausman Endogenitäts Test

Der Hausman-Test von Hausman (1978) wurde entwickelt, um β Koeffizienten von alternativen Modellen zu testen. Konkret dient der Hausman Test dazu, eine Entscheidungsgrundlage dafür zu liefern, ob das Random-Effects Modell passend ist oder ob ein Fixed-Effects-Modell gewählt werden soll. Zwar handelt es sich bei den Koeffizienten des Random-Effects Modells ($\widehat{\beta}^R$) um das effizientere Modell, allerdings werden die Schätzer inkonsistent, sobald Modellannahmen verletzt sind. Eine entscheidende Modellannahmen ist die Unkorreliertheit zwischen dem Random-Intercept und den Kovariablen. Daher testet der Hausman Test, ob es zu systematischen Unterschieden zwischen dem Random-Effect-Modell und den Koeffizienten des Fixed-Effects Modells ($\widehat{\beta}^W$) kommt.

Der Hausman-Test testet Variablen, die eine Korrelation sowohl zwischen den Gruppen als auch innerhalb der Gruppen aufweisen.

$$h = (\widehat{\beta}^W - \widehat{\beta}^R)' \left(\widehat{Cov}(\widehat{\beta}^W) - \widehat{Cov}(\widehat{\beta}^R) \right)^{-1} (\widehat{\beta}^W - \widehat{\beta}^R) \tag{37}$$

Der Hausman Test folgt einer χ^2-Verteilung und die Anzahl der Variablen mit Varianz zwischen als auch in den Gruppen gibt die Zahl der Freiheitsgrade an. Mit dem Befehl `hausman` wird der Test ausgeführt. Zuvor werden die $\widehat{\beta}^W$ Koeffizienten mit `xtreg` und der `fe` Option gewonnen und die $\widehat{\beta}^R$ mit `xtreg` und der `re` Option gewonnen.

```
* Fixed-Effekte
xtreg birwt smoke mage black, i(momid) fe
estimates store fixed
* Random Effekte
xtreg birwt smoke mage black, i(momid) re
estimates store random
* Hausman Test
hausman fixed random
```

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+------------------------------------------------------------
       smoke |   -105.6992    -249.0647        143.3655        23.85004
        mage |    23.11536     10.70324        12.41213        2.805553
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
```

```
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       58.63
                Prob>chi2 =      0.0000
```

Die Hypothese `Ho: difference in coefficients not systematic` kann zum 5 % Signifikanzniveau abgelehnt werden. D. h. die Schätzer des Random-Effects-Modells sind nicht effizient und das Modell könnte falsch spezifiziert worden sein. Den Ergebnissen zu Folge sollte man ein Fixed-Effects-Modell dem Random-Effects-Modell vorziehen. Man kann aber auch das Modell genauer spezifizieren. So ist denkbar, die Gruppenmittelwerte, also die Mittelwerte jeder Mutter für `mage` und `smoke` in das Random-Effects Modell einzufügen.

```
* Mittelwerte Erzeugen
egen mn_smoke = mean(smoke), by(momid)
egen mn_mage = mean(mage), by(momid)

*Modelle berechnen
xtreg birwt smoke mn_smoke mage mn_mage black, i(momid) fe
estimates store fixed_1

xtreg birwt smoke mn_smoke mage mn_mage black, i(momid) re
estimates store random_1
```

Der Output des Fixed-Effects Modells zeigt, dass die Gruppenmittelwerte keine Varianz innerhalb der Gruppe haben und unterdrückt daher diese Variablen

```
xtreg birwt smoke mn_smoke mage mn_mage black, i(momid) fe
note: mn_smoke omitted because of collinearity
note: mn_mage omitted because of collinearity
note: black omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

R-sq:  within  = 0.0149                         Obs per group: min =         2
       between = 0.0441                                        avg =       2.2
       overall = 0.0378                                        max =         3

                                                F(2,4624)          =     34.95
corr(u_i, Xb)  = -0.0808                        Prob > F           =    0.0000
```

```
------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -105.6992   29.53343    -3.58   0.000    -163.5989   -47.79961
    mn_smoke |  (omitted)
        mage |   23.11536   3.050865     7.58   0.000     17.13421    29.09652
     mn_mage |  (omitted)
       black |  (omitted)
       _cons |   2823.812   87.39633    32.31   0.000     2652.473     2995.15
-------------+----------------------------------------------------------------
     sigma_u |  442.92373
     sigma_e |  374.73056
         rho |  .58282488   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(3977, 4624) =     2.78          Prob > F = 0.0000

estimates store fixed_1


 xtreg birwt smoke mn_smoke mage mn_mage black, i(momid) re

Random-effects GLS regression                   Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

R-sq:  within  = 0.0149                         Obs per group: min =         2
       between = 0.1011                                        avg =       2.2
       overall = 0.0801                                        max =         3

Random effects u_i ~ Gaussian                   Wald chi2(5)       =    521.95
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -105.6992   29.51831    -3.58   0.000    -163.5541   -47.84441
    mn_smoke |  -227.2295    36.5382    -6.22   0.000     -298.843   -155.6159
        mage |   23.11536   3.049303     7.58   0.000     17.13884    29.09189
     mn_mage |   -15.5921   3.317603    -4.70   0.000    -22.09448   -9.089715
       black |  -260.0303   26.61759    -9.77   0.000    -312.1998   -207.8607
       _cons |   3318.902   38.94571    85.22   0.000     3242.569    3395.234
-------------+----------------------------------------------------------------
     sigma_u |  341.66312
     sigma_e |  374.73056
         rho |  .45393994   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
estimates store random_1
```

Der Output des Random-Effects Modells mit Gruppenmittelwerten zeigt, dass sich die Gruppenmittelwerte (`mn_smoke, mn_mage`) signifikant von den individuellen Schätzern (`smoke, mage`) unterscheiden. Die Schätzer der Effekte innerhalb der Gruppen sind identisch zu den Fixed-Effects Schätzern, wie der Hausman-Test zeigt.

```
 hausman fixed_1 random_1

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |    fixed_1      random_1      Difference          S.E.
-------------+------------------------------------------------------------
       smoke |   -105.6992    -105.6992        2.41e-07        .9450424
       mage |    23.11536     23.11536       -1.49e-07        .0976248
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        0.00
                Prob>chi2 =      1.0000
```

Der Grund für die Spezifizierung des Modells ist, dass man durch die Hinzunahme von Gruppenmittelwerten für ausgewählte Variablen einen effizienten Schätzer erhält und das Random-Effects Modell angewendet werden kann. Man kann dadurch Variablen, die die Gruppenunterschiede modellieren zusätzlich erfassen. Da Forschungsfragen oft den Effekt von Gruppenvariablen modellieren möchten, ist dieses Modell einem Fixed-Effects Modell vorzuziehen.

10.4 Random Koeffizienten Modell

Das Random-Intercept Modell wird zum Random-Koeffizienten Modell, wenn man die Steigung der Gruppenmittelwerte variieren lässt. Wie angedeutet, können nun auch die (Steigungs)Koeffizienten (*slope* = Steigung) variieren und sind nicht mehr, wie im Random-Intercept Modell, fix. Damit variiert der Effekt der Kovariablen nun auch pro Gruppe. Ein Modell, das sowohl aus Random-Intercept als auch aus Random-Slopes besteht, wird Random-Koeffizienten Modell genannt.

Das Beispiel stammt aus (Rabe-Hesketh und Skrondal 2008: Kapitel 4) und untersucht den Zusammenhang zweier Testergebnisse (Scores) von Schülern unterschiedlicher

Schulen. Zum einen liegt der Score des "Graduate Certificate of Secondary Education" (`gcse`) vor, der mit 16 Jahren ermittelt wird, zum andern gibt es Daten über den "London Reading Test" (`lrt`), der mit 10 Jahren bei den Schülern ermittelt wird. Von Interesse ist nicht nur der Zusammenhang dieser beiden Scores sondern auch, wie dieser Zusammenhang zwischen den Schulen variiert. Als Cluster bzw. Gruppen gelten daher 65 Schulen (Ebene 2), Ebene 1 bilden die Schüler.

Für eine Schule (# 65) wird der schulspezifische Intercept und der Zusammenhang der beiden Leistungsscores für die Schüler der Schule ermittelt.

```
regress gcse lrt if school == 65
```

```
      Source |       SS       df       MS              Number of obs =      80
-------------+------------------------------           F(  1,    78) =   59.59
       Model |  2719.57841     1  2719.57841           Prob > F      =  0.0000
    Residual |   3559.9141    78  45.6399243           R-squared     =  0.4331
-------------+------------------------------           Adj R-squared =  0.4258
       Total |  6279.49251    79   79.487247           Root MSE      =  6.7557

------------------------------------------------------------------------------
        gcse |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lrt |   .5684563   .0736408     7.72   0.000     .4218487     .715064
       _cons |  -1.749008   .7749431    -2.26   0.027    -3.291801   -.2062151
------------------------------------------------------------------------------
```

Der positive Zusammenhang kann auch graphisch mit einem Scatterplot dargestellt werden, wie Abbildung 11 zeigt.

```
twoway (scatter gcse lrt) (lfit gcse lrt) if school == 65, ///
xtitle(Schlaubi Schlumpf - LRT) ytitle(Daniel Düsentrieb - GCSE)
```

Diese Graphik kann auch für mehr als eine Schulen erstellt werden (Abb: 12). Es ist zu sehen, dass der Zusammenhang positiv ist, dass aber die Steigung (= Slope) bei einigen Schulen steiler als bei anderen Schulen ist und das der Intercept – der Schnittpunkt mit dem Wert ($x = 0$) variiert.

```
twoway  (lfit gcse lrt if school==1) ///
(lfit gcse lrt if school==2) ///
(lfit gcse lrt if school==3) ///
(lfit gcse lrt if school==4)  ///
(lfit gcse lrt if school==15)  ///
(lfit gcse lrt if school==20)  ///
(lfit gcse lrt if school==30)  ///
(lfit gcse lrt if school==55)
```

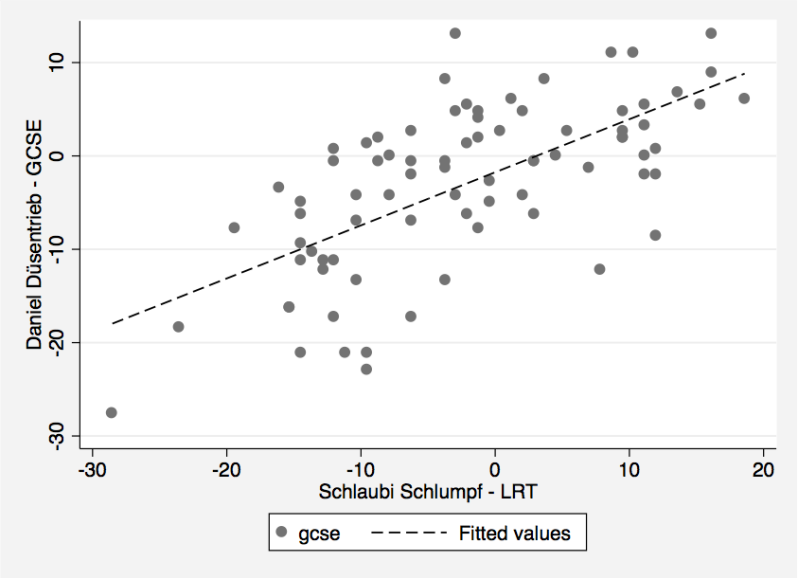


Abbildung 11: Scatterplot des Zusammenhangs von GCSE und LRT Testscores in 65 Schulen

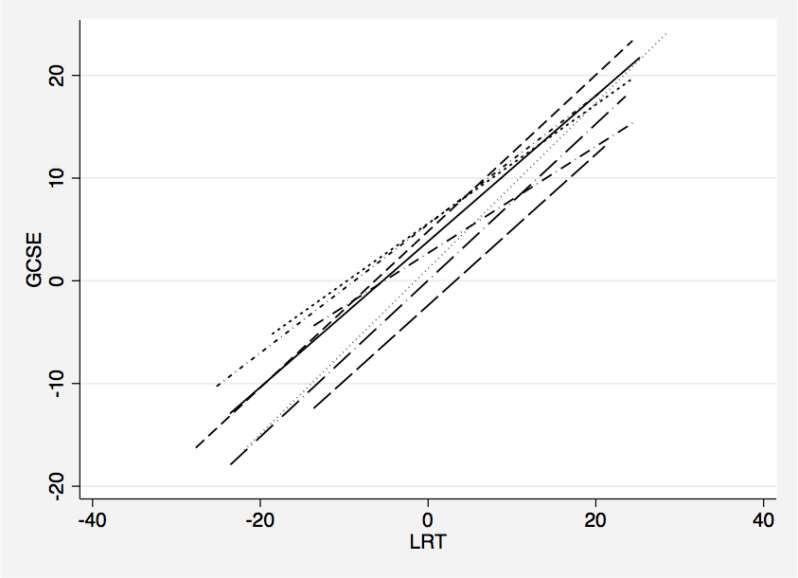


Abbildung 12: Scatterplot des Zusammenhangs von GCSE und LRT Testscores für acht Schulen

Die Gerade der geschätzten Werte erhält man für jede Schule mit:

```
twoway  (lfit gcse lrt, by(school)), xtitle(LRT) ytitle(GCSE)
```

Nun soll die Steigung und der Intercept Wert einer bivariaten Regression für jede Schule gespeichert werden und anschliessend in einem Graph ausgegeben werden. Dabei wird das `statsby` Kommando verwendet. Für jede Schule wird dabei eine Regression durchgeführt und die Werte des Intercept (`_b[_cons]`) und der Steigung (`_b[lrt]`) in einem neuen Datensatz `ols_schools.dta` gepeichert. statsby

```
statsby intercept=_b[_cons] steigung=_b[lrt], by(school) saving(ols_schools, replace) ///
         :regress gcse lrt if num > 4
(running regress on estimation sample)

      command:  regress gcse lrt if num > 4
    intercept:  _b[_cons]
     steigung:  _b[lrt]
           by:  school

Statsby groups
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..............
```

Die neuen Variablen `intercept` und `steigung` werden nun mit dem Befehl `merge` in den Datensatz `gcse.dta` eingelesen, wobei der Datensatz zuerst nach der Variable `school` sortiert wird. Mit der Option `m:1` wird allen Personen mit einer Schule der gleiche Intercept und Steigungswert zugeordnet.

```
 * Sortieren nach Schulnummer
sort school
* Zusammenfügen der Daten
merge m:1 school using ols_schools

    Result                           # of obs.
    -----------------------------------------
    not matched                             2
        from master                         2  (_merge==1)
        from using                          0  (_merge==2)

    matched                             4,057  (_merge==3)
    -----------------------------------------
drop _merge // automatisch erzeugte Variable
```

Zwei Beobachtungen konnten nicht zugeordnet werden, da für diese Werte keinen Steigungs- und Interceptwerte erzeugt wurden, da die Klasse nur zwei Schüler hat.

Um sich die Variabilität von Intercept und Steigung zwischen den Schulen graphisch anzusehen kann man sich einen Scatterplot ausgeben lassen (Abb: 13). Zu erkennen

ist, dass es erhebliche Unterschiede zwischen den Schulen gibt. Bei gleichem Intercept variiert die Steigung stark. tag()

```
twoway (scatter steigung intercept), xtitle(Intercept) ytitle(Steigung)
```

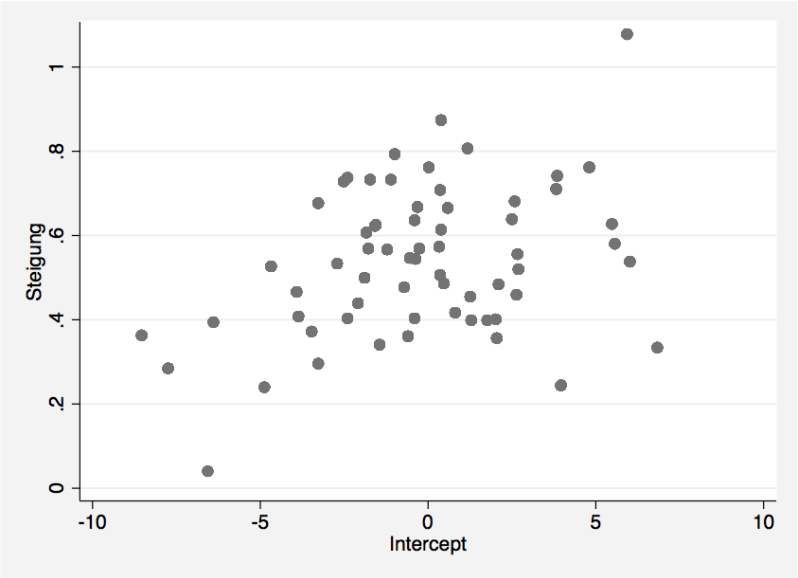


Abbildung 13: Unterschiede in Steigung und Intercept in Schulen mit mehr als 4 Schülern

Praktisch ist der Befehl `egen` mit der Option `tag(`*`groupvar`*`)`. Mit dem Befehl wird eine Beobachtung pro Gruppe mit einer 1 versehen und anschliessend mit der Bedingung `if pickone == 1` nur eine Person pro Gruppe für die Berechnungen ausgewählt.

```
* Dummy der nur einer Person pro Gruppe eine 1 zuordnen allen anderen 0
egen pickone = tag(school)
```

Anschliessend kann man statistische Masszahlen wie bspw. Korrelationen der Gruppenvariablen berechnen.

```
egen pickone = tag(school)
summarize intercept steigung if pickone==1
pwcorr intercept steigung if pickone ==1, sig
```

Zur besseren Veranschaulichung wird ein Graph mit allen Steigungskoeffizienten pro Schule produziert. Hierfür wird eine Variable `pred` erzeugt, die den geschätzten Regressionswert der beobachteten Variable `gcse` in Abhängigkeit von `lrt` enthält. Anschliessend werden die Daten erst nach Schulen und dann innerhalb der Schulen nach `lrt` aufsteigend sortiert. Der Graph wird erzeugt, indem einen Linie zwischen den Punkten gezogen wird solange höhere Werte vorhanden sind (`connect(ascending)`), dann wird eine neue Linie begonnen (Abb: 14). Als Schätzmodell sollte man sich dabei $\widehat{y}_{ij} = \widehat{\beta}_{0j} + \widehat{\beta}_{1j} x_{ij}$ vor Augen halten.

```
generate pred = intercept + steigung*lrt
sort school lrt
twoway (line pred lrt, connect(ascending)), ///
xtitle(LRT) ytitle(Geschätzte Regressionsgerade)
```

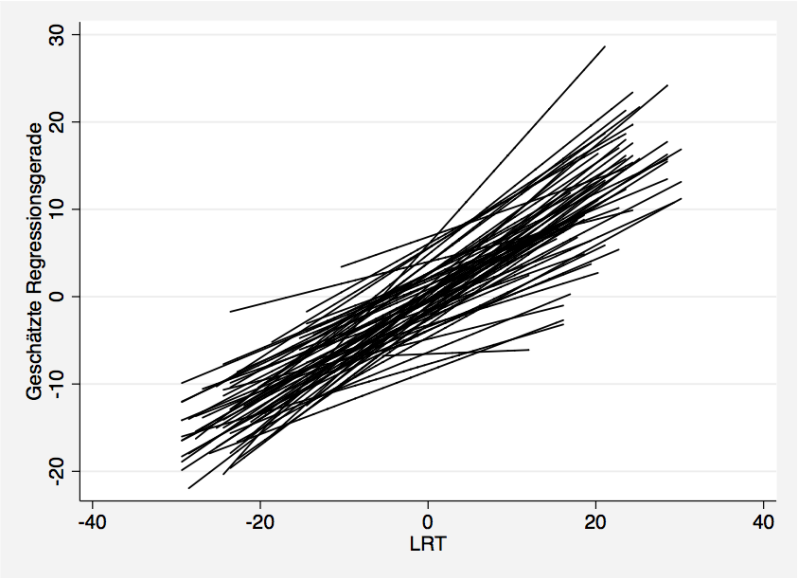


Abbildung 14: Geschätzte Regressionsgerade pro Schule

10.4.1 Modellierung des Random-Koeffizienten Modells

Das Ziel dieses Kapitels ist es ein Modell zu finden, das den Zusammenhang zwischen zwei Testscores (`gcse` und `lrt`) darstellt, wobei schulspezifische Merkmale nicht nur durch einen variierenden Intercept, sondern auch durch einen variierenden Steigungskoeffizienten ausgedrückt werden sollen. Von Interesse ist dabei ein verallgemeinerter Intercept ζ_{1j} und Slope ζ_{2j} für die Population der Schulen.

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + \zeta_{1j} + \zeta_{2j} x_{ij} + \varepsilon_{ij} \\ &= (\beta_0 + \zeta_{1j}) + (\beta_1 + \zeta_{2j}) x_{ij} + \varepsilon_{ij} \end{aligned} \tag{38}$$

ζ_{1j} drückt die schulspezifische Abweichung vom durchschnittlichen Intercept β_0 aus und ζ_{2j} ist die Abweichung der j-ten Schule von der durschnittlichen Steigung β_1. Alle drei Koeffizienten sind unkorreliert mit den einzelnen Beobachtungen. ε_{ij} korreliert nicht mit ζ_{1j} und ζ_{2j}, ferner sind ζ_{1j} und ζ_{2j} unabhängig zwischen den Schulen und die Residuen der Ebene 1 ε_{ij} sind unabhängig zwischen Schulen und Schülern.

Das Random-Koeffizienten Modell setzt sich zusammen aus einem Regressionsmodell ohne Gruppierungsvariablen.

$$E(y_{ij}|x_{ij}) = \beta_0 + \beta_1 x_{ij}$$

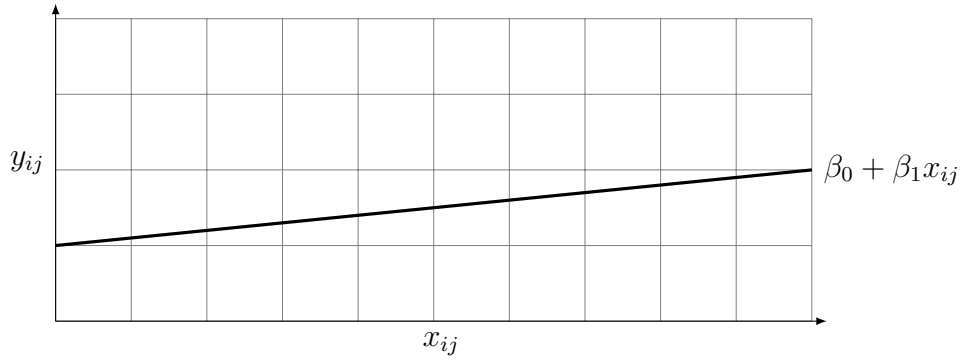


Abbildung 15: Regressionsmodell

Das Random-Intercept Modell verläuft parallel zum Populationsmittelwert mit Abweichungen pro Schule um den Gruppenmittelwert (Die Beobachtungspunkte sind hier nicht eingetragen, aber wie in Abbildung 10 gedacht.)

$$E(y_{ij}|x_{ij}, \zeta_j) = (\beta_0 + \zeta_j) + \beta_1 x_{ij}$$

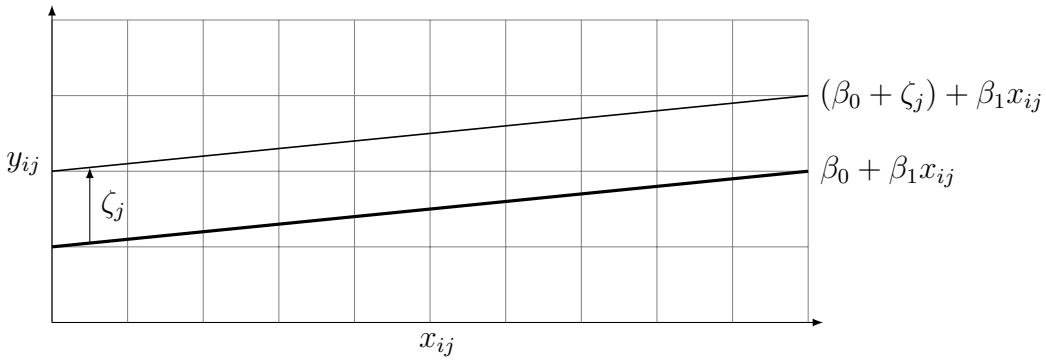


Abbildung 16: Entwicklung vom Regressions- zum Random-Slope Modell: Random-Intercept Modell

Das Radom-Koeffizientenmodell (die gestrichelte Linie) verläuft nicht mehr parallel zum Gruppenmittelwert, da die Steigung pro Schule variiert. In dem Beispiel ist die

Steigung grösser und setzt sich zusammen aus $\beta_1 + \zeta_{2j}$. Der Intercept setzt sich aus $\beta_0 + \zeta_{1j}$ zusammen. Unschwer zu erkennen ist, dass $\zeta_j = \zeta_{1j} + \zeta_{2j}$ als Interaktionseffekt betrachtet werden kann.

$$E(y_{ij}|x_{ij}, \zeta_{1j}, \zeta_{2j}) = (\beta_0 + \zeta_{1j}) + (\beta_1 + \zeta_{2j})x_{ij}$$

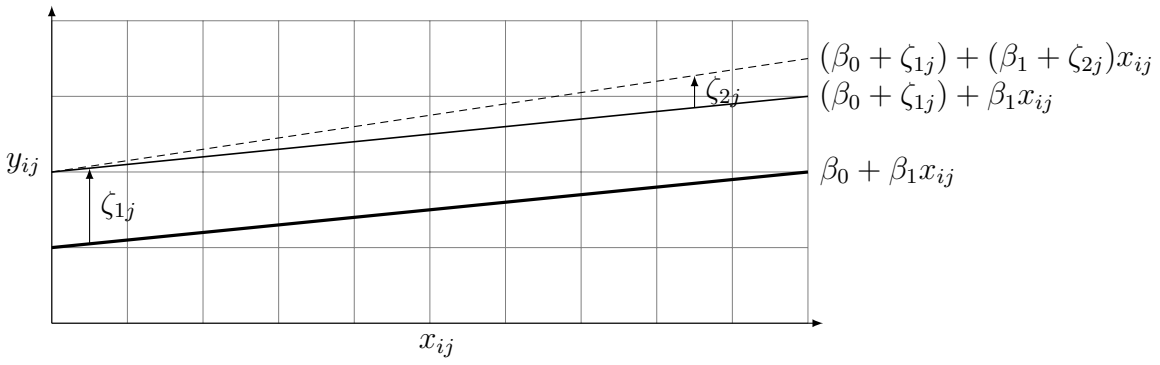


Abbildung 17: Entwicklung vom Regressions- zum Random-Koefizienten Modell: Random-Intercept Modell

Die Gesamtvarianz ξ_{ij} ("xi") des Random-Koeffizienten Modells setzt sich nicht mehr, wie im Random-Intercept Modell aus den konstanten Termen ψ (Varianz der Ebene 2) und θ (Varianz der Ebene 1) zusammen, sondern besteht aus vier, teilweise variierenden Termen.

$$Var(y_{ij}|x_{ij}) = Var(\xi_{ij}|x_{ij}) = \psi_{11} + 2\psi_{21}x_{ij} + \psi_{22}x_{ij}^2 + \theta \tag{39}$$

Zu erkennen ist, dass die Varianz von den x_{ij}-Werten der Kovariablen abhängt, daher ist das Gesamtresidual *heteroskedastisch*. Es ist nicht sinnvoll Determinationskoeffizienten (R^2, R_2^2, R_1^2) zu bestimmen, da die Werte von x_{ij} abhängen. Ebenso variiert die Varianz ψ_{11} bzw. die Kovarianz ψ_{21} von der Transformation der Variable. Je nach Nullpunkt der Kovariable (hier `lrt`) variiert die Varianz der Intercepte.

Die unterschiedlichen Varianzkomponenten ψ_{nm} des Modells stammen aus der Kovarianzmatrix.

$$\Psi = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix} = \begin{bmatrix} Var(\zeta_{1j}|x_{ij}) & Cov(\zeta_{1j}, \zeta_{2j}|x_{ij}) \\ Cov(\zeta_{2j}, \zeta_{1j}|x_{ij}) & Var(\zeta_{2j}|x_{ij}) \end{bmatrix}, \quad \psi_{12} = \psi_{21}$$

10.5 Random-Koeffizientenmodell mit Stata

Random-Koeffizientenmodell kann man in `Stata` mit `xtmixed` und `gllamm` berechnen. `gllamm` hat zusätzliche Diagnosemöglichkeiten, ist allerdings bei der Berechnung nicht so effizient wie `xtmixed`.

10.5.1 Random-Intercept Modell mit `xtmixed`

Zuerst wird ein Random-Intercept Modell berechnet, dass die Eigenschaft hat, dass die Steigungsvarianz $\zeta_{2j} = 0$ so das die Kovarianz von Steigung und Intercept $\psi_{12} = \psi_{12} = 0$. Das resultierende Modell lautet:

$$y_{ij} = (\beta_0 + \zeta_{1j}) + \beta_1 x_{ij} + \varepsilon_{ij} \tag{40}$$

```
xtmixed gcse lrt ||school:, mle
```

```
Mixed-effects ML regression                     Number of obs      =      4059
Group variable: school                          Number of groups   =        65

                                                Obs per group: min =         2
                                                               avg =      62.4
                                                               max =       198


                                                Wald chi2(1)       =   2042.57
Log likelihood = -14024.799                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        gcse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lrt |   .5633697   .0124654    45.19   0.000     .5389381    .5878014
       _cons |   .0238706   .4002255     0.06   0.952     -.760557    .8082982
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Identity             |
                   sd(_cons) |   3.035269   .3052513      2.492261    3.696587
-----------------------------+------------------------------------------------
                sd(Residual) |   7.521481   .0841759      7.358295    7.688285
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   403.27 Prob >= chibar2 = 0.0000
```

```
estimates store random_intercept
```

Die Ergebnisse werden mit `estimates store` gespeichert.

Die Varianzanalyse zeigt, dass die Schulen im Mittel mit einer geschätzten Standardabweichung von 3.05 variieren. Innerhalb der Schulen variieren die Residuen um die schulspezifische Regressionslinie mit 7.52. Die Korrelation innerhalb der Schulen, nach Kontrolle der Kovariablen `lrt`, wird mit 0.14 berechnet.

```
scalar rho_ri = 3.0353^2/(3.0353^2+7.521^2)
display rho_ri
.14006169
```

Dieser Wert kann ebenso als die Heterogenität zwischen den Gruppen betrachtet werden. Wenn die Streuung innerhalb der Gruppen gering ist wird der Wert $\widehat{\rho} = \frac{\widehat{\psi}_{11}}{(\widehat{\psi}_{11}+\widehat{\theta})}$ gross, ist die Streuung innerhalb der Gruppe dagegen sehr hoch, wie in unserem Fall, geht der Wert gegen 0.

10.5.2 Random-Koeffizienten Modell mit `xtmixed`

Ausgangsmodell ist nun die Annahme, dass auch die Steigung variiert ($\zeta_{2j} \neq 0$).

$$y_{ij} = (\beta_0 + \zeta_{1j}) + (\beta_1 + \zeta_{2j})x_{ij} + \varepsilon_{ij} \tag{41}$$

In dem Beispiel soll für `lrt` ein Random-Slope modelliert werden. In `xtmixed` wird dieser Term durch `school: lrt` spezifiziert. Zusätzlich wird die Option `covariance(unstructured)` gewählt. Anderenfalls würde die Kovarianz ψ_{21} Null gesetzt werden.

```
 xtmixed gcse lrt ||school: lrt, mle cov(unstructured)

Mixed-effects ML regression                     Number of obs      =      4059
Group variable: school                          Number of groups   =        65

                                                Obs per group: min =         2
                                                               avg =      62.4
                                                               max =       198
```

```
                                                Wald chi2(1)       =    779.79
Log likelihood = -14004.613                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        gcse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lrt |    .556729   .0199368    27.92   0.000     .5176535    .5958044
       _cons |   -.115085   .3978346    -0.29   0.772    -.8948264    .6646564
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Unstructured         |
                     sd(lrt) |   .1205646   .0189827      .0885522    .1641498
                   sd(_cons) |   3.007444   .3044148      2.466258    3.667385
              corr(lrt,_cons) |   .4975415   .1487427      .1572768    .7322094
-----------------------------+------------------------------------------------
                sd(Residual) |   7.440787   .0839482      7.278058    7.607155
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =   443.64   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

* Speichern der Ergbenisse
estimates store random_koeffizient
```

Die Ergebnisse werden wiederum mit `estimates store` gespeichert. Man erhält die Werte der Kovarianzmatrix mit dem Postestimationbefehl `etat recovariance`

```
estat recovariance

Random-effects covariance matrix for level school

             |       lrt      _cons
-------------+----------------------
         lrt |  .0145358
       _cons |  .1804042    9.04472
```

Die Werte erhält man auch, wenn man die Option `variance` im `xtmixed` Befehl verwendet.

```
xtmixed gcse lrt ||school: lrt, mle cov(unstructured) variance
```

10.5.3 Modell Test

Zuerst wird getestet, ob die Modellierung der Steigung zu einer Verbesserung des Modells führt oder ob ein Random-Intercept Modell nicht ausreicht. Getestet wird die Nullhypothese, dass die Random-Slopes Null sind.

$$H_0 : \psi_{22} = \psi_{21} = 0$$

Hierfür wird ein Likelihood-Ratio Test mit den gespeicherten Daten durchgeführt.

```
lrtest random_koeffizient random_intercept

Likelihood-ratio test                                     LR chi2(2)  =     40.37
(Assumption: random_inter~t nested in random_koeff~t)     Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis
is not on the boundary of the parameter space.  If this is not true,
 then the reported test is conservative.
```

Die Nullhypothese kann abgelehnt werden, so dass das Random-Intercept Modell abgelehnt wird und ein Random-Slope Modell vorzuziehen ist. Der Hinweis (`Note:`) im Output sagt, dass der Test konservativ ist. In der Tat werden naive p-Werte angegeben. Die korrekten p-Werte erhält man, indem man den p-Wert durch 2 teilt – was in diesem Fall nichts an den Ergebnissen ändert.

10.5.4 Interpretation der Koeffizienten im Random-Koeffizienten Modell

Die mittlere Steigung des Random-Koeffizientenmodell für die Variable `lrt` ist gleich zu dem Koeffizienten im Random-Intercept Modell. Die Standardabweichungen für Ebene 1 Residuen und für den Intercept sind etwas kleiner, da durch die variierenden Slopes pro Gruppe, die Anpassung an die Daten besser gelingt (Im Random-Koeffizienten Modell wird die Restriktion der parallelen Regressionsgeraden aufgehoben).

Um die geschätzte Standardabweichung der Random-Intercepts und der Random-Slopes besser interpretieren zu können, werden Intervalle gebildet. Die Intervalle geben an, dass 95 % der Random-Intercepts und Random-Slopes aller Schulen in dem Intervall vermutlich liegen. Dabei handelt es sich nicht um Konfidenzintervalle, da Konfidenzintervalle von einem *unbekannten Parameter* ausgehen, sondern man spricht von Realisierungen einer *Zufallsvariable*.

Die Formel zur Berechnung des Intervalls des Intercepts ist $\widehat{\beta}_0 \pm 1.96\sqrt{\psi_{11}}$ und für die Steigungsgeraden $\widehat{\beta}_1 \pm 1.96\sqrt{\psi_{22}}$.

Man erhält also für die Intercepts ein Intervall von -6.0 bis 5.8 ($-0.12 \pm 1.96 \times 3.01$). Damit kann man sagen, dass der Schulmittelwert des GCSE-Scores mit durchschnittlichem LRT-Wert (`lrt`=0) variiert zwischen -6.0 und 5.8. Die Slopes variieren zwischen 0.32 und 0.80 ($0.56 \pm 1.96 \times 0.12$). Damit liegen 95 % der schulspezifischen Slopes zwischen 0.32 und 0.85. Problematisch wird die Interpretation, wenn die Slopes je nach Schule die Vorzeichen wechseln würden. Die Bandbreite von 0.32 bis 0.85 in den Steigungsparametern ist gross. Dies bedeutet, dass eine Schule mit hohem Steigungsparamter dazu führt, dass ein Schüler mit geringem LRT-Wert einen hohen GCSE-Wert erhalten kann. Schulen mit kleiner Steigung führen dazu, dass Schüler mit hohem LRT-Wert einen relativ niedrigeren GCSE-Wert erhalten. Das heisst, dass manche Schulen bei gleichem LRT-Wert "mehr Wert produzieren", also zu einem höheren GCSE-Wert führen.

Die geschätzte Korrelation $\widehat{\rho}_{21} = 0.4975$ (`corr(lrt, _cons)`) sagt, dass Schulen mit grösserem GCSE-Mittelwert für Schüler mit durchschnittlichem LRT-Wert verglichen mit anderen Schulen auch einen höheren Steigungsparameter haben. Kurz, Intercept und Steigungsparamter korrelieren positiv.

Man kann sich die Ergebnisse auch visualisieren, siehe Abbildung 18.

```
estimates restore random_intercept
predict muri, fitted
sort school lrt
twoway(line muri lrt, connect(ascending)), xtitle(LRT) ///
ytitle(Regressionslinien für Random Intercept) name(ri, replace)

* Regressionsgeraden für Random Koeffizienten Modell
estimates restore random_koeffizient
predict murk, fitted
sort school lrt
twoway(line murk lrt, connect(ascending)), xtitle(LRT) ///
ytitle(Regressionslinien für Random Koeffizient) name(rk, replace)
* Graphen kombinieren
graph combine ri rk
```

Die Graphik zeigt die Variabilität des Modells, wenn die Annahme der parallelen Regressionsgeraden für jede Schule aufgegeben wird. Hierfür werden mit dem `predict` Befehl und der Option `fitted` Regressionsgeraden für jede Schule geschätzt.

Die Geschätzten Koeffizienten kann man sich auch mit dem Befehl `predict` und der Option `reffects` ausgeben lassen, wobei zwei neue Variablen spezifiziert werden: die erste Variable ist der Steigungskoeffizient pro Schule ζ_{2j} die letzte Variable ist der Intercept ζ_{1j}. Das Verfahren zur Schätzung der Koeffizienten wird mit "Empirical Bayes Vorhersage" bezeichnet.

```
* Ergebnisse wieder laden
```

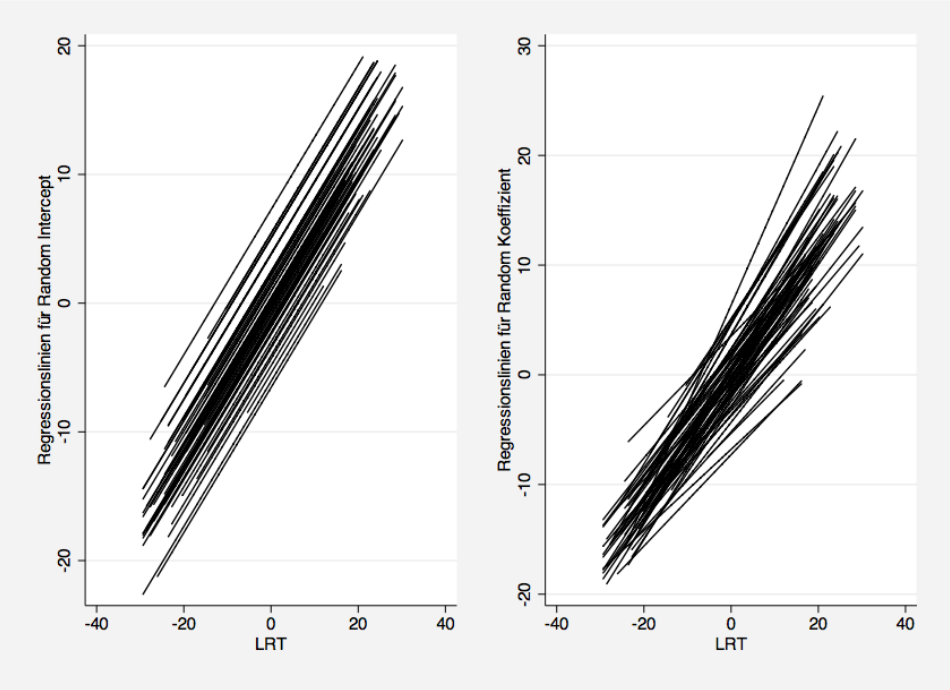


Abbildung 18: Visualisierung der geschätzten Regressionsgeraden

```
estimates restore random_koeffizient
* Koeffizienten berechnen
predict steigung intercept, reffects
* Ausgabe der Koeffizienten
list school steigung intercept if pickone==1 & school <20, noobs
  +-------------------------------+
  | school   steigung   intercept |
  |-------------------------------|
  |     19   .7928017    -.986566 |
  |      4   .7614464    .0375372 |
  |      2   .7612875    4.822753 |
  |     15   .7349201   -2.392471 |
  |      1   .7093406    3.833302 |
  |-------------------------------|
  |      5   .6800166     2.60444 |
  |     13   .6246431   -1.524703 |
  |     14   .6069078   -1.825554 |
  |      3   .5789855    5.577505 |
  |      8   .5674071     -.25197 |
  |-------------------------------|
  |      6   .5353431    6.032066 |
  |     17   .4976305   -1.865862 |
  |     12   .4768683   -.6831335 |
```

```
|     11   .4586335    2.659585 |
|     16   .4069399   -3.856449 |
|------------------------------|
|      9   .4011957   -2.373684 |
|     18   .3593853   -.5734447 |
|     10   .2950493   -3.252382 |
|      7    .242275    3.985227 |
+------------------------------+
```

10.6 Random-Koeffizienten Modell

Verwendet man ein Random-Koeffizienten Modell, so sollte zu dem Random-Slope auch immer ein Random-Intercept eingefügt werden. Mit der Verwendung von Random-Slopes sollte man sparsam umgehen, da die Anzahl der zu schätzenden Koeffizienten stark ansteigt. Für jeden Random-Effekt gibt es einen Varianzparameter und für jede Kombination der Random-Effekte gibt es ein Kovarianzparameter, so dass man beispielsweise bei drei Random-Slopes bereits 11 Parameter erhält. Inhaltlich macht die Verwendung von Random-Slopes nur Sinn, wenn es theoretische Gründe in der bisherigen Forschungsliteratur gibt, dass man variierende Steigungskoeffizienten modellieren sollte.

Übersicht Stata-Befehle

Befehl	Option	Erklärung
`hausman`		Hausman-Test
`xtreg`	`re`, `fe`	Die Option `re` steht für Random Effects und `fe` für Fixed Effects.
`statsby`		Speichert Ergebnisse eines Befehls (wie OLS-Regression) für eine Gruppenvariable, die mit `by()` spezifiziert wird.
`egen`	`tag(`*groupvar*`)`	Nur eine Beobachtung pro Gruppe mit einer 1 versehen. Alle anderen Gruppenmitglieder erhalten eine 0.
`predict` *newvar*	`fitted`	Geschätzte Koeffizienten pro Gruppe erzeugen.
`predict` *newvar_steigung newvar_intercept*	`reffects`	Empirical Bayes Prediction der Steigungs- und Interceptkoeffizienten.

Aufgabe

Vorbereitung der Abschlusspräsentation. Präsentation bitte am Montag bis 12h vor der Präsentationan an mich schicken, ihr erhaltet kurzes Feedback zurück. Die Präsentation sollte folgende Struktur enthalten:

- Forschungshypothese
- Überblick über Stand der Forschung mit mindestens 5 Literaturangaben zu Hypothesen, den verwendeten Daten und den Methoden zum Test der Hypothesen
- Informationen zu den Daten für den eigenen Hypothesentest
- Beschreibung der statistischen Methode und des Regressionsmodells für den eigenen Hypothesentest
- offene Fragen

Literatur

Anscombe, Francis J, 1973: Graphs in Statistical Analysis. American Statistician 27: 17–21. 45

Baum, Christopher F., 2006: An introduction to modern econometrics using Stata. Stata Corp. 71

Best, Henning, und *Christof Wolf*, 2010: Logistische Regression. S. 827–854 in: *Christof Wolf* und *Henning Best* (Hg.), Handbuch der Sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften. 103

Cameron, Adrian Colin, und *Pravin K Trivedi*, 2010: Microeconometrics using Stata. Texas: Stata Press. 26, 44, 45, 58, 64, 65

Fox, John, 2008: Applied regression analysis and generalized linear models. Sage Publications, Inc, London. 7, 56, 72, 73, 74, 76

Fox, John, und *Sanford Weisberg*, 2010: An R Companion to Applied Regression. o. O.: Sage Publications, Inc. 133

Hausman, Jerry A, 1978: Specification Tests in Econometrics. Econometrica 46 (6): 1251–1271. 158

Jann, Ben, 2009: Diagnostik von Regressionsschätzungen bei kleinen Stichproben (mit einem Exkurs zu logistischer Regression). Wiesbaden. 58, 60, 61

Jann, Ben, 2005: Making regression tables from stored estimates. Stata Journal 5 (3): 288. 11

Jann, Ben, 2007: Making regression tables simplified. Stata Journal 7 (2): 227–244. 11

Kohler, Ulrich, und *Frauke Kreuter*, 2008: Datenanalyse mit Stata. München: Oldenbourg. 45, 51, 58, 59, 94, 104

Langer, Wolfgang, 2010: Mehrebenenanalyse mit Querschnittsdaten. S. 741–744 in: *Christof Wolf* und *Henning Best* (Hg.), Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften. 136

Long, J Scott, und *Jeremy Freese*, 2006: Regression models for categorical dependent variables using Stata. o. O.: Stata Press. 133

Maloney, Michael P, *Michael P Ward* und *Nicholas G Braucht*, 1975: A Revised Scale for the Measurement of Ecological Attitudes and Knowledge. American Psychologist S. 787–790. 82

Rabe-Hesketh, Sophia, und *Anders Skrondal*, 2008: Multilevel and longitudinal modeling using Stata. Texas. 20, 36, 136, 141, 147, 161

RSU, Rat der Sachverständigen für Umweltfragen, 1978: Umweltgutachten 1978. Bd. 8/1978. Stuttgart: Kohlhammer. 82

Snijders, Tom A. B., und *Roel J. Bosker*, 1999: Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: SAGE Publications. 136, 155

Tutz, Gerhard, 2010: Regression für Zählvariablen. S. 887–904 in: *Christof Wolf* und *Henning Best* (Hg.), Handbuch der Sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften. 112

Wolff, Hans-Georg, und *Johann Bacher*, 2010: Hauptkomponentenanalyse und explorative Faktorenanalyse. S. 333–365 in: *Christof Wolf* und *Henning Best* (Hg.), Hand-

buch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften. 84, 89