LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Kapitel 4:
Das multiple lineare Regressionsmodell

Das multiple lineare Regressionsmodell

Das Modell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

- ε_i ist ein Fehlerterm
- β_0 heißt Regressionskonstante
- Die anderen Regressionskoeffizienten definieren eine p-dimensionale Regressionsebene
- Interpretation: β_j gibt an,
  um wie viel Einheiten sich Y ändert,
  wenn sich X_j um eine Einheit erhöht,
  unter Kontrolle der anderen im Modell enthaltenen X-Variablen
  - Synonym: β_j sagt uns, welcher Effekt verbleibt, wenn wir die anderen X-Variablen konstant halten

OLS Schätzung

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

mit :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Annahmen :

$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ vier Annahmen über die Fehlerverteilung (normalverteilt, im Mittel 0, Homoskedastizität, keine Autokorrelation)

$Cov(\mathbf{x}, \boldsymbol{\varepsilon}) = \mathbf{0}$ Exogenität von X

$rg(\mathbf{X}) = p + 1$ keine linearen Abhängigkeiten (bzw. Multikollinearität)

OLS Schätzer : $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Multiples R^2

- Die vorhergesagten Werte ergeben sich aus

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

- Multiples Bestimmtheitsmaß

$$R^2 = \frac{\text{erklärte Streuung}}{\text{gesamte Streuung}} = \frac{MSS}{TSS} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

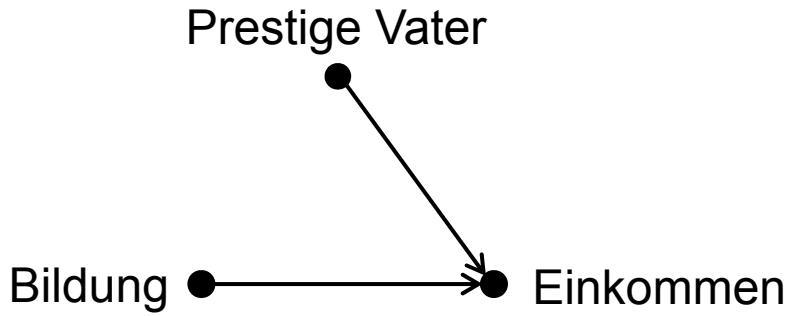
- Es besagt, welcher Anteil der Varianz von Y durch alle Regressoren zusammen erklärt wird
  - Fügt man einen weiteren Regressor hinzu, so ist das Bestimmtheitsmaß des erweiterten Modells mindestens genauso groß wie zuvor
  - Ist allerdings die Erklärungskraft der hinzugefügten Variable, gegeben die bereits im Modell enthaltenen Variablen, gering, so wird sich R^2 nur minimal erhöhen
  - Das Hinzufügen weiterer Variablen verbessert das Modell somit nur, wenn diese Variablen einen eigenständigen Erklärungsbeitrag leisten

Signifikanztests

- Test eines einzelnen Regressionskoeffizienten
  - Nullhypothese: X_j hat keinen Einfluss auf Y (kein Zusammenhang)
    $$H_0: \beta_j = 0$$
  - Die Teststatistik (t-Wert) ist
    $$T = \frac{\hat{\beta}_j}{\hat{\sigma}_j} \sim t(n-p-1)$$
  - Die H_0 wird abgelehnt, falls $|T| > t_{1-\alpha/2}(n-p-1)$
    - Ab n>30 das $z_{1-\alpha/2}$ Quantil (Faustregel für α=5%: |T| > 2)
- Test des gesamten Modells: overall F-Test
  - Nullhypothese: keine X-Variable hat einen Einfluss auf Y
    $$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
  - Die Teststatistik (F-Wert) ist
    $$F = \frac{\text{MSS}/p}{\text{RSS}/(n-p-1)} \sim F(p, n-p-1)$$
  - Die H_0 wird verworfen, falls: $F > F_{1-\alpha}(p,n-p-1)$

Beispiel: Statuszuweisungsmodell

- Blau/Duncan (1967) "The American Occupational Structure"
  - Wie erlangt man seine soziale Position?
    Durch „achievement“ oder Statusvererbung?
  - ALLBUS 2002:
    - Abhängige Variable: monatliches Netto-Einkommen in Euro
      (nur Westdeutsche, Vollzeit)
    - Status des Vaters: Magnitude-Prestigeskala (Werte von 20-187)
    - „Achievement“: eigene Schul- und Berufsbildung (Werte von 8-23,5)

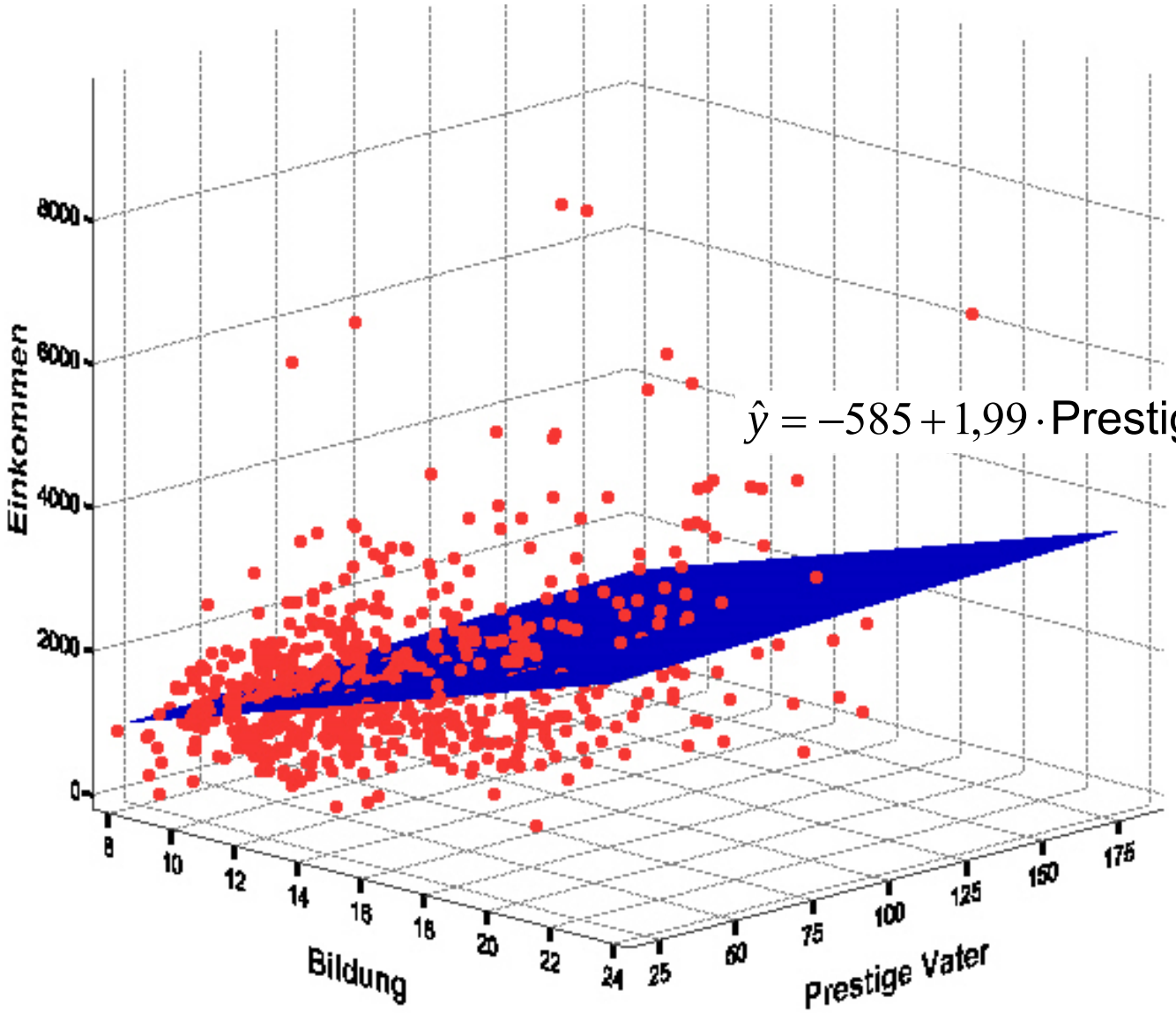


Beispiel: Stata-Output

```
. regress    eink prestv bild

      Source |       SS       df       MS              Number of obs =     670
-------------+------------------------------           F(  2,   667) =   64.99
       Model |   215244302     2   107622151           Prob > F      =  0.0000
    Residual |  1.1045e+09   667  1655904.67           R-squared     =  0.1631
-------------+------------------------------           Adj R-squared =  0.1606
       Total |  1.3197e+09   669  1972694.65           Root MSE      =  1286.8

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prestv |    1.99162   2.004457     0.99   0.321    -1.944187    5.927426
        bild |   180.2345   17.73502    10.16   0.000     145.4113    215.0577
       _cons |  -585.4664   224.3659    -2.61   0.009    -1026.015   -144.9179
------------------------------------------------------------------------------
```

Daten: ALLBUS 2002
Do-File: 2 LinReg Modell.do

Beispiel: Regressionsebene



$$\hat{y} = -585 + 1{,}99 \cdot \text{Prestige} + 180 \cdot \text{Bildung}$$

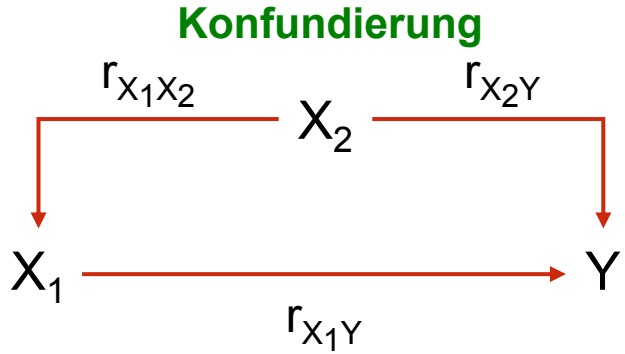
Was bedeutet „unter Kontrolle“?

- β_j ist der Effekt von X_j unter Kontrolle der anderen im Modell enthaltenen X-Variablen
  - Man sagt auch: „unter Konstanthaltung“ der anderen im Modell enthaltenen X-Variablen
- Der bivariate Effekt wird von Konfundierungen „bereinigt“

bivariater Effekt **Konfundierung**

$$\hat{\beta}_1^* = \frac{r_{X_1Y} - r_{X_1X_2} r_{X_2Y}}{1 - r_{X_1X_2}^2}$$

Standardisierung



- Spezialfall: X_1 und X_2 sind nicht korreliert (Multikausalität)

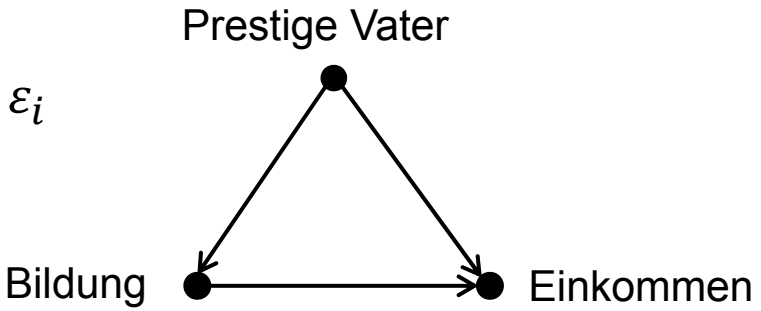
$$\hat{\beta}_1^* = r_{X_1Y}$$

  - Der multiple Regressionskoeffizient ist gleich dem bivariaten

Was bedeutet „unter Kontrolle“?

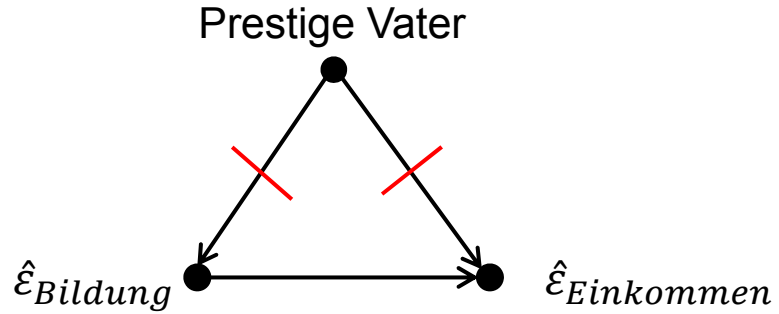
- Das multiple Modell mit einem Confounder (C)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 c_i + \varepsilon_i$$



- Der Effekt des Confounders (C) wird „herauspartialisiert“
  - $y_i = \alpha + \beta c_i + \varepsilon_{iY}$ → $\widehat{\varepsilon_{iY}} = y_i - \hat{\alpha} - \hat{\beta} c_i$
    - Die auf C beruhende Variation ist herausgerechnet: $Cov(\widehat{\varepsilon_{iY}}, c_i) = 0$
  - $x_i = \gamma + \delta c_i + \varepsilon_{iX}$ → $\widehat{\varepsilon_{iX}} = x_i - \hat{\gamma} - \hat{\delta} c_i$
    - Die auf C beruhende Variation ist herausgerechnet: $Cov(\widehat{\varepsilon_{iX}}, c_i) = 0$
- Die Regression mit den Residuen liefert ebenfalls $\widehat{\beta_1}$

$$\widehat{\varepsilon_{iY}} = \beta_1 \widehat{\varepsilon_{iX}} + \varepsilon_i$$



Beispiel: Kontrolle der Herkunft

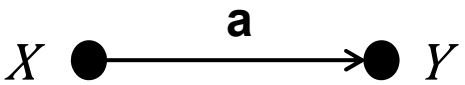
	(1) Bivariate Regression	(2) Multiple Regression	(3) Residuen Regression
bild	186.82***	180.23***	180.23***
	(16.45)	(17.74)	(17.72)
prestv		1.99	
		(2.00)	
_cons	-555.64*	-585.47**	0.00
	(222.35)	(224.37)	(49.68)
N	670	670	670
R-sq	0.162	0.163	0.134

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

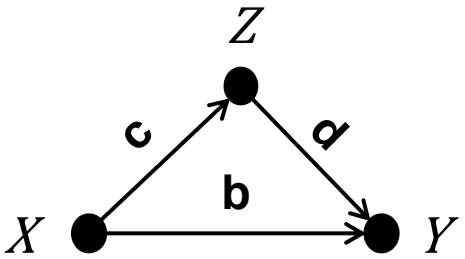
Daten: ALLBUS 2002
Do-File: 2 LinReg Modell.do

Regel I: Welche Variablen kontrollieren?

- Oft wird ein Standard-Set an „Kontrollvariablen“ verwendet
  - Geschlecht, West/Ost, Alter, Herkunft, Beruf, Bildung, …
  - Die „kontrolliere Alles auf Verdacht“ Strategie
  - Das ist keine sinnvolle Vorgehensweise!
- Regel: um den (totalen) Kausaleffekt zu identifizieren, kontrolliere nur (!) Confounder
- Wenn man Mediatoren mitkontrolliert: overcontrol-bias
  - Es wird nicht der totale, sondern nur der direkte Effekt geschätzt



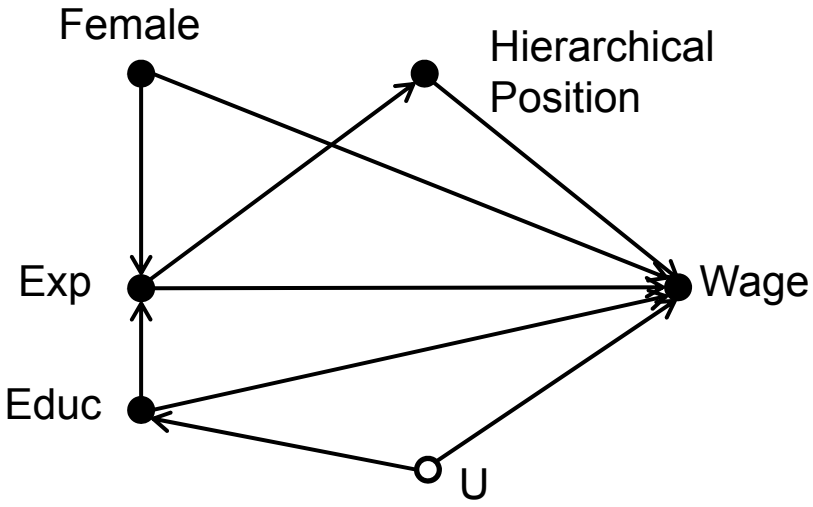
$$y_i = \beta_0 + \boldsymbol{a} x_i + \varepsilon_i$$



$$y_i = \beta_0 + \boldsymbol{b} x_i + d z_i + \varepsilon_i$$

Regel I: Welche Variablen kontrollieren?

- Ziel: Einen (totalen) Kausaleffekt zu identifizieren
  1. Ein theoretisches Modell (um den Kausaleffekt) entwickeln
  2. Daraus ein maßgeschneidertes Regressionsmodell ableiten
- Beispiel: Kausaleffekt der Berufserfahrung auf Lohn



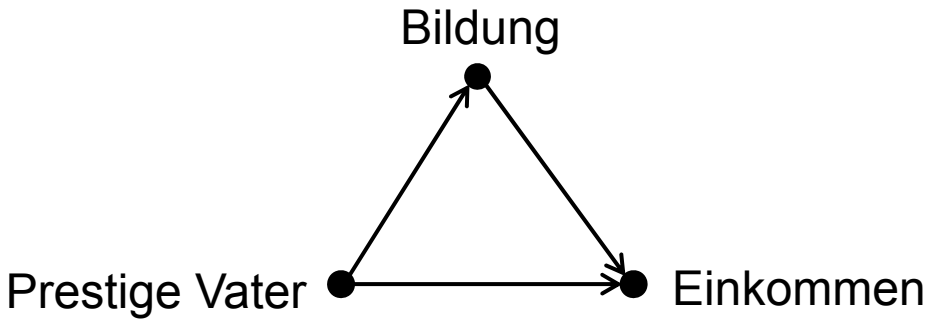
- Maßgeschneiderte Regression

$$\ln(\text{Wage}) = \alpha + \beta\text{Exp} + \gamma\text{Educ} + \delta\text{Female}$$

  - γ und δ sind keine (totalen) Kausaleffekte!

Regel II: Mediationsanalyse

- Wenn man einen Kausaleffekt gefunden hat, kann man im nächsten Schritt fragen: Was ist der Mechanismus?
  - Statistisch ist dies die Frage nach der Mediation: Welche Mediatorvariable(n) kann den Effekt erklären?
- Bsp. Statuszuweisung: Kann die Bildung den Effekt der Herkunft erklären?



- Vorgehen: man fügt der Regression die Mediatorvariable(n) hinzu
  - Reduziert sich der Kausaleffekt (signifikant)?

Beispiel: Statuszuweisungsmodell

	(1)	(2)
prestv	9.61***	1.99
	(2.00)	(2.00)
bild		180.23***
		(17.74)
_cons	1343.23***	-585.47**
	(128.51)	(224.37)
N	670	670
R-sq	0.034	0.163

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Der signifikante Kausaleffekt von „Prestige Vater“ wird unter Kontrolle des Mediators „Bildung“ deutlich kleiner und ist nicht mehr signifikant.
Oft wird in so einem Fall von „signifikanter Mediation“ gesprochen.
Das ist aber voreilig, denn dafür braucht es einen Signifikanztest für den indirekten Effekt. Diesen liefert der Sobel-Test.
In unserem Fall werden 79% des totalen Effektes signifikant mediiert.

Sobel-Goodman Mediation Tests

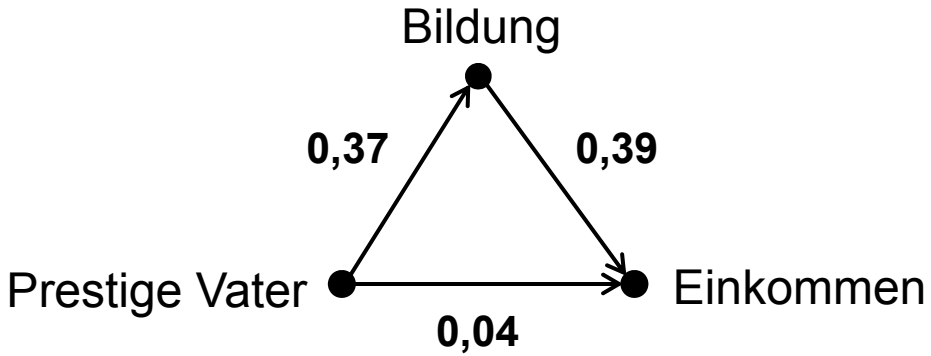
	Coef	Std Err	Z	P>|Z|
Indirect effect =	7.61593	1.04689	7.27483	3.5e-13
Direct effect =	1.99162	2.00446	.993595	.32042
Total effect =	9.60755	1.99636	4.81255	1.5e-06

Proportion of total effect that is mediated: .79270271

Daten: ALLBUS 2002
Do-File: 2 LinReg Modell.do

Beispiel: Statuszuweisung

- Das gesamte Kausalmodell mit den standardisierten Regressionskoeffizienten
  - Origin, Education, Destination: OED-Dreieck
  - Der direkte Herkunftseffekt ist schwach (und nicht signifikant)
  - Es gibt einen starken indirekten Effekt über Bildung

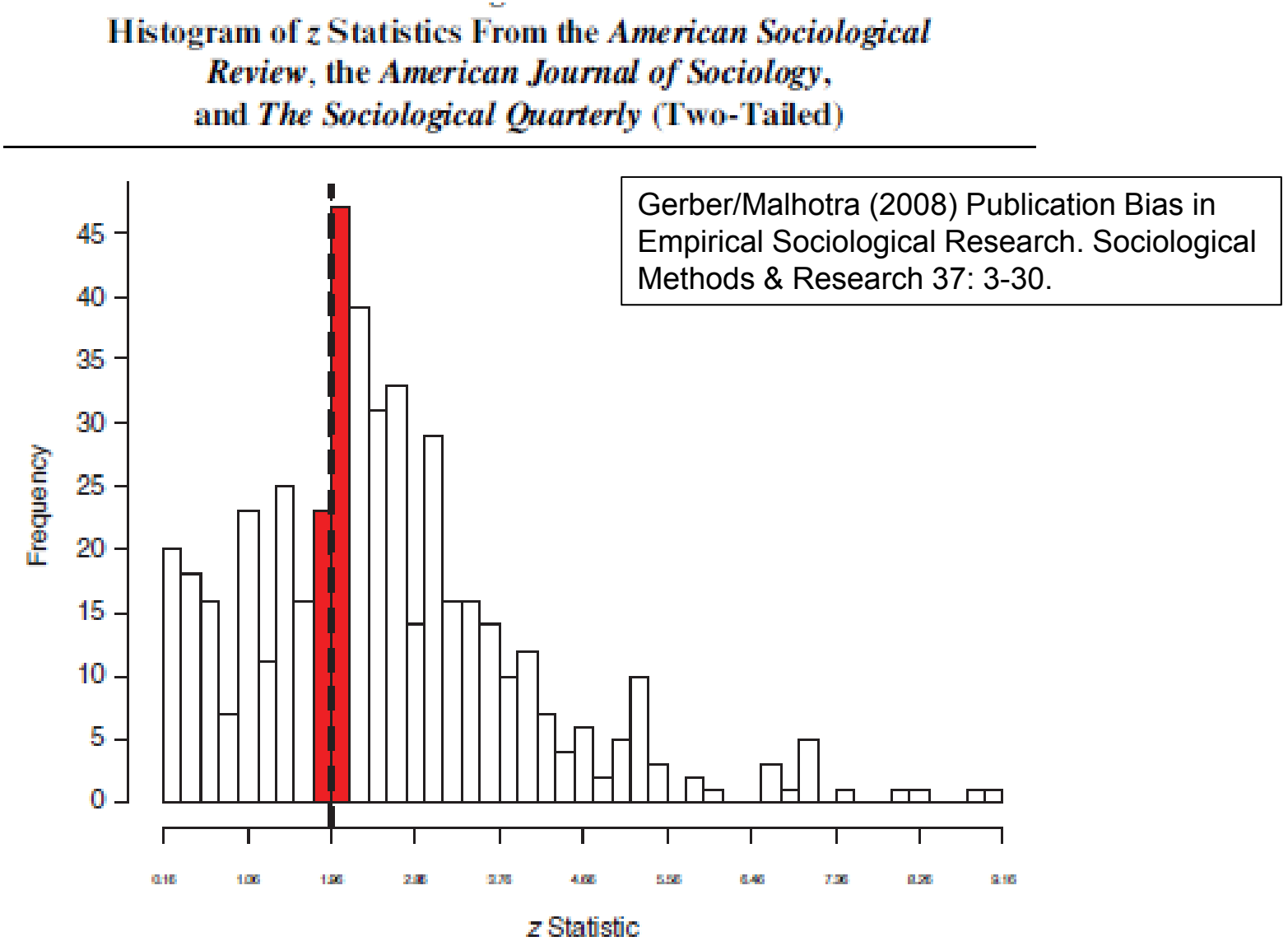


Daten: ALLBUS 2002
Do-File: 2 LinReg Modell.do

Das Signifikanztest-Ritual

- Unsinnige Anwendung der Signifikanztests in der Praxis
  - Multiples Testen und Publikations-Bias
    - Auch wenn in Wirklichkeit kein relevanter Effekt vorliegt, so werden von 100 Forschern irrtümlicherweise 5 einen „signifikanten“ Effekt finden und genau diese 5 „signifikanten“ Effekte werden publiziert
    - Analog: Variablen-/Modellselektion anhand von t-Tests
    - Nicht-signifikante Ergebnisse werden nicht publiziert (s. nächste Folie)
    - Folge: viele publizierte Ergebnisse sind zufällig zustande gekommen (also falsch, obwohl sie „signifikant“ sind)
  - Viele Forscher schauen nur noch auf die „Sternchen“
    - Aber: „Signifikanz ist nicht gleich Relevanz“
- Das Signifikanztestritual hat in den Sozialwissenschaften eine Menge an unsinnigen Ergebnissen produziert. Die Jagd nach Signifikanzen hat die Jagd nach der Realität verdrängt. Signifikanztests sollten deshalb abgeschafft werden!
  (Ziliak/McCloskey, 2008)

Publikations-Bias



Gerber/Malhotra (2008) Publication Bias in Empirical Sociological Research. Sociological Methods & Research 37: 3-30.

Regel III: Welches Signifikanzniveau?

- Statt Abschaffung, Verbesserung der Praxis (s.a. Krämer 2011)
  - Kein „Sternchenquetschen“: kein 10%-Signifikanzniveau
  - Achte mehr auf Effektstärke (und Effektrichtung!)
  - Keine Variablenselektion anhand von Signifikanztests
  - Auch nicht-signifikante Ergebnisse sind wichtig!
  - Statt R^2-Zentrierung hin zur X-Zentrierung der Sozialforschung
    - Konzentriere dich auf einen Effekt und versuche den mittels harter Spezifikationstests zu widerlegen (Falsifikationismus!)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Kapitel 5:
Interpretation von Regressionskoeffizienten

Interpretation von Regressionskoeffizienten

- Das multiple Regressionsmodell (ohne Personenindex i)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Daraus ergibt sich der bedingte Erwartungswert

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Verschiedene Interpretationsmöglichkeiten
  - Marginaleffekt (marginal effect, **ME**)

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j$$

  - Effekt der Veränderung von X_j um eine Einheit (discrete change, **DC**)

$$\frac{\Delta E(y|\boldsymbol{x})}{\Delta x_j} = E\left(y\middle|\boldsymbol{x}, x_j + 1\right) - E(y|\boldsymbol{x}, x_j) = \beta_j$$

  - Fazit: Im linearen Modell sind DC und ME identisch gleich dem Regressionskoeffizienten!

Graphische Präsentation von Regressionsergebnissen

- Anstatt einer Tabelle mit vielen Zahlen: Regressionsgraphen
  (Bauer, 2015)
- Im Prinzip drei Typen:
  1. Plotten der Marginaleffekte von X: (Effektplot) **Koeffizientenplot**
     - ME für metrische Variablen, DC für Dummies
  2. Plotten der vorhergesagten Werte von Y: **Profile-Plot**
     - Welcher Wert von Y ergibt sich für verschiedene X-Werte?
     - Für alle Fälle in den Daten werden die beobachteten Werte eingesetzt, $\widehat{y_i}$ berechnet und dann gemittelt (predictive margins)
  3. Plotten der Effekte von X gegeben Z: **konditionaler Effektplot**
     - Wie verändert sich der ME von X mit Z?
     - Hilfreich für Interaktionen

Präsentation der Regressionskoeffizienten

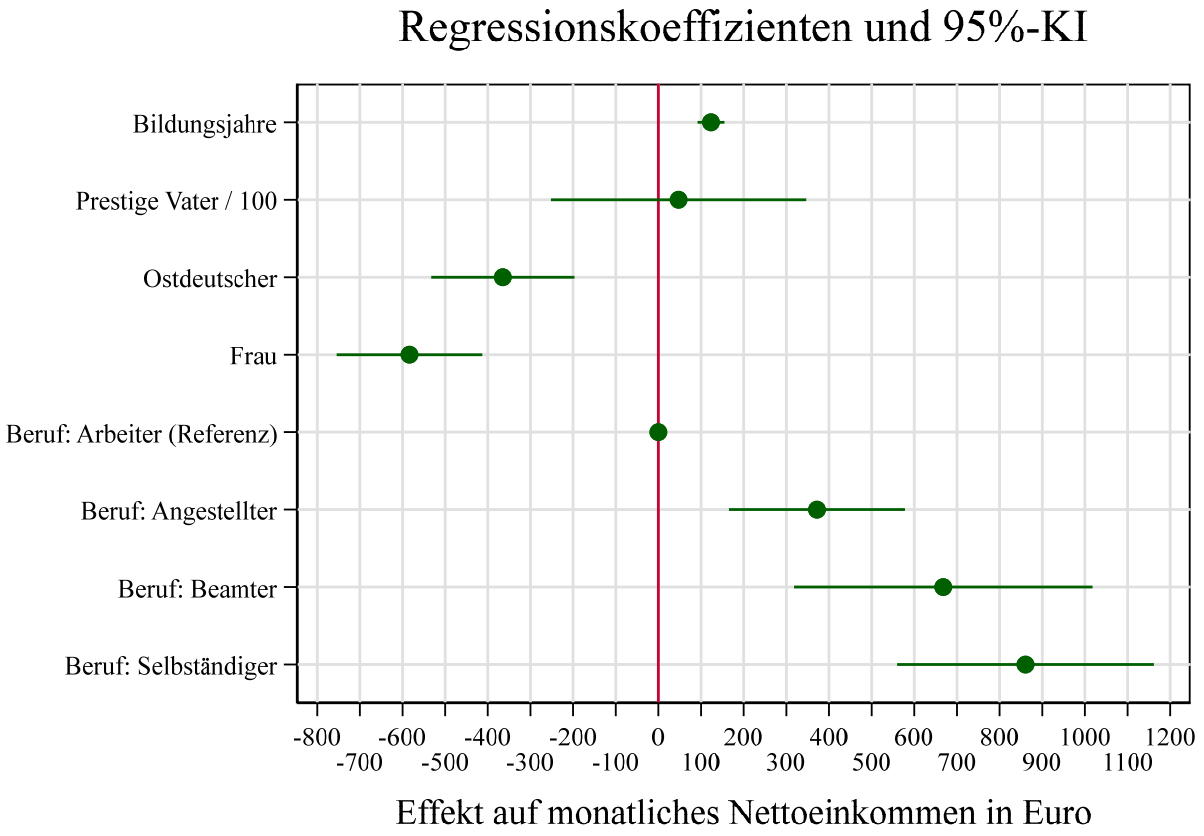
- Bsp. Einkommensregression
  - Monatliches Nettoeinkommen in Euro (nur Vollzeiterwerbstätige)
  - Bildungsjahre, Prestige Vater/100, Ostdeutscher, Frau, berufliche Stellung

```
regress eink bild prestv ost frau i.beruf
esttab using "RegTabelle.rtf", r2 b(%6.1f)
```

```
coefplot, drop(_cons) xline(0) base
```

Bildungsjahre	123.3***
	(7.67)
Prestige Vater / 100	47.4
	(0.31)
Ostdeutscher	-364.7***
	(-4.26)
Frau	-583.5***
	(-6.71)
Ber.: Arbeiter (Ref.)	
Ber.: Angestellter	371.8***
	(3.54)
Ber.: Beamter	667.9***
	(3.75)
Ber.: Selbständiger	860.7***
	(5.62)
Konstante	163.2
N	948
R^2	0.214

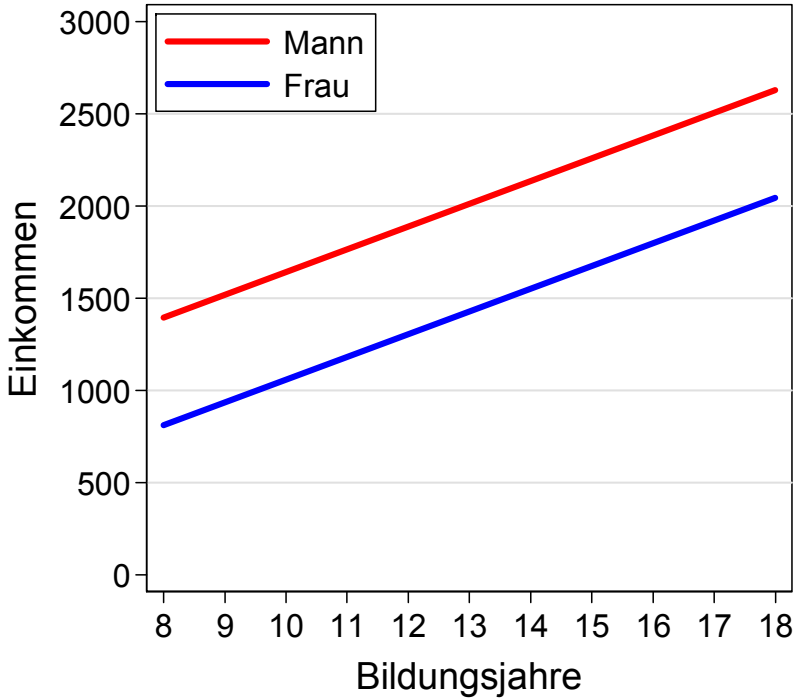
t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



Daten: ALLBUS 2002
Do-File: 3 LinReg Interpretation.do

Profile-Plot

- Hilfreich zur Veranschaulichung von Regressionskoeffizienten
  - Man plottet die vorhergesagten Werte der Outcome-Variable
    - Z.B. die geschätzte Regressionsgerade für eine metrische Variable
    - Evtl. für verschiedene inhaltlich interessierende Gruppen
  - Predictive Margins
    - Für jede Beobachtung wird mit ihren Kovariatenwerten ein Vorhersagewert berechnet
    - Nur die „marginvars“ werden auf fixierte Werte gesetzt
    - Anschließend wird über alle Vorhersagewerte gemittelt
    - Dies ist eine Art „kontrafaktisches“ Vorgehen!



Veranschaulichung des Bildungs- und des Geschlechtseffektes

„frau“ und „bild“ sind hier die marginvars

```
margins frau, at(bild=(8 18))
marginsplot,   noci
```

$$\hat{y}_M = 409 + 123 \times \text{Bild}$$
$$\hat{y}_F = 409 + 123 \times \text{Bild} - 584$$

Daten: ALLBUS 2002
Do-File: 3 LinReg Interpretation.do

Interpretation einer Polynomregression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \text{Exp} + \beta_3 \text{Exp}^2 + \epsilon$$

Daten: ALLBUS 2002
Do-File: 3 LinReg Interpretation.do

$$\frac{\partial E(y)}{\partial \text{Exp}} = \beta_2 + 2 \times \beta_3 \text{Exp}, \qquad \text{Exp}_{\text{max/min}} = -\frac{\beta_2}{2 \times \beta_3}$$

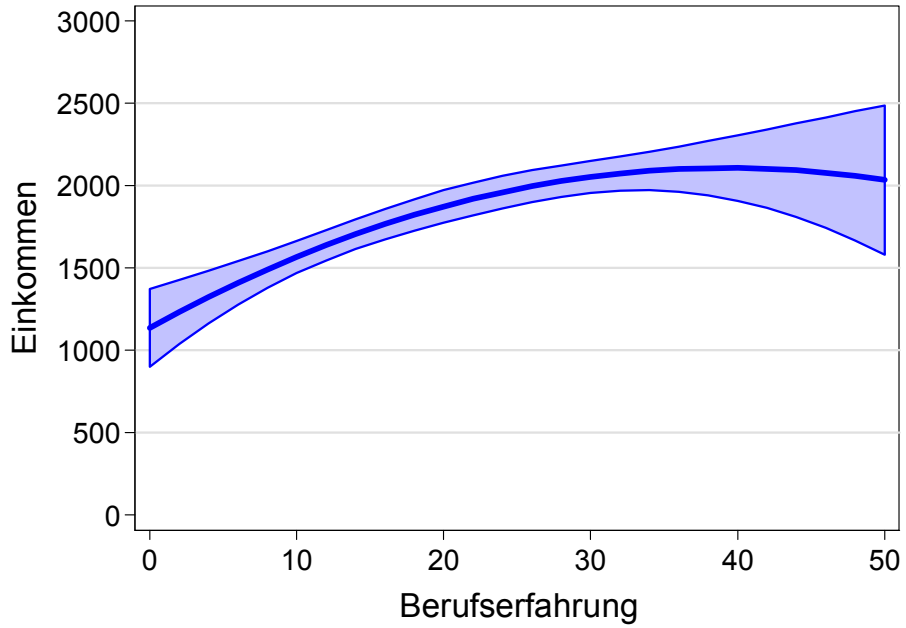
```
. regress    eink bild ost frau c.exp c.exp#c.exp

      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  5,  1112) =   65.83
       Model |   474909648     5  94981929.7           Prob > F      =  0.0000
    Residual |  1.6045e+09  1112  1442907.24           R-squared     =  0.2284
-------------+------------------------------           Adj R-squared =  0.2249
       Total |  2.0794e+09  1117  1861613.69           Root MSE      =  1201.2

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |      180.95      12.36    14.64   0.000       156.69      205.20
         ost |    -448.92       77.23    -5.81   0.000      -600.45     -297.39
        frau |    -435.94       75.63    -5.76   0.000      -584.33     -287.55
         exp |      49.53       12.57     3.94   0.000        24.88       74.19
 c.exp#c.exp |      -0.63        0.29    -2.19   0.029        -1.20       -0.07
       _cons |    -971.85     202.76    -4.79   0.000     -1369.68     -574.01
------------------------------------------------------------------------------
```

Interpretation einer Polynomregression

```
. margins, dydx(exp) at(exp=(0 10 20 30 40 50))
------------------------------------------------
    |            Delta-method
    |      dy/dx   Std. Err.      z    P>|z|
----+-------------------------------------------
 0  |      49.53      12.57     3.94   0.000
10  |      36.91       7.18     5.14   0.000
20  |      24.28       3.38     7.18   0.000
30  |      11.66       6.13     1.90   0.057
40  |      -0.97      11.41    -0.08   0.932
50  |     -13.59      17.00    -0.80   0.424
------------------------------------------------
```

```
. test exp exp#exp

( 1)  exp = 0
( 2)  c.exp#c.exp = 0

 F(  2,  1112) =   26.21
      Prob > F =    0.0000
```

$$\text{Exp}_{\text{max}} = -\frac{49{,}53}{2 \times -0{,}63} = 39{,}3$$



Profile-Plot

```
margins, at(exp=(0(2)50))
marginsplot
```

Daten: ALLBUS 2002
Do-File: 3 LinReg Interpretation.do

Das semi-logarithmische Regressionsmodell

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\leftrightarrow y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon)$$

$$\frac{\partial E(y)}{\partial x_1} = E(y)\beta_1$$

Marginaleffekt

$$\frac{\Delta E(y)}{\Delta x_1} = E(y)\left(e^{\beta_1} - 1\right) \quad \Rightarrow \quad \frac{\frac{\Delta E(y)}{\Delta x_1}}{\mathrm{E(y)}} = e^{\beta_1} - 1$$

Discrete Change prozentuale Veränderung

$(e^{\beta_1} - 1) \times 100$ ist die prozentuale Veränderung von Y bei Erhöhung von X um eine Einheit.

$\beta_1 \times 100$ ist eine gute Näherung, falls $|\beta_1| < 0{,}1$.

Das semi-logarithmische Regressionsmodell

```
. regress  lneink bild ost frau c.exp c.exp#c.exp

      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  5,  1112) =  145.12
       Model |  138.195766     5  27.6391532           Prob > F      =  0.0000
    Residual |   211.79019  1112  .190458804           R-squared     =  0.3949
-------------+------------------------------           Adj R-squared =  0.3921
       Total |  349.985956  1117  .313326729           Root MSE      =  .43642

------------------------------------------------------------------------------
      lneink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |      0.0863     0.0045    19.21   0.000       0.0775      0.0951
         ost |     -0.2496     0.0281    -8.90   0.000      -0.3047     -0.1946
        frau |     -0.2317     0.0275    -8.43   0.000      -0.2856     -0.1777
         exp |      0.0431     0.0046     9.43   0.000       0.0341      0.0520
 c.exp#c.exp |     -0.0006     0.0001    -6.13   0.000      -0.0008     -0.0004
       _cons |      5.8124     0.0737    78.90   0.000       5.6678      5.9569
------------------------------------------------------------------------------
```

Die exakten „Discrete Change“ %-Effekte

- Bildungsrendite: +9,1%
- Ostdeutscher: -22,1%
- Frau: -20,7%

Daten: ALLBUS 2002
Do-File: 3 LinReg Interpretation.do

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Kapitel 6:
Regression mit Dummies

Regression mit Dummy

Einkommen in Abhängigkeit von Bildung und Geschlecht
Dummy-Kodierung: 0 = Mann, 1 = Frau

```
. regress eink bild frau

      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  2,  1115) =  116.00
       Model |   358146421     2   179073211           Prob > F      =  0.0000
    Residual |  1.7213e+09  1115  1543745.36           R-squared     =  0.1722
-------------+------------------------------           Adj R-squared =  0.1707
       Total |  2.0794e+09  1117  1861613.69           Root MSE      =  1242.5

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |   169.2551   12.44852    13.60   0.000     144.8299    193.6803
        frau |  -511.2477   77.71202    -6.58   0.000    -663.7259   -358.7694
       _cons |  -264.8927   172.6304    -1.53   0.125    -603.6097    73.82433
------------------------------------------------------------------------------
```

Interpretation: Bei gleicher Bildung verdienen Frauen im Schnitt 511 Euro weniger als Männer.

Daten: ALLBUS 2002
Do-File: 4 LinReg Dummies.do

Kategoriale uV mit mehr als 2 Ausprägungen

- Auch kategoriale Variablen sind in der multiplen Regression möglich
  - Aber erst nach Dummy-Bildung!
  - Grundprinzip: Für jede Ausprägung wird eine Dummy gebildet
  - In die Regression werden alle Dummies außer einer (Referenzkategorie) aufgenommen
- Beispiel: berufliche Stellung

berufliche Stellung	D1	D2	D3	D4
Arbeiter	1	0	0	0
Angestellter	0	1	0	0
Beamter	0	0	1	0
Selbständiger	0	0	0	1

  - Werden keine weiteren unabhängigen Variablen berücksichtigt, so entspricht die Konstante β_0 dem Mittelwert der Referenzkategorie
  - Die Koeffizienten β_j der Dummies geben den Mittelwertunterschied der betreffenden Kategorie zur Referenzkategorie an

Generierung der Dummies

```
. tabulate beruf, gen(d)

       beruf |      Freq.     Percent        Cum.
-------------+-----------------------------------
    Arbeiter |        360       29.22       29.22
 Angestellter |        622       50.49       79.71
     Beamter |         90        7.31       87.01
Selbständiger |        160       12.99      100.00
-------------+-----------------------------------
       Total |      1,232      100.00

. tabulate beruf d1
               |    beruf==Arbeiter
       beruf |         0          1 |     Total
-------------+----------------------+----------
    Arbeiter |         0        360 |       360
 Angestellter |       622          0 |       622
     Beamter |        90          0 |        90
Selbständiger |       160          0 |       160
-------------+----------------------+----------
       Total |       872        360 |     1,232
```

Daten: ALLBUS 2002
Do-File: 4 LinReg Dummies.do

Interpretation der Dummies

```
. table beruf, contents(sum d1 sum d2 sum d3 sum d4)

----------------------------------------------------------------
        beruf |    sum(d1)     sum(d2)     sum(d3)     sum(d4)
--------------+-------------------------------------------------
     Arbeiter |         360           0           0           0
 Angestellter |           0         622           0           0
      Beamter |           0           0          90           0
Selbständiger |           0           0           0         160
----------------------------------------------------------------
```

```
. table beruf, contents(mean eink)

--------------------------
        beruf | mean(eink)
--------------+-----------
     Arbeiter |   1332.902
 Angestellter |   1894.345
      Beamter |   2480.987
Selbständiger |   2714.033
--------------------------
```

```
. regr eink d2 d3 d4

-----------------------------
        eink |      Coef.
-------------+---------------
          d2 |   561.4422
          d3 |   1148.084
          d4 |   1381.131
       _cons |   1332.903
-----------------------------
```

Daten: ALLBUS 2002
Do-File: 4 LinReg Dummies.do

Regression mit kategorialer uV

```
. regress eink bild i.beruf

      Source |       SS       df       MS              Number of obs =    1072
-------------+------------------------------           F(  4,  1067) =   52.89
       Model |   330638913     4  82659728.2            Prob > F      =  0.0000
    Residual |  1.6674e+09  1067  1562726.39            R-squared     =  0.1655
-------------+------------------------------           Adj R-squared =  0.1624
       Total |  1.9981e+09  1071  1865609.69            Root MSE      =  1250.1

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------------------------
        bild |   130.5443   14.63054     8.92   0.000
             |
       beruf |
          2  |   216.6888   95.98289     2.26   0.024
          3  |   562.1304   171.8194     3.27   0.001
          4  |   919.5047   143.5743     6.40   0.000
             |
       _cons |  -141.8791   179.5353    -0.79   0.430
------------------------------------------------------------------------------
```

Bivariate Effekte

bild	165
angest	561
beamt	1148
selbst	1381

`i.beruf` sagt Stata, dass „beruf“ eine Indikatorvariable ist. Stata bildet dann die „virtuellen“ Dummies „2.beruf“, „3.beruf“ und „4.beruf“. Referenzkategorie ist automatisch der kleinste Wert 1=Arbeiter.

Daten: ALLBUS 2002
Do-File: 4 LinReg Dummies.do

Regression mit kategorialer uV

```
. testparm i.beruf

 ( 1)  2.beruf = 0
 ( 2)  3.beruf = 0
 ( 3)  4.beruf = 0

       F(  3,  1067) =   15.01
            Prob > F =    0.0000
```

Signifikanz des Berufs?

```
. regress eink bild ib3.beruf
```

Beamte als Referenzkategorie

eink	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
bild	130.5443	14.63054	8.92	0.000	101.8364	159.2522
beruf						
1	-562.1304	171.8194	-3.27	0.001	-899.2727	-224.9881
2	-345.4416	154.3745	-2.24	0.025	-648.3538	-42.52949
4	357.3743	183.0629	1.95	0.051	-1.829857	716.5784
_cons	420.2513	271.3571	1.55	0.122	-112.2028	952.7055

Daten: ALLBUS 2002
Do-File: 4 LinReg Dummies.do

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Kapitel 7:
Interaktionseffekte

Berücksichtigung von Interaktionseffekten

- Der Effekt von X hängt vom Wert von Z ab
  - Z nennt man auch Moderator-Variable (Interaktion = Moderation)
- Berücksichtigung in einer Regression
  - Nimm die Hauptterme in die Regression (X und Z)
  - Füge einen Interaktionsterm hinzu
    - Das Produkt von X und Z (deshalb auch „Produktterm“)
    - Hierbei unterstellt man eine multiplikative Interaktion

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 (x * z) + \epsilon$$

Der (konditionale) Marginaleffekt von X

$$\frac{\partial E(y)}{\partial x} = \beta_1 + \beta_3 z$$

$$z = 0\text{: } \frac{\partial E(y)}{\partial x} = \beta_1$$

$$z = 1\text{: } \frac{\partial E(y)}{\partial x} = \beta_1 + \beta_3$$

Der (konditionale) Marginaleffekt von Z

$$\frac{\partial E(y)}{\partial z} = \beta_2 + \beta_3 x$$

$$x = 0\text{: } \frac{\partial E(y)}{\partial z} = \beta_2$$

$$x = 1\text{: } \frac{\partial E(y)}{\partial z} = \beta_2 + \beta_3$$

Der Interaktionseffekt

$$\frac{\partial^2 E(y)}{\partial x \, \partial z} = \beta_3$$

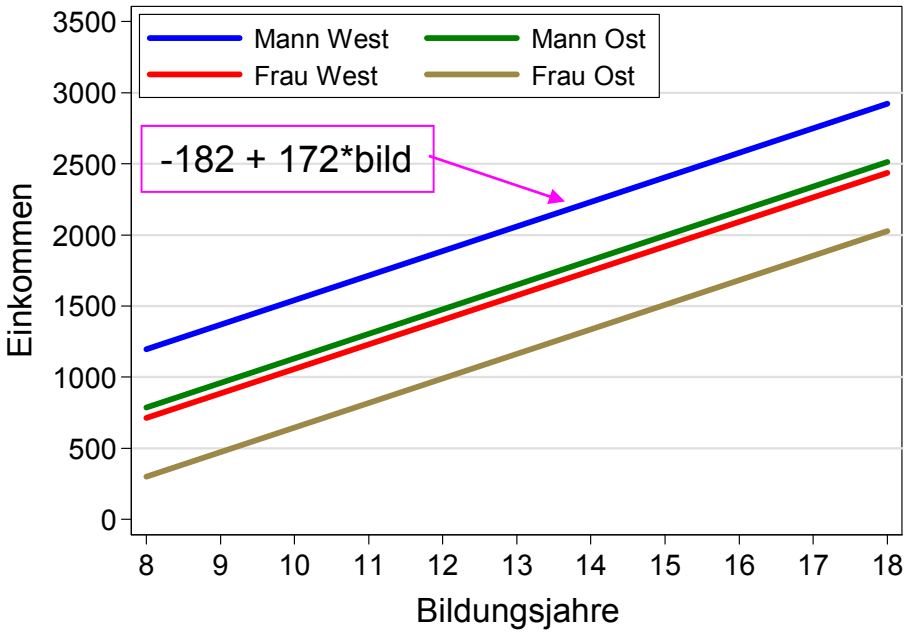
Dummy-Interaktion: Geschlecht/Wohnort

```
* Ohne Interaktion
. regress eink bild frau ost
------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |   172.4277   12.31945    14.00   0.000     148.2557    196.5996
        frau |  -484.5974   76.98204    -6.29   0.000    -635.6436   -333.5513
         ost |  -410.0585   78.53404    -5.22   0.000    -564.1498   -255.9672
       _cons |  -182.3318   171.3636    -1.06   0.288    -518.5635       153.9
------------------------------------------------------------------------------

. * Mit Interaktion
. regress eink bild i.frau i.ost frau#ost
                                                Number of obs =    1118
                                                R-squared     =  0.1947

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |   171.3628    12.3170    13.91   0.000     147.1957    195.5299
        frau |  -592.1200    95.0501    -6.23   0.000    -778.6176   -405.6225
         ost |  -526.8585    99.1837    -5.31   0.000    -721.4665   -332.2504
    frau#ost |   311.5265   161.9030     1.92   0.055      -6.1431    629.1960
       _cons |  -132.5139   173.1033    -0.77   0.444    -472.1594    207.1317
------------------------------------------------------------------------------
```

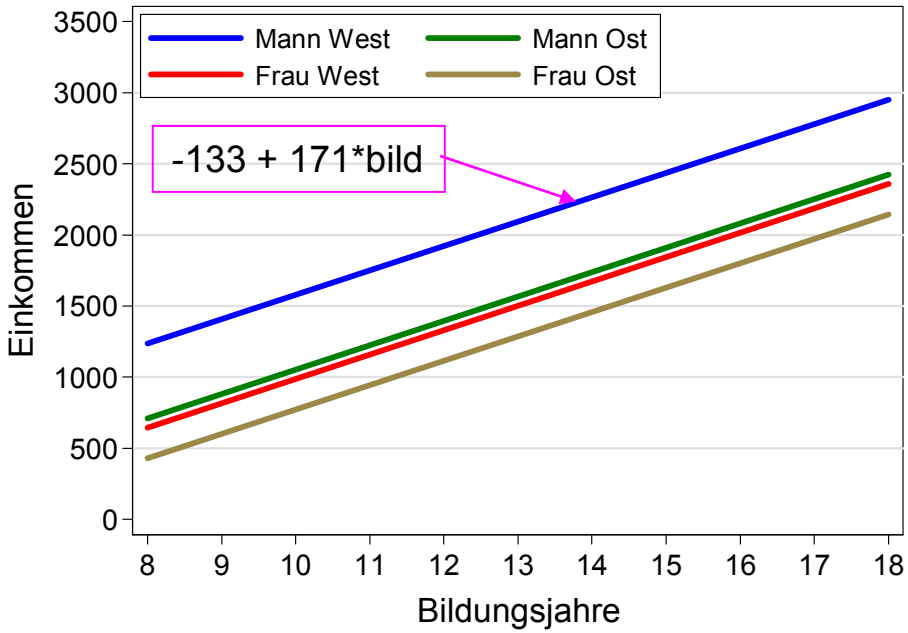
Daten: ALLBUS 2002
Do-File: 5 LinReg Interaktion.do

Dummy-Interaktion: Geschlecht/Wohnort

ohne Interaktion



mit Interaktion



Designmatrix	Frau -592	Ost -527	Ofrau 312	Einkommens- unterschied
Mann West	0	0	0	0
Mann Ost	0	1	0	-527
Frau West	1	0	0	-592
Frau Ost	1	1	1	-807

Referenzgruppe

Daten: ALLBUS 2002
Do-File: 5 LinReg Interaktion.do

Slope-Interaktion: Geschlecht/Bildung

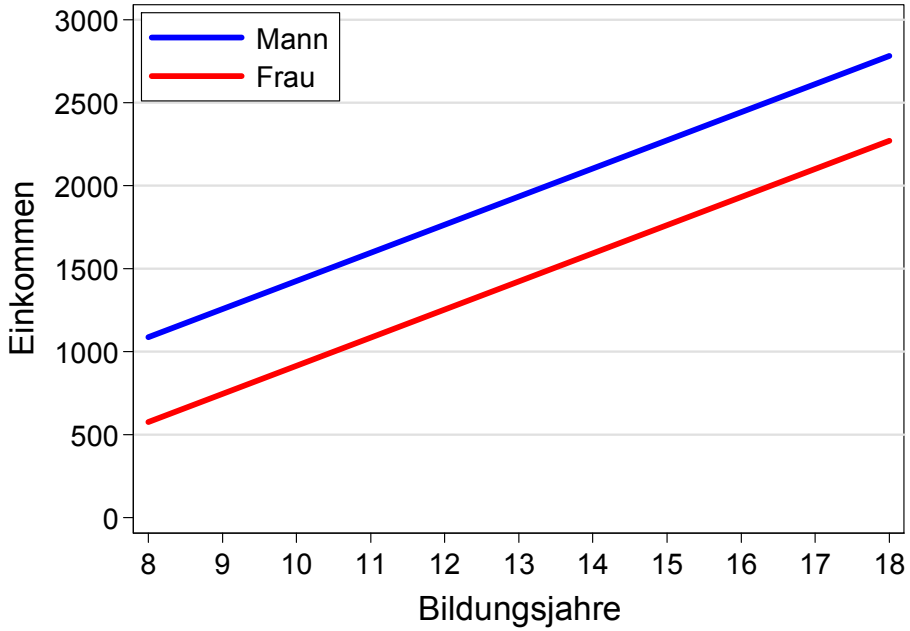
```
. * Ohne Interaktion
. regress eink c.bild i.frau

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |   169.2551   12.44852    13.60   0.000     144.8299    193.6803
        frau |  -511.2477   77.71202    -6.58   0.000    -663.7259   -358.7694
       _cons |  -264.8927   172.6304    -1.53   0.125    -603.6097    73.82433
------------------------------------------------------------------------------

. * Mit Interaktion
. regress eink c.bild i.frau   i.frau#c.bild

                                                Number of obs =    1118
                                                R-squared     =  0.1765

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |   190.2989   15.17511    12.54   0.000     160.5239    220.0739
        frau |   335.2419   359.1157     0.93   0.351    -369.3776    1039.861
 frau#c.bild |  -63.77517   26.41776    -2.41   0.016    -115.6094   -11.94099
       _cons |  -546.0541   207.9355    -2.63   0.009    -954.0433   -138.0648
------------------------------------------------------------------------------
```

Frauen verdienen mehr als Männer!?

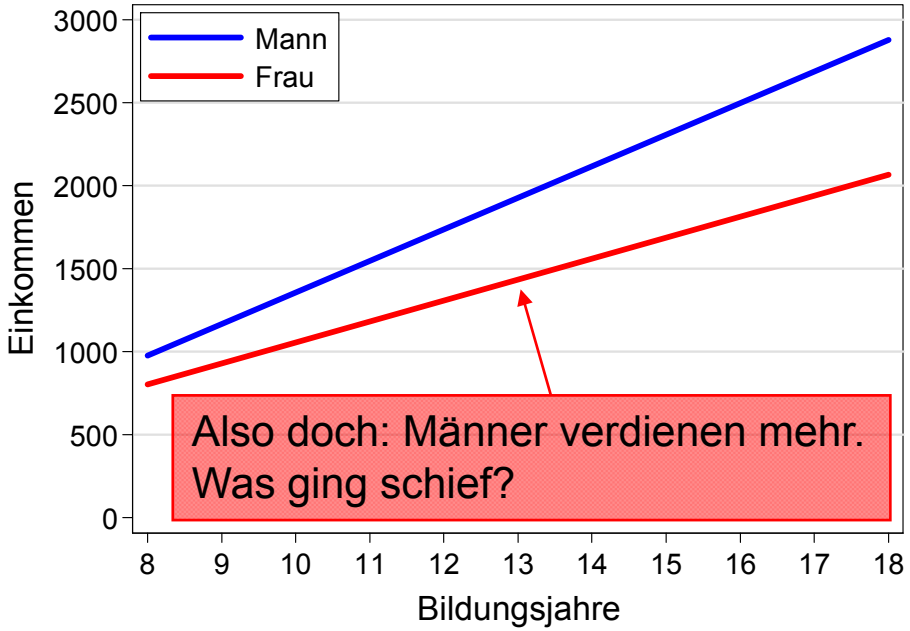
Daten: ALLBUS 2002
Do-File: 5 LinReg Interaktion.do

Slope-Interaktion: Geschlecht/Bildung

ohne Interaktion



mit Interaktion



Designmatrix	Konstante -546	Frau 335	Bild 190	Fbild -64	Regressions- gerade
Mann	1	0	1	0	-546 + 190•Bild
Frau	1	1	1	1	-211 + 126•Bild

Daten: ALLBUS 2002
Do-File: 5 LinReg Interaktion.do

Slope-Interaktion: Zentrierung

- Warum wurden wir in die Irre geführt?
  - Der Effekt von „Frau“ ist bei Bild=0 zu interpretieren
  - Dies ist offensichtlich eine sinnlose Interpretation
  - Problem tritt immer auf, wenn die metrische Interaktionsvariable keinen sinnvollen 0-Wert hat (z.B. auch bei Alter)
  - Abhilfe: Zentrieren der metrischen Variable (cbild = bild – mean(bild))

```
. regress eink c.cbild i.frau   i.frau#c.cbild

      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  3,  1114) =   79.61
       Model |   367104408     3   122368136           Prob > F      =  0.0000
    Residual |  1.7123e+09  1114  1537089.85           R-squared     =  0.1765
-------------+------------------------------           Adj R-squared =  0.1743
       Total |  2.0794e+09  1117  1861613.69           Root MSE      =  1239.8

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       cbild |   190.2989   15.17511    12.54   0.000     160.5239    220.0739
        frau |  -513.8864   77.55202    -6.63   0.000    -666.0509   -361.7219
frau#c.cbild |  -63.77517   26.41776    -2.41   0.016    -115.6094   -11.94099
       _cons |   1987.661   46.14574    43.07   0.000     1897.119    2078.204
------------------------------------------------------------------------------
```

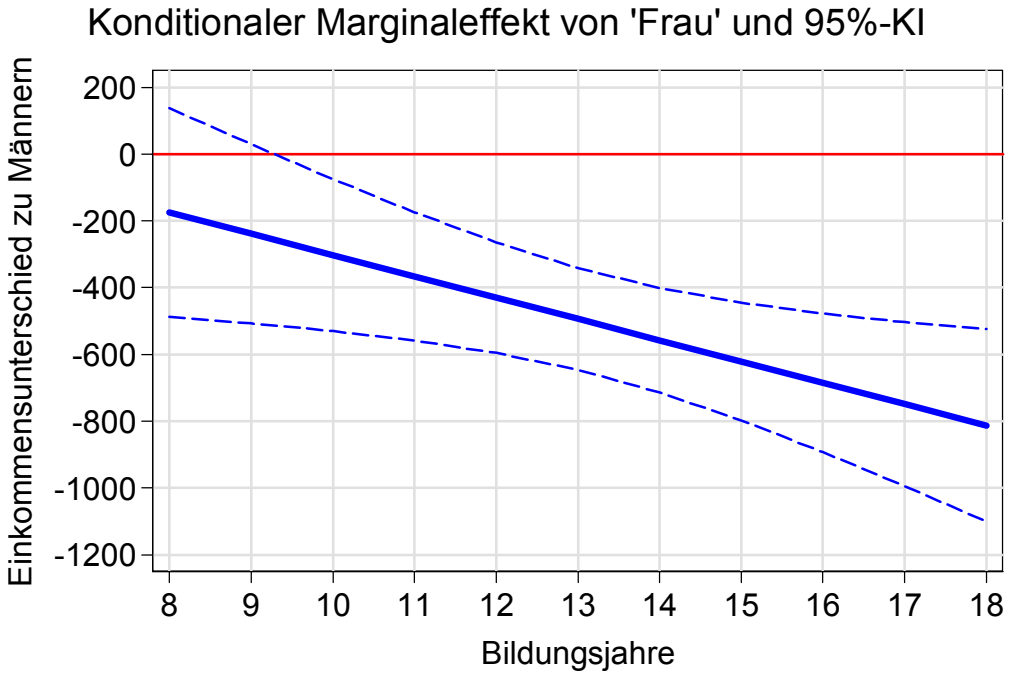
Berichte auch die konditionalen Marginaleffekte I

- Wir wissen nun, dass sich die Bildungsrenditen signifikant unterscheiden
- Oft wollen wir aber auch die Bildungsrendite in den beiden Gruppen wissen
  - Dies ist die Frage nach den konditionalen Marginaleffekten
  - `margins frau, dydx(cbild)`
- Man erhält sie auch über eine alternative Parametrisierung (**nested effects**)
- Im Beispiel: Bildungsrendite für Männer und Frauen getrennt
  - „Bild“ aus dem Modell nehmen, ersetzen durch:
    „Bild_F“: Bildung für Frauen, 0 sonst (zentriert)
    „Bild_M“: Bildung für Männer, 0 sonst (zentriert)

```
. regress eink  cbild_m  cbild_f  frau
      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  3,  1114) =   79.61
       Model |   367104408     3   122368136           Prob > F      =  0.0000
    Residual |  1.7123e+09  1114  1537089.85           R-squared     =  0.1765
-------------+------------------------------           Adj R-squared =  0.1743
       Total |  2.0794e+09  1117  1861613.69           Root MSE      =  1239.8

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cbild_m |   190.2989   15.17511    12.54   0.000     160.5239    220.0739
     cbild_f |   126.5237   21.62439     5.85   0.000     84.09456    168.9528
        frau |  -513.8864   77.55202    -6.63   0.000    -666.0509   -361.7219
       _cons |   1987.661   46.14574    43.07   0.000     1897.119    2078.204
------------------------------------------------------------------------------
```

Berichte auch die konditionalen Marginaleffekte II

- Symmetrie: Bildung moderiert den Effekt von „Frau“
  - Konditionale Marginaleffekte „Frau“: 335 – 64 * Bild
    - Ab Bild=6 ist der Effekt negativ. Dann geht die „Schere“ weiter auf
    - Achtung: der Interaktionseffekt (-64) ist zwar signifikant, das sagt uns aber nichts über die Signifikanz des Fraueneffekts!
    - Deshalb: ab welchem Bildungsniveau verdienen Frauen signifikant weniger als Männer?



Conditional-Effects-Plot

In Stata 12 leicht zu produzieren

```
margins, at(bild=(8(1)18)) dydx(frau)
marginsplot
```

Daten: ALLBUS 2002
Do-File: 5 LinReg Interaktion.do

Vollständige Interaktion mit Ost

- Will man Interaktionen aller Variablen zulassen
  - Interaktionsterme für alle Variablen (+ Hauptterm)
    - Test der Signifikanz der Interaktion für jede Variable einzeln
    - Signifikanztest für alle Interaktionseffekte (+Haupteffekt) (entspricht Signifikanztest für getrennte Modelle, Chow-Test)

```
. regress eink   i.ost##(i.frau c.cbild)
------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------------------------
       cbild |   181.2513   14.78765    12.26   0.000
        frau |  -588.6129   95.07484    -6.19   0.000
         ost |  -524.2035    99.1876    -5.28   0.000
   ost#cbild |  -32.26021   26.70952    -1.21   0.227
    ost#frau |   313.2981   161.8763     1.94   0.053
       _cons |   2148.898   54.73564    39.26   0.000
------------------------------------------------------------------------------
```

```
. * CHOW TEST
. contrast ost ost#i.frau ost#c.cbild, overall

------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     Overall |          3       10.83     0.0000
------------------------------------------------
```

Daten: ALLBUS 2002
Do-File: 5 LinReg Interaktion.do

Regeln für den Umgang mit Interaktionen

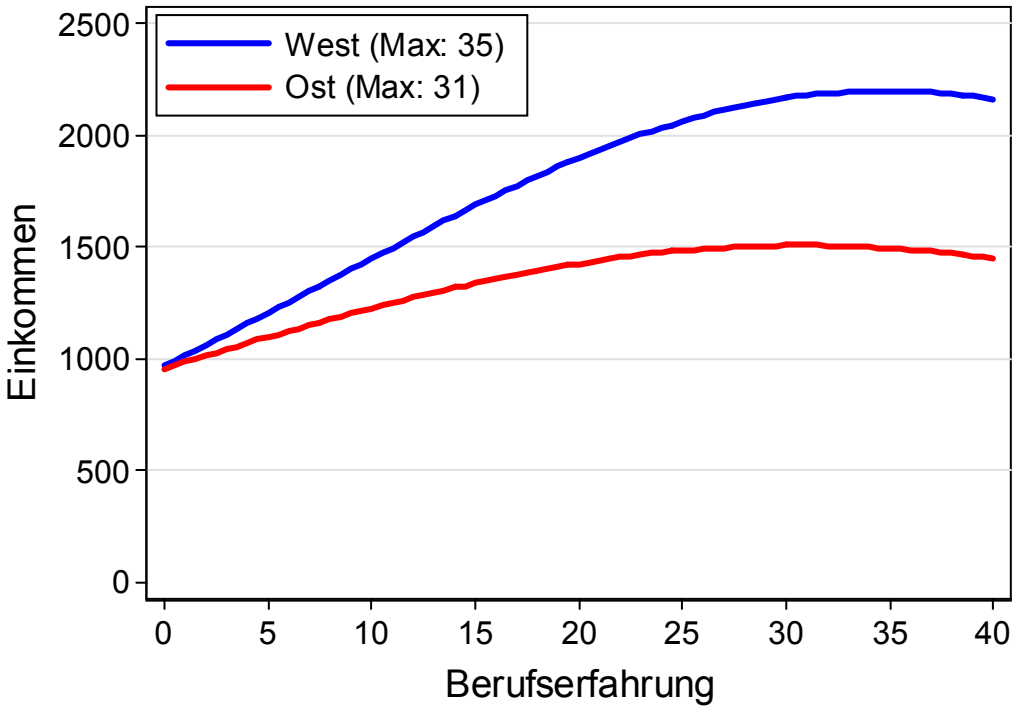
- Beide Hauptterme müssen im Modell sein
  - Ansonsten kaum realistische Restriktionen und schwer interpretierbar
  - Kollinearität zwischen Haupttermen und Interaktionstermen ist ein Datenproblem, kein Spezifikationsproblem! Bei hoher Kollinearität braucht man halt mehr Daten. Es macht aber keinen Sinn, den Hauptterm zu eliminieren.
- Zentriere metrische Interaktionsvariablen
  - Regressionskoeffizienten sind keine durchschnittlichen Marginaleffekte, sondern die Marginaleffekte an der Stelle Z = 0 (bzw. X = 0)
  - Bzw. wenn man Z zentriert hat: an der Stelle Z = mean(Z)
    - Zentrieren macht die Interpretation einfacher!
- Berichte inhaltlich bedeutsame Marginaleffekte (plus KI)
  - Z kategorial: berichte die Marginaleffekte von X in den Kategorien von Z
  - Z metrisch: plotte den Marginaleffekt von X gegen Z (Conditional-Effects-Plot)

Schließlich: Ein Humankapitalmodell getrennt für West/Ost

AV: logarithmiertes Einkommen

	West	Ost
bild	0.089***	0.082***
	(16.50)	(10.15)
exp	0.047***	0.029***
	(8.86)	(3.35)
exp^2	-0.001***	-0.000*
	(-5.56)	(-2.38)
frau	-0.246***	-0.186***
	(-7.25)	(-3.96)
_cons	5.723***	5.793***
	(64.68)	(43.08)
N	752	366
R^2	0.424	0.281

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



N.B.: Die Tabelle wurde direkt aus Stata mit dem Befehl „esttab“ erzeugt.

Daten: ALLBUS 2002
Do-File: 5 LinReg Interaktion.do

Ein komplexes Modell

Daten: ALLBUS 2002
Do-File: 3a LinReg Regressionsplots.do

```
. regress eink bild frau i.ost##(i.beruf c.exp##c.exp)
-----------------+------------------------------------------------------------
            bild |   142.4947    14.5792     9.77   0.000     113.8873    171.1021
            frau |  -470.5558    80.4638    -5.85   0.000    -628.4426   -312.6690
          1.ost |   458.5089   330.7956     1.39   0.166    -190.5811   1107.5989
                |
           beruf |
   Angestellter |   431.1194   112.9495     3.82   0.000     209.4889    652.7499
        Beamter |   387.5722   192.8944     2.01   0.045       9.0731    766.0714
  Selbständiger |  1138.4152   169.4858     6.72   0.000     805.8487   1470.9817
                |
             exp |    57.3432    15.9833     3.59   0.000      25.9807     88.7057
                |
     c.exp#c.exp |    -0.8235     0.3554    -2.32   0.021      -1.5209     -0.1261
                |
       ost#beruf |
 1#Angestellter |  -264.1148   176.2388    -1.50   0.134    -609.9322     81.7026
      1#Beamter |   846.6789   345.1562     2.45   0.014     169.4104   1523.9474
1#Selbständiger |  -599.7550   263.7621    -2.27   0.023    -1.12e+03    -82.1988
                |
       ost#c.exp |   -65.5124    30.2477    -2.17   0.031    -124.8648     -6.1601
                |
 ost#c.exp#c.exp |     1.1592     0.6706     1.73   0.084      -0.1567      2.4751
-----------------+------------------------------------------------------------
```

Die Ergebnisse dieses Modells sind in Tabellenform praktisch nicht zu interpretieren

Systematik der Regressionsplots

*** I) Koeffizientenplots**

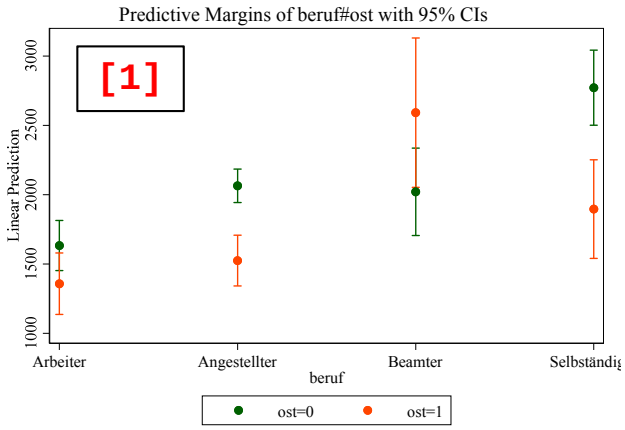
```
coefplot, drop(_cons) xline(0) base //hier nicht sehr hilfreich

margins, dydx(*)   // Plot der "Average Marginal Effects" (AME)
marginsplot, horizontal xline(0) plotopts(connect(i))  // auch
nicht sehr informativ
```

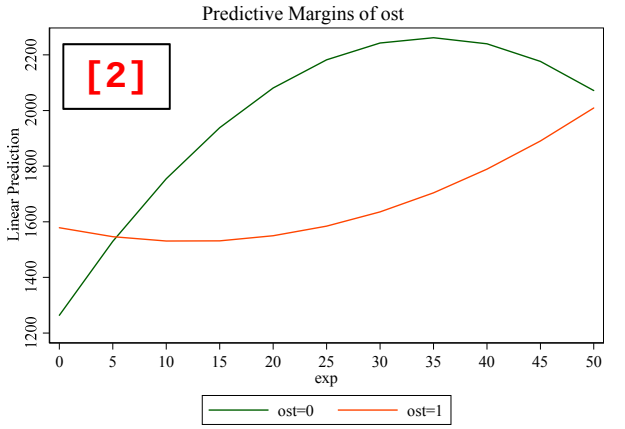
*** II) Profile Plot (PP)**

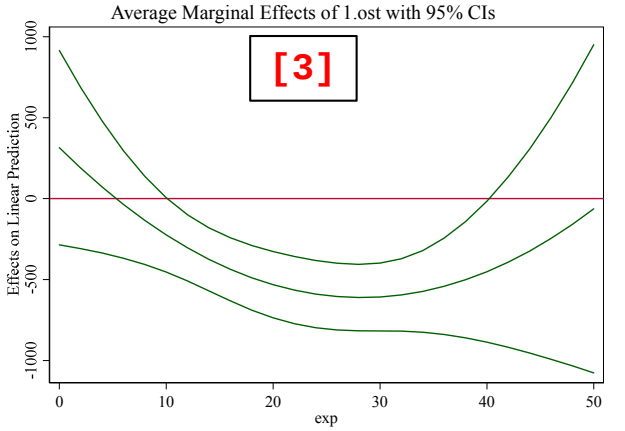
```
margins beruf                    //PP mit kategorialer Variable
marginsplot, plotopts(connect(i))

margins, at(exp=(0(5)50))        //PP mit metrischer Variable
marginsplot, recast(line) recastci(rarea)

margins beruf#ost                //Interaktion kategor. Variablen
marginsplot, plotopts(connect(i))  [1]

margins ost, at(exp=(0(5)50))  //Interaktion metr./kateg. Var.
marginsplot, noci recast(line)     [2]
```

*** III) Conditional-Effects Plot (CEP)**

```
margins ost, dydx(exp)                //AMEs Exp nach Ost (CEP I)
marginsplot, horizontal xline(0) plotopts(connect(i))

margins, dydx(ost) at(exp=(0(2)50)) //AMEs Ost nach Exp (CEPII)
marginsplot, recast(line) recastci(rline) yline(0)    [3]
```







LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Kapitel 8:
Regressionsdiagnostik

- Linearität
- Homoskedastizität
- Normalverteilung
- Ausreißer

Regressionsdiagnostik

- Die Schätzung der Regressionskoeffizienten und die Tests auf ihre Signifikanz sind von Annahmen abhängig
- Deshalb sollte auch immer überprüft werden, ob diese Annahmen gerechtfertigt sind. Im Folgenden:
  - Multikollinearität
  - Linearitätsannahme A1/A2
  - Homoskedastizitäts-Annahme A3
  - Normalverteilungsannahme A6
  - zusätzlich: Ausreißerdiagnostik
- Meist analysiert man dazu die Residuen (Residuenanalyse)
  - Die Residuen sind Schätzer für die Fehlerterme

$$\hat{\varepsilon}_i = \hat{\varepsilon}(y_i|x_i) = y_i - \hat{y}_i$$

Multikollinearität

- Perfekte Kollinearität: lineare Abhängigkeit unter den Variablen
  - Modell nicht schätzbar (ab $r \geq 0{,}99$ wird es kritisch)
  - Stata lässt automatisch Variablen weg, um Kollinearität zu beheben
- „Mäßige“ Multikollinearität ($r < 0{,}99$)
  - OLS schätzbar und konsistent
  - Aber S.E.s der betroffenen Variablen größer (Schätzung unpräzise)
    - Das sehen viele Forscher als Problem („Sternchenjagd“)
- Diagnose: variance-inflation-factor (VIF)

$$\hat{V}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(n-1)s_{x_j}{}^2} \frac{1}{1-R_j{}^2} \quad \rightarrow \quad VIF = \frac{1}{1-R_j{}^2}$$

  - Die S.E.s sind „inflationiert“ mit Faktor $\sqrt{VIF}$
  - Faustregel: problematisch falls $VIF > 30$
- Maßnahme
  - Spezifikation überprüfen (aber nicht: einfach Variable weglassen)
  - Mehr Daten erheben, damit die Schätzung präziser wird
  - Index bilden (betroffene Variablen messen dasselbe?)

Multikollinearität

Interaktionsterme korrelieren stark

```
regress eink   c.bild##frau
               c.exp##c.exp

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
        bild |      1.54    0.649513
      1.frau |     21.48    0.046559
 frau#c.bild |     21.83    0.045815
         exp |     14.55    0.068728
 c.exp#c.exp |     14.72    0.067929
-------------+----------------------
    Mean VIF |     14.82
```

Mit zentrierten Variablen ist das Problem geringer!

```
regress eink   c.bild_cen##frau
               c.exp##c.exp

. estat vif

    Variable   |       VIF       1/VIF
---------------+----------------------
    bild_cen   |      1.54    0.649513
      1.frau   |      1.01    0.991694
frau#c.bild_cen|      1.49    0.669306
         exp   |     14.55    0.068728
 c.exp#c.exp   |     14.72    0.067929
---------------+----------------------
    Mean VIF   |      6.66
```

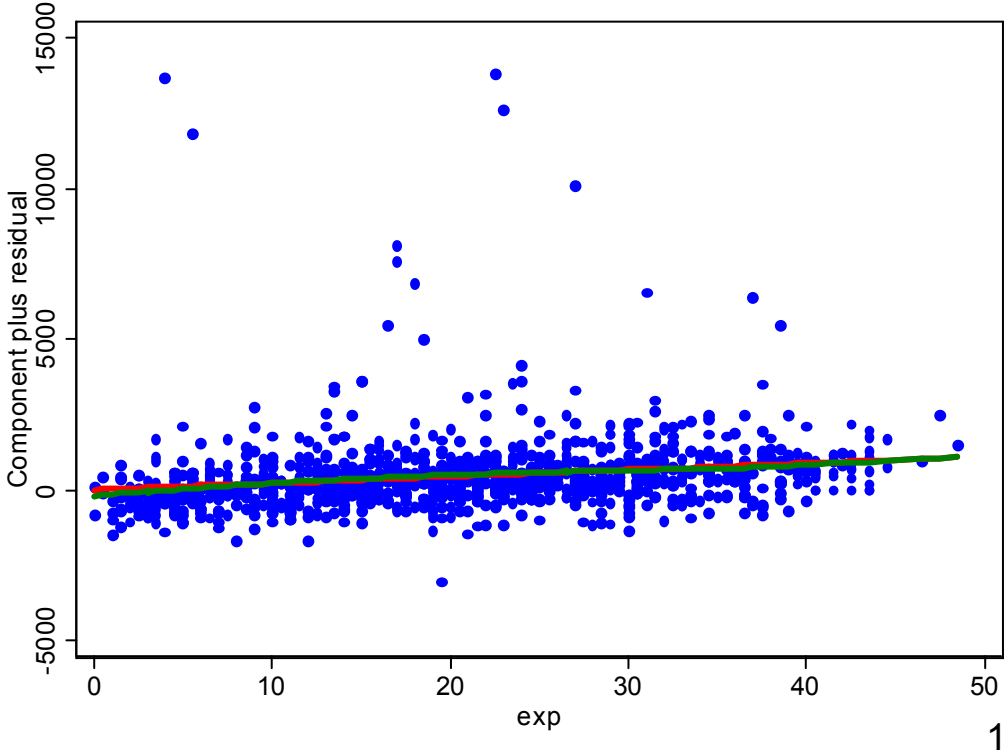
Linearität

Daten: ALLBUS 2002
Do-File: 6 LinReg Diagnostik.do

- Nicht-Linearität erkennt man in einem Residuen-Plot
  - Residuen gegen die uV auftragen
    - Abweichungen von der Null-Linie Anzeichen von Nicht-Linearität
  - In STATA: component-plus-residual plot (cprplot)
    - hilfreich: nicht-parametrischer Smoother (lowess)
    - weicht der Lowess von Regressionsgerade ab, dann Nicht-Linearität

```
* Beispiel: Berufserfahrung

regress eink bild exp frau
cprplot exp, lowess
```

Der Lowess (grün) zeigt nur geringfügige Abweichungen von der Gerade.
Es liegt also keine Nicht-Linearität vor.

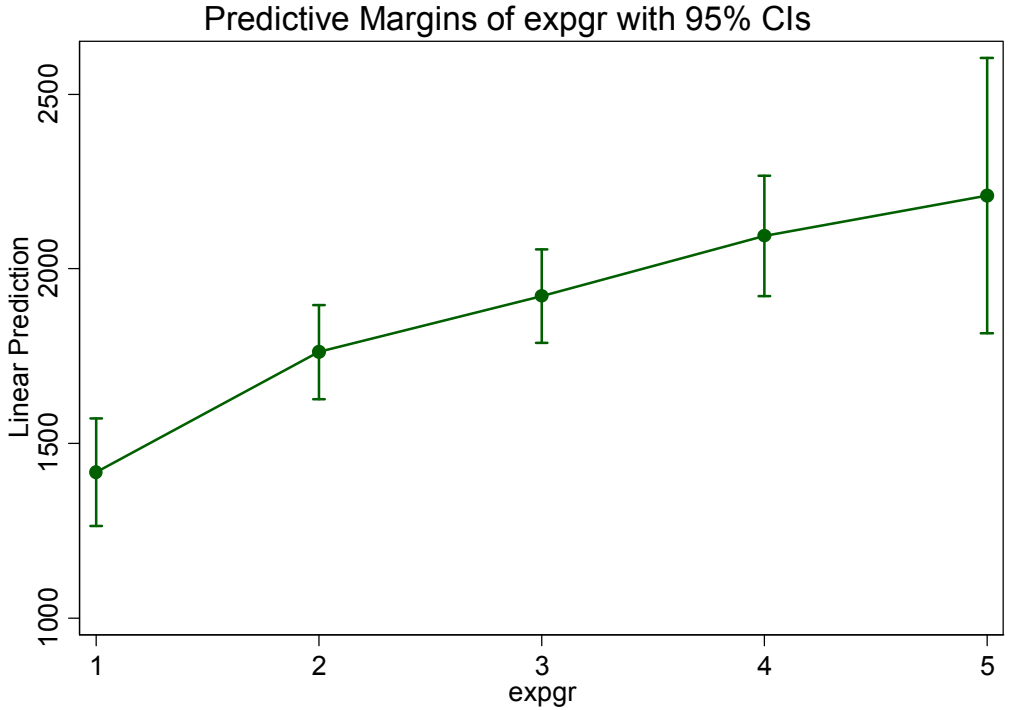


Linearität

Daten: ALLBUS 2002
Do-File: 6 LinReg Diagnostik.do

- Alternative: Man gruppiert die Variable und plottet die für die Gruppen vorhergesagten Werte
  - So kann man graphisch die Linearität beurteilen
  - Man kann auch testen, welche Polynomterme signifikant sind

```
recode expgr 0/10=1 10/20=2 20/30=3 30/40=4 40/50=5
regress eink bild frau i.expgr
margins expgr                      //Vorhergesagtes Einkommen in den Gruppen
marginsplot                        //Plot der vorhergesagten Werte
contrast p.expgr, asobserved       //Test, ob Polynomterme signifikant sind
```




```
------------------------------------------
             | df         F       P>F
-------------+----------------------------
      expgr  |
   (linear)  |  1     18.08    0.0000
(quadratic)  |  1      0.92    0.3371
    (cubic)  |  1      0.17    0.6825
  (quartic)  |  1      0.18    0.6715
      Joint  |  4     10.44    0.0000
             |
    Residual | 1111
------------------------------------------
```

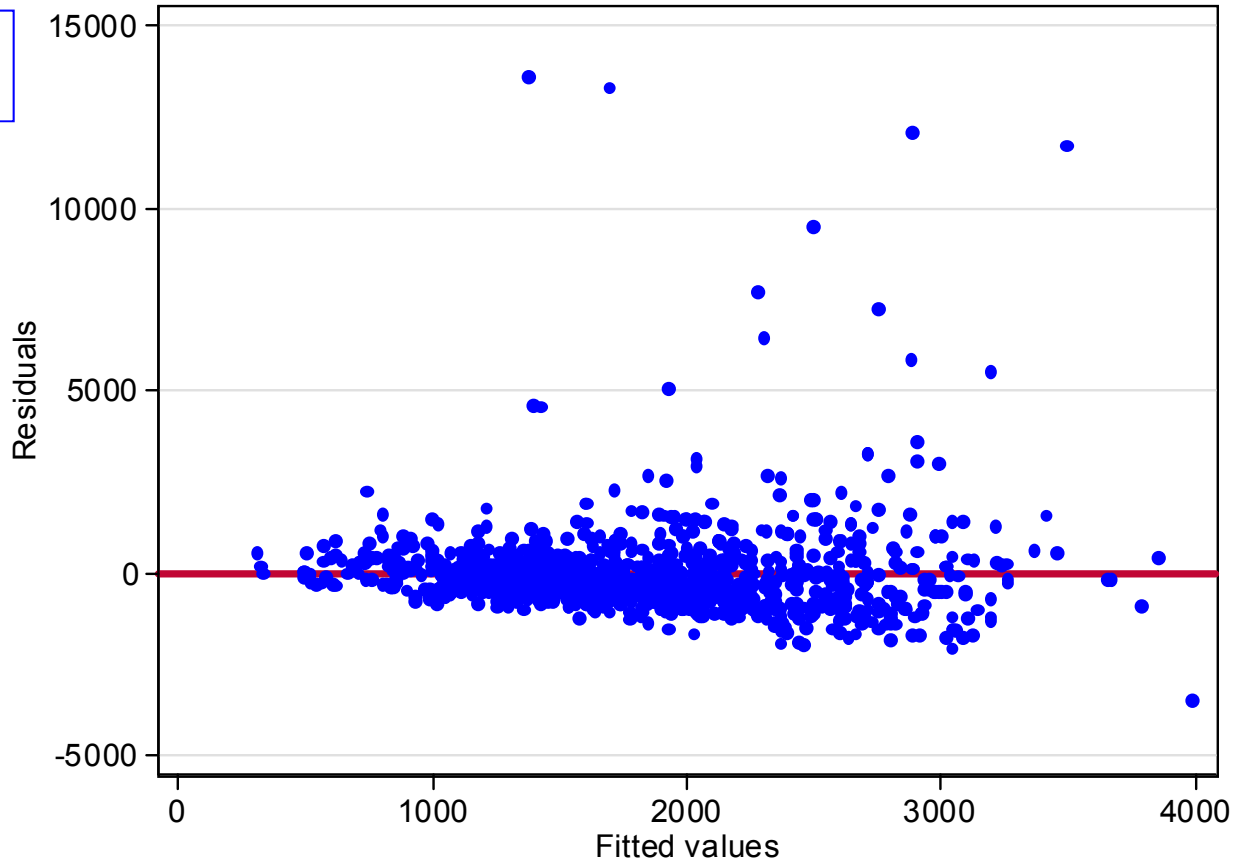
Homoskedastizität

Daten: ALLBUS 2002
Do-File: 6 LinReg Diagnostik.do

- Heteroskedastizität: Residuen streuen unterschiedlich
  - STATA: residual-versus-fitted-values Plot (rvfplot)

```
regress eink bild exp frau
rvfplot, yline(0)
```

Deutlicher Trichter erkennbar: Streuung der Residuen bei großen Werten von y-Dach höher.
Grund: rechtsschiefe Einkommensverteilung.
Abhilfe: Transformation (s.u.)

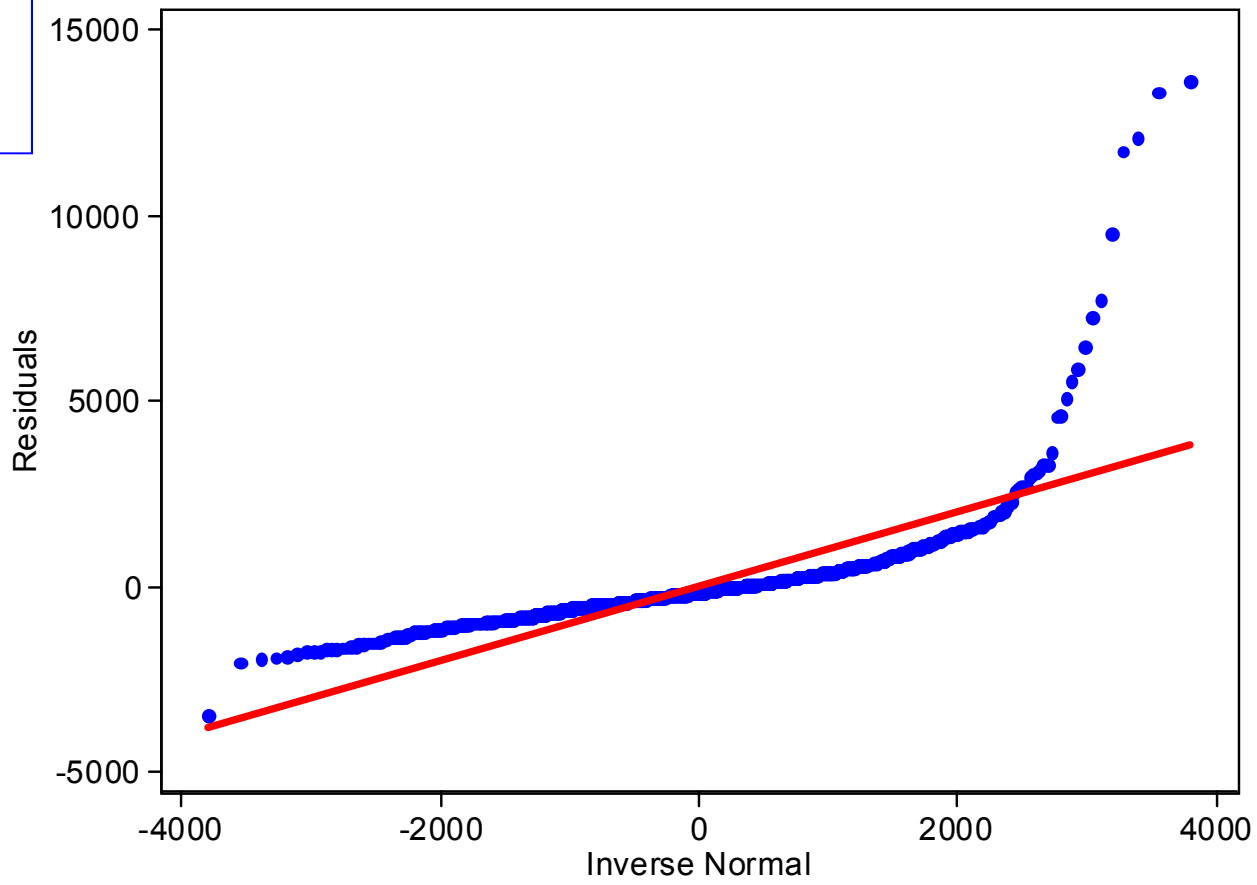


Normalverteilungsannahme

- Folgen die Residuen einer Normalverteilung?
  - STATA: Normal-Probability Plot (qnorm)

Daten: ALLBUS 2002
Do-File: 6 LinReg Diagnostik.do

```
regress eink bild exp frau
predict res1, residual
qnorm res1
```

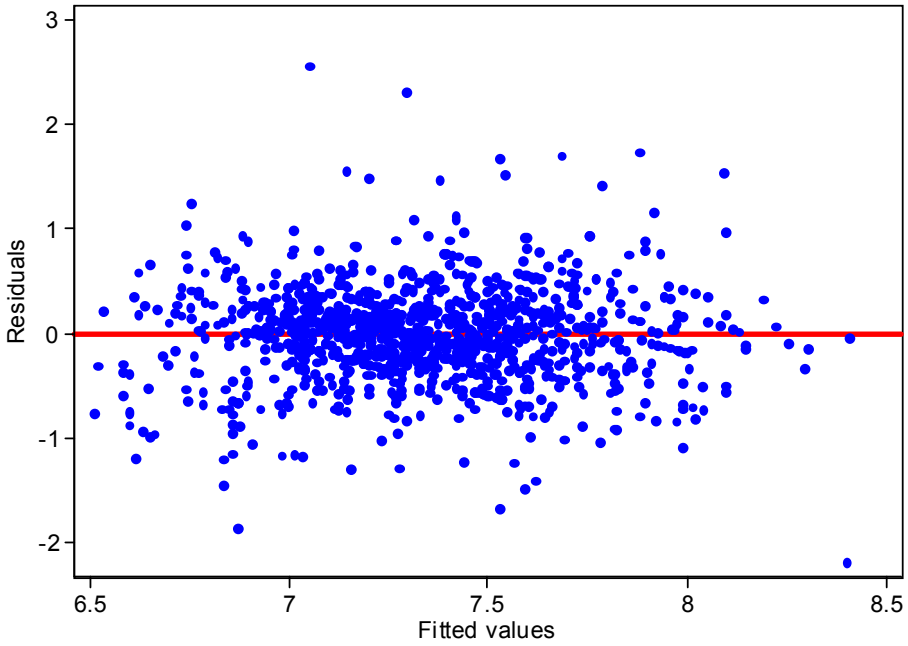


Die Residuen weichen deutlich von der roten Referenzlinie ab. Die Normalverteilungsannahme ist verletzt.
Grund: rechtsschiefe Einkommensverteilung
Abhilfe: logarithmische Transformation

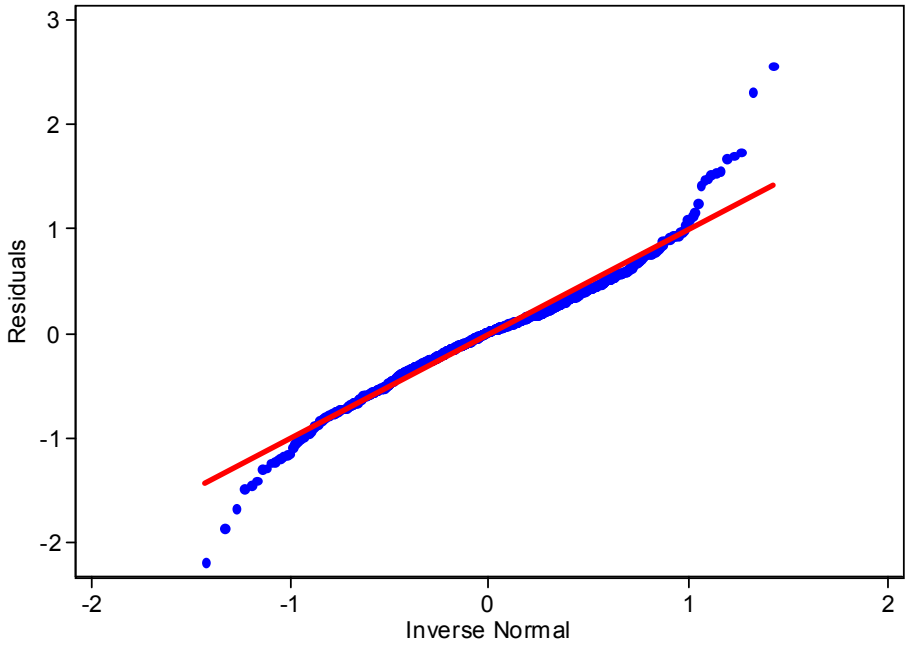
Semi-logarithmische Einkommensregression

```
. * logarithmische Transformation der aV
. generate lneink = ln(eink)
. regress lneink bild exp frau
```

Daten: ALLBUS 2002
Do-File: 6 LinReg Diagnostik.do



Kein augenfälliges Muster erkennbar: Homoskedastie



Nur mehr geringe Abweichungen an den Ränder: Residuen fast normalverteilt

Robuste Standardfehler

- Man kann die S.E.s „robust“ berechnen, d.h. sie sind robust gegen Verletzungen von A3
  - Huber-White-Sandwich-Estimator (option: `vce(robust)`)

	(1) normale S.E.s	(2) robuste S.E.s
bild	181.58***	181.58***
	(12.36)	(18.75)
exp	22.15***	22.15***
	(3.38)	(3.18)
frau	-474.57***	-474.57***
	(76.49)	(70.69)
N	1118	1118
R-sq	0.203	0.203

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Normaler Standardfehler:

$$\hat{V}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

$$\text{wobei} \quad \hat{\sigma}^2 = \frac{\sum_i \hat{\epsilon}_i^2}{n-k}.$$

Huber-White-Sandwich-Estimator:

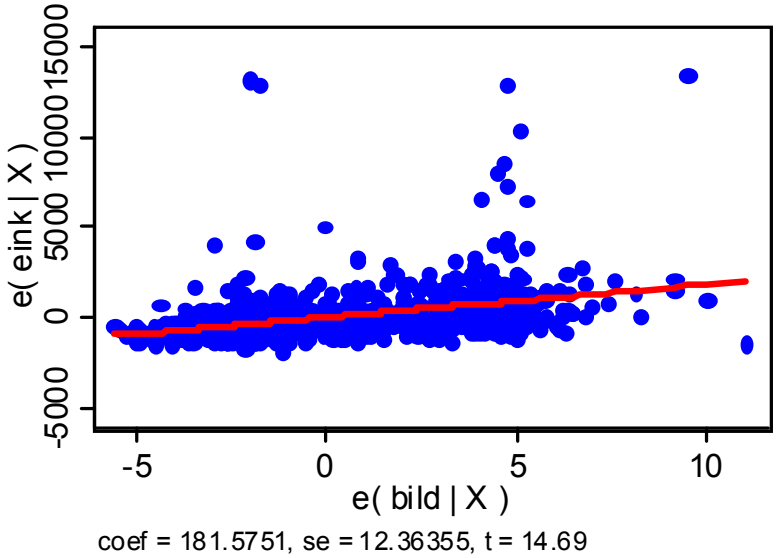
$$\hat{V}_{\mathrm{W}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

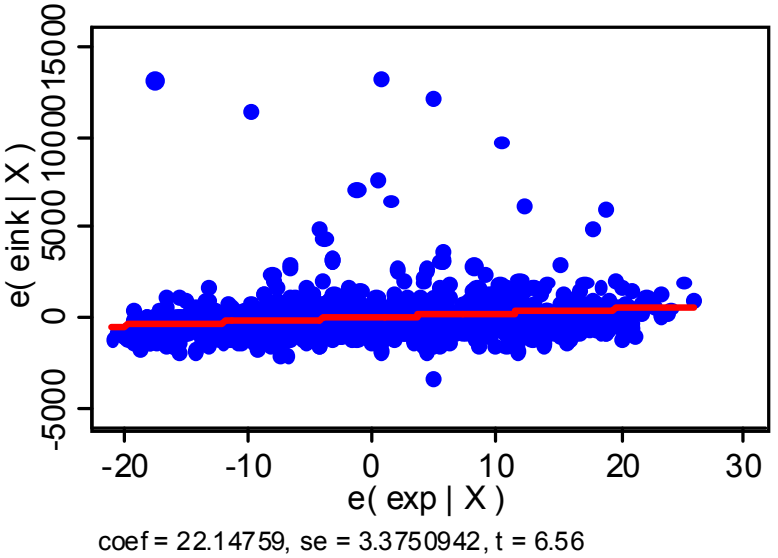
$$\text{wobei } \boldsymbol{D} = \operatorname{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2)$$

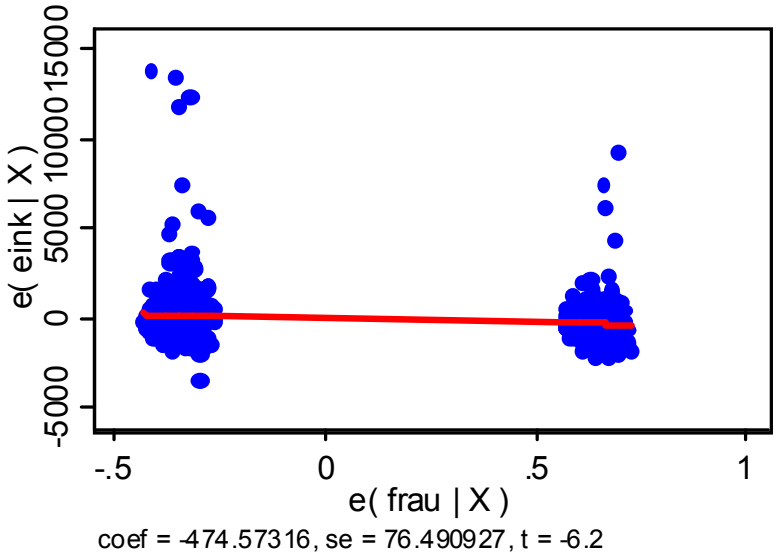
Ausreißerdiagnostik

- Ein Datenpunkt ist einflussreich, wenn seine Beseitigung die Ergebnisse der Regression deutlich verändert
  - Fälle mit ungewöhnlichem X- und Y-Wert (Ausreißer) haben Einfluss
  - Problem: das Ergebnis repräsentiert evtl. nur wenige Ausreißer
- Einflussdiagnostik
  - Im Streudiagramm erkennt man einflussreiche Datenpunkte
  - Im multiplen Fall: Partielles-Regressions Streudiagramm
  - Cook‘s D: Veränderung der Regressionskoeffizienten, wenn man einen Fall weglässt. Fälle mit besonders hohem D haben starken Einfluss.
- Abhilfe
  - Ist der einflussreiche Datenpunkt korrekt vercodet?
  - Fehlspezifikation? Was haben die einflussreichen Datenpunkte gemeinsam?
  - Weglassen ist keine Lösung, das ist Manipulation!

Partielle-Regressions Streudiagramme





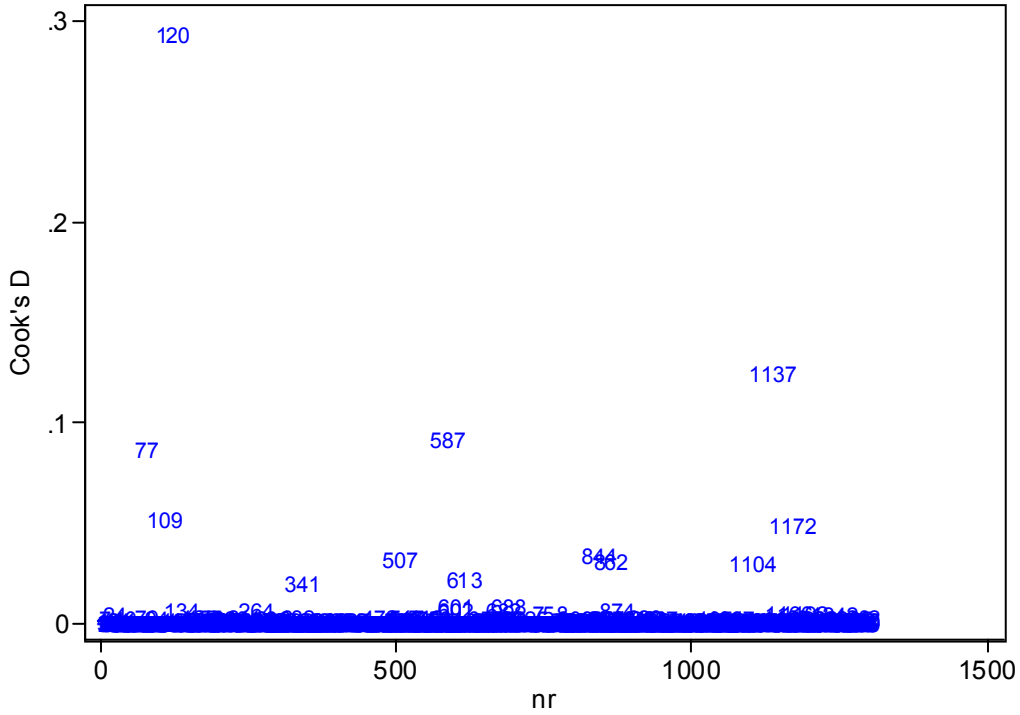


Plottyp: Added-variable plots

`avplots`

Daten: ALLBUS 2002
Do-File: 6 LinReg Diagnostik.do

Einflussreiche Datenpunkte




```
* Indexplot von Cook's D
gen nr=_n
scatter D nr,  msymbol(i) mlabel(nr)
mlabposition(0)
```

Besonderen Einfluss hat Fall 120.
Wir schauen uns die Fälle über 0,1 an.

```
. list eink bild exp frau if D>0.1 & D~=.

     +---------------------------+
     |  eink   bild   exp   frau |
     |---------------------------|
 120. | 15200   23.5   5.5      0 |
1137. | 15000     12     4      0 |
     +---------------------------+
```

Es handelt sich um „Großverdiener“.
Überprüfen, ob deren Einkommen richtig vercodet wurde.

Daten: ALLBUS 2002
Do-File: 6 LinReg Diagnostik.do