Vorlesung Multivariate Analyse

Kapitel I

Datenauswertung mit STATA

Prof. Dr. Josef Brüderl
Universität Mannheim

Herbstsemester 2007

Methoden-Curriculum B.A. Soziologie

Basismodul: Methoden und Statistik: 22

VL Datenerhebung (2): 5
ÜK (2): 3

VL Datenauswertung (2): 5
Ü (2): 2

VL Multivariate Analyse (2): 5
Ü (2): 2

Aufbaumodul: Methoden der empirischen Sozialforschung: 13

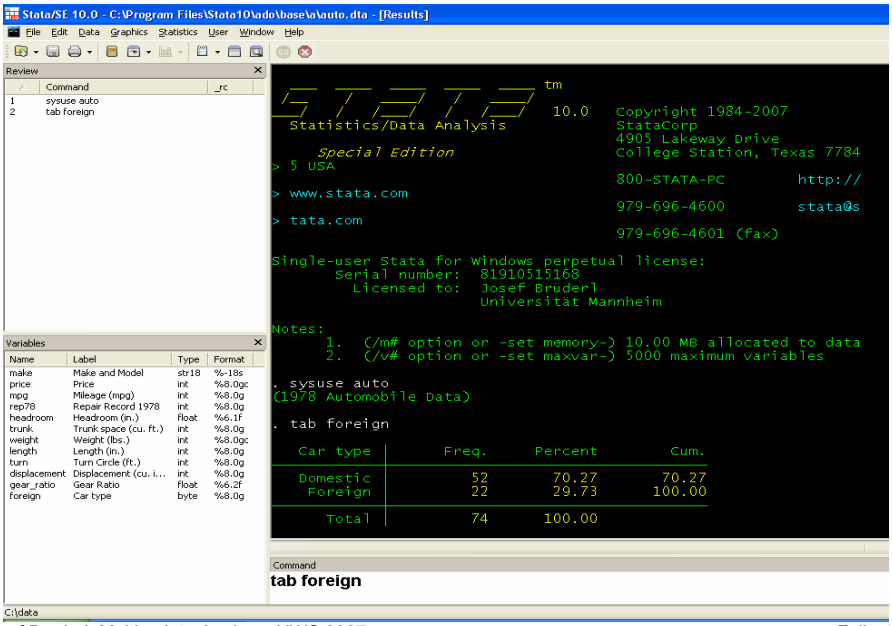
ÜK Datenerhebungsseminar (2): 5

ÜK Datenanalyseseminar (4): 8

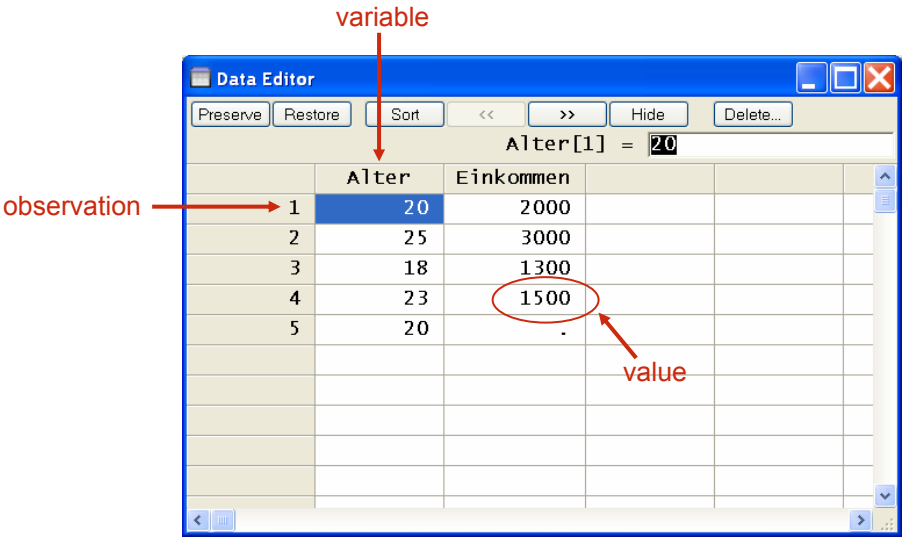
Lernziele

- Beherrschung einfacher Datenauswertungen mit STATA
  - Einführung in der Vorlesung
  - Übungen am PC im Tutorium
- Kenntnis multivariater Analyseverfahren
  - Tabellenanalyse
  - Lineare Regression
  - Logistische Regression
- Ergebnisse multivariater Verfahren interpretieren können
  - STATA output

STATA



Dateneingabe



Grundauszählung: Häufigkeitstabelle

```
. codebook Einkommen

-------------------------------------------------------------------------------
Einkommen                                                      Netto pro Monat
-------------------------------------------------------------------------------
                  type:  numeric (int)

                 range:  [1300,3000]                  units:  100
        unique values:  4                        missing .:  1/5

           tabulation:  Freq.  Value
                           1  1300
                           1  1500
                           1  2000
                           1  3000
                           1  .
```

```
. tabulate Alter

      Alter |
  Befragter |      Freq.     Percent        Cum.
------------+-----------------------------------
         18 |          1       20.00       20.00
         20 |          2       40.00       60.00
         23 |          1       20.00       80.00
         25 |          1       20.00      100.00
------------+-----------------------------------
      Total |          5      100.00
```

ALLBUS 2002

- Bevölkerungsumfrage alle 2 Jahre seit 1980 (N ~ 3.000)
  - Von ZUMA als Service für die Sozialforschung
  - Trenddaten
- ALLBUS 2002
  - Einwohnermelderegisterstichprobe, Ossis überrepräsentiert
  - GG: alle deutschsprachigen Personen über 18, wohnhaft in D in Privathaushalten
  - Ausschöpfung: 47%
  - Mündliches Interview (CAPI)
  - Infos: http://www.gesis.org/Datenservice/ALLBUS/index.htm
- Die Daten des ALLBUS 2002 (N=2.820)
  - Im CIP-Pool auf Laufwerk K:
  - Zusätzlich das Codebuch und der Fragebogen
  - Auswertungen West/Ost getrennt (oder gewichtet mit v718)

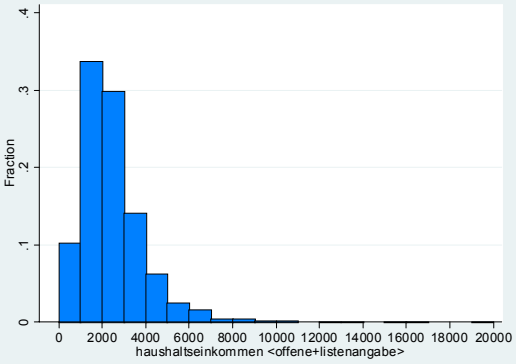
Univariate Kennzahlen

```
. codebook v441
-------------------------------------------------------------------------------
v441                                          haushaltseinkommen <offene+listenangabe>
-------------------------------------------------------------------------------
                 type:  numeric (int)
                label:  v441, but 239 nonmissing values are not labeled
                range:  [85,20000]                   units:  1
        unique values:  239                      missing .:  544/2820
```

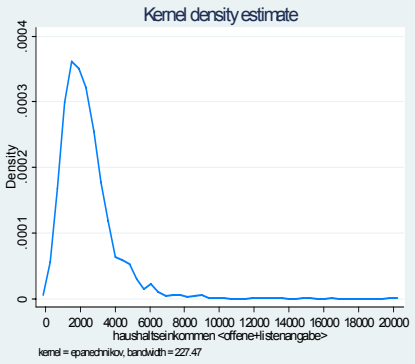
```
. summarize v441, detail

          haushaltseinkommen <offene+listenangabe>
-------------------------------------------------------------
      Percentiles      Smallest
 1%           360             85
 5%           700            125
10%           968            130       Obs                2276
25%          1400            205       Sum of Wgt.        2276

50%          2000                      Mean             2379.44
                        Largest       Std. Dev.        1535.666
75%          3000          13500
90%          4000          15000       Variance         2358270
95%          5000          16500       Skewness        2.870172
99%          8000          20000       Kurtosis         20.8778
```

Verteilungsgraphen

```
. histogram v441, width(1000) start(0)
(bin=20, start=0, width=1000)
```




```
. kdensity v441
```



Datenaufbereitung

- Vor der Datenauswertung ist praktisch immer eine Datenaufbereitung nötig
  - Fallauswahl
    - permanent: `keep if v3==1  drop if v3>1`
    - temporär: `summ v441 if v3==1`
  - Neue Variablen bilden

```
generate auslfeind = (v69 + v70) / 2
```

  - Ersetzen bestehender Variablen

```
replace auslfeind = sqrt(auslfeind^2)
```

  - Rekodierung bestehender Variablen

```
recode auslfeind 1/3.5=1 4/7=2
```

  - Labeln der Werte einer neuen Variable

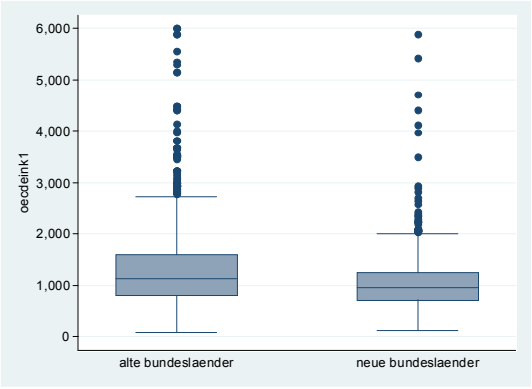
```
label variable  auslfeind   "Ausländerfeindlichkeit"
label define    ausllbl     1 "niedrig"   2 "hoch"
label value     auslfeind   ausllbl
tab             auslfeind
```

Beispiel: Armut in Deutschland

```
* monatliches Netto-HHeinkommen
generate   hheink = v441

* Bestimmung der Personenzahl im HH
tab v363                                //HHgröße
generate   hhvorst = 1                  //jeder HH hat einen Haushaltsvorstand
replace    hhvorst = . if v363==.
generate   sonst   = 0                  //Initialisierung "Zahl sonstiger Personen über 14"
generate   kind14  = 0                  //Initialisierung "Zahl Kinder unter 15"
replace    kind14  = kind14+1 if v369<15   //Bestimmung der Zahl der Kinder unter 15
replace    kind14  = kind14+1 if v379<15
replace    kind14  = kind14+1 if v389<15
replace    kind14  = kind14+1 if v399<15
replace    kind14  = kind14+1 if v409<15
replace    kind14  = kind14+1 if v419<15
replace    kind14  = kind14+1 if v429<15
replace    kind14  = . if v363==.
replace    sonst   = v363 - kind14 - 1  //Bestimmung der sonstigen Personen als Residualgröße

* Kontrolle
tab1     hhvorst sonst kind14
generate hhgr = hhvorst+sonst+kind14
tab1     hhgr v363

* Berechnung des bedarfsgewichteten Pro-Kopf-Einkommens
* (Äquivalenzeinkommen) nach der OECD-Skala. Die OECD-Gewichte sind
* Haushaltsvorstand:         1.0
* weitere Personen über 14:  0.7
* Kinder unter 15:           0.5
generate   oecdeink = hheink / (1*hhvorst + 0.7*sonst + 0.5*kind14)
summ       oecdeink, detail
```


Beispiel: Armut in Deutschland

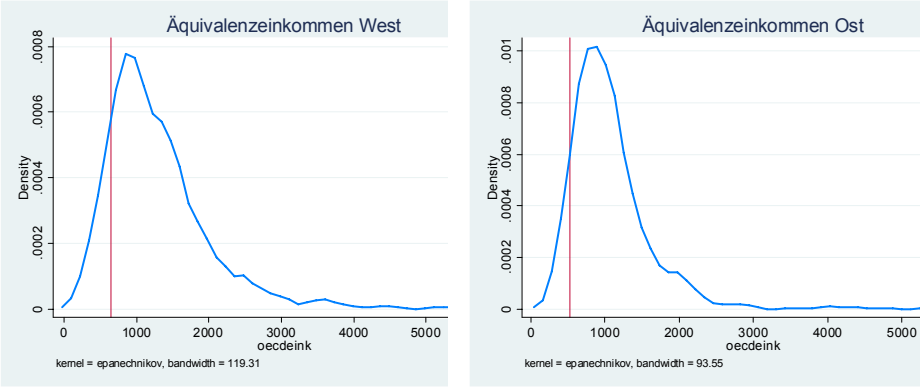
```
* Gruppiertes Boxplot West/Ost (der Aussreisser wird rekodiert)
gen    oecdeink1 = oecdeink
recode oecdeink1 6000/max=6000
graph box oecdeink1, over(v3)
```



Armutsgrenzen (Euro netto monatlich)		
	West	Ost
arithm. Mittel	649	531
Median	568	479

Armutsquoten		
	West	Ost
arithm. Mittel	14,0%	10,0%
Median	9,4%	6,1%

Beispiel: Armut in Deutschland



Eingezeichnet sind die auf dem arithmetischen Mittel basierenden Armutsgrenzen

Konfidenzintervalle

- Das 95%-Konfidenzintervall ist gegeben durch

$$\bar{x} \pm 1{,}96 \cdot \sigma_{\bar{X}} \quad \text{wobei} \quad \sigma_{\bar{X}} = \sigma / \sqrt{n}$$

```
. sort v3
. ci oecdeink, by(v3)

-------------------------------------------------------------------------------------------
-
-> v3 = alte bun

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    oecdeink |       1532    1298.624    20.10557        1259.187    1338.061

-------------------------------------------------------------------------------------------
-
-> v3 = neue bun

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    oecdeink |        742     1062.13    20.91594        1021.069    1103.192
```

Bivariate Datenanalyse: t-Test

Annahme

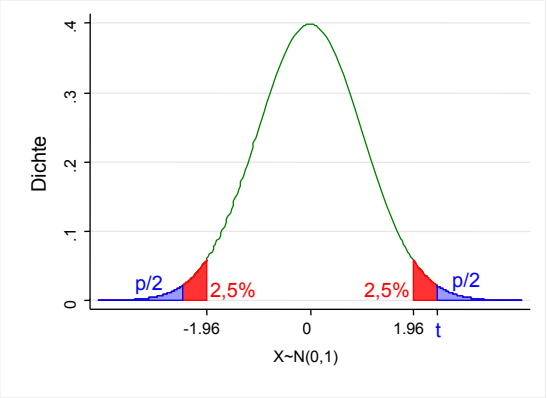
```
. ttest oecdeink, by(v3)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
alte bun |    1532    1298.624    20.10557    786.9475    1259.187    1338.061
neue bun |     742     1062.13    20.91594    569.7435    1021.069    1103.192
---------+--------------------------------------------------------------------
combined |    2274    1221.457    15.34202    731.6068    1191.371    1251.543
---------+--------------------------------------------------------------------
    diff |            236.4935    32.35108                173.0527    299.9343
------------------------------------------------------------------------------
    diff = mean(alte bun) - mean(neue bun)                        t =   7.3102
Ho: diff = 0                                     degrees of freedom =     2272

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

zweiseitiger Test
p-Wert

Teststatistik

Exkurs: p-Wert

- Der p-Wert gibt die Wahrscheinlichkeit an, dass die Teststatistik den berechneten Wert oder einen, der noch weiter in Richtung der Alternativhypothese liegt, annimmt
- Die Nullhypothese wird dann verworfen, wenn **p ≤ α**



H_0 : diff = 0

Bei 2272 df ist die t-Verteilung eine Standardnormalverteilung

Die kritischen Werte sind auf dem 5%-Niveau -1,96 und 1,96

Liegt die Teststatistik **t** z.B. bei 2,4, so kann die H_0 abgelehnt werden

Rechts von **t** liegt p/2 Whs. (die Whs., dass noch was Extremeres rauskommt)

Da offensichtlich p < 0,05, kann die H_0 abgelehnt werden

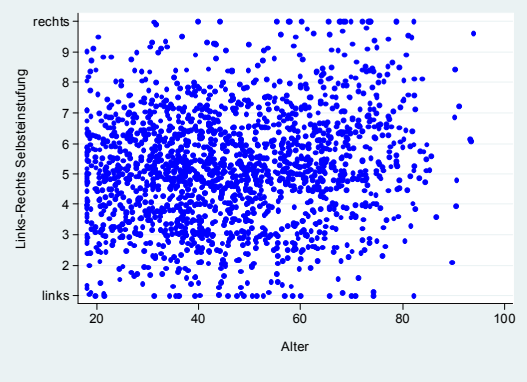
Die Kreuztabelle

```
. recode v521 8=.
. tab v521 v3, col chi2 V

Wahlabsicht          | alte bund  neue bund |     Total
---------------------+----------------------+----------
             cdu-csu |       516        206 |       722
                     |     36.24      31.02 |     34.58
---------------------+----------------------+----------
                 spd |       460        200 |       660
                     |     32.30      30.12 |     31.61
---------------------+----------------------+----------
              f.d.p. |       216         68 |       284
                     |     15.17      10.24 |     13.60
---------------------+----------------------+----------
   buendnis90-gruene |       169         32 |       201
                     |     11.87       4.82 |      9.63
---------------------+----------------------+----------
        republikaner |        18          9 |        27
                     |      1.26       1.36 |      1.29
---------------------+----------------------+----------
                 pds |        28        133 |       161
                     |      1.97      20.03 |      7.71
---------------------+----------------------+----------
       andere partei |        17         16 |        33
                     |      1.19       2.41 |      1.58
---------------------+----------------------+----------
               Total |     1,424        664 |     2,088
                     |    100.00     100.00 |    100.00

          Pearson chi2(6) = 231.5946  Pr = 0.000
               Cramér's V =   0.3330
```

p-Wert

Korrelation: Alter und Links-Rechts

```
. twoway (scatter v106 v185 if v3==1, jitter(10))
```




```
. correlate v106 v185 if v3==1
(obs=1820)

       |     v106     v185
-------+------------------
  v106 |   1.0000
  v185 |   0.1656   1.0000


. spearman v106 v185 if v3==1

Number of obs  =    1820
Spearman's rho =   0.1487

Test of Ho: v106 and v185 are
independent
    Prob > |t| =       0.0000
```

Regression: Alter und Links-Rechts

```
. regress v106 v185 if v3==1

      Source |       SS       df       MS              Number of obs =    1820
-------------+------------------------------           F(  1,  1818) =   51.25
       Model |  168.820066     1  168.820066            Prob > F      =  0.0000
    Residual |  5989.10081  1818  3.29433488            R-squared     =  0.0274
-------------+------------------------------           Adj R-squared =  0.0269
       Total |  6157.92088  1819  3.38533308            Root MSE      =   1.815

------------------------------------------------------------------------------
        v106 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        v185 |   .0177624   .0024813     7.16   0.000     .0128959    .0226288
       _cons |   4.364548   .1233541    35.38   0.000     4.122617    4.606479
------------------------------------------------------------------------------
```

R^2

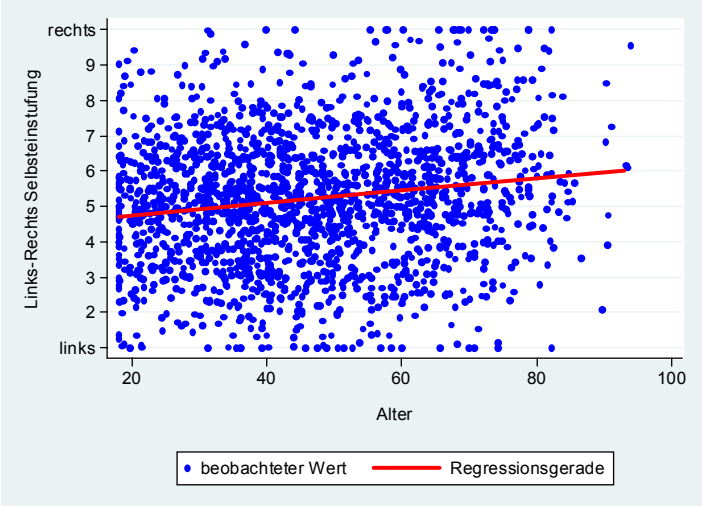
Regressionskoeffizient

t-Wert

p-Wert

Regression: Alter und Links-Rechts

```
. twoway (scatter v106 v185) (lfit v106 v185) if v3==1
```



Vorlesung Multivariate Analyse

Kapitel II

Multivariate Datenanalyse

Prof. Dr. Josef Brüderl
Universität Mannheim

Herbstsemester 2007

Kausalität

- Wissenschaft sucht nach kausalen Beziehungen
  - Ursache-Wirkungs-Beziehungen
- Die wissenschaftliche Methode: das Experiment
  - Versuchs- und Kontrollgruppe
  - Randomisierung
    - Dadurch unterscheiden sich Versuchs- und Kontrollgruppe nicht
    - Keine unbeobachtete Heterogenität
  - Kontrollierte Stimulussetzung durch Forscher
    - Damit ist sichergestellt, dass die uV der aV zeitlich vorgeht
    - Keine Endogenität
- Ein sauber durchgeführtes Experiment erlaubt einen sicheren Kausalschluss
- Experimente sind aber in den Sozialwissenschaften meist nicht praktikabel

Korrelation und Kausalität

- Deshalb erhebt man oft Daten über X und Y ex-post-facto und berechnet deren Korrelation
- Korrelation ist aber nicht gleich Kausalität
- Um von einer Korrelation auf Kausalität schließen zu können, müssen folgende Bedingungen gelten:
  - X und Y sind korreliert
  - X geht Y zeitlich voran (keine Endogenität)
    - Paneldaten nötig
    - Bei Querschnittsdaten hilft nur Theorie
  - Die Korrelation von X und Y bleibt erhalten, auch wenn man für dritte Variablen kontrolliert (keine unbeobachtete Heterogenität)
    - Dies macht man mit multivariaten Analyseverfahren
- Damit kommt der multivariaten Analyse in der Sozialforschung eine zentrale Rolle zu:
  Sie ist **der** Ersatz für das Experiment

Beispiel: Studienzulassung

- Diskriminierung von Frauen bei der Studienzulassung?
- Anfang der 70er Jahre wurde gegen die University of California at Berkely geklagt, weil diese relativ weniger Frauen als Männer zum Studium zuließ (s.: Krämer, W., 1995: Denkste! Campus-Verlag)

fiktives (!) Beispiel:

	M	F	Σ
nicht zug.	400	450	850
zug.	100 (20%)	50 (10%)	150
Σ	500	500	1000

Φ = (-) 0.14

Beispiel: Studienzulassung

Der Universität gelang es jedoch nachzuweisen, dass Frauen in Wirklichkeit gar nicht benachteiligt waren...

Mathe

	M	F	Σ
nicht zug.	100	10	110
zug.	80 (44%)	10 (50%)	90
Σ	180	20	200

Φ = (+) 0.03

SoWi

	M	F	Σ
nicht zug.	300	440	740
zug.	20 (6%)	40 (8%)	60
Σ	320	480	800

Φ = (+) 0.04

Beispiel: Studienzulassung

..., vielmehr war der Zusammenhang zwischen Geschlecht und Zulassung darin begründet, dass...

...Frauen sich häufiger für solche Fächer bewarben...

	M	F	Σ
Mathe	180	20	200
Sowi	320	480	800
Σ	500	500	1000

Φ = (+) 0.40

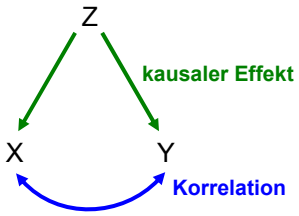
..., die insgesamt niedrigere Zulassungsquoten hatten.

	Mathe	Sowi	Σ
nicht zug.	110	740	850
zug.	90	60	150
Σ	200	800	1000

Φ = (-) 0.42

Hauptproblem: Scheinkorrelation

- X und Y korrelieren zwar, aber Grund hierfür ist eine dritte Variable Z, die sowohl X als auch Y kausal verursacht
  - Die Korrelation ist „echt“, aber die Kausalität ist „scheinbar“ (Scheinkausalität)
- Schematisch anhand eines Pfaddiagramms



- Z ist eine „antezedierende“ Variable
- Durch die beiden Kausaleffekte entsteht eine Korrelation von X und Y
- Es wäre ein Fehler diese Korrelation als kausal zu interpretieren
- Durch Kontrolle von Z (Drittvariablen-kontrolle) kann man das Problem beheben

- Beispiele
  - Zahl der Störche in einer Region korreliert positiv mit der Zahl der Geburten
  - Menge der Haare auf Kopf von Männern korreliert negativ mit Einkommen
  - Techno-Liebhaber wählen eher GRÜN
  - Regelmäßiges Zähneputzen senkt das Herzinfarktrisiko

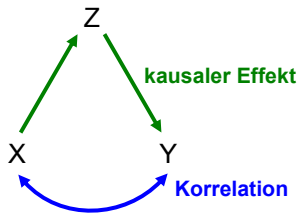
Die Lösung: Drittvariablenkontrolle

- Man muss für Z (statistisch) kontrollieren
  - Multivariate Analyseverfahren
- Z konstant halten: Konditionale Kreuztabellen (Partialtabellen)
  - Für jede Ausprägung von Z wird eine eigene Kreuztabelle (X × Y) erstellt (dreidimensionale Kreuztabelle)
  - Damit erhält man für jede Ausprägung von Z einen eigenen, konditionalen Korrelationskoeffizienten:

    $r_{XY \bullet Z_1}$, $r_{XY \bullet Z_2}$, usw.
  - Die messen die Korrelation von X und Y unter Kontrolle von Z
- Scheinkorrelation liegt vor, wenn die konditionalen Korrelationskoeffizienten gleich null sind
- Beispiel Studienzulassung (Folie 5)
  - Die konditionalen Phi-Koeffizienten sind praktisch null
  - (Ist aber keine Scheinkorrelation, s. nächste Folie)

Was noch passieren kann: Intervention

- Ist Z intervenierend, so liegt eine Intervention vor
- Schematisch anhand eines Pfaddiagramms



- Z ist eine „intervenierende“ Variable
- X hat einen „indirekten“ Kausaleffekt auf Y
- Kontrolliert man für Z, so werden die konditionalen Korrelationskoeffizienten null
- Damit hat man aber nicht „scheinbare“ Kausalität aufgedeckt, sondern einen „kausalen Mechanismus“ aufgedeckt

- Wichtig ist also, sich zu überlegen, ob Z antezedierend oder intervenierend ist
- Beispiel
  - Zulassungsbeispiel: Geschlecht – Fach – Zulassungsquote

Beispiel: Kirchgangshäufigkeit

- Mit dem ALLBUS 1994 untersuchen wir, wie sich der Wohnort (West/Ost) auf den Kirchgang auswirkt

Kirchgang nach Wohnort

	West	Ost
Selten/nie	55%	84%
Öfter	45%	16%
N	2339	1104

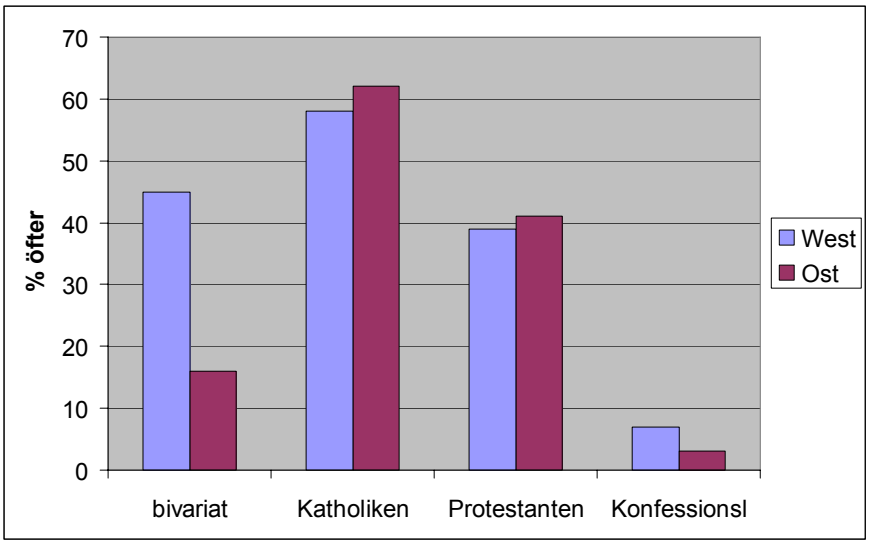
V = 0,28

- Um zu überprüfen, ob die Korrelation nur „scheinbar“ ist, kontrollieren wir für Konfession (2 × 2 × 3-Tabelle)

Kirchgang nach Konfession und Wohnort

	Katholiken		Protestanten		Konfessionslose	
	West	Ost	West	Ost	West	Ost
Selten/nie	42%	38%	61%	59%	93%	97%
Öfter	58%	62%	39%	41%	7%	3%
	V = 0,01		V = 0,01		V = 0,07	

Beispiel: Kirchgangshäufigkeit



Beispiel: Kirchgangshäufigkeit

- Um den „Kausalmechanismus“ ganz zu verstehen, erstellen wir noch die Kreuztabellen X × Z und Z × Y

Konfession nach Wohnort

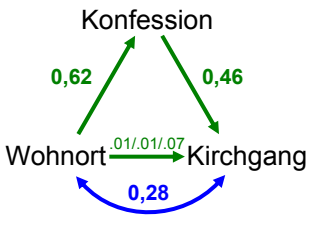
	West	Ost
Katholik	47%	3%
Protestant	41%	26%
Konfessionsl.	12%	71%

V = 0,62

Kirchgang nach Konfession

	selten	öfter
Katholik	41%	59%
Protestant	60%	40%
Konfessionsl.	96%	4%

V = 0,46



Das gesamte Kausalmodell präsentieren wir übersichtlich in einem Pfaddiagramm.
An die Pfeile schreiben wir die bivariaten Korrelationskoeffizienten.
Am Pfeil von Wohnort auf Kirchgang stehen die konditionalen Korrelationsk. Die sind fast null und machen deutlich, dass hier praktisch kein direkter Kausaleffekt vorliegt (also Intervention).

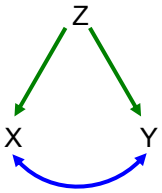
Systematik der Drittvariablenkontrolle

Bestätigung

Z

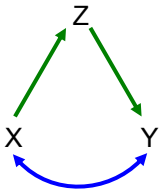
$$r_{XY} \approx r_{XY \bullet Z_1} \approx r_{XY \bullet Z_2} \neq 0, \quad r_{ZY} \approx r_{ZY \bullet X_1} \approx r_{ZY \bullet X_2} = 0$$

Scheinkorrelation



$$r_{XY} \neq 0, \quad r_{XY \bullet Z_1} \approx r_{XY \bullet Z_2} \approx 0$$

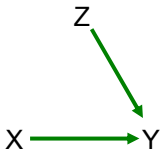
Intervention



$$r_{XY} \neq 0, \quad r_{XY \bullet Z_1} \approx r_{XY \bullet Z_2} \approx 0$$

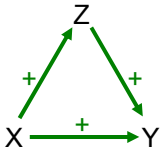
Systematik der Drittvariablenkontrolle

Multikausalität: additiver Effekt



$$r_{XY} \approx r_{XY \bullet Z_1} \approx r_{XY \bullet Z_2} \neq 0, \quad r_{ZY} \approx r_{ZY \bullet X_1} \approx r_{ZY \bullet X_2} \neq 0$$

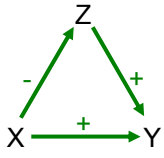
Multikausalität: Konfundierung



$$r_{XY} > r_{XY \bullet Z_1} \approx r_{XY \bullet Z_2}$$

Multikausalität: Suppression

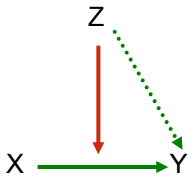
(verdeckte Korrelation)



$$r_{XY} < r_{XY \bullet Z_1} \approx r_{XY \bullet Z_2}$$

Systematik der Drittvariablenkontrolle

- Interaktion
  - Die Beziehung von X und Y fällt unterschiedlich aus, je nachdem welchen Wert Z annimmt (Z heißt auch „Moderator“)



$r_{XY \bullet Z_1} \neq r_{XY \bullet Z_2}$

Beispiele:
- Sport (X), Gesundheit (Y), Erkältung (Z)
- Einsatz (X), Erfolg (Y), Motivation (Z)

Beispiel: M. Halbwachs (1930) Les Causes du Suicide
Halbwachs stellte fest, dass eine Korrelation zwischen Konfession und Selbstmordrate besteht:
Katholiken 19,9 Selbstmorde (pro 100.000),
Protestanten 39,6 Selbstmorde (pro 100.000).
Kontrolliert man den Wohnort, so verschwindet die Korrelation für Städte, auf dem Land nimmt sie sogar noch zu.

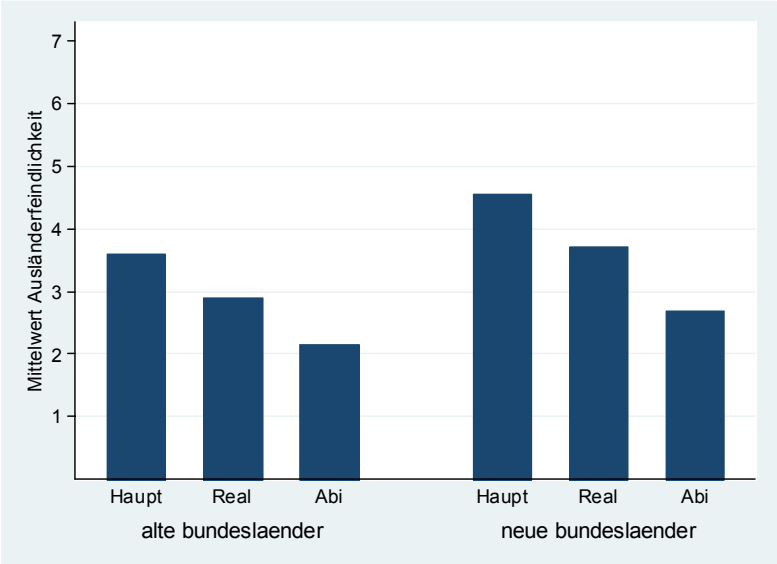
Wohnort	Katholik	Protestant
Stadt	39,9	37,8
Land	8,8	41,4
Alle	19,9	39,6

STATA-Beispiel: Einstellung zu Ausländern

```
********************** BSP DRITTVAR.DO ************************

* Allgemeine Syntax ******************
* by Z, sort: tab Y X, col chi2 V
************************************

* Tabelle: Einstellung zu Ausländern (v70) nach Bildung
codebook       v187
generate       bild = v187
recode bild    1 2=1   3=2    4 5=3   6 7=.
label var      bild "Schulabschluss"
label define   lblbild 1 "Haupt" 2 "Real" 3 "Abi"
label value    bild lblbild

tab            v70 bild, col chi2 g
spearman       v70 bild

* Drittvariablenkontrolle nach West/Ost
by v3, sort: tab    v70 bild, col chi2 g
```

Ergebnis: Der Effekt von Bildung auf die Einstellung zu Ausländern bestätigt sich (nach Kontrolle des Wohnortes)

STATA-Beispiel: Einstellung zu Ausländern



Selbstselektion

- Bei einem Experiment werden die Vpn vom Forscher den beiden Gruppen per Randomisierung zugewiesen
- Bei ex-post facto Designs überlässt man es den Personen selbst, in welche Gruppe sie gehen (Selbstselektion)
- Das führt leicht zu unbeobachteter Heterogenität
- Selbstselektion ist ein allgegenwärtiges methodisches Problem in der Sozialforschung!
- Beispiel: Ehemänner leben länger (W. Krämer, Denkste!)
  Von ledigen Männern zwischen 45 und 54 werden in den nächsten 10 Jahren 23% sterben, von verheirateten Männern nur 11%.
  Hat also die Ehe einen positiven Kausaleffekt auf Gesundheit?
  Nein! Mit ziemlicher Sicherheit haben wir es hier mit Selbstselektion nach der Variable „Gesundheit“ zu tun: Fitte Männer haben höhere Heiratschancen und leben länger.
  Man muss also unbedingt die Drittvariable „Gesundheit“ kontrollieren.

Vorlesung Multivariate Analyse
Kapitel III
Varianzanalyse

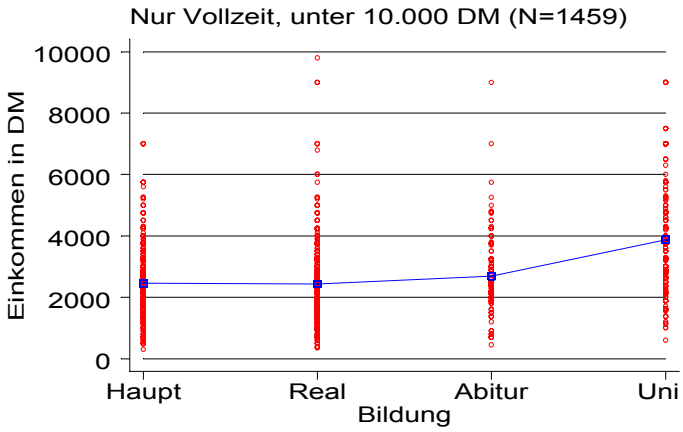
Prof. Dr. Josef Brüderl
Universität Mannheim

Herbstsemester 2007

Varianzanalyse

- Die Varianzanalyse ist ein geeignetes Verfahren, wenn man eine metrische aV und eine (einfaktorielle VA) oder mehrere (mehrfaktorielle) kategoriale uV vorliegen hat
- Im einfaktoriellen Fall liefert die VA insbesondere auch ein Maß für den bivariaten Zusammenhang zwischen einem kategorialen Merkmal und einem metrischen Merkmal
  - sonst v.a. Probleme bei Zushg. zwischen nominalskalierten und metrischen Merkmalen; denkbar: Cramer's V mit Gruppierung, aber: Informationsverlust
- Die Varianzanalyse steht in einem engen Verhältnis zur linearen Regression

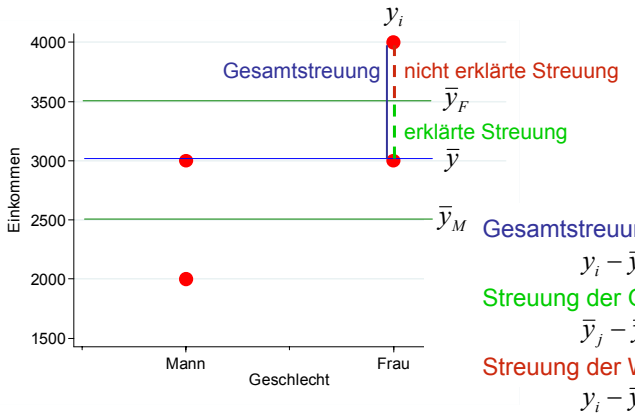
Bildung und Einkommen (ALLBUS 1994)



- Zusammenhang bei der VA heißt, dass sich die Gruppenmittelwerte unterscheiden
  - Er ist stärker, je weniger die Werte innerhalb der Gruppen um den Gruppenmittelwert streuen

Die Streuungszerlegung: Grundidee

- X hat r Ausprägungen (r (Sub)gruppen)
  - n_j bezeichne den Umfang der Gruppe $j \in \{1, ..., r\}$, d.h. $n = \sum_{j=1}^{r} n_j$
  - $\bar{y}_j$ bezeichne das Mittel der Gruppe j
- Die Gesamtstreuung wird zerlegt in zwei Terme:



Gesamtstreuung
$y_i - \bar{y}$

Streuung der Gruppenmittel
$\bar{y}_j - \bar{y}$

Streuung der Werte in den Gruppen
$y_i - \bar{y}_j$

Die Streuungszerlegung

$$(y_i - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_i - \bar{y}_j)$$

- Quadrieren liefert

$$(y_i - \bar{y})^2 = (\bar{y}_j - \bar{y})^2 + 2(\bar{y}_j - \bar{y})(y_i - \bar{y}_j) + (y_i - \bar{y}_j)^2$$

- Summation über alle Mitglieder der Gruppe j:

$$\sum_{i=1}^{n_j}(y_i - \bar{y})^2 = \sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2 + 2(\bar{y}_j - \bar{y})\sum_{i=1}^{n_j}(y_i - \bar{y}_j) + \sum_{i=1}^{n_j}(y_i - \bar{y}_j)^2$$

- Wegen der Schwerpunktregel gilt:

$$2(\bar{y}_j - \bar{y})\sum_{i=1}^{n_j}(y_i - \bar{y}_j) = 0$$

- Summation über alle Gruppen führt zu:

$$\sum_{j=1}^{r}\sum_{i=1}^{n_j}(y_i - \bar{y})^2 = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{r}\sum_{i=1}^{n_j}(y_i - \bar{y}_j)^2$$

$$\text{TSS} \quad = \quad \text{MSS} \quad + \quad \text{RSS}$$

s. Streuungszerlegung bei der Regression

Das Zusammenhangsmaß η^2 (ETA^2)

- η^2 misst den Anteil der durch X erklärten Varianz von Y

$$\eta^2 = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = \frac{\text{MSS}}{\text{TSS}} = \frac{\sum_{j=1}^{r} n_j(\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{r}\sum_{i=1}^{n_j}(y_i - \bar{y})^2}$$

- η^2 hat kein Vorzeichen
- $\eta^2 = 0$ bedeutet, dass die Gruppenzugehörigkeit X keinerlei Streuung der abhängigen Variable Y erklärt (Gruppenmittelwerte sind gleich)
- $\eta^2 = 1$ bedeutet, dass die Gruppenzugehörigkeit X die gesamte Streuung der abhängigen Variable Y erklärt (innerhalb der Gruppen keine Streuung mehr)

Rechenbeispiel

X	Y	$(y_i - \bar{y})^2$	$\bar{y}_j$	$(\bar{y}_j - \bar{y})^2$	$n_j(\bar{y}_j - \bar{y})^2$
1	2	1	3	0	0
1	4	1			
2	3	0	4	1	2
2	5	4			
3	2	1	2	1	2
3	2	1			
	18	8 (=TSS)			4 (=MSS)

$$n = 6$$

$$\bar{y} = \frac{18}{6} = 3$$

$$\eta^2 = \frac{4}{8} = 0{,}5$$

Signifikanztest

- Unterscheiden sich die Gruppenmittelwerte signifikant?
- Man testet:
  H_0: $\mu_1 = \mu_2 = ... = \mu_r$ H_1: $\mu_i \neq \mu_j$ für mind. ein Paar i,j
- Unter der H_0 streuen die Gruppenmittel nicht, also MSS=0
  - Große Werte der MSS sprechen also gegen die H_0
  - Man verwendet deshalb den normierten MSS als Teststatistik

$$F = \frac{\mathrm{MSS}/(r-1)}{\mathrm{RSS}/(n-r)} \sim F(r-1, n-r)$$

- Die H_0 wird verworfen, falls: $F > F_{1-\alpha}(r-1,n-r)$
- Bemerkung: Auch dieser Test beruht auf der Annahme gleicher Varianzen. Diese Annahme kann man ebenfalls testen (Bartlett-Test)

F-Verteilung

- Sind X und Y unabhängige Zufallsvariablen mit X ~ $\chi^2(m)$ und Y ~ $\chi^2(n)$, so ist die Zufallsvariable

$$Z = \frac{X/m}{Y/n}$$

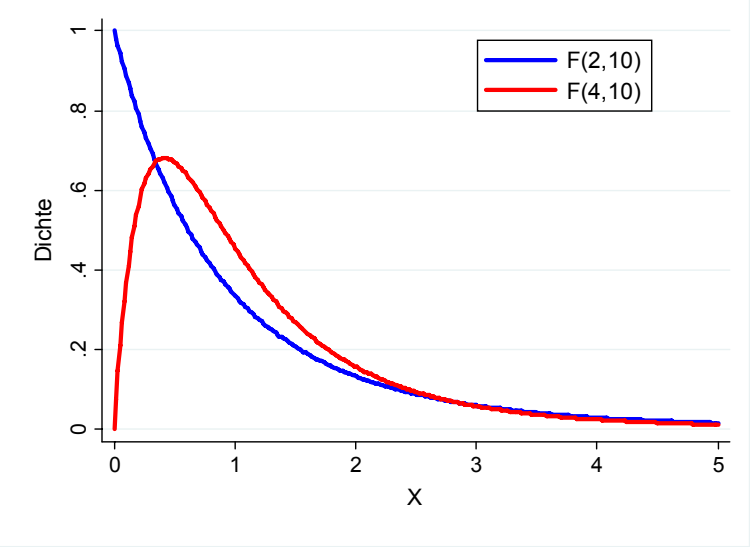
F-verteilt mit m und n Freiheitsgraden

- Man schreibt auch Z ~ F(m,n)
- Es gilt (n>2 bzw. n>4):

$$E(Z) = \frac{n}{n-2}, \quad Var(Z) = \frac{2n^2(n+m-2)}{m(n-4)(n-2)^2}$$

- $F(1,n) = t(n)^2$, $F(1,\infty) = N(0,1)^2$

F-Verteilung



F-Verteilung

Tabelle A–3.4: Ausgewählte Quantile der F–Verteilung

df_2	$df_1 = 1$ 90.0%	95.0%	97.5%	99.0%	99.5%
1	39.86	161.4	647.8	4052	16211
2	8.526	18.51	38.51	98.50	198.5
3	5.538	10.13	17.44	34.12	55.55
4	4.545	7.709	12.22	21.20	31.33
5	4.060	6.608	10.01	16.26	22.78
6	3.776	5.987	8.813	13.75	18.63
7	3.589	5.591	8.073	12.25	16.24
8	3.458	5.318	7.571	11.26	14.69
9	3.360	5.117	7.209	10.56	13.61
10	3.285	4.965	6.937	10.04	12.83
12	3.177	4.747	6.554	9.330	11.75
15	3.073	4.543	6.200	8.683	10.80
30	2.881	4.171	5.568	7.562	9.180
60	2.791	4.001	5.286	7.077	8.495
120	2.748	3.920	5.152	6.851	8.179
∞	2.706	3.841	5.024	6.635	7.879

df_2	$df_1 = 2$ 90.0%	95.0%	97.5%	99.0%	99.5%
1	49.50	199.5	799.5	5000	20000
2	9.000	19.00	39.00	99.00	199.0
3	5.462	9.552	16.04	30.82	49.80
4	4.325	6.944	10.65	18.00	26.28
5	3.780	5.786	8.434	13.27	18.31
6	3.463	5.143	7.260	10.92	14.54
7	3.257	4.737	6.542	9.547	12.40
8	3.113	4.459	6.059	8.649	11.04
9	3.006	4.256	5.715	8.022	10.11
10	2.924	4.103	5.456	7.559	9.427
12	2.807	3.885	5.096	6.927	8.510
15	2.695	3.682	4.765	6.359	7.701
30	2.489	3.316	4.182	5.390	6.355
60	2.393	3.150	3.925	4.977	5.795
120	2.347	3.072	3.805	4.787	5.539
∞	2.303	2.996	3.689	4.605	5.298

$F_{0,90}(1,5) = 4,06$ $F_{0,95}(2,30) = 3,316$

Rechenbeispiel (Fortsetzung)

- MSS = 4
  RSS = TSS – MSS = 8 – 4 = 4
  n = 6
  r = 3
- F = 4/(3-1) / 4/(6-3) = 1,5
- $F_{0,95}(2,3) = 9,552$
- Die Teststatistik ist also kleiner als der kritische Wert
  - Wir können die Nullhypothese nicht zurückweisen
  - Die Gruppenmittelwerte unterscheiden sich nicht signifikant
  - (bei der kleinen Fallzahl ist dieses Ergebnis wenig überraschend)

STATA-Beispiel

```
******************** BSP VARIANZANALYSE.DO ************************

* Nur Westdeutschland, hauptberuflich ganztags, bis Alter 65
keep if  v3   == 1
keep if  v200 == 1
keep if  v185 <  66

* Monatsnettoeinkommen
generate       eink = v361

* Bildung
generate       bild = v187
recode         bild   1 2=1   3=2    4 5=3   6 7=.
replace        bild = 4  if(v196==1 | v197==1)
label define   lblbild 1 "Haupt" 2 "Real" 3 "Abitur" 4 "Uni"
label value    bild lblbild

* Varianzanalyse
oneway eink bild, tabulate
```


STATA-Beispiel

```
. oneway eink bild, tabulate

           |           Summary of eink
      bild |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
     Haupt |   1637.1478    852.5794         230
      Real |   1632.7432    1142.491         222
    Abitur |   1716.8333   947.44769         120
       Uni |   2859.3977   2067.6175         176
-----------+------------------------------------
     Total |   1936.2126   1420.1979         748

                       Analysis of Variance
    Source              SS         df      MS (=SS/df)     F      Prob > F
------------------------------------------------------------------------------
Between groups  MSS   196790859      3     65596953     37.26     0.0000
 Within groups  RSS  1.3099e+09    744   1760591.25
------------------------------------------------------------------------------
    Total       TSS  1.5067e+09    747   2016962.18

Bartlett's test for equal variances:  chi2(3) = 193.1542  Prob>chi2 = 0.000
```

F=MMS/RMS

H_0 verwerfen

H_0 verwerfen
Varianzen ungleich

Vorlesung Multivariate Analyse

Kapitel IV

Lineare Regressionsanalyse

Prof. Dr. Josef Brüderl
Universität Mannheim

Herbstsemester 2007

Lineare Regressionsanalyse

- Die Regressionsanalyse ist das zentrale Analyseverfahren in den Sozialwissenschaften
- Im einfachsten Fall (einfache lineare Regression) handelt es sich um ein Verfahren zur bivariaten Zusammenhangsanalyse metrischer Variablen
- Ihre besondere Bedeutung erlangt die Methode
  - einerseits dadurch, dass sie einen bequemen Übergang zur multivariaten Datenanalyse erlaubt (multiple Regression),
  - andererseits dadurch, dass sie sich durchaus auch auf kategoriale Variablen erweitern lässt (uV: Dummy-Variablen; aV: logistische Regression, ordinales Logitmodell, multinomiales Logitmodell u.a.)

Die einfache Regression

- Sind aV (Regressand) und uV (Regressor) beide metrisch, so kann man zur Zusammenhangsanalyse das lineare Regressionsmodell einsetzen.
- Man formuliert folgendes lineare Modell des Zusammenhangs:

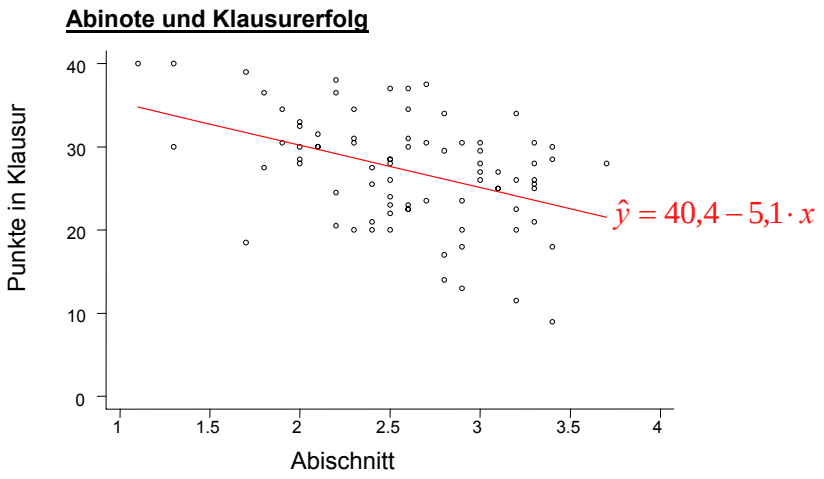
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- α und β sind die Regressionskoeffizienten
  - α: Achsenabschnitt, β: Steigung
  - β: um wie viel Einheiten ändert sich Y, wenn X um eine Einheit steigt
  - β ist damit ein einfach zu interpretierendes Maß des Zusammenhangs
- ε_i ist der Fehlerterm (Abweichung der Daten von der Regressionsgerade)
- Man schätzt die Regressionskoeffizienten, indem man die Fehlerquadratsumme minimiert (OLS)

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Die OLS-Schätzer (Formeln s.u.) notieren wir mit $\hat{\alpha}$ und $\hat{\beta}$
- Die vom Regressionsmodell vorhergesagten Werte sind $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$
- Die geschätzten Fehler (Residuen) sind damit $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$

Beispiel



OLS-Schätzer

- α und β werden so festgelegt, dass die Datenpunkte möglichst wenig von der Regressionsgeraden abweichen
- Präzise: Man schätzt die Regressionskoeffizienten, indem man die Fehlerquadratsumme minimiert (ordinary least squares, OLS)

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \varepsilon_i^{\,2} = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Ableiten dieses Ausdrucks, Nullsetzen und Auflösen der beiden daraus resultierenden Gleichungen, liefert die OLS-Schätzer:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

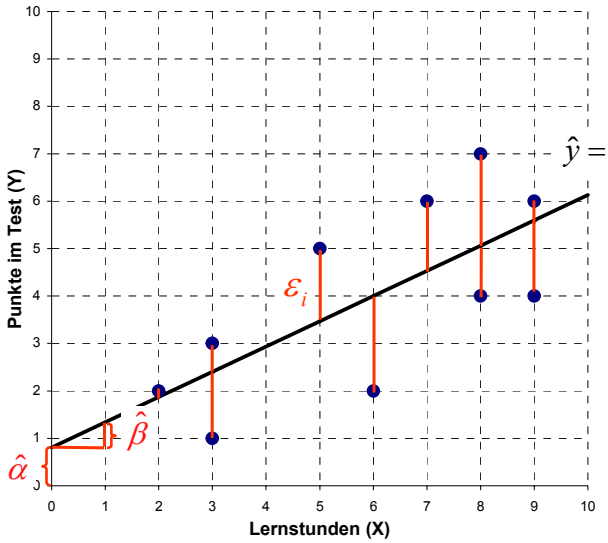
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^{\,2}} \quad = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2}$$

Beispiel: Berechnung der OLS-Schätzer

x_i Lernstunden	y_i Punkte im Test	$x_i - 6$	$y_i - 4$	$(x_i - 6)^2$	$(x_i-6)\cdot(y_i-4)$
2	2	-4	-2	16	8
3	1	-3	-3	9	9
3	3	-3	-1	9	3
5	5	-1	1	1	-1
6	2	0	-2	0	0
7	6	1	2	1	2
8	4	2	0	4	0
8	7	2	3	4	6
9	6	3	2	9	6
9	4	3	0	9	0
60	40			62	33

$$\hat{\beta} = \frac{33}{62} = 0{,}532 \qquad \hat{\alpha} = 4 - 0{,}532 \cdot 6 = 0{,}806$$

Beispiel: Die OLS-Regressionsgerade



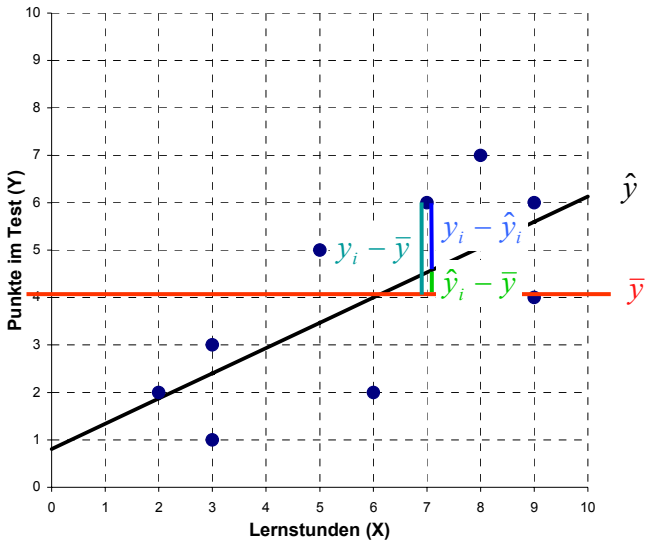
Das Bestimmtheitsmaß R^2

- Wie gut passt das Regressionsmodell auf die Daten?
- Die Grundidee ist: Welcher Anteil der Streuung von Y wird durch das Regressionsmodell „erklärt“?
- Streuungszerlegung
  - Total sum of squares (TSS): $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - Model sum of squares (MSS): $MSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
  - Residual sum of squares (RSS): $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
  - Die gesamte Streuung kann damit in zwei Teile zerlegt werden

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = MSS + RSS$$

Graphische Interpretation der Streuungszerlegung



Das Bestimmtheitsmaß R^2

- Das Bestimmtheitsmaß ist nun definiert als

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

- Es gilt: $0 \leq R^2 \leq 1$
- R^2 lässt sich interpretieren als der Anteil der Varianz, der durch die Regressionsgerade (und damit durch X) erklärt wird
- Es gilt: $R^2 = r^2$

Beispiel

Lernstunden und Punkte im Test

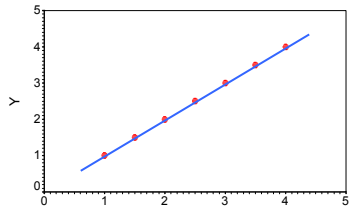
x_i	y_i	$\hat{y}_i$	$(y_i-4)^2$	$(\hat{y}_i-4)^2$
2	2	1.870968	4	4.532777
3	1	2.403226	9	2.549687
3	3	2.403226	1	2.549687
5	5	3.467742	1	0.283299
6	2	4	4	0.000000
7	6	4.532258	4	0.283299
8	4	5.064516	0	1.133194
8	7	5.064516	9	1.133194
9	6	5.596774	4	2.549687
9	4	5.596774	0	2.549687
	40		36	17.564512

$\hat{\alpha}$ = 0.806452

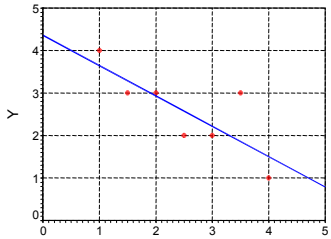
$\hat{\beta}$ = 0.532258

R^2 = 17.56 / 36 = .4879

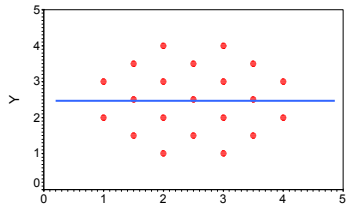
r und R^2



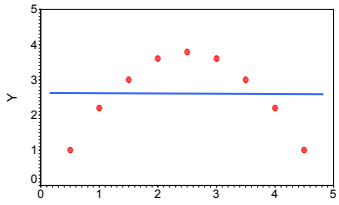
r = 1, R^2 = 1



r = -0.79 , R^2 = 0,62



r = 0 , R^2 = 0



r = 0 , R^2 = 0

Signifikanztest für $\hat{\beta}$

- $\hat{\beta}$ ist ein Schätzer
  - Mit einer Stichprobenverteilung
  - Und einem Standardfehler $\hat{\sigma}_{\hat{\beta}}$
- Damit kann man auch ein Konfidenzintervall schätzen
- Ebenso kann man einen Signifikanztest durchführen
  - Nullhypothese: X hat keinen Einfluss auf Y (kein Zusammenhang)
    H_0: β = 0
  - Die Teststatistik (t-Wert) ist

$$T = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \sim t(n-2)$$

  - Die H_0 wird abgelehnt, falls $|T| > t_{1-\alpha/2}(n-2)$
    - Ab n>30 das $z_{1-\alpha/2}$ Quantil (Faustregel für α=5%: |T| > 2)
  - Können wir die H_0 verwerfen, so spricht man davon, dass X einen signifikanten Einfluss auf Y hat

Test der Signifikanz des Gesamtmodells

- Ist das Regressionsmodell signifikant?
- Overall F-Test
  - Nullhypothese: der Steigungskoeffizient ist gleich null
    H_0: β = 0
- Unter der H_0 erklärt das Modell nichts, also MSS=0
  - Große Werte der MSS sprechen also gegen die H_0
  - Man verwendet deshalb den normierten MSS als Teststatistik (F-Wert)

$$F = \frac{\text{MSS}}{\text{RSS}/(n-2)} \sim F(1, n-2)$$

- Die H_0 wird verworfen, falls: $F > F_{1-\alpha}(1,n-2)$
- Man spricht dann davon, dass das Regressionsmodell signifikant ist

STATA Beispiel: Abinote und Klausurerfolg

```
. regress klpunkte abinote

      Source |       SS       df       MS              Number of obs =      81
-------------+------------------------------           F(  1,    79) =   17.45
       Model |  620.991439     1  620.991439           Prob > F      =  0.0001
    Residual |  2811.32955    79    35.58645           R-squared     =  0.1809
-------------+------------------------------           Adj R-squared =  0.1706
       Total |  3432.32099    80  42.9040123           Root MSE      =  5.9654

------------------------------------------------------------------------------
    klpunkte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     abinote |  -5.085771   1.217464    -4.18   0.000    -7.509073   -2.662468
       _cons |   40.35849   3.229647    12.50   0.000     33.93004    46.78694
------------------------------------------------------------------------------
```


Annahmen der Regression

- A1: Linearitätsannahme

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

- A2: Im Mittel ist der „Fehler“ null

$$E(\varepsilon_i) = 0, \quad \text{für alle i}$$

- A3: Die Fehlervarianz ist konstant (Homoskedastizität)

$$V(\varepsilon_i) = \sigma^2, \quad \text{für alle i}$$

- A4: Die Fehlerkovarianzen sind null (keine Autokorrelation)

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad \text{für alle i} \neq \text{j}$$

- A5: Regressor und Fehler sind unkorreliert
  - Der Regressor darf nicht mit weiteren unbeobachteten Variablen korrelieren (keine unbeobachtete Heterogenität)

$$Cov(x_i, \varepsilon_j) = 0, \quad \text{für alle i und j}$$

- A6: Normalverteilungsannahme (für Signifikanztests)

$$\varepsilon_i \sim N(0, \sigma^2)$$

Eigenschaften der OLS-Schätzer

- Bei Gültigkeit von A1 bis A5 haben die OLS-Schätzer gewisse wünschenswerte Eigenschaften: Sie sind
- unverzerrt (erwartungstreu): $E(\hat{\beta}) = \beta$
- in der Klasse der linearen, unverzerrten Schätzer die mit der kleinsten Stichprobenvarianz
  - best linear unbiased estimate (BLUE)
  - Gauß-Markov Theorem
- Dabei bedeutet:
  - „linear“: die Schätzer lassen sich als lineare Funktionen der Daten berechnen
  - „unbiased“: die Schätzer sind erwartungstreu
  - „best“: die Schätzer sind effizienter als alle anderen linearen Schätzer

Standardisierte Regressionskoeffizienten

- β hängt von der Maßeinheit von X und Y ab
- Um Vergleichbarkeit herzustellen, wählt man manchmal die Standardabweichung als Maßeinheit
  - Standardisierung von Y und X (Z-Transformation)

$$y_i^* = \frac{y_i - \bar{y}}{s_Y}, \quad x_i^* = \frac{x_i - \bar{x}}{s_X}$$

  - Die Regressionsgleichung lautet nun

$$y_i^* = \alpha^* + \beta^* x_i^* + \varepsilon_i^*$$

  - Für die standardisierten Regressionskoeffizienten ergibt sich

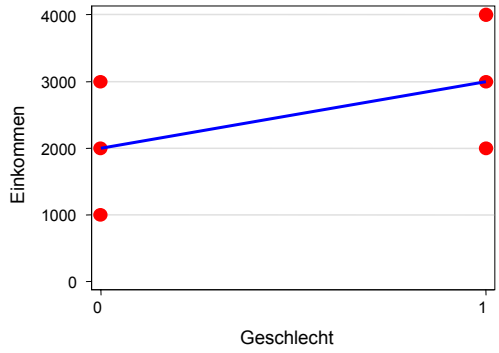
$$\hat{\alpha}^* = \bar{y}^* - \hat{\beta}^* \bar{x}^* = 0$$

$$\hat{\beta}^* = \frac{s_{X^*Y^*}}{s_{X^*}^2} = r$$

  - Beispiel „Abinote und Klausurerfolg“: r = -0,43
    - Steigt die Note um eine Standardabweichung, so verringert sich die Punktzahl um 0,43 Standardabweichungen

Regression mit Dummy-Variable

- Voraussetzung für eine Regression ist metrisches X
- Kategoriales, dichotomes X geht aber auch
  - Dummy-Codierung sinnvoll: eine Ausprägung 0, die andere 1
  - Die 0-Kategorie wird auch als Referenzkategorie bezeichnet
- Beispiel: Geschlecht und Einkommen (0=Mann, 1=Frau)



Wie sieht die Regressionsgerade aus?
RSS wird minimal, falls

$$\hat{y}_0 = \bar{y}_0 = 2000$$

$$\hat{y}_1 = \bar{y}_1 = 3000$$

Gleichzeitig gilt

$$\hat{y}_0 = \hat{\alpha}$$

$$\hat{y}_1 = \hat{\alpha} + \hat{\beta}$$

Damit ergibt sich

$$\hat{\alpha} = \bar{y}_0 = 2000$$

$$\hat{\beta} = \bar{y}_1 - \bar{y}_0 = 1000$$

Regression mit Dummy-Variable

x_i	y_i	x_i^2	$x_i y_i$
0	1000	0	0
0	2000	0	0
0	3000	0	0
1	2000	1	2000
1	3000	1	3000
1	4000	1	4000
3	15000	3	9000

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{9000 - 6 \cdot 0.5 \cdot 2500}{3 - 6 \cdot 0.5^2} = \frac{1500}{1.5} = 1000$$

$$\hat{\alpha} = \bar{y} - b\bar{x} = 2500 - 1000 \cdot 0.5 = 2000$$

STATA: Regression mit Dummy-Variable

```
. oneway Einkommen Geschlecht

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups         1500000      1      1500000      1.50     0.2879
 Within groups         4000000      4      1000000
------------------------------------------------------------------------
    Total              5500000      5      1100000

. regr Einkommen Geschlecht

      Source |       SS       df       MS              Number of obs =       6
-------------+------------------------------           F(  1,     4) =    1.50
       Model |     1500000     1     1500000           Prob > F      =  0.2879
    Residual |     4000000     4     1000000           R-squared     =  0.2727
-------------+------------------------------           Adj R-squared =  0.0909
       Total |     5500000     5     1100000           Root MSE      =    1000

------------------------------------------------------------------------------
   Einkommen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Geschlecht |       1000   816.4966     1.22   0.288    -1266.958    3266.958
       _cons |       2000   577.3503     3.46   0.026     397.0187    3602.981
------------------------------------------------------------------------------
```


Regression mit Dummy-Variable

- Führt man eine Regression auf eine dichotome Variable (Dummy-Codierung) durch, so gilt:
- Die Konstante $\hat{\alpha}$ ist gleich dem Mittelwert der Referenzkategorie (X=0)
- Der Steigungsparameter $\hat{\beta}$ ist gleich der Differenz zwischen dem Mittelwert der Kategorie X=1 und dem Mittelwert der Referenzkategorie
- Das Bestimmtheitsmaß R^2 entspricht dem Maß η^2 einer Varianzanalyse
- Der overall F-Test der Regression und der F-Test der Varianzanalyse sind identisch

Multiple Regression

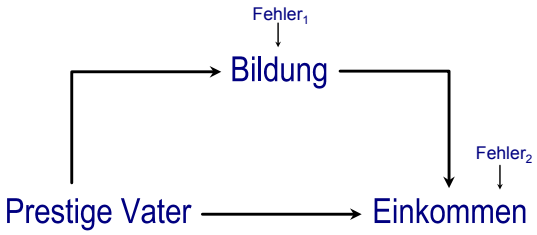
- Die Regression wird zu einem multivariaten Analyse-verfahren, wenn man mehr als eine uV einbezieht:

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

  - 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
  - β_j sagt uns, welcher Effekt verbleibt, wenn wir für die anderen uVs kontrollieren
  - Multivariate Analyse
    - Wir fügen dem Modell alle sinnvollen Drittvariablen hinzu
    - Verschwindet β_j, so liegt eine Scheinkorrelation bzw. Intervention vor

Beispiel: Statuszuweisungsmodell

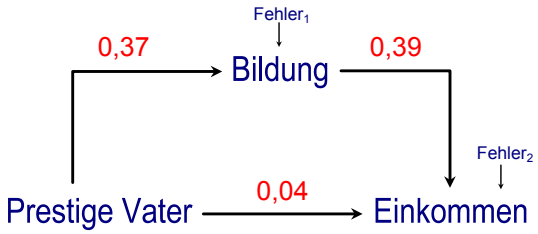
- Blau/Duncan (1967) "The American Occupational Structure"
  - Wie erlangt man seine soziale Position? Durch „achievement“ oder Statusvererbung?
  - ALLBUS 2002:
    - Abhängige Variable: Einkommen (nur Westdeutsche, Vollzeit)
    - Status des Vaters: Magnitude-Prestigeskala (Werte von 20-187)
    -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



Beispiel: Statuszuweisungsmodell

	(1)	(2)
Konstante	1343	-585
Prestige Vater	9,6	2,0
Bildung		180
R^2	0,03	0,16
N	670	670

Bivariates Modell (1)
"Prestige Vater" hat einen starken Effekt
Multiples Modell (2)
Der wird unter Kontrolle der "Bildung" deutlich kleiner (Intervention)



Pfaddiagramm
Hier sind die standardisierten Regressionskoeff. eingetragen.

Die trivariate Regression

- Ähnlich wie im bivariaten Fall werden die Koeffizienten nach der Methode der kleinsten Quadrate bestimmt
- Im trivariaten Fall, d.h. wenn eine abhängige Variable Y und zwei unabhängige Variablen X_1 bzw. X_2 vorliegen, ergibt sich

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2$$

$$\hat{\beta}_1 = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \frac{s_Y}{s_{X_1}}$$

$$\hat{\beta}_2 = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \frac{s_Y}{s_{X_2}}$$

standardisierter Regressionskoeff. $\hat{\beta}^*_j = \hat{\beta}_j \frac{s_{X_j}}{s_Y}$

STATA Beispiel: Statuszuweisungsmodell

```
************** Bsp MultipleRegr I.do ***********************

* Nur Westdeutschland, hauptberuflich ganztags, bis Alter 65
keep if  v3   == 1
keep if  v200 == 1
keep if  v185 <  66

* Monatsnettoeinkommen
generate     eink   = v361
* Magnitudeprestige Vater
generate     prestv = v312
* Bildung in Jahren
recode   v187 1=8 2=9 3=10 4=12 5=13 6=10 7=.
generate bild=v187
replace  bild=bild+0.5 if (v188==1)
replace  bild=bild+1   if (v189==1)
replace  bild=bild+1.5 if (v190==1)
replace  bild=bild+1.5 if (v191==1)
replace  bild=bild+0.5 if (v192==1)
replace  bild=bild+1.5 if (v193==1)
replace  bild=bild+2   if (v194==1)
replace  bild=bild+3   if (v195==1)
replace  bild=bild+4   if (v196==1)
replace  bild=bild+5   if (v197==1)
replace  bild=bild+0.5 if (v198==1)
```


STATA Beispiel: Statuszuweisungsmodell

```
. corr         eink prestv bild
(obs=670)

             |    eink   prestv     bild
             |    (Y)     (X1)      (X2)
-------------+---------------------------
    eink (Y) |   1.0000
  prestv (X1)|   0.1831   1.0000
    bild (X2)|   0.4023   0.3739   1.0000

. summ eink prestv bild if e(sample)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        eink |       670    1905.813    1404.526         150      15200
      prestv |       670    58.55672    26.76102          20      186.8
        bild |       670    13.17537    3.024599           8       23.5
```

$$\hat{\beta}_2 = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \frac{s_Y}{s_{X_2}} = \frac{0{,}40 - 0{,}18 \cdot 0{,}37}{1 - 0{,}37^2} \frac{1405}{3} = 180{,}9$$

$$\hat{\beta}^*_2 = \hat{\beta}_2 \frac{s_{X_2}}{s_Y} = \frac{0{,}40 - 0{,}18 \cdot 0{,}37}{1 - 0{,}37^2} = 0{,}386$$

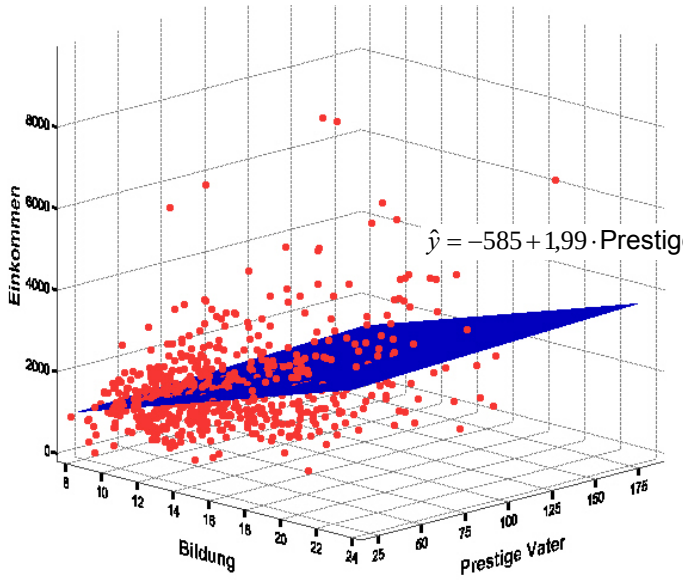
STATA Beispiel: Statuszuweisungsmodell

```
. regress      eink prestv bild, beta

      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|                     Beta
-------------+----------------------------------------------------------------
      prestv |     1.99162   2.004457     0.99   0.321                .0379471
        bild |    180.2345   17.73502    10.16   0.000                .3881287
       _cons |   -585.4664   224.3659    -2.61   0.009                       .
------------------------------------------------------------------------------
```

Der über die Korrelationskoeffizienten errechnete Regressionskoeffizient (180,9), ist bis auf Rundungsfehler identisch mit dem OLS-Schätzer (180,2)

Standardisierte Regressionskoeffizienten

Regressionsebene



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

Multiples R^2

- Die vorhergesagten Werte ergeben sich aus

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

z.B.(Prestige = 100, Bildung = 10): $-585 + 1{,}99 \cdot 100 + 180 \cdot 10 = 1414$

- 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: $F > F_{1-\alpha}(p, n-p-1)$

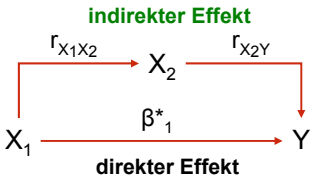
Was bedeutet „Kontrolle“?

- β_j gibt den 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 indirekten Effekten „bereinigt“

bivariater Effekt — indirekter Effekt

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

Standardisierung



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

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

  - Der multiple Regressionskoeffizient ist gleich dem Bivariaten

Partielle Korrelation

- Eng verwandt mit dem standardisierten Regressionskoeff. ist der partielle Korrelationskoeffizient

$$r_{YX_1 \bullet X_2} = \frac{r_{YX_1} - r_{X_1X_2} r_{YX_2}}{\sqrt{(1 - r_{X_1X_2}^2)(1 - r_{YX_2}^2)}}$$

  - Auch hier wird der indirekte Effekt herausgerechnet
  - Nur die Standardisierung ist anders

```
. pcorr        eink prestv bild
(obs=670)

Partial correlation of eink with

    Variable |    Corr.      Sig.
-------------+------------------
      prestv |   0.0384    0.321
        bild |   0.3662    0.000
```

Mittels des partiellen Korrelationskoeffizienten

$r_{YX \bullet Z}$

Kann man ebenso wie mit den konditionalen Korrelationskoeffizienten

$r_{YX \bullet z_1}$, $r_{YX \bullet z_2}$

Kausalanalyse betreiben.

Vorteil: nur eine Kennzahl

Nachteil: Interaktionen können nicht erkannt werden (gilt auch für Regression!)

Multiple Regression

- Auch wenn mehr als zwei unabhängige Variablen (p>2) berücksichtigt werden, können die (p+1) Koeffizienten der Regressionsgleichung mit Hilfe der Methode der kleinsten Quadrate bestimmt werden
- Die Lösungsformeln werden hier sehr unübersichtlich; in der Regel wird daher die Matrix-Schreibweise verwendet (siehe: „Regressionsanalyse" -> M.A. Soziologie)
- Es gilt:

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

mit :

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

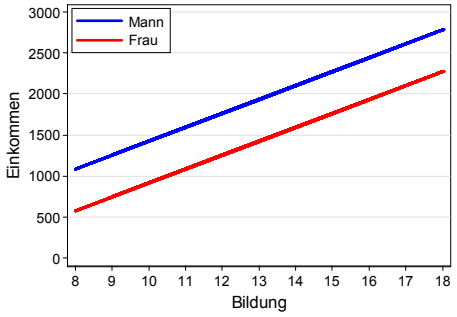
Gliederung der folgenden Abschnitte

- Praktische Regressionsanalyse
  - Regression mit Dummies
  - Interaktionseffekte in der Regression
- Regressionsdiagnostik
  - Linearitätsannahme
  - Homoskedastizität
  - Normalverteilungsannahme
  - Ausreißerdiagnostik

Kovarianzanalyse

```
********** Bsp MultipleRegr II.do ******************
. regress eink bild frau

R-squared     =  0.1722
------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------
        bild |   169.2551   12.44852    13.60   0.000
        frau |  -511.2477   77.71202    -6.58   0.000
       _cons |  -264.8927   172.6304    -1.53   0.125
------------------------------------------------------------

. gen einkmann = _b[_cons] + _b[bild]*bild
. gen einkfrau = _b[_cons] + _b[bild]*bild + _b[frau]
```

Eine muliple Regression, bei der eine der uVs eine Dummy ist
Beispiel: Einkommen in Abhängigkeit von Bildung und Geschlecht (ALLBUS 2002)



Conditional-Effect Plot:

$$\hat{y}_M = -265 + 169 \cdot \text{Bild}$$
$$\hat{y}_F = -776 + 169 \cdot \text{Bild}$$

Kategoriale uV mit mehr als 2 Ausprägungen

- Auch solche Variablen können bei der multiplen Regression als uV benutzt werden
- 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 Koffizienten β_j der Dummies geben den Mittelwertunterschied der betreffenden Kategorie zur Referenzkategorie an

Generierung der Dummies

```
recode v201 1/3=4 4=3 5=2 6=1 7/9=.
generate beruf = v201
label define lblber 1 "Arbeiter" 2 "Angestellter" 3 "Beamter" 4 "Selbständiger"
label value  beruf lblber

. 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
```


Regression mit 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
-------------------------
```

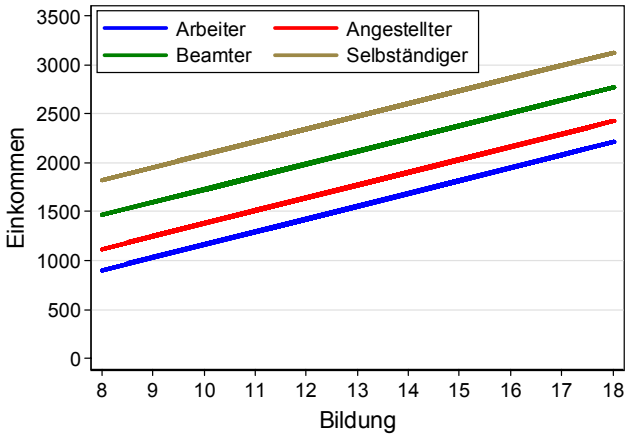
```
. regr eink d2 d3 d4

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

Regression mit Dummies

```
. rename d2 angest
. rename d3 beamt
. rename d4 selbst
. regress eink bild angest beamt selbst

      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
      angest |   216.6888   95.98289     2.26   0.024
       beamt |   562.1304   171.8194     3.27   0.001
      selbst |   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
 _cons |  1333
------------------
```


Regression mit Dummies




```
--------------------
  eink |      Coef.
-------+------------
  bild |   130.5443
angest |   216.6888
 beamt |   562.1304
selbst |   919.5047
 _cons |  -141.8791
--------------------
```

Berücksichtigung von Interaktionseffekten

- Der Effekt von z.B. X_1 hängt vom Wert von X_2 ab
- Berücksichtigung in einer Regression
  - Multipliziere X_1 und X_2
  - Nimm diese Produktvariable in die Regression auf
  - Hierbei unterstellt man eine multiplikative Interaktion

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3(x_1 \cdot x_2)+\varepsilon$$

$$\frac{\partial E(y)}{\partial x_1} = \beta_1+\beta_3 x_2$$

$$x_2 = 0: \frac{\partial E(y)}{\partial x_1} = \beta_1$$

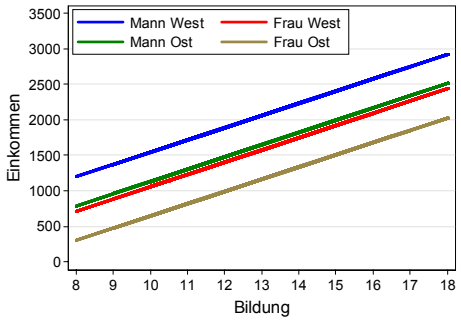
$$x_2 = 1: \frac{\partial E(y)}{\partial x_1} = \beta_1+\beta_3$$

Dummy-Interaktion: Geschlecht/Wohnort

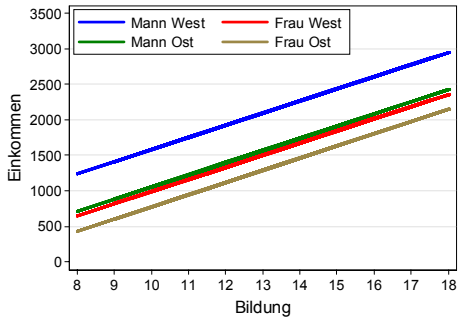
```
. * Bildung der Interaktionsvariable: Ost/Frau
. generate  ost_fr=ost*frau
. * ohne Interaktion
. regress eink bild frau ost
------------------------------------------------------------------------------
       eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
       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 frau ost ost_fr
      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  4,  1113) =   67.27
       Model |   404835698     4   101208925           Prob > F      =  0.0000
    Residual |  1.6746e+09  1113  1504570.35           R-squared     =  0.1947
-------------+------------------------------           Adj R-squared =  0.1918
       Total |  2.0794e+09  1117  1861613.69           Root MSE      =  1226.6

------------------------------------------------------------------------------
       eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
       bild |   171.3628   12.31697    13.91   0.000     147.1957    195.5299
       frau |    -592.12   95.05008    -6.23   0.000    -778.6176   -405.6225
        ost |  -526.8585   99.18368    -5.31   0.000    -721.4665   -332.2504
     ost_fr |   311.5265    161.903     1.92   0.055    -6.143091    629.196
      _cons |  -132.5139   173.1033    -0.77   0.444    -472.1594    207.1317
------------------------------------------------------------------------------
```

Dummy-Interaktion: Geschlecht/Wohnort

ohne Interaktion



mit Interaktion



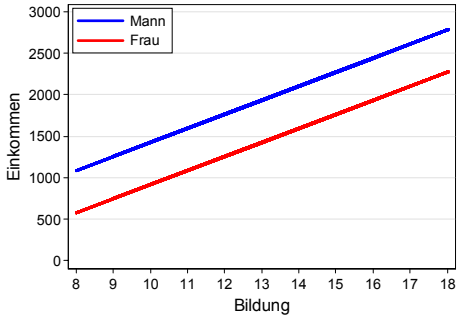
Designmatrix	Frau -592	Ost -527	Ost_fr 312	Einkommens-unterschied	
Mann West	0	0	0	0	Referenzgruppe
Mann Ost	0	1	0	-527	
Frau West	1	0	0	-592	
Frau Ost	1	1	1	-807	

Slope-Interaktion: Geschlecht/Bildung

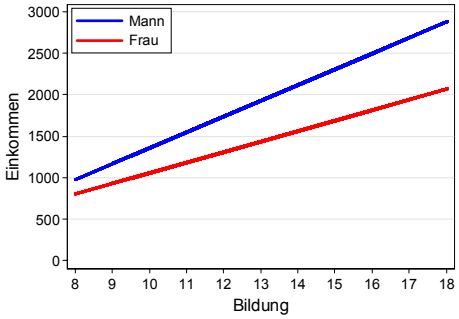
```
. * Bildung der Interaktionsvariable: Frau/Bildung
. generate  fr_bild=frau*bild
. * ohne Interaktion
. regress eink bild frau

------------------------------------------------------------------------------
        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 bild frau fr_bild

      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]
-------------+----------------------------------------------------------------
        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
     fr_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
------------------------------------------------------------------------------
```

Slope-Interaktion: Geschlecht/Bildung

ohne Interaktion



mit Interaktion



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

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:
  - A1: Linearität
  - A3: Homoskedastizität
  - A6: Normalverteilungsannahme
  - zusätzlich: Ausreißerdiagnostik
- Dazu analysiert man die Residuen
  - Die Fehlerterme selbst sind nicht beobachtbar
  - Die Residuen sind allerdings Schätzer für die Fehlerterme

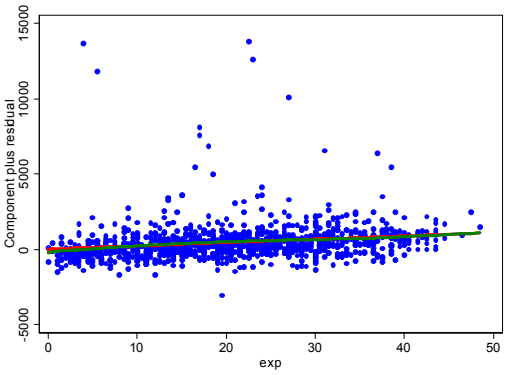
$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Linearität

- 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
generate    exp  = v185-bild-6
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

- Abhilfe bei Vorliegen von Nicht-Linearität
  - Variable gruppieren: Dummies
  - Polynomregression: $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_2^2+\varepsilon$

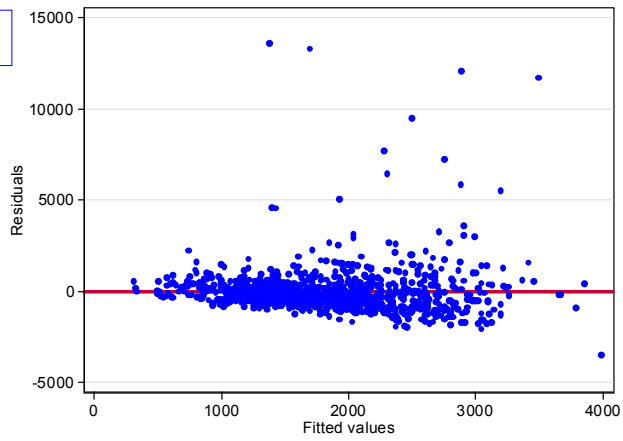
```
. generate    exp2 = exp*exp
. regress eink bild exp exp2 frau

      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  4,  1113) =   71.72
       Model |   426153858     4   106538464           Prob > F      =  0.0000
    Residual |  1.6533e+09  1113  1485416.57           R-squared     =  0.2049
-------------+------------------------------           Adj R-squared =  0.2021
       Total |  2.0794e+09  1117  1861613.69           Root MSE      =  1218.8

------------------------------------------------------------------------------
        eink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |   178.1442   12.53227    14.21   0.000     153.5546    202.7337
         exp |   42.08305   12.68307     3.32   0.001     17.19763    66.96847
        exp2 |  -.4741049   .2907692    -1.63   0.103    -1.044622    .0964126
        frau |  -468.5161   76.52427    -6.12   0.000    -618.6642    -318.368
       _cons |  -1002.292   205.6562    -4.87   0.000     -1405.81    -598.775
------------------------------------------------------------------------------
```

Homoskedastizität

- 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.)

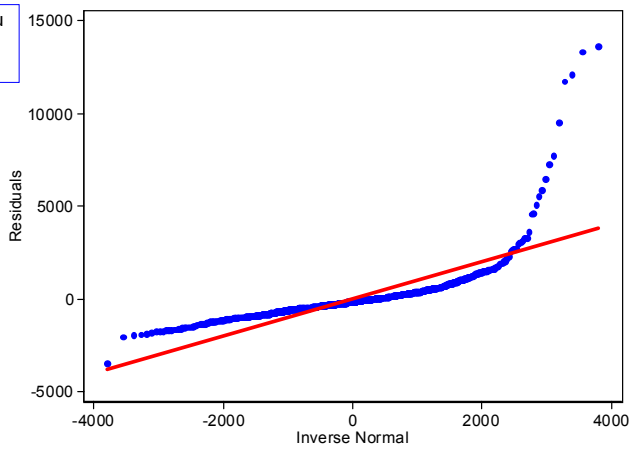


Normalverteilungsannahme

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

```
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

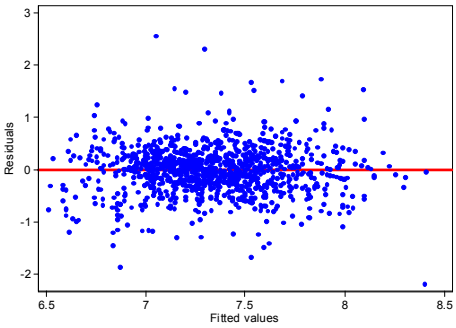
```
. * logarithmische Transformation der aV
. generate lneink = ln(eink)
. regress lneink bild exp frau

      Source |       SS       df       MS              Number of obs =    1118
-------------+------------------------------           F(  3,  1114) =  188.24
       Model |  117.736293     3  39.2454309            Prob > F      =  0.0000
    Residual |  232.249663  1114  .208482642           R-squared     =  0.3364
-------------+------------------------------           Adj R-squared =  0.3346
       Total |  349.985956  1117  .313326729           Root MSE      =   .4566

------------------------------------------------------------------------------
      lneink |      Coef.   Std. Err.      t    P>|t|     exp(Coeff.)-1
-------------+----------------------------------------------------------------
       bild |     .08873   .0046284    19.17   0.000     0.093
        exp |     .01564   .0012635    12.38   0.000     0.016
       frau |  -.2568402    .028635    -8.97   0.000    -0.226
      _cons |   5.922547   .0728721    81.27   0.000
------------------------------------------------------------------------------
```

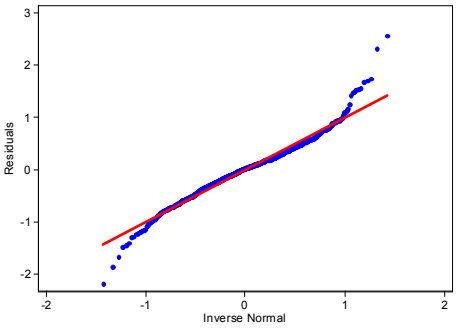
$$\ln(y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$$

$$y = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon}$$

$e^{\beta_j}-1$ ist die prozentuale Veränderung von Y bei Erhöhung von X um eine Einheit.

Semi-logarithmische Einkommensregression





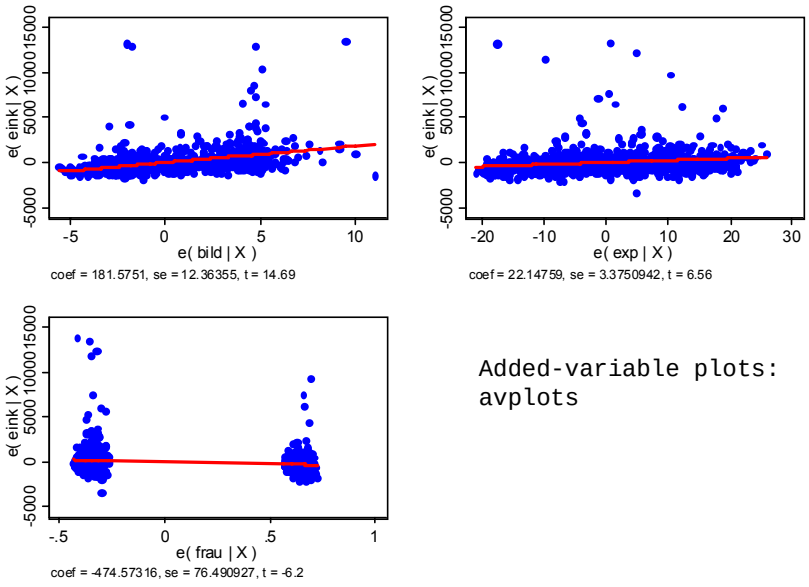
Kein augenfälliges Muster erkennbar: Homoskedastie

Nur mehr geringe Abweichungen an den Rändern: Residuen fast normalverteilt

Ausreißerdiagnostik

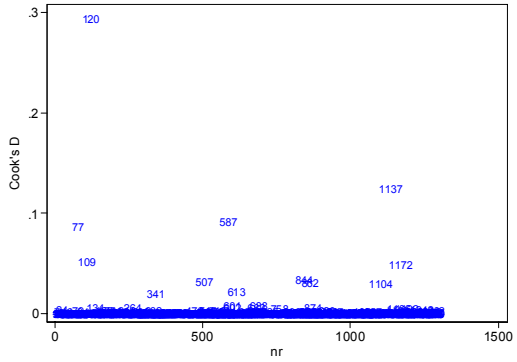
- Ein Datenpunkt ist einflussreich, wenn seine Beseitigung die Ergebnisse der Regression deutlich verändert
  - Fälle mit ungewöhnlichem X- und Y-Wert (Ausreißer) haben Einfluss
  - Problem: das Ergebnis repräsentiert evtl. nur wenige Ausreißer
- Einflussdiagnostik
  - Im Streudiagramm erkennt man einflussreiche 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




```
Added-variable plots:
avplots
```

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 selbst if D>0.1 & D~=.

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

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

Humankapitalmodell

```
. regress lneink prestv bild exp exp2 frau ost angest beamt selbst

      Source |       SS       df       MS              Number of obs =     948
-------------+------------------------------           F(  9,   938) =   63.35
       Model |  91.0060449     9  10.1117828           Prob > F      =  0.0000
    Residual |  149.711276   938  .159606904           R-squared     =  0.3781
-------------+------------------------------           Adj R-squared =  0.3721
       Total |  240.717321   947  .254189356           Root MSE      =  .39951

------------------------------------------------------------------------------
      lneink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prestv |   .0001876   .0005008     0.37   0.708    -.0007953    .0011705
        bild |   .0570828   .0054071    10.56   0.000     .0464713    .0676942
         exp |   .0267595   .0048941     5.47   0.000     .0171548    .0363642
        exp2 |  -.0003813   .0001105    -3.45   0.001    -.0005983   -.0001644
        frau |   -.282162   .0285084    -9.90   0.000    -.3381096   -.2262143
         ost |  -.2236793   .0281563    -7.94   0.000    -.2789361   -.1684226
      angest |   .2304521   .0345453     6.67   0.000     .1626571    .2982471
       beamt |   .3816213   .0583679     6.54   0.000     .2670745     .496168
      selbst |   .2912989   .0503601     5.78   0.000     .1924673    .3901304
       _cons |   6.248556    .081503    76.67   0.000     6.088607    6.408505
------------------------------------------------------------------------------
```

Vorlesung Multivariate Analyse
Kapitel V
Logistische Regression

Prof. Dr. Josef Brüderl
Universität Mannheim

Herbstsemester 2007

Regression mit kategorialer aV

- Bisher gingen wir davon aus, dass die aV metrisch ist.
- Bei kategorialen aVs kommen spezielle Regressionsverfahren zum Einsatz
  - Ordinale aV: ordinale logistische Regression (ordinales Logit)
  - Nominale aV: multinomiale logistische Regression
  - Dichotome aV: (binäre) logistische Regression (Logit)
- Die aV ist sinnvollerweise 0/1-kodiert

- Beispiele:
  - Wer wählt CDU?
  - Wer wird arbeitslos?
  - Wer macht eine Uni-Ausbildung?

Das lineare Wahrscheinlichkeitsmodell

- Man kann eine dichotome aV linear modellieren

$$E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \boldsymbol{\beta}'\mathbf{x}$$

- Da gilt

$$E(y) = P(Y=0)\cdot 0 + P(Y=1)\cdot 1 = P(Y=1)$$

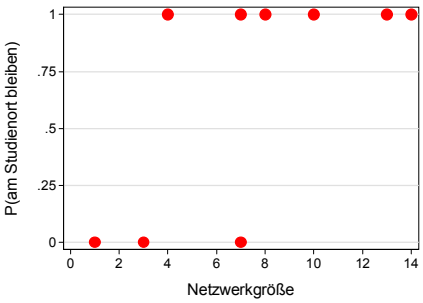
- Erhält man das lineare Wahrscheinlichkeitsmodell (LWM)

$$P(Y=1) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \boldsymbol{\beta}'\mathbf{x}$$

- Fiktives Bsp.: Bleiben am Studienort in Abhängigkeit von der Netzwerkgröße

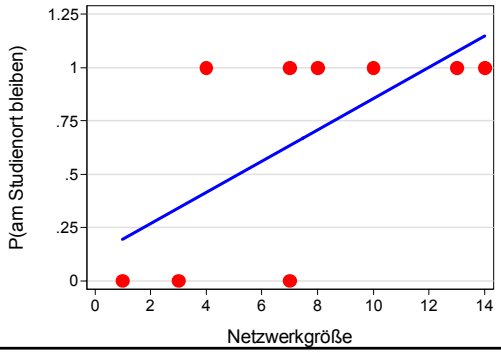
Y=1: Student will am Studienort nach Beendigung des Studiums bleiben.

Y=0: Student will Studienort nach Beendigung des Studiums verlassen.



Das lineare Wahrscheinlichkeitsmodell

```
. regress bleiben netzgr
      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =    5.00
       Model |  .832853026     1  .832853026           Prob > F      =  0.0605
    Residual |  1.16714697     7  .166735282           R-squared     =  0.4164
-------------+------------------------------           Adj R-squared =  0.3331
       Total |           2     8         .25           Root MSE      =  .40833

------------------------------------------------------------------------------
     bleiben |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      netzgr |    .073487   .0328806     2.23   0.061    -.0042633    .1512374
       _cons |   .1195965   .2800758     0.43   0.682    -.5426775    .7818706
------------------------------------------------------------------------------
```



Kennt man niemanden (netzgr=0), so ist die Bleibewhs. gleich 12%. Mit jedem zusätzlichen Bekannten steigt sie um 7 Prozentpunkte.

Folge: bei 12 Bekannten liegt die Bleibewhs. über Eins!

Das logistische Modell

- Offensichtlich führt das LWM in manchen Fällen zu unsinnigen Prognosen
  - Außerdem sind die Fehler heteroskedastisch
- Zur Modellierung von Wahrscheinlichkeiten ist es sinnvoll, Funktionen zu verwenden, deren Wertebereich [0,1] ist
  - Normalveteilungsfunktion: Probit-Modell
  - Logistische Verteilungsfunktion: Logit-Modell

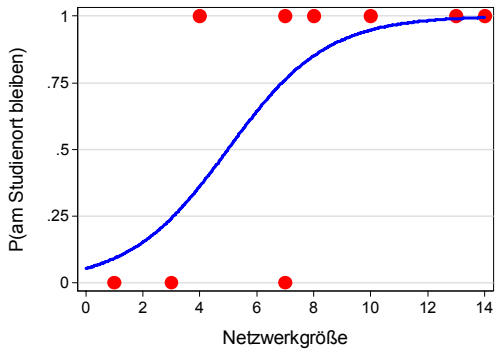
$$P(Y=1)=\frac{e^{\beta' x}}{1+e^{\beta' x}}=\frac{1}{1+e^{-\beta' x}}$$

$$P(Y=0)=1-P(Y=1)=\frac{1}{1+e^{\beta' x}}$$

Interpretation der Koeffizienten nicht einfach (s.u.).
Vorzeicheninterpretation: ein positiver Koeffizient besagt, dass mit steigendem X P(Y=1) zunimmt.

Das logistische Modell

```
. logit bleiben netzgr

Logistic regression                                     Number of obs   =          9
------------------------------------------------------------------------------
     bleiben |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      netzgr |   .5769341   .3808377     1.51   0.130    -.1694942    1.323362
       _cons |  -2.876696   2.319757    -1.24   0.215    -7.423335    1.669943
------------------------------------------------------------------------------
```



$$\hat{\beta}' x = -2.88 + 0.58 \cdot x$$

Das Modell macht nun Sinn.

Der Effekt der Netzwerkgröße ist positiv, d.h. je mehr Bekannte, desto höher P(Bleiben).

Maximum-Likelihood Schätzung

- OLS bei nominalskalierter aV nicht anwendbar
  - „Abstandsquadrate“ sind nicht sinnvoll
- Schätzung mittels Maximum-Likelihood (ML)
  - Daten: (y_i, x_i)
  - Regressionsmodell: f(Y=y_i|x_i,**ß**)
  - Schätzprinzip: Bestimme ß so, dass die Wahrscheinlichkeit diese Daten zu beobachten, maximal wird
  - Die Wahrscheinlichkeit (Likelihood) der Daten unter dem gegebenen Modell und unabhängiger Stichprobenziehung ist

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i, x_i; \boldsymbol{\beta})$$

  - Für die Berechnung ist es vorteilhaft, die Log-Likelihood zu maximieren

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln f(y_i, x_i; \boldsymbol{\beta})$$

  - Ableiten und Null-Setzen liefert die ML-Schätzer

ML-Schätzer des binären Logit-Modells

- Anwendung auf das binäre Logit-Modell
  - Whs. Y=1: P(Y=1)
  - Whs. Y=0: P(Y=0)
  - Likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \left[\frac{e^{\boldsymbol{\beta}'\mathbf{x_i}}}{1+e^{\boldsymbol{\beta}'\mathbf{x_i}}} \right]^{y_i} \cdot \left[\frac{1}{1+e^{\boldsymbol{\beta}'\mathbf{x_i}}} \right]^{(1-y_i)} \right\}$$

  - Logarithmieren, Ableiten und Null-Setzen liefert die Schätzgleichungen

$$\sum_{i=1}^{n} y_i \mathbf{x_i} = \sum_{i=1}^{n} \frac{e^{\hat{\boldsymbol{\beta}}'\mathbf{x_i}}}{1+e^{\hat{\boldsymbol{\beta}}'\mathbf{x_i}}} \mathbf{x_i}$$

  - Dies ist ein nicht-lineares Gleichungssystem, Lösung deshalb mittels iterativer numerischer Algorithmen

Signifikanztests

- Test eines einzelnen Regressionskoeffizienten
  - Nullhypothese: X_j hat keinen Einfluss auf Y (kein Zusammenhang)
    $H_0: \beta_j = 0$
  - Die Teststatistik (z-Wert) ist $Z = \frac{\hat{\beta}_j}{\hat{\sigma}_j} \overset{a}{\sim} N(0,1)$
  - Die H_0 wird abgelehnt, falls $|Z| > z_{1-\alpha/2}$
    - Ab n>100 sinnvoll (Faustregel für α=5%: |Z| > 2)
- Test des gesamten Modells: Likelihood-Ratio Test
  - Nullhypothese: keine X-Variable hat einen Einfluss auf Y
    $H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$
  - Die Teststatistik (LR-Wert) ist $\chi^2 = -2\ln\left(\frac{L_0}{L_1}\right) = 2\left(\ln L_1 - \ln L_0\right)$
    - L_0: Likelihood des Modells nur mit Konstante (Nullmodell)
    - L_1: Likelihood des Gesamtmodells
  - Die H_0 wird verworfen, falls: $\chi^2 > \chi^2_{1-\alpha}$ (p)

Modellfit: Pseudo-R^2

- R^2 nicht sinnvoll, da keine sinnvolle Streuungszerlegung
- In Analogie: Pseudo-R^2 Maße
  - Wie viel von der Likelihood des Nullmodells wird durch das Gesamtmodell „erklärt“
    - Null, wenn die weiteren X-Variablen nichts erklären
    - Maximum allerdings kleiner Eins
  - McFadden Pseudo-R^2

$$R^2_{MF} = \frac{\ln L_0 - \ln L_1}{\ln L_0}$$

    - nicht: Anteil erklärter Varianz
    - Relative Log-Likelihood Verbesserung (im Vergleich zum Nullmodell)
    - Fällt eher kleiner aus, als das R^2 des LWM

Interpretation

- Das Logit-Modell hat drei äquivalente Formulierungen:

$$\text{Whs.:} \quad P(Y=1) = \frac{e^{\beta' \mathbf{x}}}{1+e^{\beta' \mathbf{x}}}$$

$$\text{Odds:} \quad \frac{P(Y=1)}{P(Y=0)} = e^{\beta' \mathbf{x}}$$

$$\text{Logit:} \quad \ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \boldsymbol{\beta}' \mathbf{x}$$

- Deshalb drei mögliche Interpretationen:
  - $ß_j$ ist der lineare, additive Effekt auf das Logit (unverständlich)
  - $\exp(ß_j)$ ist der multiplikative Effekt auf die Odds (komplex)
    - $\exp(ß_j \cdot (x+1)) = \exp(ß_j \cdot x) \cdot \exp(ß_j)$
  - Wahrscheinlichkeitseffekte am anschaulichsten, müssen aber ausgerechnet werden und sind vom Wert der X-Variablen abhängig
    - Am einfachsten: Vorzeicheninterpretation

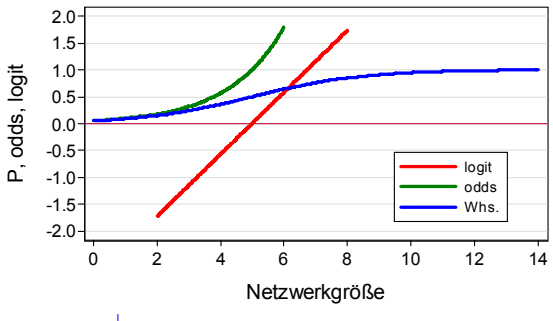
Beispiel: Bleiben am Studienort

```
. drop _all
. set obs 15
. generate netzgr = _n - 1
. generate logit = -2.876696 + 0.5769341*netzgr
. generate odds  = exp(logit)
. generate Whs   = exp(logit) / (1+exp(logit))
. format Whs odds logit %6.2f
. list netzgr Whs odds logit, noobs

  +-------------------------------+
  | netzgr    Whs     odds   logit |
  |-------------------------------|
  |      0   0.05     0.06   -2.88 |
  |      1   0.09     0.10   -2.30 |
  |      2   0.15     0.18   -1.72 |
  |      3   0.24     0.32   -1.15 |
  |      4   0.36     0.57   -0.57 |
  |-------------------------------|
  |      5   0.50     1.01    0.01 |
  |      6   0.64     1.79    0.58 |
  |      7   0.76     3.20    1.16 |
  |      8   0.85     5.69    1.74 |
  |      9   0.91    10.13    2.32 |
  |-------------------------------|
  |     10   0.95    18.04    2.89 |
  |     11   0.97    32.12    3.47 |
  |     12   0.98    57.20    4.05 |
  |     13   0.99   101.84    4.62 |
  |     14   0.99   181.34    5.20 |
  +-------------------------------+
```

$$\text{logit} = -2.88 + 0.58 \cdot x$$

$$\text{odds} = e^{\text{logit}}$$

$$\text{Whs} = \frac{e^{\text{logit}}}{1+e^{\text{logit}}}$$



Beispiel: Bleiben am Studienort

- Logit-Interpretation
  - Das Logit steigt um 0,58 mit jedem zusätzlichen Bekannten
  - Das ist sehr unanschaulich
- Odds-Interpretation
  - Das Odds (die „Chance“ zu Bleiben) nimmt mit jedem Bekannten um den Faktor exp(0,58) = 1,78 zu
    - Bsp.: Odds(6) = 1,79 = 1,01•1,78 = Odds(5) •1,78
    - Auch Prozentinterpretation: exp($ß_j$)-1
      Das odds nimmt mit jedem Bekannten um 78% zu
    - Falsch: die Whs. erhöht sich um 78%!
  - Odds („Chancen“) sind leider auch ziemlich unanschaulich
- Whs.-Interpretation
  - Der Effekt hängt von X ab
    - Bsp.: netzgr=5; die Whs. des Bleibens steigt um 14 Prozentpunkte

$$P(Y=1|6)-P(Y=1|5)=\frac{1}{1+e^{2.88-0.58\cdot 6}}-\frac{1}{1+e^{2.88-0.58\cdot 5}}=0.14$$

STATA-Beispiel: Arbeitslosigkeit

- Im ALLBUS 2002 wurden Erwerbstätige gefragt, ob sie die letzten 10 Jahre arbeitslos waren (1=Arbeitslosigkeit)
  - Hinzu: die gegenwärtig Arbeitslosen
  - Sinnvoll: Einschränkung auf 30-65 Jährige

```
. logit arblos bild

Iteration 0:   log likelihood =  -850.0254 [ln(L0)]
Iteration 1:   log likelihood = -835.29614
Iteration 2:   log likelihood = -835.20727
Iteration 3:   log likelihood = -835.20726

Logistic regression                     Number of obs =    1329
                                        LR chi2(1)    =   29.64
                                        Prob > chi2   =  0.0000
Log likelihood = -835.20726 [ln(L1)]    Pseudo R2     =  0.0174

------------------------------------------------------------------------------
      arblos |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |  -.1083058   .0204825    -5.29   0.000    -.1484507   -.0681609
       _cons |   .7369764   .2693713     2.74   0.006     .2090184    1.264934
------------------------------------------------------------------------------
```

LR – Teststatistik: $2(-835+850)$

R^2_{MF}: $\frac{-850+835}{-850}$

z-Wert

STATA-Beispiel: Arbeitslosigkeit

- Logit-Interpretation
  - Mit jedem Bildungsjahr sinkt das Logit um 0,108
- Odds-Interpretation (Odds-Ratio, OR)

```
. logit arblos bild, or

------------------------------------------------------------------------------
     arblos | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
       bild |   .8973532     .01838    -5.29   0.000     .8620425    .9341102
------------------------------------------------------------------------------
```

  - Mit jedem Bildungsjahr sinkt das Odds („Arbeitslosigkeitschance“?) um 10,3%
- Wahrscheinlichkeits-Interpretation

$$P(Y=1|14)-P(Y=1|13)=\frac{1}{1+e^{-0.737+0.108\cdot 14}}-\frac{1}{1+e^{-0.737+0.108\cdot 13}}=0.315-0.339=-0.024$$

  - Ausgehend vom Bildungsmittel (~13) sinkt mit einem weiteren Bildungsjahr die Arbeitslosigkeitswhs. um 2,4 Prozentpunkte

STATA-Beispiel: Arbeitslosigkeit

- Das LWM kommt praktisch zum gleichen Ergebnis

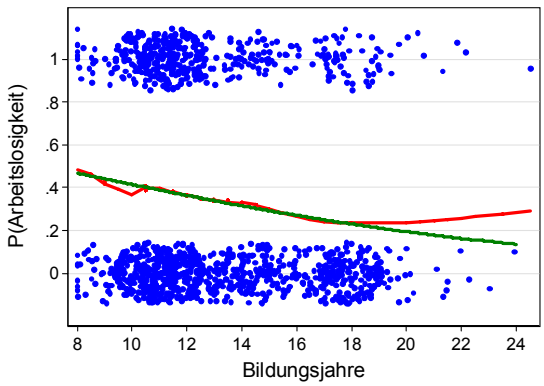
```
. regress arblos bild

      Source |       SS       df       MS              Number of obs =    1329
-------------+------------------------------           F(  1,  1327) =   29.16
       Model |  6.39192099     1  6.39192099           Prob > F      =  0.0000
    Residual |  290.914324  1327  .219227072           R-squared     =  0.0215
-------------+------------------------------           Adj R-squared =  0.0208
       Total |  297.306245  1328  .223875185           Root MSE      =  .46822

------------------------------------------------------------------------------
      arblos |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bild |  -.0228028    .004223    -5.40   0.000    -.0310872   -.0145183
       _cons |   .6386421   .0571673    11.17   0.000      .526494    .7507901
------------------------------------------------------------------------------
```

Multiple logistische Regression

```
. logit arblos bild alter frau ost

Iteration 0:   log likelihood =  -850.0254
Iteration 1:   log likelihood = -776.09162
Iteration 2:   log likelihood = -774.80455
Iteration 3:   log likelihood = -774.80199

Logistic regression                                     Number of obs   =       1329
                                                        LR chi2(4)      =     150.45
                                                        Prob > chi2     =     0.0000
Log likelihood = -774.80199                             Pseudo R2       =     0.0885

------------------------------------------------------------------------------
      arblos |      Coef.   Std. Err.      z    P>|z|     Odds         Whs.
-------------+----------------------------------------------------------------
       bild |  -.1305347   .0219059    -5.96   0.000     0.88       -0.027
      alter |  -.0380065   .0073747    -5.15   0.000     0.96       -0.008
       frau |  -.0385114   .1246573    -0.31   0.757     0.96       -0.008
        ost |   1.220968   .1260793     9.68   0.000     3.39        0.274
      _cons |    2.22248   .4519836     4.92   0.000
------------------------------------------------------------------------------
```

- Analog zur linearen Regression: die Effekte der anderen uVs sind „herauspartialisiert“
- Bei der Berechnung der Whs.Effekte sind alle X-Variablen auf ihren Mittelwert gesetzt
  - Metrische uV: Effekt, wenn X +1; Dummy: Effekt, wenn von 0 zu 1
  - (berechnet mit Ado von Scott Long: prchange)

Diagnostik: funktionale Form



Streudiagramm nach Bildung
- Mit Jitter
- Rot: Lowess
- Grün: logistisches Modell

Problem: nur bivariat

Das logistische Modell repräsentiert den Zusammenhang in den Daten ganz gut.
Der Zusammenhang ist annähernd linear, weshalb das LWM hier auch gut passen würde.
Bei hoher Bildung erkennt man eine Abweichung vom Lowess. Die ist allerdings von einem Ausreißer verursacht.