universität innsbruck

Einführung OLS

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Grundlagen

- Deskriptive Regressionsanalyse – warum?
  - Informationsverdichtung
  - einfacher Einstieg
  - "Mechanik" der Regressionsanalyse
- Es geht um **Zusammenhänge**
  - Allgemein: Funktionen

$$y = f(x_1, x_2, \ldots, x_k)$$

  - Einfachster Fall: lineare Funktion

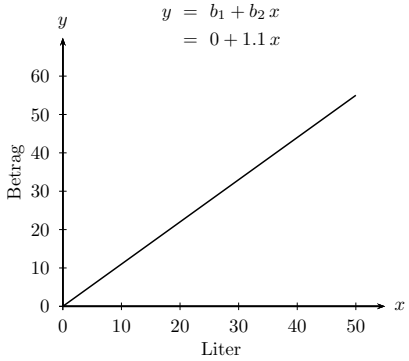
$$y = b_1 + b_2 x$$

  - Ableitung (*marginaler Effekt*):

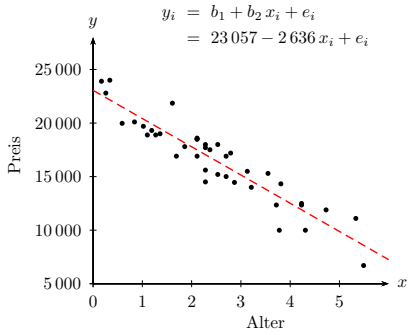
$$\frac{dy}{dx} = b_2$$

Grundlagen

z.B. Tanken . . .



z.B. Gebrauchtautos . . .



Grundlagen

Preise (in Euro) und Alter (in Jahren) von Gebrauchtautos

Obs.	Preis	Alter	AlterJ	km
1	10000	3.78	4	188000
2	21850	1.61	2	25900
3	14500	2.28	2	83300
4	11100	5.33	5	120300
5	6700	5.49	5	142000
⋮	⋮	⋮	⋮	⋮
38	22800	0.26	0	5000
39	12350	4.23	4	73000
40	17800	1.86	2	35000

$$\begin{aligned} y_1 &= b_1 + b_2 x_1 + e_1 \\ y_2 &= b_1 + b_2 x_2 + e_2 \\ &\vdots \qquad\qquad \vdots \\ y_n &= b_1 + b_2 x_n + e_n \end{aligned}$$

$y_i = b_1 + b_2 x_i + e_i$

mit $i = 1, 2, \ldots, n$

Grundlagen

$$y_i = b_1 + b_2 x_i + e_i, \qquad \text{mit } i = 1, 2, \ldots, n$$

Zwei Fragen:

1. Was ist *beobachtbar*?
2. Was ist *beobachtungsspezifisch*?

Grundlagen

Interzept und Regressionskonstante:

Vektorschreibweise:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix}$$

Grundlagen

Alternative Bezeichnungen für y und x:

y	x
– links-stehende Variable (*'left-hand side variable'*)	rechts-stehende Variable (*'right-hand side variable'*)
– **abhängige Variable** (*'dependent variable'*)	[unabhängige Variable] (*'independent variable'*)
– erklärte Variable (*'explained variable'*)	**erklärende Variable** (*'explanatory variable'*)
– Regressand (*'regressand'*)	**Regressor** (*'regressor'*)
– Antwortvariable (*'response variable'*)	Kovariable (*'covariate'*)
– Effektvariable (*'effect variable'*)	Kontrollvariable (*'control variable'*)

Die OLS Methode

Grundlagen der Ökonometrie

www.hsto.info/econometrics

OLS Methode

- **OLS**: *'**O**rdinary **L**east **S**quares'*, bzw. Methode der (Gewöhnlichen) Kleinsten Quadrate
- OLS ist eine **Zerlegungsmethode**

$$y_i = \underbrace{b_1 + b_2 x_i}_{\substack{\text{systematische} \\ \text{Komponente } \widehat{y}_i}} + \underbrace{e_i}_{\substack{\text{Resi-} \\ \text{duen}}}$$

- Gefittete Werte:

$$\widehat{y}_i = b_1 + b_2 x_i$$

Zerlegung von y

- Hypothese in linearer Form:
  - Zerlegung in eine systematische, durch x erklärte Komponente
  - und eine unerklärte Komponente e_i.

$$y_i = b_1 + b_2 x_i + e_i$$
$$y_i = \underline{b_1 + b_2 x_i} + e_i$$
$$y_i = b_1 + b_2 x_i + \underline{e_i}$$

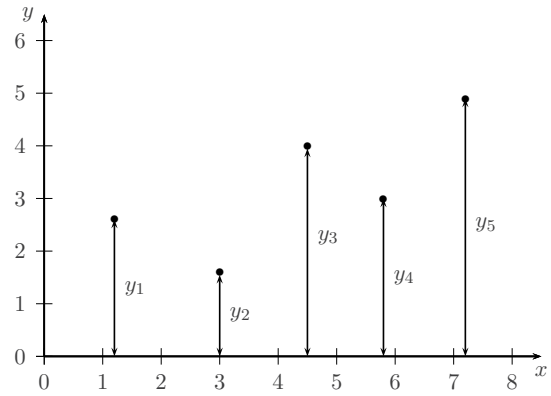
$$y_i = b_1 + b_2 x_i + e_i$$

- Die Ableitung

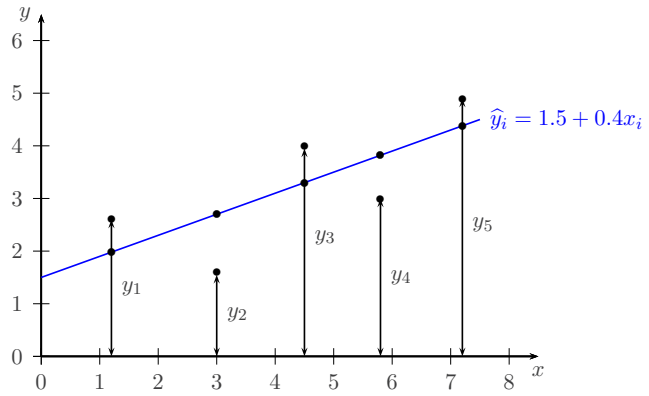
$$b_2 = \frac{dy}{dx}$$

misst die *marginalen* Auswirkungen einer Änderung von x auf y, wenn $x \perp e$!

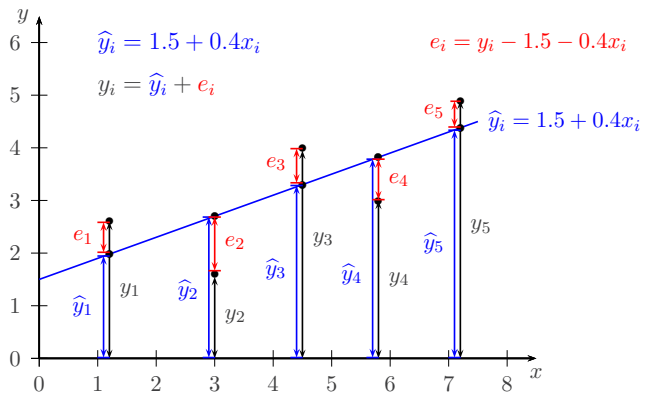
Zerlegung von y



Zerlegung von y



Zerlegung von y

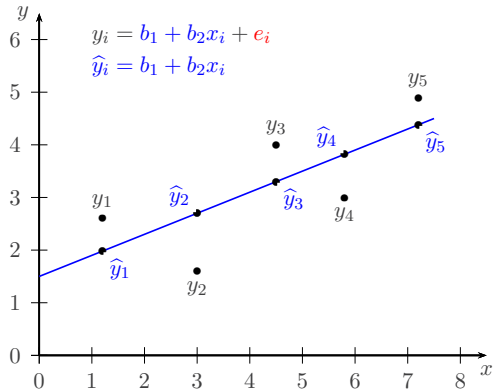


OLS Methode

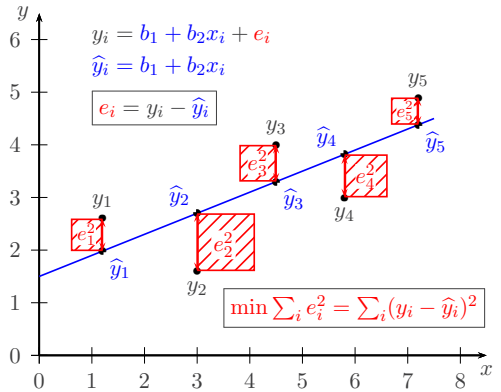
Frage: wie findet man das b_1 und b_1, damit der *systematische* (erklärte) Anteil von y möglichst groß wird?

- $\min \sum_i e_i$ funktioniert nicht $(+, -)$
- $\min \sum_i |e_i|$ ist 'schwierig'
- $\min \sum_i e_i^2$ definiert OLS!

Berechnung von b_1 und b_2



Berechnung von b_1 und b_2



OLS Methode

Einfache Minimierungsaufgabe:

- $y_i = b_1 + b_2 x_i + e_i \quad \rightarrow \quad e_i = y_i - b_1 - b_2 x_i$
- Quadratsumme:
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$
- Optimierung:
$$\min_{b_1,b_2} \sum_{i=1}^{n} e_i^2 = \min_{b_1,b_2} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

  $\Rightarrow$ die ersten Ableitungen nach b_1 und b_2 Null setzen und lösen!

OLS Methode

$$\min_{b_1,b_2} \sum_{i=1}^{n} e_i^2 = \min_{b_1,b_2} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

Lösung: **FOC** (*'first order conditions'*):

$$\frac{\partial \sum_i e_i^2}{\partial b_1} = 2 \sum_i \underbrace{(y_i - b_1 - b_2 x_i)}_{e_i} (-1) = -2 \sum_i e_i = 0$$

$$\frac{\partial \sum_i e_i^2}{\partial b_2} = 2 \sum_i \underbrace{(y_i - b_1 - b_2 x_i)}_{e_i} (-x_i) = -2 \sum_i x_i e_i = 0$$

(Kettenregel!)

OLS Methode

FOC

$$\begin{aligned} \sum_i e_i &= 0 \\ \sum_i x_i e_i &= 0 \end{aligned} \qquad \Rightarrow b_1, b_2$$

mit $\quad e_i := y_i - b_1 - b_2 x_i$

- Die mit OLS Koeffizienten berechneten Residuen erfüllen diese FOC immer!
- Fast alle wichtigen Eigenschaften des OLS Modells folgen aus diesen Bedingungen erster Ordnung!

OLS Methode

Die Lösung der beiden **FOC** liefert die **OLS Koeffizienten**

OLS Koeffizienten

$$b_2 = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$$

$$b_1 = \bar{y} - b_2 \bar{x}$$

Die detaillierte Herleitung finden Sie im Skript.
Bitte genau anschauen!

OLS Methode

Übungsbeispiele:

- Leiten Sie die OLS-Schätzfunktion für eine Regression *nur* auf die Regressionskonstante her!

$$y_i = b_1 + e_i \qquad \min \sum_i e_i^2$$

- Leiten Sie die OLS-Schätzfunktion für eine Regression durch den Ursprung (d.h. *ohne* Regressionskonstante) her!

$$y_i = b_2 x_i + e_i \qquad \min \sum_i e_i^2$$

Thanx . . .

OLS: Bedingte Mittelwerte

Grundlagen der Ökonometrie

www.hsto.info/econometrics

OLS & bedingte Mittelwerte

- Wiederholung:
  OLS ist eine Zerlegungsmethode

$$y_i = \widehat{y}_i + e_i$$

- Aus den FOC folgt für die OLS Residuen:

$$\text{cov}(x, e) = 0 \qquad \text{(warum?)}$$

- Für die Interpretation interessiert uns v.a. der *systematische* Teil $\widehat{y}$

$$\widehat{y}_i = b_1 + b_2 x_i$$

  mit dem *marginalen Effekt*

$$\frac{d\widehat{y}}{dx} = b_2 \quad \leftrightarrow \quad \frac{dy}{dx} = b_2 \quad \text{nur wenn } \text{cov}(x, e) = 0 \quad (x \perp e)$$

OLS & bedingte Mittelwerte

- Erinnern wir uns: Regression auf Regressionskonstante liefert Mittelwert

$$y_i = b_1 + e_i \quad \Rightarrow \quad b_1 = \overline{y}$$

- Dies kann verallgemeinert werden: *bedingte Mittelwerte*
- **bedingte Mittelwerte:** arithm. Mittel eine Untergruppe.
  (d.h. einer Teilmenge, die eine bestimmte *Bedingung* erfüllt)
- Ein Beispiel zum Einstieg . . .

  Zusätzliche Intuition/Beispiele im Kapitel zu ‘Dummyvariablen’)

Bedingte Mittelwerte: Autobeispiel

Obs.	Preis	Alter	AlterJ	km
1	10000	3.78	4	188000
2	21850	1.61	2	25900
3	14500	2.28	2	83300
4	11100	5.33	5	120300
5	6700	5.49	5	142000
6	24000	0.34	0	5500
7	10000	4.31	4	100500
8	16900	1.69	2	31000
9	18000	2.53	3	23000
10	15300	3.55	4	73000
11	19980	0.59	1	1500
12	15600	2.28	2	21700
13	17200	2.79	3	27570
14	18890	1.10	1	13181
15	23900	0.17	0	1800
16	14320	3.81	4	67210
17	11900	4.73	5	73900
18	15200	2.53	3	27000
19	14450	2.87	3	90000
20	18600	2.11	2	27000

Obs.	Preis	Alter	AlterJ	km
21	15000	2.70	3	51500
22	18500	2.11	2	25880
23	18500	2.11	2	19230
24	12350	3.72	4	75000
25	16900	2.70	3	22000
26	18000	2.28	2	35000
27	18890	1.27	1	22500
28	20100	0.84	1	18000
29	19700	1.02	1	12600
30	17500	2.37	2	35900
31	19300	1.19	1	5000
32	15500	3.13	3	39000
33	14000	3.21	3	56400
34	16900	2.11	2	55000
35	17700	2.28	2	25100
36	12500	4.23	4	59200
37	19000	1.36	1	19000
38	22800	0.26	0	5000
39	12350	4.23	4	73000
40	17800	1.86	2	35000

Bedingte Mittelwerte: Autobeispiel

	AlterJ = 0	AlterJ = 1	AlterJ = 2	AlterJ = 3	AlterJ = 4	AlterJ = 5
	24000	19980	21850	18000	10000	11100
	23900	18890	14500	17200	10000	6700
P	22800	18890	16900	15200	15300	11900
r		20100	15600	14450	14320	
e		19700	18600	15000	12350	
i		19300	18500	16900	12500	
s		19000	18500	15500	12350	
e			18000	14000		
			17500			
			16900			
			17700			
			17800			
n	3	7	12	8	7	3
$\overline{y}$	23567	19409	17696	15781	12403	9900
$\Delta\overline{y}$		-4158	-1713	-1915	-3378	-2503
$\widehat{y}$	22709	20192	17675	15158	12641	10124
$\Delta\widehat{y}$		-2517	-2517	-2517	-2517	-2517

Bedingte Mittelwerte: Autobeispiel

Für AlterJ (wiederholte Beobachtungen):

Bedingte Mittelwerte:

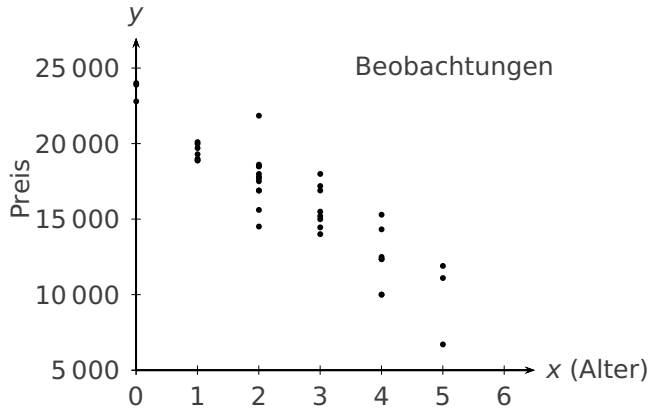
$$(\overline{\text{Preis}}|\text{AlterJ}) = \begin{cases} 23567 & \text{für AlterJ = 0} \\ 19409 & \text{für AlterJ = 1} \\ 17696 & \text{für AlterJ = 2} \\ 15781 & \text{für AlterJ = 3} \\ 12403 & \text{für AlterJ = 4} \\ 9900 & \text{für AlterJ = 5} \end{cases}$$

Gefittete Werte:

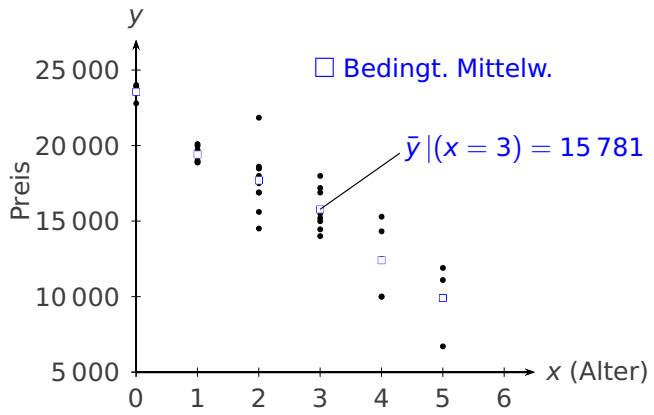
$$(\widehat{\text{Preis}}|\text{AlterJ}) = \begin{cases} 22709 & \text{für AlterJ = 0} \\ 20192 & \text{für AlterJ = 1} \\ 17675 & \text{für AlterJ = 2} \\ 15158 & \text{für AlterJ = 3} \\ 12641 & \text{für AlterJ = 4} \\ 10124 & \text{für AlterJ = 5} \end{cases}$$

$$\widehat{\text{Preis}}_i = 22\,709 - 2\,517\,\text{AlterJ}_i$$

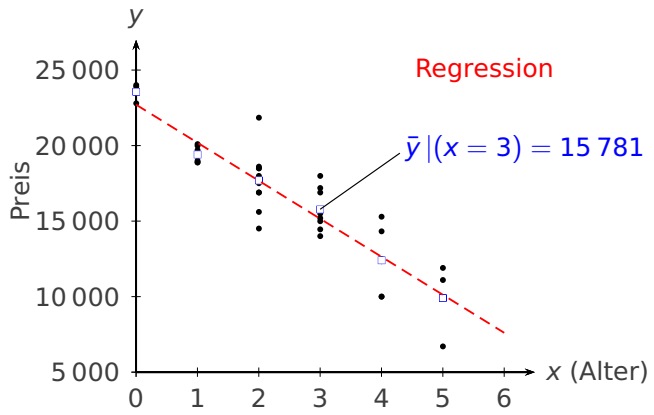
Bedingte Mittelwerte



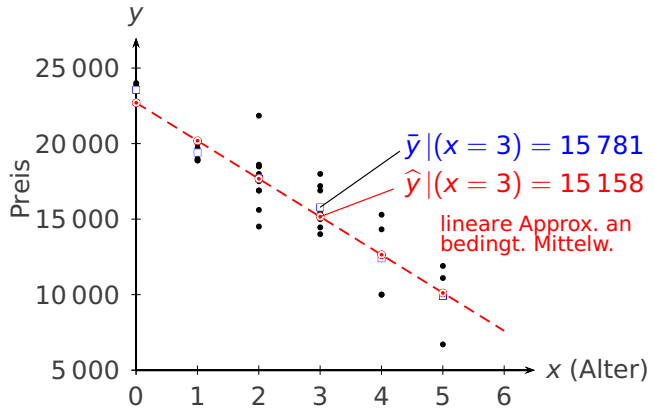
Bedingte Mittelwerte



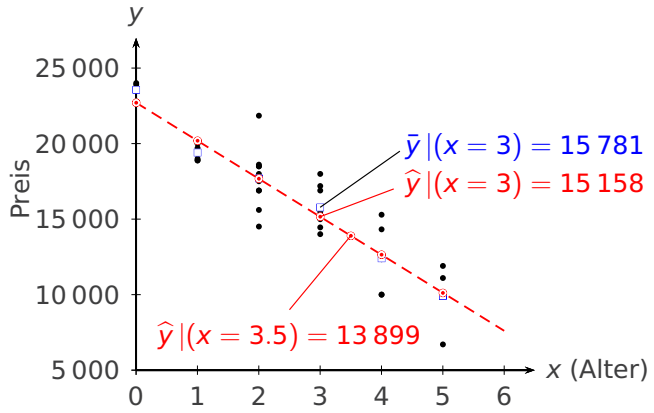
Bedingte Mittelwerte



Bedingte Mittelwerte



Bedingte Mittelwerte



Bedingte Mittelwerte: Autobeispiel

Schlussfolgerung: (etwas sloppy)

- Wir können die gefitteten Werte $\widehat{y}$ als *lineare Approximation* an die bedingten Mittelwerte interpretieren:

$$\widehat{y}|(x = \underline{x}) \overset{\text{lin}}{\approx} \bar{y}|(x = \underline{x})$$

  ($\underline{x}$ ist eine konkrete Ausprägung von x)
- bzw. etwas allgemeiner, die OLS Schätzfunktion als eine *lineare Approximation* and die bedingte Mittelwertfunktion.
- Marginaler Effekt:

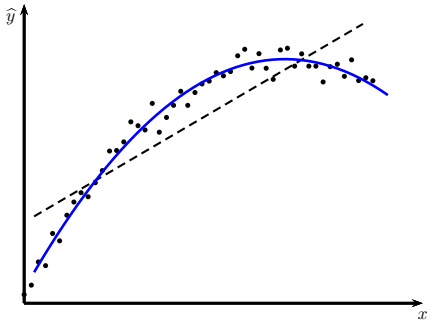
$$\frac{d\widehat{y}}{dx} = b_2 \quad \rightarrow \quad \frac{d\,\widehat{\text{Preis}}}{d\,\text{AlterJ}} = 2\,517$$

  Marg. Effekt beschreibt nicht Individuum, sondern (lineare Approx. an) durchschnittliche Änderung!

OLS & bedingte Mittelwerte

Nichtlineare Zusammenhänge: einfach(st)er Fall:

$$\widehat{y} = b_1 + b_2 x + b_3 x^2 \quad \rightarrow \quad \text{marg. Eff.:} \quad \frac{d\,\widehat{y}}{dx} = b_2 + 2b_3 x$$



Darstellung von Regressionsgleichungen

Konventionell:

In Zeilenform: (z.B. Lehrbücher)

$$\begin{aligned} \text{Preis} \;=\; & 22\,709.303 \;-\; 2\,517.267\ \text{AlterJ} \\ & (532.689)^{***} \qquad (190.125)^{***} \\ & R^2 = 0.822, \qquad n = 40 \end{aligned}$$

Darstellung von Regressionsgleichungen

In Tabellenform: (v.a. Publikationen)

	Dependent variable:	
	Preis	
	(1)	(2)
Constant	$22,709.300^{***}$	$23,056.710^{***}$
	(532.689)	(468.871)
AlterJ	$-2,517.267^{***}$	
	(190.125)	
Alter		$-2,635.669^{***}$
		(166.935)
Observations	40	40
R^2	0.822	0.868
Note:	*p<0.1; **p<0.05; ***p<0.01	

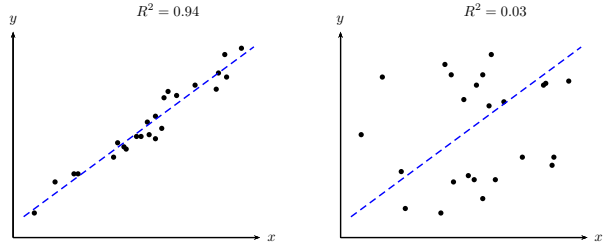
Das Bestimmtheitsmaß R^2

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Bestimmtheitsmaß R^2

Der Zusammenhang zwischen zwei Variablen kann durch eine Regressionsgerade mehr oder weniger gut beschrieben werden:



⇒ Streuungszerlegung ...

Bestimmtheitsmaß R^2

- Jede OLS Regressionsgerade *mit Interzept* verläuft durch die Mittelwerte $\overline{y}$ und $\overline{x}$!

  Warum?

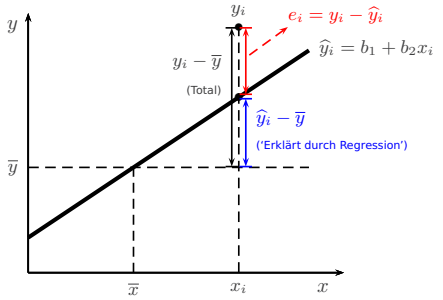
- $y_i = b_1 + b_2 x_i + e_i$
  Wie setzen für x_i den Mittelwert $\overline{x}$ ein und für das Interzept die OLS Schätzfunktion $b_1 = \overline{y} - b_2 \overline{x}$

$$\begin{aligned} \widehat{y}_{\overline{x}} &= b_1 + b_2 \overline{x} \\ &= \underbrace{\overline{y} - b_2 \overline{x}}_{b_1} + b_2 \overline{x} \\ &= \overline{y} \end{aligned}$$

Also ...

Bestimmtheitsmaß R^2

Für eine Beobachtung (Punkt):



$$(y_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i)$$

Varianzzerlegung

Für eine Beobachtung:

$$\begin{aligned}(y_i - \overline{y}) &= (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i) \\ (y_i - \overline{y})^2 &= [(\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i)]^2 \\ &= (\widehat{y}_i - \overline{y})^2 + (y_i - \widehat{y}_i)^2 + \\ &\quad +2(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i)\end{aligned}$$

Streuung über alle Beobachtungen:

$$\begin{aligned}\sum \left[(y_i - \overline{y})^2\right] &= \sum \left[(\widehat{y}_i - \overline{y})^2\right] + \sum \left[(y_i - \widehat{y}_i)^2\right] + \\ &\quad +2 \underbrace{\sum (\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i)}_{\text{cov}(\widehat{y}, e) = 0}\end{aligned}$$

(wenn $b_1 \neq 0$, s. Skript)

R^2 & Varianzzerlegung

$$\underbrace{\sum (y_i - \overline{y})^2}_{\substack{\text{Total Sum Squared} \\ \text{TSS}}} = \underbrace{\sum (\widehat{y}_i - \overline{y})^2}_{\substack{\text{Explained Sum Squared} \\ \text{ESS}}} + \underbrace{\sum (y_i - \widehat{y}_i)^2}_{\substack{\text{Residual Sum Squared} \\ \text{RSS}}}$$

bzw. als Varianzen

$$\begin{aligned}\frac{1}{n}\sum (y_i - \overline{y})^2 &= \frac{1}{n}\sum (\widehat{y}_i - \overline{y})^2 + \frac{1}{n}\sum (y_i - \widehat{y}_i)^2 \\ \text{var}(y) &= \text{var}(\widehat{y}) + \text{var}(e)\end{aligned}$$

R^2 & Varianzzerlegung

$$\begin{aligned}\text{TSS} &= \text{ESS} + \text{RSS} \\ \text{ESS} &= \text{TSS} - \text{RSS}\end{aligned}$$

Das **Bestimmtheitsmaß** R^2 ist definiert als
Anteil der erklärten Streuung ESS **an der gesamten Streuung** TSS:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{var}(e)}{\text{var}(y)}$$

bzw.

$$\boxed{R^2 = 1 - \frac{\sum (y_i - \widehat{y}_i)^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2}}$$

Beispiel

i	x	y	$\widehat{y}$	e	e^2	$(y - \overline{y})^2$
1	1.2	2.6	2.0	0.6	0.39	0.38
2	3.0	1.6	2.7	-1.1	1.18	2.62
3	4.5	4.0	3.3	0.7	0.51	0.61
4	5.8	3.0	3.8	-0.8	0.64	0.05
5	7.2	4.9	4.4	0.5	0.30	2.82
$\sum$	21.7	16.1		0.0	3.03	6.49

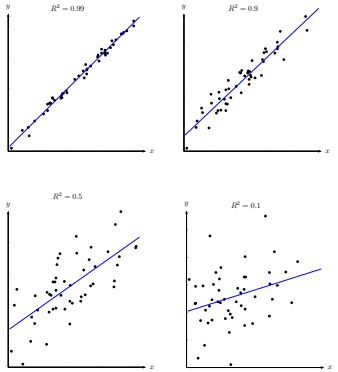
$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{3.03}{6.49} = 0.53$$

Bestimmtheitsmaß R^2

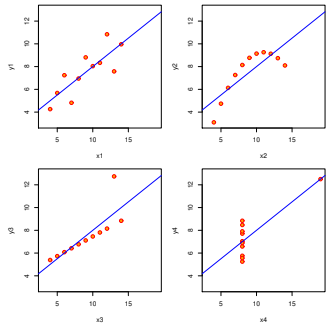
- Das Bestimmtheitsmaß R^2 gibt an, welcher Anteil der Varianz von y durch die Regression (bzw. durch die x) erklärt wird.
- Das R^2 liegt für Regressionsgleichungen mit Interzept immer zwischen Null und Eins.
  (dies muss nicht für Regressionsgleichungen ohne Interzept gelten! Warum?)
- R^2 kann auch als das Quadrat des (Pearsonschen) Korrelationskoeffizienten zwischen den beobachteten Werte y und der gefitteten Werte $\widehat{y}$ interpretiert werden

$$R^2 = \left[\frac{\operatorname{cov}(y, \widehat{y})}{\sqrt{\operatorname{var}(y)\operatorname{var}(\widehat{y})}}\right]^2$$

Unterschiedliche R^2:



Gleiche R^2: (Anscombe Daten)

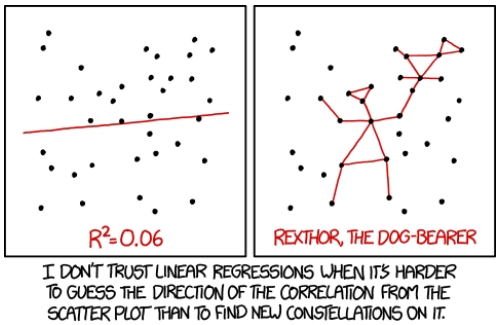


alle vier Regressionen weisen fast das gleiche $R^2 = 0.666$ auf!

Achtung!

- Das Bestimmtheitsmaß R^2 ist nur ein deskriptives Maß für die Güte der Anpassung der Regressionsgerade (*Goodness of Fit*), kein Maß für die *'Qualität'* der Schätzung.
- Ein hohes R^2 kann auch auftreten, wenn die die Schätzung grobe Probleme aufweist (z.B. Scheinkorrelation, Autokorrelation, Fehlspezifikation, . . .)
- Nur interpretierbar, wenn die Regression ein Interzept (b_1) enthält!

Bestimmtheitsmaß R^2



Quelle: xkcd, `http://xkcd.com/1725/`

Thanks . . .

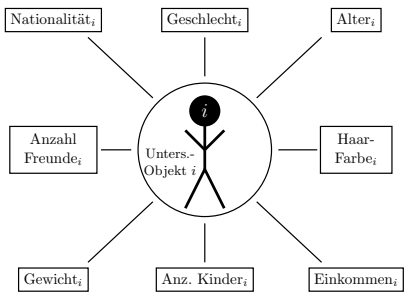
Multiple Regression

Grundlagen der Ökonometrie

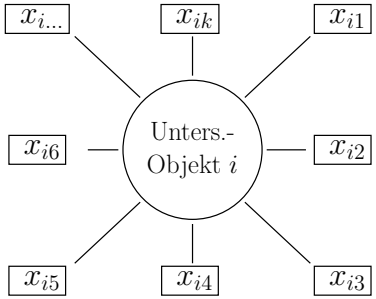
www.hsto.info/econometrics

Multiple Regression: Untersuchungsobjekte & Merkmale

Beispiel:



Allgemeiner:



Multiple Regression: Untersuchungsobjekte & Merkmale

Beispiel: Umfrage
$n = 5$ Beobachtungen, $k = 4$ Variablen

		Merkmale (Variablen)			
	i	Alter	Geschlecht	Eink.	Nat.
U.-Objekte	1	21	weibl	2400	ITA
	2	18	weibl	1800	AUT
	3	20	männl	2600	AUT
	4	18	weibl	1500	DEU
	5	19	männl	1800	AUT

Allgemein:
n Beobachtungen, k Variablen

	Merkmale (Variablen)				
i	$X_{\cdot 1}$	$X_{\cdot 2}$	$X_{\cdot 3}$	$\cdots$	$X_{\cdot k}$
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1k}
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2k}
3	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{n1}	x_{n2}	x_{n3}	$\cdots$	x_{nk}

Multiple Regression

- Autopreise hängen nicht nur vom Alter ab, sondern auch von km-Zahl, Zustand, Unfälle, Ausstattung, . . . (bisher in Residuen!)
- Eine kleine Erweiterung $\rightarrow$ zusätzliche Regressoren

$$y_i = b_1 + b_2 x_{i2} + \cdots + b_h x_{ih} + \cdots b_k x_{ik} + e_i$$

- Regressoren benötigen *zwei* Indizes x_{ih}, mit
  - $i = 1, \ldots n$ für die Zeile (Beobachtungen), und
  - $h = 1, \ldots k$ für die Spalte (Variablen).
  - Bedingung: $k \leq n$
- Neu ist nur die *ceteris paribus* Interpretation!

Multiple Regression

Für zwei Regressoren: ($k = 3$; inkl. Regressionskonstante!)

$$\min_{b_1,b_2,b_3} \sum e_i^2 = \min_{b_1,b_2,b_3} \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})^2$$

FOC:

$$\begin{aligned}
\frac{\partial \sum e_i^2}{\partial b_1} &= 2 \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-1) \stackrel{!}{=} 0 \\
\frac{\partial \sum e_i^2}{\partial b_2} &= 2 \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-x_{i2}) \stackrel{!}{=} 0 \\
\frac{\partial \sum e_i^2}{\partial b_3} &= 2 \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-x_{i3}) \stackrel{!}{=} 0
\end{aligned}$$

$\Rightarrow \sum e_i = 0$, $\sum e_i x_{i2} = 0$ und $\sum e_i x_{i3} = 0$

Multiple Regression

Lösungen sehen kompliziert aus . . .

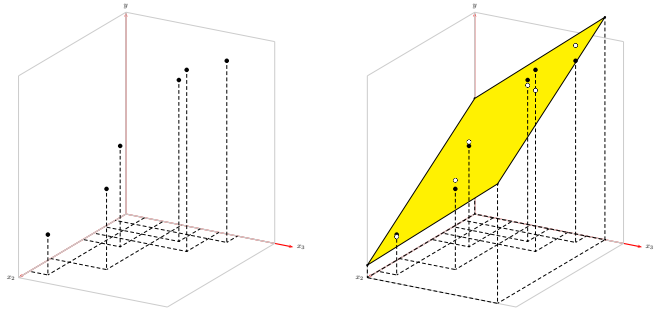
$$\begin{aligned}
b_2 &= \frac{(\sum \ddot{y}_i \ddot{x}_{i2})(\sum \ddot{x}_{i3}^2) - (\sum \ddot{y}_i \ddot{x}_{i3})(\sum \ddot{x}_{i2} \ddot{x}_{i3})}{(\sum \ddot{x}_{i2}^2)\sum \ddot{x}_{i3}^2) - (\sum \ddot{x}_{i2} \ddot{x}_{i3})^2} \\
b_3 &= \frac{(\sum \ddot{y}_i \ddot{x}_{i3})(\sum \ddot{x}_{i2}^2) - (\sum \ddot{y}_i \ddot{x}_{i2})(\sum \ddot{x}_{i2} \ddot{x}_{i3})}{(\sum \ddot{x}_{i2}^2 \sum \ddot{x}_{i3}^2) - (\sum \ddot{x}_{i2} \ddot{x}_{i3})^2} \\
b_1 &= \bar{y} - b_2 \bar{x}_2 - b_3 \bar{x}_3
\end{aligned}$$

in Matrixschreibweise geht's einfacher

$$\boxed{\mathbf{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{y}}$$

Multiple Regression

y	x_2	x_3
2	9	1
5	4	2
4	7	3
8	2	4
9	3	5
9	1	6

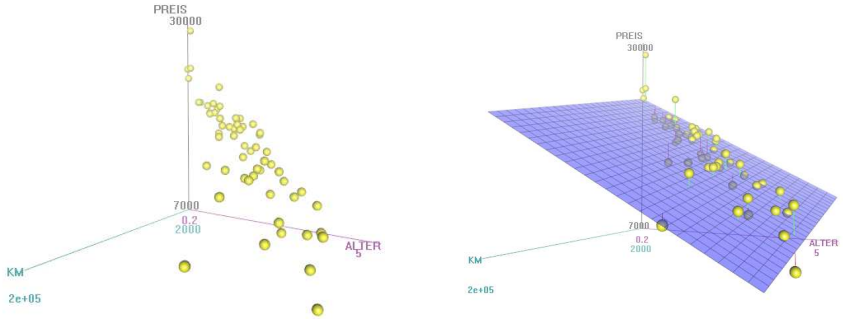


$$y_i = 5.73 - 0.51 x_{i2} + 0.76 x_{i3} + e_i$$

. . . höherdimensionale Fälle können grafisch nicht mehr dargestellt werden.

Multiple Regression

Autobeispiel: $\widehat{\text{Preis}} = 22649.9 - 1896.3\,\text{Alter} - 0.03\,\text{km}$



Fox, J., and Bouchet-Valat, M. (2020). **Rcmdr**: R Commander. R package version 2.7-1

Multiple Regression

***Ceteris paribus* Interpretation:**

$$\widehat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3}$$

Marginale Effekte:

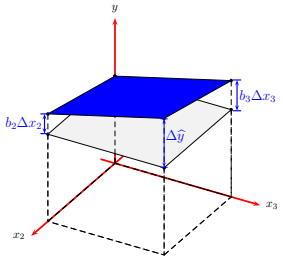
$$b_2 = \left.\frac{d\hat{y}}{dx_2}\right|_{dx_3=0} = \frac{\partial \widehat{y}}{\partial x_2} \quad \text{und} \quad b_3 = \left.\frac{d\hat{y}}{dx_3}\right|_{dx_2=0} = \frac{\partial \widehat{y}}{\partial x_3}$$

b_2: um wie viele Einheiten ändert sich $\widehat{y}$, wenn x_2 um eine Einheit zunimmt *und* x_3 *konstant bleibt*, d.h. $dx_3 = 0$, oder in anderen Worten: *ceteris paribus!*

Analog für b_3 und weitere Regressoren

Multiple Regression

***Ceteris paribus* Interpretation:**



$\widehat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3} \quad \rightarrow \quad$ Differenzen $\quad \rightarrow \quad \Delta\widehat{y} = b_2 \Delta x_2 + b_3 \Delta x_3$

$$b_2 = \left.\frac{\Delta \widehat{y}}{\Delta x_2}\right|_{\Delta x_3=0} = \frac{\partial \widehat{y}}{\partial x_2}, \quad \text{und} \quad b_3 = \left.\frac{\Delta \widehat{y}}{\Delta x_3}\right|_{\Delta x_2=0} = \frac{\partial \widehat{y}}{\partial x_3}$$

Multiple Regression

Autobeispiel:

$\widehat{\text{Preis}}$	=	22649.884	−	1896.264 Alter	−	0.031 km
		(411.87)***		(235.215)***		(0.008)***

$$R^2 = 0.907, \qquad n = 40$$

$\widehat{y}$: lineare Approximation an bedingte Mittelwerte

Welchen Preis würden wir für ein 4 Jahre altes Auto mit hunderttausend km erwarten?

$$(\widehat{y}|x_2 = 4, x_3 = 100000) = 22649.884 - 1896.264 * 4 - 0.031 * 100000 = 11\,963.79$$

Marginale Effekte: *ceteris paribus* (!)

$$\frac{\partial\,\widehat{\text{Preis}}}{\partial\,\text{Alter}} = 1896.264 \qquad \frac{\partial\,\widehat{\text{Preis}}}{\partial\,\text{km}} = 0.031$$

Multiple Regression

Achtung:

1. Ceteris-paribus Interpretation bezieht sich *ausschließlich* auf die *im Modell explizit berücksichtigten* x Variablen!!! (d.h. kein *'echtes'* ceteris paribus!)
   - $\Rightarrow$ keine Kausalaussagen!
   - $\Rightarrow$ zentraler Unterschied zu *Randomisierten kontrollierten Experimenten* (RCT)

2. Marginale Effekte beziehen sich auf gefittetes $\widehat{y}_i$ (d.h. auf die lineare Approximation an den *bedingten* Mittelwert), *nicht* auf das y_i eines Individuums!

Multiple Regression

Bestimmtheitsmaß:

- Zusätzliche Regressoren können $\sum_i e_i^2$ nicht vergrößern, d.h.

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2}$$

  kann mit Anzahl der Regressoren nicht abnehmen, nur zunehmen!
- Eine mögliche Korrektur: **Korrigiertes Bestimmtheitsmaß** $\overline{\mathbf{R}}^{\mathbf{2}}$:

$$\overline{R}^2 = 1 - \frac{\frac{\sum_i e_i^2}{(n-k)}}{\frac{\sum_i (y_i - \overline{y})^2}{(n-1)}} = 1 - \left(\frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2}\right)\left(\frac{n-1}{n-k}\right)$$

  Kann mit zunehmende Anzahl von Regressoren abnehmen!
  Keine Interpretation als Anteil erklärter Streuung.
- Alternativen: *Akaike Informationskriterium* (AIC), *Bayessches Informationskriterium* (BIC) (theoret. besser fundiert)

Interpretation von Regressionsgleichungen

- Bisher nur deskriptive Regression: kompakte Beschreibung vorliegender Daten.
- *Stochastische Regression:* Wir interessieren uns für eine *unbekannte* Grundgesamtheit, und beobachten nur eine Stichprobe daraus. Vorstellung wiederholter Stichprobenziehungen (*'repeated sampling'*) aus unbekannter Grundgesamtheit . . . Details folgen
- Wichtige Kennzahlen: **Standardfehler der Koeffizienten**
  (meist in Klammern *unter* den Koeffizienten angegeben)
  - Indikatoren für die Genauigkeit der Schätzungen bei (hypothetisch) wiederholten Stichprobenziehungen.
  - erlauben Hypothesentests für einzelne Koeffizienten
  - **Grobe Faustregel:** Koeffizient 'sollte' mindestens *doppelt so groß* ist wie sein Standardfehler (erlaubt H_0 auf dem 5% Niveau zu verwerfen, dass in einer Grundgesamtheit *kein* Zusammenhang zwischen x und y besteht)
  - Signifikanzniveau wird häufig durch Sterne neben Koeffizienten (oder Standardfehler) angedeutet (meist: *. . . 10%, **. . . 5%, ***. . . 1%)

Interpretation von Regressionsgleichungen

Keine festen Regeln, aber typische Vorgangsweise:

1. Welcher Zusammenhang wird durch Regression dargestellt?
   Es müssen nicht unbedingt alle Regressoren diskutiert werden, aber zumindest die interessierenden!
2. Welche Vorzeichen der *interessierenden* Koeffizienten würden wir a priori erwarten?
   * Positives Vorzeichen: positiver (gleichgerichteter) Zusammenhang,
   * Negatives Vorzeichen: negative (gegengerichteter) Zusammenhang.
3. Welche Vorzeichen beobachten wir tatsächlich?
4. Sind die Koeffizienten der *interessierenden* Variablen statistisch signifikant von Null verschieden?
   - *wenn ja:* interpretieren Sie die quantitative Größenordnung des Koeffizienten
   - *wenn nein:* merken Sie das kurz an und interpretieren Sie diesen Koeffizienten nicht weiter.

Interpretation von Regressionsgleichungen

5. Das Interzept wird nur interpretiert, wenn besondere Gründe dafür sprechen.
6. Erwähnen Sie kurz die Güte der Anpassung (R^2) (ohne ihm zu großes Gewicht beizumessen) und event. weitere Kennziffern.
7. Diskutieren Sie mögliche Probleme (z.B. fehlende Variablen oder andere Verletzungen von OLS Annahmen, wie z.B. *Heteroskedastizität*).

Beispiel:

$$\widehat{\text{Preis}} = \underset{(411.87)^{***}}{22649.884} - \underset{(235.215)^{***}}{1896.264\ \text{Alter}} - \underset{(0.008)^{***}}{0.031\ \text{km}}$$

$$R^2 = 0.907, \qquad n = 40$$

Thanx . . .

Fehlende relevante Variablen (*'omitted variables'*)

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Omitted Variables

- Eines der weitreichendsten und wichtigsten Probleme der Regressionanalyse
- Kaffee $\overset{?}{\rightarrow}$ Krebs (WHO 1991, 2016)
  `https://www.hsph.harvard.edu/nutritionsource/food-features/coffee/`
- Oft Ursache für Scheinkorrelationen (*'spurious correlation'*)
- Einer der Gründe, weshalb Regressionanalysen in der Regel keine Kausalaussagen erlauben
- In der Statistik häufig *'Confounding'* genannt.
- Spätestens seit George Udny Yule (1903) bekannt (Biometrika 2: 121-134).

Ein Beispiel . . .

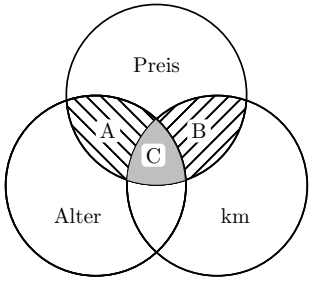
Omitted Variables

Autobeispiel:

Abh.Var.: Preis	(1)	(2)	(3)
Const.	22 649.884	23 056.714	20 279.226
Alter	−1 896.264	−2 635.669	
km	−0.031		−0.082
R^2	0.907	0.868	0.743
n	40	40	40

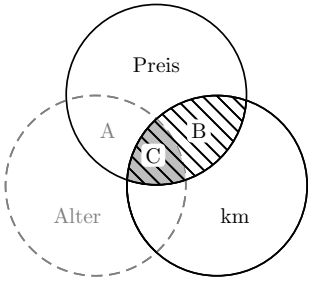
Omitted Variables

Langes (‘besseres’) Modell:



$$\widehat{\text{Preis}} = b_1^l + b_2^l \text{Alter} + b_3^l \text{km}$$

Kurzes (fehlspezifiziertes) Modell:



$$\widehat{\text{Preis}} = b_1^k + b_2^k \text{km}$$

Omitted Variables

Algebra:

- Langes (‘besseres’) Modell:

$$\ddot{y}_i = b_2^l \ddot{x}_{i2} + b_3^l \ddot{x}_{i3} + e_i^l$$

- Kurzes (fehlspezifiziertes) Modell:

$$\ddot{y}_i = b_2^k \ddot{x}_{i2} + e_i^k$$

  d.h. $\ddot{x}_3$ ist *‘omitted variable’*

- Steigungskoeffizient des *kurzen* Modells:

$$b_2^k = \frac{\text{cov}(x_2, y)}{\text{var}(x_2)} = \frac{\sum_i \ddot{x}_{i2}\ddot{y}_i}{\sum_i \ddot{x}_{i2}^2}$$

- Wir setzten das $\ddot{y}$ des langen (‘richtig spezifizierten’) Modells in b_2^k ein . . .

Omitted Variables

Algebra:

$$\begin{aligned} b_2^k &= \frac{\sum_i \ddot{x}_{i2}(\overbrace{b_2^l \ddot{x}_{i2} + b_3^l \ddot{x}_{i3} + e_i^l}^{\ddot{y}})}{\sum_i \ddot{x}_{i2}^2} \\ &= \frac{\sum_i \ddot{x}_{i2} b_2^l \ddot{x}_{i2} + \sum_i \ddot{x}_{i2} b_3^l \ddot{x}_{i3} + \sum_i \ddot{x}_{i2} e_i^l}{\sum_i \ddot{x}_{i2}^2} \\ &= \frac{b_2^l \sum_i \ddot{x}_{i2}^2 + b_3^l \sum_i \ddot{x}_{i2}\ddot{x}_{i3} + \sum_i \ddot{x}_{i2} e_i^l}{\sum_i \ddot{x}_{i2}^2} \\ &= b_2^l + b_3^l \frac{\sum_i \ddot{x}_{i2}\ddot{x}_{i3}}{\sum_i \ddot{x}_{i2}^2} + \frac{\sum_i \ddot{x}_{i2} e_i^l}{\sum_i \ddot{x}_{i2}^2} \end{aligned}$$

FOC: $\sum_i \ddot{x}_{i2} e_i^l = 0$

$$\boxed{b_2^k = b_2^l + b_3^l \frac{\sum_i \ddot{x}_{i2}\ddot{x}_{i3}}{\sum_i \ddot{x}_{i2}^2} = b_2^l + b_3^l \frac{\text{cov}(x_2, x_3)}{\text{var}(x_2)}}$$

Omitted Variables

Omitted Variables

$$b_2^k = b_2^l + b_3^l \frac{\text{cov}(x_2, x_3)}{\text{var}(x_2)}$$

Beachte:

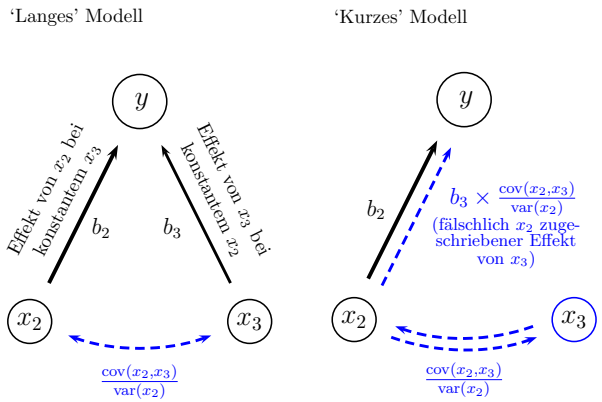
$$x_3 = a_1 + a_2 x_2 + e^*, \quad \Rightarrow \quad a_2 = \frac{\text{cov}(x_2, x_3)}{\text{var}(x_2)}$$

das heißt

$$\boxed{b_2^k = b_2^l + b_3^l \, a_2}$$

b_2^l, b_3^l und a_2 *nicht* beobachtbar!

Omitted Variables



Identifikationsproblem!

Omitted Variables

Achtung:

- 'Omitted Variables Bias' tritt *nur* auf, wenn nicht berücksichtigte Variable (hier x_3) mit
  1. der abhängigen Variable y **und**
  2. mit mindestens einer berücksichtigten erklärenden Variable (x_2) korreliert ist!
- Multiple Regression: Eine einzige fehlende relevante Variable führt in der Regel dazu, dass alle Koeffizienten verzerrt sind!
- *'Omitted Variables Bias'* kann fast nie ausgeschlossen werden!
- Oft kann zumindest das Vorzeichen des Fehlers abgeschätzt werden . . .

Omitted Variables

$$y = b_1^k + b_2^k x_2 + e^k \quad \Leftrightarrow \quad y = b_1^l + b_2^l x_2 + b_3^l x_3 + e^l$$

$$b_2^k = b_2^l + b_3^l \frac{\text{cov}(x_2, x_3)}{\text{var}(x_2)}$$

	$\text{cov}(x_2, x_3) > 0$	$\text{cov}(x_2, x_3) < 0$
$b_3^l > 0$	$b_2^k > b_2^l$	$b_2^k < b_2^l$
$b_3^l < 0$	$b_2^k < b_2^l$	$b_2^k > b_2^l$

- Std.-Löhne: $\widehat{\text{StdL}} = b_1^k + b_2^k\,\text{Bildg} \quad \Leftrightarrow \quad \widehat{\text{StdL}} = b_1^l + b_2^l\,\text{Bildg} + b_3^l\,\text{IQ}$
- Autobeispiel:

  $\widehat{\text{Preis}} = 23\,057 - 2\,636\text{Alter} \quad \Leftrightarrow \quad \widehat{\text{Preis}} = 22\,650 - 1\,896\,\text{Alter} - 0.031\,\text{km}$

Thanx . . .

Frisch-Waugh-Lovell Theorem

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Frisch-Waugh-Lovell Theorem

Computer: es war einmal . . .

Frisch-Waugh-Lovell Theorem

- Ein Klassiker der Ökonometrie:
- Ragnar Frisch und Frederick V. Waugh: Econometrica (1933); Lovell (1963)

- Einsichten in 'OLS-Mechanik' . . .
- Zahlreiche Anwendungen, z.B.
  - mittelwerttransformierte Daten

$$y = b_1 + b_2 x_2 + e \quad \Leftrightarrow \quad \ddot{y} = b_2 \ddot{x} + e$$

  - Panelökonometrie (*fixed effects model*)
  - Grafische Darstellung multipler Regressionen

Frisch-Waugh-Lovell Theorem

Partitionierte Regression:

- Interessierender Zusammenhang:

$$y = b_1 + b_2 x_2 + b_3 x_3 + e$$

- Partitionierung: der lineare Einfluss von x_3

$$\begin{aligned} y &= c_1 + c_2 x_3 + e_y \\ x_2 &= a_1 + a_2 x_3 + e_{x_2} \end{aligned}$$

- *beachte:* $e_y = y - (y|x_3)$ und $e_{x_2} = x_2 - (x_2|x_3)$

  mit $(y|x_3) = \widehat{y} = c_1 + c_2 x_3$ und $(x_2|x_3) = \hat{x}_2 = a_1 + a_2 x_3$

- Eine *bivariate* Regression der Residuen aufeinander liefert den Steigungskoeffizienten b_2 und die Residuen des langen Modells:

$$\boxed{e_y = b_2\, e_{x_2} + e}$$

Frisch-Waugh-Lovell Theorem

- Beweis: siehe Manuskript!
- Beispiel: Mittelwerttransformation $\ddot{y} = b_2 \ddot{x} + e$ (mit $\ddot{y}_i := y_i - \overline{y}$, usw.)
  - Regression auf Regressionskonstante liefert Mittelwert:

$$y = c_1 + e_y \quad \Rightarrow \quad c_1 = \overline{y} \quad \Rightarrow \quad y = \overline{y} + e_y \quad \Rightarrow \quad e_y = y - \overline{y} := \ddot{y}$$

    analog für x: $e_x = x - \overline{x} := \ddot{x}$
  - Also:

$$y = b_1 + b_2 x + e \quad \Leftrightarrow \quad \ddot{y} = b_2 \ddot{x} + e$$

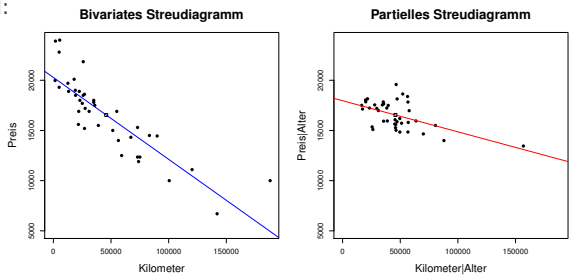
    FWL: beide Regressionen liefern das selbe b_2 und die selben Residuen e!
- Anwendungen: Paneldaten ('*fixed effects models*), Partielle Streudiagramme . . .

Frisch-Waugh-Lovell Theorem

Partielle Streudiagramme:

- Bivariate Darstellungen leiden an *'omitted variables bias'*!
- FWL: Plot der Residuen erlaubt korrekte 2-dimensionale Darstellung

Autobeispiel:



Frisch-Waugh-Lovell Theorem

R Code:

```
Auto <- read.csv2("http://www.hsto.info/econometrics/data/auto40.csv")
Auto$res_Preis <- resid(lm(Preis ~ Alter, data = Auto)) + mean(Auto$Preis)
Auto$res_km    <- resid(lm(km ~ Alter, data = Auto))    + mean(Auto$km)

# Grafik: partieller Scatterplot | Alter
x11()
par(mfrow=c(1,2))
plot(Auto$km, Auto$Preis, main = "Bivariates Streudiagramm",
     xlab = "Kilometer", ylab = "Preis", pch=16,
     xlim = c(0, max(Auto$km)), ylim = c(5000, max(Auto$Preis)))
abline(lm(Preis ~ km, data = Auto), lwd = 2, col = "blue")
#
plot(Auto$res_km, Auto$res_Preis, main = "Partielles Streudiagramm",
     xlab = "Kilometer|Alter", ylab = "Preis|Alter", pch=16,
     xlim = c(0, max(Auto$km)), ylim = c(5000, max(Auto$Preis)))
abline(lm(res_Preis ~ res_km, data = Auto), lwd = 2, col = "red")
```

thanx . . .

Dummy Variablen und kategorische Variablen

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Dummy Variablen

- Bisher: links (y) und rechts (x) *metrisch skalierte* Variablen (z.B. $x_{i2} \in \mathbb{R}$)

$$y_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + e_i \quad \text{(mit } i = 1, \dots n\text{)}$$

- *Jetzt:* links (y) *metrisch skalierte*, aber
  $\rightarrow$ rechts auch *kategoriale* Variable ($d_{ih} \in \mathbb{N}$)

$$y_i = b_1 + b_2 \boldsymbol{d}_{i2} + b_3 x_{i3} + e_i \quad \text{(mit } i = 1, \dots n\text{)}$$

Kategorial: Einteilung in Kategorien, Gruppen . . . $\rightarrow$ Klassifizierung

Dummy Variablen

- **Dummy Variablen**: einfachster Fall kategorialer Variablen, nur zwei Ausprägungen
  1 $\rightarrow$ gehört dazu (TRUE); **0** $\rightarrow$ gehört *nicht* dazu (FALSE).
- alternative Bezeichnungen: *Indikatorvariablen*, *Binäre Variablen*, *Boolesche Variablen*, *0-1 Variablen*, ...
- zum Beispiel:

$$w_i = \begin{cases} 1 & \text{wenn Person } i \text{ eine Frau ist,} \\ 0 & \text{sonst (d.h. nicht Frau)} \end{cases} \qquad m_i = \begin{cases} 1 & \text{wenn Mann, und} \\ 0 & \text{sonst} \end{cases}$$

- *Achtung:*
  Dummy Variablen als *abhängige* Variable (y) sind komplizierter ...
  OLS nur eingeschränkt anwendbar! (*'linear probability model'*)

Dummy Variablen: Beispiel

i	StdL	m	w	v	u	Bildg
1	16	1	0	0	1	17
2	12	0	1	0	1	16
3	16	1	0	1	0	18
4	14	1	0	1	0	13
5	12	1	0	0	1	8
6	12	0	1	1	0	15
7	18	1	0	1	0	19
8	14	0	1	0	1	17
9	14	0	1	1	0	16
10	14	1	0	1	0	9
11	10	0	1	1	0	11
12	13	0	1	0	1	15

Mittelw. StdL: $\overline{\text{StdL}} = (16+12+16+\cdots+13)/12 = 13.75$

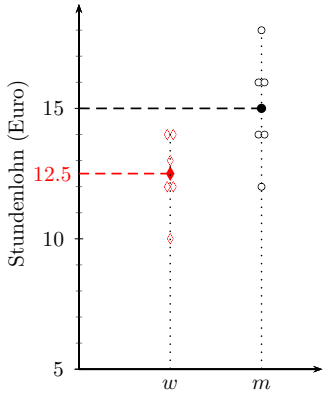
Bedingte Mittelwerte von StdL:

$$\begin{aligned} (\overline{\text{StdL}}|m=1) &= (16+16+14+12+18+14)/6 = 15 \\ (\overline{\text{StdL}}|w=1) &= (12+12+14+14+10+13)/6 = 12.5 \\ (\overline{\text{StdL}}|v=1) &= (16+14+12+\cdots+10)/7 = 14 \\ (\overline{\text{StdL}}|v=0) &= (16+12+12+14+13)/5 = 13.4 \end{aligned}$$

$$\widehat{\text{StdL}}_i = b_1 + b_2 m_i = 12.5 + 2.5\, m_i$$

$$\overline{\text{StdL}}|(m=0) = b_1 = 12.5$$
$$\overline{\text{StdL}}|(m=1) = b_1 + b_2 = 12.5 + 2.5 = 15$$

Dummy Variablen: Beispiel



Referenzkategorie:
Kategorie mit Dummy/ies $= 0$

$$\widehat{\text{StdL}}_i = 12.5 + 2.5\, m_i$$

bzw.

$$\widehat{\text{StdL}}_i = 15 - 2.5\, w_i$$

Dummy Variablen

- Kategoriale Variablen sind *diskret* $\rightarrow$ marginale Effekte nicht definiert!
  $\Rightarrow$ **Partielle Effekte**
- mit d für Dummy ($d_i = 0 \rightarrow$ 'Referenzkategorie')

$$\widehat{y} = b_1 + b_2 d$$

mit

$$\begin{aligned} \widehat{y}|(d=1) &= b_1 + b_2 \\ \widehat{y}|(d=0) &= b_1 \end{aligned}$$

- Differenz $\rightarrow$ *"partieller Effekt"*:

$$[\widehat{y}|(d=1)] - [\widehat{y}|(d=0)] = b_1 + b_2 - b_1 = b_2$$

- ansonsten analog zu marginalen Effekten

Dummy Variablen

- *Dummyvariablenfalle:* wenn alle Ausprägungen einer Dummy *und* Regressionskonstante (1) regressiert werden

$$y_i = b_1 \times 1 + b_2 w_i + b_3 m_i + e_i$$

Lineare Abhängigkeit der Regressoren: $1 = w_i + m_i \quad \forall\, i$
$\Rightarrow$ Es existiert keine Lösung!!!

- Alternativ: kein Interzept, liefert Mittelwerte für beide Ausprägungen

$$\begin{aligned} \widehat{\text{StdL}}_i &= b_2 w_i + b_3 m_i \\ &= 12.5\, w_i + 15\, m_i \end{aligned}$$

. . . eher unüblich

Dummy und metrische Regressoren:

Unterschiede im Interzept:

$$\begin{array}{lcccccc} \text{StdL} & = & 5.78 & + & 2.95\ \text{m} & + & 0.45\ \text{Bildg} \\ & & (1.17)^{***} & & (0.50)^{***} & & (0.07)^{***} \end{array}$$

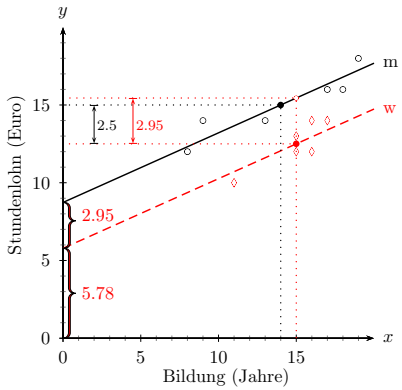
$$R^2 = 0.873, \qquad n = 12$$

Vergleich: StdL = 12.5 + 2.5 m

Kurzes Modell $\rightarrow$ *'omitted variable bias'*!

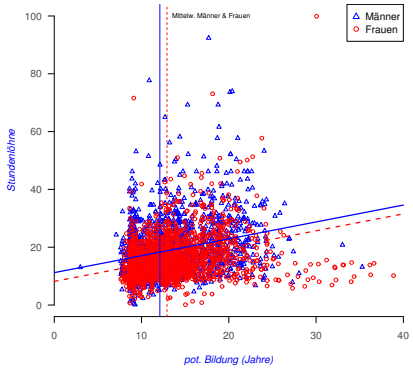
Dummy Variablen: nur Interzept

$\text{StdL} = 5.78 + 2.95m + 0.45\text{Bildg}$



Dummy Variablen: nur Interzept

Österreich 2018 (EU-Silc):



Dummy Variablen: nur Steigung

Interaktionseffekt: $m \times x$

$$\begin{aligned} \widehat{y} &= b_1 + b_2 x + b_3 (m \times x) \\ \widehat{y}|(m=1) &= b_1 + (b_2 + b_3) x \\ \widehat{y}|(m=0) &= b_1 + b_2 x \end{aligned}$$

- Marginale Effekte (Steigungen):

$$\frac{\partial \widehat{y}|(m=1)}{\partial x} = b_2 + b_3; \qquad \frac{\partial \widehat{y}|(m=0)}{\partial x} = b_2$$

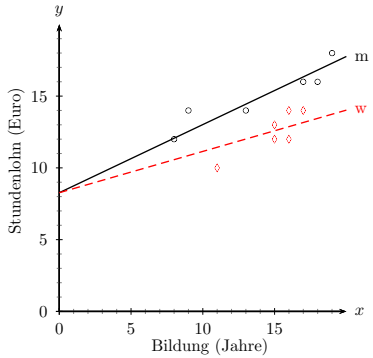
- *Unterschied der Steigungen*: b_3

$$\frac{\partial \widehat{y}|(m=1)}{\partial x} - \frac{\partial \widehat{y}|(m=0)}{\partial x} = b_3$$

Dummy Variablen: nur Steigung

Beispiel:

$$\widehat{\text{StdL}} = 8.27 + 0.29\,\text{Bildg} + 0.19(m \times \text{Bildg})$$



gleiches Interzept: unplausible Annahme!

Dummy Variablen: Interzept **und** Steigung

$$\begin{aligned} \widehat{y} &= b_1 + b_2 x + b_3 m + b_4 (m \times x) \\ \widehat{y}|(m=1) &= (b_1 + b_3) + (b_2 + b_4) x \\ \widehat{y}|(m=0) &= b_1 + b_2 x \end{aligned}$$

Unterschied zwischen den Kategorien:

$$\widehat{y}|(m=1) - \widehat{y}|(m=0) = b_3 + b_4 x$$

Gibt gleiche Koeffizienten wie bei getrennten Regressionen

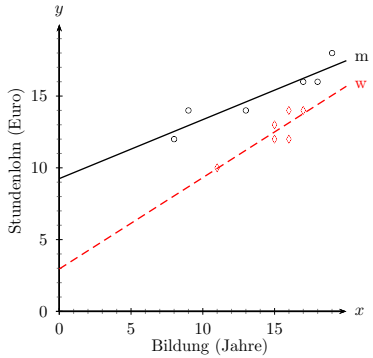
$$\begin{aligned} \text{für } m=0: \quad \widehat{y}^0 &= b_1 + b_2 x \\ \text{für } m=1: \quad \widehat{y}^1 &= c_1 + c_2 x \end{aligned}$$

mit $c_1 = b_1 + b_3$ und $c2 = b_2 + b_4$

Dummy Variablen: Interzept und Steigung

Beispiel:

$$\widehat{\text{StdL}} = 2.95 + 6.30m + 0.64\,\text{Bildg} - 0.23(m \times \text{Bildg})$$



Dummy Variablen: Interzept und Steigung

- Dummyvariablenmodell und getrennte Regressionen liefern exakt gleiche Koeffizienten und Residuen . . . *aber* nicht die gleichen Standardfehler!
- Männer und Frauen (Dummy):

$$\begin{array}{lrcrcrcl} \text{StdL} & = & 2.955 & + & 6.295\, m & + & 0.636\ \text{Bildg} & - & 0.226\ (m * \text{Bildg}) \\ & & (2.721) & & (2.961)^{*} & & (0.18)^{***} & & (0.197) \end{array}$$

$$R^2 = 0.891, \quad n = 12$$

- nur Frauen:

$$\begin{array}{lrcrcl} \text{StdL} & = & 2.955 & + & 0.636\ \text{Bildg} \\ & & (2.595) & & (0.172)^{**} \end{array}$$

$$R^2 = 0.775, \quad n = 6$$

- nur Männer:

$$\begin{array}{lrcrcl} \text{StdL} & = & 9.25 & + & 0.411\ \text{Bildg} \\ & & (1.22)^{***} & & (0.083)^{***} \end{array}$$

$$R^2 = 0.859, \quad n = 6$$

Mehrere Ausprägungen/Kategorien

- **Eine kategoriale Variable mit mehr als zwei Ausprägungen** (z.B. Nationalität, höchster Bildungsabschluss) kann in eine *Referenzkategorie* und einzelne Dummies ‘zerlegt’ werden.
  - Jedem Individuum kann genau eine Ausprägung/Dummy zugeordnet werden.
  - Koeffizient der Dummy misst Unterschied zum Mittelwert der Referenzkategorie!
  - Kann in einer Variable abgebildet werden,

z.B: EU-SILC: `Höchster Bildungsabschluss (P137000)`

```
-3  Weiß nicht
-1  keine Angabe
 1  Pflichtschule
 2  Lehre mit Berufsschule
 3  Fach- oder Handelsschule
 4  Matura
 5  Abschluss an einer Universität, (Fach-)Hochschule
 6  Anderer Abschluss nach der Matura
```


Beispiel 1: Stundenlöhne und Bildung (EU-Silc 2019)

	log(Stundenlöhne)	
	mit Interzept	ohne Interzept
Interzept	2.460^{***}	
Pflichtschule		2.460^{***}
Lehre mit Berufsschule	0.207^{***}	2.667^{***}
Fach- oder Handelsschule	0.274^{***}	2.734^{***}
Matura	0.382^{***}	2.842^{***}
Abschluss an einer Universität, FH	0.590^{***}	3.050^{***}
Anderer Abschluss nach der Matura	0.551^{***}	3.011^{***}
Observations	4,748	4,748
Adjusted R^2	0.169	[0.981]
Note:	*p<0.1; **p<0.05; ***p<0.01	

Beispiel 2: Preise von Gebrauchtautos und Alter

Alter auf ganze Jahre gerundet:

	Abhängige Variable: Autopreis		
	(1)	(2)	(3)
Interzept	22 709	23 567	
AlterJ (Jahre)	−2 517		
AlterJ$= 0$			23 567
AlterJ$= 1$		−4 158	19 409
AlterJ$= 2$		−5 871	17 696
AlterJ$= 3$		−7 785	15 781
AlterJ$= 4$		−11 164	12 403
AlterJ$= 5$		−13 667	9 900
n	40	40	40
R^2	0.82	0.84	[0.99]

Beispiel 2: Preise von Gebrauchtautos und Alter (gerundet)

Vergleich (Whg.):

	AlterJ = 0	AlterJ = 1	AlterJ = 2	AlterJ = 3	AlterJ = 4	AlterJ = 5
	24000	19980	21850	18000	10000	11100
	23900	18890	14500	17200	10000	6700
P	22800	18890	16900	15200	15300	11900
r		20100	15600	14450	14320	
e		19700	18600	15000	12350	
i		19300	18500	16900	12500	
s		19000	18500	15500	12350	
e			18000	14000		
			17500			
			16900			
			17700			
			17800			
n	3	7	12	8	7	3
$\overline{y}$	23567	19409	17696	15781	12403	9900
$\Delta\overline{y}$		-4158	-1713	-1915	-3378	-2503
$\widehat{y}$	22709	20192	17675	15158	12641	10124
$\Delta\widehat{y}$		-2517	-2517	-2517	-2517	-2517

Mehrere Ausprägungen/Kategorien

- **Mehrere kategoriale Variablen**
  - Jedes Individuum kann in mehrere *sich nicht gegenseitig ausschließende* Kategorien fallen (z.B. Geschlecht und Familienstand)
  - Wenn *jede mögliche Kombination* eigene Dummy (*'saturated model'*) ⇒ Koeffizienten messen Mittelwerte (bzw. Abweichungen von Referenzkategorie), sonst *'omitted variables'*.
  - Benötigt mehrere Variablen zur Abbildung,

    z.B. Merkmale: *männlich* und *verheiratet*

		männlich	
		ja (1)	nein (0)
ver-	ja (1)	...	...
heiratet	nein (0)	...	...

Mehrere kategoriale Variablen: Beispiel (Wiederholg.)

i	StdL	m	w	v	u	Bildg
1	16	1	0	0	1	17
2	12	0	1	0	1	16
3	16	1	0	1	0	18
4	14	1	0	1	0	13
5	12	1	0	0	1	8
6	12	0	1	1	0	15
7	18	1	0	1	0	19
8	14	0	1	0	1	17
9	14	0	1	1	0	16
10	14	1	0	1	0	9
11	10	0	1	1	0	11
12	13	0	1	0	1	15

Mittelw. StdL: $\overline{\text{StdL}} = (16+12+16+\cdots+13)/12 = 13.75$

$$m_i = \begin{cases} 1 & \text{wenn Mann, und} \\ 0 & \text{sonst} \end{cases} \qquad v_i = \begin{cases} 1 & \text{wenn verheiratet,} \\ 0 & \text{sonst} \end{cases}$$

Bedingte Mittelwerte von StdL:

$$\begin{aligned} (\overline{\text{StdL}}|m=1) &= (16+16+14+12+18+14)/6 = 15 \\ (\overline{\text{StdL}}|w=1) &= (12+12+14+14+10+13)/6 = 12.5 \\ (\overline{\text{StdL}}|v=1) &= (16+14+12+\cdots+10)/7 = 14 \\ (\overline{\text{StdL}}|v=0) &= (16+12+12+14+13)/5 = 13.4 \end{aligned}$$

Mehrere kategoriale Variablen: Beispiel

		männlich	
		ja (1)	nein (0)
ver-	ja (1)	15.5	12
heiratet	nein (0)	14	13

Zwei äquivalente Spezifikationsmöglichkeiten:

- alle Interaktionen:

$$\widehat{\text{StdL}} = 14.0 + 1.5\,mv - 1.0\,wu - 2.0\,wv$$

- Haupt- und Interaktionseffekte:

$$\widehat{\text{StdL}} = 13.0 + 1.0\,m - 1.0\,v + 2.5\,(m \times v)$$

Mehrere kategoriale Variablen: Beispiel

$$\widehat{\text{StdL}} = 13.0 + 1.0\,m - 1.0\,v + 2.5\,(m \times v)$$

- $(\widehat{\text{StdL}}|m = 0, v = 0) = 13.0 + 1.0 \times 0 - 1.0 \times 0 + 2.5\,(0 \times 0) = 13$
- $(\widehat{\text{StdL}}|m = 1, v = 0) = 13.0 + 1.0 \times 1 - 1.0 \times 0 + 2.5\,(1 \times 0) = 13 + 1 = 14$
- $(\widehat{\text{StdL}}|m = 0, v = 1) = 13.0 + 1.0 \times 0 - 1.0 \times 1 + 2.5\,(0 \times 1) = 12$
- $(\widehat{\text{StdL}}|m = 1, v = 1) = 13.0 + 1.0 \times 1 - 1.0 \times 1 + 2.5\,(1 \times 1) = 15.5$

- **Dummies wirken wie *'Ein- Ausschalter'*!**

Thanx . . .

Beispiel: Heterogenität und Simpsons Paradoxon

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Beispiel: Heterogenität und Simpsons Paradoxon

- Untersuchungsgegenstände unterscheiden sich in der Regel in unzähligen Merkmalen, und aus praktischen Gründen sind wir fast immer gezwungen, einen Großteil dieser Heterogenität in unseren Untersuchungen zu ignorieren.
- Solche *nicht berücksichtigte Heterogenität* (*unobserved heterogeneity*) kann dramatische Folgen haben, wie das folgende hypothetische Beispiel zeigt.

Beispiel: Heterogenität und Simpsons Paradoxon

Angenommen, eine Universität mit zwei überfüllten Studiengängen 'Psychologie' und 'Mathematik' richtet Aufnahmeprüfungen ein.

Resultat:

	Frauen	Männer
Bewerber	500	500
davon aufgenommen	200	300
in Prozent	40%	60%

Ist die Zulassungsprüfung unfair gegenüber Frauen???

Beispiel: Heterogenität und Simpsons Paradoxon

Schauen wir uns die Ergebnisse für die beiden Kurse separat an:

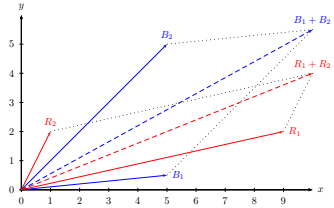
Psychologie	Frauen	Männer
Bewerber	100	400
davon aufgenommen	80	280
in Prozent	80%	70%

Mathematik	Frauen	Männer
Bewerber	400	100
davon aufgenommen	120	20
in Prozent	30%	20%

Wie ist das möglich??? $\rightarrow$ **Simpson Paradoxon**

Beispiel: Heterogenität und Simpsons Paradoxon

- Dieses Paradoxon wurde nach Edward Simpson benannt, der 1951 die Möglichkeit eines solchen paradoxen Ergebnisses veröffentlichte.
- Dies ist jedoch nur ein Sonderfall von *omitted variables*, die bereits viel früher von Karl Pearson (1899) und insbesondere Udny Yule (1903) beschrieben wurden.
- Grafische Veranschaulichung am Beispiel einer Vektoraddition



Beide roten Vektoren sind steiler als die blauen Vektoren, aber die *Vektorsumme* der blauen Vektoren ist steiler als die rote *Vektorsumme*!

Beispiel: Heterogenität und Simpsons Paradoxon

- Regression: 3 Dummy Variablen
  - aufgenommen ja/nein
  - weiblich ja/nein
  - Psychologie ja/nein
- Eine Regression auf Dummy Variablen liefert die Anteile, und da es sich um ein gesättigtes Dummy Variablenmodell handelt, erhalten wir die exakten Anteile.

Beispiel: Heterogenität und Simpsons Paradoxon

Abhängige Variable: aufgenommen		
Konstante	0.600^{***}	0.200^{***}
	(0.022)	(0.035)
fem	-0.200^{***}	0.100^{***}
	(0.031)	(0.035)
psych		0.500^{***}
		(0.035)
Beobachtungen	1,000	1,000
R^2	0.040	0.200

'Omitted variable': $b_2^{(1)} = b_2^{(2)} + b_3^{(2)} \frac{\text{cov(fem,psych)}}{\text{var(psych)}} = 0.1 + 0.5 * (-0.6) = -0.2$

Danke!

Paneldaten: Das LSDV & *'fixed effects'* Modell

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Paneldaten

Querschnitts- und Zeitdimension:

z.B. für 3 Individuen, 4 Zeitperioden, 2 Regressoren (x_2, x_3)
im *'wide'* Format:

	$i=1$ $\mathbf{y_1}$	$i=2$ $\mathbf{y_2}$	$i=3$ $\mathbf{y_3}$	$i=1$ $\mathbf{x_{21}}$	$i=2$ $\mathbf{x_{22}}$	$i=3$ $\mathbf{x_{23}}$	$i=1$ $\mathbf{x_{31}}$	$i=2$ $\mathbf{x_{32}}$	$i=3$ $\mathbf{x_{33}}$
$t=1$	y_{11}	y_{21}	y_{31}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
$t=2$	y_{12}	y_{22}	y_{32}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
$t=3$	y_{13}	y_{23}	y_{33}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
$t=4$	y_{14}	y_{24}	y_{34}	x_{214}	x_{224}	x_{234}	x_{314}	x_{324}	x_{334}
mean	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	$\bar{y}_{3\bullet}$	$\bar{x}_{21\bullet}$	$\bar{x}_{22\bullet}$	$\bar{x}_{23\bullet}$	$\bar{x}_{31\bullet}$	$\bar{x}_{32\bullet}$	$\bar{x}_{33\bullet}$

Paneldaten

Querschnitts- und Zeitdimension: (3 Indiv., 4 Zeitperioden)

zum Beispiel:

	GDP_DEU	GDP_AUT	GDP_ITA	C_DEU	C_AUT	C_ITA	I_DEU	I_AUT	I_ITA
2020	y_{11}	y_{21}	y_{31}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
2021	y_{12}	y_{22}	y_{32}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
2022	y_{13}	y_{23}	y_{33}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
2023	y_{14}	y_{24}	y_{34}	x_{214}	x_{224}	x_{234}	x_{314}	x_{324}	x_{334}

Paneldaten

Querschnitts- und Zeitdimension:

z.B. für 3 Individuen, 4 Zeitperioden, 2 Regressoren (x_2, x_3)

	$i=1$ $\mathbf{y_1}$	$i=2$ $\mathbf{y_2}$	$i=3$ $\mathbf{y_3}$	$i=1$ $\mathbf{x_{21}}$	$i=2$ $\mathbf{x_{22}}$	$i=3$ $\mathbf{x_{23}}$	$i=1$ $\mathbf{x_{31}}$	$i=2$ $\mathbf{x_{32}}$	$i=3$ $\mathbf{x_{33}}$
$t=1$	y_{11}	y_{21}	y_{31}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
$t=2$	y_{12}	y_{22}	y_{32}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
$t=3$	y_{13}	y_{23}	y_{33}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
$t=4$	y_{14}	y_{24}	y_{34}	x_{214}	x_{224}	x_{234}	x_{314}	x_{324}	x_{334}
mean	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	$\bar{y}_{3\bullet}$	$\bar{x}_{21\bullet}$	$\bar{x}_{22\bullet}$	$\bar{x}_{23\bullet}$	$\bar{x}_{31\bullet}$	$\bar{x}_{32\bullet}$	$\bar{x}_{33\bullet}$

Verschiedene Schätzmöglichkeiten:

- Querschnittsregression über *Zeitmittelwerte*: $\rightarrow$ *'between model'*

$$\bar{y}_{i\bullet} = b_1 + b_2\bar{x}_{2i\bullet} + b_3\bar{x}_{3i\bullet} + e_i$$

- eigene Zeitreihenregression für jedes Individuum
- Pool Regression (s. nächste Seite)

Paneldaten

Reorganisation der Daten: *'long'* Format (3 Indiv., 4 Zeitperioden)

i	t	y	x_2	x_3
1	1	y_{11}	x_{211}	x_{311}
1	2	y_{12}	x_{212}	x_{312}
1	3	y_{13}	x_{213}	x_{313}
1	4	y_{14}	x_{214}	x_{314}
2	1	y_{21}	x_{221}	x_{321}
2	2	y_{22}	x_{222}	x_{322}
2	3	y_{23}	x_{223}	x_{323}
2	4	y_{24}	x_{224}	x_{324}
3	1	y_{31}	x_{231}	x_{331}
3	2	y_{32}	x_{232}	x_{332}
3	3	y_{33}	x_{233}	x_{333}
3	4	y_{34}	x_{234}	x_{334}

i	t	GDP	Cons	Inv
DEU	2020	y_{11}	x_{211}	x_{311}
DEU	2021	y_{12}	x_{212}	x_{312}
DEU	2022	y_{13}	x_{213}	x_{313}
DEU	2023	y_{14}	x_{214}	x_{314}
AUT	2020	y_{21}	x_{221}	x_{321}
AUT	2021	y_{22}	x_{222}	x_{322}
AUT	2022	y_{23}	x_{223}	x_{323}
AUT	2023	y_{24}	x_{224}	x_{324}
ITA	2020	y_{31}	x_{231}	x_{331}
ITA	2021	y_{32}	x_{232}	x_{332}
ITA	2022	y_{33}	x_{233}	x_{333}
ITA	2023	y_{34}	x_{234}	x_{334}

Paneldaten: gepoolt

Pool-Modell: ('*stacked*'), z.B. 3 Individuen, 4 Zeitperioden ($i = 1, \ldots, 3,\ t = 1, \ldots, 4$)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_{2\,11} \\ x_{2\,12} \\ x_{2\,13} \\ x_{2\,14} \\ x_{2\,21} \\ x_{2\,22} \\ x_{2\,23} \\ x_{2\,24} \\ x_{2\,31} \\ x_{2\,32} \\ x_{2\,33} \\ x_{2\,34} \end{pmatrix} + b_3 \begin{pmatrix} x_{3\,11} \\ x_{3\,12} \\ x_{3\,13} \\ x_{3\,14} \\ x_{3\,21} \\ x_{3\,22} \\ x_{3\,23} \\ x_{3\,24} \\ x_{3\,31} \\ x_{3\,32} \\ x_{3\,33} \\ x_{3\,34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

R package: `reshape2`

Paneldaten

Notation: Pool-Modell

$$y_{it} = b_1 + b_2 x_{2it} + \cdots b_h x_{hit} + \cdots b_k x_{kit} + e_{it}$$

$i = 1, \ldots, n \ \rightarrow$ Individuen
$t = 1, \ldots, T \ \rightarrow$ Zeit
$h = 1, \ldots, k \ \rightarrow$ Variablen

x_{hit} ... Beobachtung von Variable h von Individuum i zum Zeitpunkt t

insgesamt $n \times T$ Beobachtungen (für *'balanced panel'*)

Paneldaten

Mehrere Möglichkeiten für Schätzung, z.B.

- gepoolt über alle Beobachtungen (OLS), unterstellt gleiche Koeffizienten für alle Individuen und Perioden!
- gepoolt, aber individuenspezifische Varianzen (GLS) $\rightarrow$ 'random effects model'
- Individuenspezifische Interzepte, aber identische Steigungskoeffizienten: **LSDV** (*'Least Squares Dummy Variable' Modell*); numerisch äquivalent zu *'fixed effects model'* (FWL-Theorem); sehr häufig angewandt!

Paneldaten: LSDV

‘Least Squares Dummy Variable Model’: individuenspezifische Dummies!

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + b_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_4 \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \\ x_{31} \\ x_{32} \\ x_{33} \\ x_{34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

Dummies kontrollieren für alles, was sich im Beobachtungszeitraum nicht geändert hat (*zeitinvariante* Effekte)!

Paneldaten: LSDV

‘Least Squares Dummy Variable Model’ ↔ ‘Fixed Effects Model’

- FWL-Theorem: LSDV Modell kann alternativ (d.h. numerisch äquivalent) und einfacher mit individuenspezifisch mittelwerttransformierten Daten geschätzt werden
  - Regression auf Dummyvariablen liefert Mittelwerte
    $\rightarrow$ Residuen sind mittelwerttransformierte Werte
  - Einfache Mittelwerttransformation ist rechnerisch einfacher zu lösen als Regression auf Dummies.
  - FWL-Theorem erlaubt Zerlegung: Regression auf mittelwerttransformierte Daten liefert identische Steigungskoeffizienten wie LSDV Modell!
    $\Rightarrow$ ***fixed effects* Modell** (*‘within’* Transformation)
  - viel einfacher in der Durchführung,
    Dummy-Koeffizienten sind selten von Interesse
  - *Achtung:* Standardfehler unterscheiden sich!
    (Freiheitsgrade; Programme korrigieren automatisch dafür)

Paneldaten: LSDV

‘Least Squares Dummy Variable Model’ ↔ ‘Fixed Effects Model’

- Schreibweise: LSDV

$$y_{it} = b_1 + \sum_{i=2}^{n} a_i d_i + b_2 x_{2it} + \cdots + b_k x_{kit} + e_{it}$$

(Beachte: Dummies d_i benötigen keinen Zeitindex!)

oder kürzer (*‘fixed effects’*)

$$y_{it} = c_i + b_2 x_{2it} + \cdots + b_k x_{kit} + e_{it}$$

mit $c_i = b_1 + \sum_{i=2}^{n} a_i d_i$

Paneldaten: LSDV

LSDV & *fixed effcts* Modell:

- Vorteil: Individuen-Dummies kontrollieren für *alle zeitinvarianten* Effekte!
- deshalb in der Regel *mehr Kontrolle* als gepooltes Modell, aber immer noch keine perfekte Kontrolle $\Rightarrow$ *keine* Kausalaussagen!
- Nachteil: Zeitinvariante Effekte können *nicht einzeln* geschätzt werden!
- z.B. Lohngleichung: $\log(\text{StdL}) = b_1 + b_2 \text{Bildg} + b_3 \text{Erf} + b_4 \text{weibl} + e$

  Geschlecht und absolvierte Ausbildung sind in der Regel *zeitinvariant*!
  $\Rightarrow \quad \log(\text{StdL})_{it} = c_i + b_2 \text{Erf}_{it} + e_{it}$ (alle zeitinvarianten Effekte in c_i!)
- LSDV und *‘fixed effects’* beruhen auf *‘within’* Streuung!
- Details & weitere Beispiele folgen . . .

Thanx . . .

Beispiel: Difference-in-Differences

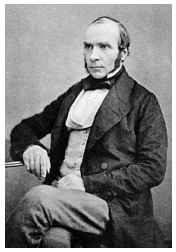
Grundlagen der Ökonometrie

www.hsto.info/econometrics

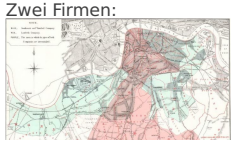
Difference-in-Differences (DiD)

- 1853/-54 Cholera Ausbruch in Sohio (London), ca. 14 000 Tote
- **John Snow** (1813-1858): Arzt & Pionier des Diff-in-diff Ansatzes

  Doppelte Differenz: Treatment – Control & Vorher – Nachher



Zwei Firmen:



Hot-spots:

Intervention:

Difference-in-Differences (DiD)

- DiD: Methode für Kausalanalysen
- **Kausalität:** viele philosoph. Definitionen und Ansätze; derzeit wichtig . . .

  Kontrafaktischer Vergleich einer Intervention: was wäre passiert, wenn Maßnahme (*Treatment*) *nicht* ergriffen worden wäre → Kontrollgruppe
- ‘Doppelte’ Differenzen: Treatment – Control & Before – After

	Treatment Group	Control Group
Before	T_B	C_B
After	T_A	C_A

“Difference-in-Differences” $= (T_A - T_B) - (C_A - C_B)$

Difference-in-Differences (DiD)

- Mit Regression:

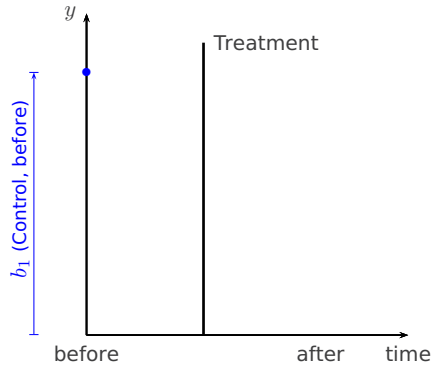
$$y_i = b_1 + b_2 \text{treat} + b_3 \text{after} + b_4 (\text{treat} \times \text{after}) + b_5 x_i + e_i$$

$$\text{treat} = \begin{cases} 1 & \text{wenn in ‘Treatment Group’,} \\ 0 & \text{wenn in ‘Control Group’.} \end{cases} \qquad \text{after} = \begin{cases} 0 & \text{vor ‘Treatment’,} \\ 1 & \text{nach ‘Treatment’.} \end{cases}$$

- Koeffizient des *Interaktionsterms* zwischen der Treatment- und After-Dummy misst Difference-in Differences

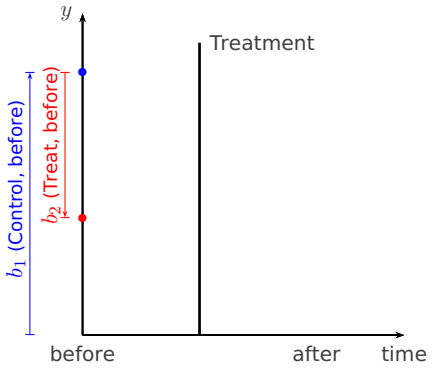
	Treatment Group	Control Group	Difference
Before	$b_1 + b_2 + b_5 x$	$b_1 + b_5 x$	b_2
After	$b_1 + b_2 + b_3 + b_4 + b_5 x$	$b_1 + b_3 + b_5 x$	$b_2 + b_4$
Difference	$b_3 + b_4$	b_3	b_4

Difference-in-Differences: $\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$



$\widehat{y}_i = b_1$

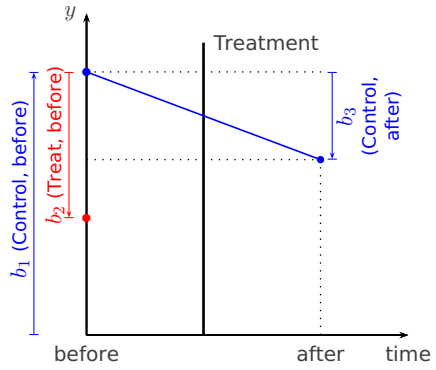
Difference-in-Differences: $\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$



$\widehat{y}_i = b_1 + b_2\text{treat}$

b_2: Diff. between Control und Treatment before!

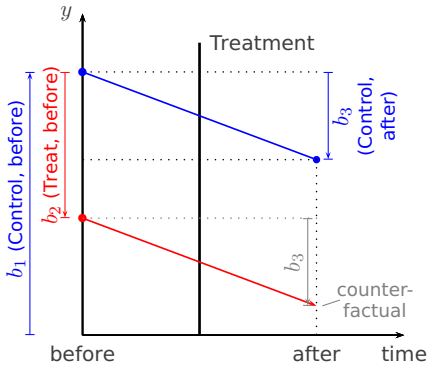
Difference-in-Differences: $\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$



$\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after}$

$b_1 + b_3$: Control after

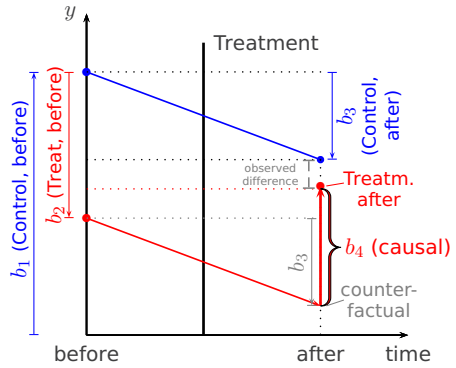
Difference-in-Differences: $\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$



$\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after}$

$b_1 + b_2 + b_3$: Counterfactual! *(NOT observable!)*

Difference-in-Differences: $\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$



$$\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$$

b_4: measures "causal effect"! (if all assumpt. fullfilled!)

Difference-in-Differences (DiD)

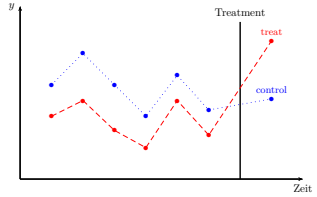
Annahmen:

- Parallel Trends Assumption: *Ohne* Treatment wäre in der Treatment Gruppe (annähernd) das gleiche passiert wie in der Control Gruppe (dies impliziert u.a. keine *omitted variables* u.ä.)
- Keine Antizipation: keine Erwartungsbildung

Difference-in-Differences (DiD)

Annahmen:

- weder *Parallel Trends Assumption* noch *keine Antizipation* können direkt empirisch überprüft werden!
  (das *Counterfactual* ist immer unbeobachtbar!)
- aber man kann prüfen, ob die Trends von Control- und Treatment Gruppe *vor* dem Treatment annähernd parallel verlaufen sind!



- kein Beweis, aber stärkt das Vertrauen . . .

Difference-in-Differences: Beispiel

Beschäftigungseffekte von Mindestlöhnen (Card & Krueger 1994)

Erhöhung des Mindestlohns in New Jersey (NJ) April 1992.
Benachbartes Pennsylvania (PA) dient als Kontrollgruppe.

	State		
	PA	NJ	Diff.
Feb	23.33	20.44	-2.89
Nov	21.17	21.03	-0.14
Diff.	2.17	-0.59	2.75

Difference-in-Differences: Beispiel

Beschäftigungseffekte von Mindestlöhnen (Card & Krueger 1994)

$$\begin{array}{lcccccc} \text{Empl} & = & 23.33 & - & 2.89\ \text{NJ} & - & 2.17\ \text{Nov} & + & 2.754\ \text{Nov}\times\text{NJ} \\ & & (1.072)^{***} & & (1.194)^{**} & & (1.516) & & (1.688) \end{array}$$

$R^2 = 0.007, \quad n = 794$

(Standardfehler in Klammern)

Mit firmen-fixen Effekten ($\rightarrow$ LSDV Modell):

$$\begin{array}{lcccc} \text{Emp} & = & -2.283\ \text{Nov} & + & 2.75\ \text{Nov}\times\text{NJ} \\ & & (1.036)^{**} & & (1.154)^{**} \end{array}$$

$R^2 = 0.015, \quad n = 794$

(firmen-fixe Effekte, Standardfehler in Klammern)

Difference-in-Differences (DiD)

Literaturtipp:

Manning, Alan (2021) "The Elusive Employment Effect of the Minimum Wage." *Journal of Economic Perspectives*, 35 (1): 3-26.

`https://www.aeaweb.org/articles?id=10.1257/jep.35.1.3`

Thanks . . .

Logarithmische Transformationen

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Marginale Effekte im linearen Modell

Wiederholung: Lineares Modell

$$\widehat{y} = b_1 + b_2 x_2 + \cdots + b_h x_h + \cdots + b_k x_k$$

$$\frac{\partial \widehat{y}}{\partial x_h} = b_h = \left.\frac{\Delta \widehat{y}}{\Delta x_h}\right|_{\text{ceteris paribus}}$$

- $\Delta x_h = 1$; ceteris paribus: alle *anderen* Regressoren konstant!
  d.h. $\Delta x_1 = \ldots = \Delta x_{h-1} = \Delta x_{h+1} = \ldots = \Delta x_k = 0$;
- Interpretation: Wenn x_h um *eine Einheit* (z.B. Euro) zunimmt verändert sich $\widehat{y}$ *ceteris paribus* um b_h *Einheiten* (z.B. Kilogramm).
- *Hinweis:* geben Sie bei der Interpretation immer die konkreten Einheiten an!
- *Achtung:* Die ceteris paribus Annahme gilt *nur* für berücksichtigte Variablen!

Logarithmische Transformation:

Drei wichtige Eigenschaften:

1. Exponentialfunktionen werden durch Logarithmierung zu linearen Funktionen

$$\begin{aligned} \log(xy) &= \log(x) + \log(y) \qquad \text{für } x, y > 0 \\ \log\left(\frac{x}{y}\right) &= \log(x) - \log(y); \quad \log(1/x) := \log(x^{-1}) = -\log(x) \\ \log(x^a) &= a \log(x) \end{aligned}$$

2. Differenz zwischen zwei logarithmierten Werten entspricht näherungsweise der *relativen* Änderung der ursprünglichen Werte, d.h.

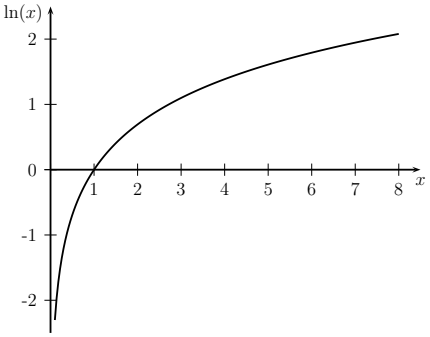
$$\Delta \log x := \log(x_2) - \log(x_1) \approx \frac{x_2 - x_1}{x_1} := \frac{\Delta x}{x} \qquad \text{für } x_2, x_1 > 0$$

3. Durch Logarithmierung werden kleine Zahlenwerte gespreizt, große Zahlenwerte gestaucht

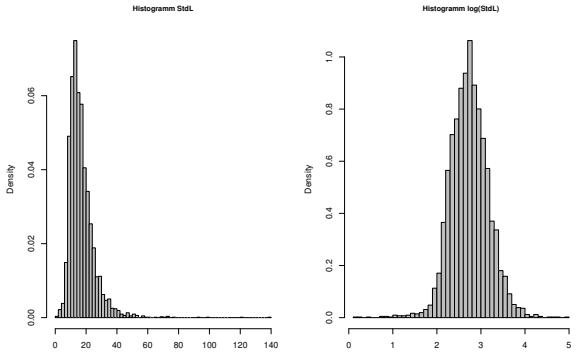
Logarithmische Transformation

Durch Logarithmierung werden kleine Zahlenwerte gespreizt, große Zahlenwerte gestaucht:

$$\begin{aligned} \log(0) &= -\infty \\ \log(0.000001) &= -13.816 \\ \log(0.01) &= -4.605 \\ \log(0.1) &= -2.303 \\ \log(1) &= 0 \\ \log(10) &= 2.303 \\ \log(100) &= 4.605 \\ \log(1000) &= 6.908 \\ \log(1000000) &= 13.816 \end{aligned}$$



Logarithmische Transformation



$\log$(StdL) sind symmetrischer verteilt als StdL $\rightarrow$ `mean`(StdL) als Kennzahl!

1) Log-log Modell: Marginaler Effekt

Zur Erinnerung: $\boxed{\Delta \log(y) \approx \frac{\Delta(y)}{y}, \quad \Delta \log(x) \approx \frac{\Delta(x)}{x}}$

(d.h. logarithm. Differenz ist ungefähr relative Änderung)

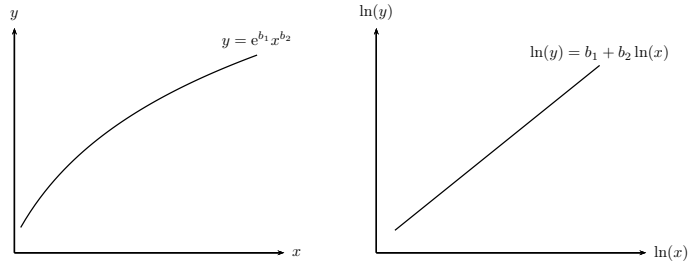
Log-log Modell:

$$\widehat{\log y} = b_1 + b_2 x_2 + \cdots + b_h \log x_h + \cdots + b_k x_k$$

$$b_h = \frac{\Delta \widehat{\log y}}{\Delta \log x_h}\bigg|_{\text{c.p.}} \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\frac{\Delta x_h}{x_h}} = \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\frac{\Delta x_h}{x_h} \times 100} := \frac{\%\Delta \widehat{y}}{\%\Delta x} \rightarrow \text{Elastizität}$$

- Interpretation: Wenn x_h um *ein Prozent* zunimmt verändert sich $\widehat{y}$ ceteris paribus um b_h *Prozent*.

1) Log-log Modell: Marginaler Effekt



Bei log-log Funktionen $\log y = b_1 + b_2 \log x$ ist nicht die Steigung dy/dx konstant, sondern die *Elastizität*

$$\frac{d \log y}{d \log x} = \frac{\frac{dy}{y}}{\frac{dx}{x}} = \frac{dy}{dx} \frac{x}{y} = b_2$$

ist über den gesamten Funktionsverlauf konstant!

1) Log-log Modell

Beispiel: Cobb-Douglas Produktionsfunktion (1928)

$$\widehat{\log(Q_i)} = \underset{(0.434)}{-0.177} + \underset{(0.145)}{0.807} \log(L_i) + \underset{(0.064)}{0.233} \log(K_i)$$

$$R^2 = 0.957, \quad n = 24 \ \ (1899 - 1922)$$

(Standardfehler in Klammern)

$$\frac{\partial \widehat{\log(Q)}}{\partial \log(L)} \approx \frac{\frac{\Delta \widehat{Q}}{\widehat{Q}}}{\frac{\Delta L}{L}} := \frac{\%\Delta \widehat{Q}}{\%\Delta L} = 0.807$$

$\Rightarrow$ Wenn der Arbeitsinput L um *ein Prozent* erhöht wird, nimmt der gefittete Output $\widehat{Q}$ ceteris paribus um ca. 0.8 *Prozent* zu (Elastizität).

2) Log-level Modell: Marginaler Effekt

Log-level Modell: (links $\log$, rechts level)

$$\widehat{\log y} = b_1 + b_2 x_2 + \cdots + b_h x_h + \cdots + b_k x_k$$

$$b_h = \left. \frac{\Delta \widehat{\log y}}{\Delta x_h} \right|_{\text{c.p.}} \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\Delta x_h}$$

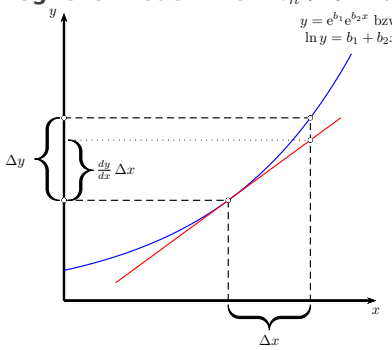
oder

$$100 \times b_h \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\Delta x_h} = \frac{\%\Delta \widehat{y}}{\Delta x_h}$$

- Interpretation: Wenn x_h um *eine Einheit* (z.B. Euro) zunimmt verändert sich $\widehat{y}$ ceteris paribus um $100 \times b_h$ *Prozent*.
- Wenn $b_h > 0.1$ Korrektur erforderlich!

2) Log-level Modell

Log-level Modell: Wenn $b_h > 0.1$ Korrektur erforderlich



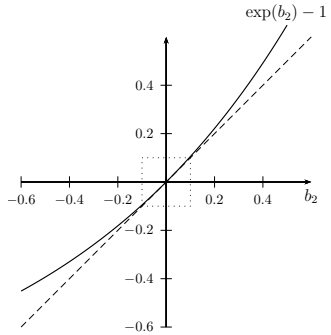
Interpretation wenn $b_h > 0.1$: Wenn x_h um *eine Einheit* (z.B. Euro) zunimmt ändert sich $\widehat{y}$ ceteris paribus um $[\exp(b_h) - 1] \times 100$ *Prozent*.

$$\%\Delta \widehat{y} := \left(\frac{\Delta \widehat{y}}{\widehat{y}} \right) \times 100 = [\exp(b_h) - 1] \times 100 \qquad \text{für } \Delta x_h = 1$$

2) Log-level Modell

Für kleine Werte von b_2 (z.B. $|b_2| \leq 0.1$) ist der Unterschied zwischen b_2 und $\exp(b_2) - 1$ oft vernachlässigbar:

b_2	$\exp(b_2) - 1$
-0.50	-0.393
-0.30	-0.259
-0.10	-0.095
-0.01	-0.010
0.01	0.010
0.05	0.051
0.10	0.105
0.30	0.350
0.50	0.649



2) Log-level Modell

Beispiel 1: österreichische EU-Silc Daten (2018)

$$\begin{array}{rccccc} \log(\text{StdL}) = & 1.871 & + 0.037\ \text{potBildg} & + 0.028\ \text{Erf} & - 0.000366\ \text{Erf}^2 \\ & (0.035)^{***} & (0.002)^{***} & (0.002)^{***} & (0.00005)^{***} \end{array}$$

$$R^2 = 0.15, \qquad n = 4104$$

$$\frac{\Delta\widehat{\log(\text{StdL})}}{\Delta\text{potBildg}} = 0.037 \quad \rightarrow \quad \frac{\%\Delta\widehat{\log(\text{StdL})}}{\Delta\text{potBildg}} = 100 \times 0.037 = 3.7\%$$

oder etwas genauer (log-level)

$$(\exp(0.037) - 1) * 100 = 3.77\%$$

Durchschnittl. Stundenlohn (StdL) nimmt c.p. mit jedem zusätzlichen Jahr potentieller Bildung (potBildg) um ca. 3.77% zu.

2) Log-level Modell

Beispiel 2: Wachstumsraten y wächst mit konstanter Wachstumsrate g

$$y_t = y_0(1+g)^t \quad \Rightarrow \quad \log y_t = \log y_0 + \log(1+g) \times t$$

$$\log y_t = \underbrace{\log y_0}_{b_1} + \underbrace{\log(1+g)}_{b_2} \times \text{Trend}_t$$

Schätzung:

$$\log \widehat{y}_t = b_1 + b_2 \text{Trend}_t$$

$$b_2 = \log(1+g) \quad \text{folgt} \quad g = \exp(b_2) - 1$$

Prozentuelle durchschnittliche diskrete Wachstumsrate:

$$\boxed{g\% := g \times 100 = [\exp(b_2) - 1] \times 100}$$

2) Log-level Modell

Beispiel 2: Wachstumsrate China 1995 – 2019
Wachstum BIP pro Kopf (Datenquelle: World dataBank, WDI; Dependent variable: GDP, PPP, constant 2005 international dollar.)

$$\begin{array}{rccc} \log(\text{GDPpc}) & = & - 160.669 & + & 0.084\ \text{Trend} \\ & & (2.71)^{***} & & (0.001)^{***} \end{array}$$

$$R^2 = 0.994, \qquad n = 25 \quad (1995 - 2019)$$

(Standardfehler in Klammern)

$$g\% = (\exp(0.084) - 1) \times 100 \approx 8.81 \textit{ Prozent}$$

Pro Kopf Einkommen verdoppelte sich ca. alle $70/8.8 \approx 8$ Jahre!

3) Level-log Modell: Marginaler Effekt

Level-log Modell: (links level, rechts log)

$$\widehat{y} = b_1 + b_2 x_2 + \cdots + b_h \log x_h + \cdots + b_k x_k$$

$$b_h = \frac{\Delta \widehat{y}}{\Delta \log x_h}\bigg|_{\text{c.p.}} \approx \frac{\Delta \widehat{y}}{\frac{\Delta x_h}{x_h}}$$

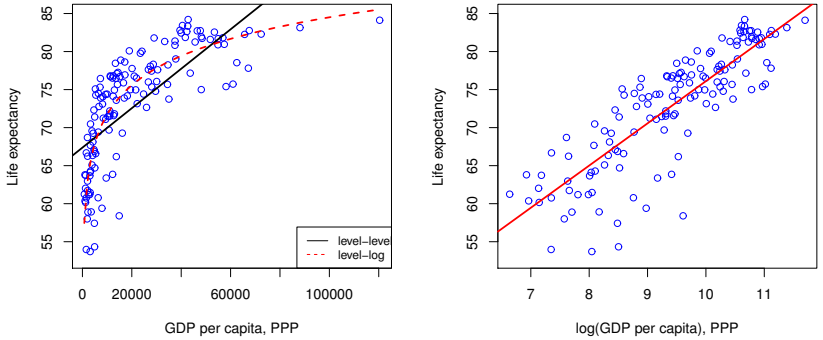
oder

$$\frac{1}{100} \times b_h \approx \frac{\Delta \widehat{y}}{\frac{\Delta x_h}{x_h} \times 100} = \frac{\Delta \widehat{y}}{\%\Delta x_h}$$

- Interpretation: Wenn x_h um *ein Prozent* zunimmt ändert sich $\widehat{y}$ ceteris paribus um $0.01 \times b_h$ *Einheiten* (z.B. Kilogramm).

3) Level-log Modell

Beispiel: Lebenserwartung und Einkommen, Länderquerschnitt (WDI 2015)



3) Level-log Modell: Beispiel

	Dependent variable:	
	LifeExp	
	(1)	(2)
Constant	19.303***	66.820***
	(2.936)	(0.724)
log(GDPpc)	5.683***	
	(0.314)	
GDPpc/1000		0.266***
		(0.026)
Observations	149	149
R^2	0.691	0.425
Note:	*p<0.1; **p<0.05; ***p<0.01	

Wann logarithmieren?

Mindestanforderung: alle Ausprägungen strikt positiv!

- Häufig logarithmiert werden . . .
  - monetäre Variablen, wie z.B. BIP, Vermögen, Löhne
  - Variablen, die Bestände messen, z.B. Bevölkerung, Flächen, . . .
  - . . .
- Eher *selten bis nicht* logarithmiert werden . . .
  - Variablen mit Zeitdimension
  - Veränderungsraten
  - . . .

- nicht in Stein gemeisselt . . .

Prozent vs. Prozent*punkte*

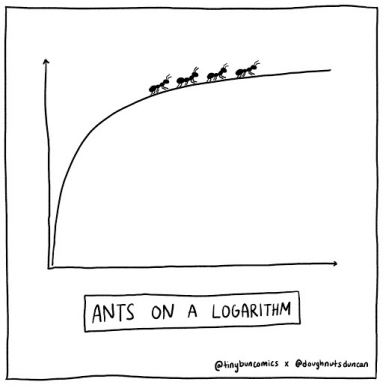
- Zunahme von 5 auf 6: 20% $(= (6-5)/5 * 100)$, aber 1 Prozent*punkt*!
- Anteil: A (mit $0 \leq A \leq 1$)

$$\log(\widehat{y}) = b_1 + b_2 A$$

$$b_2 \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\Delta A} = \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\Delta A \times 100} = \frac{\text{prozentuelle Änderung von } \widehat{y}}{\text{Zunahme von } A \text{ um einen Prozent}punkt}$$

(genauer: $[\exp(b_2) - 1] \times 100$)

- Beobachtungen mit (vielen) Nullen: schwieriger . . .
  (Tobit- oder Poisson Modelle)



Thanx . . .

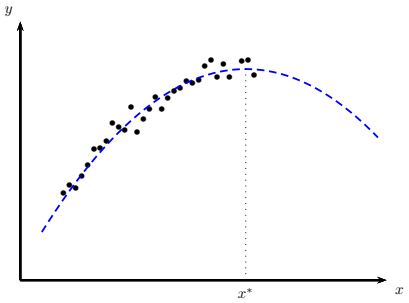
Quadratische Funktionen & Interaktionen

Grundlagen der Ökonometrie

www.hsto.info/econometrics

Quadratische Modelle und Polynome

Quadratische Modelle: $\widehat{y} = b_1 + b_2 x + b_3 x^2$



Steigung ist nicht konstant, sondern ändert sich mit x!

Marginaler Effekt:

$$\frac{\partial \widehat{y}}{\partial x} = b_2 + 2b_3 x$$

Man kann den marginalen Effekt z.B. im Mittelwert von x berechnen, oder den marginalen Effekt grafisch zeigen.

Vorsicht bei Prognosen!

Quadratische Modelle: Beispiel

Beispiel: österreichische EU-Silc Daten (2018)

$$\begin{array}{lccccc} \text{log(StdL)} = & 1.871 & + 0.037\ \text{potBildg} & + 0.028\ \text{Erf} & - 0.000366\ \text{Erf}^2 \\ & (0.035)^{***} & (0.002)^{***} & (0.002)^{***} & (0.00005)^{***} \end{array}$$

$$R^2 = 0.15, \qquad n = 4104$$

$$\frac{\Delta \widehat{\log(\text{StdL})}}{\Delta \text{Erf}} = 0.028 - 2 * 0.000366\,\text{Erf}$$

Wenn Erf = 0 (also für Berufseinsteiger) erwarten wir im ersten Jahr ceteris paribus eine Zunahme des Stundenlohnes um ca. 2.8%
(für eine genauere Abschätzung siehe Abschnitt log-level Modelle im Skript).

Dieser marginale Effekt wird mit zunehmender Erfahrung kleiner.

Quadratische Modelle: Beispiel

Beispiel: österreichische EU-Silc Daten (2018)

$$\begin{array}{lccccc} \text{log(StdL)} = & 1.871 & + 0.037\ \text{potBildg} & + 0.028\ \text{Erf} & - 0.000366\ \text{Erf}^2 \\ & (0.035)^{***} & (0.002)^{***} & (0.002)^{***} & (0.00005)^{***} \end{array}$$

$$R^2 = 0.15, \qquad n = 4104$$

Absolut gesehen nimmt der gefittete StdL bis zu einem Maximum bei ca. 38.25 Jahren zu, da

$$\frac{\Delta \widehat{\log(\text{StdL})}}{\Delta \text{Erf}} = 0.028 - 2 * 0.000366\,\text{Erf} = 0$$

$$\text{Erf}^{\max} = \frac{b_3}{-2b_4} = \frac{0.028}{-2(-0.000366)} = 38.25$$

nach diesem Maximum nimmt er absolut gesehen ab!

Interaktions-Modelle

Interaktions-Modelle: Als erklärende Variablen werden *Produkte* erklärender Variablen verwendet (auch Moderatorvariable genannt)

$$\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 x_2 x_3$$

- Die marginalen Effekte sind *nicht konstant*, der marginale Effekt von x_2 hängt vom numerischen Wert von x_3 ab, und umgekehrt.
- **Marginaler Effekt von** x_2

$$\frac{\partial \widehat{y}}{\partial x_2} = b_2 + b_4 x_3; \qquad b_4 = \frac{\frac{\partial \widehat{y}}{\partial x_2}}{\partial x_3} = \frac{\frac{\partial \widehat{y}}{\partial x_3}}{\partial x_2} = \frac{\partial^2 \widehat{y}}{\partial x_2 \partial x_3}$$

- Die marginalen Effekte können wieder im Mittelwert (oder Median) der Variablen angegeben werden, bzw. grafisch dargestellt werden.

  Besser: Alternative Parametrisierung, siehe Skript.

Beispiel: Entwicklungshilfe

- Alte Diskussion: **hilft Entwicklungshilfe?**
- Burnside & Dollar (2000): 'Aid, policies, and growth',
  American Economic Review 90(4), 847–868:
  > *"We find that aid has a positive impact on growth in developing countries with good fiscal, monetary, and trade policies but has little effect in the presence of poor policies."*
- Regression (vereinfacht): (mit $G \ldots$ *growth*, $A \ldots$ *aid*, $P \ldots$ *policy*)

  $$G = b_1 + b_2 A + b_3 P + b_4 (A \times P) + \cdots + e$$

  Marginaler Effekt der Entwicklungshilfe:

  $$\frac{\partial \widehat{G}}{\partial A} = b_2 + b_4 P$$

- Effekte erwiesen sich später als nicht robust . . .

Beispiel: Lohngleichung für Österreich

$$\widehat{\log(\text{StdL})} = b_1 + b_2\,\text{Bildg} + b_3\,\text{Erf} + b_4\,\text{Erf}^2 + b_5\,\text{Bildg} \times \text{Erf}$$

Marginale Effekte:

- Bildung:
$$\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Bildg}} = b_2 + b_5 \text{Erf}$$
- Erfahrung:
$$\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Erf}} = b_3 + 2b_4\,\text{Erf} + b_5 \text{Bildg}$$

Beispiel: Lohngleichung für Österreich

Abhängige Var.: $\log(\text{StdL})$	(1)	(2)
Constant	2.31315^{***}	1.88338^{***}
	(0.05849)	(0.03429)
Bildg	0.00851^{**}	0.04287^{***}
	(0.00342)	(0.00176)
Erf	0.00024	0.01909^{***}
	(0.00379)	(0.00252)
Erf^2	-0.00015^{***}	-0.00015^{***}
	(0.00005)	(0.00005)
$(\text{Bildg} \times \text{Erf})$	0.00151^{***}	
	(0.00016)	
$(\text{Bildg} - \overline{\text{Bildg}}) \times (\text{Erf} - \overline{\text{Erf}})$		0.00151^{***}
		(0.00016)
Observations	4 104	4 104
R^2	0.167	0.167

Beispiel: Lohngleichung für Österreich

- Marginaler Effekt von Bildg. für $\overline{\text{Erf}} = 22.79$
$$\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Bildg}} = 0.00851 + 0.00151 \times 22.79 = 0.04287$$
- Marginaler Effekt von Erf. für $\overline{\text{Bildg}} = 12.51$ und $\overline{\text{Erf}} = 22.79$
$$\begin{aligned}\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Erf}} &= 0.0002371 + 2 \times (-0.000154) \times 22.79 + 0.00151 \times 12.51 \\ &\approx 0.0121\end{aligned}$$
- *Marg. Effekt:* Wenn Erf *im Mittelwert von Erf und Bildg* um ein weiteres Jahr zunimmt, nimmt der gefittete StdL c.p. um ca. 1.21% zu
(genauer: siehe “Quadratische Log-level Modelle” im Skript!)
- Bei Nichtlinearitäten sind marginale Effekte abhängig von Werten anderer Variablen!

Beispiel: Lohngleichung für Österreich

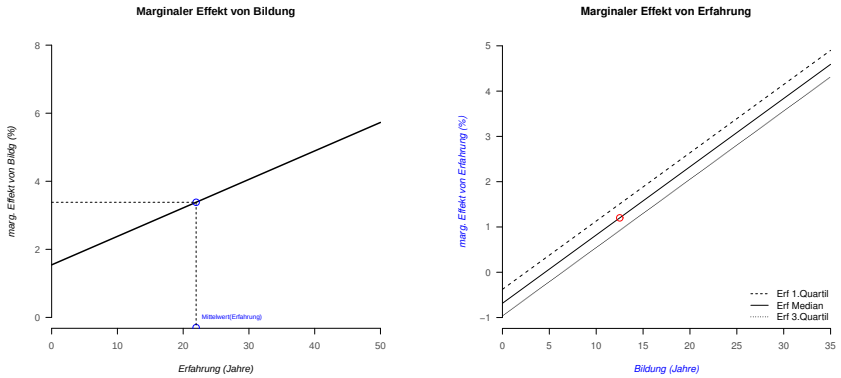
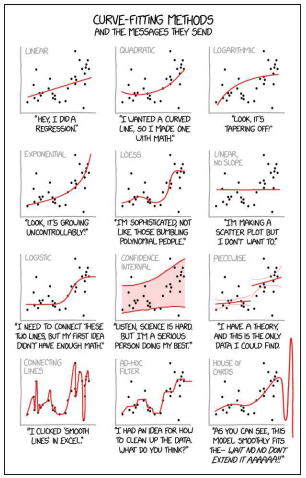


Abbildung: Marginale Effekte mit Interaktionseffekten für Lohngleichung.

Interaktionseffekte

Achtung:

- Interpretieren Sie die Koeffizienten von Interaktionsvariablen nie als unbedingte marginale Effekte!
- Falls Sie an den marginalen Effekten interessiert sind berechnen Sie diese für interessierende Werte der anderen Variablen, oder stellen sie diesen grafisch dar.
- Berechnen Sie den marginalen Effekt für relevante Werte der Variablen, z.B. im Mittelwert oder Median. Eine einfache Möglichkeit dazu bietet eine alternative Parametrisierung (s. Skript).
- Für Signifikanztest: testen Sie die *gemeinsame* Signifikanz der entsprechenden Koeffizienten!



Quelle: xkcd, https://xkcd.com/2048/

Thanx . . .