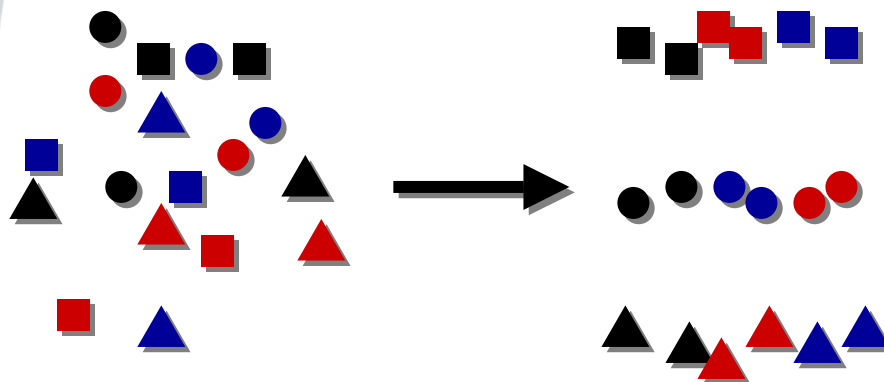


# Statistische Verfahren mit STATA

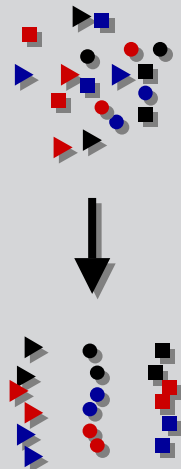
## Teil 1 – Clusteranalyse

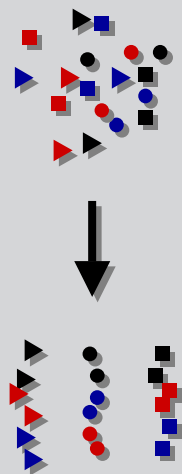


25.01.2004

## Sinn und Zweck

- > struktur-entdeckendes Verfahren
  - ⇒ keine Annahme über Kausalitäten
- > Ziel: Zusammenfassung von Objekten zu in sich ähnlichen Gruppen („Clustern“)
  - ⇒ Objekte einer Gruppe möglichst ähnlich,  
Gruppen untereinander möglichst unähnlich
- > Beispiele:
  - Länder mit ähnlichen wohlfahrtsstaatlichen Arrangements (Esping-Andersen)
  - Personen mit ähnlichen sozio-strukturellen Merkmalen
- > Ergänzung: Diskriminanzanalyse (strukturprüfendes Verfahren)





# Vorgehensweise

## 1. Wahl eines Proximitätsmaßes

- Bestimmung der Ähnlichkeit oder Distanz der Objekte
- Ergebnis: Ähnlichkeits- bzw. Distanzmatrix

## 2. Wahl eines Fusionierungsalgorithmus'

- Zusammenfassung der Objekte zu Gruppen
- Ergebnis: Zuordnungsvariable  $\text{Objekt}$

Eigenschaften

	1	2	3
Objekt 1			
Objekt 2			
Objekt 3			



Objekt

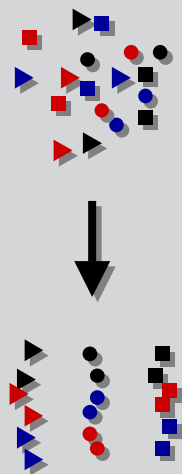
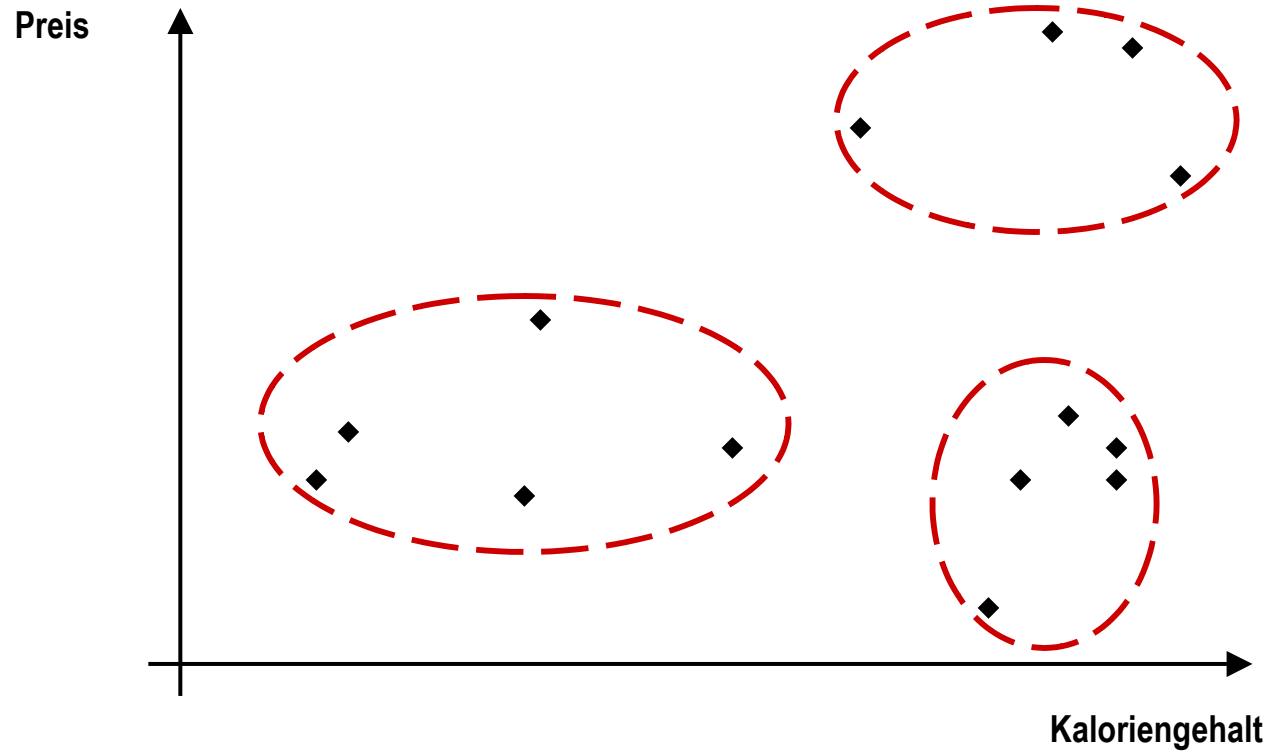
	1	2	3
Objekt 1			
Objekt 2			
Objekt 3			



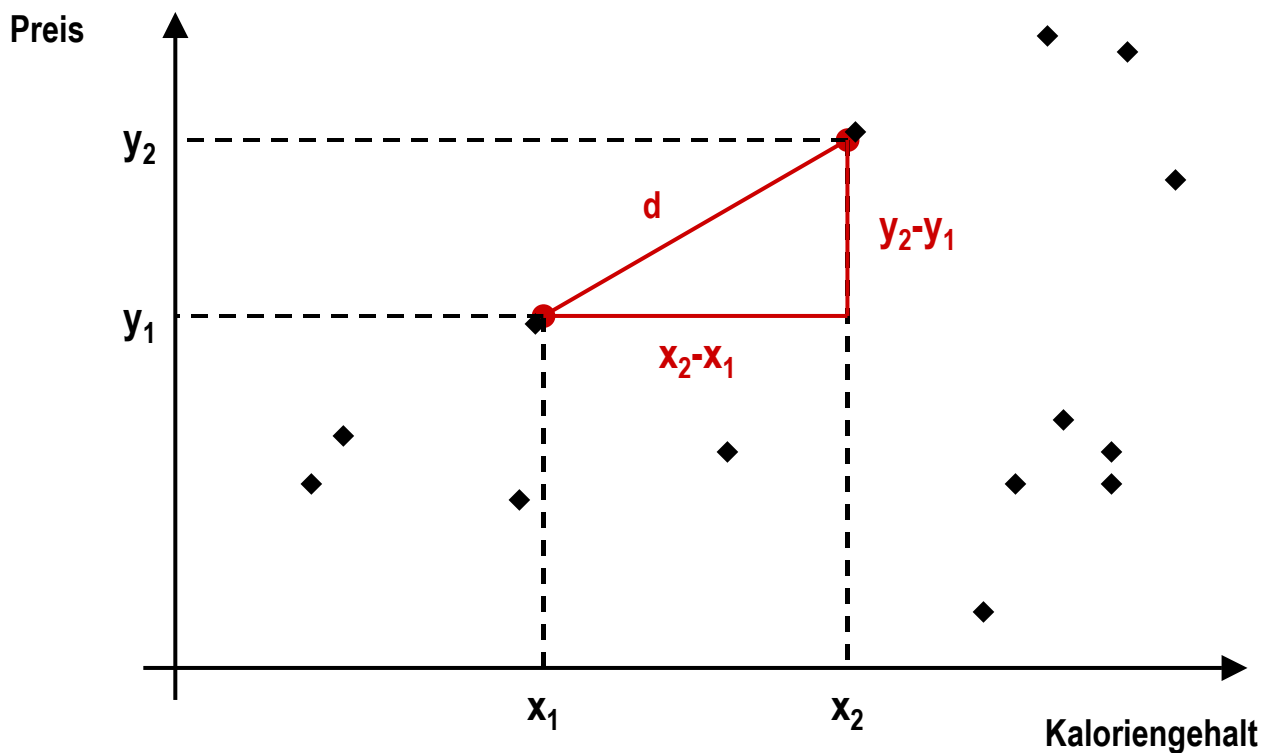
Gruppe

	1	2	3
Objekt 1	X		
Objekt 2	X		
Objekt 3		X	

# Proximitätsmaße - Clustern von Biersorten I

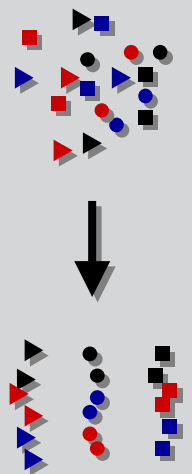
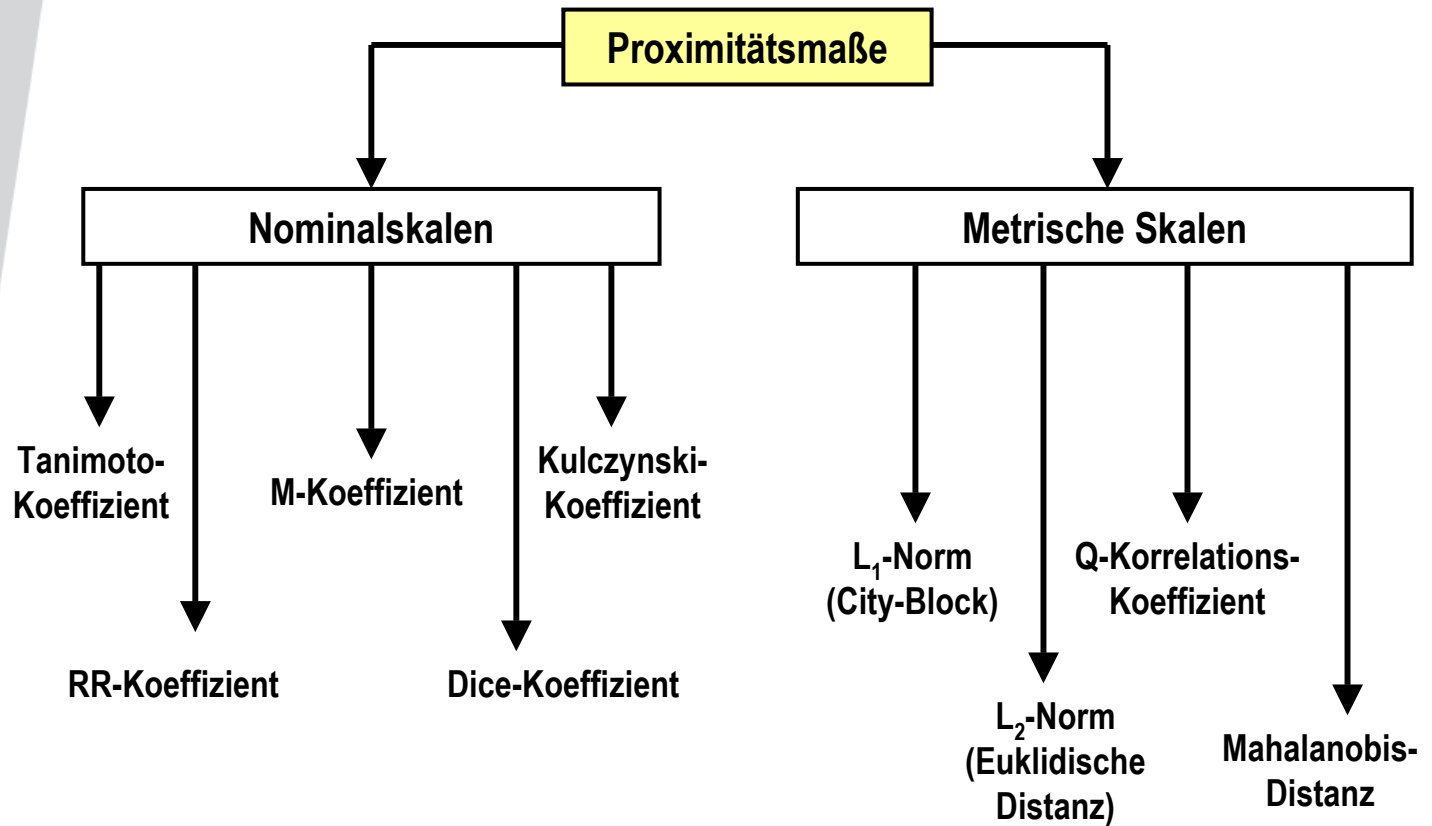


# Proximitätsmaße - Clustern von Biersorten II

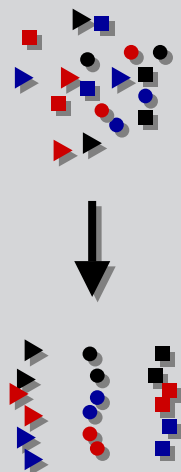
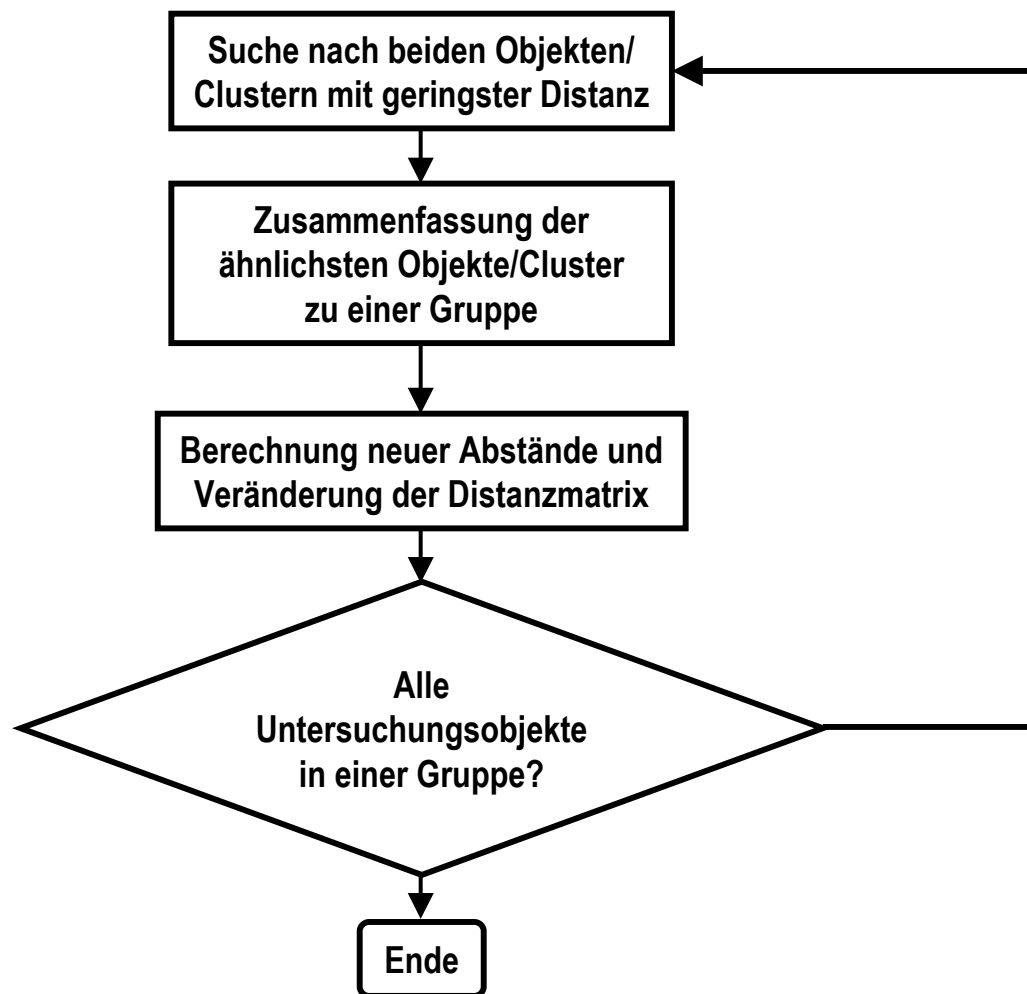


$$a^2 + b^2 = c^2 \Rightarrow (x_2 - x_1)^2 + (y_2 - y_1)^2 = d^2 \Rightarrow d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

# Proximitätsmaße - Übersicht



# Fusionierungsalgorithmen - Beispiel



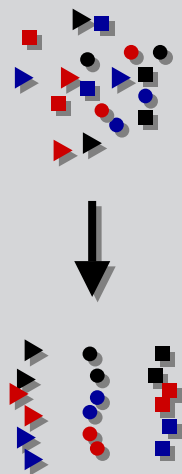
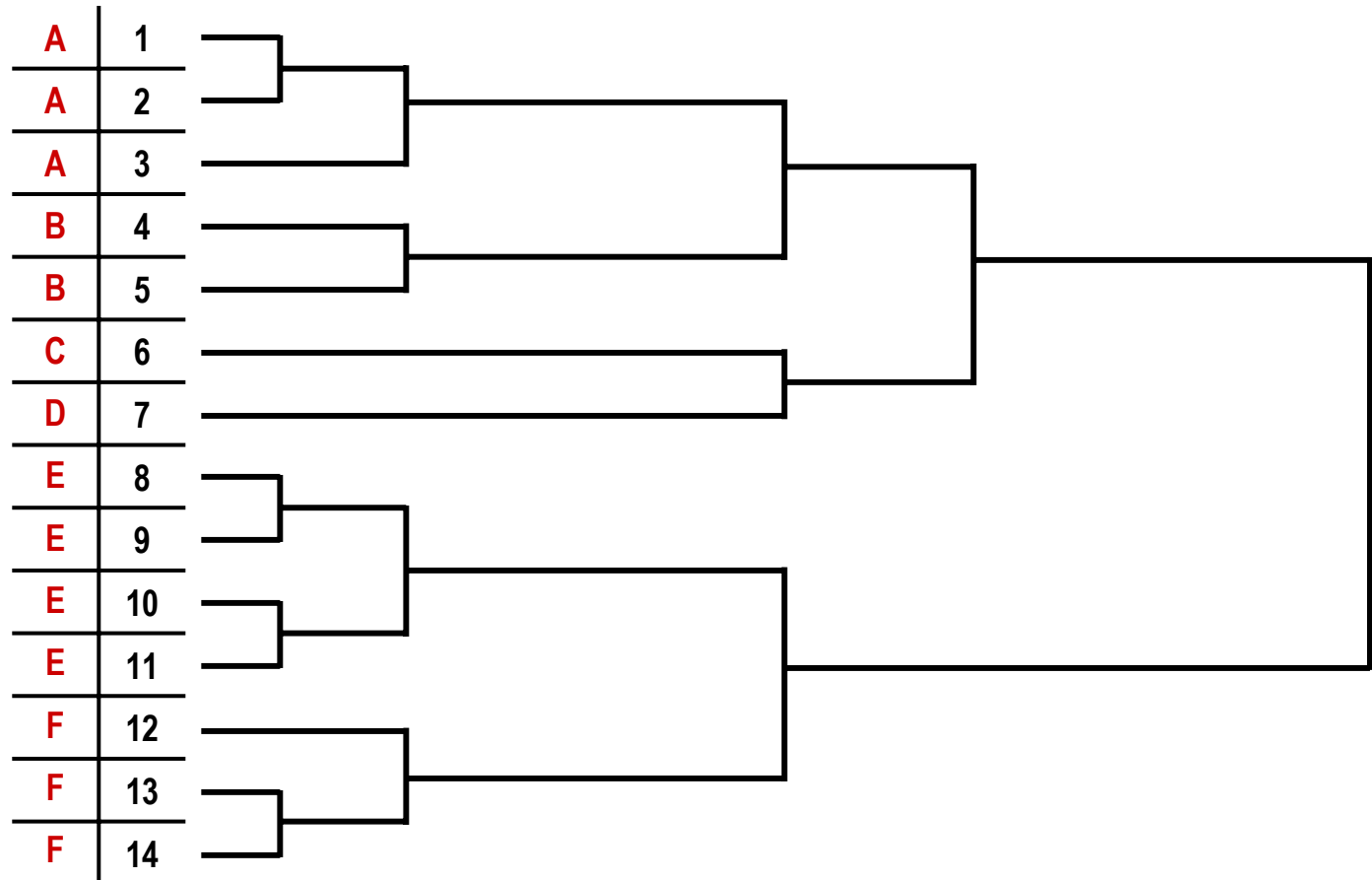
Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

Dozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04

# Fusionierungsalgorithmen - Dendrogramm



Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

Dozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04

# STATA-Befehle

cluster [subcmd]

kmeans

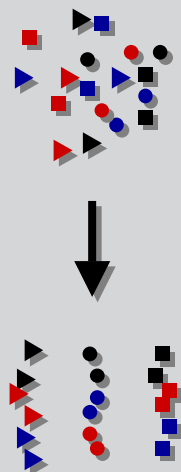
singlelinkage

completelinkage

averagelinkage

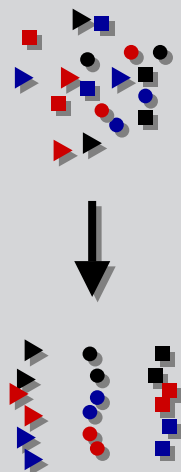
dendogramm

[...]



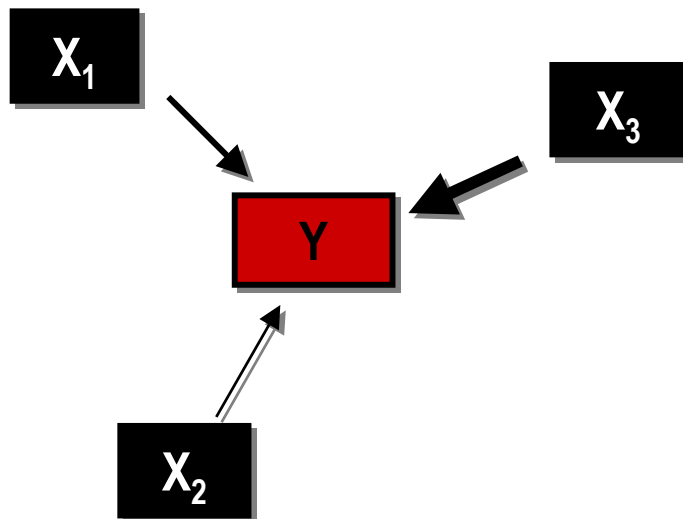
## Literaturhinweise

- > Backhaus u.a. (2000): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung (Kapitel 7); Springer-Verlag, Berlin, Heidelberg, New York; Seite 328-389.
- > Bühl / Zöfel (2000): SPSS Version 9. Einführung in die moderne Datenanalyse unter Windows (Kapitel 20); Addison-Wesley, München u.a.; Seite 434-462.
- > Deichsel / Trampisch (1985): Clusteranalyse und Diskriminanzanalyse; Gustav Fischer Verlag, Stuttgart.

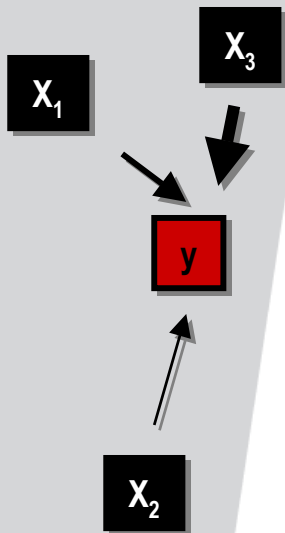


# Statistische Verfahren mit STATA

## Teil 2 – Regressionsanalyse

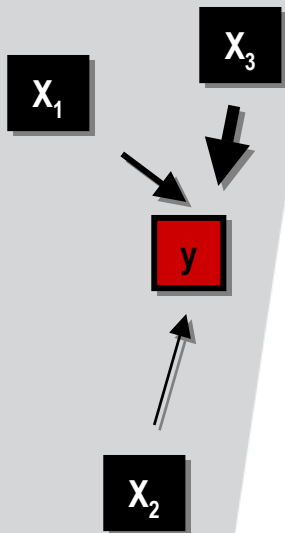


25.01.2004



## Was kann Regression?

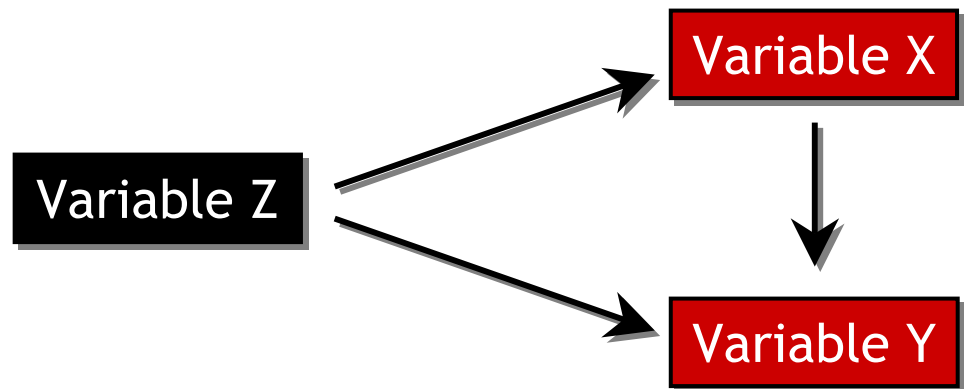
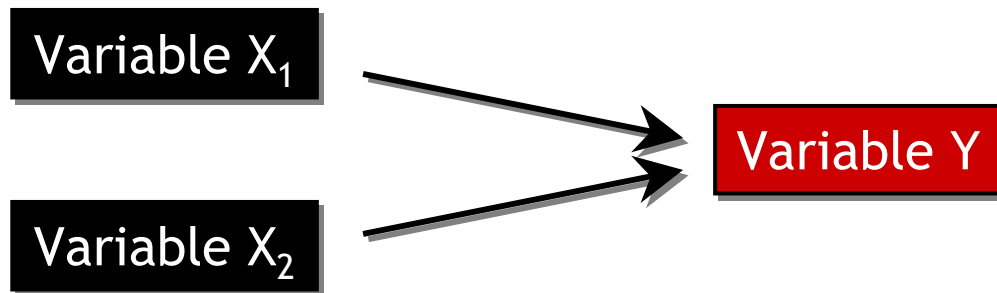
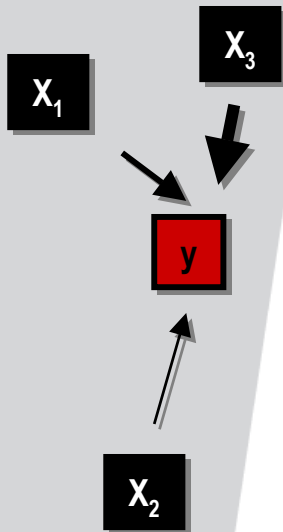
- > struktur-prüfendes Verfahren
  - Kausalität wird postuliert
- > Ziele:
  - Identifizierung eines funktionalen (linearen) Zusammenhanges zwischen  $X$  und  $Y$
  - Prognose der Ausprägung von  $Y$  anhand der Ausprägung von  $X$
  - Bestimmung der Stärke des Einflusses von  $X$  auf  $Y$

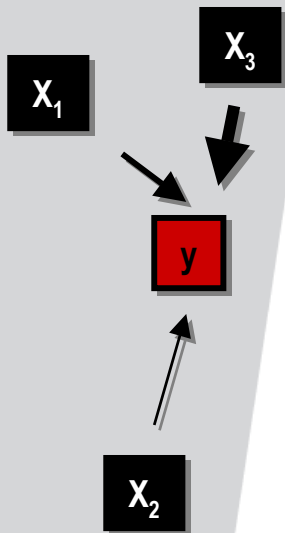


## Vorgehensweise

1. Bestimmung eines Ursache-Wirkungs-Modells (Theorie)
2. Überprüfung der Voraussetzungen für eine lineare Regression (Prämissen)
3. Ermittlung des Zusammenhangs zwischen  $X$  und  $Y$  in einer Stichprobe (Schätzung der Regressionsfunktion)
4. Prüfung der Regressionsfunktion (Hypothesentest)

# Ursache-Wirkungs-Modell





# Prämissen der linearen Regression

- > Verwendung intervallskalierter Variablen
- > Festlegung der abhängigen und der unabhängigen Variablen
- > Unterstellung eines linearen Zusammenhangs
- > [...]

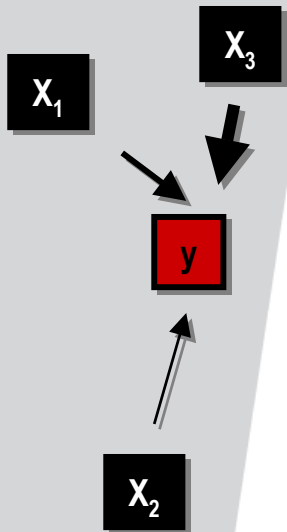
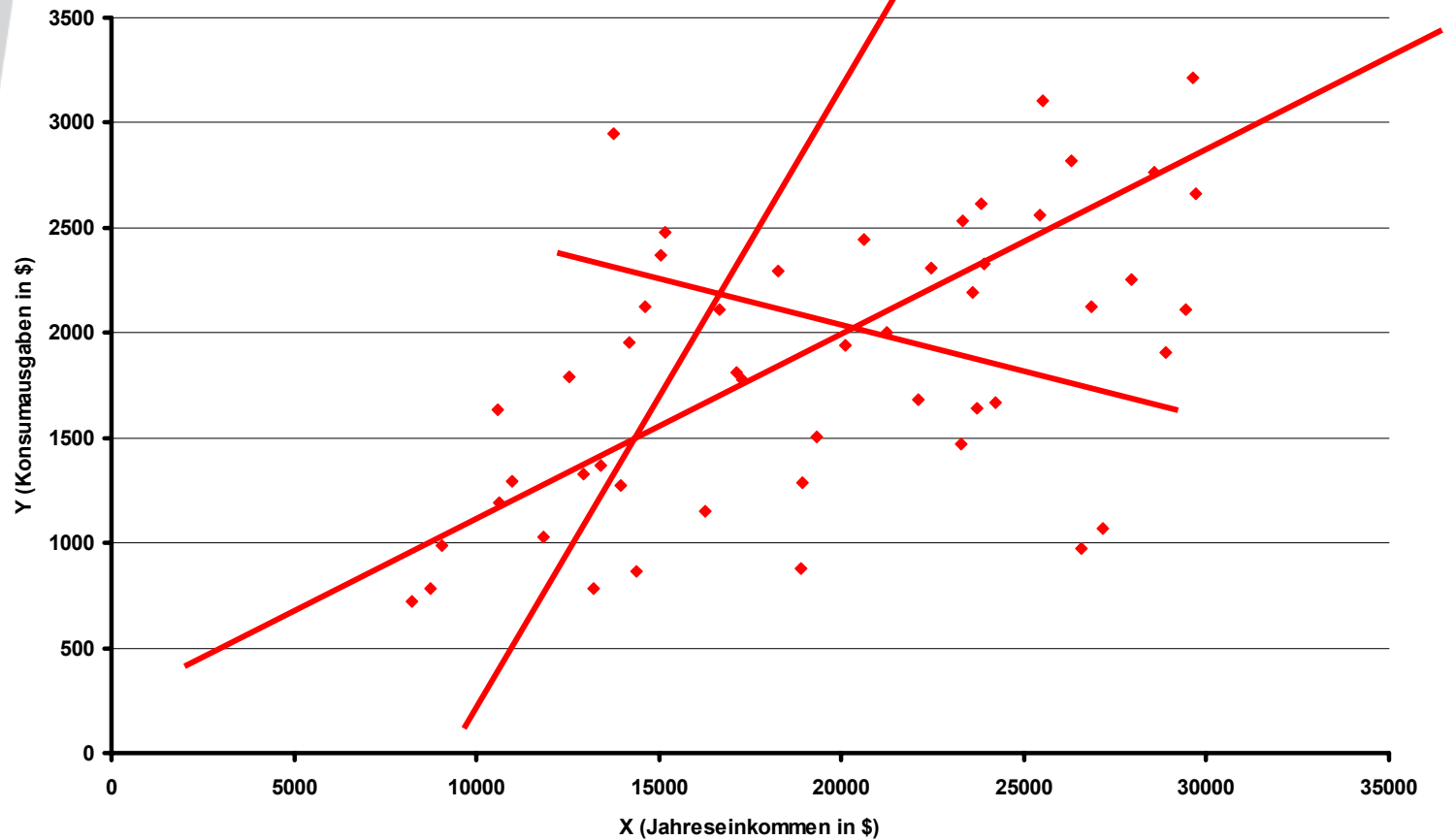
Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

Dozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04

# Streudiagramm I



# Regressionsgleichung I

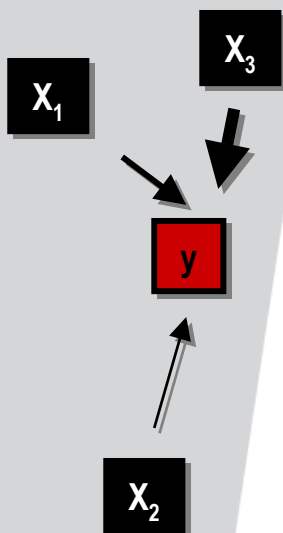
Regressions-  
konstante

Regressions-  
koeffizient

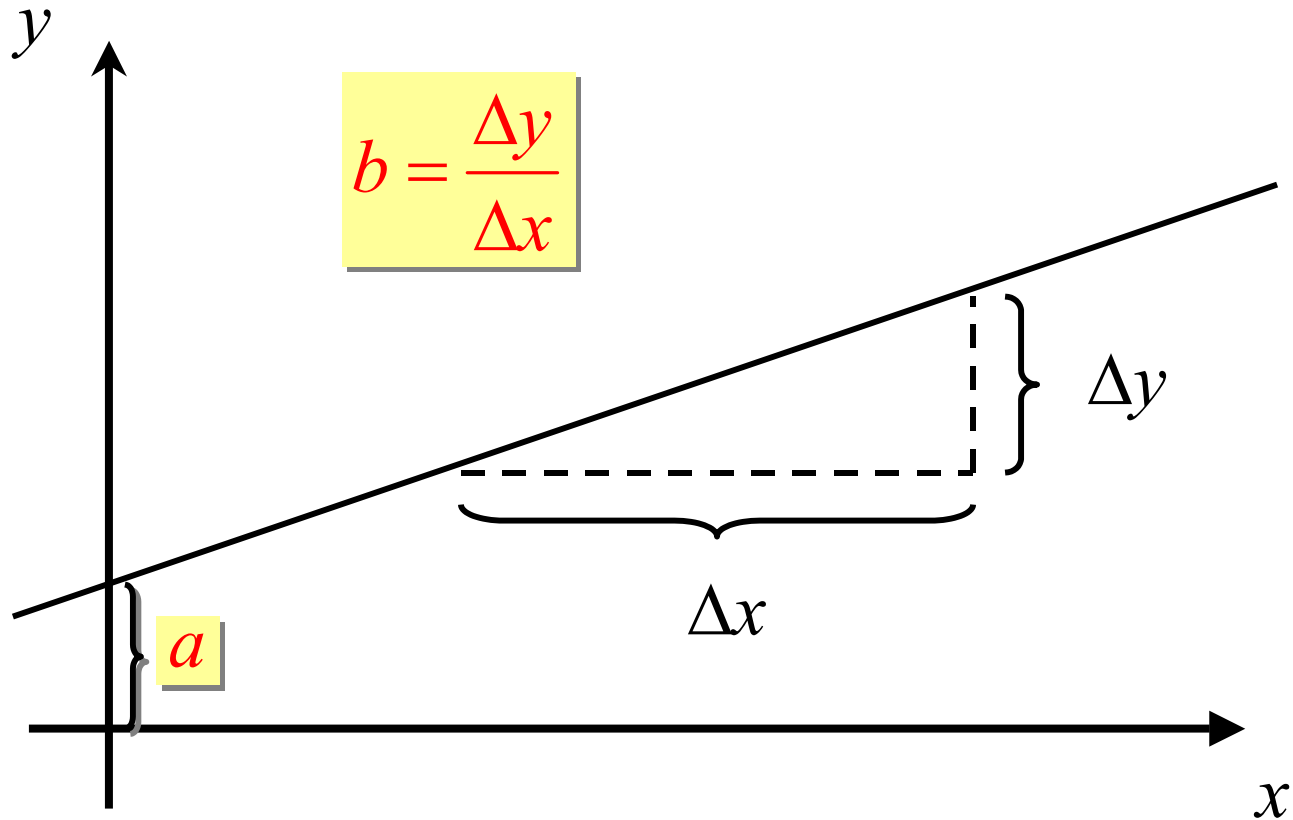
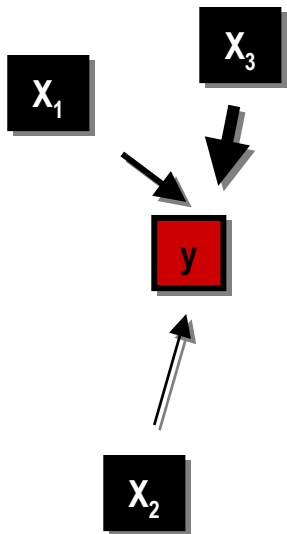
$$y = a + b x$$

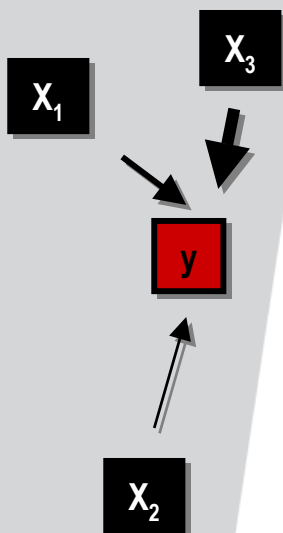
abhängige Variable  
(Regressand,  
endogene Variable)

unabhängige Variable  
(Regressor, exogene  
Variable)



# Regressionsgerade





## Regressionsgleichung II

Zusammenhang in der Realität:

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

Zusammenhang in der Schätzung:

$$\hat{y}_i = a + bx_i + e_i$$

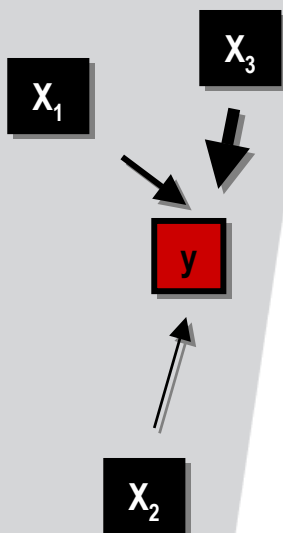
# Residuen I

**Schätzwert**  
(durch Regressionsfunktion  
ermittelt)

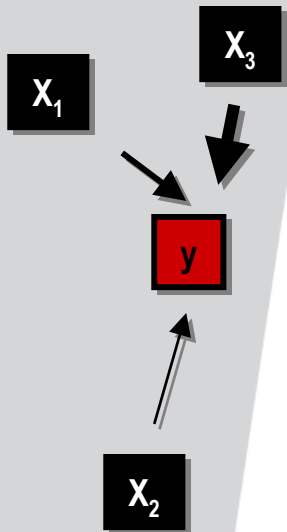
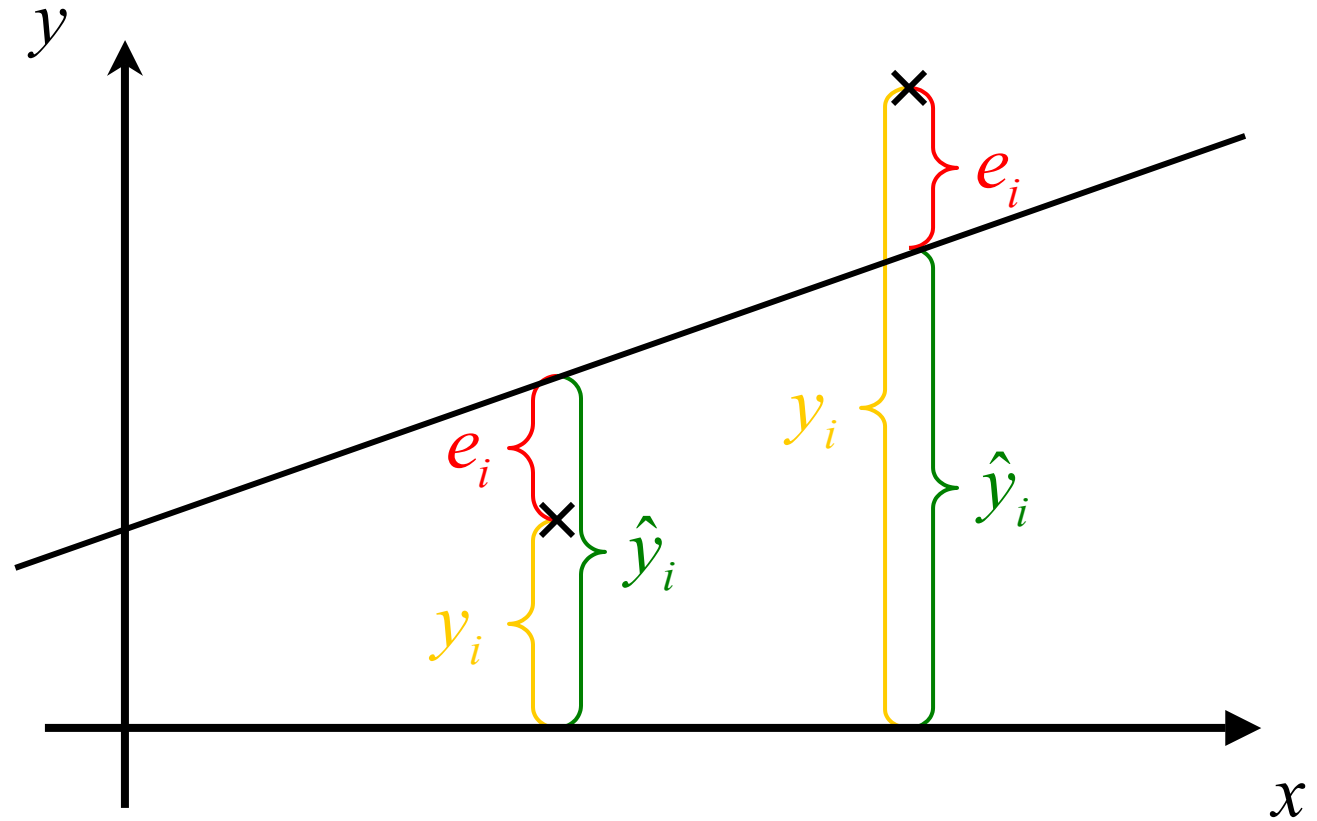
nicht erklärte  
Abweichung

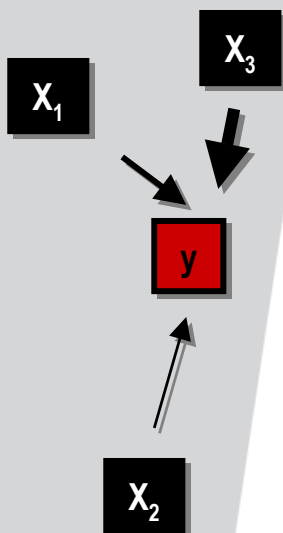
$$y_i - \hat{y}_i = e_i$$

**Beobachtungswert**  
(gemessen bzw. beobachtet)



# Residuen II





## Ordinary Least Squares

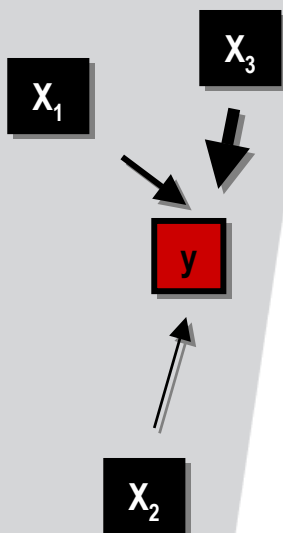
$$\sum_{i=1}^I e_i^2 = \min!$$

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

$$\sum_{i=1}^I e_i^2 = \sum_{i=1}^I (y_i - \hat{y}_i)^2 = \min!$$

$$\hat{y}_i = a + bx_i$$

$$\sum_{i=1}^I e_i^2 = \sum_{i=1}^I [y_i - (a + bx_i)]^2 = \min!$$



## Normalgleichungen

Regressionskoeffizient:

$$b = \frac{I(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{I(\sum x_i^2) - (\sum x_i)^2}$$

Regressionskonstante:

$$a = \bar{y} - b\bar{x}$$

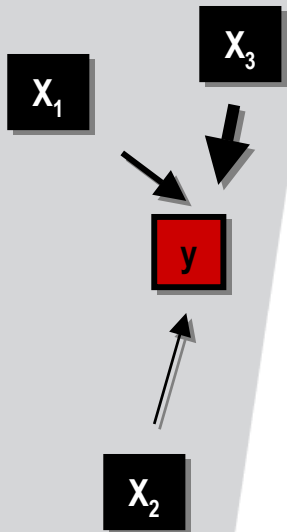
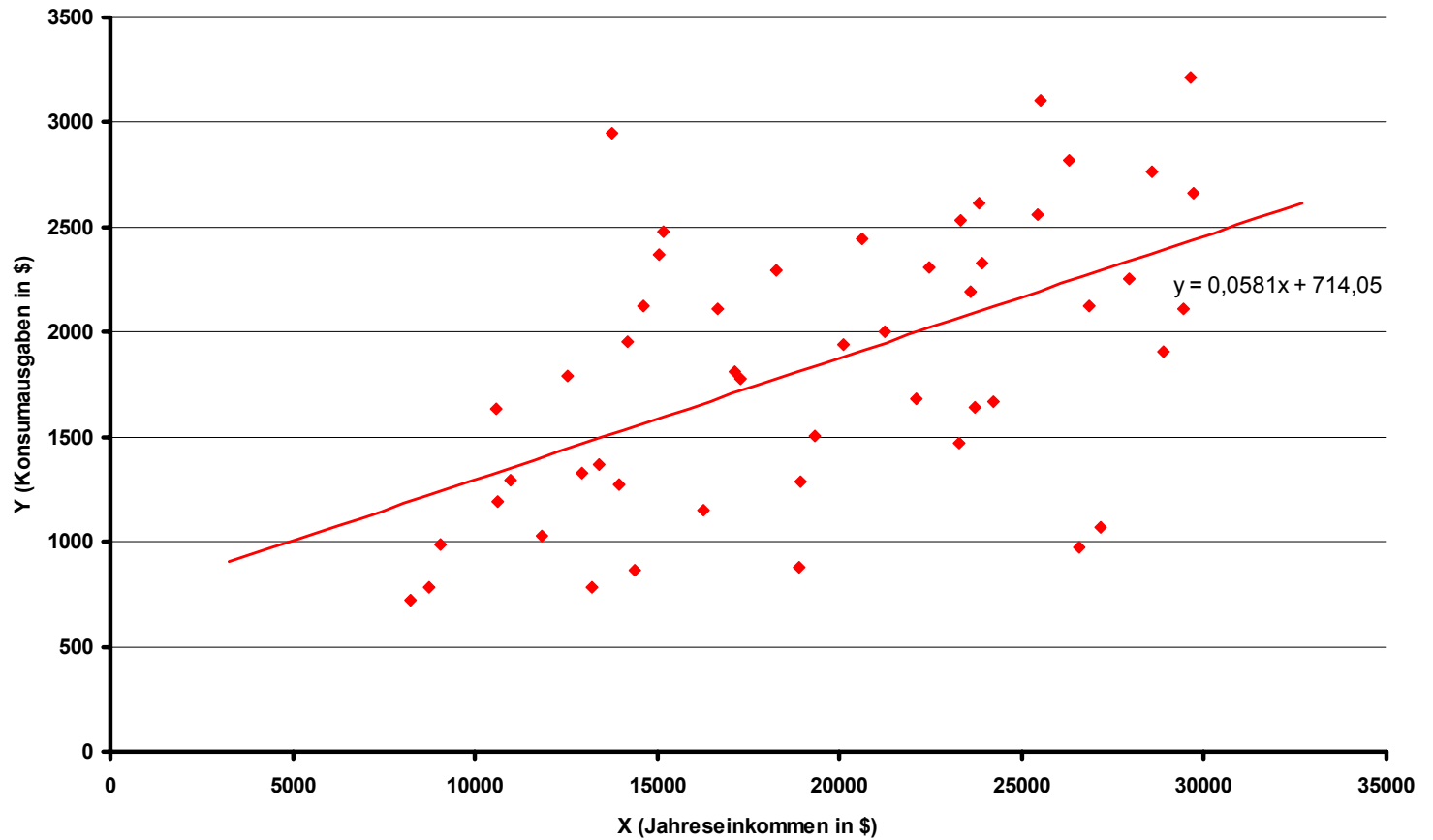
Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

Dozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04

# Streudiagramm II



## Deutung der Regressionsgleichung

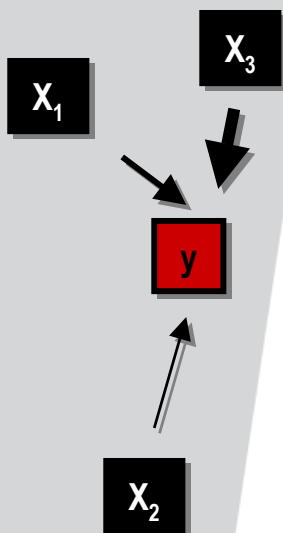
Konsumausgaben  
bei 0\$ Einkommen

Zunahme der  
Konsumausgaben für 1\$  
Einkommenszuwachs

$$Y = 714,05 + 0,0581 * X$$

Konsumausgaben

Einkommen

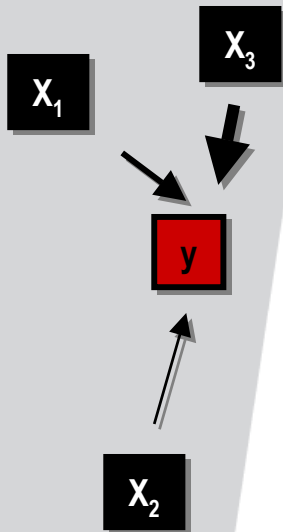


Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

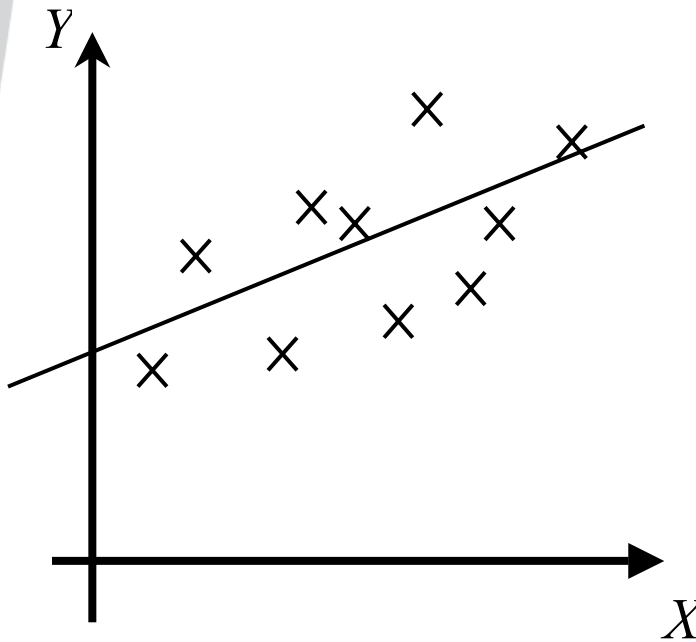
Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

Dozenten:  
Christian Brzinsky  
Christoph Hilbert

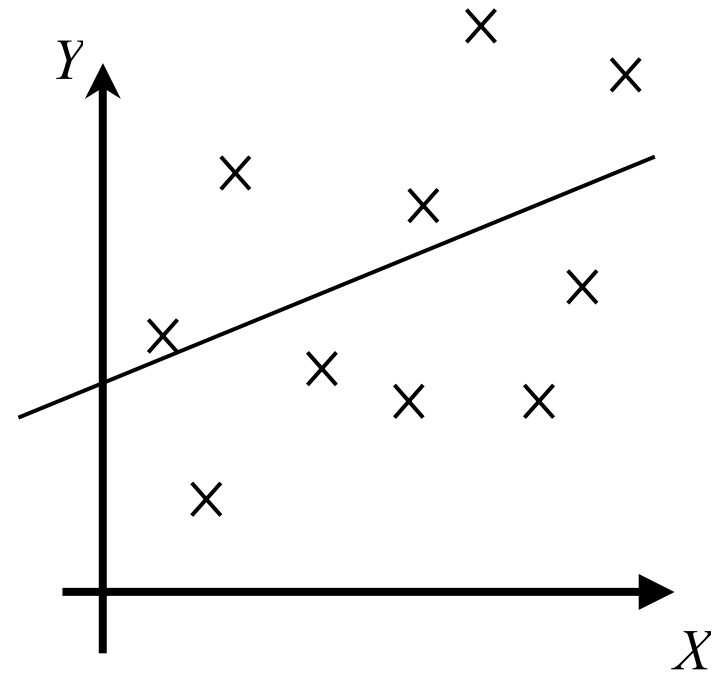
Wintersemester 03/04



# Qualität einer Schätzung

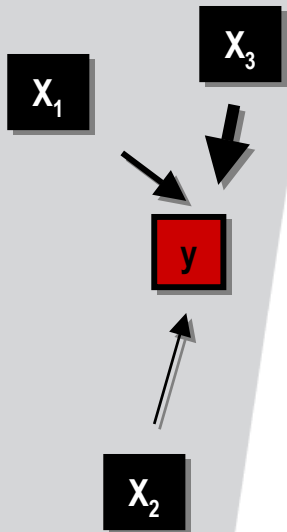
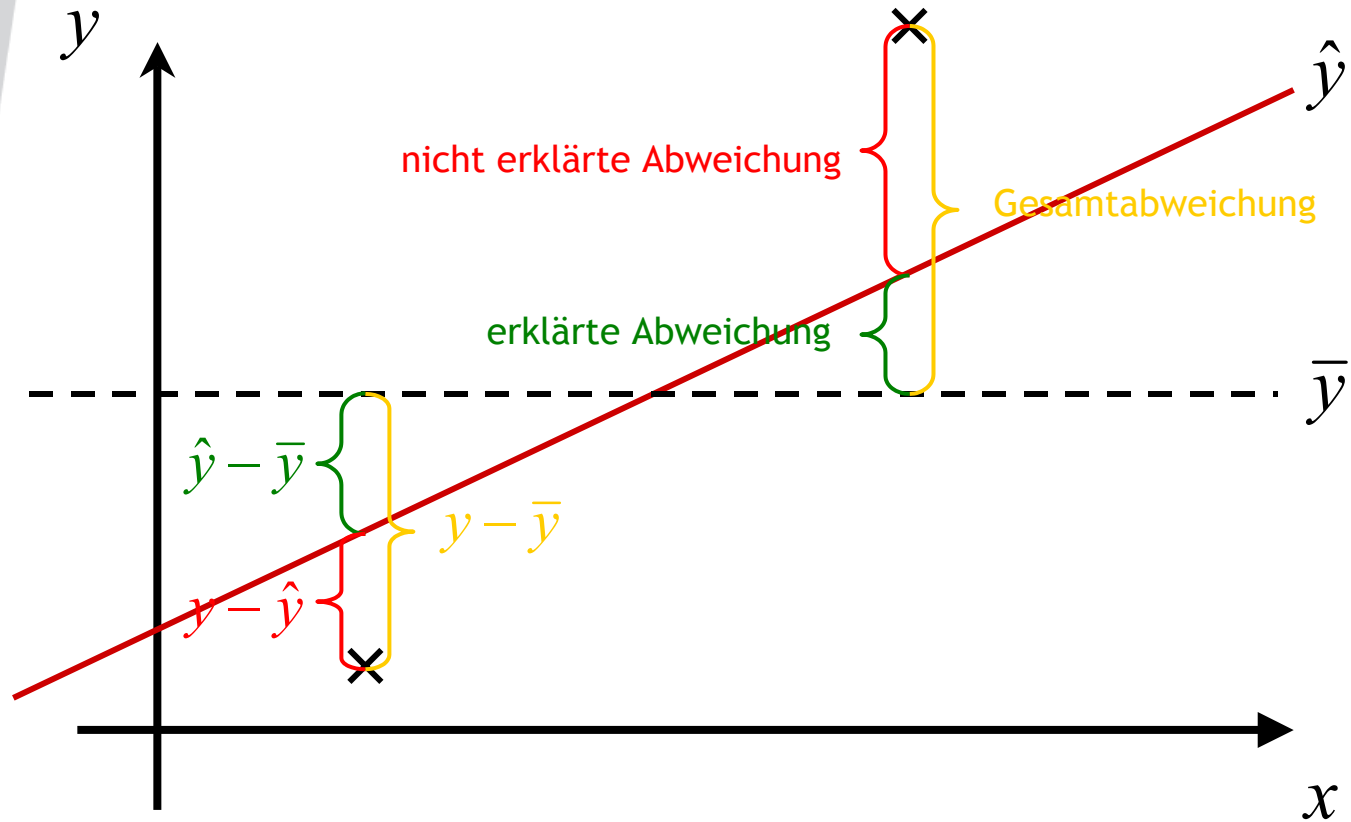


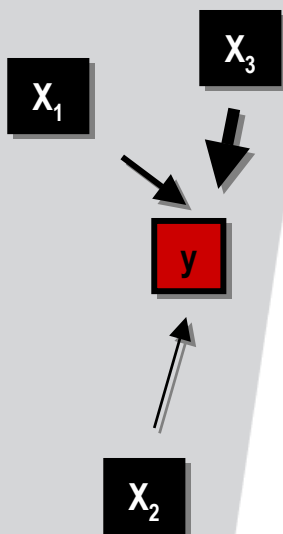
„bessere“ Schätzung



„schlechtere“ Schätzung

# Bestimmtheitsmaß I





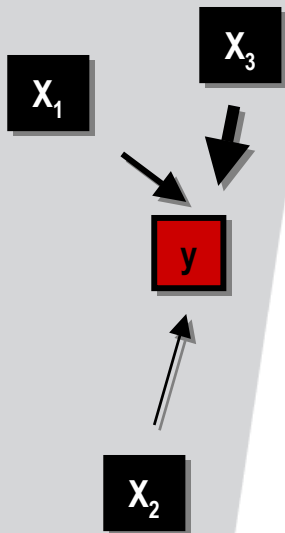
## Bestimmtheitsmaß II

Bestimmtheitsmaß  $r^2$  ist das Verhältnis von erklärter zu Gesamtstreuung

$$r^2 = \frac{\sum_{i=1}^I (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^I (y_i - \bar{y})^2}$$

$r^2 = 1 \Rightarrow$  gesamte Streuung erklärt

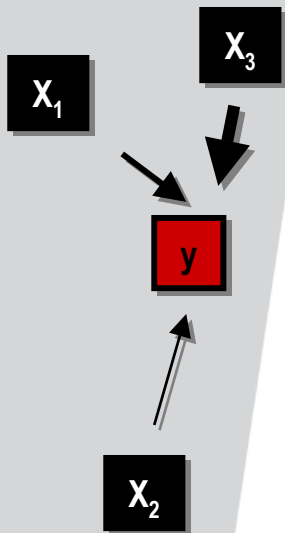
$r^2 = 0 \Rightarrow$  nichts erklärt



## Korrigiertes Bestimmtheitsmaß

- > bei hoher Regressorenzahl und wenigen Beobachtungswerten verschlechtern sich Schätzeigenschaften des Modells
- >  $r^2$  bleibt bei steigender Regressorenzahl mindestens gleich, korrigiertes  $r^2$  kann auch kleiner werden
- > Berechnung:

$$r_{korr}^2 = r^2 - \frac{J \times (1 - r^2)}{K - J - 1}$$



# Multiple Regression

einfache Regressionsgleichung:

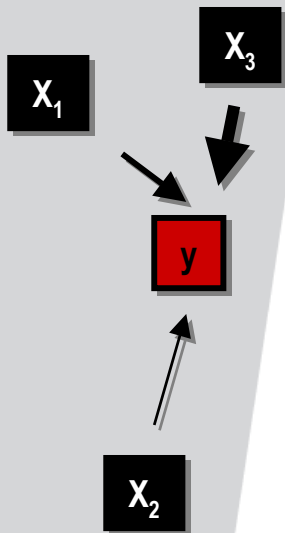
$$\hat{y}_i = a + bx_i + e_i$$

$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^N [y_i - (a + bx_i)]^2 = \min!$$

multiple Regressionsgleichung:

$$\hat{y}_i = a + b_1x_{1i} + b_2x_{2i} + \dots + b_Jx_{Ji} + e_i$$

$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^N [y_i - (a + b_1x_{1i} + b_2x_{2i} + \dots + b_Jx_{Ji})]^2 = \min!$$



## Interpretation der multiplen Regressionskoeffizienten

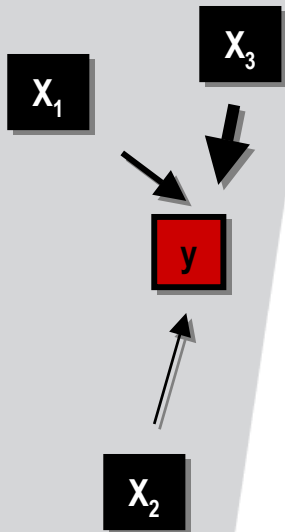
- > alle relevant erscheinenden Variablen müssen in das Modell miteinbezogen werden
- > Koeffizienten sind *ceteris paribus* zu interpretieren
- > multiple Regressionskoeffizienten sind „näher an der Realität“
- > Vergleich des Einflusses von unabhängigen Variablen nur mit standardisierten Koeffizienten möglich

Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

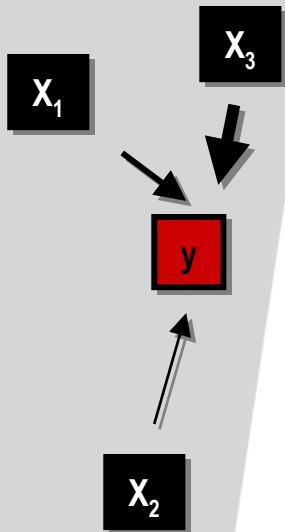
Dozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04



## Probleme der multiplen Regression

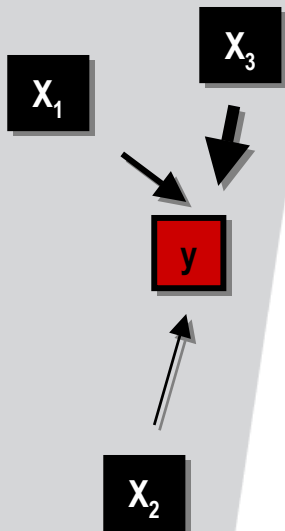
- > hoher Rechenaufwand
- > grafische Darstellung nur bei zwei unabhängigen Variablen möglich
- > Linearitätsprämisse muss für alle Variablen gelten
- > numerische Werte sind nicht direkt vergleichbar



## Standardisierte Koeffizienten

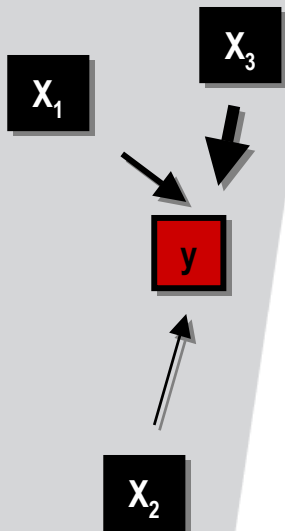
- > Vergleichbarkeit erfordert Eliminierung der unterschiedlichen Dimensionen
- > alternative Bezeichnung: beta-Koeffizient
- > Berechnung:

$$b^* = b \times \frac{s_x}{s_y}$$



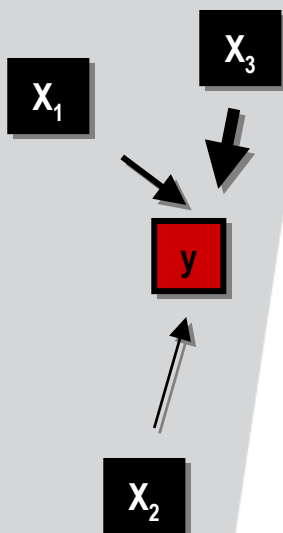
## Hypothesentest

- > F-Test: Prüfung des Determinations-koeffizienten  $r^2$
- > t-Test: Hypothesentest prüft Aussagen über eine unbekannte Grundgesamtheit anhand einer Stichprobe
  - > Ergebnis: Irrtumswahrscheinlichkeit von 5% (=„signifikant“) bzw. 1% („sehr signifikant“)
- > Konfidenzintervall: Maß für die Mögliche Abweichung des Regressionskoeffizienten der Stichprobe von der Grundgesamtheit



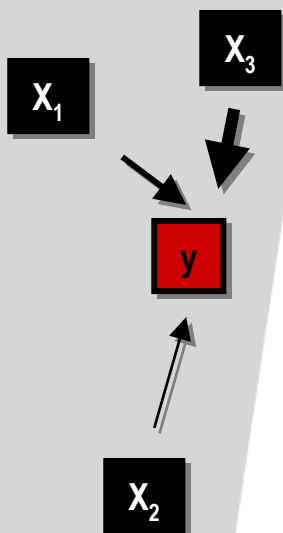
## Vorgehensweise (F-Test, t-Test)

1. Formulierung einer Alternativhypothese  $H_A$
2. Formulierung einer Nullhypothese  $H_0$
3. Festlegung der Irrtumswahrscheinlichkeit  $\alpha$  sowie des Ablehnungsbereichs der Nullhypothese
4. Berechnung der Prüfgröße
5. Entscheidung über Ablehnung der Nullhypothese



## Alternativ- und Nullhypothese

		Grundgesamtheit	
		$H_0$	$H_A$
Stichprobe	$H_0$	<i>richtig</i>	$\beta$ -Fehler
	$H_A$	$\alpha$ -Fehler	<i>richtig</i>



## Prüfgrößen

- > Berechnung des empirischen F-Wertes:

$$F_{emp} = \frac{\frac{r^2}{J}}{\frac{1-r^2}{K-J-1}}$$

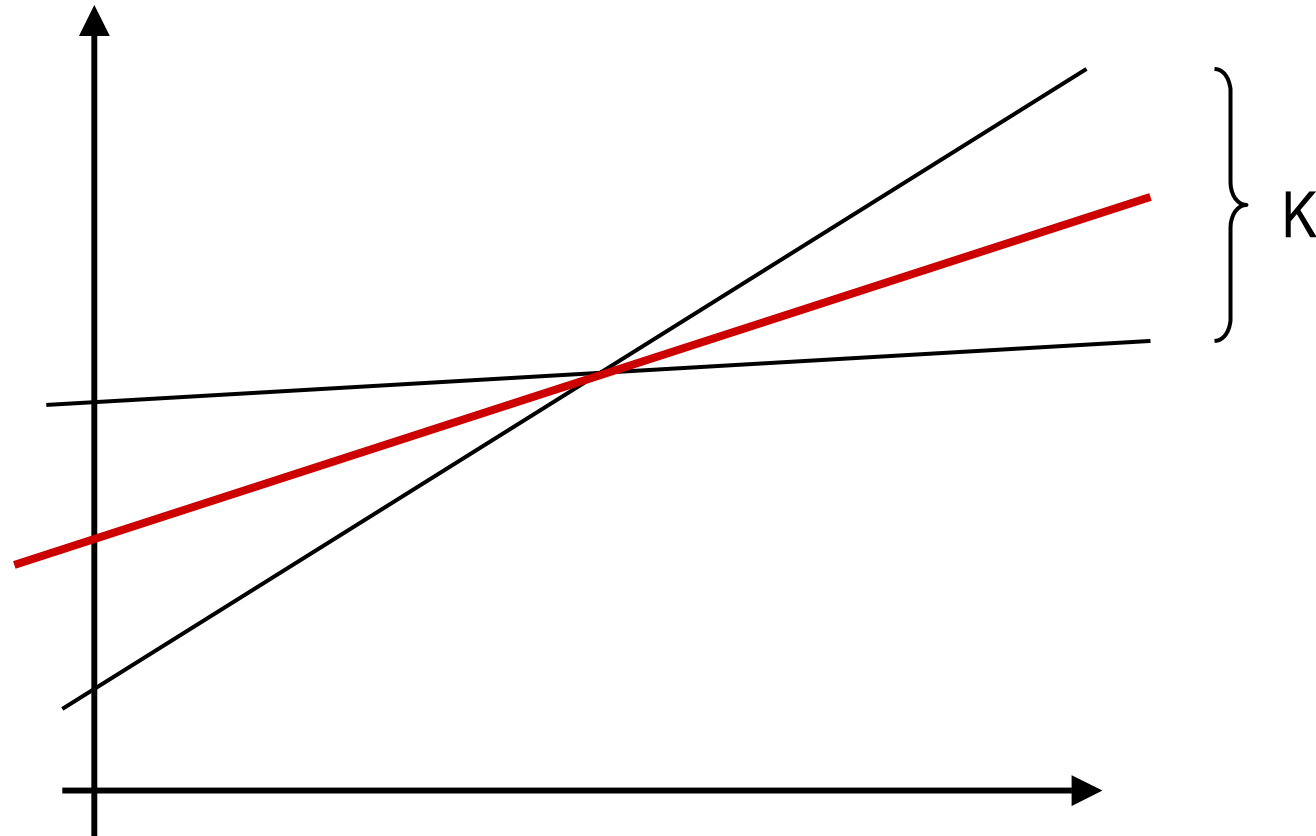
- > Berechnung des empirischen t-Wertes:

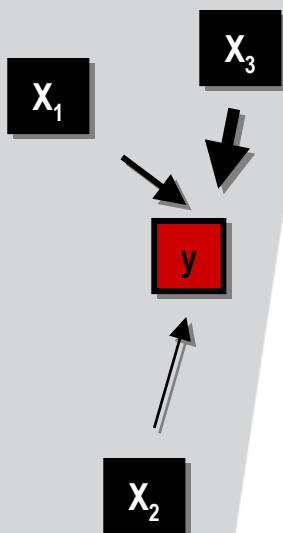
$$t_{emp} = \frac{b_J - \beta_J}{s_{b_J}}$$

Abteilung  
Arbeitsmarktpolitik  
und BeschäftigungHauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATADozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04

# Konfidenzintervall I





## Konfidenzintervall II

Berechnung:

$$b_J - t * s_{b_J} \leq \beta \leq b_J + t * s_{b_J}$$

- Interpretation: Je größer das Konfidenzintervall, desto unsicherer die Schätzung der Steigung in der Grundgesamtheit (besonders bei Vorzeichenwechsel)

Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

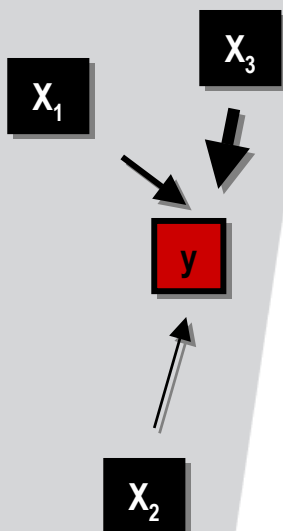
Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

Dozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04

# Prämissen II

Betrachtete Variable(n)	Prämissen	Verletzung	Folgen
Residuen	Normalverteilung	nicht normalverteilt	F-Test und t-Test ungültig
abhängige + unabhängige	Linearität	Nichtlinearität	Verzerrung der Schätzwerte
unabhängige	nicht linear abhängig	Multikollinearität	Ineffizienz
Residuen	nicht korreliert	Autokorrelation	Ineffizienz
unabhängige + Residuenstreuung	nicht korreliert	Heteroskedastizität	Ineffizienz

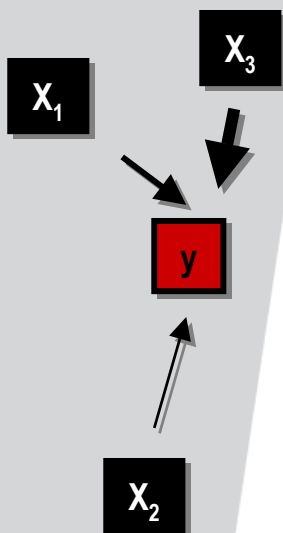


Abteilung  
Arbeitsmarktpolitik  
und Beschäftigung

Hauptseminar:  
Analyse von Längsschnittdaten  
mit GSOEP und STATA

Dozenten:  
Christian Brzinsky  
Christoph Hilbert

Wintersemester 03/04

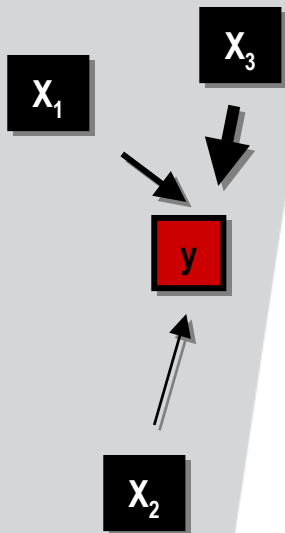


# Maßzahlen-Übersicht

Abkürzung	Name	Interpretation
a	Regressionskonstante	Betrag von Y, wenn X = 0
$b_1, b_2, \dots b_j$	Regressionskoeffizienten	Veränderung von Y bei Erhöhung von $X_j$ um 1
r	Korrelationskoeffizient	$-1 \leq r \leq 1$ , vergleichbares Maß für das Miteinander-Variieren von X und Y
$r^2$	Bestimmtheitsmaß, Determinationskoeffizient	Anteil der durch die Regression erklärten Variation von Y an der Gesamtvariation
beta	standardisierte Regressionskoeffizienten	miteinander vergleichbare Regressionskoeffizienten
$r^2_{\text{korr}}$	korrigiertes Bestimmtheitsmaß	um den Einfluss von geringer Stichprobengröße und großer Regressorenzahl korrigiertes $r^2$
F	(empirischer) F-Wert	Test für Determinationskoeffizient
T	(empirischer) t-Wert	Test für Regressionskoeffizienten

# Korrelation $\leftrightarrow$ Regression

- > Korrelation + Regression ermitteln linearen Zusammenhang von  $X$  und  $Y$  anhand ihrer Streuungen (keine Kausalität)
- > Regression setzt Unterscheidung von unabhängiger + abhängiger Variable voraus
- > Regression gibt zusätzliche Informationen für Prognose



# STATA-Befehle

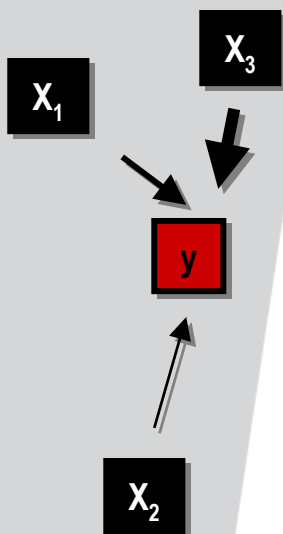
`regress [dependent variable] [varlist] if [exp] in [range]`

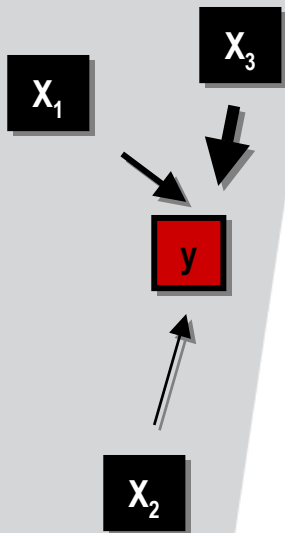
`predict [type] [new var] if [exp] in [range]`

(anova)

(xtintreg)

(xtreg)





## Literaturhinweise

- > Backhaus u.a. (2000): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung (Kapitel 1); Springer-Verlag, Berlin, Heidelberg, New York; Seite 1-69.
- > Gujarati (2001): Basic Econometrics; McGraw-Hill, New York u.a.
- > Kleinbaum u.a. (1998): Applied Regression Analysis And Other Multivariable Methods; Duxbury Press, Albany u.a.
- > Chatterjee (1991): Regression Analysis By Example; John Wiley & Sons, New York u.a.