

Regressionsdiagnostik

Ben Jann

Kolloquium „Soziologie“
ETH Zürich, 30. Mai 2006

Aufbau

1 Ausgangslage

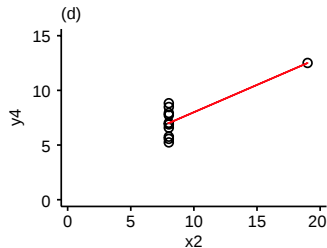
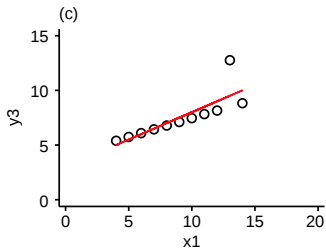
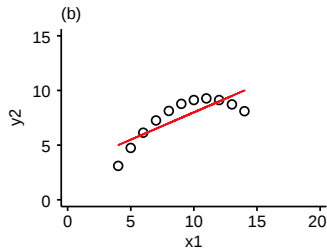
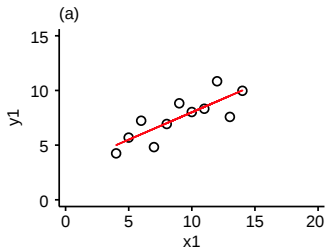
- Viele Wege führen nach Rom
- Das lineare Regressionsmodell und seine Annahmen

2 Einfache grafische Mittel

3 Identifikation einflussreicher Datenpunkte

- Instrumente
- Anwendungsbeispiel
- Massnahmen

Anscombes Quartett



$$\hat{\beta}_0 = 3.0, \hat{\beta}_1 = 0.5, se(\hat{\beta}_1) = 0.118, R^2 = 0.67, \bar{X} = 9.0, \bar{Y} = 7.5,$$

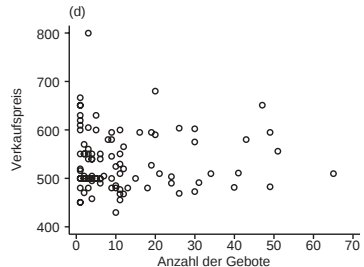
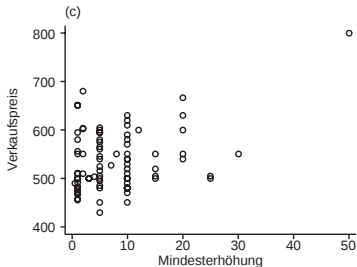
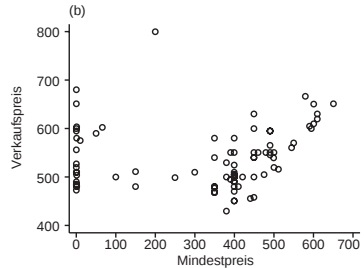
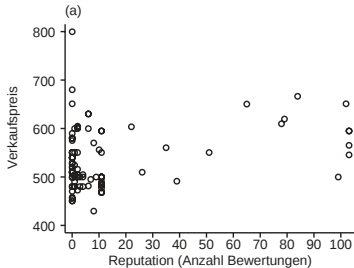
Lineare Regression

$$Y = X\beta + \epsilon, \quad E(\epsilon|X) = 0$$

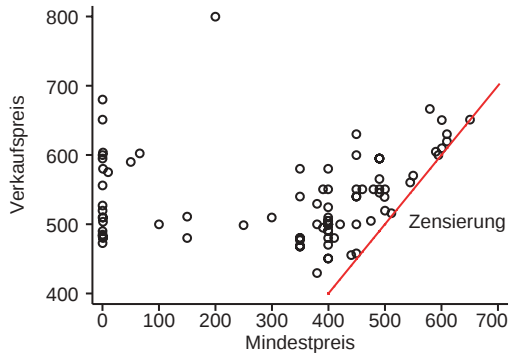
Annahmen und Anforderungen an die Daten:

- Korrektes Modell / korrekte Spezifikation:
 - linearer Zusammenhang / korrekte funktionale Form
 - keine „vernachlässigten“ Variablen
 - keine Messfehler
- Homoskedastizität und keine Autokorrelation: $E(\epsilon\epsilon') = \sigma^2 I$
- Normalverteilte Fehler $\epsilon \sim N(0, \sigma^2)$
- Genügend Variabilität / keine Multikollinearität
- Mehr oder weniger regelmässig verteilte Daten; keine Ausreisser / einflussreiche Punkte

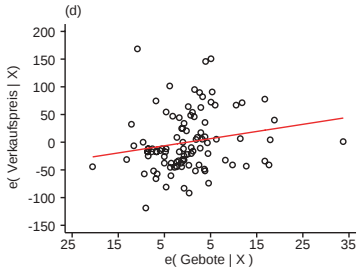
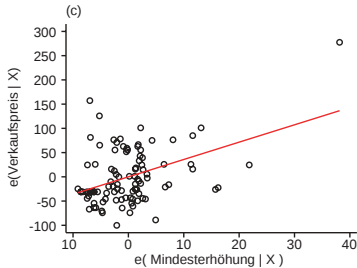
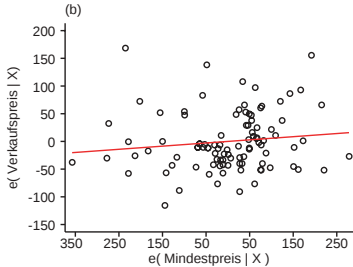
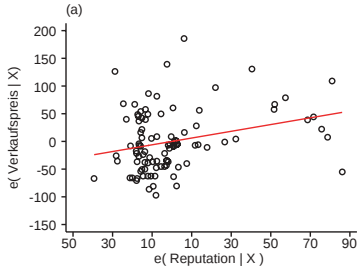
Einfaches Streudiagramm



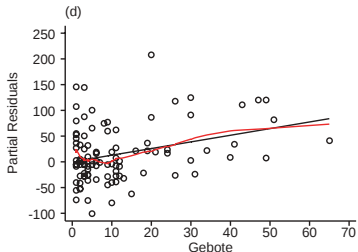
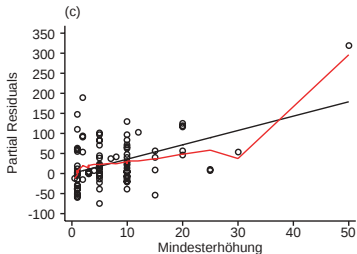
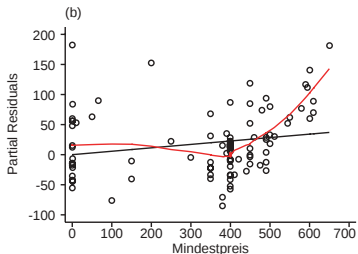
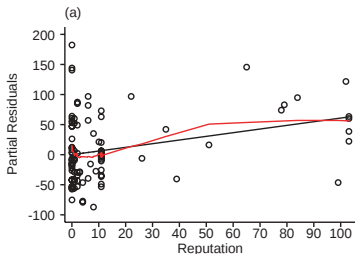
Einfaches Streudiagramm



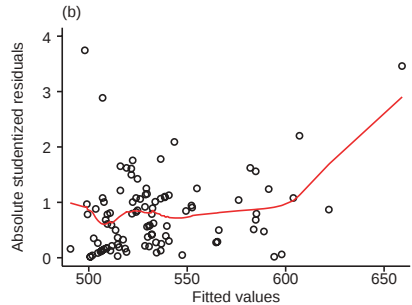
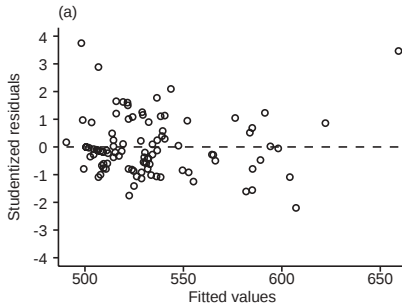
Partielle Regressionsplots (added variable plots)



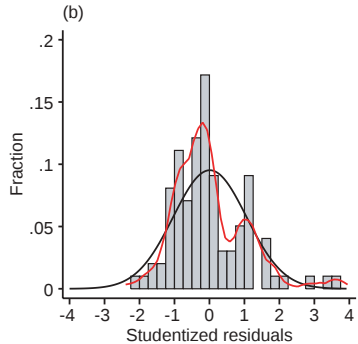
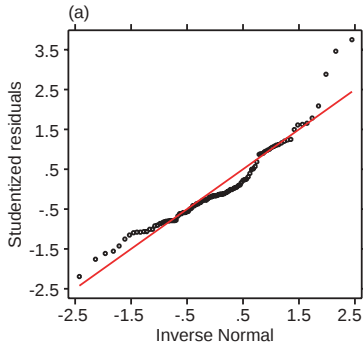
Partielle Residuenplots (component-plus-residual plot)



Tukey-Anscombe Plot (Residuen vs. Vorhersagewerte)



Normal-Quantil-Plot und Histogramm der Residuen



Das Problem einflussreicher Datenpunkte

“We are usually happier about asserting a regression relation if the relation is still apparent after a few observations (any ones) have been deleted – that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two.” (Anscombe 1973)

“An influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of the various estimates (coefficients, standard errors, t-values, etc.) than is the case for most of the other observations.” (Belsley et al. 1980)

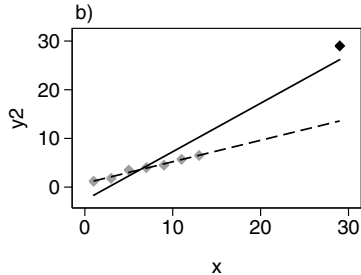
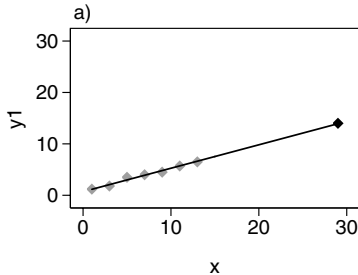
Grundlegende Konzepte

■ **Leverage** (Hebelwirkung)

$$\hat{\beta} = (X'X)^{-1}X'Y \quad \Rightarrow \quad \hat{Y} = X\hat{\beta} = HY$$

Ein Diagonalelement h_i von H (*hat matrix*) entspricht dem (potentiellen) Einfluss von Punkt i auf die Regressionsschätzung. Diese „Hebelwirkung“ ist um so grösser, je weiter Punkt i vom „Zentrum“ von X entfernt ist.

Grundlegende Konzepte



Grundlegende Konzepte

■ Studentisierte Residuen

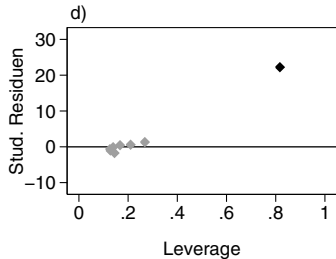
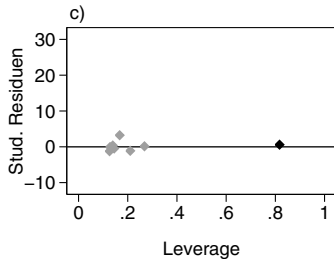
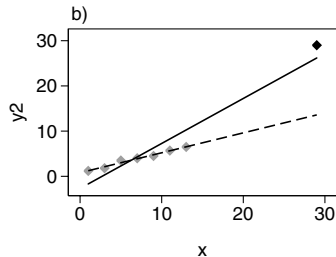
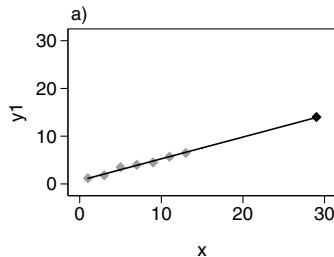
Die Residuen

$$\hat{\epsilon} = Y - \hat{Y}$$

messen die Abweichung eines Datenpunktes von der Regressionsgerade. Da hohe Leverage aufgrund der Hebelwirkung die Residuen verkleinert, ist es sinnvoll, standardisierte oder studentisierte Residuen zu betrachten:

$$\hat{\epsilon}'_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}, \quad \hat{\epsilon}^*_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1 - h_i}}$$

Grundlegende Konzepte



Masszahlen

Die Masszahlen beziehen sich auf die Veränderung der Regressionsschätzung, wenn Punkt i aus den Daten gelöscht wird.

- Einfluss auf die Koeffizienten: Cook's D

$$D_i = \frac{\hat{\epsilon}_i'^2}{k} \cdot \frac{h_i}{1 - h_i}$$

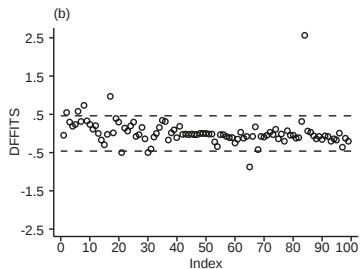
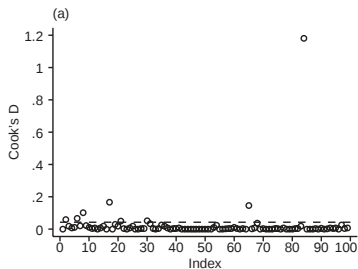
- Einfluss auf die Vorhersagewerte: $DFFITS$

$$DFFITS_i = \hat{\epsilon}_i^* \sqrt{\frac{h_i}{1 - h_i}}$$

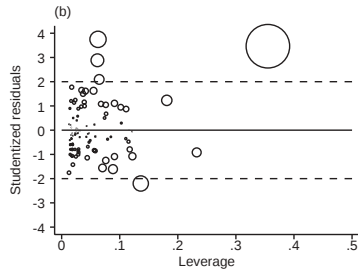
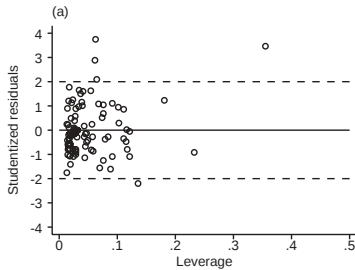
- Einfluss auf einzelne Koeffizienten: $DFBETAS$

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{SE_{(-i)}(\hat{\beta}_j)}$$

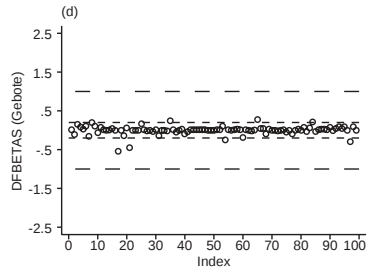
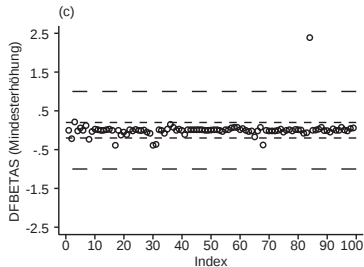
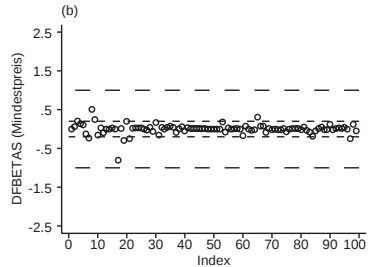
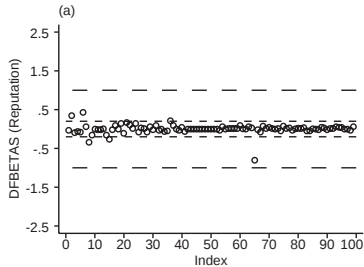
Masszahlen



Masszahlen



Masszahlen



Wu and Xie (2003)

Table 4. OLS Coefficients from the Multiple Linear Regression of Monthly Earnings on Selected Independent Variables, Urban China, 1996: Three-Worker-Type Analysis

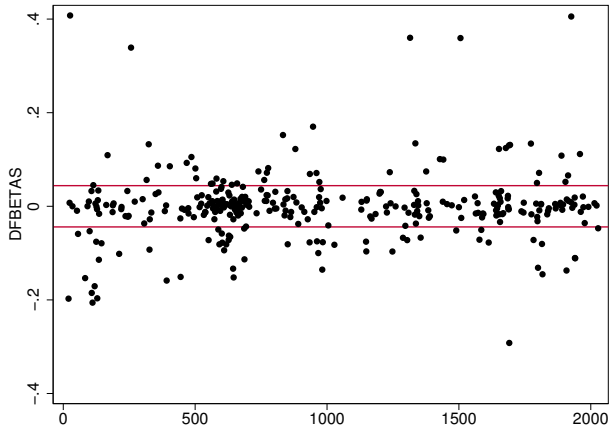
Variable	Restrictive Measure		Broad Measure		Comprehensive Measure	
	Model 4a	Model 5a	Model 4b	Model 5b	Model 4c	Model 5c
Education (years of schooling)	.049*** (.009)	.045*** (.010)	.053*** (.008)	.045*** (.009)	.057*** (.007)	.047*** (.006)
Experience	.010** (.005)	.009* (.005)	.010* (.004)	.008 (.005)	.014*** (.004)	.013*** (.004)
(Experience) ² × 1,000	-.153* (.074)	-.144* (.070)	-.160* (.074)	-.099 (.070)	-.203** (.070)	-.0167** (.060)
Party member (yes = 1)	.121** (.038)	.126** (.037)	.138*** (.035)	.145*** (.035)	.142*** (.037)	.149*** (.037)
Sex (male = 1)	.218*** (.040)	.213*** (.040)	.220*** (.038)	.210*** (.039)	.225*** (.038)	.216*** (.038)
Later entrants ^a	.312* (.144)	-.732 (.370)	.238*** (.068)	-.263 (.193)	.313*** (.071)	-.175 (.182)
Early birds	.553 (.439)	.130 (.602)	.184 (.230)	-.124 (.266)	.151 (.206)	-.067 (.249)
Later entrants × Education	—	.122* (.047)	—	.060* (.022)	—	.056** (.019)
Early birds × Education	—	.051 (.103)	—	.037 (.031)	—	.025 (.024)
Constant	5.305*** (.156)	5.348*** (.165)	5.238*** (.140)	5.333*** (.153)	5.124*** (.106)	5.230*** (.116)
Number of cases	2,072		2,061		2,060	
R ²	.117	.129	.114	.123	.127	.136

Notes: Numbers in parentheses are standard errors adjusted for clustering on counties. Data are weighted.

^a "Stayers" is the reference category; market losers are omitted from the analysis because of the small number of cases (N = 19).

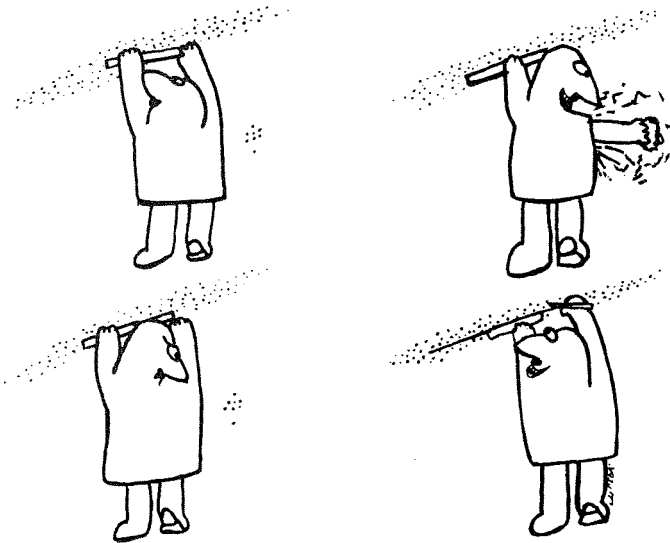
p* < .05 *p* < .01 ****p* < .001 (two-tailed tests)

Wu and Xie (2003)



Wu and Xie (2003)

	Modell 1	Modell 2
educ	0.048*** (0.007)	0.048*** (0.007)
...	...	...
educlater	0.061** (0.020)	0.028 (0.016)
educearly	0.026 (0.027)	0.025 (0.027)
_cons	5.185*** (0.111)	5.181*** (0.111)
N	2057	2051
r2	0.142	0.127



(Fox 1991:77)

Massnahmen

- 1 Identifizierte Ausreisser einer detaillierten Betrachtung unterziehen.
- 2 Ausreisser korrigieren bzw. ausschliessen, wenn es sich um offensichtliche Messfehler handelt.
- 3 Einfluss von Ausreissern begrenzen (Vergleich von Modellen mit und ohne Ausreisser; robuste Regression).
- 4 Modell / Theorie verbessern, denn ...

Massnahmen

An apparently wild (or otherwise anomalous) observation is a signal that says: "Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study." (Kruskal 1960)