

6.6 Residuen in Multiple Regression

$$\epsilon_i \sim N(0, \sigma^2) \quad u.i.v.$$

$$\begin{aligned} e &= y - X\hat{b} = y - \hat{y} \\ &= y - X(X'X)^{-1}X'y \\ &= (I - H)y \end{aligned}$$

$$V(e) = (I - H)\sigma^2$$

$$V(e_i) = (1 - h_i)\sigma^2$$

d.h. die Varianz des Residuums einer Beobachtung mit größerer Hebelwirkung wird klein sein. Um die Verteilung der Residuen zu standardisieren, betrachtet man studentisierte Residuen:

$$r_i = \frac{e_i}{s\sqrt{1 - h_i}} \sim t_{n-p-1}$$

”Student” weil man s statt σ benutzt.

Es gibt zwei Varianten, interne und externe. Bei den internen werden alle Daten benutzt. Bei den externen schätzt man σ^2 mit $s^2(i)$, ” s^2 nicht i ”, d.h. ohne Beobachtung i .

6.7 Auswahl der erklärenden Variablen

Gegeben sind Y und X_1, \dots, X_p .

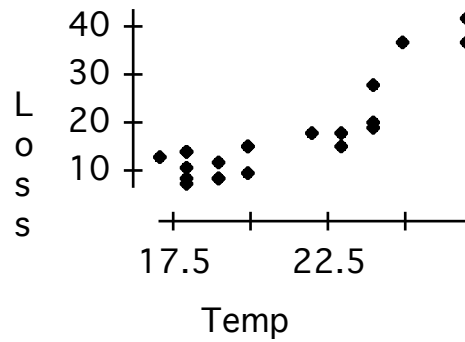
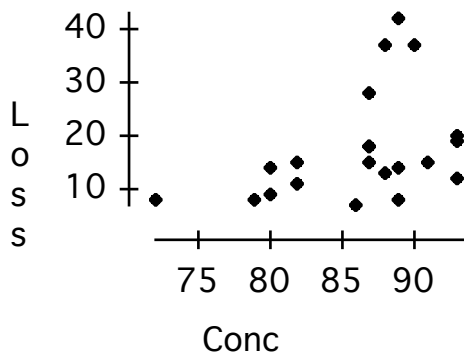
Man könnte Modelle berechnen:

1. vorwärts schrittweise
2. rückwärts schrittweise
3. alle (!) mögliche Kombinationen (2^p)

Wenn man vorwärts rechnet, versucht man immer die nächstbeste erklärende Variable auszusuchen. Wenn man rückwärts rechnet, wirft man die Variable weg, die am wenigsten zum Modell beiträgt. Es gibt verschiedene Algorithmen dafür, aber alle sind theoretisch unzufriedenstellend, weil

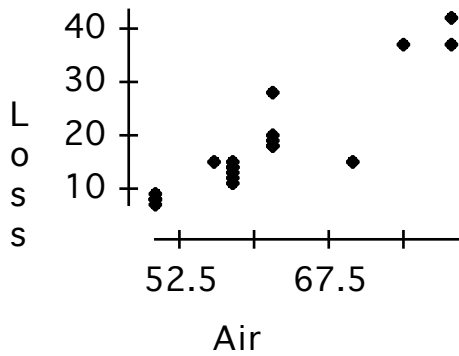
- (a) Tests werden mit denselben Daten durchgeführt
- (b) das "beste" Modell wird nicht unbedingt gefunden.

Stackloss Datensatz (1)



Pearson Product-Moment Correlation

No Selector



	Loss	Temp	Air	Conc
Loss	1.000			
Temp	0.876	1.000		
Air	0.920	0.782	1.000	
Conc	0.400	0.391	0.500	1.000

Dependent variable is: Loss

No Selector

R squared = 91.4% R squared (adjusted) = 89.8%

s = 3.243 with 21 - 4 = 17 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1890.41	3	630.136	59.9
Residual	178.830	17	10.5194	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-39.9197	11.90	-3.36	0.0038
Temp	1.29529	0.3680	3.52	0.0026
Air	0.715640	0.1349	5.31	≤ 0.0001
Conc	-0.152123	0.1563	-0.973	0.3440

Stackloss (2)

Dependent variable is: Loss

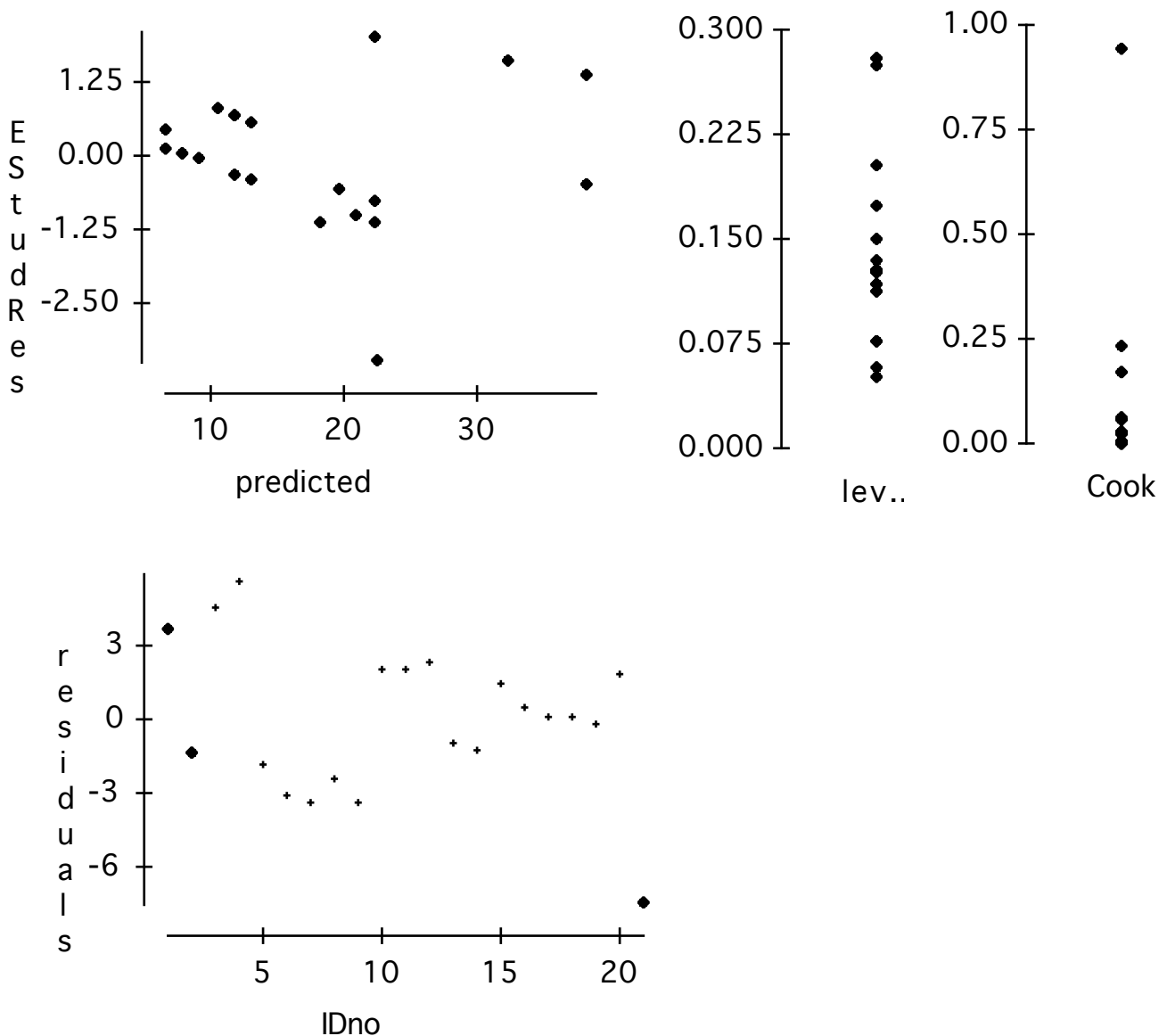
No Selector

R squared = 90.9% R squared (adjusted) = 89.9%

s = 3.239 with 21 - 3 = 18 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1880.44	2	940.221	89.6
Residual	188.795	18	10.4886	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-50.3588	5.138	-9.80	≤ 0.0001
Temp	1.29535	0.3675	3.52	0.0024
Air	0.671154	0.1267	5.30	≤ 0.0001



Die Punkte mit den größten Hebelwirkungen sind selektiert

Stackloss (3)

Dependent variable is: Loss

No Selector

R squared = 84.6% R squared (adjusted) = 83.8%

s = 4.098 with 21 - 2 = 19 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1750.12	1	1750.12	104
Residual	319.116	19	16.7956	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-44.1320	6.106	-7.23	≤ 0.0001
Air	1.02031	0.1000	10.2	≤ 0.0001

Ohne den Punkt mit dem größten Cooks D Wert

Dependent variable is: Loss

cases selected according to MinusCookMax

21 total cases of which 1 are missing

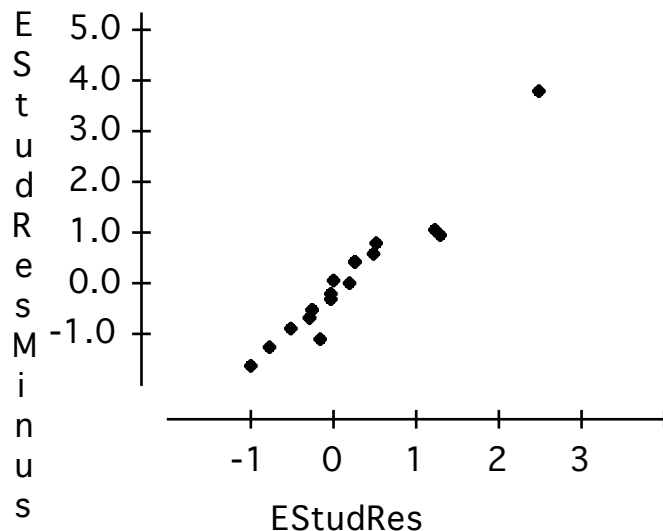
R squared = 92.7% R squared (adjusted) = 92.3%

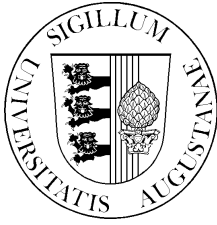
s = 2.895 with 20 - 2 = 18 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1911.65	1	1911.65	228
Residual	150.905	18	8.38359	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-48.1893	4.408	-10.9	≤ 0.0001
Air	1.09824	0.0727	15.1	≤ 0.0001

Vergleich der Residuen





Vergleich von Regressionsmodellen für den Stackloss Datensatz

Modell	df	R²	R² (adj)
T, A, C	17	91.4	89.8
T,A	18	90.9	89.9
A	19	84.6	83.8
A, minus	18	92.7	92.3

aber ist irgendein der Modelle gut?

Table 3

Climate and phenology influences on sugar, and acid for Cabernet Sauvignon and Merlot: equations and statistical tests^a

Independent variables	Dependent variables (g/l)			
	CABSUG	CABACID	MERSUG	MERACID
DPET	–	–	19.64 log(x) (3.11)	1.265.2(1/x) (6.97)
BPET	–	1014.0(1/x) (3.78)	–	–
BPREC	1260.4(1/x) (2.61)	0.81 log(x) (6.84)	–	–
FPET	–	–	–	5115.9(1/x) (8.13)
FPREC	–0.06x (4.32)	0.007x (8.99)	–0.09x (5.76)	0.48 log(x) (4.35)
FTEMP25	–926.2(1/x) (6.07)	–1.08 log(x) (4.91)	–442.53(1/x) (2.03)	–
VPET	–	–	–	–2.01 log(x) (5.98)
VPREC	–11.0 log(x) (9.84)	0.48 log(x) (7.26)	–11.15 log(x) (6.48)	–
VTEMP25	–	–	0.52x (3.10)	–
VTEMP30	–	1.82 (1/x) (4.69)	–	2.07(1/x) (5.31)
CONSTANT	267.8 (34.22)	–	132.0 (2.49)	9.35 (3.87)
<i>Test statistics</i>				
S.E.	3.30	0.154	3.92	0.130
R ²	0.931	0.974	0.952	0.980

^a Model calculations: time 1980–1994, *T*-ratios in parentheses, and variables as described in Table 2.

6.8 Warum ein F-Test im Regressionsoutput?

Betrachten wir die Tabelle

	QS	Freiheitsgrade	MQS
Modell	QS_M	p	$\frac{QS_M}{p}$
Residuen	$\sum (y_i - \hat{y}_i)^2$	$n - p - 1$	$\frac{QS_R}{n - p - 1}$
Insgesamt	$\sum (y_i - \bar{y})^2$	$n - 1$	

$$QS_M = \sum (\hat{y}_i - \bar{y})^2$$

$$\frac{QS_R}{n - p - 1} = s^2 \quad \text{ist unser Schätzer für } \sigma^2$$

Falls $b_i = 0 \quad \forall i$ (H_0), dann wäre $\frac{QS_M}{p}$ auch ein Schätzer für σ^2 . Dann ist die Statistik

$$F = \frac{QS_M/p}{QS_R/(n - p - 1)}$$

eine Teststatistik für

$$H_0 : b_i = 0 \quad \forall i$$

mit einer $F_{(p, n-p-1)}$ -Verteilung. Dieser Test testet das *Gesamtmodell* und ist deshalb fast immer signifikant.

6.9 Beurteilung eines Modells

1. R^2 (oder ein anderes globales Kriterium)
2. Residuenplots
3. Diagnostiken
 - (a) Hebelwirkung
 - (b) Einfluß
4. Ist das Modell sinnvoll?

6.9.1 Modellwerte und Diagnostiken in der Regression

$$\hat{y} = X\hat{b}$$

$$\hat{b} = (X'X)^{-1}X'y$$

$$\begin{aligned}\hat{y} &= X(X'X)^{-1}X'y \\ &= Hy\end{aligned}$$

H heißt der Hutmatrix und $H = X(X'X)^{-1}X'$.

H ist

$n \times n$

symmetrisch ($h_{ij} = h_{ji}$)

idempotent ($H^2 = H$)

und $0 \leq h_{ii} \leq 1$

$$\sum h_{ii} = p + 1$$

im Durchschnitt $h_{ii} = \frac{p+1}{n}$

Hebelwirkung $h_i = h_{ii}$ ist die Hebelwirkung der i .ten Beobachtung und mißt wie weit Beobachtung i von den anderen im X -Raum liegt.

6.9.2 Einfluß — Cook's D

Der Einfluß einer Beobachtung i , dieselbe Formel aber 3 verschiedene Interpretationen:

1. Unterschiede zwischen vorausgesagten Werten

$$D_i = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{(p + 1)s^2}$$

2. gewichtete Unterschiede zwischen Parameterschätzern

$$D_i = \frac{(\hat{b}_{(i)} - \hat{b})'(X'X)(\hat{b}_{(i)} - \hat{b})}{(p + 1)s^2}$$

weil $\hat{y} = X\hat{b}$ und $\hat{y}_{(i)} = X\hat{b}_{(i)}$

3. Beobachtung i allein

$$D_i = \frac{r_i^2}{(p + 1)} \frac{h_i}{(1 - h_i)}$$

weil

$$\hat{b}_{(i)} - \hat{b} = -(X'X)^{-1}x_i'(1 - h_i)^{-1}e_i$$

Die Schwäche dieser Diagnostiken ist, dass sie sich immer nur auf eine Beobachtung beziehen.

Regression — Zusammenfassung

1. Theorie

- (a) gut für einfache Situationen
- (b) "grau" für komplizierte Modelle

2. Praxis

- (a) Auswahl Probleme
- (b) Interpretations Probleme
- (c) Sehr (sehr) oft angewandt