Lineare Regression

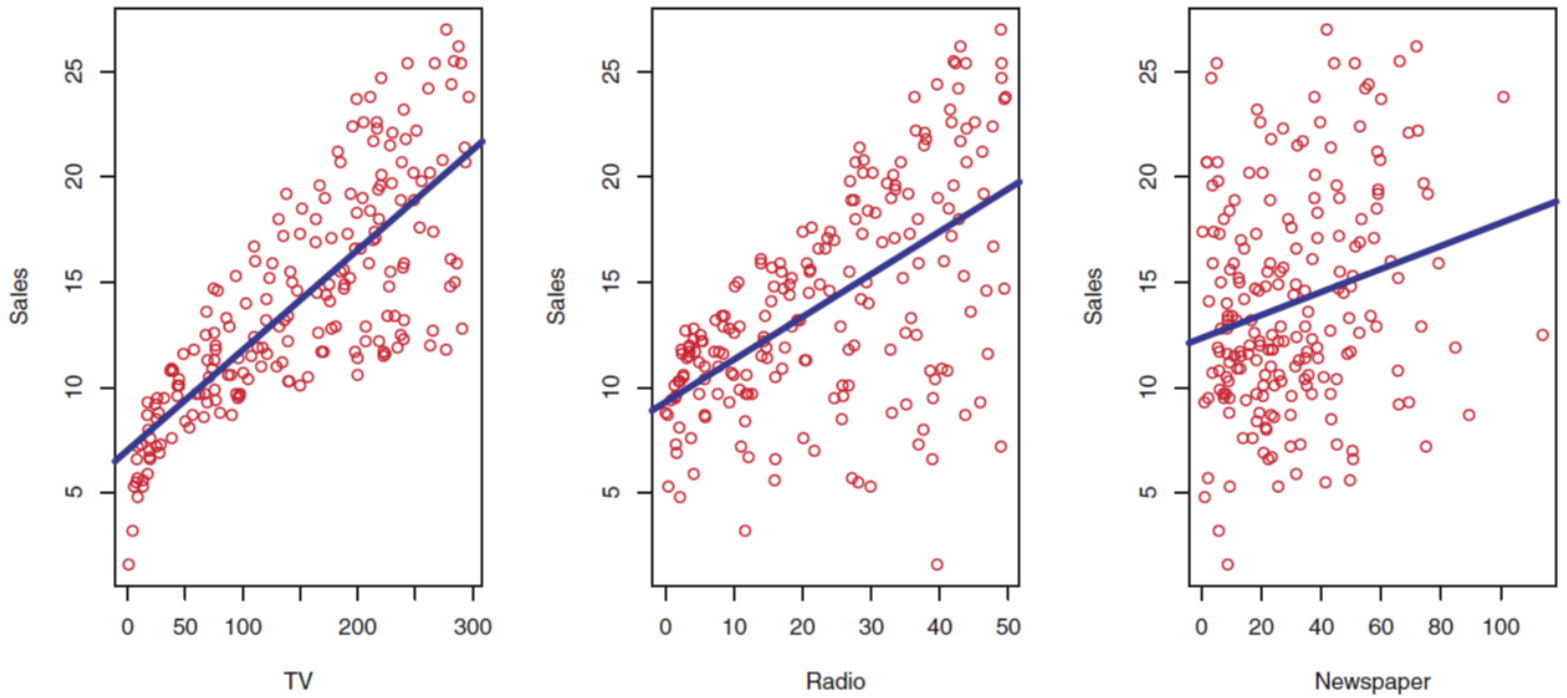


FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

Vgl.: James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert: An Introduction to Statistical Learning. With Applications in R., 8. verbesserte Auflage. New York 2013. S.19 (ILSR)

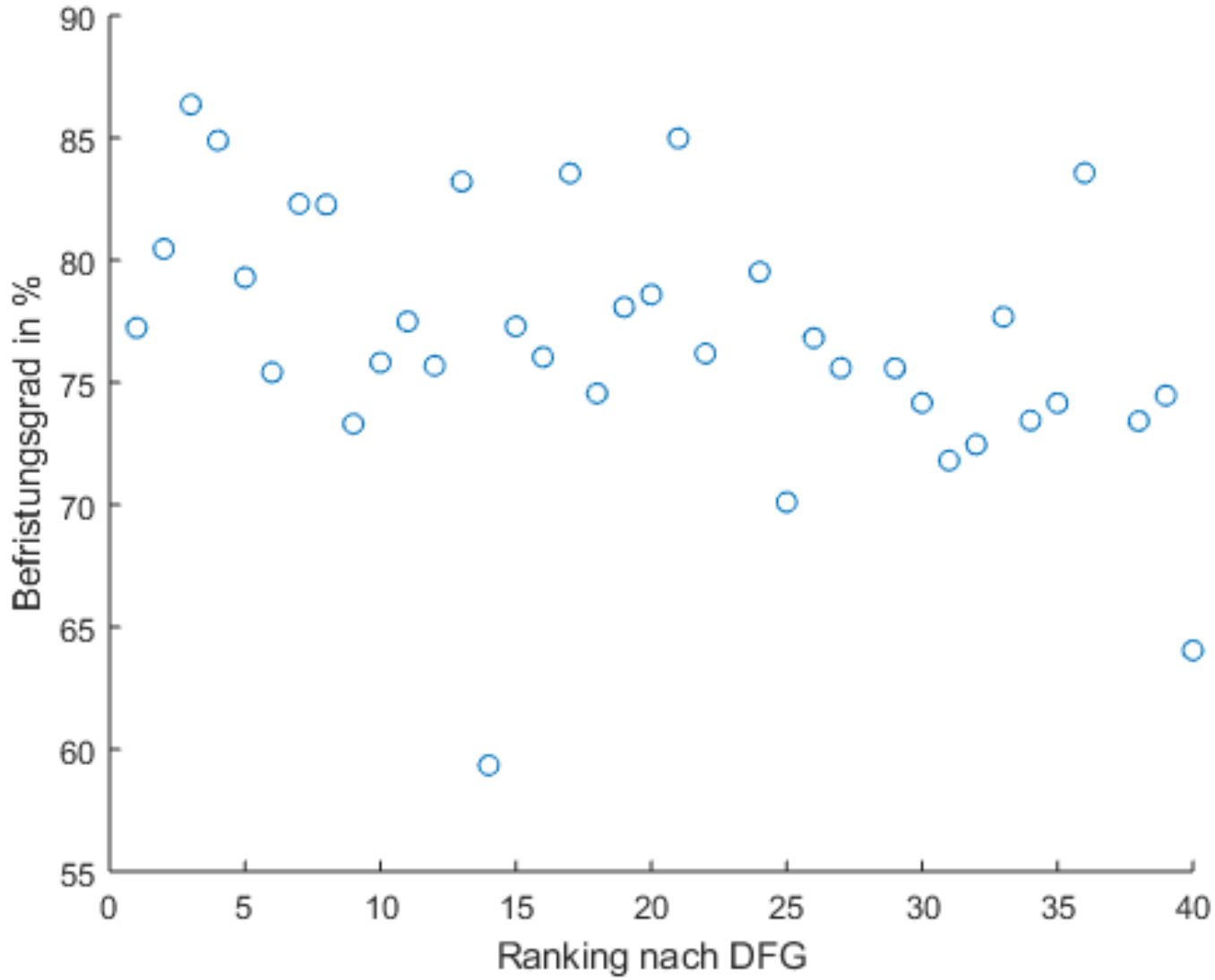
Motivation

	Befristungsanteil (in Prozent, absteigend)	DFG-Mittel 2011–2013 (in Mio.)	Platzierung nach DFG-Mitteln
RWTH Aachen	87,36	272,5	3
TU Berlin	84,98	121,5	21
TU München	84,89	259,9	4
TU Dortmund	83,56	74,9	36
TU Darmstadt	83,55	145,3	17
U Münster	83,21	174,8	13
U Freiburg	82,30	239,6	7
KIT	82,27	198,2	8
U Heidelberg	80,46	274,7	2
TU Braunschweig	79,78	Nicht aufgeführt	k. P.
U Hannover	79,52	115,5	24
Freie U Berlin	79,29	252,2	5
U Stuttgart	78,59	128,4	20
U Würzburg	78,08	141,3	19
U Bielefeld	77,68	84,5	33
U Erlangen-Nürnberg	77,49	186,7	11
U Frankfurt am Main	77,29	162,8	15
U München	77,23	277,8	1
U Kiel	76,81	106,7	26
U Paderborn	76,59	Nicht aufgeführt	k. P.
U Bochum	76,18	120,5	22
U Köln	76,03	158,8	16
TU Dresden	**75,81**	**191,6**	**10**
U Kassel	75,76	Nicht aufgeführt	k. P.
U Bonn	75,68	184,4	12
U Bremen	75,59	100,5	27
U Düsseldorf	75,58	88,1	29
U Göttingen	75,41	247,6	6
U Potsdam	74,66	Nicht aufgeführt	k. P.
U Hamburg	74,55	143,9	18
U Gießen	74,46	72,2	39
U Marburg	74,16	86,2	30
U Jena	74,15	81,6	35
U Duisburg-Essen	73,44	82,7	34
U Saarbrücken	73,42	72,3	38
HU Berlin	73,30	195,8	9
U Regensburg	72,46	85,1	32
U Leipzig	71,80	85,3	31
U Augsburg	71,69	Nicht aufgeführt	k. P.
U Mainz	70,10	107,4	25
U Siegen	68,71	Nicht aufgeführt	k. P.
U Halle-Wittenberg	64,04	68,4	40
U Wuppertal	63,79	Nicht aufgeführt	k. P.
U Hagen	63,17	Nicht aufgeführt	k. P.
U Tübingen	59,34	169,0	14

Anmerkungen: k. P. = keine Platzierung unter den bei der DFG-Mitteleinwerbung erfolgreichsten 40 Universitäten. Die grau abgestuften Schattierungen weisen die Quartile der Verteilung aus. Die fett gedruckte und hell hinterlegte Universität markiert den Median der Verteilung.

Quellen: Statistisches Bundesamt, Stichtag: 31.12.2013 und DFG-Förderatlas 2015

Vgl.: Leschner, Franziska; Krüger, Anne u.a.: Beschäftigungsbedingungen und Personalpolitik an der Universitäten in Deutschland im Vergleich.
IN: Keller, Andreas (Hg.): Von Pakt zu Pakt. Perspektiven der Hochschul-und Wissenschaftsförderung. = GEW. Material aus Hochschule und Forschung.123. Bielefeld 2017. S.183f


90
85
80
75
70
65
60
55
Befristungsgrad in %
0
5
10
15
20
25
30
35
40
Ranking nach DFG

Grundannahme der Linearen Regression

$$Y = \beta_1 + \beta_2 X$$

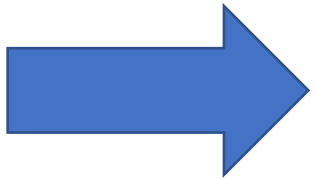
Berechnung der Koeffizienten

Ziel: Minimiere die Summe

der quadratischen Abweichung

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{n} e_i^2$$

$$= \min_{\beta_1,\beta_2} \sum_{i=1}^{n} (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i))^2$$



$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i + \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

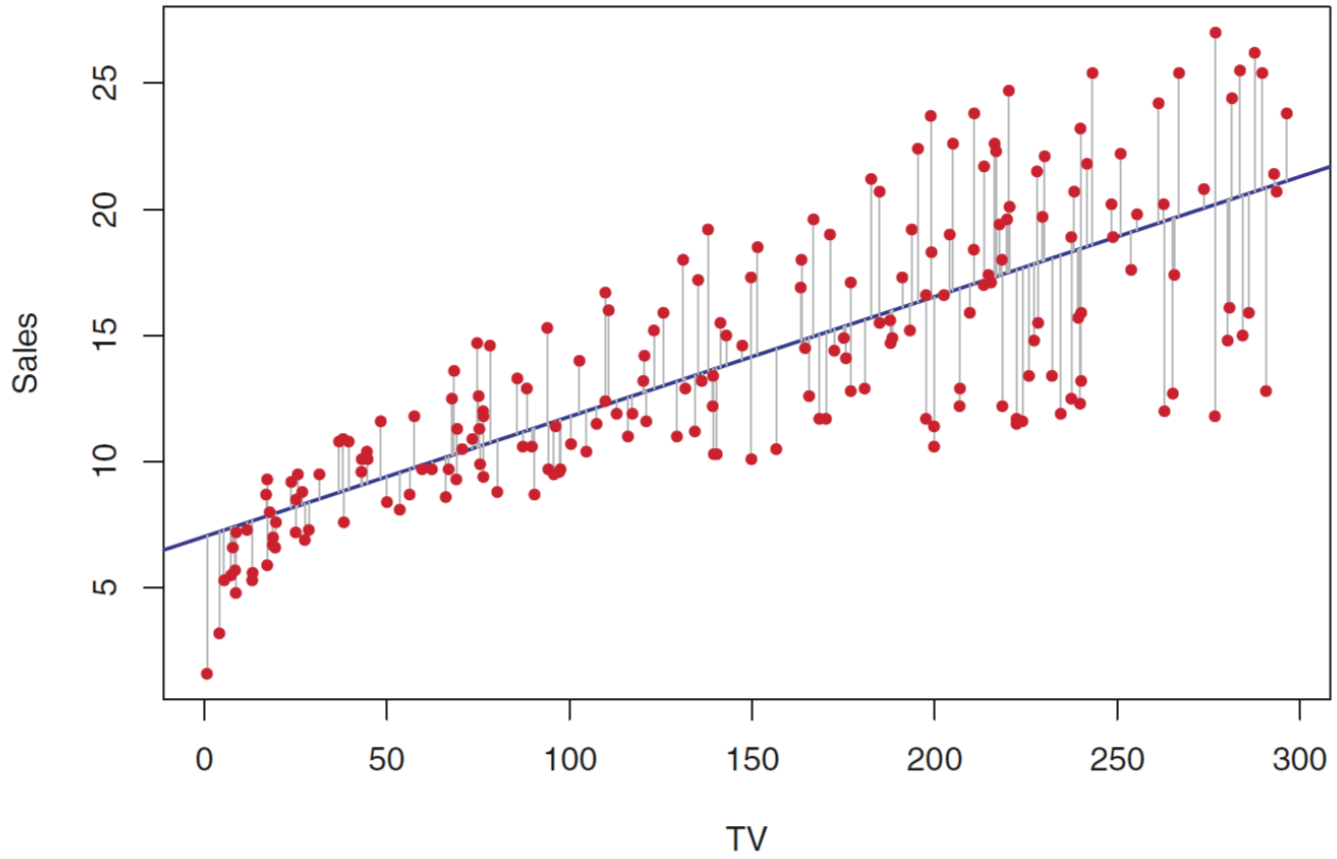


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

Vgl.: ISLR. S.62

Standardfehler des Residuums

$$\sigma^2 = Var(\epsilon),$$

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}$$

Standardfehler der Koeffizienten

$$SE(\hat{\beta}_1)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i + \bar{x})^2} \right]$$

$$SE(\hat{\beta}_2)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i + \bar{x})^2}$$

Mit Hilfe des Standardfehlers können wir nun abschätzen, dass β_1 und β_2 mit 95% Wahrscheinlichkeit in den Intervallen

$$\left[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1) \right]$$

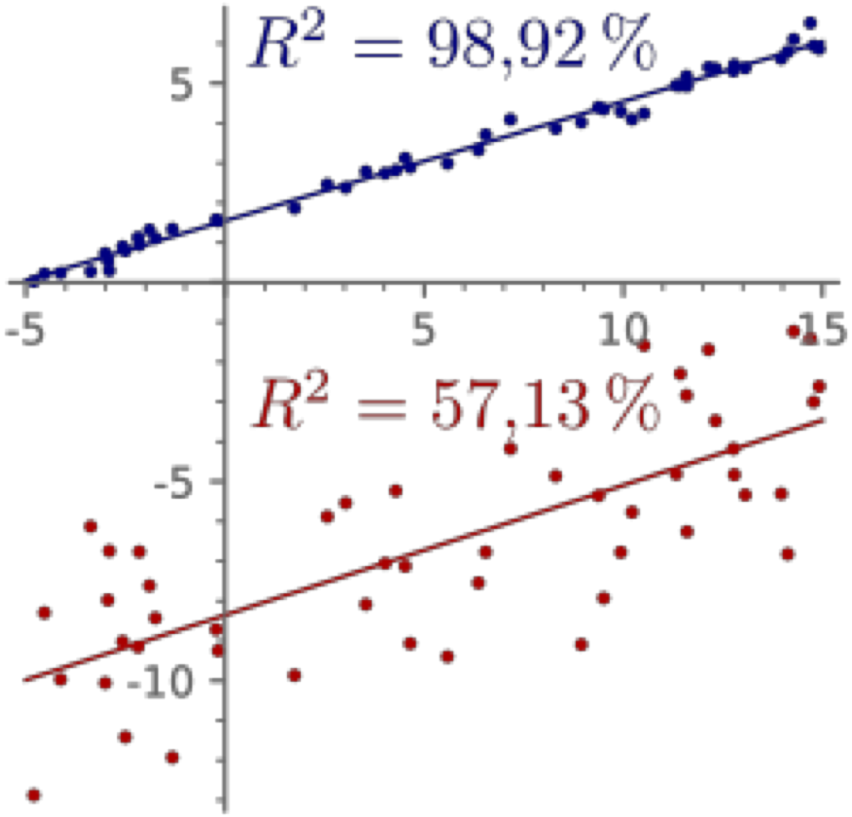
$$\left[\hat{\beta}_2 - 2 \cdot SE(\hat{\beta}_2), \hat{\beta}_2 + 2 \cdot SE(\hat{\beta}_2) \right]$$

liegen.

Bestimmungsmaß

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0, 1]$$



Vgl.: https://de.wikipedia.org/wiki/Bestimmtheitsmaß

Zurück zum Anfang

Die Lineare Regression ergibt die folgende Funktion:

$$Y = 80.4 - 0.1855 \cdot X$$