7 Zusammenfassung

Zusammenfassung des Blocks Multiple lineare Regression (Regression 1)

7.1 Einfache lineare Regression

a Das **Modell** der einfachen linearen Regression lautet

$$Y_i = \alpha + \beta x_i + E_i \ .$$

Die x_i sind feste Zahlen. Die E_i sind zufällig und werden zufällige Abweichungen oder Zufallsfehler genannt. Es wird (normalerweise) angenommen, dass

$$E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle \ , \qquad E_i, E_k \text{ unabhängig}$$

sei. (Man spricht auch bei anderen Annahmen über die Zufallsfehler von einfacher linearer Regression.)

Die Parameter des Modells sind die **Koeffizienten** α, β und die Standardabweichung σ des Zufallsfehlers.

Abbildung 7.1.a veranschaulicht das Modell. Es ist nützlich, sich simulierte Datensätze zum Modell vorzustellen(Abbildung 2.1.i) .

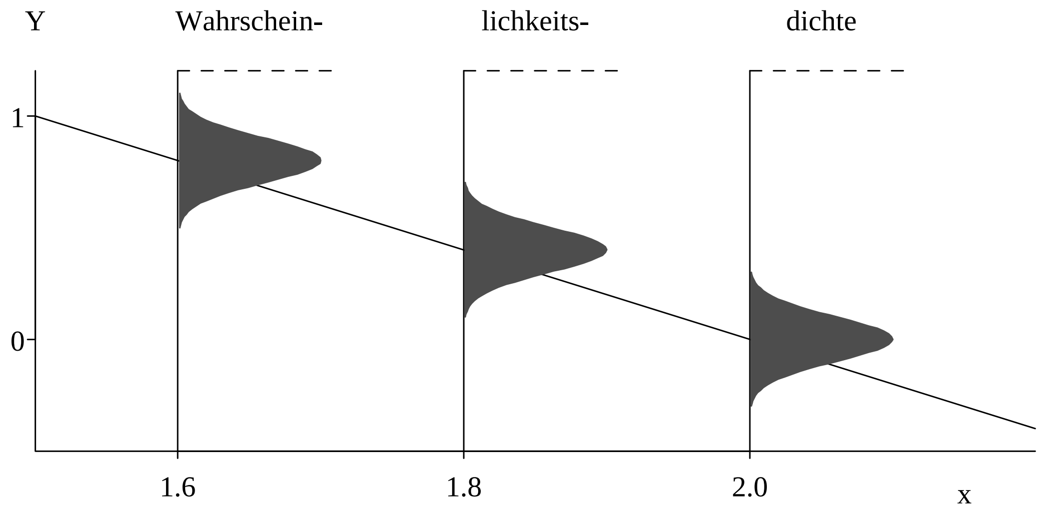

Abbildung 7.1.a: Veranschaulichung des Wahrscheinlichkeitsmodells $Y_i = 4 - 2x_i + E_i$ für drei Beobachtungen Y_1, Y_2 und Y_3 zu den x-Werten $x_1 = 1.6$, $x_2 = 1.8$ und $x_3 = 2$

b Die **Schätzung der Koeffizienten** erfolgt über das Prinzip der **Kleinsten Quadrate**, das man aus dem Prinzip der Maximalen Likelihood herleiten kann. Das ergibt

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \;, \qquad \widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\bar{x} \;.$$

Die Schätzungen sind normalverteilt,

$$\widehat{\beta} \sim \mathcal{N}\langle\beta,\, \sigma^2/\mathrm{SSQ}^{(X)}\rangle \;, \qquad \widehat{\alpha} \sim \mathcal{N}\left\langle \alpha,\, \sigma^2\left(\frac{1}{n} + \bar{x}^2/\mathrm{SSQ}^{(X)}\right)\right\rangle \;,$$

$$\mathrm{SSQ}^{(X)} = \sum\nolimits_{i=1}^{n}(x_i - \bar{x})^2 \;.$$

Sie sind also erwartungstreu. Ihre Varianz ist, wenn das Modell stimmt, die kleinstmögliche (unter den erwartungstreuen Schätzungen).

c Die Abweichungen der beobachteten Werte Y_i von den geschätzten oder angepassten Werten $\widehat{\alpha} + \widehat{\beta}x_i$ heissen **Residuen** und sind „Schätzungen" für die Zufallsfehler E_i.

Sie führen zur Schätzung der Standardabweichung σ des Zufallsfehlers,

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} R_i^2 \;.$$

d **Test** für die Nullhypothese $\beta = \beta_0$: Die Testgrösse

$$T = \frac{\widehat{\beta} - \beta_0}{\mathrm{se}^{(\beta)}} \;, \quad \mathrm{se}^{(\beta)} = \sqrt{\widehat{\sigma}^2/\mathrm{SSQ}^{(X)}}$$

ist t-verteilt mit $n-2$ Freiheitsgraden.

Daraus erhält man das Vertrauensintervall

$$\widehat{\beta} \pm q_{0.975}^{t_{n-2}}\,\mathrm{se}^{(\beta)} \;, \quad \mathrm{se}^{(\beta)} = \widehat{\sigma}/\sqrt{\mathrm{SSQ}^{(X)}} \;.$$

Programm-Ausgabe: siehe multiple Regression.

e Das „**Vertrauensband**" für den Wert der Regressionsfunktion verbindet die Endpunkte der Vertrauensintervalle für $\mathcal{E}\langle Y|x\rangle = \alpha + \beta x$.

Ein **Prognose-Intervall** soll einen (noch unbekannten) Messwert Y_0 für gegebenes x_0 enthalten – mit der vorgegebenen „statistischen Sicherheit" (von meistens 95%). Verbindet man die Endpunkte für verschiedene x_0, so erhält man das „Prognoseband".

7.2 Multiple lineare Regression

a Das **Modell** lautet

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + E_i \\ E_i &\sim \mathcal{N}\langle 0, \sigma^2 \rangle , \qquad E_i, E_k \text{ unabhängig} . \end{aligned}$$

In **Matrix-Schreibweise**:

$$\underline{Y} = \widetilde{\boldsymbol{X}} \underline{\widetilde{\beta}} + \underline{E} , \qquad \underline{E} \sim \mathcal{N}_n \langle \underline{0}, \sigma^2 \boldsymbol{I} \rangle .$$

b Die Schätzung erfolgt wieder über das Prinzip der Kleinsten Quadrate,

$$\underline{\widehat{\beta}} = (\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^T \underline{Y} .$$

Aus der Verteilung der geschätzten Koeffizienten

$$\widehat{\beta}_j \sim \mathcal{N} \left\langle \beta_j, \sigma^2 \left((\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}})^{-1} \right)_{jj} \right\rangle$$

erhält man t-Tests und Vertrauensintervalle für einzelne Koeffizienten.

Die Standardabweichung σ wird geschätzt durch

$$\widehat{\sigma}^2 = \sum_{i=1}^n R_i^2 \Big/ (n-p) .$$

c Tabelle 7.2.c zeigt eine **Programm-Ausgabe**, angereichert durch die mathematischen Symbole.

Die **multiple Korrelation** R ist die Korrelation zwischen den angepassten $\widehat{y}_i$ und den beobachteten Werten Y_i. Ihr Quadrat misst auch den „durch die Regression erklärten Anteil der Varianz“

$$R^2 = 1 - \mathrm{SSQ}^{(E)} / \mathrm{SSQ}^{(Y)}$$

und heisst deshalb **Bestimmtheitsmass**.

d **Vielfalt der Fragestellungen**. Das Modell der multiplen linearen Regression kann viele Situationen beschreiben:

- **Transformationen** der X- (und Y-) Variablen können aus ursprünglich nicht-linearen Zusammenhängen lineare machen.
- Ein Vergleich von zwei Gruppen lässt sich mit einer zweiwertigen X-Variablen, von mehreren Gruppen mit einem „Block“ von dummy Variablen als multiple Regression schreiben. Auf diese Art werden **nominale erklärende Variable** in ein Regressionsmodell aufgenommen.
- Die Vorstellung von verschiedenen linearen Abhängigkeiten für verschiedene Gruppen von Daten kann als ein einziges Modell hingeschrieben werden. Allgemeiner können **Wechselwirkungen** zwischen erklärenden Variablen durch zusätzliche Terme ins Modell aufgenommen werden.
- Die **polynomiale Regression** ist ein Spezialfall der multiplen linearen (!) Regression.

Coefficients:

	Value	Std. Error	t value	Pr(> |t|)
(Intercept)	19.7645	2.6339	7.5039	0.0000
pH	-1.7530	0.3484	-5.0309	0.0000
lSAR	-1.2905	0.2429	-5.3128	0.0000

Residual standard error: $\widehat{\sigma} = 0.9108$ on $n - p = 120$ degrees of freedom
Multiple R-Squared: $R^2 = 0.5787$

Analysis of variance

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Regression	$m =$ 2	$SSQ^{(R)} = 136.772$	68.386	$T = 82.43$	0.0000
Residuals	$n-p = 120$	$SSQ^{(E)} = 99.554$	$\widehat{\sigma}^2 = 0.830$		*P-Wert*
Total	122	$SSQ^{(Y)} = 236.326$			

Tabelle 7.2.c: Computer-Output für das Beispiel der basischen Böden mit Varianzanalyse-Tabelle und der im folgenden verwendeten Notation

e Der **F-Test zum Vergleich von Modellen** ermöglicht es, zu prüfen, ob mehrere Koeffizienten =0 sind. Das ist nötig, um zu testen, ob eine nominale Variable einen Einfluss auf die Zielgrösse hat.

7.3 Residuen-Analyse

a Die Annahmen des Modells der multiplen linearen Regression kann man auflösen in

(a) ihr Erwartungswert ist $\mathcal{E}\langle E_i \rangle = 0$ (oder: die Regressionsfunktion ist korrekt),

(b) sie haben alle die gleiche Streuung, $\text{var}\langle E_i \rangle = \sigma^2$,

(c) sie sind normalverteilt.

(d) Die E_i sind unabhängig,

Diese Voraussetzungen sollen überprüft werden, um

- aus Abweichungen auf bessere Modelle zu schliessen,
- Tests und Vertrauensintervalle zu rechtfertigen.

Abweichungen werden mit grafischen Darstellungen entdeckt. Tests spielen eine untergeordnete Rolle.

b Die folgenden grafischen Darstellungen sind nützlich:

(a) **Nicht-Linearitäten**: Streudiagramme der (unstandardisierten) Residuen gegen angepasste Werte **(Tukey-Anscombe plot)** und gegen die (ursprünglichen) **erklärenden Variablen**.
Wechselwirkungen: Pseudo-dreidimensionales Diagramm der (unstandardisierten) Residuen gegen je zwei erklärende Variable.

(b) **Gleiche Streuungen**: Streudiagramme der (standardisierten) absoluten Residuen gegen angepasste Werte und gegen die (ursprünglichen) erklärenden Variablen (meist nicht speziell dargestellt, mit den Streudiagrammen unter (a) mitbetrachtet).

(c) **Normalverteilung: Q-Q-plot** (oder Histogramm) der (standardisierten) Residuen.

(d) **Unabhängigkeit**: (Unstandardisierte) Residuen gegen die Zeit oder gegen den Ort auftragen.

(*) **Einflussreiche Beobachtungen** für die gesamte Anpassung: Streudiagramm der (standardisierten) Residuen gegen die **leverage**.
Einflussreiche Beobachtungen für einzelne Koeffizienten: added-variable plot.

(*) **Kollinearitäten**: Scatterplot matrix (Streudiagramme der erklärenden Variablen gegeneinander) und numerische Werte (R_j^2 oder VIF_j oder „tolerance“).

c Massnahmen zur Verbesserung eines Modells:

- **Transformation der Zielgrösse**: bei schiefer Verteilung, Nicht-Linearitäten, ungleichen Streuungen.
- **Transformation** (nicht-lineare) von **erklärenden Variablen**: bei Nicht-Linearitäten, Hebelpunkten (schiefe Verteilung der erklärenden Variablen und einzelne hohe leverages) und Wechselwirkungen.
- **Zusätzliche Terme**: bei Nicht-Linearitäten und Wechselwirkungen.
- Lineare Transformation von mehreren erklärenden Variablen: bei **Kollinearitäten**.
- **Gewichtete Regression**: bei ungleichen Streuungen.
- Überprüfung der Korrektheit von Beobachtungen: bei **Ausreissern**.
- Verwerfung von Ausreissern: wenn solche vorhanden sind und man nicht robust rechnen will (siehe unten).

Wo geht's weiter? Noch nicht behandelte Methoden:

- Verallgemeinerte Kleinste Quadrate: bei stochastischen Abhängigkeiten der Zufallsfehler.
- Nicht-lineare Regression: bei Nicht-Linearitäten, wenn Transformationen nicht zum Ziel führen oder vom Anwendungszweck her nicht zulässig sind.
- Robuste Regression: immer; vor allem bei Ausreissern und langschwänzigen Verteilungen.

d **Eine Regressions-Analyse ohne Residuen-Analyse ist eine unnütze Rechnung!**

7.4 Modellwahl

a Explorative Analyse versucht, das **Modell aus den Daten** zu entwickeln – hoffentlich in enger „Zusammenarbeit" mit dem Fachwissen. (Manchmal steht das Modell natürlich bereits am Anfang fest.)

b Zur **automatisierten Variablenwahl** gibt es „schrittweise" Verfahren (rückwärts, vorwärts, gemischt) und die Prüfung aller möglichen Gleichungen (all subsets).

Für die Güte von Modellen gibt es verschiedene Kriterien. Bei gleicher Anzahl x-Variablen geben alle Kriterien die gleiche Reihenfolge, also insbesondere das gleiche beste Modell.

Jede zusätzliche Variable führt zu einem höheren Bestimmtheitsmass. Für den Vergleich von Modellen mit verschieden vielen x-Variablen muss man sich deshalb überlegen, wie stark die Aufnahme einer zusätzlichen Variablen „bestraft" werden soll. Ein gebräuchliches Kriterium wird C_p genannt. Die übliche Regel, auf die **Signifikanz der einzelnen Terme** zu achten, ist wegen des multiplen Testens **fragwürdig**.

Fazit: Die automatisierte Variablenauswahl soll benützt werden, um mehrere von den Daten her geeignete Modelle auszuwählen, unter denen mit Fachwissen ausgewählt wird.

7.5 Ergänzungen

a **Fehlerbehaftete erklärende Variable**. Sofern die interessierende erklärende Variable nicht genau, sondern, wie die Zielgrösse, nur mit zufälligen Abweichungen beobachtet werden kann, und man an der Schätzung der Koeffizienten interessiert ist, muss man das Modell und die Schätzmethoden ändern. Mit Kleinsten Quadraten schätzt man eine zu flache Gerade.

b **Eichproblem**: Im üblichen Fall ist für die „Trainingsdaten" ein Messfehler in der einfach beobachtbaren Variablen zu postulieren. Sie wird deshalb als Zielgrösse eingeführt. In der Anwendung wird die Beziehung in umgekehrter Richtung benützt (inverse Regression). Das führt zu speziellen Überlegungen für die Schätzung der Parameter, der Bestimmung von x-Werten zu gegebenem Ablesewert Y und die Beurteilung von deren Genauigkeit.

Literatur zur linearen Regression

a Kurze Einführungen in die Regression in deutscher Sprache gibt es etliche:

- Linder und Berchtold (1982, Kap. 3) führen auf einfachem Niveau in die Methoden ein.
- Vincze (1984) beschreibt in Band 2 auf 90 Seiten die lineare Regression in recht theoretischer Art.

In englischer Sprache enthält das auch sonst empfehlenswerte Einführungsbuch von Devore (1991) zwei Kapitel zur Regression.

b Die Literatur zum Thema Regression ist äusserst umfangreich, besonders im englischen Sprachbereich.

- Ein neueres Buch, das auch in allgemeinere Regressionmodelle einführt, ist Ryan (1997).
- Weisberg (1990) betont die explorative Suche nach einem geeigneten Modell – eine empfehlenswerte Einführung in die Praxis der Regressionsanalyse mit vielen Beispielen.
- Draper and Smith (1998): Ein klassisches Einführungsbuch, das der überprüfung der Voraussetzungen die nötige Beachtung schenkt.
- Daniel and Wood (1980): Empfehlenswertes, anwendungsorientiertes Buch, das zur Entwicklung der explorativen Datenanalyse beigetragen hat und deshalb bereits zu den Klassikern gehört.
- Sen and Srivastava (1990): Mathematische Theorie und Anwendungsaspekte werden diskutiert. Empfohlen für mathematisch Interessierte.
- ... und viele andere.

c **Spezielle Hinweise**

- Wetherill (1986) behandelt einige spezielle Probleme der multiplen linearen Regression ausführlicher, insbesondere die **Kollinearität**.
- In Cook and Weisberg (1999) wird gezeigt, wie man mit modernen grafischen Mitteln Modelle (nicht nur lineare) von Grund auf entwickeln kann. Es führt in ein dafür entwickeltes, einfach zu bedienendes Computer-Paket (R-code) ein, das mit dem Buch mitgeliefert wird.
- Mosteller and Tukey (1977): Explorative Datenanalyse mit vielen Ideen.

d In den Büchern über allgemeinere Regressionsmodelle wird die multiple lineare Regression normalerweise mit diskutiert. Man könnte also die Literaturliste zu jenen Themen hier anfügen.

Literaturverzeichnis

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, N.Y.

Agresti, A. (1996). *Introduction to categorical data analysis*, Wiley Series in Probability & Math. Statistics, Wiley, New York.

Christensen, R. (1990). *Log-linear models*, Springer, N.Y.

Cleveland, W. S. (1994). *The Elements of Graphing Data*, 2nd edn, Hobart Press, Summit, New Jersey.

Clogg, C. C. and Shihadeh, E. S. (1994). *Statistical models for ordinal variables*, Sage, Thousand Oaks, CA.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table, *Communications in Statistics – Theory and Methods* **A9**: 1025–1041.

Collet, D. (1991, 1999). *Modelling binary data*, Chapman & Hall/CRC Press LLC, Boca Raton, Florida.

Cook, R. D. and Weisberg, S. (1999). *Applied regression including computing and graphics*, Wiley, N.Y.

Cox, D. R. (1989). *Analysis of Binary Data*, 2nd edn, Chapman and Hall, London.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*, Chapman and Hall, London.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, N.Y.

Davies, P. (1995). Data features, *Statistica Neerlandica* **49**: 185–245.

Devore, J. L. (1991). *Probability and Statistics for Engineering and the Sciences*, 3rd edn, Duxbury Press, Belmont, California.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Draper, N. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edn, Wiley, N.Y.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer-Verlag, New York.

Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics, *Journal of the American Statistical Association* **87**: 178–183.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N.Y.

Haaland, P. D. (1989). *Experimental Design in Biotechnology*, Marcel Dekker, N.Y.

Hartung, J., Elpelt, B. und Klösener, K. (1998). *Statistik. Lehr- und Handbuch der angewandten Statistik*, 11. Aufl., Oldenbourg, München.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*, Wiley, N.Y.

Linder, A. und Berchtold, W. (1982). *Statistische Methoden II: Varianzanalyse und Regressionsrechnung*, Birkhäuser, Basel.

Lindsey, J. K. (1995). *Modelling Frequency and Count Data*, number 15 in *Oxford Statistical Science Series*, Clarendon Press, Oxford.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading, Massachusetts.

Myers, R. H., Montgomery, D. C. and Vining, G. G. (2001). *Generalized Linear Models. With Applications in Engineering and the Sciences*, Wiley Series in Probability and Statistics, Wiley, NY.

Ryan, T. P. (1997). *Modern Regression Methods*, Series in Probability and Statistics, Wiley, N.Y. *includes disk*

Sachs, L. (1997). *Angewandte Statistik*, 8. Aufl., Springer, Berlin.

Sen, A. and Srivastava, M. (1990). *Regression Analysis; Theory, Methods, and Applications*, Springer-Verlag, N.Y.

Stahel, W. A. (2000). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 3. Aufl., Vieweg, Wiesbaden.

Stahel, W. A. (2002). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 4. Aufl., Vieweg, Wiesbaden.

van der Waerden, B. L. (1971). *Mathematische Statistik*, 3. Aufl., Springer, Berlin.

Vincze, I. (1984). *Mathematische Statistik mit industriellen Anwendungen*, Band1, 2, 2. Aufl., Bibliograhisches Institut, Mannheim.

Weisberg, S. (1990). *Applied Linear Regression*, 2nd edn, Wiley, N.Y.

Wetherill, G. (1986). *Regression Analysis with Applications*, number 27 in *Monographs on Statistics and Applied Probability*, Chapmann and Hall, London.