Vorwort zur zweiten Auflage

Die zweite Auflage unserer Einführung in die multivariate Statistik weist im Wesentlichen drei Änderungen auf: Wir orientieren uns nun erstens an den Prozeduren und Ausgaben der *Open-Source*-Programmiersprache R. Im Text werden die Befehle und Ausgaben der entsprechenden R-Funktionen zur Erläuterung abgedruckt. Zu jeder Analyse findet sich im *Online-Plus-Material*[1] (vgl. dazu auch den Anhang) der entsprechende Datensatz sowie ein R-*Notebook*, welches die entsprechenden Befehle enthält. So kann jede Analyse selbst nachvollzogen werden. Die an SPSS orientierten *online*-Materialien bleiben verfügbar. Zweitens wurde das Kapitel über die Hierarchischen Linearen Modelle deutlich erweitert. Die Einführung in diese Modelle ist jetzt in Kap. 7 zu finden. Drittens wurde das Buch generell überarbeitet, korrigiert und ergänzt.

Herzlich bedanken möchten wir uns bei Joachim Coch und Dr. Angelika Schulz vom Springer-Verlag für die kompetente Unterstützung sowie Prof. Dr. Jürgen Kriz für die Einladung zu diesem Buch und für seine hilfreichen Anregungen.

Saarbrücken, Deustchland
August 2022

Dirk Wentura
Benedikt Wirth
Markus Pospeschill

[1] http://www.lehrbuch-psychologie.springer.com/.

Lineare Regression

2

Bei der linearen Regression wird eine Kriteriumsvariable *Y* auf die Prädiktorvariable *X* „zurückgeführt", indem die beste lineare Gleichung

$$\hat{Y} = b_0 + b_1 X$$

gesucht wird. Was heißt hierbei „beste" Gleichung? Es lassen sich sicherlich mehrere Kriterien denken; aus verschiedenen Gründen bietet sich das *Kriterium der kleinsten Quadrate* an, das heißt. die Parameter b_0 und b_1 werden so bestimmt, dass die Summe der quadrierten Abweichungen der vorhergesagten *Y*-Werte von den tatsächlichen *Y*-Werten minimiert wird. Diese Abweichungen nennen wir Residuen. Wir können diese auch explizit in die Gleichung aufnehmen:[1]

$$Y = b_0 + b_1 X + e$$

Ein Grund für die Wahl dieses Kriteriums liegt darin, dass die Fehlervarianz (also die nicht vorhergesagte Varianz von *Y*) minimiert wird; ein zweiter, dass durch die Quadrierung „zwanglos" das Vorzeichen der Abweichungen eliminiert wird. Der Algorithmus zur Bestimmung der Funktionsparameter braucht uns hier nicht zu interessieren (vgl. z. B. Bortz & Schuster, 2010, Kap. 11), da wir wissen, welches Kriterium er realisiert. Ein Wort noch zur Terminologie: Um die lineare Regression mit nur einem Prädiktor von der multiplen Regression abzugrenzen, die wir im

[1] Beachten Sie: In der oberen Gleichung steht links „Y-Dach", das heißt, die Variable der vorhergesagten Werte; in der zweiten Gleichung ist Y die Variable der gemessenen Werte.

D. Wentura et al., *Multivariate Datenanalyse mit R*, Basiswissen Psychologie,
https://doi.org/10.1007/978-3-662-65522-1_2

nächsten Kapitel behandeln, spricht man auch von *bivariater linearer Regression* (bivariat, da der Zusammenhang nur zweier Variablen bestimmt wird).

Schauen wir uns ein Beispiel an: Die Durchschnittsnote von 120 Schülern (Variable *Schule*) sei die abhängige Variable; sie wird auf den Intelligenzwert der Schüler (Variable *IQ*) regrediert. Abb. 2.1 zeigt das Streudiagramm der Daten. (Die Daten sind fiktiv und in mancherlei Hinsicht unrealistisch) Zur Berechnung nutzen wir die R-Funktion lm().[2]

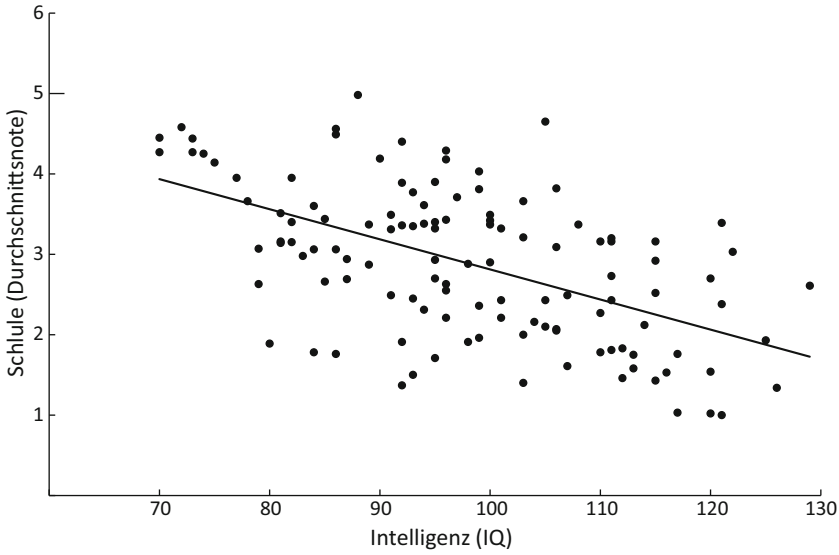

Abb. 2.1 Streudiagramm Schulnote und Intelligenz (fiktive Daten)

[2] „lm" steht für *linear model*. Wir folgen in diesem Buch der üblichen Konvention und geben R-Funktionsnamen immer mit einer offenen und geschlossenen Klammer an. Dies zeigt an, dass der Funktion im Anwendungsfall noch Argumente übergeben werden müssen (siehe Anhang I und das *Online-Plus*-Material für weitere Informationen zur Anwendung von R-Funktionen). Funktionen sind immer Teil von *Packages*. Eine Liste der in diesem Buch genutzten Funktionen mit der Nennung ihrer Packages findet sich am Ende von Anhang I (Tab. A.1).

```
> bivreg_intschule <- lm(formula = SCHULE ~ IQ, data = intschule)
> summary(bivreg_intschule)

Call:
lm(formula = SCHULE ~ IQ, data = intschule)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74212 -0.54420  0.00716  0.58126  2.02098

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  6.555844   0.520062  12.606 < 0.0000000000000002 ***
IQ          -0.037436   0.005258  -7.119      0.0000000000916 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7885 on 118 degrees of freedom
Multiple R-squared:  0.3005,  Adjusted R-squared:  0.2945
F-statistic: 50.68 on 1 and 118 DF,  p-value: 0.000000000009156

> lm.beta(bivreg_intschule)
       IQ
-0.5481515

> anova(bivreg_intschule)
Analysis of Variance Table

Response: SCHULE
           Df Sum Sq Mean Sq F value             Pr(>F)
IQ          1 31.510 31.5099  50.685 0.000000000009156 ***
Residuals 118 73.359  0.6217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Abb. 2.2 R-Ausgabe der Prozedur *lm* (plus Zusatzausgaben)

Wir erhalten hier eine Fülle von Informationen, die einzeln besprochen werden sollen (Abb. 2.2).

Regressionsgewichte (1) An dieser Stelle sind die Parameter b_0 und b_1 der Regressionsgleichung angegeben. Die beste Schätzung für die Schulnote ergibt sich somit aus (gerundete Werte):

$$\hat{S}chule = 6.556 - 0.037 \cdot IQ$$

Welche Note ist die beste Schätzung für einen durchschnittlich intelligenten Schüler? Da 100 der Durchschnittswert eines Standard-Intelligenztests ist, ist die Vorhersage 2.86 ($= 6.556 - 0.037 \cdot 100$). Welche Note ist die beste Schätzung für einen Schüler, der zwei Norm-Standardabweichungen über dem Mittelwert liegt? Da die Standardabweichung des Intelligenztests 15 beträgt, ist die Vorhersage 1.75 ($= 6.556 - 0.037 \cdot 130$).

Standardschätzfehler der Regressionsgewichte (2) Sie geben die Genauigkeit an, mit der aus den Stichprobendaten die Ausprägung der Regressionsparameter geschätzt werden kann: Würden wir unsere Erhebung (immer mit 120 Schülern) viele Male wiederholen, so hätte die Verteilung des Regressionsparameters diese Standardabweichung.

t-*Wert des Signifikanztests (3)* Der Wert ergibt sich – ganz analog zum Einstichproben-*t*-Test (Kap. 1) – durch:

$$t = \frac{b}{s_b}$$

Es wird also die Hypothese getestet, ob der entsprechende Regressionsparameter bedeutsam von Null abweicht.

Wahrscheinlichkeitsniveau des t-Wertes (4) Das Wahrscheinlichkeitsniveau des *t*-Wertes. Hier ist $p < .001$; es ist also sehr unwahrscheinlich, ein solches oder (vom Betrag) noch größeres Regressionsgewicht zu erhalten, wenn in der Population das Gewicht Null beträgt. Der *p*-Wert wird stets zweiseitig angegeben. Hat man eine einseitige Hypothese, so kann der Wert halbiert werden.

Beta-Gewicht (Standardpartial-Regressionskoeffizient; 5) Dieser Parameter wird nicht in der Standardausgabe von lm() ausgegeben; er muss mittels der Funktion lm.beta() extra angefordert werden (Abb. 2.2). Zum Verständnis dieses standardisierten Koeffizienten ist es nützlich zu wissen, dass (1) bei der bivariaten linearen Regression das Beta-Gewicht mit der Produkt-Moment-Korrelation identisch ist und (2) bei *z*-Standardisierung von Kriterium und Prädiktor das Regressionsgewicht b_1 gleich dem Beta-Gewicht ist (während die Konstante b_0 den Wert Null annimmt). Insbesondere bei den multiplen Regressionen, die später erläutert werden, wird in der Regel das Beta-Gewicht berichtet, wenn der Beitrag eines Prädiktors in Richtung und Ausprägung prägnant benannt werden soll. Der Zusammenhang zwischen b_1 und Beta-Gewicht ergibt sich nach folgender einfacher Formel:

$$b_1 = \beta \frac{s_y}{s_x}$$

Beachten Sie aber, dass das Beta-Gewicht zwar in Standardfällen im Bereich von −1 bis +1 liegt (wie die Korrelation), aber formal nicht auf dieses Intervall begrenzt ist. In manchen Fällen der multiplen Regression, die wir später noch kennenlernen werden, kann es Werte außerhalb dieses Bereichs annehmen.

Multiples Korrelationsquadrat (R^2; 6) Um diesen Wert zu verstehen, muss man zunächst die *multiple Korrelation* einführen. Die multiple Korrelation ist die Korrelation zwischen dem Kriterium *Y* und dem durch die Regressionsgleichung geschätzten Kriterium *Ý*. Im Fall der bivariaten Regression ist diese Korrelation identisch mit der Produkt-Moment-Korrelation von Prädiktor und Kriterium (und damit auch identisch mit dem Beta-Gewicht. Das muss auch so sein, wie eine einfache Überlegung deutlich macht: Die Korrelation zwischen *Schule* und *Śchule* (= 6.556 – 0.037 · IQ) muss identisch mit der Korrelation zwischen *Schule* und *IQ* sein, da *Śchule* lediglich eine Lineartransformation von *IQ* ist. Korrelationen sind aber invariant gegenüber Lineartransformationen der beteiligten Variablen. Wie der Name *multiples Korrelationsquadrat* sagt, handelt es sich bei diesem Wert um das Quadrat der multiplen Korrelation. Es lässt sich leicht zeigen, dass dieser Wert ein Index der „erklärten Varianz" des Kriteriums durch den Prädiktor ist. Aus diesem Grund wird er auch *Determinationskoeffizient* genannt. Um den Begriff der „erklärten Varianz" besser zu verstehen, nehmen wir ihn ganz wörtlich. Wir bilden zwei neue Variablen: (1) *S_SCHULE* (durch die Regressionsgleichung) und *R_SCHULE* (die Differenz zwischen *SCHULE* und *S_SCHULE* – die sogenannten Residuen; zu den entsprechenden R-Kommandos vgl. Abb. 2.3 und *Online Plus*; s. Anhang).

$$\text{S\_SCHULE} = 6.556 \; - \; 0.037 \cdot \; \text{IQ}$$

$$\text{R\_SCHULE} = \text{SCHULE} \; - \; \text{S\_SCHULE}$$

Berechnet man jetzt die Varianzen der Variablen *SCHULE*, *S_SCHULE* und *R_SCHULE*, sieht man, was mit „erklärter Varianz" gemeint ist. Abb. 2.3 enthält die Ausgabe der Varianzen.

Teilen Sie die Varianz von *S_SCHULE* (0.265) durch die Varianz von *SCHULE* (0.881), erhalten Sie den Wert des multiplen Korrelationsquadrats. Die Varianzen von *S_SCHULE* („erklärte" Varianz), *R_SCHULE* (Fehlervarianz) ergänzen sich zur Varianz von *SCHULE. W*enn wir also die Varianz von *R_SCHULE* (0.616)

```
> intschule$S_SCHULE <- 6.555844 -0.037436 *intschule$IQ
> intschule$R_SCHULE <- intschule$SCHULE - intschule$S_SCHULE
> sapply(X = intschule[,c('SCHULE','S_SCHULE', 'R_SCHULE')],
                       FUN = var)

   SCHULE  S_SCHULE  R_SCHULE
0.8812486 0.2647889 0.6164598
```

Abb. 2.3 Ausgabe der Varianzen (*intschule* ist der Name des Datensatzes)

durch die Varianz von *SCHULE* (0.881) teilen und das Ergebnis von 1 abziehen, erhalten wir ebenfalls das multiple Korrelationsquadrat. Diese Darstellung werden wir gleich unten noch einmal benötigen.

Standardabweichung der Residuen (Populationsschätzer; 7) Die Varianz (und damit die Standardabweichung) der Residuen ist in der Abb. 2.3 auf die übliche Art (d. h. Quadratsumme geteilt durch n-1) bestimmt worden. Dies ist aber keine erwartungstreue Schätzung der Residuen, wie eine einfache Überlegung zeigt: So wie wir bei der Bestimmung der Varianz einer gemessenen Variable gesagt hatten, nur $n-1$ Werte der Quadratsumme können frei variieren (da der Mittelwert schon aus den gemessenen Werten bestimmt wurde, vgl. Kap. 1), so müssen wir jetzt feststellen, dass nur $n-2$ Residualwerte frei variieren können, da Kriteriums- und Prädiktorvariable in die Bestimmung der Residuen eingehen. Der „Residual standard error" – wie es im R-Protokoll heißt – ist somit einfach die Wurzel der durch die richtige Anzahl von Freiheitsgraden (hier: $n-2$) geteilten Quadratsumme der Residuen (9).

Das adjustierte multiple Korrelationsquadrat (8) Wegen des gerade erwähnten Freiheitsgradproblems ist das multiple R^2 kein erwartungstreuer Schätzer des Populations-R^2. Wie wir oben gesagt hatten, erhalten wir R^2 dadurch, dass wir das Verhältnis von Residuenvarianz zu Kriteriumsvarianz (d. h. die nicht erklärte Varianz) von 1 abziehen. Setzen wir statt der Residuenvarianz die Populationsschätzung der Residuenvarianz ein – das heißt, das Quadrat der gerade eingeführten *Standardabweichung der Residuen* (s. oben) – so erhalten wir das adjustierte R^2. Dieser Wert erhält eine wichtige Funktion vor allem bei der multiplen Regression (Kap. 3).

Quadratsummen (9) Der Ergebnisausdruck der univariaten Statistiken kann noch zu einer weiteren Erläuterung verwendet werden. Bekanntlich ergibt sich die Varianz (genauer: eine „erwartungstreue Schätzung der Populationsvarianz") durch folgenden Ausdruck:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

Wie wir wissen, wird der Ausdruck im Zähler auch als *Quadratsumme* bezeichnet. Wenn die Varianzen von *S_SCHULE* und *R_SCHULE* mit 119 (= *N*−*1*) multipliziert werden, erhält man die Quadratsummen (QS) für „Regression" und „Residuen", die auch im (erweiterten) Ergebnisausdruck (Abb. 2.2) zu finden sind. Wie man sich leicht überlegen kann, gilt dann auch:

$$R^2 = \frac{QS_{regression}}{QS_{regression} + QS_{residual}}$$

F-*Wert (10)* Während der *t*-Test, der jedem Regressionsparameter zugeordnet ist, eben diesen auf Abweichung von Null testet (s. oben), liefert der *F*-Test eine Entscheidungshilfe darüber, ob das Ausmaß der erklärten Varianz als statistisch signifikant angesehen werden soll. Der *F*-Wert ist der Quotient der mittleren Quadratsummen für „Regression" und „Residuen" (12), die ihrerseits durch Relativierung der entsprechenden Quadratsummen auf die Freiheitsgrade (13) berechnet werden.

Wahrscheinlichkeitsniveau des* F-*Wertes (11) Es ist zu beachten, dass auf einen *F*-Wert die Unterscheidung *einseitig vs. zweiseitig* prinzipiell nicht anwendbar ist, da mit dem *F*-Test Varianzverhältnisse getestet werden, die keine Richtungsunterschiede mehr enthalten. Im Übrigen ist an dem Beispiel aber zu erkennen, dass der *F*-Test (auf signifikante Varianzaufklärung) offenbar zu der gleichen Wahrscheinlichkeitsaussage führt wie der *t*-Test (auf Abweichung des Regressionsparameters von Null). In der Tat lassen sich diese beiden Tests ineinander überführen, wenn der *F*-Wert nur einen Zählerfreiheitsgrad hat (also nur ein Prädiktor getestet wird), wobei gilt:

$$t(df_n) = \sqrt{F(1, df_n)}$$

(mit 1 als Zählerfreiheitsgrad des *F*-Wertes, df_n als Nennerfreiheitsgrade). Wegen dieser Äquivalenz von *t*-Test und *F*-Test (mit einem Zählerfreiheitsgrad) kann mitunter auch ein *F*-Test einseitig interpretiert werden (Maxwell et al., 2017, S. 236 f.).

Mittlere Quadratsummen (12) Die mittleren Quadratsummen ergeben sich durch die Relativierung der Quadratsummen auf die Freiheitsgrade.