Lehr- und Handbücher der Statistik

Herausgegeben von
Universitätsprofessor Dr. Rainer Schlittgen

Böhning, Allgemeine Epidemiologie
Caspary · Wichmann, Lineare Modelle
Chatterjee · Price (Übers. Lorenzen), Praxis der Regressionsanalyse, 2. Auflage
Degen · Lorscheid, Statistik-Aufgabensammlung, 3. Auflage
Hartung, Modellkatalog Varianzanalyse
Har ey (Übers. Untiedt), Ökonometrische Analyse von Zeitreihen, 2. Auflage
Har ey (Übers. Untiedt), Zeitreihenmodelle, 2. Auflage
Heiler · Michels, Deskriptive und Explorative Datenanalyse
Kockelkorn, Lineare statistische Methoden
Miller (Übers. Schlittgen), Grundlagen der Angewandten Statistik
Nae e, Stochastik für Informatik
Oerthel · Tuschl, Statistische Datenanalyse mit dem Programmpaket SAS
Pflaumer · Heine · Hartung, Statistik für Wirtschaft- und Sozialwissenschaften: Deskriptive Statistik
Pflaumer · Heine · Hartung, Statistik für Wirtschafts- und Sozialwissenschaften: Induktive Statistik
Pokropp, Lineare Regression und Varianzanalyse
Rasch · Herrendörfer u.a., Verfahrensbibliothek, Band I und Band 2
Riedwyl · Ambühl, Statistische Auswertungen mit Regressionsprogrammen
Rinne, Wirtschafts- und Bevölkerungsstatistik, 2. Auflage
Rinne, Statistische Analyse multivariater Daten – Einführung
Rüger, Induktive Statistik, 3. Auflage
Rüger, Test- und Schätztheorie, Band I: Grundlagen
Schlittgen, Statistik, 9. Auflage
Schlittgen, Statistische Inferenz
Schlittgen · Streitberg, Zeitreihenanalyse, 8. Auflage
Schürger, Wahrscheinlichkeitstheorie
Tutz, Die Analyse kategorialer Daten

Fachgebiet Biometrie

Herausgegeben von Dr. Rolf Lorenz

Bisher erschienene Werke:

Bock, Bestimmung des Stichprobenumfangs
Brunner · Langer, Nichtparametrische Analyse longitudinaler Daten

Statistische Auswertungen mit Regressionsprogrammen

Lineare Regression und Verwandtes
Multivariate Statistik
Planung und Auswertung von Versuchen

Von
Universitätsprofessor
Dr. Hans Riedwyl
und
Dipl.-Statistiker Mathias Ambühl

R. Oldenbourg Verlag München Wien

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Riedwyl, Hans:
Statistische Auswertungen mit Regressionsprogrammen : lineare Regression und Verwandtes, multivariate Statistik, Planung und Auswertung von Versuchen / von Hans Riedwyl und Mathias Ambühl. - München ; Wien : Oldenbourg, 2000
(Lehr- und Handbücher der Statistik)
ISBN 3-486-25532-0

Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
www.oldenbourg-verlag.de

Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: Druckhaus „Thomas Müntzer“ GmbH, Bad Langensalza

ISBN 3-486-25532-0

Inhaltsverzeichnis

Vorwort

Die Methode der linearen Regression hat sich in den vergangenen Jahrzehnten als statistisches Analyseverfahren durchgesetzt und hat selbst unter gelegentlichen Anwendern von Statistik einen beachtlichen Bekanntheitsgrad erreicht. Infolgedessen sind Programme zur Berechnung der Kleinstquadrateschätzer, deren Standardabweichungen sowie der darauf basierenden t- und F-Tests der einfachen und multiplen Regression in zahlreichen Softwarepaketen implementiert. Dies gilt einerseits für alle Statistik-Softwarepakete, aber auch für Tabellenkalkulationsprogramme wie Excel. Somit darf davon ausgegangen werden, dass jeder Computerbenützer Zugang hat zu einem Programm, welches die wichtigsten Funktionen der Regressionsrechnung bewältigen kann.

Weniger bekannt ist die Tatsache, dass sich zahlreiche Probleme aus verschiedenen Anwendungsgebieten der angewandten Statistik durch einfache Tricks und Datenmanipulationen auf die Situation einer linearen Regression zurückführen lassen. Die Kenntnis dieser Kunstgriffe erweitert somit den Spielraum erheblich, den ein einfaches Regressionprogramm in der statistische Datenanalyse bietet. Zusätzlich zu diesem rein praktischen Vorteil kann das Nachvollziehen bestehender Parallelen zur linearen Regression auf der Ebene der theoretischen Modellbildung die Einsicht in die Bedeutung der Modellparameter in der jeweiligen Anwendung fördern, und trägt damit zu einem besseren Verständnis der Ergebnisse und zu deren Interpretation bei.

Das vorliegende Buch richtet sich sowohl an angewandte Statistiker als auch an Leser aus anderen Fachgebieten, die mit der Anwendung von Statistik konfrontiert werden. Grundlegende Kenntnisse statistischer Begriffe, etwa im Umfang einer ein- bis zweisemestrigen Einführungsvorlesung, werden vorausgesetzt. Auch eine gewisse Vertrautheit mit der angewandten Regressionsrechnung ist von Vorteil. Ein solides Verständnis der knappen Einführung, die im Abschnitt 1.1 gegeben wird, stellt diesbezüglich ein Minimalwissen dar. Da der Schwerpunkt bei der Anwendung der vorgestellten Verfahren liegt, haben wir auf detaillierte Herleitungen und Beweise verzichtet, hingegen wurde Wert gelegt auf die klare und vollständige Wiedergabe der jeweils unterstellten statistischen Modelle.

Das Buch gliedert sich in drei Kapitel, die je einen Hauptthemenbereich behandeln. Im ersten Kapitel werden die einfache sowie die multiple lineare Regression eingeführt und ihre Handhabung bei verschiedenen Problemlagen aus dem Alltag der statistischen Praxis, wie dem

Vergleich mehrerer Regressionsgeraden oder dem Umgang mit Ausreissern, beschrieben. Im zweiten Kapitel wenden wir uns einer Auswahl von Verfahren aus der multivariaten Statistik zu. Nach einem Abschnitt, der wichtige Eigenschaften von Linearkombinationen einführt, werden Diskriminanzanalyse, Identifikationsanalyse, Spezifikationsanalyse sowie Hauptkomponentenanalyse besprochen. Im dritten Kapitel wird die Planung und Auswertung von Versuchen in industrieller Entwicklung, Landwirtschaft, Qualitätskontrolle oder anderen Anwendungsgebieten erläutert. Hier wird nicht nur gezeigt, wie die aus einem Versuch gewonnenen Daten ausgewertet werden können, sondern es wird besonderer Wert auf das Vorgehen bei der Versuchsplanung gelegt, denn bei diesem Schritt wird noch vor dem eigentlichen Mess- oder Erhebungsvorgang der Grundstein gelegt zum Erreichen einer optimalen Aussagekraft der Resultate unter Berücksichtigung der zur Verfügung stehenden Ressourcen.

In jedem Abschnitt wird die Anwendung des vorgestellten Verfahrens anhand eines konkreten Datenbeispiels aus der statistischen Literatur demonstriert. Die Darstellung der Auswertung ist mit (fiktiven) Computeroutputs, in denen die üblicherweise von Regressionsprogrammen berechneten Grössen aufgeführt sind, und mit Abbildungen illustriert. Mit Ausnahme der Hauptkomponentenanalyse (Abschnitt 2.5) können sämtliche Berechnungen mit einem Programm der multiplen lineraren Regression bewältigt werden. Der Leser erhält so die Möglichkeit, die Resultate am Computer selber nachzuvollziehen und allfällige Alternativen zu den gegebenen Lösungsvorschlägen zu finden. Am Ende des Buches sind eine Auswahl von bewährten Lehrbüchern in englischer und deutscher Sprache sowie die Quellennachweise der Beispielsdatensätze gegeben.

Dieses Buch ist aus Skripten zu verschiedenen Vorlesungen und Kursen entstanden. Zahlreiche Kursteilnehmer und Studenten haben durch Korrekturen und Anregungen zu diesem Werk beigetragen. Ihnen sei an dieser Stelle unser bester Dank ausgesprochen.

September 2000

Mathias Ambühl
Hans Riedwyl

Kapitel 1

Lineare Regression und Verwandtes

1.1 Regressionsgerade mit einer Einflussgrösse und einer Zielgrösse

1.1.1 Problemstellung

Wir betrachten eine Variable x mit festen Werten und eine zufällige Variable Y, von denen vermutet wird, dass x einen Einfluss auf Y ausüben könnte. Ausgehend von der Annahme eines linearen Zusammenhangs soll nun dieser Einfluss näher untersucht werden. x heisst *Einflussgrösse* oder *unabhängige Variable*, Y heisst *Zielgrösse* oder *abhängige Variable*. Die Werte der Einflussgrösse x werden in vielen Fällen vom Untersucher bestimmt.

1.1.2 Ein Zahlenbeispiel

Heute findet man bekanntlich in zahlreichen Zeitschriften seitenweise Kontaktinserate, wo immer wieder bestimmte Wünsche im Bezug auf Charakter, Bildung oder Interessen der gesuchten Person angeführt werden. Häufig sind auch Angaben über das gewünschte Alter des oder der Zukünftigen. Uns interessiert der Zusammenhang zwischen dem Alter der Person, von der ein Inserat stammt, und dem gewünschten Alter des Partners. Da wir davon ausgehen, dass das Wunschverhalten für Frauen und Männer nicht übereinstimmt, wurden nur Inserate von Inserentinnen berücksichtigt. Tabelle 1.1 gibt für 94 Inserentinnen das eigene Alter x_i und das bevorzugte Alter y_i des Gesuchten an.

1.1.3 Modell und Hypothesen

Der verwendete Modellansatz lautet folgendermassen:

Zu einem gegebenen Wert x der Einflussgrösse ist die Zielgrösse Y normalverteilt mit einem Mittelwert von $\mu_Y(x) = \alpha + \beta x$ und einer von x unabhängigen Standardabweichung von σ:

$$Y|x \sim N(\alpha + \beta x,\, \sigma)$$

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
20.5	26	26.5	28.5	31.5	34	40.5	45
21.5	32.5	26.5	28	31.5	37.5	41.5	47
22.5	27.5	26.5	33.5	31.5	35	41.5	40
22.5	27.5	27.5	31.5	32.5	32.5	42.5	46.5
22.5	28.5	27.5	32.5	32.5	35	42.5	42
22.5	24.5	27.5	31.5	32.5	34	42.5	50
22.5	26	27.5	32.5	32.5	35	43.5	52.5
22.5	26	27.5	32.5	34.5	37.5	43.5	52.5
22.5	30	28.5	32.5	34.5	40	43.5	44.5
23.5	27	28.5	28.5	35.5	37.5	43.5	45
23.5	26.5	28.5	30.5	36.5	41	45.5	57.5
23.5	26	28.5	32.5	36.5	45	45.5	47.5
23.5	30	28.5	34	36.5	40	45.5	47.5
23.5	27.5	28.5	34	36.5	39	46.5	50
23.5	29	29.5	32.5	36.5	42.5	47.5	52.5
24.5	27.5	29.5	35	37.5	42.5	47.5	48.5
24.5	27.5	29.5	32.5	38.5	50	48.5	50
24.5	29.5	30.5	37.5	38.5	41.5	50.5	60
25.5	30.5	30.5	37	38.5	41.5	50.5	55
25.5	33	30.5	42.5	38.5	40	51.5	55
25.5	27.5	30.5	34	39.5	44	55.5	60.5
25.5	31.5	30.5	32.5	39.5	41	60.5	63
25.5	36	30.5	32	40.5	44	62.5	62.5
26.5	32.5	31.5	31				

Tabelle 1.1: Alter x von Inserentinnen und Wunschalter y des Partners.

Dies lässt sich anders schreiben als

$$Y_i = \alpha + \beta x_i + E_i \ , (i = 1, \ldots, n), \tag{1.1}$$

wobei die E_i *unabhängig und identisch verteilte Zufallsgrössen* (englisch *indepentently identically distributed*, kurz i.i.d.) sind mit Verteilungsgesetz

$$E_i \sim N(0, \sigma)$$

und n den Stichprobenumfang beschreibt. Die E_i heissen **Residuen**. Sie messen die Abweichung der Y_i von ihrem Mittelwert.

Die durch die Gleichung $y = \alpha + \beta x$ definierte Gerade heisst **Regressionsgerade**. Der Parameter α gibt also an, wo die Regressionsgerade die y-Achse kreuzt und wird als Nullpunktordinate bezeichnet. β ist der Steigungsparameter der Regressionsgeraden. Er sagt uns, um wieviele Einheiten der erwartete Wert der Zielgrösse zunimmt, wenn die Einflussgrösse

um eine Einheit erhöht wird. σ schliesslich ist die Standardabweichung der Residuen, d.h. sie beurteilt, wie weit die Punkte (x_i, y_i) um die Regressionsgerade herum verstreut sind. Die Parameter α, β und σ sind nicht bekannt, sollen also geschätzt werden.

Wir werden die folgenden Testsituationen untersuchen:

a) $H_0: \ \alpha = \alpha_0 \quad$ gegen $\quad H_1: \ \alpha \neq \alpha_0$

b) $H_0: \ \beta = \beta_0 \quad$ gegen $\quad H_1: \ \beta \neq \beta_0$

mit bekannten Parameterwerten α_0 und β_0. Wir betrachten also nur die Fälle, wo die Alternativhypothese zweiseitig ist. Besonders häufig von Interesse sind diese Hypothesen mit $\alpha_0 = 0$ beziehungsweise $\beta_0 = 0$.

1.1.4 Lösungsansatz

Parameterschätzungen

Die Modellparameter α, β und σ werden anhand der Methode der kleinsten Quadrate geschätzt: Wir suchen diejenigen Werte von a und b, für welche die resultierende Summe der quadrierten Abweichungen,

$$S(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 ,$$

minimal wird. Dieser Ansatz liefert die folgenden Parameterschätzungen:

- für β:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} ,$$

- für α:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} ,$$

- für σ:

$$\hat{\sigma} = \sqrt{\frac{S_{Min}}{n-2}} .$$

Dabei wurde von den Bezeichnungen

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \ ,$$

$$S_{yy} = \sum_{I=1}^{n}(y_i - \bar{y})^2 = \sum_{I=1}^{n} y_i^2 - n\bar{y}^2$$

und

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

mit

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{und} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

sowie

$$S_{Min} = S_{yy} - (S_{xy})^2 / S_{xx}$$

Gebrauch gemacht, die auch in späteren Kapiteln verwendet werden.

Hat man die Parameterschätzungen $\hat{\alpha}$ und $\hat{\beta}$ einmal berechnet, so ergeben sich daraus die aus der Einflussgrösse geschätzten Werte der Zielgrösse,

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \quad (i = 1, \ldots, n) ,$$

und die beobachteten Werte der Residuen,

$$\hat{e}_i = y_i - \hat{y}_i \quad (i = 1, \ldots, n) .$$

Somit werden die beobachteten Werte von Y gemäss

$$y_i = \hat{y}_i + \hat{e}_i$$

additiv zerlegt in ein bei gegebenem $x = x_i$ zu erwartendes $\hat{y}_i$ und die Abweichung $\hat{e}_i$ davon.

Analyse der Residuen

Im Hinblick auf die Beurteilung der gemachten Voraussetzungen müssen wir stets prüfen, ob die Residuen annähernd als Realisierung einer unabhängig und identisch normalverteilten Stichprobe gelten können. Ein Histogramm der Residuen ist ein mögliches Hilfsmittel zur Visualisierung dieser Annahme. Wichtig ist auch ein Punktediagramm der Wertepaare $(\hat{y}_i, \hat{e}_i)$, also der Residuen gegen die Schätzwerte, welches oft Aufschluss darüber bringt, ob das zugrundeliegende lineare Modell korrekt ist. Bei erfüllten Voraussetzungen sollten sich die Residuen im ganzen Bereich von $\hat{y}$ ohne erkennbare Struktur um 0 scharen. Nichtkonstante Streuung der Residuen, nichtlineare Abhängigkeit und weitere Verletzungen der Modellannahmen können durch diese Zeichnung aufgedeckt werden.
Sind die Modellvoraussetzungen nicht erfüllt, so haben die hier vorgestellten Schätz- und Testverfahren keine Gültigkeit. Es muss dann nach einem geeigneteren Modellansatz gesucht werden.

Tests

Für die beiden Hypothesen a) und b) verwenden wir F-Tests, die auf einem Vergleich der Fehlerquadratsumme im allgemeinen Modell (1.1), S_{Min}, mit derjenigen im durch die zu testende Nullhypothese beschränkten Modell, S^0_{Min}, beruhen. Die Testgrösse lautet

$$F(H_1) = \frac{S^0_{Min} - S_{Min}}{(S_{Min})/(n-2)} \tag{1.2}$$

und folgt bei Gültigkeit der erwähnten Voraussetzungen einer F-Verteilung mit einem Freiheitsgrad im Zähler und $n-2$ Freiheitsgraden im Nenner. Der Wert der Statistik $F(H_1)$ muss also bei Zulassung einer Fehlerwahrscheinlichkeit erster Art von α (nicht zu verwechseln mit dem Parameter α!) mit dem $(1-\alpha)$-Quantil aus der betreffenden F-Verteilung verglichen werden: H_0 wird verworfen, sobald $F(H_1)$ einen Wert annimmt, der grösser ist als dieses Quantil. Grosse Werte von $F(H_1)$ weisen also auf die Richtigkeit der Alternativhypothese H_1 hin. Falls nicht ausdrücklich anders erwähnt, wird im Folgenden immer eine Fehlerwahrscheinlichkeit erster Art (Signifikanzniveau) von $\alpha = 5\%$ eingeräumt.

Die Fehlerquadratsumme im vollen Modell (1.1), auch minimales Summenquadrat genannt, lautet

$$S_{Min} = S_{yy} - (S_{xy})^2 / S_{xx}, \tag{1.3}$$

während die Fehlerquadratsumme S^0_{Min} im beschränkten Modell, d.h. unter der Voraussetzung, dass die Nullhypothese erfüllt ist, davon abhängt, welche Nullhypothese wir betrachten.

Hypothesen $\alpha = 0$ und $\beta = 0$

Die Fehlerquadratsummen bei Gültigkeit der Hypothesen $H_0 : \alpha = 0$ beziehungsweise $H_0 : \beta = 0$ lauten:

a) im Fall $H_0 : \alpha = 0$:

$$S^0_{Min} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^n x_i y_i\right)^2}{\sum_{i=1}^n x_i^2}$$

b) im Fall $H_0 : \beta = 0$:

$$S^0_{Min} = S_{yy}$$

In beiden Fällen ist die Ungleichung

$$S^0_{Min} \geq S_{Min}$$

immer erfüllt, so dass die F-Statistik (1.2) nur nichtnegative Werte annimmt.

Hypothesen $\alpha = \alpha_0$ und $\beta = \beta_0$

Die Tests für diese Nullhypothesen lassen sich durch einen einfachen Trick auf die obigen Spezialfälle $\alpha_0 = 0$ resp. $\beta_0 = 0$ in Modellen mit einer modifizierten Zielgrösse zurückführen:

a) Fall $H_0: \alpha = \alpha_0$:

Wir subtrahieren in der Modellgleichung (1.1)

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

auf beiden Seiten α_0 und erhalten

$$Y_i - \alpha_0 = (\alpha - \alpha_0) + \beta x_i + E_i = \alpha' + \beta x_i + E_i$$

mit $\alpha' = \alpha - \alpha_0$. Daraus wird ersichtlich, dass das Testen der Hypothese $H_0: \alpha = \alpha_0$ im Modell (1.1) äquivalent ist zu einem Test der Hypothese

$$H_0: \alpha' = 0$$

im modifizierten Modell

$$Y_i' = \alpha' + \beta x_i + E_i \tag{1.4}$$

mit der Zielgrösse $Y_i' = Y_i - \alpha_0$. Somit lässt sich die Fehlerquadratsumme berechnen als

$$\begin{aligned} S^0_{Min} &= \sum_{i=1}^{n} (y_i')^2 - \frac{\left(\sum_{i=1}^n x_i y_i'\right)^2}{\sum_{i=1}^n x_i^2} \\ &= \sum_{i=1}^{n} (y_i - \alpha_0)^2 - \frac{\left(\sum_{i=1}^n x_i (y_i - \alpha_0)\right)^2}{\sum_{i=1}^n x_i^2}\,, \end{aligned}$$

wo $y_i' = y_i - \alpha_0$ die Werte der transformierten Zielgrösse sind.

b) Fall $H_0: \beta = \beta_0$:

Analog zu a) subtrahieren wir im Regressionsmodell (1.1) auf beiden Seiten $\beta_0 x_i$ und erhalten

$$Y_i - \beta_0 x_i = \alpha + (\beta - \beta_0) x_i + E_i = \alpha + \beta' x_i + E_i$$

mit $\beta' = \beta - \beta_0$. So wird ersichtlich, dass die gesuchte Testgrösse konstruiert werden kann als Testgrösse für die Nullhypothese

$$H_0: \beta' = 0$$

im Modell

$$Y_i' = \alpha + \beta' x_i + E_i\,, \tag{1.5}$$

wo die Zielgrösse die Werte $y'_i = y_i - \beta_0 x_i$ besitzt und als Steigungsparameter $\beta' = \beta - \beta_0$ gesetzt wird. Die reduzierte Fehlerquadratsumme beträgt somit

$$\begin{aligned} S^0_{Min} &= S_{y'y'} = \sum_{i=1}^{n} (y'_i - \bar{y}')^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 x_i - (\bar{y} - \beta_0 \bar{x}))^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 x_i)^2 - n\,(\bar{y} - \beta_0 \bar{x})^2 \quad . \end{aligned}$$

Für Computer-Anwender besteht also die Möglichkeit, auch mit einem Programm, das nur Tests für die Nullhypothesen $\alpha = 0$ und $\beta = 0$ liefert, die Testgrössen für die allgemeineren Tests $\alpha = \alpha_0$ oder $\beta = \beta_0$ durch den Computer rechnen zu lassen. Der Aufwand dafür beschränkt sich auf eine einfache Datenmanipulation, nämlich die Einführung der jeweiligen Variablen Y'.

Bemerkung zu den obigen F-Tests: In zahlreichen Lehrbüchern und Statistik-Softwarepaketen figuriert anstelle des F-Tests (1.2) ein zweiseitiger t-Test mit $n-2$ Freiheitsgraden, der sich für die Praxis als äquivalent herausstellt. Zwischen den beiden Testgrössen gilt die Relation

$$t^2 = F(\mathit{Parameter} \neq \mathit{hypothetischer\ Wert})$$

und für die Quantile der beiden verwendeten Verteilungen gilt

$$t^2_{1-\frac{\alpha}{2}}(n-2) = F_{1-\alpha}(1, n-2) \quad .$$

Daraus folgt

$$|\,t\,| > t_{1-\frac{\alpha}{2}}(n-2) \iff F(\mathit{Parameter} \neq \mathit{hypothetischer\ Wert}) > F_{1-\alpha}(1, n-2),$$

und da der F-Test einseitig, der t-Test jedoch zweiseitig erfolgt, resultiert aus beiden Tests die gleiche Entscheidung.

Standardabweichungen der Parameterschätzungen

Die Parameterschätzungen $\hat{\alpha}$ und $\hat{\beta}$ sind als Funktionen von verschiedenen Zufallsvariablen selbst wieder Zufallsvariablen. Sie sind *erwartungstreu*, d.h. ihr Erwartungswert entspricht dem jeweiligen wahren Parameterwert:

$$E(\hat{\alpha}) = \alpha \quad \text{und} \quad E\left(\hat{\beta}\right) = \beta \quad .$$

Ihre Standardabweichungen lassen sich aus dem Schätzwert und der F-Teststatistik für die Nullhypothese

$$H_0: \quad \mathit{Parameter} = 0$$

mittels der Gleichung

$$\text{Standardabweichung der Schätzung} = \frac{|Schätzwert|}{\sqrt{F(Parameter \neq 0)}} \tag{1.6}$$

bestimmen. Ausdrücklich kann also die Standardabweichung der Parameter $\hat{\alpha}$ und $\hat{\beta}$ geschätzt werden als

$$\text{S.A.}(\hat{\alpha}) = \frac{|\hat{\alpha}|}{\sqrt{F(\alpha \neq 0)}} \quad \text{bzw.} \quad \text{S.A.}\left(\hat{\beta}\right) = \frac{\left|\hat{\beta}\right|}{\sqrt{F(\beta \neq 0)}} \ ,$$

wobei $\hat{\alpha}$ bzw. $\hat{\beta}$ auf der linken Seite der Gleichung als Zufallsvariable, auf der rechten jedoch als deren Realisation aufzufassen ist.
Die Standardabweichung der bedingten Parameterschätzung des Parameters α nach der Annahme der Hypothese $\beta = \beta_0$ kann einfacher als mit Formel (1.6) berechnet werden. Die Formel für diesen Fall wird im folgenden Abschnitt angegeben.

Bedingte Parameterschätzung nach Annahme einer Nullhypothese

Falls aus einem Test die Annahme einer Nullhypothese H_0 resultiert, so muss dieses Ergebnis in der Schätzung der Modellgleichung mitberücksichtigt werden, d.h. die Parameter müssen neu geschätzt werden unter der Nebenbedingung, dass H_0 gilt.
Wir nehmen hier die gleiche Fallunterscheidung vor wie im vorangehenden Abschnitt:

1. a) Fall $H_0 : \alpha = 0$:
   Das Modell (1.1) der einfachen linearen Regression lässt sich nach Annahme der Hypothese $\alpha = 0$ vereinfachen zu

   $$Y_i = \beta x_i + E_i \ , (i = 1, \ldots, n).$$

   Dieses Modell beschreibt eine proportionale Beziehung zwischen der Einflussgrösse X und der Zielgrösse Y. Man spricht hier von einer *Regressionsgeraden ohne Nullpunktordinate* (englisch *intercept*), oder auch von einer *Regression durch den Koordinatenursprung.* Der Parameter β wird mit der Methode der kleinsten Quadrate geschätzt mit

   $$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \ .$$

   b) Fall $H_0 : \beta = 0$:
   In diesem Fall lautet das vereinfachte Modell, in dem $\beta = 0$ gesetzt wird

   $$Y_i = \alpha + E_i \ , (i = 1, \ldots, n).$$

Die Minimum-Quadrat-Schätzung von α entspricht hier dem arithmetischen Mittelwert der y_i:

$$\hat{\alpha} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ .$$

Die Standardabweichung von $\hat{\alpha}$ erhält man einfacher als mit Formel (1.6) aus der empirischen Standardabweichung s_y der y_i als

$$\text{S.A.}\,(\hat{\alpha}) = \frac{s_y}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2} \ .$$

2. a) Fall $H_0 : \alpha = \alpha_0$:
Wie weiter oben gezeigt wurde, entspricht die Annahme der Hypothese $\alpha = \alpha_0$ im Regressionsmodell (1.1) der Annahme der Hypothese $\alpha' = 0$ im abgeänderten Modell (1.4):

$$Y_i' = \alpha' + \beta x_i + E_i \quad \text{mit} \quad Y_i' = Y_i - \alpha_0 \quad \text{und} \quad \alpha' = \alpha - \alpha_0 \ .$$

Der Parameter β kann somit im Modell (1.4) wie unter 1.a) geschätzt werden, also mit der Formel

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i'}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i (y_i - \alpha_0)}{\sum_{i=1}^{n} x_i^2} \ .$$

Dies ist auch die Minimum-Quadrat-Schätzung von β im Modell (1.1), unter der Bedingung, dass $\alpha = \alpha_0$ gilt. Der Wert des Parameters α muss nicht geschätzt werden, da der Test gezeigt hat, dass er gleich α_0 gesetzt werden darf.

b) Fall $H_0 : \beta = \beta_0$:
Diese Hypothese ist äquivalent zur Hypothese $\beta' = 0$ im modifizierten Modell (1.5):

$$Y_i' = \alpha + \beta' x_i + E_i \quad \text{mit} \quad Y_i' = Y_i - \beta_0 x_i \quad \text{und} \quad \beta' = \beta - \beta_0 \ .$$

Wir können also unter Berücksichtigung von $\beta' = 0$ den Parameter α im Modell (1.5) und seine Standardabweichung mit den Formeln aus 1.b) schätzen:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i' = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 x_i) = \bar{y} - \beta_0 \bar{x} \quad \text{und}$$

$$\text{S.A.}\,(\hat{\alpha}) = \frac{s_{y'}}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \beta_0 x_i - \bar{y} + \beta_0 \bar{x})^2} \ .$$

Diese Formeln gelten auch für die Schätzung von α im Modell (1.1) gegeben $\beta = \beta_0$.

Die bedingten Schätzungen der Modellparameter können somit auch mit einem Computerprogramm ermittelt werden, das nur über eine Routine zur Berechnung der (nicht bedingten) Regressionsgeraden verfügt. Dazu muss der Weg über die modifizierten Modelle (1.4) beziehungsweise (1.5) begangen werden. Für die Fälle 1.a) und 2.a) muss die Möglichkeit der Berechnung einer Regressionsgeraden durch den Koordinatenursprung gegeben sein, in den beiden anderen Fällen besteht die Parameterschätzung in einer Mittelwertberechnung.

Bestimmtheitsmass und Korrelationskoeffizient

Aus der Formel für das minimale Summenquadrat (1.3),

$$S_{Min} = S_{yy} - (S_{xy})^2 / S_{xx} \ ,$$

erkennt man, dass das totale Summenquadrat in den y_i, S_{yy}, aufgeteilt wird in einen durch die Regression erklärten Teil $(S_{xy})^2/S_{xx}$ und einen nicht erklärten Teil, der dem minimalen Summenquadrat S_{Min} entspricht. Man bezeichnet den Anteil von S_{yy}, den die Regression zu erklären vermag, als **Bestimmtheitsmass** R^2:

$$R^2 = \frac{S_{yy} - S_{Min}}{S_{yy}} \ . \tag{1.7}$$

Das Bestimmtheitsmass entspricht gleichzeitig dem Varianzanteil der Zielgrösse, der durch die Einflussgrösse erklärt wird. Es liegt immer zwischen 0 und 1 und beschreibt die Güte des linearen Zusammenhangs zwischen Einfluss- und Zielgrösse. R^2 ist identisch mit dem Quadrat des empirischen Korrelationskoeffizienten zwischen den Zufallsgrössen Y und $\hat{Y}$:

$$R^2 = r^2_{y\hat{y}} \ .$$

Im vorliegenden Fall einer einfachen linearen Regression entspricht das Bestimmtheitsmass gleichzeitig dem Quadrat des Korrelationskoeffizienten zwischen den x_i und den y_i, d.h. $R^2 = r^2_{xy}$.

Achtung: In einem Regressionsmodell ohne Nullpunktordinate wird das Bestimmtheitsmass nicht gemäss Formel (1.7) berechnet, sondern als

$$R^2 = \frac{\sum_{i=1}^n y_i^2 - S_{Min}}{\sum_{i=1}^n y_i^2} \ ,$$

d.h. es gibt den durch die Einflussgrösse erklärten Anteil der quadratischen Variabilität in den y_i ohne Korrektur bezüglich des Mittelwertes $\bar{y}$ an. Damit ist das Bestimmtheitsmass eines Modells nach Nullsetzen des Parameters α nur bedingt mit demjenigen aus einem Modell mit Nullpunktordinate vergleichbar.

Regression	
Zielgrösse	WUNSCH
Bestimmtheitsmass	0.9152
Anzahl Beobachtungen	94

	Koeff.	S.A.(Koeff.)	F-partial	p-Wert
Nullpunktordinate	5.4749	1.0682	26.2715	0.0000
Steigung(ALTER)	0.96361	0.03057	993.4689	0.0000

ANOVA

	F.G.	Summenquadrat	F-global	p-Wert
Regression	1	7661.0329	993.4689	0.0000
Residuen	92	709.4485		
Insgesamt	93	8370.4814		

Tabelle 1.2: Computer-Output zum Beispiel Wunschalter des Partners.

1.1.5 Lösungsvorschlag zum Beispiel

In unserem Beispiel über Kontaktinserate wollen wir die Regressionsgerade für $Y =$ *Wunschalter des Partners* in Abhängigkeit von $x =$ *Alter der Inserentin* berechnen. Anschliessend soll die Verträglichkeit des vorliegenden Datensatzes mit den Hypothesen

1. $H_0: \ \alpha = 0$
2. $H_0: \ \beta = 0$
3. $H_0: \ \beta = 1$

getestet werden.
Das Beispiel ist im Punkteschwarm in Abbildung 1.1 illustriert und die Resultate sind in den Tabellen 1.2 und 1.3 zusammengestellt. Die in diesen Tabellen angegebenen Resultate bilden den Kern einer Regressionsanalyse und sind auf ähnliche Weise in den Outputs der gebräuchlichsten Statistik-Softwarepakete zu finden.

Schätzung der Regressionsgeraden

Mit den Formeln aus Abschnitt 1.1.4 erhält man die geschätzte Regressionsgerade

$$\mu_Y(x) = \underset{(1.07)}{5.47} \ + \underset{(0.0306)}{0.9636}\, x \ .$$

In Klammern unter den Parameterschätzungen stehen ihre geschätzten Standardabweichungen. Als Rundungsregel empfehlen wir, den Wert der Standardabweichungen der Parameterschätzungen auf drei gültige Stellen genau anzugeben und die Schätzwerte auf ebenso viele

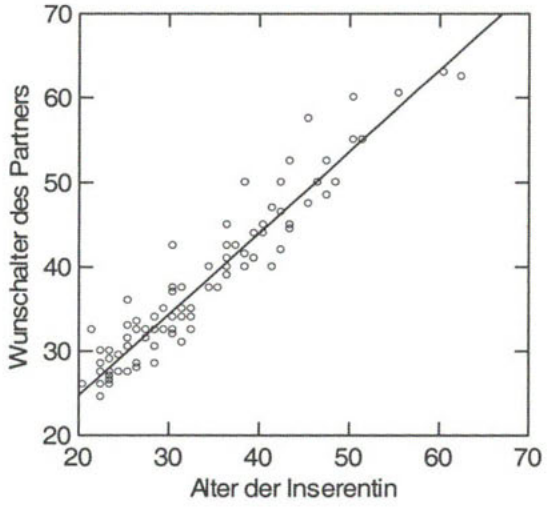

Abbildung 1.1: Punkteschwarm und Regressionsgerade im Beispiel über Wunschalter des Partners.

Dezimalen zu runden. Die Standardabweichung der Residuen wird auf

$$\hat{\sigma} = \sqrt{\frac{S_{Min}}{n-2}} = \sqrt{\frac{709.4485}{92}} = 2.777$$

geschätzt. Das Bestimmtheitsmass beträgt

$$R^2 = 0.915 \ ,$$

was den aus dem Bild ersichtlichen starken linearen Zusammenhang wiedergibt.

Tests

1. $H_0: \ \alpha = 0$: Unter dieser Nullhypothese ist ein Modell der Form

   $$\mu_Y(x) = \beta x$$

   zulässig, d.h. das gewünschte Alter des Partners lässt sich als ein konstantes Vielfaches des Alters der Inserentin selbst beschreiben.
   Der Wert der Testgrösse,

   $$F(\alpha \neq 0) = 26.271,$$

   liegt über dem 95%-Quantil der F-Verteilung mit einem und $n-2 = 92$ Freiheitsgraden von $F_{0.95}(1, 92) = 3.95$, d.h. H_0 wird verworfen und eine Modellvereinfachung dieser Form ist nicht geeignet.

Regression	
Zielgrösse	DIFFERENZ
Bestimmtheitsmass	0.123
Anzahl Beobachtungen	94

	Koeff.	S.A.(Koeff.)	F-partial	p-Wert
Nullpunktordinate	5.474914	1.068157	26.2715	0.0000
Steigung(ALTER)	-0.036391	0.030572	1.41693	0.23697

ANOVA

	F.G.	Summenquadrat	F-global	p-Wert
Regression	1	10.926477	1.41693	0.23697
Residuen	92	709.448523		
Insgesamt	93			

Tabelle 1.3: Computer-Output zum Beispiel mit der Variablen DIFFERENZ (= WUNSCH−ALTER) als Zielgrösse.

2. $H_0: \beta = 0$: Falls diese Nullhypothese zutrifft, reduziert sich das Modell (1.1) zu

$$\mu_Y(x) = \alpha,$$

was bedeuten würde, dass das Wunschalter des Partners bei den Inserentinnen aller Altersklassen identisch wäre. Wegen

$$F(\beta \neq 0) = 993.49 > F_{0.95}(1, 92) = 3.95$$

wird auch diese Hypothese abgelehnt.

3. $H_0: \beta = 1$: Diese Hypothese unterstellt, dass das gewünschte Alter des Partners um eine feste Anzahl Jahre vom Alter der Inserentin abweicht, d.h. dass die von der Inserentin gewünschte Altersdifferenz nicht von ihrem Alter abhängt.
Die geeignete Teststatistik ist dieselbe, die zur Überprüfung der Hypothese

$$H_0 : \beta' = \beta - 1 = 0$$

im Modell

$$\mu_{Y'}(x) = \alpha + \beta' x$$

mit der modifizierten Variablen

$$Y' = Y - x = \text{Wunschalter des Partners} - \text{Alter der Inserentin}$$

ermittelt würde. Es gilt (vgl. Tabelle 1.3)

$$F(\beta \neq 1) = F(\beta' \neq 0) = 1.42 < F_{0.95}(1, 92) = 3.95,$$

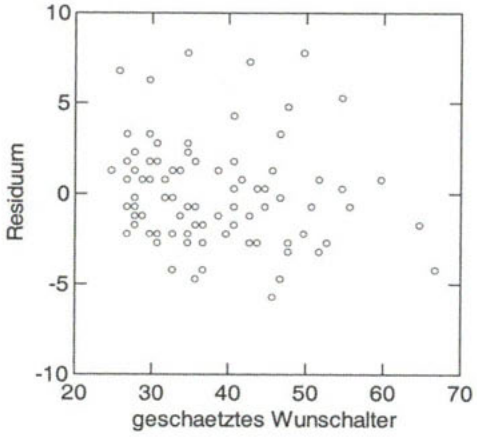

Abbildung 1.2: Residuenplot der Punkte $(\hat{y}_i, \hat{e}_i)$ im Beispiel über Wunschalter des Partners.

somit besteht kein Anlass, diese Hypothese zu verwerfen.
Nun muss der Parameter α neu geschätzt werden unter der Bedingung $\beta = 1$. Man findet das Modell

$$\mu_{\text{Wunschalter des Partners}} = \underset{(0.287)}{4.250} + \text{Alter der Inserentin} \ ,$$

was bedeutet, dass eine 'typische' Inserentin unabhängig von ihrem Alter einen um gut vier Jahre älteren Partner sucht.

Zur Überprüfung der Modellvoraussetzungen (für das Modell mit $\beta = 1$) ist in Abbildung 1.2 ein Punkteschwarm der Residuen $\hat{e}_i$ in Abhängigkeit der Schätzwerte $\hat{y}_i$ gezeichnet. Die Verteilung der Residuen erscheint leicht (rechts-) schief, also nicht normalverteilt. Diesem Umstand können wir Rechnung tragen, indem wir die Regressionsgerade so wählen, dass nicht das arithmetische Mittel der Residuen, sondern ihr Zentralwert den Wert 0 annimmt. Die Bestimmung der Parameter dieser Geraden geschieht folgendermassen: Der Steigungsparameter β wird normal mit der Methode der kleinsten Quadrate festgelegt. Um den Achsenabschnitt α zu bestimmen, nimmt man dann nicht den Mittelwert der $y_i - \beta x_i$, sondern deren Median. Die so definierte Gerade besitzt die oben erwähnte Eigenschaft.
In unserem Beispiel wählen wir das vereinfachte Modell mit Steigungsparameter $\beta = 1$ als Ausgangspunkt. So finden wir als Achsenabschnitt den Wert von $\alpha = 4$, was uns zu folgender Schätzungsgleichung führt:

$$\textit{Wunschalter des Partners} = 4 + \textit{Alter der Inserentin} \ .$$

Diese Gerade verläuft etwas unterhalb der klassischen Regressionsgeraden.