6 Ergänzungen

6.1 Fehlerbehaftete erklärende Variable

a Die erklärenden Variablen erscheinen in den besprochenen Modellen nicht als Zufallsvariable, obwohl sie oft ebenso zufällig sind wie die Zielgrösse. Wir haben dies bisher vernachlässigt und immer so getan, als ob die x-Werte feste, vorgegebene Zahlen seien. Eine formale Begründung dafür besteht darin, dass die Verteilungen gemäss Modell als bedingte Verteilungen, gegeben die x_i-Werte, aufgefasst werden.

b Wir wollen nun untersuchen, was geschieht, wenn die erklärende Variable, deren Einfluss auf die Zielgrösse von Interesse ist, nur ungenau gemessen oder beobachtet werden kann. Wir stellen uns zwei **„latente“ Variable** u und v vor, die deterministisch zusammenhängen – im einfachsten Fall linear,

$$v = \widetilde{\alpha} + \widetilde{\beta} u \; .$$

Sie können aber beide nicht exakt beobachtet werden, sondern nur mit zufälligen Fehlern, also

$$X_i = u_i + D_i \; , \qquad Y_i = v_i + E_i = \widetilde{\alpha} + \widetilde{\beta} u_i + E_i \; .$$

Die Fehler D_i sollen ebenso wie die Messfehler E_i normalverteilt sein,

$$D_i \sim \mathcal{N}\langle 0, \sigma_D^2 \rangle \; , \qquad E_i \sim \mathcal{N}\langle 0, \sigma_E^2 \rangle$$

– und unabhängig. Die u_i und damit auch die v_i seien feste Zahlen – wie es in der linearen Regression die x_i sind. Unser Interesse gilt dem Koeffizienten $\widetilde{\beta}$ und eventuell auch $\widetilde{\alpha}$.

Für $\sigma_D^2 = 0$ wird u_i gleich der beobachtbaren Variablen X_i, und man erhält das Modell der einfachen linearen Regression.

c Das beschriebene Modell ist der einfachste Fall einer Regression mit fehlerbehafteten erklärenden Variablen (**errors-in-variables regression**). Man spricht auch von einer **funktionalen Beziehung** (**functional relationship**). Wenn die wahren Werte u_i der erklärenden Variablen als zufällig statt als fest aufgefasst werden, dann heisst das Modell eine **structural relationship**.

d Den Unterschied zwischen dem Modell der funktionalen Beziehung und der einfachen linearen Regression wird in Abbildung 6.1.d an einem simulierten Beispiel gezeigt.

Vergleicht man die Beobachtungen mit den Punkten, die man erhalten hätte, wenn die erklärende Variable u ohne Messfehler verfügbar wäre, dann sieht man, dass sich die Streuung der Punkte in x-Richtung ausdehnt.

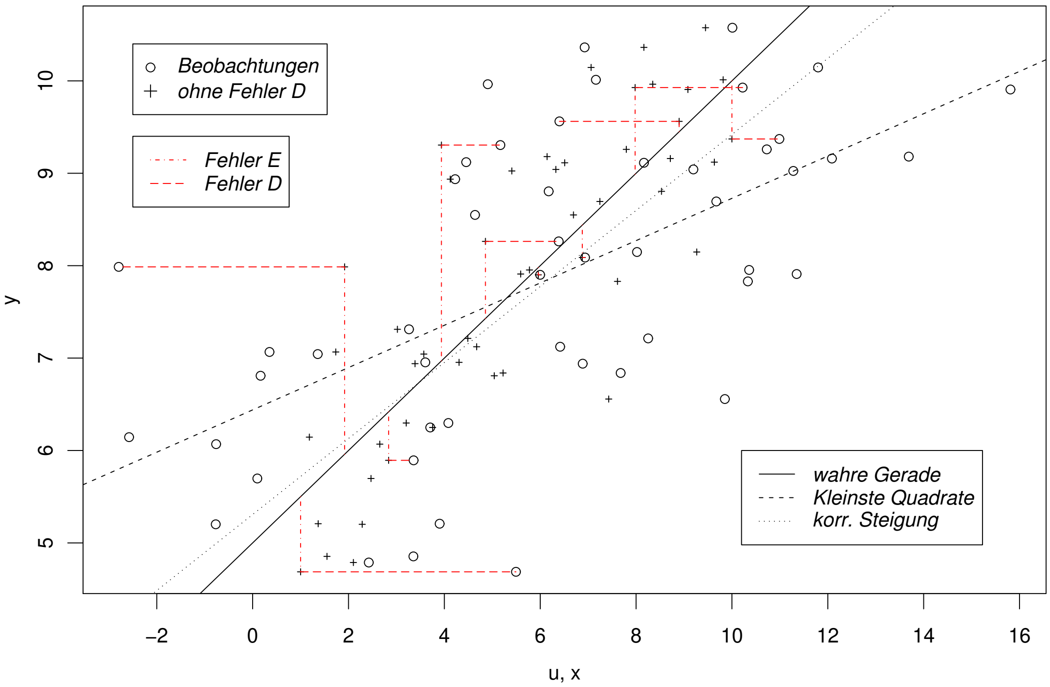

Abbildung 6.1.d: Veranschaulichung des Modells mit einer fehlerbehafteten erklärenden Variablen. 50 Beobachtungen wurden mit dem Modell $v = 5 + 0.5 \cdot u$, $\sigma_D = 3$ und $\sigma_E = 1$ simuliert. Die Beobachtungen ($\circ$) streuen in x-Richtung stärker als die „Beobachtungen ohne Fehler in x-Richtung"($+$), die aus der Simulation hier bekannt sind. Zusätzlich zur „wahren" Geraden sind die mit Kleinsten Quadraten geschätzte und die korrigierte Gerade eingezeichnet.

e Die Steigung der Regressionsgeraden, die mit Kleinsten Quadraten bestimmt wird, ist gleich

$$\widehat{\beta}_{LS} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\widehat{\text{cov}}\langle X, Y\rangle}{\widehat{\text{var}}\langle X\rangle} ,$$

also gleich dem Quotienten aus der (empirischen) Kovarianz zwischen X und Y und der (empirischen) Varianz von X. In Abbildung 6.1.d zeigt sich, dass die geschätzte Gerade viel flacher ist als die wahre. Ist das Zufall?

Um die gewünschte Steigung $\widetilde{\beta}$ zu bestimmen, müssten wir die X_i-Werte durch die u_i ersetzen können. Was würde sich ändern? Da die Zufallsfehler D_i unabhängig sind von den E_i und den u_i und damit auch von den $Y_i = \widetilde{\beta} u_i + E_i$, verändert sich die Kovarianz nicht (genauer: die empirische Kovarianz zwischen U und Y hat den gleichen Erwartungswert wie diejenige zwischen X und Y). Die empirische Varianz der u_i ist dagegen im Erwartungswert um σ_D^2 kleiner als die empirische Varianz der X_i. Deshalb wird der Nenner in der obigen Formel zu gross, während der Zähler den richtigen Erwartungswert hat. Das führt zu einer systematisch zu flachen Geraden.

Der systematische Fehler lässt sich aber leicht korrigieren, wenn σ_D bekannt ist: Wir setzen im Nenner $\widehat{\text{var}}\langle X\rangle - \sigma_D^2$ statt $\widehat{\text{var}}\langle X\rangle$ ein. Anders gesagt,

$$\widehat{\beta} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2 - \sigma_D^2} = \widehat{\beta}_{LS}/\widehat{\kappa}$$
$$\widehat{\kappa} = \frac{\widehat{\text{var}}\langle X\rangle - \sigma_D^2}{\widehat{\text{var}}\langle X\rangle}$$

Die Grösse $\widehat{\kappa}$ schreiben wir mit Hut ($\widehat{\ }$), da sie (über die u_i) von der Stichprobe abhängt. Wenn die „wahren" Werte u_i der erklärenden Variablen selbst als Zufallsvariable modelliert werden, ist der Modellparameter, der durch $\widehat{\kappa}$ geschätzt wird gleich $\kappa = \text{var}\langle U\rangle / \text{var}\langle X\rangle$.

Die Grösse κ wird in der Literatur als „Abschwächungs-Koeffizient" *(attenuation coefficient)* bezeichnet. Er misst, wie viel flacher die mit der üblichen Methode geschätzte Steigung wird als die gesuchte Steigung $\widetilde{\beta}$. Er wird auch *reliability ratio* genannt, da er die „Verlässlichkeit" der Variablen X als Mass für die gewünschte Variable U misst.

f Den zweiten Parameter $\widetilde{\alpha}$, den Achsenabschnitt der gesuchten Geraden, schätzt man wie früher nach der Formel $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{X}$ (2.2.c) – hier natürlich mit der soeben eingeführten erwartungstreuen Schätzung $\widehat{\beta}$.

Bevor wir den Fall diskutieren, in dem σ_D nicht bekannt ist, soll ein Beispiel folgen.

g Im **Beispiel der Schadstoffe im Tunnel** (1.1.d) sollen die Emissionsfaktoren für die beiden Fahrzeugklassen „Personenwagen" und „Lastwagen" bestimmt werden. In der erwähnten Untersuchung im Gubrist-Tunnel konnte die Anzahl Fahrzeuge einer Fahrzeugklasse nicht genau bestimmt werden. Die systematische Abweichung (systematische Unterschätzung des Anteils der Lastwagen am Gesamtverkehr durch die Schlaufen-Klassierung) kann durch „Eichung" (siehe 1.1.f und 6.2 unten) korrigiert werden, aber der Erfassungsfehler wird auch zufällig streuen. Die Daten, die zur Eichung dienen, liefern auch eine Schätzung der Varianz dieser zufälligen Fehler, also von σ_D^2, nämlich 0.0213^2.

Wenn die Schätzung diese zufälligen Fehler nicht berücksichtigt, wird die Gerade zu flach geschätzt, wie wir gesehen haben. Für Schadstoffe, die von den Lastwagen stärker emittiert werden, bewirkt das, dass ihre Emissionen unterschätzt und jene der Personenwagen überschätzt werden – und umgekehrt für Schadstoffe, die von Personenwagen in grösserer Menge ausgestossen werden. Abbildung 6.1.g zeigt die Daten der Studie, die für die Berechnung der Emissionsfaktoren brauchbar waren. In den Nachtstunden herrschte geringer Verkehr, was zu so kleinen Luftgeschwindigkeiten führt, dass die Emissionen nicht mehr richtig berechnet werden konnten. (Die Rechnung setzt laminare Luftströmung voraus.) Die flachere eingezeichnete Gerade resultiert aus einer robusten Schätzung ohne Berücksichtigung der Fehler der erklärenden Variablen; die steilere ist die korrigierte. Der Korrekturfaktor $1/\kappa$ für die Steigung beträgt 1.12. Der Achsenabschnitt, der den Emissionsfaktor für die Personenwagen misst, wird geringfügig von 1254 auf 1169 korrigiert, während der geschätzte Emissionsfaktor für die Lastwagen ($\widehat{\alpha} + \widehat{\beta}$) von 14580 um 10% auf 16065 klettert.

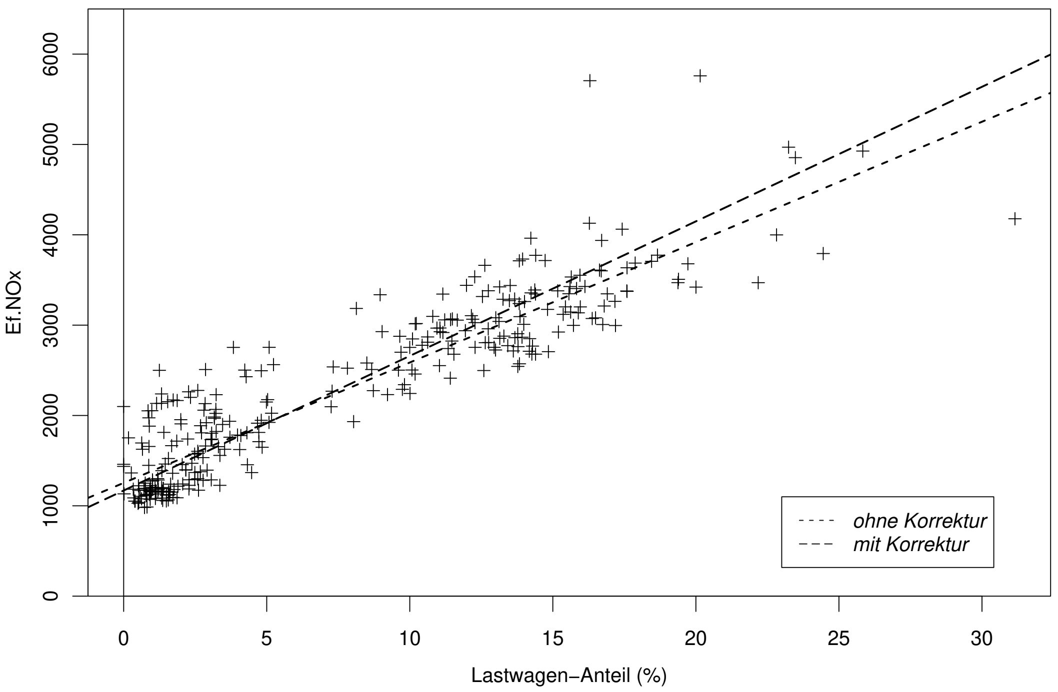

Abbildung 6.1.g: Emissionsfaktor für NO_X und Lastwagen-Anteil im Beispiel der Schadstoffe im Tunnel, für die Zeitabschnitte mit genügender Luftgeschwindigkeit. Die Geraden stellen die Schätzung mit und ohne Berücksichtigung der Messfehler des Lastwagen-Anteils dar.

h Im Umweltbereich gibt es viele ähnliche Fragestellungen, vor allem auch auf dem Gebiet des Zusammenhangs von **Gesundheitsschäden** mit der **Exposition gegenüber Risikostoffen**: Die Schädigungen werden systematisch unterschätzt, wenn die Ungenauigkeit der Erfassung der Exposition nicht berücksichtigt wird.

i Statt der Ungenauigkeit der erklärenden Variablen X kann auch das **Verhältnis** $\gamma = \sigma_E/\sigma_D$ der Ungenauigkeiten von X und Y (näherungsweise) **bekannt** sein. Durch Umskalierung der einen Variablen ($X \to \gamma X$) lässt sich dann erreichen, dass beide gemäss Annahme die gleiche Genauigkeit aufweisen. Dann liefert die orthogonale Regression die richtige Schätzung.

j Die **orthogonale Regression** minimiert statt der Quadratsumme der vertikalen Abweichungen $r_i\langle a,b\rangle$ (Methode der Kleinsten Quadrate) diejenige der orthogonalen Abstände $d_i\langle a,b\rangle$ (Abbildung 6.1.j).

Das ergibt eine steilere Gerade als die Kleinsten Quadrate der r_i. (* Sie fällt mit der ersten **Hauptkomponente** einer Hauptkomponenten-Analyse zusammen – ein Thema der Multivariaten Statistik.)

Wenn die Masseinheit von X oder Y geändert wird, ändert sich die mit orthogonaler Regression bestimmte Gerade in einer Weise, die schwierig interpretierbar ist. (Probieren Sie Extremfälle aus!) Man soll diese Art der Regression daher nur auf geeignet standardisierte Daten anwenden.

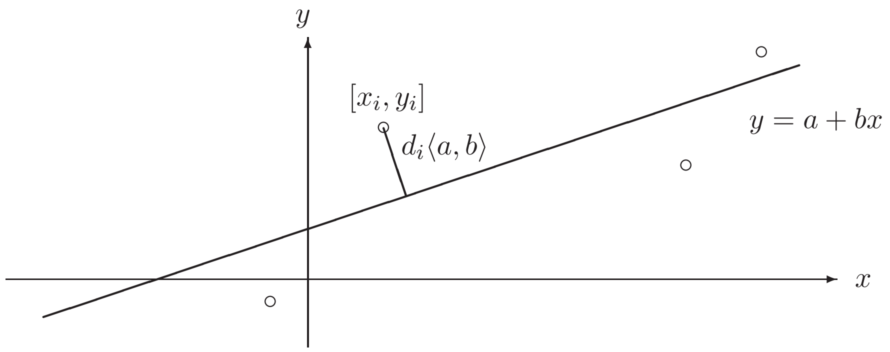

Abbildung 6.1.j: Zur Definition der orthogonalen Regression

Wenn X und Y auf empirische Standardabweichung 1 transformiert werden, ergibt sich immer eine Steigung von $+1$ oder -1 für die optimale Gerade, unabhängig von der „Stärke" des Zusammenhangs. (Wenn die Korrelation 0 ist, ist die Gerade für standardisierte Variable unbestimmt.)

k Die bisher besprochenen Schätzmethoden setzen voraus, dass die Varianz σ_D^2 der Zufallsfehler D_i oder das Verhältnis σ_E/σ_D bekannt sei. Wenn über die **Varianzen** σ_D und σ_E **nichts bekannt** ist, wird das Problem in einem grundlegenden Sinn schwierig. Wenn die wahren Werte u_i als normalverteilte Zufallsvariable $U_i \sim \mathcal{N}\langle\mu, \sigma_U^2\rangle$ modelliert werden, dann lässt sich zeigen, dass die Parameter auch mit unendlich vielen Beobachtungen nicht geschätzt werden können. Es führen dann nämlich verschiedene Parametersätze ($[\widetilde{\beta},\ \widetilde{\alpha},\ \sigma_D\ , \sigma_E\ , \sigma_U]$) zur genau gleichen Verteilung der Beobachtungen $[X_i, Y_i]$. Das Modell ist „**nicht identifizierbar**".

Bei anderen Annahmen über die u_i ist die Identifizierbarkeit zwar theoretisch gegeben, aber für vernünftige Stichprobenumfänge nicht wirklich erreichbar. Man braucht in der Praxis also eine zusätzliche Information.

Kennt man wenigstens eine obere Schranke („grösser als ... kann σ_D nicht sein"), dann kann man den schlimmsten Fall durchrechnen und aus dem Unterschied zu den Resultaten für $\sigma_D = 0$ abschätzen, ob das Problem bedeutsam sei oder nicht.

l Wieso wird diese Methodik so **selten behandelt** und noch weniger angewandt? Nicht nur wegen mangelndem Wissen!

Wenn man Y „**vorhersagen**" oder interpolieren will, so macht dies meistens nur für gegebene X-Werte Sinn, nicht für gegebene u-Werte, da man diese ja nicht beobachten kann. Dann ist die gewöhnliche Regressionsrechnung angebracht. Allerdings muss gewährleistet sein, dass die X-Werte für die neuen Beobachtungen auf gleiche Weise zustande kommen wie die Daten, mit denen das Modell angepasst wurde.

Wenn die Frage interessiert, ob ein **Einfluss von** u **auf** Y (oder v) vorhanden sei, so muss man die Nullhypothese $\widetilde{\beta} = 0$ **testen**. Wenn die Hypothese gilt, ist auch die Steigung im Regressionsmodell von Y auf X null, und man kann den Test der gewöhnlichen Regressionsrechnung anwenden.

m *Literatur:* Wetherill (1986) gibt eine kurze, kritische Darstellung. Fuller (1987) ist ein umfassendes Werk über dieses Thema.

6.2 Eichung

a „Ausgleichs-Geraden“ werden oft verwendet, um eine Mess-Methode zu **eichen** oder um aus dem Resultat einer (billigen) Mess-Methode das Resultat einer anderen (teuren) zu „schätzen“.

Für die Bestimmung des Zusammenhangs geht man meist von bekannten „wahren“ Werten x_i (oder Werten der präzisen, teuren Mess-Methode) aus und bestimmt dazu die Werte Y_i der zu untersuchenden Methode. Es wird beispielsweise jeweils für eine chemische Lösung mit bekannter Konzentration die Absorption von Licht bei einer bestimmten Wellenlänge gemessen. (Meistens muss zunächst eine Reaktion durchgeführt werden, die die interessierende chemische Substanz in eine optisch erfassbare Substanz verwandelt.)

In der Anwendung der Eich-Geraden (oder -Kurve) ist umgekehrt der Wert Y der fraglichen Messmethode vorgegeben, und man will den zugehörigen wahren Wert x schätzen. Im Beispiel will man aus der Absorption die Konzentration der Lösung ausrechnen. Man verwendet die Regressions-Beziehung also in der „falschen“ Richtung. Daraus ergeben sich Probleme. Ihre Behandlung findet man auch unter dem Titel **inverse regression** oder **calibration**.

b Wir wollen hier eine einfache Behandlung vorstellen, die ein brauchbares Resultat ergibt, wenn der Zusammenhang eng (das Bestimmtheitsmass gross, beispielsweise über 0.95) ist.

Zunächst nehmen wir an, dass die x-Werte keine Messfehler aufweisen. Das erreicht man, indem man im Beispiel sehr sorgfältig erstellte Eich-Lösungen verwendet. Für mehrere solche Lösungen mit möglichst unterschiedlichen Konzentrationen führt man jeweils mehrere (möglichst) unabhängige Messungen (Aufbereitung und Ablesung des optischen Messgerätes) der Grösse Y durch. Daraus bestimmt man mit den besprochenen Methoden eine einfache lineare Regressionsgleichung – sofern Linearität vorhanden ist. Dies führt zu Schätzungen der Parameter α, β und σ und zu geschätzten Standardfehlern von $\widehat{\alpha}$ und $\widehat{\beta}$.

Wenn nun für eine zu messende Probe der Wert y abgelesen wird, ist klar, wie ein zugehöriger x-Wert bestimmt wird:

$$\widehat{x} = (y - \widehat{\alpha})/\widehat{\beta} \; .$$

c Die Frage stellt sich, wie genau dieser Wert ist.

Die Antwort lässt sich formulieren, indem wir x als Parameter ansehen, für den ein Vertrauensintervall gesucht ist. Ein solches Intervall ergibt sich (wie immer) aus einem Test. Nehmen wir als Nullhypothese $x = x_0$ an! Wie wir im Abschnitt über Vorhersage gesehen haben, liegt Y mit Wahrscheinlichkeit 0.95 in im Vorhersage-Intervall

$$\widehat{\alpha} \; + \; \widehat{\beta} x_0 \pm b \qquad \text{mit } b = q_{0.975}^{t_{n-2}} \, \widehat{\sigma} \, \sqrt{1 + \tfrac{1}{n} + (x_0 - \bar{x})^2/\mathrm{SSQ}^{(X)}} \; ,$$

das in Abbildung 6.2.c wie in Abbildung 2.4.c – gleich für alle möglichen x_0 – dargestellt ist. Das Intervall bildet deshalb ein Annahmeintervall für die Grösse Y (die hier die Rolle einer Teststatistik spielt) unter der Nullhypothese $x = x_0$.

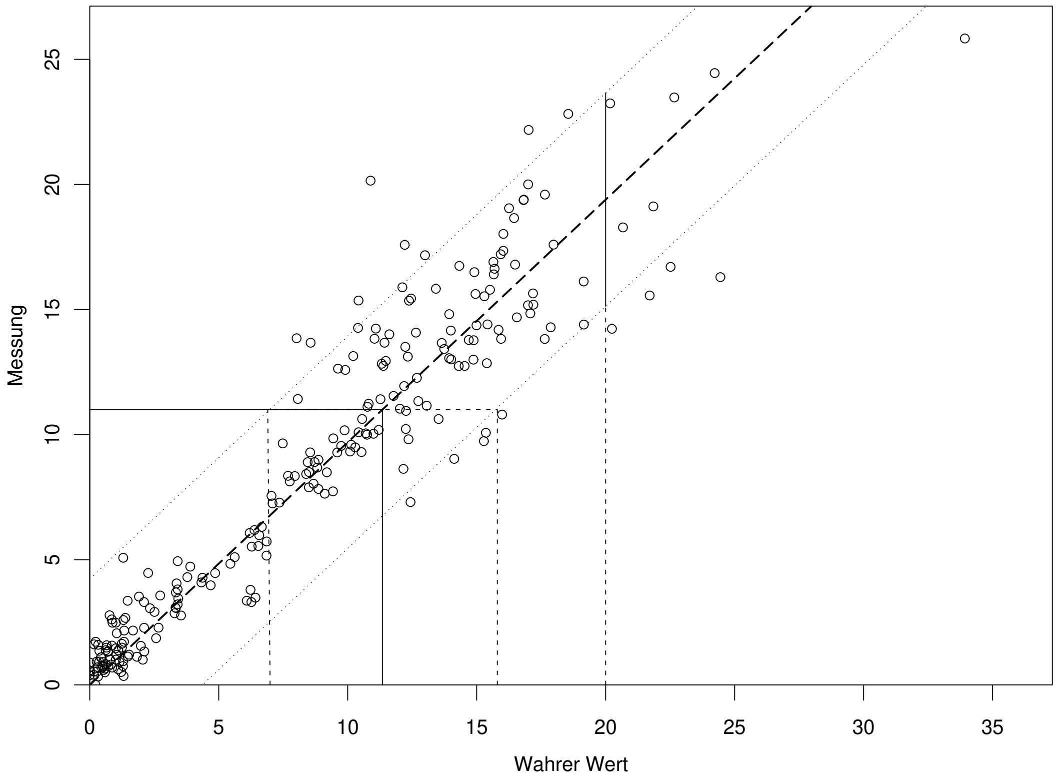

Abbildung 6.2.c: Veranschaulichung der Verwendung einer Eichgeraden für einen Messwert von 11. Zum Vergleich die Verwendung für eine Vorhersage des Messwertes bei einem wahren Wert von 20.

Die Abbildung veranschaulicht nun den weiteren Gedankengang: Messwerte y sind mit Parameterwerten x_0 vereinbar im Sinne des Tests, wenn der Punkt $[x_0, y]$ zwischen den eingezeichneten Kurven liegt. In der Figur kann man deshalb ohne Schwierigkeiten die Menge der x_0-Werte bestimmen, die mit der Beobachtung y verträglich sind. Sie bilden das eingezeichnete Intervall – das Vertrauensintervall für x_0. In sehr guter Näherung hat dies den Mittelpunkt $\widehat{x}$ und die Breite $2 \cdot b/\widehat{\beta}$, ist also gleich

$$(y - \widehat{\alpha})/\widehat{\beta} \pm b/\widehat{\beta} \; .$$

d* Einige weitere Stichworte:

- Fehlerbehaftete x-Werte: Man verwende eine Schätzung der „wahren Geraden“ $\widetilde{\alpha} + \widetilde{\beta}\, x$.
- Überprüfung der Linearität und anderer Modell-Annahmen ist wichtig!
- Periodische Eichung: sollte nicht mit Einzelmessungen erfolgen.

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.