Lineare Regression

Werner Stahel
Seminar für Statistik, ETH Zürich

Januar 2006

Unterlagen zum Block Rg1 des Kurses in Angewandter Statistik

Inhaltsverzeichnis

1 Einführung in die statistische Regressionsrechnung

1.1 Beispiele zur linearen Regression

a In der Wissenschaft, in der Technik und im Alltag fragen wir immer wieder danach, wie eine Grösse, die uns speziell interessiert, von anderen Grössen abhängt. Diese grundlegende Frage behandelt die statistische Regression, die deshalb wohl (neben einfachen grafischen Darstellungen) die am meisten verwendete Methodik der Statistik darstellt.

In diesem Abschnitt soll mittels Beispielen zur „gewöhnlichen" linearen Regression in die Problemstellung eingeführt werden, bevor ein Überblick über die verschiedenen, allgemeineren Regressions-Modelle geboten wird.

b ▷ ***Beispiel Sprengungen***. *Beim Bau eines Strassentunnels zur Unterfahrung einer Ortschaft muss gesprengt werden. Die Erschütterung der Häuser darf dabei einen bestimmten Wert nicht überschreiten. In der Nähe der Häuser muss daher vorsichtig gesprengt werden, was natürlich zu erhöhten Kosten führt. Es lohnt sich, eine Regel zu entwickeln, die angibt, wie stark in welcher Situation gesprengt werden darf.*

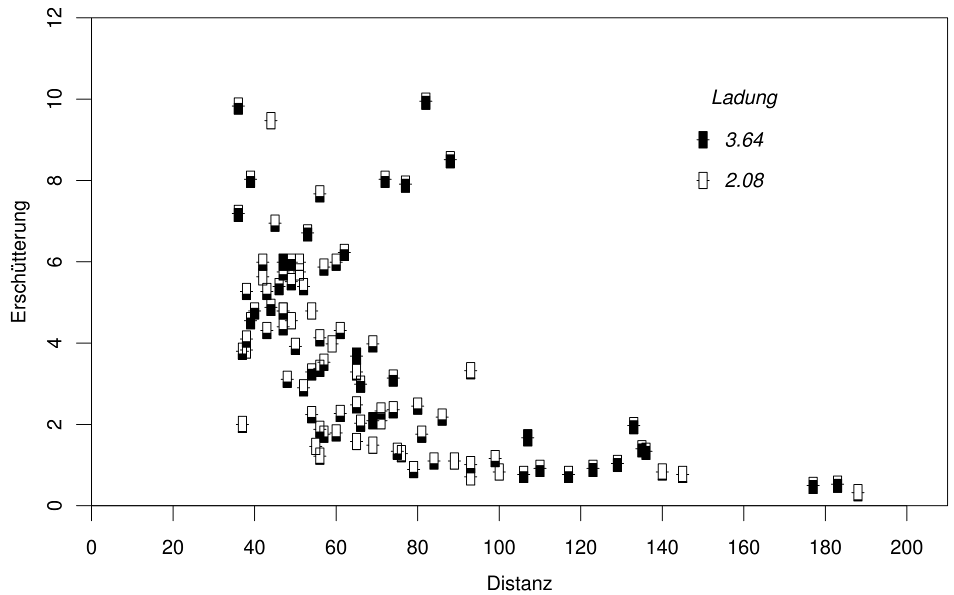

Abbildung 1.1.c: Erschütterung in Abhängigkeit von der Distanz für verschiedene Ladungen

Version WL Jan 2006, © W. Stahel

Die Erschütterung ist abhängig von der Sprengladung, von der Distanz zwischen dem Spreng- und dem Messort, von der Art des Untergrund-Materials zwischen diesen Punkten, vom Ort der Sprengung im Tunnelprofil und möglicherweise von weiteren Grössen. Wäre die Erschütterung eine exakte, bekannte Funktion dieser Grössen und könnte man sie bei einer geplanten Sprengung alle genau erfassen, dann könnte man die Sprengladung ausrechnen, die zu einer gerade noch tolerierbaren Erschütterung führt.

c Beginnen wir, mathematische Symbole und Sprachregelungen einzuführen! Die **Zielgrösse** y (englisch *target variable)* – die Erschütterung – hängt über eine Funktion h von den **Ausgangsgrössen** oder **erklärenden Variablen** $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ *(explanatory variables)* – Ladung, Distanz, Spreng-Situation, Untergrundart – ab. (Die ebenfalls gebräuchlichen Ausdrücke **„unabhängige Variable"** für die $x^{(j)}$ und **„abhängige Variable"** für y sind irreführend, da sie mit stochastischer Unabhängigkeit nichts zu tun haben.)

Im Idealfall sollte also

$$y_i = h\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \rangle$$

für jede **Beobachtung** i (jede Sprengung) gelten.

Leider existiert eine solche Formel nicht, und das Untergrundmaterial ist sowieso nicht genau genug erfassbar. Abbildung 1.1.c zeigt die Erschütterung in Abhängigkeit von der Distanz für verschiedene Ladungen. (Die Daten stammen vom Bau der Unterfahrung von Schaffhausen. Sie wurden freundlicherweise vom Ingenieurbüro Basler und Hoffmann, Zürich, zur Verfügung gestellt.)

Die statistische Regressionsrechnung geht davon aus, dass eine Formel wenigstens „ungefähr" gilt – bis auf Abweichungen, die „zufällig" genannt werden. Wir schreiben

$$Y_i = h\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \rangle + E_i$$

und nennen die E_i die **Zufallsfehler**. Die Vorstellungen, wie gross solche Abweichungen sind, werden mit einer Wahrscheinlichkeits-Verteilung formuliert. Oft wird dafür die Normalverteilung verwendet.

Man wird mit Hilfe dieses Modells trotz der Unsicherheit eine Regel für die zu wählende Grösse der Sprengladung herleiten können. Allerdings muss man zulassen, dass gemäss Modell auch eine zu grosse Erschütterung mit einer gewissen Wahrscheinlichkeit auftreten kann. Will man diese Wahrscheinlichkeit klein halten, so muss man entsprechend vorsichtig sprengen. Die statistische Regressionsrechnung gibt einen Zusammenhang zwischen der Ladung und der Wahrscheinlichkeit einer zu grossen Erschütterung bei einer bestimmten Distanz an.

Dieses Beispiel wird uns in den kommenden Abschnitten begleiten. Auf die Antworten müssen Sie deshalb noch eine Weile warten.

d ▷ ***Beispiel Schadstoffe im Tunnel**. Die Schadstoffe, die vom motorisierten Verkehr ausgestossen werden, bilden einen wesentlichen Bestandteil der Belastung der Luft. Um die Grösse dieser Belastung zu schätzen, werden für die Fahrzeuge so genannte **Emissionsfaktoren** bestimmt. Dies kann einerseits auf dem Prüfstand geschehen, auf dem die Strasse mit Rollen simuliert wird. Der Widerstand der Rol-*

len wird dabei variiert, so dass ein typischer „Fahrzyklus" durchgespielt werden kann. – Andererseits eignen sich Strassentunnels mit Ein-Richtungs-Verkehr für Messungen unter realen Bedingungen. Misst man Schadstoff-Konzentrationen am Anfang und am Schluss des Tunnels und zählt, wie viele Fahrzeuge durch den Tunnel fahren, so kann man ebenfalls Emissionsfaktoren ausrechnen. Allerdings erhält man zunächst nur einen gemittelten Faktor für jeden gemessenen Schadstoff, und dieser lässt sich nicht ohne zusätzliche Erkenntnisse auf andere Strassenabschnitte übertragen. Wenn man die Anzahl der Fahrzeuge nach Fahrzeug-Kategorien aufteilen kann, dann kann man immerhin mit Regressionsrechnung zu einem Emissionsfaktor für jede Fahrzeug-Kategorie kommen.

Während einer Woche im September 1993 wurden in der Südröhre des Gubrist-Tunnels nördlich von Zürich solche Messungen durchgeführt. Die Schadstoff-Konzentrationen am Anfang und am Ende wurden gemessen und die Luftströmung erfasst. Daraus lässt sich die Schadstoff-Emission Y pro Kilometer für alle durchgefahrenen Fahrzeuge zusammen berechnen. Von einem Schlaufen-Detektor im Strassenbelag wurden die Fahrzeuge in zwei Kategorien gezählt: Auf Grund des Abstands von Vorder- und Hinterachse wurden die Lastwagen von den übrigen Fahrzeugen getrennt. Es bezeichne $x^{(1)}$ die Anzahl „Nicht-Lastwagen" und $x^{(2)}$ die Anzahl Lastwagen. Die gesamten Emissionen in der Zeitperiode i setzen sich zusammen gemäss

$$Y_i = \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + E_i \ ,$$

wobei θ_1 die durchschnittliche Emission pro Nicht-Lastwagen und θ_2 diejenige pro Lastwagen bedeutet – also die Grössen, an denen wir in der Studie primär interessiert sind. Die „Zufallsfehler" E_i entstehen durch Variationen in Bauart und Zustand der Fahrzeuge, durch zeitliche Abgrenzungs-Schwierigkeiten und durch Mess-Ungenauigkeiten.

e ▷ *Die Formel lässt sich in eine üblichere und vielleicht noch einfachere Form bringen: Wir dividieren Y_i, $x_i^{(1)}$ und $x_i^{(2)}$ durch die gesamte Anzahl Fahrzeuge $x_i^{(1)} + x_i^{(2)}$ und erhalten $\widetilde{Y}_i = \theta_1 \widetilde{x}_i^{(1)} + \theta_2 \widetilde{x}_i^{(2)} + \widetilde{E}_i$, wobei $\widetilde{Y}_i$ der „mittlere Emissionsfaktor" für die Zeitperiode i und $\widetilde{x}_i^{(1)}$ und $\widetilde{x}_i^{(2)}$ die Anteile der Nicht-Lastwagen und der Lastwagen bedeuten. Da $\widetilde{x}_i^{(1)} = 1 - \widetilde{x}_i^{(2)}$ ist, gilt*

$$\widetilde{Y}_i = \theta_1 + (\theta_2 - \theta_1)\widetilde{x}_i^{(2)} + \widetilde{E}_i \ .$$

Mit weniger komplizierten Symbolen geschrieben sieht das so aus:

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

*Dies ist das Modell einer so genannten **einfachen linearen Regression**. Die Konstanten α und β nennen wir **Koeffizienten** oder **Parameter** des Modells. Wir wollen sie aus den Daten der Studie bestimmen, also **schätzen**.*

In Abbildung 1.1.d zeigt sich als Tendenz eine lineare Zunahme des mittleren Emissionsfaktors für NO_x mit zunehmendem Lastwagen-Anteil, wie es dem besprochenen Modell entspricht.

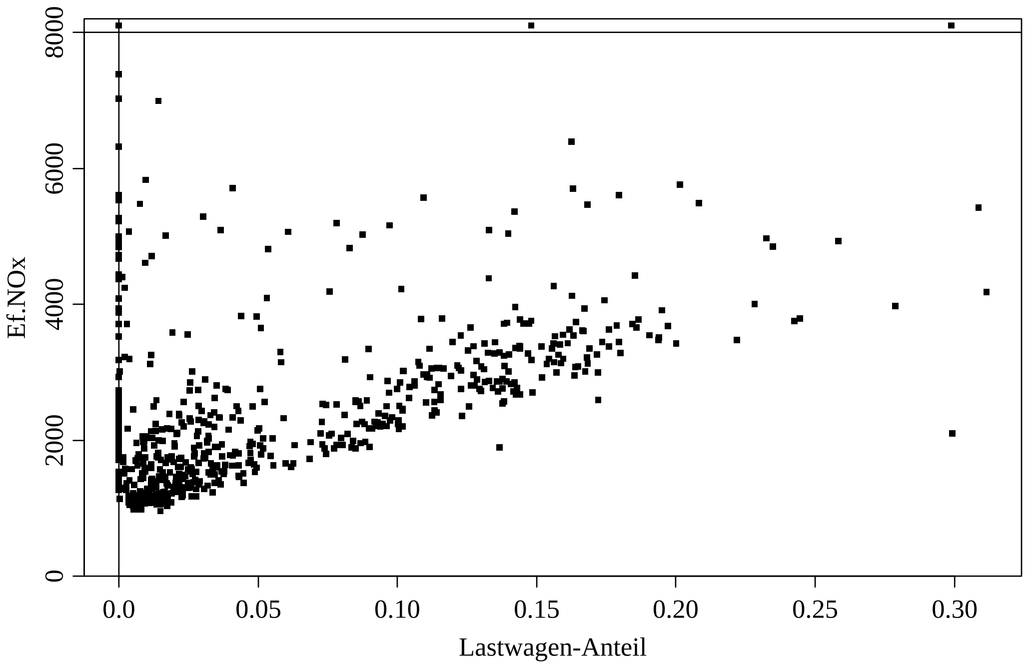

Abbildung 1.1.d: Emissionsfaktor für NO_X und Lastwagen-Anteil, gemittelt über jeweils 15 Minuten, im Beispiel der Schadstoffe im Tunnel. Drei extrem hohe Y-Werte sind im Bildrand dargestellt.

f ▷ ***Beispiel Lastwagen-Anteil****.* *Der Schlaufen-Detektor zählt zwar die gesamte Zahl der Fahrzeuge zuverlässig, kann aber den Anteil der Lastwagen nur ungenau erfassen. Deshalb (unter anderem) wurde der Verkehr zeitweise mit Video aufgenommen und der Lastwagen-Anteil auf diesen Aufnahmen genau ausgezählt. Da dies teurer war, konnte nicht der ganze Zeitraum abgedeckt werden. Abbildung 1.1.f zeigt, dass die Schlaufen-Zählung systematische und zufällige Abweichungen von der Video-Zählung aufweist. Die zufälligen Abweichungen kommen teilweise zustande, weil die Schlaufe am Anfang, die Kamera aber am Ende des Tunnels installiert war, und die Abgrenzung der Mess-Intervalle nicht entsprechend korrigiert wurde. (Die Fahrzeit beträgt etwa 3 Minuten, die Intervalle dauerten 15 Minuten.)*

Es ergibt sich die weit verbreitete Situation, dass der Wert einer interessierenden Grösse auf Grund der Messung einer mit ihr zusammenhängenden anderen Grösse mittels einer Umrechnungsformel ermittelt werden soll. Dabei kann die Messung auf einer ganz anderen Skala erfolgen; beispielsweise wird eine Konzentration mittels einer optischen Durchlässigkeit erfasst.

Man geht zunächst davon aus, dass für einen gegebenen exakten Wert x_i *die Messung* Y_i *sich aus einem „Idealwert“* $h\langle x_i\rangle$ *und einem Messfehler* E_i *zusammensetzt. Das entspricht einem Regressionsmodell. Man bestimmt die Funktion* h *mittels Messungen* Y_i*, für die der zugehörige Wert* x_i *bekannt ist. In der Anwendung wird aber nicht von* x *auf* Y*, sondern von einem Messwert* Y *auf den gesuchten Wert* x *geschlossen. Aus dieser Umkehrung ergeben sich gewisse zu-*

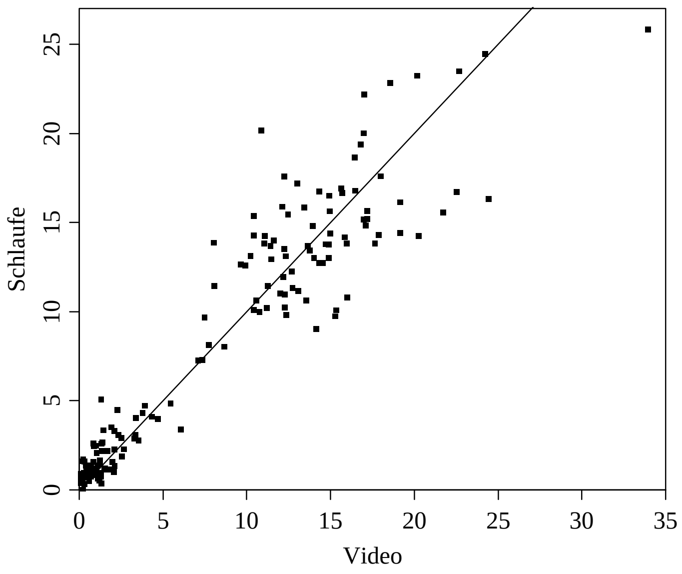

Abbildung 1.1.f: Lastwagen-Anteil (in Prozenten) gemäss Schlaufen- und Videozählung. Die Gerade stellt die Gleichheit $(y = x)$ dar.

sätzliche Probleme.

Dieses Vorgehen entspricht der ***Eichung*** *eines Messgeräts. Man misst Proben mit bekanntem exaktem Wert (z. B. bekannter Konzentration) und liest die Messung ab. Dann wird die Ablese-Skala ajustiert, was der Schätzung und Verwendung der Funktion h in unserem allgemeineren Zusammenhang entspricht.*

g ▷ ***Beispiel basische Böden.*** *In Indien behindern basische Böden, also tiefe Säurewerte oder hohe pH-Werte, Pflanzen beim Wachstum. Es werden daher Baumarten gesucht, die eine hohe Toleranz gegen solche Umweltbedingungen haben. In einem Freilandversuch wurden auf einem Feld mit grossen lokalen Schwankungen des pH-Wertes 120 Bäume einer Art gepflanzt und ihre Höhe Y_i nach 3 Jahren gemessen. Abbildung 1.1.g zeigt die Ergebnisse mit den zugehörigen pH-Werten $x_i^{(1)}$ des Bodens zu Beginn des Versuchs. Zusätzlich wurde eine Variable $x_i^{(2)}$ gemessen, die einen etwas anderen Aspekt der „Basizität“ erfasst (der Logarithmus der so genannten sodium absorption ratio, SAR). Dieses Beispiel hat also zwei Ausgangsgrössen.*

Ein Hauptziel der Untersuchung besteht darin, für gegebene Werte der beiden Ausgangsgrössen an einem möglichen Pflanzort bestimmen zu können, wie gut ein solcher Baum dort wohl wachsen wird. Es stellt sich zusätzlich die Frage, ob die Messung der zweiten Grösse $x^{(2)}$ dazu überhaupt etwas beiträgt, oder ob der pH $(x^{(1)})$ allein auch genügt.

h ▷ ***Beispiel Antikörper-Produktion.*** *Grössere Mengen von Antikörpern werden in biotechnologischen Prozessen gewonnen. Dazu werden biotechnologisch veränderte Zellen, die den entsprechenden Antikörper produzieren können, Wirtstieren (z. B. Mäusen) injiziert. Nach einer gewissen Zeit beginnen diese Zellen Antikör-*

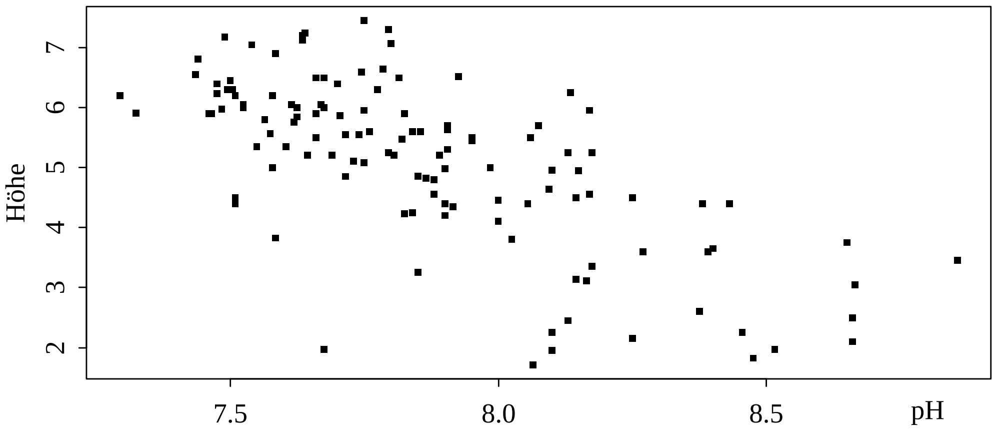

Abbildung 1.1.g: Baumhöhe in Abhängigkeit vom pH für das Beispiel der basischen Böden

per zu produzieren und auszuscheiden. Die ausgeschiedene Flüssigkeit wird dann eingesammelt und weiter verarbeitet. Dieses Beispiel wird ausführlich in Haaland (1989) dargestellt und analysiert. Es dient uns hier nur zur Illustration der Fragestellung.

Die Zellen können erfahrungsgemäss nur Antikörper produzieren, wenn das Immunsystem der Wirtstiere geschwächt wird. Dies kann durch 4 Faktoren geschehen. Es wird zudem vermutet, dass die Menge der injizierten Zellen und deren Entwicklungsstand die Antikörper-Produktion beeinflusst.

Da es für so komplexe biologische Prozesse keine theoretischen Modelle gibt, werden die relevanten Prozessfaktoren durch ein Experiment ermittelt. Ein solches Experiment braucht viele Mäuse, ist zeitaufwändig und kostet Geld. Mit einer geschickten Versuchsanordnung können unter geringstmöglichem Aufwand die wichtigen Prozessfaktoren ermittelt werden. Hier hilft die ***statistische Versuchsplanung****.*

i ▷ *Als relevante Prozessfaktoren wurden in dieser Studie zwei Prozessfaktoren identifiziert, nämlich die Dosis von* Co^{60} *Gamma-Strahlen und die Anzahl Tage zwischen der Bestrahlung und der Injektion eines reinen Öls (englische Bezeichnung* pristane*). Diese beiden Prozessfaktoren sollen nun so eingestellt werden, dass eine möglichst optimale Menge von Antikörpern durch die veränderten Zellen produziert wird.*

Dazu wollen wir ein empirisches Modell $Y_i = h\langle x_i^{(1)}, x_i^{(2)}\rangle + E_i$ *finden, das die Ausbeute* Y *von Antikörpern möglichst gut aus den beiden Prozessfaktoren* $x^{(1)}$ *und* $x^{(2)}$ *vorhersagt. Als Funktion* h *wird oft ein quadratisches Polynom in den Variablen* $x^{(1)}$ *und* $x^{(2)}$ *verwendet. Mit dem aus den Daten bestimmten Modell lässt sich dann die optimale Einstellung* $[x_o^{(1)}, x_o^{(2)}]$ *der Prozessfaktoren bestimmen.*

1.2 Fragestellungen

a Von der Problemstellung her können die Anwendungen der Regression in Gruppen eingeteilt werden:
• **Vorhersage, Prognose, Interpolation**. Im Beispiel der Sprengungen soll eine Formel helfen, für gegebene Distanz und Ladung die Erschütterung „vorherzusagen". Es interessiert nicht nur der mittlere zu erwartende Wert, sondern auch eine obere Grenze, über der die Erschütterung nur mit kleiner Wahrscheinlichkeit liegen wird. (Die Begriffe Vorhersage und Prognose werden meistens für eine zeitliche Extrapolation in die Zukunft verwendet. Hier spielt die Zeit keine Rolle – ausser dass die Problemstellung nur wesentlich ist, wenn die Sprengung noch nicht erfolgt ist.)

b • **Schätzung von Parametern**. Im Beispiel des Gubrist-Tunnels sollen zwei Konstanten, die Emissionsfaktoren für Lastwagen und für übrige Fahrzeuge, bestimmt werden.

c • **Bestimmung von Einflussgrössen**. Im Beispiel der Antikörper-Produktion müssen zunächst aus mehreren in Frage kommenden Ausgangsgrössen diejenigen herausgefunden werden, die die Zielvariable wesentlich beeinflussen. In vielen Forschungs-Projekten steht diese Frage ebenfalls im Vordergrund: Von welchen Grössen wird eine Zielgrösse eigentlich beeinflusst?

d • **Optimierung**. Im Beispiel der Antikörper-Produktion sollten optimale Produktionsbedingungen gefunden werden. In allen Bereichen der Produktion ist diese Frage offensichtlich von grundlegender Bedeutung.

e • **Eichung**. Auf Grund der ungenauen und systematisch verfälschten Angabe des Schlaufen-Detektors soll der Anteil der Lastwagen bestimmt werden. Diese Problemstellung kombiniert Elemente der Vorhersage und der Schätzung von Parametern.

f Der Block Regression 1 wird sich vor allem mit den ersten drei Fragen befassen.

1.3 Ausblick

a In der **linearen Regression**, die im Folgenden behandelt wird, setzt man voraus,
• dass die Zielgrösse eine kontinuierliche Variable ist,
• dass die zufälligen Abweichungen E_i einer Normalverteilung folgen und von einander statistisch unabhängig sind
• und dass die Funktion h von einer einfachen Form ist, nämlich in einem gewissen Sinne linear (siehe 3.2.w). Die gleichen Fragestellungen werden auch in der Varianzanalyse 1 behandelt, mit anderen Schwerpunkten bezüglich der Art der Ausgangsgrössen.

b Am Ende dieses Blockes und in späteren Blöcken wird dieser Ansatz in vielen Richtungen erweitert:

• Wenn die Funktion h nicht im erwähnten Sinne linear ist, kommt die **nichtlineare Regression** zum Zug.

c • Wenn die Beobachtungen der Zielgrösse und der erklärenden Grössen in einer zeitlichen Abfolge auftreten, entstehen normalerweise besondere Probleme durch entsprechende Korrelationen. Diese Besonderheiten werden in der Theorie der **Zeitreihen** behandelt.

d • Man kann an mehreren Zielgrössen interessiert sein. Eine einfache Art, damit umzugehen, besteht darin, für jede von ihnen eine separate Regressionsrechnung durchzuführen. Die multivariate Statistik zeigt, wie man bei gemeinsamer Betrachtung mit **multivariater Regression und Varianzanalyse** noch etwas darüber hinaus gewinnen kann.

e • Die Annahme der Normalverteilung für die E_i ist oft nur näherungsweise erfüllt. Die Methoden, die wir im Folgenden kennen lernen, sind dann nicht mehr gut geeignet. Besser fährt man mit den Methoden der **robusten Regression**.

f • Die interessierende Zielgrösse kann eine zweiwertige Variable (Ja/Nein) sein. Das führt zur **logistischen Regression**. Ist die Zielvariable eine Zählgrösse, eine diskrete geordnete oder eine nominale Variable, so sind die **verallgemeinerten linearen Modelle** anzuwenden, zu denen auch das gewöhnliche und das logistische Regressionmodell gehören.

g • Zeiten bis zum Ausfall eines Gerätes oder bis zum Eintreffen eines anderen Ereignisses folgen meist anderen Verteilungen als der üblicherweise verwendeten Normalverteilung. Ausserdem werden solche Ereignisse oft nicht für alle Beobachtungseinheiten abgewartet, was zu so genannt zensierten Daten führt. Es gibt auch für solche Daten geeignete Regressionsmethoden, die im Gebiet der **Überlebenszeiten** (*survival* oder *failure time data*) behandelt werden.

h • In der linearen Regression werden nur die Abweichungen E_i als Zufallsvariable modelliert. Manchmal kann es auch sinnvoll sein, die **Parameter** selbst durch **Zufallsgrössen** zu ersetzen. Dies kommt vor allem in einem weiterführenden Gebiet der Varianzanalyse (repeated measures und „Spaltanlagen“, *split plot designs)* zum Zug, wo man von **zufälligen Effekten** spricht.

i • In all diesen Modellen ist die Regressionsfunktion ein Mitglied einer Schar von vorgegebenen Funktionen, die durch einen oder mehrere Parameter charakterisiert ist. Es geht dann darum, diese(n) Parameter zu bestimmen. Was wir intuitiv oft wollen, ist kein in solcher Weise vorgegebener Funktionstyp, sondern einfach eine „glatte Funktion“. Man spricht von „**Glättung**“ der Daten. Wie man eine solche Idee mathematisch formuliert und die entsprechende Funktion schätzt, untersucht die **nichtparametrische Regression**.

j In all diesen Verallgemeinerungen erscheinen immer wieder die gleichen Grundideen, die wir nun an Hand der linearen Regression – zunächst mit einer einzigen erklärenden Variablen, nachher mit mehreren – einführen wollen.

Die folgenden Unterlagen für die einfache Regression enthalten Repetitions-Abschnitte zu den Begriffen der Schliessenden Statistik. Sie sollen den Einstieg vor allem jenen erleichtern, die nicht gerade den entsprechenden Block des Nachdiplomkurses hinter sich haben.

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.