Angewandte statistische Regression

Marianne Müller
Zürcher Hochschule Winterthur

26. Oktober 2005

Inhaltsverzeichnis

1 Einführung

- Wann macht man eine Regressionsanalyse?
- Was ist ein lineares Modell?
- Welche Modelltypen gibt es?

- Ist Cadmium gesundheitsschädigend?
- Welche Faktoren haben den grössten Einfluss auf den Ozongehalt?
- Welche Baumart wächst am schnellsten auf basischen Böden?
- Wieso variieren die Kosten pro behandelte Person in verschiedenen Spitälern?
- Wer ist bereit, für eine verbesserte Nutztierhaltung mehr Geld aufzuwenden?
- Bei welchen Personen ist das Risiko einer postoperativen Venenthrombose erhöht und sollte deshalb prophylaktisch angegangen werden?

Worin besteht das Gemeinsame und was sind die Unterschiede zwischen diesen Beispielen?

1.1 Das lineare Modell

Jedes der obigen Beispiele kann formuliert werden als Frage nach dem Zusammenhang zwischen einer *Zielvariablen* Y und einer oder mehrerer *erklärender Variablen* $x_1,\ldots,x_p$.

Bsp.	Zielvariable	erklärende Variablen
1	Lungenkapazität	Expositionsdauer, Alter
2	Ozongehalt	Meteorologische Daten, Region, Verkehr
3	Baumhöhe	ph-Wert, Bodentyp
4	Kosten	Stadt-Land, Altersverteilung, Ärztedichte, mittleres Einkommen
5	Höhe der Zahlungsbereitschaft	Einkommen, politische Haltung, Geschlecht
6	% Thrombose	Alter, BMI, Fibrinogen

Man versucht die funktionale Beziehung zwischen der Zielvariablen Y und den möglichen erklärenden Variablen $x_1,\ldots,x_p$ durch ein Modell zu beschreiben. Meist beschränkt man sich dabei auf *lineare Modelle*, sodass einige Variablen eventuell zuerst tranformiert werden müssen. Die mathematische Schreibweise für den systematischen Teil des linearen Modells sieht folgendermassen aus:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{1.1}$$

$\beta_0,\ldots,\beta_p$ sind die *Regressionskoeffizienten*, welche in der Regressionsanalyse aus den vorhandenen Daten geschätzt werden.

1.2 Ziel einer Regressionsanalyse

Es gibt verschiedene Gründe, eine Regressionsanalyse durchführen zu wollen:

- Verständnis für den kausalen Zusammenhang
- Vorhersage

In Beispiel 1 möchte man nachweisen, dass die Cadmium-Exposition eine Reduktion der Lungenkapazität verursacht. In einer Regression sollten also die gemessenen Lungenfunktionswerte mit zunehmender Expositionsdauer abnehmen. Weil schon länger beschäftigte Personen aber tendenziell älter sind und ältere Leute schlechtere Lungenfunktionswerte aufweisen, ist es wichtig (aber nicht ganz einfach), einen Expositionseffekt unabhängig vom Alter nachweisen zu können.

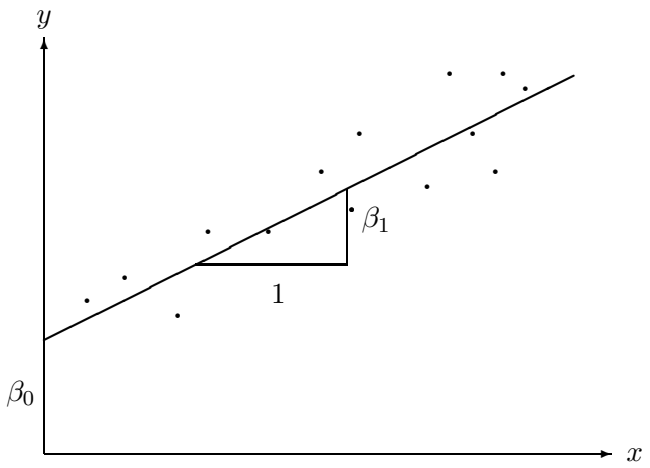
Auch bei den Spitalkosten sind wir an Ursachen interessiert. Durch geeignete Manipulation von erklärenden Variablen sollen Kosten gesenkt werden können. Eine absolut klare Aussage bezüglich Kausalität erhält man natürlich nur mit einem kontrollierten Experiment wie es in Beispiel 3 denkbar wäre.

Beispiel 2 kann auch als Vorhersageproblem angeschaut werden. Es gibt wahrscheinlich mehrere verschiedene Regressionsmodelle, die ähnlich gute Prognosen liefern, obwohl sie jeweils andere erklärende Variablen beinhalten.

1.3 Modelltypen

Multiple Regression, Varianzanalyse, logistische Regression, das sind alles Spezialfälle des linearen Modells (1.1). Welcher Modelltyp benutzt werden soll, hängt vor allem von der Art der vorhandenen Daten ab. Wir unterscheiden zwischen Binärdaten, z. B. Geschlecht, kategoriellen Daten, z. B. sozio-oekonomische Schicht, und stetigen Daten, z. B. Baumhöhe.

Einfache lineare Regression:
Untersucht wird der lineare Zusammenhang zwischen zwei stetigen Variablen y und x. Die folgende Figur zeigt einen Scatterplot mit angepasster Gerade $y = \beta_0 + \beta_1 x$, wobei β_0 den Achsenabschnitt und β_1 die Steigung bezeichnet. Wenn also x um eine Einheit wächst, nimmt y um β_1 zu.



Beispiel: Lungenfunktion y in Abhängigkeit von der Expositionsdauer x.

Multiple Regression:
Hier werden mehr als eine stetige erklärende Variable betrachtet. Das einfachste Beispiel einer multiplen Regression enthält also zwei erklärende Variablen x_1 und x_2. Es wird dann eine Ebene an die Daten angepasst: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Auch einzelne kategorielle Variablen können als sogenannte *Dummy-Variablen* in die Regressionsgleichung aufgenommen werden. Die Interpretation der Regressionskoeffizienten ist ähnlich wie zuvor. Wenn x_1 um eine Einheit zunimmt, verändert sich y um β_1, vorausgesetzt x_2 bleibt konstant. Beispiel: Lungenfunktion y in Abhängigkeit von Expositionsdauer x_1 und Alter x_2.

Varianzanalyse:
Die Zielvariable ist stetig. Alle erklärenden Variablen sind binär oder kategoriell, sogenannte *Faktoren*. Wenn nur ein Faktor untersucht wird, spricht man von *Ein-Weg-Varianzanalyse*, sonst von *Mehr-Weg-Varianzanalyse*. Werden zusätzlich noch ein paar wenige stetige erklärende Variablen als sog. *Covariablen* mitberücksichtigt, so ergibt sich eine *Covarianzanalyse*.

Das Modell der Varianzanalyse kann zwar in der allgemeinen Form eines linearen Modells (1.1) geschrieben werden, in der Praxis zieht man aber eine andere Schreibweise vor. Die Ergebnisse werden in einer Varianzanalyse-Tabelle dargestellt und auf die Angabe der Regressionskoeffizienten β_i wird verzichtet. Beispiel: Baumhöhe in Abhängigkeit von Bodentyp, ph-Wert (hoch/tief), Klimatyp.

Logistische Regression:
Die Zielvariable ist in diesem Modell binär. Die erklärenden Variablen können stetig oder kategoriell sein. In der medizinischen und sozialwissenschaftlichen Forschung werden diese Modelle sehr häufig benutzt, da sehr oft Binärvarablen wie „geheilt/nicht geheilt“ oder „stimmt zu/stimmt nicht zu“ untersucht werden. Die Regressionskoeffizienten können in *odds ratios* transformiert werden. Beispiel: Variable „Thrombose ja/nein“ in Abhängigkeit von BMI, Altersgruppe und Fibrinogen.

Loglineare Modelle, Poissonregression:
Die Zielvariable ist eine Anzahl oder Rate. Die erklärenden Variablen können stetig oder kategoriell sein. Loglineare Modelle werden für die Analyse von mehrdimensionalen Kontingenztafeln verwendet. Wiederum entsprechen die Regressionskoeffizienten odds ratios. Beispiel: Anzahl gemeldeter Schadensfälle in Abhängigkeit von der Region, dem Jahr, der wirtschaftlichen Lage oder Zusammenhang zwischen Spitalkosten („hoch/mittel /tief“) und Ärztedichte („hoch/mittel/tief“), einem Faktor „Patientenmix“ und dem Faktor Kanton.

Cox' Proportional Hazard Modell:
Die Zielvariable ist eine Überlebenszeit. Die erklärenden Variablen können stetig oder kategoriell sein. Beispiel: Überlebenszeit einer elektronischen Komponente in Abhängigkeit von der Art der Benutzung, dem verwendeten Material, der Herstellungsart, usw.

Die wichtigsten Analysemethoden für multivariate Datensätze können also unter dem Oberbegriff **verallgemeinerte lineare Modelle** zusammengefasst werden. Bei der konkreten Berechnung von Schätzungen und Vertrauensintervallen und für die Modellüberprüfung sind dann aber je nach Modelltyp andere Methoden verfügbar. Wir beschränken uns zunächst im folgenden auf die einfache und die multiple lineare Regression.

Fragen, die wir zu beantworten versuchen, sind:

- Wie werden die β_i 's geschätzt und dazugehörige Vertrauensintervalle berechnet?
- Was für Voraussetzungen sind nötig, damit die Methoden zulässig sind, und wie werden diese Voraussetzungen überprüft?
- Wie gut passt das Modell? Was tun, wenn das Modell nicht passt?
- Wie wählen wir das „beste“ Modell?
- Wann ist eine Regressionsanalyse überhaupt geeignet und wann nicht?

2 Einfache lineare Regression

- Was ist die Methode der kleinsten Quadrate?
- Wie sieht eine Varianzanalyse-Tabelle aus?
- Wie wird das Modell überprüft?

Beispiel:
Bei 40 Industriearbeitern, die unterschiedlich lange Cadmiumdämpfen ausgesetzt waren, wurden Lungenfunktionsmessungen durchgeführt. Die folgende Tabelle enthält neben diesen Messungen das Alter der 40 Männer.

Exposition $>$ 10 Jahre		Exposition $<$ 10 Jahre			
Alter	**Vitalkapazität [l]**	**Alter**	**Vitalkapazität [l]**	**Alter**	**Vitalkapazität [l]**
39	4.62	29	5.21	38	3.64
40	5.29	29	5.17	38	5.09
41	5.52	33	4.88	43	4.61
41	3.71	32	4.50	39	4.73
45	4.02	31	4.47	38	4.58
49	5.09	29	5.12	42	5.12
52	2.70	29	4.51	43	3.89
47	4.31	30	4.85	43	4.62
61	2.70	21	5.22	37	4.30
65	3.03	28	4.62	50	2.70
58	2.73	23	5.07	50	3.50
59	3.67	35	3.64	45	5.06
		48	4.06	51	4.51
		46	4.66	58	2.88

Bevor wir untersuchen wollen, ob die länger exponierten Männer schlechtere Lungenfunktionswerte besitzen als die kürzer Exponierten, studieren wir den Zusammenhang zwischen Vitalkapazität und Alter. Die Graphik 2.1 stellt die 40 Beobachtungen in einem Streudiagramm (scatterplot) dar.

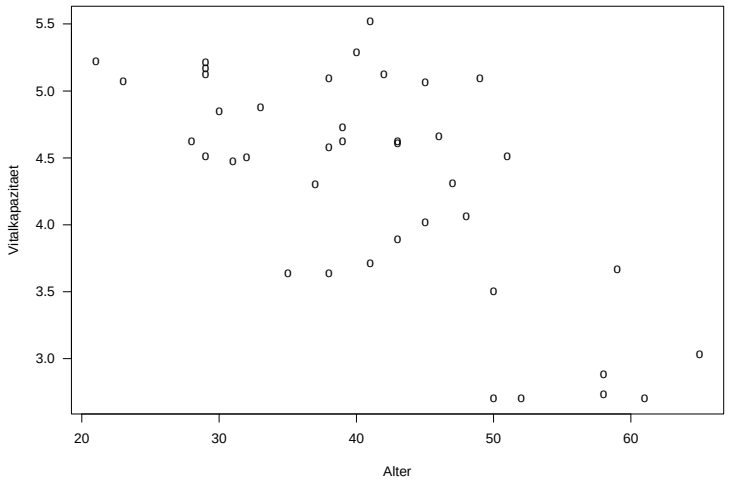


Abbildung 2.1: Lungenfunktionsmessungen von Cadmium-Arbeitern

Mit zunehmendem Alter nimmt die Vitalkapazität tendenziell ab. Der Zusammenhang ist genähert linear.

2.1 Das Modell

Der Zusammenhang zwischen einer erklärenden Variablen x und der Zielvariablen Y wird folgendermassen beschrieben:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{2.1}$$

Y_i ist die Zielvariable der i-ten Beobachtung.

x_i ist die erklärende Variable der i-ten Beobachtung. Die Variable x_i wird als feste, nicht zufällige Grösse betrachtet.

β_0, β_1 sind unbekannte *Parameter*, die sog. Regressionskoeffizienten. Diese sollen mit Hilfe der vorhandenen Daten geschätzt werden.

ϵ_i ist der *zufällige Rest* oder *Fehler*, d. h. die zufällige Abweichung von Y_i von der Geraden. Es wird vorausgesetzt, dass der Erwartungswert $E(\epsilon_i) = 0$ und die Varianz $Var(\epsilon_i) = \sigma^2$ ist und dass die ϵ_i unkorreliert sind: $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$.

Das Modell (2.1) heisst *einfach*, weil nur eine erklärende Variable im Modell enthalten ist. Es heisst **nicht** linear, weil eine Gerade angepasst werden soll. Das Wort *linear* bezieht sich auf die Regressionsparameter, d. h. die Gleichung (2.1) ist linear in β_0 und β_1. Das bedeutet, dass zum Beispiel auch $y = \beta_0 + \beta_1 x^2 + \epsilon$ ein einfaches lineares Modell ist.

Die Zielgrösse Y_i ist dann eine Zufallsvariable mit

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i \\ Var(Y_i) &= Var(\beta_0 + \beta_1 x_i + \epsilon_i) = \sigma^2 \end{aligned}$$

Y_i und Y_j sind unkorreliert für $i \neq j$.

Zur Erinnerung: Rechnen mit Erwartungswerten, Varianzen und Covarianzen

Seien X und Y Zufallsvariablen, a, b, c und d Konstanten. Dann gilt:

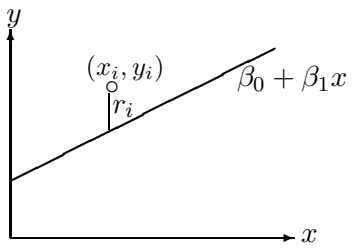
$$\begin{aligned} &E(a + bX) = a + bE(X) \\ &Var(a + bX) = b^2 Var(X) \\ &Cov(a + bX, c + dY) = bdCov(X, Y) \end{aligned}$$

2.2 Parameterschätzungen

Welche Gerade beschreibt die n Wertepaare am besten? Für jeden Punkt (x_i, y_i) betrachten wir die vertikale Abweichung von der Geraden $\beta_0 + \beta_1 x$:

$$r_i = y_i - (\beta_0 + \beta_1 x_i)$$

Die r_i heissen *Residuen* und sollen möglichst klein sein.



Wir bestimmen nun diejenige Gerade, d. h. $\hat{\beta}_0$ und $\hat{\beta}_1$, für die die Quadratsumme der Residuen

$$Q(\beta_o, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \quad \text{minimal wird.}$$

Dieses Verfahren heisst *Methode der Kleinsten Quadrate (Least Squares Method)*. Man erhält $\hat{\beta}_0$ und $\hat{\beta}_1$, indem man $Q(\beta_0, \beta_1)$ nach β_0 und nach β_1 ableitet, die beiden Ableitungen gleich Null setzt und nach β_0 und β_1 auflöst:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i)) \quad = 0 \\ \frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))x_i = 0 \end{aligned}$$

Umformen ergibt die *Normalgleichungen*:

$$\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \tag{2.2}$$

Die Lösung ist:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{und} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.3}$$

Wir erhalten daraus die Regressionsgerade (Least squares fit): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

$\hat{y}$ ist der vom Modell geschätzte Wert der Zielgrösse (*fitted or predicted value*) zu einem gegebenen x. Die Residuen r_i sind die Differenzen zwischen beobachtetem und geschätztem Wert von y, $r_i = y_i - \hat{y}_i$.

Statt der Quadratsumme könnte auch die Summe der absoluten Abweichungen $\sum |r_i|$, die sogenannte L_1-Norm, minimiert werden. Das entsprechende Verfahren ist robuster gegenüber extremen y-Werten.

Beispiel:
In unserem Beispiel erhalten wir die folgenden LS-Schätzungen:

$$\hat{\beta}_0 = 6.54 \quad \hat{\beta}_1 = -0.054$$

Die Abbildung 2.2 zeigt nochmals die 40 Beobachtungen, zusammen mit der Regressionsgeraden.

Aufgabe 2.1

a) Wie sieht die Gleichung der Regressionsgerade aus?
b) Wie gross ist die erwartete Abnahme der Vitalkapazität pro 10 Jahre?
c) Wie hoch schätzen Sie die mittlere Vitalkapazität von 40-jährigen Arbeitern?

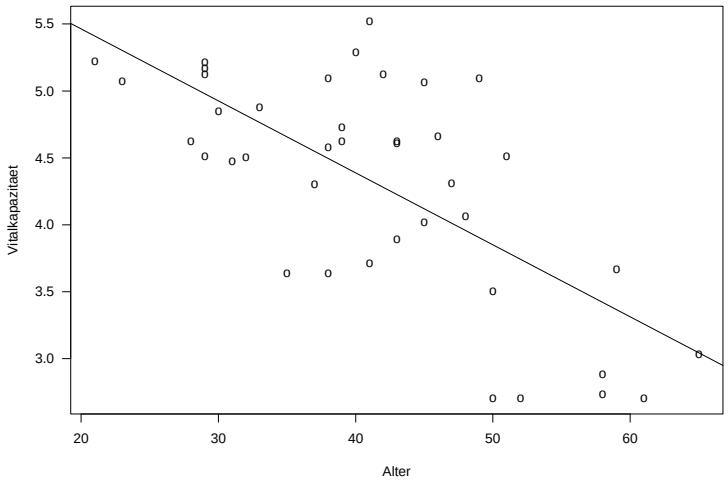


Abbildung 2.2: Lungenfunktionsmessungen von Cadmium-Arbeitern

Eigenschaften der LS-Schätzer

Gute Gründe sprechen für die Wahl der Kleinsten-Quadrate-Methode. Das **Gauss-Markov-Theorem** besagt, dass unter den Bedingungen von Modell (2.1) $\hat{\beta}_0$ und $\hat{\beta}_1$ erwartungstreu sind, d. h. $E(\hat{\beta}_0) = \beta_0$ und $E(\hat{\beta}_1) = \beta_1$, und unter allen erwartungstreuen, linearen Schätzern minimale Varianz haben.

Man kann zeigen, dass

$$\begin{aligned} Var(\hat{\beta}_0) &= \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}) \\ Var(\hat{\beta}_1) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \end{aligned} \tag{2.4}$$

Schätzung von σ^2

Neben den Regressionsparametern ist auch noch σ^2, die Varianz der zufälligen Fehler, zu schätzen. Eine solche Schätzung wird für alle möglichen Tests und Vertrauensintervalle benötigt. Eine unverzerrte Schätzung basiert auf der Quadratsumme der Residuen $SSE = \sum r_i^2 = \sum(y_i - \hat{y}_i)^2$. Die Abkürzung SSE steht für *error sum of squares*. Als

Schätzung für σ^2 verwendet man

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2} = MSE, \tag{2.5}$$

die mittlere Residuenquadratsumme (*mean squares of errors*).

2.3 Tests und Vertrauensintervalle

Bis jetzt haben wir keinerlei Verteilungsannahmen für die zufälligen Fehler ϵ_i gemacht. Um zu testen, ob die Variable x einen *signifikanten Einfluss* hat auf die Zielvariable Y, und um Vertrauensintervalle, resp. Prognoseintervalle zu konstruieren, brauchen wir aber jetzt eine Verteilungsannahme. Wir setzen im folgenden voraus, dass die ϵ_i normalverteilt, d. h. $\epsilon_i \sim N(0, \sigma^2)$, und unabhängig sind.

Das Modell kann nun so geschrieben werden:

$$\begin{aligned} &Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \\ &Y_i \text{ und } Y_j \text{ sind unabhängig für } i \neq j \end{aligned} \tag{2.6}$$

Die LS-Schätzer sind unter Annahme der Normalverteilung identisch mit den Maximum-Likelihood-Schätzern. Die ML-Methode wird bei den Modelltypen mit nichtstetigen Zielvariablen verwendet (Bsp: logistische Regression).

Um zu entscheiden, ob ein linearer Zusammenhang besteht zwischen x und Y, testet man die Nullhypothese $H_0 : \beta_1 = 0$. Testgrösse für den allgemeinen Fall $H_0 : \beta_1 = \beta$ gegen $H_A : \beta_1 \neq \beta$ ist

$$t^* = \frac{\hat{\beta}_1 - \beta}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta}{\sqrt{\hat{\sigma}^2 / \sum(x_i - \bar{x})^2}} \tag{2.7}$$

t^* hat eine t-Verteilung mit $n-2$ Freiheitsgraden. Die Grösse $se(\hat{\beta}_1)$ ist die geschätzte Standardabweichung von $\hat{\beta}_1$ und heisst *Standardfehler (standard error)* von $\hat{\beta}_1$.

Ein 95%-Vertrauensintervall für β_1 ist:

$$\hat{\beta}_1 \pm t_{97.5\%, n-2} \cdot \sqrt{\hat{\sigma}^2 / \sum(x_i - \bar{x})^2} \tag{2.8}$$

Tests und Vertrauensintervalle für β_0 werden analog konstruiert.

Aufgabe 2.2

a) Die Genauigkeit der Schätzung $\hat{\beta}_1$ hängt von den x-Werten ab. Welche Wahl von x-Werten gibt die effizienteste Schätzung? Konkret: Wenn Sie 40 Arbeiter beliebigen Alters untersuchen können, welche Altersverteilung wählen Sie?

b) Der t-Test für $H_0 : \beta_1 = 0$ ist nicht signifikant ausgefallen. Was schliessen Sie daraus?

Varianzanalyse-Tabelle

Der Computer-Output einer Regressionsanalyse enthält in der Regel neben den geschätzten Koeffizienten (inkl. Standardfehlern und t-Tests) eine Varianzanalyse-Tabelle (anova table). Diese Tabelle basiert auf der Zerlegung der Quadratsummen:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST &= SSR \quad + \quad SSE \end{aligned} \tag{2.9}$$

SST heisst *total sum of squares*, SSR *regression sum of squares* und SSE, das wir schon früher angetroffen haben, *error sum of squares*.

Dividiert man die sum of squares durch die entsprechende Anzahl Freiheitsgrade, dann erhält man die *mean squares* und daraus den *F-Test* mit der Teststatistik:

$$F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} \tag{2.10}$$

F^* hat unter $H_0 : \beta_1 = 0$ eine F-Verteilung mit 1 und $n-2$ Freiheitsgraden und ist im Falle einer einfachen linearen Regression gleich dem quadrierten Wert der t-Statistik. Grosse Werte von F^* sprechen gegen H_0, d. h. der Test ist einseitig.

All das wird in der Anova-Tabelle zusammengefasst:

Source of Variation	Sum of squares	Degrees of Freedom	Mean square	F^*
Regression	SSR	1	MSR	MSR/MSE
Residual	SSE	$n-2$	MSE	
Total	SST	$n-1$		

Neben dem Wert der F- oder t-Statistik wird oft auch das *Bestimmtheitsmass* R^2 angegeben. Das ist der Anteil an der Gesamtvariabilität, der „durch die Regression erklärt wird“:

$$R^2 = 1 - \frac{SSE}{SST} \tag{2.11}$$

Es gilt $R^2 = r^2$, wobei r die Korrelation zwischen x und y ist. Bei der Interpretation von R^2 ist deshalb die gleiche Vorsicht geboten wie bei r.

Beispiel: Berechnung mit dem Statistikprogramm R

```
# Daten einlesen
> library(foreign)
> lung=read.dta("D:/Kurse/biostat/Kurs2a/lung.dta")

# Daten anschauen/kontrollieren
> summary(lung)
      age             vit             exp
 Min.   :21.00   Min.   :2.700   Min.   :0.0
 1st Qu.:32.75   1st Qu.:3.700   1st Qu.:0.0
 Median :41.00   Median :4.545   Median :0.0
 Mean   :41.38   Mean   :4.315   Mean   :0.3
 3rd Qu.:48.25   3rd Qu.:5.063   3rd Qu.:1.0
 Max.   :65.00   Max.   :5.520   Max.   :1.0

> lung$exp=factor(lung$exp)
> levels(lung$exp)=c("10 Jahre und mehr", "weniger als 10 Jahre")
> summary(lung)
      age             vit                              exp
 Min.   :21.00   Min.   :2.700   10 Jahre und mehr    :28
 1st Qu.:32.75   1st Qu.:3.700   weniger als 10 Jahre :12
 Median :41.00   Median :4.545
 Mean   :41.38   Mean   :4.315
 3rd Qu.:48.25   3rd Qu.:5.063
 Max.   :65.00   Max.   :5.520


> plot(vit~age,data=lung)             # Plot wie Abbildung 2.1

# Einfache lineare Regression rechnen
> mod1=lm(vit~age,data=lung)
> mod1

Call:
lm(formula = vit ~ age, data = lung)

Coefficients:
(Intercept)          age
    6.53915     -0.05376
```

```
> summary(mod1)

Call:
lm(formula = vit ~ age, data = lung)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15136 -0.40553  0.03428  0.32242  1.18489

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.539152   0.388926  16.813  < 2e-16 ***
age         -0.053756   0.009112  -5.899 7.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6042 on 38 degrees of freedom
Multiple R-Squared: 0.478,      Adjusted R-squared: 0.4643
F-statistic:  34.8 on 1 and 38 DF,  p-value: 7.821e-07

> names(mod1)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"


> names(summary(mod1))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"


> plot(lung$age,lung$vit)
> abline(mod1)              # Plot wie Abbildung 2.2
```

Beantworten Sie die folgenden Fragen mit Hilfe des obigen Computer-Outputs.

a) Wie gross sind $\hat{\beta}_0$, $\hat{\beta}_1$ und $\hat{\sigma}^2$?
b) Ist der lineare Zusammenhang signifikant? Wie gross ist die Teststatistik?
c) Wie gross ist die Korrelation zwischen Alter und Vitalkapazität?

2.4 Prognosebereiche

Auf Seite 10 haben Sie die mittlere Vitalkapazität eines 40jährigen Arbeiters geschätzt. Eine Möglichkeit ist $\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 6.539 - 0.0538 \cdot 40 = 4.387$. Wie gut ist diese Schätzung?

Das Vertrauensintervall für $\beta_0 + \beta_1 x_0$ ist

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{97.5\%,n-2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{2.12}$$

Da das für alle x_0 gilt, können wir um die Regressionsgerade $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ ein Band einzeichnen und dann das Vertrauensintervall für beliebige x direkt ablesen. Dieses Band ist in der Mitte schmaler als an den Rändern.

Die Frage, in welchem Bereich eine neue Beobachtung Y_0 liegt, ist damit allerdings noch nicht beantwortet. Wir haben ja erst ein Vertrauensintervall für den erwarteten Wert von Y an der Stelle x_0 berechnet. Die einzelnen Beobachtungen streuen noch zusätzlich um den mittleren Wert herum.

Das *Prognoseintervall* für Y_0 ist

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{97.5\%,n-2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{2.13}$$

Zur Bezeichnung: Ein Vertrauensintervall gibt einen Bereich für einen Parameter an, ein Prognoseintervall einen Bereich für eine Zufallsvariable.

Prognose- und Vertrauensbereich sind in Abbildung 2.3 eingezeichnet.

Eine Prognose ausserhalb des x-Bereichs, für den Beobachtungen vorliegen, ist gefährlich.

2.5 Residuenanalyse

In jeder Regressionanalyse müssen nach der Schätzung die Modellannahmen überprüft werden. Diese sind:

- Der Zusammenhang zwischen y und x ist genähert linear.
- Die Fehler ϵ_i haben Erwartungswert 0.
- Die Fehler ϵ_i haben konstante Varianz σ^2.
- Die Fehler ϵ_i sind unkorreliert.
- Die Fehler ϵ_i sind normalverteilt.

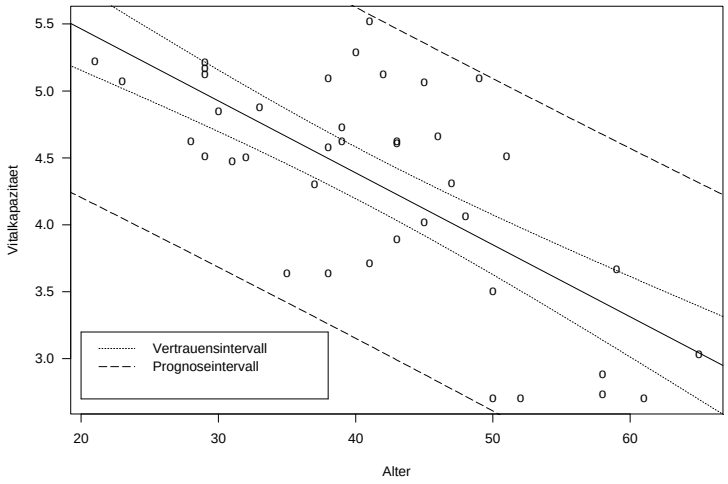


Abbildung 2.3: Vertrauens- und Prognosebereich

Die Überprüfung geschieht am einfachsten mit Hilfe von Residuenplots. Dabei werden die Residuen $r_i = y_i - \hat{y}_i$ gegen verschiedene andere Variablen graphisch dargestellt. Wenn Abweichungen von den Annahmen gefunden werden, so führt das im Idealfall zu einer Verbesserung des Modells. Danach folgt wieder die Parameterschätzung, die Modellüberprüfung, usw. Meist sind mehrere Durchgänge nötig, bis man bei einem befriedigenden Resultat angekommen ist.

In der einfachen linearen Regression werden die meisten Verletzungen von Modellannahmen schon im Streudiagramm y gegen x sichtbar. Die verschiedenen Plots sind deshalb vor allem in der multiplen Regression nützlich.

Normal Plot

Mit einem Normalplot der Residuen kann man die Normalverteilungsannahme überprüfen. Dabei plottet man die geordneten Residuen gegen die entsprechenden Quantile der Normalverteilung. Wenn die Fehler ϵ_i normalverteilt sind, dann sind das auch die Residuen r_i. Die Punkte im Normalplot sollten demnach ungefähr auf einer Geraden liegen.

Die Residuen haben im Gegensatz zu den Fehlern aber nicht konstante Varianz und sie sind korreliert. Wenn n klein ist, arbeitet man deshalb oft mit *standardisierten Residuen*. Leider wird diese Bezeichnung unterschiedlich verwendet.

Abbildung 2.4 zeigt typische Abweichungen.

Wenn der Normalplot eine schiefe Verteilung aufzeigt, hilft meist eine Transformation

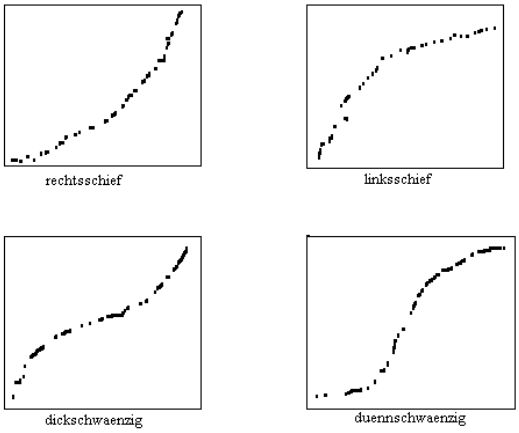


Abbildung 2.4: Normalplots

der Zielgrösse. Am häufigsten verwendet werden Logarithmus- und Wurzeltransformation. Relativ oft zeigt der Plot auch ein paar Beobachtungen mit extrem grossen oder kleinen Residuen. Bei solchen Ausreissern muss zunächst abgeklärt werden, ob es sich um grobe Fehler handelt (z. B. Abschreibfehler). Mit Hilfe von sogenannten *diagnostics* kann der Einfluss einer einzelnen Beobachtung auf die Schätzungen und Tests studiert werden.

Plot von r_i gegen $\hat{y}_i$

Mit diesem Plot können verschiedene Verletzungen von Modellannahmen entdeckt werden. Im Idealfall befinden sich alle Residuen in einem horizontalen Band konstanter Breite wie das in Abbildung 2.5 (a) dargestellt ist.

Bei ungleichen Varianzen wie in Abbildung 2.5 (b) und (c) hilft entweder eine Transformation oder es muss eine *gewichtete Regression* durchgeführt werden. Auch bei Nichtlinearität ist eine Transformation vielleicht hilfreich oder das Modell muss durch quadratische Terme oder andere Variablen verbessert werden.

Plot von r_i gegen x_i

Diese Plots können ähnlich aussehen wie die vorherigen. Wieder zeigen sich hier ungleiche Varianzen, diesmal in Abhängigkeit von der Grösse von x, und Nichtlinearität. Ist letzteres der Fall, hilft vielleicht ein quadratischer Term.

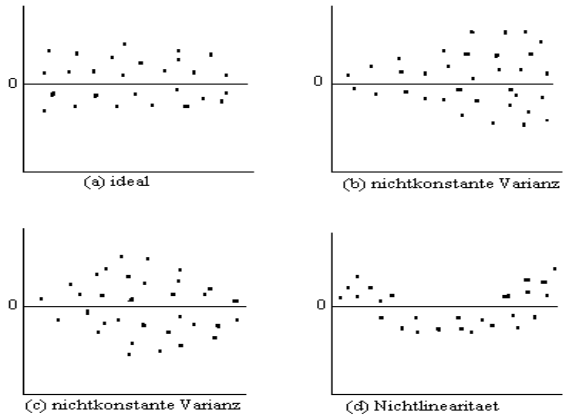


Abbildung 2.5: Plot von r_i gegen $\hat{y_i}$

Plot von r_i gegen i

Wenn der Index zum Beispiel der zeitlichen Reihenfolge entspricht, in der die Beobachtungen gemacht worden sind, dann kann dieser Plot korrelierte Fehler aufzeigen. In diesem Fall sind spezielle Methoden notwendig.

2.6 Transformationen

Nichtlinearität kann im Residuenplot entdeckt werden. Manchmal ist auch aus theoretischen Gründen ein anderes Modell vorzuziehen, z. B. weil eine relative statt absoluten Veränderung in y in Abhängigkeit von x mehr Sinn macht.

Viele funktionelle Zusammenhänge sind mit Hilfe von geeigneten Transformationen linearisierbar:

a)	$y = \beta_0 x^{\beta_1}$	$y' = log(y), x' = log(x) \Longrightarrow y' = log(\beta_0) + \beta_1 x'$
b)	$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y) \Longrightarrow y' = ln\beta_0 + \beta_1 x$
c)	$y = \beta_0 + \beta_1 log(x)$	$x' = log(x) \Longrightarrow y = \beta_0 + \beta_1 x'$
d)	$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = 1/y, x' = 1/x \Longrightarrow y' = \beta_0 - \beta_1 x'$

Das nichtlineare Modell $y = \beta_0 e^{\beta_1 x} \epsilon$ kann also mit passenden Transformationen in ein lineares Modell überführt werden.

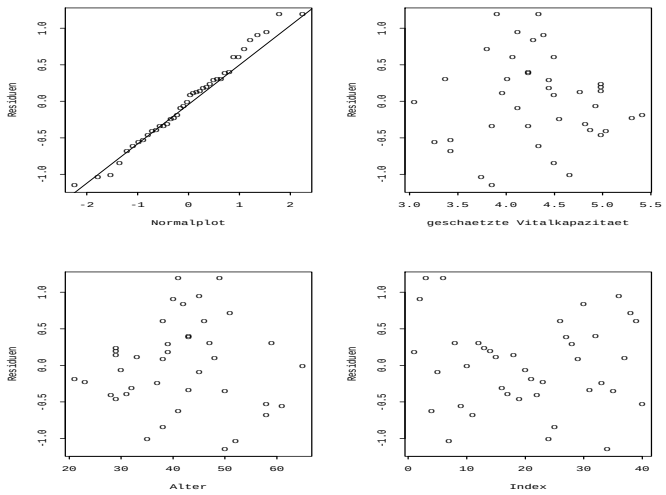


Abbildung 2.6: Residuenplots für die Cadmiumarbeiter

Varianzstabilisierende Transformationen

Manchmal ist die Voraussetzung, dass die Y_i konstante Varianz haben, verletzt. Wenn die Zielgrösse beispielsweise poissonverteilt ist, dann gilt $E(Y) = Var(Y)$, d. h. Var(y) wächst oder fällt mit x. Gesucht ist eine Transformation, die zu konstanter Varianz führt.

Sei Y eine Zufallsvariable und $Z = g(Y)$ mit einer festen Funktion g. Betrachte die Taylorapproximation von Z an der Stelle μ_Y:

$$Z = g(Y) \approx g(\mu_Y) + (Y - \mu_Y)g'(\mu_Y)$$

Dann gilt für Erwartungswert und Varianz von Z genähert:

$$\begin{aligned} \mu_Z &\approx g(\mu_Y) \\ \sigma_Z^2 &\approx \sigma_Y^2[g'(\mu_Y)^2] \end{aligned} \tag{2.14}$$

Nun wird g so gewählt, dass $\sigma_Y^2[g'(\mu_Y)^2]$ konstant wird.

Wenn Y poissonverteilt ist, muss also $\lambda \cdot g'(\lambda)^2$ konstant sein, d. h. $g(Y) = \sqrt{}(Y)$ ist eine passende Transformation.

Transformationen können auch analytisch bestimmt werden. Für die *Box-Cox-Transformationen* Y^λ kann der Parameter λ gleichzeitig mit den Regressionskoeffizienten geschätzt werden. $\lambda = 0$ bedeutet dabei die Logarithmus-Transformation.

3 Multiple lineare Regression

- Wie wird der Einfluss von mehreren Variablen gleichzeitig untersucht?
- Welche Tests sind sinnvoll?
- Was sind Ausreisser und einflussreiche Beobachtungen?

Beispiel:
Um den Einfluss der Luftverschmutzung auf die allgemeine Mortalität zu untersuchen, wurden in einer US-Studie (finanziert von General Motors) Daten aus 60 verschiedenen Regionen zusammengetragen. Neben der altersstandardisierten Mortalität und der Belastung durch CO, NOx und SO_2 wurden verschiedene demographische und meteorologische Variablen erfasst.

Eine einfache lineare Regression von Mortalität auf SO_2 zeigt, dass mit zunehmender SO_2-Konzentration die allgemeine Sterblichkeit signifikant ansteigt. Aber auch der Zusammenhang zwischen Mortalität und allgemeinem Bildungsstand, Bevölkerungsdichte, %-Nichtweisse, Einkommen, Niederschlagsmenge ist jeweils signifikant.

Statt viele einzelne einfache Regressionen zu rechnen, ist es besser, den Zusammenhang mit mehreren erklärenden Variablen gleichzeitig zu untersuchen.

3.1 Das Modell

Das multiple lineare Regressionsmodell beschreibt den Zusammenhang zwischen einer Zielvariablen Y und p erklärenden Variablen $x_1, \ldots, x_p$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad i = 1, \ldots, n \tag{3.1}$$

Y_i ist die Zielvariable der i-ten Beobachtung.

$x_{i1}, \ldots, x_{ip}$ sind die erklärenden Variablen der i-ten Beobachtung. Sie werden als feste, nicht zufällige Grössen betrachtet.

$\beta_0, \ldots, \beta_p$ sind unbekannte Parameter, die sog. Regressionskoeffizienten. Diese sollen mit Hilfe der vorhandenen Daten geschätzt werden.

ϵ_i ist der *zufällige Rest* oder *Fehler*. Es wird vorausgesetzt, dass $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ und $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$.

Für Tests und Vertrauensintervalle wird zudem angenommen, dass die ϵ_i normalverteilt sind. Dann gilt $Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$ und $Cov(Y_i, Y_j) = 0$ für $i \neq j$.

In der Luftverschmutzungsstudie ist zum Beispiel:

Y_i die altersstandardisierte Mortalität (Anzahl Todesfälle pro 100'000 Einw.) in der Region i

x_{i1} die mittlere SO_2- Konzentration in der Region i

x_{i2} der Anteil der nichtweissen Population in der Region i

x_{i3} die mittlere jährliche Niederschlagsmenge (in inches) in der Region i

Die Abbildung 3.1 zeigt den Zusammenhang zwischen y und den erklärenden Variablen in Streudiagrammen.

Die SO_2-Werte sind ziemlich schief verteilt und der Zusammenhang mit y sieht nicht gerade linear aus. Eine Logarithmus-Transformation nützt.

Die Regressionsgleichungen der drei einfachen Regressionen und der multiplen Regression sehen folgendermassen aus:

$$\begin{aligned}
\hat{y} &= 886.85 + 16.73 \cdot \log SO_2 \\
\hat{y} &= 887.06 + 4.49 \cdot \%\text{-Nichtweisse} \\
\hat{y} &= 849.53 + 2.37 \cdot \text{Niederschlag} \\
\hat{y} &= 776.22 + 16.9 \cdot \log SO_2 + 3.66 \cdot \%\text{-Nichtweisse} + 1.73 \cdot \text{Niederschlag}
\end{aligned}$$

Wie sind die geschätzten Regressionskoeffizienten $\hat{\beta}_j$ zu interpretieren? Die Schätzungen für dieselbe Variable sind verschieden in der einfachen und in der multiplen Regression. Hat eine Zunahme der nichtweissen Bevölkerung um 10% dieselbe Auswirkung auf die Sterblichkeit wie eine Zunahme der nichtweissen Bevölkerung um nur 5%, zusammen mit 10 inches mehr Regen?

Nein. Der Regressionskoeffizient gibt die Veränderung in Y bei einem Anstieg von x_j um eine Einheit an, vorausgesetzt alle andern Variablen bleiben konstant. Die Sprechweise „... unter Berücksichtigung der anderen Variablen ... " ist nicht ganz eindeutig.

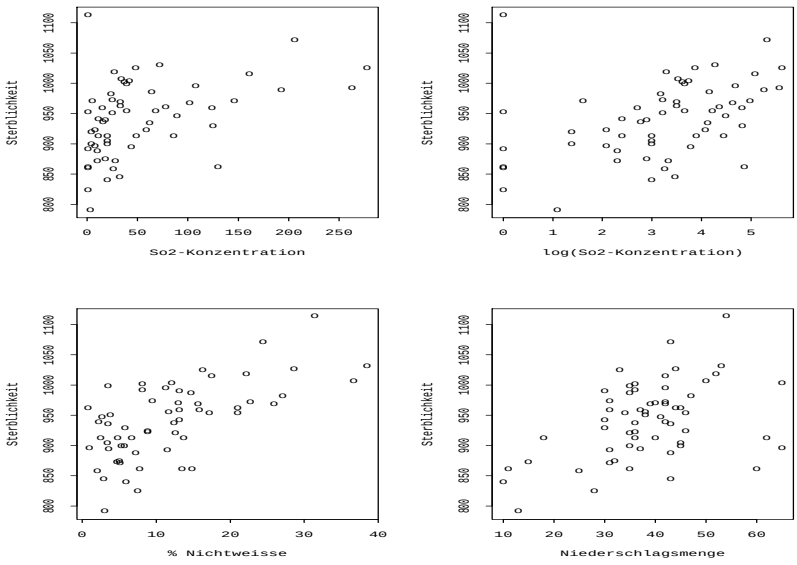


Abbildung 3.1: Luftverschmutzung und Mortalität

Die Methode der kleinsten Quadrate kann verallgemeinert werden für mehrere erklärende Variable. Gesucht sind $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ so, dass die Quadratsumme der Residuen

$$Q(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2$$

minimal wird.

Man erhält die Lösung, indem man Q nach $\beta_0, \beta_1, \ldots, \beta_p$ ableitet und die Ableitungen gleich Null setzt. Das ergibt nicht nur zwei, wie bei der einfachen linearen Regression, sondern $p+1$ Normalgleichungen in $p+1$ Unbekannten:

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \ldots + \beta_p x_{ip})) = 0 \\
\frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \ldots + \beta_p x_{ip})) x_{i1} = 0 \\
&\vdots \quad \vdots \\
\frac{\partial Q}{\partial \beta_p} &= -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \ldots + \beta_p x_{ip})) x_{ip} = 0
\end{aligned}$$

Das bereits erwähnte Gauss-Markov-Theorem gilt auch im mehrdimensionelen Fall: die LS-Schätzungen sind erwartungstreu und haben unter allen lineare, erwartungstreuen Schätzern minimale Varianz. Unter Annahme der Normalverteilung fallen die LS-Schätzer mit den Maximum-Likelihood-Schätzern zusammen.

Das Lösen der Gleichungssysteme und die Berechnung von Teststatistiken und Vertrauensintervallen ist ohne weitere algebraische Hilfsmittel ziemlich mühsam und die Ergebnisse der Rechnungen können fast nicht lesbar aufgeschrieben werden. In jedem ausführlicheren Text über multiple Regression (auch in Software-Manuals) finden Sie deshalb die entsprechenden Resultate in Matrixschreibweise.

Matrixalgebra stellt aber nicht nur eine elegante Schreibweise zur Verfügung, sondern ermöglicht auch das Verständnis für viele theoretische und praktische Schwierigkeiten in der multiplen Regression und der multivariaten Statistik überhaupt. Wir stützen uns deshalb im folgenden auf die einfachsten Resultate der Matrixalgebra. Eine Zusammenstellung befindet sich in Anhang A.

Das Regressionsmodell in Matrixschreibweise

Mit Hilfe von Matrizen können wir das multiple Regressionsmodell (3.1) so schreiben:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

$\mathbf{y}$ ist der Zielvariablenvektor der Länge n.

$\mathbf{X}$ ist die Designmatrix der Dimension $n \times (p+1)$. In den Spalten von $\mathbf{X}$ stehen die erklärenden Variablen. $\mathbf{X}$ ist fest.

$\boldsymbol{\beta}$ ist der Parametervektor der Länge $(p+1)$. Dieser soll mit Hilfe der vorhandenen Daten geschätzt werden.

$\boldsymbol{\epsilon}$ ist der Fehlervektor. Es wird vorausgesetzt, dass $E(\boldsymbol{\epsilon}) = \mathbf{0}$ und $Cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

Für Tests und Vertrauensintervalle wird zudem angenommen, dass die ϵ_i normalverteilt sind. Dann gilt $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Die Normalgleichungen (3.2) sehen jetzt so aus:

$$\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$$

oder

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y} \tag{3.3}$$

Die Normalgleichungen haben genau dann eine eindeutige Lösung, wenn die Matrix $\mathbf{X}^t\mathbf{X}$ invertierbar ist, d. h. wenn alle Spalten von $\mathbf{X}$ linear unabhängig sind. Es darf also keine erklärende Variable Linearkombination der übrigen Variablen sein. Eine notwendige Bedingung für die Invertierbarkeit von $\mathbf{X}^t\mathbf{X}$ ist $p < n$.

Multiplizieren wir beide Seiten mit dem Inversen von $\mathbf{X}^t\mathbf{X}$, so erhalten wir die Least-Squares-Schätzungen:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{3.4}$$

Es ist $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ und $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ und unter Annahme der Normalverteilung ergibt sich $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$.

Wenn man $SSE = \sum r_i^2 = (\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}})$ durch die Anzahl Freiheitsgrade dividiert (Anzahl Beobachtungen − Anzahl geschätzte Parameter), dann erhält man eine erwartungstreue Schätzung für σ^2 :

$$\hat{\sigma}^2 = \frac{SSE}{n - p - 1} = MSE \tag{3.5}$$

Die geschätzten Werte $\hat{\mathbf{y}}$ bekommt man durch eine Matrixmultiplikation aus den beobachteten $\mathbf{y}$-Werten:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{H}\mathbf{y} \tag{3.6}$$

$\mathbf{H}$ heisst *Hat-Matrix* : sie setzt dem y einen Hut auf. Die Residuen $\mathbf{r}$ lassen sich ebenfalls mit Hilfe der Hat-Matrix schreiben:

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Beispiel: Einfache lineare Regression
Es ist

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{und} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

sowie

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{und} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Für die Normalgleichungen brauchen wir $\mathbf{X}^t\mathbf{X}$ und $\mathbf{X}^t\mathbf{y}$:

$$\begin{aligned}
\mathbf{X}^t\mathbf{X} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \\
\mathbf{X}^t\mathbf{y} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}
\end{aligned}$$

Die Normalgleichungen $\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y}$ sind also:

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

oder ausgeschrieben:

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i &= \sum y_i \\
\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 &= \sum x_i y_i
\end{aligned}$$

Das entspricht den Normalgleichungen, die wir im Kapitel 2 (siehe Seite 10) angegeben haben.

Das Inverse von $\mathbf{X}^t\mathbf{X}$ ist:

$$(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{n\sum(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

und wir erhalten:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \\
&= \frac{1}{n\sum(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\
&= \frac{1}{n\sum(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n \sum x_i y_i \end{pmatrix}
\end{aligned}$$

Vereinfachen und umformen führt zu den LS-Schätzern (2.3) von Kapitel 2.

3.2 Tests und Vertrauensintervalle

Für alle Tests und Vertrauensintervalle setzen wir normalverteilte Fehler voraus. Die Resultate werden wie bei der einfachen linearen Regression in einer Anova-Tabelle (siehe Seite 13) zusammengefasst, ergänzt mit den einzelnen Koeffizientenschätzungen und Standardfehlern.

Die Anova-Tabelle:

Source of Variation	Sum of squares	Degrees of Freedom	Mean square	F^*
Regression	$SSR = \sum(\hat{y}_i - \bar{y})^2$	p	MSR	MSR/MSE
Residual	$SSE = \sum(y_i - \hat{y}_i)^2$	$n-1-p$	MSE	
Total	$SST = \sum(y_i - \bar{y})^2$	$n-1$		

Globaler F-Test

Als erstes soll geprüft werden, ob insgesamt ein Zusammenhang besteht mit den erklärenden Variablen. Die entsprechende Teststatistik steht in der letzten Spalte der obigen Anova-Tabelle. Mit $F^* = MSR/MSE$ wird die Nullhypothese

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

gegen die Alternativhypothese

$$H_A : \text{ mindestens ein } \beta_j \neq 0$$

getestet. F^* hat unter H_0 eine F-Verteilung mit p und $n-1-p$ Freiheitsgraden. H_0 wird verworfen, wenn F^* grösser als das 95%-Perzentil der entsprechenden F-Verteilung ist.

Bestimmtheitsmass R^2

Das *multiple Bestimmheitsmass* ist wie im einfachen Fall definiert:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Es ist der Anteil der Variabilität der y-Werte, der durch die Regression erklärt wird. Es gilt: $0 \leq R^2 \leq 1$, wobei $R^2 = 0$, wenn $\beta_j = 0$ für alle j, und $R^2 = 1$, wenn alle Residuen gleich Null sind („perfekter Fit“).

Wenn mehr erklärende Variablen ins Modell aufgenommen werden, kann R^2 nur grösser werden, niemals kleiner. Deshalb betrachtet man oft eine korrigierte Version von R^2, die die Anzahl erklärender Variablen im Modell berücksichtigt. Das *adjusted R-squared* ist definiert als

$$adjR^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\frac{SSE}{SST} \tag{3.7}$$

Tests von individuellen Parametern

Da $\hat{\boldsymbol{\beta}}$ und somit $\hat{\beta}_j$ für alle j normalverteilt sind, kann

$$H_0 : \beta_j = 0 \quad \text{gegen} \quad H_A : \beta_j \neq 0$$

mit einem t-Test getestet werden. Die Teststatistik

$$t^* = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^t\mathbf{X})^{-1}_{jj}}}$$

hat eine t-Verteilung mit $n - p - 1$ Freiheitsgraden.

Die Frage, die mit diesem Test beantwortet wird, lautet, ob die erklärende Variable x_j einen signifikanten Zusammenhang mit y hat, *gegeben alle andern Variablen*. Ob x_j für sich allein einen Zusammenhang mit y hat, wird in einer einfachen Regression untersucht.

Vertrauens- und Prognosebereiche

Ein 95%-Vertrauensintervall für β_j ist gegeben durch:

$$\hat{\beta}_j \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{(\mathbf{X}^t\mathbf{X})^{-1}_{jj}} \tag{3.8}$$

Die Wahrscheinlichkeit, dass alle Regressionsparameter gleichzeitig in den so berechneten Intervallen liegen, ist aber nicht mehr 95%, sondern wird mit zunehmender Parameterzahl immer kleiner. Man kann zwar einen *gemeinsamen* 95%-*Vertrauensbereich* für den Parametervektor $\boldsymbol{\beta}$ bestimmen, aber die Rechnung ist nicht ganz einfach. Eine andere Möglichkeit bietet die *Bonferroni-Regel*: Für einen Vertrauensbereich für g Parameter nimmt man in (3.8) das $100(1 - \alpha'/2)$. Perzentil der t-Verteilung mit $\alpha' = \alpha/g$.

Man kann auch ein Vertrauensintervall für den erwarteten Wert von y oder ein Prognoseintervall für eine zukünftige Beobachtung zu gegebenen $x_{01}, \ldots, x_{0p}$ berechnen.

Ein 95%-Vertrauensintervall für $E(y_0)$ ist gegeben durch

$$\hat{y_0} \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0} \tag{3.9}$$

und ein 95%-Prognoseintervall für eine zukünftige Beobachtung ist

$$\hat{y_0} \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{1 + \mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0} \tag{3.10}$$

wobei $\mathbf{x}_0^t = (1\, x_{01}\, x_{02}\, \cdots\, x_{0p})^t$.

Beispiel:
In der Luftverschmutzungsstudie haben wir eine Regression mit den erklärenden Variablen $\log(SO_2)$, %-Nichtweisse und Niederschlagsmenge gerechnet (siehe Seite 22). Der R–Output sieht folgendermassen aus:

```
Call: lm(formula = mort ~ log(so2) + nonwhite + rain, data = smsa)

Residuals:
     Min       1Q   Median       3Q      Max
-75.9671 -25.0296   0.5792  20.7507 128.3527

Coefficients:
            Estimate Std.Error tvalue Pr(>|t|)
(Intercept) 776.225  21.248  36.532  < 2e-16 ***
log(so2)     16.949   3.348   5.060 4.84e-06 ***
nonwhite      3.665   0.586   6.248 5.98e-08 ***
rain          1.732   0.456   3.796 0.000363 ***
---

Residual standard error: 38.17 on 56 df Multiple R-Squared: 0.6428,
Adjusted R-squared: 0.6237 F-statistic:  33.6 on 3 and 56 df,
p-value: 1.48e-012
```

Der globale F-Test und alle drei t-Tests sind signifikant, insbesondere ist auch der Koeffizient von $\log(SO_2)$ signifikant von Null verschieden. Ob Schwefeldioxid eine erhöhte Sterblichkeit verursacht, ist damit natürlich noch nicht entschieden. Je mehr andere erklärende Variablen, die einen Einfluss auf die Sterblichkeit haben, ins Modell einbezogen sind, desto stärker werden aber die Argumente für eine kausale Wirkung der Luftverschmutzung, weil dann die Signifikanz gilt unter Berücksichtigung aller andern erklärenden Variablen.

Rechnen wir also eine multiple Regression mit allen verfügbaren demographischen und meteorologischen Variablen:

jantemp	Mittlere Januar-Temperatur in Fahrenheit
julytemp	Mittlere Juli-Temperatur in Fahrenheit
relhum	Mittlere relative Luftfeuchtigkeit um 13 Uhr
rain	Mittlere jährliche Niederschlagsmenge in Inches
educ	Median der absolvierten Schuljahre aller über 25-Jährigen
dens	Bevölkerungsdichte pro Quadratmeile
nonwhite	Anteil der nichtweissen Bevölkerung in %
wc	Anteil “white-collar worker” in %
pop	Bevölkerung
house	Mittlere Anzahl Personen pro Haushalt
income	Median des Einkommens

```
Call: lm(formula = mort ~ educ + jantemp + julytemp + relhum +
rain + dens + nonwhite + wc + pop + house + income + log(so2),
data = smsa, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-70.915 -20.941  -2.773  18.859 105.931

Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept) 1.16e+03  2.94e+02  3.96  0.00026 ***
educ       -1.11e+01  9.45e+00 -1.17  0.24698
jantemp    -1.67e+00  7.93e-01 -2.10  0.04079 *
julytemp   -1.17e+00  1.94e+00 -0.60  0.55021
relhum      7.02e-01  1.11e+00  0.63  0.52864
rain        1.22e+00  5.49e-01  2.23  0.03074 *
dens        5.62e-03  4.48e-03  1.25  0.21594
nonwhite    5.08e+00  1.01e+00  5.02  8.3e-06 ***
wc         -1.93e+00  1.26e+00 -1.52  0.13462
pop         2.07e-06  4.05e-06  0.51  0.61180
house      -2.22e+01  4.04e+01 -0.55  0.58607
income      2.43e-04  1.33e-03  0.18  0.85562
log(so2)    6.83e+00  5.43e+00  1.26  0.21426
---

Residual standard error: 36.2 on 46 df
Multiple R-Squared: 0.733, Adjusted R-squared: 0.664
F-statistic: 10.5 on 12 and 46 df, p-value: 1.42e-009
```

Zunächst fällt auf, dass nur noch 59 Städte in die Analyse aufgenommen worden sind. Offenbar fehlen bei einer Stadt einzelne x-Werte. Die grosse Enttäuschung aber kommt weiter unten: der Zusammenhang mit $log(SO_2)$ ist nicht mehr signifikant, der P-Wert ist 0.214. Heisst das nun, dass die Luftverschmutzung mit SO_2 doch keine Auswirkung auf die Mortalität hat?

Bevor man irgendwelche voreiligen Schlüsse zieht, sollte man die Modellannahmen überprüfen (siehe Abschnitt 3.3).

Zudem ist die relative Bedeutung einzelner erklärender Variablen schwierig zu beurteilen, wenn die x-Variablen untereinander, oder mit andern Variablen, die im Modell fehlen, korreliert sind: sog. *Multicollinearität* der erklärenden Variablen. Sind die Variablen x_1 und x_2 unkorreliert, dann bleiben die Schätzungen für β_1, bzw. β_2 gleich,

unabhängig davon, ob die jeweils andere Variable im Modell ist. Bei korrelierten x-Variablen ändern sich die geschätzten Koeffizienten, je nachdem welche Variablen im Modell sind (vgl. Seite 22). Auch die Testergebnisse sind manchmal etwas verblüffend. Es kann vorkommen, dass der globale F-Test signifikant und alle einzelnen t-Tests nicht signifikant sind, weil eine einzelne Variable nicht mehr viel zusätzlich bringt, wenn die andern Variablen schon im Modell sind.

In einem kontrollierten Experiment wird man die Versuchsbedingungen so wählen, dass die erklärenden Variablen unkorreliert sind, bei beobachtenden Studien muss man mit den Korrelationen leben. Wir kommen im Kapitel über Modellwahl darauf zurück. Die Streudiagramme 3.2 zeigen die Variablenpaare mit den höchsten Korrelationen.

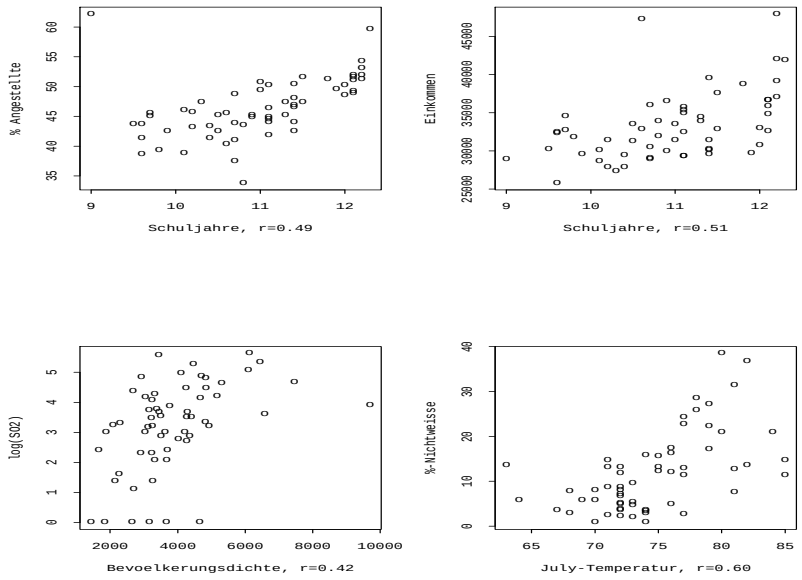


Abbildung 3.2: Luftverschmutzung und Mortalität

Partielle F-Tests

Statt auf einzelne Regressionskoeffizienten zu testen, kann man auch eine Gruppe von x-Variablen gemeinsam betrachten. In der Luftverschmutzungsstudie kann man zum Beispiel fragen, ob die meteorologischen Variablen insgesamt einen signifikanten Effekt haben.

Von den p erklärenden Variablen, wollen wir den Effekt von $p-q$ Variablen gemeinsam testen. Dazu partitioniert man den Parametervektor und die Designmatrix wie folgt:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \\ \beta_{q+1} \\ \vdots \\ \beta_p \end{pmatrix} \qquad \text{und} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix},$$

wobei $\mathbf{X}_1$ die Dimension $n \times (q+1)$ und $\mathbf{X}_2$ die Dimension $n \times (p-q)$ hat.

Das Modell kann dann geschrieben werden als

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

und das Testproblem ist

$$H_0: \ \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{gegen} \quad H_A: \ \boldsymbol{\beta}_2 \neq \mathbf{0}$$

in Worten: H_0:“ die $p-q$ Variablen haben keinen Effekt” gegen H_A: “mindestens eine der $p-q$ Variablen hat einen Effekt”.

Wenn man zwei multiple Regressionen rechnet, einmal mit allen p Variablen und nachher mit der reduzierten Auswahl von q Variablen , so erhält man zwei verschiedene Regressions-Summenquadrate, SSR_{H_A} für das *volle Modell* und SSR_{H_0} für das *reduzierte Modell*. Intuitiv ist klar, dass die Nullhypothese nicht verworfen werden kann, wenn die Differenz zwischen diesen beiden sum of squares “klein” ist.

Die entsprechende F-Statistik ist gleich

$$F^* = \frac{(SSR_{H_A} - SSR_{H_0})/(p-q)}{SSE_{H_A}/(n-p-1)} \tag{3.11}$$

Die Hypothese H_0 wird verworfen, wenn $F^* > F_{95\%,p-q,n-p-1}$.

Beispiel:
Testen wir, ob die vier meteorologischen Variablen Januartemperatur, Julitemperatur, relative Luftfeuchtigkeit und Niederschlagsmenge im Modell auf Seite 30 gemeinsam signifikant sind. Die Teststatistik hat den Wert $F^* = 2.92$ und unter H_0 eine F-Verteilung mit 4 und 46 Freiheitsgraden. Der P-Wert ist 0.031, d. h. H_0 wird verworfen, die meterologischen Variablen haben gemeinsam einen signifikanten Effekt auf die Mortalität.

Die besprochenen partiellen F-Tests können weiter verallgemeinert werden, um sogenannte *lineare Kontraste* zu testen . Damit können Linearkombinationen der Regressionskoeffizienten wie zum Beispiel $\beta_1 = \beta_2 = \beta_3$ getestet werden.

3.3 Modelldiagnostik

Residuenplots

Für die Überprüfung der Modellannahmen sehr nützlich sind die bereits bei der einfachen, linearen Regression besprochenen Residuenplots:

- Normalplot der Residuen r_i
- Residuen r_i gegen geschätzte y-Werte $\hat{y}_i$
- Residuen r_i gegen eine erklärende Variable x_i des Modells
- Residuen r_i gegen eine neue Variable x_i', die nicht im Modell ist
- Residuen r_i gegen den Index i

Ausreisser und einflussreiche Beobachtungen

Manchmal werden die Schätzungen der Koeffizienten in einer Regressionanalyse von ein paar wenigen Beobachtungen stark beeinflusst. Falls das so ist, möchte man natürlich diese Beobachtungen identifizieren. In den Residuenplots erkennt man aber die einflussreichen Beobachtungen (influential points) nur, wenn sie gleichzeitig auch Ausreisser (outlier) sind.

Eine dritte wichtige Kategorie in diesem Zusammenhang sind die Hebelpunkte (leverage points). Das sind Beobachtungen mit extremen x-Werten. Die Abbildung 3.3 illustriert ein paar verschiedenen Situationen. In (a) ist nichts besonderes los, in (b) ist ein Hebelpunkt ohne Einfluss, (c) enthält einen Hebelpunkt mit Einfluss und in (d) hat es einen Ausreisser ohne Einfluss.

Der Einfluss einer Beobachtung auf Schätzungen und Tests kann im Prinzip festgestellt werden, indem die Analyse ohne die fragliche Beobachtung gemacht wird. In der Praxis ist es aber nicht nötig, die Analyse n-mal zu wiederholen. Mit ein paar rechnerischen Kniffs können die wichtigen Grössen ohne viel Rechenaufwand bestimmt werden. Neben Parameterschätzungen, geschätzten y-Werten und Residuen, berechnet jeweils ohne die i-te Beobachtung, betrachtet man vor allem zwei Grössen: **Leverages** und die **Cook's Distanz**.

Die *Leverages* sind die Diagonalelemente h_{ii} der Hat-Matrix $\mathbf{H}$, die wir auf Seite 25 kennengelernt haben. Sie sind ein Mass dafür, wie extrem Beobachtungen bezüglich der erklärenden Variablen sind. Es gilt $0 \leq h_{ii} \leq 1$ für alle i. Beobachtungen mit Werten grösser als $2(p+1)/n$ werden als Hebelpunkte angesehen. Punkte mit grossem Residuum r_i und grossem h_{ii} sind gefährlich. Ein entsprechender Plot r_i gegen h_{ii} kann hilfreich sein.

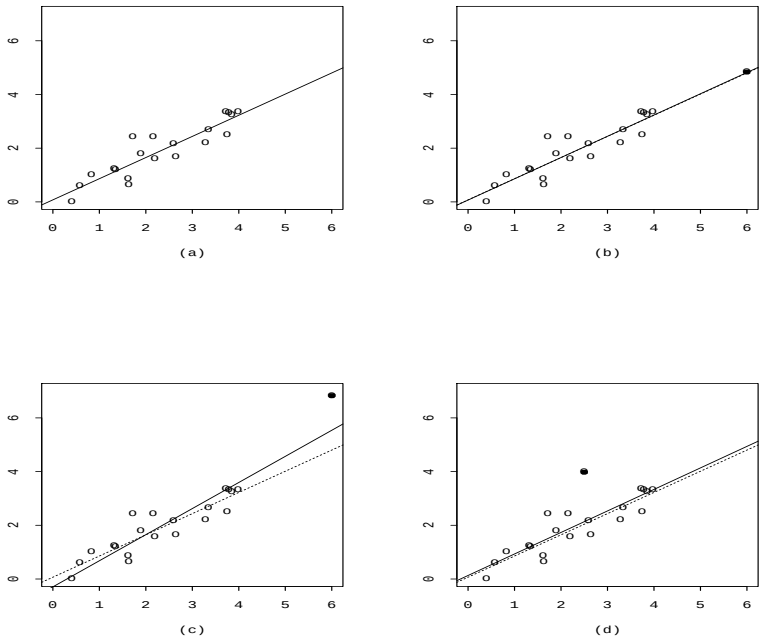


Abbildung 3.3: Ausreisser, einflussreiche Beobachtungen und Hebelpunkte

Die *Cook's Distanz* für Beobachtung i ist definiert als

$$D_i = \frac{\sum(\hat{y_j} - y_{\hat{j(i)}})^2}{(p+1)\hat{\sigma}^2} \tag{3.12}$$

wobei $y_{\hat{j(i)}}$ den geschätzten y-Wert ohne Beobachtung i bezeichnet.

Die D_i können ohne viel Rechenaufwand aus den h_{ii} und den r_i bestimmt werden, denn es gilt:

$$D_i = \frac{h_{ii}}{1 - h_{ii}} \cdot \frac{r_i^{*2}}{p+1}$$

r_i^* ist ein auf gleiche Varianzen standardisiertes Residuum, meist *studentisiertes Residuum* genannt:

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{3.13}$$

Punkte mit $D_i > 1$ sollten genauer untersucht werden.

Beispiel:
Wir wollen das Modell von Seite 30 überprüfen. Ein Normalplot der Residuen und die Graphik mit den studentisierten Residuen (Abb. 3.4 und 3.5) zeigen zwei klare Ausreisser "New Orleans" und "Albany". ("New Orleans" hat seine sehr hohe Mortalität). Die Graphiken 3.6 und 3.7 zeigen mögliche Hebelpunkte. "Los Angeles" und "York" haben grosse h_{ii}-Werte. Gefährlich ist "York", weil es Hebelpunkt und (leichter) Ausreisser zugleich ist. In der Graphik 3.8 entpuppt sich "York" auch als der einflussreichste Punkt, obschon alle Cook's Distanzen, auch diejenige von "York", ziemlich klein sind. Wenn man "York" genauer untersucht, dann entdeckt man zwei Auffälligkeiten: die extrem hohe Bevölkerungsdichte (Abb. 3.9) und ein Bildungsdefizit bei gleichzeitig hohem "white-collar worker"-Anteil (Abb. 3.10). Die extreme Bevölkerungsdichte folgt aus einer unglücklichen Distriktdefinition, wie eine Nachkontrolle ergeben hat. Das Bildungsdefizit wird mit dem hohen Anteil von Amischen begründet.

Es scheint sinnvoll, die Analyse auch ohne "York" und "New Orleans" zu machen.

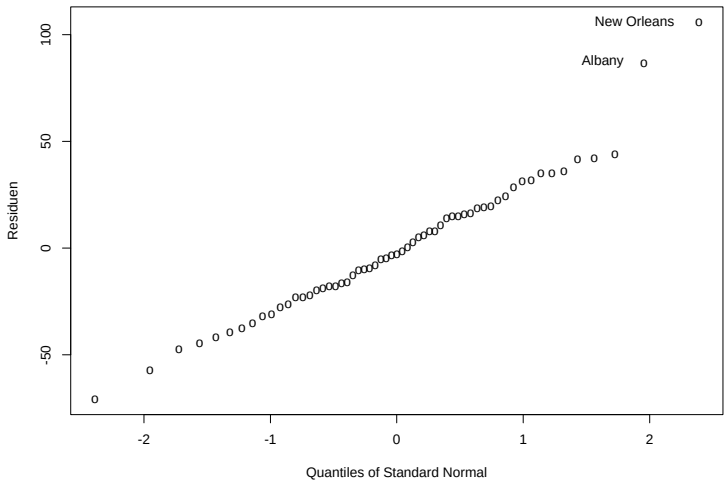


Abbildung 3.4: Normalplot

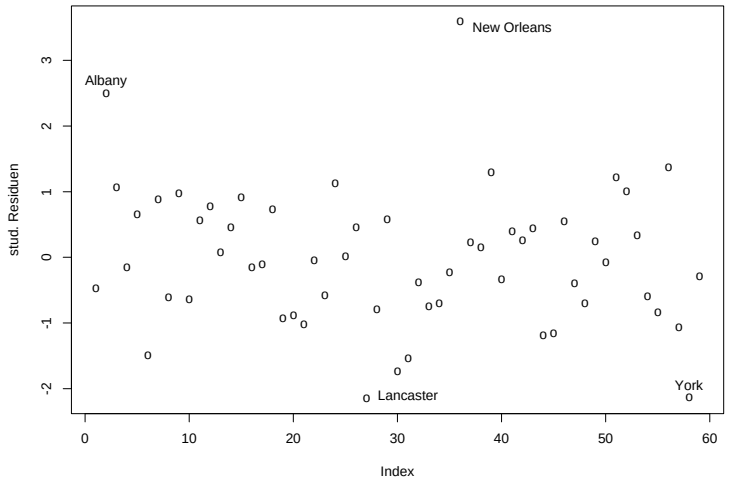


Abbildung 3.5: Studentisierte Residuen

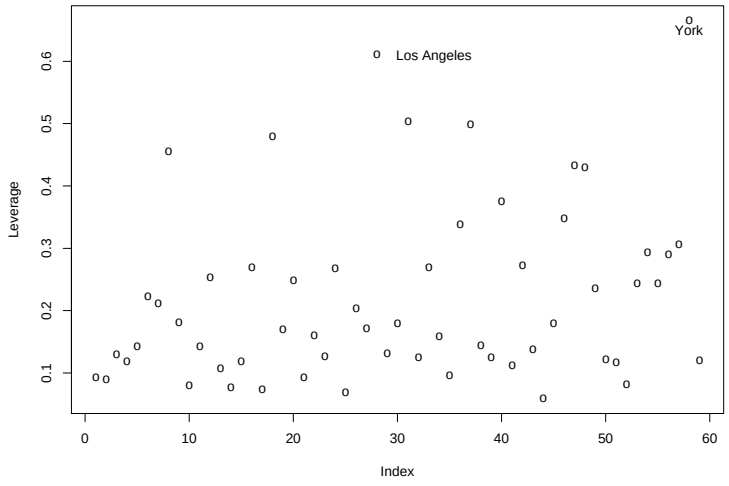


Abbildung 3.6: Hebelpunkte

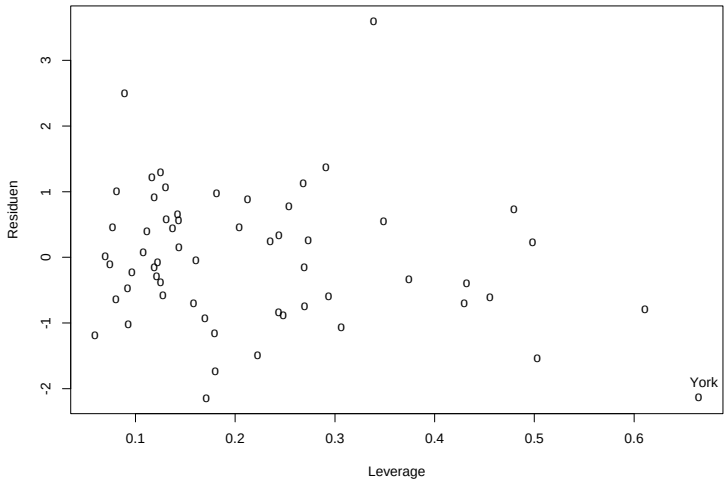


Abbildung 3.7: Residuen gegen Hebelpunkte

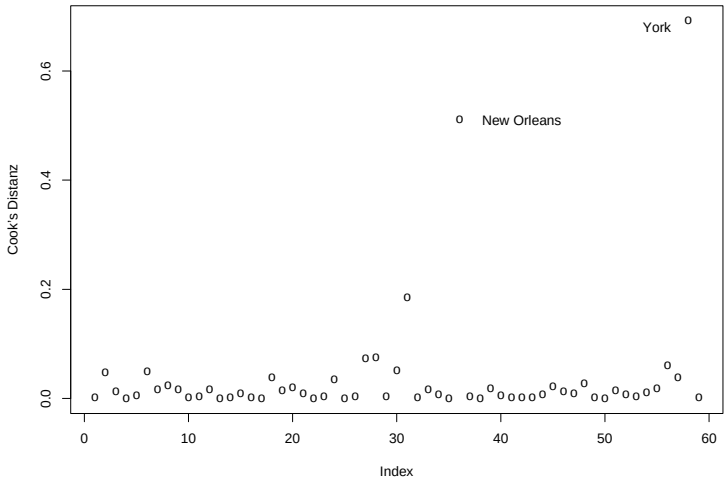


Abbildung 3.8: Cook's Distanzen

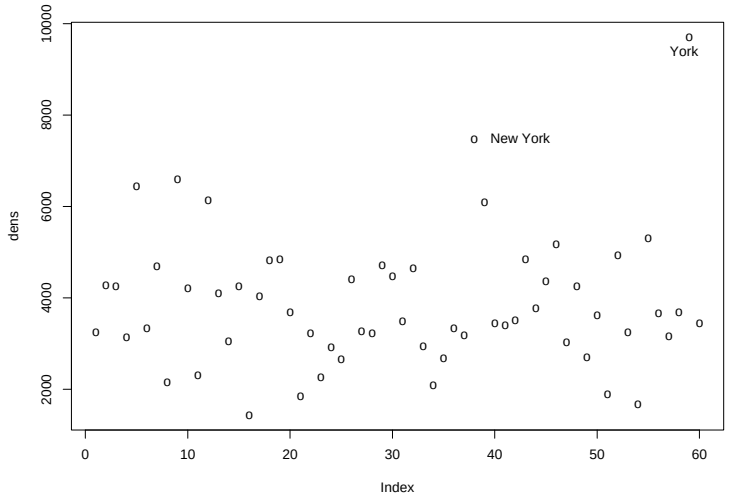


Abbildung 3.9: Bevölkerungsdichte

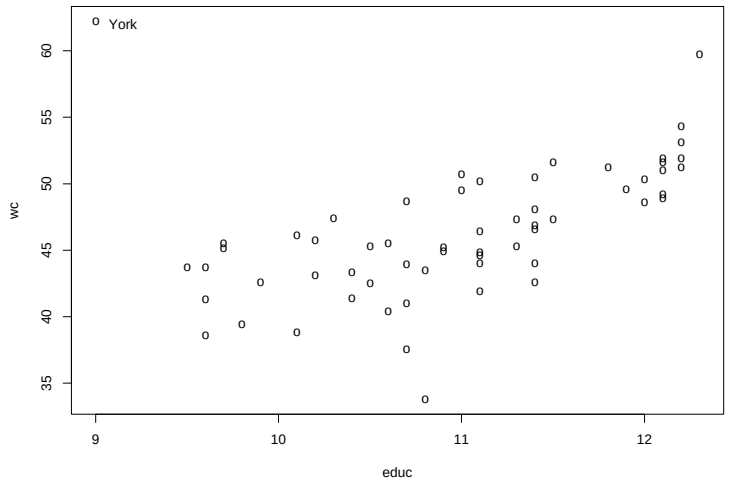


Abbildung 3.10: % Angestellte gegen Anzahl Schuljahre

Das Modell von Seite 30, ohne “York” und “New Orleans” gerechnet, ergibt:

```
Call:
lm(formula = mort ~ jantemp + julytemp + relhum + rain + dens +
    nonwhite + wc + pop + house + income + log(so2) + educ,
    data = smsa[-c(37,59), ])

Residuals:
     Min       1Q   Median       3Q      Max
-80.0176 -21.1399   0.2163  19.2546  74.4091

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.025e+02  2.564e+02   3.521 0.001016 **
jantemp     -1.246e+00  6.714e-01  -1.856 0.070168 .
julytemp    -1.317e-01  1.693e+00  -0.078 0.938339
relhum       3.984e-01  9.286e-01   0.429 0.670023
rain         1.399e+00  4.630e-01   3.022 0.004174 **
dens         9.360e-03  4.210e-03   2.223 0.031377 *
nonwhite     3.651e+00  9.021e-01   4.048 0.000206 ***
wc          -1.046e+00  1.371e+00  -0.763 0.449775
pop         -1.175e-06  3.478e-06  -0.338 0.737058
house        1.390e+00  3.430e+01   0.041 0.967857
income      -9.580e-05  1.118e-03  -0.086 0.932089
log(so2)     1.388e+01  5.151e+00   2.695 0.009926 **
educ        -5.788e+00  9.571e+00  -0.605 0.548430
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.31 on 44 degrees of freedom
Multiple R-Squared: 0.7929,     Adjusted R-squared: 0.7364
F-statistic: 14.04 on 12 and 44 DF,  p-value: 2.424e-11
```

4 Polynomiale Regression

- Was ist eine quadratische Regression?
- Wann passt man Polynome höherer Ordnung an?
- Sind die Korrelationen zwischen $x, x^2, \ldots$ ein Problem?

Wenn der Zusammenhang zwischen einer Zielvariablen y und einer erklärenden Variablen x nichtlinear ist, kann man versuchen, das Modell mit quadratischen oder Termen höherer Ordnung von x zu verbessern. Das führt zur *polynomialen Regression.*

4.1 Ein Modell mit einer erklärenden Variablen

Statt dem einfachen linearen Modell (2.1)

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

betrachten wir das polynomiale Modell

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i \qquad i = 1, \ldots, n \tag{4.1}$$

Die höchste vorkommende Potenz p von x heisst *Grad* des Polynoms. Polynome 2. Grades nennt man *quadratisch*, solche vom Grad 3 *kubisch*. Das Modell ist von *p-ter Ordnung.* Die Abbildung 4.1 zeigt verschiedene Polynome.

Wenn die verschiedenen Potenzen von x als eigenständige erklärende Variablen angesehen werden, können wir die polynomiale Regression als Spezialfall der multiplen Regression interpretieren mit $x_1 = x, x_2 = x^2, \ldots, x_p = x^p$. Für n beliebige Punkte gibt es immer ein Polynom vom Grad $n - 1$, das durch alle Punkte geht. Das Ziel ist, ein Polynom kleinstmöglichen Grades zu wählen. In der Praxis geht man kaum je über die dritte Potenz hinaus.

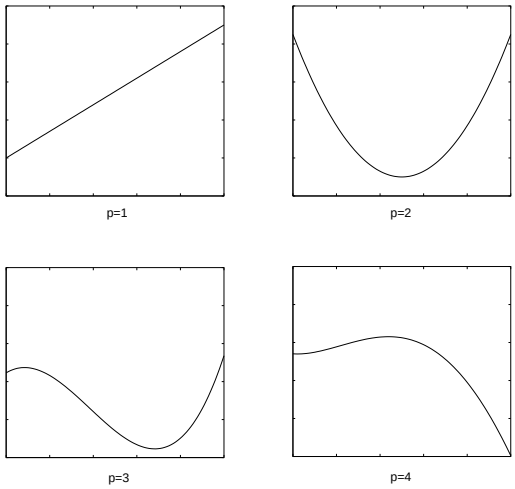


Abbildung 4.1: Polynome 1. bis 4. Grades

Beispiel: Cadmiumarbeiter

In Kapitel 2 haben wir den Zusammenhang zwischen Lungenfunktion und Alter bei Arbeitern in der Cadmiumindustrie untersucht. In einer zweiten Messperiode hat man neun weitere, vorwiegend junge Personen untersucht. Die Abbildung 4.2 zeigt alle 49 Beobachtungen.

Die Vitalkapazität scheint jetzt nicht mehr linear mit dem Alter abzunehmen. Ein Modell mit einem quadratischen Term könnte die Daten besser beschreiben.

```
Call:
lm(formula = vit ~ age + I(age^2), data = lung)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.545     .842      4.210    0.000 ***
age            0.089     .044      2.031    0.048 *
I(age^2)       -.002     .001     -3.003    0.004 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5415 on 46 degrees of freedom
Multiple R-Squared: 0.5096,     Adjusted R-squared: 0.4883
F-statistic:  23.9 on 2 and 46 DF,  p-value: 7.634e-08
```

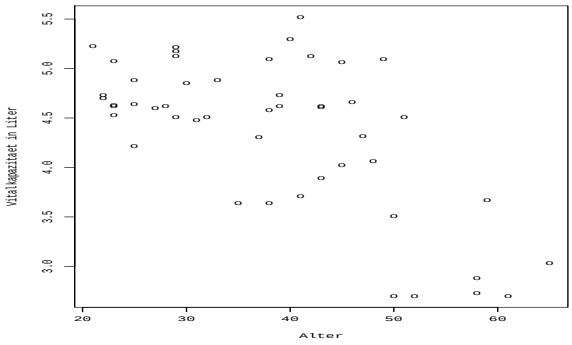


Abbildung 4.2: Lungenfunktionswerte von Cadmium-Arbeitern

Eine kleine Schwierigkeit ergibt sich allerdings, wenn Potenzen von x in ein Regressionsmodell eingebaut werden. Die Variablen x und x^2, sowie höhere Potenzen sind in der Regel hochkorreliert. Das erschwert, wie wir schon gesehen haben, die Interpretation der Testresultate und der Schätzungen. Wenn die Korrelationen extrem hoch ausfallen, gibt es numerische Schwierigkeiten bei der Invertierung von $\mathbf{X}^t\mathbf{X}$. Besser ist es, ein Modell mit transformierten erklärenden Variablen $z_i = x_i - \bar{x}, z_i^2 = (x_i - \bar{x})^2, \ldots$ zu rechnen. Die Korrelation zwischen z und z^2 ist oft viel kleiner als diejenige zwischen x und x^2.

In unserem Beispiel ist die Korrelation zwischen Alter und Alter^2 : 0.989. Wenn man stattdessen die Abweichungen vom Mittelwert betrachtet, fällt die Korrelation auf 0.32.

Der R–Output mit den transformierten Variablen sieht so aus:

```
Call:
lm(formula = vit ~ age + I(age^2), data = lung)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.591      .107     42.99    0.000 ***
age-38.16          -.035      .007     -4.95    0.000 ***
I((age-38.16)^2)   -.002      .001     -3.00    0.004 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5415 on 46 degrees of freedom
Multiple R-Squared: 0.5096,     Adjusted R-squared: 0.4883
F-statistic:  23.9 on 2 and 46 DF,  p-value: 7.634e-08
```

Bemerkung

Der Vollständigkeit halber erwähnen wir noch das einfachste kompliziertere Modell, das Modell 2. Ordnung mit zwei erklärenden Variablen:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \epsilon_i \qquad (4.2)$$
$$i = 1, \ldots, n$$

Neben den linearen und quadratischen Komponenten von x_1 und x_2 enthält dieses Modell einen Term für die *Interaktion* zwischen x_1 und x_2.

5 Indikatorvariablen

- Wie werden qualitative erklärende Variablen berücksichtigt?

5.1 Variablen mit zwei Kategorien

Bis jetzt haben wir ausschliesslich stetige Variablen betrachtet. Häufig interessiert aber auch der Effekt von qualitativen Variablen auf eine stetige Zielgrösse. Beispiele sind Geschlecht, Region, Sozialschicht oder Schweregrad der Erkrankung. Da im multiplen Regressionsmodell keine Voraussetzungen über die x-Variablen gemacht werden, sind kategorielle Variablen durchaus erlaubt. Sie werden im Modell durch Indikatoren für die verschiedenen Kategorien dargestellt.

Neben 40 exponierten Arbeitern sind weitere 44 Industriearbeiter untersucht worden, die keinen Cadmiumdämpfen ausgesetzt worden sind. Wir interessieren uns für den Zusammenhang zwischen Vitalkapazität (y) und Alter (x_1) sowie Exposition. Die zweite erklärende Variable ist binär, sie nimmt nur die Werte „nicht exponiert“ und „exponiert“ an. Wir definieren eine *Indikatorvariable*

$$x_2 = \begin{cases} 0 & : \quad \text{nicht exponiert} \\ 1 & : \quad \text{exponiert} \end{cases}$$

Indikatorvariablen werden auch *Dummy Variablen* genannt.

Das einfachste Modell sieht so aus:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \qquad i = 1, \ldots, n \tag{5.1}$$

Das gibt für nicht exponierte Arbeiter ($x_2 = 0$): $Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$
und für exponierte Arbeiter ($x_2 = 1$): $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 + \epsilon_i,$

also zwei parallele Regressionsgeraden mit Steigung β_1. Der Unterschied im Achsenabschnitt ist β_2.

Mit einem solchen Modell nehmen wir an, dass die Exposition die Lungenfunktion um eine konstante Grösse reduziert, unabhängig vom Alter. Umgekehrt hat das Alter in beiden Arbeitergruppen denselben Effekt. Ob eine solche Annahme vernünftig ist, wird die Analyse zeigen.

Mit Matrizen kann das Modell (5.1) folgendermassen geschrieben werden:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

mit

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & 0 \\ 1 & x_{21} & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{n_1 1} & 0 \\ 1 & x_{n_1+1,1} & 1 \\ 1 & x_{n_1+2,1} & 1 \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & 1 \end{pmatrix} \qquad \text{und } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Beispiel:
Rechnen wir das Modell (5.1) mit den Daten der 84 Cadmiumarbeiter. Wir erhalten die folgende Regressionsgleichung:

$$\hat{y} = 6.0208 - 0.0402 \cdot x_1 - 0.0835 \cdot x_2$$

Die Regressionsgerade für die nicht exponierten Arbeiter ist also: $\hat{y} = 6.0208 - 0.0402 x_1$, für die exponierten Arbeiter: $\hat{y} = 5.9373 - 0.0402 \cdot x_1$ Der Koeffizient der Variablen x_2 ist aber nicht signifikant (P-Wert 0.57).

Die Abbildung 5.1 zeigt die 84 Beobachtungen mit den parallelen Regressionsgeraden.

Modelle mit Interaktionen

Im nächst komplizierteren Modell nimmt man an, dass neben dem Achsenabschnitt auch die Steigungen der Regressionsgeraden für die beiden Gruppen verschieden sind. Das Modell sieht dann so aus:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \qquad i = 1, \ldots, n \tag{5.2}$$

Der zusätzliche Term $\beta_3 x_1 x_2$ modelliert eine *Interaktion* oder *Wechselwirkung* zwischen x_1 und x_2. Der Effekt von x_1 ist jetzt verschieden in den beiden durch x_2 definierten Gruppen.

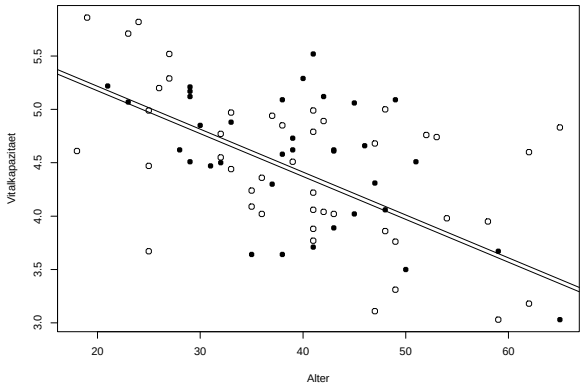


Abbildung 5.1: o Nichtexponierte, • Exponierte

Aufgabe 5.1
Wie sehen die Regressionsgleichungen für die beiden Gruppen in diesem Modell aus?

Mit Matrizen kann das Modell (5.2) folgendermassen geschrieben werden:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

mit

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ 1 & x_{21} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1 1} & 0 & 0 \\ 1 & x_{n_1+1,1} & 1 & x_{n_1+1,1} \\ 1 & x_{n_1+2,1} & 1 & x_{n_1+2,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & 1 & x_{n1} \end{pmatrix} \qquad \text{und } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Die vierte Spalte von $\mathbf{X}$ ist das Produkt der zweiten und dritten Spalte: $\mathbf{x}_4 = \mathbf{x}_2^t \mathbf{x}_3$.

Beispiel
Der R–Output für das Regressionsmodell (5.2) sieht folgendermassen aus:

```
> summary(lm(vit~(age+exp)^2,data=lung1))
```

```
Call:
lm(formula = vit ~ (age + exp)^2, data = lung1)

Residuals:
      Min        1Q    Median        3Q       Max
-1.244974 -0.415557  0.004294  0.403347  1.184889

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.680291   0.315814  17.986  < 2e-16 ***
age         -0.030613   0.007605  -4.025 0.000128 ***
exp1         0.858861   0.498226   1.724 0.088600 .
age:exp1    -0.023143   0.011804  -1.961 0.053409 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5987 on 80 degrees of freedom
Multiple R-Squared: 0.3981,     Adjusted R-squared: 0.3756
F-statistic: 17.64 on 3 and 80 DF,  p-value: 7.002e-09
```

Die Interaktion ist knapp nicht signifikant. Die beiden Regressionsgleichungen sind:
Nichtexponierte: $5.68 - 0.0307 \cdot$ age und
Exponierte: $6.539 - 0.0538 \cdot$ age
Die Abbildung 5.2 zeigt die beiden Geraden.

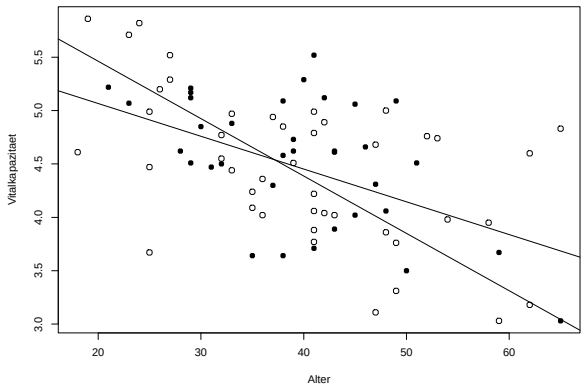


Abbildung 5.2: o Nichtexponierte, • Exponierte

5.2 Variablen mit mehr als zwei Kategorien

Bei den Cadmiumarbeitern ist unterschieden worden zwischen mehr als 10 Jahre Exponierte und weniger als 10 Jahre Exponierte. Man kann also eine erklärende Variable mit drei Kategorien bilden: keine Exposition, weniger als 10 Jahre, mehr als 10 Jahre Exposition. Um diese Variable in ein Regressionsmodell aufzunehmen, braucht es zwei Indikatorvariablen x_2 und x_3.

$$x_2 = \begin{cases} 1 & : & < \text{ 10 Jahre exponiert} \\ 0 & : & \text{sonst} \end{cases}$$

$$x_3 = \begin{cases} 1 & : & > \text{ 10 Jahre exponiert} \\ 0 & : & \text{sonst} \end{cases}$$

Das ergibt
$x_2 = 0, x_3 = 0$ für die Gruppe „keine Exposition“
$x_2 = 1, x_3 = 0$ für die Gruppe „< 10 Jahre exponiert“
$x_2 = 0, x_3 = 1$ für die Gruppe „> 10 Jahre exponiert“

Für eine Variable mit k Kategorien braucht es $k - 1$ Indikatorvariablen.

Ein Modell, das neben dieser kategoriellen Variablen auch noch Interaktionen berücksichtigt, sieht so aus:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \epsilon_i \qquad i = 1, \ldots, n \qquad (5.3)$$

Beispiel:
Die entsprechende Regressionsanalyse mit R liefert:

```
> summary(lm(vit~(age+exp)^2,data=lung1))

Call:
lm(formula = vit ~ (age + exp)^2, data = lung1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24497 -0.36929  0.01977  0.43681  1.13953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  5.680291   0.313426  18.123  < 2e-16 ***
age         -0.030613   0.007547  -4.056 0.000117 ***
exp1         0.549740   0.575884   0.955 0.342728
exp2         2.503148   1.041842   2.403 0.018655 *
age:exp1    -0.015919   0.014547  -1.094 0.277170
age:exp2    -0.054498   0.021070  -2.587 0.011554 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5942 on 78 degrees of freedom
Multiple R-Squared: 0.422,      Adjusted R-squared: 0.385
F-statistic: 11.39 on 5 and 78 DF,  p-value: 2.871e-08
```

Der Effekt einer kategoriellen Variablen mit mehr als zwei Klassen sollte nicht auf Grund der P-Werte für die einzelnen Koeffizienten beurteilt werden. Es wäre falsch zu schliessen, dass Exposition signifikant ist (P-Wert=0.019), man aber auf die Zwischenkategorie „exp=1“ verzichten kann, weil P-Wert=0.34.

Eine richtige Beurteilung ist mit Hilfe einer modifizierten Anova-Tabelle möglich (siehe Seite 50). Dabei wird die Regression sum of squares weiter unterteilt in drei Summen entsprechend den drei erklärenden Variablen. Ein Vergleich der Mean squares mit MSE ergibt drei partielle F-Tests. Der F-Test in der dritten Zeile, der die Wechselwirkung untersucht, testet die Nullhypothese $H_0 : \beta_4 = \beta_5 = 0$. Es wird bestätigt, dass die Exposition je nach Alter unterschiedlich wirkt.

```
> anova(lm(vit~(age+exp)^2,data=lung1))
Analysis of Variance Table

Response: vit
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 17.4446 17.4446 49.4159 6.918e-10 ***
exp        2  0.1617  0.0808  0.2290   0.79584
age:exp    2  2.4995  1.2497  3.5402   0.03376 *
Residuals 78 27.5352  0.3530
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Bemerkungen

a) Auch quantitative Variablen können mit Hilfe von Indikatoren modelliert werden. Man kann zum Beispiel Altersgruppen bilden und diese kategorielle Variable statt

des genauen Alters ins Modell aufnehmen. Der funktionale Zusammenhang mit der Zielvariable muss dann nicht mehr festgelegt werden. Dafür müssen mehr Parameter geschätzt werden.

b) Wenn die Mehrheit der erklärenden Variablen kategoriell ist, betrachtet man das Modell besser im Rahmen der Varianzanalyse.

6 Modellwahl

- Wie wählt man aus einer Reihe konkurrenzierender Modelle das „beste“ aus?
- Welche erklärenden Variablen sind im Modell unbedingt nötig, welche sind überflüssig?

Es gibt verschiedene Strategien, das „beste“ Modell zu finden und verschiedene Kriterien, was das „beste“ Modell ist. Meistens gibt es allerdings, gleich nach welchem Kriterium, nicht ein „bestes“ Modell, sondern mehrere gleich „gute“.

6.1 Strategien

Rückwärts-Elimination

Bei der Rückwärts-Elimination (backward elimination) beginnt man mit dem vollständigen Modell, d. h. mit allen zur Verfügung stehenden, erklärenden Variablen. Man eliminiert diejenige Variable mit dem kleinsten F-Wert, sofern dieser kleiner als eine vorgegebene Schranke von z. B. $F_{OUT} = 3$ ist. Dann berechnet man eine neue Regression und eliminiert die nächst unwichtigste Variable im Modell, bis keine Variable mehr einen F-Wert unterhalb der Schranke besitzt. Diese Strategie ist nur durchführbar, wenn die Anzahl vorhandener erklärender Variablen deutlich kleiner ist als die Anzahl Beobachtungen.

Vorwärts-Selektion

Bei der Vorwärts-Selektion (forward selection) beginnt man mit dem „leeren“ Modell (keine erklärende Variablen) und nimmt schrittweise jeweils die wichtigste, zusätzliche Variable in das Modell auf, solange diese eine vorgegebene Schranke von z. B. $F_{IN} = 3$ überschreitet. Die Vorwärtsselektion braucht viel weniger Rechenaufwand als die Rückwärtsmethode.

Schrittweise Regression

Die schrittweise Regression (stepwise regression) ist eine Kombination von Vorwärts- und Rückwärtsstrategie. Man beginnt vorwärts, überprüft nach jeder Aufnahme einer neuen Variable aber die F-Werte der anderen Variablen. Es ist also möglich, dass einmal aufgenommene Variablen wieder eliminiert werden, oder dass eliminierte Variablen später wieder aufgenommen werden.

„Alle Gleichungen“

Bei diesem Verfahren wird unter allen Regressionsmodellen mit den vorhandenen erklärenden Variablen das „beste“ gesucht (all subsets). Die Zahl der Modelle wächst mit der Anzahl Variablen allerdings ziemlich schnell an. Bsp: mit 10 Variablen gibt es $2^{10} = 1024$ Modelle. Intelligente Algorithmen ersparen es sich, alle Modelle durchzurechnen.

6.2 Gütekriterien

In Frage kommen die folgenden Grössen:

1. Maximales Korrigiertes Bestimmtheitsmass :

$$adjR^2 = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right) = 1 - \frac{MSE}{MST}$$

2. Minimaler Mean Square Error:

$$MSE = \frac{SSE}{n-p-1}$$

3. Maximaler Wert der Teststatistik des globalen F-Tests:

$$F^* = \frac{MSR}{MSE}$$

4. Minimale **PRESS-Statistik** : Dieses Kriterium misst die Güte des Modells an seinem Vorhersagewert. Man rechnet das Regressionsmodell jeweils ohne die i-te Beobachtung und vergleicht dann den geschätzten Wert für die i-te Beobachtung mit dem tatsächlichen y-Wert.

   Das i-te PRESS-Residuum ist definiert als

$$r_{i,-i} = y_i - \hat{y}_{i,-i}$$

mit dem geschätzten y-Wert für $\mathbf{x}_i^t$:

$$\hat{y}_{i,-i} = \mathbf{x}_i^t \boldsymbol{\beta}_{-i}.$$

$\boldsymbol{\beta}_{-i}$ bezeichnet die LS-Lösung ohne die i-te Beobachtung.

Führt man diese Berechnungen für jede Beobachtung aus (also n mal) und summiert die quadrierten PRESS-Residuen auf, so erhält man die PRESS-Statistik

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} r_{i,-i}^2$$

Glücklicherweise ist es nicht nötig, alle n Regressionen durchzuführen, denn es gilt:

$$r_{i,-i} = \frac{r_i}{1 - h_{ii}}$$

Somit erhält man

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{r_i}{1 - h_{ii}} \right)^2$$

Die PRESS-Statistik lässt sich also aus den gewöhnlichen Residuen $r_i$und den Leverages h_{ii} (Diagonalelementen der Hat-Matrix) berechnen.

5. Mallows C_q:

   Die Statistik von Mallows ist gegeben durch

   $$C_q = \frac{SSE_q}{\hat{\sigma}_f^2} - n + 2q$$

   Dabei ist SSE_q das Fehlersummenquadrat des betrachteten Modells, q ist die Anzahl der Parameter des entsprechenden Modells ($q = p+1$), und $\hat{\sigma}_f^2$ ist die geschätzte Varianz unter dem vollen Modell mit allen erklärenden Variablen. Modelle mit C_q nahe bei q sind gute Kandidaten für die Modellwahl.

Für eine feste Anzahl von Variablen im Modell führen all diese Kriterien zum gleichen Ergebnis. Wenn hingegen Modelle miteinander verglichen werden mit einer unterschiedlichen Anzahl Variablen, dann können verschiedene „beste“ Modelle herauskommen. Es gibt eben auch nicht ein „bestes“ oder gar „richtiges“ Modell und alles andere ist schlechter oder falsch. Neben der automatisierten Suche aufgrund eines objektiven Kriteriums braucht es immer auch viel Fachwissen und subjektive Entscheidungen, um zu einem oder mehreren „geeigneten“ Modellen zu gelangen.

Beispiel: Luftverschmutzung und Mortalität

Wir haben insgesamt 14 erklärende Variablen zur Verfügung. Alle Modelle auszurechnen, ist zu aufwendig. Wenden wir deshalb die Vorwärts-, Rückwärts- und Stepwisestrategie auf unser Beispiel an, wobei wir "York" und "New Orleans" wie vorher weglassen.

Vorwärts-Selektion:

```
Call: lm(formula = mort ~ nonwhite + log(so2) +
log(hc) + rain + dens + wc + jantemp + log(nox),
data = smsa[-c(21, 37, 59), ])

Residuals:
     Min       1Q   Median       3Q      Max
-74.8222 -20.1081  -0.7881  21.1415  64.9520

Coefficients:
            Estimate Std.Err tvalue Pr(>|t|)
(Intercept) 890.2641 48.9959 18.170 < 2e-16 ***
jantemp      -1.1327  0.5865 -1.931 0.05937 .
rain          1.3446  0.4861  2.766 0.00803 **
dens          0.0082  0.0036  2.291 0.02638 *
nonwhite      3.5219  0.5496  6.409 5.9e-08 ***
wc           -1.5430  0.8928 -1.728 0.09036 .
log(hc)     -18.3978 10.9789 -1.676 0.10029
log(nox)     16.2242 11.4449  1.418 0.16277
log(so2)     14.5321  6.1516  2.362 0.02226 *
---
Residual standard error: 28.49 on 48 df
Mult. R-Squared: 0.8003, Adj. R-squared: 0.7671
```

Rückwärts-Elimination:

```
Call: lm(formula = mort ~ jantemp + rain + dens
+ nonwhite + wc + log(hc) + log(nox) + log(so2),
      data = smsa[-c(21, 37, 59), ])

Residuals:
     Min       1Q   Median       3Q      Max
-74.8222 -20.1081  -0.7881  21.1415  64.9520
```

```
Coefficients:
            Estimate Std.Err tvalue Pr(>|t|)
(Intercept) 890.2641 48.9959 18.170 < 2e-16 ***
jantemp      -1.1327  0.5865 -1.931 0.05937 .
rain          1.3446  0.4861  2.766 0.00803 **
dens          0.0082  0.0036  2.291 0.02638 *
nonwhite      3.5219  0.5496  6.409 5.9e-08 ***
wc           -1.5430  0.8928 -1.728 0.09036 .
log(hc)     -18.3978 10.9789 -1.676 0.10029
log(nox)     16.2242 11.4449  1.418 0.16277
log(so2)     14.5321  6.1516  2.362 0.02226 *
---
Residual standard error: 28.49 on 48 df
Mult. R-Squared: 0.8003, Adj. R-squared: 0.7671
F-stat.: 24.05 on 8 and 48 df,p-value: 2.5e-014
```

Vorwärts- und Rückwärts-Strategie ergeben in diesem Fall dasselbe. Allgemein ist bei der Vorwärtsstrategie die Gefahr aber am grössten, bei einem suboptimalem Modell zu landen.

A Matrizen und Vektoren

A.1 Definition

Eine *Matrix* ist eine rechteckige Anordnung von Zahlen in Zeilen und Spalten

$$\mathbf{A} = \begin{pmatrix} 55.7 & 4.1 \\ 58.2 & 1.6 \\ 56.9 & 2.7 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

$\mathbf{A}$ hat die Dimension 3×2 (Anzahl Zeilen $\times$ Anzahl Spalten). a_{ij} ist das Element in der i-ten Zeile und j-ten Spalte von $\mathbf{A}$.

Die *transponierte Matrix* von $\mathbf{A}$ ist

$$\mathbf{A^t} = \begin{pmatrix} 55.7 & 58.2 & 56.9 \\ 4.1 & 1.6 & 2.7 \end{pmatrix}$$

d. h. die Zeilen von $\mathbf{A}$ werden zu Spalten von $\mathbf{A^t}$ und umgekehrt. $\mathbf{A^t}$ hat die Dimension 2x3.

Beispiele:

1. Eine Matrix der Dimension 1×1 ist einfach eine Zahl, auch *Skalar* genannt.
2. Eine quadratische Matrix hat gleichviele Spalten wie Zeilen, also Dimension $n \times n$.
3. Eine Matrix, die nur aus einer Spalte besteht, ist ein *Vektor*:

$$\mathbf{z} = \begin{pmatrix} 7.13 \\ 8.82 \\ 8.34 \end{pmatrix}, \qquad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Die Transponierte von $\mathbf{z}$ ist ein Zeilenvektor

$$\mathbf{z^t} = \begin{pmatrix} 7.13 & 8.82 & 8.34 \end{pmatrix}.$$

4. Wenn $\mathbf{A} = \mathbf{A^t}$, so heisst $\mathbf{A}$ *symmetrisch*.

$$\mathbf{A} = \begin{pmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{pmatrix}$$

   Eine symmetrische Matrix muss natürlich quadratisch sein.

5. Eine *Diagonalmatrix* ist eine symmetrische Matrix, bei der alle Elemente ausserhalb der Diagonalen 0 sind.

$$\mathbf{D} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

6. Die *Einheitsmatrix* $\mathbf{I}$ ist eine Diagonalmatrix mit lauter Einsen in der Diagonale.

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

A.2 Wie lässt sich mit Matrizen rechnen?

Addition und Subtraktion

Die Matrizen müssen gleiche Dimension haben. Die Operation wird elementweise ausgeführt.

$$\mathbf{A} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 2 & 1 \end{pmatrix} \quad \mathbf{A} + \mathbf{B} = \begin{pmatrix} 2 & 6 \\ 4 & 8 \\ 5 & 7 \end{pmatrix} \quad \mathbf{A} - \mathbf{B} = \begin{pmatrix} 0 & 2 \\ 0 & 2 \\ 1 & 5 \end{pmatrix}$$

Multiplikation einer Matrix mit einem Skalar

wird ebenfalls elementweise ausgeführt:

$$3 \cdot \mathbf{A} = \begin{pmatrix} 3 & 12 \\ 6 & 15 \\ 9 & 18 \end{pmatrix} \qquad \mathbf{B} \cdot 2 = \begin{pmatrix} 2 & 4 \\ 4 & 6 \\ 4 & 2 \end{pmatrix}$$

Multiplikation zweier Matrizen

Betrachten wir ein Beispiel:

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 1 & 4 & 6 \\ 1 & 5 & 8 \end{pmatrix}$$

Das Produkt $\mathbf{A} \cdot \mathbf{B}$ hat die Dimension 2×3 (Anzahl Zeilen von $\mathbf{A}\times$ Anzahl Spalten von $\mathbf{B}$). Das Element von $\mathbf{A} \cdot \mathbf{B}$ an der Stelle $(1, 1)$ wird so berechnet:

$$\begin{pmatrix} \boxed{\begin{matrix} 2 & 3 \end{matrix}} \\ \begin{matrix} 4 & 1 \end{matrix} \end{pmatrix} \cdot \begin{pmatrix} \boxed{\begin{matrix} 1 \end{matrix}} & 4 & 6 \\ \begin{matrix} 1 \end{matrix} & 5 & 8 \end{pmatrix} = \begin{pmatrix} 5 & . & . \\ . & . & . \end{pmatrix}$$
$$2 \cdot 1 + 3 \cdot 1 = 5$$

Nun berechnen wir das Element an der Stelle $(1, 2)$:

$$\begin{pmatrix} \boxed{\begin{matrix} 2 & 3 \end{matrix}} \\ \begin{matrix} 4 & 1 \end{matrix} \end{pmatrix} \cdot \begin{pmatrix} 1 & \boxed{\begin{matrix} 4 \end{matrix}} & 6 \\ 1 & \begin{matrix} 5 \end{matrix} & 8 \end{pmatrix} = \begin{pmatrix} 5 & 23 & . \\ . & . & . \end{pmatrix}$$
$$2 \cdot 4 + 3 \cdot 5 = 23$$

Allgemein berechnet man das Element an der Stelle (i, j) aus der i-ten Zeile von $\mathbf{A}$und der j-ten Spalte von $\mathbf{B}$. Das Resultat ist:

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} 5 & 23 & 36 \\ 5 & 21 & 32 \end{pmatrix}$$

Das Produkt von $\mathbf{A}$ und $\mathbf{B}$ ist nur definiert, wenn die Anzahl Spalten von $\mathbf{A}$ gleich der Anzahl Zeilen von $\mathbf{B}$ ist. Die Produkte $\mathbf{A} \cdot \mathbf{B}$ und $\mathbf{B} \cdot \mathbf{A}$ können verschieden sein oder möglicherweise ist nur eines der beiden Produkte definiert.

Noch ein paar Beispiele:

a) $$\begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 2\beta_0 + 3\beta_1 \\ 4\beta_0 + \beta_1 \end{pmatrix}$$

b) $$\begin{pmatrix} 2 & 1 & 4 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 2^2 + 1^2 + 4^2 \end{pmatrix} = 21$$

c) $$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}$$

Für alle Matrizen $\mathbf{A}$ gilt: $\mathbf{I} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{I} = \mathbf{A}$

d) Um das einfache lineare Regressionsmodell mit Matrizen zu schreiben, setzen wir

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{und} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

sowie

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{und} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Dabei ist $\mathbf{y}$ der Zielvariablenvektor, $\mathbf{X}$ die *Designmatrix* der Dimension $n \times 2$, $\boldsymbol{\beta}$ der Koeffizientenvektor und $\boldsymbol{\epsilon}$ der Fehlervektor.

Das Modell sieht dann so aus:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Frage:

Wie sehen $\mathbf{y}$ und $\mathbf{X}$ aus im Cadmium-Beispiel von Kapitel 2, wenn wir uns auf die länger als 10 Jahre exponierten Arbeiter beschränken?

e) Eine wichtige Rolle spielt die Matrix $\mathbf{X}^t\mathbf{X}$ in der Regression. Im einfachen linearen Modell ist das eine 2×2-Matrix:

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

A.3 Lineare Unabhängigkeit und inverse Matrizen

Lineare Unabhängigkeit

Betrachten wir die folgende Matrix etwas genauer:

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 4 & 2 \\ 0 & 3 & 1 & 1 \\ 2 & 4 & 7 & 3 \end{pmatrix}$$

B besteht aus vier Spaltenvektoren, zwischen denen ein spezieller Zusammenhang existiert. Die dritte Spalte ist nämlich eine *Linearkombination* der übrigen Spalten:

$$\begin{pmatrix} 4 \\ 1 \\ 7 \end{pmatrix} = 2 \cdot \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$$

Die Spalten von B heissen *linear abhängig*. Umgekehrt heisst eine Menge von Vektoren *linear unabhängig*, wenn keiner der Vektoren als Linearkombination der übrigen geschrieben werden kann.

Das Inverse einer Matrix

Für bestimmte Matrizen ist auch eine Art Division definiert. Das *Inverse* einer quadratischen Matrix $\mathbf{A}$ wird mit $\mathbf{A^{-1}}$ bezeichnet und ist folgendermassen definiert:

$$\mathbf{A^{-1} \cdot A = A \cdot A^{-1} = I}$$

Nicht jede quadratische Matrix ist invertierbar. Ein Inverses existiert genau dann, wenn alle Spalten, resp. Zeilen linear unabhängig sind.

Ein Beispiel:

$$\mathbf{A} = \begin{pmatrix} 2 & 4 \\ 3 & 1 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{A^{-1}} = \begin{pmatrix} -0.1 & 0.4 \\ 0.3 & -0.2 \end{pmatrix}$$

Überprüfen Sie selbst, ob $\mathbf{A^{-1} \cdot A = A \cdot A^{-1}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Die Berechnung des Inversen ist, ausser in ein paar Spezialfällen, schwierig sobald $n \geq 4$. Ein einfacher Spezialfall sind Diagonalmatrizen. Das Inverse ist wieder eine Diagonalmatrix mit reziproken Werten in der Diaonale.

$$\mathbf{D} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{D^{-1}} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/3 \end{pmatrix}$$

Inverse Matrizen werden für das Lösen von linearen Gleichungssystemen benutzt. Ein Gleichungssystem in Matrixschreibweise ist:

$$\mathbf{Ay = c}$$

Multiplizieren wir beide Seiten mit $\mathbf{A}^{-1}$ so erhalten wir

$$\mathbf{A}^{-1}\mathbf{Ay} = \mathbf{A}^{-1}\mathbf{c} \quad \text{und somit} \quad \mathbf{y} = \mathbf{A}^{-1}\mathbf{c}.$$

A.4 Zufallsvektoren und Kovarianzmatrizen

Aus den Zufallsvariablen $Y_1, Y_2, \ldots$ können wir einen *Zufallsvektor* bilden:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \end{pmatrix}$$

Der Erwartungswert von $\mathbf{Y}$ ist dann ebenfalls ein Vektor:

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \end{pmatrix}$$

Die Varianzen der einzelnen Variabeln und die Kovarianzen zwischen je zwei Variablen werden in der sogenannten *Kovarianzmatrix* zusammengefasst.

$$Cov(\mathbf{Y}) = \begin{pmatrix} VarY_1 & Cov(Y_1, Y_2) & Cov(Y_1, Y_3) & \cdots \\ Cov(Y_1, Y_2) & VarY_2 & Cov(Y_2, Y_3) & \cdots \\ Cov(Y_1, Y_3) & Cov(Y_2, Y_3) & VarY_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Zur Erinnerung: $Cov(Y_1, Y_2) = \rho \cdot \sqrt{VarY_1}\sqrt{VarY_2}$, wobei ρ die Korrelation zwischen Y_1 und Y_2 ist.

Kovarianzmatrizen sind symmetrisch.

Rechenregeln:

$$\begin{aligned} E(\mathbf{a} + \mathbf{B} \cdot \mathbf{Y}) &= \mathbf{a} + \mathbf{B} \cdot E(\mathbf{Y}) \\ Cov(\mathbf{a} + \mathbf{B} \cdot \mathbf{Y}) &= \mathbf{B} \cdot Cov(\mathbf{Y}) \cdot \mathbf{B}^t \end{aligned}$$

wobei $\mathbf{a}$ ein konstanter Vektor und $\mathbf{B}$ eine konstante Matrix ist.

A.5 Mehrdimensionale Verteilungen

Die Wahrscheinlichkeitsverteilung eines Zufallsvektors ist die gemeinsame Verteilung der einzelnen Variablen. Am häufigsten benutzt wird die multivariate Normalverteilung. Was man sich unter einer zweidimensionalen, sogenannte bivariaten Normalverteilung vorstellen soll, zeigen die Abbildungen A.1 und A.2.

X und Y, wie auch U und V haben univariate Normalverteilungen. Daneben bestimmt die Korrelation zwischen den Variablen die genaue Form der gemeinsamen Verteilung.

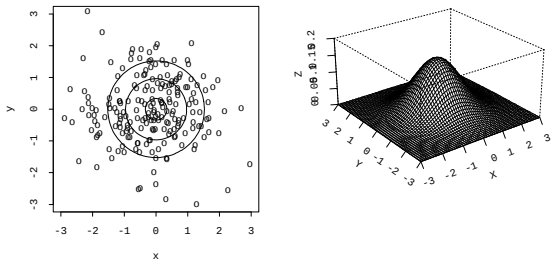


Abbildung A.1: Bivariate Normalverteilungen mit $\rho = 0$

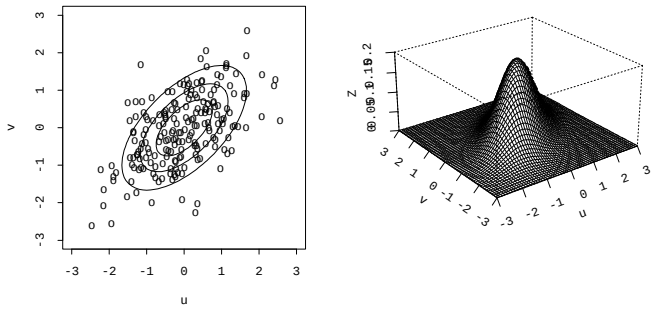


Abbildung A.2: Bivariate Normalverteilungen mit $\rho = 0.6$

Je grösser die Korrelation ρ, desto enger werden die elliptischen Kontourlinien (Punkte mit gleicher Dichte).

Eine bivariate Normalverteilung wird demnach durch fünf Parameter festgelegt: $\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho_{xy}$. Für eine 3-dimensionale Normalverteilung braucht es schon 9 Parameter und die Liste wird länger und länger mit wachsender Dimension. Benutzt man die Matrixnotation, so genügt die Angabe des Vektors der Erwartungswerte und der Kovarianzmatrix.

Also zum Beispiel:

$$\mathbf{Z} \sim N(\boldsymbol{\mu}, Cov(\mathbf{Z}))$$

mit

$$\mathbf{Z} = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{und} \quad Cov(\mathbf{Z}) = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$