Statistik-Praktikum: Lineare Modelle

Sebastian Mentemeier

06.10.2010

Albert Einstein erhielt den Nobelpreis 1921 nicht für die Entwicklung der Relativitätstheorie, sondern für die Erklärung des photoelektrischen Effektes: Fällt Licht geeigneter Frequenz auf eine Metallplatte, so werden Elektronen herausgelöst. Zwischen der Frequenz f der Photonen und der kinetischen Energie E_{kin} der Elektronen (sowie der nötigen Austrittsarbeit W_A) besteht die fundamentale Beziehung

$$E_{kin} = h \cdot f - W_A,$$

wobei h das Plancksche Wirkungsquantum bezeichnet. Da die kinetische Energie der Elektronen gemessen werden kann, und die Frequenz vorgebbar ist, ermöglicht diese Identität eine experimentelle Bestimmung des Wertes von h. Da typischerweise Messungenauigkeiten $\varepsilon^{(i)}$ auftreten, gilt für die gemessenen Daten die Gleichung

$$E_{kin}^{(i)} = h \cdot f^{(i)} - W_A + \varepsilon^{(i)},$$

hierbei indiziert i die Messungen. Dies mag als Motivation für die Beschäftigung mit *Linearen Modellen* dienen, deren primäres Ziel die Bestimmung obiger Konstanten ist.

Allgemein wird in einem Linearen Modell angenommen, dass zwischen zwei zwei Größen X und Y ein *zufällig gestörter* funktionaler Zusammenhang

$$Y = f(X) + \varepsilon$$

besteht, d.h. Y ist im wesentlichen eine *deterministische* Funktion f von X, allerdings mit einem additiven (zufälligen) Fehler ε. Hierauf bezieht sich der Begriff *linear*: Nicht f muss eine lineare Funktion sein, sondern der Fehler muss additiv eingehen (es wäre ja auch ein Zusammenhang $Y = f(X, \varepsilon)$ denkbar).

Ist f aber eine lineare Funktion, so befinden wir uns im Spezialfall der linearen Regression, auf die wir uns um der Einfachheit der Darstellung wegen größtenteils beschränken werden.

1 Lineare Zusammenhänge erkennen

1.1 Streudiagramme

Der erste Schritt zur Datenanalyse sollte immer die graphische Darstellung sein, so auch hier. Gegeben zwei Beobachtungen x und y, kann ein Streudiagramm deutliche Hinweise geben, welche funktionalen Abhängigkeiten zwischen den Größen X und Y vorliegen mögen.

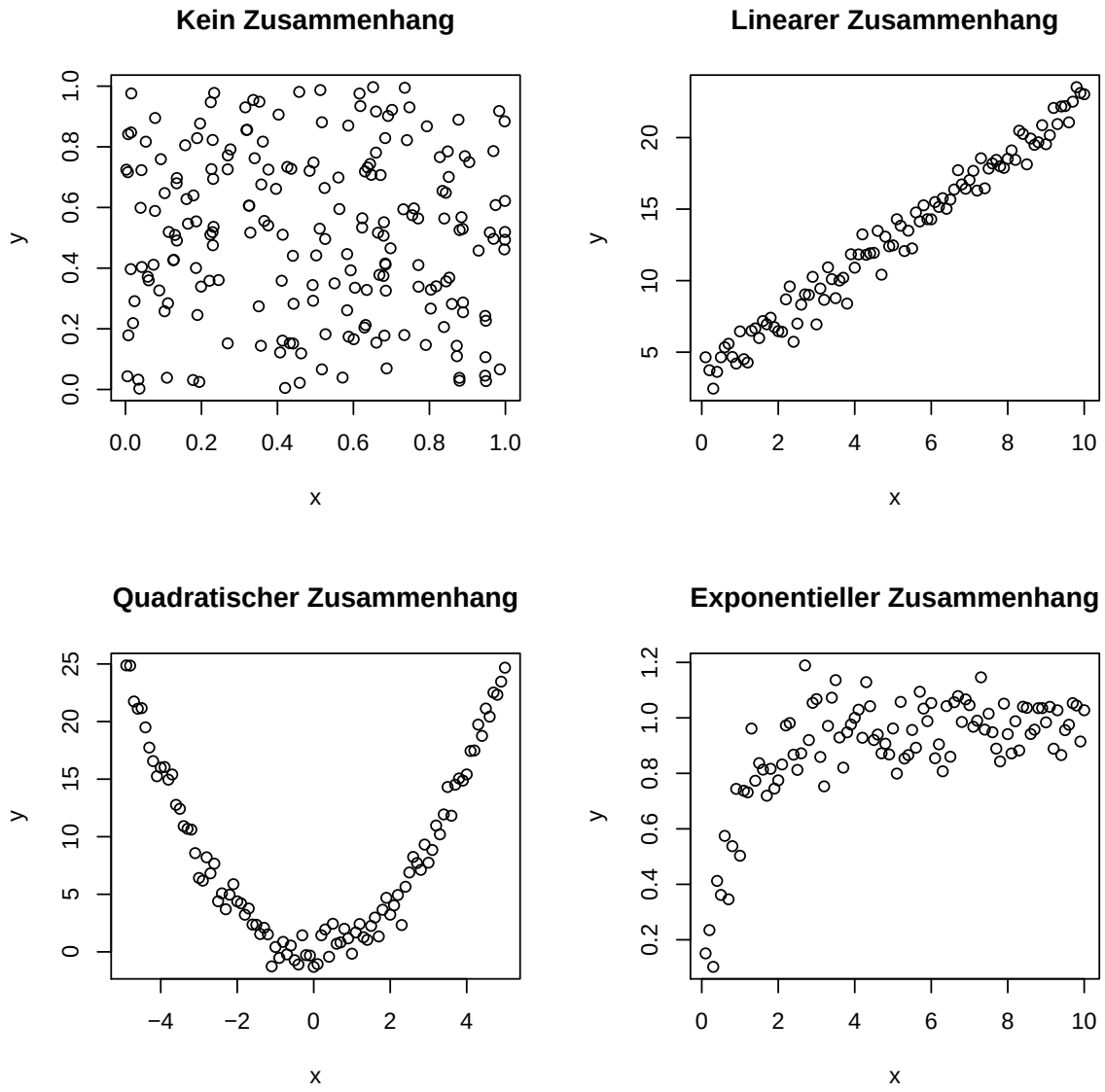

Abbildung 1: Verschiedene Arten funktionaler Abhängigkeiten

Zur Erinnerung: Streudiagramme werden in R mit `plot(x,y)` erzeugt.

1.2 Korrelationskoeffizient

Deutet das Streudiagramm auf einen linearen Zusammenhang hin, so ist der *empirische Korrelationskoeffizient*, oder auch *Bravais-Pearson-Korrelationskoeffizient* genannt, ein Maß für die Stärke dieses Zusammenhangs. Wer sich an den Korrelationskoeffizienten aus der Stochastik erinnert, dies ist einfach das empirische Gegenstück, (Ko-)Varianzen werden durch die Stochproben(ko)varianzen ersetzt:

Definition 1.1 Der *empirische Korrelationskoeffizient nach Bravais-Pearson* ist definiert durch

$$r \;=\; r_{XY} \;=\; \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}} \;=\; \frac{\tilde{s}_{XY}}{\tilde{s}_X\tilde{s}_Y},$$

wobei

$$\tilde{s}_{XY} \;:=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

die Stichprobenkovarianz ist.

Dieses r_{XY} wird in R mit dem Befehl `cor(x,y)` berechnet. Der Korrelationskoeffizient r nimmt Werte im Intervall $[-1,1]$ an.

- $r>0$: positive Korrelation,
  gleichsinniger linearer Zusammenhang,
  Tendenz: Werte (x_i,y_i) um eine Gerade mit positiver Steigung liegend
- $r<0$: negative Korrelation,
  gegensinniger linearer Zusammenhang,
  Tendenz: Werte (x_i,y_i) um eine Gerade mit negativer Steigung liegend
- $r=0$: keine Korrelation,
  unkorreliert, kein linearer Zusammenhang

In einem groben Raster lassen sich Korrelationen einordnen (siehe [1, S.139]) durch

„schwache Korrelation“	$\lvert r\rvert < 0.5$
„mittlere Korrelation“	$0.5 \le \lvert r\rvert < 0.8$
„starke Korrelation“	$0.8 \le \lvert r\rvert$.

Es gilt $|r|=1$ genau dann, wenn die Werte (x_i,y_i) auf einer Geraden liegen.

Vermutet man keinen linearen sondern einen monotonen Zusammenhang, so verwendet man häufig den *Spearman'schen Korrelationskoeffizienten*. Hierzu bildet man x_i und y_i auf ihre Ränge $\mathrm{rg}(x_i)$ und $\mathrm{rg}(y_i)$ im Tupel $(x_1,\ldots,x_n)$ bzw. $(y_1,\ldots,y_n)$ ab und bildet den Bravais-Pearson'schen Korrelationskoeffizienten für $((\mathrm{rg}(x_1),\mathrm{rg}(y_1)),\ldots,(\mathrm{rg}(x_n),\mathrm{rg}(y_n)))$.

```
(7 5 9 6)     (8 9 6 7)
    ↓             ↓
(2 1 4 3)     (3 4 1 2)
```

Der Rang ist dabei wie folgt definiert: $\mathrm{rg}(x_i) = \#\{j : x_j \leq i\}$, falls die x_j alle verschieden sind. Stimmen mehrere x_j überein, so kommen mehrere Zahlen als Rang infrage. $\mathrm{rg}(x_i)$ ist dann als das arithmetische Mittel dieser Ränge definiert.

Definition 1.2 Der *Korrelationskoeffizient nach Spearman* ist durch

$$r_{\mathrm{SP}} = \frac{\sum_{i=1}^{n}(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_X)(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_Y)}{\sqrt{\sum_{i=1}^{n}(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_X)^2 \sum_{i=1}^{n}(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_Y)^2}}$$

definiert.

Um den Spearman'schen Korrelationskoeffizienten zu bestimmen, muss das optionale Argument `method` auf `"spearman"` gesetzt werden. Der Korrelationskoeffizient r_{SP} nimmt ebenfalls Werte im Intervall $[-1, 1]$ an.

- $r_{\mathrm{SP}} > 0$: gleichsinniger monotoner Zusammenhang,
  Tendenz: x groß $\Leftrightarrow$ y groß, x klein $\Leftrightarrow$ y klein
- $r_{\mathrm{SP}} < 0$: gegensinniger monotoner Zusammenhang,
  Tendenz: x groß $\Leftrightarrow$ y klein, x klein $\Leftrightarrow$ y groß
- $r_{\mathrm{SP}} = 0$: kein monotoner Zusammenhang

2 Lineare Modelle

Wir präzisieren das eingangs beschriebene Modell im Fall eines linearen Zusammenhangs, und stellen insbesondere Modellannahmen über die Fehler ε auf:

Standardmodell der linearen Einfachregression

Es gilt $\quad Y_i = \alpha + \beta x_i + \varepsilon_i$.
Dabei sind:
$Y_1, \ldots, Y_n$ beobachtbare metrische Zufallsvariablen
$x_1, \ldots, x_n$ gegebene deterministische Werte oder Realisierungen einer metrischen Zufallsvariablen X.
$\varepsilon_1, \ldots, \varepsilon_n$ unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $\mathbb{E}(\varepsilon_i) = 0$ und $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Aus den Eigenschaften der Fehlervariablen folgen entsprechende Eigenschaften für die Zielvariablen. Es gilt

$$\mathbb{E}(Y_i) = a + \beta x_i, \quad \mathrm{Var}(Y_i) = \sigma^2.$$

Die Eigenschaft gleicher Varianz der Fehlervariablen ε_i wird als *Homoskedastizität* bezeichnet. Sie wird oft dadurch verletzt, dass die Varianzen der ε_i und damit der Y_i mit größer werdenden x-Werten ebenfalls zunehmen. Ob die Annahme der Homoskedastizität kritisch ist, sieht man oft schon aus den Streudiagrammen. Sie sollte aber ebenso *a posteriori* anhand der Residuen $\hat{\varepsilon}_i$ (s.u.) überprüft werden.

2.1 Kleinste-Quadrate-Schätzer

Das hier übliche Vorgehen besteht darin, die Summe der Quadrate der Entfernungen der Punkte y_i von den Prognosen $\hat{y}_i = \alpha + \beta x_i$ zu minimieren, d. h., wir suchen

$$\text{argmin}_{(\alpha,\beta)\in\mathbb{R}^2} Q(\alpha,\beta) = \text{argmin}_{(\alpha,\beta)\in\mathbb{R}^2} \frac{1}{n}\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2.$$

Durch eine gewöhnliche Kurvendiskussion (siehe [1, S. 155]) erhält man

$$\begin{aligned} \hat{\alpha} &= \overline{y} - \hat{\beta}\overline{x}, \\ \hat{\beta} &= \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \end{aligned}$$

Die geschätzten Fehler $\hat{\varepsilon}_i = y_i - \hat{y}_i$ mit $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ bezeichnet man als *Residuen.*

Anhand der Residuen können nun die Modellannahmnen überprüft werden: Die Residuen sollten mit ähnlicher Schwankungsbreite (Homoskedastizität) um den Nullpunkt streuen ($\mathbb{E}(\varepsilon_i) = 0$). Dies kann mittels eines Residualplots geschehen:

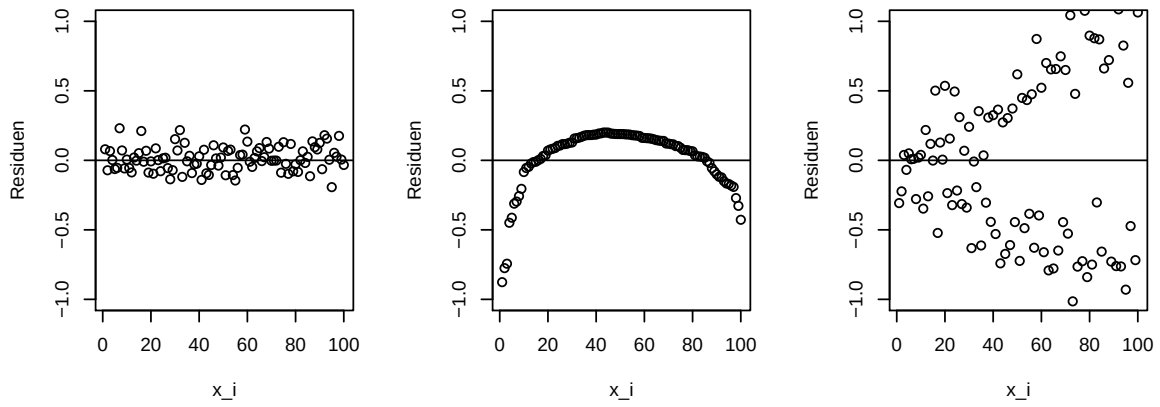

Abbildung 2: Verschiedene Residualplots

Die erste Abbildung zeigt ein ideales Verhalten der Residuen: Sie schwanken unsystematisch um die horizontale Achse und sind nahe bei Null. Der Verlauf der Residuen in der zweiten Abbildung legt die Vermutung nahe, dass eine nicht lineare Abhängigkeit zwischen den Merkmalen besteht, die nicht durch das Modell erfasst wird. Die dritte Abbildung weist darauf hin, dass die Bedingung der Homoskedastizität verletzt ist. Hier verändert sich die Variabilität der Residuen mit den Werten der Einflussgröße X. In diesem Fall hilft manchmal eine Transformation der beobachteten Variablen, $Y \to \exp(Y)$.

2.2 Das Bestimmtheitsmaß R^2

Mit Hilfe der Residuen kann man nun für jeden einzelnen Datenpunkt überprüfen, wie gut er aufgrund des Modells vorhergesagt worden wäre. Damit haben wir aber noch kein Maß gefunden, mit dem wir die Güte des Modells insgesamt beurteilen können. Ein solches Maß können

wir über die sogenannte *Streuungszerlegung* erhalten. Die dahinterstehende Frage ist: Welcher Anteil der Streuung der y_i (um ihr arithmetisches Mittel $\overline{y}$) lässt sich durch die Regression von Y auf X erklären? Dazu seien

$$\begin{aligned} \text{SQT} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad (\textbf{S}\text{um of S}\textbf{q}\text{uares }\textbf{T}\text{otal}), \\ \text{SQE} &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \qquad (\textbf{S}\text{um of S}\textbf{q}\text{uares }\textbf{E}\text{xplained}), \\ \text{SQR} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (\textbf{S}\text{um of S}\textbf{q}\text{uares }\textbf{R}\text{esiduals}). \end{aligned}$$

Satz 2.1 (von der Streuungszerlegung)

$$\text{SQT} = \text{SQE} + \text{SQR},$$

d. h.,

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Liegen alle Punkte auf einer Geraden, so gilt $\text{SQR} = 0$.

Je größer die Residualstreuung ist, desto schlechter beschreibt das lineare Regressionsmodell die Daten.

Als Maßzahl für die Güte der Modellanpassung verwendet man das Bestimmtheitsmaß bzw. den Determinationskoeffizienten R^2.

Definition 2.2 Das Bestimmtheitsmaß R^2 ist durch

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = \frac{\text{SQT} - \text{SQR}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}}$$

definiert.

Das Bestimmtheitsmaß steht in direkter Beziehung zum Korrelationskoeffizienten nach Bravais-Pearson, was auch die Benennung erklärt:

Satz 2.3

$$R^2 = r_{XY}^2.$$

Beweis. Zeige zuerst, dass der Mittelwert der prognostizierten Werte $\hat{y}_i$ mit dem der beobachteten Werte y_i übereinstimmt:

$$\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i) = \hat{\alpha} + \hat{\beta}\overline{x} = (\overline{y} - \hat{\beta}\overline{x}) + \hat{\beta}\overline{x} = \overline{y}.$$

Damit folgt

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2 = \sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} + \hat{\beta}\overline{x})^2 = \hat{\beta}^2\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

und somit für R^2

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{\hat{\beta}^2 \sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{s_{XY}^2 \cdot s_X^2}{(s_X^2)^2 \cdot s_Y^2} = \left(\frac{s_{XY}}{s_X s_Y}\right)^2 = r_{XY}^2.$$

□

3 Gütekriterien für Schätzer, Schätzer für Varianz, Standardfehler

So wie wir die Störungen ε_i der funktionalen Abhängigkeit als Realisierungen von unabhängig identisch verteilten Zufallsgrößen modelliert haben, so können wir auch die anhand der Beobachtungen berechneten Werte für $\hat{\alpha}$ und $\hat{\beta}$ als Realisierungen von Zufallsgrößen A bzw. B ansehen, in der Tat sind beides Funktionen der Y_i, und damit der ε_i. Ein Maß für die Schwankungsbreite der Schätzwerte ist dann durch die Varianz von A bzw. B gegeben:

$$\begin{aligned} B &= \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n x_i Y_i - n\overline{x}\overline{Y}}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ A &= \overline{Y} - B\overline{x}. \end{aligned}$$

Damit folgt

$$\begin{aligned} \mathrm{Var}(B) &= \frac{1}{(s_X^2)^2}\left(\mathrm{Var}(\sum_{i=1}^n x_i Y_i) - 2\mathrm{Cov}(\sum_{i=1}^n x_i Y_i, n\overline{x}\overline{Y}) + \mathrm{Var}(n\overline{x}\overline{Y})\right) \\ &= \frac{1}{(s_X^2)^2}\left(\sum_{i=1}^n x_i^2\sigma^2 - 2\overline{x}\sum_{i=1}^n x_i \,\mathrm{Cov}(Y_i, \sum_{j=1}^n Y_j) + n^2\overline{x}^2\frac{\sigma^2}{n}\right) \\ &= \frac{1}{(s_X^2)^2}\left(\sum_{i=1}^n x_i^2\sigma^2 - 2n\overline{x}^2\sigma^2 + n\overline{x}^2\sigma^2\right) \\ &= \frac{1}{(s_X^2)^2}\left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right)\sigma^2 = \frac{s_X^2}{(s_X^2)^2}\sigma^2 = \frac{\sigma^2}{s_X^2}, \end{aligned}$$

und analog

$$\mathrm{Var}(A) = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n \cdot s_X^2}.$$

3.1 Gleichmäßig beste erwartungstreue Schätzer

Die Schätzer $\hat{a}$ und $\hat{b}$ sind *erwartungstreue Schätzer*, d.h.

$$\mathbb{E}(A) = \alpha, \quad \mathbb{E}(B) = \beta.$$

Die Varianz eines erwartungstreuen Schätzers ist somit gleich der erwarteten (d.h. unter dem durch die Modellannahmen spezifizierten Wahrscheinlichkeitsmaß) quadratischen Abweichung des geschätzten Wertes vom erwarteten Wert:

$$\mathrm{Var}(A) = \mathbb{E}(A - \mathbb{E}(A))^2 = \mathbb{E}(A - \alpha)^2.$$

Da wir stets nur eine Realisierung $\hat{\alpha}^2$ des Schätzers A beobachten, sollte dessen Varianz möglichst klein sein. Dies motiviert folgende Definition:

Definition 3.1 Ein *gleichmäßig bester erwartungstreuer Schätzer* A^* für den Parameter a ist ein erwartungstreuer Schätzer, so dass für jeden weiteren erwartungstreuen Schätzer A' gilt und jeden möglichen tatsächlichen Parameterwert a gilt

$$\mathbb{E}(A^* - a)^2 \leq \mathbb{E}(A' - a)^2.$$

Dies in unserem Fall nachzuprüfen, benötigt die Theorie suffizienter und vollständiger Statistiken, weshalb wir nur ohne Beweis notieren, dass unter der Zusatzannahme, dass die Fehler ε_i unabhängig identisch *normalverteilt* sind, die oben definierten Schätzer $\hat{\alpha}$ und $\hat{\beta}$ gleichmäßig beste erwartungstreue Schätzer für α bzw. β sind. Beachte, dass sie bereits per Konstruktion die quadratischen Abweichungen minimieren.

3.2 Konsistenz

Erinnern wir uns an den Beweis des schwachen Gesetzes der großen Zahlen, so folgt aus $\lim_{n\to\infty} \mathrm{Var}(A) = 0$, dass A in Wahrscheinlichkeit gegen den zu schätzenden Wert α konvergiert. Dabei meint $n \to \infty$ eine Erhöhung der Stichprobenzahl, formal korrekt müsste die Folge A_n von Schätzern für Stichproben vom Umfang n bei wachsendem n betrachtet werden.

Definition 3.2 Ein Schätzer, für den $A \xrightarrow{\mathbb{P}} \alpha$ für $n \to \infty$ gilt, heißt *schwach konsistent*.

Starke Konsistenz ist analog als $\mathbb{P}$-*f.s.*-Konvergenz von A gegen α definiert.

Für die Konsistenz obiger Schätzer ist folgende *Konsistenzbedingung* notwendig und hinreichend:

$$\lim_{n\to\infty} s_X^2 = \infty.$$

Sie besagt, dass die x-Werte hinreichend stark um ihr arithmetisches Mittel streuen, offensichtlich ist dies notwendig für eine schärfere Bestimmung der Steigung, und damit auch des Achsenabschnittes.

3.3 Schätzung von σ^2

Typischerweise ist σ^2 nicht bekannt, wir nehmen lediglich an, dass alle ε_i dieselbe Varianz besitzen. Unter dieser Annahme kann σ^2 jedoch erwartungstreu geschätzt werden, und zwar mittels

$$\hat{\sigma^2} := \frac{1}{(n-2)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Der Faktor $(n-2)^{-1}$ (anstelle von n^{-1}) sorgt für die Erwartungstreue des Schätzers (hängt mit verbleibenden Freiheitsgraden zusammen - in die Schätzung von σ^2 sind schon die Schätzungen $\hat{\alpha}$ und $\hat{\beta}$ über $\hat{y}_i$ eingeflossen). Im Falle normalverteilter Fehler ist dies bereits schon ein GBES für σ^2.

3.4 Standardfehler

Mit der Schätzung für σ^2 können wir nun auch $\mathrm{Var}(A)$ und $\mathrm{Var}(B)$ schätzen. Diese Schätzer bezeichnen wir mit $\hat{\sigma}_{\hat{\alpha}}$ und $\hat{\sigma}_{\hat{\beta}}$, und nennen sie die *Standardfehler* von $\hat{\alpha}$ bzw. $\hat{\beta}$.

3.5 Ein Beispiel

In unserem anfänglichen Beispiel des Photoeffektes könnte ein Versuch mit Bleisulfid folgende Messwerte ergeben:

f in 10^{14} Hz	5.187	5.491	6.876	7.402	8.191
E_{kin} in 10^{-19} J	1.13683	1.32419	2.24884	2.52618	3.00719

Die Untersuchung geschieht in R mittels des Aufrufes `lm(y ~ x)`, d.h. wir übergeben die funktionale Abhängigkeit (y *ist proportional zu* x). Der einfache Aufruf liefert nur spärliche Auskünfte, detaillierte Informationen liefert der anschließende Aufruf von `summary(...)`.

```
> Modell<-lm(Energie.in.J~Frequenz.in.Hz)
> summary(Modell)

Call:
lm(formula = Energie.in.J ~ Frequenz.in.Hz)

Residuals:
         1          2          3          4          5
-8.176e-22 -1.127e-21  4.570e-21 -6.487e-22 -1.977e-21

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.105e-19  7.954e-21  -26.46 0.000118 ***
Frequenz.in.Hz  6.265e-34  1.183e-35   52.98 1.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.009e-21 on 3 degrees of freedom
Multiple R-squared: 0.9989,     Adjusted R-squared: 0.9986
F-statistic:  2806 on 1 and 3 DF,  p-value: 1.481e-05
```

Schauen wir uns die Ausgabe einmal genauer an. Zunächst wiederholt R das spezifizierte Modell. Dann werden die Residuen beschrieben (bei einer größeren Stichprobenzahl werden nur die Quantile angegeben). Unter `Coefficients` werden nun die berechneten Schätzer angegeben. $\hat{\alpha}$ ist `(Intercept)`, also der Achsenabschnitt, $\hat{\beta}$ ist der geschätzte Koeffizient vor der Frequenz, wird also als Koeffizient von `Frequenz.in.Hz` ausgegeben. Unter `Estimate` steht der Schätzwert, unter `Std. Error` der Standardfehler, die Werte unter `t-value` und `Pr(>|t|)` sowie die Signifikanzsterne dahinter beziehen sich auf einen sog. t-Test auf die Hypothese, dass der zugehörige Koeffizient Null ist (die normalisierten Schätzer $\frac{\hat{\alpha}-\alpha}{\hat{\sigma}_\alpha}$ sind $t(n-2)$ verteilt) - mehr dazu in der Testtheorie. Uns genügen an dieser Stelle die Sternchen... Sie beschreiben die Wahrscheinlichkeit dafür, dass der Koeffizient in Wirklichkeit 0 ist, also irrelevant wäre.

Der `Residual standard error ... on` $(n-2)$ `degrees of freedom` ist unser $\hat{\sigma^2}$. Das R^2 wird als `Multiple R-squared` ausgegeben. Das `Adjusted R-squared` $\overline{R}^2$ wird nach einer leicht abgeänderten Formel berechnet:

$$\overline{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2)$$

(siehe [2, S.160]). Die Koeffizienten rühren daher, dass in die Berechnung von $\overline{R}^2$ erwartungstreue Schätzer für SQR und SQT einfließen (Freiheitsgrade...).

Die `F-statistic` und der `p-value` gehören zum sogenannten *Goodness of fit-Test*, der die Hypothese, dass alle Koeffizienten (außer dem Achsenabschnitt) Null sind, überprüft. Bei einem p-Wert größer als 0.05 sollte die Modellierung verworfen werden!

4 Lineare Modelle - polynomiale Regression

Definition 4.1 Ein lineares statistisches Modell liegt vor, wenn der Beobachtungsvektor Y der Regressionsgleichung

$$Y = \mathbf{A}\theta + \varepsilon$$

darstellen, dabei ist

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

der Vektor der abhängigen Variablen,

$$\mathbf{A} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix}$$

die Matrix der unabhängigen Variablen (*Designmatrix*),

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$$

der Vektor der Regressionskoeffizienten der mit A beschriebenen Variablen sowie

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

der Vektor der Fehler, die unabhängige Zufallsgrößen mit Erwartungswert 0 und Varianzen $\sigma_1^2, \dots, \sigma_n^2$ sind. Das Modell heisst homoskedastisch, falls $\sigma_1^2 = \dots = \sigma_n^2$, und sonst heteroskedastisch. Eine Funktion $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^p$ heisst dann *kleinste Quadrate-Schätzer (KQS) für* θ, wenn

$$\left\| x - \mathbf{A}\hat{\theta}(x) \right\| = \min_{\theta \in \mathbb{R}^p} \|x - \mathbf{A}\theta\| \tag{4.1}$$

für alle $x \in \mathbb{R}^n$ gilt.

Im eben betrachteten Fall der linearen Regression haben wir das Modell mit

$$\mathbf{A} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

untersucht.

Der Unterraum $\{\mathbf{A}\theta : \theta \in \mathbb{R}^p\}$ ist gerade das Bild von $\mathbf{A}$, und Gleichung (4.1) besagt gerade, dass $\mathbf{A}\hat{\theta}(x)$ der Projektion $\mathrm{pr}(x)$ von x auf das Bild $\mathrm{Im}(\mathbf{A})$ von $\mathbf{A}$ entsprechen muss. Nun gilt aber

$$\begin{aligned} \mathbf{A}\hat{\theta}(x) = \mathrm{pr}(x) \quad &\Leftrightarrow \quad (\mathbf{A}e_j)^\top (x - \mathbf{A}\hat{\theta}(x)) = 0 \ \text{ für } j = 1, \ldots, p \\ &\Leftrightarrow \quad e_j^\top \mathbf{A}^\top (x - \mathbf{A}\hat{\theta}(x)) = 0 \text{ für } j = 1, \ldots, p \\ &\Leftrightarrow \quad \mathbf{A}^\top x - \mathbf{A}^\top \mathbf{A}\hat{\theta}(x) = 0 \\ &\Leftrightarrow \quad \mathbf{A}^\top \mathbf{A}\hat{\theta}(x) = \mathbf{A}^\top x \end{aligned}$$

Die letzte Gleichung heisst *Normalengleichung*, aus ihr lässt sich unter der Voraussetzung, dass $\mathbf{A} \in M(n \times p, \mathbb{R})$ vollen Rang besitzt, der KQS berechnen:

Satz 4.2 In einem linearen Modell $Y = \mathbf{A}\theta + \varepsilon$ vollen Ranges (d.h. die Designmatrix besitze vollen Rang) gilt: Der KQS $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^p$ für θ ist durch

$$\hat{\theta}(x) = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top x$$

gegeben.

Im Falle normalverteilter Fehler gelten analoge Aussagen zu oben; der $\beta^\top \hat{\theta}$ ist in diesem Fall bereits ein gleichmäßig bester erwartungstreuer Schätzer für $\beta^\top \theta$ für jedes $\beta \in \mathbb{R}^d$ (verschärfter Satz von Gauß-Markow).

5 Logit-Modelle

Lineare Modelle eignen sich besonders für Regressionsanalysen, bei denen die Zielvariable stetig ist und - möglicherweise nach einer geeigneten Transformation - zumindest approximativ durch eine Normalverteilung modelliert werden kann. Zusätzlich muss sich der Erwartungswert der Zielvariablen durch eine Linearkombination von - möglicherweise ebenfalls transformierten - Kovariablen darstellen lassen. In vielen Anwendungen ist die Zielvariable jedoch nicht stetig, sondern binär, bspw. als Antwort auf die Frage

Haben Sie einen Riester-Renten-Vertrag abgeschlossen?

Ziel einer binären Regressionsanalyse ist die Modellierung und Schätzung des Effekts der Kovariablen auf die (bedingte) Wahrscheinlichkeit

$$\pi = \mathbb{P}\left(Y_i = 1 | x_{i1}, ..., x_{ip}\right) = \mathbb{E}\left[Y_i | x_{i1}, ..., x_{ip}\right]$$

für das Auftreten von $Y_i = 1$, gegeben die Kovariablenwerte $x_{i1}, ..., x_{ip}$. Die Zielvariablen werden dabei als (bedingt) unabhängig angenommen. Der intuitive Ansatz,

$$\pi_i = \theta_0 + \sum_{k=1}^{p} \theta_k x_{ik} + \varepsilon_i$$

zu modellieren, schlägt jedoch fehl, da $\pi_i \in [0,1]$ liegen muss, was jedoch für die rechte Seite keinesfalls vorausgesetzt werden kann.

Alle üblichen binären Regressionsmodelle verknüpfen daher die Wahrscheinlichkeit π_i durch eine Beziehung der Form

$$\pi_i = h(\eta_i) = h(\theta_0 + \sum_{k=1}^{p} \theta_k x_{ik})$$

mit dem linearen Prädiktor

$$\eta_i = \theta_0 + \sum_{k=1}^{p} \theta_k x_{ik}.$$

Dabei muss die *Responsefunktion* eine monoton wachsende Funktion $h : \mathbb{R} \to [0,1]$ sein, weshalb Verteilungsfunktionen die geeigneten Kandidaten sind. Im *logit*-Modell wird die *logistische Verteilungsfunktion*

$$h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

gewählt. Dann wird das lineare Modell

$$h^{-1}(Y) = \mathbf{A}\theta + \varepsilon$$

untersucht. Beachte, dass die absoluten Werte der geschätzten Koeffizienten keine sinnvolle Interpretation besitzen, wohl aber deren Quotienten - es kann also festgestellt werden, welche Kovariable den größten Einfluss unter den beobachteten ausübt.

In R werden logit-Modelle mit dem Aufruf

```
glm(Y ~ x1 + x2 + ... + xp, family = binomial(link="logit"))
```

aufgerufen. Wie bei linearen Modellen liefert der Aufruf von `summary` Signifikanzniveaus dafür, ob ein Koeffizient von 0 verschieden ist, d.h., ob die zugehörige Kovariable einen Einfluss auf die Zielgröße ausübt.

6 Verwendete R-Befehle

`lm` Berechnet Schätzer in einem linearen Modell, eine vermutete Abhängigkeit der Form $X = C + f(Y) + \varepsilon$ wird dabei als `x~f(y)` übergeben. Soll kein Achsenabschnitt (`Interception`) berechnet werden, so muss `x~f(y)+0` übergeben werden. Wird eine Abhängigkeit von mehreren Variablen untersucht, so wird bei Aufruf mit `x~y*z` eine Abhängigkeit der Form $X = C + Y + Z + Y \cdot Z + \varepsilon$ betrachtet, und bei Aufruf `x~y:z` nur die Abhängigkeit von $Y \cdot Z$. Mittels `summary(lm(...))` erhält man weitere Informationen.

`glm` R-Funktion zur Untersuchung generalisierter linearer Modelle. Das hier behandelte Logit-Modell

$$\pi_i = h(\eta_i) = h(\theta_0 + \sum_{k=1}^{p} \theta_k x_{ik})$$

mit der logistischen Verteilungsfunktion als Link-Funktion h wird mittels

```
glm(Y ~ x1 + x2 + ... + xp, family = binomial(link="logit"))
```

aufgerufen.

`predict` Nimmt ein lineares Modell, z.B. `lm(y~x)` als Argument und berechnet Werte für `x` anhand der geschätzten Steigungen und Achsenabschnitte, also gerade die $\hat{y}_i$.

`abline` Fügt eine Gerade in einen bestehenden Plot ein, insbesondere für lineare Regression: `abline(lm(y~x))` zeichnet die Regressionsgerade $\hat{y} = \hat{\alpha} + \hat{\beta}x$.

Literatur

[1] Fahrmeir, Künstler, Pigeot, Tutz: *Statistik. Der Weg zur Datenanalyse.* Springer 2007

[2] Fahrmeir, Kneib, Lang: *Regression.* Springer 2009