Regression

von Peter Pfaffelhuber

Version: 31. Oktober 2015

1 Einleitung

Oftmals will man mit Daten Zusammenhänge bestimmen, etwa zwischen der Größe einer Wohnung und dem Mietpreis, oder der Verkehrsdichte und der Durschschnittsgeschwindigkeit von Fahrzeugen. Weiter kann die *Zielvariable* oder *Beobachtung* (hier etwa Mietpreis und Durschschnittsgeschwindigkeit) von weiteren Einflüssen (*Covariate* oder *unabhängige Variable*) abhängen, etwa von der Lage der Wohnung, oder der Breite der Straße. In den meisten Situationen kann man ein Regressionsmodell verwenden, um Korrelationen zwischen Covariaten und Zielvariablen herauszufinden. Dieses ist für n Messungen (also etwa n Wohnungen oder n bestimmte Durchschnitsgeschwindigkeiten) und k Einflussgrößen von der Form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, ..., n \tag{1.1}$$

Hier sind $y_1, ..., y_n$ die Beobachtungen und $x_{i1}, ..., x_{ik}$ sind die Werte der Einflussgrößen auf die i-te Beobachtung. Um dies in ein statistisches Modell umzuwandeln, seien $\epsilon_1, ..., \epsilon_n$ (und damit auch $y_1, ..., y_n$) Zufallsvariable, und mit $x_{i0} := 1$ schreiben wir besser mit Vektoren[1]

$$Y_i = x_{i\cdot}\beta + \epsilon_i, \qquad i = 1, ..., n$$

oder mit $x = (x_{ij})_{i=1,...,n,j=0,...,k}$

$$Y = x\beta + \epsilon.$$

□

Bemerkung 1.1 (Einfache Regression). Der einfachste Fall tritt ein, wenn es nur eine einzige Covariate gibt; siehe auch Beipsiel 1.2. In diesem Fall verändert sich das Regressionsmodell zu

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, ..., n$$

Im Gegensatz dazu nennt man (1.1) für $k > 1$ *multiple Regression*.

Beispiel 1.2 (Regressionsanalyse mit `R`). Wir verwenden einen Datensatz `faithful` aus den 1980er Jahren, der in [R] verfügbar ist und dessen ersten Zeilen wir mittels

```
> head(faithful)
```

[1] Für uns ist im Folgenden x ein Spaltenvektor und $x^\top$ ein Zeilenvektor.

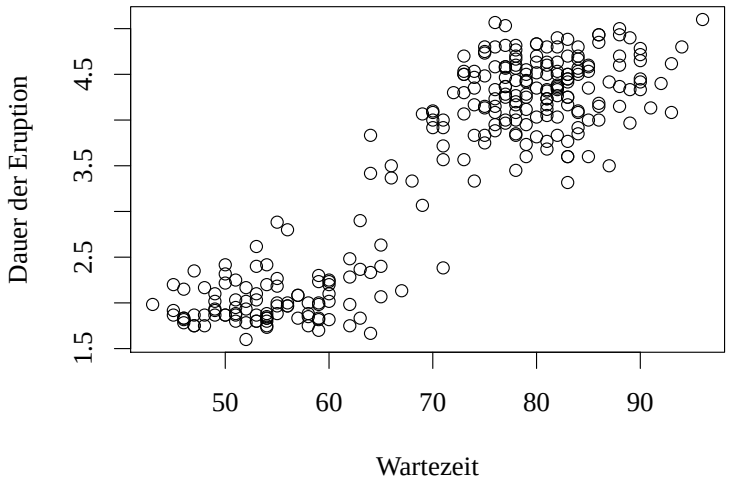

Abbildung 1.1: Das Datenbeispiel aus dem `faithful`-Datensatz, der in R zur Verfügung steht.

ansehen[2], was

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

liefert. Die Größe `waiting` steht für die Wartezeit bis zur nächsten Eruption des *Old Faithful Gaysier* im Yellowstone National Park der USA und `eruptions` für dessen Dauer. Um uns einen ersten Eindruck zu verschaffen, ob diese beiden Größen korreliert sind, plotten wir einfach mal die Datenpunkte. Mit

```
> duration = faithful$eruptions
> waiting = faithful$waiting
```

weisen wir die beiden Spalten des Datensatzes den Vektoren `duration` und `waiting` zu. Den gewünschten Plot erzeugen wir durch

```
> plot(waiting, duration, xlab="Wartezeit", ylab="Dauer der Eruption")
```

Das Ergebnis ist in Abbildung 1.1 abgebildet. [3]

[2] Das Kommando `head` liefert nur die ersten Zeilen des Datensatzes. Will man den Datensatz ganz ansehen, gibt man `faithful` ein.

[3] Um das Bild in ein Skript wie dieses hier einzubetten, ist es natürlich praktisch, wenn es als `pdf` vorliegt. In R habe ich deswegen die Befehle

```
> pdf(file = "fig1.pdf", width=7, height=5, family="Times", onefile=FALSE)
> par(mar=c(5,4,1,1), cex=1.5)
```

Offenbar besteht ein Zusammenhang zwischen der Wartezeit und der Dauer der Eruption. Wir werden in den folgenden Kapiteln herleiten, wie man sinnvollerweise eine *Regressionsgerade* durch die Datenwolke legt, die gut passt. Der entsprechende R-Befehl wird

```
> lm(eruptions ~ waiting, data=faithful)
```

lauten. Dies liefert den Output

```
Coefficients:
(Intercept)      waiting
   -1.87402      0.07563
```

Das bedeutet, dass R die Gerade

$$\hat{Y} = -1.87402 + 0.07563x$$

für die Wartezeit Y und die Dauer der Eruption x gefunden hat. Dies können wir auch grafisch in Abbildung 1.2 veranschaulichen.[4] Im Folgenden wollen wir diese Regressiongerade und ihre Eigenschaften diskutieren. Wir gehen dabei gleich zum Fall der multiplen Regression, in dem `waiting` auch mehr als eine Variable beinhalten hätte können. In Kapitel 7 kommen wir noch einmal auf das Beispiel zurück.

2 Das Modell

Das statistische Modell besteht aus den Daten Y und deren Verteilungen. Letztere hängen nur von den Werten β und den Verteilungen von ϵ ab. Wir bezeichnen die Verteilungen deswegen auch mit $\mathbb{P}_\beta$ (und spezifizieren damit die Abhängigkeit von der Verteilung von ϵ nicht genauer). Oftmals werden wir Annahmen über die Verteilung von ϵ treffen.

Annahme 2.1 (Gauß-Markov-Bedingungen). *Es gilt für ein* $\sigma^2 > 0$

$$\mathbb{E}_\beta[\epsilon_i] = 0, \qquad \mathbb{COV}_\beta[\epsilon_i, \epsilon_j] = \sigma^2 \delta_{ij}.$$

Hierfür schreiben wir auch

$$\mathbb{E}_\beta[\epsilon] = 0, \qquad \mathbb{COV}_\beta[\epsilon, \epsilon] = \mathbb{E}_\beta[\epsilon\epsilon^\top] = \sigma^2 I$$

für die $k \times k$*-Einheitsmatrix* I*, wobei alle Gleichungen in Vektorschreibweise gelesen werden.*

---

vor den `plot`-Befehl gestellt. (Der `par`-Befehl verkleinert die Ränder des Bildes für eine bessere Optik.) Nicht vergessen darf man allerdings, nach dem `plot`-Befehl auch noch

```
> dev.off()
```

einzugeben, erst dann kann die `pdf`-Datei fehlerfrei dargestellt werden.

[4] Praktisch ist hier der Befehl `abline`. Um die Gerade zu plotten, habe ich die Befehle

```
> coeffs=coefficients(lm(eruptions ~ waiting, data=faithful))
> coeffs=as.vector(coeffs)
> abline(coeffs)
```

benutzt. Der erste Befehl gibt die beiden Koeffizienten in einer Liste aus, der zweite wandelt diese in einen Vektor um und der dritte zeichnet die Regressionsgerade.

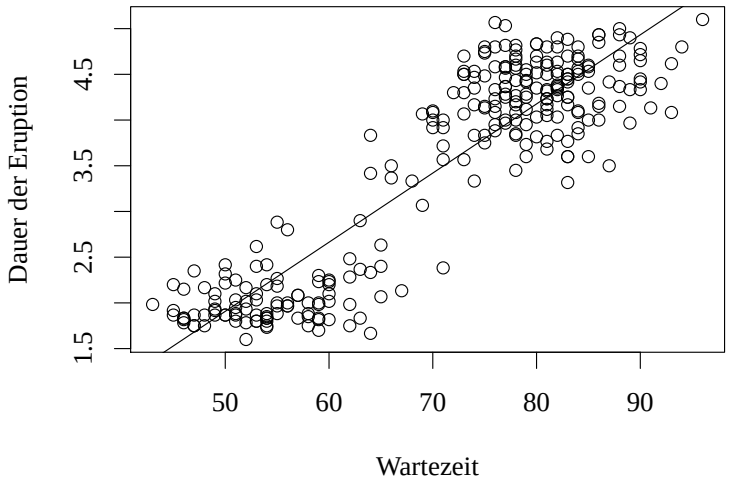

Abbildung 1.2: Die von R berechnete Regressionsgerade im `faithful`-Datensatz.

Stärker ist die Annahme, dass die Daten sogar unabhängig normalverteilt sind und gleiche Varianz haben.

Annahme 2.2 (Normalverteilungsannahme). *Für ein σ^2 ist $\epsilon_1, ..., \epsilon_n$ unabhängig und nach $\mathcal{N}(0, \sigma^2)$ verteilt. (Insbesondere sind alle Varianzen identisch.)*

Ein erstes Ziel ist es, die Parameter β zu bestimmen bzw. zu schätzen. Als Konsequenz erhält man dann die Vorhersage $\hat{Y} = x\hat{\beta}$. Der Fit des Modells ist umso besser, je kleiner die Residuen $Y - \hat{Y}$ sind. Deshalb versucht man, die Summe der Residuenquadrate zu minimieren, also suchen wir β, so dass[5]

$$RSS(\beta) := \sum_{i=1}^{n} (Y_i - x_{i\cdot}\beta)^2 = (Y - x\beta)^\top (Y - x\beta) = Y^\top Y - 2Y^\top x\beta + \beta^\top x^\top x\beta$$

minimal wird. Wir nehmen im Folgenden immer an, dass $x^\top x$ invertierbar ist (ansonsten müssen wir mit Pseudo-Inversen arbeiten). Eine notwendige Bedingung ist damit

$$0 = \tfrac{1}{2}\nabla RSS(\beta) = -Y^\top x + \beta^\top x^\top x = (x^\top x\beta - x^\top Y)^\top,$$

also ist ein Extremum von $\beta \mapsto RSS(\beta)$ bei

$$\hat{\beta} = (x^\top x)^{-1} x^\top Y.$$

Theorem 2.3 (Multiple Regression). *Falls $x^\top x$ invertierbar ist, so ist das Minimum von $RSS(\beta)$ eindeutig und bei*

$$\hat{\beta} = (x^\top x)^{-1} x^\top Y.$$

Für die Vorhersage

$$\hat{Y} := x\hat{\beta} (= x(x^\top x)^{-1} x^\top Y)$$

[5] RSS steht für *Residual Sum of Squares.*

gilt

$$Y - \hat{Y} = (I - x(x^\top x)^{-1}x^\top)\epsilon.$$

Außerdem stehen die Residuen $Y - \hat{Y}$ sowohl auf den Vorhersagen $\hat{Y}$, als auch auf den Spalten von x senkrecht.

Bemerkung 2.4 (Minimales RSS). Den minimalen Wert der *Residual Sum of Squares* bezeichnen wir mit

$$RSS := RSS(\hat{\beta}) = \sum_{i=1}^{n}(Y_i - x_{i\cdot}\hat{\beta})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (Y - \hat{Y})^\top(Y - \hat{Y}).$$

Beweis von Theorem 2.3. Die Tatsache, dass der Gradient von $RSS(\beta)$ bei $\hat{\beta}$ verschwindet, haben wir oben bereits nachgerechnet. Weiter ist die Hesse-Matrix von $RSS(\beta)$ (für alle β) durch $x^\top x$ gegeben ist, also durch eine positiv-definite Matrix. Für die zweite Behauptung setzen wir $\hat{Y}$ in das Modell ein und erhalten

$$\begin{aligned} Y - \hat{Y} &= (I - x(x^\top x)^{-1}x^\top)Y = (I - x(x^\top x)^{-1}x^\top)(x\beta + \varepsilon) \\ &= x\beta + \epsilon - x\beta - x(x^\top x)^{-1}x^\top\epsilon = (I - x(x^\top x)^{-1}x^\top)\epsilon. \end{aligned} \tag{$\circ$}$$

Weiter schreiben wir

$$\begin{aligned} (Y - \hat{Y})^\top\hat{Y} &= Y^\top x(x^\top x)x^\top Y - Y^\top x(x^\top x)^{-1}x^\top x(x^\top x)^{-1}x^\top Y = 0, \\ (Y - \hat{Y})^\top x &= Y^\top x - Y^\top x(x^\top x)^{-1}x^\top x = 0, \end{aligned}$$

woraus die behauptete Orthogonalität folgt. □

3 Schätzung der Modellparameter

Zwar haben wir nun Schätzer für $\hat{\beta}$ erhalten, allerdings wissen wir noch nichts über ihre Eigenschaften, etwa die Unverzerrtheit und Konsistenz. In diesem Abschnitt zeigen wir, dass $\hat{\beta}$ beide Eigenschaften besitzt (Theorem 3.1), und geben einen unverzerrten und konsistenten Schätzer für σ^2 an (Theorem 3.2).

Theorem 3.1 (Unverzerrtheit, Konsistenz von $\hat{\beta}$). *Gelten die Gauß-Markov-Bedingungen, so ist $\mathbb{E}_\beta[Y] = x\beta$ und $\hat{\beta}$ ist ein unverzerrter Schätzer für β. Weiter gilt*

$$\mathbb{COV}_\beta[\hat{\beta}, \hat{\beta}] = \sigma^2(x^\top x)^{-1}.$$

Gilt $tr((x^\top x)^{-1}) \xrightarrow{n\to\infty} 0$, so ist $\hat{\beta}$ ein konsistenter Schätzer für β.

Beweis. Es gilt

$$\mathbb{E}_\beta[\hat{\beta}] = (x^\top x)^{-1}x^\top(x\beta) = \beta,$$

woraus die Unverzerrtheit von $\hat{\beta}$ folgt. Weiter ist

$$\begin{aligned} \mathbb{COV}_\beta[\hat{\beta}, \hat{\beta}] &= ((x^\top x)^{-1}x^\top)\mathbb{COV}_\beta[Y, Y]x(x^\top x)^{-1}) \\ &= ((x^\top x)^{-1}x^\top)\mathbb{COV}_\beta[\epsilon, \epsilon]x(x^\top x)^{-1}) \\ &= ((x^\top x)^{-1}x^\top)\sigma^2 I x(x^\top x)^{-1}) = \sigma^2(x^\top x)^{-1}. \end{aligned}$$

Für die Konsistenz ist zunächst klar, dass $\mathbb{V}_\beta[\hat{\beta}_i] = \sigma^2((x^\top x)^{-1})_{ii}$. Da $(x^\top x)^{-1}$ als positiv-definite Matrix positive Diagonaleinträge hat, so folgt aus der Bedingung $\mathrm{tr}((x^\top x)^{-1}) \xrightarrow{n\to\infty} 0$, dass für $i = 1, ..., k$

$$\mathbb{V}_\beta[\hat{\beta}_i] \xrightarrow{n\to\infty} 0$$

und die Behauptung folgt. □

Zwar haben wir nun einen unverzerrten und konsistenten Schätzer für β, jedoch sollten wir auch in der Lage sein, σ^2 zu schätzen.

Theorem 3.2 (Ein Schätzer für σ^2). *Gelten die Gauß-Markov-Bedingungen, so ist*

$$\widehat{\sigma^2} := \frac{1}{n-k} RSS = \frac{1}{n-k-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ein unverzerrter und konsistenter Schätzer für σ^2.

Beweis. Zunächst ist mit $(\circ)$

$$\begin{aligned} RSS &= (Y - \hat{Y})^\top (Y - \hat{Y}) = \epsilon^\top (I - x(x^\top x)^{-1} x^\top)(I - x(x^\top x)^{-1} x^\top)\epsilon \\ &= \epsilon^\top (I - x(x^\top x)^{-1} x^\top)\epsilon. \end{aligned} \tag{$*$}$$

Wir berechnen mit Theorem 2.3[6]

$$\begin{aligned} \mathbb{E}_\beta[RSS] = \mathbb{E}_\beta[\epsilon^\top (I - x(x^\top x)^{-1} x^\top)\epsilon] &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}_\beta\Big[\epsilon_i (I - x(x^\top x)^{-1} x^\top)_{ij} \epsilon_j\Big] \\ &= \sigma^2 \sum_{i=1}^{n} ((I - x(x^\top x)^{-1} x^\top))_{ii} = \sigma^2 \mathrm{tr}(I - x(x^\top x)^{-1} x^\top) \\ &= \sigma^2 (\mathrm{tr}(I) - \mathrm{tr}(x^\top x (x^\top x)^{-1})) = \sigma^2 (n - k - 1), \end{aligned}$$

woraus die Unverzerrtheit folgt. Für die Konsistenz schreiben wir mit $(*)$

$$\widehat{\sigma^2} = \frac{1}{n-k-1} (\epsilon^\top \epsilon - \epsilon^\top x (x^\top x)^{-1} x^\top \epsilon).$$

Nach dem Gesetz der großen Zahlen ist $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \xrightarrow{n\to\infty}_{fs} \sigma^2$, also auch $\frac{1}{n-k-1}\epsilon^\top\epsilon \xrightarrow{n\to\infty}_{fs} \sigma^2$. Außerdem ist $(x^\top x)^{-1}$ positiv semi-definit und damit

$$\begin{aligned} \mathbb{E}_\beta[|\epsilon^\top x (x^\top x)^{-1} x^\top \epsilon|] &= \mathbb{E}_\beta[\epsilon^\top x (x^\top x)^{-1} x^\top \epsilon] = \mathbb{E}_\beta[\mathrm{tr}(\epsilon^\top x (x^\top x)^{-1} x^\top \epsilon)] \\ &= \mathrm{tr}(x (x^\top x)^{-1} x^\top \mathbb{E}_\beta[\epsilon\epsilon^\top]) = \sigma^2 \mathrm{tr}(x (x^\top x)^{-1} x^\top) = \sigma^2 (k+1), \end{aligned}$$

also $\frac{1}{n}\epsilon^\top x (x^\top x)^{-1} x^\top \epsilon \xrightarrow{n\to\infty}_{L^1} 0$. Insgesamt folgt also die Konsistenz

$$\widehat{\sigma^2} \xrightarrow{n\to\infty}_{p} \sigma^2.$$

□

[6] Wir verwenden hier die wohlbekannten Tatsachen aus der linearen Algebra, dass für Matrizen A, B

$$\begin{aligned} \mathrm{tr}(A + B) &= \mathrm{tr}(A) + \mathrm{tr}(B), \\ \mathrm{tr}(AB) &= \sum_i \sum_j A_{ij} B_{ji} = \sum_i \sum_j A_{ji} B_{ij} = \mathrm{tr}(BA). \end{aligned}$$

In Theorem 2.3 und im Beweis des letzten Theorems spielte die Matrix $I - x(x^\top x)^{-1}x^\top$ eine zentrale Rolle. Sie hat wichtige Eigenschaften, die wir nun sammeln. Wir wiederholen zunächst den Begriff der Idempotenz.

Bemerkung 3.3 (Idempotente Matrix). Eine quadratische Matrix A heißt idempotent, wenn $A^2 = A$. Eine solche Matrix hat als Eigenwerte nur 0 und 1.
Denn: Ist $Av = \lambda v$ für ein $v \neq 0$, so gilt auch $Av = A^2v = \lambda Av = \lambda^2 v$ und damit $\lambda = \lambda^2$. Dies ist aber nur für $\lambda \in \{0, 1\}$ möglich.

Lemma 3.4 (Eigenschaften von $I - x(x^\top x)^{-1}x^\top$). *Die Matrix*

$$\Sigma := I - x(x^\top x)^{-1}x^\top$$

ist idempotent, symmetrisch und positiv semi-definit. Weiter ist $(x(x^\top x)^{-1}x)_{ii} \leq 1$ für alle i und $rg(\Sigma) = n - k - 1$.

Beweis. Die Symmetrie und Idempotenz von Σ leitet man direkt her. Weiter ist klar, dass im letzten Beweis $RSS \geq 0$, ganz egal, welche Werte ϵ annimmt. Nun folgt die positive Semi-Definitheit von Σ aus $(*)$. Für die nächste Behauptung bemerken wir, dass die Diagonaleinträge einer positiv semi-definiten Matrix nicht-negativ sind. (Wäre der i-te Diagonaleintrag Σ_{ii}, so wäre $e_i^\top \Sigma e_i = \Sigma_{ii} < 0$, ein Widerspruch.) Es bleibt, die Aussage über den Rang von Σ zu zeigen. Da als Eigenwerte von Σ nur 0 und 1 in Betracht kommen (siehe Bemerkung 3.3), genügt es zu zeigen, dass die Summe der Eigenwerte von Σ gerade $n-k-1$ ist. Hierfür genügt es, $\mathrm{tr}(\Sigma) = n-k-1$ zu zeigen, wobei $\mathrm{tr}(\Sigma)$ die Spur von Σ ist (und bekanntermaßen invariant unter Ähnlichkeitstransformationen ist). Die Behauptung folgt nun aus

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(I) - \mathrm{tr}(x(x^\top x)^{-1}x^\top) = n - \mathrm{tr}(x^\top x(x^\top x)^{-1}) = n - k - 1.$$

□

4 Fit der Regressionsgeraden

Wir wollen nun untersuchen, wie gut der Fit der Regressionsgeraden $\hat{Y} = x\beta$ an die Daten Y ist. Am besten geht dies durch die empirische Korrelation von Y und $\hat{Y}$.

Definition 4.1 (Bestimmtheitsmaß). *Das Bestimmtheitsmaß ist definiert als*

$$R^2 = \frac{\left(\sum_{i=1}^n (Y_i - \overline{Y})(\hat{Y}_i - \overline{Y})\right)^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}.$$

Wir wollen das Bestimmtheitsmaß nun durch die RSS ausdrücken, um einen klareren Zusammenhang zu sehen.

Proposition 4.2 (Darstellung des Bestimmtheitsmaßes). *Es gilt*

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{RSS}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

Bemerkung 4.3 (Interpretation). Liegt ein Bestimmtheitsmaß von R^2 vor, so sagt man auch, dass die Regressionsgerade einen Anteil von R^2 an der Varianz der Daten erklärt. Grund hierfür ist die erste Darstellung aus der Proposition. Die *erklärte Varianz* ist ja gerade $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$, und die *Gesamtvarianz* ist $\sum_{i=1}^n (Y_i - \overline{Y})^2$.

Beweis. Zunächst zeigen wir die beiden Identitäten

$$RSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - (\hat{Y}_i - \overline{Y})^2,$$
$$\sum_{i=1}^{n}(Y_i - \overline{Y})(\hat{Y}_i - \overline{Y}) = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2.$$

Sind diese gezeigt, so folgt die Aussage einfach aus

$$R^2 = \frac{\left(\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2\right)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{RSS}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}.$$

Für die erste Identität wissen wir aus Theorem 2.3, dass $\hat{Y} - Y$ auf der ersten Spalte von x, also auf 1, senkrecht steht. Deshalb ist $\sum_{i=1}^{n}\hat{Y}_i = \sum_{i=1}^{n} Y_i$. Damit ergibt sich, da $\hat{Y}^\top$ auf $Y - \hat{Y}$ senkrecht steht, $\hat{Y}^\top\hat{Y} = (\hat{Y} - Y + Y)^\top\hat{Y} = Y^\top\hat{Y}$ und

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - (\hat{Y}_i - \overline{Y})^2 &= \sum_{i=1}^{n} Y_i^2 - \hat{Y}_i^2 = Y^\top Y - \hat{Y}^\top\hat{Y} \\ &= Y^\top(Y - \hat{Y}) = (Y - \hat{Y})^\top(Y - \hat{Y}) = RSS.\end{aligned}$$

Für die zweite Identität schreiben wir

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \overline{Y})(\hat{Y}_i - \overline{Y}) &= (Y - \overline{Y}I)^\top(\hat{Y} - \overline{Y}I) = (Y - \hat{Y} + \hat{Y} - \overline{Y}I)^\top(\hat{Y} - \overline{Y}I) \\ &= (\hat{Y} - \overline{Y}I)^\top(\hat{Y} - \overline{Y}I) = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2\end{aligned}$$

und alle Aussagen sind gezeigt. □

5 Das Gauß-Marov-Theorem

Die Schätzer $\hat{\beta}$ haben wir mit der Methode der kleinsten Quadrate erhalten. Nun geben wir einen berühmten Satz, dass dieses Vorgehen in der Tat in gewissem Sinn optimal ist.

Definition 5.1 (BLUE). *Im Regressionsmodell heißt jeder Schätzer $y \mapsto c_\ell^\top y$* linear. *Er heißt unverzerrt (für β), falls*

$$\mathbb{E}_\beta[c_\ell^\top Y] = \ell^\top\beta$$

für alle ℓ. Weiter heißt er Best Linear Unbiased Estimator (BLUE) *(für β), wenn er unverzerrt ist und $\mathbb{V}_\beta[c_\ell^\top Y] \leq \mathbb{V}_\beta[d_\ell^\top Y]$ für jeden linearen unverzerrten Schätzer $y \mapsto d_\ell^\top y$ ist.*

Bemerkung 5.2 (Ein linearer unverzerrter Schätzer). Aus Theorem 3.1 wissen wir bereits, dass $\hat{\beta} = (x^\top x)^{-1}x^\top Y$ ein unverzerrter Schätzer (für β) ist. Setzen wir $c_\ell = x(x^\top x)^{-1}x^\top\ell$, so ist damit $\mathbb{E}_\beta[c_\ell^\top Y] = \ell^\top\mathbb{E}_\beta[\hat{\beta}] = \ell^\top\beta$ und damit ist $y \mapsto c_\ell^\top y$ ein unverzerrter, linearer Schätzer. Das folgende Resultat zeigt, dass es sich auch um einen BLUE handelt.

Theorem 5.3 (Gauß-Markov-Theorem). *Sei $\hat{\beta} = (x^\top x)^{-1}x^\top Y$. Falls die Gauß-Markov-Bedingungen gelten, ist $y \mapsto \ell^\top(x^\top x)^{-1}x^\top y = \ell^\top\hat{\beta}$ ein BLUE.*

Beweis. Sei $y \mapsto d_\ell^\top y$ ein weiterer linearer, unverzerrtter Schätzer für β, also

$$\ell^\top \beta = \mathbb{E}_\beta[d_\ell^\top Y] = d_\ell^\top x\beta.$$

Da dies für alle ℓ gelten muss, ist also $x^\top d_\ell = \ell$. Wir schreiben nun mit Hilfe von Theorem 3.1

$$\begin{aligned}\mathbb{V}_\beta[d_\ell^\top Y] - \mathbb{V}_\beta[\ell^\top \hat{\beta}] &= d_\ell^\top \mathbb{COV}_\beta[Y,Y]d_\ell - \ell^\top \mathbb{COV}_\beta[\hat{\beta},\hat{\beta}]\ell \\ &= \sigma^2 d_\ell^\top d_\ell - \sigma^2 d_\ell^\top x(x^\top x)^{-1}x^\top d_\ell = \sigma^2 d_\ell^\top (I - x(x^\top x)^{-1}x^\top)d_\ell \geq 0\end{aligned}$$

wegen Lemma 3.4. □

6 Statistische Tests im Regressionsmodell

Oft will man herausfinden, ob man bei einer Regression auch mit weniger Covariaten auskommt. Könnte man etwa auf die i-te Covariate verzichten, so würde das auf ein Modell mit $\beta_i = 0$ hinauslaufen. Mit anderen Worten wollen wir im Regressionsmodell $H_0 : \beta_i = 0$ gegen $H_1 : \beta_i \neq 0$ testen. Etwas allgemeiner beschreiben wir im Folgenden Tests von $H_0 : A\beta - \gamma = 0$ für $A \in \mathbb{R}^{m\times(k+1)}$ mit Rang $m \leq k+1$ und $\gamma \in \mathbb{R}^m$ gegen $H_1 : A\beta - \gamma \neq 0$. Die Teststatistik wird dann eine F-Verteilung besitzen, die wir zunächst definieren.

Definition 6.1 (F-Verteilung). *Seien $X_1, ..., X_k, Y_1, ..., Y_l$ unabhängig und nach $\mathcal{N}(0,1)$ verteilt. Dann heißt die Verteilung von*

$$\frac{(X_1^2 + \cdots + X_k^2)/k}{(Y_1^2 + \cdots + Y_l^2)/l}$$

F-Verteilung mit Freiheitsgraden k und l oder $F_{k,l}$. Ihr p-Quantil bezeichnen wir mit $F_{k,l,p}$.

Bemerkung 6.2 (Äquivalente Formulierung). Bekanntermaßen hat $X_1^2 + \cdots + X_k^2$ (für $X_1, ..., X_k$ unabhängig nach $\mathcal{N}(0,1)$ verteilt) gerade eine χ_k^2-Verteilung (d.h. eine χ^2-Verteilung mit k Freiheitsgraden. Sind also $Z_1 \sim \chi_k^2$ und $Z_2 \sim \chi_l^2$ zwei unabhängige χ^2-Verteilungen, so ist

$$\frac{Z_1/k}{Z_2/l} \sim F_{k,l}.$$

Wir werden zwei Eigenschaften von mehrdimensionalen Normalverteilugen benötigen, die wir nun wiederholen.

Bemerkung 6.3 (Mehrdimensionale Normalverteilung). Sei $b \in \mathbb{R}^k$ und Σ symmetrisch und positiv semi-definit.

1. Ist $Y \sim \mathcal{N}(b, \Sigma)$, dann ist $AY \sim \mathcal{N}(Ab, A\Sigma A^\top)$.
   Denn: Es gilt $\mathbb{E}[AY] = Ab$ und

   $$\mathbb{COV}[AY, AY] = \mathbb{E}[(AY - Ab)(AY - Ab)^\top] = \mathbb{E}[AYY^\top A^\top] - Abb^\top A^\top = A\Sigma A^\top$$

2. Ist $Y \sim \mathcal{N}(0, \Sigma)$ und Σ eine idempotente Matrix von Rang r. Dann ist $Y^\top \Sigma Y \sim \chi_r^2$.
   Denn: Da Σ symmetrisch ist, und Σ nach Bemerkung 3.3 als Eigenwerte nur 0 und 1 hat, gibt es ein O orthogonal und $D = \mathrm{diag}(1, ..., 1, 0, ..., 0)$ mit $\mathrm{rg}(D) = r$, so dass $ODO^\top = \Sigma$. Damit ist $O^\top Y \sim \mathcal{N}(0, O^\top \Sigma O) = \mathcal{N}(0, D)$ und $Y^\top \Sigma Y = Y^\top ODO^\top Y \sim \chi_r^2$.

Theorem 6.4 (χ^2-Verteilungen im Regressionsmodell). *Es gelte Annahme 2.2. Ist $A\beta - \gamma = 0$, so ist unter $\mathbb{P}_\beta$ mit $\hat{\beta} = (x^\top x)^{-1}x^\top Y$*

$$\frac{1}{\sigma^2}(A\hat{\beta}-\gamma)^\top(A(x^\top x)^{-1}A^\top)^{-1}(A\hat{\beta}-\gamma) \sim \chi^2_m,$$
$$\frac{1}{\sigma^2}Y^\top(I - x(x^\top x)^{-1}x^\top)Y \sim \chi^2_{n-k-1}$$

und die Zufallsvariablen in den beiden Zeilen sind unabhängig.

Teilt man die beiden Zufallsvariablen des letzten Theorems durcheinander, so erhält man sofort eine F-verteilte Zufallsgröße, die später als Teststatistik dient.

Korollar 6.5 (Verteilung der Teststatistik). *Es gelte Annahme 2.2. Ist $A\beta - \gamma = 0$, so ist unter $\mathbb{P}_\beta$*

$$F := \frac{(A\hat{\beta}-\gamma)^\top(A(x^\top x)^{-1}A)^{-1}(A\hat{\beta}-\gamma)}{m\widehat{\sigma^2}} \sim F_{m,n-k-1}$$

mit $\widehat{\sigma^2}$ wie in Theorem 3.2.

Beweis von Theorem 6.4. Nach Theorem 3.1 ist $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(x^\top x)^{-1})$. Damit ist, falls $A\beta - \gamma = 0$ nach Bemerkung 6.3.1

$$A\hat{\beta} - \gamma \sim \mathcal{N}(A\beta - \gamma, \sigma^2 A(x^\top x)^{-1}A^\top) = \mathcal{N}(0, \sigma^2 A(x^\top x)^{-1}A^\top).$$

Da $A(x^\top x)^{-1}A^\top$ positiv definit ist, gibt es $(A(x^\top x)^{-1}A^\top)^{-1}$ und auch die Wurzel $(A(x^\top x)^{-1}A^\top)^{-1/2}$. Es ist

$$\tfrac{1}{\sqrt{\sigma^2}}(A(x^\top x)^{-1}A^\top)^{-1/2}(A\hat{\beta}-\gamma) \sim \mathcal{N}(0, I),$$

also auch

$$\tfrac{1}{\sigma^2}(A\hat{\beta}-\gamma)^\top(A(x^\top x)^{-1}A^\top)^{-1}(A\hat{\beta}-\gamma) \sim \chi^2_m.$$

Für die zweite Zufallsvariable erinnern wir an Lemma 3.4, wo wir gezeigt haben, dass

$$\Sigma := I - x(x^\top x)x^\top$$

symmetrisch, nicht-negativ definit, idempotent und von Rang $n-k-1$ ist. Da $\frac{1}{\sqrt{\sigma^2}}(I - x(x^\top x)^{-1}x^\top)Y = \frac{1}{\sqrt{\sigma^2}}\Sigma Y \sim N(0, \Sigma)$, ist nach Bemerkung 6.3.2

$$\frac{1}{\sigma^2}Y^\top(I - x(x^\top x)^{-1}x^\top)Y = \frac{1}{\sigma^2}Y^\top\Sigma Y \sim \chi^2_{n-k-1}.$$

Um die Unabhängigkeit einzusehen, schreiben wir

$$\begin{aligned}\mathbb{COV}_\beta[\hat{\beta}, \Sigma Y] &= \mathbb{E}_\beta[(x^\top x)^{-1}x^\top\epsilon\epsilon^\top(I - x(x^\top x)^{-1}x^\top)] \\ &= \sigma^2((x^\top x)^{-1}x^\top - (x^\top x)^{-1}x^\top x(x^\top x)^{-1}x^\top) = 0.\end{aligned}$$

Damit sind die beiden normalverteilten Zufallsvariablen $\hat{\beta}$ und $(I - x(x^\top x)^{-1}x^\top)Y$ unabhängig. Die Zufallsvariable der ersten Zeile des Theorems ist eine Funktion von $\hat{\beta}$ und die in der zweiten Zeile ist wegen

$$Y^\top(I - x(x^\top x)^{-1}x^\top)Y = Y^\top\Sigma Y = (\Sigma Y)^\top\Sigma Y$$

eine Funktion von ΣY. Damit sind beide Zufallsvariablen unabhängig. □

Beispiel 6.6 (Test auf $\beta_i = 0$). Wollen wir die Nullhypothese $H_0 : \beta_i = 0$ testen, So setzen wir $A = e_i^\top$ (dem i-ten kanonischen Basisvektor) und $\gamma = 0$. Im Beweis von Theorem 6.4 haben wir gesehen, dass

$$\frac{1}{\sqrt{\sigma^2}}(e_i^\top (x^\top x)^{-1} e_i)^{-1/2} e_i^\top \hat{\beta} = \frac{1}{\sqrt{\sigma^2((x^\top x)^{-1})_{ii}}}\hat{\beta}_i \sim \mathcal{N}(0,1)$$

unabhängig von $\frac{1}{\sigma^2}\widehat{\sigma^2} \sim \chi^2_{n-k-1}$ ist. Damit ist

$$T_i := \frac{\hat{\beta}_i}{\sqrt{((x^\top x)^{-1})_{ii}\widehat{\sigma^2}}} \sim t_{n-k-1}.$$

Deshalb ist für $\alpha \in (0,1)$ das Tupel (T, C) mit $C = (-\infty, t_{n-k-1,\alpha/2}) \cup (t_{n-k-1,1-\alpha/2}, \infty)$ ein Test von H_0 gegen $H_1 : \beta_i \neq 0$ zum Niveau α.

Beispiel 6.7 (Test auf $\beta = 0$). Will man testen, ob überhaupt ein Zusammenhang zwischen den Covariaten und Zielvariablen besteht, so überprüft man die Nullhypothese $H_0 : \beta_1 = \cdots = \beta_k = 0$. (Man beachte, dass $\beta_0 \neq 0$ zugelassen ist.) Hierzu verwenden wir in Korollar 6.5

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

und $\gamma = 0$. Damit ist (F, C) mit $C = (F_{k,n-k-1,1-\alpha}, \infty)$ ein Test von H_0 zum Signifikanzniveau α.

7 Ein R-Beispiel

Wir kommen noch einmal zurück zu den Daten von Geysir-Ausbrüchen aus Beispiel 1.2. Wir nehmen an, dass die Daten normalverteilt sind. Hierzu sehen wir uns nun die Ausgabe von

```
> summary(lm(eruptions ~ waiting, data=faithful))
```

an:

```
Call:
lm(formula = eruptions ~ waiting, data = faithful)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Zunächst wird hier eine Zusammenfassung der Residuen (`residuals`), also $Y - \hat{Y}$ angegeben. Dies geschieht durch Angabe des minimalen und maximalen Wertes, sowie durch Angabe der drei Quartile. Als nächstes werden die Werte $\hat{\beta}_0$ und $\hat{\beta}_1$ im Modell

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

angegeben. Die hier berichteten Standardfehler (`Std. Error`) sind durch

$$\widehat{\text{s.e.}}(\hat{\beta}_i) := \sqrt{\widehat{\sigma^2}(x^\top x)^{-1}_{ii}}$$

mit $\widehat{\sigma^2}$ aus Theorem 3.2 gegeben. Diese Formel begründet sich mit Theorem 3.1, wobei σ^2 durch einen Schätzer ersetzt wurde. Der nachfolgende t-Wert (`t value`) ist wie in Beispiel 6.6 berechnet. Der entsprechende p-Wert (`Pr(>|t|)`) ist sowohl für β_0 als auch für β_1 so klein, dass selbst zu einem sehr kleinen Signifikanzniveau die Hypothese $\beta_0 = 0$ bzw. $\beta_1 = 0$ nicht abgelehnt werden kann. Der residuale Standardfehler (`Residual standard error`) ist gerade $\sqrt{\widehat{\sigma^2}}$, Das Bestimmtheitsmaß (`Multiple R-squared`) haben wir in Proposition 4.2 bestimmt. (Der `Adjusted R-squared` ergibt sich dabei aus $1 - \widehat{\sigma^2}(n-1)/(\sum_{i=1}^n Y_i - \overline{Y})^2$; vergleiche mit Proposition 4.2) Schließlich wird die F-Statistik angegeben, die sich beim Test von $\beta_1 = 0$ zu $\beta_1 \neq 0$ ergibt; siehe Beispiel 6.7.

Literatur

[Sen90] A. Sen and M. Srivastava. Regression Analysis. Theory, Methods, and Applications. *Springer*, 1990.

[Fah07] L. Fahrmeir, T. Kneib and S. Lang. Regression. Modelle, Methoden und Anwendungen. *Springer*, 2. Auflage, 2009.

[R] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`, 2013.

Wichtige Formeln

$$
\begin{aligned}
Y &= x\beta + \epsilon && \\
\hat{Y} &= x\hat{\beta} && \text{Theorem 2.3} \\
\hat{\beta} &= (x^\top x)^{-1} x^\top Y && \text{Theorem 2.3} \\
Y - \hat{Y} &= (I - x(x^\top x)^{-1}x^\top)Y = (I - x(x^\top x)^{-1}x^\top)\epsilon && \text{Theorem 2.3 und } (\circ) \\
RSS &= (Y - \hat{Y})^\top (Y - \hat{Y}) = \epsilon^\top (I - x(x^\top x)^{-1}x^\top)\epsilon && \text{Bemerkung 2.4 und } (*) \\
\mathbb{COV}_\beta[\hat{\beta}, \hat{\beta}] &= \sigma^2 (x^\top x)^{-1} && \text{Theorem 3.1} \\
\widehat{\sigma^2} &= \frac{1}{n-k-1} RSS && \text{Theorem 3.2} \\
R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{RSS}{\sum_{i=1}^n (Y_i - \overline{Y})^2} && \text{Proposition 4.2}
\end{aligned}
$$