Angewandte Regression — Serie 1

1. Ein geldgieriger Autor will wissen, ob und wie die Seitenzahl eines Buches dessen Verkaufspreis beeinflusst. Er sucht dazu aus dem Katalog der "American Government books", Frühling 1988, folgende Daten von Paperback-Büchern, alle herausgegeben vom gleichen Verleger, heraus:

Seitenzahl	112	260	382	175	340	322	188	240	425	378
Preis	10.25	14.25	17.50	12.00	16.25	16.50	12.50	16.75	17.50	17.75

a) Lesen Sie die obigen Daten aus dem Datensatz
`http://stat.ethz.ch/Teaching/Datasets/WBL/buchpreis.dat`
ein und stellen Sie diese in einem Streudiagramm `preis` vs. `seiten` dar.

b) Rechnen Sie eine einfache lineare Regression
$y_i = \beta_0 + \beta_1 x_i + E_i$, $E_i \sim \mathcal{N}(0, \sigma^2)$ für diese Fragestellung.
Geben Sie die Standardfehler von $\hat{\beta}_0$ und $\hat{\beta}_1$ und die Gleichung der Regressionsgeraden an.

c) Zeichnen Sie ins Streudiagramm die Regressionsgerade ein.
Hat die Seitenzahl einen Einfluss auf den Verkaufspreis? Testen Sie die Nullhypothese: "Der Verkaufspreis wird von der Seitenzahl nicht beeinflusst." auf dem 5%-Niveau. Begründen Sie Ihre Antwort mit Hilfe des Regressions-Outputs.

d) Was kann man anhand der Parameter über die Preispolitik des Verlegers sagen?

e) Führen Sie eine Residuenanalyse durch. Falls es Ausreisser gibt, markieren Sie diese in der Skizze und geben Sie die Beobachtungsnummer an. (Hinweis: benütze `identify()`)

f) Geben Sie ein 95%-Vertrauensintervall für die Steigung an.

g) Was würde nach diesem Modell ein 600-seitiges Buch kosten? Wie stark vertrauen Sie dieser Prognose? Was ist das Problem?

h) Erklären Sie in Worten das Vertrauens- und Vorhersageband für $h\langle x\rangle$.

- Berechnen Sie von Hand das Vertrauens- und Vorhersageintervall für $x_0 = 340$.
- Zeichnen Sie das Vertrauens- und Vorhersageband in das Streudiagram und kontrollieren Sie das Ergebnis mit Ihrer Berechnung.
  R-Hinweise: Die R-Funktion `predict(..., se.fit=T, ...)` liefert eine Liste mit 2 Vektoren: Die Vorhersage $\hat{y}$ an bestimmten Stützstellen, die zugehörigen Standardfehler $\widehat{s.e.}(\hat{y})$ und zwei Zahlen (Anzahl Freiheitsgrade der Regression, geschätzte Standardabweichung des Fehlers). Zur Eingabe von Stützstellen $s = (s_0, \cdots, s_n)$ benütze man `predict(..., se.fit=T, newdata=data.frame(seiten=s))`

(Daten aus Sen, A. and Srivastava, M. (1990), *Regression Analysis; Theory, Methods, and Applications*)

2. In dieser Aufgabe betrachten wir vier Y-Variablen und vier X-Variablen, welche von Anscombe konstruiert wurden. Sie sind im Datensatz `anscombe` aufgeführt.

Betrachten Sie die vier Modelle $Y_i^{(k)} = \alpha + \beta \cdot X_i^{(k)} + E_i$ für $k = 1, ..., 4$.

Einlesen der Daten:

```
t.url <- "http://stat.ethz.ch/Teaching/Datasets/WBL/anscombe.dat"
d.anscombe <- read.table(t.url, header=TRUE)
```

a) Berechnen Sie für alle vier Modelle den Achsenabschnitt, die Steigung und die zugehörigen Standardfehler. Berechnen Sie ebenfalls $\widehat{\sigma}^2$ und R^2. Tabellieren Sie diese Werte.
Hinweis: Speichern Sie ausser dem Resultat `t.r` von `lm` auch das `summary` ab, `t.rs <- summary(t.r)`. Mit `str(t.rs)` sehen Sie, was dieses Objekt enthält. Sie finden u.a. die Komponenten `$coefficients` und `$sigma`.

b) Stellen Sie für alle vier Modelle die Regressionsgerade im Streudiagramm dar.

c) Kommentieren Sie die Resultate aus a) und b).

Quelle: F.J. Anscombe, Graphs in statistical analysis, American Statistician 27, 17-21 (1973)

3. Wir möchten die Verteilung der Schätzwerte für die Parameter α und β simulieren. Unser Modell sei $Y_i = 4 + 2x_i + E_i$ mit den folgenden x-Werten:

i	1	2	3	4	5	6	7	8	9	10
x_i	0	3	4	8	10	11	13	16	17	20

Der Messfehler E_i für einen Messwert Y_i sei normalverteilt mit $\mu = 0, \sigma^2 = 2$.

a) Simulieren Sie die 10 Messwerte Y_i nach dem Modell $Y_i = 4 + 2x_i + E_i$ 100-mal.

b) Schätzen Sie jeweils die Parameter α und β.

c) Betrachten Sie die Verteilung der geschätzten Parameter mit Hilfe eines Histogramms. Kommentar!

Bemerkung: Die 100 Simulationen erzeugen Sie am elegantesten wie folgt:

1. Fehlermatrix E der Dimension (10×100) erzeugen. Pro Spalte sind die Fehler eines Experiments enthalten.
   `t.E <- matrix(rnorm(10*1000), ncol=...)`
2. Daraus die simulierten Beobachtungen berechnen. (Beachten Sie die Eigenart von R, Objekte von "falscher" Dimension zyklisch zu verwenden!)
   `t.Y <- t.E + t.y`
3. Eine Resultatmatrix der Dimension 100×2 definieren.
   Entweder mit einer for-Schleife die 100 Experimente auswerten und die Koeffizienten pro Experiment in einer Zeile der Resultatmatrix speichern,

   ```
   r.coef <- matrix(nrow=100, ncol=2)
   for (i in 1:100) {
     r.coef[i,] <- lm(t.Y[,i] ~ t.x)$coefficients
   }
   ```

   oder das Ganze etwas eleganter mit `apply` lösen

   ```
   r.coef <- apply(t.Y, 2, FUN=function(y) lm(y~t.x)$coefficients)
   ```