

Regression I — Serie 1

1. Wir betrachten in dieser Aufgabe den Datensatz von Forbes. Er enthält die Werte von Siedepunkt (in Fahrenheit) und Luftdruck (inch of mercury), welche an 17 Orten in den Alpen und in Schottland gemessen wurden. Forbes wollte den Siedepunkt benutzen, um damit den Luftdruck zu schätzen. Mit dem Luftdruck wollte er dann wiederum die Höhe berechnen.

Sie finden die Daten im Dataframe `forbes` mit den Variablen `y` = Siedepunkt und `druck` = Luftdruck auf dem Internet. Mit den folgenden R-Befehle können Sie die Daten einlesen:

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

Unter den Teilaufgaben finden Sie nützliche Hinweise zum Lösen der Aufgabe.

- Fügen Sie die Variable `x = log10(druck)` zum Dataframe `forbes` hinzu und tragen Sie die `y`-Werte gegen die `x`-Werte auf. Kommentar!
- Berechnen Sie die Kleinste-Quadrate-Schätzungen für die Regression von `y` auf `x`. Betrachten Sie die Regressionsgerade im Scatterplot. Kommentar?
- Testen Sie $H_0 : \beta = 0$ gegen $H_A : \beta \neq 0$ im Modell $Y_i = \alpha + \beta x_i + E_i$ mit Hilfe des Regressions-Outputs auf dem 5%-Niveau.
- Berechnen Sie ein 95%-Konfidenzintervall für die Steigung β .
- Berechnen Sie die Kleinste-Quadrate-Schätzer für die Regression von `y` auf `x`, wenn die 12. Beobachtung weggelassen wird. Vergleichen Sie die Werte von $\hat{\alpha}$, $\hat{\beta}$, $se(\hat{\alpha})$, $se(\hat{\beta})$ und $\hat{\sigma}^2$ mit denjenigen aus Aufgabe b).

R-Hinweise:

- Sie können eine weitere Variable an das Dataframe `d.forbes` mit dem folgenden R-Befehl hinzufügen: `d.forbes[, "NeuerName"] <- ...`. Wichtig ist, dass die Länge des Vektors auf der rechten Seite von `<-` gleich wie die Anzahl Zeilen von `d.forbes` ist (`nrow(d.forbes)`).
- Die lineare Regression von `y` auf `x` führen Sie mit dem R-Befehl `lm` durch (lesen Sie dazu im R-Skript das Kapitel "Lineare Regression").
- Man kann direkt `abline(lm(...))` schreiben, um die Regressionsgerade einzuzeichnen.

Quelle: S. Weisberg, *Applied Linear Regression*, Wiley (1990), p. 3

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 $k = 1, \dots, 4$.

Einlesen der Daten:

```
t.url <- "http://stat.ethz.ch/Teaching/Datasets/NDK/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 $\hat{\sigma}^2$ 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.

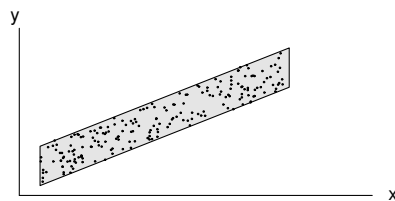
- c) Kommentieren Sie die Resultate aus a) und b).

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

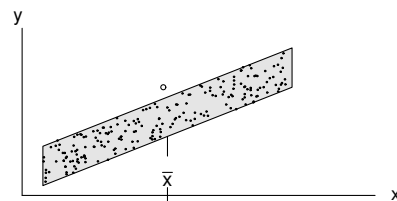
3. Betrachten Sie die folgenden Abbildungen. Geben Sie für die Figuren (b) – (f) an, welche Auswirkung das Hinzufügen des Punktes, resp. der Punkte, $\circ$ auf die Werte von $SS_X = \sum_i (x_i - \bar{x})^2$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$, $\widehat{\text{Var}}(\hat{\alpha})$ und $\widehat{\text{Var}}(\hat{\beta})$ hat.

Benützen Sie die Kategorien:

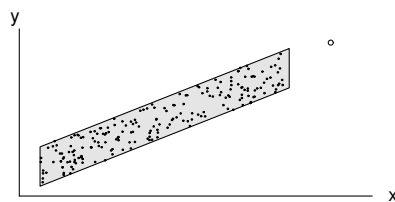
-- (für starke Verkleinerung), --, =, +, ++ (für starke Vergrößerung)



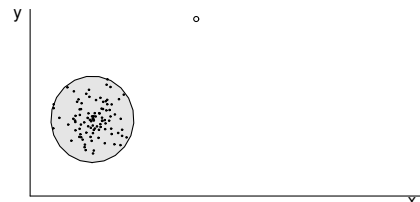
(a)



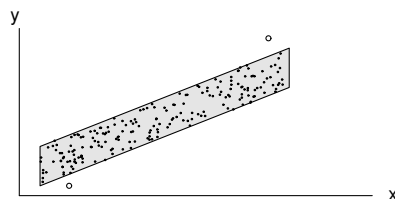
(b)



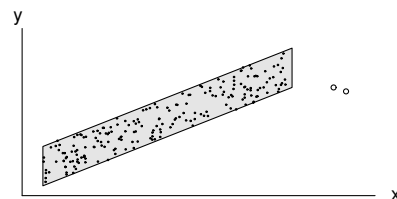
(c)



(d)



(e)



(f)