1 Lineare Modelle

Aufgabe 1:

Lesen Sie den Datensatz 1 (erhältlich von der Übungshomepage) ein. Dieser enthält für 32 Automobiltypen das Gewicht (1. Spalte, in amerikanischen Pfund) und den Benzinverbrauch (2. Spalte, in Meilen pro Gallone).

(a) Schätzen Sie unter Verwendung statistischer Software (z.B. mit R) den Einfluss des Gewichtes auf den Verbrauch und interpretieren Sie die Parameterschätzer.

(b) Der unten angegebene R-Output ist unvollständig. Berechnen Sie den Wert, der an Stelle von `A1` im Output stehen müsste.

```
##
## Call: lm(formula = MPG ~ Weight, data = auto)
##
## Coefficients:
##             Estimate    Std. Error  t value  Pr(>|t|)
## (Intercept) "37.29"     "1.878"     "A1"     "8.242e-19"
## Weight      "-0.005344" "0.0005591" "-9.559" "A2"
##
## Multiple R-squared: 0.7528,Adjusted R-squared: 0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

(c) Erläutern Sie, wie man den Eintrag der mit `A2` markierten Lücke erhält. Gehen Sie dabei insbesondere auf die benötigte Verteilung einschließlich eventueller Parameter ein.

(d) Wieso scheint es inhaltlich sinnvoller, den Verbrauch nicht in Meilen pro Gallone, sondern in Gallonen pro Meile anzugeben? Schätzen Sie das entsprechende Modell und überprüfen Sie, ob diese Transformation der Daten eine bessere Anpassung liefert. Verschaffen Sie sich auch graphisch einen Eindruck vom Zusammenhang zwischen den Variablen.

(e) Leiten Sie die Umrechnungsformel für die KQ-Schätzer nach einer linearen Variablentransformation

$$x_i \to t_i = a_0 + a_1 x_i, \quad \text{mit } a_1 \neq 0 \tag{1}$$
$$y_i \to u_i = b_0 + b_1 y_i, \quad \text{mit } b_1 \neq 0 \tag{2}$$

her.

(f) Wie ändern sich die Parameterschätzungen und p-Werte, wenn Sie das Gewicht in Kilogramm und den Verbrauch in Liter pro 100 km angeben (1 kg = 2.2046 am. Pfund, 1 km = 0.6214 Meilen, 1 Liter = 0.22 Gallonen)? Berechnen Sie hierzu die neuen Schätzungen an Hand der Ergebnisse aus Teilaufgabe b). Hätten Sie auch die Ergebnisse aus Teilaufgabe a) verwenden können?

Aufgabe 2*:

(a) Betrachten Sie wieder das Modell von Blatt 1, Aufgabe 1 d). Zentrieren Sie die Einflussvariable `Gewicht` und passen Sie das Modell mit dieser *zentrierten* Variable neu an. Interpretieren Sie die Parameter. *Hinweis*: Die zentrierte Variable ergibt sich aus der ursprünglichen durch $\tilde{x}_i = x_i - \bar{x}$.

(b) Berechnen Sie die Parameterschätzer nun direkt aus den Parameterschätzern des Modells in Aufgabe 1 d). Vergleichen Sie hierzu auch Blatt1, Aufgabe 1 e).

(c) Betrachten Sie den Output ihres Modells. Zu welchem Test gehört der am Ende des Outputs dargestellte p-Wert. Wie lauten H_0 und H_1? Berechnen Sie den Wert in **R**.
Hinweis: Nutzen Sie dabei die Funktion `?pf`.

(d) Betrachten Sie nun das Modell $\widetilde{GPM}_i = \gamma_0 + \gamma_1 \widetilde{Weight}_i + \varepsilon_i$, wobei $\widetilde{GPM}$ und $\widetilde{Weight}$ die jeweils *standardisierten* Versionen der entsprechenden Variablen sind. Vergleichen Sie den Koeffizientenschätzer $\hat{\gamma}_1$ mit dem Korrelationskoeffizienten zwischen den beiden standardisierten Variablen. Was fällt Ihnen auf?
Hinweis: Eine Variable $\tilde{x}$ ist standardisiert, wenn $\bar{\tilde{x}} = 0$ und $s_{\tilde{x}} = 1$, d.h. $\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{s_x^2}}$, mit $s_x^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$.

Aufgabe 3:

Betrachten Sie das einfache lineare Regressionsmodell (VL, Folie 14):

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i;\ i = 1, \ldots, n \tag{3}$$

mit den Annahmen

$$\begin{aligned} E(\epsilon_i) &= 0;\ i = 1, \ldots, n & (4)\\ V(\epsilon_i) &= \sigma^2;\ i = 1, \ldots, n & (5)\\ \{\epsilon_i \mid i = 1, \ldots, n\} &\quad \text{stoch. unabhängig} & (6)\\ \epsilon_i &\sim N(0, \sigma^2);\ i = 1, \ldots, n & (7) \end{aligned}$$

a) Zeigen Sie, dass der KQ-Schätzer für (β_0, β_1) auch ML-Schätzer ist.

b) Wie lautet der ML-Schätzer für σ^2?

c) Leiten Sie die Varianzen $\text{Var}(\hat{\beta}_0)$ und $\text{Var}(\hat{\beta}_1)$ her.

Aufgabe 4:

Die Methode der kleinsten Quadrate kann auch als rein deskriptives Verfahren zur Beschreibung des linearen Zusammenhangs zwischen zwei Merkmalen X und Y aufgefasst werden. Man kann dann die zu $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ gehörige Umkehrregression $X_i = \alpha_0 + \alpha_1 y_i + \delta_i$ betrachten und die Parameter beider Gleichungen mit Hilfe der KQ-Methode bestimmen.

(a) Zeigen Sie: Zeichnet man die Gerade der Umkehrregression in das übliche (x,y)-Koordinatensystem ein, so schneidet diese die gewöhnliche Regressionsgerade im Punkt $(\bar{x}, \bar{y})$, hat aber eine dem Betrag nach größere (oder gleiche) Steigung als die gewöhnliche Regressionsgerade.

(b) Zeigen Sie den Unterschied zwischen den beiden Methoden graphisch für das Beispiel $(x_1, y_1) = (2, 2), (x_2, y_2) = (4, 2), (x_3, y_3) = (6, 4), (x_4, y_4) = (8, 4)$.

(c) Zeigen Sie außerdem: Die beiden Geraden stehen genau dann aufeinander senkrecht, wenn der Pearson'sche Korrelationskoeffizient $r_{xy} = 0$ ist und sind deckungsgleich genau dann, wenn $r_{xy} \in \{-1, 1\}$. Somit eignet sich der Winkel zwischen den beiden Geraden auch als Maß für die Anpassungsgüte ($r_{xy}^2 = R^2$).