Regressionsanalyse – Übungen: Blatt 2

1. Die folgenden Daten bringen die Produktion von Biomasse von Sojabohnen mit der aufgefangenen, kumulierten Sonnenstrahlung in Beziehung.

Solarstrahlung (x)	Pflanzenbiomasse (y)
29.7	16.6
68.4	49.1
120.7	121.7
217.2	219.6
313.5	375.5
419.1	570.8
535.9	648.2
642.5	755.6

   (a) Berechne $\hat{\beta}_0$ und $\hat{\beta}_1$ des SLR für die Pflanzenbiomasse in Abhängigkeit von der Solarstrahlung. Schreibe explizit die Regressionsgleichung auf.

   (b) Berechne 95% Konfidenzintervalle für β_0 und β_1. Interpretiere die beiden Intervalle.

   (c) Teste H_0: $\beta_1 = 1$ gegen H_1: $\beta_1 \neq 1$ mittels eines geeigneten t-Tests mit $\alpha = 0.10$. Ist das Resultat des t-Tests konsistent mit dem Konfidenzintervall von zuvor?

   (d) Verwende einen t-Test um H_0: $\beta_0 = 0$ gegen H_1: $\beta_0 \neq 0$ zu testen. Interpretiere dieses Ergebnis.

   (e) Berechne den Schätzer der Varianz von $\hat{\beta}_1$ unter einem Modell ohne Intercept. Vergleiche die Varianzen der Schätzer für die Steigungen $\hat{\beta}_1$ unter beiden Modellen (mit/ohne β_0). Welches Modell liefert die bessere Präzision für den Steigungsschätzer?

   (f) Berechne ein 95% Konfidenzintervall für die mittlere Biomassenproduktion bei $x = 30$ und für $x = 600$ unter beiden Modellen (mit/ohne β_0). Diskutiere Ursachen für den Unterschied in diesen beiden Intervallen.

2. Der data frame `faithful` im externen `R` Paket MASS (`library(MASS); data(faithful); faithful`) besteht aus 272 Beobachtungen zweier Variablen, Ausbruch (`eruptions`) und Wartezeit (`waiting`). Der Wert von `eruptions` beschreibt die Dauer eines Ausbruchs des Old Faithful Geysirs in Minuten und der entsprechende Wert für `waiting` beschreibt die Dauer der Wartezeit bis zum nächsten Ausbruch. Angenommen ein Ausbruch mit Dauer 4 Minuten und 30 Sekunden wurde gerade beobachtet. Berechne ein 95% Prädiktionsintervall für die Wartezeit bis zum nächsten Ausbruch. Welche Annahmen werden hierbei benötigt, um dieses Intervall zu rechtfertigen? Welche Störungen dieser Annahmen sind bei unserer Datensituation wahrscheinlich? Decken Residuenplots irgend etwas auf?

3. Zeige explizit ohne Verwendung von Matrizen, dass für ein lineares Modell die folgende Identität hält:

$$\sum_{i=1}^{n} r_i \hat{\mu}_i = 0$$

   mit Residuen $r_i = y_i - \hat{\mu}_i$.