



Statistik II

Übungsblatt 1

Abgabe: Do. 30.10.2003, 10.00 Uhr, Briefkasten: Statistik II.

Bei jeder Aufgabe können maximal 5 Punkte erreicht werden.

1. "Tree"-Datensatz:

Dieser Datensatz gibt Messungen der Baumstärke ("Girth" = Durchmesser in 4 ft 6 in Höhe), der Höhe und des Volumens von 31 Schwarzkirschbäumen an. Die Daten sind in Zoll, Fuß und Kubikfuß angegeben.

- (a) Führen Sie eine lineare Regression der Form

$$y = \beta_0 + \beta_1 \text{Höhe} + \beta_2 \text{Durchmesser}$$

für das Volumen in Abhängigkeit von Höhe und Durchmesser durch. Wie interpretieren Sie die Parameter β_1 und β_2 ?

- (b) Führen Sie eine polynomiale Regression der Form

$$y = \beta_0 + \beta_1 \text{Höhe} + \beta_2 \text{Durchmesser} + \beta_3 \text{Durchmesser}^2$$

für das Volumen in Abhängigkeit von Höhe und Durchmesser durch. Wie interpretieren Sie jetzt die Parameter β_1 , β_2 und β_3 ?

- (c) Logarithmieren Sie die Variablen Volumen, Höhe und Durchmesser und führen Sie dann eine lineare Regression wie in (a) durch, also

$$\ln y = \beta_0 + \beta_1 \ln(\text{Höhe}) + \beta_2 \ln(\text{Durchmesser}).$$

Wie interpretieren Sie jetzt die Parameter β_1 und β_2 ?

- (d) Geben Sie für die Unterschiede der Modelle in (a), (b) und (c) eine plausible Erklärung.

2. "Tree"-Datensatz:

Führen Sie für die Daten aus Aufgabe 1 eine Box-Cox-Transformation durch. Welche Transformation liefert die "beste" Modellanpassung? Vergleichen Sie dieses Ergebnis mit der besten Anpassung aus Aufgabe 1.

3. "Cars-Datensatz": Sie analysieren einen Datensatz mit technischen Daten von Autos. Hier interessiert der Zusammenhang zwischen dem Verbrauch (in Miles per Gallon, MPG), Leistung (Horsepower) und Hubraum (Engine Displacement).

- (a) Schreiben Sie das zugrundeliegende mathematische Modell auf ("Regressionsgleichung").
- (b) Kommentieren Sie die Güte der Regression. Erläutern Sie dabei genau, wie Sie zu Ihrer Einschätzung kommen.
- (c) Erstellen Sie einen Residuenplot und betrachten Sie diesen genauer. Möglicherweise könnte eine nochmalige Regression, diesmal auf den Residuen, die letzten linearen Strukturen entfernen. Würden Sie das empfehlen? Begründen Sie Ihre Antwort.
- (d) Hängt die Erwartungstreue der Parameterschätzer von der häufig getroffenen Normalverteilungsannahme ab, und wenn ja, wie?
- (e) Der Parameter für "Engine Displacement" ist statistisch signifikant, aber fast Null. Erläutern Sie den Unterschied.

4. "Cars"-Datensatz:

Analysieren Sie den cars--Datensatz mit Hilfe von multipler linearer Regression weiter und beantworten Sie dabei die folgenden Fragen:

- (a) Was halten Sie persönlich für das "beste" Modell? Begründen Sie Ihre Auswahl.
- (b) Haben Transformationen (z.B. MPG \rightarrow 1/MPG) eher positive oder negative Auswirkungen auf Ihre Modellierung? Was würden Sie empfehlen?
- (c) Treten Probleme mit dem Datensatz oder der verwendeten Methode (hier: multiple Regression) auf?

5. In der Literatur werden verschiedene Definitionen für die extern studentisierten Residuen gegeben. Zeigen Sie, daß folgende Definitionen äquivalent sind.

(a) $e_{i(i)}^* = \frac{e_i}{s_{(i)}\sqrt{1-h_i}}$

(b) $e_{i(i)}^* = \frac{e_{i(i)}}{s_{(i)}\sqrt{1+h_{i(i)}}}$.

Hinweis: Sie dürfen dabei, falls erforderlich, folgende Gleichung **ohne Beweis** verwenden:

$$(A - bb')^{-1} = A^{-1} + \frac{1}{1 - b'A^{-1}b} A^{-1}bb'A^{-1},$$

wobei A eine Matrix und b ein Spaltenvektor mit passenden Dimensionen sind.