h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Antje Jahn
Multivariate Statistik
Sommersemester 2019

Arbeitsblatt 4

A 1 In einer Kneipe wurde bei 7 Gästen die Höhe des Trinkgelds prozentual im Verhältnis zur Gesamtrechnung erfasst. Die nachfolgende Tabelle beschreibt die Ergebnisse und die Information, zu welcher Uhrzeit der Kneipenbesuch stattfand und ob eine männliche oder weibliche Bedienung kassiert hatte.

Trinkgeld_in_Prozent	5.1	15.6	28.2	11.1	4.0	31.5	19.5
Uhrzeit	18.00	19.50	20.00	20.50	21.25	21.50	22.00
Geschlecht	m	w	w	m	m	w	w

Mit einem linearen Regressionsmodell soll untersucht werden, ob Uhrzeit und/oder Geschlecht die Trinkgeldhöhe beeinflussen. Dabei ist (bei einer Kodierung von männlich=0 und weiblich=1) $\hat{\sigma} = 6.8$, $R^2 = 0.73$ und

$$(X^T X)^{-1} = \begin{pmatrix} 39.836 & -1.983 & 1.320 \\ -1.983 & 0.100 & -0.083 \\ 1.320 & -0.083 & 0.652 \end{pmatrix}.$$

a) Formulieren Sie das Regressionsmodell und erstellen Sie die Designmatrix X
b) Schätzen Sie die Regressionskoeffizienten und interpretieren Sie diese.
c) Um wieviel ändert sich das zu erwartende Trinkgeld pro Stunde bei Männern? Um wieviel bei Frauen?
d) Führen Sie einen statistischen Test zum Signifikanzniveau $\alpha = 0.05$ durch, um zu überprüfen, ob die Höhe des Trinkgelds vom Geschlecht der Bedienung abhängt. Formulieren Sie die Testhypothesen über die Modellparamter.
e) Es wurde eine zweite lineare Regression durchgeführt, bei der die abhängige Variable nicht das prozentuale, sondern das absolute Trinkgeld war. Die Resttreuung, SS_E, war in diesem Modell deutlich höher. Können Sie hieraus folgern, dass Uhrzeit und Geschlecht die absolute Trinkgeldhöhe schlechter erklären als die prozentuale? Begründen Sie Ihre Entscheidung.

A 2 Laden Sie das Paket ISLR, hierin ist der Datensatz Auto enthalten mit den Variablen:

- mpg: Meilen per Gallon
- cylinders: Zylinder
- displacement: Hubraum
- horsepower: PS
- weight: Gewicht
- acceleration: Beschleunigung
- year: Modelljahr
- origin: Ursprungsland (1=Amerika, 2=Europa, 3=Japan)
- name: Fahrzeugname

Es soll der Zusammenhang zwischen (logarithmierten) Meilen per Gallon und den übrigen Variablen mit Regressionsmethoden untersucht werden.

a) Untersuchen Sie zunächst die Verteilung der Variable mpg und log(mpg) mittels Histogramm und QQ-Plot. Welche Variable scheint eher einer Normalverteilung zu genügen? Arbeiten Sie mit log(mpg) als abhängiger Variable weiter.

b) Führen Sie eine univariate lineare Regression mit log(mpg) als Zielgröße und PS als Prädiktor durch.
   - Gibt es einen Zusammenhang zwischen Prädiktor und Zielgröße?
   - Ist dieser positiv oder negativ?
   - Ist der Zusammenhang statistisch signifikant zum 5%-Signifikanzniveau?
   - Welchen mpg-Wert würde man anhand dieses Ergebnisses bei einem Auto mit 98 PS erwarten?
   - Um wieviel unterscheidet sich erwartungsgemäß der logarithmierte Verbrauch log(mpg) zwischen zwei Autos, die sich um 20 PS unterscheiden?
   - Extrahieren Sie den Vektor der Residuen und überprüfen Sie diesen auf Normalverteilung, indem Sie einen QQ-Plot und ein Histogramm erstellen sowie einen Shapiro-Wilk-Test auf Normalverteilung durchführen
   - Erstellen Sie ein Streudiagramm für Zielgröße und Prädiktor. Zeichnen Sie die Regressionsgerade in den Plot.

c) Führen Sie nun jeweils eine multiple lineare Regression durch mit log(mpg) als Zielgröße und A) PS, Jahr und Ursprungsland als Prädiktoren und B) allen Variablen (ausser Fahrzeugname) als Prädiktoren. Behandeln Sie dabei die Variable Ursprungsland als nominale Variable (as.factor)

- Welche Einflußgrößen zeigen jeweils einen statistisch signifikanten Zusammenhang mit der Zielgröße?
- Wie ist in diesen Modellen der Regressionskoeffizient zur Variable PS zu interpretieren? Wie unterscheiden sich nach diesen Ergebnissen zwei Autos, die sich um 20 PS unterscheiden bei sonst gleichen Einflussvariablen?
- Wie unterscheidet sich der Standardfehler des Koeffizientenschätzers zur Variablen PS zwischen den drei Modellen (univariat, multivariat A, multivariat B)? Finden Sie eine mögliche Erklärung für die Unterschiede?
- Inwiefern beeinflusst der Standardfehler die Teststatistik und den p-Wert?