Multiple lineare Regression

Aufgabe 1:

Der Datensatz `teengamb` im **R**-package `faraway` enthält Informationen zum Glücksspielverhalten von Teenagern in Großbritanien:

Variable	Beschreibung
`gamble`	Ausgaben für Glücksspiel (in Pounds/Jahr)
`status`	Sozioökonomischer Status (basierend auf dem Beruf der Eltern)
`income`	Einkommen (in Pounds/Woche)
`verbal`	Test-Score hinsichtlich der Anzahl richtig definierter Wörter (max. 12)
`sex`	Geschlecht (0 = männlich, 1 = weiblich)

(a) Passen Sie ein multiples lineares Modell (mit Absolutglied β_0) an die Daten an, welches beschreibt, wie sich der Status (`status`), das Einkommen (`income`) und der Test-Score (`verbal`) auf das Glücksspielverhalten (`gamble`) von Jügendlichen auswirken. Geben Sie dabei zunächst die Modellspezifikation an, nutzen Sie anschließend die Funktion `lm`, um das Modell anzupassen.

(b) Interpretieren Sie alle Parameterschätzungen ausführlich.

Aufgabe 2:

Nehmen Sie an Sie arbeiten in der statistischen Beratung und sollen eine Studie auswerten, bei der mit Hilfe eines psychologischen Tools das allgemeine Engagement für den Studiengang (`dedication`) sowie die selbst empfundene Lebenszufriedenheit (`happiness`) von 93 Studenten gemessen worden ist. Außerdem wurde ermittelt wie viel Prozent der Übungsaufgaben sie während des Semesters bearbeitet haben (`diligence`) und schließlich, wie viele Punkte sie im Abschlusstest erreicht haben (`final`).

(a) Ein Beispieldatensatz findet sich auf der Veranstaltungshomepage (`CAstudy`). Lesen Sie diesen in **R** ein und verschaffen Sie sich einen ersten Überblick über die Zusammenhänge.

(b) Modellieren Sie die erreichten Punkte (`final`) in Abhängigkeit der bearbeiteten Übungen (`diligence`). Geben Sie hierfür zunächst die Modellgleichung an. Stellen Sie die Ergebnisse graphisch dar, interpretieren Sie die geschätzten Parameter.

(c) Sie möchten nun auch noch die verbliebenen Variablen `dedication` und `happiness` in das Modell aufnehmen. Geben Sie zunächst die Modellgleichung mit Index-Notation und in Matrixnotation an, skizzieren Sie die entsprechenden Matrizen und geben Sie deren Dimensionen an.

(d) Passen Sie das Modell in R an. Interpretieren Sie die Parameter der neu hinzugekommenen Variablen. Wie erklären Sie sich die Änderung im Vergleich zu Aufgabe 2 b)?

Aufgabe 3:

Unterstellen Sie die Gültigkeit des linearen Modells

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta_p} + \boldsymbol{\epsilon} \tag{1}$$

wobei $\mathbf{X}$ (neben einer Konstanten) p unabhängige Variablen enthält. Irrtümlicherweise schätzen Sie aber das falsche Modell

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta_{p-1}} + \boldsymbol{\epsilon} \tag{2}$$

wobei man die Regressormatrix $\tilde{\mathbf{X}}$ in (2) aus $\mathbf{X}$ in (1) durch Weglassen der $(p+1)$-ten Spalte erhält, d.h. es gilt $\mathbf{X} = [\tilde{\mathbf{X}}, \mathbf{x_p}]$, $\mathbf{x_p} = [x_{1p}, \ldots, x_{np}]'$ und $\boldsymbol{\beta_p} = [\boldsymbol{\beta'_{p-1}}, \beta_p]'$. Mithin „vergessen“ Sie also die p-te Einflussgröße in Ihrer Schätzung. Zeigen Sie, dass der aufgrund von (2) ermittelte KQ-Schätzer des Parametervektors $\boldsymbol{\beta_{p-1}}$ in der Regel nicht erwartungstreu ist. Ermitteln und interpretieren Sie den Bias, d.h. die Verzerrung $E(\hat{\boldsymbol{\beta}}_{\boldsymbol{p-1}}) - \boldsymbol{\beta_{p-1}}$. Wann ist der Bias gleich Null?