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

Antje Jahn
Nichtparametrische und nichtlineare Modelle
Sommersemester 2019

Arbeitsblatt 1

A 1 Nehmen Sie an, dass für die Zufallsvariablen (X, Y) ein lineares Modell

$$Y = bX + \epsilon \quad \text{mit } b = 2, Var(\epsilon) = \sigma^2 = 4, \epsilon \text{ unabhängig von X}$$

gilt. Ein aus einer Stichprobe entwickeltes Prognosemodell liegt etwas daneben und schätzt $E(Y|x) = \hat{b}x = 2.2 \cdot x$.

a) Berechnen Sie den erwarteten quadratischen Prognosefehler $E\Big((Y - \hat{b}x)^2\Big)$ dieses Prognosemodells im Punkt $x = 1$. Betrachten Sie dabei Y als Zufallsvariable und x und $\hat{b}$ als feste Werte (Konstanten).

b) Vergleichen Sie den Prognosefehler mit dem Fehler eines optimalen Prognosemodells $E(Y|x) = \hat{b}x = 2 \cdot x$. Begründen Sie, warum der Unterschied hier verhältnismäßig klein ausfällt.

Hinweis zu a): Addieren und subtrahieren Sie im quadratischen Term (links) den Erwartungswert E(Y), wenden Sie danach eine binomische Formel an und nutzen Sie aus, dass der verbleibende nicht-quadratische Term 0 ergibt.

A 2 Überprüfen Sie in einer Simulationsstudie wie Bias und Varianz von Regressionskoeffizientenschätzern von der Variablenauswahl und damit der Modellkomplexität abhängen können:

i) Simulieren Sie für $n = 100$ unabhängige Beobachtungen die Realisierungen von zwei jeweils standardnormalverteilten korrelierten Variablen (X_1, X_2) mit $Cor(X_1, X_2) = 0.8$ [mvtnorm]

ii) Simulieren Sie $n = 100$ unabhängige Realisierungen von jeweils 48 standardnormalverteilten und unkorrelierten Variablen $(X_3, \ldots X_{50})$

iii) Simulieren Sie $n = 100$ unabhängige standardnormalverteilte Fehlerterme ϵ

iv) Generieren Sie eine Outcome-Variable, die entsprechend eines linearen Regressionsmodell von X_1 und X_2 abhängt:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \epsilon$$

Legen Sie dabei die Parameter b_0, b_1 und b_2 selbst fest. Diese "wahren"
Werte sollen später mit den Schätzwerten aus den verschiedenen linearen Regressionsmodellen verglichen werden.

v) Schätzen Sie nun für die von Ihnen simulierten Daten, den Parameter b_1 aus verschiedenen linearen Regressionsmodellen:

- Modell A: lmA <- lm($Y \sim X_1$)
- Modell B: lmB <- lm($Y \sim X_1 + X_2$)
- Modell C3: lmC3 <-lm($Y \sim X_1 + X_2 + X_3$)
- Modell C4: lmC4 <-lm($Y \sim X_1 + X_2 + X_3 + X_4$)
- Modell C5-C50 entsprechend

vi) Speichern Sie die geschätzten Regressionskoeffizienten $\hat{b_1}$ aus den 50 verschiedenen statistischen Modellen ab.

Führen Sie Schritt i) bis vi) 1.000-mal durch. Speichern Sie dabei die insgesamt 50.000 geschätzten Regressionskoeffizienten (1.000 pro statistischem Modell) ab.

a) Bestimmen Sie für jedes der 50 Modelle den Durchschnitt der Schätzwerte $\hat{b_1}$ (mean) und die Differenz zwischen Durchschnittswert und wahrem Wert b_1 (Bias).
b) Bestimmen Sie für jedes Modell die empirische Varianz der Schätzwerte $\hat{b_1}$ (var).
c) Plotten Sie Bias und Varianz von $\hat{b_1}$ gegen die Modellkomplexität (d.h. gegen die Anzahl an Variablen im Modell).
d) Erklären Sie Ihr Ergebnis.