12. Übung am 24. Juni 2024

UV Angewandte Statistik (405.170)
Link Ankreuzliste: siehe www.trutschnig.net/courses
Mit 'F' versehene Aufgaben sind freiwillig, mit * versehene Aufgaben haben einen erhöhten Schwierigkeitsgrad.
Alle Verweise beziehen sich auf das Statistik oder das Angewandte Statistik Skriptum.

Übungsaufgabe 58. Zeigen Sie die in Bemerkung 6.5 behauptete Darstellung von R^2 im Falle univariater linearer Regression.

Übungsaufgabe 59 (Fortsetzung von Aufgabe 35)**.** Generieren Sie Stichproben wie in Aufgabe 35 beschrieben und berechnen Sie dann die Schätzer $\hat{\theta}_0$ und $\hat{\theta}_1$ mit Hilfe der R-Funktion *lm*. Wiederholen Sie den Vorgang $R = 1000$ und plotten Sie die erhaltenen Werte $(\hat{\theta}_0^i, \hat{\theta}_1^i), i = 1 \dots R$ - was ist zu beobachten? Was ändert sich, wenn (i) die Stichprobengröße n erhöht wird, und wenn (ii) die Varianz von ε verringert wird?

Übungsaufgabe 60. Laden Sie den Lebenserwartungs-Datensatz der Weltbank herunter (`https://data.worldbank.org/indicator/SP.DYN.LE00.IN`) und lesen Sie ihn in R ein.

1. Filtern Sie auf die USA und bringen Sie den Datensatz in die folgende Form:

	country	year	life_expectancy
1	United States	1960	69.77
2	United States	1961	70.27
3	United States	1962	70.12
4	United States	1963	69.92
5	United States	1964	70.17
6	United States	1965	70.21

2. Plotten Sie die Zeitreihe in ggplot2.
3. Passen Sie mit Hilfe von `lm` ein lineares Modell an und prognostizieren damit Sie die Lebenserwartung in den USA im Jahr 2025 und im Jahr 2035. Verwenden Sie dafür die Funktion `predict`.
4. Wiederholen Sie die ersten drei Schritte für Österreich und für Japan. Was ist zu erkennen?

Übungsaufgabe 61. In der Praxis ist oft nicht einmal die parametrischer Form der Regressionsfunktion gegeben, nichtsdestotrotz gibt es nichtparametrische Methoden, die dann (zumindest im niedrigdimensionalen Setting) eingesetzt werden können. Für $(x_1, y_1), \dots, (x_n, y_n)$ kann beispielsweise die sog. Kernregression (Nadaraya Watson Regression) berechnet werden, i.e. als Schätzer für die Regressionsfunktion $r^*(x)$ an der Stelle x betrachtet man

$$\hat{r}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}, \tag{7.1}$$

wobei als K beispielsweise die Dichte von $\mathcal{N}(0,1)$ verwendet werden kann. Illustrieren Sie die Performanz des obigen Schätzers durch eine einfache Simulationsstudie, in dem Sie wiefolgt vorgehen:

- Wählen Sie eine konkrete stetige[vii] Regressionfunktion $r^* : [0, 10] \to \mathbb{R}$, betrachten Sie $(\varepsilon_i)_{i=1}^n$ i.i.d. mit $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$ für ein von Ihnen gewähltes $\sigma^2 > 0$ sowie eine Stichprobe $x_1, \ldots, x_n$ von $\mathcal{U}(0, 10)$ und setzen Sie $y_i := r^*(x_i) + \varepsilon_i$.
- Berechnen Sie dann $\hat{r}$ gemäß Gleichung (7.1) auf einem äquidistanten Gitter in $[0, 10]$ und vergleichen Sie r^* und $\hat{r}$ für kleines und für großes n. Was ist zu beobachten?

Übungsaufgabe 62. Lesen Sie Abschnitt 6.2 im Skriptum. Wählen Sie dann $\theta_1^*, \theta_0^* \in \mathbb{R}$ und generieren Sie Daten $(x_1, y_1), \ldots, (x_n, y_n)$ des logistischen Modells $\mathbb{P}(Y = 1|X = x) = l_{\theta^*}(x)$. Verwenden Sie *glm* um $l_{\hat{\theta}}$ zu berechnen. Wiederholen Sie den obigen Vorgang $R = 1000$ mal und plotten Sie die erhaltenen R Paare $(\hat{\theta}_1^j, \hat{\theta}_0^j)$, $j \in 1, \ldots, R$, gemeinsam mit (θ_1^*, θ_0^*). Vergleichen Sie die erhaltenen Scatterplots für $n = 20$, $n = 200$ und $n = 2000$ - was ist zu sehen?

Übungsaufgabe 63. Importieren Sie den sog. Titanic Datensatz (abrufbar unter `https://github.com/datasciencedojo/datasets/blob/master/titanic.csv`) in R. Finden Sie mittels (univariater) logistischer Regression heraus, ob das Feature 'age' einen Einfluss auf die Überlebenswahrscheinlichkeit hat. Wie sieht es mit der Variable 'fare' aus?

[vii] aber nicht notwendigerweise lineare oder polynomiale