10. Übung am 26. Juni 2023

Übungsaufgabe 52. Zeigen Sie die in Bemerkung 6.5 behauptete Darstellung von R^2 im Falle univariater linearer Regression.

Übungsaufgabe 53 (Fortsetzung von Aufgabe 49)**.** Generieren Sie Stichproben wie in Aufgabe 49 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 54. Wenn Sie (wie in der vorigen Aufgabe) die R-Funktion *lm* anwenden und dann via *summary* die Zusammenfassung des geschätzten linearen Modells printen, dann wird neben den Schätzern $\hat{\theta}_0, \hat{\theta}_1$ auch noch weitere Information ausgegeben. Finden Sie heraus (kleine Literaturrecherche), was die angegebenen Werte bedeuten. Recherchieren Sie weiters, was unter dem Begriff Multikollinearität zu verstehen ist und veranschaulichen Sie das Problem anhand einer linearen Regression mit 2 erklärenden (und stark korrelierten) Variablen.

Übungsaufgabe 55. 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[x] 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 $x_1, \dots, x_n$ von $\mathcal{U}(0, 10)$ und setzen Sie $y_i := r^*(x_i) + \varepsilon_i$.
- Berechnen Sie dann $\hat{r}$ gemäß Gleichung (7.2) 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 56. Wählen Sie $\theta_1^*, \theta_0^* \in \mathbb{R}$ und generieren Sie Daten $(x_1, y_1), \dots, (x_n, y_n)$ des logistischen Modells $\mathbb{P}(Y = 1|X = x) = l_{\theta^*}(x)$. Verwenden Sie dann *glm* um $l_{\hat{\theta}}$ zu berechnen. Wiederholen Sie den obigen Vorgang $R = 1000$ mal und plotten Sie die erhaltenen R Paare $(\hat{\theta}_1^j, \hat{\theta}_0^j)$, $j \in 1, \dots, R$, gemeinsam mit (θ_1^*, θ_0^*). Vergleichen Sie die erhaltenen Scatterplots für $n = 20$, $n = 200$ und $n = 2000$ - was ist zu sehen?

Übungsaufgabe 57. Importieren Sie den sog. Titanic Datensatz (abrufbar hier) 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?

[x] aber nicht notwendigerweise lineare oder polynomiale