12. Übung am 27. Juni 2022

UV Angewandte Statistik (405.170)

Übungsaufgabe 67. Wählen Sie $\theta_1^*, \theta_0^* \in \mathbb{R}$ 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 dann *glm* um $l_{\hat{\theta}}$ zu berechnen. Wiederholen Sie den obigen Vorgang $R = 1000$ mal und plotten Sie die erhaltenen R Paare $(\hat{\theta}_1^j, \hat{\theta}_0^j)$, $j \in 1, \ldots, R$, gemeinsam mit (θ_1^*, θ_0^*). Vergleichen Sie die erhaltenen Scatterplots für $n = 20$, $n = 200$ und $n = 2000$ - was ist zu sehen

Übungsaufgabe 68. 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?

Übungsaufgabe 69. Es seien X_1, X_2, X_3, X_4 unabhängige Zufallsvariablen mit einer festen, von Ihnen gewählten Verteilung; weiters erfülle der Fehler $\varepsilon \sim \mathcal{N}(0, 0.5)$ und es gelte $Y = X_1 + X_2^2 - X_3 + \varepsilon$[viii]. Verwenden Sie das R-package *qmd* in seiner letzten, hier abrufbaren Version 1.1.0 und die darin enthaltene Funktion *feature_selection*, um qmd die relevanten Features für Y herausfinden zu lassen. Ändern Sie dann das Modell um zu überprüfen, ob qmd clever genug ist, die relevanten Features auch in diesem Fall zu erkennen.
Hinweis: Die obere, von qmd produzierte Grafik zeigt die Abhängigkeitswerte (schwarze Linie, je höher desto besser), die untere die gemäß Wichtigkeit selektierten Variablen.

Definition 7.2. Wir betrachten wie bisher $\emptyset \neq \Theta_0 \subset \Theta$ und eine Stichprobe $\mathbf{X} = (X_1, \ldots, X_n)$ von $X \sim (P_\theta)_{\theta \in \Theta}$. Mit $\mathcal{X} := Rg(X)^n$ bezeichnen wir den Stichprobenraum. Weiters sei $\mathcal{G}$ eine endliche Gruppe von Transformationen $g : \mathcal{X} \to \mathcal{X}$. Wir sagen, dass *unter H_0 die Randomisierungseigenschaft* gilt, genau dann, wenn für jedes $X \sim P_\theta$ mit $\theta \in \Theta_0$, jede Stichprobe $\mathbf{X} = (X_1, \ldots, X_n)$ von X, und jedes $g \in \mathcal{G}$ gilt: $\mathbf{X}$ und $g \circ \mathbf{X}$ haben die selbe Verteilung.

Übungsaufgabe 70. Sei $(X_1, \ldots, X_m)$ eine Stichprobe von $X \sim \mathcal{N}(\mu_1, \sigma^2)$, $(Y_1, \ldots, Y_n)$ eine Stichprobe von $Y \sim \mathcal{N}(\mu_2, \sigma^2)$, und gelte $H_0 : \mu_1 = \mu_2$. Wir setzen $\mathcal{X} = \mathbb{R}^{m+n}$ und $\mathbf{X} = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$. Für jede Permutation π von $\{1, \ldots, m+n\}$ sei $g_\pi : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ definiert durch

$$g_\pi(z_1, \ldots, z_{m+n}) = (z_{\pi(1)}, \ldots, z_{\pi(m+n)}),$$

$\mathcal{G}$ bezeichne die Gruppe all dieser Transformationen. Zeigen Sie, dass in diesem Setting die Randomisierungseigenschaft gilt.

Übungsaufgabe 71. Wir betrachten das Setting von Definition 7.2 und schreiben $M := \#G$ für die Kardinalität von G. Zusätzlich sei T eine beliebige Teststatistik, für jede Stichprobe $\mathbf{x} = (x_1, \ldots, x_n)$ bezeichne $T_{(1)}(\mathbf{x}) \leq T_{(2)}(\mathbf{x}) \leq \cdots \leq T_{(n)}(\mathbf{x})$ die Ordnungsstatistik der Werte $\{T \circ g(\mathbf{x}) : g \in \mathcal{G}\}$. Für jedes $\alpha \in (0, 1)$ setzen wir $k := M - \lfloor M\alpha \rfloor$ sowie

$$\begin{aligned} M^0(\mathbf{x}) &= \#\Big\{j \in \{1, \ldots, M\} : T_{(j)}(\mathbf{x}) = T_{(k)}(\mathbf{x})\Big\} \\ M^+(\mathbf{x}) &= \#\Big\{j \in \{1, \ldots, M\} : T_{(j)}(\mathbf{x}) > T_{(k)}(\mathbf{x})\Big\}, \end{aligned}$$

[viii] kein Tippfehler, die vierte Variable kommt nicht vor

und definieren einen 'drei'-wertigen Test φ durch

$$\varphi(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) > T_{(k)}(\mathbf{x}), \\ \frac{M\alpha - M^+(\mathbf{x})}{M^0(\mathbf{x})} & \text{if } T(\mathbf{x}) = T_{(k)}(\mathbf{x}), \\ 0 & \text{if } T(\mathbf{x}) < T_{(k)}(\mathbf{x}). \end{cases}$$

Beweisen Sie, dass dann für jedes $\theta \in \Theta_0$ die folgende Gleichheit gilt:

$$\mathbb{E}_\theta(\varphi \circ \mathbf{X}) = \alpha.$$