Einführung in die Statistik

HAUSÜBUNG 4 - Lineare Regression

Alexandra Posekany

WS 2023

Für alle Beispiele gelten folgende Aufgabenstellungen:

- Überprüfen Sie alle erforderlichen statistischen Voraussetzungen für die Gültigkeit dieses Modells mtihilfe der quality plots der Residuen und gegebenenfalls Scatterplots.
- Führen Sie eine Modellselektion durch und wählen anhand statistischer Kriterien ein optimales Modell aus. Argumentieren Sie anhand Kriterien für die Signifkanz von Koeffizienten und gegebenenfalls zusätzlich von Modellen.
- Schreiben Sie das Regressionsmodell und die angepasste Modellgleichung des optimalen Modells explizit an.
- Interpretieren Sie die Werte die Koeffizienten im Sachzusammenhang.

Datentransformation

Wählen Sie den Datensatz UN aus der library car. Filtern Sie erst 'NA' mit der Funktion na.omit. Erklären Sie dann infant mortality durch gross domestic product. Explorieren Sie die Daten, bevor Sie ein Modell anpassen.

Schweiz

Wir kehren zurück zu den Variablen "Fertility", "Agriculture", "Education", "Catholic" und "Infant. Mortality" aus dem R Datensatz swiss des R package utils. Passen Sie für die oben genannten Variablen ein Modell an, das Education durch die übrigen Variablen erklärt, soweit dies zulässig ist.

USA

Wir kehren zurück zu den Variablen "Population", "Income", "Illiteracy", "Life.Exp", "Murder", "HS Grade" und "Frost" aus dem R Datensatz state.x77. Passen Sie für die oben genannten Variablen ein lineares Modell (lm) an, das "Murder" durch die übrigen Variablen erklärt, soweit dies zulässig ist.

Lake Huron

Wir kehren zurück zum Datensatz "LakeHuron". Passen Sie ein Modell an, das den Zeittrend modelliert. Überprüfen Sie alle erforderlichen statistischen Voraussetzungen für die Gültigkeit dieses Modells mtihilfe der quality plots der Residuen.

Pima Indians

Laden Sie den Datensatz 'Pima.tr' aus der library 'MASS'. Ermittle ein logistisches Regressionsmodell, dass das Auftreten von Diabetes ('type') durch die übrigen unabhängigen Variablen Alter (age), Anzahl der Schwangerschaften (npreg), BMI, Glukosespiegel (glu), Blutdruck (bp), familiäre Häufung von Diabetesfällen (ped) und Hautfaltendickemessung am Oberarm (skin) erklärt. Schreibe die Modellgleichung an und interpretiere die Werte der Koeffizienten im Kontext.

Ermitteln Sie die prädiktive Qualität des Modells mithilfe einer Receiver Operating Characteristic (ROC) Kurve. Führen Sie auch die False Positive, False Negative, True Positive und True Negative Raten in einer Tabelle (Konfusionsmatrix) an.