Aufgaben zur Hausarbeit 1

Aufgabe 1

Datensatz

Laden Sie den Datensatz "kidney.csv" herrunter. Die Variablenbeschreibung finden Sie unter **UCI Machine Learning Repository**. Lesen Sie die Daten in R ein. Der Datensatz enthält 11 numerische Variablen und 14 nominalskalierte Variablen. Zuerst entfernen Sie aus dem Datensatz alle Stichproben (also Zeilen) mit den fehlenden Werten. Wandeln Sie die Variable `classification` in eine binäre Variable um, so dass Ausprägungen *ckd* den Wert 1 bekommen und Ausprägungen *notckd* den Wert 0.

Aufgaben

1. Schätzen Sie ein logistisches Regressionsmodell mit der Zielvariable `classification` und den Kovariablen `bp`, `sg` und `pot` (3., 4. und 15. Spalte im Datensatz). Sind alle Kovariablen in diesem Modell signifikant? Was können Sie über die Güte der Anpassung dieses Modells sagen? (2 Punkte)

2. Testen Sie mit einem geeigneten Verfahren, ob das kleinere Modell ohne Variable `pot` genau so gut wie das Modell aus Punkt 1 ist. Welches Modell würden Sie bevorzugen? (2 Punkte)

3. Gruppieren Sie die Daten des bevorzugten Modells aus Punkt 2 und schätzen Sie das entsprechende Modell. Was kann man über die Anpassung dieses Modells sagen? Führen Sie auch die Residualanalyse durch und kommentieren Sie das Ergebnis. (4 Punkte)

Aufgabe 2

Datensatz

Laden Sie den Datensatz "KenyaDHS.txt" herrunter.

Aufgaben

1. Schätzen Sie ein verallgemeinertes lineares Modell mit `numberlivingchild` als Zielvariable und `assetindex`, `BMI`, `agefirstbirth`, `breastfeeding` und `yearsofedu` als Kovariablen. Nehmen Sie an, dass die Zielvariable Poisson-verteilt ist und benutzen Sie eine kanonische Linkfunktion. Beurteilen Sie die Anpassung des Modells. Sind alle Kovariablen signifikant? Interpretieren Sie die geschätzten Koeffizienten. Gibt es Anzeichen einer Überdispersion in den Daten? Führen Sie eine Residualanalyse durch. Ist die Annahme einer Poisson Verteilung der Zielvariable gerechtfertigt? (6 Punkte)

2. Schätzen Sie das Modell aus Punkt 1 unter der Annahme, dass die Zielvariable einer negativen Binomialverteilung folgt. Beurteilen Sie die Anpassung des Modells und vergleichen Sie dieses Modell mit einem geeigneten Kriterium mit dem Modell aus Punkt 1. Führen Sie die Residualanalyse durch und interpretieren Sie das Ergebnis. (3 Punkte)

Aufgabe 3

Datensatz

Laden Sie den Datensatz "RealEstate.txt" herrunter. Die Variablenbeschreibung finden Sie unter **UCI Machine Learning Repository**. Belassen Sie im Datensatz nur die Variablen `HouseAge, Distance, NumberStores, Latitude, Price`.

Aufgaben

1. Schätzen Sie das einfache lineare Modell mit der Zielvariable `Preis` und den Kovariablen `HouseAge, Distance, NumberStores`. Sind alle Kovariablen im Modell signifikant? Interpretieren Sie die geschätzte Koeffizienten. Beurteilen Sie die Anpassung des Modells, führen Sie die Residualanalyse durch und interpretieren Sie das Ergebnis. (4 Punkte)

2. Ergänzen Sie das Modell aus Punkt 1 mit quadrierten Werten von `HouseAge` und `Distance`. Sind diese signifikant? Warum ist es sinnvoll für diese Variablen auch die quadrierten Werte aufzunehmen? Beurteilen Sie die Anpassung dieses Modells und führen Sie die Residualanalyse durch. Vergleichen Sie die Ergebnisse mit denen aus Punkt 1. (3 Punkte)

3. Da `Preis` eine positive Variable ist, schätzen Sie ein verallgemeinertes lineares Modell mit den Variablen aus Punkt 2, indem Sie `family=Gamma` setzen und die kanonische Linkfunktion einsetzen. Beurteilen Sie die Anpassung dieses Modells und führen Sie die Residualanalyse durch. Vergleichen Sie dieses Modell mit dem Modell aus Punkt 2 mit einem geeigneten Kriterium. Welches Modell passt die Daten besser an? Interpretieren Sie die geschätzten Koeffizienten. (6 Punkte)