Aufgaben zur Hausarbeit 1

Aufgabe 1

Datensatz

Laden Sie den Datensatz "KenyaDHS.txt" herrunter. Lesen Sie die Daten in R ein und bereiten Sie den Datensatz wie folgt vor.

1. Behandeln Sie die Ausprägungen "not dejure resident" der Variable `water` als fehlende Werte und entfernen Sie die entsprechenden Stichproben (also Zeilen) aus dem Datensatz.

2. Für die Variable `wealth` ersetzen Sie die beiden Ausprägungen *poorer, poorest* durch *poor* und die Ausprägungen *richer, richest* durch *rich*.

3. Für die Variable `placedelivery` ersetzen Sie die Ausprägungen *govt. dispensary, govt. health center, govt. hospital, other public* durch *govt*, die Ausprägungen *mission hospital/clinic, other private medica, private hosp/clinic* durch *private* und die verbliebenen Ausprägungen durch *home*.

4. Für die Variable `water` ersetzten Sie die Ausprägungen *bottled water, piped into compound/plot, piped into dwelling, public tap* durch *piped*, die Ausprägungen *covered public well, covered well in compound/plot, open public well, open well in compound/plot* durch *well* und die verbliebenen Ausprägungen durch *open*.

Die fehlenden Werte im Datensatz werden automatisch von `aov` entfernt.

Aufgaben

1. Führen Sie zuerst eine einfaktorielle Varianzanalyse für `childweight` mit dem Faktor `water` durch. Stellen Sie die Hypothesen auf und kommentieren Sie das Ergebnis. Überprüfen Sie alle Annahmen einer einfaktoriellen Varianzanalyse. Gibt es Annahme(n), die verletzt sind? Führen Sie einen nicht-parmetrischen Kruskall-Wallis-Test durch. Wie lautet die Nullhypothese von diesem Test? Unterscheidet sich das Ergebnis dieses Tests von der Varianzanalyse? Führen Sie einen Tukey-Test durch, um zu überprüfen welche Paare der Faktorenausprägungen signifikant sind. Sind alle Annahmen dieses Testes erfüllt? Interpretieren Sie das Ergebnis. (6 Punkte)

2. Betrachten Sie den Einfluss der Faktoren `water` und `wealth` auf `childweight`. Erstellen Sie entsprechendene Interaktions-Plots. Sprechen diese Plots für einen signifikanten Interaktionseffekt? Führen Sie eine zweifaktorielle Varianzanalyse durch. Ist der Interaktionseffekt signifikant? Was können Sie über die einzelne Effekte von `water` und `wealth` sagen? Überprüfen Sie alle Annahmen einer zweifaktoriellen Varianzanalyse und kommentieren Sie das Ergebnis. (4 Punkte)

3. Betrachten Sie nun den Einfluss der Faktoren `water` und `placedelivery` auf `childweight`. Erstellen Sie entsprechendene Interaktions-Plots. Sprechen diese Plots für einen signifikanten Interaktionseffekt? Führen Sie eine zweifaktorielle Varianzanalyse für `childweight` und die Faktoren `water` und `placedelivery` durch. Ist der Interaktionseffekt signifikant? Was können Sie in diesem Fall über einzelne Effekte sagen? Überprüfen Sie alle Annahmen einer zweifaktoriellen Varianzanalyse und kommentieren Sie das Ergebnis. Führen Sie einen Tukey-Test durch, um zu überprüfen welche Paare der Interaktionseffekte signifikant sind. Sind die Annahmen dieses Tests erfüllt? Listen Sie fünf Paare der Interaktionseffekte, die den kleinsten $p-$Wert haben und interpretieren Sie diese. (6 Punkte)

Aufgabe 2

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. Anschließend, nehmen Sie nur die 11 numerischen Variablen für die Analyse (Spalten 2,3 und von 11 bis 19).

Aufgaben

1. Erklären Sie, warum die Hauptkomponentenanalyse auf der Kovarianzmatrix für diese Daten nicht sinnvoll ist. Führen Sie eine Hauptkomponentenanalyse auf der Korrelationsmatrix durch. Beschreiben Sie die ersten zwei empirischen Hauptkomponenten und interpretieren Sie diese. Wie viel der gesamten Varianz erklären die ersten beiden Hauptkomponenten? (5 Punkte)

2. Plotten Sie die erste und die zweite empirische Hauptkomponente (Scores) gegeneinander und markieren Sie jeden Punkt mit der Variable `class`, also, ob der Patient eine chronische Nierenerkrankung hat oder nicht. Was fällt Ihnen auf? (2 Punkte)

3. Führen Sie nun eine Faktorenanalyse mit 2 Faktoren durch, indem Sie das Modell mit der Maximum-Likelihood-Methode schätzen, keine Rotation benutzen und die Faktorenwerte mit dem Thomson-Schätzer bestimmen. Schreiben Sie das geschätzte Modell auf und erklären Sie alle Werte im R-Output. Beschreiben Sie die beiden Scores und interpretieren Sie diese; vergleichen Sie diese mit den ersten zwei empirischen Hauptkomponenten aus Teilaufgabe 1. Sind 2 Faktoren für das Modell ausreichend? Warum? Plotten Sie den ersten und den zweiten Score gegeneinander und markieren Sie jeden Punkt mit der Variable `class`. Vergleichen Sie das Ergebnis mit dem Plot aus Punkt 2. Kommentieren Sie die Unterschiede, wenn vorhanden. (7 Punkte)