

Statistik 3

Übung 4

Einleitung und Ziele der Übung

Herzlich Willkommen zur vierten Übung Statistik 3. Diese Übung trainiert mehrere neue Befehle um anschliessend verschiedene Kodierungen in R durchzuführen.

Agenda

- R-Befehle
- Dummykodierung
- Ungewichtete Effektkodierung
- Gewichtete Effektkodierung
- Vergleich mit Mittelwerten

1. R-Befehle - Einführung 1

Führen Sie zunächst folgenden R-Code aus, um einen Beispieldatensatz für die Befehle zu erhalten:

```
1 Beispiel <- data.frame(VP = 1:10,  
2                       Geschlecht = c("m", "m", "w", "d", "m", "d", "w", "w", "w", "w"),  
3                       Neugier = c(22, 31, 23, 21, 32, 42, 32, 18, 15, 16))
```

1. R-Befehle - Einführung 2

`aggregate(x, data, FUN)`

x

- Formel, die angibt, welche Spalte nach den Stufen anderer Spalten zusammengefasst werden soll
- Beispiele (mehr UVs/AVs möglich)
 - V1~V2 (fasst die Werte der Spalte V1 für die Stufen der Spalte V2 getrennt zusammen)
 - V1~V2+V3 (fasst die Werte der Spalte V1 getrennt nach Stufenkombinationen von V1 und V2 zusammen)

data

- Name des Datensatz

FUN

- Funktion die auf Werte mit der gleichen UV-Kombination angewandt werden soll
- häufig mean oder sd

1. R-Befehle - Übung 1

Bestimmen Sie mittels `aggregate()` die Mittelwerte und Standardabweichungen der Neugier je Geschlecht.

$\overline{x}_m$	=	
$\overline{x}_w$	=	
$\overline{x}_d$	=	
s_m	=	
s_w	=	
s_d	=	

1. R-Befehle - Einführung 3

table(x)

x

- Vektor oder Variable
- es wird eine Häufigkeitstabelle für diese Variable ausgegeben

unique(x)

x

- Vektor oder Variable
- es wird jede Ausprägung der Variable genau 1x ausgegeben

1. R-Befehle - Einführung 4

Datensatz[Zeilen,Spalten]

Zeilen

- Regel/Bedingung für die Zeilen angeben, die auszuwählen sind
- keine Angabe = alle Zeilen

Spalten

- Regel/Bedingung für die Spalten angeben, die auszuwählen sind
- keine Angabe = alle Spalten

Beispiel

- Datensatz[1:3,2:4] gibt für die ersten drei Zeilen Werte der 2. bis 4. Spalte zurück
- Datensatz[Datensatz\$Variable == "Text"] gibt alle Spalten für die Zeilen des Datensatzes zurück, bei denen die Spalte Variable den Wert "Text" hat.
- Datensatz\$Variable2[Datensatz\$Variable3 == 3] gibt die Werte von Variable2 zurück für alle Zeilen die in Variable3 den Wert 3 haben

1. R-Befehle - Übung 2

Führen Sie die folgenden Aufträge durch

- lassen sie sich alle Werte von VP (1x ausgeben)
- Bestimmen Sie die Häufigkeiten von Geschlecht
- Lassen Sie sich alle Zeilen ausgeben, deren VP-Nummer grösser als 5 ist
- Lassen Sie sich die Werte der Neugier ausgeben für Personen weiblichen Geschlechts
- Ersetzen Sie die Werte der Neugier für Personen, die sich nicht dem männlichen Geschlecht zugehörig fühlen mit 999
- legen Sie eine neue Spalte Test an, jede Person soll den Wert 80 haben

2. Dummykodierung - Einführung

Jede der Kodierungen wird verwendet um den Einfluss der Ausprägungen nominalskaliertter Variablen auf den Regressanden bestimmen zu können.

Bei der Dummykodierung geht es darum, den Einfluss einer Ausprägung (Referenz) mit den Einflüssen der anderen Ausprägungen zu vergleichen.

Dazu erstellen wir für jede Ausprägung (ausser der Referenz) eine neue (Dummy-)Variable. Eine Person bekommt in einer bestimmten Dummyvariable den Wert 1 zugewiesen, wenn Sie selber diese Ausprägung aufweist, ansonsten den Wert 0.

2. Dummykodierung - Übung 1

Importieren Sie zunächst den Datensatz `Uebung_Statistik3_4a.csv` und verschaffen Sie sich einen Überblick über die darin enthaltenen Variablen/Spalten.

Wir werden eine Dummykodierung für die Variable `Lieblingsessen` vornehmen um anschliessend die Offenheit der Personen vorherzusagen. Nutzen Sie `table()` um alle möglichen Ausprägungen der Variable `Lieblingsessen` zu bestimmen. Wie viele verschiedene Ausprägungen sind vorhanden?

Wie viele Dummyvariablen sind somit nötig?

Führen Sie die Dummykodierung durch

- Die Referenz soll zunächst noch offen sein. Daher legen Sie für jede Ausprägung des `Lieblingsessen` eine neue Spalte (Dummyvariable) im Datensatz an. Der Name der Spalte sei identisch mit der jeweiligen Ausprägung.
- Weisen Sie zunächst allen Personen bei jeder Dummyvariable den Wert 0 zu
- Weisen Sie anschliessend jeder Dummyvariable den Wert 1 zu, für jede Person, deren `Lieblingsessen` der Ausprägung der Dummyvariable entspricht.
- Benutzen Sie `table()` mit jeder der Dummyvariablen um zu überprüfen ob die Anzahl der Personen mit dem Wert 1, der tatsächlichen Häufigkeit der Ausprägung in Stichprobe entspricht.

2. Dummykodierung - Übung 2

Normalerweise wird für die Referenz keine Dummyvariable angelegt. Führen Sie trotzdem eine Regression mit Offenheit als AV und allen Dummyvariablen als UV durch und lassen Sie sich die Ergebnisse mittels `summary()` anzeigen. Was fällt Ihnen auf?

Führen Sie die Regression einmal mit Pommes und einmal mit Pizza als Referenz durch (die jeweilige Dummyvariable sollte also nicht mehr enthalten sein). Schauen Sie sich jeweils den Haupteffekt von Sushi an. Was fällt Ihnen auf und wie kommt es dazu?

Wie sind die Regressionskoeffizienten zu interpretieren? Was sagt uns ein signifikanter/ nicht signifikanter Koeffizient jeweils?

3. Ungewichtete Effektkodierung - Einführung

Bei der ungewichteten Effektkodierung geht es darum, den Einfluss jeder Ausprägung (ausser der Referenz) mit dem ungewichteten Mittelwert über die Ausprägungen zu vergleichen.

Dazu erstellen wir für jede Ausprägung (ausser der Referenz) eine neue (Effekt-)Variable. Eine Person bekommt in einer bestimmten Effektvariable den Wert

- 1 zugewiesen, wenn sie selber diese Ausprägung aufweist.
- 0 zugewiesen, wenn sie eine andere Ausprägung als die Ausprägung der Effektvariable und der Referenz aufweist.
- -1 zugewiesen, wenn sie als Ausprägung die Referenz aufweist.

3. Ungewichtete Effektkodierung - Übung 1

Führen Sie die ungewichtete Effektkodierung mit Pommes als Referenz in R durch. Das heisst, für Pommes ist keine Effektvariable notwendig.

Überprüfen Sie die Häufigkeiten erneut mittels `table()`.

3. Ungewichtete Effektkodierung - Übung 2

Führen Sie die Regression der Offenheit auf die (ungewichteten) Effektvariablen durch. Welche Hypothesen wurden diesmal für die einzelnen Regressionskoeffizienten geprüft.

Wie sind die Regressionskoeffizienten zu interpretieren? Was sagt uns ein signifikanter/ nicht signifikanter Koeffizient jeweils?

4. Gewichtete Effektkodierung - Einführung

Bei der gewichteten Effektkodierung geht es darum, den Einfluss jeder Ausprägung (ausser der Referenz) mit dem gewichteten Mittelwert über die Ausprägungen zu vergleichen.

Dazu erstellen wir für jede Ausprägung (ausser der Referenz) eine neue (Effekt-)Variable. Eine Person bekommt in einer bestimmten Effektvariable den Wert

- 1 zugewiesen, wenn sie selber diese Ausprägung aufweist.
- 0 zugewiesen, wenn sie eine andere Ausprägung als die Ausprägung der Effektvariable und der Referenz aufweist.
- $\frac{n_{\text{Ausprägung}_{\text{Effektvariable}}}}{n_{\text{Ausprägung}_{\text{Referenz}}}}$ zugewiesen, wenn sie als Ausprägung die Referenz aufweist.

2. Gewichtete Effektkodierung - Übung 1

Führen Sie die gewichtete Effektkodierung mit Pommies als Referenz in R durch. Das heisst, für Pommies ist keine Effektvariable notwendig.

Überprüfen Sie die Häufigkeiten erneut mittels `table()`.

4. Gewichtete Effektkodierung - Übung 2

Führen Sie die Regression der Offenheit auf die (gewichteten) Effektvariablen durch. Welche Hypothesen wurden diesmal für die einzelnen Regressionskoeffizienten geprüft.

Wie sind die Regressionskoeffizienten zu interpretieren? Was sagt uns ein signifikanter/ nicht signifikanter Koeffizient jeweils?

4. Vergleich mit Mittelwerten - Einführung

Alle unsere Hypothesen beziehen sich auf verschiedene Mittelwerte oder Mittelwertsdifferenzen. Im letzten Abschnitt werden Sie verschiedene Mittelwerte bestimmen und diese mit Ihren Regressionskoeffizienten vergleichen.

4. Vergleich mit Mittelwerten - Übung

Bestimmen Sie folgende Mittelwerte mittels der Funktionen `mean()` und/oder `aggregate()`.

- den Mittelwert der Offenheit über die Gesamtstichprobe
- die Mittelwerte der Offenheit getrennt nach den Lieblingsessen
- den Mittelwert über die “Mittelwerte der Offenheit getrennt nach den den Lieblingsessen”

Vergleichen Sie diese Mittelwerte mit den Achsenabschnitten ihrer Regressionen. Was fällt Ihnen auf?

Können Sie auch die anderen Regressionskoeffizienten aus den Mittelwerten herleiten?