

ASMNW - Lösung 3

Peter von Rohr

2016-04-26

Kontrollfrage 1

1. Aus welchen Komponenten besteht ein lineares Modell?
2. Welcher Zusammenhang besteht zwischen diesen Komponenten

Lösung

1. Zielgrösse, erklärende Variablen und Resteffekte
2. Zielgrösse soll als lineare Funktion der erklärenden Variablen dargestellt werden

Kontrollfrage 2

1. Welche fixen und welche zufälligen Effekte gibt es in einer multiplen linearen Regression
2. Wie lauten Erwartungswerte und Varianzen der zufälligen Effekte

Lösung

1. erklärende Variablen sind fix und Resteffekte sind zufällig
2. Erwartungswert $E(\epsilon) = \mathbf{0}$ und die Co-Varianz-Matrix beträgt: $Var(\epsilon) = \mathbf{I} * \sigma^2$

Kontrollfrage 3

Welche vier Ziele wollen wir mit einer multiplen linearen Regression erreichen?

Lösung

1. gute **Anpassung** des Modells, so dass Abweichungen zu beobachteten Zielgrössen möglichst klein
2. gute **Schätzung** der Parameter, Änderungen bei den erklärenden Variablen sollen möglichst eng mit Änderungen in Zielgrössen zusammenhängen
3. gute **Voraussage** für neue Werte der erklärenden Variablen innerhalb des Wertebereichs des Datensatzes, keine Extrapolation
4. Unsicherheiten sollen durch Tests und Vertrauensintervalle quantifiziert werden

Kontrollfrage 4

1. Was bedeutet der Ausdruck

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2$$

?

2. Welcher Schätzer resultiert aus dem Ausdruck unter 1?

Lösungen

1. Der Ausdruck $\|\mathbf{y} - \mathbf{X}\beta\|^2$ wird nach β abgeleitet und 0 gesetzt und der Wert für β an diesem Minimum wird als Schätzer $\hat{\beta}$ verwendet.
- 2.

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Aufgabe 1:

Während der Vorlesung habe wir das Beispiel mit der Zunahme nach dem Absetzen als Zielgrösse betrachtet. Wir haben aber nirgends die Koeffizienten und die geschätzte Restvarianz berechnet. Stellen Sie für dieses kleine Beispiel das lineare Modell gemäss Vorlesungsunterlagen auf und berechnen Sie die Koeffizienten des linearen Modells und die geschätzte Restvarianz.

Lösung

Zuerst werden die Daten in ein `data.frame` abgelegt.

```
dfBwWgData <- data.frame(BW = c(35.0, 25.3, 34.2, 31.2, 38.7),
                        WWG = c(0.90, 0.58, 0.78, 0.70, 1.00),
                        PWG = c(1.36, 1.00, 1.36, 1.20, 1.50))
```

Mit diesem `data.frame` können wir schon ein lineares Modell anpassen.

```
lmBwWg <- lm(PWG ~ ., data = dfBwWgData)
summary(lmBwWg)

##
## Call:
## lm(formula = PWG ~ ., data = dfBwWgData)
##
## Residuals:
##      1      2      3      4      5
## -9.919e-05  6.950e-03  2.004e-02 -2.439e-02 -2.497e-03
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.008119   0.101168   0.080   0.9433
## BW           0.041413   0.008749   4.733   0.0419 *
## WWG        -0.108294   0.266154  -0.407   0.7235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02292 on 2 degrees of freedom
## Multiple R-squared:  0.9928, Adjusted R-squared:  0.9856
## F-statistic: 137.9 on 2 and 2 DF, p-value: 0.007201
```

Die Modellanpassung ohne Achsenabschnitt

```
lmBwWgNI <- lm(PWG ~ -1 + BW + WWG, data = dfBwWgData)
summary(lmBwWgNI)
```

```
##
## Call:
## lm(formula = PWG ~ -1 + BW + WWG, data = dfBwWgData)
##
## Residuals:
##      1      2      3      4      5
## 0.0002699  0.0084706  0.0191781 -0.0246233 -0.0028784
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## BW  0.041997   0.003965  10.592  0.0018 **
## WWG -0.122413   0.163336  -0.749  0.5080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01875 on 3 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9998
## F-statistic: 1.194e+04 on 2 and 3 DF,  p-value: 1.409e-06
```

Anstatt das Modell ohne Achsenabschnitt ganz neu anzugeben, gibt es auch die Möglichkeit ein bestehendes Modell mit der Funktion `update()` anzupassen. Dadurch können wir mit folgendem Befehl das ursprüngliche Modell in `lmBwWg` anpassen. Dabei ist das erste Argument von `update()` das ursprüngliche Objekt, welches das lineare Modell enthält (bei uns ist das `lmBwWg`). Das zweite Argument ist die neue Modellformel. Im nachfolgenden Aufruf von `update()` wird eine abgekürzte Schreibweise verwendet, wobei die Punkte als Platzhalter für die Komponenten im schon bestehenden Modell in `lmBwWg` stehen. Somit heisst die Abkürzung `. ~ . -1`: nimm das bestehende Modell und entferne den Achsenabschnitt.

```
lmBwGwUpd <- update(lmBwWg, . ~ . -1)
summary(lmBwGwUpd)
```

```
##
## Call:
## lm(formula = PWG ~ BW + WWG - 1, data = dfBwWgData)
##
## Residuals:
##      1      2      3      4      5
## 0.0002699  0.0084706  0.0191781 -0.0246233 -0.0028784
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## BW  0.041997   0.003965  10.592  0.0018 **
## WWG -0.122413   0.163336  -0.749  0.5080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01875 on 3 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9998
## F-statistic: 1.194e+04 on 2 and 3 DF,  p-value: 1.409e-06
```

Die Resultate der Modellanpassung mit und ohne Achsenabschnitte mögen auf den ersten Blick verwirrend aussehen. Aufgrund des höheren R^2 -Wertes und basierend auf der tieferen Restvarianz des Modells ohne Achsenabschnitt, würde man dieses Modell ohne Achsenabschnitt als das bessere Modell bezeichnen. Aufgrund der biologischen Zusammenhänge macht aber ein Modell ohne Achsenabschnitt wenig Sinn. Als Erklärung dafür gilt wohl, dass nur fünf Beobachtungen wohl zu wenig sind um zwischen zwei Modellen sicher unterschieden zu können. Des Weiteren sind die Unterschiede zwischen den Modellen sehr gering und würden eine Rangierung der Modelle kaum erlauben.