

# MSc. Research Methods - Statistikteil Lösungen 2018

Gian-Andrea Egeler

November 2018

## Übung 4.2S: multiple logistische Regression

```
# Generiert eine Dummyvariable: Fleisch 1, kein Fleisch 0
df <- nova # kopiert originaler Datensatz
df$meat <- ifelse(nova$label_content == "Fleisch", 1, 0)
df_ <- df[df$label_content != "Buffet", ] # entfernt Personen die sich ein Buffet Teller gekauft haben

# Löscht alle Missings bei der Variable "Fleisch"
df_ <- df_[!is.na(df_$meat), ]

# sieht euch die Verteilung zwischen Fleisch und kein Fleisch
table(df_$meat)

##
##    0    1
## 387 564

# definiert das logistische Modell und wende es auf den Datensatz an
mod0 <- glm(meat ~ gender + member + age, data = df_, binomial("logit"))
summary.lm(mod0) # Member und Alter scheinen keinen Einfluss zu nehmen, lassen wir also weg

##
## Call:
## glm(formula = meat ~ gender + member + age, family = binomial("logit"),
##      data = df_)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6134 -1.0258  0.7174  0.7443  1.1998
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.407608   0.395553   1.030   0.303
## genderM        0.733556   0.141743   5.175 2.78e-07 ***
## memberStudierende -0.218506   0.197620  -1.106   0.269
## age            -0.012299   0.009312  -1.321   0.187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.002 on 947 degrees of freedom
## Multiple R-squared:  0.001772, Adjusted R-squared: -0.00139
## F-statistic: 0.5603 on 3 and 947 DF, p-value: 0.6413

# neues Modell ohne Alter und Hochschulzugehörigkeit
mod1 <- update(mod0, ~. -member - age)
summary.lm(mod1)

##
```

```
## Call:
## glm(formula = meat ~ gender, family = binomial("logit"), data = df_)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3687 -0.9506  0.7306  0.7306  1.0520
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1014     0.1128  -0.899   0.369
## genderM       0.7291     0.1403   5.198 2.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.001 on 949 degrees of freedom
## Multiple R-squared:  0.001681, Adjusted R-squared:  0.0006287
## F-statistic: 1.598 on 1 and 949 DF, p-value: 0.2065

# Modeldiagnostik (wenn nicht signifikant, dann OK)
1 - pchisq(mod1$deviance, mod1$df.resid) # hochsignifikant, d.h. kein guter Modellfit??

## [1] 5.353906e-11

#Modellgüte (pseudo-R²)
1 - (mod1$dev / mod1$null) # sehr kleines pseudo-R²

## [1] 0.02121993

# Konfusionsmatrix vom Datensatz
# Model Vorhersage
predicted <- predict(mod1, df_, type = "response")

# erzeugt eine Tabelle mit den beobachteten
# Fleischesser/Nichtfleischesser und den Vorhersagen des Modells
km <- table(df_$meat, predicted > 0.5)
dimnames(km) <- list(
  c("Beobachtung kein Fleisch", "Beobachtung Fleisch"),
  c("Modell kein Fleisch", "Modell Fleisch"))
km

##              Modell kein Fleisch Modell Fleisch
## Beobachtung kein Fleisch              166              221
## Beobachtung Fleisch              150              414

# kalkuliert die Missklassifizierungsrate
mf <- 1-sum(diag(km)/sum(km)) # ist mit knapp 40% eher hoch
mf

## [1] 0.3901157
```

## Methoden

Die Kriteriumsvariable “Fleischkonsum” ist eine binäre Variable. Demnach wird eine multiple logistische Regression mit den Prädiktoren “Alter”, “Geschlecht” und “Hochschulzugehörigkeit” gerechnet. Die Modelldiagnostik und das pseudo- $R^2$  zeigen allerdings, dass das Modell nicht gut zu den empirischen Daten passt.

## Ergebnisse

Mit der logistischen Regression kann der Fleischkonsum weder durch das Geschlecht, die Hochschulzugehörigkeit noch das Alter vorhergesagt werden. Die Tests für die Modelldiagnostik und das kleine pseudo- $R^2$  unterstützen diesen Befund. Auch die hohe Missklassifizierungsrate (39%) deutet auf ein Modell, welches nicht zu den Daten passt. Es sollte nach einem weiteren adäquateren Modell gesucht werden. Bei näherer Betrachtung der Daten erkennt man, dass einige Personen wiederholt im Datensatz auftauchen (siehe `card_num`). Bei einer weiteren Analyse müssten die einzelnen Individuen ebenfalls berücksichtigt werden z. B. mit genesteten Modellen (siehe Statistik 5).