

Tutorial 4

Logistische Regression

Aufgabe 1: Titanic

In diesem Tutorial arbeiten wir mit personenbezogenen Daten zu $n = 884$ Passagieren der Jungfernfahrt der Titanic. Nutzen Sie zunächst folgenden Befehl, um die Daten einzulesen:

```
titanic_data = read.csv("https://tinyurl.com/3zfdz99m")
```

Die Variable **Ueberlebt** gibt an, ob der/die Passagier*in überlebt hat (= 1) oder nicht (= 0). Die Variable **Preis** gibt die Kosten des Tickets der betreffenden Person an.

- a) Wir wollen zunächst ein bisschen explorative Datenanalyse betreiben.
 - Wie viele Personen haben überlebt?
 - Wie groß ist das Durchschnittsalter?
 - Wie viele Männer bzw. Frauen sind im Datensatz enthalten?
 - Wie viele Frauen/Männer haben überlebt?
- b) Erstellen Sie einen Scatterplot der Variablen **Ueberlebt** (y-Achse) vs. **Preis** (x-Achse). Bestimmen Sie mit `cor()` auch den empirischen Korrelationskoeffizienten dieser beiden Variablen.
- c) Mit welcher Verteilung ist die Variable **Ueberlebt** zu modellieren? Geben Sie die Gleichung eines entsprechenden Regressionsmodells an, in dem **Preis** als erklärende Variable fungiert. Rufen Sie sich dabei noch einmal in Erinnerung, warum diese spezielle Form der Regressionsfunktion gewählt wird.
- d) Schätzen Sie mit `glm()` das in c) formulierte Modell — das angepasste Modell soll als Objekt `mod` gespeichert werden (für eine spätere Teilaufgabe) — und betrachten Sie die Modellzusammenfassung. Welcher Zusammenhang besteht zwischen dem Ticketpreis und der Überlebenswahrscheinlichkeit?
- e) Geben Sie auch eine Interpretation von $\hat{\beta}_1$ an.
- f) Wie hoch war gemäß des Modells die Überlebenswahrscheinlichkeit einer Person, die 100 Dollar für ihr Ticket gezahlt hat?
- g) Zeichnen Sie die angepasste Regressionsfunktion in den Scatterplot aus b) ein und kontrollieren Sie grafisch Ihr Ergebnis aus f):

```
curve(plogis(mod$coeff[1] + mod$coeff[2] * x), lwd = 2, col = "#E69F00",
      add = TRUE)
```

Lesen Sie zudem rein visuell aus der Grafik ab, für welchen gezahlten Preis eine Person gemäß des Modells eine Überlebenswahrscheinlichkeit von 50% gehabt hätte.

h) Schätzen Sie nun mithilfe der `glm()`-Funktion das multiple logistische Regressionsmodell

$$\text{Ueberlebt}_i \sim \text{Bern}(\pi_i), \pi_i = \frac{e^{\beta_0 + \beta_1 \text{Preis}_i + \beta_2 \text{Alter}_i + \beta_3 \text{Alter}_i^2 + \beta_4 \text{Geschlecht}_i}}{1 + e^{\beta_0 + \beta_1 \text{Preis}_i + \beta_2 \text{Alter}_i + \beta_3 \text{Alter}_i^2 + \beta_4 \text{Geschlecht}_i}}.$$

Beziehen Sie in Anbetracht der Modellzusammenfassung Stellung zu der Vermutung, dass die Plätze in den (zu) wenigen Rettungsbooten zunächst Frauen, Kindern und älteren Passagieren vorbehalten waren.

i) Zusatzaufgabe: mit dem nachfolgenden Code erhalten Sie eine Visualisierung des in h) angepassten Modells — copy-pasten Sie den Code und kontrollieren Sie grafisch Ihre Ergebnisse aus h).

```
x1 = seq(min(titanic_data$Preis), max(titanic_data$Preis), length = 40)
x2 = seq(min(titanic_data$Alter), max(titanic_data$Alter), length = 40)
predictor1 = function(x1, x2){
  plogis(mod$coeff[1] + mod$coeff[2] * x1 + mod$coeff[3] * x2 +
    mod$coeff[4] * x2 ^ 2)
}
predictor2 = function(x1, x2){
  plogis(mod$coeff[1] + mod$coeff[2] * x1 + mod$coeff[3] * x2 +
    mod$coeff[4] * x2 ^ 2 + mod$coeff[5])
}
z1 = outer(x1, x2, predictor1)
z2 = outer(x1, x2, predictor2)
ind = which(titanic_data$Geschlecht == "female")
par(mfrow = c(1, 2))
res = persp(x1, x2, z1, theta = -50, phi = 5, xlab = "Preis", ylab = "Alter",
  zlab = "Ueberlebt", zlim = c(0,1), ticktype="detailed",
  col = "orange", main = "weibliche Passagiere")
points(trans3d(titanic_data$Preis[ind], titanic_data$Alter[ind],
  titanic_data$Ueberlebt[ind], pmat = res),
  col = "black", pch = 16, cex = 0.6)
res = persp(x1, x2, z2, theta = -50, phi = 5, xlab = "Preis",
  ylab = "Alter", zlab = "Ueberlebt", zlim = c(0,1),
  ticktype = "detailed", col = "orange",
  main = "männliche Passagiere")
```

```
points(trans3d(titanic_data$Preis[-ind], titanic_data$Alter[-ind],  
              titanic_data$Ueberlebt[-ind], pmat = res), col = "black",  
       pch = 19, cex = 0.6)
```