Tut 4

Aufgabe 1

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

Die Variable Ueberlebt gibt an, ob der/die Passagier 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 Durschnittsalter?
- Wie viele Männer bzw. Frauen sind im Datensatz enthalten?
- Wie viele Frauen/Männer haben überlebt?

```
# Wie viele Personen haben überlebt?

sum(Ueberlebt == 1)
```

```
[1] 339
```

```
# Es haben 339 Passagiere überlebt, 545 Passagiere haben nicht überlebt.
```

```r
# Wie groß ist das Durchschnittsalter?

mean(Alter)
```

```
[1] 29.45155
```

```r
# Das Durchschnittsalter liegt bei etwa 29.5 Jahren.
```

```r
# Wie viele Männer bzw. Frauen sind im Datensatz enthalten?

table(Geschlecht)
```

```
Geschlecht
female   male
   313    571
```

```r
# Der Datensatz enthält 313 Frauen und 571 Männer
```

```r
# Wie viele Frauen/Männer haben überlebt?

table(Geschlecht[Ueberlebt == 1])
```

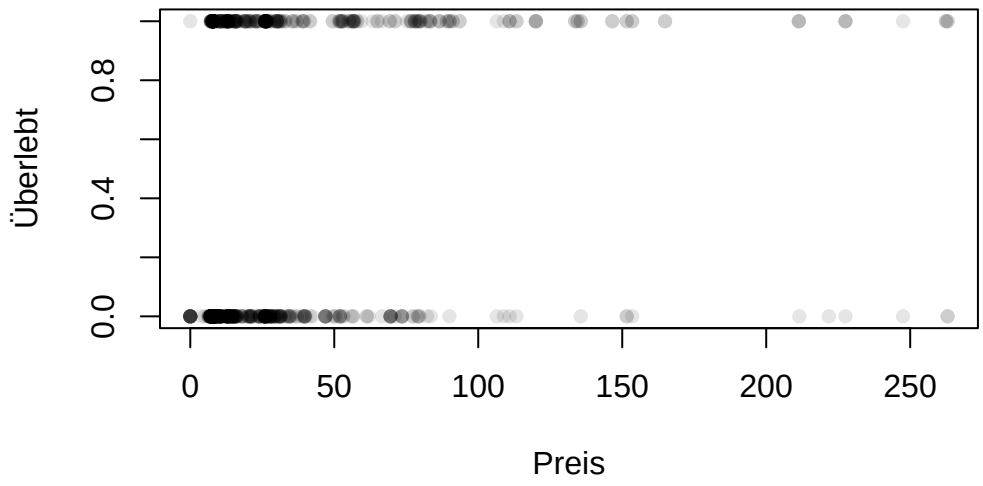
```
female   male
   232    107
```

```r
# Es haben 232 Frauen und 107 Männer 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.

```r
plot(Preis, Ueberlebt, pch = 16, col = "#0000001A", ylab="Überlebt",xlab="Preis")
```




```
cor(Preis, Ueberlebt)
```

```
[1] 0.2604613
```

```
# Die Korrelation beträgt in etwa 0.26
# Wir haben also eine schwache positive Korrelation
```

c)

Mit welcher Verteilung ist die Variable Ueberlebt zu modellieren? Geben Sie die Gle ichung 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.

Antwort:

Die Variable "Ueberlebt" ist Bernoulli verteilt und kann mit einer logistischen Regression modelliert werden.

$\text{Ueberlebt}_i \sim \text{Bern}(\pi_i), \quad \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}, \quad \eta_i = \beta_0 + \beta_1 \cdot \text{Preis}_i$

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?

```r
mod <- glm(Ueberlebt ~ Preis, family = binomial(link = "logit"))
# Wenn Daten nicht attached sind, muss "data = titanic" in der Funktion
# ergänzt werden

summary(mod)
```

```
Call:
glm(formula = Ueberlebt ~ Preis, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4821  -0.8914  -0.8562   1.3497   1.5898

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.931520   0.095280  -9.777  < 2e-16 ***
Preis        0.015076   0.002232   6.755 1.43e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1177.0  on 883  degrees of freedom
Residual deviance: 1114.6  on 882  degrees of freedom
AIC: 1118.6

Number of Fisher Scoring iterations: 4
```

$\hat{\beta}_1 > 0$, es besteht also ein positiver Zusammenhang zwischen dem Preis und der Überlebenswahrscheinlichkeit.

e)

Geben Sie auch eine Interpretation von $\hat{\beta}_1$ an.

```
exp(mod$coef[2])
```

```
  Preis
1.01519
```

Zahlt man eine Preiseinheit mehr für das Ticket, so erhöhen sich die Odds multiplikativ um $e^{\hat{\beta}_1}$ (1.01519).

f)

Wie hoch war gemäß des Modells die Überlebenswahrscheinlichkeit einer Person, die 100 Dollar für ihr Ticket gezahlt hat?

```
exp(mod$coef[1] + mod$coef[2] * 100) / (1 + exp(mod$coef[1] + mod$coef[2] * 100))
```

```
(Intercept)
  0.6401557
```

```
plogis(mod$coef[1] + mod$coef[2] * 100)
```

```
(Intercept)
  0.6401557
```

```
predict(mod, newdata = data.frame(Preis = 100), type = "response")
```

```
        1
0.6401557
```

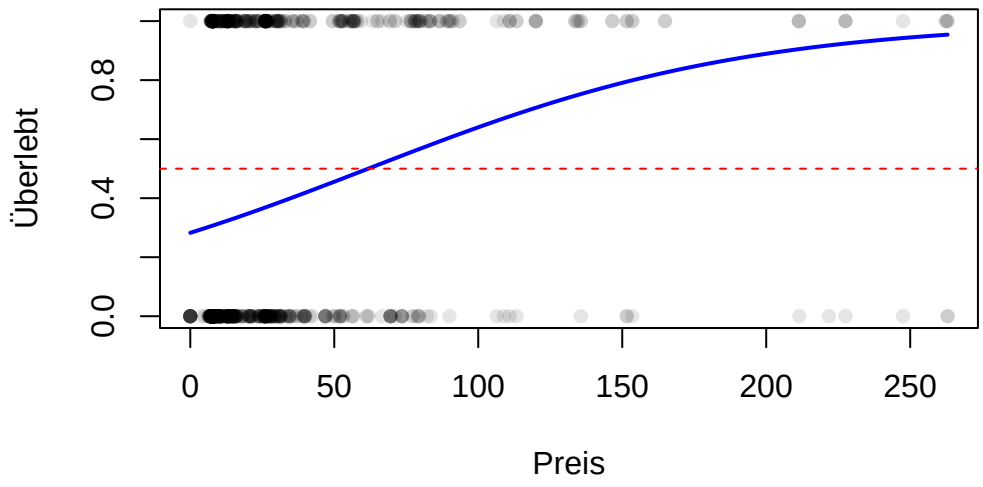
```
# response = Ausgabe der Wahrscheinlichkeiten

# Das Modell gibt bei einem Ticketpreis von 100 Preiseinheiten eine
# Überlebenswahrscheinlichkeit von etwa 64% an.
```

g)

Zeichnen Sie die angepasste Regressionsfunktion in den Scatterplot aus b) ein und kon trollieren Sie grafisch Ihr Ergebnis aus f):

```
plot(Preis, Ueberlebt, pch = 16, col = "#0000001A", ylab="Überlebt",xlab="Preis")
curve(plogis(mod$coef[1] + mod$coef[2] * x), lwd = 2, col = "blue", add = TRUE)
abline(h = 0.5, col = "red", lty = 2)
```



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

Antwort:

Das Modell schätzt eine Überlebenswahrscheinlichkeit von etwa 50% bei einem Ticketpreis von etwa 60 Preiseinheiten.

f)

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

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

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.

```
mod <-glm (Ueberlebt ~ Preis + Alter + I(Alter^2) + Geschlecht,
           family="binomial")

summary(mod)
```

```
Call:
glm(formula = Ueberlebt ~ Preis + Alter + I(Alter^2) + Geschlecht,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3641  -0.6149  -0.5715   0.8153   1.9769

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.0740772  0.3252678   3.302  0.00096 ***
Preis           0.0112609  0.0023800   4.732 2.23e-06 ***
Alter          -0.0252221  0.0193486  -1.304  0.19238
I(Alter^2)      0.0002804  0.0002873   0.976  0.32908
Geschlechtmale -2.3975778  0.1715620 -13.975  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1177.0  on 883  degrees of freedom
Residual deviance:  880.4  on 879  degrees of freedom
AIC: 890.4

Number of Fisher Scoring iterations: 4
```

Das geschätzte Modell scheint den Ehrenkodex "Kinder, Alte und Frauen zuerst" grundsätzlich zu bestätigen (denn $\hat{\beta}_4$ ist negativ und die Tatsache, dass $\hat{\beta}_3$ positiv ist, bedeutet, dass wir für den Effekt des Alters eine U-Form haben). Allerdings ist das Alter nicht signifikant, der Ticketpreis war offensichtlich viel entscheidender — vermutlich weil die günstigsten Tickets zu Kabinen weiter unten im Schiffsrumpf gehörten.

Zusatzaufgabe

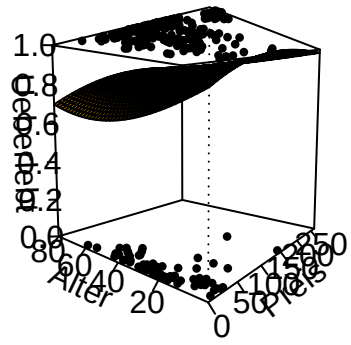
Zusatzaufgabe: mit dem nachfolgenden Code erhalten Sie eine Visualisierung des in h) angepassten Modells — copy-pasten Sie den Code und kontrollieren Sie grafisch Ihre Ergebnisse aus h).

```
x1 = seq(min(Preis), max(Preis), length = 40)
 x2 = seq(min(Alter), max(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(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(Preis[ind], Alter[ind],
 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")
 2
```

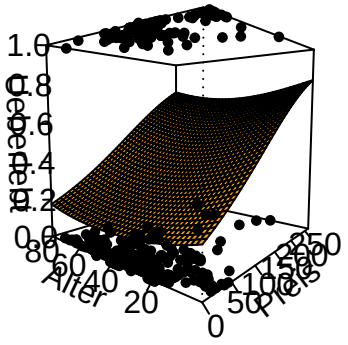
```
[1] 2
```

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

weibliche Passagiere



männliche Passagiere



- man sieht den starken Effekt des Geschlechts
- positiver Effekt vom Preis ist erkennbar
- der schwache quadratische Effekt ist erkennbar