Tutorial 4

Lars Abt

Aufgabe 1a)

Wir beginnen damit, die Daten zu den Passagieren der Titanic einzulesen, sowie alle für uns relevanten Packages zu laden. Ich nutze dafür erneut die 'p_load()' Funktion aus dem 'pacman' Paket, und greife damit auf 'data.table' und 'ggplot2' zu.

```
pacman::p_load(data.table, ggplot2)

titanic_data <- fread("https://tinyurl.com/3zfdz99m")
```

Über diese Daten möchten wir nun mehr erfahren:

1. Wie viele Personen haben überlebt?

```
sum(titanic_data$Ueberlebt)
```

```
[1] 339
```

```
table(titanic_data$Ueberlebt)
```

```
  0   1
545 339
```

2. Wie hoch ist das Durchschnittsalter?

```
mean(titanic_data$Alter)
```

```
[1] 29.45155
```

3. Wie viele Männer bzw. Frauen sind im Datensatz enthalten?

```r
table(titanic_data$Geschlecht)
```

```
female   male
   313    571
```

4. Wie viele Frauen/Männer haben überlebt?

```r
sum(titanic_data[titanic_data$Geschlecht == "male",]$Ueberlebt)
```

```
[1] 107
```

```r
sum(titanic_data[titanic_data$Geschlecht == "female",]$Ueberlebt)
```

```
[1] 232
```

```r
# bzw. als relativer Anteil:
mean(titanic_data[titanic_data$Geschlecht == "male",]$Ueberlebt)
```

```
[1] 0.1873905
```

```r
mean(titanic_data[titanic_data$Geschlecht == "female",]$Ueberlebt)
```

```
[1] 0.7412141
```

```r
# Alternative:
tapply(titanic_data$Ueberlebt, titanic_data$Geschlecht, FUN = mean)
```

```
   female      male
0.7412141 0.1873905
```

Aufgabe 1b)

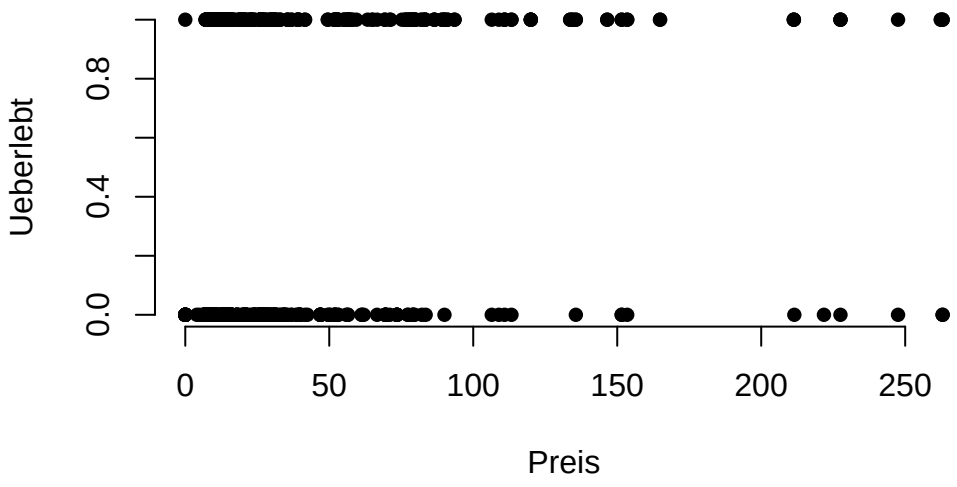
Wir erstellen nun ein Streudiagramm der Variablen "Ueberlebt" (y-Achse) vs. "Preis" (x–Achse). Anschließend bestimmen wir den empirischen Korrelationskoeffizienten dieser beiden Variablen:

```r
cor(titanic_data$Preis, titanic_data$Ueberlebt)
```

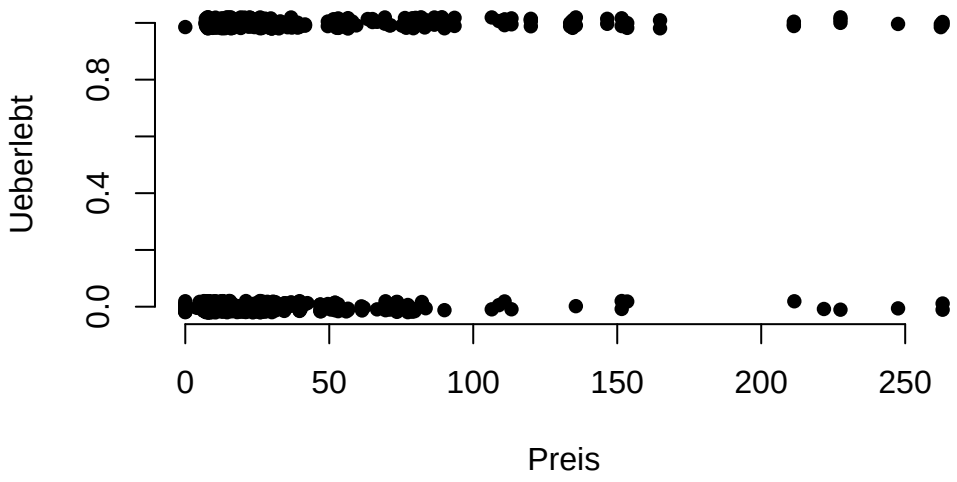
```
[1] 0.2604613
```

Wir sehen, dass wir als Ergebnis eine Wert von 0.26 erhalten, also durchaus eine gewissen Korrelation zwischen den beiden Variablen Preis und Überlebt besteht. Dies ergibt intuitiv Sinn, da wir annehmen können, dass wohlhabendere Passagiere eher überlebt haben als ärmere Passagiere. Anschließend möchten diesen Zusammenhang grafisch darstellen.

```r
plot(titanic_data$Preis, titanic_data$Ueberlebt,
     pch = 16, bty = "n", ylab = "Ueberlebt", xlab = "Preis")
```



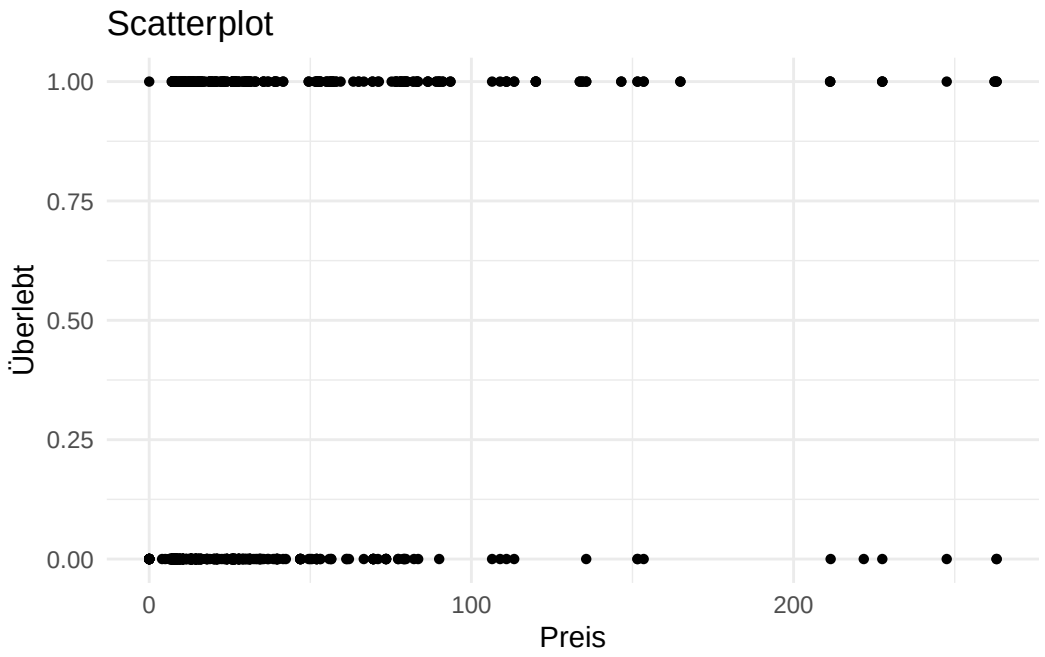

```r
plot(titanic_data$Preis, jitter(titanic_data$Ueberlebt, 0.1),
     pch = 16, bty = "n", ylab = "Ueberlebt", xlab = "Preis")
```



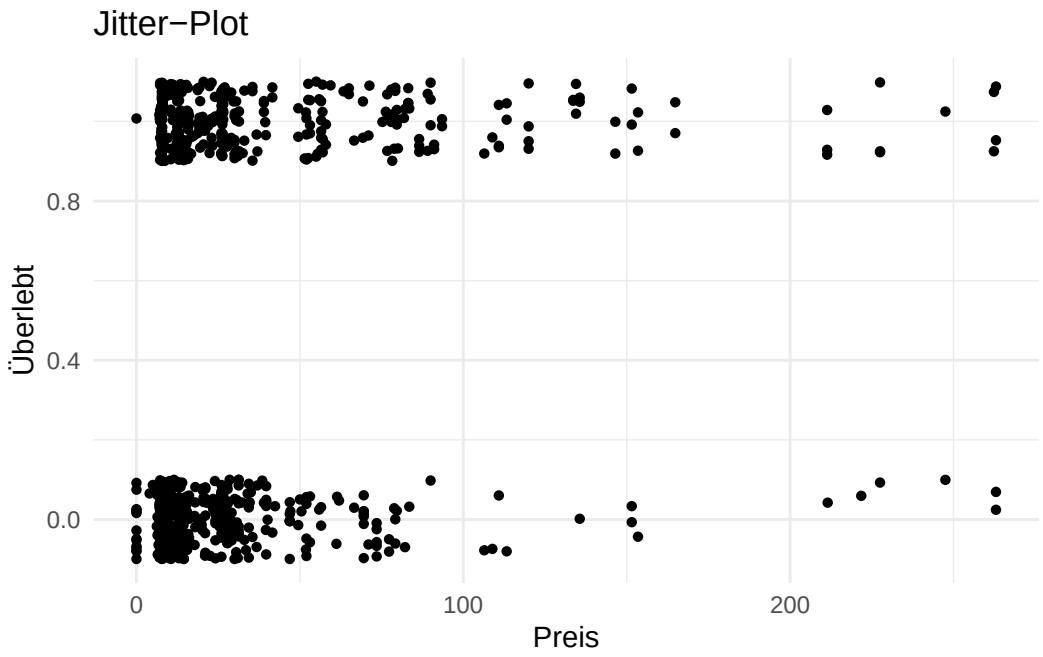
Unser Output gibt uns zwei Graphen, welche grundsätzlich dasselbe Bild vermitteln. Die zweite Version ist lediglich gejittered, was dabei helfen kann, die Datenpunkte besser zu erkennen, insbesondere die große Menge an Punkten am unteren Ende. So ist es leichter, die Dichte der Punkte zu erkennen, was bei einer Menge von insgesamt 884 Punkten(Passagieren) durchaus ratsam sein kann.

Nun möchten wir auch noch einmal ggplot2 nutzen, um denselben Zusammenhang darzustellen:

```
gg_scatterplot <-
  ggplot(titanic_data, aes(x = Preis, y = Ueberlebt)) +
  geom_point(shape = 16) +
  labs(x = "Preis", y = "Überlebt", title = "Scatterplot") +
  theme_minimal()
gg_scatterplot
```




```
gg_jitterplot <-
  ggplot(titanic_data, aes(x = Preis, y = Ueberlebt)) +
  geom_jitter(width = 0, height = 0.1, shape = 16) +
  labs(x = "Preis", y = "Überlebt", title = "Jitter-Plot") +
  theme_minimal()
gg_jitterplot
```



Aufgabe 1c)

Wie können wir die Variable "Ueberlebt" nun modellieren?

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

Die Variable "Ueberlebt" ist binär, also nimmt sie nur die Werte 0 ("nicht überlebt") oder 1 ("überlebt") an. Wir können also eine logistische Regression verwenden, um die Überlebenswahrscheinlichkeit in Abhängigkeit vom Preis zu modellieren. Dafür transfomrmieren wir den linearen Prädiktor _i mit Hilfe der logistischen Funktion, welche lediglich Werte zwischen 0 und 1 annehmen kann.

Aufgabe 1d)

Nun schätzen wir eben dieses Modell mit Hilfe der 'glm()' Function, welche wir bereits von der Poisson-Regression kennen.

```
mod <- glm(Ueberlebt ~ Preis, family = "binomial", data = titanic_data)

summary(mod)
```

```
Call:
glm(formula = Ueberlebt ~ Preis, family = "binomial", data = titanic_data)

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
```

Ganz wichtig! Das 'family' Argument muss nun das Argument "binomial" enthalten, anstatt "poisson". Die Anführungszeichen sind hier übrigens nicht entscheidend, es funktioniert auch ohne sie. Wir sehen nun in unserem Modell-Summary, dass der Preis einen positiven Einfluss auf die Überlebenswahrscheinlichkeit hat, genau wie unsere Korrelationsberechnung es bereits vermuten ließ. Zudem sehen wir, dass diese Schätzung auf einem Signifikanzniveau von 0.001 signifikant ist, was bedeutet, dass wir mit einer Wahrscheinlichkeit von 99.9% sagen können, dass der Preis einen Einfluss auf die Überlebenswahrscheinlichkeit hat. Dies ist ein sehr starkes Ergebnis und lässt darauf schließen, dass wohlhabendere Passagiere eine höhere Überlebenswahrscheinlichkeit hatten als ärmere Passagiere. Die exakte Effektgröße ist nun jedoch etwas komplizierter zu interpretieren, da wir hier nicht mehr mit einer linearen Regression arbeiten.

Aufgabe 1e)

Die erste Möglichkeit, den Wert von _1 zu interpretieren, sind die Odds. Diese berechnen wir folgendermaßen:

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

```
  Preis
1.01519
```

Die Odds zu Überleben erhöhen sich also um den Faktor 1,015 für einen um einen Dollar erhöhten Preis (und z.B. um den Faktor $e^{1 \cdot 100} = 4{,}52$ für einen um 100 Dollar höheren Preis). Ein Nachteil dieser Interpretation ist die fehlende Intuition für Odds.

Aufgabe 1f)

Nun berechnen wir, wie hoch die Überlebenswahrscheinlichkeit einer Person gemäß unserers Modells war, welche 100 Dollar für ihr Ticket gezahlt hat.

$$\pi = \frac{e^{-0{,}931520+0{,}015076*100}}{1+e^{-0{,}931520+0{,}015076*100}} \approx 0,64$$

Oder, einfach in R:

```
# Klassisch mit den bekannten Werten:
exp(-0.931520 + 0.015076 * 100)/(1 + exp(-0.931520 + 0.015076 * 100))
```

```
[1] 0.6401649
```

```
# Etwas dynamischer mit Hilfe des Modells:
exp(coef(mod)[1] + coef(mod)[2] * 100) / (1 + exp(coef(mod)[1] + coef(mod)[2] * 100))
```

```
(Intercept)
  0.6401557
```

```
# Oder mit Hilfe von 'plogis()':
plogis(-0.931520 + 0.015076 * 100)
```

```
[1] 0.6401649
```

```
# Oder mit 'predict()' (keine Ungenauigkeiten durch Rundung):
predict(mod, newdata = data.frame(Preis = 100), type = "response")
```

```
        1
0.6401557
```

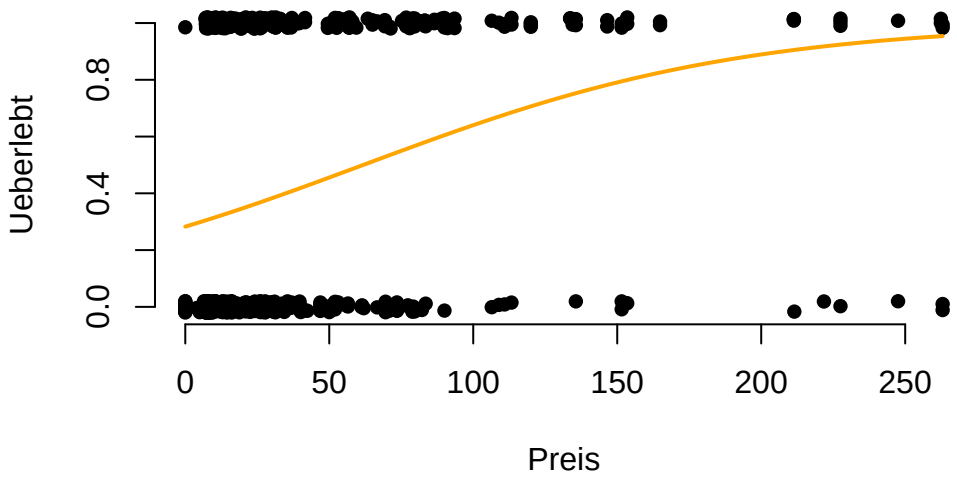
```
  # funktioniert auch mit data.table statt data.frame
```

Folgen wir unserem Modell, so sehen wir, dass eine Person, welche 100 Dollar für ihr Ticket gezahlt hat, eine Überlebenswahrscheinlichkeit von circa 0,64, also von 64% hatte.

Aufgabe 1g)

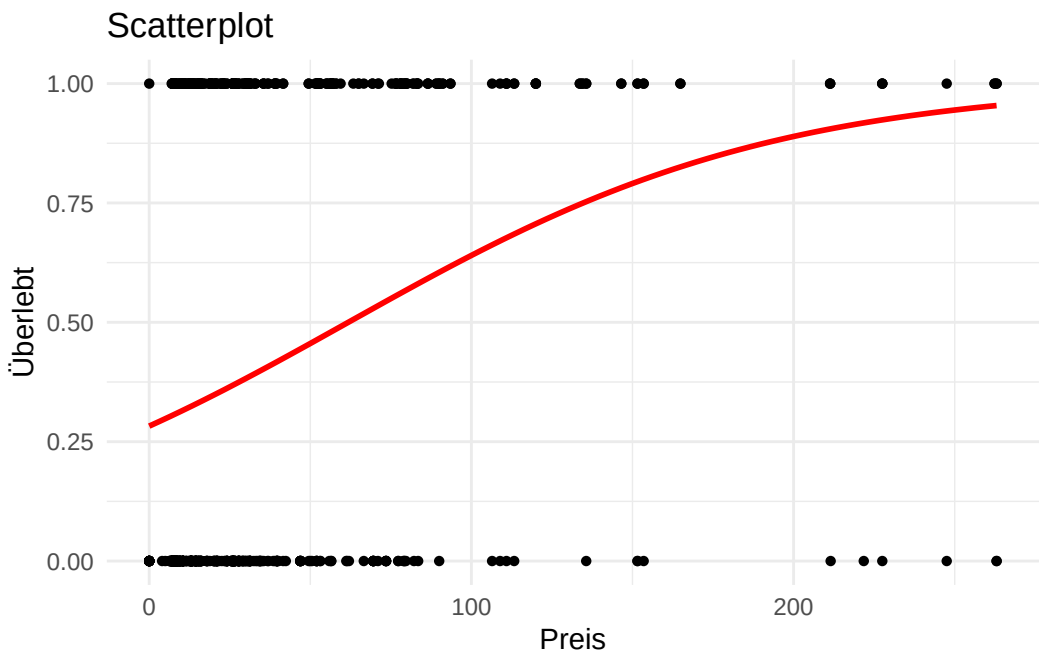
Jetzt möchten wir unser Regressionsergebnis visualisieren, und mit unserem Jitter-Plot von vorher kombinieren.

```
# Mit normalen R Befehlen:
plot(titanic_data$Preis, jitter(titanic_data$Ueberlebt, 0.1),
     pch = 16, bty = "n", ylab = "Ueberlebt", xlab = "Preis")
curve(plogis(mod$coeff[1] + mod$coeff[2] * x),
      lwd = 2, col = "orange", add = TRUE)
```



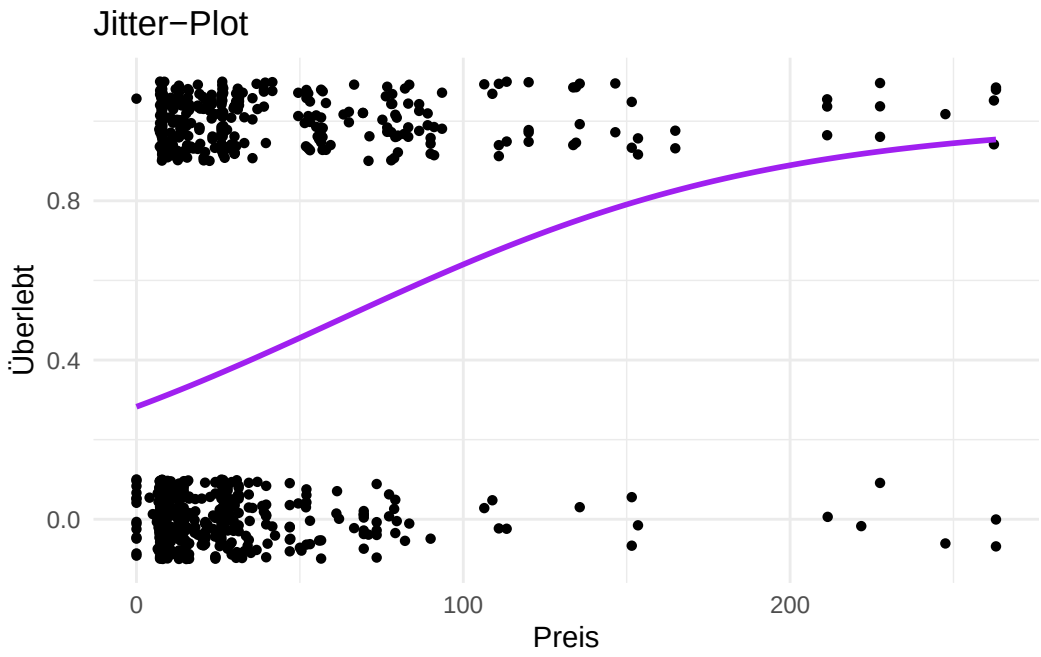

```
# Mit ggplot2:
gg_scatterplot +
  geom_smooth(method = "glm", method.args = list(family = "binomial"),
              se = FALSE, color = "red")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
gg_jitterplot +
  geom_smooth(method = "glm", method.args = list(family = "binomial"),
              se = FALSE, color = "purple")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Wir sehen, dass die farbige Linie die Überlebenswahrscheinlichkeit in Abhängigkeit vom Preis darstellt. Diese Linie ist also die logistische Regression, welche wir zuvor geschätzt haben. Wir sehen, dass die Überlebenswahrscheinlichkeit mit steigendem Preis zunimmt, was auch unsere Schätzung des Modells bestätigt. Bei einem Preis von circa 60 Dollar scheint die Überlebenswahrscheinlichkeit ungefähr 50:50 gewesen zu sein.

Bei den Plots, welche mit ggplot2 erstellt wurden, sehen wir nun auch, die modulare Struktur von ggplot. Wenn wir weitere Effekte/Darstellungen wünschen, können wir sie einfach zu unseren bisherigen Plots hinzufügen. Sehr praktisch!

Aufgabe 1h)

Nun möchten wir noch ein multiples logistisches Regressionsmodell schätzen, um zusätzlich den Einfluss des Geschlechts und des Alters auf die Überlebenswahrscheinlichkeit zu untersuchen.

```
# Wir nenenn dieses neue, erweiterte Modell "mod2"
mod2 <- glm(Ueberlebt ~ Preis + Alter + I(Alter^2) + Geschlecht,
            family = "binomial", data = titanic_data)
summary(mod2)
```

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

```
    family = "binomial", data = titanic_data)

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
```

'I(Alter^2)' wird hier verwendet, um einfach den Wert von Alter zu quadrieren. Wird das ganze nicht mit I() verpackt, so versteht R den Befehl anders, und wird versuchen, alle Interaktionen bis zur zweiten Ordnung zu berechnen. Hier interessiert uns aber explizit der quadrierte Wert des Alters, als penalty-term ("Strafterm") um sehr hohes Alter und sehr niedriges Alter effektiv zu differenzieren.

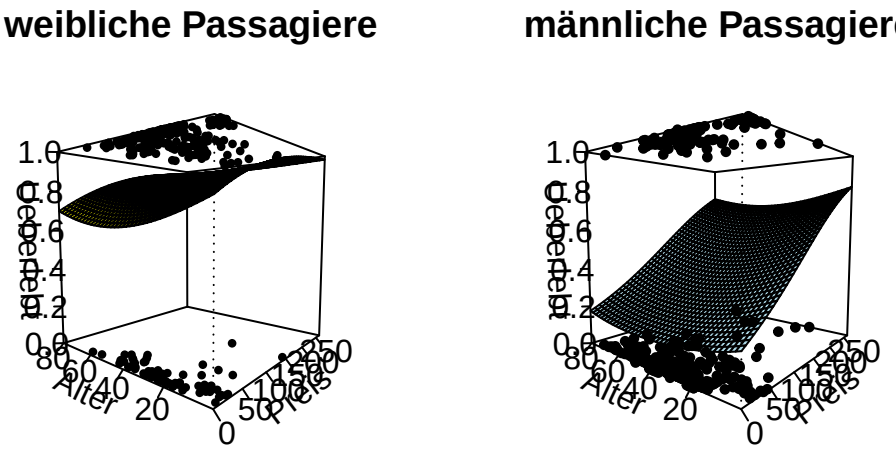
Unser Summary liefert uns nun die Schätzungen für alle Variablen, und wir sehen, dass ein höherer Preis einen positiven Einfluss auf die Überlebenswahrscheinlichkeit hat (wie wir bereits wissen), und sehen nun, dass männliche Passagiere eine deutlich geringere Überlebenswahrscheinkichkeit gehabt haben. Alter scheint ebenfalls einen leicht negativen Effekt haben, was "Kinder, Alte und Frauen Zuerst" zu bestätigen scheint, jedoch wird dieser Effekt als nicht signifikant eingestuft. Der Ticketpreis und das Geschlecht sind hier die wichtigen Prädiktoren!

Aufgabe 1i)

Wir visualisieren diese Ergebnisse nun mit Hilfe des zur Verfügung gestellten Codes. Wichtiger Hinweis: Im Code vom Aufgabenblatt wird das verwendete Modell "mod" genannt, hier verwenden wir jedoch "mod2", da wir das zweite Modell verwenden möchten, wo wir nicht nur den Einfluss von Preis, sondern auch von Alter und Geschlecht geschätzt haben. Habt ihr diese Schätzung anders benannt als ich (bei mir heißt sie mod2), dann müsst ihr das im folgenden Code auch ersetzen!

```r
# Wir betrachten die Spannweite von Preis und Alter für unsere Schätzung und Visualisierung
x1 <- seq(min(titanic_data$Preis), max(titanic_data$Preis), length = 40)
x2 <- seq(min(titanic_data$Alter), max(titanic_data$Alter), length = 40)

# Wir bauen einen Schätzer für die weiblichen Passagiere (ohne den Faktor für Männer)
predictor1 <- function(x1, x2){
  plogis(mod2$coeff[1] + mod2$coeff[2] * x1 + mod2$coeff[3] * x2 +
         mod2$coeff[4] * x2^2)
  }

# Wir bauen einen Schätzer für die männlichen Passagiere (mit dem Faktor für Männer)
predictor2 <- function(x1, x2){
  plogis(mod2$coeff[1] + mod2$coeff[2] * x1 + mod2$coeff[3] * x2 +
         mod2$coeff[4] * x2^2 + mod2$coeff[5])
  }

# Wir berechnen Werte und schauen, welche Passagiere weiblich waren (als Index)
z1 <- outer(x1, x2, predictor1)
z2 <- outer(x1, x2, predictor2)
ind <- which(titanic_data$Geschlecht == "female")

# Wir bereiten ein Fenster für beide Plots nebeneinander vor (1 Zeile, 2 Spalten)
par(mfrow = c(1, 2))

# Wir erstellen den Plot mit einer Schätzung für weibliche Passagiere
res <- persp(x1, x2, z1, theta = -50, phi = 5,
             xlab = "Preis", ylab = "Alter", zlab = "Ueberlebt",
             zlim = c(0,1), ticktype = "detailed",
             col="yellow", main = "weibliche Passagiere")

# Wir ergänzen den Plot mit den Datenpunkten der Frauen
points(trans3d(titanic_data$Preis[ind], titanic_data$Alter[ind],
               titanic_data$Ueberlebt[ind], pmat = res),
               col = "black", pch = 16, cex = 0.6)

# Wir erstellen den Plot mit einer Schätzung für männliche Passagiere
res <- persp(x1, x2, z2, theta = -50, phi = 5,
             xlab = "Preis", ylab = "Alter", zlab = "Ueberlebt",
             zlim = c(0,1), ticktype = "detailed",
             col = "lightblue", main = "männliche Passagiere")

# Wir ergänzen den Plot mit den Datenpunkten der Männer
```

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



Wir können sehen, dass weibliche Passagiere, ungeachtet aller anderen Variablen, eine relativ hohe Überlebenschance hatten. Insbesondere der Preis zeigt einen zusätzlich positiven Effekt, während das Alter für mittlere Jahrgänge leicht negativ ins Gewicht fällt. Männer hingegen haben eine deutlich geringere Überlebenchance, und zwar für alle Gruppen und Variablen. Das Alter zeigt einen ähnlichen, kaum ins Gewicht fallenden Verlauf, während der Preis für Männer umso mehr ins Gewicht zu fallen scheint. Eventuell kann es interessant sein, hier noch einen Interaktionseffekt zwischen Geschlecht und Preis zu beleuchten.