set.seed(42)

Aufgabe 1

a)

```
library(MASS)
sig = 6.8
R = 0.73
tgeld = c(5.1,15.6,28.2,11.1,4.0,31.5,19.5)
y = tgeld
time = c(18,19.5,20,20.5,21.25,21.5,22)
gender = c(0,1,1,0,0,1,1)
eins = rep(1,7)

X = matrix(
  c(eins, time, gender),
  nrow = 7,
  ncol = 3,
  byrow = FALSE
)

XT = t(X)

XTX = XT %*% X

iXTX = ginv(XTX)

b = iXTX %*% XT %*% y
b = round(b, 2)

model <- lm(y~time+gender)
model$coefficients
#> (Intercept)        time      gender
#> -11.8775934   0.9344398  16.1879668
```

b)

Wir erhalten dann eine Schätzung der Regressionskoeffizienten mittels:

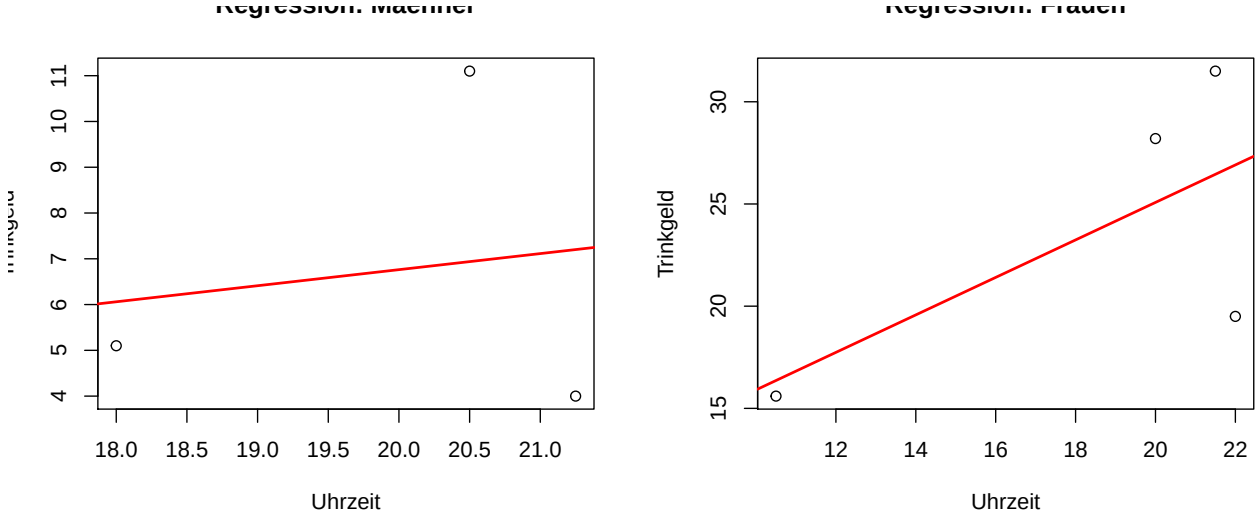
$$\hat{b} = (X^T X)^{-1} X^T y \approx (-11.88, 0.93, 16.19)$$

Sowohl Uhrzeit als auch Geschlecht (Frauen) haben einen positiven Einfluss auf die abhängige Variable Trinkgeld. Dabei ist das Geschlecht deutlich größer gewichtet als die Uhrzeit.

c)

```
y_m <- c(5.1, 11.1, 4)
X_m <- matrix(c(1, 1, 1, 18, 20.5, 21.25), nrow = 3, ncol = 2)
XTX_m <- solve(t(X_m)%*%X_m)
b_hat_m <- XTX_m%*%t(X_m)%*%y_m
```

```
y_w <- c(15.6, 28.2, 31.5, 19.5)
X_w <- matrix(c(1, 1, 1, 1, 10.5, 20, 21.5, 22), nrow = 4, ncol = 2)
XTX_w <- solve(t(X_w)%*%X_w)
b_hat_w <- XTX_w%*%t(X_w)%*%y_w

par(mfrow=c(1,2))

plot(c(18, 20.5, 21.25), y_m, xlab = 'Uhrzeit', ylab = 'Trinkgeld', main = 'Regression: Maenner')
abline(b_hat_m, lw=2, col='red')

plot(c(10.5, 20, 21.5, 22), y_w, xlab = 'Uhrzeit', ylab = 'Trinkgeld', main = 'Regression: Frauen')
abline(b_hat_w, lw=2, col='red')
```




```
f = t(b) %*% c(0,1,1)
m = t(b) %*% c(0,1,0)
```

Die Änderung des zu erwartenden Trinkgelds pro Stunde ist die Steigung der Regression, das heißt:

Trinkgeld pro Stunde Männer: $y = \hat{b}_1 + 1 \cdot \hat{b}_2 = 0.93$

Trinkgeld pro Stunde Frauen: $y = \hat{b}_1 + 1 \cdot \hat{b}_2 = 17.12$

d)

H0: “Es besteht kein signfikanter Unterschied der Höhe des Trinkgelds unter den Geschlechtern.” H1: “Es besteht ein signfikanter Unterschied der Höhe des Trinkgelds unter den Geschlechtern.”

```
n <- 7
p <- 3
beta <- b[3]

test <- beta/(sig*sqrt(XTX[3,3]))

t_value <-  qt(c(0.025, 0.975), df=n-p)

if (t_value[1] < test & test < t_value[2]){
  print('H0 wird nicht verworfen!')
  } else {
    print('H0 wird verworfen!')
```

```
  }
#> [1] "H0 wird nicht verworfen!"
```

Wir erhalten einen Testwert von 1.19 und dieser liegt nicht im Ablehnungsbereich -2.78, 2.78. Folglich lehnen wir die Nullhypothese nicht ab.

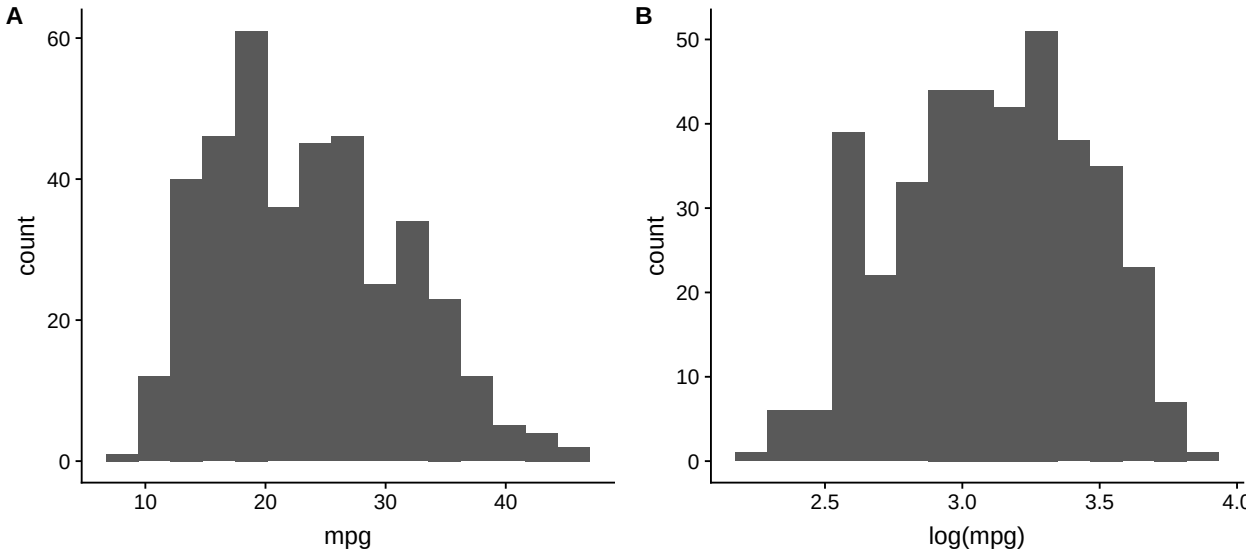
e)

Der SSE ist die Quadratsumme der Residuen und damit unmittelbar abhängig von der Skalierung der Zielwerte, deswegen kann man daraus nicht folgern, dass die absolute Trinkgeldhöhe schlechter erklärt wird, als die prozentuale. Als Beispiel kann man die Werte hier betrachten, diese liegen im unteren 2-stelligen Bereicht. Nimmt man nun das absolute Trinkgeld in Yen, so befinden sich die Werte im höheren 3-stelligen oder sogar im 4-stelligen Bereich folglich wären bei gleicher Modellgüte die Beträge der Residuen deutlich größer.

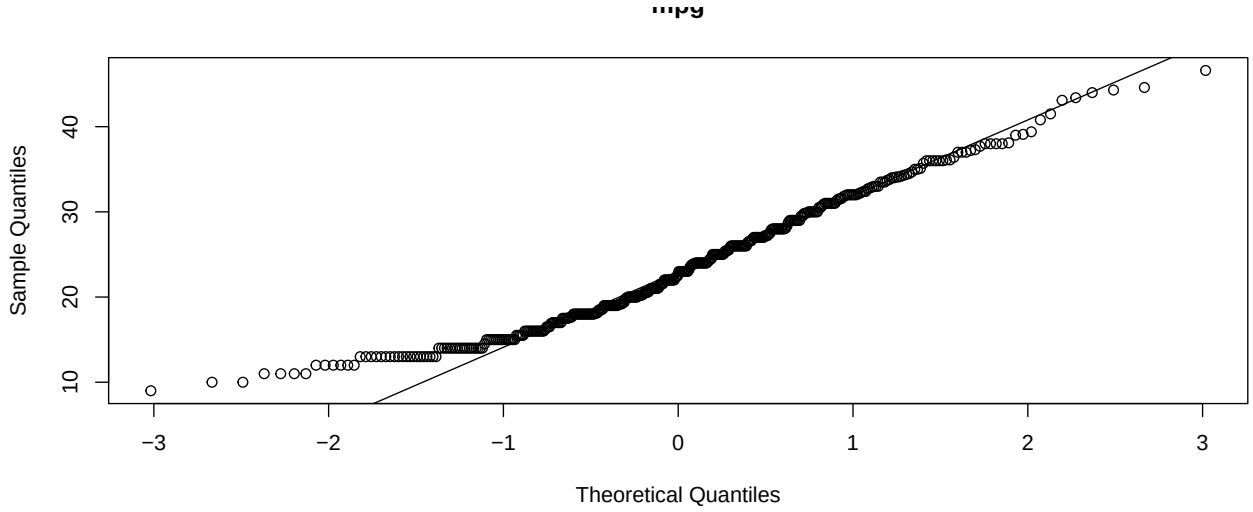
Aufgabe 2

a)

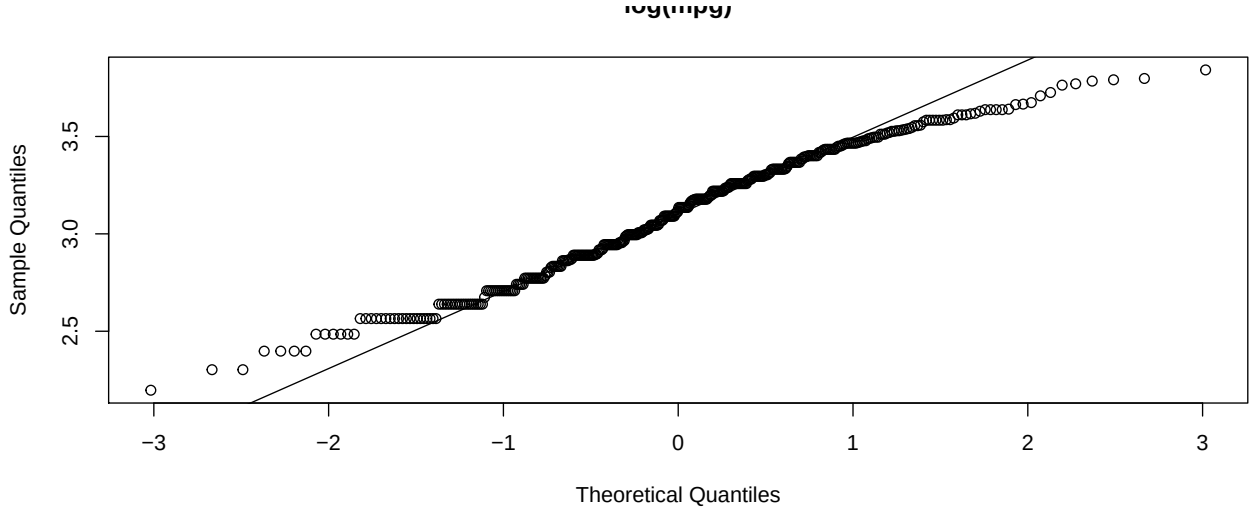
```
library(ISLR)
library(ggplot2)
library(cowplot)
par(mfrow=c(1,2))

mpg_hist <- ggplot(Auto, aes(x=mpg)) + geom_histogram(bins = 15)
log_mpg_hist <- ggplot(Auto, aes(x=log(mpg))) + geom_histogram(bins = 15)
plot_grid(mpg_hist, log_mpg_hist, labels = "AUTO")
```




```
qqnorm(Auto$mpg, main='mpg')
qqline(Auto$mpg)
```



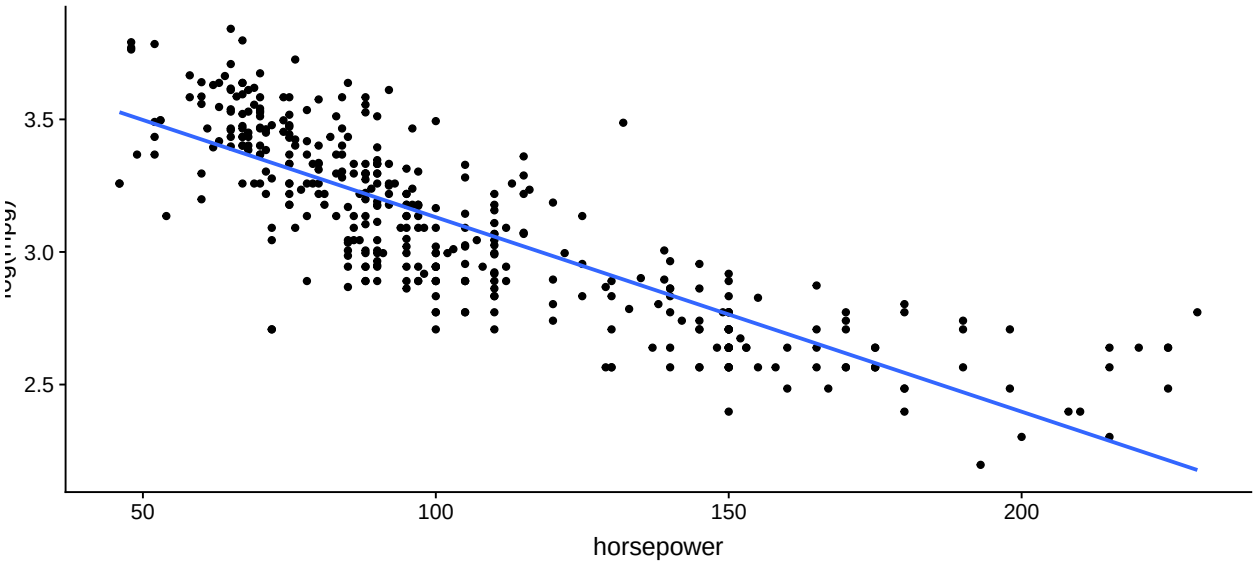

```
qqnorm(log(Auto$mpg), main='log(mpg)')
qqline(log(Auto$mpg))
```



Laut Histogramm gleicht mpg einer rechtsschiefen Verteilung, log(mpg) kommt laut Histogramm einer Normalverteilung näher. Die QQ-Plots sind schwieriger auszuwerten, beide Variablen weichen an den Rändern deutlich von der Geraden ab. Allerdings weichen die Punkte von log(mpg) an den Rändern "symmetrischer" von der Geraden ab, als bei mpg.

b)

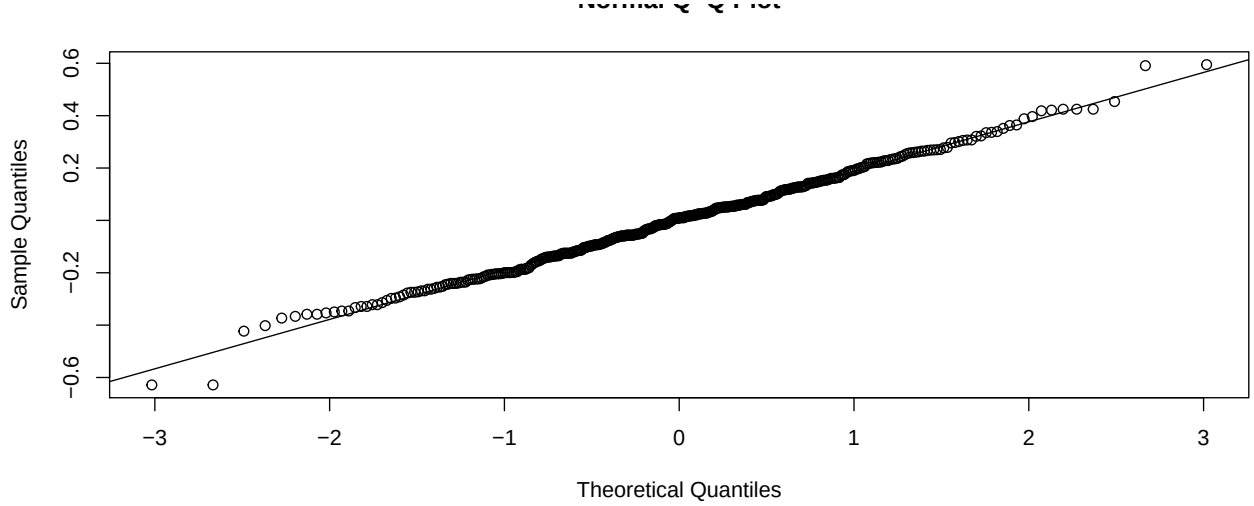
```
X <- matrix(NA, nrow = nrow(Auto), ncol = 2)
X[,1] <- rep(1, nrow(Auto))
X[,2] <- Auto$horsepower
XTX <- solve(t(X)%*%X)
b_hat <- XTX%*%t(X)%*%log(Auto$mpg)
b_hat[2,]
#> [1] -0.007333764

ggplot(Auto, aes(x=horsepower, y=log(mpg))) + geom_point() + geom_smooth(method = 'lm', se = FALSE)
```



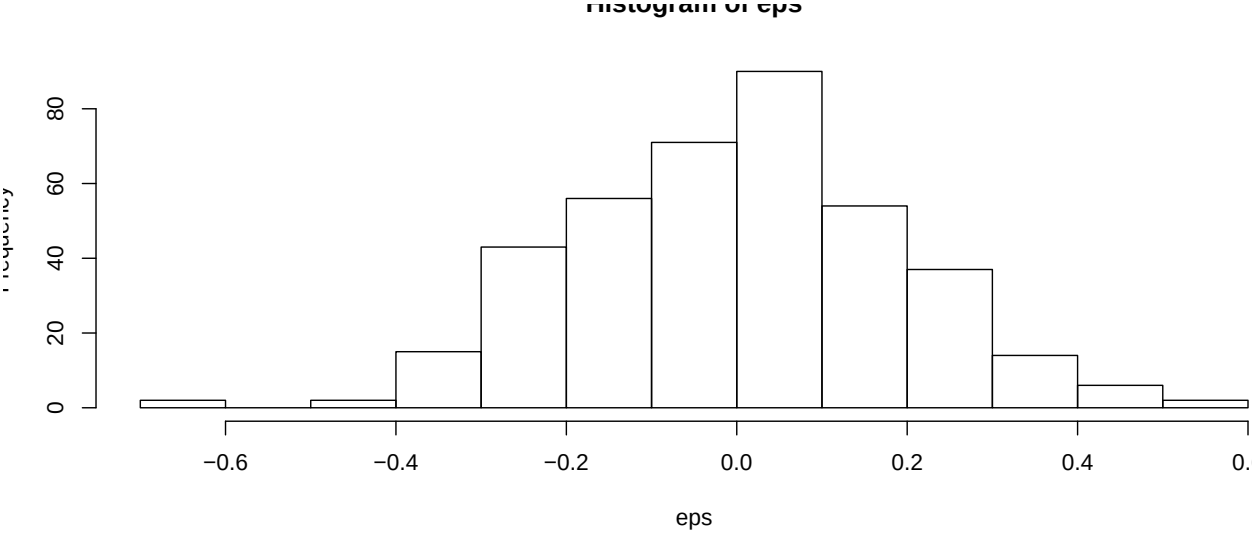
Auf den ersten Blick deutet ein beta von -0.007 auf keinen Zusammenhang zwischen Prädiktor und Zielgröße hin. Plottet man sich Prädiktor und Zielgröße allerdings, ist ein klarer negativer linearer Zusammenhang erkennbar. Da horsepower und log(mpg) unterschiedlich skaliert sind, ist ein Ablesen des Zusammenhangs anhand beta1 irreführend. horsepower müsste ebenfalls logarithmiert werden.

```
X_log <- matrix(NA, nrow = nrow(Auto), ncol = 2)
X_log[,1] <- rep(1, nrow(Auto))
X_log[,2] <- log(Auto$horsepower)
XTX_log <- solve(t(X_log)%*%X_log)
b_hat_log <- XTX_log%*%t(X_log)%*%log(Auto$mpg)
b_hat_log[2,]
#> [1] -0.841847

n <- nrow(Auto)
p <- 3
beta <- b_hat[2,]
sig_hat <- sd(log(Auto$mpg))

test <- beta/(sig_hat*sqrt(XTX[2,2]))
test
#> [1] -16.41525

t_value <-  qt(c(0.025, 0.975), df=n-p)

if (t_value[1] < test & test < t_value[2]){
  print('H0 wird nicht verworfen!')
} else {
  print('H0 wird verworfen!')
}
#> [1] "H0 wird verworfen!"

mpg98 <- exp(b_hat[1,] + b_hat[2,]*98)
mpg98
#> [1] 23.23728

diff_20 <- abs(20*b_hat[2,])
```

```
diff_20
#> [1] 0.1466753

y_hat <- c()
for(i in 1:nrow(Auto)){
  y_hat[i] <- (b_hat[1,] + b_hat[2,]*Auto$horsepower[i])
}

eps <- log(Auto$mpg) - y_hat
qqnorm(eps)
qqline(eps)
```




```
hist(eps)
```




```
shapiro.test(eps)
#>
#>  Shapiro-Wilk normality test
#>
```

```
#> data:  eps
#> W = 0.99607, p-value = 0.4421
```

Laut Shapiro-Wilk-Test wird H0 (“Die Residuen sind normalverteilt.”) nicht verworfen. Wir können von einer Normalverteilung ausgehen. Die Plots bestätigen unsere Vermutung.

c)

TO DO! Max pls :)