Übung 1

Computational Statistics

Sommersemester 2019

April 4, 2019

J. Groos (FBMN, h_da)

Name:

Übung 1: Multiple lineare Regression

1. Führe eine lineare Regression auf dem Datensatz Donald durch

x1: Geschlecht (0 = w, 1 = m)

x2: Alter

x3: Minderheit (0 = Nein, 1 = Ja)

x4: Fremdenfeindlich (0-11)

x5: IQ

y: Zustimmung

(Funktion: lm)

2. Vergleiche den Einfluss der Variablen -> Standardisierte Parameter (Funktion: lm.beta (Paket QuantPsyc))
3. Bestimme die KI für die Parameter
4. Bestimme den VIF-Wert (Multikollinearität) (Funktion. vif (Paket car))
5. Überprüfe die Verteilung der Residuen und Linearität (Funktionen: plot, list, residualPlot, avPlots, . . .)
6. Prognose für sich selbst

1.

```
# Lade Daten
load(file='Donald.RData')
data <- Donald_1

# Fitte die Regression
fit <- lm(Trump ~ Alter + Geschlecht + Minderheit + Fremdenfeindlich + IQ, data = data)

# Ergebnis
summary(fit)
#>
#> Call:
#> lm(formula = Trump ~ Alter + Geschlecht + Minderheit + Fremdenfeindlich +
#>     IQ, data = data)
#>
#> Residuals:
```

```
#>      Min       1Q   Median       3Q      Max
#> -15.6835  -4.5286  -0.0023   4.1231  13.0974
#>
#> Coefficients:
#>                  Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      29.73834    3.60973   8.238 9.73e-14 ***
#> Alter             0.18153    0.03049   5.954 1.91e-08 ***
#> Geschlecht        5.75572    0.99977   5.757 4.97e-08 ***
#> Minderheit       -6.57586    1.82843  -3.596 0.000443 ***
#> Fremdenfeindlich  9.34984    0.16325  57.272  < 2e-16 ***
#> IQ               -0.40135    0.03016 -13.309  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.065 on 144 degrees of freedom
#> Multiple R-squared:  0.9713, Adjusted R-squared:  0.9703
#> F-statistic: 973.6 on 5 and 144 DF,  p-value: < 2.2e-16
```

2.

```
# installieren des QuantPsyc package
packageTest('QuantPsyc')

# standardisiere die Parameter des Regressionsmodells
fit.beta <- lm.beta(fit)
print(fit.beta)
#>            Alter       Geschlecht       Minderheit Fremdenfeindlich
#>       0.08513441       0.08190695      -0.05790182       0.91808061
#>               IQ
#>      -0.19120337
```

Der Parameter “Fremdenfeindlich” hat den größten Effekt auf die abhängige Variable. Je höher der Parameterwert, desto größer die Zustimmung zu Trump (in %). Der Parameter “IQ” hat einen moderaten negativen Effekt auf die Ausprägung der abhängigen Variable. Je höher der Parameterwert, desto geringer die Zustimmung zu Trump (in %). DIe Parameter “Geschlecht”, “Minderheit” und “Alter” haben jeweils einen geringen Effekt auf die Ausprägung der abhängigen Variable.

3.

```
KI <- confint(object = fit, level = 0.95)
KI
#>                      2.5 %     97.5 %
#> (Intercept)       22.6034487 36.8732397
#> Alter              0.1212600  0.2417933
#> Geschlecht         3.7795957  7.7318438
#> Minderheit       -10.1898944 -2.9618306
#> Fremdenfeindlich   9.0271567  9.6725195
#> IQ                -0.4609559 -0.3417435
```

In der Ausgabe sehen wir die Intervallgrenzen für das 95%-Konfidenzintervall.

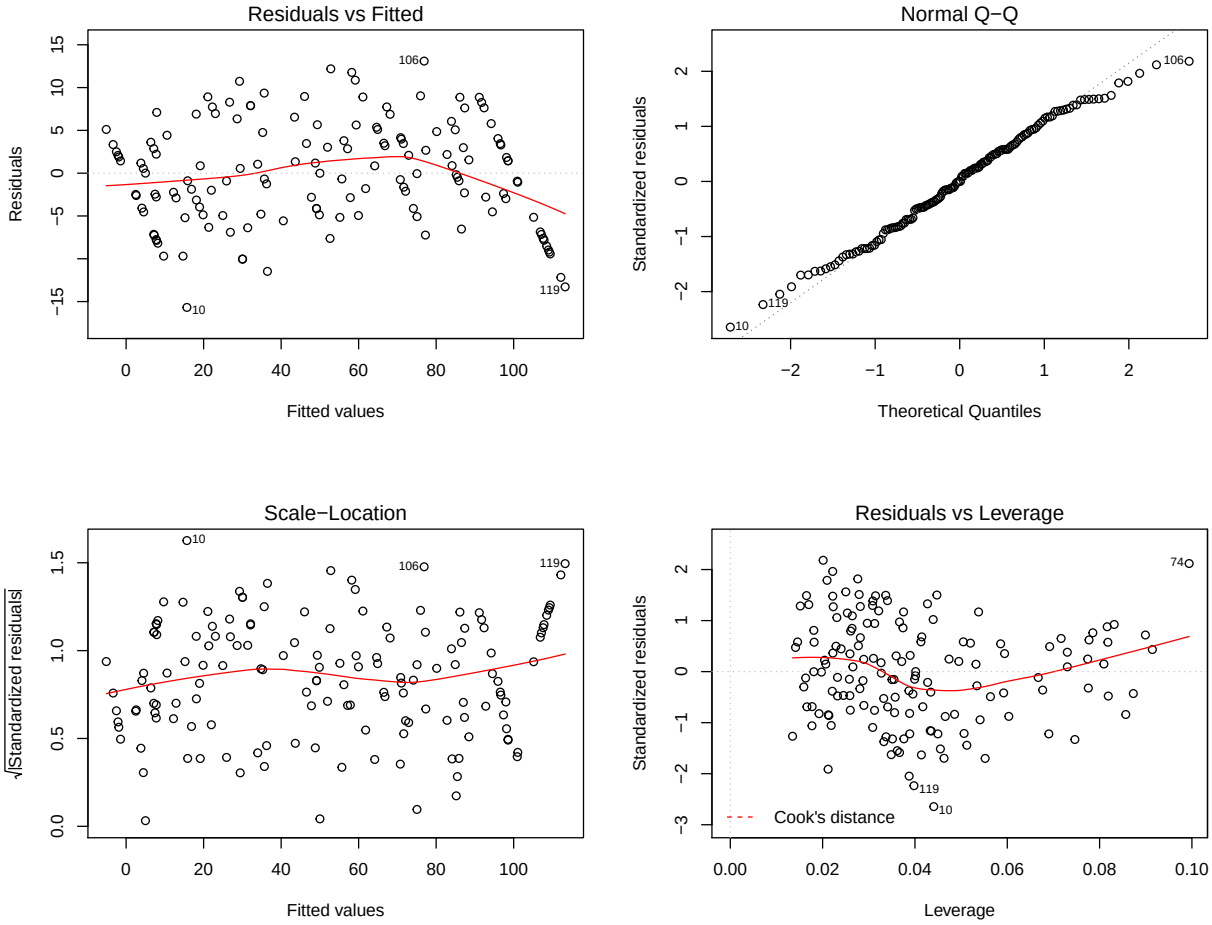
4.

```
packageTest('car')

vif_fit <- vif(mod = fit)
vif_fit
#>            Alter       Geschlecht       Minderheit Fremdenfeindlich
#>         1.024821         1.014460         1.299054         1.287851
#>               IQ
#>         1.034409
```

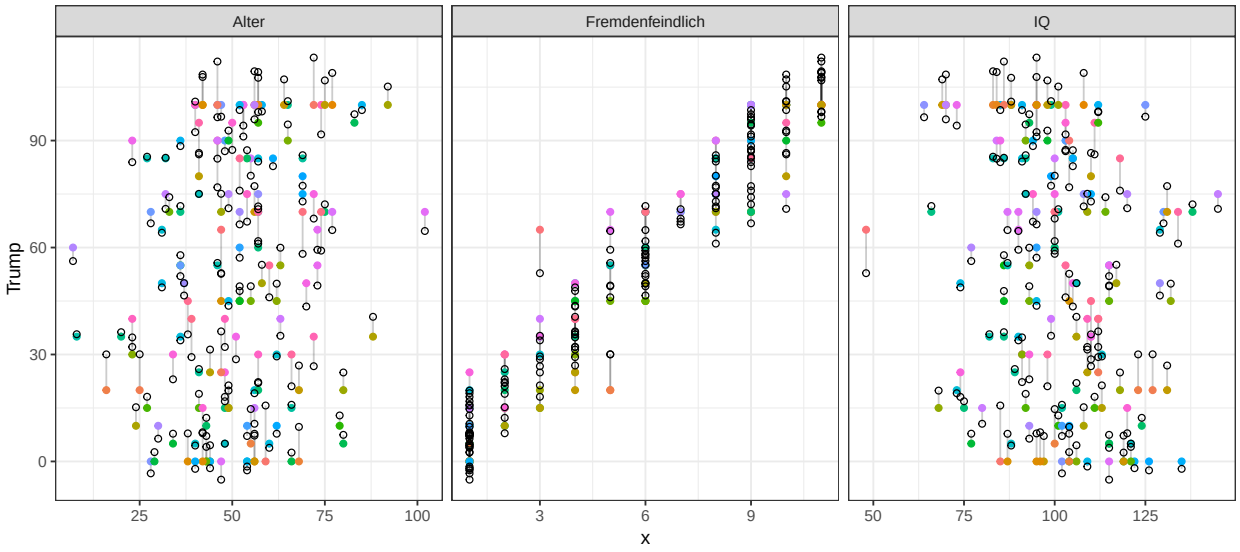
Die VIF-Werte liegen alle deutlich unter 5(10), es liegen also keine Hinweise für Multikolliniarität zwischen den Modellparametern vor.

5.

```
packageTest('magrittr')
packageTest('tidyr')
packageTest('ggplot2')

par(mfrow = c(2, 2))
plot(fit)
```



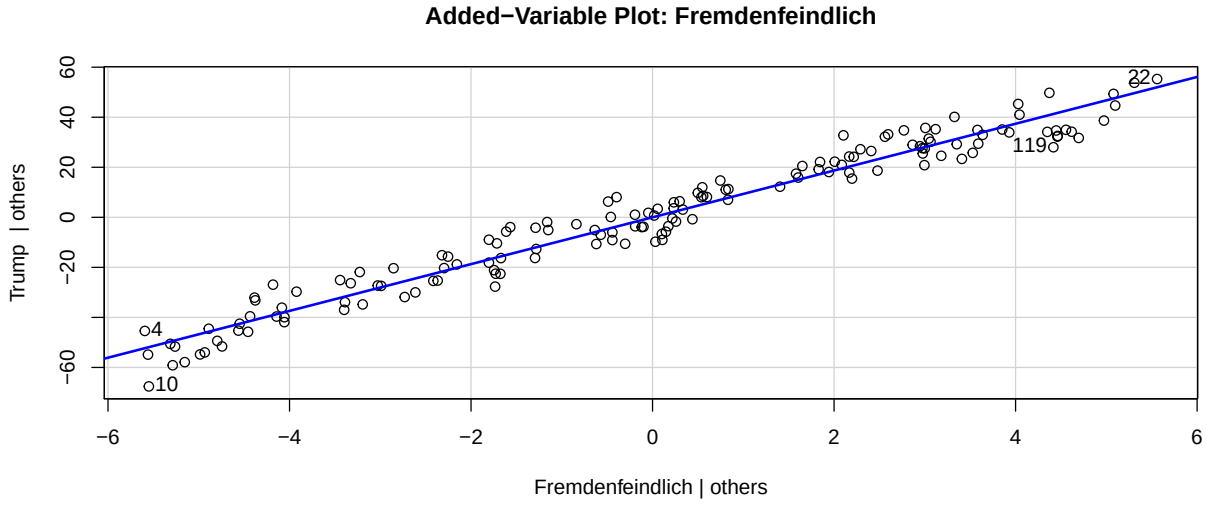

```
data$predicted <- predict(fit)
data$residuals <- residuals(fit)

data %>%
  gather(key = "iv", value = "x", -Trump, -predicted, -residuals, -Minderheit, -Geschlecht) %>%
  ggplot(aes(x = x, y = Trump)) +
  geom_segment(aes(xend = x, yend = predicted), alpha = .2) +
  geom_point(aes(color = factor(residuals))) +
  guides(color = FALSE) +
  geom_point(aes(y = predicted), shape = 1) +
  facet_grid(~ iv, scales = 'free_x') +
  theme_bw()
```



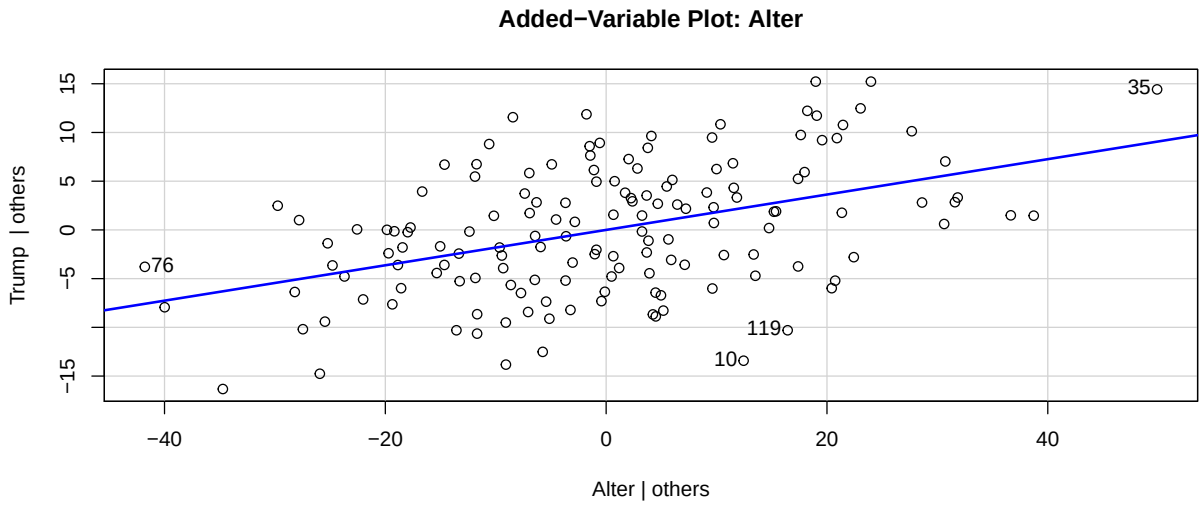
Die Residuenanalyse ergibt, dass die Residuen näherungsweise normalverteilt sind und es keine klar erkennbaren Muster in den Residuenplots gibt.

Zudem wurden die Ausprägungen der (nicht-binären) unabhängigen Parameter den Ausprägungen der abhängigen Variablen durch Scatterplots gegenübergestellt. Lediglich der Parameter "Fremdenfeindlich" weist einen klar erkennbaren linearen Zusammenhang zur abhängigen Variabel auf. Dies bestätigt das Ergebnis aus 2.

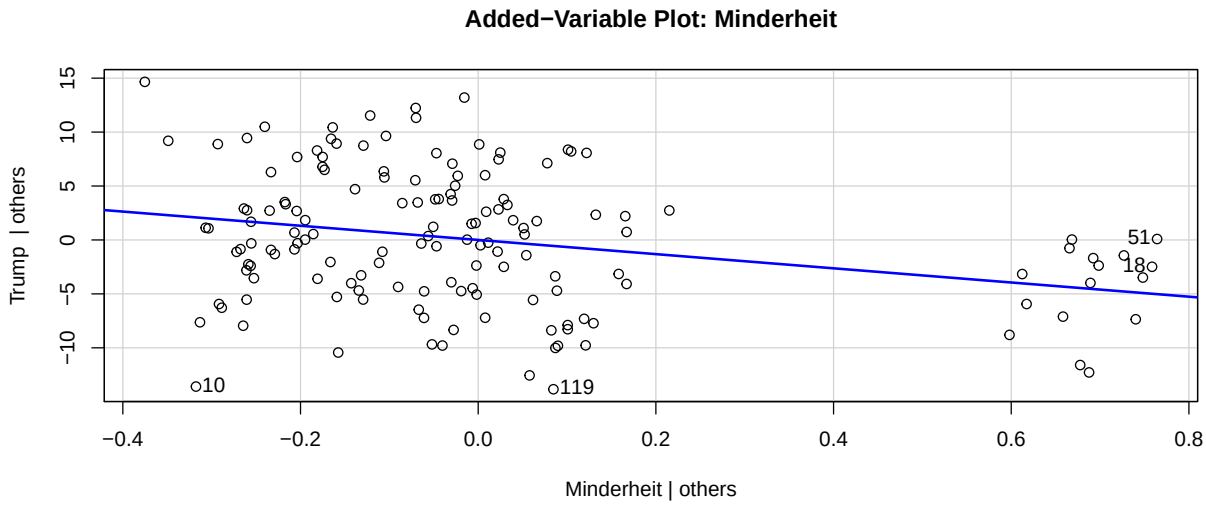
```
avPlot(fit, "Fremdenfeindlich")
```



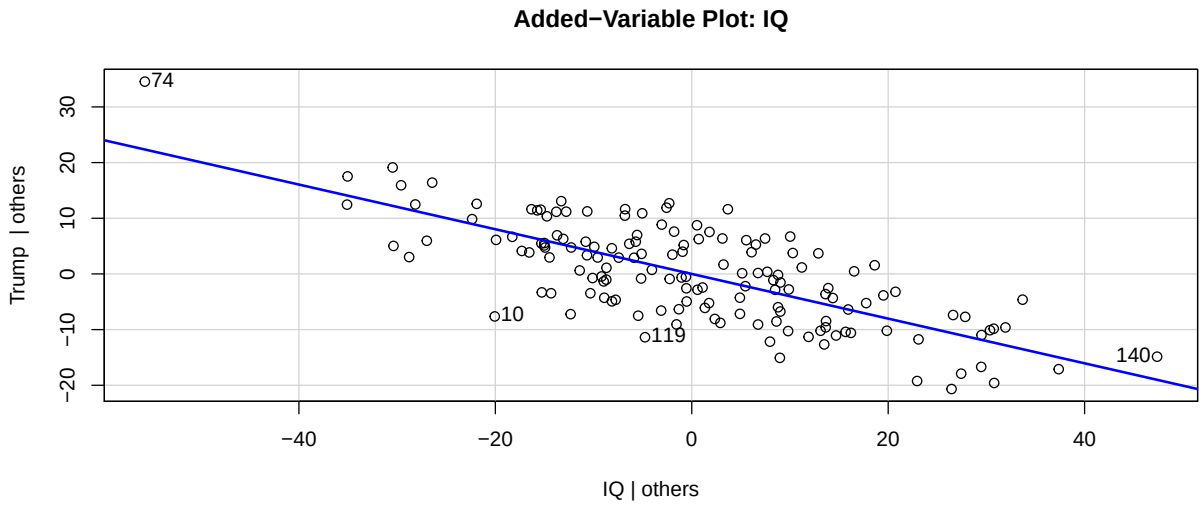

```
avPlot(fit, "Alter")
```



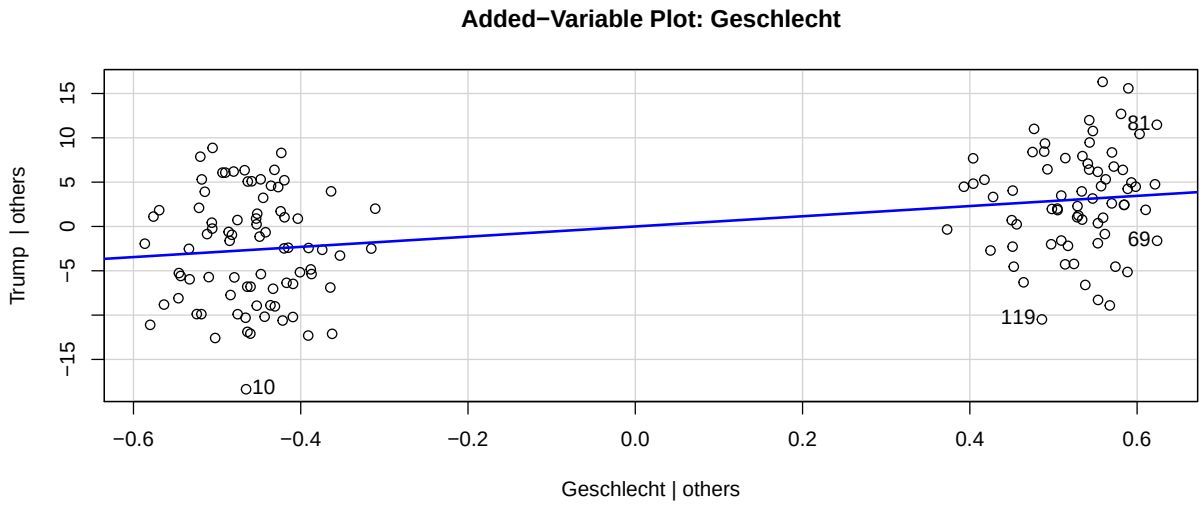

```
avPlot(fit, "Minderheit")
```




```
avPlot(fit, "IQ")
```




```
avPlot(fit, "Geschlecht")
```



An den Added-Variable Plot wird nochmals deutlich, dass “Fremdenfeindlich”, “IQ” und “Alter” einen linearen Trend aufweisen. Die beiden binären Variablen “Minderheit” und “Geschlecht” kann man nicht linear erklären.

6.

```
my_data <- data.frame("Geschlecht" = 1, "Alter" = 24,
                      "Minderheit" = 0, "Fremdenfeindlich" = 5, "IQ" = 100)
my_data$predicted <- predict(fit, newdata =  my_data)


my_data2 <- data.frame("Geschlecht" = 1, "Alter" = 27,
                      "Minderheit" = 1, "Fremdenfeindlich" = 3, "IQ" = 100)
my_data2$predicted <- predict(fit, newdata =  my_data2)
```

```r
my_data3 <- data.frame("Geschlecht" = 1, "Alter" = 29,
                       "Minderheit" = 0, "Fremdenfeindlich" = 8, "IQ" = 90)
my_data3$predicted <- predict(fit, newdata =  my_data3)

pred_sum <- rbind(my_data, my_data2, my_data3)
pred_sum
#>   Geschlecht Alter Minderheit Fremdenfeindlich  IQ predicted
#> 1          1    24          0                5 100  46.46492
#> 2          1    27          1                3 100  21.73396
#> 3          1    29          0                8  90  79.43557
```

Das Modell prognostiziert uns unterschiedliche Ergebnisse. Man erkennt klar, dass “Fremdenfeindlich” die größte Auswirkung auf die Zustimmungsrate hat.