Skript zur Vorlesung 5: Das lineare Regressionsmodell

Luis Erhardt

2025

Contents

1 Einleitung

In diesem Dokument werden die Abbildungen aus der fünften Vorlesung repliziert.

Folgende Pakete wurden verwendet:

```
library(tidyverse)
library(rmarkdown)
library(testthat)
library(ggplot2)
library(knitr)
library(data.table)
library(ggpubr)
library(car)
```

2 Einfache Regression zum Zusammenhang zwischen Gewerkschaften und Innovationstätigkeit

Die Daten werden eingelesen und zwei Spalten angepasst.

```
oecd_data <- fread(here::here("data/T5/oecd_data.csv"))

oecd_data <- dplyr::mutate(oecd_data, "GDPpc"=GDP/POP*1000000)

oecd_data <- dplyr::mutate(oecd_data, "LME"=dplyr::if_else(Country %in% c("AUS","USA"),1,0))
```

Nun formulieren wir die Regression und geben deren Werte aus:

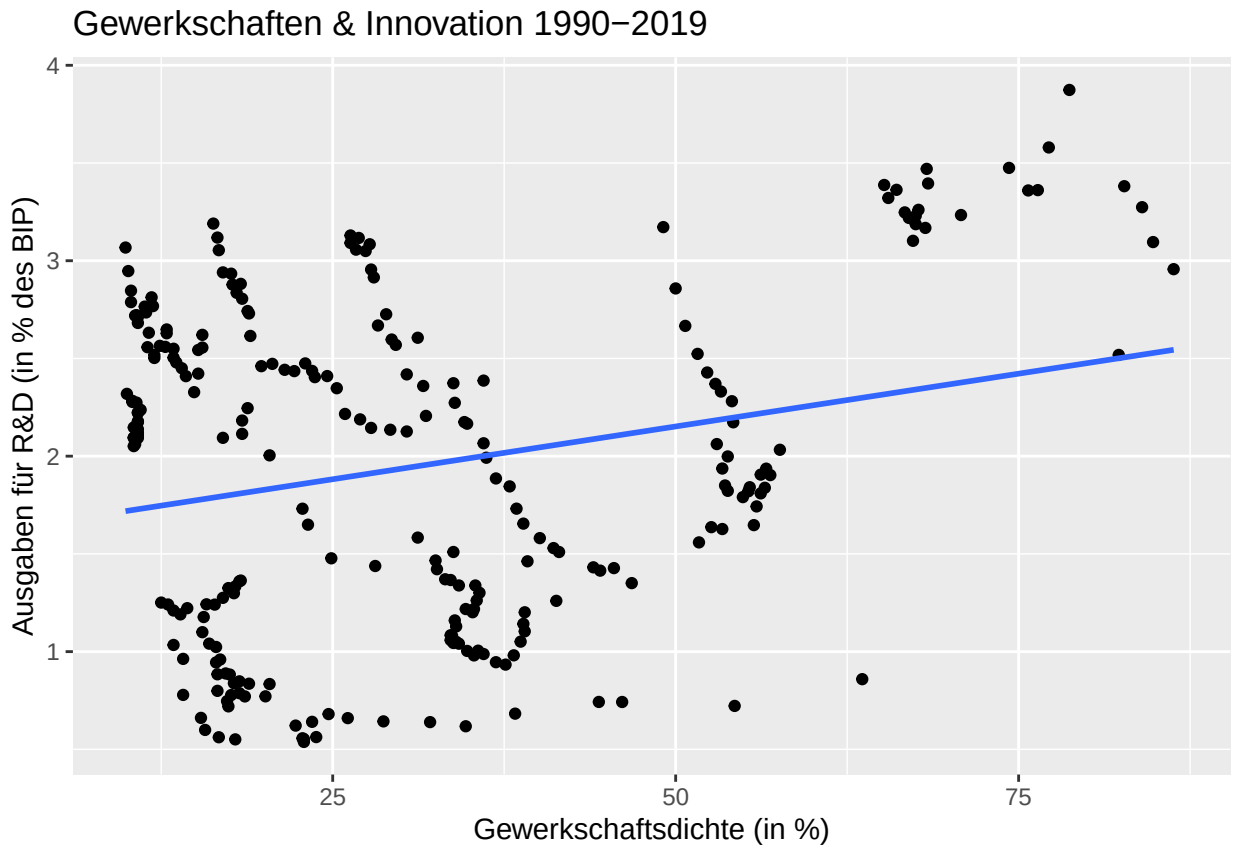
```
techmodel1 <- lm(Tech ~ UnionDensity, data=oecd_data)
summary(techmodel1)
```

```
##
## Call:
## lm(formula = Tech ~ UnionDensity, data = oecd_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4763 -0.7305  0.1139  0.7593  1.4117
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.61192    0.09779   16.48  < 2e-16 ***
## UnionDensity  0.01080    0.00268    4.03 7.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8111 on 256 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.05967,    Adjusted R-squared:  0.056
## F-statistic: 16.24 on 1 and 256 DF,  p-value: 7.346e-05
```

2.1 Scatterplot

```
scatterplot_tm1 <- ggplot2::ggplot(
  data = oecd_data,
  mapping = aes(
    x=UnionDensity,
    y=Tech)
  ) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity"
    ) +
  ggplot2::geom_smooth(
    method = "lm", se = FALSE) +
  ggplot2::scale_x_continuous(name = "Gewerkschaftsdichte (in %)") +
  ggplot2::scale_y_continuous(name = "Ausgaben für R&D (in % des BIP)") +
  ggplot2::scale_color_discrete(name="Land") +
  ggplot2::labs(title = "Gewerkschaften & Innovation 1990-2019") +
  ggplot2::coord_cartesian() +
  ggplot2::facet_null()

scatterplot_tm1
```



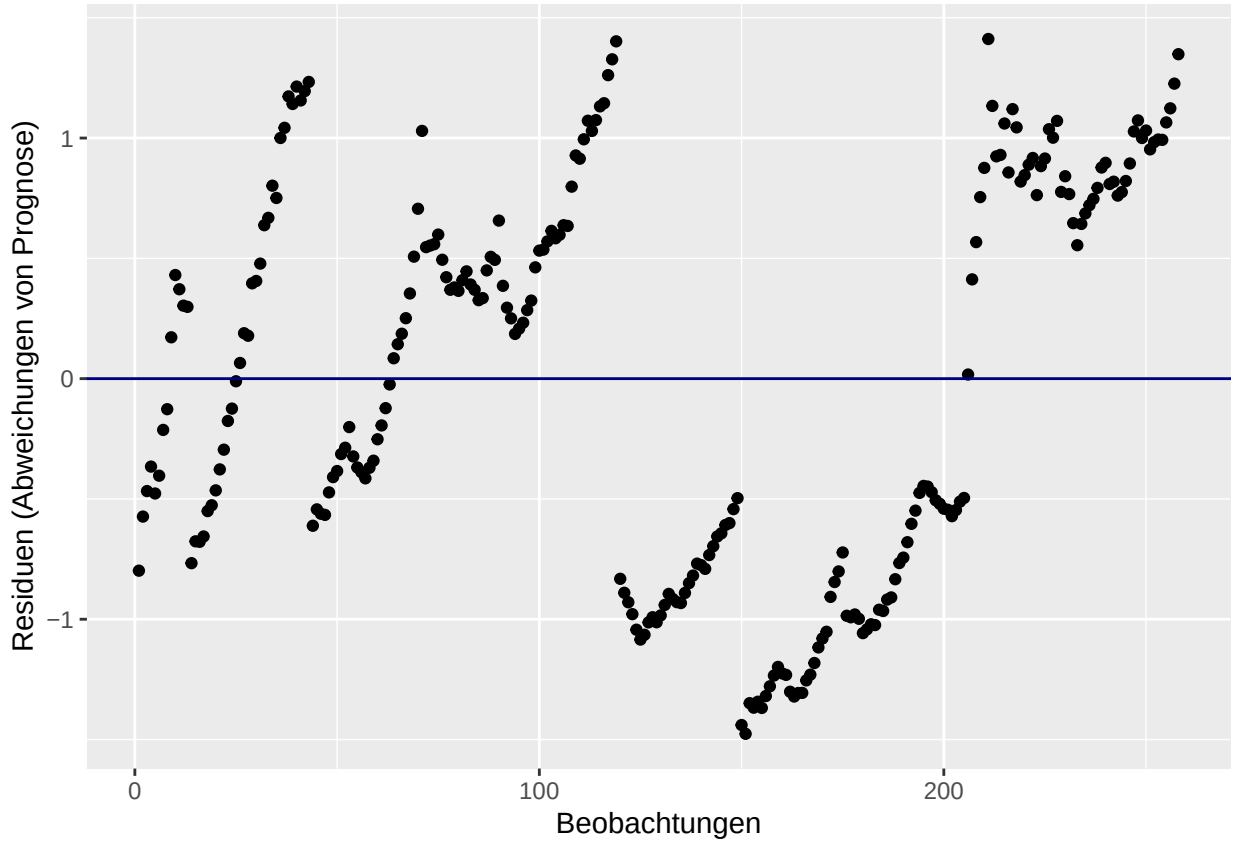
2.2 Residuenplot

```
resids <- as.data.table(techmodel1[["residuals"]])
typeof(resids)
```

```
## [1] "list"
```

```
Techm1_Residuen <- ggplot2::ggplot(
  data = resids,
  mapping = aes(
   x=1:258,
    y=V1)) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
 ggplot2::geom_abline(color='navy',
                      intercept = 0,
                      slope = 0)+
  ggplot2::scale_x_continuous(name = "Beobachtungen") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Techm1_Residuen
```

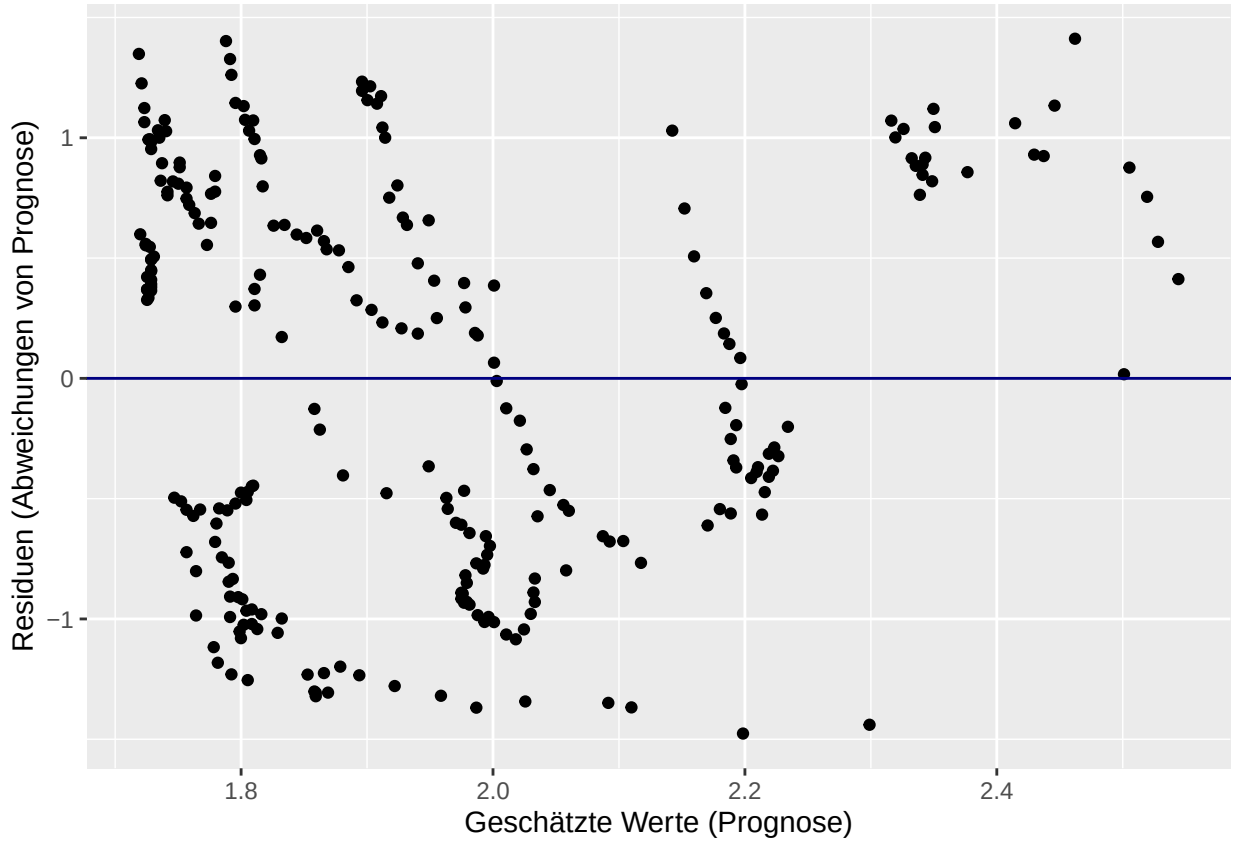


2.3 Tukey Anscombe Plot

```
TAdata <- data.table("resids"=techmodel1[["residuals"]], "fittedvalues"=predict(techmodel1))

Techm1_TA <- ggplot2::ggplot(
  data = TAdata,
  mapping = aes(
   x=fittedvalues,
    y=resids)
  ) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
   ggplot2::geom_abline(color='navy',
                  intercept = 0,
                  slope = 0)+
  ggplot2::scale_x_continuous(name = "Geschätzte Werte (Prognose)") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Techm1_TA
```

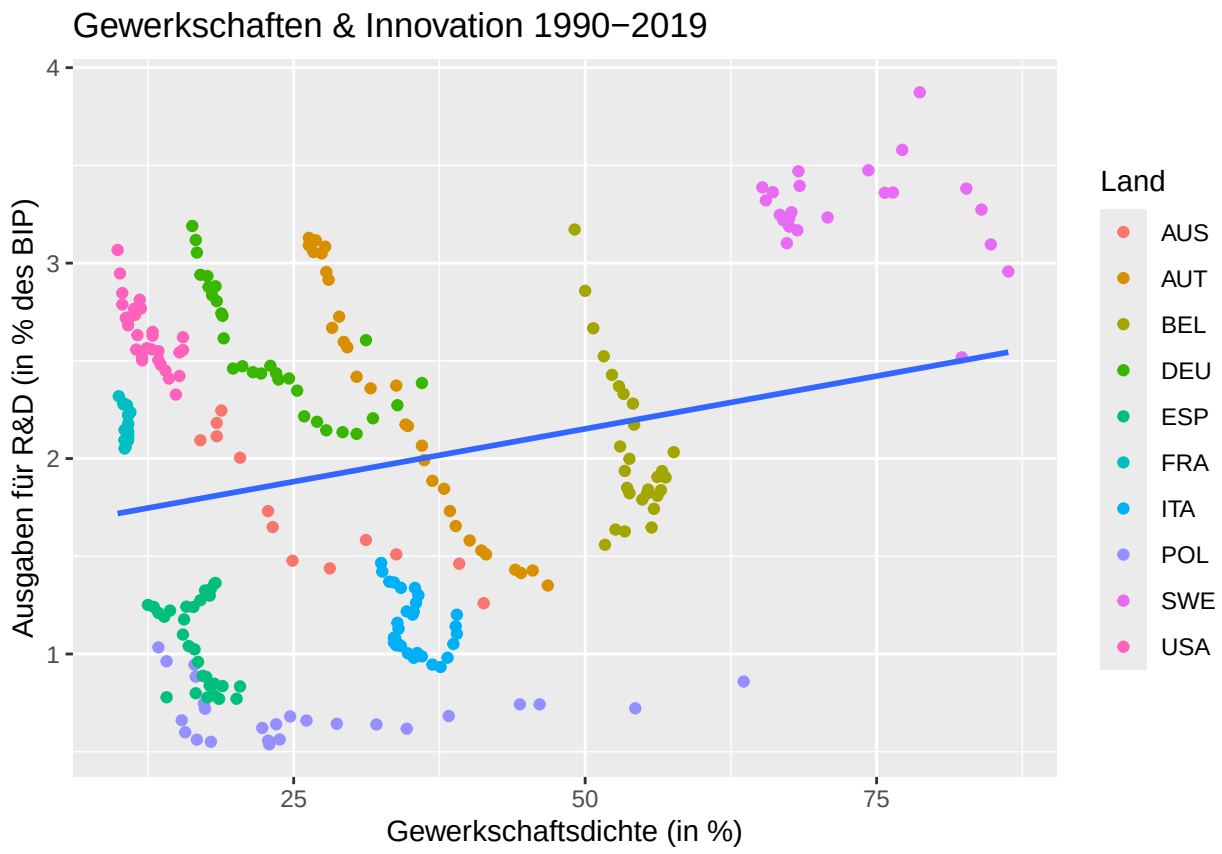


2.4 Scatterplot mit Farben für verschiedene Länder

```
scatterplot_tm1_c <- ggplot2::ggplot(
  data = oecd_data,
  mapping = aes(
    x=UnionDensity,
    y=Tech)
  ) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity",
    mapping = aes(color=Country)
    ) +
  ggplot2::geom_smooth(
    method = "lm", se = FALSE
    ) +

  ggplot2::scale_x_continuous(name = "Gewerkschaftsdichte (in %)") +
    ggplot2::scale_y_continuous(name = "Ausgaben für R&D (in % des BIP)") +
  ggplot2::scale_color_discrete(name="Land") +
  ggplot2::labs(title = "Gewerkschaften & Innovation 1990-2019") + #
  ggplot2::coord_cartesian() +
  ggplot2::facet_null()

scatterplot_tm1_c
```



3 Simulierte Regression

Wir führen eine Regression mit einem simulierten Datensatz durch. Dazu ziehen wir Zufallswerte, die normalverteilt um die echte Regressionsgerade herum liegen.

```
set.seed(123)
true_DGP <- function(x, b0, b1){
  y <- b0 + b1*x + rnorm(length(x), 0, 5)
  return(y)
}
beta_0_wahr <- 3
beta_1_wahr <- 2
sample_size <- 500
x <- runif(sample_size, 0, 10)
```

```
set.seed(123)
n_datensaetze <- 1
beta_0_estimates <- rep(NA, n_datensaetze)
beta_1_estimates <- rep(NA, n_datensaetze)

for (i in 1:n_datensaetze){
  daten_satz <- data.frame(
    x = x,
    y = true_DGP(x, beta_0_wahr, beta_1_wahr)
```

```
  )
  schaetzung_2 <- lm(y~x, data = daten_satz)
  beta_0_estimates[i] <- schaetzung_2[["coefficients"]][1]
  beta_1_estimates[i] <- schaetzung_2[["coefficients"]][2]
}

daten_satz
```

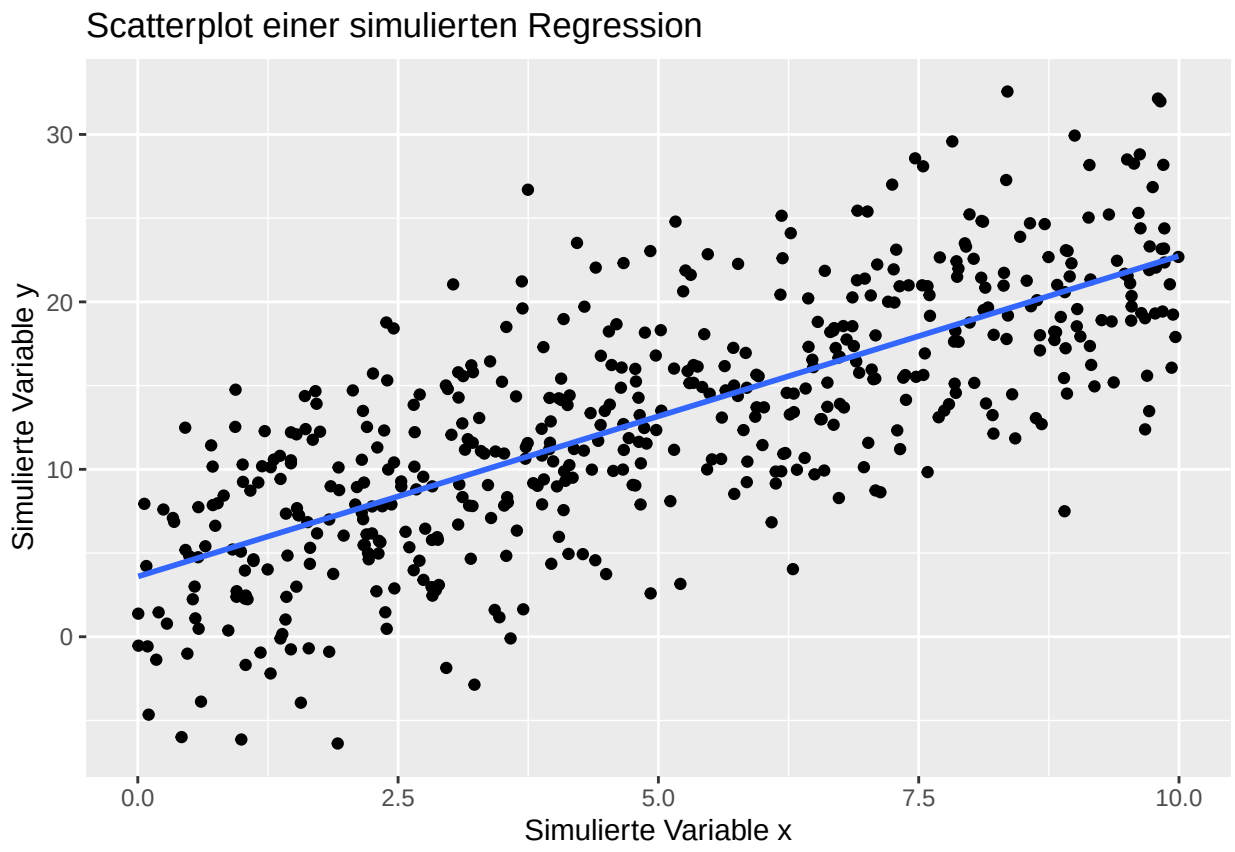
```
Simulation2 <- lm(y ~ x, daten_satz)
summary(Simulation2)
```

```
##
## Call:
## lm(formula = y ~ x, data = daten_satz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6480  -3.1004  -0.0766   3.1939  15.9292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.59641    0.43701    8.23 1.66e-15 ***
## x            1.91450    0.07653   25.02  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.863 on 498 degrees of freedom
## Multiple R-squared:  0.5568, Adjusted R-squared:  0.5559
## F-statistic: 625.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

3.1 Scatterplot

```
scatterplot_Simulation2 <- ggplot2::ggplot(
  data = Simulation2,
  mapping = aes(
    x=x,
    y=y)
  ) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity"
    ) +
  ggplot2::geom_smooth(
    method = "lm", se=FALSE) +
  ggplot2::scale_x_continuous(name = "Simulierte Variable x") +
  ggplot2::scale_y_continuous(name = "Simulierte Variable y") +
  ggplot2::labs(title = "Scatterplot einer simulierten Regression") +
  ggplot2::coord_cartesian() +
  ggplot2::facet_null()
```

```
scatterplot_Simulation2
```

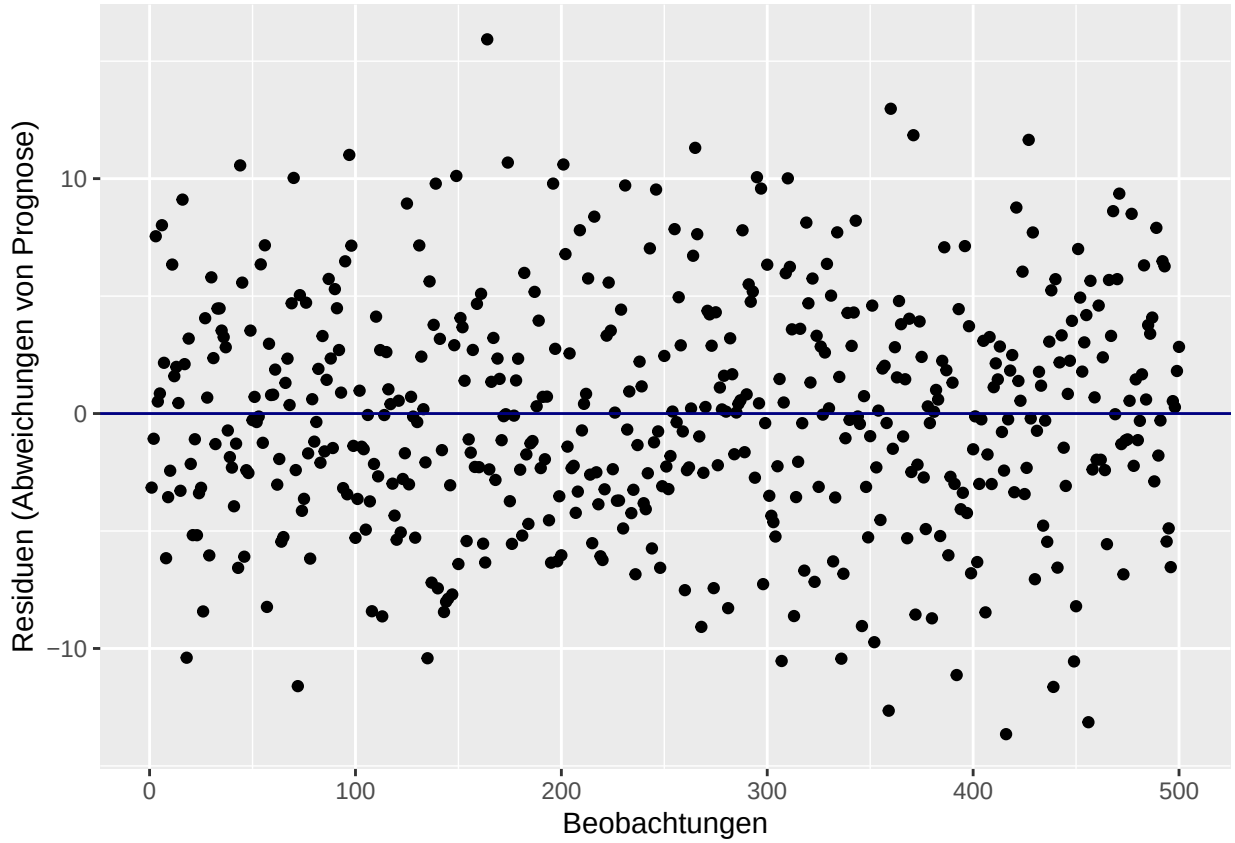


3.2 Residuen

```
residsS2 <- as.data.table(Simulation2[["residuals"]])

Simulation2_Residuen <- ggplot2::ggplot(
  data = residsS2,
  mapping = aes(
   x=1:500,
    y=V1)) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
 ggplot2::geom_abline(color='navy',
                  intercept = 0,
                  slope = 0)+
  ggplot2::scale_x_continuous(name = "Beobachtungen") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Simulation2_Residuen
```

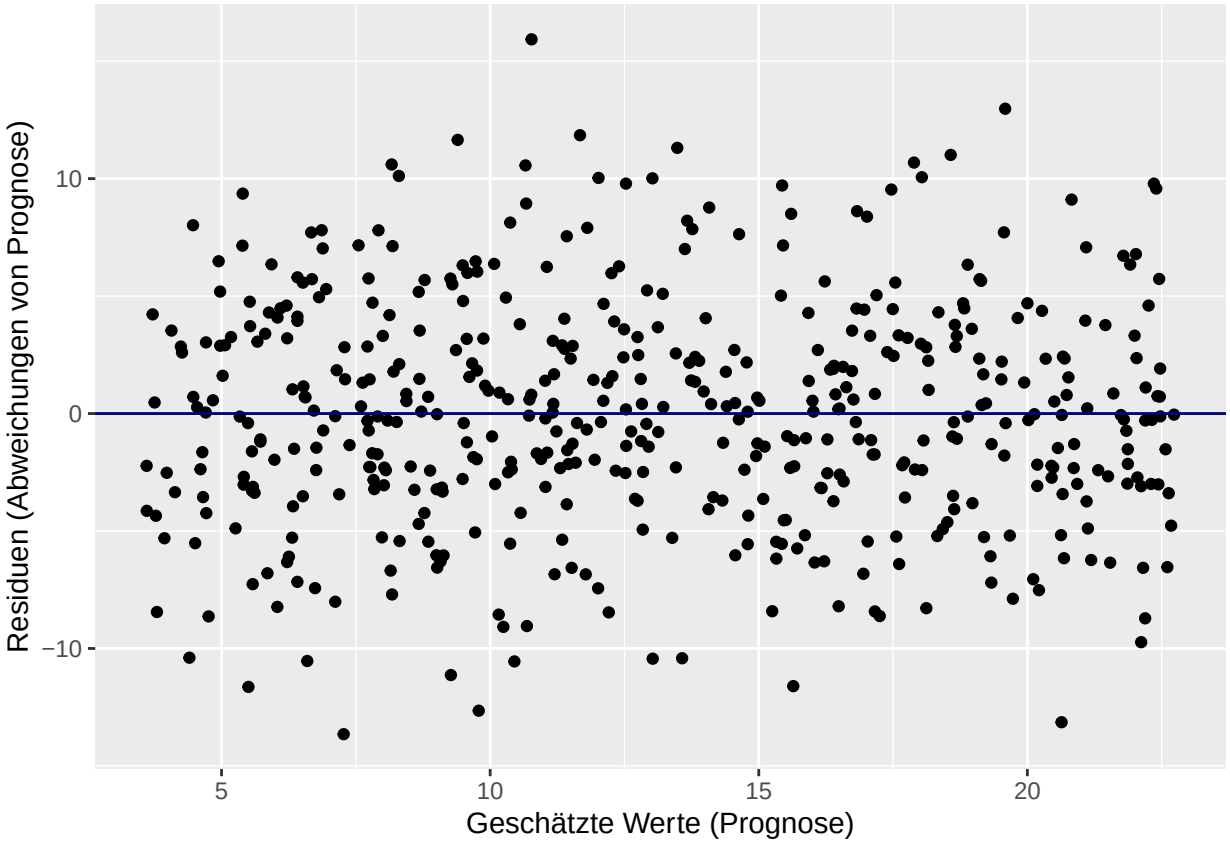


3.3 Tukey Anscombe Plot

```
TAdataSimulation2 <- data.table("resids"=Simulation2[["residuals"]], "fittedvalues"=predict(Simulation2)

Simulation2_TA <- ggplot2::ggplot(
  data = TAdataSimulation2,
  mapping = aes(
   x=fittedvalues,
    y=resids)
  ) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
   ggplot2::geom_abline(color='navy',
                  intercept = 0,
                  slope = 0)+
  ggplot2::scale_x_continuous(name = "Geschätzte Werte (Prognose)") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Simulation2_TA
```



4 Erweiterte Regression mit mehreren erklärenden Variablen

```
techmodel2 <- lm(Tech ~ UnionDensity+GDPpc+TAX, data=oecd_data)
summary(techmodel2)
```

```
##
## Call:
## lm(formula = Tech ~ UnionDensity + GDPpc + TAX, data = oecd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81954 -0.34235 -0.03506  0.29677  1.16416
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.240e-01  1.186e-01   2.732  0.00674 **
## UnionDensity  1.394e-02  1.870e-03   7.458 1.40e-12 ***
## GDPpc         5.384e-05  2.227e-06  24.180  < 2e-16 ***
## TAX          -3.505e-02  5.766e-03  -6.078 4.45e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4282 on 253 degrees of freedom
```

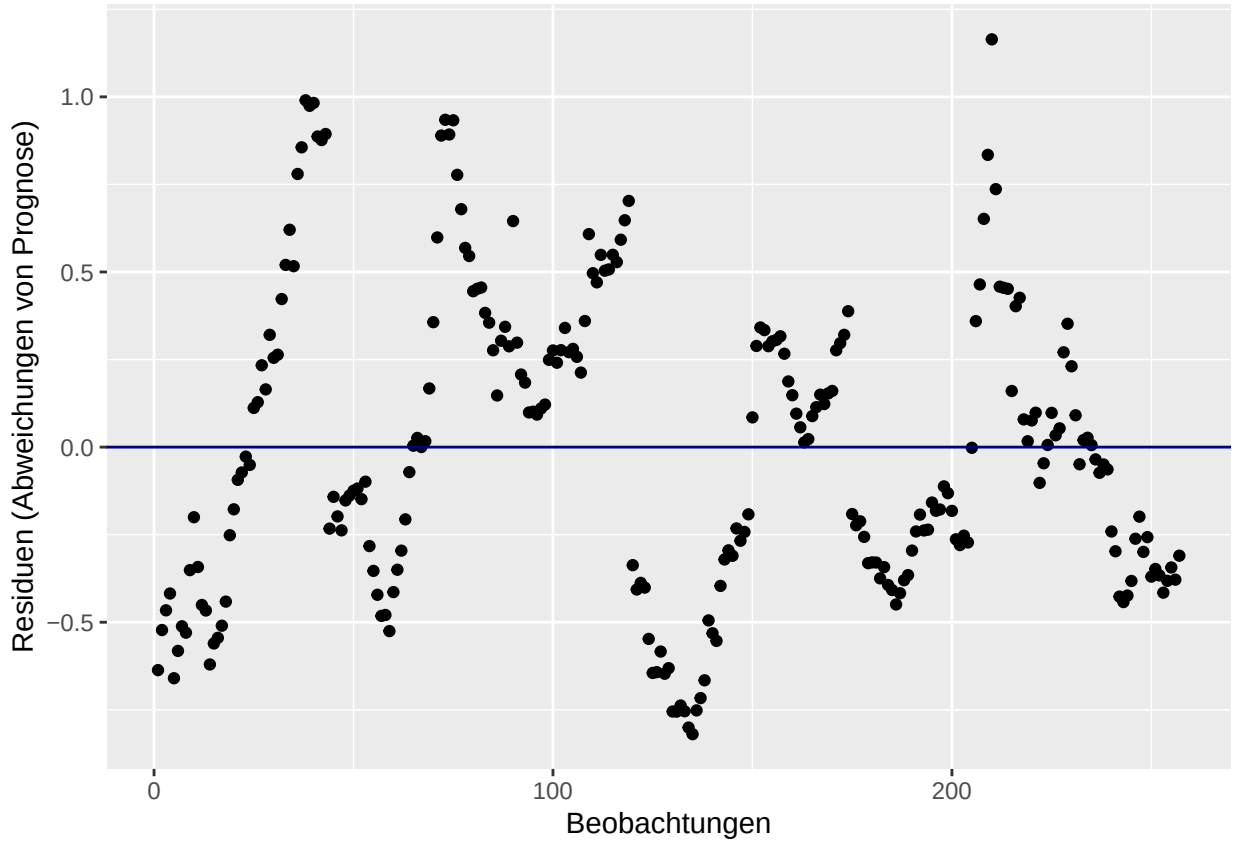
```
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.7392, Adjusted R-squared:  0.7362
## F-statistic: 239.1 on 3 and 253 DF,  p-value: < 2.2e-16
```

4.1 Residuenplot

```
resids2 <- as.data.table(techmodel2[["residuals"]])

Techm2_Residuen <- ggplot2::ggplot(
  data = resids2,
  mapping = aes(
   x=1:257,
    y=V1)) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
 ggplot2::geom_abline(color='navy',
                  intercept = 0,
                  slope = 0)+
  ggplot2::scale_x_continuous(name = "Beobachtungen") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Techm2_Residuen
```

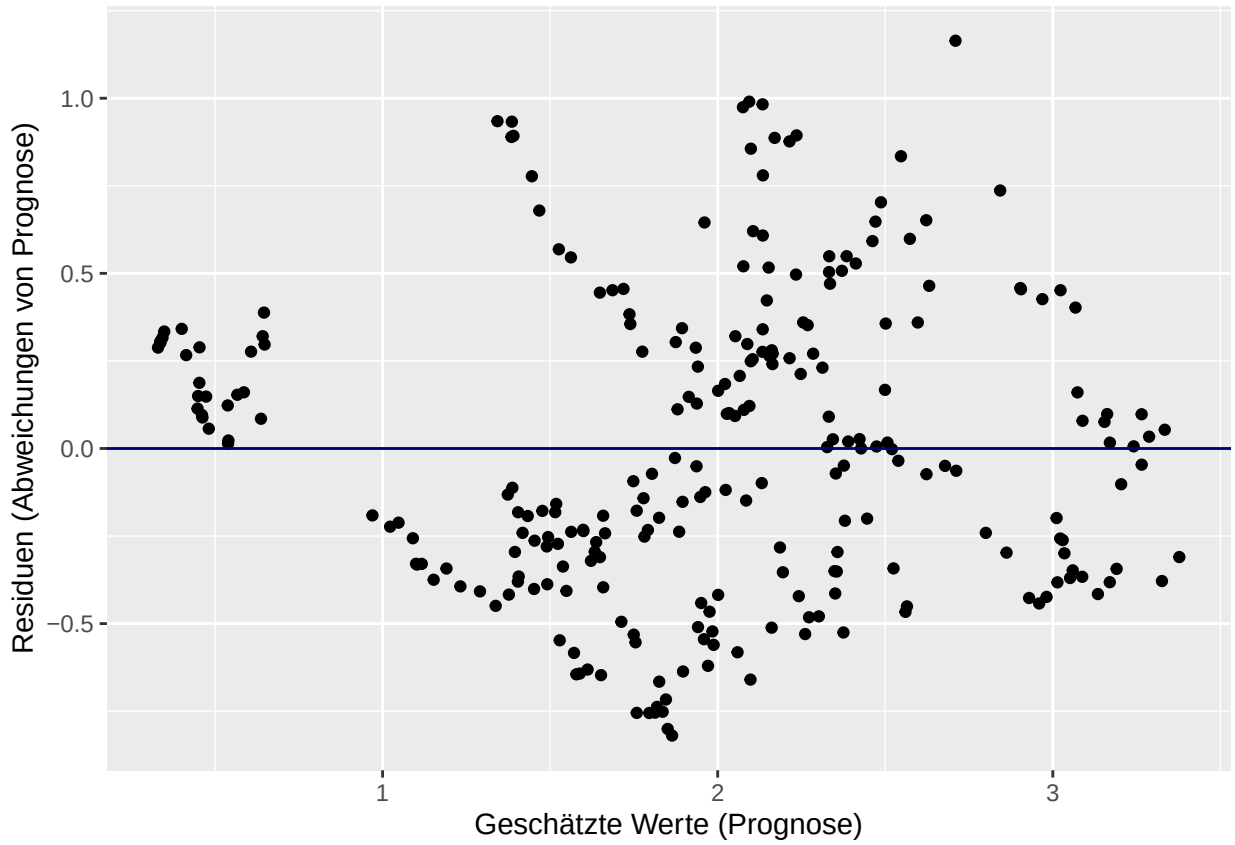


4.2 Tukey Anscombe Plot

```
TAdata2 <- data.table("resids"=techmodel2[["residuals"]], "fittedvalues"=predict(techmodel2))

Techm2_TA <- ggplot2::ggplot(
  data = TAdata2,
  mapping = aes(
   x=fittedvalues,
    y=resids)
  ) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
   ggplot2::geom_abline(color='navy',
                  intercept = 0,
                  slope = 0)+
  ggplot2::scale_x_continuous(name = "Geschätzte Werte (Prognose)") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Techm2_TA
```



5 Erweiterte Regression mit Dummy-Variablen

```
techmodel3 <- lm(Tech ~ UnionDensity+GDPpc+TAX+LME, data=oecd_data)
summary(techmodel3)
```

```
##
## Call:
## lm(formula = Tech ~ UnionDensity + GDPpc + TAX + LME, data = oecd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88318 -0.26758 -0.02478  0.27850  1.21890
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.201e-01  1.101e-01   2.907  0.00398 **
## UnionDensity  1.046e-02  1.819e-03   5.748 2.60e-08 ***
## GDPpc         6.312e-05  2.521e-06  25.040  < 2e-16 ***
## TAX          -4.102e-02  5.434e-03  -7.549 8.03e-13 ***
## LME          -5.668e-01  8.807e-02  -6.435 6.17e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3976 on 252 degrees of freedom
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.7761, Adjusted R-squared:  0.7725
## F-statistic: 218.3 on 4 and 252 DF,  p-value: < 2.2e-16
```

6 Erweiterung mit funktionaler Form

```
techmodel4 <- lm(Tech ~UnionDensity+I(UnionDensity^2)+GDPpc+TAX, data=oecd_data)
summary(techmodel4)
```

```
##
## Call:
## lm(formula = Tech ~ UnionDensity + I(UnionDensity^2) + GDPpc +
##     TAX, data = oecd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65519 -0.33891 -0.04773  0.31051  0.98236
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.661e-01  1.310e-01   5.848 1.54e-08 ***
## UnionDensity      -2.659e-02  6.686e-03  -3.977 9.12e-05 ***
## I(UnionDensity^2)  4.633e-04  7.378e-05   6.280 1.48e-09 ***
## GDPpc              5.079e-05  2.131e-06  23.830  < 2e-16 ***
## TAX               -1.862e-02  5.976e-03  -3.115  0.00205 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.399 on 252 degrees of freedom
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.771
## F-statistic: 216.4 on 4 and 252 DF,  p-value: < 2.2e-16
```

7 Erweiterung mit Logarithmus

```
techmodel5 <- lm(Tech ~UnionDensity+log(GDPpc)+TAX, data=oecd_data)
summary(techmodel5)
```

```
##
## Call:
## lm(formula = Tech ~ UnionDensity + log(GDPpc) + TAX, data = oecd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95954 -0.44694  0.08531  0.32182  1.17362
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.722301   0.647821 -15.008  < 2e-16 ***
## UnionDensity  0.015702   0.002198   7.144 9.65e-12 ***
## log(GDPpc)    1.154740   0.061864  18.666  < 2e-16 ***
## TAX          -0.040321   0.006788  -5.940 9.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5054 on 253 degrees of freedom
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.6368, Adjusted R-squared:  0.6325
## F-statistic: 147.9 on 3 and 253 DF,  p-value: < 2.2e-16
```

8 Finale Regression mit Interaktionstermen

```
techmodel6 <- lm(Tech ~ UnionDensity+GDPpc+TAX+UnionDensity*GDPpc, data=oecd_data)
summary(techmodel6)
```

```
##
## Call:
## lm(formula = Tech ~ UnionDensity + GDPpc + TAX + UnionDensity *
##     GDPpc, data = oecd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79262 -0.30974 -0.03812  0.27383  1.12091
```

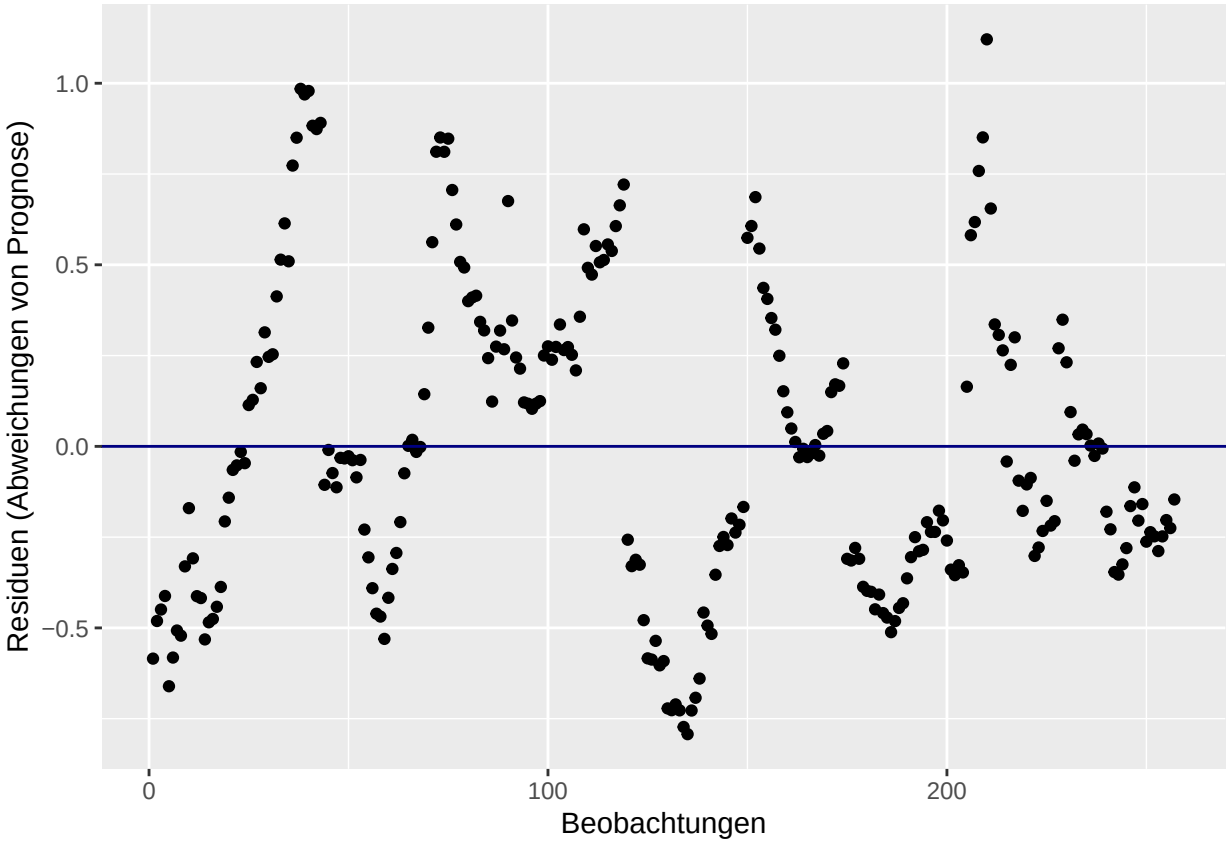
```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.218e-01  1.833e-01   4.483 1.12e-05 ***
## UnionDensity        -5.549e-03  5.850e-03  -0.948 0.343806
## GDPpc                4.125e-05  4.198e-06   9.826  < 2e-16 ***
## TAX                 -3.480e-02  5.642e-03  -6.167 2.75e-09 ***
## UnionDensity:GDPpc   4.876e-07  1.390e-07   3.508 0.000535 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.419 on 252 degrees of freedom
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.7474
## F-statistic: 190.4 on 4 and 252 DF,  p-value: < 2.2e-16
```

8.1 Residuen

```
resids6 <- as.data.table(techmodel6[["residuals"]])


Techm6_Residuen <- ggplot2::ggplot(
  data = resids6,
  mapping = aes(
   x=1:257,
    y=V1)) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
 ggplot2::geom_abline(color='navy',
                  intercept = 0,
                  slope = 0)+
  ggplot2::scale_x_continuous(name = "Beobachtungen") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Techm6_Residuen
```

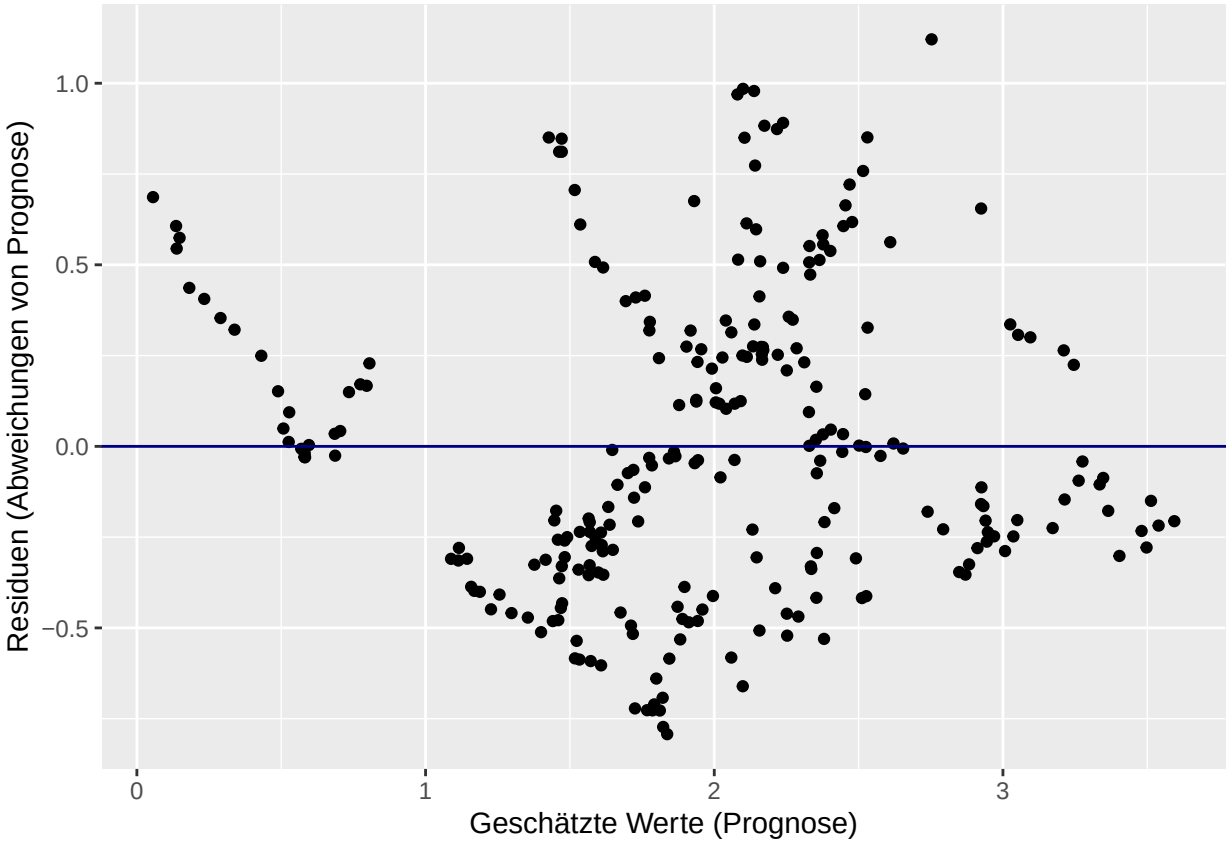


8.2 Tukey Anscombe Plot

```
TAdata6 <- data.table("resids"=techmodel6[["residuals"]], "fittedvalues"=predict(techmodel6))

Techm6_TA <- ggplot2::ggplot(
  data = TAdata6,
  mapping = aes(
   x=fittedvalues,
    y=resids)
  ) +
  ggplot2::layer(
    geom = "point",
    stat = "identity",
    position = "identity") +
   ggplot2::geom_abline(color='navy',
                  intercept = 0,
                  slope = 0)+
  ggplot2::scale_x_continuous(name = "Geschätzte Werte (Prognose)") +
  ggplot2::scale_y_continuous(name = "Residuen (Abweichungen von Prognose)")

Techm6_TA
```


Residuen (Abweichungen von Prognose)
1.0
0.5
0.0
−0.5
0
1
2
3
Geschätzte Werte (Prognose)