Box-Cox Transformationen

Thomas Klebel & Daniel Kreimer

2020-05-26

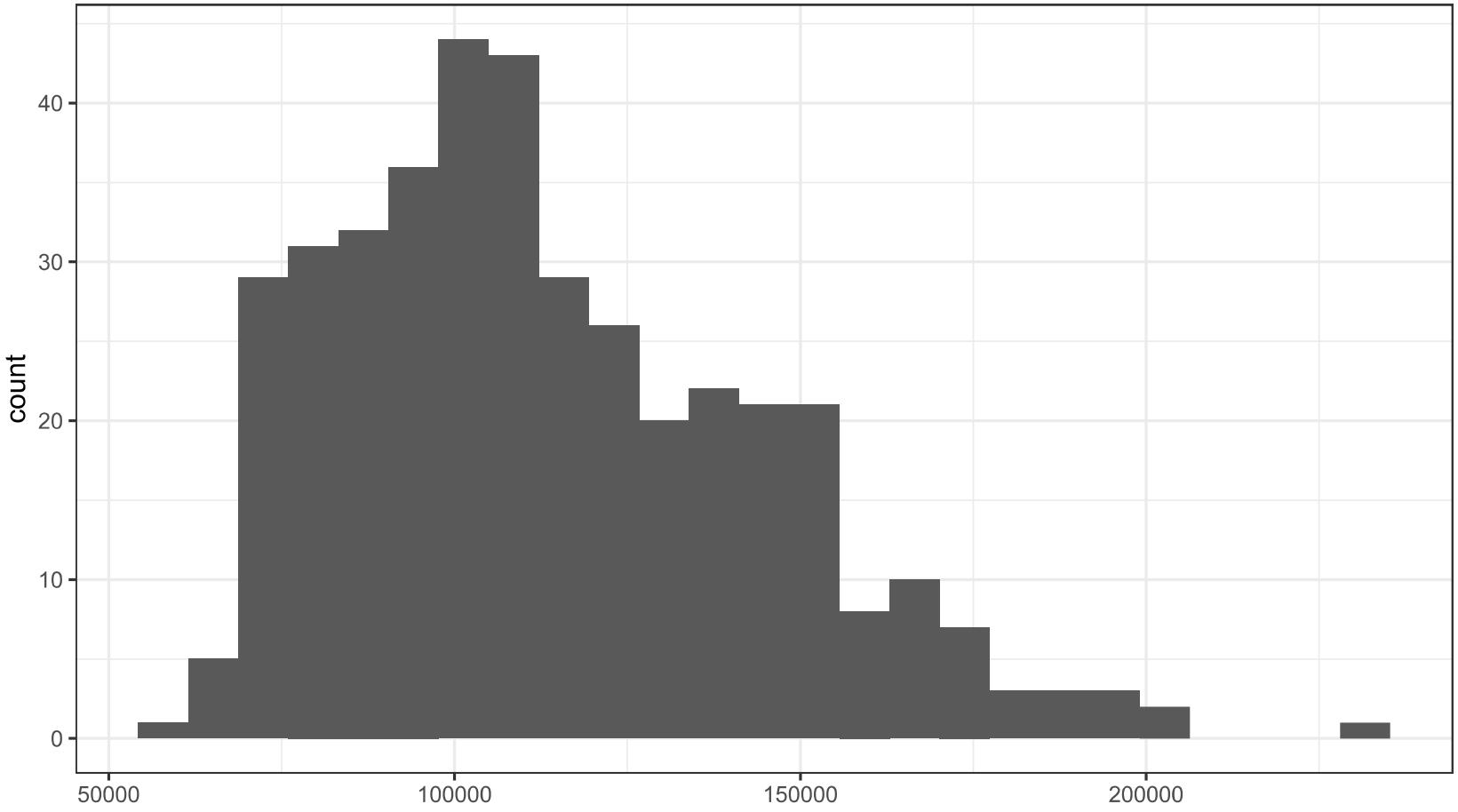
Agenda

- Problemstellung anhand eines praktischen Beispiels
  - Regressionsmodell
  - QQ-Plot
- Grundlagen der Transformation
  - Berechnung von λ
  - Auswirkung auf die Residuen
  - Interpretation
- Anwendung der Transformation auf das praktische Beispiel

Motivation

Verletzung der Normalverteilungsannahme im Regressionsmodell:

- "Die Störgrößen u_t sind normalverteilt." (von Auer 2005:413)



Beispiel: Einkommen von Uni-Professoren

Datensatz 'Salaries' (aus dem Package 'car')

```
library(car)
data(Salaries)
```

```
head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

Modellspezifikation

```
model <- lm(salary ~ rank + yrs.service + yrs.since.phd +
              discipline, data = Salaries)
broom::tidy(summary(model))
```

```
## # A tibble: 6 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     69869.     3332.     21.0  5.83e-66
## 2 rankAssocProf   12832.     4148.      3.09 2.12e- 3
## 3 rankProf        45288.     4237.     10.7  1.44e-23
## 4 yrs.service      -477.      212.     -2.25 2.50e- 2
## 5 yrs.since.phd     535.      241.      2.22 2.72e- 2
## 6 disciplineB     14505.     2343.      6.19 1.52e- 9
```

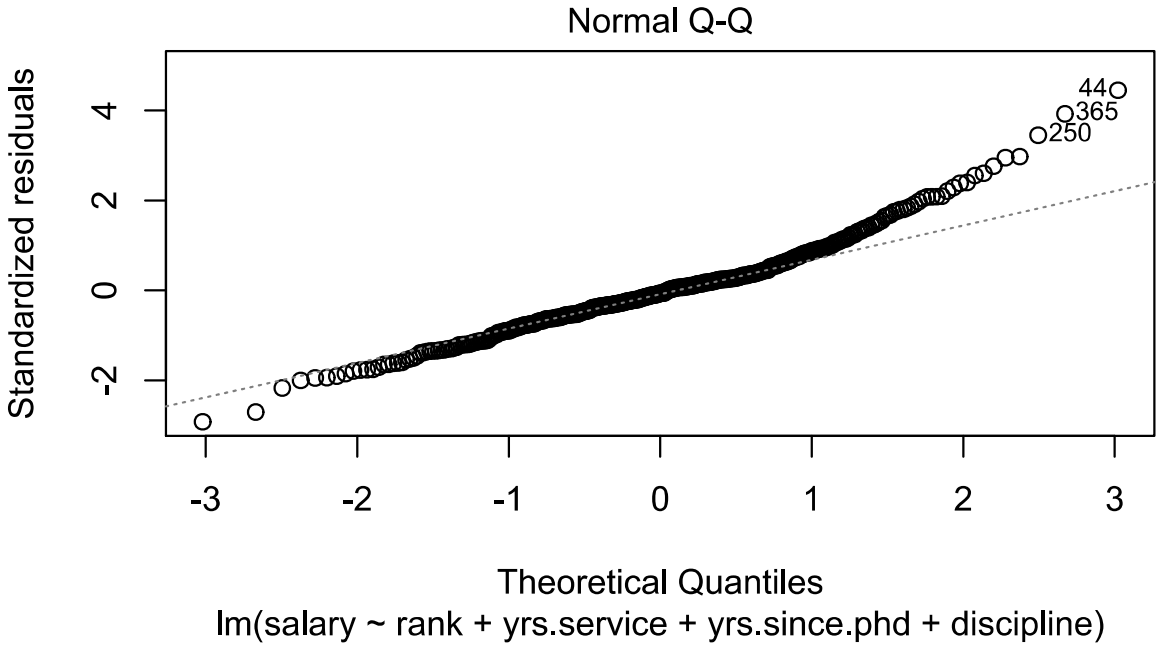
```
broom::glance(summary(model))
```

```
## # A tibble: 1 x 6
##   r.squared adj.r.squared  sigma statistic  p.value    df
##       <dbl>         <dbl>  <dbl>     <dbl>    <dbl> <int>
## 1     0.453         0.446 22554.      64.6 4.51e-49     6
```

Prüfung der NV-Annahme der Residuen

QQ-Plot: visualisiert die theoretische Position der Residuen, unter der Annahme der Normalverteilung, und stellt diese als Gerade dar. Darauf werden die beobachteten Residuen des Modells gelegt.

```
plot(model, 2)
```



Grundlagen der Transformation

Grundlagen der Transformation

Box-Cox-Modell:

$$Y_i^{(\lambda)} = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$$

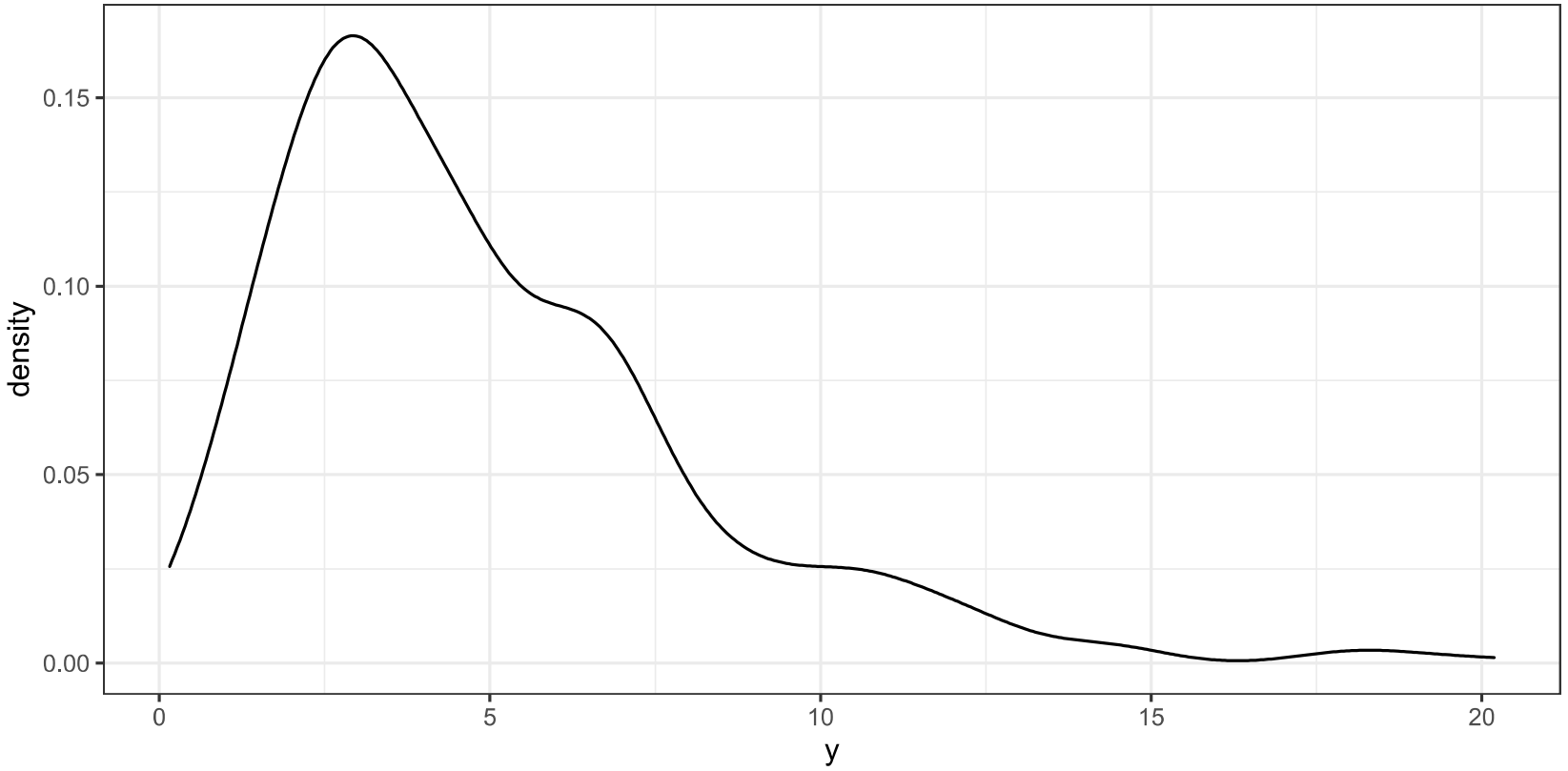
mit $\epsilon \sim N(0, \sigma_\epsilon^2)$ und

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{wenn } \lambda \neq 0 \\ log(Y_i) & \text{wenn } \lambda = 0 \end{cases}$$

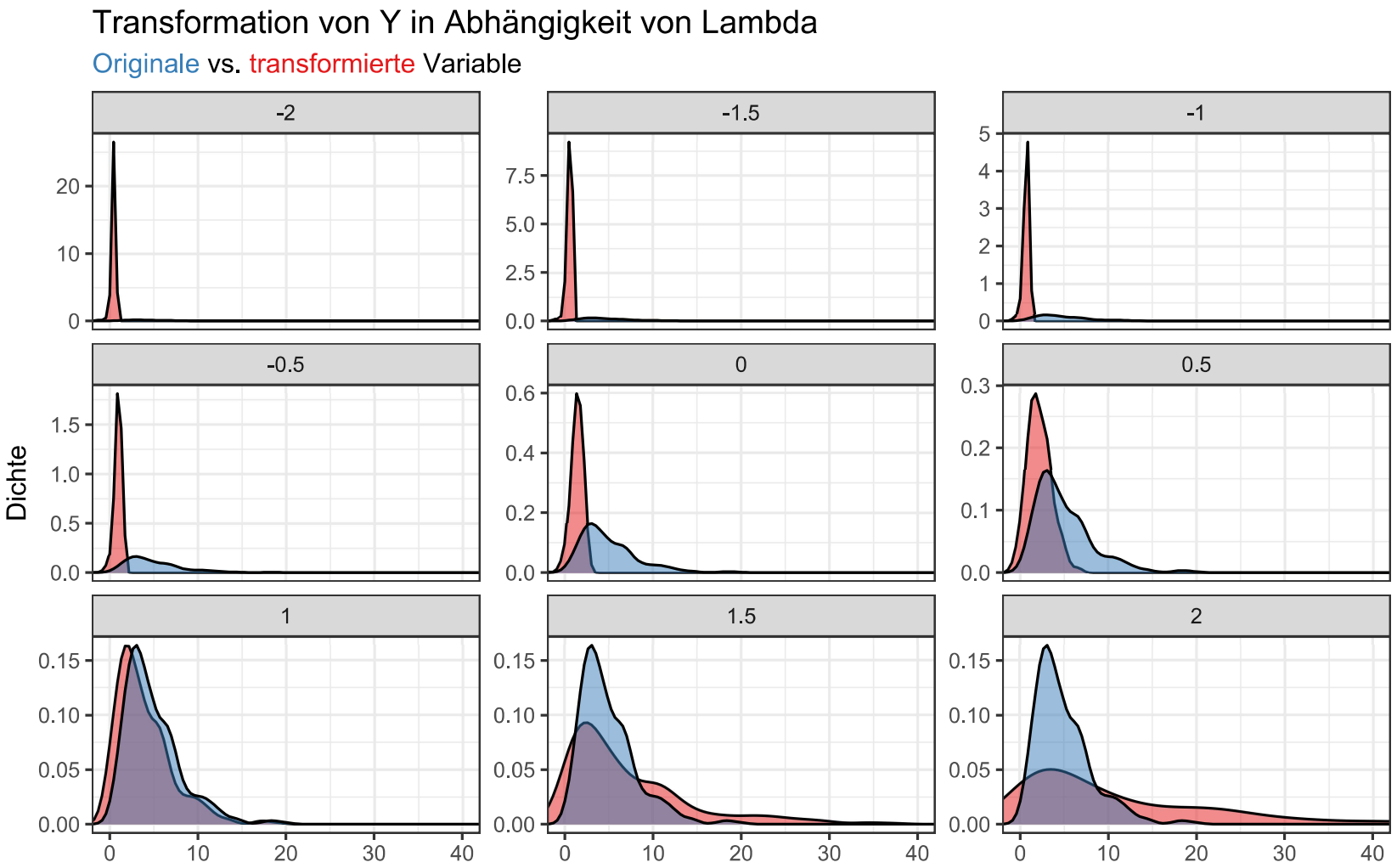
Bedingung: Alle Y-Werte müssen positiv sein.

Vergleich der Verteilungen (1)

```r
# Simluierte Werte
df <- data.frame(y = rchisq(n = 500, df = 5))
```

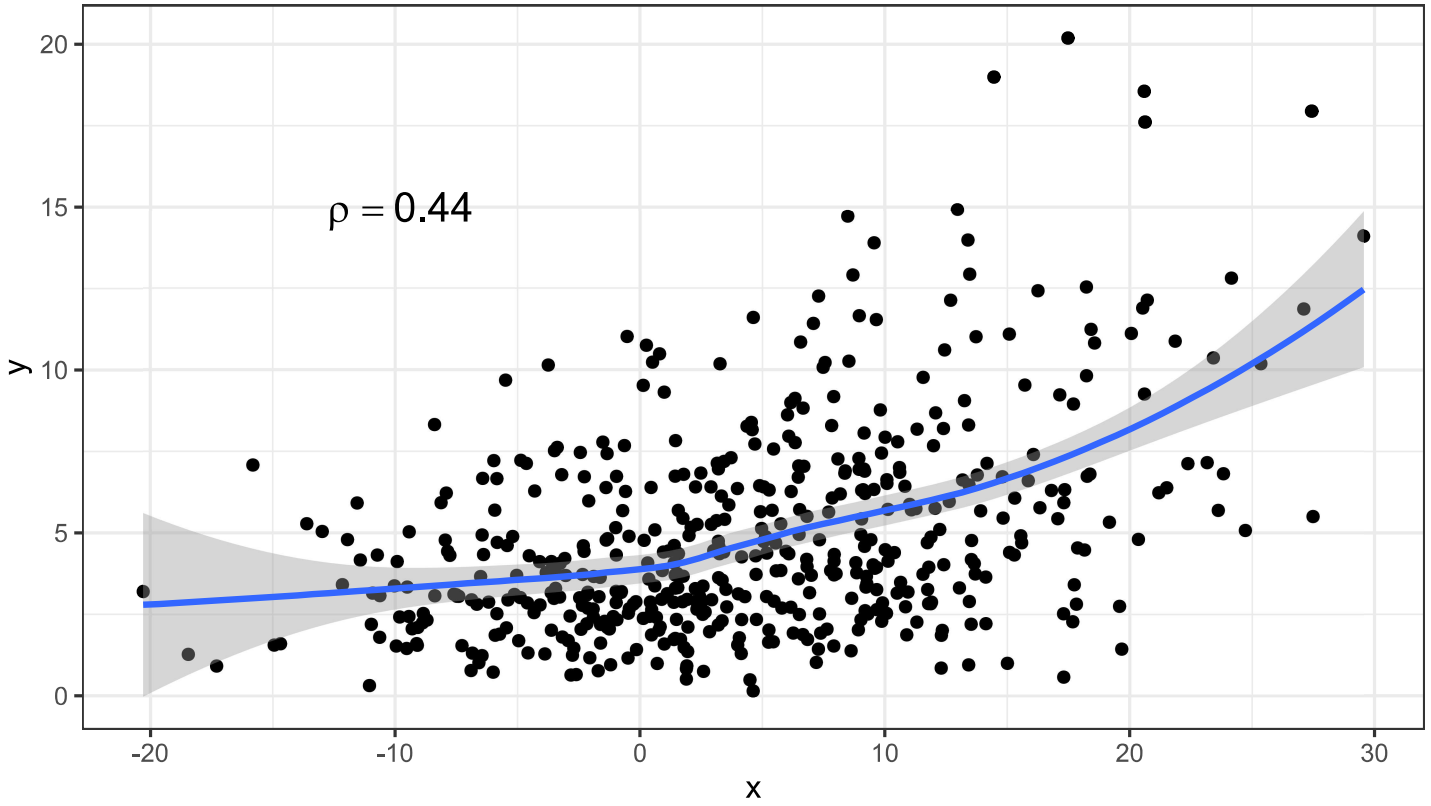


Vergleich der Verteilungen (2)



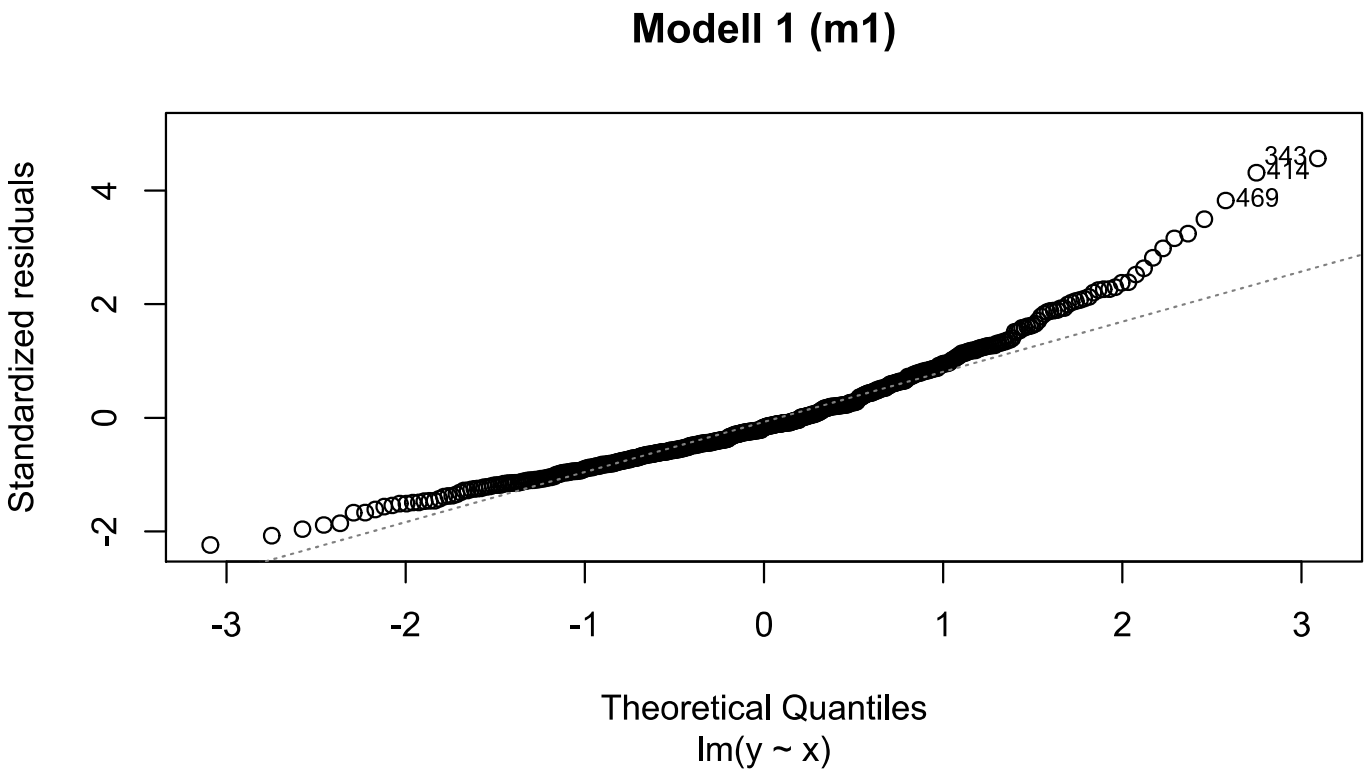
Berechnung von λ (1)

```
df <- df %>%
  mutate(x = y + rnorm(500, mean = 0, sd = 8))
```



Berechnung von λ (2)

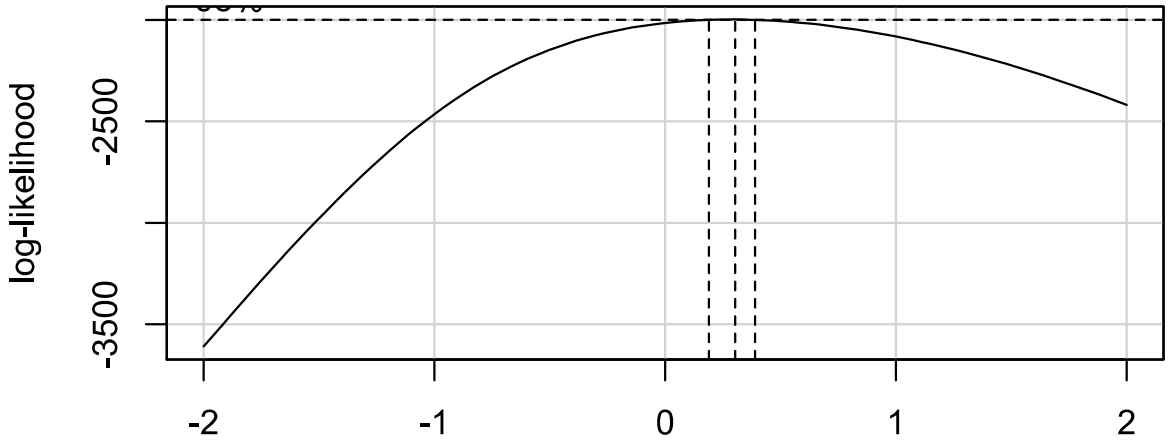
```
m1 <- lm(y ~ x, data = df)
```



Berechnung von λ (3)

`car::boxCox` berechnet λ via Maximum-Likelihood-Schätzung (basierend auf den Residuen)

```
bc <- car::boxCox(m1)
```

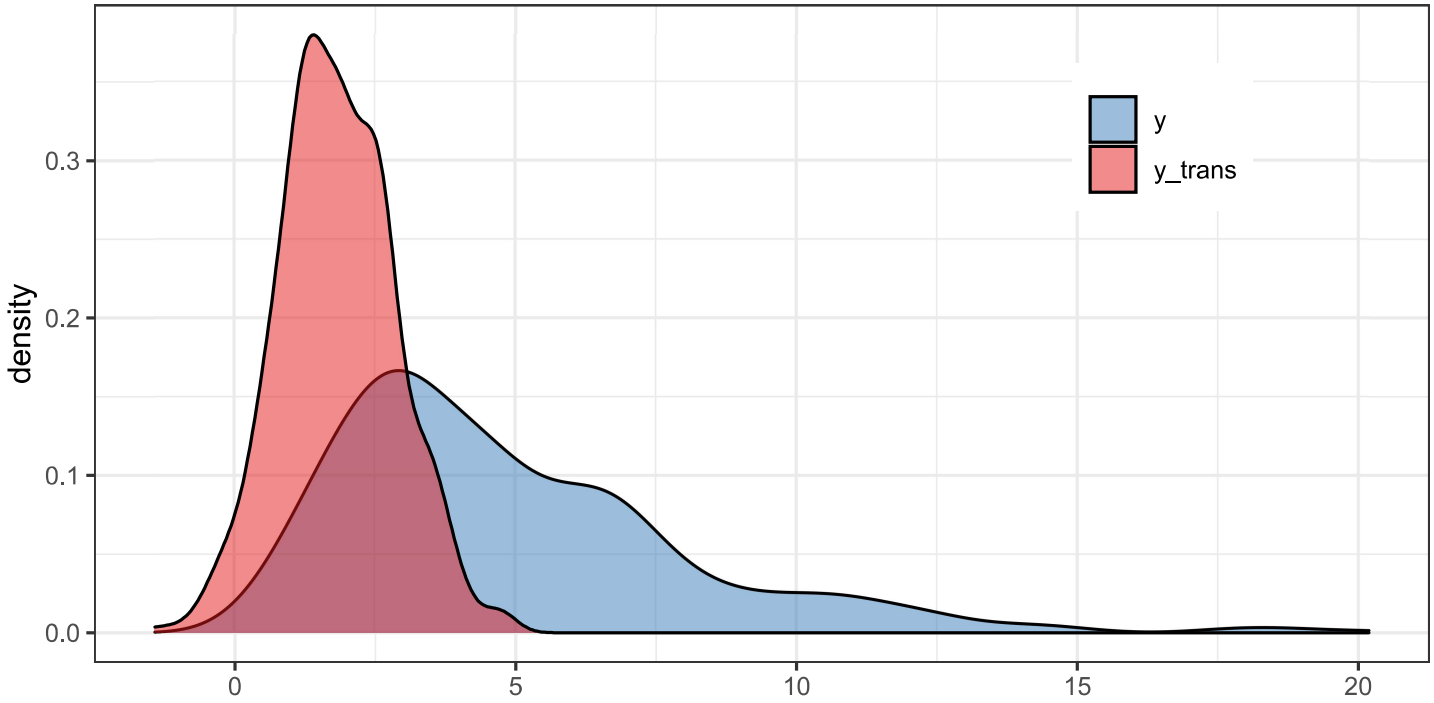



```
best.lambda <- bc$x[which(bc$y == max(bc$y))]
best.lambda
```

```
## [1] 0.3030303
```

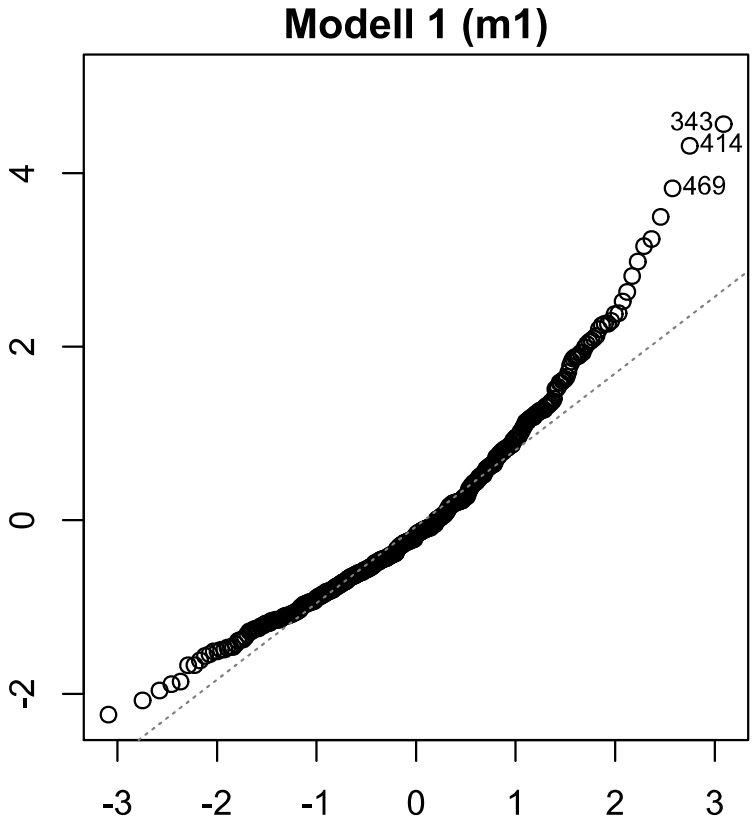
Neues Modell mit Transformation (1)

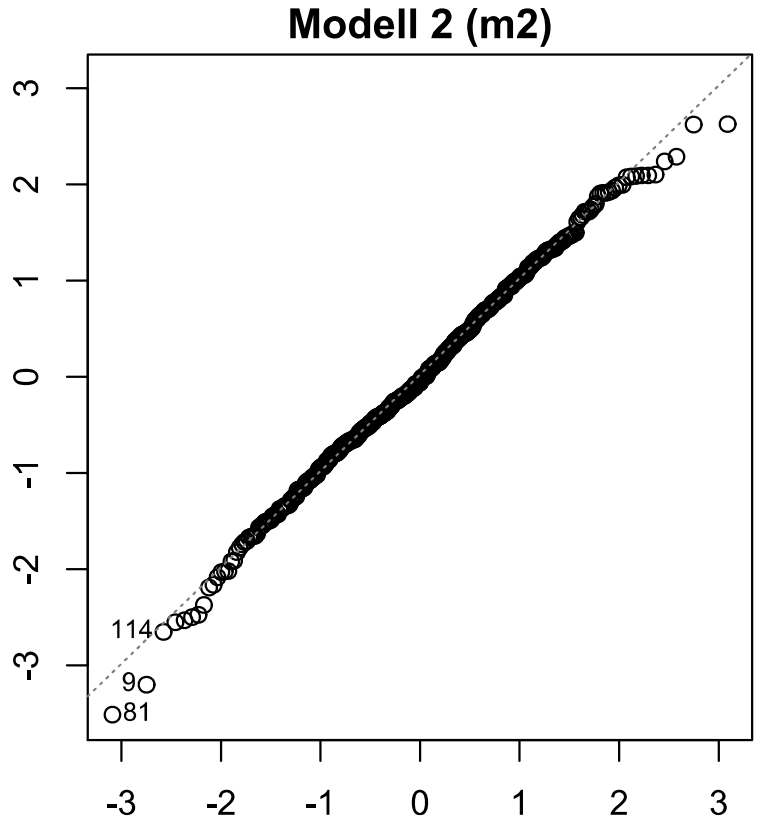
```r
transform_box_cox <- function(y, lambda) {
  (y ^ lambda - 1)/lambda
}

df <- mutate(df, y_trans = transform_box_cox(y, best.lambda))
```



Neues Modell mit Transformation (2)

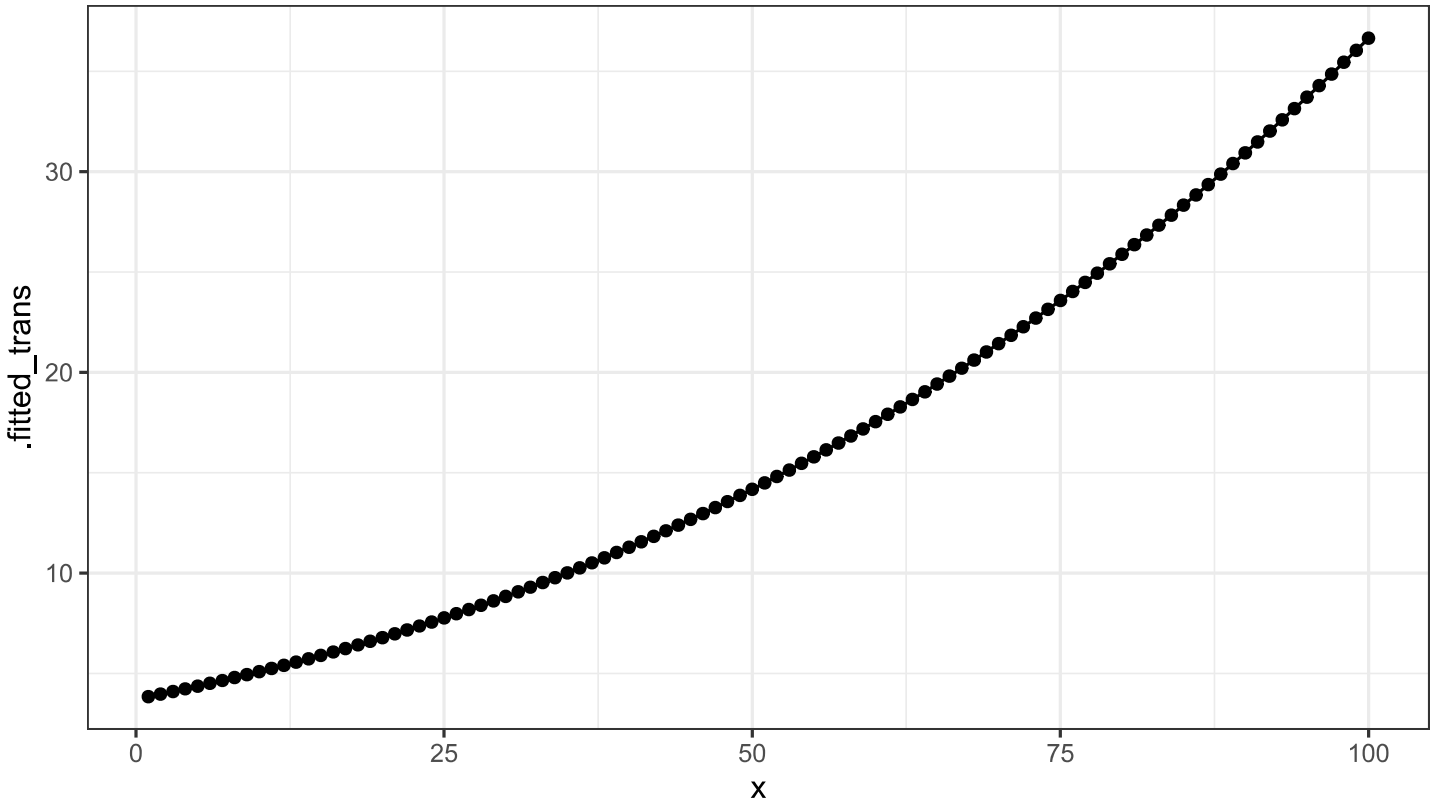
```
m2 <- lm(y_trans ~ x, data = df)
```





Interpretation

- Direkte Interpretation der Koeffizienten schwierig, außer für bekannte Fälle wie $log(Y)$
- Alternativ: Vorhersage für plausible X_i und Re-Transformierung der Vorhersage



Fortsetzung praktisches Beispiel

Box-Cox-Modell als Heuristik

> We shall choose λ partly in the light of the information provided by the data and partly from general considerations of simplicity, ease of interpretation, etc. For instance, it would be quite possible for the formal analysis to show that say $\sqrt{y}$ is the best scale for normality and constancy of variance, but for us to decide that there are compelling arguments of ease of interpretation for working say with $log(y)$. [...] the method developed below for finding a transformation is useful as a guide, but is, of course, not to be followed blindly. (Box and Cox 1964:213)

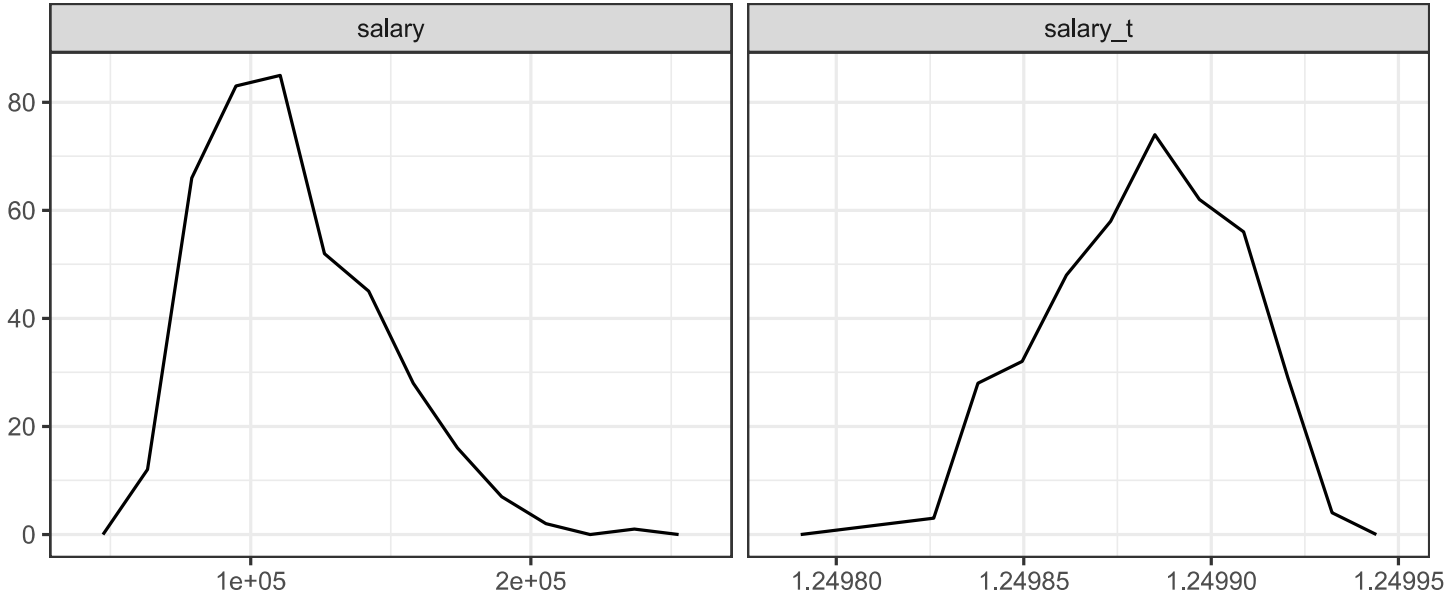
Box, G. E. P., and D. R. Cox. 1964. “An Analysis of Transformations.” Journal of the Royal Statistical Society. Series B (Methodological) 26 (2): 211–52.

Transformieren der abhängigen Variable

```
bc <- car::boxCox(model, plotit = FALSE)
(best.lambda <- bc$x[which(bc$y == max(bc$y))])
```

```
## [1] -0.8
```

```
Salaries$salary_t <- transform_box_cox(Salaries$salary, best.lambda)
```



Neues Modell

```
new_model <- lm(salary_t ~ rank + yrs.service + yrs.since.phd +
                  discipline, data = Salaries)
broom::tidy(summary(new_model))
```

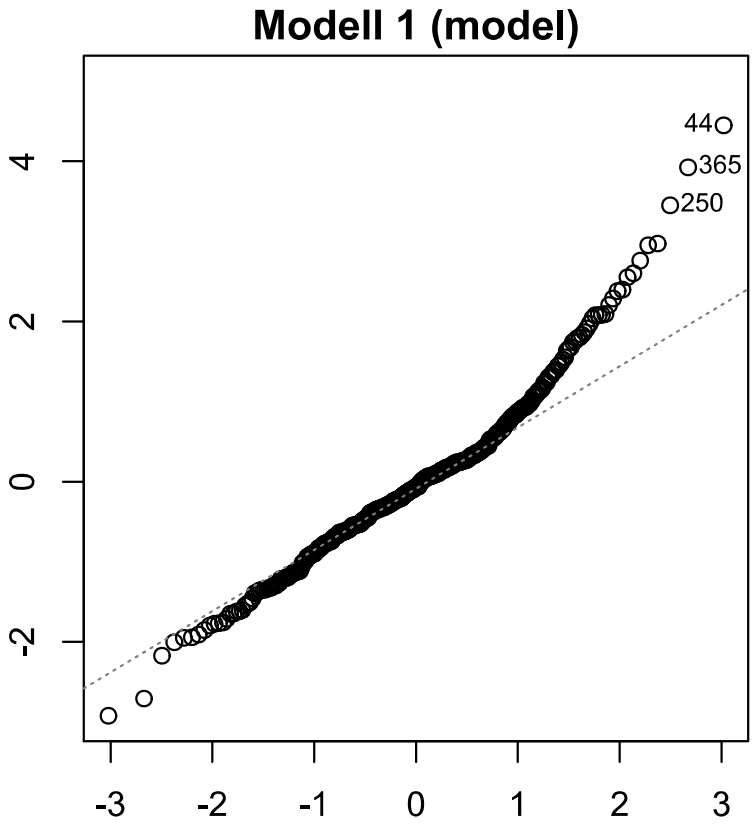
```
## # A tibble: 6 x 5
##   term              estimate   std.error statistic  p.value
##   <chr>                <dbl>       <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.25        0.00000240  520828.   0.
## 2 rankAssocProf  0.0000173   0.00000299       5.80 1.34e- 8
## 3 rankProf       0.0000460   0.00000305      15.1  8.51e-41
## 4 yrs.service   -0.000000334 0.000000153     -2.19 2.90e- 2
## 5 yrs.since.phd  0.000000215 0.000000174      1.24 2.16e- 1
## 6 disciplineB    0.0000129   0.00000169       7.62 1.91e-13
```

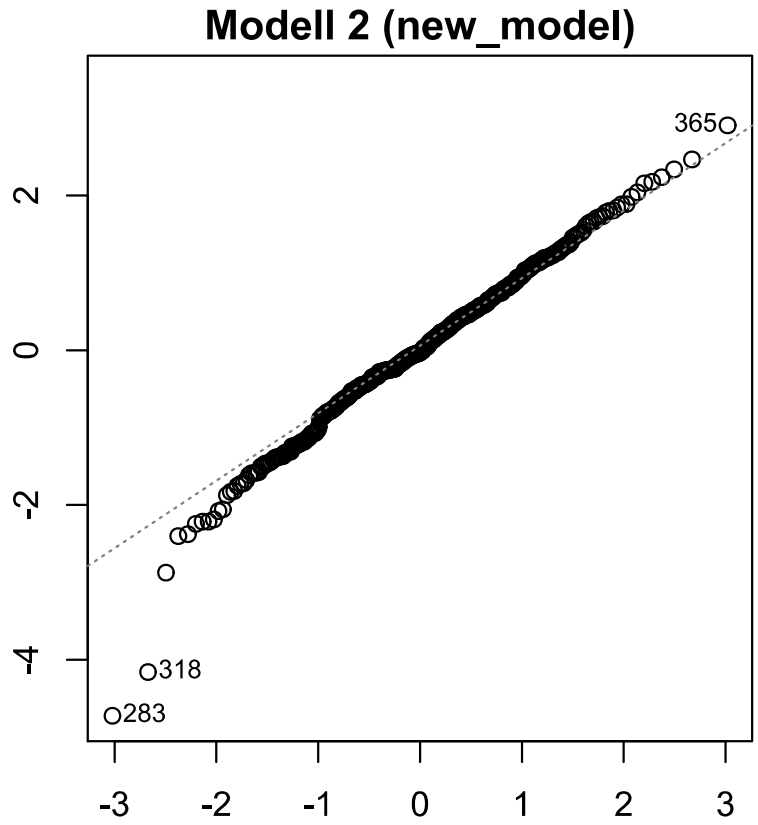
```
broom::glance(summary(new_model))
```

```
## # A tibble: 1 x 6
##   r.squared adj.r.squared     sigma statistic  p.value    df
##       <dbl>         <dbl>     <dbl>     <dbl>    <dbl> <int>
## 1     0.561         0.556 0.0000162      100. 1.05e-67     6
```

QQ-Plot

Verbesserte Normalverteilung der Residuen bei transformiertem Y.





Slides und Code

https://github.com/tklebel/box_cox_introduction