Remedial Measures

Dima Mikhaylov

9/29/2021

Mammals dataset

The data set mammals from the MASS package contains the average brain and body weights for 62 species of land mammals. We wish to see how body weight (x) could explain the brain weight (y) of land mammals.

```
library(MASS)
#data(package = 'mammals')
head(mammals)
```

```
##                    body brain
## Arctic fox        3.385  44.5
## Owl monkey        0.480  15.5
## Mountain beaver   1.350   8.1
## Cow             465.000 423.0
## Grey wolf        36.330 119.5
## Goat             27.660 115.0
```

1. Create a scatter plot of brain weight against body weight of land mammals. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

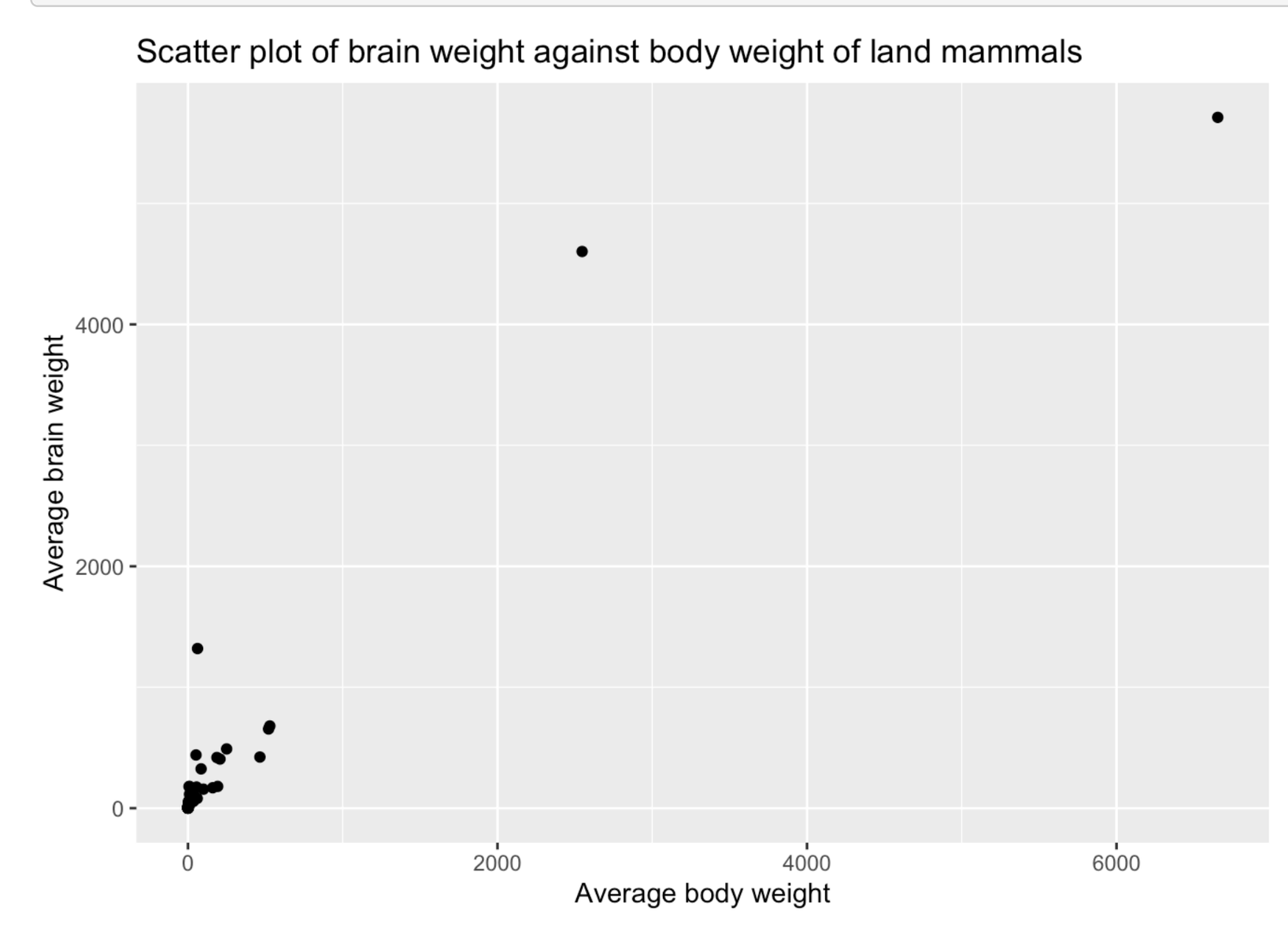
```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.3     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
ggplot(mammals, aes(x=body, y=brain))+
  geom_point()+
  labs(x="Average body weight", y="Average brain weight",
       title="Scatter plot of brain weight against body weight of land mammals")
```



Comment on appearance of the plot: 2 influencial observations on the far right may cause a problem when fitting the least squares regression line. It is not 100% obvious if assumptions of SLR were actually violated. Several observations:

- General pattern may be assumed linear, NOT the last 2 observations;
- Data points are NOT evenly scattered around OLS line;
- Vertical variation of data points is NOT constant.

2. Fit a simple linear regression to the data, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

```
library(tidyverse)
ggplot(mammals, aes(x=body, y=brain))+
  geom_point()+
  geom_smooth(method = "lm", se = TRUE)+
  labs(x="Average body weight", y="Average brain weight",
       title="Regression line with residuals for brain weight against body weight of land mammals")
```

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

3. Based on your answers to parts 1 and 2, do we need to transform at least one of the variables? Briefly explain.

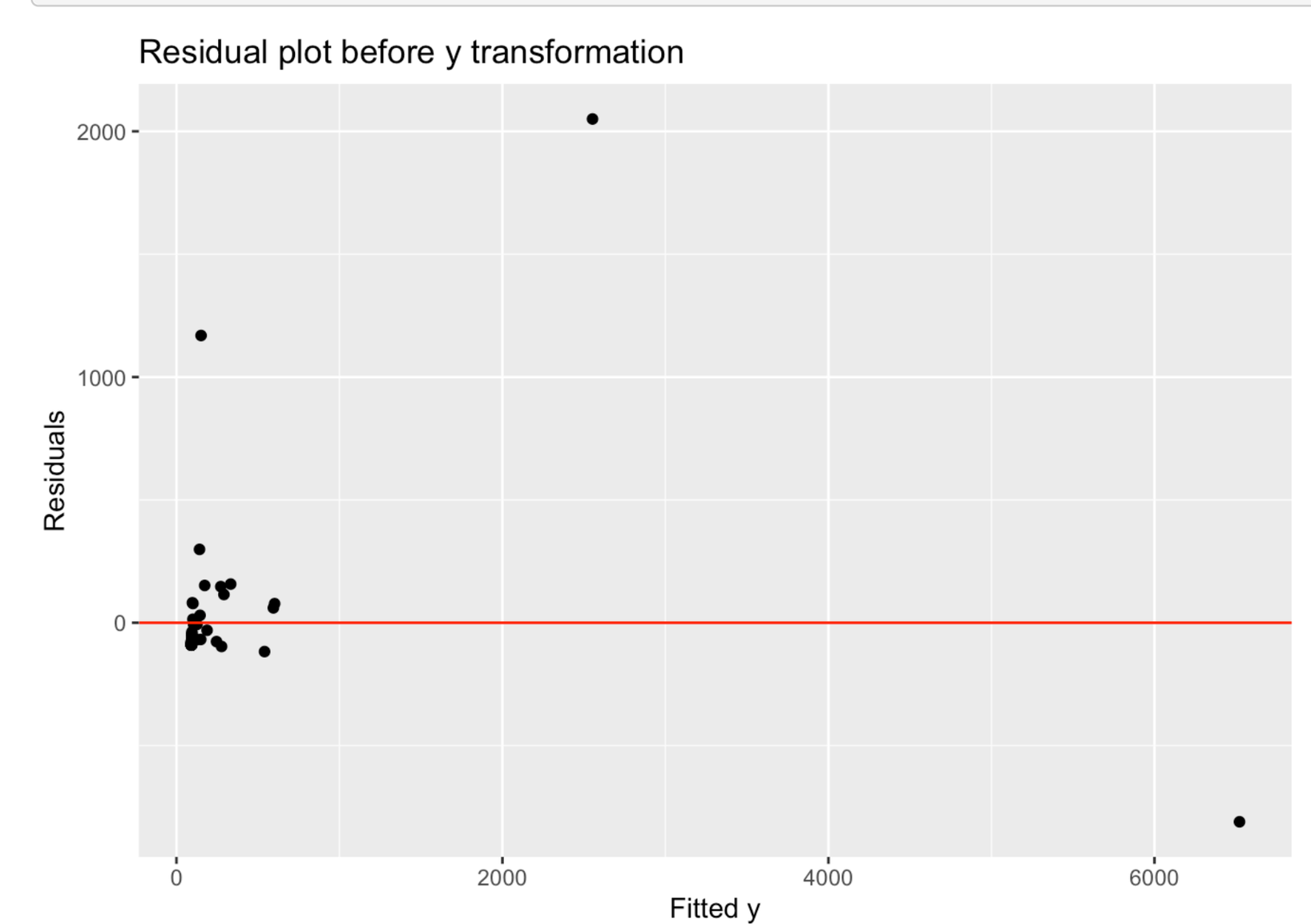
* Yes, at least one, variance is increasing => to remedy non-constant variance (problem #2) – errors need to be i.i.d. On scatterplot this should result in residuals 1) being randomly scattered, 2) not displaying any pattern (mean 0), and 3) spread of residuals for each fitted value of x should be constant (constant variance)

* Transforming the response variable is used to mitigate non-constant variable (problem #2), so taking a log of y may help. This could also potentially help imporove issues with non-zero mean (problem #1).

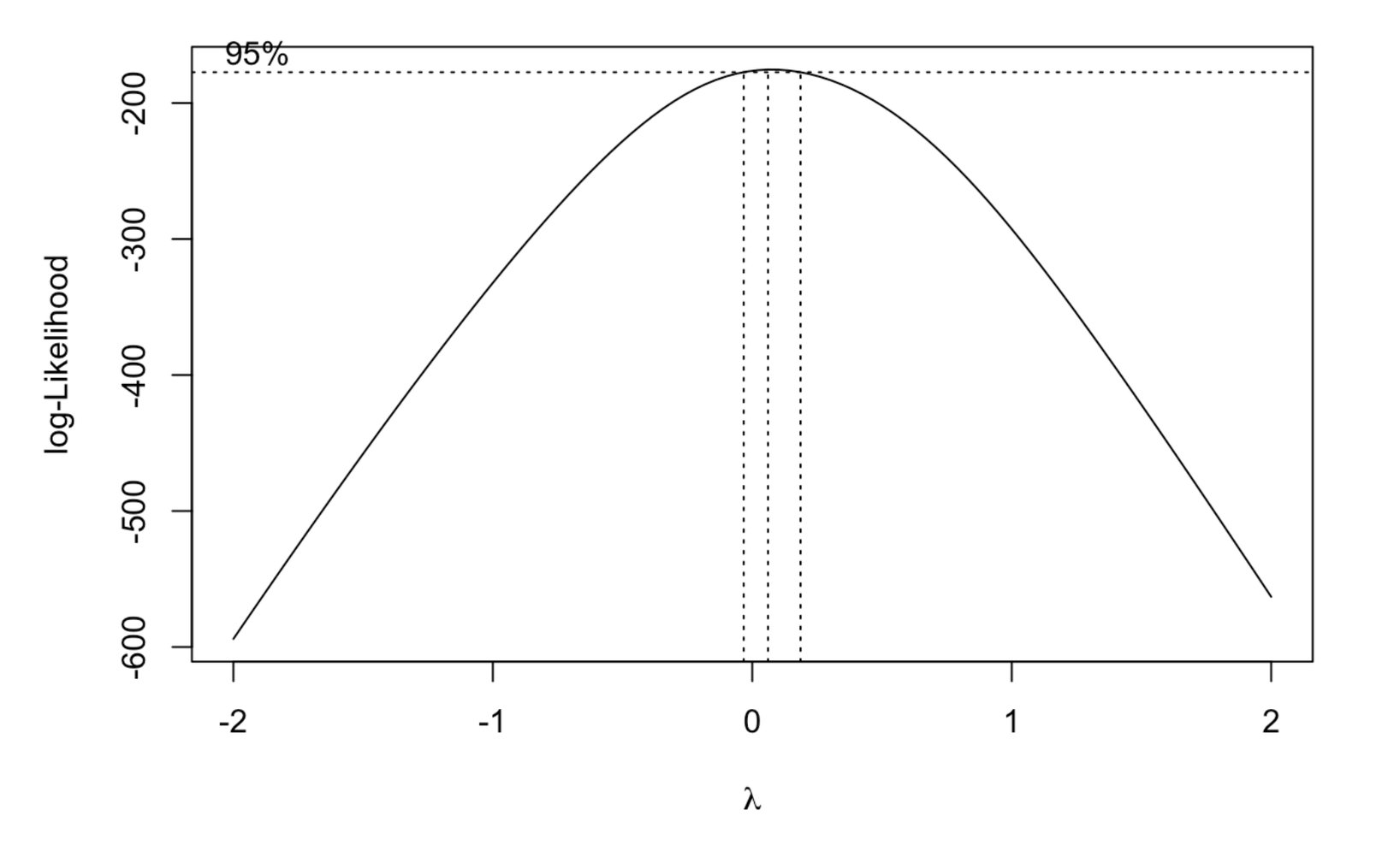
4. For the simple linear regression in part 2, create a Box Cox plot. What transformation, if any, would you apply to the response variable? Briefly explain.

* Log transformation of the response variable is used to remedy non-constant variance.

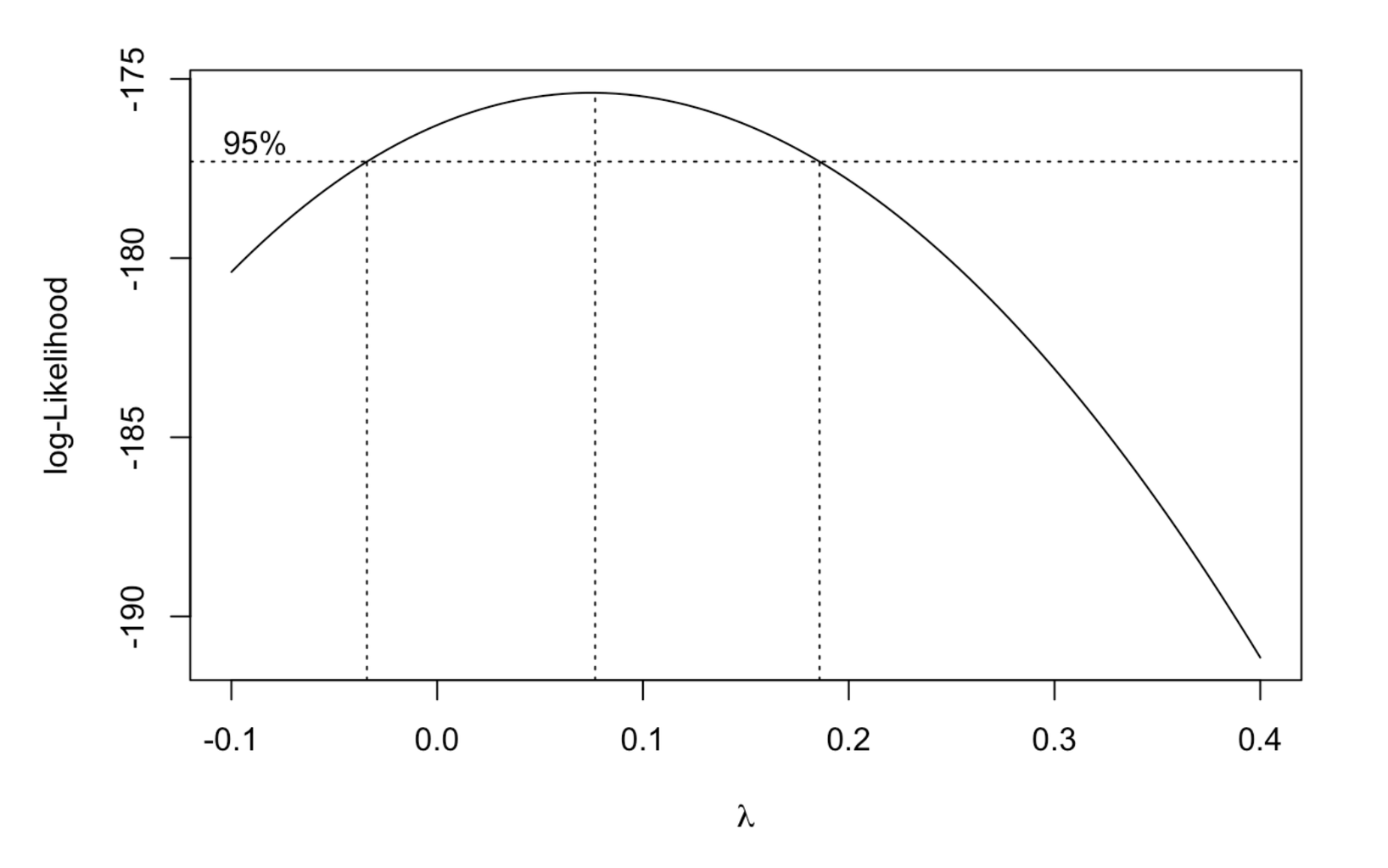
```
result = lm(brain ~ body, data=mammals)
y_hat <- result$fitted.values
res <- result$residuals
mammals <- data.frame(mammals, y_hat, res)

ggplot(mammals, aes(x=y_hat, y=res))+
  geom_point()+
  geom_hline(yintercept = 0, color='red')+
  labs(x="Fitted y", y="Residuals", title= "Residual plot before y transformation")
```




```
boxcox(result)
```




```
boxcox(result, lambda = seq(-0.1, 0.4, 1/10) )
```

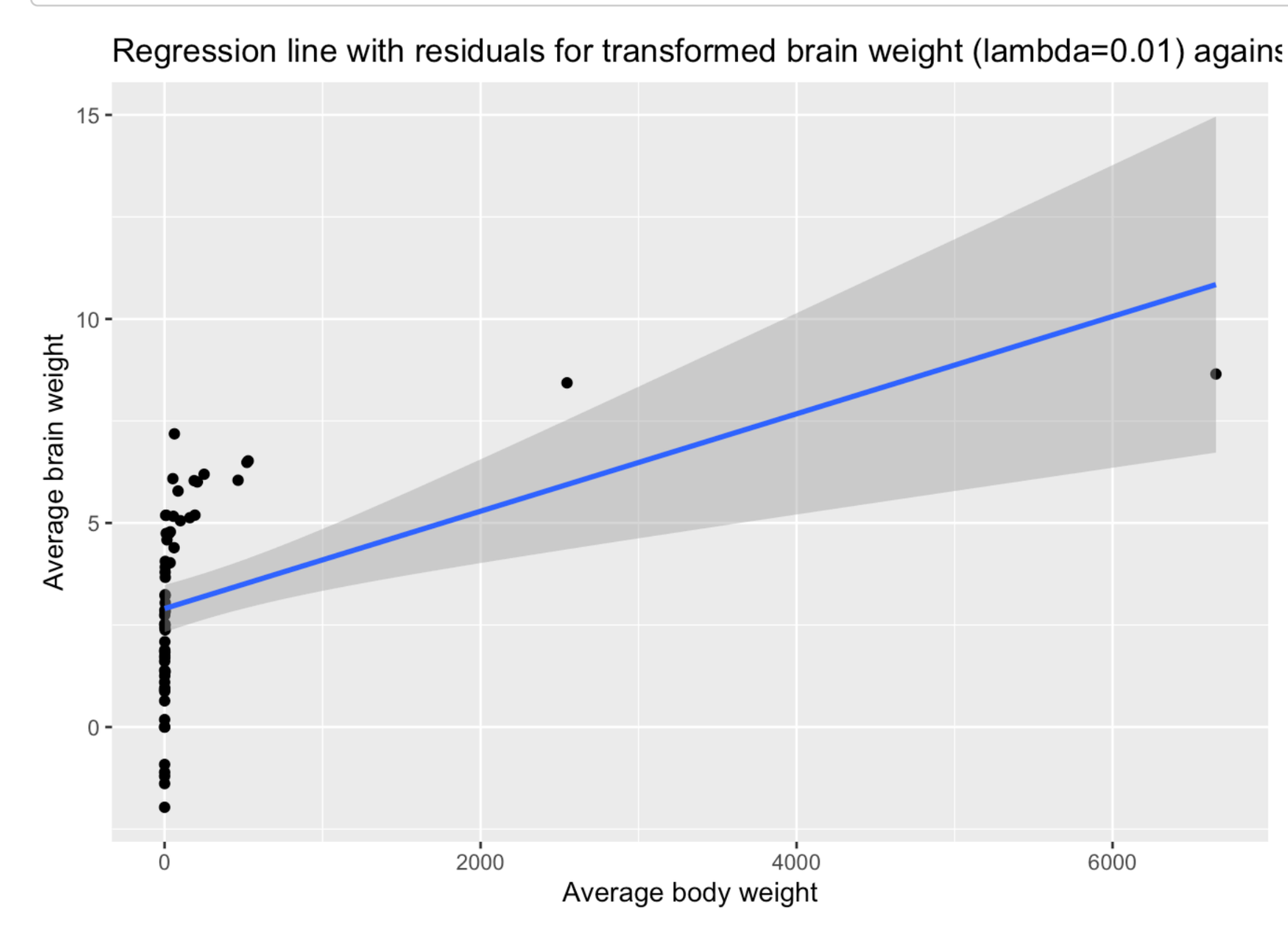


From Box Cox above, lambda of 0.08 seems like an optimal choice, for siplification purposes can try lambda=0.1 as well.

5. Apply the transformation you specified in part 4, and let y_* denote the transformed response variable. Create a scatterplot of y_* against x. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

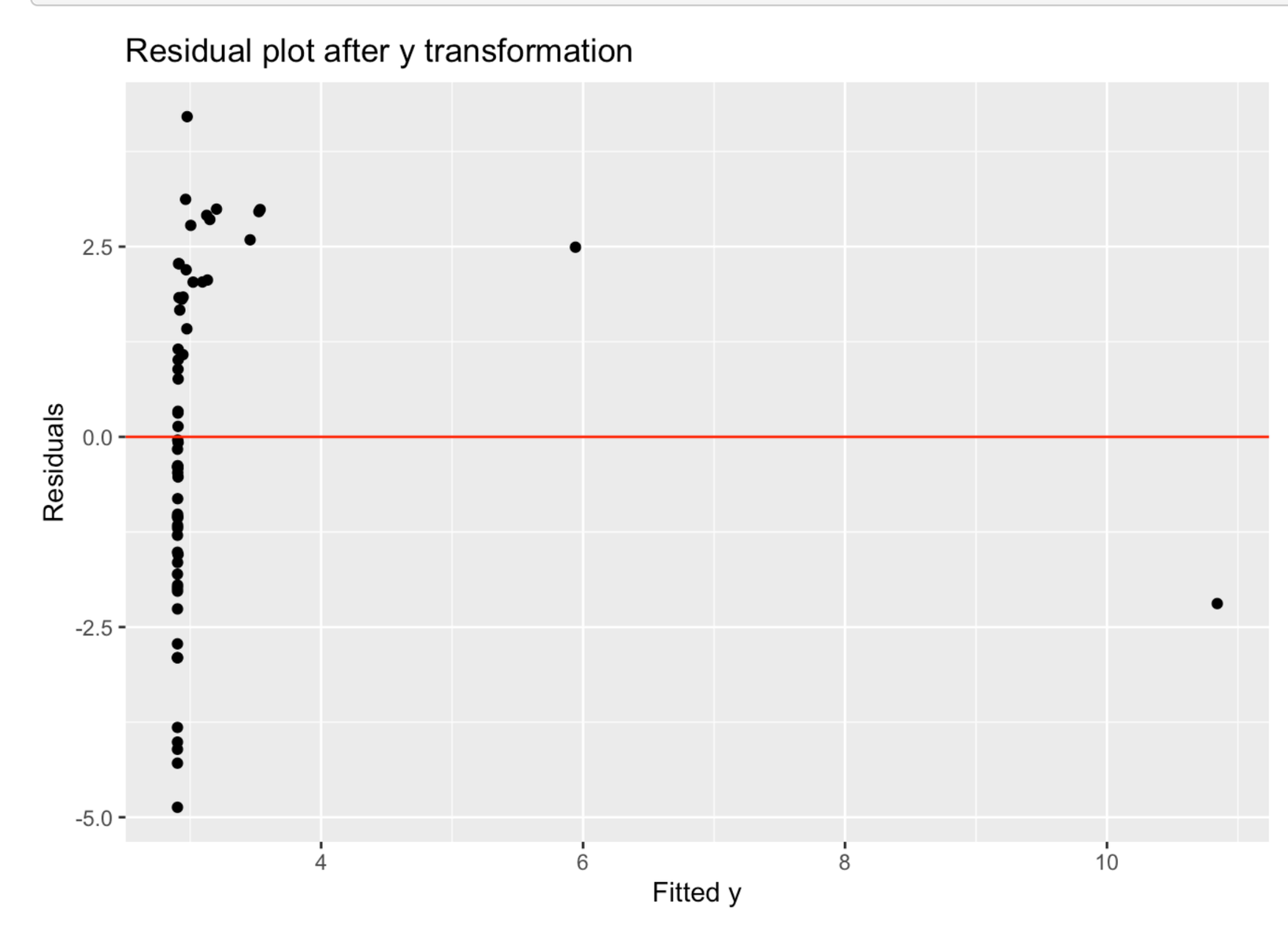
```
y_star <- log (mammals$brain)
mammals <- data.frame(mammals, y_star)

ggplot(mammals, aes(x=body, y=y_star))+
  geom_point()+
  geom_smooth(method = "lm", se = TRUE)+
  labs(x="Average body weight", y="Average brain weight",
       title="Regression line with residuals for transformed brain weight (lambda=0.01) against body weight of la
nd mammals")
```

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



6. Fit a simple linear regression to y_* against x, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

```
result.y_star <- lm(y_star ~ body, data=mammals)

y_hat2 <- result.y_star$fitted.values
res2 <- result.y_star$residuals

ggplot(mammals, aes(x=y_hat2, y=res2))+
  geom_point()+
  geom_hline(yintercept = 0, color='red')+
  labs(x="Fitted y", y="Residuals", title= "Residual plot after y transformation")
```



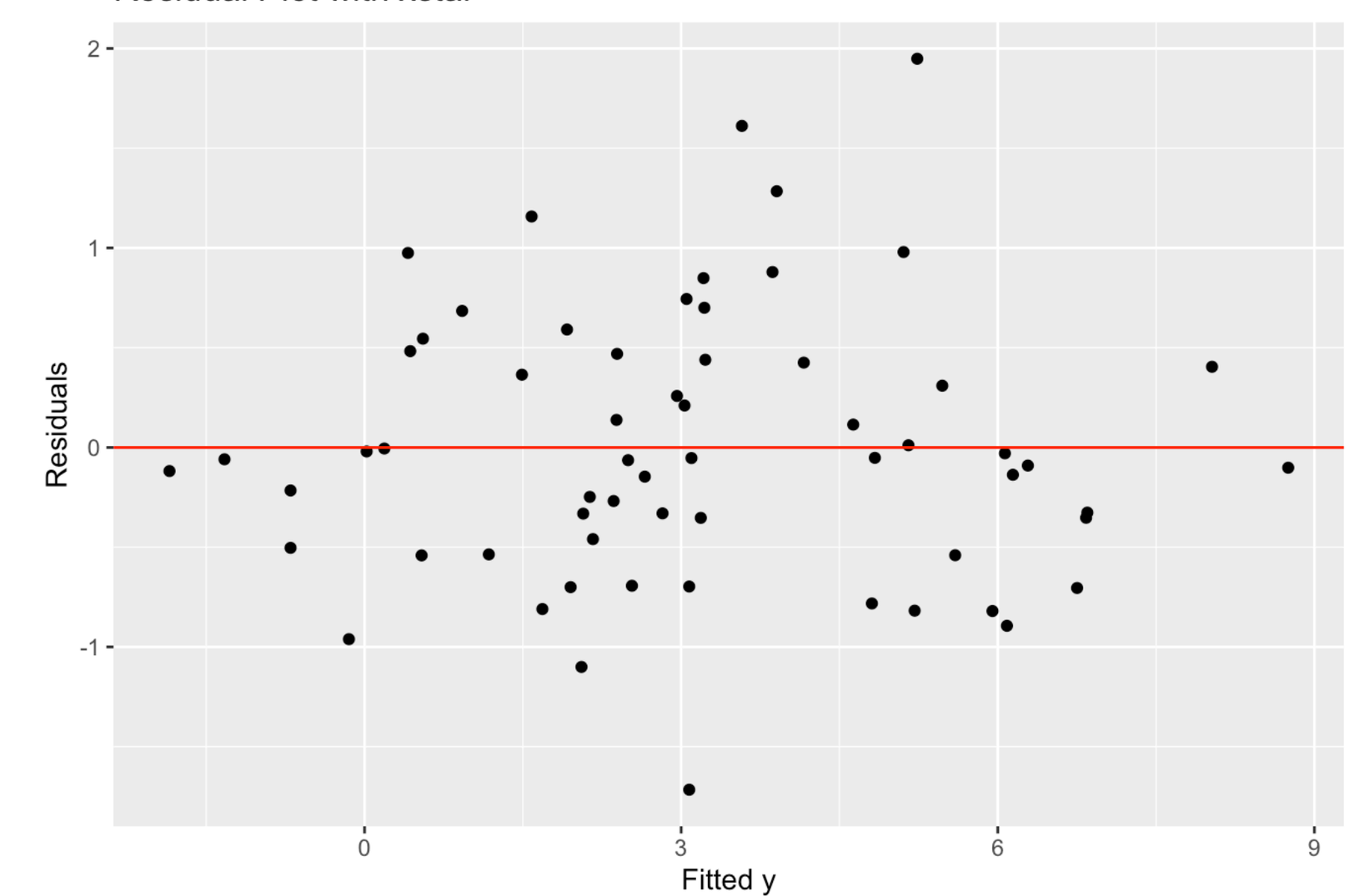
Observation: probably we also have issue #1, mean=0 does not hold true as well. Checking with Box Cox once again to make sure no further issue #2 transformations are warranted.

```
#boxcox(result.y_star)
```

7. Do we need to transform the x variable? If yes, what transformation(s) would you try? ** Briefly explain. Create a scatterplot of y against x . Do any assumptions for simple linear regression appear to be violated? If so, which ones?

May be issue #1 now? => transform x

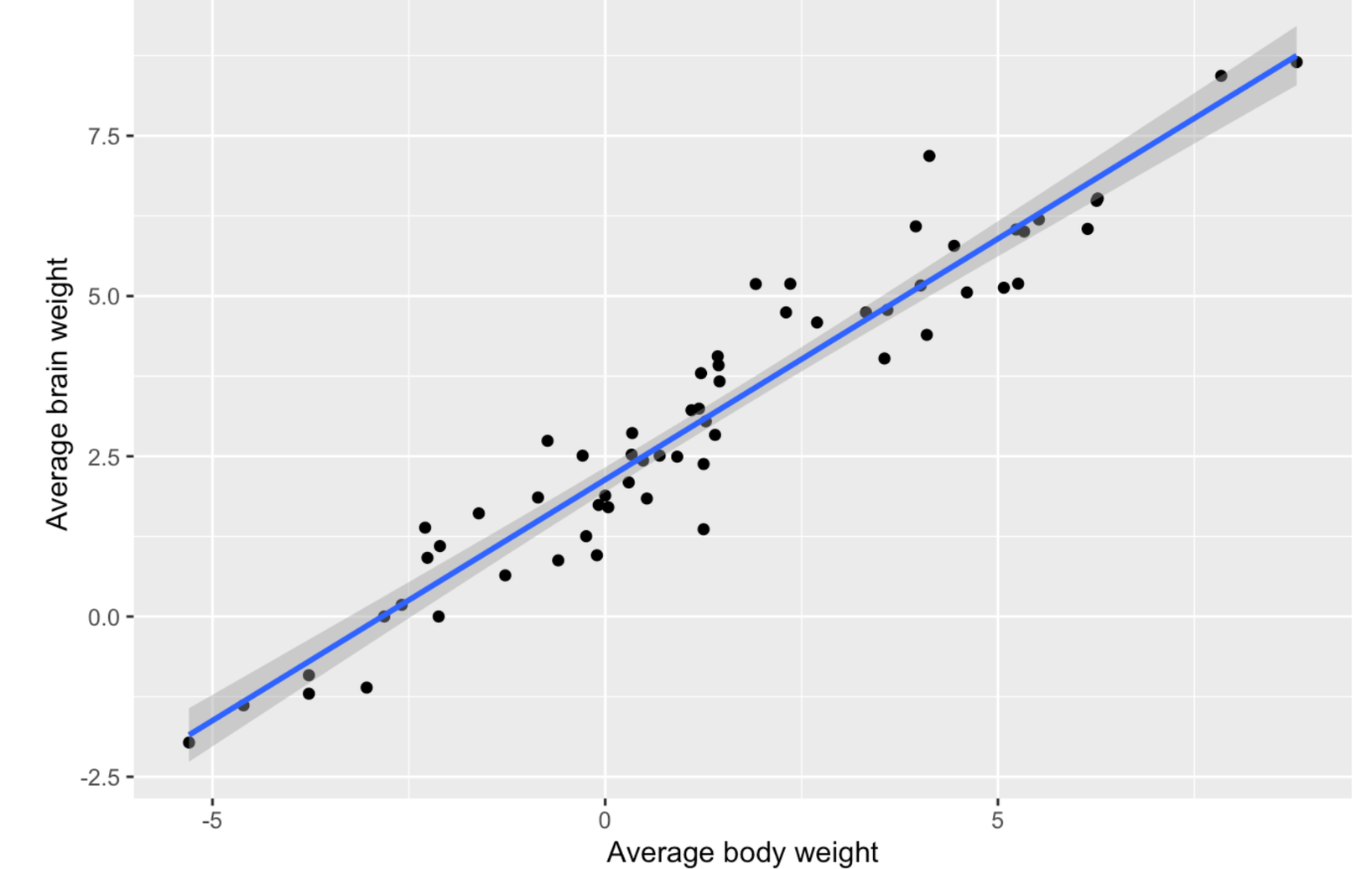
```
x_star<-log(mammals$body)
mammals<-data.frame(mammals,x_star)

result.x_star<-lm(y_star ~ x_star, data=mammals)

y_hat3<-result.x_star$fitted.values
res3<-result.x_star$residuals
mammals<-data.frame(mammals,y_hat3,res3)

##residual plot with xstar
ggplot(mammals, aes(x=y_hat3,y=res3))+
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted y", y="Residuals", title="Residual Plot with xstar")
```



8. Fit a simple linear regression to y_* against x_*, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones? If the assumptions are not met, repeat with a different transformation on the predictor until you are satisfied.

```
ggplot(mammals, aes(x=x_star, y=y_star))+
  geom_point()+
  geom_smooth(method = "lm", se = TRUE)+
  labs(x="Average body weight", y="Average brain weight",
       title="Regression line with residuals for transformed brain weight (lambda=0.01) against transformed body
weight of land mammals")
```

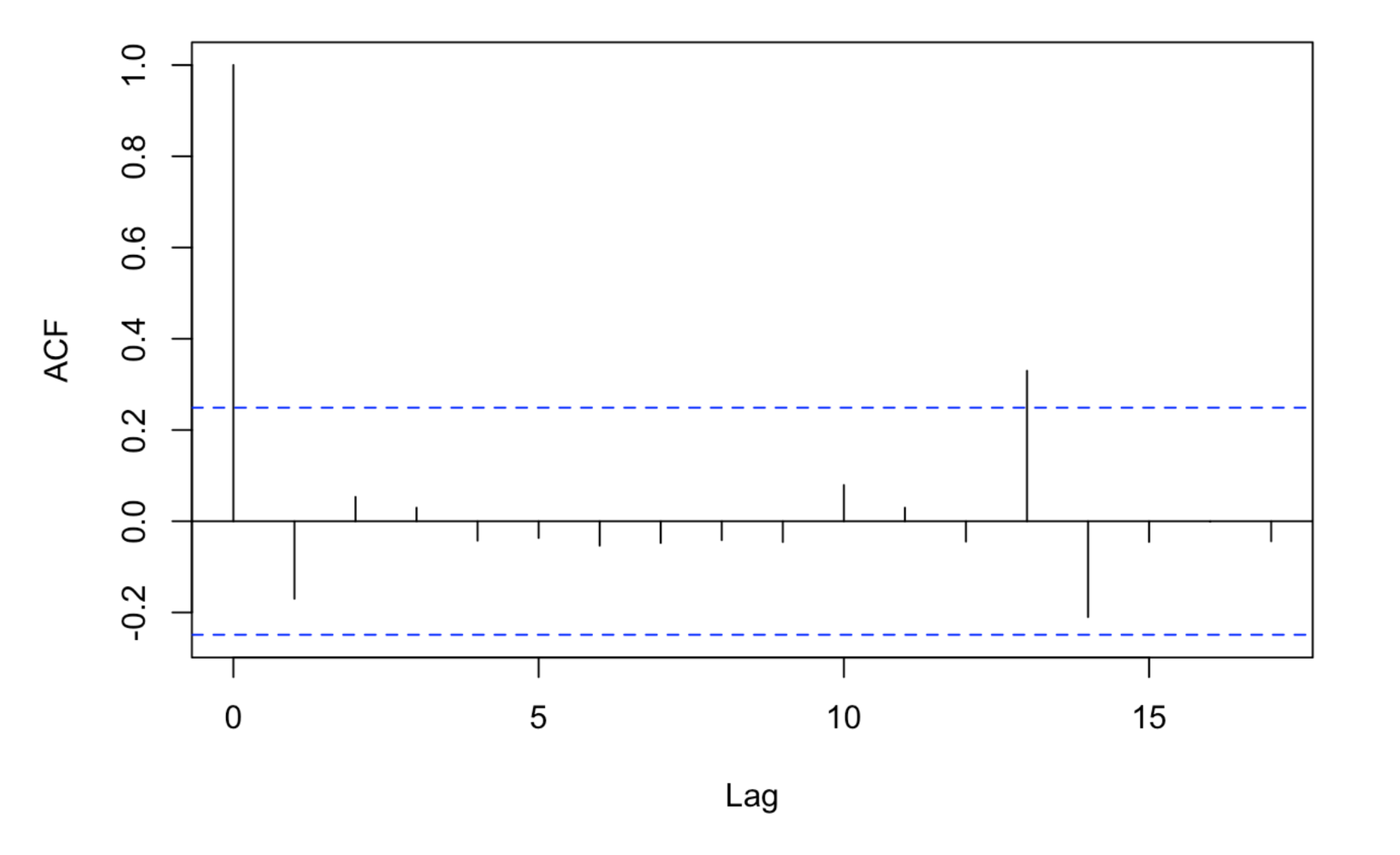
```
## `geom_smooth()` using formula 'y ~ x'
```



9. Create an ACF plot of the residuals. Comment if assumptions are met for linear regression.

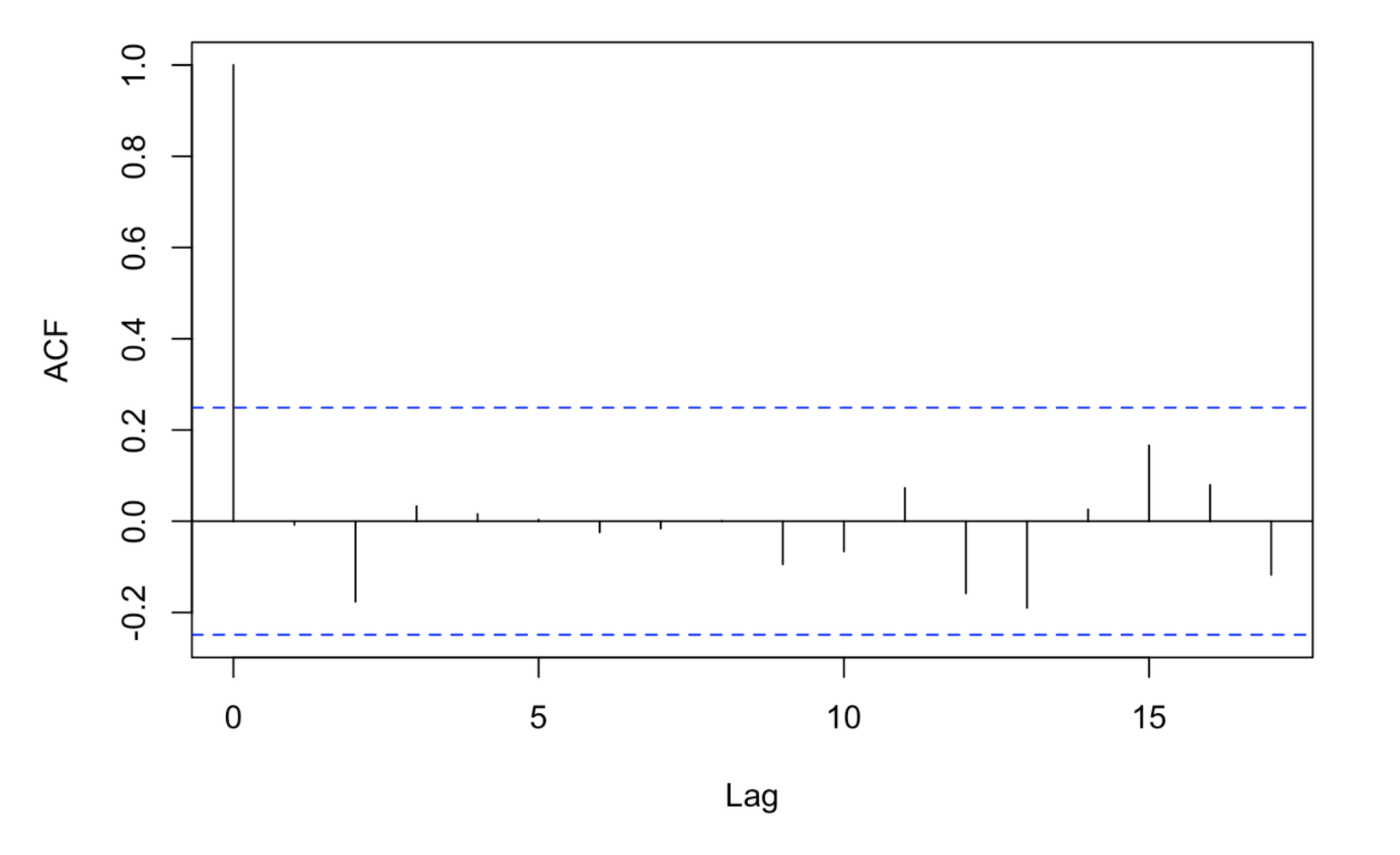
```
acf(res, main="ACF Plot of Residuals with origiinal x")
```

ACF Plot of Residuals with origiinal x




```
acf(res3, main="ACF Plot of Residuals with xstar")
```

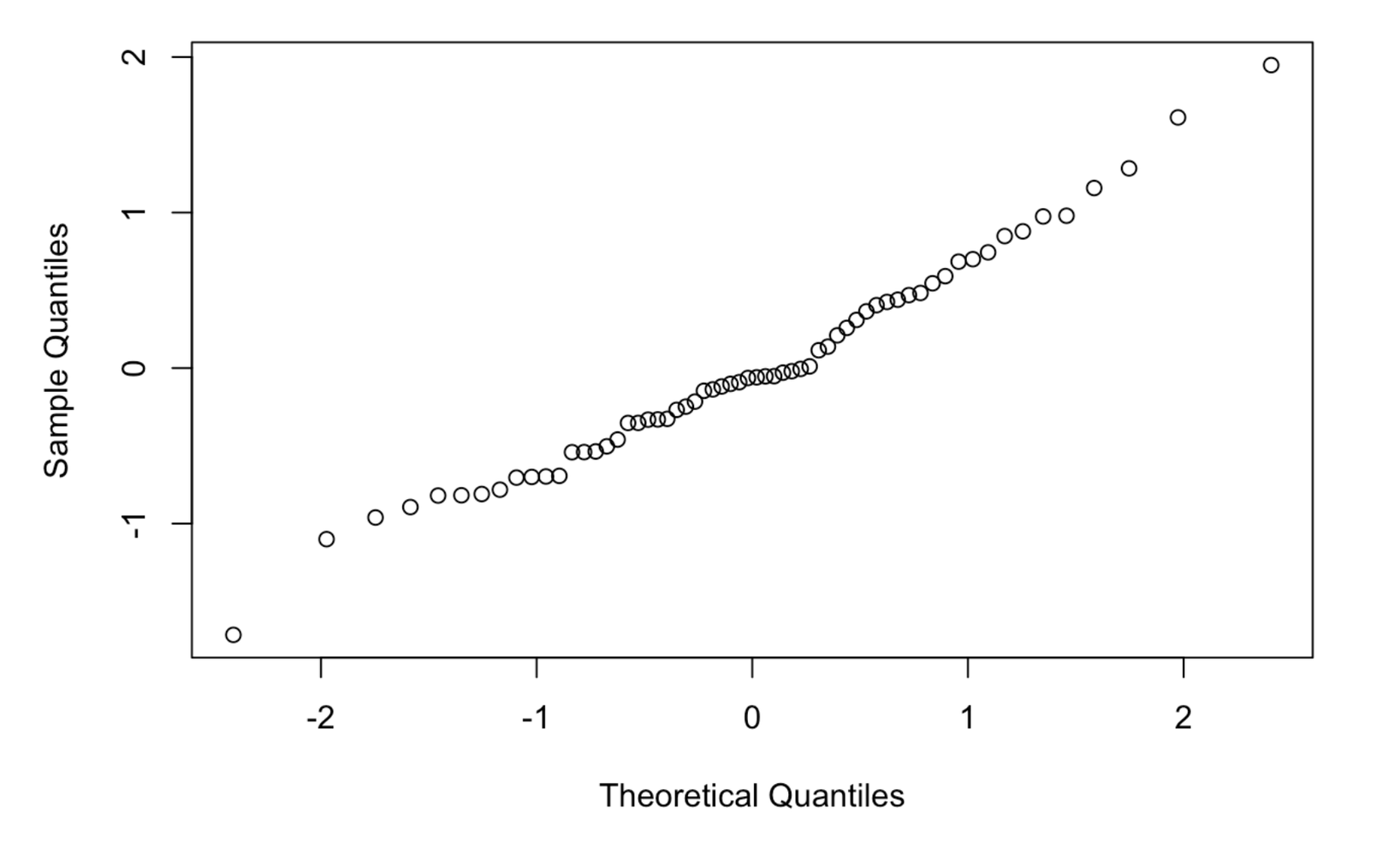
ACF Plot of Residuals with xstar



10. Create a QQ plot of the residuals. Comment if assumptions are met for linear regres- sion.

```
qqnorm(res3)
```

Normal Q-Q Plot




```
#qqline(res3, col="red")
```

11. Write out the regression equation, and if possible, interpret the slope of the regression.

```
final_result <- lm(formula = y_star ~ x_star, data = mammals)
summary(final_result)
```

```
##
## Call:
## lm(formula = y_star ~ x_star, data = mammals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71550 -0.49228 -0.06162  0.43597  1.94829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.13479    0.09604   22.23   <2e-16 ***
## x_star       0.75169    0.02846   26.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6943 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Model: y_star = 2.13 + 0.75 * x_star, where y_star = log(brain)m x_star = log(mass)

Since both variables were log transformed, we interpret the slope of 0.75 as, for a 1% increase in body weight, the weight of the brain increases by approximately 0.75%. We note that based on the residual plot, ACF plot of residuals, and QQ plot of residuals in parts 8, 9, and 10, the assumptions for this regression model are met.