Stat 6021: Addressing Guided Question Set for Module 5: Model Diagnostics and Remedial Measures in SLR

Tom Lever

09/24/22

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.

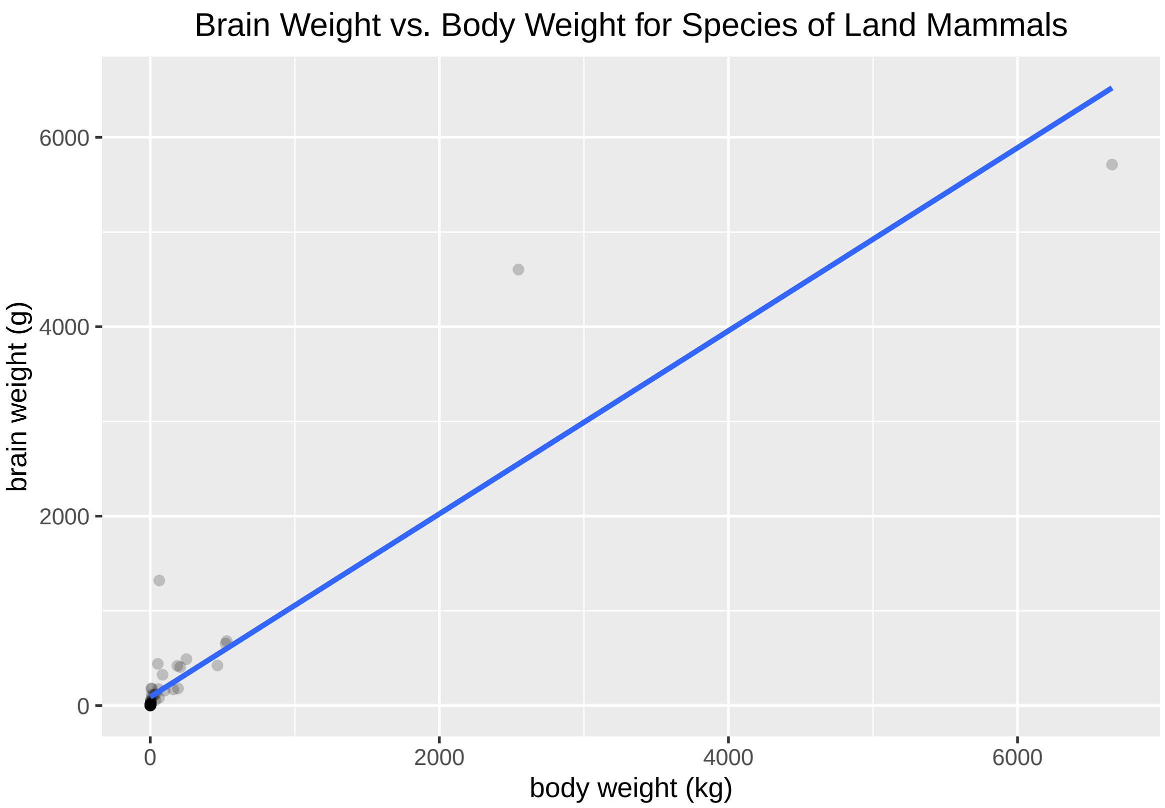
1. Create a scatter plot of brain weight against body weight of land mammals. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

```
library(MASS)
library(dplyr)
library(ggplot2)
library(TomLeversRPackage)

species_body_weight_and_brain_weight <-
    MASS::mammals %>% rename(body_weight = body, brain_weight = brain)
head(species_body_weight_and_brain_weight, n = 3)
```

```
##                  body_weight brain_weight
## Arctic fox             3.385         44.5
## Owl monkey             0.480         15.5
## Mountain beaver        1.350          8.1
```

```
data_set <- species_body_weight_and_brain_weight
plot(
    ggplot(data_set, aes(x = body_weight, y = brain_weight)) +
        geom_point(alpha = 0.2) +
        geom_smooth(method = "lm", se = FALSE) +
        labs(
            x = "body weight (kg)",
            y = "brain weight (g)",
            title = "Brain Weight vs. Body Weight for Species of Land Mammals"
        ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
)
```



The relationship between response / brain weight y and predictor / regressor / body weight x appears nonlinear.

Assumptions for simple linear regression appear to be violated.

a. The assumption that the relationship between response / brain weight y and predictor / regressor / body weight x is linear, at least approximately, is not met. The relationship appears to be nonlinear. Residuals are not evenly scattered aroun $e = 0$. A Box-Cox plot suggests a maximum-likelihood estimate of parameter λ around 0.
b. The assumption that the error term ϵ of the linear model has mean 0 is not met. Observations are not scattered evenly around the fitted line. Residuals are not evenly scattered around $e = 0$.
c. The assumption that the error term ϵ of the linear model has constant variance is not met. The vertical variation of observations is not constant. Residuals do not have similar vertical variation across $e = 0$.
d. The assumption that the errors ϵ_i / residuals e_i are uncorrelated is not met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since the ACF value for lag 13 is significant, we have sufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We have sufficient evidence to conclude that the residuals of the linear model are correlated. We have sufficient evidence to conclude that the assumption that the errors ϵ_i / residuals e_i are uncorrelated is not met.
e. Assumptions that the errors ϵ_i / residuals e_i are normally distributed is not met. A linear model is robust to these assumptions. Given sharp downward and upward curves at extremes of a plot of samples quantiles versus theoretical quantiles for the residuals of the linear model, the tails of the probability vs. externally studentized residuals plot / distribution are too light for this distribution to be considered normal. The assumption that the errors ϵ_i / residuals e_i are normally distributed is not met.

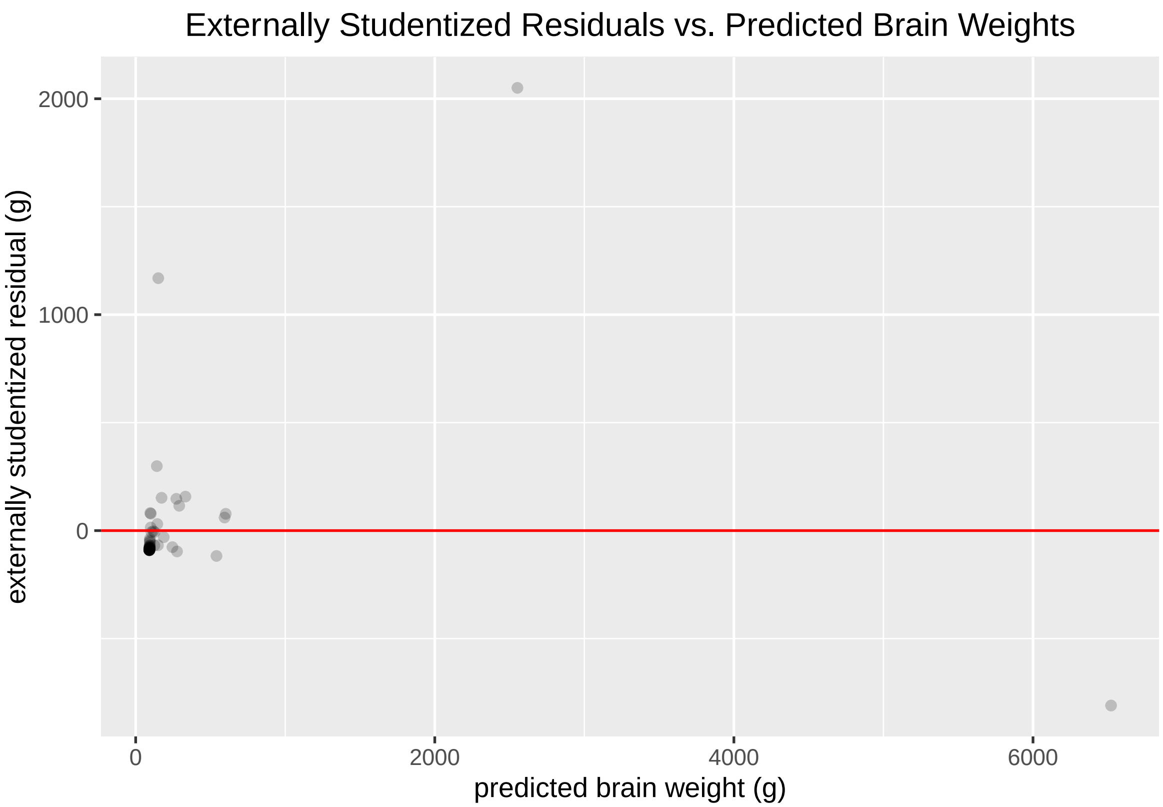
2. Fit a simple linear regression to the data, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

See above analysis.

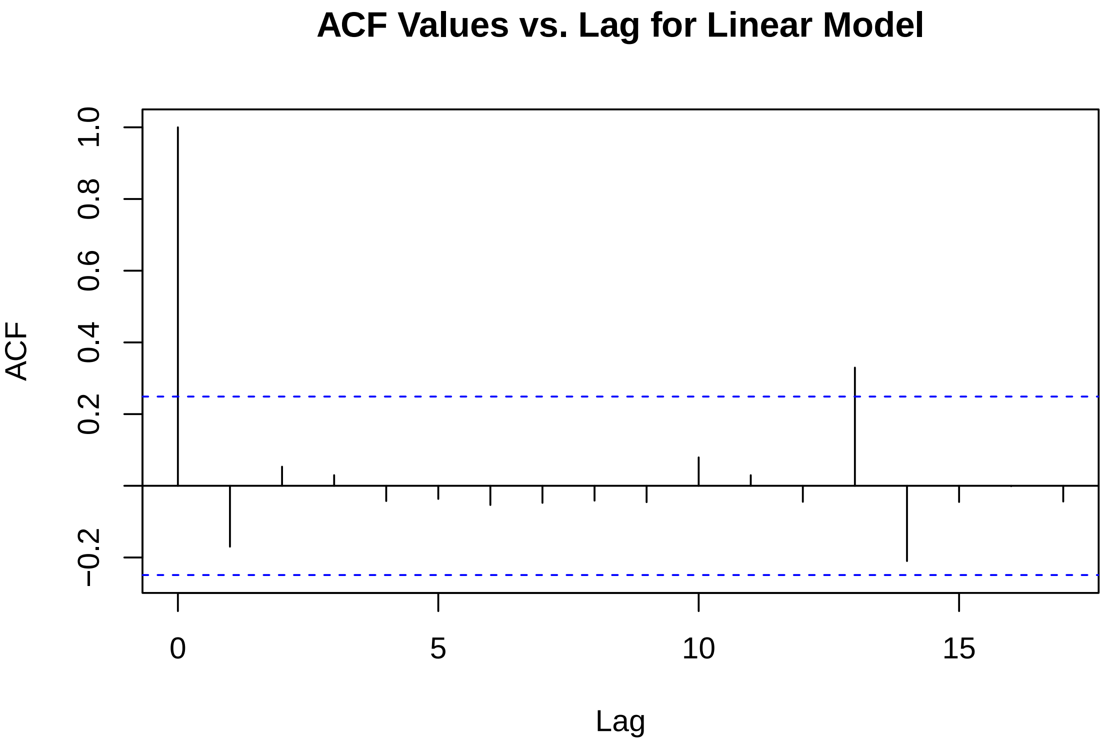
```
linear_model <- lm(brain_weight ~ body_weight, data = data_set)
print(summarize_linear_model(linear_model))
```

```
##
## Call:
## lm(formula = brain_weight ~ body_weight, data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -810.07  -88.52  -79.64  -13.02 2050.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 91.00440   43.55258    2.09   0.0409 *
## body_weight  0.96650    0.04766   20.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 334.7 on 60 degrees of freedom
## Multiple R-squared:  0.8727, Adjusted R-squared:  0.8705
## F-statistic: 411.2 on 1 and 60 DF,  p-value: < 2.2e-16
##
## E(y | x) = B_0 + B_1 * x = 91.0043962074069 + 0.966496367672576 * x
## Number of observations: 62
## Estimated variance of errors: 112037.31759175
## Multiple R:  0.934163842323354   Adjusted R:  0.933027215953214
```

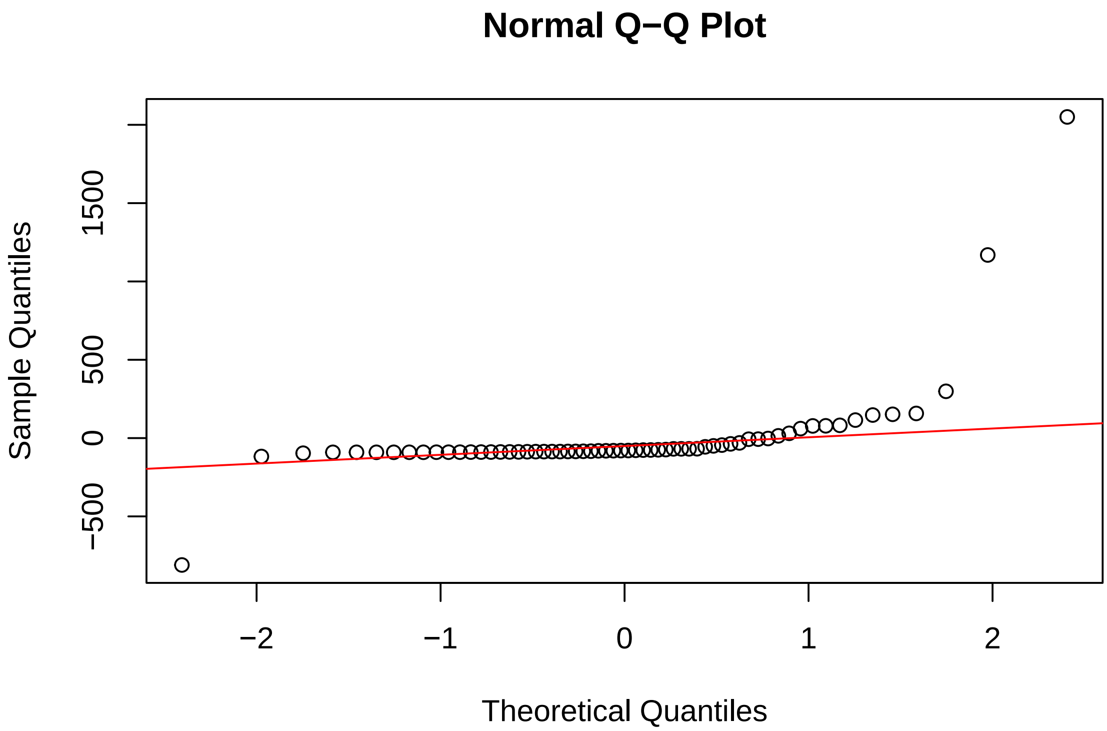
```
ggplot(
    data.frame(
        externally_studentized_residuals = linear_model$residuals,
        predicted_brain_weights = linear_model$fitted.values
    ),
    aes(x = predicted_brain_weights, y = externally_studentized_residuals)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted brain weight (g)",
        y = "externally studentized residual (g)",
        title = "Externally Studentized Residuals vs. Predicted Brain Weights"
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```



```
acf(linear_model$residuals, main = "ACF Values vs. Lag for Linear Model")
```



```
qqnorm(linear_model$residuals)
qqline(linear_model$residuals, col = "red")
```

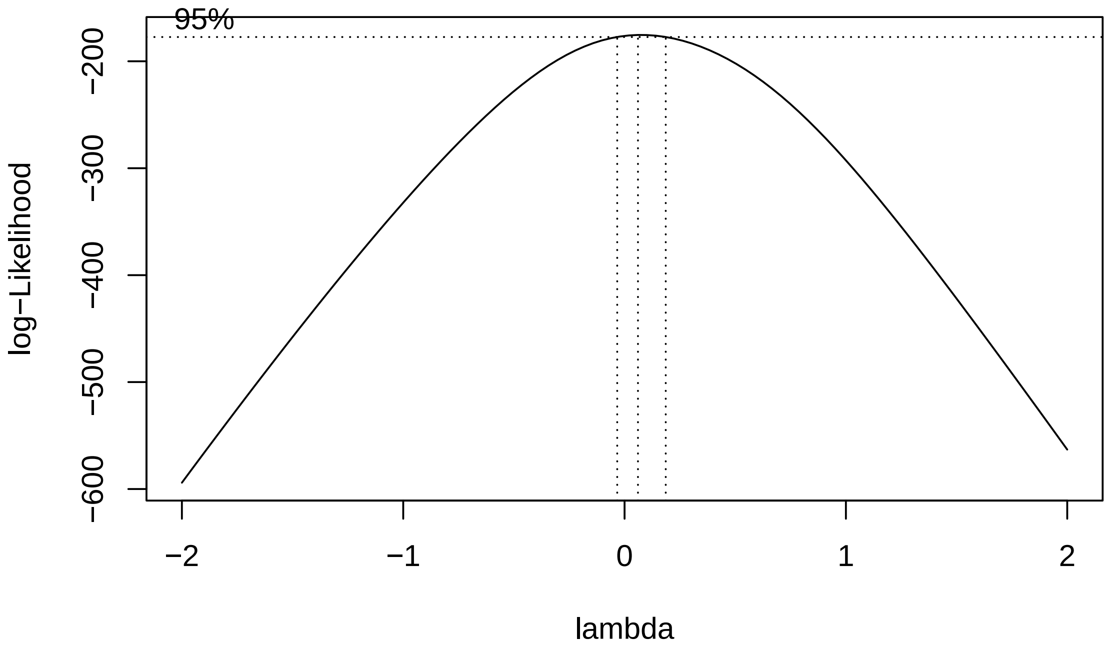


3. Based on your answers to parts 1 and 2, do we need to transform at least one of the variables?

   Yes. Because the assumption that the error term ϵ of the linear model has constant variance is not met, we will transform estimated response / brain weight $\hat{y}$. Because the assumption that the error term ϵ of the linear model has mean 0 is not met, we will transform predictor / body weight x after transforming estimated response / brain weight $\hat{y}$ if necessary.

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?

```
result_of_Box_Cox_Method <- perform_Box_Cox_Method(linear_model)
```



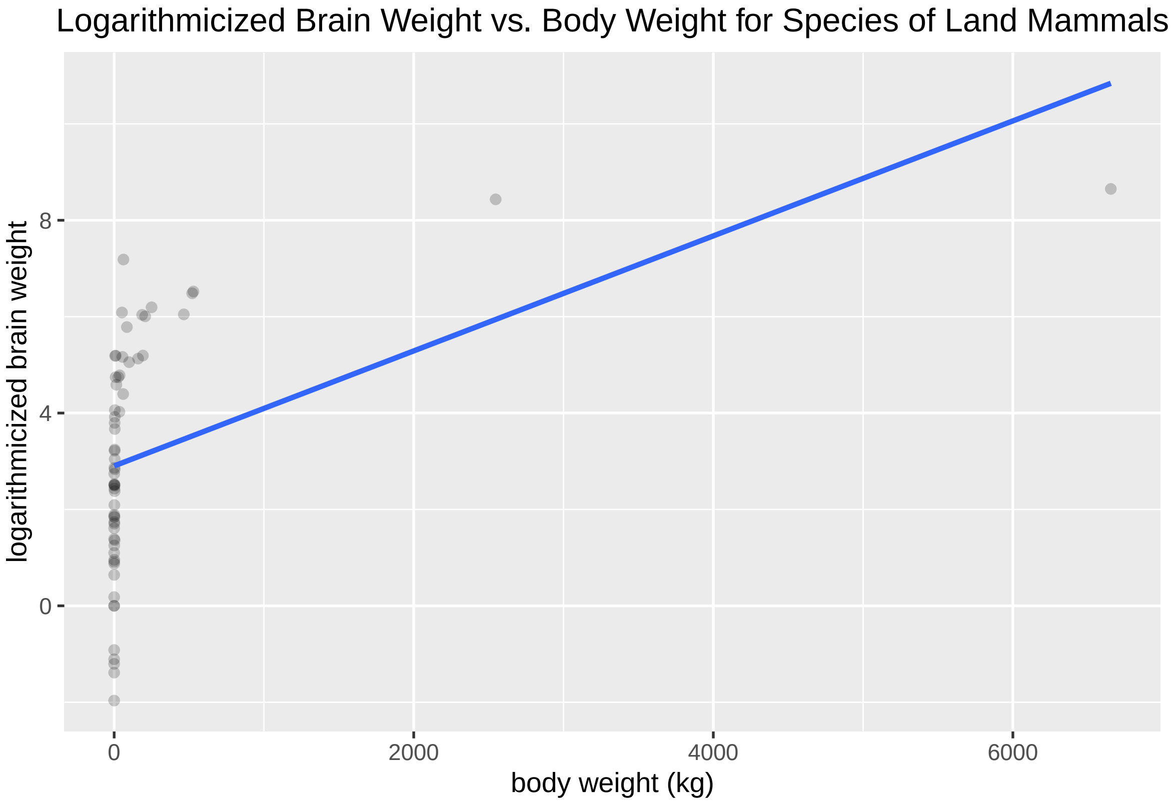
```
print(result_of_Box_Cox_Method$maximum_likelihood_estimate_of_parameter_lambda)
```

```
## [1] 0.06060606
```

For the set of observations of brain weight and body weight, the maximum-likelihood estimate of λ is close to parameter $\lambda = 0$. Since parameter λ is close to 0, we may use the transformation $\hat{y}' = ln(\hat{y})$.

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?

```
data_set <- data_set %>% mutate(logarithmicized_brain_weight = log(brain_weight))
ggplot(data_set, aes(x = body_weight, y = logarithmicized_brain_weight)) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "body weight (kg)",
        y = "logarithmicized brain weight",
        title = paste(
            "Logarithmicized Brain Weight ",
            "vs. Body Weight for Species of Land Mammals",
            sep = ""
        )
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```



Assumptions for simple linear regression appear to be violated.

a. The assumption that the relationship between response / brain weight y and predictor / regressor / body weight x is linear, at least approximately, is not met. The relationship appears to be nonlinear. Residuals are not evenly scattered aroun $e = 0$.
b. The assumption that the error term ϵ of the linear model has mean 0 is not met. Observations are not scattered evenly around the fitted line. Residuals are not evenly scattered around $e = 0$.
c. The assumption that the error term ϵ of the linear model has constant variance is not met. The vertical variation of observations is not constant. Residuals do not have similar vertical variation across $e = 0$.
d. The assumption that the errors ϵ_i / residuals e_i are uncorrelated is not met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since the ACF value for lag 13 is significant, we have sufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We have sufficient evidence to conclude that the residuals of the linear model are correlated. We have sufficient evidence to conclude that the assumption that the errors ϵ_i / residuals e_i are uncorrelated is not met.
e. Assumptions that the errors ϵ_i / residuals e_i are normally distributed is not met. A linear model is robust to these assumptions. Given sharp downward and upward curves at extremes of a plot of samples quantiles versus theoretical quantiles for the residuals of the linear model, the tails of the probability vs. externally studentized residuals plot / distribution are too light for this distribution to be considered normal. The assumption that the errors ϵ_i / residuals e_i are normally distributed is not met.

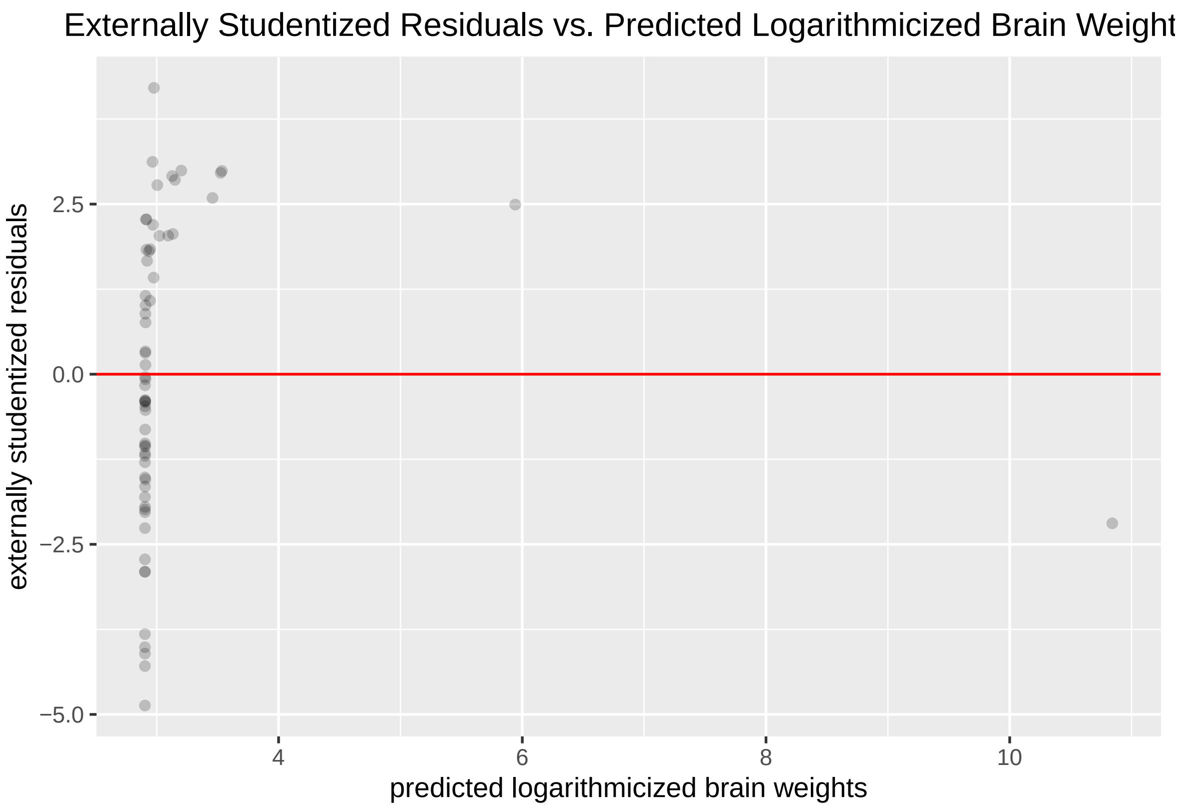
6. Fit a simple linear regression to $y*$ against x, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

   See above analysis.

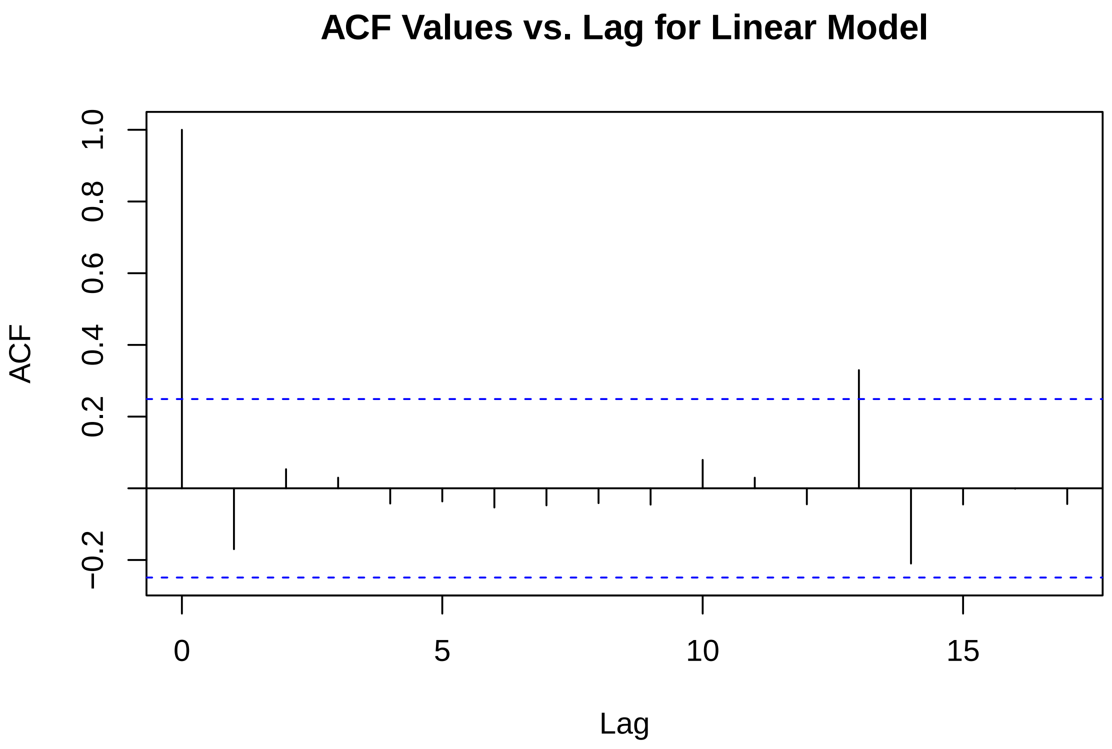
```r
semilogarithmicized_linear_model <-
    lm(logarithmicized_brain_weight ~ body_weight, data = data_set)
print(summarize_linear_model(semilogarithmicized_linear_model))
```

```
##
## Call:
## lm(formula = logarithmicized_brain_weight ~ body_weight, data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8691 -1.5389 -0.1187  1.9847  4.2084
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.9030241  0.2884703  10.064 1.68e-14 ***
## body_weight 0.0011931  0.0003157   3.779 0.000364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.217 on 60 degrees of freedom
## Multiple R-squared:  0.1923, Adjusted R-squared:  0.1788
## F-statistic: 14.28 on 1 and 60 DF,  p-value: 0.0003642
##
## E(y | x) = B_0 + B_1 * x = 2.90302410266777 + 0.00119308731468424 * x
## Number of observations: 62
## Estimated variance of errors: 4.91515305588938
## Multiple R:  0.438491217374058   Adjusted R:  0.422862219691891
```

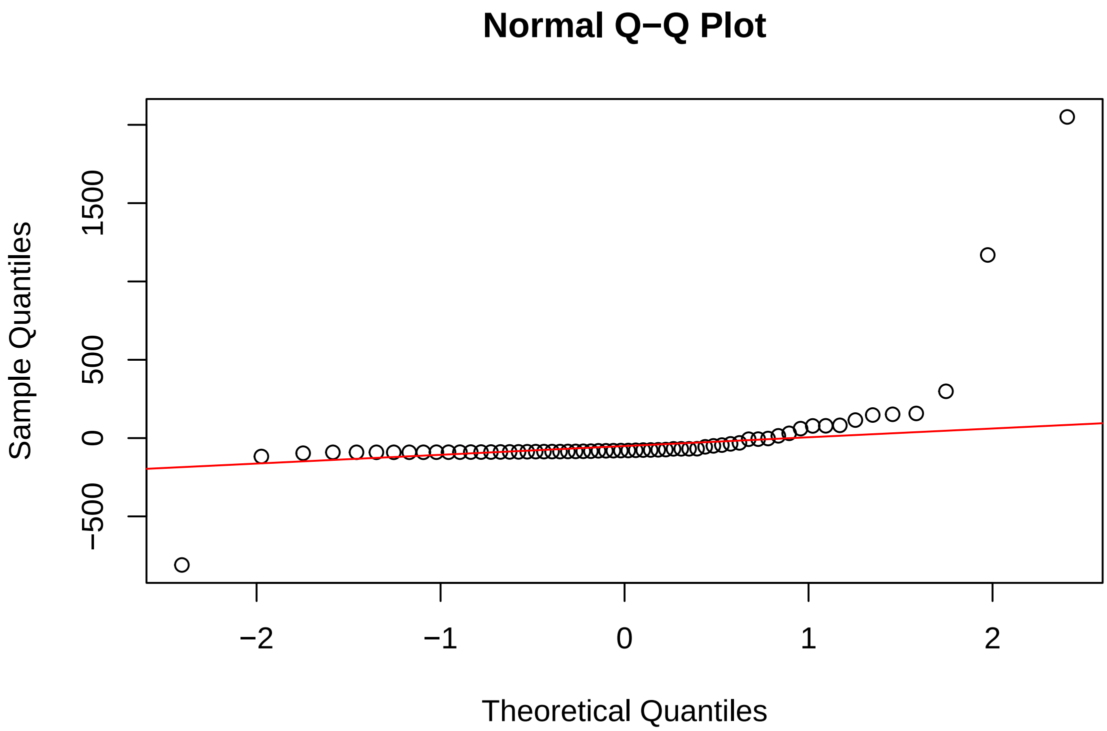
```r
ggplot(
    data.frame(
        externally_studentized_residuals =
            semilogarithmicized_linear_model$residuals,
        fitted_values = semilogarithmicized_linear_model$fitted.values
    ),
    aes(x = fitted_values, y = externally_studentized_residuals)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted logarithmicized brain weights",
        y = "externally studentized residuals",
        title = paste(
            "Externally Studentized Residuals vs. Predicted Logarithmicized Brain Weights",
            sep = ""
        )
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```



```
acf(linear_model$residuals, main = "ACF Values vs. Lag for Linear Model")
```



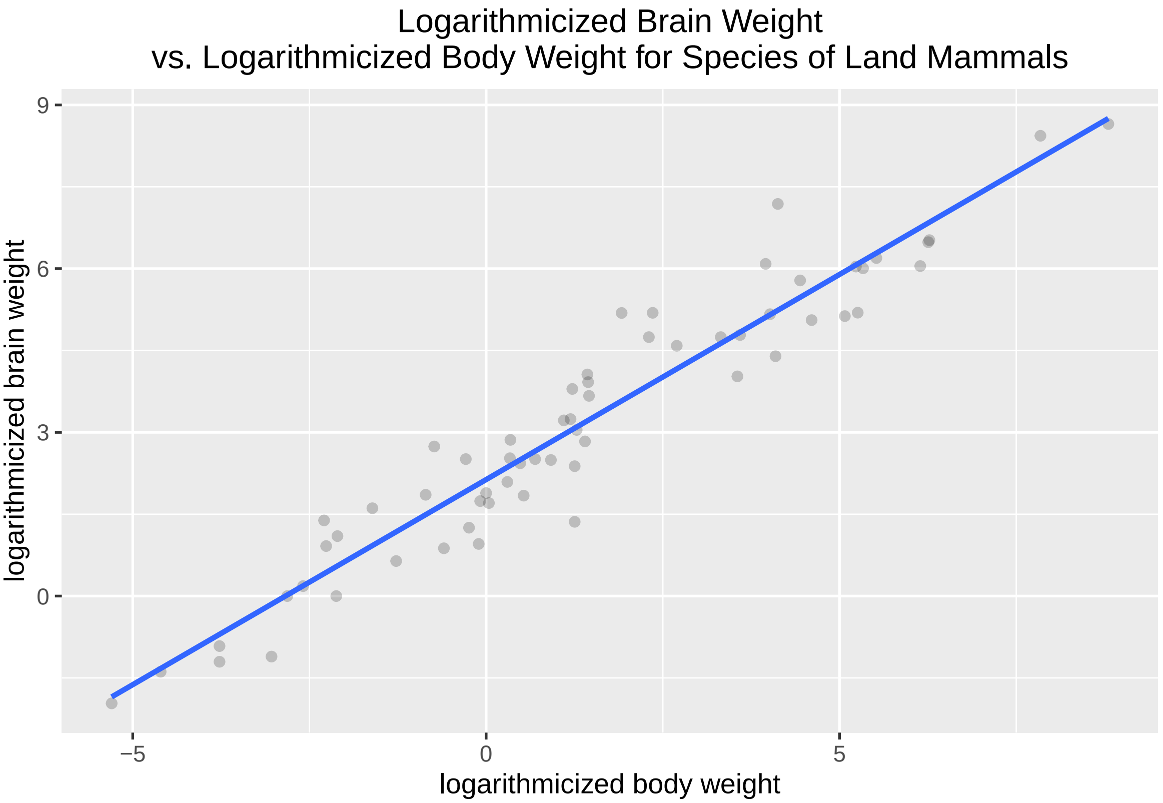
```
qqnorm(linear_model$residuals)
qqline(linear_model$residuals, col = "red")
```



7. Do we need to transform the x variable? If yes, what transformations would you try? Create a scatterplot of $y*$ versus $x*$. Do any assumptions for simple linear regression appear violated? If so, which ones?

   Yes. Logarithmicized Brain Weight vs. Body Weight for Species of Land Mammals is linearizable by logarithmicizing predictor / regressor / body weight x.

```
data_set <- data_set %>% mutate(logarithmicized_body_weight = log(body_weight))
ggplot(
    data_set,
    aes(x = logarithmicized_body_weight, y = logarithmicized_brain_weight)
) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "logarithmicized body weight",
        y = "logarithmicized brain weight",
        title = paste(
            "Logarithmicized Brain Weight\n",
            "vs. Logarithmicized Body Weight for Species of Land Mammals",
            sep = ""
        )
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```



Assumptions for simple linear regression appear to be met.

a. The assumption that the relationship between response / brain weight y and predictor / regressor / body weight x is linear, at least approximately, is met. The relationship appears to be linear. Residuals are not evenly scattered around $e = 0$.
b. The assumption that the error term ϵ of the linear model has mean 0 is met. Observations are scattered evenly around the fitted line. Residuals are evenly scattered around $e = 0$.
c. The assumption that the error term ϵ of the linear model has constant variance is met. The vertical variation of observations is constant. Residuals have similar vertical variation across $e = 0$.
d. The assumption that the errors ϵ_i / residuals e_i are uncorrelated is met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since the ACF values for all lags are insignificant, we have insufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We do not have sufficient evidence to conclude that the residuals of the linear model are correlated. We do not have sufficient evidence to conclude that the assumption that the errors ϵ_i / residuals e_i are uncorrelated is not met.
e. Assumptions that the errors ϵ_i / residuals e_i are normally distributed is met. A linear model is robust to these assumptions. Given observations in a plot of samples quantiles versus theoretical quantiles for the residuals of the linear model lie near the corresponding line of best fit, the probability vs. externally studentized residuals plot / distribution is normal. The assumption that the errors ϵ_i / residuals e_i are normally distributed is met.

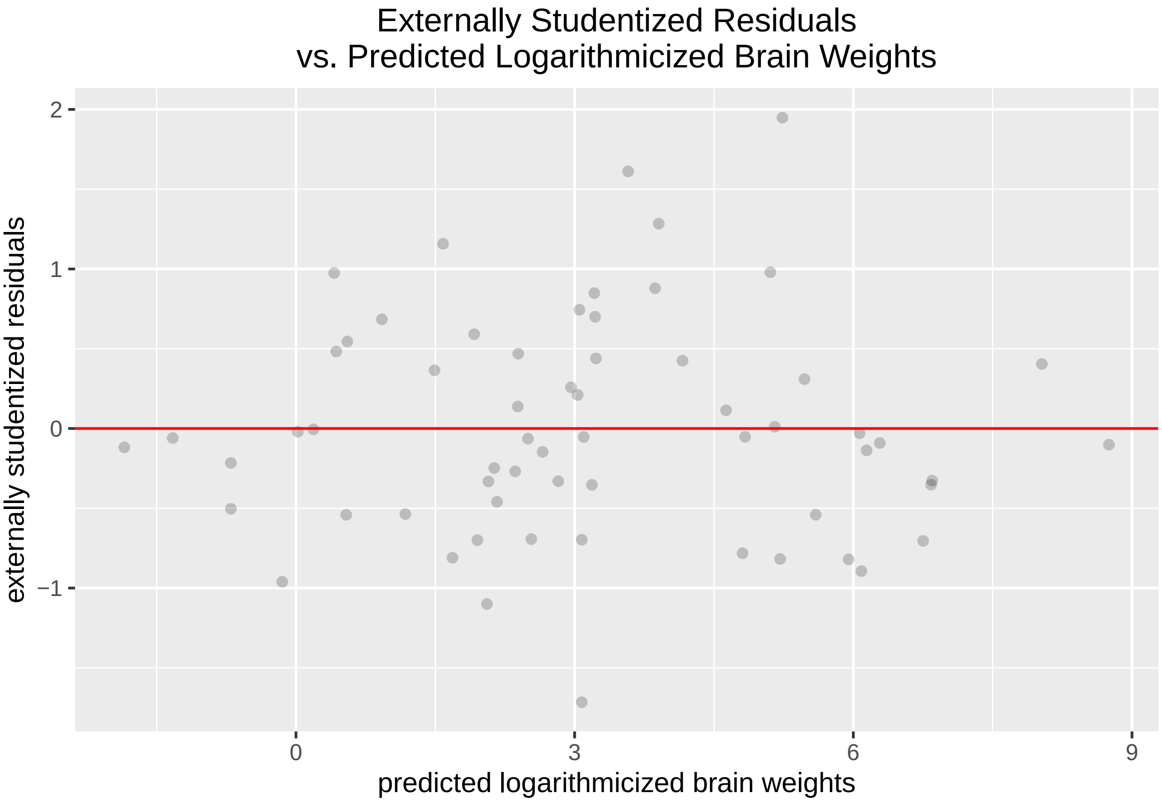
8. Fit a simple linear regression to $y*$ versus $x*$, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones? If the assumptions are not met, repeat with a different transformation on the predictor until you are satisfied.

   See above analysis.

```r
logarithmicized_linear_model <-
    lm(logarithmicized_brain_weight ~ logarithmicized_body_weight, data = data_set)
print(summarize_linear_model(logarithmicized_linear_model))
```

```
##
## Call:
## lm(formula = logarithmicized_brain_weight ~ logarithmicized_body_weight,
##     data = data_set)
##
## 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 ***
## logarithmicized_body_weight  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
##
## E(y | x) = B_0 + B_1 * x = 2.13478867676464 + 0.751685936241901 * x
## Number of observations: 62
## Estimated variance of errors: 0.482045173691011
## Multiple R:  0.95957475837098    Adjusted R:  0.958886565511151
```

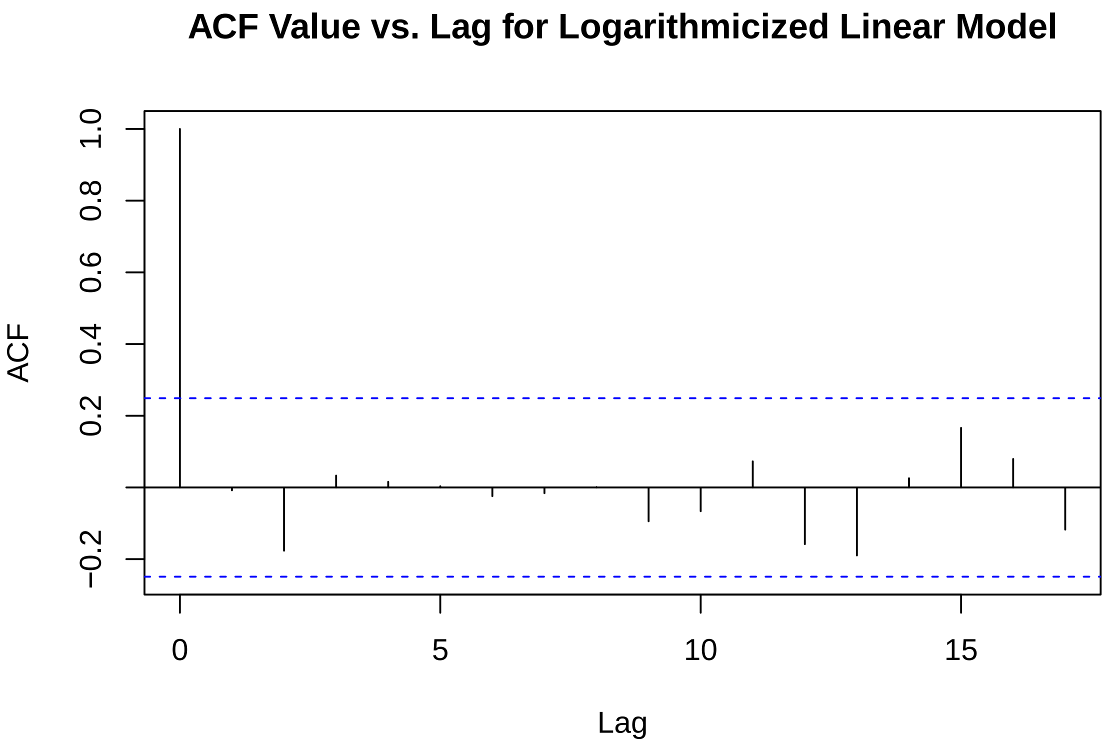
```r
ggplot(
    data.frame(
        externally_studentized_residuals = logarithmicized_linear_model$residuals,
        fitted_values = logarithmicized_linear_model$fitted.values
    ),
    aes(x = fitted_values, y = externally_studentized_residuals)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted logarithmicized brain weights",
        y = "externally studentized residuals",
        title = paste(
            "Externally Studentized Residuals\n",
            "vs. Predicted Logarithmicized Brain Weights",
            sep = ""
        )
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```



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

   See above analysis.

```
acf(
    logarithmicized_linear_model$residuals,
    main = "ACF Value vs. Lag for Logarithmicized Linear Model"
)
```



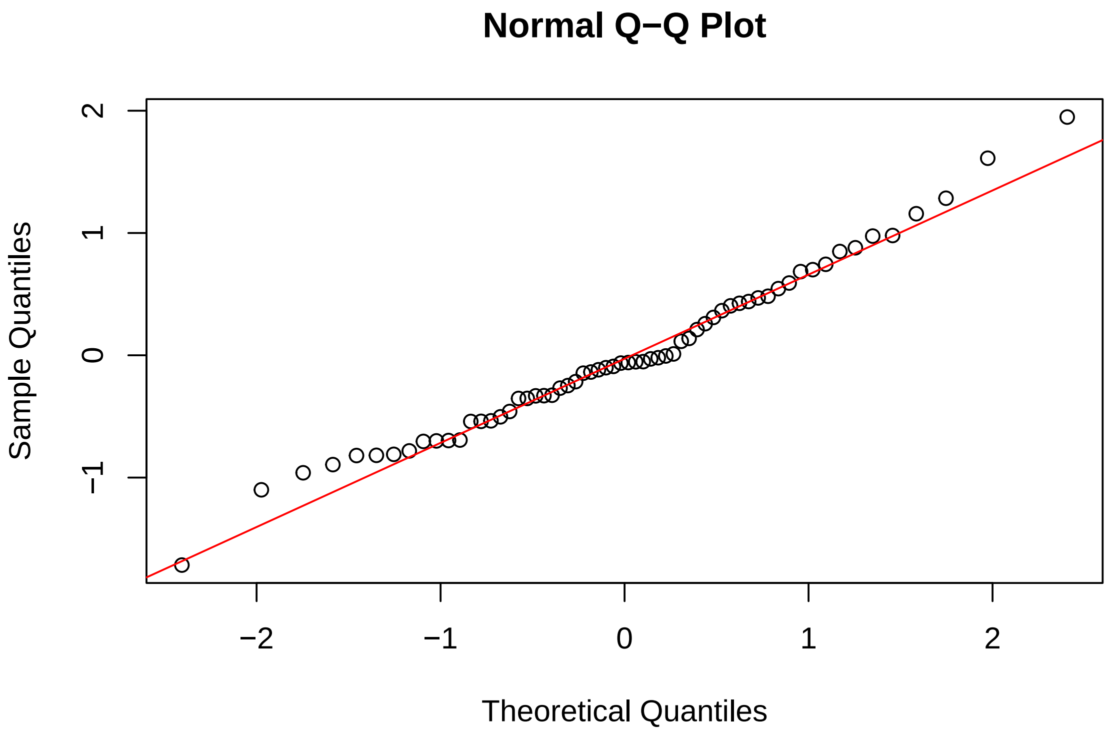
Jeffrey Woo recommends ensuring assumptions d and e are met and creating an ACF plot of residuals after we are satisfied that assumptions b and c are met.

If the errors / residuals are uncorrelated, the observations are independent.

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

See above analysis.

```
qqnorm(logarithmicized_linear_model$residuals)
qqline(logarithmicized_linear_model$residuals, col = "red")
```



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

$$\beta_0 = 2.13479$$

$$\beta_1 = 0.75169$$

$$ln(y) = \beta_0 + \beta_1 \; ln(x)$$

$$y = exp\left(\beta_0 + \beta_1 \; ln(x)\right) = exp\left(\beta_0\right) exp\left(\beta_1 \; ln(x)\right) = exp\left(\beta_0\right) exp\left(ln\left(x^{\beta_1}\right)\right) = exp\left(\beta_0\right) x^{\beta_1}$$

$$y_+ = exp\left(\beta_0\right) \left[(1+p) \; x\right]^{\beta_1}$$

$$\frac{y_+}{y} = \frac{exp\left(\beta_0\right) \left[(1+p) \; x\right]^{\beta_1}}{exp\left(\beta_0\right) x^{\beta_1}} = \frac{\left[(1+p) \; x\right]^{\beta_1}}{x^{\beta_1}} = \left[\frac{(1+p) \; x}{x}\right]^{\beta_1} = (1+p)^{\beta_1}$$

If x increases by proportion p to $x + px = (1+p)x$, y increases by a factor of $(1+p)^{\beta_1}$.