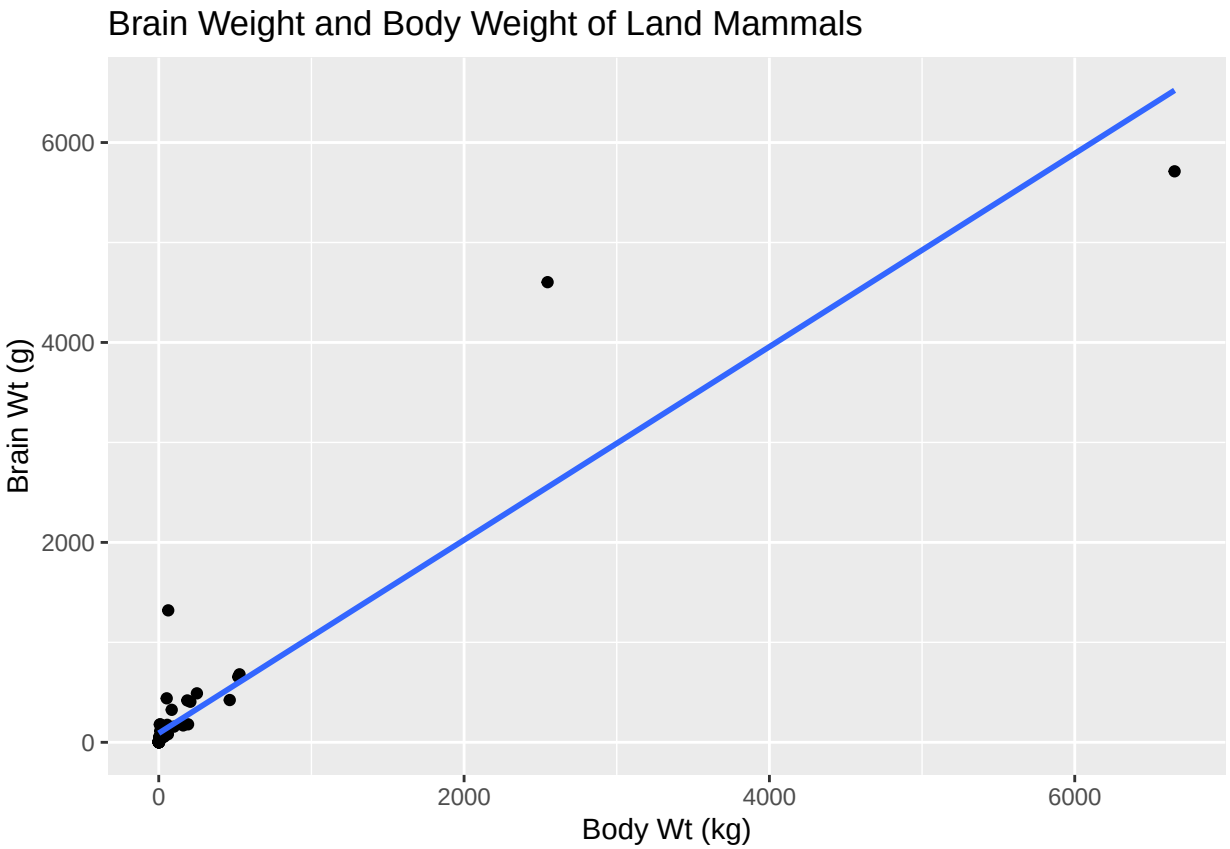
Guided Question Set 5 Solutions

```
library(tidyverse)
library(MASS)
Data<-mammals
```

1)

```
ggplot(Data, aes(x=body,y=brain))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Body Wt (kg)", y="Brain Wt (g)",
       title="Brain Weight and Body Weight of Land Mammals")
```

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



Generally, there appears to be an increasing association between body weight and brain weight of land mammals. The heavier the mammal, the heavier the brain. However, the relationship may be more logarithmic rather than linear. There are a couple of mammals with heavy body weights (around 2500kg and 6500kg) that make this distinction between a logarithmic or linear relationship difficult.

So assumption 1, that the relationship is linear, may not be met.

The heavier mammals tend to deviate further from the "regression line", so the variance may be increasing, so assumption 2, that the variance of the errors is constant, may not be met.

We create a residual plot to see if we get a clearer picture.

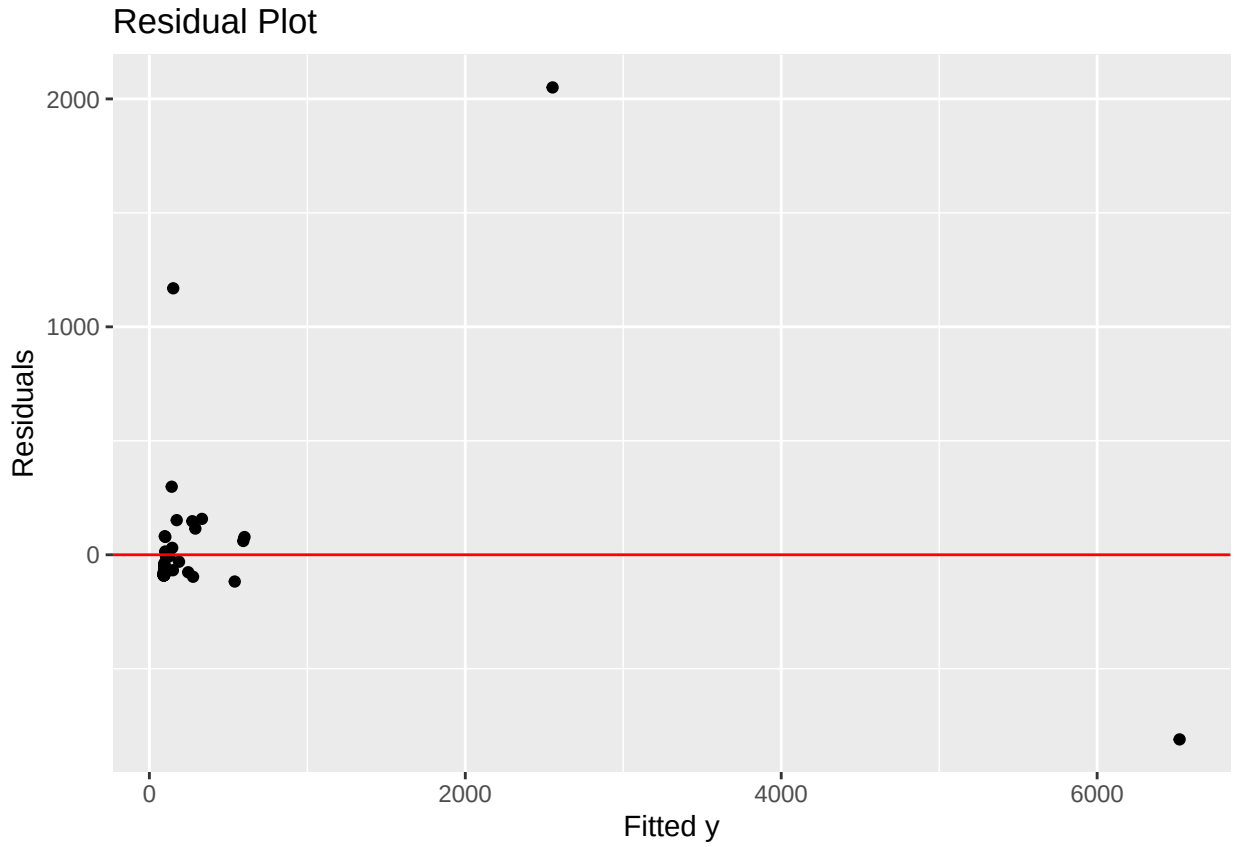
2)

```
result<-lm(brain~body, data=Data)

Data$yhat<-result$fitted.values
Data$res<-result$residuals

ggplot(Data, aes(x=yhat,y=res))+
```

```
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted y", y="Residuals", title="Residual Plot")
```



Similar to the scatterplot, this residual plot is a little difficult to assess due to the presence of two observations with large fitted values (probably the 2 heavy mammals identified earlier).

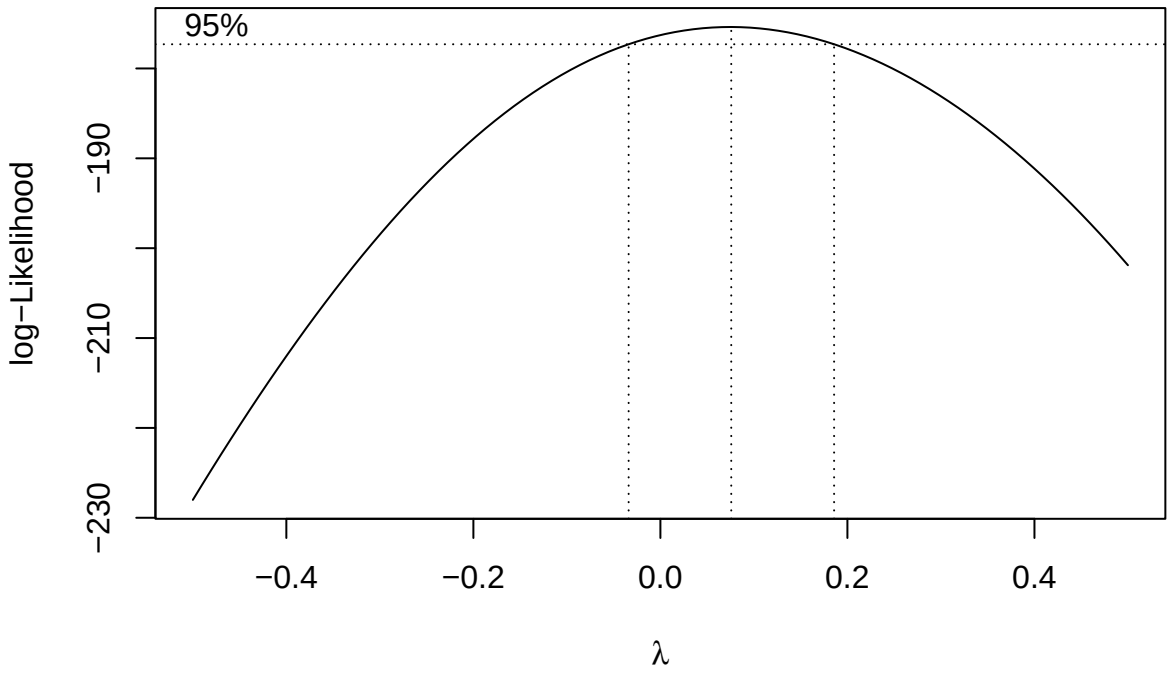
The variance is likely not be constant, as the residuals deviate further away from the horizontal axis, for larger values of $\hat{y}$.

3)

We should think about transforming the predictor, since the linearity of the relationship is questionable. The plots were difficult to evaluate if the constant variance assumption was met. Looking at the Box-Cox plot may help us decide if we need to transform the response variable (and if so, suggests the variance is not constant).

4)

```
boxcox(result, lambda=seq(-0.5,0.5,0.1))
```



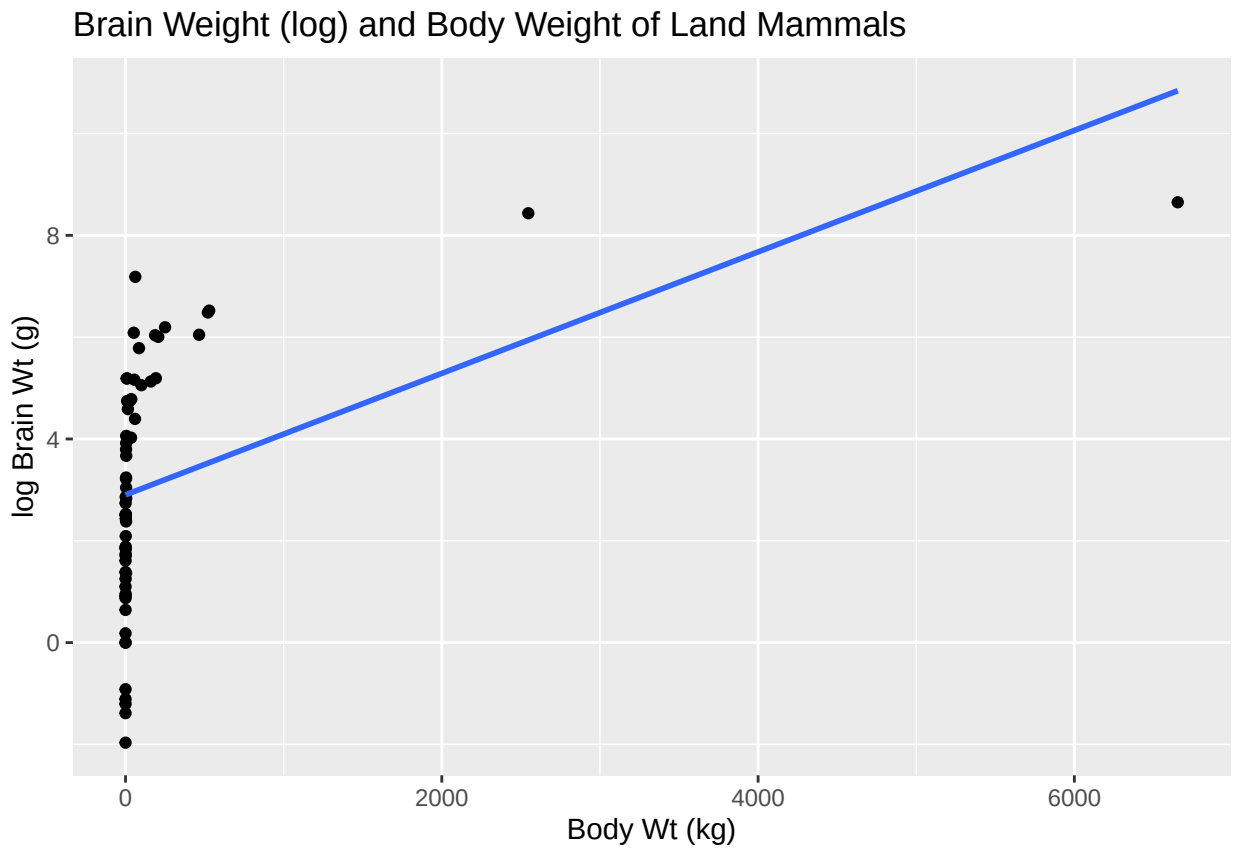
Based on the Box Cox plot, a log transformation on the response variable should be performed, so $y^* = \log(y)$.

5)

```
##log transform y
Data$log.brain<-log(Data$brain)

##scatter plot of y* against x
ggplot(Data, aes(x=body,y=log.brain))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Body Wt (kg)", y="log Brain Wt (g)",
       title="Brain Weight (log) and Body Weight of Land Mammals")
```

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



The relationship between y^* and x does not appear to be linear. The relationship appears logarithmic.

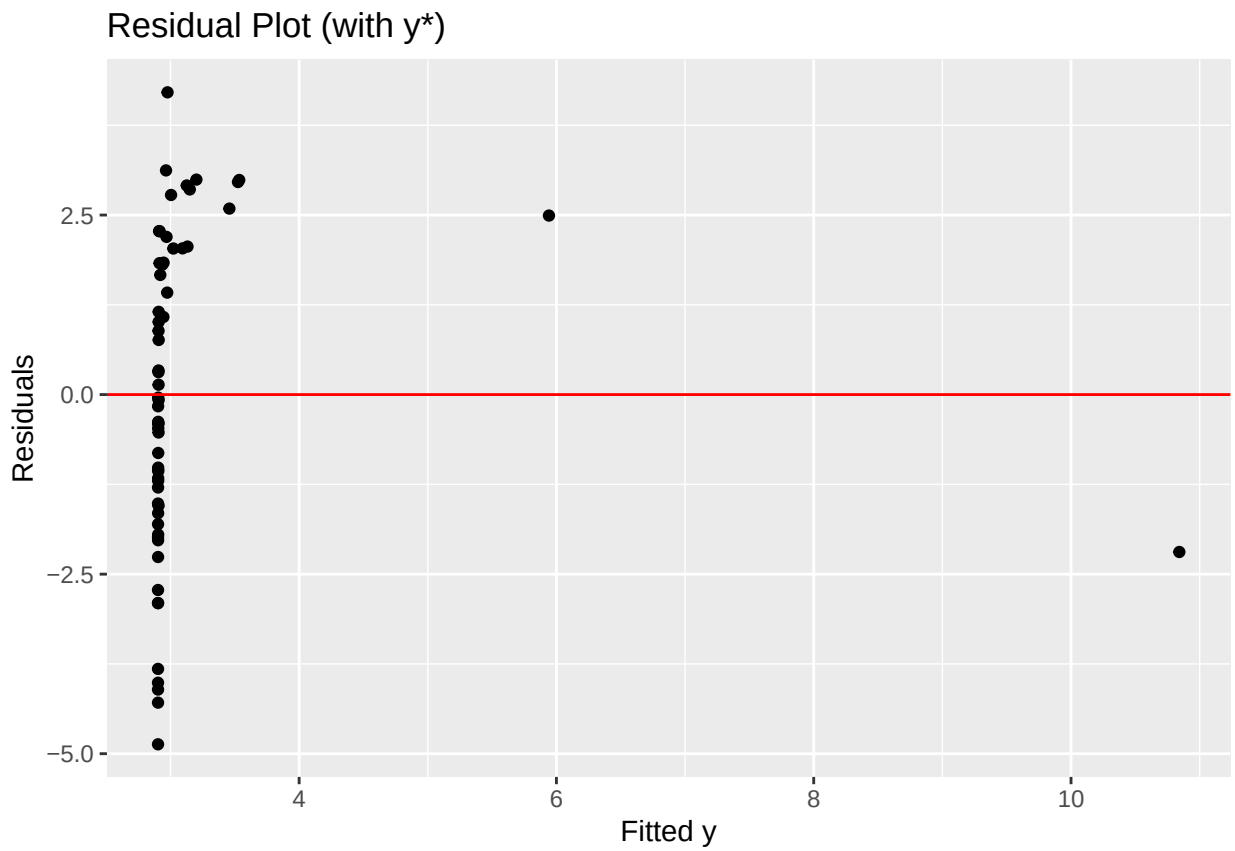
The constant variance assumption is difficult to evaluate with this scatterplot. So we take a look at the residual plot.

6)

```
result2<-lm(Data$log.brain~Data$body, data=Data)

##create residual plot
Data$yhat2<-result2$fitted.values
Data$res2<-result2$residuals

ggplot(Data, aes(x=yhat2,y=res2))+
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted y", y="Residuals", title="Residual Plot (with y*)")
```



The mean of the residuals do not appear to be 0 when fitted y is great than 2 and below 10. This indicates that the relationship between y^* and x is not linear.

Again, the constant variance assumption is difficult to evaluate with this scatterplot. Although the deviation of the residuals from the horizontal axis when $\hat{y}$ is large is more similar to the deviation when $\hat{y}$ is small, so the variance is a lot more constant than before. So assumption 2 is less of an issue now, and we next focus on assumption 1.

7)

Since we already performed the transformation on the response variable based on the Box Cox plot, we consider transforming the predictor to handle assumption 1.

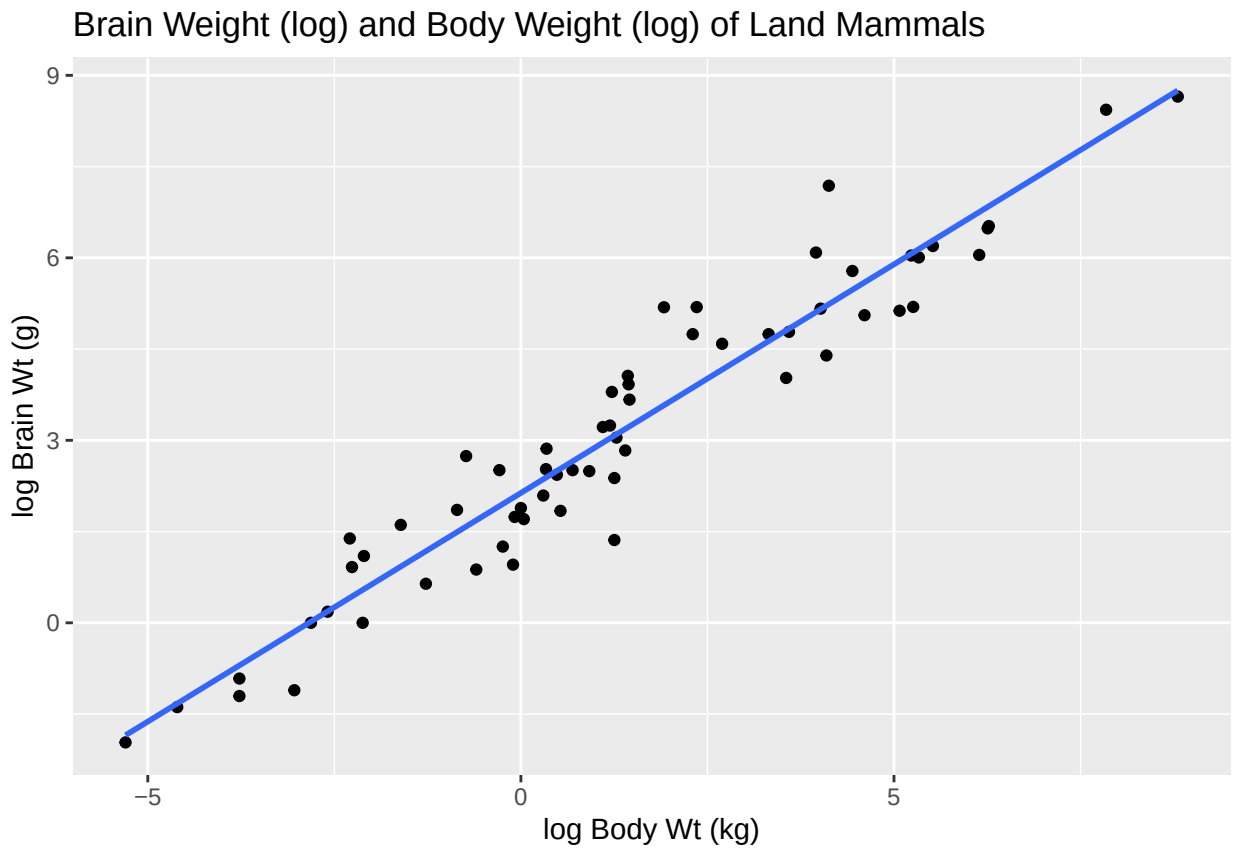
Based on the scattterplot in part 5, we should perform a log transformation on the predictor, so $x^* = \log(x)$.

```
##transform x
Data$log.body<-log(Data$body)

##scatterplot of y* against x*
ggplot(Data, aes(x=log.body,y=log.brain))+
```

```
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="log Body Wt (kg)", y="log Brain Wt (g)",
       title="Brain Weight (log) and Body Weight (log) of Land Mammals")
```

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



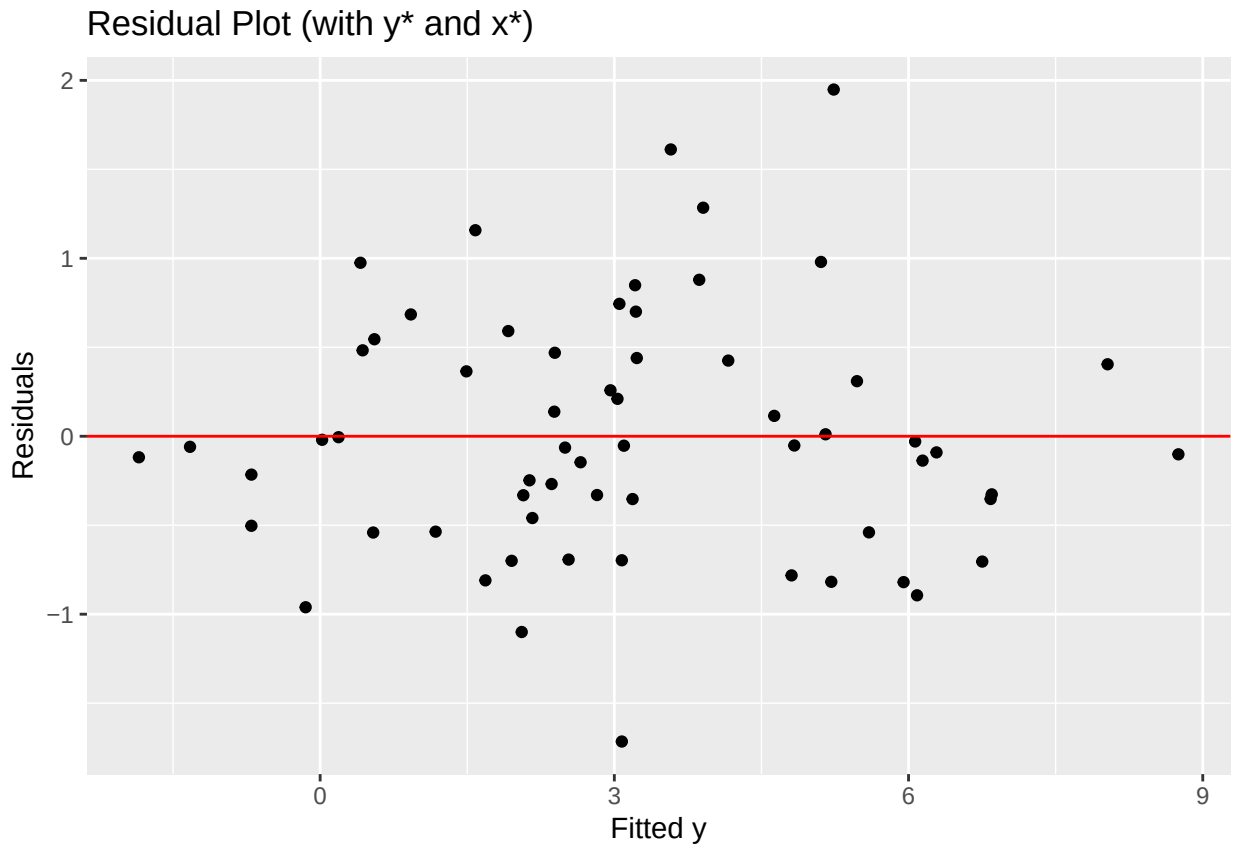
Looking at the scatterplot of y^* against x^*, the relationship appears to be linear (and positive). The constant variance assumption appears to be reasonably met as we don't see the variance increasing or decreasing.

8)

```
result3<-lm(log.brain~log.body, data=Data)

##create residual plot
yhat3<-result3$fitted.values
res3<-result3$residuals
```

```
ggplot(Data, aes(x=yhat3,y=res3))+
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted y", y="Residuals", title="Residual Plot (with y* and x*)")
```

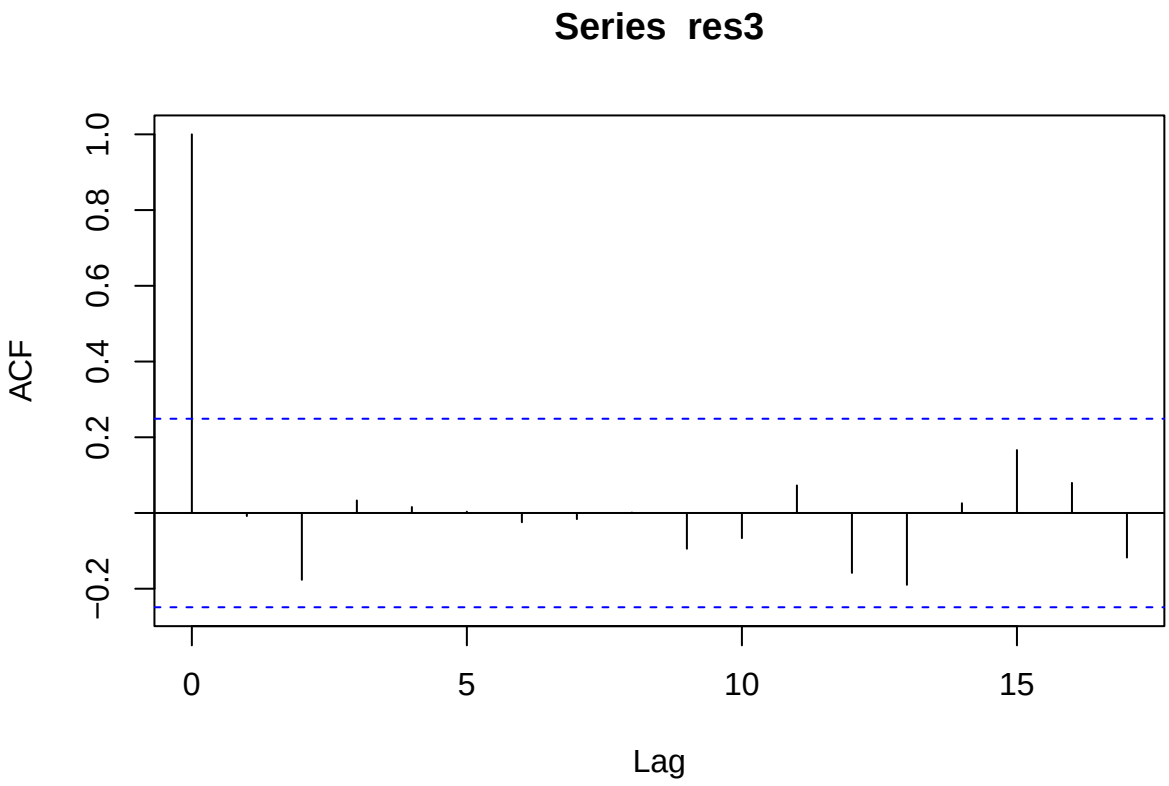


We see huge improvements in the residual plot of y^* against x^*. The residuals are generally evenly scattered on both sides of the x-axis, so the assumption that the errors have 0 mean is met.

The spread of the residuals also is fairly constant as we move across the residual plot. We do not see the variance increasing or decreasing. So the constant variance assumption for the errors is met.

9)

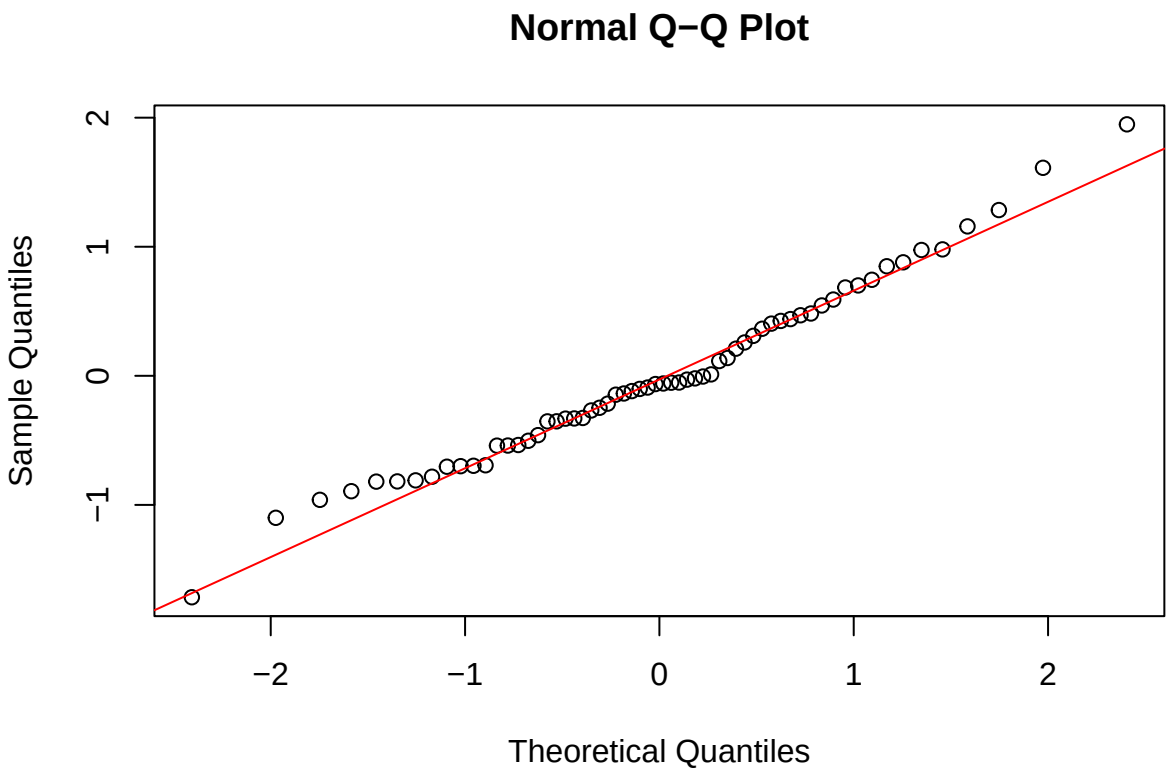
```
acf(res3)
```



Based on the ACF plot, the residuals are uncorrelated, so we don't have evidence that the errors are dependent.

10)

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



The plots fall closely to the line, so the residuals follow a normal distribution.

11)

```
summary(result3)
```

```
##
## Call:
## lm(formula = log.brain ~ log.body, data = Data)
##
## 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 ***
## log.body     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
```

We have $y^* = 2.13 + 0.75x^*$, where $y^* = \log(y), x^* = \log(x)$.

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.