# M05Guided

Alanna Hazlett

2024-02-26

```
Data<-mammals
```

Problem 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?

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

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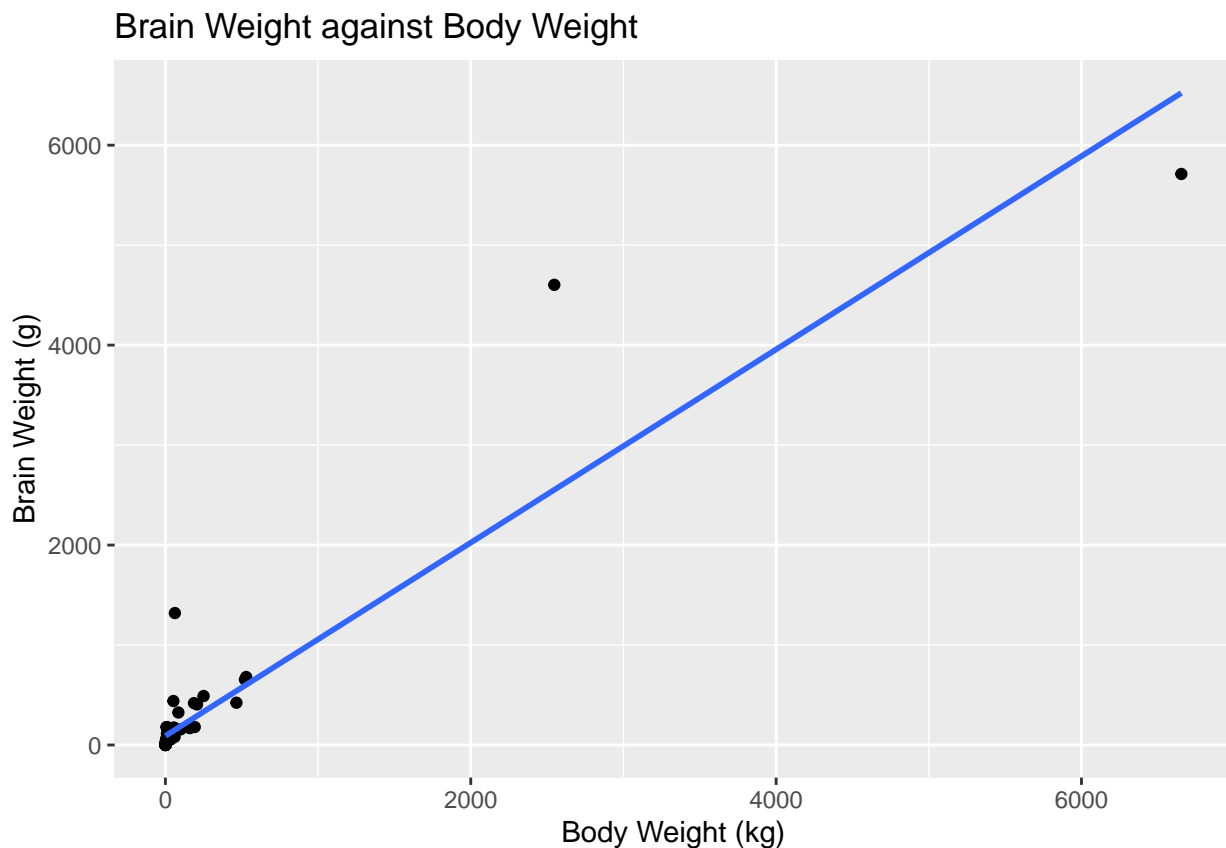

The majority of the data points fall within the lower left part of the scatterplot with 3 outliers present. Assumption 1 is not met. The data points are not evenly spread on either side of the regression line. Assumption 2 is not met. The variance on the left side of the scatterplot is small and then gets larger at

about 60, then gets smaller again at 200.
We can not comment on assumption 3 or 4 from the scatterplot.

Problem 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?

```
result<-lm(brain~body,data=Data)
par(mfrow=c(2,2))
plot(result)
```
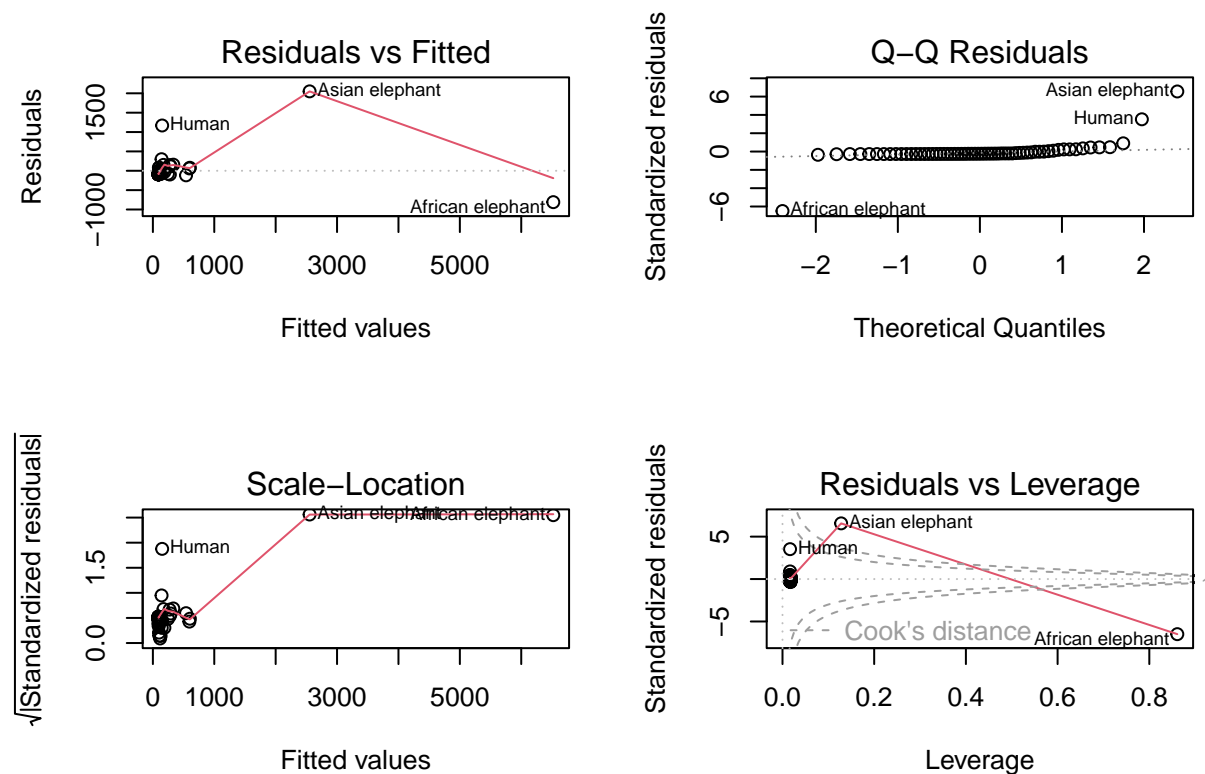

Assumption 1 is not met, we can see that the errors do not have a mean of 0 by the fact that the Residuals vs Fitted does not have a straight horizontal line.
Assumption 2 is not met, the variance of the data points is not uniform.
Assumption 4 - the errors are normally distributed, excluding the outliers of African elephant, Asian elephant, and Human.
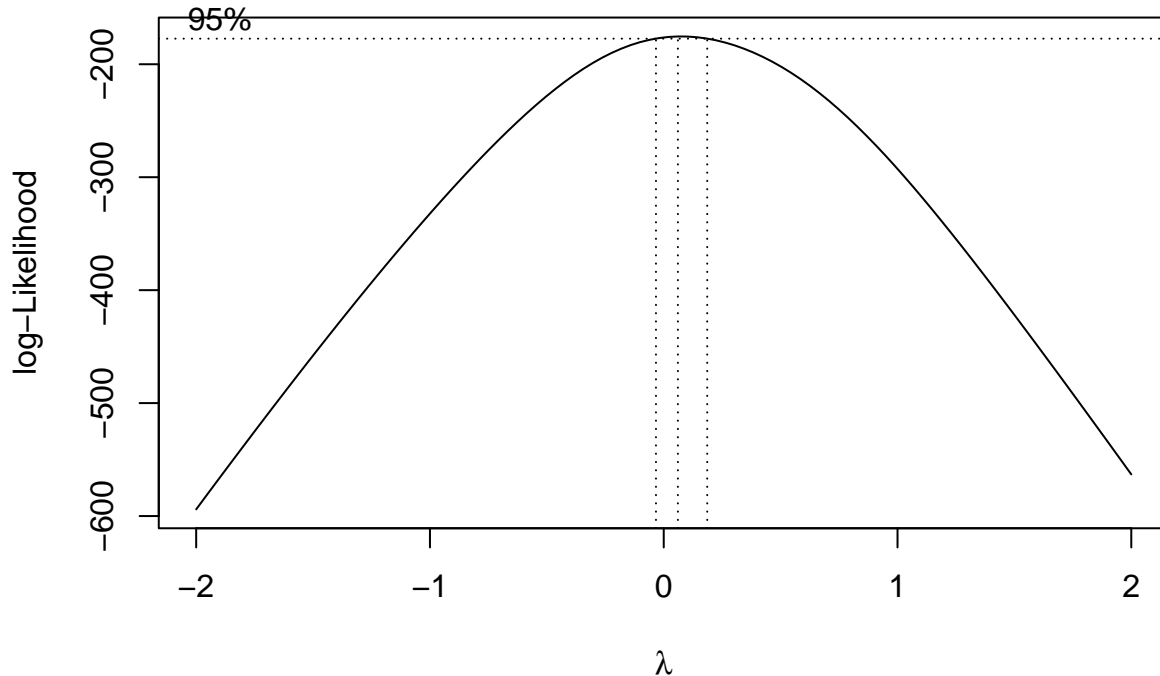
Problem 3

Based on your answers to parts 1 and 2, do we need to transform at least one of the variables? Briefly explain.
We should start by transforming y, as transforming it will affect the outcome of Assumptions 1 & 2.

Problem 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.

```
boxcox(result)
```


We will do a logarithmic transformation on y, because zero is within our confidence interval for lambda and this will allow us to be able to interpret the regression coefficients still.

Problem 5

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

```
ystar<-log(Data$brain)
Data<-data.frame(Data,ystar)
ggplot(Data, aes(x=body,y=ystar))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)+
  labs(x="Body Weight (kg)", y="Log Transformation of Brain Weight", title="Ystar against Body Weight")
```
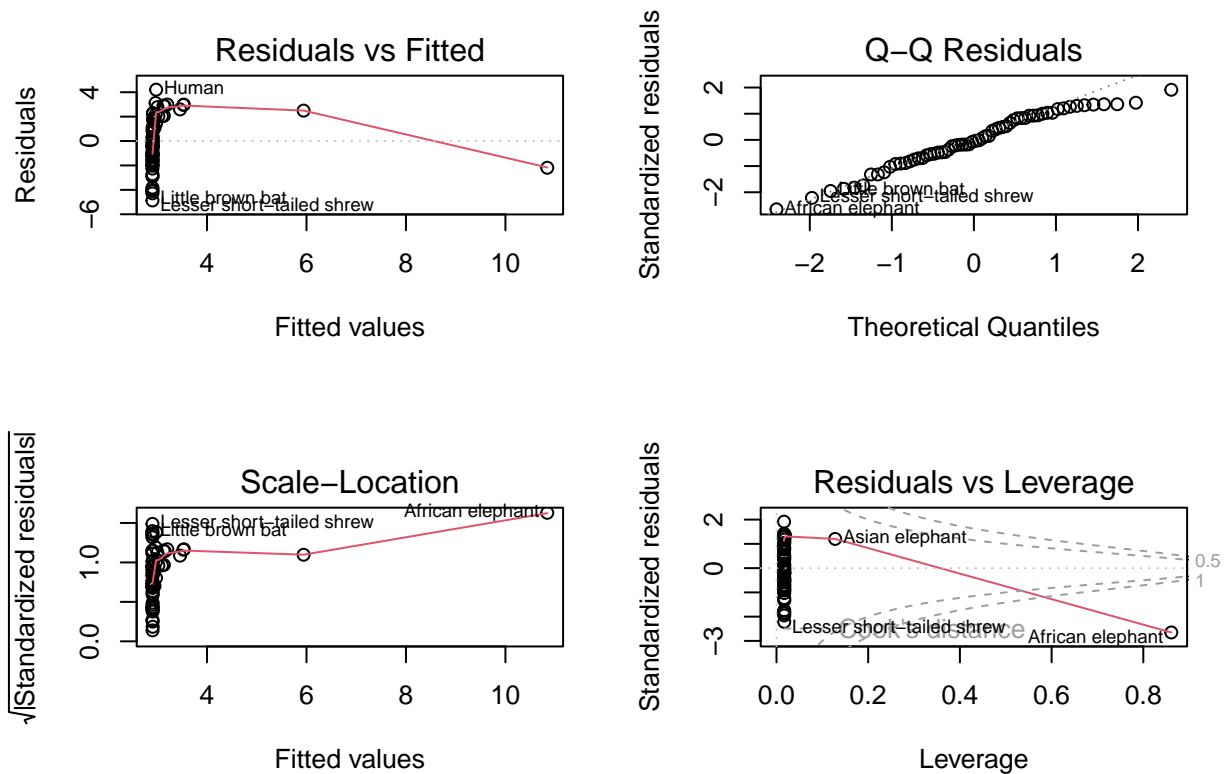
```
## `geom_smooth()` using formula = 'y ~ x'
```

Ystar against Body Weight

Assumption 1 appears to be met. The data points appear even distributed on either side of the regression line.
* Assumption 2 is not met, the variance of the data points is not uniform. On the left side of the scatterplot the variance is large and becomes less on the right.

Problem 6

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

```
result.ystar<-lm(ystar~body,data=Data)
par(mfrow=c(2,2))
plot(result.ystar)
```

Residuals vs Fitted

Residuals

Human
Little brown bat
Lesser short-tailed shrew

4   6   8   10

Fitted values

Q–Q Residuals

Standardized residuals

Little brown bat
Lesser short-tailed shrew
African elephant

−2   −1   0   1   2

Theoretical Quantiles

Scale–Location

√|Standardized residuals|

Lesser short-tailed shrew
Little brown bat
African elephant

4   6   8   10

Fitted values

Residuals vs Leverage

Standardized residuals

Asian elephant
Lesser short-tailed shrew
Cook's distance
African elephant

0.0   0.2   0.4   0.6   0.8

Leverage

Assumption 1: Is difficult to determine as the majority of the values all fall at or near the same fitted value. Assumption 2: This is also difficult to determine with the majority of the values at or near the same fitted value. Assumption 4: The Q-Q Residuals Plot actually decreased in accuracy of our residuals being on the dashed line.
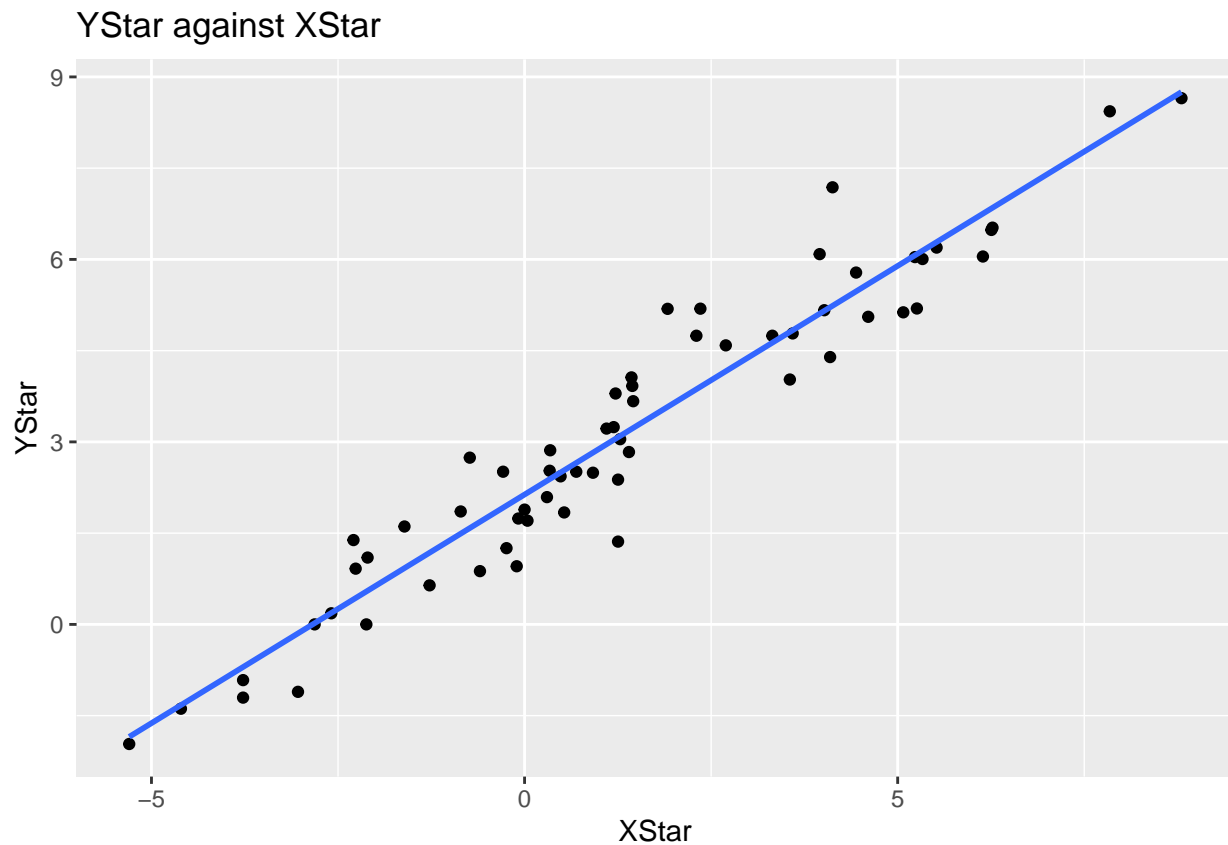
Problem 7

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

* I think that we do need to transform the x variable. I choose to transform in as log(x) due to the shape of the scatterplot of ystar vs x.

```
xstar<-log(Data$body)
ggplot(Data, aes(x=xstar,y=ystar))+
  geom_point()+
  geom_smooth(method="lm", se=FALSE)+
  labs(x="XStar",y="YStar",title="YStar against XStar")
```

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

YStar against XStar

* Assumption 1: Appears to be met, data point spread across the regression line evenly.
* Assumption 2: Appears to be met, variance is even from left to right on the scatterplot.

Problem 8

Fit a simple linear regression to ystar against xstar, 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.

```
result.xstar<-lm(ystar~xstar)
par(mfrow=c(2,2))
plot(result.xstar)
```
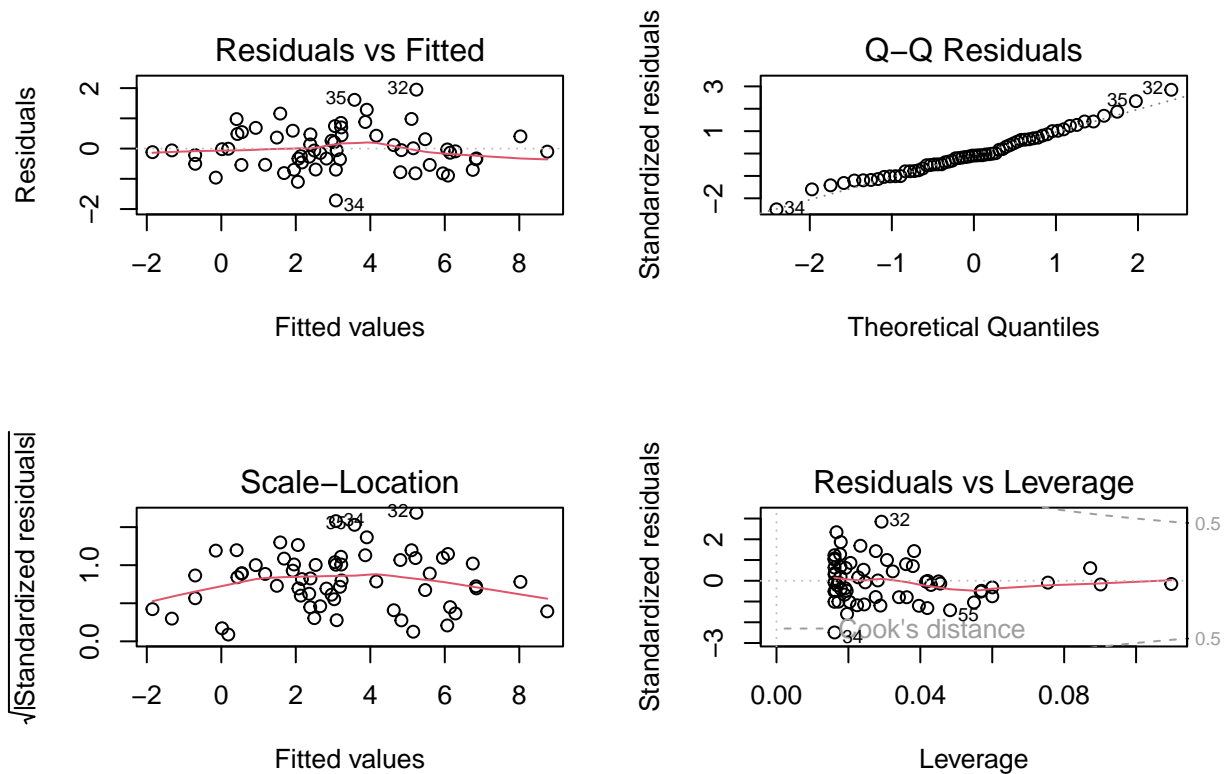
* Assumption 1: Error of means = 0 is met, the red line in Residuals vs Fitted is approximately horizontal at 0. * Assumption 2: I would say the variance is relatively even across the plot.

Problem 9

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

```
summary(result.xstar)
```

```
##
## Call:
## lm(formula = ystar ~ xstar)
##
## 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 ***
## xstar        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
```

The regression equation is:

$$\hat{y} = 2.13 + 0.75x$$

Since we only performed logarithmic transformations our statistical inference has not been affected and we can interpret the slope in context.
For every kg increase in body weight the brain weight increases by 0.75g.