M12 Problem Set: Simple Linear Regression for TAs

2024-04-17

Female specimens of the amphipod crustacean **Platorchestia platensis** were collected from a beach near Stony Brook, Long Island (New, York, USA), in April 1987 (McDonald, 1989). The eggs were removed, counted, and then discarded. The females were freeze-dried. Later in the laboratory the females were weighed.

Is the number of eggs produced by a mother related to the mother's mass?

1. Use simple linear regression to find the best fit. What is the equation for that line? (round your coefficient to 2 decimals).
   Answer: Y=12.69 +1.60X

2. What is the r2 of your model, is it meaningful?
   Answer: 0.2055 0 0.21. No, it is not meaningful.

3A. What is the RMSE of your model?
3B. What is the PRESS_RMSE of your model?
3C. Is your model stable? Remember to compare the PRESS_RMSE with the RMSE.
Answer: RMSE= 3.600417
Answer: PRESS_RMSE= 3.887804
Answer: Yes the model is stable.

4. Using the Cook's D statistical test, is there evidence of potential outliers or high leverage observations in your model? If so, how many? You don't need to remove them because we don't know if they are genuine data points, but you do need to identify them.
   Answer: 3

5. Write a concise one paragraph summary of this analysis.
   Remember that any summary should include the following:
   5a. Hypothesis or research objectives clearly stated.
   5b. Correct statistical test selected and clearly stated . 5b.a. Simple Linear Least Squares Regression(no assumption for normality of variables, only of residuals (actual vs. predicted).
   5c. Include the model equation in Y = b + mX format, to describe the model include:
   5c.a r2: how meaningful is the mode.
   5c.b RMSE: how accurate is the model.
   5c.c PRESS RMSE: how stable is the model.
   5c.d p-value: how significant is the relationship the model is based on.
   5d. Interpretation.How useful do you think this model will be for its intended purpose (link back to the original objective). Consider all of the metrics you described above as well as the range over which the model is valid.

Answer: The objective of the analysis was to investigate the relationship between the number of eggs produced by a mother and the mother's mass. To address this, a simple linear least squares regression model was employed. The resulting model equation, eggs=12.6890+1.6017×weight, indicates that for each unit increase in the mother's mass, there is an estimated increase of 1.6017 units in the number of eggs produced.

The coefficient of determination (R2) value of 0.2055 suggests that approximately 20.55% of the variability in egg production can be explained by the mother's mass. The residual standard error (RMSE) of 3.600417 represents the average deviation of observed values from the fitted values. The p-value for the coefficient of the weight variable is 0.0154, indicating a significant relationship between egg production and the mother's mass at a 5% significance level.

Assessment of potential outliers using Cook's distance suggests that their presence may impact the model's coefficients, accuracy, and significance. Thus, caution is warranted when interpreting results, especially if outlier removal is considered. Despite this, the model demonstrates usefulness in predicting egg production within the observed range of mother masses. However, its predictive power may be limited beyond this range or in the presence of outliers. Therefore, while informative, it's crucial to consider additional factors and exercise caution when applying the model to new data or extrapolating beyond the observed range.

#1. Import libraries and load packages

```
library(tidyverse)
library(dplyr)
library(readxl)
```

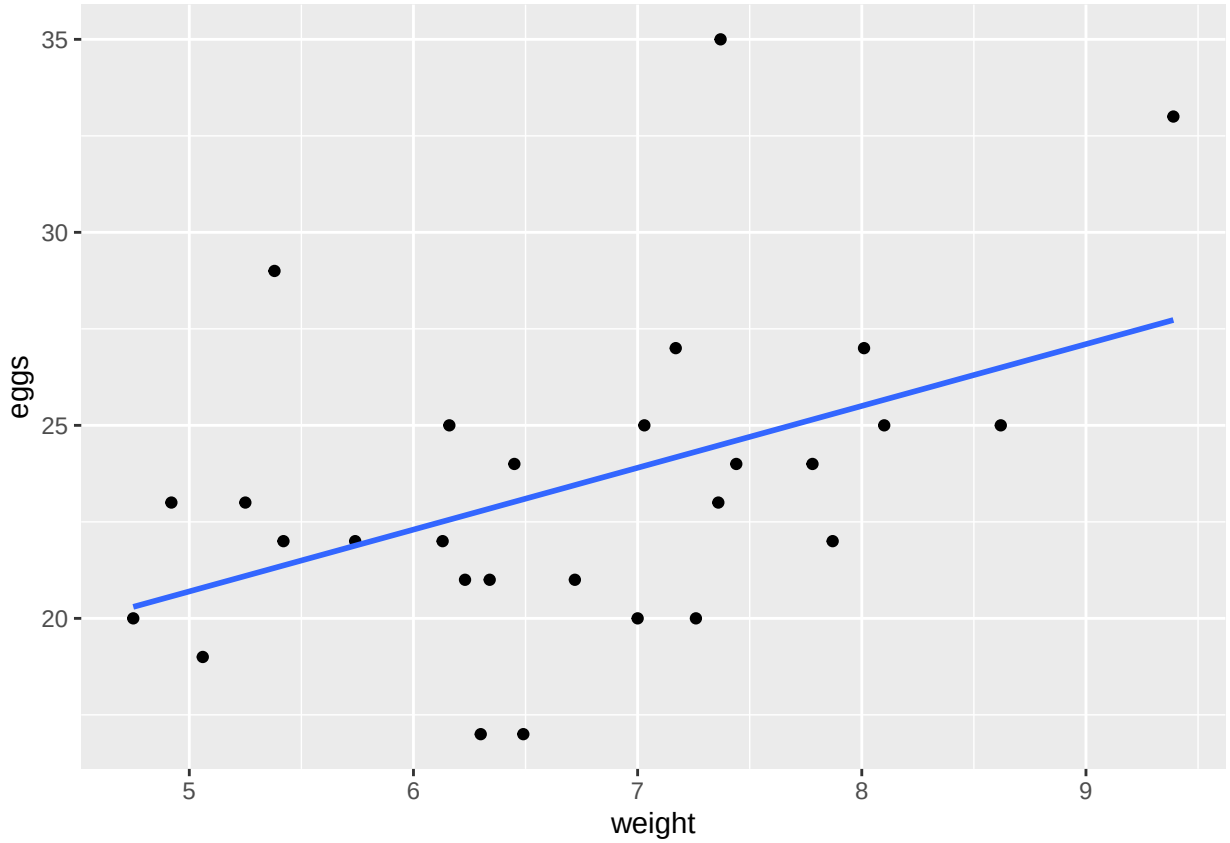
#2. importing our data

```
eggs_weight <- read_excel("eggs_weight.xlsx")
```

#3. Explore your data

```
ggplot(eggs_weight, aes(x=weight, y=eggs)) +
  geom_point() + geom_smooth(method='lm', se=F)
```

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



#4. Run the model

```
mod = lm(eggs ~ weight, data = eggs_weight)
summary(mod)
```

```
##
## Call:
## lm(formula = eggs ~ weight, data = eggs_weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0842 -1.8063 -0.5567  1.5864 10.5063
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.6890     4.2009   3.021   0.0056 **
## weight        1.6017     0.6176   2.593   0.0154 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.736 on 26 degrees of freedom
## Multiple R-squared:  0.2055, Adjusted R-squared:  0.175
## F-statistic: 6.726 on 1 and 26 DF,  p-value: 0.0154
```

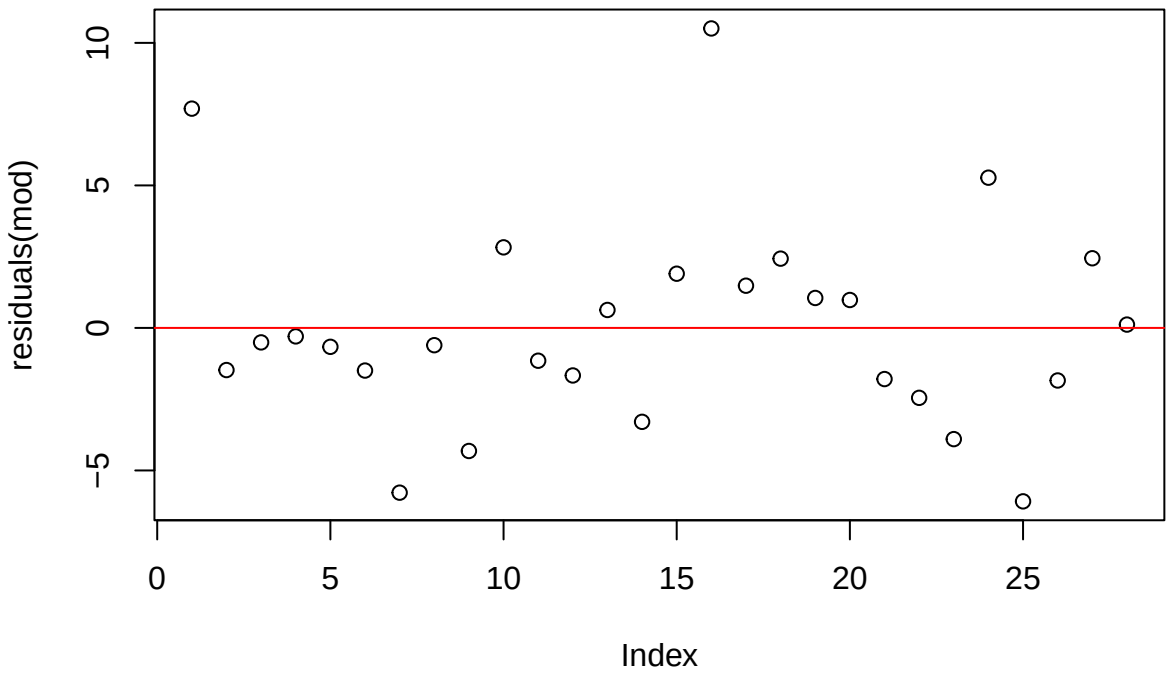
#5. Check Normality of the residuals.

```r
shapiro.test(mod$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod$residuals
## W = 0.93817, p-value = 0.0993
```

#6a. Plot the Residuals

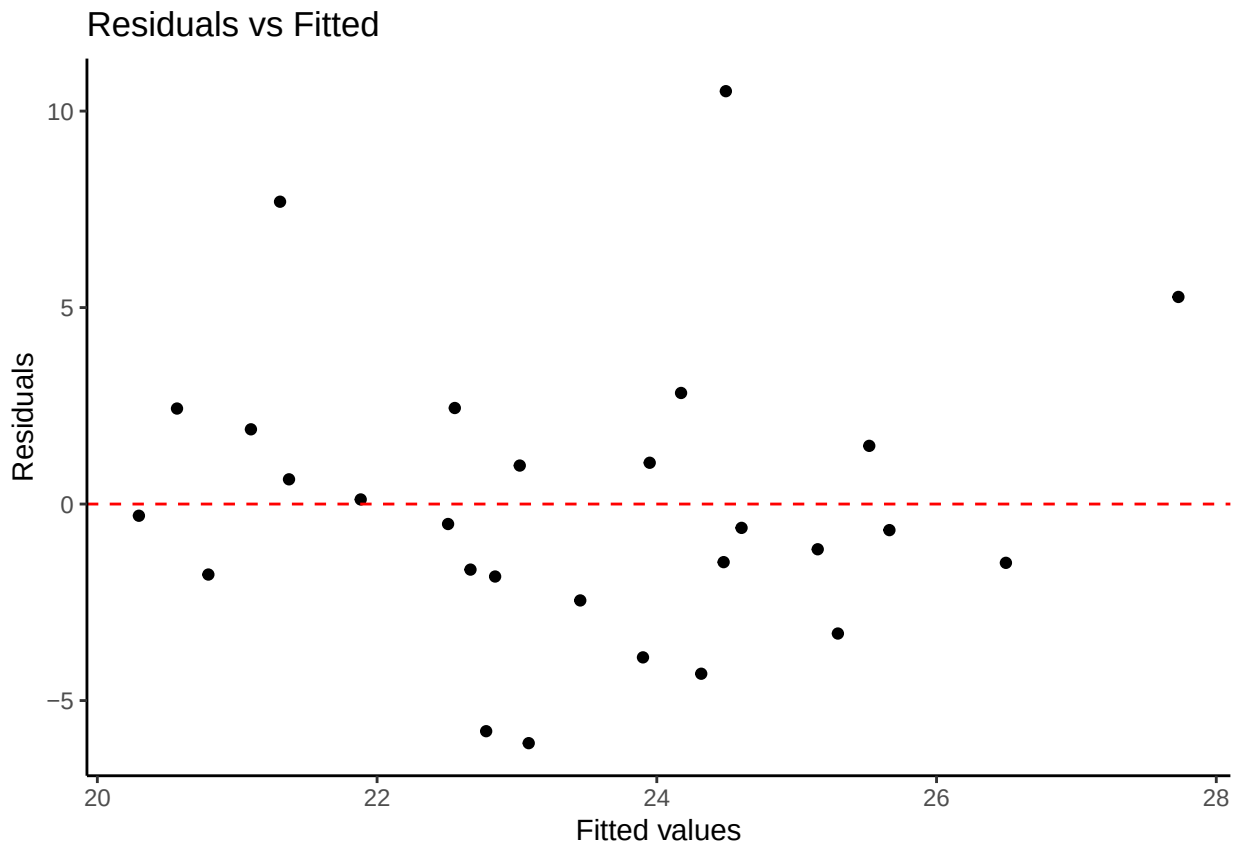
```r
plot(residuals(mod))
abline(h = 0, col = "red")
```



#6b. Plot the Residuals in ggplot.

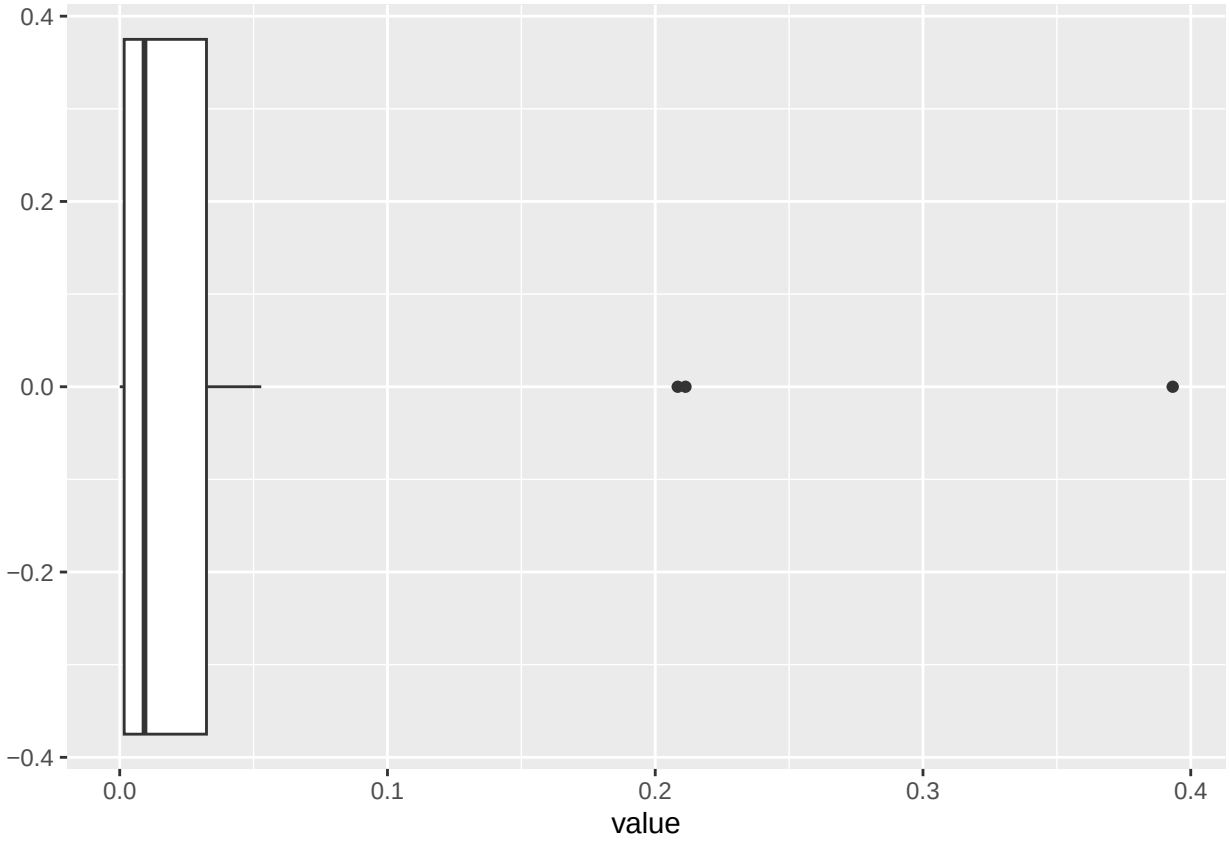
```r
# Create a data frame with residuals and fitted values
residuals_df <- data.frame(
  Residuals = residuals(mod),
  Fitted = fitted(mod)
)

# Plot the residuals
ggplot(residuals_df, aes(x = Fitted, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
```

```
  labs(x = "Fitted values", y = "Residuals") +
  ggtitle("Residuals vs Fitted") +
  theme_classic()
```

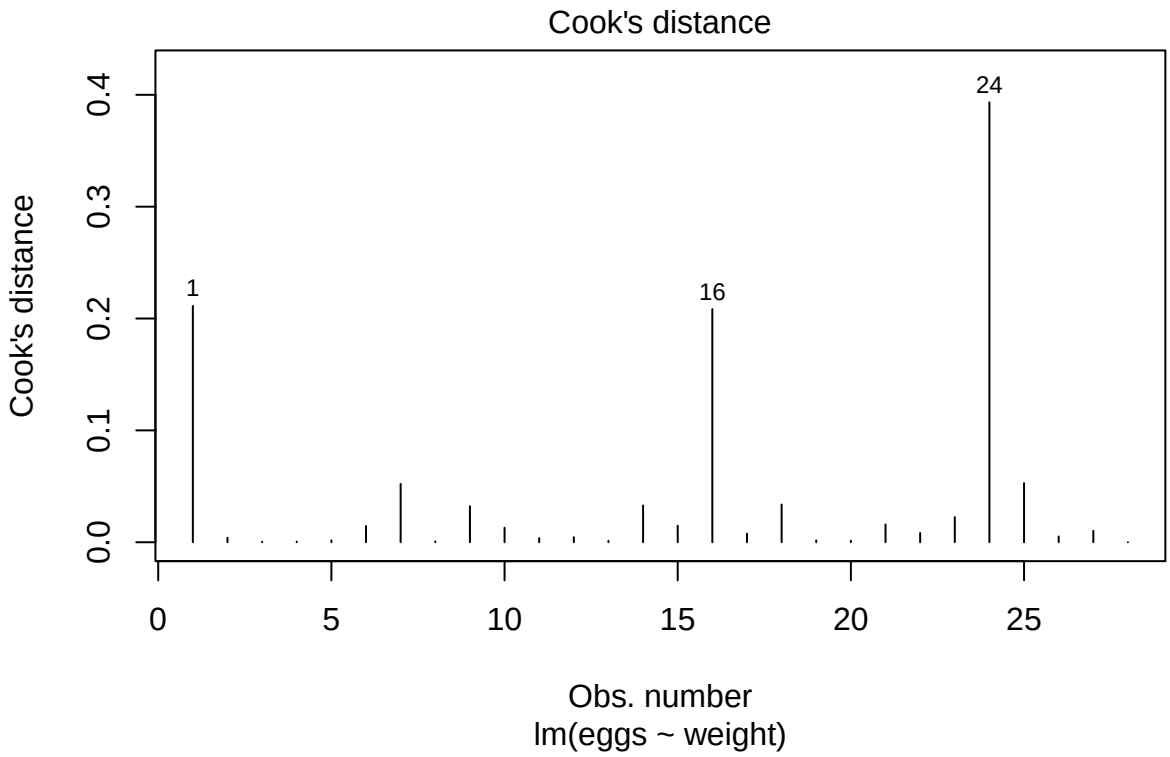


#7a. Cooks D

```
#Cooks D
ggplot(as_tibble(cooks.distance(mod)), aes(value)) + geom_boxplot()
```

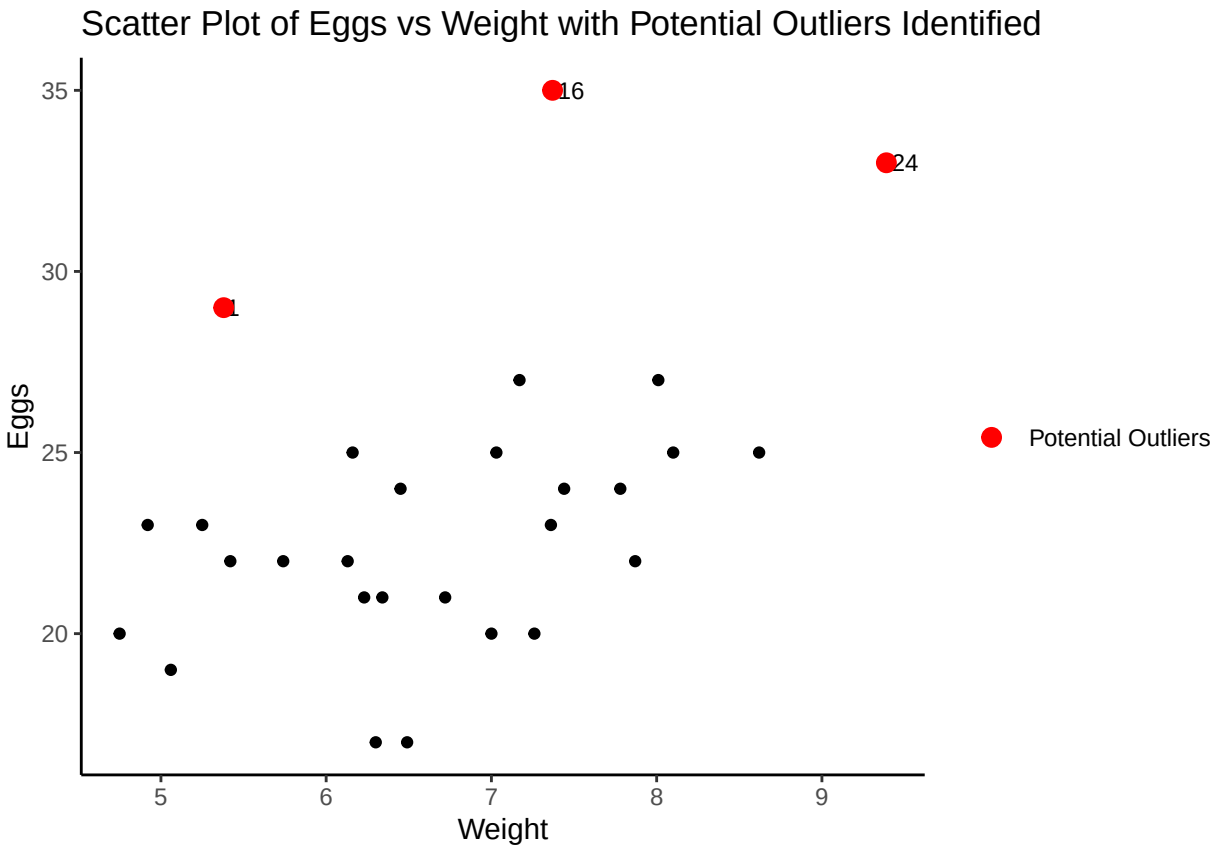



```
plot(mod,which=4)
```



#7b. Hihglight the potential ouliers with ggplot

```
# Calculate Cook's distance
cooks_distance <- cooks.distance(mod)

# Create a data frame with observation numbers and Cook's distance
cooks_df <- data.frame(
  Observation = 1:length(cooks_distance),
  Cooks_Distance = cooks_distance
)

# Set a threshold for identifying potential outliers
threshold <- 4 / length(cooks_distance)

# Identify potential outliers
outlier_indices <- which(cooks_distance > threshold)

# Create a ggplot of the data with potential outliers identified and labeled
ggplot(eggs_weight, aes(x = weight, y = eggs)) +
  geom_point() +
  geom_text(data = eggs_weight[outlier_indices, ], aes(label = outlier_indices), hjust = -0.2, vjust = 
  geom_point(data = eggs_weight[outlier_indices, ], aes(color = "Potential Outliers"), size = 3) +
  labs(x = "Weight", y = "Eggs", color = "") +
  scale_color_manual(values = c("red", "red")) +
  ggtitle("Scatter Plot of Eggs vs Weight with Potential Outliers Identified") +
  theme_classic()
```




```
#Residual sum of squares
RSS <- c(crossprod(mod$residuals))
# Mean squared error:
MSE <- RSS / length(mod$residuals)
RMSE <- sqrt(MSE)
RMSE
```

```
## [1] 3.600417
```

```
# PRESS ------------------------------------------------------------------
# (r <- resid(mod))
pr <- resid(mod)/(1 - lm.influence(mod)$hat)
PRESS <- sum(pr^2)
PRESS
```

```
## [1] 423.2206
```

```
# PRESS_RMSE
PRESS_RMSE <- sqrt(PRESS/length(mod$residuals))
PRESS_RMSE
```

```
## [1] 3.887804
```

References https://www.biostathandbook.com/linearregression.html

McDonald, J.J. (1987). Repeated geographic variation at three enzyme loci in the amphipod Platorchestia platensis. Evolution 41(2):438-441.