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?

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.

#1. Import libraries and load packages

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

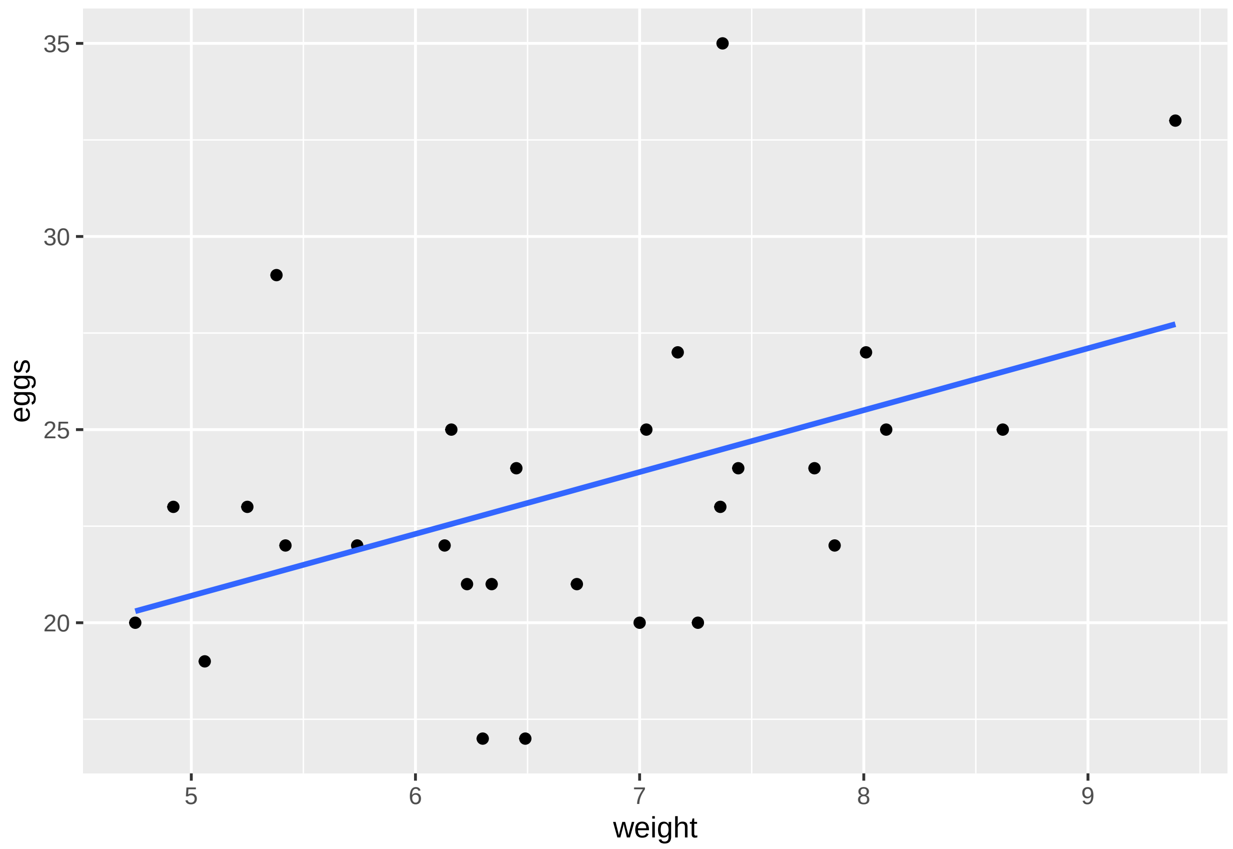
#2. importing our data

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

#3. Explore your data

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

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



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

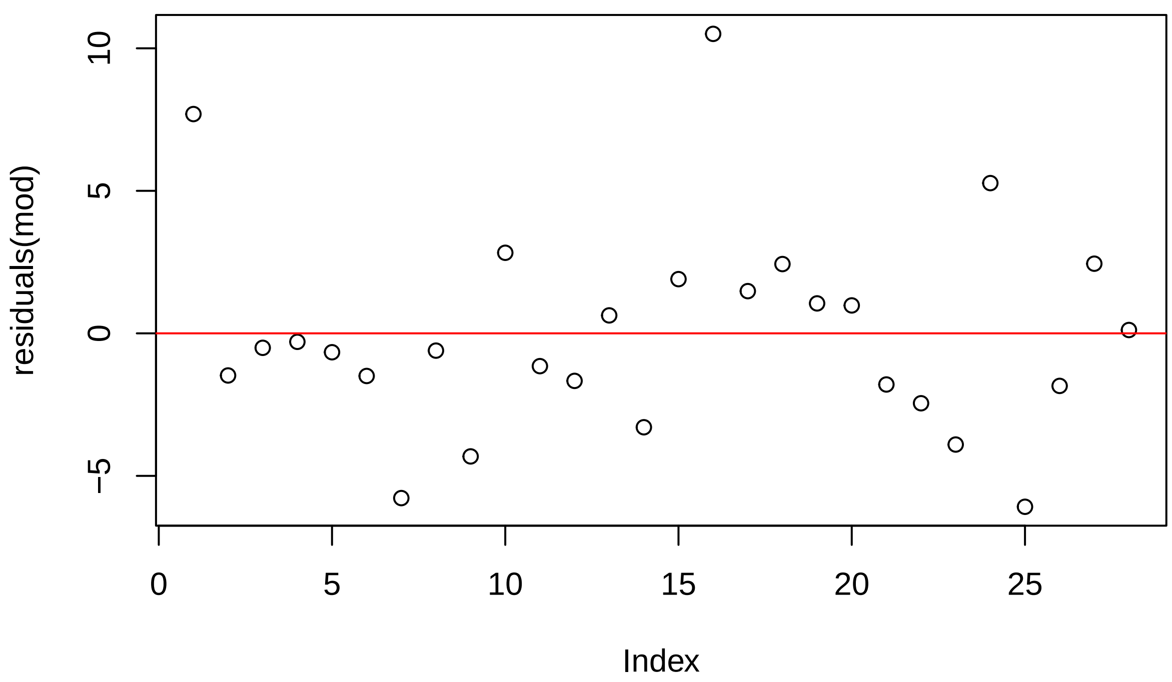
```
##
## Call:
## lm(formula = eggs ~ weight, data = eggs_weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0842 -1.8063 -0.5567  1.5864 10.5063
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.6890     4.2009   3.021   0.0056 **
## weight        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
```

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

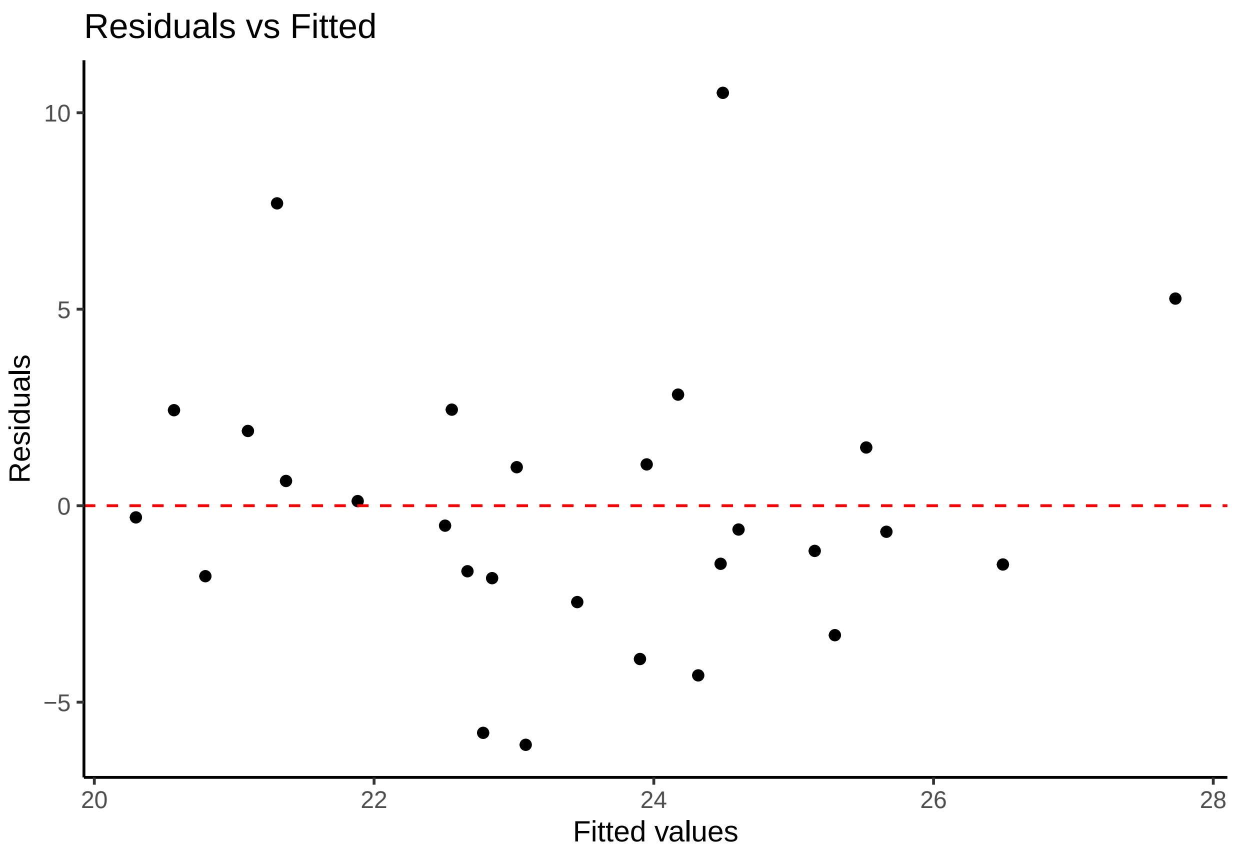
```
##
```

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

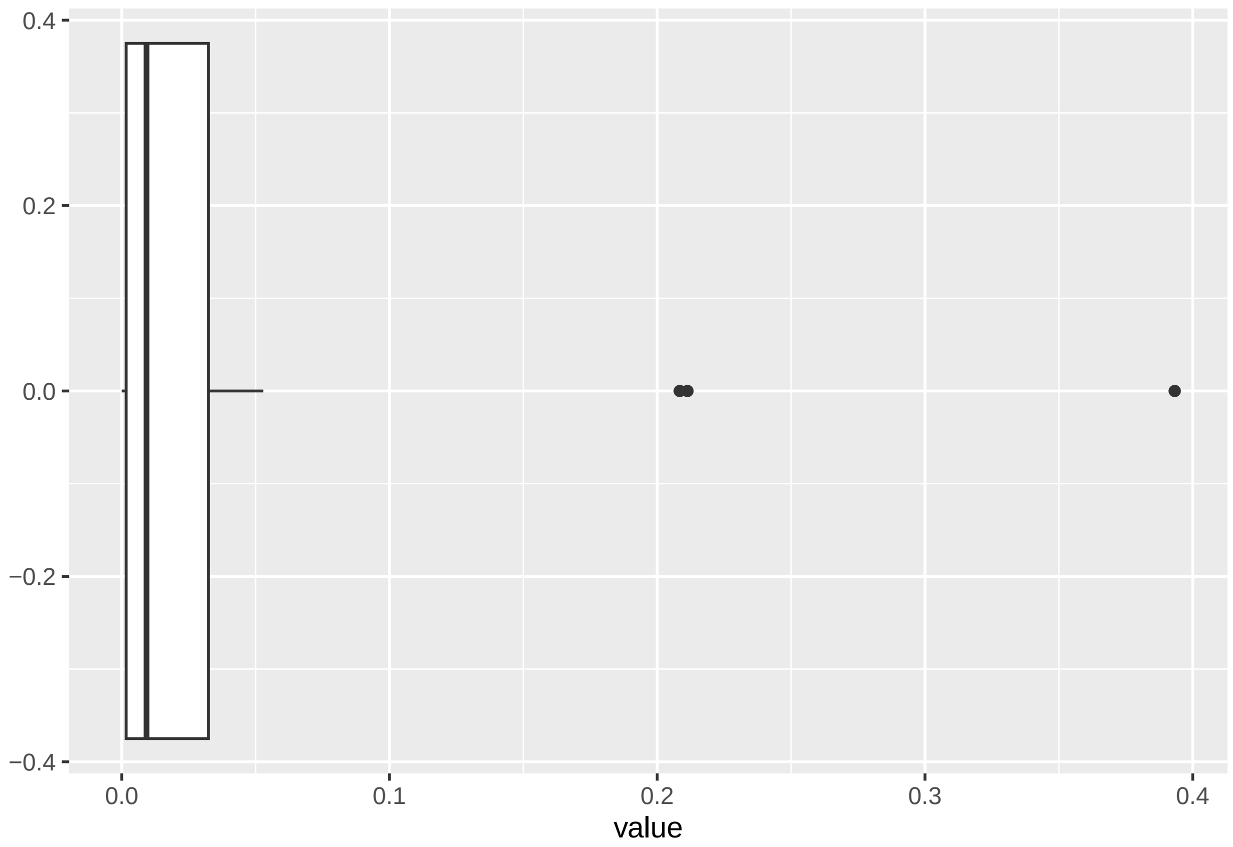
```
plot(residuals(mod))
abline(h = 0, col = "red")
```



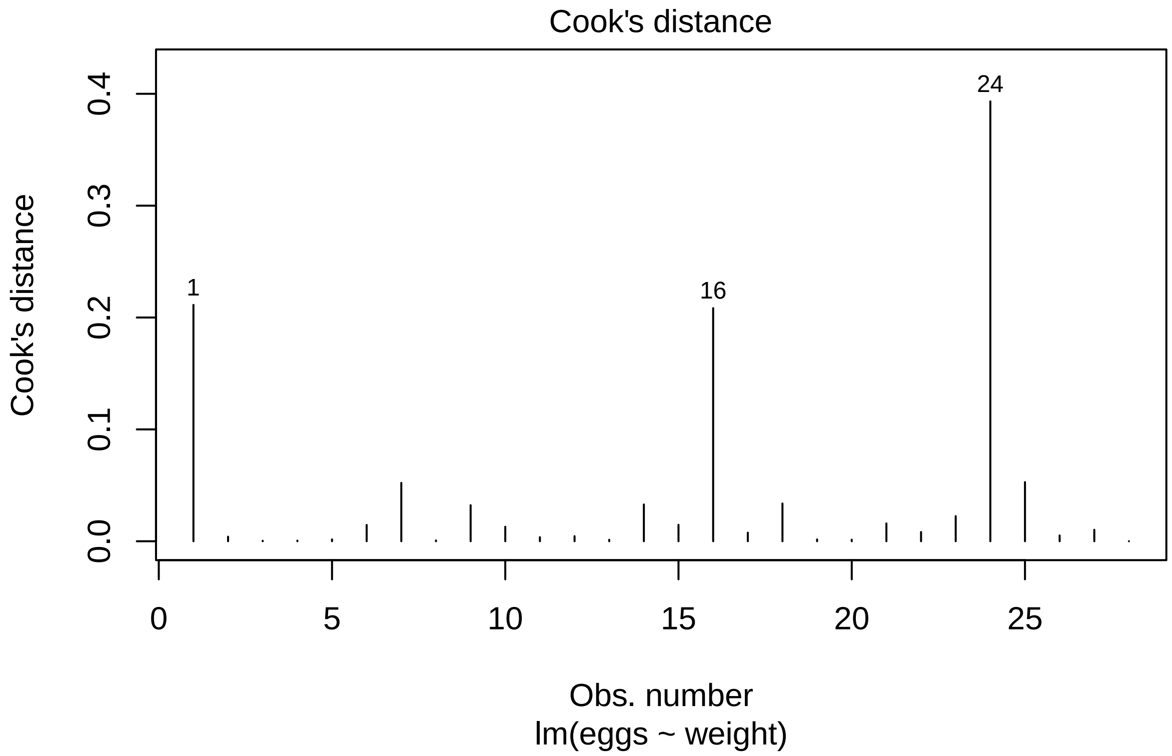
```
# Create a data frame with residuals and fitted values
residuals_df <- data.frame(
  Residuals = residuals(mod),
  Fitted = fitted(mod)
)

# 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()
```



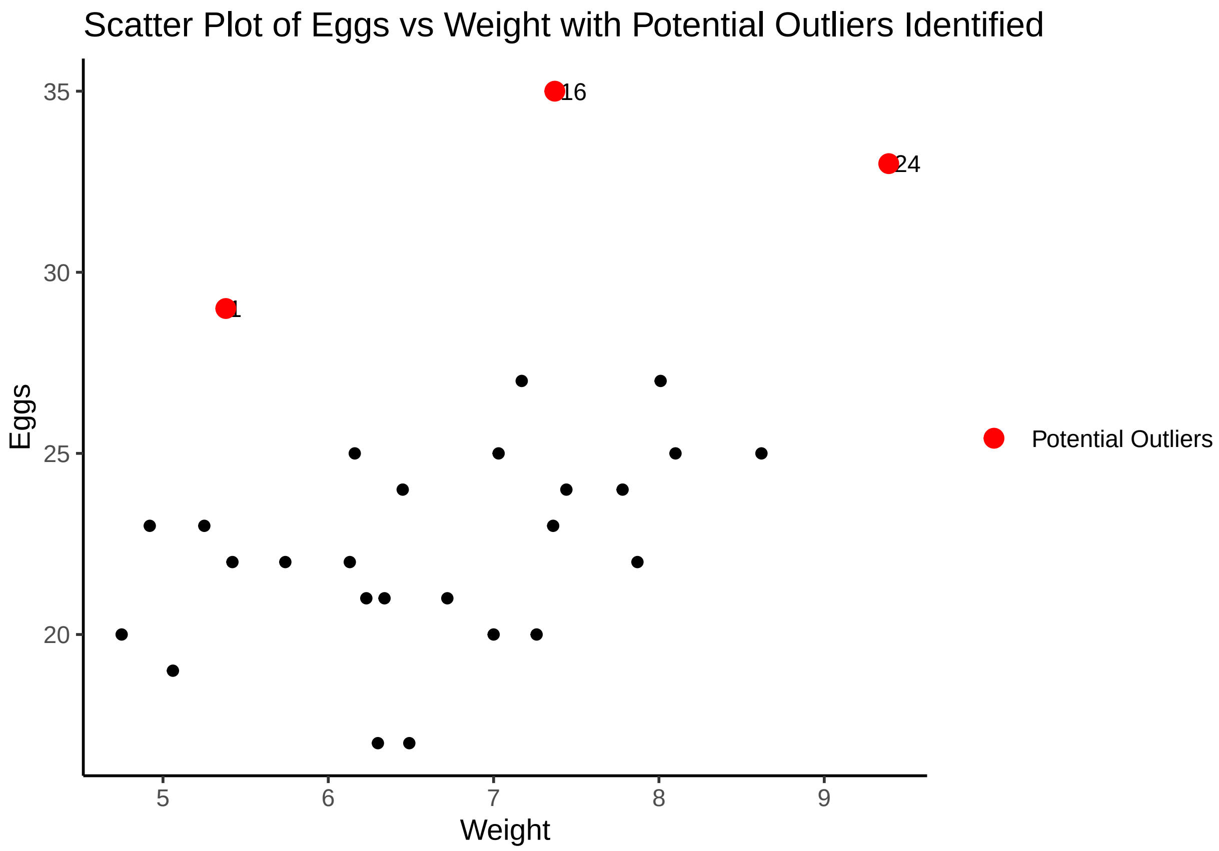
```r
#Cooks D
ggplot(as_tibble(cooks.distance(mod)), aes(value)) + geom_boxplot()
```



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



```r
library(ggplot2)

# Fit the linear regression model
mod <- lm(eggs ~ weight, data = eggs_weight)

# 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 = 0
  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()
```



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.