In-Class Worksheet - Solutions

STAT011 with Prof Suzy

Week 4: Fitting a Linear Regression Model

Instructions: Wildlife researchers are monitoring a Florida alligator population by taking areal photographs and attempting to estimate the weights of the gators based on the length of the gators in the images. The data set `Gators.csv` contains the variables `Length` and `Weight` for a sample of alligators who have been captured and studied. This data is shown in the scatterplot below. Import the `Gators` data set into either Excel or RStudio and then answer the following questions.

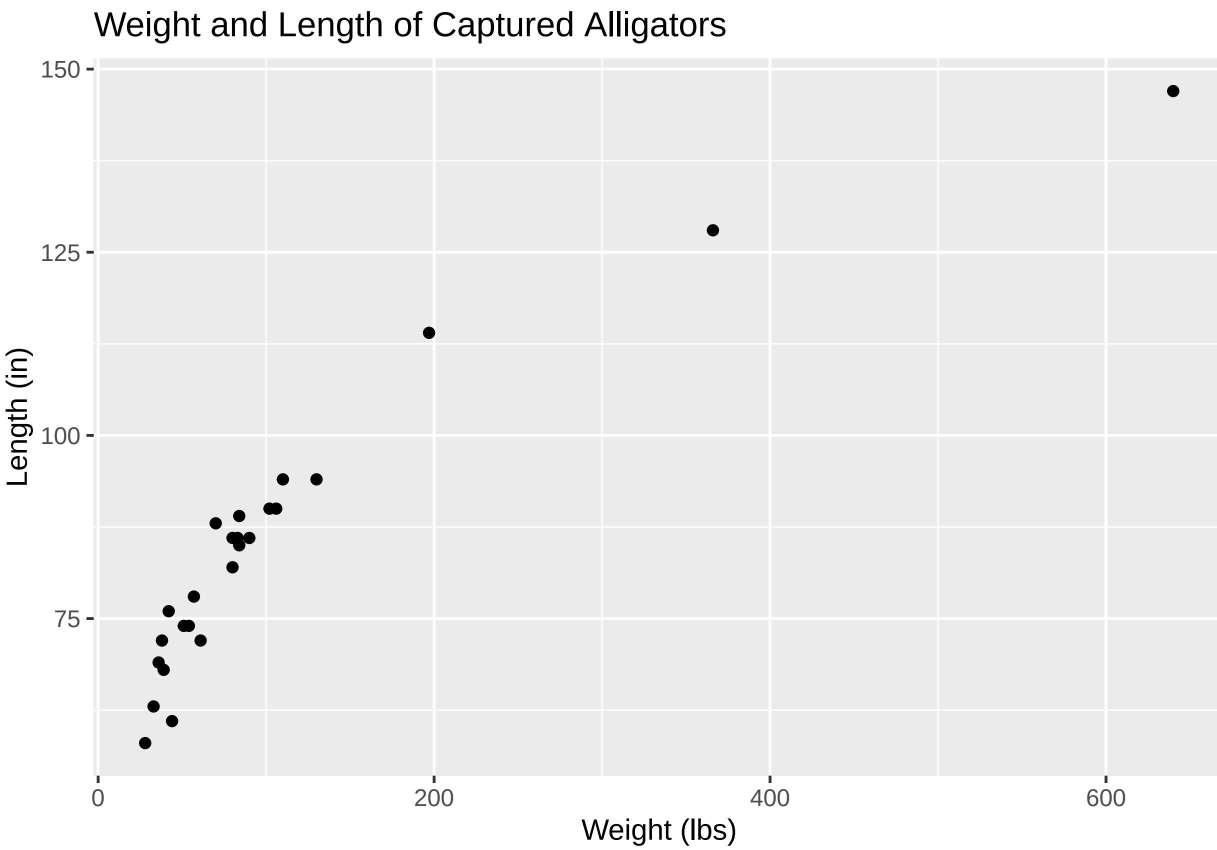
Import data in Excel

Copy and past the data from the following link: https://raw.githubusercontent.com/dr-suz/Stat11/main/Data/Gators.csv

Import data in R

```
gators <- read.csv("https://raw.githubusercontent.com/dr-suz/Stat11/main/Data/Gators.csv")
```

Scatter plot of the data



1. Choose which variable should be the response and justify this choice in 1-2 sentences.

2. What are the slope and intercept of the line of best fit through this data? What is the interpretation of the slope within this context?

```r
slr <- lm(gators$Weight.lbs. ~ gators$Length.in.)
summary(slr)
```

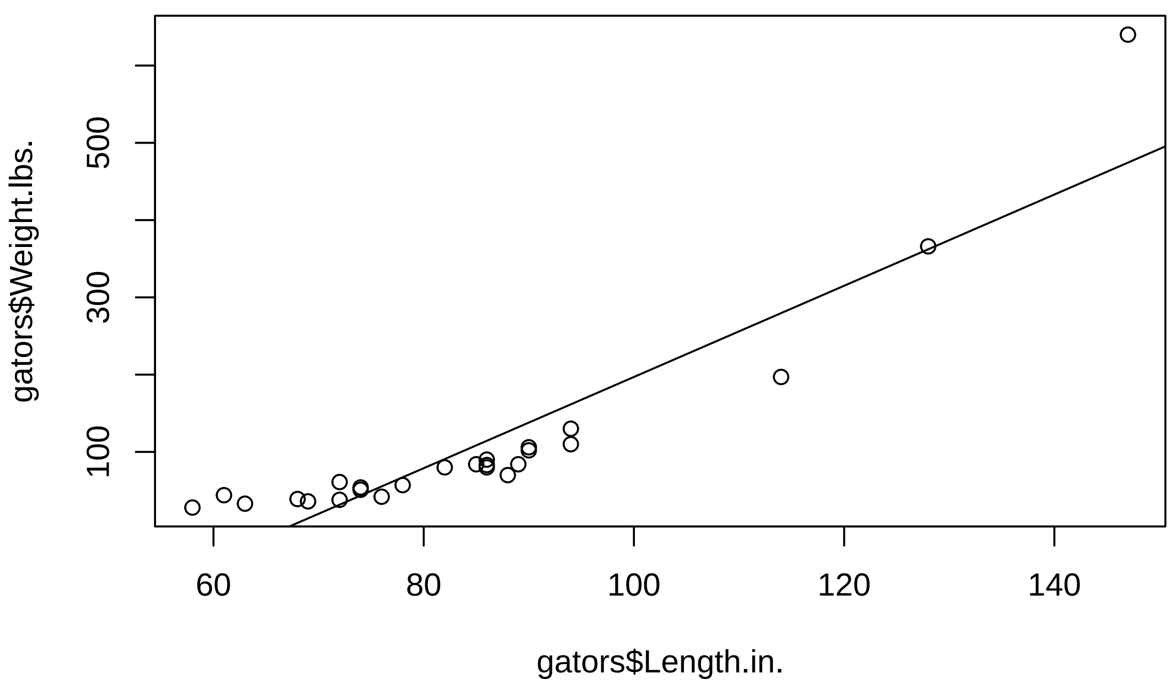
```
##
## Call:
## lm(formula = gators$Weight.lbs. ~ gators$Length.in.)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82.60 -31.95 -10.73  22.00 165.62
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -393.2640    47.5341  -8.273 2.40e-08 ***
## gators$Length.in.     5.9024     0.5448  10.833 1.65e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.01 on 23 degrees of freedom
## Multiple R-squared:  0.8361, Adjusted R-squared:  0.829
## F-statistic: 117.4 on 1 and 23 DF,  p-value: 1.654e-10
```

3. Does the linear model seem to be a good fit for this data? If so, describe why. If not, what could be done to make a linear model more appropriate?

If we plot the data and the regression line, it becomes obvious that the relationship between length (in inches) and weight (in lbs) is non-linear.

```r
plot(gators$Length.in., gators$Weight.lbs.)
abline(slr)
```



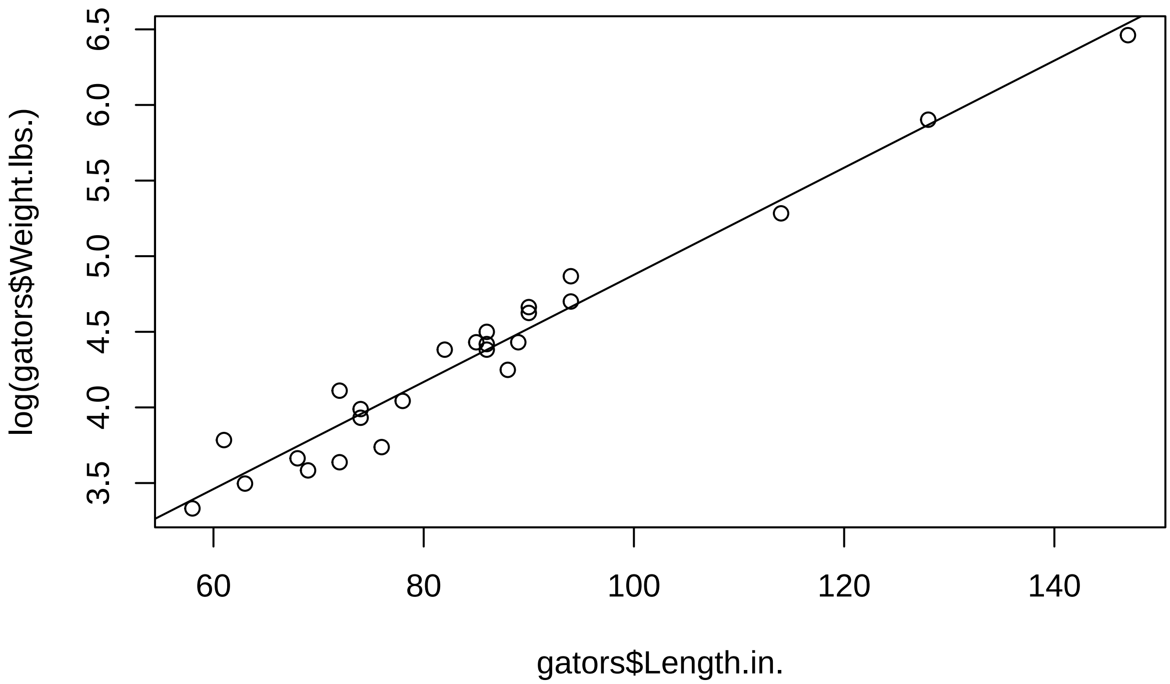
This is a sign that we should try transforming one or more of the variables to see if we can find a more linear relationship. For example, I'm going to transform the response variable by taking the natural logarithm (the function in R for this is `log()`):

```r
slr_transformed <- lm(log(gators$Weight.lbs.) ~ gators$Length.in.)
summary(slr_transformed)
```

```
##
## Call:
## lm(formula = log(gators$Weight.lbs.) ~ gators$Length.in.)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.289266 -0.079989  0.000933  0.102216  0.288491
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.335335   0.131394   10.16 5.63e-10 ***
## gators$Length.in. 0.035416   0.001506   23.52  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1493 on 23 degrees of freedom
## Multiple R-squared:  0.9601, Adjusted R-squared:  0.9583
## F-statistic:   553 on 1 and 23 DF,  p-value: < 2.2e-16
```

```r
plot(gators$Length.in., log(gators$Weight.lbs.))
abline(slr_transformed)
```



This transformation worked! Our estimated regression equation for the transformed data is

$$ln(\hat{Weight}) = 1.34 + 0.035(length).$$

Now the interpretation of the slope is: for each additional inch in length (in the photographs), we expect the natural logarithm of the weight of the aligator to increase by 0.035 lbs. We can un-do the transformation by raising e to the power of the left hand side and the right hand side of our regression equation:

$$e^{ln(\hat{Weight})} = \hat{Weight} = e^{1.34+0.035(length)} = e^{1.34} \times (e^{0.035})^{length}.$$

From simplifying this equation we can tell that the effect of an additional inche in length is actually a multiplicative increase in weight by $e^{0.035} = 1.036$ lbs, on average.

4. The largest residual has a value of 165.62. Explain the meaning of this value in 1-2 sentences.

5. If you were a wildlife researcher who needed to know the different weights of alligators, would you decide to use this method? Give a statistically informed justification of your answer. (Hint: Report and interpret the R^2 value and/or the correlation coefficient.)

Even though the R^2 values is large, I would not use the original linear regression equation to estimate gator weight because the relationship between length (in inches) and weight (in lbs) is non-linear.