**Module 5: Additional Exercises with Answers**

**Example1: Simple Regression Model (1 Variable)**

In this example, we will consider the case of simple linear regression with one response variable and a single independent variable. The data used for this example is from a study in central Florida where 15 alligators were captured and two measurements were made on each of the alligators. The weight (in pounds) was recorded with the snout vent length (in inches – this is the distance between the back of the head to the end of the nose).

The purpose of using this data is to determine whether there is a relationship, described by a simple linear regression model, between the weight and snout vent length. We first create a data frame for this study:

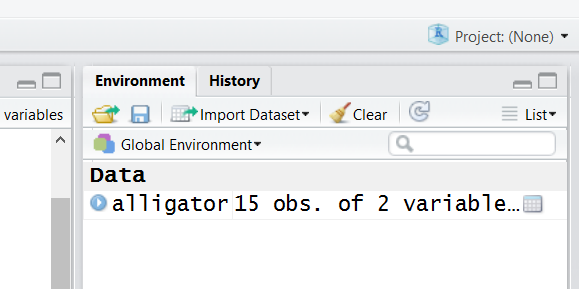
alligator = data.frame(

Length = c(47.94, 36.96, 75.94, 30.87, 45.15, 46.06, 31.81, 42.94, 33.11, 35.87, 66.02, 43.81, 40.85, 41.67, 43.816),

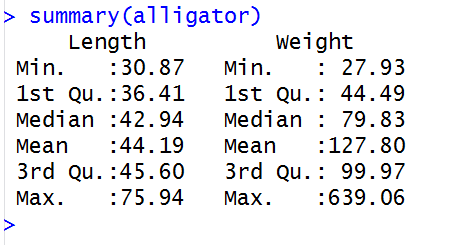
Weight = c(130.32,50.90, 639.06106, 27.93, 79.83, 109.94, 33.11, 90.01, 35.87, 38.09, 365.03, 83.93, 79.83, 83.09, 70.105)

)

By copy and pasting the above lines into the R console you should have a dataframe called alligator in your environment.



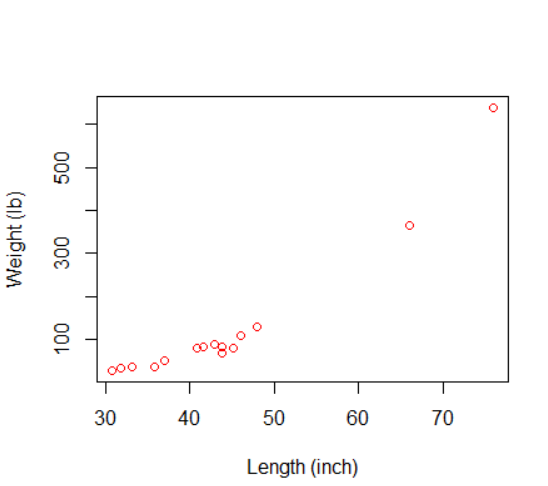
Let’s examine the data:



**Questions**

**Q1. Produce a plot of alligators’ weight against their snout vent length.**

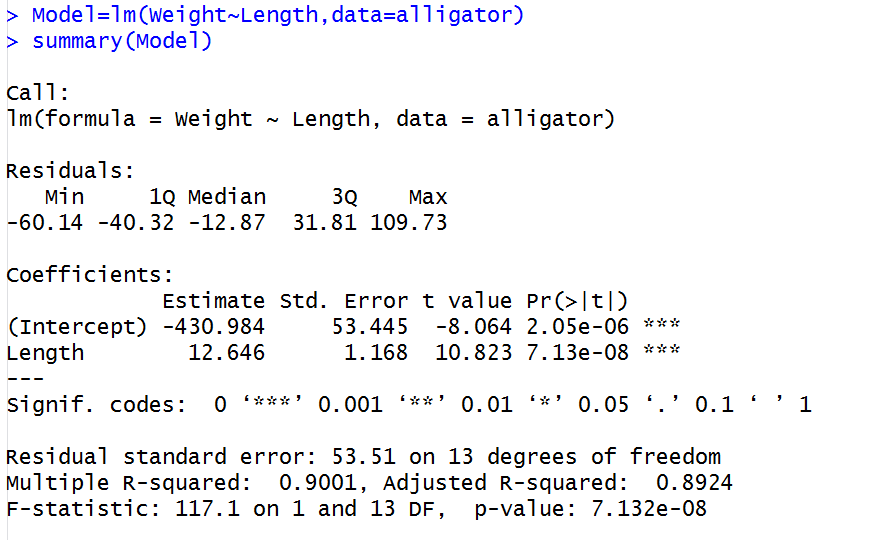
plot(alligator$Weight~alligator$Length, xlab='Length (inch)',ylab='Weight (lb)',col='red')



**Q2. Build a simple linear regression model to express the alligators’ weight based on their snout vent length.**

Model=lm(Weight~Length,data=alligator)

summary(Model)



**Q3. Based on the summary() output, what is the accuracy of this model?**

R2 is 0.90 which means the model explains 90% variability the target (response) variable i.e. the snout vent length is a very good predictor of the weights of the alligators.

**Q4. Is Length (snout vent length) has a statistically significant relationship with weight?**

Yes, the t-value of the coefficient for ‘Length’ variable is 10.823 implying a p-value of 7.13e-8 that means that we can very comfortably reject the default null hypothesis that the coefficient of the ‘Length’ is zero i.e. there is no relationship between Weight and Length.

**Q5. What is the F-statistic is testing?**

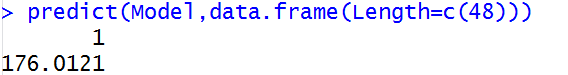
The default null hypothesis that all coefficients (intercept and the coefficient for length) are zero i.e. the whole model is completely useless. The F value is 117.1 which implies a p-value of 7.132e-8 which is very easy to reject this null hypothesis. Of course we already knew from R2 which was very high. A random model can never explain 90% variability of the target variable!

**Q6. Formulate the equation of the weight of the alligators based on their snout vent length. Use the formula to predict the weight of a given alligator which has a snout vent length of 48in.**

Weight(lb)=-430.984+12.646\* length (in)

Weight=-430.984+12.646\* 48= 176.024 (lb)

OR

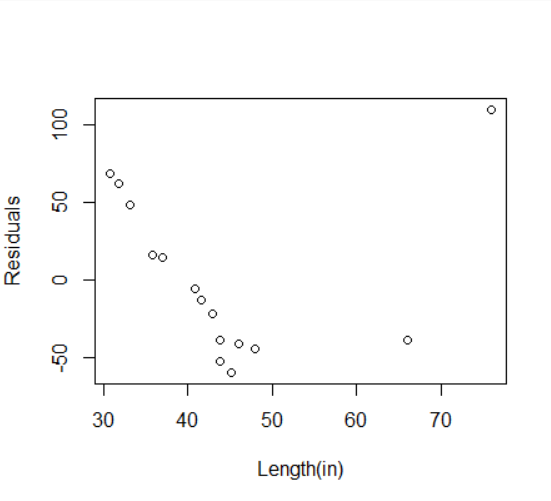


**Q7. The coefficient of Length (snout vent length) is 12.646. What does it mean?**

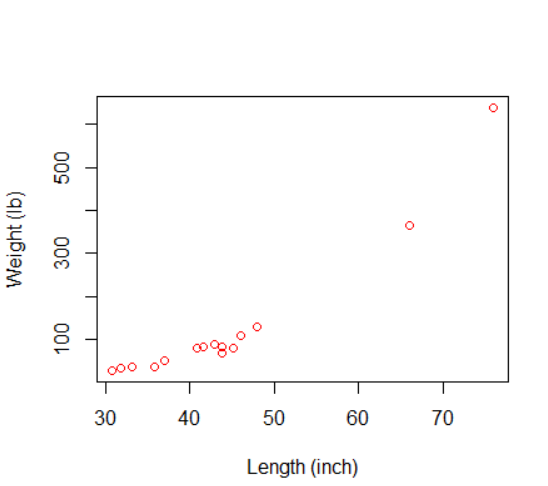
It means that for every additional inch of snout vent length, the weight would be 12.646 pound higher.

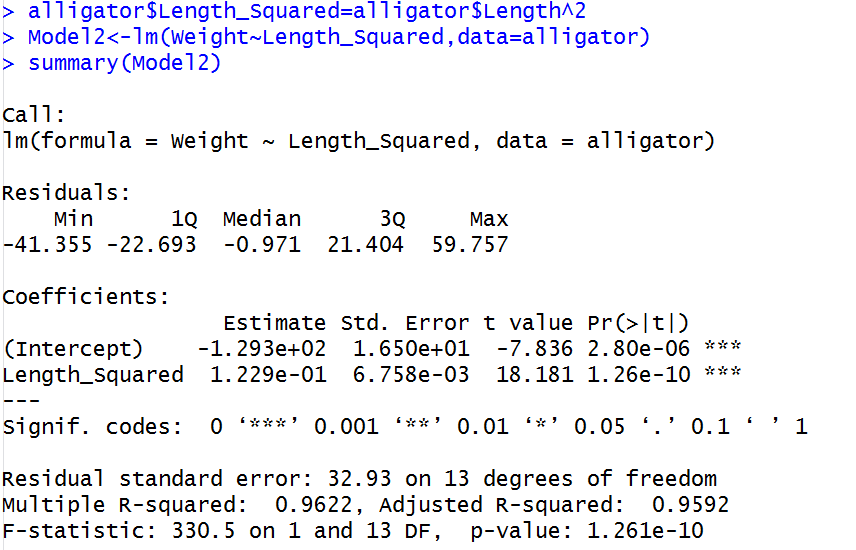
**Q8. Plot the residuals of the model against Length values. Based on this, do you think a simple linear regression model was a right choice here?**

plot(alligator$Length,Model$residuals,xlab='Length(in)',ylab='Residuals')



We can easily see that there is a pattern in the residuals, so linear model is not the best choice here. Looking at the plot of the weight versus length again, we can see that a degree 2 polynomial i.e. y=b0+b1x2 is a better fit. Let’s try that (see next page)





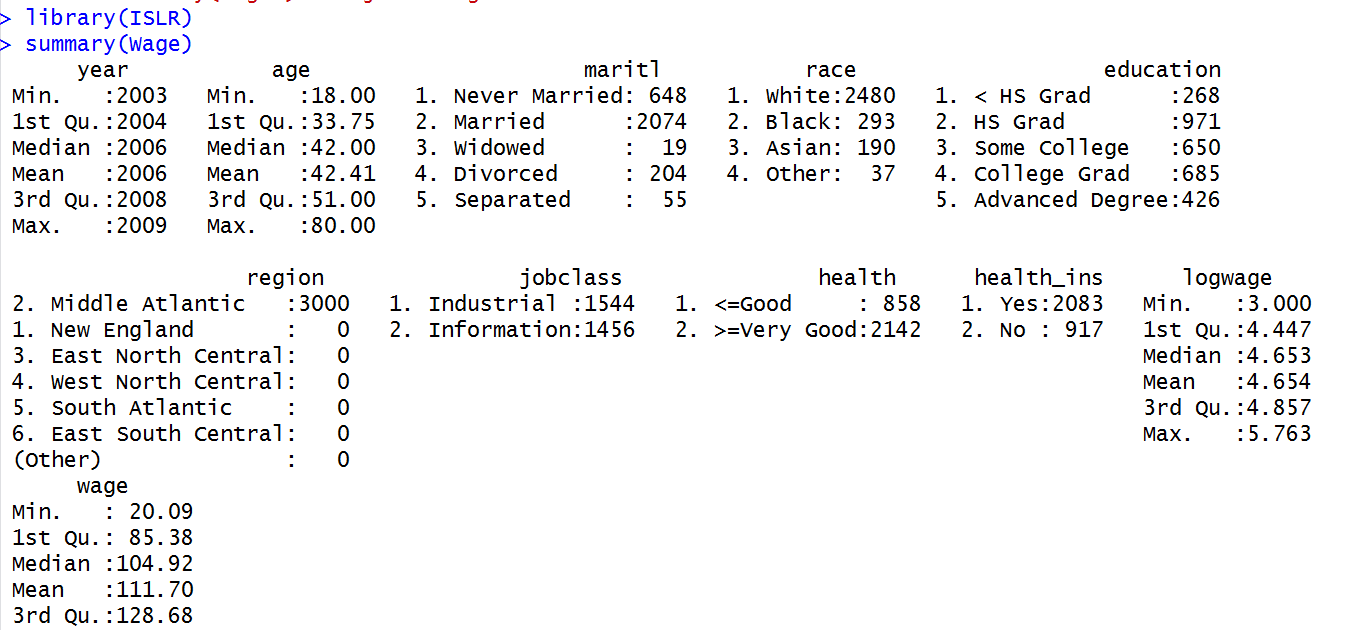
First we created a new variable called Length\_Squared defined as Length^2 and then build a model of Weight based on that. The R2 is improved from 0.90 to 0.962!

**Example 2: Multiple Regression Model (Multiple Variables)**

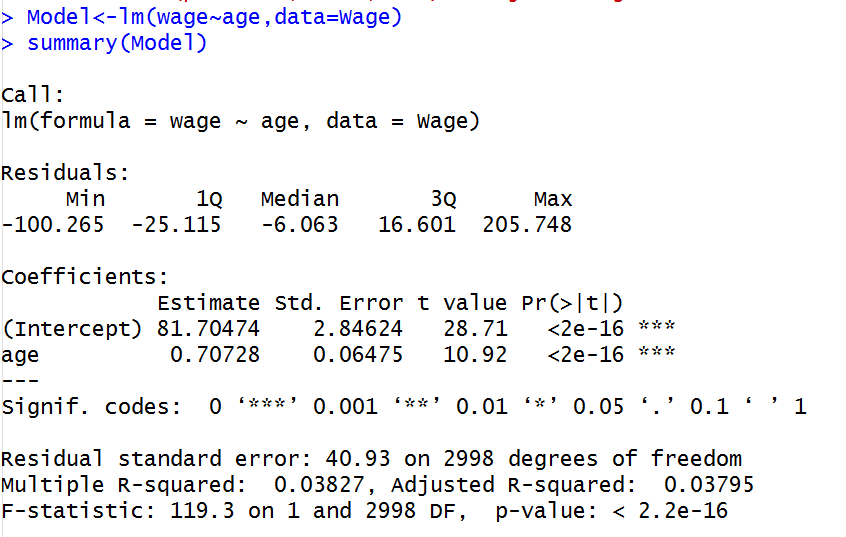
Let’s see if we can build a model to predict the wage of individuals. We use the ‘Wage’ dataset available in the ‘ISLR’ package. If you have not installed the ‘ISLR’ package you can do so using

install.packages(‘ISLR’)

The following shows the summary of the dataset

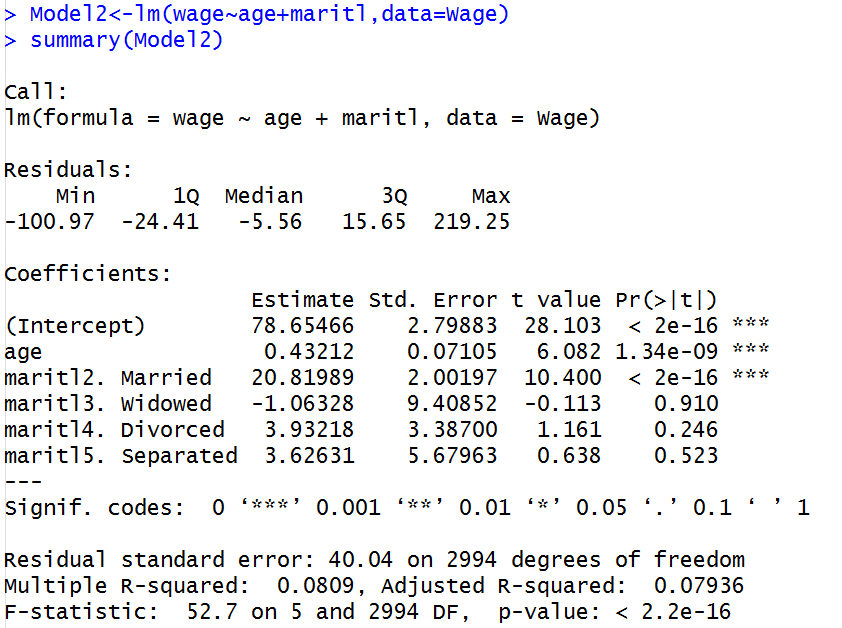


**Q1. Build a model to predict the wage based o the age. How accurate is this model?**



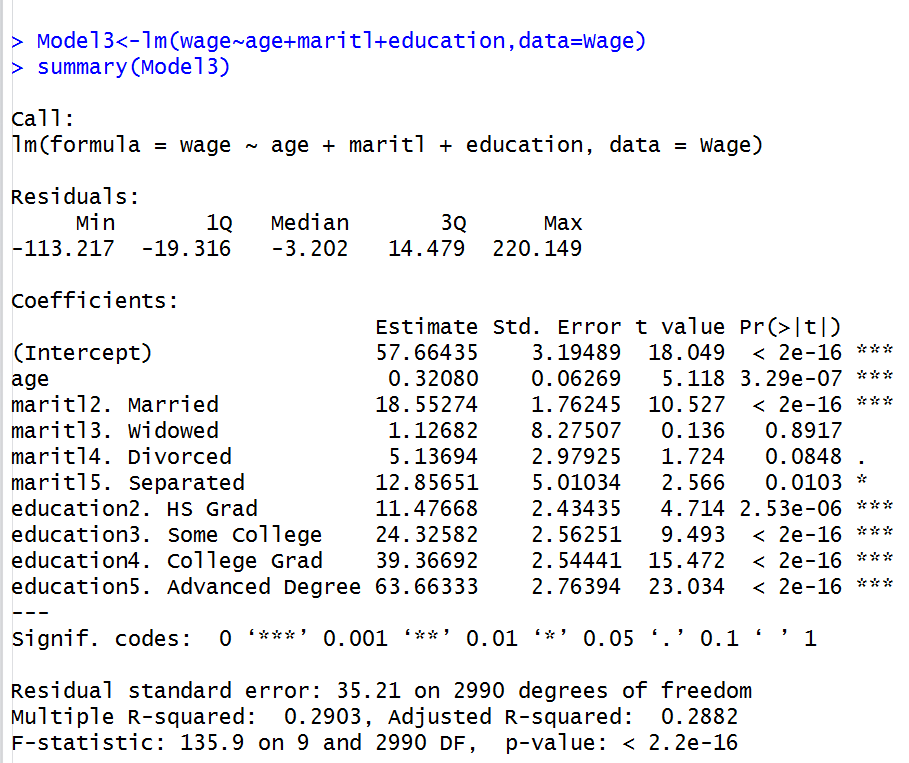
The model is very poor as R2 is only around 3%.

**Q2. Try to improve your model by additionally including the martial staus. Does this improve the model accuracy?**



We have a slightly a better model as the R2 has improved to 8%. Also because marital status ‘maritl’ is a categorical variable, the first level (i.e. never married – see the data frame summary table) is considered as the default for the base model and the coefficients for other values are shown accordingly. This means that, for example, if the individual is married, we add 20.8198 unit to our estimates of wage.

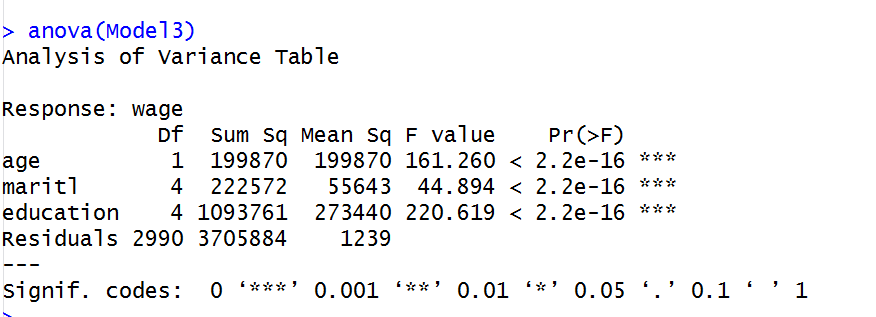
**Q3. Try to improve your model by additionally including the education. Does this improve the model accuracy?**



This is a much better model, with R2 improved to 29%. We can also see that the coefficient for education categories make sense as they increase with the education level. Again, in this case the first education level (i.e. < HS Grad (see summary table) that represents less than High School Graduate) is considered as the default for the base model.

**Q4. What is the order of importance of variables?**

We use the Analysis of Variance (ANOVA) to answer that.



We can see that the variability (sum squared) explained by the education variable is significantly higher than that of age or marital status. We could guess this as adding the education, significantly improved the model. Still we can see that a large portion of the variability is unexplained, that is shown by residuals

**Q5. Can you tell what the value of R2 is by simply looking at the anova output?**

Yes, R2 is the percentage of the total variance that is explained by the model as oppose to what is left out as residuals. The total variability explained by the model in our example is 199870+222572+1093761 the ratio of this against the variability, including what was not capture by the model (i.e. residual) is R2

R2= (199870+222572+1093761)/( 199870+222572+1093761+3705884)= 0.2903443 which is the same number we saw in the model summary screen shot.