Linear Regression: Additional Examples with Answers

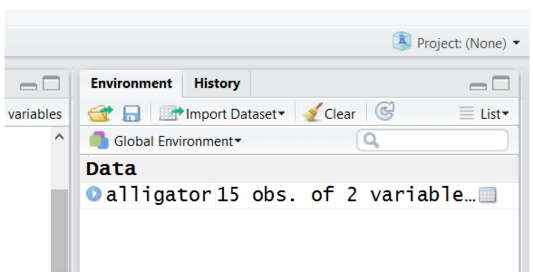
Example1: Simple Regression Model (1 Variable)

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.

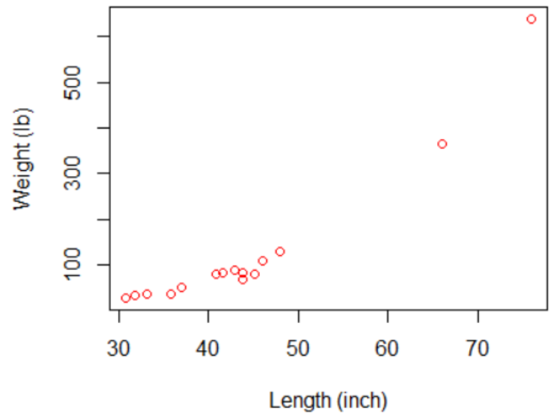


Let's examine the data:

```
> summary(alligator)
     Length          Weight
 Min.   :30.87   Min.   : 27.93
 1st Qu.:36.41   1st Qu.: 44.49
 Median :42.94   Median : 79.83
 Mean   :44.19   Mean   :127.80
 3rd Qu.:45.60   3rd Qu.: 99.97
 Max.   :75.94   Max.   :639.06
>
```

Let's visualize the data.

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



Let's build a simple linear regression model

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

```
> Model=lm(Weight~Length,data=alligator)
> summary(Model)

Call:
lm(formula = Weight ~ Length, data = alligator)

Residuals:
   Min     1Q Median     3Q    Max
-60.14 -40.32 -12.87  31.81 109.73

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -430.984     53.445  -8.064 2.05e-06 ***
Length        12.646      1.168  10.823 7.13e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.51 on 13 degrees of freedom
Multiple R-squared:  0.9001, Adjusted R-squared:  0.8924
F-statistic: 117.1 on 1 and 13 DF,  p-value: 7.132e-08
```

Based on the summary() output, what is the accuracy of this model? R^2 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.

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.

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 R^2 which was very high. A random model can never explain 90% variability of the target variable!

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

```
> predict(Model,data.frame(Length=c(48)))
       1
176.0121
```

What is the 90% and 95% confidence interval of the above prediction?

For 95% interval:

```
> predict(Model,data.frame(Length=c(48)),interval = 'prediction', level=0.95)
       fit      lwr      upr
1 176.0121 56.23499 295.7892
```

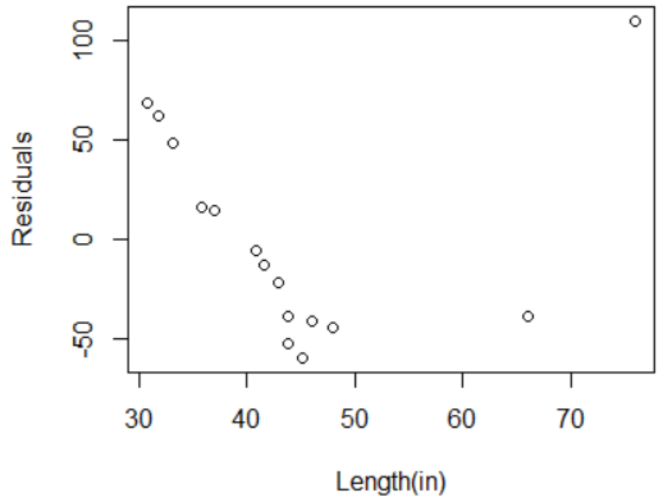
For 90% interval:

```
> predict(Model,data.frame(Length=c(48)),interval = 'prediction', level=0.9)
       fit      lwr      upr
1 176.0121 77.82641 274.1977
```

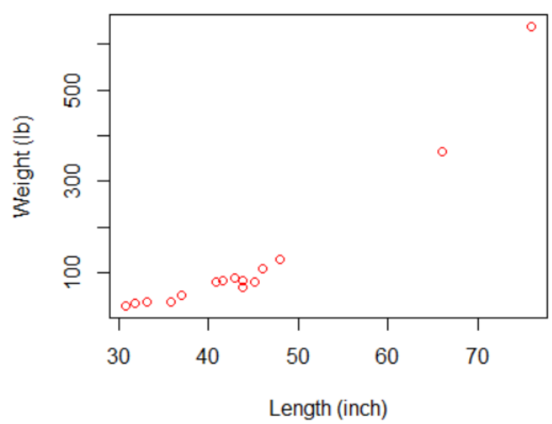
The coefficient of Length (snout vent length) is 12.646. What does it mean? It means that for every additional inch of snout vent length, the weight would be 12.646 pound higher.

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. $y=b_0+b_1x^2$ is a better fit. Let's try that (see next page)



```
> alligator$Length_Squared=alligator$Length^2
> Model2<-lm(Weight~Length_Squared,data=alligator)
> summary(Model2)

Call:
lm(formula = Weight ~ Length_Squared, data = alligator)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.355 -22.693  -0.971  21.404  59.757 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -1.293e+02  1.650e+01  -7.836 2.80e-06 ***
Length_Squared  1.229e-01  6.758e-03  18.181 1.26e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.93 on 13 degrees of freedom
Multiple R-squared:  0.9622,	Adjusted R-squared:  0.9592 
F-statistic: 330.5 on 1 and 13 DF,  p-value: 1.261e-10
```

First we created a new variable called Length_Squared defined as Length^2 and then build a model of Weight based on that. The R^2 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

```
> library(ISLR)
> summary(Wage)
      year           age                       maritl           race                   education
 Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480   1. < HS Grad      :268
 1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293   2. HS Grad        :971
 Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190   3. Some College   :650
 Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37   4. College Grad   :685
 3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55                   5. Advanced Degree:426
 Max.   :2009   Max.   :80.00

                   region                  jobclass             health        health_ins       logwage
 2. Middle Atlantic   :3000   1. Industrial :1544   1. <=Good     : 858   1. Yes:2083   Min.   :3.000
 1. New England       :   0   2. Information:1456   2. >=Very Good:2142   2. No : 917   1st Qu.:4.447
 3. East North Central:   0                                                             Median :4.653
 4. West North Central:   0                                                             Mean   :4.654
 5. South Atlantic    :   0                                                             3rd Qu.:4.857
 6. East South Central:   0                                                             Max.   :5.763
 (Other)              :   0
      wage
 Min.   : 20.09
 1st Qu.: 85.38
 Median :104.92
 Mean   :111.70
 3rd Qu.:128.68
```

Let's try to build a model based on age.

```
> Model<-lm(wage~age,data=Wage)
> summary(Model)

Call:
lm(formula = wage ~ age, data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max
-100.265  -25.115   -6.063   16.601  205.748

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.70474    2.84624   28.71   <2e-16 ***
age          0.70728    0.06475   10.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.93 on 2998 degrees of freedom
Multiple R-squared:  0.03827,	Adjusted R-squared:  0.03795
F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
```

The model is very poor as R^2 is only around 3%. So let's see if we can improve the model by additionally including the marital status.

```
> Model2<-lm(wage~age+maritl,data=Wage)
> summary(Model2)

Call:
lm(formula = wage ~ age + maritl, data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max 
-100.97  -24.41   -5.56   15.65  219.25 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        78.65466    2.79883  28.103  < 2e-16 ***
age                 0.43212    0.07105   6.082 1.34e-09 ***
maritl2. Married   20.81989    2.00197  10.400  < 2e-16 ***
maritl3. Widowed   -1.06328    9.40852  -0.113    0.910    
maritl4. Divorced   3.93218    3.38700   1.161    0.246    
maritl5. Separated  3.62631    5.67963   0.638    0.523    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.04 on 2994 degrees of freedom
Multiple R-squared:  0.0809,	Adjusted R-squared:  0.07936 
F-statistic:  52.7 on 5 and 2994 DF,  p-value: < 2.2e-16
```

Ok slightly a better model as the R^2 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.

R2 of 8% is still too weak, let’s say if we can improve the model further by adding the education.

```
> Model3<-lm(wage~age+maritl+education,data=Wage)
> summary(Model3)

Call:
lm(formula = wage ~ age + maritl + education, data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max 
-113.217  -19.316   -3.202   14.479  220.149 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 57.66435    3.19489  18.049  < 2e-16 ***
age                          0.32080    0.06269   5.118 3.29e-07 ***
maritl2. Married            18.55274    1.76245  10.527  < 2e-16 ***
maritl3. Widowed             1.12682    8.27507   0.136   0.8917    
maritl4. Divorced            5.13694    2.97925   1.724   0.0848 .  
maritl5. Separated          12.85651    5.01034   2.566   0.0103 *  
education2. HS Grad         11.47668    2.43435   4.714 2.53e-06 ***
education3. Some College    24.32582    2.56251   9.493  < 2e-16 ***
education4. College Grad    39.36692    2.54441  15.472  < 2e-16 ***
education5. Advanced Degree 63.66333    2.76394  23.034  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.21 on 2990 degrees of freedom
Multiple R-squared:  0.2903,	Adjusted R-squared:  0.2882 
F-statistic: 135.9 on 9 and 2990 DF,  p-value: < 2.2e-16
```

This is a much better model, with R^2 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.

What is the order of importance of variables?

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

```
> anova(Model3)
Analysis of Variance Table

Response: wage
            Df  Sum Sq Mean Sq F value    Pr(>F)    
age          1  199870  199870 161.260 < 2.2e-16 ***
maritl       4  222572   55643  44.894 < 2.2e-16 ***
education    4 1093761  273440 220.619 < 2.2e-16 ***
Residuals 2990 3705884    1239                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Can you tell what the value of R^2 is by simply looking at the anova output? Yes, R^2 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 R^2

R^2= (199870+222572+1093761)/(199870+222572+1093761+3705884)= 0.2903443 which is the same number we saw in the model summary screen shot.