ASSIGNMENT 6 – Linear Regression 2

In this assignment we will implementing two multiple linear regression models. The dataset we will be working with will be the same used in Assignment 5 i.e. **countyComplete.csv**. We will be using the same variable i.e. **poverty** as our response variable and two or more variables as our explanatory variable.

We start by reading the file by using the **read.csv** command.

```
require( mosaic )
county = read.csv( "countyComplete.csv" )
```

We will now use the **median population income** variable along with **per capita income** as our explanatory variables for the poverty estimate.

In R, we use the **lm()** command to generate multiple linear regression models like the simple linear models. We will generate the regression model in the following way.

```
model = lm(county$poverty~county$median_household_income+county$per_capita_income)
```

We will now use the **summary()** function to get the information stored in our model.

```
summary(model)
```

```
Call:
lm(formula = county$poverty ~ county$median_household_income +
    county$per_capita_income)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9081  -2.7607  -0.4896   2.1561  28.0040

Coefficients:
                               Estimate Std. Error t value
(Intercept)                   3.489e+01  3.237e-01 107.786
county$median_household_income -2.966e-04  1.320e-05 -22.470
county$per_capita_income       -2.783e-04  2.818e-05  -9.875
                              Pr(>|t|)
(Intercept)                     <2e-16 ***
county$median_household_income  <2e-16 ***
county$per_capita_income        <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.219 on 3140 degrees of freedom
Multiple R-squared:  0.5636,  Adjusted R-squared:  0.5633
F-statistic:  2027 on 2 and 3140 DF,  p-value: < 2.2e-16
```

The above output shows the information contained in our model.

Before explaining this model let us see how the explanatory variables perform individually in a simple linear model. We will first take the **median household income** variable. To create a linear model using this variable, we run the code below and get its summary.

```
m1 =lm(county$poverty~county$median_household_income)
```

```
summary(m1)
```

```
Call:
lm(formula = county$poverty ~ county$median_household_income)

Residuals:
    Min      1Q  Median      3Q     Max
-12.4094  -2.8108  -0.4418   2.1569  29.8510

Coefficients:
                              Estimate Std. Error t value
(Intercept)                  3.365e+01  3.027e-01  111.15
county$median_household_income -4.100e-04  6.617e-06  -61.96
                             Pr(>|t|)
(Intercept)                   <2e-16 ***
county$median_household_income   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.283 on 3141 degrees of freedom
Multiple R-squared:   0.55,         Adjusted R-squared:  0.5499
F-statistic:  3839 on 1 and 3141 DF,  p-value: < 2.2e-16
```

Let us now consider the **per capita income** variable. We will create the model in the same way we created for the previous variable.

m2 =lm(county$poverty~county$per_capita_income)

summary(m2)

```
Call:
lm(formula = county$poverty ~ county$per_capita_income)

Residuals:
   Min     1Q  Median     3Q    Max
-12.931  -3.017  -0.571   2.260  32.681

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.416e+01  3.469e-01   98.45   <2e-16
county$per_capita_income -8.291e-04  1.499e-05  -55.31   <2e-16

(Intercept)           ***
county$per_capita_income ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.545 on 3141 degrees of freedom
Multiple R-squared:  0.4934,      Adjusted R-squared:  0.4932
F-statistic:  3059 on 1 and 3141 DF,  p-value: < 2.2e-16
```

Now we will do a comparative analysis of all the three models to see which model performs better in terms of explaining the **poverty** variable.

Like the previous assignment we will be explaining the individual sections in the summary table.

```
lm(formula = county$poverty ~ county$median_household_income +
    county$per_capita_income)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9081  -2.7607  -0.4896   2.1561  28.0040
```

The first item shows the Call functionality. This only points out to our formula for generating the linear model. The next item are the residuals.

The residuals are the difference between the actual value observed versus the value predicted by the model. If the residuals are symmetrical with the median around 0, it means the model is a good fit. Here we can observe that the that the residuals are not uniformly distributed for the model. Though the median is close to zero the uniformity of the residual distributed signifies a better fit.

```
Coefficients:
                                Estimate
(Intercept)                    3.489e+01
county$median_household_income -2.966e-04
county$per_capita_income       -2.783e-04
```

The next section is the coefficient outputs. The first section is the intercept for the model. The intercept in our case is the expected value of the percentage below poverty level by taking the average median household income. Here its comes out to be more than 34%. The second value in the estimate column shows how much change can we observe in the poverty variable given a unit change in the median household income and per capita income. Here it comes out to be 0.00029% and 0.00027% decrease in percentage for the respective variables which is minute.

```
Std. Error
3.237e-01
1.320e-05
2.818e-05
```

The next column in the coefficient is the standard error. A standard error is the estimate of the difference in the coefficient values if we run the model repeatedly. It is a measure of the error in our estimates of the coefficients. We want them to be lower as compared to our coefficients. In this case for the expected poverty percentage it is 0.323 %.

```
Multiple R-squared:  0.5636,  Adjusted R-squared:  0.5633
```

We now come to the R-squared statistic which provides us a measure of how well the model is fitting the actual data. It takes in the form of how much variance in our response variable is explained by our explanatory variable. Here in our case it is 0.5636 which is 56%. So, our model is able to explain 56% of the variance in the poverty percentage variable.

Observe that the amount of variance explained in simple linear models came out to be 55% and 49% for the median household income and per capita income respectively. This is a slight improvement over **per capita income** model but almost same in case of **median household income** model. So, we can observe that this multiple linear model does not perform sufficiently better that one of the individual variables. Let us take another example.

We use the variable **home ownership** as one of the other explanatory variable. We will see how using it performs to predict the poverty variable. The code along with output is below.

m3 =lm(county$poverty~county$home_ownership)

summary(m3)

```
Call:
lm(formula = county$poverty ~ county$home_ownership)

Residuals:
   Min     1Q  Median     3Q    Max
-30.650  -4.101  -0.920   3.244  32.230

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         34.74999    1.01446   34.26   <2e-16 ***
county$home_ownership -0.26276    0.01377  -19.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.044 on 3141 degrees of freedom
Multiple R-squared:  0.1039,      Adjusted R-squared:  0.1036
F-statistic: 364.2 on 1 and 3141 DF,  p-value: < 2.2e-16
```

We can observe for the multiple R-square value i.e. 10% that this is the lowest in terms of all the models we have seen thus far. Let us develop a multiple linear model using **per capita income** and **home ownership** variables. Also note that we have already using **home ownership** variable along with **median household income** variable previously in assignment 5. It performed really well as compared to simple linear models. Let us see if it also holds true with **the per capita income** variable. We create the model and gets its summary as below.

model2 = lm(county$poverty~county$per_capita_income+county$home_ownership)

summary(model2)

```
Call:
lm(formula = county$poverty ~ county$per_capita_income + county$home_ownership)

Residuals:
    Min      1Q   Median      3Q     Max
-15.2467  -2.4732  -0.3789   1.9941  25.0363

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.461e+01  7.401e-01   73.79   <2e-16
county$per_capita_income -8.384e-04  1.319e-05  -63.57   <2e-16
county$home_ownership    -2.763e-01  9.108e-03  -30.34   <2e-16

(Intercept)             ***
county$per_capita_income ***
county$home_ownership    ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.997 on 3140 degrees of freedom
Multiple R-squared:  0.6082,      Adjusted R-squared:  0.608
F-statistic:  2437 on 2 and 3140 DF,  p-value: < 2.2e-16
```

Here we can see that this model better explains the variance in the poverty percentage variable. The adjusted R-square value is 60.82% which is better than the previous 55% we had while using only the **median household income** variable. It is also better than the 56% R-square value we got from the first multiple linear model.

Let us now consider all the variables to explain the response variable. We generate this multiple linear model in the following way.

model3 = lm(county$poverty ~ county$median_household_income + county$per_capita_income + county$home_ownership)

summary(model3)

```
Call:
lm(formula = county$poverty ~ county$median_household_income +
   county$per_capita_income + county$home_ownership)

Residuals:
    Min      1Q  Median      3Q     Max
-21.7173  -2.3286  -0.3903   1.8178  22.5055

Coefficients:
                              Estimate Std. Error t value
(Intercept)                  5.294e+01  7.031e-01   75.29
county$median_household_income -2.369e-04  1.199e-05  -19.76
county$per_capita_income      -3.976e-04  2.554e-05  -15.57
county$home_ownership         -2.458e-01  8.728e-03  -28.16
                             Pr(>|t|)
(Intercept)                   <2e-16 ***
county$median_household_income   <2e-16 ***
county$per_capita_income       <2e-16 ***
county$home_ownership          <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.77 on 3139 degrees of freedom
Multiple R-squared:  0.6516,      Adjusted R-squared:  0.6512
F-statistic:  1957 on 3 and 3139 DF,  p-value: < 2.2e-16
```

From the R-square output value i.e. 65.16% we can see that this is the best performing linear model we have seen so far. We can observe from the above analysis that multiple linear regression models are able to perform better than the simple linear models in terms of explaining the response variable, but it also depends on the choice of variable used as an explanatory variables. We really need to choose our explanatory variables carefully to get a better performing model. In this case using all the variables helped in getting a model that performs better that each of the simple linear models.