Homework Assignment 6 Nihit Hindlekar

Construct two linear models, each with two or more explanatory variable using the poverty measure in countyComplete as the response variable. The choice of explanatory variables should be guided by your earlier work on assignments 5 and 3. For example, you could combine in one model two explanatory variables you examined separately.
Compute the amount of variance accounted for in the model using r-squared (the square of the correlation between the fitted values and the response variable). See how much additional variance is accounted when you add the additional explanatory variables. If you wish to comment on the relative importance of the variables in the model, don't rely on the model coefficients but instead compare correlations between the variables and the fitted values.
Write a 1-3 page discussion of your analysis, revisiting and comparing your results with the results you reported in assignments 3 and 5. What new insights does the multiple regression model offer on the data and its relevance to poverty at the county level?

What is Multiple Linear regression?

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for y vary about their means μ_y and are assumed to have the same standard deviation σ. The fitted values $b_0, b_1, \ldots, b_p$ estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the population regression line.

Since the observed values for y vary about their means μ_y, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$. The "RESIDUAL" term represents the deviations of the observed values y from their means μ_y, which are normally distributed with mean 0 and variance σ. The notation for the model deviations is ε.

Formally, the model for multiple linear regression, given *n* observations, is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i$ for $i = 1,2, \ldots n$.

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then

summed, there are no cancellations between positive and negative values. The least-squares estimates b_0, b_1, ... b_p are usually computed by statistical software.

The values fit by the equation $b_0 + b_1x_{i1} + ... + b_px_{ip}$ are denoted $\hat{y}_i$, and the residuals e_i are equal to $y_i - \hat{y}_i$, the difference between the observed and fitted values. The sum of the residuals is equal to zero.

The variance σ^2 may be estimated by $s^2 = \frac{\sum e_i^2}{n - p - 1}$, also known as the mean-squared error (or MSE).
The estimate of the standard error s is the square root of the MSE. [1]

Model 1:

In this model, we are taking into consideration the two predictor variables (hs_grad and age_under_18) from the countyComplete data set to predict the outcome 'poverty'.

Steps(Code):

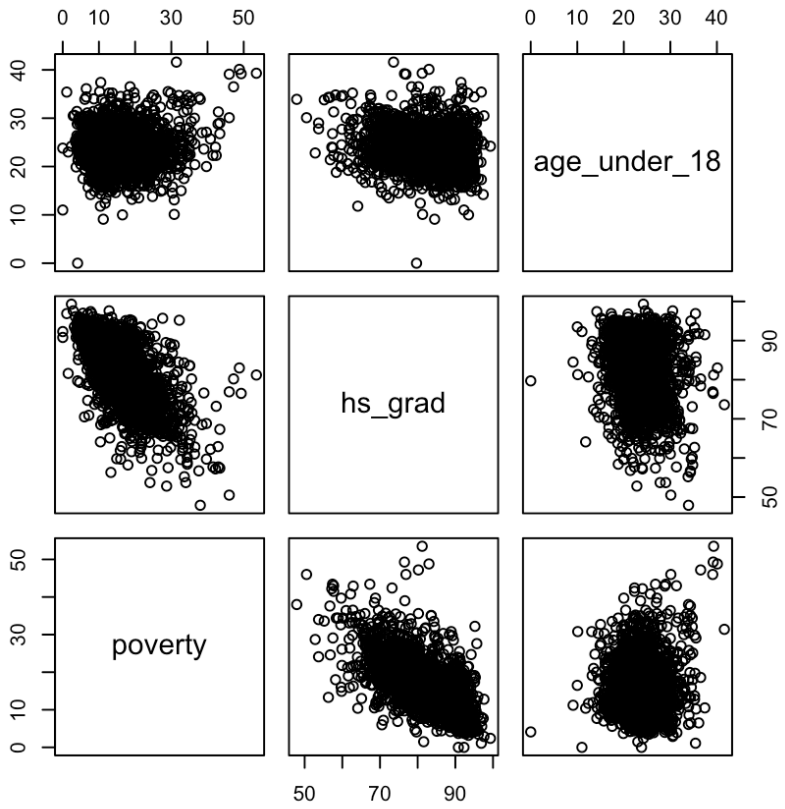
1.First we are reading the "countyComplete" data set into a variable named pov.

```
pov <- read.csv("countyComplete.csv")
```

2.We are plotting the graph to show the relation between the variables. We are using the pairs() function in R.

```
> pairs(poverty ~ hs_grad + age_under_18 ,data = pov, row1attop=FALSE)
```

A diagnostic scatterplot is plotted by default as the variables used are quantitative:



From the above plot we can infer relations between our predictor variables and the response variable as well as the relationship between our two predictor variables to check collinearity.

3. We then construct the linear model with our variables to predict the poverty outcome as below

```
> lmod <- lm(poverty ~ hs_grad + age_under_18, data= pov)
```

4. After the defining the linear model we check the coefficients of the model by using the coef() function which returns the coefficients of all the predictor variables that we are using in the model as well as the intercept.

```
> coef(lmod)

 (Intercept)      hs_grad age_under_18

 66.61745501  -0.59583430  -0.06832745
```

We are also calculating the fitted values to interpret the correlation between the variables used in the model by using the fitted() command in R as below

```
> fitted2 <- fitted(lmod)
```

5. Later we Compute the amount of variance accounted for in the model using r-squared (the square of the correlation between the fitted values and the response variable).

```
> cor(fitted2, pov$poverty) ^ 2

[1] 0.4647489
```

We will later compare this variance with our accounted variance in the model 2.

6. We then check the correlation between the variables and fitted values to interpret the relative importance of each variable with the model outcome.

```
> cor(fitted2, pov$hs_grad)

[1] -0.9986323

> cor(fitted2, pov$age_under_18)

[1] 0.1105146
```

As we can see the output we are getting for the variable (hs_grad) is a negative value, which means that when the predictor variable "hs_grad" is increasing, the value of response variable "poverty" is decreasing. Hence, it is strong negative correlation and we can assume the predictor variable hs_grad is relatively more more important in the model as compared to the other predictor variable.

Model 2:

In model 1, we are taking into consideration the two predictor variables (hs_grad and age_under_18) with the addition of two more predictors(per_capita_income & age_over_65)from the countyComplete data set to predict the outcome 'poverty'.

Steps(Code):

1. First we are reading the "countyComplete" data set into a variable named pov.

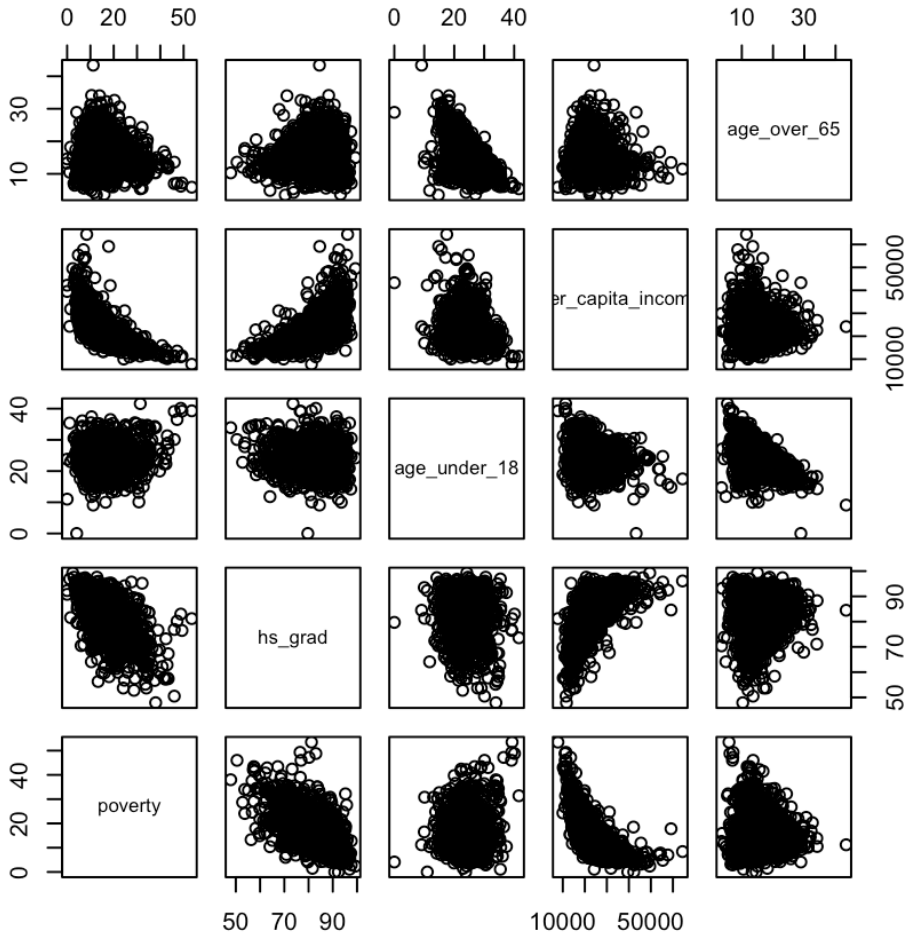
```
pov <- read.csv("countyComplete.csv")
```

2.We are plotting the graph to show the relation between the variables. We are using the pairs()

function in R.

```
> pairs(poverty ~ hs_grad + age_under_18 + per_capita_income + age_over_65 ,data = pov, row1attop=FALSE)
```

A diagnostic scatterplot is plotted by default as the variables used are quantitative:



From the above plot we can infer relations between our predictor variables and the response variable as well as the relationship in between our predictor variables to check collinearity.

3. We then construct the linear model with our variables to predict the poverty outcome as below

```
> lmod2 <- lm(poverty ~ hs_grad + age_under_18 + per_capita_income + age_over_65, data= pov)
```

Here we are adding two more predictor variables to the model namely; per_capita_income and age_over_65 to test the accuracy of variance.

4. After the defining the linear model we check the coefficients of the model by using the coef() function which returns the coefficients of all the predictor variables that we are using in the model.

```
> coef(lmod2)

     (Intercept)          hs_grad     age_under_18  per_capita_income

   66.2126005037   -0.3183458363   -0.2708995739   -0.0005955181

      age_over_65

   -0.2839680177

> fitted3 <- fitted(lmod2)
```

5. Later we Compute the amount of variance accounted for in the model using r-squared (the square of the correlation between the fitted values and the response variable) using the cor() function.

```
> cor(fitted3, pov$poverty) ^ 2

[1] 0.609649
```

Here we see that the computed variance is higher than the accounted variance for model 1.

Since the first model has an r-squared less than the second model, we would prefer the second model to the first for predicting the outcome. Hence, adding the two more explanatory variables(per_capita_income & age_over_65) to our model we see that the accuracy of variance increases as well as the predictability of the model.

6. We then check the correlation between the predictor variables and fitted values to interpret the relative importance of each variable with the model outcome.

```
> cor(fitted3, pov$per_capita_income)

[1] -0.8996078

> cor(fitted3, pov$age_over_65)

[1] -0.1118169
```

```
> cor(fitted3, pov$hs_grad)

[1] -0.8719164

> cor(fitted3, pov$age_under_18)

[1] 0.09649147
```

As we can see the output we are getting for the two variables (per_capita_income & hs_grad) is a negative value, which means that when the predictor variables "hs_grad" & "per_capita_income" are increasing, the value of response variable "poverty" is decreasing. Hence, it is strong negative correlation and we can assume both the predictor variables are relatively important in the model as compared to the other two predictor variables.

Comparison:

If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome poverty and predictor hs_grad earlier using simple linear regression, we were unable to control for other variables like the number of people under some age ranges(age_under_18 & age_over_65) included in the county. That model was biased. When we use all variables, this particular underlying and unintentional bias is reduced or eliminated (though bias from other confounding variables may still remain).
As we inferred from the earlier assignments that linear models are more accurate and useful in predicting outcomes as it measures the strength in the linear relationship between the two variables and it becomes easier to predict the outcome at any interval in time for the given explanatory variable. Because as discussed earlier, linear model helps in the prediction of future observations. Certain things such as the fitted line in linear model helps in understanding the linear relationship between the two variables more easily than the group-based models. Hence, linear models are more preferred to the group based models in practicality.
But to add to our latest prediction with multiple linear models or multiple linear regression, we can clearly say that the multiple linear model helps calculate or predict the outcome more accurately because as mentioned earlier above, we are able to control predictions from all the variables in one model reducing or eliminating the bias from the earlier simple linear or group-based models.

Citations:

1. http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm