Construct a linear model with one explanatory variable using the poverty measure in **countyComplete** as the response variable. Use one of the explanatory variables you chose in Homework 3, as long as at least one of those is a quantitative variable.

Write a 1-3 page discussion of your analysis, revisiting and comparing your results with the results you reported in Homework 3. What new insights does the linear model offer on the data and its relevance to poverty at the county level?

What is a Linear model?

Linear model is used for explaining or modeling the relationship between a single variable Y, called the response, output or dependent variable; and one or more predictor, input, independent or explanatory variables, X1 ,…, Xp . When p=1, it is called simple regression but when p>1 it is called multiple regression or sometimes multivariate regression. When there is more than one Y, then it is called multivariate multiple regression which we will not be covering explicitly here, although we can just do separate regressions on each Y.
The response must be a continuous variable, but the explanatory variables can be continuous, discrete or categorical.[1]

Regression analyses have several possible objectives including:
1. Prediction of future observations
2. Assessment of the effect of, or relationship between, explanatory variables and the response
3. A general description of data structure

The general equation for a linear model is:

$$y = \beta_0 + \beta_1 x$$

where β_0, β_1 represent the two linear model parameters.
where β_0 and β_1 represent two model parameters. These parameters are estimated using data, and we write their point estimates as b_0 and b_1. When we use x to predict y, we usually call x the explanatory or predictor variable, and we call y the response.

In this document, we will be assessing the effect of the number of high school graduates (“hs_grad”) on the rate of poverty in the county(“countyComplete”).
We are using only one explanatory variable(x) to model the relationship between “x” and the response variable(y). In this case, from the countyComplete data set:
x = hs_grad
y = poverty

Steps(Code):

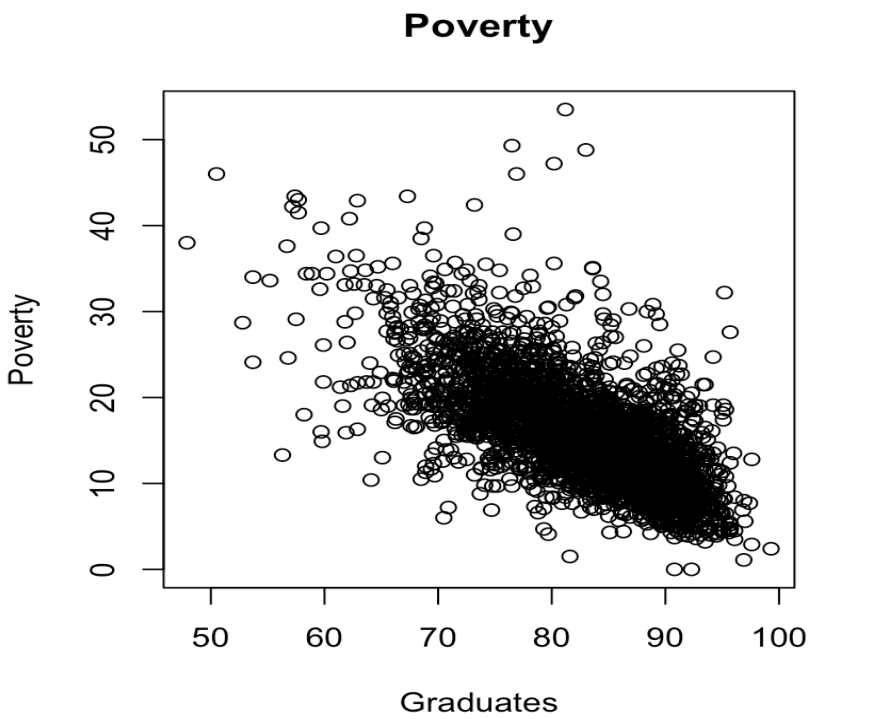
1.First we are reading the “countyComplete” data set into a variable named pov.

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

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

```
plot(pov$poverty~pov$hs_grad, xlab='Graduates', ylab='Poverty', main = 'Poverty')
```

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



3. We are then checking the correlation between the two variables for finding out the strength in the relationship between the two. We can compute the correlation using a formula, just as we did with the sample mean and standard deviation previously. However, we are using the cor() function in R to calculate the correlation between the two variables as below:

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

[1] -0.6807925
```

where we are passing the arguments as our two variables. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.

As we see the output we are getting 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.

4. Now we are using the lm() function in R to construct a linear model for the two variables as below:

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

Call:
lm(formula = poverty ~ hs_grad, data = pov)

Coefficients:
```

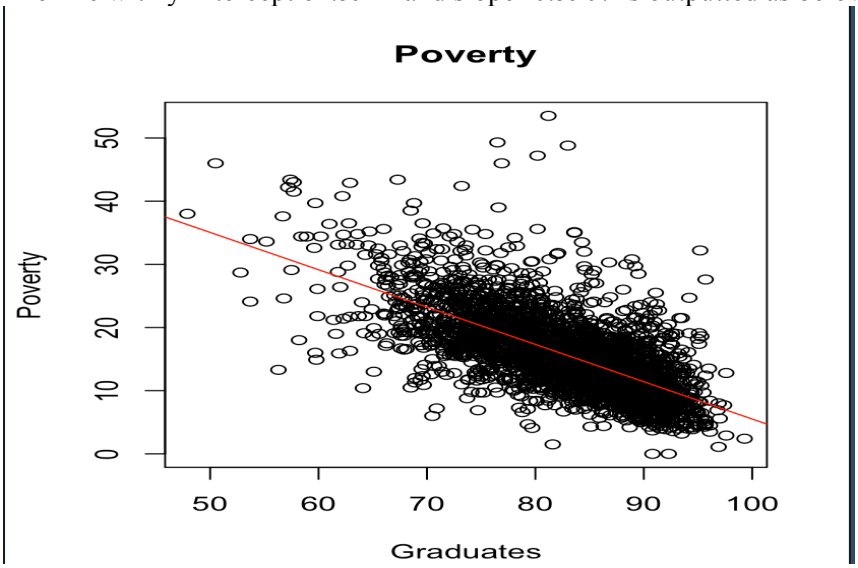
```
(Intercept)      hs_grad
   64.5944      -0.5907
```

(where the formula contains the first variable used as the response while the second as the predictor from the dataset and the output shows the β_0 and β_1 values as the intercept 64.5944 and slope -0.5907 respectively.)

5. We then plot the linear model along with the best fit line as below:

```
>abline(linear_model, col='red')
```

The line with y-intercept 64.5944 and slope -0.5907 is outputted as below:



6. We then use the summary() command as below:

```
> summary(lmod)

Call:
lm(formula = poverty ~ hs_grad, data = pov)

Residuals:
    Min      1Q  Median      3Q     Max
-18.035  -3.034  -0.434   2.405  36.874

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.59437    0.94619   68.27   <2e-16 ***
hs_grad     -0.59075    0.01134  -52.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.677 on 3141 degrees of freedom
Multiple R-squared:  0.4635,      Adjusted R-squared:  0.4633
F-statistic:  2713 on 1 and 3141 DF,  p-value: < 2.2e-16
```

Here, we are summarizing all the necessary details of the linear model that we had constructed earlier. We understand the residual value with the limits within which they fall, the y-intercept and the slopes for the line along with the residual standard error.

Comparison:

In comparison to the group based model that we had constructed for measuring the poverty using the explanatory variable "hs_grad" previously, the linear model gives us the strengthening relationship between the two variables. As a linear model is an approximation of the real relationship between two variables.
According to our model, the relationship between the two variables is declining. i.e. when the number of hs_grads in the county increases, the number of poverty in the county falls. As the number of people in a state having completed high school studies contributes directly somehow to the poverty in that particular county. Since, the maximum number of people who have not completed their high school tend to remain jobless without earning any respectable amount of income further contributing lesser to the overall state income which in turn leads into more poverty. For example, higher the rate of illiterate people in a state, more the poverty in that particular state. Hence, this model looks as a good fit 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.

Screens of the code:

```
#read the data
pov <- read.csv("countyComplete.csv")

#plotting the relation between the variables
plot(pov$poverty~pov$hs_grad, xlab='Graduates', ylab='Poverty', main = 'Poverty')

#calculating correlation
cor(pov$poverty,pov$hs_grad)

#linear model
lmod <- lm(poverty ~ hs_grad, data= pov)

#fitting the line
abline(linear_model, col='red')
```

Citations:

1. http://www.utstat.toronto.edu/~brunner/books/LinearModelsWithR.pdf