BIG DATA MINI PROJECT- PRESTIGE DATASET ANALYSIS

Group Members- Vaibhav Bagri, Lekha Sharma, Advait Lad, Niket Parekh, Siddhesh Pai

Linear regression is a type of supervised statistical learning approach that is useful for predicting a quantitative response Y. It can take the form of a single regression problem (where you use only a single predictor variable X) or a multiple regression (when more than one predictor is used in the model). It is one of the simplest and most straightforward approaches available and it is a starting point for more advanced modelling and predictive exercises.

The project taken up will illustrate the application of a linear regression exercise using one predictor (Simple Linear Regression) as well as two predictors.

The dataset used is called Prestige and comes from the car package library(car). The Prestige dataset is a data frame with 102 rows and 6 columns. Each row is an observation that relates to an occupation. The columns relate to predictors such as average years of education, percentage of women in the occupation, prestige of the occupation, etc. The occupations illustrated are of from the nation of Canada and hence the following analysis is in relation with the general Canadian public working class.

The six columns corresponding to a particular job are education, income,number of  women, prestige, census type. We mainly make use of Education, income, number of women and prestige columns to co-relate and deduce are theories. To clarify on the columns being used, the prestige column literally indicates the prestige and high regard of the job in the Canadian society based upon various social surveys. The higher the number is of prestige, more prestigious the job is.

Hence, we come to are one of our fundamental analysis from the data set which compares Income with Prestige and Education. The Simple Linear Regression applied helps up understand us better how income and prestige are intertwined in modern society and helps establish us an important relationship between various critical factors in the ocean of jobs and employability. We find a similar such relationship of income with education and thus try to emphasize of the importance of education.

Dataset:

Education Income Women Prestige Census Type

gov.administrators    13.11  12351 11.16      68.8 1113 prof

general.managers    12.26  25879  4.02      69.1 1130 prof

accountants              12.77    9271 15.70      63.4 1171 prof

purchasing.officers  11.42    8865  9.11      56.8 1175 prof

chemists                  14.62    8403 11.68      73.5 2111 prof

… (97 more such rows)

In our analysis, we did not need the Census and Type columns, so we made a new dataset that removed these columns using -

data = Prestige[,c(1:4)]

qplot(education, data=data, geom="histogram", binwidth=1) +

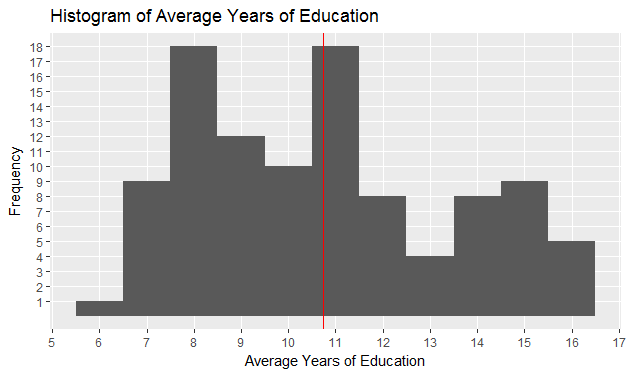
labs(title = "Histogram of Average Years of Education") +

labs(x ="Average Years of Education") +

labs(y = "Frequency") +

scale\_y\_continuous(breaks=c(1:20), minor\_breaks=NULL) +

scale\_x\_continuous(breaks=c(5:17), minor\_breaks=NULL) geom\_vline(xintercept=mean(data$education), show.legend=TRUE, color="red" )



In this histogram, we have provided some information about the Average years of Education completed by the residents of Canada. ”labs” stands for label. The x axis has been scaled with a range from 5 years to 17 years of education. The y axis, scaled from 1 to 20 provides us the frequency, that is the number of times a particular value in the x axis that has been occurred. The red vertical line in the histogram is the mean line. It helps us to get a rough idea about the average years of education completed with respect to the frequency and also helps us to divide the graph in two parts.

qplot(income, data=data, geom="histogram", binwidth=1000) +

labs(title = "Histogram of Average Income") +

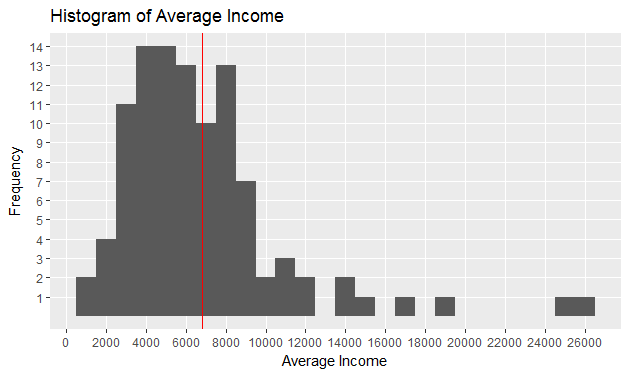
labs(x ="Average Income") +

labs(y = "Frequency") +

scale\_y\_continuous(breaks=c(1:20), minor\_breaks=NULL) +

scale\_x\_continuous(breaks=c(0, 2000, 4000, 6000, 8000, 10000, 12000, 14000, 16000, 18000, 20000, 22000, 24000, 26000), minor\_breaks=NULL) +

geom\_vline(xintercept=mean(data$income), show.legend=TRUE, color="red" )



In this histogram, we have provided some information about the Average years of Education completed by the residents of Canada. ”labs” stands for label. The x axis has been scaled with a range from 0 to 26000 Canadian dollars. The y axis, scaled from 1 to 20 provides us the frequency, that is the number of times a particular value in the x axis that has been occurred. The red vertical line in the histogram is the mean line. It helps us to get a rough idea about the average income with respect to the frequency and also helps us to divide the graph in two parts.

The education and prestige column have a wide range and are both positive. We use the following command to scale down the values and center them about 0 in order to create a better plot.

education.center = scale(data$education, center=TRUE, scale=FALSE)

prestige.center = scale(data$prestige, center=TRUE, scale=FALSE)

We have used regression analysis to find the relation between income and education. The following command creates a regression model:

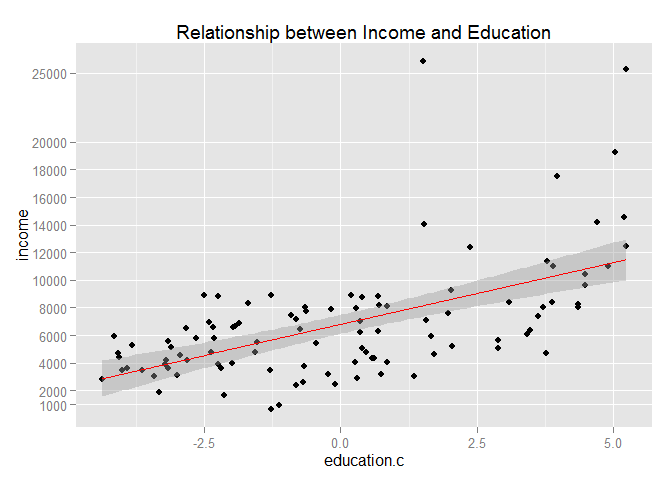
model = lm(income~education.center, data=data)

Now we plot the graph for this regression analysis.

qplot(education.center, income, data = data, main = "Relationship between Income and Education") +

stat\_smooth(method="lm", col="red") +

scale\_y\_continuous(breaks = c(1000, 2000, 4000, 6000, 8000, 10000, 12000, 14000, 16000, 18000, 20000, 25000), minor\_breaks = NULL)

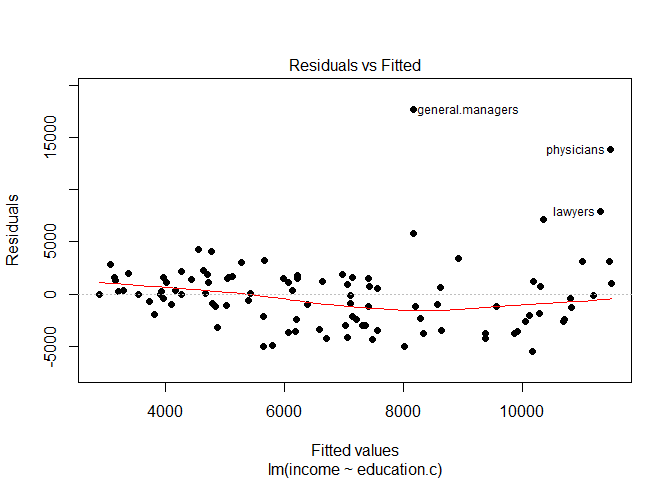


From the model output and the scatterplot we can make some interesting observations:

Visually the scatterplot indicates the relationship between income and education does not follow a straight line. While this visual inspection alone is not a sufficient indication of non-linearity, this may suggest the relationship is in fact non-linear. Observe that our fitted line does not seem to follow pattern observed across all points.

While we can see a significant p-value (very close to zero), the model generated does not yield a strong R2R2.R2R2 (or coefficient of determination) is a measure that indicates the proportional variance of income explained by education. The closer the number is to 1, the better the model explains the variance shown. In our model results, the R2R2 we get is 0.33, a pretty low score. This suggests the linear model we just fit in the data is explaining a mere 33% of the variance observed in the data.

plot(model, pch=16, which=1)



The graph above shows the model residuals (which is the average amount that the response will deviate from the true regression line) plotted against the fitted values (the model’s predicted value of income).

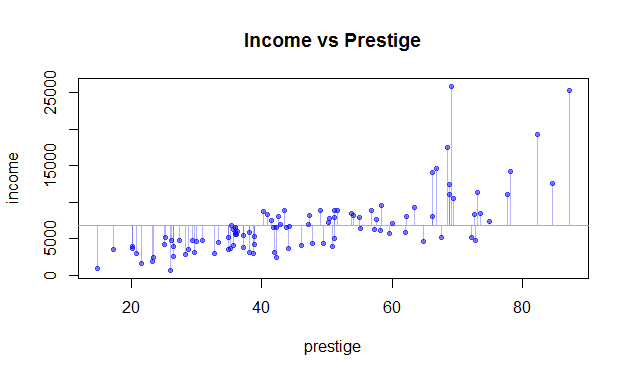
Ideally, when the model fits the data well, the residuals would be randomly scattered around the horizontal line. In our case here, there is strong evidence a non-linear pattern is present in the relationship. Also, there are points standing far away from the horizontal line. This could indicate the presence of outliers (note how the points for general managers, physicians and lawyers are way out there!).

plot(newdata, col="red", type="n",main="Income vs Prestige")

segments(x0=newdata$prestige, y0=newdata$income, x1=newdata$prestige, y1=rep(mean(newdata$income), length(newdata$income)),col=rgb(0,0,1,0.3))

abline(h=mean(newdata$income), col="orange")

points(newdata, col=rgb(0,0,1,0.5), pch=20,cex=1)



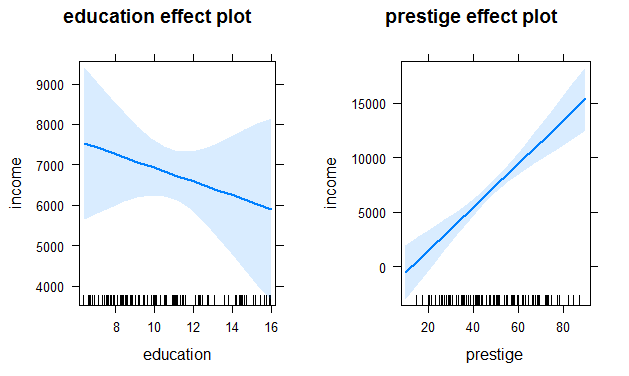
We want to check whether jobs having higher prestige have higher income or not and vice versa. Also, we want to check how much difference the highest and lowest income have from the average income of people. For both these results we use a plot function that plots a graph of income vs prestige with an orange line parallel to x axis indicating the average income. The points are plotted according to the coordinates and then there are multiple blue lines parallel to the y axis from these points to the average income line. Thus from the graph we conclude that the job with the highest income lies in the range of prestige having value(60-80) while the lowest income job lies in the range of prestige(0-20). Thus we conclude that higher prestige jobs have higher income. Also there is a huge difference between the highest and average income where as there is lesser difference between the lowest and average income.

We now find how do income and prestige together affect the income. So we use a concept of multivariable regression analysis. This is achieved by the following command, which the same as previously used, except we add the extra variable prestige:

model1 = lm(income~education+prestige, data=data)

Now we plot this regression as two separate graphs.

plot(allEffects(model1))



We have fitted a linear model using lm function of dependent variable income with independent variables eduction and prestige. Then we plot the linear model using allEffects function. From the plot we see that income and education are inversely related and income and prestige are directly related according to the data we have. From our summary of the model we see that the the slope of education is -169.63 and that of prestige is 199.30. Thus our model predicts that if we increase eduction time by one year the income decreases by 169.63 dollars and if the prestige point goes up by 1, income increases by 199.30 dollars. Thus the plot.

CONCLUSION:

Thus by this project, we can conclude that for the dataset prestige, as the prestige and education increases, the income for the particular job increases. However, this relation is not exactly linear, so the most prestigious job or the most educated person might not have the highest income. Also, we can infer that when prestige and education are taken together as the variables affecting income, education becomes inversely proportional to income. This might sound very absurd, and something which we couldn’t have thought of by ourselves. But the dataset proves it and it goes to show the importance of data analysis in our lives. Things that are seemingly absurd can be proven true with the help of proper data analysis.