# MET CS 555 Assignment 4 – 20 points

Fall 2, 2019

**SUBMISSION REQUIREMENTS: Please submit a single document (word or PDF) for submission. Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append to the end of your submission). Please use R for the calculations whenever possible. You will lose points if you are not utilizing R. You will also lose 10 points per day for late submissions unless prior arrangements are made with your facilitator.**

**The data on the next two pages is from a Canadian 1970 census which collected information about specific occupations. Data collected was used to develop a regression model to predict prestige for all occupations. Use R to calculate the quantities and generate the visual summaries requested below.**

(1) Save the data to excel and read into R for analysis.

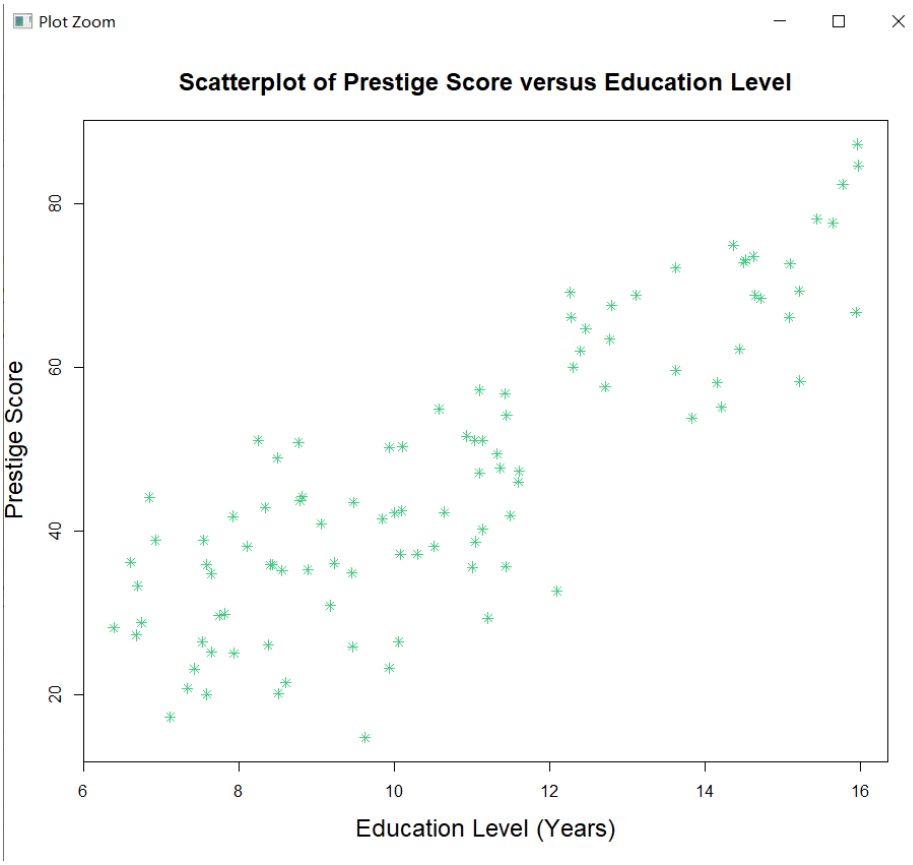
**My answer:**

See my R script below.

(2) To get a sense of the data, generate a scatterplot to examine the association between prestige score and years of education. Briefly describe the form, direction, and strength of the association between the variables. Calculate the correlation.

**My answer:**

Scatterplot of Prestige Score versus Education Level:




In the above scatterplot, I chose to place the education level on the x-axis and the prestige score on the y-axis, since in this case, the relationship between the two factors is apparent that the prestige score is a response variable and the education level is an explanatory variable. In this plot, its form is linear as the points tends toward a straight line, and its direction can be described as “positively associated” since as the education level increases, the prestige score also tends to increase in value. As regards of strength of the association, it is obvious that there is a strong positively relationship between the two variables. Moreover, we can calculate the correlation in R:

```
cor(ocpt$edu, ocpt$prestige_score)
```

The result is 0.850

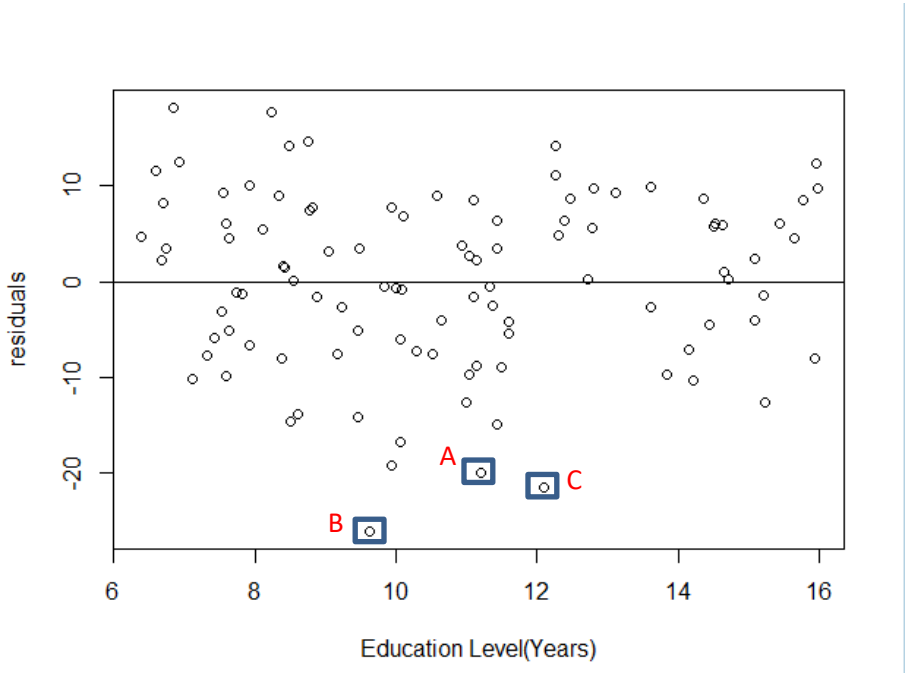
(3) Perform a simple linear regression. Generate a residual plot. Assess whether the model assumptions are met. Are there any outliers or influence points? If so, identify them by ID and comment on the effect of each on the regression.

**My answer:**

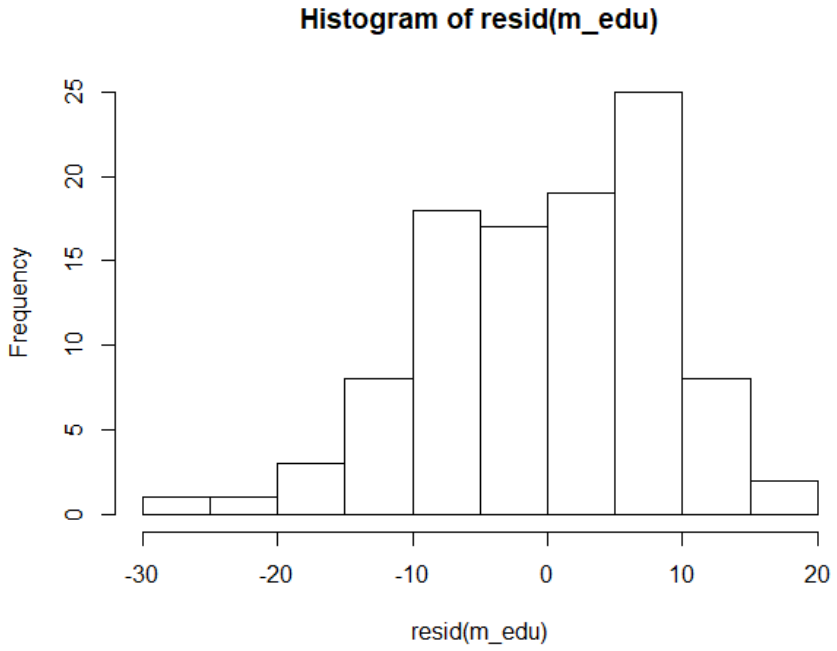
I select the education level as the explanatory variable and set a linear model, then generate a residual plot.

```
m_edu <- lm(ocpt$prestige_score~ocpt$edu)

plot(ocpt$edu,resid(m_edu), axes=TRUE, frame.plot=TRUE, xlab = "Education Level",
ylab="residuals")

abline(h=0)
```

Residual Plot:

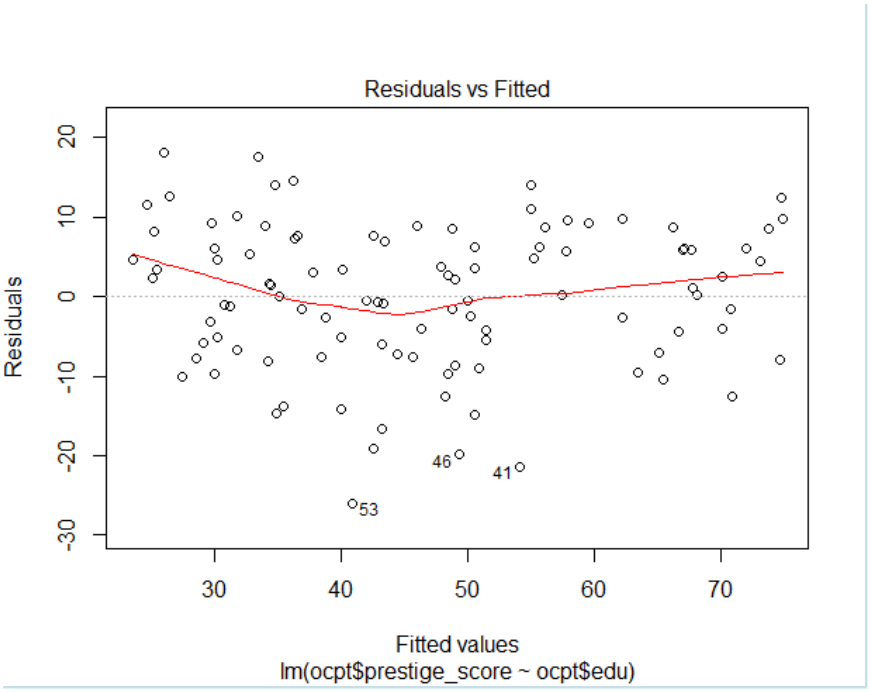



Histogram of Residual Plot (for checking the Normality):




As we seen from the residual plot, the linearity is generally met since the relationship between the factor and residual is neither curved nor non-linear. Secondly, we make the assumption that the observations are independent, since we collect the data from different occupations. Third, as for the constant variance, since the variability of the outcome in our residual plot show approximately the same amount of scatter left to right, so the constant variance is met. Lastly, in order to check the Normality, the histogram of the residuals can be used to display the distribution of the residuals, and it is obvious that the residual does follow a normal distribution and thus the normality is met.

According to the residual plot, there appears to be some outliers and I marked them as A, B and C in the residual plot. All are outliers in the y-direction, each of them is far from the regression line and has a much lower score than other people in the sample. After the removal of B, we calculate the value of R-squared and it increases from 0.72 to 0.734, which means the removal of outliers can enhance the correlation. Likewise, the removal of A also results in the increase of correlation, and they have same influential. (Using R, it shows us the Residuals vs Fitted plot below, and points out the outliers with ID = 53, 46, 41)




(4) Calculate the least squares regression equation that predicts prestige from education, income and percentage of women. Formally test whether the set of these predictors are associated with prestige at the $\alpha$= 0.05 level.

**My answer:**

m_all <- lm(ocpt$prestige_score~ocpt$edu+ocpt$income+ocpt$women_pctl)

summary(m_all)

The result:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.7943342  3.2390886  -2.098   0.0385 *
ocpt$edu        4.1866373  0.3887013  10.771  < 2e-16 ***
ocpt$income     0.0013136  0.0002778   4.729 7.58e-06 ***
ocpt$women_pctl -0.0089052  0.0304071  -0.293   0.7702
```

Now, according to the estimate value given above, we get the beta_0 = -6.794, and beta_1 = 4.187, beta_2 = 0.001, beta_3 = -0.009.

The least squares regression equation of education, income and percentage of women is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

$$\hat{y} = -6.794 + 4.187 x_{education} + 0.0013 x_{income} - 0.0089 x_{percent\ of\ women}$$

Here is the F-test.

1. Set up the hypotheses and select the alpha level
   $H_0$ : $\beta_{edu} = \beta_{income} = \beta_{women\_pctl} = 0$ (education, income, and percentage of women are not predictors of prestige score.)

   $H_1$ : $\beta_{edu} \neq 0$ or $\beta_{income} \neq 0$ or $\beta_{women\_pctl} \neq 0$ (at least one of the slope coefficients is different than 0; education and/or income and/or percentage of women are predictors/is a predictor of prestige score)

   $\alpha = 0.05$

2. Select the appropriate test-statistic

   $F = \frac{MS\ Reg}{MS\ Res}$ df = 3, n-k-1

3. State the decision rule
   Decision Rule: Reject $H_0$ if $p \leq \alpha$. Otherwise, do not reject $H_0$

4. Compute the test statistic

   summary(m_all) gives

   ```
   F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
   ```

5. Conclusion

   Reject $H_0$ since $p \leq \alpha$. We have significant evidence at the $\alpha = 0.05$ level that education, income, and percentage of women when taken together are predictive of prestige score. That is, there is evidence of a linear association between prestige score and education, income, and percentage of women (here, $p < 0.001$).

(5) If the overall model was significant, summarize the information about the contribution of each variable separately at the same significance level as used for the overall model (no need to do a formal 5-step procedure for each one, just comment on the results of the tests). Provide interpretations for any estimates that were significant. Calculate 95% confidence intervals where appropriate.

**My answer:**

According to the conclusion and F-test, the overall model was significant since the null hypothesis was rejected because at least one of the slope parameters is different from 0. The next step is doing t-test, and performing testing on each individual parameter to identify the relative contribution of each independent variable. In this case, we first calculate the t-value with n-k-1 degrees of freedom, here our df = 98, qt(0.95, df=98), we get the t-value = 1.660, and each t-value of three variables separately.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.7943342  3.2390886  -2.098   0.0385 *
ocpt$edu        4.1866373  0.3887013  10.771  < 2e-16 ***
ocpt$income     0.0013136  0.0002778   4.729 7.58e-06 ***
ocpt$women_pctl -0.0089052  0.0304071  -0.293   0.7702
```

We can summarize that the variables of education and income are significant predictor of prestige score, since the t-value of education is 10.771 and 4.729 for income, which are greater than 1.660. However, the t-value of percentage of women is -0.293 < 1.660, which means that the percentage of women is not a significant predictor of prestige score after adjusting the education level and income. Also, we can judge the p-value from the above table, and the conclusion is the same.

confint(m_all, level = .95)

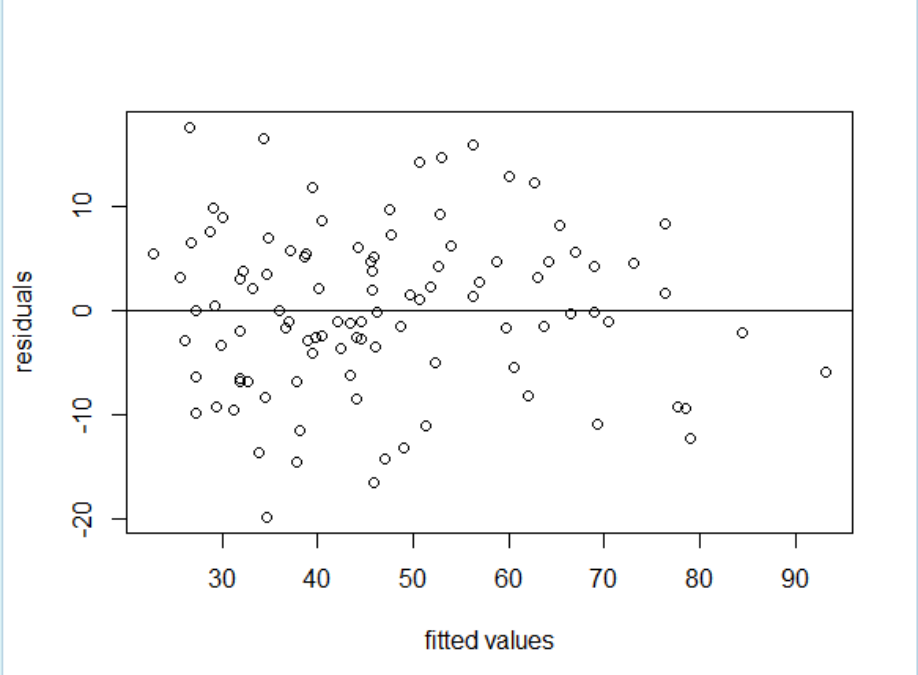
```
                          2.5 %       97.5 %
(Intercept)     -1.322220e+01 -0.366468202
ocpt$edu         3.415272e+00  4.958002277
ocpt$income      7.623127e-04  0.001864808
ocpt$women_pctl -6.924697e-02  0.051436660
```

We are 95% confident that the true value of $\beta_{edu}$ is between 3.4 and 4.9, the $\beta_{income}$ is between 0.000762 and 0.0018. That is, for every addition year of education, we are 99% confident that the prestige score is generally between 3.4 and 4.9 higher, also for every additional unit of income, the prestige score is generally slightly higher (less than 0.01)

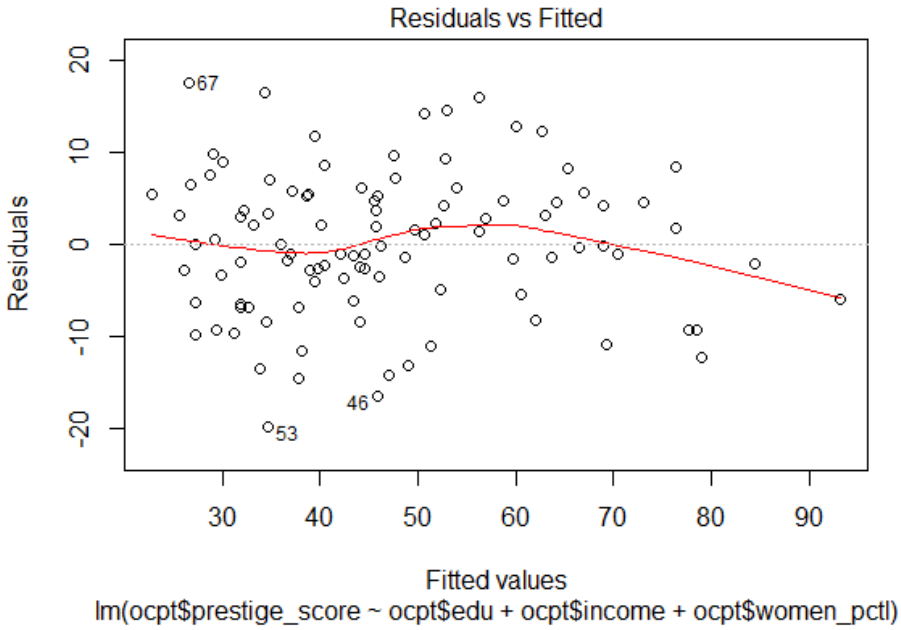
(6) Generate a residual plot showing the fitted values from the regression against the residuals. Is the fit of the model reasonable? Are there any outliers or influence points?

**My answer:**

The residual plot of the fitted value from the regression against the residuals:




According to the residual plot of the fitted value from the regression against the residuals, I don't think the fit of the model is reasonable, since it violates the rule of constant variance, that is, the variability decreases as the fitted value increase (especially when fitted values greater than 60). Also, the residual plot above violates the rule of linearity, because it looks like somewhat curved. There appears to be some outliers or influence points, which their id = 46, 53 and 67 seen from the plot below.



R script:

```
setwd("C:\\Users\\Lin\\Desktop\\2019 Fall\\555 DAV_ R\\Homework\\hw4")

# 1.read the csv file
ocpt = read.csv("occupation.csv")
ocpt

summary(ocpt)

# 2.draw scatterplot and examine the association

plot(ocpt$edu, ocpt$prestige_score,
    main="Scatterplot of Prestige Score versus Education Level",
    ylab = "Prestige Score", xlab="Education Level (Years)",
    pch = 8, col="seagreen3",
    cex=1, cex.lab = 1.5, cex.main = 1.5)
abline(m_edu,lty=3,col="black")
# calculate the correlation
cor(ocpt$edu, ocpt$prestige_score)
cor(ocpt$income, ocpt$prestige_score)
cor(ocpt$women_pctl, ocpt$prestige_score)

# 3. perform a simple linear regression
m_edu <- lm(ocpt$prestige_score~ocpt$edu)
summary(m_edu)
```

```
plot(ocpt$edu,resid(m_edu), axes=TRUE, frame.plot=TRUE, xlab = "Education Level(Years)",
ylab="residuals")

abline(h=0)


hist(resid(m_edu))


# test R^2 the removal of influence point

ocpt_copy = ocpt

ocpt <- ocpt[-c(53),]

ocpt <- ocpt[-c(67),]


# 4. multiple linear regression

m_all <- lm(ocpt$prestige_score~ocpt$edu+ocpt$income+ocpt$women_pctl)

summary(m_all)

confint(m_all, level = .95)


# 6. residual plot

plot(fitted(m_all),resid(m_all), axes=TRUE, frame.plot=TRUE, xlab = "fitted values",
ylab="residuals")

abline(h=0)

plot(m_all)
```