A study on occupation prestige rating in Canada with multiple regression model – country level study

Ziyue Zhong

4/11/2018

Introduction

Pineo-Porter prestige score is collected by a national survey study in Canada throw phone interview in 2005. The national occupation classification (NOC) is a structure that sociologically divide working force into different major groups last updated in 2001. The survey required every participant to give prestige rating for the 26 major groups in the NOC. Based on the survey result, every occupation in the Canadian labor force is give a prestige score. The score is a reliable measure for the prestige rating because the scale is large and scientific.

In order to understand how people's perception on job are influenced, a multiple regression model is formed to evaluate the correlation between prestige rating of an occupation and other related factors including education, income and type.

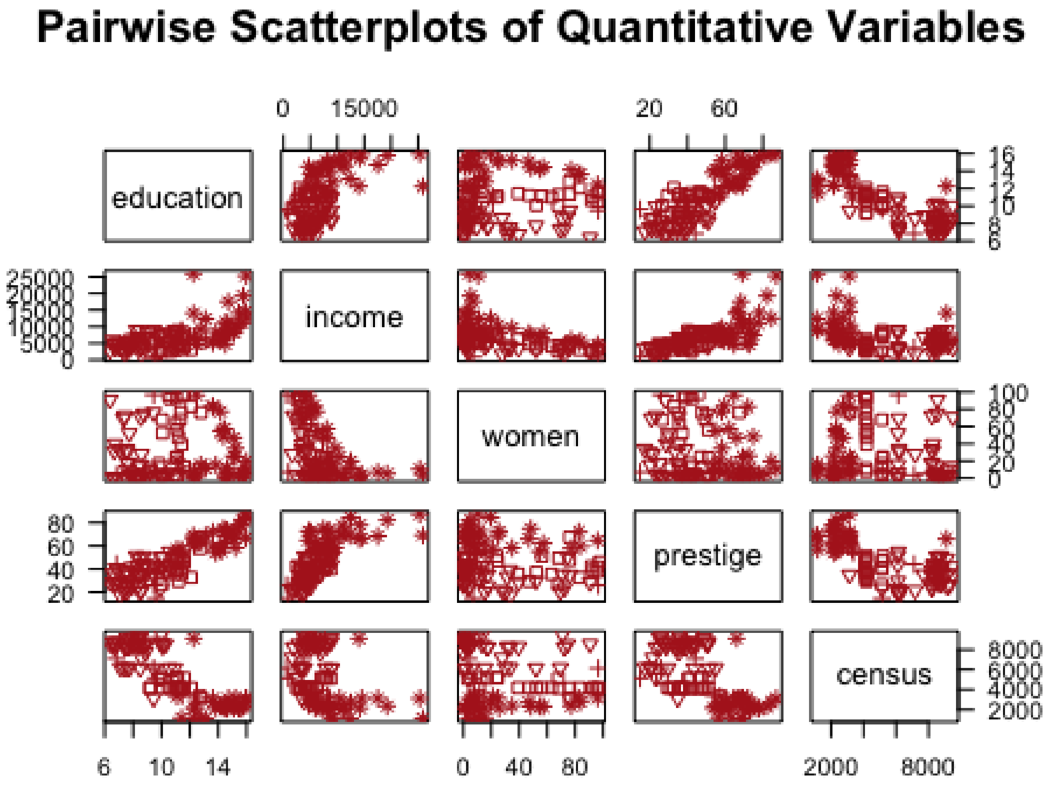
Data Description

Variable Name	Data Type	Description
Prestige score	Numeric	Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s
Education	Numeric	average education of occupational incumbents, years, in 1971
Income	Numeric	average income of incumbents, dollars, in 1971
Women	Numeric	percentage of incumbents who are women
Census	Char	Canadian Census occupational code
Type	Categorical	type of occupation: "bc"=blue collar, "prof"= professional/managerial/technical, "wc"=white collar

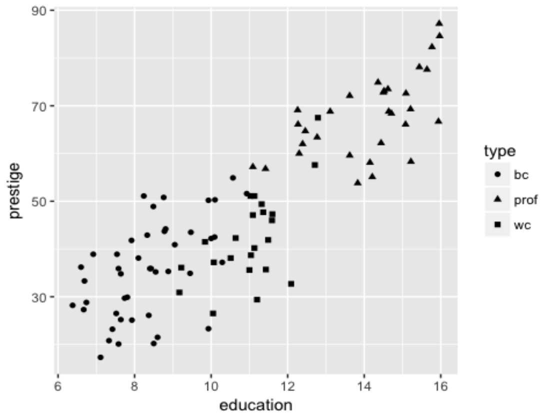
Variable Selection

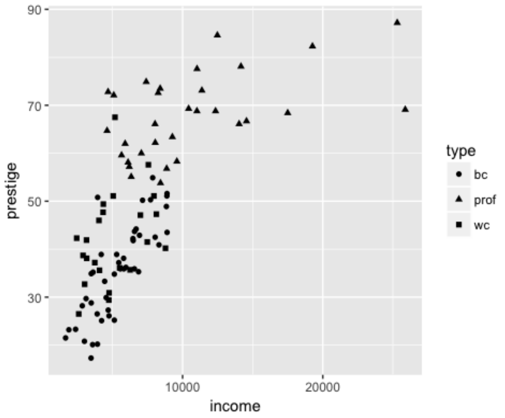
From the scatterplot of all the quantitative variables, we can see that there is a strong linear relationship between prestige & education and prestige & income. Also, it shows that women has weak linear relationship with prestige meaning a low correlation. Also, we can notice that for each occupation type, they have different distribution pattern in the scatterplot for prestige which indicates a strong correlation.

It is reasonable to restrict our analysis to only education, income and type, because these three variables all have strong correlation with the prestige variable. For the other variables women and census, women has low correlation and census is not related to the occupation itself. Therefore, we drop them from the model.


Pairwise Scatterplots of Quantitative Variables

We draw scatter plots for prestige & income and prestige & education using different point shape to represent different occupation type. From the plot, we can notice that each occupation have different distribution on the plot, ‘bc’ and ‘wc’ have more similar distribution compared to ‘prof’. Therefore we believe that type have interation with education and income, especially education.





Model Building

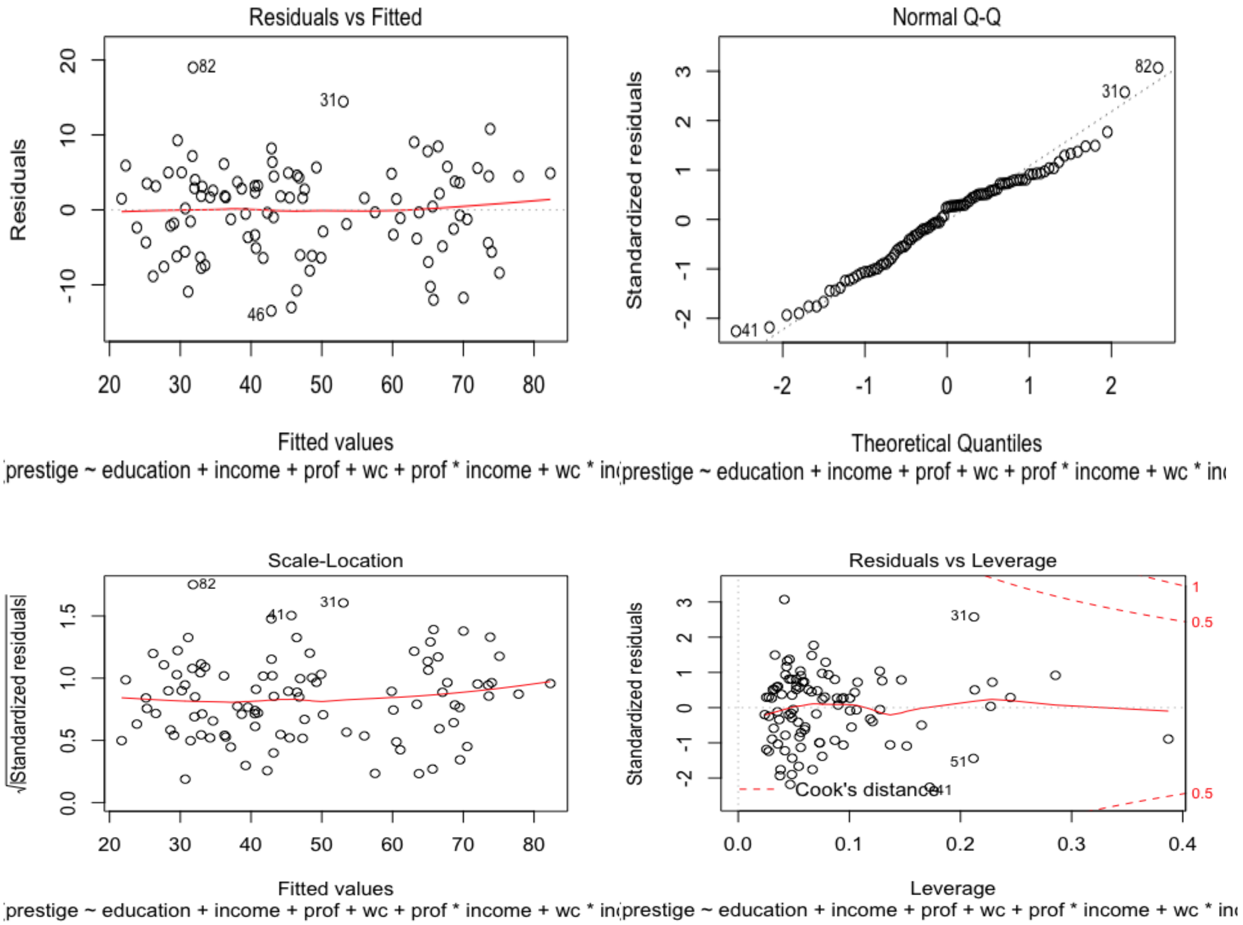
1. First include all three variables education, income, type and all interaction variables into the model lm.1.
2. Use partial F-test to test the null hypothesis. Since p-value is smaller than 0.05, we will reject the null hypothesis and keep the interaction variables.

```
## Analysis of Variance Table
##
## Model 1: prestige ~ education + income + prof + wc + prof * income + wc *
##     income + prof * education + wc * education
## Model 2: prestige ~ education + income + prof + wc
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     89 3552.9
## 2     93 4681.3 -4   -1128.4 7.0668 5.479e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = prestige ~ education + income + prof + wc + prof *
##     income + wc * income + prof * education + wc * education,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.462  -4.225   1.346   3.826  19.631
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.276e+00  7.057e+00   0.323   0.7478
## education        1.713e+00  9.572e-01   1.790   0.0769 .
## income           3.522e-03  5.563e-04   6.332 9.62e-09 ***
## prof             1.535e+01  1.372e+01   1.119   0.2660
## wc              -3.354e+01  1.765e+01  -1.900   0.0607 .
## income:prof    -2.903e-03  5.989e-04  -4.847 5.28e-06 ***
## income:wc      -2.072e-03  8.940e-04  -2.318   0.0228 *  
## education:prof  1.388e+00  1.289e+00   1.077   0.2844    
## education:wc    4.291e+00  1.757e+00   2.442   0.0166 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.318 on 89 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8634 
## F-statistic: 77.64 on 8 and 89 DF,  p-value: < 2.2e-16
```

3. Using t-test to drop insignificant variables from lm.1 until all the variables are significant and get model lm.2. From the summary table, we found that only education*prof is not significant. So we drop it and keep others in the model.
4. Use t-test to check all the variables in model lm.2 are significant. Since all variables are significant, we keep them all in the model.

```
## 
## Call:
## lm(formula = prestige ~ education + income + prof + wc + prof * 
##     income + wc * income + wc * education, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -13.462  -4.396   1.471   4.231  19.003 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -2.922e+00  5.153e+00  -0.567 0.572046    
## education     2.479e+00  6.414e-01   3.865 0.000209 ***
## income        3.298e-03  5.164e-04   6.387 7.27e-09 ***
## prof          2.883e+01  5.611e+00   5.139 1.59e-06 ***
## wc           -2.834e+01  1.700e+01  -1.667 0.098919 .  
## income:prof  -2.633e-03  5.447e-04  -4.835 5.47e-06 ***
## income:wc    -1.848e-03  8.702e-04  -2.124 0.036455 *  
## education:wc  3.525e+00  1.608e+00   2.192 0.030983 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.324 on 90 degrees of freedom
## Multiple R-squared:  0.873,  Adjusted R-squared:  0.8632 
## F-statistic: 88.41 on 7 and 90 DF,  p-value: < 2.2e-16
```



Model (e) Interpretation

The regression model: lm.2 <- lm(prestige ~ education + income + prof + wc + prof*income + wc*income + wc*education, data=data)

If occupation is professional: prestige = 2.479 * edu + 0.000665 * income +25.908

If occupation is white collar: prestige = 6.004 * education + 0.00145 * income - 5.756

If occupation is blue collar: prestige = 2.479 * education + 0.003298 * income - 2.922

```
## [1] "AVG VIF is:"

## [1] 1.491621
```

Check regression assumption:

1. x variables are fixed and measured without error The x variables are fixed when survey is conducted but it may have some errors compared to the real value because this result is collected through a survey so they are not perfectly measured.

2. constant variance According to the ‘squared std.residual & fitted’ value plot, the red line is flat when you go from left to right, so this assumption is satisfied.

3. nonlinearity According to the ‘std. residual & fitted’ value plot, there is no obvious trend line, so this assumption is satisfied.

4. normality The theoretical Quantiles shows a S-shape, therefore this assumption is not satisfied.

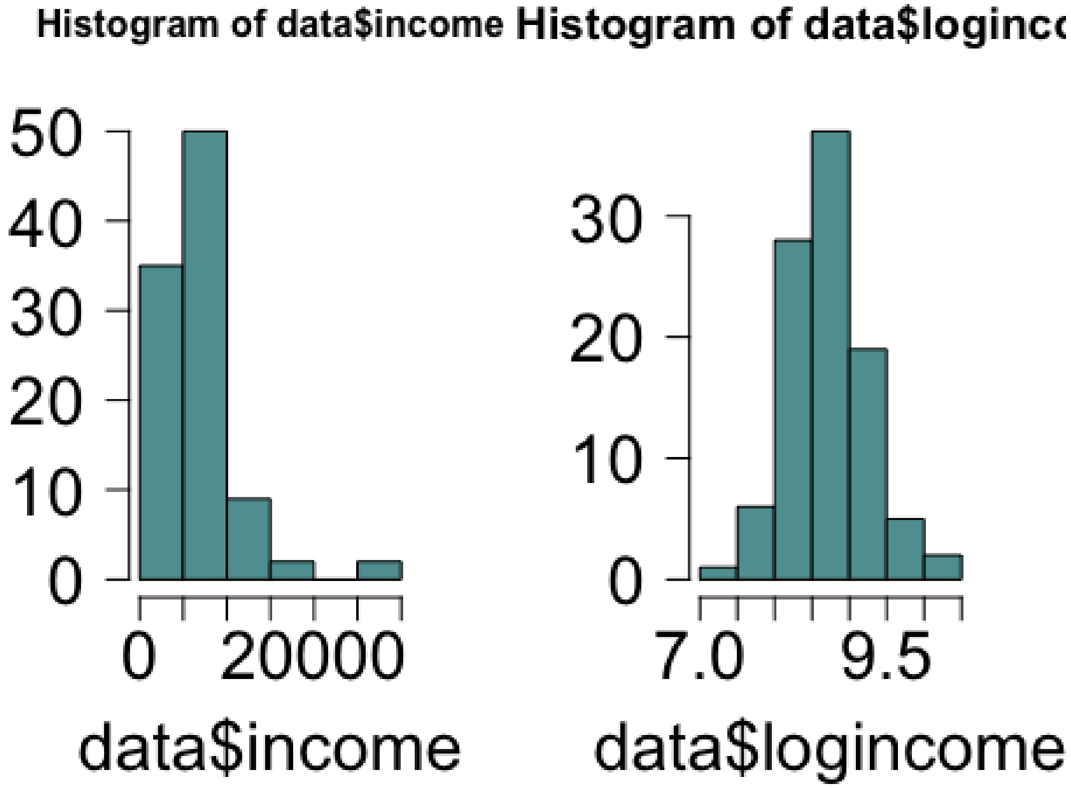
5. Independence of x We assume that the participants of the survey are randomly selected so each instance should be independent.

6. No collinearity/multicollinearity Because we uses that interaction variable, so we don’t have to include type into analysis. Although the correlation between income and education is 0.574 which slightly high, the VIF is low, about 1.5, therefor this consumption is also satisfied.

7. Residuals vs leverage plot shows that there is no strong outlier in this regression model. (Cook’s distance < 0.5)

Model Optimization (Variable transformaiton)

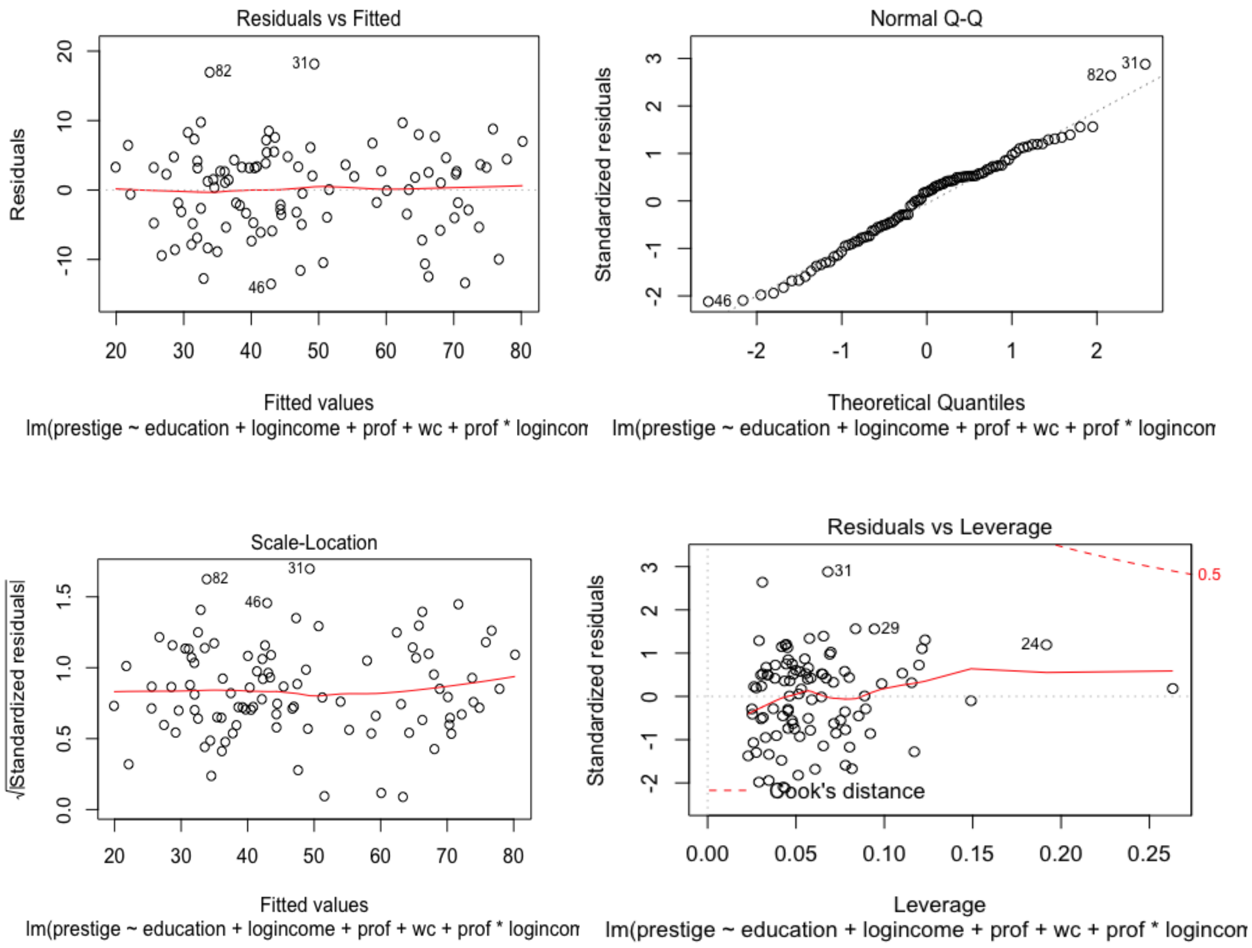
Before transformation, income is left skewed. After transformation, log.income is closer to normal distribution.



1. Apply all the variables into the regression model but use log.income instead of income this time and we get model lm.3.
2. Using T-test to drop variables that are insignificant one by one.
3. Check the significance of all variables in final model lm.4.

```
##
## Call:
## lm(formula = prestige ~ education + logincome + prof + wc + prof *
##     logincome + wc * logincome + +prof * education + wc * education,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.970  -4.124   1.206   3.829  18.059
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -120.0459    20.1576  -5.955 5.07e-08 ***
## education         2.3357     0.9277   2.518  0.01360 *
## logincome        15.9825     2.6059   6.133 2.32e-08 ***
## prof             85.1601    31.1810   2.731  0.00761 **
## wc               30.2412    37.9788   0.796  0.42800
## logincome:prof   -9.4288     3.7751  -2.498  0.01434 *
## logincome:wc     -8.1556     4.4029  -1.852  0.06730 .
## education:prof    0.6974     1.2895   0.541  0.58998
## education:wc      3.6400     1.7589   2.069  0.04140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.409 on 89 degrees of freedom
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.8595
## F-statistic: 75.15 on 8 and 89 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = prestige ~ education + logincome + prof + wc + prof *
##     logincome, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.535  -4.540   1.165   3.802  18.151
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -102.1269    17.0032  -6.006 3.75e-08 ***
## education         3.3143     0.5983   5.539 2.87e-07 ***
## logincome        12.9164     2.0700   6.240 1.32e-08 ***
## prof             66.0206    29.4281   2.243   0.0273 *
## wc               -1.3823     2.3393  -0.591   0.5560
## logincome:prof   -6.6620     3.2835  -2.029   0.0454 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.528 on 92 degrees of freedom
## Multiple R-squared:  0.8617, Adjusted R-squared:  0.8542
## F-statistic: 114.6 on 5 and 92 DF,  p-value: < 2.2e-16
```



Model (e) Interpretation

The Regression Model: lm(formula = prestige ~ education + logincome + prof + wc + prof * logincome, data = data)

If occupation is professional/managerial/technical: prestige = 3.3143*education + 6.2544*logincome - 36.1063

If occupation is white collar: prestige = 3.3143 * education + 12.9146 * log(income) - 103.6513

If occupation is blue collar: prestige = 3.3143 * education + 12.9146 * log(income) - 102.1269

```
##           education income  women prestige census
## education     1.000  0.574  0.062    0.866 -0.826
## income        0.574  1.000 -0.448    0.703 -0.362
## women         0.062 -0.448  1.000   -0.110 -0.242
## prestige      0.866  0.703 -0.110    1.000 -0.649
## census       -0.826 -0.362 -0.242   -0.649  1.000

## [1] "AVG VIF is:"

## [1] 1.491621
```

Check regression assumption:

1. x variables are fixed and measured without error The x variables are fixed when survey is conducted but it may have some errors compared to the real value because this result is collected through a survey so they are not perfectly measured.

2. constant variance According to the ‘squared std.residual & fitted’ value plot, the red line is flat when you go from left to right, so this assumption is satisfied.

3. nonlinearity According to the ‘std. residual & fitted’ value plot, there is no obvious trend line, so this assumption is satisfied.

4. normality The theoretical Quantiles is almost a straight line, therefore the residual are normally distributed and this assumption is satisfied.

5. Independence of x We assume that the participants of the survey are randomly selected so each instance should be independent.

6. No collinearity/multicollinearity Because we uses that interaction variable, so we don’t have to include type into analysis. Although the correlation between income and education is slightly high, the VIF is low, about 1.5, therefor this consumption is also satisfied.

7. Residuals vs leverage plot shows that there is no strong outlier in this regression model. (Cook’s distance < 0.5)

Model Comparison & Conclusion

R squared for model(e): 0.8632 R squared for model(g): 0.8542

Although model (e) has a slightly higher than model(g), model(g)’s residual is more normally distributed therefore have less overfitting problems. Meanwhile, model(g) is simpler than model(e) with less variables. Therefore, model in (g) is better.