Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Halios Georgios 2. Haw, Kar Whing 3. Sun, Xuehui	4. Turki, Yasmine 5. Yi, Jinze (Eric) 6. Yuan, Ziyun (Isabella) **GROUP NUMBER:** **3**
MSc in: Business Analytics	
Module Code: SMM 637	
Module Title: Quantitative Methods	
Lecturer: Radice, Rosalba	**Submission Date:** 2/11/2020

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: 66 %

$16\frac{1}{2}/25$

Table of Contents

Q1: JUSTIFICATION OF THE CHOSEN REGRESSION MODEL SPECIFICATION.... 2

DATA MODIFICATION 2

MULTICOLLINEARITY..... 2

ANOVA TEST 2

RESIDUAL ANALYSIS 3

Q2: SUMMARY AND INTERPRETATION..... 4

Q3: RECOMMENDATION AND LIMITATION 6

LIMITATIONS..... 6

RECOMMENDATIONS..... 6

Q4: REFLECTION AND SUGGESTION..... 7

APPENDIX 8

Tables..... *8*

Figures *9*

Q1: Justification of the chosen regression model specification

In this report, we are examining the effects of various variables on wage in the US. Table 1 explains our data set.

Data Modification

Before fitting a regression model, modification is made on two of the variables: wage and experience. In the original data set, experience is computed by age - education - 6, therefore, some of the data are recorded as negative values. Since, an individual having a negative experience essentially means the person has no experience, so we decided to replace zero with the negative values for experience. Furthermore, we are using the log of wage, instead of wage, as the independent variable because this improves the model's adjusted R^2. The fitted model (model.1) can be shown as below.

Equation 1 — model.1

$$\log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exp_i + \beta_3 Eth_i + \beta_4 SMSA_i + \beta_5 Reg_i + \beta_6 PT_i + u_i$$

Multicollinearity

The VIF values of all variables are less than 5 (table 2), so the presence of multicollinearity is not detected in the model.

ANOVA Test

The ANOVA test is employed to test the null hypothesis that each variable's coefficient is insignificant. We performed ANOVA test on model.1, and the p-value of the quantitative variables: education and experience, are significant at the 5% level (table 3). In addition, F-test in ANOVA is used to determine the significance of the categorical variables. Similarly, the p-values obtained from the F-test (table 3) suggest the rejection of the null hypothesis for all the variables. Hence, all the independent variables in model.1 are significant.

Adding Interaction Terms and Performing ANOVA Test on the New Fitted Model

After carefully analysing the possible relationship between qualitative variables, we chose three interaction terms and added into our model：

a. Education * SMSA: The SMSA variable refers to the residence in SMSA, an area requiring relatively higher living expenses but offering better education facilities including schools and extracurricular activities. People living in SMSA can access to better education and gain advantage in business so we think the education variable may apply various effect on the dependent variable according to the value of SMSA
b. Experience * SMSA: Besides better education facilities, SMSA also offers a better business environment. People living in SMSA can access to professional opportunities with better salary and welfare so we think the experience variable may apply various effect on the dependent variable according to the value of SMSA
c. Experience * Parttime: The parttime variable refers whether the individual works part-time. In the real world, HR considers different weights for full-time and part-time experiences of candidates so we think the experience variable may apply various effect on the dependent variable according to the value of parttime

The new fitted model (model.it) is shown as below.

Equation 2 — model.it

$$x_i = \beta_0 + \beta_1 Ed_i + \beta_2 Exp_i + \beta_3 Eth_i + \beta_4 SMSA_i + \beta_5 Reg_i + \beta_6 PT_i$$

$$\log(wage_t) = x_i + \gamma_1(Ed_i * SMSA_i) + \gamma_2(Exp_i * SMSA_i) + \gamma_3(Exp_i * PT_i) + u_i$$

According to the ANOVA test, the p-value of all the variables in model.it are significant at the 5% level (table 4).

Residual Analysis

A new model (model.res_analy) consisting only of quantitative variables: dependent variable (log of wage) and independent variables (experience and education), is constructed for residual analysis.

Equation 3 — model.res_analy

$$\log(wage_t) = \beta_0 + \beta_1 Ed_i + \beta_2 Exp_i + u_i$$

From the residuals vs fitted plot, the residuals are scattered around the zero line, implying the assumption of linearity in the model. Also, there is an indication of small increasing variance, in absolute terms, with mean fitted value in Residual vs Fitted plot and the residuals are not fully evenly spread around the vertical line in the Scale-Location plot, so the assumption of constant variance is questionable. Additionally, the assumption of normality does not hold because some of the residuals deviate from the straight dotted line in the QQ plot. Hence, normality of the plot is questionable as we have limited data points to be conclusive. Lastly, all the residuals are within the 0.5 cook's distance line in the Residuals vs Leverage plot. Even though there are few high leverage points but their actual influence on the fit is not unduly large.

Q2: Summary and Interpretation

The assumptions we are making for equation 2 (model.it) are that the errors are normally distributed, have a zero mean, a constant variance and are independent. Moreover, the following interpretations of our estimates for each variable hold if and only if, all other variables are held at some fixed values.

Table 5 — Summary of the final model & interpretations

Estimates	P-values	Transformation of β_i and γ_i	Interpretation
$\beta_0 = 4.6165522$	< 2e-16 ***	$\exp(\beta_0) \approx 101.14470$	The average wage is 101 USD per week when all the variables are set to 0.
$\beta_1 = 0.0791728$	< 2e-16 ***	$\exp(\beta_1) \approx 1.08239$	For a one-year increase in education, we expect to see about a 8% increase in wage.
$\beta_2 = 0.0163285$	< 2e-16 ***	$\exp(\beta_2) \approx 1.01646$	For a one-year increase in experience, we expect to see about a 2% increase in wage.
$\beta_3 = 0.2169753$	< 2e-16 ***	$\exp(\beta_3) \approx 1.24231$	The wage will be 24% higher for the caucasian workers than for the african american workers.
$\beta_4 = -0.1153485$	0.00559 **	$\exp(\beta_4) \approx 0.89105$ $\exp(\beta_4) - 1 \approx -0.108944$	The wage will be 11% lower for the workers residing in a SMSA than for the workers who don't.

$\beta_3 Reg_{northeast}$ =0.3828 $\beta_3 Reg_{south}$ = -0.05125 $\beta_3 Reg_{west}$ = 0.02073	6.33e-05 *** 9.24e-09 *** 0.03174 *	$\exp(\beta_3 Reg_{northeast}) \approx 1.039$ $\exp(\beta_3 Reg_{south}) \approx 0.95$ $\exp(\beta_3 Reg_{west}) \approx 1.021$	Having as categorical base region midwest we can say that, when i = northeast, wage increases by 3.9%, when i = south, wage decrease by 5%, and when i = west, wage increase by 2.1%
β_0 = -0.9906870	< 2e-16 ***	$\exp(\beta_0) \approx 0.37132$ $\exp(\beta_0)$ -1 ≈ -0.628678	The wage will be 63% lower for a worker working part-time compared to a worker who doesn't.
γ_1 = 0.0178532	3.12e-10 ***	$\exp(\gamma_1) \approx 1.01801$	Interaction indicates that residents in standard metropolitan statistical areas have 1.8% more wage compared to the ones which are not. The effect of being in this category increases when the years of education increase.
γ_2 = 0.0023288	9.38e-05 ***	$\exp(\gamma_2) \approx 1.00233$	Interaction indicates that residents in standard metropolitan statistical areas have 0.23% more wage compared to the ones which do not. The effect of being in this category increases when the years of experience increase.
γ_3 = -0.0053842	8.18e-16 ***	$\exp(\gamma_3) \approx 0.99463$ $\exp(\gamma_3)$ - 1 ≈ -0.00536	Interaction indicates that individuals that work part time have 0.5% less wage compared to the ones which are not part time. The effect of being in this category decreases when the years of experience increase.
Residual standard error: 0.5474 on 28143 degrees of freedom , Multiple R-squared: 0.4156, Adjusted R-squared: 0.4153 , F-statistic: 1819 on 11 and 28143 DF, p-value: < 2.2e-16			

The P-values for our estimates are all significant at the 5% significance level. The adjusted R-squared is a statistical measure of how close the data are to the fitted regression line. In our model, the adjusted $R^2 = 0.4153$ which means that our model explains the 42% of the variability of our wage variable. As seen from the first pair of graphs in figure 2, residents in standard metropolitan statistical areas after years of education tend to have higher wages, the same applies to the second pair on experience. Last pair shows that people in full time tend to get more money than those in part time.

By searching for insights from the graphs it is noticed that we can question interactions since plots within pairs seem to have parallel trends and shapes[1].

[1] Suresh HP " The significance of Interaction Plots in Statistics". Medium.[Online] [Viewed on November 2nd, ,2020] Available on : https://medium.com/@hpsuresh12345/the-significance-of-interaction-plots-in-statistics-6f2d3a6f77a3

Q3: Recommendation and Limitation

Limitations

1. Assumption of normality does not fully hold in this case according to QQ plot
2. The adjusted R^2 is 0.4153. The fitted model only describes about 41.53% of the data
3. We set all the negative numbers in 'experience' to be 0, which might affect the regression result. Experience was calculated from a "guessing equation".
4. We did not test all the interaction factors. We chose 7 pairs of interaction factors that we think might make sense in the real world.
5. For the variable 'region', we only given the data of four parts: northeast, midwest, south, and west, while in real life we also need to consider the effect on other parts such as northern, eastern and others.
6. The nature of linear regression is exploring only linear relationships between dependent and independent variables.

3.5/5

Recommendations

1. Search for a distribution, other than normal distribution, that can better fit our model
2. We could include more variables that affect wages such as the cost of living, the bargain of trade unions, the productivity, the cost of training or the sector[2]. This would increase the R^2 and the adjusted R^2.
3. We could include a triple way interaction.

Elaborate more

Interpretability?

[2] Smriti Chand "Top 8 Factors Influencing the Determination of Wage Rates". Your actual library. [Online] [Viewed on November 1rst,2020] Available on : https://www.yourarticlelibrary.com/employee-management/wages/top-8-factors-influencing-the-determination-of-wage-rates/34666

Q4: Reflection and Suggestion

From our analysis, we learnt that many factors enter in the determination of wages and many of these factors can be correlated. At first, we were worried about the value of the R-squared since our model only explains 41.53% of the variation of our wage variable. After a small research we found that analyses which attempt to predict human behavior usually have R-squared values lower than 50, where physical process analysis might have values somewhere close to 90% (if good measurements are taken)[3].

3/5

We cannot expect when doing an analysis that every assumption will be strictly held. There are cases that need to compromise or search for a better alternative approach/method.

Analysis could be improved by maybe adding polynomial terms. Moreover, the analysis could be improved by increasing the sample. The sample is composed of 28,155 observations and this cannot fully represent the whole wage situation in the US, given that it is a country with a population higher than 300 million people.

on which covariate?

How?

Overall good use of visualization tools

4/5

[3] Jim Frost, 2018. "How High Does R-squared Need to Be?". Statistics by Jim. [Online] [Viewed on November 1rst,2020] Available on : https://statisticsbyjim.com/regression/how-high-r-squared/

Appendix

Tables

Table 1: Dataset Variables and Description

Variable	Description
Wage	Wage in US dollars per week.
Ed (education)	Years of education.
Exp (experience)	Years of working experience.
Eth (ethnicity)	A factor with levels Caucasion (cauc) & African-American (afam).
SMSA	Residence in a standard metropolitan statistical area.
Reg (region)	The region within the United States of America.
PT (Parttime)	Whether the individual works part-time.

Table 2: VIF Values of the Variables in Model.1

Variables	GVIF	DF	GVIF ^ (1 / (2*Df))
Ed	1.108266	1	1.052742
Exp	1.102207	1	1.049860
Eth	1.040533	1	1.020065
SMSA	1.029443	1	1.014615
Reg	1.055363	3	1.009021
PT	1.011249	1	1.005609

Table 3: ANOVA Test

Variables	P-value from Anova Test
Ed	2.20E-16
Exp	2.20E-16

Models	Fitted Model is model.1 Excluding Variable:	P-value from ANOVA F-Test Against model.1
model.2	Education	2.20E-16
model.3	Ethnicity	2.20E-16
model.4	SMSA	1.73E-10
model.5	Region	1.73E-10
model.6	Part-time	2.20E-16

Table 4: ANOVA Test on the fitted model with interaction terms (model.it)

Variables	P-value from Anova Test
Ed	2.20E-16
Exp	2.20E-16
Eth	2.20E-16
SMSA	2.20E-16
Reg	2.20E-16
PT	2.20E-16
Ed:SMSA	8.52E-08
Exp:SMSA	9.12E-05
Exp:PT	8.18E-16

Figures

Figure 1: Residual Analysis Plot on model.it

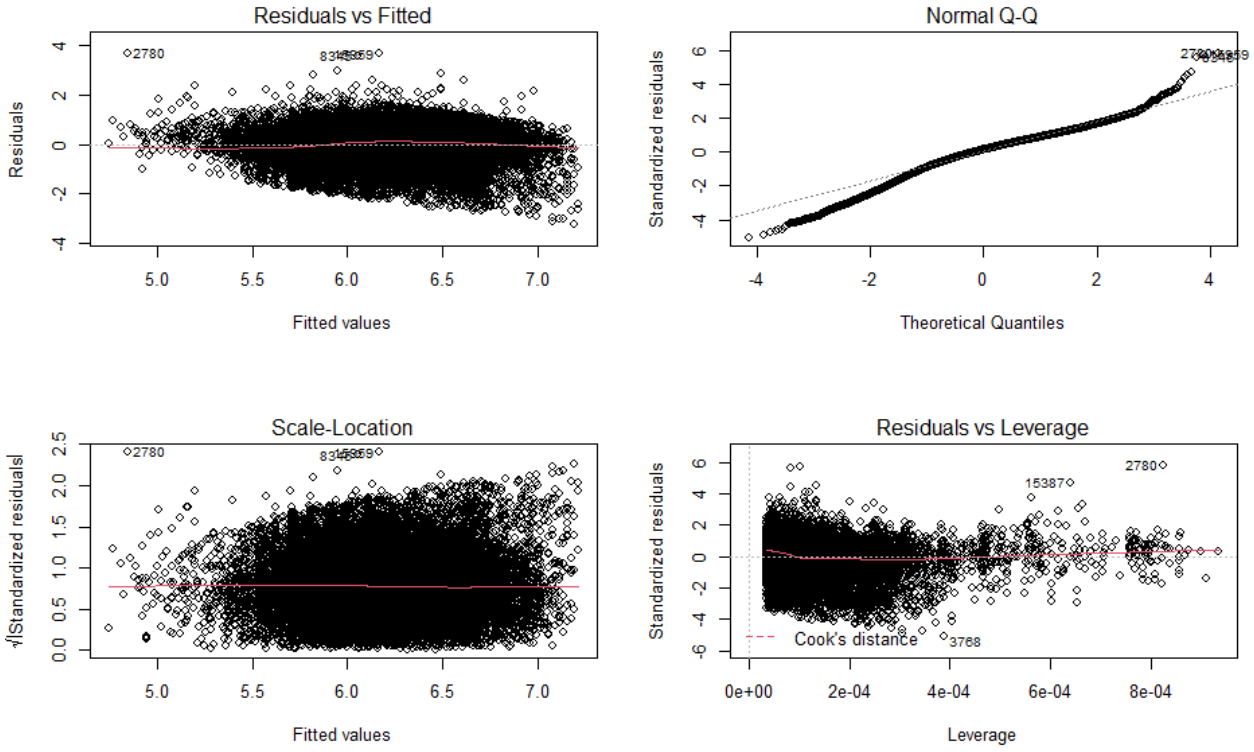

Figure 2: Visual representation of the interaction terms