Math (Final Project)

Data Set 2

Wong Choiyanheki(2004 0850)

Summary

This project is analysis the data set of 534 observations in order to identify the significant indicating factor for the wage level. The objective is to propose a new conceptual multiple regression model based on the regression results. Educational level and working experience is found to be the significant predictors for wage level. It is foreseeable that the further result of the new model can help the employer set the wage level.

Background Information

This dataset contains 524 observations on 11 variables sampled from the current population survey. The variables include years of education, southern region, sex, working experience, union membership, wage, age, race, occupation, sector and also the marital status. In these variables, only education, working experience, age and wage are continuous variable and others are all categorical variables.

Objective

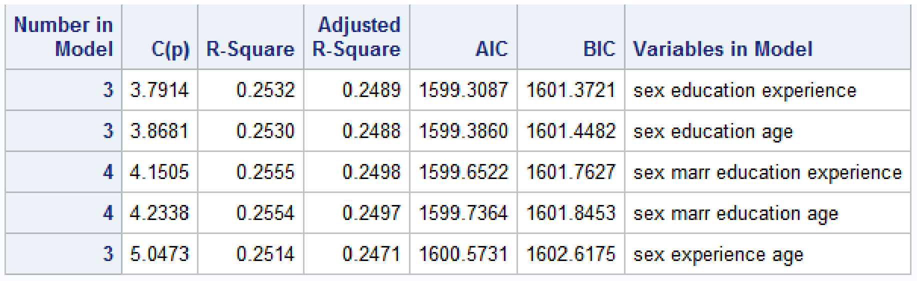
The objective of this study is to examine the determinants of wage (dollars per hour) in Hong Kong by using multiple regressions. We will do regression analysis in order to determine which model is the best model to identify the factors that determine the wage. To this end, a new conceptual multiple regression model will be proposed based on the analysis results.

Methodoloy and Results

Model selection

First of all, we carry out model selection to find out which model is the best model to identify the factors that determine the wage. The model with the smallest AIC represents the best approximation to the true model. Adjusted R-square value derived to reflect the extent of how the independent variables contribute to the variation in the dependent variable. We can see that the model that includes sex, education and experience, is the best model, as this model has a high adjusted R^2(0.2489), a lowest Mallows' Cp value(3.7914), and the lowest AIC value(1599.3087).

Table1: Model summary



Number in Model	C(p)	R-Square	Adjusted R-Square	AIC	BIC	Variables in Model
3	3.7914	0.2532	0.2489	1599.3087	1601.3721	sex education experience
3	3.8681	0.2530	0.2488	1599.3860	1601.4482	sex education age
4	4.1505	0.2555	0.2498	1599.6522	1601.7627	sex marr education experience
4	4.2338	0.2554	0.2497	1599.7364	1601.8453	sex marr education age
3	5.0473	0.2514	0.2471	1600.5731	1602.6175	sex experience age

Test the interaction effect

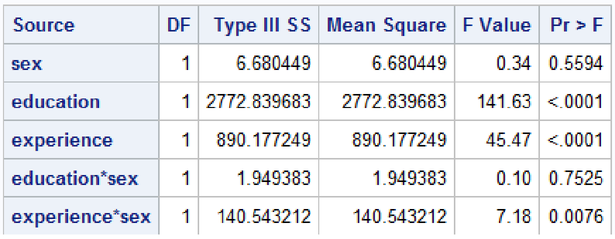
From the previous result, the best model is:

$$\hat{y}_{wage} = \hat{\beta}_0 + \hat{\alpha}_1 x_{sex} + \hat{\beta}_2 x_{experience} + \hat{\beta}_3 x_{education} + \hat{\alpha}_4 x_{sex} * x_{experience} + \hat{\alpha}_5 x_{sex} * x_{education}$$

The null hypothesis is that there is no interaction effect between the two variables. From the table, we can can see that $\text{education} * \text{sex}$ have a P-value 0.7525(greater than 0.05) and so we will reject H_0at significant level 0.05. $\text{experience} * \text{sex}$ have a P-value 0.0076 (less than 0.05) and so we fail to reject H_0at significant level 0.05. It implied that there is no interaction effect between sex and education but there is interaction effect between sex and experience. In this way, we will discard $x_{sex} * x_{education}$ from the above model. The New Model now becomes:

$$\hat{y}_{wage} = \hat{\beta}_0 + \hat{\alpha}_1 x_{sex} + \hat{\beta}_2 x_{experience} + \hat{\beta}_3 x_{education} + \hat{\alpha}_4 x_{sex} * x_{experience}$$

Table2: interaction effect of variables

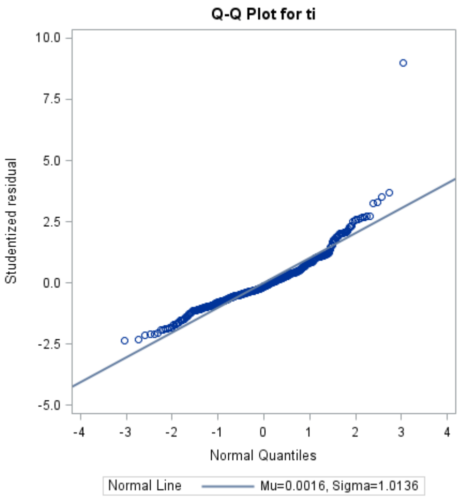


Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	6.680449	6.680449	0.34	0.5594
education	1	2772.839683	2772.839683	141.63	<.0001
experience	1	890.177249	890.177249	45.47	<.0001
education*sex	1	1.949383	1.949383	0.10	0.7525
experience*sex	1	140.543212	140.543212	7.18	0.0076

Check the regression assumptions –test for the normality

In regression models, we assume that the random error terms for each population are from the normal distribution, and all of them have a common variance. We need to check if this model violates any regression assumptions. To test the normality of residual, Q-Q plot is shown in Graph1. It is observed that the points follow a nonlinear pattern and it implied the studentized residual does not follow normal. Box-Cox transformation is then tried to transform the studentized residual into normally distributed. Seeing from Graph2, the linearity of the points suggests that the data are normally distributed after applying transformation. In tests for normality, it is seen from table 3 that three out of four of the test have P-value greater than 0.05, which means the studentized residual follow normal now.

Graph1: Q-Q plot for ti



Graph2: Q-Q plot for ti after Box-Cox transform

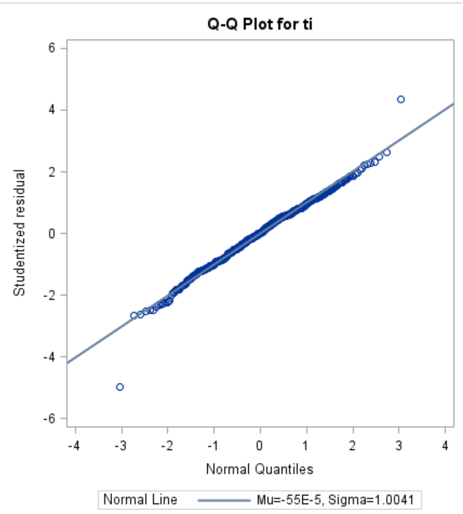
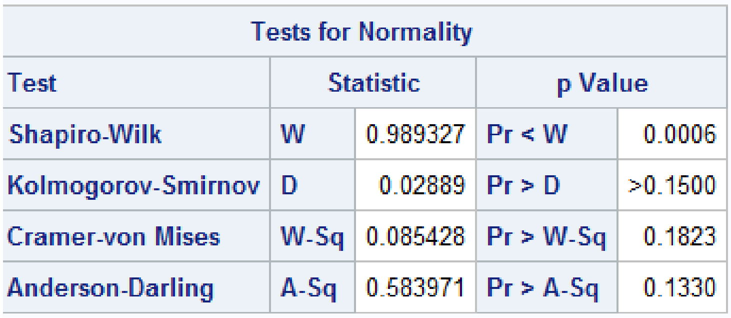


Table3: Tests for normality of residuals

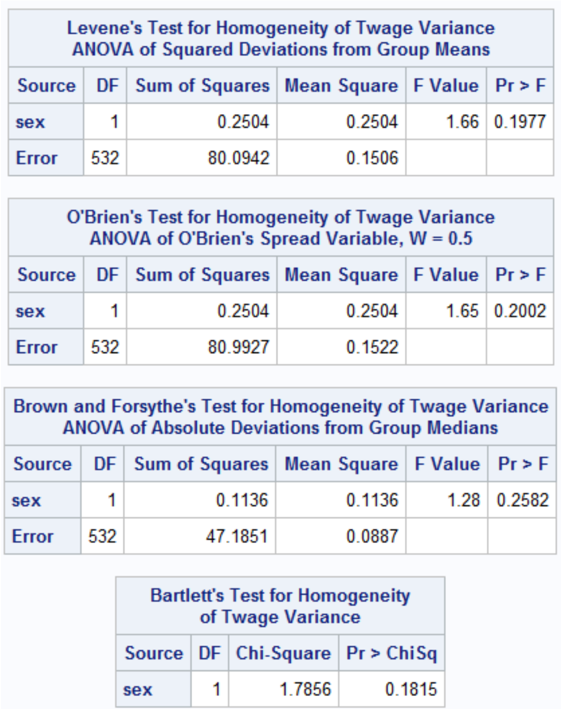


Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.989327	Pr < W	0.0006
Kolmogorov-Smirnov	D	0.02889	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.085428	Pr > W-Sq	0.1823
Anderson-Darling	A-Sq	0.583971	Pr > A-Sq	0.1330

Check the regression assumptions- test for the homogeneity of variances

Table 4: Test for Homogeneity of variances



Levene's Test for Homogeneity of Twage Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
sex	1	0.2504	0.2504	1.66	0.1977
Error	532	80.0942	0.1506		

O'Brien's Test for Homogeneity of Twage Variance
ANOVA of O'Brien's Spread Variable, W = 0.5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
sex	1	0.2504	0.2504	1.65	0.2002
Error	532	80.9927	0.1522		

Brown and Forsythe's Test for Homogeneity of Twage Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
sex	1	0.1136	0.1136	1.28	0.2582
Error	532	47.1851	0.0887		

Bartlett's Test for Homogeneity of Twage Variance

Source	DF	Chi-Square	Pr > ChiSq
sex	1	1.7856	0.1815

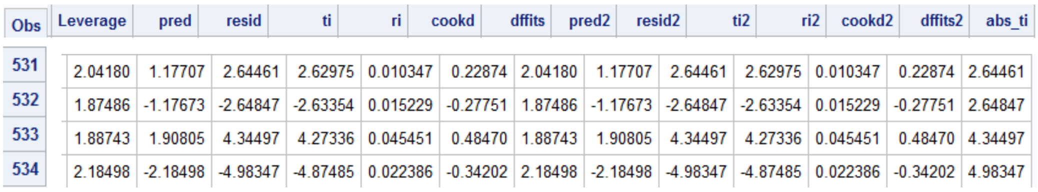
To test whether studentized residual have common variance or not, four tests are carried out. From the Table4, it is showed that four P-value is greater than 0.05, denoted that the null hypothesis is true. It means the equality of the variances of studentized residuals is not violated.

Test the Outlier

Outliers in the model can be caused by measurement errors or may be the result of inherent variability of the data. Thus, we should find out those data, which are considerably exceptional or inconsistent with the rest of the data after we fitted the model. For determine outlier, studentized residual is used as an indicator. Normally, we use 3 as the cutoff value.

From the table5, we can see that there are two observations have large studentized residual (one is 4.34497, and one is 4.98347), which are larger than 4. Since this value is found after absolute the studentized residual and the data set is rearranged in order, we need to find back the observations from the original data set. It is then found that the 171st and 200th observations are outlier. Thus, we should delete this observation and fit the model again.

Table 5: absolute the studendized residual and rearranged the dataset order



Obs	Leverage	pred	resid	ti	ri	cookd	dffits	pred2	resid2	ti2	ri2	cookd2	dffits2	abs_ti
531		2.04180	1.17707	2.64461	2.62975	0.010347	0.22874	2.04180	1.17707	2.64461	2.62975	0.010347	0.22874	2.64461
532		1.87486	-1.17673	-2.64847	-2.63354	0.015229	-0.27751	1.87486	-1.17673	-2.64847	-2.63354	0.015229	-0.27751	2.64847
533		1.88743	1.90805	4.34497	4.27336	0.045451	0.48470	1.88743	1.90805	4.34497	4.27336	0.045451	0.48470	4.34497
534		2.18498	-2.18498	-4.98347	-4.87485	0.022386	-0.34202	2.18498	-2.18498	-4.98347	-4.87485	0.022386	-0.34202	4.98347

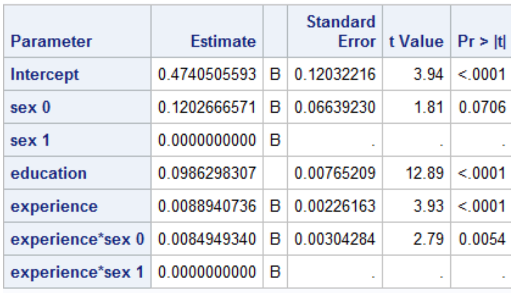
After removing the outlier

From the SAS result, the lamda is 0, so the new fitted model for each level of the factor variable is:

$$\begin{cases} \widehat{\log(y_{wage1})} = (\hat{\beta}_{01} + \hat{\alpha}_1) + (\hat{\beta}_2 + \hat{\alpha}_4)x_{experience} + \hat{\beta}_3 x_{education} \\ \widehat{\log(y_{wage2})} = \hat{\beta}_{02} + \hat{\beta}_2 x_{experience} + \hat{\beta}_3 x_{education} \end{cases}$$

$$\begin{cases} \widehat{\log(y_{wage1})} = 0.5943 + 0.01739x_{experience} + 0.09863x_{education}\ (for\ male) \\ \widehat{\log(y_{wage2})} = 0.4741 + 0.008894x_{experience} + 0.09863x_{education}\ (for\ female) \end{cases}$$

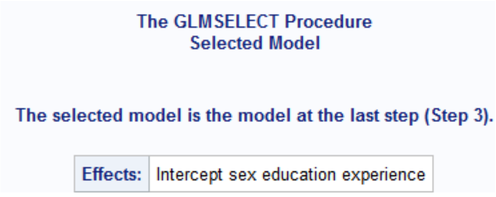
Table 6: parameter estimate



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	0.4740505593	B	0.12032216	3.94	<.0001
sex 0	0.1202666571	B	0.06639230	1.81	0.0706
sex 1	0.0000000000	B	.	.	.
education	0.0986298307		0.00765209	12.89	<.0001
experience	0.0088940736	B	0.00226163	3.93	<.0001
experience*sex 0	0.0084949340	B	0.00304284	2.79	0.0054
experience*sex 1	0.0000000000	B	.	.	.

By using the GLMSELECT procedure in table7, we again get the best model with variables sex, education and experience.

Table7: GLMSELECT procedure



The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept sex education experience

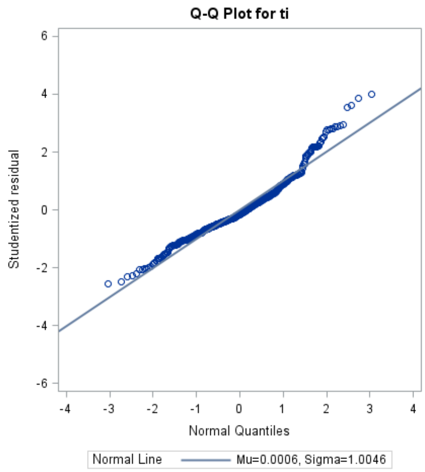
The marital status and age is not selected as model predictors, is an expected result because whether you have married or not does not represent that you are capable in your job. Age also does not consider as a model predictor as even an elder people may have less efficiency and creativity as young people. Nowadays, some successful businessman is young man. So, age may not be a main factor that determines the wage. However, elder people have more experience may result in a high wage. In this way, experience, which can let employees be more practically familiar with the job, may become one of the important determinants in our model.

Moreover, people have higher educational level nowadays comparing to the past. The results of this trend is now creating a hour glass shaped work force in which the middle class is being squeezed out, leaving a large group of highly educated and experienced workers at the top of the labor market and another large group of unskilled workers at the bottom of the market. So this trend may accounts for the model selection results and that is why wage is dependent with the education and experience.

The beta value represent how does the independent variables change with the dependent variables. As see from the above model, it is observed that the beta value in education of male is the same as female. One of the reasons accounts for this may be the education system. The education system forces boys and girls to study since kindergarten to secondary. Nowadays, number of boys and number of girls who enter into university may approximately about the same because the education system forces everyone to study hard and aim to enter university no matters boys or girls. So, education level of male and female may not appear to have significant difference nowadays and so education of male and female has the same degree of influence to the wage.

After refitting this new model, we check the assumption of normality of residual again. Seeing from the QQ-plot in Graph3, it is not following normal distribution. However, the studentized residual again follow normal after Boxcox transformation. In the table8, four of the test have P-value greater than 0.05, which is a better result then the model before removing the outliers. So, we can now conclude that education and experience are the significant indicating factors for wage level for both female and male.

Graph3: Q-Q plot of ti



Graph 4: Q-Q plot for ti after transformation

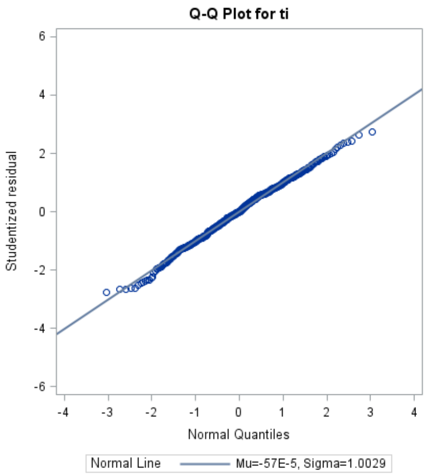
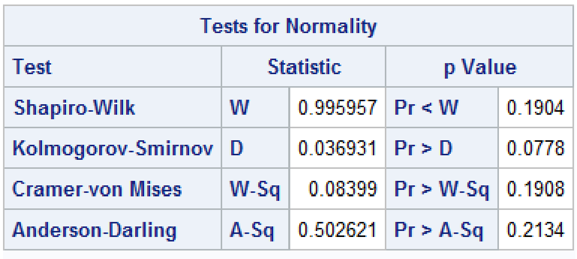


Table8: Test for normality



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.995957	Pr < W	0.1904
Kolmogorov-Smirnov	D	0.036931	Pr > D	0.0778
Cramer-von Mises	W-Sq	0.08399	Pr > W-Sq	0.1908
Anderson-Darling	A-Sq	0.502621	Pr > A-Sq	0.2134

Conclusion

Predictors for the new adjusted wage model will be sex, education and experience. It is foreseeable that experience and education level will become more and more important factors to the wage level. It is better that the new graduate students who aim to get into top of the labor market, should study hard and get a better results in university and should also work hard for internship and part-time in order to gain more working experience.

Limitation on the interpretation and application

In this datafile, there are only 534 observations from the current population survey. In order to get a more accuracy model, more data should be collected. Also, the data should be updated, as the latest population survey is already 5 years before. We should use more variables to fit the model instead of only three variables, since it can then get a more precise result in finding the determinants of wage level.