Spring 2020 | BUAN6337.002 | HW 3

Predictive Analytics using SAS

Group2

Adam Butcher

Dong In Kim

Maria Phetteplace

Mahmoud Ali

Q1-

Part1:

The p-values for 'Number of Kids' and 'Unemployment' are both greater than 0.05 indicating that neither variable is significant at the 95% confidence level.

Our model excludes 'Number of Kids' and 'Unemployment.'

We are left with 'Education,' 'Hr,' Self-Employed,' 'Salary,' 'Age,' 'Married.'

The best model will be:

Regression Model
Ln(wage) = 1.28 +0.07 Edu +0.0001 Hr -0.35 Self +0.295 Sal +0.01 Age +0.16 Mar

Part2:

Checking the correlation coefficients, VIF, and TOL, there is no indication of multicollinearity between any independent variables. All the variance inflation values are less than 10 and the lowest eigenvalue is 0.00924.

Q2-

We tried a variety of models with different squared variables.

We concluded that using education, education^2, hour, hour^2, self-employed, salary, and age allowed all variables to become significant at the 95% level.

We can see that the age sq for example are one of the variables that have nonlinear effect on the model.

Q3-

Answer:

Report of findings:

a. The R2 value is 0.2844 and the adjusted R2 is 0.2793.
   According to the adjusted R2 value, 27.93% of the variance in wages is explaine d by the independent variables.

b. Meaning of coefficients:

I. With each additional year of **education**, wages first decrease 9.52% per hour. After a turning point, wages then increase 7.13% per hour. We suspect that this is due to someone not reaching an educational level that would produce extra earning until having a college degree.

II. With each additional work **hour per year**, wages first increase 0.014% per hour. After a turning point, wage almost does not decrease per hour. We suspect that is because someone who is working many hours may be compensating for making a lower wage.

III. If someone is **self-employed**, wages decrease 34.21% per hour compared to someone who is not self-employed.

IV. If someone is a **salaried employee**, wages increase 22.23% per hour compared to someone who is not a salaried employee.

V. With each additional year of **age**, wages increase 1.19% per hour.

c. Because we have added 'education squared' and 'hours squared,' we will have m ulticollinearity between the linear and non-linear variables.

We have to include both the linear and non-linear variables to be able to interpret those correctly.

VIF values are significant for "**Age_sq**" and "**Edu_sq**"; almost 101 and 17.

We excluded "**Age_sq**" and kept on the "**Edu_sq"** in order to analyze the impact of education as required.

d. According to results of White's Test and Breusch-Pagan test, we reject the null-hypothesis of no heteroskedasticity with the confidence level of 95%.

Since there is heteroskedasticity in the model, we use robust standard errors to in terpret the t-test results.

All the coefficients are significantly different from zero at 95% confidence level.

We realized that white model returned broader std errors. The most significant on e is the hr variable.

Q4-

Rand and Fixed Model					OLS
Variable	**RanONE**	**RanTWO**	**FixedONE**	**FixedTWO**	
Intercept	**1.52633**	**1.526303**	1.472576	2.194828	1.35583
Edu	**0.076103**	**0.076102**	0	0	0.06793
Hr	**-0.00025**	**-0.00025**	**-0.0003**	**-0.0003**	-0.00014230
Self	**-0.27036**	**-0.27037**	**-0.23383**	**-0.23498**	-0.35113
Sal	**0.201757**	**0.201777**	**0.124978**	**0.126807**	0.29299
Age	**0.010295**	**0.010295**	0.016916	0	0.01225
Mar	0.125945	0.125952	0.103028	0.104906	0.13834

Q-5
We did the Hausman test in order to find the best test to use. The result was to reject the null hypothesis which indicate that the best test is the fixed test.

We can see difference between the fixed model and ols model in the following variables:

- **Self:** almost **30% decrease** through the fixed model compared to the OLS.
- **Sal:** almost **65% decrease** through the fixed model compared to the OLS.

For the paneled data, most of the variables changed in the random effects model.
Though the variable **education** has no effect for the fixed effects model as it is time invariant.

Q-6

There is a significance for the variable **edu** as there is a higher coefficient compared to the OLS coefficient.

As we can see in both random effects model evaluations, both values come out to be 0.076, which means that there is a 7.6% increase in income for each additional year of education.