Hourly wages prediction

BY:
BHARATH VENKATESH SRINIVASAN (BXS210006)

Among those nine explanatory variables, only 5 variables are significant namely, edu (education), hr (work hours per year), selfempl (selfemployed), salaried and age. The t-values for all the variables can be inferred from the below output.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: wage1

Number of Observations Read	1002
Number of Observations Used	1002

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	125.91339	15.73917	41.33	<.0001
Error	993	378.13964	0.38081		
Corrected Total	1001	504.05303			

Root MSE	0.61709	R-Square	0.2498
Dependent Mean	2.59348	Adj R-Sq	0.2438
Coeff Var	23.79408		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.35583	0.18072	7.50	<.0001
edu	1	0.06793	0.00722	9.40	<.0001
hr	1	-0.00014230	0.00002524	-5.64	<.0001
selfempl	1	-0.35113	0.05191	-6.76	<.0001
salaried	1	0.29299	0.04433	6.61	<.0001
married	1	0.13834	0.07509	1.84	0.0657
numkid	1	0.02635	0.01955	1.35	0.1781
age	1	0.01225	0.00213	5.75	<.0001
locunemp	1	-0.01494	0.01133	-1.32	0.1876

The best regression model is the one with the 5 explanatory variables – edu, hr, selfempl, salaried and age with the highest R-squared (0.2432) and adjust R-squared (0.2394) value.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: wage1

Number of Observations Read	1002
Number of Observations Used	1002

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	122.60011	24.52002	64.02	<.0001
Error	996	381.45292	0.38298		
Corrected Total	1001	504.05303			

Root MSE	0.61886	R-Square	0.2432
Dependent Mean	2.59348	Adj R-Sq	0.2394
Coeff Var	23.86208		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.41104	0.13934	10.13	<.0001
edu	1	0.06756	0.00719	9.40	<.0001
hr	1	-0.00013233	0.00002491	-5.31	<.0001
selfempl	1	-0.34411	0.05197	-6.62	<.0001
salaried	1	0.29067	0.04417	6.58	<.0001
age	1	0.01173	0.00198	5.93	<.0001

Wage = 1.41104 + 0.06756*edu - 0.00013233*hr - 0.29067*selfempl + 0.29067*salaried + 0.01175*age

Multicollinearity can be inferred by checking the correlation between independent variables. Based on results of VIF and COLLIN, it is inferred that there are no correlation or relationship between the independent variables in the model used to explain the target variable.

VIF diagnosis:

Further interpreting the VIF values, the Variance inflation values for the explanatory variables are less than 10. Hence, it is confirmed that there is no presence of multicollinearity between the eight variables in the model.

COLLIN diagnosis:

Based on the results of COLLIN Results table, the condition indexes are smaller than the threshold value of 100. Thus, it could be inferred that the is no presence of multicollinearity between the variables in the model.

Collinearity Diagnostics								
			Proportion of Variation					
Number	Eigenvalue	Condition Index	Intercept	edu	hr	selfempl	salaried	age
1	4.57784	1.00000	0.00089991	0.00176	0.00502	0.00942	0.01201	0.00229
2	0.84843	2.32286	0.00003168	0.00011490	6.39076E-7	0.65728	0.11648	0.00006084
3	0.40613	3.35737	0.00346	0.00224	0.00411	0.27455	0.73130	0.01077
4	0.10957	6.46368	0.00314	0.01055	0.78099	0.03137	0.02394	0.09494
5	0.04428	10.16799	0.00472	0.48275	0.07583	0.00289	0.05112	0.46860
6	0.01375	18.24647	0.98775	0.50258	0.13405	0.02450	0.06514	0.42335

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.41104	0.13934	10.13	<.0001	0
edu	1	0.06756	0.00719	9.40	<.0001	1.16653
hr	1	-0.00013233	0.00002491	-5.31	<.0001	1.12500
selfempl	1	-0.34411	0.05197	-6.62	<.0001	1.12039
salaried	1	0.29067	0.04417	6.58	<.0001	1.25023
age	1	0.01173	0.00198	5.93	<.0001	1.02932

Using the finalized model with the 5 variables, nonlinearity is checked by adding a squared term of continuous explanatory variables in the model and the output is tabulated below,

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: wage1

Number of Observations Read	1002
Number of Observations Used	1002

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	145.19669	18.14959	50.22	<.0001
Error	993	358.85634	0.36139		
Corrected Total	1001	504.05303			

Root MSE	0.60115	R-Square	0.2881
Dependent Mean	2.59348	Adj R-Sq	0.2823
Coeff Var	23.17945		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.15861	0.42622	2.72	0.0067
edu	1	-0.09797	0.02706	-3.62	0.0003
hr	1	0.00014532	0.00006352	2.29	0.0224
selfempl	1	-0.34517	0.05067	-6.81	<.0001
salaried	1	0.21895	0.04388	4.99	<.0001
age	1	0.05480	0.01904	2.88	0.0041
hr1	1	-6.43299E-8	1.364452E-8	-4.71	<.0001
edu1	1	0.00719	0.00115	6.24	<.0001
age1	1	-0.00049329	0.00021790	-2.26	0.0238

The required results are tabulated below,

Model and its respective squared variable	P > 0.05	Result
Model 1 – Education (edu)	False, fail to reject null	The relationship between Wage and education is non-linear
Model 2 – Work hours (hr)	False, fail to reject null	The relationship between Wage and work hours is non-linear
Model 3 – Age (age)	False, fail to reject null	The relationship between Wage and age is non-linear

Conclusion:

From the results, it is noted that the p-value of edu*edu, age*age and hr*hr is <0.05, so we reject the null hypothesis, signifying that those variables are causing non-linearity. So, we include the squared term of edu, hr and age to add non-linearity to the model.

R-squared inference:

From the below screenshot we could infer that the R-squared & adjusted R-squared for the best fitting model is 0.28 and 0.28 respectively.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: wage1

Number of Observations Read	1002
Number of Observations Used	1002

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	145.19669	18.14959	50.22	<.0001
Error	993	358.85634	0.36139		
Corrected Total	1001	504.05303			

Root MSE	0.60115	R-Square	0.2881
Dependent Mean	2.59348	Adj R-Sq	0.2823
Coeff Var	23.17945		

The R-squared and adjusted R-squared values are similar for the model. 28% of the variance or the variability in the target variable - log(wages) is explained by the explanatory variables in the model.

T- value interpretation:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.15861	0.42622	2.72	0.0067
edu	1	-0.09797	0.02706	-3.62	0.0003
hr	1	0.00014532	0.00006352	2.29	0.0224
selfempl	1	-0.34517	0.05067	-6.81	<.0001
salaried	1	0.21895	0.04388	4.99	<.0001
age	1	0.05480	0.01904	2.88	0.0041
hr1	1	-6.43299E-8	1.364452E-8	-4.71	<.0001
edu1	1	0.00719	0.00115	6.24	<.0001
age1	1	-0.00049329	0.00021790	-2.26	0.0238

The above screenshot indicates that each of the variables in the model (including the square terms that were added to handle non-linearity) are significant with t-values greater than 1.96 (as per 95% CI).

Variable Coefficients:

1. Edu (Education in years)

   Since the coefficient of edu is -0.097 and coefficient of edu*edu (squared term -edu1) = 0.0071 then the shape of the curve first decreases and then increases (U shaped).

   Since we have added a square/quadratic term to handle nonlinearity, the percentagechange depends on the education value. So, we could interpret edu's coefficient as,

   If the year of education increases by one year, then the dollar wage hour changes by 100*(-0.097+2*0.0071*edu) %

2. Hr (Work hours)

   Since the coefficient of Hr is 0.00014 and coefficient of Hr*Hr (squared term -Hr1) = -6.43299E-8 then the shape of the curve first increases and then decreases (Inverted U shaped).

Since we have added a square/quadratic term to handle nonlinearity, the percentage change depends on the education value. So, we could interpret Hr's coefficient as,

If the year of education increases by one year, then the dollar wage hour changes by 100*(0.00014 +2*-6.43299E-8 *Hr) %

3. Selfempl

   The dollar wage per hour will be 34.5% more for people who are not self-employed when compared to the self-employed people

4. Salaried

   The dollar wage per hour will be 21 % more for people who are salaried when compared to the people who are not salaried.

5. Age

   Since the coefficient of age is 0.054 and coefficient of age*age (squared term -age1) = -0.0004 then the shape of the curve first increases and then decreases (Inverted U shaped).

   Since we have added a square/quadratic term to handle nonlinearity, the percentage change depends on the education value. So, we could interpret age's coefficient as,

   If the year of education increases by one year, then the dollar wage hour changes by 100*(-0. 054 +2*0. 0004 *age) %

Collinearity diagnosis:

VIF diagnosis:

Further interpreting the VIF values, the Variance inflation values for the explanatory variables are less than 10. Hence, it is confirmed that there is no presence of multicollinearity between the eight variables in the model.

COLLIN diagnosis:

Based on the results of COLLIN Results table, the condition indexes are smaller than the threshold value of 100. Thus, it could be inferred that the is no presence of multicollinearity between the variables in the model.

Since the coefficient of edu is -0.097 and coefficient of edu*edu (squared term -edu1) = 0.0071 then the shape of the curve first decreases and then increases (U shaped).

Since we have added a square/quadratic term to handle nonlinearity, the percentage change depends on the education value. So, we could interpret edu's coefficient as,

If the year of education increases by one year, then the dollar wage hour changes by 100*(-0.097+2*0.0071*edu) %

Note: Here the % change depends on education value since we are having quadratic term.

Table of Coefficients:

Variables	OLS Regression	Fixed one	Fixed Two	Random one	Random Two
Intercept	1.15861	-0.78993	4.97093	1.322901	1.322944
edu	**-0.09797**	0	0	**-0.09432**	**-0.09432**
hr	**0.00014532**	**-0.00041**	**-0.00041**	**-0.00022**	**-0.00022**
selfempl	**-0.34517**	**-0.22899**	**-0.22991**	**-0.27204**	**-0.27205**
salaried	**0.21895**	**0.12572**	**0.12723**	**0.18169**	**0.1817**
age	**0.0548**	**0.13443**	0	**0.068983**	**0.068978**
hr1	**-6.43E-08**	2.17E-08	2.16E-08	-7.33E-09	-7.33E-09
edu1	**0.00719**	0	0	**0.007299**	**0.007299**
age1	**-0.00049329**	**-0.00141**	**-0.0014**	**-0.00068**	**-0.00068**

OLS & Fixed One:

Variables	OLS Regression	Fixed one	Percentage change
Intercept	1.15861	-0.78993	168.1791112
edu	-0.09797	0	
hr	0.00014532	-0.00041	382.1359758
selfempl	-0.34517	-0.22899	33.65877683
salaried	0.21895	0.12572	42.57867093
age	0.0548	0.13443	145.3083942
hr1	-6.43E-08	2.17E-08	133.7012804
edu1	0.00719	0	
age1	-0.00049329	-0.00141	185.835918

The coefficients of all the variables have changed in the OLS and Fixed One model and the percentage change between the coefficients are mentioned in the table above.

Hr is the variable with a coefficient that has varied the highest among all variables.

The edu is a time invariant variable and is not estimated as part of the Fixed one model.

OLS & Fixed Two:

Variables	OLS Regression	Fixed Two	Percentage change
Intercept	1.15861	4.97093	329.0426459
edu	-0.09797	0	
hr	0.00014532	-0.00041	382.1359758
selfempl	-0.34517	-0.22991	33.3922415
salaried	0.21895	0.12723	41.89038593
age	0.0548	0	
hr1	-6.43E-08	2.16E-08	133.5458317
edu1	0.00719	0	
age1	-0.00049329	-0.0014	183.8087129

The coefficients of all the variables have changed in the OLS and Fixed Two model and the percentage change between the coefficients are mentioned in the table above.
The hr (Work hours per year) variable's coefficient has recorded the highest percentage difference/change between the two models.
Time invariant variables Age and Edu are not estimated in the Fixed two model as they will get cancelled out.

OLS & Random One:

Variables	OLS Regression	Random one	Percentage change
Intercept	1.15861	1.322901	14.1800088
edu	-0.09797	-0.09432	3.725630295
hr	0.00014532	-0.00022	251.3900358
selfempl	-0.34517	-0.27204	21.18666164
salaried	0.21895	0.18169	17.01758392
age	0.0548	0.068983	25.88138686
hr1	-6.43E-08	-7.33E-09	88.60560952
edu1	0.00719	0.007299	1.515994437
age1	-0.00049329	-0.00068	37.84994628

The coefficients of all the variables have between the OLS and Random One model.
Edu (Education in years) has the least change in its coefficient when compared to the rest.

The hr (Work hours per year) variable's coefficient has varied the most between the two models.

OLS & Random Two:

Variables	OLS Regression	Random Two	Percentage change
Intercept	1.15861	1.322944	14.18372015
edu	-0.09797	-0.09432	3.725630295
hr	0.00014532	-0.00022	251.3900358
selfempl	-0.34517	-0.27205	21.18376452
salaried	0.21895	0.1817	17.01301667
age	0.0548	0.068978	25.87226277
hr1	-6.43E-08	-7.33E-09	88.60560952
edu1	0.00719	0.007299	1.515994437
age1	-0.00049329	-0.00068	37.84994628

The coefficients of all the variables have changed in the OLS and Random Two model and the percentage change between the coefficients are mentioned in the table above.

The coefficients of all the variables have changed between the OLS and Random Two model and the percentage change between the coefficients are mentioned in the table above.

The change percentage for hr (Work hours per year) is the highest and for edu (Education in years) is the least.

Non estimable parameters:

In fixed effects model, the parameters Education in years (Edu) and Age are not estimated as they both are time invariant variables, which does not change over time. Those variables tend to cancel out when subtracted in the fixed intercept estimator.

Fixed one

The coefficient of education is 0 in both the cases, this can be because education stays mostly constant over period of time and in fixed effects model it gets eliminated

Random one

With 1 year increase in education, wages change by 100*(-0.094+2*0.0073*edu)%

Fixed two

The coefficient of education is 0 in both the cases, this can be because education stays mostly constant over period of time and in fixed effects model it gets eliminated

Random two

With 1 year increase in education, wages change by 100*(-0.094+2*0.0073*edu)%

"Note: In all the cases, changes can be increase or decrease depending on number of years of education

Ex: If years of education is 10, wages will have a positive coefficient but if it is 1 year, it will have a negative coefficient"

When edu = 1, The coefficient is changed by 0.42% across regression model and random effects model. Though the change seems to be small, its better to use random effects model since we observe some difference in comparison to regression model and also as education increases the difference will be more.

Using Random effects model helps us avoid unobserved heterogeneity.

The correct estimate would be with 1 year increase in education, wages increases by 100*(-0.094+2*0.0073*edu)%