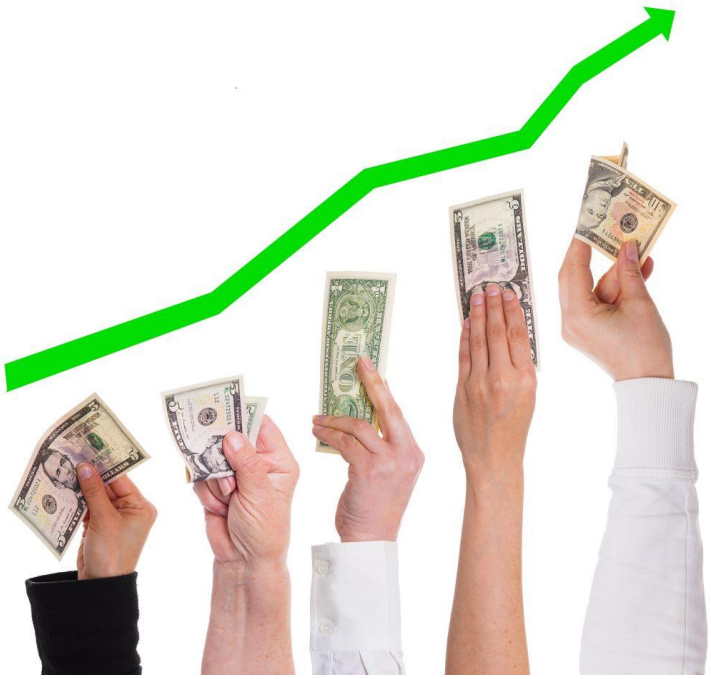

PREDICTING EMPLOYEE WAGES

DETAILS

Handling non-linearity, multicollinearity, heteroscedasticity effects and predicting wages on employee wages data set using Panel Data.

Course

Predictive Analytics using SAS

OBJECTIVE

We need to do a regression to understand the determinants of “natural log (wages)” that is {ln(wage)}.

We need to understand the effect of the following variables: age, edu, numkid, hr, mar, sal, self, unemp.

1. **Find the best linear regression model. Check for multicollinearity and take appropriate actions.**

The REG Procedure
Model: MODEL1
Dependent Variable: lwage

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|	Tolerance	Variance Inflation
Intercept	1	1.35583	0.18072	7.50	<.0001	.	0
age	1	0.01225	0.00213	5.75	<.0001	0.83236	1.20140
edu	1	0.06793	0.00722	9.40	<.0001	0.84330	1.18582
numkid	1	0.02635	0.01955	1.35	0.1781	0.80053	1.24918
hr	1	-0.00014230	0.00002524	-5.64	<.0001	0.86035	1.16232
married	1	0.13834	0.07509	1.84	0.0657	0.90720	1.10229
salaried	1	0.29299	0.04433	6.61	<.0001	0.78944	1.26672
selfempl	1	-0.35113	0.05191	-6.76	<.0001	0.88954	1.12418
locunemp	1	-0.01494	0.01133	-1.32	0.1876	0.97126	1.02959

R squared value = 24.9%. The variables age, edu, hr, salaried, selfempl are found to be significant out of all the variables as they have a p-value of <0.05%.

We ran another model including “famearn” which resulted in a higher R—squared value. Though, it rendered the “numkid” & “married” variables to be significant which from the outside might not be agreed upon as we generally believe that a person’s wage is not dependent on marriage or the number of kids they have. But since the R-squared value increase is relatively high, we have decided to include “famearn” in our model.

Our **best linear regression model** is:

The REG Procedure
Model: MODEL1
Dependent Variable: lwage

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	9	308.19052	34.24339	173.44	<.0001
Error	992	195.86251	0.19744		
Corrected Total	1001	504.05303			

Root MSE	0.44434	R-Square	0.6114
Dependent Mean	2.59348	Adj R-Sq	0.6079
Coeff Var	17.13315		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	1.92126	0.13146	14.62	<.0001	.	0
age	1	0.00783	0.00154	5.08	<.0001	0.82492	1.21224
edu	1	0.01588	0.00548	2.90	0.0038	0.76079	1.31443
numkid	1	0.06169	0.01413	4.37	<.0001	0.79510	1.25771
hr	1	-0.00027012	0.00001866	-14.48	<.0001	0.81661	1.22457
married	1	-0.17208	0.05502	-3.13	0.0018	0.87593	1.14165
salaried	1	0.14827	0.03227	4.59	<.0001	0.77225	1.29492
selfempl	1	-0.24663	0.03754	-6.57	<.0001	0.88207	1.13369
locunemp	1	0.00876	0.00820	1.07	0.2854	0.96247	1.03900
famearn	1	0.00001595	5.248756E-7	30.38	<.0001	0.72704	1.37543

R squared value = 61.1%. The variables age, edu, hr, salaried, selfempl, married, famearn, numkid are found to be significant out of all the variables as they have a p-value of <0.05%.

Root MSE	0.44394	R-Square	0.6125
Dependent Mean	2.59348	Adj R-Sq	0.6086
Coeff Var	17.11764		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	1.67788	0.19600	8.56	<.0001	.	0
age	1	0.00815	0.00155	5.25	<.0001	0.81185	1.23176
edu	1	0.03307	0.01164	2.84	0.0046	0.16810	5.94867
eduhr	1	-0.00000861	0.00000514	-1.67	0.0947	0.04139	24.15770
numkid	1	0.06449	0.01421	4.54	<.0001	0.78408	1.27537
hr	1	-0.00015771	0.00006973	-2.26	0.0239	0.05835	17.13770
married	1	-0.16671	0.05507	-3.03	0.0025	0.87295	1.14553
salaried	1	0.15407	0.03243	4.75	<.0001	0.76343	1.30988
selfempl	1	-0.24667	0.03750	-6.58	<.0001	0.88207	1.13369
locunemp	1	0.00827	0.00819	1.01	0.3128	0.96126	1.04030

We ran some other models with interaction terms (like edu * hr, age * edu). However, the R-squared value was not significantly increased when compared to the model which we ran before. So, we still believe that the model before was the best regression model.

Collinearity Diagnostics												
			Proportion of Variation									
Number	Eigenvalue	Condition Index	Intercept	age	edu	numkid	hr	married	salaried	selfempl	locunemp	famearn
1	7.65397	1.00000	0.00018877	0.00069212	0.00055839	0.00373	0.00165	0.00112	0.00382	0.00290	0.00085651	0.00305
2	0.86514	2.97440	0.00000154	0.00000856	0.00004207	0.00311	0.00004299	0.00001565	0.10709	0.61975	4.022863E-7	0.00646
3	0.57142	3.65987	0.00005865	0.00008237	0.00012036	0.39012	9.130141E-8	0.00144	0.17486	0.16592	0.00092530	0.02370
4	0.40220	4.36240	0.00093418	0.00963	0.00104	0.31444	5.12744E-11	0.00400	0.44072	0.12050	0.00751	0.00203
5	0.24072	5.63883	0.00121	0.00621	0.00054991	0.03105	0.00008738	0.00034419	0.14343	0.00044741	0.01484	0.75769
6	0.10263	8.63575	0.00043670	0.03253	0.00450	0.05303	0.86969	0.01868	0.02872	0.04152	0.00000152	0.02683
7	0.06708	10.68198	0.00187	0.00167	0.04017	0.03938	0.01552	0.71417	0.01165	0.00270	0.16459	0.03731
8	0.04800	12.62794	0.00304	0.21552	0.09400	0.04659	0.03915	0.12761	0.04190	0.02231	0.56203	0.05282
9	0.03990	13.85097	0.00180	0.41857	0.44880	0.07267	0.01662	0.08223	0.04005	0.01038	0.02324	0.05674
10	0.00895	29.23956	0.99046	0.31508	0.41022	0.04588	0.05724	0.05038	0.00797	0.01358	0.22601	0.03338

We found that multicollinearity was not present in our model. Since variance inflation factor of the independent variables is < 10 and the condition index value of the variables is < 100, we concluded that there is no multicollinearity.

2. **Develop a model to test if there are nonlinear effects for some variables. Which variables have non-linear effect on ln(wages).**
   Edu and famearn have non-linear effect on ln(wages).

Root MSE	0.44321	R-Square	0.6138
Dependent Mean	2.59348	Adj R-Sq	0.6099
Coeff Var	17.08924		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.17875	0.16749	13.01	<.0001
age	1	0.00761	0.00154	4.95	<.0001
edu	1	-0.03170	0.02002	-1.58	0.1136
sqedu	1	0.00213	0.00086325	2.47	0.0136
numkid	1	0.06072	0.01410	4.31	<.0001
hr	1	-0.00026680	0.00001866	-14.30	<.0001
married	1	-0.17289	0.05488	-3.15	0.0017
salaried	1	0.13598	0.03257	4.17	<.0001
selfempl	1	-0.25233	0.03751	-6.73	<.0001
locunemp	1	0.00850	0.00818	1.04	0.2990
famearn	1	0.00001571	5.323027E-7	29.51	<.0001

Interaction term “sqedu” has a t-value > 1.96 and hence it is significant which means that edu has a non linear effect on ln(wages)

Root MSE	0.41414	R-Square	0.6628
Dependent Mean	2.59348	Adj R-Sq	0.6594
Coeff Var	15.96851		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.85023	0.12266	15.08	<.0001
age	1	0.00868	0.00144	6.04	<.0001
edu	1	0.00574	0.00517	1.11	0.2675
sqfamearn	1	-6.5853E-11	5.35941E-12	-12.29	<.0001
numkid	1	0.06168	0.01317	4.68	<.0001
hr	1	-0.00030920	0.00001768	-17.49	<.0001
married	1	-0.25843	0.05176	-4.99	<.0001
salaried	1	0.11073	0.03023	3.66	0.0003
selfempl	1	-0.18474	0.03534	-5.23	<.0001
locunemp	1	0.01140	0.00764	1.49	0.1361
famearn	1	0.00002745	0.00000106	25.99	<.0001

Interaction term “sqfamearn” has a t-value >1.96 and hence it is significant which means that famearn has a non-linear effect on ln(wages)

3. **Write a report on your findings. Interpret model fit, t-values, meaning of coefficients, collinearity diagnostics, White test, Breusch-Pagan test etc.**

Model fit can be explained by R squared and Adjusted R-squared values. R squared value shows that 61.1% of the variation in the dependent variable ln(wage) can be explained by the variation in independent variables (age, edu, hr, salaried, numkid, married, selfempl, famearn, and locunemp).

T-values for age, edu, numkid, hr, married, salaried, self empl, and famearnare greater than 1.96 and hence they are significant. Thus, we reject null which means that the coefficients of the respective variables are not equal to zero.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	1.92126	0.13146	14.62	<.0001	.	0
age	1	0.00783	0.00154	5.08	<.0001	0.82492	1.21224
edu	1	0.01588	0.00548	2.90	0.0038	0.76079	1.31443
numkid	1	0.06169	0.01413	4.37	<.0001	0.79510	1.25771
hr	1	-0.00027012	0.00001866	-14.48	<.0001	0.81661	1.22457
married	1	-0.17208	0.05502	-3.13	0.0018	0.87593	1.14165
salaried	1	0.14827	0.03227	4.59	<.0001	0.77225	1.29492
selfempl	1	-0.24663	0.03754	-6.57	<.0001	0.88207	1.13369
locunemp	1	0.00876	0.00820	1.07	0.2854	0.96247	1.03900
famearn	1	0.00001595	5.248756E-7	30.38	<.0001	0.72704	1.37543

As age increases by a year, there is a corresponding increase in wage by 0.7%.

As education increases by one year, wage increases by 1.58%.

As number of hours worked per year increases by one, wage decreases by 0.02%

For married people, the wage is higher by 17.2% when compared to unmarried people.

For salaried employees, the wage is higher by 14.82% when compared to non-salaried employees.

For self employed people, the wages are lower by 24.66% when compared to people who are not self-employed.

When famearn increases by one dollar per year, there is a increase in wage by 0.001%.

Parameter	DF	Estimate	Standardized Estimate	Standard Error	t Value
Intercept	1	1.042852	0	0.458599	2.27
age	1	0.007756	0.109708	0.001580	4.91
numkid 0	1	0.724036	0.501353	0.446303	1.62
numkid 1	1	0.711718	0.419393	0.446913	1.59
numkid 2	1	0.859979	0.528136	0.447447	1.92
numkid 3	1	0.915971	0.365531	0.448738	2.04
numkid 4	1	0.932644	0.205098	0.454962	2.05
numkid 5	0	0	0	.	.
hr	1	-0.000273	-0.320313	0.000018650	-14.63
married 0	1	0.159296	0.061222	0.055924	2.85
married 1	0	0	0	.	.
edu	1	0.013885	0.057531	0.005472	2.54
salaried 0	1	-0.147629	-0.103023	0.032082	-4.60
salaried 1	0	0	0	.	.
selfempl 0	1	0.247068	0.138709	0.037392	6.61
selfempl 1	0	0	0	.	.
famearn	1	0.000015890	0.702699	0.000000521	30.50

The person with no kids has a wage increase of 72% when compared to a person with 5 kids.

The person with one kid has a wage increase of 71% when compared to a person with 5 kids.

The person with two kids has a wage increase of 85% when compared to a person with 5 kids.

The person with three kids has a wage increase of 91% when compared to a person with 5 kids.

The person with four kids has a wage increase of 93% when compared to a person with 5 kids.

Based on VIF, COLLIN, we were able to conclude that our model was free from multicollinearity.

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr> ChiSq	Variables
lwage	White's Test	236.0	51	<.0001	Cross of all vars
	Breusch-Pagan	86.64	9	<.0001	1, age, edu, numkid, hr, married, salaried, selfempl, locunemp, famearn

Based on White-test and breush-pagan test, we found that the p-value was <0.05, which shows that it is significant, and we can reject the null hypothesis (variances are equal), Thus we can confirm that the model has heteroskedasticity.

4. **Using the same model as above, run fixed effects models and random effects models i.e., FIXEDONE, FIXEDTWO, RANONE, RANTWO.**

FIXONE		FIXTWO		RANONE		RANTWO	
Variable	Estimate	Variable	Estimate	Variable	Estimate	Variable	Estimate
Intercept	2.896107	Intercept	2.3948	Intercept	2.21335	Intercept	2.21321
age	-0.01133	age	0	**age**	**0.00577**	**age**	**0.00591**
numkid	0.031537	numkid	0.03163	**numkid**	**0.05281**	**numkid**	**0.05298**
hr	**-0.00038**	**hr**	**-0.0004**	**hr**	**-0.0003**	**hr**	**-0.0003**
married	-0.02619	married	-0.0231	married	-0.0876	married	-0.0874
edu	0	edu	0	**edu**	**0.0185**	**edu**	**0.0184**
salaried	-0.00051	salaried	0.00233	**salaried**	**0.07243**	**salaried**	**0.07306**
selfempl	**-0.1938**	**selfempl**	**-0.1954**	**selfempl**	**-0.2117**	**selfempl**	**-0.2122**
famearn	0.000017	famearn	1.7E-05	famearn	1.6E-05	famearn	1.6E-05
locunemp	**-0.03916**	**locunemp**	**-0.0383**	locunemp	-0.0126	locunemp	-0.0132

5. **What is the effect of panel data models on the coefficients? What parameters have changed and by what percentage?**

For a person/employee, the number of years of education remains constant across the year 1984-1986. So, the Fixed One-way effects model does not take those in account for the model as it does not have any within-group variance. Whereas in Fixed two-way effects model, age of a person is not explained by the model as we know age increases uniformly along with the year which results in no variance across the time period.

In Fixed effects model, the variables “hr”, “selfempl” and “locunemp” are significant as their p-value <0.05%. All the significant variables have negative effect on ln(wage).

In Random effects model, the variables “age”,”numkid”,”hr”,”edu”,”salaried”,”selfempl” are significant as their p-value is <0.05%. The variables “hr” and “selfempl” have negative effect on ln(wage).

The parameters which have changed when compared to the Fixed Effects model are the variables ‘Age’, ‘numkid’, ’Edu’, ‘Salaried’,’selfempl’ & ‘locunemp’. So, ‘Salaried’, ‘numkid’,‘Age’ and ‘Edu’, ‘selfempl’ & ‘locunemp’ has been changed by approximately 7%, 2%,1.6%, 1.8%, 2%, 5% respectively.

6. **We are especially interested in the effect of education on wages. How much (%) has this coefficient changed across the different models?What is the correct estimate of the effect of education on wages?**

For linear regression model, when education increases by one year, wage increases by 1.58 %.

In Fixed Effects model, we are not able to account for the effect of Education as the number of years of education received remained the same for each person.

For Random One-Way Effects model, when education increases by one year, wage increases by 1.8 %.

We have also observed that from Hausman Test, the p-value was <0.05 which means that it is significant, and we reject the null hypothesis of no collinearity between the error terms.

Therefore, Hausman's test suggests us to use fixed effects model. Since the random effects model gives higher percentage when compared to linear regression, the correct estimate of the effect of education on wages would be 1.8% (random effects model).