**GROUP 8**

**Homework 2 – Part A**

**Ray Wong (yxw180006)  
Shambhavi Singh (sxs180107)  
Megha Syam Gutti (mxg180014)  
Yueting Gao (yxg180015)  
Tolulope Oduwole (too170130)**

Loading the data after adding ID and Year to it.

**data** hw2;

set work.WAGE19;

**run**;

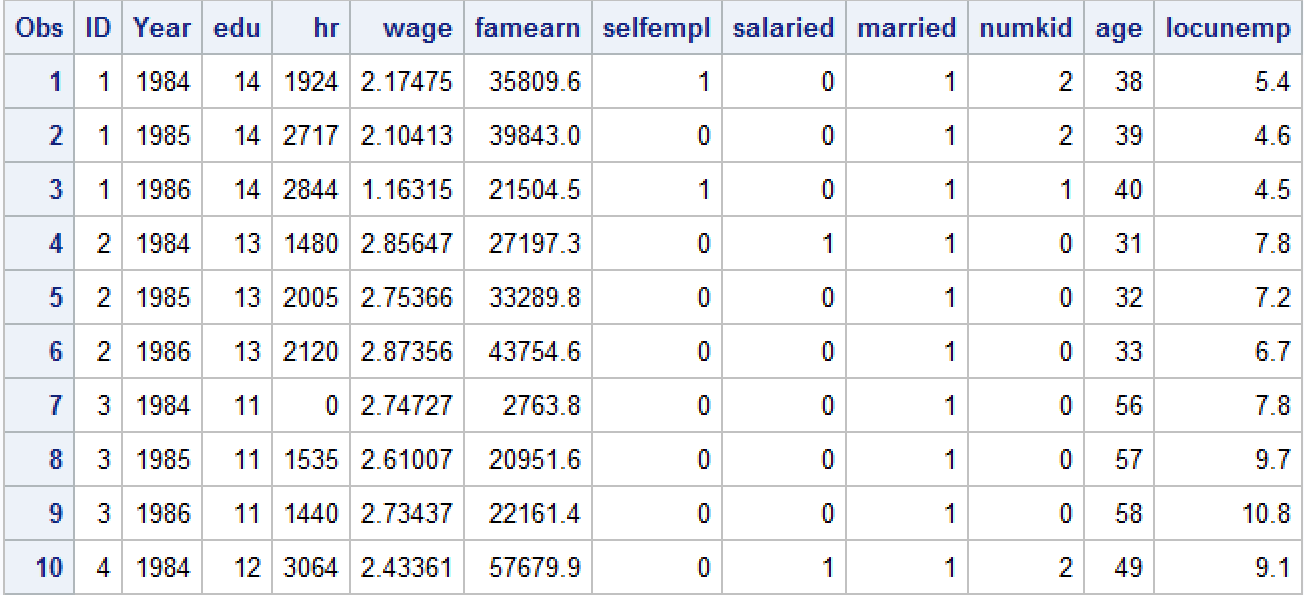
Applying natural log to wage variable.

**data** hw;

set hw2;

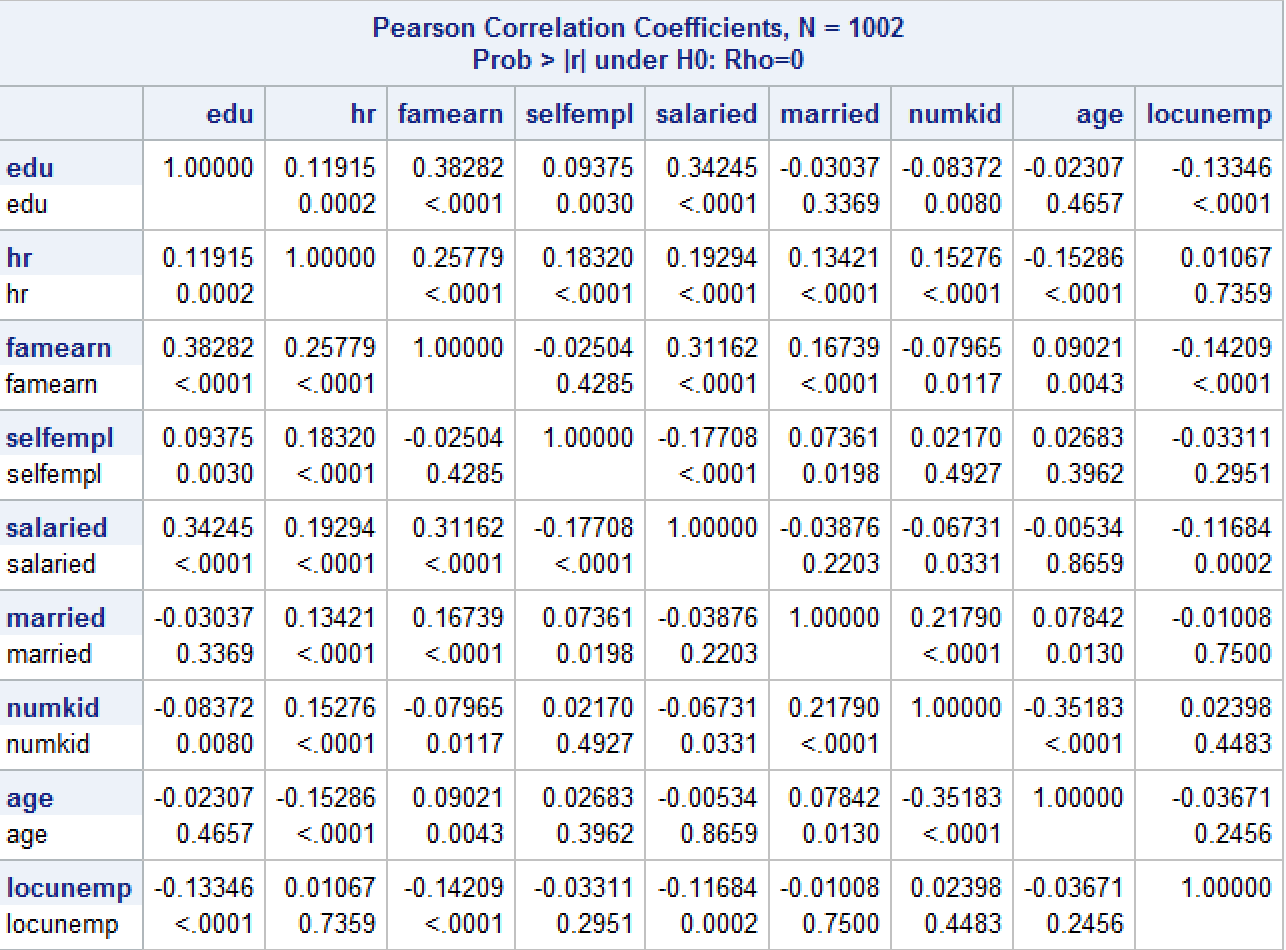
wage = log(wage);

**run**;

A look at the data after making the changes to it:

Q1. Find the best linear regression model. Check for multicollinearity and take appropriate actions. Interpret the estimates and write a report on your findings. Interpret model fit, t-values, meaning of coefficients, collinearity diagnostics, tests for heteroscedasticity etc.

Checking correlation among independent variables:

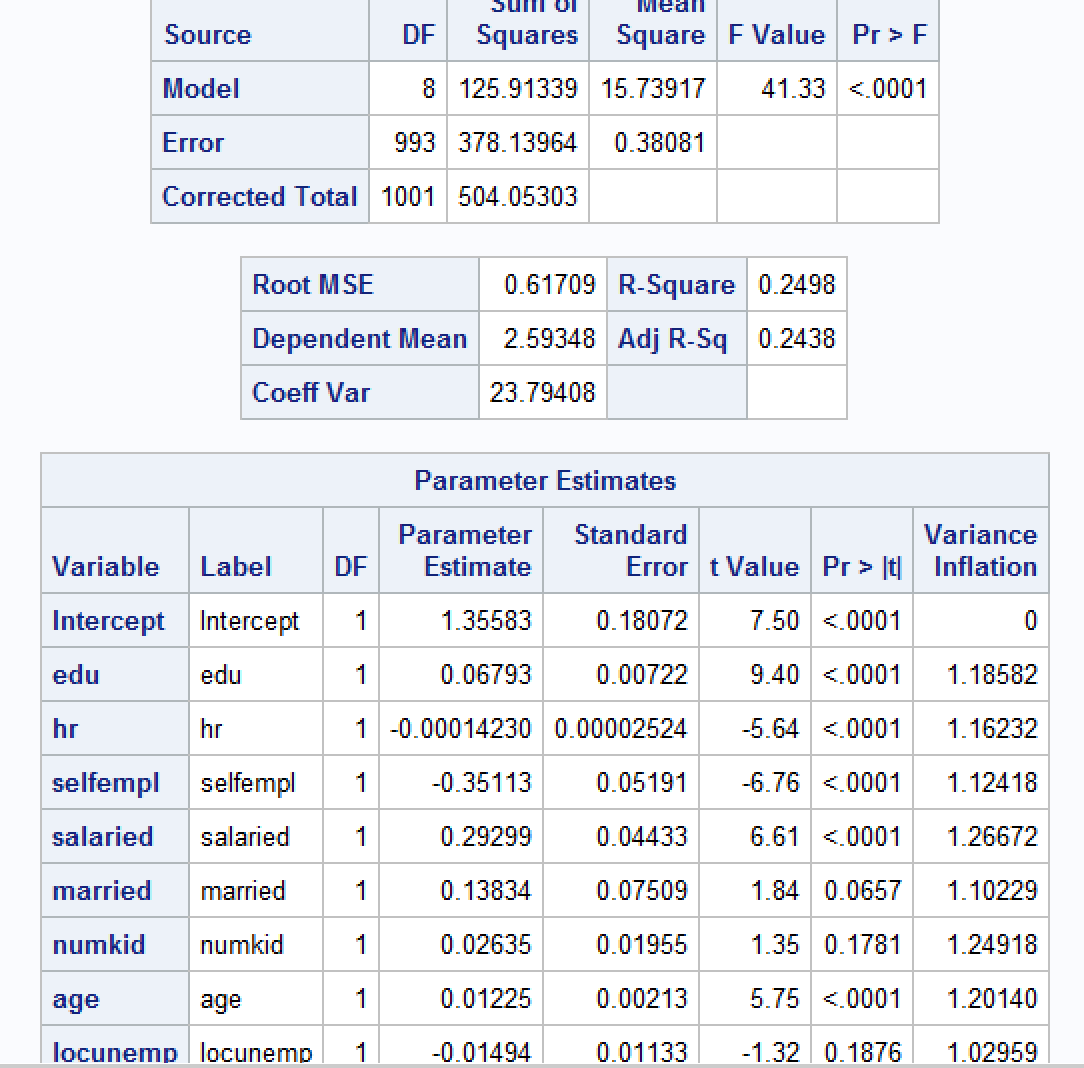


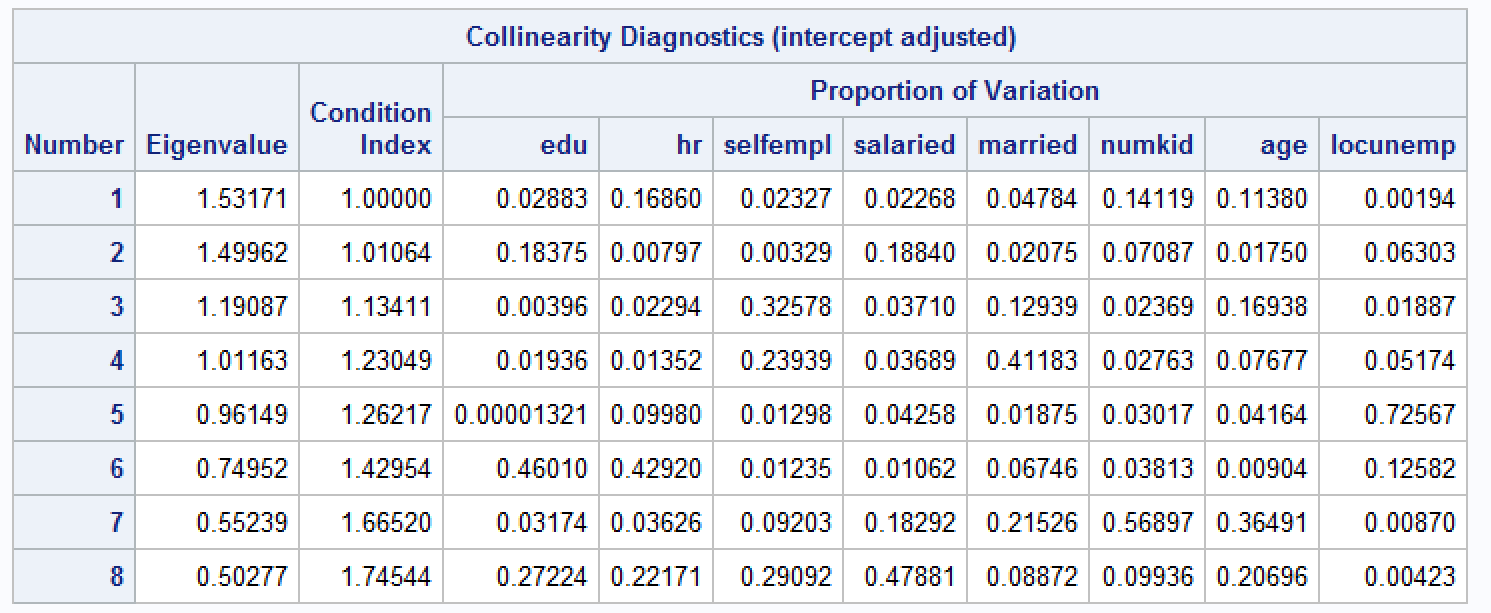
None of the variables are highly correlated from the table above, but this doesn’t ensure that there is no multicollinearity.

**Model 1:**

Wage = f(edu, hr, selfempl, salaried, married, numkid, age, locunemp)

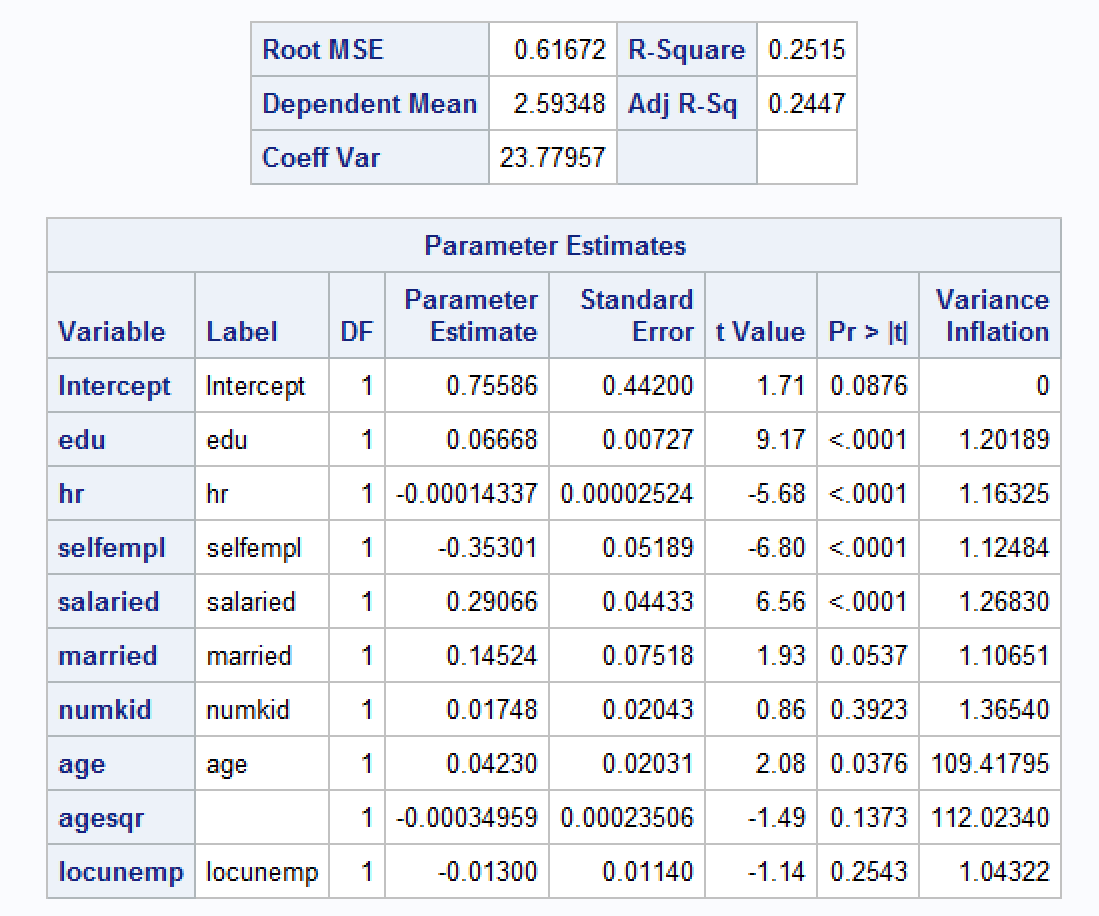
Running the pooled regression with all independent variables with vif and collinoint to check multicollinearity.





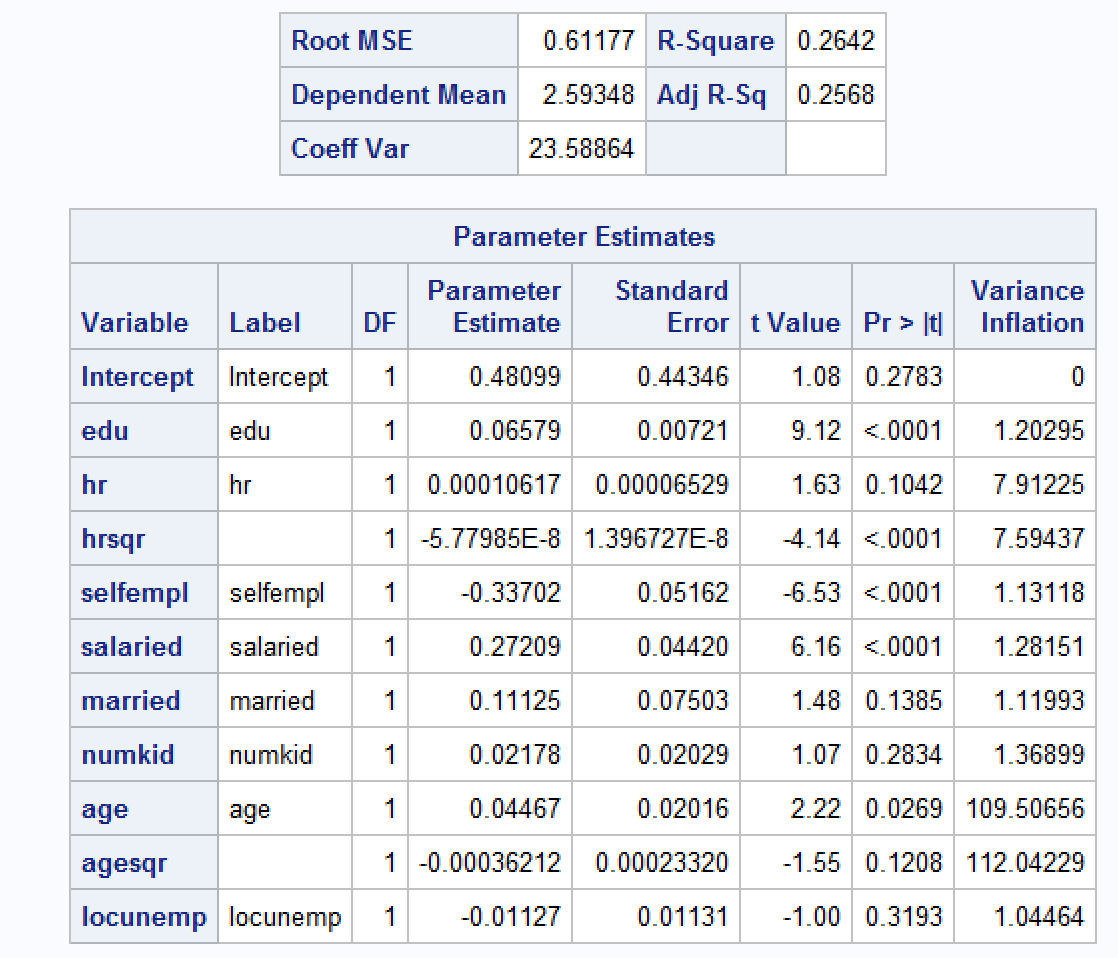
**Model2:**

Wage = f(edu, hr, selfempl, salaried, married, numkid, age, agesqr, locunemp)



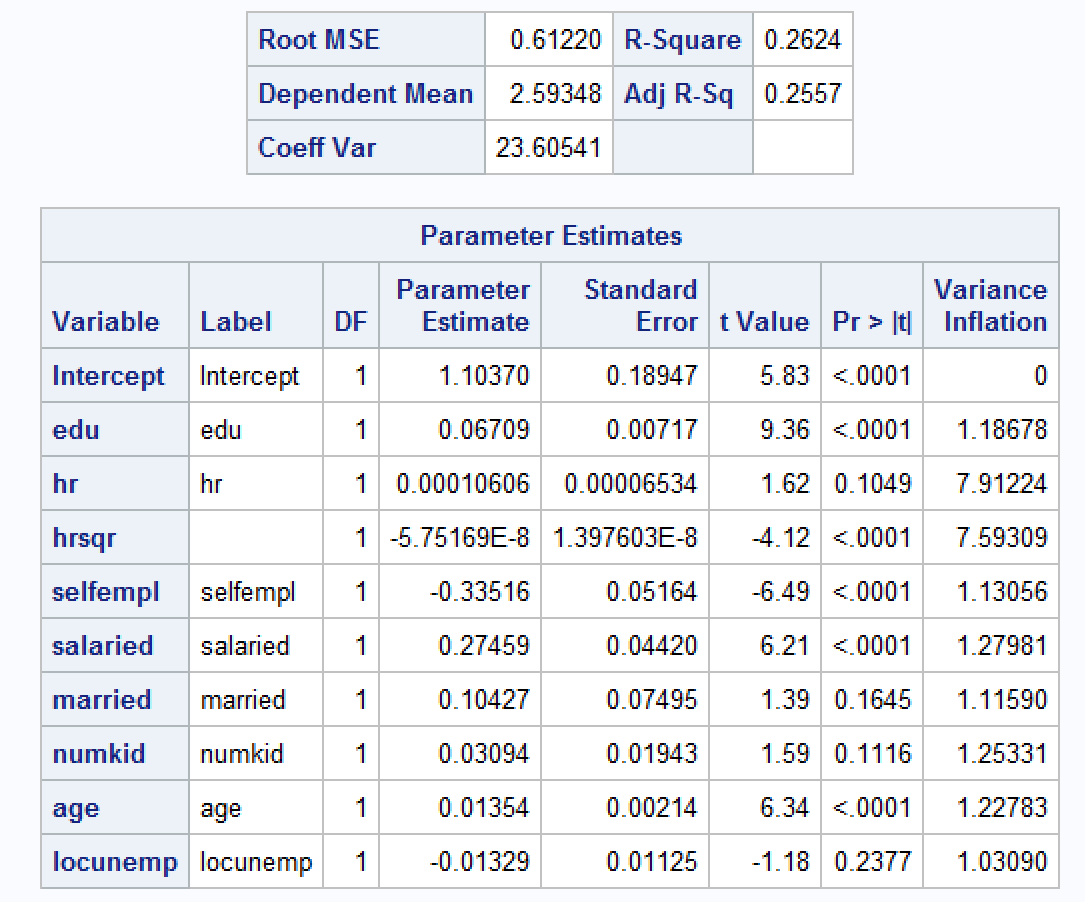
**Model 3:**

Wage = f(edu, hr, hrsqr, selfempl, salaried, married, numkid, age, agesqr, locunemp)



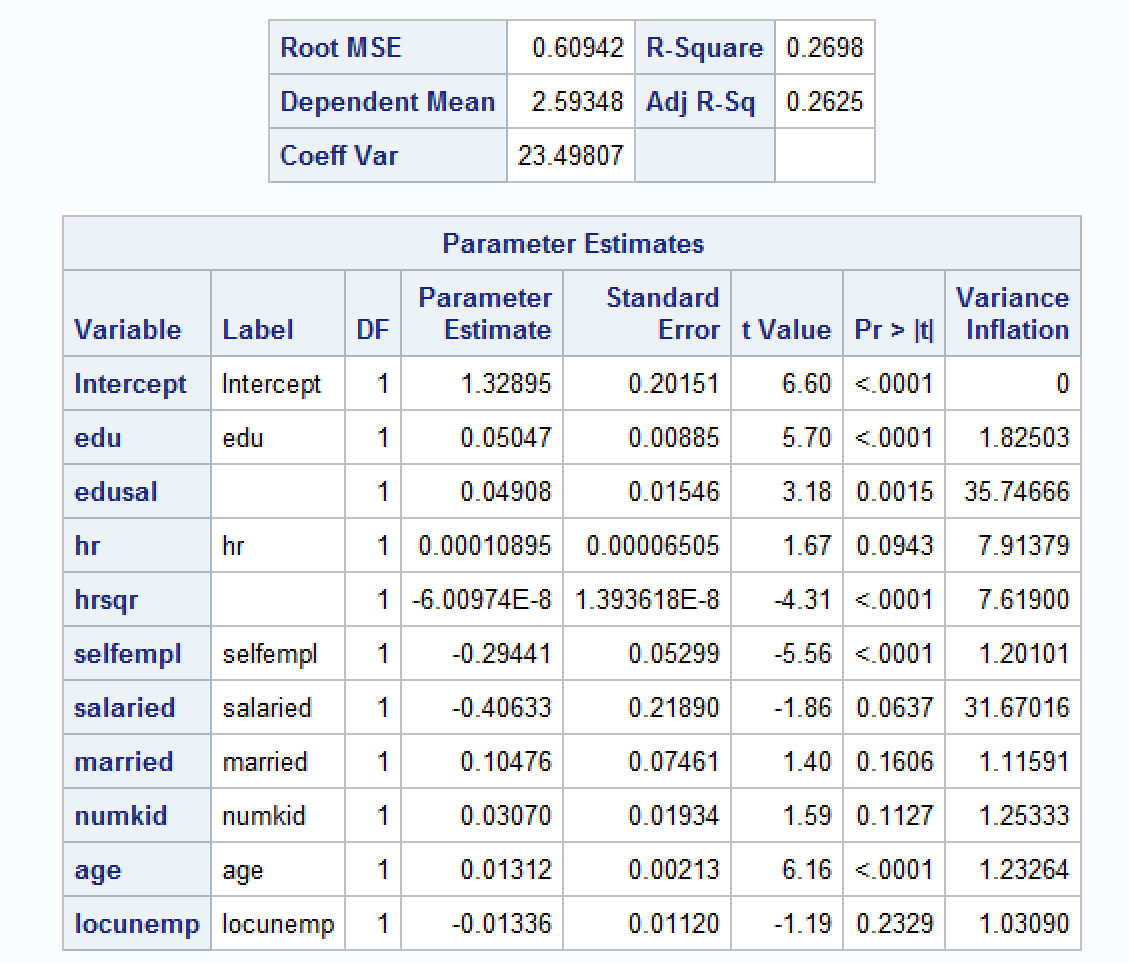
**Model 4:**

Wage = f(edu, hr, hrsqr, selfempl, salaried, married, numkid, age, locunemp)

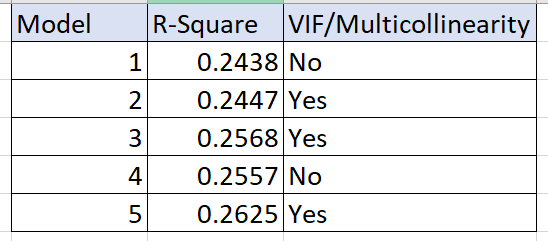


**Model 5:**Wage = f(edu, edusal, hr, hrsqr, selfempl, salaried, married, numkid, age, locunemp)

Edusal is an interaction variable of variables edu and salaried, created in an assumption that a person with high education and being salaried will result in more wage.



**Model Comparison:**



All the models have almost similar R- Square, Model 5 has highest value, but this model has Multicollinearity.   
Only Model 1 and 4 has no multicollinearity but adding hr2 variable to Model 4 is making hr insignificant.

**Model 1 vs Model 4:**

There is no significant difference in R-square between both the models, so it is up to the analyst to choose the model. We are going by Model 1 because, an addition of one significant variable will give a lot of insights and meaningful interpretations.

**Conclusions based on Model 1:**

1. T values and the standard errors look normal, they are not big values, which is not the case when there is multicollinearity.
2. As shown above, none of the independent variables have Variance inflation factor greater than 10. Also, Condition index is not greater than 30.
3. The variables Numkid and locunemp are not statistically significant, the variable married is insignificant at 95% confidence level but might be significant at 90% confidence level.

**Coefficients meaning:**

Edu: one additional year of education will increase the wage by about 7%.  
selfempl: self-employed people will have 35% less wage than non-self-employed people.  
salaried: Salaried people will have about 30% more wage then non-salaried people.  
married: Married people will have about 14% more wage than unmarried people (This interpretation cannot be made at 95% confidence level).  
numkid: this variable is insignificant, so we cannot make an interpretation.  
age: an increase in the age of one year will increase the wage by 1%.   
locunemp: this variable is insignificant, so we cannot make an interpretation.

**Test for heteroscedasticity:**

White test:

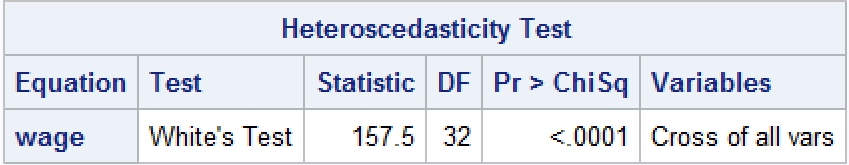
**proc** **model** data = hw;

parms b0 b1 b2 b3 b4 b5 b6 b7 b8;

wage = b0 + b1\*edu + b2\*famearn + b3\*selfempl + b4\*salaried + b5\*married + b6\*numkid+ b7\*age+ b8\*locunemp;

fit wage/white;

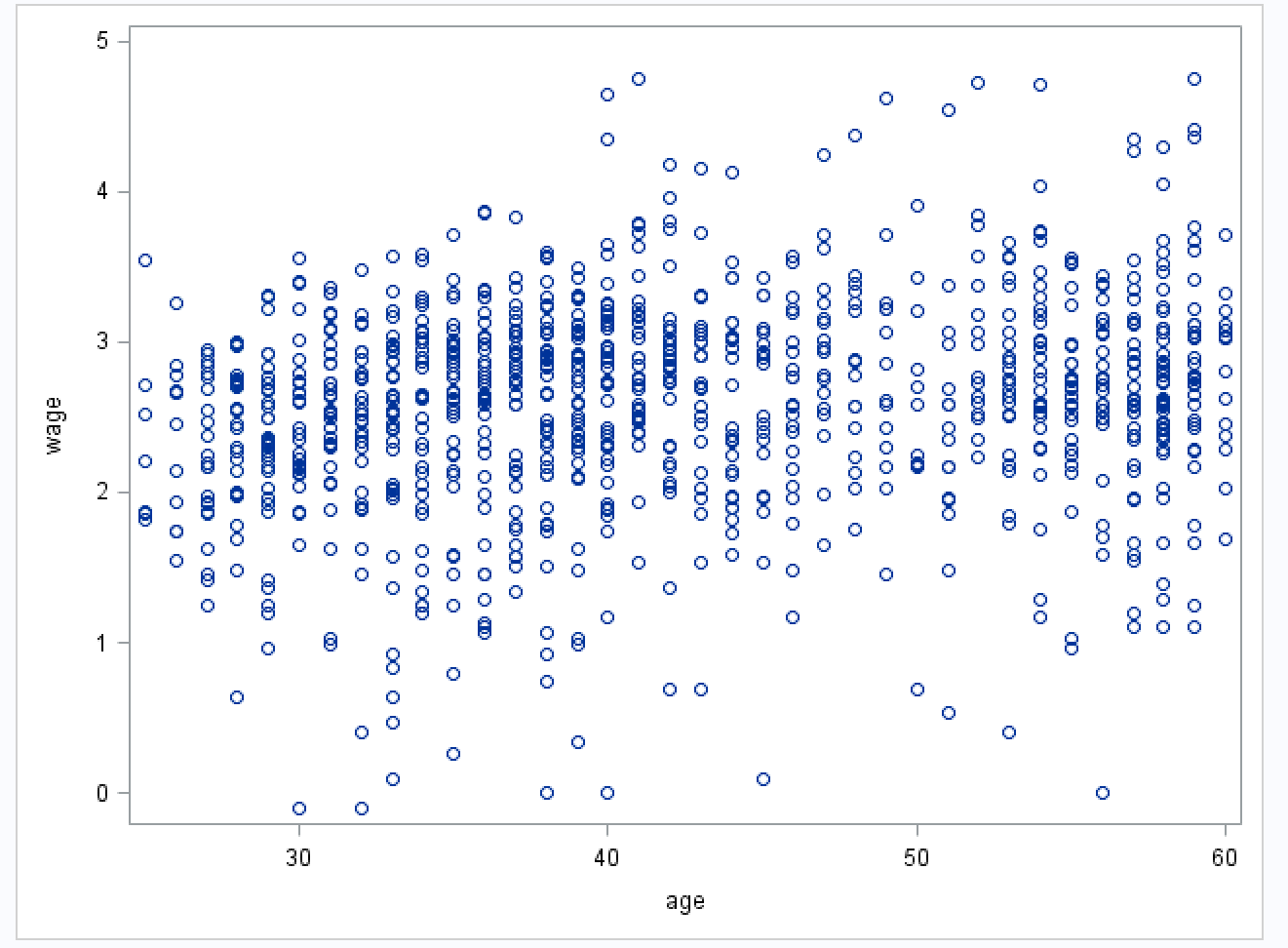
**run**;



White test statistic is statistically significant, so we reject the Null hypothesis that the model is homoscedastic. This model has heteroscedasticity.

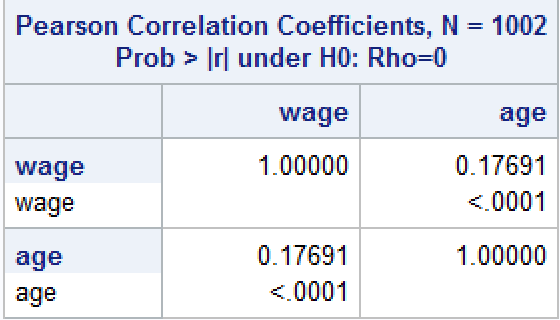
**Q2. Develop a model to test if there are nonlinear effects for age. What do you conclude?**

Scatter plot of wage vs age:

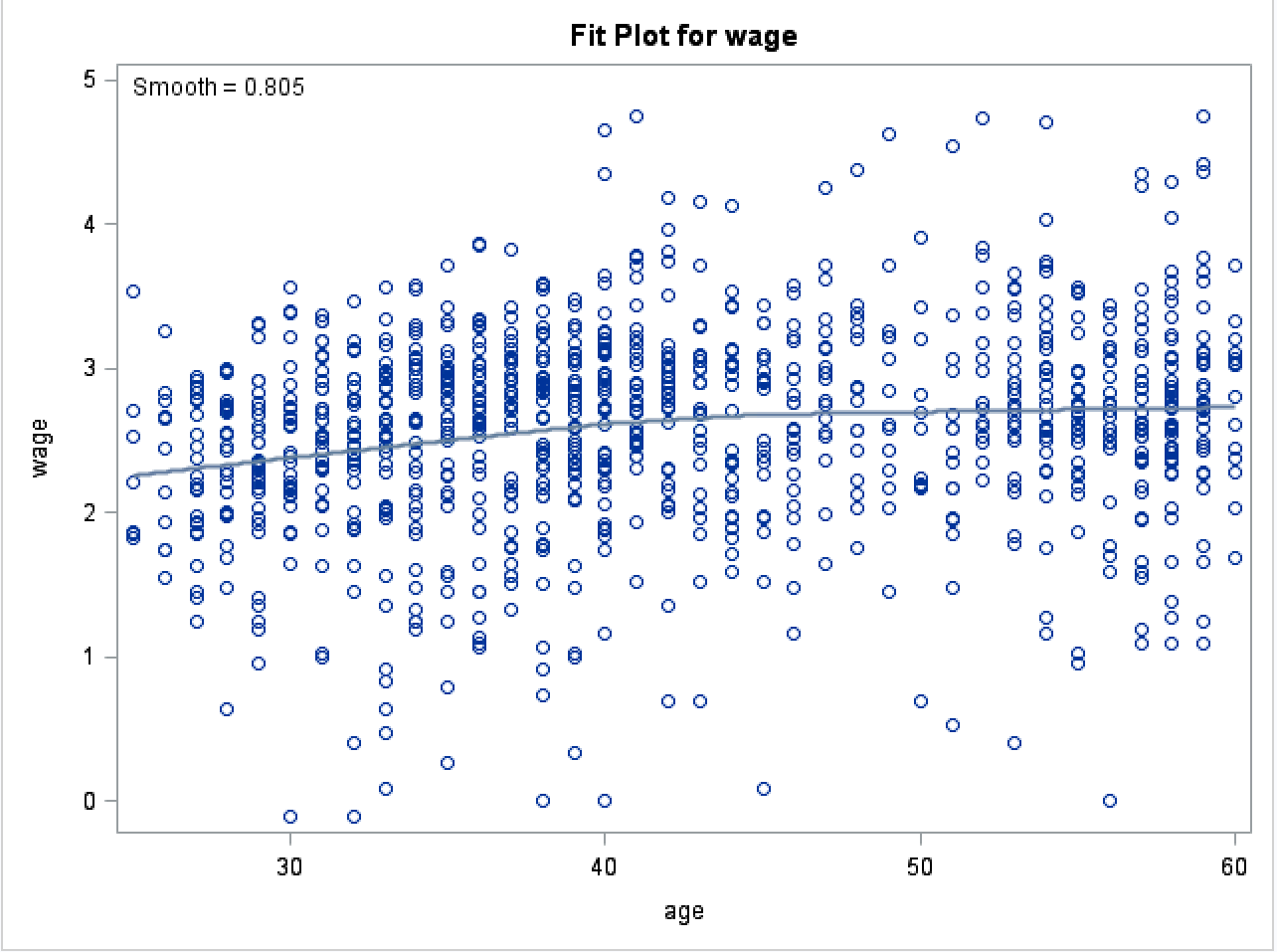


Scatter plot shows that age has weak linear relationship with wage.

**Correlation between wage and age:**

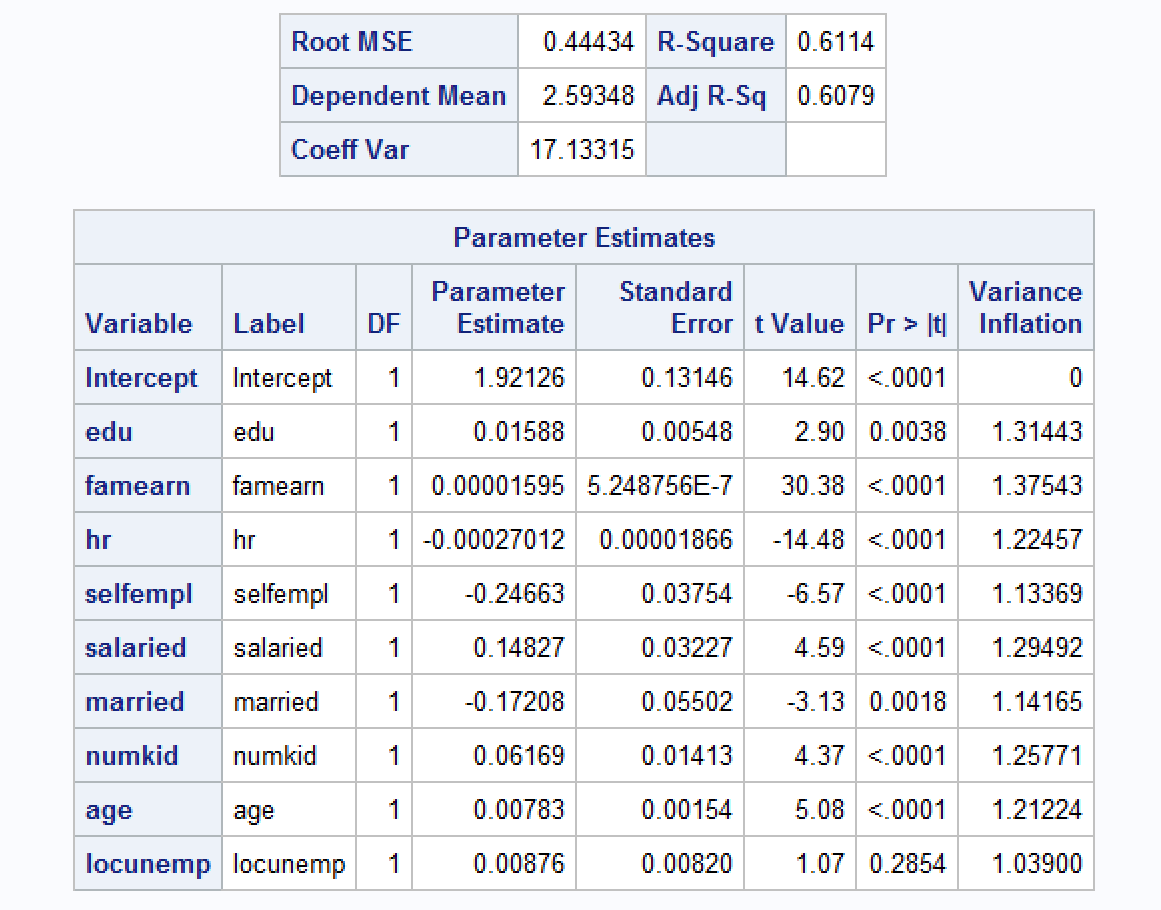
  
Correlation between wage and age is only 0.17, so there is a weak linear relationship between wage and age.

Finding the best fit plot:



The graph above shows that age has slight polynomial relation with wage. Also, the variable agesqr is significant in the “wage = age agesqr” model. The equation between wage and age could be like Wage = a + b\*age – c\*age2.

Q3: **Should you include family earnings in the explanatory variables? Why or why not?**



**Conclusions**:

* Famearn is a statistically significant variable.
* Adding Famearn to the model increases the adj-R2 from 0.24 to 0.60, which means, a lot of dependent variable variance is captured by the famearn variable.
* Removing Famearn from the model also changes the coefficients of other independent variables significantly. So, famearn should be included in the model.

Q4**. Using the same regression model as in Q1, run fixed effects models and random effects models i.e., FIXEDONE, FIXEDTWO, RANONE, RANTWO. Create a table of coefficients side-by side with significant coefficients shown in bold (you may do this in Excel).**

  
(Note: This is a spreadsheet, double click to access the complete spreadsheet)

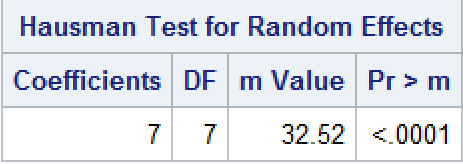
**Q5.** **What is the effect of panel data models on the coefficients? What parameters have changed and by what percentage?**

As shown in the spreadsheet above, as compared to original pooled regression model:

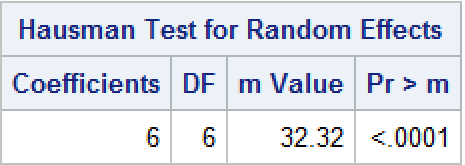
* Edu coefficient is 0 in Fixed Effects model and increased by 11.4% in Random effects model.
* Hr decreased by about 110% in FE and 82.7% in RE model.
* Selfempl increased by 31% in FE and 21.7% in RE model.
* Salaried increased by 58% in FE and 32.5% in RE model.
* Married is insignificant across all the models.
* Numkid is insignificant in OLS and FE, but significant in RE and decreased by 75%.
* Age increased by 58% for FE\_1way (Insignificant) and 6% in RE model.
* Locunemp decreased by 174% in FE, 89% in RE model.

**Q6. 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?**

The coefficient of education in Fixed effects model is 0, because, there is no change in the length of education for all the individuals over the years. In the Random effects model, edu is decreased by 11.4% from 0.067 to 0.075.

Ranone:  
  


Rantwo:



H0 = No correlation between ui and X variables.  
HA = ui and X variables are correlated.

Hausman test for both the Random effects models rejects the Null hypothesis. So, Fixed effects model is the best model in this case.

The effect of education on wages is zero.