Computer Assignment- 2A: Multiple Regression Model: Estimation of earnings function

This analysis is based on the second wave (2011-12) of the Indian Human Development Survey Data. For this particular study we are excluding the state of Madhya Pradesh with STATEID 23. Furthermore, we are taking in few conditions where we are keeping values where the age is between 15 to 65, indus_grp = 9 & 999, dwg = 1.

Annual Earnings is mentioned as AE and Monthly Earnings is given by ME. Total hours worked in the last one year is given by the variable ***wrkhr*** which has zero values in it, therefore we exclude the observations with zero workhours, as the individuals with zero workhours cannot have a positive wage.

After giving the following conditions and creating necessary variables, the original dataset which had 204569 obs. & 61 variables, is now with a total of 70 variables & 47574 obs.

1. **Estimate the Mincer's earnings function where log annual earnings is the dependent variable, experience, experience-square, eduyrs, eduyrs-square are the regressors. Write the econometric equation for this model. Report the results in Table 1 under column heading Model 1. Question: Discuss the significance, the signs and interpretation of the estimated coefficients including the intercept.**

Model 1:

$\widehat{\log(AE)} = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 eduyrs + \beta_4 eduyrs^2 + U_i$

FITTED EQUATION:

$\widehat{\log(AE)} = 8.5593042 + 0.0678988exper - 0.0009397exper^2 + 0.0580324eduyrs + 0.0056221eduyrs^2$

COMMENT ON THE SIGNIFICANCE, SIGNS AND INTERPRETATION OF THE COEFFCIENT ESTIMATES:

We use t-test for each of the estimated coefficient, to determine the significance of each coefficient. We can conclude that the estimated coefficient is statistically significant if: $\mathbf{t_{cal} > t_{tab}}$.

The t_{tab} value for model 1 is 1.644886, the t_{cal} for each of the coefficient estimates can be obtained by using the formula: $\frac{\widehat{\beta}}{se(\widehat{\beta})}$ **(note: the critical value for the exper^2 is in negative terms)**

Calculation of t_{cal} for each coefficient estimates: (at 5% level of significance)

$$\mathbf{H_0: \beta_j = 0 \quad H_1: \beta_j \neq 0 \text{ where } j = 1,2,3,4}$$

- $t_{exper} = \frac{0.0678988}{0.02306} = 2.945$, as $\mathbf{t_{cal} > t_{tab}}$, we reject null hypothesis concluding that experience has an impact on log(AE).
- $t_{exper^2} = \left|\frac{-0.0009397}{0.00002394}\right| = 39.252$, by taking a two-tailed test and the absolute value, $\mathbf{|t_{cal}| > t_{tab}}$, we reject null hypothesis stating that exper^2 has an impact on log(AE).
- $t_{eduyrs} = \frac{0.05803}{0.003279} = 17.698$, as $\mathbf{t_{cal} > t_{tab}}$, we reject null hypothesis concluding that eduyrs has an impact on log(AE)
- $t_{eduyrs^2} = \frac{0.005622}{0.00022} = 25.554$, as $\mathbf{t_{cal} > t_{tab}}$, we reject null hypothesis and concluding that eduyrs^2 has an impact on log(AE)

Comment on Signs of the Coefficient Estimates:

- All the coefficient estimates hold a positive sign except for exper^2 which states that experience increases at a diminishing rate for a larger increase in experience which in our model is the square of experience. Therefore, showing that for more and more increase in experience, the log(AE) tend to have a negative impact from the variations in square of experience.

- We can also state that the signs of other coefficient estimates is positive stating that there is positive linear relationship between log(AE) and experience, education years and square of education years.

Interpretation:

- **Intercept:** The intercept of the model 1 can be interpreted in terms of the base wages, irrespective of the attributes (independent variables) of an individual i.e. when the coefficient of the independent variables is all zero. Therefore, in our model the base log(AE) is **8.5593042** dollars which when converted to dollar terms is **5215.5 dollars.**

- **Coefficient Estimates:**

  1. **Experience:** For an one unit increase in experience, the log(AE) **increases** by **0.0678988** dollars.

  2. **Experience^2:** For an one unit increase in exper^2, the log(AE) **decreases** by **0.0009397** dollars.

  3. **Education years:** For an one unit increase in education years, the log(AE) **increases** by **0.0580324** dollars.

  4. **Education years^2:** For an one unit increase in eduyrs^2, the log(AE) **increases** by **0.0056221** dollars.

2. In the above specification, retain all other variables as it is and change the following: (a) the dependent variable is monthly earnings instead of annual earnings to and (b) exper-square is transformed to exper-square divided by 100.
Estimate this regression model and add the results in a new column with heading Model 2 in Table 1. The columns should report the estimated coefficients and p-values for each of the models. Further, mark with a *, ** and * respectively if the coefficient is statistically significant at 10%,5%, and 1% level of significance.**
Question: Compare the changes in the estimated coefficients, standard errors and p-values if any. Derive the changes in the results due to the transformation of the variables for the OLS estimator and compare the estimated values in Columns 1 and 2.

Model 2

$\widehat{\log(ME)} = \beta_0 + \beta_1 \text{exper} + \beta_2(\text{exper}^2/100) + \beta_3 \text{eduyrs} + \beta_4 \text{eduyrs}^2$

$\widehat{\log(ME)} = 6.074398 + 0.067899\text{exper} - 0.093973(\text{exper}^2/100) + 0.058032\text{eduyrs} + 0.005622\text{eduyrs}^2$

COMPARISON OF CHANGES IN THE MODEL:

The **estimated intercept** in model 1 and model 2 has changed, as we transformed the dependant variable from log(AE) to log(ME) by which the measure of the model has now changed to monthly terms has induced **no changes in any of the p-value, coefficient estimates and standard error.**

From the overall transformations of the dependant variable to monthly earnings and the independent variable exper^2 to $\frac{\text{exper}^2}{100}$ has just supressed the value of the annual earnings and the exper^2 with no loss in the ordering of the values due to the transformations.

Further, the R^2 value of both the models are same with **0.241,** i.e. the model estimates **24%** of variations in the log(AE) & log(ME), stating that, the transformations above does not prove to have any improvement in the model and thereby, concluding to our interest that transformations have not changed the coefficient estimates of the model with no loss of generality in the ordering of the model.

3. **Replace exper and (exper-square/100) with age and age-square/100. Keep all the other variables as in Question (2) above. Add these results in additional column of Table 1 under column heading Model 3.**
   Question: Which of the two models in columns 2 and 3 will you prefer to estimate and why? Provide justification using (a) an appropriate measure of goodness of fit and (b) by interpreting the coefficient of education respectively in models 2 and 3.

Model 3

$\widehat{\log(ME)} = \beta_0 + \beta_1(age) + \beta_2(age^2/100) + \beta_3(eduyrs) + \beta_4(eduyrs^2) + U_i$

$\widehat{\log(ME)} = 5.43966 + 0.07806(age) - 0.08109(age^2) + 0.05735(eduyrs) + 0.00436(eduyrs^2)$

a. Goodness of fit:

Table1.1: Goodness of fit

	Model 2	Model 3
$\bar{R}^2$	0.241	0.2263

- From Table1.1 we can easily say that the model 2 is preferred over model 3 by the fact that **model 2 has 24% variations on log(ME)** over **model 3 which has 22% variations on log(ME)**
- We can also add on to support our answer by showing that when **exper^2** is added to the model, it has **lower diminishing returns** to log(ME) with coefficient estimate being negative at **-0.093973** while comparing it to **age^2** which has a coefficient estimate of **-0.08109,** this maybe the reason for **model 2's edge over model 3**.

b. Interpretation of coefficient estimate of education:

- **Case 1(model 2): Experience added to the model with education:** An effect is visible between education on log(ME) with **0.058032** increase in log(ME) for an unit increase in education.
- **Case 2(model 3): Age added to the model with education:** An effect is visible between age on log(ME) with **0.057535** increase in log(ME) for an unit increase in age.

Therefore, Model 2 is preferred over Model 3 with an effect in education to the models.

c. $\ln(AE) = 8.5600313 + 0.0678672exper - 0.0009392exper^2 + 0.0579346eduyrs + 0.0056271eduyrs^2$

Table 1

	Model1	P-Value	Model2	P-Value	LOS	Model3
Y	log(AE)		log(ME)			log(ME)
β_0	8.5593042	2e-16	6.074398	2e-16	***	5.43966
β_1	0.0678988	2e-16	0.067899	2e-16	***	0.07806
β_2	-0.0009397	2e-16	-0.093973	2e-16	***	-0.08109
β_3	0.0580324	2e-16	0.058032	2e-16	***	0.05735
β_4	0.0056221	2e-16	0.005622	2e-16	***	0.00436

4. (i) Now include log of hours worked in Model 3 report the results in column 2 of Table 2.
(ii) Estimate the restricted model if you have to test the hypothesis for the significance of log hours using F-test approach.
(iii) Estimate the restricted model if you have to test the hypothesis that the coefficient of log hours is one using the F-test approach.
In Table 2 present the results for the unrestricted model as in (i) in column 1, restricted model as in (ii) in column 2 and restricted model as in (iii) in column 3. Add separate rows with values of TSS, RSS, R^2, $\bar{R}^2$(adjusted R^2) below the respective model columns. Make sure to write the appropriate econometric equations and the tests of hypothesis while you answer the questions.
Question:
(a) Are the TSS values the same across these three models, why or why not? Comment on when these three models can be compared using $R^2/\bar{R}^2$ and when not. Write the restricted and unrestricted econometric models and the respective decomposition of sum of squares to justify your answer.
(b) Use the RSS and the R^2 formula of the F-test to test the hypotheses in (ii) and (iii). Report the F-stats value in two additional rows in Table 2 for the respective models. Why does the F-stats values calculated using the two different formulas remain the same in one model while it changes for another model? Which of the two F-stats should be used when the values are different?
(c) Now, add a row with tabulated F values along with the degrees of freedom at 1% level of significance and add one more row with its p-value for a two-sided test. Use R commands to generate the tabulated values and the p-values. Comment on the findings of the results for rejecting/accepting the null hypothesis based on calculated and tabulated F-values to decide about the acceptance/rejection of the respective null hypothesis against that alternative hypothesis.

UNRESTRICTED MODEL (MODEL 4):

$\widehat{\log(ME)} = \beta_0 + \beta_1 age + \beta_2(age^2/100) + \beta_3\ eduyrs + \beta_4 eduyrs^2 + \beta_5 \log(wrkhr) + U_i$

$\widehat{\log(ME)} = -1.7115268 + 0.0187197age - 0.0001208(age^2/100) + 0.0067853eduyrs + 0.0050383eduyrs^2 + 1.1588020\log(wrkhr)$

RESTRICTED MODEL (MODEL 3 :NO CHANGE IN DEPENDANT VARIABLE):

$$H_0: \beta_5 = 0 \qquad H_1: \beta_5 > 0$$

$\widehat{\log(ME)} = \beta_0 + \beta_1 age + \beta_2(age^2/100) + \beta_3\ eduyrs + \beta_4 eduyrs^2 + \beta_5 \log(wrkhr) + U_i$

$\widehat{\log(ME)} = 5.43966 + 0.07806age - 0.08109\ (age^2/100) + 0.05735eduyrs + 0.00436eduyrs^2$

RESTRICTED MODEL (MODEL 4_1: CHANGE IN DEPENDANT VARIABLE):

$$H_0: \beta_5 = 1 \qquad H_1: \beta_5 \neq 1$$

$\log(ME) = \beta_0 + \beta_1 age + \beta_2(age^2/100) + \beta_3\ eduyrs + \beta_4 eduyrs^2 + 1 * \log(wrkhr) + U_i$

$\log(ME) - \log(wrkhr) = \beta_0 + \beta_1 age + \beta_2(age^2/100) + \beta_3\ eduyrs + \beta_4 eduyrs^2 + U_i$

$\log\widehat{\frac{ME}{wrkhrs}} = \beta_0 + \beta_1 age + \beta_2(age^2/100) + \beta_3\ eduyrs + \beta_4 eduyrs^2 + U_i$

$\log\widehat{\frac{ME}{wrkhrs}} = -0.731529 + 0.026852age - 0.021533age^2 + 0.013715eduyrs + 0.004945eduyrs^2$

Table 2

		Model4(UN)	Model3(R)	Model4_1(R)
TSS		**78394**	**78394**	**42803**
RSS		**33794**	**60650**	**34299**
R^2		**0.5689**	**0.2264**	**0.1987**
$\bar{R}^2$		**0.5689**	**0.2263**	**0.1986**
F	RSS		**37804.0540**	**710.8671**
	R^2		**37791.7815**	**-**
$F_{tab\ (0.01,\ 47569)}$			**6.635429**	**6.635429**
P Value (2 sided)			**4.4e-16**	**4.4e-16**

(a) <u>Comment on TSS:</u>

The TSS for the unrestricted model (model 4) and the restricted model (model 3) with no change in the dependant variable have the same TSS values. The dependent variable, log(ME) is the same across the above two specified models. TSS for the above models is calculated as follows:

$$\text{TSS} = \{log(ME) - \overline{log(ME)}\}^2$$

Whereas, in the restricted model (model 4_1) with the change in the dependant variable to $\widehat{\log\frac{\text{ME}}{\text{wrkhrs}}}$, states that the TSS is now:

$$\text{TSS} = \{\log\frac{\text{ME}}{\text{wrkhrs}} - \overline{\log\frac{ME}{wrkhrs}}\}^2$$

By this we can state that the change in TSS in this model is due to the change in dependant variable unlike the unchanged dependant variable in the previous models 3 and 4

<u>Comment on R^2 and $\bar{R}^2$ between models</u>:

The R^2 and $\bar{R}^2$ values from our model have not much of a difference, apart from the fact that the $\bar{R}^2$ value is used for the restrictions in the model made to account for the F-statistic.

We should be having $\mathbf{R^2 \geq \bar{R}^2}$ for further proceedings, as the no. of independent variables increases there will be more explanation on the dependant variable, giving a higher R^2 value compared to the restricted model ($\bar{R}^2$ will be lower)

For testing the model for statistical significance, we use the formula:

$$F_{cal} = \frac{R^2_{UR} - R^2_R / q}{1 - R^2_{UR} / (n-k-1)_{UR}}$$

We use this formula for model 3 (restricted) and model 4 (unrestricted) for determining the statistical significance as they have the same dependant variable but we can't compare model 4 (unrestricted) and model 4_1 (restricted) as their dependant variable has changed and so we use the RSS formula for calculating the F-statistic as the restricted model is now expressed in a different explanation with the dependant variable.

(b) The F-stat calculated for model 4 (unrestricted model) and model 3 (restricted model) using both the R^2 and RSS formula yields the same result and is ultimately due to the fact that the result of the null hypothesis for checking the statistical significance for both the models are same: The F-stat calculated for the model 4 (unrestricted model) and model 3 (restricted model) using both the R^2 and RSS formula yields the same result is ultimately due to the similarity in the dependant variable and also to the fact that the null hypothesis for checking the statistical significance for both the models are same:

$$\mathbf{H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0}$$

Whereas, in model 4 (unrestricted model) and model4_1 (restricted model), as there is a mismatch between the dependant variable, which is log(ME) for model 4 and $\log\frac{\text{ME}}{\text{wrkhrs}}$ for model 4_1, for calculating the F-statistic, R^2 formula cannot be used for estimating, instead we have to use the RSS formula for testing the significance of the model. The null hypothesis of model 4_1 has now changed to:

$$\mathbf{H_0: \beta_2 = \beta_3 = \beta_4 = 0}$$

Based on the above information, we can conclude that the difference between the R^2 and RSS formula in estimating for the F-statistic is **due to the change in dependant variable** and the **usage of RSS formula would be more appropriate** for estimating the F-statistic value.

$$F_{cal} = \frac{R^2_{UR} - R^2_R / q}{1 - R^2_{UR} / (n-k-1)_{UR}} \qquad F_{cal} = \frac{RSS_R - RSS_{UR} / q}{RSS_{UR} / (n-k-1)_{UR}}$$

(c) Based on the results from table 2, we can conclude that **both the models are statistically significant at 1% level of significance**, given the null and alternative hypothesis:

Model 4 and Model 3 (RSS):

$$\mathbf{F_{cal} > F_{tab}}$$

$$\mathbf{37804.0540 > 6.635429}$$

Therefore, we reject the null hypothesis stating that, the coefficient estimate $\beta_5 > 0$ of log(wrkhr)

Model 4 and Model 4_1 (RSS):

$$\mathbf{F_{cal} > F_{tab}}$$

$$\mathbf{710.8671 > 6.635429}$$

Therefore, we reject the null hypothesis stating that, the coefficient estimate $\beta_5 \neq 1$ of log(wrkhr)

5. **Model with quadratics: Refer to the text in pages 192-197 to answer this question. Use the model specified in Wooldridge C.6.2 in page 219 (question starts in p.218) but use *age* and *age-squared* instead of *exper* and *exper-squared*. Question: Answer the questions as in (iii) and (iv) of the book's question. Show clearly the derivations to arrive at the results mentioned in these two parts of the question.**

(iii) Using the approximation

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

(iv) At what value of *exper* does additional experience actually lower predicted log(*wage*)? How many people have more experience in this sample?

We write the estimated equation as:

$$\widehat{\log(\text{wage})} = \hat{\beta}_0 + \hat{\beta}_1 \text{educ} + \hat{\beta}_2 \text{age} + \hat{\beta}_3 \text{age}^2 + \text{u}$$

$$\widehat{\log(\text{wage})} = 7.7096618 + 0.1162840\text{educ} + 0.0837502\text{age} - 0.0008648\text{age}^2$$

Then we have the approximation

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 age)\Delta age$$

$$\%\Delta\widehat{wage} \approx 100(0.0837502 + 2(-0.0008648)*age)\Delta age$$

$$= (8.37502 + (-0.17296)*age)\Delta age$$

The approximate return to the fifth year of experience:

$$= (8.37502 + (-0.17296)*5)$$

$$= \mathbf{7.51022\%}$$

The approximate return to the twentieth year of experience:

$$= (8.37502 + (-0.17296)*20)$$

$$= \mathbf{4.91582\%}$$

To find the value of experience where additional age lowers the predicted log(wage), is when the slope of the model is 0. Slope: $\hat{\beta}_2 + 2\hat{\beta}_3 \text{age} = 0$

$0.0837502 + 2 * 0.0008648\text{age} = 0$

$0.0837502 + 0.0017296\text{age} = 0$

Age = $\left|\frac{0.0837502}{0.0017296}\right|$ = **48.42**

Therefore, after the age of 48.42, the log(wage) tends to fall for an additional year of age for the individual.

6. **Question: Answer question C6.3 except part (iv). Refer to section "Models with interaction terms" in 197-199 along with the example 6.3 to answer this question.**

C6.3 Consider a model where the return to education depends upon the amount of work experience (and vice versa):

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{educ} \cdot \text{exper} + u.$$

(i) Show that the return to another year of education (in decimal form), holding *exper* fixed, is $\beta_1 + \beta_3 exper$.

(ii) State the null hypothesis that the return to education does not depend on the level of exper. What do you think is the appropriate alternative?

(iii) Use the data to test the null hypothesis in (ii) against your stated alternative.

(i) Wage equation:

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{educ} \cdot \text{exper} + u$$

To find the return to another year of education, we differentiate the wage equation with respect to education:

$$\frac{\Delta \log(wage)}{\Delta \text{educ}} = \beta_1 + \beta_3 \text{exper}$$

This expression represents the additional return to education, accounting for the interaction with experience.

(ii) the null hypothesis that the return to education does not depend on the level of exper and the appropriate alternate hypothesis that the return to education does depend on the level of experience can be given as follows:

H_0: $\beta_3 = 0$
H_1: $\beta_3 \neq 0$

(iii) $T_{cal} = \frac{0.002007}{0.00008087} =$ **24.821**

$T_{tab} =$ **1.644886**

$T_{cal} > T_{tab}$ therefore we reject H_0