Statistics and Econometrics: Problem Set 4

Siow Meng Low

4 November 2016

Question 1

Question 1a: Return of Another Year of Education

The equation is given as:

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

It can be rewritten as:

$$log(wage) = \beta_0 + (\beta_1 + \beta_3 exper) educ + \beta_2 exper + u$$

Holding *exper* and u fixed, the return of another year of education is $(\beta_1 + \beta_3 exper)$, in decimal form. In percentage form, it is $100 \cdot (\beta_1 + \beta_3 exper)\%$

Question 1b: Null and Alternative Hypothesis

The null hypothesis states that the return to education does not depend on the level of *exper*. This can be expressed as: H_0: $\beta_3 = 0$

For the alternative hypothesis, we cannot be sure if education and experience interact positively or negatively. The return of education may increase or decrease with higher level of experience. Therefore, we should use the two-sided alternative: H_1: $\beta_3 \neq 0$

Question 1c: Hypothesis Testing

Table 1 displays the model information of regressing monthly wage (in log form) on education, experience, and the interaction term between education and experience.

Table 1: Regression Model for Question 1c

	Log(Wage)
educ	0.0440**
	(0.0174)
exper	−0.0215
	(0.0200)
educ:exper	0.0032**
	(0.0015)
Constant	5.9495***
	(0.2408)
N	935
R^2	0.1349
Adjusted R^2	0.1321
Residual Std. Error	0.3923 (df = 931)
F Statistic	48.4069*** (df = 3; 931)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

The null hypothesis and alternative hypothesis are:

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

From the above, it is clear that we would do an two-tail test. The critical value at 5% significance level is thus 1.96 (assume normal distribution since we have large degree of freedom, > 120).

The t-statistic is $\dfrac{\hat{\beta}_3}{se(\hat{\beta}_3)} = 2.0946$. Since this value is larger than the critical value 1.96, we reject the null hypothesis and conclude that the return to education depends on the level of *exper* (i.e. $\beta_3 \neq 0$).

Question 1d: Expected Wage Prediction of an Average Person

The expected $\widehat{log(wage)}$ for an average person (with $educ = 12$ and $exper = 10$) is predicted to be 6.647

To get the predicted expected wage for an average person, we will need to multiply $exp(\widehat{log(wage)}) = 770.82$ by $E(exp(u))$. Without assuming u to be normally distributed, we can obtain the sample estimate of $E(exp(u))$ using $n^{-1}\Sigma_{i=1}^{n} exp(\hat{u}_i) = 1.077$

Therefore, the predicted expected wage, $\widehat{wage}$ for an average person (with $educ = 12$ and $exper = 10$) is predicted to be 829.9

Question 2

Question 2a: OLS Using Quadratic Term

The equation is given as:

$$log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

Table 2 displays the model information of regressing birth weight (in log form) on the number of prenatal visits (both linear term and quadratic term).

Table 2: Regression Model for Question 2a

	Log(Birth Weight)
npvis	0.0189^{***}
	(0.0037)
npvissq	-0.0004^{***}
	(0.0001)
Constant	7.9579^{***}
	(0.0273)
N	1,764
R^2	0.0213
Adjusted R^2	0.0201
Residual Std. Error	0.2031 (df = 1761)
F Statistic	19.1202^{***} (df = 2; 1761)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

From the above table, although the coefficient of the quadratic term is small, it is very significant due to its small standard error. The p-Value, for $H_0 : \beta_{npvis^2} = 0$, is extremely small: 0.00036. As a result, we can see that the quadratic term is statistically significant even at 1% significance level.

Question 2b: Number of Prenatal Visits that Maximizes log(*bwght*)

To calculate the number of prenatal visits that maximizes log(*bwght*), we use the first order derivative of $log(bwght)$ with respect to *npvis*. Assuming the error term u is uncorrelated with *npvis*, the derivative is: $\frac{\partial(log(bwght))}{\partial(npvis)} = \beta_1 + 2\beta_2 npvis$

Next, set the derivative to zero and we have:

$$\beta_1 + 2\beta_2 npvis = 0$$

$$npvis = \frac{-\beta_1}{2\beta_2}$$

Using the estimated coefficients in Table 2, the estimated value of *npvis*, which maximizes log(*bwght*), is $\frac{-0.0189}{2(-0.000429)} = 22.06$

It is shown that the number of prenatal visits that maximizes log(*bwght*) is estimated to be about 22. Note that there are 68 records with *npvis* = *NA* in the dataset. After removing these 68 records, there are 21 women who had at least 22 prenatal visits in the given sample.

Question 2c: Decline of Birth Weight after 22 Prenatal Visits

Yes, it makes sense that birth weight is predicted to decline after 22 prenatal visits.

While more prenatal visits may indicate better prenatal care (and hence healthier babies with higher birth weights), a very high number of prenatal visits (in our data, more than 22 visits) may signal other pregnancy issues (e.g. health issues with mother or baby) that requires doctors' assistance. This could result in lower birth weights.

Question 2d: Effects of Mother's Age

The new equation is:

$$log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + \beta_3 mage + \beta_4 mage^2 + u$$

The right column of Table 3 displays the model information of regressing birth weight (in log form) on the number of prenatal visits (both linear term and quadratic term) and the mother's age (both linear term and quadratic term).

Table 3: Regression Model for Question 2d

	Log(Birth Weight)	
	Question 2a Model	Question 2d Model
	(1)	(2)
npvis	0.0189***	0.0180***
	(0.0037)	(0.0037)
npvissq	−0.0004***	−0.0004***
	(0.0001)	(0.0001)
mage		0.0254***
		(0.0093)
magesq		−0.0004***
		(0.0002)
Constant	7.9579***	7.5837***
	(0.0273)	(0.1371)
N	1,764	1,764
R^2	0.0213	0.0256
Adjusted R^2	0.0201	0.0234
Residual Std. Error	0.2031 (df = 1761)	0.2027 (df = 1759)
F Statistic	19.1202*** (df = 2; 1761)	11.5640*** (df = 4; 1759)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

To calculate the mother's age that maximizes log(*bwght*), we use the first order derivative of $log(bwght)$ with respect to *mage*. Since *npvis* is held fixed, we can treat it as constant while deriving the derivative. Assuming the error term u is uncorrelated with *mage*, the derivative is thus: $\frac{\partial(log(bwght))}{\partial(mage)} = \beta_3 + 2\beta_4 mage$

Next, set the derivative to zero and we have:

$$\beta_3 + 2\beta_4 mage = 0$$

$$mage = \frac{-\beta_3}{2\beta_4}$$

Using the estimated coefficients in Table 3, the value of *mage*, which maximizes $\log(bwght)$, is $\frac{-0.0254}{2(-0.000412)} = 30.83$

From the above, the mother's age that maximizes $\log(bwght)$, with *npvis* held fixed, is estimated to be about 31 years old.

Note that there are 68 records with $npvis = NA$ in the given dataset and these 68 records are omitted while performing linear regression shown in Table 3. Excluding these 68 observations, the total number of samples is 1764, and the number of mothers with age:

- 31 Years Old and Above: 720 women, the fraction is 40.82%
- 32 Years Old and Above: 584 women, the fraction is 33.11%

Note that the numbers shown in above bulleted points exclude the 68 observations with $npvis = NA$.

Question 2e: Variation in log(*bwght*)

From Table 3, the model (with regressors: *npvis*, *npvissq*, *mage* and *magesq*) has a R^2 value of 0.0256 and Adjusted R^2 value of 0.0234

This means that the model only explained around 2.56% of the variations in $\log(bwght)$. Obviously this is a very low value, and I would say that mother's age and number of prenatal visits explain very little of the variation in $\log(bwght)$. We could consider adding in more regressors for a better fit.