HOME ASSIGNMENT 3

DANIIL BUCHKO

Question (1)**.** :

(a) Comment on the effect of *profmarg* on CEO salary
(b) Does market value have a significant effect? Explain.
(c) Interpret the coefficients on *ceoten* and *comten*. Are the variables statistically significant? What do you make of the fact that longer tenure with the company, holding the other factors fixed, is associated with a lower salary?

Solution. In the reasoning that follows we will calculate *t-stat* by the following formula:

$$t_{\hat{\beta}_k} = \frac{\hat{\beta}_k}{\sigma(\hat{\beta}_k)}$$

(a) Since *profmarg* is not statistically significant (t-statistics ~ 1 is much less than 1.96) we cant really use it to describe the connections inside the data.
(b) Since t-statistics is higher than 1.96 we can point out that market value has significant positive effect. We can explain coefficient 0.1 as follows: if we increase market vakue by 1% salary statistically increases on average by $\sim 10\%$
(c) *ceoten* and *comten* have separately big t-stats which therefore implies that they are significant. We can see that coefficient before *comten* is negative and therefore longer tenure with the company, holding the other factors fixed, is associated with a lower salary. This could be explained by the lack of higher orders of polynomials of *comten*. Maybe in reality, there is some point in time after which earnings begin to increase which could have been demostrated by higher orders of *comten*.

□

Question (2)**.** :

(a) Suppose you get an OLS estimate for β_1 and corresponding t-statistics. Also you get an OLS estimate for γ_1 and corresponding t-statistics from the following regression:

$$y = \gamma_0 + \gamma_1 \tilde{x} + v$$

where $\tilde{x}$ is demeaned value of x: $\tilde{x} = x - \bar{x}$ and $\bar{x}$ is the sample mean. Which one of two t-statistics is larger?

(b) Suppose that you have performed OLS estimation and obtained the following estimate of the conditional variance-covariance matrix for the parameter estimates:

$$Var\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}2 & 1\\ 1 & 3\end{pmatrix}$$

What is the value of $\hat{\sigma}_u^2$?

(c) Under the condition of part (b), for which value of α does the random variable $\hat{\beta}_0 + \alpha\hat{\beta}_1$ have minimal variance?

Solution. :

(a) t-statistics by definition is

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\sigma(\hat{\beta}_1)}$$

And estimators are represented as follows

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{\gamma}_1 = \frac{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})(y_i - \bar{y})}{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})^2}$$

Now lets notice that

$$\bar{\tilde{x}} = \frac{1}{n}\sum_{i=1}^n \tilde{x}_i = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Thus substitution of $\tilde{x}_i$ yields:

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\beta}_1$$

Now since estimators are equal, their standard deviations are also equal and so are *t-stats*.

(b) We will solve the problem by writing down the definitions of components of the covariance matrix, namely for $\mathrm{Var}(\hat{\beta}_1)$ and $\mathrm{Var}(\hat{\beta}_0)$ and $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\\ &= \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = 3\end{aligned}$$

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y} - \hat{\beta}_1\bar{x}) = \underbrace{\mathrm{Var}(\bar{y})}_{?} + (\bar{x})^2\mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\,\underbrace{\mathrm{Cov}(\bar{y}, \hat{\beta}_1)}_{?}$$

Lets begin by establishing that $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$:

$$\begin{aligned}
\text{Cov}(\bar{y}, \hat{\beta}_1) &= \text{Cov}\left[\frac{1}{n}\sum_{i=1}^{n} y_i,\ \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\text{Cov}\left[\sum_{i=1}^{n} y_i,\ \sum_{i=1}^{n}(x_i - \bar{x})y_i\right] \\
&= \frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})\sum_{j=1}^{n}\text{Cov}\left[y_i, y_j\right] \\
&= \frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0}\sigma_u^2 = 0
\end{aligned}$$

Now lets find $\text{Var}(\bar{y})$:

$$\begin{aligned}
\text{Var}(\bar{y}) = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}(y_i, y_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Cov}(y_i, y_i) = \frac{\sigma_u^2}{n}
\end{aligned}$$

And finally getting the result:

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma_u^2}{n} + 3(\bar{x})^2 = 2$$

Now lets proceed with **covariance**:

$$\begin{aligned}
\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}\left[\bar{y} - \hat{\beta}_1\bar{x},\ \hat{\beta}_1\right] &= \underbrace{\text{Cov}\left[\bar{y}, \hat{\beta}_1\right]}_{=0} - \bar{x}\text{Var}(\hat{\beta}_1) \\
&= -3\bar{x} = 1 \implies \bar{x} = -\frac{1}{3}
\end{aligned}$$

and by inserting this value back to $\text{Var}(\hat{\beta}_0)$ we get:

$$2 = \frac{\sigma_u^2}{102} + \frac{1}{3} \implies \sigma_u^2 = 170$$

(c)

$$\begin{aligned}
\text{Var}(\hat{\beta}_0 + \alpha\hat{\beta}_1) &= \text{Var}(\hat{\beta}_0) + 2\alpha\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \alpha^2\text{Var}(\hat{\beta}_1) \\
&= 2 + 2\alpha + 3\alpha^2
\end{aligned}$$

which is minimal at point $\alpha = -1/3$

□

Question (3)**.** Refer to Problem 5 in Problem Set 2. Now, use the log of the housing price as the dependent variable:

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

(a) You are interested in estimating and obtaining a confidence interval for the percentage change in price when a 150-square-foot bedroom is added to a house. In decimal form, this is $\theta_1 = 150\beta_1 + \beta_2$. Use the data in *hprice1.csv* to estimate θ_1.
(b) Write β_2 in terms of θ_1 and β_1 and plug this into the $\ln(price)$ equation.
(c) Use part (ii) to obtain a standard error for $\hat{\theta}_1$ and use this standard error to construct a 95% confidence interval.

Solution. :

(a) After doing some transformations and estimating model below we get that $\hat{\theta} = 0.0858$
(b)

$$\log(price) = \beta_0 + \beta_1 sqrft + (\theta_1 - 150\beta_1)bdrms + u$$
$$\log(price) = \beta_0 + \beta_1(sqrft - 150bdrms) + \theta_1 bdrms + u$$

(c) Estimated standard error (using robust errors) is $\hat{\sigma}(\hat{\theta}_1) = 0.027742$ and confidence intervals are

TABLE 1. confidence interval for θ_1

	2.5 %	97.5 %
bdrms	0.033	0.139

□

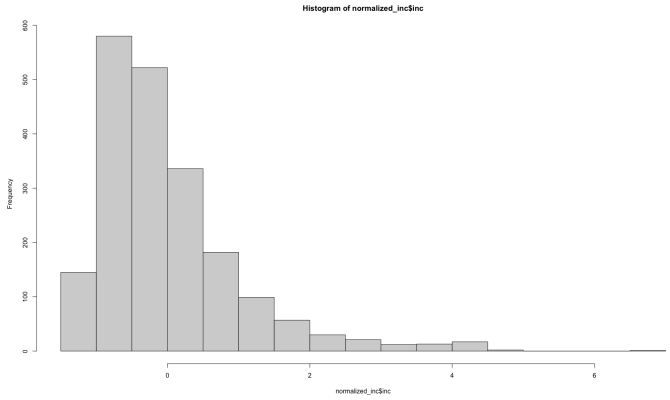
Question (4)**.** :

(a) First use the data set *401ksubs.csv*, keeping only observations with $fsize = 1$. Find the skewness measure for *inc*. Do the same for $\ln(inc)$. Which variable has more skewness and therefore seems less likely to be normally distributed?
(b) Next use *bwght2.csv*. Find the skewness measures for *bwght* and $\log(bwght)$. What do you conclude?
(c) Evaluate the following statement: "The logarithmic transformation always makes a positive variable look more normally distributed."

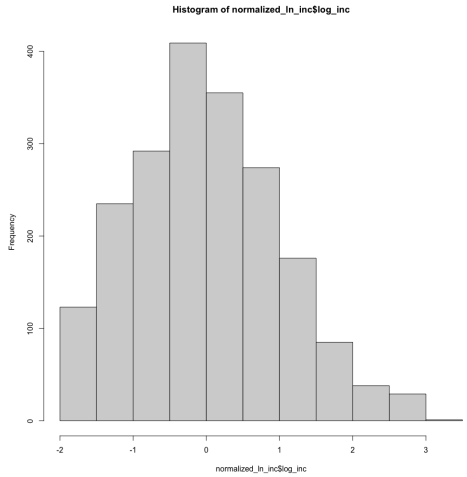
(d) If we are interested in the normality assumption in the context of regression, should we be evaluating the unconditional distributions of y and $\log(y)$? Explain.

Solution. :

(a) By looking at normalized values of *inc* we can notice that it is probably not normally distributed and skewness measure will probably differ from zero. Skewness measure value is 1.862259.



Then we take logarithm of *inc* and get the histogram as below. Its skewness measure is 0.3606031.



(b) Non-logarithmic data has skewness of -0.6001563 and logarithmic has -2.948616

(c) We have just seen that skewness measure of log transformed data was bigger than non-log transformed. Moreover log transformation tends to make left-skewed distributions even more left-skewed, despite being positive.

□

Question (5)**.** In this problem set, you continue your analysis of the returns to education and the gender gap in earnings. The data set for this analysis `cps99_ps3.csv` is the full sample of full-time workers from the March 1999 CPS. Estimate a regression of $\ln(ahe)$ on *yrseduc*, *age*, *female*, and *hsdipl*.

(a) Explain the meaning of the coefficients on *yrseduc*, *age*, and *female*.
(b) Explain the meaning of the root mean squared error in this regression.
(c) Using this regression, provide a 95% confidence interval for the gender gap in earnings (controlling for *age*, *years* of *education*, and obtaining a high school diploma).
(d) What is the estimated marginal value of 12 years of education culminating in a HS diploma, relative to 12 years of education and no HS diploma? Test (at the 5% significance level) the hypothesis that this marginal value is zero, against the hypothesis that it is not. Is this marginal value large or small in a real-world sense? Explain.
(e) Let V denote the marginal value of 12 years of education plus a high school diploma, relative to the value of having only 10 years of education, holding constant age and gender. That is, $V = 0.2$ means that high school graduates on average earn 20% more than someone of the same age and gender but with only 10 years of education. Using the regression, provide an estimate of V . Is your estimate large in a real-world sense?
(f) Construct a 95% confidence interval for V.

Solution. :

(a) By looking at the table below we can see that: *yrseduc* is significantly positive meaning that earnings on average increases with each additional year by 8%. Similar interpretation has *age* variable: earnings increases with each additional age by 0.7%. On contrast, coefficient before *female* is significantly negative, meaning that on average females had lower earnings than men by 24%.
(b) $RMSE = 0.468947$ meaning that predicted earnings were on average 1.59 times as higher (both directions).
(c) Gender gap lies between -0.2536607 and -0.2345203.
(d) Estimated model is once again:

$$\ln(ahe) = \hat{\beta}_0 + \hat{\beta}_1 yrseduc + \hat{\beta}_2 age + \hat{\beta}_3 female + \hat{\beta}_4 hsdipl$$

TABLE 2. Results of OLS estimation

	Dependent variable:
	log(ahe)
yrseduc	0.081^{***}
	(0.001)
age	0.007^{***}
	(0.0002)
female	-0.244^{***}
	(0.005)
hsdipl	0.127^{***}
	(0.010)
Constant	1.183^{***}
	(0.017)
Note:	*p<0.1; **p<0.05; ***p<0.01

Marginal effect of 12 years of education with diploma relative to just 12 years of education is the the following:

$$\frac{12\hat{\beta}_1 + \hat{\beta}_4}{12\hat{\beta}_1} = \frac{1.099}{0.972} = 1.13$$

Marginal effect is zero when the following constraint holds:

$$\underbrace{12\beta_1 + \beta_4}_{=\theta} = 0$$

Lets rewrite coefficients as in task 3: $\beta_4 = \theta - 12\beta_1$ and estimate slightly different model:

$$\begin{aligned}\ln(ahe) = \beta_0 + \beta_1 yrseduc + \beta_2 age + \\ + \beta_3 female + (\theta - 12\beta_1) hsdipl + u\end{aligned}$$

$$\begin{aligned}\ln(ahe) = \beta_0 + \beta_1 (yrseduc - 12hsdipl) + \beta_2 age + \\ + \beta_3 female + \theta hsdipl + u\end{aligned}$$

And the p-value for $\hat{\theta}$ is 0, which means that it is significant from zero and thus marginal effect is statistically different from zero.

Interpretation: percentage increase of earnings by 13% higher for those who obtained diploma versus those who didnt. This is not a huge difference in the real-world scenario.

(e) To solve this lets estimate the following constraint:

$$\theta = 2\beta_1 + \beta_4 \implies \beta_4 = \theta - 2\beta_1$$

Once again the model is then:

$$\begin{aligned}\ln(ahe) = \beta_0 + \beta_1 yrseduc + \beta_2 age + \\ + \beta_3 female + (\theta - 2\beta_1) hsdipl + u\end{aligned}$$

$$\begin{aligned}\ln(ahe) = \beta_0 + \beta_1 (yrseduc - 2hsdipl) + \beta_2 age + \\ + \beta_3 female + \theta hsdipl + u\end{aligned}$$

After ols estimation we got that $\hat{\theta} = 0.29$ is positive and significant from zero. Thus the 2 additional years of education and degree brings you by 29% more earnings. Difference is not big in real-world scenario.

(f) Interval for V is: $(0.272679,\ 0.3070001)$

□

Question (6)**.** Estimate a regression of ln(*ahe*) on *yrseduc*, *age*, *age2*, *female*, and *hsdipl*.

(a) What are the signs of the coefficients on age and on age^2? In qualitative terms, what does this tell you about the relationship between earnings and age?

(b) Test (at the 5% significance level) the hypothesis that the relation between ln(*ahe*) and *age* (controlling for *yrseduc*, *female*, and *hsdipl*) is linear against the alternative hypothesis that it is quadratic.

(c) Does including the term age^2 change the estimated gender gap, in a real-world way? What is the econometric reason for this change (or lack of change)?

(d) Test (at the 5% significance level) the hypothesis that the relation between ln(*ahe*) and *age* (controlling for *yrseduc*, *female*, and *hsdipl*) is quadratic, against the alternative hypothesis that it is cubic. (You will need to specify and estimate another regression.)

(e) Test (at the 5% significance level) the hypothesis that the relation between ln(*ahe*) and *age* (controlling for *yrseduc*, *female*, and *hsdipl*) is linear, against the alternative hypothesis that it is a polynomial of up to cubic degree. Based on the results for parts (b) – (d), which of the specifications would you recommend using: linear in age, quadratic in age, or cubic in age?

Solution. :

(a) Since both coefs before *age* are significant they have the following interpretation: earnings increases with the age only until reaching some maximum value of increase, after which they tend to decrease.

TABLE 3. OLS regression for task 6

	Dependent variable:
	ln(ahe)
yrseduc	0.081^{***}
	(0.001)
age	0.045^{***}
	(0.002)
age2	-0.0004^{***}
	(0.00002)
female	-0.245^{***}
	(0.005)
hsdipl	0.122^{***}
	(0.010)
Constant	0.441^{***}
	(0.043)
Note:	*p<0.1; **p<0.05; ***p<0.01

(b) By using linear hypothesis test we calculate F-stats of 329.26 which implies that model is of quadratic form.

(c) Adding squared *age* changes coefficient before *female* from -0.24409 to -0.24473. Econometric reason: female is positively correlated with quadratic age and thus adding omitted variable influences the coefficient in this way.

(d) We need to estimate the following model:

$$\ln(ahe) = \beta_0 + \beta_1 yrseduc + \beta_2 age + \\ + \beta_3 female + \beta_4 hsdipl + \beta_5 age^2 + \beta_6 age^3 + u$$

and test constraint for $\hat{\beta}_6 = 0$ using F statistics. Result: $F = 18.761$ and p-value is 0, which is in favour of cubic model.

(e) Estimating model:

$$\ln(ahe) = \beta_0 + \beta_1 yrseduc + \beta_2 age + \\ + \beta_3 female + \beta_4 hsdipl + \beta_5 age^2 + \beta_6 age^3 + u$$

and test constraint for $\hat{\beta}_6 = 0$, $\hat{\beta}_5 = 0$ using F statistics. Result: $F = 182.37$ and p-value is 0. I would definetly recommend using cubic model since tests show that it is right specification and moreover RMSE is lower for cubic model.

□

Question (7)**.** The entries in the following table are the predicted marginal percentage increases in average hourly earnings, as a worker gains an additional year of age, evaluated for different ages and based on three different specifications; all the estimates control for years of education, gender, and whether the worker has a HS diploma. For example, the top left entry in the table is the predicted difference in earnings, in percentage terms, between a worker aged 31 and one aged 30, controlling for years of education, gender, and a HS diploma.

(a) Fill in the table
(b) Do the entries differ in a real-world sense going from the linear to the quadratic specification? From the quadratic to the cubic?
(c) Which row or rows of entries make the most sense to you, based on economic reasoning? Explain.

Solution. :

Linear model:

$$\ln y = \hat{\beta}_0 + \hat{\beta}_1 x \implies \Delta y = 100\% \hat{\beta}_1 \Delta x$$

thus lets estimate the following model:

$$\ln(ahe) = \beta_0 + \beta_1 age + \beta_4 yrseduc + \\ + \beta_5 female + \beta_6 hsdipl + u$$

Marginal effect of age on earnings doesnt rely on age and remains constant at differnt data points:

$$\text{ME}(age = 30) = \text{ME}(age = 45) = \text{ME}(age = 60) = 100\%\hat{\beta}_1 = 0.7\%$$

Quadratic model:

$$\ln y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \implies \Delta y = 100\%(\hat{\beta}_1 \Delta x + \hat{\beta}_2 (x_1^2 - x_0^2))$$

thus lets estimate the following model:

$$\ln(ahe) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_4 yrseduc + \\ + \beta_5 female + \beta_6 hsdipl + u$$

$$\text{ME}(age = 30) = 1.8\%$$
$$\text{ME}(age = 45) = 0.4\%$$
$$\text{ME}(age = 60) = -0.8\%$$

Cubic model:

$$\ln y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 \implies \Delta y = 100\%(\hat{\beta}_1 \Delta x + \hat{\beta}_2 (x_1^2 - x_0^2) + \hat{\beta}_3 (x_1^3 - x_0^3))$$

thus lets estimate the following model:

$$\ln(ahe) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 yrseduc + \\ + \beta_5 female + \beta_6 hsdipl + u$$

$$\text{ME}(age = 30) = 2\%$$
$$\text{ME}(age = 45) = 0.2\%$$
$$\text{ME}(age = 60) = -0.1\%$$

Entries differ from linear to quadratic and from quadratic to cubic model. We can see that linear model predicts that each additional age happens to correlate with 0.7% earnings growth, wheras quadratic and cubic models demostrate bigger values of earnings growth in the early age and lower in the late age. This is perfectly ok for real-world scenario, since after 60 people tend to go on pension and thus earn less on average. Quadratic form seems most appropriate for me since in Russia pensioners receives extremely less when they retire. □