

## Tutorial 8

**keywords:** binary variables, dummy variables, intercept, slope, conditional expectation, regression line, F-test, prediction intervals, prediction uncertainty, estimation uncertainty, variation in error, sum of squared residuals, Chow test, confidence intervals, standard errors

**estimated reading time:** 36 minutes

Quang Bui

September 11, 2018

# Question 1

## Logarithmic and quadratic model

EViews workfile: *wage1tute4.wf1* from Tutorial 4

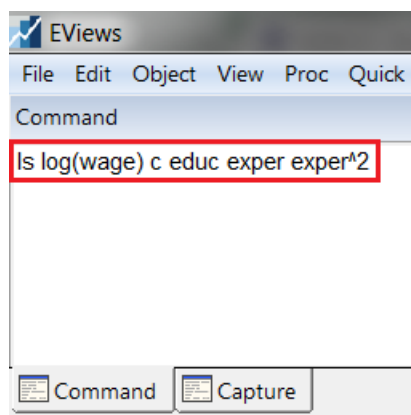
(a) Use OLS to estimate the equation

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + u$$

and report the results using the usual format. Name this **eq01**. [Note that in most other statistical software, you have to generate  $\log(\text{wage})$  and  $\text{exper}^2$  first, give them names like *lwage* and *expersq*, then use them in the regression command. EViews allows you to do this in the regression command, which is a great advantage.]

To estimate this model from the **Command window** in EViews,

$$\log(\text{wage}) \text{ c educ exper exper}^2$$



(Press Enter to execute code)

To name this equation **eq01**,

*Name* → *Name to identify object* : eq01

Equation: UNTITLED    Workfile: WAGE1TUTE4::Wa...

View   Proc   Object   Print   **Name**   Freeze   Estimate   Forecast   Stats   Resids

Dependent Variable: LOG(WAGE)  
 Method: Least Squares  
 Date: 04/23/18   Time: 11:11  
 Sample: 1 526  
 Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.127998	0.105932	1.208296	0.2275
EDUC	0.090366	0.007468	12.10041	0.0000
EXPER	0.041009	0.005197	7.891606	0.0000
EXPER^2	-0.000714	0.000116	-6.163888	0.0000

R-squared	0.300273	Mean dependent var	1.623268
Adjusted R-squared	0.296251	S.D. dependent var	0.531538
S.E. of regression	0.445906	Akaike info criterion	1.230158
Sum squared resid	103.7904	Schwarz criterion	1.262594
Log likelihood	-319.5316	Hannan-Quinn criter.	1.242858
F-statistic	74.66829	Durbin-Watson stat	1.785009
Prob(F-statistic)	0.000000		

Object Name

Name to identify object

eq01    24 characters maximum,  
16 or fewer recommended

Display name for labeling tables and graphs (optional)

OK    Cancel

Dependent Variable: LOG(WAGE)

Method: Least Squares

Sample: 1 526

Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.127998	0.105932	1.208296	0.2275
EDUC	0.090366	0.007468	12.10041	0.0000
EXPER	0.041009	0.005197	7.891606	0.0000
EXPER^2	-0.000714	0.000116	-6.163888	0.0000
R-squared	0.300273	Mean dependent var	1.623268	
Adjusted R-squared	0.296251	S.D. dependent var	0.531538	
S.E. of regression	0.445906	Akaike info criterion	1.230158	
Sum squared resid	103.7904	Schwarz criterion	1.262594	
Log likelihood	-319.5316	Hannan-Quinn criter.	1.242858	
F-statistic	74.66829	Durbin-Watson stat	1.785009	
Prob(F-statistic)	0.000000			

$$\widehat{\log(wage)} = 0.1280 + 0.0904educ + 0.0410exper - 0.0007exper^2$$

(0.1059)      (0.0075)      (0.0052)      (0.0001)

(b) Is  $exper^2$  statistically significant at the 1% level?

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)} = \frac{\hat{\beta}_3 - 0}{se(\hat{\beta}_3)} = \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} \sim t_{n-k-1} \quad \text{under } H_0$$

$$\text{degrees of freedom : } n - k - 1 = 526 - 3 - 1 = 522$$

$$t_{calc} = \frac{\hat{\beta}_4}{se(\hat{\beta}_4)} = -6.1639$$

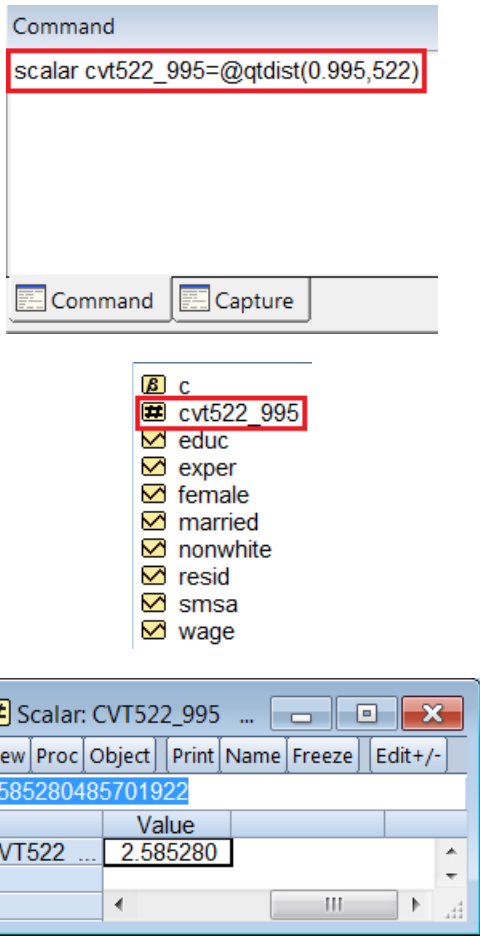
$$p\text{-value} = 0.0000$$

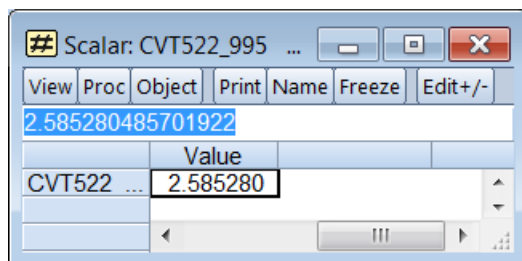
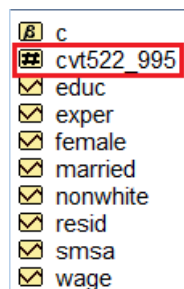
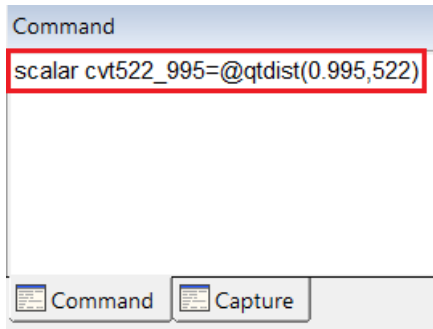
Testing at the 1% significance level,

$$\alpha = 0.01$$

$$+t_{crit} = t_{0.995,522} =$$

$$-t_{crit} = t_{0.005,522} =$$

To obtain  $+t_{crit}$  in EViews, 



By comparing  $t_{calc}$  with  $\pm t_{crit}$ , we reject  $H_0$  if,

$$t_{calc} > t_{crit}$$

OR

$$t_{calc} < -t_{crit}$$

Alternatively, we can compare the p-value with  $\alpha$ , rejecting  $H_0$  if,

$$p - value < \alpha$$

Since  $p - value = 0.0000 < \alpha = 0.01$  we reject  $H_0$  and conclude that there is sufficient evidence from our sample to suggest that  $exper^2$  is statistically significant at explaining  $\log(wage)$  at the 1% significance level.

(c) 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?

To obtain this approximation,

- Take the derivative of  $\log(\widehat{wage})$  with respect to  $exper$ ,  $\frac{\partial \log(\widehat{wage})}{\partial exper}$
- Replace infinitesimally small change  $\partial$  with finite change  $\Delta$ ,  $\frac{\Delta \log(\widehat{wage})}{\Delta exper}$
- Multiple 100 on both sides
- Multiple  $\Delta exper$  on both sides

Return to 5<sup>th</sup> year of experience:

$$\begin{aligned} exper &= \\ \Delta exper &= \\ \% \Delta \widehat{wage} &\approx \\ &= \end{aligned}$$

Return to 20<sup>th</sup> year of experience:

$$\begin{aligned} exper &= \\ \Delta exper &= \\ \% \Delta \widehat{wage} &\approx \\ &= \end{aligned}$$

We can see that that percentage return on wage for each additional year of experience diminishes.

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

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

(Discuss in class)

## Question 2

EViews workfile: *marks.wf1*

The data set *marks.wf1* contains data on students' performance in a previous year in Introductory Econometrics. The variables are:

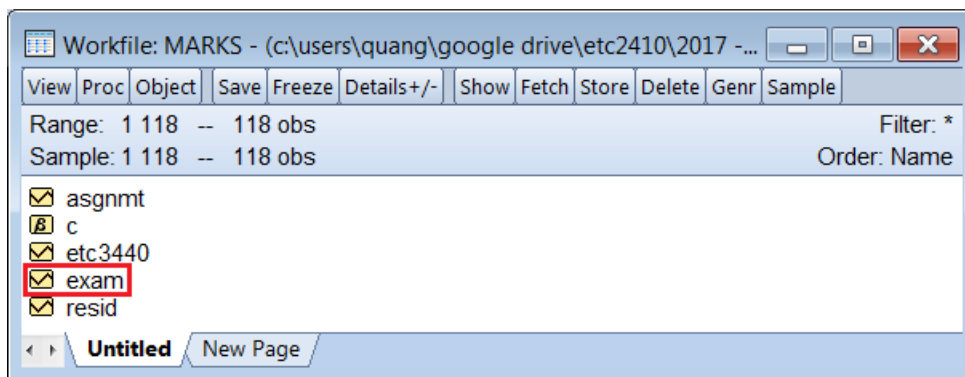
*exam* – final mark in exam (%)  
*asgnmt* – mark on assignments (%)  
*etc3440* – 1 for students in ETC3440, 0 for students in ETC2410

There were no students taking Introductory Econometrics under any other code in that year. The goal is to make a predictive model that uses assignment mark to predict final exam mark.

(a) “Look at the data.” Do you see any anomalies in the final exam marks? If so, discuss what you would do about it.

To visualise the distribution of final exam marks (histogram) and obtain some descriptive statistics,

*Double click on exam*



*View → Descriptive Statistics & Tests → Histograms and Stats*



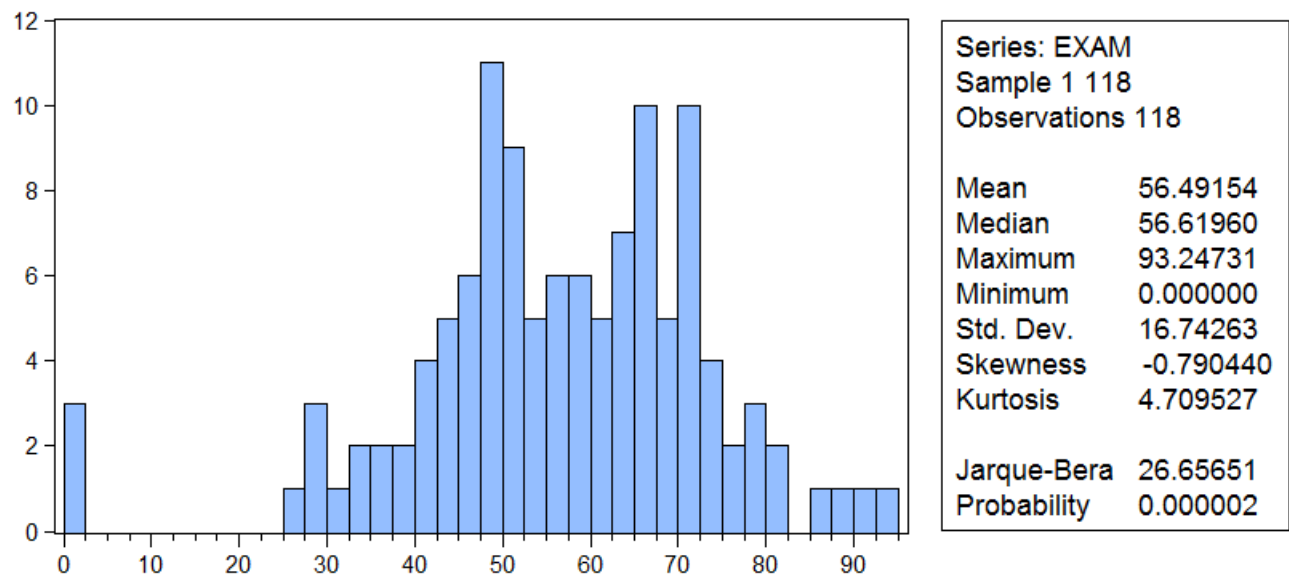
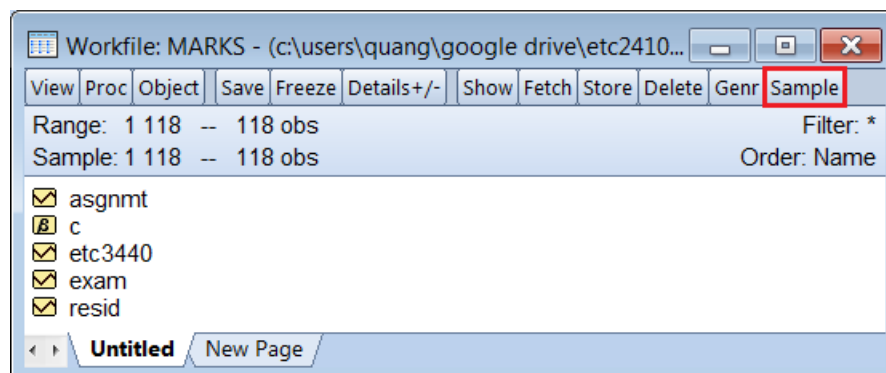


Figure 1: Histogram and descriptive statistics of final exam mark (%).

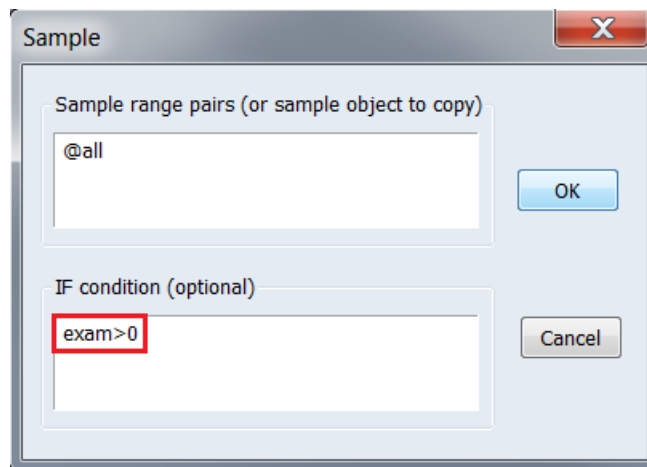
We observe 3 students with an exam mark of 0. These may have been students with very special consideration or ones that did not attend the exam, so we will omit these observations as they will affect the estimate of correlation between assignment mark and final exam mark, and  $\therefore$  our estimated predictive model.

To omit these observations from our EViews sample,

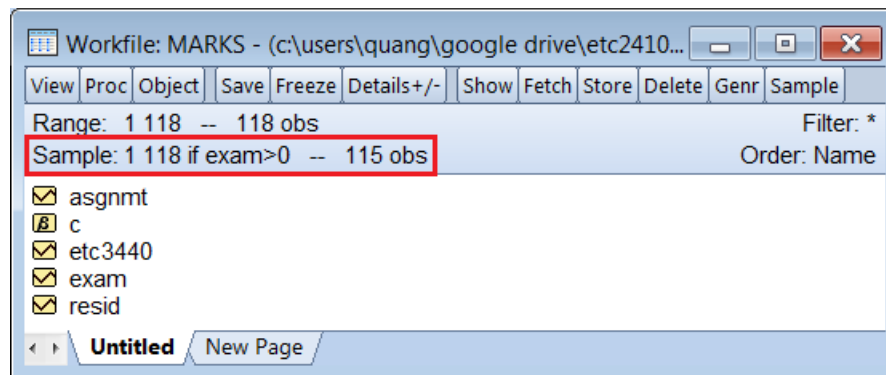
*Workfile*  $\rightarrow$  *Sample*



*IF condition (optional) : exam > 0*



The workfile sample now contains only students who scored higher than 0 in their final exam.



(b) By using the dummy variable *etc3440* in a regression, test if there is any difference between the expected exam mark in ETC2410 and ETC3440.

## Background

### Conditional expectation with dummy variables

For the simple regression model,

$$exam = \beta_0 + \beta_1 etc3440 + u$$

the expectation of exam score conditional on whether the student is enrolled in ETC2410 or ETC3440 is given by,

$$E(exam|etc3440) = \beta_0 + \beta_1 etc3440$$

$\therefore$  the expected exam score if the student is enrolled in ETC3440,

$$E(exam|etc3440 = 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

and the expected exam score if the student is enrolled in ETC2410,

$$E(exam|etc3440 = 0) = \beta_0 + \beta_1 \times 0 = \beta_0$$

If there is no difference between the expected exam score for ETC3440 and ETC2410 students then,

$$E(exam|etc3440 = 1) = E(exam|etc3440 = 0)$$

i.e.,

$$\beta_1 = 0$$

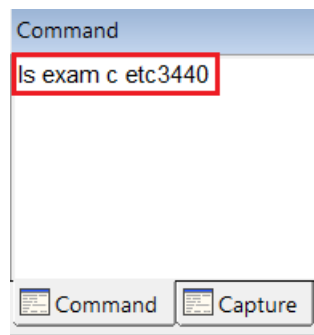
otherwise,

$$\beta_1 \neq 0$$

$$exam = \beta_0 + \beta_1 etc3440 + u$$

To estimate this model from the **Command window**,

*ls exam c etc3440*



(Press Enter to execute code)

Equation: UNTITLED Workfile: MARKS::Untitled\

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: EXAM  
Method: Least Squares  
Date: 08/20/17 Time: 10:59  
Sample: 1 118 IF EXAM>0  
Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	59.03932	1.734912	34.03014	0.0000
ETC3440	-2.573333	2.685381	-0.958275	0.3400

R-squared 0.008061 Mean dependent var 57.96523  
Adjusted R-squared -0.000717 S.D. dependent var 14.19578  
S.E. of regression 14.20087 Akaike info criterion 8.161722  
Sum squared resid 22788.11 Schwarz criterion 8.209460  
Log likelihood -467.2990 Hannan-Quinn criter. 8.181098  
F-statistic 0.918291 Durbin-Watson stat 0.341546  
Prob(F-statistic) 0.339970

To name (save) the estimated equation,

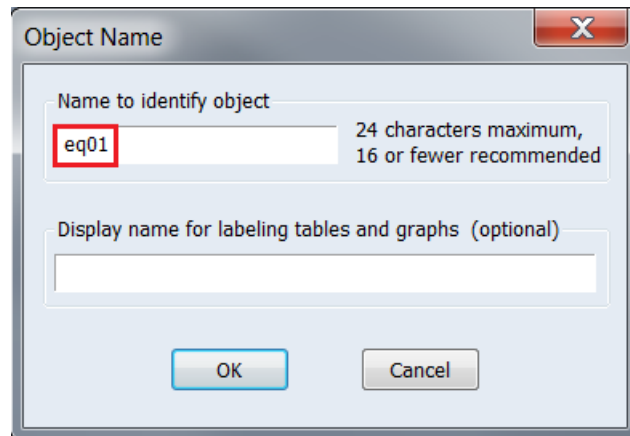
Name → Name to identify object : eq01

(This names the equation **eq01**)

Equation: UNTITLED Workfile: MARKS::Untitled\

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: EXAM  
Method: Least Squares  
Date: 08/20/17 Time: 10:59  
Sample: 1 118 IF EXAM>0  
Included observations: 115



Dependent Variable: EXAM  
 Method: Least Squares  
 Sample: 1 118 IF EXAM>0  
 Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	59.03932	1.734912	34.03014	0.0000
ETC3440	-2.573333	2.685381	-0.958275	0.3400
<hr/>				
R-squared	0.008061	Mean dependent var	57.96523	
Adjusted R-squared	-0.000717	S.D. dependent var	14.19578	
S.E. of regression	14.20087	Akaike info criterion	8.161722	
Sum squared resid	22788.11	Schwarz criterion	8.209460	
Log likelihood	-467.2990	Hannan-Quinn criter.	8.181098	
F-statistic	0.918291	Durbin-Watson stat	0.341546	
Prob(F-statistic)	0.339970			

$$\widehat{exam} = 59.0393 - 2.5733etc3440$$

(1.7349)      (2.6854)

**State the null and alternative hypothesis**

$H_0 : \beta_1 = 0$  (no difference in exam mark between ETC3440 and ETC2410 students)

$H_1 : \beta_1 \neq 0$  (difference in exam mark between ETC3440 and ETC2410 students)

**The test statistic and its distribution under  $H_0$**

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-k-1} \quad \text{under } H_0$$

$n = \text{sample size} = 115$

$k = \text{number of regressors} = 2$

### Calculate the test statistic

$$t_{calc} = -\frac{2.5733}{2.6854} = -0.9583$$

Dependent Variable: EXAM

Method: Least Squares

Date: 08/20/17 Time: 10:59

Sample: 1 118 IF EXAM>0

Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	59.03932	1.734912	34.03014	0.0000
ETC3440	-2.573333	2.685381	-0.958275	0.3400
R-squared	0.008061	Mean dependent var	57.96523	
Adjusted R-squared	-0.000717	S.D. dependent var	14.19578	
S.E. of regression	14.20087	Akaike info criterion	8.161722	
Sum squared resid	22788.11	Schwarz criterion	8.209460	
Log likelihood	-467.2990	Hannan-Quinn criter.	8.181098	
F-statistic	0.918291	Durbin-Watson stat	0.341546	
Prob(F-statistic)	0.339970			

### p-value

$$p - \text{value} = 0.3400$$

Note: The *Prob.* values from the EViews regression output are p-values for a two-sided t-test to test if a regressor is statistically significant, holding the other regressors constant.

### Critical value and rejection region

To obtain the critical value using the Stats Table, locate the t distribution table,

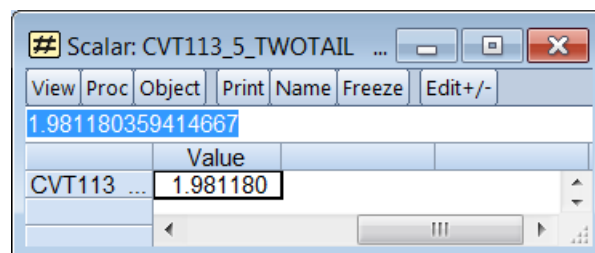
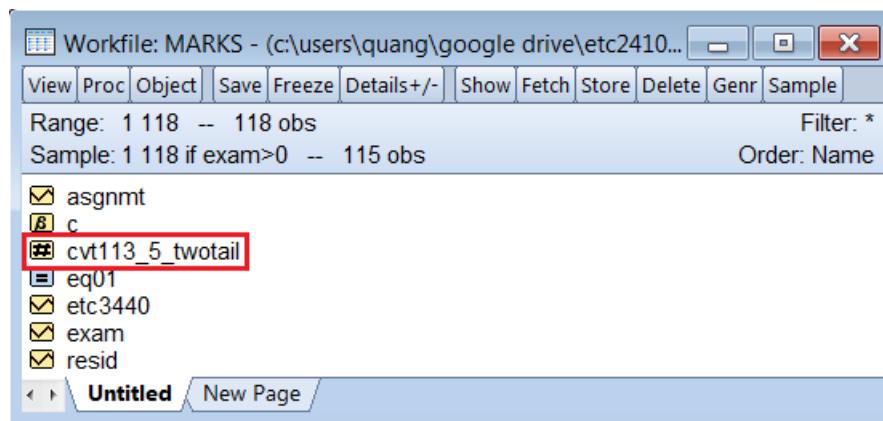
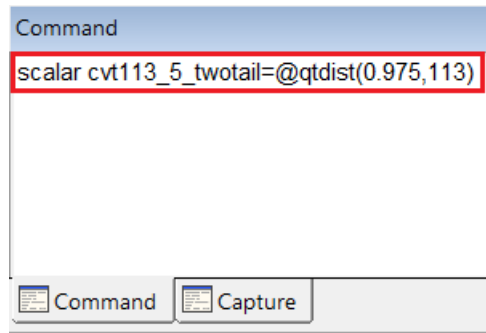
$$\text{degrees of freedom} = 113$$

Since 169 is not in the table, we take a conservative approach and choose the closest available degrees of freedom less than 113 i.e.  $d.o.f = 90$ .

TABLE G.2 Critical Values of the <i>t</i> Distribution						
		Significance Level				
1-Tailed:		.10	.05	.025	.01	.005
2-Tailed:		.20	.10	.05	.02	.01
D e g r e e s  o f  F r e e d o m	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

To obtain the critical value using EViews,

Command window : `scalar cvt113_5_twotail = @qtdist(0.975,113)`



From Stat Table:

$$+t_{crit} = 1.987$$

$$-t_{crit} = -1.987$$

From EViews:

$$+t_{crit} = 1.9812$$

$$-t_{crit} = -1.9812$$

Rejection rule:

Comparing the calculated test statistic with the critical value, we reject  $H_0$  if,

$$t_{calc} > +t_{crit}$$



*or*

$$t_{calc} < -t_{crit}$$

Comparing the p-value with the significance level, we reject  $H_0$  if,

$$p - value < \alpha = 0.05$$

### **Conclusion**

Since  $p - value = 0.3400 > \alpha = 0.05$ , we do not reject the null at the 5% significance level and conclude that there is insufficient evidence from our sample to suggest that there is a difference in expected exam mark between ETC3440 and ETC2410 students.

(c) Estimate a regression of *exam* on a constant and *asgnmt*,

$$exam = \beta_0 + \beta_1 asgnmt + u$$

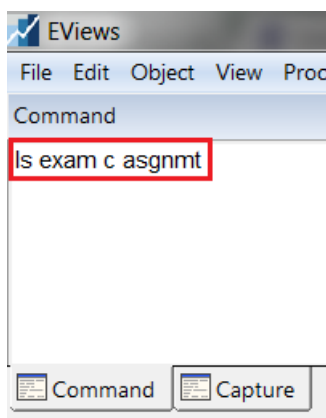
along with the scatter plot with regression line added.

Based on this regression, provide predictions for the exam marks of a student who has obtained 40% on the assignment, and another student who has obtained 80% on the assignment.

Provide 95% prediction intervals for the exam mark, once ignoring estimation uncertainty, and once properly accounting for estimation uncertainty.

To estimate this model from the Command window,

*ls exam c asgnmt*

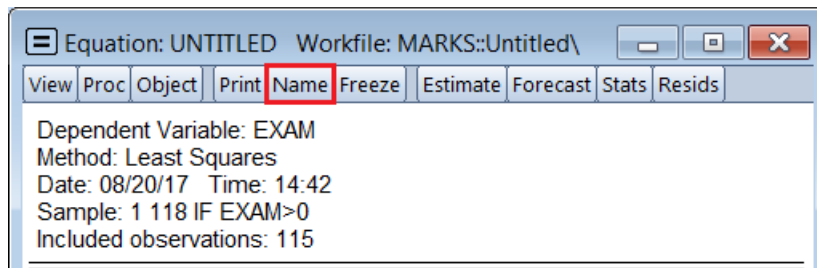


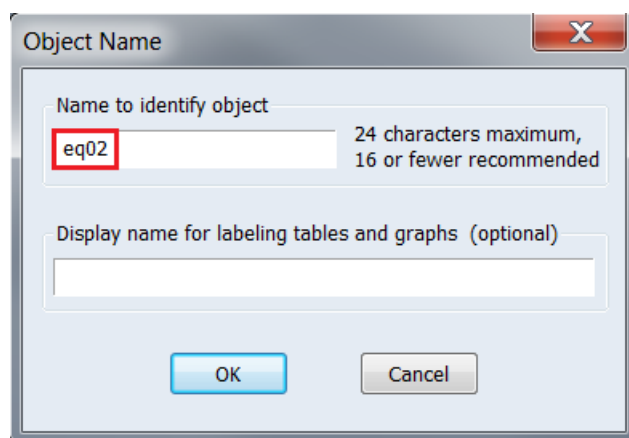
*(Press Enter to execute code)*

To name (save) the estimated equation,

*Name → Name to identify object : eq02*

*(This names the equation **eq02**)*





Dependent Variable: EXAM

Method: Least Squares

Date: 04/23/18 Time: 13:40

Sample: 1 118 IF EXAM>0

Included observations: 115

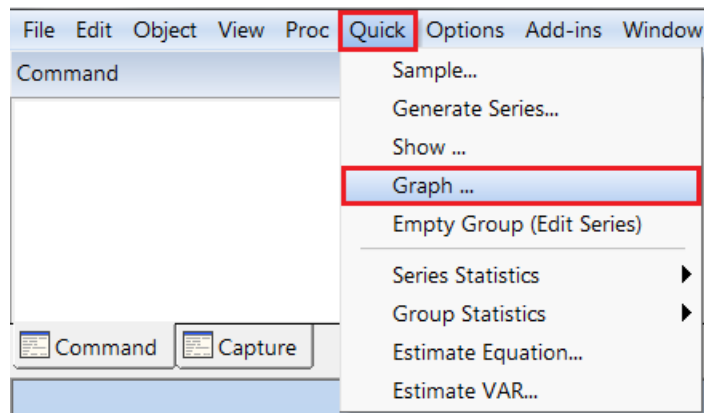
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20.04634	4.876848	4.110512	0.0001
ASGNMT	0.638077	0.080088	7.967197	0.0000
R-squared	0.359687	Mean dependent var	57.96523	
Adjusted R-squared	0.354021	S.D. dependent var	14.19578	
S.E. of regression	11.40955	Akaike info criterion	7.724017	
Sum squared resid	14710.10	Schwarz criterion	7.771755	
Log likelihood	-442.1310	Hannan-Quinn criter.	7.743394	
F-statistic	63.47622	Durbin-Watson stat	1.240656	
Prob(F-statistic)	0.000000			

$$\widehat{exam} = 20.0463 + 0.6381asgmt$$

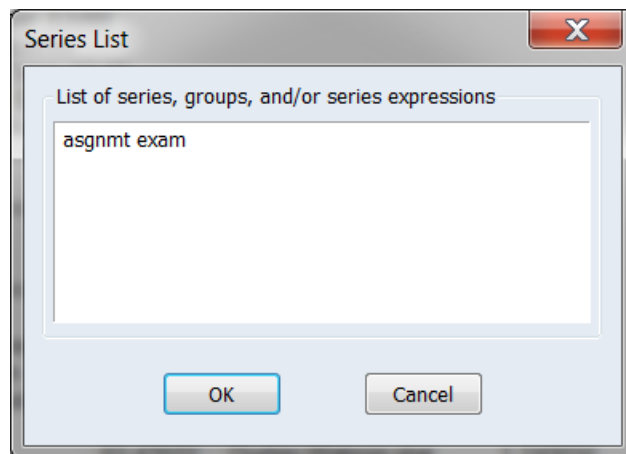
(4.8768)
(0.0801)

To obtain a scatter plot of *exam* against *asgmt* with a regression line,

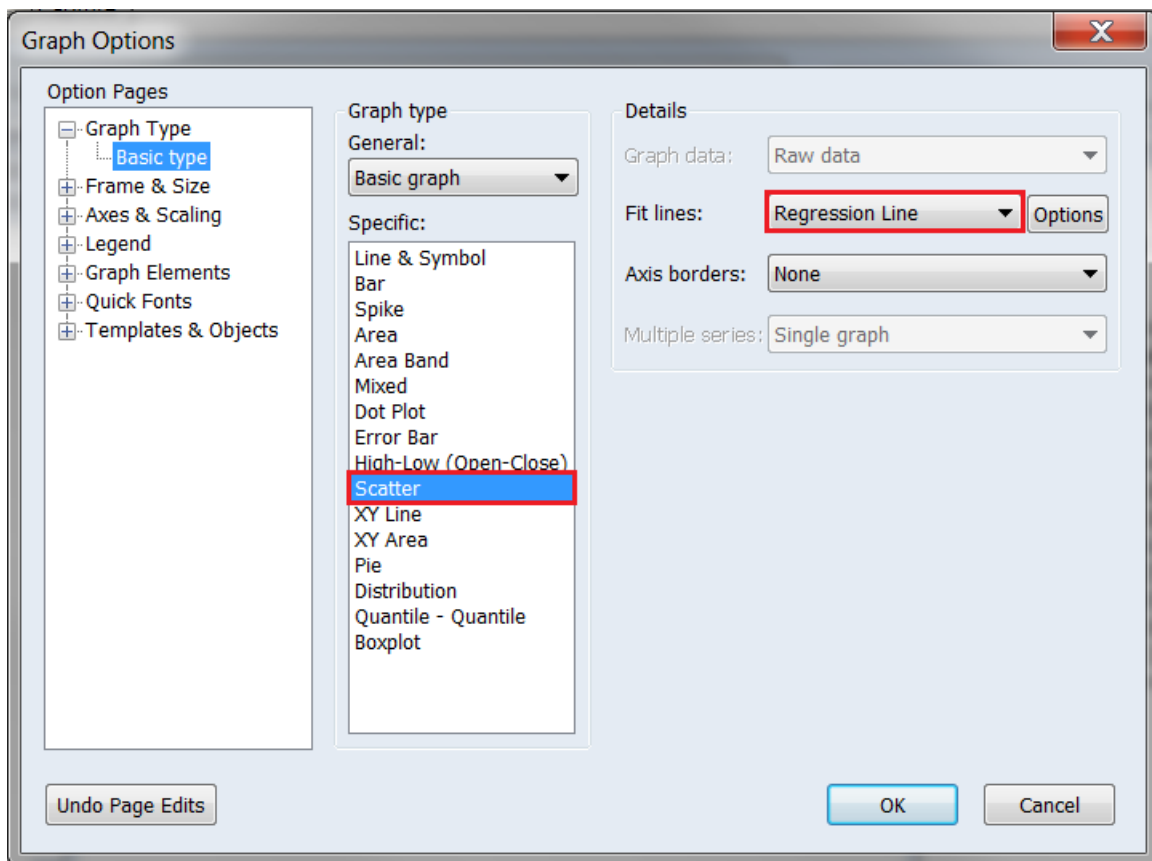
*Quick* → *Graph...*



*Series List : asgnmt exam*



*Specific : Scatter  $\rightarrow$  Fit lines : Regression Line*



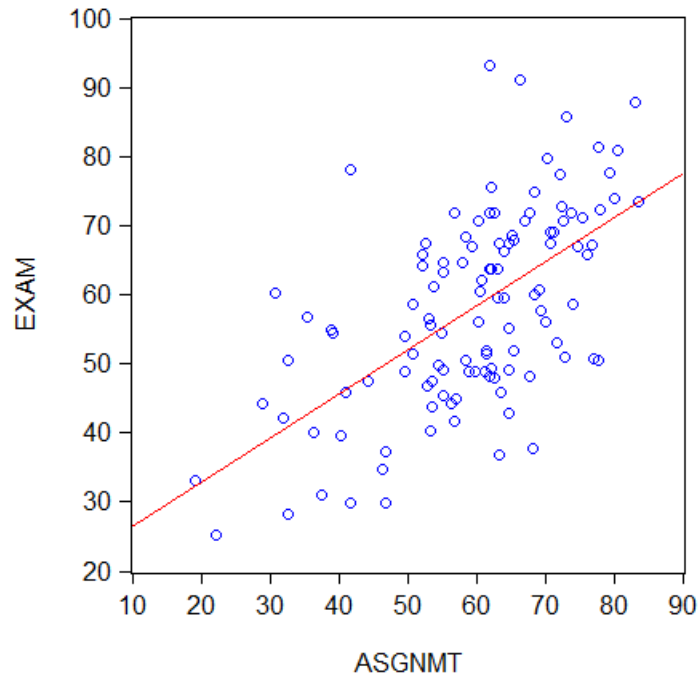


Figure 2: Scatter plot of final exam mark against assignment mark.

Based on this regression, provide predictions for the exam marks of a student who has obtained 40% on the assignment, and another student who has obtained 80% on the assignment.

$$exam = \beta_0 + \beta_1 asgnmt + u$$

$$E(exam|asgnmt) = \beta_0 + \beta_1 asgnmts$$

$$\begin{aligned}\widehat{exam} &= E(\widehat{exam}|asgnmt) \\ &= \hat{\beta}_0 + \hat{\beta}_1 asgnmt\end{aligned}$$

$\widehat{exam}$  plays two roles:

- It is the prediction of final exam score given an assignment score
- And also the estimated expectation of final exam score conditional on assignment score

Prediction of exam mark for a student that scored 40% on the assignment:

$$\widehat{exam} = E(\widehat{exam}|asgnmt = 40)$$

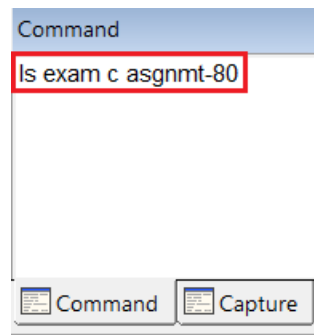
$$\begin{aligned}
&= 20.0463 + 0.6381 \times 40 \\
&= 45.5703
\end{aligned}$$

Prediction of exam mark for a student that scored 80% on the assignment:

$$\begin{aligned}
\widehat{exam} &= E(exam | \widehat{asgnmt} = 80) \\
&= 20.0463 + 0.6381 \times 80 \\
&= 71.0925
\end{aligned}$$

By regressing *exam* on a constant and (*asgnmt* − 80),

$$exam = \beta_0 + \beta_1(asgnmt - 80) + u$$



Dependent Variable: EXAM

Method: Least Squares

Sample: 1 118 IF EXAM>0

Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	71.09253	1.961324	36.24721	0.0000
ASGNMT-80	0.638077	0.080088	7.967197	0.0000
R-squared	0.359687	Mean dependent var	57.96523	
Adjusted R-squared	0.354021	S.D. dependent var	14.19578	
S.E. of regression	11.40955	Akaike info criterion	7.724017	
Sum squared resid	14710.10	Schwarz criterion	7.771755	
Log likelihood	−442.1310	Hannan-Quinn criter.	7.743394	
F-statistic	63.47622	Durbin-Watson stat	1.240656	
Prob(F-statistic)	0.000000			

$$\widehat{exam} = \underset{(1.9613)}{71.0925} + \underset{(0.0801)}{0.6381}(asgnmt - 80)$$

$\hat{\beta}_0$  becomes the prediction of final exam mark for a student that scored 80% on the assignment, and  $se(\hat{\beta}_0)$  becomes the standard error of the prediction of final exam mark for a student scored 80% on the assignment.

Provide 95% prediction intervals for the exam mark, once ignoring estimation uncertainty, and once properly accounting for estimation uncertainty.

## Background

### Prediction Uncertainty

For the following simple regression model,

$$exam = \beta_0 + \beta_1 asgmt + u$$

The expectation of *exam* conditional on *asgmt* is given by,

$$\begin{aligned} E(exam|asgmt) &= E(\beta_0 + \beta_1 asgmt + u|asgmt) \\ &= \beta_0 + \beta_1 asgmt + E(u|asgmt) \\ &= \beta_0 + \beta_1 asgmt \end{aligned}$$

The true final exam mark given that *assignment score* is 80,

$$exam = \beta_0 + 80\beta_1 + u$$

Let  $\widehat{exam}$  be the prediction of final exam mark for an *assignment score* of 80,

$$\widehat{exam} = \hat{\beta}_0 + 80\hat{\beta}_1$$

which is also equal to the estimated expectation of exam mark given an *assignment score* of 80,

$$E(exam|\widehat{asgmt} = 80) = \hat{\beta}_0 + 80\hat{\beta}_1$$

Since  $\widehat{exam} = E(exam|\widehat{asgmt} = 80)$  depends on  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , the OLS estimator for  $\beta_0$  and  $\beta_1$ ,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

is subject to sampling variation (since it is a random variable and varies from sample to sample), so it follows that,

$$\widehat{exam} = \hat{\beta}_0 + 80\hat{\beta}_1$$

(the prediction of final exam mark for a student with an assignment score of 80) is also subject to sampling variation. This source of uncertainty in our prediction is called **estimation uncertainty**. (It comes from the sampling variability in  $\hat{\beta}$  which leads to uncertainty in  $\widehat{exam}$ ).



Since,

$$\widehat{exam} = \hat{\beta}_0 + 80\hat{\beta}_1$$

is a prediction of,

$$exam = \beta_0 + 80\beta_1 + u$$

The second source of prediction uncertainty, which is unexplained by our model, is in the **unobserved error  $u$** . High variation in  $u$  increases prediction uncertainty.

Note that a confidence interval of  $E(exam|asgmt = 80)$ , depends on the variability of  $E(exam|\widehat{asgmt} = 80)$  but not on the variability of  $u$ , however, the prediction interval of  $exam$  when  $asgmt = 80$ , depends on both the variability in  $E(exam|\widehat{asgmt} = 80)$  and  $u$ .

- $E(exam|asgmt = 80) = \beta_0 + \beta_1 80$  &  $E(exam|\widehat{asgmt} = 80) = \hat{\beta}_0 + \hat{\beta}_1 80$
- $exam = \beta_0 + \beta_1 80 + u$  &  $\widehat{exam} = \hat{\beta}_0 + \hat{\beta}_1 80$

### Prediction Interval

The 95% prediction interval of exam score for an assignment score of 80 is given by,

$$\widehat{exam} \pm t_{0.975, n-k-1} \times se(\hat{e})$$

where  $\hat{e}$  represents the prediction error i.e. the difference between true final exam mark for *assignment score of 80* and the prediction of final exam mark for an *assignment score of 80*,

$$\begin{aligned}\hat{e} &= exam - \widehat{exam} \\ &= (\beta_0 + 80\beta_1 + u) - (E(exam|\widehat{asgmt} = 80)) \\ &= (\beta_0 + 80\beta_1 + u) - (\hat{\beta}_0 + 80\hat{\beta}_1)\end{aligned}$$

and is itself a random variable.

(Not to be confused with the OLS residual  $\hat{u}$ , which is the difference between final exam marks from the individuals in our sample and the predicted final exam marks for these individuals.)

Whether we wish to account for estimation uncertainty in our prediction interval depends on how we formulate the standard error of the prediction,

$$\begin{aligned}se(\hat{e}) &= \sqrt{\widehat{var}(\hat{e}|\widehat{asgmt} = 80)} \\ &= \sqrt{\widehat{var}(exam - \widehat{exam}|\widehat{asgmt} = 80)} \\ &= \sqrt{\hat{\sigma}^2 + \widehat{var}(\widehat{exam}|\widehat{asgmt} = 80) + 0} \\ &= \sqrt{\hat{\sigma}^2 + (se(\widehat{exam}|\widehat{asgmt} = 80))^2}\end{aligned}$$

$se(\hat{e})$  is the statistic that captures both source of prediction uncertainty:

- Estimation uncertainty i.e. the variability in our estimation given by  $(se(\widehat{exam}|x = 80))^2$
- Uncertainty from unobserved error  $u$  measured by  $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{n - k - 1}$$

In EViews,  $\hat{\sigma}$  is reported under the name **Standard Error of Regression**.

95% prediction interval for the exam mark of a student who obtained 80% with estimation uncertainty,

$$\widehat{exam} \pm t_{0.975, n-k-1} \times se(\hat{e})$$

where,

$$\begin{aligned}\widehat{exam} &= \\ n &= 115 \\ k &= 2 \\ t_{0.975, 113} &= 1.9812 \\ se(\hat{e}) &= \\ &= \\ &= \\ &= \end{aligned}$$

Dependent Variable: EXAM  
Method: Least Squares  
Date: 08/20/17 Time: 15:57  
Sample: 1 118 IF EXAM>0  
Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	71.09253	1.961324	36.24721	0.0000
ASGNMT-80	0.638077	0.080088	7.967197	0.0000
R-squared	0.359687	Mean dependent var	57.96523	
Adjusted R-squared	0.354021	S.D. dependent var	14.19578	
S.E. of regression	11.40955	Akaike info criterion	7.724017	
Sum squared resid	14710.10	Schwarz criterion	7.771755	
Log likelihood	-442.1310	Hannan-Quinn criter.	7.743394	
F-statistic	63.47622	Durbin-Watson stat	1.240656	
Prob(F-statistic)	0.000000			

therefore,

$$71.0925 \pm 1.9812 \times 11.57$$

$$[48.191, 94.009]$$

We predict with 95% confidence that ...

95% prediction interval for the exam mark of a student who obtained 80% without estimation uncertainty,

$$\begin{aligned} & exam \pm t_{0.975, n-k-1} \times se(\hat{e}) \\ & exam \pm t_{0.975, n-k-1} \times \sqrt{\hat{\sigma}^2 + (se(exam|assignment = 80))^2} \end{aligned}$$

(d) Specify and estimate a regression model to test whether both the intercept and slope in the regression in part (c) is different for ETC2410 and ETC3440.

By including the dummy variable etc3440, we have a regression model where the intercept of expected exam score differs between ETC2410 and ETC3440 students holding assignment score constant,

$$exam = \beta_0 + \beta_1 asgnmt + \delta_0 etc3440 + u$$

$$E(exam|asgnmt, etc3440) = \beta_0 + \beta_1 asgnmt + \delta_0 etc3440$$

$$\begin{aligned} E(exam|asgnmt, etc3440 = 0) &= \beta_0 + \beta_1 asgnmt + \delta_0 \times 0 \\ &= \beta_0 + \beta_1 asgnmt \end{aligned}$$

$$\begin{aligned} E(exam|asgnmt, etc3440 = 1) &= \beta_0 + \beta_1 asgnmt + \delta_0 \times 1 \\ &= (\beta_0 + \delta_0) + \beta_1 asgnmt \end{aligned}$$

$\delta_0$  is the difference in intercept.

By including both the dummy variable and the interaction between etc3440 and asgnmt, we have a regression model where both intercept and slope of expected exam score differs between ETC2410 and ETC3440 students holding assignment score constant,

$$exam = \beta_0 + \beta_1 asgnmt + \delta_0 etc3440 + \delta_1 etc3440 * asgnmt + u$$

$$E(exam|asgnmt, etc3440) = \beta_0 + \beta_1 asgnmt + \delta_0 etc3440 + \delta_1 etc3440 * asgnmt$$

$$\begin{aligned} E(exam|asgnmt, etc3440 = 0) &= \beta_0 + \beta_1 asgnmt + \delta_0 \times 0 + \delta_1 \times 0 \\ &= \beta_0 + \beta_1 asgnmt \end{aligned}$$

$$\begin{aligned} E(exam|asgnmt, etc3440 = 1) &= \beta_0 + \beta_1 asgnmt + \delta_0 \times 1 + \delta_1 asgnmt \\ &= (\beta_0 + \delta_0) + (\beta_1 + \delta_1) asgnmt \end{aligned}$$

$\delta_0$  and  $\delta_1$  represent the different in intercept and slope respectively.

Testing whether the intercept and/or slope is different between ETC3440 and ETC2410 students.

If there is no difference between intercepts and slopes for the expectation of exam mark between ETC3440 and ETC2410 students (holding assignment mark constant),

$$\delta_0 = \delta_1 = 0$$

and if the intercept and/or slope is different,

$$\delta_0 \neq 0 \text{ and/or } \delta_1 \neq 0$$

Since this is a test of 2 linear restrictions and the alternative is a two-sided, we use the F-test. When performing an F-test, we need to estimate the unrestricted and restricted model to obtain  $SSR_{ur}$  and  $SSR_r$ .

**Unrestricted model (the model before imposing restrictions):**

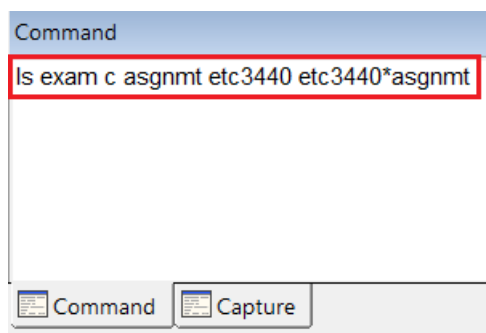
$$exam = \beta_0 + \beta_1 asgnmt + \delta_0 etc3440 + \delta_1 etc3440 * asgnmt + u$$

**Restricted model (the model after imposing restrictions):**

$$\begin{aligned} exam &= \beta_0 + \beta_1 asgnmt + 0 \times etc3440 + 0 \times etc3440 * asgnmt + u \\ &= \beta_0 + \beta_1 asgnmt + u \end{aligned}$$

To estimate the unrestricted model from the Command window,

*ls exam c asgnmt etc3440 etc3440\*asgnmt*



Dependent Variable: EXAM  
Method: Least Squares  
Sample: 1 118 IF EXAM>0  
Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.72129	5.657110	2.425495	0.0169
ASGNMT	0.768005	0.093077	8.251292	0.0000
ETC3440	21.87665	10.38634	2.106291	0.0374
ETC3440*ASGNMT	-0.420274	0.170347	-2.467161	0.0151
R-squared	0.404915	Mean dependent var	57.96523	
Adjusted R-squared	0.388831	S.D. dependent var	14.19578	
S.E. of regression	11.09788	Akaike info criterion	7.685548	
Sum squared resid	13671.08	Schwarz criterion	7.781024	
Log likelihood	-437.9190	Hannan-Quinn criter.	7.724301	
F-statistic	25.17594	Durbin-Watson stat	1.282800	
Prob(F-statistic)	0.000000			

$$SSR_{ur} = 13671.08$$

Dependent Variable: EXAM  
Method: Least Squares  
Sample: 1 118 IF EXAM>0  
Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20.04634	4.876848	4.110512	0.0001
ASGNMT	0.638077	0.080088	7.967197	0.0000
R-squared	0.359687	Mean dependent var	57.96523	
Adjusted R-squared	0.354021	S.D. dependent var	14.19578	
S.E. of regression	11.40955	Akaike info criterion	7.724017	
Sum squared resid	14710.10	Schwarz criterion	7.771755	
Log likelihood	-442.1310	Hannan-Quinn criter.	7.743394	
F-statistic	63.47622	Durbin-Watson stat	1.240656	
Prob(F-statistic)	0.000000			

$$SSR_r = 14710.10$$

**State the null and alternative hypothesis**

$$H_0 : \delta_0 = \delta_1 = 0$$

$$H_1 : \delta_0 \neq 0 \text{ and/or } \delta_1 \neq 0$$

**The test statistic and its distribution under  $H_0$**

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)} = \frac{(SSR_r - SSR_{ur})/2}{SSR_{ur}/(115 - 3 - 1)} \sim F_{2,111} \quad \text{under } H_0$$

$$n = \text{sample size} = 115$$

$$k = \text{number of regressors in the unrestricted model} = 4$$

$$q = \text{number of restrictions} = 2$$

$$SSR_r = \text{sum of squared residuals from estimated restricted model}$$

$$SSR_{ur} = \text{sum of squared residuals from estimated unrestricted model}$$

**Calculate the test statistic**

$$F_{calc} = \frac{(SSR_r - SSR_{ur})/2}{SSR_{ur}/(111)} = \frac{(14710.10 - 13671.08)/2}{13671.08/(111)} = 4.2181$$

**Critical value and rejection region**

$$5\% \text{ significance level} \rightarrow \alpha = 0.05$$

To obtain the critical value using the Stats Table, locate the F distribution table at the 5% significance level,

$$\text{Numerator d.o.f} = 2$$

$$\text{Denominator d.o.f} = 111$$

Since 111 is not in the table, we take a conservative approach and choose the closest available degrees of freedom less than 111 i.e.  $d.o.f = 90$ .



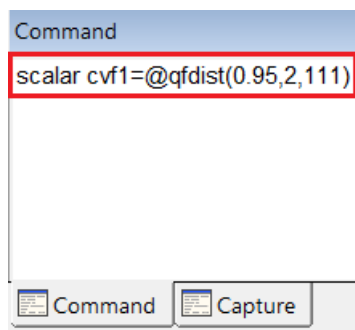
TABLE G.3b 5% Critical Values of the <i>F</i> Distribution											
		Numerator Degrees of Freedom									
		1	2	3	4	5	6	7	8	9	10
D e n o m i n a t o r	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
D e g r e e s	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
F r e e d o m	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
$\infty$		3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

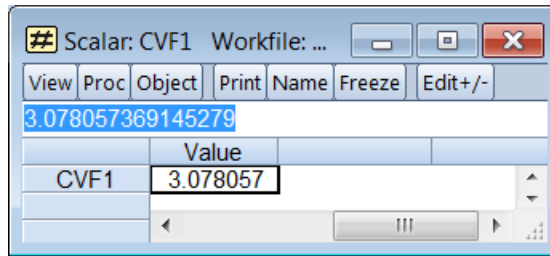
Example: The 5% critical value for numerator  $df = 4$  and large denominator  $df(\infty)$  is 2.37.

Source: This table was generated using the Stata® function invFtail.

To obtain the critical value using EViews,

Command window : `scalar cvf1 = @qfdist(0.95,2,111)`





$$F_{crit} \text{ (from Stat Table)} = 3.10$$

$$F_{crit} \text{ (from EViews)} = 3.0781$$

Rejection rule:

Comparing the calculated test statistic with the critical value, we reject  $H_0$  if,

$$F_{calc} > F_{crit}$$

## Conclusion

Since  $F_{calc} = 4.2181 > F_{crit} = 3.0781$ , we reject the null at the 5% significance level and conclude that there is sufficient evidence from our sample to suggest that the expectation of exam mark differs for ETC3440 and ETC2410 students either in the intercept and/or the slope, holding assignment mark constant.

From the unrestricted regression results, obtain the estimate of the intercept and slope for ETC2410 and for ETC3440 regression lines.

Dependent Variable: EXAM  
Method: Least Squares  
Sample: 1 118 IF EXAM>0  
Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.72129	5.657110	2.425495	0.0169
ASGNMT	0.768005	0.093077	8.251292	0.0000
ETC3440	21.87665	10.38634	2.106291	0.0374
ETC3440*ASGNMT	-0.420274	0.170347	-2.467161	0.0151
R-squared	0.404915	Mean dependent var	57.96523	
Adjusted R-squared	0.388831	S.D. dependent var	14.19578	
S.E. of regression	11.09788	Akaike info criterion	7.685548	
Sum squared resid	13671.08	Schwarz criterion	7.781024	
Log likelihood	-437.9190	Hannan-Quinn criter.	7.724301	
F-statistic	25.17594	Durbin-Watson stat	1.282800	
Prob(F-statistic)	0.000000			

Regression line of final exam mark against assignment mark for ETC2410 students:

$$\begin{aligned}
E(\widehat{exam} | asgnmt, etc3440 = 0) &= \hat{\beta}_0 + \hat{\beta}_1 asgnmt \\
&= 13.7213 + 0.7680 asgnmt
\end{aligned}$$

Regression line of final exam mark against assignment mark for ETC3440 students:

$$\begin{aligned}
E(\widehat{exam} | asgnmt, etc3440 = 1) &= (\hat{\beta}_0 + \hat{\delta}_0) + (\hat{\beta}_1 + \hat{\delta}_1) asgnmt \\
&= (13.7213 + 21.8767) + (0.7680 - 0.4203) asgnmt \\
&= 35.5979 + 0.3477 asgnmt
\end{aligned}$$

(e) Estimate two separate models to predict exam mark based on assignment mark, one for ETC2410 students and another for ETC3440 students.

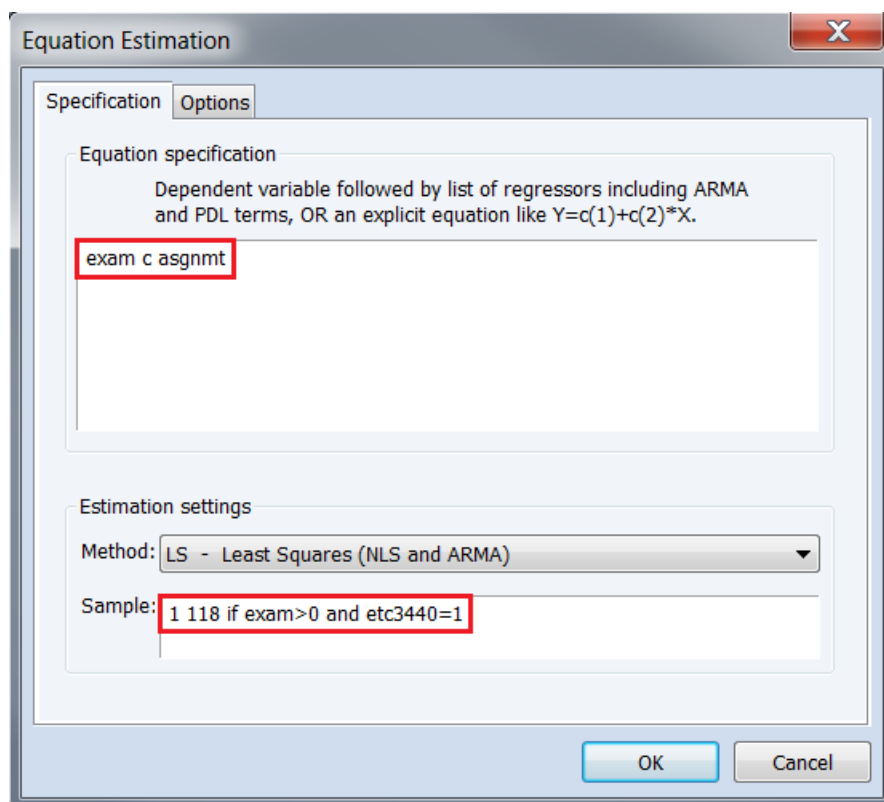
$$exam = \beta_0 + \beta_1 asgnmt + u$$

To estimate the model based on sample of ETC3440 students only:

*Quick → Estimate Equation*

*Equation estimation : exam c asgnmt*

*Sample : 1 118 if exam > 0 and etc3440 = 1*



Dependent Variable: EXAM

Method: Least Squares

Sample: 1 118 IF EXAM>0 AND ETC3440=1

Included observations: 48

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	35.59794	8.108920	4.389973	0.0001
ASGNMT	0.347731	0.132817	2.618119	0.0119
R-squared	0.129687	Mean dependent var	56.46599	
Adjusted R-squared	0.110767	S.D. dependent var	10.95598	
S.E. of regression	10.33139	Akaike info criterion	7.549025	
Sum squared resid	4909.934	Schwarz criterion	7.626992	
Log likelihood	-179.1766	Hannan-Quinn criter.	7.578489	
F-statistic	6.854548	Durbin-Watson stat	0.971504	
Prob(F-statistic)	0.011930			

$$\widehat{exam} = \underset{(8.1089)}{35.5979} + \underset{(0.1328)}{0.3477} asgnmt$$

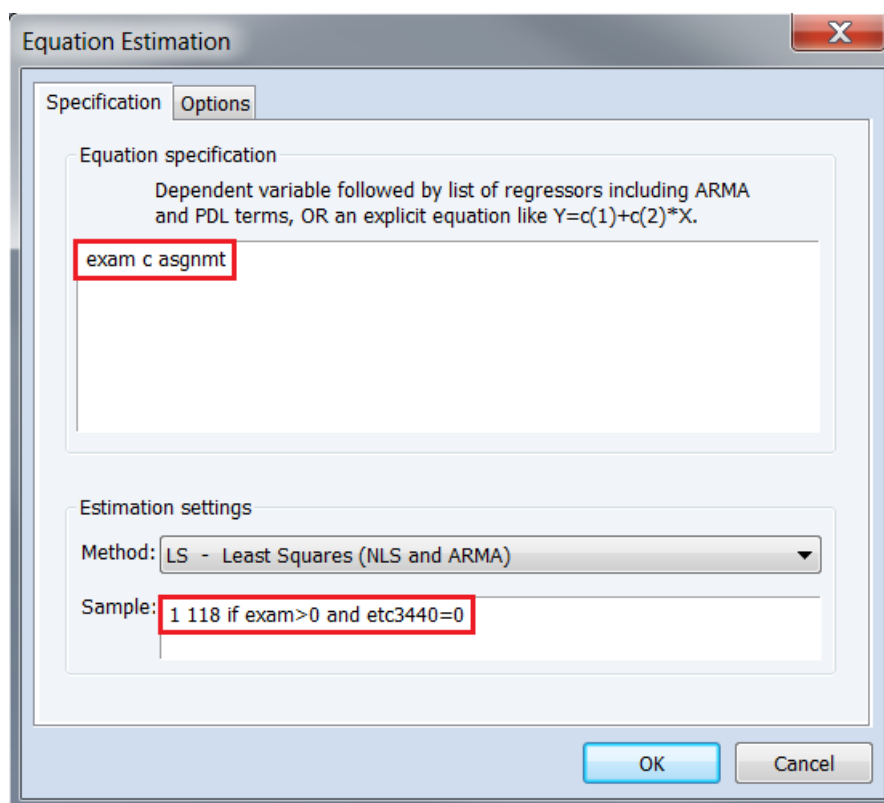
$$SSR_{with \ etc3440 \ students \ only} = 4909.934$$

To estimate the model based on sample of ETC2410 students only:

*Quick  $\rightarrow$  Estimate Equation*

*Equation estimation : exam c asgnmt*

*Sample : 1 118 if exam > 0 and etc3440 = 1*



Dependent Variable: EXAM

Method: Least Squares

Sample: 1 118 IF EXAM>0 AND ETC3440=0

Included observations: 67

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.72129	5.918047	2.318550	0.0236
ASGNMT	0.768005	0.097370	7.887477	0.0000
R-squared	0.489043	Mean dependent var	59.03932	
Adjusted R-squared	0.481182	S.D. dependent var	16.11819	
S.E. of regression	11.60977	Akaike info criterion	7.770968	
Sum squared resid	8761.143	Schwarz criterion	7.836779	
Log likelihood	-258.3274	Hannan-Quinn criter.	7.797009	
F-statistic	62.21230	Durbin-Watson stat	1.442208	
Prob(F-statistic)	0.000000			

$$\widehat{exam} = 13.7213 + 0.7680asgnmt$$

(5.9180)      (0.0974)

$$SSR_{with \text{ etc2410 students only}} = 8761.143$$

Compare the estimates of the intercept and slope you obtain here with what you obtained in part (d) and discuss.

From part (d), we regressed *exam* on a constant, *asgnmt*, *etc3440* and *etc3440\*asgnmt* using the sample with both ETC3440 and ETC2410 students,

$$\widehat{exam} = 13.7213 + 0.7680asgnmt + 21.8767etc3440 - 0.4204etc3440*asgnmt$$

(5.6571)      (0.0931)      (10.3863)      (0.1703)

$$SSR_{ur} = 13671.08$$

ETC2410 students only,

$$\widehat{exam} = 13.7213 + 0.7680asgnmt$$

(5.9180)      (0.0974)

ETC3440 students only,

$$\widehat{exam} = 35.5979 + 0.3477asgnmt$$

(8.1089)      (0.1328)

What we find when comparing each model,

$$\hat{\beta}_{0 \text{ with etc2410 students only}} = \hat{\beta}_{0 \text{ part}(d)}$$

$$\hat{\beta}_{1 \text{ with etc2410 students only}} = \hat{\beta}_{1 \text{ part}(d)}$$

$$\hat{\beta}_{0 \text{ with etc3440 students only}} = \hat{\beta}_{0 \text{ part}(d)} + \hat{\delta}_{0 \text{ part}(d)}$$

$$\hat{\beta}_{1 \text{ with etc3440 students only}} = \hat{\beta}_{1 \text{ part}(d)} + \hat{\delta}_{1 \text{ part}(d)}$$

The intuitive reason for why we observe this relationship:

- The parameters for the ETC2410 group (the parameters you see when the dummy variable equals to 0) only affects the *SSR* of the ETC2410 group.
- Similarly, the parameters for the ETC3440 group (the parameters you see when the dummy variable equals to 1) only affects the *SSR* of the ETC3440 group.
- There is no common parameter that enters both i.e. each group's parameters are separate.

As such, the OLS estimator achieves the smallest *SSR* for the model of both groups, when the *SSR* of the model of each individual group is minimised.

$$SSR_{with \text{ etc2410 students only}} + SSR_{with \text{ etc3440 students only}} = SSR_{ur}$$

$$8761.143 + 4909.934 = 13671.08$$

## Background

### Chow test

The Chow test is an F-test of the difference between intercept and/or slope by running separate regressions for each group and then computing  $SSR_{ur}$  by summing each group's SSR.

Regress with sample of ETC2410 students only:

$$exam = \beta_0 + \beta_1 asgnmt + u$$

$$SSR_{2410} = 8761.143$$

Regress with sample of ETC3440 students only:

$$exam = \beta_0 + \beta_1 asgnmt + u$$

$$SSR_{3440} = 4909.934$$

Regress with sample of both ETC2410 and ETC3440 students :

$$exam = \beta_0 + \beta_1 asgnmt + u$$

$$SSR_r = ???$$



(f) Consider the estimated intercept and slope of ETC2410 students in part (d) and part (e). While the numerical values are the same, their standard errors are not. Discuss why they are different, and which one you would prefer to use to construct a 95% confidence interval for the slope parameter for ETC2410 students.

The standard error of  $\hat{\beta}$ ,

$$\widehat{var}(\hat{\beta}|\mathbf{X}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \quad (\text{estimated variance - covariance matrix of } \hat{\beta})$$

$$se(\hat{\beta}) = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}}$$

depends on,

$$\hat{\sigma}^2 = \frac{\sum_i^n \hat{u}_i^2}{n - k - 1} = \frac{SSR}{n - k - 1} = \text{unbiased estimator of } var(\mathbf{u}|\mathbf{X})$$

and since  $\hat{\sigma}^2$  differs between each model,

$$\hat{\sigma}_{with\ etc2410\ students\ only}^2 = 11.6098^2 \neq \hat{\sigma}_{ur}^2 = 11.0997^2$$

We need to consider which standard error to use when constructing the 95% confidence interval of the slope parameter for ETC2410 students,

$$\hat{\beta}_1 \pm t_{0.975} \times se(\hat{\beta}_1)_{with\ etc2410\ students\ only}$$

OR

$$\hat{\beta}_1 \pm t_{0.975} \times se(\hat{\beta}_1)_{unrestricted\ with\ both\ etc2410\ and\ etc3440\ sample}$$

From question (d), we found that there was sufficient evidence to suggest that the conditional expectation of exam mark, after controlling for assignment, mark differs between ETC2410 and ETC3440 students,

$$E(exam|asgmt, etc3440 = 1) \neq E(exam|asgmt, etc3440 = 0)$$

so there is no compelling reason to insist that  $var(\mathbf{u}|\mathbf{X})$  (which is equal to  $var(\mathbf{y}|\mathbf{X})$ ) should be the same across both groups  $\therefore$  we should base our confidence interval of the slope parameter of ETC2410 students using  $se(\hat{\beta}_1)_{with\ etc2410\ students\ only}$  and not  $se(\hat{\beta}_1)_{unrestricted\ with\ both\ etc2410\ and\ etc3440\ sample}$ .