INTRODUCTORY ECONOMETRICS (12156)

February/06

All questions are compulsory.

Part one should be answered, very briefly, in a maximum of two sheets of paper.

PART ONE

1. Suppose that a score on a final exam, *score*, depends on classes attended, *attend*, and unobserved factors that affect exam performance (such as student ability)

$$score = \beta_0 + \beta_1 attend + u$$

When would you expect this model to satisfy $E(u \mid attend) = 0$?

2. In the following example, the OLS fitted line explaining college GPA, *colGPA*, in terms of high school GPA, *hsGPA*, and ACT score, *ACT*, is

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$

If the average high school GPA in the sample is about 3.4 and the average ACT score in the sample is about 24.2, what is the average college GPA in the sample? (**Hint:** Use the algebraic property concerning sample averages of all the variables and the OLS regression line)

3. Suppose you estimate a regression model and obtain $\hat{\beta}_1 = .56$ and p-value $= .086$ for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$. What is the p-value for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 > 0$?

4. In a regression model with a large sample size ($n \to \infty$), what is an approximate 95% confidence interval for $\hat{\beta}_j$ under MLR.1 through MLR.5 (*Gauss-Markov assumptions*)? We call this an **asymptotic confidence interval**.

5. Explain why choosing a model by maximizing $\bar{R}^2$ or minimizing $\hat{\sigma}$ (the standard error of the regression) is the same thing.

PART TWO

6. Let y, p and m denote the real output in 1989 pesetas, the price level and the quantity of money, respectively, all measured in natural logs, in the following models:

$$y = \beta_0 + \beta_1 m + u \quad (1)$$

$$p = \beta_0 + \beta_1 m + w \quad (2)$$

a) Using model (1), state the null hypothesis that money does not affect real output, against the alternative that if money increases, real output increases.
b) Using model (2), state the null hypothesis that the price level is strictly proportional to the quantity of money, against the alternative that if money increases, the increase in the price level is less than proportional.
c) Let

$$\underset{}{\hat{p}} = \underset{(0.015)}{0.003} + \underset{(0.017)}{0.971}\ m$$

be the estimated model (2) for a sample of 62 observations, standard errors in brackets. Test the null hypothesis of question b) above, at the 5% and 10% significance levels.
d) If real output were measured in €, how are the estimates and t-statistics of model (1) affected? (**Hint:** Remember that all the variables are in natural logs.)

7. Let W, *educ*, and *exper* be the nominal monthly wage, years of education and years of experience, respectively. Let *male* be a dummy variable, which takes value one when the individual is male, and zero otherwise. Using data from a recent survey, the following models were estimated:

Dependent Variable: W **Equation 1**
Method: Least Squares
Date: 01/30/06 Time: 08:47
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	106.8320	7.522148	14.20233	0.0000
MALE	14.92081	10.86735	1.372994	0.1704
EDUC	24.93752	0.950112	26.24692	0.0000
EDUC*MALE	15.19665	1.392210	10.91549	0.0000
EXPER	34.63248	0.346800	99.86303	0.0000
EXPER*MALE	15.41967	0.501065	30.77379	0.0000

R-squared	0.986585	Mean dependent var	861.9887
Adjusted R-squared	0.986449	S.D. dependent var	428.3035
S.E. of regression	49.85851	Akaike info criterion	10.66818
Sum squared resid	1228020.	Schwarz criterion	10.71876
Log likelihood	-2661.046	F-statistic	7265.904
Durbin-Watson stat	1.959561	Prob(F-statistic)	0.000000

Dependent Variable: W **Equation 2**
Method: Least Squares
Date: 01/30/06 Time: 08:48
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	106.8655	19.21086	5.562766	0.0000
EDUC	32.51901	2.456772	13.23648	0.0000
EXPER	42.51501	0.885262	48.02531	0.0000

R-squared	0.830946	Mean dependent var	861.9887
Adjusted R-squared	0.830265	S.D. dependent var	428.3035
S.E. of regression	176.4561	Akaike info criterion	13.19000
Sum squared resid	15474974	Schwarz criterion	13.21529
Log likelihood	-3294.501	F-statistic	1221.442
Durbin-Watson stat	2.034541	Prob(F-statistic)	0.000000

Dependent Variable: W **Equation 3**
Method: Least Squares
Date: 01/30/06 Time: 08:49
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	31.39722	7.876566	3.986156	0.0001
MALE	259.9578	6.059557	42.90046	0.0000
EDUC	28.86252	0.942038	30.63838	0.0000
EXPER	42.21498	0.339553	124.3250	0.0000

R-squared	0.973574	Mean dependent var	888.7960
Adjusted R-squared	0.973414	S.D. dependent var	414.9693
S.E. of regression	67.66111	Akaike info criterion	11.27487
Sum squared resid	2270701.	Schwarz criterion	11.30858
Log likelihood	-2814.717	F-statistic	6091.191
Durbin-Watson stat	1.963510	Prob(F-statistic)	0.000000

a) Estimate the wage equation separately for males and females. Are there notable differences in the two estimated equations?
b) Compute the Chow test for equality of the parameters in the wage equation for males and females, specifying the restricted and the unrestricted models, and the relevant degrees of freedom.
c) Now, allow for a different intercept for males and females and determine whether the interaction terms are jointly significant.

INTRODUCTORY ECONOMETRICS (12156)

ANSWERS

February/06

PART ONE

1. When the student ability, motivation, age, health, and other factors in u are not related to attendance, then $\mathrm{E}(u \mid attend) = 0$ would hold. This seems unlikely to be the case.

2. We use the property of OLS concerning sample averages of all the variables and the OLS regression line: when we plug the average values of all independent variables into the OLS regression line, we obtain the average value of the dependent variable[1]. This is,

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \ldots + \hat{\beta}_k \bar{x}_k$$

So in this case,

$$\begin{aligned}\overline{colGPA} &= 1.29 + .453\,\overline{hsGPA} + .0094\,\overline{ACT} \\ &= 1.29 + .453(3.4) + .0094(24.2) \approx 3.06\end{aligned}$$

This can be checked, to the second decimal place, obtaining the average value of *colGPA* in the file GPA1.RAW from Wooldridge´s (2003) book.

3. Because $\hat{\beta}_1 = .56 > 0$ and we are testing against H_1: $\beta_1 > 0$, the one-sided p-value is one-half of the two-sided p-value, or .043.

4. $\hat{\beta}_j \pm 1.96 se(\hat{\beta}_j)$ is the asymptotic 95% confidence interval. Or, we can replace 1.96 with 2.
Note that MLR.6, *Normality*, is not needed for this result.

5. From the definition of $\bar{R}^2$,

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{SST/(n-1)}$$

since $\hat{\sigma}^2 = SSR/(n-k-1)$.
For a given sample, n, and a given dependent variable, $SST/(n-1)$ is fixed. When we use different sets of explanatory variables, only $\hat{\sigma}^2$ changes. As $\hat{\sigma}^2$ decreases, $\bar{R}^2$ increases. If we make $\hat{\sigma}$, and therefore $\hat{\sigma}^2$, as small as possible, we are making $\bar{R}^2$ as large as possible.
Note that if we restrict attention to models with the same number of explanatory variables, k, the same claim can be made about R^2 and *SSR*, so OLS by minimizing the *SSR* is in fact maximizing the R^2.

---

[1] In other words, the point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ is always on the OLS regression line.

PART TWO

6. a) H_0: $\beta_1 = 0$ against H_1: $\beta_1 > 0$.

b) Note that p and m are in natural logs, $p = \log P$ and $m = \log M$ respectively, so β_1 in (2) is the elasticity of prices with respect to money, $\beta_1 = \frac{\%\Delta P}{\%\Delta M}$.

This means that, holding w constant, $\Delta w = 0$, a 1 percent change in the quantity of money, M, will increase the price level, P, by β_1 percent. Strict proportionality is obtained when $\beta_1 = 1$.
Hence H_0: $\beta_1 = 1$ against H_1: $\beta_1 < 1$.

Note that taking exponentials in (2) we obtain

$$P = e^{\beta_0} M^{\beta_1} e^{w}$$

which makes clear that strict proportionality means $\beta_1 = 1$.

c) The t statistic for testing the null hypothesis in b) is

$$t = \frac{0.971 - 1}{0.017} \approx -1.706$$

The critical value for the one sided test at the 5% and the 10% significance level, with 60 degrees of freedom, are –1.671 and –1.296 respectively. In both cases the t statistic, –1.706, is lower than the critical values, –1.671 and –1.296, so we reject H_0: $\beta_1 = 1$ in favour of H_1: $\beta_1 < 1$ in both cases.

Note that rejection at 5% implies rejection at a higher significance level, for example at 10%, but not necessarily at a lower significance level, for example the 1% critical value is –2.390, and in this case we fail to reject H_0: $\beta_1 = 1$ in favour of H_1: $\beta_1 < 1$.

Note also that the t statistic is very close to the critical value at 5%, so the statistical test is almost inconclusive at this significance level. In fact the p-value of the test is just .047 or 4.7%.

d) This involves a change in the scale of the real output, but since the dependent variable in (1) is in log form, $y = \log Y$, only the estimate of β_0 and its t statistic are affected.

Let c represent the constant in going from pesetas to €, so $Y' = \frac{Y}{c}$, then $y' = \log Y' = y - \log(c)$, so the new equation is

$$y' = y - \log(c) = \left(\beta_0 - \log(c)\right) + \beta_1 m + u$$

hence the new intercept is $\beta_0 - \log(c)$.

Note that the change in the origin does not affect neither the variance of β_0 nor the standard error, so the t statistic is reduced by $\frac{\log(c)}{se(\hat{\beta}_0)}$.

7. a) Note that the base group or category is *female*, since the included variable is *male*, which is one for males and zero for females. This means that comparisons are made against females.

Using two decimal places:

<u>Females</u>:

$W = 106.83 + 24.94\ educ + 34.63\ exper$

<u>Males</u>:

$W = (106.83 + 14.92) + (24.94 + 15.20)\ educ + (34.63 + 15.42)\ exper$

$W = 121.75 + 40.13\ educ + 50.05\ exper$

There are in fact notable differences in the estimated equations.

(*i*) The slope dummy variables are highly statistically significant and also practically significant, since the size of the estimated coefficients is quantitatively important.

(*ii*) The intercept dummy variable is statistically significant only at a higher than 17% significance level (p-value of the corresponding t-statistic is 0.1704), which does not provide too much evidence against the null. The magnitude of the coefficient is also relatively small. However because the intercept of the equation is affected by the origin of the variables is not advisable to drop the intercept dummy variable.

b) The Chow statistic can be computed by testing H_0: $\delta_0 = \delta_1 = \delta_2 = 0$ where the δ′s are the parameters associated to the dummy variables in **Equation 1**. Hence the restricted model is

$$W = \beta_0 + \beta_1\ educ + \beta_2\ exper + u$$

estimated as **Equation 2**. And the unrestricted model is

$$W = \beta_0 + \delta_0\ male + \beta_1\ educ + \delta_1\ educ{\cdot}male + \beta_2\ exper + \delta_2\ educ{\cdot}exper + u$$

estimated as **Equation 1**.

This involves the comparison of the SSR of the equations 1 and 2, so the F test is,

$$F = \frac{15474974 - 1228020}{1228020} . \frac{500-6}{3} \approx 1910.39$$

which allows us to reject H_0 at any significance level.

Note that the same statistic could be computed from the R^2 of both regressions,

$$F = \frac{0.986585 - 0.830946}{1 - 0.986585} . \frac{500-6}{3} \approx 1910.44$$

where the difference comes from rounding.

The degrees of freedom, df, of the test are $500 - 6 = 494$, and the number of restrictions $q = 3$.

c) This involves testing the hypothesis H_0: $\delta_1 = \delta_2 = 0$. So we have to compare the *SSR* of the equations 1 and 3, since the restricted model is now,

$$W = \beta_0 + \delta_0\, male + \beta_1\, educ + \beta_2\, exper + u$$

estimated as **Equation 3**.
The F test is

$$F = \frac{2270701 - 1228020}{1228020} . \frac{500-6}{2} \approx 209.72$$

which allows us to reject H_0 at any significance level.

Note that the same statistic could be computed from the R^2 of both regressions,

$$F = \frac{0.986585 - 0.973574}{1 - 0.986585} . \frac{500-6}{2} \approx 239.56$$

where the difference comes from rounding[2].

Overall, **Equation 1** seems to be the preferred model.

[2] Take care that, such a big difference between both ways to calculate the same statistic cannot be due to rounding only, so there is probably an error in the reporting of **Equation 3**. This miss-reporting is in fact evident if you compare the mean and standard deviation of the dependent variable in equation 3 and in equations 1 and 2.
Note however that this does not affect the conclusion about H_0 in anyway.