SOLUTIONS

Ph.D. Econometrics I
Heinz School, Carnegie Mellon University
90-906, Spring 2004

Final

Instructions You may use any books, notes, calculators, and other aids you like. You may not converse, nor may you cooperate.

Please complete all questions.

Please show all relevant work.

Please interpret your results in plain English.

Use the backs of pages if you need more space.

Each part of each question is worth 6 points, except for parts 3a-d which are each worth 7 points.

Typo: Q2: Assume $Cov(\nu, \epsilon) = 0$ for all i, t.

SOLUTIONS

Ph.D. Econometrics I
Heinz School, Carnegie Mellon University
90-906, Spring 2004

Final

Instructions You may use any books, notes, calculators, and other aids you like. You may not converse, nor may you cooperate.

Please complete all questions.

Please show all relevant work.

Please interpret your results in plain English.

Use the backs of pages if you need more space.

Each part of each question is worth 6 points, except for parts 3a-d which are each worth 7 points.

Typo: Q2: Assume $Cov(\nu, \epsilon) = 0$ for all i, t.

1. The Medical Expenditure Panel Survey is an annual survey which collects information about medical expenditures, income, employment, demographics, health information, &c for a representative sample of Americans.

   The data extract in this example is different from the homework extract.

   In this extract, I have chosen a group of people aged 19-64 in 1996 and followed them over two years (1996 and 1997). For each person, there are exactly two observations in the dataset, one for 1996 and one for 1997.

   The following variables appear on the data:

Variable	Meaning
age	age of person in years
age2	age squared
sex	sex of person, 1=male & 0=female
yred	years of education (12=HS diploma)
income	**earned** (wages + profits) income in $
health	perceived health status, higher is sicker
health_??	dummies for perceived health status
	ex=excellent, ... , pr=poor

   We will begin with the regression on page 4 of the output, and we will begin with the assumption that the classical linear regression model is true for these data and this model.

(a) Give an estimate and 95% confidence interval for the effect of a year's education on earned income.

From page 4:

β_{yred}: $3332 \pm 1.96 \cdot 75$

3332 ± 147

(b) Please test the hypothesis that people in good and excellent health have equal income, other things in the regression equal.

PAGE 4:

$\hat{\beta}_{VG} = -961$

std error $= 396$

t-stat $= -2.43$

P-value $= 0.0152$

Since $p < 0.05$, we can reject at 5% level.

(c) Please test the hypothesis that age has no effect on income.

Page 4:

$H_0: \beta_{age} = 0$ and $\beta_{age2} = 0$

$$\text{F-stat} = \left(\hat\beta_{age} \quad \hat\beta_{age2}\right) \begin{bmatrix} 10707 & -123 \\ -123 & 1.44 \end{bmatrix}^{-1} \begin{pmatrix} \hat\beta_{age} \\ \hat\beta_{age2} \end{pmatrix} \Big/ 2$$

$$= \frac{1}{10707(1.44) - 123 \cdot 123} \cdot \frac{1}{2} \cdot \left(2689 \quad -31\right) \begin{bmatrix} 1.44 & 123 \\ 123 & 10707 \end{bmatrix} \begin{pmatrix} 2 \\ -3 \end{pmatrix}$$

$= 352$ (refer to $F_{2,\infty}$ table)

H_0 is rejected at any reasonable significance level

(d) A critic claims that there is heteroskedasticity in this regression. What will the effect of this be if true, and how would you deal with the critique?

If true, parameter estimates are OK, but variance estimates are not. So all CI and tests in a-c are wrong if heteroskedasticity is present.

I would run White's test. If I reject the null of no heteroskedasticity, I will use White/Huber robust standard errors for all my CI and hyp tests.

(e) Is there any particular reason to believe that any of the RHS variables in this regression might be correlated with the error? What effect will this have, if it is true?

If unmeasured ability affects income and education, then yred will be correlated with error.

If income affects health status, say because high income people can afford better/more health care, then the health status dummies will be correlated with the error.

If true all coefficient estimates are biased & inconsistent.

2. A critic complains that the regression in the previous question suffers from the repeated measures problem — that observations from 1996 and 1997 for the same individual are correlated. He proposes a model like this:

$$Y_{it} = X^{it}\beta + \nu_i + \epsilon_{it} \tag{1}$$

Here, i refers to individual i and t refers to the time period. i runs from 1 to N and t from 1 to T, so that there are NT total observations. (Here, t would equal 1 for 1996 and 2 for 1997).

ϵ_{it} is a well-behaved error, so that $V(\epsilon) = \sigma_\epsilon^2 I$.

The other error term is ν_i. It is also a random variable, and it captures the serial correlation within individuals. ν_i has a mean of zero, a variance of σ_ν^2 and is uncorrelated across individuals.

To be absolutely clear, the very same ν_i appears in the two observations for individual i.

(a) What is the covariance between two error terms (error term here means $\nu_i + \epsilon_{it}$) from the same individual?

$$Cov(\nu_i + \epsilon_{i96}, \nu_i + \epsilon_{i97}) =$$

$$Cov(\nu_i, \nu_i) + Cov(\nu_i, \epsilon_{i97}) +$$

$$\cancel{+ Cov(\nu}$$

$$+ Cov(\epsilon_{i96}, \nu_i)$$

$$+ Cov(\epsilon_{i96}, \epsilon_{i97})$$

$$Cov(\nu_i, \nu_i) = \sigma_\nu^2$$

(b) What is the covariance between two error terms from different individuals?

$$Cov(\nu_i + \epsilon_{it}, \nu_j + \epsilon_{jt'}) =$$

$$Cov(\nu_i, \nu_j) + Cov(\nu_i, \epsilon_{jt'}) +$$
$$Cov(\epsilon_{it}, \nu_j) +$$
$$Cov(\epsilon_{it}, \epsilon_{jt'})$$

$$= 0$$

(c) Does this model create a serial correlation problem? What would be its consequence for the answers to the previous question?

Yes, since some of the covariances between errors are $\neq 0$ there is a serial correlation problem.

OLS estimates are OK, but all CI, hypothesis tests, and std errors are wrong.

(d) A helpful fellow researcher suggests that you can eliminate your serial correlation problem by only using a single year of data — as in the regressions in the output labeled "96 only" or "97 only". That is she suggests you throw away half the data. Will this fix the problem?

Yes. With, say, only 1996 data all Covariances between different observations' errors become 0. Recall from a) and b) that the only problematic correlations are:

$Cov(\nu_i + \epsilon_{i1996}, \nu_i + \epsilon_{i1997})$

Since $\nu_i + \epsilon_{i1997}$ is not in the data when we only use 1996 data, there is no problem any more.

(e) Another helpful fellow researcher suggests that you can eliminate your serial correlation problem by averaging your data. That is, have one observation per person, with the LHS variable being the average of that person's LHS variable for the years 1996 and 1997 and each of the RHS variables being the average of the respective RHS variable for 1996 and 1997. Like the regression in the output labeled 96/7 averaged. Will this fix the problem?

Yes. Now we have:

$$\bar{Y}_i = \bar{X}^i \beta + v_i + \bar{\epsilon}_i$$

$$Cov(v_i + \bar{\epsilon}_i, v_j + \bar{\epsilon}_j) =$$

$$Cov(v_i, v_j) + Cov(v_i, \bar{\epsilon}_j) + Cov(\bar{\epsilon}_i, v_j) + Cov(\bar{\epsilon}_i, \bar{\epsilon}_j)$$

All these are 0.

Again, the problem only arises ~~w/~~ between 2 obs from the same individual. Here there is only 1 obs per individual.

So, no problem.

(f) Which of these two methods is better and why?

The second is better.

Intuitively, it does not throw away data & therefore information.

Practically, looking at pages 8, 10, 14, the standard errors are smaller for the averaged data.

Theoretically, the error terms in the "throw away" case are $\nu_i + \epsilon_{i1996}$ with a variance of $\sigma_\nu^2 + \sigma_\epsilon^2$. In the "average" case, the error terms are $\cancel{\sigma_\nu^2 +}$ $\nu_i + \frac{1}{2}(\epsilon_{i1996} + \epsilon_{i1997})$ with a variance of $\sigma_\nu^2 + \frac{1}{2}\sigma_\epsilon^2$. Lower error variance is better

$(V(\hat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1})$

3. Consider the following very simple regression models:

$$Y = \beta_1 + \beta_2 * \text{Female} + \epsilon \tag{2}$$

$$Y = \beta_3 + \beta_4 * \text{Male} + \epsilon \tag{3}$$

For the parts of this question, you can get full credit only with proofs of your answers.

(a) What will be the relationship between $\hat{\beta}_{2,OLS}$ and $\hat{\beta}_{4,OLS}$?

$$\hat{\beta}_{2OLS} = \frac{\sum (F-\bar{F})(Y-\bar{Y})}{\sum (F-\bar{F})^2} \quad \#$$

$$\hat{\beta}_{4OLS} = \frac{\sum (M-\bar{M})(Y-\bar{Y})}{\sum (M-\bar{M})^2}$$

$$M = 1-F$$
$$\bar{M} = 1-\bar{F}$$

$$\hat{\beta}_{4OLS} = \frac{\sum \left((1-F)-(1-\bar{F})\right)(Y-\bar{Y})}{\sum \left((1-F)-(1-\bar{F})\right)^2}$$

$$= \frac{\sum (-F+\bar{F})(Y-\bar{Y})}{\sum (-F+\bar{F})^2}$$

$$= -\frac{\sum (F-\bar{F})(Y-\bar{Y})}{\sum (F-\bar{F})^2}$$

$$= -\hat{\beta}_{2OLS}$$

(b) What will be the relationship between $V(\hat{\beta}_{2,OLS})$ and $V(\hat{\beta}_{4,OLS})$?

$$V(\hat{\beta}_{2\,OLS}) = \sigma^2 / \Sigma(F - \bar{F})^2$$

$$= \sigma^2 / \Sigma\big((1-M) - (1-\bar{M})\big)^2$$

$$= \sigma^2 / \Sigma(-M + \bar{M})^2$$

$$= \sigma^2 / \Sigma(M - \bar{M})^2$$

$$= V(\hat{\beta}_{4\,OLS})$$

(c) What will be the relationship between the $\hat{Y}$ from the two regressions?

We know $\hat{\beta}_{2OLS} = -\hat{\beta}_{4OLS}$.

We also know:

$$\hat{\beta}_{1OLS} = \bar{Y} - \hat{\beta}_{2OLS}\bar{F}$$

$$\begin{aligned}
\hat{\beta}_{3OLS} &= \bar{Y} - \hat{\beta}_{4OLS}\bar{M} \\
&= \bar{Y} + \hat{\beta}_{2OLS}(1-\bar{F}) \\
&= \bar{Y} - \hat{\beta}_{2OLS}\bar{F} + \hat{\beta}_{2OLS} \\
&= \hat{\beta}_{1OLS} + \hat{\beta}_{2OLS}
\end{aligned}$$

From M:

$$\begin{aligned}
\hat{Y} &= \hat{\beta}_{3OLS} + \hat{\beta}_{4OLS} M \\
&= \hat{\beta}_{1OLS} + \hat{\beta}_{2OLS} - \hat{\beta}_{2OLS}(1-F) \\
&= \hat{\beta}_{1OLS} + \hat{\beta}_{2OLS} F \\
&= \hat{Y} \text{ from } F
\end{aligned}$$

(d) What will be the relationship between R^2 from the two regressions and $\hat{\sigma}^2_{OLS}$ from the two regressions?

If $\hat{Y}$ is the same, then $\hat{\sigma}^2_{OLS}$ must be the same, since $\hat{\sigma}^2_{OLS} = \frac{1}{N-2} e'e$ and $e = Y - \hat{Y}$.

Similarly, if $\hat{Y}$ is the same then Regression Sum of Squares is the same and thus $R^2 = \frac{RSS}{TSS}$ is the same.

(e) What will be the relationship between $\hat{V}(\hat{\beta}_{2,OLS})$ and $\hat{V}(\hat{\beta}_{4,OLS})$?

$$\hat{V}(\hat{\beta}_{2\,OLS}) = \hat{\sigma}^2_{OLS} / \Sigma(F-\bar{F})^2$$

$$\hat{V}(\hat{\beta}_{4\,OLS}) = \hat{\sigma}^2_{OLS} / \Sigma(M-\bar{M})^2$$

We already saw in b) that denominators are =.
We saw in d) that numerators are equal, too.
Thus they are equal:

$$\hat{V}(\hat{\beta}_{2\,OLS}) = \hat{V}(\hat{\beta}_{4\,OLS}).$$