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.

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.

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

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

(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?

(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?

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?

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

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

(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?

(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?

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

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 $\hat{\beta}_{4,OLS}$?

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

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

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

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