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

Midterm, **Solution**

| | |
|---|---|
| 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 worth 10 points, except for 4a, which is worth 20. |

1. Consider the following regression model in which we assume that all the classical assumptions are satisfied:

$$Y = \beta_1 X + \epsilon \tag{1}$$

Consider the following estimator for $\beta_1$:

$$\widetilde{\beta}_1 = \frac{\sum X_i^2 Y_i}{\sum X_i^3}$$

(a) Is $\widetilde{\beta}_1$ unbiased?

**Yes, it is. Observe:**

$$\begin{aligned}
\widetilde{\beta}_1 &= \frac{\sum X_i^2 Y_i}{\sum X_i^3} \\
&= \frac{\sum X_i^2 (\beta_1 X_i + \epsilon_i)}{\sum X_i^3} \\
&= \frac{\sum X_i^2 (\beta_1 X_i)}{\sum X_i^3} + \frac{\sum X_i^2 \epsilon_i}{\sum X_i^3} \\
&= \beta_1 \frac{\sum X_i^2 (X_i)}{\sum X_i^3} + \frac{\sum X_i^2 \epsilon_i}{\sum X_i^3} \\
&= \beta_1 + \frac{\sum X_i^2 \epsilon_i}{\sum X_i^3} \\
E\left(\widetilde{\beta}_1\right) &= \beta_1 + \frac{\sum X_i^2 E(\epsilon_i)}{\sum X_i^3} \\
&= \beta_1
\end{aligned}$$

(b) What is the variance of $\widetilde{\beta}_1$?

$$
\begin{aligned}
V\left(\widetilde{\beta}_1\right) &= V\left(\beta_1 + \frac{\sum X_i^2 \epsilon_i}{\sum X_i^3}\right) \\
&= V\left(\frac{\sum X_i^2 \epsilon_i}{\sum X_i^3}\right) \\
&= \frac{V\left(\sum X_i^2 \epsilon_i\right)}{\left(\sum X_i^3\right)^2} \\
&= \frac{\sum X_i^4 V(\epsilon_i) + \underset{\sum\sum X_i^2 X_j^2 Cov(\epsilon_i, \epsilon_j)}{i \neq j}}{\left(\sum X_i^3\right)^2} \\
&= \frac{\sum X_i^4 \sigma^2}{\left(\sum X_i^3\right)^2} \\
&= \sigma^2 \frac{\sum X_i^4}{\left(\sum X_i^3\right)^2}
\end{aligned}
$$

(c) Is the variance of the OLS estimator greater than, less than, or the same as the variance of $\widetilde{\beta}_1$?

**Well, $\widetilde{\beta}_1$ is a linear unbiased estimator of $\beta_1$. Thus, by the Gauss-Markov Theorem, its variance must be no less than the variance of $\widehat{\beta}_{1,OLS}$. And, since the two generally have different variances, generally OLS will have a lower variance.**

2. Consider a factory producing pairs of shoes. We have data on the monthly costs of running the factory and the number of shoes produced. We want to estimate the following model by OLS:

$$\text{Cost}_i = \beta_1 + \beta_2 \text{Left Shoes} + \beta_3 \text{Right Shoes} + \epsilon$$

Is it a good plan to try to estimate this model by OLS? Why or why not?

**It is a bad idea. Since shoes are produced in pairs, Right Shoes = Left Shoes. This is a linear dependence in $X$. This violates one of the assumptions of the CLRM. Furthermore, it means that OLS cannot be estimated, since $X'X$ is not invertible.**

**Notice a critical distinction. It may be that left shoes are more expensive to produce than are right shoes, so that the model above may make a kind of sense. However, since shoes are always produced in pairs, there is no way we can estimate the difference in the shoe costs this way. If we could find a factory which did not always produce shoes in pairs, we could find out about the difference in costs for left and right shoes.**

3. Suppose you are interested in trying to predict how much money people will spend on their next new car. You have a dataset documenting how much a large, relevant sample of people spent on their cars and how much they characteristically spend on a bottle of wine. So, you estimate the following model by OLS:

$$P_{\text{car}} = \beta_1 + \beta_2 P_{\text{wine}} + \epsilon$$

Then, for a similar group of people, you are informed of each person's $P_{\text{wine}}$, and you use your estimates of $\beta_1$ and $\beta_2$ to predict car purchase price, $P_{\text{car}}$.

A critic complains that the model above does not conform to the assumptions of the classical linear regression model. Is the critic right about this?

The critic goes on to claim that, since the assumptions of CLRM are violated, you should not use OLS for these purposes. Assuming that the critic is right about the assumption violation, is she right about the uselessness of OLS here?

**The critic is right about the assumption violation. Both the price of wine and the price of the car are caused by underlying characteristics of the person (say income and/or taste for luxury). However, the critic is wrong about the use of OLS. All we are trying to do here is *predict* car price based on wine price. Since OLS is essentially always the best linear predictor, it is perfectly appropriate to use it here.**

**This is an important distinction to get right. If all we care about is predicting $Y$ using $X$ and we have no good reason to believe that a linear predictor is a bad thing, we should use OLS. If, on the other hand, we are trying to discover the size of a causal relationship between $Y$ and $X$ (that is, we care about the magnitude and meaning of $\beta$) we have a persuasive argument in favor of using OLS only when the CLRM is true.**

4. 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.

   I have prepared an extract of these data for 1996. The following variables appear on the data:

| Variable | Meaning |
|---:|---:|
| age | age of person in years |
| sex | sex of person, 1=male & 0=female |
| income | income in 1996 $ |
| employed | 1=employed, 0=not employed |
| insured | 1=had health insurance, 0=not |
| health | perceived health status, higher is sicker |
| spending | spending on health care, 1996 $ |

(a) Consider the regression on page 5 of the output. Please interpret each of the estimates there, excluding the constant. Which ones "make sense" intuitively?

**They all make intuitive sense. A one dollar increase in income (holding const the other variables) is associated with a 0.5 cent increase in health spending. A one year increase in age (...) is associated with a 23 dollar increase in health spending. Male sex is associated with 48 dollars less in health spendingl. Employment is associated with 1502 dollars less in health spending (sensible since very sick people are usually not employed). Insurance is associated with 1588 dollars more in health spending (sensible since you consume more of something that is less expensive to you and insurance lowers the price of health spending). Finally, people who perceive themselves as sicker spend more.**

(b) For which of the coefficients can we reject, at the 5% level, the null hypothesis that the coefficient is equal to zero?

**Look for t value $\geq$ 1.96 or P value $<$ 0.05. This is true for age, employment, insurance, and health status.**

(c) Suppose an unemployed and uninsured person took a job with health insurance. What is your best estimate of the change in their health spending? Can we conclude at the 10% level that this change is different from 0?

**My best estimate (best by G-M Thm) would be 1588-1502=86. To make a 90% CI, I need the variance of** $\widehat{\beta}_{emp,OLS} + \widehat{\beta}_{ins,OLS}$ **and this variance is equal to** $V\left(\widehat{\beta}_{emp,OLS}\right) + V\left(\widehat{\beta}_{ins,OLS}\right) + 2Cov\left(\widehat{\beta}_{emp,OLS}, \widehat{\beta}_{ins,OLS}\right)$**.**

**Looking at the output on page 5, we see that** $V\left(\widehat{\beta}_{emp,OLS}\right) = 44934$**,** $V\left(\widehat{\beta}_{ins,OLS}\right) = 48773$**, and** $Cov\left(\widehat{\beta}_{emp,OLS}, \widehat{\beta}_{ins,OLS}\right) = -556$**.**

**Thus, the variance of the sum is** $44934 + 48773 - 1112 = 92595$**, and the standard error of the sum is 304. So, our 90% confidence interval is** $86 \pm 1.65(304)$ **or** $86 \pm 502$**. Since zero is inside this confidence interval, we cannot conclude at the 10% significance level that the effect of going from unemployed and uninsured to employed and insured on health spending is different from zero.**

(d) If we were to regress spending on income alone, would the slope coefficient be positive or negative and why?

**Negative. Look at page 3 at the correlation matrix. The correlation between income and spending is -0.02. Since the bivariate regression coefficient for the slope has the same sign as the simple correlation, we know that the bivariate regression coefficient will be negative.**