Intro to Econometric Theory
Heinz School, Carnegie Mellon University
90-906, Spring 2005-6

Solution #2

1. **Consider a sample of people who were asked to report their sex (Everyone answers, and everyone answers either male or female). From their answers, we construct the following matrix:**

$$X = \begin{bmatrix} \mathbf{1} & M & F \end{bmatrix} \tag{1}$$

   In the matrix, each row corresponds to one person's response; 1 is a column of 1s, M is a column with a 1 in the ith row if the ith person answered male and a zero otherwise, and F is a column with a 1 in the ith row if person i answered female and a zero otherwise.

   Is $X'X$ invertible. Why or why not?

   No, $X'X$ is not invertible. Notice that $M + F = \mathbf{1}$. This is a linear dependency in the columns of X. Thus, the rank of X must be less than three. Since the rank of $X'X$ (a 3×3 matrix) is the same as the rank of X, its rank must be less than 3. Thus, it is not full rank. Thus it is not invertible.

2. **We discussed in class that variance matrixes are always positive semi-definite but not always positive definite. In the case of scalar random variables, a singular variance matrix means a zero variance — that is, that the random variable is not really random at all, but fixed.**

   Now, let's think about what a positive semi-definite but not positive definite (hereafter PSDNPD) variance matrix means for the random vector it corresponds to. Let x be a random vector with PSDNPD variance matrix V.

   (a) **Must there be a fixed non-zero vector α, such that $\alpha' V \alpha$=0? Why or why not?**

   Yes. If there were no such vector, then V would be positive definite, and we are told that it is not.

   (b) **Suppose there is such an α, what would the variance of $\alpha' x$ be?**

   Zero. $V(\alpha' x) = \alpha' V \alpha = 0$.

(c) **What does this tell us about $\alpha' x$?**

It is fixed, non-random.

(d) **Does this mean that $\alpha' x = 0$?**

No, it could just as easily be 8 or 0.73 or any fixed number. The variance of 0 is zero, but the variance of 8 or 0.73 are is zero.

(e) **So, what does it mean for a random vector to have a PSD-NPD variance matrix?**

When a random vector has a PSDNDP variance matrix, this means that there are, in some sense, fewer random variables in the vector than appear to be. Consider a scalar random variable z and a random vector:

$$x = \begin{pmatrix} z \\ -z \end{pmatrix}$$

As you can see, although x has two elements it does not *really* contain two random variables, just z.

The variance matrix of x is:

$$V(x) = \begin{bmatrix} \sigma_z^2 & -\sigma_z^2 \\ -\sigma_z^2 & \sigma_z^2 \end{bmatrix}$$

This matrix is singular, and:

$$\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{bmatrix} \sigma_z^2 & -\sigma_z^2 \\ -\sigma_z^2 & \sigma_z^2 \end{bmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0$$

and:

$$V((1 \quad 1)x) = 0$$

Another way of saying the same thing is that if V is PSDNPD, then you can always write one of the elements of x as a linear function of the other elements plus a fixed number (this follows from $V(\alpha' x) = 0$).

3. **You are interested in the effect of going to the emergency room on a person's life expectancy. There is a sample of people collected on a particular day, some of whom went to an emergency room and some of whom did not. They are followed over time until they die. From each person there are two items in the dataset: T_i the length of time from the day of the survey until death and ER_i a variable which equals one if they went to the emergency room and zero if they did not. You wish to run a regression, but the people who conducted the study will not let you have the data, only summary statistics. What should you ask for and how will you use it?**

I can run the regression using only $\overline{T}$, $\overline{ER}$, $\widehat{\sigma_T^2}$, $\widehat{\sigma_{ER}^2}$, and $\widehat{Cov}(T, ER)$. Actually, if I only want to regress T on ER, I could do without $\widehat{\sigma_{ER}^2}$. Consider the regression model $T = \beta_1 + \beta_2 ER + \epsilon$. I will use the sample statistics to calculate the regression coefficients as follows:

$$\hat{\beta}_{2,OLS} = \frac{\widehat{Cov}(T, ER)}{\widehat{\sigma^2_{ER}}}$$
$$\hat{\beta}_{1,OLS} = \overline{T} - \hat{\beta}_{2,OLS}\overline{ER} \tag{2}$$

4. **Considering the example in problem 3, consider two linear models:**

   (a) $T_i = \beta_1 + \beta_2 ER_i + \epsilon_i$

   (b) $ER_i = \beta_3 + \beta_4 T_i + \epsilon_i$

   Let's call the OLS estimates you would get from regressions run on these equations $\widehat{\beta}_{1,OLS}, \widehat{\beta}_{2,OLS}, \ldots$. What relationship will there be between $\widehat{\beta}_{2,OLS}, \widehat{\beta}_{4,OLS}$? What sign ($\pm$) do you expect for $\widehat{\beta}_{2,OLS}, \widehat{\beta}_{4,OLS}$?

   Just by examining the formula for OLS estimates, we can see that the relationship will be:

$$\hat{\beta}_{2,OLS} = \hat{\beta}_{4,OLS}\frac{\widehat{\sigma^2_T}}{\widehat{\sigma^2_{ER}}} \tag{3}$$

   Since people who go to the emergency room are probably sick, I expect $Cov(T, ER) < 0$ and therefore that both $\hat{\beta}_{2,OLS} < 0$ and $\hat{\beta}_{4,OLS} < 0$.

5. **We usually interpret regression models as embodying some causal story (the RHS causing the LHS); however, OLS does not "know" this: it produces estimates regardless of whether our implicit causal story is right. Discuss this fact in the context of the models in question 4 and your expectations about the signs of the relevant coefficients.**

   Well, the model in question 4a would produce the result that emergency rooms kill people if we were to take its implicit causal story seriously; whereas, the model in question 4b would produce the result that sick people go to the emergency room if we were to take its implicit causal story seriously. Assuming for the moment that we don't believe that emergency rooms kill people, obviously we can go astray if we don't think carefully about what causal model we are assuming when we write down and interpret regression models. That is, if we wrote down, estimated, and interpreted the model of question 4a we would be led to the (false) conclusion that emergency rooms kill people; whereas, if we wrote down, estimated, and interpreted the model of question 4b we would be led to the (true) conclusion that sick people go to emergency rooms.

Regressions only *use* correlation in making their estimates; therefore, they can not *tell* you about causation. You must already have a causal theory before you run your regression if you want to draw causal conclusions from your results.

6. **You may recall that, in the bivariate regression model,**

$$Y = \beta_1 + \beta_2 X_2 + \epsilon \tag{4}$$

the OLS estimate of β_2 is $\hat{\beta}_{2,OLS} = \frac{\widehat{Cov}(Y,X_2)}{\widehat{V(X_2)}}$. This means that if the correlation (or covariance) between Y and X_2 is positive, then $\hat{\beta}_2$ is positive. Consider now a slightly more complex model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{5}$$

Let's suppose that $\overline{Y} = \overline{X_1} = \overline{X_2} = 0$. Please calculate an expression for $\hat{\beta}_{1,OLS}$ and $\hat{\beta}_{2,OLS}$ in terms of the variances and covariances of Y, X_1, X_2. Using your formula, discuss whether or not it can happen that $\widehat{Cov}(Y, X_1) > 0$ while $\hat{\beta}_{1,OLS} < 0$. If it cannot, why not? If it can, then how? Your goal is both to construct a correct argument and to explain in *intuitive* English why your result holds.

The formula for OLS says:

$$\begin{aligned}
\begin{pmatrix} \hat{\beta}_{1,OLS} \\ \hat{\beta}_{2,OLS} \end{pmatrix} &= \begin{bmatrix} \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 X_1 & \sum X_2^2 \end{bmatrix}^{-1} \begin{pmatrix} \sum Y X_1 \\ \sum Y X_2 \end{pmatrix} \\
&= \begin{bmatrix} \widehat{V(X_1)} & \widehat{Cov}(X_1, X_2) \\ \widehat{Cov}(X_1, X_2) & \widehat{V(X_2)} \end{bmatrix}^{-1} \begin{pmatrix} \widehat{Cov}(Y, X_1) \\ \widehat{Cov}(Y, X_2) \end{pmatrix} \\
&= \frac{1}{\widehat{V(X_1)}\widehat{V(X_2)} - \widehat{Cov}(X_1, X_2)^2} * \\
&\quad \begin{bmatrix} \widehat{V(X_2)} & -\widehat{Cov}(X_1, X_2) \\ -\widehat{Cov}(X_1, X_2) & \widehat{V(X_1)} \end{bmatrix} \begin{pmatrix} \widehat{Cov}(Y, X_1) \\ \widehat{Cov}(Y, X_2) \end{pmatrix} \\
&= \frac{1}{\widehat{V(X_1)}\widehat{V(X_2)} - \widehat{Cov}(X_1, X_2)^2} * \\
&\quad \begin{pmatrix} \widehat{Cov}(Y, X_1)\widehat{V(X_2)} - \widehat{Cov}(Y, X_2)\widehat{Cov}(X_1, X_2) \\ \widehat{Cov}(Y, X_2)\widehat{V(X_1)} - \widehat{Cov}(Y, X_1)\widehat{Cov}(X_1, X_2) \end{pmatrix}
\end{aligned}$$

Between the first and second equals signs I multiply the $(X'X)^{-1}$ matrix by N and the $X'Y$ matrix by $1/N$.

Since $X'X$ is positive definite, its determinant (the fraction above) is positive; this result is in the book, pg 47. So, $\hat{\beta}_{1,OLS}$ will have the same sign as $\widehat{Cov}(Y, X_1)\widehat{V(X_2)} - \widehat{Cov}(Y, X_2)\widehat{Cov}(X_1, X_2)$. By looking at the first term, we can see that one factor tending to make $\hat{\beta}_{1,OLS}$ positive would be $\widehat{Cov}(Y, X_1)$ positive, since $\widehat{V(X_2)}$ must be positive.

However, it is possible for this factor to be overturned by the second factor in the sum, $-\widehat{Cov}(Y, X_2)\widehat{Cov}(X_1, X_2)$. If, for example, the covariance between Y and X_2 is also positive and the covariance between X_1 and X_2 is positive then it could happen that $\hat{\beta}_{1,OLS}$ is negative.

So, $\hat{\beta}_{1,OLS}$ could be negative even though the covariance between Y and X_1 is positive if there is a strong positive relationship between Y and X_2 and between X_1 and X_2. This would happen if all three of Y, X_1, X_2 moved together, obviously. This might happen, for example, if increases in X_2 "drive" both increases in Y and X_1 but X_1 itself has a negative influence on Y.

Consider the following example. The covariance (across countries or over time) between life expectancy and the consumption of red meat per capita is positive. Thinking of life expectancy as Y and meat/capita as X_1 and income as X_2, you can see how this kind of situation can come about (assuming for purposes of argument we believe that red meat kills you prematurely). People often call a variable playing the role of X_2 here a "confounder" because it hides the true relationship between Y and X_1.