HG
Oct. 14

ECON 4130 H14

Extra exercises for no-seminar week 42

(Solutions will be put on the net at the end of the week)

Exercise 1. Show that the sample correlation, $r = \frac{S_{XY}}{S_X S_Y}$ is a consistent estimator for the population correlation, $\rho = \text{corre}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$, based on a random sample, $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ (meaning that the *n* pairs are independent and have all the same joint distribution), where $E(X_i)=\mu_1$, $E(Y_i)=\mu_2$, $\text{var}(X_i)=\sigma_1^2$, $\text{var}(Y_i)=\sigma_2^2$, and $\text{cov}(X_i,Y_i)=\sigma_{12}$.

[Hint: The consistency of S_X, S_Y for σ_1,σ_2 respectively has been proven in the lecture (or see **example 3** in the lecture notes to Rice chapter 5). To prove the consistency of the sample covariance, write

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \bar{X}\bar{Y}\right]$$

(The last equality you can prove in the same way as in **exercise 3a** from seminar week 41)

Note that, since (X_i,Y_i) are *iid* pairs, then the rv's $Z_i = X_iY_i$, $i=1,2,\ldots,n$, must also be *iid*. Find first $E(Z_i)$ expressed by σ_{12},μ_1,μ_2. Then use the note just before section **1.2** in "lecture notes to chapter 5". Also, don't forget the continuity properties of limits in probability.]

Exercise 2

In this exercise we will study what happens when the explanatory variable in a regression model is observed with error. An example of this was given in the exercise from no-seminar week 39, where we tried to explain the expenditure, Y, by the income, *X*, based on micro data from Hong Kong, but where, in lack of true income data, we used X = total expenditure instead. Hence, the explanatory variable, income, was observed with error - as it often is in econometric models.

Let (X_i, Y_i, Z_i) $i = 1, 2, \ldots, n$ be n *iid* triples of rv's, having a common joint pdf, $f(x, y, z)$. (This implies that there is independence between variables from different triples, although there may be dependencies between X_i, Y_i, Z_i for the same *i*.)

To fix ideas imagine that *i* refers to household no. *i* in a random sample of households drawn from a certain large population. Assume further that for household *i*

Y_i is the observed expenditure (in a given period)
Z_i is "the true income" (not directly observable)
X_i is the observed "income" (i.e., total expenditure)

We now assume a simple regression relationship between Y_i and Z_i

(1) $Y_i = \alpha + \beta Z_i + e_i \qquad i = 1, 2, \ldots, n$

where α, β are unknown constants and the error term, e_i, is assumed to satisfy

(2) $\mathrm{E}(e_i \mid z_i) = 0$ and $\mathrm{var}(e_i \mid z_i) = \sigma^2$ (implying $\mathrm{cov}(e_i, Z_i) = 0$ as in (12) in the week-39 exercise).

(1) and (2) constitutes our econometric model. The task is to estimate β from the information in the observed data ((X_i, Y_i) $i = 1, 2, \ldots, n$). The problem here is that we don't know the values of Z_i (a non-observable variable is often called a *latent* variable in econometric literature). Instead we observe X_i which we assume is near Z_i but with some (random) error, expressed by the following assumption

(3) $X_i = Z_i + v_i$ where $\mathrm{E}(v_i \mid z_i) = 0$ and $\mathrm{var}(v_i \mid z_i) = \sigma_v^2$. In addition we assume that v_i and e_i are uncorrelated since, intuitively, there is no reason to expect any dependence between the regression error, e_i, and the error in measuring *Z*. I.e., $E(e_i v_i) = \mathrm{cov}(e_i, v_i) = 0$.

Substituting (3) in (1), we get

$$Y_i = \alpha + \beta(X_i - v_i) + e_i = \alpha + \beta X_i + (e_i - \beta v_i)$$

Hence

(4) $Y_i = \alpha + \beta X_i + u_i$ where $u_i = e_i - \beta v_i$ is an error term.

a. Let μ_X, μ_Y, μ_Z denote expected values and $\sigma_X^2, \sigma_Y^2, \sigma_Z^2$ variances of *X, Y, Z* respectively. The week-39 exercise (10)-(12) shows that the error terms e_i, v_i, u_i all have expected value 0 (why?), which implies (why?) that $\mu_Y = \alpha + \beta \mu_X$.

b. Show that u_i and X_i are correlated, i.e. show that

(5) $\text{cov}(u_i, X_i) = \text{E}(u_i X_i) = -\beta\sigma_v^2$

c. We are interested to estimate β in particular. Using the ordinary least squares (OLS) method, we get the OLS estimator (as in the week-39 exercise):

$\hat{\beta} = \dfrac{S_{XY}}{S_X^2}$ where S_{XY}, S_X^2 are the usual sample estimators for the covariance and variance respectively. As in exercise 1 we obtain (explain why):

$$\hat{\beta} \underset{n\to\infty}{\overset{P}{\to}} \frac{\text{cov}(X_i, Y_i)}{\sigma_X^2}$$

Now show that $\text{cov}(X_i, Y_i) = \beta(\sigma_X^2 - \sigma_v^2)$.

[**Hint:** $\text{cov}(X_i, Y_i) = \text{E}(Y_i - \mu_Y)(X_i - \mu_X) = \text{E}(\alpha + \beta X_i + u_i - \alpha - \beta\mu_X)(X_i - \mu_X) =$
$= \cdots\cdots$fill in $\cdots\cdots = \beta(\sigma_X^2 - \sigma_v^2)$]

Then explain why

$$\hat{\beta} \underset{n\to\infty}{\overset{P}{\to}} \beta\left(1 - \frac{\sigma_v^2}{\sigma_X^2}\right)$$

Hence the OLS estimator $\hat{\beta}$ is an inconsistent estimator for β unless $\sigma_v^2 = \text{var}(v_i) = 0$ (in which case surely $v_i = 0$, i.e. $P(v_i = 0) = 1$; see **B(iii)** in the week-39 exercise). If $\sigma_v^2 > 0$, the OLS estimator is biased in terms of probability limits. Since the bias, $1 - \dfrac{\sigma_v^2}{\sigma_X^2} < 1$, $\hat{\beta}$ tends to underestimate β. We have thus shown that the OLS estimator $\hat{\beta}$ in a simple regression model is consistent if and only if the explanatory variable, *X*, can be observed without error.

d. What if the response variable, *Y*, is observed with error while the explanatory variable, *Z*, is observed without error? Will this also lead to biased regression estimates? To be a little more precise, let Y_i^* denote the true (and not observed) response variable and Y_i the observed one. The model now is

$Y_i = Y_i^* + v_i$, where the errors $v_i, i = 1, 2, \ldots$ are *iid* and independent of the Z_i's, and
$Y_i^* = \alpha + \beta Z_i + e_i$

Answer the question under this model.

[**Note:** The assumption (2) implies (as in the week-39 exercise)

(6) $\mathrm{E}(e_i) = 0, \quad \mathrm{var}(e_i) = \sigma^2, \quad \text{and} \ \ \mathrm{cov}(e_i, Z_i) = 0$

Likewise, the assumption (3) implies

(7) $X_i = Z_i + v_i \quad \text{where} \quad \mathrm{E}(v_i) = 0, \quad \mathrm{var}(v_i) = \sigma_v^2, \quad \text{and} \ \ \mathrm{cov}(v_i, Z_i) = 0.$

The assumptions (6) and (7) are slightly weaker than assumptions (2) and (3) respectively (i.e., we cannot prove (2) and (3) from (6) and (7) without extra assumptions). On the other hand, if we replace (2) and (3) by (6) and (7), we can still prove the limit results above by the same arguments as above. This is maybe the main reason why (6) and (7) are more common in econometric literature as assumptions in connection with the simple regression model than (2) and (3).]