Elements of Econometrics. Lecture 11. Stochastic Regressors. Measurement Errors.

FCS, 2022-2023

NONSTOCHASTIC AND STOCHASTIC REGRESSORS

In the Model A we assumed that the explanatory variables in a regression model are nonstochastic (not having random components). Reason: to simplify the analysis of the properties of the regression estimators.

Now: we progress to Model B (the case of stochastic regressors).

New key assumption: *the values of the regressors are drawn randomly from certain populations.*

This is much more realistic framework for regressions with cross-sectional data.

It is *not* assumed that the regressors are independent of each other. There are joint probability distributions for populations, and their sample values may be correlated.

Another new assumption: *the regressors are distributed independently from the disturbance term.*

All other Model B assumptions are similar to those of the Model A.

STOCHASTIC REGRESSORS. ASSUMPTIONS FOR MODEL B

MODEL B ASSUMPTIONS

B.1 The model is linear in parameters and correctly specified.

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

B.2 The values of the regressors are drawn randomly from fixed populations

B.3 There does not exist an exact linear relationship among the regressors

B.4 The disturbance term has zero expectation (Gauss-Markov 1)

B.5 The disturbance term is homoscedastic (Gauss-Markov 2).

B.6 The values of the disturbance term have independent distributions (Gauss-Markov 3)

B.7 The disturbance term is distributed independently of the regressors (Gauss-Markov 4)

B.8 The disturbance term has a normal distribution

MODEL B. PROPERTIES OF THE REGRESSION

COEFFICIENTS: UNBIASEDNESS

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \beta_2 + \sum a_i u_i \qquad a_i = \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$$

$$\mathrm{E}(\hat{\beta}_2) = \beta_2 + \mathrm{E}\left(\sum a_i u_i\right) = \beta_2 + \sum \mathrm{E}(a_i u_i)$$

$$\mathrm{E}\{f(X)g(Y)\} = \mathrm{E}\{f(X)\}E\{g(Y)\} \qquad f(X) = a_i,\ g(Y) = u_i$$

$$\mathrm{E}(a_i u_i) = \mathrm{E}(a_i)\mathrm{E}(u_i) = 0$$

$$\mathrm{E}(\hat{\beta}_2) = \beta_2 + \sum \mathrm{E}(a_i)\mathrm{E}(u_i) = \beta_2$$

Since E(u_i) = 0 for all *I* (Assumption B.4), $\hat{\beta}_2$ is unbiased (assuming E(a_i) exists). There must be some variation in *X* in the sample (Assumption B.3). Otherwise $\hat{\beta}_2$ would not exist.

MODEL B. PROPERTIES OF THE REGRESSION COEFFICIENTS: CONSISTENCY

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \beta_2 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2} = \beta_2 + \sum a_i u_i$$

$$\text{plim}\,\hat{\beta}_2 = \beta_2 + \text{plim}\left(\frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2}\right)$$

$$= \beta_2 + \text{plim}\left(\frac{\frac{1}{n}\sum(X_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum(X_i - \bar{X})^2}\right)$$

$$= \beta_2 + \frac{\text{plim}\left(\frac{1}{n}\sum(X_i - \bar{X})(u_i - \bar{u})\right)}{\text{plim}\left(\frac{1}{n}\sum(X_i - \bar{X})^2\right)} = \beta_2 + \frac{\text{Cov}(X, u)}{\text{Var}(X)}$$

$$\textbf{plim}\,\frac{A}{B} = \frac{\textbf{plim}\,A}{\textbf{plim}\,B}$$ **provided both** $\textbf{plim}\,A$ **and** $\textbf{plim}\,B$ **exist**

MODEL B. PROPERTIES OF THE REGRESSION COEFFICIENTS: CONSISTENCY

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Demonstrate that $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$

is a consistent estimator of β_1 in the simple regression model. (is a consistent estimator of β_2.)

$$\operatorname{plim} \hat{\beta}_1 = \operatorname{plim} (\beta_1 + \beta_2 \bar{X} + \bar{u}) - \operatorname{plim} \hat{\beta}_2 \operatorname{plim} \bar{X} = \beta_1$$

since $\operatorname{plim} \bar{u} = 0.$

Alternatively, we have seen that $\hat{\beta}_1$ is unbiased and that its variance is

$$\sigma^2_{\hat{\beta}_1} = \sigma^2_u \left(\frac{1}{n} + \frac{\bar{X}^2}{n \operatorname{MSD}(X)} \right)$$

Since the variance tends to zero as *n* tends to infinity, the estimator is consistent.

MEASUREMENT ERRORS. MEASUREMENT ERROR IN EXPLANATORY VARIABLE.

$$Y = \beta_1 + \beta_2 Z + v \quad - \ actual\ relationship$$

$$X = Z + w \quad - \ measured\ \ with\ \ random\ \ error\ \ w$$

$$Y = \beta_1 + \beta_2 (X - w) + v = \beta_1 + \beta_2 X + v - \beta_2 w =$$
$$= \beta_1 + \beta_2 X + u$$

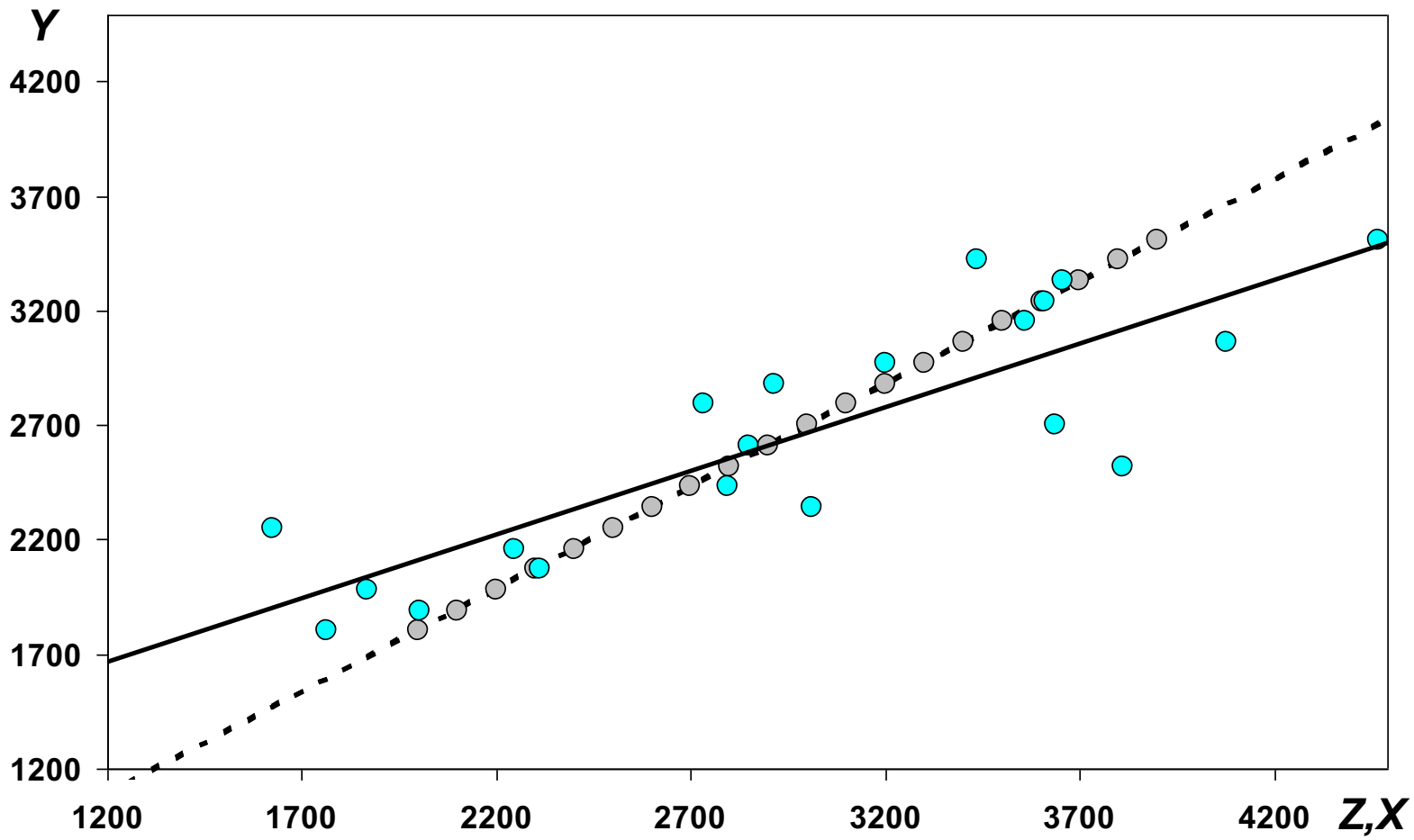
w ***w***

X **has a random component, the measurement error *w*. Suppose E(w)=0.**

w **is also one of the components of the compound disturbance term.**

Hence *u* is not distributed independently of *X*. Assumption B.7 is violated.

MEASUREMENT ERROR IN EXPLANATORY VARIABLE: GRAPHICAL ILLUSTRATION.



The regression line underestimates the true positive slope,

or overestimate the true negative one.

MEASUREMENT ERROR IN EXPLANATORY VARIABLE.

$$Y = \beta_1 + \beta_2 Z + v \qquad Y = \beta_1 + \beta_2 X + u$$

$$X = Z + w \qquad u = v - \beta_2 w$$

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})(\beta_2[X_i - \bar{X}] + u_i - \bar{u})}{\sum(X_i - \bar{X})^2}$$

$$= \beta_2 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2} = \beta_2 + \frac{\frac{1}{n}\sum(X_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum(X_i - \bar{X})^2}$$

$$= \beta_2 + \frac{\widehat{\mathrm{Cov}}\,(X, u)}{\widehat{\mathrm{Var}}\,(X)}$$

We decompose the slope coefficient into the true value and an error term as usual.

MEASUREMENT ERROR IN EXPLANATORY VARIABLE.

$$Y = \beta_1 + \beta_2 Z + v \qquad Y = \beta_1 + \beta_2 X + u$$

$$X = Z + w \qquad u = v - \beta_2 w$$

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \beta_2 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2}$$

$$\text{plim}\,\hat{\beta}_2 = \beta_2 + \frac{\text{plim}\left(\frac{1}{n}\sum(X_i - \bar{X})(u_i - \bar{u})\right)}{\text{plim}\left(\frac{1}{n}\sum(X_i - \bar{X})^2\right)} = \beta_2 + \frac{\text{Cov}(X, u)}{\text{Var}(X)}$$

Having divided by *n*, it can be shown that the plim of the numerator is Cov(*X*, *u*) and the plim of the denominator is Var(*X*).

MEASUREMENT ERROR IN EXPLANATORY VARIABLE.

$$Y = \beta_1 + \beta_2 Z + v \qquad Y = \beta_1 + \beta_2 X + u$$

$$X = Z + w \qquad u = v - \beta_2 w$$

$$\text{plim}\,\hat{\beta}_2 = \beta_2 + \frac{\text{Cov}(X,u)}{\text{Var}(X)} = \beta_2 - \beta_2 \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$$

$$\begin{aligned}\text{Cov}(X,u) &= \text{Cov}\big((Z+w),(v-\beta_2 w)\big) = \\ &= \text{Cov}(Z,v) + \text{Cov}(w,v) + \text{Cov}(Z,-\beta_2 w) + \text{Cov}(w,-\beta_2 w) = \\ &= 0 + 0 + 0 - \beta_2 \sigma_w^2\end{aligned}$$

$$\begin{aligned}&\text{Var}(X) = \text{Var}(Z+w) = \text{Var}(Z) + \text{Var}(w) + 2\,\text{Cov}(Z,w) \\ &= \sigma_Z^2 + \sigma_w^2 + 0\end{aligned}$$

Thus in large samples, $\hat{\beta}_2$ is biased towards 0 and the size of the bias depends on the relative sizes of the variances of *w* and *Z*. It is also biased for small samples. Due to violation of Assumption B.7, the standard errors, *t* tests and *F* test are invalid.

MEASUREMENT ERROR IN EXPLANATORY VARIABLE.

$$Y = \beta_1 + \beta_2 Z + v \qquad Y = \beta_1 + \beta_2 X + u$$

$$X = Z + w \qquad u = v - \beta_2 w$$

$$\mathrm{E}(w) = 0 \Rightarrow \mathrm{E}(X) = \mathrm{E}(Z). \qquad \mathrm{E}(u) = \mathrm{E}(v) - \beta_2 \mathrm{E}(w) = 0 - 0 = 0.$$

$$\operatorname{plim} \hat{\beta}_2 = \beta_2 - \beta_2 \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2} = \beta_2 \frac{\sigma_Z^2}{\sigma_Z^2 + \sigma_w^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = \beta_1 + \beta_2 \bar{X} + \bar{u} - \hat{\beta}_2 \bar{X} = \beta_1 + \beta_2 \bar{X} + \bar{v} - \beta_2 \bar{w} - \hat{\beta}_2 \bar{X}$$

$$\operatorname{plim} \hat{\beta}_1 = \beta_1 + (\beta_2 - \operatorname{plim} \hat{\beta}_2) \operatorname{plim} \bar{X} + \operatorname{plim} \bar{v} - \beta_2 \operatorname{plim} \bar{w}$$

$$\operatorname{plim} \hat{\beta}_1 = \beta_1 + (\beta_2 - \operatorname{plim} \hat{\beta}_2) \operatorname{plim} \bar{X} =$$

$$= \beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2} \mathrm{E}(X) = \beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2} E(Z)$$

Measurement error in explanatory variable: Multiple Regression

$$x_1 = x_1^* + e_1$$ ← Mis-measured value = True value + Measurement error

$$y = \beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k + u$$ ← Population regression

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (u - \beta_1 e_1)$$ ← Estimated regression

Classical errors-in-variables assumption: $Cov(x_1^*, e_1) = 0$ ← Error uncorrelated to true value

$$\Rightarrow \quad Cov(x_1, u - \beta_1 e_1) = -\beta_1 Cov(x_1, e_1) = -\beta_1 \sigma_{e_1}^2$$ ← The mismeasured variable x_1 is correlated with the error term

$$\text{plim}\, \hat{\beta}_1 = \beta_1 \frac{\sigma_{r_1}^2}{\sigma_{r_1}^2 + \sigma_{e_1}^2}$$ ←

This factor (which involves the error variance of a regression of the true value of x_1 on the other explanatory variables) will always be between zero and one

$x_1^* = \alpha_0 + \alpha_1 x_2 + \ldots + \alpha_k x_k + r_1^*$

Even if x_1 is the only mismeasured variables, all OLS estimators are biased and inconsistent. The directions and sizes of the bias of all coefficients other than $\hat{\beta}_1$ may differ.

MEASUREMENT ERROR IN DEPENDENT VARIABLE

$$Q = \beta_1 + \beta_2 X + v \quad - \text{actual dependent variable}$$

$$Y = Q + r \quad - \quad \text{measured variable}$$

$r -$ mearurement error. Suppose $E(r) = 0$

$$Y - r = \beta_1 + \beta_2 X + v$$

$$Y = \beta_1 + \beta_2 X + v + r = \beta_1 + \beta_2 X + u \qquad u = v + r$$

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma_u^2}{n\sigma_X^2} = \frac{\sigma_v^2 + \sigma_r^2}{n\sigma_X^2}$$

The standard errors and tests remain valid. But the standard errors tend to be larger than with no measurement error.

FRIEDMAN'S CRITIQUE OF OLS ESTIMATION OF THE CONSUMPTION FUNCTION

True model	$Q = \beta_1 + \beta_2 Z + v$	$C^P = \beta_1 + \beta_2 Y^P$
Measurement errors	$X = Z + w$ $Y = Q + r$	$Y = Y^P + Y^T$ $C = C^P + C^T$
True model, in measured variables	$Y = \beta_1 + \beta_2 X + v + r - \beta_2 w = \beta_1 + \beta_2 X + u$	$C = \beta_1 + \beta_2 Y + C^T - \beta_2 Y^T = \beta_1 + \beta_2 Y + u$
$\text{plim}\ \hat{\beta}_2$	$\beta_2 - \beta_2 \dfrac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$	$\beta_2 - \beta_2 \dfrac{\sigma_{Y^T}^2}{\sigma_{Y^P}^2 + \sigma_{Y^T}^2}$

To simplify the analysis, we will assume that the transitory components of consumption and income are independent of their permanent components and of each other. MPC tends to be underestimated.

FRIEDMAN'S CRITIQUE: MONTE CARLO SIMULATION

True model	$C^P = \beta_1 + \beta_2 Y^P$	$C^P = 0 + 0.9 Y^P$
Data for the regressor(s)		$Y^P = 2000, 2100, \ldots, 3900$
Measurement errors	$Y = Y^P + Y^T$ $C = C^P + C^T$	$Y^T = 400N(0,1)$ $C^T = 0$
$\text{plim } \hat{\beta}_2$	$\beta_2 - \beta_2 \dfrac{\sigma^2_{Y^T}}{\sigma^2_{Y^P} + \sigma^2_{Y^T}}$	$0.9 - 0.9 \dfrac{160000}{325000 + 160000}$ $= 0.9 - 0.29 = \mathbf{0.61}$

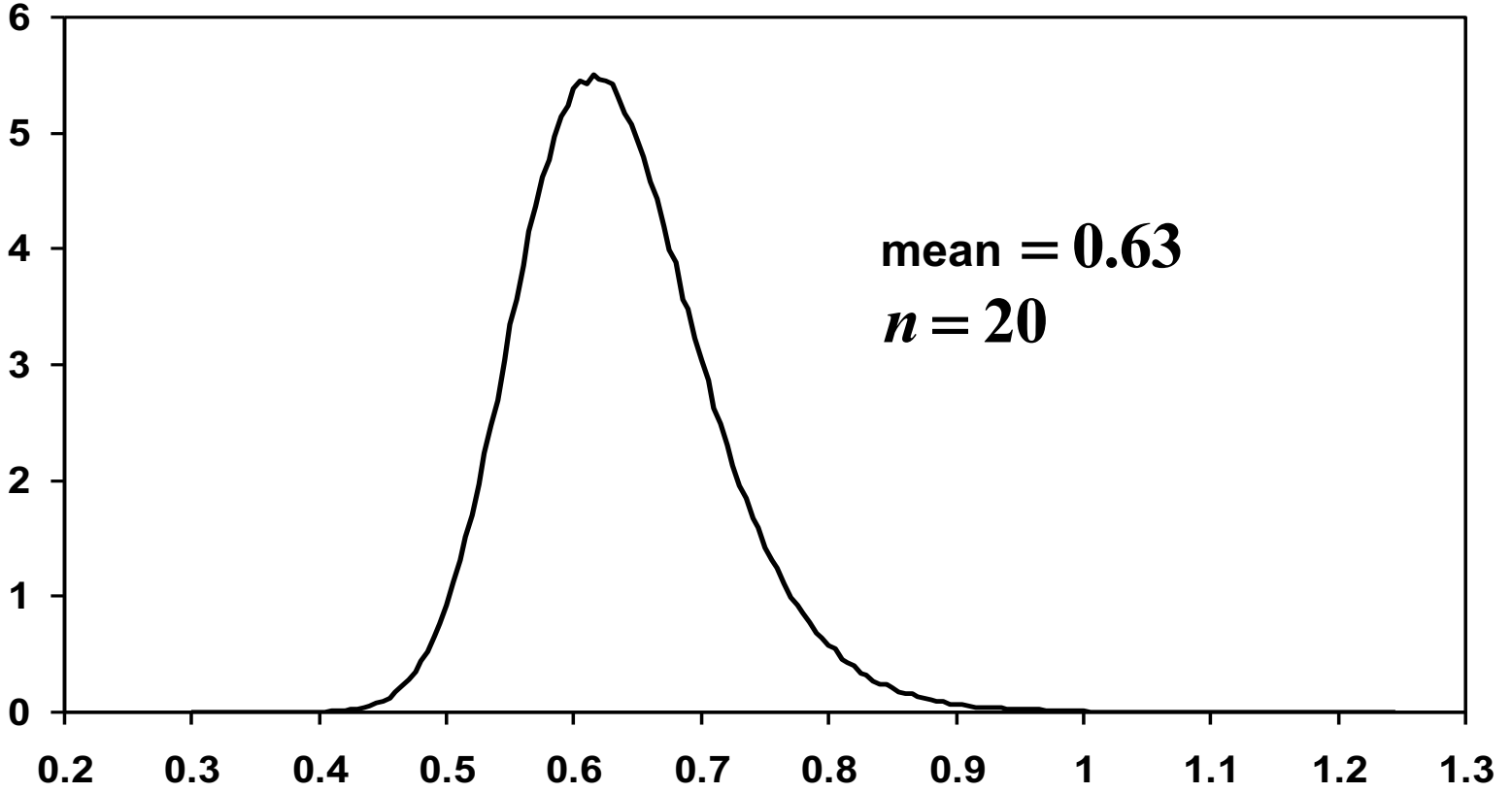
The scaling factor of 400 was chosen, after a bit of trial and error, because it produced a visible bias. Given our choice of parameters and data, $\hat{\beta}_2$ will tend to 0.61 in large samples.

FRIEDMAN'S CRITIQUE OF OLS ESTIMATION OF THE CONSUMPTION FUNCTION

Sample	$\hat{\beta}_1$	$s.e.(\hat{\beta}_1)$	$\hat{\beta}_2$	$s.e.(\hat{\beta}_2)$
1	1,001	251	0.56	0.08
2	755	357	0.62	0.11
3	756	376	0.68	0.13
4	668	290	0.66	0.09
5	675	179	0.64	0.06
6	982	289	0.57	0.10
7	918	229	0.56	0.07
8	625	504	0.66	0.16
9	918	181	0.58	0.06
10	679	243	0.65	0.08

The table shows 10 results of the experiment. The true values of β_1 and β_2 are 0 and 0.9. There is a strong downward bias in $\widehat{\beta}_2$ and its values do seem to be distributed around the limiting value of 0.61. $\widehat{\beta}_1$ is upwards biased. The standard errors are invalidated by the measurement error bias and we should not perform tests.

FRIEDMAN'S CRITIQUE OF OLS ESTIMATION OF THE CONSUMPTION FUNCTION



$$\text{plim}(\hat{\beta}_2) = 0.61$$

This chart shows the distribution of the estimates of the slope coefficient when the Monte Carlo simulation was repeated 1,000,000 times. The probability limit for the slope coefficient, 0.61, is close though the sample size, 20, is quite small.

ASSUMPTION B7 VIOLATION:

1. Reasons: Measurement Errors (special case). More general – Endogeneity.

2. Consequences: inconsistent OLS estimators, standard statistics wrongly calculated, tests invalid.

3. Detection: Durbin-Wu-Hausman test

4. Remedial measures: Instrumental Variables