Ch. 7 Violations of the Ideal Conditions

(December 3, 2020)

The previous chapter dealt with the estimation and hypothesis procedures typically used in a regression model satisfying the full ideal conditions. This chapter will consider violations of the above conditions **one at a time**, in the sense that it will consider a violation of one of these conditions while assuming that all the others still hold.

Consideration of nonspherical disturbance, however, will be put off until Chapter 8-10.

1 Specification

1.1 Selection of Variables

Consider an initial model, which we assume that

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}, \tag{7-1}$$

It is not unusual to begin with some formulation and then contemplate adding more variable (regressors) to the model:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{7-2}$$

Let R_1^2 be the R-square of the model with fewer regressor as in (7-1), and R_{12}^2 be the R-square of the model with more regressors as in (7-2). It is apparent as we have

shown earlier that $R_{12}^2 > R_1^2$. Clearly, it would be possible to push R^2 as high as desired by adding regressors. This problem motivates the use of the adjusted R-square,

$$\bar{R}^2 = 1 - \frac{T-1}{T-k}(1-R^2).$$

as a method to selection of regressors.

It has been suggested that the adjusted R-square does not penalize the loss of degree of freedom heavily, three alternatives have frequently been proposed for comparing models are

(a).

$$\tilde{R}_j^2 = \frac{T+k_j}{T-k_j}(1-R_j^2),$$

(b). Akaike's information criterion (min AIC)):

$$AIC_j = \ln\left(\frac{\mathbf{e}_j'\mathbf{e}_j}{T}\right) + \frac{2k_j}{T} = \ln\hat{\sigma}_j^2 + \frac{2k_j}{T},$$

(c). Scharz's Bayesian information criterion (min BIC):

$$BIC_j = \ln\left(\frac{\mathbf{e}_j'\mathbf{e}_j}{T}\right) + k_j \cdot \frac{\ln T}{T} = \ln\hat{\sigma}_j^2 + \frac{k_j \ln T}{T},$$

where k_j is number of regressors.

(d). Hansen (2007)'s Model Averaging. □

Although intuitively appealing, these measures are a bit unorthodox in that they have no firm basis in theory (unless that are used in time series analysis model). Perhaps a somewhat more palatable alternative is the method of *stepwise regression*. This procedure evaluates each variable in turn on the basis of its significance level and accumulates the model by adding or deleting variables sequentially. At the second step, to say nothing for the latter steps of this procedure, the classical inference procedures break down. Suppose, for example (as it is common), that the method is used to ensure that all the variables ultimately included have F statistics (the square of the conventional t ratio) larger than four. It is hardly appropriate to call theses F statistics and base inference on them as if they are drawn from an F distribution; it is known a

priori that they all will be larger than four. For this reason, economists have tended to avoid stepwise regression method for the break down of inference procedures. There are good reasons to be skeptical of a "model" that is constructed entirely by mechanical means.

A model constructed by mechanical methods is likely to suffer *data snooping bias*. The data snooping bias is a statistical bias that appears when exhaustively searching for combinations of variables, the probability that a result arose by chance grow with the number of combinations tested. For example in the selection of variable to be included in the regressor, data snooping may arise because, when many models are evaluated individually, some are bound to be "statistically" superior by chance alone even though they are not better than a benchmark. Suppose there are 100 false models that are mutually independent, and we apply a t-test to each model with the significance level 5%. The probability of falsely rejecting at least one correct null hypothesis[1] is $1-(0.95)^{100} \approx 0.994$. It is thus highly likely that an individual test may incorrectly suggest a false (or an inferior) model to be a significant one.

1.2 Omission of Relevant variables

Suppose that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where the two parts of $\mathbf{X}$ have k_1 and k_2 columns, respectively. If we regress $\mathbf{y}$ on $\mathbf{X}_1$ without including $\mathbf{X}_2$, that is you have estimated the model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$

and obtain the estimator as (you do not have $\hat{\boldsymbol{\beta}}_2$)

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\underbrace{(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon})}_{\text{true } \mathbf{y}} \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}. \qquad (7\text{-}3)\end{aligned}$$

[1] The null hypothesis is $\beta = 0$ in this model which is correct because this regressor should not be included.

1.2.1 Bias of the OLS estimator

Taking the expectation of (7-3), we see that unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$, the OLS $\hat{\boldsymbol{\beta}}_1$ is biased:

$$E(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2.$$

For statistical inference, it would be necessary to estimate σ^2. Proceeding as usual, we would use

$$s^2 = \frac{\mathbf{e}_1'\mathbf{e}_1}{T - k_1}.$$

But now

$$\mathbf{e}_1 = \mathbf{M}_1\mathbf{y} = \mathbf{M}_1(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) = \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\boldsymbol{\varepsilon}.$$

Thus,

$$E[\mathbf{e}_1'\mathbf{e}_1] = \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \sigma^2 tr(\mathbf{M}_1) = \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \sigma^2(T - k_1).$$

It is simple to see that $\boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ is positive (how ?) so s^2 is biased upward.

1.2.2 Variance of the Coefficients Estimators

The variance of $\hat{\boldsymbol{\beta}}_1$ can be easily computed as

$$Var(\hat{\boldsymbol{\beta}}_1) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

If we had computed the correct regression, including $\mathbf{X}_2$, then the slope estimator on $\mathbf{X}_1$, denoted by $\hat{\boldsymbol{\beta}}_{1,\mathbf{X}}$ would have a covariance matrix equal to the upper left block of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, i.e.

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} Var(\hat{\boldsymbol{\beta}}_{1,\mathbf{X}}) \\ Var(\hat{\boldsymbol{\beta}}_{2,\mathbf{X}}) \end{bmatrix} &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \\ &= \sigma^2 \begin{bmatrix} [\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1} & * \\ * & * \end{bmatrix}, \end{aligned}$$

or

$$Var(\hat{\boldsymbol{\beta}}_{1,\mathbf{X}}) = \sigma^2[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1}.$$

We can compare the covariance matrix of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_{1,\mathbf{X}}$ more easily by comparing their inverses:

$$Var(\hat{\boldsymbol{\beta}}_1)^{-1} - Var(\hat{\boldsymbol{\beta}}_{1,\mathbf{X}})^{-1} = (1/\sigma^2)\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1,$$

which is nonnegative definite (see the following lemma below). We conclude that although $\hat{\boldsymbol{\beta}}_1$ is biased, it has a smaller variance than $\hat{\boldsymbol{\beta}}_{1,\mathbf{X}}$ (since the inverse of its variance is larger).

Lemma.
Let $\mathbf{A}$ be a positive definite $(n \times n)$ matrix and let $\mathbf{B}$ denote any nonzero $(n \times m)$ matrix. Then $\mathbf{B}'\mathbf{A}\mathbf{B}$ is nonnegative definite.

Proof.
Let $\mathbf{x}$ be any nonzero vector. Define

$$\tilde{\mathbf{x}} \equiv \mathbf{B}\mathbf{x}.$$

Then $\tilde{\mathbf{x}}$ can be any vector including the zero vector. Then

$$\mathbf{x}'\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{x} = \tilde{\mathbf{x}}'\mathbf{A}\tilde{\mathbf{x}} \geq 0$$

from the positive definiteness of matrix $\mathbf{A}$. ■

The conclusion of this subsection is that if we *omit* relevant variables from the regression, then our estimate of both $\boldsymbol{\beta}_1$ and σ^2 are biased although it is possible that $\hat{\boldsymbol{\beta}}_1$ is more precise than $\hat{\boldsymbol{\beta}}_{1,\mathbf{X}}$.

1.3 Inclusion of Irrelevant Variables

If the regression model is correct given by

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}, \tag{7-4}$$

but we estimate it by

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{7-5}$$

That is, we include the irrelevant regressors $\mathbf{X}_2$. In fact $\boldsymbol{\beta}_2 = \mathbf{0}$. From partitioned regression estimator, we obtain that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y} = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\underbrace{(\mathbf{X}_1\boldsymbol{\beta}_1+\boldsymbol{\varepsilon})}_{\text{true } \mathbf{y}} \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\boldsymbol{\varepsilon},\end{aligned}$$

and (we have an estimation of $\boldsymbol{\beta}_2$)

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1(\mathbf{X}_1\boldsymbol{\beta}_1+\boldsymbol{\varepsilon}) \\ &= \mathbf{0} + (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\boldsymbol{\varepsilon}.\end{aligned}$$

Therefore,

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}_1) &= \boldsymbol{\beta}_1, \text{ and} \\ E(\hat{\boldsymbol{\beta}}_2) &= \mathbf{0}.\end{aligned}$$

Exercise 1.
Show that s^2 is unbiased:

$$E\left(\frac{\mathbf{e}'\mathbf{e}}{T-k_1-k_2}\right) = \sigma^2,$$

where $\mathbf{e} = \mathbf{y} - \mathbf{X_1}\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2$ is the residual from (7-5) when the true model is from (7-4). ■

Then what's the problem ? It would seem that one would generally want to "overfit" the model. However the cost is the reduction of the precision of the estimation. As we have seen that the covariance matrix of the shorter (correct) regressors is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.

2 Functional Form

2.1 Sample Splitting (Shifts of Function)

2.1.1 Dummy Variables

One of the most useful devices in regression analysis is the binary, or *dummy* variables, which takes value of only 0 and 1. It is a convenient means of building discrete shifts of the function into a regression model.

Consider the following illustration example of dummy variables application in comparing two means. If a model describes the salary function by

$$Y = \mu + \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \tag{7-6}$$

where μ can be regard as "starting salary" to anyone (even) with different academic degree. This model can more realistic if dividing the sample in "starting salary" (μ) into two categories: individuals attending college and not attending college. Formally (7-6) can be rewritten as

$$Y_i = \mu + \delta d_i + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, ..., N,$$

$$\text{where} \begin{cases} d_i = 1, & if\ attending\ college, \\ d_i = 0, & if\ not\ attending\ college. \end{cases}$$

Logically, $\delta > 0$ and d_i is the *dummy variable*.

The above model can also be written equivalently as

$$Y_i = \delta d_{1i} + \eta d_{2i} + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, ..., N,$$

$$\text{where} \begin{cases} d_{1i} = 1, & if\ attending\ college \\ d_{1i} = 0, & if\ not\ attending\ college \end{cases}$$

$$\text{and} \begin{cases} d_{2i} = 0, & if\ attending\ college \\ d_{2i} = 1, & if\ not\ attending\ college, \end{cases}$$

but not

$$Y_i = \mu + \delta d_{1i} + \eta d_{2i} + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{7-7}$$

to avoid **dummy variable trap**.[2]

For this reason, if one wants to remove the seasonal effect,[3] we need 4 dummies without a common mean or use 3 dummies with a common mean in a linear regression model.

Example.
See eq. (6.1) on p.108 of Greene 6th Edition. ■

2.1.2 Splitting by Other Variables, Hansen (2000)'s Threshold Regression

Sometimes the subsamples are selected on categorical variables, such as gender, but in other cases the subsamples are selected based on continuous variables, such as firm size. In the latter case, some decision must be made concerning what is the appropriate threshold (i.e., how big must a firm be to be categorized as "large") at which to split the sample. When this value is unknown, some method must be employed in its selection.

The threshold regression model of Hansen (2000) takes the form

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_i, \quad q_i \leq \gamma, \tag{7-8}$$

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta}_2 + \varepsilon_i, \quad q_i > \gamma, \tag{7-9}$$

where q_i may be called the threshold variable, and is used to split the sample into two groups, which we may call "classes", or "regimes" depending on the context. This model allows the regression parameters to differ depending on the value of q_i. In the threshold regression model, we observe $\{Y_i, \mathbf{x}_i', q_i\}_{i=1}^N$. where Y_i and q_i are real-valued and $\mathbf{x}_i$ is an k-vector.

[2] In this case, the sample matrix of the regressor would be

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & \mathbf{x}_1' \\ 1 & 0 & 1 & \mathbf{x}_2' \\ 1 & 1 & 0 & \mathbf{x}_3' \\ . & . & . & . \\ . & . & . & . \\ 1 & 1 & 0 & \mathbf{x}_{N-1}' \\ 1 & 0 & 1 & \mathbf{x}_N' \end{bmatrix}.$$

The rank of $\mathbf{X}$ is not full.

[3] Assume that the seasonal effects can be removed from the constant.

To write the model in a single equation, define the dummy variable

$$d_i(\gamma) = 1, \text{when } q_i \leq \gamma,$$

and set

$$\mathbf{x}_i(\gamma) = \mathbf{x}_i d_i(\gamma)$$

so that (7-8)-(7-9) equals

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_i'(\gamma)\boldsymbol{\delta} + \varepsilon_i, \tag{7-10}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_2$ and $\boldsymbol{\delta} = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$.[4]

To express the model in matrix notation, define the $N \times 1$ vectors $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ by stacking the variables Y_i and ε_i , and the $N \times k$ matrices $\mathbf{X}$ and $\mathbf{X}_\gamma$ by stacking the vectors $\mathbf{x}'$ and $\mathbf{x}_i'(\gamma)$. Then (7-10) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_\gamma\boldsymbol{\delta} + \boldsymbol{\varepsilon}. \tag{7-11}$$

The regression parameters are $(\boldsymbol{\beta}, \boldsymbol{\delta}, \gamma)$, and the natural estimator is least squares (LS). Let

$$S(\boldsymbol{\beta}, \boldsymbol{\delta}, \gamma) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_\gamma\boldsymbol{\delta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_\gamma\boldsymbol{\delta}) \tag{7-12}$$

be the sum of squared errors function. Then by definition the LS estimators $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \hat{\gamma}$ jointly minimize (7-12). For this minimization, γ is assumed to be restricted to a bounded set $[\underline{\gamma}, \overline{\gamma}] = \Gamma$.

The computationally easiest method to obtain the LS estimates is through concentration. Conditional on γ, is linear in $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, yielding the conditional OLS estimators $\hat{\boldsymbol{\beta}}(\gamma), \hat{\boldsymbol{\delta}}(\gamma)$ by regressing $\mathbf{y}$ on $\mathbf{X}_\gamma^* = [\mathbf{X} \;\; \mathbf{X}_\gamma]$. The concentrated sum of squared errors function is

$$S(\gamma) = S(\hat{\boldsymbol{\beta}}(\gamma), \hat{\boldsymbol{\delta}}(\gamma), \gamma) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}_\gamma^*(\mathbf{X}_\gamma'^*\mathbf{X}_\gamma^*)^{-1}\mathbf{X}_\gamma'^*\mathbf{y},$$

and $\hat{\gamma}$ is the value that minimizes $S(\gamma)$. Since $S(\gamma)$ takes on less than N distinct values, $\hat{\gamma}$ can be defined uniquely as

$$\hat{\gamma} = \text{argmin}_{\gamma \in \Gamma}\, S(\gamma),$$

which requires less than N function evaluations. The slope estimates can be computed via $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\gamma})$ and $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(\hat{\gamma})$.

[4] In the original Hensen (2000) paper, Hensen denote $\boldsymbol{\delta} = (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)$. See p. 576.

Hansen (2000) develops a statistical theory for threshold estimation in the regression context. An asymptotic distribution theory for the regression estimates (the threshold and the regression slopes) is developed. It is found that the distribution of the threshold estimate is nonstandard. A method to construct asymptotic confidence intervals is developed by inverting the likelihood ratio statistic. It is shown that this yields asymptotically conservative confidence regions.

Example. (Hossfeld and MacDonald (2014) "Carry funding and safe haven currencies: a threshold regression approach", eq(2).)

$$\begin{aligned}\Delta e_t &= \beta_{s,1}(i^*_{t-1} - i_{t-1}) + \beta_{s,2}(i^*_{t-1} - i_{t-1}) \cdot VXO_t \\ &\quad +\beta_{s,3}(\pi^*_{t-1} - \pi_{t-1}) + \beta_{s,4}u_{t-1} + \beta_{s,5}\Delta msciw_t + \varepsilon_t\end{aligned}$$

where

$$s = \begin{cases} l, & if \quad VOX_t \leq \gamma \\ h, & if \quad VOX_t > \gamma, \end{cases}$$

where Δe_t is the (effective) exchange rate return in period t (in terms of foreign currency units per unit of domestic currency), $(i^*_{t-1} - i_{t-1})$ the lagged (effective) interest differential, VXO_t the CBOE $S\&P$ 100 Volatility Index in period t, the $(\pi^*_{t-1} - \pi_{t-1})$ lagged (effective) inflation differential, and $\Delta msciw_t$ the return on the MSCI world index in local currencies in period t; and $\beta_{s,i}$ denote the regime-specific regression parameters, γ the threshold value (in our case, the value of the VXO), which splits the sample into two regimes, l and h are acronyms (l=low, h=high) used to denote the specific stress-regime, s. ■

2.2 Nonlinearity in the Variables

The linear model we proposed is not as "limited" as the first glance. By using logarithms, exponential, reciprocal, transcendental functions and polynomials, and so on, this "linear" model is also useful in the general form:

$$\begin{aligned}g(Y) &= \beta_1 f_1(Z_1) + \beta_2 f_2(Z_2) + ... + \beta_k f_k(Z_k) + \varepsilon \\ &= \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon.\end{aligned}$$

which can be tailored to any number of situations.

2.2.1 Log-Linear Model

A commonly used form of regression model is the log-linear model:

$$W = \alpha \prod_k Z_k^{\beta_k} \exp(\varepsilon)$$

or

$$\begin{aligned} \ln W = Y &= \ln \alpha + \sum_k \beta_k \ln Z_k + \varepsilon \\ &= \beta_1 + \sum_k \beta_k X_k + \varepsilon. \end{aligned}$$

All you have to do is take natural logarithms of the data before the regression.

In this model, the coefficients are elasticities:

$$\left(\frac{\partial W}{\partial Z_k}\right)\left(\frac{Z_k}{W}\right) = \frac{\partial \ln W}{\partial \ln Z_k} = \frac{\partial Y}{\partial X_k} = \beta_k.$$

2.2.2 The Semi-Log Model

A hybrid of the linear and log-linear models is the semilog equation

$$Y = e^{\beta_1 + \beta_2 X + \varepsilon}$$

or

$$\ln Y = \beta_1 + \beta_2 X + \varepsilon.$$

A common use of semilog formulation is in exponential growth curves. If X is "time" t, then

$$\begin{aligned} \frac{d \ln Y}{dt} &= \beta_2 \\ &= \text{average rate of growth of Y.} \end{aligned}$$

Example.

Macroeconomic models are often formulated with autonomous time trends. For example, aggregate models of productivity will usually include a time trend variables, as in

$$\ln\left(\frac{Q}{L}\right)_t = \beta_1 + \ln\left(\frac{K}{L}\right)_t + \delta t + \varepsilon_t.$$

In this equation, δ is the rate of growth of average product not attributed to increase in the use of capital. The autonomous growth in productivity usually attributed to "technical change". ■

2.2.3 Regression with Interaction Terms

Another useful formulation of the regression model is one with *interaction terms*. For example, a model relating breaking distance D to speed S and road wetness W might be

$$D = \beta_1 + \beta_2 S + \beta_3 W + \beta_4 (S \cdot W) + \varepsilon.$$

In this model,

$$\frac{\partial D}{\partial S} = \beta_2 + \beta_4 W,$$

which implies that the marginal effect of higher speed on breaking distance will increase when the road is wetter (assuming that β_4 is positive). Likewise for $\partial D/\partial W$.

3 Stochastic Regressors

To this point we have considered the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where the regressors $\mathbf{X}$ are treated as fixed in repeated samples. In many cases, however, this assumption is untenable. The explanatory variables economists and other social scientists use are often generated by stochastic processes beyond their control. This section we will consider the consequences of relaxing the fixed $\mathbf{X}$ assumption. It will be assumed that the full ideal conditions hold except that the regressor matrix $\mathbf{X}$ is random.

3.1 The Exogenous Regressors Model

A weaker assumption than $\mathbf{X}$ being nonstochastic is that the explanatory variables form the columns of $\mathbf{X}$ is *exogenous*. It implies that any randomness in the DGP that generated $\mathbf{X}$ is independent of the disturbances $\boldsymbol{\varepsilon}$ in the DGP for $\mathbf{y}$. This independence in turn implies that[5]

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}. \tag{7-13}$$

With this assumption, the classical ideal conditions can be rewritten as follows.

Assumption 2. (Stochastic but Exogenous Regressor)

(a). The model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is correct; (no model misspecification)

(b). Rank E($\mathbf{X}$) = k; (for model identification)

(c). $\mathbf{X}$ is stochastic and $plim_{T\to\infty}(\mathbf{X}'\mathbf{X}/T) = \underline{\mathbf{Q}}$, where $\underline{\mathbf{Q}}$ is a finite and nonsingular matrix.

[5]A substantially weaker than assumption (7-13) is $E(\varepsilon_t|\mathbf{x}_t) = 0$. While (7-13) rules out the possibility that ε_t may dependent on the values of the regressors for *any* observations, this conditions merely rules out the possibility that it may depend on their values for the current observations. We refer this condition as *predeterminedness* condition.

(d).

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \begin{bmatrix} E(\varepsilon_1|\mathbf{X}) \\ E(\varepsilon_2|\mathbf{X}) \\ \vdots \\ E(\varepsilon_T|\mathbf{X}) \end{bmatrix} = \mathbf{0}.$$

The zero conditional mean implies that the unconditional mean is also zero:

$$E(\varepsilon_i) = E_{\mathbf{X}}[E(\varepsilon_i|\mathbf{X})] = E_{\mathbf{X}}[0] = 0.$$

(e). $Var(\boldsymbol{\varepsilon}|\mathbf{X}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2 \cdot \mathbf{I}$; (disturbances have same variance and are not autocorrelated)

(f). $\boldsymbol{\varepsilon}|\mathbf{X}$ is normal distributed. □

We now investigate the statistical properties of OLS estimator under these assumptions.

3.1.1 Unbiasedness ?

The least squares of slope estimators are still unbiased in every sample. To show this, write

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \end{aligned}$$

Using law of iterated expectation, the expected value of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) = E_{\mathbf{X}}[E(\hat{\boldsymbol{\beta}}|\mathbf{X})] &= E_{\mathbf{X}}\{E[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}]\} \\ &= E_{\mathbf{X}}[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}|\mathbf{X})] \\ &= E_{\mathbf{X}}(\boldsymbol{\beta}) = \boldsymbol{\beta}. \end{aligned}$$

The OLS estimator of the disturbance variance $s^2 = \frac{\mathbf{e}'\mathbf{e}}{T-k}$ remains unbiased since

$$E(\mathbf{e}'\mathbf{e}) = E_{\mathbf{X}}[E(\boldsymbol{\varepsilon}'\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}|\mathbf{X})] = E_{\mathbf{X}}(\sigma^2(T-k)) = \sigma^2(T-k),$$

therefore $E(s^2) = \sigma^2$.

3.1.2 Efficiency ?

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is slightly different from previous model, however.

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] \\ &= E_{\mathbf{X}}\{E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}]\} \\ &= E_{\mathbf{X}}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\} \\ &= E_{\mathbf{X}}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\} \\ &= \sigma^2 E_{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 E(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

provided, of course, that $\sigma^2 E(\mathbf{X}'\mathbf{X})^{-1}$ exists. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is σ^2 times the expected value of $(\mathbf{X}'\mathbf{X})^{-1}$ since $(\mathbf{X}'\mathbf{X})^{-1}$ takes different values with new random samples.

The Gauss-Markov theorem can be established logically from the results of the preceding paragraph. We have showed that

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) - Var(\tilde{\boldsymbol{\beta}}|\mathbf{X})$$

is nonpositive for any $\hat{\boldsymbol{\beta}} \neq \tilde{\boldsymbol{\beta}}$ and for the specific $\mathbf{X}$ in our sample. But if this inequality holds for every particular $\mathbf{X}$, then it must hold for

$$Var(\hat{\boldsymbol{\beta}}) = E_{\mathbf{X}}[Var(\hat{\boldsymbol{\beta}}|\mathbf{X})].$$

That is, if it holds for every particular $\mathbf{X}$, then it must hold over the average value of $\mathbf{X}$.

Theorem. (Gauss-Markov Theorem with Stochastic Regressors)
In the classical linear regression model, the least squares estimator $\hat{\boldsymbol{\beta}}$ is the minimum variance linear unbiased estimator of $\boldsymbol{\beta}$ whether $\mathbf{X}$ is stochastic or non-stochastic. ■

3.1.3 Consistency ?

From notation we know that

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_T' \end{bmatrix},$$

then

$$\frac{1}{T}\mathbf{X}'\mathbf{X} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'.$$

If we assume that

$$plim\frac{1}{T}\mathbf{X}'\mathbf{X} = plim\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' = \underline{\mathbf{Q}}$$

is finite and nonsingular, then by law of large number, $\underline{\mathbf{Q}} = \frac{1}{T}\sum_{t=1}^{T} E(\mathbf{x}_t\mathbf{x}_t')$,[6] that is, the second moments of the regressors is finite.[7]

The exogenous regressors assumption $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ implies that $plim(\mathbf{X}'\boldsymbol{\varepsilon}/T) = \mathbf{0}$. This is because

$$plim\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T}\right) \xrightarrow{a.s.} \frac{1}{T}E(\mathbf{X}'\boldsymbol{\varepsilon}) = \mathbf{0}, \tag{7-14}$$

where we have used the fact that

$$E(\mathbf{X}'\boldsymbol{\varepsilon}) = E_{\mathbf{X}}[E(\mathbf{X}'\boldsymbol{\varepsilon})|\mathbf{X}] = E_{\mathbf{X}}[\mathbf{X}'E(\boldsymbol{\varepsilon}|\mathbf{X})] = E_{\mathbf{X}}[\mathbf{0}] = \mathbf{0}.$$

Recall that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T},$$

therefore the OLS slope estimator is consistent,

$$plim\ \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \underline{\mathbf{Q}}^{-1}plim\ \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T} = \boldsymbol{\beta} + \underline{\mathbf{Q}}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}.$$

[6] When $\mathbf{x}_t$ is *i.i.d.* sample, then $\underline{\mathbf{Q}} = E(\mathbf{x}_t\mathbf{x}_t')$.

[7] This assumption is violated when $\mathbf{X}$ is an "$I(1)$" or an unit root process. See Chapter 21 for details.

Remark**.
To prove the unbiasedness of OLS under stochastic regressors, we need the assumption that $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$. However, to prove the consistency of OLS under stochastic regressors, we *only* need to show that $plim(\mathbf{X}'\boldsymbol{\varepsilon}/T) = \mathbf{0}$ which only requires that $plim(\mathbf{X}'\boldsymbol{\varepsilon}/T) = plim\ 1/T\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t = 1/T\sum_{t=1}^{T}E(\mathbf{x}_t\varepsilon_t) = \mathbf{0}$ or $E(X_{ti}\varepsilon_t) = 0,\ i = 1, 2, ..., k$. That is each regressor is *uncorrelated* with the disturbance at time t. i.e. *non contemporaneous correlation*. Hence if it is unbiased, then it is consistent:

$$\begin{aligned} E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0} \Longrightarrow E(\mathbf{X}'\boldsymbol{\varepsilon}) &= E(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_T \end{bmatrix} \\ &= E(\mathbf{x}_1\varepsilon_1 + \mathbf{x}_2\varepsilon_2 + ... + \mathbf{x}_T\varepsilon_T) = \mathbf{0}, \end{aligned}$$

but not vice versa. ■

Remark**.
If $\mathbf{x}_t$ and ε_t are independent and $E(\varepsilon_t) = 0$, then $E(\varepsilon_t|\mathbf{x}_t) = 0$.

Proof.

$$\begin{aligned} E(\varepsilon_t|\mathbf{x}_t) &= \int \varepsilon_t \frac{f(\varepsilon_t\mathbf{x}_t)}{f(\mathbf{x}_t)}d\varepsilon_t \\ &= \int \varepsilon_t \frac{f(\varepsilon_t)f(\mathbf{x}_t)}{f(\mathbf{x}_t)}d\varepsilon_t \\ &= \int \varepsilon_t f(\varepsilon_t)\varepsilon_t \\ &= E(\varepsilon_t) \\ &= 0. \end{aligned}$$

■

Remark**.
If $E(\varepsilon_t|\mathbf{x}_t) = 0$, then $E(\mathbf{x}_t\varepsilon_t) = \mathbf{0}$.

Proof.

$$\begin{aligned} E(\mathbf{x}_t\varepsilon_t) &= E\left[E(\mathbf{x}\varepsilon_t|\mathbf{x}_t)\right] \\ &= E\left[\mathbf{x}\cdot E(\varepsilon_t|\mathbf{x}_t)\right] \\ &= E(\mathbf{x}_t\cdot 0) \\ &= \mathbf{0}. \end{aligned}$$

■

Example. (Consistency but not Unbiasedness)[8]
Consider the $AR(1)$ model

$$Y_t = \phi Y_{t-1} + \varepsilon_t,$$

where $|\phi| < 1$ and ε_t is a white noise. Because Y_t dependent on Y_{t-1} and so on, there we lost the assumption $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$. It can be easily seen that

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \begin{bmatrix} E(\varepsilon_1|\mathbf{X}) = E(\varepsilon_1|Y_0, Y_1, ..., Y_T) \\ E(\varepsilon_2|\mathbf{X}) = E(\varepsilon_2|Y_0, Y_1, ..., Y_T) \\ \vdots \\ E(\varepsilon_T|\mathbf{X}) = E(\varepsilon_T|Y_0, Y_1, ..., Y_T) \end{bmatrix} \neq \mathbf{0}.$$

The OLS estimator $\hat{\phi}$ is not unbiased, but it is consistent in T. It is because the assumption for consistency that $E(Y_{t-1}\varepsilon_t) = 0$ is still valid. ■

Example.
The following Figures are to illustrate the bias but consistency of OLS from AR(1) process. The data generating process is

$$Y_t = 0.8Y_{t-1} + \varepsilon_t,\ t = 1, 2, ..., T,$$

where $\varepsilon_t \sim i.i.d.N(0, 2)$. It can be seen that the OLS is biased when a sample of size $T = 50$, but it is consistent as sample is large.
(a). Plot $T = 50$ for biasedness.
(b). Plot, $T = 50, 100, 300, 500$ and 1000 for consistency. Both are repeated with 1000 trials.

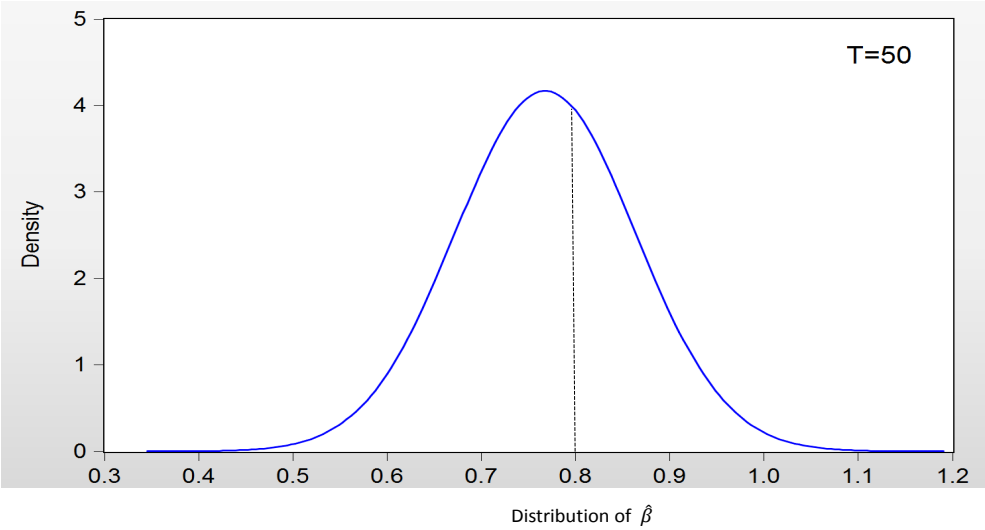

Figure (7-1a). Biasedness of $AR(1)$ Estimator, $\hat{\beta}$.

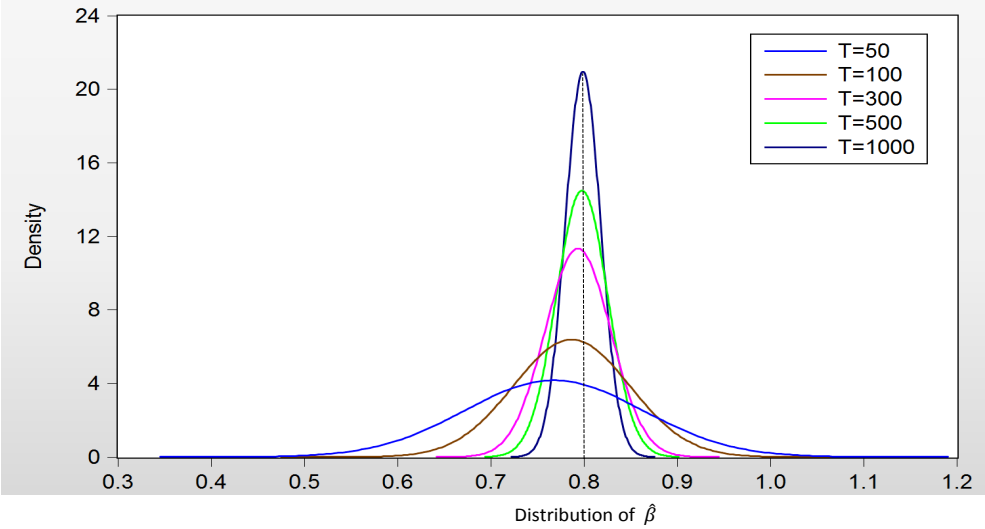

Figure (7-1b). Consistency of $AR(1)$ Estimator, $\hat{\beta}$ as T is large.

■

There are three common circumstances that one of the stochastic regressors at time t is correlated that with the disturbance terms of the same period, i.e. $E(X_{ti}\varepsilon_t) \neq 0$:

[8]See an illustration of the (downward) bias by Asatoshi Maeshiro, *The Journal of Economic Education* (2000), pp. 76-80.

(a). the lagged dependent variables with serial correlation disturbance,

(b). unobservable model and

(c). simultaneous equation model.[9] □

Under these circumstance, The OLS is not a consistent estimator and the *Instrumental Variables* (IV) estimators is proposed to encounter this problems. See the following section.

3.1.4 Distribution of the Estimators

The distribution of the estimators can be found by a two step procedure. The first step evaluates the distribution conditional on $\mathbf{X}$; that is, it treats $\mathbf{X}$ as deterministic just as in the earlier analysis. The second step multiplies by the density of $\mathbf{X}$ and integrates over $\mathbf{X}$ to find the unconditional distribution.

Result.
Let the random vector $(u, v)'$ with joint pdf $f(u, v)$, then the marginal density of u can be obtained by

$$f(u) = \int f(u, v)dv = \int f(u|v)g(v)dv,$$

here, $g(v)$ is the marginal density function of v. ■

For example, it is easy to see that

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

If the density is multiplied by the density of $\mathbf{X}$ and integrated over $\mathbf{X}$, the result is no longer a Gaussian distribution; thus, $\hat{\boldsymbol{\beta}}$ is non-Gaussian under Assumption 2.

On the other hand, it implies that

$$\mathbf{e}'\mathbf{e}|\mathbf{X} \sim \sigma^2 \cdot \chi^2(T - k).$$

[9]For example, let $Y_t = \beta X_t + \varepsilon_t$ and $X_t = \gamma Y_t + \eta_t$, where $E(\varepsilon_t \eta_t) = 0$. However it is easily to see that $E(X_t \varepsilon_t) = E((\gamma Y_t + \eta_t) \cdot \varepsilon_t) = \gamma \sigma_\varepsilon^2 \neq 0$.

But this density is the same for all $\mathbf{X}$. Thus, when we multiply the density of $\mathbf{e}'\mathbf{e}|\mathbf{X}$ by the density of $\mathbf{X}$ and integrate, we will get exact the same density. Hence unconditionally,

$$\frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{(T-k)s^2}{\sigma^2} \sim \chi^2(T-k).$$

The distribution of $\hat{\boldsymbol{\beta}}$ therefore depend on the stochastic properties of $\mathbf{X}$. It is pessimistic to say that the usual test of hypothesis will not be valid. However as we will see in the following , it is not the case.

3.1.5 Hypothesis Test

We now consider the validity of our sample test statistics and inference procedures when $\mathbf{X}$ is stochastic. First consider the conventional t test statistics for testing $H_0 : \beta_i = \beta_i^0$. Under the null hypothesis

$$t|\mathbf{X} = \frac{(\hat{\beta}_i - \beta_i^0)}{[s^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}]^{1/2}} \sim t_{T-k}. \tag{7-15}$$

However, what interests us is the marginal, that is the unconditional distribution of t. Since (7-15) holds for *any* given $\mathbf{X}$, it must hold for all $\mathbf{X}$ and (7-15) is an unconditional probability statement as well. A formal proof is as follow.

Proof.

Remember that if $W \sim t(n)$ then the density function of W would be:

$$f(w;n) = \frac{1}{\sqrt{(n\pi)}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+\frac{w^2}{n}\right)^{[(n+1)/2]}} \; n > 0, \; w \in \mathbb{R}.$$

Therefore we see that the distribution function $f(t|\mathbf{X})$ of the random variables $t|\mathbf{X}$ is not a function of $\mathbf{X}$.

Let $g(\mathbf{x})$ be the density function of $\mathbf{X}$, and the joint pdf of $\mathbf{X}$ and t is

$$f(t, \mathbf{X}) = f(t|\mathbf{X})g(\mathbf{X}),$$

therefore the marginal density of t is

$$\begin{aligned} f(t) &= \int f(t, \mathbf{X}) d\mathbf{X} \\ &= \int f(t|\mathbf{X})g(\mathbf{X})d\mathbf{X} \quad since\ f(t|\mathbf{X})\ is\ not\ a\ function of\ \mathbf{X} \\ &= f(t|\mathbf{X}) \int g(\mathbf{X})d\mathbf{X} = f(t|\mathbf{X}). \end{aligned}$$

■

We have the surprising results that, regardless of the distribution of $\mathbf{X}$, or even of whether $\mathbf{X}$ is stochastic or nonstochastic, the marginal distribution of t-ratio statistics is still t distribution. The same reason can be used to deduce that the usual F-ratio used for testing linear restrictions are valid whether $\mathbf{X}$ is stochastic or not.

Remark** .
The above conclusions only happen in the assumption that the disturbances are conditionally normally distributed. ■

3.2 Instrumental Variables Estimator

We again consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with the case in which the regressor matrix $\mathbf{X}$ is random. Frequently in linear models describing economic process, one or more of the regressors are *contemporaneously correlated* with the disturbance term. That is,

$$plim \frac{1}{T}\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{a.s.} \frac{1}{T}\sum_{t=1}^{T} E(\mathbf{x}_t\varepsilon_t) \neq 0,$$

so that the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X}/T)^{-1}(\mathbf{X}'\boldsymbol{\varepsilon}/T)$ is inconsistent.

Example. (Lagged Dependent Variables with Autocorrelated Disturbances)
Let the regression be

$$\begin{aligned} y_t &= \beta y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t. \end{aligned}$$

In this model the regressor and the disturbance are correlated:

$$\begin{aligned} Cov(y_{t-1}, \varepsilon_t) &= Cov(y_{t-1}, \rho\varepsilon_{t-1} + u_t) \\ &= \rho Cov(y_{t-1}, \varepsilon_{t-1}) \\ &= \frac{\rho\sigma_u^2}{(1-\beta\rho)(1-\rho^2)}. \end{aligned}$$

Since

$$plim\ \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{Cov(y_{t-1}, \varepsilon_t)}{var(y_t)},$$

therefore,

$$plim\ \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{\rho(1-\beta^2)}{1+\beta\rho}.$$

The OLS estimator $\hat{\boldsymbol{\beta}}$ is an inconsistent estimator of $\boldsymbol{\beta}$. ■

Example.

The following Figures are to illustrate the bias and inconsistency of OLS from AR(1) process. The data generating process is

$$Y_t = 0.8Y_{t-1} + \varepsilon_t,\ \varepsilon_t = 0.5\varepsilon_{t-1} + u_t,\ t = 1, 2, ..., T,$$

where $\varepsilon_t \sim i.i.d.N(0, 2)$. It can be seen that the OLS is biased with a sample of size $T = 50$, but it is also inconsistent even with a large sample.

(a). Plot $T = 50$ for biasedness.

(b). Plot, $T = 100, 300, 500$ and 1000 for inconsistency. Both are repeated with 1000 trials.

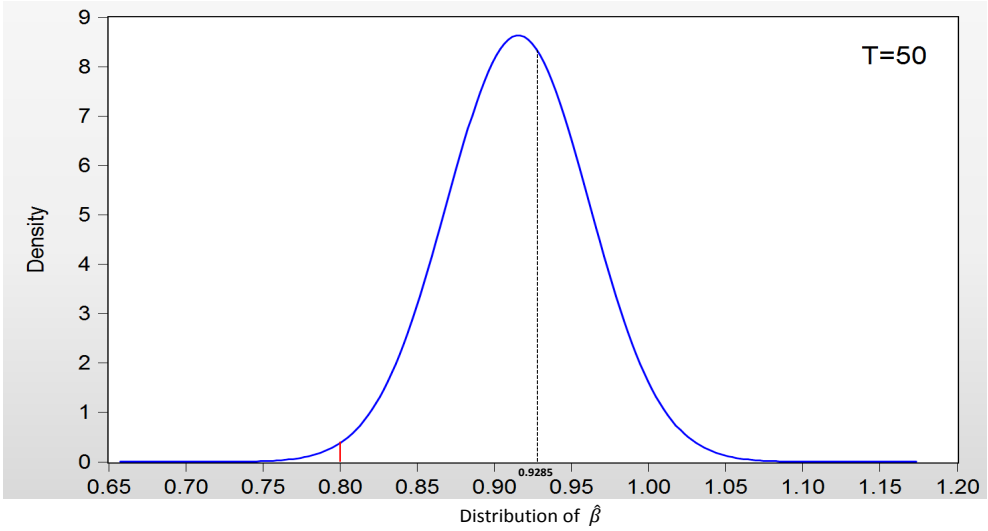

Figure (7-2a). Biasedness of $ARMA(1,1)$ Model Estimator, $E(\hat{\beta}) \neq 0.8$.

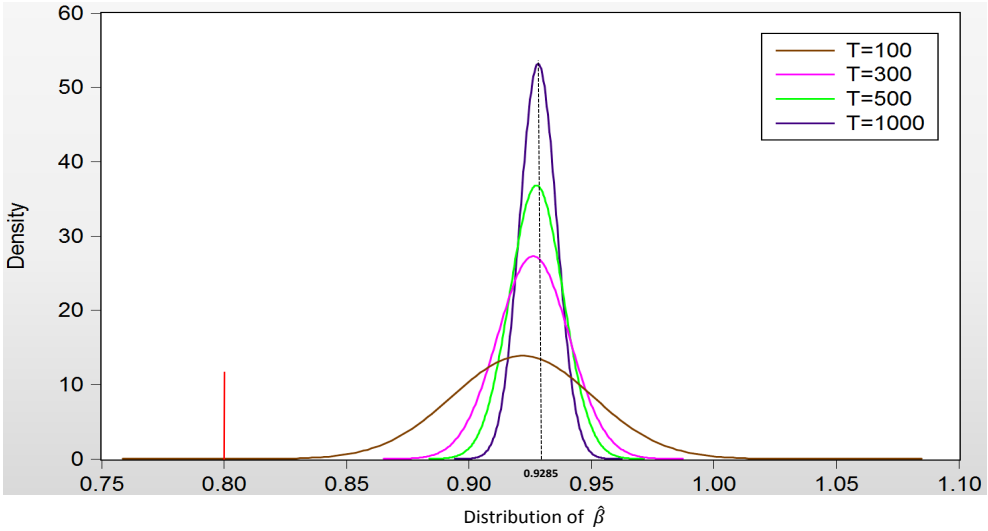

Figure (7-2b). Inconsistency of $ARMA(1,1)$ Model Estimator, $\hat{\beta} \xrightarrow{p} 0.9285$ (rather than $\hat{\beta} \xrightarrow{p} 0.8$) even as T is large.

■

To find a consistent estimator of $\boldsymbol{\beta}$ when the regressors and the disturbance are contemporaneously correlated, i.e. $E(\mathbf{x}_t\varepsilon_t) \neq \mathbf{0}$, we may find a set of variables $\mathbf{z}_t$ to "replace" $\mathbf{x}_t$ such that $\mathbf{z}_t$ is uncorrelated with ε_t but it is correlated with $\mathbf{x}_t$. Formally, it is stated in the following theorem.

Theorem (Instrument Variables Estimator).
Assume the linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $plim\frac{1}{T}\mathbf{X}'\boldsymbol{\varepsilon} \neq 0$. Suppose that there exists a set of T observations on k variables $\mathbf{z}_t$, $t = 1, 2, ..., T$; *i.e.* $\mathbf{Z} = (\mathbf{z}_1\ \mathbf{z}_2\ ....\ \mathbf{z}_T)'$ such that

$$\begin{aligned} plim\ \frac{1}{T}\mathbf{Z}'\mathbf{Z} &= \underline{\mathbf{Q}}_{ZZ}, \text{ a finite, positive definite matrix (well behaved data)}, \\ plim\ \frac{1}{T}\mathbf{Z}'\mathbf{X} &= \underline{\mathbf{Q}}_{ZX}, \text{ a finite, } k \times k \text{ matrix with rank k (relevance)}, \\ plim\ \frac{1}{T}\mathbf{Z}'\boldsymbol{\varepsilon} &= \mathbf{0}, \text{ such that } \frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{\sqrt{T}} \xrightarrow{L} N(0, \boldsymbol{\Phi}), \text{ (exogeneity)}, \end{aligned}$$

then the estimator[10]

$$\tilde{\boldsymbol{\beta}}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

is consistent, and the asymptotic distribution of $\sqrt{T}(\tilde{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta})$ is $N(\mathbf{0}, \underline{\mathbf{Q}}_{ZX}^{-1}\boldsymbol{\Phi}\ \underline{\mathbf{Q}}_{ZX}'^{-1})$. The estimator $\tilde{\boldsymbol{\beta}}_{IV}$ is the instrumental variables (IV) estimator of $\boldsymbol{\beta}$. The matrix $\mathbf{Z}$ is the set of instruments for $\mathbf{X}$.

Proof.
Note that

$$\tilde{\boldsymbol{\beta}}_{IV} = \boldsymbol{\beta} + \left(\frac{\mathbf{Z}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{T}.$$

Hence

$$\begin{aligned} plim\ \tilde{\boldsymbol{\beta}}_{IV} &= \boldsymbol{\beta} + plim\left(\frac{\mathbf{Z}'\mathbf{X}}{T}\right)^{-1} plim\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{T}\right) \\ &= \boldsymbol{\beta} + \underline{\mathbf{Q}}_{ZX}^{-1} \cdot \mathbf{0} \\ &= \boldsymbol{\beta}. \end{aligned}$$

To prove the second part of the theorem, note that

$$\sqrt{T}(\tilde{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) = \left(\frac{\mathbf{Z}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{\sqrt{T}}.$$

[10]Consider the "normal equation": $\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\varepsilon}$, the IV estimator follows intuitively.

The result follows immediately. ■

Result.

The typical case is the one in which $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$, then $\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{\sqrt{T}}$ has asymptotic distribution $N(\mathbf{0}, \sigma^2\underline{\mathbf{Q}}_{ZZ})$, and

$$\sqrt{T}(\tilde{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \sigma^2\underline{\mathbf{Q}}_{ZX}^{-1}\underline{\mathbf{Q}}_{ZZ}\underline{\mathbf{Q}}_{ZX}'^{-1})$$

or

$$\tilde{\boldsymbol{\beta}}_{IV} \xrightarrow{asy} N\left(\boldsymbol{\beta},\ \frac{\sigma^2}{T}\underline{\mathbf{Q}}_{ZX}^{-1}\underline{\mathbf{Q}}_{ZZ}\underline{\mathbf{Q}}_{ZX}'^{-1}\right). \quad ■$$

To estimate the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}_{IV}$ in the above special case, we will require an estimator of σ^2. The natural estimator is

$$\tilde{\sigma}_{IV}^2 = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \mathbf{x}_t'\tilde{\boldsymbol{\beta}}_{IV})^2,$$

and we will estimate $asy\ Var(\tilde{\boldsymbol{\beta}}_{IV})$ with

$$asy\ Var(\tilde{\boldsymbol{\beta}}_{IV}) = \tilde{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}.$$

3.2.1 Two-Stage Least Square Estimator, 2SLS

If $\mathbf{Z}$ contains more variable ($l > k$) than $\mathbf{X}$, one obvious possibility is simply to choose k variables among the l variables in $\mathbf{Z}$. But intuition correctly suggests that throwing away the information contained in the remaining $l - k$ columns is inefficient. A better way we may choose the projection of the column of $\mathbf{X}$ in the column space of $\mathbf{Z}$:[11]

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

[11] To see this, we try to find $\hat{\mathbf{X}}$ to "represent" $\mathbf{X}$ from $\mathbf{Z}$ which can be formed from a linear combination of $\mathbf{Z}$, i.e $\hat{\mathbf{X}}_{T\times k} = \mathbf{Z}_{T\times l}\hat{\boldsymbol{\Lambda}}_{l\times k}$ such that the "forecast error" $\mathbf{X} - \hat{\mathbf{X}}$ is orthogonal to the "information" $\mathbf{Z}$, i.e.

$$\mathbf{Z}'(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\Lambda}}) = \mathbf{0}.$$

Hence, $\hat{\boldsymbol{\Lambda}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ and therefore $\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\Lambda}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$.

With the choice of instrumental variable, $\hat{\mathbf{X}}$ for $\mathbf{Z}$ we have

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{IV} = \tilde{\boldsymbol{\beta}}_{2SLS} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.\end{aligned}$$

That is, $\tilde{\boldsymbol{\beta}}_{IV}$ can be computed in two steps, first by computing $\hat{\mathbf{X}}$, then by the least squares regression. For this reason, this is called the two-stage least squares (2SLS) estimator.

To estimate the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}_{2SLS}$, we will require an estimator of σ^2. The natural estimator is

$$\tilde{\sigma}^2_{2SLS} = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \mathbf{x}_t'\tilde{\boldsymbol{\beta}}_{2SLS})^2,$$

and we will estimate $asy\ Var(\tilde{\boldsymbol{\beta}}_{2SLS})$ with

$$asy\ Var(\tilde{\boldsymbol{\beta}}_{2SLS}) = \tilde{\sigma}^2_{2SLS}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}. \tag{7-16}$$

One should be careful of this approach, however, in the computation of the asymptotic covariance matrix; $\tilde{\sigma}^2_{2SLS}$ should not be based on $\hat{\mathbf{X}}$. The estimator

$$s^2_{2SLS} = \frac{1}{T}(\mathbf{y} - \hat{X}\tilde{\boldsymbol{\beta}}_{2SLS})'(\mathbf{y} - \hat{X}\tilde{\boldsymbol{\beta}}_{2SLS}) = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{\mathbf{x}}_t'\tilde{\boldsymbol{\beta}}_{2SLS})^2,$$

is inconsistent for σ^2, with or without a correction for degrees of freedom.

Exercise 2.

Prove the result in equation (7-16). ■

Exercise 3.

Reproduce the results in Table 12.1 (p.321) of Greene 6th edition. ■

4 Non-Normal Disturbance

In this section we will suppose that all of the ideal conditions hold except that the disturbances are not normally distributed. In particular, we still suppose that the disturbance ε_t is independent and identically distributed with zero mean and finite variance σ^2; however, we no longer suppose that their distribution is normal.

4.1 Unbiasedness, Efficiency, and Consistency?

It is easy to show the OLS properties in the following.

Theorem.
The OLS coefficient estimator, $\hat{\boldsymbol{\beta}}$, is unbiased, BLUE, consistent, and has covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The variance estimator, s^2, is unbiased and consistent. ■

4.2 Hypothesis Testing in Finite Sample

Theorem.
Since $\boldsymbol{\varepsilon}$ is not normally distributed, $\hat{\boldsymbol{\beta}}$ is not normally distributed and therefore $(T-k)s^2/\sigma^2$ doesn't have a χ^2 distribution. Therefore the test hypothesis t and F statistics developed in section 4 of Chapter 6 are not valid when $\boldsymbol{\varepsilon}$ is not normally distributed.■

4.3 Hypothesis Testing in Large Sample, Asymptotically Normality

Although the finite sample test statistics are no longer valid when the disturbance is not normally distributed, the following results show that the usual test procedures are *asymptotically* justified whether or not the disturbance is normal.

Lemma.
Assume that $\varepsilon_t, t = 1, 2, ..., T$ are independent and identically distributed with mean 0 and variance $\sigma^2 < \infty$, and $\lim_{T\to\infty}(\mathbf{X}'\mathbf{X}/T) = \underline{\mathbf{Q}}$, a finite positive definite matrix, then

$$\left(\frac{1}{\sqrt{T}}\mathbf{X}'\boldsymbol{\varepsilon}\right) \xrightarrow{d} N(\mathbf{0}, \sigma^2\underline{\mathbf{Q}}).$$

Proof.
Denote

$$\left(\frac{1}{\sqrt{T}}\mathbf{X}'\boldsymbol{\varepsilon}\right) = \sqrt{T}(\bar{\mathbf{w}} - E(\bar{\mathbf{w}})),$$

where

$$\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t$$

is the average of T independent random vector $\mathbf{x}_t\varepsilon_t$ with mean $E(\mathbf{x}_t\varepsilon_t) = \mathbf{x}_t E(\varepsilon_t) = \mathbf{0}$ (hence $E(\bar{\mathbf{w}}) = \mathbf{0}$); and variance $Var(\mathbf{x}_t\varepsilon_t) = \sigma^2\mathbf{x}_t\mathbf{x}_t' = \sigma^2\mathbf{Q}_t$.[12] The mean of $\sqrt{T}\bar{\mathbf{w}}$ is $\mathbf{0}$ and the variance of $\sqrt{T}(\bar{\mathbf{w}})$ is

$$\begin{aligned} Var\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right) &= \sigma^2\left(\frac{1}{T}\right)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' \\ &= \sigma^2\left(\frac{1}{T}\right)[\mathbf{Q}_1 + \mathbf{Q}_2 + ... + \mathbf{Q}_T] \\ &= \sigma^2\underline{\mathbf{Q}}_T \\ &= \sigma^2\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right). \end{aligned}$$

Assume that $\lim_{T\to\infty}(\mathbf{X}'\mathbf{X}/T) = \underline{\mathbf{Q}}$, a positive definite matrix, then

$$\lim_{T\to\infty}\sigma^2\underline{\mathbf{Q}}_T = \sigma^2\underline{\mathbf{Q}}.$$

By Lindberg-Feller CLT to the vector $\sqrt{T}\bar{\mathbf{w}}$, we obtain the results. ■

[12] It is to be noted that here $\mathbf{w}_t$'s are independent but heterogeneously distributed. We apply the Lindberg-Feller rather Lindberg-Levy CLT in the following.

Theorem.
The asymptotic distribution of $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is $N(\mathbf{0}, \sigma^2 \underline{\mathbf{Q}}^{-1})$.

Proof.
Note that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T},$$

or

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T}.$$

Multiple the above equation by $\sqrt{T}$ we have

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\sqrt{T}\cdot\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T} = \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{1}{\sqrt{T}}\mathbf{X}'\boldsymbol{\varepsilon}.$$

Then as $T \to \infty$,

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\xrightarrow{d} \underline{\mathbf{Q}}^{-1}\cdot N(\mathbf{0}, \sigma^2\underline{\mathbf{Q}}) \\ &\equiv N(\mathbf{0},\ \sigma^2\underline{\mathbf{Q}}^{-1}\underline{\mathbf{Q}}\underline{\mathbf{Q}}^{-1}) \\ &= N(\mathbf{0},\ \sigma^2\underline{\mathbf{Q}}^{-1}).\end{aligned}$$

■

The above results show that $\hat{\boldsymbol{\beta}}$ has the same asymptotic distribution whether or not the distribution is normal, as long as they are *i.i.d.* with mean zero and finite variance.

4.3.1 Tests of a Single Linear Combination on $\boldsymbol{\beta}$: the t-ratio Test Statistics

The following results show that this implies that the usual t tests are asymptotically valid; the t distribution is of course replaced by the $N(0,1)$ distribution, however.

Theorem.
Suppose that the element of $\boldsymbol{\varepsilon}$ are *i.i.d.* with zero mean and finite variance, and that

$\underline{\mathbf{Q}}$ is finite and nonsingular. Let R be a known $1 \times k$ vector, and r be a known scalar. Then under the null hypothesis that $R\boldsymbol{\beta} = r$, the t-ratio test statistic

$$\frac{R\hat{\boldsymbol{\beta}} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \xrightarrow{d} N(0,1).$$

Proof.

Under the null hypothesis, $R\hat{\boldsymbol{\beta}} - r = R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, so the test statistic is

$$\frac{\sqrt{T}R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X}/T)^{-1}R'}}.$$

Now, since $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0,\ \sigma^2\underline{\mathbf{Q}}^{-1})$, it is clear that

$$\sqrt{T}R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0,\ \sigma^2 R\underline{\mathbf{Q}}^{-1}R').$$

Also, the denominator of the t-ratio test statistic has probability limit

$$\sqrt{\sigma^2 R\underline{\mathbf{Q}}^{-1}R'},$$

since $s^2 \to \sigma^2$ and $(\mathbf{X}'\mathbf{X}/T)^{-1} \to \underline{\mathbf{Q}}^{-1}$. But this is a scalar which is in effect the standard deviation of the asymptotic distribution of the numerator random variable $\sqrt{T}R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Hence the test statistics do indeed converge asymptotically to $N(0,1)$.■

As should be clear, the test is only asymptotically valid. In any finite sample the normal distribution is only an approximation to the distribution of the test statistic.

4.3.2 Tests of Several Linear Restrictions on $\boldsymbol{\beta}$: the F-ratio Test Statistics

The above result showed that the usual t-ratio test is asymptotically justified even if the disturbance is not normal, as long as they are *i.i.d.*. The following result shows that this is also true of the usual F-ratio test.

Theorem.

Suppose that the elements of $\boldsymbol{\varepsilon}$ are *i.i.d.* with zero mean and finite variance, and that $\underline{\mathbf{Q}}$ is finite and nonsingular. Let $\mathbf{R}$ be a known $m \times k$ vector, and $\mathbf{q}$ be a known vector. Then under the null hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the Wald test statistic

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = m \cdot \text{F-statistics} \xrightarrow{d} \chi^2_m.$$

Proof.

Because $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0},\ \sigma^2\underline{\mathbf{Q}}^{-1})$, it is clear that under the null hypothesis

$$\sqrt{T}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) \xrightarrow{d} N(\mathbf{0},\ \sigma^2\mathbf{R}\underline{\mathbf{Q}}^{-1}\mathbf{R}').$$

Recall from the results on p.63 of Chapter 2 (distribution of a full rank quadratic form) we have

$$T(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'(\sigma^2\mathbf{R}\underline{\mathbf{Q}}^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) \xrightarrow{d} \chi^2_m. \tag{7-17}$$

Because

$$plim\ s^2\left(\frac{1}{T}\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2\underline{\mathbf{Q}}^{-1},$$

then the statistic obtained by replacing $\sigma^2\underline{\mathbf{Q}}^{-1}$ by $s^2\left(\frac{1}{T}\mathbf{X}'\mathbf{X}\right)^{-1}$ in (7-17) has the same limiting distribution. The T's cancel, and we are left with the results proposed. ■

The Tunnel Entrance to NSYSU.

End of this Chapter