Advanced Econometrics

Professor: Julie Shi

Timekeeper: Zircon

Fall 2023

Contents

1 The Linear Regression Model

The multiple linear regression model is used to study the relationship between a dependent variable and one or more independent variables. The generic form of the linear regression model is

$$\begin{aligned} y &= f(x_1, x_2, \cdots, x_K) + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \varepsilon \end{aligned}$$

where y is the dependent or explained variable and $x_1, \cdots, x_K$ are the independent or explanatory variables. One's theory will specify $f(x_1, x_2, \cdots, x_K)$. This function is commonly called the **population regression equation** of y on $x_1, \cdots, x_K$. In this setting, y is the *regressand* and $x_k (k = 1, \cdots, K)$ are the *regressors* or covariates. The underlying theory will specify the dependent and independent variables in the model, though it is not always obvious which is appropriately defined as each of these. The term ε is a random disturbance, so named because it "disturbs" an otherwise stable relationship. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model, no matter how elaborate.

We assume that each observation in a sample $(y_i, x_{i1}, x_{i2}, \cdots, x_{iK})$ for $i = 1, \cdots, n$, is generated by an underlying process described by

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$$

The observed value of y_i is the sum of two parts, a deterministic part and the random part ε_i. Our objective is to estimate the unknown parameters of the model, use the data to study the validity of the theoretical propositions, and perhaps use the model to predict the variable y. How we proceed from here depends crucially on what we assume about the stochastic process that has led to our observations of the data in hand.

1.1 Assumptions of Linear Regression Model

1.1.1 Linearity

Let the column vector $\mathbf{x}_k$ be the n observations on the variable x_k, for $k = 1, \cdots, K$, and assemble these data in a $n \times K$ data matrix X. In most contexts, the first column of X is assumed to be a column of 1s so that β_1 is the constant term in the model. Let y be the n observations, $y_1, \cdots, y_n$, and let ε be the column vector containing the n disturbances. The linear model as it applies to all n observations can now be written as:

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots \mathbf{x}_K\beta_K + \varepsilon$$

Or in the matrix form of Assumption 1:

$$y = X\beta + \varepsilon$$

The linearity assumption is not so narrow as it might first appear. In the regression context, linearity refers to the manner in which the parameters and the disturbance enter the equation, not necessarily to the relationship among the variables.

1.1.2 Full Rank

Assumption 2 is that there are no exact linear relationships among the variables.

$$X \text{ is an } n \times K \text{ matrix with rank } K.$$

Hence, X has full column rank; the columns of X are linearly independent and there are at least K observations. This assumption is known as an **identification condition**. If there are fewer than K observations, then X cannot have full rank. Hence, we make the (redundant) assumption that n is at least as large as K.
In a two-variable linear model with a constant term, the full rank assumption means that there must be variation in the regressor x. If there is no variation in x, then all our observations will lie on a vertical line. This situation does not invalidate the other assumptions of the model; presumably, it is a flaw in the data set. The possibility that this suggests is that we *could have* drawn a sample in which there was variation in x, but in this instance, we did not. Thus, the model still applies, but we cannot learn about it from the data set in hand. Another interpretation is, in this case the invariant variable is collinear with the constant term.

Technically speaking, the OLS estimator for coefficients, $b = (X'X)^{-1}X'y$, requires the inverse of $(X'X)^{-1}$ exists. Since $\text{rank}\left((X'X)^{-1}\right) = \text{rank}\,(X)$, so naturally we require X be full column rank.

1.1.3 (Zero) Condition Mean

Assumption 3 is that, the disturbance is assumed to have conditional expected value zero at every observation, which we write as

$$\mathrm{E}\left[\varepsilon_i | X\right] = 0$$

For the full set of observations, we write Assumption 3 as

$$\mathrm{E}\left[\varepsilon | X\right] = \begin{bmatrix} \mathrm{E}\left[\varepsilon_1 | X\right] \\ \mathrm{E}\left[\varepsilon_2 | X\right] \\ \vdots \\ \mathrm{E}\left[\varepsilon_n | X\right] \end{bmatrix} = \mathbf{0}$$

There is a subtle point worth noting. The left-hand side states, in principle, that the mean of each ε_i conditioned on *all observations* $\mathbf{x}_i$ is zero. This conditional mean assumption states, in words, that no observations on X convey information about the

expected value of the disturbance. It is conceivable—for example, in a time-series setting—that although $\mathbf{x}_i$ might provide no information about $\mathrm{E}[\varepsilon_i|\cdot]$, $\mathbf{x}_j$ at some other observation, such as in the next time period, might. Our assumption at this point is that there is no information about $\mathrm{E}[\varepsilon_i|\cdot]$ contained in any observation $\mathbf{x}_j$. We will also assume that the disturbances convey no information about each other. That is, $\mathrm{E}[\varepsilon_i|\varepsilon_1, \cdots, \varepsilon_{i-1}, \varepsilon_{i+1}, \cdots, \varepsilon_n] = 0$. In sum, at this point, we have assumed that the disturbances are purely random draws from some population.
The zero conditional mean implies that the unconditional mean is also zero, since

$$\mathrm{E}[\varepsilon_i] = \mathrm{E}_X[\mathrm{E}[\varepsilon_i|X]] = \mathrm{E}_X[0] = 0$$

This assumption also means there is no correlation between regressors and disturbances:

$$\mathrm{Cov}[\mathrm{E}[\varepsilon_i|X], X] = \mathrm{Cov}[\varepsilon_i, X] = 0$$

Notice that the converse is not true: $\mathrm{E}[\varepsilon_i] = 0$ cannot imply $\mathrm{E}[\varepsilon_i|X] = 0$.

Later we will see, zero conditional mean assumption is critical for unbiasedness of OLS estimators.

In most cases, the zero overall mean assumption is not restrictive. The mean could have been something else. But, if the original model does not contain a constant term, then assuming $\mathrm{E}[\varepsilon_i] = 0$ could be substantive. This suggests that there is a potential problem in models without constant terms. As a general rule, regression models should not be specified without constant terms unless this is specifically dictated by the underlying theory. Arguably, if we have reason to specify that the mean of the disturbance is something other than zero, we should build it into the systematic part of the regression, leaving in the disturbance only the unknown part of ε. Assumption 3 also implies that

$$\mathrm{E}[y|X] = X\beta$$

Assumptions 1 and 3 comprise the *linear regression model*. The regression of y on X is the conditional mean, $\mathrm{E}[y|X]$, so that without assumption 3, $X\beta$ is not the conditional mean function.

The remaining assumptions will more completely specify the characteristics of the disturbances in the model and state the conditions under which the sample observations on $\mathbf{x}$ are obtained.

1.1.4 Spherical Disturbances

Assumption 4 concerns the variances and covariances of the disturbances:

$$\begin{cases} \text{Var}\left[\varepsilon_i|X\right] = \sigma^2, \forall i = 1, \cdots, n \\ \text{Cov}\left[\varepsilon_i, \varepsilon_j\right] = 0, \forall i \neq j \end{cases}$$

Constant variance is labeled homoskedasticity. Uncorrelatedness across observations is labeled generically nonautocorrelation. Disturbances that meet the assumptions of homoskedasticity and nonautocorrelation are sometimes called spherical disturbances. The two assumptions imply that

$$\text{E}\left[\varepsilon\varepsilon'|X\right] = \begin{bmatrix} \text{E}\left[\varepsilon_1\varepsilon_1'|X\right] & \text{E}\left[\varepsilon_1\varepsilon_1'|X\right] & \cdots & \text{E}\left[\varepsilon_1\varepsilon_1'|X\right] \\ \text{E}\left[\varepsilon_2\varepsilon_1'|X\right] & \text{E}\left[\varepsilon_2\varepsilon_1'|X\right] & \cdots & \text{E}\left[\varepsilon_2\varepsilon_n'|X\right] \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}\left[\varepsilon_n\varepsilon_1'|X\right] & \text{E}\left[\varepsilon_n\varepsilon_1'|X\right] & \cdots & \text{E}\left[\varepsilon_n\varepsilon_n'|X\right] \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

$$\iff \text{Var}\left[\text{E}\left[\varepsilon|X\right]\right] = \sigma^2\mathbf{I}$$

Homoskedasticity guarantees the efficiency of OLS estimators, as we will see in Gauss-Markov theorem.

By using the variance decomposition formula, we find

$$\text{Var}\left[\varepsilon\right] = \text{E}\left[\text{Var}\left[\varepsilon|X\right]\right] + \text{Var}\left[\text{E}\left[\varepsilon|X\right]\right] = \sigma^2\mathbf{I}$$

Once again, we should emphasize that this assumption describes the information about the variances and covariances among the disturbances that is provided by the independent variables. For the present, we assume that there is none. We will also drop this assumption later when we enrich the regression model. We are also assuming that the disturbances themselves provide no information about the variances and covariances. Although a minor issue at this point, it will become crucial in our treatment of time-series applications.

1.2 OLS Estimation

1.2.1 Derivation of Coefficients Estimation

The unknown parameters of the stochastic relationship $y_i = x_i'\beta + \varepsilon_i$ are the objects of estimation. It is necessary to distinguish between population quantities, such as β and ε_i, and sample estimates of them, denoted as b and e_i. The population regression is

$$\text{E}\left[y_i|x_i\right] = x_i'\beta$$

whereas our estimate of $\text{E}\left[y_i|x_i\right]$ is denoted:

$$\widehat{y}_i = x_i'b$$

The disturbance associated with the i-th data point is

$$\varepsilon_i = y_i - x_i'\beta$$

For any value of b, we shall estimate ε_i with the *residual* defined as

$$e_i = y_i - x_i b$$

From the definitions,

$$y_i = x_i'\beta + \varepsilon_i = x_i'b + e_i$$

The population quantity β is a vector of unknown parameters of the probability distribution of y_i whose values we hope to estimate with our sample data $(y_i, x_i), i = 1, 2, \cdots, n$. This is a problem of statistical inference. However, it is instructive to begin by considering the purely algebraic problem of choosing a vector b so that the fitted line $x_i'b$ is close to the data points. The measure of closeness (or, deviations) constitutes a fitting criterion. Among them, the one used most frequently is **least squares**.

The least squares coefficient vector minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n} (y_i - x_i'b_0)^2$$

where b_0 denotes the choice for the coefficient vector. In matrix terms, minimizing the sum of squares requires us to choose b_0 to:

$$\begin{aligned}\min_{b_0} S(b_0) = e_0'e_0 &= (y - Xb_0)'(y - Xb_0) \\ &= y'y - b_0'X'y - y'Xb_0 + b_0'X'Xb_0 \\ &= y'y - 2y'Xb_0 + b_0'X'Xb_0\end{aligned}$$

Notice the third equality holds because both $b_0'X'y$ and $y'Xb_0$ are numbers, or 1×1 matrix.

The necessary condition for a minimum is:

$$\frac{\partial S(b_0)}{\partial b_0} = -2X'y + 2X'Xb_0 = \mathbf{0}$$

Let b be the solution. Then we find that b satisfies the **least squares normal equation**:

$$X'Xb = X'y$$

The normal equation lays the foundation for coefficient estimation and is therefore very important.

If the inverse of $X'X$ exists, which follows from the full column rank assumption of X, then the solution is:

$$b = (X'X)^{-1}X'y$$

From this solution to minimize the sum of squares,

$$\frac{\partial^2 S(b_0)}{\partial b_0 \partial b_0'} = 2X'X$$

which must be a positive definite matrix. To see this, let $q = c'X'Xc$ for some arbitrary nonzero vector c. Then

$$q = c'X'Xc = (Xc)'(Xc) = ||Xc||^2$$

Since X is of full column rank, then Xc cannot be $\mathbf{0}$ for any nonzero c. Thus, q must be positive for any nonzero vector c. Therefore, we can safely conclude that if X has full column rank, then the least squares solution b is unique and minimizes the sum of squared residuals.

Remarks: Our deduction after getting b aims to prove that b does minimize the sum of squared residuals, and b is unique. This is because in the preceding parts, although we have obtained a local optimizer, we do not know whether b is a maximizer and a minimizer, and whether b is unique.

Here we summarize our results in this part:

OLS Estimator and Normal Equation Suppose we hope to estimate $y = X\beta + \varepsilon$, with criterion of minimizing sum of squared residuals. The normal equation is given by

$$X'Xb = X'y$$

Under assumption of X's full column rank, it follows that

$$b = (X'X)^{-1}X'y$$

1.2.2 Projector and Residual Maker

The vector of least squares residuals is

$$\begin{aligned} e &= y - Xb \\ &= y - X(X'X)^{-1}X'y \\ &= \{I - X(X'X)^{-1}X'\}y \\ &:= My \end{aligned}$$

The $n \times n$ matrix M defined as $M = I - X(X'X)^{-1}X'$ is fundamental in regression analysis. In view of $e = My$, we can interpret M as a matrix that produces the vector of least squares residuals in the regression of y on X when it premultiplies any vector y. Thus, M is called the "residual maker". The residual maker is perpendicular to column space of X:

$$MX = \mathbf{0}$$

One way to interpret this result is that, if X is regressed on X, a perfect fit will result and the residuals will naturally be $\mathbf{0}$.

Moreover, we can check that M is symmetric and idempotent, and M is then non-negative definite. The properties of M are quite useful in following theories on least squares.

Coincidentally, our definition of estimated linear model means that y is partitioned into two meaningful parts: the fitted values $\hat{y} = Xb$ and the residual e. Since $MX = 0$, the two parts are orthogonal.

$$\hat{y} = y - e = (I - M)y = X(X'X)^{-1}X'y = Py$$

The matrix P is a **projection matrix**. It is the matrix formed from X such that when a vector y is premultiplied by P, the result is the fitted values in the least squares regression of y on X. This is also the *projection* of the vector y onto the column space of X. From the projection meaning in P, it is quick to verify that

$$PX = X$$

In meanings of vectors spaces, space of regressors and space of residuals are orthogonal to each other:

$$PM = MP = 0$$

The residual maker and projector sum up to construct the original space:

$$P + M = I$$

Same as M, P is symmetric and idempotent, and P is thus non-negative definite. The transformation of P and M are crucial in regression analysis.

By far, we can see the partition of y is quite meaningful:

$$y = Py + My = \text{Projection} + \text{Residual}$$

Graphically speaking, the column space of X is perpendicular to the residuals'. In mathematical forms, we can also see the Pythagorean theorem at work in the sum of squares:

$$\begin{aligned} y'y &= y'P'Py + y'M'My \\ &= \hat{y}'\hat{y} + e'e \end{aligned}$$

In manipulating equations involving least squares results, the following equivalent expressions for the sum of squares residuals are often useful:

$$\begin{aligned} e'e &= y'M'My = y'My = y'e = e'y \\ e'e &= y'y - b'X'Xb = y'y - b'X'y = y'y - y'Xb \end{aligned}$$

Here we summarize our results in this part:

Projector and Residual Maker If we regress y on X, we define the projector P and residual maker M as

$$\begin{cases} P = X(X'X)^{-1}X' \\ M = \mathbf{I} - X(X'X)^{-1}X' \end{cases}$$

Both P and M are symmetric, idempotent and thus non-negative definite. Moreover, Py will return the fitted values in regression of y on X, and My gives the residuals. P and M correspond to two orthogonal spaces of variation of y. And

$$PX = X, MX = \mathbf{0}$$

1.2.3 Partitioned Regression

1.2.3.1 FWL Theorem Suppose that the regression involves two sets of variables X_1 and X_2. Thus,

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

In order to obtain the algebraic solutions for b_1 and b_2, we start from normal equation in partitioned form:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

A solution can be obtained using the partitioned inverse matrix. Alternatively, the two partitioned equations can be manipulated directly to solve for b_2. We first solve for b_1:

$$b_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2b_2 = (X_1'X_1)^{-1}X_1'(y - X_2b_2)$$

This solution says that b_1 is the set of coefficients in the regression of y on X_1, *minus* a correction vector. However, if $X_1'X_2 = \mathbf{0}$, i.e., X_1 and X_2 are orthogonal, then b_1 can be simplified to: $b_1 = (X_1'X_1)^{-1}X_1'y$. The general result is given in the following theorem.

Orthogonal Partitioned Regression In the multiple linear least squares regression of y on two sets of variables, X_1 and X_2, if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of y on X_1 alone and y on X_2 alone.

In most cases, the two sets of variables X_1 and X_2 are not orthogonal, then the solution for b_1 and b_2 would be much more involved than just simple regressions. The more general solution is given by the following theorem.

Frisch-Waugh-Lovell Theorem In the linear least squares regression of vector y on two sets of variables, X_1 and X_2, the subvector b_2 is the set of coefficients obtained when the residuals from a regression of y on X_1 alone are regressed on the set of residuals obtained when each column of X_2 is regressed on X_1.

Proof From the two sub-equations in the partitioned normal equation, we obtain b_1 from one equation and plug in expression of b_1 into another equation. After collecting terms, the solution for b_2 is:

$$\begin{aligned} b_2 &= \{X_2(I - X_1(X_1'X_1)^{-1}X_1')X_2\}^{-1}\{X_2'(I - X_1(X_1'X_1)^{-1}X_1')y\} \\ &= (X_2'M_1X_2)^{-1}(X_2'M_1y) \end{aligned}$$

where M_1 is the residual maker for regression on X_1. Notice that M_1X_2 is a matrix of residuals; each column of M_1X_2 is a vector of residuals in the regression of the corresponding column of X_2 on the variables in X_1.

By exploiting the fact that M_1 is symmetric and idempotent, the result can be rewritten as:

$$b_2 = (X_2^{*'}X_2^*)^{-1}X_2^{*'}y^*$$

where $X_2^* = M_1X_2$, and $y^* = M_1y$.

This process described by FWL theorem is commonly called *partialling out* or *netting out* the effect of X_1. For this reason, the coefficients in a multiple regression are often called the *partial regression coefficients*.

It is not hard to compute the expectation and variance of b_2, and verify that b_2 is

unbiased.

$$
\begin{aligned}
\mathrm{E}\left[b_2|X\right] &= \mathrm{E}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1y\right)|X\right] = \mathrm{E}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1\left(X_1\beta_1+X_2\beta_2+\varepsilon\right)\right)|X\right] \\
&= \mathrm{E}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1X_1\beta_1+X_2'M_1X_2\beta_2+X_2'M_1\varepsilon\right)|X\right] \\
&= \mathrm{E}\left[0+\beta_2+\left(X_2'M_1X_2\right)^{-1}X_2'M_1\varepsilon|X\right] \\
&= \beta_2 \\
\mathrm{Var}\left[b_2|X\right] &= \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1y\right)|X\right] = \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1\left(X_1\beta_1+X_2\beta_2+\varepsilon\right)\right)|X\right] \\
&= \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1X_1\beta_1+X_2'M_1X_2\beta_2+X_2'M_1\varepsilon\right)|X\right] \\
&= \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1\varepsilon\right)|X\right] \\
&= \left(X_2'M_1X_2\right)^{-1}X_2'M_1\left(\mathrm{Var}\left[\varepsilon|X\right]\right)M_1'X_2\left(X_2'M_1X_2\right)^{-1} \\
&= \sigma^2\left(X_2'M_1X_2\right)^{-1}
\end{aligned}
$$

An application of FWL theorem is to compute a single coefficient.

Individual Regression Coefficients The coefficient on z in a multiple regression of y on $W = [X, z]$ is computed as $c = (z'Mz)^{-1}(z'My) = (z^{*'}z^*)^{-1}z^{*'}y^*$, where z^* and y^* are the residual vectors from least squares regressions of z and y on X; $z^* = Mz$ and $y^* = My$ where M is residual maker for X: $M = \mathbf{I} - X(X'X)^{-1}X'$.

Besides, consider the case in which X_1 is $\mathbf{1}$, namely a constant term which is a column of 1s in the first column of X. For each vectors regressed on $\mathbf{1}$, the resulted residuals are deviations from the sample mean. The interesting discovery is summarized as in the corollary.

Regression with a Constant Term The slopes in a multiple regression that contains a constant term are obtained by transforming the data to deviations from their means and then regressing the variable y in deviation form on the explanatory variables, also in deviation form.

From the two theorems in partitioned regression and the corollary of regression with a constant term, we formalize our integrated ideas into the following theorem.

Orthogonal Regression If the multiple regression of y on X contains a constant term and the variables in the regression are uncorrelated, then the multiple regression slopes are the same as the slopes in the individual simple regressions of y on a constant and each variable in turn.

The use of multiple regression involves a conceptual experiment that we might not be able to carry out in practice, the *ceteris paribus*. FWL theorem rationalizes all the thought experiment with *ceteris paribus*, even if the sample contains no such pair of observations.

1.2.3.2 Partial Correlation Coefficient From the perspective of *ceteris paribus*, to see the correlation between two variables, we should first purge out effects of other variables, and use a **partial correlation coefficient** as a measure. Follow the preceding notations, the partial correlation coefficient is defined as

$$r_{yz}^{*2} = \frac{(z_*' y_*)^2}{(z_*' z_*)(y_*' y_*)}$$

Note that there is a convenient shortcut.

Relationship of Partial Correlation Coefficient and t Statistic Once the multiple regression is computed, the t ratio for testing the hypothesis that the coefficient of z equals zero can be used to compute

$$r_{yz}^{*2} = \frac{t_z^2}{t_z^2 + \text{degrees of freedom}}$$

where the degrees of freedom is equal to $n - (K + 1)$.

Proof By definition, the squared t ratio for z's coefficient is

$$\begin{aligned} t_z^2 &= \frac{c^2}{s^2 (W'W)^{-1}_{K+1,K+1}} \\ &= \frac{c^2}{\left(\dfrac{u'u}{n-(K+1)}\right)(W'W)^{-1}_{K+1,K+1}} \end{aligned}$$

where $(W'W)_{K+1,K+1}$ is the $(K + 1)$, namely last, diagonal element of $(W'W)^{-1}$. To see it specifically, we introduce the following useful partitioned inverse formula:

Partitioned Inverse Formula For the general 2×2 partitioned matrix, one form of the partitioned inverse is

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1}(\mathbf{I} + A_{12}F_2A_{21}A_{11}^{-1}) & -A_{11}^{-1}A_{12}F_2 \\ -F_2A_{21}A_{11}^{-1} & F_2 \end{bmatrix}$$

where

$$F_2 = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$$

The upper left block could also be written as

$$F_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$$

Specifically, finding the inverse of $X'X$ is the most common in regression analysis, where $X = [X_1, X_2]$.

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} = \begin{bmatrix} (X_1'M_2X_1)^{-1} & -(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1} \\ -(X_2'M_1X_2)^{-1}X_2'X_1(X_1'M_1X_1)^{-1} & (X_2'M_1X_2)^{-1} \end{bmatrix}$$

where $M_1 = \mathbf{I} - X_1(X_1'X_1)^{-1}X_1', M_2 = \mathbf{I} - X_2(X_2'X_2)^{-1}X_2'$. The off-diagonal elements are symmetric.

By partitioned inverse formula, the last lower right element of the matrix equals $(z'Mz)^{-1} = (z_*'z_*)^{-1}$. From FWL theorem, we also have that $c = (z_*'z_*)^{-1}z_*'y_* = \frac{(z_*'y_*)}{(z_*'z_*)}$. For convenience, let $df = n - (K+1)$. Then,

$$t_z^2 = \frac{\frac{(z_*'y_*)^2}{(z_*'z_*)^2}}{\frac{u'u}{df} \cdot (z_*'z_*)^{-1}} = \frac{(z_*'y_*)^2 \cdot df}{(u'u)(z_*'z_*)}$$

$$\Longrightarrow \frac{t_z^2}{t_z^2 + df} = \frac{\frac{(z_*'y_*)^2 \cdot df}{(u'u)(z_*'z_*)}}{\frac{(z_*'y_*)^2 \cdot df}{(u'u)(z_*'z_*)} + df} = \frac{\frac{(z_*'y_*)^2}{(u'u)(z_*'z_*)}}{\frac{(z_*'y_*)^2}{(u'u)(z_*'z_*)} + 1} = \frac{(z_*'y_*)^2}{(z_*'y_*)^2 + (u'u)(z_*'z_*)}$$

Divide numerator and denominator by $(z_*'z_*)(y_*'y_*)$ to move closer to r_{yz}^{*2}:

$$\begin{aligned} \frac{t_z^2}{t_z^2 + df} &= \frac{(z_*'y_*)^2}{(z_*'y_*)^2 + (u'u)(z_*'z_*)} \\ &= \frac{\frac{(z_*'y_*)^2}{(z_*'z_*)(y_*'y_*)}}{\frac{(z_*'y_*)^2}{(z_*'z_*)(y_*'y_*)} + \frac{(u'u)(z_*'z_*)}{(z_*'z_*)(y_*'y_*)}} \\ &= \frac{r_{yz}^{*2}}{r_{yz}^{*2} + \frac{u'u}{y_*'y_*}} \end{aligned}$$

By the theorem of change in the sum of squares when a variable is added to a regression,

$$u'u = e'e - c^2(z_*'z_*) = y_*'y_* - \frac{(z_*'y_*)^2}{z_*'z_*}$$

Inserting this into our previous result to get:

$$\begin{aligned}
\frac{t_z^2}{t_z^2+df} &= \frac{r_{yz}^{*2}}{r_{yz}^{*2}+\dfrac{u'u}{y_*'y_*}} \\
&= \frac{r_{yz}^{*2}}{r_{yz}^{*2}+\dfrac{y_*'y_*-\dfrac{(z_*'y_*)^2}{z_*'z_*}}{y_*'y_*}} \\
&= \frac{r_{yz}^{*2}}{r_{yz}^{*2}+(1-r_{yz}^{*2})} \\
&= r_{yz}^{*2}
\end{aligned}$$

1.2.4 Goodness of Fit

Variation of the dependent variable is defined in terms of deviations from its mean. The *total variation* in y is the sum of squared deviations:

$$SST = \sum_{i=1}^{n}(y_i-\bar{y})^2$$

In terms of the regression analysis, we may write the full set of observations as

$$y = Xb + e = \hat{y} + e$$

Notice that here normal equation plays an important role. The first observation comes from the assumption that $E[\varepsilon|X]=0$, that is, the regressors are uncorrelated with the disturbances,

$$\begin{aligned}
& X'Xb = X'y \\
\iff & X'(y-Xb) = 0 \\
\iff & X'e = 0
\end{aligned}$$

Therefore, if the regression contains a constant term, the residuals will add up to be 0. From this, the mean of the predicted values of y_i will equal the mean of the actual values, that is,

$$\bar{\hat{y}} = \bar{y}$$

From this result we can immediately derive:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i = (x_i - \bar{x})'b + e_i$$

The second observation is that, as defined before, projector P and residual maker M are perpendicular to each other, and y can be partitioned into the two orthogonal parts:

$$y = \hat{y} + e = Py + My$$

So are the variations in each part (where M_0 is the residual maker for $\mathbf{1}$).

$$\begin{aligned} & y = Xb + e \\ \Longrightarrow & M_0 y = M_0 Xb + M_0 e \\ \Longrightarrow & y' M_0 y = b' X' M_0 Xb + e' e \\ \Longrightarrow & SST = SSR + SSE \end{aligned}$$

In the deduction we have made good use of the fact that $X'e = e'X = 0$. We can now obtain a measure of how well the regression line fits:

$$R^2 = \frac{SSR}{SST} = \frac{b' X' M_0 Xb}{y' M_0 y} = 1 - \frac{e'e}{y' M_0 y}$$

As we have shown, R^2 must be between 0 and 1, and it measures the proportion of the total variation in y that is accounted for by variation in the regressors. It equals zero if the regression is a horizontal line, that is, if all the elements of b except the constant term are zero. In this case, the predicted values of y are always $\bar{y}$, so deviations of x from its mean do not translate into different predictions for y. As such, x has no explanatory power. The other extreme, $R^2 = 1$, occurs if the values of x and y all lie in the same hyperplane (on a straight line for a two variable regression) so that the residuals are all zero. If all the values of y_i lie on a vertical line, then R^2 has no meaning and cannot be computed.

Regression analysis is often used for forecasting. In this case, we are interested in how well the regression model predicts movements in the dependent variable. With this in mind, an equivalent way to compute R^2 is also useful. First

$$\begin{cases} \hat{y} = Xb \\ y = \hat{y} + e \\ X'e = 0 \\ \bar{e} = 0 \Longrightarrow M_0 e = 0 \end{cases} \qquad \Longrightarrow \hat{y}' M_0 \hat{y} = \hat{y}' M_0 y$$

$$\begin{aligned} R^2 &= \frac{SSR}{SST} = \frac{b' X' M_0 Xb}{y' M_0 y} \\ \Longrightarrow \quad &= \frac{\hat{y}' M_0 \hat{y}}{y' M_0 y} = \frac{\hat{y}' M_0 \hat{y}}{y' M_0 y} \cdot \frac{\hat{y}' M_0 y}{\hat{y}' M_0 \hat{y}} \\ &= \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2}{\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right] \left[\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2\right]} \end{aligned}$$

which is the squared correlation between the observed values of y and the predicted values.

There are some problems with the use of R^2 in analyzing goodness of fit. The first concerns the number of degrees of freedom used up in estimating the parameters. R^2 will never decrease when another variable is added to a regression equation. First, we will see change in the sum of squares will never increase when adding a new variable into a regression.

Change in the Sum of Squares When a Variable is Added to a Regression If $e'e$ is the sum of squared residuals when y is regressed on X and $u'u$ is the sum of squared residuals when y is regressed on X and z, then

$$u'u = e'e - c^2(z_*'z_*) \leq e'e$$

where c is the coefficient on z in the long regression of y on $[X, z]$ and $z_* = Mz$ is the vector of residuals when z is regressed on X.

Proof In the long regression of y on X and z, the vector of residuals is $u = y - Xd - zc$. Note that unless $X'z = \mathbf{0}$, d will not equal $b = (X'X)^{-1}X'y$. Moreover, unless $c = \mathbf{0}$, u will not equal $e = y - Xb$. From FWL theorem, $c = (z_*'z_*)^{-1}(z_*'y_*)$. From the equation in partitioned regression, we know

$$d = (X'X)^{-1}X'(y - zc) = b - (X'X)^{-1}X'zc$$

Inserting this expression for d in that for u gives

$$u = y - Xd - zc = y - Xb - X(X'X)^{-1}X'zc = e - Mzc = e - z_*c$$

Then,

$$u'u = e'e + c^2(z_*'z_*) - 2c(z_*'e)$$

Notice that $e = My = y_*$, and $z_*'e = z_*'y_* = c(z_*'z_*)$. Inserting this result in $u'u$ immediately above gives the result in the theorem.

The result can be extended to the case when there is a set of variables added to a regression.

Change in the Sum of Squares When a Set of Variables is Added to a Regression If $e'e$ is the sum of squares when y is regressed on X_1, and $u'u$ is the sum of squared residuals when y is regressed on $[X_1, X_2]$, then

$$u'u = e'e - b_2'X_2'M_1X_2b_2$$

Proof First in the setup, we estimate the following two equations:

$$\begin{aligned} y =& X_1 b + e \\ y =& X_1 b_1 + X_2 b_2 + u \end{aligned}$$

From FWL theorem, we know that

$$b_2 = (X_2' M_1 X_2)^{-1} (X_2' M_1 y)$$

From partitioned regression formula, we can solve for b_1 from b_2:

$$\begin{aligned} b_1 &= (X_1' X_1)^{-1} X_1' (y - X_2 b_2) \\ &= (X_1' X_1)^{-1} X_1 y - (X_1' X_1)^{-1} X_1' X_2 b_2 \\ &= b - (X_1' X_1)^{-1} X_1' X_2 b_2 \end{aligned}$$

Therefore, we can link the residuals of long regression and residuals of the short regression:

$$\begin{aligned} u &= y - X_1 b_1 - X_2 b_2 \\ &= y - X_1 (b - (X_1' X_1)^{-1} X_1' X_2 b_2) - X_2 b_2 \\ &= (y - X_1 b) + X_1 (X_1' X_1)^{-1} X_1' X_2 b_2 - X_2 b_2 \\ &= e + X_1 (X_1' X_1)^{-1} X_1' X_2 b_2 - X_2 b_2 \\ &= e - M_1 X_2 b_2 \\ \Longrightarrow u'u &= e'e - e' M_1 X_2 b_2 - b_2' X_2' M_1 e + b_2' X_2' M_1 X_2 b_2 \\ &= e'e - 2 b_2' X_2' M_1 e + b_2' X_2' M_1 X_2 b_2 \end{aligned}$$

To simplify the result, we should notice

$$\begin{cases} e = M_1 y \\ b_2 = (X_2' M_1 X_2)^{-1} (X_2' M_1 y) \end{cases} \Longrightarrow b_2' X_2' M_1 e = (y' M_1 X_2')(X_2' M_1 X_2)^{-1} (X_2' M_1 y)$$

$$\begin{aligned} b_2' X_2' M_1 X_2 b_2 &= (y' M_1 X_2')(X_2' M_1 X_2)^{-1} (X_2' M_1 X_2)(X_2' M_1 X_2)^{-1} (X_2' M_1 y) \\ &= (y' M_1 X_2')(X_2' M_1 X_2)^{-1} (X_2' M_1 y) \end{aligned}$$

Therefore, we find that $b_2' X_2' M_1 e = b_2' X_2' M_1 X_2 b_2$, which can be used to simplify the result. However, since $b_2' X_2' M_1 X_2 b_2$ is more meaningful, we keep this term to get:

$$u'u = e'e - b_2' X_2' M_1 X_2 b_2$$

Based on the results of change in sum of squares, we are straightforward to see change in R^2 when a variable is added to a regression.

Change in R^2 When a Variable Is Added to a Regression Let R^2_{Xz} be the coefficient of determination in the regression of y on X and an additional variable z, let R^2_X be the same for the regression of y on X alone, and let r^{*2}_{yz} be the partial correlation between y and z, controlling for X. Then

$$R^2_{Xz} = R^2_X + (1 - R^2_X) r^{*2}_{yz}$$

Proof Following the same notations, we have shown

$$u'u = e'e - c^2(z'_* z_*)$$

By definition, $r^{*2}_{yz} = \frac{(z'_* y_*)^2}{(z'_* z_*)(y'_* y_*)}$; so we have

$$u'u = e'e - c^2(z'_* z_*) = e'e(1 - r^{*2}_{yz})$$

Now divide through both sides of the equality by $y'M_0 y$. Notice that $1 - R^2_X = \dfrac{u'u}{y'M_0 y}$ and $1 - R^2_{Xz} = \dfrac{e'e}{y'M_0 y}$. Rearrange the result produces the equation illustrated above.

Thus, the R^2 in the longer regression cannot be smaller. It is tempting to exploit this result by just adding variables to the model; R^2 will continue to rise to its limit of 1. Therefore, we introduce the **adjusted** R^2 (for degrees of freedom), which incorporates a penalty for adding superfluous variables:

$$\bar{R}^2 = 1 - \left(\frac{e'e}{n-K}\right) \Big/ \left(\frac{y'M_0 y}{n-1}\right)$$

For computational purposes, the connection between R^2 and $\bar{R}^2$ is

$$\bar{R}^2 = 1 - \frac{n-1}{n-K} \cdot (1 - R^2)$$

The adjusted R^2 may decline when a variable is added to the set of independent variables. Indeed, $\bar{R}^2$ may even be negative. To consider an admittedly extreme case, suppose that x and y have a sample correlation of zero. Then the adjusted R^2 will equal $-\dfrac{1}{n-2}$. (Thus, the name "adjusted R-squared" is a bit misleading—$\bar{R}^2$ is not actually computed as the square of any quantity.) Whether $\bar{R}^2$ rises or falls depends on whether the contribution of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom. The general result (the proof of which is left as an exercise) is as follows.

Change in $\bar{R}^2$ When a Variable is Added to a Regression In a multiple regression, $\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if the square to the t ratio associated with this variable is greater (less) than 1.

Proof Follow the preceding notations. Suppose at first we have the regression on X, and then we consider adding a new variable z (or equivalently think of the reverse procedure). We construct the difference of $\bar{R}^2$ from the definition of R^2:

$$\begin{cases} R^2_{Xz} = 1 - \frac{u'u}{y'M_0y} \\ R^2_X = 1 - \frac{e'e}{y'M_0y} \end{cases}, \quad \begin{cases} \bar{R}^2_{Xz} = 1 - \frac{n-1}{n-(K+1)}(1 - R^2_{Xz}) \\ \bar{R}^2_X = 1 - \frac{n-1}{n-K}(1 - R^2_X) \end{cases}$$

Since the t ratio is rather complicated, we construct t by proof:

$$\begin{aligned} & \bar{R}^2_{Xz} - \bar{R}^2_X = \frac{n-1}{n-K}(1 - R^2_X) - \frac{n-1}{n-(K+1)}(1 - R^2_{Xz}) \geq 0 \\ \Longleftarrow & \frac{n-1}{n-K} \cdot \frac{e'e}{y'M_0y} \geq \frac{n-1}{n-(K+1)} \cdot \frac{u'u}{y'M_0y} \\ \Longleftarrow & (n-K-1) \cdot e'e \geq (n-K) \cdot u'u \\ \Longleftarrow & (n-K-1) \cdot (u'u + c^2(z'_*z_*)) \geq (n-K) \cdot u'u \\ \Longleftarrow & (n-K-1) \cdot c^2(z'_*z_*) \geq u'u \\ \Longleftarrow & \frac{c^2}{\frac{u'u}{n-K-1} \cdot (z'_*z_*)^{-1}} \geq 1 \\ \Longleftarrow & \left(\frac{c}{\sqrt{\frac{u'u}{n-K-1} \cdot (z'_*z_*)^{-1}}} \right)^2 \geq 1 \end{aligned}$$

In fact, t ratio for z is

$$t_z = \frac{c}{\sqrt{\frac{u'u}{n-K-1} \cdot (z'_*z_*)^{-1}}}$$

Therefore, we conclude that the adjusted R^2 would increase after adding a new variable z if and only if $t_z^2 \geq 1$.

A second difficulty with R^2 concerns the constant term in the model. The proof tat $0 \leq R^2 \leq 1$ requires X to contain a column of 1s. Or more generally speaking, the orthogonal decomposition of variance of y stands upon the fact that a constant term should be included in the model. If NOT, then $M_0e \neq e$ and $e'M_0X \neq \mathbf{0}$, and the term $2e'M_0Xb$ in $y'M_0y = (M_0Xb + M_0e)'(M_0Xb + M_0e)$ in the expansion would not drop

out. Consequently, when we compute R^2 as

$$R^2 = 1 - \frac{e'e}{y'M_0y}$$

The result would be unpredictable, that is, R^2 is no longer meaningful. It may even be negative. However, even if you consider an alternative computation,

$$R^2 = \frac{b'X'M_0y}{y'M_0y}$$

This is equally problematic. Again, this calculation will differ from the one obtained with the constant term included; this time, R^2 may be larger than 1. Some computer packages bypass these difficulties by reporting a third "R^2", the squared sample correlation between the actual values of y and the fitted values from the regression. However, even though this time the determinant coefficient falls within 0 and 1, its meaning might be deceptive; it is no longer the proportion of variance explained by the model.

2 Properties of OLS Estimators

2.1 Finite Sample Properties of Least Squares Estimators

An "estimator" is a strategy, or formula for using the sample data that are drawn from a population. The "properties" of that estimator are a description of how that estimator can be expected to behave when it is applied to a sample of data. To consider an example, the concept of unbiasedness implies that "on average" an estimator (strategy) will correctly estimate the parameter in question; it will not be systematically too high or too low. It seems less than obvious how one could know this if they were only going to draw a single sample of data from the population and analyze that one sample. The argument adopted in classical econometrics is provided by the sampling properties of the estimation strategy. A conceptual experiment lies behind the description. One imagines "repeated sampling" from the population and characterizes the behavior of the "sample of samples". The underlying statistical theory of the estimator provides the basis of the description.

2.1.1 Unbiasedness

The least squares estimator is unbiased in every sample.

$$\begin{aligned} b &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta+\varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \\ \Longrightarrow &\mathrm{E}\left[b|X\right] = \beta + \mathrm{E}\left[(X'X)^{-1}X'\varepsilon|X\right] \end{aligned}$$

By assumption of exogeneity, the second term is $\mathbf{0}$, so

$$\mathrm{E}\left[b|X\right] = \beta$$

Therefore,

$$\mathrm{E}\left[b\right] = \mathrm{E}_X\left[\mathrm{E}\left[b|X\right]\right] = \mathrm{E}_X\left[\beta\right] = \beta$$

The interpretation of this result is that, for any particular set of observations X, the least squares estimator has expectation β. Therefore, when we average this over the possible values of X, we find the unconditional mean is β as well.

2.1.1.1 Bias From Variable Omission From what we have discussed above, we can also see bias caused by omission of relevant variables. The most common one of specification errors are the omission of relevant variables and the inclusion of superfluous (irrelevant) variables. Suppose that a correctly specified regression model would be

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where the two parts of X have K_1 and K_2 columns, respectively. If we regress y on X_1, without including X_2, then the estimator is

$$b_1 = (X_1'X_1)^{-1}X_1'y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$$

Taking the expectation, we see that unless $X_1'X_2 = \mathbf{0}$ or $\beta_2 = \mathbf{0}$, b_1 is biased. The well-known result is the **omitted variable formula**:

$$\mathrm{E}\left[b_1|X\right] = \beta_1 + P_{1,2}\beta_2, \text{ where } P_{1,2} = (X_1'X_1)^{-1}X_1'X_2$$

Note that each column of the $K_1 \times K_2$ matrix $P_{1,2}$ is the column of slopes in the least squares regression of the corresponding column of X_2 on the columns of X_1.

Alternatively, we can view the omission of a set of variables as equivalent to imposing an incorrect restriction on the correct regression model. In particular, omitting X_2 is equivalent to *incorrectly* estimating the correct model subject to the restriction $\beta_2 = \mathbf{0}$. Incorrectly imposing a restriction produces a biased estimator. Another way to view this error is to note that it amounts to incorporating incorrect information in our estimation. Suppose, however, that our error is simply a failure to use some information that is correct. In this view, we can assert without formal proof that inclusion of irrelevant variables in the regression will not affect unbiasedness, but may face issues of overfitting and a larger covariance matrix for OLS estimators then.

2.1.1.2 Multicollinearity As a response to what appears to be a "multicollinearity problem," it is often difficult to resist the temptation to drop what appears to be an offending variable from the regression, if it seems to be the one causing the problem. This "strategy" creates a subtle dilemma for the analyst. Consider the partitioned multiple regression

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

If we regress y only on X_1, the estimator is biased:

$$\begin{aligned} b_1 &= (X_1'X_1)^{-1}X_1'y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon \\ \Longrightarrow &\begin{cases} \mathrm{E}\left[b_1|X\right] = \beta_1 + P_{1,2}\beta_2 \\ b_1 - \mathrm{E}\left[b_1|X\right] = (X_1'X_1)^{-1}X_1'\varepsilon \end{cases} \end{aligned}$$

The covariance matrix of this estimator is

$$\mathrm{Var}\left[b_1|X\right] = \sigma^2(X_1'X_1)^{-1}$$

Note that the covariance matrix is around the $\mathrm{E}\left[b_1|X\right]$ instead of β_1 (Coincidentally in the unbiased estimator case, $\mathrm{E}\left[b_k\right] = b_k$). If β_2 is not actually zero, then in the multiple regression of y on (X_1, X_2), the variance of $b_{1.2}$ around its mean, β_1, would be

$$\begin{aligned} \mathrm{Var}\left[b_{1.2}|X\right] &= \sigma^2\left(X_1'M_2X_1\right)^{-1} \\ &= \sigma^2\left[X_1'X_1 - X_1X_2(X_2{}'X_2)^{-1}X_2{}'X_1\right]^{-1} \end{aligned}$$

We compare the two covariance matrices. It is simpler to compare the inverses. Thus,

$$\{\text{Var}\left[b_1|X\right]\}^{-1} - \{\text{Var}\left[b_{1.2}|X\right]\}^{-1} = \frac{1}{\sigma^2} X_1' X_2 (X_2{}' X_2)^{-1} X_2{}' X_1$$

The difference matrix is a non-negative definite matrix. The implication is that, the variance of b_1 is not larger than the variance of $b_{1.2}$. It follows that although b_1 is biased, its variance is never larger than the variance of the unbiased estimator. In any realistic case, $X_1' X_2 \neq \mathbf{0}$, in fact it will be smaller.

2.1.2 Efficiency

If the regressors can be treated as non-stochastic, as they would be in an experimental situation in which the analyst chooses the values in X, then the sampling variance of the least squares estimator can be derived by treating X as a matrix of constants. Alternatively, we can allow X to be stochastic, do the analysis conditionally on the observed X, then consider averaging over X as we did in analysis of unbiasedness.

$$b = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

From the result we can see, b is a linear function of the disturbances (which by definition we will see, makes it a linear estimator). As we have seen, the expected value of the second term is $\mathbf{0}$. Therefore, regardless of the distribution of ε, under our other assumptions, b is a linear, unbiased estimator of β. By assumption 4, $\text{Var}\left[\varepsilon|X\right] = \sigma^2\mathbf{I}$. Thus, conditional covariance matrix of the least squares slope estimator is

$$\begin{aligned}
\text{Var}\left[b|X\right] &= \text{E}\left[(b-\beta)(b-\beta)'|X\right] \\
&= \text{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X\right] \\
&= (X'X)^{-1}X'\text{E}\left[\varepsilon\varepsilon'|X\right]X(X'X)^{-1} \\
&= (X'X)^{-1}X'\left(\sigma^2\mathbf{I}\right)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

Specifically, suppose that X contains only a constant term and a single regressor $\mathbf{x}$, the lower-right element of $\sigma^2(X'X)^{-1}$ is

$$Var[b|\mathbf{x}] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Like what we have done in unbiasedness part, if we seek the unconditional covariance matrix:

$$\begin{aligned}
&\text{Var}\left[b|X\right] = \sigma^2(X'X)^{-1} \\
\Longrightarrow &\text{Var}\left[b\right] = \text{E}_{\mathbf{X}}\left[\text{Var}\left[b|X\right]\right] + \text{Var}_{\mathbf{X}}\left[\text{E}\left[b|X\right]\right] \\
\Longrightarrow &\text{Var}\left[b\right] = \text{E}_{\mathbf{X}}\left[\sigma^2(X'X)^{-1}\right] + \text{Var}_{\mathbf{X}}\left[\beta\right] \\
\Longrightarrow &\text{Var}\left[b\right] = \text{E}\left[\sigma^2(X'X)^{-1}\right] = \sigma^2\text{E}\left[(X'X)^{-1}\right]
\end{aligned}$$

Our earlier conclusion is altered slightly. We must replace $(X'X)^{-1}$ with its expected value to get the appropriate covariance matrix, which brings a subtle change in the interpretation of these results. The unconditional variance of b can only be described in terms of the average behavior of X, so to proceed further, it would be necessary to make some assumptions about the variances and covariances of the regressors. We will return to this subject later.

2.1.2.1 Gauss-Markov Theorem After obtaining the result of covariance matrix, we now obtain a general result for the class of linear unbiased estimator of β regarding its efficiency.

Gauss-Markov Theorem In the linear regression model with regressor matrix X, the least squares estimator b is the minimum variance linear unbiased estimator of β. For any vector of constants w, the minimum variance linear unbiased estimator of $w'\beta$ in the regression model is $w'b$, where b is the least squares estimator.

Proof A direct approach to proving this important theorem would be to *define the class of linear and unbiased estimators* ($b_L = Cy$ such that $\mathrm{E}[b_L|X] = \beta$) and then find the member of that class which has the smallest variance.

Let $b_0 = Cy$ be another linear unbiased estimator of β, where C is a $K \times n$ matrix. If b_0 is unbiased, then

$$\begin{aligned}\mathrm{E}[b_0|X] &= \mathrm{E}[Cy|X] = \mathrm{E}[C(X\beta+\varepsilon)|X] = \mathrm{E}[CX\beta + C\varepsilon|X] \\ &= \mathrm{E}[CX\beta|X] = \beta\end{aligned}$$

which implies that $CX = \mathbf{I}$. There are many candidates. Now construct a "difference matrix" D, $D = C - (X'X)^{-1}X'$. (This idea is somewhat based on the conjecture that our OLS estimator $b = (X'X)^{-1}X'y$ is the best one so we are naturally interested in the difference.) So, $Dy = b_0 - b$. Then,

$$\begin{aligned}\mathrm{Var}[b_0|X] &= \mathrm{Var}[Dy + b|X] \\ &= \mathrm{Var}[D(X\beta+\varepsilon) + b|X] \\ &= \mathrm{Var}[DX\beta + D\varepsilon + b|X] \\ &= \mathrm{Var}\left[DX\beta + D\varepsilon + \beta + (X'X)^{-1}X'\varepsilon|X\right] \\ &= \mathrm{Var}\left[(DX+I)\beta + (D+(X'X)^{-1}X')\varepsilon|X\right] \\ &= \mathrm{Var}\left[(D+(X'X)^{-1}X')\varepsilon|X\right] \\ &= (D+(X'X)^{-1}X')\,\mathrm{Var}[\varepsilon|X]\,(D+(X'X)^{-1}X')' \\ &= \sigma^2(D+(X'X)^{-1}X')(D+(X'X)^{-1}X')' \\ &= \sigma^2(X'X)^{-1} + \sigma^2 DD'\end{aligned}$$

The last equality holds because since we know that $CX = \mathbf{I} = DX + (X'X)^{-1}(X'X)$, so DX must equal $\mathbf{0}$. This way is a bit tedious, another way to see the result is through

the direct covariance matrix of b_0:

$$\begin{aligned}\text{Var}\left[b_0|X\right] &= \text{Var}\left[Cy|X\right] = \text{Var}\left[CX\beta + C\varepsilon|X\right] = \text{Var}\left[C\varepsilon|X\right] \\ &= C\text{Var}\left[\varepsilon|X\right]C' = \sigma^2 CC' \\ &= \sigma^2\left(D + (X'X)^{-1}X'\right)\left(D + (X'X)^{-1}X'\right)' \\ &= \sigma^2(X'X)^{-1} + \sigma^2 DD'\end{aligned}$$

Since a quadratic form in DD' is $q'DD'q = (D'q)'(D'q) = ||D'q||^2 \geq 0$, the conditional covariance matrix of b_0 equals that of b plus a non-negative definite matrix. Therefore, every quadratic form in $Var[b_0|X]$ is larger than the corresponding quadratic from in $Var[b|X]$, which establishes the first result.

The proof of the second statement follows from the previous derivation, since the varaince of $w'b$ is a quadratic form in $Var[b|X]$, and likewise for any b_0 and proves that each individual slope estimator b_k is the best linear unbiased estimator of β_k.

Remarks: Gauss-Markov theorem makes no use of assumption of normality of the distribution of the disturbances. Only the first four assumptions are necessary.

We have shown in preceding Gauss-Markov theorem that $\text{Var}\left[b|X\right] \leq \text{Var}\left[b_0|X\right]$ for any linear and unbiased $b_0 \neq b$ and for the specific X in our sample. But if this inequality holds for every particular X, together with unconditional covariance that $\text{Var}\left[b\right] = \sigma^2 \text{E}_{\mathbf{X}}\left[(X'X)^{-1}\right]$, then it must hold for $\text{Var}\left[b\right] = \text{E}_{\mathbf{X}}\left[\text{Var}\left[b|X\right]\right]$. That is, if it holds for every particular X, then it must hold over the average value(s) of X.

The conclusion, therefore, is that the important results we have obtained thus far for the least squares estimator, unbiasedness, and the Gauss–Markov theorem hold whether or not we condition on the particular sample in hand or consider, instead, sampling broadly from the population.

Gauss–Markov Theorem (Concluded) In the linear regression model, the least squares estimator b is the minimum variance linear unbiased estimator of β whether X is stochastic or nonstochastic, so long as the other assumptions of the model continue to hold.

From here on, we will be encountering many times the comparison between covariance matrices. So it shall be better to know something about matrices comparison.

Matrices Comparison

Derivations in econometrics often focus on whether one matrix is "larger" than another.We now consider how to make such a comparison. As a starting point, the two matrices must have the same dimensions. A useful comparison is based on

$$d = x'Ax - x'Bx = x'(A - B)x$$

If d is always positive for any nonzero vector x, then by this criterion, we can say that A is larger than B. The reverse would apply if d is always negative.

It follows from the definition that, if $d > 0$ for all nonzero $\mathbf{x}$, then $A - B$ is positive definite.

If d is only greater than or equal to zero, then $A - B$ is nonnegative definite. Notice that the ordering is not complete. For some pairs of matrices, d could have either sign, depending on $\mathbf{x}$. In this case, there is no simple comparison.

A particular case of the general result which we will encounter frequently is that, if A is positive definite and B is non-negative definite, then $A + B \geq A$. This is often used in variance matrix comparison and see which estimator is more efficient.

Finally, in comparing matrices, it may be more convenient to compare their inverses. The result analogous to a familiar result for scalars is:

$$A > B \Longrightarrow B^{-1} > A^{-1}$$

2.1.2.2 Variance Estimation If we wish to test hypotheses about β or to form confidence intervals, then we will require a sample estimate of the covariance matrix, $\text{Var}[b|X] = \sigma^2(X'X)^{-1}$. The population parameter σ^2 remains to be estimated. Since σ^2 is the expected value of ε_i^2 and e_i is an estimate of ε_i, by analogy,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

would seem to be a natural estimator. But the least squares residuals are imperfect estimates of their population counterparts; $e_i = y_i - x_i'b = \varepsilon_i - x_i'(b - \beta)$. **The estimator is distorted (as might be expected) because β is not observed directly**. The expected square on the right-hand side involves a second term that might not have expected value zero.
The least squares residuals are

$$e = My = M(X\beta + \varepsilon) = M\varepsilon$$

As $MX = \mathbf{0}$. An estimator of σ^2 will be based on the sum of squared residuals:

$$e'e = \varepsilon' M\varepsilon$$

The expected value of this quadratic form is $\text{E}[e'e|X] = \text{E}[\varepsilon' M\varepsilon|X]$. Notice that the scalar $\varepsilon' M\varepsilon$ is a 1×1 matrix, namely a number, so it is equal to its trace. By using the result on cyclic permutations,

$$\begin{aligned}
\text{E}[\text{tr}(\varepsilon' M\varepsilon)|X] &= \text{E}[\text{tr}(M\varepsilon\varepsilon')|X] \\
&= \text{tr}(M\text{E}[\varepsilon\varepsilon'|X]) \\
&= \text{tr}(M(\sigma^2\mathbf{I})) \\
&= \sigma^2\text{tr}(M)
\end{aligned}$$

The trace of M is

$$\begin{aligned}\operatorname{tr}(M) &= \operatorname{tr}\left(\mathbf{I}_n - X(X'X)^{-1}X'\right) \\ &= \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}\left(X(X'X)^{-1}X'\right) \\ &= \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}\left[(X'X)^{-1}(X'X)\right] \\ &= \operatorname{tr}(()\,\mathbf{I}_n) - \operatorname{tr}(\mathbf{I}_K) \\ &= n - K\end{aligned}$$

Therefore,

$$\mathrm{E}\left[e'e|X\right] = (n-K)\sigma^2$$

An unbiased estimator of σ^2 is:

$$s^2 = \frac{e'e}{n-K}$$

So the natural estimator, $\frac{1}{n}\sum_{i=1}^{n} e_i^2$ is biased toward zero, although the bias becomes smaller as the sample size increases. Notice that the estimator s^2 is unbiased unconditionally as well, since $\mathrm{E}\left[s^2\right] = \mathrm{E}_{\mathbf{X}}\left[\mathrm{E}\left[s^2|X\right]\right] = \mathrm{E}_{\mathbf{X}}\left[\sigma^2\right] = \sigma^2$. The standard error of the regression is then s, the square root of s^2. With s^2, we can then compute

$$\text{Est.Var}\left[b|X\right] = s^2(X'X)^{-1}$$

Henceforth, we shall use the notation Est.Var$[\cdot]$ to indicate a sample estimate of the sampling variance of an estimator. The square root of the k-th diagonal element of this matrix, $\left\{s^2(X'X)^{-1}\right\}^{\frac{1}{2}}$, is the standard error of the estimator b_k, which is often denoted simply "the standard error of b_k".

2.1.3 Normality

To this point, our specification and analysis of the regression model are semiparametric. We have not used Assumption A6, normality of ε, in any of our results. The assumption is useful for constructing statistics for forming confidence intervals. As we have shown, b is a linear function of the disturbance vector ε. If we assume that ε has a multivariate normal distribution, then

$$\begin{aligned} b &= \beta + (X'X)^{-1}X'\varepsilon \\ \Longrightarrow b|X &\sim N\left[\beta, \sigma^2(X'X)^{-1}\right]\end{aligned}$$

This specifies a multivariate normal distribution, so each element of $b|X$ is normally distributed:

$$b_k|X \sim N\left[\beta_k, \sigma^2(X'X)^{-1}_{kk}\right]$$

The distribution of b is conditioned on X. The normal distribution of b in a finite sample is a consequence of our specific assumption of normally distributed disturbances.

Without this assumption, and without some alternative specific assumption about the distribution of ε, we will not be able to make any definite statement about the exact distribution of b, conditional or otherwise. Later, however, we will be able to obtain an approximate normal distribution for b, with or without assuming normally distributed disturbances and whether the regressors are stochastic or not.

2.2 Large Sample Properties of Least Squares Estimators

Unbiasedness is a useful starting point for assessing the virtues of an estimator. It assures the analyst that their estimator will not persistently miss its target, either systematically too high or too low. However, as a guide to estimation strategy, it has two shortcomings. First, save for the least squares slope estimator we are discussing in this chapter, it is relatively rare for an econometric estimator to be unbiased. In nearly all cases beyond the multiple regression model, the best one can hope for is that the estimator improves in the sense suggested by unbiasedness as more information (data) is brought to bear on the study. As such, we will need a broader set of tools to guide the econometric inquiry. Second, the property of unbiasedness does not, in fact, imply that more information is better than less in terms of estimation of parameters. The sample means of random samples of 100 and $10,000$ are all unbiased estimators of a population mean—by this criterion all are equally desirable. Logically, one would hope that a larger sample is better than a smaller one in some sense that we are about to define (and, by extension, an extremely large sample should be much better, or even perfect). The property of consistency improves on unbiasedness in both of these directions.

2.2.1 Consistency

To begin, we leave the data generating mechanism for X unspecified—X may be any mixture of constants and random variables generated independently of the process that generates ε. We do make two crucial assumptions. The first is a modification of Assumption A5:

$(\mathbf{x}_i, \varepsilon_i), i = 1, \cdots, n$ is a sequence of *independent* observations.

The second concerns the behavior of the data in large samples:

$$\text{plim } \frac{X'X}{n} = Q, \text{ a positive definite matrix.}$$

The least squares estimator may be written as:

$$\begin{aligned} b &= \beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right) \end{aligned}$$

If Q^{-1} exists, then

$$\text{plim } b = \beta + Q^{-1}\text{plim } \left(\frac{X'\varepsilon}{n}\right)$$

This stands because the inverse is a continuous function of the original matrix. We require the probability limit of the last term. Let

$$\frac{X'\varepsilon}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \bar{\mathbf{w}}$$

Then,

$$\text{plim } b = \beta + Q^{-1}\text{plim } \bar{\mathbf{w}}$$

From the exogeneity assumption, we have

$$\begin{aligned}
\text{E}\left[\mathbf{w}_i\right] &= \text{E}_{\mathbf{x}}\left[\text{E}\left[\mathbf{w}_i|\mathbf{x}_i\right]\right] \\
&= \text{E}_{\mathbf{x}}\left[\mathbf{x}_i\text{E}\left[\varepsilon_i|\mathbf{x}_i\right]\right] \\
&= \mathbf{0} \\
\Longrightarrow \text{E}\left[\bar{\mathbf{w}}\right] &= \mathbf{0}
\end{aligned}$$

For any element in $\mathbf{x}_i$ that is non-stochastic, the zero expectations follow from the marginal distribution of ε_i. We now consider the variance:

$$\begin{aligned}
\text{Var}\left[\bar{\mathbf{w}}\right] &= \text{E}\left[\text{Var}\left[\bar{\mathbf{w}}|X\right]\right] + \text{Var}\left[\text{E}\left[\bar{\mathbf{w}}|X\right]\right] \\
&= \text{E}\left[\text{Var}\left[\bar{\mathbf{w}}|X\right]\right] + \text{Var}\left[\mathbf{0}\right] \\
&= \text{E}\left[\text{Var}\left[\bar{\mathbf{w}}|X\right]\right]
\end{aligned}$$

To obtain the first term, $\text{E}\left[\text{Var}\left[\bar{\mathbf{w}}|X\right]\right]$, w use $\text{E}\left[\varepsilon\varepsilon'|X\right] = \sigma^2\mathbf{I}$, so

$$\begin{aligned}
\text{Var}\left[\bar{\mathbf{w}}|X\right] &= \text{E}\left[\bar{\mathbf{w}}\bar{\mathbf{w}}'|X\right] \\
&= \text{E}\left[\left(\frac{X'\varepsilon}{n}\right)\left(\frac{X'\varepsilon}{n}\right)'\middle|X\right] \\
&= \frac{1}{n}X'\text{E}\left[\varepsilon\varepsilon'|X\right]X\frac{1}{n} \\
&= \left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right)
\end{aligned}$$

Therefore,

$$\text{Var}\left[\bar{\mathbf{w}}\right] = \left(\frac{\sigma^2}{n}\right)\text{E}\left[\frac{X'X}{n}\right]$$

The variance will collapse to zero if the expectation in parentheses is (or converges to) a constant matrix, so that the leading scalar will dominate the product as n increases. It then follows that

$$\lim_{n\to\infty} \text{Var}\left[\bar{\mathbf{w}}\right] = 0 \cdot \frac{X'X}{n} = \mathbf{0}$$

Since the mean of $\bar{\mathbf{w}}$ is identically zero and its variance converges to zero, $\bar{\mathbf{w}}$ converges in mean square to zero, so we establish:

$$\text{plim } \bar{\mathbf{w}} = \mathbf{0} \iff \text{plim } \frac{X'\varepsilon}{n} = \mathbf{0}$$

So

$$\text{plim } b = \beta + Q^{-1} \cdot \mathbf{0} = \beta$$

The result establishes that, under the first 4 assumptions and the additional assumption of plim $\frac{X'X}{n} = Q$, b is a consistent estimator of β in the linear regression model.

Time-series settings that involve time trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about X that is broad enough to include most of these is the **Grenander conditions**. The conditions ensure that the data matrix is "well behaved" in large samples. The assumptions are very weak and likely to be satisfied by almost any data set encountered in practice.

2.2.2 Asymptotic Normality

As a guide to estimation, consistency is an improvement over unbiasedness. Since we are in the process of relaxing the more restrictive assumptions of the model, including normality of the disturbances, we will also lose the normal distribution of the estimator that will enable us to form confidence intervals (covered later). It seems that the more general model we have built here has come at a cost. In this section, we will find that normality of the disturbances is not necessary for establishing the distributional results we need to allow statistical inference including confidence intervals and testing hypotheses. Under generally reasonable assumptions about the process that generates the sample data, large sample distributions will provide a reliable foundation for statistical inference in the regression model.

To derive the asymptotic distribution of the least squares estimator, we need to make use of some basic central limit theorems. So in addition to assumption of exogeneity, we will assume that observations are *independent*. First rewrite the b as

$$b = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon$$

$$\iff \sqrt{n}\,(b-\beta) = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon$$

Since the inverse matrix is a continuous function of the original matrix, plim $\left(\frac{X'X}{n}\right)^{-1} = Q^{-1}$. Therefore, if the limiting distribution of the random vector $\left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon$ exists, then that limiting distribution is the same as that of

$$\left[\text{plim } \left(\frac{X'X}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon = Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon$$

Thus, we must establish the limiting distribution of

$$\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon = \sqrt{n}\left(\bar{\mathbf{w}} - \mathrm{E}\left[\bar{\mathbf{w}}\right]\right)$$

where $\mathrm{E}\left[\bar{\mathbf{w}}\right] = \mathbf{0}$. We can use the multivariate Lindeberg-Feller version of the central limit theorem to obtain the limiting distribution of $\sqrt{n}\bar{\mathbf{w}}$. By definition of $\bar{\mathbf{w}}$, $\bar{\mathbf{w}}$ is the average of n independent random vectors $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$, with means $\mathbf{0}$ and variances

$$\mathrm{Var}\left[\mathbf{w}_i\right] = \mathrm{Var}\left[\mathbf{x}_i\varepsilon_i\right] = \sigma^2\mathrm{E}\left[\mathbf{x}_i\mathbf{x}_i'\right] = \sigma^2 Q_i$$

The variance of $\sqrt{n}\bar{\mathbf{w}}$ is

$$\sigma^2\bar{Q}_n = \sigma^2\left(\frac{1}{n}\right)\left[Q_1 + Q_2 + \cdots + Q_n\right]$$

As long as the sum is not dominated by any particular term and the regressors are well behaved, which in this case means plim $\frac{X'X}{n} = Q$ holds,

$$\lim_{n\to\infty}\sigma^2\bar{Q}_n = \sigma^2 Q$$

Therefore, we may apply the Lindeberg–Feller central limit theorem to the vector $\sqrt{n}\bar{\mathbf{w}}$. We now have the elements we need for a formal result. If $\left[\mathbf{x}_i\varepsilon_i\right], i = 1, 2, \cdots, n$ are independent vectors distributed with mean $\mathbf{0}$ and variance $\sigma^2 Q_i$, and if plim $\frac{X'X}{n} = Q$ holds, then

$$\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \xrightarrow{\mathrm{d}} N\left[\mathbf{0}, \sigma^2 Q\right]$$

It then follows that

$$\begin{aligned}
& Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \xrightarrow{\mathrm{d}} N\left[Q^{-1}\mathbf{0}, Q^{-1}\left(\sigma^2 Q\right)Q^{-1}\right] \\
\iff & Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \xrightarrow{\mathrm{d}} N\left[\mathbf{0}, \sigma^2 Q^{-1}\right] \\
\iff & \sqrt{n}\left(b - \beta\right) \xrightarrow{\mathrm{d}} N\left[\mathbf{0}, \sigma^2 Q^{-1}\right]
\end{aligned}$$

Thus, we obtain the asymptotic distribution of b.

Asymptotic Distribution of b with Independent Observations If $\{\varepsilon_i\}$ are independently distributed with mean zero and finite variance σ^2, and x_{ik} is such that the Grenander conditions are met, then

$$b \overset{a}{\sim} N\left[\beta, \frac{\sigma^2}{n} Q^{-1}\right]$$

In practice, it is necessary to estimate $\frac{1}{n}Q^{-1}$ (recall that plim $\frac{X'X}{n} = Q$) with $(X'X)^{-1}$ and σ^2 with $\frac{e'e}{n-K}$.

Remarks: If ε is normally distributed, then normality of $b|X$ holds in *every* sample, so it holds asymptotically as well. The important implication of this derivation is that if the regressors are well behaved and observations are independent, then the asymptotic normality of the least squares estimator does not depend on normality of the disturbances; it is a consequence of the central limit theorem.

To complete the derivation of the asymptotic properties of b, we require an estimator of Asy.Var $[b] = \frac{\sigma^2}{n}Q^{-1}$. So we still need to assess the consistency of s^2 as an estimator of σ^2.

$$\begin{aligned} s^2 &= \frac{e'e}{n-K} = \frac{\varepsilon' M \varepsilon}{n-K} \\ &= \frac{1}{n-K} \cdot [\varepsilon'\varepsilon - \varepsilon X (X'X)^{-1} X'\varepsilon] \\ &= \frac{n}{n-K} \cdot \left[\frac{\varepsilon'\varepsilon}{n} - \left(\frac{\varepsilon' X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right)\right] \end{aligned}$$

The leading constant clearly converges to 1. As we have shown, plim $\left(\frac{X'X}{n}\right)^{-1} = Q^{-1}$, plim $\left(\frac{X'\varepsilon}{n}\right) = 0$. Using the product rule for probability limits, we assert that the second term in the brackets converges to 0. That leaves

$$\bar{\varepsilon^2} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$$

This is a narrow case in which the random variables ε_i^2 are independent with the same finite mean σ^2, so not much is required to get the mean to converge almost surely to $\sigma^2 = \mathrm{E}\left[\varepsilon_i^2\right]$. Only weak conditions about ε_i^2 are needed (skipped for discussions). This gives our result:

$$\text{plim } s^2 = \sigma^2$$

So the appropriate estimator of the asymptotic covariance matrix of b is

$$\text{Est.Asy.Var}\left[b\right] = s^2(X'X)^{-1}$$

2.3 Interval Estimation

2.3.1 Confidence Interval Construction

The objective of interval estimation is to present the best estimate of a parameter with an explicit expression of the uncertainty attached to that estimate. A general approach, for estimation of a parameter θ, would be

$$\hat{\theta} \pm \text{ sampling variability}$$

(We are assuming that the interval of interest would be symmetric around $\hat{\theta}$.) Following the logic that the range of the sampling variability should convey the degree of (un)certainty, we consider the logical extremes. We can be absolutely (100 percent) certain that the true value of the parameter we are estimating lies in the range $\hat{\theta} \pm \infty$. Of course, this is not particularly informative. At the other extreme, we should place no certainty (0 percent) on the range $\hat{\theta} \pm 0$. The probability that our estimate precisely hits the true parameter value should be considered zero. The point is to choose a value of α – 0.05 or 0.01 is conventional—such that we can attach the desired confidence (prob- ability), $100(1-\alpha)$ percent to the interval. We consider how to find that range and then apply the procedure to three familiar problems, *interval estimation for one of the regression parameters*, *estimating a function of the parameters* and *predicting the value of the dependent variable in the regression using a specific setting of the independent variables*. For this purpose, we depart from Assumption A6 that the disturbances are normally distributed. We will then relax that assumption and rely instead on the asymptotic normality of the estimator.

Under assumption 6 that the disturbances are normally and independently distributed,

$$b_k \sim N\left[\beta_k, \sigma^2 S^{kk}\right]$$
$$\Longrightarrow z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \sim N\left[0, 1\right]$$

Note that z_k, which is a function of b_k, β_k, σ^2 and S^{kk}, nonetheless has a distribution that involves none of the model parameters or the data; z_k is a pivotal statistic.

Using our conventional α (say $\alpha = 95\%$) percent confidence level, we know from distribution of z_k that

$$\Pr\left[-z_{\alpha/2} \leq z_k \leq z_{\alpha/2}\right] = \alpha$$

where $\alpha = 0.05$. By a simple manipulation, we find that

$$\Pr\left[b_k - z_{1-\alpha/2} \cdot \sqrt{\sigma^2 S^{kk}} \leq z_k \leq b_k + z_{1-\alpha/2} \cdot \sqrt{\sigma^2 S^{kk}}\right] = 1 - \alpha$$

Remarks: This is a statement about the probability that the random interval contains β_k, not the probability that β_k lies in the specified interval.

We would have our desired confidence interval as above, save for the complication that σ^2 is unknown, so the interval is not operational. It would seem natural to use s^2 from the regression. This is, indeed, an appropriate approach, but later we will see it follows a different distribution, t distribution with $(n - K)$ degrees of freedom.

To start with our adjustments of σ^2 by s^2, the quantity

$$\frac{(n-K)\,s^2}{\sigma^2} = \frac{e'e}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$$

Notice this quantity is an idempotent quadratic form in a standard normal vector, $\left(\frac{\varepsilon}{\sigma}\right)$. Therefore, it has a chi-squared distribution with degrees of freedom equal to the $\text{rank}(M) = \text{tr}(M) = n - K$.

Till now, to construct the t statistic, we still need the independency of $\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$ and $\left(\frac{b-\beta}{\sigma}\right)$. To prove this, it suffices to show that $\left(\frac{b-\beta}{\sigma}\right) = (X'X)^{-1}X'\left(\frac{\varepsilon}{\sigma}\right)$ is independent of $\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$. We use the following result to finish the proof.

Independence of b and s^2 If ε is normally distributed, then the least squares coefficient estimator b is statistically independent of the residual vector e and therefore, all functions of e, including s^2.

Proof A sufficient condition for the independence of a linear form $L\mathbf{x}$ and an idempotent quadratic form $\mathbf{x}'A\mathbf{x}$ in a standard normal vector $\mathbf{x}$ is that, $LA = \mathbf{0}$. Letting $\left(\frac{\varepsilon}{\sigma}\right)$ be the $\mathbf{x}$. The requirement here would be $(X'X)^{-1}X'M = \mathbf{0}$. Notice that $MX = \mathbf{0}$, so the requirement is met. So we have proved the independency of $\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$ and $\left(\frac{b-\beta}{\sigma}\right)$. Since $\left(\frac{b-\beta}{\sigma}\right)$ is just a function of b, and $s^2 = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$, so we can end our proof to the general result.

Therefore, the ratio

$$t_k = \frac{(b_k - \beta_k)/\sqrt{\sigma^2 S^{kk}}}{\sqrt{\left[(n-K)s^2/\sigma^2\right]/(n-K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}}$$

has a t distribution with $(n-K)$ degrees of freedom. We can use t_k to test hypothesis or form confidence intervals about the individual elements of β.

The result of t_k differs from z_k in the use of s^2 instead of σ^2, and in the pivotal distribution, t with $(n-K)$ degrees of freedom, rather than standard normal. It follows that a confidence interval for β_k can be formed using

$$\left[b_k - t_{1-\alpha/2,n-K} \cdot \sqrt{s^2 S^{kk}} \leq \beta_k \leq b_k + t_{1-\alpha/2,n-K} \cdot \sqrt{s^2 S^{kk}}\right] = 1-\alpha$$

Notice here, the distribution of the pivotal statistic depends on the sample size through $(n-K)$, but once again, not on the parameters or the data. The practical advantage of t_k is that it does not involve any unknown parameters.

2.3.2 Prediction

Suppose that we wish to predict the value of y_0 associated with a regressor vector $\mathbf{x}_0$. The actual value would be

$$y_0 = \mathbf{x}_0'\beta + \varepsilon_0$$

It follows from the Gauss-Markov theorem that $\widehat{y}_0 = \mathbf{x}_0'b$ is the minimum variance linear unbiased estimator of $\mathrm{E}[y_0|\mathbf{x}_0] = \mathbf{x}_0'\beta$. The prediction error is

$$e_0 = \widehat{y}_0 - y_0 = \mathbf{x}_0'(b-\beta) + \varepsilon_0$$

The prediction variance of this estimator is

$$\begin{aligned}
\mathrm{Var}[e_0|X,\mathbf{x}_0] &= \sigma^2 + \mathrm{Var}[\mathbf{x}_0'(b-\beta)|X,\mathbf{x}_0] \\
&= \sigma^2 + \mathrm{Var}[\mathbf{x}_0'(X'X)^{-1}X'\varepsilon|X,\mathbf{x}_0] \\
&= \sigma^2 + \mathbf{x}_0'\left[\sigma^2(X'X)^{-1}\right]\mathbf{x}_0 \\
&= \mathrm{Var}[b|X] + \mathbf{x}_0'\mathrm{Var}[b|X]\,\mathbf{x}_0
\end{aligned}$$

The prediction variance can be estimated by using s^2 in place of σ^2. A confidence (prediction) interval for y_0 would then be formed using

$$\widehat{y}_0 \pm t_{1-\alpha/2,n-K} \cdot \mathrm{se}(e_0)$$

3 Hypothesis Testing

The general linear hypothesis is a *set* of J restrictions on the linear regression model $y = X\beta + \varepsilon$, which can be written as:

$$\begin{aligned} & R\beta = q \\ \iff & \begin{cases} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K = q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K = q_2 \\ \vdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K = q_J \end{cases} \end{aligned}$$

The matrix R has K columns to be conformable with β, J rows for a total of J restrictions. R must be *full row rank* to be meaningful, so J must be less than or equal to K. However, the case of $J = K$ must also be ruled out. If the K coefficients satisfy $J = K$ restrictions, then R is square and non-singular, so $\beta = R^{-1}q$. There would be no estimation or inference problem. The restrictions $R\beta = q$ imposes J restrictions on K otherwise free parameters. Hence, with the restrictions imposed, only $K - J$ free parameters are remained.

The hypothesis implied by the restrictions is written

$$\begin{aligned} H_0 &: R\beta - q = \mathbf{0} \\ H_1 &: R\beta - q \neq \mathbf{0} \end{aligned}$$

We will consider two approaches to testing the hypothesis, Wald tests and fit based tests. The hypothesis characterizes the population. If the hypothesis is correct, then the sample statistics should mimic that description. The tests will proceed as follows:

- Wald tests: The hypothesis states that $R\beta - q$ equals $\mathbf{0}$. The least squares estimator, b, is an unbiased and consistent estimator of β. If the hypothesis is correct, then the sample discrepancy, $R\beta - q$ should be close to zero.
- Fit based tests: We obtain the best possible fit—highest R^2—by using least squares without imposing the restrictions. We will show later that the sum of squares will never decrease when we impose the restrictions.

To develop the test statistics in this section, we will assume normally distributed disturbances. With this assumption, we will be able to obtain the exact distributions of the test statistics.

3.1 Wald Test

The Wald test is the most commonly used procedure. It is often called a "significance test." The operating principle of the procedure is to fit the regression without the restrictions, and then assess whether the results appear, within sampling variability, to agree with the hypothesis.

3.1.1 Single Coefficient

The Wald distance of a coefficient estimate from a hypothesized value is the linear distance, measured in standard deviation units. Thus, the distance of b_k from β_k^0 would be

$$W_k = \frac{b_k - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}$$

The following procedure is the same with constructing confidence interval, thus omitted.

3.1.2 Multiple Coefficients

Given the least squares estimator b, our interest centers on the *discrepancy vector* $Rb - q := m$. It is unlikely that m will be exactly $\mathbf{0}$. The statistical question is whether the deviation of m from $\mathbf{0}$ can be attributed to sampling error or whether it is significant.

Since b is normally distributed and m is a linear function of b, m is also normally distributed. If the null hypothesis is true, then $R\beta - q = \mathbf{0}$ and

$$\begin{aligned} \mathrm{E}\left[m|X\right] &= R\mathrm{E}\left[b|X\right] - q = R\beta - q = \mathbf{0} \\ \mathrm{Var}\left[m|X\right] &= \mathrm{Var}\left[Rb - q|X\right] = \mathrm{Var}\left[Rb|X\right] = R\mathrm{Var}\left[b|X\right]R' = \sigma^2 R(X'X)^{-1}R' \end{aligned}$$

We can base a test of H_0 on the *Wald criterion*. Conditioned on X, we find:

$$\begin{aligned} W &= m'\left\{\mathrm{Var}\left[m|X\right]\right\}^{-1} m \\ &= (Rb-q)'\left[\sigma^2 R(X'X)^{-1}R'\right]^{-1}(Rb-q) \\ &= \frac{(Rb-q)'\left[R(X'X)^{-1}R'\right]^{-1}(Rb-q)}{\sigma^2} \sim \chi^2\left[J\right] \end{aligned}$$

The statistic W has a chi-squared distribution with J degrees of freedom if the hypothesis is correct. Intuitively, the larger m is—that is, the worse the failure of least squares to satisfy the restrictions—the larger the chi-squared statistic. Therefore, a large chi-squared value will weigh against the hypothesis.
However, the chi-squared statistic of W is not usable because of the unknown σ^2. By using s^2 instead of σ^2 and dividing the result by J, we obtain a usable F statistic with J and $n - K$ degrees of freedom.

$$\begin{aligned} F &= \frac{W}{J} \cdot \frac{\sigma^2}{s^2} \\ &= \left(\frac{(Rb-q)'\left[R(X'X)^{-1}R'\right]^{-1}(Rb-q)}{\sigma^2}\right) \cdot \frac{1}{J} \cdot \frac{\sigma^2}{s^2} \cdot \frac{n-K}{n-K} \\ &= \frac{(Rb-q)'\left[\sigma^2 R(X'X)^{-1}R'\right]^{-1}(Rb-q)/J}{\left[(n-K)\,s^2/\sigma^2\right]/(n-K)} \end{aligned}$$

If the null hypothesis is true, that is, $R\beta = q$, then

$$Rb - q = Rb - R\beta = R(b - \beta) = R(X'X)^{-1}X'\varepsilon$$

Let $C = R(X'X)^{-1}R'$, then we have

$$\frac{Rb - q}{\sigma} = R(X'X)^{-1}X'\left(\frac{\varepsilon}{\sigma}\right) := D\left(\frac{\varepsilon}{\sigma}\right)$$

Then the numerator of F can be simplified as

$$\begin{aligned}
\frac{(Rb-q)'\left[\sigma^2 R(X'X)^{-1}R'\right]^{-1}(Rb-q)}{J} &= \frac{\left(\frac{Rb-q}{\sigma}\right)'\left\{R(X'X)^{-1}R'\right\}^{-1}\left(\frac{Rb-q}{\sigma}\right)}{J} \\
&= \frac{\left(\frac{\varepsilon}{\sigma}\right)'D'\left\{R(X'X)^{-1}R'\right\}^{-1}D\left(\frac{\varepsilon}{\sigma}\right)}{J} \\
&= \frac{\left(\frac{\varepsilon}{\sigma}\right)'D'C^{-1}D\left(\frac{\varepsilon}{\sigma}\right)}{J} \\
&:= \frac{\left(\frac{\varepsilon}{\sigma}\right)'T\left(\frac{\varepsilon}{\sigma}\right)}{J}
\end{aligned}$$

where $T = D'C^{-1}D$. The numerator is $\dfrac{W}{J}$ and is distributed as $\dfrac{1}{J}$ times a $\chi^2[J]$. Also, as we have done in constructing t statistic for testing significance of a single coefficient, the denominator of F what we are familiar with

$$\frac{\dfrac{(n-K)s^2}{\sigma^2}}{n-K} = \frac{\dfrac{\varepsilon' M\varepsilon}{\sigma^2}}{n-K} = \frac{1}{n-K}\cdot\left(\frac{\varepsilon}{\sigma}\right)'M\left(\frac{\varepsilon}{\sigma}\right) \sim \frac{1}{n-K}\cdot\chi^2[n-K]$$

Therefore, the F statistic is the ratio of two chi-squared variables each divided by its degrees of freedom. Since $M\left(\dfrac{\varepsilon}{\sigma}\right)$ and $T\left(\dfrac{\varepsilon}{\sigma}\right)$ are both normally distributed and their covariance TM is $\mathbf{0}$, the vectors of the quadratic forms are independent. The numerator and denominator of F are functions of independent random vectors and are therefore independent. This completes the proof of the F distribution.

Cancelling terms leaves the F statistic for testing a linear hypothesis:

$$F[J, n-K|X] = \frac{(Rb-q)'\left\{R\left[s^2(X'X)^{-1}\right]R'\right\}^{-1}(Rb-q)}{J}$$

Notice that degrees of freedom in the denominator is set for a certain regression, since $(n-K)$ is "used" to make up for substitute s^2 for σ^2. The only change with regard to

different hypothesis is the degrees of freedom in the numerator, which depends on the number of restrictions.

For testing one linear restriction of the form

$$H_0 : r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = r'\beta$$

The F statistic is them

$$F\left[1, n-K\right] = \frac{\left(\sum_j r_j b_j - q\right)^2}{\sum_j \sum_k r_j r_k \text{Est.Cov}\left[b_j, b_k\right]}$$

If the hypothesis is that the j-th coefficient is equal to a particular value, then R has a single row with a 1 in the j-th position and 0s elsewhere, so $R(X'X)^{-1}R'$ is the j-th diagonal element of the inverse matrix $(X'X)^{-1}$, and $Rb-q$ is (b_j-q). The F statistic is then

$$F\left[1, n-K\right] = \frac{\left(b_j - q\right)^2}{\text{Est.Var}\left[b_j\right]}$$

Consider an alternative approach to test a single restriction of testing $r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = r'\beta$. The sample estimate of $r'\beta$ is

$$r_1 b_1 + r_2 b_2 + \cdots + r_K b_K = \hat{q} = r'b$$

If $\hat{q}$ differs significantly from q, then we conclude that the sample data are not consistent with the hypothesis. It is natural to base the test on

$$t = \frac{\hat{q} - q}{\text{se}\left(\hat{q}\right)}$$

In words, t statistic is the distance in standard error units between the hypothesized function of the true coefficients and the same function of our estimates of them. If the hypothesis is true, then our estimates should reflect that, at least within the range of sampling variability. Thus, if the absolute value of the preceding t ratio is larger than the appropriate critical value, then doubt is cast on the hypothesis.

We require an estimate of the standard error of $\hat{q}$. Since $\hat{q}$ is a linear function of b and we have an estimate of the covariance matrix of b, which is $s^2(X'X)^{-1}$, we can estimate the variance of $\hat{q}$ with

$$\text{Est.Var}\left[\hat{q}|X\right] = r'\left[s^2(X'X)^{-1}\right]r$$

Therefore, we find

$$\begin{aligned} t &= \frac{\hat{q}-q}{\text{se}\left(\hat{q}\right)} = \frac{r'b-q}{\sqrt{r'\left[s^2(X'X)^{-1}\right]r}} \\ t^2 &= \frac{\left(\hat{q}-q\right)^2}{\left[\text{se}\left(\hat{q}\right)\right]^2} = \frac{\left(r'b-q\right)\left(r'b-q\right)}{r'\left[s^2(X'X)^{-1}\right]r} \\ &= \frac{\left(r'b-q\right)\left\{r'\left[s^2(X'X)^{-1}\right]r\right\}^{-1}\left(r'b-q\right)}{1} \\ &= F \end{aligned}$$

It follows, therefore, that for testing a single restriction, the t statistic is the square root of the F statistic that would be used to test that hypothesis.

3.2 Fit-Based Test

A different approach to hypothesis testing focuses on the fit of the regression. Recall that the least squares vector b was chosen to minimize the sum of squared deviations, $e'e$. Since R^2 equals $1 - \dfrac{e'e}{y'M_0y}$ and $y'M_0y$ is a constant that does not involve b, it follows that b is chosen to maximize R^2. One might ask whether choosing some other value for the slopes of the regression leads to a significant loss of fit. To develop the test statistic, we first examine the computation of the least squares estimator subject to a set of restrictions. We will then construct a test statistic that is based on comparing the R^2 s from the two regressions.

3.2.1 Constrained Estimator

Suppose that we explicitly impose the restriction of the general linear hypothesis in the regression. The restricted least squares estimator is obtained as the solution to

$$\begin{aligned} &\min_{b_0} S(b_0) = (y - Xb_0)'(y - Xb_0) \\ &\text{s.t. } Rb_0 = q \end{aligned}$$

A Lagrangian function for this problem can be written

$$\mathcal{L}(b_0, \lambda) = \left(y - Xb_0\right)'\left(y - Xb_0\right) + 2\lambda'\left(Rb_0 - q\right)^2$$

(Notice that since λ is not restricted, we formulate the constraints in terms of 2λ for convenience of scaling later.)

The solutions b_0 and λ will satisfy the necessary conditions

$$\frac{\partial \mathcal{L}}{\partial b_0} = -2X'\left(y - Xb_0\right) + 2R'\lambda = \mathbf{0}$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = 2\left(Rb_0 - q\right) = \mathbf{0}$$

which can be expressed in the following partitioned matrix equation:

$$\begin{bmatrix} X'X & R' \\ R & \mathbf{0} \end{bmatrix} \begin{bmatrix} b_0 \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ q \end{bmatrix}$$

If, in addition, $X'X$ is not singular, then explicit solutions for b_0 and λ can be obtained by using the partitioned inverse formula:

$$b_0 = b - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(Rb - q)$$
$$\lambda = \left[R(X'X)^{-1}R'\right]^{-1}(Rb - q)$$

However, the partitioned inverse formula is unfriendly to remember, so another useful way is to smartly solve the necessary condition equations:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial b_0} = \mathbf{0} \\ \frac{\partial \mathcal{L}}{\partial \lambda} = \mathbf{0} \end{cases} \Longrightarrow \begin{cases} -X'y + X'Xb_0 + R'\lambda = \mathbf{0} \\ Rb_0 = q \end{cases}$$

To make use of the restriction that $Rb_0 = q$, if $X'X$ is invertible, we premultiply $R(X'X)^{-1}$ in the first equation:

$$\begin{aligned} & -X'y + X'Xb_0 + R'\lambda = \mathbf{0} \\ \Longrightarrow & -R(X'X)^{-1}X'y + R(X'X)^{-1}X'Xb_0 + R(X'X)^{-1}R'\lambda = \mathbf{0} \\ \Longleftrightarrow & -Rb + Rb_0 + R(X'X)^{-1}R'\lambda = \mathbf{0} \\ \Longleftrightarrow & -Rb + q + R(X'X)^{-1}R'\lambda = \mathbf{0} \\ \Longrightarrow & \lambda = \left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q) \end{aligned}$$

After solving $\lambda = \left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q)$, we insert this into the first equation to obtain b_0:

$$\begin{aligned} & -X'y + X'Xb_0 + R'\lambda = \mathbf{0} \\ \Longleftrightarrow & X'Xb_0 = X'y - R'\lambda \\ \Longleftrightarrow & X'Xb_0 = X'y - R'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q) \\ \Longrightarrow & b_0 = (X'X)^{-1}X'y - (X'X)^{-1}R'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q) \\ & \quad = b - (X'X)^{-1}R'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q) \end{aligned}$$

Intuitively, the constrained solution b_0 is equal to the unconstrained solution b *minus* a term that accounts for the failure of the unrestricted solution to satisfy the constraints.

To get the variance of b_0, as we have done before, we need to work out the random part of b_0. Under the hypothesis of restriction, $R\beta = q$, so

$$\begin{aligned}
b_0 &= b - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q) \\
&= b - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - R\beta) \\
&= b - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(b - \beta) \\
&= b - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'\varepsilon \\
&= \beta + (X'X)^{-1}X'\varepsilon - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'\varepsilon \\
&= \beta + \left\{(X'X)^{-1}X' - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'\right\}\varepsilon
\end{aligned}$$

Notice that we should not forget b also contains the random part, so we need to segment it into $b = \beta + (X'X)^{-1}X'\varepsilon$. Then we forward to get the variance of b_0:

$$\begin{aligned}
\text{Var}\left[b_0|X\right] &= \left\{(X'X)^{-1}X' - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'\right\}\left\{\text{Var}\left[\varepsilon|X\right]\right\}\left\{(X'X)^{-1}X' - \right. \\
&= \sigma^2\left\{(X'X)^{-1}X' - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'\right\}\left\{(X'X)^{-1}X' - (X'X)^{-1}\right. \\
&= \sigma^2\left\{(X'X)^{-1}X' - (X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'\right\}\left\{X(X'X)^{-1} - X(X'X)^{-}\right. \\
&= \sigma^2\left\{(X'X)^{-1}X'X(X'X)^{-1}\right\} \\
&\quad - \sigma^2\left\{(X'X)^{-1}X'X(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}\right\} \\
&\quad - \sigma^2\left\{(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'X(X'X)^{-1}\right\} \\
&\quad + \sigma^2\left\{(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'X(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'\right. \\
&= \sigma^2(X'X)^{-1} - \sigma^2\left\{(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}\right\} - \sigma^2\left\{(X'X)^{-1}R'\left\{R(X'X\right.\right. \\
&\quad + \sigma^2\left\{(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}\right\} \\
&= \sigma^2(X'X)^{-1} - \sigma^2\left\{(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}\right\} \\
&= \text{Var}\left[b|X\right] - \sigma^2\left\{(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}\right\}
\end{aligned}$$

We hope to prove that the second term, $\sigma^2\left\{(X'X)^{-1}R' \left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}\right\}$, is a non-negative definite matrix. From the core out, since $X'X$ is a positive definite matrix, so $(X'X)^{-1}$ is a positive matrix. Because R has full row

rank, $R(X'X)^{-1}R'$ is then a positive definite matrix, so is $\left\{R(X'X)^{-1}R'\right\}^{-1}$. In the same logic, $R'\left\{R(X'X)^{-1}R'\right\}^{-1}R$ is non-negative definite, and finally $(X'X)^{-1}R'\left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}$ is non-negative definite.

3.2.2 Loss-of-Fit Statistics

To develop a test based on the restricted least squares estimator, we consider a single coefficient first and then turn to the general case of J linear restrictions. Consider the change in the fit of a multiple regression when a variable z is added to a model that already contains $K-1$ variables, $\mathbf{x}$. We showed in before that the effect on the fit would be given by

$$R^2_{Xz} = R^2_X + (1 - R^2_X)r^{*2}_{yz}$$

where R^2_{Xz} is the new R^2 after z is added, R^2_X is the original R^2 and r^*_{yz} is the partial correlation between y and z, controlling for $\mathbf{x}$. So as we knew, the fit improves. In deriving the partial correlation coefficient between y and z in preceding sections, we obtained the convenient result

$$r^{*2}_{yz} = \frac{t^2_z}{t^2_z + (n-K)}$$

where t_z is the square of the t ratio for testing the hypothesis that the coefficient on z is zero in the *multiple* regression of y on X and z. Based on the two equations, we have

$$t^2_z = \frac{\left(R^2_{Xz} - R^2_x\right)/1}{\left(1 - R^2_{Xz}\right)/(n-K)}$$

We have developed in Wald test that in testing for a single restriction,

$$F\left[1, n-K\right] = t^2\left[n-K\right]$$

We see that the squared t statistic (i.e., the F statistic) is computed using the change in the R^2. By interpreting the preceding as the result of removing z from the regression, we see that we have proved a result for the case of testing whether a single slope is zero. By this construction, we see that for a single restriction, F is a measure of the loss of fit that results from imposing that restriction. Next, we will proceed to the general case of J linear restrictions, which will include one restriction as a special case.

The fit of the restricted least squares coefficients cannot be better than that of the unrestricted solution.

Change in Sum of Squares with Different Coefficient Vectors Suppose that b is the least squares coefficient vector in the regression of y on X, and that c is any other $K \times 1$ vector. The difference in the two sums of squared residual is

$$(y - Xc)'(y - Xc) - (y - Xb)'(y - Xb) = (c - b)'X'X(c - b)$$

As we can see directly, this difference is positive.

Let e_0 equal $y - Xb_0$. Then, using a familiar device,

$$\begin{aligned} e_0 &= y - Xb_0 = y - Xb - X(b_0 - b) \\ &= e - X(b_0 - b) \end{aligned}$$

The new sum of squared deviation is (recall the result that $X'e = 0$)

$$e_0'e_0 = e'e + (b_0 - b)'X'X(b_0 - b) \geq e'e$$

The loss of fit is

$$\begin{aligned} e_0'e_0 - e'e &= (b_0 - b)'X'X(b_0 - b) \\ &= \left((X'X)^{-1}R'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q)\right)'X'X\left((X'X)^{-1}R'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q)\right) \\ &= (Rb - q)'\left\{R(X'X)^{-1}R'\right\}^{-1}R(X'X)^{-1}X'X(X'X)^{-1}R'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q) \\ &= (Rb - q)'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q) \end{aligned}$$

Recall that the expression, $(Rb - q)'\left\{R(X'X)^{-1}R'\right\}^{-1}(Rb - q)$, appears in the numerator of F statistic. Inserting the remaining parts, we obtain

$$F[J, n - K] = \frac{(e_0'e_0 - e'e)/J}{e'e/(n - K)}$$

Finally, by dividing both numerator and denominator of F by $y'M_0y$, we obtain the general result involving R^2:

$$F[J, n - K] = \frac{(R^2 - R_0^2)/J}{(1 - R^2)/(n - K)}$$

This form has some intuitive appeal in that the difference in the fits of the two models is directly incorporated in the test statistic. As an example of this approach, consider the joint test that all the slopes in the model are zero. In this case, $R_0^2 = 0$, the F statistic is then $F[K, n - K] = \dfrac{R^2/K}{(1 - R^2)/(n - K)}$. For imposing a set of exclusion

restrictions such as $\beta_k = 0$ for one or more coefficients, the obvious approach is simply to omit the variables from the regression and base the test on the sums of squared residuals for the restricted and unrestricted regressions.

Remarks: The R^2-form of F statistic is based on regressions for the same dependent variable. If the set of restriction requires an equivalent restricted regression with a different dependent variable, then R^2-form is invalid.

Analytically, if we are to test the hypothesis that a subset of coefficients, say β_2, are all zero, which is constructed using $R = [\mathbf{0}; \mathbf{I}]$, $q = \mathbf{0}$, and number of restrictions J is equal to K_2, the number of elements in β_2, then the matrix $R(X'X)^{-1}R'$ is the $K_2 \times K_2$ lower right block of the full inverse matrix. From partitioned inverse formula,

$$R(X'X)^{-1}R' = (X_2'M_1X_2)^{-1}$$

And clearly, $Rb - q = [\mathbf{0}; \mathbf{I}] \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} - \mathbf{0} = b_2$. Jointly,

$$\begin{aligned} e_0'e_0 - e'e &= (Rb - q)' \left\{R(X'X)^{-1}R'\right\}^{-1} (Rb - q) \\ &= b_2'X_2'M_1X_2b_2 \end{aligned}$$

(Notice that the result has in fact been proved from another perspective in section of goodness of fit.)

The procedure for computing the appropriate F statistic amounts simply to comparing the sums of squared deviations from the "short" and "long" regressions, which we saw earlier.

4 Endogeneity and Instrumental Variable

4.1 Violation of Exogeneity

Three basic assumptions for linear model:

1. X has full rank
2. Exogeneity: $E[\varepsilon|X] = 0$
3. Homoskedasticity: $Var(\varepsilon|X) = E[\varepsilon\varepsilon'|X] = \sigma^2\mathbf{I}$.

Endogeneity happens when exogeneity is violated, i.e., $E[\varepsilon|X] \neq 0$. Sources of endogeneity usually are:

- Omitted variable bias: endogenous treatment effects, omitted parameter heterogeneity
- Simultaneous/Reverse causality
- Measurement error
- Sample selection bias: non-random sampling, attrition (survivorship bias)

In violation of exogeneity, there must be some correlation between the disturbances and the independent variables. We assume:

$$E[\varepsilon|X] = \eta$$

which means that the regressors now provide information about the expectations of the disturbances. An important implication of this assumption is that, the disturbances and the regressors are now correlated, specifically

$$E[x_i\varepsilon_i] = \gamma$$

for some nonzero γ. If the data are "well-behaved", then we can apply Khinchine's theorem to asset that

$$\text{plim } \frac{X'\varepsilon}{n} = \gamma$$

Khinchine's Weak Law of Large Numbers If $x_i(i = 1, 2, \cdots, n)$ is a random (i.i.d.) sample from a distribution with finite mean $E[x_i] = \mu$, then

$$\text{plim } \bar{x}_n = \mu$$

Then the estimator of b is biased:

$$\begin{aligned} b &= (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon \\ \Longrightarrow E[b|X] &= \beta + (X'X)^{-1}X'\eta \neq \beta \end{aligned}$$

Original OLS estimator b is also inconsistent:

$$\begin{aligned}\text{plim } b &= \beta + \text{plim } \left(\frac{X'X}{n}\right)^{-1} \cdot \text{plim } \left(\frac{X'\varepsilon}{n}\right) \\ &= \beta + Q_{XX}^{-1}\gamma \neq \beta\end{aligned}$$

The inconsistency of least squares is not confined to the endogenous variable(s). Even though only one of the variables in X is correlated with ε, all of the elements of b are inconsistent, not just the estimator of the coefficient on the endogenous variable. The inconsistency due to the endogeneity of the one variable is smeared across all of the least squares estimators.

4.2 IV Estimation

Luckily, beyond X, there is an additional set of variables, Z, that have two properties:

1. Exogeneity: IV uncorrelated with the disturbance.
2. Relevance: IV correlated with the independent variable X.

In the context of our model, variables that have these two properties are instrumental variables. We assume the following:

1. $[x_i, z_i, \varepsilon_i](i = 1, \cdots, n)$ are i.i.d. sequence of random variables.
2. $E[x_{ik}^2] = Q_{XX}$, a finite constant.
   - plim $\frac{X'X}{n} = Q_{XX}$
3. $E[z_{il}^2] = Q_{ZZ}$, a finite constant.
   - plim $\frac{Z'Z}{n} = Q_{ZZ}$
4. $E[z_{il}x_{ik}] = Q_{ZX}$, a finite constant.
   - plim $\frac{Z'X}{n} = Q_{ZX}$
5. $E[\varepsilon_i|z_i] = 0$
   - plim $\frac{Z'\varepsilon}{n} = 0$

The requirements for IV can also be described as:

1. Exogeneity: plim $\frac{Z'\varepsilon}{n} = 0$.
2. Relevance: plim $\frac{Z'X}{n} = Q_{XX}$, a finite $L \times K$ matrix with rank K.
   - Baseline for coefficient estimation: similar to full column rank of X before.

3. Well-behaved data: plim $\frac{Z'Z}{n} = Q_{ZZ}$, a positive definite matrix.

We discuss two cases, where $L = K$ or $L > K$.

4.2.1 $L = K$

First consider the situation when $L = K$. We partition X into x_1, a set of K_1 exogenous variables; and x_2, a set of K_2 endogenous variables, then $Z = [x_1, z_2]$, where z_2 are the instrumental variables for x_2, and x_1 are the instrumental variables for themselves.

Because $E[z_i\varepsilon_i] = 0$ and all terms have finite variances, we have:

$$\begin{aligned}
\text{plim } \left(\frac{Z'\varepsilon}{n}\right) &= \text{plim } \left(\frac{Z'(y - X\beta)}{n}\right) \\
&= \text{plim } \left(\frac{Z'y}{n}\right) - \text{plim } \left(\frac{Z'X\beta}{n}\right) \\
&= 0 \\
\Longrightarrow \text{plim } \left(\frac{Z'y}{n}\right) &= \text{plim } \left(\frac{Z'X\beta}{n}\right) = \beta \cdot \text{plim } \left(\frac{Z'X}{n}\right)
\end{aligned}$$

We have assumed that Z has the same number of variables as X, so $Z'X$ is a square matrix. Moreover, the rank of $Z'X$ is also K, therefore plim $\left(\frac{Z'X}{n}\right)$ is invertible. Then we have:

$$\left[\text{plim } \left(\frac{Z'X}{n}\right)\right]^{-1} \text{plim } \left(\frac{Z'y}{n}\right) = \beta$$

which leads us to the *instrumental variable estimator*:

$$b_{IV} = (Z'X)^{-1}Z'y$$

From the preceding deduction of IV estimator we have already proved that b_{IV} is consistent. Alternatively, in the traditional way we can see:

$$\begin{aligned}
b_{IV} &= (Z'X)^{-1}Z'y \\
&= (Z'X)^{-1}Z'(X\beta + \varepsilon) \\
&= \beta + (Z'X)^{-1}Z'\varepsilon
\end{aligned}$$

It is clear that b_{IV} is consistent since plim $\frac{Z'\varepsilon}{n} = 0$.

We now turn to the asymptotic distribution of b_{IV}. We will use the same method as in previous asymptotic distribution of b_{LS}. First,

$$\sqrt{n}(b_{IV} - \beta) = \left(\frac{Z'X}{n}\right)^{-1} \frac{1}{\sqrt{n}} Z'\varepsilon$$

which has the same limiting distribution as $Q_{ZX}^{-1} \cdot \frac{1}{\sqrt{n}} Z'\varepsilon$. Analysis of distribution of $Z'\varepsilon$ can be the same as that of $X'\varepsilon$, so it follows that:

$$\begin{aligned} &\left(\frac{1}{\sqrt{n}} Z'\varepsilon\right) \xrightarrow{\mathrm{d}} N\left[0, \sigma^2 Q_{ZZ}\right] \\ \Longrightarrow &\left(\frac{Z'X}{n}\right)^{-1} \left(\frac{1}{\sqrt{n}} Z'\varepsilon\right) \xrightarrow{\mathrm{d}} N\left[0, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right] \end{aligned}$$

Thus, the asymptotic distribution of b_{IV} is

$$b_{IV} \overset{a}{\sim} N\left[\beta, \frac{\sigma^2}{n} Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right]$$

The asymptotic covariance matrix is estimated as:

$$\begin{aligned} Est.Asy.Var[b_{IV}] &= \hat{\sigma}^2 (Z'X)^{-1} (Z'Z) (X'Z)^{-1} \\ &= \hat{\sigma}^2 \left[X'Z (Z'Z)^{-1} Z'X\right]^{-1} \end{aligned}$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i' b_{IV}\right)^2$$

$\hat{\sigma}^2$ is a *consistent* estimator of σ^2. A correction for degrees of freedom is superfluous here, as all results here are asymptotic.

Equivalently, we can see the variance matrix of b_{IV} from its expression:

$$\begin{aligned} b_{IV} &= (Z'X)^{-1} Z'y = (Z'X)^{-1} Z'(X\beta + \varepsilon) = \beta + (Z'X)^{-1} Z'\varepsilon \\ \Longrightarrow Var(b_{IV}|X) &= (Z'X)^{-1} Z' Var(\varepsilon|X) Z (X'Z)^{-1} \\ &= \sigma^2 (Z'X)^{-1} Z'Z (X'Z)^{-1} \\ &= \sigma^2 [X'Z (Z'Z)^{-1} Z'X]^{-1} \end{aligned}$$

In sum, IV estimation share some technical part in common with OLS. See motivation of IV from linear model:

$$\begin{aligned} &y = X\beta + \varepsilon \\ \Longrightarrow &X'y = X'X\beta + X'\varepsilon \\ \Longrightarrow &\text{plim } \frac{X'y}{n} = \text{plim } \frac{X'X}{n} \cdot \beta + \text{plim } \frac{X'\varepsilon}{n} \\ \Longrightarrow &Q_{Xy} = Q_{XX}\beta + \gamma \end{aligned}$$

From $Q_{Xy} = Q_{XX}\beta + \gamma$ we can see that, β and γ cannot be jointly identified without further assumptions. In standard OLS assumption, $\gamma = 0$ is a key assumption of exogeneity, then β can be estimated. However in violation of exogeneity, we introduce IV with assumption of plim $\frac{Z'\varepsilon}{n} = 0$ for estimation:

$$\begin{aligned} & y = X\beta + \varepsilon \\ \Longrightarrow & Z'y = Z'X\beta + Z'\varepsilon \\ \Longrightarrow & \text{plim } \frac{Z'y}{n} = \text{plim } \frac{Z'X}{n} \cdot \beta + \text{plim } \frac{Z'\varepsilon}{n} \\ \Longrightarrow & Q_{Zy} = Q_{ZX}\beta \end{aligned}$$

4.2.2 $L > K$

The crucial results in all the preceding of IV is based on

$$\text{plim } \frac{Z'\varepsilon}{n} = \mathbf{0}$$

That is, every column of Z is asymptotically uncorrelated with ε. That also means that every linear combination of the columns of Z is also uncorrelated with ε, which suggests that one approach would be to choose K linear combinations of the columns of Z. Specially when we consider the situation when $L > K$, i.e., the instrumental variables are more than the endogenous variables. To make use of the L-many IVs, a better choice is the **projection of the columns of X in the column space of Z**:

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

Intuitively, $\hat{X}$ is the predicted value of X when regressing X on Z. After projected into column space of Z, $\hat{X}$ is exogenous and uncorrelated with the disturbance. From our motivation, $\hat{X}$ has K columns of IVs that are combinations of Z. We then have (substitute $\hat{X}$ for "Z" in expression of b_{IV} since $\hat{X}$ is a set of K IVs):

$$\begin{aligned} b_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'y \end{aligned}$$

The variance matrix of b_{IV} is the same as before,

$$\begin{aligned} b_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'y \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'(X\beta + \varepsilon) \\ &= \beta + \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\varepsilon \\ \Longrightarrow \text{Var}\left[b_{IV}\right] &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\text{Var}\left[\varepsilon|X\right] Z(Z'Z)^{-1}Z'X \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} \\ &= \sigma^2 \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} \end{aligned}$$

Remarks: In computation of the estimated asymptotic covariance matrix, $\hat{\sigma}^2$ should not be based on $\hat{X}$. The estimator $s_{IV}^2 = \dfrac{\left(y - \hat{X}b_{IV}\right)' \left(y - \hat{X}b_{IV}\right)}{n}$ is inconsistent for σ^2, with or without a correction for degrees of freedom.

Alternatively, we can rewrite b_{IV} as

$$\begin{aligned} b_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\ &= (X'P_Z'X)^{-1}\hat{X}'y \\ &= (X'P_Z'P_ZX)^{-1}\hat{X}'y \\ &= \left[(P_ZX)'(P_ZX)\right]^{-1}\hat{X}'y \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \end{aligned}$$

$b_{IV} = (\hat{X}'X)^{-1}\hat{X}'y$ has practical meanings, which is the theoretical foundation for Two Stage Least Squares (2SLS). In practice, b_{IV} can be estimated in two steps:

1. X is regressed on Z, and predict $\hat{X}$.
2. y is regressed on $\hat{X}$ to get $b_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$.

2SLS estimator is not only intuitive and easy to compute, but also efficient in the IV class of linear combinations of Z.

Efficiency of 2SLS Estimator Of all the different linear combinations of Z, $\hat{X}$ is the most efficient in the sense that the asymptotic covariance matrix of an IV estimator based on a linear combination ZF is smaller when $F = (Z'Z)^{-1}Z'X$ than with any other F that uses all L columns of Z.

Proof Denote a combination of columns of Z as $\tilde{Z} = ZF$, where F is a $L \times K$ matrix. IV estimator with $\tilde{Z}$ is

$$\begin{aligned} b_{\tilde{Z}} &= (\tilde{Z}'X)^{-1}\tilde{Z}'y \\ &= (\tilde{Z}'X)^{-1}\tilde{Z}'(X\beta + \varepsilon) \\ &= \beta + (\tilde{Z}'X)^{-1}\tilde{Z}'\varepsilon \end{aligned}$$

Therefore, the $(\tilde{Z}'X)^{-1}\tilde{Z}'\varepsilon$ part in $b_{\tilde{Z}}$ is its random part. We have

$$\begin{aligned} Var[b_{\tilde{Z}}|X] &= (\tilde{Z}'X)^{-1}\tilde{Z}'Var[\varepsilon|X]\tilde{Z}(X'\tilde{Z})^{-1} \\ &= \sigma^2(\tilde{Z}'X)^{-1}\tilde{Z}'\tilde{Z}(X'\tilde{Z})^{-1} \end{aligned}$$

Since $\tilde{Z}'X, \tilde{Z}'Z$, and $X'\tilde{Z}$ are square and full rank matrices, thus invertible. We rewrite the covariance matrix for $b_{\tilde{Z}}$ as:

$$\begin{aligned} Var[b_{\tilde{Z}}|X] &= \sigma^2(\tilde{Z}'X)^{-1}\tilde{Z}'\tilde{Z}(X'\tilde{Z})^{-1} \\ &\sigma^2\{X'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'X\}^{-1} \end{aligned}$$

The 2SLS IV is $\hat{Z} = P_Z X = Z(Z'Z)^{-1}Z'X$ (the same as preceding $\hat{X}$; write as $\hat{Z}$ for uniformity in notations), and its coefficients estimator is:

$$b_{\hat{Z}} = (\hat{Z}'X)^{-1}\hat{Z}'y = \beta + (\hat{Z}'X)^{-1}\hat{Z}'\varepsilon$$

Same as before, we focus on the random part in $b_{\hat{Z}}$ to obtain its covariance matrix:

$$\begin{aligned} Var[b_{\hat{Z}}|X] =& \sigma^2(\hat{Z}'X)^{-1}\hat{Z}'\hat{Z}(X'\hat{Z})^{-1} \\ =& \sigma^2\{X'Z(Z'Z)^{-1}Z'X\}^{-1}\{X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X\}\{X'Z(Z'Z)^{-1}Z'X\}^{-1} \\ =& \sigma^2\{X'Z(Z'Z)^{-1}Z'X\}^{-1} \end{aligned}$$

Denote $P_Z = Z(Z'Z)^{-1}Z', P_{\tilde{Z}} = \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'$. So the covariance matrices for $b_{\tilde{Z}}$ and $b_{\hat{Z}}$ can be simplified to

$$\begin{aligned} Var[b_{\tilde{Z}}|X] &= \sigma^2\{X'P_{\tilde{Z}}X\}^{-1} \\ Var[b_{\hat{Z}}|X] &= \sigma^2\{X'P_Z X\}^{-1} \end{aligned}$$

To compare the two covariance matrices, we only need to compare $X'P_{\tilde{Z}}X$ and $X'P_Z X$. To compare $X'P_{\tilde{Z}}X$ and $X'P_Z X$, we only need to compare $P_{\tilde{Z}}$ and P_Z. Hence, we first construct the "difference matrix":

$$D = P_Z - P_{\tilde{Z}}$$

Apparently, since both P_Z and $P_{\tilde{Z}}$ are projection matrices, they are symmetric and idempotent. So D is immediately symmetric. We are then interested to see if D is also idempotent.

$$\begin{aligned} D^2 = DD &= (P_Z - P_{\tilde{Z}})(P_Z - P_{\tilde{Z}}) \\ &= P_Z - P_Z p_{\tilde{Z}} - P_{\tilde{Z}}P_Z + P_{\tilde{Z}} \end{aligned}$$

Thus, our interest flows to $P_Z P_{\tilde{Z}}$ and $P_{\tilde{Z}}P_Z$:

$$\begin{aligned} P_Z P_{\tilde{Z}} &= Z(Z'Z)^{-1}Z'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}' \\ &= Z(Z'Z)^{-1}Z'ZF(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}' \\ &= ZF(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}' \\ &= \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}' \\ &= P_{\tilde{Z}} \\ P_{\tilde{Z}}P_Z &= \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'Z(Z'Z)^{-1}Z' \\ &= \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}F'Z'Z(Z'Z)^{-1}Z' \\ &= \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}F'Z' \\ &= \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}' \\ &= P_{\tilde{Z}} \end{aligned}$$

Plug in those two results to D^2, we can see:

$$\begin{aligned} D^2 = DD &= (P_Z - P_{\tilde{Z}})(P_Z - P_{\tilde{Z}}) \\ &= P_Z - P_Z p_{\tilde{Z}} - P_{\tilde{Z}} P_Z + P_{\tilde{Z}} \\ &= P_Z - P_{\tilde{Z}} \\ &= D \end{aligned}$$

Therefore, our conjecture of D being idempotent has been proved!

Since D is both symmetric and idempotent, then take any vector x,

$$x'Dx = X'DDx = X'D'Dx = (Dx)'(DX) = ||DX|| \geq 0$$

So D is non-negative definite. From this result, we can see that $X'DX$ is also non-negative definite:

$$y'X'DXy = (Xy)'D(Xy) \geq 0$$

Or alternatively, $y'X'DXy = y'X'D'DXy = (DXy)'(DXy) = ||DXy|| \geq 0$.

Since $X'DX$ is non-negative, that is, semi-positive definite, we have $Var[b_{\hat{Z}}|X] \leq Var[b_{\tilde{Z}}|X]$. In conclusion, the 2SLS estimator is (asymptotically) efficient in the class of all IV estimators using linear combination of Z. ■

4.3 Specification Tests

4.3.1 Hausman Test

Before introducing IV, we should conduct tests to inspect endogeneity or exogeneity. Under classical Guass-Markov assumptions, if X is uncorrelated with ε, then b_{LS} is the most efficient estimator. So if we do not have endogeneity, going with b_{LS} is apparently a better choice with more estimation precision. Consider a comparison of the two covariance matrices under the hypothesis that both estimators are consistent, that is, assuming plim $\frac{X'\varepsilon}{n} = 0$. The difference between the asymptotic covariance matrices of the two estimators is

$$\begin{aligned} \text{Asy.Var}\left[b_{IV}\right] - \text{Asy.Var}\left[b_{LS}\right] &= \frac{\sigma^2}{n} \cdot \text{plim } \left(\frac{X'Z(Z'Z)^{-1}Z'X}{n}\right) - \frac{\sigma^2}{n} \cdot \text{plim } \left(\frac{X'X}{n}\right)^{-1} \\ &= \frac{\sigma^2}{n} \cdot \text{plim } \left(n\left[\left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} - (X'X)^{-1}\right]\right) \end{aligned}$$

To compare the two matrices in the brackets, we can compare their inverses. The inverse of the first is $X'Z(Z'Z)^{-1}Z'X = X'\left(1 - M_Z\right)X = X'X - X'M_ZX$. Because M_Z is a non-negative definite matrix, it follows that $X'M_ZX$ is also. So $X'Z(Z'Z)^{-1}Z'X$ equals $X'X$ minus a non-negative definite matrix. Because $X'Z(Z'Z)^{-1}Z'X$ is smaller,

in the matrix sense, than $X'X$, so its inverse is larger. Under the hypothesis, the asymptotic covariance matrix of the LS estimator is never larger than that of the IV estimator, and it will actually be smaller unless all the columns of X are perfectly predicted by regressions on Z. Thus, we have established that if plim $\frac{X'\varepsilon}{n} = \mathbf{0}$—that is, if LS is consistent—then it is a preferred estimator.

Our interest in the difference between these two estimators goes beyond the question of efficiency. The null hypothesis of interest will usually be specifically whether plim $\frac{X'\varepsilon}{n} = \mathbf{0}$. Seeking the covariance between X and ε through $\frac{X'e}{n}$ is fruit- less, of course, because the normal equations produce $X'e = \mathbf{0}$. The logic of Hausman's test is as follows.

- Null hypothesis: X is uncorrelated with ε;
- Alternative hypothesis: X is correlated with ε.

Under the null hypothesis, both b_{LS} and b_{IV} are consistent estimators of β. While under the alternative hypothesis. Only b_{IV} is consistent. We construct $d = b_{IV} - b_{LS}$. Under the null hypothesis, plim $d = 0$, whereas under the alternative, plim $d \neq 0$. Hence, we have a Wald statistic,

$$H = d'\{Est.Asy.Var[d]\}^{-1}d$$

The asymptotic covariance matrix we need for the test is

$$\begin{aligned}\text{Asy.Var}\left[b_{IV} - b_{LS}\right] =& \text{Asy.Var}\left[b_{IV}\right] + \text{Asy.Var}\left[b_{LS}\right] \\ & - \text{Asy.Cov}\left[b_{IV}, b_{LS}\right] - \text{Asy.Cov}\left[b_{LS}, b_{IV}\right]\end{aligned}$$

Hausman gives a fundamental result that allows us to proceed.

Hausman Result The covariance between an efficient estimator b_E of a parameter vector β and its difference from an inefficient estimator b_I of the same parameter vector $b_E - b_I$ is zero, which means

$$\begin{aligned}& Cov(b_E, b_E - b_I) = 0 \\ \iff & Cov(b_E, b_I) = Var(b_E)\end{aligned}$$

Hence, plug the result back to asymptotic variance matrix of $d = b_{IV} - b_{LS}$, we have

$$\begin{aligned}\text{Asy.Var}\left[b_{IV} - b_{LS}\right] =& \text{Asy.Var}\left[b_{IV}\right] + \text{Asy.Var}\left[b_{LS}\right] \\ & - \text{Asy.Cov}\left[b_{IV}, b_{LS}\right] - \text{Asy.Cov}\left[b_{LS}, b_{IV}\right] \\ =& \text{Asy.Var}\left[b_{IV}\right] + \text{Asy.Var}\left[b_{LS}\right] \\ & - \text{Asy.Var}\left[b_{LS}\right] - \text{Asy.Var}\left[b_{LS}\right] \\ =& \text{Asy.Var}\left[b_{IV}\right] - \text{Asy.Var}\left[b_{LS}\right]\end{aligned}$$

Inserting this useful and critical result into our Wald statistic and reverting to our empirical estimates of these quantities, we have

$$H = (b_{IV} - b_{LS})' \left\{ \text{Asy.Var}\left[b_{IV}\right] - \text{Asy.Var}\left[b_{LS}\right] \right\}^{-1} (b_{IV} - b_{LS}) = (b_{IV} - b_{LS})' \left\{ \sigma_{IV}^2 (\hat{X}'\hat{X})^{-1} - \sigma_{LS}^2 (X'X$$

Remarks:

1. Under large samples, $s_{IV}^2 = s_{LS}^2 = s^2$ and both are consistent estimators to σ_{IV}^2 and σ_{LS}^2, respectively. Here we do not make adjustments for substituting s^2 for the unknown σ^2 in the statistic. Because rigorously speaking, if we do make such adjustments, the H is then reverted to a F statistic, and the corresponding F distribution is asymptotically a χ^2 distribution as if we had not done adjustments.

2. It is tempting to invoke our results for the full rank quadratic form in a normal vector and conclude the degrees of freedom for this chi-squared statistic is K. But that method will usually be incorrect, and worse yet, unless X and Z have no variables in common, the rank of the matrix in this statistic is less than K, and the ordinary inverse will not even exist. In most cases, at least some of the variables in X will also appear in Z. (In almost any application, X and Z will both contain the constant term.) That is, some of the variables in X are known to be uncorrelated with the disturbances. In this case, our hypothesis, plim $\frac{X'X}{n} = \mathbf{0}$, does not really involve all K variables, because a subset of the elements in this vector are known to be zero, say K_0-many. As such, the quadratic form in the Wald test is being used to test only $K^* = K - K_0$ hypothesis. In the meantime, it is easy and useful to show that, H is in fact a rank K^* quadratic form. Since $Z(Z'Z)^{-1}Z'$ is an idempotent matrix, $\left(\hat{X}'\hat{X}\right) = \hat{X}'X$. Using this result and expanding d, we find

$$\begin{aligned} d &= b_{IV} - b_{LS} \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'y - (X'X)^{-1}X'y \\ &= (\hat{X}'\hat{X})^{-1}\left[\hat{X}'y - (\hat{X}'\hat{X})(X'X)^{-1}X'y\right] \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'\left[y - X(X'X)^{-1}X'y\right] \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'e \end{aligned}$$

where e is the vector of least squares residuals. Recall that for exogenous variables in X, they are IV for themselves, and satisfy $X'e = 0$. Denote the endogenous part of X as $\hat{X}^*$, then d can be expressed in the form:

$$d = (\hat{X}'\hat{X})^{-1} \begin{pmatrix} 0 \\ \hat{X}^{*\prime} e \end{pmatrix} = (\hat{X}'\hat{X})^{-1} \begin{pmatrix} 0 \\ q \end{pmatrix}$$

Finally, denote the entire matrix in H by W. (Because that the ordinary inverse may not exist, this matrix will have to be a generalized inverse.) Then, denoting the whole matrix product by P, we obtain

$$H = \begin{bmatrix} \mathbf{0}' & q' \end{bmatrix} (\hat{X}'\hat{X})^{-1} W (\hat{X}'\hat{X})^{-1} \begin{bmatrix} \mathbf{0} \\ q \end{bmatrix} = \begin{bmatrix} \mathbf{0}' & q' \end{bmatrix} P \begin{bmatrix} \mathbf{0} \\ q \end{bmatrix} = q' P_{**} q$$

where P_{**} is the lower right $K^* \times K^*$ submatrix of P. We now have the end result, that algebraically H is actually a quadratic form in a K^* vector, so K^* is the degrees of freedom for the test.

1. Hausman test can be applied to extensive situations. Suppose we have a pair of estimators $\hat{\theta}_E$ and $\hat{\theta}_I$, such that under the null hypothesis that $\hat{\theta}_E$ and $\hat{\theta}_I$ are both consistent, while under the alternative hypothesis only $\hat{\theta}_I$ remains consistent and $\hat{\theta}_E$ is inconsistent, then we can form a test of the hypothesis by referring the Hausman statistic:

$$H = \left(\hat{\theta}_I - \hat{\theta}_E\right)' \left\{ \text{Est.Asy.Var}\left[\hat{\theta}_I\right] - \text{Est.Asy.Var}\left[\hat{\theta}_E\right] \right\} \left(\hat{\theta}_I - \hat{\theta}_E\right) \xrightarrow{\text{d}} \chi^2 [J]$$

where the appropriate degrees of freedom for the test, J, will depend on the context.

4.3.2 Wu Test

The preceding Wald test requires a generalized inverse, so it is going to be a bit cumbersome. In fact, one need not actually approach the test in this form, and it can be carried out with any regression program. The alternative variable addition test approach devised by Wu (1973) is simpler. An F statistic with K^* and $n - K - K^*$ degrees of freedom can be used to test the joint significance of the elements of γ in the augmented regression:

$$y = X\beta + \hat{X}^*\gamma + \varepsilon^*$$

where $\hat{X}^*$ are the fitted values in the regressions of the variables in X^* on Z. This result is equivalent to the Hausman test for this model.

4.4 Measurement Error

The general assessment of measurement error problem is not particularly optimistic. The bias introduced by measurement error can be rather severe. There are almost no known finite-sample results for the models of measurement error; nearly all the results that have been developed are asymptotic.

Consider a regression model with a single regressor and no constant term:

$$y^* = \beta x^* + \varepsilon$$

where y^* and x^* are not available. Instead we can only observe y and x:

$$y = y^* + v, \text{ with } v \sim N[0, \sigma_v^2]$$
$$x = x^* + v, \text{ with } u \sim N[0, \sigma_u^2]$$

First assume for the moment that only y^* is measured with error:

$$y = \beta x^* + \varepsilon + v = \beta x^* + \varepsilon'$$

This result conforms to the assumptions of the classical regression model. As long as the regressor is measured properly, measurement error on y^* can be absorbed in the disturbance of the regression and hence ignored.

Then suppose x has measurement error (y without), we have

$$\begin{aligned} y &= \beta x^* + \varepsilon \\ &= \beta(x - u) + \varepsilon \\ &= \beta x + (\varepsilon - \beta u) \end{aligned}$$

To check exogeneity, we inspect covariance of x and the disturbance term $(\varepsilon - \beta u)$ as a whole:

$$\text{Cov}\left[x, \varepsilon - \beta u\right] = \text{Cov}\left[x^* + u, \varepsilon - \beta u\right] = -\beta\sigma_u^2 \neq 0$$

The nonzero correlation violates assumption of exogeneity. As $b = \dfrac{\frac{1}{n}\sum_{i=1}^n x_i y_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}$, we have

$$\text{plim } b = \frac{\text{plim } \frac{1}{n}\sum_{i=1}^n (x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\text{plim } \frac{1}{n}\sum_{i=1}^n (x_i^* + u_i)^2}$$

Because x^*, ε and u are mutually independent, this equation reduces to

$$\text{plim } b = \frac{\beta Q^*}{Q^* + \sigma_u^2} = \frac{\beta}{1 + \sigma_u^2/Q*} < \beta$$

where $Q^* = \text{plim } \dfrac{\sum_{i=1}^n x_i^2}{n}$.

As long as σ_u^2 is positive, b is then inconsistent, with a persistent bias towards 0. Clearly, the greater the variability in the measurement error, the worse the bias. The effect of biasing the coefficient towards zero is called **attenuation**.

In a multiple regression model, matters only get worse. When only a single variable is measured with error (assume the first variable), we have

$$\text{plim } b_1 = \frac{\beta_1}{1 + \sigma_u^2 q^{*11}}$$
$$\text{plim } b_k = \beta_k - \beta_1 \left[\frac{\sigma_u^2 q^{*k1}}{1 + \sigma_u^2 q^{*11}} \right], \text{ for } k \neq 1$$

where q^{*k1} is the $(k, 1)$th element in $(Q^*)^{-1}$.

The result in multiple regression case is not required. Notice that the coefficient on the badly measured variable is still biased toward zero. The other coefficients are all biased as well, although in *unknown directions*. A badly measured variable contaminates all the least squares estimates. If more than one variable is measured with error, there is very little that can be said.

5 Generalized Regression Model and Heteroskedasticity

Generalized linear regression model applies when homoskedasticity is violated, i.e.,

$$E[\varepsilon\varepsilon'|X] = \sigma^2\Omega = \Sigma$$

Remarks: For mathematical convenience, we normalize the variance matrix so that $\text{tr}\,(\Omega) = n$.

The generalized linear regression model is

$$y = X\beta + \varepsilon$$

with assumptions of

$$\begin{cases} E[\varepsilon|X] = 0 \\ E[\varepsilon\varepsilon'|X] = \sigma^2\Omega = \Sigma \end{cases}$$

The two leading cases we will consider in detail are heteroskedasticity and autocorrelation. Disturbances are heteroskedastic when they have different variances. Heteroskedasticity arises in volatile high-frequency time-series data and in cross-section data. Microeconomic data usually has heteroskedasticity. In heteroskedasticity, the disturbances are still assumed to be uncorrelated across observations, so $\sigma^2\Omega$ would be

$$\sigma^2\Omega = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Autocorrelation is usually found in time-series data. Economic time series often display a "memory" in that variation around the regression function is not independent from one period to the next. Time-series data are usually homoskedastic, so $\sigma^2\Omega$ might be:

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}$$

Panel data may exhibit both heteroskedasticity and autocorrelation.

$$\Gamma = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix}$$

$$\text{and } \sigma^2\Omega = \begin{bmatrix} \Gamma_1 & 0 & \cdots & 0 \\ 0 & \Gamma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Gamma_n \end{bmatrix}$$

In comparison with classical linear model, the quick conclusion is that, classical model with spherical disturbances, i.e., $\mathrm{E}[\varepsilon|X] = \mathbf{0}$ and $\mathrm{E}[\varepsilon\varepsilon'|X] = \sigma^2\mathbf{I}$, then the OLS estimator is best linear unbiased, consistent, and asymptotically normally distributed; **when homoskedasticity is violated, OLS estimator is still unbiased, consistent and asymptotically normally distributed, but no longer efficient, so the usual inference procedures are no longer appropriated.**

5.1 Properties of OLS Estimator

5.1.1 Unbiasedness

$$b = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon$$

If $E[\varepsilon|X] = \mathbf{0}$, then

$$\mathrm{E}[b] = \mathrm{E}_X[\mathrm{E}[b|X]] = \beta$$

The covariance matrix of the disturbance vector has played no role here; unbiasedness is a property of the means.

5.1.2 (No-Longer) Efficiency

$$\begin{aligned} b &= \beta + (X'X)^{-1}X'\varepsilon \\ \Longrightarrow \mathrm{Var}[b|X] &= \mathrm{E}[(b - E[b])(b - E[b])'|X] \\ &= \mathrm{E}[(b-\beta)(b-\beta)'|X] \\ &= \mathrm{E}[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\Omega X\right)\left(\frac{1}{n}X'X\right)^{-1} \end{aligned}$$

As we know, under classical linear model's assumptions, especially the homoskedasticity assumption, the OLS estimator has variance matrix $\mathrm{Var}[b|X] = \sigma^2(X'X)^{-1}$ and is efficient. In the case of heteroskedasticity, the variance matrix is

$\sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$, which is different from $\sigma^2(X'X)^{-1}(X'\mathbf{I}X)(X'X)^{-1}$, therefore, the estimator is no longer efficient.

5.1.3 Consistency

5.1.3.1 Consistency of b Since b is still unbiased, one natural idea is that, if $Var[b|X]$ converges to zero, then b is mean square consistent. With well-behaved regressors, $\left(\frac{X'X}{n}\right)^{-1}$ will converge to a constant matrix, so we consider

$$\frac{\sigma^2}{n}\left(\frac{X'\Omega X}{n}\right) = \left(\frac{\sigma^2}{n}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}x_i x_j'}{n}\right)$$

We see that though the leading constant will by itself converge to 0; the following matrix is a sum of n^2 terms, divided by n. Thus, the product is a scalar that if $O(\frac{1}{n})$ times a matrix that is at least at this juncture, $O(n)$, which results in $O(1)$. So, it does appear at first blush that if the product of $Var[b|X]$ does converge, it might converge to a matrix of nonzero constants. In this case, the covariance matrix of the least squares estimator would not converge to zero, and consistent would be hard to establish through mean square convergence. However, if only $\left(\frac{X'\Omega X}{n}\right)$ converges to a constant matrix, then $Var[b|X]$ would then converge to zero.

Consistency of OLS in the Generalized Regression Model If $Q = \text{plim}\left(\frac{X'X}{n}\right)$ and plim $\left(\frac{X'\Omega X}{n}\right)$ are both finite positive definite matrices, then b is consistent for β. Under the assumed conditions,

$$\text{plim } b = \beta$$

5.1.3.2 Consistency of s^2 Under assumption of plim $\frac{X'\Omega X}{n}$ being a finite positive matrix, we inspect the consistency of s^2:

$$s^2 = \frac{e'e}{n-K} = \frac{\varepsilon' M\varepsilon}{n-K}$$

As $M = \mathbf{I} - X(X'X)^{-1}X'$, we have

$$s^2 = \frac{\varepsilon'\varepsilon}{n-K} - \frac{\varepsilon X(X'X)^{-1}X'\varepsilon}{n-K}$$

For the first part, as $\text{tr}(\Omega) = n$ (which is a man-made normalization), we have

$$\text{E}\left[\frac{\varepsilon'\varepsilon}{n-K}\middle| X\right] = \frac{\text{tr}\left(\text{E}\left[\varepsilon\varepsilon'|X\right]\right)}{n-K} = \frac{n\sigma^2}{n-K}$$

For the second part, we have

$$\begin{aligned}
\mathrm{E}\left[\frac{\varepsilon' X(X'X)^{-1}X'\varepsilon}{n-K}\middle| X\right] &= \frac{\mathrm{tr}\left\{\mathrm{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon' X|X\right]\right\}}{n-K} \\
&= \frac{\mathrm{tr}\left\{\sigma^2\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\right\}}{n-K} \\
&= \frac{\sigma^2}{n-K}\mathrm{tr}\left\{\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\right\}
\end{aligned}$$

As $n \to \infty$, the first part will converge to σ^2; the second part will converge to zero because both matrices in the product are finite. Therefore, if b is consistent, then $\lim_{n\to\infty} s^2 = \sigma^2$.

In addition, if the fourth moment of every disturbance is finite and all other assumptions are met, then

$$\lim_{n\to\infty} \mathrm{Var}\left[\frac{e'e}{n-K}\right] = \lim_{n\to\infty} \mathrm{Var}\left[\frac{\varepsilon'\varepsilon}{n-K}\right] = 0$$

The result implies that

$$\mathrm{plim}\ b = \beta \Longrightarrow \mathrm{plim}\ s^2 = \sigma^2$$

5.1.3.3 Heteroskedasticity and Var $[b|X]$

$$\begin{aligned}
\mathrm{Var}\left[b|X\right] &= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\
&= (X'X)^{-1}\left(\sigma^2\sum_{i=1}^{n}\omega_i x_i x_i'\right)(X'X)^{-1}
\end{aligned}$$

The difference between the conventional estimator and the appropriate (true) covariance matrix for b is

$$\mathrm{Est.Var}\left[b|X\right] - \mathrm{Var}\left[b|X\right] = s^2(X'X)^{-1} - \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$$

In large sample cases, $s^2 \approx \sigma^2$, so the difference is approximately equal to

$$\begin{aligned}
\sigma^2\frac{X'X}{n} - \sigma^2\frac{X'X}{n}(X'\Omega X)\frac{X'X}{n} &= \sigma^2\frac{X'X}{n}(X'X)\frac{X'X}{n} - \sigma^2\frac{X'X}{n}(X'\Omega X)\frac{X'X}{n} \\
&= \sigma^2\frac{X'X}{n}\left\{(X'X) - (X'\Omega X)\right\}\frac{X'X}{n} \\
&= \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i)x_i x_i'\right)\left(\frac{X'X}{n}\right)^{-1}
\end{aligned}$$

If the heteroskedasticity is not correlated with the **variables** (instead of observations) in the model, then at least in large samples, the ordinary least squares computations, although not the optimal way to use the data, will not be misleading.

The preceding is a useful result, but one should not be overly optimistic. First, it remains true that ordinary least squares is demonstrably inefficient. Second, if the primary assumption of the analysis—that the heteroskedasticity is unrelated to the variables in the model—is incorrect, then the conventional standard errors may be quite far from the appropriate values.

5.1.4 Asymptotic Normal Distribution

Asymptotic Distribution of b in the Generalized Regression Model If the regressors are sufficiently well behaved and the off-diagonal terms in Ω diminish sufficiently, then the OLS estimator is asymptotically normally distributed with mean β and covariance matrix

$$\text{Asy.Var}\,[b] = \frac{\sigma^2}{n} Q^{-1} \text{plim}\ \left(\frac{X'\Omega X}{n}\right) Q^{-1}$$

Remarks: The condition "off-diagonal tears in ω diminish sufficiently" is required for applying the CLT, which requires a sequence of independent random variables.

5.1.5 IV Estimator

$$\begin{aligned} b_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'y \\ &= \beta + \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\varepsilon \end{aligned}$$

Define $Q_{XX,Z}$ for convenience of matrix expression:

$$Q_{XX,Z} = \text{plim}\ \left\{ \left(\frac{X'Z}{n}\right) \left(\frac{Z'Z}{n}\right)^{-1} \left(\frac{Z'X}{n}\right) \right\}^{-1} \left(\frac{X'Z}{n}\right) \left(\frac{Z'Z}{n}\right)^{-1}$$

If the random term in b_{IV}, $\left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\varepsilon$, vanishes asymptotically, then

$$\text{plim}\ b_{IV} = \beta + Q_{XX,Z}\text{plim}\ \left(\frac{Z'\varepsilon}{n}\right) = \beta$$

And

$$b_{IV} \overset{a}{\sim} N\left[\beta, \frac{\sigma^2}{n}(Q_{XX,Z})\text{plim}\ \left(\frac{Z'\Omega Z}{n}\right)(Q'_{XX,Z})\right]$$

5.2 Efficient Estimation by GLS

We start by considering the case in which Ω is a **known**, symmetric and positive definite matrix.

We have $\Omega = C\Lambda C'$. Let $T = C\Lambda^{1/2}$, and thus $\Omega = TT'$. Also, let $P = \Lambda^{-1/2}C', P' = C\Lambda^{-1/2}$, we have

$$\begin{cases} \Omega^{-1} & = (C\Lambda C')^{-1} = (C')^{-1}\Lambda^{-1}C^{-1} = C\Lambda^{-1}C' \\ P'P & = C\Lambda^{-1}C' \end{cases} \Longrightarrow \Omega^{-1} = P'P$$

Consider the linear model of $y = X\beta + \varepsilon$, with $\mathrm{E}[\varepsilon\varepsilon'|X] = \sigma^2\Omega$. We do the following transformation:

$$\begin{aligned} y &= X\beta + \varepsilon \\ \Longrightarrow Py &= PX\beta + P\varepsilon \\ (y_*) &= (X_*\beta) + (\varepsilon_*) \end{aligned}$$

Luckily, we will show that **the transformed linear model is homoskedastic**:

$$\begin{aligned} \mathrm{E}[\varepsilon_*\varepsilon_*'|X] &= \mathrm{E}[P\varepsilon\varepsilon'P'|X_*] \\ &= \mathrm{E}[P\varepsilon\varepsilon'P'|X] \\ &= P(\sigma^2\Omega)P' \\ &= \sigma^2\left(\Lambda^{-1/2}C'\right)(C\Lambda C')\left(C\Lambda^{-1/2}\right) \\ &= \sigma^2\mathbf{I} \end{aligned}$$

Remarks: From here we can understand the construction of P more intuitively. Our goal is to make the transformed model stick with the homoskedasticity assumption. To realize this, we shall "scale" the disturbances in a way that make the scaled version has an identity variance matrix. That is, we scale ε by P, and hope $\mathrm{Var}[P\varepsilon|X] = P\mathrm{Var}[\varepsilon|X]P' = \sigma^2 P\Omega P' = \sigma^2\mathbf{I}$. Hence, P works like the inverse of the square root of Ω, so $P = \Lambda^{-1/2}C'$.

The OLS estimator b is then

$$\begin{aligned} b &= (X_*'X_*)^{-1}X_*'y_* \\ &= \left[(PX)'(PX)\right]^{-1}(PX)'(Py) \\ &= (X'P'PX)^{-1}(X'P'Py) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \end{aligned}$$

The variance of this estimator is then

$$\begin{aligned} b &= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y \\ &= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\left(X\beta+\varepsilon\right) \\ &= \beta + \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\varepsilon \\ \text{Or equivalently, } b &= (X_*'X_*)^{-1}X_*'y_* \\ &= (X_*'X_*)^{-1}X_*'\left(X\beta+\varepsilon\right) \\ &= \beta + (X_*'X_*)^{-1}X_*'\varepsilon \\ \Longrightarrow \text{Var}\left[b|X\right] &= \sigma^2(X_*'X_*)^{-1} \\ &= \sigma^2\left(X'P'PX\right)^{-1} \\ &= \sigma^2\left(X'\Omega X\right)^{-1} \end{aligned}$$

Since $\text{E}\left[\varepsilon_*\varepsilon_*'|X\right] = \sigma^2\mathbf{I}$, the classical regression model applies to this transformed model. Hence the $b = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega y$ is the efficient estimator of β. This estimator is the **generalized least squares (GLS)** or Aitken estimator of β. This estimator is in contrast to the ordinary least squares estimator, which uses a "weighting matrix", $\mathbf{I}$ instead of Ω^{-1}.

Properties of the GLS Estimator

1. If $\text{E}\left[\varepsilon|X\right] = \mathbf{0}$, then the GLS estimator is unbiased.
2. The GLS estimator is consistent if $\text{plim}\ \dfrac{X_*'X_*}{n} = Q_*$, where Q_* is a finite positive definite matrix. (That is, we require the transformed data $X_* = PX$, not the original data, to be well-behaved.)
3. The GLS estimator is asymptotically normally distributed, with mean β and sampling variance $\text{Var}\left[b|X\right] = \sigma^2(X_*'X_*)^{-1} = \sigma^2\left(X'\Omega^{-1}X\right)^{-1}$.
4. The GLS estimator is the minimum variance linear unbiased estimator in the generalized regression model. (This statement follows by applying the Gauss-Markov theorem to the transformed model. In fact, the broad result includes the Gauss-Markov theorem as a special case when $\Omega = \mathbf{I}$.)

Moreover, all problems regarding this way of heteroskedasticity with known Ω can be addressed by the transformed linear model, because the transformed linear regression model again satisfies the classical linear model assumptions.

There is no precise counterpart to R^2 in the generalized regression model. Alternatives have been proposed, but care must be taken when using them. For example, one choice is the R^2 in the transformed regression. But this regression need not have a constant term, so the R^2 is not bounded by zero and one. Even if there is a constant term, the transformed regression is a computational device, not the model of interest. That a good

(or bad) fit is obtained in the transformed model may be of no interest; the dependent variable in that model, y_*, is different from the one y in the model as originally specified. The usual R^2 often suggests that the fit of the model is improved by a correction for heteroskedasticity and degraded by a correction for autocorrelation, but both changes can often be attributed to the computation of y_*. A more appealing fit measure might be based on the residuals from the original model once the GLS estimator is in hand, such as

$$R_G^2 = 1 - \frac{(y - Xb_{GLS})'(y - Xb_{GLS})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Like the earlier contender, however, this measure is not bounded in the unit interval. In addition, this measure cannot be reliably used to compare models. The generalized least squares estimator minimizes the generalized sum of squares

$$\varepsilon_*'\varepsilon_* = (y - X\beta)\,\Omega^{-1}(y - X\beta)$$

not $\varepsilon'\varepsilon$. As such, there is no assurance, for example, that dropping a variable from the model will result in a decrease in R_G^2 , as it will in R^2. The R^2-like measures in this setting are purely descriptive. That being the case, the squared sample correlation between the actual and predicted values, $r_{y,\hat{y}}^2 = corr^2(y, \hat{y}) = corr^2\left(y, x'\hat{\beta}\right)$, would likely be a useful descriptor. Note, though, that this is not a proportion of variation explained, as is R^2; it is a measure of the agreement of the model predictions with the actual data.

5.3 Feasible GLS

When Ω is **unknown**, but has **specific structures,** we can estimate β using feasible generalized least squares (FGLS). A typical problem involves a set of parameters α such that $\Omega = \Omega(\alpha)$. A typical example is autocorrelation: $\Omega(\rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho^2 & \cdots & \rho^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$. Also, another usually-assumed heteroskedasticity is formed as: $\sigma_i^2 = \sigma^2 z_i^{\theta}$.

Procedure of Feasible GLS works as:

- Estimate $\hat{\alpha}$, a consistent estimator of α; (We will consider later how such an estimator might be obtained.)
- Obtain $\hat{\Omega} = \Omega(\hat{\alpha})$ and estimate β using $\hat{\beta} = \left(X'\hat{\Omega}^{-1}X\right)^{-1}X'\hat{\Omega}^{-1}y$, when satisfying

$$\begin{cases} \text{plim } \left[\frac{1}{n}X'\hat{\Omega}^{-1}X - \frac{1}{n}X'\Omega^{-1}X\right] = \mathbf{0} \\ \text{plim } \left[\frac{1}{\sqrt{n}}X'\hat{\Omega}^{-1}\varepsilon - \frac{1}{\sqrt{n}}X'\Omega^{-1}\varepsilon\right] = \mathbf{0} \end{cases}$$

The first of these equations states that if the weighted sum of squares matrix based on the true Ω converges to a positive definite matrix, then the one based on $\hat{\Omega}$ converges to the same matrix. We are assuming that this is true. In the second condition, if the transformed regressors are well behaved, then the right-hand-side sum will have a limiting normal distribution.

Efficiency of the FGLS Estimator An asymptotically efficient FGLS estimator does not require that we have an efficient estimator of α; only a consistent one is required to achieve full efficiency for the FGLS estimator.

5.3.1 Weighted Least Squares

In the most general case, given that $\sigma_i^2 = \sigma^2 w_i$, we have

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix}$$

$$\implies \Omega^{-1} = \begin{bmatrix} \frac{1}{\omega_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\omega_n} \end{bmatrix}$$

$$P = P' = \begin{bmatrix} \frac{1}{\sqrt{\omega_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\omega_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\omega_n}} \end{bmatrix}$$

Therefore, the GLS estimator is obtained by regressing

$$Py = \begin{bmatrix} \frac{y_1}{\sqrt{\omega_1}} \\ \frac{y_2}{\sqrt{\omega_2}} \\ \vdots \\ \frac{y_n}{\sqrt{\omega_n}} \end{bmatrix} \text{ on } PX = \begin{bmatrix} \frac{x_1'}{\sqrt{\omega_1}} \\ \frac{x_2'}{\sqrt{\omega_2}} \\ \vdots \\ \frac{x_n'}{\sqrt{\omega_n}} \end{bmatrix}$$

Hence, the **weighted least square (WLS)** estimator is

$$\hat{\beta} = \left[\sum_{i=1}^{n} w_i x_i x_i'\right]^{-1} \left[\sum_{i=1}^{n} w_i x_i y_i\right], \text{ where } w_i = \frac{1}{\omega_i}$$

The logic of the computation is that observations with smaller variances receive a larger weight in the computations of the sums and therefore have greater influence in the estimates obtained.

5.3.1.1 WLS with Known Ω A common specification is that the variance is proportional to one of the regressors or its square. Say if $\sigma_i^2 = \sigma^2 x_{ik}^2$, then the transformed regression model for GLS is

$$\frac{y}{x_k} = \beta_k + \beta_1 \left(\frac{x_1}{x_k}\right) + \beta_2 \left(\frac{x_2}{x_k}\right) + \cdots + \frac{\varepsilon}{x_k}$$

However, it is rarely possible to be certain about the nature of the heteroskedasticity in a regression model. In one respect, this problem is minor, since the WLS estimator $\hat{\beta} = \left[\sum_{i=1}^n w_i x_i x_i'\right]^{-1} \left[\sum_{i=1}^n w_i x_i y_i\right]$ is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances.

5.3.1.2 WLS with Unknown Ω Two-Step estimators:

1. Obtain estimates $\hat{\sigma}_i$ using $e_i = y_i - x_i'b$ based on OLS regression;

   The OLS estimator of β, although inefficient, is still consistent. As such, statistics computed using the OLS residuals, $e_i = y_i - x_i'b$, will have the same asymptotic properties as those computed using the true disturbances, $\varepsilon_i = y_i - x_i'\beta$.

2. Obtain $\hat{\hat{\beta}}$ using

$$\hat{\hat{\beta}} = \left[\sum_{i=1}^n \left(\frac{1}{\hat{\sigma}_i^2}\right) x_i x_i'\right] \left[\sum_{i=1}^n \left(\frac{1}{\sigma_i^2} x_i y_i\right)\right]$$

Remarks: The two-step estimator may be iterated by recomputing the residuals after computing the FGLS estimates and then reentering the computation.

5.4 White heteroskedasticity consistent estimator

We treat $\sigma^2\Omega$ as a whole parameter, and assume $\text{tr}\Omega = n$, which is just a normalization. $\sigma^2\Omega$ contains $1 + \cdots + n = \dfrac{n\,(n+1)}{2}$ unknown parameters, which is almost impossible to estimate with a moderately large sample. We define

$$Q^* = \text{plim}\ \frac{X'\left(\sigma^2\Omega\right)X}{n} = \text{plim}\ \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}^2 x_i x_j'$$

The least square estimator b is a consistent estimator of β, which implies that the least squares e_i are "pointwise" consistent estimators of their population counterparts ε_i. The

general approach is to use X and e to devise an estimator of Q for the heteroskedasticity case, i.e., $\sigma_{ij} = 0$ when $i \neq j$.

White shows that, under very general conditions, the estimator $S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'$ has plim $S_0 = Q^*$. Hence, White heteroskedasticity consistent estimator is

$$\text{Est.Asy.Var}\left[b\right] = \frac{1}{n}\left(\frac{X'X}{n}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'\right]\left(\frac{X'X}{n}\right)^{-1}$$

which can be used to estimate the asymptotic covariance matrix of b.

This result implies that without actually specifying the type of heteroskedasticity, we can still make appropriate inferences based on the least squares estimator. This implication is especially useful if we are unsure of the precise nature of the heteroskedasticity (which is probably most of the time). However, the estimator is too optimistic when the sample is small.

5.5 White Test for Heteroskedasticity

Tests for heteroskedasticity are based on the following strategy. Ordinary least squares is a consistent estimator of β even in the presence of heteroskedasticity. As such, the ordinary least squares residuals will mimic, albeit imperfectly because of sampling variability, the heteroskedasticity of the true disturbances. Therefore, tests designed to detect heteroskedasticity will, in general, be applied to the ordinary least squares residuals.

$$H_0 : \sigma_i^2 = \sigma^2, \forall i = 1, 2, \cdots, n$$
$$H_1 : \text{Not } H_0$$

The intuition of White test for heteroskedasticity is that, if there is no heteroskedasticity, then $V = s^2(X'X)^{-1}$ will give a consistent estimator of $\text{Var}\left[b|X\right] = \sigma^2(X'X)^{-1}\left(X'\Omega X\right)(X'X)^{-1}$, which as we have seen can be estimated using White's heteroskedasticity consistent estimator.

Operationally speaking, White test is an F test to test the hypothesis that $\gamma_1 = \gamma_2 = 0$ in the following regression:

$$e_i^2 = \gamma_0 + x_i'\gamma_i + \left(x_i \otimes x_i\right)'\gamma_2 + v_i$$

Remarks: The White test is extremely general. To carry it out, we need not make any specific assumptions about the nature of the heteroscedasticity. Although this characteristic is a virtue, it is, at the same time, a potentially serious shortcoming. The test may reveal heteroskedasticity, but it may instead simply identify some other specification error (such as the omission of x^2 from a simple regression). Except in the context of a specific problem, little can be said about the power of White's test; it may

be very low against some alternatives. In addition, unlike some of the other tests we shall discuss, the White test is **nonconstructive**. If we reject the null hypothesis, then the result of the test gives no indication of what to do next.

5.6 Clustering and the Moulton Factor

A bivariate model,

$$y_{ig} = \beta_0 + \beta_1 x_g + \varepsilon_{ig}$$

where $\varepsilon_{ig} = v_g + \eta_{ig}$. We assume that v and η are independent, $v_g \sim N\left(0, \sigma_v^2\right)$ and $\eta_{ig} \sim N\left(0, \sigma_\eta^2\right)$.

Therefore,

$$\text{Cov}\left[\varepsilon_{ig}, \varepsilon_{jg}\right] = \text{Cov}\left[v_g + \eta_{ig}, v_g + \eta_{jg}\right] = \sigma_v^2$$

As $\text{Var}\left[\varepsilon_{ig}\right] = \sigma_\varepsilon^2 = \sigma_v^2 + \sigma_\eta^2$, and denote $\rho = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} = \dfrac{\sigma_v^2}{\sigma_\varepsilon^2}$, so we have

$$\text{Cov}\left[\varepsilon_{ig}, \varepsilon_{jg}\right] = \sigma_v^2 = \rho\sigma_\varepsilon^2 > 0$$

The conclusion is that, if we do not cluster the standard errors to the right level, the variance will be mistakenly cut down by a Moulton factor:

$$\frac{V\left(\hat{\beta}_1\right)}{V_c\left(\hat{\beta}_1\right)} = 1 + (n-1)\,\rho$$

From the setting, we can first clarify the variables of interest:

$$y_g = \begin{bmatrix} y_{1g} \\ y_{2g} \\ \vdots \\ y_{n_g g} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix}$$

$$\varepsilon_g = \begin{bmatrix} \varepsilon_{1g} \\ \varepsilon_{2g} \\ \vdots \\ \varepsilon_{n_g g} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_G \end{bmatrix}$$

$$X = \begin{bmatrix} l_1 x_1 \\ l_2 x_2 \\ \vdots \\ l_G x_G \end{bmatrix}$$

where l_g is a column vector of n_g-many 1s, and G is the number of groups.

$$\mathrm{E}\left[\varepsilon\varepsilon'|X\right] = \Omega = \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{bmatrix}$$

$$\begin{aligned} \Omega_g &= \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \\ &= \sigma_\varepsilon^2 \left[(1-\rho)\,\mathbf{I} + \rho \cdot \left(l_g l_g'\right)\right] \end{aligned}$$

where $\left(l_g l_g'\right)$ is a $g \times g$ matrix full of 1s, and $\rho = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}$.

As we know, the heteroskedastic covariance matrix is $(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$, and the homoskedastic covariance matrix is $\sigma^2(X'X)^{-1}$, so we compute the following two

matrices $X'X$ and $X'\Omega X$:

$$\begin{aligned}
X'X &= \begin{bmatrix} x_1'l_1' & x_2'l_2' & \cdots & x_G'l_G' \end{bmatrix} \begin{bmatrix} l_1x_1 \\ l_2x_2 \\ \vdots \\ l_Gx_G \end{bmatrix} \\
&= \sum_{g=1}^{G} n_g x_g' x_g \\
X'\Omega X &= \begin{bmatrix} x_1'l_1' & x_2'l_2' & \cdots & x_G'l_G' \end{bmatrix} \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{bmatrix} \begin{bmatrix} l_1x_1 \\ l_2x_2 \\ \vdots \\ l_Gx_G \end{bmatrix} \\
&= \sum_{g=1}^{G} x_g'l_g'\Omega_g l_g x_g \\
&= \sum_{g=1}^{G} x_g'l_g' \left\{\sigma_\varepsilon^2 \left[(1-\rho)\,\mathbf{I} + \rho l_g l_g'\right]\right\} l_g x_g \\
&= \sigma_\varepsilon^2 \sum_{g=1}^{G} \left\{x_g' \left[(1-\rho)\,l_g'l_g + \rho l_g'l_g l_g'l_g\right]\right\} x_g \\
&= \sigma_\varepsilon^2 \sum_{g=1}^{G} \left\{x_g' \left[(1-\rho)\,n_g + \rho n_g^2\right]\right\} x_g \\
&= \sigma_\varepsilon^2 \sum_{g=1}^{G} n_g \left[1 + \left(n_g - 1\right)\rho\right] x_g'x_g \\
&:= \sigma_\varepsilon^2 \sum_{g=1}^{G} n_g \tau_g x_g'x_g
\end{aligned}$$

If the group sizes are equal, that is, $n_g = n$ for all $g = 1, \cdots, G$; then $\tau := \tau_g = 1 + (n-1)\,\rho$. In this case,

$$\begin{aligned}
V\left(\hat{\beta}\right) &= \sigma_\varepsilon^2 \tau \left(\sum_g n x_g'x_g\right)^{-1} \left(\sum_g n x_g'x_g\right) \left(\sum_g n x_g'x_g\right)^{-1} \\
&= \sigma_\varepsilon^2 \tau \left(\sum_g n x_g'x_g\right)^{-1} \\
&= \tau V_c\left(\hat{\beta}\right)
\end{aligned}$$

So we jointly have

$$\frac{V\left(\widehat{\beta}\right)}{V_c\left(\widehat{\beta}\right)} = \tau = 1 + (n-1)\,\rho$$

which we called the Moulton factor.

The basic framework of a regression model for panel data is

$$\begin{aligned} y_{it} &= x_{it}'\beta + z_i'\alpha + \varepsilon_{it} \\ &= x_{it}'\beta + c_i + \varepsilon_{it} \end{aligned}$$

There are K regressors in x_{it}, not including a constant term. The heterogeneity, or individual effect is $z_i'\alpha$ where z_i contains a constant term and a set of individual or group-specific variables which are constant over time t. The heterogeneity can be either observed (e.g., sex and education) or unobserved (e.g., family specific characteristics, individual heterogeneity in skill or preference). If z_i is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares. The complications arise when c_i is unobserved, which will be the case in most applications.

The main objective of the analysis will be consistent and efficient estimation of the partial effects,

$$\beta = \frac{\partial \mathrm{E}\left[y_{it}|x_{it}\right]}{\partial x_{it}}$$

Whether this is possible depends on the assumptions about the unobserved effects. We begin with a strict exogeneity assumption for the independent variables. Strict exogeneity requires that

$$\mathrm{E}\left[\varepsilon_{it}|x_{i1}, x_{i2}, \cdots, c_i\right] = \mathrm{E}\left[\varepsilon_{it}|X_i, c_i\right] = 0$$

That is, the current disturbance is uncorrelated with the independent variables in every period, past, present, and future. The crucial aspect of the model concerns the heterogeneity. A particularly convenient assumption would be mean independence, that is, the unobserved variables are uncorrelated with the included variables:

$$\mathrm{E}\left[c_i|x_{i1}, x_{i2}, \cdots\right] = \alpha$$

If the missing variable(s) are uncorrelated with the included variables, then, as we shall see, they may be included in the disturbance of the model. This is the assumption that underlies the random effects model, as we will explore later. It is, however, a particularly strong assumption—it would be unlikely in the labor market and health care examples mentioned previously. The alternative would be

The unobserved variables are correlated with the included variables if

$$\mathrm{E}\left[c_i | x_{i1}, x_{i2}, \cdots\right] = h\left(x_{i1}, x_{i2}, \cdots\right) = h\left(X_i\right)$$

This formulation is more general, but at the same time, considerably more complicated, the more so since it may require yet further assumptions about the nature of the function.

6 Panel Data

6.1 Basic Idea for Panel Data

The basic framework of a regression model for panel data is

$$y_{it} = x_{it}'\beta + z_i'\alpha + \varepsilon_{it}$$

- Pooled regression: If z_i contains only a constant term;
- Fixed effects: If z_i is unobserved, but correlated with x_{it}, the model becomes $y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it}$;
- Random effects: If z_i is unobserved, and uncorrelated with x_{it}, the model becomes $y_{it} = x_{it}'\beta + \alpha + u_i + \varepsilon_{it}$;
- Random parameters: Coefficients vary randomly across individuals, i.e., $y_{it} = x_{it}(\beta + h_i) + (\alpha + u_i) + \varepsilon_{it}$.

We begin the analysis by assuming the simplest version of the model, the pooled model.

$$\begin{aligned} y_{it} =& \alpha + x_{it}'\beta + \varepsilon_{it}, \quad \text{for } i = 1, \cdots, n; t = 1, \cdots, T_i \\ & \text{with } \begin{cases} \mathrm{E}\left[\varepsilon_{it} | x_{i1}, x_{i2}, \cdots, x_{iT_i}\right] = 0 \\ \mathrm{Var}\left[\varepsilon | X\right] = \sigma^2 \mathbf{I} \end{cases} \end{aligned}$$

Remarks: In the panel data context, this is also called the population averaged model under the assumption that any latent heterogeneity has been averaged out.

In this form, if the remaining assumptions of the classical model are met (zero conditional mean of ε_{it}, homoskedasticity, independence across observations, i, and strict exogeneityof x_{it}), then no further analysis beyond the chapter of classical linear model is needed. Ordinary least squares is the efficient estimator and inference can reliably proceed along the lines developed before.

The crux of the panel data analysis in this chapter is that the assumptions underlying ordinary least squares estimation of the pooled model are unlikely to be met. The question, then, is what can be expected of the estimator when the heterogeneity does differ across individuals?

The fixed effects case is obvious. As we will examine later, omitting (or ignoring) the heterogeneity when the fixed effects model is appropriate renders the least squares estimator inconsistent—sometimes wildly so. In the random effects case, in which the

true model is $y_{it} = c_i + x'_{it}\beta + \varepsilon_{it}$, where $\mathrm{E}[c_i|X_i] = \alpha$, we can rewrite the model

$$\begin{aligned} y_{it} &= \alpha + x'_{it}\beta + \varepsilon_{it} \\ &= \alpha + x'_{it}\beta + \varepsilon_{it} + (c_i - \mathrm{E}[c_i|X_i]) \\ &= \alpha + x'_{it}\beta + \varepsilon_{it} + u_i \\ &= \alpha + x'_{it}\beta + w_{it} \end{aligned}$$

In this form, we can see that the unobserved heterogeneity induces autocorrelation, that is, $\mathrm{E}[w_{it}w_{is}] = \sigma_u^2 \neq 0$ when $t \neq s$.

6.1.1 Robust Covariance Matrix Estimation

Here we stack the T_i observations for individual i in a single equation:

$$y_i = X_i\beta + w_i$$

where β now includes the constant term.

In this setting, there may be heteroskedasticity across individuals. However, in a panel data set, the more substantive problem is crossobservation correlation, or autocorrelation. In a longitudinal data set, the group of observations may all pertain to the same individual, so any latent effects left out of the model will carry across all periods. Suppose, then, we assume that the disturbance vector consists of ε_{it} plus these omitted components. Then,

$$\mathrm{Var}[w_i|X_i] = \sigma_\varepsilon^2 \mathbf{I}_{T_i} + \Sigma_i = \Omega_i$$

The OLS estimator of β is

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= \left[\sum_{i=1}^{n} X'_i X_i\right]^{-1} \left[\sum_{i=1}^{n} X'_i y_i\right] \\ &= \left[\sum_{i=1}^{n} X'_i X_i\right]^{-1} \left[\sum_{i=1}^{n} X'_i (X_i\beta + w_i)\right] \\ &= \beta + \left[\sum_{i=1}^{n} X'_i X_i\right]^{-1} \sum_{i=1}^{n} X'_i w_i \end{aligned}$$

Consistency can be established along the lines developed as before. The true asymptotic

covariance matrix would take the form we saw for the generalized regression model:

$$\begin{aligned}\text{Asy.Var}\left[b\right] &= \frac{1}{n}\text{plim}\ \left[\frac{1}{n}\sum_{i=1}^{n}X_i'X_i\right]^{-1}\ \text{plim}\ \left[\sum_{i=1}^{n}X_i'w_iw_i'X_i\right]\ \text{plim}\ \left[\frac{1}{n}\sum_{i=1}^{n}X_i'X_i\right]^{-1}\\ &= \frac{1}{n}\text{plim}\ \left[\frac{1}{n}\sum_{i=1}^{n}X_i'X_i\right]^{-1}\ \text{plim}\ \left[\sum_{i=1}^{n}X_i'\Omega_iX_i\right]\ \text{plim}\ \left[\frac{1}{n}\sum_{i=1}^{n}X_i'X_i\right]^{-1}\end{aligned}$$

It is quite likely that the more important issue for appropriate estimation of the asymptotic covariance matrix is the correlation across observations, not heteroscedasticity. As such, it is quite likely that the White estimator is not the solution to the inference problem here.

6.1.2 Robust Estimation Using Group Means

The pooled regression model can be estimated using the sample means of the data. The implied regression model is obtained by premultiplying each group by $\left(\frac{1}{T}\right)\mathbf{1}'$ where $\mathbf{1}'$ is a row vector of 1s:

$$\begin{aligned}&\left(\frac{1}{T}\right)\mathbf{1}'y_i = \left(\frac{1}{T}\right)\mathbf{1}'X_i\beta + \left(\frac{1}{T}\right)\mathbf{1}'w_i\\ \iff\quad &\bar{y}_i = \bar{x}_i'\beta + \bar{w}_i\end{aligned}$$

In the transformed linear regression, the disturbances continue to have zero conditional means but heteroskedastic variances $\sigma_i^2 = \frac{1}{T^2}\mathbf{1}'\Omega_i\mathbf{1}$. With Ω_i unspecified, this is a heteroskedastic regression for which we would use the White estimator for appropriate inference.

The first difference model is,

$$\begin{aligned}\Delta y_{it} &= \Delta\alpha + \left(\Delta x_{it}\right)'\beta + \Delta\varepsilon_{it}\\ &= \Delta\alpha + \left(\Delta x_{it}\right)'\beta + \left(\varepsilon_{it} - \varepsilon_{i,t-1}\right)\\ &= \left(\Delta x_{it}\right)'\beta + u_{it}\end{aligned}$$

6.1.3 The Within- And Between-Groups Estimators

We can formulate the pooled regression model in three ways. First, the general regression is:

$$y_{it} = \alpha + x_{it}'\beta + \varepsilon_{it}$$

In terms of the group means

$$\bar{y}_i = \alpha + \bar{x}_i'\beta + \bar{\varepsilon}_i$$

While in terms of deviations from the group means:

$$(y_{it} - \bar{y}) = (x_{it} - \bar{x}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

We are assuming there are no time-invariant variables.

$$\begin{aligned}
S_{xx}^{Total} &= \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{\bar{x}}) (x_{it} - \bar{\bar{x}})' \\
S_{xy}^{Total} &= \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{\bar{x}}) (y_{it} - \bar{\bar{y}})' \\
S_{xx}^{Within} &= \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}) (x_{it} - \bar{x})' \\
S_{xy}^{Within} &= \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}) (y_{it} - \bar{y})' \\
S_{xx}^{Between} &= \sum_{i=1}^{n} T (\bar{x}_i - \bar{\bar{x}}) (\bar{x}_i - \bar{\bar{x}})' \\
S_{xy}^{Between} &= \sum_{i=1}^{n} T (\bar{x}_i - \bar{\bar{x}}) (\bar{y}_i - \bar{\bar{y}})'
\end{aligned}$$

It is easy to verify that

$$\begin{aligned}
S_{xx}^{Total} &= S_{xx}^{Within} + S_{xx}^{Between} \\
S_{xy}^{Total} &= S_{xy}^{Within} + S_{xy}^{Between}
\end{aligned}$$

Therefore, there are three possible least squares estimators of β corresponding to the decomposition. By FWL theorem, we see that the least squares estimator is

$$b^{Total} = (S_{xx}^{Total})^{-1} S_{xy}^{Total} = (S_{xx}^{Within} + S_{xx}^{Between})^{-1} (S_{xy}^{Within} + S_{xy}^{Between})$$

The within-groups estimator is

$$b^{Within} = (S_{xx}^{Within})^{-1} S_{xy}^{Within}$$

An alternative estimator would be the between-groups estimator

$$b^{Between} = (S_{xx}^{Between})^{-1} S_{xy}^{Between}$$

This is the group means estimator.

From the preceding expressions, we can rewrite as

$$\begin{cases} b^{Within} = \left(S_{xx}^{Within}\right)^{-1} S_{xy}^{Within} \\ b^{Between} = \left(S_{xx}^{Between}\right)^{-1} S_{xy}^{Between} \end{cases} \Longrightarrow \begin{cases} S_{xy}^{Within} = S_{xx}^{Within} b^{Within} \\ S_{xy}^{Between} = S_{xx}^{Between} b^{Between} \end{cases}$$

Inserting these into the expression of b^{Total}, we see that the least squares estimator is a matrix weighted average of the within- and between-groups estimators:

$$b^{Total} = F^{Within} b^{Within} + F^{Between} b^{Between}$$

where $F^{Within} = \left(S_{xx}^{Within} + S_{xx}^{Between}\right)^{-1} S_{xx}^{Within} = \mathbf{I} - F^{Between}$.

6.2 Fixed Effects Model

Recall that the general model is $y_{it} = x_{it}'\beta + c_i + \varepsilon_{it}$. The fixed effects model arises from the assumption that the omitted effects, c_i, are correlated with the included variables. In a general form, we assume $\mathrm{E}\left[c_i | X_i\right] = h\left(X_i\right)$.

The fixed effects model has the form

$$y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it}$$

Remarks: "Fixed effects" does NOT mean any variable is "fixed" in this context and random elsewhere. It means differences across groups can be captured in differences in the constant term. α_i is allowed to be correlated with x_{it}, and each α_i is treated as an unknown parameter to be estimated. As discussed, the coefficients on the time-invariant variables cannot be estimated.

For each group, i.e., for individual i, we have

$$y_i = X_i\beta + \mathbf{1}\alpha_i + \varepsilon_i$$

where y_i and X_i are the T observations for the i-th unit, $\mathbf{1}$ is a $T \times 1$ ones, and ε_i is the associated $T \times 1$ vector of disturbances.

For all the groups,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \beta + \begin{bmatrix} \mathbf{1} & 0 & \cdots & 0 \\ 0 & \mathbf{1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The model can be written as

$$y = \begin{bmatrix} X & d_1 & d_2 & \cdots & d_n \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \varepsilon$$

Let $D = \begin{bmatrix} d_1 & d_2 & \cdots & d_n \end{bmatrix}$, which is a $nT \times n$ matrix, the model becomes

$$y = X\beta + D\alpha + \varepsilon$$

which is called *Least Squares Dummy Variables (LSDV)* model.

6.2.1 Estimating LSDV Model

$$\begin{aligned}
D'D &= \begin{bmatrix} d_1' \\ d_2' \\ \vdots \\ d_n' \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \\
&= \begin{bmatrix} d_1'd_1 & d_1'd_2 & \cdots & d_1'd_n \\ d_2'd_1 & d_2'd_2 & \cdots & d_2'd_n \\ \vdots & \vdots & \ddots & \vdots \\ d_n'd_1 & d_n'd_2 & \cdots & d_n'd_n \end{bmatrix} \\
&= \begin{bmatrix} T & 0 & \cdots & 0 \\ 0 & T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T \end{bmatrix} \\
&= T\mathbf{I}_n
\end{aligned}$$

$$\begin{aligned}
D(D'D)^{-1}D' &= \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \left[T\mathbf{I}_n\right]^{-1} \begin{bmatrix} d_1' \\ d_2' \\ \vdots \\ d_n' \end{bmatrix} \\
&= \frac{1}{T} \begin{bmatrix} \mathbf{1} & 0 & \cdots & 0 \\ 0 & \mathbf{1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{1}' & 0 & \cdots & 0 \\ 0 & \mathbf{1}' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}' \end{bmatrix} \\
&= \frac{1}{T} \begin{bmatrix} \mathbf{1}\mathbf{1}' & 0 & \cdots & 0 \\ 0 & \mathbf{1}\mathbf{1}' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}\mathbf{1}' \end{bmatrix}
\end{aligned}$$

$$M_0 := \mathbf{I} - \frac{1}{T}\mathbf{1}\mathbf{1}'$$

$$M_0 z_i = \left(\mathbf{I} - \frac{1}{T}\mathbf{1}\mathbf{1}'\right) z_i = z_i - \mathbf{1}\left(\frac{1}{T}\left(\mathbf{1}' z_i\right)\right) = z_i - \mathbf{1}\left(\frac{1}{T}\sum_{i=1}^{T} z_i\right) = z_i - \bar{z}\mathbf{1}$$

$$M_D = \mathbf{I} - D(D'D)^{-1}D' = \begin{bmatrix} M_0 & 0 & \cdots & 0 \\ 0 & M_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_0 \end{bmatrix}$$

$$M_D X = \begin{bmatrix} M_0 & 0 & \cdots & 0 \\ 0 & M_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} M_0 x_1 \\ M_0 x_2 \\ \vdots \\ M_0 x_n \end{bmatrix} = \begin{bmatrix} x_1 - \bar{x}_1\mathbf{1} \\ x_2 - \bar{x}_2\mathbf{1} \\ \vdots \\ x_n - \bar{x}_n\mathbf{1} \end{bmatrix}$$

$$M_D y = \begin{bmatrix} M_0 & 0 & \cdots & 0 \\ 0 & M_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} M_0 y_1 \\ M_0 y_2 \\ \vdots \\ M_0 y_n \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y}_1\mathbf{1} \\ y_2 - \bar{y}_2\mathbf{1} \\ \vdots \\ y_n - \bar{y}_n\mathbf{1} \end{bmatrix}$$

$$b = \left[X' M_D X\right]^{-1}\left[X' M_D Y\right] = b^{Within}$$

$$\begin{aligned} y &= Xb + Da + e \\ D'y &= D'Xb + D'Da + D'e \\ \iff (D'D)\,a &= D'\left(y - Xb\right) \\ \alpha &= (D'D)^{-1}D\left(y - Xb\right) \end{aligned}$$

$$\begin{aligned} (D'D)^{-1}D'y &= (T\mathbf{I}_n)^{-1}\begin{bmatrix} d_1' \\ d_2' \\ \vdots \\ d_n' \end{bmatrix} y \\ &= (T\mathbf{I}_n)^{-1}\begin{bmatrix} \sum_{t=1}^{T} y_{1t} \\ \sum_{t=1}^{T} y_{2t} \\ \vdots \\ \sum_{t=1}^{T} y_{nt} \end{bmatrix} \\ &= (T\mathbf{I}_n)^{-1}\begin{bmatrix} T\bar{y}_1 \\ T\bar{y}_2 \\ \vdots \\ T\bar{y}_n \end{bmatrix} \\ &= \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_n \end{bmatrix} \end{aligned}$$

Similarly, it follows that

$$\begin{aligned}(D'D)^{-1}DXb &= (T\mathbf{I}_n)^{-1}\begin{bmatrix} d_1' \\ d_2' \\ \vdots \\ d_n' \end{bmatrix} Xb \\ &= (T\mathbf{I}_n)^{-1}\begin{bmatrix} \sum_{t=1}^{T} x_{1t}b \\ \sum_{t=1}^{T} x_{2t}b \\ \vdots \\ \sum_{t=1}^{T} x_{nt}b \end{bmatrix} \\ &= (T\mathbf{I}_n)^{-1}\begin{bmatrix} T\bar{x}_1 b \\ T\bar{x}_2 b \\ \vdots \\ T\bar{x}_n b \end{bmatrix} \\ &= \begin{bmatrix} \bar{x}_1 b \\ \bar{x}_2 b \\ \vdots \\ \bar{x}_n b \end{bmatrix}\end{aligned}$$

Jointly, so we have

$$\alpha_i = \bar{y}_i - \bar{x}_i' b$$

Covariance matrix for b is

$$\text{Est.Asy.Var}[b] = s^2 \left[X' M_D X\right]^{-1} = s^2 \left[S_{XX}^{Within}\right]^{-1}$$

$$\text{where } s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\beta - \alpha_i\right)}{nT - n - k}$$

6.2.2 Testing the Significance of Group Effects

Hypothesis

$$H_0 : \text{All } \alpha_i \text{ are equal.}$$
$$H_1 : \text{NOT all } \alpha_i \text{ are equal.}$$

Statistic:

$$F\left[n-1, nT-n-K\right] = \frac{\left(R_{LSDV}^2 - R_{Pooled}^2\right)/(n-1)}{\left(1 - R_{LSDV}^2\right)/(nT-n-K)}$$

$\ddot{X}$ can be treated as an instrument of X:

- Exogeneity: We have assumed that plim $\left(\frac{1}{nT}\right) X'\varepsilon = \mathbf{0}$, then we must have plim $\left(\frac{1}{nT}\right) X' M_D \varepsilon =$ plim $\left(\frac{1}{nT}\right) X' \left(M_D \varepsilon\right) = \mathbf{0}$. That is, if X is uncorrelated with ε, it will be uncorrelated with ε in deviations from its group means.
- Relevance:

6.3 Random Effects Model

The random effects model is

$$y_{it} = \alpha + x_{it}'\beta + u_t + \varepsilon_{it}$$

with assumptions

$$\begin{cases} \mathrm{E}\left[\varepsilon_{it}|X\right] = \mathrm{E}\left[u_i|X\right] = 0 \\ \mathrm{E}\left[\varepsilon_{it}^2|X\right] = \sigma_\varepsilon^2, \mathrm{E}\left[u_i^2|X\right] = \sigma_u^2 \\ \mathrm{E}\left[\varepsilon_{it}\varepsilon_{js}|X\right] = 0 \text{ for } t \neq s \text{ or } i \neq j \\ \mathrm{E}\left[u_i u_j\right] = 0 \text{ for } i \neq j \\ \mathrm{E}\left[\varepsilon_{it} u_j\right] = 0 \text{ for all } i, t, j \end{cases}$$

For each individual (or viewed as a group) i,

$$\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} \\ &= \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{1}_T \mathbf{1}_T' \end{aligned}$$

and for the nT observations,

$$\Omega = \begin{bmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} = \mathbf{I}_n \otimes \Sigma$$

Four models

$$\begin{cases} y_{it} = \alpha + x_{it}'\beta + u_i + \varepsilon_{it} \\ y_{it} - \bar{y}_i = \left(x_{it} - \bar{x}_i\right)'\beta + \varepsilon_{it} - \bar{\varepsilon}_i \\ y_{it} - y_{i,t-1} = \left(x_{it} - x_{i,t-1}\right)'\beta + \varepsilon_{it} - \varepsilon_{i,t-1} \\ \bar{y}_i = \alpha + \bar{x}_i'\beta + u_i + \bar{\varepsilon}_i \end{cases} \overset{\text{Est.Var}}{\Longrightarrow} \begin{cases} \sigma_\varepsilon^2 + \sigma_u^2 \\ \sigma_{\varepsilon^2}\left(1 - \frac{1}{T}\right) \\ 2\sigma_\varepsilon^2 \\ \frac{1}{T}\sigma_\varepsilon^2 + \sigma_u^2 \end{cases}$$

$$\Omega^{-1/2} = \left[\mathbf{I}_n \otimes \Sigma\right]^{-1/2} = \mathbf{I}_n \otimes \Sigma^{-1/2}$$

$$\text{with } \Sigma^{-1/2} = \frac{1}{\sigma_\varepsilon}\left[\mathbf{I}_T - \frac{\theta_i}{T}\mathbf{1}_T\mathbf{1}_T'\right]$$

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}} = 1 - \frac{1}{\sqrt{1 + T\left(\frac{\sigma_u}{\sigma_\varepsilon}\right)^2}}$$

The transformation of y_i and X_i for GLS is

$$\Sigma^{-1/2} y_i = \frac{1}{\sigma_\varepsilon} \begin{bmatrix} y_{i1} - \theta \bar{y}_i \\ y_{i2} - \theta \bar{y}_i \\ \vdots \\ y_{in} - \theta \bar{y}_i \end{bmatrix}$$

The GLS estimator is

$$b^{Total} = \hat{F}$$

$$\hat{\sigma}_\varepsilon^2 = s_{LSDV}^2 = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \left(e_{it} - \bar{e}_i \right)^2}{nT - n - K}$$

Lagrange multiplier test:

$$H_0 : \sigma_u^2 = 0$$
$$H_1 : \sigma_u^2 \neq 0$$

The test statistic is

$$LM = \frac{nT}{2\left(T-1\right)} \cdot \left[\frac{\sum_{i=1}^{n} \left[\sum_{t=1}^{T} e_{it}^2 \right]}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} \right]^2 \sim \chi$$

6.4 Hausman's Specification Test

Hausman's test is used to decide which model, fixed or random effects,

The model is

$$y_{it} = \alpha + x_{it}' \beta + u_i + \varepsilon_{it}$$

$H_0 : x_{it}$ is uncorrelated with u_i

$H_1 : x_{it}$ is correlated with u_i

7 Maximum Likelihood Estimation

7.1 The Likelihood Function

The probability density function (p.d.f. in short), for a random variable y, conditioned on a set of parameters θ, is denoted $f(y|\theta)$.

The jointly density, or likelihood function is

$$f(y_1, \cdots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y)$$

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

Note that we write the joint density as a function of the data conditioned on the parameters whereas when we form the likelihood function, we will write this function in reverse, as a function of the parameters, conditioned on the data. Though the two functions are the same, it is to be emphasized that the likelihood function is written in this fashion to highlight our interest in the parameters and the information about them that is contained in the observed data. However, it is understood that the likelihood function is not meant to represent a probability density for the parameters.

MLE estimation aims to choose θ to maximize log-likelihood function. The necessary condition is

$$\frac{\partial \ln L(\theta|y, X)}{\partial \theta} = 0$$

which is called the likelihood equation.

Definition (Identification) The parameter vector θ is identified (or estimable) if for any other parameter vector $\theta^* \neq \theta$, for some data y, $L(\theta^*|y) \neq L(\theta|y)$.

7.2 Properties of MLE

Commonly-used densities:

- Normal

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\cdot\left(\frac{y-\mu}{\sigma}\right)^2}$$

- Bernoulli

$$f(y) = p^y (1-p)^{1-y}$$

- Exponential

$$f(y) = \lambda e^{-\lambda y}$$

- Poisson

$$f(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Example (Normal Distribution Case) In sampling from a normal distribution with mean μ and variance σ^2, the MLE estimator are

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_n$$
$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

Example (OLS Revisited)

By assumption,

$$\begin{aligned} f(\varepsilon_i) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) \\ \Longrightarrow \ln f(\varepsilon_i) &= -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2}{2} - \frac{\varepsilon_i^2}{2\sigma^2} \\ \Longrightarrow \ln L(\beta, \sigma^2) &= -\frac{n\ln 2\pi}{2} - \frac{n\ln \sigma^2}{2} - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \end{aligned}$$

As we can see from the expression of $\ln L(\beta, \sigma^2)$, the determinant part is the same as minimizing least squares, $(y - X\beta)'(y - X\beta)$. By resorting to the first-order condition, we can get the same result:

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\hat{\sigma}^2 = \frac{\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)}{n}$$

Notice that the $\hat{\sigma}^2$ here is biased but consistent.

Under the regularity, the maximum likelihood estimator (MLE) has the following asymptotic properties:

- Consistency: plim $\widehat{\theta} = \theta_0$.
- Asymptotic normality: $\widehat{\theta} \overset{\text{a}}{\sim} N\left[\theta_0, \{I(\theta_0)\}^{-1}\right]$, where $I(\theta_0) = -\mathrm{E}_0\left[\dfrac{\partial^2 \ln L}{\partial\theta_0 \partial\theta_0'}\right]$.
- Asymptotic efficiency: $\widehat{\theta}$ is asymptotically efficient and achieves Cramer-Rao lower bound for consistent estimators.

An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

Notations

For each observations, we have log-density $\ln f(y_i|\theta)$. Denote

$$g_i = \frac{\partial \ln f(y_i|\theta)}{\partial\theta}$$
$$H_i = \frac{\partial^2 \ln f(y_i|\theta)}{\partial\theta\partial\theta'}$$

Then we have

$$g = \frac{\partial \ln L(\theta|y)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta)}{\partial\theta} = \sum_{i=1}^{n} g_i$$
$$H = \frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\theta)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} H_i$$

Remarks: H is called *Hessian Matrix*.

7.3 Likelihood Equation

Likelihood equation

$$\mathrm{E}_0\left[\frac{\partial \ln L(\theta_0)}{\partial\theta_0}\right] = \mathrm{E}_0[g_0] = 0$$

Proof

$$\begin{aligned}
\mathrm{E}_0\left[g_{i0}\right] &= \mathrm{E}_0\left[\frac{\partial \ln f\left(y_i|\theta_0\right)}{\partial \theta_0}\right] \\
&= \int \frac{\partial \ln f\left(y_i|\theta_0\right)}{\partial \theta_0} \cdot f\left(y_i|\theta_0\right) \mathrm{d}y_i \\
&= \int \frac{1}{f\left(y_i|\theta_0\right)} \cdot \frac{\partial f\left(y_i|\theta_0\right)}{\partial \theta_0} \cdot f\left(y_i|\theta_0\right) \mathrm{d}y_i \\
&= \int \frac{\partial f\left(y_i|\theta_0\right)}{\partial \theta_0} \mathrm{d}y_i \\
&= \frac{\partial}{\partial \theta_0} \int f\left(y_i|\theta_0\right) \mathrm{d}y_i \\
&= \frac{\partial}{\partial \theta_0}\left(1\right) = 0
\end{aligned}$$

By definition, we have

$$g_0 = \sum_{i=1}^{n} g_{i0} = \sum_{i=1}^{n} \frac{\partial \ln f\left(y_i|\theta_0\right)}{\partial \theta_0} = \frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0}$$

Since we have proved that $\mathrm{E}_0\left[g_{i0}\right] = 0$, it follows that

$$\mathrm{E}_0\left[\frac{\partial \ln L\left(\theta_0\right)}{\partial \theta_0}\right] = \mathrm{E}_0\left[g_0\right] = \mathrm{E}_0\left[\sum_{i=1}^{n} g_{i0}\right] = \sum_{i=1}^{n} \mathrm{E}_0\left[g_{i0}\right] = 0$$

7.4 Information Matrix Equality

Information Matrix Equality

$$\begin{aligned}
\mathrm{Var}\left[\frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0}\right] &= \mathrm{E}_0\left[\left(\frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0}\right)\left(\frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0'}\right)\right] \\
&= -\mathrm{E}_0\left[\frac{\partial^2 \ln L\left(\theta_0|y\right)}{\partial \theta_0 \partial \theta_0'}\right]
\end{aligned}$$

Or equivalently, in our earlier notations, $\mathrm{Var}\left[g_0\right] = -\mathrm{E}_0\left[H_0\right]$.

Proof We only show a proof of this equality in the scalar case. First we have

$$\begin{aligned}\frac{\partial^2 \ln f\left(y_i|\theta\right)}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left(\frac{\partial \ln f\left(y_i|\theta\right)}{\partial\theta}\right) \\ &= \frac{\partial}{\partial\theta}\left(\frac{1}{f_i}\cdot\frac{\partial f_i}{\partial\theta}\right) \\ &= \frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right) - \left(\frac{1}{f_i}\right)^2\left(\frac{\partial f_i}{\partial\theta}\right)^2 \\ &= \frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right) - \left(\frac{\partial \ln f_i}{\partial\theta}\right)^2\end{aligned}$$

Taking expectations on both sides, we have

$$\mathrm{E}\left[\frac{\partial^2 \ln f_i}{\partial\theta^2}\right] = \mathrm{E}\left[\frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right)\right] - \mathrm{E}\left[\left(\frac{\partial \ln f_i}{\partial\theta}\right)^2\right]$$

Compare such result with the equality we hope to prove, it suffices to show that

$$\begin{aligned}\mathrm{E}\left[\frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right)\right] &= \int \frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right)\cdot f_i \mathrm{d}y_i \\ &= \int \frac{\partial^2 f_i}{\partial\theta^2}\mathrm{d}y_i \\ &= \frac{\partial^2}{\partial\theta^2}\int f_i \mathrm{d}y_i \\ &= \frac{\partial^2}{\partial\theta^2}\left(1\right) = 0\end{aligned}$$

So we immediately have

$$\mathrm{E}\left[\frac{\partial^2 \ln f_i}{\partial\theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial \ln f_i}{\partial\theta}\right)^2\right]$$

Since by definition,

$$\begin{aligned}\frac{\partial \ln L\left(\theta|y\right)}{\partial\theta} &= \sum_{i=1}^{n}\frac{\partial \ln f\left(y_i|\theta\right)}{\partial\theta} \\ \frac{\partial^2 \ln L\left(\theta|y\right)}{\partial\theta^2} &= \sum_{i=1}^{n}\frac{\partial^2 \ln f\left(y_i|\theta\right)}{\partial\theta^2}\end{aligned}$$

So we can conclude that

$$\mathrm{E}\left[\frac{\partial^2 \ln L\left(\theta|y\right)}{\partial\theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial \ln L\left(\theta|y\right)}{\partial\theta}\right)^2\right]$$

7.5 Consistency of MLE

Consistency of MLE Let θ_0 be the true value of the parameter, $\hat{\theta}$ is the MLE, and θ be any other estimator in the set Θ. Then MLE $\hat{\theta}$ is consistent.

$$\text{plim}\ \hat{\theta} = \theta_0$$

Proof

Jensen's Inequality If $g(x)$ is a concave function, then $\mathrm{E}[g(x)] \leq g(\mathrm{E}[x])$. The inequality is held strict when $g(x)$ is strictly concave.

According the Jensen's inequality, we have

$$\mathrm{E}_0\left[\ln \frac{L(\theta|y)}{L(\theta_0|y)}\right] < \ln \mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right]$$

For the left-hand side,

$$\begin{aligned}
\mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right] &= \int \frac{L(\theta|y)}{L(\theta_0|y)} \cdot L(\theta_0|y)\,\mathrm{d}y \\
&= \int L(\theta|y)\,\mathrm{d}y \\
&= 1
\end{aligned}$$

By the inequality above, it follows that

$$\begin{aligned}
1 = \mathrm{E}_0\left[\ln \frac{L(\theta|y)}{L(\theta_0|y)}\right] &< \ln \mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right] \\
\Longrightarrow \mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right] &< 0 \\
\Longleftrightarrow \mathrm{E}_0\left[\ln L(\theta|y)\right] &< \mathrm{E}_0\left[\ln L(\theta_0|y)\right]
\end{aligned}$$

The intuition is that, the likelihood function gets the maximum when $\theta = \theta_0$.

Now consider the sample analogy,

$$\frac{1}{n}\left[\ln f(y|\theta) - \ln f(y|\theta_0)\right] = \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta) - \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta_0)$$

As sample mean is a consistent estimator of the expectation, we have

$$\begin{aligned}
\text{plim}\ \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta) &= \mathrm{E}_0\left[\frac{1}{n}\ln L(\theta|y)\right] \\
\text{plim}\ \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta_0) &= \mathrm{E}_0\left[\frac{1}{n}\ln L(\theta_0|y)\right]
\end{aligned}$$

It follows that

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta\right) - \text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right) < 0$$

According to the definition of $\hat{\theta}$ that maximizes the likelihood function, for any finite sample, we have

$$\frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\hat{\theta}\right) \geq \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$$

Hence, it must be that

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\hat{\theta}\right) \geq \text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$$

The two inequalities of plim $\frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\hat{\theta}\right)$ and plim $\frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$ can hold if and only if

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta\right) = \text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$$

which implies that plim $\hat{\theta} = \theta_0$.

7.6 Asymptotic Normality

Asymptotic Normality The MLE $\hat{\theta}$ has an asymptotic normal distribution,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\text{d}} N\left[0, \left\{-\text{E}_0\left[\frac{1}{n}H\left(\theta_0\right)\right]\right\}^{-1}\right]$$

So we have

$$\hat{\theta} \overset{\text{a}}{\sim} N\left[\theta_0, \left\{I\left(\theta_0\right)\right\}^{-1}\right], \text{ where } I\left(\theta_0\right) = -\text{E}_0\left[H\left(\theta_0\right)\right]$$

Proof

At the maximum likelihood estimator, the gradient of the log likelihood equals zero (which is by definition), so $g\left(\hat{\theta}\right) = 0$. We expand this set of equations in a Taylor series around the true parameters θ_0. We will use the mean value theorem to truncate the Taylor series for each element of $g\left(\hat{\theta}\right)$ at the second order (which is Lagrange theorem),

$$g\left(\hat{\theta}\right) = g\left(\theta_0\right) + H\left(\bar{\theta}\right)\cdot\left(\hat{\theta} - \theta_0\right) = 0$$

The K rows of the Hessian are each evaluated at a point $\bar{\theta}_k$ that is between $\hat{\theta}$ and θ_0. Specifically.

Because plim $\left(\hat{\theta}-\theta_0\right)=0$, plim $\left(\hat{\theta}-\bar{\theta}\right)=0$ as well. The second derivatives are continuous functions . Therefore, if the limiting distribution exists, then

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{\mathrm{d}} -\left[H\left(\theta_0\right)\right]^{-1}\left[\sqrt{n}g\left(\theta_0\right)\right]$$

By dividing $H\left(\theta_0\right)$ and $g\left(\theta_0\right)$ by n, we obtain

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{\mathrm{d}} -\left[\frac{1}{n}H\left(\theta_0\right)\right]^{-1}\left[\sqrt{n}\bar{g}\left(\theta_0\right)\right]$$

We may apply the Lindeberg-Levy CLT to $\sqrt{n}\bar{g}\left(\theta_0\right)$, because it is $\sqrt{n}$ times the mean of a random sample.

Example (Information Matrix for a Normal Distribution) For a normal distribution with mean μ and variance σ^2,

$$\ln L\left(\mu,\sigma^2\right)=-\frac{1}{2}\sum_{i=1}^{n}\left[\ln 2\pi+\ln\sigma^2+\frac{\left(y_i-\mu\right)^2}{\sigma^2}\right]$$

$$\Longrightarrow\begin{cases}\frac{\partial^2\ln L}{\partial\mu^2}=-\frac{n}{\sigma^2}\\ \frac{\partial^2\ln L}{\partial\left(\sigma^2\right)^2}=\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\\ \frac{\partial^2\ln L}{\partial\mu\partial\left(\sigma^2\right)}=-\frac{1}{\sigma^4}\sum_{i=1}^{n}\left(y_i-\mu\right)\end{cases}$$

For the asymptotic variance of the maximum likelihood estimator, we need the expectations of these derivatives.

The first is non-stochastic, and the third has expectation of 0, since $\mathrm{E}\left[y_i\right]=\mu$. That leaves the second, which we can verify has expectation of $-\frac{n}{2\sigma^4}$, because each of the n terms $\left(y_i-\mu\right)^2$ has expected value σ^2.

Collecting these in the information matrix, reversing the sign, and inverting the matrix gives the asymptotic covariance matrix for the maximum likelihood estimators,

$$\left\{-\mathrm{E}_0\left[H_0\right]\right\}^{-1}=\begin{bmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{bmatrix}$$

7.7 Discrete Choice Model

Let U_a and U_b represent an individual's utility of two choices:

$$U_a=w'\beta_1+z_a'\gamma_a+\varepsilon_a$$
$$U_b=w'\beta_b+z_b'\gamma_b+\varepsilon_b$$

If we denote by $Y = 1$ the consumer's choice of alternative a, we infer from $Y = 1$ that $U_a > U_b$.

$$\begin{aligned}\Pr[Y = 1|w, z_a, z_b] &= \Pr[U_a > U_b] \\ &= \Pr[w'(\beta_a - \beta_b) + z_a'\gamma_a - z_b'\gamma_b + (\varepsilon_a - \varepsilon_b)|w, z_a, z_b] \\ &= \Pr[X\beta + \varepsilon > 0|X]\end{aligned}$$

We model the net benefit of a choice as an variable y^* such taht

$$y^* = x'\beta + \varepsilon$$

where ε has mean zero and has either a standardized logistic or normal distribution.

However we do not observe y^* directly; instead, our observation is

$$\begin{cases} y = 1, \text{ if } y^* > 0 \\ y = 0, \text{ if } y^* \leq 0 \end{cases}$$

Then we have

$$\begin{aligned}\Pr[y = 1|x] &= \Pr[y^* > 0|x] \\ &= \Pr[x'\beta + \varepsilon > 0|x] \\ &= \Pr[\varepsilon > -x'\beta|x] \\ &= \Pr[\varepsilon < x'\beta|x] \\ &= F(x'\beta)\end{aligned}$$

Remarks:

- The assumptions of known variance is an innocent normalization.
  - Suppose the variance of ε is scaled by an unrestricted parameter σ^2. The latent regression will be

$$\begin{aligned} y^* &= x'\beta + \sigma\varepsilon \\ \Longrightarrow \left(\frac{y^*}{\sigma}\right) &= x'\left(\frac{\beta}{\sigma}\right) + \varepsilon \end{aligned}$$

  - The transformed model is the same model with the same data. The observed data will be unchanged; y is still 0 or 1, only depending on the sign of y^* instead of its scale.
  - This also means that there is no information about σ in the sample data so σ cannot be estimated.
- The assumption of zero cutoff is another innocent normalization.

– Let a be the supposed nonzero cutoff and α be the unknown constant term in the model, then

$$\begin{aligned}\Pr[y^* > a|x] &= \Pr[\alpha + x'\beta + \varepsilon > a|x] \\ &= \Pr[(\alpha - a) + x'\beta + \varepsilon > 0|x]\end{aligned}$$

– Since α is unknown, the difference $(\alpha - a)$ remains an unknown parameter. If the model contains a constant term, it is unchanged by the choice of the cutoff.

7.7.1 Model Setup

The binary outcome suggests a regression model

$$F(x'\beta) = x'\beta$$

with $\mathrm{E}[y|x] = \{0 \cdot [1 - F(x'\beta)]\} + [1 \cdot F(x'\beta)] = F(x'\beta)$.

This implies the linear probability model:

$$\begin{aligned}y &= \mathrm{E}[y|x] + (y - \mathrm{E}[y|x]) \\ &= x'\beta + \varepsilon\end{aligned}$$

- Shortcoming:
  - We cannot constrain $x'\beta$ to the $[0, 1]$ interval.
  - Heteroskedasticity: $\mathrm{Var}[\varepsilon_i|X] = (x_i'\beta)(1 - x_i'\beta)$, as $x'\beta + \varepsilon$ must equal 0 or 1, and ε equals either $x'\beta$ or $1 - x'\beta$, with probability $1 - F$ and F, respectively.
- Advantages:
  - Simplicity: The coefficient is easy to be interpreted.
  - Robustness: The assumptions of normality or logisticality are fragile while linearity is distribution free.

Generally, we want to construct a model which produces predictions consistent with the underlying theory

$$\Pr[y^* > 0|x] = \Pr[\varepsilon < x'\beta|x] = F(x'\beta)$$

and we expect that

$$\begin{aligned}\lim_{x'\beta \to +\infty} \Pr[Y = 1|x] &= \lim_{x'\beta \to +\infty} F(x'\beta) = 1 \\ \lim_{x'\beta \to -\infty} \Pr[Y = 1|x] &= \lim_{x'\beta \to -\infty} F(x'\beta) = 0\end{aligned}$$

The normal distribution has been commonly used, denoted as the probit model

$$\Pr\left[Y=1|x\right]=\int_{-\infty}^{x'\beta}\phi\left(t\right)\mathrm{d}t=\Phi\left(x'\beta\right)$$

Another commonly used model is the logit model, assuming logistic distribution:

$$\begin{aligned}\Pr\left[Y=1|x\right]&=\frac{1}{1+e^{-x'\beta}}=\Lambda\left(x'\beta\right)\\&=\left(\int_{-\infty}^{x'\beta}\Lambda\left(x'\beta\right)\left(1-\Lambda\left(x'\beta\right)\right)\mathrm{d}\left(x'\beta\right)\right)\end{aligned}$$

The probability model is

$$\mathrm{E}\left[y|x\right]=F\left(x'\beta\right)$$

The parameters of the model are not necessarily the marginal effects

$$\frac{\partial\mathrm{E}\left[y|x\right]}{\partial x}=\frac{\mathrm{d}F\left(x'\beta\right)}{\mathrm{d}\left(x'\beta\right)}\times\frac{\mathrm{d}\left(x'\beta\right)}{\mathrm{d}x}=f\left(x'\beta\right)\times\beta$$

For the probit model,

$$\frac{\partial\mathrm{E}\left[y|x\right]}{\partial x}=\phi\left(x'\beta\right)\times\beta$$

For the logit model,

$$\frac{\partial\mathrm{E}\left[y|x\right]}{\partial x}=\Lambda\left(x'\beta\right)\left[1-\Lambda\left(x'\beta\right)\right]\times\beta$$

The partial effects at the average (PEA)

$$PEA=\hat{\gamma}\left(\bar{x}\right)=f\left(\bar{x}'\hat{\beta}\right)\hat{\beta}$$

The average partial effects (APE):

$$APE=\hat{\bar{\gamma}}=\frac{1}{n}\sum_{i=1}^{n}f\left(x_i'\hat{\beta}\right)\hat{\beta}$$

For the logit model, the odds "in favor of $Y=1$" are

$$\begin{aligned}Odds&=\frac{\Pr\left(Y=1|x\right)}{\Pr\left(Y=0|x\right)}\\&=\exp\left(x'\beta\right)\end{aligned}$$

Consider the effect on the odds of the change of a dummy variable d (whose coefficient is δ),

$$\begin{aligned} OddsRatio &= \frac{Odds\left(x, d=1\right)}{Odds\left(x, d=0\right)} \\ &= \exp\left(\delta\right) \end{aligned}$$

For discrete choice model, the likelihood function is

$$\Pr\left[Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n\right] = \prod_{y_i=0}\left[1 - F\left(x_i'\beta\right)\right] \prod_{y_i=1} F\left(x_i'\beta\right)$$

It can be written as (which is the most important application):

$$\begin{aligned} L\left(\beta|\mathbf{y}\right) &= \prod_{i=1}^{n} F\left(x_i'\beta\right)^{y_i}\left[1 - F\left(x_i'\beta\right)\right]^{1-y_i} \\ \Longrightarrow \ln L &= \sum_{i=1}^{n}\left\{y_i \ln F\left(x_i'\beta\right) + \left(1 - y_i\right)\ln\left[1 - F\left(x_i'\beta\right)\right]\right\} \end{aligned}$$

The likelihood equations is then

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left[\frac{y_i f_i}{F_i} - \left(1 - y_i\right)\cdot\frac{f_i}{1 - F_i}\right]x_i = 0$$

Remarks: Take care when taking derivatives with respect to $\ln F\left(x_i'\beta\right)$: $\dfrac{\partial \ln F\left(x'\beta\right)}{\partial \beta} = \dfrac{1}{F\left(x'\beta\right)}\cdot f\left(x'\beta\right)\cdot x$.

The likelihood equation for a logit model:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left(y_i - \Lambda_i\right)x_i = 0$$

Likelihood Ratio Index:

$$LRI = 1 - \frac{\ln L}{\ln L_o}$$

7.7.2 Hypothesis Testing

If J is the number of restrictions,

- Likelihood ratio test

$$-2\ln\frac{\hat{L}_R}{\hat{L}_U} \sim \chi^2\left(J\right)$$

- Wald test

$$W = \left[c\left(\hat{\theta}\right) - q\right]' \left\{\text{Asy.Var}\left[c\left(\hat{\theta}\right) - q\right]\right\}^{-1} \left[c\left(\hat{\theta}\right) - q\right] \sim \chi^2\left(J\right)$$

- Lagrange multiplier test

$$LM = \left(\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \hat{\theta}_R}\right)' \left\{I\left(\hat{\theta}_R\right)^{-1}\right\} \left(\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \hat{\theta}_R}\right) \sim \chi^2\left(J\right)$$

8 Time Series Analysis

Time series data are *not* random samples (for example, GDPs in adjacent years are not independent, nor are they drawn from identical distributions). Indeed, a time series should be understood as a *single* occurrence of a random event.

Standard models/methods may not apply/work.

$$y_t = \beta x_t + u_t$$

- The random sampling assumption fails, almost for sure.
- There might be endogeneity problem: $\mathrm{E}\left[u_t | x_t\right] \neq 0$.
- The disturbances might be serially correlated.

Autoregression: $\mathrm{Cov}\left[y_t, y_s\right] \neq 0, \text{ for } t \neq s$.

Serial correlation: $\mathrm{Cov}\left[\varepsilon_t, \varepsilon_s\right] \neq 0, \text{ for } t \neq s$.

8.1 Stationarity and Ergodicity

Unconditional mean:

$$\mu_t = \mathrm{E}\left[y_t\right] = \text{plim}_{\,l\to\infty} \frac{1}{l} \sum_{i=1}^{l} y_t$$

Unconditional jth autocovariance ($j = 0, \pm 1, \pm 2, \cdots$):

$$\gamma_{jt} = \mathrm{E}\left[\left(y_t - \mu_t\right)\left(y_{t-j} - \mu_{t-j}\right)\right] = \text{plim}_{\,l\to\infty} \frac{1}{l} \sum_{i=1}^{l} \left[y_t^{(i)} - \mu_t\right]\left[y_{t-j}^{(i)} - \mu_{t-j}\right]$$

Covariance-stationary or weakly stationary: μ_t and γ_{jt} do not depend on date t:

$$\begin{aligned} \mu_t &= \mu, \forall t \\ \gamma_{jt} &= \gamma_j, \forall t, j \end{aligned}$$

White noise

$$\begin{cases} \mathrm{E}\left[\varepsilon_t\right] = 0 \\ \mathrm{E}\left[\varepsilon_t^2\right] = \mathrm{Var}\left[\varepsilon_t\right] = \sigma^2 \\ \mathrm{E}\left[\varepsilon_t \varepsilon_\tau\right] = \mathrm{Cov}\left[\varepsilon_t, \varepsilon_\tau\right] = 0, \forall t \neq \tau \end{cases}$$

Gaussian white noise

$$\varepsilon_t \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

8.2 Moving Average Process

MA(1) process:

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \text{ with } \varepsilon_t \sim \text{w.n.}\left(\sigma^2\right)$$

MA(1) is necessarily stationary. Moments are

$$\mathrm{E}\left[y_t\right] = \mu$$

$$\begin{cases} \gamma_0 = \sigma^2 + \theta^2\sigma^2 = \left(1+\theta^2\right)\sigma^2 \\ \gamma_1 = \mathrm{E}\left[\left(y_t - \mu\right)\left(y_{t-1} - \mu\right)\right] = \mathrm{E}\left[\left(\varepsilon_t + \theta\varepsilon_{t-1}\right)\left(\varepsilon_{t-1} + \theta\varepsilon_{t-2}\right)\right] = \theta\mathrm{E}\left[\varepsilon_{t-1}^2\right] = \theta\sigma^2 \\ \gamma_j = 0, \forall j \geq 2 \end{cases}$$

$$\begin{cases} \rho_0 = 1 \\ \rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta}{1+\theta^2} \\ \rho_j = 0, \forall j \geq 2 \end{cases}$$

MA(∞) process

$$y_t = \mu + \sum_{j=0}^{\infty}\theta_j\varepsilon_{t-j} = \mu + \theta_0\varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots, \text{ with } \varepsilon_t \sim \text{w.n.}\left(\sigma^2\right)$$

MA(∞) is not necessarily stationary. A sufficient condition for stationary: θ's are absolutely summable, i.e., $\sum_{j=0}^{\infty}\left|\theta_j\right| < \infty$.

Based on MA processes, it is straightforward to calculate the expected marginal effect of current innovation on future endogenous variable. Impulse response function (IRF):

$$IRF(j) = \frac{\partial \mathrm{E}_t\left[y_{t+j}\right]}{\partial\varepsilon_t} = \theta_j, \forall j = 0, 1, \cdots$$

8.3 Autoregressive (AR) Process

AR(1) process:

$$y_t = c + \phi y_{t-1} + \varepsilon_t, \text{ with } \varepsilon_t \sim \text{w.n.}\left(\sigma^2\right)$$

AR(1) is not necessarily stationary. Stationary condition: $|\phi| < 1$. If yes, moments are

$$\mathrm{E}[y_t] = c + \phi \mathrm{E}[y_{t-1}] \Longrightarrow \mathrm{E}[y_t] = \frac{c}{1-\phi}$$

$$\begin{cases} \gamma_0 = \mathrm{Var}[y_t] = \phi^2 \mathrm{Var}[y_{t-1}] + \sigma^2 = \phi^2 \gamma_0 + \sigma^2 \Longrightarrow \gamma_0 = \dfrac{\sigma^2}{1-\phi^2} \\ \gamma_1 = \mathrm{E}[(y_t - \mu)(y_{t-1} - \mu)] = \mathrm{E}[(\phi(y_{t-1} - \mu) + \varepsilon_t)(y_{t-1} - \mu)] = \phi\gamma_0 \\ \rho_1 = \dfrac{\gamma_1}{\gamma_0} = \dfrac{\phi\gamma_0}{\gamma_0} = \phi \end{cases}$$

$$\begin{cases} \gamma_2 = \mathrm{E}[(y_t - \mu)(y_{t-2} - \mu)] = \mathrm{E}[(\phi(y_{t-1} - \mu) + \varepsilon_t)(y_{t-2} - \mu)] = \phi\gamma_1 = \phi^2\gamma_0 \\ \rho_2 = \dfrac{\gamma_2}{\gamma_0} = \dfrac{\phi^2\gamma_0}{\gamma_0} = \phi^2 \end{cases}$$

Let $\mu = \mathrm{E}[y_t] = \dfrac{c}{1-\phi}$,

$$y_t = c + \phi y_{t-1} + \varepsilon_t \Longrightarrow y_t - \mu = \phi(y_{t-1} - \mu) + \varepsilon_t$$

AR(1) is a special case of AR(p) process:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \text{ with } \varepsilon_t \sim \text{w.n.}(\sigma^2)$$

AR(p) is not necessarily stationary. Stationary condition: all roots of the pth-order equation $1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$ lie outside unit circle, i.e., all have a modulus strictly greater than 1. If yes, moments are

$$\begin{cases} \mathrm{E}[y_t] = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p) \\ \gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma^2 \\ \gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p}, \forall j = 1, 2, \cdots \end{cases}$$

Unconditional and conditional moments of a stationary AR(1):

$$y_t = c + \phi y_{t-1} + \varepsilon_t, \text{ with } \varepsilon_t \sim \text{w.n.}(\sigma^2), |\phi| < 1$$

Unconditional moments:

$$\begin{cases} \mathrm{E}[y_t] = \dfrac{c}{1-\phi} \\ \gamma_j = \dfrac{\phi^j}{1-\phi^2}\sigma^2, \forall j = 0, 1, \cdots \\ \rho_j = \phi^j, \forall j = 0, 1, \cdots \end{cases}$$

Conditional moments:

$$\begin{aligned}
y_{t+j} &= c + \phi y_{t+j-1} + \varepsilon_{t+j} \\
&= c + \phi\left(c + \phi y_{t+j-2} + \varepsilon_{t+j-1}\right) + \varepsilon_{t+j} \\
&= (1+\phi)\, c + \phi^2\left(c + \phi y_{t+j-3} + \varepsilon_{t+j-2}\right) + \phi\varepsilon_{t+j-1} + \varepsilon_{t+j} \\
&= \left(1 + \phi + \cdots + \phi^{j-1}\right) c + \phi^j y_t + \left(\phi^{j-1}\varepsilon_{t+1} + \cdots + \varepsilon_{t+j}\right) \\
\Longrightarrow \mathrm{E}_t\left[y_{t+j}\right] &= \mathrm{E}_t\left[\left(1 + \phi + \cdots + \phi^{j-1}\right) c + \phi^j y_t\right] \\
&= \frac{1-\phi^j}{1-\phi} c + \phi^j y_t \\
&= \left(1-\phi^j\right)\mathrm{E}_t\left[y_t\right] + \phi^j y_t, \forall j = 0, 1, \cdots \\
\mathrm{Var}\left[y_{t+j}\right] &= \mathrm{Var}\left[\phi^{j-1}\varepsilon_{t+1} + \cdots + \varepsilon_{t+j}\right] \\
&= \left(\phi^{2(j-1)} + \phi^{2(j-2)} + \cdots + \phi^{2\cdot 1} + 1\right)\sigma^2 \\
&= \frac{1-\phi^{2j}}{1-\phi^2}\sigma^2 \\
&= \left(1-\phi^{2j}\right)\gamma_0, \forall j = 0, 1, \cdots
\end{aligned}$$

Remarks:

- $\mathrm{E}\left[y_{t+j}\right]$ is a weighted average of $\mathrm{E}_t\left[y_t\right]$ and y_t, converging to $\mathrm{E}_t\left[y_t\right]$ in the long run.
- $\mathrm{Var}_t\left[y_{t+j}\right]$ rises in j, converging to γ_0 in the long run.

8.4 Autoregressive Moving Average (ARMA) Process

ARMA(p, q) process:

$$\begin{aligned}
y_t =& c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \\
&+ \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}, \\
&\text{with } \varepsilon_t \sim \text{w.n.}\left(\sigma^2\right)
\end{aligned}$$

ARMA(p, q) is not necessarily stationary. Its stationarity relies entirely on AR coefficients $\left(\phi_1, \phi_2, \cdots, \phi_p\right)$, not on MA coefficients $\left(\theta_1, \theta_2, \cdots, \theta_q\right)$; indeed, ARMA$(p, q)$ and the corresponding AR(p) share the same stationarity condition: all roots of the pth-order equation $1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$ lie outside unit circle, i.e., all have a modulus strictly greater than 1. If yes, moments are

$$\begin{cases}
\mathrm{E}\left[y_t\right] = c / \left(1 - \phi_1 - \phi_2 - \cdots - \phi_p\right) \\
\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p}, \forall j > q \\
\gamma_j \text{ has no analytical expression for } j = 0, 1, \cdots q
\end{cases}$$

8.5 Lag Polynomial Representation

Lag operator is defined as

$$Ly_t = y_{t-1}, L^2 y_t = y_{t-2}, \cdots, L^p y_t = y_{t-p}$$
$$L\beta = \beta, L(\beta y_t) = \beta L y_t, L(y_{1t} \pm y_{2t}) = L y_{1t} \pm L y_{2t}, L^p L^q y_t = L^{p+q} y_t$$

MA(q) in a lag polynomial representation:

$$\begin{aligned} & y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \\ \Longrightarrow & y_t = \mu + \theta(L)\varepsilon_t, \text{ where } \theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q \end{aligned}$$

AR(p) in a lag polynomial representation:

$$\begin{aligned} & y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \\ \Longrightarrow & \phi(L) = c + \varepsilon_t, \text{ where } \phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p \end{aligned}$$

ARMA(p, q) in a lag polynomial representation:

$$\begin{aligned} y_t = & c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \\ & + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \\ \Longrightarrow \phi(L) y_t = & c + \theta(L)\varepsilon_t \end{aligned}$$

For a log polynomial, say $\phi(L)$, if all roots of $\phi(z) = 0$ lie outside unit circle, then there exists a lag polynomial with absolutely summable coefficients, denoted by $\phi(L)^{-1}$, such that $\phi(L)^{-1}\phi(L) y_t = y_t$. Example:

$$\begin{aligned} & \phi(L) = 1 - \phi L, |\phi| < 1 \\ \Longrightarrow & \phi(L)^{-1} = 1 + \phi L + \phi^2 L^2 + \cdots \end{aligned}$$

A stationary AR(p) has a stationary MA(∞) representation:

$$\begin{aligned} & \phi(L) y_t = c + \varepsilon_t \\ \Longrightarrow & y_t = \phi(L)^{-1} c + \phi(L)^{-1} \varepsilon_t \end{aligned}$$

Take AR(1) for example:

$$\begin{aligned} & y_t = c + \phi y_{t-1} + \varepsilon_t \iff (1 - \phi L) y_t = c + \varepsilon_t \\ \Longrightarrow & y_t = (1 - \phi L)^{-1} c + (1 - \phi L)^{-1} \varepsilon_t \\ & = (1 + \phi L + \phi^2 L^2 + \cdots) c + (1 + \phi L + \phi^2 L^2 + \cdots) \varepsilon_t \\ & = (1 + \phi + \phi^2 + \cdots) c + (1 + \phi L + \phi^2 L^2 + \cdots) \varepsilon_t \\ & = \frac{c}{1 - \phi} + \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} \cdots \end{aligned}$$

Take MA(1) as another example:

$$y_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1}, |\theta| < 1$$

Note that here we rewrite the term $\theta\varepsilon_{t-1}$ as a negative one, which convenient our computation without distorting its mathematical meaning.

$$\begin{aligned}
& y_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1}, |\theta| < 1 \\
\Longrightarrow & y_t - \mu = (1 - \theta L)\,\varepsilon_t \\
\Longrightarrow & (1 - \theta L)^{-1}\,(y_t - \mu) = \varepsilon_t \\
\Longrightarrow & \left(1 + \theta L + \theta^2 L^2 + \cdots\right) y_t - \frac{\mu}{1-\theta} = \varepsilon_t \\
\Longrightarrow & y_t = \frac{\mu}{1-\theta} + \varepsilon_t - \theta y_{t-1} - \theta^2 y_{t-2} - \cdots
\end{aligned}$$

which is an AR(∞).

8.6 Estimation of ARMA: MLEs

Case 1: Gaussian AR(1)

$$y_t = c + \phi y_{t-1} + \varepsilon_t, \varepsilon_t \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

This is a Markov process, the probability density conditioning on all histories is equivalent to conditioning on current history only. In order to obtain conditional distributions, we focus the "random" part. In AR(1), we turn to $\varepsilon_t = y_t - c - \phi y_{t-1}$, which follows a distribution of $N\left(0, \sigma^2\right)$.

$$\begin{aligned}
f\left(y_2|y_1\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_2 - c - \phi y_1\right)^2}{2\sigma^2}\right] \\
f\left(y_3|y_1, y_2\right) = f\left(y_3|y_2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_3 - c - \phi y_2\right)^2}{2\sigma^2}\right] \\
&\cdots \\
f\left(y_T|y_1, y_2, \cdots, y_T\right) = f\left(y_T|y_{T-1}\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_T - c - \phi y_{T-1}\right)^2}{2\sigma^2}\right]
\end{aligned}$$

The conditional joint density is:

$$f\left(y_2, y_3, \cdots, y_T|y_1\right) = \prod_{t=2}^{T} f\left(y_t|y_{t-1}\right) = \prod_{t=2}^{T} \left\{\exp\left[-\frac{\left(y_t - c - \phi y_{t-1}\right)^2}{2\sigma^2}\right]\right\}$$

The conditional log likelihood function is:

$$\mathcal{L}\left(c,\phi,\sigma^2\right)=-\frac{T-1}{2}\ln 2\pi-\frac{T-1}{2}\ln\sigma^2-\sum_{t=2}^{T}\frac{\left(y_t-c-\phi y_{t-1}\right)^2}{2\sigma^2}$$

Remarks:

- For AR processes, MLEs of coefficients are identical to OLS estimates; because OLS estimates are consistent, so are MLEs –this fact doesn't hinge on normality.
- For MA processes, MLEs have no analytical solution; MLEs must be found by numerical optimization.

8.7 Error Correction Model

ECM: It captures the idea that changes in a variable partially reflect adjustments of the variable to its deviations from a (long-run) equilibrium level/relationship.

Take AR(1) as an example.

$$\begin{aligned} y_t &= c+\phi y_{t-1}+\varepsilon_t \\ \mu &= \mathrm{E}\left[y_t\right]=\frac{c}{1-\phi} \\ \Longrightarrow \Delta y_t &= -\left(1-\phi\right)\left(y_{t-1}-\mu\right)+\varepsilon_t \end{aligned}$$

Example: An ADL model

$$y_t=\beta_0+\beta_1 y_{t-1}+\gamma_0 x_t+\gamma_1 x_{t-1}+\varepsilon_t, |\beta_1|<1$$

Long-run relationship (by assumption):

$$y^*=a+bx^*, \text{ where } y^*=\mathrm{E}\left[y_t\right], x^*=\mathrm{E}\left[x_t\right]$$

Easy to show that

$$a=\frac{\beta_0}{1-\beta_1}, b=\frac{\gamma_0+\gamma_1}{1-\beta_1}$$

ECM representation:

$$\Delta y_t=\gamma_0\Delta x_t-\left(1-\beta_1\right)\left(y_{t-1}-a-bx_{t-1}\right)+\varepsilon_t$$

8.8 Multivariate Processes

8.8.1 Vector Autoregressions (VAR): Stationarity

VAR(p) process:

$$\vec{y}_t = \vec{c} + \Phi_1 \vec{y}_{t-1} + \Phi_2 \vec{y}_{t-2} + \cdots + \Phi_p \vec{y}_{t-p} + \vec{\varepsilon}_t, \text{ with } \vec{\varepsilon} \sim \text{w.n.}(\Sigma)$$

Lag polynomial representation of VAR(p):

$$\Phi(L)\,\vec{y}_t = \vec{c} + \vec{\varepsilon}_t, \text{ where } \Phi(L) = \mathbf{I}_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p$$

VAR(p) is not necessarily stationary. Stationarity condition: all roots of the npth-order equation $\det[\Phi(z)] = \det\left(\mathbf{I}_n - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p\right) = 0$ lie outside unit circle, i.e., all have a modulus strictly greater than 1.

VAR(1) process:

$$\vec{y}_t = \vec{c} + \Phi_1 \vec{y}_{t-1} + \vec{\varepsilon}_t, \text{ with } \varepsilon_t \sim \text{w.n.}(\sigma^2)$$

Why VAR?

- Better fit; and can do a better forecasting job than early macro models.
- Solutions to DSGEs can have a VAR representation.

VAR: Estimation

- The idea is pretty much the same as in the univariate case.
- For coefficients, conditional MLE is equivalent to OLS estimates equation by equation.