CHAPTER 3 CLASSICAL LINEAR REGRESSION MODELS

Key words: Classical linear regression, Conditional heteroskedasticity, Conditional homoskedasticity, F-test, GLS, Hypothesis testing, Model selection criterion, OLS, R^2, t-test

Abstract: In this chapter, we will introduce the classical linear regression theory, including the classical model assumptions, the statistical properties of the OLS estimator, the t-test and the F-test, as well as the GLS estimator and related statistical procedures. This chapter will serve as a starting point from which we will develop the modern econometric theory.

3.1 Assumptions

Suppose we have an observed random sample $\{Z_t\}_{t=1}^n$ of size n, where $Z_t = (Y_t, X_t')'$, Y_t is a scalar, $X_t = (1, X_{1t}, X_{2t}, ..., X_{kt})'$ is a $(k+1) \times 1$ vector, t is an index (either cross-sectional unit or time period) for observations, and n is the sample size. We are interested in modelling the conditional mean $E(Y_t|X_t)$ using an observed realization (i.e., a data set) of the random sample $\{Y_t, X_t'\}', t = 1, ..., n$.

Notations:

Throughtout this book, we set $K \equiv k + 1$, the number of regressors for which there are k economic variables and an intercept. The index t may denote an individual unit (e.g., a firm, a household, a country) for cross-sectional data, or denote a time period (e.g., day, weak, month, year) in a time series context.

We first list and discuss the assumptions of the classical linear regression theory.

Assumption 3.1 [Linearity]:

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

where β^o, is a $K \times 1$ unknown parameter vector, and ε_t is an unobservable disturbance.

Remarks:

In Assumption 3.1, Y_t is the dependent variable (or regressand), X_t is the vector of regressors (or independent variables, or explanatory variables), and β^o is the regression

coefficient vector. When the linear model is correctly specified for the conditional mean $E(Y_t|X_t)$, i.e., when $E(\varepsilon_t|X_t) = 0$, the parameter $\beta^o = \frac{\partial}{\partial X_t}E(Y_t|X_t)$ can be interpreted as the marginal effect of X_t on Y_t.

The key notion of *linearity* in the classical linear regression model is that the regression model is linear in β^o rather than in X_t.In other words, linear regression models cover some models for Y_t which have a nonlinear relationship with X_t.

Question: Does Assumption 3.1 imply a causal relationship from X_t to Y_t?

Not necessarily. As Kendall and Stuart (1961, Vol.2, Ch. 26, p.279) point out, "a statistical relationship, however strong and however suggestive, can never establish causal connection. Our ideas of causation must come from outside statistics ultimately, from some theory or other." Assumption 3.1 only implies a predictive relationship: Given X_t, can we predict Y_t linearly?

Denote

$$\begin{aligned} Y &= (Y_1, ..., Y_n)', \qquad n \times 1, \\ \varepsilon &= (\varepsilon_1, ..., \varepsilon_n)', \qquad n \times 1, \\ \mathbf{X} &= (X_1, ..., X_n)', \qquad n \times K. \end{aligned}$$

where the t-th row of $\mathbf{X}$ is $X_t' = (1, X_{1t}, ..., X_{kt})$. With these matrix notations, we have a compact expression for Assumption 3.1:

$$\begin{aligned} Y &= \mathbf{X}\beta^o + \varepsilon, \\ n \times 1 &= (n \times K)(K \times 1) + n \times 1. \end{aligned}$$

The second assumption is a strict exogeneity condition.

Assumption 3.2 [Strict Exogeneity]:

$$E(\varepsilon_t|\mathbf{X}) = E(\varepsilon_t|X_1, ..., X_t, ..., X_n) = 0, \qquad t = 1, ..., n.$$

Remarks:

Among other things, Assumption 3.2 implies correct model specification for $E(Y_t|X_t)$. This is because Assumption 3.2 implies $E(\varepsilon_t|X_t) = 0$ by conditional expectation. It also implies $E(\varepsilon_t) = 0$ by the law of iterated expectations.

Under Assumption 3.2, we have $E(X_s\varepsilon_t) = 0$ for any (t, s), where $t, s \in \{1, ..., n\}$. This follows because

$$\begin{aligned} E(X_s\varepsilon_t) &= E[E(X_s\varepsilon_t|\mathbf{X})] \\ &= E[X_s E(\varepsilon_t|\mathbf{X})] \\ &= E(X_s \cdot 0) \\ &= 0. \end{aligned}$$

Note that (i) and (ii) imply $\text{cov}(X_s, \varepsilon_t) = 0$ for all $t, s \in \{1, ..., n\}$.

Because $\mathbf{X}$ contains regressors $\{X_s\}$ for both $s \leq t$ and $s > t$, Assumption 3.2 essentially requires that the error ε_t do not depend on the past and future values of regressors if t is a time index. This rules out dynamic time series models for which ε_t may be correlated with the future values of regressors (because the future values of regressors depend on the current shocks), as is illustrated in the following example.

Example 1: Consider a so-called AutoRegressive AR(1) model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t, \qquad t = 1, ..., n, \\ &= X_t'\beta + \varepsilon_t, \\ \{\varepsilon_t\} &\sim \text{i.i.d.}(0, \sigma^2), \end{aligned}$$

where $X_t = (1, Y_{t-1})'$. This is a dynamic regression model because the term $\beta_1 Y_{t-1}$ represents the "memory" or "feedback" of the past into the present value of the process, which induces a correlation between Y_t and the past. The term autoregression refers to the regression of Y_t on its own past values. The parameter β_1 determines the amount of feedback, with a large absolute value of β_1 resulting in more feedback. The disturbance ε_t can be viewed as representing the effect of "new information" that is revealed at time t. Information that is truly new cannot be anticipated so that the effects of today's new information should be unrelated to the effects of yesterday's news in the sense that $E(\varepsilon_t|X_t) = 0$. Here, we make a stronger assumption that we can model the effect of new information as an i.i.d.$(0, \sigma^2)$ sequence.

Obviously, $E(X_t\varepsilon_t) = E(X_t)E(\varepsilon_t) = 0$ but $E(X_{t+1}\varepsilon_t) \neq 0$. Thus, we have $E(\varepsilon_t|\mathbf{X}) \neq 0$, and so Assumption 3.2 does not hold. Here, the lagged dependent variable Y_{t-1} in the regressor vector X_t is called a predetermined variable, since it is orthogonal to ε_t but depends on the past history of $\{\varepsilon_t\}$.

In Chapter 5 later, we will consider linear regression models with dependent observations, which will include this example as a special case. In fact, the main reason of imposing Assumption 3.2 is to obtain a finite sample distribution theory. For a large

sample theory (i.e., an asymptotic theory), the strict exogeneity condition will not be needed.

In econometrics, there are some alternative definitions of strict exogeneity. For example, one definition assumes that ε_t and $\mathbf{X}$ are independent. An example is that $\mathbf{X}$ is nonstochastic. This rules out conditional heteroskedasticity (i.e., $\text{var}(\varepsilon_t|\mathbf{X})$ depends on $\mathbf{X}$). In Assumption 3.2, we still allow for conditional heteroskedasticity, because we do not assume that ε_t and $\mathbf{X}$ are independent. We only assume that the conditional mean $E(\varepsilon_t|\mathbf{X})$ does not depend on $\mathbf{X}$.

Question: What happens to Assumption 3.2 if $\mathbf{X}$ is nonstochastic?

If $\mathbf{X}$ is nonstochastic, Assumption 3.2 becomes

$$E(\varepsilon_t|\mathbf{X}) = E(\varepsilon_t) = 0.$$

An example of nonstochastic $\mathbf{X}$ is $X_t = (1, t, ..., t^k)'$. This corresponds to a time-trend regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \sum_{j=0}^{k} \beta_j^o t^j + \varepsilon_t. \end{aligned}$$

Question: What happens to Assumption 3.2 if $Z_t = (Y_t, X_t')'$ is an independent random sample (i.e., Z_t and Z_s are independent whenever $t \neq s$, although Y_t and X_t may not be independent)?

When $\{Z_t\}$ is i.i.d., Assumption 3.2 becomes

$$\begin{aligned} E(\varepsilon_t|\mathbf{X}) &= E(\varepsilon_t|X_1, X_2, ...X_t, ..., X_n) \\ &= E(\varepsilon_t|X_t) \\ &= 0. \end{aligned}$$

In other words, when $\{Z_t\}$ is i.i.d., $E(\varepsilon_t|\mathbf{X}) = 0$ is equivalent to $E(\varepsilon_t|X_t) = 0$.

Assumption 3.3 [Nonsingularity]: *(a) The minimum eigenvalue of the $K \times K$ square matrix $X'X = \sum_{t=1}^n X_t X_t'$ is nonsingular, and (b)*

$$\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty \text{ as } n \to \infty$$

with probability one.

Remarks:

Assumption 3.3(a) rules out multicollinearity among the $(k+1)$ regressors in X_t. We say that there exists multicollinearity (sometimes called the exact or perfect multicolinearity in the literature) among the X_t if for all $t \in \{1, ..., n\}$, the variable X_{jt} for some $j \in \{0, 1, ..., k\}$ is a linear combination of the other $K-1$ column variables $\{X_{it}, i \neq j\}$. In this case, the matrix $\mathbf{X}'\mathbf{X}$ is singular, and as a consequence, the true model parameter β^o in Assumption 3.1 is not identifiable.

The nonsingularity of $\mathbf{X}'\mathbf{X}$ implies that $\mathbf{X}$ must be of full rank of $K = k+1$. Thus, we need $K \leq n$. That is, the number of regressors cannot be larger than the sample size. This is a necessary condition for identification of parameter β^o.

The eigenvalue λ of a square matrix A is characterized by the system of linear equations:

$$\det(A - \lambda I) = 0,$$

where $\det(\cdot)$ denotes the determinant of a square matrix, and I is an identity matrix with the same dimension as A.

It is well-known that the eigenvalue λ can be used to summarize information contained in a matrix (recall the popular principal component analysis). Assumption 3.3 implies that new information must be available as the sample size $n \to \infty$ (i.e., X_t should not only have same repeated values as t increases).

Intuitively, if there are no variations in the values of the X_t, it will be difficult to determine the relationship between Y_t and X_t (indeed, the purpose of classical linear regression is to investigate how a change in $\mathbf{X}$ causes a change in Y). In certain sense, one may call $\mathbf{X}'\mathbf{X}$ the "information matrix" of the random sample $\mathbf{X}$ because it is a measure of the information contained in $\mathbf{X}$. The magnitude of $\mathbf{X}'\mathbf{X}$ will affect the preciseness of parameter estimation for β^o. Indeed, as will be shown below, the condition that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ ensures that variance of the OLS estimator will vanish to zero as $n \to \infty$. This rule out a possibility called near-multicolinearity that there exists an approximate linear relationship among the sample values of explanatory variables in X_t such that although $\mathbf{X}'\mathbf{X}$ is nonsingular, its minimum eigenvalue $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n. When $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with n, the OLS estimator is well-defined and has a well-behaved finite sample distribution, but its variance never vanishes to zero as $n \to \infty$. In other words, in the near multicolinearity case where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with n, the OLS estimator will never converges to the true parameter value β^o, although it will still have a well-defined finite sample distribution.

Question: Why can the eigenvalue λ be used as a measure of the information contained in $\mathbf{X}'\mathbf{X}$?

Assumption 3.4 [Spherical error variance]:

(a) [conditional homoskedasticity]:

$$E(\varepsilon_t^2|\mathbf{X}) = \sigma^2 > 0, \qquad t = 1, ..., n;$$

(b) [conditional non-autocorrelation]:

$$E(\varepsilon_t \varepsilon_s|\mathbf{X}) = 0, \qquad t \neq s, t, s \in \{1, ..., n\}.$$

Remarks:

We can write Assumption 3.4 as

$$E(\varepsilon_t \varepsilon_s|\mathbf{X}) = \sigma^2 \delta_{ts},$$

where $\delta_{ts} = 1$ if $t = s$ and $\delta_{ts} = 0$ otherwise. In mathematics, δ_{ts} is called the Kronecker delta function. Under this assumption, we have

$$\begin{aligned} var(\varepsilon_t|\mathbf{X}) &= E(\varepsilon_t^2|\mathbf{X}) - [E(\varepsilon_t|\mathbf{X})]^2 \\ &= E(\varepsilon_t^2|\mathbf{X}) \\ &= \sigma^2 \end{aligned}$$

and

$$\begin{aligned} cov(\varepsilon_t, \varepsilon_s|\mathbf{X}) &= E(\varepsilon_t \varepsilon_s|\mathbf{X}) \\ &= 0 \text{ for all } t \neq s. \end{aligned}$$

By the law of iterated expectations, Assumption 3.4(b) implies that $\text{var}(\varepsilon_t) = \sigma^2$ for all $t = 1, ..., n$, the so-called unconditional homoskedasticity. Similarly, Assumption 3.4(a) implies $\text{cov}(\varepsilon_t, \varepsilon_s) = 0$ for all $t \neq s$. Thus, there exists no serial correlation between ε_t and its lagged values when t is an index for time, or there exists no spatial correlation between the disburbances associated with different cross-sectional units when t is an index for the cross-sectional unit (e.g., consumer, firm, household, etc).

Assumption 3.4 does not imply that ε_t and $\mathbf{X}$ are independent. It allows the possibility that the conditional higher order moments (e.g., skewness and kurtosis) of ε_t depend on $\mathbf{X}$.

We can write Assumptions 3.2 and 3.4 compactly as follows:

$$E(\varepsilon|\mathbf{X}) = 0 \text{ and } E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 I,$$

where $I \equiv I_n$ is a $n \times n$ identity matrix.

3.2 OLS Estimation

Question: How to estimate β^o using an observed data set generated from the random sample $\{Z_t\}_{t=1}^n$, where $Z_t = (Y_t, X_t')'$?

Definition [OLS estimator]: *Suppose Assumptions 3.1 and 3.3(a) hold. Define the sum of squared residuals (SSR) of the linear regression model $Y_t = X_t'\beta + u_t$ as*

$$\begin{aligned} SSR(\beta) &\equiv (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) \\ &= \sum_{t=1}^{n} (Y_t - X_t'\beta)^2. \end{aligned}$$

Then the Ordinary Least Squares (OLS) estimator $\hat{\beta}$ is the solution to

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^K} SSR(\beta).$$

Note that $SSR(\beta)$ is the sum of squred model errors $\{u_t = Y_t - X_t'\beta\}$, with equal weighting for each t.

Theorem 1 [Existence of OLS]: *Under Assumptions 3.1 and 3.3, the OLS estimator $\hat{\beta}$ exists and*

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= \left(\frac{1}{n} \sum_{t=1}^{n} X_t X_t' \right)^{-1} \frac{1}{n} \sum_{t=1}^{n} X_t Y_t. \end{aligned}$$

The last expression will be useful for our asymptotic analysis in subsequent chapters.

Proof: Using the formula that for an $K \times 1$ vector A and and $K \times 1$ vector β, the derivative

$$\frac{\partial (A'\beta)}{\partial \beta} = A,$$

we have

$$\begin{aligned}
\frac{dSSR(\beta)}{d\beta} &= \frac{d}{d\beta}\sum_{t=1}^{n}(Y_t - X_t'\beta)^2 \\
&= \sum_{t=1}^{n}\frac{\partial}{\partial\beta}(Y_t - X_t'\beta)^2 \\
&= \sum_{t=1}^{n}2(Y_t - X_t'\beta)\frac{\partial}{\partial\beta}(Y_t - X_t'\beta) \\
&= -2\sum_{t=1}^{n}X_t(Y_t - X_t'\beta) \\
&= -2\mathbf{X}'(Y - \mathbf{X}\beta).
\end{aligned}$$

The OLS must satisfy the FOC:

$$\begin{aligned}
-2\mathbf{X}'(Y - \mathbf{X}\hat{\beta}) &= 0, \\
\mathbf{X}'(Y - \mathbf{X}\hat{\beta}) &= 0, \\
\mathbf{X}'Y - (\mathbf{X}'\mathbf{X})\hat{\beta} &= 0.
\end{aligned}$$

It follows that

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'Y.$$

By Assumption 3.3, $\mathbf{X}'\mathbf{X}$ is nonsingular. Thus,

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

Checking the SOC, we have the $K \times K$ Hessian matrix

$$\begin{aligned}
\frac{\partial^2 SSR(\beta)}{\partial\beta\partial\beta'} &= -2\sum_{t=1}^{n}\frac{\partial}{\partial\beta'}\left[(Y_t - X_t'\beta)X_t\right] \\
&= 2\mathbf{X}'\mathbf{X} \\
&\sim \text{positive definite}
\end{aligned}$$

given $\lambda_{\min}(\mathbf{X}'\mathbf{X}) > 0$. Thus, $\hat{\beta}$ is a global minimizer. Note that for the existence of $\hat{\beta}$, we only need that $\mathbf{X}'\mathbf{X}$ is nonsingular, which is implied by the condition that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ but it does not require that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$. This completes the proof.

Remarks:

Suppose $Z_t = \{Y_t, X_t'\}', t = 1, ..., n$, is an independent and identically distributed (i.i.d.) random sample of size n. Consider the sum of squared residual scaled by n^{-1} :

$$\frac{SSR(\beta)}{n} = \frac{1}{n}\sum_{t=1}^{n}(Y_t - X_t'\beta)^2$$

and its minimizer

$$\hat{\beta} = \left(\frac{1}{n}\sum_{t=1}^{n} X_t X_t'\right)^{-1} \frac{1}{n}\sum_{t=1}^{n} X_t Y_t.$$

These are the sample analogs of the population MSE criterion

$$MSE(\beta) = E(Y_t - X_t'\beta)^2$$

and its minimizer

$$\beta^* \equiv [E(X_t X_t')]^{-1} E(X_t Y_t).$$

That is, $SSR(\beta)$, after scaled by n^{-1}, is the sample analogue of $MSE(\beta)$, and the OLS $\hat{\beta}$ is the sample analogue of the best LS approximation coefficient β^*.

Put $\hat{Y}_t \equiv X_t'\hat{\beta}$. This is called the fitted value (or predicted value) for observation Y_t, and $e_t \equiv Y_t - \hat{Y}_t$ is the estimated residual (or prediction error) for observation Y_t. Note that

$$\begin{aligned} e_t &= Y_t - \hat{Y}_t \\ &= (X_t'\beta^o + \varepsilon_t) - X_t'\hat{\beta} \\ &= \varepsilon_t - X_t'(\hat{\beta} - \beta^o), \end{aligned}$$

where ε_t is the unavoidable true disturbance ε_t, and $X_t'(\hat{\beta} - \beta^o)$ is an estimation error $X_t'(\hat{\beta} - \beta^o)$, which is made smaller when a larger data set is available (so $\hat{\beta}$ becomes closer to β^o).

The FOC implies that the estimated residual $e = Y - \mathbf{X}\hat{\beta}$ is orthogonal to regressors $\mathbf{X}$ in the sense that

$$\mathbf{X}'e = \sum_{t=1}^{n} X_t e_t = 0.$$

This is the consequence of the very nature of OLS, as implied by the FOC of $\min_{\beta \in R^K} SSR(\beta)$. It always holds no matter whether $E(\varepsilon_t|\mathbf{X}) = 0$ (recall that we do not impose Assumption 3.2 in the Theorem above). Note that if X_t contains the intercept, then $\mathbf{X}'e = 0$ implies $\Sigma_{t=1}^{n} e_t = 0$.

Some useful identities

To investigate the statistical properties of $\hat{\beta}$, we first state some useful lemmas.

Lemma: *Under Assumptions 3.1 and 3.3(a), we have:*

(i)

$$\mathbf{X}'e = 0;$$

(ii)

$$\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon;$$

(iii) Define a $n \times n$ projection matrix

$$P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and

$$M = I_n - P.$$

Then both P and M are symmetric (i.e., $P = P'$ and $M = M'$) and idempotent (i.e., $P^2 = P, M^2 = M$), with

$$\begin{aligned} P\mathbf{X} &= \mathbf{X}, \\ M\mathbf{X} &= 0. \end{aligned}$$

(iv)

$$SSR(\hat{\beta}) = e'e = Y'MY = \varepsilon'M\varepsilon.$$

Proof: (i) The result follows immediately by the FOC of the OLS estimator.

(ii) Because $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ and $Y = \mathbf{X}\beta^o + \varepsilon$, we have

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

(iii) P is idempotent because

$$\begin{aligned} P^2 &= PP \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= P. \end{aligned}$$

Similarly we can show $M^2 = M$.

(iv) By the definition of M, we have

$$\begin{aligned} e &= Y - \mathbf{X}\hat{\beta} \\ &= Y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= [I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']Y \\ &= MY \\ &= M(\mathbf{X}\beta^o + \varepsilon) \\ &= M\mathbf{X}\beta^o + M\varepsilon \\ &= M\varepsilon \end{aligned}$$

given $M\mathbf{X} = 0$. It follows that

$$\begin{aligned} SSR(\hat{\beta}) &= e'e \\ &= (M\varepsilon)'(M\varepsilon) \\ &= \varepsilon' M^2 \varepsilon \\ &= \varepsilon' M \varepsilon, \end{aligned}$$

where the last equality follows from $M^2 = M$.

3.3 Goodness of fit and model selection criterion

Question: How well does the linear regression model fit the data? That is, how well does the linear regression model explain the variation of the observed data of $\{Y_t\}_{t=1}^n$?

We need some criteria or some measures to characterize goodness of fit.

We first introduce two measures for goodness of fit. The first measure is called the uncentered squared multi-correlation coefficient R^2

Definition [Uncentered R^2] : *The uncentered squared multi-correlation coefficient is defined as*

$$R^2_{uc} = \frac{\hat{Y}'\hat{Y}}{Y'Y} = 1 - \frac{e'e}{Y'Y},$$

where the second equality follows from the first order condition of the OLS estimation.

Remarks:

The measure R^2_{uc} has a nice interpretation: The proportion of the uncentered sample quadratic variation in the dependent variables $\{Y_t\}$ that can be attributed to the uncentered sample quadratic variation of the predicted values $\{\hat{Y}_t\}$. Note that we always have $0 \leq R^2_{uc} \leq 1$.

Next, we define a closely related measure called Centered R^2.

Definition [Centered R^2 : Coefficient of Determination]: *The coefficient of determination*

$$R^2 \equiv 1 - \frac{\sum_{t=1}^n e_t^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2},$$

where $\bar{Y} = n^{-1} \sum_{t=1}^n Y_t$ *is the sample mean.*

Remarks:

When X_t contains the intercept, we have the following orthogonal decomposition:

$$\begin{aligned}
\sum_{t=1}^n (Y_t - \bar{Y})^2 &= \sum_{t=1}^n (\hat{Y}_t - \bar{Y} + Y_t - \hat{Y}_t)^2 \\
&= \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n e_t^2 \\
&\quad +2 \sum_{t=1}^n (\hat{Y}_t - \bar{Y}) e_t \\
&= \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n e_t^2,
\end{aligned}$$

where the cross-product term

$$\begin{aligned}
\sum_{t=1}^n (\hat{Y}_t - \bar{Y}) e_t &= \sum_{t=1}^n \hat{Y}_t e_t - \bar{Y} \sum_{t=1}^n e_t \\
&= \hat{\beta}' \sum_{t=1}^n X_t e_t - \bar{Y} \sum_{t=1}^n e_t \\
&= \hat{\beta}' (\mathbf{X}' e) - \bar{Y} \sum_{t=1}^n e_t \\
&= \hat{\beta}' \cdot 0 - \bar{Y} \cdot 0 \\
&= 0,
\end{aligned}$$

where we have made use of the facts that $\mathbf{X}'e = 0$ and $\sum_{t=1}^n e_t = 0$ from the FOC of the OLS estimation and the fact that X_t contains the intercept (i.e., $X_{0t} = 1$). It follows

that

$$\begin{aligned} R^2 &\equiv 1 - \frac{e'e}{\sum_{t=1}^n (Y_t - \bar{Y})^2} \\ &= \frac{\sum_{t=1}^n (Y_t - \bar{Y})^2 - \sum_{t=1}^n e_t^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2} \\ &= \frac{\sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2}. \end{aligned}$$

and consequently we have

$$0 \leq R^2 \leq 1.$$

Question: Can R^2 be negative?

Yes, it is possible! If X_t does not contain the intercept, then the orthogonal decomposition identity

$$\sum_{t=1}^n (Y_t - \bar{Y})^2 = \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n e_t^2$$

no longer holds. As a consequence, R^2 may be negative when there is no intercept! This is because the cross-product term

$$2\sum_{t=1}^n (\hat{Y}_t - \bar{Y}) e_t$$

may be negative.

When X_t contains an intercept, the centered R^2 has a simlar interpretation to the uncentered R_{uc}^2. That is, R^2 measures the proportion of the sample variance of $\{Y_t\}_{t=1}^n$ that can be explained by the linear predictor of X_t.

Example [Capital Asset Pricing Model (CAPM)] The classical CAPM is characterized by the equation

$$r_{pt} - r_{ft} = \alpha_p + \beta_p (r_{mt} - r_{ft}) + \varepsilon_{pt}, \qquad t = 1, ..., n,$$

where r_{pt} is the return on portfolio (or asset) p, r_{ft} is the return on a risk-free asset, and r_{mt} is the return on the market portfolio. Here, $r_{pt} - r_{ft}$ is the risk premium of portfolio p, $r_{mt} - r_{ft}$ is the risk premium of the market portfolio, which is the only systematic market risk factor, and ε_{pt} is the individual-specific risk which can be eliminated by diversification if the ε_{pt} are uncorrelated across different assets. In this model, R^2 has an interesting economic interpretation: It is the proportion of the risk of portfolio

p (as measured by the sample variance of its risk premium $r_{pt} - r_{ft}$) that is attributed to the market risk factor $(r_{mt} - r_{ft})$. In contrast, $1 - R^2$ is the proportion of the risk of portfolio p that is contributed by individual-specific risk factor ε_{pt}.

For any given random sample $\{Y_t, X_t'\}', t = 1, ..., n$, R^2 is nondecreasing in the number of explanatory variables X_t. In other words, the more explanatory variables are added in the linear regression, the higher is R^2. This is always true no matter whether X_t has any true explanatory power for Y_t.

Theorem: *Suppose* $\{Y_t, X_{1,t}, ..., X_{k+q,t}\}', t = 1, ..., n$, *is a random sample, and Assumptions 3.1 and 3.3(a) hold. Let* R_1^2 *be the centered* R^2 *from the linear regression*

$$Y_t = X_t'\beta + u_t,$$

where $X_t = (1, X_{1t}, ..., X_{kt})'$, *and* β *is a* $K \times 1$ *parameter vector; also,* R_2^2 *is the centered* R^2 *from the extended linear regression*

$$Y_t = \tilde{X}_t'\gamma + v_t,$$

where $\tilde{X}_t = (1, X_{1t}, ..., X_{kt}, X_{k+1,t}, ..., X_{k+q,t})'$,*and* γ *is a* $(K + q) \times 1$ *parameter vector. Then* $R_2^2 \geq R_1^2$.

Proof: By definition, we have

$$\begin{aligned} R_1^2 &= 1 - \frac{e'e}{\sum_{t=1}^n (Y_t - \bar{Y})^2}, \\ R_2^2 &= 1 - \frac{\tilde{e}'\tilde{e}}{\sum_{t=1}^n (Y_t - \bar{Y})^2}, \end{aligned}$$

where e is the estimated residual vector from the regression of Y on $\mathbf{X}$, and $\tilde{e}$ is the estimated residual vector from the regression of Y on $\tilde{\mathbf{X}}$. It suffices to show $\tilde{e}'\tilde{e} \leq e'e$. Because the OLS estimator $\hat{\gamma} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Y$ minimizes $SSR(\gamma)$ for the extended model, we have

$$\tilde{e}'\tilde{e} = \sum_{t=1}^n (Y_t - \tilde{X}_t'\hat{\gamma})^2 \leq \sum_{t=1}^n (Y_t - \tilde{X}_t'\gamma)^2 \text{ for all } \gamma \in \mathbb{R}^{K+q}.$$

Now we choose

$$\gamma = (\hat{\beta}', 0')',$$

where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ is the OLS from the first regression. It follows that

$$\begin{aligned}\tilde{e}'\tilde{e} &\leq \sum_{t=1}^{n}\left(Y_t - \sum_{j=0}^{k}\hat{\beta}_j X_{jt} - \sum_{j=k+1}^{k+q} 0 \cdot X_{jt}\right)^2 \\ &= \sum_{t=1}^{n}(Y_t - X_t'\hat{\beta})^2 \\ &= e'e.\end{aligned}$$

Hence, we have $R_1^2 \leq R_2^2$. This completes the proof.

Question: What is the implication of this theorem?

The measure R^2 can be used to compare models with the same number of predictors, but it is not a useful criterion for comparing models of different sizes because it is biased in favor of large models.

The measure R^2 is not a suitable criterion for correct model specification. It is a measure for sampling variation rather than a measure of population. A high value of R^2 does not necessarily imply correct model specification, and correct model specification also does not necessarily imply a high value of R^2.

Strictly speaking, R^2 is a measure merely of aasociation with nothing to say about causality. High values of R^2 are often very easy to achieve when dealing with economic time series data, even when the causal link between two variables is extremely teneous or perhaps nonexistent. For example, in the spurious regressions where the dependent variable Y_t and the regressors X_t have no causal relationship but they diaplay similar trending behaviors over time, it is often found that R^2 is close to unity.

Finally, R^2 is a measure of the strength of linear association between the dependent variable Y_t and the regressors X_t (see Exercise 3.2). It is not a suitable measure for goodness of fit of a nonlinear regression model where $E(Y_t|X_t)$ is a nonlinear function of X_t.

Question: How to interpret R^2 for the linear regression model

$$\ln Y_t = \beta_0 + \beta_1 \ln L_t + \beta_2 \ln K_t + \varepsilon_t,$$

where Y_t is output, L_t is labor and K_t is capital?

Answer: R^2 is the proprotion of the total sample variations in $\ln Y_t$ that can be attributed to the sample variations in $\ln L_t$ and $\ln K_t$. It is not the proportion of the

sample quadratic variation in Y_t that can be attributed to the sample variations of L_t and K_t.

Question: Does a high R^2 value imply a precise estimation for β^o?

Two popular model selection criteria

Often, a large number of potential predictors are available, but we do not necessarily want to include all of them. There are two conflicting factors to consider: On one hand, a larger model has less systematic bias and it would give the best predictions if all parameters could be estimated without error. On the other hand, when unknown parameters are replaced by estimates, the prediction becomes less accurate, and this effect is worse when there are more parameters to estimate. An important idea in statistics is to use a simple model to capture essential information contained in data as much as possible. This is often called the KISS principle, namely "Keep It Sophistically Simple"!

Below, we introduce two popular model selection criteria that reflect such an idea.

Akaike Information Criterion [AIC]:

A linear regression model can be selected by minimizing the following AIC criterion with a suitable choice of K :

$$\begin{aligned} AIC &= \ln(s^2) + \frac{2K}{n} \\ &\sim \text{goodness of fit + model complexity} \end{aligned}$$

where

$$s^2 = e'e/(n-K),$$

is called the residual varianc estimator for $E(\varepsilon_t^2) = \sigma^2$and $K = k + 1$ is the number of regressors. AIC is proposed by Akaike (1973).

Bayesian Information Criterion [BIC, Schwarz (1978)]:

A linear regression model can be selected by minimizing the following criterion with a suitable choice of K :

$$BIC = \ln(s^2) + \frac{K\ln(n)}{n}.$$

This is called the Baysian information criterion (BIC), proposed by Schwarz (1978).

Both AIC and BIC try to trade off a goodness of fit to data measured by $\ln(s^2)$ with the desire to use as few paramerers as possible. When $\ln n \geq 2$, which is the case when $n > 7$, BIC gives a heavier penalty for model complexity than AIC, which is measured by the number of estimated parameters (relative to the sample size n). As a consequence, BIC will choose a more parsimonious linear regression model than AIC.

The difference between AIC and BIC is due to the way they are constructed. AIC is deisgned to select a model that will predict best and is less concerned than BIC with having a few too many parameters. BIC is designed to select the true value of K excatly. Under certain regularity conditions, BIC is strongly consistent in the sense that it determines the true model asymptotically (i.e., as $n \to \infty$), whereas for AIC an overparameterized model will emerge no matter how large the sample is. Of course, such properties are not necessarily guaranteed in finite samples. In practice, te best AIC model is usually close to the best BIC model and often they deliver the same model.

In addition to AIC and BIC, there are other criteria such as $\bar{R}^2$, the so-called adjusted R^2 that can also be used to select a linear regression model. The adjusted $\bar{R}^2$ is defined as

$$\bar{R}^2 = 1 - \frac{e'e/(n-K)}{(Y-\bar{Y})'(Y-\bar{Y})/(n-1)}.$$

This differs from

$$R^2 = 1 - \frac{e'e}{(Y-\bar{Y})'(Y-\bar{Y})}.$$

In $\bar{R}^2$, the adjustmet is made according to the degrees of freedom, or the number of explanatory variables in X_t. It may be shown that

$$\bar{R}^2 = 1 - \left[\frac{n-1}{n-K}(1-R^2)\right].$$

we note that $\bar{R}^2$ may take a negative value although there is an intercept in X_t.

All model criteria are structured in terms of the estimated residual variance $\hat{\sigma}^2$ plus a penalty adjustment involving the number of estimated parameters, and it is in the extent of this penalty that the criteria differ from. For more discussion about these, and other selection criteria, see Judge *et al.* (1985, Section 7.5).

Question: Why is it not a good practice to use a complicated model?

A complicated model contains many unknown parameters. Given a fixed amount of data information, parameter estimation will become less precise if more parameters have to be estimated. As a consequence, the out-of-sample forecast for Y_t may become less precise

than the forecast of a simpler model. The latter may have a larger bias but more precise parameter estimates. Intuitively, a complicated model is too flexible in the sense that it may not only capture systematic components but also some features in the data which will not show up again. Thus, it cannot forecast futures well.

3.4 Consistency and Efficiency of OLS

We now investigate the statistical properties of $\hat{\beta}$. We are interested in addressing the following basic questions:

- Is $\hat{\beta}$ a good estimator for β^o(consistency)?
- Is $\hat{\beta}$ the best estimator (efficiency)?
- What is the sampling distribution of $\hat{\beta}$ (normality)?

Question: What is the sampling distribution of $\hat{\beta}$?

The distribution of $\hat{\beta}$ is called the sampling distribution of $\hat{\beta}$, because $\hat{\beta}$ is a function of the random sample $\{Z_t\}_{t=1}^n$, where $Z_t = (Y_t, X_t')'$.

The sampling distribution of $\hat{\beta}$ is useful for any statistical inference involving $\hat{\beta}$, such as confidence interval estimation and hypothesis testing.

We first investigate the statistical properties of $\hat{\beta}$.

Theorem: *Suppose Assumptions 3.1-3.3(a) and 3.4 hold. Then*
(i) [Unbiasedness] $E(\hat{\beta}|\mathbf{X}) = \beta^o$ *and* $E(\hat{\beta}) = \beta^o$.
(ii) [Vanishing Variance]

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= E\left[(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})'|\mathbf{X}\right] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

If in addition Assumption 3.3(b) holds, then for any $K \times 1$ *vector* τ *such that* $\tau'\tau = 1$, *we have*

$$\tau' \text{var}(\hat{\beta}|\mathbf{X})\tau \to 0 \text{ as } n \to \infty.$$

(iii) [Orthogonality between e *and* $\hat{\beta}$*]*

$$\text{cov}(\hat{\beta}, e|\mathbf{X}) = E\{[\hat{\beta} - E(\hat{\beta}|\mathbf{X})]e'|\mathbf{X}\} = 0.$$

(iv) [Gauss-Markov]

$$\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X}) \text{ is positive semi-definite (p.s.d.)}$$

for any unbiased estimator $\hat{b}$ that is linear in Y with $E(\hat{b}|\mathbf{X}) = \beta^o$.

(v) [Residual variance estimator]

$$s^2 = e'e/(n-K) = \frac{1}{n-K}\sum_{t=1}^{n} e_t^2$$

is unbiased for $\sigma^2 = E(\varepsilon_t^2)$. That is, $E(s^2|\mathbf{X}) = \sigma^2$.

Proof: (i) Given $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$, we have

$$\begin{aligned} E[(\hat{\beta} - \beta^o)|\mathbf{X}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'0 \\ &= 0. \end{aligned}$$

(ii) Given $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$ and $E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 I$, we have

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

Note that Assumption 3.4 is crucial here to obtain the expression of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ for $\text{var}(\hat{\beta}|\mathbf{X})$. Moreover, for any $\tau \in \mathbb{R}^K$ such that $\tau'\tau = 1$, we have

$$\begin{aligned} \tau'\text{var}(\hat{\beta}|\mathbf{X})\tau &= \sigma^2\tau'(\mathbf{X}'\mathbf{X})^{-1}\tau \\ &\leq \sigma^2\lambda_{\max}[(\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2\lambda_{\min}^{-1}(\mathbf{X}'\mathbf{X}) \\ &\to 0 \end{aligned}$$

given $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ with probability one. Note that the condition that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ ensures that $\text{var}(\hat{\beta}|\mathbf{X})$ vanishes to zero as $n \to \infty$.

(iii) Given $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$, $e = Y - \mathbf{X}\hat{\beta} = MY = M\varepsilon$ (since $M\mathbf{X} = 0$), and $E(e) = 0$, we have

$$\begin{aligned}
\text{cov}(\hat{\beta}, e|\mathbf{X}) &= E\left[(\hat{\beta} - E\hat{\beta})(e - Ee)'|\mathbf{X}\right] \\
&= E\left[(\hat{\beta} - \beta^o)e'|\mathbf{X}\right] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon' M|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon'|\mathbf{X})M \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 IM \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'M \\
&= 0.
\end{aligned}$$

Again, Assumption 3.4 plays a crucial role in ensuring zero correlation between $\hat{\beta}$ and e.

(iv) Consider a linear estimator

$$\hat{b} = C'Y,$$

where $C = C(\mathbf{X})$ is a $n \times K$ matrix depending on $\mathbf{X}$. It is unbiased for β^o regardless of the value of β^o if and only if

$$\begin{aligned}
E(\hat{b}|\mathbf{X}) &= C'\mathbf{X}\beta^o + C'E(\varepsilon|\mathbf{X}) \\
&= C'\mathbf{X}\beta^o \\
&= \beta^o.
\end{aligned}$$

This follows if and only if

$$C'\mathbf{X} = I.$$

Because

$$\begin{aligned}
\hat{b} &= C'Y \\
&= C'(\mathbf{X}\beta^o + \varepsilon) \\
&= C'\mathbf{X}\beta^o + C'\varepsilon \\
&= \beta^o + C'\varepsilon,
\end{aligned}$$

the variance of $\hat{b}$

$$\begin{aligned}
\text{var}(\hat{b}) &= E\left[(\hat{b} - \beta^o)(\hat{b} - \beta^o)'|\mathbf{X}\right] \\
&= E\left[C'\varepsilon\varepsilon' C|\mathbf{X}\right] \\
&= C'E(\varepsilon\varepsilon'|\mathbf{X})C \\
&= C'\sigma^2 IC \\
&= \sigma^2 C'C.
\end{aligned}$$

Using $C'\mathbf{X} = I$, we now have

$$\begin{aligned}
\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X}) &= \sigma^2 C'C - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2[C'C - C'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'C] \\
&= \sigma^2 C'[I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']C \\
&= \sigma^2 C'MC \\
&= \sigma^2 C'MMC \\
&= \sigma^2 C'M'MC \\
&= \sigma^2 (MC)'(MC) \\
&= \sigma^2 D'D \\
&= \sigma^2 \sum_{t=1}^{n} D_t D_t' \\
&\sim \text{p.s.d.}
\end{aligned}$$

where we have used the fact that for any real-valued matrix D, the squared matrix $D'D$ is always p.s.d. [**Question:** How to show this?]

(v) Now we show $E[e'e/(n-K)] = \sigma^2$. Because $e'e = \varepsilon' M \varepsilon$ and $\text{tr}(AB) = \text{tr}(BA)$, we have

$$\begin{aligned}
E(e'e|\mathbf{X}) &= E(\varepsilon' M \varepsilon|\mathbf{X}) \\
&= E[\text{tr}(\varepsilon' M \varepsilon)|\mathbf{X}] \\
[\text{putting } A &= \varepsilon' M, B = \varepsilon] \\
&= E[\text{tr}(\varepsilon\varepsilon' M)|\mathbf{X}] \\
&= \text{tr}[E(\varepsilon\varepsilon'|\mathbf{X})M] \\
&= \text{tr}(\sigma^2 I M) \\
&= \sigma^2 \text{tr}(M) \\
&= \sigma^2(n-K)
\end{aligned}$$

where

$$\begin{aligned}
\text{tr}(M) &= \text{tr}(I_n) - \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \text{tr}(I_n) - \text{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\
&= n - K,
\end{aligned}$$

using $\text{tr}(AB) = \text{tr}(BA)$ again. It follows that

$$\begin{aligned} E(s^2|\mathbf{X}) &= \frac{E(e'e|\mathbf{X})}{n-K} \\ &= \frac{\sigma^2(n-K)}{(n-K)} \\ &= \sigma^2. \end{aligned}$$

This completes the proof.

Remarks:

Both Theorems (i) and (ii) imply that the conditional MSE

$$\begin{aligned} MSE(\hat{\beta}|\mathbf{X}) &= E[(\hat{\beta}-\beta^o)(\hat{\beta}-\beta^o)'|\mathbf{X}] \\ &= \text{var}(\hat{\beta}|\mathbf{X}) + \text{Bias}(\hat{\beta}|\mathbf{X})\text{Bias}(\hat{\beta}|\mathbf{X})' \\ &= \text{var}(\hat{\beta}|\mathbf{X}) \\ &\to 0 \text{ as } n \to \infty, \end{aligned}$$

where we have used the fact that

$$Bias(\hat{\beta}|\mathbf{X}) \equiv E(\hat{\beta}|\mathbf{X}) - \beta^o = 0.$$

Recall that MSE measures how close an estimator $\hat{\beta}$ is to the target parameter β^o.

Theorem (iv) implies that $\hat{\beta}$ is the best linear unbiased estimator (BLUE) for β^o because $\text{var}(\hat{\beta}|\mathbf{X})$ is the smallest among all unbiased linear estimators for β^o.

Formally, we can define a related concept for comparing two unbiased estimators:

Definition [Efficiency]: *An unbiased estimator $\hat{\beta}$ of parameter β^o is more efficient than another unbiased estimator $\hat{b}$ of parameter β^o if*

$$\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X}) \text{ is p.s.d.}$$

When $\hat{\beta}$ is more efficient than $\hat{b}$, we have that for any $\tau \in \mathbb{R}^K$ such that $\tau'\tau = 1$,

$$\tau' \left[\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X})\right] \tau \geq 0.$$

Choosing $\tau = (1, 0, ..., 0)'$, for example, we have

$$\text{var}(\hat{b}_1) - \text{var}(\hat{\beta}_1) \geq 0.$$

We note that the OLS estimator $\hat{\beta}$ is still BLUE even when there exists near-multicolinearity, where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n, and $\text{var}(\hat{\beta}|\mathbf{X})$ does not vanish to zero as $n \to \infty$. Near-multicolinearity is essentially a sample or data problem which we cannot remedy or improve upon when the objective is to estimate the unknown parameter β^o.

3.5 The Sampling Distribution of $\hat{\beta}$

To obtain the finite sample sampling distribution of $\hat{\beta}$, we impose the normality assumption on ε.

Assumption 3.5: $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I)$.

Remarks:

Assumption 3.5 implies both Assumptions 3.2 ($E(\varepsilon|\mathbf{X}) = 0$) and 3.4 ($E(\varepsilon\varepsilon|\mathbf{X}) = \sigma^2 I$). Moreover, under Assumption 3.5, the conditional *pdf* of ε given $\mathbf{X}$ is

$$f(\varepsilon|\mathbf{X}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{\varepsilon'\varepsilon}{2\sigma^2}\right) = f(\varepsilon),$$

which does not depend on $\mathbf{X}$, so the disturbance ε is independent of $\mathbf{X}$. Thus, every conditional moment of ε given $\mathbf{X}$ does not depend on $\mathbf{X}$.

The normal distribution is also called the Gaussian distribution named after the German mathematician and astronomer Carl F. Gauss. It is assumed here so that we can drive the finite sample distributions of $\hat{\beta}$ and related statistics, i.e., the distributions of $\hat{\beta}$ and related statistics when the sample size n is a finite integer. This assumption may be reasonable for observations that are computed as the averages of the outcomes of many repeated experiments, due to the effect of the so-called central limit theorem (CLT). This may occur in physics, for example. In economics, the normality assumption may not always be reasonable. For example, many high-frequency financial time series usually display heavy tails (with kurtosis larger than 3).

Question: What is the sampling distribution of $\hat{\beta}$?

We write

$$\begin{aligned}
\hat{\beta} - \beta^o &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
&= (\mathbf{X}'\mathbf{X})^{-1}\sum_{t=1}^{n} X_t\varepsilon_t \\
&= \sum_{t=1}^{n} C_t\varepsilon_t,
\end{aligned}$$

where the weighting vector

$$C_t = (\mathbf{X}'\mathbf{X})^{-1}X_t$$

is called the leverage of observation X_t.

Theorem [Normality of $\hat{\beta}$]: *Under Assumptions 3.1, 3.3(a) and 3.5,*

$$(\hat{\beta} - \beta^o)|\mathbf{X} \sim N[0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

Proof: Conditional on $\mathbf{X}$, $\hat{\beta} - \beta^o$ is a weighted sum of independent normal random variables $\{\varepsilon_t\}$, and so is also normally distributed.

We note that the OLS estimator $\hat{\beta}$ still has the conditional finite sample normal distribution $N(\beta^o, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ even when there exists near-multicolinearity, where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n and $\text{var}(\hat{\beta}|\mathbf{X})$ does not vanish to zero as $n \to \infty$.

The corollary below follows immediately.

Corollary [Normality of $R(\hat{\beta} - \beta^o)$]: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then for any nonstochastic $J \times K$ matrix R, we have*

$$R(\hat{\beta} - \beta^o)|\mathbf{X} \sim N[0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'].$$

Proof: Conditional on $\mathbf{X}$, $\hat{\beta} - \beta^o$ is normally distributed. Therefore, conditional on $\mathbf{X}$, the linear combination $R(\hat{\beta} - \beta^o)$ is also normally distributed, with

$$E[R(\hat{\beta} - \beta^o)|\mathbf{X}] = RE[(\hat{\beta} - \beta^o)|\mathbf{X}] = R \cdot 0 = 0$$

and

$$\begin{aligned}
\text{var}[R(\hat{\beta} - \beta^o)|\mathbf{X}] &= E\left[R(\hat{\beta} - \beta^o)(R(\hat{\beta} - \beta^o))'|\mathbf{X}\right] \\
&= E\left[R(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'R'|\mathbf{X}\right] \\
&= RE\left[(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'|\mathbf{X}\right]R' \\
&= R\text{var}(\hat{\beta}|\mathbf{X})R' \\
&= \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'.
\end{aligned}$$

It follows that

$$R(\hat{\beta} - \beta^o)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R').$$

Question: What is the role of the $J \times K$ nonstochastic matrix R?

Answer: The $J \times K$ matrix R is a selection matrix. For example, when $R = (1, 0, ..., 0)$, we then have $R(\hat{\beta} - \beta^o) = \hat{\beta}_0 - \beta_0^o$.

Question: Why would we like to know the sampling distribution of $R(\hat{\beta} - \beta^o)$?

This is mainly for confidence interval estimation and hypothesis testing.

3.6 Estimation of the Covariance Matrix of $\hat{\beta}$

Since $\text{var}(\varepsilon_t) = \sigma^2$ is unknown, $\text{var}[R(\hat{\beta} - \beta^o)|\mathbf{X}] = \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'$ is unknown. We need to estimate σ^2. We can use the residual variance estimator

$$s^2 = e'e/(n - K).$$

Theorem [Residual Variance Estimator]: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then we have for all $n > K$, (i)*

$$\frac{(n-K)s^2}{\sigma^2}|\mathbf{X} = \frac{e'e}{\sigma^2}|\mathbf{X} \sim \chi^2_{n-K},$$

where χ^2_{n-K} denotes the Chi-square distribution with $n - K$ degrees of freedom;

(ii) conditional on $\mathbf{X}$, s^2 and $\hat{\beta}$ are independent.

Proof: (i) Because $e = M\varepsilon$, we have

$$\frac{e'e}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right).$$

In addition, because $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I_n)$, and M is an idempotent matrix with rank $= n - K$ (as has been shown before), we have the quadratic form

$$\frac{e'e}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2}|\mathbf{X} \sim \chi^2_{n-K}$$

by the following lemma.

Lemma [Quadratic form of normal random variables]: *If $v \sim N(0, I_n)$ and Q is an $n \times n$ nonstochastic symmetric idempotent matrix with rank $q \leq n$, then the quadratic form*

$$v'Qv \sim \chi^2_q.$$

In our application, we have $v = \varepsilon/\sigma \sim N(0, I)$, and $Q = M$. Since $\text{rank}(M) = n - K$, we have

$$\frac{e'e}{\sigma^2}|\mathbf{X} \sim \chi^2_{n-K}.$$

(ii) Next, we show that s^2 and $\hat{\beta}$ are independent. Because $s^2 = e'e/(n-K)$ is a function of e, it suffices to show that e and $\hat{\beta}$ are independent. This follows immediately because both e and $\hat{\beta}$ are jointly normally distributed and they are uncorrelated. It is well-known that for a joint normal distribution, zero correlation is equivalent to independence.

It remains to show that e and $\hat{\beta}$ jointly normally distributed? For this purpose, we write

$$\begin{aligned}\begin{bmatrix} e \\ \hat{\beta} - \beta^o \end{bmatrix} &= \begin{bmatrix} M\varepsilon \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \end{bmatrix} \\ &= \begin{bmatrix} M \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{bmatrix} \varepsilon.\end{aligned}$$

Because $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I)$, the linear combination of

$$\begin{bmatrix} M \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{bmatrix} \varepsilon$$

is also normally distributed conditional on $\mathbf{X}$. It follows that e and $\hat{\beta}$ are independent given $\text{cov}(\hat{\beta}, e|\mathbf{X}) = 0$. This completes the proof.

Question: What is a χ^2_q distribution?

Definition [Chi-square Distribution, χ^2_q] *Suppose $\{Z_i\}_{i=1}^q$ are i.i.d.N(0,1) random variables. Then the random variable*

$$\chi^2 = \sum_{i=1}^{q} Z_i^2$$

will follow a χ^2_q distribution.

The χ^2_q distribution is nonsymmetric and has long right tails. For a χ^2_q random variable, we have $E(\chi^2_q) = q$ and $\text{var}(\chi^2_q) = 2q$.

Based on these properties of a χ^2 distribution, Theorem (i) implies

$$\begin{aligned} E\left[\frac{(n-K)s^2}{\sigma^2}|\mathbf{X}\right] &= n-K. \\ \frac{(n-K)}{\sigma^2}E(s^2|\mathbf{X}) &= n-K. \end{aligned}$$

It follows that $E(s^2|\mathbf{X}) = \sigma^2$. Note that we have shown this result with a different method but under a more general condition.

Theorem (i) also implies

$$\begin{aligned} \text{var}\left[\frac{(n-K)s^2}{\sigma^2}|\mathbf{X}\right] &= 2(n-K), \\ \text{var}(s^2|\mathbf{X}) &= \frac{2\sigma^4}{n-K} \\ &\to 0 \end{aligned}$$

as $n \to \infty$.

Both Theorems (i) and (ii) imply that the conditional MSE of s^2

$$\begin{aligned} MSE(s^2|\mathbf{X}) &= E\left[(s^2-\sigma^2)^2|\mathbf{X}\right] \\ &= \text{var}(s^2|\mathbf{X}) + [E(s^2|\mathbf{X}) - \sigma^2]^2 \\ &\to 0. \end{aligned}$$

Thus, s^2 is a good estimator for σ^2.

The independence between s^2 and $\hat{\beta}$ is crucial for us to obtain the sampling distribution of the popular t-test and F-test statistics, which will be introduced shortly.

The sample residual variance $s^2 = e'e/(n-K)$ is a generalization of the sample variance $S_n^2 = (n-1)^{-1}\Sigma_{t=1}^n (Y_t - \bar{Y})^2$ for the random sample $\{Y_t\}_{t=1}^n$. The factor $n-K$ is called the degrees of freedom of the estimated residual sample $\{e_t\}_{t=1}^n$. To gain tuition why the degrees of freedom is equal to $n-K$, note that the orginal sample $\{Z_t\}_{t=1}^n = \{Y_t, X_t'\}_{t=1}^n$ has n

observations, which can be viewed to have n degrees of freedom. Now when estimating σ^2, we have to use the estimated residual sample $\{e_t\}_{t=1}^n$. These n estimated residuals are not linearly independent because they have to satisfy the FOC of the OLS estimation, namely,

$$\begin{aligned} \mathbf{X}'e &= 0. \\ (K \times n) \times (n \times 1) &= K \times 1. \end{aligned}$$

The FOC imposes K restrictions on $\{e_t\}_{t=1}^n$, conditional on $\mathbf{X}$. These K restrictions are needed in order to estimate K unknown paramerers β^o. They can used to obtain the remaining K estmated residuals $\{e_{T-K+1}, ..., e_T\}$ from the first $n-K$ estimated residuals $\{e_1, ..., e_{n-K}\}$ if the latter have been available. Thus, the remaining degrees of freedom of e is $n-K$. Note that the sample variance S_n^2 is the residual variance estimator with $Y_t = \beta_0 + \varepsilon_t$.

Question: Why are these sampling distributions of $\hat{\beta}$ and s^2 useful in practice?

They are useful in confidence interval estimation and hypothesis testing on model parameters. In this book, we will focus on hypothesis testing on model parameters. Statistically speaking, confidence interval estimation and hypothesis testong on model parameters are just two sides of the same coin.

3.7 Hypothesis Testing

We now use the sampling distributions of $\hat{\beta}$ and s^2 to develop test procedures for hypotheses of interest. We consider testing the following linear hypothesis in form of

$$\begin{aligned} \mathbf{H}_0 &: R\beta^o = r, \\ (J \times K)(K \times 1) &= J \times 1, \end{aligned}$$

where R is called the selection matrix, and J is the number of restrictions. We assume $J \leq K$.

It is important to emphasize that we will test $\mathbf{H}_0$ under correct model specification for $E(Y_t|X_t)$.

Motivation

We first provide a few motivating examples for hypothesis testing.

Example 1 [Reforms have no effect]: Consider the extended production function

$$\ln(Y_t) = \beta_0 + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) + \beta_3 AU_t + \beta_4 PS_t + \varepsilon_t,$$

where AU_t is a dummy variable indicating whether firm t is granted autonomy, and PS_t is the profit share of firm t with the state.

Suppose we are interested in testing whether autonomy AU_t has an effect on productivity. Then we can write the null hypothesis

$$\mathbf{H}_0^a : \beta_3^o = 0$$

This is equivalent to the choices of:

$$\begin{aligned} \beta^o &= (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'. \\ R &= (0, 0, 0, 1, 0), \\ r &= 0. \end{aligned}$$

If we are interested in testing whether profit sharing has an effect on productivity, we can consider the null hypothesis

$$\mathbf{H}_0^b : \beta_4^o = 0.$$

Alternatively, to test whether the production technology exhibits the constant return to scale (CRS), we can write the null hypothesis as follows:

$$\mathbf{H}_0^c : \beta_1^o + \beta_2^o = 1.$$

This is equivalent to the choice of $R = (0, 1, 1, 0, 0)$ and $r = 1$.

Finally, if we are interested in examining the joint effect of both autonomy and profit sharing, we can test the hypothesis that neither autonomy nor profit sharing has impact:

$$\mathbf{H}_0^d : \beta_3^o = \beta_4^o = 0.$$

This is equivalent to the choice of

$$\begin{aligned} R &= \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\ r &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

Example 2 [Optimal Predictor for Future Spot Exchange Rate]: Consider

$$S_{t+\tau} = \beta_0 + \beta_1 F_t(\tau) + \varepsilon_{t+\tau}, \qquad t = 1, ..., n,$$

where $S_{t+\tau}$ is the spot exchange rate at period $t + \tau$, and $F_t(\tau)$ is the forward exchange rate, namely the period t's price for the foreign currency to be delivered at period $t + \tau$. The null hypothesis of interest is that the forward exchange rate $F_t(\tau)$ is an optimal predictor for the future spot rate $S_{t+\tau}$ in the sense that $E(S_{t+\tau}|I_t) = F_t(\tau)$, where I_t is the information set available at time t. This is actually called the *expectations hypothesis* in economics and finance. Given the above specification, this hypothesis can be written as

$$\mathbf{H}_0^e : \beta_0^o = 0, \beta_1^o = 1,$$

and $E(\varepsilon_{t+\tau}|I_t) = 0$. This is equivalent to the choice of

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, r = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

All examples considered above can be formulated with a suitable specification of R, where R is a $J \times K$ matrix in the null hypothesis

$$\mathbf{H}_0 : R\beta^o = r,$$

where r is a $J \times 1$ vector.

Basic Idea of Hypothesis Testing

To test the null hypothesis

$$\mathbf{H}_0 : R\beta^o = r,$$

we can consider the statistic:

$$R\hat{\beta} - r$$

and check if this difference is significantly different from zero.

Under $\mathbf{H}_0 : R\beta^o = r$, we have

$$\begin{aligned} R\hat{\beta} - r &= R\hat{\beta} - R\beta^o \\ &= R(\hat{\beta} - \beta^o) \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

because $\hat{\beta} - \beta^o \to 0$ as $n \to \infty$ in the term of MSE.

Under the alternative to $\mathbf{H}_0$, $R\beta^o \neq r$, but we still have $\hat{\beta} - \beta^o \to 0$ in the term of MSE. It follows that

$$\begin{aligned} R\hat{\beta} - r &= R(\hat{\beta} - \beta^o) + R\beta^o - r \\ &\to R\beta^o - r \neq 0 \end{aligned}$$

as $n \to \infty$, where the convergence is in term of MSE. In other words, $R\hat{\beta} - r$ will converge to a nonzero limit, $R\beta^o - r$.

The fact that the behavior of $R\hat{\beta} - r$ is different under $\mathbf{H}_0$ and under the alternative hypothesis to $\mathbf{H}_0$ provides a basis to construct hypothesis tests. In particular, we can test $\mathbf{H}_0$ by examining whether $R\hat{\beta} - r$ is significantly different from zero.

Question: How large should the magnitude of the absolute value of the difference $R\hat{\beta} - r$ be in order to claim that $R\hat{\beta} - r$ is significantly different from zero?

For this purpose, we need a decision rule which specifies a threshold value with which we can compare the (absolute) value of $R\hat{\beta} - r$. Because $R\hat{\beta} - r$ is a random variable

and so it can take many (possibly an infinite number of) values. Given a data set, we only obtain one realization of $R\hat{\beta} - r$. Whether a realization of $R\hat{\beta} - r$ is close to zero should be judged using the critical value of its sampling distribution, which depends on the sample size n and the significance level $\alpha \in (0, 1)$ one preselects.

Question: What is the sampling distribution of $R\hat{\beta} - r$ under $\mathbf{H}_0$?

Because

$$R(\hat{\beta} - \beta^o)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'),$$

we have that conditional on $\mathbf{X}$,

$$\begin{aligned} R\hat{\beta} - r &= R(\hat{\beta} - \beta^o) + R\beta^o - r \\ &\sim N(R\beta^o - r, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R') \end{aligned}$$

Corollary: *Under Assumptions 3.1, 3.3 and 3.5, and* $\mathbf{H}_0 : R\beta^o = r$*, we have for each* $n > K$*,*

$$(R\hat{\beta} - r)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R').$$

The difference $R\hat{\beta} - r$ cannot be used as a test statistic for $\mathbf{H}_0$, because σ^2 is unknown and there is no way to calculate the critical values of the sampling distribution of $R\hat{\beta} - r$.

Question: How to construct a feasible (i.e., computable) test statistic?

The forms of test statistics will differ depending on whether we have $J = 1$ or $J > 1$. We first consider the case of $J = 1$.

CASE I: t-Test ($J = 1$):

Recall that we have

$$(R\hat{\beta} - r)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'),$$

When $J = 1$, the conditional variance

$$\text{var}[(R\hat{\beta} - r)|\mathbf{X}] = \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'$$

is a scalar (1×1). It follows that conditional on $\mathbf{X}$, we have

$$\begin{aligned} \frac{R\hat{\beta} - r}{\sqrt{\text{var}[(R\hat{\beta} - r)|\mathbf{X}]}} &= \frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \\ &\sim N(0, 1). \end{aligned}$$

Question: What is the unconditional distribution of

$$\frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}?$$

The unconditional distribution is also N(0,1).

However, σ^2 is unknown, so we cannot use the ratio

$$\frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}$$

as a test statistic. We have to replace σ^2 by s^2, which is a good estimator for σ^2. This gives a feasible (i.e., computable) test statistic

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

However, the test statistic T will be no longer normally distributed. Instead,

$$\begin{aligned} T &= \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \\ &= \frac{\frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2}/(n-K)}} \\ &\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-K}/(n-K)}} \\ &\sim t_{n-K}, \end{aligned}$$

where t_{n-K} denotes the Student t-distribution with $n-K$ degrees of freedom. Note that the numerator and denominator are mutually independent conditional on $\mathbf{X}$, because $\hat{\beta}$ and s^2 are mutually independent conditional on $\mathbf{X}$. The feasible statistic T is called a t-test statistic because it follows a t_{n-K} distribution.

Question: What is a Student t_q distribution?

Definition [Student's t-distribution]: *Suppose $Z \sim N(0,1)$ and $V \sim \chi^2_q$, and both Z and V are independent. Then the ratio*

$$\frac{Z}{\sqrt{V/q}} \sim t_q.$$

The t_q-distribution is symmetric about 0 with heavier tails than the $N(0,1)$ distribution. The smaller muber of the degrees of freedom, the heavier tails it has. When $q \to \infty$, $t_q \xrightarrow{d} N(0,1)$, where $\xrightarrow{d}$ denotes convergence in distribution. This implies that we have

$$T = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

This result has a very important implication in practice: For a large sample size n, it makes no difference to use either the critical values from t_{n-K} or from $N(0,1)$.

Question: What is convergence in distribution?

Definition [Convergence in distribution]: *Suppose $\{Z_n, n = 1, 2, \cdots\}$ is a sequence of random variables/vectors with distribution functions $F_n(z) = P(Z_n \leq z)$, and Z is a random variable/vector with distribution $F(z) = P(Z \leq z)$. We say that Z_n converges to Z in distribution if the distribution of Z_n converges to the distribution of Z at all continuity points; namely,*

$$\begin{aligned} \lim_{n\to\infty} F_n(z) &= F(z) \text{ or} \\ F_n(z) &\to F(z) \text{ as } n \to \infty \end{aligned}$$

for any continuity point z (i.e., for any point at which $F(z)$ is continuous). We use the notation $Z_n \xrightarrow{d} Z$. The distribution of Z is called the asymptotic or limiting distribution of Z_n.

In practice, Z_n is a test statistic or a parameter estimator, and often its sampling distribution $F_n(z)$ is either unknown or very complicated, but $F(z)$ is known or very simple. As long as $Z_n \xrightarrow{d} Z$, then we can use $F(z)$ as an approximation to $F_n(z)$. This gives a convenient procedure for statistical inference. The potential cost is that the approximation of $F_n(z)$ to $F(z)$ may not be good enough in finite samples (i.e., when n is finite). How good the approximation will depend on the data generating process and the sample size n.

Example: Suppose $\{\varepsilon_n, n = 1, 2, \cdots\}$ is an *i.i.d.* sequence with distribution function $F(z)$. Let ε be a random variable with the same distribution function $F(z)$. Then $\varepsilon_n \xrightarrow{d} \varepsilon$.

With the obtained sampling distribution for the test statistic T, we can now describe a decision rule for testing $\mathbf{H}_0$ when $J = 1$.

Decision Rule of the T-test Based on Critical Values

(i) Reject $\mathbf{H}_0 : R\beta^o = r$ at a prespecified significance level $\alpha \in (0,1)$ if

$$|T| > C_{t_{n-K,\frac{\alpha}{2}}},$$

where $C_{t_{n-K,\frac{\alpha}{2}}}$ is the so-called upper-tailed critical value of the t_{n-K} distribution at level $\frac{\alpha}{2}$, which is determined by

$$P\left[t_{n-K} > C_{t_{n-K,\frac{\alpha}{2}}}\right] = \frac{\alpha}{2}$$

or equivalently

$$P\left[|t_{n-K}| > C_{t_{n-K,\frac{\alpha}{2}}}\right] = \alpha.$$

(ii) Accept $\mathbf{H}_0$ at the significance level α if

$$|T| \leq C_{t_{n-K,\frac{\alpha}{2}}}.$$

Remark: In testing $\mathbf{H}_0$, there exist two types of errors, due to the limited information about the population in a given random sample $\{Z_t\}_{t=1}^n$. One possibility is that $\mathbf{H}_0$ is true but we reject it. This is called the "Type I error". The significance level α is the probability of making the Type I error. If

$$P\left[|T| > C_{t_{n-K,\frac{\alpha}{2}}} | \mathbf{H}_0\right] = \alpha,$$

we say that the decision rule is a test with size α.

On the other hand, the probability $P[|T| > C_{t_{n-K,\frac{\alpha}{2}}} | \mathbf{H}_0 \text{ is false}]$ is called the power function of a size α test. When

$$P\left[|T| > C_{t_{n-K,\frac{\alpha}{2}}} | \mathbf{H}_0 \text{ is false}\right] < 1,$$

there exists a possibility that one may accept $\mathbf{H}_0$ when it is false. This is called the "Type II error".

Ideally one would like to minimize both the Type I error and Type II error, but this is impossible for any given finite sample. In practice, one usually presets the level for Type I error, the so-called significance level, and then minimizes the Type II error. Conventional choices for significance level α are 10%, 5% and 1% respectively.

Next, we describe an alternative decision rule for testing $\mathbf{H}_0$ when $J = 1$, using the so-called p-value of test statistic T.

An Equivalent Decision Rule Based on p-values

Given a data set $\mathbf{z}^n = \{y_t, x_t'\}_{t=1}^{\prime n}$, which is a realization of the random sample $\mathbf{Z}^n = \{Y_t, X_t'\}_{t=1}^{\prime n}$, we can compute a realization (i.e., a number) for the t-test statistic T, namely

$$T(\mathbf{z}^n) = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{x}'\mathbf{x})^{-1}R'}}.$$

Then the probability

$$p(\mathbf{z}^n) = P\left[|t_{n-K}| > |T(\mathbf{z}^n)|\right],$$

is called the p-value (i.e., probability value) of the test statistic T given that $\{Y_t, X_t'\}_{t=1}^n = \{y_t, x_t'\}_{t=1}^n$ is observed, where t_{n-K} is a Student t random variable with $n-K$ degrees of freedom, and $T(\mathbf{z}^n)$ is a realization for test statistic $T = T(\mathbf{Z}^n)$ given the observed data $\mathbf{z}^n$. Intuitively, the p-value is the smallest value of significance level α for which the null hypothesis is rejected. Here, it is the tail probability that the absolute value of a Student t_{n-K} random variable take values larger than the absolute value of the test statistic $T(\mathbf{z}^n)$. If this probability is very small relative to the significance level, then it is unlikely that the test statistic $T(\mathbf{Z}^n)$ will follow a Student t_{n-K} distribution. As a consequence, the null hypothesis is likely to be false.

The above decision rule can be described equivalently as follows:

Decision Rule Based on the p-value

(i) Reject $\mathbf{H}_0$ at the significance level α if $p(\mathbf{z}^n) < \alpha$.

(ii) Accept $\mathbf{H}_0$ at the significance level α if $p(\mathbf{z}^n) \geq \alpha$.

Remark:

A small p-value is evidence against the null hypothesis. A large p-value shows that the data are consistent with the null hypothesis.

Question: What are the advantages/disadvantages of using p-values versus using critical values?

p-values are more informative than only rejecting/accepting the null hypothesis at some significance level α. A p-value is the smallest significance level at which a null hypothesis can be rejected. It not only tells us whether the null hypothesis should be accepted or rejected, but it also tells us whether the decision to accept or reject the null hypothesis is a close call.

Most statistical software reports p-values of parameter estimates. This is much more convenient than asking the user to specify significance level α and then reporting whether the null hypothesis is accepted or rejected for that α.

When we reject a null hypothesis, we often say there is a statistically significant effect. This does not mean that there is an effect of practical importance (i.e., an effect of economic importance). This is because when large samples are used, small and practically unimportant effects are likely to be statistically significant.

The t-test and associated procedures just introduced are valid even when there exists near-multicolinearity, where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n and $\text{var}(\hat{\beta}|\mathbf{X})$ does not vanish to zero as $n \to \infty$. However, the degree of near-multicolineary, as measured by sample correlations between explanatory variables, will affect the the precision of the OLS estimator $\hat{\beta}$. Other things being equal, the higher degree of near-multicolinearity, the larger the variance of $\hat{\beta}$. As a result, the t-statistic is often insignificant even when the null hypthesis $\mathbf{H}_0$ is false.

Examples of t-tests

Example 1 [Reforms have no effects (continued.)]

We first consider testing the null hypothesis

$$\mathbf{H}_0^a : \beta_3 = 0,$$

where β_3 is the coefficient of the autonomy AU_t in the extended production function regression model. This is equivalent to the selection of $R = (0, 0, 0, 1, 0)$. In this case, we have

$$\begin{aligned} s^2 R(\mathbf{X}'\mathbf{X})^{-1}R' &= \left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(4,4)} \\ &= S^2_{\hat{\beta}_3} \end{aligned}$$

which is the estimator of $\text{var}(\hat{\beta}_3|\mathbf{X})$. The squared root of $\text{var}(\hat{\beta}_3|X)$ is called the standard error of estimator $\hat{\beta}_3$, and $S_{\hat{\beta}_3}$ is called the estimated standard error of $\hat{\beta}_3$. The t-test statistic

$$\begin{aligned} T &= \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \\ &= \frac{\hat{\beta}_3}{\sqrt{S^2_{\hat{\beta}_3}}} \\ &\sim t_{n-K}. \end{aligned}$$

Next, we consider testing the CRS hypothesis

$$\mathbf{H}_0^c : \beta_1 + \beta_2 = 1,$$

which corresponds to $R = (0, 1, 1, 0, 0)$ and $r = 1$. In this case,

$$\begin{aligned} s^2 R(\mathbf{X}'\mathbf{X})^{-1}R' &= S^2_{\hat{\beta}_1} + S^2_{\hat{\beta}_2} + 2\hat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2) \\ &= \left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(2,2)} \\ &\quad + \left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(3,3)} \\ &\quad + 2\left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(2,3)} \\ &= S^2_{\hat{\beta}+\hat{\beta}_2}, \end{aligned}$$

which is the estimator of $\text{var}(\hat{\beta}_1 + \hat{\beta}_2|\mathbf{X})$. Here, $\hat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2)$ is the estimator for $\text{cov}(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X})$, the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ conditional on $\mathbf{X}$.

The t-test statistic is

$$\begin{aligned} T &= \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \\ &= \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{S_{\hat{\beta}_1+\hat{\beta}_2}} \\ &\sim t_{n-K}. \end{aligned}$$

CASE II: F-testing ($J > 1$)

Question: How to construct a test statistic for H_0 if $J > 1$?

We first state a useful lemma.

Lemma: *If $Z \sim N(0, V)$, where $V = var(Z)$ is a nonsingular $J \times J$ variance-covariance matrix, then*

$$Z'V^{-1}Z \sim \chi^2_J.$$

Proof: Because V is symmetric and positive definite, we can find a symmetric and invertible matrix $V^{1/2}$ such that

$$\begin{aligned} V^{1/2}V^{1/2} &= V, \\ V^{-1/2}V^{-1/2} &= V^{-1}. \end{aligned}$$

(Question: What is this decomposition called?) Now, define

$$Y = V^{-1/2}Z.$$

Then we have $E(Y) = 0$, and

$$\begin{aligned}
\text{var}(Y) &= E\left\{[Y - E(Y)][Y - E(Y)]'\right\} \\
&= E(YY') \\
&= E(V^{-1/2}ZZ'V^{-1/2}) \\
&= V^{-1/2}E(ZZ')V^{-1/2} \\
&= V^{-1/2}VV^{-1/2} \\
&= V^{-1/2}V^{1/2}V^{1/2}V^{-1/2} \\
&= I.
\end{aligned}$$

It follows that $Y \sim N(0, I)$. Therefore, we have

$$Y'Y \sim \chi^2_q.$$

Applying this lemma, and using the result that

$$(R\hat{\beta} - r)|\mathbf{X} \sim N[0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R']$$

under $\mathbf{H}_0$, we have the quadratic form

$$(R\hat{\beta} - r)'[\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r) \sim \chi^2_J$$

conditional on $\mathbf{X}$, or

$$\frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi^2_J$$

conditional on $\mathbf{X}$.

Because χ^2_J does not depend on $\mathbf{X}$, therefore, we also have

$$\frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi^2_J$$

unconditionally.

Like in constructing a t-test statistic, we should replace σ^2 by s^2 in the left hand side:

$$\frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)}{s^2}.$$

The replacement of σ^2 by s^2 renders the distribution of the quadratic form no longer Chi-squared. Instead, after proper scaling, the quadratic form will follow a so-called F-distribution with degrees of freedom equal to $(q, n - K)$.

Why?

To explain this, we observe

$$\begin{aligned} & \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{s^2} \\ = & J \cdot \frac{\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{\sigma^2}/J}{\frac{(n-K)s^2}{\sigma^2}/(n-K)} \\ \sim & J \cdot F_{J,n-K}, \end{aligned}$$

where $F_{J,n-K}$ denotes the F distribution with degrees of J and $n-K$ distributions.

Question: What is a $F_{J,n-K}$ distribution?

Definition: *Suppose $U \sim \chi^2_p$ and $V \sim \chi^2_q$, and both U and V are independent. Then the ratio*

$$\frac{U/p}{V/q} \sim F_{p,q}$$

is called to follow a $F_{p,q}$ distribution with degrees of freedom (p, q).

This distribution is called F-distribution because it is named after Professor Fisher, a well-known statistician in the 20th century. It is similar to the shape of the χ^2 distribution with a long right tail. An $F_{p,q}$ random variable has the following properties:

(i) If $F \sim F_{p,q}$, then $F^{-1} \sim F_{q,p}$.

(ii) $t_q^2 \sim F_{1,q}$.

$$t_q^2 = \frac{\chi_1^2/1}{\chi^2/q} \sim F_{1,q}$$

(iii) Given any fixed integer p, $p \cdot F_{p,q} \to \chi_p^2$ as $q \to \infty$.

Property (ii) implies that when $J = 1$, using either the t-test or the F-test will deliver the same conclusion. Property (iii) implies that the conclusions based on $F_{p,q}$ and on $p \cdot F_{p,q}$ using the χ_p^2 approximation will be approximately the same when q is sufficiently large.

We now define the following F-test statistic to test H_0 :

$$\begin{aligned} F & \equiv \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)/J}{s^2} \\ & \sim F_{J,n-K}. \end{aligned}$$

Theorem: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then under $\mathbf{H}_0 : R\beta^o = r$,*

we have

$$\begin{aligned} F &= \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)/J}{s^2} \\ &\sim F_{J,n-K} \end{aligned}$$

for all $n > K$.

Alternative Expression for the F-Test Statistic

A practical issue now is how to compute the F-statistic. One can of course compute the F-test statistic using the above definition of the F test statistic. However, there is a very convenient way to compute the F-test statistic. We now introduce this method.

Theorem: *Suppose Assumptions 3.1 and 3.3(a) hold. Let* $SSR_u = e'e$ *be the sum of squared residuals from the unrestricted model*

$$Y = \mathbf{X}\beta^o + \varepsilon.$$

Let $SSR_r = \tilde{e}'\tilde{e}$ *be the sum of squared residuals from the restricted model*

$$Y = \mathbf{X}\beta^o + \varepsilon$$

subject to

$$R\beta^o = r,$$

where $\tilde{\beta}$ *is the restricted OLS estimator. Then under* $\mathbf{H}_0$*, the* F*-test statistic can be written as*

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/J}{e'e/(n-K)} \sim F_{J,n-K}.$$

Proof: Let $\tilde{\beta}$ be the OLS under $\mathbf{H}_0$; that is,

$$\tilde{\beta} = \arg\min_{\beta \in R^K} (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta)$$

subject to the constraint that $R\beta = r$. We first form the Lagrangian function

$$L(\beta, \lambda) = (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) + 2\lambda'(r - R\beta),$$

where λ is a $J \times 1$ vector called the Lagrange multiplier vector.

We have the following FOC:

$$\begin{aligned} \frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial \beta} &= -2\mathbf{X}'(Y - \mathbf{X}\tilde{\beta}) - 2R'\tilde{\lambda} = 0, \\ \frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial \lambda} &= 2(r - R\tilde{\beta}) = 0. \end{aligned}$$

With the unconstrained OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$, and from the first equation of FOC, we can obtain

$$\begin{aligned} -(\hat{\beta} - \tilde{\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}R'\tilde{\lambda}, \\ R(\mathbf{X}'\mathbf{X})^{-1}R'\tilde{\lambda} &= -R(\hat{\beta} - \tilde{\beta}). \end{aligned}$$

Hence, the Lagrange multiplier

$$\begin{aligned} \tilde{\lambda} &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}R(\hat{\beta} - \tilde{\beta}). \\ &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r), \end{aligned}$$

where we have made use of the constraint that $R\tilde{\beta} = r$. It follows that

$$\hat{\beta} - \tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}R'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r).$$

Now,

$$\begin{aligned} \tilde{e} &= Y - \mathbf{X}\tilde{\beta} \\ &= Y - \mathbf{X}\hat{\beta} + \mathbf{X}(\hat{\beta} - \tilde{\beta}) \\ &= e + \mathbf{X}(\hat{\beta} - \tilde{\beta}). \end{aligned}$$

It follows that

$$\begin{aligned} \tilde{e}'\tilde{e} &= e'e + (\hat{\beta} - \tilde{\beta})'\mathbf{X}'\mathbf{X}(\hat{\beta} - \tilde{\beta}) \\ &= e'e + (R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r). \end{aligned}$$

We have

$$(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r) = \tilde{e}'\tilde{e} - e'e$$

and

$$\begin{aligned} F &= \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2} \\ &= \frac{(\tilde{e}'\tilde{e} - e'e)/J}{e'e/(n-K)}. \end{aligned}$$

This completes the proof.

Remarks:

The F-statistic is a convenient test statistic! One only needs to compute SSR in order to compute the F-test statistic. Intuitively, the sum of squared residuals SSR_u

of the unrestricted regression model is always larger than or at least equal to that of the restricted regression model. When the null hypothesis $\mathbf{H}_0$ is true (i.e., when the parameter restriction is valid), the sum of squared residuals SSR_r of the restricted model is more or less similar to that of the unrestricted model, subject to the difference due to sampling variations. If SSR_r is sufficiently larger than SSR_u, then there exists evidence against $\mathbf{H}_0$.How large a difference between SSR_r and SSR_u is considered as sufficiently large to reject $\mathbf{H}_0$ is determined by the critical value of the associated F distribution.

Question: What is the interpretation for the Lagrange multiplier $\tilde{\lambda}$?

Recall that we have obtained the relation that

$$\begin{aligned}\tilde{\lambda} &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}R(\hat{\beta}-\tilde{\beta}) \\ &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r).\end{aligned}$$

Thus, $\tilde{\lambda}$ is an indicator of the departure of $R\hat{\beta}$ from r. That is, the value of $\tilde{\lambda}$ will indicate whether $R\hat{\beta}-r$ is significantly different from zero.

Question: What happens to the distribution of F when $n \to \infty$?

Recall the important property of the $F_{p,q}$ distribution that $p \cdot F_{p,q} \to^d \chi^2_p$ when $q \to \infty$. Since our F-statistic for $\mathbf{H}_0$ follows a $F_{J,n-K}$ distribution, it follows that under $\mathbf{H}_0$, the quadratic form

$$\begin{aligned}J \cdot F &= (R\hat{\beta}-r)'\left[s^2R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\beta}-r) \\ &\to {}^d\chi^2_J\end{aligned}$$

as $n \to \infty$. We formally state this result below.

Theorem: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then under* $\mathbf{H}_0$, *we have the Wald test statistic*

$$W = \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{s^2} \xrightarrow{d} \chi^2_J$$

as $n \to \infty$.

This result implies that when n is sufficiently large, using the F-statistic and the exact $F_{J,n-K}$ distribution and using the quadratic form W and the simpler χ^2_J approximation will make no essential difference in statistical inference.

3.8 Applications

We now consider some special but important cases often countered in economics and finance.

Case 1: Testing for the Joint Significance of Explanatory Variables

Consider a linear regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \beta_0^o + \sum_{j=1}^{k} \beta_j^o X_{jt} + \varepsilon_t. \end{aligned}$$

We are interested in testing the combined effect of all the regressors except the intercept. The null hypothesis is

$$\mathbf{H}_0 : \beta_j^o = 0 \text{ for } 1 \leq j \leq k,$$

which implies that none of the explanatory variables influences Y_t.
The alternative hypothesis is

$$\mathbb{H}_A : \ \beta_j^o \neq 0 \text{ at least for some } \beta_j^o, \qquad j = 1, \cdots, k.$$

One can use the F-test and

$$F \sim F_{k,n-(k+1)}.$$

In fact, the restricted model under $\mathbf{H}_0$ is very simple:

$$Y_t = \beta_0^o + \varepsilon_t.$$

The restricted OLS estimator $\tilde{\beta} = (\bar{Y}, 0, \cdots, 0)'$. It follows that

$$\tilde{e} = Y - \mathbf{X}\tilde{\beta} = Y - \bar{Y}.$$

Hence, we have

$$\tilde{e}'\tilde{e} = (Y - \bar{Y})'(Y - \bar{Y}).$$

Recall the definition of R^2 :

$$\begin{aligned} R^2 &= 1 - \frac{e'e}{(Y - \bar{Y})'(Y - \bar{Y})} \\ &= 1 - \frac{e'e}{\tilde{e}'\tilde{e}}. \end{aligned}$$

It follows that

$$\begin{aligned} F &= \frac{(\tilde{e}'\tilde{e} - e'e)/k}{e'e/(n-k-1)} \\ &= \frac{(1 - \frac{e'e}{\tilde{e}'\tilde{e}})/k}{\frac{e'e}{\tilde{e}'\tilde{e}}/(n-k-1)} \\ &= \frac{R^2/k}{(1-R^2)/(n-k-1)}. \end{aligned}$$

Thus, it suffices to run one regression, namely the unrestricted model in this case. We emphasize that this formula is valid only when one is testing for $\mathbf{H}_0 : \beta_j^o = 0$ for all $1 \leq j \leq k$.

Example 1 [Efficient Market Hypothesis]: Suppose Y_t is the exchange rate return in period t, and I_{t-1} is the information available at time $t-1$. Then a classical version of the efficient market hypothesis (EMH) can be stated as follows:

$$E(Y_t|I_{t-1}) = E(Y_t)$$

To check whether exchange rate changes are unpredictable using the past history of exchange rate changes, we specify a linear regression model:

$$Y_t = X_t'\beta^o + \varepsilon_t,$$

where

$$X_t = (1, Y_{t-1}, ..., Y_{t-k})'.$$

Under EMH, we have

$$\mathbf{H}_0 : \beta_j^o = 0 \text{ for all } j = 1, ..., k.$$

If the alternative

$$\mathbb{H}_A : \beta_j^o \neq 0 \text{ at least for some } j \in \{1, ..., k\}$$

holds, then exchange rate changes are predictable using the past information.

Question: What is the appropriate interpretation if $\mathbf{H}_0$ is not rejected?

Note that there exists a gap between the efficiency hypothesis and $\mathbf{H}_0$, because the linear regression model is just one of many ways to check EMH. Thus, $\mathbf{H}_0$ is not rejected, at most we can only say that no evidence against the efficiency hypothesis is found. We should not conclude that EMH holds.

Strictly speaking, the current theory (Assumption 3.2: $E(\varepsilon_t|\mathbf{X}) = 0$) rules out this application, which is a dynamic time series regression model. However, we will justify in Chapter 5 that

$$\begin{aligned} k \cdot F &= \frac{R^2}{(1-R^2)/(n-k-1)} \\ &\xrightarrow{d} \chi^2_k \end{aligned}$$

under conditional homoskedasticity even for a linear dynamic regression model.

In fact, we can use a simpler version when n is large:

$$(n-k-1)R^2 \xrightarrow{d} \chi^2_k.$$

This follows from the Slutsky theorem because $R^2 \to^p 0$ under $\mathbf{H}_0$. Although Assumption 3.5 is not needed for this result, conditional homoskedasticity is still needed, which rules out autoregressive conditional heteroskedasticity (ARCH) in the time series context.

Below is a concrete numerical example.

Example 1 [Consumption Function and Wealth Effect]: Let Y_t = consumption, X_{1t} = labor income, X_{2t} = liquidity asset wealth. A regression estimation gives

$$\begin{aligned} Y_t &= 33.88 - 26.00X_{1t} + 6.71X_{2t} + e_t, \qquad R^2 = 0.742, n = 25. \\ &\quad [1.77] \qquad [-0.74] \qquad [0.77] \end{aligned}$$

where the numbers inside $[\cdot]$ are t-statistics.

Suppose we are interested in whether labor income or liquidity asset wealth has impact on consumption. We can use the F-test statistic,

$$\begin{aligned} F &= \frac{R^2/2}{(1-R^2)/(n-3)} \\ &= (0.742/2)/[(1-0.742)/(25-3)] \\ &= 31.636 \\ &\sim F_{2,22} \end{aligned}$$

Comparing it with the critical value of $F_{2,22}$ at the 5% significance level, we reject the null hypothesis that neither income nor liquidity asset has impact on consumption at the 5% significance level.

Case 2: Testing for Omitted Variables (or Testing for No Effect)

Suppose $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, where $\mathbf{X}^{(1)}$ is a $n \times (k_1 + 1)$ matrix and $\mathbf{X}^{(2)}$ is a $n \times k_2$ matrix.

A random vector $X_t^{(2)}$ has no explanatory power for the conditional expectation of Y_t if

$$E(Y_t|X_t) = E(Y_t|X_t^{(1)}).$$

Alternatively, it has explanatory power for the conditional expectation of Y_t if

$$E(Y_t|X_t) \neq E(Y_t|X_t^{(1)}).$$

When $X_t^{(2)}$ has explaining power for Y_t but is not included in the regression, we say that $X_t^{(2)}$ is an omitted random variable or vector.

Question: How to test whether $X_t^{(2)}$ is an omitted variable in the linear regression context?

Consider the restricted model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_{k_1} X_{k_1 t} + \varepsilon_t.$$

Suppose we have additional k_2 variables $(X_{(k_1+1)t}, \cdots, X_{(k_1+k_2)t})$, and so we consider the unrestricted regression model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_{1t} + ... + \beta_{k_1} X_{k_1 t} \\ &\quad + \beta_{k_1+1} X_{(k_1+1)t} + \cdots + \beta_{(k_1+k_2)} X_{(k_1+k_2)t} + \varepsilon_t. \end{aligned}$$

The null hypothesis is that the additional variables have no effect on Y_t. If this is the case, then

$$\mathbf{H}_0 : \beta_{k_1+1} = \beta_{k_1+2} = \cdots = \beta_{k_1+k_2} = 0.$$

The alternative is that at least one of the additional variables has effect on Y_t.

The F-Test statistic is

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/k_2}{e'e/(n - k_1 - k_2 - 1)} \sim F_{k_2, n-(k_1+k_2+1)}.$$

Question: Suppose we reject the null hypothesis. Then some important explanatory variables are omitted, and they should be included in the regression. On the other hand, if the F-test statistic does not reject the null hypothesis $\mathbf{H}_0$, can we say that there is no omitted variable?

No. There may exist a nonlinear relationship for additional variables which a linear regression specification cannot capture.

Example 1 [Testing for the Effect of Reforms]:

Consider the extended production function

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) \\ &\quad + \beta_3 AU_t + \beta_4 PS_t + \beta_5 CM_t + \varepsilon_t, \end{aligned}$$

where AU_t is the autonomy dummy, PS_t is the profit sharing ratio, and CM_t is the dummy for change of manager. The null hypothesis of interest here is that none of the three reforms has impact:

$$\mathbf{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0.$$

We can use the F-test, and $F \sim F_{3,n-6}$ under $\mathbf{H}_0$.

Suppose rejection occurs. Then there exists evidence against $\mathbf{H}_0$. However, if no rejection occurs, then we can only say that we find no evidence against $\mathbf{H}_0$ (which is not the same as the statement that reforms have no effect). It is possible that the effect of $X_t^{(2)}$ is of nonlinear form. In this case, we may obtain a zero coefficient for $X_t^{(2)}$, because the linear specification may not be able to capture it.

Example 2 [Testing for Granger Causality]:

Consider two time series $\{Y_t, Z_t\}$, where t is the time index, $I_{t-1}^Y = \{Y_{t-1}, ..., Y_1\}$ and $I_{t-1}^Z = \{Z_{t-1}, ..., Z_1\}$. For example, Y_t is the GDP growth, and Z_t is the money supply growth. We say that Z_t does not Granger-cause Y_t in conditional mean with respect to $I_{t-1} = \{I_{t-1}^{(Y)}, I_{t-1}^{(Z)}\}$ if

$$E(Y_t | I_{t-1}^{(Y)}, I_{t-1}^{(Z)}) = E(Y_t | I_{t-1}^{(Y)}).$$

In other words, the lagged variables of Z_t have no impact on the level of Y_t.

In time series analysis, Granger causality is defined in terms of incremental predictability rather than the real cause-effect relationship. From an econometric point of view, it is a test of omitted variables in a time series context. It is first introduced by Granger (1969).

Question: How to test Granger causality?

Consider now a linear regression model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} \\ &\quad + \beta_{p+1} Z_{t-1} + \cdots + \beta_{p+q} Z_{t-q} + \varepsilon_t. \end{aligned}$$

Under non-Granger causality, we have

$$\mathbf{H}_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0.$$

The F-test statistic

$$F \sim F_{q,n-(p+q+1)}.$$

The current econometric theory (Assumption 3.2: $E(\varepsilon_t|\mathbf{X}) = 0$) actually rules out this application, because it is a dynamic regression model. However, we will justify in Chapter 5 that under H_0,

$$q \cdot F \xrightarrow{d} \chi^2_q$$

as $n \to \infty$ under conditional homoskedasticity even for a linear dynamic regression model.

Example 2 [Testing for Structural Change (or testing for regime shift)]

Consider a bivariate regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t,$$

where t is a time index, and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent. Suppose there exist changes after $t = t_0$, i.e., there exist structural changes. We can consider the extended regression model:

$$\begin{aligned} Y_t &= (\beta_0 + \alpha_0 D_t) + (\beta_1 + \alpha_1 D_t) X_{1t} + \varepsilon_t \\ &= \beta_0 + \beta_1 X_{1t} + \alpha_0 D_t + \alpha_1 (D_t X_{1t}) + \varepsilon_t, \end{aligned}$$

where $D_t = 1$ if $t > t_0$ and $D_t = 0$ otherwise. The variable D_t is called a dummy variable, indicating whether it is a pre- or post-structral break period.

The null hypothesis of no structral change is

$$\mathbf{H}_0 : \alpha_0 = \alpha_1 = 0.$$

The alternative hypothesis that there exists a structral change is

$$\mathbb{H}_A : \alpha_0 \neq 0 \text{ or } \alpha_1 \neq 0.$$

The F-test statistic

$$F \sim F_{2,n-4}.$$

The idea of such a test is first proposed by Chow (1960).

Case 3: Testing for linear restrictions

Example 1 [Testing for CRS]:

Consider the extended production function

$$\ln(Y_t) = \beta_0 + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) + \beta_3 AU_t + \beta_4 PS_t + \beta_5 CM_t + \varepsilon_t.$$

We will test the null hypothesis of CRS:

$$\mathbf{H}_0 : \beta_1 + \beta_2 = 1.$$

The alternative hypothesis is

$$\mathbf{H}_0 : \beta_1 + \beta_2 \neq 1.$$

What is the restricted model under $\mathbf{H}_0$? It is given by

$$\ln(Y_t) = \beta_0 + \beta_1 \ln(L_t) + (1 - \beta_1) \ln(K_t) + \beta_3 AU_t + \beta_4 PS_t + \beta_5 CM_t + \varepsilon_t$$

or equivanetly

$$\ln(Y_t/K_t) = \beta_0 + \beta_1 \ln(L_t/K_t) + \beta_3 AU_t + \beta_4 CON_t + \beta_5 CM_t + \varepsilon_t.$$

The F-test statistic

$$F \sim F_{1,n-6}.$$

Because there is only one restriction, both t- and F- tests are applicable to test CRS.

Example 2 [Wage Determination]: Consider the wage function

$$\begin{aligned} W_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 U_t \\ &\quad + \beta_4 V_t + \beta_5 W_{t-1} + \varepsilon_t, \end{aligned}$$

where W_t = wage, P_t = price, U_t = unemployment, and V_t = unfilled vacancies.

We will test the null hypothesis

$$\mathbf{H}_0 : \beta_1 + \beta_2 = 0, \beta_3 + \beta_4 = 0, \text{ and } \beta_5 = 1.$$

Question: What is the economic interpretation of the null hypothesis $\mathbf{H}_0$?

Under $\mathbf{H}_0$, we have the restricted wage equation:

$$\Delta W_t = \beta_0 + \beta_1 \Delta P_t + \beta_4 D_t + \varepsilon_t,$$

where $\Delta W_t = W_t - W_{t-1}$ is the wage growth rate, $\Delta P_t = P_t - P_{t-1}$ is the inflation rate, and $D_t = V_t - U_t$ is an index for job market situation (excess job supply). This implies that the wage increase depends on the inflation rate and the excess labor supply.

The F-test statistic for $\mathbf{H}_0$ is

$$F \sim F_{3,n-6}.$$

Case 4: Testing for Near-Multicollinearity

Example 1 [Consumption Function (Cont.)]:

Consider the following estimation results for three separate regressions based on the same data set with $n = 25$. The first is a regression of consumption on income:

$$\begin{aligned} Y_t &= 36.74 + 0.832X_{1t} + e_{1t}, \qquad R^2 = 0.735 \\ &\quad [1.98][7.98] \end{aligned}$$

The second is a regression of consumption on wealth:

$$\begin{aligned} Y_t &= 36.61 + 0.208X_{2t} + e_{2t}, \qquad R^2 = 0.735 \\ &\quad [1.97][7.99] \end{aligned}$$

The third is a regression of consumption on both income and wealth:

$$\begin{aligned} Y &= 33.88 - 26.00X_{1t} + 6.71X_{2t} + e_t, \qquad R^2 = 0.742, \\ &\quad [1.77][-0.74][0.77] \end{aligned}$$

Note that in the first two separate regressions, we can find significant t-test statistics for income and wealth, but in the third joint regression, both income and wealth are insignificant. This may be due to the fact that income and wealth are highly multicolinear! To test neither income nor wealth has impact on consumption, we can use the F-test:

$$\begin{aligned} F &= \frac{R^2/2}{(1-R^2)/(n-3)} \\ &= \frac{0.742/2}{(1-0.742)/(25-3)} \\ &= 31.636 \\ &\sim F_{2,22}\,. \end{aligned}$$

This F-test shows that the null hypothesis is firmly rejected at the 5% significance level, because the critical value of $F_{2,22}$ at the 5% level is 3.44.

3.9 Generalized Least Squares Estimation

Question: The classical linear regression theory crucially depends on the assumption that $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I)$, or equivalently $\{\varepsilon_t\} \sim i.i.d.N(0, \sigma^2)$, and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent. What may happen if some classical assumptions do not hold?

Question: Under what conditions, the existing procedures and results are still approximately true?

Assumption 3.5 is unrealistic for many economic and financial data. Suppose Assumption 3.5 is replaced by the following condition:

Assumption 3.6: $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 V)$, *where* $0 < \sigma^2 < \infty$ *is unknown and* $V = V(\mathbf{X})$ *is a known* $n \times n$ *symmetric, finite and positive definite matrix.*

Remarks:
Assumption 3.6 implies that

$$\begin{aligned} \text{var}(\varepsilon|\mathbf{X}) &= E(\varepsilon\varepsilon'|\mathbf{X}) \\ &= \sigma^2 V = \sigma^2 V(\mathbf{X}) \end{aligned}$$

is known up to a constant σ^2. It allows for conditional heteroskedasticity of known form.

In Assumption 3.6, it is possible that V is not a diagonal matrix. Thus, $\text{cov}(\varepsilon_t, \varepsilon_s|\mathbf{X})$ may not be zero. In other words, Assumption 3.6 allows conditional autocorrelation of known form. If t is a time index, this implies that there exists serial correlation of unknown form. If t is an index for cross-sectional units, this implies that there exists spatial correlation of unknown form.

However, the assumption that V is known is still very restrictive from a practical point of view. In practice, V usually has an unknown form.

Question: What is the statistical property of OLS $\hat{\beta}$ under Assumption 3.6?

Theorem: *Suppose Assumptions 3.1, 3.3(a) and 3.6 hold. Then*
(i) unbiasedness: $E(\hat{\beta}|\mathbf{X}) = \beta^o$.
(ii) variance:

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\neq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

(iii)

$$(\hat{\beta} - \beta^o)|\mathbf{X} \sim N(0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}).$$

(iv) $cov(\hat{\beta}, e|\mathbf{X}) \neq 0$ *in general.*

Proof: (i) Using $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$, we have

$$\begin{aligned} E[(\hat{\beta} - \beta^o)|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'0 \\ &= 0. \end{aligned}$$

(ii)

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= E[(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'|\mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Note that we cannot further simplify the expression here because $V \neq I$.

(iii) Because

$$\begin{aligned} \hat{\beta} - \beta^o &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\ &= \sum_{t=1}^{n} C_t\varepsilon_t, \end{aligned}$$

where the weighting vector

$$C_t = (\mathbf{X}'\mathbf{X})^{-1}X_t,$$

$\hat{\beta} - \beta^o$ follows a normal distribution given $\mathbf{X}$, because it is a sum of a normal random variables. As a result,

$$\hat{\beta} - \beta^o \sim N(0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}).$$

(iv)

$$\begin{aligned} \text{cov}(\hat{\beta}, e|\mathbf{X}) &= E[(\hat{\beta} - \beta^o)e'|\mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'M|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon'|\mathbf{X})M \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'VM \\ &\neq 0 \end{aligned}$$

because $\mathbf{X}'VM \neq 0$. We can see that it is conditional heteroskedasticity and/or autocorrelation in $\{\varepsilon_t\}$ that cause $\hat{\beta}$ to be correlated with e.

Remarks:

OLS $\hat{\beta}$ is still unbiased and one can show that its variance goes to zero as $n \to \infty$ (see Question 6, Problem Set 03). Thus, it converges to β^o in the sense of MSE.

However, the variance of the OLS estimator $\hat{\beta}$ does no longer have the simple expression of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ under Assumption 3.6. As a consequence, the classical t- and F-test statistics are invalid because they are based on an incorrect variance-covariance matrix of $\hat{\beta}$. That is, they use an incorrect expression of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ rather than the correct variance formula of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.

Theorem (iv) implies that even if we can obtain a consistent estimator for $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and use it to construct tests, we can no longer obtain the Student t-distribution and F-distribution, because the numerator and the denominator in defining the t- and F-test statistics are no longer independent.

Generalized Least Square (GLS) Estimation

To introduce GLS, we first state a useful lemma.

Lemma: *For any symmetric positive definite matrix V, we can always write*

$$\begin{aligned} V^{-1} &= C'C, \\ V &= C^{-1}(C')^{-1} \end{aligned}$$

where C is a $n \times n$ nonsingular matrix.

Question: What is this decomposition called? Note that C may not be symmetric.

Consider the original linear regression model:

$$Y = \mathbf{X}\beta^o + \varepsilon.$$

If we mutitiply the equation by C, we obtain the transformed regression model

$$\begin{aligned} CY &= (C\mathbf{X})\beta^o + C\varepsilon, \text{ or} \\ Y^* &= \mathbf{X}^*\beta^o + \varepsilon^*, \end{aligned}$$

where $Y^* = CY, \mathbf{X}^* = C\mathbf{X}$ and $\varepsilon^* = C\varepsilon$. Then the OLS of this transformed model

$$\begin{aligned} \hat{\beta}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}Y^* \\ &= (\mathbf{X}'C'C\mathbf{X})^{-1}(\mathbf{X}'C'CY) \\ &= (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}Y \end{aligned}$$

is called the Generalized Least Square (GLS) estimator.

Question: What is the nature of GLS?

Observe that

$$\begin{aligned} E(\varepsilon^*|\mathbf{X}) &= E(C\varepsilon|\mathbf{X}) \\ &= CE(\varepsilon|\mathbf{X}) \\ &= C \cdot 0 \\ &= 0. \end{aligned}$$

Also, note that

$$\begin{aligned} \text{var}(\varepsilon^*|\mathbf{X}) &= E[\varepsilon^*\varepsilon^{*\prime}|\mathbf{X}] \\ &= E[C\varepsilon\varepsilon' C'|\mathbf{X}] \\ &= CE(\varepsilon\varepsilon'|\mathbf{X})C' \\ &= \sigma^2 CVC' \\ &= \sigma^2 C[C^{-1}(C')^{-1}]C' \\ &= \sigma^2 I. \end{aligned}$$

It follows from Assumption 3.6 that

$$\varepsilon^*|\mathbf{X} \sim N(0, \sigma^2 I).$$

The transformation makes the new error ε^* conditionally homoskedastic and serially uncorrelated, while maintaining the normality distribution. Suppose that for t, ε_t has a large variance σ_t^2. The transformation $\varepsilon_t^* = C\varepsilon_t$ will discount ε_t by dividing it by its conditional standard deviation so that ε_t^* becomes conditionally homoskedastic. In addition, the transformation also removes possible correlation between ε_t and $\varepsilon_s, t \neq s$. As a consequence, GLS becomes the best linear LS estimator for β^o in term of the Gauss-Markov theorem.

To appreciate how the transformation by matrix C removes conditional heteroskedasticity and eliminates serial correlation, we now consider two examples.

Example 1 [Removing Heteroskedasticity]: Suppose

$$V = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \sigma_n^2 \end{bmatrix},$$

Then

$$C = \begin{bmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \sigma_n^{-1} \end{bmatrix}$$

where $\sigma_i^2 = \sigma_i^2(\mathbf{X}), i = 1, ..., n$, and

$$\varepsilon^* = C\varepsilon = \begin{bmatrix} \frac{\varepsilon_1}{\sigma_1} \\ \frac{\varepsilon_2}{\sigma_2} \\ \cdots \\ \frac{\varepsilon_n}{\sigma_n} \end{bmatrix}.$$

The transformed regression model is

$$Y_t^* = X_t^* \beta^o + \varepsilon_t^*, \qquad t = 1, ..., n,$$

where

$$\begin{aligned} Y_t^* &= Y_t/\sigma_t, \\ X_t^* &= X_t/\sigma_t, \\ \varepsilon_t^* &= \varepsilon_t/\sigma_t. \end{aligned}$$

Example 2 [Eliminating Serial Correlation] Suppose

$$V = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-3} & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-4} & \rho^{n-4} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \cdots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & \rho & 1 \end{bmatrix}.$$

Then we have

$$V^{-1} = \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & \rho^{n-3} & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & \rho^{n-4} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

and

$$C = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

It follows that

$$\varepsilon^* = C\varepsilon = \begin{bmatrix} \sqrt{1-\rho^2}\varepsilon_1 \\ \varepsilon_2 - \rho\varepsilon_1 \\ \cdots \\ \varepsilon_n - \rho\varepsilon_{n-1} \end{bmatrix}.$$

The transformed regression model is

$$Y_t^* = X_t^{*\prime}\beta^o + \varepsilon_t^*, t = 1, ..., n,$$

where

$$\begin{aligned} Y_1^* &= \sqrt{1-\rho^2}Y_1, \qquad Y_t^* = Y_t - \rho Y_{t-1}, t = 2, ..., n, \\ X_1^* &= \sqrt{1-\rho^2}X_1, \qquad X_t^* = X_t - \rho X_{t-1}, t = 2, ..., n, \\ \varepsilon_1^* &= \sqrt{1-\rho^2}\varepsilon_1, \qquad \varepsilon_t^* = \varepsilon_t - \rho\varepsilon_{t-1}, t = 2, ..., n. \end{aligned}$$

The $\sqrt{1-\rho^2}$transformation for $t = 1$ is called the Prais-Winsten transformation.

Theorem: *Under Assumptions* 3.1, 3.3(a) *and 3.6,*
(i) $E(\hat{\beta}^*|\mathbf{X}) = \beta^o$;
(ii) $var(\hat{\beta}^*|\mathbf{X}) = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} = \sigma^2(\mathbf{X}'V^{-1}\mathbf{X})^{-1}$;
(iii) $cov(\hat{\beta}^*, e^*|\mathbf{X}) = 0$, *where* $e^* = Y^* - \mathbf{X}^*\hat{\beta}^*$;
(iv) $\hat{\beta}^*$ *is BLUE.*
(v) $E(s^{*2}|\mathbf{X}) = \sigma^2$, *where* $s^{*2} = e^{*\prime}e^*/(n-K)$.

Proof: Results in (i)–(iii) follow because the GLS is the OLS of the transformed model.

(iv) The transformed model satisfies 3.1, 3.3 and 3.5 of the classical regression assumptions with $\varepsilon^*|\mathbf{X}^* \sim N(0, \sigma^2 I_n)$. It follows that GLS is BLUE by the Gauss-Markov theorem. Result (v) also follows immediately. This completes the proof.

Remarks:

Because $\hat{\beta}^*$ is the OLS of the transformed regression model with i.i.d. $N(0, \sigma^2 I)$ errors, the t-test and F-test are applicable, and these test statistics are defined as follows:

$$\begin{aligned} T^* &= \frac{R\hat{\beta}^* - r}{\sqrt{s^{*2} R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} R'}} \sim t_{n-K}, \\ F^* &= \frac{(R\hat{\beta}^* - r)'[R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R']^{-1}(R\hat{\beta}^* - r)/J}{s^{*2}} \\ &\sim F_{J,n-K}. \end{aligned}$$

It is very important to note that we still have to estimate the proportionality σ^2 in spite of the fact that $V = V(X)$ is known.

When testing whether all coefficients except the intercept are jointly zero, we have $(n-K)R^{*2} \to^d \chi^2_J$.

Because GLS $\hat{\beta}^*$ is BLUE and OLS $\hat{\beta}$ differs from $\hat{\beta}^*$, OLS $\hat{\beta}$ cannot be BLUE.

$$\begin{aligned} \hat{\beta}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}Y^*, \\ &= (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}Y. \\ \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y. \end{aligned}$$

In fact, the most important message of GLS is the insight it provides into the impact of conditional heteroskedasticity and serial correlation on the estimation and inference of the linear regression model. In practice, GLS is generally not feasible, because the $n \times n$ matrix V is of unknown form, where $\text{var}(\varepsilon|\mathbf{X}) = \sigma^2 V$.

Question: What are feasible solutions?

Two Approaches

(i) First Approach: Adaptive feasible GLS

In some cases with additional assumptions, we can use a nonparametric estimator $\hat{V}$ to replace the unknown V, we obtain the adaptive feasible GLS

$$\hat{\beta}^*_a = (\mathbf{X}'\hat{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{V}^{-1}Y,$$

where $\hat{V}$ is an estimator for V. Because V is an $n \times n$ unknown matrix and we only have n data points, it is impossible to estimate V consistently using a sample of size n if we do not impose any restriction on the form of V. In other words, we have to impose some restrictions on V in order to estimate it consistently. For example, suppose we assume

$$\begin{aligned} \sigma^2 V &= \text{diag}\{\sigma_1^2(\mathbf{X}), ..., \sigma_n^2(\mathbf{X})\} \\ &= \text{diag}\{\sigma^2(X_1), ..., \sigma^2(X_n)\}, \end{aligned}$$

where diag$\{\cdot\}$ is a $n \times n$ diagonal matrix and $\sigma^2(X_t) = E(\varepsilon_t^2|X_t)$ is unknown. The fact that $\sigma^2 V$ is a diagonal matrix can arise when $\text{cov}(\varepsilon_t \varepsilon_s|\mathbf{X}) = 0$ for all $t \neq s$, i.e., when there is no serial correlation. Then we can use the nonparametric kernel estimator

$$\begin{aligned}\hat{\sigma}^2(x) &= \frac{\frac{1}{n}\sum_{t=1}^n e_t^2 \frac{1}{b} K\left(\frac{x-X_t}{b}\right)}{\frac{1}{n}\sum_{t=1}^n \frac{1}{b} K\left(\frac{x-X_t}{b}\right)} \\ &\to {}^p\sigma^2(x),\end{aligned}$$

where e_t is the estimated OLS residual, and $K(\cdot)$ is a kernel function which is a specified symmetric density function (e.g., $K(u) = (2\pi)^{-1/2}\exp(-\frac{1}{2}u^2)$ if x is a scalar), and $b = b(n)$ is a bandwidth such that $b \to 0, nb \to \infty$ as $n \to \infty$. The finite sample distribution of $\hat{\beta}_a^*$ will be different from the finite sample distribution of $\hat{\beta}^*$, which assumes that V were known. This is because the sampling errors of the estimator $\hat{V}$ have some impact on the estimator $\hat{\beta}_a^*$. However, under some suitable conditions on $\hat{V}$, $\hat{\beta}_a^*$ will share the same asymptotic property as the infeasible GLS $\hat{\beta}^*$ (i.e., the MSE of $\hat{\beta}_a^*$ is approximately equal to the MSE of $\hat{\beta}^*$). In other words, the first stage estimation of $\sigma^2(\cdot)$ has no impact on the asymptotic distribution of $\hat{\beta}_a^*$. For more discussion, see Robinson (1988) and Stinchcombe and White (1991).

(ii) Second Approach

Continue to use OLS $\hat{\beta}$, obtaining the correct formula for

$$\text{var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

as well as a consistent estimator for $\text{var}\left(\hat{\beta}|\mathbf{X}\right)$. The classical definitions of t and F-tests cannot be used, because they are based on an incorrect formula for $\text{var}(\hat{\beta}|\mathbf{X})$. However, some modified tests can be obtained by using a consistent estimator for the correct formula for $\text{var}(\hat{\beta}|\mathbf{X})$. The trick is to estimate $\sigma^2\mathbf{X}'V\mathbf{X}$, which is a $K \times K$ nknown matrix, rather than to estimate V, which is a $n \times n$ unknown matrix. However, only asymptotic distributions can be used in this case.

Question: Suppose we assume

$$\begin{aligned}E(\varepsilon\varepsilon'|\mathbf{X}) &= \sigma^2 V \\ &= \text{diag}\{\sigma_1^2(\mathbf{X}), ..., \sigma_n^2(\mathbf{X})\}.\end{aligned}$$

As pointed out earlier, this essentially assumes $E(\varepsilon_t\varepsilon_s|\mathbf{X}) = 0$ for all $t \neq s$. That is, there is no serial correlation in $\{\varepsilon_t\}$ conditional on $\mathbf{X}$. Instead of estimating $\sigma_t^2(\mathbf{X})$, one can estimate the $K \times K$ matrix $\sigma^2\mathbf{X}'V\mathbf{X}$ directly.

Then, how to estimate

$$\sigma^2 \mathbf{X}'V\mathbf{X} = \sum_{t=1}^{n} X_t X_t' \sigma_t^2(\mathbf{X})?$$

We can use the folllowing estimator

$$\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X} = \sum_{t=1}^{n} X_t X_t' e_t^2,$$

where $\mathrm{D}(e) = \mathrm{diag}(e_1, ..., e_n)$ is a $n \times n$ diagonal matrix with all off-diagonal elements being zero. This is called White's (1980) heteroskedasticity-consistent variance-covariance estimator. See more discussion in Chapter 4.

Question: For $J = 1$, do we have

$$\frac{R\hat{\beta} - r}{\sqrt{R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'}} \sim t_{n-K}?$$

For $J > 1$, do we have

$$\begin{aligned} & (R\hat{\beta} - r)' \left[R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1} (R\hat{\beta} - r)/J \\ \sim \quad & F_{J,n-K}? \end{aligned}$$

No. Although we have standardized both test statistics by the correct variance estimators, we still have $\mathrm{cov}(\hat{\beta}, e|\mathbf{X}) \neq 0$ under Assumption 3.6. This implies that $\hat{\beta}$ and e are not independent, and therefore, we no longer have a t-distribution or an F-distribution in finite samples.

However, when $n \to \infty$, we have
(i) Case I ($J = 1$) :

$$\frac{R\hat{\beta} - r}{\sqrt{R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'}} \to^d N(0,1).$$

This can be called a robust t-test.
(ii) Case II ($J > 1$) :

$$(R\hat{\beta} - r)' \left[R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1} (R\hat{\beta} - r) \to^d \chi_J^2.$$

This is a robust Wald test statistic.

The above two feasible solutions are based on the assumption that $E(\varepsilon_t\varepsilon_s|\mathbf{X}) = 0$ for all $t \neq s$.

In fact, we can also consistently estimate the limit of $X'VX$ when there exists conditional heteroskedasticity and autocorrelation. This is called heteroskedasticity and autocorrelation consistent variance-covariance estimation. When there exists serial correlation of unknown form, an alternative solution should be provided. This is discussed in Chapter 6. See also Andrews (1991) and Newey and West (1987, 1994).

3.10 Summary and Conclusion

In this chapter, we have presented the econometric theory for the classical linear regression models. We first provide and discuss a set of assumptions on which the classical linear regression model is built. This set of regularity conditions will serve as the starting points from which we will develop modern econometric theory for linear regression models.

We derive the statistical properties of the OLS estimator. In particular, we point out that R^2 is not a suitable model selection criterion, because it is always nondecreasing with the dimension of regressors. Suitable model selection criteria, such as AIC and BIC, are discussed. We show that conditional on the regressor matrix $\mathbf{X}$, the OLS estimator $\hat{\beta}$ is unbiased, has a vanishing variance, and is BLUE. Under the additional conditional normality assumption, we derive the finite sample normal distribution for $\hat{\beta}$, the Chi-squared distribution for $(n-K)s^2/\sigma^2$, as well as the independence between $\hat{\beta}$ and s^2.

Many hypotheses encountered in economics can be formulated as linear restrictions on model parameters. Depending on the number of parameter restrictions, we derive the t-test and the F-test. In the special case of testing the hypothesis that all slope coefficients are jointly zero, we also derive an asymptotically Chi-squared test based on R^2.

When there exist conditional heteroskedasticity and/or autocorrelation, the OLS estimator is still unbiased and has a vanishing variance, but it is no longer BLUE, and $\hat{\beta}$ and s^2 are no longer mutually independent. Under the assumption of a known variance-covariance matrix up to some scale parameter, one can transform the linear regression model by correcting conditional heteroskedasticity and eliminating autocorrelation, so that the transformed regression model has conditionally homoskedastic and uncorrelated errors. The OLS estimator of this transformed linear regression model is called the GLS estimator, which is BLUE. The t-test and F-test are applicable. When the variance-covariance structure is unknown, the GLS estimator becomes infeasible. However, if the

error in the original linear regression model is serially uncorrelated (as is the case with independent observations across t), there are two feasible solutions. The first is to use a nonparametric method to obtain a consistent estimator for the conditional variance $\text{var}(\varepsilon_t|X_t)$, and then obtain a feasible plug-in GLS. The second is to use White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator for the OLS estimator $\hat{\beta}$. Both of these two methods are built on the asymptotic theory. When the error of the original linear regression model is serially correlated, a feasible solution to estimate the variance-covariance matrix is provided in Chapter 6.

References

Andrews, D. (1991)

Newey, W. and K. West (1987)

Newey, W. and K. West (1994)

Robinson, P. (1988)

Stinchombe, M. and H. White (1991)

EXERCISES

3.1. Consider a bivariate linear regression model

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

where $X_t = (X_{0t}, X_{1t})' = (1, X_{1t})'$, and ε_t is a regression error.

(a) Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ be the OLS estimator. Show that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1$, and

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{t=1}^n (X_{1t} - \bar{X}_1)(Y_t - \bar{Y})}{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2} \\ &= \frac{\sum_{t=1}^n (X_{1t} - \bar{X}_1) Y_t}{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2} \\ &= \sum_{t=1}^n C_t Y_t, \end{aligned}$$

where $C_t = (X_{1t} - \bar{X}_1)/\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2$.

(b) Suppose $\mathbf{X} = (X_{11}, ..., X_{1n})'$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ are independent. Show that $\text{var}(\hat{\beta}_1 | \mathbf{X}) = \sigma_\varepsilon^2/[(n-1)S_{X_1}^2]$, where $S_{X_1}^2$ is the sample variance of $\{X_{1t}\}_{t=1}^n$ and σ_ε^2 is the variance of ε_t.Thus, the more variations in $\{X_{1t}\}$, the more accurate estimation for β_1^o.

(c) Let $\hat{\rho}$ denote the sample correlation between Y_t and X_{1t}; namely,

$$\hat{\rho} = \frac{\sum_{t=1}^n (X_{1t} - \bar{X}_1)(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2 \sum_{t=1}^n (Y_t - \bar{Y})^2}}.$$

Show that $R^2 = \hat{\rho}^2$.Thus, the squared sample correlation between Y and X_1 is the fraction of the sample variation in Y that can be predicted using the linear predictor of X_1.This resuly also implies that R^2 is a measure of the strength of sample linear association between Y_t and X_{1t}.

3.2. For the OLS estimation of the linear regression model $Y_t = X_t'\beta^o + \varepsilon$, where X_t is a $K \times 1$ vector, show $R^2 = \hat{\rho}_{Y\hat{Y}}^2$, the squared sample correlation between Y_t and $\hat{Y}_t$.

3.2. Suppose $X_t = Q$ for all $t \geq m$, where m is a fixed integer, and Q is a $K \times 1$ constant vector. Do we have $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$? Explain.

3.3. The adjusted R^2, denoted as $\bar{R}^2$, is defined as follows:

$$\bar{R}^2 = 1 - \frac{e'e/(n-K)}{(Y - \bar{Y})'(Y - \bar{Y})/(n-1)}.$$

Show

$$\bar{R}^2 = 1 - \left[\frac{n-1}{n-K}(1 - R^2)\right].$$

3.4. [Effect of Multicolinearity] Consider a regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t.$$

Suppose Assumptions 3.1–3.4 hold. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$ be the OLS estimator. Show

$$\begin{aligned} \text{var}(\hat{\beta}_1|X) &= \frac{\sigma^2}{(1-\hat{r}^2)\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2}, \\ \text{var}(\hat{\beta}_2|X) &= \frac{\sigma^2}{(1-\hat{r}^2)\sum_{t=1}^n (X_{2t} - \bar{X}_2)^2}, \end{aligned}$$

where $\bar{X}_1 = n^{-1}\sum_{t=1}^n X_{1t}, \bar{X}_2 = n^{-1}\sum_{t=1}^n X_{2t}$, and

$$\hat{r}^2 = \frac{\left[\sum_{t=1}^n (X_{1t} - \bar{X}_1)(X_{2t} - \bar{X}_2)\right]^2}{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2 \sum_{t=1}^n (X_{2t} - \bar{X}_2)^2}.$$

3.5. Consider the linear regression model

$$Y_t = X_t'\beta^o + \varepsilon_t,$$

where $X_t = (1, X_{1t}, ..., X_{kt})'$. Suppose Assumptions 31.–3.3 hold. Let R_j^2 is the coefficient of determination of regressing variable X_{jt} on all the other explanatory variables $\{X_{it}, 0 \le i \le k, i \ne j\}$. Show

$$\text{var}(\hat{\beta}_j|\mathbf{X}) = \frac{\sigma^2}{(1-R_j^2)\sum_{t=1}^n (X_{jt} - \bar{X}_j)^2},$$

where $\bar{X}_j = n^{-1}\sum_{t=1}^n X_{jt}$. The factor $1/(1-R_j^2)$ is called the variance inflation factor (VIF); it is used to measure the degree of multicolinearity among explanatory variables in X_t.

3.6. Consider the following linear regression model

$$Y_t = X_t'\beta^o + u_t, \qquad t = 1, ..., n, \tag{4.1}$$

where

$$u_t = \sigma(X_t)\varepsilon_t,$$

where $\{X_t\}$ is a nonstochastic process, and $\sigma(X_t)$ is a positive function of X_t such that

$$\Omega = \begin{bmatrix} \sigma^2(X_1) & 0 & 0 & ... & 0 \\ 0 & \sigma^2(X_2) & 0 & ... & 0 \\ 0 & 0 & \sigma^2(X_3) & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \sigma^2(X_n) \end{bmatrix} = \Omega^{\frac{1}{2}}\Omega^{\frac{1}{2}},$$

with

$$\Omega^{\frac{1}{2}} = \begin{bmatrix} \sigma(X_1) & 0 & 0 & ... & 0 \\ 0 & \sigma(X_2) & 0 & ... & 0 \\ 0 & 0 & \sigma(X_3) & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \sigma(X_n) \end{bmatrix}.$$

Assume that $\{\varepsilon_t\}$ is i.i.d. $N(0,1)$. Then $\{u_t\}$ is i.i.d. $N(0, \sigma^2(X_t))$. This differs from Assumption 3.5 of the classical linear regression analysis, because now $\{u_t\}$ exhibits conditional heteroskedasticity.

Let $\hat{\beta}$ denote the OLS estimator for β^o.

(a) Is $\hat{\beta}$ unbiased for β^o?

(b) Show that $\text{var}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.

Consider an alternative estimator

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}Y \\ &= \left[\sum_{t=1}^{n} \sigma^{-2}(X_t)X_tX_t'\right]^{-1} \sum_{t=1}^{n} \sigma^{-2}(X_t)X_tY_t. \end{aligned}$$

(c) Is $\tilde{\beta}$ unbiased for β^o?

(d) Show that $\text{var}(\tilde{\beta}) = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}$.

(e) Is $\text{var}(\hat{\beta}) - \text{var}(\tilde{\beta})$ positive semi-definite (p.s.d.)? Which estimator, $\hat{\beta}$ or $\tilde{\beta}$, is more efficient?

(f) Is $\tilde{\beta}$ the Best Linear Unbiased Estimator (BLUE) for β^o? [Hint: There are several approaches to this question. A simple one is to consider the transformed model

$$Y_t^* = X_t^{*\prime}\beta^o + \varepsilon_t, \qquad t = 1, ..., n, \tag{4.2}$$

where $Y_t^* = Y_t/\sigma(X_t), X_t^* = X_t/\sigma(X_t)$. This model is obtained from model (4.1) after dividing by $\sigma(X_t)$. In matrix notation, model (4.2) can be written as

$$Y^* = \mathbf{X}^*\beta^o + \varepsilon,$$

where the $n \times 1$ vector $Y^* = \Omega^{-\frac{1}{2}}Y$ and the $n \times k$ matrix $\mathbf{X}^* = \Omega^{-\frac{1}{2}}\mathbf{X}$.]

(g) Construct two test statistics for the null hypothesis of interest $\mathbf{H}_0 : \beta_2^o = 0$. One test is based on $\hat{\beta}$, and the other test is based on $\tilde{\beta}$. What are the finite sample distributions of your test statistics under $\mathbf{H}_0$? Can you tell which test is better?

(h) Construct two test statistics for the null hypothesis of interest $\mathbf{H}_0 : R\beta^o = r$, where R is a $J \times k$ matrix with $J > 0$. One test is based on $\hat{\beta}$, and the other test is based on $\tilde{\beta}$. What are the finite sample distributions of your test statistics under $\mathbf{H}_0$?

3.7. Consider the following classical regression model

$$Y_t = X_t'\beta^o + \varepsilon_t.$$

Suppose that we are interested in testing the null hypothesis

$$\mathbf{H}_0 : R\beta^o = r,$$

where R is a $J \times K$ matrix, and r is a $J \times 1$ vector. The F-test statistic is defined as

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2}.$$

Show that

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/J}{e'e/(n-k-1)}.$$

where $e'e$ is the sum of squared residuals from the unrestricted model, and $\tilde{e}'\tilde{e}$ is the sum of squared residuals from the restricted regression model subject to the restriction $R\beta = r$.

3.8. The F-test statistic is defined as follows:

$$F = \frac{(R\hat{\beta} - r)'[R(\mathbf{X'X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2}.$$

Show that

$$\begin{aligned} F &= \frac{\sum_{t=1}^{n}(\hat{Y}_t - \tilde{Y}_t)^2/J}{s^2} \\ &= \frac{(\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})/J}{s^2}, \end{aligned}$$

where $\hat{Y}_t = X_t'\hat{\beta}, \tilde{Y}_t = X_t'\tilde{\beta}$, and $\hat{\beta}, \tilde{\beta}$ are the unrestricted and restricted OLS estimators respectively.

3.9. Consider the following classical regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \beta_0^o + \sum_{j=1}^{k}\beta_j^o X_{jt} + \varepsilon_t, \qquad t = 1, ..., n. \end{aligned} \tag{7.1}$$

Suppose that we are interested in testing the null hypothesis

$$\mathbf{H}_0 : \beta_1^o = \beta_2^o = \cdots = \beta_k^o = 0.$$

Then the F-statistic can be written as

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/k}{e'e/(n-k-1)}.$$

where $e'e$ is the sum of squared residuals from the unrestricted model (7.1), and $\tilde{e}'\tilde{e}$ is the sum of squared residuals from the restricted model (7.2)

$$Y_t = \beta_0^o + \varepsilon_t. \tag{7.2}$$

(a) Show that under Assumptions 3.1 and 3.3,

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)},$$

where R^2 is the coefficient of determination of the unrestricted model (7.1).

(b) Suppose in addition Assumption 3.5 holds. Show that under $\mathbf{H}_0$,

$$(n-k-1)R^2 \xrightarrow{d} \chi_k^2.$$

3.10. The F-test statistic is defined as folllows:

$$F = \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2}.$$

Show that F

$$F = \frac{(1/J)\sum_{t=1}^{n}(\hat{Y}_t - \tilde{Y}_t)^2}{s^2} = \frac{(\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})/J}{s^2},$$

where $\hat{Y}_t = X_t'\hat{\beta}, \tilde{Y}_t = X_t'\tilde{\beta}$, and $\hat{\beta}, \tilde{\beta}$ are the unrestricted and restricted OLS estimators respectively.

3.11. **[Structral Change]** Suppose Assumptions 3.1 and 3.3 hold. Consider the following model on the whole sample:

$$Y_t = X_t'\beta^o + (D_t X_t)'\alpha^o + \varepsilon_t, t = 1, ..., n,$$

where the time dummy variable $D_t = 0$ if $t \leq n_1$ and $D_t = 1$ if $t > n_1$. This model can be written as two separate models:

$$Y_t = X_t'\beta^o + \varepsilon_t, t = 1, ..., n_1$$

and

$$Y_t = X_t'(\beta^o + \alpha^o) + \varepsilon_t, t = n_1 + 1, ..., n.$$

Let SSR_u, SSR_1, SSR_2 denotes the sums of squared residuals of the above three regression models via OLS. Show

$$SSR_u = SSR_1 + SSR_2.$$

This identity implies that estimating the first regression mdoel with time dummy variable D_t via OLS is equivalent to estimating two separate regression models over two subsample periods respectively.

3.12. Suppose $\mathbf{X}'\mathbf{X}$ is a $K \times K$ matrix, and V is a $n \times n$ matrix, and both $\mathbf{X}'\mathbf{X}$ and V are symmetric and nonsingular, with the minimum eigenvalue $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ and $0 < c \leq \lambda_{\max}(V) \leq C < \infty$. Show that for any $\tau \in R^K$ such that $\tau'\tau = 1$,

$$\tau' \text{var}(\hat{\beta}|\mathbf{X})\tau = \sigma^2 \tau'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\tau \to 0$$

as $n \to \infty$. Thus, $\text{var}(\hat{\beta}|\mathbf{X})$ vanishes to zero as $n \to \infty$ under conditional heteroskedasticity.

3.13. Suppose the conditions in 3.9 hold. It can be shown that the variances of the OLS $\hat{\beta}$ and GLS $\hat{\beta}^*$ are respectively:

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \\ \text{var}(\hat{\beta}^*|\mathbf{X}) &= \sigma^2(\mathbf{X}'V^{-1}\mathbf{X})^{-1}. \end{aligned}$$

Show that $\text{var}(\hat{\beta}|\mathbf{X}) - \text{var}(\hat{\beta}^*|\mathbf{X})$ is positive semi-definite.

3.14. Suppose a data generating process is given by

$$Y_t = \beta_1^o X_{1t} + \beta_2^o X_{2t} + \varepsilon_t = X_t'\beta^o + \varepsilon_t,$$

where $X_t = (X_{1t}, X_{2t})'$, $E(X_t X_t')$ is nonsingular, and $E(\varepsilon_t|X_t) = 0$. For simplicity, we further assume $E(X_{2t}) = 0$ and $E(X_{1t}X_{2t}) \neq 0$.

Now consider the following bivariate linear regression model

$$Y_t = \beta_1^o X_{1t} + u_t.$$

(a) Show that if $\beta_2^o \neq 0$, then $E(Y_1|X_t) = X_t'\beta^o \neq E(Y_{1t}|X_{1t})$. That is, there exists an omitted variable (X_{2t}) in the bivariate regression model.

(b) Show that $E(Y_t|X_{1t}) \neq \beta_1 X_{1t}$ for all β_1. That is, the bivariate linear regression model is misspecified for $E(Y_t|X_{1t})$.

(c) Is the best linear least squares approximation cofficient β_1^* in the bivariate linear regression model equal to β_1^o?

3.15. Suppose a data generating process is given by

$$Y_t = \beta_1^o X_{1t} + \beta_2^o X_{2t} + \varepsilon_t = X_t'\beta^o + \varepsilon_t,$$

where $X_t = (X_{1t}, X_{2t})'$, and Assumptions 3.1–3.4 hold. (For simplicity, we have assumed no intercept.) Denote the OLS estimator by $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$.

If $\beta_2^o = 0$ and we know it. Then we can consider a simpler regression

$$Y_t = \beta_1^o X_{1t} + \varepsilon_t.$$

Denote the OLS of this simpler regression as $\tilde{\beta}_1$.

Please compare the relative efficiency between $\hat{\beta}_1$ and $\tilde{\beta}_1$. That is, which estimator is better for β_1^o? Give your reasoning.

3.16. Suppose Assumption 3.6 is replaced by the following assumption:
Assumption 3.6′ : $\varepsilon|\mathbf{X} \sim N(0, V)$, *where* $V = V(\mathbf{X})$ *is a known* $n \times n$ *finite and positive definite matrix.*

Compared to Assumption 3.6, Assumption 3.6′ assumes that $\text{var}(\varepsilon|\mathbf{X}) = V$ is completely known and there is no unknown proportionality σ^2. Define GLS $\hat{\beta}^* = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}Y$.

(a) Is $\hat{\beta}^*$ BLUE?

(b) Put $X^* = CX$ and $s^{*2} = e^{*\prime}e^*/(n-K)$, where $e^* = Y - X^*\hat{\beta}^*$, $C'C = V$. Do the the usual t-test and F-test defined as

$$\begin{aligned} T^* &= \frac{R\hat{\beta}^* - r}{\sqrt{s^{*2}R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R'}}, \text{ for } J = 1, \\ F^* &= \frac{(R\hat{\beta}^* - r)'[R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R']^{-1}(R\hat{\beta}^* - r)/J}{s^{*2}} \end{aligned}$$

follow the t_{n-K} and $F_{J,n-K}$ distributions respectively under the null hypothesis that $R\beta = r$? Explain.

(c) Construct two new test statistics:

$$\begin{aligned} \tilde{T}^* &= \frac{R\hat{\beta}^* - r}{\sqrt{R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R'}}, \text{ for } J = 1, \\ \tilde{Q}^* &= (R\hat{\beta}^* - r)'[R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R']^{-1}(R\hat{\beta}^* - r). \end{aligned}$$

What distributions will these test statistics follow under the null hypothesis that $R\beta = r$? Explain.

(d) Which set of tests, (T^*, F^*) or $(\tilde{T}^*, \tilde{Q}^*)$, are more powerful at the same significance level? Explain. [*Hint: The t-distribution has a heavier tail than* $N(0,1)$ *and so has a larger critical value at a given significance level.*]

3.17. Consider a linear regression model

$$Y_t = X_t'\beta^0 + \varepsilon_t, \qquad t = 1, 2, ..., n,$$

where $\varepsilon_t = \sigma(X_t)v_t$, X_t is a $K \times 1$ nonstochastic vector, and $\sigma(X_t)$ is a positive function of X_t, and $\{v_t\}$ is i.i.d. $N(0,1)$.

Let $\hat{\beta} = (X'X)^{-1}X'Y$ denote the OLS estimator for β^0, where X is a $n \times K$ matrix whose t-th row is X_t, and Y is a $n \times 1$ vector whose t-th component is Y_t.

(a) Is $\hat{\beta}$ unbiased for β^0?

(b) Find $\text{var}(\hat{\beta}) = E[(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})']$. You may find the following notation useful: $\Omega = \text{diag}\{\sigma^2(X_1), \sigma^2(X_2), ..., \sigma^2(X_n)\}$, i.e., Ω is a $n \times n$ diagonal matrix with the t-th diagonal component equal to $\sigma^2(X_t)$ and all off-diagonal components equal to zero.

Consider the transformed regression model

$$\frac{1}{\sigma(X_t)}Y_t = \frac{1}{\sigma(X_t)}X_t'\beta^0 + v_t$$

or

$$Y_t^* = X_t^{*\prime}\beta^0 + v_t,$$

where $Y_t^* = \sigma^{-1}(X_t)Y_t$ and $X_t^* = \sigma^{-1}(X_t)X_t$.

Denote the OLS estimator of this transformed model as $\tilde{\beta}$.

(c) Show

$$\tilde{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.$$

(d) Is $\tilde{\beta}$ unbiased for β^0?

(e) Find $\text{var}(\tilde{\beta})$.

(f) Which estimator, $\hat{\beta}$ or $\tilde{\beta}$, is more efficient in terms of the mean squared error criterion? Give your reasoning.

(g) Use the difference $R\tilde{\beta} - r$ to construct a test statistic for the null hypothesis of interest $\mathbf{H}_0 : R\beta^0 = r$, where R is a $J \times K$ matrix, r is $K \times 1$, and $J > 1$. What is the finite sample distribution of your test statistic under $\mathbf{H}_0$?