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 a 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:
(i) \( K \equiv k + 1 \) is the number of regressors for which there are \( k \) economic variables and an intercept.
(ii) 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, week, month, year) in a time series context.

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

Assumption 3.1 [Linearity]:

\[ Y_t = X'_t \beta^0 + \varepsilon_t, \quad t = 1, ..., n, \]

where \( \beta^0 \) is a \( K \times 1 \) unknown parameter vector, and \( \varepsilon_t \) is an unobservable disturbance.

Remarks:
(i) In Assumption 3.1, \( Y_t \) is the dependent variable (or regresand), \( X_t \) is the vector of regressors (or independent variables, or explanatory variables), and \( \beta^0 \) is the regression
coefficient vector. When the linear model is correctly for the conditional mean $E(Y_t|X_t)$, i.e., when $E(\varepsilon_t|X_t) = 0$, the parameter $\beta^0 = \frac{\partial}{\partial X_t} E(Y_t|X_t)$ is the marginal effect of $X_t$ on $Y_t$.

(ii) The key notion of linearity in the classical linear regression model is that the regression model is linear in $\beta^0$ rather than in $X_t$.

(iii) 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?

(iv) Matrix Notation: Denote

$$Y = (Y_1,...,Y_n)', \quad n \times 1,$$
$$\varepsilon = (\varepsilon_1,...,\varepsilon_n)', \quad n \times 1,$$
$$X = (X_1,...,X_n)', \quad n \times K.$$

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

$$Y = X\beta^0 + \varepsilon,$$
$$n \times 1 = (n \times K)(K \times 1) + n \times 1.$$

The second assumption is a strict exogeneity condition.

**Assumption 3.2 [Strict Exogeneity]:**

$$E(\varepsilon_t|X) = E(\varepsilon_t|X_1,...,X_t,...,X_n) = 0, \quad t = 1,...,n.$$

**Remarks:**

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

(ii) Assumption 3.2 also implies $E(\varepsilon_t) = 0$ by the law of iterated expectations.

(iii) 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

$$E(X_s\varepsilon_t) = E[E(X_s\varepsilon_t|X)]$$
$$= E[X_sE(\varepsilon_t|X)]$$
$$= E(X_s \cdot 0)$$
$$= 0.$$
Note that (i) and (ii) imply $\text{cov}(X_s, \varepsilon_t) = 0$ for all $t, s \in \{1, ..., n\}$.

(iv) Because $X$ contains regressors $\{X_s\}$ for both $s \leq t$ and $s > t$, Assumption 3.2 essentially requires that the error $\varepsilon_t$ 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 $\varepsilon_t$ 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

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t, \quad t = 1, ..., n,$$

$$\{\varepsilon_t\} \sim \text{i.i.d.}(0, \sigma^2),$$

where $X_t = (1, Y_{t-1})'$. 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 $E(\varepsilon_t | X) \neq 0$, and so Assumption 3.2 does not hold. In Chapter 5, 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.

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

(vi) **Question:** What happens to Assumption 3.2 if $X$ is nonstochastic?

**Answer:** If $X$ is nonstochastic, Assumption 3.2 becomes

$$E(\varepsilon_t | X) = E(\varepsilon_t) = 0.$$  

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

$$Y_t = X_t' \beta^0 + \varepsilon_t = \sum_{j=0}^{k} \beta^0_j t^j + \varepsilon_t.$$  

(vii) **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)?
Answer: When \( \{Z_t\} \) is i.i.d., Assumption 3.2 becomes

\[
E(\varepsilon_t | X) = E(\varepsilon_t | X_1, X_2, ..., X_t, ..., X_n) = E(\varepsilon_t | X_t) = 0.
\]

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

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

\[
\lambda_{\text{min}} (X'X) \to \infty \text{ as } n \to \infty
\]

*with probability one.*

**Question:** How to find eigenvalues of a square matrix \( A \)?

**Answer:** By solving 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 \( A \).

**Remarks:**

(i) Assumption 3.3 rules out multicollinearity among the \( (k+1) \) regressors in \( X_t \). We say that there exists multicollinearity 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.

(ii) This condition implies that \( X'X \) is nonsingular, or equivalently, \( X \) 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.

(iii) It is well-known that the eigenvalue \( \lambda \) 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).

(iv) 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 \( X \) causes a change in \( Y \)). In certain sense, one may call \( X'X \) the “information matrix” of the random sample \( X \) because it is a
measure of the information contained in \( X \). The degree of variation in \( X'X \) will affect the preciseness of parameter estimation for \( \beta^0 \).

**Question:** Why can the eigenvalue \( \lambda \) used as a measure of the information contained in \( X'X \)?

**Assumption 3.4 [Spherical error variance]:**

(a) [conditional homoskedasticity]:

\[
E(\varepsilon_t^2 | X) = \sigma^2 > 0, \quad t = 1, ..., n;
\]

(b) [conditional serial uncorrelatedness]:

\[
E(\varepsilon_t \varepsilon_s | X) = 0, \quad t \neq s, t, s \in \{1, ..., n\}.
\]

**Remarks:**

(i) We can write Assumption 3.4 as

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

where \( \delta_{ts} = 1 \) if \( t = s \) and \( \delta_{ts} = 0 \) otherwise. In mathematics, \( \delta_{ts} \) is called the Kronecker delta function.

(ii) \( \text{var}(\varepsilon_t | X) = E(\varepsilon_t^2 | X) - [E(\varepsilon_t | X)]^2 = E(\varepsilon_t^2 | X) = \sigma^2 \) given Assumption 3.2. Similarly, we have \( \text{cov}(\varepsilon_t, \varepsilon_s | X) = E(\varepsilon_t \varepsilon_s | X) = 0 \) for all \( t \neq s \).

(iii) By the law of iterated expectations, Assumption 3.4 implies that \( \text{var}(\varepsilon_t) = \sigma^2 \) for all \( t = 1, ..., n \), the so-called unconditional homoskedasticity. Similarly, \( \text{cov}(\varepsilon_t, \varepsilon_s) = 0 \) for all \( t \neq s \).

(iv) A compact expression for Assumptions 3.2 and 3.4 together is:

\[
E(\varepsilon_t | X) = 0 \quad \text{and} \quad E(\varepsilon_t \varepsilon_s | X) = \sigma^2 I
\]

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

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

**3.2 OLS Estimation**

**Question:** How to estimate \( \beta^0 \) 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 hold. Define the sum of squared residuals (SSR) of the linear regression model \( Y_t = X'_t \beta + u_t \) as

\[
SSR(\beta) = (Y - X \beta)'(Y - X \beta) = \sum_{t=1}^{n} (Y_t - X'_t \beta)^2.
\]

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

\[
\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^k} SSR(\beta).
\]

Remark: \( SSR(\beta) \) is the sum of squared model errors \( \{u_t = Y_t - X'_t \beta\} \) for observations, 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

\[
\hat{\beta} = (X'X)^{-1}X'Y = \left( \sum_{t=1}^{n} X_tX'_t \right)^{-1} \sum_{t=1}^{n} X_tY_t = \left( \frac{1}{n} \sum_{t=1}^{n} X_tX'_t \right)^{-1} \frac{1}{n} \sum_{t=1}^{n} X_tY_t.
\]

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

Remark: Suppose \( Z_t = \{Y_t, X'_t\}, t = 1, \ldots, n \), is an independent and identically distributed (i.i.d.) random sample of size \( n \). Compare the population MSE criterion

\[
MSE(\beta) = E(Y_t - X'_t \beta)^2
\]

and the sample MSE criterion

\[
\frac{SSR(\beta)}{n} = \frac{1}{n} \sum_{t=1}^{n} (Y_t - X'_t \beta)^2.
\]

By the weak law of large numbers (WLLN), we have

\[
\frac{SSR(\beta)}{n} \rightarrow^p E(Y_t - X'_t \beta)^2 \equiv MSE(\beta)
\]
and their minimizers

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

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

Here, \( SSR(\beta) \), 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 \( \beta^* \).

**Proof of the Theorem:** Using the formula that for both \( a \) and \( \beta \) being vectors, the derivative

\[
\frac{\partial (a' \beta)}{\partial \beta} = a,
\]

we have

\[
\frac{dSSR(\beta)}{d\beta} = \frac{d}{d\beta} \sum_{t=1}^{n} (Y_t - X_t' \beta)^2
\]

\[
= \sum_{t=1}^{n} \frac{\partial}{\partial \beta} (Y_t - X_t' \beta)^2
\]

\[
= \sum_{t=1}^{n} 2(Y_t - X_t' \beta) \frac{\partial}{\partial \beta} (Y_t - X_t' \beta)
\]

\[
= -2 \sum_{t=1}^{n} X_t (Y_t - X_t' \beta)
\]

\[
= -2 X'(Y - X \beta).
\]

The OLS must satisfy the FOC:

\[-2 X'(Y - X \hat{\beta}) = 0, \]

\[X'(Y - X \hat{\beta}) = 0, \]

\[X'Y - (X'X)\hat{\beta} = 0.\]

It follows that

\[(X'X)\hat{\beta} = X'Y.\]

By Assumption 3.3, \( X'X \) is nonsingular. Thus,

\[\hat{\beta} = (X'X)^{-1} X'Y.\]

Checking the SOC, we have the \( K \times K \) Hessian matrix

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

\[
= 2 X'X
\]

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

Remarks:

(i) $\hat{Y}_t \equiv X'_t\hat{\beta}$ 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

$$
e_t = Y_t - \hat{Y}_t = X'_t(\beta - \beta^0)$$

where the true disturbance $\epsilon_t$ is unavoidable, and the estimation error $X'_t(\hat{\beta} - \beta^0)$ can be made smaller when a larger data set is available (so $\hat{\beta}$ becomes closer to $\beta^0$).

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

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

This is the consequence of the very nature of OLS, as implied by the FOC of $\min_{\beta \in \mathbb{R}^k} SSR(\beta)$. It always holds no matter whether $E(\epsilon_t|X) = 0$ (recall that we do not impose Assumption 3.2). Note that if $X_t$ contains the intercept, then $X'e = 0$ 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, we have:

(i) $X'e = 0$;

(ii) $\hat{\beta} - \beta^0 = (X'X)^{-1}X'\epsilon$;

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

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

and

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

$$PX = X, \quad MX = 0.$$  

(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 $\hat{\beta} = (X'X)^{-1}X'Y$ and $Y = X\beta^0 + \varepsilon$, we have 

$$\hat{\beta} - \beta^0 = (X'X)^{-1}X'(X\beta^0 + \varepsilon) - \beta^0 = (X'X)^{-1}X'\varepsilon.$$ 

(iii) $P$ is idempotent because 

$$P^2 = PP \\
= [X(X'X)^{-1}X']X(X'X)^{-1}X' \\
= X(X'X)^{-1}X' \\
= P.$$ 

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

(iv) By the definition of $M$, we have 

$$e = Y - X\hat{\beta} \\
= Y - X(X'X)^{-1}X'Y \\
= [I - X(X'X)^{-1}X']Y \\
= MY \\
= M(X\beta^0 + \varepsilon) \\
= MX\beta^0 + M\varepsilon \\
= M\varepsilon$$

given $MX = 0$. It follows that 

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

9
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 $\{Y_t\}$?

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

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

**Remarks:**

Interpretation of $R^2_{uc}$: The proportion of the uncentered sample quadratic variation in the dependent variables $\{Y_t\}$ that can be explained by 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:**

(i) When $X_t$ contains the intercept, we have the following orthogonal decomposition:

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

$$+ 2 \sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})e_t = \sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^{n} e_t^2,$$
where the cross-product term

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

\[ = \hat{\beta}' \sum_{t=1}^{n} X_t e_t - \bar{Y} \sum_{t=1}^{n} e_t \]

\[ = \hat{\beta}' (X'e) - \bar{Y} \sum_{t=1}^{n} e_t \]

\[ = \hat{\beta}' \cdot 0 - \bar{Y} \cdot 0 \]

\[ = 0, \]

where we have made use of the facts that \( X'e = 0 \) 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

\[ R^2 \equiv 1 - \frac{e'e}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \]

\[ = \frac{\sum_{t=1}^{n} (Y_t - \bar{Y})^2 - \sum_{t=1}^{n} e_t^2}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \]

\[ = \frac{\sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} . \]

and consequently we have

\[ 0 \leq R^2 \leq 1. \]

(ii) 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.

(iii) For any given random sample \( \{Y_t, X_t'\}', t = 1, \ldots, n \), \( R^2 \) 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 $R^2$. This is always true no matter whether $X_t$ has any true explanatory power for $Y_t$.

**Theorem:** Suppose $\{Y_t, X_t'\}_t, t = 1, ..., n$, is a random sample, $R^2_1$ is the centered $R^2$ from the linear regression

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

where $X_t = (1, X_{1t}, ..., X_{kt})'$, and $\beta$ 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 $\gamma$ is a $(K + q) \times 1$ parameter vector. Then $R^2_2 \geq R^2_1$.

**Proof:** By definition, we have

$$R^2_1 = 1 - \frac{e'e}{\sum_{t=1}^n (Y_t - \bar{Y})^2},$$

$$R^2_2 = 1 - \frac{\tilde{e}'\tilde{e}}{\sum_{t=1}^n (Y_t - \bar{Y})^2},$$

where $e$ is the estimated residual vector from the regression of $Y$ on $X$, and $\tilde{e}$ is the estimated residual vector from the regression of $Y$ on $\tilde{X}$. It suffices to show $\tilde{e}'\tilde{e} \leq e'e$. Because the OLS estimator $\hat{\gamma} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$ 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 $\hat{\beta} = (X'X)^{-1}X'Y$ is the OLS from the first regression. It follows that

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

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

$$= e'e.$$

Hence, we have $R^2_1 \leq R^2_2$. This completes the proof.

**Question:** What is the implication of this result?
**Answer:** $R^2$ is not a suitable criterion for correct model specification. It is not a suitable criterion of model selection.

**Question:** Does a high $R^2$ imply a precise estimation for $\beta^0$?

**Question:** What are the suitable model selection criteria?

**Two popular model selection criteria**

The **KISS Principle: Keep It Sophistically Simple!**

An important idea in statistics is to use a simple model to capture essential information contained in data as much as possible. 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 $K$:

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

$$\sim \text{ goodness of fit} + \text{model complexity}$$

where

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

and $K = k + 1$ is the number of regressors.

**Bayesian Information Criterion [BIC]:**

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

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

**Remark:** When ln $n \geq 2$, 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.

**Remark:** BIC is strongly consistent in 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.

Remarks: 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

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

This differs from

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

It may be shown that

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

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?

Answer: 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 $\beta^0$ (consistency)?
- Is $\hat{\beta}$ the best estimator (efficiency)?
- What is the sampling distribution of $\hat{\beta}$ (normality)?
**Question:** What is the sampling distribution of $\hat{\beta}$?

**Answer:** 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 $Z_t = (Y_t, X_t)'$.

**Remark:** 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 3.1-3.4 hold. Then

(i) [Unbiaseness] $E(\hat{\beta}|X) = \beta^0$ and $E(\hat{\beta}) = \beta^0$.

(ii) [Vanishing Variance]

\[
\text{var}(\hat{\beta}|X) = E\left[ (\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})'|X \right]
= \sigma^2(X'X)^{-1}
\]

and for any $K \times 1$ vector $\tau$ such that $\tau'\tau = 1$, we have

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

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

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

(iv) [Gauss-Markov]

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

for any unbiased estimator $\hat{b}$ that is linear in $Y$ with $E(\hat{b}|X) = \beta^0$.

(v) [Residual variance estimator]

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

is unbiased for $\sigma^2 = E(e^2)$. That is, $E(s^2|X) = \sigma^2$.

**Question:** What is a linear estimator $\hat{b}$?

**Remarks:**

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

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

15
where we have used the fact that \( \text{Bias}(\hat{\beta}|X) \equiv E(\hat{\beta}|X) - \beta^0 = 0 \). Recall that MSE measures how close an estimator \( \hat{\beta} \) is to the target \( \beta^0 \).

(b) Theorem (iv) implies that \( \hat{\beta} \) is the best linear unbiased estimator (BLUE) for \( \beta^0 \) because \( \text{var}(\hat{\beta}|X) \) is the smallest among all unbiased linear estimators for \( \beta^0 \).

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

**Definition [Efficiency]:** An unbiased estimator \( \hat{\beta} \) of parameter \( \beta^0 \) is more efficient than another unbiased estimator \( \hat{b} \) of parameter \( \beta^0 \) if

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

**Implication:** For any \( \tau \in \mathbb{R}^K \) such that \( \tau'\tau = 1 \), we have

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

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

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

**Proof:**

(i) Given \( \hat{\beta} - \beta^0 = (X'X)^{-1}X'\varepsilon \), we have

\[
E[(\hat{\beta} - \beta^0)|X] = E[(X'X)^{-1}X'\varepsilon|X]
\]

\[
= (X'X)^{-1}X'E(\varepsilon|X)
\]

\[
= (X'X)^{-1}X'0
\]

\[
= 0.
\]

(ii) Given \( \hat{\beta} - \beta^0 = (X'X)^{-1}X'\varepsilon \) and \( E(\varepsilon\varepsilon'|X) = \sigma^2 I \), we have

\[
\text{var}(\hat{\beta}|X) \equiv E \left[ (\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})'|X \right]
\]

\[
= E \left[ (\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)'|X \right]
\]

\[
= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X]
\]

\[
= (X'X)^{-1}X'E(\varepsilon\varepsilon'|X)X(X'X)^{-1}
\]

\[
= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}
\]

\[
= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}
\]

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

\[
\tau' \text{var}(\hat{\beta}|X)\tau = \sigma^2 \tau'(X'X)^{-1}\tau \leq \sigma^2 \lambda_{\text{max}}[(X'X)^{-1}] = \sigma^2 \lambda_{\text{min}}^{-1}(X'X) \rightarrow 0
\]

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

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

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

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 $C = C(X)$ is a $n \times K$ matrix depending on $X$. It is unbiased for $\beta^0$ regardless of the value of $\beta^0$ if and only if

\[
E(\hat{b}|X) = C'X\beta^0 + C'\varepsilon = C'X\beta^0 = \beta^0.
\]

This follows if and only if

\[
C'X = I.
\]
Because

\[
\hat{b} = C'Y \\
= C'(X\beta^0 + \varepsilon) \\
= C'X\beta^0 + C'\varepsilon \\
= \beta^0 + C'\varepsilon,
\]

the variance of \( \hat{b} \)

\[
\text{var}(\hat{b}) = E \left[ (\hat{b} - \beta^0)(\hat{b} - \beta^0)'|X \right] \\
= E \left[ C'\varepsilon \varepsilon' C|X \right] \\
= C' E(\varepsilon \varepsilon'|X)C \\
= C'\sigma^2 I C \\
= \sigma^2 C'C.
\]

Using \( C'X = I \), we now have

\[
\text{var}(\hat{b}|X) - \text{var}(\hat{\beta}|X) = \sigma^2 C'C - \sigma^2 (X'X)^{-1} \\
= \sigma^2\left[ C'C - C'X(X'X)^{-1} X'C \right] \\
= \sigma^2 C'[I - X(X'X)^{-1} X']C \\
= \sigma^2 C'MC \\
= \sigma^2 C'M'MC \\
= \sigma^2 C'M'M'C \\
= \sigma^2 (MC)'(MC) \\
= \sigma^2 D'D \\
= \sigma^2 \sum_{t=1}^{n} D_tD_t' \\
\sim \text{p.s.d.}
\]

where we have used the fact that for any real-valued vector \( D \), 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
\[ E(e'e|X) = E(\varepsilon'M\varepsilon|X) \]
\[ = E[\text{tr}(\varepsilon'M\varepsilon)|X] \]
[putting \( A = \varepsilon'M, B = \varepsilon \)]
\[ = E[\text{tr}(\varepsilon\varepsilon'M)|X] \]
\[ = \text{tr}[E(\varepsilon\varepsilon'|X)M] \]
\[ = \text{tr}(\sigma^2IM) \]
\[ = \sigma^2\text{tr}(M) \]
\[ = \sigma^2(n - K) \]

where
\[ \text{tr}(M) = \text{tr}(I_n) - \text{tr}(X(X'X)^{-1}X') \]
\[ = \text{tr}(I_n) - \text{tr}(X'X(X'X)^{-1}) \]
\[ = n - K, \]

using \( \text{tr}(AB) = \text{tr}(BA) \) again. It follows that
\[ E(s^2|X) = \frac{E(e'e|X)}{n - K} \]
\[ = \frac{\sigma^2(n - K)}{(n - K)} \]
\[ = \sigma^2. \]

This completes the proof. Note that the sample residual variance \( s^2 = e'e/(n - K) \) is a generalization of the sample variance \( S_n^2 = (n - 1)^{-1}\sum_{t=1}^n(Y_t - \hat{Y})^2 \) for the random sample \( \{Y_t\}_{t=1}^n \).

3.5 The Sampling Distribution of \( \hat{\beta} \)

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

**Assumption 3.5:** \( \varepsilon|X \sim N(0, \sigma^2I) \).

**Remarks:**
(i) Under Assumption 3.5, the conditional pdf of $\varepsilon$ given $X$ is
\[
    f(\varepsilon|X) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp \left( -\frac{\varepsilon' \varepsilon}{2\sigma^2} \right) = f(\varepsilon),
\]
which does not depend on $X$, the disturbance $\varepsilon$ is independent of $X$. Thus, every conditional moment of $\varepsilon$ given $X$ does not depend on $X$.

(ii) Assumption 3.5 implies Assumptions 3.2 ($E(\varepsilon|X) = 0$) and 3.4 ($E(\varepsilon'X) = \sigma^2 I$).

(iii) The normality 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 $\hat{\beta}$?

**Sampling Distribution of $\hat{\beta}$:**
\[
    \hat{\beta} - \beta^0 = (X'X)^{-1}X'\varepsilon = (X'X)^{-1} \sum_{t=1}^{n} X_t \varepsilon_t = \sum_{t=1}^{n} C_t \varepsilon_t,
\]
where the weighting matrix
\[
    C_t = (X'X)^{-1}X_t
\]
is called the leverage of observation $X_t$.

**Theorem [Normality of $\hat{\beta}$]:** Under Assumptions 3.1, 3.3 and 3.5,
\[
    (\hat{\beta} - \beta^0|X) \sim N[0, \sigma^2(X'X)^{-1}].
\]

**Proof:** Conditional on $X$, $\hat{\beta} - \beta^0$ is a weighted sum of independent normal random variables $\{\varepsilon_t\}$, and so is also normally distributed.

The corollary below follows immediately.

**Corollary [Normality of $R(\hat{\beta} - \beta^0)$]:** Suppose Assumptions 3.1, 3.3 and 3.5 hold. Then for any nonstochastic $J \times K$ matrix $R$, we have
\[
    R(\hat{\beta} - \beta^0)|X \sim N[0, \sigma^2 R(X'X)^{-1} R'^{\top}].
\]
Question: What is the role of the nonstochastic matrix $R$?

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

Proof: Conditional on $X$, $\hat{\beta} - \beta^0$ is normally distributed. Therefore, conditional on $X$, the linear combination $R(\hat{\beta} - \beta^0)$ is also normally distributed, with

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

and

$$\text{var}[R(\hat{\beta} - \beta^0)|X] = E\left[ R(\hat{\beta} - \beta^0)(R(\hat{\beta} - \beta^0))'|X \right]$$

$$= E\left[ R(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)'R'|X \right]$$

$$= RE\left[ (\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)'|X \right] R'$$

$$= R \text{var}(\hat{\beta}|X) R'$$

$$= \sigma^2 R(X'X)^{-1} R'.$$

It follows that

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

Question: Why would we like to know the sampling distribution of $R(\hat{\beta} - \beta^0)$?

Answer: This is mainly for confidence interval estimation and hypothesis testing.

Remark: However, $\text{var}[R(\hat{\beta} - \beta^0)|X] = \sigma^2 R(X'X)^{-1} R'$ is unknown, because $\text{var}(\epsilon_t) = \sigma^2$ is known. We need to estimate $\sigma^2$.

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

Question: How to estimate $\text{var}(\hat{\beta}|x) = \sigma^2(X'X)^{-1}$?

Answer: To estimate $\sigma^2$, we can use the residual variance estimator

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

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

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

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

(i) Question: What is a $\chi^2_q$ distribution?

Definition [Chi-square Distribution, $\chi^2_q$] 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 $\chi^2_q$ distribution.

Properties of $\chi^2_q$:

- $E(\chi^2_q) = q$.
- $\text{var}(\chi^2_q) = 2q$.

Remarks:

(i) Theorem (i) implies

$$E \left[ \frac{(n - K)s^2}{\sigma^2} \mid X \right] = n - K.$$

$$(n - K) \frac{1}{\sigma^2} E(s^2 \mid X) = n - K.$$ 

It follows that $E(s^2 \mid X) = \sigma^2$.

(ii) Theorem (i) also implies

$$\text{var} \left[ \frac{(n - K)s^2}{\sigma^2} \mid X \right] = 2(n - K),$$

$$\frac{(n - K)^2}{\sigma^4} \text{var}(s^2 \mid X) = 2(n - K),$$

$$\text{var}(s^2 \mid X) = \frac{2\sigma^4}{n - K} \to 0$$

as $n \to \infty$.

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

$$MSE(s^2 \mid X) = E \left[ (s^2 - \sigma^2)^2 \mid X \right]$$

$$= \text{var}(s^2 \mid X) + [E(s^2 \mid X) - \sigma^2]^2$$

$$\to 0.$$
Thus, $s^2$ is a good estimator for $\sigma^2$.

(iv) 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 soon.

Proof of Theorem: (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 $\varepsilon'X \sim N(0, \sigma^2 I_n)$, and $M$ is an idempotent matrix with rank $= n - K$ (as has been shown before), we have the quadratic form

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

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

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

(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.

Question: Why are $e$ and $\hat{\beta}$ jointly normally distributed?

Answer: We write

$$\begin{bmatrix} e \\ \hat{\beta} - \beta^0 \end{bmatrix} = \begin{bmatrix} M\varepsilon \\ (X'X)^{-1}X'\varepsilon \end{bmatrix} = \begin{bmatrix} M \\ (X'X)^{-1}X' \end{bmatrix} \varepsilon.$$

Because $\varepsilon'X \sim N(0, \sigma^2 I)$, the linear combination of

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

23
is also normally distributed conditional on $X$.

**Question:** Why are these sampling distributions useful in practice?

**Answer:** They are useful in confidence interval estimation and hypothesis testing.

### 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

$$H_0 : R\beta^0 = r,$$

$$(J \times K)(K \times 1) = J \times 1,$$

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

**Remark:** We will test $H_0$ under correct model specification for $E(Y_t|X_t)$.

### Motivations

**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

$$H_0^a : \beta_3^0 = 0$$

This is equivalent to the choice of:

$$\beta^0 = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$$

$$R = (0, 0, 0, 1, 0),$$

$$r = 0.$$

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

$$H_0^b : \beta_4 = 0.$$
Alternatively, to test whether the production technology exhibits the constant return to scale (CRS), we can write the null hypothesis as follows:

\[ H_0^c : \beta_1^0 + \beta_2^0 = 1. \]

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:

\[ H_0^d : \beta_3^0 = \beta_4^0 = 0. \]

This is equivalent to the choice of

\[
R = \begin{bmatrix}
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}, \quad r = \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}.
\]

Example 3 [Optimal Predictor for Future Spot Exchange Rate]: Consider

\[ S_{t+\tau} = \beta_0 + \beta_1 F_t(\tau) + \varepsilon_{t+\tau}, \quad t = 1, ..., n, \]

where \( S_{t+\tau} \) is the spot exchange rate at period \( t + \tau \), and \( F_t(\tau) \) is 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

\[ H_0^c : \beta_0^0 = 0, \beta_1^0 = 1, \]

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

\[
R = \begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0 \\
0 & 0 \\
0 & 0
\end{bmatrix}, \quad r = \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
1
\end{bmatrix}.
\]

Remark: 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

\[ H_0 : R\beta^0 = r, \]

25
where \( r \) is a \( J \times 1 \) vector.

**Basic Idea of Hypothesis Testing**

To test the null hypothesis

\[
H_0 : R\beta^0 = r,
\]

we can consider the statistic:

\[
R\hat{\beta} - r
\]

and check if this difference is significantly different from zero.

Under \( H_0 : R\beta^0 = r \), we have

\[
R\hat{\beta} - r = R\hat{\beta} - R\beta^0 = R(\hat{\beta} - \beta^0) \to 0 \text{ as } n \to \infty
\]

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

Under the alternative to \( H_0 \), \( R\beta^0 \neq r \), but we still have \( \hat{\beta} - \beta^0 \to 0 \) in the term of MSE. It follows that

\[
R\hat{\beta} - r = R(\hat{\beta} - \beta^0) + R\beta^0 - r \to R\beta^0 - r \neq 0
\]

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, \( R\beta^0 - r \).

**Remark:** The behavior of \( R\hat{\beta} - r \) is different under \( H_0 \) and under the alternative hypothesis to \( H_0 \). Therefore, we can test \( H_0 \) 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?

**Answer:** We need to a decision rule which specifies a threshold value with which we can compare the (absolute) value of \( R\hat{\beta} - r \).

Note that \( R\hat{\beta} - r \) 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 \( \mathbb{H}_0 \)?

Because

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

we have that conditional on \( X \),

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

**Corollary:** Under Assumptions 3.1, 3.3 and 3.5, and \( H_0 : R\beta^0 = r \), we have for each \( n > K \),

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

**Remarks:**

(i) \( \text{var}(R\hat{\beta}|X) = \sigma^2 R(X'X)^{-1}R' \) is a \( J \times J \) matrix.

(ii) However, \( R\hat{\beta} - r \) cannot be used as a test statistic for \( \mathbb{H}_0 \), because \( \sigma^2 \) 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

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

When \( J = 1 \), the conditional variance

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

is a scalar (\( 1 \times 1 \)). It follows that conditional on \( X \), we have

\[
\frac{R\hat{\beta} - r}{\sqrt{\text{var}[(R\hat{\beta} - r)|X]}} = \frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(X'X)^{-1}R'}} \sim N(0, 1).
\]
**Question:** What is the unconditional distribution of

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

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

However, \(\sigma^2\) is unknown, so we cannot use the ratio

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

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

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

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

\[ T = \frac{R \hat{\beta} - r}{\sqrt{s^2 R(X'X)^{-1} R'}} = \frac{R \hat{\beta} - r}{\sqrt{(n-K)s^2}/(n-K)} \]

\[ \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-K}/(n-K)}} \sim t_{n-K}, \]

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 \(X\), because \(\hat{\beta}\) and \(s^2\) are mutually independent conditional on \(X\).

**Remarks:**

(i) The feasible statistic \(T\) is called a \(t\)-test statistic.

(ii) What is a Student \(t\) 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

\[ \frac{Z}{\sqrt{V/q}} \sim t_q. \]
(iii) When $q \rightarrow \infty$, $t_q \xrightarrow{d} N(0, 1)$, where $\xrightarrow{d}$ denotes convergence in distribution. This implies that we have

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

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 $N(0, 1)$.

Plots of Student $t$-distributions?

**Question:** What is meant by 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,

$$\lim_{n \rightarrow \infty} F_n(z) = F(z) \text{ or } F_n(z) \rightarrow F(z) \text{ as } n \rightarrow \infty$$

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 $Z_n$.

**Implication:** In practice, $Z_n$ is a test statistic or a parameter estimator, and 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 be 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 $\varepsilon$ 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 $H_0$ when $J = 1$.

**Decision Rule of the T-test Based on Critical Values**
(i) Reject $\mathbb{H}_0 : R\beta^0 = r$ 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 $\mathbb{H}_0$ at the significance level $\alpha$ if

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

**Remarks:**

(a) In testing $\mathbb{H}_0$, 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 $\mathbb{H}_0$ is true but we reject it. This is called the “Type I error”. The significance level $\alpha$ is the probability of making the Type I error. If

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

we say that the decision rule is a test with size $\alpha$.

(b) The probability $P[|T| > C_{t_{n-K}, \frac{\alpha}{2}}]$ is called the power function of a size $\alpha$ test. When

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

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

(c) 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 $\alpha$ are 10%, 5% and 1% respectively.

Next, we describe an alternative decision rule for testing $\mathbb{H}_0$ 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 $z^n = \{y_t, x'_t\}_{t=1}^n$, which is a realization of the random sample $Z^n = \{Y_t, X'_t\}_{t=1}^n$, we can compute a realization (i.e., a number) for the $t$-test statistic $T$, namely

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

Then the probability

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

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 $T(z^n)$ is a realization for test statistic $T = T(Z^n)$. Intuitively, the $p$-value 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 $T(z^n)$. If this probability is very small relative to the significance level, then it is unlikely that the test statistic $T(Z^n)$ 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 $H_0$ at the significance level $\alpha$ if $p(z^n) < \alpha$.

(ii) Accept $H_0$ at the significance level $\alpha$ if $p(z^n) \geq \alpha$.

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

**Remarks:**

(i) $p$-values are more informative than only rejecting/accepting the null hypothesis at some significance level. A $p$-value is the smallest significance level at which a null hypothesis can be rejected.

(ii) Most statistical software reports $p$-values of parameter estimates.

**Examples of $t$-tests**

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

We first consider testing the null hypothesis

$$H_0^a : \beta_3 = 0,$$

where $\beta_3$ 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
\[ s^2 R(X'X)^{-1} R' = \left[ s^2(X'X)^{-1} \right]_{(4,4)} \]
\[ = S^2_{\hat{\beta}} \]
which is the estimator of \( \text{var}(\hat{\beta}_3|X) \). The \( t \)-test statistic
\[ T = \frac{R \hat{\beta} - r}{\sqrt{s^2 R(X'X)^{-1} R'}} \]
\[ = \frac{\hat{\beta}_3}{\sqrt{S^2_{\hat{\beta}_3}}} \]
\[ \sim t_{n-K}. \]

Next, we consider testing the CRS hypothesis
\[ H_0^0: \beta_1 + \beta_2 = 1, \]
which corresponds to \( R = (0, 1, 1, 0, 0) \) and \( r = 1 \). In this case,
\[ s^2 R(X'X)^{-1} R' = S^2_{\hat{\beta}_1} + S^2_{\hat{\beta}_2} + 2 \hat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2) \]
\[ = [s^2(X'X)^{-1}]_{(2,2)} + [s^2(X'X)^{-1}]_{(3,3)} + 2 \cdot [s^2(X'X)^{-1}]_{(2,3)} \]
\[ = S^2_{\hat{\beta}_1 + \hat{\beta}_2}, \]
which is the estimator of \( \text{var}(\hat{\beta}_1 + \hat{\beta}_2|X) \). Here, \( \hat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2) \) is the estimator for \( \text{cov}(\hat{\beta}_1, \hat{\beta}_2|X) \), the covariance between \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) conditional on \( X \).

The \( t \)-test statistic is
\[ T = \frac{R \hat{\beta} - r}{\sqrt{s^2 R(X'X)^{-1} R'}} \]
\[ = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{S_{\hat{\beta}_1 + \hat{\beta}_2}} \]
\[ \sim t_{n-K}. \]

**CASE II: F-testing \((J > 1)\)**

**Question:** What happens if \( J > 1 \)? We first state a useful lemma.
Lemma: If \( Z \sim N(0, V) \), where \( V = \text{var}(Z) \) 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

\[
V^{1/2}V^{1/2} = V, \\
V^{-1/2}V^{-1/2} = V^{-1}.
\]

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

\[ Y = V^{-1/2}Z. \]

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

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

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

\[ (R\hat{\beta} - r)'[\sigma^2 R(X'X)^{-1}R'] \]

under \( H_0 \), we have the quadratic form

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

conditional on \( X \), or

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

conditional on \( X \).
Because $\chi_j^2$ does not depend on $X$, therefore, we also have

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

unconditionally.

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

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

The replacement of $\sigma^2$ 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 $(q, n - K)$.

**Question:** Why?

Observe

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{s^2} = J \cdot \frac{\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2}}{\frac{(n-K)s^2}{(n-K)}} / \frac{J}{(n-K)}$$

$$\sim J \cdot \frac{\chi_j^2 / J}{\chi_{n-K}^2 / (n-K)}$$

$$\sim J \cdot F_{j,n-K},$$

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 $U \sim \chi_p^2$ and $V \sim \chi_q^2$, 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 $(p,q)$.

**Question:** Why call it the $F$-distribution?

**Answer:** It is named after Professor Fisher, a well-known statistician in the 20th century.

Plots of $F$-distributions.
Remarks:

Properties of $F_{p,q}$:
(i) If $F \sim F_{p,q}$, then $F^{-1} \sim F_{q,p}$.
(ii) $t_q^2 \sim F_{1,q}$.
$$t_q^2 = \frac{\chi^2_1/1}{\chi^2_q} \sim F_{1,q}$$
(iii) Given any fixed integer $p$, $p \cdot F_{p,q} \to \chi^2_p$ as $q \to \infty$.

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

We can now define the following $F$-test statistic
$$F \equiv \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R]'^{-1}(R\hat{\beta} - r)/J}{s^2} \sim F_{J,n-K}.$$ 

**Theorem:** Suppose Assumptions 3.1, 3.3 and 3.5 hold. Then under $H_0: R\beta^0 = r$, we have

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

for all $n > K$.

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

**Alternative Expression for the $F$-Test Statistic**

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

$$Y = X\beta^0 + \varepsilon.$$
Let \( \text{SSR}_r = \tilde{e}'\tilde{e} \) be the sum of squared residuals from the restricted model

\[ Y = X\beta^0 + \varepsilon \]

subject to

\[ R\beta^0 = r; \]

where \( \tilde{\beta} \) is the restricted OLS estimator. Then under \( H_0 \), the F-test statistic can be written as

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

**Remark:** This is a convenient test statistic! One only needs to compute SSR in order to compute the F-test statistic.

**Proof:** Let \( \tilde{\beta} \) be the OLS under \( H_0 \); that is,

\[ \tilde{\beta} = \arg \min_{\beta \in \mathbb{R}^K} (Y - X\beta)'(Y - X\beta) \]

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

\[ L(\beta, \lambda) = (Y - X\beta)'(Y - X\beta) + 2\lambda'(r - R\beta), \]

where \( \lambda \) is a \( J \times 1 \) vector called the Lagrange multiplier vector.

We have the following FOC:

\[ \frac{\partial L(\tilde{\beta}, \lambda)}{\partial \beta} = -2X'(Y - X\tilde{\beta}) - 2\tilde{\beta}'\lambda = 0, \]

\[ \frac{\partial L(\tilde{\beta}, \lambda)}{\partial \lambda} = r - R\tilde{\beta} = 0. \]

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

\[-(\hat{\beta} - \tilde{\beta}) = (X'X)^{-1}R'\lambda, \]

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

Hence, the Lagrange multiplier

\[ \lambda = -[R(X'X)^{-1}R'\lambda]^{-1}R(\hat{\beta} - \tilde{\beta}), \]

\[ = -[R(X'X)^{-1}R'\lambda]^{-1}(R\tilde{\beta} - r), \]
where we have made use of the constraint that $R\hat{\beta} = r$. It follows that
\[ \hat{\beta} - \tilde{\beta} = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r). \]

Now,
\[
\tilde{e} = Y - X\tilde{\beta} \\
= Y - X\hat{\beta} + X(\hat{\beta} - \tilde{\beta}) \\
= e + X(\hat{\beta} - \tilde{\beta}).
\]

It follows that
\[
\tilde{e}'\tilde{e} = e'e + (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta}) \\
= e'e + (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r).
\]

We have
\[
(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) = \tilde{e}'\tilde{e} - e'e
\]
and
\[
F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2} = \frac{(\tilde{e}'\tilde{e} - e'e)/J}{e'e/(n - K)}.
\]

This completes the proof.

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

**Answer:** Recall that we have obtained the relation that
\[
\tilde{\lambda} = -[R(X'X)^{-1}R']^{-1}R(\hat{\beta} - \tilde{\beta}) \\
= -[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r).
\]

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 that under $H_0$, the quadratic form
\[
(R\hat{\beta} - r)' [\sigma^2 R(X'X)^{-1}R']^{-1} (R\hat{\beta} - r) \sim \chi^2_j.
\]
We have shown that
\[ J \cdot F = (R\hat{\beta} - r)' \left[ s^2 R(X'X)^{-1} R' \right]^{-1} (R\hat{\beta} - r) \]
is no longer \( \chi^2_J \) for each finite \( n > K \).

However, the following theorem shows that as \( n \to \infty \), the quadratic form \( J \cdot F \to^d \chi^2_J \). In other words, when \( n \to \infty \), the limiting distribution of \( J \cdot F \) coincides with the distribution of the quadratic form
\[ (R\hat{\beta} - r)' \left[ \sigma^2 R(X'X)^{-1} R' \right]^{-1} (R\hat{\beta} - r). \]

**Theorem:** Suppose Assumptions 3.1, 3.3 and 3.5 hold. Then under \( H_0 \), we have the Wald test statistic
\[ W = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1} R']^{-1}(R\hat{\beta} - r)}{s^2} \to^d \chi^2_J \]
as \( n \to \infty \).

**Remarks:**

(i) \( W \) is called the Wald test statistic. In the present context, we have \( W = J \cdot F \).

(ii) Recall that when \( n \to \infty \), the \( F \)-distribution \( F_{J,n-K} \), after multiplied by \( J \), will approach the \( \chi^2_J \) distribution. This result has a very important implication in practice: When \( n \) is large, the use of \( F_{J,n-K} \) and the use of \( \chi^2_J \) will deliver the same conclusion on statistical inference.

**Proof:** The result follows immediately by the so-called Slutsky theorem and the fact that \( s^2 \to^p \sigma^2 \), where \( \to^p \) denotes convergence in probability.

**Question:** What is the Slutsky theorem? What is convergence in probability?

**Slutsky Theorem:** Suppose \( Z_n \to^d Z, a_n \to^p a \) and \( b_n \to^p b \) as \( n \to \infty \). Then
\[ a_n + b_nZ_n \to^d a + bZ. \]

In our application, we have
\[ Z_n = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1} R']^{-1}(R\hat{\beta} - r)}{\sigma^2}, \]
\[ Z \sim \chi^2_J, \]
\[ a_n = 0, \]
\[ a = 0, \]
\[ b_n = \sigma^2/s^2, \]
\[ b = 1. \]
The by the Slutsky theorem, we have
\[ W = b_n Z_n \xrightarrow{d} 1 \cdot Z \sim \chi^2_j \]
because \( Z_n \xrightarrow{d} Z \) and \( b_n \xrightarrow{p} 1 \).

What we need is to show that \( s^2 \xrightarrow{p} \sigma^2 \) and then apply the Slutsky theorem.

**Question:** What is the concept of convergence in probability?

**Definition [Convergence in Probability]** Suppose \( \{Z_n\} \) is a sequence of random variables and \( Z \) is a random variable (including a constant). Then we say that \( Z_n \) converges to \( Z \) in probability, denoted as \( Z_n \xrightarrow{p} Z \) or \( Z_n - Z = o_p(1) \) if for every small constant \( \delta > 0 \),
\[
\lim_{n \to \infty} \Pr[|Z_n - Z| > \delta] = 0, \text{ or } \frac{\Pr(|Z_n - Z| > \delta)}{n} \to 0 \text{ as } n \to \infty.
\]

**Interpretation:** Define the event
\[
A_n(\delta) = \{ \omega \in \Omega : |Z_n(\omega) - Z(\omega)| > \delta \},
\]
where \( \omega \) is a basic outcome in sample space \( \Omega \). Then convergence in probability says that the probability of event \( A_n(\delta) \) may be nonzero for any finite \( n \), but such a probability will eventually vanish to zero as \( n \to \infty \). In other words, it becomes less and less likely that the difference \( |Z_n - Z| \) is larger than a prespecified constant \( \delta > 0 \). Or, we have more and more confidence that the difference \( |Z_n - Z| \) will be smaller than \( \delta \) as \( n \to \infty \).

**Question:** What is the interpretation of \( \delta \)?

**Answer:** It can be viewed as a prespecified tolerance level.

**Remark:** When \( Z = b \), a constant, we can write \( Z_n \xrightarrow{p} b \), and we say that \( b \) is the probability limit of \( Z_n \), written as
\[ b = \lim_{n \to \infty} Z_n. \]

Next, we introduce another convergence concept which can be conveniently used to show convergence in probability.

**Definition [Convergence in Quadratic Mean]:** We say that \( Z_n \) converges to \( Z \) in quadratic mean, denoted as \( Z_n - Z \xrightarrow{q.m} 0 \), or \( Z_n - Z = o_{q.m}(1) \), if
\[
\lim_{n \to \infty} \frac{\text{E}[|Z_n - Z|^2]}{n} = 0.
\]
Lemma: If $Z_n - Z \xrightarrow{a.s.} 0$, then $Z_n - Z \xrightarrow{p} 0$.

Proof: By Chebyshev’s inequality, we have

$$P(|Z_n - Z| > \delta) \leq \frac{E[Z_n - Z]^2}{\delta^2} \to 0$$

for any given $\delta > 0$ as $n \to \infty$.

Example: Suppose Assumptions 3.1, 3.3 and 3.5 hold. Does $s^2$ converge in probability to $\sigma^2$?

Solution: Under the given assumptions,

$$(n - K) \frac{s^2}{\sigma^2} \sim \chi^2_{n-K},$$

and therefore we have $E(s^2) = \sigma^2$ and $\text{var}(s^2) = \frac{2\sigma^4}{n-K}$. It follows that $E(s^2 - \sigma^2)^2 = 2\sigma^4/(n - K) \to 0$, $s^2 \xrightarrow{a.s.} \sigma^2$ and so $s^2 \xrightarrow{p} \sigma^2$ because convergence in quadratic mean implies convergence in probability.

Question: If $s^2 \xrightarrow{p} \sigma^2$, do we have $s \xrightarrow{p} \sigma$?

Answer: Yes. Why? It follows from the following lemma with the choice of $g(s^2) = \sqrt{s^2} = s$.

Lemma: Suppose $Z_n \xrightarrow{p} Z$, and $g(\cdot)$ is a continuous function. Then

$$g(Z_n) \xrightarrow{p} g(Z).$$

Proof: Because $g(\cdot)$ is continuous, we have that for any given constant $\epsilon > 0$, there exists $\delta = \delta(\epsilon)$ such that whenever $|Z_n(\omega) - Z(\omega)| < \delta$, we have

$$|g[Z_n(\omega)] - g[Z(\omega)]| < \epsilon.$$

Define

$$A(\delta) = \{ \omega : |Z_n(\omega) - Z(\omega)| > \delta \},$$

$$B(\epsilon) = \{ \omega : |g(Z_n(\omega)) - g(Z(\omega))| > \epsilon \}.$$

Then the continuity of $g(\cdot)$ implies that $B(\epsilon) \subseteq A(\delta)$. It follows that

$$P[B(\epsilon)] \leq P[A(\delta)] \to 0$$

40
as \( n \to \infty \), where \( P[A(\delta)] \to 0 \) given \( Z_n \to^p Z \). Because \( \epsilon \) and \( \delta \) are arbitrary, we have

\[
g(Z_n) \to^p g(Z).
\]

This completes the proof.

### 3.8 Applications

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

**Case 1: Testing for the Joint Significance of Explanatory Variables**

Consider a linear regression model

\[
Y_t = X_t^\prime \beta_0^0 + \varepsilon_t = \beta_0^0 + \sum_{j=1}^{k} \beta_j^0 X_{jt} + \varepsilon_t.
\]

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

\[
H_0 : \beta_j^0 = 0 \text{ for } 1 \leq j \leq k,
\]

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

\[
H_A : \beta_j^0 \neq 0 \text{ at least for some } \beta_j^0, \quad j = 1, \ldots, k.
\]

One can use the \( F \)-test and

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

In fact, the restricted model under \( H_0 \) is very simple:

\[
Y_t = \beta_0^0 + \varepsilon_t.
\]

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

\[
\tilde{\varepsilon} = Y - X \tilde{\beta} = Y - \bar{Y}.
\]

Hence, we have

\[
\tilde{\varepsilon}' \tilde{\varepsilon} = (Y - \bar{Y})' (Y - \bar{Y}).
\]
Recall the definition of $R^2$:

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

$$= 1 - \frac{e'e}{\bar{e}'\bar{e}}.$$

It follows that

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

$$= \frac{(1 - \frac{e'e}{\bar{e}'\bar{e}})/k}{\frac{e'e}{\bar{e}'\bar{e}}/(n - k - 1)}$$

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

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 $H_0 : \beta_j^0 = 0$ 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:

$$Y_t = X_t'\beta^0 + \varepsilon_t,$$

where

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

Under EMH, we have

$$H_0 : \beta_j^0 = 0 \text{ for all } j = 1, \ldots, k.$$

If the alternative

$$H_A : \beta_j^0 \neq 0 \text{ at least for some } j \in \{1, \ldots, k\}$$

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

**Remarks:**
(i) What is the appropriate interpretation if \( H_0 \) is not rejected? Note that there exists a gap between the efficiency hypothesis and \( H_0 \), because the linear regression model is just one of many ways to check EMH. Thus, \( H_0 \) 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.

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

\[
k \cdot F = \frac{R^2}{(1 - R^2)/(n - k - 1)} \xrightarrow{d} \chi^2_k
\]

under conditional homoskedasticity even for a linear dynamic regression model.

(iii) 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 \( R^2 \rightarrow^p 0 \) under \( H_0 \). 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.

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

\[
Y_t = 33.88 - 26.00X_{1t} + 6.71X_{2t} + \varepsilon_t, \quad R^2 = 0.742, n = 25.
\]

\[
[1.77] \quad [-0.74] \quad [0.77]
\]

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,

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

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

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

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

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

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

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

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

Consider the restricted model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_{k_1} X_{k_1t} + \varepsilon_t.$$

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

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_{k_1} X_{k_1t} + \beta_{k_1+1} X_{(k_1+1)t} + \cdots + \beta_{(k_1+k_2)} X_{(k_1+k_2)t} + \varepsilon_t.$$

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

$$H_0 : \beta_{k_1+1} = \beta_{k_1+2} = \cdots = \beta_{k_1+k_2} = 0.$$

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

The $F$-Test statistic is

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

**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 $H_0$, can we say that there is no omitted variable?
**Answer:** 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

\[ Y_t = \beta_0 + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) + \beta_3 AU_t + \beta_4 PS_t + \beta_5 CM_t + \varepsilon_t, \]

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:

\[ H_0: \beta_3 = \beta_4 = \beta_5 = 0. \]

We can use the \( F \)-test, and \( F \sim F_{3, n-6} \) under \( H_0 \).

**Interpretation:** If rejection occurs, there exists evidence against \( H_0 \). If no rejection occurs, then we find no evidence against \( H_0 \) (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, \( I_{t-1}^Y = \{Y_{t-1}, \ldots, Y_1\} \) and \( I_{t-1}^Z = \{Z_{t-1}, \ldots, Z_1\} \). 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 \( Y_t \).

**Remark:** 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

\[ Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \beta_{p+1} Z_{t-1} + \cdots + \beta_{p+q} Z_{t-q} + \varepsilon_t. \]
Under non-Granger causality, we have

\[ H_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0. \]

The \( F \)-test statistic

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

**Remark:** The current theory (Assumption 3.2: \( E(\varepsilon_t | X) = 0 \)) 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)]**

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

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:

\[
Y_t = (\beta_0 + \alpha_0 D_t) + (\beta_1 + \alpha_1 D_t) X_{1t} + \varepsilon_t \\
= \beta_0 + \beta_1 X_{1t} + \alpha_0 D_t + \alpha_1 (D_t X_{1t}) + \varepsilon_t,
\]

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-structural break period.

The null hypothesis of no structural change is

\[ H_0 : \alpha_0 = \alpha_1 = 0. \]

The alternative hypothesis that there exist a structural change is

\[ H_A : \alpha_0 \neq 0 \text{ or } \alpha_1 \neq 0. \]

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:
\[ H_0 : \beta_1 + \beta_2 = 1. \]
The alternative hypothesis is
\[ H_0 : \beta_1 + \beta_2 \neq 1. \]
What is the restricted model under \( H_0 \)? 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 equivalently
\[ \ln(Y_t/K_t) = \beta_0 + \beta_1 \ln(L_t/K_t) + \beta_3 AU_t + \beta_4 CON_t + \beta_5 CM_t + \varepsilon_t. \]
The \( F \)-test statistic
\[ F \sim F_{1, n-6}. \]
Remark: Both \( t \)- and \( F \)-tests are applicable to test CRS.

Example 2 [Wage Determination]: Consider the wage function
\[ W_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 U_t + \beta_4 V_t + \beta_5 W_{t-1} + \varepsilon_t, \]
where
- \( W_t \) = wage,
- \( P_t \) = price,
- \( U_t \) = unemployment,
- \( V_t \) = unfilled vacancies.

We will test the null hypothesis
\[ H_0 : \beta_1 + \beta_2 = 0, \beta_3 + \beta_4 = 0, \text{ and } \beta_5 = 1. \]
Question: What is the economic interpretation of the null hypothesis \( H_0 \)?
What is the restricted model? Under \( H_0 \), 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 $H_0$ is 

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

**Case 4: Testing for Near-Multicolinearity**

**Example 1:**

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:

$$Y_t = 36.74 + 0.832X_{1t} + e_{1t}, \quad R^2 = 0.735$$

[1.98][7.98]

The second is a regression of consumption on wealth:

$$Y_t = 36.61 + 0.208X_{2t} + e_{2t}, \quad R^2 = 0.735$$

[1.97][7.99]

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

$$Y = 33.88 - 26.00X_{1t} + 6.71X_{2t} + e_t, \quad R^2 = 0.742,$$

[1.77][−0.74][0.77]

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 stock wealth has impact on consumption, we can use the $F$-test:

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

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 ???.
3.9 Generalized Least Squares Estimation

**Question:**

The classical linear regression theory crucially depends on the assumption that $\varepsilon|X \sim N(0, \sigma^2I)$, 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:** $\varepsilon|X \sim N(0, \sigma^2V)$, where $0 < \sigma^2 < \infty$ is unknown and $V = V(X)$ is a known $n \times n$ finite and positive definite matrix.

**Question:** Is this assumption realistic in practice?

**Remarks:**

(i) Assumption 3.6 implies that

$$\text{var}(\varepsilon|X) = E(\varepsilon\varepsilon'|X) = \sigma^2V = \sigma^2V(X)$$

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

(iii) It is possible that $V$ is not a diagonal matrix. Thus, $\text{cov}(\varepsilon_t, \varepsilon_s|X)$ may not be zero. In other words, Assumption 3.6 allows conditional serial correlation of known form.

(iv) 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 and 3.6 hold. Then

(i) unbiasedness: $E(\hat{\beta}|X) = \beta^0$.

(ii) variance:

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

$\neq \sigma^2(X'X)^{-1}$.
\begin{align*}
(iii) & \quad (\hat{\beta} - \beta^0) | X \sim N(0, \sigma^2(X'X)^{-1}X'VX(X'X)^{-1}). \\
(iv) & \quad \text{cov}(\hat{\beta}, \epsilon|X) \neq 0 \text{ in general.}
\end{align*}

Remarks:

(i) 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 $\beta^0$ in the sense of MSE.

(ii) However, the variance of the OLS estimator $\hat{\beta}$ does no longer have the simple expression of $\sigma^2(X'X)^{-1}$ 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 $\sigma^2(X'X)^{-1}$ rather than the correct variance formula of $\sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$.

(iii) Theorem (iv) implies that even if we can obtain a consistent estimator for $\sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$ 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.

Proof: (i) Using $\hat{\beta} - \beta^0 = (X'X)^{-1}X'\epsilon$, we have

\begin{align*}
E[(\hat{\beta} - \beta^0)|X] &= (X'X)^{-1}X'E(\epsilon|X) \\
&= (X'X)^{-1}X'0 \\
&= 0.
\end{align*}

(ii)

\begin{align*}
\text{var}(\hat{\beta}|X) &= E[(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)'|X] \\
&= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X] \\
&= (X'X)^{-1}X'E(\epsilon\epsilon'|X)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'VX(X'X)^{-1}.
\end{align*}

Note that we cannot further simplify the expression here because $V \neq \sigma^2I$.

(iii) Because

\begin{align*}
\hat{\beta} - \beta^0 &= (X'X)^{-1}X'\epsilon \\
&= \sum_{t=1}^{n} C_t \varepsilon_t,
\end{align*}
where
\[ C_t = (X'X)^{-1}X_t, \]

\( \hat{\beta} - \beta^0 \) follows a normal distribution given \( X \), because it is a sum of a normal random variables. As a result,

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

(iv)

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

because \( X'VM \neq 0 \). We can see that it is conditional heteroskedasticity and/or auto-correlation in \( \{\varepsilon_t\} \) that cause \( \hat{\beta} \) to be correlated with \( e \).

**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

\[
V^{-1} = C'C, \\
V = C^{-1}(C')^{-1}
\]

where \( C \) is a \( n \times n \) nonsingular matrix.

**Question:** What is this decomposition called?

**Remark:** \( C \) may not be symmetric.

Consider the original linear regression model:

\[ Y = X\beta^0 + \varepsilon. \]

If we multiply in the equation by \( C \), we obtain the transformed regression model

\[
CY = (CX)\beta^0 + C\varepsilon, \text{ or} \\
Y^* = X^*\beta^0 + \varepsilon^*.
\]
where \( Y^* = CY, X^* = CX \) and \( \varepsilon^* = C\varepsilon \). Then the OLS of this transformed model

\[
\hat{\beta}^* = (X^*X^*)^{-1}X^*Y^* = (X'C'CX)^{-1}(X'C'CY) = (X'V^{-1}X)^{-1}X'V^{-1}Y
\]

is called the Generalized Least Square (GLS) estimator.

**Question:** What is the nature of GLS?

Observe that

\[
E(\varepsilon^*|X) = E(C\varepsilon|X) = CE(\varepsilon|X) = C \cdot 0 = 0.
\]

Also, note that

\[
\text{var}(\varepsilon^*|X) = E[\varepsilon^*\varepsilon^*'|X] = E[C\varepsilon\varepsilon'|X] = CE(\varepsilon^*|X)C' = \sigma^2CVC' = \sigma^2C[C^{-1}(C')^{-1}]C' = \sigma^2I.
\]

It follows from Assumption 3.6 that

\[
\varepsilon^*|X \sim N(0, \sigma^2I).
\]

The transformation makes the new error \( \varepsilon^* \) conditionally homoskedastic and serially uncorrelated, while maintaining the normality distribution. Suppose that for \( t, \varepsilon_t \) has a large variance \( \sigma_t^2 \). The transformation \( \varepsilon_t^* = C\varepsilon_t \) will discount \( \varepsilon_t \) by dividing it by its conditional standard deviation so that \( \varepsilon_t^* \) becomes conditionally homoskedastic. In addition, the transformation also remove possible correlation between \( \varepsilon_t \) and \( \varepsilon_s, t \neq s \). As a consequence, GLS becomes the best linear LS estimator for \( \beta^0 \) 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

$$V = \begin{bmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & h_n \end{bmatrix}.$$  

Then

$$C = \begin{bmatrix} h_1^{-1/2} & 0 & \cdots & 0 \\ 0 & h_2^{-1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & h_n^{-1/2} \end{bmatrix}$$

and

$$\varepsilon^* = C\varepsilon = \begin{bmatrix} \frac{\varepsilon_1}{\sqrt{h_1}} \\ \frac{\varepsilon_2}{\sqrt{h_2}} \\ \vdots \\ \frac{\varepsilon_n}{\sqrt{h_n}} \end{bmatrix}.$$  

**Example 2 [Eliminating Serial Correlation]** Suppose

$$V = \begin{bmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ 0 & \rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \rho \end{bmatrix}.$$  

Then

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

**Theorem:** Under Assumptions 3.1, 3.3 and 3.6,

(i) $E(\hat{\beta}^*|X) = \beta^0$;

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

(iii) $\text{cov}(\hat{\beta}^*, e^*|X) = 0$, where $e^* = Y^* - X^*\hat{\beta}^*$;

(iv) $\hat{\beta}^*$ is BLUE.
\[ E(s^2|X) = \sigma^2, \text{ where } s^2 = e^*e^*/(n - K). \]

**Proof:**

(i) GLS is the OLS of the transformed model.

(ii) The transformed model satisfies 3.1, 3.3 and 3.5 of the classical regression assumptions with \( \varepsilon^*|X^* \sim N(0, \sigma^2I_n) \). It follows that GLS is BLUE by the Gauss-Markov Theorem.

**Remarks:**

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

\[
T^* = \frac{\hat{R}\hat{\beta}^* - r}{\sqrt{s^2 R(X^*X^*)^{-1}R^t}} \sim t_{n-K},
\]

\[
F^* = \frac{(\hat{R}\hat{\beta}^* - r)(R(X^*X^*)^{-1}R^t)^{-1}(\hat{R}\hat{\beta}^* - r)/J}{s^2} \sim F_{J,n-K}.
\]

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

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

\[
\hat{\beta}^* = (X^*X^*)^{-1}X^*Y^* = (X'V^{-1}X)^{-1}X'V^{-1}Y.
\]

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 \( \text{var}(\varepsilon|X) = \sigma^2V \).

What are feasible solutions?

**Two solutions**

(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

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

54
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

$$
\sigma^2 V = \text{diag}\{\sigma_1^2(X), \ldots, \sigma_n^2(X)\} = \text{diag}\{\sigma^2(X_1), \ldots, \sigma^2(X_n)\},
$$

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

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

where $e_t$ is the estimated OLS residual, and $K(\cdot)$ is a kernel function which is a specified symmetric density function (e.g., $K(u) = (2\pi)^{-1/2} \exp(-\frac{1}{2}u^2)$), 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 $\hat{\beta}_a^*$, 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 $\hat{\beta}_a^*$ (i.e., the MSE of $\hat{\beta}_a^*$ is approximately equal to the MSE of $\hat{\beta}_a^*$). 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).

Simulation studies?

(ii) Second Approach

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

$$\text{var}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}X'VXX'(X'X)^{-1}\right) \text{. The classical definitions of } t \text{ and } F\text{-tests cannot be used, because they are based on an incorrect formula for } \text{var}(\hat{\beta}|X). \text{ However, some modified tests can be obtained by using a consistent estimator for the correct formula for } \text{var}(\hat{\beta}|X). \text{ The trick is to estimate } \sigma^2X'VX, \text{ which is a } K \times K \text{ unknown}$$

55
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

$$E(\varepsilon \varepsilon | X) = \sigma^2 V = \text{diag}\{\sigma_1^2(X), ..., \sigma_n^2(X)\}.$$ 

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

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

$$R\hat{\beta} - r \sim t_{n-K}?$$ 

For $J > 1$, do we have

$$(R\hat{\beta} - r)' [R(X'X)^{-1}X'D(e)D(e)'X(X'X)^{-1}R']^{-1} (R\hat{\beta} - r)/J \sim F_{J,n-K}?$$ 

**Answer:** No. Why? Recall $\text{cov}(\hat{\beta}, e|X) \neq 0$ under Assumption 3.6. It follows that $\hat{\beta}$ and $e$ are not independent.

However, when $n \to \infty$, we have

(i) Case I ($J = 1$):

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

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

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

This is a robust Wald test statistic.

The above two feasible solutions are based on the assumption that \(E(\varepsilon_t\varepsilon_s|X) = 0\) for all \(t \neq s\). When there exists serial correlation of unknown form, an alternative solution should be provided. This is discussed in Chapter 6.

### 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 \(X\), 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 is provided in Chapter 6.

References

1. Consider a bivariate linear regression model

\[ Y_t = X'_t \beta^0 + \varepsilon_t, \quad t = 1, \ldots, n, \]

where \( X_t = (X_{0t}, X_{1t}) = (1, X_{1t})' \), and \( \varepsilon_t \) is a regression error. Let \( \hat{\rho} \) denote the sample correlation between \( Y_t \) and \( X_{1t} \); namely,

\[ \hat{\rho} = \frac{\sum_{i=1}^{n} x_{1i} y_i}{\sqrt{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} y_i^2}}, \]

where \( y_t = Y_t - \bar{Y}, \) \( x_{1t} = X_{1t} - \bar{X}_1 \), and \( \bar{Y} \) and \( \bar{X}_1 \) are the sample means of \( \{Y_t\} \) and \( \{X_{1t}\} \). Show \( R^2 = \hat{\rho}^2 \).

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 \( \lambda_{\min}(X'X) \to \infty \) as \( n \to \infty \)? Explain.

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

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

4. Consider the following linear regression model

\[ Y_t = X'_t \beta^0 + u_t, \quad t = 1, \ldots, n, \]  \hfill (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

\[ \Omega = \begin{bmatrix} \sigma^2(X_1) & 0 & 0 & \ldots & 0 \\ 0 & \sigma^2(X_2) & 0 & \ldots & 0 \\ 0 & 0 & \sigma^2(X_3) & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sigma^2(X_n) \end{bmatrix} = \Omega^{1/2} \Omega^{1/2}, \]

with
\[
\Omega^\frac{1}{2} = 
\begin{bmatrix}
\sigma(X_1) & 0 & 0 & \ldots & 0 \\
0 & \sigma(X_2) & 0 & \ldots & 0 \\
0 & 0 & \sigma(X_3) & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \sigma(X_n)
\end{bmatrix}
\]

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 \(\beta^0\).

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

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

Consider an alternative estimator

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

(c) Is \(\tilde{\beta}\) unbiased for \(\beta^0\)?

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

(e) Is \(\text{var}(\tilde{\beta}) - \text{var}(\hat{\beta})\) 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 \(\beta^0\)? [Hint: There are several approaches to this question. A simple one is to consider the transformed model

\[
Y^*_t = X^*_t\beta^0 + \varepsilon_t, \quad t = 1, \ldots, n,
\]

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

\[
Y^* = X^*\beta^0 + \varepsilon,
\]

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

(g) Construct two test statistics for the null hypothesis of interest \(H_0: \beta_2^0 = 0\). 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 \(H_0\)? Can you tell which test is better?

(h) Construct two test statistics for the null hypothesis of interest \(H_0: R\beta^0 = r\), 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 \(H_0\)?
5. Consider the following classical regression model

\[ Y_t = X'_t \beta^0 + \varepsilon_t \]

\[ = \beta^0_0 + \sum_{j=1}^{k} \beta^0_j X_{jt} + \varepsilon_t, \quad t = 1, \ldots, n. \] (5.1)

Suppose that we are interested in testing the null hypothesis

\[ H_0 : \beta^0_1 = \beta^0_2 = \cdots = \beta^0_k = 0. \]

Then the \( F \)-statistic can be written as

\[ F = \frac{(\bar{e}' \bar{e} - e'e)/(n-k-1)}{e'e/(n-k-1)}. \]

where \( e'e \) is the sum of squared residuals from the unrestricted model (5.1), and \( \bar{e}' \bar{e} \) is the sum of squared residuals from the restricted model (5.2)

\[ Y_t = \beta^0_0 + \varepsilon_t. \] (5.2)

(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 (5.1).

(b) Suppose in addition Assumption 3.5 holds. Show that under \( H_0 \),

\[ (n-k-1)R^2 \overset{d}{\to} \chi^2_k. \]

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

\[ \tau'\text{var}(\hat{\beta}|X)\tau = \sigma^2 \tau'(X'X)^{-1}X'VX(X'X)^{-1}\tau \to 0 \]

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

7. Suppose a data generating process is given by

\[ Y_t = \beta^0_{1t}X_{1t} + \beta^0_{2t}X_{2t} + \varepsilon_t = X'_t \beta^0 + \varepsilon_t, \]

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

$$Y_t = \beta_1^0 X_{1t} + \varepsilon_t.$$

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 $\beta_1^0$? Give your reasoning in detail.