Chapter 3. Least Squares Estimation - Finite-Sample Properties[*]

This chapter studies finite-sample properties of the LSE. Related materials can be found in Chapter 1 of Hayashi (2000) and Chapter 3 of Hansen (2007).

1 Terminology and Assumptions

Recall that the linear regression model is

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + u, \\ E[u|\mathbf{x}] &= 0. \end{aligned}$$

We first summarize the terminology for y and $\mathbf{x}$ in Table 1. u is called the **error term**, **disturbance** or **unobservable**.

y	$\mathbf{x}$
Dependent variable	Independent Variable
Explained variable	Explanatory variable
Response variable	Control (Stimulus) variable
Predicted variable	Predictor variable
Regressand	Regressor
LHS variable	RHS variable
Endogenous variable	Exogenous variable
·	Covariate
·	Conditioning variable

Table 1: Terminology for Linear Regression

We maintain the following assumptions in this chapter.

Assumption OLS.0 (random sampling): $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, n$, are independent and identically distributed (i.i.d.).

Assumption OLS.1 (full rank): $\text{rank}(\mathbf{X}) = k$.

Assumption OLS.2 (first moment): $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$.

[*] Email: pingyu@hku.hk

Assumption OLS.3 (second moment): $E[u^2] < \infty$.

Assumption OLS.3$'$ (homoskedasticity): $E[u^2|\mathbf{x}] = \sigma^2$.

Assumption OLS.2 is equivalent to $y = \mathbf{x}'\boldsymbol{\beta} + u$ (linear in parameters) plus $E[u|\mathbf{x}] = 0$ (zero conditional mean). To study the finite-sample properties of the LSE, such as the unbiasedness, we always assume Assumption OLS.2, i.e., the model is linear regression. Assumption OLS.3$'$ is stronger than Assumption OLS.3. The linear regression model under Assumption OLS.3$'$ is called the **homoskedastic linear regression model**,

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + u, \\ E[u|\mathbf{x}] &= 0, \\ E[u^2|\mathbf{x}] &= \sigma^2. \end{aligned}$$

On the other hand, if $E[u^2|\mathbf{x}] = \sigma^2(\mathbf{x})$ depends on $\mathbf{x}$ we say u is **heteroskedastic**.

Exercise 1 *Suppose $y = \mathbf{x}'\boldsymbol{\beta} + u$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$ and u need not satisfy $E[\mathbf{x}u] = \mathbf{0}$.*[1] *Show that if $E\left[\|\mathbf{x}\|^2\right] < \infty$ and $E[u^2] < \infty$, then $E[y^2] < \infty$.*

Exercise 2 *In Exercise 11 of Chapter 2, is the linear regression model homoskedastic? (Hint: If $P(y = j) = \frac{\exp(-\lambda)\lambda^j}{j!}$, then $Var(y) = \lambda$.)*

2 Goodness of Fit

Express

$$y_i = \widehat{y}_i + \widehat{u}_i, \tag{1}$$

where $\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ is the **predicted value**, and $\widehat{u}_i = y_i - \widehat{y}_i$ is the **residual**.[2] Often, the error variance $\sigma^2 = E[u^2]$ is also a parameter of interest. It measures the variation in the "unexplained" part of the regression. Its method of moments (MoM) estimator is the sample average of the squared residuals,

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{u}_i^2 = \frac{1}{n}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}.$$

An alternative estimator uses the formula

$$s^2 = \frac{1}{n-k}\sum_{i=1}^n \widehat{u}_i^2 = \frac{1}{n-k}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}.$$

This estimator adjusts the **degree of freedom** (df) of $\widehat{\mathbf{u}}$. The square root of s^2 is called the **standard error of the regression** (SER).

[1] This means that $\boldsymbol{\beta}$ need not be the true coefficient $\boldsymbol{\beta}_0$ in linear projection.

[2] $\widehat{u}_i$ is different from u_i. The later is unobservable while the former is a by-product of OLS estimation.

Exercise 3 *Consider the OLS regression of the $n \times 1$ vector $\mathbf{y}$ on the $n \times k$ matrix $\mathbf{X}$. Consider an alternative set of regressors $\mathbf{Z} = \mathbf{XC}$, where $\mathbf{C}$ is a $k \times k$ non-singular matrix. Thus, each column of $\mathbf{Z}$ is a mixture of some of the columns of $\mathbf{X}$. Compare the OLS estimates and residuals from the regression of y on $\mathbf{X}$ to the OLS estimates from the regression of $\mathbf{y}$ on $\mathbf{Z}$.*

If $\mathbf{X}$ includes a column of ones, $\mathbf{1}'\widehat{\mathbf{u}} = \sum_{i=1}^{n} \widehat{u}_i = 0$, so $\overline{y} = \overline{\widehat{y}}$. Subtracting $\overline{y}$ from both sides of (1), we have

$$\widetilde{y}_i \equiv y_i - \overline{y} = \widehat{y}_i - \overline{y} + \widehat{u}_i \equiv \widetilde{\widehat{y}}_i + \widehat{u}_i$$

Since $\widetilde{\widehat{\mathbf{y}}}'\widehat{\mathbf{u}} = \widehat{\mathbf{y}}'\widehat{\mathbf{u}} - \overline{y} \cdot \mathbf{1}'\widehat{\mathbf{u}} = \widehat{\boldsymbol{\beta}}\left(\mathbf{X}'\widehat{\mathbf{u}}\right) - \overline{y} \cdot \mathbf{1}'\widehat{\mathbf{u}} = 0$,

$$SST \equiv \left\|\widetilde{\mathbf{y}}\right\|^2 = \widetilde{\mathbf{y}}'\widetilde{\mathbf{y}} = \left\|\widetilde{\widehat{\mathbf{y}}}\right\|^2 + 2\widetilde{\widehat{\mathbf{y}}}'\widehat{\mathbf{u}} + \left\|\widehat{\mathbf{u}}\right\|^2 = \left\|\widetilde{\widehat{\mathbf{y}}}\right\|^2 + \left\|\widehat{\mathbf{u}}\right\|^2 \equiv SSE + SSR, \tag{2}$$

where SST, SSE and SSR mean the *total sum of squares*, the *explained sum of squares*, and the *residual sum of squares* (or the sum of squared residuals), respectively. Dividing SST on both sides of (2),[3] we have

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}.$$

The R**-squared** of the regression, sometimes called the **coefficient of determination**, is defined as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}_y^2}.$$

R^2 is defined only if $\mathbf{x}$ includes a constant. It is usually interpreted as the fraction of the sample variation in y that is explained by (nonconstant) $\mathbf{x}$. R^2 can also be treated as an estimator of

$$\rho^2 = 1 - \sigma^2/\sigma_y^2.$$

It is often useful in algebraic manipulation of some statistics. An alternative estimator of ρ^2 proposed by Theil (1961)[4] called **adjusted** R**-squared** or "R-bar-squared" is

$$\overline{R}^2 = 1 - \frac{s^2}{\widetilde{\sigma}_y^2} = 1 - (1 - R^2)\frac{n-1}{n-k},$$

where $\widetilde{\sigma}_y^2 = \widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}/(n-1)$. $\overline{R}^2$ adjusts the degrees of freedom in the numerator and denominator of R^2 and is smaller than R^2 when $k > 1$.

It is mysterious why we call $n-k$ and $n-1$ as the degree of freedom of $\widehat{\mathbf{u}}$ and $\widetilde{\mathbf{y}}$. We briefly explain the reason here. Roughly speaking, the degree of freedom is the dimension of the space where a vector can stay, or how "freely" a vector can move. For example, $\widehat{\mathbf{u}}$, as a n-dimensional vector, can only stay in a subspace with dimension $n-k$. Why? This is because $\mathbf{X}'\widehat{\mathbf{u}} = 0$, so k constraints are imposed on $\widehat{\mathbf{u}}$, and $\widehat{\mathbf{u}}$ cannot move completely freely and loses k degree of freedom.

[3] When can we conduct this operation, i.e., $SST \neq 0$?

[4] Henri Theil (1924-2000) was a Dutch econometrician. Besides $\overline{R}^2$, he is most famous for the two-stage least squares estimation which we will cover in Chapter 7.

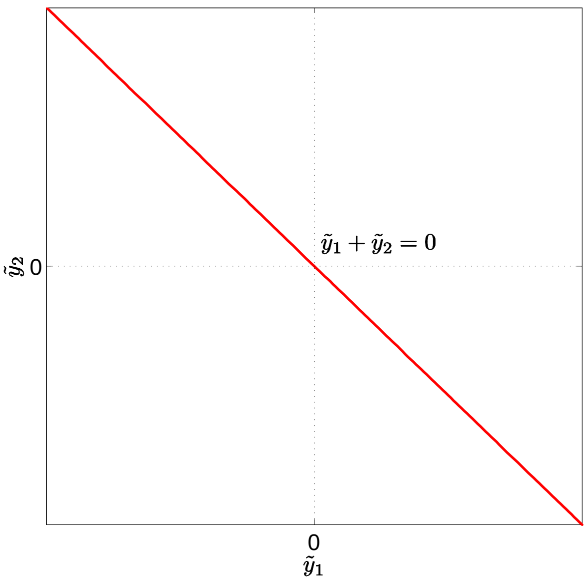

Figure 1: Although $\dim(\widetilde{\mathbf{y}}) = 2$, $\text{df}(\widetilde{\mathbf{y}}) = 1$, where $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \widetilde{y}_2)$

Similarly, the degree of freedom of $\widetilde{\mathbf{y}}$ is $n-1$. Figure 1 illustrates why the degree of freedom of $\widetilde{\mathbf{y}}$ is $n-1$ when $n=2$. Table 2 summarizes the degrees of freedom for the three terms in (2).

Variation	Notation	df
SSE	$\widetilde{\widehat{\mathbf{y}}}'\widetilde{\widehat{\mathbf{y}}}$	$k-1$
SSR	$\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$	$n-k$
SST	$\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}$	$n-1$

Table 2: Degrees of Freedom for Three Variations

Exercise 4 *(i) Explain why the df of SSE is $k-1$. (Hint: $\widetilde{\widehat{y}}_i = \widetilde{\underline{\mathbf{x}}}_i'\widehat{\underline{\boldsymbol{\beta}}}$, where $\underline{\mathbf{x}}$ is the nonconstant covariate, $\underline{\boldsymbol{\beta}}$ is the associated coefficient.) (ii) Show that $SSE = \widehat{\underline{\boldsymbol{\beta}}}'\widetilde{\underline{\mathbf{X}}}'\widetilde{\underline{\mathbf{X}}}\widehat{\underline{\boldsymbol{\beta}}} = \widehat{\underline{\boldsymbol{\beta}}}'\widetilde{\underline{\mathbf{X}}}'\mathbf{y}$.*

Exercise 5 *(i) Show that R^2 remains the same if we demean y_i as $y_i - \overline{y}$. (ii) Prove that R^2 is the square of the simple correlation between $\mathbf{y}$ and $\widehat{\mathbf{y}}$. (iii) Show that $R^2 = \widehat{Corr}(y, \underline{\mathbf{x}})'\widehat{Corr}(\underline{\mathbf{x}}, \underline{\mathbf{x}})^{-1}\widehat{Corr}(y, \underline{\mathbf{x}})$, where $\widehat{Corr}(y, \underline{\mathbf{x}})$ is the sample correlation vector of y with each element of $\underline{\mathbf{x}}$, and $\widehat{Corr}(\underline{\mathbf{x}}, \underline{\mathbf{x}})$ is the sample correlation matrix between every two elements of $\underline{\mathbf{x}}$. When $\widehat{Corr}(\underline{\mathbf{x}}, \underline{\mathbf{x}}) = \mathbf{I}_{k-1}$, $R^2 = \sum_{l=1}^{k-1} \widehat{Corr}(y, \underline{\mathbf{x}}_l)^2$, so R, the square root of R^2, is also called the **multiple correlation coefficient** in statistics.*

To define R^2, we must assume there is a constant term in $\mathbf{x}_i$. Sometimes, when there is no constant term in $\mathbf{x}_i$, we need to define so-called **uncentered** R^2, denoted as R_u^2. Similar to R^2,

$$R_u^2 = \frac{\widehat{\mathbf{y}}'\widehat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}}.$$

Exercise 6 *Show that* $1-R^2=\left(1+\frac{n\overline{y}^2}{\sum_{i=1}^n(y_i-\overline{y})^2}\right)(1-R_u^2)$.

3 Bias and Variance

As mentioned, Assumption OLS.2 implies that

$$y=\mathbf{x}'\boldsymbol{\beta}+u,E[u|\mathbf{x}]=0.$$

Then

$$E[\mathbf{u}|\mathbf{X}]=\begin{pmatrix}\vdots\\E[u_i|\mathbf{X}]\\\vdots\end{pmatrix}=\begin{pmatrix}\vdots\\E[u_i|\mathbf{x}_i]\\\vdots\end{pmatrix}=\mathbf{0},$$

where the second equality is from the assumption of independent sampling (Assumption OLS.0). Now,

$$\widehat{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)=\boldsymbol{\beta}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u},$$

so

$$E\left[\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}|\mathbf{X}\right]=E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}\right]=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}]=\mathbf{0},$$

i.e., $\widehat{\boldsymbol{\beta}}$ is unbiased.

Exercise 7 *In the linear regression model under the restrictions* $\mathbf{R}'\boldsymbol{\beta}=\mathbf{c}$, *let* $\widetilde{\mathbf{u}}$ *be the vector of RLS residuals. Show that under* $\mathbf{R}'\boldsymbol{\beta}=\mathbf{c}$,

(a) $\mathbf{R}'\widehat{\boldsymbol{\beta}}-\mathbf{c}=\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$

(b) $\widehat{\boldsymbol{\beta}}_R-\boldsymbol{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$

(c) $\widetilde{\mathbf{u}}=(\mathbf{I}-\mathbf{P}+\mathbf{A})\mathbf{u}$ *for* $\mathbf{P}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ *and some matrix* $\mathbf{A}$ *(find this matrix* $\mathbf{A}$*).*

(d) *Show that* $\mathbf{A}$ *is symmetric and idempotent,* $\mathrm{tr}(\mathbf{A})=q$, *and* $\mathbf{PA}=\mathbf{A}$.

Furthermore

$$\begin{aligned}Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)&=Var\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}\right)\\&=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Var\left(\mathbf{u}|\mathbf{X}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\&\equiv\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}.\end{aligned}$$

Note that

$$Var\left(u_i|\mathbf{X}\right)=Var\left(u_i|\mathbf{x}_i\right)=E\left[u_i^2|\mathbf{x}_i\right]-E\left[u_i|\mathbf{x}_i\right]^2=E\left[u_i^2|\mathbf{x}_i\right]\equiv\sigma_i^2,$$

and

$$\begin{aligned} Cov(u_i, u_j|\mathbf{X}) &= E\left[u_i u_j|\mathbf{X}\right] - E\left[u_i|\mathbf{X}\right]E\left[u_j|\mathbf{X}\right] \\ &= E\left[u_i u_j|\mathbf{x}_i, \mathbf{x}_j\right] - E\left[u_i|\mathbf{x}_i\right]E\left[u_j|\mathbf{x}_j\right] \\ &= E\left[u_i|\mathbf{x}_i\right]E\left[u_j|\mathbf{x}_j\right] - E\left[u_i|\mathbf{x}_i\right]E\left[u_j|\mathbf{x}_j\right] = 0, \end{aligned}$$

so $\mathbf{D}$ is a diagonal matrix:

$$\mathbf{D} = \text{diag}\left(\sigma_1^2, \cdots, \sigma_n^2\right).$$

It is useful to note that

$$\mathbf{X}'\mathbf{D}\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\sigma_i^2.$$

In the homoskedastic case, $\sigma_i^2 = \sigma^2$ and $\mathbf{D} = \sigma^2\mathbf{I}_n$, so $\mathbf{X}'\mathbf{D}\mathbf{X} = \sigma^2\mathbf{X}'\mathbf{X}$, and

$$Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Exercise 8 *Show that $Var\left(\widehat{\beta}_j|\mathbf{X}\right) = \sum_{i=1}^n w_{ij}\sigma_i^2/SSR_j$, $j = 1, \cdots, k$, where $w_{ij} > 0$, $\sum_{i=1}^n w_{ij} = 1$, and SSR_j is the SSR in the regression of x_j on all other regressors. (Hint: Use the FWL theorem.)*

From this exercise, we can see that $Var\left(\widehat{\beta}_j|\mathbf{X}\right) = \sigma^2/SSR_j = \sigma^2/\left[SST_j(1-R_j^2)\right]$, $j = 1, \cdots, k$, under homoskedasticity (why?), where SST_j is the SST of x_j, and R_j^2 is the R-squared from the simple regression of x_j on the remaining regressors (which includes an intercept).

We now calculate the finite-sample bias of the MoM estimator $\widehat{\sigma}^2$ for σ^2. Recall that $\widehat{\mathbf{u}} = \mathbf{Mu}$, where we abbreviate $\mathbf{M}_\mathbf{X}$ as $\mathbf{M}$, so by the properties of projection matrices and the trace operator, we have

$$\widehat{\sigma}^2 = \frac{1}{n}\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = \frac{1}{n}\mathbf{u}'\mathbf{MMu} = \frac{1}{n}\mathbf{u}'\mathbf{Mu} = \frac{1}{n}\text{tr}\left(\mathbf{u}'\mathbf{Mu}\right) = \frac{1}{n}\text{tr}\left(\mathbf{Muu}'\right).$$

Then

$$E\left[\widehat{\sigma}^2\middle|\mathbf{X}\right] = \frac{1}{n}\text{tr}\left(E\left[\mathbf{Muu}'|\mathbf{X}\right]\right) = \frac{1}{n}\text{tr}\left(\mathbf{M}E\left[\mathbf{uu}'|\mathbf{X}\right]\right) = \frac{1}{n}\text{tr}\left(\mathbf{MD}\right).$$

In the homoskedastic case, $\mathbf{D} = \sigma^2\mathbf{I}_n$, so

$$E\left[\widehat{\sigma}^2\middle|\mathbf{X}\right] = \frac{1}{n}\text{tr}\left(\mathbf{M}\sigma^2\right) = \sigma^2\left(\frac{n-k}{n}\right).$$

Thus $\widehat{\sigma}^2$ underestimates σ^2. As an alternative, the estimator $s^2 = \frac{1}{n-k}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$ is unbiased for σ^2. This is the justification for the common preference of s^2 over $\widehat{\sigma}^2$ in empirical practice. However, it is important to remember that this estimator is only unbiased in the special case of the homoskedastic linear regression model. It is not unbiased in the absence of homoskedasticity or in the projection model.

Exercise 9 *(i) In the homoskedastic linear regression model under the restrictions $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$, find* *an unbiased estimator of* σ^2*. (Hint:* $\frac{1}{n-k+q}\sum_{i=1}^n \widetilde{u}_i^2$*, where* $\widetilde{u}_i$ *is defined in Exercise 7.). (ii) In the heteroskedastic linear regression model with* $Var(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{\Omega}$*, find an unbiased estimator of* σ^2*. (Hint:* $\frac{1}{n-k}\widetilde{\mathbf{u}}'\mathbf{\Omega}^{-1}\widetilde{\mathbf{u}}$*, where* $\widetilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}'\widetilde{\boldsymbol{\beta}}$*, and* $\widetilde{\boldsymbol{\beta}}$ *is the GLS estimator with the weight matrix* $\mathbf{W} = \mathbf{\Omega}^{-1}$*.)*

4 The Gauss-Markov Theorem

The LSE has some optimality properties among a restricted class of estimators in a restricted class of models. The model is restricted to be the homoskedastic linear regression model, and the class of estimators are restricted to be linear unbiased. Here, "linear" means the estimator is a linear function of $\mathbf{y}$. In other words, the estimator, say, $\widetilde{\boldsymbol{\beta}}$, can be written as

$$\widetilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\mathbf{u},$$

where $\mathbf{A}$ is any $n \times k$ matrix of $\mathbf{X}$. Unbiasedness implies that $E[\widetilde{\boldsymbol{\beta}}|\mathbf{X}] = E[\mathbf{A}'\mathbf{y}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ or $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$. In this case, $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{A}'\mathbf{u}$, so under homoskedasticity,

$$Var\left(\widetilde{\boldsymbol{\beta}}|\mathbf{X}\right) = \mathbf{A}'Var(\mathbf{u}|\mathbf{X})\mathbf{A} = \mathbf{A}'\mathbf{A}\sigma^2.$$

The Gauss-Markov Theorem states that the best choice of $\mathbf{A}'$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in the sense that this choice of $\mathbf{A}$ achieves the smallest variance.

Theorem 1 *In the homoskedastic linear regression model, the best (minimum-variance) linear unbiased estimator (BLUE) is the LSE.*

Proof. Given that the variance of the LSE is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ and that of $\widetilde{\boldsymbol{\beta}}$ is $\mathbf{A}'\mathbf{A}\sigma^2$. It is sufficient to show that $\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} \geq 0$. Set $\mathbf{C} = \mathbf{A} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. Note that $\mathbf{X}'\mathbf{C} = 0$. Then we calculate that

$$\begin{aligned}
\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} &= \left(\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right)'\left(\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) - (\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}'\mathbf{C} + \mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}'\mathbf{C} \geq 0.
\end{aligned}$$

■

The scope of the Gauss-Markov Theorem is quite limited given that it requires the class of estimators to be linear unbiased and the model to be homoskedastic. This leaves open the possibility that a nonlinear or biased estimator could have lower mean squared error (MSE) than the LSE in a heteroskedastic model, where the MSE of an estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as $E\left[\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right]$. To exclude such possibilities, we need asymptotic arguments. Chamberlain (1987) shows that in the model $y = \mathbf{x}'\boldsymbol{\beta} + u$, if the *only* available information is $E[\mathbf{x}u] = 0$ or ($E[u|\mathbf{x}] = 0$ and $E[u^2|\mathbf{x}] = \sigma^2$), then among *all* estimators, the LSE achieves the lowest asymptotic MSE;[5] in other words, the LSE is semi-parametrically efficient. To study the efficiency theory, you may start from Newey (1990a). For a more comprehensive treatment, see Bickel et al. (1998).

Exercise 10 *Show that* $E\left[\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\right)'\right] = Var\left(\widetilde{\boldsymbol{\beta}}\right)+\left(E\left[\widetilde{\boldsymbol{\beta}}\right]-\boldsymbol{\beta}\right)\left(E\left[\widetilde{\boldsymbol{\beta}}\right]-\boldsymbol{\beta}\right)'$. *When* $\dim(\boldsymbol{\beta})=1$, *how to interpret this decomposition?*

4.1 Geometry of the Gauss-Markov Theorem (*)

We are studying the variance matrices of estimators. Ellipse can represent *all* of the information in general variance matrices. Given the random vector $\mathbf{y}\in\mathbb{R}^n$ with the variance matrix $\mathbf{\Omega}$, the **variance ellipse** $\mathbb{V}_{\mathbf{y}}$ of $\mathbf{y}$ is the set

$$\mathbb{V}_{\mathbf{y}}=\left\{\mathbf{w}=\mathbf{\Omega a}|\mathbf{a}\in\mathbb{R}^n,\mathbf{a}'\mathbf{\Omega a}\leq 1\right\},$$

where $\mathbf{a}'\mathbf{\Omega a}=Var(\mathbf{a}'\mathbf{y})$.[6] $\mathbb{V}_{\mathbf{y}}$ is a subset in the linear space *span*$(\mathbf{\Omega})$ mainly because $P(\mathbf{y}-E[\mathbf{y}]\in span(\mathbf{\Omega}))=1$.[7] To understand $\mathbb{V}_{\mathbf{y}}$, note that $\mathbb{V}_{\mathbf{y}}$ can be reexpressed as $\left\{\mathbf{w}\in\mathbb{R}^n|\mathbf{w}'\mathbf{\Omega}^{-1}\mathbf{w}\leq 1\right\}$ which is a contour of the density of $\mathbf{y}$ when $\mathbf{y}\sim N(\mathbf{0},\mathbf{\Omega})$. The upper bound of 1 is chosen for convenience. Any constant would serve the same purpose. It is the shape and relative size of a variance ellipse that interest us. In the one-dimensional case,

$$\mathbb{V}_y=\left\{w\in\mathbb{R}|w^2/\sigma^2\leq 1\right\}=\left\{w|-\sigma\leq w\leq\sigma\right\}.$$

Exercise 11 *In the two-dimensional case, show that* $\mathbb{V}_{\mathbf{y}}$ *is an ellipse tangent to a box with dimensions that coincide with the lengths (standard deviations) of the random variables. When* $Corr(y_1,y_2)=1$, *what does* $\mathbb{V}_{\mathbf{y}}$ *look like?*

For a linear transformation of $\mathbf{y}$, say, $\mathbf{z}=\mathbf{Ay}\in\mathbb{R}^k$ for a constant matrix $\mathbf{A}$, the variance ellipse of $\mathbf{z}$ is the image of the variance ellipse of $\mathbf{y}$ under the linear transformation $\mathbf{A}$:

$$\mathbb{V}_{\mathbf{z}}=\left\{\mathbf{w}=\mathbf{A\Omega A}'\mathbf{a}|\mathbf{a}\in\mathbb{R}^k,\mathbf{a}'\mathbf{A\Omega A}'\mathbf{a}\leq 1\right\}=\left\{\mathbf{w}=\mathbf{Av}|\mathbf{v}\in\mathbb{V}_{\mathbf{y}}\right\}.$$

In the homoskedastic linear regression model,

$$\mathbb{V}_{\mathbf{y}}=\left\{\mathbf{w}\in\mathbb{R}^n|\mathbf{w}'\mathbf{w}\leq\sigma^2\right\}.$$

[5] See Back and Brown (1992) for a different derivation.

[6] Malinvaud (1970, pp160-165) gives a derivation of the definition of variance ellipse. According to Malinvaud, Darmois (1945) originally defined the variance ellipse, but he called it the *concentration ellipsoid.*

[7] See Lemma 7.2 of Ruud (2000).

and the variance ellipse of $\Pi(\mathbf{y})$ is

$$\begin{aligned}\mathbb{V}_{\Pi(\mathbf{y})} &= \left\{\mathbf{w}=\sigma^2\mathbf{P}\mathbf{a}|\mathbf{a}\in\mathbb{R}^n, \sigma^2\mathbf{a}'\mathbf{P}\mathbf{a}\le 1\right\} \\ &= \left\{\mathbf{w}=\mathbf{P}\mathbf{b}|\mathbf{b}\in\mathbb{R}^n, (\mathbf{P}\mathbf{b})'(\mathbf{P}\mathbf{b})\le\sigma^2\right\} \\ &= \left\{\mathbf{w}\in span(\mathbf{X})|\mathbf{w}'\mathbf{w}\le\sigma^2\right\} = \left\{\mathbf{w}=\mathbf{P}\mathbf{v}|\mathbf{v}\in\mathbb{V}_{\mathbf{y}}\right\},\end{aligned}$$

where we abbreviate $\mathbf{P}_{\mathbf{X}}$ as $\mathbf{P}$. $\mathbb{V}_{\Pi(\mathbf{y})}$ is a sphere in $span(\mathbf{X})$ and has a similar structure as $\mathbb{V}_{\mathbf{y}}$.

Exercise 12 *What is $\mathbb{V}_{\widehat{\boldsymbol{\beta}}}$ in the homoskedastic linear regression model? Is it a sphere?*

To intuitively understand the Gauss-Markov Theorem, we consider the following simple example. Suppose $n=2$, and $x_i=1$, $i=1,2$. In this case,

$$\widehat{\beta}=\frac{y_1+y_2}{2},$$

and

$$\Pi(\mathbf{y})=\left(\widehat{\beta},\widehat{\beta}\right)'=\mathbf{1}\widehat{\beta}.$$

Given that $\Pi(\mathbf{y})_1=\Pi(\mathbf{y})_2=\widehat{\beta}$, $\mathbb{V}_{\Pi(\mathbf{y})}$ represents the variation in $\widehat{\beta}$. In Figure 2, we use a circle centering at $\mathbf{1}\beta_0$ to represent $\mathbb{V}_{\mathbf{y}}$ since $\widehat{\beta}$ is unbiased to β_0. $\mathbb{V}_{\Pi(\mathbf{y})}$ is the orthogonal projection of all the points in this circle onto $span(\mathbf{X})$; this projection is the interval $[A,B]$ given by the intersection of $\mathbb{V}_{\mathbf{y}}$ and $span(\mathbf{X})$. This interval is also centered at $\mathbf{1}\beta_0$. The variation of a nonorthogonal projection $\mathbf{1}\widetilde{\beta}$ is also the corresponding projection of the interior of $\mathbb{V}_{\mathbf{y}}$ onto $span(\mathbf{X})$. The interval $[A',B']$ in Figure 2 is an example. The nonorthogonal projection along the direction $(1,0)$ yields a larger interval, also centered at $\mathbf{1}\beta_0$, that contains the interval of variation of $\Pi(\mathbf{y})$. So $\widehat{\beta}$ has the smallest variance among all linear unbiased estimators of β_0.

5 Multicollinearity

If rank$(\mathbf{X}'\mathbf{X})<k$, then $\widehat{\boldsymbol{\beta}}$ is not uniquely defined. This is called **strict** (or **exact**) **multicollinearity**. This happens when the columns of $\mathbf{X}$ are linearly dependent, i.e., there is some $\boldsymbol{\alpha}\neq\mathbf{0}$ such that $\mathbf{X}\boldsymbol{\alpha}=\mathbf{0}$. Most commonly, this arises when sets of regressors are included which are identically related. For example, if $\mathbf{X}$ includes a column of ones and both dummies for male and female.[8] When this happens, the applied researcher quickly discovers the error as the statistical software will be unable to construct $(\mathbf{X}'\mathbf{X})^{-1}$. Since the error is discovered quickly, this is rarely a problem for applied econometric practice.

The more relevant issue is **near multicollinearity**, which is often called "multicollinearity" for brevity. This is the situation when the $\mathbf{X}'\mathbf{X}$ matrix is near singular, or when the columns of $\mathbf{X}$ are *close* to be linearly dependent. This definition is not precise, because we have not said

[8] Such multicollinearity is often called the **dummy variable trap**. See Exercise 12 in Chapter 2.

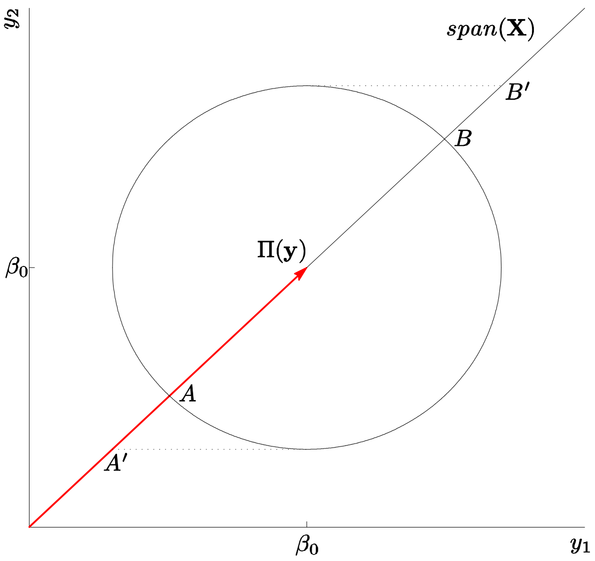

Figure 2: Intuition for the Gauss-Markov Theorem

what it means for a matrix to be "near singular". This is one difficulty with the definition and interpretation of multicollinearity.

One implication of near singularity of matrices is that the numerical reliability of the calculations is reduced. In extreme cases it is possible that the reported calculations will be in error due to floating-point calculation difficulties.

A more relevant implication of near multicollinearity is that individual coefficient estimates will be imprecise. We can see this most simply in a homoskedastic linear regression model with two regressors

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + u_i,$$

and

$$\frac{1}{n}\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In this case,

$$Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n(1-\rho^2)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

The correlation indexes collinearity, since as ρ approaches 1 the matrix becomes singular. We can see the effect of collinearity on precision by observing that the variance of a coefficient estimate $\sigma^2/n(1-\rho^2)$ approaches infinity as ρ approaches 1. Thus the more "collinear" are the regressors, the worse the precision of the individual coefficient estimates. In the general model

$$y_i = x_{1i}\beta_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + u_i,$$

from the discussion following Exercise 8,

$$Var\left(\widehat{\beta}_1|\mathbf{X}\right) = \frac{\sigma^2}{SST_1(1-R_1^2)} = \frac{\sigma^2}{SSR_1}, \tag{3}$$

where SST_1 is the SST of x_1, and R_1^2 is the R-squared from the simple regression of x_1 on $\mathbf{x}_2$ (which includes an intercept). Because the R-squared measures goodness of fit, a value of R_1^2 close to one indicated that $\mathbf{x}_2$ explains much of the variation in x_1 in the sample. This means that x_1 and $\mathbf{x}_2$ are highly correlated. When R_1^2 approaches 1, the variance of $\widehat{\beta}_1$ explodes. $1/(1-R_1^2)$ is often termed as the **variance inflation factor** (VIF). Usually, a VIF larger than 10 should arise our attention.

What is happening is that when the regressors are highly dependent, it is statistically difficult to disentangle the impact of β_1 from that of $\boldsymbol{\beta}_2$. As a consequence, the precision of individual estimates is reduced. Intuitively, β_1 means the effect on y as x_1 changes one unit, holding $\mathbf{x}_2$ fixed. When x_1 and $\mathbf{x}_2$ are highly correlated, you cannot change x_1 while holding $\mathbf{x}_2$ fixed, so β_1 cannot be estimated precisely.

Multicollinearity is a small-sample problem. As larger and larger data sets are available nowadays, i.e., $n >> k$, it is seldom a problem in current econometric practice.

6 Influential Observations and Quantile Regression (*)

The ith observation is **influential** on the least-squares estimate if the deletion of the observation from the sample results in a meaningful change in $\widehat{\boldsymbol{\beta}}$. To investigate the possibility of influential observations, define the leave-one-out least-squares estimator of $\boldsymbol{\beta}$, that is, the OLS estimator based on the sample excluding the ith observation. This equals

$$\widehat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}_{(-i)}\mathbf{y}_{(-i)}, \tag{4}$$

where $\mathbf{X}_{(-i)}$ and $\mathbf{y}_{(-i)}$ are the data matrices omitting the ith row. The relationship between $\widehat{\boldsymbol{\beta}}_{(-i)}$ and $\widehat{\boldsymbol{\beta}}$ is stated in the following exercise.

Exercise 13 *(i) Show that*

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - (1-h_i)^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\widehat{u}_i, \tag{5}$$

where

$$h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

is the ith diagonal element of the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$*. (Hint: If* $\mathbf{T} = \mathbf{A} + \mathbf{CBD}$*, then* $\mathbf{T}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{CB}(\mathbf{B}+\mathbf{BDA}^{-1}\mathbf{CB})^{-1}\mathbf{BDA}^{-1}$*.) (ii) Show that* $0 \le h_i \le 1$ *and* $\sum_{i=1}^n h_i = k$*.*

We can also define the leave-one-out residual

$$\widehat{u}_{i,-i} = y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_{(-i)} = (1 - h_i)^{-1} \widehat{u}_i. \tag{6}$$

A simple comparison yields that

$$\widehat{u}_i - \widehat{u}_{i,-i} = (1 - h_i)^{-1} h_i \widehat{u}_i.$$

As we can see, the change in the coefficient estimate by deletion of the ith observation depends critically on the magnitude of h_i. If the ith observation has a large value of h_i (note that $0 \leq h_i \leq 1$ from the above exercise), then this observation is a **leverage point** and has the potential to be an influential observation. Investigations into the presence of influential observations can plot the values of (6), which is considerably more informative than plots of the uncorrected residuals $\widehat{u}_i$.

Exercise 14 *Let $\widehat{\boldsymbol{\beta}}_n = (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n y_n$ denote the OLS estimate when $\mathbf{y}_n$ is $n \times 1$ and $\mathbf{X}_n$ is $n \times k$. A new observation $(y_{n+1}, \mathbf{x}_{n+1})$ becomes available. Prove that the OLS estimate computed using this additional observation is*

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}}_n + \frac{1}{1 + \mathbf{x}_{n+1}' (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{x}_{n+1}} (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{x}_{n+1} (y_{n+1} - \mathbf{x}_{n+1}' \widehat{\boldsymbol{\beta}}_n).$$

Carefully examining why an observation in the least squares estimation can be influential, we would conclude that this is essentially because the "squared" residuals are used in the objective function of the LSE, $\sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2$, such that the effect of an outlier is magnified. If the objective function is changed to $\sum_{i=1}^n |y_i - \mathbf{x}_i' \boldsymbol{\beta}|$, we expect the effect of an outlier is neglectable. Indeed, this objective function is estimating the conditional median of y_i which is well known to be robust to outliers. Actually, we can extend this objective function to estimate conditional quantiles of y_i. Specifically, given the minimizer $\widehat{\boldsymbol{\beta}}$ in the following **quantile regression** (QR)

$$\min_{\boldsymbol{\beta}} \sum\nolimits_{i=1}^n \rho_\tau \left(y_i - \mathbf{x}_i' \boldsymbol{\beta} \right), \tag{7}$$

$x_i' \widehat{\boldsymbol{\beta}}$ is estimating the τth conditional quantile of y_i given x_i, where $\tau \in (0, 1)$, and

$$\rho_\tau(u) = u(\tau - 1(u \leq 0)) = \begin{cases} -(1-\tau)u & u \leq 0; \\ \tau u & u > 0; \end{cases}$$

is called the **check function**. When $\tau = 0.5$, $\rho_\tau(\cdot)$ is equivalent to the absolute value function. To compare with the objective function of the LSE, we put the "check function" of the least squares u^2 and the check functions for $\tau = 0.5$ and 0.25 in Figure 3. From Figure 3, the contribution of residuals is much less in quantile regression than in least squares when they are off the neighborhood of zero. Quantile regression is introduced by Koenker and Bassett (1978); see Koenker (2005) for an introduction to this rapidly growing topic. Besides robustness, quantile regression provides a more complete picture on the conditional distribution of y_i comparing to the least squares estimation

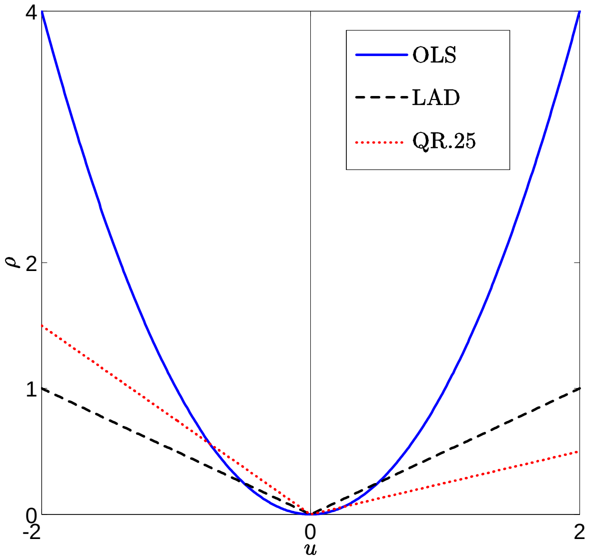

Figure 3: Objective Function for OLS, LAD and QR with $\tau = 0.25$

which considers only the conditional mean.

Exercise 15 *For any predictor $g(\mathbf{x}_i)$ for y_i, the mean absolute error (MAE) is $E\left[|y_i - g(\mathbf{x}_i)|\right]$. Show that the function $g(\mathbf{x})$ which minimizes the MAE is the conditional median $m(\mathbf{x})$ =Med$(y_i|\mathbf{x}_i)$. Based on your answer, can you provide any intuition why (7) will pick out the τth conditional quantile of y_i.*

7 Hypothesis Testing: An Introduction

All the above sections are about estimation. We formally introduce hypothesis testing in this section. Different from an estimation problem where nothing is known about the true parameter, in hypothesis testing, some restrictions about the true parameter are assessed. In other words, there is already a target to attack. Nevertheless, hypothesis testing and estimation are closely related since some test statistics are based on estimators.

The **null hypothesis**, written as H_0, is often a **point hypothesis** $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta})$ is a $q \times 1$ parameter of interest and $\boldsymbol{\theta}_0$ is a hypothesized value.[9] For example, $\boldsymbol{\theta}$ may be a single coefficient, e.g., $\theta = \beta_j$, the difference between two coefficients, e.g., $\theta = \beta_j - \beta_l$, or the ratio of two coefficients, e.g., $\theta = \beta_j/\beta_l$. The complement of the null hypothesis is called the **alternative**

[9] Related concepts are **simple hypothesis** and **composite hypothesis**. A simple hypothesis is any hypothesis which specifies the population distribution completely. A composite hypothesis is any hypothesis which does not specify the population distribution *completely*. A point hypothesis can be either a simple hypothesis or a composite hypothesis (why?). The **one-sided hypothesis** such as $\theta \leq \theta_0$ or $\theta \geq \theta_0$ is not a point hypothesis. We consider only the point null hypothesis in this course.

hypothesis. So the alternative hypothesis, written as H_1, is $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. More generally, letting Θ be the parameter space of $\boldsymbol{\theta}$, we express a null hypothesis as H_0: $\boldsymbol{\theta} \in \Theta_0$ and the alternative hypothesis as H_1: $\boldsymbol{\theta} \in \Theta_1$, where Θ_0 is a proper subset of Θ, $\Theta_0 \cap \Theta_0 = \varnothing$, and $\Theta_0 \cup \Theta_0 = \Theta$. For simplicity, we often refer to the hypotheses as "the null" and "the alternative". In the return-to-schooling example, we may want to know whether education will affect wage. The null is then $\beta_{edu} = 0$, and the alternative is $\beta_{edu} \neq 0$, where β_{edu} is the coefficient of education.

A hypothesis test either accepts the null hypothesis, or rejects the null hypothesis in favor of the alternative hypothesis. The rejection/acceptance dichotomy is associated with the Neyman-Pearson approach to hypothesis testing. We can describe these two decisions as "Accept H_0" and "Reject H_0". The decision is based on the data, and so is a mapping from the sample space to the decision set. This splits the sample space into two regions S_0 and S_1 such that if the observed sample falls into S_0 we accept H_0, while if the sample falls into S_1 we reject H_0. The set S_0 can be called the **acceptance region** and the set S_1 the **rejection** or **critical region**. It is convenient to express this mapping as a real-valued function called a **test statistic**

$$T_n = T_n\left((y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)\right)$$

relative to a **critical value** c. The hypothesis test then consists of the decision rule

1. Accept H_0 if $T_n \leq c$,
2. Reject H_0 if $T_n > c$.

A test statistic T_n should be designed so that small values are likely when H_0 is true and large values are likely when H_1 is true. There is a well developed statistical theory concerning the design of optimal tests. We will not review that theory here, but refer the reader to Lehmann and Romano (2005).

A false rejection of the null hypothesis H_0 (rejecting H_0 when H_0 is true) is called a **Type I error**. The probability of a Type I error is

$$P\left(\text{Reject } H_0 | H_0 \text{ is true}\right) = P\left(T_n > c | H_0 \text{ is true}\right). \tag{8}$$

The finite sample **size** of the test is defined as the supremum of (8) across all data distributions which satisfy H_0. A primary goal of test construction is to limit the incidence of Type I error by bounding the size of the test. A false acceptance of the null hypothesis H_0 (accepting H_0 when H_1 is true) is called a **Type II error**. The rejection probability under the alternative hypothesis is called the **power** of the test, and equals 1 minus the probability of a Type II error:

$$\pi_n\left(\boldsymbol{\theta}\right) = P\left(\text{Reject } H_0 | H_1 \text{ is true}\right) = P\left(T_n > c | H_1 \text{ is true}\right).$$

We call $\pi_n\left(\boldsymbol{\theta}\right)$ the **power function** and is written as a function of $\boldsymbol{\theta}$ to indicate its dependence on the true value of the parameter $\boldsymbol{\theta}$.

In the dominant approach to hypothesis testing, the goal of test construction is to have high power, subject to the constraint that the size of the test is lower than the pre-specified significance level. Generally, the power of a test depends on the true value of the parameter $\boldsymbol{\theta}$, and for a well behaved test the power is increasing both as π moves away from the null hypothesis $\boldsymbol{\theta}_0$ and as the sample size n increases.

Given the two possible states of the world (H_0 or H_1) and the two possible decisions (Accept H_0 or Reject H_0), there are four possible pairings of states and decisions as is depicted in Table 3.

	Accept H_0	Reject H_0
H_0 is true	Correct Decision	Type I Error
H_1 is true	Type II Error	Correct Decision

Table 3: Hypothesis Testing Decisions

Given a test statistic T_n, increasing the critical value c increases the acceptance region S_0 while decreasing the rejection region S_1. This decreases the likelihood of a Type I error (decreases the size) but increases the likelihood of a Type II error (decreases the power). Thus the choice of c involves a trade-off between size and the power. This is why the significance level of the test cannot be set arbitrarily small. (Otherwise the test will not have meaningful power.)

It is important to consider the power of a test when interpreting hypothesis tests, as an overly narrow focus on size can lead to poor decisions. For example, it is trivial to design a test which has perfect size yet has trivial power. Specifically, for any hypothesis we can use the following test: Generate a random variable $U \sim U[0,1]$ and reject H_0 if $U < \alpha$. This test has exact size of α. Yet the test also has power precisely equal to α. When the power of a test equals the size, we say that the test has **trivial power**. Nothing is learned from such a test.

To determine the critical value c, we need to pre-select a **significance level** α such that $P\left(T_n > c | H_0 \text{ is true}\right) = \alpha$, yet there is no objective scientific basis for choice of α. Nevertheless, the common practice is to set $\alpha = 0.05$ (5%). Alternative values are $\alpha = 0.10$ (10%) and $\alpha = 0.01$ (1%). These choices are somewhat the by-product of traditional tables of critical values and statistical software. The informal reasoning behind the choice of a 5% critical value is to ensure that Type I errors should be relatively unlikely - that the decision "Reject H_0" has scientific strength - yet the test retains power against reasonable alternatives. The decision "Reject H_0" means that the evidence is inconsistent with the null hypothesis, in the sense that it is relatively unlikely (1 in 20) that data generated by the null hypothesis would yield the observed test result. In contrast, the decision "Accept H_0" is not a strong statement. It does not mean that the evidence supports H_0, only that there is insufficient evidence to reject H_0. Because of this, it is more accurate to use the label "Do not Reject H_0" instead of "Accept H_0". When a test rejects H_0 at the 5% significance level it is common to say that the statistic is **statistically significant** and if the test accepts H_0 it is common to say that the statistic is not **statistically significant** or that it is **statistically insignificant**. It is helpful to remember that this is simply a way of saying "Using the statistic T_n, the hypothesis H_0 can [cannot] be rejected at the 5% level." When the null hypothesis H_0: $\theta = 0$ is rejected it is common to say that the coefficient θ is statistically significant, because the test has rejected the hypothesis that the coefficient is equal to zero.

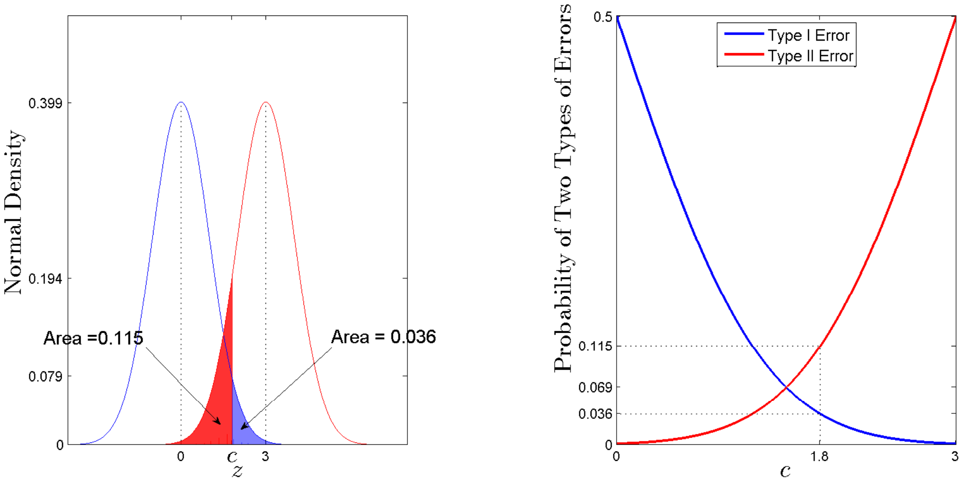

Figure 4: Trade-Off Between the Type I Error and Type II Error

We use a simple example to illustrate the basic concepts of hypothesis testing. Suppose we have only one data point z in hand and we know $z \sim N(\mu, 1)$. We want to test H_0: $\mu = 0$ against H_1: $\mu > 0$. A natural test is to reject H_0 if z is large. Rigorously, the test is $1(z > c)$, where $1(\cdot)$ is the indicator function which equals 1 if the event in the parenthesis is true and 0 otherwise, 1 indicates rejection and 0 acceptance, the test statistic $T_n = z$, and the critical value is c. Set the significance level $\alpha = 0.05$; then c is chosen such that $P(z > c|\mu = 0) = E[1(z > c)|\mu = 0] = 1 - \Phi(c) = 0.05$, i.e., $c = 1.645$. So if $z > 1.645$, we will reject H_0; otherwise, we cannot reject H_0. The power function $\pi(\mu) = P(z > c|\mu) = P(z - \mu > c - \mu) = 1 - \Phi(c - \mu)$ which is an increasing function of μ and a decreasing function of c. It is understandable that $\pi(\mu)$ is increasing with μ since when μ is larger, it is easier to detect $\mu > 0$. That $\pi(\mu)$ is decreasing in c indicates a trade-off between size and the power. Since the power equals 1 minus the probability of the Type II error, it is equivalent to study the trade-off between the probabilities of the Type I error and Type II error. Figure 4 shows this trade-off when the true μ equals 3. The left panel illustrates the probabilities of the Type I error and Type II error when $c = 1.8$, and the right panel illustrates these probabilities as a function of c. From Figure 4, it is obvious that we cannot reduce these two types of errors simultaneously.

In the simple example above, the acceptance region is $z \leq c$ and the critical region is $z > c$, which are quite trivial. To illustrate these regions in a more complicated example, suppose two data points y_1 and y_2 are observed and they follow $N(\mu, 2)$. We want to test H_0: $\mu = 0$ against H_1: $\mu \neq 0$. A natural test is to reject H_0 if the absolute value of $\overline{y} = (y_1 + y_2)/2$ is large. Given that $\overline{y}$ follows $N(\mu, 1)$, the 5% critical value is 1.96 since $P(|\overline{y}| > 1.96|\mu = 0) = 0.05$. The acceptance region is $\{(y_1, y_2)| \, |\overline{y}| \leq 1.96\}$ or $\{(y_1, y_2)| - 3.92 - y_1 \leq y_2 \leq 3.92 - y_1\}$. Figure 5 shows this region based on (y_1, y_2) and $\overline{y}$.

In summary, one hypothesis testing includes the following steps. First, specify the null and

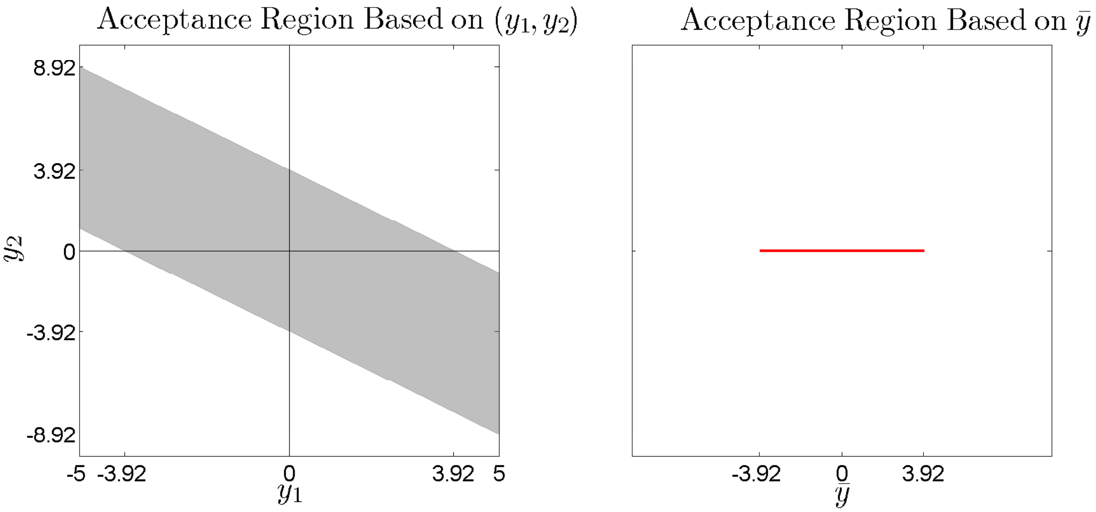

Figure 5: Acceptance Region Based on (y_1, y_2) and $\overline{y}$

alternative. Second, construct the test statistic. Third, derive the distribution of the test statistic under the null. Fourth, determine the decision rule (acceptance and rejection regions) by specifying a level of significance. Fifth, study the power of the test.

8 LSE as a MLE

Another motivation for the LSE can be obtained from the **normal regression model**.[10] This is the linear regression model with the additional assumption

Assumption OLS.4 (normality): $u|\mathbf{x} \sim N(0, \sigma^2)$.

That is, the error u_i is independent of $\mathbf{x}_i$ and has the distribution $N(0, \sigma^2)$. This is a parametric model, where likelihood methods can be used for estimation, testing, and distribution theory. In this setup, we can get *exact* distributions of many statistics. In other words, the advantage of knowing the distribution of the error term is to make use of the finite-sample information. But if the distribution assumption is not right, then the exact distribution is not right, so we need study the robustness of the model. Anyway, this is part of the classical theory in least squares estimation.

We restate the assumptions of the model in matrix notation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},\ \text{rank}(\mathbf{X}) = k,\ \mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n\sigma^2).$$

Note that $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n\sigma^2)$ implies that $\mathbf{u}$ is independent of $\mathbf{X}$ (which obviously implies $E[u|\mathbf{x}] = 0$ and $E[u^2|\mathbf{x}] = \sigma^2$). This is a very strong assumption, e.g., for time series or panel data, this

[10] Gauss proposed the normal regression model, and derived the LSE as the MLE for this model. Gauss's assumption of normality was justified by Laplace's simultaneous discovery of the central limit theorem (which will be discussed in the next chapter). This is also why the normal distribution is also called the Gaussian distribution.

assumption is hard to hold. In the discussion below, we always condition on $\mathbf{X}$; i.e., we always assume $\mathbf{X}$ is fixed.

Since we have the distribution of $\mathbf{y}$ conditional on $\mathbf{X}$, we can estimate $\boldsymbol{\beta}$ and σ^2 by the conditional MLE. Here, the average log likelihood is

$$\ell_n\left(\boldsymbol{\beta},\sigma^2\right)=\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right)\right)=-\frac{1}{2}\log\left(2\pi\right)-\frac{1}{2}\log\left(\sigma^2\right)-\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2},$$

so

$$\left(\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2\right)_{MLE}=\arg\max_{\boldsymbol{\beta},\sigma^2}\ell_n\left(\boldsymbol{\beta},\sigma^2\right)=\left(\begin{array}{c}\widehat{\boldsymbol{\beta}}_{OLS}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\\ \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{n}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{OLS})^2\end{array}\right).$$

Before studying the properties of the MLE, we first define a term, UMVUE.

Definition 1 (UMVUE) *An estimator $\widehat{\boldsymbol{\theta}}$ is the* best unbiased estimator *or the* uniform minimum variance unbiased estimator *of $\boldsymbol{\theta}$ if $E_{\boldsymbol{\theta}}\left[\widehat{\boldsymbol{\theta}}\right]=\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$ and, for any other estimator $\widetilde{\boldsymbol{\theta}}$ with $E_{\boldsymbol{\theta}}\left[\widetilde{\boldsymbol{\theta}}\right]=\boldsymbol{\theta}$, we have $Var_{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}\right)\leq Var_{\boldsymbol{\theta}}\left(\widetilde{\boldsymbol{\theta}}\right)$ for all $\boldsymbol{\theta}$, where the subscript $\boldsymbol{\theta}$ indicates that the expectation is taken under $\boldsymbol{\theta}$.*

The following theorem shows that although the MLE of $\boldsymbol{\beta}$ is the UMVUE, the MLE of σ^2 is not since it is not unbiased. Nevertheless, the bias-corrected version of $\widehat{\sigma}^2$, s^2, is the UMVUE of σ^2.

Theorem 2 (Finite-Sample Properties) *In the normal regression model,*
(i) $\forall$ $\mathbf{a}\in\mathbb{R}^k$, $\mathbf{a}\cdot\widehat{\boldsymbol{\beta}}$ is the UMVUE of $\mathbf{a}\cdot\boldsymbol{\beta}$;
(ii) s^2 is the UMVUE of σ^2.

The proof of this result is out of the scope of this course; interested readers are referred to Theorem 3.7 of Shao (2003) for the details.

We now discuss the distribution theory in the normal regression model. We will discuss two tests, the t-test and F-test. The former is named after W.S. Gosset (1876-1937), and the later is named after R.A. Fisher. Since we only know the distribution of $\mathbf{u}$, every statistic must be converted to a function of $\mathbf{u}$. The basic result is

$$\left(\begin{array}{c}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\mathbf{u}}\end{array}\right)=\left(\begin{array}{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\\ \mathbf{M}\end{array}\right)\mathbf{u}\sim N\left(\mathbf{0},\left(\begin{array}{cc}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}&\mathbf{0}\\ \mathbf{0}&\sigma^2\mathbf{M}\end{array}\right)\right). \tag{9}$$

$\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ is independent of $\widehat{\mathbf{u}}$ since $\mathbf{M}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}=\mathbf{0}$.

We first discuss the famous t-statistic which is designed to test H_0: $\beta_j=\beta_{j0}$ against H_1: $\beta_j\neq\beta_{j0}$. The t-statistic is defined as

$$t=\frac{\widehat{\beta}_j-\beta_{j0}}{s\left(\widehat{\beta}_j\right)}$$

where $s\left(\widehat{\beta}_j\right)=s\sqrt{[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$ is a natural estimator of the standard deviation of $\widehat{\beta}_j$, $\sqrt{Var\left(\widehat{\beta}_j\right)}$, given that $E[s^2|\mathbf{X}] = \sigma^2$. $s\left(\widehat{\beta}_j\right)$ is called a **standard error** for $\widehat{\beta}_j$ (a standard error for an estimator is an estimate of the standard deviation of that estimator).

Theorem 3 *In the normal regression model, under H_0,*

(i) $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$, *where χ^2_{n-k} is the chi-square distribution with $n-k$ degrees of freedom.*

(ii) $t \sim t_{n-k}$, *where t_{n-k} is the t-distribution with $n-k$ degrees of freedom.*

Proof. Recall that for any idempotent matrix $\mathbf{M}$ with rank m, there exists an orthogonal matrix $\mathbf{H}$ such that $\mathbf{M} = \mathbf{H}\begin{pmatrix} \mathbf{I}_m & 0 \\ 0 & 0 \end{pmatrix}\mathbf{H}'$. So

$$\begin{aligned} \frac{(n-k)\,s^2}{\sigma^2} &= \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{\sigma^2} = \frac{\mathbf{u}'\mathbf{M}\mathbf{u}}{\sigma^2} = \frac{1}{\sigma^2}\mathbf{u}'\mathbf{H}\begin{pmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{H}'\mathbf{u} \\ &= \mathbf{v}'\begin{pmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{v} \sim \chi^2_{n-k}. \end{aligned}$$

Here $\mathbf{v} = \mathbf{H}'\mathbf{u}/\sigma \sim N(\mathbf{0}, \mathbf{H}'\mathbf{H}) = N(\mathbf{0}, \mathbf{I}_n)$. Note that when n goes to infinity, $\frac{(n-k)s^2}{\sigma^2}$ should diverge (why?).

$$\frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} = \frac{\widehat{\beta}_j - \beta_j}{s\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}} \sim \frac{N(0, \sigma^2\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj})}{\sqrt{\frac{\sigma^2}{n-k}\chi^2_{n-k}}\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-k}}{n-k}}} \sim t_{n-k}.$$

The last step is from (9), i.e., $\left(\widehat{\beta}_j - \beta_j\right)$ is independent of $s(\widehat{\beta}_j)$. An interesting observation is that this t-distribution converge to standard normal when n goes to infinity (why?). ■

Exercise 16 *Show that $\frac{SSE}{\sigma^2} \sim \chi^2_{k-1}$ if $\underline{\boldsymbol{\beta}} = \mathbf{0}$, where $\underline{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}$ excluding the intercept.*

Exercise 17 *Suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$ with $\mathbf{D}$ being a known diagonal matrix. Show that $WSSR_U \equiv \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{D}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \sim \chi^2_{n-k}$, where $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}$ is a GLS estimator.*

We next consider the general linear hypothesis testing problem[11]:

$$\begin{aligned} H_0 &: \mathbf{R}'\boldsymbol{\beta} = \mathbf{c}, \\ H_1 &: \mathbf{R}'\boldsymbol{\beta} \neq \mathbf{c}, \end{aligned}$$

[11] Why we do not consider the nonlinear hypothesis testing in the finite-sample environment?

where $\mathbf{R}$ is a $k \times q$ matrix with $q \leq k$. Suppose the estimate under H_0 is $\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2\right)$; then

$$\begin{aligned}
\ell_n\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2\right) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\left(\frac{SSR_U}{n}\right), \\
\ell_n\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2\right) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\left(\frac{SSR_R}{n}\right),
\end{aligned}$$

where SSR_U is the SSR of the unrestricted model and SSR_R is the SSR of the restricted model.

Exercise 18 *Show that*

$$\begin{aligned}
SSR_R - SSR_U &= \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right)'(\mathbf{X}'\mathbf{X})\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right) = \left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left[\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right) \\
&= \widehat{\boldsymbol{\lambda}}'\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\widehat{\boldsymbol{\lambda}} = \left(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)'\mathbf{P}\left(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right).
\end{aligned}$$

(Hint: Use the results in Exercise 26 of Chapter 2).

Theorem 4 *In the normal regression model, under* H_0*,*

$$\begin{aligned}
F &= \frac{\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left[\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)/q}{s^2} \textit{(Wald Principle)} \\
&= \left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left[\mathbf{R}' \cdot \widehat{Var(\widehat{\boldsymbol{\beta}}|\mathbf{X})} \cdot \mathbf{R}\right]^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)/q \\
&= \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k)} \textit{(Likelihood-Ratio Principle)} \\
&= \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)/q}{\widehat{\sigma}^2/(n-k)} = \frac{\left(R_U^2 - R_R^2\right)/q}{(1 - R_U^2)/(n-k)} \sim F_{q,n-k}.
\end{aligned}$$

Proof. Under H_0, $\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right) \sim N\left(\mathbf{0}, \sigma^2\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right)$, so

$$\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left[\sigma^2\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right) \sim \chi_q^2$$

From Theorem 3(i), $\frac{(n-k)s^2}{\sigma^2} \sim \chi_{n-k}^2$. Also, from (9), they are independent, so

$$\frac{\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left[\sigma^2\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)/q}{\frac{(n-k)s^2}{\sigma^2}/(n-k)} = F \sim F_{q,n-k}$$

■

Some special cases of the F test are listed as follows:

(i) $\mathbf{R}' = [0, \cdots, 0, 1, 0, \cdots, 0]$, where 1 is in the jth position, $c = \beta_{j0}$, and $q = 1$. It is easy to check that $F = t^2$, so $F_{1,n-k} = t_{n-k}^2$.

(ii) $\mathbf{R}' = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{(k-1)\times k}$, $\mathbf{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{(k-1)\times 1}$. This is usually called the *significance test* since it tests whether all slopes are zero. Since $R_R^2 = 0$ in this case, $F = \frac{R_U^2/q}{(1-R_U^2)/(n-k)}$. F associated with this choice of $\mathbf{R}$ is a popular statistic in the early days of econometric reporting, when sample sizes were very small and researchers wanted to know if there was "any explanatory power" to their regression. This is rarely an issue today, as sample sizes are typically sufficiently large that this F statistic is highly "significant". While there are special cases where this F statistic is useful, these cases are atypical.

Exercise 19 *In the setup of Exercise 17, show that F with R^2 replaced by Buse (1973)'s R^2,*

$$R^2 = \frac{WSSR_R - WSSR_U}{WSSR_R}$$

follows $F_{k-1,n-k}$ under the null (ii) above, where $WSSR_R = \mathbf{y}'\mathbf{D}^{-1}\mathbf{y} - \frac{(\mathbf{y}'\mathbf{D}^{-1}\mathbf{1})^2}{\mathbf{1}'\mathbf{D}^{-1}\mathbf{1}}$.

(iii) Chow (1960) considered the test $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ in the regression $y = d \cdot \mathbf{x}'\boldsymbol{\beta}_1 + (1-d)\mathbf{x}'\boldsymbol{\beta}_2 + u$, where d is a dummy variable. This test intends to check whether the independent variables have different impacts on different subgroups of the population. The Chow test is most commonly used in time series analysis to test for the presence of a structural break. Figure 6 shows the restricted and unrestricted model, where the left panel is for the structural break case ($d = 1(x \leq \gamma)$ for some known constant γ). In this test, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' \in \mathbb{R}^{2k}$, $\mathbf{R}' = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & -1 \end{pmatrix}_{k\times 2k}$, $\mathbf{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{k\times 1}$. SSR_R is the SSR in the restricted regression $y = \mathbf{x}'\boldsymbol{\beta} + u$, and $SSR_U = SSR_0 + SSR_1$, where SSR_0 is the SSR in the regression $y = \mathbf{x}'\boldsymbol{\beta} + u$ using only the data with $d = 0$, and SSR_1 is similarly defined. A key assumption for the F statistic to follow $F_{k,n-2k}$ is homoskedasticity, that is, the error variances for the two groups are the same. Extensions of the Chow test to the heteroskedasticity and unknown γ case are available in the current econometric literature.

If you are presented with an F statistic but don't have access to critical values, a useful rule of thumb is to know that for large n, the 5% asymptotic critical value is decreasing as q increases, and is less than 2 for $q \geq 7$.

The powers of the t-test and F-test are related to the non-central t-distribution and the non-central F-distribution and will not be discussed in this course since they are not that popular nowadays.

The t-test and F-test are finite-sample tests, i.e., they assume that n is fixed. While elegant, the difficulty in applying these tests is that the normality assumption is too restrictive to be empirically

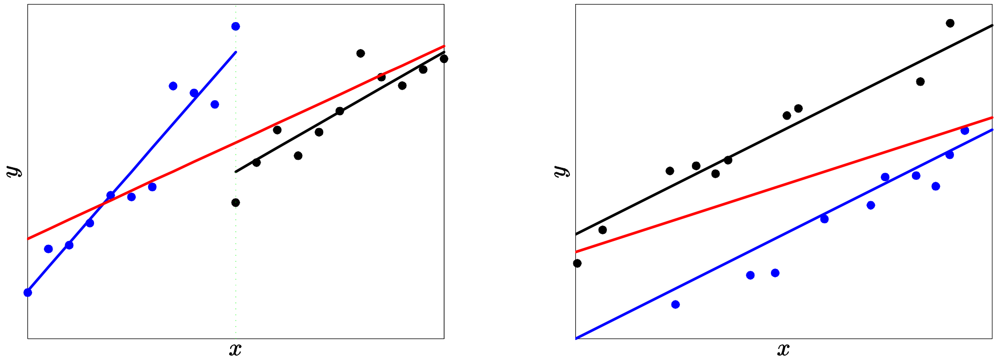

Figure 6: Restricted and Unrestricted Model in the Chow Test

plausible, and therefore inference based on these tests has no guarantee of accuracy. We develop an alternative inference theory based on large sample (asymptotic) approximations in the following chapter.

Exercise 20 (Empirical) *Use the data from the empirical exercise in Chapter 2. Numerically calculate the following:*

(a) $\sum_{i=1}^{n} \widehat{u}_i^2$

(b) $\widehat{\sigma}^2$ *and* s^2

(c) R^2, $\overline{R}^2$ *and* R_u^2

(d) *In the second-stage residual regression, (the regression of the residuals on the residuals), calculate the equation* R^2 *and sum of squared errors. Do they equal the values from the initial OLS regression? Explain.*