Chapter 8
Restricted Estimation

8.1 Introduction

In the linear projection model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

a common task is to impose a constraint on the coefficient vector. For example, partitioning

$$\boldsymbol{x}_i' = \left(\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}'\right) \text{ and } \boldsymbol{\beta}' = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right),$$

a typical constraint is an exclusion restriction of the form beta_2 = 0: In this case the constrained model is

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0}. \end{aligned}$$

At first glance this appears the same as the linear projection model, but there is one important difference: **the error e_i is uncorrelated with the entire regressor vector x’_i = (x’_1i; x’_2i) not just the included regressor x_1i**.

In general, a set of q linear constraints on beta takes the form

$$\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}$$

where R is k-by-q; rank(R) = q < k and c is q-by-1.

The assumption that R is **full rank** means that **the constraints are linearly independent** (**there are no redundant or contradictory constraints**). We can define the restricted parameter space B_R as the set of values of which satisfy the linear constraint, that is

$$\boldsymbol{B}_{\boldsymbol{R}} = \left\{\boldsymbol{\beta} : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}\right\}.$$

We will call the set of linear constraints **a constraint** or **a restriction**.

We will call estimators which satisfy the linear constraints **constrained estimators** or **restricted estimators**.

The constraint beta_2 = 0 discussed above is a special case of the constraint with

$$\boldsymbol{R} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{I}_{k_2} \end{pmatrix},$$

a **selector matrix**, and c = 0.

Another common restriction is that a set of coefficients sum to a known constant, i.e. beta_1 + beta_2 =1. For example, this constraint arises in a constant-return-to-scale production function.

Other common restrictions include the equality of coefficients beta_1 = beta_2, and equal and offsetting coefficients beta_1 = -beta_2.

A typical reason to impose a constraint is that **we believe (or have information) that the constraint is true**. By imposing the constraint, we hope to improve **estimation efficiency**. The goal is to obtain **consistent estimates** with **reduced variance** relative to the unconstrained estimator.

The questions then arise: How should we estimate the coefficient vector imposing the linear restriction? If we impose such constraints, what is the sampling distribution of the resulting estimator? How should we calculate standard errors? These are the questions explored in this chapter.

8.2 Constrained Least Squares

An intuitively appealing method to estimate a constrained linear projection is to minimize the least-squares criterion subject to the constraint

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\, SSE(\boldsymbol{\beta})$$

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

We call beta~_cls the constrained least-squares (CLS) estimator.

One method to find the solution to (8.3) uses the technique of Lagrange multipliers. The problem is equivalent to the minimization of the Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} SSE(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{c}\right)$$

over (beta,lambda), where lambda is an s-by-1 vector of Lagrange multipliers.

The first-order conditions for minimization of the Lagrangian are

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\lambda}}_{\text{cls}}) = -\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{X}'\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\text{cls}} + \boldsymbol{R}\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \mathbf{0}$$

$$\frac{\partial}{\partial \boldsymbol{\lambda}}\mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\lambda}}_{\text{cls}}) = \boldsymbol{R}'\widetilde{\boldsymbol{\beta}} - \boldsymbol{c} = \mathbf{0}.$$

Premultiplying (8.6) by R’(X’X)^{-1} we obtain

$$-\boldsymbol{R}'\widehat{\boldsymbol{\beta}} + \boldsymbol{R}'\widetilde{\boldsymbol{\beta}}_{\text{cls}} + \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \mathbf{0}$$

where

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

is the unrestricted least-squares estimator.

Imposing

$$\boldsymbol{R}'\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{c} = \mathbf{0}$$

we find

$$\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \left[\boldsymbol{R}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right).$$

Notice that (X’X)^{-1} > 0 and R full rank imply that R’(X’X)^{-1}R > 0 and is hence invertible.

We find the solution to the constrained minimization problem

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}}_{\text{ols}} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{R}\left[\boldsymbol{R}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{c}\right).$$

(See Exercise 8.5 to verify that the CLS estimator satisfies the constraint.)

This is a general formula for the CLS estimator. It also can be written as

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}} = \widehat{\boldsymbol{\beta}}_{\mathrm{ols}} - \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\mathrm{ols}} - \boldsymbol{c}\right).$$

The CLS residuals are

$$\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}}$$

and the n-by-1 vector of residuals are written in vector notation as e~.

To illustrate, we generated a random sample of 100 observations for the variables (y_i, x_1i, x_2i) and calculated the sum of squared errors function for the regression of yi on x_1i and x_2i.

Figure 8.1 displays contour plots of the sum of squared errors function. The center of the contour plots is the least squares minimizer beta^_ols = (0.33; 0.26)'.

Suppose it is desired to estimate the coefficients subject to the constraint beta_1 + beta_2 = 1. This constraint is displayed in the figure by the straight line.

The constrained least squares estimator is the point on this straight line which yields the smallest sum of squared errors, which is the point which intersects with the lowest contour plot.

The solution is the point where a contour plot is tangent to the constraint line, and marked as beta~_cls = (0:52; 0:48)'.

In Stata, constrained least squares is implemented using the cnsreg command.

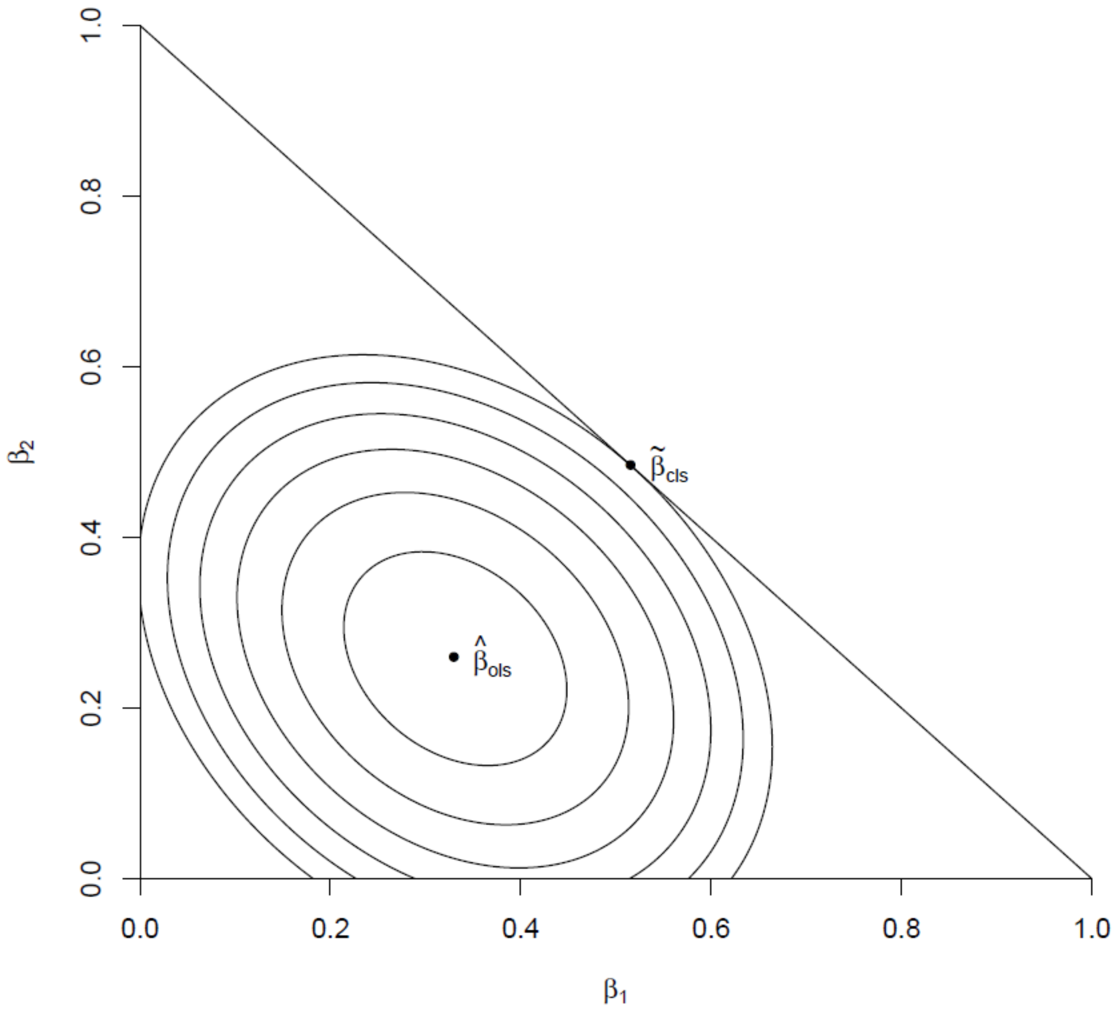


Figure 8.1: Imposing a Constraint on the Least Squares Criterion

8.3 Exclusion Restriction

While (8.8) is a general formula for the CLS estimator, in most cases the estimator can be found by applying least-squares to a **reparameterized** equation.

To illustrate, let us return to the first example presented at the beginning of the chapter - a simple exclusion restriction. Recall the unconstrained model is

$$y_i = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 + \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2 + e_i$$

the exclusion restriction is beta_2 = 0, and the constrained equation is

$$y_i = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 + e_i.$$

In this setting the CLS estimator is OLS of y_i on x_1i. (See Exercise 8.1.) We can write this as

$$\widetilde{\boldsymbol{\beta}}_1 = \left(\sum_{i=1}^{n} \boldsymbol{x}_{1i}\boldsymbol{x}'_{1i}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_{1i}y_i\right).$$

The CLS estimator of the entire vector is

$$\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix}.$$

It is not immediately obvious, but (8.8) and (8.13) are algebraically (and numerically) equivalent.

To see this, the first component of (8.8) with (8.2) is

$$\widetilde{\boldsymbol{\beta}}_1 = \begin{pmatrix} \boldsymbol{I}_{k_2} & \mathbf{0} \end{pmatrix} \left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \begin{pmatrix} \mathbf{0} \\ \boldsymbol{I}_{k_2} \end{pmatrix} \left[\begin{pmatrix} \mathbf{0} & \boldsymbol{I}_{k_2} \end{pmatrix} \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \begin{pmatrix} \mathbf{0} \\ \boldsymbol{I}_{k_2} \end{pmatrix} \right]^{-1} \begin{pmatrix} \mathbf{0} & \boldsymbol{I}_{k_2} \end{pmatrix} \widehat{\boldsymbol{\beta}} \right].$$

Using (3.40) this equals

$$
\begin{aligned}
\widetilde{\boldsymbol{\beta}}_1 &= \widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{Q}}^{12}\left(\widehat{\boldsymbol{Q}}^{22}\right)^{-1}\widehat{\boldsymbol{\beta}}_2 \\
&= \widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\boldsymbol{\beta}}_2 \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2y}\right) \\
&+ \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\left(\widehat{\boldsymbol{Q}}_{2y} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\right)\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y} \\
&= \widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}
\end{aligned}
$$

which is (8.13) as originally claimed.

8.4 Finite Sample Properties

In this section we explore some of the properties of the CLS estimator in the linear regression model

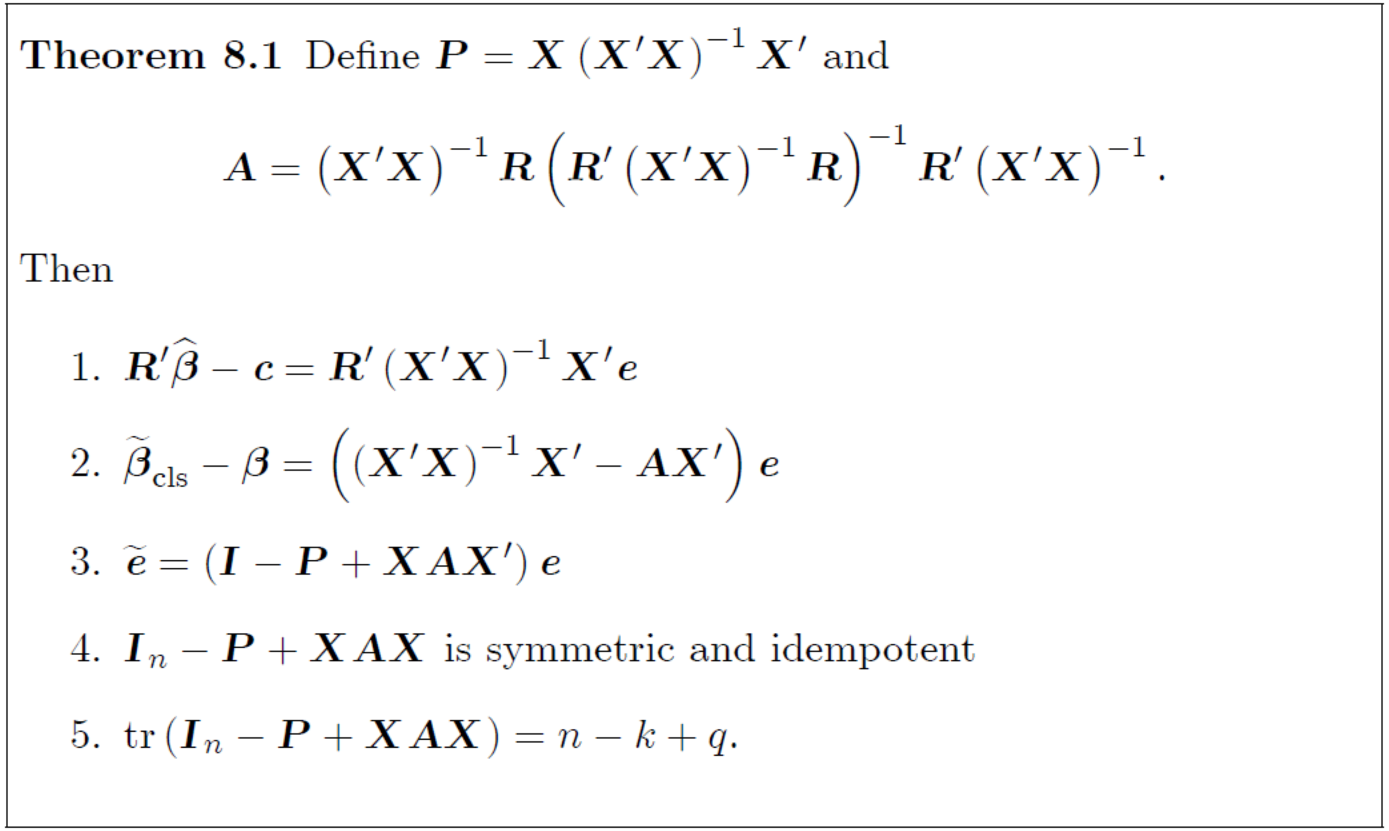
$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned}$$

First, it is useful to write the **estimator** and the **residuals** as **linear functions of the error vector**. These are algebraic relationships and **do not rely on the linear regression assumptions**.

For a proof, see Exercise 8.6.

Given the linearity of Theorem 8.1.2, it is not hard to show that the CLS estimator is unbiased for beta.

For a proof, see Exercise 8.7.

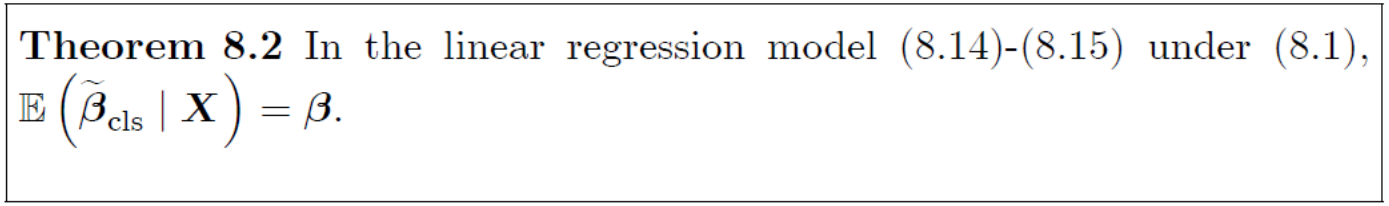


Theorem 8.1 Define $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ and

$$\boldsymbol{A} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

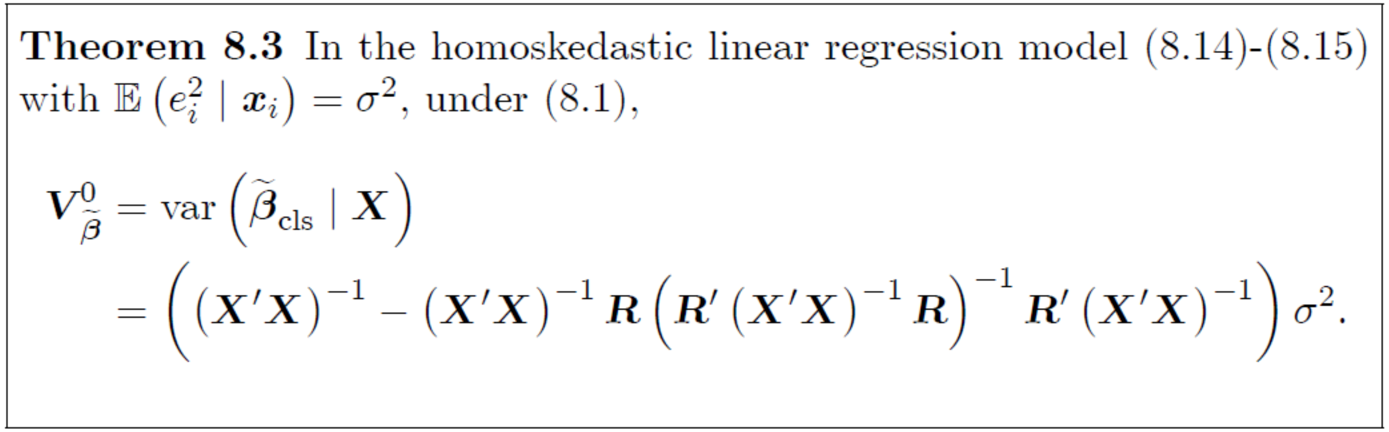
Then

1. $\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$
2. $\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta} = \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$
3. $\widetilde{\boldsymbol{e}} = \left(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$
4. $\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}$ is symmetric and idempotent
5. $\operatorname{tr}\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}\right) = n - k + q.$



Theorem 8.2 In the linear regression model (8.14)-(8.15) under (8.1), $\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$.

Given the linearity we can also calculate the variance matrix of beta~_cls. For this we will add the assumption of conditional homoskedasticity to simplify the expression. For a proof, see Exercise 8.8.



Theorem 8.3 In the homoskedastic linear regression model (8.14)-(8.15) with $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, under (8.1),

$$\begin{aligned}\boldsymbol{V}^0_{\widetilde{\boldsymbol{\beta}}} &= \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} \mid \boldsymbol{X}\right) \\ &= \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\sigma^2.\end{aligned}$$

We use the V^0_{beta~_cls} notation to emphasize that this is the variance matrix under the assumption of **conditional homoskedasticity**.

For inference we need an estimate of V^0_{beta~_cls}. A natural estimator is

$$\widehat{\mathbf{V}}^0_{\widetilde{\beta}} = \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R} \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \right) s^2_{\text{cls}}$$

where

$$s^2_{\text{cls}} = \frac{1}{n-k+q} \sum_{i=1}^{n} \widetilde{e}_i^2$$

is a **biased-corrected** estimator of sigma^2. Standard errors for the components of beta are then found by taking the squares roots of the diagonal elements of V^_{beta~}, for example

$$s(\widehat{\beta}_j) = \sqrt{\left[\widehat{\mathbf{V}}^0_{\widetilde{\beta}}\right]_{jj}}.$$

The estimator (8.16) has the property that it is unbiased for sigma^2 under conditional homoskedasticity. To see this, using the properties of Theorem 8.1,

$$\begin{aligned}(n-k+q)\, s_{\text{cls}}^2 &= \widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}} \\ &= \boldsymbol{e}'\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e} \\ &= \boldsymbol{e}'\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}.\end{aligned}$$

We defer the remainder of the proof to Exercise 8.9.

Theorem 8.4 In the homoskedastic linear regression model (8.14)-(8.15) with $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, under (8.1), $\mathbb{E}\left(s_{\text{cls}}^2 \mid \boldsymbol{X}\right) = \sigma^2$ and $\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widetilde{\beta}}^0 \mid \boldsymbol{X}\right) = \boldsymbol{V}_{\widetilde{\beta}}^0$.

Now consider the distributional properties in the **normal regression model**

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$$
$$e_i \sim \mathrm{N}(0, \sigma^2).$$

By the linearity of Theorem 8.1.2, conditional on X, e

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta}$$

is **normal**. Given Theorems 8.2 and 8.3, we deduce that

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} \sim \mathrm{N}(\boldsymbol{\beta}, \mathbf{V}^0_{\widetilde{\boldsymbol{\beta}}}).$$

Similarly, from Exercise 8.1 we know

$$\widetilde{\boldsymbol{e}} = (\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}')\,\boldsymbol{e}$$

is linear in e so is also **conditionally normal**. Furthermore, since

$$\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \boldsymbol{X}\boldsymbol{A}\right) = 0$$

e~ and beta~_cls are **uncorrelated** and thus **independent**. Thus s^2_cls and beta~_cls are **independent**.

From (8.17) and the fact that I_n - P + XAX' is idempotent with rank n - k + q, it follows that

$$s^2_{\text{cls}} \sim \sigma^2 \chi^2_{n-k+q} / \left(n - k + q\right)$$

It follows that the t-statistic has the exact distribution

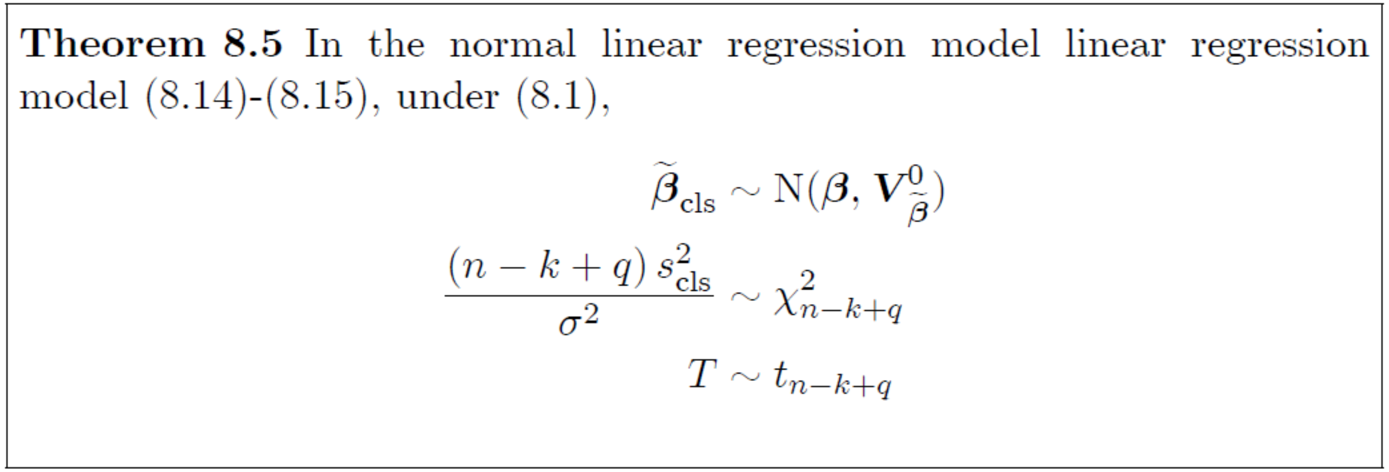
$$\begin{aligned} T &= \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} \\ &\sim \frac{\mathrm{N}\left(0,1\right)}{\sqrt{\chi^2_{n-k+q} / \left(n-k+q\right)}} \\ &\sim t_{n-k+q} \end{aligned}$$

a student t distribution with n - k + q degrees of freedom.

The relevance of this calculation is that the “**degrees of freedom**” for a CLS regression problem equal n - k+q rather than n - k as in the OLS regression problem.

Essentially, the model has k - q free parameters instead of k. Another way of thinking about this is that estimation of a model with k coefficients and q restrictions is equivalent to estimation with k - q coefficients.

We summarize the properties of the normal regression model.



Theorem 8.5 In the normal linear regression model linear regression model (8.14)-(8.15), under (8.1),

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} \sim \mathrm{N}(\boldsymbol{\beta}, \mathbf{V}^{0}_{\widetilde{\boldsymbol{\beta}}})$$

$$\frac{(n-k+q)\, s^2_{\text{cls}}}{\sigma^2} \sim \chi^2_{n-k+q}$$

$$T \sim t_{n-k+q}$$

An interesting relationship is that in the **homoskedastic regression model**

$$
\begin{aligned}
\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) &= \mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta}\right)'\right) \\
&= \mathbb{E}\left(\left(\boldsymbol{A}\boldsymbol{X}'\right)\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} - \boldsymbol{X}\boldsymbol{A}\right)\right)\sigma^2 = 0
\end{aligned}
$$

So (beta^_ols - beta~_cls) and (beta~_cls) are uncorrelated and hence independent. One corollary is

$$
\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}, \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) = \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)
$$

A second corollary is

$$
\begin{aligned}
\operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right) - \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2.
\end{aligned}
$$

This also shows us the difference between the CLS and OLS variances

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right)-\operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2\geq 0$$

the final equality meaning positive semi-definite. It follows that

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right)\geq\operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)$$

in the positive definite sense, and thus **CLS is more efficient than OLS**. **Both estimators are unbiased** (in the **linear regression model**), and **CLS has a lower variance matrix** (in the **linear homoscedastic regression model**).

The relationship (8.18) is rather interesting and will appear again. The expression says that the variance of the difference between the estimators is equal to the difference between the variances. This is rather special. It occurs (generically) when we are comparing an efficient and an inefficient estimator. We call (8.18) the **Hausmann Equality** as it was first pointed out in econometrics by Hausman (1978).

8.5 Minimum Distance

The previous section explored the finite sample distribution theory under the assumptions of the linear regression model, homoskedastic regression model, and normal regression model. We now return to **the general projection model** where **we do not impose linearity, homoskedasticity, nor normality**.

We are interested in the question: Can we do better than CLS in this setting?

A minimum distance estimator tries to find a parameter value which satisfies the constraint which is as close as possible to the unconstrained estimate. Let beta^ be the unconstrained least-squares estimator, and for some k-by-k positive definite weight matrix W^ > 0 define the **quadratic criterion function**

$$J(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \widehat{\mathbf{W}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

This is a (squared) weighted Euclidean distance between beta^ and beta.

J(beta) is small if beta is close to beta^, and is minimized at zero only beta = beta^. A minimum distance estimator beta~_md for minimizes J (beta) subject to the constraint (8.1), that is,

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}} \; J\left(\boldsymbol{\beta}\right).$$

The CLS estimator is the **special case** when

$$\widehat{\mathbf{W}} = \widehat{\boldsymbol{Q}}_{xx}$$

and we write this criterion function as

$$J^{0}\left(\boldsymbol{\beta}\right) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

To see the equality of CLS and minimum distance, rewrite the least-squares criterion as follows. Substitute the unconstrained least-squares fitted equation y_i = x_ibeta^ + e^_i in SSE(beta) to obtain

$$
\begin{aligned}
SSE(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
&= \sum_{i=1}^{n} \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{e}_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
&= \sum_{i=1}^{n} \widehat{e}_i^2 + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&= n\widehat{\sigma}^2 + J^0\left(\boldsymbol{\beta}\right)
\end{aligned}
$$

where the third equality uses the fact that

$$\sum_{i=1}^{n} \boldsymbol{x}_i\widehat{e}_i = 0$$

and the last line uses

$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = n\widehat{\mathbf{Q}}_{\boldsymbol{xx}}$

The expression (8.21) only depends on through J^0(beta). Thus minimization of SSE(beta) and J^0(beta) are equivalent, and hence

$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \widetilde{\boldsymbol{\beta}}_{\text{cls}}$

when

$\widehat{\mathbf{W}} = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}$

We can solve for beta~_md explicitly by the method of Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} J\left(\boldsymbol{\beta}, \widehat{\boldsymbol{W}}\right) + \boldsymbol{\lambda}'\left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{c}\right)$$

which is minimized over (beta, lambda). The solution is

$$\widetilde{\boldsymbol{\lambda}}_{\mathrm{md}} = n\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)$$

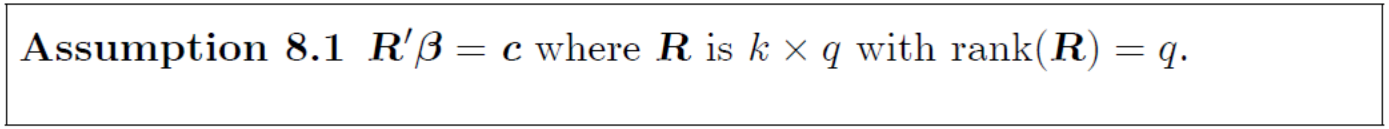
$$\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)$$

(See Exercise 8.10.) Comparing (8.23) with (8.9) we can see that beta~_md specializes to beta~_cls when we set W^ = Q^_xx.

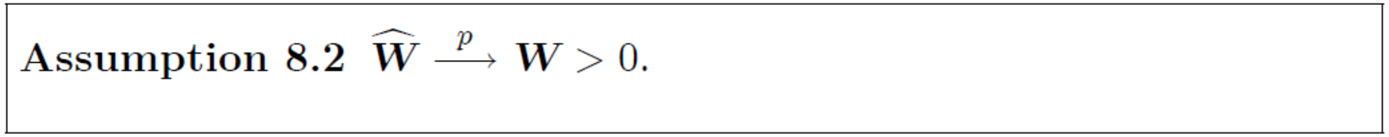
An obvious question is which weight matrix W^ is best. We will address this question after we derive the asymptotic distribution for a general weight matrix.

8.6 Asymptotic Distribution

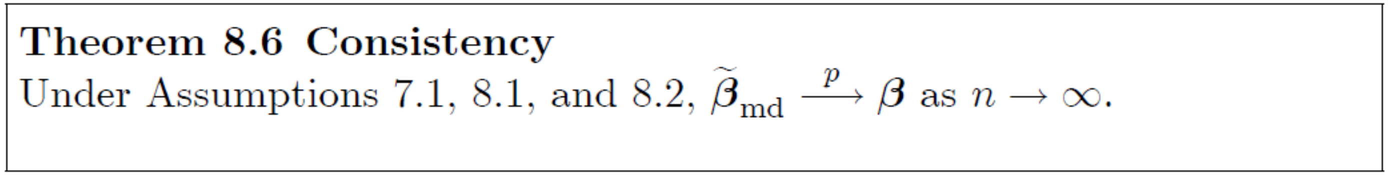
We first show that the class of minimum distance estimators are consistent for the population parameters when the constraints are valid.



Assumption 8.1 $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}$ where $\boldsymbol{R}$ is $k \times q$ with $\operatorname{rank}(\boldsymbol{R}) = q$.



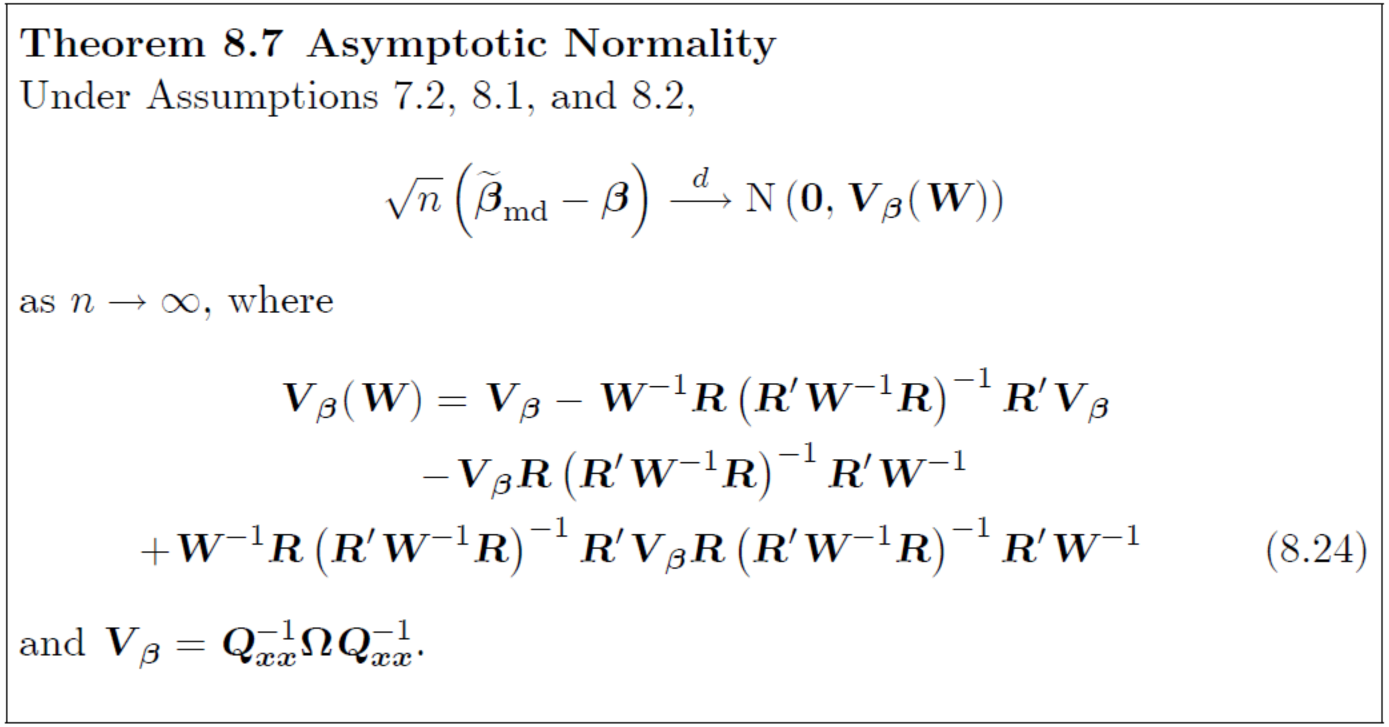
Assumption 8.2 $\widehat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W} > 0$.



Theorem 8.6 Consistency
Under Assumptions 7.1, 8.1, and 8.2, $\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} \xrightarrow{p} \boldsymbol{\beta}$ as $n \to \infty$.

For a proof, see Exercise 8.11. Theorem 8.6 shows that consistency holds for **any weight matrix with a positive definite limit**, so the result includes the CLS estimator.

Similarly, the constrained estimators are asymptotically normally distributed.



Theorem 8.7 Asymptotic Normality

Under Assumptions 7.2, 8.1, and 8.2,

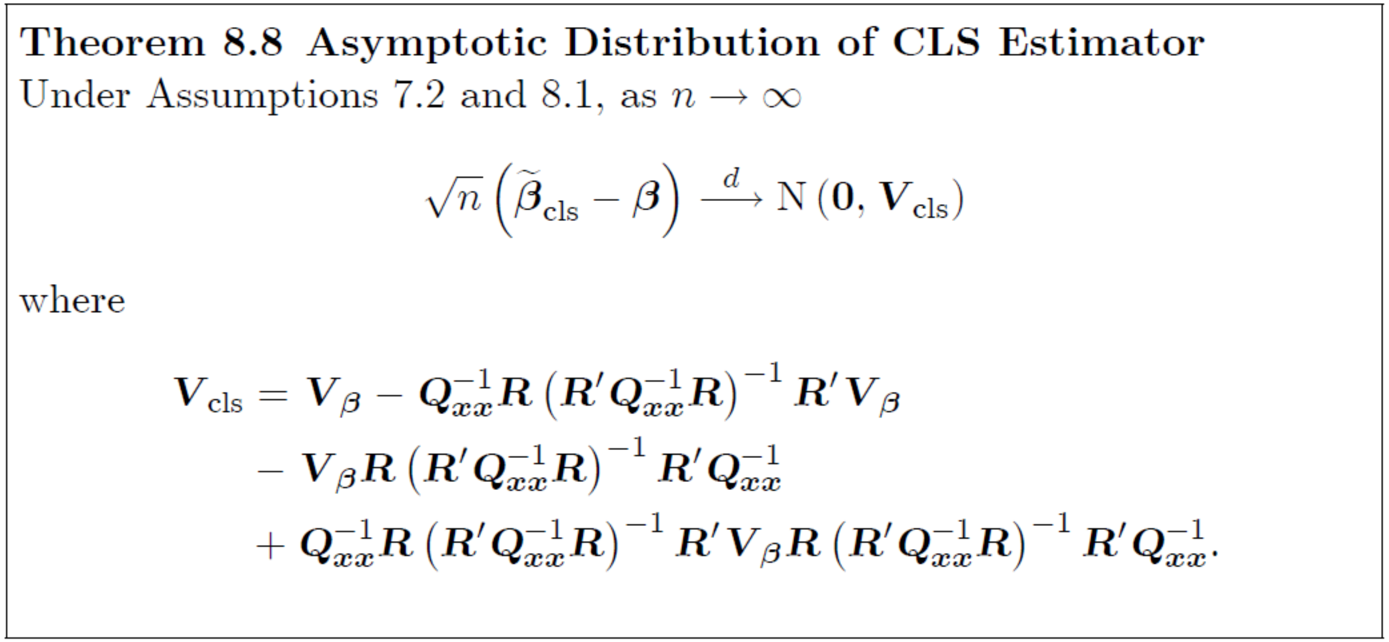
$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}-\boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right)$$

as $n \to \infty$, where

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}} \\ -\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \\ +\boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \end{aligned} \tag{8.24}$$

and $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$.

For a proof, see Exercise 8.12. Theorem 8.7 shows that the minimum distance estimator is asymptotically normal for **all positive definite weight matrices**. The asymptotic variance depends on W. The theorem includes the CLS estimator as a special case by setting W = Q_xx.



Theorem 8.8 Asymptotic Distribution of CLS Estimator
Under Assumptions 7.2 and 8.1, as $n \to \infty$

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\mathrm{cls}}\right)$$

where

$$\begin{aligned}\boldsymbol{V}_{\mathrm{cls}} &= \boldsymbol{V}_{\beta} - \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\beta} \\ &\quad - \boldsymbol{V}_{\beta}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1} \\ &\quad + \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\beta}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}.\end{aligned}$$

For a proof, see Exercise 8.13.

8.7 Variance Estimation and Standard Errors

Earlier we introduced the covariance matrix estimator under the assumption of conditional homoskedasticity. We now introduce an estimator which **does not impose homoskedasticity**.

The asymptotic covariance matrix V_cls may be estimated by replacing V_beta with a consistent estimator such as b V^_beta. A more efficient estimator is obtained by using the restricted coefficient estimator. Given the constrained least-squares squares residuals

$$\widetilde{e}_i = y_i - \boldsymbol{x}_i' \widetilde{\boldsymbol{\beta}}_{\text{cls}}$$

we can estimate the matrix

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right)$$

by

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{n-k+q} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \widetilde{e}_i^2$$

Notice that we have defined Omega~ using an adjusted degrees of freedom. This is an ad hoc adjustment designed to mimic that used for estimation of the error variance sigma^2. Given Omega~ the moment estimator of V_beta is

$$\widetilde{\boldsymbol{V}}_{\beta} = \widehat{\boldsymbol{Q}}_{xx}^{-1} \widetilde{\boldsymbol{\Omega}} \widehat{\boldsymbol{Q}}_{xx}^{-1}$$

and that for V_cls is

$$\begin{aligned} \widetilde{\boldsymbol{V}}_{\text{cls}} &= \widetilde{\boldsymbol{V}}_{\beta} - \widehat{\boldsymbol{Q}}_{xx}^{-1} \boldsymbol{R} \left(\boldsymbol{R}' \widehat{\boldsymbol{Q}}_{xx}^{-1} \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \widetilde{\boldsymbol{V}}_{\beta} \\ &\quad - \widetilde{\boldsymbol{V}}_{\beta} \boldsymbol{R} \left(\boldsymbol{R}' \widehat{\boldsymbol{Q}}_{xx}^{-1} \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \widehat{\boldsymbol{Q}}_{xx}^{-1} \\ &\quad + \widehat{\boldsymbol{Q}}_{xx}^{-1} \boldsymbol{R} \left(\boldsymbol{R}' \widehat{\boldsymbol{Q}}_{xx}^{-1} \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \widetilde{\boldsymbol{V}}_{\beta} \boldsymbol{R} \left(\boldsymbol{R}' \widehat{\boldsymbol{Q}}_{xx}^{-1} \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \widehat{\boldsymbol{Q}}_{xx}^{-1} \end{aligned}$$

We can calculate standard errors for any linear combination h'beta~_cls so long as **h does not lie in the range space of R**. A standard error for h'beta~ is

$$s(\boldsymbol{h}' \widetilde{\boldsymbol{\beta}}_{\text{cls}}) = \left(n^{-1} \boldsymbol{h}' \widetilde{\boldsymbol{V}}_{\text{cls}} \boldsymbol{h} \right)^{1/2}$$

8.8 Efficient Minimum Distance Estimator

Theorem 8.7 shows that minimum distance estimators, which include CLS as a special case, are asymptotically normal with an asymptotic covariance matrix which depends on the weight matrix W.

The asymptotically optimal weight matrix is the one which minimizes the asymptotic variance V_beta(W). This turns out to be

$$\boldsymbol{W} = \boldsymbol{V}_{\beta}^{-1}$$

as is shown in Theorem 8.9 below.

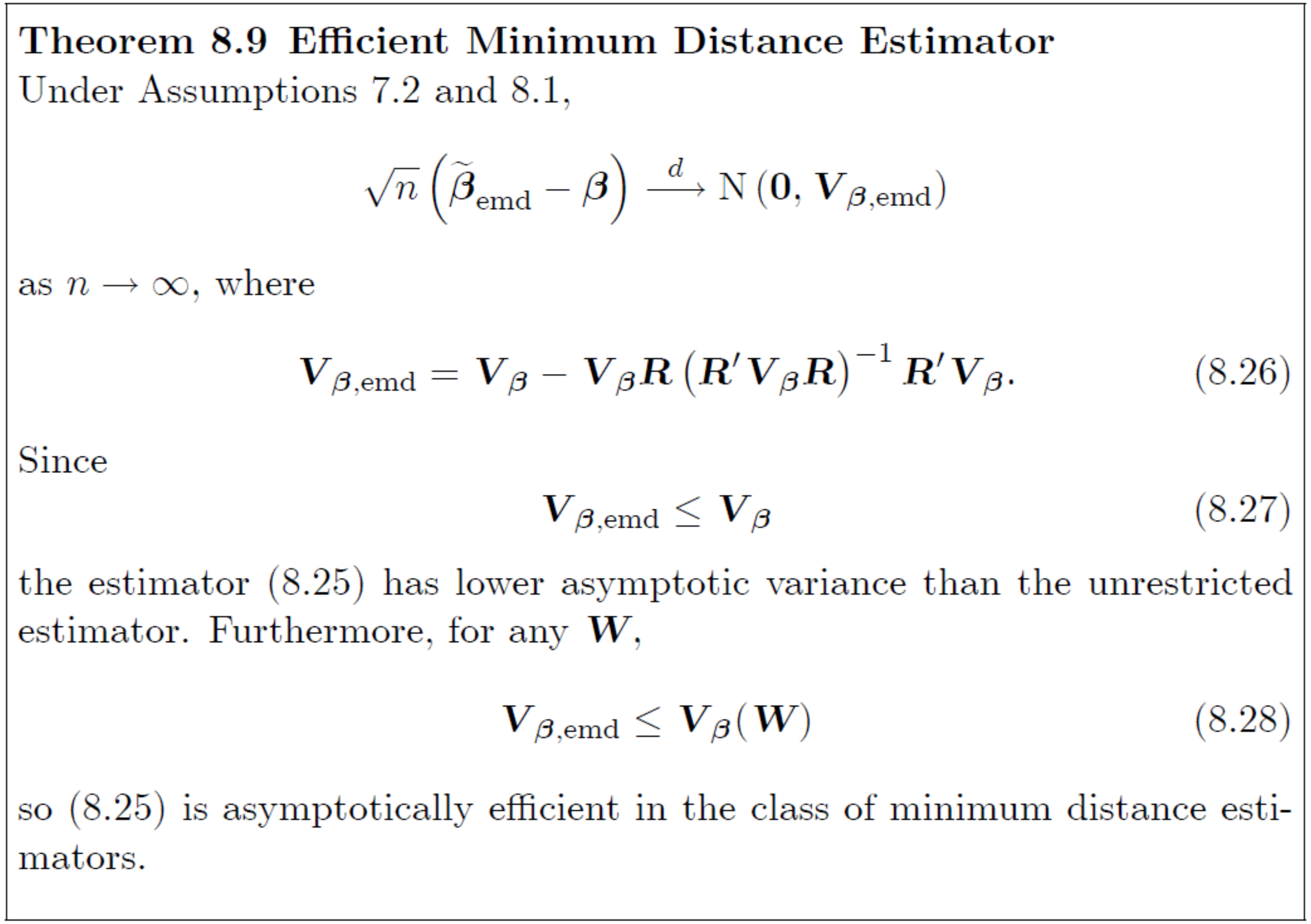
Since (V_beta)^{-1} is unknown this weight matrix cannot be used for a feasible estimator, but we can replace (V_beta)^{-1} with a consistent estimate (V^_beta)^{-1} and the asymptotic distribution (and efficiency) are unchanged. We call the minimum distance estimator setting

$$\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_{\beta}^{-1}$$

the efficient minimum distance estimator and takes the form

$$\widetilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\boldsymbol{\beta}} - \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \boldsymbol{R} \left(\boldsymbol{R}' \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \boldsymbol{R} \right)^{-1} \left(\boldsymbol{R}' \widehat{\boldsymbol{\beta}} - \boldsymbol{c} \right)$$

The asymptotic distribution of (8.25) can be deduced from Theorem 8.7. (See Exercises 8.14 and 8.15, and the proof in Section 8.16.)



Theorem 8.9 Efficient Minimum Distance Estimator
Under Assumptions 7.2 and 8.1,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\text{emd}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta},\text{emd}}\right)$$

as $n \to \infty$, where

$$\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}} = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}. \tag{8.26}$$

Since

$$\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}} \leq \boldsymbol{V}_{\boldsymbol{\beta}} \tag{8.27}$$

the estimator (8.25) has lower asymptotic variance than the unrestricted estimator. Furthermore, for any $\boldsymbol{W}$,

$$\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}} \leq \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) \tag{8.28}$$

so (8.25) is asymptotically efficient in the class of minimum distance estimators.

Theorem 8.9 shows that the minimum distance estimator with the smallest asymptotic variance is (8.25). One implication is that **the constrained least squares estimator is generally inefficient**. **The interesting exception is the case of conditional homoskedasticity**, in which case the optimal weight matrix is W = (V^0_beta)^{-1} so in this case CLS is an efficient minimum distance estimator.

Otherwise when the error is **conditionally heteroskedastic**, there are asymptotic efficiency gains by using minimum distance rather than least squares.

The fact that CLS is generally inefficient is **counter-intuitive** and requires some reflection to understand. Standard intuition suggests to apply the same estimation method (least squares) to the unconstrained and constrained models, and this is the most common empirical practice.

But Theorem 8.9 shows that this is not the efficient estimation method. Instead, the efficient minimum distance estimator has a smaller asymptotic variance. Why?

The reason is that **the least-squares estimator does not make use of the regressor x_2i**. **It ignores the information E(x_2i e_i) = 0**. This information is relevant when **the error is heteroskedastic** and **the excluded regressors are correlated with the included regressors**.

Inequality (8.27) shows that the efficient minimum distance estimator beta~_md has a smaller asymptotic variance than the unrestricted least squares estimator beta^. This means that efficient estimation is attained by imposing correct restrictions when we use the minimum distance method.

8.9 Exclusion Restriction Revisited

We return to the example of estimation with a simple exclusion restriction. The model is

$$y_i = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 + \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2 + e_i$$

with the exclusion restriction beta_2 = 0. We have introduced three estimators of beta_1.

The first is unconstrained least-squares applied to (8.10), which can be written as

$$\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{1y\cdot 2}$$

From Theorem 7.25 and equation (7.14) its asymptotic variance is

$$\operatorname{avar}(\widehat{\boldsymbol{\beta}}_1) = \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\boldsymbol{\Omega}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{21} - \boldsymbol{\Omega}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1}$$

The second estimator of beat_1 is the CLS estimator, which can be written as

$$\widetilde{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{Q}}_{11}^{-1} \widehat{\boldsymbol{Q}}_{1y}$$

Its asymptotic variance can be deduced from Theorem 8.8, but it is simpler to apply the CLT directly to show that

$$\operatorname{avar}(\widetilde{\boldsymbol{\beta}}_1) = \boldsymbol{Q}_{11}^{-1} \boldsymbol{\Omega}_{11} \boldsymbol{Q}_{11}^{-1}$$

The third estimator of beta_1 is the efficient minimum distance estimator. Applying (8.25), it equals

$$\overline{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{V}}_{12}\widehat{\boldsymbol{V}}_{22}^{-1}\widehat{\boldsymbol{\beta}}_2$$

where we have partitioned

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left[\begin{array}{cc} \widehat{\boldsymbol{V}}_{11} & \widehat{\boldsymbol{V}}_{12} \\ \widehat{\boldsymbol{V}}_{21} & \widehat{\boldsymbol{V}}_{22} \end{array} \right]$$

From Theorem 8.9 its asymptotic variance is

$$\operatorname{avar}(\overline{\boldsymbol{\beta}}_1) = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}$$

See Exercise 8.16 to verify equations (8.29), (8.30), and (8.31).

In general, the **three estimators are different**, and **they have different asymptotic variances**. It is instructive to compare the variances to assess whether or not the constrained estimator is necessarily more efficient than the unconstrained estimator.

First, consider the case of **conditional homoskedasticity**. In this case the two covariance matrices simplify to

$$\operatorname{avar}(\widehat{\boldsymbol{\beta}}_1) = \sigma^2 \boldsymbol{Q}_{11\cdot 2}^{-1}$$

$$\operatorname{avar}(\widetilde{\boldsymbol{\beta}}_1) = \sigma^2 \boldsymbol{Q}_{11}^{-1}$$

If **Q_12 = 0** (so **x_1i and x_2i are orthogonal**) then these two variance matrices are equal and the two estimators have equal asymptotic efficiency.

Otherwise, since

$$\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} \geq 0$$

Then

$$\boldsymbol{Q}_{11} \geq \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$$

and consequently

$$\boldsymbol{Q}_{11}^{-1}\sigma^2 \leq \left(\boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)^{-1}\sigma^2$$

This means that under **conditional homoskedasticity**, **beta~_1 has a lower asymptotic variance matrix than beta^_1**.

Therefore in this context, constrained least-squares is more efficient than unconstrained least-squares. This is consistent with our intuition that imposing a correct restriction (excluding an irrelevant regressor) improves estimation efficiency.

However, in the general case of **conditional heteroskedasticity** this ranking is not guaranteed. In fact what is really amazing is that **the variance ranking can be reversed**. The CLS estimator can have a larger asymptotic variance than the unconstrained least squares estimator.

To see this let's use the simple heteroskedastic example from Section 7.4. In that example,

$$Q_{11} = Q_{22} = 1,\ Q_{12} = \frac{1}{2},\ \Omega_{11} = \Omega_{22} = 1, \text{ and } \Omega_{12} = \frac{7}{8}$$

We can calculate (see Exercise 8.17) that

$$Q_{11\cdot 2} = \frac{3}{4}$$

$$\begin{aligned} \operatorname{avar}(\widehat{\boldsymbol{\beta}}_1) &= \frac{2}{3} \\ \operatorname{avar}(\widetilde{\boldsymbol{\beta}}_1) &= 1 \end{aligned}$$

$$\operatorname{avar}(\overline{\boldsymbol{\beta}}_1) = \frac{5}{8}$$

Thus the restricted least-squares estimator beta~_1 has a larger variance than the unrestricted least-squares estimator beta^_1. The minimum distance estimator has the smallest variance of the three, as expected.

What we have found is that when the estimation method is least-squares, deleting the irrelevant variable x_2i can actually increase estimation variance; or equivalently, adding an irrelevant variable can decrease the estimation variance.

To repeat this unexpected finding, we have shown in a very simple example that it is possible for least-squares applied to the short regression (8.11) to be less efficient for estimation of beta_1 than least-squares applied to the long regression (8.10), **even though the constraint beta_2 = 0 is valid**.

This result is **strongly counter-intuitive**. It seems to contradict our initial motivation for pursuing constrained estimation - to improve estimation efficiency.

It turns out that **a more refined answer is appropriate**. **Constrained estimation is desirable**, **but not constrained least-squares estimation**. While **least-squares is asymptotically efficient for estimation of the unconstrained projection model**, **it is not an efficient estimator of the constrained projection model**.

8.10 Variance and Standard Error Estimation

We have discussed covariance matrix estimation for the CLS estimator, but not yet for the EMD estimator.

The asymptotic covariance matrix (8.26) may be estimated by replacing V_beta with a consistent estimate. It is best to construct the variance estimate using beta~_emd. The EMD residuals are

$$\widetilde{e}_i = y_i - \boldsymbol{x}_i' \widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}$$

Using these we can estimate the matrix

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{n-k+q} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \widetilde{e}_i^2$$

Following the formula for CLS we recommend an adjusted degrees of freedom. Given Omega~ the moment estimator of V_beta is

$$\widetilde{\boldsymbol{V}}_{\beta} = \widehat{\boldsymbol{Q}}_{xx}^{-1} \widetilde{\boldsymbol{\Omega}} \widehat{\boldsymbol{Q}}_{xx}^{-1}$$

Given this, we construct the variance estimator

$$\widetilde{\mathbf{V}}_{\boldsymbol{\beta},\mathrm{emd}} = \widetilde{\mathbf{V}}_{\boldsymbol{\beta}} - \widetilde{\mathbf{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widetilde{\mathbf{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\mathbf{V}}_{\boldsymbol{\beta}}$$

A standard error for h'beta~ is then

$$s(\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}) = \left(n^{-1}\boldsymbol{h}'\widetilde{\mathbf{V}}_{\boldsymbol{\beta},\mathrm{emd}}\boldsymbol{h}\right)^{1/2}$$

8.11 Hausman Equality

Form (8.25) we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}-\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)=\widehat{\mathbf{V}}_{\beta}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\mathbf{V}}_{\beta}\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}}-\boldsymbol{c}\right)$$

$$\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\,\mathbf{V}_{\beta}\boldsymbol{R}\left(\boldsymbol{R}'\mathbf{V}_{\beta}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\mathbf{V}_{\beta}\right).$$

It follows that the asymptotic variances of the estimators satisfy the relationship

$$\operatorname{avar}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}-\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)=\operatorname{avar}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right)-\operatorname{avar}\left(\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)$$

We call (8.37) the Hausman Equality: the asymptotic variance of the difference between an efficient and inefficient estimator is the difference in the asymptotic variances.

8.13 Misspecification

What are the consequences for a constrained estimator beta~ if the constraint (8.1) is incorrect? To be specific, suppose that the truth is

$$\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}^*$$

where c* is not necessarily equal to c.

This situation is a generalization of the analysis of "**omitted variable bias**" from Section 2.24, where we found that the short regression (e.g. (8.12)) is estimating a different projection coefficient than the long regression (e.g. (8.10)).

One mechanical answer is that we can use the formula (8.23) for the minimum distance estimator to find that

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} \xrightarrow{p} \boldsymbol{\beta}^*_{\mathrm{md}} = \boldsymbol{\beta} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^* - \boldsymbol{c})$$

The second term

$$\boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^{*}-\boldsymbol{c})$$

shows that imposing an incorrect constraint leads to inconsistency - an asymptotic bias.

We can call the limiting value beta*_md the **minimum-distance projection coefficient** or the **pseudo-true value implied by the restriction**.

However, we can say more.

For example, we can describe some characteristics of the approximating projections. The CLS estimator projection coefficient has the representation

$$\boldsymbol{\beta}^*_{\text{cls}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}} \, \mathbb{E}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

the best linear predictor subject to the constraint (8.1).

The minimum distance estimator converges in probability to

$$\boldsymbol{\beta}^*_{\text{md}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}} \left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)' \mathbf{W} \left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)$$

where beta_0 is the true coefficient. That is, beta*_md is the coefficient vector satisfying (8.1) closest to the true value in the weighted Euclidean norm.

These calculations show that the constrained estimators are **still reasonable** in the sense that **they produce good approximations to the true coefficient, conditional on being required to satisfy the constraint**.

We can also show that beta~_md has an asymptotic normal distribution. The trick is to define the pseudo-true value

$$\boldsymbol{\beta}_n^* = \boldsymbol{\beta} - \widehat{\mathbf{W}}^{-1}\boldsymbol{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^* - \boldsymbol{c})$$

(Note that (8.38) and (8.39) are different!) Then

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta}_n^*\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \widehat{\mathbf{W}}^{-1}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^*\right) \\
&= \left(\boldsymbol{I} - \widehat{\mathbf{W}}^{-1}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&\xrightarrow{d} \left(\boldsymbol{I} - \mathbf{W}^{-1}\mathbf{R}\left(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right) \\
&= \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\mathbf{W})\right).
\end{aligned}$$

In particular

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}} - \boldsymbol{\beta}_n^*\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}^*\right)$$

This means that even when the constraint (8.1) is misspecified, **the conventional covariance matrix estimator (8.35) and standard errors (8.36) are appropriate measures of the sampling variance**, though **the distributions are centered at the pseudo-true values** (projections) beta*_n rather than beta.

The fact that the estimators are biased is an unavoidable consequence of misspecification.

An alternative approach to the asymptotic distribution theory under misspecification uses the concept of **local alternatives**. It is a technical device which might seem a bit artificial, but it is a powerful method to derive useful distributional approximations in a wide variety of contexts.

The idea is to index the true coefficient beta_n by n via the relationship

$$\boldsymbol{R}'\boldsymbol{\beta}_n = \boldsymbol{c} + \boldsymbol{\delta} n^{-1/2}$$

Equation (8.41) specifies that beta_n violates (8.1) and thus the constraint is misspecified. However, the constraint is “close” to correct, as the difference

$$\boldsymbol{R}'\boldsymbol{\beta}_n - \boldsymbol{c} = \boldsymbol{\delta} n^{-1/2}$$

is “small” in the sense that it decreases with the sample size n. We call (8.41) **local misspecification**.

The asymptotic theory is then derived as n → infintiy under the sequence of probability distributions with the coefficient beta_n. The way to think about this is that the true value of the parameter is beta_n, and it is "close" to satisfying (8.1).

The reason why the deviation is proportional to n^{-1/2} is because **this is the only choice under which the localizing parameter appears in the asymptotic distribution but does not dominate it**. The best way to see this is to work through the asymptotic approximation.

Since beta_n is the true coefficient value, then y_i = x'_i beta_n + e_i and we have the standard representation for the **unconstrained** estimator, namely

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_n\right)=\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i e_i\right)$$
$$\xrightarrow{d} \mathrm{N}\left(\mathbf{0},\, \boldsymbol{V}_{\boldsymbol{\beta}}\right).$$

There is **no difference under fixed (classical) or local asymptotics**, since the right-hand-side is independent of the coefficient beta_n.

A difference arises for the constrained estimator. Using (8.41),

$$\boldsymbol{c} = \mathbf{R}'\boldsymbol{\beta}_n - \boldsymbol{\delta} n^{-1/2}$$

$$\mathbf{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c} = \mathbf{R}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) + \boldsymbol{\delta} n^{-1/2}$$

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} &= \widehat{\boldsymbol{\beta}} - \widehat{\mathbf{W}}^{-1}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R}\right)^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) \\ &= \widehat{\boldsymbol{\beta}} - \widehat{\mathbf{W}}^{-1}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) + \widehat{\mathbf{W}}^{-1}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R}\right)^{-1}\boldsymbol{\delta} n^{-1/2}\end{aligned}$$

It follows that

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta}_n\right) &= \left(\boldsymbol{I} - \widehat{\mathbf{W}}^{-1}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) \\ &\quad + \widehat{\mathbf{W}}^{-1}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R}\right)^{-1}\boldsymbol{\delta}.\end{aligned}$$

The first term is asymptotically normal (from 8.42)). The second term converges in probability to a constant. This is because the n^{-1/2} local scaling in (8.41) is exactly balanced by the n^{-1/2} scaling of the estimator. No alternative rate would have produced this result.

Consequently, we find that the asymptotic distribution equals

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}-\boldsymbol{\beta}_n\right) &\xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)+\boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta} \\ &= \mathrm{N}\left(\boldsymbol{\delta}^*, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right)\end{aligned}$$

$$\boldsymbol{\delta}^* = \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta}$$

The asymptotic distribution (8.43) is an approximation of the sampling distribution of the restricted estimator under misspecification. The distribution (8.43) contains an asymptotic bias component delta*.

The approximation is not fundamentally different from (8.40) - **they both have the same asymptotic variances**, and both reflect the bias due to misspecification. The difference is that (8.40) puts the bias on the left-side of the convergence arrow, while (8.43) has the bias on the right-side. There is no substantive difference between the two, **but (8.43) is more convenient for some purposes**, such as **the analysis of the power of tests**, as we will explore in the next chapter.

8.14 Nonlinear Constraints

In some cases it is desirable to impose nonlinear constraints on the parameter vector. They can be written as

$$\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}$$

where r: R^r → R^q. This includes the linear constraints (8.1) as a special case. An example of (8.44) which cannot be written as (8.1) is

$$\beta_1 \beta_2 = 1$$

which is (8.44) with

$$r(\boldsymbol{\beta}) = \beta_1 \beta_2 - 1$$

The constrained least-squares and minimum distance estimators of beta subject to (8.44) solve the minimization problems

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmin}} \, SSE(\boldsymbol{\beta})$$

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmin}} \; J\,(\boldsymbol{\beta})$$

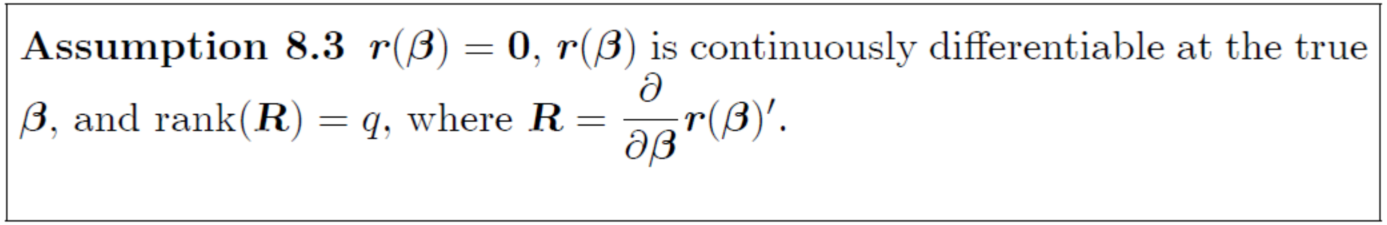
where SSE(beta) and J (beta) are defined in (8.4) and (8.19), respectively. The solutions minimize the Lagrangians

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{1}{2}SSE(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\boldsymbol{r}(\boldsymbol{\beta})$$

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{1}{2}J\,(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\boldsymbol{r}(\boldsymbol{\beta})$$

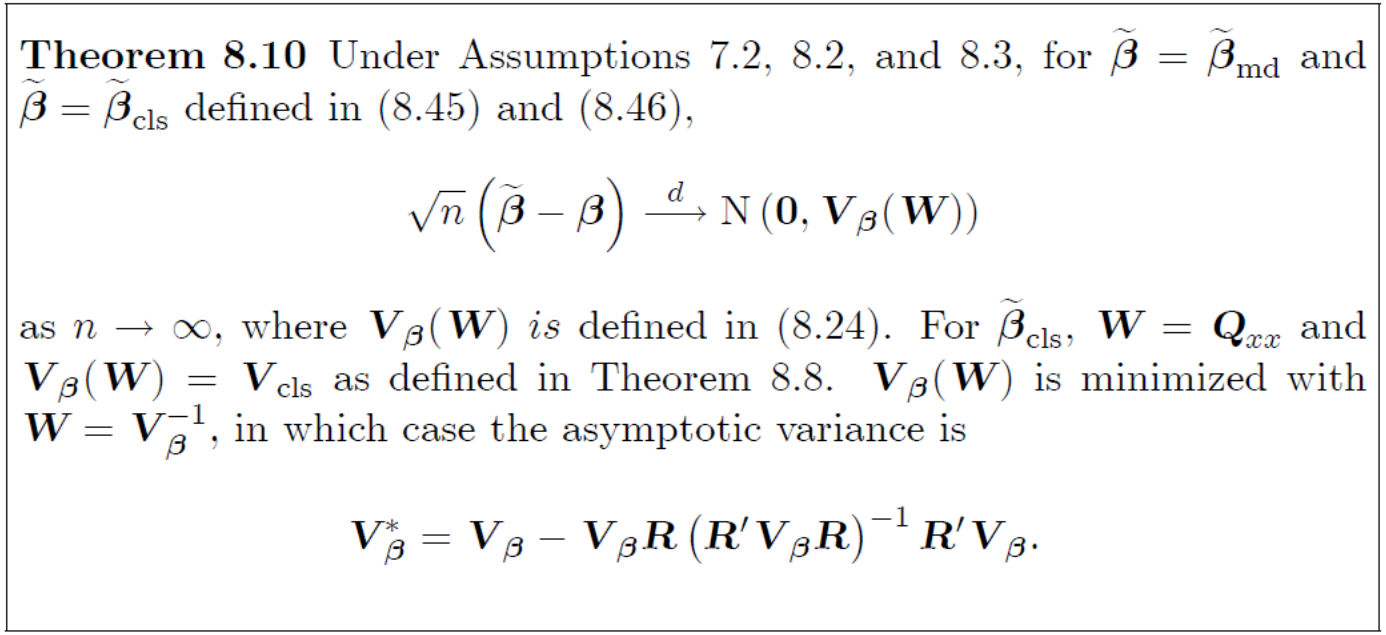
over (beta, lambda).

Computationally, there is **no general closed-form solution** for the estimator so they must be found numerically. Algorithms to numerically solve (8.45) and (8.46) are known as constrained optimization methods, and are available in programming languages including MATLAB, GAUSS and R.



Assumption 8.3 $\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}$, $\boldsymbol{r}(\boldsymbol{\beta})$ is continuously differentiable at the true $\boldsymbol{\beta}$, and $\operatorname{rank}(\boldsymbol{R}) = q$, where $\boldsymbol{R} = \dfrac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\boldsymbol{\beta})'$.

The asymptotic distribution is a simple generalization of the case of a linear constraint, but the proof is more delicate.



Theorem 8.10 Under Assumptions 7.2, 8.2, and 8.3, for $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\text{md}}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\text{cls}}$ defined in (8.45) and (8.46),

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right)$$

as $n \to \infty$, where $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ *is* defined in (8.24). For $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$, $\boldsymbol{W} = \boldsymbol{Q}_{xx}$ and $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) = \boldsymbol{V}_{\text{cls}}$ as defined in Theorem 8.8. $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ is minimized with $\boldsymbol{W} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$, in which case the asymptotic variance is

$$\boldsymbol{V}_{\boldsymbol{\beta}}^{*} = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}.$$

The asymptotic variance matrix for the efficient minimum distance estimator can be estimated by

$$\widehat{\boldsymbol{V}}_{\beta}^{*} = \widehat{\boldsymbol{V}}_{\beta} - \widehat{\boldsymbol{V}}_{\beta}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\beta}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\beta}$$

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{r}(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}})'$$

Standard errors for the elements of beta~_md are the square roots of the diagonal elements of V^*_{beta~} = n^{-1}V^*_{beta}.

8.15 Inequality Restrictions

Inequality constraints on the parameter vector beta take the form

$$\boldsymbol{r}(\boldsymbol{\beta}) \geq \mathbf{0}$$

for some function r: R^k → R^q. The most common example is a non-negative constraint

$$\beta_1 \geq 0$$

The constrained least-squares and minimum distance estimators can be written as

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{r}(\boldsymbol{\beta}) \geq \mathbf{0}}{\operatorname{argmin}} \, SSE(\boldsymbol{\beta})$$

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \underset{\boldsymbol{r}(\boldsymbol{\beta}) \geq \mathbf{0}}{\operatorname{argmin}} \; J\left(\boldsymbol{\beta}\right)$$

Except in special cases the constrained estimators do not have simple algebraic solutions. An important exception is when there is a single non-negativity constraint, e.g. beta_1 >= 0 with q = 1.

In this case the constrained estimator can be found by **two-step** approach. First compute the unconstrained estimator beta^. If beta^_>= 0, then beta~ = beta^. Second, if beta^_1<0, then impose beta_1 = 0 (eliminate the regressor X_1) and re-estimate.

This yields the constrained least-squares estimator. While this method works when there is **a single non-negativity constraint**, it does not immediately generalize to other contexts.

The computational problems (8.50) and (8.51) are examples of **quadratic programming problems**. Quick and easy computer algorithms are available in programming languages including MATLAB, GAUSS and R.

Inference on inequality-constrained estimators is **unfortunately quite challenging**.

The conventional asymptotic theory gives rise to the following **dichotomy**.

If the true parameter satisfies the strict inequality r(beta) > 0, then asymptotically the estimator is not subject to the constraint and the inequality-constrained estimator has an asymptotic distribution equal to the unconstrained case.

However, if the true parameter is on the boundary, e.g. r(beta) = 0, then the estimator has a truncated structure.

This is easiest to see in the **one-dimensional** case. If we have an estimator beta^ which satisfies

$$\sqrt{n}\left(\widehat{\beta}-\beta\right) \xrightarrow{d} \mathrm{Z} = \mathrm{N}\left(0, V_{\beta}\right)$$

$$\beta = 0$$

then the constrained estimator

$$\widetilde{\beta} = \max[\widehat{\beta}, 0]$$

will have the asymptotic distribution

$$\sqrt{n}\widetilde{\beta} \xrightarrow{d} \max[\mathrm{Z}, 0]$$

a “half-normal” distribution.