Chapter 8

Restricted Estimation

8.1 Introduction

In the linear projection model

$$\begin{aligned} Y &= X'\beta + e \\ \mathbb{E}[Xe] &= 0 \end{aligned}$$

a common task is to impose a constraint on the coefficient vector β. For example, partitioning $X' = (X_1', X_2')$ and $\beta' = (\beta_1', \beta_2')$ a typical constraint is an exclusion restriction of the form $\beta_2 = 0$. In this case the constrained model is

$$\begin{aligned} Y &= X_1'\beta_1 + e \\ \mathbb{E}[Xe] &= 0. \end{aligned}$$

At first glance this appears the same as the linear projection model but there is one important difference: the error e is uncorrelated with the entire regressor vector $X' = (X_1', X_2')$ not just the included regressor X_1.

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

$$\boldsymbol{R}'\beta = \boldsymbol{c} \tag{8.1}$$

where $\boldsymbol{R}$ is $k \times q$, $\operatorname{rank}(\boldsymbol{R}) = q < k$, and $\boldsymbol{c}$ is $q \times 1$. The assumption that $\boldsymbol{R}$ is full rank means that the constraints are linearly independent (there are no redundant or contradictory constraints). We define the restricted parameter space B as the set of values of β which satisfy (8.1), that is

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

Sometimes we will call (8.1) a **constraint** and sometimes a **restriction**. They are the same thing. Similarly sometimes we will call estimators which satisfy (8.1) **constrained estimators** and sometimes **restricted estimators**. They mean the same thing.

The constraint $\beta_2 = 0$ discussed above is a special case of the constraint (8.1) with

$$\boldsymbol{R} = \begin{pmatrix} 0 \\ \boldsymbol{I}_{k_2} \end{pmatrix}, \tag{8.2}$$

a selector matrix, and $\boldsymbol{c} = 0$.

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 β imposing the linear restriction (8.1)? 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 $\boldsymbol{R}'\beta = \boldsymbol{c}$.

The constrained least squares estimator is

$$\widetilde{\beta}_{\text{cls}} = \underset{\boldsymbol{R}'\beta=\boldsymbol{c}}{\operatorname{argmin}} \operatorname{SSE}(\beta) \tag{8.3}$$

where

$$\operatorname{SSE}(\beta) = \sum_{i=1}^{n} \left(Y_i - X_i'\beta\right)^2 = \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{Y}'\boldsymbol{X}\beta + \beta'\boldsymbol{X}'\boldsymbol{X}\beta. \tag{8.4}$$

The estimator $\widetilde{\beta}_{\text{cls}}$ minimizes the sum of squared errors over all $\beta \in B$, or equivalently such that the restriction (8.1) holds. We call $\widetilde{\beta}_{\text{cls}}$ the **constrained least squares** (CLS) estimator. We use the convention of using a tilde "~" rather than a hat "^" to indicate that $\widetilde{\beta}_{\text{cls}}$ is a restricted estimator in contrast to the unrestricted least squares estimator $\widehat{\beta}$ and write it as $\widetilde{\beta}_{\text{cls}}$ to be clear that the estimation method is CLS.

One method to find the solution to (8.3) is the technique of Lagrange multipliers. The problem (8.3) is equivalent to finding the critical points of the Lagrangian

$$\mathscr{L}(\beta,\lambda) = \frac{1}{2}\operatorname{SSE}(\beta) + \lambda'\left(\boldsymbol{R}'\beta - \boldsymbol{c}\right) \tag{8.5}$$

over (β, λ) where λ is an $s \times 1$ vector of Lagrange multipliers. The solution is a saddlepoint. The Lagrangian is minimized over β while maximized over λ. The first-order conditions for the solution of (8.5) are

$$\frac{\partial}{\partial\beta}\mathscr{L}(\widetilde{\beta}_{\text{cls}}, \widetilde{\lambda}_{\text{cls}}) = -\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{X}'\boldsymbol{X}\widetilde{\beta}_{\text{cls}} + \boldsymbol{R}\widetilde{\lambda}_{\text{cls}} = 0 \tag{8.6}$$

and

$$\frac{\partial}{\partial\lambda}\mathscr{L}(\widetilde{\beta}_{\text{cls}}, \widetilde{\lambda}_{\text{cls}}) = \boldsymbol{R}'\widetilde{\beta} - \boldsymbol{c} = 0. \tag{8.7}$$

Premultiplying (8.6) by $\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ we obtain

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

where $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is the unrestricted least squares estimator. Imposing $\boldsymbol{R}'\widetilde{\beta}_{\text{cls}} - \boldsymbol{c} = 0$ from (8.7) and solving for $\widetilde{\lambda}_{\text{cls}}$ we find

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

Notice that $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} > 0$ and $\boldsymbol{R}$ full rank imply that $\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R} > 0$ and is hence invertible. (See Section A.10.)

Substituting this expression into (8.6) and solving for $\widetilde{\beta}_{\mathrm{cls}}$ we find the solution to the constrained minimization problem (8.3)

$$\widetilde{\beta}_{\mathrm{cls}} = \widehat{\beta}_{\mathrm{ols}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\beta}_{\mathrm{ols}} - \boldsymbol{c}\right). \tag{8.8}$$

(See Exercise 8.5 to verify that (8.8) satisfies (8.1).)

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

$$\widetilde{\beta}_{\mathrm{cls}} = \widehat{\beta}_{\mathrm{ols}} - \widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta}_{\mathrm{ols}} - \boldsymbol{c}\right). \tag{8.9}$$

The CLS residuals are $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}_{\mathrm{cls}}$ and are written in vector notation as $\widetilde{\boldsymbol{e}}$.

To illustrate we generated a random sample of 100 observations for the variables (Y, X_1, X_2) and calculated the sum of squared errors function for the regression of Y on X_1 and X_2. Figure 8.1 displays contour plots of the sum of squared errors function. The center of the contour plots is the least squares minimizer $\widehat{\beta}_{\mathrm{ols}} = (0.33, 0.26)'$. 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. This 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 is marked as $\widetilde{\beta}_{\mathrm{cls}} = (0.52, 0.48)'$.

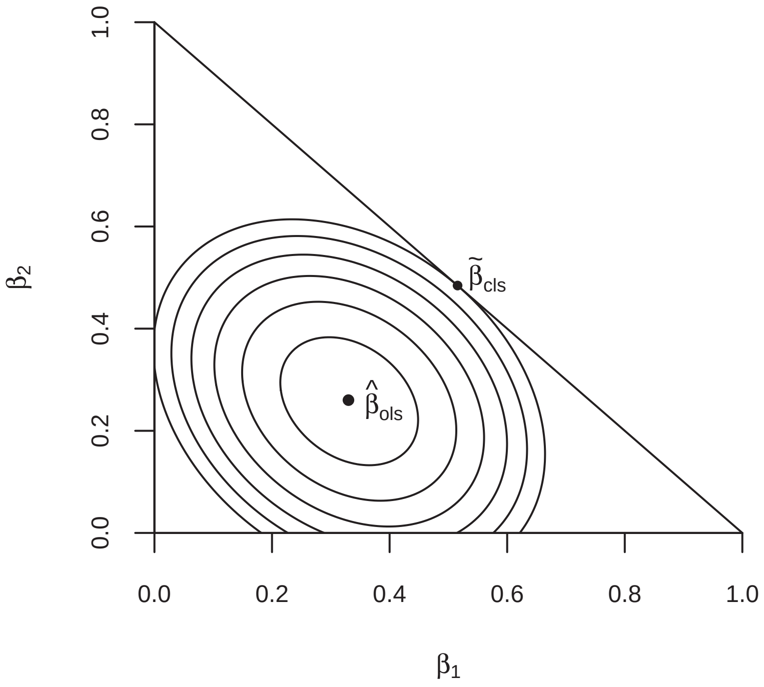

Figure 8.1: Constrained Least Squares Criterion

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

8.3 Exclusion Restriction

While (8.8) is a general formula for CLS, 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 that the unconstrained model is

$$Y = X_1'\beta_1 + X_2'\beta_2 + e, \tag{8.10}$$

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

$$Y = X_1'\beta_1 + e. \tag{8.11}$$

In this setting the CLS estimator is OLS of Y on X_1. (See Exercise 8.1.) We can write this as

$$\widetilde{\beta}_1 = \left(\sum_{i=1}^n X_{1i}X_{1i}'\right)^{-1}\left(\sum_{i=1}^n X_{1i}Y_i\right). \tag{8.12}$$

The CLS estimator of the entire vector $\beta' = \left(\beta_1', \beta_2'\right)$ is

$$\widetilde{\beta} = \begin{pmatrix} \widetilde{\beta}_1 \\ 0 \end{pmatrix}. \tag{8.13}$$

It is not immediately obvious but (8.8) and (8.13) are algebraically identical. To see this the first component of (8.8) with (8.2) is

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

Using (3.39) this equals

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

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

$$Y = X'\beta + e \tag{8.14}$$

$$\mathbb{E}\left[e \mid X\right] = 0. \tag{8.15}$$

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.

Theorem 8.1 The CLS estimator satisfies

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

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

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 β.

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

For a proof see Exercise 8.7.

We can also calculate the covariance matrix of $\widetilde{\beta}_{\text{cls}}$. First, for simplicity take the case of conditional homoskedasticity.

Theorem 8.3 In the homoskedastic linear regression model (8.14)-(8.15) with $\mathbb{E}\left[e^2\mid X\right]=\sigma^2$, under (8.1),

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

For a proof see Exercise 8.8.

We use the $\boldsymbol{V}_{\widetilde{\beta}}^{0}$ notation to emphasize that this is the covariance matrix under the assumption of conditional homoskedasticity.

For inference we need an estimate of $\boldsymbol{V}_{\widetilde{\beta}}^{0}$. A natural estimator is

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

where

$$s_{\text{cls}}^{2}=\frac{1}{n-k+q}\sum_{i=1}^{n}\widetilde{e}_i^{2}\qquad(8.16)$$

is a biased-corrected estimator of σ^2. Standard errors for the components of β are then found by taking the squares roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widetilde{\beta}}$, for example

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

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

$$\begin{aligned}\left(n-k+q\right)s^2_{\text{cls}} &= \widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}} \\ &= \boldsymbol{e}'\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e} \\ &= \boldsymbol{e}'\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}. \end{aligned} \tag{8.17}$$

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 $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$, under (8.1), $\mathbb{E}\left[s^2_{\text{cls}} \mid \boldsymbol{X}\right] = \sigma^2$ and $\mathbb{E}\left[\widehat{\boldsymbol{V}}^0_{\widetilde{\beta}} \mid \boldsymbol{X}\right] = \boldsymbol{V}^0_{\widetilde{\beta}}$.

Now consider the distributional properties in the normal regression model $Y = X'\beta + e$ with $e \sim \mathrm{N}(0,\sigma^2)$. By the linearity of Theorem 8.1.2, conditional on $\boldsymbol{X}$, $\widetilde{\beta}_{\text{cls}} - \beta$ is normal. Given Theorems 8.2 and 8.3 we deduce that $\widetilde{\beta}_{\text{cls}} \sim \mathrm{N}\left(\beta, \boldsymbol{V}^0_{\widetilde{\beta}}\right)$.

Similarly, from Exericise 8.1 we know $\widetilde{\boldsymbol{e}} = \left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$ is linear in $\boldsymbol{e}$ so is also conditionally normal. Furthermore, since $\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \boldsymbol{X}\boldsymbol{A}\right) = 0$, $\widetilde{\boldsymbol{e}}$ and $\widetilde{\beta}_{\text{cls}}$ are uncorrelated and thus independent. Thus s^2_{cls} and $\widetilde{\beta}_{\text{cls}}$ are independent.

From (8.17) and the fact that $\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'$ 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

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

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

The relevance of this calculation is that the “degrees of freedom” for CLS regression equal $n-k+q$ rather than $n-k$ as in OLS. 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 (8.14)-(8.15) with constraint (8.1),

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

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

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

An interesting relationship is that in the homoskedastic regression model

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

This means that $\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{cls}}$ and $\widetilde{\beta}_{\text{cls}}$ are conditionally uncorrelated and hence independent. A corollary is

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

A second corollary is

$$\begin{aligned}
\operatorname{var}\left[\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] &= \operatorname{var}\left[\widehat{\beta}_{\text{ols}} \mid \boldsymbol{X}\right] - \operatorname{var}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.
\end{aligned} \tag{8.18}$$

This also shows that the difference between the CLS and OLS variances matrices equals

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

the final equality meaning positive semi-definite. It follows that $\operatorname{var}\left[\widehat{\beta}_{\text{ols}} \mid \boldsymbol{X}\right] \geq \operatorname{var}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right]$ 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 covariance matrix (in the linear homoskedastic 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 **Hausman 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 satisfying the constraint which is as close as possible to the unconstrained estimator. Let $\widehat{\beta}$ be the unconstrained least squares estimator, and for some $k \times k$ positive definite weight matrix $\widehat{\boldsymbol{W}}$ define the quadratic criterion function

$$J(\beta) = n\left(\widehat{\beta} - \beta\right)'\widehat{\boldsymbol{W}}\left(\widehat{\beta} - \beta\right). \tag{8.19}$$

This is a (squared) weighted Euclidean distance between $\widehat{\beta}$ and β. $J(\beta)$ is small if β is close to $\widehat{\beta}$, and is minimized at zero only if $\beta=\widehat{\beta}$. A **minimum distance estimator** $\widetilde{\beta}_{\text{md}}$ for β minimizes $J(\beta)$ subject to the constraint (8.1), that is,

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

The CLS estimator is the special case when $\widehat{\boldsymbol{W}}=\widehat{\boldsymbol{Q}}_{XX}$ and we write this criterion function as

$$J^0(\beta)=n\left(\widehat{\beta}-\beta\right)'\widehat{\boldsymbol{Q}}_{XX}\left(\widehat{\beta}-\beta\right). \tag{8.20}$$

To see the equality of CLS and minimum distance rewrite the least squares criterion as follows. Substitute the unconstrained least squares fitted equation $Y_i=X_i'\widehat{\beta}+\widehat{e}_i$ into SSE(β) to obtain

$$\begin{aligned}
\text{SSE}(\beta)&=\sum_{i=1}^{n}\left(Y_i-X_i'\beta\right)^2\\
&=\sum_{i=1}^{n}\left(X_i'\widehat{\beta}+\widehat{e}_i-X_i'\beta\right)^2\\
&=\sum_{i=1}^{n}\widehat{e}_i^2+\left(\widehat{\beta}-\beta\right)'\left(\sum_{i=1}^{n}X_iX_i'\right)\left(\widehat{\beta}-\beta\right)\\
&=n\widehat{\sigma}^2+J^0(\beta)
\end{aligned} \tag{8.21}$$

where the third equality uses the fact that $\sum_{i=1}^{n}X_i\widehat{e}_i=0$, and the last line uses $\sum_{i=1}^{n}X_iX_i'=n\widehat{\boldsymbol{Q}}_{XX}$. The expression (8.21) only depends on β through $J^0(\beta)$. Thus minimization of SSE(β) and $J^0(\beta)$ are equivalent, and hence $\widetilde{\beta}_{\text{md}}=\widetilde{\beta}_{\text{cls}}$ when $\widehat{\boldsymbol{W}}=\widehat{\boldsymbol{Q}}_{XX}$.

We can solve for $\widetilde{\beta}_{\text{md}}$ explicitly by the method of Lagrange multipliers. The Lagrangian is

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

The solution to the pair of first order conditions is

$$\widetilde{\lambda}_{\text{md}}=n\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{c}\right) \tag{8.22}$$

$$\widetilde{\beta}_{\text{md}}=\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta}-\boldsymbol{c}\right). \tag{8.23}$$

(See Exercise 8.10.) Comparing (8.23) with (8.9) we can see that $\widetilde{\beta}_{\text{md}}$ specializes to $\widetilde{\beta}_{\text{cls}}$ when we set $\widehat{\boldsymbol{W}}=\widehat{\boldsymbol{Q}}_{XX}$.

An obvious question is which weight matrix $\widehat{\boldsymbol{W}}$ 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 $\boldsymbol{R}'\beta=\boldsymbol{c}$ where $\boldsymbol{R}$ is $k\times q$ with 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, $\widetilde{\beta}_{\mathrm{md}} \xrightarrow[p]{} \beta$ 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 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,

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

as $n \to \infty$, where

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

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

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 $\boldsymbol{W}$. The theorem includes the CLS estimator as a special case by setting $\boldsymbol{W} = \boldsymbol{Q}_{XX}$.

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

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

where

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

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 $\boldsymbol{V}_{\text{cls}}$ may be estimated by replacing $\boldsymbol{V}_\beta$ with a consistent estimator such as $\widehat{\boldsymbol{V}}_\beta$. A more efficient estimator can be obtained by using the restricted coefficient estimator which we now show. Given the constrained least squares squares residuals $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}_{\text{cls}}$ we can estimate the matrix $\Omega = \mathbb{E}\left[XX'e^2\right]$ by

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

Notice that we have used an adjusted degrees of freedom. This is an *ad hoc* adjustment designed to mimic that used for estimation of the error variance σ^2. The moment estimator of $\boldsymbol{V}_\beta$ is

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

and that for $\boldsymbol{V}_{\text{cls}}$ is

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

We can calculate standard errors for any linear combination $h'\widetilde{\beta}_{\text{cls}}$ such that h does not lie in the range space of $\boldsymbol{R}$. A standard error for $h'\widetilde{\beta}$ is

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

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 $\boldsymbol{W}$. The asymptotically optimal weight matrix is the one which minimizes the asymptotic variance $\boldsymbol{V}_\beta(\boldsymbol{W})$. This turns out to be $\boldsymbol{W} = \boldsymbol{V}_\beta^{-1}$ as is shown in Theorem 8.9 below. Since $\boldsymbol{V}_\beta^{-1}$ is unknown this weight matrix cannot be used for a feasible estimator but we can replace $\boldsymbol{V}_\beta^{-1}$ with a consistent estimator $\widehat{\boldsymbol{V}}_\beta^{-1}$ and the asymptotic distribution (and efficiency) are unchanged. We call the minimum distance estimator with $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_\beta^{-1}$ the **efficient minimum distance estimator** and takes the form

$$\widetilde{\beta}_{\text{emd}} = \widehat{\beta} - \widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}\right). \tag{8.25}$$

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,

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

as $n \rightarrow \infty$, where

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

Since

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

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

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

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 $\boldsymbol{W}=\left(\boldsymbol{V}_{\beta}^{0}\right)^{-1}$ 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. Standard intuition suggests to apply the same estimation method (least squares) to the unconstrained and constrained models and this is the common empirical practice. But Theorem 8.9 shows that this is inefficient. Why? The reason is that the least squares estimator does not make use of the regressor X_2. It ignores the information $\mathbb{E}\left[X_2 e\right]=0$. 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 $\widetilde{\beta}_{\mathrm{emd}}$ has a smaller asymptotic variance than the unrestricted least squares estimator $\widehat{\beta}$. 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

$$Y=X_1^{\prime} \beta_1+X_2^{\prime} \beta_2+e$$

with the exclusion restriction $\beta_2=0$. We have introduced three estimators of β_1. The first is unconstrained least squares applied to (8.10) which can be written as $\widehat{\beta}_1=\widehat{\boldsymbol{Q}}_{11 \cdot 2}^{-1} \widehat{\boldsymbol{Q}}_{1 Y \cdot 2}$. From Theorem 7.25 and equation (7.14) its asymptotic variance is

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

The second estimator of β_1 is CLS, which can be written as $\widetilde{\beta}_1 = \widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1Y}$. Its asymptotic variance can be deduced from Theorem 8.8, but it is simpler to apply the CLT directly to show that

$$\operatorname{avar}\left[\widetilde{\beta}_1\right] = \boldsymbol{Q}_{11}^{-1}\Omega_{11}\boldsymbol{Q}_{11}^{-1}. \tag{8.29}$$

The third estimator of β_1 is efficient minimum distance. Applying (8.25), it equals

$$\overline{\beta}_1 = \widehat{\beta}_1 - \widehat{\boldsymbol{V}}_{12}\widehat{\boldsymbol{V}}_{22}^{-1}\widehat{\beta}_2 \tag{8.30}$$

where we have partitioned

$$\widehat{\boldsymbol{V}}_{\beta} = \begin{bmatrix} \widehat{\boldsymbol{V}}_{11} & \widehat{\boldsymbol{V}}_{12} \\ \widehat{\boldsymbol{V}}_{21} & \widehat{\boldsymbol{V}}_{22} \end{bmatrix}.$$

From Theorem 8.9 its asymptotic variance is

$$\operatorname{avar}\left[\overline{\beta}_1\right] = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}. \tag{8.31}$$

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 more efficient than the unconstrained estimator.

First, assume conditional homoskedasticity. In this case the two covariance matrices simplify to $\operatorname{avar}\left[\widehat{\beta}_1\right] = \sigma^2\boldsymbol{Q}_{11\cdot 2}^{-1}$ and $\operatorname{avar}\left[\widetilde{\beta}_1\right] = \sigma^2\boldsymbol{Q}_{11}^{-1}$. If $\boldsymbol{Q}_{12} = 0$ (so X_1 and X_2 are uncorrelated) 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 $\widetilde{\beta}_1$ has a lower asymptotic covariance matrix than $\widehat{\beta}_1$. 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, $Q_{11} = Q_{22} = 1$, $Q_{12} = \frac{1}{2}$, $\Omega_{11} = \Omega_{22} = 1$, and $\Omega_{12} = \frac{7}{8}$. We can calculate (see Exercise 8.17) that $Q_{11\cdot 2} = \frac{3}{4}$ and

$$\operatorname{avar}\left[\widehat{\beta}_1\right] = \frac{2}{3} \tag{8.32}$$

$$\operatorname{avar}\left[\widetilde{\beta}_1\right] = 1 \tag{8.33}$$

$$\operatorname{avar}\left[\overline{\beta}_1\right] = \frac{5}{8}. \tag{8.34}$$

Thus the CLS estimator $\widetilde{\beta}_1$ has a larger variance than the unrestricted least squares estimator $\widehat{\beta}_1$! 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_2 can actually increase estimation variance, or equivalently, adding an irrelevant variable can decrease the estimation variance.

To repeat this unexpected finding, we have shown that it is possible for least squares applied to the short regression (8.11) to be less efficient for estimation of β_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 necessarily CLS. 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 CLS but not yet for the EMD estimator.

The asymptotic covariance matrix (8.26) may be estimated by replacing $\boldsymbol{V}_\beta$ with a consistent estimator. It is best to construct the variance estimate using $\widetilde{\beta}_{\text{emd}}$. The EMD residuals are $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}_{\text{emd}}$. Using these we can estimate the matrix $\Omega = \mathbb{E}\left[XX'e^2\right]$ by

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

Following the formula for CLS we recommend an adjusted degrees of freedom. Given $\widetilde{\Omega}$ the moment estimator of $\boldsymbol{V}_\beta$ is $\widetilde{\boldsymbol{V}}_\beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widetilde{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}$. Given this, we construct the variance estimator

$$\widetilde{\boldsymbol{V}}_{\beta,\text{emd}} = \widetilde{\boldsymbol{V}}_\beta - \widetilde{\boldsymbol{V}}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\widetilde{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\boldsymbol{V}}_\beta. \tag{8.35}$$

A standard error for $h'\widetilde{\beta}$ is then

$$s\left(h'\widetilde{\beta}\right) = \left(n^{-1}h'\widetilde{\boldsymbol{V}}_{\beta,\text{emd}}h\right)^{1/2}. \tag{8.36}$$

8.11 Hausman Equality

Form (8.25) we have

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right) &= \widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\beta}_{\text{ols}} - \boldsymbol{c}\right)\\ &\xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta\right).\end{aligned}$$

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

$$\operatorname{avar}\left[\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right] = \operatorname{avar}\left[\widehat{\beta}_{\text{ols}}\right] - \operatorname{avar}\left[\widetilde{\beta}_{\text{emd}}\right]. \tag{8.37}$$

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

8.12 Example: Mankiw, Romer and Weil (1992)

We illustrate the methods by replicating some of the estimates reported in a well-known paper by Mankiw, Romer, and Weil (1992). The paper investigates the implications of the Solow growth model using cross-country regressions. A key equation in their paper regresses the change between 1960 and 1985 in log GDP per capita on (1) log GDP in 1960, (2) the log of the ratio of aggregate investment to

Table 8.1: Estimates of Solow Growth Model

	$\widehat{\beta}_{\text{ols}}$	$\widehat{\beta}_{\text{cls}}$	$\widehat{\beta}_{\text{emd}}$
$\log GDP_{1960}$	−0.29	−0.30	−0.30
	(0.05)	(0.05)	(0.05)
$\log \frac{I}{GDP}$	0.52	0.50	0.46
	(0.11)	(0.09)	(0.08)
$\log\left(n+g+\delta\right)$	−0.51	−0.74	−0.71
	(0.24)	(0.08)	(0.07)
log(School)	0.23	0.24	0.25
	(0.07)	(0.07)	(0.06)
Intercept	3.02	2.46	2.48
	(0.74)	(0.44)	(0.44)

Standard errors are heteroskedasticity-consistent

GDP, (3) the log of the sum of the population growth rate n, the technological growth rate g, and the rate of depreciation δ, and (4) the log of the percentage of the working-age population that is in secondary school (*School*), the latter a proxy for human-capital accumulation.

The data is available on the textbook webpage in the file `MRW1992`.

The sample is 98 non-oil-producing countries and the data was reported in the published paper. As g and δ were unknown the authors set $g+\delta = 0.05$. We report least squares estimates in the first column of Table 8.1. The estimates are consistent with the Solow theory due to the positive coefficients on investment and human capital and negative coefficient for population growth. The estimates are also consistent with the convergence hypothesis (that income levels tend towards a common mean over time) as the coefficient on intial GDP is negative.

The authors show that in the Solow model the 2^{nd}, 3^{rd} and 4^{th} coefficients sum to zero. They reestimated the equation imposing this constraint. We present constrained least squares estimates in the second column of Table 8.1 and efficient minimum distance estimates in the third column. Most of the coefficients and standard errors only exhibit small changes by imposing the constraint. The one exception is the coefficient on log population growth which increases in magnitude and its standard error decreases substantially. The differences between the CLS and EMD estimates are modest.

We now present Stata, R and MATLAB code which implements these estimates.

You may notice that the Stata code has a section which uses the Mata matrix programming language. This is used because Stata does not implement the efficient minimum distance estimator, so needs to be separately programmed. As illustrated here, the Mata language allows a Stata user to implement methods using commands which are quite similar to MATLAB.

Stata do File

```
use "MRW1992.dta", clear
gen lndY = log(Y85)-log(Y60)
gen lnY60 = log(Y60)
gen lnI = log(invest/100)
gen lnG = log(pop_growth/100+0.05)
gen lnS = log(school/100)
* Unrestricted regression
reg lndY lnY60 lnI lnG lnS if N==1, r
* Store result for efficient minimum distance
mat b = e(b)'
scalar k = e(rank)
mat V = e(V)
* Constrained regression
constraint define 1 lnI+lnG+lnS=0
cnsreg lndY lnY60 lnI lnG lnS if N==1, constraints(1) r
* Efficient minimum distance
mata{
    data = st_data(.,("lnY60","lnI","lnG","lnS","lndY","N"))
    data_select = select(data,data[.,6]:==1)
    y = data_select[.,5]
    n = rows(y)
    x = (data_select[.,1..4],J(n,1,1))
    k = cols(x)
    invx = invsym(x'*x)
    b_ols = st_matrix("b")
    V_ols = st_matrix("V")
    R = (0 \ 1 \ 1 \ 1 \ 0)
    b_emd = b_ols-V_ols*R*invsym(R'*V_ols*R)*R'*b_ols
    e_emd = J(1,k,y-x*b_emd)
    xe_emd = x:*e_emd
    xe_emd'*xe_emd
    V2 = (n/(n-k+1))*invx*(xe_emd'*xe_emd)*invx
    V_emd = V2 - V2*R*invsym(R'*V2*R)*R'*V2
    se_emd = diagonal(sqrt(V_emd))
    st_matrix("b_emd",b_emd)
    st_matrix("se_emd",se_emd)}
mat list b_emd
mat list se_emd
```

R Program File

```
data <- read.table("MRW1992.txt",header=TRUE)
N <- matrix(data$N,ncol=1)
lndY <- matrix(log(data$Y85)-log(data$Y60),ncol=1)
lnY60 <- matrix(log(data$Y60),ncol=1)
lnI <- matrix(log(data$invest/100),ncol=1)
lnG <- matrix(log(data$pop_growth/100+0.05),ncol=1)
lnS <- matrix(log(data$school/100),ncol=1)
xx <- as.matrix(cbind(lnY60,lnI,lnG,lnS,matrix(1,nrow(lndY),1)))
x <- xx[N==1,]
y <- lndY[N==1]
n <- nrow(x)
k <- ncol(x)
# Unrestricted regression
invx <-solve(t(x)%*%x)
b_ols <- solve((t(x)%*%x),(t(x)%*%y))
e_ols <- rep((y-x%*%beta_ols),times=k)
xe_ols <- x*e_ols
V_ols <- (n/(n-k))*invx%*%(t(xe_ols)%*%xe_ols)%*%invx
se_ols <- sqrt(diag(V_ols))
print(beta_ols)
print(se_ols)
# Constrained regression
R <- c(0,1,1,1,0)
iR <- invx%*%R%*%solve(t(R)%*%invx%*%R)%*%t(R)
b_cls <- b_ols - iR%*%b_ols
e_cls <- rep((y-x%*%b_cls),times=k)
xe_cls <- x*e_cls
V_tilde <- (n/(n-k+1))*invx%*%(t(xe_cls)%*%xe_cls)%*%invx
V_cls <- V_tilde - iR%*%V_tilde - V_tilde%*%t(iR) +iR%*%V_tilde%*%t(iR)
print(b_cls)print(se_cls)
# Efficient minimum distance
Vr <- V_ols%*%R%*%solve(t(R)%*%V_ols%*%R)%*%t(R)
b_emd <- b_ols - Vr%*%b_ols
e_emd <- rep((y-x%*%b_emd),times=k)
xe_emd <- x*e_emd
V2 <- (n/(n-k+1))*invx%*%(t(xe_emd)%*%xe_emd)%*%invx
V_emd <- V2 - V2%*%R%*%solve(t(R)%*%V2%*%R)%*%t(R)%*%V2
se_emd <- sqrt(diag(V_emd))
```

MATLAB Program File

```
data = xlsread(’MRW1992.xlsx’);
N = data(:,1);
Y60 = data(:,4);
Y85 = data(:,5);
pop_growth = data(:,7);
invest = data(:,8);
school = data(:,9);
lndY = log(Y85)-log(Y60);
lnY60 = log(Y60);
lnI = log(invest/100);
lnG = log(pop_growth/100+0.05);
lnS = log(school/100);
xx = [lnY60,lnI,lnG,lnS,ones(size(lndY,1),1)];
x = xx(N==1,:);
y = lndY(N==1);
[n,k] = size(x);
% Unrestricted regression
invx = inv(x’*x);
beta_ols = (x’*x)\(x’*y);
xe_ols = x.*(y-x*beta_ols);
V_ols = (n/(n-k))*invx*(xe_ols’*xe_ols)*invx;
se_ols = sqrt(diag(V_ols));
display(beta_ols);
display(se_ols);
% Constrained regression
R = [0;1;1;1;0];
iR = invx*R*inv(R’*invx*R)*R’;
beta_cls = beta_ols - iR*beta_ols;
xe_cls = x.*(y-x*beta_cls);
V_tilde = (n/(n-k+1))*invx*(xe_cls’*xe_cls)*invx;
V_cls = V_tilde - iR*V_tilde - V_tilde*(iR’) + iR*V_tilde*(iR’);
se_cls = sqrt(diag(V_cls));
display(beta_cls);display(se_cls);
% Efficient minimum distance
beta_emd = beta_ols-V_ols*R*inv(R’*V_ols*R)*R’*beta_ols;
xe_emd = x.*(y-x*beta_emd);
V2 = (n/(n-k+1))*invx*(xe_emd’*xe_emd)*invx;
V_emd = V2 - V2*R*inv(R’*V2*R)*R’*V2;
se_emd = sqrt(diag(V_emd));
display(beta_emd);display(se_emd);
```

8.13 Misspecification

What are the consequences for a constrained estimator $\widetilde{\beta}$ if the constraint (8.1) is incorrect? To be specific suppose that the truth is

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

where $\boldsymbol{c}^*$ is not necessarily equal to $\boldsymbol{c}$.

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 answer is to apply formula (8.23) to find that

$$\widetilde{\beta}_{\text{md}} \xrightarrow[p]{} \beta^*_{\text{md}} = \beta - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{c}^* - \boldsymbol{c}\right). \tag{8.38}$$

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 β^*_{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

$$\beta^*_{\text{cls}} = \underset{\boldsymbol{R}'\beta=\boldsymbol{c}}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - X'\beta\right)^2\right],$$

the best linear predictor subject to the constraint (8.1). The minimum distance estimator converges in probability to

$$\beta^*_{\text{md}} = \underset{\boldsymbol{R}'\beta=\boldsymbol{c}}{\operatorname{argmin}} \left(\beta - \beta_0\right)' \boldsymbol{W}\left(\beta - \beta_0\right)$$

where β_0 is the true coefficient. That is, β^*_{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 $\widetilde{\beta}_{\text{md}}$ has an asymptotic normal distribution. The trick is to define the pseudo-true value

$$\beta^*_n = \beta - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{c}^* - \boldsymbol{c}\right). \tag{8.39}$$

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

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\beta}_{\text{md}} - \beta^*_n\right) &= \sqrt{n}\left(\widehat{\beta} - \beta\right) - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}^*\right) \\
&= \left(\boldsymbol{I}_k - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\beta} - \beta\right) \\
&\xrightarrow[d]{} \left(\boldsymbol{I}_k - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{N}\left(0, \boldsymbol{V}_\beta\right) \\
&= \mathrm{N}\left(0, \boldsymbol{V}_\beta(\boldsymbol{W})\right).
\end{aligned} \tag{8.40}$$

In particular

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

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) β_n^* rather than β. 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 β_n by n via the relationship

$$\boldsymbol{R}'\beta_n = \boldsymbol{c} + \delta n^{-1/2}. \tag{8.41}$$

for some $\delta \in \mathbb{R}^q$. Equation (8.41) specifies that β_n violates (8.1) and thus the constraint is misspecified. However, the constraint is "close" to correct as the difference $\boldsymbol{R}'\beta_n - \boldsymbol{c} = \delta n^{-1/2}$ is "small" in the sense that it decreases with the sample size n. We call (8.41) **local misspecification**.

The asymptotic theory is derived as $n \to \infty$ under the sequence of probability distributions with the coefficients β_n. The way to think about this is that the true value of the parameter is β_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 δ appears in the asymptotic distribution but does not dominate it. The best way to see this is to work through the asymptotic approximation.

Since β_n is the true coefficient value, then $Y = X'\beta_n + e$ and we have the standard representation for the unconstrained estimator, namely

$$\sqrt{n}\left(\widehat{\beta} - \beta_n\right) = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i e_i\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right). \tag{8.42}$$

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

A difference arises for the constrained estimator. Using (8.41), $\boldsymbol{c} = \boldsymbol{R}'\beta_n - \delta n^{-1/2}$ so

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

and

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

It follows that

$$\sqrt{n}\left(\widetilde{\beta}_{\mathrm{md}} - \beta_n\right) = \left(\boldsymbol{I}_k - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\beta} - \beta_n\right) + \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\delta.$$

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 $\sqrt{n}$ scaling of the estimator. No alternative rate would have produced this result.

Consequently we find that the asymptotic distribution equals

$$\sqrt{n}\left(\widetilde{\beta}_{\mathrm{md}} - \beta_n\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right) + \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\delta = \mathrm{N}\left(\delta^*, \boldsymbol{V}_\beta(\boldsymbol{W})\right) \tag{8.43}$$

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

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 δ^*. 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. However, (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 β. They can be written as

$$r(\beta)=0 \tag{8.44}$$

where $r:\mathbb{R}^k\to\mathbb{R}^q$. 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 $r(\beta)=\beta_1\beta_2-1$.

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

$$\widetilde{\beta}_{\text{cls}}=\underset{r(\beta)=0}{\operatorname{argmin}}\,\text{SSE}(\beta) \tag{8.45}$$

$$\widetilde{\beta}_{\text{md}}=\underset{r(\beta)=0}{\operatorname{argmin}}\;J\left(\beta\right) \tag{8.46}$$

where SSE(β) and $J\left(\beta\right)$ are defined in (8.4) and (8.19), respectively. The solutions solve the Lagrangians

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

or

$$\mathscr{L}(\beta,\lambda)=\frac{1}{2}J\left(\beta\right)+\lambda' r(\beta) \tag{8.47}$$

over (β,λ).

Computationally there is no general closed-form solution 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 and R. See Chapter 12 of *Probability and Statistics for Economists.*

Assumption 8.3

1. $r(\beta)=0$.
2. $r(\beta)$ is continuously differentiable at the true β.
3. rank($\boldsymbol{R}$) $=q$, where $\boldsymbol{R}=\dfrac{\partial}{\partial\beta}r(\beta)'$.

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 $\widetilde{\beta} = \widetilde{\beta}_{\text{md}}$ and $\widetilde{\beta} = \widetilde{\beta}_{\text{cls}}$ defined in (8.45) and (8.46),

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

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

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

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

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

where

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial\beta} r(\widetilde{\beta}_{\text{md}})'. \tag{8.48}$$

Standard errors for the elements of $\widetilde{\beta}_{\text{md}}$ are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widetilde{\beta}}^{*} = n^{-1}\widehat{\boldsymbol{V}}_{\beta}^{*}$.

8.15 Inequality Restrictions

Inequality constraints on the parameter vector β take the form

$$r(\beta) \geq 0 \tag{8.49}$$

for some function $r: \mathbb{R}^k \to \mathbb{R}^q$. 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

$$\widetilde{\beta}_{\text{cls}} = \underset{r(\beta)\geq 0}{\operatorname{argmin}} \operatorname{SSE}(\beta) \tag{8.50}$$

and

$$\widetilde{\beta}_{\text{md}} = \underset{r(\beta)\geq 0}{\operatorname{argmin}}\ J\left(\beta\right). \tag{8.51}$$

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. $\beta_1 \geq 0$ with $q = 1$. In this case the constrained estimator can be found by the following approach. Compute the uncontrained estimator $\widehat{\beta}$. If $\widehat{\beta}_1 \geq 0$ then $\widetilde{\beta} = \widehat{\beta}$. Otherwise if $\widehat{\beta}_1 < 0$ then impose $\beta_1 = 0$ (eliminate the regressor X_1) and re-estimate. This method 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 computation problems (8.50) and (8.51) are examples of **quadratic programming**. Quick computer algorithms are available in programming languages including MATLAB 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 $\widehat{\beta}$ which satisfies $\sqrt{n}\left(\widehat{\beta}-\beta\right) \xrightarrow[d]{} Z = \mathrm{N}\left(0,V_\beta\right)$ and $\beta=0$, then the constrained estimator $\widetilde{\beta}=\max[\widehat{\beta},0]$ will have the asymptotic distribution $\sqrt{n}\widetilde{\beta} \xrightarrow[d]{} \max[Z,0]$, a "half-normal" distribution.

8.16 Technical Proofs*

Proof of Theorem 8.9, equation (8.28) Let $\boldsymbol{R}_\perp$ be a full rank $k\times(k-q)$ matrix satisfying $\boldsymbol{R}_\perp' \boldsymbol{V}_\beta \boldsymbol{R}=0$ and then set $\boldsymbol{C}=[\boldsymbol{R},\boldsymbol{R}_\perp]$ which is full rank and invertible. Then we can calculate that

$$\boldsymbol{C}'\boldsymbol{V}_\beta^*\boldsymbol{C}=\left[\begin{array}{cc}\boldsymbol{R}'\boldsymbol{V}_\beta^*\boldsymbol{R} & \boldsymbol{R}'\boldsymbol{V}_\beta^*\boldsymbol{R}_\perp \\ \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*\boldsymbol{R} & \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*\boldsymbol{R}_\perp\end{array}\right]=\left[\begin{array}{cc}0 & 0\\ 0 & \boldsymbol{R}_\perp'\boldsymbol{V}_\beta\boldsymbol{R}_\perp\end{array}\right]$$

and

$$\begin{aligned}&\boldsymbol{C}'\boldsymbol{V}_\beta(\boldsymbol{W})\boldsymbol{C}\\ &=\left[\begin{array}{cc}\boldsymbol{R}'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R} & \boldsymbol{R}'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R}_\perp\\ \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R} & \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R}_\perp\end{array}\right]\\ &=\left[\begin{array}{cc}0 & 0\\ 0 & \boldsymbol{R}_\perp'\boldsymbol{V}_\beta\boldsymbol{R}_\perp+\boldsymbol{R}_\perp'\boldsymbol{W}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}_\perp\end{array}\right].\end{aligned}$$

Thus

$$\begin{aligned}&\boldsymbol{C}'\left(\boldsymbol{V}_\beta(\boldsymbol{W})-\boldsymbol{V}_\beta^*\right)\boldsymbol{C}\\ &=\boldsymbol{C}'\boldsymbol{V}_\beta(\boldsymbol{W})\boldsymbol{C}-\boldsymbol{C}'\boldsymbol{V}_\beta^*\boldsymbol{C}\\ &=\left[\begin{array}{cc}0 & 0\\ 0 & \boldsymbol{R}_\perp'\boldsymbol{W}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}_\perp\end{array}\right]\\ &\geq 0\end{aligned}$$

Since $\boldsymbol{C}$ is invertible it follows that $\boldsymbol{V}_\beta(\boldsymbol{W})-\boldsymbol{V}_\beta^*\geq 0$ which is (8.28). ■

Proof of Theorem 8.10 We show the result for the minimum distance estimator $\widetilde{\beta}=\widetilde{\beta}_{\mathrm{md}}$ as the proof for the constrained least squares estimator is similar. For simplicity we assume that the constrained estimator is consistent $\widetilde{\beta}\xrightarrow[p]{}\beta$. This can be shown with more effort, but requires a deeper treatment than appropriate for this textbook.

For each element $r_j(\beta)$ of the q-vector $r(\beta)$, by the mean value theorem there exists a β_j^* on the line segment joining $\widetilde{\beta}$ and β such that

$$r_j(\widetilde{\beta})=r_j(\beta)+\frac{\partial}{\partial\beta}r_j(\beta_j^*)'\left(\widetilde{\beta}-\beta\right). \tag{8.52}$$

Let $\boldsymbol{R}_n^*$ be the $k\times q$ matrix

$$\boldsymbol{R}^*=\left[\begin{array}{cccc}\frac{\partial}{\partial\beta}r_1(\beta_1^*) & \frac{\partial}{\partial\beta}r_2(\beta_2^*) & \cdots & \frac{\partial}{\partial\beta}r_q(\beta_q^*)\end{array}\right].$$

Since $\widetilde{\beta} \underset{p}{\longrightarrow} \beta$ it follows that $\beta_j^* \underset{p}{\longrightarrow} \beta$, and by the CMT, $\boldsymbol{R}^* \underset{p}{\longrightarrow} \boldsymbol{R}$. Stacking the (8.52), we obtain

$$r(\widetilde{\beta}) = r(\beta) + \boldsymbol{R}^{*\prime}\left(\widetilde{\beta} - \beta\right).$$

Since $r\left(\widetilde{\beta}\right) = 0$ by construction and $r(\beta) = 0$ by Assumption 8.1 this implies

$$0 = \boldsymbol{R}^{*\prime}\left(\widetilde{\beta} - \beta\right). \tag{8.53}$$

The first-order condition for (8.47) is $\widehat{\boldsymbol{W}}\left(\widehat{\beta} - \widetilde{\beta}\right) = \widehat{\boldsymbol{R}}\widetilde{\lambda}$ where $\widehat{\boldsymbol{R}}$ is defined in (8.48). Premultiplying by $\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}$, inverting, and using (8.53), we find

$$\widetilde{\lambda} = \left(\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}^{*\prime}\left(\widehat{\beta} - \widetilde{\beta}\right) = \left(\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}^{*\prime}\left(\widehat{\beta} - \beta\right).$$

Thus

$$\widetilde{\beta} - \beta = \left(\boldsymbol{I}_k - \widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\left(\boldsymbol{R}_n^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}_n^{*\prime}\right)\left(\widehat{\beta} - \beta\right). \tag{8.54}$$

From Theorem 7.3 and Theorem 7.6 we find

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

■

8.17 Exercises

Exercise 8.1 In the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$, show directly from definition (8.3) that the CLS estimator of $\beta = (\beta_1, \beta_2)$ subject to the constraint that $\beta_2 = 0$ is the OLS regression of Y on X_1.

Exercise 8.2 In the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$, show directly from definition (8.3) that the CLS estimator of $\beta = (\beta_1, \beta_2)$ subject to the constraint $\beta_1 = \boldsymbol{c}$ (where $\boldsymbol{c}$ is some given vector) is OLS of $Y - X_1'\boldsymbol{c}$ on X_2.

Exercise 8.3 In the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$, with β_1 and β_2 each $k \times 1$, find the CLS estimator of $\beta = (\beta_1, \beta_2)$ subject to the constraint that $\beta_1 = -\beta_2$.

Exercise 8.4 In the linear projection model $Y = \alpha + X'\beta + e$ consider the restriction $\beta = 0$.

(a) Find the CLS estimator of α under the restriction $\beta = 0$.

(b) Find an expression for the efficient minimum distance estimator of α under the restriction $\beta = 0$.

Exercise 8.5 Verify that for $\widetilde{\beta}_{\text{cls}}$ defined in (8.8) that $\boldsymbol{R}'\widetilde{\beta}_{\text{cls}} = \boldsymbol{c}$.

Exercise 8.6 Prove Theorem 8.1.

Exercise 8.7 Prove Theorem 8.2, that is, $\mathbb{E}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] = \beta$, under the assumptions of the linear regression regression model and (8.1). (Hint: Use Theorem 8.1.)

Exercise 8.8 Prove Theorem 8.3.

Exercise 8.9 Prove Theorem 8.4. That is, show $\mathbb{E}\left[s^2_{\text{cls}} \mid \boldsymbol{X}\right] = \sigma^2$ under the assumptions of the homoskedastic regression model and (8.1).

Exercise 8.10 Verify (8.22), (8.23), and that the minimum distance estimator $\widetilde{\beta}_{\text{md}}$ with $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{XX}$ equals the CLS estimator.

Exercise 8.11 Prove Theorem 8.6.

Exercise 8.12 Prove Theorem 8.7.

Exercise 8.13 Prove Theorem 8.8. (Hint: Use that CLS is a special case of Theorem 8.7.)

Exercise 8.14 Verify that (8.26) is $\boldsymbol{V}_\beta(\boldsymbol{W})$ with $\boldsymbol{W} = \boldsymbol{V}_\beta^{-1}$.

Exercise 8.15 Prove (8.27). Hint: Use (8.26).

Exercise 8.16 Verify (8.29), (8.30) and (8.31).

Exercise 8.17 Verify (8.32), (8.33), and (8.34).

Exercise 8.18 Suppose you have two independent samples each with n observations which satisfy the models $Y_1 = X_1'\beta_1 + e_1$ with $\mathbb{E}[X_1 e_1] = 0$ and $Y_2 = X_2'\beta_2 + e_2$ with $\mathbb{E}[X_2 e_2] = 0$ where β_1 and β_2 are both $k \times 1$. You estimate β_1 and β_2 by OLS on each sample, with consistent asymptotic covariance matrix estimators $\widehat{\boldsymbol{V}}_{\beta_1}$ and $\widehat{\boldsymbol{V}}_{\beta_2}$. Consider efficient minimum distance estimation under the restriction $\beta_1 = \beta_2$.

(a) Find the estimator $\widetilde{\beta}$ of $\beta = \beta_1 = \beta_2$.

(b) Find the asymptotic distribution of $\widetilde{\beta}$.

(c) How would you approach the problem if the sample sizes are different, say n_1 and n_2?

Exercise 8.19 Use the `cps09mar` dataset and the subsample of white male Hispanics.

(a) Estimate the regression

$$\begin{aligned}\widehat{\log(wage)} &= \beta_1\, education + \beta_2\, experience + \beta_3\, experience^2/100 + \beta_4 married_1 \\ &\quad + \beta_5 married_2 + \beta_6 married_3 + \beta_7 widowed + \beta_8 divorced + \beta_9 separated + \beta_{10}\end{aligned}$$

where $married_1$, $married_2$, and $married_3$ are the first three marital codes listed in Section 3.22.

(b) Estimate the equation by CLS imposing the constraints $\beta_4 = \beta_7$ and $\beta_8 = \beta_9$. Report the estimates and standard errors.

(c) Estimate the equation using efficient minimum distance imposing the same constraints. Report the estimates and standard errors.

(d) Under what constraint on the coefficients is the wage equation non-decreasing in experience for experience up to 50?

(e) Estimate the equation imposing $\beta_4 = \beta_7$, $\beta_8 = \beta_9$, and the inequality from part (d).

Exercise 8.20 Take the model

$$\begin{aligned} Y &= m(X)+e \\ m(x) &= \beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_p x^p \\ \mathbb{E}\left[X^j e\right] &= 0, \qquad j=0,\dots,p \\ g(x) &= \frac{d}{dx} m(x) \end{aligned}$$

with i.i.d. observations (Y_i, X_i), $i=1,\dots,n$. The order of the polynomial p is known.

(a) How should we interpret the function $m(x)$ given the projection assumption? How should we interpret $g(x)$? (Briefly)

(b) Describe an estimator $\widehat{g}(x)$ of $g(x)$.

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{g}(x)-g(x)\right)$ as $n\to\infty$.

(d) Show how to construct an asymptotic 95% confidence interval for $g(x)$ (for a single x).

(e) Assume $p=2$. Describe how to estimate $g(x)$ imposing the constraint that $m(x)$ is concave.

(f) Assume $p=2$. Describe how to estimate $g(x)$ imposing the constraint that $m(u)$ is increasing on the region $u\in[x_L,x_U]$.

Exercise 8.21 Take the linear model with restrictions $Y=X'\beta+e$ with $\mathbb{E}[Xe]=0$ and $\boldsymbol{R}'\beta=\boldsymbol{c}$. Consider three estimators for β:

- $\widehat{\beta}$ the unconstrained least squares estimator
- $\widetilde{\beta}$ the constrained least squares estimator
- $\overline{\beta}$ the constrained efficient minimum distance estimator

For the three estimator define the residuals $\widehat{e}_i=Y_i-X_i'\widehat{\beta}$, $\widetilde{e}_i=Y_i-X_i'\widetilde{\beta}$, $\overline{e}_i=Y_i-X_i'\overline{\beta}$, and variance estimators $\widehat{\sigma}^2=n^{-1}\sum_{i=1}^n\widehat{e}_i^2$, $\widetilde{\sigma}^2=n^{-1}\sum_{i=1}^n\widetilde{e}_i^2$, and $\overline{\sigma}^2=n^{-1}\sum_{i=1}^n\overline{e}_i^2$.

(a) As $\overline{\beta}$ is the most efficient estimator and $\widehat{\beta}$ the least, do you expect $\overline{\sigma}^2<\widetilde{\sigma}^2<\widehat{\sigma}^2$ in large samples?

(b) Consider the statistic

$$T_n=\widehat{\sigma}^{-2}\sum_{i=1}^n\left(\widehat{e}_i-\widetilde{e}_i\right)^2.$$

Find the asymptotic distribution for T_n when $\boldsymbol{R}'\beta=\boldsymbol{c}$ is true.

(c) Does the result of the previous question simplify when the error e_i is homoskedastic?

Exercise 8.22 Take the linear model $Y=X_1\beta_1+X_2\beta_2+e$ with $\mathbb{E}[Xe]=0$. Consider the restriction $\dfrac{\beta_1}{\beta_2}=2$.

(a) Find an explicit expression for the CLS estimator $\widetilde{\beta}=(\widetilde{\beta}_1,\widetilde{\beta}_2)$ of $\beta=(\beta_1,\beta_2)$ under the restriction. Your answer should be specific to the restriction. It should not be a generic formula for an abstract general restriction.

(b) Derive the asymptotic distribution of $\widetilde{\beta}_1$ under the assumption that the restriction is true.