Chapter 5. Least Squares Estimation - Large-Sample Properties*

In Chapter 3, we assume $u|\mathbf{x} \sim N(0, \sigma^2)$ and study the conditional distribution of $\widehat{\boldsymbol{\beta}}$ given $\mathbf{X}$. In general the distribution of $u|\mathbf{x}$ is unknown and even if it is known, the unconditional distribution of $\widehat{\boldsymbol{\beta}}$ is hard to derive since $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is a complicated function of $\{\mathbf{x}_i\}_{i=1}^n$. Asymptotic (or large sample) methods approximate sampling distributions based on the limiting experiment that the sample size n tends to infinity. A preliminary step in this approach is the demonstration that estimators converge in probability to the true parameters as the sample size gets large. The second step is to study the distributional properties of $\widehat{\boldsymbol{\beta}}$ in the neighborhood of the true value, that is, the asymptotic normality of $\widehat{\boldsymbol{\beta}}$. The final step is to estimate the asymptotic variance which is necessary in statistical inferences such as hypothesis testing and confidence interval (CI) construction. In hypothesis testing, it is necessary to construct test statistics and derive their asymptotic distributions under the null. We will study the t-test and three asymptotically equivalent tests under both homoskedasticity and heteroskedasticity. It is also standard to develop the local power function for illustrating the power properties of the test.

This chapter concentrates on asymptotic properties related to the LSE. Related materials can be found in Chapter 2 of Hayashi (2000), Chapter 4 of Cameron and Trivedi (2005), Chapter 4 of Hansen (2007), and Chapter 4 of Wooldridge (2010).

1 Asymptotics for the LSE

We first show that the LSE is CAN and then re-derive its asymptotic distribution by treating it as a MoM estimator.

1.1 Consistency

It is useful to express $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \tag{1}$$

To show $\widehat{\boldsymbol{\beta}}$ is consistent, we impose the following additional assumptions.

Assumption OLS.1′: $\text{rank}(E[\mathbf{x}\mathbf{x}']) = k$.

Assumption OLS.2′: $y = \mathbf{x}'\boldsymbol{\beta} + u$ with $E[\mathbf{x}u] = \mathbf{0}$.

*Email: pingyu@hku.hk

Note that Assumption OLS.1′ implicitly assumes that $E\left[\|\mathbf{x}\|^2\right] < \infty$. Assumption OLS.1′ is the large-sample counterpart of Assumption OLS.1, and Assumption OLS.2′ is weaker than Assumption OLS.2.

Theorem 1 *Under Assumptions OLS.0, OLS.1′, OLS.2′ and OLS.3,* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Proof. From (1), to show $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, we need only to show that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \xrightarrow{p} 0$. Note that

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i\right) \\ &= g\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i', \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i\right) \xrightarrow{p} E[\mathbf{x}_i\mathbf{x}_i']^{-1}E[\mathbf{x}_i u_i] = 0.\end{aligned}$$

Here, the convergence in probability is from (I) the WLLN which implies

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} E[\mathbf{x}_i\mathbf{x}_i'] \text{ and } \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i \xrightarrow{p} E[\mathbf{x}_i u_i]; \tag{2}$$

(II) the fact that $g(\mathbf{A}, \mathbf{b}) = \mathbf{A}^{-1}\mathbf{b}$ is a continuous function at $(E[\mathbf{x}_i\mathbf{x}_i'], E[\mathbf{x}_i u_i])$. The last equality is from Assumption OLS.2′.

(I) To apply the WLLN, we require (i) $\mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{x}_i u_i$ are i.i.d., which is implied by Assumption OLS.0 and that functions of i.i.d. data are also i.i.d.; (ii) $E\left[\|\mathbf{x}\|^2\right] < \infty$ (OLS.1′) and $E[\|\mathbf{x}u\|] < \infty$. $E[\|\mathbf{x}u\|] < \infty$ is implied by the Cauchy-Schwarz inequality,[1]

$$E[\|\mathbf{x}u\|] \le E\left[\|\mathbf{x}\|^2\right]^{1/2} E\left[|u|^2\right]^{1/2},$$

which is finite by Assumption OLS.1′ and OLS.3. (II) To guarantee $\mathbf{A}^{-1}\mathbf{b}$ to be a continuous function at $(E[\mathbf{x}_i\mathbf{x}_i'], E[\mathbf{x}_i u_i])$, we must assume that $E[\mathbf{x}_i\mathbf{x}_i']^{-1}$ exists which is implied by Assumption OLS.1′.[2] ■

Exercise 1 *Take the model* $y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_i$ *with* $E[\mathbf{x}_i u_i] = \mathbf{0}$. *Suppose that* $\boldsymbol{\beta}_1$ *is estimated by regressing* y_i *on* $\mathbf{x}_{1i}$ *only. Find the probability limit of this estimator. In general, is it consistent for* $\boldsymbol{\beta}_1$*? If not, under what conditions is this estimator consistent for* $\boldsymbol{\beta}_1$*?*

We can similarly show that the estimators $\widehat{\sigma}^2$ and s^2 are consistent for σ^2.

Theorem 2 *Under the assumptions of Theorem 1,* $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ *and* $s^2 \xrightarrow{p} \sigma^2$.

[1] Cauchy-Schwarz inequality: For any random $m \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$, $E\left[\|\mathbf{X}'\mathbf{Y}\|\right] \le E\left[\|\mathbf{X}\|^2\right]^{1/2} E\left[\|\mathbf{Y}\|^2\right]^{1/2}$, where the inner product is defined as $\langle \mathbf{X}, \mathbf{Y} \rangle = E\left[\|\mathbf{X}'\mathbf{Y}\|\right]$.

[2] If $x_i \in \mathbb{R}$, $E[x_i x_i']^{-1} = E[x_i^2]^{-1}$ is the reciprocal of $E[x_i^2]$ which is a continuous function of $E[x_i^2]$ only if $E[x_i^2] \ne 0$.

Proof. Note that

$$\begin{aligned}
\widehat{u}_i &= y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}} \\
&= u_i + \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\widehat{\boldsymbol{\beta}} \\
&= u_i - \mathbf{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).
\end{aligned}$$

Thus

$$\widehat{u}_i^2 = u_i^2 - 2u_i\mathbf{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\mathbf{x}_i\mathbf{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \tag{3}$$

and

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \widehat{u}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^n u_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n u_i\mathbf{x}_i'\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&\xrightarrow{p} \sigma^2,
\end{aligned}$$

where the last line uses the WLLN, (2), Theorem 1 and the CMT.

Finally, since $n/(n-k) \to 1$, it follows that

$$s^2 = \frac{n}{n-k}\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$$

by the CMT. ■

One implication of this theorem is that multiple estimators can be consistent for the population parameter. While $\widehat{\sigma}^2$ and s^2 are unequal in any given application, they are close in value when n is very large.

1.2 Asymptotic Normality

To study the asymptotic normality of $\widehat{\boldsymbol{\beta}}$, we impose the following additional assumption.

Assumption OLS.5: $E[u^4] < \infty$ and $E\left[\|\mathbf{x}\|^4\right] < \infty$.

Theorem 3 *Under Assumptions OLS.0, OLS.1′, OLS.2′, OLS.3 and OLS.5,*

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

where $\mathbf{V} = \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}$ *with* $\mathbf{Q} = E\left[\mathbf{x}_i\mathbf{x}_i'\right]$ *and* $\boldsymbol{\Omega} = E\left[\mathbf{x}_i\mathbf{x}_i'u_i^2\right]$.

Proof. From (1),

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{x}_i u_i\right).$$

Note first that

$$E\left[\left\|\mathbf{x}_i\mathbf{x}_i'u_i^2\right\|\right] \leq E\left[\left\|\mathbf{x}_i\mathbf{x}_i'\right\|^2\right]^{1/2} E\left[u_i^4\right]^{1/2} \leq E\left[\|\mathbf{x}_i\|^4\right]^{1/2} E\left[u_i^4\right]^{1/2} < \infty, \tag{4}$$

where the first inequality is from the Cauchy-Schwarz inequality, the second inequality is from the Schwarz matrix inequality,[3] and the last inequality is from Assumption OLS.5. So by the CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i u_i \xrightarrow{d} N\left(\mathbf{0},\boldsymbol{\Omega}\right).$$

Given that $n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \mathbf{Q}$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{d} \mathbf{Q}^{-1}N\left(\mathbf{0},\boldsymbol{\Omega}\right) = N(\mathbf{0},\mathbf{V})$$

by Slutsky's theorem. ■

In the homoskedastic model, $\mathbf{V}$ reduces to $\mathbf{V}^0 = \sigma^2\mathbf{Q}^{-1}$. We call $\mathbf{V}^0$ the **homoskedastic covariance matrix**. Sometimes, to state the asymptotic distribution of part of $\widehat{\boldsymbol{\beta}}$ as in the residual regression, we partition $\mathbf{Q}$ and $\boldsymbol{\Omega}$ as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{21} \end{pmatrix}, \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}. \tag{5}$$

Recall from the proof of the FWL theorem,

$$\mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{Q}_{11.2}^{-1} & -\mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22.1}^{-1}\mathbf{Q}_{21}\mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22.1}^{-1} \end{pmatrix},$$

where $\mathbf{Q}_{11.2} = \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}$ and $\mathbf{Q}_{22.1} = \mathbf{Q}_{22} - \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}$. Thus when the error is homoskedastic, $n \cdot AVar\left(\widehat{\boldsymbol{\beta}}_1\right) = \sigma^2\mathbf{Q}_{11.2}^{-1}$, and $n \cdot ACov\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2\right) = -\sigma^2\mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}$. We can also derive the general formulas in the heteroskedastic case, but these formulas are not easily interpretable and so less useful.

Exercise 2 *Of the variables $(y_i^*, y_i, \mathbf{x}_i)$ only the pair $(y_i, \mathbf{x}_i)$ are observed. In this case, we say that y_i^* is a* latent *variable. Suppose*

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + u_i, \\ E[\mathbf{x}_i u_i] &= \mathbf{0}, \\ y_i &= y_i^* + v_i, \end{aligned}$$

where v_i is a measurement error satisfying $E[\mathbf{x}_i v_i] = 0$ and $E[y_i^* v_i] = 0$. Let $\widehat{\boldsymbol{\beta}}$ denote the OLS coefficient from the regression of y_i on $\mathbf{x}_i$.

[3] Schwarz matrix inequality: For any random $m \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$, $\|\mathbf{X}'\mathbf{Y}\| \leq \|\mathbf{X}\|\,\|\mathbf{Y}\|$. This is a special form of the Cauchy-Schwarz inequality, where the inner product is defined as $\langle\mathbf{X},\mathbf{Y}\rangle = \|\mathbf{X}'\mathbf{Y}\|$.

(i) *Is $\boldsymbol{\beta}$ the coefficient from the linear projection of y_i on $\mathbf{x}_i$?*

(ii) *Is $\widehat{\boldsymbol{\beta}}$ consistent for $\boldsymbol{\beta}$?*

(iii) *Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$.*

1.3 LSE as a MoM Estimator

The LSE is a MoM estimator. The corresponding moment conditions are the orthogonal conditions

$$E\left[\mathbf{x}u\right]=\mathbf{0},$$

where $u=y-\mathbf{x}'\boldsymbol{\beta}$. So the sample analog is the normal equation

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)=\mathbf{0},$$

the solution of which is exactly the LSE. Now, $\mathbf{M}=-E\left[\mathbf{x}_i\mathbf{x}_i'\right]=-\mathbf{Q}$, and $\boldsymbol{\Omega}=E\left[\mathbf{x}_i\mathbf{x}_i'u_i^2\right]$, so

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

the same as in Theorem 3. Note that the asymptotic variance $\mathbf{V}$ takes the *sandwich* form. The larger the $E\left[\mathbf{x}_i\mathbf{x}_i'\right]$, the smaller the $\mathbf{V}$. Although the LSE is a MoM estimator, it is a *special* MoM estimator because it can be treated as a "projection" estimator.

We provide more intuition on the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ below. Consider a simple linear regression model

$$y_i=\beta x_i+u_i,$$

where $E[x_i]$ is normalized to be 0. From introductory econometrics courses,

$$\widehat{\beta}=\frac{\sum\limits_{i=1}^{n}x_iy_i}{\sum\limits_{i=1}^{n}x_i^2}=\frac{\widehat{Cov}(x,y)}{\widehat{Var}(x)},$$

and under homoskedasticity,

$$AVar\left(\widehat{\beta}\right)=\frac{\sigma^2}{nVar(x)}.$$

So the larger the $Var(x)$, the smaller the $AVar\left(\widehat{\beta}\right)$. Actually, $Var(x)=\left|\frac{\partial E[xu]}{\partial\beta}\right|$, so the intuition in introductory courses matches the general results of the MoM estimator.

Similarly, we can derive the asymptotic distribution of the **weighted least squares** (WLS) **estimator**, a special GLS estimator with a diagonal weight matrix. Recall that

$$\widehat{\boldsymbol{\beta}}_{GLS}=(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

which reduces to

$$\widehat{\boldsymbol{\beta}}_{WLS} = \left(\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} w_i \mathbf{x}_i y_i\right)$$

when $\mathbf{W} = \text{diag}\{w_1, \cdots, w_n\}$. Note that this estimator is a MoM estimator under the moment condition

$$E\left[w_i \mathbf{x}_i u_i\right] = \mathbf{0},$$

so

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

where $\mathbf{V}_{\mathbf{W}} = E\left[w_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} E\left[w_i^2 \mathbf{x}_i \mathbf{x}_i' u_i^2\right] E\left[w_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1}$.

Exercise 3 *Suppose* $w_i = \sigma_i^{-2}$, *where* $\sigma_i^2 = E[u_i^2|\mathbf{x}_i]$. *Derive the asymptotic distribution of* $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{WLS} - \boldsymbol{\beta}\right)$.

2 Covariance Matrix Estimators

Since $\mathbf{Q} = E\left[\mathbf{x}_i \mathbf{x}_i'\right]$ and $\boldsymbol{\Omega} = E\left[\mathbf{x}_i \mathbf{x}_i' u_i^2\right]$,

$$\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = \frac{1}{n}\mathbf{X}'\mathbf{X},$$

and

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \widehat{u}_i^2 = \frac{1}{n}\mathbf{X}'\text{diag}\left\{\widehat{u}_1^2, \cdots, \widehat{u}_n^2\right\}\mathbf{X} \equiv \frac{1}{n}\mathbf{X}'\widehat{\mathbf{D}}\mathbf{X} \tag{6}$$

are the MoM estimators for $\mathbf{Q}$ and $\boldsymbol{\Omega}$, where $\{\widehat{u}_i\}_{i=1}^n$ are the OLS residuals. Given that $\mathbf{V} = \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}$, it can be estimated by

$$\widehat{\mathbf{V}} = \widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1},$$

and $AVar(\widehat{\boldsymbol{\beta}})$ is estimated by $\widehat{\mathbf{V}}/n$. As in (5), we can partition $\widehat{\mathbf{Q}}$ and $\widehat{\boldsymbol{\Omega}}$ accordingly and the corresponding notations are just put a hat on. Recall from Exercise 8 of Chapter 3, $Var\left(\widehat{\beta}_j|\mathbf{X}\right) = \sum_{i=1}^n w_{ij}\sigma_i^2/SSR_j$. Since $\widehat{\mathbf{V}}/n = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{D}}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ just replaces σ_i^2 in $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ by $\widehat{u}_i^2$, $\widehat{AVar}\left(\widehat{\beta}_j\right) = \sum_{i=1}^n w_{ij}\widehat{u}_i^2/SSR_j$, $j = 1, \cdots, k$, where $w_{ij} > 0$, $\sum\limits_{i=1}^n w_{ij} = 1$, and SSR_j is the SSR in the regression of x_j on all other regressors.

Although this estimator is natural nowadays, it took some time to come into existence because Ω is usually expressed as $E\left[\mathbf{x}_i \mathbf{x}_i' \sigma_i^2\right]$ whose estimation requires estimating n conditional variances. $\widehat{\mathbf{V}}$ appeared first in the statistical literature Eicker (1967) and Huber (1967), and was introduced into econometrics by White (1980c). So this estimator is often called the "Eicker-Huber-White formula" or something of the kind. Other popular names for this estimator include the "heteroskedasticity-consistent (or robust) convariance matrix estimator" or the "sandwich-form convariance matrix estimator". The following theorem provides a direct proof of the consistency of $\widehat{\mathbf{V}}$ (although it is a trivial corollary of the consistency of the MoM estimator).

Exercise 4 *In the model*

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + u_i, \\ E[\mathbf{x}_i u_i] &= 0, \boldsymbol{\Omega} = E\left[\mathbf{x}_i\mathbf{x}_i'u_i^2\right], \end{aligned}$$

find the MoM estimators of $(\boldsymbol{\beta}, \boldsymbol{\Omega})$.

Theorem 4 *Under the assumptions of Theorem 3,* $\widehat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$.

Proof. From the WLLN, $\widehat{\mathbf{Q}}$ is consistent. As long as we can show $\widehat{\boldsymbol{\Omega}}$ is consistent, by the CMT $\widehat{\mathbf{V}}$ is consistent. Using (3)

$$\begin{aligned} \widehat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\widehat{u}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'u_i^2 - \frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\mathbf{x}_i u_i + \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\mathbf{x}_i\right)^2. \end{aligned}$$

From (4), $E\left[\left\|\mathbf{x}_i\mathbf{x}_i'u_i^2\right\|\right] < \infty$, so by the WLLN, $n^{-1}\sum_{i=1}^n x_i x_i' u_i^2 \xrightarrow{p} \Omega$. We need only to prove the remaining two terms are $o_p(1)$.

Note that the second term satisfies

$$\begin{aligned} \left\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\mathbf{x}_i u_i\right\| &\le \frac{2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i\mathbf{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\mathbf{x}_i u_i\right\| \\ &\le \frac{2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i\mathbf{x}_i'\right\|\left|\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\mathbf{x}_i\right||u_i| \\ &\le \left(\frac{2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i\right\|^3|u_i|\right)\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|, \end{aligned}$$

where the first inequality is from the triangle inequality,[4] and the second and third inequalities are from the Schwarz matrix inequality. By Hölder's inequality,[5]

$$E\left[\left\|\mathbf{x}_i\right\|^3|u_i|\right] \le E\left[\left\|\mathbf{x}_i\right\|^4\right]^{3/4} E\left[|u_i|^4\right]^{1/4} < \infty,$$

so by the WLLN, $n^{-1}\sum_{i=1}^n \left\|\mathbf{x}_i\right\|^3|u_i| \xrightarrow{p} E\left[\left\|\mathbf{x}_i\right\|^3|u_i|\right] < \infty$. Given that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$, the second

[4] Triangle inequality: For any $m \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$, $\|\mathbf{X} + \mathbf{Y}\| \le \|\mathbf{X}\| + \|\mathbf{Y}\|$.

[5] Hölder's inequality: If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any random $m \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$, $E\left[\|\mathbf{X}'\mathbf{Y}\|\right] \le E\left[\|\mathbf{X}\|^p\right]^{1/p} E\left[\|\mathbf{Y}\|^q\right]^{1/q}$.

term is $o_p(1)O_p(1) = o_p(1)$. The third term satisfies

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' \left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\mathbf{x}_i\right)^2 \right\| &\le \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{x}_i\mathbf{x}_i'\right\| \left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\mathbf{x}_i\right)^2 \\ &\le \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{x}_i\right\|^4 \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2 = o_p(1),\end{aligned}$$

where the steps follow from similar arguments as in the second term. ∎

In the homoskedastic case, we can estimate $\mathbf{V}$ by $\widehat{\mathbf{V}}^0 = \widehat{\sigma}^2\widehat{\mathbf{Q}}^{-1}$, and correspondingly, $\widehat{AVar}\left(\widehat{\beta}_j\right) = n^{-1}\sum_{i=1}^{n}\widehat{u}_i^2/SSR_j$. In other words, the weights in the general formula take a special form, $w_{ij} = n^{-1}$. It is hard to judge which formula, homoskedasticity-only or heteroskedasticity-robust, is larger (why?). Although either way is possible in theory, the heteroskedasticity-robust formula is usually larger than the homoskedasticity-only one in practice; Kauermann and Carroll (2001) show that the former is actually more variable than the latter, which is the price paid for robustness.

2.1 Alternative Covariance Matrix Estimators (*)

MacKinnon and White (1985) suggest a small-sample corrected version of $\widehat{\mathbf{V}}$ based on the jackknife principle which is introduced by Quenouille (1949, 1956) and Tukey (1958). Recall in Section 6 of Chapter 3 the definition of $\widehat{\boldsymbol{\beta}}_{(-i)}$ as the least-squares estimator with the i'th observation deleted. From equation (3.13) of Efron (1982), the jackknife estimator of the variance matrix for $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\mathbf{V}}^* = (n-1)\sum_{i=1}^{n}\left(\widehat{\boldsymbol{\beta}}_{(-i)}-\overline{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}}_{(-i)}-\overline{\boldsymbol{\beta}}\right)', \tag{7}$$

where

$$\overline{\boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}_{(-i)}.$$

Using formula (5) in Chapter 3, we can show that

$$\widehat{\mathbf{V}}^* = \left(\frac{n-1}{n}\right)\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}^*\widehat{\mathbf{Q}}^{-1}, \tag{8}$$

where

$$\widehat{\boldsymbol{\Omega}}^* = \frac{1}{n}\sum_{i=1}^{n}(1-h_i)^{-2}\mathbf{x}_i\mathbf{x}_i'\widehat{u}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}(1-h_i)^{-1}\mathbf{x}_i\widehat{u}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}(1-h_i)^{-1}\mathbf{x}_i\widehat{u}_i\right)'$$

and $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$. MacKinnon and White (1985) present numerical (simulation) evidence that $\widehat{\mathbf{V}}^*$ works better than $\widehat{\mathbf{V}}$ as an estimator of $\mathbf{V}$. They also suggest that the scaling factor $(n-1)/n$ in (8) can be omitted.

Exercise 5 *Show that the two expressions of $\widehat{\mathbf{V}}^*$ in (7) and (8) are equal.*

Andrews (1991) suggests a similar estimator based on cross-validation, which is defined by replacing the OLS residual $\widehat{u}_i$ in (6) with the leave-one-out estimator $\widehat{u}_{i,-i} = (1-h_i)^{-1}\widehat{u}_i$. With this substitution, Andrews' proposed estimator is

$$\widehat{\mathbf{V}}^{**} = \widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}^{**}\widehat{\mathbf{Q}}^{-1},$$

where

$$\widehat{\boldsymbol{\Omega}}^{**} = \frac{1}{n}\sum_{i=1}^{n}(1-h_i)^{-2}\mathbf{x}_i\mathbf{x}_i'\widehat{u}_i^2.$$

It is similar to the MacKinnon-White estimator $\widehat{\mathbf{V}}^{*}$, but omits the mean correction. Andrews (1991) argues that simulation evidence indicates that $\widehat{\mathbf{V}}^{**}$ is an improvement on $\widehat{\mathbf{V}}^{*}$.

The jackknife represents a linear approximation of the bootstrap proposed in Efron (1979). See Hall (1994) and Horowitz (2001) for an introduction of the bootstrap in econometrics. Other popular books on the bootstrap and resampling methods include Efron (1982), Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995), and Davison and Hinkley (1997).

3 Restricted Least Squares Revisited (*)

In Chapter 2, we derived the RLS estimator. We now study the asymptotic properties of the RLS estimator. Since the RLS estimator is a special MD estimator, we concentrate on the MD estimator under the constraints $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$ in this section. From Exercise 26 and the discussion in Section 5.1 of Chapter 2, we can show

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

To derive its asymptotic distribution, we impose the following regularity conditions.

Assumption RLS.1: $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$ where $\mathbf{R}$ is $k \times q$ with $\text{rank}(\mathbf{R}) = q$.

Assumption RLS.2: $\mathbf{W}_n \xrightarrow{p} \mathbf{W} > 0$.

Theorem 5 *Under the Assumptions of Theorem 1, RLS.1 and RLS.2,* $\widehat{\boldsymbol{\beta}}_{MD} \xrightarrow{p} \boldsymbol{\beta}$. *Under the Assumptions of Theorem 3, RLS.1 and RLS.2,*

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

where

$$\begin{aligned}\mathbf{V}_{\mathbf{W}} &= \mathbf{V} - \mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V} - \mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}\mathbf{R}'\mathbf{W}^{-1} \\ &\quad + \mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}\mathbf{R}'\mathbf{W}^{-1},\end{aligned}$$

and $\mathbf{V} = \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}$.

From this theorem, the RLS estimator is CAN and its asymptotic variance is $\mathbf{V_Q}$. Unless the model is homoskedastic, it is hard to compare $\mathbf{V_Q}$ and $\mathbf{V}$.

Exercise 6 *Prove the above theorem.*

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{MD}$ depends on $\mathbf{W}$. A natural question is which $\mathbf{W}$ is the best in the sense of minimizing $\mathbf{V_W}$. This turns out to be $\mathbf{V}^{-1}$ as shown in the following theorem. Since $\mathbf{V}^{-1}$ is unknown, we can replace $\mathbf{V}^{-1}$ with a consistent estimate $\widehat{\mathbf{V}}^{-1}$ and the asymptotic distribution (and efficiency) are unchanged. We call the MD estimator setting $\mathbf{W}_n = \widehat{\mathbf{V}}^{-1}$ the **efficient minimum distance** (EMD) **estimator**, which takes the form

$$\widehat{\boldsymbol{\beta}}_{EMD} = \widehat{\boldsymbol{\beta}} - \widehat{\mathbf{V}}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}\right)^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right). \tag{9}$$

Theorem 6 *Under the Assumptions of Theorem 3 and RLS.1,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{EMD} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{V}^*\right),$$

where

$$\mathbf{V}^* = \mathbf{V} - \mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}.$$

Since

$$\mathbf{V}^* \leq \mathbf{V},$$

$\widehat{\boldsymbol{\beta}}_{EMD}$ *has lower asymptotic variance than the unrestricted estimator. Furthermore, for any* $\mathbf{W}$,

$$\mathbf{V}^* \leq \mathbf{V_W},$$

so $\widehat{\boldsymbol{\beta}}_{EMD}$ *is asymptotically efficient in the class of MD estimators.*

Exercise 7 *(i) Show that* $\mathbf{V}_{\mathbf{V}^{-1}} = \mathbf{V}^*$. *(ii*) Show that* $\mathbf{V}^* \leq \mathbf{V_W}$ *for any* $\mathbf{W}$.

Exercise 8 *Consider the exclusion restriction* $\boldsymbol{\beta}_2 = \mathbf{0}$ *in the linear regression model* $y_i = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + u_i$.

(i) *Derive the asymptotic distribution of* $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_{1R}$ *and* $\widehat{\boldsymbol{\beta}}_{1,EMD}$ *and show that* $AVar(\widehat{\boldsymbol{\beta}}_{1,EMD}) \leq AVar(\widehat{\boldsymbol{\beta}}_1)$ *and* $AVar(\widehat{\boldsymbol{\beta}}_{1,EMD}) \leq AVar(\widehat{\boldsymbol{\beta}}_{1R})$.

(ii) *When the model is homoskedastic, show that* $AVar(\widehat{\boldsymbol{\beta}}_{1,EMD}) = AVar(\widehat{\boldsymbol{\beta}}_{1R}) \leq AVar(\widehat{\boldsymbol{\beta}}_1)$. *When will the equality hold? (Hint:* $\mathbf{Q}_{12} = \mathbf{0}$*)*

(iii) *When the model is heteroskedastic, provide an example where* $AVar(\widehat{\boldsymbol{\beta}}_{1R}) > AVar(\widehat{\boldsymbol{\beta}}_1)$.

This theorem shows that the MD estimator with the smallest asymptotic variance is $\widehat{\boldsymbol{\beta}}_{EMD}$. One implication is that the RLS estimator is generally inefficient. The interesting exception is the case of conditional homoskedasticity, in which the optimal weight matrix is $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{0-1}$ and thus the RLS estimator is an efficient MD estimator. When the error is conditionally heteroskedastic, there are asymptotic efficiency gains by using minimum distance rather than least squares.

The fact that the RLS estimator is generally inefficient appears 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 the above theorem shows that this is not the efficient estimation method. Instead, the EMD estimator has a smaller asymptotic variance. Why? Consider the RLS estimator with the exclusion restrictions. In this case, the least squares estimation does not make use of the regressor $\mathbf{x}_{2i}$. It ignores the information $E\left[\mathbf{x}_{2i}u_i\right] = \mathbf{0}$. This information is relevant when the error is heteroskedastic and the excluded regressors are correlated with the included regressors.

Finally, note that all asymptotic variances can be consistently estimated by their sample analogs, e.g.,

$$\widehat{\mathbf{V}}^* = \widehat{\mathbf{V}} - \widehat{\mathbf{V}}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{V}},$$

where $\widehat{\mathbf{V}}$ is a consistent estimator of $\mathbf{V}$.

3.1 Orthogonality of Efficient Estimators

One important property of an efficient estimator is the following orthogonality property popularized by Hausman (1978).[6]

Theorem 7 (Orthogonality of Efficient Estimators) *Let $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ be two CAN estimators of $\boldsymbol{\beta}$, and $\widetilde{\boldsymbol{\beta}}$ is efficient. Then the limiting distributions of $\sqrt{n}\widehat{\Delta}$ and $\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ have zero covariance, where $\widehat{\Delta} \equiv \widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$.*

Proof. Suppose $\widehat{\Delta}$ and $\widetilde{\boldsymbol{\beta}}$ are not orthogonal. Since $\text{plim}\left(\widehat{\Delta}\right) = \mathbf{0}$, consider a new estimator $\overline{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} + a\mathbf{A}\widehat{\Delta}$, where a is a scalar and $\mathbf{A}$ is an arbitrary matrix to be chosen. The new estimator is CAN with asymptotic variance

$$Var(\overline{\boldsymbol{\beta}}) = Var(\widetilde{\boldsymbol{\beta}}) + a\mathbf{A}Cov(\widetilde{\boldsymbol{\beta}}, \widehat{\Delta}) + aCov(\widetilde{\boldsymbol{\beta}}, \widehat{\Delta})'\mathbf{A}' + a^2\mathbf{A}Var(\widehat{\Delta})\mathbf{A}'.$$

Since $\widetilde{\boldsymbol{\beta}}$ is efficient, the minimizer of $Var(\overline{\boldsymbol{\beta}})$ with respect to a is achieved at $a = 0$. Taking the first-order derivative with respect to a yields

$$\mathbf{AC} + \mathbf{C}'\mathbf{A}' + 2a\mathbf{A}Var(\widehat{\Delta})\mathbf{A}',$$

where $\mathbf{C} \equiv Cov(\widetilde{\boldsymbol{\beta}}, \widehat{\Delta})$. Choosing $\mathbf{A} = -\mathbf{C}'$, we have

$$-2\mathbf{C}'\mathbf{C} + 2a\mathbf{C}'Var(\widehat{\Delta})\mathbf{C}.$$

[6] See also Lehmann and Casella (1998, Theorem 1.7, p85) and Rao (1973, Theorem 51.2.(i), p317) for a similar result. Lehmann and Casella cite Barankin (1949), Stein (1950), Bahadur (1957), and Luenberger (1969) as early references.

Therefore, at $a = 0$, this derivative is equal to $-2\mathbf{C}'\mathbf{C} \leq 0$. Unless $\mathbf{C} = \mathbf{0}$, we may have a better estimator than $\widetilde{\boldsymbol{\beta}}$ by choosing a a little bit deviating from 0. ■

Exercise 9 (Finite-Sample Orthogonality of Efficient Estimators) *In the homoskedastic linear regression model, check* $Cov\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_R, \widehat{\boldsymbol{\beta}}_R \middle| \mathbf{X}\right) = \mathbf{0}$ *directly. (Hint: recall that* $\widehat{\boldsymbol{\beta}}_R = \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}\mathbf{S}}\widehat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}\mathbf{S}})\mathbf{s}$.*)*

A direct corollary of the above theorem is as follows.

Corollary 1 *Let* $\widehat{\boldsymbol{\beta}}$ *and* $\widetilde{\boldsymbol{\beta}}$ *be two CAN estimators of* $\boldsymbol{\beta}$*, and* $\widetilde{\boldsymbol{\beta}}$ *is efficient. Then* $AVar(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) = AVar(\widehat{\boldsymbol{\beta}}) - AVar(\widetilde{\boldsymbol{\beta}})$.

3.2 Misspecification

We usually have two methods to study the effects of misspecification of the restrictions. In Method I, we assume the truth is $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}^*$ with $\mathbf{c}^* \neq \mathbf{c}$; in Method II, we assume the truth is $\mathbf{R}'\boldsymbol{\beta}_n = \mathbf{c} + n^{-1/2}\boldsymbol{\delta}$. The specification in Method II need some explanation. In this specification, the constraint is "close" to correct, as the difference $\mathbf{R}'\boldsymbol{\beta}_n - \mathbf{c} = n^{-1/2}\boldsymbol{\delta}$ is "small" in the sense that it decreases with sample size n. We call such a misspecification as **local misspecification**. The reason why the deviation is proportional to $n^{-1/2}$ is because this is the only choice under which the localizing parameter $\boldsymbol{\delta}$ appears in the asymptotic distribution but does not dominate it. We will give more discussions on this choice of rate in studying the local power of tests.

First discuss Method I. From the expression of $\widehat{\boldsymbol{\beta}}_{MD}$, it is not hard to see that

$$\widehat{\boldsymbol{\beta}}_{MD} \xrightarrow{p} \boldsymbol{\beta}^*_{MD} = \boldsymbol{\beta} - \mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}(\mathbf{c}^* - \mathbf{c}).$$

The second term, $\mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}(\mathbf{c}^* - \mathbf{c})$, shows that imposing an incorrect constraint leads to inconsistency - an asymptotic bias. We call the limiting value $\boldsymbol{\beta}^*_{MD}$ the minimum-distance projection coefficient or the pseudo-true value implied by the restriction. There are more to say. Define

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

(Note that $\boldsymbol{\beta}^*_n$ is different from $\boldsymbol{\beta}^*_{MD}$.) Then

$$\begin{aligned} \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{MD} - \boldsymbol{\beta}^*_n\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \mathbf{W}_n^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}_n^{-1}\mathbf{R})^{-1}\sqrt{n}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}^*\right) \qquad (10) \\ &= \left(\mathbf{I} - \mathbf{W}_n^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}_n^{-1}\mathbf{R})^{-1}\mathbf{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &\xrightarrow{d} \left(\mathbf{I} - \mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}\mathbf{R}'\right)N(\mathbf{0}, \mathbf{V}) \\ &= N(\mathbf{0}, \mathbf{V}_{\mathbf{W}}). \end{aligned}$$

In particular,

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

This means that even when the constraint $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$ is misspecified, the conventional covariance matrix estimator is still an appropriate measure of the sampling variance, though the distribution is centered at the pseudo-true values (or projections) $\boldsymbol{\beta}_n^*$ rather than $\boldsymbol{\beta}$. The fact that the estimators are biased is an unavoidable consequence of misspecification.

In Method II, since the true model is $y_i = \mathbf{x}_i'\boldsymbol{\beta}_n + u_i$, it is not hard to show that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}) \tag{11}$$

which is the same as when $\boldsymbol{\beta}$ is fixed. A difference arises in the constrained estimator. Since $\mathbf{c} = \mathbf{R}'\boldsymbol{\beta}_n - n^{-1/2}\boldsymbol{\delta}$,

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

and

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

It follows that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{MD} - \boldsymbol{\beta}_n\right) \xrightarrow{d} \left(\mathbf{I} - \mathbf{W}_n^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}_n^{-1}\mathbf{R})^{-1}\mathbf{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) + \mathbf{W}_n^{-1}\mathbf{R}\left(\mathbf{R}'\mathbf{W}_n^{-1}\mathbf{R}\right)^{-1}\boldsymbol{\delta}.$$

The first term is asymptotically normal by (11). The second term converges in probability to a constant. This is because the $n^{-1/2}$ local scaling 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

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{MD} - \boldsymbol{\beta}_n\right) \xrightarrow{d} N\left(0, \mathbf{V}_{\mathbf{W}}\right) + \mathbf{W}^{-1}\mathbf{R}\left(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R}\right)^{-1}\boldsymbol{\delta} = N\left(\boldsymbol{\delta}^*, \mathbf{V}_{\mathbf{W}}\right), \tag{12}$$

where

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

The asymptotic distribution (12) is an approximation of the sampling distribution of the restricted estimator under misspecification. The distribution (12) contains an asymptotic bias component $\boldsymbol{\delta}^*$. The approximation is not fundamentally different from (10) - they both have the same asymptotic variances, and both reflect the bias due to misspecification. The difference is that (10) puts the bias on the left-side of the convergence arrow, while (12) has the bias on the right-side. There is no substantive difference between the two, but (12) is more convenient for some purposes, such as the analysis of the power of tests, as we will explore in the last section of this chapter.

3.3 Nonlinear and Inequality Restrictions

In some cases it is desirable to impose nonlinear constraints on the parameter vector $\boldsymbol{\beta}$. They can be written as

$$\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}, \tag{13}$$

where $\mathbf{r}$: $\mathbb{R}^k \to \mathbb{R}^q$. This includes the linear constraints as a special case. An example of (13) which cannot be written in a linear form is $\beta_1\beta_2 = 1$, which is (13) with $r(\boldsymbol{\beta}) = \beta_1\beta_2 - 1$.

The RLS and MD estimators of $\boldsymbol{\beta}$ subject to (13) solve the minimization problems

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_R &= \arg \min_{\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}} SSR(\boldsymbol{\beta}), \\ \widehat{\boldsymbol{\beta}}_{MD} &= \arg \min_{\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}} J_n(\boldsymbol{\beta}). \end{aligned}$$

The solutions can be achieved by the Lagrangian method. Computationally, there is in general no explicit expression for the solutions so they must be found numerically. Algorithms to numerically solve such Lagrangian problems are known as *constrained optimization* methods, and are available in programming languages such as Matlab.

The asymptotic distributions of $\widehat{\boldsymbol{\beta}}_R$ and $\widehat{\boldsymbol{\beta}}_{MD}$ are the same as in the linear constraints case except that $\mathbf{R}$ is replaced by $\partial\mathbf{r}(\boldsymbol{\beta})'/\partial\boldsymbol{\beta}$, but the proof is more delicate.

We sometimes impose inequality constraints on $\boldsymbol{\beta}$,

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

The most common example is a non-negative constraint $\beta_1 \geq 0$. The RLS and MD estimators of $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_R &= \arg \min_{\mathbf{r}(\boldsymbol{\beta})\geq\mathbf{0}} SSR(\boldsymbol{\beta}), \\ \widehat{\boldsymbol{\beta}}_{MD} &= \arg \min_{\mathbf{r}(\boldsymbol{\beta})\geq\mathbf{0}} J_n(\boldsymbol{\beta}). \end{aligned}$$

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 a two-step approach. First compute the unconstrained estimator $\widehat{\boldsymbol{\beta}}$. If $\widehat{\beta}_1 \geq 0$ then $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MD} = \widehat{\boldsymbol{\beta}}$. Second, if $\widehat{\beta}_1 < 0$ 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 with inequality constraints are examples of *quadratic programming* problems. Quick and easy computer algorithms are available in programming languages such as Matlab.

Exercise 10 (Ridge Regression) *Suppose the nonlinear constraints are* $\boldsymbol{\beta}'\boldsymbol{\Lambda}\boldsymbol{\beta} \leq B$, *where* $\boldsymbol{\Lambda} >$

0 *and* $B>0$. *Show that*

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

where $\widehat{\lambda}$ *is the Lagrange multiplier for the constraint* $\boldsymbol{\beta}'\boldsymbol{\Lambda}\boldsymbol{\beta}\leq B$.

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 $\mathbf{r}(\boldsymbol{\beta})>\mathbf{0}$, then asymptotically the estimator is not subject to the constraint and the inequality-constrained estimator has an asymptotic distribution equal to the unconstrained one. However if the true parameter is on the boundary, e.g., $\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}$, 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 $\sqrt{n}\left(\widehat{\beta}-\beta\right)\xrightarrow{d}Z=N(0,V)$ and $\beta=0$, then the constrained estimator $\widehat{\beta}_R(=\widehat{\beta}_{MD})=\max\left\{\widehat{\beta},0\right\}$ will have the asymptotic distribution $\sqrt{n}\widehat{\beta}_R\xrightarrow{d}\max\{Z,0\}$, a "half-normal" distribution.

4 Functions of Parameters

Sometimes we are interested in some lower-dimensional function of the parameter vector $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_k)'$. For example, we may be interested in a single coefficient β_j or a ratio β_j/β_l. In these cases we can write the parameter of interest as a function of $\boldsymbol{\beta}$. Let $\mathbf{r}$: $\mathbb{R}^k\to\mathbb{R}^q$ denote this function and let

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

denote the parameter of interest. A natural estimate of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}}=\mathbf{r}\left(\widehat{\boldsymbol{\beta}}\right)$. To derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$, we impose the following assumption.

Assumption RLS.1′: $\mathbf{r}(\cdot)$ is continuously differentiable at the true value $\boldsymbol{\beta}$ and $\mathbf{R}=\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{r}(\boldsymbol{\beta})'$ has rank q.

This assumption is an extension of Assumption RLS1.

Theorem 8 *Under the assumptions of Theorem 3 and Assumption RLS.1′,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d}N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\theta}}\right),$$

where $\mathbf{V}_{\boldsymbol{\theta}}=\mathbf{R}'\mathbf{V}\mathbf{R}$.

Proof. By the CMT, $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$. By the Delta method, if $\mathbf{r}(\cdot)$ is differentiable at the true value $\boldsymbol{\beta}$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=\sqrt{n}\left(\mathbf{r}\left(\widehat{\boldsymbol{\beta}}\right)-\mathbf{r}\left(\boldsymbol{\beta}\right)\right)\xrightarrow{d}\mathbf{R}'N\left(\mathbf{0},\mathbf{V}\right)=N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\theta}}\right)$$

where $\mathbf{V}_{\boldsymbol{\theta}}=\mathbf{R}'\mathbf{V}\mathbf{R}>0$ if $\mathbf{R}$ has full rank q. ■

A natural estimator of $\mathbf{V}_{\boldsymbol{\theta}}$ is

$$\widehat{\mathbf{V}}_{\boldsymbol{\theta}}=\widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{R}},\tag{14}$$

where $\widehat{\mathbf{R}} = \partial \mathbf{r}\left(\widehat{\boldsymbol{\beta}}\right)' / \partial \boldsymbol{\beta}$. If $\mathbf{r}(\cdot)$ is a $C^{(1)}$ function, then by the CMT, $\widehat{\mathbf{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \mathbf{V}_{\boldsymbol{\theta}}$ (why?).

In many cases, the function $\mathbf{r}(\boldsymbol{\beta})$ is linear:

$$\mathbf{r}(\boldsymbol{\beta}) = \mathbf{R}'\boldsymbol{\beta}$$

for some $k \times q$ matrix $\mathbf{R}$. In this case, $\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\boldsymbol{\beta})' = \mathbf{R}$ and $\widehat{\mathbf{R}} = \mathbf{R}$, so $\widehat{\mathbf{V}}_{\boldsymbol{\theta}} = \mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}$. For example, if $\mathbf{R}$ is a "selector matrix"

$$\mathbf{R} = \begin{pmatrix} \mathbf{I}_{q \times q} \\ \mathbf{0}_{(k-q) \times q} \end{pmatrix},$$

so that if $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, then $\boldsymbol{\theta} = \mathbf{R}'\boldsymbol{\beta} = \boldsymbol{\beta}_1$ and

$$\widehat{\mathbf{V}}_{\boldsymbol{\theta}} = (\mathbf{I}, \mathbf{0})\, \widehat{\mathbf{V}} \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} = \widehat{\mathbf{V}}_{11},$$

the upper-left block of $\widehat{\mathbf{V}}$. When $q = 1$ (so $r(\boldsymbol{\beta})$ is real-valued), the standard error for $\widehat{\theta}$ is the square root of $n^{-1}\widehat{V}_{\theta}$, that is, $s\left(\widehat{\theta}\right) = n^{-1/2}\sqrt{\widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{R}}}$.

5 The t Test

Let $\theta = r(\boldsymbol{\beta})$: $\mathbb{R}^k \to \mathbb{R}$ be any parameter of interest (for example, θ could be a single element of $\boldsymbol{\beta}$), $\widehat{\theta}$ its estimate and $s\left(\widehat{\theta}\right)$ its asymptotic standard error. Consider the studentized statistic

$$t_n(\theta) = \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)}.$$

Since $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N(0, V_{\theta})$ and $\sqrt{n}s\left(\widehat{\theta}\right) \xrightarrow{p} \sqrt{V_{\theta}}$, by Slutsky's theorem, we have

Theorem 9 *Under the assumptions of Theorem 8,* $t_n(\theta) \xrightarrow{d} N(0, 1)$.

Thus the asymptotic distribution of the t-ratio $t_n(\theta)$ is the standard normal. Since the standard normal distribution does not depend on the parameters, we say that $t_n(\theta)$ is **asymptotically pivotal**. In special cases (such as the normal regression model), the statistic t_n has an exact t distribution, and is therefore exactly free of unknowns. In this case, we say that t_n is an **exactly pivotal** statistic. In general, however, pivotal statistics are unavailable and so we must rely on asymptotically pivotal statistics.

The most common one-dimensional hypotheses are the null

$$H_0 : \theta = \theta_0, \tag{15}$$

against the alternative

$$H_1 : \theta \neq \theta_0, \tag{16}$$

where θ_0 is some pre-specified value. The standard test for H_0 against H_1 is based on the absolute value of the t-statistic,

$$t_n = t_n(\theta_0) = \frac{\widehat{\theta} - \theta_0}{s\left(\widehat{\theta}\right)}.$$

Under H_0, $t_n \xrightarrow{d} Z \sim N(0,1)$, so $|t_n| \xrightarrow{d} |Z|$ by the CMT. $G(u) = P(|Z| \leq u) = \Phi(u) - (1 - \Phi(u)) = 2\Phi(u) - 1 \equiv \overline{\Phi}(u)$ is called the **asymptotic null distribution**.

The **asymptotic size** of the test is defined as the asymptotic probability of a Type I error:

$$\lim_{n\to\infty} P(|t_n| > c | H_0 \text{ true}) = P(|Z| > c) = 1 - \overline{\Phi}(c).$$

We see that the asymptotic size of the test is a simple function of the asymptotic null distribution G and the critical value c. As mentioned in Chapter 3, in the dominant approach to hypothesis testing, the researcher pre-selects a significance level $\alpha \in (0,1)$ and then selects c so that the (asymptotic) size is no larger than α. We call c the **asymptotic critical value** because it has been selected from the asymptotic null distribution. Let $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of the standard normal distribution. That is, if $Z \sim N(0,1)$, then $P(Z > z_{\alpha/2}) = \alpha/2$ and $P(|Z| > z_{\alpha/2}) = \alpha$. For example, $z_{.025} = 1.96$ and $z_{.05} = 1.645$. A test of asymptotic significance α rejects H_0 if $|t_n| > z_{\alpha/2}$. Otherwise the test does not reject, or "accepts" H_0.

The alternative hypothesis (16) is sometimes called a “two-sided” alternative. Sometimes we are interested in testing for one-sided alternatives such as

$$H_1\colon\ \theta > \theta_0 \tag{17}$$

or

$$H_1\colon\ \theta < \theta_0. \tag{18}$$

Tests of (15) against (17) or (18) are based on the signed t-statistic t_n. The hypothesis (15) is rejected in favor of (17) if $t_n > c$ where c satisfies $\alpha = 1 - \Phi(c)$. Negative values of t_n are not taken as evidence against H_0, as point estimates $\widehat{\theta}$ less than θ_0 do not point to (17). Since the critical values are taken from the single tail of the normal distribution, they are smaller than for two-sided tests. Specifically, the asymptotic 5% critical value is $c = 1.645$. Thus, we reject (15) in favor of (17) if $t_n > 1.645$. Testing against (18) can be conducted in a similar way.

There seems to be an ambiguity. Should we use the two-sided critical value 1.96 or the one-sided critical value 1.645? The answer is that we should use one-sided tests and critical values only when the parameter space is known to satisfy a one-sided restriction such as $\theta \geq \theta_0$. This is when the test of (15) against (17) makes sense. If the restriction $\theta \geq \theta_0$ is not known a priori, then imposing this restriction to test (15) against (17) does not make sense. Since linear regression coefficients typically do not have a priori sign restrictions, we conclude that two-sided tests are generally appropriate.

Exercise 11 *Prove that if an additional regressor $\mathbf{X}_{k+1}$ is added to $\mathbf{X}$, Theil's adjusted $\overline{R}^2$ increases* *if and only if* $|t_{k+1}| > 1$*, where* $t_{k+1} = \widehat{\beta}_{k+1}/s\left(\widehat{\beta}_{k+1}\right)$ *is the t-ratio for* $\widehat{\beta}_{k+1}$ *and*

$$s\left(\widehat{\beta}_{k+1}\right) = \left(s^2\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]_{k+1,k+1}\right)^{1/2}$$

is the homoskedasticity-formula standard error. (Hint: Use the FWL theorem)

6 p-Value

An alternative approach, associated with R.A. Fisher, is to report an asymptotic p-value. The asymptotic p-value for $|t_n|$ is constructed as follows. Define the tail probability, or asymptotic p-value function

$$p(t) = P\left(|Z| > t\right) = 1 - G(t) = 2\left(1 - \Phi(t)\right),$$

where $G(\cdot)$ is the cdf of $|Z|$. Then the asymptotic p-value of the statistic $|t_n|$ is

$$p_n = p(|t_n|).$$

So the p-value is the probability of obtaining a test statistic result at least as extreme as the one that was actually observed or the smallest significance level at which the null would be rejected, assuming that the null is true. Since the distribution function G is monotonically increasing, the p-value is a monotonically decreasing function of t_n and is an equivalent test statistic. Figure 1 shows how to find p_n when $|t_n| = 1.85$ (the left panel) and p_n as a function of $|t_n|$ (the right panel). An important caveat is that the p-value p_n should not be interpreted as the probability that either hypothesis is true. For example, a common mis-interpretation is that p_n is the probability "that the null hypothesis is false." This is incorrect. Rather, p_n is a measure of the strength of information against the null hypothesis.

A researcher will often "reject the null hypothesis" when the p-value turns out to be less than a predetermined significance level, often 0.05 or 0.01. Such a result indicates that the observed result would be highly unlikely under the null hypothesis. In a sense, p-values and hypothesis tests are equivalent since $p_n < \alpha$ if and only if $|t_n| > z_{\alpha/2}$. Thus an equivalent statement of a Neyman-Pearson test is to reject at the α level if and only if $p_n < \alpha$. The p-value is more general, however, in that the reader is allowed to pick the level of significance α, in contrast to Neyman-Pearson rejection/acceptance reporting where the researcher picks the level.

Another helpful observation is that the p-value function has simply made a unit-free transformation of the test statistic. That is, under H_0, $p_n \xrightarrow{d} U[0,1]$, so the "unusualness" of the test statistic can be compared to the easy-to-understand uniform distribution, regardless of the complication of the distribution of the original test statistic. To see this fact, note that the asymptotic distribution

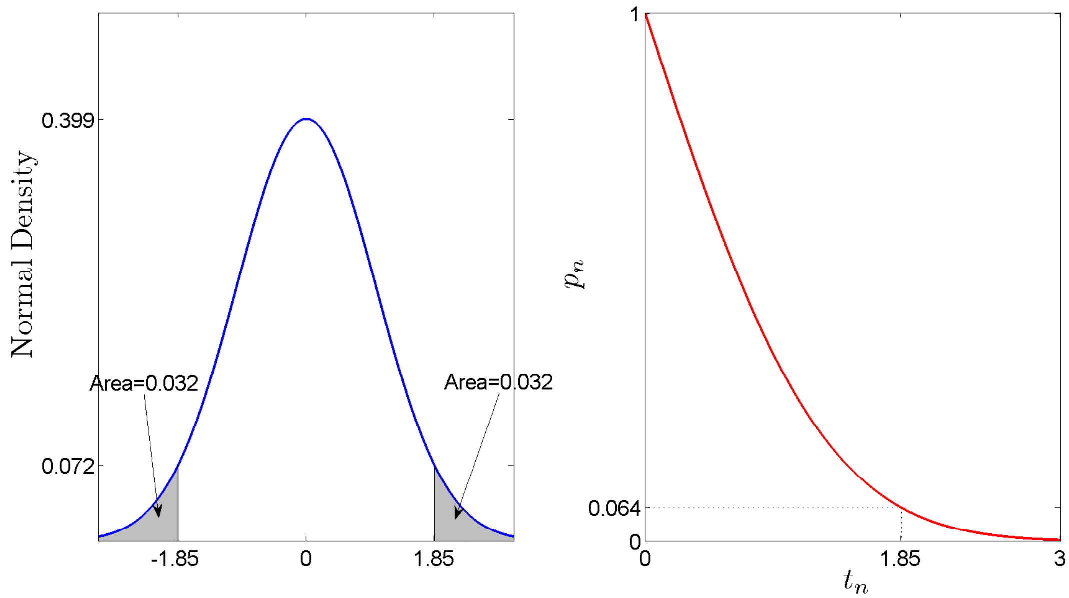

Figure 1: Obtaining the p-Value in a Two-Sided t-Test

of $|t_n|$ is $G(x) = 1 - p(x)$. Thus

$$\begin{aligned} P(1 - p_n \leq u) &= P(1 - p(|t_n|) \leq u) = P(G(|t_n|) \leq u) \\ &= P\left(|t_n| \leq G^{-1}(u)\right) \to G(G^{-1}(u)) = u, \end{aligned}$$

establishing that $1 - p_n \xrightarrow{d} U[0,1]$, from which it follows that $p_n \xrightarrow{d} U[0,1]$.

7 Confidence Interval

A confidence interval (CI) C_n is an interval estimate of $\theta \in \mathbb{R}$ which is assumed to be *fixed*. It is a function of the data and hence is random. So it is not correct to say that "θ will *fall* in C_n with high probability", rather, C_n is designed to *cover* θ with high probability. Either $\theta \in C_n$ or $\theta \notin C_n$. The coverage probability is $P(\theta \in C_n)$.

We typically cannot calculate the exact coverage probability $P(\theta \in C_n)$. However we often can calculate the asymptotic coverage probability $\lim_{n\to\infty} P(\theta \in C_n)$. We say that C_n has asymptotic $(1-\alpha)$ coverage for θ if $P(\theta \in C_n) \to 1 - \alpha$ as $n \to \infty$.

A good method for constructing a confidence interval is collecting parameter values which are not rejected by a statistical test, so-called "*test statistic inversion*" method. The t-test in Section 5 rejects H_0: $\theta = \theta_0$ if $|t_n(\theta_0)| > z_{\alpha/2}$. A confidence interval is then constructed using the values

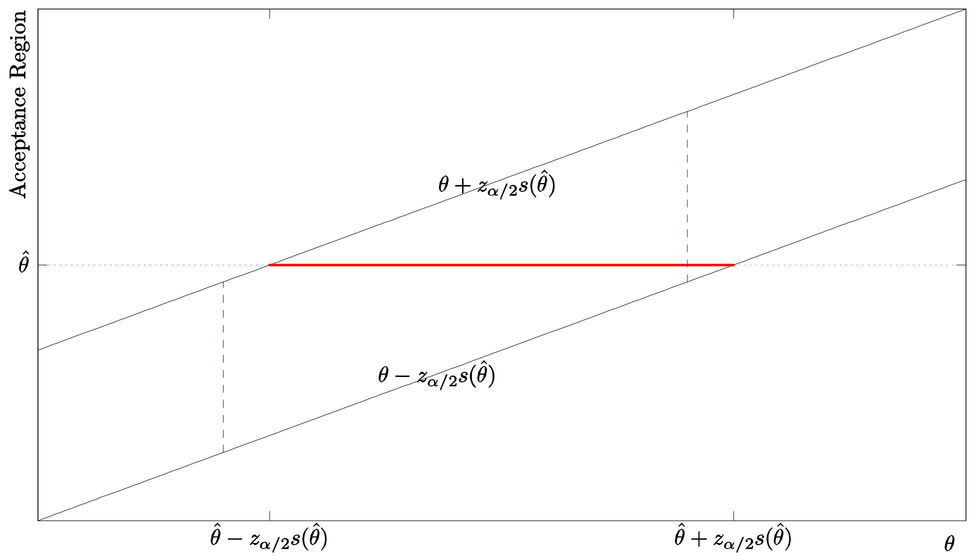

Figure 2: Test Statistic Inversion

for which this test does not reject:

$$C_n = \left\{\theta \middle| \left|t_n(\theta)\right| \le z_{\alpha/2}\right\} = \left\{\theta \middle| -z_{\alpha/2} \le \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)} \le z_{\alpha/2}\right\} = \left[\widehat{\theta} - z_{\alpha/2}s\left(\widehat{\theta}\right), \widehat{\theta} + z_{\alpha/2}s\left(\widehat{\theta}\right)\right].$$

Figure 2 illustrates the idea of inverting a test statistic. In Figure 2, the acceptance region for $\widehat{\theta}$ at θ is $\left[\theta - z_{\alpha/2}s\left(\widehat{\theta}\right), \theta + z_{\alpha/2}s\left(\widehat{\theta}\right)\right]$, which is the region for $\widehat{\theta}$ such that the hypothesis that the true value is θ cannot be rejected or is "accepted".

While there is no hard-and-fast guideline for choosing the coverage probability $1-\alpha$, the most common professional choice is 95%, or $\alpha = .05$. This corresponds to selecting the confidence interval $\left[\widehat{\theta} \pm 1.96s\left(\widehat{\theta}\right)\right] \approx \left[\widehat{\theta} \pm 2s\left(\widehat{\theta}\right)\right]$. Thus values of θ within two standard errors of the estimated $\widehat{\theta}$ are considered "reasonable" candidates for the true value θ, and values of θ outside two standard errors of the estimated $\widehat{\theta}$ are considered unlikely or unreasonable candidates for the true value.

Finally, the interval has been constructed so that as $n \to \infty$,

$$P\left(\theta \in C_n\right) = P\left(\left|t_n(\theta)\right| \le z_{\alpha/2}\right) \to P\left(|Z| \le z_{\alpha/2}\right) = 1 - \alpha,$$

so C_n is indeed an asymptotic $(1-\alpha)$ confidence interval.

(*) Coverage accuracy is a basic requirement for a CI. Another property of a CI is its length (if it is fixed) or expected length (if it is random). Since at most one point of the interval is the true value, the expected length of the interval is a measure of the "average extent" of the false values included In the technical appendix, we discuss this issue following Pratt (1961).

8 The Wald Test

Sometimes $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta})$ is a $q \times 1$ vector, and it is desired to test the joint restrictions simultaneously. In this case the t-statistic approach does not work. We have the null and alternative

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ vs } H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$[7]

The natural estimate of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = \mathbf{r}\left(\widehat{\boldsymbol{\beta}}\right)$. Suppose $\widehat{\mathbf{V}}_{\boldsymbol{\theta}}$ is an estimate of the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$, e.g., $\widehat{\mathbf{V}}_{\boldsymbol{\theta}}$ in (14); then the Wald statistic for H_0 against H_1 is

$$W_n = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

We have known that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}}\right)$, and $\widehat{\mathbf{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \mathbf{V}_{\boldsymbol{\theta}}$ under the null. So by Example 2 of Chapter 4, $W_n \xrightarrow{d} \chi_q^2$ under the null. We have established:

Theorem 10 *Under the assumptions of Theorem 8, $W_n \xrightarrow{d} \chi_q^2$ under H_0.*

When $\mathbf{r}$ is a linear function of $\boldsymbol{\beta}$, i.e., $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{R}'\boldsymbol{\beta}$, the Wald statistic takes the form

$$W_n = n\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}\right)^{-1} \left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right).$$

When $q = 1$, $W_n = t_n^2$. Correspondingly, the asymptotic distribution $\chi_1^2 = N(0,1)^2$.

An asymptotic Wald test rejects H_0 in favor of H_1 if W_n exceeds $\chi_{q,\alpha}^2$, the upper-α quantile of the χ_q^2 distribution. For example, $\chi_{1,.05}^2 = 3.84 = z_{.025}^2$. The Wald test fails to reject if W_n is less than $\chi_{q,\alpha}^2$. The asymptotic p-value for W_n is $p_n = p(W_n)$, where $p(x) = P(\chi_q^2 \geq x)$ is the tail probability function of the χ_q^2 distribution. As before, the test rejects at the α level if $p_n < \alpha$, and p_n is asymptotically $U[0,1]$ under H_0.

9 Confidence Region

Similarly, we can construct confidence regions for multiple parameters, e.g., $\boldsymbol{\theta} = r(\boldsymbol{\beta}) \in \mathbb{R}^q$. By the test statistic inversion method, an asymptotic $(1-\alpha)$ confidence region for $\boldsymbol{\theta}$ is

$$\mathbf{C}_n = \left\{\boldsymbol{\theta} | W_n(\boldsymbol{\theta}) \leq \chi_q^2(\alpha)\right\},$$

where $W_n(\boldsymbol{\theta}) = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \widehat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$. Since $\widehat{\mathbf{V}}_{\boldsymbol{\theta}} > 0$, $\mathbf{C}_n$ is an ellipsoid in the $\boldsymbol{\theta}$ plane.

To show $\mathbf{C}_n$ intuitively, assume $q = 2$ and $\boldsymbol{\theta} = (\beta_1, \beta_2)'$. In this case, $\mathbf{C}_n$ is an ellipse in the (β_1, β_2) plane as shown in Figure 3. In Figure 3, we also show the $(1-\alpha)$ CIs for β_1 and β_2. It is tempting to use the rectangular region, say $\mathbf{C}_n'$, as a confidence region for (β_1, β_2). However,

[7] Different from t-tests, Wald tests are hard to apply for one-sided alernatives.

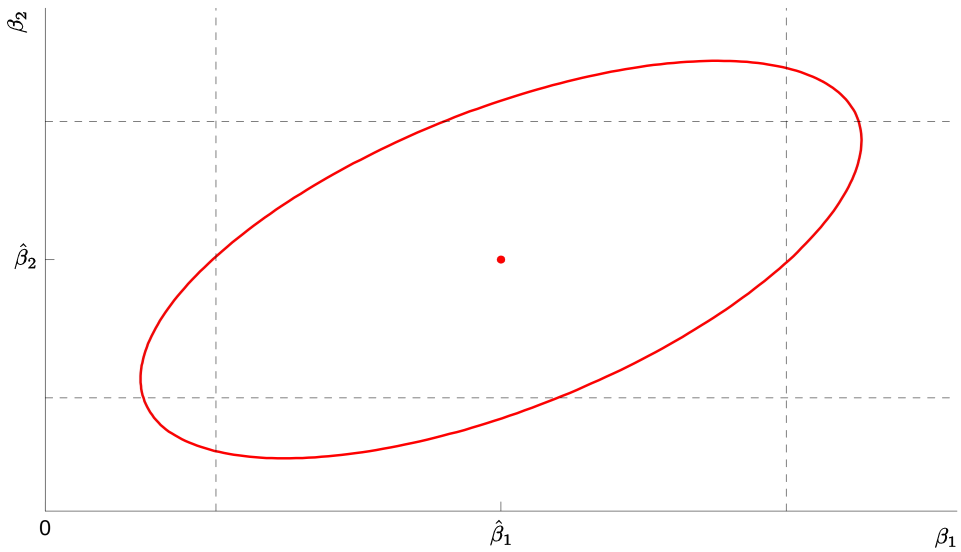

Figure 3: Confidence Region for (β_1, β_2)

$P((\beta_1, \beta_2) \in \mathbf{C}'_n)$ may not converge to $1-\alpha$. For example, suppose $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are asymptotically independent; then $P((\beta_1, \beta_2) \in \mathbf{C}'_n) \to (1-\alpha)^2 < (1-\alpha)$.

Exercise 12 *Show that when $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are asymptotically independent, $P((\beta_1, \beta_2) \in \mathbf{C}'_n) \to (1-\alpha)^2$.*

10 Problems with Tests of Nonlinear Hypotheses

While the t test and Wald tests work well when the hypothesis is a linear restriction on $\boldsymbol{\beta}$, they can work quite poorly when the restrictions are nonlinear. This can be seen in a simple example introduced by Lafontaine and White (1986). Take the model

$$y_i = \beta + u_i, u_i \sim N(0, \sigma^2)$$

and consider the hypothesis

$$H_0\colon\ \beta = 1.$$

Let $\widehat{\beta}$ and $\widehat{\sigma}^2$ be the sample mean and variance of y_i. The standard Wald test for H_0 is

$$W_n = n\frac{\left(\widehat{\beta} - 1\right)^2}{\widehat{\sigma}^2}.$$

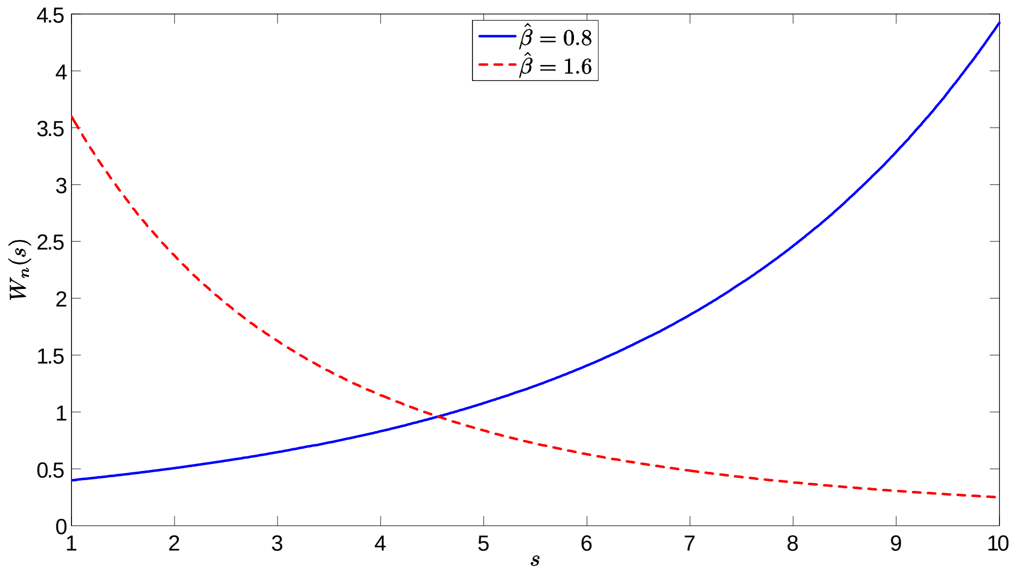

Figure 4: Wald Statistic as a function of s

Now notice that H_0 is equivalent to the hypothesis

$$H_0(s)\colon\ \beta^s = 1,$$

for any positive integer s. Letting $r(\beta) = \beta^s$, and noting $R = s\beta^{s-1}$, we find that the standard Wald test for $H_0(s)$ is

$$W_n(s) = n\frac{\left(\widehat{\beta}^s - 1\right)^2}{\widehat{\sigma}^2 s^2 \widehat{\beta}^{2s-2}}.$$

While the hypothesis $\beta^s = 1$ is unaffected by the choice of s, the statistic $W_n(s)$ varies with s. This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we plot in Figure 4 the Wald statistic $W_n(s)$ as a function of s, setting $n/\widehat{\sigma}^2 = 10$. The increasing solid line is for the case $\widehat{\beta} = 0.8$ and the decreasing dashed line is for the case $\widehat{\beta} = 1.6$. It is easy to see that in each case there are values of s for which the test statistic is significant relative to asymptotic critical values, while there are other values of s for which the test statistic is insignificant.[8] This is distressing since the choice of s is arbitrary and irrelevant to the actual hypothesis.

Our first-order asymptotic theory is not useful to help pick s, as $W_n(s) \xrightarrow{d} \chi_1^2$ under H_0 for any s. This is a context where **Monte Carlo simulation** can be quite useful as a tool to study and compare the exact distributions of statistical procedures in finite samples. The method uses random simulation to create artificial datasets, to which we apply the statistical tools of interest. This

[8] Breusch and Schmidt (1988) show that any positive value for the Wald test statistic is possible by rewriting H_0 in an algebraically equivalent form.

produces random draws from the statistic's sampling distribution. Through repetition, features of this distribution can be calculated. In the present context of the Wald statistic, one feature of importance is the Type I error of the test using the asymptotic 5% critical value 3.84 - the probability of a false rejection, $P(W_n(s) > 3.84|\beta = 1)$. Given the simplicity of the model, this probability depends only on s, n, and σ^2. In Table 1 we report the results of a Monte Carlo simulation where we vary these three parameters: the value of s is varied from 1 to 10, n is varied among 20, 100 and 500, and σ is varied among 1 and 3. The table reports the simulation estimate of the Type I error probability from $50,000$ random samples. Each row of the table corresponds to a different value of s - and thus corresponds to a particular choice of test statistic. The second through seventh columns contain the Type I error probabilities for different combinations of n and σ. These probabilities are calculated as the percentage of the $50,000$ simulated Wald statistics $W_n(s)$ which are larger than 3.84. The null hypothesis $\beta^s = 1$ is true, so these probabilities are Type I error.

	$\sigma = 1$			$\sigma = 3$		
s	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
1	.06	.05	.05	.07	.05	.05
2	.08	.06	.05	.15	.08	.06
3	.10	.06	.05	.21	.12	.07
4	.13	.07	.06	.25	.15	.08
5	.15	.08	.06	.28	.18	.10
6	.17	.09	.06	.30	.20	.11
7	.19	.10	.06	.31	.22	.13
8	.20	.12	.07	.33	.24	.14
9	.22	.13	.07	.34	.25	.15
10	.23	.14	.08	.35	.26	.16

Table 1: Type I Error Probability of Asymptotic 5% $W_n(s)$ Test

Note: Rejection frequencies from 50,000 simulated random samples

To interpret the table, remember that the ideal Type I error probability is 5%(.05) with deviations indicating distortion. Type I error rates between 3% and 8% are considered reasonable. Error rates above 10% are considered excessive. Rates above 20% are unacceptable. When comparing statistical procedures, we compare the rates row by row, looking for tests for which rejection rates are close to 5% and rarely fall outside of the $3\% - 8\%$ range. For this particular example the only test which meets this criterion is the conventional $W_n = W_n(1)$ test. Any other choice of s leads to a test with unacceptable Type I error probabilities.

In Table 1 you can also see the impact of variation in sample size. In each case, the Type I error probability improves towards 5% as the sample size n increases. There is, however, no magic choice of n for which all tests perform uniformly well. Test performance deteriorates as s increases, which is not surprising given the dependence of $W_n(s)$ on s as shown in Figure 4.

In this example it is not surprising that the choice $s = 1$ yields the best test statistic. Other choices are arbitrary and would not be used in practice. While this is clear in this particular example, in other examples natural choices are not always obvious and the best choices may in fact appear counter-intuitive at first. This point can be illustrated through another example which is similar to one developed in Gregory and Veall (1985). Take the model

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + u_i, E[\mathbf{x}_i u_i] = \mathbf{0} \tag{19}$$

and the hypothesis

$$H_0\colon \frac{\beta_1}{\beta_2} = \theta_0$$

where θ_0 is a known constant. Equivalently, define $\theta = \beta_1/\beta_2$, so the hypothesis can be stated as $H_0\colon \theta = \theta_0$. Let $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2\right)'$ be the least-squares estimates of (19), $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}}$ be an estimate of the covariance matrix for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\theta} = \widehat{\beta}_1/\widehat{\beta}_2$.[9] Define

$$\widehat{\mathbf{R}}_1 = \left(0, \frac{1}{\widehat{\beta}_2}, -\frac{\widehat{\beta}_1}{\widehat{\beta}_2^2}\right)'$$

so that the standard error for $\widehat{\theta}$ is $s(\widehat{\theta}) = \left(\widehat{\mathbf{R}}_1'\widehat{\mathbf{V}}\widehat{\mathbf{R}}_1\right)^{1/2}$. In this case, a t-statistic for H_0 is

$$t_{1n} = \frac{\widehat{\beta}_1/\widehat{\beta}_2 - \theta_0}{s(\widehat{\theta})}.$$

An alternative statistic can be constructed through reformulating the null hypothesis as

$$H_0\colon \beta_1 - \theta_0\beta_2 = 0.$$

A t-statistic based on this formulation of the hypothesis is

$$t_{2n} = \frac{\widehat{\beta}_1 - \theta_0\widehat{\beta}_2}{\left(\mathbf{R}_2'\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}}\mathbf{R}_2\right)^{1/2}},$$

where $\mathbf{R}_2 = (0, 1, -\theta_0)'$.

To compare t_{1n} and t_{2n} we perform another simple Monte Carlo simulation. We let x_{1i} and x_{2i} be mutually independent $N(0,1)$ variables, u_i be an independent $N(0,\sigma^2)$ draw with $\sigma = 3$, and normalize $\beta_0 = 1$ and $\beta_1 = 1$. This leaves β_2 as a free parameter, along with sample size n. We vary β_2 among $.1, .25, .50, .75$, and 1.0 and n among 100 and 500. The one-sided Type I error probabilities $P(t_n < -1.645)$ and $P(t_n > 1.645)$ are calculated from $50,000$ simulated samples. The results are presented in Table 2. Ideally, the entries in the table should be 0.05. However, the rejection rates for the t_{1n} statistic diverge greatly from this value, especially for small values of

[9] If $\widehat{\mathbf{V}}$ is used to estimate the asymptotic variance of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$, then $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}} = \widehat{\mathbf{V}}/n$.

β_2. The left tail probabilities $P(t_{1n} < -1.645)$ greatly exceed 5%, while the right tail probabilities $P(t_{1n} > 1.645)$ are close to zero in most cases. In contrast, the rejection rates for the linear t_{2n} statistic are invariant to the value of β_2, and are close to the ideal 5% rate for both sample sizes. The implication of Table 2 is that the two t-ratios have dramatically different sampling behaviors.

	$n = 100$				$n = 500$			
	$P(t_n < -1.645)$		$P(t_n > 1.645)$		$P(t_n > 1.645)$		$P(t_n > 1.645)$	
β_2	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}
.10	.47	.06	.00	.06	.28	.05	.00	.05
.25	.26	.06	.00	.06	.15	.05	.00	.05
.50	.15	.06	.00	.06	.10	.05	.00	.05
.75	.12	.06	.00	.06	.09	.05	.00	.05
1.00	.10	.06	.00	.06	.07	.05	.02	.05

Table 2: Type I Error Probability of Asymptotic 5% t-tests

The common message from both examples is that Wald statistics are sensitive to the algebraic formulation of the null hypothesis. In all cases, if the hypothesis can be expressed as a linear restriction on the model parameters, this formulation should be used. If no linear formulation is feasible, then the "most linear" formulation should be selected (as suggested by the theory of Phillips and Park (1988)), and alternatives to asymptotic critical values should be considered. It is also prudent to consider alternative tests to the Wald statistic, such as the minimum distance statistic developed in the next section.

11 Minimum Distance Test (*)

The likelihood ratio (LR) test is valid only under homoskedasticity. The counterpart of the LR test in the heteroskedastic environment is the minimum distance test. Based on the idea of the LR test, the minimum distance statistic is defined as

$$J_n = \min_{\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}} J_n(\boldsymbol{\beta}) - \min_{\boldsymbol{\beta}} J_n(\boldsymbol{\beta}) = \min_{\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}} J_n\left(\widehat{\boldsymbol{\beta}}_{MD}\right),$$

where $\min_{\boldsymbol{\beta}} J_n(\boldsymbol{\beta}) = 0$ if no restrictions are imposed (why?). $J_n \geq 0$ measures the cost (on $J_n(\boldsymbol{\beta})$) of imposing the null restriction $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}$. Usually, $\mathbf{W}_n$ in $J_n(\boldsymbol{\beta})$ is chosen to be the efficient weight matrix $\widehat{\mathbf{V}}^{-1}$, and the corresponding J_n is denoted as J_n^* with

$$J_n^* = n\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{EMD}\right)' \widehat{\mathbf{V}}^{-1}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{EMD}\right).$$

Consider the class of linear hypotheses H_0: $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$. In this case, we know from (9) that

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

so

$$
\begin{aligned}
J_n^* &= n\left(\mathbf{R}'\widehat{\boldsymbol{\beta}}-\mathbf{c}\right)'\left(\mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{V}}\mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}\right)^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}}-\mathbf{c}\right)\\
&= n\left(\mathbf{R}'\widehat{\boldsymbol{\beta}}-\mathbf{c}\right)'\left(\mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}\right)^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}}-\mathbf{c}\right)=W_n.
\end{aligned}
$$

Thus for linear hypotheses, the efficient minimum distance statistic J_n^* is identical to the Wald statstic W_n which is heteroskedastic-robust.

Exercise 13 *Show that $J_n^* = W$ under homoskedasticity when the null hypothesis is H_0: $\mathbf{R}'\boldsymbol{\beta}=\mathbf{c}$, where W is the homoskedastic form of the Wald statistic defined in Chapter 4.*

For nonlinear hypotheses, however, the Wald and minimum distance statistics are different. We know from Section 10 that the Wald statistic is not robust to the formulation of the null hypothesis. However, like the LR test statistic, the minimum distance statistic is invariant to the algebraic formulation of the null hypothesis, so is immune to this problem. Consequently, a simple solution to the problem associated with W_n in Section 10 is to use the minimum distance statistic J_n, which equals W_n with $s=1$ in the first example, and $|t_{2n}|$ in the second example there. Whenever possible, the Wald statistic should not be used to test nonlinear hypotheses.

Exercise 14 *Show that $J_n^* = W_n(1)$ in the first example and $J_n^* = t_{2n}^2$ in the second example of Section 10.*

Newey and West (1987a) established the asymptotic null distribution of J_n^* for linear and nonlinear hypotheses.

Theorem 11 *Under the assumptions of Theorem 8, $J_n^* \xrightarrow{d} \chi_q^2$ under H_0.*

12 The Heteroskedasticity-Robust LM Test (*)

The validity of the LM test in the normal regression model depends on the assumption that the error is homoskedastic. In this section, we extend the homoskedasticity-only LM test to the heteroskedasticity-robust form. Suppose the null hypothesis is H_0: $\boldsymbol{\beta}_2 = 0$, where $\boldsymbol{\beta}$ is decomposed as $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $k_1\times 1$ and $k_2 \times 1$ vectors, respectively, $k=k_1+k_2$, and $\mathbf{x}$ is decomposed as $(\mathbf{x}_1', \mathbf{x}_2')'$ correspondingly with $\mathbf{x}_1$ including the constant.

Exercise 15 *Show that testing any linear constraints $\mathbf{R}'\boldsymbol{\beta}=\mathbf{c}$ is equivalent to testing some original coefficients being zero in a new regression with redefined $\mathbf{X}$ and $\mathbf{y}$, where $\mathbf{R}\in\mathbb{R}^{k_2\times k}$ is full rank.*

After some algebra we can write

$$
LM = \left(n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{r}}_i\widetilde{u}_i\right)'\left(\widetilde{\sigma}^2 n^{-1}\sum_{i=1}^n \widehat{\mathbf{r}}_i\widehat{\mathbf{r}}_i'\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{r}}_i\widetilde{u}_i\right),
$$

where $\widetilde{\sigma}^2 = n^{-1}\sum_{i=1}^n \widetilde{u}_i^2$ and each $\widehat{\mathbf{r}}_i$ is a $k_2 \times 1$ vector of OLS residuals from the (multivariate) regression of $\mathbf{x}_{i2}$ on $\mathbf{x}_{i1}$, $i = 1, \cdots, n$. This statistic is not robust to heteroskedasticity because the matrix in the middle is not a consistent estimator of the asymptotic variance of $n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{r}}_i\widetilde{u}_i$ under heteroskedasticity. A heteroskedasticity-robust statistic is

$$\begin{aligned} LM_n &= \left(n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{r}}_i\widetilde{u}_i\right)'\left(n^{-1}\sum_{i=1}^n \widetilde{u}_i^2\widehat{\mathbf{r}}_i\widehat{\mathbf{r}}_i'\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{r}}_i\widetilde{u}_i\right) \\ &= \left(\sum_{i=1}^n \widehat{\mathbf{r}}_i\widetilde{u}_i\right)'\left(\sum_{i=1}^n \widetilde{u}_i^2\widehat{\mathbf{r}}_i\widehat{\mathbf{r}}_i'\right)^{-1}\left(\sum_{i=1}^n \widehat{\mathbf{r}}_i\widetilde{u}_i\right). \end{aligned}$$

Dropping the i subscript, this is easily obtained as $n - SSR_0$ from the OLS regression (without intercept)[10]

$$1 \text{ on } \widetilde{u}\cdot\widehat{\mathbf{r}}, \tag{20}$$

where $\widetilde{u}\cdot\widehat{\mathbf{r}} = (\widetilde{u}\cdot\widehat{r}_1, \cdots, \widetilde{u}\cdot\widehat{r}_{k_2})'$ is the $k_2 \times 1$ vector obtained by multiplying $\widetilde{u}$ by each element of $\widehat{\mathbf{r}}$ and SSR_0 is just the usual sum of squared residuals from the regression. Thus, we first regress each element of $\mathbf{x}_2$ onto all of $\mathbf{x}_1$ and collect the residuals in $\widehat{\mathbf{r}}$. Then we form $\widetilde{u}\cdot\widehat{\mathbf{r}}$ (observation by observation) and run the regression in (20); $n - SSR_0$ from this regression is distributed asymptotically as $\chi^2_{k_2}$. For more details, see Davidson and MacKinnon (1985, 1993) or Wooldridge (1991, 1995).

13 Test Consistency

We now define the test consistency against fixed alternatives. This concept was first introduced by Wald and Wolfowitz (1940).

Definition 1 *A test of* H_0*:* $\boldsymbol{\theta} \in \Theta_0$ *is* consistent against fixed alternatives *if for all* $\boldsymbol{\theta} \in \Theta_1$, $P\left(\text{Reject } H_0|\boldsymbol{\theta}\right) \to 1$ *as* $n \to \infty$*.*

To understand this concept, consider the following simple example. Suppose that y_i is i.i.d. $N(\mu, 1)$. Consider the t-statistic $t_n(\mu) = \sqrt{n}\left(\overline{y} - \mu\right)$, and tests of H_0: $\mu = 0$ against H_1: $\mu > 0$. We reject H_0 if $t_n = t_n(0) > c$. Note that

$$t_n = t_n(\mu) + \sqrt{n}\mu$$

and $t_n(\mu) \equiv Z$ has an exact $N(0,1)$ distribution. This is because $t_n(\mu)$ is centered at the true mean μ, while the test statistic $t_n(0)$ is centered at the (false) hypothesized mean of 0. The power of the test is

$$P\left(t_n > c|\theta\right) = P\left(Z + \sqrt{n}\mu > c\right) = 1 - \Phi\left(c - \sqrt{n}\mu\right).$$

This function is monotonically increasing in μ and n, and decreasing in c. Notice that for any c and $\mu \neq 0$, the power increases to 1 as $n \to \infty$. This means that for $\mu \in H_1$, the test will reject H_0

[10] If there is an intercept, then the regression is trivial.

with probability approaching 1 as the sample size gets large. This is exactly test consistency.

For tests of the form "Reject H_0 if $T_n > c$", a sufficient condition for test consistency is that T_n diverges to positive infinity with probability one for all $\boldsymbol{\theta} \in \Theta_1$. In general, the t-test and Wald test are consistent against fixed alternatives. For example, in testing H_0: $\theta = \theta_0$,

$$t_n = \frac{\widehat{\theta} - \theta_0}{s\left(\widehat{\theta}\right)} = \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)} + \frac{\sqrt{n}\left(\theta - \theta_0\right)}{\sqrt{\widehat{V}_\theta}} \tag{21}$$

since $s\left(\widehat{\theta}\right) = \sqrt{\widehat{V}_\theta / n}$. The first term on the right-hand-side converges in distribution to $N(0,1)$. The second term on the right-hand-side equals zero if $\theta = \theta_0$, converges in probability to $+\infty$ if $\theta > \theta_0$, and converges in probability to $-\infty$ if $\theta < \theta_0$. Thus the two-sided t-test is consistent against H_1: $\theta \neq \theta_0$, and one-sided t-tests are consistent against the alternatives for which they are designed. For another example, The Wald statistic for H_0: $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ against H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ is

$$W_n = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

Under H_1, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta} \neq \theta_0$. Thus $\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{p} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{V}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) > 0$. Hence under H_1, $W_n \xrightarrow{p} \infty$. Again, this implies that Wald tests are consistent tests.

(*) Andrews (1986) introduces a testing analogue of estimator consistency, called **complete consistency**, which he shows is more appropriate than test consistency. It is shown that a sequence of estimators is consistent, if and only if certain tests based on the estimators (such as Wald or likelihood ratio tests) are completely consistent, for all simple null hypotheses.

14 Asymptotic Local Power

Consistency is a good property for a test, but does not give a useful approximation to the power of a test. To approximate the power function we need a distributional approximation. The standard asymptotic method for power analysis uses what are called **local alternatives**. This is similar to our analysis of restriction estimation under misspecification. The technique is to index the parameter by sample size so that the asymptotic distribution of the statistic is continuous in a localizing parameter. We first consider the t-test and then the Wald test.

In the t-test, we consider parameter vectors $\boldsymbol{\beta}_n$ which are indexed by sample size n and satisfy the real-valued relationship

$$\theta_n = r(\boldsymbol{\beta}_n) = \theta_0 + n^{-1/2}h, \tag{22}$$

where the scalar h is called a **localizing parameter**. We index $\boldsymbol{\beta}_n$ and θ_n by sample size to indicate their dependence on n. The way to think of (22) is that the *true* value of the parameters are $\boldsymbol{\beta}_n$ and θ_n. The parameter θ_n is close to the hypothesized value θ_0, with deviation $n^{-1/2}h$. Such a sequence of local alternatives θ_n is often called a **Pitman** (1949) **drift** or a **Pitman sequence**.[11]

[11] See McManus (1991) for who invented local power analysis.

We know for a fixed alternative, the power will converge to 1 as $n \to \infty$. To offset the effect of increasing n, we make the alternative harder to distinguish from H_0 as n gets larger. The rate $n^{-1/2}$ is the correct balance between these two forces. In the statistical literature, such alternatives are termed as "contiguous" local alternatives.

The specification (22) states that for any fixed h, θ_n approaches θ_0 as n gets large. Thus θ_n is "close" or "local" to θ_0. The concept of a localizing sequence (22) might seem odd at first as in the actual world the sample size cannot mechanically affect the value of the parameter. Thus (22) should not be interpreted literally. Instead, it should be interpreted as a technical device which allows the asymptotic distribution of the test statistic to be continuous in the alternative hypothesis.

Similarly as in (21),

$$t_n = \frac{\widehat{\theta} - \theta_0}{s\left(\widehat{\theta}\right)} = \frac{\widehat{\theta} - \theta_n}{s\left(\widehat{\theta}\right)} + \frac{\sqrt{n}\,(\theta_n - \theta_0)}{\sqrt{\widehat{V}_\theta}} \xrightarrow{d} Z + \delta$$

under the local alternative (22), where $Z \sim N(0,1)$ and $\delta = h/\sqrt{V_\theta}$. In testing the one-sided alternative H_1: $\theta > \theta_0$, a t-test rejects H_0 for $t_n > z_\alpha$. The **asymptotic local power** of this test is the limit of the rejection probability under the local alternative (22),

$$\begin{aligned} \lim_{n\to\infty} P\left(\text{Reject } H_0 | \theta = \theta_n\right) &= \lim_{n\to\infty} P\left(t_n > z_\alpha | \theta = \theta_n\right) \\ &= P\left(Z + \delta > z_\alpha\right) = 1 - \Phi(z_\alpha - \delta) = \Phi(\delta - z_\alpha) \equiv \pi_\alpha\left(\delta\right). \end{aligned}$$

We call $\pi_\alpha(\delta)$ the **local power function**.

Exercise 16 *Derive the local power function for the two-sided t-test.*

In Figure 5 we plot the local power function $\pi_\alpha(\delta)$ as a function of $\delta \in [0,4]$ for tests of asymptotic size $\alpha = 0.10, \alpha = 0.50$, and $\alpha = 0.01$. We do not consider $\delta < 0$ since θ_n should be greater than θ_0. $\delta = 0$ corresponds to the null hypothesis so $\pi_\alpha(0) = \alpha$. The power functions are monotonically increasing in both δ and α. The monotonicity with respect to α is due to the inherent trade-off between size and power. Decreasing size induces a decrease in power, and vice versa. The coefficient δ can be interpreted as the parameter deviation measured as a multiple of the standard error $s\left(\widehat{\theta}\right)$. To see this, recall that $s\left(\widehat{\theta}\right) = n^{-1/2}\sqrt{\widehat{V}_\theta} \approx n^{-1/2}\sqrt{V_\theta}$ and then note that

$$\delta = \frac{h}{\sqrt{V_\theta}} \approx \frac{n^{-1/2}h}{s\left(\widehat{\theta}\right)} = \frac{\theta_n - \theta_0}{s\left(\widehat{\theta}\right)},$$

meaning that δ equals the deviation $\theta_n - \theta_0$ expressed as multiples of the standard error $s\left(\widehat{\theta}\right)$. Thus as we examine Figure 5, we can interpret the power function at $\delta = 1$ (e.g., 26% for a 5% size test) as the power when the parameter θ_n is one standard error above the hypothesized value.

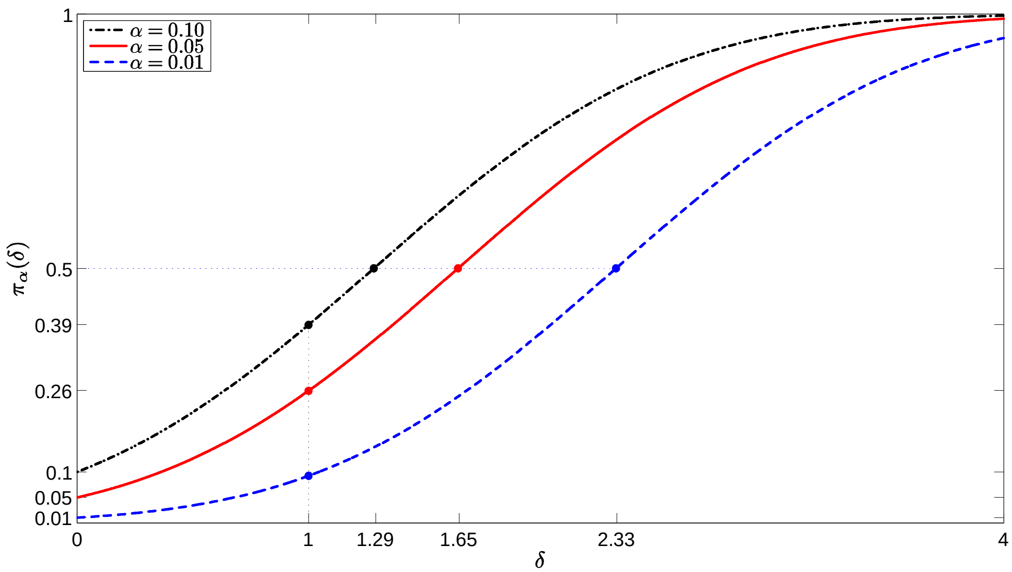

Figure 5: Asymptotic Local Power Function of One-Sided t-Test

Exercise 17 *Suppose we have n_0 data points, and we want to know the power when the true value of θ is ϑ. Which δ we should confer in Figure 5?*

The difference between power functions can be measured either vertically or horizontally. For example, in Figure 5 there is a vertical dotted line at $\delta = 1$, showing that the asymptotic local power $\pi_\alpha(1)$ equals 39% for $\alpha = 0.10$, equals 26% for $\alpha = 0.05$ and equals 9% for $\alpha = 0.01$. This is the difference in power across tests of differing sizes, holding fixed the parameter in the alternative. A horizontal comparison can also be illuminating. To illustrate, in Figure 5 there is a horizontal dotted line at 50% power. 50% power is a useful benchmark, as it is the point where the test has equal odds of rejection and acceptance. The dotted line crosses the three power curves at $\delta = 1.29$ ($\alpha = 0.10$), $\delta = 1.65$ ($\alpha = 0.05$), and $\delta = 2.33$ ($\alpha = 0.01$). This means that the parameter θ must be at least 1.65 standard errors above the hypothesized value for the one-sided test to have 50% (approximate) power. The ratio of these values (e.g., $1.65/1.29 = 1.28$ for the asymptotic 5% versus 10% tests) measures the relative parameter magnitude needed to achieve the same power. (Thus, for a 5% size test to achieve 50% power, the parameter must be 28% larger than for a 10% size test.) Even more interesting, the square of this ratio (e.g., $(1.65/1.29)^2 = 1.64$) can be interpreted as the increase in sample size needed to achieve the same power under fixed parameters. That is, to achieve 50% power, a 5% size test needs 64% more observations than a 10% size test. This interpretation follows by the following informal argument. By definition and (22) $\delta = h/\sqrt{V_\theta} = \sqrt{n}\left(\theta_n - \theta_0\right)/\sqrt{V_\theta}$. Thus holding θ and V_θ fixed, we can see that δ^2 is proportional to n.

We next generalize the local power analysis to the case of vector-valued alternatives. Now the local parametrization takes the form

$$\boldsymbol{\theta}_n = \mathbf{r}(\boldsymbol{\beta}_n) = \boldsymbol{\theta}_0 + n^{-1/2}\mathbf{h}, \tag{23}$$

where $\mathbf{h}$ is a $q \times 1$ vector. Under (23),

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\right) + \mathbf{h} \xrightarrow{d} \mathbf{Z}_{\mathbf{h}} \sim N(\mathbf{h}, \mathbf{V}_{\boldsymbol{\theta}}),$$

a normal random vector with mean $\mathbf{h}$ and variance matrix $\mathbf{V}_{\boldsymbol{\theta}}$. Applied to the Wald statistic we find

$$W_n = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathbf{Z}_{\mathbf{h}}'\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{Z}_{\mathbf{h}} \sim \chi_q^2(\lambda).$$

where $\chi_q^2(\lambda)$ is a **non-central chi-square distribution** with q degrees of freedom and non-central parameter (or noncentrality) $\lambda = \mathbf{h}'\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{h}$. Under the null, $\mathbf{h} = \mathbf{0}$, and the $\chi_q^2(\lambda)$ distribution then degenerates to the usual χ_q^2 distribution. In the case of $q = 1$, $|Z + \delta|^2 \sim \chi_1^2(\lambda)$ with $\lambda = \delta^2$. The asymptotic local power of the Wald test at the level α is

$$P\left(\chi_q^2(\lambda) > \chi_{q,\alpha}^2\right) \equiv \pi_{q,\alpha}(\lambda).$$

Figure 6 plots $\pi_{q,.05}(\lambda)$ (the power of asymptotic 5% tests) as a function of λ for $q = 1$, 2 and 3. The power functions are monotonically increasing in λ and asymptote to one. Figure 6 also shows the power loss for fixed non-centrality parameter λ as the dimensionality of the test increases. The power curves shift to the right as q increases, resulting in a decrease in power. This is illustrated by the dotted line at 50% power. The dotted line crosses the three power curves at $\lambda = 3.85$ ($q = 1$), $\lambda = 4.96$ ($q = 2$), and $\lambda = 5.77$ ($q = 3$). The ratio of these λ values correspond to the relative sample sizes needed to obtain the same power. Thus increasing the dimension of the test from $q = 1$ to $q = 2$ requires a 28% increase in sample size, or an increase from $q = 1$ to $q = 3$ requires a 50% increase in sample size, to obtain a test with 50% power. Intuitively, when testing more restrictions, we need more deviation from the null (or equivalently, more data points) to achieve the same power.

Exercise 18 *(i) Show that* $\widehat{\mathbf{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \mathbf{V}_{\boldsymbol{\theta}}$ *under the local alternative* $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\mathbf{b}$. *(ii) If the local alternative* $\boldsymbol{\beta}_n$ *is specified as as in (i), what is the local power?*

Exercise 19 (*) *Derive the local power function of the minimum distance test with local alternatives (23). Is it the same as the Wald test?*

Exercise 20 (Empirical) *The data set invest.dat contains data on 565 U.S. firms extracted from Compustat for the year 1987. The variables, in order, are*

- I_i *Investment to Capital Ratio (multiplied by 100).*
- Q_i *Total Market Value to Asset Ratio (Tobin's Q).*

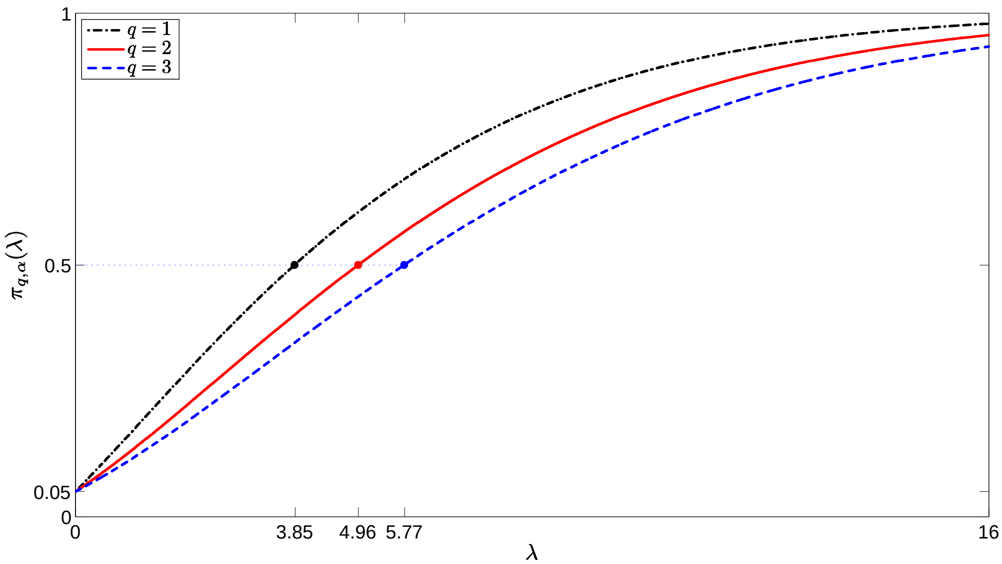

Figure 6: Asymptotic Local Power Function of the Wald Test

- *C_i Cash Flow to Asset Ratio.*
- *D_i Long Term Debt to Asset Ratio.*

The flow variables are annual sums for 1987. The stock variables are beginning of year.

(a) *Estimate a linear regression of I_i on the other variables. Calculate appropriate standard errors.*

(b) *Calculate asymptotic confidence intervals for the coefficients.*

(c) *This regression is related to Tobin's q theory of investment, which suggests that investment should be predicted solely by Q_i. Thus the coefficient on Q_i should be positive and the others should be zero. Test the joint hypothesis that the coefficients on C_i and D_i are zero. Test the hypothesis that the coefficient on Q_i is zero. Are the results consistent with the predictions of the theory?*

(d) *Now try a non-linear (quadratic) specification. Regress I_i on Q_i, C_i, D_i, Q_i^2, C_i^2, D_i^2, Q_iC_i, Q_iD_i, C_iD_i. Test the joint hypothesis that the six interaction and quadratic coefficients are zero.*

Exercise 21 (Empirical) *In a paper in 1963, Marc Nerlove analyzed a cost function for 145 American electric companies. The data file nerlov.dat contains his data. The variables are described as follows,*

- *Column 1: total costs (call it TC) in millions of dollars*

- *Column 2: output (Q) in billions of kilowatt hours*
- *Column 3: price of labor (PL)*
- *Column 4: price of fuels (PF)*
- *Column 5: price of capital (PK)*

Nerlove was interested in estimating a cost function: $TC = f(Q, PL, PF, PK)$.

(a) *First estimate an unrestricted Cobb-Douglas specification*

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + u_i. \tag{24}$$

Report parameter estimates and standard errors.

(b) *Using a Wald statistic, test the hypothesis* H_0*:* $\beta_3 + \beta_4 + \beta_5 = 1$.

(c) *Estimate (24) by least-squares imposing this restriction by substitution. Report your parameter estimates and standard errors.*

(d) *Estimate (24) subject to* $\beta_3 + \beta_4 + \beta_5 = 1$ *using the RLS estimator. Do you obtain the same estimates as in part (c)?*

Technical Appendix: Expected Length of a Confidence Interval

Suppose $L \leq \theta \leq U$ is a CI for θ, where L and U are random. The expected length of the interval is $E[U - L]$, which turns out equal to

$$\int_{\theta \neq \theta_0} P(L \leq \theta \leq U)\, d\theta,$$

the integrated probability of the interval covering the false value, where $P(\cdot)$ is the probability measure under the truth θ_0. To see why, note that

$$\begin{aligned} E[U - L] &= \int \int 1(L \leq \theta \leq U)\, d\theta dP \\ &= \int \int 1(L \leq \theta \leq U)\, dP d\theta \\ &= \int P(L \leq \theta \leq U)\, d\theta = \int_{\theta \neq \theta_0} P(L \leq \theta \leq U)\, d\theta. \end{aligned}$$

So minimizing the expected length of a CI is equivalent to minimizing the coverage probability for each $\theta \neq \theta_0$. From the discussion in the main text,

$$P(L \leq \theta \leq U) = P(X \in A(\theta)),$$

where X is the data and $A(\theta)$ is the acceptance region for the null that θ is the true value. As a result, for $\theta \neq \theta_0$, $P(L \leq \theta \leq U)$ is the type II error in testing $H_0 : \theta$ vs $H_1 : \theta_0$, and minimizing $P(L \leq \theta \leq U)$ is equivalent to maximizing the power. By the Neyman-Pearson Lemma, the most powerful test is

$$1\left(\frac{f(X)}{f_\theta(X)} > c(\theta)\right), \tag{25}$$

where $f(X)$ is the true density of X (or density under θ_0), and $c(\theta)$ is the critical value for $H_0 : \theta$. Collecting all θ's such that $\frac{f(X)}{f_\theta(X)} \leq c(\theta)$ is the length minimizing CI.

The arguments above assume θ_0 were known, but if it were known, why do we need a CI for it? A more natural measure for the interval length is the average expected length $\int E_\vartheta [U - L]\, d\nu(\vartheta)$, where ϑ is the parameter in the model and $\nu(\cdot)$ is a measure for ϑ representing some prior information. Similarly, we can show

$$\int E_\vartheta [U - L]\, d\nu(\vartheta) = \int\int P_\vartheta (X \in A(\theta))\, d\nu(\vartheta)\, d\theta.$$

Minimizing the average expected length is equivalent to maximizing the power in testing $H_0 : \theta$ vs $H_1 : P(\cdot)$, where $P(X \in A) = \int P_\vartheta (X \in A)\, d\nu(\vartheta)$ has the density $\int f_\vartheta(X) d\nu(\vartheta)$ with $f_\vartheta(\cdot)$ being the density associated with $P_\vartheta(\cdot)$. For any test φ, the average power

$$\int E_\vartheta [\varphi]\, d\nu(\vartheta) = \int\int \varphi f_\vartheta(X) dX d\nu(\vartheta) = \int \varphi \int f_\vartheta(X) d\nu(\vartheta)\, dX = \int \varphi dP(X)$$

is just the power in the above test, so the test maximizing the power against $P(\cdot)$ is equivalent to maximizing the average power $\int E_\vartheta [\varphi]\, d\nu(\vartheta)$. The corresponding test is the same as (25) but replaces $f(X)$ by $\int f_\vartheta(X) d\nu(\vartheta)$.