Least Squares Estimation- Large-Sample Properties

Ping Yu

School of Economics and Finance
The University of Hong Kong

Introduction

- If $u|\mathbf{x} \sim N(0, \sigma^2)$, we have shown that $\widehat{\beta}|\mathbf{X} \sim N\left(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$.
- In general the distribution of $u|\mathbf{x}$ is unknown.
- Even if it is known, the unconditional distribution of $\widehat{\beta}$ is hard to derive since $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is a complicated function of $\{\mathbf{x}_i\}_{i=1}^n$.
- The asymptotic (or large sample) method approximates (unconditional) sampling distributions based on the limiting experiment that the sample size n tends to infinity.
- It does not require any assumption on the distribution of $u|\mathbf{x}$, and only some moments restrictions are imposed.
- Three steps: consistency, asymptotic normality and estimation of the covariance matrix.

Asymptotics for the LSE

Consistency

- Express $\widehat{\beta}$ as

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

- To show $\widehat{\beta}$ is consistent, we impose the following additional assumptions.
- **Assumption OLS.1′**: $\text{rank}(E[\mathbf{x}\mathbf{x}']) = k$.
- **Assumption OLS.2′**: $y = \mathbf{x}'\beta + u$ with $E[\mathbf{x}u] = \mathbf{0}$.
- 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.
- Assumption OLS.2′ is weaker than Assumption OLS.2.

Theorem

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

Proof.

From (1), to show $\widehat{\beta} \xrightarrow{p} \beta$, 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 $\left(E[\mathbf{x}_i\mathbf{x}_i'], E[\mathbf{x}_i u_i]\right)$. The last equality is from Assumption OLS.2′. □

Proof.

[Proof continue] (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,[a]

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

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

[a] Cauchy-Schwarz inequality: For any random $m \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$, $E[\|\mathbf{X}'\mathbf{Y}\|] \le E\left[\|\mathbf{X}\|^2\right]^{1/2} E\left[\|\mathbf{Y}\|^2\right]^{1/2}$, where the inner product is defined as $\langle \mathbf{X}, \mathbf{Y}\rangle = E[\|\mathbf{X}'\mathbf{Y}\|]$, and for a $m \times n$ matrix $\mathbf{A}$, $\|\mathbf{A}\| = \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2\right)^{1/2} = [\text{trace}(\mathbf{A}'\mathbf{A})]^{1/2}$.

[b] 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 $E[x_i^2] \neq 0$.

Consistency of $\widehat{\sigma}^2$ and s^2

Theorem

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

Proof.

Note that

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

Thus

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

and

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

□

Proof.

[Proof continue]

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

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.

Asymptotic Normality

- To study the asymptotic normality of $\widehat{\beta}$, we impose the following additional assumption.
- **Assumption OLS.5**: $E[u^4] < \infty$ and $E\left[\|\mathbf{x}\|^4\right] < \infty$.

Theorem

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

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

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

Proof.

From (1),

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

□

Proof.

[Proof continue] Note first that

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

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

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

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

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

by Slutsky's theorem. □

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

- 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**.

Partitioned Formula of $\mathbf{V}^0$

- Sometimes, to state the asymptotic distribution of part of $\widehat{\beta}$ as in the residual regression, we partition $\mathbf{Q}$ and Ω as

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

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

LSE as a MoM Estimator

- The LSE is a MoM estimator, and the moment conditions are

$$E[\mathbf{x}u] = \mathbf{0} \text{ with } u = y - \mathbf{x}'\beta.$$

- The sample analog is the normal equation

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

  the solution of which is exactly the LSE.

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

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

- 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.

Intuition

- 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,

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

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, $Var(x) = \left|\frac{\partial E[xu]}{\partial \beta}\right|$.

Asymptotics for the Weighted Least Squares (WLS) Estimator

- The WLS estimator is a special GLS estimator with a diagonal weight matrix.
- Recall that

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

which reduces to

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

when $\mathbf{W} = \text{diag}\{w_1, \cdots, w_n\}$.

- Note that this estimator is a MoM estimator under the moment condition (check!)

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

so

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

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

Covariance Matrix Estimators

Sample Analogs

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

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

and

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\widehat{u}_i^2 = \frac{1}{n}\mathbf{X}'\text{diag}\left\{\widehat{u}_1^2,\cdots,\widehat{u}_n^2\right\}\mathbf{X} \equiv \frac{1}{n}\mathbf{X}'\widehat{\mathbf{D}}\mathbf{X} \tag{5}$$

are the MoM estimators (exercise) for $\mathbf{Q}$ and Ω, where $\left\{\widehat{u}_i\right\}_{i=1}^n$ are the OLS residuals.

- Given that $\mathbf{V} = \mathbf{Q}^{-1}\Omega\mathbf{Q}^{-1}$, it can be estimated by

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

and $AVar\left(\widehat{\beta}\right)$ is estimated by

$$\widehat{\mathbf{V}}/n = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{D}}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

History and Limitation

- People thought $\Omega = E\left[\mathbf{x}_i\mathbf{x}_i'\sigma_i^2\right]$ 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"
- In the homoskedastic case, we can estimate $\mathbf{V}$ by $\widehat{\mathbf{V}}^0 = \widehat{\sigma}^2\widehat{\mathbf{Q}}^{-1}$.
- **Practical Suggestion**: Use $\widehat{\mathbf{V}}$ rather than $\widehat{\mathbf{V}}^0$ whenever possible.
- $\widehat{AVar}\left(\widehat{\beta}_j\right) \overset{\text{why?}}{=} \sum_{i=1}^n w_{ij}\widehat{u}_i^2/SSR_j \overset{\text{homo}}{\longrightarrow} \widehat{AVar}\left(\widehat{\beta}_j\right) = n^{-1}\sum_{i=1}^n \widehat{u}_i^2/SSR_j$.
- 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.
- It can be shown that the former is actually more variable than the latter, which is the price paid for robustness.

Consistency of $\widehat{\mathbf{V}}$

Theorem

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 $\widehat{\Omega}$ is consistent, by the CMT $\widehat{\mathbf{V}}$ is consistent. Using (3)

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

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

Proof.

The second term satisfies

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

where the first inequality is from the triangle inequality,[a] and the second and third inequalities are from the Schwarz matrix inequality. □

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

Proof.

[Proof continue] By Hölder's inequality,[a]

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

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

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

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

[a] 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 $\mathbf{Y}$, $E\left[\|\mathbf{X}'\mathbf{Y}\|\right] \leq E\left[\|\mathbf{X}\|^p\right]^{1/p} E\left[\|\mathbf{Y}\|^q\right]^{1/q}$.

Functions of Parameters

Functions of Parameters

- Let $\theta = \mathbf{r}(\beta)$ denote the parameter of interest, where $\mathbf{r}$: $\mathbb{R}^k \to \mathbb{R}^q$.
- **Assumption RLS.1′**: $\mathbf{r}(\cdot)$ is continuously differentiable at the true value β and $\mathbf{R} = \frac{\partial}{\partial \beta}\mathbf{r}(\beta)'$ has rank q.

Theorem

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

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

where $\widehat{\theta} = \mathbf{r}\left(\widehat{\beta}\right)$, *and* $\mathbf{V}_\theta = \mathbf{R}'\mathbf{V}\mathbf{R}$.

Proof.

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

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

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

Estimation of $\mathbf{V}_\theta$

- A natural estimator of $\mathbf{V}_\theta$ is
$$\widehat{\mathbf{V}}_\theta = \widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{R}}, \tag{6}$$
where $\widehat{\mathbf{R}} = \partial \mathbf{r}\left(\widehat{\beta}\right)' / \partial \beta$. If $\mathbf{r}(\cdot)$ is a $C^{(1)}$ function, then by the CMT, $\widehat{\mathbf{V}}_\theta \xrightarrow{p} \mathbf{V}_\theta$ (why?).
- If $\mathbf{r}(\beta)$ is linear:
$$\mathbf{r}(\beta) = \mathbf{R}'\beta$$
for some $k \times q$ matrix $\mathbf{R}$. In this case, $\frac{\partial}{\partial \beta}\mathbf{r}(\beta)' = \mathbf{R}$ and $\widehat{\mathbf{R}} = \mathbf{R}$, so $\widehat{\mathbf{V}}_\theta = \mathbf{R}'\widehat{\mathbf{V}}\mathbf{R}$.
- 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 $\beta = (\beta_1', \beta_2')'$, then $\theta = \mathbf{R}'\beta = \beta_1$ and
$$\widehat{\mathbf{V}}_\theta = (\mathbf{I}, \mathbf{0})\widehat{\mathbf{V}}\begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} = \widehat{\mathbf{V}}_{11},$$
the upper-left block of $\widehat{\mathbf{V}}$.

The t Test

The Studentized Statistic

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

Under the assumptions of Theorem 5, $t_n(\theta) \xrightarrow{d} N(0,1)$.

- 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**.

[1] A standard error for an estimator is an estimate of the standard deviation of that estimator

The t Test

- The most common one-dimensional hypotheses are the null

$$H_0 : \theta = \theta_0, \tag{7}$$

  against the alternative

$$H_1 : \theta \neq \theta_0, \tag{8}$$

  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**.

Asymptotic Size and Asymptotic Critical Value

- The **asymptotic size** of the test is defined as the asymptotic probability of a Type I error:
$$\lim_{n\to\infty} P(|t_n| > c|H_0 \text{ true}) = P(|Z| > c) = 1 - \overline{\Phi}(c).$$
- The asymptotic size of the test is a simple function of the asymptotic null distribution G and the critical value c.
- c is called 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.

One-Sided Alternative

- The alternative hypothesis (8) is called a "two-sided" alternative.
- One-sided alternatives:

$$H_1: \theta > \theta_0 \tag{9}$$

or

$$H_1: \theta < \theta_0. \tag{10}$$

- Tests of (7) against (9) or (10) are based on the signed t-statistic t_n.
- The hypothesis (7) is rejected in favor of (9) if $t_n > c$ (why?) where c satisfies $\alpha = 1 - \Phi(c)$.
- The critical values are smaller than for two-sided tests (why?). Specifically, the asymptotic 5% critical value is $c = 1.645$. Thus, we reject (7) in favor of (9) if $t_n > 1.645$.
- 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 $\theta \geq \theta_0$. Since linear regression coefficients typically do not have a priori sign restrictions, we conclude that two-sided tests are generally appropriate.

p-Value

p-Value

- The rejection/acceptance dichotomy is associated with the Neyman-Pearson approach to hypothesis testing; p-value is associated with R.A. Fisher.
- Define the tail probability, or asymptotic p-value function

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

  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]
- **Caveat**: the p-value p_n should not be interpreted as the probability that either hypothesis is true. For example, p_n is NOT the probability "that the null hypothesis is false." Rather, p_n is a measure of the strength of information against the null hypothesis.

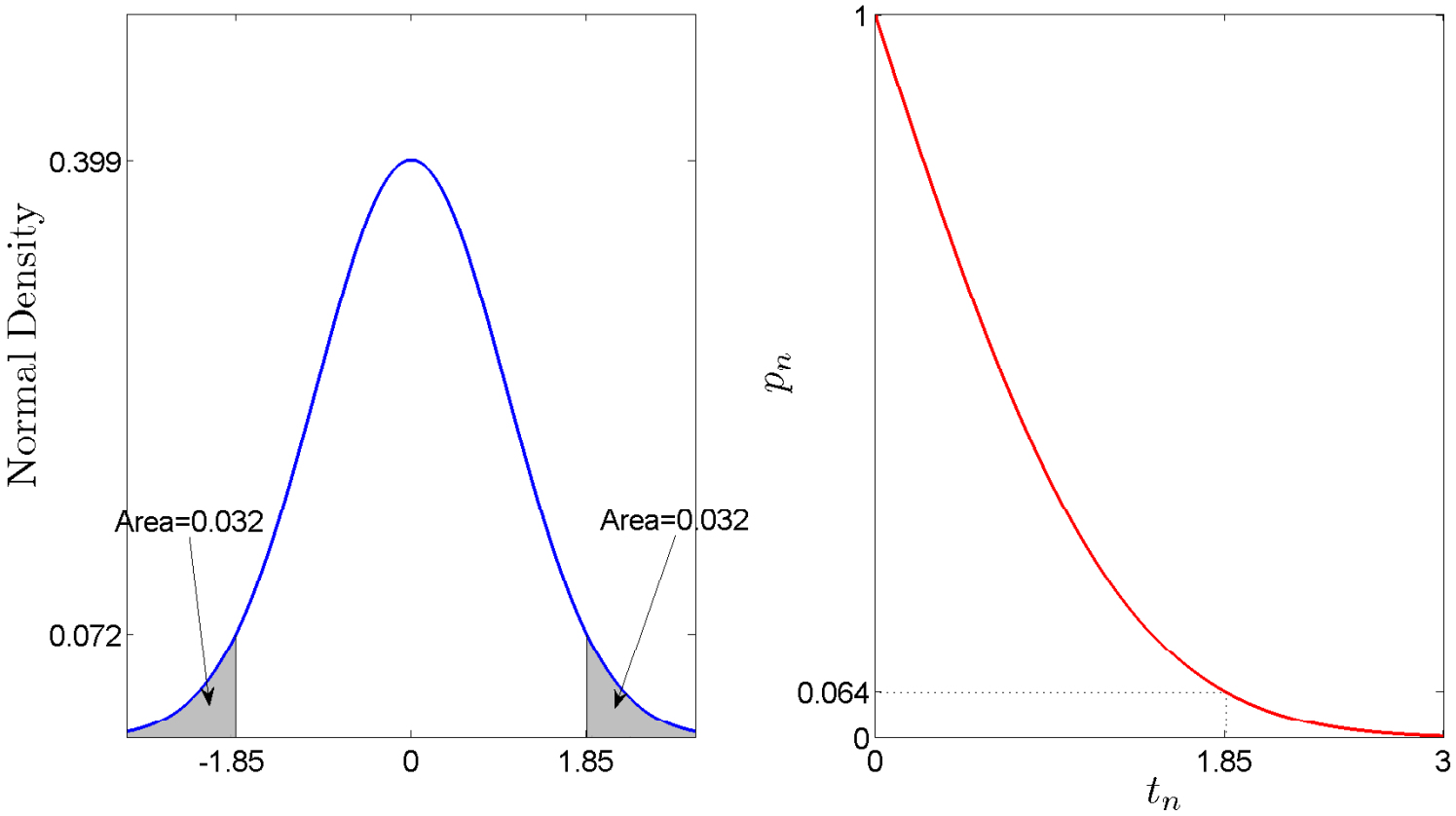

Figure: Obtaining the p-Value in a Two-Sided t-Test: $|t_n| = 1.85$

continue...

- An equivalent statement of a Neyman-Pearson test is to reject at the α level if and only if $p_n < \alpha$.
- In this sense, p-values and hypothesis tests are *equivalent* since $p_n < \alpha$ if and only if $|t_n| > z_{\alpha/2}$.
- 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.
- The p-value function has simply made a unit-free transformation of the test statistic. That is, under H_0, $p_n \xrightarrow{d} U[0,1]$, regardless of the complication of the distribution of the original test statistic.
- Why? The asymptotic distribution of $|t_n|$ is $G(x) = 1 - p(x)$. Thus

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

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

Confidence Interval

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 $n \to \infty$.

Test Statistic Inversion

- *test statistic inversion* method: collecting parameter values which are not rejected by a statistical test.
- The t-test rejects H_0: $\theta = \theta_0$ if $|t_n(\theta_0)| > z_{\alpha/2}$. A CI is then constructed using the values for which this test does not reject [Figure 2]:

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

- The most common professional choice for $1-\alpha$ is 95%, or $\alpha = .05$. This corresponds to selecting the CI $\left[\widehat{\theta} \pm 1.96 s\left(\widehat{\theta}\right)\right] \approx \left[\widehat{\theta} \pm 2 s\left(\widehat{\theta}\right)\right]$.
- The interval has been constructed so that as $n \to \infty$,

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

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

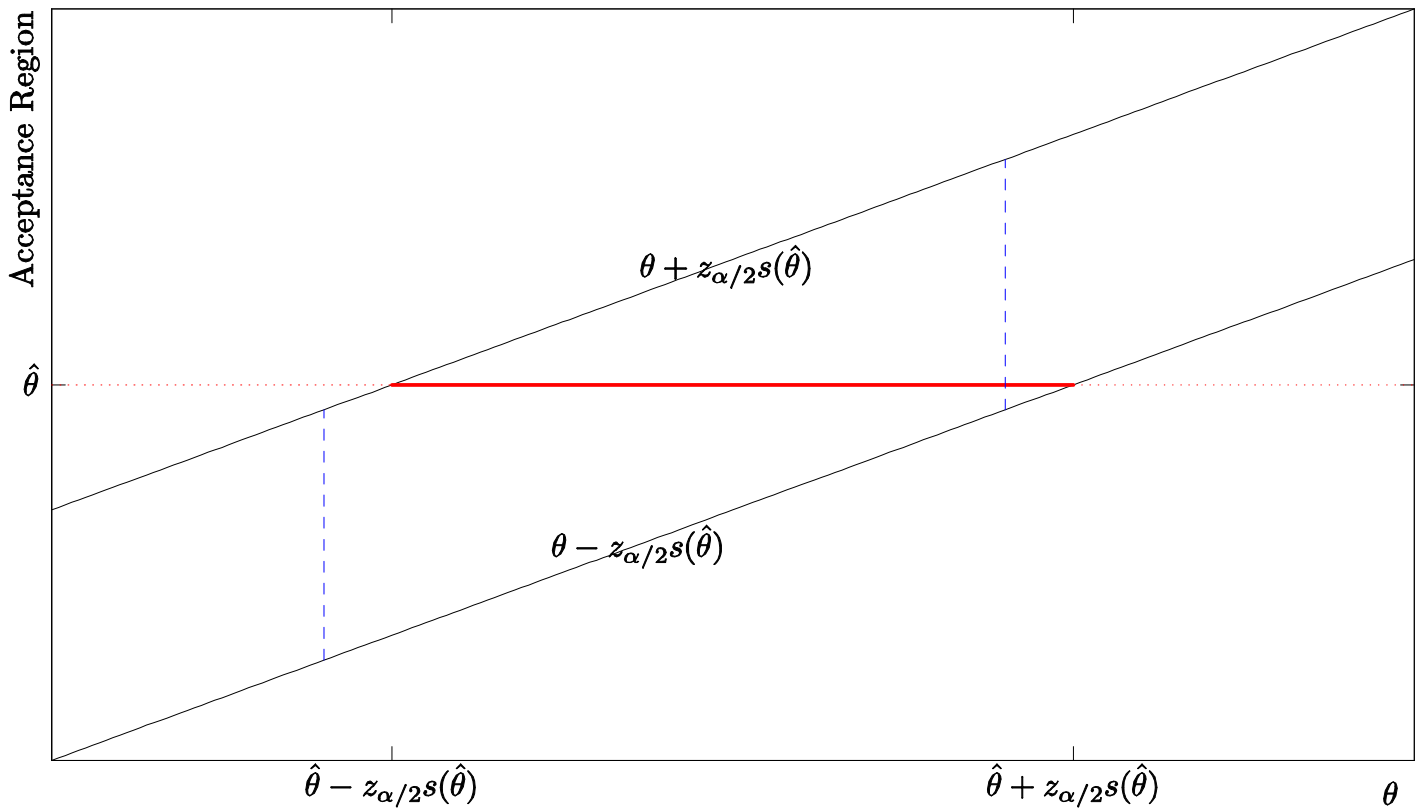

Figure: Test Statistic Inversion: acceptance region for $\widehat{\theta}$ at θ is $\left[\theta - z_{\alpha/2}s\left(\widehat{\theta}\right), \theta + z_{\alpha/2}s\left(\widehat{\theta}\right)\right]$

The Wald Test

The Wald Test

- When $\theta = \mathbf{r}(\beta)$ is a $q \times 1$ vector, 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

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0.$$

- The Wald statistic for H_0 against H_1 is

$$W_n = n\left(\widehat{\theta} - \theta_0\right)' \widehat{\mathbf{V}}_\theta^{-1} \left(\widehat{\theta} - \theta_0\right).$$

- We have known that $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_\theta)$, and $\widehat{\mathbf{V}}_\theta \xrightarrow{p} \mathbf{V}_\theta$ under H_0. So $W_n \xrightarrow{d} \chi_q^2$ under the null (why?).
- When $q = 1$, $W_n = t_n^2$ (check). 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 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.

Confidence Region

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

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

  where $W_n(\theta) = n\left(\widehat{\theta} - \theta\right)' \widehat{\mathbf{V}}_\theta^{-1}\left(\widehat{\theta} - \theta\right)$. Since $\widehat{\mathbf{V}}_\theta > 0$, $\mathbf{C}_n$ is an ellipsoid in the θ plane.
- Assume $q = 2$ and $\theta = (\beta_1, \beta_2)'$; then $\mathbf{C}_n$ is an ellipse in the (β_1, β_2) plane as shown in Figure 3.
- $\mathbf{C}_n' \equiv \text{CI}(\beta_1) \times \text{CI}(\beta_2)$ does not work! that is, $P((\beta_1, \beta_2) \in \mathbf{C}_n') \neq 1-\alpha$ in general.

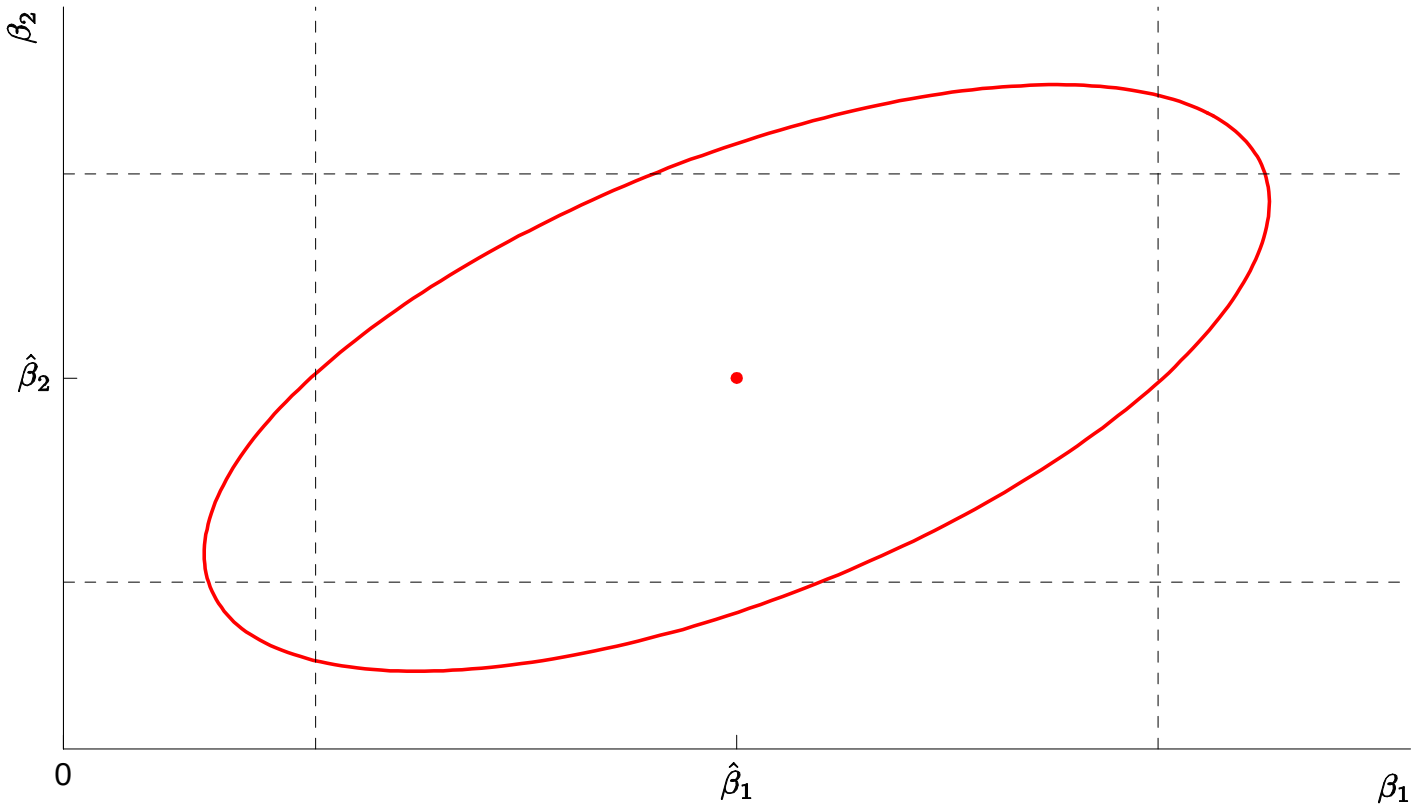

Figure: Confidence Region for (β_1, β_2)

Problems with Tests of Nonlinear Hypotheses

Problems with Tests of Nonlinear Hypotheses

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

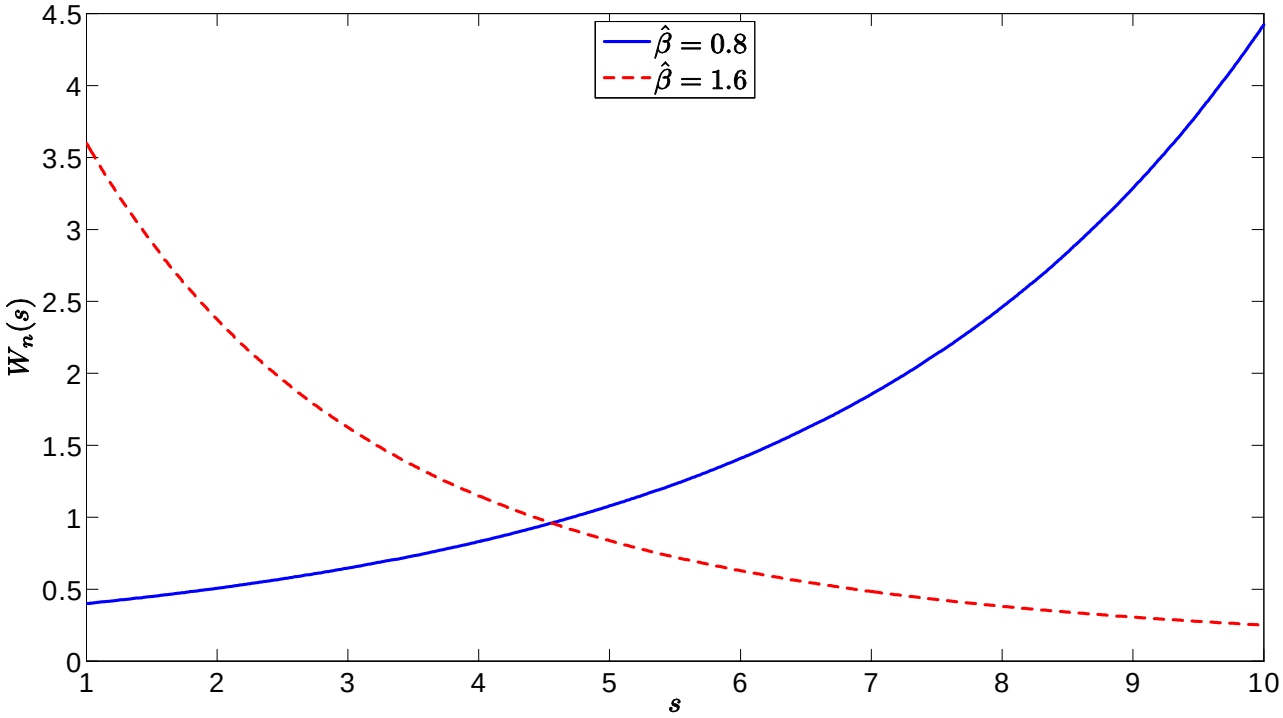

Figure: Wald Statistic as a function of s: $n/\widehat{\sigma}^2 = 10$

- 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.

Finite-Sample Distribution

- The 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, while **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 produces random draws from the statistic's sampling distribution. Through repetition, features of this distribution can be calculated.
- Let's focus on the Type I error of the test using the asymptotic 5% critical value 3.84 - the probability of a false rejection, $P(W_n(s) > 3.84|\beta = 1)$.
- This probability depends only on s, n, and σ^2 in this simple model.
- Table 1 reports the simulation estimate of the Type I error probability from $50,000$ random samples.
- The probabilities in Table 1 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

Results from Table 1

- 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; as s increases test performance deteriorates.
- Impact of variation in sample size n: 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.

A More Complicated Example

- Take the model
$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + u_i, E[\mathbf{x}_i u_i] = \mathbf{0} \tag{11}$$
and the hypothesis
$$H_0\colon \theta \equiv \frac{\beta_1}{\beta_2} = \theta_0.$$
- $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2\right)'$, $\widehat{\mathbf{V}}_{\widehat{\beta}} = \widehat{\mathbf{V}}/n$, and $\widehat{\theta} = \widehat{\beta}_1/\widehat{\beta}_2$.
- A t-statistic for H_0 is
$$t_{1n} = \frac{\widehat{\beta}_1/\widehat{\beta}_2 - \theta_0}{s(\widehat{\theta})}.$$
where $s(\widehat{\theta}) = \left(\widehat{\mathbf{R}}_1'\widehat{\mathbf{V}}_{\widehat{\beta}}\widehat{\mathbf{R}}_1\right)^{1/2}$ with $\widehat{\mathbf{R}}_1 = \left(0, \frac{1}{\widehat{\beta}_2}, -\frac{\widehat{\beta}_1}{\widehat{\beta}_2^2}\right)'$.
- An equivalent null hypothesis is
$$H_0\colon \beta_1 - \theta_0\beta_2 = 0.$$

continue...

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

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

  where $\mathbf{R}_2 = (0, 1, -\theta_0)'$.
- Monte Carlo Simulation: 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.
- Ideally, the entries in Table 2 should be 0.05. However, the rejection rates for the t_{1n} statistic diverge greatly from this value, especially for small values of 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

Solutions

- The common message from both examples is that Wald statistics are sensitive to the algebraic formulation of the null hypothesis.
- **Solution I**: 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, and alternatives to asymptotic critical values should be considered.
- **Solution II**: Consider alternative tests to the Wald statistic, such as the minimum distance statistic.

Test Consistency

Test Consistency

Definition

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

- Suppose that y_i is i.i.d. $N(\mu, 1)$. Consider the t-statistic $t_n(\mu) = \sqrt{n}(\overline{y} - \mu)$, 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
$$P(t_n > c|\theta) = P\left(Z + \sqrt{n}\mu > c\right) = 1 - \Phi\left(c - \sqrt{n}\mu\right).$$
This function is monotonically increasing in μ and n, and decreasing in c.
- 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 with probability approaching 1 as the sample size gets large.

continue...

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

$$t_n = \frac{\widehat{\theta} - \theta_0}{s\left(\widehat{\theta}\right)} = \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)} + \frac{\sqrt{n}\left(\theta - \theta_0\right)}{\sqrt{\widehat{V}_\theta}} \equiv \text{Term I} + \text{Term II} \tag{12}$$

since $s\left(\widehat{\theta}\right) = \sqrt{\widehat{V}_\theta / n}$.

- Term I $\xrightarrow{d} N(0,1)$. Term II $= 0$ if $\theta = \theta_0$, and $\xrightarrow{p} +\infty$ if $\theta > \theta_0$, and $\xrightarrow{p} -\infty$ if $\theta < \theta_0$. Thus both the two-sided t-test and one-sided t-test are consistent.
- For another example, The Wald statistic for H_0: $\theta = \mathbf{r}(\beta) = \theta_0$ against H_1: $\theta \neq \theta_0$ is

$$W_n = n\left(\widehat{\theta} - \theta_0\right)' \widehat{\mathbf{V}}_\theta^{-1} \left(\widehat{\theta} - \theta_0\right).$$

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

Asymptotic Local Power

Local Alternatives

- 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 use **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.
- In the t-test, we consider parameter vectors β_n which are indexed by sample size n and satisfy the real-valued relationship

$$\theta_n = r(\beta_n) = \theta_0 + n^{-1/2}h, \tag{13}$$

  where the scalar h is called a **localizing parameter**, and the sequence of local alternatives θ_n is called a **Pitman drift** or a **Pitman sequence**.
- We index β_n and θ_n by sample size to indicate their dependence on n. The way to think of (13) is that the *true* value of the parameters are β_n and θ_n. The parameter θ_n is close to the hypothesized value θ_0, with deviation $n^{-1/2}h$.

How to Understand Local Alternatives?

- The specification (13) 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.
- 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.
- The concept of a localizing sequence (13) might seem odd at first as in the actual world the sample size *cannot* mechanically affect the value of the parameter.
- Thus (13) 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.

Local Power Function

- Similarly as in (12),

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

  under the local alternative, where $Z \sim N(0,1)$ and $\delta = h/\sqrt{V_\theta}$.

- In testing the one-sided alternative $H_1\colon \theta > \theta_0$, 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,

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

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

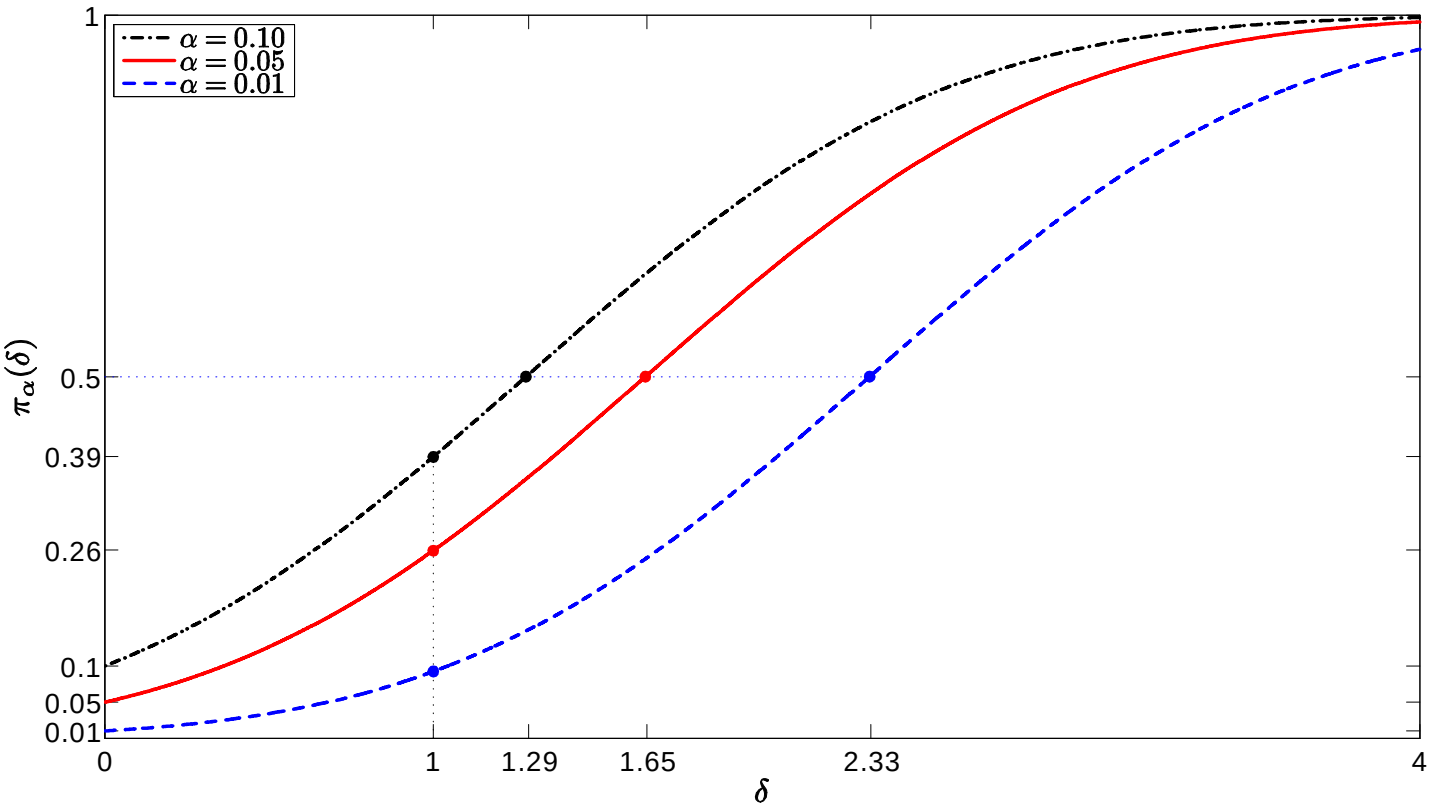

Figure: Asymptotic Local Power Function of One-Sided t-Test with Different Asymptotic Sizes

How to Read the Local Power Function?

- 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.
- The coefficient δ can be interpreted as the parameter deviation measured as a multiple of the standard error $s\left(\widehat{\theta}\right)$.
- Why? $s\left(\widehat{\theta}\right) = n^{-1/2}\sqrt{\widehat{V}_\theta} \approx n^{-1/2}\sqrt{V_\theta}$, so

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

- Thus in the figure, 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.

Read Vertically or Horizontally?

- In the figure 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.
- In the figure there is also 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.)
- 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.
- Why? $\delta = h/\sqrt{V_\theta} = \sqrt{n}(\theta_n - \theta_0)/\sqrt{V_\theta}$. Holding θ and V_θ fixed, $\delta^2 \sim n$.

Local Power in the Wald Test

- Local parametrization:
$$\theta_n = \mathbf{r}(\beta_n) = \theta_0 + n^{-1/2}\mathbf{h}, \tag{14}$$
where $\mathbf{h}$ is a $q \times 1$ vector.
- Under (14),
$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = \sqrt{n}\left(\widehat{\theta} - \theta_n\right) + \mathbf{h} \xrightarrow{d} \mathbf{Z_h} \sim N(\mathbf{h}, \mathbf{V}_\theta).$$
- Applied to the Wald statistic,
$$W_n = n\left(\widehat{\theta} - \theta_0\right)' \widehat{\mathbf{V}}_\theta^{-1}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathbf{Z}_\mathbf{h}' \mathbf{V}_\theta^{-1} \mathbf{Z_h} \sim \chi_q^2(\lambda).$$
where $\chi_q^2(\lambda)$ is a **non-central chi-square distribution** with q degrees of freedom and non-central parameter $\lambda = \mathbf{h}'\mathbf{V}_\theta^{-1}\mathbf{h}$.
- 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
$$P\left(\chi_q^2(\lambda) > \chi_{q,\alpha}^2\right) \equiv \pi_{q,\alpha}(\lambda).$$

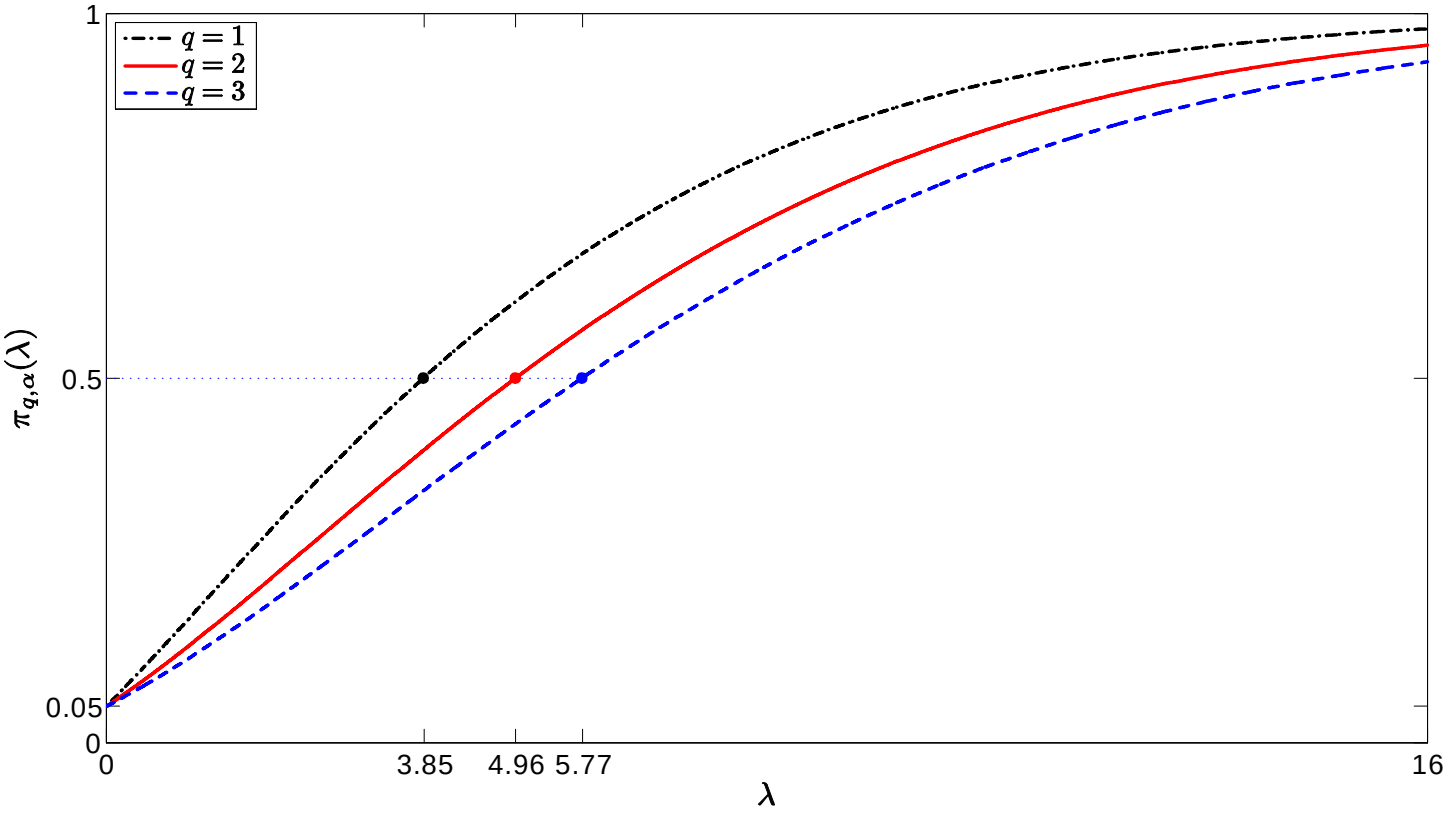

Figure: Asymptotic Local Power Function of the Wald Test

Read the Local Power Function

- The power functions are monotonically increasing in λ and asymptote to one.
- The figure 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 (why?). 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.
- Intuition: when testing more restrictions, we need more deviation from the null (or equivalently, more data points) to achieve the same power.