Large-Sample Tests

Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}$. The u_i are assumed IID. Regressors may be predetermined and/or exogenous.

The t statistic for the hypothesis $\beta_2 = \beta_2^0$ is

$$t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2^0}{\sqrt{s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}}} = \frac{N^{1/2}(\hat{\beta}_2 - \beta_2^0)}{\sqrt{s^2 N(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}}}, \tag{1}$$

where $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}$ is the last element on the main diagonal of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

Theorem 4.3 tells us that:

- $N^{1/2}(\hat{\beta}_2 - \beta_2^0)$ is asymptotically normal with variance the last element on the main diagonal of $\sigma_0^2(\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})^{-1}$;
- s^2 times $N(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}$ consistently estimates this variance.

Recall that, asymptotically, $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}} = N^{-1}\boldsymbol{X}^\top\boldsymbol{X}$.

We conclude that

$$t_{\beta_2} \overset{a}{\sim} \mathrm{N}(0,1), \tag{2}$$

where "$\overset{a}{\sim}$" means "**is asymptotically distributed as**."

The result (2) justifies the use of t tests outside the confines of the classical normal linear model.

We can compute asymptotic P values or critical values using either the standard normal or $t(N-k)$ distributions.

When the model is $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$, the usual F statistic is asymptotically valid.

Rewriting $F_{\boldsymbol{\beta}_2}$ (under the null) in terms of quantities that are $O_p(1)$,

$$F_{\boldsymbol{\beta}_2} = \frac{N^{-1/2}\boldsymbol{\epsilon}^\top \boldsymbol{M}_1\boldsymbol{X}_2(N^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{X}_2)^{-1}N^{-1/2}\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{\epsilon}/r}{\boldsymbol{\epsilon}^\top \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\epsilon}/(N-k)}, \tag{3}$$

where $\boldsymbol{\epsilon} \equiv \boldsymbol{u}/\sigma_0$ and $r = k_2$, the dimension of $\boldsymbol{\beta}_2$.

We now show that rF_{β_2} is asymptotically distributed as $\chi^2(r)$, which follows from Theorem 4.3.

The denominator of (3) is $\boldsymbol{\epsilon}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{\epsilon} / (N-k)$, which is just s^2/σ_0^2.

Since s^2 is consistent for σ_0^2, this denominator must tend to 1 asymptotically.

The numerator of rF_{β_2}, multiplied by r, is

$$N^{-1/2} \boldsymbol{\epsilon}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 (N^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1} N^{-1/2} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{\epsilon}. \tag{4}$$

Let $\boldsymbol{v} = N^{-1/2} \boldsymbol{X}^\top \boldsymbol{\epsilon}$. Then a CLT shows that $\boldsymbol{v} \overset{a}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}})$.

If we partition $\boldsymbol{v}$ into two subvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, we have

$$N^{-1/2} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{\epsilon} = \boldsymbol{v}_2 - N^{-1} \boldsymbol{X}_2^\top \boldsymbol{X}_1 (N^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{v}_1. \tag{5}$$

This tends to $\boldsymbol{v}_2 - \boldsymbol{S}_{21} \boldsymbol{S}_{11}^{-1} \boldsymbol{v}_1$ as $N \to \infty$. Here $\boldsymbol{S}_{11}$ and $\boldsymbol{S}_{21}$ are submatrices of $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$, with $N^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \to \boldsymbol{S}_{11}$ and $N^{-1} \boldsymbol{X}_2^\top \boldsymbol{X}_1 \to \boldsymbol{S}_{21}$.

Since v is asymptotically multivariate normal, so is $v_2 - S_{21}S_{11}^{-1}v_1$.

The vector $v_2 - S_{21}S_{11}^{-1}v_1$has mean $\mathbf{0}$, so its covariance matrix is the expectation of

$$v_2v_2^\top + S_{21}S_{11}^{-1}v_1v_1^\top S_{11}^{-1}S_{12} - 2v_2v_1^\top S_{11}^{-1}S_{12}. \tag{6}$$

Asymptotically, we can replace $v_1v_1^\top$, $v_2v_2^\top$, and $v_2v_1^\top$ by their plims, which are S_{11}, S_{22}, and S_{21}, respectively. To see this, observe that

$$\operatorname{plim}(v_1v_1^\top) = \operatorname{plim}(N^{-1}X_1^\top \epsilon\epsilon^\top X_1) = \operatorname{plim} N^{-1}(X_1^\top X_1) = S_{11}. \tag{7}$$

This allows us to conclude that

$$\operatorname{Var}(v_2 - S_{21}S_{11}^{-1}v_1) = S_{22} - S_{21}S_{11}^{-1}S_{12}. \tag{8}$$

Thus the numerator of the F statistic is asymptotically equal to

$$(v_2 - S_{21}S_{11}^{-1}v_1)^\top (S_{22} - S_{21}S_{11}^{-1}S_{12})^{-1}(v_2 - S_{21}S_{11}^{-1}v_1). \tag{9}$$

(9) is a quadratic form in the r-vector $\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1$, which is asymptotically multivariate normal, and the inverse of $\text{Var}(\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1)$, its covariance matrix.

By Theorem 4.1, expression (9) is asymptotically distributed as $\chi^2(r)$.

Because the denominator of the F statistic tends to 1, we conclude that

$$rF_{\boldsymbol{\beta}_2} \overset{a}{\sim} \chi^2(r) \tag{10}$$

under H_0 with predetermined regressors.

Since $1/r$ times a $\chi^2(r)$ random variable is distributed as $F(r, \infty)$, we may also conclude that $F_{\boldsymbol{\beta}_2} \overset{a}{\sim} F(r, N-k)$.

(10) justifies **asymptotic F tests** when disturbances are not normally distributed and some regressors may be predetermined.

We can use either the $\chi^2(r)$ or $F(r, N-k)$ distributions. If we use the $\chi^2(r)$ distribution, we have to multiply the F statistic by r.

Wald Tests

A vector of r linear restrictions on a parameter vector $\boldsymbol{\beta}$ can always be written as

$$\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}, \tag{11}$$

where $\boldsymbol{R}$ is an $r \times k$ matrix and $\boldsymbol{r}$ is an r-vector.

For example, if $k = 3$ and the restrictions were that $\beta_1 = 0$ and $\beta_2 = -1$, equations (11) would be

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}. \tag{12}$$

The elements of the matrix $\boldsymbol{R}$ and the vector $\boldsymbol{r}$ must be known. They are not functions of the data, and they are often integers.

Suppose that a k-vector of estimates $\hat{\boldsymbol{\beta}}$ has covariance matrix $\text{Var}(\hat{\boldsymbol{\beta}})$.

It is easy see that $\text{Var}(\boldsymbol{R}\hat{\boldsymbol{\beta}})$ must be $\boldsymbol{R}\,\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top$.

We can test the restrictions (11) by calculating the **Wald statistic**

$$W(\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \big(\boldsymbol{R}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \tag{13}$$

where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ estimates $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ consistently.

Inserting appropriate powers of N, we can rewrite (13) as

$$W(\hat{\boldsymbol{\beta}}) = \big(N^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\big)^\top \big(\boldsymbol{R}N\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top\big)^{-1}\big(N^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\big). \tag{14}$$

By Theorem 4.3, $\mathrm{N}^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ is asymptotically multivariate normal.

Therefore, we can apply Theorem 4.1 (asymptotically) to (14) and conclude that $W(\hat{\boldsymbol{\beta}}) \overset{a}{\sim} \chi^2(r)$ under the null hypothesis.

These results are not just for OLS. Equations (13) and (14) would still define a Wald statistic if $\hat{\boldsymbol{\beta}}$ were any root-N consistent estimator and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ were any valid estimator of its covariance matrix.

Properties of Asymptotic Tests

Asymptotic tests are almost never exact in finite samples.

They may either **over-reject** (reject a true null more than $100\alpha\%$ of the time) or **under-reject** (reject it less than $100\alpha\%$ of the time).

Whether an asymptotic test over-rejects or under-rejects, and how severely, depends on many things, including:

- the sample size;
- the distribution of the disturbances;
- the number of explanatory variables and their properties;
- the number of restrictions;
- the covariance matrix estimator employed, which may depend on assumptions about the disturbances;

Wald tests of many restrictions are particularly problematic. They often over-reject severely.

Performing Several Hypothesis Tests

When we perform two or more tests, we are engaged in **multiple testing**. P values cannot be interpreted in the usual way.

An unusually large test statistic (unusually small P value) is more likely to be obtained by accident when several tests are performed.

Suppose we perform m exact tests at level α. Let the **familywise error rate** α_m be the probability that at least one of the tests rejects.

For independent tests, the familywise error rate is one minus the probability that none of the tests rejects:

$$\alpha_m = 1 - (1 - \alpha)^m. \tag{15}$$

When m is large, α_m can be much larger than α. For example, if $\alpha = 0.05$, then $\alpha_2 = 0.0975$, $\alpha_4 = 0.18549$, $\alpha_8 = 0.33658$, and $\alpha_{16} = 0.55987$.

The simplest method to control familywise error rate is the **Bonferroni procedure**. It rejects whenever the smallest P value is less than α/m.

This procedure is based on the **Bonferroni inequality**

$$\Pr\left(\cup_{j=1}^{m}\left(P_j \leq \alpha/m\right)\right) \leq \alpha, \tag{16}$$

where P_j is the P value for the j^{th} test.

For large values of m, α/m is very much smaller than α.

It is even smaller than the value α that solves (15) for a given α_m, which would be appropriate if all the tests were independent.

When the P_j are positively correlated, α/m can be much too small.

If there is perfect dependence, all the tests yield identical P values, and the familywise error rate is just α.

One way to control the familywise error rate is to use the bootstrap.

The **Simes procedure** is less conservative than the Bonferroni one.

Order the P values from the smallest, $P_{(1)}$, to the largest, $P_{(m)}$. Then reject the joint null hypothesis whenever

$$P_{(j)} \leq j\alpha/m \text{ for any } j = 1, \ldots, m, \tag{17}$$

where α is the desired familywise error rate.

If $P_{(1)} < \alpha/m$, Simes and Bonferroni both reject.

But Simes can also reject when the second-smallest P value is less than $2\alpha/m$, the third-smallest is less than $3\alpha/m$, and so on. It must reject if the largest P value is less than α, i.e., if every one of the tests rejects.

When many coefficients in the same regression are of interest, or when we estimate many different regressions with just one coefficient of interest in each of them, m can be large.

In either case, a single "large" t statistic, with a small P value, may not mean very much.

The Power of Hypothesis Tests

An F statistic for $\boldsymbol{\beta}_2 = \mathbf{0}$ is the ratio of $\|\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{y}\|^2$ to $\|\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}\|^2$, each divided by its appropriate number of degrees of freedom.

For the classical normal linear model, under both H_0 and H_1, the numerator and denominator of this ratio are independent χ^2 variables, divided by their degrees of freedom.

Under either hypothesis, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$. The difference in distribution under H_0 and H_1 must come from the numerator alone.

The numerator of F, times r/σ^2, is

$$\frac{1}{\sigma^2}\boldsymbol{y}^\top \boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{y}. \tag{18}$$

The vector $\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{y}$ is normal under both hypotheses.

Its mean is $\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2$, which vanishes when $\boldsymbol{\beta}_2 = \mathbf{0}$, and its covariance matrix is $\sigma^2\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{X}_2$.

Under the alternative, (18) follows the **noncentral chi-squared distribution**.

- If the m-vector $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, then $\|\boldsymbol{z}\|^2 = \boldsymbol{z}^\top \boldsymbol{z}$ is distributed as (central) chi-squared with m degrees of freedom.
- Similarly, if $\boldsymbol{x} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$, then $\boldsymbol{x}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{x} \sim \chi^2(m)$.
- If instead $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$, then $\boldsymbol{z}^\top \boldsymbol{z}$ follows the noncentral chi-squared distribution with m degrees of freedom and **noncentrality parameter**, or **NCP**, $\Lambda \equiv \boldsymbol{\mu}^\top \boldsymbol{\mu}$.
- This distribution is written as $\chi^2(m, \Lambda)$. The expectation of $\boldsymbol{z}^\top \boldsymbol{z}$ is $m + \Lambda$.
- When $\boldsymbol{\mu} = \mathbf{0}$, the $\chi^2(m, 0)$ distribution is just the central $\chi^2(m)$ distribution.

If $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\nu}, \boldsymbol{\Omega})$, then $\boldsymbol{x}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{x} \sim \chi^2(m, \boldsymbol{\nu}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\nu})$. The distribution depends on $\boldsymbol{\nu}$ and $\boldsymbol{\Omega}$ only through the NCP, $\Lambda = \boldsymbol{\nu}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\nu}$.

In general, (18) is noncentral chi-squared, with r degrees of freedom and NCP

$$\Lambda \equiv \frac{1}{\sigma^2} \boldsymbol{\beta}_2^\top \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 \tag{19}$$

$$= \frac{1}{\sigma^2} \boldsymbol{\beta}_2^\top \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2. \tag{20}$$

Under the null hypothesis, $\Lambda = 0$.

Under either hypothesis, the distribution of the denominator of the F statistic, divided by σ^2, is central $\chi^2(N-k)$, and it is independent of the numerator.

The F statistic is thus the ratio of two random variables, divided by their degrees of freedom. The numerator is noncentral $\chi^2(r)$, and the denominator is central $\chi^2(N-k)$.

Under the normality assumption, these two random variables are independent of each other

We can write the distribution of the F statistic as

$$\frac{\chi^2(r, \Lambda)/r}{\chi^2(N-k)/(N-k)}, \tag{21}$$

with numerator and denominator mutually independent.

This is called the (singly) **noncentral F distribution**, with r and $N-k$ degrees of freedom and NCP Λ.

- The difference between the distributions of the F statistic under H_0 and H_1 depends only on Λ.
- The numerator of the F statistic in (18), multiplied by r/σ^2, is (perhaps only asymptotically) distributed as $\chi^2(r, \Lambda)$. The NCP Λ depends on σ, $\boldsymbol{X}$, and $\boldsymbol{\beta}_2$.
- As Λ increases, the distribution of the asymptotic test statistic moves to the right and becomes more spread out, causing the power of the test to increase. Power also depends on r.

Under the alternative, (18) is equal to

$$\frac{1}{\sigma^2}\boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{u} \tag{22}$$

$$+ \frac{2}{\sigma^2}\boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \frac{1}{\sigma^2}\boldsymbol{\beta}_2^\top \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2. \tag{23}$$

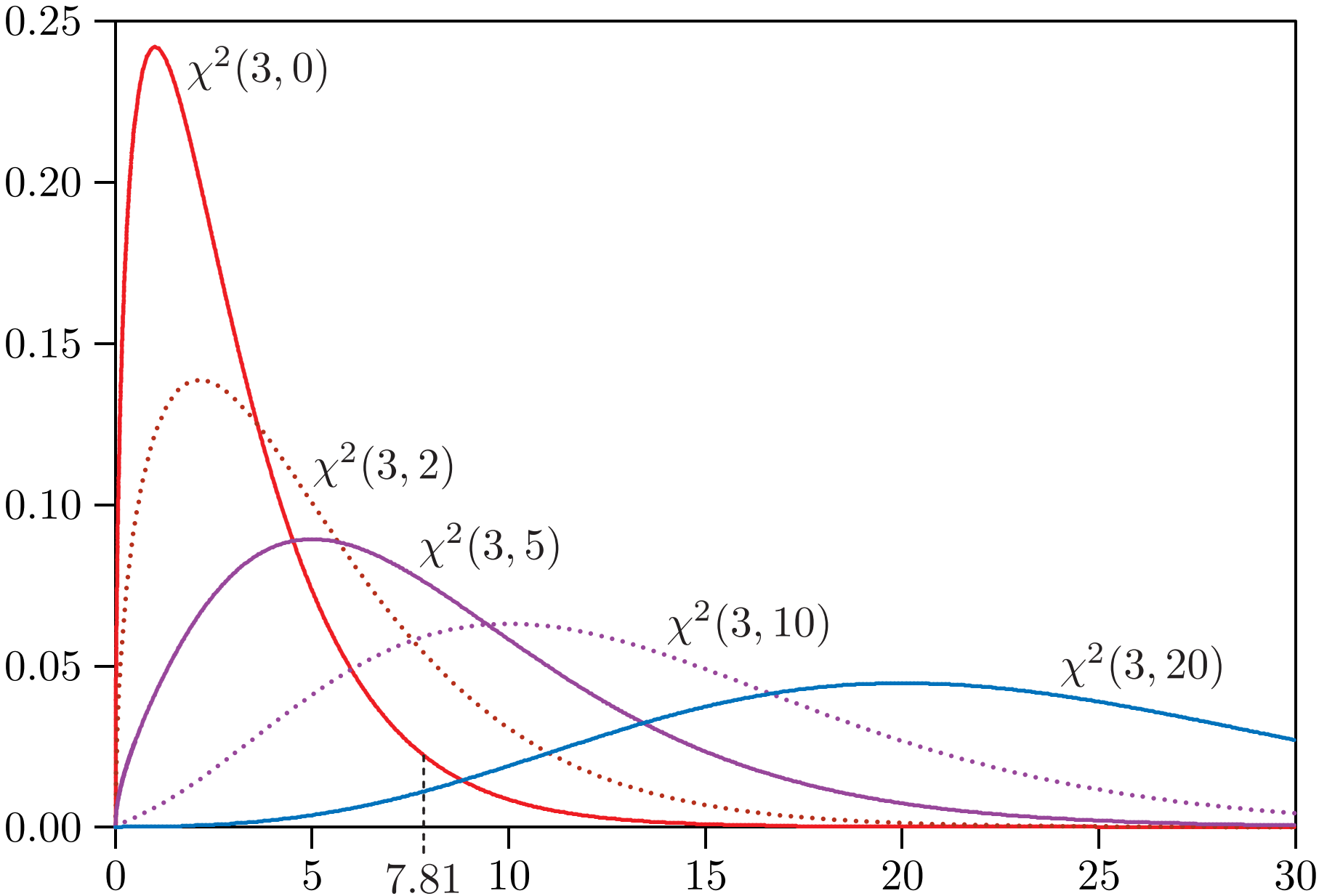
The first term follows the central $\chi^2(r)$ distribution.

The second term is a random scalar which is normally distributed with mean zero and variance 4Λ. It causes the $\chi^2(r, \Lambda)$ distribution to become more spread out as Λ increases.

The third term is the NCP, Λ. It is what causes the $\chi^2(r, \Lambda)$ distribution to move to the right as Λ increases.

Thus, informally, (18) equals $\chi^2(r) + \mathrm{N}(0, 4\Lambda) + \Lambda$.

The next figure shows the density of the $\chi^2(3, \Lambda)$ distribution for noncentrality parameters of 0, 2, 5, 10, and 20.


0.25
0.20
0.15
0.10
0.05
0.00
$\chi^2(3, 0)$
$\chi^2(3, 2)$
$\chi^2(3, 5)$
$\chi^2(3, 10)$
$\chi^2(3, 20)$
0
5
7.81
10
15
20
25
30

The signed square root of an F statistic for a single restriction is the t statistic:

$$t_{\beta_2} = \frac{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{s(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}}. \tag{24}$$

The numerator, $\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}$, is normally distributed under both H_0 and H_1, with variance $\sigma^2 \boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2$ and mean $\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2 \beta_2$.

Thus s/σ times (24) is normal with variance 1 and mean

$$\lambda \equiv \frac{1}{\sigma}(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2} \beta_2. \tag{25}$$

It follows that t_{β_2} has a distribution which can be written as

$$\frac{\mathrm{N}(\lambda, 1)}{\left(\chi^2(N-k)/(N-k)\right)^{1/2}}, \tag{26}$$

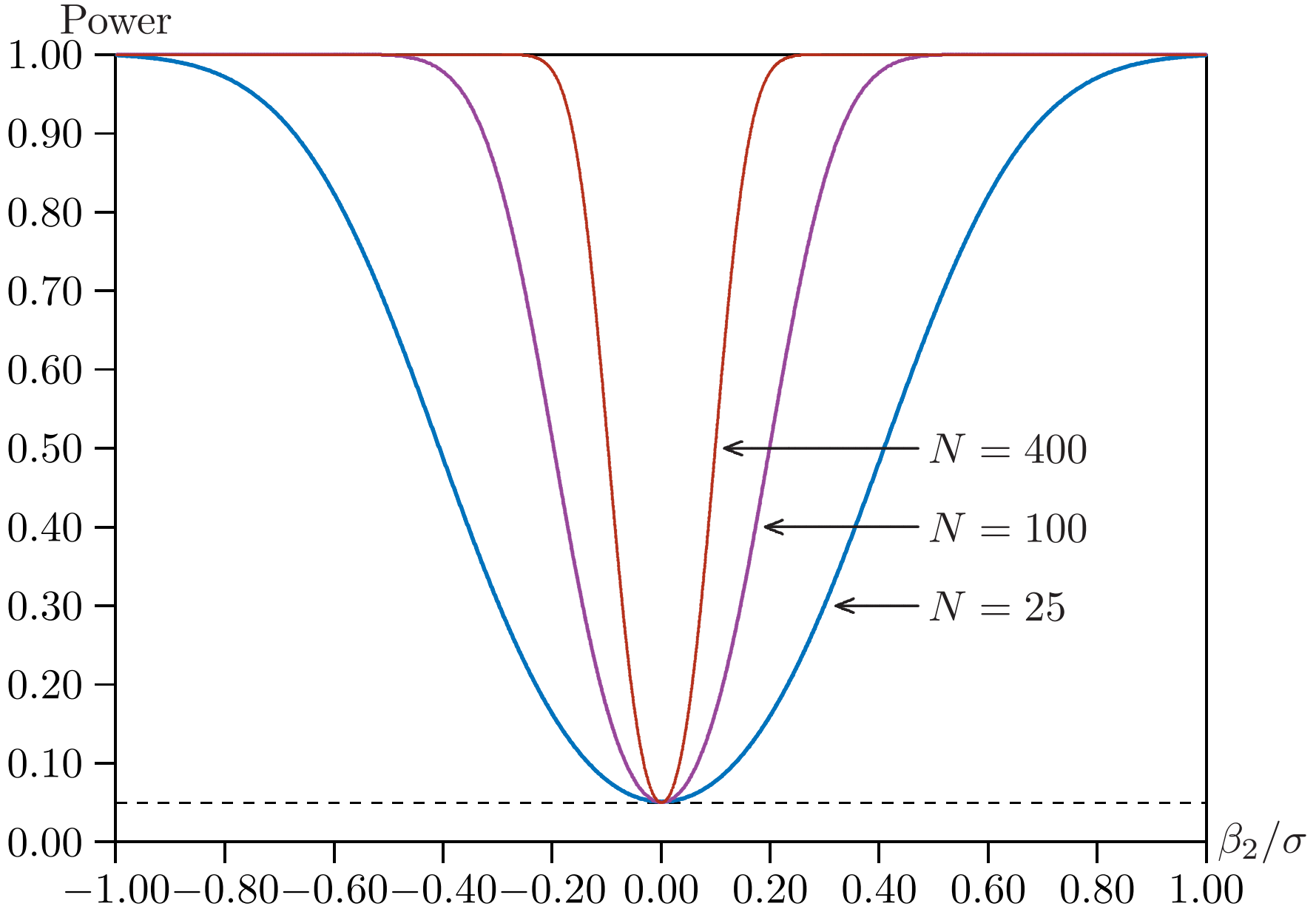
with independent numerator and denominator.

The distribution (26) is the **noncentral t distribution**, with $N-k$ degrees of freedom and noncentrality parameter λ; it is written as $t(N-k,\lambda)$.

- $\lambda^2 = \Lambda$, where Λ is the NCP of the corresponding F statistic.
- The $t(N-k,\lambda)$ distribution is similar to the $\mathrm{N}(\lambda,1)$ distribution.
- It is also like an ordinary (**central**) t distribution with its mean shifted from the origin to (25), but it has a bit more variance.
- The distribution of t_{β_2} under H_1 depends only on the NCP λ.
- For a given regressor matrix $\mathbf{X}$ and sample size N, λ in turn depends on the parameters only through the ratio β_2/σ; see (25).
- Therefore, the power of the t test depends on β_2/σ.

The **power function** of a test gives the power of the test at any given level as a function of the parameters of the alternative hypothesis.

The next figure shows power functions for tests at the .05 level. The only regressor, x_2, is a constant.


Power
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−1.00
−0.80
−0.60
−0.40
−0.20
0.00
0.20
0.40
0.60
0.80
1.00
β_2/σ
$N = 400$
$N = 100$
$N = 25$

- Since the test is exact, all the power functions equal .05 when $\beta_2 = 0$. Power then increases as β_2 moves away from 0.
- Power when $N = 400$ exceeds power when $N = 100$, which in turn exceeds power when $N = 25$, for every value of $\beta_2 \neq 0$.
- As $N \to \infty$, the power function converges to a T, with the foot of the vertical segment at .05 and the horizontal segment at 1.0.
- Asymptotically, the test rejects with probability 1 whenever the null is false. In finite samples, however, $N^{1/2}\beta_2/\sigma$ must be sufficiently large for rejection to be likely.

In finite samples, asymptotic tests may have power functions that do not look like the figure.

- Power may be greater or less than .05 when H_0 holds, and it may be minimized at a value that does not correspond to the null.
- Power may not even tend to unity as the parameter under test becomes infinitely far from the null hypothesis.

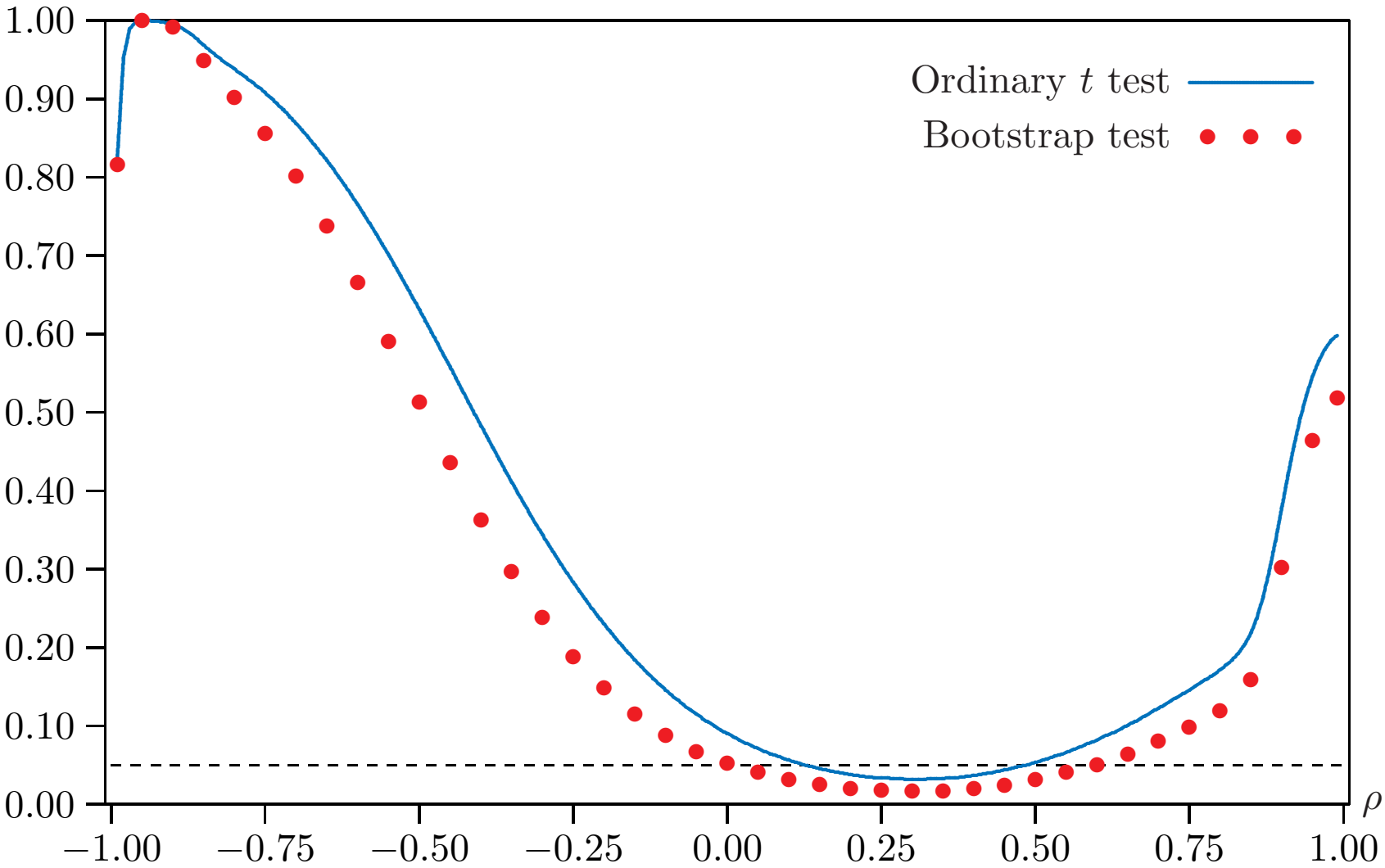
As an example of a power function that does not look the way asymptotic theory suggests, consider the model

$$y_t = \beta_1 + \sum_{j=2}^{4} \beta_j X_{tj} + \delta y_{t-1} + u_t, \quad u_t = \rho u_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathrm{N}(0, \sigma^2), \quad (27)$$

where $T = 20$, the X_{tj} follow an AR(1) process with coefficient 0.75, all the β_j are equal to 1, $\sigma = 0.1$, and $\delta = 0.9$.

The next figure shows the power of asymptotic t and bootstrap tests for $\rho = 0$ as a function of ρ, which varies between -0.99 and 0.99.

- For neither test does the power function achieve its minimum at $\rho = 0$ or increase monotonically as $|\rho|$ increases.
- Power actually declines sharply as ρ approaches -1.
- The asymptotic test has less power for values of ρ between 0 and about 0.62 than it does for $\rho = 0$.
- The bootstrap test rejects less than 5% of the time for values between 0 and about 0.61.


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
ρ
Ordinary t test
Bootstrap test

Pre-Testing

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}). \tag{28}$$

Here $\boldsymbol{\beta}$ is a k-vector, γ is an r-vector, and the regressors in $\boldsymbol{X}$ and $\boldsymbol{Z}$ are assumed, for simplicity, to be exogenous.

The parameters of interest are the k elements of $\boldsymbol{\beta}$. We do not care about γ. They are **nuisance parameters**.

The unrestricted OLS estimator of $\boldsymbol{\beta}$ and its covariance matrix are

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y} \quad \text{and} \quad \text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1}. \tag{29}$$

Similarly, the restricted OLS estimator and its covariance matrix are

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad \text{and} \quad \text{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{30}$$

Except when $\boldsymbol{X}$ and $\boldsymbol{Z}$ are orthogonal, $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$.

However, because $\tilde{\boldsymbol{\beta}}$ is biased if $\boldsymbol{\gamma} \neq \mathbf{0}$, its MSE matrix is larger than $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$ in that case.

$$\mathrm{MSE}(\tilde{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\boldsymbol{Z}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{31}$$

which is $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$ plus the bias vector times itself transposed.

Since $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\gamma}$ is zero, it seems natural to test whether $\boldsymbol{\gamma} = \mathbf{0}$ and use $\hat{\boldsymbol{\beta}}$ when the test rejects and $\tilde{\boldsymbol{\beta}}$ when it does not.

This test is called a **preliminary test**, or **pre-test** for short. It implicitly defines a new estimator, which is called a **pre-test estimator**:

$$\acute{\boldsymbol{\beta}} = \mathbb{I}(F_{\boldsymbol{\gamma}=\mathbf{0}} > c_\alpha)\hat{\boldsymbol{\beta}} + \mathbb{I}(F_{\boldsymbol{\gamma}=\mathbf{0}} \leq c_\alpha)\tilde{\boldsymbol{\beta}}, \tag{32}$$

where $F_{\boldsymbol{\gamma}=\mathbf{0}}$ is the F statistic for $\boldsymbol{\gamma} = \mathbf{0}$, and c_α is the critical value for an F test with r and $N - k - r$ degrees of freedom at level α.

Properties of Pre-Test Estimators

Consider the model

$$\boldsymbol{y} = \beta \boldsymbol{x} + \gamma \boldsymbol{z} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{NID}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{33}$$

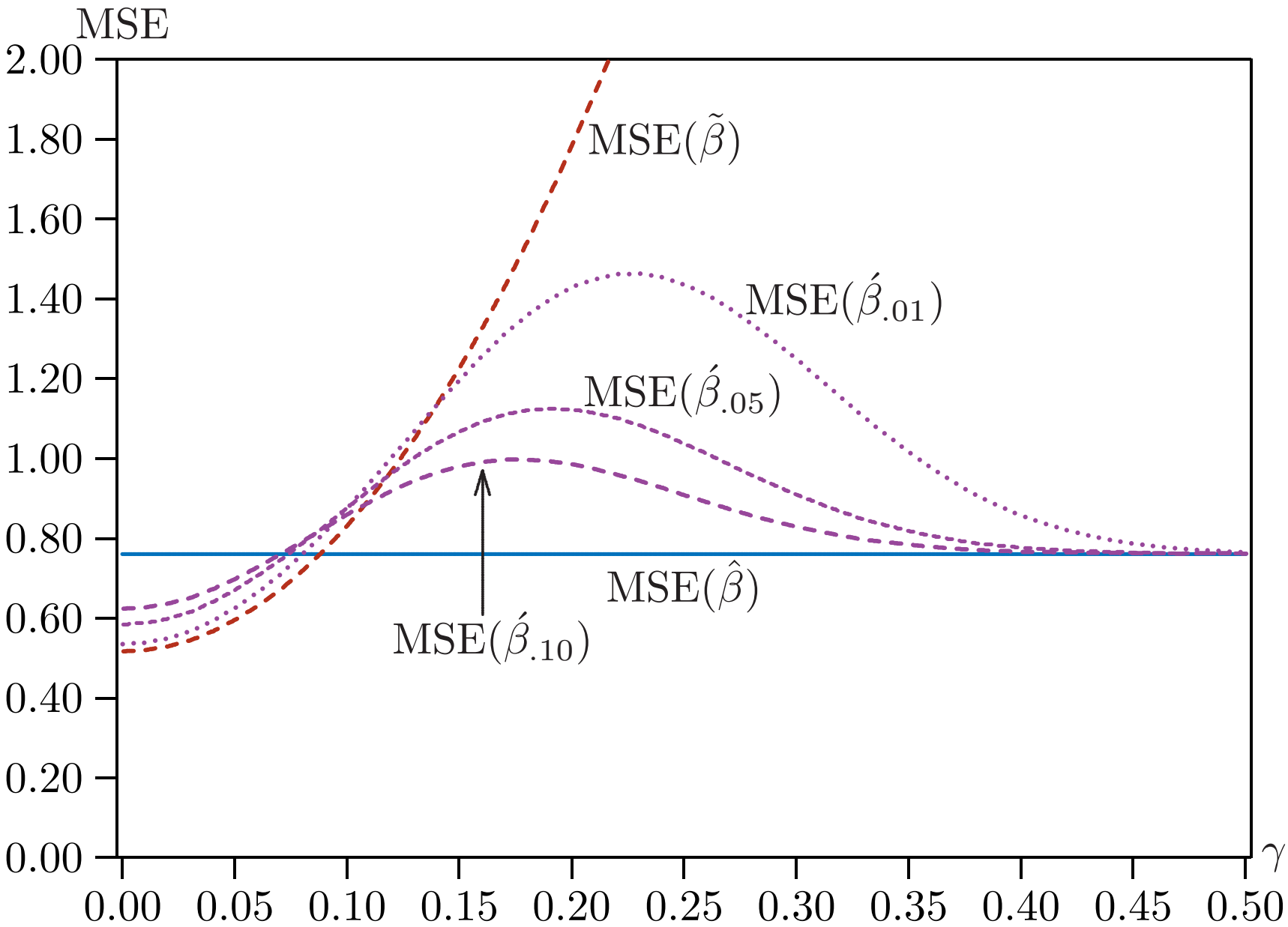
for which MSE is a scalar rather than a matrix

The two regressors are bivariate normal with correlation $\rho = 0.5$.

The potential reduction in variance from using $\tilde{\beta}$ or $\acute{\beta}$ rather than $\hat{\beta}$ is evidently increasing in $|\rho|$, but so is the potential bias.

The figure shows MSE for five estimators of β as functions of γ. They are $\hat{\beta}$, $\tilde{\beta}$, and three pre-test estimators, $\acute{\beta}_{.01}$, $\acute{\beta}_{.05}$, and $\acute{\beta}_{.10}$.

The figure would look different if the NCP of the test were on the horizontal axis, but since the NCP is proportional to γ^2, both figures would contain the same information.


MSE
2.00
1.80
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
γ
MSE($\tilde{\beta}$)
MSE($\acute{\beta}_{.01}$)
MSE($\acute{\beta}_{.05}$)
MSE($\hat{\beta}$)
MSE($\acute{\beta}_{.10}$)

The MSE of $\hat{\beta}$ is a horizontal line. It does not depend on γ.

MSE($\tilde{\beta}$) is lower than MSE($\hat{\beta}$) when γ is sufficiently small, but it increases in proportion to γ^2 (i.e., in proportion to the NCP).

- For small values of γ, the pre-test estimators have smaller MSE than $\hat{\beta}$ but greater MSE than $\tilde{\beta}$. For very large values of γ, they perform essentially the same as $\hat{\beta}$.
- There is a large region in the middle of the figure where the pre-test estimators perform better than $\tilde{\beta}$ but less well than $\hat{\beta}$.
- Near the point where MSE($\tilde{\beta}$) crosses MSE($\hat{\beta}$), the pre-test estimator performs worse than either $\tilde{\beta}$ or $\hat{\beta}$.

The level of the pre-test is important. When the level is high, the potential gain in efficiency for small values of γ is small, but so is the potential increase in MSE due to bias for intermediate values.

There is no reason to use a "conventional" significance level like .05 when pretesting. It is probably safer to use a higher level.