Chapter 9
Hypothesis Testing

9.1 Hypotheses

In Chapter 8 we discussed estimation subject to restrictions, including **linear** restrictions (8.1), **nonlinear** restrictions (8.44), and **inequality** restrictions (8.49).

In this chapter we discuss tests of such restrictions. Hypothesis tests attempt to assess whether there is evidence to contradict a proposed parametric restriction. Let

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

be a q-by-1 parameter of interest where

$$\boldsymbol{r} : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$$

is some transformation.

For example, theta may be a single coefficient, e.g. theta = beta_j, the different between two coefficients, e.g. theta = beta_j – beta_l, or the ratio of two coefficients, e.g. theta = beta_j/beta_l.

A **point hypothesis** concerning theta is a proposed restriction such as

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0$$

where theta_0 is a hypothesized (known) value.

More generally, letting

$$\boldsymbol{\beta} \in \boldsymbol{B} \subset \mathbb{R}^k$$

be the parameter space, a hypothesis is a restriction

$$\boldsymbol{\beta} \in \boldsymbol{B}_0$$

Where B_0 is a proper subset of B. This specializes to (9.1) by setting

$$\boldsymbol{B}_0 = \{\boldsymbol{\beta} \in \boldsymbol{B} : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0\}$$

In this chapter we will focus exclusively on **point hypotheses** of the form (9.1) as they are the most common and relatively simple to handle.

The hypothesis to be tested is called the **null hypothesis**.

Definition 9.1 The **null hypothesis**, written $\mathbb{H}_0$, is the restriction $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ or $\boldsymbol{\beta} \in \boldsymbol{B}_0$.

We often write the null hypothesis as

$$\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ or } \mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$$

The complement of the null hypothesis (the collection of parameter values which do not satisfy the null hypothesis) is called the **alternative hypothesis**.

Definition 9.2 The **alternative hypothesis**, written $\mathbb{H}_1$, is the set $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0\}$ or $\{\boldsymbol{\beta} \in \boldsymbol{B} : \boldsymbol{\beta} \notin \boldsymbol{B}_0\}$.

We often write the alternative hypothesis as

$$\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \text{ or } \mathbb{H}_1 : \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$$

For simplicity, we often refer to the hypotheses as “the null” and “the alternative”.

In hypothesis testing, we assume that there is a true (but unknown) value of theta and this value either satisfies H_0 or does not satisfy H_0: The goal of hypothesis testing is to assess whether or not H_0 is true, by asking if H_0 is consistent with the observed data.

To be specific, take our example of wage determination and consider the question: Does union membership affect wages? We can turn this into a hypothesis test by specifying the null as the restriction that a coefficient on union membership is zero in a wage regression.

Consider, for example, the estimates reported in Table 4.1. The coefficient for “Male Union Member” is 0.095 (a wage premium of 9.5%) and the coefficient for “Female Union Member” is 0.022 (a wage premium of 2.2%). These are estimates, not the true values. The question is: Are the true coefficients zero? To answer this question, the testing method asks the question: Are the observed estimates compatible with the hypothesis, in the sense that the deviation from the hypothesis can be reasonably explained by stochastic variation? Or are the observed estimates incompatible with the hypothesis, in the sense that that the observed estimates would be highly unlikely if the hypothesis were true?

9.2 Acceptance and Rejection

A hypothesis test either **accepts** the null hypothesis or **rejects** the null hypothesis in favor of the alternative hypothesis. We can describe these two decisions as “Accept H_0” and “Reject H_0”.

In the example given in the previous section, the decision would be either to accept the hypothesis that union membership does not affect wages, or to reject the hypothesis in favor of the alternative that union membership does affect wages.

The decision is based on the data, and so **is a mapping from the sample space to the decision set**. **This splits the sample space into two regions S_0 and S_1** such that if the observed sample falls into S_0 we accept H_0, while if the sample falls into S1 we reject H_0. The set S_0 is called the **acceptance region** and the set S1 the **rejection or critical region**.

It is convenient to express this mapping as a real-valued function called a **test statistic**

$$T = T\left(\left(y_1, \boldsymbol{x}_1\right), ..., \left(y_n, \boldsymbol{x}_n\right)\right)$$

relative to a **critical value** c.

The hypothesis test then consists of the decision rule

1. Accept $\mathbb{H}_0$ if $T \leq c$.
2. Reject $\mathbb{H}_0$ if $T > c$.

A test statistic T should be designed so that **small values are likely when H_0 is true** and **large values are likely when H_1 is true**.

There is a well-developed statistical theory concerning **the design of optimal tests**. We will not review that theory here, but instead refer the reader to Lehmann and Romano (2005). In this chapter we will summarize the main approaches to **the design of test statistics**.

The most commonly used test statistic is **the absolute value of the t-statistic**

$$T = |T(\theta_0)|$$

$$T(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})}$$

is the t-statistic from (7.33), theta^ is a point estimate, and s(theta^) its standard error.

T is an appropriate statistic when testing hypotheses on individual coefficients or real-valued parameters theta = h(beta) and theta_0 is the hypothesized value.

Quite typically, theta_0 = 0; as interest focuses on whether or not a coefficient equals zero, but this is not the only possibility. For example, interest may focus on whether an elasticity theta equals 1, in which case we may wish to test H_0: theta = 1.

9.3 Type I Error

A false rejection of the null hypothesis H_0 (**rejecting H_0 when H_0 is true**) is called a Type I error. The probability of a Type I error is

$$\mathbb{P}\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ true}\right) = \mathbb{P}\left(T > c \mid \mathbb{H}_0 \text{ true}\right)$$

The finite sample **size** of the test is defined as **the supremum of (9.4) across all data distributions which satisfy H_0**.

A primary goal of test construction is to limit the incidence of Type I error by bounding the size of the test.

For the reasons discussed in Chapter 7, in typical econometric models **the exact sampling distributions of estimators and test statistics are unknown** and hence we cannot explicitly calculate (9.4).

Instead, **we typically rely on asymptotic approximations**. Suppose that the test statistic has an **asymptotic distribution under H_0**.

That is, when H_0 is true

$$T \xrightarrow{d} \xi$$

as n → infinity for some continuously-distributed random variable xi. This is not a substantive restriction as most conventional econometric tests satisfy (9.5). Let G(u) = P(xi <= u) denote the distribution of xi. We call xi (or G) **the asymptotic null distribution**.

It is **generally desirable** to design test statistics T whose **asymptotic null distribution G is known** and **does not depend on unknown parameters**. In this case we say that the statistic T is **asymptotically pivotal**.

For example, if the test statistic equals the absolute t-statistic from (9.2), then we know from Theorem 7.11 that if theta = theta_0 (that is, the null hypothesis holds), then T converges in distribution to |Z| as n → infinity where Z ~ N(0,1). This means that G(u) = P(|Z| <= u) = 2Phi(u) – 1, the distribution of the absolute value of the standard normal as shown in (7.34). This distribution does not depend on unknowns and is pivotal.

We define the **asymptotic size** of the test as the **asymptotic probability of a Type I error**:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(T > c \mid \mathbb{H}_0 \text{ true}\right) &= \mathbb{P}\left(\xi > c\right) \\ &= 1 - G(c).\end{aligned}$$

We see that the asymptotic size of the test is a simple function of the asymptotic null distribution G and the critical value c. For example, the asymptotic size of a test based on the absolute t-statistic with critical value c is 2 (1 – Phi(c)).

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 alpha.

When **the asymptotic null distribution G is pivotal**, we can accomplish this by **setting c equal to the 1 - alpha quantile of the distribution G**.

(If the distribution G is not pivotal, more complicated methods must be used, pointing out the great convenience of using asymptotically pivotal test statistics.)

We call c the **asymptotic critical value** because it has been selected from the asymptotic null distribution.

For example, since 2 (1 – Phi(1.96)) = 0.05, it follows that the 5% asymptotic critical value for the absolute t-statistic is c = 1:96.

Calculation of normal critical values is done numerically in statistical software. For example, in MATLAB the command is norminv(1 - alpha/2).

9.4 t-tests

As we mentioned earlier, the most common test of the **one-dimensional hypothesis**

$$\mathbb{H}_0 : \theta = \theta_0$$

against the alternative

$$\mathbb{H}_1 : \theta \neq \theta_0$$

is the absolute value of the t-statistic (9.3). We now formally state its asymptotic null distribution, which is a simple application of Theorem 7.11.

Theorem 9.1 Under Assumptions 7.2, 7.3, and $\mathbb{H}_0 : \theta = \theta_0$,

$$T(\theta_0) \xrightarrow{d} \mathrm{Z}.$$

For c satisfying $\alpha = 2\left(1 - \Phi(c)\right)$,

$$\mathbb{P}\left(|T(\theta_0)| > c \mid \mathbb{H}_0\right) \longrightarrow \alpha,$$

and the test "Reject $\mathbb{H}_0$ if $|T(\theta_0)| > c$" has asymptotic size α.

The theorem shows that asymptotic critical values can be taken from the normal distribution.

As in our discussion of asymptotic confidence intervals (Section 7.13), the critical value could alternatively be taken from the **student t distribution**, which would be the **exact test** in the **normal regression model** (Section 5.15).

Indeed, t critical values are the default in packages such as Stata.

Since the critical values from the student t distribution are (slightly) larger than those from the normal distribution, using student t critical values decreases the rejection probability of the test.

In practical applications the difference is typically unimportant unless the sample size is quite small (in which case the asymptotic approximation should be questioned as well).

The alternative hypothesis theta not = theta_0 is sometimes called a **“two-sided” alternative**.

In contrast, sometimes we are interested in testing for **one-sided alternatives** such as

$$\mathbb{H}_1 : \theta > \theta_0$$

$$\mathbb{H}_1 : \theta < \theta_0.$$

Tests of theta = theta_0 against theta > theta_0 or theta < theta_0 are based on the **signed** t-statistic T = T(theta_0).

The hypothesis theta = theta_0 is rejected in favor of theta > theta_0 if T > c where c satisfies alpha = 1 - Phi(c). **Negative values of T** are not taken as evidence against H_0, as point estimates theta^ less than theta_0 do not point to theta > theta_0.

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 theta = theta_0 in favor of theta > theta_0 if T > 1.645.

Conversely, **tests of theta = theta_0 against theta < theta_0 reject H_0 for negative t-statistics**, e.g. if T <= -c. For these alternative large positive values of T are not evidence against H_0. An asymptotic 5% test rejects if T < -1.645.

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 **in most cases the two-sided critical value is appropriate**.

We should use the one-sided critical values only when the parameter space is known to satisfy a one-sided restriction such as theta >= theta_0: This is when the test of theta = theta_0 against theta > theta_0 makes sense.

If the restriction theta > theta_0 is not **known a priori**, then imposing this restriction to test theta = theta_0 against theta > theta_0 does not makes sense.

Since linear regression coefficients typically do not have a priori sign restrictions, the standard convention is to use two-sided critical values.

This may seem contrary to the way testing is presented in statistical textbooks, which often focus on one-sided alternative hypotheses. The latter focus is primarily for pedagogy, as the one-sided theoretical problem is cleaner and easier to understand.

9.5 Type II Error and Power

A false acceptance of the null hypothesis H_0 (**accepting H_0 when H_1 is true**) is called a Type II error.

The rejection probability under the alternative hypothesis is called **the power of the test**, and equals **1 minus the probability of a Type II error**.

$$\pi(\boldsymbol{\theta}) = \mathbb{P}\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_1 \text{ true}\right) = \mathbb{P}\left(T > c \mid \mathbb{H}_1 \text{ true}\right)$$

We call pi(theta) the **power function** and is written as **a function of theta** to indicate **its dependence on the true value of the parameter theta**.

In the dominant approach to hypothesis testing, the goal of test construction is to have **high power** subject to the constraint that **the size of the test is lower than the pre-specified significance level**.

Generally, the power of a test **depends on the true value of the parameter theta**, and for **a well-behaved test** the power **is increasing** both as **theta moves away from the null hypothesis theta_0** and **as the sample size n increases**.

Given the two possible states of the world (H_0 or H_1) and the two possible decisions (Accept H_0 or Reject H_0), there are four possible pairings of states and decisions as is depicted in Table 9.1.

Table 9.1: Hypothesis Testing Decisions

	Accept $\mathbb{H}_0$	Reject $\mathbb{H}_0$
$\mathbb{H}_0$ true	Correct Decision	Type I Error
$\mathbb{H}_1$ true	Type II Error	Correct Decision

Given a test statistic T, **increasing the critical value** c **increases the acceptance region** S_0 while **decreasing the rejection region S_1**. This **decreases the likelihood of a Type I error** (**decreases the size**) but **increases the likelihood of a Type II error** (**decreases the power**).

Thus the choice of c involves **a trade-off between size and the power**. This is why **the significance level alpha of the test cannot be set arbitrarily small**. (Otherwise the test will not have meaningful power.)

It is important to consider the power of a test when interpreting hypothesis tests, as an overly narrow focus on size can lead to poor decisions. For example, it is easy to design a test which has **perfect size** yet has **trivial power**.

Specifically, for **any hypothesis** we can use the following test: Generate a random variable U ~ U[0,1] and reject H_0 if U < alpha. This test has exact **size** of alpha. Yet the test also has **power** precisely equal to alpha. When **the power of a test equals the size**, we say that the test has **trivial power**. Nothing is learned from such a test.

9.6 Statistical Significance

Testing requires a pre-selected choice of significance level, yet **there is no objective scientific basis for choice of alpha**.

Nevertheless **the common practice is to set alpha = 0.05 (5%)**. Alternative values are alpha = 0:10 (10%) and alpha = 0:01 (1%). These choices are somewhat **the by-product of traditional tables of critical values and statistical software**.

The informal reasoning behind the choice of a 5% critical value is to ensure that Type I errors should be relatively unlikely - that the decision "Reject H_0" has scientific strength - yet the test retains power against reasonable alternatives.

The decision "Reject H_0" means that the evidence is inconsistent with the null hypothesis, in the sense that **it is relatively unlikely** (1 in 20) that **data generated by the null hypothesis would yield the observed test result**.

In contrast, the decision "Accept H0" is not a strong statement. It does not mean that the evidence supports H_0, only that there is insufficient evidence to reject H_0. Because of this, it is more accurate to use the label "Do not Reject H_0" instead of "Accept H_0".

When a test rejects H_0 at the 5% significance level it is common to say that the statistic is **statistically significant** and if the test accepts H_0 it is common to say that the statistic is **not statistically significant** or that it is statistically insignificant.

It is helpful to remember that this is simply **a compact way** of saying "**Using the statistic T, the hypothesis H_0 can [cannot] be rejected at the asymptotic 5% level**.

Furthermore, when the null hypothesis **H_0: theta = 0 is rejected** it is common to say that **the coefficient theta is statistically significant**, because the test has rejected the hypothesis that the coefficient is equal to zero.

Let us return to the example about the union wage premium as measured in Table 4.1.

The absolute t-statistic for the coefficient on "Male Union Member" is 0.095/0.020 = 4.7, which is greater than the 5% asymptotic critical value of 1.96. Therefore we reject the hypothesis that union membership does not affect wages for men.

In this case, we can say that union membership is statistically significant for men. However, the absolute t-statistic for the coefficient on "Female Union Member" is 0.023/0.020 = 1/2, which is less than 1.96 and therefore we do not reject the hypothesis that union membership does not affect wages for women. In this case we find that membership for women is not statistically significant.

When a test accepts a null hypothesis (when a test is not statistically significant) a common misinterpretation is that this is evidence that the null hypothesis is true. This is incorrect.

Failure to reject is by itself not evidence. Without an analysis of power, we do not know the likelihood of making a Type II error, and thus are uncertain.

In our wage example, it would be a mistake to write that "the regression finds that female union membership has no effect on wages". This is an incorrect and most unfortunate interpretation. The test has failed to reject the hypothesis that the coefficient is zero, but that does not mean that the coefficient is actually zero.

When a test rejects a null hypothesis (when a test is statistically significant) it is strong evidence against the hypothesis (since if the hypothesis were true then rejection is an unlikely event).

Rejection should be taken as evidence against the null hypothesis. However, we can never conclude that the null hypothesis is indeed false, as we cannot exclude the possibility that we are making a Type I error.

Perhaps more importantly, there is an important distinction between **statistical** and **economic** significance.

If we correctly reject the hypothesis H_0: theta = 0 it means that the true value of theta is non-zero. This includes the possibility that theta may be non-zero but close to zero in magnitude. This only makes sense if we interpret the parameters in the context of their relevant models.

In our wage regression example, we might consider wage effects of 1% magnitude or less as being "close to zero". In a log wage regression this corresponds to a dummy variable with a coefficient less than 0.01. If **the standard error is sufficiently small** (less than 0.005) then a coefficient estimate of 0.01 will be **statistically significant** but **not economically significant**. This occurs frequently in applications with **very large sample sizes** where **standard errors can be quite small**.

The solution is to focus whenever possible on **confidence intervals** and the economic meaning of the coefficients. For example, if the coefficient estimate is 0.005 with a standard error of 0.002 then a 95% confidence interval would be [0.001, 0.009] indicating that **the true effect is likely between 0% and 1%**, and hence **is slightly positive but small**. This is much more informative than the misleading statement "the effect is statistically positive".

9.7 P-Values

Continuing with the wage regression estimates reported in Table 4.1, consider another question: Does marriage status affect wages?

To test the hypothesis that marriage status has no effect on wages, we examine the t-statistics for the coefficients on “Married Male” and “Married Female” in Table 4.1, which are 0.211/0.010 = 22 and 0.016/0.010 = 1.7; respectively.

The first exceeds the asymptotic 5% critical value of 1.96, so we reject the hypothesis for men. The second is smaller than 1.96, so we fail to reject the hypothesis for women.

Taking a second look at the statistics, we see that the statistic for men (22) is exceptionally high, and that for women (1.7) is only slightly below the critical value. Suppose that the t-statistic for women were slightly increased to 2.0. This is larger than the critical value so would lead to the decision “Reject H_0” rather than “Accept H_0”.

Should we really be making a different decision if the t-statistic is 2.0 rather than 1.7?

The difference in values is small, shouldn’t the difference in the decision be also small? Thinking through these examples it seems unsatisfactory to simply report “Accept H_0” or “Reject H_0”. These two decisions do not summarize the evidence.

Instead, the magnitude of the statistic T suggests a "**degree of evidence**" against H_0. How can we take this into account?

The answer is to report what is known as the **asymptotic p-value**

$$p = 1 - G(T).$$

Since **the distribution function G is monotonically increasing**, **the p-value is a monotonically decreasing function of T** and **is an equivalent test statistic**.

Instead of rejecting H_0 at the significance level alpha if T > c, we can **reject H_0 if p < alpha**. Thus **it is sufficient to report p, and let the reader decide**.

In practice, **the p-value is calculated numerically**. For example, in MATLAB the command is 2*(1-normalcdf(abs(t))).

In is instructive to interpret p as the **marginal significance level**: **the smallest value of alpha for which the test T "rejects" the null hypothesis**.

That is, p = 0.11 means that T rejects H_0 for all significance levels **greater than** 0.11, but fails to reject H_0 for significance levels **less than** 0.11.

Furthermore, the asymptotic p-value has a very convenient asymptotic null distribution.

Since T converges in distribution to xi under H_0, then p = 1 - G(T) converges in distribution to 1 - G(xi), which has the distribution

$$\begin{aligned}\mathbb{P}\left(1-G(\xi)\leq u\right) &= \mathbb{P}\left(1-u\leq G(\xi)\right)\\ &= 1-\mathbb{P}\left(\xi\leq G^{-1}(1-u)\right)\\ &= 1-G\left(G^{-1}(1-u)\right)\\ &= 1-(1-u)\\ &= u,\end{aligned}$$

which is the uniform distribution on [0,1].

(This calculation assumes that **G(u) is strictly increasing** which is **true for conventional asymptotic distributions such as the normal**.)

Thus p converges in distribution to U[0,1].

This means that **the "unusualness" of p** is easier to interpret than **the "unusualness" of T**.

An important caveat is that **the p-value p should not be interpreted as the probability that either hypothesis is true**.

A common mis-interpretation is that p is the probability "that the null hypothesis is true." This is incorrect.

Rather, p is the **marginal significance level** - **a measure of the strength of information against the null hypothesis**.

For a t-statistic, the p-value can be calculated either using the normal distribution or the student t distribution, the latter presented in Section 5.15.

p-values calculated using the student t will be slightly larger, though the difference is small when the sample size is large.

Returning to our empirical example, for the test that the coefficient on "Married Male" is zero, the p-value is 0.000. This means that it would be **nearly impossible to observe a t-statistic as large as 22 when the true value of the coefficient is zero**.

When presented with such evidence we can say that we "**strongly reject**" the null hypothesis, that the test is "**highly significant**", or that "**the test rejects at any conventional critical value**".

In contrast, the p-value for the coefficient on “Married Female” is 0.094.

In this context it is typical to say that the test is “**close to significant**”, meaning that the p-value is larger than 0.05, but not too much larger.

A related (but **inferior**) empirical practice is to append asterisks (*) to coefficient estimates or test statistics to indicate the level of significance.

A common practice to append a **single asterisk** (*) for an estimate or test statistic which **exceeds the 10% critical value** (i.e., is significant at the 10% level), append a **double asterisk** (**) for a test which **exceeds the 5% critical value**, and append a **triple asterisk** (***) for a test which **exceeds the 1% critical value**.

Such a practice can be better than a table of raw test statistics as the asterisks permit a quick interpretation of significance.

On the other hand, **asterisks are inferior to p-values**, which are also easy and quick to interpret. The goal is essentially the same; it seems wiser to report p-values whenever possible and avoid the use of asterisks.

Our recommendation is that the best empirical practice is to compute and report the asymptotic p-value p rather than simply the test statistic T, the binary decision Accept/Reject, or appending asterisks.

The p-value is a simple statistic, easy to interpret, and contains more information than the other choices.

We now summarize the main features of hypothesis testing.

1. Select a significance level α.
2. Select a test statistic T with asymptotic distribution $T \xrightarrow{d} \xi$ under $\mathbb{H}_0$.
3. Set the asymptotic critical value c so that $1 - G(c) = \alpha$, where G is the distribution function of ξ.
4. Calculate the asymptotic p-value $p = 1 - G(T)$.
5. Reject $\mathbb{H}_0$ if $T > c$, or equivalently $p < \alpha$.
6. Accept $\mathbb{H}_0$ if $T \leq c$, or equivalently $p \geq \alpha$.
7. Report p to summarize the evidence concerning $\mathbb{H}_0$ versus $\mathbb{H}_1$.

9.8 t-ratios and the Abuse of Testing

In Section 4.19, we argued that a good applied practice is to report **coefficient estimates** theta^ and **standard errors** s(theta^) for all coefficients of interest in estimated models. With theta^ and s(theta^) the reader can easily construct confidence intervals [theta^ +/- 2s(theta^)] and t-statistics (theta^-theta_0)/s(theta^) for hypotheses of interest.

Some applied papers (especially older ones) report t-ratios T = theta^/s(theta^) instead of standard errors. This is poor econometric practice.

While the same information is being reported (you can back out standard errors by division, e.g. s(theta^) = theta^/T), standard errors are generally more helpful to readers than t-ratios.

Standard errors help the reader focus on the **estimation precision** and **confidence intervals**, while t-ratios focus attention on **statistical significance**.

While statistical significance is important, it is less important that the parameter estimates themselves and their confidence intervals.

The focus should be on **the meaning of the parameter estimates**, **their magnitudes**, and **their interpretation**, not on listing which variables have significant (e.g. non-zero) coefficients.

In many modern applications, **sample sizes are very large so standard errors can be very small**. Consequently **t-ratios can be large even if the coefficient estimates are economically small**.

In such contexts it may not be interesting to announce "The coefficient is non-zero!"

Instead, what is interesting to announce is that "The coefficient estimate is economically interesting!"

In particular, some applied papers report coefficient estimates and t-ratios, and limit their discussion of the results to describing which variables are "significant" (meaning that their t-ratios exceed 2) and the **signs** of the coefficient estimates. This is very poor empirical work, and should be studiously avoided.

It is also a recipe for banishment of your work to lower tier economics journals.

Fundamentally, the common t-ratio is a test for the hypothesis that a coefficient equals zero. This should be reported and discussed when this is an interesting economic hypothesis of interest. But if this is not the case, it is distracting.

One problem is that standard packages, such as Stata, by default report t-statistics and p-values for every estimated coefficient. While this can be useful (as a user doesn't need to explicitly ask to test a desired coefficient) it can be misleading as it may unintentionally suggest that the entire list of t-statistics and p-values are important. Instead, **a user should focus on tests of scientifically motivated hypotheses**.

In general, when a coefficient theta is of interest, it is constructive to focus on the point estimate, its standard error, and its confidence interval.

The point estimate gives our **"best guess" for the value**. The standard error is **a measure of precision**. The confidence interval gives us **the range of values consistent with the data**.

If **the standard error is large** then **the point estimate is not a good summary about theta**.

The endpoints of the confidence interval describe the bounds on the likely possibilities.

If the confidence interval embraces too **broad** a set of values for theta, then the dataset is not sufficiently informative to render useful inferences about theta.

On the other hand if the confidence interval is **tight**, then the data have produced an accurate estimate, and the focus should be on the value and interpretation of this estimate.

In contrast, the statement "the t-ratio is highly significant" has little interpretive value.

The above discussion requires that the researcher knows what the coefficient theta means (in terms of the economic problem) and can interpret values and magnitudes, not just signs. This is critical for good applied econometric practice.

For example, consider the question about the effect of marriage status on mean log wages. We had found that the effect is “highly significant” for men and “close to significant” for women. Now, let’s construct asymptotic 95% confidence intervals for the coefficients. The one for men is [0.19,0.23] and that for women is [-0.00,0.03]: This shows that average wages for married men are about 19-23% higher than for unmarried men, which is substantial, while the difference for women is about 0-3%, which is small. **These magnitudes are more informative than the results of the hypothesis tests**.

9.9 Wald Tests

The t-test is appropriate when the null hypothesis is **a real-valued restriction**. More generally, there may be **multiple restrictions** on the coefficient vector beta.

Suppose that we have q > 1 restrictions which can be written in the form (9.1). It is natural to estimate theta = r(theta) by the plug-in estimator theta^ = r(beta^).

To test H_0: theta = theta_0 against H_1: theta not= theta_0 one approach is to measure **the magnitude of the discrepancy** theta^ - theta_0. As this is a vector, **there is more than one measure of its length**. One simple measure is **the weighted quadratic form** known as the **Wald statistic**.

This is (7.37) evaluated at the null hypothesis

$$W = W(\boldsymbol{\theta}_0) = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}$$

is an estimator of V_{theta^} and

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\widehat{\boldsymbol{\beta}})'$$

Notice that we can write W alternatively as

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

using the asymptotic variance estimator V^_theta, or we can write it directly as a function of beta^ as

$$W = \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)$$

Also, when r(beta) = R'beta is a linear function of beta, then the Wald statistic simplifies to

$$W = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)$$

The Wald statistic W is **a weighted Euclidean measure of the length of the vector theta^ - theta_0**.

When q = 1 then W = T^2, the square of the t-statistic, so hypothesis tests based on W and |T| are equivalent.

The Wald statistic (9.6) is a generalization of the t-statistic to the case of multiple restrictions. As **the Wald statistic is symmetric** in the argument theta - theta_0, **it treats positive and negative alternatives symmetrically**. Thus, the inherent alternative is always **two-sided**.

As shown in Theorem 7.13, when beta satisfies r(beta) = theta_0, then W converges in distribution to Chi^2_q, a chi-square random variable with q degrees of freedom.

Let G_q(u) denote the Chi^2_q distribution function. For a given significance level alpha, the asymptotic critical value c satisfies alpha = 1 - G_q(c).

For example, the 5% critical values for q = 1, q = 2, and q = 3 are 3.84, 5.99, and 7.82, respectively, and in general the level alpha critical value can be calculated in MATLAB as chi2inv(1-alpha,q).

An asymptotic test rejects H_0 in favor of H_1 if W > c. As with t-tests, it is conventional to describe a Wald test as "significant" if W exceeds the 5% asymptotic critical value.

Theorem 9.2 Under Assumptions 7.2, 7.3, 7.4, and $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, then

$$W \xrightarrow{d} \chi^2_q,$$

and for c satisfying $\alpha = 1 - G_q(c)$,

$$\mathbb{P}\left(W > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $W > c$" has asymptotic size α.

Notice that the asymptotic distribution in Theorem 9.2 depends solely on q, the number of restrictions being tested. It does not depend on k; the number of parameters estimated.

The asymptotic p-value for W is p = 1 - G_q(W), and this is particularly useful when testing multiple restrictions.

For example, if you write that a Wald test on eight restrictions (q = 8) has the value W = 11.2, it is difficult for a reader to assess the magnitude of this statistic unless they have quick access to a statistical table or software.

Instead, if you write that the p-value is p = 0.19 (as is the case for W = 11.2 and q = 8) then it is simple for a reader to interpret its magnitude as “insignificant”.

To calculate the asymptotic p-value for a Wald statistic in MATLAB, use the command 1-chi2cdf(w,q).

Some packages (including Stata) and papers report F versions of Wald statistics. That is, for any Wald statistic W which tests a q-dimensional restriction, the F version of the test is

$$F = W/q.$$

When F is reported, it is conventional to use F_{q,n-k} critical values and p-values rather than Chi^2_q values.

The connection between Wald and F statistics is demonstrated in Section 9.14, we show that when Wald statistics are calculated using a **homoskedastic** covariance matrix, then F = W/q is identical to the F statistic of (5.22).

While there is **no formal justification to using the F_{q,n-k} distribution** for **non-homoskedastic** covariance matrices, the F_{q,n-k} distribution provides continuity with the exact distribution theory under normality and is a bit more conservative than the Chi^2_q distribution. (Furthermore, the difference is small when n - k is moderately large.)

To implement a test of zero restrictions in Stata, an easy method is to use the command “test X1 X2” where X1 and X2 are the names of the variables whose coefficients are hypothesized to equal zero. This command should be executed after executing a regression command. The F version of the Wald statistic is reported, using the covariance matrix calculated using the method specified in the regression command. A p-value is reported, calculated using the F_{q,n-k} distribution.

To illustrate, consider the empirical results presented in Table 4.1. The hypothesis “Union membership does not affect wages” is the joint restriction that both coefficients on “Male Union Member” and “Female Union Member” are zero. We calculate the Wald statistic for this joint hypothesis and find W = 23 (or F = 12:5) with a p-value of p = 0:000. Thus, we reject the null hypothesis in favor of the alternative that at least one of the coefficients is non-zero. This does not mean that both coefficients are non-zero, just that one of the two is non-zero. Therefore, examining both the joint Wald statistic and the individual t-statistics is useful for interpretation.

As a second example from the same regression, take the hypothesis that married status has no effect on mean wages for women. This is the joint restriction that the coefficients on “Married Female” and “Formerly Married Female” are zero. The Wald statistic for this hypothesis is W = 6.4 (F = 3:2) with a p-value of 0.04. Such a p-value is typically called “marginally significant”, in the sense that it is slightly smaller than 0.05.

9.10 Homoskedastic Wald Tests

If the error is known to be homoskedastic, then it is appropriate to use the homoskedastic Wald statistic (7.38) which replaces V^_{theta^} with the homoskedastic estimator V^^0_{theta^}. This statistic equals

$$\begin{aligned} W^0 &= \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ &= \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right) / s^2. \end{aligned}$$

In the case of linear hypotheses H_0: R'beta = theta_0 we can write this as

$$W^0 = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) / s^2.$$

We call either a homoskedastic Wald statistic as it is an appropriate test when the errors are conditionally homoskedastic.

As for W, when q = 1 then W^0 = T^2, the square of the t-statistic where the latter is computed with a homoskedastic standard error.

Theorem 9.3 Under Assumptions 7.2 and 7.3, $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2 > 0$, and $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, then

$$W^0 \xrightarrow{d} \chi^2_q,$$

and for c satisfying $\alpha = 1 - G_q(c)$,

$$\mathbb{P}\left(W^0 > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $W^0 > c$" has asymptotic size α.

9.11 Criterion-Based Tests

The **Wald statistic** is based on the length of the vector theta^ - theta_0: **the discrepancy between the estimate theta^ = r(beta^) and the hypothesized value theta_0**.

An alternative class of tests is based on **the discrepancy between the criterion function minimized with and without the restriction**.

Criterion-based testing applies when we have a criterion function, say J(beta) with beta in B, which is minimized for estimation, and the goal is to test H_0: “beta is in the set B_0” versus H_1: beta is not in the set B_0, where B_0 is a subset of B. Minimizing the criterion function over B and B_0 we obtain the unrestricted and restricted estimators

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \boldsymbol{B}}{\operatorname{argmin}} \; J(\boldsymbol{\beta})$$

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \boldsymbol{B}_0}{\operatorname{argmin}} \; J(\boldsymbol{\beta}) .$$

The **criterion-based statistic** for H_0 versus H_1 is proportional to

$$J = \min_{\beta \in B_0} J(\beta) - \min_{\beta \in B} J(\beta)$$
$$= J(\tilde{\beta}) - J(\hat{\beta}).$$

有没有限制式

The criterion-based statistic J is sometimes called a **distance statistic**, a **minimum-distance statistic**, or a **likelihood-ratio-like statistic**.

Since B_0 is a subset of B, J(beta^) >= J(beta) and thus J >= 0. The statistic J measures **the cost (on the criterion) of imposing the null restriction** that beta is in B_0.

9.12 Minimum Distance Tests

The minimum distance test is a criterion-based test where **J (beta) is the minimum distance criterion** (8.19)

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

with beta^ the unrestricted (LS) estimator.

The restricted estimator beta~_emd minimizes (9.8) subject to beta in B_0. Observing that J(beta^) = 0, the minimum distance statistic simplifies to

$$J = J(\widetilde{\boldsymbol{\beta}}_{\text{md}}) = n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{md}}\right)' \widehat{\boldsymbol{W}} \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{md}}\right).$$

The efficient minimum distance estimator beta~_emd is obtained by setting W^ = V^^{-1}_{beta} in (9.8) and (9.9). The efficient minimum distance statistic for H_0: “beta is in B_0” is therefore

$$J^* = n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right).$$

Consider the class of linear hypotheses H_0: R’beta = theta_0. In this case we know from (8.25) that the efficient minimum distance estimator beta~_emd subject to the constraint R’beta = theta_0 is

$$\widetilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)$$

$$\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right).$$

Substituting into (9.10) we find

$$\begin{aligned} J^* &= n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{V}}_\beta^{-1}\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) \\ &= n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) \\ &= W, \end{aligned}$$

which is the **Wald statistic** (9.6).

For **linear hypotheses** H_0: R'beta = theta_0, **the efficient minimum distance statistic J* is identical to the Wald statistic** (9.6).

For **non-linear hypotheses**, however, **the Wald and minimum distance statistics are different**.

Newey and West (1987a) established **the asymptotic null distribution** of J* for **linear** and **non-linear** hypotheses.

Theorem 9.4 Under Assumptions 7.2, 7.3, 7.4, and $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, then $J^* \xrightarrow{d} \chi^2_q$.

Testing using the minimum distance statistic J* is similar to testing using the Wald statistic W. Critical values and p-values are computed using the Chi^2_q distribution. H_0 is rejected in favor of H1 if J* exceeds the level alpha critical value, which can be calculated in MATLAB as chi2inv(1-alpha,q). The asymptotic p-value is p = 1 – G_q(J*). In MATLAB, use the command 1-chi2cdf(J,q).

We now demonstrate Theorem 9.4. The conditions of Theorem 8.10 hold, since H_0 implies Assumption 8.1. From (8.54) with W^ = V^_beta, we see that

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) &= \widehat{\boldsymbol{V}}_{\beta}\widehat{\boldsymbol{R}}\left(\boldsymbol{R}_n^{*\prime}\widehat{\boldsymbol{V}}_{\beta}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}_n^{*\prime}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\\ &\xrightarrow{d} \boldsymbol{V}_{\beta}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\beta}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\mathrm{N}(\boldsymbol{0},\boldsymbol{V}_{\beta})\\ &= \boldsymbol{V}_{\beta}\boldsymbol{R}\ \mathrm{Z}\end{aligned}$$

where Z ~ N(0, (R'V_beta R)^{-1}). Thus

$$\begin{aligned}J^* &= n\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)'\widehat{\boldsymbol{V}}_{\beta}^{-1}\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)\\ &\xrightarrow{d} \mathrm{Z}'\boldsymbol{R}'\boldsymbol{V}_{\beta}\boldsymbol{V}_{\beta}^{-1}\boldsymbol{V}_{\beta}\boldsymbol{R}\ \mathrm{Z}\\ &= \mathrm{Z}'\left(\boldsymbol{R}'\boldsymbol{V}_{\beta}\boldsymbol{R}\right)\ \mathrm{Z} = \chi_q^2\end{aligned}$$

9.13 Minimum Distance Tests Under Homoskedasticity

If we set W^ = (Q^_xx)/(s^2) in (9.8) we obtain the criterion (8.20)

$$J^0(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) / s^2.$$

A minimum distance statistic for H_0: beta in B_0 is

$$J^0 = \min_{\boldsymbol{\beta} \in \boldsymbol{B}_0} J^0(\boldsymbol{\beta}).$$

Equation (8.21) showed that

$$SSE(\boldsymbol{\beta}) = n\widehat{\sigma}^2 + s^2 J^0(\boldsymbol{\beta})$$

and so the minimizers of SSE(beta) and J^0(beta) are identical.

Thus the constrained minimizer of J^0(beta) is constrained least-squares

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{\beta}\in\boldsymbol{B}_0}{\operatorname{argmin}}\, J^0\left(\boldsymbol{\beta}\right) = \underset{\boldsymbol{\beta}\in\boldsymbol{B}_0}{\operatorname{argmin}}\, SSE(\boldsymbol{\beta})$$

and therefore

$$\begin{aligned} J_n^0 = &\ J_n^0(\widetilde{\boldsymbol{\beta}}_{\text{cls}}) \\ &= n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)' \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) / s^2. \end{aligned}$$

In the special case of linear hypotheses H_0: R'beta = theta_0, the constrained least-squares estimator subject to R'beta = theta_0 has the solution (8.9)

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)$$

and solving we find

$$J^0 = n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)/s^2 = W^0.$$

This is **the homoskedastic Wald statistic** (9.7). Thus for testing linear hypotheses, homoscedastic minimum distance and Wald statistics agree. For nonlinear hypotheses they disagree, but have the same null asymptotic distribution.

Theorem 9.5 Under Assumptions 7.2 and 7.3, $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2 > 0$, and $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, then $J^0 \xrightarrow{d} \chi_q^2$.

9.14 F Tests

In Section 5.16 we introduced the F test for **exclusion restrictions** in the normal regression model. More generally, the F statistic for testing H_0: “beta is in B_0” is

$$F = \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)/q}{\widehat{\sigma}^2/(n-k)}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^2$$

sigma^ and beta^ are the unconstrained estimators of beta and sigma.

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)^2$$

sigma~ and beta~ are the constrained estimators of beta and sigma.

q is the number of restrictions and k is the number of unconstrained coefficients.

We can alternatively write

$$F = \frac{SSE(\widetilde{\boldsymbol{\beta}}_{\text{cls}}) - SSE(\widehat{\boldsymbol{\beta}})}{qs^2}$$

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

is the sum-of-squared errors. Thus **F is a criterion-based statistic**. Using (8.21) we can also write

$$F = J^0/q,$$

so the F statistic is identical to **the homoskedastic minimum distance statistic** divided by **the number of restrictions q**.

As we discussed in the previous section, in the special case of linear hypotheses H_0: R'beta = theta_0, J^0 = W^0: It follows that in this case F = W^0/q. Thus for linear restrictions the F statistic equals the homoskedastic Wald statistic divided by q: It follows that they are equivalent tests for H_0 against H_1.

Theorem 9.6 For tests of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$,

$$F = W^0/q$$

the F statistic equals the homoskedastic Wald statistic divided by the degrees of freedom. Thus under 7.2, $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2 > 0$, and $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, then

$$F \xrightarrow{d} \chi_q^2/q.$$

When using an F statistic, it is conventional to use the F_{q,n-k} distribution for critical values and p-values. Critical values are given in MATLAB by finv(1-alpha,q,n-k), and p-values by 1-fcdf(F,q,n-k). Alternatively, the Chi^2/q distribution can be used, using chi2inv(1-alpha,q)/q and 1-chi2cdf(F*q,q), respectively. Using the F_{q,n-k} distribution is a prudent small sample adjustment which yields exact answers if the errors are normal, and otherwise slightly increasing the critical values and p-values relative to the asymptotic approximation. Once again, **if the sample size is small enough that the choice makes a difference, then probably we shouldn't be trusting the asymptotic approximation anyway**!

An elegant feature about (9.12) or (9.13) is that they are directly computable from the standard output from two simple OLS regressions, as the sum of squared errors (or regression variance) is a typical printed output from statistical packages, and is often reported in applied tables. Thus F can be calculated by hand from standard reported statistics even if you don't have the original data (or if you are sitting in a seminar and listening to a presentation!).

If you are presented with an F statistic (or a Wald statistic, as you can just divide by q) but don't have access to critical values, a useful rule of thumb is to know that for large n; the 5% asymptotic critical value is decreasing as q increases, and is less than 2 for q >= 7.

A word of warning: In many statistical packages, when an OLS regression is estimated an “F-statistic” is automatically reported, even though no hypothesis test was requested. What the package is reporting is an F statistic of the hypothesis that all slope coefficients (except the intercept) are zero. This was a popular statistic in the early days of econometric reporting when sample sizes were very small and researchers wanted to know if there was “any explanatory power” to their regression. This is rarely an issue today, as sample sizes are typically sufficiently large that this F statistic is nearly always highly significant. While there are special cases where this F statistic is useful, these cases are not typical. As a general rule, there is no reason to report this F statistic.

9.15 Hausman Tests

Hausman (1978) introduced a general idea about how to test a hypothesis H0. If you have two estimators, one which is **efficient under H_0** but **inconsistent under H_1**, and another which is **consistent under H_1**, then construct a test as **a quadratic form in the differences of the estimators**.

In the case of testing a hypothesis H_0: r(beta) = theta_0 let beta^_ols denote the unconstrained least-squares estimator and let beta~_emd denote the efficient minimum distance estimator which imposes r(beta) = theta_0.

Both estimators are consistent under H0, but beta~_emd is asymptotically efficient.

Under H1, beta^_ols is consistent for beta but beta~_emd is inconsistent.

The difference has the asymptotic distribution

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) \xrightarrow{d} \text{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\right).$$

Let A- denote the Moore-Penrose generalized inverse. The Hausman statistic for H_0 is

$$
\begin{aligned}
H &= \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \widehat{\text{avar}}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)^{-} \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) \\
&= n\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \left(\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\right)^{-} \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right).
\end{aligned}
$$

The matrix

$$
\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}
$$

idempotent so its generalized inverse is itself. (See Section A.11.) It follows that

$$
\begin{aligned}
\left(\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\right)^{-} &= \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2}\left(\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}\right)^{-}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2} \\
&= \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2} \\
&= \widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'.
\end{aligned}
$$

Thus the Hausman statistic is

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

In the context of **linear restrictions**, R^ = R and R'beta~ = theta_0 so the statistic takes the form

$$H = n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{R}}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\theta}_0\right),$$

which is precisely **the Wald statistic**.

With **nonlinear restrictions** W and H can differ.

In either case we see that that the asymptotic null distribution of the Hausman statistic H is Chi^2_q, so the appropriate test is to reject H_0 in favor of H_1 if H > c where c is a critical value taken from the Chi^2_q distribution.

Theorem 9.7 For general hypotheses the Hausman test statistic is

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

Under Assumptions 7.2, 7.3, 7.4, and $\mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$,

$$H \xrightarrow{d} \chi^2_q.$$

9.16 Score Tests

Score tests are traditionally derived in **likelihood analysis**, but can more generally be constructed from **first-order conditions** evaluated **at restricted estimates**.

We focus on the likelihood derivation. Given the log likelihood function log L(beta, sigma^2), a restriction H_0: r(beta) = theta_0, and restricted estimators beta~ and sigma~^2, **the score statistic for H_0** is defined as

$$S = \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) \right)' \left(-\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) \right)^{-1} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) \right).$$

The idea is that **if the restriction is true, then the restricted estimators should be close to the maximum of the log-likelihood** where **the derivative should be small**.

However **if the restriction is false then the restricted estimators should be distant from the maximum** and **the derivative should be large**.

Hence small values of S are expected under H_0 and large values under H_1. Tests of H_0 thus reject for large values of S.

We explore the score statistic in the context of the **normal regression model** and **linear hypotheses** r(beta) = R'beta.

Recall that in the normal regression log-likelihood function is

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2.$$

The constrained MLE under linear hypotheses is constrained least squares

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

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

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widetilde{e}_i^2.$$

We can calculate that the derivative and Hessian are

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) = \frac{1}{\widetilde{\sigma}^2} \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \boldsymbol{x}_i' \widetilde{\boldsymbol{\beta}} \right) = \frac{1}{\widetilde{\sigma}^2} \boldsymbol{X}' \widetilde{\boldsymbol{e}}$$

$$-\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) = \frac{1}{\widetilde{\sigma}^2} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \frac{1}{\widetilde{\sigma}^2} \boldsymbol{X}' \boldsymbol{X}.$$

Since e~ = y -Xbeta~ we can further calculate that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) &= \frac{1}{\widetilde{\sigma}^2} \left(\boldsymbol{X}' \boldsymbol{X} \right) \left(\left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{y} - \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{X} \widetilde{\boldsymbol{\beta}} \right) \\
&= \frac{1}{\widetilde{\sigma}^2} \left(\boldsymbol{X}' \boldsymbol{X} \right) \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}} \right) \\
&= \frac{1}{\widetilde{\sigma}^2} \boldsymbol{R} \left[\boldsymbol{R}' \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \boldsymbol{R} \right]^{-1} \left(\boldsymbol{R}' \widehat{\boldsymbol{\beta}} - \boldsymbol{c} \right).
\end{aligned}$$

Together we find that

$$S = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)' \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) / \widetilde{\sigma}^2.$$

This is identical to **the homoskedastic Wald statistic**, with **s^2 replaced by sigma~^2**. We can also write S as **a monotonic transformation of the F statistic**, since

$$S = n\frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)}{\widetilde{\sigma}^2} = n\left(1 - \frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2}\right) = n\left(1 - \frac{1}{1 + \frac{q}{n-k}F}\right).$$

The test “Reject H_0 for large values of S” is identical to the test “Reject H_0 for large values of F”, so they are identical tests.

Since for the normal regression model the exact distribution of F is known, it is better to use the F statistic with F p-values.

In more complicated settings **a potential advantage of score tests** is that **they are calculated using the restricted parameter estimates** rather than the unrestricted estimates.

Thus when the restricted estimate is relatively easy to calculate there can be a preference for score statistics.

This is not a concern for linear restrictions.

More generally, score and score-like statistics can be constructed from **first-order conditions** evaluated at **restricted parameter estimates**.

Also, when test statistics are constructed using covariance matrix estimators which are calculated using restricted parameter estimates (e.g. restricted residuals) then these are often described as score tests.

An example of the latter is the Wald-type statistic

$$W = \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)$$

where the covariance matrix estimate V~_{beta^} is calculated using the restricted residuals e~_i = y_i – x'_ibeta~.

This may be done when beta and theta are **high-dimensional**, so there is worry that the estimator V^_{beta^} is imprecise.

9.17 Problems with Tests of Nonlinear Hypotheses

While the t and Wald tests **work well** when **the hypothesis is a linear restriction** on beta, they can **work quite poorly** when **the restrictions are nonlinear**.

This can be seen by a simple example introduced by Lafontaine and White (1986). Take the mode

$$y_i = \beta + e_i$$
$$e_i \sim \mathrm{N}(0, \sigma^2)$$

and consider the hypothesis

$$\mathbb{H}_0 : \beta = 1.$$

Let beta^ and (sigma^)^2 be the sample mean and variance of y_i. The standard Wald test for H_0 is

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

Now notice that H_0 is equivalent to the hypothesis

$$\mathbb{H}_0(s) : \beta^s = 1$$

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

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

While **the hypothesis beta^s = 1 is unaffected by the choice of s**, **the statistic W(s) varies with s**.

This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we have plotted in Figure 9.1 the Wald statistic W(s) as a function of s, setting n/((sigma^)^2) = 10.

The increasing solid line is for the case beta^ = 0.8: The decreasing dashed line is for the case 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**.

This is distressing since the choice of s is arbitrary and irrelevant to the actual hypothesis.

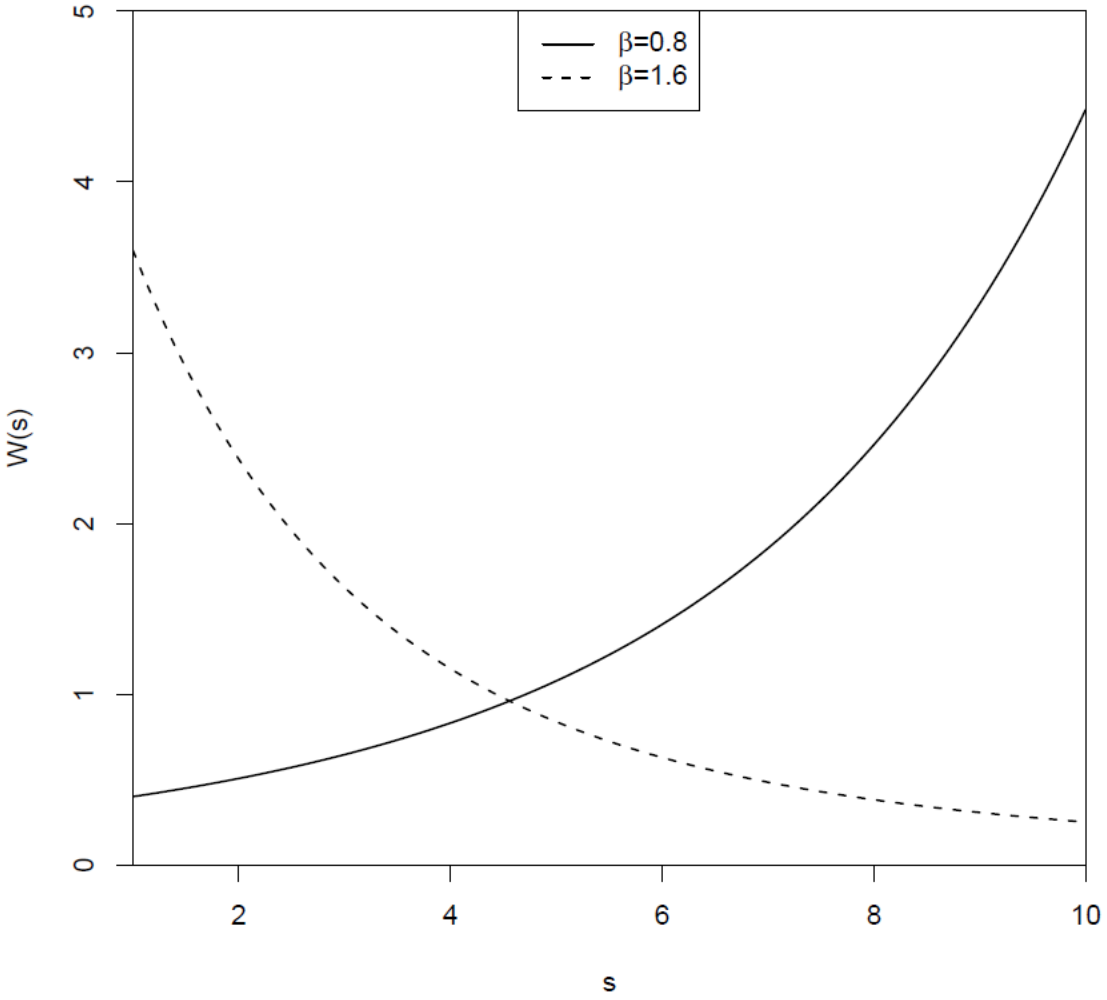

Figure 9.1: Wald Statistic as a Function of s

Our **first-order asymptotic theory** is not useful to help pick s, as W(s) converges in distribution to Chi^2_1 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 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(s) > 3:84 | beta = 1). Given the simplicity of the model, this probability depends only on s, n, and sigma^2.

In Table 9.2 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 sigma 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 sigma.

These probabilities are calculated as the percentage of the 50,000 simulated Wald statistics W(s) which are larger than 3.84.

The null hypothesis s = 1 is true, so these probabilities are Type I error.

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 = W(1) test. Any other choice of s leads to a test with unacceptable Type I error probabilities.

Table 9.2: Type I Error Probability of Asymptotic 5% $W(s)$ Test

s	$\sigma = 1$			$\sigma = 3$		
	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
1	0.05	0.05	0.05	0.05	0.05	0.05
2	0.07	0.06	0.05	0.14	0.08	0.06
3	0.09	0.06	0.05	0.21	0.12	0.07
4	0.12	0.07	0.05	0.25	0.15	0.08
5	0.14	0.08	0.06	0.27	0.18	0.10
6	0.16	0.09	0.06	0.30	0.20	0.12
7	0.18	0.10	0.06	0.32	0.22	0.13
8	0.20	0.12	0.07	0.33	0.24	0.14
9	0.21	0.13	0.07	0.34	0.25	0.16
10	0.23	0.14	0.08	0.35	0.26	0.17

Rejection frequencies from 50,000 simulated random samples.

In Table 9.2 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(s) on s as shown in Figure 9.1.

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.

The common message from both examples is that **Wald statistics are sensitive to the algebraic formulation of the null hypothesis**.

A simple solution is to use the minimum distance statistic J, which equals W with r = 1 in the first example, and |T_2| in the second example.

The minimum distance statistic is invariant to the algebraic formulation of the null hypothesis, so **is immune to this problem**.

Whenever possible, the Wald statistic should not be used to test nonlinear hypotheses. Theoretical investigations of these issues include Park and Phillips (1988) and Dufour (1997).

9.19 Confidence Intervals by Test Inversion

There is a close relationship between hypothesis tests and confidence intervals. We observed in Section 7.13 that the standard 95% asymptotic confidence interval for a parameter theta is

$$\begin{aligned}\widehat{C} &= \left[\widehat{\theta} - 1.96 \cdot s(\widehat{\theta}), \quad \widehat{\theta} + 1.96 \cdot s(\widehat{\theta})\right] \\ &= \{\theta : |T(\theta)| \leq 1.96\}.\end{aligned}$$

That is, we can describe C^ as "**The point estimate plus or minus 2 standard errors**" or "**The set of parameter values not rejected by a two-sided t-test**."

The second definition, known as **test statistic inversion**, is a general method for finding confidence intervals, and typically produces confidence intervals with excellent properties.

Given a test statistic T(theta) and critical value c, the acceptance region "Accept if T(theta) <= c" is identical to the confidence interval C^ = {theta: T(theta) <= c}.

Since the regions are identical, the probability of coverage P(theta in C^) equals the probability of correct acceptance P(Accept | theta) which is exactly 1 minus the Type I error probability. Thus inverting a test with good Type I error probabilities yields a confidence interval with good coverage probabilities.

Now suppose that the parameter of interest theta = r(beta) is a **nonlinear** function of the coefficient vector. In this case the standard confidence interval for theta is the set C^ as in (9.16) where theta^ = r(beta^) is the point estimator and s(theta^) = (R^V^_{beta^}R^)^{-1/2} is the delta method standard error.

This confidence interval is inverting the t-test based on the nonlinear hypothesis r(beta) = theta.

The trouble is that in Section 9.17 we learned that **there is no unique t-statistic for tests of nonlinear hypotheses** and that **the choice of parameterization matters greatly**.

For example, if theta = beta_1/beta_2 then the coverage probability of the standard interval (9.16) is 1 minus the probability of the Type I error, which as shown in Table 8.2 can be far from the nominal 5%.

In this example **a good solution** is the same as discussed in Section 9.17 - **to rewrite the hypothesis as a linear restriction**. The hypothesis theta = beta_1 = beta_2 is the same as theta(beta_2) = beta_1.

The t-statistic for this restriction is

$$T(\theta) = \frac{\widehat{\beta}_1 - \widehat{\beta}_2\theta}{\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\widehat{\beta}}\boldsymbol{R}\right)^{1/2}}$$

$$\boldsymbol{R} = \begin{pmatrix} 1 \\ -\theta \end{pmatrix}$$

and V^_{beta^} is the covariance matrix for (beta^_1 beta^_2).

A 95% confidence interval for theta = beta_1/beta_2 is the set of values of theta such that |T(theta)| <= 1.96. Since T(theta) is a non-linear function of theta one method to find the confidence set is by grid search over theta.

9.20 Multiple Tests and Bonferroni Corrections

In most applications, economists examine a large number of estimates, test statistics, and p-values. **What does it mean** (or **does it mean anything**) if one statistic appears to be "significant" after examining a large number of statistics? This is known as the problem of **multiple testing** or **multiple comparisons**.

To be specific, suppose we examine a set of k coefficients, standard errors and t-ratios, and consider the "significance" of each statistic.

Based on conventional reasoning, for each coefficient we would reject the hypothesis that the coefficient is zero with asymptotic size alpha if the absolute t-statistic exceeds the 1 - alpha critical value of the normal distribution, or equivalently if the p-value for the t-statistic is smaller than alpha.

If we observe that one of the k statistics is "significant" based on this criterion, that means that one of the p-values is smaller than alpha, or equivalently, that the smallest p-value is smaller than alpha.

We can then rephrase the question: Under the joint hypothesis that a set of k hypotheses are all true, **what is the probability that the smallest p-value is smaller than alpha**?

In general, we cannot provide a precise answer to this question, but the **Bonferroni correction** bounds this probability by k.

The Bonferroni method furthermore suggests that if we want the familywise error probability (the probability that one of the tests falsely rejects) to be bounded below alpha, then an appropriate rule is to reject only if the smallest p-value is smaller than alpha/k.

Equivalently, the Bonferroni familywise p-value is k(min{p_j; j<=k}).

Formally, suppose we have k hypotheses H_j, j = 1, …, k. For each we have a test and associated p-value p_j with the property that when H_j is true lim_{n→infinity} P(p_j < alpha) = alpha.

We then observe that among the k tests, one of the k will appear “significant” if min{p_j; j <= k}). This event can be written as

$$\left\{\min_{j\leq k} p_j < \alpha\right\} = \bigcup_{j=1}^{k} \{p_j < \alpha\}.$$

Boole’s inequality states that for any k events A_j,

$$\mathbb{P}\left(\bigcup_{j=1}^{k} A_j\right) \leq \sum_{j=1}^{k} \mathbb{P}(A_k).$$

Thus

$$\mathbb{P}\left(\min_{j\leq k} p_j < \alpha\right) \leq \sum_{j=1}^{k} \mathbb{P}(p_j < \alpha) \longrightarrow k\alpha$$

as stated. This demonstrates that the familywise rejection probability is at most k times the individual rejection probability.

Furthermore,

$$\mathbb{P}\left(\min_{j\leq k} p_j < \frac{\alpha}{k}\right) \leq \sum_{j=1}^{k} \mathbb{P}\left(p_j < \frac{\alpha}{k}\right) \longrightarrow \alpha.$$

This demonstrates that the family rejection probability can be controlled (bounded below alpha) if each individual test is subjected to the stricter standard that a p-value must be smaller than alpha/k to be labeled as "significant."

To illustrate, suppose we have two coefficient estimates, with individual p-values 0.04 and 0.15. Based on a conventional 5% level, the standard individual tests would suggest that the first coefficient estimate is "significant" but not the second.

A Bonferroni 5% test, however, does not reject as it would require that the smallest p-value be smaller than 0.025, which is not the case in this example.

Alternatively, the Bonferroni familywise p-value is 0.08, which is not significant at the 5% level.

In contrast, if the two p-values are 0.01 and 0.15, then the Bonferroni familywise p-value is 0.02, which is significant at the 5% level.

9.21 Power and Test Consistency

The **power of a test** is **the probability of rejecting H_0 when H_1 is true**.

For simplicity suppose that y_i is i.i.d. N(theta, sigma^2) with sigma^2 known, consider the t-statistic T(theta) = (ybar - theta)^{-1/2}/sigma, and tests of H_0: theta = 0 against H_1: theta > 0. We reject H_0 if T = T(0) > c. Note that

$$T = T(\theta) + \sqrt{n}\theta/\sigma$$

and T(theta) has an exact N(0,1) distribution. This is because T(theta) is centered at the true mean theta, while the test statistic T(0) is centered at the (false) hypothesized mean of 0.

The power of the test is

$$\mathbb{P}\left(T > c \mid \theta\right) = \mathbb{P}\left(\mathrm{Z} + \sqrt{n}\theta/\sigma > c\right) = 1 - \Phi\left(c - \sqrt{n}\theta/\sigma\right).$$

This function is monotonically increasing in theta and n; and decreasing in sigma and c.

Notice that for any c and theta not equal to 0; the power increases to 1 as n → infinity.

This means that **for theta in H_1**, the test will reject H_0 with probability approaching 1 as the sample size gets large. We call this property **test consistency**.

Definition 9.3 A test of $\mathbb{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ is **consistent against fixed alternatives** if for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, $\mathbb{P}(\text{Reject } \mathbb{H}_0 \mid \theta) \to 1$ as $n \to \infty$.

For tests of the form "Reject H_0 if T > c", **a sufficient condition** for test consistency is that **the T diverges to positive infinity with probability one for all theta in Theta_1**.

Definition 9.4 We say that $T \xrightarrow{p} \infty$ as $n \to \infty$ if for all $M < \infty$, $\mathbb{P}(T \leq M) \to 0$ as $n \to \infty$. Similarly, we say that $T \xrightarrow{p} -\infty$ as $n \to \infty$ if for all $M < \infty$, $\mathbb{P}(T \geq -M) \to 0$ as $n \to \infty$.

In general, **t-tests and Wald tests are consistent against fixed alternatives**.

Take a t-statistic for a test of H_0: theta = theta_0

$$T = \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})}$$

where theta_0 is a known value and

$$s(\widehat{\theta}) = \sqrt{n^{-1}\widehat{V}_\theta}\ .$$

Note that

$$T = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} + \frac{\sqrt{n}\,(\theta - \theta_0)}{\sqrt{\widehat{V}_\theta}}.$$

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 +infinity if theta > theta_0, and converges in probability to -infinity if theta < theta_0.

Thus the two-sided t-test is consistent against H_1: theta not equal to theta_0, and one-sided t-tests are consistent against the alternatives for which they are designed.

Theorem 9.8 Under Assumptions 7.2, 7.3, and 7.4, for $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$ and $q = 1$, then $|T| \xrightarrow{p} \infty$, so for any $c < \infty$ the test "Reject $\mathbb{H}_0$ if $|T| > c$" *is* consistent against fixed alternatives.

The Wald statistic for H_0: theta = r(beta) = theta_0 against H_1 : theta not equal to theta_0 is

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

Under H_1, theta^ → in probability to theta not equal to theta_0. Thus

$$\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{p} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{V}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) > 0.$$

Hence under H_1, W → in probability to infinity. Again, this implies that Wald tests are consistent tests.

Theorem 9.9 Under Assumptions 7.2, 7.3, and 7.4, for $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$, then $W \xrightarrow{p} \infty$, so for any $c < \infty$ the test "Reject $\mathbb{H}_0$ if $W > c$" *is* consistent against fixed alternatives.

9.22 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 (Section 8.13).

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 this section we consider t-tests on real-valued parameters and in the next section consider Wald tests. Specifically, we consider parameter vectors n which are indexed by sample size n and satisfy the real-valued relationship

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

where the scalar h is called a localizing parameter. We index beta_n and theta_n by sample size to indicate their dependence on n.

The way to think of (9.17) is that the true value of the parameters are beta_n and theta_n. The parameter theta_n is close to the hypothesized value theta_0, with deviation n^{-1/2}h.

The specification (9.17) states that for any fixed h, theta_n approaches theta_0 as n gets large. Thus theta_n is “close” or “local” to theta_0.

The concept of a localizing sequence (9.17) might seem odd since in the actual world the sample size cannot mechanically affect the value of the parameter.

Thus (9.17) **should not be interpreted literally**.

Instead, it should be interpreted as **a technical device** which allows **the asymptotic distribution to be continuous in the alternative hypothesis**.

To evaluate the asymptotic distribution of the test statistic we start by examining the scaled estimate centered at the hypothesized value theta_0. Breaking it into a term centered at the true value theta_n and a remainder we find

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}-\theta_0\right) &= \sqrt{n}\left(\widehat{\theta}-\theta_n\right)+\sqrt{n}\left(\theta_n-\theta_0\right)\\ &= \sqrt{n}\left(\widehat{\theta}-\theta_n\right)+h\end{aligned}$$

where the second equality is (9.17). The first term is asymptotically normal:

$$\sqrt{n}\left(\widehat{\theta}-\theta_n\right)\xrightarrow{d}\sqrt{V_\theta}\mathrm{Z}$$

where Z ~ N(0; 1). Therefore

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}\sqrt{V_\theta}\mathrm{Z}+h\sim\mathrm{N}(h,V_\theta).$$

This asymptotic distribution depends continuously on the localizing parameter h.

Applied to the t statistic we find

$$
\begin{aligned}
T &= \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \\
&\xrightarrow{d} \frac{\sqrt{V_\theta}\mathrm{Z} + h}{\sqrt{V_\theta}} \\
&\sim \mathrm{Z} + \delta
\end{aligned}
$$

Where delta = h/(V_theta)^{1/2}.

This generalizes Theorem 9.1 (which assumes H_0 is true) to allow for local alternatives of the form (9.17).

Consider a t-test of H0 against the one-sided alternative H1 : > 0 which rejects H0 for T > c where (c) = 1 □
. The asymptotic local power of this test is the limit (as the sample size diverges) of the rejection probability under the local alternative (9.17)

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(\text{Reject } \mathbb{H}_0\right) &= \lim_{n\to\infty} \mathbb{P}\left(T > c\right) \\ &= \mathbb{P}\left(\mathrm{Z} + \delta > c\right) \\ &= 1 - \Phi\left(c - \delta\right) \\ &= \Phi\left(\delta - c\right) \\ &\overset{def}{=} \pi(\delta).\end{aligned}$$

We call pi(delta) the asymptotic local power function.

In Figure 9.2 we plot the local power function pi(delta) as a function of delta in [-1,4] for tests of asymptotic size alpha = 0.10, alpha = 0.05, and alpha = 0.01. delta = 0 corresponds to the null hypothesis so pi(delta) = alpha. The power functions are monotonically increasing in delta. Note that the power is lower than alpha for delta < 0 due to the one-sided nature of the test.

We can see that the three power functions are ranked by alpha so that the test with alpha = 0.10 has higher power than the test with alpha = 0.01.

This is the inherent **trade-off between size and power**.

Decreasing size induces a decrease in power, and conversely.

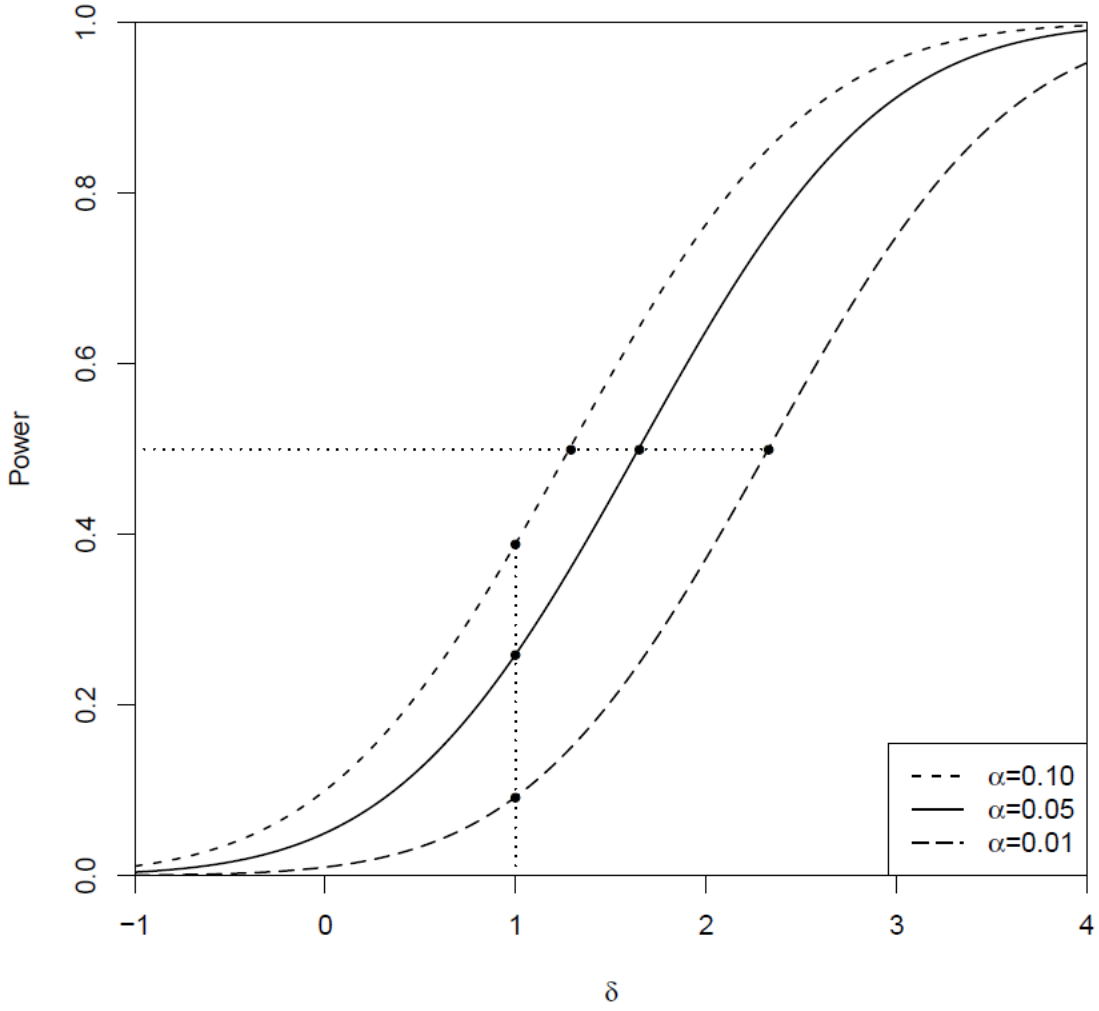

Figure 9.2: Asymptotic Local Power Function of One-Sided t Test

The coefficient delta can be interpreted as the parameter deviation measured as a multiple of the standard error s(theta^). To see this, recall that

$$s(\widehat{\theta}) = n^{-1/2}\sqrt{\widehat{V}_\theta} \simeq n^{-1/2}\sqrt{V_\theta}$$

and then note that

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

Thus delta approximately equals the deviation (theta_n - theta_0) expressed as multiples of the standard error s(theta^).

Thus as we examine Figure 9.2, we can interpret the power function at delta = 1 (e.g. 26% for a 5% size test) as the power when the parameter theta_n is one standard error above the hypothesized value.

For example, from Table 4.1 the standard error for the coefficient on “Married Female” is 0.010. Thus in this example, delta = 1 corresponds to theta_n = 0.010 or an 1.0% wage premium for married females. Our calculations show that the asymptotic power of a one-sided 5% test against this alternative is about 26%.

The difference between power functions can be measured either **vertically** or **horizontally**.

For example, in Figure 9.2 there is a vertical dotted line at delta = 1, showing that the asymptotic local power function pi(delta) 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 size, holding fixed the parameter in the alternative.

A horizontal comparison can also be illuminating. To illustrate, in Figure 9.2 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 theta must be at least 1.65 standard errors above the hypothesized value for a one-sided 5% 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 (9.17) delta = h/(V_{theta})^{1/2} = n^{-1/2}(theta_n - theta_0) = (V_theta)^{1/2}. Thus holding theta and V_theta fixed, delta^2 is proportional to n.

The analysis of a two-sided t test is similar. (9.18) implies that

$$T = \left| \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \right| \xrightarrow{d} |\mathrm{Z} + \delta|$$

and thus the local power of a two-sided t test is

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\left(\text{Reject } \mathbb{H}_0\right) &= \lim_{n\to\infty} \mathbb{P}\left(T > c\right) \\ &= \mathbb{P}\left(|\mathrm{Z} + \delta| > c\right) \\ &= \Phi\left(\delta - c\right) + \Phi\left(-\delta - c\right) \end{aligned}$$

which is monotonically increasing in |delta|.

Theorem 9.10 Under Assumptions 7.2, 7.3, 7.4, and $\theta_n = r(\boldsymbol{\beta}_n) = r_0 + n^{-1/2}h$, then

$$T(\theta_0) = \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \xrightarrow{d} \mathrm{Z} + \delta$$

where $\mathrm{Z} \sim \mathrm{N}(0,1)$ and $\delta = h/\sqrt{V_\theta}$. For c such that $\Phi(c) = 1 - \alpha$,

$$\mathbb{P}\left(T(\theta_0) > c\right) \longrightarrow \Phi\left(\delta - c\right).$$

Furthermore, for c such that $\Phi(c) = 1 - \alpha/2$,

$$\mathbb{P}\left(|T(\theta_0)| > c\right) \longrightarrow \Phi\left(\delta - c\right) + \Phi\left(-\delta - c\right).$$

9.23 Asymptotic Local Power, Vector Case

In this section we extend the local power analysis of the previous section to the case of vector-valued alternatives. We generalize (9.17) to allow theta_n to be vector-valued.

The local parameterization takes the form

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

where h is q-by-1.

Under (9.19),

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

a normal random vector with mean h and variance matrix V_theta.

Applied to the Wald statistic we find

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

Where lambda = h'V^{-1}h. Chi^2_q(lambda) is a non-central chi-square random variable with non-centrality parameter lambda. (See Section 5.3 and Theorem 5.11.)

The convergence (9.20) shows that under the local alternatives (9.19), W → in distribution to Chi^2_q(lambda). This generalizes the null asymptotic distribution which obtains as the special case lambda = 0. We can use this result to obtain a continuous asymptotic approximation to the power function. For any significance level alpha > 0 set the asymptotic critical value c so that P(Chi^2_q > c) = alpha. Then as n → infinity,

$$\mathbb{P}\left(W > c\right) \longrightarrow \mathbb{P}\left(\chi_q^2(\lambda) > c\right) \stackrel{def}{=} \pi(\lambda).$$

The asymptotic local power function pi(lambda) depends only on alpha, q, and lambda.

Theorem 9.11 Under Assumptions 7.2, 7.3, 7.4, and $\boldsymbol{\theta}_n = \boldsymbol{r}(\boldsymbol{\beta}_n) = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h}$, then

$$W \xrightarrow{d} \chi^2_q(\lambda)$$

where $\lambda = \boldsymbol{h}' \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \boldsymbol{h}$. Furthermore, for c such that $\mathbb{P}\left(\chi^2_q > c\right) = \alpha$,

$$\mathbb{P}\left(W > c\right) \longrightarrow \mathbb{P}\left(\chi^2_q(\lambda) > c\right).$$

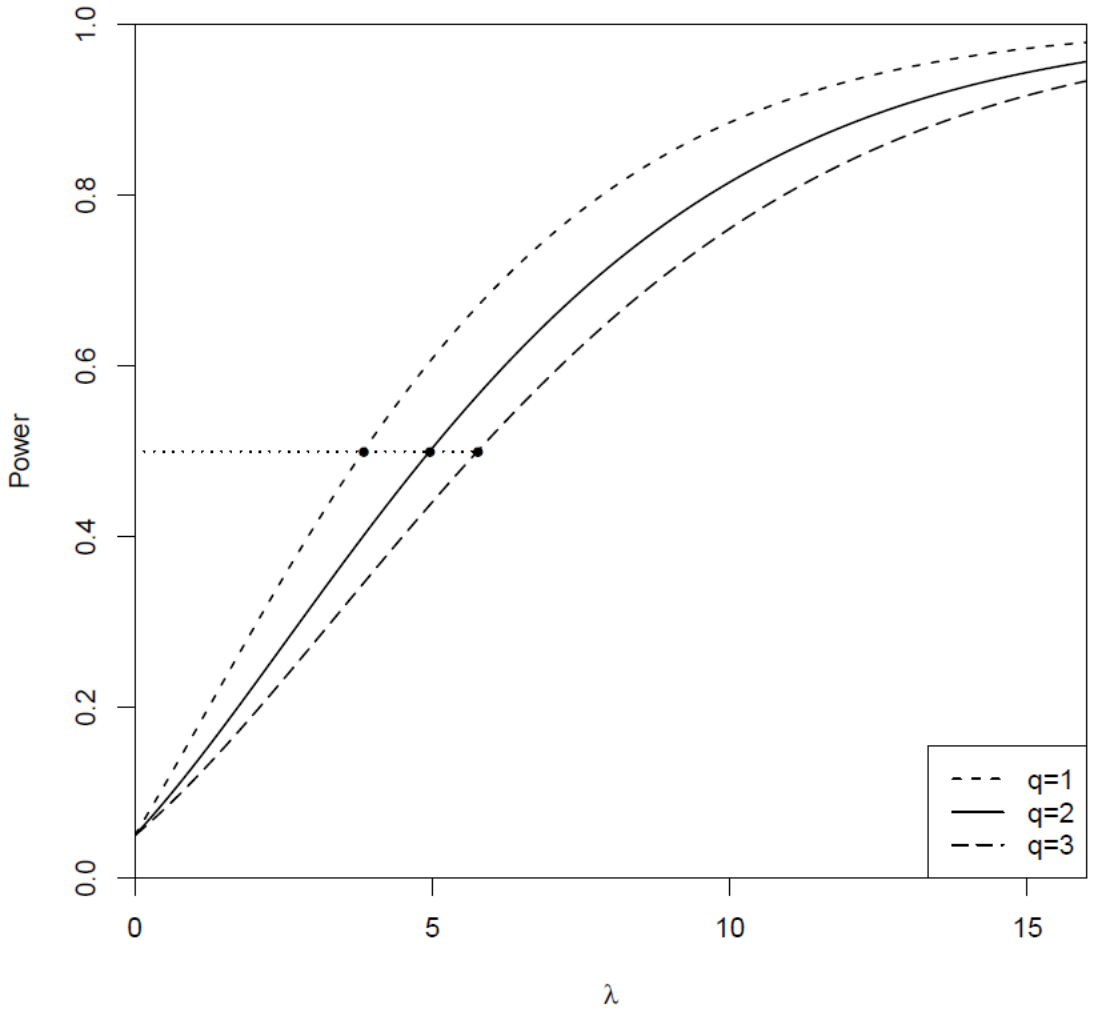

Figure 9.3: Asymptotic Local Power Function, Varying q

Figure 9.3 plots pi(lambda) as a function of lambda for q = 1, q = 2, and q = 3, and alpha = 0.05. The asymptotic power functions are monotonically increasing in lambda and asymptote to one.

Figure 9.3 also shows the power loss for fixed non-centrality parameter lambda 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 lambda 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.