POLSCI 630 Introduction to Regression Methods Spring 2026

Lecture 4 — Large Sample Property

Jiawei Fu, Duke University *Scribe: Jiawei Fu*

1 Overview

Finite-sample properties hold for any sample size. However, from a modern perspective these results often rely on relatively strong assumptions. In this lecture, we study OLS by letting the sample size grow to infinity. We will see that many desirable properties can be obtained under weaker assumptions. In particular, by the central limit theorem, we can derive the asymptotic distribution of the estimator without imposing a parametric distributional assumption on the error term. This allows us to conduct hypothesis tests, construct confidence intervals, and so on.

2 Review Some Tools

For asymptotic analysis, we typically care more about consistency than unbiasedness. Informally, consistency means that the probability that the distance between a random vector Z_n and its limit Z (denoted $d(Z_n, Z)$) is small approaches one as n grows.

Definition 1 (Convergence in Probability)**.** *A sequence of random vectors $Z_n \in \mathbb{R}^k$ converges in probability to Z as $n \to \infty$, denoted $Z_n \to_p Z$, if for all $\delta > 0$,*

$$\lim_{n\to\infty} P[d(Z_n, Z) \leq \delta] = 1.$$

Definition 2 (Consistency)**.** *An estimator $\hat{\beta}$ of β is consistent if $\hat{\beta} \to_p \beta$ as $n \to \infty$.*

A useful tool for proving consistency is the weak law of large numbers.

Theorem 3 (WLLN)**.** *If $Y_i \in \mathbb{R}^k$ are iid and $\mathbb{E}\,||Y|| < \infty$, then as $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \to_p \mathbb{E}\,[Y]$$

It shows that sample mean converges in probability to the true population expectation.

We also want to characterize the asymptotic distribution of estimators. Let Z_n be a sequence of random vectors with distribution functions $F_n(u) = \mathbb{P}[Z_n \leq u]$.

Definition 4 (Convergence in Distribution)**.** *A sequence of random vectors $Z_n \in \mathbb{R}^k$ converges in distribution to Z as $n \to \infty$, denoted $Z_n \to_d Z$, if for all u at which $F(u)$ is continuous, $F_n(u) \to F(u)$ as $n \to \infty$.*

A useful tool for proving convergence in distribution is the (Lindeberg–Lévy) central limit theorem. The WLLN shows that when we scale by n, the sample mean converges to a constant, $\mathbb{E}[Y]$. If instead we scale by $\sqrt{n}$, the fluctuations no longer vanish; under suitable conditions they converge in distribution to a normal random vector.

Theorem 5 (CLT). *If $Y_i \in \mathbb{R}^k$ are iid and $\mathbb{E}||Y||^2 < \infty$, then as $n \to \infty$,*

$$\sqrt{n}(\overline{Y} - \mu) \to_d N(0, V)$$

where $\mu = \mathbb{E}[Y]$ and $V = \mathbb{E}[(Y-\mu)(Y-\mu)']$.

We call V the asymptotic variance of $\sqrt{n}(\overline{Y} - \mu)$. Equivalently, we can treat $\overline{Y}$ as approximately

$$\overline{Y} \sim N(\mu, \frac{V}{n}).$$

The matrix $\frac{V}{n}$ is the asymptotic variance of $\overline{Y}$.

Another useful tool is the continuous mapping theorem. Often we want to know the limiting behavior of a continuous function of a random vector; the CMT provides a convenient way to obtain it.

Theorem 6 (CMT). *Let $Z_n \in \mathbb{R}^k$ and $g(u) : \mathbb{R}^k \to \mathbb{R}^q$.*

1. *If $Z_n \to_p c$, and $g(u)$ is continuous at c, then $g(Z_n) \to_p g(c)$.*
2. *If $Z_n \to_d Z$ and g has the set of discontinuity points D_g s.t. $P[Z \in D_g] = 0$, then $g(Z_n) \to_d g(Z)$.*

3 Consistency of OLS

Before we discuss the properties of the estimator, we need to ensure that the parameter is identified. Recall that the structural parameter

$$\beta = (\mathbb{E}[X_i X_i'])^{-1}\mathbb{E}[X_i Y_i]$$

is identified under the following two assumptions:

Assumption 7 (Population orthogonality). $\mathbb{E}[X_i e_i] = 0$

Assumption 8 (Full Rank). *$rank(X_i X_i') = K$*

The BLP parameter $\beta = (\mathbb{E}[X_i X_i'])^{-1}\mathbb{E}[X_i Y_i]$ is also identified under the full-rank assumption. (In Lecture 1, we said we need $\mathbb{E}[XX']$ to be positive definite. Since $\mathbb{E}[X_i X_i']$ is always positive semidefinite, the additional requirement that it has full rank implies it is positive definite and hence invertible.)

Under these identification assumptions, the OLS estimator can be written as

$$\hat{\beta} = (X'X)^{-1}X'Y = \Big(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\Big).$$

In asymptotic analysis we prefer this summation form because the WLLN and CLT are stated for sample averages. Throughout, we implicitly assume standard regularity conditions (e.g., i.i.d. sampling and suitable moment conditions) so that the WLLN and CLT apply to X_iX_i' and X_iY_i.

We can use the WLLN to show that $\hat{\beta} \to_p \beta$. By the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n} X_iX_i' \to_p \mathbb{E}\left[X_iX_i'\right] \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_iY_i \to_p \mathbb{E}\left[X_iY_i\right].$$

Then the CMT allows us to combine these limits:

$$\hat{\beta} = \Big(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} X_iY_i\Big) \ \to_p \ (\mathbb{E}\left[X_iX_i'\right])^{-1}\mathbb{E}\left[X_iY_i\right] = \beta.$$

In words, the OLS estimator converges in probability to the population parameter as the sample size gets large.

Another useful demonstration is to plug in the model $Y_i = X_i'\beta + e_i$:

$$\begin{aligned}
\hat{\beta} &= (\frac{1}{n}\sum_{i=1}^{n} X_iX_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} X_iY_i) \\
&= \beta + (\frac{1}{n}\sum_{i=1}^{n} X_iX_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} X_ie_i) \\
&\to \beta + (\mathbb{E}\left[X_iX_i'\right])^{-1}\mathbb{E}\left[X_ie_i\right] \\
&= \beta
\end{aligned}$$

The last equality follows from $\mathbb{E}\left[X_ie_i\right] = 0$. In a structural model, this is an assumption; for the BLP projection, it holds by definition.

4 Asymptotic Normality of OLS

To show OLS estimator is asymptotically normal, we rely on CLT. From the idea of CLT, we need $\sqrt{n}$ in the dominator to balance the increasing sum of random variables. Therefore, let us write

$$\sqrt{n}(\hat{\beta} - \beta) = (\frac{1}{n}\sum_{i=1}^{n} X_iX_i')^{-1}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_ie_i)$$

When you see $\sqrt{n}$, you should realize you can apply CLT. Look at the second part first. Note that $\mathbb{E}[X_ie_i] = 0$.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_ie_i \to_d N(0, \Omega)$$

where $\Omega = \mathbb{E}\left[(X_ie_i)(X_ie_i)'\right] = \mathbb{E}\left[X_iX_i'e_i^2\right]$ is $k \times k$ covariance matrix.

For the first part $(\frac{1}{n}\sum_{i=1}^{n} X_iX_i')^{-1}$, we know it converges to $(\mathbb{E}\left[X_iX_i'\right])^{-1}$. Now, combine them together, we conclude that

$$\sqrt{n}(\hat{\beta} - \beta) \to_d N(0, V_\beta)$$

where $V_\beta = (\mathbb{E}[X_i X_i'])^{-1}\Omega(\mathbb{E}[X_i X_i'])^{-1}$. It looks like a sandwich, right? Note, V_β is the asymptotic varaince for $\sqrt{n}(\hat{\beta} - \beta)$. Also note that asymptotic variance for $\hat{\beta}$ is $\frac{1}{n}V_\beta$.

4.1 Asymptotic Variance

Recall that in the previous lecture we examined homoskedastic and heteroskedastic variance. We do the same here.

Homoskedasticity means that the variance of e_i does not depend on X_i. Here, we can express this as $\mathrm{Cov}(X_i X_i', e_i^2) = 0$. Then $\Omega = \mathbb{E}[X_i X_i' e_i^2] = \sigma^2 \mathbb{E}[X_i X_i']$, where $\sigma^2 = \mathbb{E}[e_i^2]$. In this case,

$$V_\beta = (\mathbb{E}[X_i X_i'])^{-1}\Omega(\mathbb{E}[X_i X_i'])^{-1} = \sigma^2(\mathbb{E}[X_i X_i'])^{-1} = V_\beta^0.$$

V_β^0 is called the homoskedastic asymptotic covariance matrix.

Since this variance is defined by a population moment, how should we estimate it? In the finite-sample setting, we examined two estimators for σ^2. Here we show that both $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$ and $s^2 = \frac{1}{n-k}\sum_{i=1}^n \hat{e}_i^2$ are consistent for σ^2.

Proposition 9. *$\hat{\sigma}^2 \to \sigma^2$ and $s^2 \to \sigma^2$.*

Proof. Recall that $\hat{e}_i = Y_i - X_i'\hat{\beta} = e_i - X_i'(\hat{\beta} - \beta)$. Therefore,

$$\hat{e}_i^2 = e_i^2 - 2e_i X_i'(\hat{\beta} - \beta) + (\hat{\beta} - \beta)' X_i X_i'(\hat{\beta} - \beta).$$

Taking averages yields

$$\frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 - 2(\frac{1}{n}\sum_{i=1}^n e_i X_i')(\hat{\beta} - \beta) + (\hat{\beta} - \beta)'(\frac{1}{n}\sum_{i=1}^n X_i X_i')(\hat{\beta} - \beta)$$

By the WLLN, $\frac{1}{n}\sum_{i=1}^n e_i^2 \to \sigma^2$, $\frac{1}{n}\sum_{i=1}^n e_i X_i' \to \mathbb{E}[eX'] = 0$, and $\frac{1}{n}\sum_{i=1}^n X_i X_i' \to \mathbb{E}[XX']$. Moreover, $\hat{\beta} \to \beta$.

Therefore, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 \to \sigma^2$.

Because $\frac{n}{n-k} \to 1$, it follows that $s^2 = \frac{n}{n-k}\hat{\sigma}^2 \to \sigma^2$. □

Given this, we can propose a consistent estimator for the asymptotic variance of $\hat{\beta}$. Under homoskedasticity, $V_\beta^0 = \sigma^2(\mathbb{E}[X_i X_i'])^{-1}$, so we use the sample analogue

$$\hat{V}_\beta^0 = s^2\Big(\frac{1}{n}\sum_{i=1}^n X_i X_i'\Big)^{-1} = s^2\Big(\frac{1}{n}X'X\Big)^{-1}.$$

This is consistent because each component is consistent; therefore, $\hat{V}_\beta^0 \to V_\beta^0$. Note that the asymptotic variance of $\hat{\beta}$ is $\frac{1}{n}V_\beta^0$; therefore, a consistent estimator is $s^2(X'X)^{-1}$. This is the same as the $\hat{V}_{\hat{\beta}}^0$ derived in the finite-sample variance lecture.

Under heteroskedasticity, recall that $V_\beta = (\mathbb{E}[X_i X_i'])^{-1}\mathbb{E}[X_i X_i' e_i^2](\mathbb{E}[X_i X_i'])^{-1}$. Again, we can use sample analogues for each component:

$$\hat{V}_\beta^{HC0} = \Big(\frac{1}{n}\sum_{i=1}^n X_i X_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n X_i X_i' \hat{e}_i^2\Big)\Big(\frac{1}{n}\sum_{i=1}^n X_i X_i'\Big)^{-1}.$$

This is consistent, i.e., $\hat{V}_\beta^{HC0} \to V_\beta$. Accordingly, the HC0 variance estimator for $\hat{\beta}$ is:

$$\frac{1}{n}\hat{V}_\beta^0 = (X'X)^{-1}(\sum_{i=1}^n X_i X_i' \hat{e}_i^2)(X'X)^{-1}$$

Note that this is equivalent to $\hat{V}_{\hat{\beta}}^{HC0}$ derived in the finite-sample variance lecture.

We can define HC1, HC2, and HC3 similarly; all are consistent. For example, $\hat{V}_\beta^{HC1} = \frac{n}{n-k}\hat{V}_\beta^{HC0} \to V_\beta$.

For (asymptotic) standard errors $s^*(\hat{\beta}_j)$ of $\hat{\beta}_j$ (here * denotes a population quantity that must be estimated), we take the square root of the jth diagonal element of the covariance matrix, $\sqrt{[V_{\hat{\beta}}]_{jj}}$. We estimate this by $s(\hat{\beta}_j) = \sqrt{[\hat{V}_{\hat{\beta}}]_{jj}}$.

5 Confidence Interval

Because $\hat{\beta}$ is estimated from data—and the data are random—$\hat{\beta}$ is random. Previously, we focused on point estimation, where we aim to obtain a single estimate. Accounting for sampling uncertainty, here we instead construct a set of plausible values.

We focus on the variable of interest, β_1. We want to find a confidence interval $\hat{C} = [\hat{L}, \hat{U}]$ such that $\mathbb{P}[\beta_1 \in \hat{C}] = 1-\alpha$; common choices for α are 0.10, 0.05, and 0.01. Note that β_1 is treated as a fixed parameter, while $\hat{C}$ is random. Thus, $1-\alpha$ is the coverage probability: the probability that $\hat{C}$ covers the true parameter.

When $\hat{\beta}_1$ is asymptotically normal with standard error $s^*(\hat{\beta}_1)$, the conventional confidence interval for β_1 takes the form

$$[\hat{\beta}_1 - c \times s^*(\hat{\beta}_1), \hat{\beta}_1 + c \times s^*(\hat{\beta}_1)]$$

How do we choose c? We know that $\sqrt{n}(\hat{\beta}_1 - \beta_1) \to N(0, V_{\beta_1})$. We choose c such that

$$\mathbb{P}\Big[\Big|\frac{\sqrt{n}(\hat{\beta}_1 - \beta_1)}{\sqrt{V_{\beta_1}}}\Big| < c\Big] = 1-\alpha.$$

Under the standard normal distribution, $c = z_{1-\alpha/2}$, the $(1-\alpha/2)$ quantile. Therefore, the $1-\alpha$ confidence interval is

$$[\hat{\beta}_1 - z_{1-\alpha/2} \times s^*(\hat{\beta}_1), \hat{\beta}_1 + z_{1-\alpha/2} \times s^*(\hat{\beta}_1)]$$

For $\alpha = 0.05$, $c \approx 1.96$. To estimate the standard error $s^*(\hat{\beta}_1)$, we use the estimated variance $s(\hat{\beta}_1) = \sqrt{[\hat{V}_{\hat{\beta}}]_{11}}$. Since $\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}$ still converges to $N(0,1)$, the previous derivation remains valid.

In practice, people often use the t distribution to choose c. The idea is that we estimate the standard error rather than use the true standard error. Suppose the errors are exactly normal and homoskedastic. Consider the statistic

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \\ &= \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2[XX']_{11}}} \\ &= \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2[XX']_{11}}} / \sqrt{\frac{(n-k)s^2}{\sigma^2(n-k)}} \\ &= \frac{N(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} \\ &\sim t_{n-k} \end{aligned}$$

because $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$. Note that this only applies to the t-statistic constructed using the homoskedastic standard error; it does not apply to a t-statistic constructed with robust standard errors. The difference between Student-t and normal critical values is typically small in practice (since sample sizes are usually large in economic applications). However, because Student-t critical values are larger, they yield slightly wider confidence intervals, which can be viewed as a conservative approximation.

6 Hypothesis Testing

Besides confidence intervals, standard statistical inference includes hypothesis testing. In this course, we are interested in whether the effect is actually zero.

Definition 10. *A hypothesis is a statement about a population parameter.*

The hypothesis to be tested is called the null hypothesis: $\beta_1 = 0$. The complement of the null hypothesis is called the alternative hypothesis: $\beta_1 \neq 0$. This is a two-sided alternative. More generally, we write $H_0 : \beta \in \Theta_0$ and $H_1 : \beta \in \Theta_0^c$.

Definition 11. *A hypothesis testing procedure or hypothesis test is a rule that specifies*

1. *For which sample values the decision is made to accept H_0*
2. *For which sample values H_0 is rejected (and H_1 is accepted).*

The subset of the sample space for which H_0 will be rejected is called the rejection region (or critical region). The complement of the rejection region is called the acceptance region.

Typically, a hypothesis test is specified in terms of a test statistic, a function of the sample. For example, a common choice is the t statistic, $|T| = \left|\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}\right| = \left|\frac{\hat{\beta}_1}{s(\hat{\beta}_1)}\right|$ under $H_0 : \beta_1 = 0$. If H_0 is true then we expect $|T|$ to be small, but if H_1 is true then we would expect $|T|$ to be large. Thus the hypothesis test takes the form

$$\text{Reject } H_0 \text{ if } |T| > c$$

where c is called the critical value.

There are two possible errors. If H_0 is true but the hypothesis test incorrectly reject H_0 (false rejection of the null hypothesis), then the test has made a Type I error. The probability of a Type I error is

$$\mathbb{P}[\text{Reject } H_0 | H_0]$$

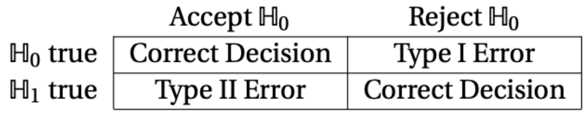
On the other hand, if H_1 is true but the test accepts H_0 (false acceptance of the null hypothesis), then a Type II error has been made.

Definition 12. *The power function of a hypothesis test is the function of β defined by $\pi(\beta) = P[\textit{reject } H_0 | \beta]$.*

Ideally, the power function is 0 for $\beta \in \Theta_0$ and 1 for $\beta \in \Theta_0^c$. The power of a test is the rejection probability under the alternative hypothesis:

$$\mathbb{P}[\text{Reject } H_0 | H_1]$$

Note that this equals 1 minus the probability of a Type II error.



	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

Figure 1: Caption

Power generally depends on the sample size. Also, given a test statistic T, increasing the critical value c increases the acceptance region while decreasing the rejection region. This reduces the likelihood of a Type I error (reduces the size) but increases the likelihood of a Type II error (reduces the power). Thus the choice of c involves a trade-off between size (Type I error) and power. This is why the significance level α of the test cannot be set arbitrarily small; otherwise the test will not have meaningful power.

In science, we often seek evidence of causal effects, which are sometimes assumed to be zero by default. Compared to a Type II error, a Type I error is often viewed as more severe. Therefore, a primary goal of test construction is to limit the incidence of Type I errors. Given the constraint that the size of the test is no larger than the pre-specified significance level α, we then want the test to have high power.

Definition 13. *For $0 \leq \alpha \leq 1$, a test with power function $\pi(\beta)$ is a size α test if $\sup_{\beta \in \Theta_0} = \alpha$.*

For $0 \leq \alpha \leq 1$, a test with power function $\pi(\beta)$ is a level α test if $\sup_{\beta \in \Theta_0} \pi(\beta) \leq \alpha$.

To calculate these probabilities, we need a distribution. As we have seen, the exact distribution is hard to derive without strong assumptions, so we focus on asymptotic size and level. Returning to the previous example, how should we choose c (an asymptotic critical value)? Its value is selected to control the probability of false rejections (Type I errors). Researchers typically pre-select a significance level $\alpha \in (0, 1)$ so that, with an appropriate critical value c, the asymptotic size is no larger than α.

We know under H_0, the distribution of T is asymptotically N(0,1) (or ad hoc conservative adjustment, t_{n-k}). Therefore,

$$\begin{aligned}\mathbb{P}[\text{Reject } H_0|H_0] &= \mathbb{P}[|T| > c|H_0] \\ &= \mathbb{P}[T > c|H_0] + \mathbb{P}[T < -c|H_0] \\ &\to 1 - F(c) + F(-c) \\ &= 2(1 - F(c))\end{aligned}$$

where F is the CDF of the standard normal distribution (or a t distribution). We set $2(1-F(c)) = \alpha$, so $F(c) = 1 - \frac{\alpha}{2}$. Therefore, $c = z_{1-\alpha/2}$ or $c = t_{1-\alpha/2,n-k}$. This test has asymptotic size α. Note that this is the same c we derived for confidence intervals, which is not surprising.

When a test rejects H_0 at a given significance level, it is common to say that the statistic is statistically significant. If the test accepts H_0, it is common to say that the statistic is not statistically significant. Furthermore, when the null hypothesis $H_0 : \beta = 0$ is rejected, it is common to say that the coefficient β is statistically significant, because the test rejects the hypothesis that the coefficient equals zero.

When a test rejects a null hypothesis (when a test is statistically significant) it is strong evidence against the hypothesis (because 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 significance and scientific (or substantive) significance. If we correctly reject $H_0 : \beta = 0$, then the true value of β is nonzero. This still allows β to be nonzero but close to zero in magnitude. Interpretation therefore requires considering the parameter in the context of the model and the units of measurement.

6.1 P value

To report the result of a hypothesis test, we typically pre-determine the significance level α in order to calculate the critical value c. This can be inconvenient and arbitrary. Moreover, if α is small, rejecting H_0 is fairly convincing; if α is large, rejecting H_0 is less convincing because the test allows a higher probability of making that decision incorrectly. An alternative is to report the p-value of the test. A p-value summarizes the strength of evidence against H_0 (but it is not the probability that H_0 is true). This section is mainly based on Casella & Berger (2024).

There are two common ways to understand the p-value. Suppose we use $O = \{(X_i, Y_i)\}_{i=1}^n$ to denote the sample.

Definition 14. *A p-value $p(O)$ is a test statistic satisfying $0 \le p(o) \le 1$ for every sample point o. Small values of p give evidence that H_1 is true. A p-value is valid if, for every $\beta \in \Theta_0$ and every $0 \le \alpha \le 1$,*

$$P_\beta[p(O) \le \alpha] \le \alpha$$

If $p(O)$ is a valid p-value, it is easy to construct a level-α test based on $p(O)$. Instead of rejecting H_0 at significance level α when $|T| > c$, we can reject H_0 when $p(O) \le \alpha$. An advantage of reporting the p-value is that each reader can choose the α they consider appropriate, compare the reported $p(o)$ to α, and determine whether the data lead to rejection of H_0. Thus it is sufficient to report $p(o)$ and let the reader decide. Furthermore, the smaller the p-value, the stronger the evidence against H_0.

It is also instructive to interpret p as the marginal significance level: the smallest value of α for which the test statistic rejects the null hypothesis. For example, $p = 0.11$ means the test rejects H_0 for all significance levels greater than 0.11, but fails to reject H_0 for significance levels less than 0.11.

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.

The second (and most common) way to define a valid p-value is given below.

Theorem 15. *Let $W(O)$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point o, define*

$$p(o) = \sup_{\beta \in \Theta_0} P_\beta[W(O) \geq W(o)]$$

Then, $p(O)$ is a valid p-value.

It says that the p-value is the probability (under H_0) of observing a test statistic at least as extreme as the one computed from the sample. If this probability is large, then the sample is not unusual under H_0, so we do not reject the null. If the p-value is small, then the sample is unusual under H_0, providing evidence against H_0.

Let us return to the previous example. Suppose H_0 is true. The test statistic is $W(O) = |T| = \left|\frac{\hat{\beta}_1}{s(\hat{\beta})}\right|$. We know $\frac{\hat{\beta}_1}{s(\hat{\beta})}$ converges in distribution to $N(0,1)$. Given the data o, we can compute the realized value of this statistic, $W(o)$. What is the p-value? It is the probability of observing a value of $W(O)$ at least as extreme as $W(o)$. Here, "extreme" means either large or small, so

$$P[W(O) \geq W(o)] = P\left[|T| \geq \left|\frac{\hat{\beta}_1}{s(\hat{\beta})}\right|\right] = 2P\left[T \geq \left|\frac{\hat{\beta}_1}{s(\hat{\beta})}\right|\right] = 2\left(1 - F\left(\left|\frac{\hat{\beta}_1}{s(\hat{\beta})}\right|\right)\right),$$

where F is the CDF of the statistic T (e.g., the standard normal or the t distribution).

The t-test is appropriate when the null hypothesis is a single real-valued restriction. More generally, there may be multiple restrictions on the coefficient vector; such tests are called Wald tests. For further discussion, see Hansen (2022).

6.2 Test based on the fit

Previously, the test depended on the difference between the estimated parameter and the true parameter. Now we introduce another approach based on model fit.

We can think of the null hypothesis as imposing constraints on the linear model. The model you estimate is the unrestricted model, for example,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e.$$

Suppose we want to test whether the last q regressors have zero effects: $H_0 : \beta_{k-q+1} = \cdots = \beta_k = 0$. Imposing these restrictions yields the restricted model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-q} X_{k-q} + e.$$

For the test, we can look at the change in RSS. If H_0 is true, we may expect SSR not to change much. The F statistic is defined as

$$F = \frac{(RSS_r - RSS_{ur})/q}{RSS_{ur}/(n-k-1)}$$

where RSS_r is the residual sum of squares from the restricted model and RSS_{ur} is the residual sum of squares from the unrestricted model. Note that RSS_r cannot be smaller than RSS_{ur}. Intuitively, F measures the relative increase in SSR when moving from the unrestricted to the restricted model. Here, q is the number of restrictions.

F statistics follow the F distribution, $F_{q,n-k-1}$. In general, the F distribution with d_1 and d_2 degrees of freedom is the distribution of $\frac{U_1/d_1}{U_2/d_2}$, where U_1 and U_2 are independent chi-square random variables with d_1 and d_2 degrees of freedom, respectively.

We reject H_0 in favor of H_1 when F is sufficiently large.

References

Casella, G. & Berger, R. (2024). *Statistical inference.* Chapman and Hall/CRC.

Hansen, B. (2022). *Econometrics.* Princeton University Press.