Interval estimation and hypothesis tests

So far our focus has been on estimation of the parameter vector β in the linear model

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i \\ &= x_i'\beta + u_i \qquad\qquad \text{for } i = 1, 2, ..., n \end{aligned}$$

or

$$y = X\beta + u$$

We have introduced the OLS estimator $\widehat{\beta}_{OLS}$ and shown that this gives an unbiased and efficient estimator (in the class of linear, unbiased estimators) in the classical linear regression models

The OLS estimator gives us **point estimates** of the β_k parameters

For example, if we estimate $\widehat{\beta}_2 = 1.032$ in a particular application, our 'best guess' for the value of the β_2 parameter may well be 1.032

But we know that this guess will almost certainly be wrong

Often we are more interested in characterizing an interval within which the true value of the parameter lies with a specified probability

For example, we may want to say that there is a 95% probability that β_2 lies in an interval like $(0.988, 1.076)$

This is known as **interval estimation**, and the interval in this example is called a **95% confidence interval**

Closely related to this, if we estimate $\widehat{\beta}_2 = 1.032$, we may want to test whether the true value of the β_2 parameter could plausibly be one

If the 95% confidence interval is $(0.988, 1.076)$, which includes the value one, we would conclude that a true value of one is quite likely, and we would not reject the hypothesis that $\beta_2 = 1$ (at the conventional 5% significance level)

This is known as **hypothesis testing**

Our result that the conditional variance $\text{Var}(\widehat{\beta}_{\text{OLS}}|X) = \sigma^2(X'X)^{-1}$ in the classical linear regression model with stochastic regressors will be useful for characterizing confidence intervals and developing hypothesis tests

But knowing the conditional expectation and variance of the OLS estimator is not sufficient

We also need results about the (conditional) **distribution** of the OLS estimator before we can use these to estimate confidence intervals and conduct hypothesis tests

There are two approaches we can take here

- adding the assumption that $u|X$ (or equivalently $y|X$) has a normal distribution will allow us to derive the exact, finite sample distribution of $\widehat{\beta}_{\text{OLS}}|X$ in the classical linear regression model with stochastic regressors
- without specifying the form of the distribution for $u|X$ (or $y|X$), we can still characterize the distribution of $\widehat{\beta}_{\text{OLS}}$ in large samples (more precisely, as the sample size n tends to infinity), and we can use this asymptotic result to approximate the distribution of $\widehat{\beta}_{\text{OLS}}$ in large finite samples

One motivation for assuming that the stochastic error terms are drawn from a normal distribution may be the idea that the error term captures the combined effects of many different factors that we have left out of our model

We can then appeal to a central limit theorem to suggest that the sum of many independent random variables will have a normal distribution

Adding the assumption that $u|X$ has a normal distribution to the assumptions we made to specifiy the classical linear regression model with stochastic regressors gives us the classical linear regression model with normally distributed errors (from now on, unless otherwise specified, we will be treating the explanatory variables as stochastic)

Classical linear regression model with normally distributed errors

This can be stated as

$$y = X\beta + u$$

$$u|X \sim \mathrm{N}(0, \sigma^2 I)$$

$$X \text{ has full rank (with probability one)}$$

or equivalently as

$$y|X \sim \mathrm{N}(X\beta, \sigma^2 I)$$

$$X \text{ has full rank (with probability one)}$$

The key result is that

$$\widehat{\beta}_{\text{OLS}}|X \sim \text{N}(\beta, \sigma^2(X'X)^{-1})$$

We already knew that $\text{E}(\widehat{\beta}_{\text{OLS}}|X) = \beta$ and $\text{Var}(\widehat{\beta}_{\text{OLS}}|X) = \sigma^2(X'X)^{-1}$, since we had established those results in the classical linear regression model without using the normality assumption

The new ingredient is that the random vector $\widehat{\beta}_{\text{OLS}}|X$ also has a **normal distribution**

This follows directly from the property that, conditional on X, $\widehat{\beta}_{\text{OLS}}$ is a linear function of the random vector y, and we have now assumed that, conditional on X, the random vector y has a normal distribution

Here we use the fact that if, *conditional on* X, the $n \times 1$ vector y is stochastic, with $y|X \sim \text{N}(\mu, \Sigma)$, and if, *conditional on* X, the $K \times 1$ vector a is non-stochastic and the $K \times n$ matrix B is non-stochastic, with full row rank ($\text{rank}(B) = K$), then *conditional on* X, the $K \times 1$ vector $z = a + By$ is stochastic, with $z|X \sim \text{N}(a + B\mu, B\Sigma B')$

Since $\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y = Ay$, with A non-stochastic conditional on X and $\text{rank}(A) = K$, we can use this fact to obtain the conditional distribution of $\widehat{\beta}_{\text{OLS}}|X$

Some more **useful facts about the multivariate normal distribution** are now collected

Let $y \sim \text{N}(\mu, \Sigma)$ where y and μ are $n \times 1$ vectors and Σ is an $n \times n$ matrix

Partition

$$y = \begin{pmatrix} y_1 \\ \\ y_2 \end{pmatrix}, \ \mu = \begin{pmatrix} \mu_1 \\ \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where y_1 and μ_1 are $n_1 \times 1$ vectors, y_2 and μ_2 are $n_2 \times 1$ vectors, $n_1 + n_2 = n$, Σ_{11} is an $n_1 \times n_1$ matrix, Σ_{22} is an $n_2 \times n_2$ matrix, Σ_{12} is an $n_1 \times n_2$ matrix, and $\Sigma_{21} = \Sigma_{12}'$ is an $n_2 \times n_1$ matrix

Marginal Distribution $\quad y_1 \sim \mathrm{N}(\mu_1, \Sigma_{11})$

Conditional Distribution $\quad y_2 | y_1 \sim \mathrm{N}(a + B'y_1, \Sigma_{22} - B'\Sigma_{11}B)$

where here $B = (\Sigma_{11})^{-1}\Sigma_{12}$ is an $n_1 \times n_2$ matrix, and $a = \mu_2 - B'\mu_1$ is an $n_2 \times 1$ vector

Zero correlation implies independence: if $\Sigma_{12} = 0$ then y_1 and y_2 are independent

This does not hold in general for two random vectors y_1 and y_2, but does hold if their joint distribution is multivariate normal

Finally if the scalar random variable z has a standard normal distribution, $z \sim \mathrm{N}(0, 1)$, then

$$\mathrm{P}(-1.96 < z < 1.96) = 0.95$$

An immediate implication of the Marginal Distribution property is that since $\widehat{\beta}_{\text{OLS}}|X \sim \text{N}(\beta, \sigma^2(X'X)^{-1})$, we have

$$\widehat{\beta}_k|X \sim \text{N}(\beta_k, v_{kk}) \text{ for } k = 1, ..., K$$

where $\widehat{\beta}_k$ is again the k^{th} element of $\widehat{\beta}_{\text{OLS}}$, β_k is the k^{th} element of true parameter vector β, and $v_{kk} = \text{Var}(\widehat{\beta}_k|X)$ is the element in the k^{th} row and the k^{th} column of $\text{Var}(\widehat{\beta}_{\text{OLS}}|X) = \sigma^2(X'X)^{-1}$ (equivalently the k^{th} element of the main diagonal of $\text{Var}(\widehat{\beta}_{\text{OLS}}|X) = \sigma^2(X'X)^{-1}$)

And the linear function of $\widehat{\beta}_k$

$$z_k = \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) = \left(\frac{-\beta_k}{\sqrt{v_{kk}}}\right) + \left(\frac{1}{\sqrt{v_{kk}}}\right)\widehat{\beta}_k$$

has a standard normal distribution

$$z_k|X \sim \mathrm{N}(0,1) \;\; \text{for } k = 1, ..., K$$

Moreover, because the conditional distribution of $z_k|X$ is the same distribution (standard normal) for *any* realization of X, we also have that the unconditional distribution of z_k is standard normal

$$z_k = \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} \right) \sim \mathrm{N}(0,1) \;\; \text{for } k = 1, ..., K$$

One subtle point to note here is that the *realized value* of z_k does depend on the realized X, which affects both $\widehat{\beta}_k$ and $\sqrt{v_{kk}}$, but the *conditional distribution* does not depend on the realized X, which is sufficient to obtain the unconditional distribution result

In contrast, since the conditional distribution of $\widehat{\beta}_k|X \sim \mathrm{N}(\beta_k, v_{kk})$ does depend on the realized X (through v_{kk}), we do not have an unconditional normal distribution for the estimated parameter $\widehat{\beta}_k$ itself (in the model with stochastic regressors considered here)

But this does not matter, since we use the standardized variable z_k in the construction of confidence intervals and test statistics

If we knew the variance parameter, we could use these results directly

Imagine for a moment that we knew σ^2, and hence we knew $\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X) = \sigma^2(X'X)^{-1}$ and v_{kk}

The scalar random variable $z_k \sim \mathrm{N}(0,1)$, so that

$$\mathrm{P}(-1.96 < z_k < 1.96) = 0.95$$

$$\mathrm{P}(-1.96 < \frac{\widehat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} < 1.96) = 0.95$$

$$\mathrm{P}(-1.96\sqrt{v_{kk}} < \widehat{\beta}_k - \beta_k < 1.96\sqrt{v_{kk}}) = 0.95$$

$$\mathrm{P}(-1.96\sqrt{v_{kk}} < \beta_k - \widehat{\beta}_k < 1.96\sqrt{v_{kk}}) = 0.95$$

$$\mathrm{P}(\widehat{\beta}_k - 1.96\sqrt{v_{kk}} < \beta_k < \widehat{\beta}_k + 1.96\sqrt{v_{kk}}) = 0.95$$

The penultimate step uses the symmetry of the interval about zero

This would give our **95% confidence interval** for the true value of β_k as $(\widehat{\beta}_k - 1.96\sqrt{v_{kk}}, \widehat{\beta}_k + 1.96\sqrt{v_{kk}})$, or equivalently as $\widehat{\beta}_k \pm 1.96\sqrt{v_{kk}}$

To test the hypothesis that the true value of β_k is some candidate value β_k^0, we would form the **test statistic**

$$z_k = \frac{\widehat{\beta}_k - \beta_k^0}{\sqrt{v_{kk}}}$$

which would have a standard normal distribution if this hypothesis is correct

If this value of z_k looks like an unlikely draw from a standard normal distribution (e.g. if this value of z_k lies outside the interval $(-1.96, 1.96)$), it is unlikely that the hypothesis is correct, and we would reject this hypothesis (at the 5% significance level)

Otherwise (e.g. if our test statistic z_k is in the interval $(-1.96, 1.96)$) we would not reject the hypothesis

These approaches are related: if the hypothesized value β_k^0 lies outside the 95% confidence interval, we will reject the hypothesis that this is the true value of β_k at the 5% significance level

Similar ideas will be used in the more realistic case when we do not know the variance parameter σ^2 and so have to estimate this from the data to obtain estimates of $\text{Var}(\widehat{\beta}_{\text{OLS}}|X)$ and v_{kk}

We can use the (unbiased) OLS estimator

$$\widehat{\sigma}^2_{\text{OLS}} = \frac{\widehat{u}'\widehat{u}}{n-K}$$

and estimate $\text{Var}(\widehat{\beta}_{\text{OLS}}|X)$ using

$$\widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}}|X) = \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1}$$

Letting $\widehat{v}_{kk}$ denote the k^{th} element on the main diagonal of $\widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}}|X)$, we will use the standardized statistic

$$t_k = \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} \right) = \left(\frac{\widehat{\beta}_k - \beta_k}{se_k} \right)$$

where $se_k = \sqrt{\widehat{v}_{kk}}$ is called the **standard error** of the estimated parameter $\widehat{\beta}_k$

However, replacing the unknown $\sqrt{v_{kk}}$ by the estimated $se_k = \sqrt{\widehat{v}_{kk}}$ changes the finite sample distribution of this statistic, so we now need to derive the distribution of t_k

To obtain the distribution of the t_k statistic, we use some more properties of normally distributed random vectors

We start with properties of **functions of a standard normal vector**

The $n \times 1$ random vector $z \sim \mathrm{N}(0, I)$, where I is an $n \times n$ identity matrix, is called a standard normal vector, with elements $z_i \sim \mathrm{N}(0, 1)$ for $i = 1, 2, ..., n$ that are **independent** standard normal random variables

The density of z

$$f_z(a) = (2\pi)^{\frac{-n}{2}} \exp\left(\frac{-a'a}{2}\right) = \prod_{i=1}^{n} \left[\left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-a_i^2}{2}\right)\right]$$

is the product of n standard normal densities

1) The scalar $w = z'z = \sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$

The sum of squares of n independent standard normal random variables has a **chi-squared distribution** with n degrees of freedom

Special case $(n = 1) \Rightarrow z_i^2 \sim \chi^2(1)$ where $z_i \sim \mathrm{N}(0, 1)$

2) If the random variables $w_1 \sim \chi^2(m)$ and $w_2 \sim \chi^2(n)$ and w_1 and w_2 are independent, then the scalar

$$v = \frac{(w_1/m)}{(w_2/n)} \sim F(m, n)$$

The ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom, has a **Snedecor F-distribution** with degrees of freedom of the numerator and denominator variables respectively

3) If the random variables $z \sim \mathrm{N}(0,1)$ and $w \sim \chi^2(n)$ are independent, then the scalar

$$u = \frac{z}{\sqrt{(w/n)}} \sim t(n)$$

The ratio of a standard normal random variable to the square root of an independent chi-squared random variable divided by its degrees of freedom has a **Student t-distribution** with that degrees of freedom

Corollary: since $z^2 \sim \chi^2(1)$

$$u^2 = \frac{z^2}{(w/n)} = \frac{(z^2/1)}{(w/n)} \sim F(1,n)$$

The square of a r.v. with a $t(n)$ distribution has a $F(1,n)$ distribution

We also use two properties of **quadratic forms** of normally distributed random vectors

1) If the $n \times 1$ vector $y \sim \mathrm{N}(\mu, \Sigma)$ and the scalar $w = (y - \mu)'\Sigma^{-1}(y - \mu)$, then $w \sim \chi^2(n)$

2) If the $n \times 1$ vector $z \sim \mathrm{N}(0, I)$ and the non-stochastic $n \times n$ matrix G is symmetric and idempotent with $\mathrm{rank}(G) = r \leqslant n$, then the scalar $w = z'Gz \sim \chi^2(r)$

We can now find the **distribution of the statistic**

$$t_k = \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} \right) = \left(\frac{\widehat{\beta}_k - \beta_k}{se_k} \right)$$

in the classical linear regression model with normally distributed errors

First let $u = \sigma z$, with $z|X \sim \mathrm{N}(0, I)$ and $y = X\beta + \sigma z$

Recall that $\widehat{u} = My$ where $M = I - X(X'X)^{-1}X'$ is a symmetric, idempotent matrix with the properties that $MX = 0$ and $\mathrm{rank}(M) = n - K$

Then $\widehat{u} = My = M[X\beta + \sigma z] = MX\beta + M\sigma z = \sigma Mz$

And $\widehat{u}'\widehat{u} = (\sigma Mz)'(\sigma Mz) = \sigma^2(z'M')Mz = \sigma^2 z'MMz = \sigma^2 z'Mz$

Then since $\text{rank}(M) = n - K$ and $z|X$ is a standard normal vector, the scalar

$$w_0 = \frac{\widehat{u}'\widehat{u}}{\sigma^2} = \frac{\sigma^2 z' M z}{\sigma^2} = z' M z$$

has a conditional distribution

$$w_0|X \sim \chi^2(n - K)$$

using property (2) of quadratic forms in normal vectors, and the fact that M is not stochastic conditional on X

This is one of the key results we need to establish the distribution of t_k

Now write $\widehat{\beta}_{\text{OLS}} = Ay = A[X\beta + \sigma z] = AX\beta + A\sigma z = \beta + \sigma Az$ since $AX = I$

Equivalently $\widehat{\beta}_{\text{OLS}} - \beta = \sigma Az$

Consider the random vectors

$$d_1 = \frac{\widehat{u}}{\sigma} = \frac{\sigma Mz}{\sigma} = Mz$$

$$d_2 = \frac{\widehat{\beta}_{\text{OLS}} - \beta}{\sigma} = \frac{\sigma Az}{\sigma} = Az$$

Since M and A are not stochastic conditional on X, and since $z|X \sim \text{N}(0, I)$, we know that $d_1|X$ and $d_2|X$ are normally distributed

Moreover $d|X$ is normally distributed, where $d = (d_1, d_2)'$

Also $AM = (X'X)^{-1}X'[I - X(X'X)^{-1}X'] = (X'X)^{-1}X' -$

$(X'X)^{-1}X'X(X'X)^{-1}X' = (X'X)^{-1}X' - (X'X)^{-1}X' = 0$

It follows that $\mathrm{Cov}(d_2, d_1|X) = AM\mathrm{Var}(z|X) = 0$, so that conditional on X, d_1 and d_2 are uncorrelated, and hence also independent

But $\widehat{u} = \sigma d_1$ and $\widehat{\beta}_{\mathrm{OLS}} = \beta + \sigma d_2$, where β and σ are non-stochastic

Conditional on X, $\widehat{u}$ and $\widehat{\beta}_{\mathrm{OLS}}$ are therefore independent

And conditional on X, any function of $\widehat{u}$ and any function of $\widehat{\beta}_{\mathrm{OLS}}$ are therefore independent

This is another key result we need to find the distribution of t_k

Now write

$$t_k = \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} \right) = \left(\frac{\sqrt{v_{kk}}}{\sqrt{\widehat{v}_{kk}}} \right) \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} \right) = \left(\frac{\sqrt{v_{kk}}}{\sqrt{\widehat{v}_{kk}}} \right) z_k$$

where $z_k|X \sim \mathrm{N}(0,1)$

Since $\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X) = \sigma^2(X'X)^{-1} = \sigma^2 Q^{-1}$, we can write $v_{kk} = \sigma^2 q_{kk}$ where q_{kk} is the k^{th} element on the main diagonal of $Q^{-1} = (X'X)^{-1}$, and $\sqrt{v_{kk}} = \sigma\sqrt{q_{kk}}$

Similarly since $\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{OLS}}|X) = \widehat{\sigma}^2_{\mathrm{OLS}}(X'X)^{-1} = \widehat{\sigma}^2_{\mathrm{OLS}} Q^{-1}$, we can write $\widehat{v}_{kk} = \widehat{\sigma}^2_{\mathrm{OLS}} q_{kk}$, and $\sqrt{\widehat{v}_{kk}} = \widehat{\sigma}_{\mathrm{OLS}}\sqrt{q_{kk}}$

Then

$$\frac{\sqrt{v_{kk}}}{\sqrt{\widehat{v}_{kk}}} = \frac{\sigma\sqrt{q_{kk}}}{\widehat{\sigma}_{\text{OLS}}\sqrt{q_{kk}}} = \frac{\sigma}{\widehat{\sigma}_{\text{OLS}}}$$

Now

$$\frac{\widehat{\sigma}^2_{\text{OLS}}}{\sigma^2} = \left(\frac{1}{\sigma^2}\right)\left(\frac{\widehat{u}'\widehat{u}}{n-K}\right) = \left(\frac{1}{n-K}\right)\left(\frac{\widehat{u}'\widehat{u}}{\sigma^2}\right) = \frac{w_0}{n-K}$$

where $w_0 = \frac{\widehat{u}'\widehat{u}}{\sigma^2}$ and we showed that $w_0|X \sim \chi^2(n-K)$

Hence

$$t_k = \left(\frac{\sqrt{v_{kk}}}{\sqrt{\widehat{v}_{kk}}}\right) z_k = \left(\frac{\sigma}{\widehat{\sigma}_{\text{OLS}}}\right) z_k = \frac{z_k}{(\widehat{\sigma}_{\text{OLS}}/\sigma)} = \frac{z_k}{\sqrt{(\widehat{\sigma}^2_{\text{OLS}}/\sigma^2)}} = \frac{z_k}{\sqrt{w_0/(n-K)}}$$

Consider

$$t_k = \frac{z_k}{\sqrt{w_0/(n-K)}}$$

We know that $z_k|X \sim \mathrm{N}(0,1)$ and $w_0|X \sim \chi^2(n-K)$

Moreover $z_k = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}$ is a function of $\widehat{\beta}_{\mathrm{OLS}}$

While $w_0 = \frac{\widehat{u}'\widehat{u}}{\sigma^2}$ is a function of $\widehat{u}$

Conditional on X, we showed that any function of $\widehat{u}$ and any function of $\widehat{\beta}_{\mathrm{OLS}}$ are independent

So conditional on X, the scalar

$$t_k = \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} \right) = \frac{z_k}{\sqrt{w_0/(n-K)}}$$

is the ratio of a standard normal random variable to the square root of an independent chi-squared random variable divided by its degrees of freedom

Hence

$$t_k|X \sim t(n-K)$$

And since this conditional distribution of $t_k|X$ is the same distribution (t-distribution with $n-K$ degrees of freedom) for any realization of X, we again have that the unconditional distribution is the same, i.e.

$$t_k \sim t(n-K)$$

$$t_k = \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} \right) = \left(\frac{\widehat{\beta}_k - \beta_k}{se_k} \right) \sim t(n-K)$$

We can use this result to form **confidence intervals** and to construct **hypothesis tests**

NB. Since the sample mean $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \widehat{\beta}_1$ in the special case where $y_i = \beta_1 + u_i$, this result gives the exact finite sample distribution of

$$t_1 = \left(\frac{\widehat{\beta}_1 - \beta_1}{se_1} \right) = \left(\frac{\overline{y} - \mu}{(s/\sqrt{n})} \right) \sim t(n-1)$$

when the data on $y_1, y_2, ..., y_n$ is iid $\mathrm{N}(\mu, \sigma^2)$, i.e. we have $y_i \sim \mathrm{N}(\mu, \sigma^2)$ for each $i = 1, 2, ..., n$ and the observations are independent

The t-distributions are also symmetric about zero, with more mass in their tails ('fatter tails') compared to the standard normal distribution

To construct the 95% confidence interval, we use the lower 2.5% percentile and the upper 97.5% percentile of the t-distribution with $n - K$ degrees of freedom

By symmetry about zero, these 'critical values' have the form

$$\mathrm{P}(t_k > c_{0.025}(n-K)) = 0.025$$

$$\mathrm{P}(t_k < -c_{0.025}(n-K)) = 0.025$$

or

$$\mathrm{P}(-c_{0.025}(n-K) < t_k < c_{0.025}(n-K)) = 0.95$$

$$\mathrm{P}(-c_{0.025}(n-K) < \frac{\widehat{\beta}_k - \beta_k}{se_k} < c_{0.025}(n-K)) = 0.95$$

Again we can rearrange to obtain the 95% confidence interval for β_k as

$$\mathrm{P}(\widehat{\beta}_k - c_{0.025}(n-K)se_k < \beta_k < \widehat{\beta}_k + c_{0.025}(n-K)se_k) = 0.95$$

which may be written as the interval

$$(\widehat{\beta}_k - c_{0.025}(n-K)se_k, \widehat{\beta}_k + c_{0.025}(n-K)se_k)$$

or as $\widehat{\beta}_k \pm c_{0.025}(n-K)se_k$

Notice that, all else equal, the 95% confidence interval for β_k will be narrower if the standard error $se_k = \sqrt{\widehat{v}_{kk}}$ is smaller; a lower standard error means we have a more precise estimate of β_k

Note that the critical value depends on the degrees of freedom, and hence on the sample size (n) relative to the number of parameters in β (K)

For example, with $n - K = 20$, we have $c_{0.025}(n - K) = 2.086$, which is bigger than the corresponding 97.5% percentile of the standard normal distribution (1.96)

However, as $n - K$ increases, $c_{0.025}(n - K)$ approaches 1.96 from above

For example, with $n - K = 40$, we have $c_{0.025}(n - K) = 2.021$, and with $n - K = 100$, we have $c_{0.025}(n - K) = 1.984$

In the limit, as $(n-K) \to \infty$, the $t(n-K)$ distribution approaches the standard normal distribution, so that $c_{0.025}(n-K) \to 1.96$

In very large samples, relative to the dimension of the parameter vector, we could still construct confidence intervals based on the standard normal distribution, even though we have estimated σ^2

Other confidence intervals can be constructed in the same way

Letting $5\% = 0.05 = \alpha$, we have constructed the $(1-\alpha)$ confidence interval, using $c_{0.025}(n-K) = c_{\alpha/2}(n-K)$ such that

$$\mathrm{P}(-c_{0.025}(n-K) < t_k < c_{0.025}(n-K)) = 0.95$$

More generally, we use the critical value $c_{\alpha/2}(n-K)$ such that

$$\mathrm{P}(-c_{\alpha/2}(n-K) < t_k < c_{\alpha/2}(n-K)) = 1-\alpha$$

to obtain the $(1-\alpha)$ confidence interval as $\widehat{\beta}_k \pm c_{\alpha/2}(n-K)se_k$

For example, in very large samples, we could use the 95% percentile of the standard normal distribution to obtain the 90% confidence interval as $\widehat{\beta}_k \pm 1.645se_k$

Or we could use the 99.5% percentile of the standard normal distribution to obtain the 99% confidence interval as $\widehat{\beta}_k \pm 2.576se_k$

The same result is used to construct a **simple hypothesis test** about the true value of an individual parameter in the classical linear regression model with normally distributed errors

The hypothesis we wish to test is called the null hypothesis

For example, $H_0 : \beta_2 = 1$, or more generally $H_0 : \beta_k = \beta_k^0$ for some $k = 1, ..., K$

For a two-sided test, the alternative hypothesis would be $H_1 : \beta_2 \neq 1$, or more generally $H_1 : \beta_k \neq \beta_k^0$

We form the test statistic

$$t_k = \frac{\widehat{\beta}_k - \beta_k^0}{se_k}$$

If the null hypothesis is true, we know that

$$t_k = \frac{\widehat{\beta}_k - \beta_k^0}{se_k} = \frac{\widehat{\beta}_k - \beta_k}{se_k} \sim t(n-K) \text{ under } H_0 : \beta_k = \beta_k^0$$

We then ask whether the value of the test statistic t_k that we have constructed is an unlikely draw from a t-distribution with $n - K$ degrees of freedom

If the value of t_k is a sufficiently unlikely draw from the distribution of this statistic under the null hypothesis, we conclude that the null hypothesis is unlikely to be true

For example, we **reject** the null hypothesis at the **5% level of significance** if the probability of drawing a test statistic with an absolute value as large as $|t_k|$ from its null distribution (i.e. its distribution if the null hypothesis is true) is less than 5%

Otherwise we **do not reject** the null hypothesis at the 5% significance level

Since we have

$$\mathrm{P}(-c_{0.025}(n-K) < t_k < c_{0.025}(n-K)) = 0.95 \text{ under } H_0 : \beta_k = \beta_k^0$$

we reject the null hypothesis against the two-sided alternative at the 5% level of significance if the value of t_k lies outside the symmetric interval

$$(-c_{0.025}(n-K), c_{0.025}(n-K))$$

or equivalently if $|t_k| > c_{0.025}(n-K)$

Otherwise we do not reject the null hypothesis against the two-sided alternative at the 5% level

While it is most common to conduct hypothesis tests at the 5% significance level, the same principle can be used to test hypotheses at any desired level of significance, and results are sometimes reported at the 1% level or at the 10% level

In general, we reject $H_0 : \beta_k = \beta_k^0$ against $H_1 : \beta_k \neq \beta_k^0$ at the α level of significance if the value of t_k lies outside the interval

$$(-c_{\alpha/2}(n-K), c_{\alpha/2}(n-K))$$

or equivalently if $|t_k| > c_{\alpha/2}(n-K)$

Note that it is quite possible to reject the null hypothesis at the 10% level, but not at the 5% level

In a very large sample, this would just correspond to a t-statistic with an absolute value between 1.645 and 1.96

In such cases, the evidence against the validity of the null hypothesis is sometimes described as 'weakly significant' or 'marginally significant' (depending on whether the author thinks the hypothesis is true or false, respectively)

An alternative to reporting whether the null hypothesis is rejected or not rejected at some arbitrary level of significance is to report the significance level at which the null hypothesis would just be on the borderline between being rejected and being not rejected

For example, if we obtained a t-statistic with an absolute value of 1.96 in a very large sample, we know that the significance level for which we would just reject the null hypothesis is 5%

That is, we would reject the null at the 5.1% significance level or higher, but not at the 4.9% significance level or lower

This is referred to as the **marginal significance level** or **p-value** of the test statistic

More generally, the p-value is the probability of drawing a test statistic with an absolute value greater than or equal to the one we have obtained, if the null hypothesis is true

Conventionally the null hypothesis is then rejected if the p-value is less than 0.05 or 5%

By reporting the p-value, we allow readers to decide whether a draw with a 5% probability is sufficiently unlikely to want to reject the null hypothesis

There is a one-to-one correspondence between confidence intervals and hypothesis tests in this context

A question in one of the problem sets asks you to show that we reject $H_0 : \beta_k = \beta_k^0$ against $H_1 : \beta_k \neq \beta_k^0$ at the 5% significance level if and only if the hypothesized value β_k^0 lies outside the 95% confidence interval for β_k

A special case of the t-test we have considered arises when the null hypothesis of interest is that the coefficient on the k^{th} explanatory variable is zero, i.e. $H_0 : \beta_k = 0$ (in which case we *might* choose to omit this explanatory variable from the model)

The test statistic is then

$$t_k = \frac{\widehat{\beta}_k - 0}{se_k} = \frac{\widehat{\beta}_k}{se_k} \sim t(n-K) \text{ under } H_0 : \beta_k = 0$$

which is simply the ratio of the OLS estimate of this coefficient to its standard error

This ratio is referred to as the **t-ratio**

We reject $H_0 : \beta_k = 0$ against the two-sided alternative $H_1 : \beta_k \neq 0$ at the 5% significance level if the t-ratio lies outside the interval

$$(-c_{0.025}(n-K), c_{0.025}(n-K))$$

or if the absolute value of the t-ratio is bigger than $c_{0.025}(n-K)$

The estimated coefficient (*not* the variable) is then said to be 'significantly different from zero' at the 5% level

Sometimes this statement gets abbreviated to 'statistically significant' at the 5% level, although it is better practice to be clear about the particular null hypothesis that has been rejected when making statements of this kind

Sometimes we may want to test the null hypothesis $H_0 : \beta_k = \beta_k^0$ against a **one-sided alternative** hypothesis $H_1 : \beta_k > \beta_k^0$ (or perhaps $H_1 : \beta_k < \beta_k^0$)

For example, if the coefficient β_k has an interpretation as a mark-up parameter in a regression relating price to marginal cost, we may wish to test $H_0 : \beta_k = 1$ against the one-sided alternative $H_1 : \beta_k > 1$

Here point estimates of β_k below one may be interpreted as being consistent with a perfectly competitive market structure (price = marginal cost), while evidence that $\beta_k > 1$ may be interpreted as indicating market power (price > marginal cost)

To test $H_0 : \beta_k = \beta_k^0$ against $H_1 : \beta_k > \beta_k^0$, we conduct a **one-sided test**, asking whether the constructed test statistic t_k is an unlikely draw from the upper tail of the null distribution of the test statistic

Here we would not regard unlikely draws from the lower tail of the null distribution as providing evidence against the validity of the null hypothesis

In this case we reject H_0 against this one-sided alternative H_1 at the 5% significance level if $t_k > c_{0.05}(n - K)$, indicating a less than 5% probability of drawing the value t_k if the null hypothesis is true

[For $H_1 : \beta_k < \beta_k^0$, we would focus instead on the lower tail]

We have considered how to test a hypothesis about the value taken by an **individual element** of the parameter vector β

But often we are interested in testing a hypothesis about a **linear combination** of these parameters

For example, we may have estimated a Cobb-Douglas production function of the form

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

where y_i is the log of value added for observation i, x_{2i} is the log of labour input for observation i, and x_{3i} is the log of capital input for observation i

In this case, we may want to test the constant returns to scale restriction $\beta_2 + \beta_3 = 1$

Also we may want to test a hypothesis that imposes more than one restriction on the vector of parameters in the model

For example, the **joint hypothesis** that $\beta_2 = 0$ and $\beta_3 = 0$ imposes two linear restrictions

To conduct tests of these more general **linear restrictions** in the classical linear regression model with normally distributed errors, we first derive a distribution result for a more general test statistic

To consider $p \geqslant 1$ linear restrictions, we use the $p \times 1$ vector $\theta = H\beta$ for some non-stochastic $p \times K$ matrix H with $\text{rank}(H) = p$ (full row rank)

For example, with $K = 3$, the 1×3 matrix

$$H = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}$$

gives $\theta = \beta_2 + \beta_3$, and the 2×3 matrix

$$H = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

gives

$$\theta = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}$$

We estimate θ using $\widehat{\theta}_{\mathrm{OLS}} = H\widehat{\beta}_{\mathrm{OLS}}$, a $p \times 1$ random vector

In the classical linear regression model, we have $\mathrm{E}(\widehat{\theta}_{\mathrm{OLS}}|X) = H\mathrm{E}(\widehat{\beta}_{\mathrm{OLS}}|X) = H\beta = \theta$, so that $\widehat{\theta}_{\mathrm{OLS}}$ is an unbiased estimator of θ

And $\mathrm{Var}(\widehat{\theta}_{\mathrm{OLS}}|X) = H\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X)H' = H[\sigma^2(X'X)^{-1}]H' = \sigma^2 H(X'X)^{-1}H' = \sigma^2 D^{-1}$, where $D^{-1} = H(X'X)^{-1}H'$ is non-stochastic conditional on X

With normally distributed errors, we also have

$$\widehat{\theta}_{\mathrm{OLS}}|X \sim \mathrm{N}(\theta, \sigma^2 D^{-1})$$

Similarly we estimate $\mathrm{Var}(\widehat{\theta}_{\mathrm{OLS}}|X)$ using $\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{OLS}}|X) = H\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{OLS}}|X)H' = H[\widehat{\sigma}^2_{\mathrm{OLS}}(X'X)^{-1}]H' = \widehat{\sigma}^2_{\mathrm{OLS}}H(X'X)^{-1}H' = \widehat{\sigma}^2_{\mathrm{OLS}}D^{-1}$

Using property (1) of quadratic forms of normal random vectors, we have $w|X \sim \chi^2(p)$, where the scalar random variable

$$\begin{aligned} w &= (\widehat{\theta}_{\text{OLS}} - \theta)'[\sigma^2 D^{-1}]^{-1}(\widehat{\theta}_{\text{OLS}} - \theta) \\ &= \left(\frac{1}{\sigma^2}\right)(\widehat{\theta}_{\text{OLS}} - \theta)' D (\widehat{\theta}_{\text{OLS}} - \theta) \end{aligned}$$

Consider

$$\widehat{w} = \left(\frac{1}{\widehat{\sigma}^2_{\text{OLS}}}\right)(\widehat{\theta}_{\text{OLS}} - \theta)' D (\widehat{\theta}_{\text{OLS}} - \theta)$$

and

$$v = \frac{\widehat{w}}{p} = \left(\frac{1}{p\widehat{\sigma}^2_{\text{OLS}}}\right)(\widehat{\theta}_{\text{OLS}} - \theta)' D (\widehat{\theta}_{\text{OLS}} - \theta)$$

Write

$$v = \frac{\widehat{w}}{p} = \left(\frac{\sigma^2}{\widehat{\sigma}^2_{\text{OLS}}}\right)\left(\frac{w}{p}\right) = \left(\frac{\sigma^2}{\widehat{\sigma}^2_{\text{OLS}}}\right)\left(\frac{1}{p\sigma^2}\right)(\widehat{\theta}_{\text{OLS}} - \theta)' D (\widehat{\theta}_{\text{OLS}} - \theta)$$

As before, we have

$$\frac{\widehat{\sigma}^2_{\text{OLS}}}{\sigma^2} = \frac{w_0}{n-K} \text{ where } w_0 = \frac{\widehat{u}'\widehat{u}}{\sigma^2}$$

and we showed previously that $w_0|X \sim \chi^2(n-K)$

Thus

$$v = \left(\frac{\sigma^2}{\widehat{\sigma}^2_{\text{OLS}}}\right)\left(\frac{w}{p}\right) = \frac{(w/p)}{(\widehat{\sigma}^2_{\text{OLS}}/\sigma^2)} = \frac{(w/p)}{(w_0/(n-K))}$$

is the ratio of two chi-squared random variables, each divided by their respective degrees of freedom

$$v = \frac{(w/p)}{(w_0/(n-K))}$$

where $w|X \sim \chi^2(p)$ and $w_0|X \sim \chi^2(n-K)$

Moreover $w = \left(\frac{1}{\sigma^2}\right)(\widehat{\theta}_{\text{OLS}} - \theta)'D(\widehat{\theta}_{\text{OLS}} - \theta)$ is a function of $\widehat{\beta}_{\text{OLS}}$ since $\widehat{\theta}_{\text{OLS}} = H\widehat{\beta}_{\text{OLS}}$

While $w_0 = \frac{\widehat{u}'\widehat{u}}{\sigma^2}$ is a function of $\widehat{u}$

Conditional on X, we showed previously that any function of $\widehat{u}$ and any function of $\widehat{\beta}_{\text{OLS}}$ are independent

So conditional on X, the scalar

$$v = \frac{(w/p)}{(w_0/(n-K))}$$

is the ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom

Hence we have

$$v|X \sim F(p, n-K)$$

And since this conditional distribution of $v|X$ is the same distribution (F-distribution with p and $n-K$ degrees of freedom) for any realization of X, we again have that the unconditional distribution is the same, i.e.

$$v \sim F(p, n-K)$$

We have shown that the scalar random variable

$$v = \left(\frac{1}{p\widehat{\sigma}^2_{\text{OLS}}}\right)(\widehat{\theta}_{\text{OLS}} - \theta)'D(\widehat{\theta}_{\text{OLS}} - \theta)$$

$$= \left(\frac{1}{p}\right)(\widehat{\theta}_{\text{OLS}} - \theta)'[\widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X)]^{-1}(\widehat{\theta}_{\text{OLS}} - \theta) \sim F(p, n-K)$$

where $\theta = H\beta$ and $\widehat{\theta}_{\text{OLS}} = H\widehat{\beta}_{\text{OLS}}$

We can use this result to test p linear restrictions involving one or more of the elements of the parameter vector β

Notice that our earlier result is a special case

Letting H be the $1 \times K$ matrix

$$H = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$

with the scalar 1 as its k^{th} element and zeros elsewhere, we have $\theta = H\beta = \beta_k$ and $\widehat{\theta}_{\text{OLS}} = \widehat{\beta}_k$

$$\begin{aligned} \text{Then } v &= \left(\frac{1}{p}\right)(\widehat{\theta}_{\text{OLS}} - \theta)'[\widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X)]^{-1}(\widehat{\theta}_{\text{OLS}} - \theta) \\ &= \frac{(\widehat{\beta}_k - \beta_k)^2}{\widehat{v}_{kk}} \\ &= t_k^2 \sim F(1, n-K) \end{aligned}$$

This holds, as the square of a random variable with a $t(n-K)$ distribution was shown to have a $F(1, n-K)$ distribution

To test a hypothesis about p linear combinations of the β_k parameters, we formulate the null hypothesis in the form

$$H_0 : H\beta = \theta^0$$

where H is a non-stochastic $p \times K$ matrix with rank$(H) = p$ (full row rank)

The alternative hypothesis is simply

$$H_1 : H\beta \neq \theta^0$$

We use the estimator $\widehat{\theta}_{\text{OLS}} = H\widehat{\beta}_{\text{OLS}}$ and construct the scalar test statistic

$$v = \left(\frac{1}{p}\right)(\widehat{\theta}_{\text{OLS}} - \theta^0)'[\widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X)]^{-1}(\widehat{\theta}_{\text{OLS}} - \theta^0)$$

where $\widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X) = H\widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}}|X)H' = \widehat{\sigma}^2_{\text{OLS}}H(X'X)^{-1}H'$ in the classical linear regression model with normally distributed errors

If the null hypothesis is true, we have shown that

$$v \sim F(p, n-K)$$

We ask whether the value of the test statistic v that we obtain is an unlikely draw from this null distribution

Again we reject $H_0 : H\beta = \theta^0$ against $H_1 : H\beta \neq \theta^0$ if the probability of drawing a test statistic with a value as high as v from an F-distribution with $(p, n - K)$ degrees of freedom is below our chosen level of significance, conventionally 5%

Note that the quadratic form for v implies that the scalar v is positive, so here we are only concerned with the upper tail of the $F(p, n-K)$ distribution

We reject $H_0 : H\beta = \theta^0$ at the 5% significance level if the test statistic v lies above the upper 95% percentile of the $F(p, n - K)$ distribution

Examples

Consider the model with $K = 3$ and $x_{1i} = 1$ for all $i = 1, ..., n$

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

1) To test the joint hypothesis that $\beta_2 = 0$ and $\beta_3 = 0$ against the alternative that $\beta_2 \neq 0$ or $\beta_3 \neq 0$, we formulate the null hypothesis as

$$H_0 : H\beta = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

We use

$$\widehat{\theta}_{\text{OLS}} = H\widehat{\beta}_{\text{OLS}} = \begin{pmatrix} \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix}$$

and

$$\begin{aligned} \widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X) &= H\widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}}|X)H' \\ &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{v}_{11} & \widehat{v}_{12} & \widehat{v}_{13} \\ \widehat{v}_{12} & \widehat{v}_{22} & \widehat{v}_{23} \\ \widehat{v}_{13} & \widehat{v}_{23} & \widehat{v}_{33} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \widehat{v}_{22} & \widehat{v}_{23} \\ \widehat{v}_{23} & \widehat{v}_{33} \end{pmatrix} \end{aligned}$$

To test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, we construct the test statistic

$$
\begin{aligned}
v &= \left(\frac{1}{p}\right)(\widehat{\theta}_{\text{OLS}} - \theta^0)'[\widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X)]^{-1}(\widehat{\theta}_{\text{OLS}} - \theta^0) \\
&= \left(\frac{1}{2}\right)\begin{pmatrix} \widehat{\beta}_2 - 0 \\ \widehat{\beta}_3 - 0 \end{pmatrix}'\begin{pmatrix} \widehat{v}_{22} & \widehat{v}_{23} \\ \widehat{v}_{23} & \widehat{v}_{33} \end{pmatrix}^{-1}\begin{pmatrix} \widehat{\beta}_2 - 0 \\ \widehat{\beta}_3 - 0 \end{pmatrix} \\
&= \left(\frac{1}{2}\right)\left(\widehat{\beta}_2, \widehat{\beta}_3\right)\begin{pmatrix} \widehat{v}_{22} & \widehat{v}_{23} \\ \widehat{v}_{23} & \widehat{v}_{33} \end{pmatrix}^{-1}\begin{pmatrix} \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix} \\
&\sim F(2, n-K) \quad \text{under } H_0 : \theta = 0
\end{aligned}
$$

We reject $H_0 : \theta = 0$ at the 5% level of significance if the probability of drawing this value of v from the $F(2, n-K)$ distribution is less than 0.05, i.e. if v lies above the upper 95% percentile of the $F(2, n-K)$ distribution

For example, if $n - K = 30$, the upper 95% percentile of the $F(2, 30)$ distribution is 3.32, and values of the test statistic $v > 3.32$ would reject the null hypothesis at the 5% level

2) To test the null hypothesis that $\beta_2 + \beta_3 = 1$ against the alternative hypothesis that $\beta_2 + \beta_3 \neq 1$, we formulate the null hypothesis as

$$H_0 : H\beta = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \beta_2 + \beta_3 = 1$$

We use

$$\widehat{\theta}_{\text{OLS}} = H\widehat{\beta}_{\text{OLS}} = \widehat{\beta}_2 + \widehat{\beta}_3$$

and

$$\begin{aligned}
\widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X) &= H\widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}}|X)H' \\
&= \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \widehat{v}_{11} & \widehat{v}_{12} & \widehat{v}_{13} \\ \widehat{v}_{12} & \widehat{v}_{22} & \widehat{v}_{23} \\ \widehat{v}_{13} & \widehat{v}_{23} & \widehat{v}_{33} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \\
&= \widehat{v}_{22} + 2\widehat{v}_{23} + \widehat{v}_{33} \quad \text{(as expected)}
\end{aligned}$$

To test $H_0 : \theta = \beta_2 + \beta_3 = 1$ against $H_1 : \theta = \beta_2 + \beta_3 \neq 1$, we construct the test statistic

$$\begin{aligned} v &= \left(\frac{1}{p}\right)(\widehat{\theta}_{\text{OLS}} - \theta^0)'[\widehat{\text{Var}}(\widehat{\theta}_{\text{OLS}}|X)]^{-1}(\widehat{\theta}_{\text{OLS}} - \theta^0) \\ &= \frac{(\widehat{\beta}_2 + \widehat{\beta}_3 - 1)^2}{\widehat{v}_{22} + 2\widehat{v}_{23} + \widehat{v}_{33}} \\ &\sim F(1, n-K) \quad \text{under } H_0 : \theta = \beta_2 + \beta_3 = 1 \end{aligned}$$

Again we reject $H_0 : \beta_2 + \beta_3 = 1$ against $H_1 : \beta_2 + \beta_3 \neq 1$ at the 5% level if the value of the test statistic v lies above the upper 95% percentile of the $F(1, n-K)$ distribution

In this special case with $p = 1$, we could also use the result that

$$\begin{aligned}\sqrt{v} &= \frac{\widehat{\beta}_2 + \widehat{\beta}_3 - 1}{\sqrt{\widehat{v}_{22} + 2\widehat{v}_{23} + \widehat{v}_{33}}} \\ &\sim t(n-K) \quad \text{under } H_0 : \beta_2 + \beta_3 = 1\end{aligned}$$

We reject $H_0 : \beta_2 + \beta_3 = 1$ against the two-sided alternative $H_1 : \beta_2 + \beta_3 \neq 1$ at the 5% significance level if the absolute value of the test statistic $|\sqrt{v}|$ lies above the upper 97.5% percentile of the $t(n-K)$ distribution

In this special case, one-sided tests can also be considered

Using linear transformations to simplify hypothesis tests

In this last example, another possibility is to re-parameterize the model so that the hypothesis of interest can be tested using a simple t-test on a single regression coefficient

We have the model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

and wish to test $H_0 : \beta_2 + \beta_3 = 1$ against $H_1 : \beta_2 + \beta_3 \neq 1$

Subtracting x_{3i} from both sides of the model gives

$$y_i - x_{3i} = \beta_1 + \beta_2 x_{2i} + (\beta_3 - 1)x_{3i} + u_i$$

Now subtracting $\beta_2 x_{3i}$ from the second term on the right-hand side, and adding $\beta_2 x_{3i}$ to the third term on the right-hand side, gives

$$y_i - x_{3i} = \beta_1 + \beta_2(x_{2i} - x_{3i}) + (\beta_2 + \beta_3 - 1)x_{3i} + u_i$$

So if we regress $(y_i - x_{3i})$ on $(x_{2i} - x_{3i})$ and x_{3i}, and an intercept, we can test $H_0 : \beta_2 + \beta_3 = 1 \leftrightarrow \beta_2 + \beta_3 - 1 = 0$ using a simple t-test of the null hypothesis that the coefficient on x_{3i} is equal to zero, in this re-parameterization of the model

If we do not reject this hypothesis and wish to impose it, this also indicates that the restricted model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

$$\text{subject to } \beta_2 + \beta_3 = 1$$

can be estimated quite straightforwardly using the representation

$$y_i - x_{3i} = \beta_1 + \beta_2 (x_{2i} - x_{3i}) + u_i$$

This gives the restricted estimates of β_1 and β_2 directly as $\widehat{\beta}_1$ and $\widehat{\beta}_2$

We obtain the restricted estimate of β_3 simply as $\widehat{\beta}_3 = 1 - \widehat{\beta}_2$

Goodness of fit

Note that $y = \widehat{y} + \widehat{u}$, and that

$$\widehat{y}'\widehat{u} = (X\widehat{\beta}_{\text{OLS}})'\widehat{u} = (\widehat{\beta}_{\text{OLS}}X')\widehat{u} = \widehat{\beta}_{\text{OLS}}(X'\widehat{u}) = 0$$

since the first order conditions used to obtain $\widehat{\beta}_{\text{OLS}}$ imply that $X'\widehat{u} = 0$

The OLS residuals $\widehat{u}$ are orthogonal to both the explanatory variables X and to the fitted values $\widehat{y} = X\widehat{\beta}_{\text{OLS}}$ in the sample

Consider

$$\begin{aligned} y'y &= (\widehat{y} + \widehat{u})'(\widehat{y} + \widehat{u}) = \widehat{y}'\widehat{y} + \widehat{y}'\widehat{u} + \widehat{u}'\widehat{y} + \widehat{u}'\widehat{u} \\ &= \widehat{y}'\widehat{y} + \widehat{u}'\widehat{u} \end{aligned}$$

$$y'y = \widehat{y}'\widehat{y} + \widehat{u}'\widehat{u}$$

Or equivalently

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} \widehat{u}_i^2$$

This is an analysis (or decomposition) of sums of squares

Subtracting $n\overline{y}^2$ from both sides gives

$$\left(\sum_{i=1}^{n} y_i^2\right) - n\overline{y}^2 = \left(\sum_{i=1}^{n} \widehat{y}_i^2\right) - n\overline{y}^2 + \left(\sum_{i=1}^{n} \widehat{u}_i^2\right)$$

We also have $y_i = \widehat{y}_i + \widehat{u}_i$ for $i = 1, ..., n$, which implies

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i + \sum_{i=1}^{n} \widehat{u}_i$$

and hence (dividing by n)

$$\overline{y} = \overline{\widehat{y}} + \overline{\widehat{u}}$$

Now *provided* we have $\overline{\widehat{u}} = 0$, we obtain

$$\overline{y} = \overline{\widehat{y}}$$

$$\left(\sum_{i=1}^{n} y_i^2\right) - n\overline{y}^2 = \left(\sum_{i=1}^{n} \widehat{y}_i^2\right) - n\overline{y}^2 + \left(\sum_{i=1}^{n} \widehat{u}_i^2\right)$$

Note that $\left(\sum_{i=1}^{n} y_i^2\right) - n\overline{y}^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2$

Provided $\overline{\widehat{u}} = 0$ s.t. $\overline{y} = \overline{\widehat{y}}$, we also have $\left(\sum_{i=1}^{n} \widehat{y}_i^2\right) - n\overline{y}^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$

We then obtain

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} \widehat{u}_i^2 \\ TSS &= ESS + RSS \end{aligned}$$

This is now an analysis (or decomposition) of variation (squared deviations from sample means)

Note that we require the average of the sample residuals to be zero in order to equate the sample means of the predicted values $\overline{\widehat{y}}$ and the actual values $\overline{y}$

This condition will always be satisfied if the model includes an intercept term

Dividing both sides by $\sum_{i=1}^{n}(y_i - \overline{y})^2$ gives

$$1 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} + \frac{\sum_{i=1}^{n}\widehat{u}_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

or rearranging

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}\widehat{u}_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

$$= \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Note that $0 \leqslant R^2 \leqslant 1$, *provided* we have $\overline{\widehat{u}} = 0$

R^2 can then also be shown to be the square of the sample correlation coefficient between y_i and $\widehat{y}_i$

R^2 measures the proportion of the total variation in the dependent variable that is 'accounted for' or 'explained' by the model

$R^2 = 1$ (perfect fit) occurs when $y = X\widehat{\beta}_{\text{OLS}}$ so that $y - \widehat{y} = \widehat{u} = 0$ and so $\widehat{u}'\widehat{u} = \sum_{i=1}^{n} \widehat{u}_i^2 = 0$

$R^2 = 0$ occurs when each element of $\widehat{y} = X\widehat{\beta}_{\text{OLS}} = \overline{y}$ for $i = 1, 2, ..., n$, so that $\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = 0$

Note that this outcome occurs when the model contains *only* an intercept term

If X is an $n \times 1$ vector with each element equal to 1, then $X'\widehat{u} = \sum_{i=1}^{n} \widehat{u}_i$

Setting $X'\widehat{u} = 0$ is thus equivalent to setting $\sum_{i=1}^{n} \widehat{u}_i = 0$

Letting the coefficient on the intercept be β_1, this gives

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_1) = 0 \leftrightarrow n\widehat{\beta}_1 = \sum_{i=1}^{n} y_i \leftrightarrow \widehat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$$

Hence $\widehat{y}_i = \widehat{\beta}_1 \times 1 = \overline{y}$ for $i = 1, 2, ..., n$

R^2 thus measures the 'goodness of fit' of a model with additional explanatory variables *relative* to the fit of a model that includes only an intercept term

The inclusion of an intercept term in a linear model estimated using OLS ensures that $\sum_{i=1}^{n} \widehat{u}_i = 0 \leftrightarrow \overline{\widehat{u}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i = 0$, which in turn ensures that R^2 lies between zero and one

For linear models that do not include an intercept term (or do not otherwise ensure that $\overline{\widehat{u}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i = 0$), R^2 should not be used as a measure of 'goodness of fit'

Note that in a sample of size n, we could always obtain a perfect fit ($R^2 = 1$) simply by regressing y_i on a set of n linearly independent explanatory variables

More generally, increasing the number of explanatory variables cannot reduce the fit as measured by R^2

Adding an explanatory variable that is irrelevant in the sample (s.t. the estimated coefficient on this variable is zero) leaves the fit unchanged

While adding a variable that is not irrelevant in the sample (s.t. the estimated coefficient is not zero) improves the fit

Some researchers prefer to report an 'adjusted R^2', defined by

$$(1 - \overline{R}^2) = \frac{(n-1)(1-R^2)}{(n-K)}$$

which discounts this improvement in the fit as K increases in a given sample size n

Note that there is no strong theoretical rationale for this particular adjustment

A disadvantage of the R^2 measure of goodness of fit is that the value of R^2 is *not invariant to linear transformations* of the model

To see this, compare

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

and

$$(y_i - x_{3i}) = y_i^* = \beta_1 + \beta_2 x_{2i} + (\beta_3 - 1) x_{3i} + u_i$$

which is obtained by subtracting x_{3i} from both sides of the original model

These models are equivalent, and yield the same OLS estimates of $(\beta_1, \beta_2, \beta_3)$, noting that the coefficient on x_{3i} is $(\beta_3 - 1)$ in the transformed model

They have the same error term u_i, the same residuals $\widehat{u}_i$, and hence the same residual sum of squares $RSS = \sum_{i=1}^{n} \widehat{u}_i^2$

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

$$(y_i - x_{3i}) = y_i^* = \beta_1 + \beta_2 x_{2i} + (\beta_3 - 1) x_{3i} + u_i$$

However they have different dependent variables, and in general

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 \neq \sum_{i=1}^{n} (y_i^* - \overline{y}^*)^2$$

Hence they have different values for $R^2 = 1 - \frac{RSS}{TSS}$

The implication is that values of the R^2 statistic should not be used to *compare* the fit of models that have *different dependent variables*

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

$$(y_i - x_{3i}) = y_i^* = \beta_1 + \beta_2 x_{2i} + (\beta_3 - 1) x_{3i} + u_i$$

A more appropriate comparison in this situation would be provided by comparing estimates of $\widehat{\sigma}^2_{\text{OLS}} = \frac{\widehat{u}'\widehat{u}}{n-K}$ or $\widehat{\sigma}_{\text{OLS}}$, which provide unbiased estimates of the variance or standard deviation of the error term u_i

Using R^2 to test exclusion restrictions

Consider the linear model with $K = 4$ and $x_{1i} = 1$ for all $i = 1, ..., n$

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$$

Suppose we want to test the null hypothesis that $\beta_3 = \beta_4 = 0$

We can form a $F(2, n - K)$ test statistic in the usual way

An equivalent test can be obtained by comparing the R^2 from the unrestricted model given above and the R^2 from the restricted model

$$y_i = \beta_1 + \beta_2 x_{2i} + u_i$$

which imposes these two exclusion restrictions

Denote the R^2 obtained from the unrestricted model by R^2_U and denote the R^2 obtained from the restricted model by R^2_R, noting that $R^2_U \geqslant R^2_R$

Then we can also write the test statistic as

$$v = \left(\frac{n-K}{2}\right)\left(\frac{R^2_U - R^2_R}{1-R^2_U}\right) \sim F(2, n-K)$$

This extends naturally to testing p exclusion restrictions, replacing 2 by p

$$v = \left(\frac{n-K}{p}\right)\left(\frac{R^2_U - R^2_R}{1-R^2_U}\right) \sim F(p, n-K)$$

$$v = \left(\frac{n-K}{p}\right)\left(\frac{R_U^2 - R_R^2}{1-R_U^2}\right) \sim F(p, n-K)$$

This form of the F-test statistic is very intuitive

The null hypothesis is rejected if the exclusion of these p explanatory variables results in a sufficiently large fall in the R^2 goodness of fit measure, or in a sufficiently large deterioration in the fit of the model

This can be used to test the restriction that *all* $K-1$ of the slope coefficients in a linear model are equal to zero, i.e. to test the exclusion of all the explanatory variables except the intercept

The restricted model $y_i = \beta_1 + u_i$ has $R_R^2 = 0$

The test statistic simplifies to

$$v = \left(\frac{n-K}{K-1}\right)\left(\frac{R^2}{1-R^2}\right) \sim F(K-1, n-K)$$

where $R^2 = R_U^2$ denotes the R^2 from the unrestricted model

This is sometimes referred to as 'the F-test' for the model

Asymptotic results

If $t_k \sim t(n)$ then $t_k \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$ as $n \to \infty$

Write $t_k = \frac{z}{\sqrt{w/n}}$ where $z \sim \mathrm{N}(0,1)$ and $w \sim \chi^2(n)$

We know that $w = \sum_{i=1}^{n} z_i^2$ where $z_i \sim \mathrm{N}(0,1)$ for $i = 1, 2, ..., n$ independent standard normal random variables

Note that $\mathrm{E}(w) = n\mathrm{E}(z_i^2) = n \times 1 = n$ and $\mathrm{E}(w/n) = \mathrm{E}(w)/n = n/n = 1$

Moreover $w/n = \frac{1}{n}\sum_{i=1}^{n} z_i^2 \xrightarrow{\mathrm{P}} \mathrm{E}(z_i^2) = 1$ as $n \to \infty$ [applying a Law of Large Numbers]

Now since $t_k = \frac{z}{\sqrt{w/n}}$ and $(w/n) \xrightarrow{\text{P}} 1$ as $n \to \infty$, we have the result that t_k has the same limit distribution as $z \sim \text{N}(0,1)$ [applying generalized Slutsky's theorem]

In large samples (i.e. as $n \to \infty$ treating the number of parameters K as fixed, so that $(n - K) \to \infty$), to test one linear restriction, we can use critical values obtained from the standard normal distribution

Similarly if $v \sim F(m,n)$ then $mv \xrightarrow{\text{D}} \chi^2(m)$ as $n \to \infty$

In large samples, to test p linear restrictions, we can use the test statistic pv and critical values obtained from the $\chi^2(p)$ distribution

Write $v = \frac{w_1/m}{w_2/n}$ so that $mv = \frac{w_1}{w_2/n}$, where $w_1 \sim \chi^2(m)$ and $w_2 \sim \chi^2(n)$

Again $w_2 = \sum_{i=1}^{n} z_i^2$ with $z_i \sim \mathrm{N}(0,1)$ for $i = 1, 2, ..., n$ independent standard normal random variables, so that $\mathrm{E}(w_2) = n$ and $\mathrm{E}(w_2/n) = 1$

Again we have that $(w_2/n) = \frac{1}{n}\sum_{i=1}^{n} z_i^2 \xrightarrow{\mathrm{P}} \mathrm{E}(z_i^2) = 1$ as $n \to \infty$, so that $mv = \frac{w_1}{w_2/n}$ has the same limit distribution as $w_1 \sim \chi^2(m)$ [again applying a Law of Large Numbers and generalized Slutsky's theorem]