1 Getting an estimate of $V(\widehat{\beta})$

Let

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ E[X'\varepsilon] &= 0_K, \\ E[\varepsilon\varepsilon'] &= \sigma^2 I_N, \end{aligned}$$

and let X be full rank with a rank of K.

We know that

$$V(\widehat{\beta}) = \sigma^2 (X'X)^{-1},$$

and this is fine if we know σ^2, but if we don't where do we get an estimate of it, and what does that do to our variance?

A natural estimator for σ^2 is the average of the square residuals, where residuals are given by

$$\begin{aligned} e &= Y - X\widehat{\beta} \\ &= Y - P_X Y \\ &= (I_N - P_X)Y \\ &= M_X Y \end{aligned}$$

where M_Z is the "residual projection matrix" that creates the residuals from a regression of something on Z:

$$M_Z = I - Z(Z'Z)^{-1}Z' = I - P_Z$$

The reason it is called a projection matrix is that it has 3 familiar properties:

1. M_Z is symmetric
2. M_Z is idempotent
3. M_Z has a known rank: rank(M_Z)=rank(I)-rank(Z)

Knowing that the residual vector can be written as a projection matrix times Y is helpful when it comes to figuring out how the average of squared residuals relates to σ^2.

First, we must establish that the residuals can be written in terms of ε:

$$\begin{aligned} e &= M_X Y = M_X X\beta + M_X \varepsilon \\ &= I_N X\beta - X(X'X)^{-1}X'X\beta + M_X\varepsilon \\ &= X\beta - X\beta + M_X\varepsilon \\ &= M_X\varepsilon. \end{aligned}$$

That $M_X X\beta = 0$ should not be surprising: what are the residuals from a regression of X on X? Big fat zero.

Consider the sum of squared residuals:

$$e'e,$$

and write it in terms of the residual projection matrix and Y:

$$\begin{aligned} e'e &= \varepsilon' M_X' M_X \varepsilon \\ &\quad \varepsilon' M_Y \varepsilon \\ &= \varepsilon'\varepsilon - \varepsilon' P_X \varepsilon. \end{aligned}$$

The expectation of $e'e$ can be written by using our knowledge of the variance of ε:

$$\begin{aligned} E[e'e] &= E[\varepsilon'\varepsilon] - E[\varepsilon' P_X \varepsilon] \\ &= N\sigma^2 - rank(P_X)\sigma^2 \\ &= (N-K)\sigma^2. \end{aligned}$$

Why does $E[\varepsilon' P_X \varepsilon] = K\sigma^2$? How much ε could you predict with X. Well, since X and ε are uncorrelated, you can only pick up what regression does mechanically. Consequently, with K columns in X, you could get K perfect fits of ε.

Mathematically, we have

$$E[\varepsilon' P_X \varepsilon] = E[tr(\varepsilon' P_X \varepsilon)]$$

because $\varepsilon' P_X \varepsilon$ is a scalar,

$$E[tr(\varepsilon' P_X \varepsilon)] = trE[(P_X \varepsilon\varepsilon')]$$

because we can rearrange within a trace and because the trace is a linear operator,

$$trE[(P_X \varepsilon\varepsilon')] = tr(P_X E[(\varepsilon\varepsilon')]) = \sigma^2 I_N tr(P_X) = K\sigma^2,$$

because P_X is a fixed matrix and because $E[(\varepsilon\varepsilon')] = \sigma^2 I_N$, and because trace=rank.

A consequence of the above is that we can define an unbiased estimator s^2 of σ^2 as

$$\begin{aligned} s^2 &= e'e/N - K \\ E[s^2] &= \sigma^2, \end{aligned}$$

and define an estimate of the variance of the OLS estimator as

$$\begin{aligned} \widehat{V}(\widehat{\beta}) &= \frac{e'e}{N-K}(X'X)^{-1} \\ &= s^2(X'X)^{-1}, \end{aligned}$$

and this is an unbiased estimate of the variance of the OLS estimator. If you divided by N instead of $N-K$, the estimator would only be consistent (unbiased asymptotically), because asymptotically, $N = N - K$.

2 Digression: What is the *normal* distribution?

The pdf of the normal distribution, and standard normal distribution, are

$$N(\mu,\sigma^2) \Leftrightarrow f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}\varphi\left(\frac{(x-\mu)}{\sigma}\right)$$

$$N(0,1) \Leftrightarrow f(x;0,1) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

We can evaluate the expectation of any power of this random variable.

a chi-square distribution with J degrees of freedom, denoted χ^2_J, is the sum of J squared standard normals.

Why are people interested in Normals? Normals have some useful features. A very useful one is that linear functions of normals are also normal.

Consider a normally distributed random vector x

$$x \tilde{} N(\mu,\Sigma)$$

were μ is a $J-$vector and Σ is a $J \times J$ symmetric positive definite covariance matrix. This thing can be expressed in terms of J standard normal scalars as follows:

$$\Sigma^{-1/2}(x-\mu) \tilde{} \begin{bmatrix} N(0,1) \\ ... \\ N(0,1) \end{bmatrix}.$$

Linear combinations of normals $Ax+b$ are distributed

$$Ax + b \tilde{} N(A'\mu + b, A\Sigma A'),$$

which, since it is also a normal, is also completely specified up to the expectation of every power.

One can check that

$$\begin{aligned} \Sigma^{-1/2}(x-\mu) &= \Sigma^{-1/2}x - \Sigma^{-1/2}\mu \tilde{} N\left(\Sigma^{-1/2}\mu - \Sigma^{-1/2}\mu, \Sigma^{-1/2}\Sigma\Sigma^{-1/2}\right) \\ &\tilde{} N(0, I_J), \end{aligned}$$

which matches up to the claim 2 equations up.

Also, (finite-valued finite-length vector-) functions of non-normal finite-variance random vectors have approximately normal limiting distributions. This means that normal distributions show up a lot in asymptotic theory.

3 Where's β: Confidence Intervals

An $\alpha\%$ *confidence interval* for a parameter β is the smallest range such that there is an $\alpha\%$ probability that β lies in that range. Here, β is an unknown fixed parameter (or parameter vector), and $\widehat{\beta}$ is our estimate of it. We will use

the sampling distribution of $\widehat{\beta}$ to construct endpoints of the confidence interval for β. Those endpoints are random variables, and they center on $\widehat{\beta}$.

We have been working with estimated coefficient vectors ($\widehat{\beta}$), which are random variables, and we have learned some stuff about their sampling distributions. We have worked out the mean and variance of $\widehat{\beta}$, e.g., when

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ E\left[X'\varepsilon\right] &= 0_K, \\ E\left[\varepsilon\varepsilon'\right] &= \sigma^2 I_N, \end{aligned}$$

and X is full rank with rank K, the sampling distribution of $\widehat{\beta}$ has

$$\begin{aligned} E\left[\widehat{\beta}\right] &= \beta, \\ V\left[\widehat{\beta}\right] &= \sigma^2 (X'X)^{-1}. \end{aligned}$$

There are many sampling distributions that could match these features. To illustrate, consider a scalar $\widehat{\beta}$. These moments could be matched with $N(\beta, \frac{\sigma^2}{\sum X_i^2})$. A uniform distribution over the range a, b has a mean of $(b+a)/2$ and a variance of $(b-a)^2/12$. Thus, a uniform over the range $\left(\beta - \sqrt{\frac{3\sigma^2}{\sum X_i^2}}\right)$, $\left(\beta + \sqrt{\frac{3\sigma^2}{\sum X_i^2}}\right)$ would also have a mean of β and a variance of $\frac{\sigma^2}{\sum X_i^2}$.

The fact that $E\left[\widehat{\beta}\right] = \beta$ tells us to center our confidence interval for β on $\widehat{\beta}$. A 95% confidence interval for this fixed scalar β would be $\left[\widehat{\beta} - 1.96\sqrt{\frac{\sigma^2}{\sum X_i^2}}, \widehat{\beta} + 1.96\sqrt{\frac{\sigma^2}{\sum X_i^2}}\right]$ if $\widehat{\beta}$ were distributed normally. We compute this by finding the range of the normal distribution, centered on the mean, that covers 95% of the pdf. We would say "there is a 95% coverage probability for β in the range $\left[\widehat{\beta} - 1.96\sqrt{\frac{\sigma^2}{\sum X_i^2}}, \widehat{\beta} + 1.96\sqrt{\frac{\sigma^2}{\sum X_i^2}}\right]$".

But, if $\widehat{\beta}$ were distributed uniformly, then its distribution would be uniform over the range $\left(\beta - \sqrt{\frac{3\sigma^2}{\sum X_i^2}}\right)$, $\left(\beta + \sqrt{\frac{3\sigma^2}{\sum X_i^2}}\right)$. The 95% confidence interval for this fixed scalar β would just be the middle 95% of a range of that width, centered on $\widehat{\beta}$: $\left[\left(\widehat{\beta} - 0.95\sqrt{\frac{3\sigma^2}{\sum X_i^2}}\right), \left(\widehat{\beta} + 0.95\sqrt{\frac{3\sigma^2}{\sum X_i^2}}\right)\right]$. This is a different confidence band from that corresponding to the normally distributed version.

The point is that we need to know more about the distribution that just the mean the variance to compute the confidence band. We need to know everything about the pdf, which is equivalent to knowing the value of *every* moment of the random variable, not just the first two (mean and variance).

3.1 A Cheap Trick: Normally Distributed Disturbances

Suppose

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ &\quad \varepsilon \tilde{} N(0_N, \sigma^2 I_N), \end{aligned}$$

and X is full rank with rank K. The fact that the disturbances are independent mean-zero normals, $\varepsilon \tilde{} N(0, \sigma^2 I_N)$, implies $E[X'\varepsilon] = 0_K$ and $E[\varepsilon\varepsilon'] = \sigma^2 I_N$, so the OLS estimator still satisfies unbiasedness and has the simple variance matrix:

$$\begin{aligned} E\left[\widehat{\beta}\right] &= \beta, \\ E\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'\right] &= \sigma^2(X'X)^{-1}. \end{aligned}$$

But, now having assumed more than we did before, we get more than we got before. Write out $\widehat{\beta}$ as a function of ε:

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'Y \\ &= \beta + (X'X)^{-1}X'\varepsilon \end{aligned}$$

is a linear combination of a normally distributed vector. Since, for any vector $x \tilde{} N(\mu, \Sigma)$, $(a + Ax) \tilde{} N(a + \mu, A\Sigma A')$, we have

$$\begin{aligned} &\widehat{\beta} \tilde{} N\left(\beta + 0_K, (X'X)^{-1}X'\sigma^2 I_N X(X'X)^{-1}\right), \text{ or} \\ &\widehat{\beta} \tilde{} N\left(\beta, \sigma^2(X'X)^{-1}\right). \end{aligned}$$

If the disturbances are normally distributed, then $\widehat{\beta}$ is normally distributed. If we know σ^2, then we have enough information to construct a confidence interval. If we don't know σ^2, then we can still construct an unbiased estimate of it, s^2, and use that in place of σ^2 to construct the confidence interval. The confidence band which uses the normal distribution and s^2 in place of σ^2 would only be valid asymptotically, because $s^2 = \sigma^2$ only in the limit.

3.2 Another Cheap Trick: Central Limit Theorem

Suppose that

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ E[X'\varepsilon] &= 0_K, \\ E[\varepsilon\varepsilon'] &= \sigma^2 I_N, \end{aligned}$$

and X is full rank with rank K. The OLS estimator is linear combination of the random vector ε, and the length of $\widehat{\beta}$ does not grow with N: $\widehat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$. Consequently, one can invoke a central limit theorem: Since $\widehat{\beta}$ is a finite length vector function of a finite-variance random vector ε, as the length of ε grows to infinity, the second-order approximation of the vector function $\widehat{\beta}$ converges to the normal distribution. Thus, as N gets large, $\widehat{\beta}$ looks approximately normal,

$$\widehat{\beta} \underset{N\to\infty}{\tilde{}}^{approx} N\left(\beta, \sigma^2(X'X)^{-1}\right).$$

Here, we did not have to invoke normality of the disturbances. Rather, we required the disturbances to have finite variance, so that when you 'add a lot

of them up', you still get something with finite variance, and a Central Limit Theorem then tells you that the summation goes to something that looks pretty much like a normal.

It turns out that you can substitute in s^2 for σ^2, and it does not change the approximation of the limiting distribution, so that

$$\widehat{\beta} \overset{approx}{\underset{N\to\infty}{\sim}} N\left(\beta, s^2(X'X)^{-1}\right).$$

These methods for constructing confidence intervals allow us to get an idea of where the true β is on the basis of our observed $\widehat{\beta}$ and its variance. However, since in practise we don't observe σ^2, we must use an estimate of it, usually s^2. Thus, we typically end up with a confidence interval that is only valid asymptotically, and possibly even then only approximately.

4 Constructing a test: Tests of Equalities

There are 3 steps:

1. First specify a Null Hypothesis, usually denoted H_0, which describes a model of interest. Usually, we express H_0 as a restricted version of a more general model. In the background, we have a model

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ E\left[X'\varepsilon\right] &= 0_K, \\ E\left[\varepsilon\varepsilon'\right] &= \sigma^2 I_N, \end{aligned}$$

   where X is NxK and β is a $K-$vector, where interesting hypotheses can be expressed as restrictions on β. We will consider 3 types of tests of equalities: single linear, multiple linear, and general nonlinear. Tests of equalities are fully specified when you specify the Null hypothesis: the Null is either true or not true, and you don't care how exactly it isn't true, just that it isn't true.

   (a) A single linear test could be written as

$$R\beta + r = 0,$$

   where R is $1xK$ and r is a scalar.

      i. An exclusion restriction, e.g., that the second variable does not belong in the model would have

$$\begin{aligned} R &= \left[\begin{array}{ccccc} 0 & 1 & 0 & ... & 0 \end{array}\right], \\ r &= 0. \end{aligned}$$

      ii. A symmetry restriction, e.g., that the second and third variables had identical effects, would have

$$\begin{aligned} R &= \left[\begin{array}{cccccc} 0 & -1 & 1 & 0 & ... & 0 \end{array}\right], \\ r &= 0. \end{aligned}$$

iii. A value restriction, e.g, that the second variable's coefficient is 1, would have

$$\begin{aligned} R &= \begin{bmatrix} 0 & 1 & 0 & ... & 0 \end{bmatrix}, \\ r &= -1. \end{aligned}$$

(b) A multiple linear test could be written as

$$R\beta + r = 0,$$

where R is JxK and r is a $J-$vector.

i. A set of exclusion restrictions, e.g., that the second and third variables do not belong in the model, would have

$$\begin{aligned} R &= \begin{bmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \end{bmatrix}, \\ r &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

ii. A set of symmetry restrictions, that the first, second and third variables all have the same coefficients, would have

$$\begin{aligned} R &= \begin{bmatrix} 1 & -1 & 0 & ... & 0 \\ 0 & 1 & -1 & ... & 0 \end{bmatrix}, \\ r &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

iii. Given that we write the restriction as $R\beta + r = 0$ for both single and multiple linear hypotheses, you can think of the single hypothesis as a case of the multiple hypothesis.

(c) A multiple nonlinear test could be written as

$$c(\beta) = 0,$$

where c is a $J-$vector function of β.

i. A restriction that the product of the first and second coefficients equals 1 and the product of the third and fourth coefficients equals 1 would have

$$c(\beta) = \begin{bmatrix} \beta_1\beta_2 - 1 \\ \beta_3\beta_4 - 1 \end{bmatrix}.$$

(d) Of course you can think of single- and multiple-linear hypotheses as cases of the multiple nonlinear test.

2. Then, construct a test statistic, which is a random variable (because it is a function of other random variables) with two features:

(a) it has a known distribution under the Null Hypothesis (usually, normal or chi-square, t or F). Its distribution is known either because we assume enough about the distribution of the model disturbances to get small-sample distributions, or we assume enough to get asymptotic distributions.

(b) this known distribution may depend on data, but not on parameters (this is called pivotality: a test statistic is pivotal if it satisfies this condition).

3. Check whether or not the sample value of the test statistic is very far out in its sampling distribution. If it is very far out, then you are left with 2 options: (1) the Null hypothesis is true, but you got a really weird draw of ε leading to a really weird value of the test statistic; or (2) the Null hypothesis is false.

4.1 The Wald Test

A common test statistic, called the Wald Statistic, asks whether the hypothesis, evaluated at the sample value of $\widehat{\beta}$, is very far out in its sampling distribution. The *discrepancy vector*, $\widehat{d}$, is the sample value of the hypothesis, that is, the value of H_0 evaluated at the sample estimate of β, $\widehat{\beta}$. Using the terminology above, for a linear hypothesis,

$$\widehat{d} = R\widehat{\beta} - r,$$

and for a nonlinear hypothesis,

$$\widehat{d} = c\left(\widehat{\beta}\right).$$

Even if the hypothesis is true, we would not expect $\widehat{d}$ to be exactly zero, because $\widehat{\beta}$ is not exactly equal to β. However, if the hypothesis is true, we would expect $\widehat{d}$ to be close to 0. In contrast if the hypothesis is false, we'd have no real prior about where we'd see $\widehat{d}$.

The *Wald Statistic* is a random variable, which is a function of the data and the Null hypothesis. The Wald test asks whether or not $\widehat{d}$ is very far from zero, given the assumption that the null hypothesis is true. It does so by creating the Wald Statistic, which has a known and pivotal distribution under the Null hypothesis, and asking whether its value is very far out in the tails of the distribution.

Consider a Wald Statistic for a multiple hypothesis $R\beta - r = 0$, so that

$$\widehat{d} = R\widehat{\beta} - r.$$

To evaluate whether or not $\widehat{d}$ is very far from its hypothesized value of 0, we need to figure out its sampling distribution. Luckily, $\widehat{d}$ is a linear function of

$\widehat{\beta}$, and we know the mean and variance of $\widehat{\beta}$:

$$\begin{aligned} E[\widehat{\beta}] &= \beta \\ V\left[\widehat{\beta}\right] &= \sigma^2(X'X)^{-1}, \end{aligned}$$

and so

$$\begin{aligned} E[\widehat{d}] &= E\left[R\widehat{\beta} - r\right] \\ V\left[\widehat{d}\right] &= \sigma^2 R(X'X)^{-1}R', \end{aligned}$$

and since under the Null Hypothesis, $R\beta - r = 0$, we can substitute that in to get

$$\begin{aligned} E[\widehat{d}] &= 0 \\ V\left[\widehat{d}\right] &= \sigma^2 R(X'X)^{-1}R'. \end{aligned}$$

Unfortunately, this is not enough information to really pin down whether or not our observed value of $\widehat{d}$ is very far out in the sampling distribution of $\widehat{d}$, because many different sampling distributions could have this mean and variance. So, we need more.

Assume instead that

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ &\quad \varepsilon \tilde{} N(0_N, \sigma^2 I_N), \end{aligned}$$

so that

$$\widehat{\beta} \tilde{} N\left(\beta, \sigma^2(X'X)^{-1}\right).$$

Then,

$$\widehat{d} \tilde{} N(0, \sigma^2 R(X'X)^{-1}R')$$

under the Null Hypothesis. (Its distribution is unknown if the Null is not true, because we wouldn't know its mean.) This is enough information to figure out if $\widehat{d}$ is very far out in the tails of its sampling distribution.

There is a problem, though: $\widehat{d}$ is not a test statistic—its distribution is known, but it is not pivotal (since it depends on parameters). Thus, we need to get the parameters out of the sampling distribution.

Since $\widehat{d}$ is normally distributed, any linear combination of it is normally distributed, and there is a particular linear combination that turns it into a standard normal vector:

$$T_{Wv} = \frac{1}{\sigma}\left(R(X'X)^{-1}R'\right)^{-1/2}\widehat{d} \tilde{} N(0_J, I_J)$$

Since $(X'X)^{-1}$ is positive definite, $R(X'X)^{-1}R'$ is also positive definite, so the "minus one-half" matrix of $R(X'X)^{-1}R'$ exists.

T_{Wv} (for "Wald vector") could be used as a test statistic, because its distribution does not depend on any parameters. Unfortunately, it is comprised of things that can deviate from zero in both directions, so we need a way to aggregate it. Somehow we need to detect whether or not any element of W_v is very far from zero.

The Wald Statistic is its sum-of-squares $T'_{Wv}T_{Wv}$. It is distributed as the sum of J squared normals (aka, χ^2_J):

$$T_W = T'_{Wv}T_{Wv} = \frac{1}{\sigma^2}\widehat{d}' \left(R(X'X)^{-1}R'\right)^{-1} \widehat{d}\tilde{} \chi^2_J.$$

If we knew σ^2, we could compare the sample value of the Wald Statistic, T_W, to the χ^2_J distribution and ask whether or not it is far out in the tails of the distribution.

4.2 How Far is Far?

How can we judge whether or not T_W is far out in the distribution of the χ^2_J? There are 2 basic strategies, both based on the distinction between *Type I* and *Type II errors*.

1. A *Type I Error* is when we reject the Null even though it is true. The probability of a type I error is equal to the significance level. (aka *size*).

2. A *Type II Error* is when we fail to reject the Null even though it is false. The *power of a test* is the probability of making a Type II Error. The power of a test varies with the true value of the parameter(s).

The first strategy is to choose in advance a tolerance for Type I errors. For example, suppose we were willing to tolerate rejecting the Null 5% of the time, even when it was true. This level of tolerance is usually called "alpha", and we'd say $\alpha = 5\%$. Then, we would go out in the tails of the distribution to a value far enough that only 5% of the probability remained in values larger, and call this value the *critical value*. Then, we compare the sample value of the test statistic to the critical value: if it is bigger, we reject the hypothesis; if not, not.

The second strategy instead evaluates the probability that a test statistic as large or larger than the one observed would be drawn if the Null was true. This is called the *p-value* of the test. If this p-value seems very small, we reject the hypothesis; if not, not.

Because we are interested only in whether or not the equalities of the Null Hypothesis are true or not true (and not in whether they are untrue in particular directions), only the right-hand tail of the χ^2_J distribution matters. This is where big deviations of $\widehat{d}$ from its hypothesized value of 0, be they big negatives or big positives, will show up.

So, consider

$$T_W = 8$$

for a hypothesis with $J = 3$ restrictions. Suppose we use the first strategy, and pick an α of 5%. The 5% critical value for a χ^2_3 variable (which I just looked up in a table) is 7.8. Compare $T_W = 8$ to the critical value 7.8: it is bigger, so we reject the Null hypothesis.

Suppose we use the second strategy, and ask what is the probability that a χ^2_3 would be at least as big as 8? I just looked up that probability: it is 4.8%. I'd say that's low, and reject the Null.

4.3 But I don't know σ^2!

Well, indeed, one typically does not know σ^2; instead we have an unbiased estimator for it $s^2 = e'e/N-K$. How does that fit in? Since s^2 is asymptotically equal to σ^2, the Wald Statistic behaves asymptotically just the same if you use s^2 or σ^2:

$$\lim_{N\to\infty} \frac{1}{s^2}\widehat{d}' \left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} = \frac{1}{\sigma^2}\widehat{d}' \left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} \tilde{\chi}^2_J.$$

So, if N is "pretty big", then you can use s^2 in place of σ^2, and the test statistic will follow a χ^2_J distribution anyways. This is an asymptotic result, and it may be irksome if you think your sample is quite far from infinitely large.

When N is not big, then the Wald test using s^2 is still pretty close to a χ^2_J. Since

$$\frac{1}{\sigma^2}\widehat{d}' \left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} \tilde{\chi}^2_J,$$

we have that the Wald Test statistic is equal to the product of $\frac{\sigma^2}{s^2}$ and a χ^2_J.

$$\frac{1}{s^2}\widehat{d}' \left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} = \frac{\sigma^2}{s^2}\frac{1}{\sigma^2}\widehat{d}' \left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} = \frac{\sigma^2}{s^2}\chi^2_J$$

Unfortunately, we don't know the small-sample distribution of $\frac{\sigma^2}{s^2}$ without assuming something about the distribution of ε.

4.4 I Don't Like Asymptotics and I Do Like Normality

4.4.1 Single Linear Tests: the finite-sample t-test

Consider the Wald Statistic when we just have a single linear test, a linear model and normal ε:

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ &\quad \varepsilon \tilde{} N(0, \sigma^2), \\ H_0 &: R\beta + r = 0. \end{aligned}$$

The Wald Vector, which just has a single element in this case, is distributed normally under the Null hypothesis if ε is distributed normally:

$$T_{Wv} = \frac{1}{\sigma}\left(R(X'X)^{-1}R'\right)^{-1/2} \widehat{d} \tilde{} N(0,1).$$

If you use s in place of σ to create a Wald vector T^s_{Wv} (the superscript notes the fact that we're using s), you still have asymptotic normality by the argument above:

$$T^s_{Wv} = \frac{1}{s}\left(R(X'X)^{-1}R'\right)^{-1/2}\widehat{d}^{\sim}_{N->\infty}N(0,1).$$

4.4.2 The z-statistic and the t-statistic

In both of these cases (the Wald Vector with one restriction, using either σ or s), the test statistic follows a normal distribution as the sample size gets large. For that reason, the Wald test in this case is often called a "z-test" (because standard normal variables are often denoted "z"), and the Wald Vector (with one restriction) is often called a "z-statistic", denoted T_z:

$$T_z = T_{Wv} = \frac{1}{\sigma}\left(R(X'X)^{-1}R'\right)^{-1/2}\widehat{d}^{\sim}N(0,1),$$

or analogously for the asymptotic distribution using s instead of σ:

$$T^s_z = \frac{1}{s}\left(R(X'X)^{-1}R'\right)^{-1/2}\widehat{d}^{\sim}_{N->\infty}N(0,1).$$

4.4.3 The finite-sample t-statistic

In fact, we can say a bit more about the z-statistic T^s_z using s in place of σ. It is equal to $\frac{\sigma}{s}$ times something that is normally distributed:

$$T^s_z = T^s_{Wv} = \frac{\sigma}{s}T_{Wv} = \frac{\sigma}{s}T_z = \frac{\sigma}{s}N(0,1).$$

So, if we can figure out the distribution of $\frac{\sigma}{s}$, then we might be able to work this out.

Recall the definition of s:

$$s^2 = \frac{e'e}{N-K},$$

and consider the reciprocal of $\frac{\sigma}{s}$ (that is, $\sqrt{\frac{s^2}{\sigma^2}}$):

$$\begin{aligned}\sqrt{\frac{s^2}{\sigma^2}} &= \sqrt{\frac{1}{N-K}\frac{(e'e)}{\sigma^2}} \\ &= \sqrt{\frac{1}{N-K}\frac{(\varepsilon' M_X \varepsilon)}{\sigma^2}} \\ &= \sqrt{\frac{1}{N-K}(v' M_X v)},\end{aligned}$$

where $v = \varepsilon/\sigma$. Since $\varepsilon^{\sim}N(0,\sigma^2)$, we have that

$$v^{\sim}N(0,1).$$

Consider $v'M_Xv$:

$$\begin{aligned} v'M_Xv &= v'\left(I_N - P_X\right)v \\ &= v'v - v'P_Xv \\ &= \chi^2_N - v'P_Xv. \end{aligned}$$

The quadratic form $v'M_Xv$ is the sum of N squared normals minus a quadratic form equal to the amount of v we can predict from X. If the v were observed, we could predict K of them exactly with K columns of X, so P_Xv could equal exactly v for K elements of v, and 0 for the rest. Alternatively, it could be a linear combination of those. Thus, $v'P_Xv$ is the sum of K squared normals, a χ^2_K, so we have

$$v'M_Xv = \chi^2_N - \chi^2_K = \chi^2_{N-K}.$$

The term $v'M_Xv$ is equal to the sum of $N-K$ squared normals, so

$$\frac{s}{\sigma} = \sqrt{\frac{s^2}{\sigma^2}} = \sqrt{\frac{\chi^2_{N-K}}{N-K}},$$

the square-root of a chi-square divided by its own degrees of freedom. Returning to the z-statistic, we have

$$\begin{aligned} T^s_z &= \frac{\sigma}{s}T_z = \frac{\sigma}{s}N(0,1) \\ &= \frac{N(0,1)}{\sqrt{\frac{\chi^2_{N-K}}{N-K}}}. \end{aligned}$$

This would seem to be useless, unless someone had tabulated the distribution of a normal divided by a square root of a chi-square divided by its own degrees of freedom.

Luckily, someone did. A dude named Gosset figured it out, and called it the "Student's t" distribution, denoted t_{N-K}, where $N-K$ is the number of degrees of freedom in the denominator. Fisher popularised the name. Instead of calling the test as a "single linear Wald test vector using s", we call it a "t Test Statistic", and denote it as

$$\begin{aligned} t-test &= \frac{\sigma}{s}N(0,1) \\ &= \frac{N(0,1)}{\sqrt{\frac{\chi^2_{N-K}}{N-K}}} \\ &= t_{N-K}, \end{aligned}$$

where "t_{N-K}" means "t distribution with $N-K$ degrees of freedom" which means "a standard normal divided by the square root of a chi-square divided by its own degrees of freedom".

4.4.4 Multiple Linear Tests: the finite-sample F-test

Consider the Wald Statistic when we have a multiple linear test, a linear model and normally distributed ε:

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ &\quad \varepsilon \tilde{} N(0, \sigma^2), \\ H_0 &: R\beta + r = 0. \end{aligned}$$

The Wald Statistic is

$$T_W = T'_{Wv} T_{Wv} = \frac{1}{\sigma^2} \widehat{d}' \left(R(X'X)^{-1} R' \right)^{-1} \widehat{d} \tilde{} \chi^2_J.$$

If we substitute in s^2 for σ^2, to create T^s_W (the superscript notes the fact that we're using s), then we have

$$\begin{aligned} T^s_W &= \frac{1}{s^2} \widehat{d}' \left(R(X'X)^{-1} R' \right)^{-1} \widehat{d} = \frac{\sigma^2}{s^2} T_W \\ &= \frac{\sigma^2}{s^2} \chi^2_J. \end{aligned}$$

We can rewrite the leading term in the Wald Statistic as follows

$$T^s_W = \frac{\sigma^2}{s^2} \chi^2_J = \frac{1}{\frac{1}{N-K} \left(\frac{e'e}{\sigma^2} \right)} \chi^2_J = \frac{\chi^2_J}{\chi^2_{N-K}/N-K},$$

by the same reasoning as for the Single Linear t-test. But, it feels a little unbalanced, because we divide by the degrees of freedom in the denominator, but not in the numerator. Consider dividing the numerator by its degrees of freedom:

$$T^s_{Wald}/J = \frac{\chi^2_J/J}{\chi^2_{N-K}/N-K} \tilde{} F_{J,N-K}.$$

This ratio of chi-squareds divided by their own distribution is so commonly seen that we (actually, a guy named Snedecor) named it F, with degrees of freedom given by its numerator and denominator degrees of freedom. The F-test statistic is equal to the Wald Test statistic (using s^2 instead of σ^2) divided by the number of restrictions being tested, and the F-test statistic is distributed as an F distribution.

4.4.5 Normality and Finite-Sample Distributions

These results concerning the finite-sample distributions of test statistics all rest on the assumption of normality of the ε's. If you are not willing to specify the distribution of the ε, you can forget about finite-sample distributions. If you are willing to specify the distribution of the ε's, you may be able work out the finite-sample distribution of any test statistic, particularly those that are that linear in the ε's.

4.5 But ε isn't actually normal!

Well, indeed, one typically does not know ε is normally distributed, but one may believe that ε has finite variance. In this case, a central limit theorem can be invoked: T'_{Wv} (the Wald vector) is a finite-length (length J) vector function of the random variable ε, and so as the length of ε grows to infinity, the distribution of T'_{Wv} is approximately equal to that of a vector of normals.

Consider a multiple linear hypothesis with a linear model and possibly non-normal (but finite variance) ε:

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ E\left[X'\varepsilon\right] &= 0_K, \\ E\left[\varepsilon\varepsilon'\right] &= \sigma^2 I_N, \\ H_0 &: R\beta + r = 0. \end{aligned}$$

Here, ε may be nonnormal (for example uniform) as long as ε is finite-variance. Then, we have that the Wald vector (a finite-length vector function of the data) is asympotically approximately normal:

$$T_{Wv} = \frac{1}{\sigma}\left(R(X'X)^{-1}R'\right)^{-1/2} \widehat{d} \overset{approx}{\underset{N\to\infty}{\sim}} N(0_J, I_J).$$

Since this vector is asymptotically approximately a vector of standard normals, its inner product is asymptotically approximately a chi-square:

$$T_W = T'_{Wv}T_{Wv} = \frac{1}{\sigma^2}\widehat{d}'\left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} \overset{approx}{\underset{N\to\infty}{\sim}} \chi^2_J.$$

(The central limit theorem can be used even more weirdly: if J is really big, then T_W will approximate a normal distribution. That is, as chi-square distributions get more degrees of freedom, adding together more and more squared standard normals (which have finite variance), the chi-square distribution will start to look normal (big hump in the middle, thin tails).)

Even better, the second-order approximation is not affected by replacing σ^2 with s^2. This is because s^2 is asymptotically equal to σ^2. Thus, we have that

$$T^s_W = T^{s\prime}_{Wv}T^s_{Wv} = \frac{1}{s^2}\widehat{d}'\left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} \overset{approx}{\underset{N\to\infty}{\sim}} \chi^2_J.$$

The Wald Statistic approximately follows the chi-square distribution as the sample size gets really large, even if one uses s^2.

4.6 Nonlinear Wald Tests

Consider a model in which we do not assume normality and have a set of J nonlinear restrictions $c(\beta) = 0$ that we wish to test:

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ E\left[X'\varepsilon\right] &= 0_K, \\ E\left[\varepsilon\varepsilon'\right] &= \sigma^2 I_N, \\ H_0 &: c(\beta) = 0. \end{aligned}$$

The discrepancy vector $\widehat{d}$ gives the distance between the sample value of the hypothesis and its hypothesized value of 0:

$$\widehat{d} = c\left(\widehat{\beta}\right).$$

Since we have not assumed normality of the ε, all we have for the distribution of $\widehat{\beta}$ is a mean and variance:

$$\begin{aligned} E\left[\widehat{\beta}\right] &= \beta, \\ V\left[\widehat{\beta}\right] &= \sigma^2 (X'X)^{-1}. \end{aligned}$$

Application of the delta-method allows us to calcuate the mean and variance of the nonlinear functions $c\left(\widehat{\beta}\right)$:

$$\begin{aligned} E\left[c\left(\widehat{\beta}\right)\right] &= E\left[c\left(E\left[\widehat{\beta}\right]\right)\right] = E\left[c\left(\beta\right)\right] = 0, \\ V\left[c\left(\widehat{\beta}\right)\right] &= \left(\nabla_{\beta} c\left(\beta\right)\right)' V\left[\widehat{\beta}\right] \nabla_{\beta} c\left(\beta\right)' = \sigma^2 \nabla_{\beta'} c\left(\beta\right) \left(X'X\right)^{-1} \left(\nabla_{\beta'} c\left(\beta\right)\right)', \end{aligned}$$

where $\nabla_{\beta'} c\left(\beta\right)$ is the matrix of derivatives of the vector-function $c(\beta)$ with respect to the row-vector β' (each row of $\nabla_{\beta'} c\left(\beta\right)$ gives the derivatives of an element of $c(\beta)$ with respect to β).

Then, if $V\left[c\left(\widehat{\beta}\right)\right]$ is finite, we can use an approximate asympototic result by applying a central limit theorem:

$$c\left(\widehat{\beta}\right) \underset{N\to\infty}{\sim}^{approx} N\left(0_J, \sigma^2 \nabla_{\beta'} c\left(\beta\right) \left(X'X\right)^{-1} \left(\nabla_{\beta'} c\left(\beta\right)\right)'\right).$$

Since $\widehat{\beta}$ goes to β asymptotically, we can replace $\nabla_{\beta'} c\left(\beta\right)$ with $\nabla_{\beta'} c\left(\widehat{\beta}\right)$:

$$c\left(\widehat{\beta}\right) \underset{N\to\infty}{\sim}^{approx} N\left(0_J, \sigma^2 \nabla_{\beta'} c\left(\widehat{\beta}\right) \left(X'X\right)^{-1} \left(\nabla_{\beta'} c\left(\widehat{\beta}\right)\right)'\right).$$

Now, we use this information to create the Wald Statistic. Premultiplying the sample value of the hypothesis by the minus-one-half matrix of its variance gives the Wald Vector distributed as a vector of standard normals:

$$T_{Wv} = \frac{1}{\sigma}\left(\nabla_{\beta'} c\left(\widehat{\beta}\right) \left(X'X\right)^{-1} \left(\nabla_{\beta'} c\left(\widehat{\beta}\right)\right)'\right)^{-1/2} c\left(\widehat{\beta}\right) \underset{N\to\infty}{\sim}^{approx} N\left(0_J, I_J\right).$$

Finally, we take the inner product of this to create the Wald Statistic

$$T_W = \frac{1}{\sigma^2} c\left(\widehat{\beta}\right)' \left(\nabla_{\beta'} c\left(\widehat{\beta}\right) \left(X'X\right)^{-1} \left(\nabla_{\beta'} c\left(\widehat{\beta}\right)\right)'\right)^{-1} c\left(\widehat{\beta}\right) \underset{N\to\infty}{\sim}^{approx} \chi^2_J.$$

Since this is an approximate asymptotic result, it also works with s instead of σ:

$$T_W = \frac{1}{s^2} c\left(\widehat{\beta}\right)' \left(\nabla_{\beta'} c\left(\widehat{\beta}\right) \left(X'X\right)^{-1} \left(\nabla_{\beta'} c\left(\widehat{\beta}\right)\right)'\right)^{-1} c\left(\widehat{\beta}\right) \underset{N\to\infty}{\sim}^{approx} \chi^2_J.$$

4.6.1 How to Get Any Wald Test You Want

Consider a model in which we do not assume normality and have a 1 nonlinear restriction $\beta^2 = 0$ that we wish to test:

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ E[X'\varepsilon] &= 0_K, \\ E[\varepsilon\varepsilon'] &= \sigma^2 I_N, \\ H_0 &: \beta^2 = 0. \end{aligned}$$

Since $\nabla_{\beta'} c(\beta) = 2\beta$, so

$$\begin{aligned} T_W &= \frac{1}{s^2} c\left(\widehat{\beta}\right)' \left(\nabla_{\beta'} c\left(\widehat{\beta}\right) (X'X)^{-1} \left(\nabla_{\beta'} c\left(\widehat{\beta}\right)\right)'\right)^{-1} c\left(\widehat{\beta}\right) \\ &= \frac{1}{s^2}\widehat{\beta}^2 \left(2\widehat{\beta}(X'X)^{-1}2\widehat{\beta}\right)^{-1} \widehat{\beta}^2 \\ &= \frac{1}{4s^2}\widehat{\beta}^2 (X'X) \overset{approx}{\underset{N\to\infty}{\sim}} \chi^2_1. \end{aligned}$$

This nonlinear hypothesis has a Wald Test statistic whose distribution is known under the Null Hypothesis. That is nice.

However, there is a not-nice feature to this. The restriction $\beta^2 = 0$ is equivalent to the linear restriction $\beta = 0$, which has a Wald Test

$$\begin{aligned} T_W &= \frac{1}{s^2}\widehat{d}' \left(R(X'X)^{-1}R'\right)^{-1} \widehat{d} \\ &= \frac{1}{s^2}\widehat{\beta} \left((X'X)^{-1}\right)^{-1} \widehat{\beta} \\ &= \frac{1}{s^2}\widehat{\beta}^2 (X'X) \overset{approx}{\underset{N\to\infty}{\sim}} \chi^2_1, \end{aligned}$$

because $R = 1$ and $\widehat{d}' = \widehat{\beta}$. The two test statistics are distributed approximately asympototically χ^2_1, but one is a quarter as large as the other. If you want to not reject, use $H_0 : \beta^2 = 0$ rather than $H_0 : \beta = 0$. Hypotheses such as $\beta^3 = 0$ would shrink the test statistic even further.

4.7 Goodness of Fit

Since errors are random variables, sums of squared errors (SSR) are random variables. So, we use the fit of a regression (SSR) as a test statistic. Consider the model

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ &\quad \varepsilon \sim N(0_N, \sigma^2 I_N) \\ H_0 &: R\beta + r = 0, \end{aligned}$$

and now, instead of worrying about the sampling distribution of $\widehat{\beta}$, we try to figure out the sampling distribution of

$$SSR = \sum_{i=1}^{N} (e_i)^2$$

where

$$e_i = Y_i - X_i\widehat{\beta} = M_X Y = M_X \varepsilon.$$

Goodness of fit could be compared by comparing the SSR when we impose the Null compared to SSR when we don't impose the Null.

This is different from the spirit of a Wald Test, because to do a Wald Test, you don't have to estimate under the restriction that the Null is true. Rather, with a Wald Test, you estimate a general model and ask how large is the discrepancy from the Null. Wald Tests are based on the sampling distribution of the estimated parameters; Goodness of Fit tests are based on the sampling distribution of the SSRs.

The Goodness of Fit test has 2 steps:

1. First, you estimate under the Null, and call the sum of squared errors from this as SSR_R (the R is for "restricted").

2. Then, you estimate under the alternative, and call the sum of squared errors from this as SSR_U (the U is for "unrestricted").

Notice that under the Null, SSR_R and SSR_U are driven by the same value of β, because under the Null the restrictions are true. This means that we might consider using

$$SSR_R - SSR_U$$

as part of a test statistic because its expectation under the Null is driven solely by ε (where the true parameter vector is β), and not by differences in the underlying parameters. We know that it must be weakly positive because the unrestricted model contains the restricted model as a possibility. How is this thing distributed?

If we knew σ^2, we would have that

$$\begin{aligned}
\frac{SSR_U}{\sigma^2} &= \frac{e'e}{\sigma^2}, \\
\frac{e'e}{\sigma^2} &= \frac{\varepsilon' M_X \varepsilon}{\sigma^2} \\
&= \frac{\varepsilon'\varepsilon}{\sigma^2} - \frac{\varepsilon' P_X \varepsilon}{\sigma^2} \\
&= v'v - v' P_X v,
\end{aligned}$$

where $v \tilde{} N(0,1)$. Then, we have sums of squared standard normals, aka chi-squared random variables.

$$\begin{aligned} v'v - v'P_X v = & \quad \chi^2_N - \varepsilon' P_X \varepsilon \\ = & \quad \chi^2_N - \chi^2_K \\ = & \quad \chi^2_{N-K}. \end{aligned}$$

Here, $\varepsilon' P_X \varepsilon$ is distributed as a χ^2_K because $P_X \varepsilon$ can get exactly K perfect fits of ε, so it is the sum of squares of K of the ε's. So,

$$SSR_U/\sigma^2 \tilde{} \chi_{N-K}.$$

What about SSR_R? By the same reasoning,

$$SSR_R/\sigma^2 = v'v - v'\widetilde{P_X}v,$$

where $\widetilde{P_X}$ is the projection of X subject to the restrictions $H_0 : R\beta + r = 0$. This projection does not really have K columns in its X matrix because J linear restrictions are imposed on the parameters. Thus, $\widetilde{P_X}v$ could have $K - J$ perfect fits (J restrictions is like having J less parameters to freely choose). Consequently,

$$v'\widetilde{P_X}v \tilde{} \chi^2_{K-J},$$

and,

$$\begin{aligned} SSR_R/\sigma^2 = & \quad v'v - v'\widetilde{P_X}v, \\ = & \quad \chi^2_N - \chi^2_{K-J} \\ = & \quad \chi^2_{N-K+J} \end{aligned}$$

Now, we can construct a Goodness of Fit test statistic:

$$\frac{SSR_U - SSR_R}{\sigma^2} \tilde{} \chi^2_J.$$

This is all good if we know σ^2: $\frac{SSR_U - SSR_R}{\sigma^2}$ is distributed as a chi-square.

Even if we don't know σ^2, we can use s^2 in place of σ^2 :

$$s^2 = \frac{1}{N-k}\sum_{i=1}^{N}(e_i)^2 = SSR_U/(N-K).$$

This object converges in distribution to a spike on 1. Then, we can use the asymptotic distribution:

$$\frac{SSR_U - SSR_R}{s^2} \tilde{}_{N->\infty} \chi^2_J.$$

Above, we ignore the sampling variation in the denominator, treating it like a constant asymptotically. This is a weird thing to do: we have sums of

squared normals on top and on bottom, but we ignore the sampling variation of the bottom. Alternatively, we can model the sampling distribution of the denominator:

$$\begin{aligned}\frac{SSR_U - SSR_R}{s^2} &= \frac{\sigma^2}{s^2}\frac{SSR_U - SSR_R}{\sigma^2} \\ &= \frac{\sigma^2}{s^2}\chi^2_J \\ &= \frac{\chi^2_J}{s^2/\sigma^2},\end{aligned}$$

so that we see that the test statistic using s^2 equals the test statistic using σ^2 (which is a chi-square) divided by s^2/σ^2. We've seen s^2/σ^2 before: it is equal to a $\chi^2_{N-K}/N-K$, so

$$\frac{SSR_U - SSR_R}{s^2} \tilde{} \frac{\chi^2_J}{\chi^2_{N-K}/N-K}.$$

The numerator is a chi-square not divided by its degrees of freedom, so if we divide it by its degrees of freedom, we get a ratio of chi-squares divided by their degrees of freedom, also known as an F:

$$\frac{(SSR_R - SSR_U)/J}{s^2} = \frac{\chi^2_J/J}{\chi^2_{N-K}/N-K} \tilde{} F_{J,N-k}.$$

4.8 Inequality Tests: Single hypotheses

These strategies both depend on whether or not we care about deviations from the Null hypothesis in both directions, or just in one-direction. Consider a single linear hypothesis

$$H_0 : R\beta + r = 0,$$

where all interesting deviations are negative. For example, if we were testing whether or not an increase in price led to a decrease in demand, we might reasonably think that β, the coefficient on price, could be zero or negative, but not positive. Then, $R = [1\ 0...\ 0]$, $r = 0$ (if the price is the first regressor), and we have alternatives only in one direction

$$\begin{aligned}H_0 &: R\beta + r = 0, \\ H_A &: R\beta + r < 0.\end{aligned}$$