/ Courses ▾ Tutorials ▾

/ (/default.html) GLMs (/wws509) Multilevel (/pop510) Survival (/pop509)

Demography (/eco572) Stata (/stata) R (/R)

Germán Rodríguez

Generalized Linear Models

Lecture Notes (/wws509/notes) •••

2.3 Tests of Hypotheses

Consider testing hypotheses about the regression coefficients $\boldsymbol{\beta}$. Sometimes we will be interested in testing the significance of a single coefficient, say β_j, but on other occasions we will want to test the joint significance of several components of $\boldsymbol{\beta}$. In the next few sections we consider tests based on the sampling distribution of the maximum likelihood estimator and likelihood ratio tests.

2.3.1 Wald Tests

Consider first testing the significance of one particular coefficient, say

$$H_0 : \beta_j = 0.$$

The m.l.e. $\hat{\beta}_j$ has a distribution with mean 0 (under H_0) and variance given by the j-th diagonal element of the matrix in Equation 2.9. Thus, we can base our test on the ratio

$$t = \frac{\hat{\beta}_j}{\sqrt{\mathrm{var}(\hat{\beta}_j)}}. \qquad (2.10)$$

Note from Equation 2.9 that $\mathrm{var}(\hat{\beta}_j)$ depends on σ^2, which is usually unknown. In practice we replace σ^2 by the unbiased estimate based on the residual sum of squares.

Under the assumption of normality of the data, the ratio of the coefficient to its standard error has under H_0 a *Student's t* distribution with $n - p$ degrees of freedom when σ^2 is estimated, and a standard normal distribution if σ^2 is known. This result provides a basis for exact inference in samples of any size.

Under the weaker second-order assumptions concerning the means, variances and covariances of the observations, the ratio has approximately in large samples a standard normal distribution. This result provides a basis for approximate inference in large samples.

Many analysts treat the ratio as a Student's t statistic regardless of the sample size. If normality is suspect one should not conduct the test unless the sample is large, in which case it really makes no difference which distribution is used. If the sample size is moderate, using the t test provides a more conservative procedure. (The Student's t distribution converges to a standard normal as the degrees of freedom increases to ∞. For example the 95% two-tailed critical value is 2.09 for 20 d.f., and 1.98 for 100 d.f., compared to the normal critical value of 1.96.)

The t test can also be used to construct a confidence interval for a coefficient. Specifically, we can state with $100(1 - \alpha)$% confidence that β_j is between the bounds

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p} \sqrt{\mathrm{var}(\hat{\beta}_j)}, \tag{2.11}$$

where $t_{1-\alpha/2,n-p}$ is the two-sided critical value of Student's t distribution with $n-p$ d.f. for a test of size α.

The Wald test can also be used to test the joint significance of several coefficients. Let us partition the vector of coefficients into two components, say $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ with p_1 and p_2 elements, respectively, and consider the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = 0.$$

In this case the Wald statistic is given by the quadratic form

$$W = \hat{\boldsymbol{\beta}}_2' \, \mathrm{var}^{-1}(\hat{\boldsymbol{\beta}}_2) \, \hat{\boldsymbol{\beta}}_2,$$

where $\hat{\boldsymbol{\beta}}_2$ is the m.l.e. of $\boldsymbol{\beta}_2$ and $\mathrm{var}(\hat{\boldsymbol{\beta}}_2)$ is its variance-covariance matrix. Note that the variance depends on σ^2 which is usually unknown; in practice we substitute the estimate based on the residual sum of squares.

In the case of a single coefficient $p_2 = 1$ and this formula reduces to the square of the t statistic in Equation 2.10.

Asymptotic theory tells us that under H_0 the large-sample distribution of the m.l.e. is multivariate normal with mean vector **0** and variance-covariance matrix $\mathrm{var}(\boldsymbol{\beta}_2)$. Consequently, the large-sample distribution of the quadratic form W is chi-squared with p_2 degrees of freedom. This result holds whether σ^2 is known or estimated.

Under the assumption of normality we have a stronger result. The distribution of W is exactly chi-squared with p_2 degrees of freedom if σ^2 is known. In the more general case where σ^2 is estimated using a residual sum of squares based on $n-p$ d.f., the distribution of W/p_2 is an F with p_2 and $n-p$ d.f.

Note that as n approaches infinity for fixed p (so $n-p$ approaches infinity), the F distribution times p_2 approaches a chi-squared distribution with p_2 degrees of freedom. Thus, in large samples it makes no difference whether one treats W as chi-squared or W/p_2 as an F statistic. Many analysts treat W/p_2 as F for all sample sizes.

The situation is exactly analogous to the choice between the normal and Student's t distributions in the case of one variable. In fact, a chi-squared with one degree of freedom is the square of a standard normal, and an F with one and v degrees of freedom is the square of a Student's t with v degrees of freedom.

2.3.2 The Likelihood Ratio Test

Consider again testing the joint significance of several coefficients, say

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

as in the previous subsection. Note that we can partition the model matrix into two components $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ with p_1 and p_2 predictors, respectively. The hypothesis of interest states that the response does not depend on the last p_2 predictors.

We now build a likelihood ratio test for this hypothesis. The general theory directs us to (1) fit two nested models: a smaller model with the first p_1 predictors in $\boldsymbol{X}_1$, and a larger model with all p

predictors in $\boldsymbol{X}$; and (2) compare their maximized likelihoods (or log-likelihoods).

Suppose then that we fit the smaller model with the predictors in $\boldsymbol{X}_1$ only. We proceed by maximizing the log-likelihood of Equation 2.5 for a fixed value of σ^2. The maximized log-likelihood is

$$\max \log \mathrm{L}(\boldsymbol{\beta}_1) = c - \frac{1}{2}\mathrm{RSS}(\boldsymbol{X}_1)/\sigma^2,$$

where $c = -(n/2)\log(2\pi\sigma^2)$ is a constant depending on π and σ^2 but not on the parameters of interest. In a slight abuse of notation, we have written $\mathrm{RSS}(\boldsymbol{X}_1)$ for the residual sum of squares after fitting $\boldsymbol{X}_1$, which is of course a function of the estimate $\hat{\boldsymbol{\beta}}_1$.

Consider now fitting the larger model $X_1 + X_2$ with all predictors. The maximized log-likelihood for a fixed value of σ^2 is

$$\max \log \mathrm{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = c - \frac{1}{2}\mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2)/\sigma^2,$$

where $\mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2)$ is the residual sum of squares after fitting $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, itself a function of the estimate $\hat{\boldsymbol{\beta}}$.

To compare these log-likelihoods we calculate minus twice their difference. The constants cancel out and we obtain the likelihood ratio criterion

$$-2\log\lambda = \frac{\mathrm{RSS}(\boldsymbol{X}_1) - \mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2)}{\sigma^2}. \qquad (2.12)$$

There are two things to note about this criterion. First, we are directed to look at the reduction in the residual sum of squares when we add the predictors in $\boldsymbol{X}_2$. Basically, these variables are deemed to have a significant effect on the response if including them in the model results in a reduction in the residual sum of squares. Second, the reduction is compared to σ^2, the error variance, which provides a unit of comparison.

To determine if the reduction (in units of σ^2) exceeds what could be expected by chance alone, we compare the criterion to its sampling distribution. Large sample theory tells us that the distribution of the criterion converges to a chi-squared with p_2 d.f. The expected value of a chi-squared distribution with ν degrees of freedom is ν (and the variance is 2ν). Thus, chance alone would lead us to expect a reduction in the RSS of about one σ^2 for each variable added to the model. To conclude that the reduction exceeds what would be expected by chance alone, we usually require an improvement that exceeds the 95-th percentile of the reference distribution.

One slight difficulty with the development so far is that the criterion depends on σ^2, which is not known. In practice, we substitute an estimate of σ^2 based on the residual sum of squares of the *larger* model. Thus, we calculate the criterion in Equation 2.12 using

$$\hat{\sigma}^2 = \mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2)/(n-p).$$

The large-sample distribution of the criterion continues to be chi-squared with p_2 degrees of freedom, even if σ^2 has been estimated.

Under the assumption of normality, however, we have a stronger result. The likelihood ratio criterion $-2\log\lambda$ has an *exact* chi-squared distribution with p_2 d.f. if σ^2 is know. In the usual case where σ^2 is estimated, the criterion divided by p_2, namely

$$F = \frac{(\mathrm{RSS}(\boldsymbol{X}_1) - \mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2))/p_2}{\mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2)/(n-p)}, \tag{2.13}$$

has an exact F distribution with p_2 and $n-p$ d.f.

The numerator of F is the reduction in the residual sum of squares per degree of freedom spent. The denominator is the average residual sum of squares, a measure of noise in the model. Thus, an F-ratio of one would indicate that the variables in $\boldsymbol{X}_2$ are just adding noise. A ratio in excess of one would be indicative of signal. We usually reject H_0, and conclude that the variables in $\boldsymbol{X}_2$ have an effect on the response if the F criterion exceeds the 95-th percentage point of the F distribution with p_2 and $n-p$ degrees of freedom.

A Technical Note:\/ In this section we have built the likelihood ratio test for the linear parameters $\boldsymbol{\beta}$ by treating σ^2 as a nuisance parameter. In other words, we have maximized the log-likelihood with respect to $\boldsymbol{\beta}$ for fixed values of σ^2. You may feel reassured to know that if we had maximized the log-likelihood with respect to both $\boldsymbol{\beta}$ and σ^2 we would have ended up with an equivalent criterion based on a comparison of the *logarithms* of the residual sums of squares of the two models of interest. The approach adopted here leads more directly to the distributional results of interest and is typical of the treatment of scale parameters in generalized linear models.□

2.3.3 Student’s t, F and the Anova Table

You may be wondering at this point whether you should use the Wald test, based on the large-sample distribution of the m.l.e., or the likelihood ratio test, based on a comparison of maximized likelihoods (or log-likelihoods). The answer in general is that in large samples the choice does not matter because the two types of tests are asymptotically equivalent.

In linear models, however, we have a much stronger result: the two tests are *identical*. The proof is beyond the scope of these notes, but we will verify it in the context of specific applications. The result is unique to linear models. When we consider logistic or Poisson regression models later in the sequel we will find that the Wald and likelihood ratio tests differ.

At least for linear models, however, we can offer some simple practical advice:

- To test hypotheses about a single coefficient, use the t-test based on the estimator and its standard error, as given in Equation 2.10.
- To test hypotheses about several coefficients, or more generally to compare nested models, use the F-test based on a comparison of RSS's, as given in Equation 2.13.

The calculations leading to an F-test are often set out in an analysis of variance (anova) table, showing how the total sum of squares (the RSS of the null model) can be partitioned into a sum of squares associated with $\boldsymbol{X}_1$, a sum of squares *added by* $\boldsymbol{X}_2$, and a residual sum of squares. The table also shows the degrees of freedom associated with each sum of squares, and the mean square, or ratio of the sum of squares to its d.f.

Table 2.2 shows the usual format. We use ϕ to denote the null model. We also assume that one of the columns of $\boldsymbol{X}_1$ was the constant, so this block adds only $p_1 - 1$ variables to the null model.

Table 2.2. The Hierarchical Anova Table

Source of variation	Sum of squares	Degrees of freedom

$\boldsymbol{X}_1$	$\mathrm{RSS}(\phi) - \mathrm{RSS}(\boldsymbol{X}_1)$	$p_1 - 1$
$\boldsymbol{X}_2$ given $\boldsymbol{X}_1$	$\mathrm{RSS}(\boldsymbol{X}_1) - \mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2)$	p_2
Residual	$\mathrm{RSS}(\boldsymbol{X}_1 + \boldsymbol{X}_2)$	$n - p$
Total	$\mathrm{RSS}(\phi)$	$n - 1$

Sometimes the component associated with the constant is shown explicitly and the bottom line becomes the total (also called 'uncorrected') sum of squares: $\sum y_i^2$. More detailed analysis of variance tables may be obtained by introducing the predictors one at a time, while keeping track of the reduction in residual sum of squares at each step.

Rather than give specific formulas for these cases, we stress here that *all* anova tables can be obtained by calculating differences in RSS's and differences in the number of parameters between nested models. Many examples will be given in the applications that follow. A few descriptive measures of interest, such as simple, partial and multiple correlation coefficients, turn out to be simple functions of these sums of squares, and will be introduced in the context of the applications.

An important point to note before we leave the subject is that the order in which the variables are entered in the anova table (reflecting the order in which they are added to the model) is extremely important. In Table 2.2, we show the effect of adding the predictors in $\boldsymbol{X}_2$ to a model that already has $\boldsymbol{X}_1$. This *net* effect of X_2 after allowing for X_1 can be quite different from the *gross* effect of X_2 when considered by itself. The distinction is important and will be stressed in the context of the applications that follow.

Math rendered by (http://www.mathjax.org)