Hypothesis testing for goodness-of-fit

Stats 203, Winter 2024

Lecture 8 [Last update: February 9, 2024]

1 F-test

In last class we introduce the T-test, to test the null hypothesis that the jth predictor did not add to a model with all the other predictors included. Today we will talk about **the F-test**, which tests the null hypothesis that *none* of the predictors add to a model that has only an intercept term:

$$H_0 : \beta_1 = \ldots = \beta_p = 0. \tag{1}$$

The F-test is an example of **goodness-of-fit** test: it asks whether a very simple linear model that includes only an intercept term fits well enough to the data that the other predictors need not be included.

The basic idea behind the F-test is to compare the predictions made by the usual OLS estimates to the predictions made by the OLS estimates of a "null model", and see how much the predictions improve compared to what we would expect if H_0 were true.

The full model versus null model. To avoid confusion, we now call our usual linear model – which is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i,$$

with the usual assumptions of independent observations and mean-zero, variance σ^2 and Normal errors – the **full model**. The full-model OLS estimates are the usual

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y$$

the fitted values are $\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{H}Y$, and the mean-squared error is the mean of the squared residuals: $\mathrm{MSE}_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n}\|e\|^2$. In performing the F-test, it is often more convenient to skip dividing by n and use $\|e\|^2$, which is the sum of squared residual rather than the mean. This is sometimes called the **residual sum of squares**.

On the other hand, in the null model – meaning, the model if the null hypothesis is correct – we have

$$Y_i = \gamma_0 + \epsilon_i.$$

Estimating this (particularly boring) model by least squares, that is computing $\hat{\gamma}_0 = \operatorname{argmin} \mathrm{MSE}_n(\gamma)$, gives the least-squares estimate $\hat{\gamma}_0 = \bar{Y}$, fitted values $\bar{Y}\mathbf{1}$, and a sum of squared residuals

$$n\mathrm{MSE}_n(\hat{\gamma}_0) = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \|Y - \bar{Y}\mathbf{1}\|^2.$$

This quantity is sometimes called the **total sum of squares**.

Total sum of squares minus residual sum of squares. Remember that test statistics should measure the strength of the evidence against the null model. The motivation behind the F-statistic is that if (at least one of) the predictors truly belongs in the model, then the predictions $\hat{Y} = \mathbf{X}\hat{\beta}$ should be more accurate than the predictions made by the null model, $\bar{Y}\mathbf{1}$, in the sense of having smaller mean-squared error. Mathematically, we should expect

$$n\big(\mathrm{MSE}_n(\hat{\gamma}_0) - \mathrm{MSE}_n(\hat{\beta})\big) = \|Y - \bar{Y}\mathbf{1}\|^2 - \|Y - \hat{Y}\|^2 \tag{2}$$

to be large if the null is false. The F-statistic essentially takes the statistic in (2) and rescales it to have a simple distribution if the null hypothesis is correct. To properly understand this, we need to calculate the distribution of (2).

More geometry. Observe that the null model is “nested” inside the full model. That is, our usual linear model includes the null as a special case. This leads to some special geometric relations between the fitted values produced by the two models.

Specifically, notice that the fitted values in the null model can be written as

$$\bar{Y}\mathbf{1} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top Y := \mathbf{H}_0 Y,$$

where $\mathbf{H}_0$ projects onto the vector of all-ones. Recall that $\mathbf{H}\mathbf{1} = \mathbf{1}$ (Good bit of study prep: how did we show this?) and so $\mathbf{H}\mathbf{H}_0 = \mathbf{H}_0\mathbf{H} = \mathbf{H}_0$, and

$$(\mathbf{H} - \mathbf{H}_0)^2 = \mathbf{H} - \mathbf{H}_0.$$

In other words, $\mathbf{H} - \mathbf{H}_0$ is yet another projection matrix, which projects onto the subspace $\mathrm{col}(\mathbf{X}) \ominus \mathbf{1}$: the orthogonal complement of $\mathbf{1}$ in the column space of $\mathbf{X}$. You can confirm that $\mathrm{tr}\mathbf{H}_0 = 1$ and therefore $\mathrm{tr}(\mathbf{H} - \mathbf{H}_0) = p + 1 - 1 = p$, which is the dimension of $\mathrm{col}(\mathbf{X}) \ominus \mathbf{1}$.

To summarize: ordinary least squares in the null model is projecting onto a one-dimensional subspace of the columns of $\mathbf{X}$, the subspace spanned by the all-ones vector.

Sampling distribution of (2) under the null. Again we use the geometry of OLS. Recall that the residuals and fitted values are orthogonal: $e \perp \hat{Y}$. Recall also that the residuals have mean-zero: $e \perp \mathbf{1}$. (Good bit of study prep: how did we show these facts?) A little bit of algebra leads to the following Pythagorean relationship:

$$\begin{aligned}
\|Y - \bar{Y}\mathbf{1}\|^2 &= \|Y - \hat{Y} + \hat{Y} - \bar{Y}\mathbf{1}\|^2 \\
&= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\mathbf{1}\|^2 + 2(Y - \hat{Y})^\top(\hat{Y} - \bar{Y}\mathbf{1}) \\
&= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\mathbf{1}\|^2.
\end{aligned}$$

So (2) is simply

$$\|Y - \bar{Y}\mathbf{1}\|^2 - \|Y - \hat{Y}\|^2 = \|\hat{Y} - \bar{Y}\mathbf{1}\|^2.$$

That is, the difference between the total sum of squares and the sum of the squared residuals is the sum of the squared difference between the predictions made by the full model and those made the null model.

Now, rewriting $\hat{Y} - \bar{Y}1$ in terms of projection matrices, and *assuming the null model is correct*,

$$\begin{aligned}
\|\hat{Y} - \bar{Y}\mathbf{1}\|^2 &= \|(\mathbf{H} - \mathbf{H}_0)Y\|^2 \\
&= Y^\top(\mathbf{H} - \mathbf{H}_0)^2 Y \\
&= Y^\top(\mathbf{H} - \mathbf{H}_0) Y \\
&= (\gamma_0\mathbf{1} + \epsilon)^\top(\mathbf{H} - \mathbf{H}_0)(\gamma_0\mathbf{1} + \epsilon) \\
&= \epsilon^\top(\mathbf{H} - \mathbf{H}_0)\epsilon.
\end{aligned}$$

Remember that we showed previously that the sum of squared residuals $\|e\|^2 = \epsilon^\top(\mathbf{I} - \mathbf{H})\epsilon$, that $\mathbb{E}[\epsilon^\top(\mathbf{I} - \mathbf{H})\epsilon] = \sigma^2\mathrm{tr}(\mathbf{I} - \mathbf{H}) = \sigma^2(n - p - 1)$ and that in fact $\|e\|^2 = \epsilon^\top(\mathbf{I} - \mathbf{H})\epsilon \sim \sigma^2\chi^2_{n-p-1}$. It is not too hard to show (so do it!) the analogous property

$$\mathbb{E}\Big[\epsilon^\top(\mathbf{H} - \mathbf{H}_0)\epsilon\Big] = \sigma^2\mathrm{tr}(\mathbf{H} - \mathbf{H}_0) = \sigma^2 p.$$

In fact it turns out that $\epsilon^\top(\mathbf{H}-\mathbf{H}_0)\epsilon \sim \sigma^2\chi^2_p$. So we have shown that if the null model is correct then

$$\|\hat{Y} - \bar{Y}\mathbf{1}\|^2 \sim \sigma^2\chi^2_p \tag{3}$$

This could be used to construct a p-value and calibrate a hypothesis test for H_0 if the error variance σ^2 were known.

F-test. Since we typically don't know the error variance we have to somehow get rid of the dependence on σ^2 in (3). We do this in exactly the same way as with the T-statistic: divide by an unbiased estimate of the error variance $\hat{\sigma}^2$ leading to

$$F = \frac{\|\hat{Y} - \bar{Y}\mathbf{1}\|^2/p}{\hat{\sigma}^2},$$

In the numerator we divide by p so that $F \leq 1$ means little evidence against the null, while F much greater than 1 indicates strong evidence against the null.

To summarize, we have shown that

$$\hat{\sigma}^2 \sim \frac{\sigma^2\chi^2_{n-p-1}}{n-p-1}, \tag{4}$$

$$\|\hat{Y} - \bar{Y}\mathbf{1}\|^2/p \sim \frac{\sigma^2\chi^2_p}{p} \quad \text{if } H_0 \text{ is true.} \tag{5}$$

Under the null hypothesis, the F-statistic is thus the ratio of two independent chi-squared distributed random variables, normalized by their degrees of freedom. This distribution is called the **F-distribution**, denoted $F \sim F_{p,n-p-1}$. The F is for Fisher.

Warnings. The F-test is geometrically impressive but is prone to abuse. Some of these abuses are related to those I described for the T-tests in the last lecture notes. In the case of F, however, it is particularly tempting to treat the test as determining whether the linear model is correct. This is categorically wrong. The F-test *assumes the linear model is correct*, and tests whether the "nested" null model suffices.

In practice, if any of the assumptions of the linear model – including Normal errors – are wrong, what is likely to happen is that the F p-value will be extraordinarily small. This makes perfect literal sense since the null model is certainly not true. But it provides no evidence *in favor* of the full model, or any other linear model.

The only way to check whether the linear model is correct are diagnostics.

2 R^2

The F-statistic is a measure of **goodness-of-fit**: By far the most common measure of goodness-of-fit is R^2, which can be viewed as a quantity that arises out of the calculus of F-statistics. R^2 is defined as

$$R^2 := 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

R^2 is quite common in applied work, but it has a number of issues.

- To repeat: goodness-of-fit means something different than the model being correct. When the error variance σ^2 is large, the R^2 will be small even if the fitted model is exactly correct. On other hand, there are cases where the true model is *not* linear but R^2 is (arbitrarily) close to 1.
- Adding more predictors will almost always increase the R^2 (and it will *never* decrease the R^2) whether or not they have any true relationship with the response.
- R^2 says nothing about whether the model will predict well at a new value of the predictors.
- R^2 only measures the strength of the linear relationship between x and y. It is possible for x to perfectly predict y – meaning $y = f(x)$ is a deterministic function of x – but for the R^2 to equal 0.

Relationship between R^2 and F. A little bit of algebra shows the relation between R^2 and F:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p} \Longleftrightarrow R^2 = \frac{1}{\frac{n-p-1}{(p)F} + 1}. \tag{6}$$

Remember that values of $F \approx 1$ indicate weak evidence against the null. Suppose $F = 1$, and $n - p$ were small relative to p, in the extreme case $n - p - 1 = 1$. Then $R^2 = \frac{1}{p^{-1}+1}$. Thus, even if there is limited evidence against the null, the R^2 can be made arbitrary close to 1 if the number of predictors p is large. This provides a dramatic illustration of the second bullet point above.

Adjusted R^2. Another way to see the issue with R^2 is to notice that

$$R^2 = 1 - \frac{\hat{\sigma}^2_{\text{PLUG}}}{s^2_Y},$$

where we recall that $\hat{\sigma}^2_{\text{PLUG}} = \frac{1}{n}\|e\|^2$ was our plug-in estimate of error variance, and s^2_Y is the empirical variance of Y. So R^2 is defined with respect to *b*iased estimates of variance, with the bias of $\hat{\sigma}^2_{\text{PLUG}}$ increasing with p.

This motivates the definition of **adjusted** R^2,

$$R^2_p = 1 - \frac{\frac{1}{n-p-1}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Adjusted R^2 is better than R^2 in that it does not blindly increase as more predictors are included in the model. But later, we will see still better ways of measuring the relative goodness-of-fit of two models.