Fall 2013 Statistics 151 (Linear Models) : Lecture Nine

Aditya Guntuboyina

26 September 2013

1 Hypothesis Tests to Compare Models

Let M denote the full regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$$

which has p explanatory variables.

Let m denote a sub-model of M that is obtained by a linear constraint on the parameter $\beta = (\beta_0, \ldots, \beta_p)$ of M. Examples:

1. For the constraint $\beta_1 = 0$, the model m becomes: $y_i = \beta_0 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$.
2. For $\beta_1 = \beta_2 = 0$, the model m becomes $y_i = \beta_0 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + e_i$.
3. For $\beta_1 = \cdots = \beta_p = 0$, the model m becomes $y_i = \beta_0 + e_i$.
4. For $\beta_1 = \beta_2$, the model m becomes $y_i = \beta_0 + \beta_1(x_{i1} + x_{i2}) + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + e_i$.
5. For $\beta_1 = 3$, the model m becomes $y_i = \beta_0 + 3x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$.

How do we test the hypothesis $H_0 : m$ against $H_1 : M$? Let q be the number of explanatory variables in m. This test can be carried out by noting that if $RSS(m) - RSS(M)$ is large, then m is not a good model for the data and we therefore reject H_0. On the other hand, if $RSS(m) - RSS(M)$ is small, we do not reject H_0.

The test is therefore based on $RSS(m) - RSS(M)$. What is the distribution of this quantity under the null bypothesis? We will show that

$$\frac{RSS(m) - RSS(M)}{\sigma^2} \sim \chi^2_{p-q}$$

where q is the number of explanatory variables in m and p is the number of explanatory variables in M. Let us now prove this fact. Let the hat matrices in the two models be denoted by $H(m)$ and $H(M)$ respectively. Write

$$RSS(m) = Y^T(I - H(m))Y \quad \text{and} \quad RSS(M) = Y^T(I - H(M))Y$$

so that

$$RSS(m) - RSS(M) = Y^T(H(M) - H(m))Y.$$

We need the null distribution of $RSS(m) - RSS(M)$. So we shall assume that $Y = X(m)\beta(m) + e$ (where $X(m)$ is the X-matrix in the model m). It is important to realize that $H(m)X(m) = X(m)$ and also $H(M)X(m) = X(m)$. So

$$RSS(m) - RSS(M) = e^T(H(M) - H(m))e.$$

$H(M) - H(m)$ is a symmetric $n \times n$ matrix of rank $p - q$. It is also idempotent because

$$(H(M) - H(m))^2 = H(M) + H(m) - 2H(M)H(m). \tag{1}$$

Now $H(M)H(m) = H(m)$. To see this, it is enough to show that $H(M)H(m)v = H(m)v$ for every vector v. Now recall that $H(m)v$ denotes the projection of v onto the column space of $X(m)$. And $H(M)H(m)v$ projects $H(m)v$ projects onto the column space of $X(M)$ (which equals the original X matrix). But because the column space of $X(m)$ is contained in the column space of $X(M)$, it follows that $H(m)v$ is already contained in the column space of $X(M)$. Thus its projection onto the column space of $X(M)$ equals itself. So $H(M)H(m)v = H(m)v$.

Because $H(M)H(m) = H(m)$, it follows from (1) that

$$(H(M) - H(m))^2 = H(M) - H(m).$$

Therefore, under the null hypothesis

$$\frac{RSS(m) - RSS(M)}{\sigma^2} = \left(\frac{e}{\sigma}\right)^T (H(M) - H(m))\frac{e}{\sigma} \sim \chi^2_{p-q}.$$

Since we do not know σ^2, we estimate it by

$$\hat{\sigma}^2 = \frac{RSS(M)}{n - p - 1},$$

to obtain the test statistic:

$$\frac{RSS(m) - RSS(M)}{RSS(M)/(n - p - 1)}$$

The numerator and the denominator are independent because $RSS(m) - RSS(M) = Y^T(H(M) - H(m))Y$ and $RSS(M) = Y^T(I - H(M))Y$ and the product of the matrices

$$(H(M) - H(m))\,(I - H(M)) = H(M) - H(M)^2 - H(m) + H(m)H(M) = H(M) - H(M) - H(m) + H(m) = 0.$$

Thus under the null hypothesis

$$\frac{(RSS(m) - RSS(M))/(p - q)}{RSS(M)/(n - p - 1)} \sim F_{p-q,n-p-1}.$$

p-value can therefore be got by

$$\mathbb{P}\left(F_{p-q,n-p-1} > \frac{(RSS(m) - RSS(M))/(p - q)}{RSS(M)/(n - p - 1)}\right).$$

If the null hypothesis can be written in terms of a single linear function of β, such as $H_0 : \beta_1 + 5\beta_3 = 5$. Then it can also be tested via the t-test; using the statistic:

$$\frac{\hat{\beta}_1 + 5\hat{\beta}_3 - 5}{s.e(\hat{\beta}_1 + 5\hat{\beta}_3)}$$

which has the t-distribution with $n - p - 1$ degrees of freedom under H_0. This test and the corresponding F-test will have the same p-value.

1.1 Testing for all explanatory variables

How do we test $H_0 : \beta_1 = \cdots = \beta_p = 0$ against its complement? Just take m to be the model $y_i = \beta_0 + e_i$. In this case, $RSS(m) = TSS$ and $q = 0$ and $RSS(M) = RSS$. Thus the p-value is

$$\mathbb{P}\left\{F_{p,n-p-1} > \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}\right\}.$$