Biostatistics 140.653

Distributional Results
Justification for Sequential General F-test

I. Preliminaries

Recall that the multiple linear regression model specifies $E[Y|X] = X\beta$. Said another way, MLR specifies that the vector of outcome means must belong to the following set of possibilities:

$$\{\mu \text{ which satisfy } \mu = \beta_0 X_0 + \beta_1 X_1 + \ldots + \beta_p X_p \text{ for some } (\beta_0, \beta_1, \ldots, \beta_p)\},$$

where X_0 is an nx1 vector of ones (intercept) and X_j is the column of X which contains the values of the jth covariate. Mathematically, such a set is a "Euclidean linear vector space." Let's call it the "model space" in our special case.

We've seen that models can be built up sequentially, beginning with an intercept only, then an intercept and X_1, and so on. In what follows, "Mj" will stand for the space of models including only the first j covariates, with "M0" standing for the intercept only model.

Linear algebra fact: Any model space can be generated from an "orthonormal basis"—e.g., if there are k covariates in the model, the space can be written

$$\{\mu \text{ which satisfy } \mu = a_0 e_0 + \ldots + a_k e_k \text{ for some } (a_0, a_1, \ldots, a_k)\},$$

where $e_j^T e_{j'} = 1$ if $j=j'$ and $= 0$ otherwise.

Linear algebra fact ("Gram-Schmidt orthogonalization"): An orthonormal basis $(e_0, e_1, \ldots, e_p)$ can be built up for the full model (Mp) so that $(e_0, \ldots, e_k)$ is an orthonormal basis for Mk, k=0,...,p-1. Moreover, the data vector Y can be rewritten as $c_0e_0+c_1e_1+ \ldots +c_pe_p + \ldots +c_{n-1}e_{n-1}$, where $(e_0, e_1, \ldots, e_p)$ is the orthonormal basis for the model and $(e_0, e_1, \ldots, e_p, \ldots, e_{n-1})$ is also an orthonormal basis (for n-dimensional space).

Finally, the least squares prediction for Y under model Mj is $c_0e_0+c_1e_1+ \ldots +c_je_j$. This is because the least squares procedure minimizes the distance between Y and a given model--e.g., projects Y into Mj. We can verify that the claim is the projection by noting that (1) the prediction is a member of Mj and (2) the prediction is orthogonal to Y- $c_0e_0+c_1e_1+ \ldots +c_je_j$, which is the perpendicularity property.

II. Distributional results

A. Under MLR assumptions A1-A4, $(c_0, c_1, \ldots, c_p, \ldots, c_{n-1})$ are independent normal random variables that have variance = σ^2 and with $E[c_{j+1}] = \ldots = E[c_{n-1}] = 0$ under H_0: Mj.

Why: Y can be written as Tc, where T is a matrix whose columns are $(e_0, e_1, \ldots, e_p, \ldots e_{n-1})$. Because of the orthonormal properties, T has the property that $T^TT = I$. So, $T^TY = c$. Thus. $Var(c) = Var(T^TY) = T^TVar(Y)T = T^T\sigma^2IT = \sigma^2I$. Under Mj, $E[Y] = a_0e_0 + \ldots + a_je_j$ for some Mj; thus, $a_h = E[c_h]$ for h=0,...,j, and $E[c_h] = 0$ otherwise (h=j+1....,n-1).

B. Under MLR assumptions A1-A4, the standardized residual sum of squares (RSS/σ^2) is chi-square distributed with n-p-1 degrees of freedom.

Why: The residual sum of squares equals $(Y-\hat{Y})^T(Y-\hat{Y})$

$$= (c_{p+1}e_{p+1} + \ldots + c_{n-1}e_{n-1})^T (c_{p+1}e_{p+1} + \ldots + c_{n-1}e_{n-1})$$

$$= \sum_{j=p+1}^{n-1} c_j^2 .$$

Under A1-A4, the terms of this sum are squared independent normal with mean 0 and variance σ^2. Dividing through by σ^2, we get a sum of squared indepedent standard normal (0,1) random variables. By definition, the overall distribution is chi-squared with n-p-1 degrees of freedom (df equal to the number of terms in the sum).

c. GENERAL F-test with respect to a subset of the full model: Suppose that our hypothesis is $H_0 : \beta_{k+1} = \ldots = \beta_p = 0$ —e.g., that the last p-k covariates in the full model are not associated with Y after controlling for the first k covariates.

Claim: Under H_0, $[(RSS_k-RSS_p)/(p-k)]/[RSS_p/(n-p-1)]$ is distributed as an F random variable with p-k and n-p-1 degrees of freedom.

Why: Under (b), we already demonstrated that the denominator is distributed as

$\sigma^2\chi^2_{n-p-1}/(n-p-1)$. H_0 is just another way of saying we're assuming model Mk. Using the same argument as under (b), the residual sum of squares for that model

$$= \sum_{j=k+1}^{n-1} c_j^2 .$$

Therefore, $RSS_k-RSS_p = \sum_{j=k+1}^{p} c_j^2$. Under the null hypothesis, the terms of this sum are squared independent normal with mean 0 and variance σ^2. Dividing through by σ^2, we get a sum of squared indepedent standard normal (0,1) random variables. By definition, the overall distribution is chi-squared with p-k degrees of freedom (df equal to the number of terms in the sum). Also, all the terms in this sum are independent to the terms in the "F-statistic" denominator. Thus, the statistic has the same distribution as the scaled ratio of two independent chi-squared random variables -- one with p-k degrees in the numerator, one with n-p-1 degrees of freedom in the denominator, where the ratio is divided by (p-k)/(n-p-1). By definition, the overall distribution is F with p-k and n-p-1 degrees of freedom.

d. Sequential F-test: For concreteness, suppose we wanted to test H_0: $\beta_1=...=\beta_k = 0$, with respect to the model $Y=\beta_0+\beta_1X_1 + ... + \beta_kX_k + \epsilon$. The claim was that one gets a valid test as "F"=$[(RSS_0-RSS_k)/k]/[RSS_p/(n-p-1)]$. Here's the argument for why:

i) As before, the denominator is distributed as $\sigma^2\chi^2_{n-p-1}/(n-p-1)$. That only depends on A1-A4, where A1 is based on the full model.

ii) Using arguments as above, $RSS_0-RSS_k = \sum_{j=1}^{k} c_j^2$. The thing to keep in mind is that all distributional statements are conditional on all the covariates; this is what causes intuitional discomfort about the claimed test. Regardless of H_0, then, the c_j's are independent, independent of the denominator, normal, and have variance σ^2. Therefore, it's only the means that are of concern. To figure out the means, consider how we built up the orthonormal basis: Now: the way we built up our is such that H_0 is equivalent to assuming that the mean response is in that part of n-dimensional space which is orthogonal to the subspace whose basis is $\{e_1,....e_k\}$ -- e.g.,

$$E[Y|X_1,...,X_p] = a_0e_0 + a_{k+1}e_{k+1} + ... + a_pe_p \text{ for some choice of a's.}$$

Thus, each of the c_j's, j=1,...,k, must have mean 0—otherwise, the above mean would depend on $\{e_1,...,e_k\}$ in some way. Thus, RSS_0-RSS_k is distributed as $\sigma^2\chi^2_k/k$, and the proposed test statistic is $F_{k,n-p-1}$ as claimed.