LAB 5

The purpose of this lab is to investigate the F-test for comparing nested models.

Consider two models: Y is an $n \times 1$ column vector, and ε is an $n \times 1$ column vector if i.i.d. $N(0, \sigma^2)$ random variables.

$$\begin{aligned} Y &= \mu_1 + \varepsilon & \mu_1 &\in V_0 \\ Y &= \mu_2 + \varepsilon & \mu_2 &\in V_1 \end{aligned}$$

where $V_0 \subset V_1$ are subspaces of $\mathbb{R}^n$.

For example, V_0 is the space on the first r columns of $n \times p$ matrix X, and V_1 is the space on all p columns. In the that case, in model 1, if X_i denotes the i-th column of D,

$$\mu_1 = \beta_1 X_1 + \cdots + \beta_r X_r,$$

while in model 2,

$$\mu_2 = \beta_1 X_1 + \cdots + \beta_r X_r + \beta_{r+1} X_{r+1} + \beta_p X_p$$

Note that if P_1 is the projection onto V_1 and P_2 is the projection onto V_2, then

$$I = P_1 + (P_2 - P_1) + (I - P_1)$$

Supposing the dimensions of V_1 and V_2 are r and p, the ranks of these matrices are

$$n = r + (p - r) + (n - p)$$

Note also that

$$(P_2 - P_1)(P_2 - P_1) = P_2 - P_2 P_1 - P_1 P_2 + P_1 = P_2 - P_1,$$

So that $P_2 - P_1$ is a projection matrix, namely onto the orthogonal complement of V_1 in V_2. Thus

$$I - P_1 = (P_2 - P_1) + (I - P_2) \tag{1}$$

Setting $Z = Y - EY$, we have by the chi-squared decomposition theorem that under

$$H_0 : \text{model 1 is correct: } E(Y) \in V_0,$$

it follows that

$$Y^T (I - P_1) Y - Y^T (I - P_1) Y = Y^T (P_2 - P_1) Y$$

is chi-squared with $p - r$ degrees of freedom, independent of $Y^T (I - P_2) Y$, which is chi-squared with r degrees of freedom. Note the first is the difference in residual sum of squares between model 1 and model 2, and the second is the residual sum of squares when model 2 is fit.

Thus

$$F = \frac{(RSS_1 - RSS_2)/(\Delta df)}{RSS_2/(df_2))}$$

is an F-statistic.

Consider the data set `smsa.txt`. THe response Y is mortality rate, and the individual data points correspond to city. The X_i's are variables corresponding to meteorological data and demographic data.

It appears the Education and Rain are important for predicting mortality. COmpare the model that includes just Education and Rain with the model that includes Education, Rain, NOx, PopDensity, income, RelHum, and JanTemp. Use a F-test to test the hypothesis that in fact the other variables are not in the model (their coefficients are all zero.)

You should do this first by fitting both models, and looking at the residual sum of squares in both to calculate the F-statistic.

The use the `anova` function to carry out this comparison, and check the results agree. The function `anova`, it takes as two (or more) arguments, fitted linear models which are nested.

Now compare three nested models. So model 1 will have p_1 variables included, model 2 will have p_2 variables included, and model 3 will have p_3 variables, where $p_1 < p_2 < p_3$.

If RSS_i denotes the residual sum of squares of the i-th model, and d_i it's degrees of freedom, the statistics

$$\frac{(RSS_i - RSS_j)/(d_i - d_j)}{RSS_3/d_3}$$

has a F distribution under the hypothesis that model i is the correct model ($i < j$).

Fit three nested models and test hypotheses comparing the appropriateness of the larger models. Compare with the results of `anova` applied to the three models.

What is the appropriate analog to (1) when there are three models?