STAT 350: Lecture 11

Reading: Chapter 7 Sections 1-3, 7 and 8.

F tests and the Extra Sum of Squares

Example:

$$\begin{aligned} Y &= \text{plaster hardness} \\ s &= \text{sand content} \\ f &= \text{fibre content} \end{aligned}$$

Model:

$$Y_i = \beta_0 + \beta_1 s_i \beta_2 s_i^2 + \beta_3 f_i + \beta_4 f_i^2 + \beta_5 s_i f_i + \epsilon_i$$

In matrix form:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & s_1 & s_1^2 & f_1 & f_1^2 & s_1 f_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_n & s_n^2 & f_n & f_n^2 & s_n f_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_5 \end{bmatrix}$$

Questions:

- Do we need the $S * F$ term?
- Do we need the F or F^2 terms?
- Do we need the S or S^2 terms?

To answer these questions we test

- $H_o : \beta_5 = 0$
- $H_o : \beta_3 = \beta_4 = 0$ (or perhaps $H_o : \beta_3 = \beta_4 = \beta_5 = 0$)
- $H_o : \beta_1 = \beta_2 = 0$

Technique: we fit a sequence of models:

(a) Original "full" model.

(b) The model with **no interactions**:

$$\mu_i = \beta_0 + \beta_1 s_i \beta_2 s_i^2 + \beta_3 f_i + \beta_4 f_i^2$$

(c) The Sand only model:

$$\mu_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2$$

(d) The Fibre only model:

$$\mu_i = \beta_0 + \beta_3 f_i + \beta_4 f_i^2$$

(e) The "Empty" model:

$$\mu_i = \beta_0$$

Each model corresponds to a design matrix which is a submatrix of the full design matrix:

$$Y = [\mathbf{1}|X_1|X_2|X_3] \begin{bmatrix} \beta_0 \\ \hline \beta_1 \\ \beta_2 \\ \hline \beta_3 \\ \beta_4 \\ \hline \beta_5 \end{bmatrix} + \epsilon$$

The design matrices for the models a, b, c, d and e are given by

$$\begin{aligned} X_1 &= X \\ X_b &= [\mathbf{1}|X_1|X_2] \\ X_c &= [\mathbf{1}|X_1] \\ X_d &= [\mathbf{1}|X_2] \\ X_e &= [\mathbf{1}] \end{aligned}$$

Note that $\mathbf{1}$ is a column vector of n 1s.

F tests

We can compare two models easily if one is a special case of the other, such as for example, when the design matrix of the first model is a submatrix of the second obtained by selecting subcolumns.

For example model (b) is a special case of (a), model (c) is a special case of (a) or (b) but models (c) and (d) are not comparable.

Comparing two models: General Theory

We consider the case of a design matrix X partitioned into two pieces X_1 and X_2.

$$X = [X_1|X_2]$$

The Full Model is

$$\begin{aligned} Y &= X\beta + \epsilon \\ &= [X_1|X_2]\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon \\ &= X_1\beta_1 + X_2\beta_2 + \epsilon \end{aligned}$$

The Reduced model is

$$Y = X_1\beta_1 + \epsilon$$

The difference between these two models is the term $X_2\beta_2$ which is 0 is the null hypothesis $H_o : \beta_2 = 0$ is true.

Dimensions:

β has p parameters.

β_i has p_i parameters with $p_1 + p_2 = p$.

To test $H_o : \beta_2 = 0$ we fit both the full and the reduced models and get

$$\begin{aligned} Y &= \hat{\mu}_F + \hat{\epsilon}_F \\ &= \hat{\mu}_r 5 + \hat{\epsilon}_R \end{aligned}$$

where the subscript F refers to the full model and R to the reduced model. This leads to the decomposition of the data vector Y into the sum of three perpendicular vectors:

$$Y = \hat{\mu}_R + (\hat{\mu}_F - \hat{\mu}_R) + \hat{\epsilon}_F$$

I showed

$$\begin{aligned} \hat{\mu}_R &\perp \hat{\mu}_F - \hat{\mu}_R \\ \hat{\mu}_R &\perp \hat{\epsilon}_F \\ \hat{\mu}_F - \hat{\mu}_R &\perp \hat{\epsilon}_F \end{aligned}$$

which leads to the following ANOVA table:

Source	Sum of Squares	Degrees of Freedom
X_2	$\|\|\hat{\mu}_R\|\|^2$	p_1
$X_2\|X_1$	$\|\|\hat{\mu}_F - \hat{\mu}_R\|\|^2$	p_2
Error	$\|\|\hat{\epsilon}\|\|^2$	$n-p$
Total (Unadjusted)	$\|\|Y\|\|^2$	n

In this table the notation $X_2|X_1$ means X_2 **adjusted for** X_1 or X_2 after fitting X_1.
This table can now be used to test $H_o : \beta_2 = 0$ by computing

$$F = \frac{\text{MS}(X_2|X_1)}{\text{MSE}} = \frac{||\hat{\mu}_F - \hat{\mu}_R||^2/p_2}{||\hat{\epsilon}||^2/(n-p)}$$

and getting a P value from the $F_{p_2,n-p}$ distribution.

Another Formula for this F statistic

recall that

$$||\hat{\mu}_R||^2 + ||\hat{\epsilon}_R||^2 = ||Y||^2$$

and

$$||\hat{\mu}_R||^2 + ||\hat{\mu}_F - \hat{\mu}_R||^2 + ||\hat{\epsilon}_F||^2 = ||Y||^2$$

so that

$$||\hat{\epsilon}_R||^2 = ||\hat{\epsilon}_F||^2 + ||\hat{\mu}_F - \hat{\mu}_R||^2$$

This makes

$$\begin{aligned} F &= \frac{(\text{ESS}_R - \text{ESS}_F)/p_2}{\text{ESS}_F/(n-p)} \\ &= \frac{\text{ExtraSS}/p_2}{\text{ESS}_F/(n-p)} \end{aligned}$$

Remarks:

1. If the errors are normal then

$$\frac{\text{ESS}_F}{\sigma^2} \sim \chi^2_{n-p}$$

2. If the errors are normal **and** $H_o : \beta_2 = 0$ is true then

$$\frac{\text{ExtraSS}}{\sigma^2} \sim \chi^2_{p_2}$$

3. ESS_F is independent of the Extra SS.

4. SO:

$$\frac{\text{ExtraSS}/(\sigma^2 p_2)}{\text{ESS}_F/(\sigma^2(n-p))} = F \sim F_{p_2,n-p}$$

Example of the above: Multiple Regression

In the data set below the hardness of plaster is measured for each of 9 combinations of sand content and fibre content. Sand content was set at one of 3 levels as was fibre content and all possible combinations tried on two batches of plaster.

Here is an excerpt of the data:

```
Sand  Fibre  Hardness  Strength
   0     0        61        34
   0     0        63        16
  15     0        67        36
  15     0        69        19
  30     0        65        28
     ...
```

The complete data set is here.

I fit submodels of the following "Full" model:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 S_i^2 + \beta_3 F_i + \beta_4 F_i^2 + \beta_5 S_i F_I + \epsilon_i$$

I adopt the idea that the interaction term is probably negligible unless each of S and F have some effect and that quadratic terms will probably not be present unless linear terms are present. This limits the set of potential reasonable models. I fit each of them and report the error sum of squares in the following table

Model for μ	Error Sum of Squares	Error df
Full	81.264	12
$\beta_0 + \beta_1 S_i + \beta_2 S_i^2 + \beta_3 F_i + \beta_4 F_i^2$	82.389	13
$\beta_0 + \beta_1 S_i + \beta_2 S_i^2 + \beta_3 F_i$	104.167	14
$\beta_0 + \beta_1 S_i + \beta_2 S_i^2$	169.500	15
$\beta_0 + \beta_1 S_i$	174.194	16
$\beta_0 + \beta_1 S_i + \beta_3 F_i + \beta_4 F_i^2$	87.083	14
$\beta_0 + \beta_1 F_i + \beta_2 F_i^2$	189.167	15
$\beta_0 + \beta_1 F_i$	210.944	16
$\beta_0 + \beta_1 S_i + \beta_3 F_i$	108.861	15

I begin by asking whether the 2nd degree polynomial terms, that is, those involving β_2, β_4 and β_5 need be included. To do so I compare the top line with the model containing only $\beta_0 + \beta_1 S_i + \beta_3 F_i$. The extra SS is 108.861-81.264 on 3 degrees of freedom which gives a mean square of (108.861-81.264)/3= 9.199. The MSE is $81.264/12 = 6.772$. This gives an F-statistic of 9.199/6.772=1.358 on 3 numerator and 12 denominator degrees of freedom. This gives a P-value of 0.30 which is not significant. We would then delete the quadratic terms and consider the coefficients of S and F. We have a choice between pretending that the last line in the table is now the "Full" model and forming the F-statistics (210.944-108.861)/(108.861/15) = 14.066 and (174.194-108.861)/(108.861/15) = 9.002. The first is for testing $\beta_1 = 0$ and the second for $\beta_3 = 0$. Each is on 1 and 15 degrees of freedom. The corresponding P-values are 0.002 and 0.009. This are both highly significant and we conclude that both Sand content and Fibre content have an impact on hardness and that there is little reason to look for non-linear impacts of the the two factors.

An alternative starting point would be to check first to see if the interaction terms could be eliminated, that is, test the hypothesis that $\beta_5 = 0$. This hypothesis can be tested either using the F statistic [(82.389-81.264)/1]/[12.264/12] = 0.166 or using the t-statistic which is $\hat\beta_5/\hat\sigma_{\hat\beta_5}$ and which SAS calculates to be -0.41. Note that $(-0.41)^2 = 0.166$ to within round-off error. Algebraically $F = t^2$. Note, too, that the t test can be made one-sided while the F-test cannot.

Here is SAS CODE and output.