STAT 350: Lecture 22

Reading:

Goodness-of-fit: Pure Error Sum of Squares

If, for each (or at least sufficiently many) combination of covariates in a data set, there are several observations, we can carry out an extra sum of squares F-test to see if our regression model is adequate. Suppose that $x_1, \ldots, x_K$ are the **distinct** rows of the design matrix and suppose we have n_1 observations for which the covariate values are those in x_1, n_2 observations with covariate pattern x_2 and so on. Of course $n_1 + \cdots + n_K = n$. We compare our final fitted model with a so-called saturated model by an extra sum of squares F-test. To be precise we let α_1 be the mean value of Y when the covariate pattern is x_1, α_2 the mean corresponding to x_2 and so on. Relabel the n data points as $Y_{i,j}; j =, \ldots, n_i; i = 1, \ldots, K$ and fit a one way ANOVA model to the $Y_{i,j}$. The error sum of squares for this FULL model is

$$ESS_{FULL} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_{i,\cdot})^2$$

This ESS is called the pure error sum of squares because we have not assumed any particular relation between the mean of Y and the covariate vector x. We form the F statistic for testing the overall quality of our model by computing the "lack of fit SS" as

$$ESS_{Restricted} - ESS_{FULL}$$

where the restricted model is the final model whose fit we are checking.

As an example return to the plaster hardness data of Lecture 12. There are 9 different covariate patterns corresponding to all the possible combinations of the 3 levels of SAND and 3 levels of FIBRE. There are two ways to compute the pure error sum of squares: create a new variable with 9 levels which labels the 9 categories or fit a two way ANOVA with interactions:

DATA

0	0	1	61	34
0	0	1	63	16
15	0	2	67	36
15	0	2	69	19
30	0	3	65	28
30	0	3	74	17
0	25	4	69	49
0	25	4	69	48
15	25	5	69	43
15	25	5	74	29
30	25	6	74	31
30	25	6	72	24
0	50	7	67	55
0	50	7	69	60
15	50	8	69	45
15	50	8	74	43
30	50	9	74	22
30	50	9	74	48

SAS CODE

```
options pagesize=60 linesize=80;
data plaster;
infile 'plaster1.dat';
input sand fibre combin hardness strength;
proc glm  data=plaster;
 model hardness = sand fibre;
run;
proc glm  data=plaster;
 class sand fibre;
 model hardness = sand | fibre ;
run;
proc glm  data=plaster;
 class combin;
 model hardness = combin;
run;
```

EDITED OUTPUT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	167.41666667	83.70833333	11.53	0.0009
Error	15	108.86111111	7.25740741		
Corrected Total	17	276.27777777			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	202.77777778	25.34722222	3.10	0.0557
Error	9	73.50000000	8.16666667		
Corrected Total	17	276.27777778			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	202.77777778	25.34722222	3.10	0.0557
Error	9	73.50000000	8.16666667		
Corrected Total	17	276.27777778			

From the output we can put together a summary ANOVA table

Source	df	SS	MS	F	P
Model	2	167.417	83.708		
Lack of Fit	6	35.361	5.894	0.722	0.64
Pure Error	9	73.500	8.167		
Total (Corrected)	17	276.278			

- The F statistic is $[(108.86111111 - 73.50000000)/6]/[8.16666667]$.
- The P-value comes from the $F_{6,9}$ distribution.
- The P-value is not significant so there is no reason to reject the final fitted model which was additive and linear in each of SAND and FIBRE.
- Notice that the Error SS are the same for the two-way ANOVA with interactions, which is the second model, and for the 1 way ANOVA.
- This test is not very powerful in general; more sensitive tests are available if you know how the model might break down. For instance, most realistic alternatives will be picked up more easily by checking

for quadratic terms in a bivariate polynomial model as done in class in Lecture 11.

- Notice that the test for any effect of SAND and FIBRE carried out in the one way analysis of variance is not significant. This is an example of the lack of power found in many F-tests with large numbers of degrees of freedom in the numerator. If you can guess a reasonable functional form for the effect of the factors (either the additive two way model with no interactions or the even simpler multiple regression model which is the first model above) you will get a more sensitive test usually.