Lecture 24

ANOVA and Regression

In this lecture we discuss the nested F-test for determining which of two nested models fit the data better and we also discuss ANOVA in a regression context.

Nested F-test

Two linear regression models are nested if they model the same response variable and if all the predictors in one of the models are included in the other. The model with more predictors is called the Full model and the model with fewer predictors is called the Reduced model. The nested models F statistic is used to test whether the Full model is significantly better (explains a significantly larger amount of the variance in the response variable) than the Reduced model. The statistic is:

$$F = \frac{\left(\text{SSE}_{\text{Reduced}} - \text{SSE}_{\text{Full}}\right) / \left(\text{d.f.}_{\text{Reduced}} - \text{d.f.}_{\text{Full}}\right)}{\text{SSE}_{\text{Full}}/\text{d.f.}_{\text{Full}}}$$

The degrees of freedom referred to in the formula above is the degrees of freedom for the error. The numbers that go into the formula come from the ANOVA tables provided in each regression. The F statistic above has numerator degrees of freedom equal to the number of additional predictor variables included in the Full model as compared to the Reduced model. The denominator degrees of freedom is the degrees of freedom associated with the error sum of squares for the Full model.

As an example in class, we worked with the predictor of oxidant levels from weather measurements. The data are available at:
http://mypage.iu.edu/~ehouswor/Fall2004/Math467/L21.html

The models we compared were:
Reduced: Oxidant Levels = β_0 + β_1 Windspeed + β_2 Temperature
and
Full : Oxidant Levels = β_0 + β_1 Windspeed + β_2 Temperature + β_3 Humidity + β_4 Humidity^2

The two ANOVA tables provided by Minitab were:

Reduced:

Source	DF	SS	MS	F	P
Regression	2	819.90	409.95	47.12	0.000
Residual Error	27	234.90	8.70		
Total	29	1054.80			

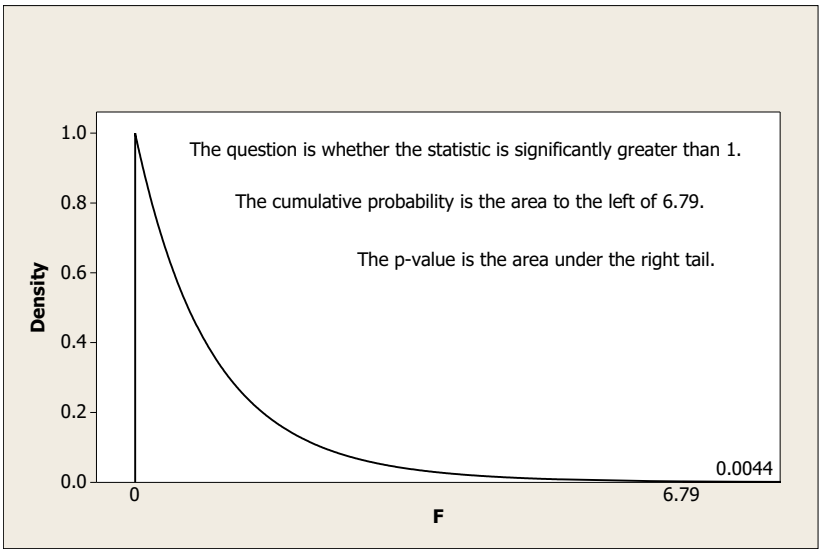
Full:

Source	DF	SS	MS	F	P
Regression	4	902.58	225.64	37.06	0.000
Residual Error	25	152.22	6.09		
Total	29	1054.80			

The nested F test statistic for comparing these two models is thus:

$$F = \frac{(234.90 - 152.22)/(27 - 25)}{152.22/25} = \frac{41.34}{6.09} = 6.79$$

This F statistic has 2 numerator degrees of freedom and 25 denominator degrees of freedom. To determine the p-value we can use Minitab's probability distribution calculators under Calc > Probability Distributions > F.., asking for the cumulative probability for an F distribution with 2 numerator and 25 denominator degrees of freedom up to the value of 6.79. This gives a cumulative probability of 0.995587. The p-value is 1 minus this number (we want the other tail) so the p-value is 0.004413.



ANOVA is a kind of Regression

Using indicator variables that are 1 if the measurement comes from a specific group or category and 0 otherwise allows the analysis of responses to continuous and categorical predictors at the same time. In fact, there is no significant difference between ANOVA and Regression. Regression may sometimes be better because it provides more information although, for precisely the same reason, ANOVA may be preferred when less details are requires for answering the question of interest. To explore these ideas, let's analyze the precent of women in the jury pools of various judge's and the trial judge in Spock's conspiracy case. The data are in the Lecture 11 notes and at:
http://mypage.iu.edu/~ehouswor/Fall2008/BioZ620/Judges.txt

To create indicator variables in Minitab, use the Calc > Make Indicator Variables.. dialog box.

When performing the regression, you need to omit ONE of the indicator variables. The reason is that if you know you are not any of 6 of the judges, then you know that you are the 7^{th} judge so that all 7 indicator variables are perfectly dependent (and redundant.) The variable you choose to leave out becomes the reference variable. The regression coefficients will change depending on which variable is the reference, but the model itself is exactly the same no matter which variable is the reference. The only thing that changes is the interpretation of the coefficients.

For the percent women in the judges' venires, using the trial judge in Spock's case as reference leads to the following regression model:
Percent Women =

$15.0 + 19.2\ \text{Judge}_A + 18.5\ \text{Judge}_B + 14.0\ \text{Judge}_C + 12.0\ \text{Judge}_D + 13.1\ \text{Judge}_E + 11.2\ \text{Judge}_F$

So that the average percent women on the trial judge's venires was 15% and every other judge had at least 11.2% more women on their venires on average.

In class, we conducted a nested F-test to determine whether the Full model above was significantly better than the nested model that used a "Not Trial Judge" indicator variable (requiring the slopes for all the Judges A-F to be the same). We determined that the reduced model was just as good as the full model. Looking at the significance of the slope in the reduced model tells us that the trial judge has a significantly different percentage of women on his venires than the other judges had on theirs:

$$\begin{array}{lccc} \text{Percent Women} = & 15.0 & + & 14.5\ \text{Not Trial Judge} \\ & (2.4) & & (2.7) \end{array}$$

Exercises for Lecture 24

1. –

2. –