Biostatistics 201A - Lecture 16 - 10/31/11

Contents:

- Recap Regression version of ANOVA
- Intro to confounding, adjustment, multicollinearity, mediation
- Announcements re lecture/hw schedule

pages = 9

Happy Halloween!

10/31/11
Lecture 16

Announcements:

- Friday discussion will be replaced w/ lecture to make up for Veteran's Day (next Friday). Will use next Wednesday's lab for discussion
- HW5 will be due Monday, Nov. 14th Labs for it start Wednesday
- Midterms will be returned W or F.

ANOVA as Regression

In ANOVA we model

$$Y = \mu_j + \varepsilon$$

μ_j: group mean

ε: variation about group mean

This can be thought of as regression with indicator variables for group membership

Y = pollution level

5 groups (locations):

- parking lot (P)
- cafeteria (C)
- yard (Y)
- air-conditioned class (A)
- window-open class (W)

Let's let yard be our reference group; define four indicators

$$X_P = \begin{cases} 1 & \text{if measure in parking lot} \\ 0 & \text{otherwise} \end{cases}$$

similarly for X_C, X_A, X_W

Regression equation becomes

$$\hat{Y} = b_0 + b_P X_P + b_C X_C + b_A X_A + b_W X_W$$

If measurement is in the yard

$$X_P = X_A = X_C = X_W = 0$$

$$\Rightarrow \hat{Y} = b_0 = \text{average value in Yard}$$

naturally it turns out that

$$b_0 = \bar{Y}_Y = \text{sample average in yard}$$

Other locations, e.g. parking lot

$X_P = 1 \quad X_C = X_A = X_W = 0$

$\hat{Y} = b_0 + b_P \cdot 1 + 0 + 0 + 0$

$= b_0 + b_P$ = average value in parking lot

estimated by $\bar{Y}_P$

$b_0 = \bar{Y}_Y \qquad b_0 + b_P = \bar{Y}_P$

$\Rightarrow b_P = \bar{Y}_P - \bar{Y}_Y$ = difference between parking lot and yard

How do we write our hypotheses?

F test:

ANOVA: $H_0: \mu_Y = \mu_P = \mu_A = \mu_C = \mu_W$

H_A: vs not all means =

Regression: $H_0: \beta_P = \beta_A = \beta_C = \beta_W = 0$

i.e. none of the other areas differs from the yard!

either way use F test.

Regression: Printout also gives us t-tests for β's, e.g.

$H_0: \beta_p = 0$

$H_A: \beta_p \neq 0$

Tests whether or not parking lot differs from yard - one of our pairwise comparisons! ($\mu_Y = \mu_p$ vs $\mu_Y \neq \mu_p$)

You get for free comparison of each group to reference group. To get other comparisons use "test" or "lincom" statements

Suppose we want to compare parking lot to cafeteria:

mean parking lot: $b_0 + b_p$

mean cafeteria: $b_0 + b_c$

Difference $= (b_0 + b_p) - (b_0 + b_c)$

$= b_p - b_c$

$H_0: \beta_p = \beta_c$

$H_A: \beta_p \neq \beta_c$

Other more complicated combos work the same way.

This handout includes both an ANOVA and a regression printout for Problem 3 from the midterm. Recall that the investigator was collecting pollution measurements at elementary schools in 5 locations: in the parking (P) lot at morning drop-off, in the play yard at afternoon recess (R), in the cafeteria (C) at lunch time, and in two classrooms right before lunch, one with the windows closed and air conditioner running (A) and one with the windows open (W).

. oneway particle location, tabulate bonferroni

location	Summary of particle Mean	Std. Dev.	Freq.
A (class, AC)	14.2	7.1	25
C (cafeteria)	19.1	6.9	25
P (parking lot)	37.9	6.7	25
W (window open)	36.8	5.1	25
R (play yard)	41.5	9.0	25
Total	29.9	13.1	125

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	15225.6	4	3806.4	76.13	0.0000
Within groups	6000.0	120	50.0		
Total	21225.6	124	171.2		

. regress particle parkinglot windowopen cafeteria aircon

Source	SS	df	MS		
Model	15225.6	4	3806.4	Number of obs	= 125
Residual	6000.0	120	50.0	F(4, 120)	= 76.13
Total	21225.5816	124	171.174046	Prob > F	= 0.0000
				R-squared	= 0.7173
				Adj R-squared	= 0.7079
				Root MSE	= 7.0711

particle	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
parkinglot	-3.626978	2	-1.81	0.072	-7.586838	.3328833
windowopen	-4.746978	2	-2.37	0.019	-8.706838	-.7871167
cafeteria	-22.42698	2	-11.21	0.000	-26.38684	-18.46712
aircon	-27.30698	2	-13.65	0.000	-31.26684	-23.34712
_cons	41.50698	1.414214	29.35	0.000	38.70693	44.30702

$$SSB = \sum_{j=1}^{K} n_j (\bar{Y}_j - \bar{\bar{Y}})^2$$

$$SSW = \sum_{j=1}^{K} (n_j - 1) s_j^2$$

The usual extension of ANOVA is something called ANCOVA (Analysis of Covariance) and it really means ANOVA with some continuous predictors added — i.e. a multiple regression with some indicators and some continuous X's. Intuitively your main interest is in comparing groups but there are other factors you want to control for. Why?

In observational studies the groups may differ on other factors (e.g. age, gender balance, SES, ---) that either obscure or falsely project the relationship of interest. We often want to "adjust" for such factors.

Example: Pollution data - turns out that the measurements in the various locations were taken at different times of day. Pollution rose w/ higher temperature - i.e. later in day

Turned out Parking lot and yard (11 a.m. / 1 p.m.) didn't look different in the original ANOVA

After adjusting for temp/time of day these locations did look different.

This leads to a set of related ideas:

1. Confounding
2. Adjustment
3. Mediation
4. Multicollinearity

---

5. Moderation

① Confounding occurs if you get a meaningfully different interpretation of the relationship between the outcome, Y, and a predictor, X, depending on whether or not a third variable, Z, (the potential confounder) is included in your model.

Note: Confounding is a threat to an interpretation of causality (i.e. if X is related to Y but after you adjust for Z that significance goes away it's hard to be sure X "causes" Y) - this doesn't change the fact of the relationship between X and Y; it just goes to how likely we feel it is that X is driving Y.

- To be a confounder, at a minimum, Z must be related to both X and Y. Some people restrict confounding to the case where Z causes/makes both X and Y.