Biostatistics 201a – Lecture 15

10/26/11

Contents

- Multicategory indicators
- ANOVA as Regression
- Exam Review

10/26/11
Lecture 15

More on Indicators + ANOVA

Recall jogging example I had defined 2 indicators

$X_s = X_{hilly} = \begin{cases} 1 & \text{if terrain hilly} \\ 0 & \text{if not} \end{cases}$

$X_q = X_{sandy} = \begin{cases} 1 & \text{sandy terrain} \\ 0 & \text{if not} \end{cases}$

flat terrain was referenerance

b_s = difference in $\bar{y}$ rate between running on hilly vs flat

b_q " sandy vs flat

	X_s	X_q
flat	0	0
hilly	1	0
sandy	0	1
(hilly+sandy	1	1)?

" coding "

Suppose we think sandy terrain is worst and we want to see the successive differences in β rate as terrain gets harder. Then we might "code" as follows

$$X_q = X_{\text{sandy}} = \begin{cases} 1 & \text{sandy} \\ 0 & \text{if not} \end{cases}$$

$$X_g = X_{\text{not flat}} = \begin{cases} 1 & \text{not flat} \\ 0 & \text{flat} \end{cases}$$

	X_g	X_q	$Y = \mu + b_g X_g + b_q X_q$
flat	0	0	stuff + 0
hilly	1	0	stuff + $b_g \cdot 1$ + ~~$b_q \cdot 0$~~
sandy	1	1	stuff + $b_g \cdot 1 + b_q \cdot 1$ = stuff + $b_g + b_q$

b_g = difference between hilly and flat

b_q = difference between sandy and hilly

sandy vs flat $b_g + b_q$

ANOVA: Just a regression in which the only predictors are indicators for what group you belong to.

Regression Model

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{\text{Average value of } Y \text{ at a given set of } X\text{'s}} + \varepsilon$$

So .... if the X's tell you what group you're in, this just becomes "average value of Y for that group"

For ANOVA we wrote

$$Y_{ij} = \mu_j + \varepsilon_{ij} = \text{group mean} + \text{error}$$

Two group example

Y = height

$X = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$

Regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Males: $Y = \beta_0 + \beta_1 \cdot 1 + \varepsilon$

$= \beta_0 + \beta_1 + \varepsilon$

Females $Y = \beta_0 + \beta_1 \cdot 0 + \varepsilon$

$= \beta_0 + \varepsilon$

β_0 = average value of Y when $X=0$

= average height of females

β_1 = Difference in average value of Y between men and women

In regular ANOVA we

would first compute
- mean for men
- mean for women

⇒ take difference men - women to compare groups

Regression:
- mean for reference group
- difference between each other group and the reference

If you want to you can work back to get other group means

e.g. men average 6 ft
women " 5.5 ft

In ANOVA $\bar{Y}_{men} = 6$

$\bar{Y}_{woman} = 5.5$

Difference $\bar{Y}_{men} - \bar{Y}_{woman} = .5$

Regression

$Y = 5.5 + .5 X_{men}$ $X_{men} = \begin{cases} 1 & \text{male} \\ 0 & \text{not} \end{cases}$

↑ mean for women ↑ difference

If you view ANOVA as regression

b_0 = mean of reference group

b_j's = difference between indicated group and reference group

$SSR = SSB$

$SS\Sigma = SSW$

$MSR = MSB$

$MS\Sigma = MSW$

F statistics match

degrees of freedom:

ANOVA: $k-1$ (SSB), $n-k$ (SSW) k = # of groups

Regression: p, $n-p-1$ where p = # of X variables

But if we have k categories we need k-1 indicator variables to include them in the regression so $p = k-1$

All tests, CIs, etc. work out identically.

X in SLR
and then as part of MLR
coefficient changes - means?

- In SLR b = change in Y for a 1 unit change in X
- In MLR b = change in Y for a 1 unit change in X after dealing with other variables

If b doesn't change it means none of the other variables explained the same thing about Y that X did

Y = weight
X_1 = height
X_2 = age

- in SLR - asking does height by itself explain weight
- in MLR - asking after adjusting for age does height provide any additional info about weight

Since height and age are related, height adds less in the MLR than it did in SLR.

Power:
Changes w/ sample size
α, sidedness, SD.

As n ↑ power ↑ - more information means more sure of results / easier to detect stuff

As α ↑ power goes up - higher α means it's easier to reject H_0, i.e. easier to claim there's a difference

As sd ↑ power ↓
higher s.d. means more noise / variability / uncertainty there is in estimates and therefore it's harder to see differences

Sidedness: 2-sided test power is lower than for a 1-sided test basically because you have to check both sides before you reject