Biostatistics 201A - Lecture 14 - 10/24/11

Contents:

- Exam announcements
- CI's and tests for β's in MLR
- Indicator variables

pages = 9

10/24/11
Lecture 14

Announcements

- Today's "labs" in office hour room (HW + Exam help)
- Wednesday 8-9 review session in lecture hall replaces lab
- Exam: Friday 8-10 in lecture hall
  Bring: calculator, writing instrument, 2 pages (front + back = 4 sides) notes
  Covers through today's lecture

CI/T-tests in MLR: Evaluate whether a particular X variable is worth adding to model:

CI: $b_j \pm t_{\alpha/2, n-p-1} \cdot s_{b_j}$

- b_j: estimated coefficient
- $t_{\alpha/2, n-p-1}$: d.f. for MSE, p = # variables
- s_{b_j}: standard error

Gives range of possible values for change in Y associated with a 1-unit change in X_j assuming all other variables held fixed.

t-test (classic version)

H_0: $\beta_j = 0$ - Variable X_j does not explain anything about Y, BEYOND what was explained by the other Xs

H_A: $\beta_j \neq 0$ - X_j is worth adding to model - it explains variance in Y which is not explained by other variables

Test statistic: $t_{obs} = \frac{b_j - 0}{s_{b_j}}$ ← H_0 value

$$p\text{-value} = 2P(t_{n-p-1} \geq |t_{obs}|)$$

If we want to do a 1-sided test, ÷ p-value in half

To test values other than 0, say β^*, just replace 0 in t_{obs} by the value we want to test

$$t_{obs} = \frac{b_j - \beta^*}{s_{b_j}}$$ (computer doesn't give these p-values)

Indicator Variables (Dummy Variables)

- Way to include a qualitative variable in a regression model
- You have a characteristic of interest. Let

$$X = \begin{cases} 1 & \text{if subject has characteristic} \\ 0 & \text{if they don't} \end{cases}$$

  - Values are arbitrary - but makes corresponding β easier to interpret
  - Also doesn't matter which category you make 1 and which 0 - just changes sign of β.

- Main issue with indicators is interpretation, of the b's / β's. It doesn't make sense to talk about a 1 unit change in something like gender! Instead b / β gives the difference in Y between people who have the characteristic ($X=1$) and people who don't ($X=0$)

Why? Jogging example X_6 = Gender

$$X_6 = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$$

For a man

$$\hat{Y} = b_0 + b_1X_1 + \cdots + b_5X_5 + b_6(1) + b_7X_7 + \cdots$$

For a woman

$$\hat{Y} = b_0 + b_1X_1 + \cdots + b_5X_5 + b_6(0) + b_7X_7 + \cdots$$

Assume other than gender, same characteristics / run conditions so all X's other than X_6 are =

Subtract:

$$\hat{Y}_{man} - \hat{Y}_{woman} = (b_0 + b_1X_1 + \cdots + b_6(1) + \cdots) - (b_0 + b_1X_1 + \cdots + b_6(0) + \cdots)$$

$$= b_6(1) - b_6(0)$$

$$= b_6$$

$b_6 = -4.0$

$\Rightarrow$ <u>All else equal</u> after jogging a man's O-rate will be 4 beats per minute <u>slower</u> (negative sign) than a comparable woman's

If we'd let $X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$

we would have gotten $b_6 = 4$

i.e. a woman's ♂ rate is 4 bpm faster than an equivalent man's

Note: The CI's and tests for coefficients of indicator variables are just like those for other variables - just modify interpretations accordingly

e.g. Jogging data, CI for β_6 was $[-7.22, -.78]$

i.e. All else equal, we are 95% sure that a man's post-run ♂ rate is between .78 to 7.22 bpm slower than a woman's. Since this CI doesn't include 0 we are confident that men on average have slower ♂ rates.

Corresponding test

$H_0: \beta_6 = 0$ - After accounting for distance, time, etc..... gender tells us nothing extra about @ rate

$H_A: \beta_6 \neq 0$ - Even after adjusting for all those other factors there is a significant difference between men and wom

$t_{obs} = -2.5$ p-value = .016

Reject H_0, conclude there is a gender difference.

What if our categorical variable has more than two possible values? e.g. terrain = flat, sandy, hilly

Use more indicators! In particular if you have c categories you need c-1 indicators. You pick one

category to serve as the reference group (all indicators=0) and 1 indicator each for the remaining categories

Get as interpretation that each b gives the difference between the indicated group and the reference.

e.g. jogging

$b_8 \Rightarrow$ "hilly" $\quad X_8 = \begin{cases} 1 & \text{hilly} \\ 0 & \text{not} \end{cases}$

$b_9 \Rightarrow$ "sandy" $\quad X_9 = \begin{cases} 1 & \text{sandy} \\ 0 & \text{not} \end{cases}$

"flat" is reference

$b_8 = 5 \Rightarrow$ compared to running on flat ground your $\bar{O}$ rate will be 5 bpm faster if you run on hills

$b_9 = 5.5 \Rightarrow$ sandy terrain leads to higher $\bar{O}$ rate than flat terrain

$b_8 - b_9 =$ difference btwn hilly + sandy

Multiple Regression Example

Consider a multiple regression with the following variables:

Y = Heart Rate in beats per minute
X_1 = Distance run in miles
X_2 = Average speed during the run in miles per hour
X_3 = Humidity as a percentage
X_4 = Altitude in thousands of feet above sea level
X_5 = Temperature in degrees Farenheit
X_6 = Gender: $X_6 = 1$ for males and 0 for females
X_7 = Fitness: $X_7 = 1$ if the person exercises regularly and 0 if not
X_8, X_9 = Terrain (flat, hilly or on sand): These are coded as $X_8 = X_9 = 0$ for flat, $X_8 = 1$ and $X_9 = 0$ for hilly and $X_8 = 0, X_9 = 1$ for on sand.

Note: $\bar{Y} = 100, n = 62, p = 9$ where n is the sample size and p is the number of predictors.

The STATA printout for the multiple regression is as follows:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_9 = 0$
H_A: not all β_j are 0

```
. reg HRate Distance Time Speed Weight RestHRate Gender Fitness Hilly Sand
      Source |       SS       df       MS              Number of obs =      62
-------------+------------------------------           F(  9,    52) =  160.889
       Model |      23168        9     2574.22         Prob > F      =  0.0000
    Residual |        832       52       16.00         R-squared     =  0.965
-------------+------------------------------           Adj R-squared =  0.959
       Total |      24000       61      393.44         Root MSE      =  4.000

------------------------------------------------------------------------------
       HRate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Distance |         3.0       1.80     1.67   0.101      -0.62       6.62
        Time |       -0.05       0.04    -1.25   0.217      -0.11       0.03
    Humidity |         0.0       1.30     0.00   1.000      -2.61       2.61
 Temperature |         0.1       0.04     2.50   0.016       0.02       0.18
    Altitude |         1.0       0.25     4.00   0.000       0.50       1.50
      Gender |        -4.0        1.6    -2.50   0.016      -7.22      -0.78
     Fitness |       -10.0        3.6    -2.80   0.004     -17.24      -2.76
       Hilly |         5.0        1.2     4.17   0.000       2.59       7.41
        Sand |         5.5        1.0     5.50   0.000       3.49       7.51
       _cons |        70.0       10.0     7.00   0.000      50.10      90.10
------------------------------------------------------------------------------
```

ANOVA table

overall F test

fit statistics

b_j's s_{b_j}'s t-obs p-val CIs for β_j's

2-sided p-values for tests of $\beta_j = 0$ vs $\beta_j \neq 0$