Biostatistics 201a - Lecture 12 - 10/19/11

Contents:

- Inference for β_0, β_1 in SLR
- CIs and PIs for Y in SLR
- Intro to Multiple Regression
- MLR example handout

pages: 11

10/19/11

Inference in SLR continued:

Last time we saw how to do CIs / tests for β_0, β_1. The most interesting test was

$H_0: \beta_1 = 0$ - no relationship between X and Y

$H_A: \beta_1 \neq 0$ - there is a relationship between X and Y

Test statistics:

- either $t_{obs} = \frac{b_1 - 0}{s_{b_1}}$ - measures how far sample slope, b_1, is from H_0 value $\beta_1 = 0$. This tells us if predictor X_1 is useful in explaining Y
- or $F_{obs} \frac{MSR}{MSE} = \frac{\text{explained}}{\text{unexplained}}$ variability answers question of whether the model as a whole does a good job explaining Y
- In SLR these tests are the same (and $F_{obs} = t_{obs}^2$) since the one variable is the whole model. This changes in MLR

We can do 1-sided tests or test values other than 0. In general we have

$H_0: \beta \begin{matrix} \leq \\ = \\ \geq \end{matrix} \beta^*$ ← value of special interest

$H_A: \beta \begin{matrix} > \\ \neq \\ < \end{matrix} \beta^*$

$$t_{obs} = \frac{b - \beta^*}{s_b}$$

only change is plugging in a different H_0 value

$$\text{p-value} = \begin{cases} P(t_{n-2} \geq t_{obs}) \\ 2P(t_{n-2} \geq |t_{obs}|) \\ P(t_{n-2} \leq t_{obs}) \end{cases}$$

The Stata printout doesn't give these except for the 2-sided tests when $\beta^* = 0$. For 1-sided tests of 0, divide STATA's p-value by 2. For other values of β^* use a contrast statement or compute t_{obs} by hand and use the ttail command.

In addition to knowing how well we estimate β_0, β_1, we might want to know how much uncertainty there is in predictions of Y.

Y = height of child
X = age

Two forms:

- Average value of Y at a given X ($\mu_{Y|X}$) = value on the population regression line
  → get a C.I.
  e.g. how sure are we about the average height of 8 year olds?
- Value of an individual Y at a given X - requires a "prediction interval" or P.I.
  What are the possible heights for my eight year old?

As usual our intervals have form:

estimate $\pm t_{\alpha/2, df} \cdot SD(\text{estimate})$

$\hat{Y}_0 = b_0 + b_1 X_0 \pm t_{\alpha/2, n-2}$ · $s_{\hat{Y}_0}$ ⇒ s.d. for estimating $\mu_{Y|X}$

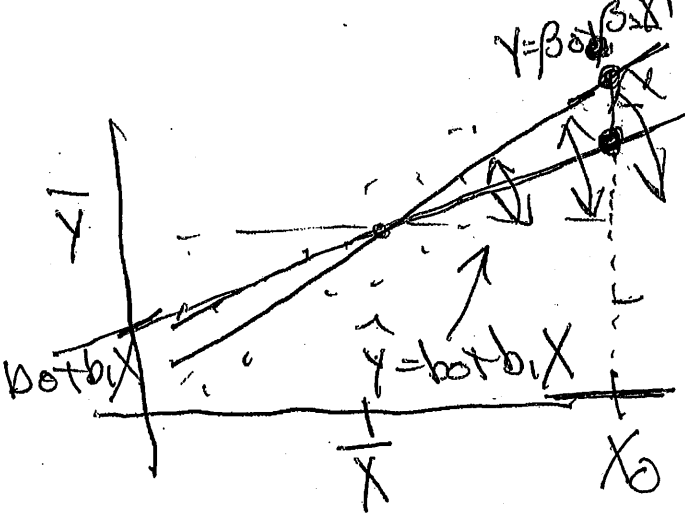
· s_{Y_0} ⇒ s.e. of prediction for single Y

d.f. that go with MSE

For C.I. for $\mu_{Y|X_0}$: X_0 is given value of X

$$s_{\hat{Y}_0} = RMSE \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SSX}}$$

Depends on - variability about line
sample size
how far X_0 is from center of data (extrapolation)
spread of X's in sample



For a P.I. the formula changes slightly:

$$s_{Y_0} = RMSE \sqrt{1 + 1/n + \frac{(X_0 - \bar{X})^2}{SSX}}$$

(uncertainty in line)

(1 ↑ accounts for individual variation about line.)

A P.I. will always be wider than a C.I. Why? For the C.I. the uncertainty is in estimating the average Y at a given X - i.e. the value on the line. A P.I. is for an individual - even if we knew the true line exactly we couldn't predict an individual perfectly. The extra 1 in the standard error calc. corresponds to an extra factor of MSE - the variation of individuals about the line

$$s_{Y_0}^2 = MSE\left(1 + 1/n + \frac{(X_0 - \bar{X})^2}{SSX}\right) = MSE + s_{\hat{Y}_0}^2$$

($s_{Y_0}^2$: individual variation; MSE: individual about line; $s_{\hat{Y}_0}^2$: line uncertainty)

Multiple Linear Regression

Same as SLR except that there are more variables which changes degrees of freedom and some of the interpretations. All interpretations have to account for the presence of the other variables in model.

Example:

Y = heart rate in beats/minute after a run

Possible X's:
- how long run was
- how fast person went
- fitness
- altitude
- ------

Model becomes

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$

Y_i = response/outcome for subject i

X_{ij} = value of the j^{th} predictor for subject i

(e.g. X_{32} = value of variable 2 for the 3rd subject)

β_0 = intercept = average value of Y when all the X's = 0

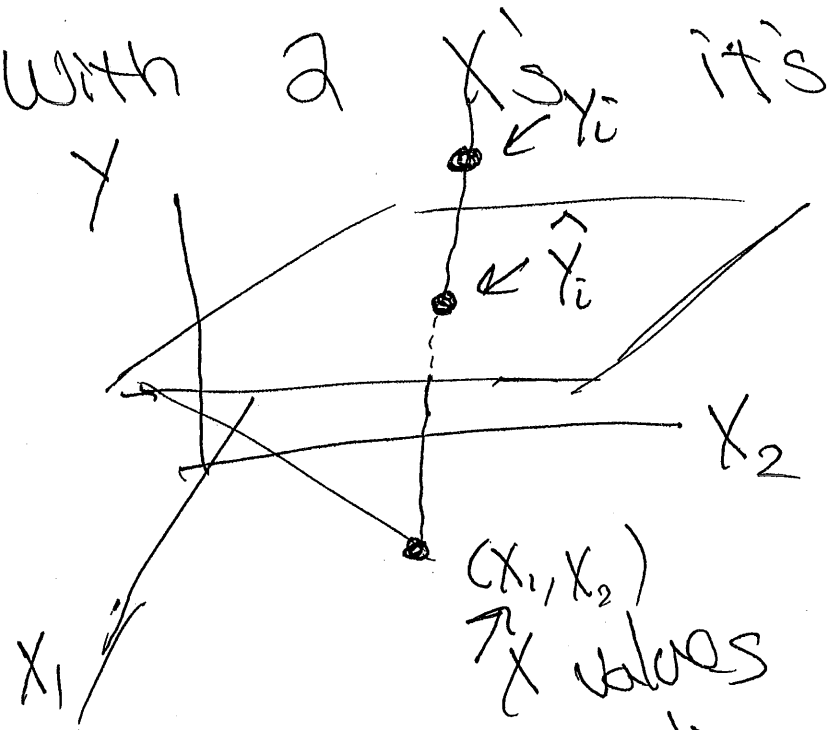
β_j = "slopes" associated with predictor X_j - 1 unit change in X_j is associated with a β_j change in Y

All Else Equal

ε_i = individual variation

p = number of predictors

Graphically what does this look like? It's a hyperplane with 2 X's, it's just a plane



↑ X values

Hard to visualize.

Calculations also messier - they involve matrices (linear algebra)

(Aside: matrix notation for MLR)

$$\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & & X_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & & & X_{pn} \end{pmatrix}$$

↑ design matrix

↑ for intercept

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\vec{Y} = X\beta + \varepsilon$$

As in simple linear regression we don't usually know the β's so we estimate them using least squares - minimizing squared distances from points to the plane

$$\text{minimize} \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip})^2$$

It turns out you can write the least squares solution in matrix form as

$$\hat{\beta} = b = \underbrace{(X^T X)^{-1}}_{\text{like } SS_X} \underbrace{X^T Y}_{\text{like } SCP}$$

note in SLR

$$\text{slope } b_1 = \frac{SCP}{SS_X} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Let STATA, SAS do this!

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:

```
. 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
------------------------------------------------------------------------------
```

Main differences are degrees of freedom and # of predictors in the parameter estimates table