Biostatistics 201a - Lecture 13 - 10/21/11

Contents

- Interpreting MLR coefficients
- MLR ANOVA table / fit stats
- MLR Inference

#pages = 12

10/21/11
Lecture 13

Interpreting a Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

(1) β_0 or b_0 - intercept: average value of Y when <u>all</u> X's = 0

Makes practical sense only if 0 is a realistic value for all variables. Otherwise just think of this as a baseline value on which to build model.

Example:

Y = heart rate

X's = distance run, speed, humidity, fitness, etc.

$b_0 = 70$ = average heart rate if a person doesn't run (so speed = 0 too), no humidity, doesn't exercise, etc.

i.e. a resting HR rate in weird conditions

(2) β_j or b_j: Slope associated with predictor X_j. Change in Y associated with a 1 unit change in X_j ASSUMING ALL THE OTHER VARIABLES ARE HELD FIXED/CONSTANT

Why? If we don't we can't tell which of the variables that changed was connected with the change in Y

e.g. change how far person runs and as a result they run slower - what "causes" any change in ♡ rate?

Note: Not always possible to get data that fixes all but one variable.

e.g. $b_1 = 3$ - distance variable

So this says on average every extra mile run is associated with a 3 beats per minute higher heart rate assuming all the

other running conditions are the same - eg, have the same person do the same run twice but one time run 1 mile further.

Predictions in MLR

Estimated model

$$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_p X_p$$

plug in values for all the X variables

How do we evaluate how good a job our MLR is doing?

ANOVA Table

SST = "sum of squares total"
= variability in Y

Doesn't depend on X's so it's the same whether SLR or MLR

Divide into two pieces

SSR = "sum of squares for regression"

= variability in Y that is explained by our group of variables

SSE = "sum of squared errors"

= variability in Y not explained by any of the X's (i.e. factors we missed or measurement error)

As before SSR + SSE = SST

Degrees of freedom:

SST: n-1 as before - just involves computing $\bar{Y}$

SSE: n-p-1 where p = # of X's

Why? we have to estimate our "plane" first before

we can look at variation about plane → need $b_0, b_1, b_2, \ldots, b_p$

$p+1$ numbers

SSR: p - one for each variable we add to the model as opposed to just guessing $\bar{Y}$ for everything

Mean Squares:

$$MSR = \frac{SSR}{p} = \text{Variability explained (on average)} \quad \frac{PER}{\text{PREDICTOR VARIABLE}}$$

$$MSE = \frac{SSE}{n-p-1} = \text{Variability not explained} \quad \frac{PER}{\text{DATA POINT BASIS}}$$

adjusted ($n-p-1$ rather than n) to give us an unbiased estimate of σ^2, the variance of points about the plane

$F_{obs} = \frac{MSR}{MSE}$ = ratio of explained to unexplained variability

Under H_0: model is useless,

i.e. $\beta_1 = \beta_2 = \cdots \beta_p = 0$

F_{obs} has F distribution with $p, n-p-1$ degrees of freedom

⇓

ANOVA table.

Source	SS	df	MS	F	Prob > F
Model	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$	
Error	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$		
Total	SST	$n-1$			

From the ANOVA table we get RMSE, R^2, R^2_{adj}, F test

RMSE = "root mean squared error"
= average error we make when using all the X's to predict Y. As before to judge how well we're doing we need to compare to Y values

Example: jogging RMSE = 4 beats/min.
on average our predictions using distance, speed, fitness, running conditions, ---- are off by 4 bpm. Printout Y = 100 was average running ♡ rate so ≈ 4% error — pretty good

R^2 and R^2_{adj} = percentage of variability in Y explained by all the Xs as a group

$$R^2 = \frac{SSR}{SST}$$

$$\rightarrow R^2_{adj} = 1 - \frac{SSE}{SST} \cdot \frac{(n-1)}{(n-p-1)}$$ ← penalizes you for adding extra useless variables

R^2_{adj} produces an unbiased estimate of the population %. In SLR this doesn't matter much since $p=1$ but MLR if p is high it matters.

Example: Jogging $R^2 = .965$, $R^2_{adj} = .959$
Not very different and very high - we've explained $\approx 96\%$ of variation in O rate.

F test: Measures overall how good a job the model (group of X variables) does

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ - none of X variables helps to explain Y (i.e. none of distance, speed, conditions, etc. explain running @ rate)

H_A: At least one $\beta_j \neq 0$ - At least one of the predictors helps to explain Y (heart rate)

Test statistic $f_{obs} = \frac{MSR}{MSE}$

$$p\text{-value} = P(F_{p,n-p-1} \geq f_{obs})$$

Example: $F_{obs} = 160.9$

$$p\text{-value } P(F_{9,52} \geq 160.9)$$

$$= .0000$$

Reject H_0, conclude at least one of our 9 predictors is related to @ rate.

Which one? t-tests!!

For that we use t-tests for individual variables:

H_0: $\beta_j = 0$ - X_j does not explain any variability in Y THAT IS NOT EXPLAINED BY OTHER Xs, i.e. X_j is not worth adding to the model if I have other variables

H_A: $\beta_j \neq 0$ - X_j is worth adding - it provides new info about Y

This is different from asking whether X_j by itself is related to Y.

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

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