STA 302 / 1001

Note Title 11/29/2011

Unambiguous, accurate, deep

No re-mark request accepted after Dec. 9 5:00 pm.

See Blackboard for Announcements regarding:

- questions you can safely ignore on old exams
- formula sheet you are given on exam

House price example

Fitted equation

$$\widehat{price} = 18.6 - 7.70\,bdr + 0.0176\,flr + 6.91\,fp + 3.90\,rms + 10.8\,st + .263\,lot + 2.37\,bth + 1.77\,gar$$

(price in $1000)

Interpret coefficients

b_j, $j = 1, \ldots, 8$ represents the estimated change in mean selling price due to a unit change in X_j all other variables held constant

Example: 100 sq feet ↑, est. mean price ↑ \$1760, everything else constant

(more bedrooms, est. mean price ↓ \$7700, everything else constant

(despite the fact that the correlation between price and bdr is positive)

t-test of H_0: $\beta_{bath} = 0$ vs H_a: $\beta_{bath} \neq 0$

$p = 0.3662$

So no evidence that bath contributes to the prediction of selling price over and above other variables (despite the fact that the p-value for the test that the price–bath correlation is 0 is 0.0034)

F-test : $H_0: \beta_1 = \dots = \beta_8 = 0$ $p < .0001$

Strong evidence that not all coef of predictor variables are 0.

MULTICOLLINEARITY

When explanatory variables are highly correlated:

- difficult or impossible to measure the individual variable's influence on the response
- the fitted equation is unstable
- the estimated regression coefficients vary widely from data set to data set (even if data sets are very similar)

and depending on which other predictor variables are in the model

- the estimated coef. may have opposite sign to what you expect (eg bdr in house price example)
- the coef. might not be statistically significantly different from 0 even though there is a strong relationship between that X and Y when only considering X and Y

"Multicollinearity" = lots of correlation among X's

"ill-conditioning"

Recall: $b = (X'X)^{-1} X'Y$

If X's perfectly correlated, $X'X$ would be singular — can't calculate b

Even if $X'X$ close to singular, the determinant of $X'X$ close to 0 $\Rightarrow$ standard errors of estimated coefficients will be large

$$Var(\underset{\sim}{b}) = \sigma^2 (X'X)^{-1}$$

So we have "inefficient" estimates.
(can't make precise statements about its value)

Quantify Multicollinearity : VARIANCE INFLATION FACTOR

$$VIF_j = \frac{1}{1 - R_j^2}$$

R_j^2 = coefficient of multiple determination obtained when j-th predictor variable is regressed against other predictor variables

Large VIF_j is a sign of multicollinearity

What is large?

Rules-of-thumb $VIF_j > 10$ - serious problem

If $5 < VIF_j < 10$ - warning that effects of multicollinearity may be seen

But don't worry about high VIF_j if the variable's coefficient is stat. sig. different from 0.

House price example $VIF_{rms} = 8.5$ $VIF_{bdr} = 6.5$

We did see effects of multicollinearity eg. opposite sign of coef of bdr

$\text{Tolerance}_j = 1 / VIF_j$

Proposed solutions to fitting regression models with serious multi-collinearity

- ridge regression
- robust regression
- principle components regression

} Mathematical approaches; sometimes difficult to interpret in a practical way.