Biostatistics 201a - Lecture 24 - 11/21/11

Contents

- Outlier recap
- Intro to model selection
- Partial F / R^2 difference test
- Stepwise procedures

pages = 9

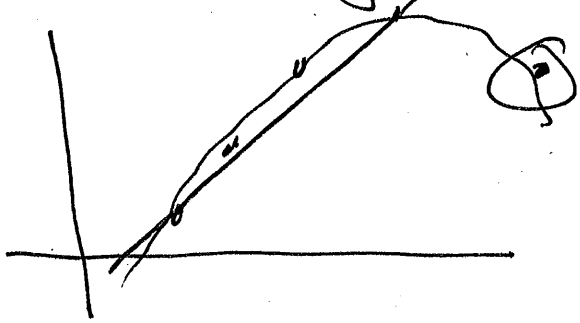
11/21/11

Outlier Recap:

- Last time we saw a bunch of methods for detecting whether a point was
  (a) unusual (e.g. large residual/ far from center of data)
  (b) influential (i.e. affected the predictions or regression coefficients)
- What do we do with such points if we find them?
  (a) Try to find out what caused it - if it seems "wrong" in some way, either fix it or remove.
  (b) If the point looks legitimate and you can't find a reason why it's illegitimate, fit model with and without the point, say what the difference in interpretation is and try to make an argument for which model is more appropriate.

Note: What counts as an outlier depends on what model you fit - both what predictors are in it and what shape it has

- e.g. Lily Lip on midterm - prediction was off w/out knowing she had hypercholesterolemia



← this would be a very influential outlier in an SLR, but in a quadratic model it's fine

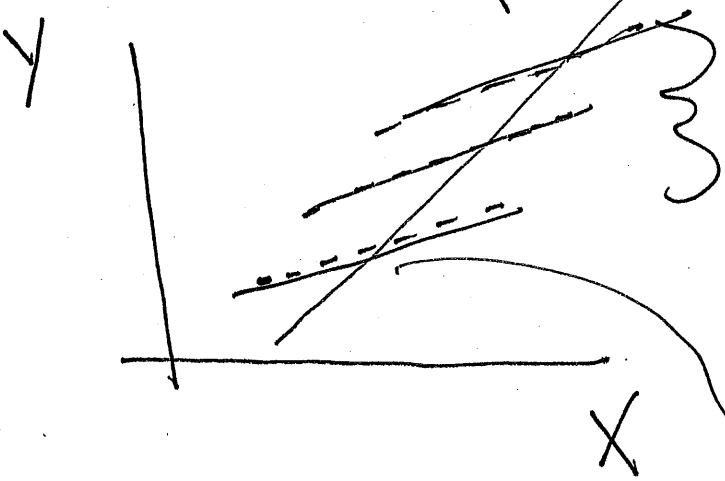
You need to identify model first and then whether you have outliers but it's an iterative process

Model Selection

So far we've been assuming we have the right set of predictors and we just need to find the best way to combine them. But what if

we have lots of possible predictors and we need to choose which to use?

- If we miss some predictors our model ends up being "underspecified"; usually leads to biased parameter estimates



If use group variable I get a shallow slope

If I don't I get a much steeper slope.

- If you have extra variables, you either tend to get overfitting leading to poor predictors or multicollinearity and its associated problems - this called over-specification

There's a tradeoff - need some criteria to identify "best" models

There are lots of criteria.....

- Summary statistics (e.g. R^2, RMSE, ...)
- Which obey model assumptions better
- Look at predictions for new data points and see how well the model does
- Formal tests to compare two (nested) models
- Procedures for picking variables based on all these things

Let's start with a formal test for comparing two models:

Partial F test or R^2 difference test

Model 1: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

Model 2: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta X_{p+1} + \cdots \beta_{p+m} X_{p+m} + \varepsilon$

Model 2 has m extra predictors - to tell if model 2 is better we need to know if predictors $X_{p+1}, X_{p+2}, \ldots, X_{p+m}$ added

extra explanatory power:

$H_0: \beta_{p+1} = \cdots = \beta_{p+m} = 0$ – model 2 does not explain any more about Y than model 1

H_A: at least one of those βs $\neq 0$ – Model 2 is a significant improvement

Looks just like overall F test, just for a subset of variables. In STATA, you'd write

test "X_{p+1}" = "X_{p+2}" = $\cdots$ = "X_{p+m}" = 0

(Variable names)

Test statistic

$$F_{obs} = \frac{(SSR_2 - SSR_1)/m}{SSE_2/n-(p+m)-1}$$

("full model" = 2; "reduced" model = 1; m = difference in # of predictors)

(Same structure as overall F test comparing actual model to model w/ no predictors)

This has an F distribution with m, $n-p-m-1$ degrees of freedom.

Also called the R^2 difference test: Divide top and bottom by SST

$$F = \frac{\left(\frac{SSR_2}{SST} - \frac{SSR_1}{SST}\right)/m}{\left(SSE_2/SST\right)/n-p-m-1}$$

$$= \frac{\left(R_2^2 - R_1^2\right)/m}{\left(1-R_2^2\right)/n-p-m-1}$$

difference in R^2's for the two models

i.e. "Did my extra variables explain a significant additional proportion of variability?"

This is useful for checking whether special blocks of variables, e.g. - socio-demographic set
- multicategory indicator set
- interaction set
- multiple powers of an X variable

Going with it is something called the hierarchical principle. If you have a set of variables representing a single concept (e.g. all racial categories; all powers of X ⇒ shape) you either keep them all in or take them all out.

How do we get a set of models to compare?

Easy + common approach (with flaws that we'll discuss next time) is "stepwise model selection"

Idea:

Forward stepwise: Start with no variables in model; first add the best single predictor (i.e. highest correlation or smallest SLR p-value). Then find the next variable that makes the biggest improvement to the current model. Keep going until nothing else worth adding.

Backwards stepwise: Start with all variables in; at each step throw out the least useful one (e.g. largest p-value) until everything left is useful