STA 302/1001

Note Title 12/1/2011

Office Hours next week

Me Mon 11-1 SS 5016A (as usual, priority to STA220)

Andrey Tues 11-1 UC 244

Me Wed 11:15-1 SS 5016A

Me Thurs 2-4 SS 5016A

Plus New College Stats Aid Centre.

Partial Regression Plots

or Added Variable Plots (+ textbook)

(or Adjusted Variable Plot or Partial Residual Plot)

For visualizing the relationship between a response and an explanatory variable over and above the other explanatory variables.

What is Plotted: Added Variable Plot for X_j:

- residuals from regression of Y vs other X's.

versus

- residuals from regression of X_j vs other X's

("partial correlation" is the correlation between those 2 sets of residuals)

Linear pattern indicates X_j useful in model (untransformed) over and above other explanatory variables

Shows strength of linear relationship between Y and X_j over and above other variables

Also useful for detecting non-linear relationships (polynomial in X_j or transform X_j to make linear?), outliers, non-constant variance, influential points

Assessing Model Assumptions

- Plots of Y vs X_j — is linear model appropriate
  - unusual points (potentially, outliers, influential)
  - also look at Added Variable Plot

Hover example — in plot of price versus bdr
- not a strong linear relationship
  - + potential influential point (8 bdr)
- SLR maybe not a good idea for price–bdr relationship but may be OK in multiple reg^n

- check if influential in added variable plot
  - linear relationship is stronger (although its now negative)
  - large negative residual is not the house with 8 bdr.

- otherwise no concerns with non-linearity

Concerns: - large negative residual in partial plots
- potential influential point in added

variable for flr

- Residual plots
  - (Standardized) residuals versus X_j, $j = 1, \ldots, p$
    Look for: curvature, influential points, (outliers)
  - (Standardized) residuals versus $\hat{Y}$
    Look for non-constant variance, outliers in residuals
  - Normal quantile plot of residuals
  (- Residuals versus other potential predictors)

House example → - one house with large negative residual (perhaps one large positive residual)
- an influential point (large's Cook's D in Diagnostics Panel)
Is 7 the house with 8 bdr?

Influence Statistics - same statistics as for Simple Regression.

For leverage $h_{ii} > \frac{2(p+1)}{n}$, $\sum h_{ii} = p+1$

For DFFITS $> 2\sqrt{\frac{p+1}{n}}$

For DFBETAS $> 2/\sqrt{n}$

AND LARGE GAP

For Cook's D $> \frac{4}{n-(p+1)}$

House Example Unusual observations:

Obs 8 - influential - largest Cook's D
- very large floor space, everything else including price was of moderate space

Obs 19 - large positive residual
- small house but middle-ish price; nice house?

Obs 20 - large negative residual
- smallest price, but all features middle-ish

Above with 8 bedrooms is neither an outlier, nor influential; It's expensive but everything is big

---

Why do Regression?

Two main reasons

1. Prediction value of a variable given value(s) of other variable(s)
   - for interpolation, or future prediction
   eg. house price example, assignment 3

Then doesn't matter which variables are in model or how they were chosen
(model building / variable selection OK)
But must consider the level of smoothing that is appropriate. If overfit, predictions may not work well on new data

② Understanding underlying mechanism

- describe the relationship between variables
- What are effects of predictor(s) on response?
- describe differences in relationships between different sets of data

eg Mauna Loa CFCs, meadowfoam, rainfall and corn yield.

This is inferential modelling
Form of model is assumed to be known but there is uncertainty about the value(s) of coefficient(s)

DON'T DO INFERENTIAL MODELING ON THE SAME DATA THAT YOU USED TO BUILD A MODEL

Why do Multiple Regression?

- multiple X's give better prediction of Y
  eg. house price
- want to fit polynomial
  eg. rainfall & corn yield
- want to compare regression line for 2 groups
  eg. meadowfoam
- want to control for some Xs
  eg. baseline value in a study

Another example: mammals & brain size