STA 302 / 1001

Note Title 10/18/2011

See Blackboard announcements for test information

Continuing from last class:

Properties of residuals: $\hat{\epsilon}_i$

$E(\hat{\epsilon}_i) = 0$

$Var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$ (not constant)

$Cov(\hat{\epsilon}_i, \hat{\epsilon}_j) = -h_{ij}\sigma^2$, $i \neq j$

(not uncorrelated)

$\hat{e}_i$'s are "more" normally distributed than e_i's (if e_i's distribution deviates from normal)

by CLT

If standardize residual,

$$r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$$

Two advantages

- r_i's no longer have unequal variance
- easier to find outlies

that are an indication of lack of model fit

$|r_i| > 3$ unusual

$|r_i| > 4$ very unusual

(Small data sets $|r_i| > 2$?)

Watch for $|r_i|$ that are much larger than all of the others.

Typically, don't see much qualitative difference between r_i and $\hat{e}_i$, except for leverage points

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

n very large $\rightarrow 0$ $\quad$ x_i near $\bar{x}$ $\quad$ h_{ii} very small

x_i very far from $\bar{x}$ $\quad$ h_{ii} large

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$$

n very large $\rightarrow 0$ $\quad$ x_i, x_j on same side of $\bar{x}$

x_i very far apart $\quad$ x_i, x_j on opposite sides of $\bar{x}$

In general, often lack of constant variance in the residuals is often accompanied by correlation

Residual Plots for Checking Model Assumptions

Looking for:

- evidence that straight line model is not appropriate
  - curvature
  - outliers
  - influential points

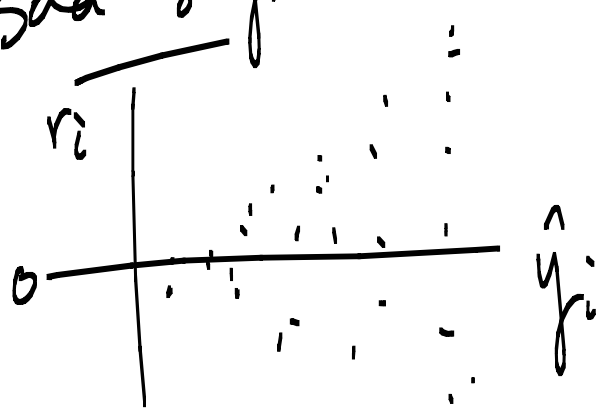
  ↳ for some or all of data
- constant variance
- normally distributed errors

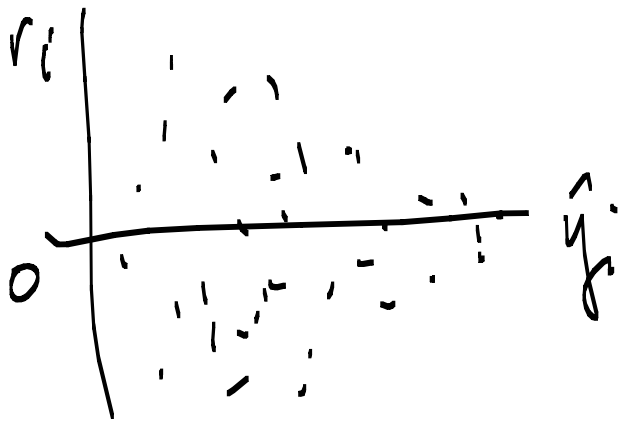
Also (if can)
- lack of independence in errors
- potential missing predictor variables

Recommended Plots

① r_i vs $\hat{y}_i$ (Standardized residuals versus fitted values)

Bad sign!





evidence of non-constant variance

Know $\sum \hat{e}_i \hat{y}_i = 0$ (Is $\sum r_i \hat{y}_i = 0$?)

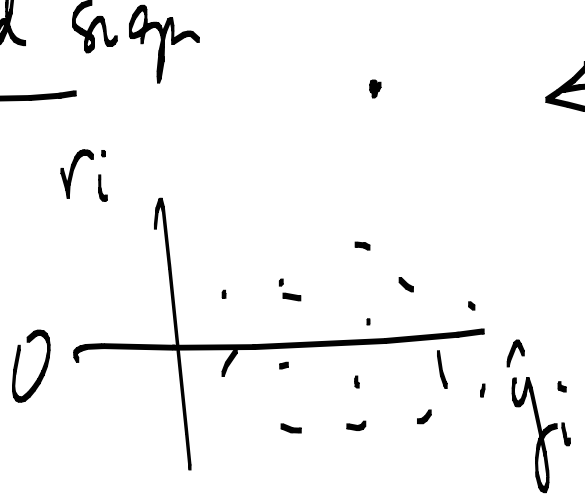
$\Rightarrow$ correlation coefficient (r) of $\hat{e}_i, \hat{y}_i$ is 0

So shouldn't be a pattern in $\hat{e}_i, \hat{y}_i$ plot

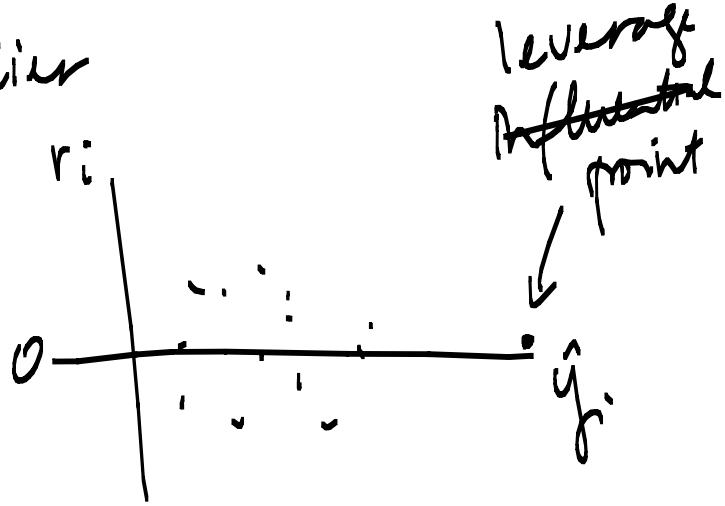
Bad sign

evidence of curvature

Bad sign



← outlier



leverage point

'Good' plot – random scatter about 0

② r_i vs x_i Residuals versus explanatory variable.

- for simple linear regression, looks exactly like plot of r_i vs $\hat{y}_i$ ($\hat{y}_i = b_0 + b_1 x_i$) except scale is different and plot is flipped $b_1 < 0$
- look for same things as r_i vs $\hat{y}_i$ plot

$\sum \hat{e}_i x_i = 0$ so $\hat{e}_i, x_i$ have 0 corr. coef.

Typically, use r_i vs x_i for curvature, unusual observations

r_i vs $\hat{y}_i$ for non-constant, unusual observation

③ Normal quantile plot
- assess normality (come back to this)
- don't bother to assess normality in presence of other problems

Other residual analyses

- univariate analysis of residuals
  - histogram should be bell-shaped
  - should satisfy 68/95/99.7% rule
  - look for outliers, skew

- Plot Square root of absolute value of residuals vs $\hat{y}$
  - increasing pattern ⇒ non-constant variance
- Residuals versus time or other sequence in observations
  - a pattern indicates residuals may be correlated

positive auto-correlation

negative auto-correlation

- Residuals versus other potential predictors
  - Want random scatter
  - any pattern indicators other predictor should be in the model

Assignment / marking

- Marks will be adjusted so each TA will have same distribution
- 3 may not be perfect
- look at solutions
- look at posted TA comments
- re-mark policy: Look at solutions; write down concerns; give to me or Andriy