STA 302 / 1001

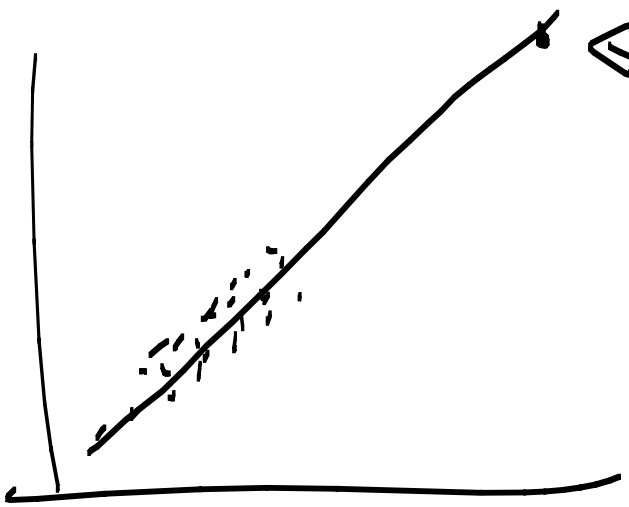
Note Title 10/13/2011

Tuesday Oct 18: Andriy will be here at 11:10 to
- return your assignments
- answer any questions you have about course material, practice problems, old tests, etc.

He'll also have some more office hours before the test.

Influential Points / Outliers / Leverage Points

Influential - take it out and makes a substantial difference in the est. of slope and intercept



This makes a huge difference in R^2

=> makes a substantial difference in fitted values

But is not influential ("Good" leverage point)

Outliers - in x direction, the y - direction or the residuals

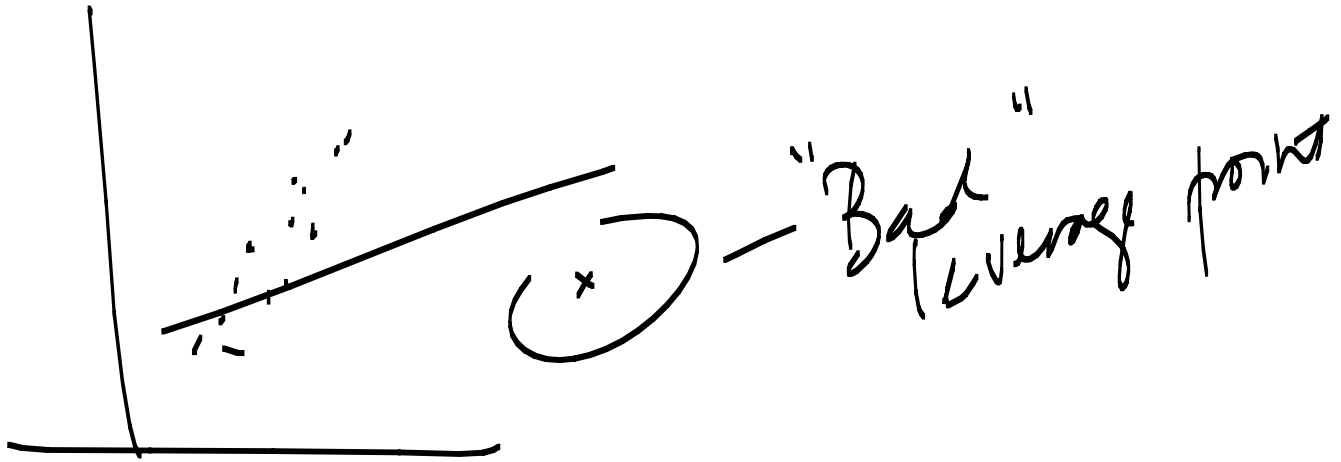
If residuals ⇒ model failure

Outliers with respect to the explanatory variable are called leverage points

They may be influential and possibly cause the s.e. of est. coefficients to be smaller than they would be if point was removed

Sheather calls leverage points that follow

the pattern of the data "good" leverage points and leverage points that are influential are "bad" leverage points.



Quantifying Leverage

"h_{ij}"

$$\hat{y}_i = b_0 + b_1 x_i$$

$$= \bar{y} + b_1 (x_i - \bar{x}), \qquad b_1 = \sum_j \frac{(x_j - \bar{x}) y_j}{S_{xx}}$$

$$= \bar{y} + \sum_j \left[\frac{(x_j - \bar{x}) y_j}{S_{xx}} \right] (x_i - \bar{x})$$

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

$$= \sum_j \left\{ \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right\} y_j$$

$$= \sum_{j=1}^{n} h_{ij} y_j , \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$$

why "h_{ij}", h is for "hat"

Exercise: Show $\sum_{j=1}^{n} h_{ij} = 1$

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

h_{ii} = leverage of i^{th} data point

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

If h_{ii} is large, x_i is far from $\bar{x}$

Suppose $h_{ii} \simeq 1$, $\sum_{j \neq i} h_{ij} \simeq 0 \Rightarrow \hat{y}_i \simeq y_i$

<u>Average</u> of h_{ii}

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

$$= \frac{1}{n} + \frac{1}{n} = 2/n$$

<u>Rule</u> of thumb for what is large for leverage point is $h_{ii} > 2 * \text{average value} = 4/n$

OR large gap between h_{ii} for i^{th} observation and its value for all other observations

Dealing with leverage points

- Remove only if invalid; there's a good reason to say that it's unlike other points
- Is model correct? Maybe a curvilinear model would be more appropriate.

Bad leverage points are an example of model failure

Good leverage points - decrease s.e. of est of β's because increase S_{xx}

- increase R^2

$\Rightarrow$ make you think you have a better fit than you actually do

Measuring Influence of i^{th} Observation

Notation: subscript (i) indicates i^{th} observation has been deleted from calculation

DFBETAS - difference in est. of b's with and without i^{th} observation

$$DFBETAS_{ik} = \frac{b_k - b_{k(i)}}{\text{s.e. of } b_{k(i)}}$$

$i = 1, \ldots, n$

$k = 0, 1$ ← intercept, → slope

$b_{1(i)}$ is est of slope for n-1 observations
(ith observation removed)

Rule of thumb $|DFBETAS_{ik}| > 2/\sqrt{n}$ for large data sets

> 1 for small data sets

or separated by large gap from the other values of DFBETAs

Crime example: $2/\sqrt{23} = .417$

For NYC, $DFBETAS_{1,23} = .98$ ← Influential

$DFFITS_i$ - how i^{th} predicted value changes with and without the inclusion of i^{th} observation

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\text{S.e. of } \hat{y}_{i(i)}}$$

$\hat{y}_{i(i)}$ is predicted at x_i calculated from $n-1$ observations (i^{th} observation excluded)

Rule of Thumb $|DFFITS_i| > 2\sqrt{2/n}$ for large data set

> 1 for small data sets

OR gap between $DFFITS_i$ and DFFITS for all other observations

Crime example $2\sqrt{2/23} = .59$, 3 / (Lawless 23)

Cook's Distance : Measure of influence of i^{th} Observation

$$D_i = \frac{\sum_{j=1} (\hat{y}_{j(i)} - \hat{y}_j)^2}{2s^2}$$

$s^2 = MSE$

Can show $D_i = \frac{r_i^2 h_{ii}}{2(1-h_{ii})}$

r_i is the i^{th} standardized residual

Large D_i can be due to large residual or leverage or both

Rule-of-thumb $D_i > 4/n-2$

OR gaps between D_i & other D_i's

Crime example $4/23-2 = .19$ ($D_i = .53$ for NJC)

SAS also prints Covariance Ratio - ratio of determinants of Var-Cov matrix of β_i with and without i^{th} observation

Residuals - How to use residuals as a diagnostic tool. to test conditions of model that are necessary to make valid inferences.

Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, \ldots, n$

Least squares line: $y_i = b_0 + b_1 x_i$

1st condition assumption: This is the right form of the model

Gauss-Markov assumptions: $E(\varepsilon_i) = 0$

$Var(\varepsilon_i) = \sigma^2$ all i

$\varepsilon_i, \varepsilon_j$ uncorrelated, $i \neq j$

ε_i's are Normally distributed.

Residuals $\hat{\varepsilon}_i = y_i - b_0 - b_1 x_i = y_i - \hat{y}_i$

What do we know about $\hat{\varepsilon}_i$?

$\rightarrow E(\hat{\varepsilon}_i) = 0$ since $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

Now

$$\sum \hat{\varepsilon}_i = 0, \quad \sum \hat{\varepsilon}_i x_i = 0, \quad \sum \hat{\varepsilon}_i \hat{y}_i = 0$$

$\sum \hat{\varepsilon}_i = 0$ because $\bar{\hat{\varepsilon}}_i = 0$

$\hat{\varepsilon}_i$ as estimators of ε_i

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

$\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ Calc" on p.61 of text

- high leverage points have residuals with lower variance

I'll show this later using matrices

$\hat{\varepsilon}_i$'s not uncorrelated (know $\sum \hat{\varepsilon}_i = 0$)

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

$\hat{\varepsilon}_i$'s are linear combinations of Y_i's so they are normally distributed

They can be more closely to normally distributed then the $\hat{e}_i$'s because of the CLT

If no leverage points, h_{ii} is small, variances of $\hat{e}_i$'s close, lack of equal variance sometimes ignored.

To adjust for unequal variance, standardize

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

SAS call this "student"

Studentized

This is internal studentization.

Sometimes people use external studentization.

$$\frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

$s_{(i)}$: s.e. of $\hat{e}_i$ with i^{th} observation removed

externally studentized residuals
have a t distribution

SAS calls externally studentized
residuals "rstudent"

END OF TEST COVERAGE