



\_

## **Outliers in Regression**

A data point far from the majority of the data (in y and/or x) may be called an outlier, especially if it does not follow the general trend of the rest of the data.

- Data points that are far from the mean of the x's are called *leverage points*.
- A data point that is far from the mean of either or both the x's and/or the y's are *influential points* if they influence the fit of the regression.
- An outlier may or may not impact the regression fit significantly, thus it may or may not be an influential point.

The upshot: Sometimes there are good reasons for excluding subsets (there were errors in the data entry; there were errors in the experiment).

Sometimes - the outlier belongs in the data. Outliers should always be examined.

Georgia Tech

3

## **Checking for Outliers**

Look at the standardized residuals:

$$r_i^* = \frac{y_i - \hat{y}_i}{\sqrt{MSE}}$$

Compare the standardized residuals to the -2 to +2 band (or -1 to +1).

- Standardized residuals bigger than 1 are large.
- Standardized residuals bigger than 2 extremely large.

Most statistics packages will calculate these automatically.



4

## Coefficient of Determination

A statistic that efficiently summarizes how well the X's can be used to predict Y is the R-square:

$$R^2 = 1 - SSE/SST$$

which is interpreted as:

SSE = 
$$\sum_{i=1}^{n} r_i^2$$
  
SST = 
$$\sum_{i=1}^{n} (y_i - \overline{y})^2$$

R<sup>2</sup> = Proportion of total variability in Y that can be explained by the regression (that uses X)

> Georgia Tech

5

## Correlation Coefficient

A statistic that efficiently summarizes how well the **X's** are linearly related to **Y** is the correlation coefficient:

$$\rho = \operatorname{cor}(X, Y) = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}} \sqrt{S_{YY}}} = \hat{\beta}_1 \sqrt{\frac{S_{XX}}{S_{YY}}}$$

Correlation coefficient and coefficient of variation:

$$\rho^2 = R^2$$

Georgia Tech

6

