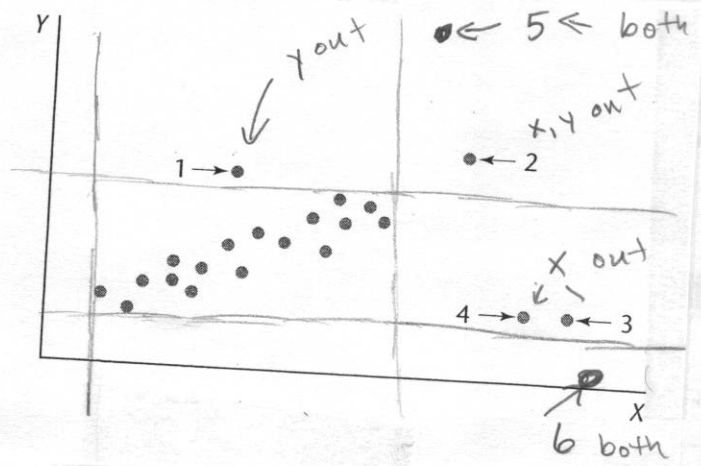
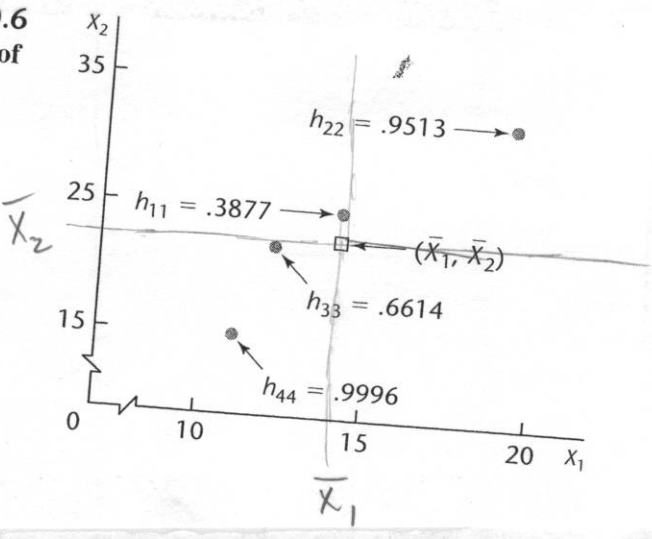


**FIGURE 10.5**  
Scatter Plot for  
Regression  
with One  
Predictor  
Variable  
Illustrating  
Outlying  
Cases.



**FIGURE 10.6**  
Illustration of  
Leverage  
Values as  
Distance  
Measures—  
Table 10.2  
Example.



I.  $Y$ -outlier

A. Deleted residual: Suppose we took out the 5th observation of a data set. If we refit the model with the remaining  $n-1$  observations and calculated  $\hat{Y}$  at the  $X$  levels of observation 5 we would have  $\hat{Y}_{5(5)}$ . The "deleted" residual would be:

$d_5 = y_5 - \hat{Y}_{5(5)}$

$\uparrow$   
means without obs 5.

B. If we divide  $d_5$  by its approx standard error, we have the studentized deleted residual for obs 5.

C. A test could be conducted for a particular observation

$H_0$ : The observation is not an outlier.

$H_A$ : The observation is an outlier

The critical value(s) would be  $t(1 - \frac{\alpha}{2n}, n-p-1)$

$p$  is the number of parameters.

SAS calls these values 'RSTUDENT.'

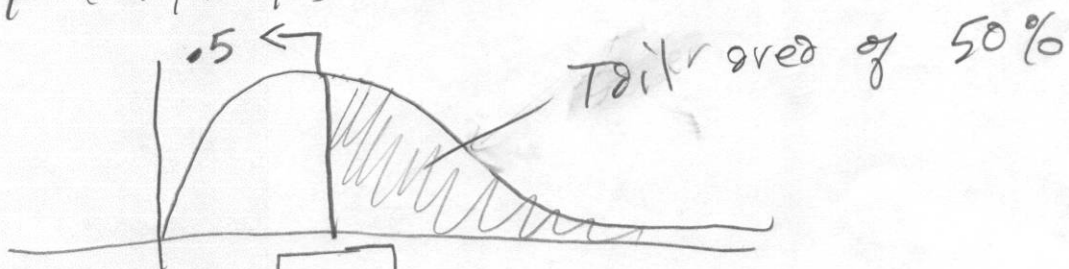
## II. X-Outlier

A. Hot Leverage values: The hot or leverage stat for an observation is a metric measuring the distance a particular observation's X values are from the centroid of a data set. It can be shown that the average hot value is  $\frac{p}{n}$ . The yardstick for an influential observation is a hot value greater than  $\frac{2p}{n}$ .

## III. X and/or Y outlier

### A. Cook's D.

Cook's D is considered an aggregate outlier statistic. There is no official 'official' rule for what value is large enough to flag an outlier. Some suggest using the  $F(n, n-p)$  distribution



A Cook's value greater than the 50th percentile would be considered influential.

## IV Influence on Prediction

A. DF FITS: Measures the influence of an observation on the fitted equation.

A guideline for an influential observation for DF FITS is a value greater than 1 for small to medium data sets and  $2\sqrt{\frac{p}{n}}$  for very large data sets.

## V Influence on Betas

A. DF Betas: A measure of the influence of the  $i$ th observation on each of the betas.

A guideline for influence is a value greater than 1 for small to medium data sets and  $\frac{2}{\sqrt{n}}$  for large data sets.

## VI. Variance Inflation. Extreme correlation among X values.

A. A great amount of correlation among predictors themselves can cause the variance to inflate for particular betas. Other problems can also arise.

2-4-10

(4)

VIF or Variance Inflation Factor is a measure of the magnitude that variation increases as a result of high correlation among predictors in a model. If the greatest VIF for a set of betas is greater than 10 or the average of all of the VIF's is much greater than 1, correlation among the  $X$ 's is likely causing problems.