Presentation Number: 545

Graziose

Title: "Locating Outliers and Influential Points using Regression Analysis and Technology"

- Definition: An **outlier** is an extreme value falling far from most of the data values in the data set.
- It is important to identify and consider outliers because they can have a strong affect on the mean and standard deviation.

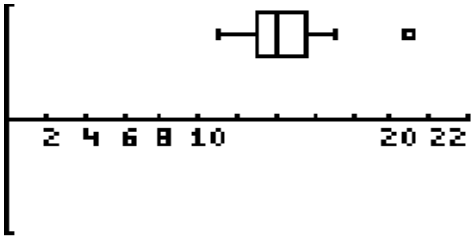
In a **modified box-plot**, a data value is an outlier if the data value,

is above Q_3 by an amount greater than $1.5 \times IQR$

is below Q_1 by an amount greater than $1.5 \times IQR$

where $IQR = Q_3 - Q_1$

Modified Box Plot: (Data from table 1)



Linear Regression Line:

- **Definition:** Given a collection of paired sample data, the **regression line** (Line of best fit, or least squares line) is the straight line that "best" fits the scatter plot of the data.

Regression equation: $\hat{y} = a + bx$
where x is the independent variable and
$\hat{y}$ is the dependent variable.

$$a = \frac{\left(\sum x\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

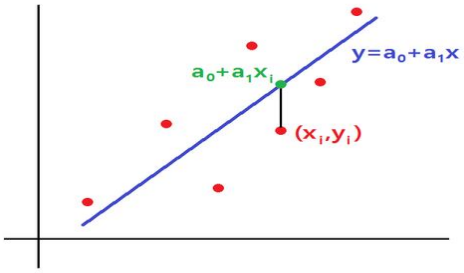
Matrix (Regression) Projection:

Projection Matrix: $P = A(A^T A)^{-1} A^T$

Fit the data points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ with the best line.

We need to minimize the errors between the data points and the line.

The error e_i is the vertical distance from the point to the line.



The linear system in matrix form: $Ax = b$

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$$

Notice that we have 3 equations and 2 unknowns this linear system has no solutions.

Since, we cannot solve the matrix equation $Ax = b$.

We need to solve: $A^T A x^* = A^T b$ for x^*.

$A^T A x^* = A^T b$ (Projection equation)

$x^* = (A^T A)^{-1} A^T b$

$P = Ax^* = A(A^T A)^{-1} A^T b$

The projection matrix: $P = A(A^T A)^{-1} A^T$

Now $x^* = (A^T A)^{-1} A^T b$,

Therefore $\begin{bmatrix} C \\ D \end{bmatrix} = (A^T A)^{-1} A^T b.$

and $x^* = \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 2/3 \\ 1/2 \end{bmatrix}.$

Cook's Distance:

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

where $MSE = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}$.

p is the number of fitted parameters in the model;
MSE is the mean square error of the regression model;
h_{ii} is the leverage;
e_i is the residual.