**Bivariate Data**

**In this lesson, we are going to study association between two numerical or quantiative variables.**

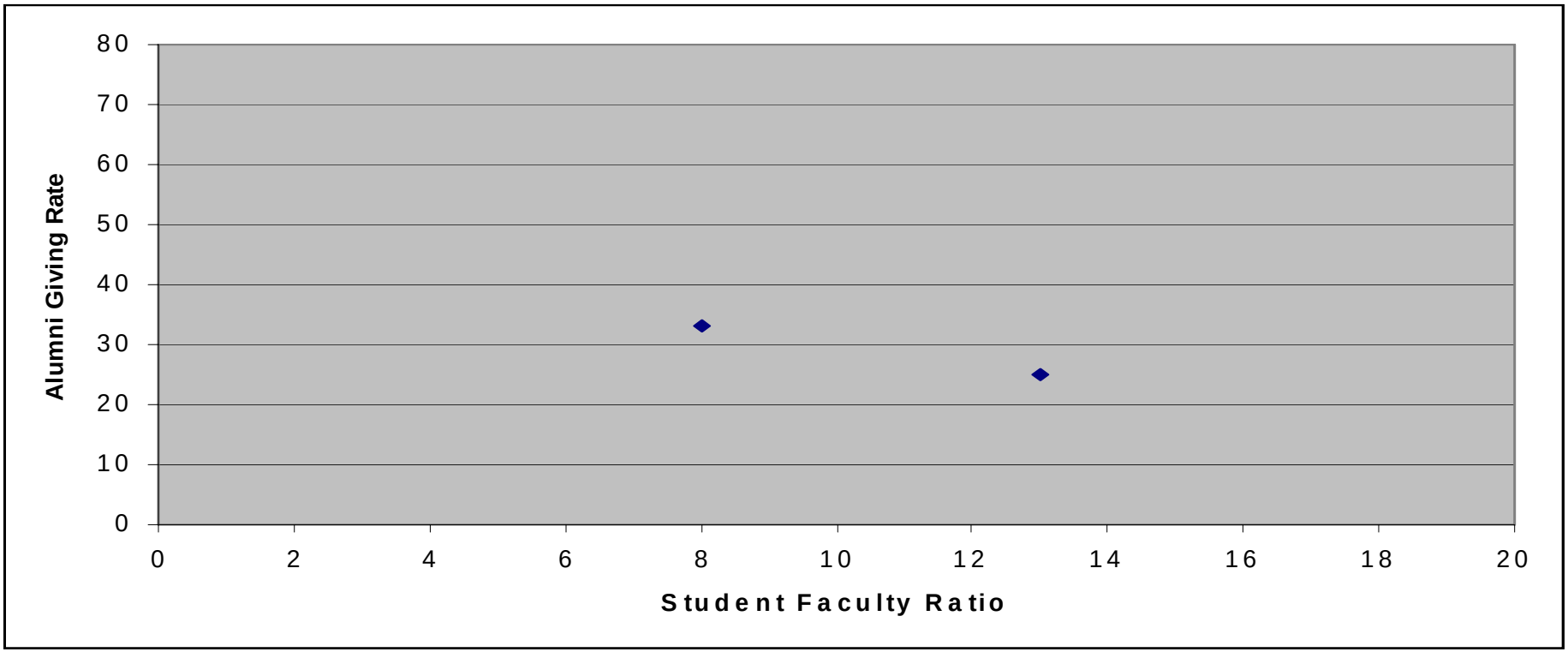
**Example 1:**

**The data file "Alumni" shows "the student faculy ratio", "% of classes with class size under 20" and "the giving rate of the alumni" for 48 universities**

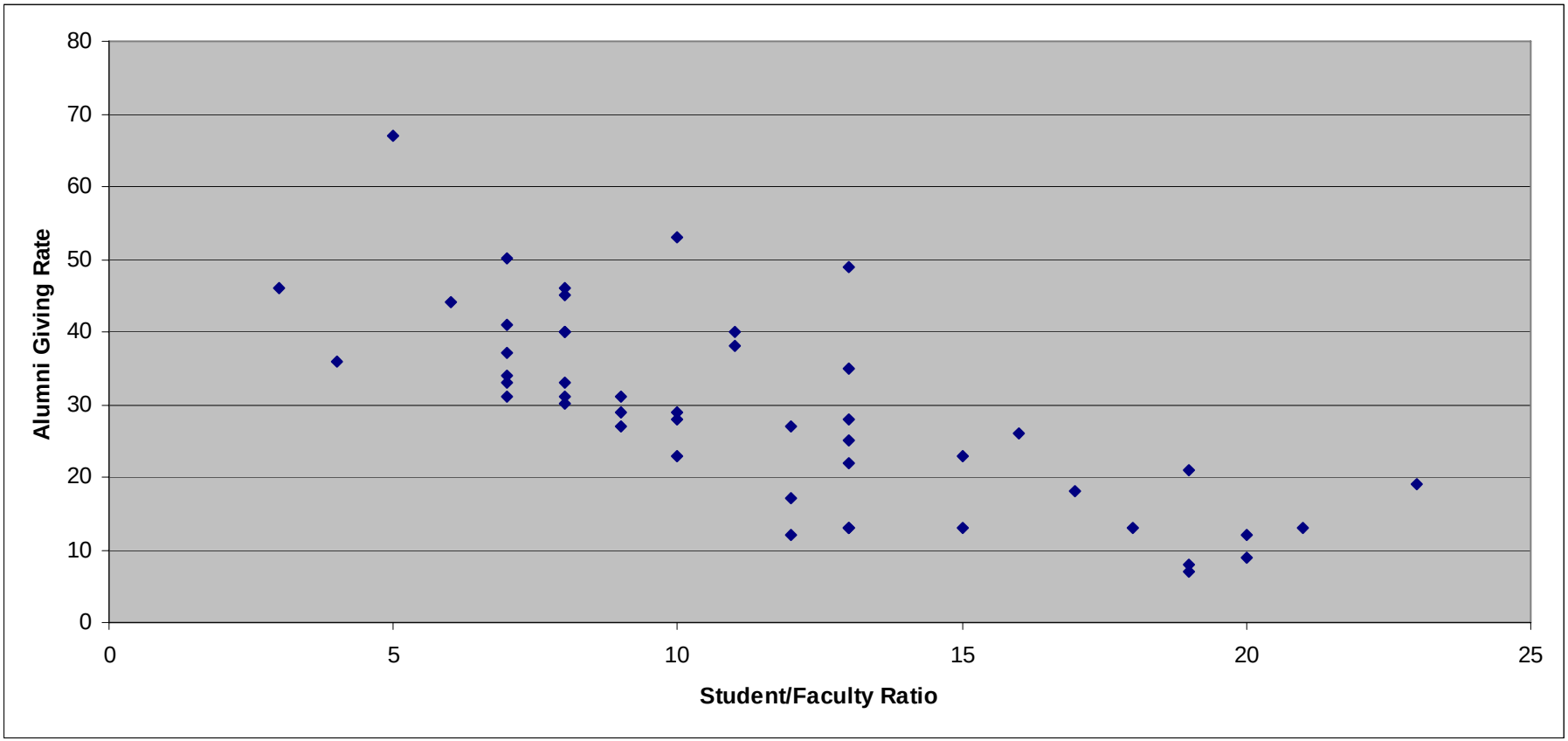
**Let us see how does "the giving rate of the alumni" respond to "the student faculy ratio." In this case, we are calling "the student faculy ratio" as the explnanatory variable and "the giving rate of the alumni" as response variable. Sometimes, an explanatory variable may be referred to as an independent variable and the response variable as independent variable.**

**We shall treat each point as an ordered pair, for example, for Boston College, the student faculty ratio is 13 and the alumni giving rate is 25. We plot the point (13,25) in a coordinate plane.**

**For Brandeis University,the student faculty ratio is 8, and the alumni giving rate is 33. We plot the point (8,33) in a coordinate plane.**

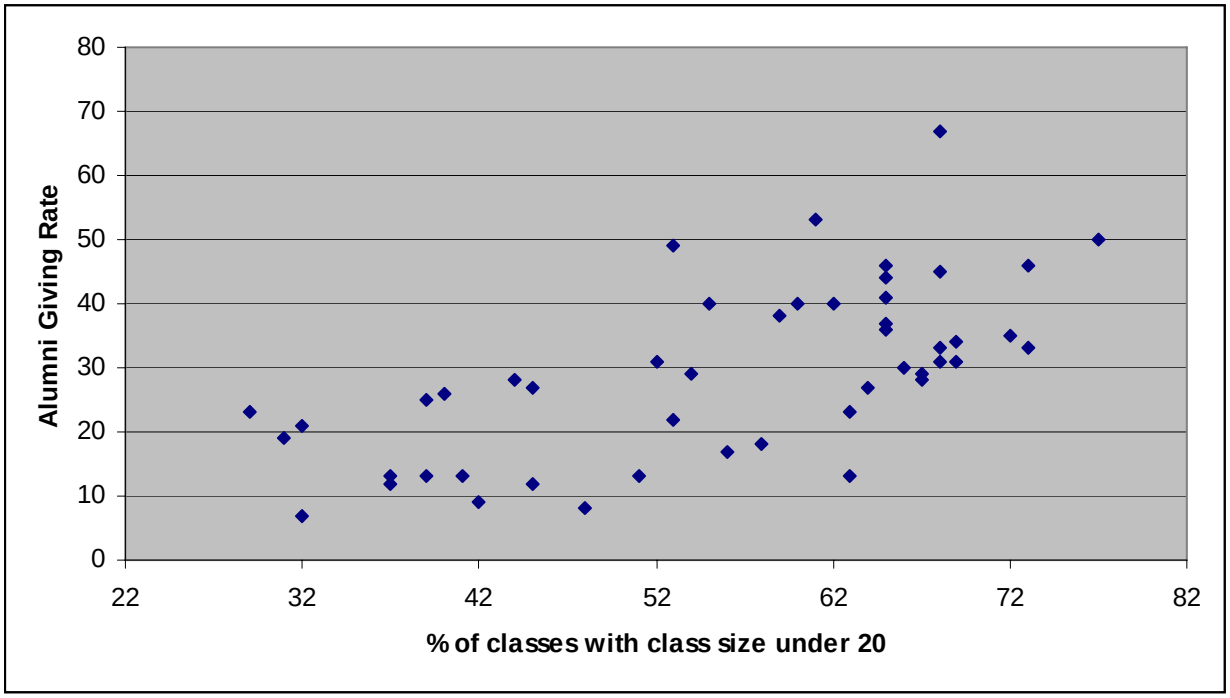



**We may complete plotting the data for all the universities as shown below**




**Note that the larger values of the variable "the student faculy ratio" are accompanied by the smaller values of the variable "the giving rate of the alumni"**
**therefore the above association is a negative association.**

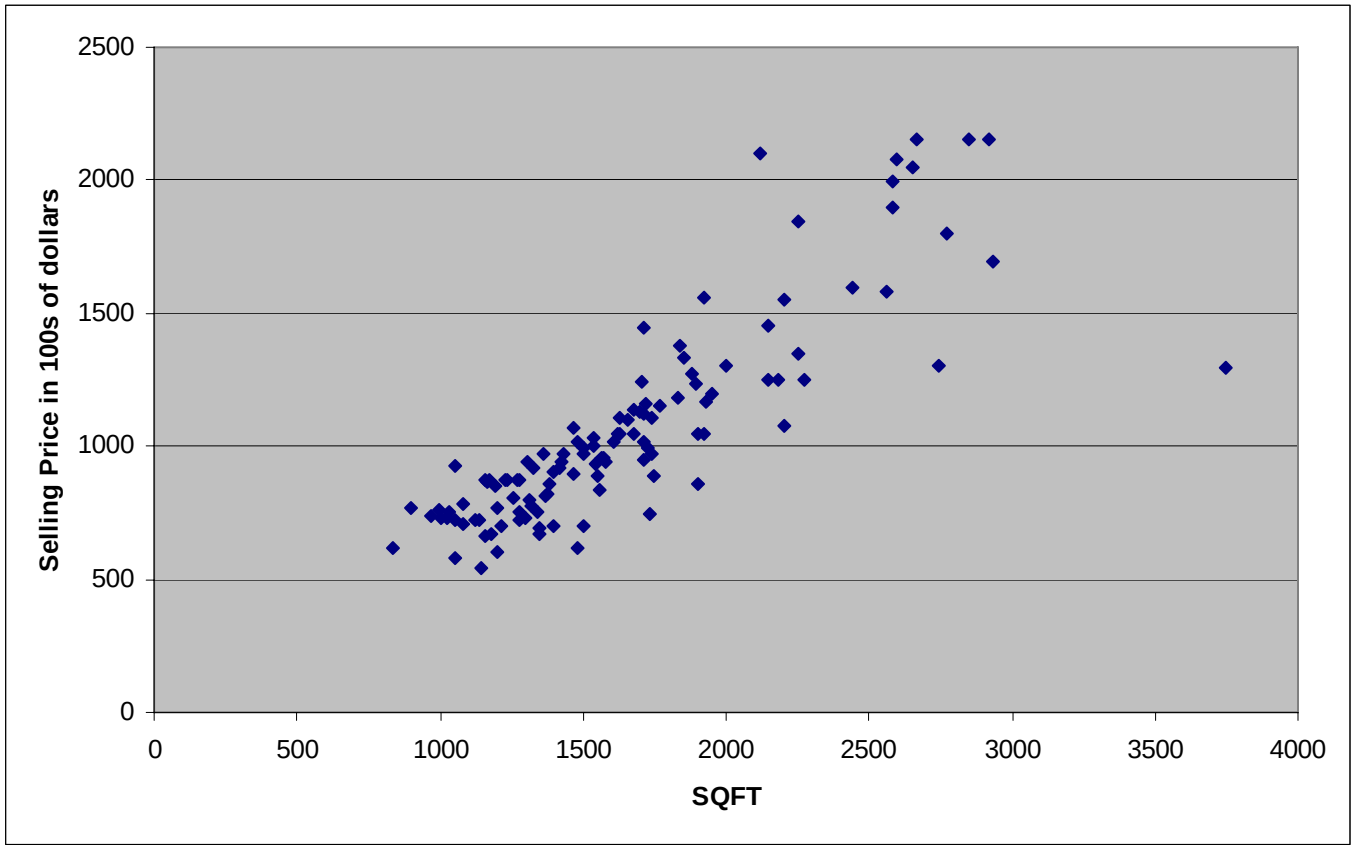
**On the other hand if we plot "% of classes with class size under 20" as the explnatory and "the giving rate of the alumni" as response, the plot looks like**




**which shows a positive association between the two variables. That is more classes the university has with small class size, higher the alumni giving rate is.**

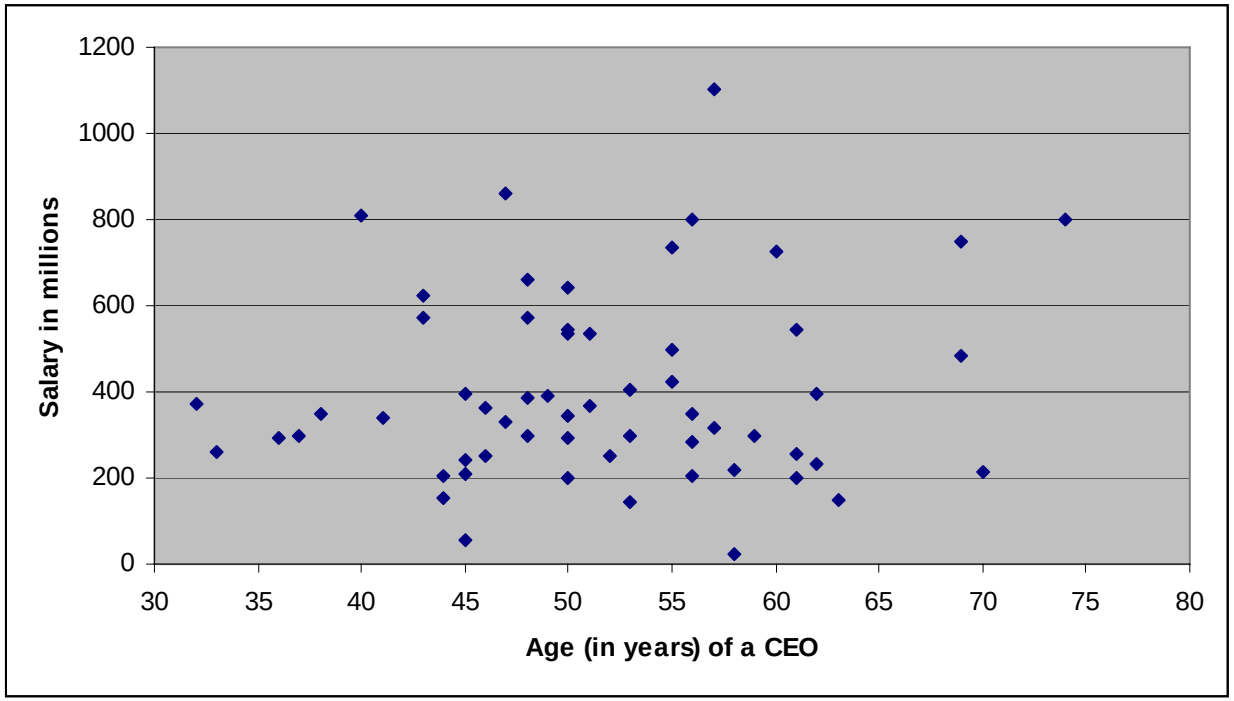
**Example 2**:

**The following scatterplot shows a positive association between the square feet of living space in a house and its price in 100s of dollars in 1993 for a sample taken in Albuquerque. The data are given in the data file "homeprices."**




**Example 3**:

**Refer to the data file CEO salaries, if we look at the association between the age of a CEO and the salary of the CEO, we hardly see any association.**



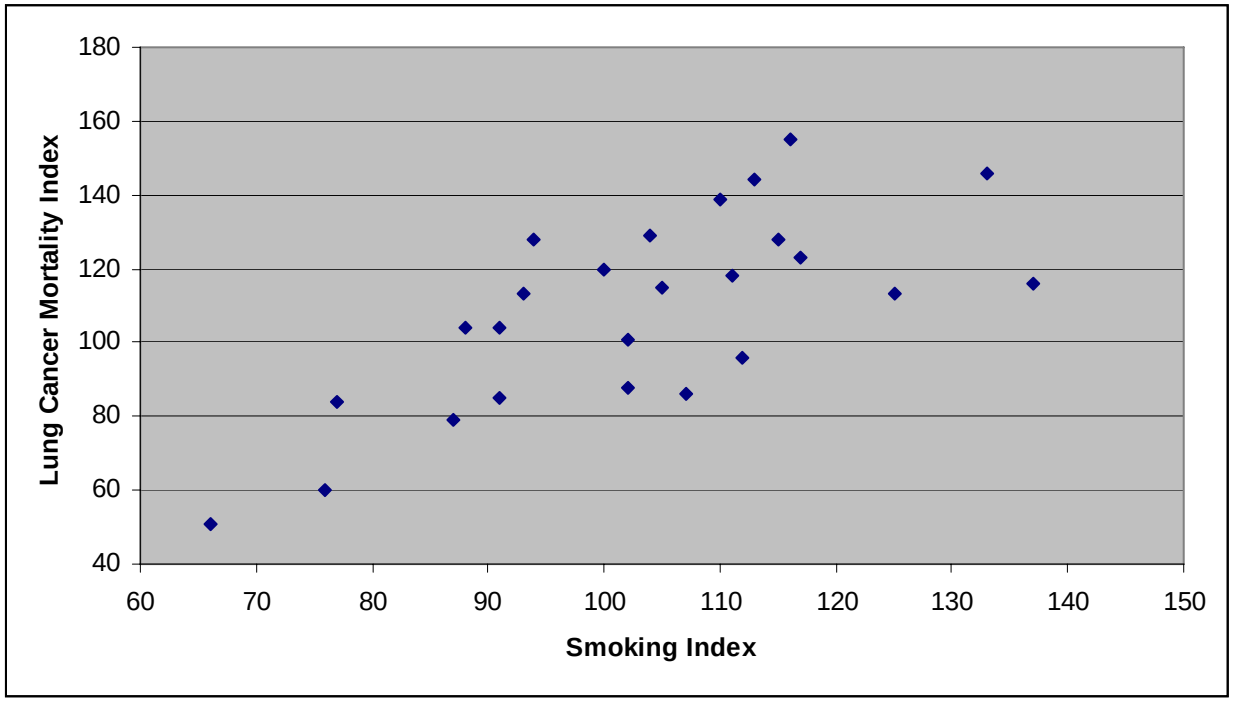

**Example 4**:

**The following data is from Chance Data and Story Library, according to the website**

**"Data summarizes a study of men in 25 occupational groups in England. Two indices are presented for each occupational group. The smoking index is the ratio of the average number of cigarettes smoked per day by men in the particular occupational group to the average number of cigarettes smoked per day by all men. The mortality index is the ratio of the rate of deaths from lung cancer among men in the particular occupational group to the rate of deaths from lung cancer among all men. "**

| Smoking Index | Lung Cancer Mortality Index |
|---|---|
| 77 | 84 |
| 137 | 116 |
| 117 | 123 |
| 94 | 128 |
| 116 | 155 |
| 102 | 101 |
| 111 | 118 |
| 93 | 113 |
| 88 | 104 |
| 102 | 88 |
| 91 | 104 |
| 104 | 129 |
| 107 | 86 |
| 112 | 96 |
| 113 | 144 |
| 110 | 139 |
| 125 | 113 |
| 133 | 146 |
| 115 | 128 |
| 105 | 115 |
| 87 | 79 |
| 91 | 85 |
| 100 | 120 |
| 76 | 60 |
| 66 | 51 |

**If we look at a scatterplot**




**We see a positive association**.

**Now that we have seen THREE different examples of positive assiciation**, **we are interested in looking at a numerical mesaure of the strength of such an association**. **Of course**, **you can get a fairly good idea of the association by looking at the scatterplot but COMBINING the scatterplt with numerical summaries will be helpful in computations**.

**WARNING**: **I am using an approach that looks different but is equivalent to the approach that is given in the text book**.

**I shall post the descritptive statistics of the textbook exercises to use this approach**

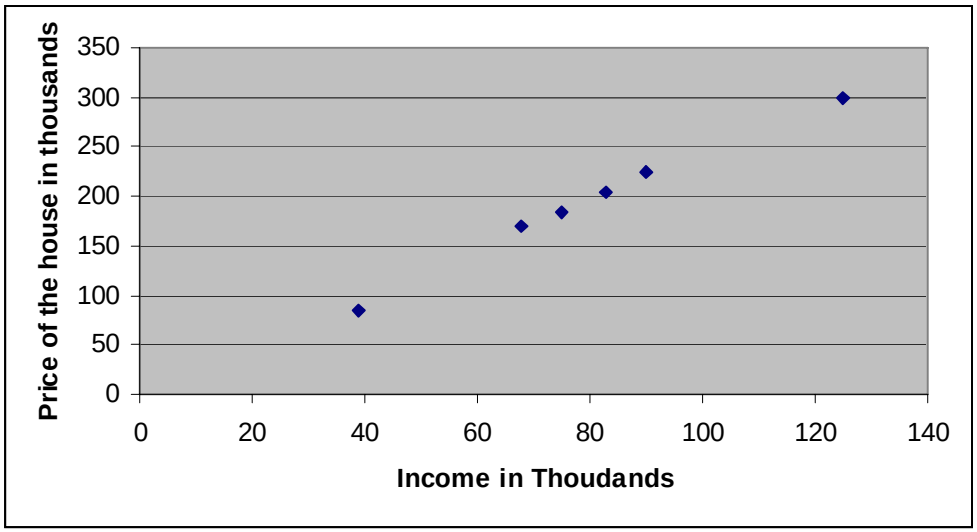
**Let us look at a formula approach of finding the strength of a linear association and also a formula approach to find an equation of the "least square regression line" that is the line that fits the scatter plot the best.**

**Example 5**:

**John is a business major and would like to study the association between the family income and the price of the house the family lives in. He finds the following from a random sample of 6 homeowners.**

| Family Income in Thousands of US Dollars | Price of the house in Thousands of US Dollars |
|---|---|
| 39 | 85 |
| 68 | 170 |
| 75 | 185 |
| 83 | 205 |
| 90 | 225 |
| 125 | 300 |

**A careful look at the table shows that more people earn, more they pay to buy a house. This will show more clearly if we make a scatterplot plot these values as ordered pairs.**




**The picture clearly shows the positive association that we had noted above**.

**Correlation**:

**To measure the numerical strength of a linear association we can use the correlation coefficient r which can be proven to equal the following expression**.

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\left(n-1\right)s_x s_y}$$

**$\bar{x}$ : The mean of the x-values**

**$\bar{y}$ : The mean of the y-values**

**$s_x$ : The standard deviation of the x-values**

**$s_y$ : The standard deviation of the y-values**

**$n$ : The number of observations in the sample**

**If the data consitutes a population instead of a sample, we use n instead of $(n-1)$ in the denominator.**

**For the above example (income vs price of the house) suppose that we have calculated the following**

| Variable | N | Mean | Median | TrMean | StDev |
|---|---|---|---|---|---|
| Income | 6 | 80.0 | 79.0 | 80.0 | 28.2 |
| Homeprice | 6 | 195.0 | 195.0 | 195.0 | 70.5 |

**Now complete the table**

| x | y | (x-80) | (y-195) | (x-80)(y-195) | |
|---|---|---|---|---|---|
| 39 | 85 | -41 | -110 | 4510 | |
| 68 | 170 | -12 | -25 | 300 | |
| 75 | 185 | -5 | -10 | 50 | |
| 83 | 205 | 3 | 10 | 30 | |
| 90 | 225 | 10 | 30 | 300 | |
| 125 | 300 | 45 | 105 | 4725 | |
| | | | | 9915 | SUM |

$$r = \frac{9915}{(6-1)\times 28.2\times 70.5} \rightarrow r \cong 0.997$$

**About this correlation coefficient r**, **note that**

Notes about r:

1. r measures the strength of association of a linear relationship.
2. $-1 \le r \le 1$
3. r does not have any units.
4. r is affected by outliers.

**Finding the Line of Best Fit**:

**The line of best fit is the line that describes the straight line relationship between the two variables the best**, **as shown in class**, **it is the line for which the sum of the squares of the differences between an observed value**

**and corresponding predicted value is the smallest**. **It can be shown that the line is given by**
**y=a+bx**
**where**

$$b = r\frac{s_y}{s_x} \text{ and } a = \bar{y} - b\bar{x}$$

For the income verses home price data

$$b = .997\frac{70.5}{28.2} \rightarrow b \cong 2.4925$$

**and**

$$a = \bar{y} - b\bar{x} = 195 - 2.4925 \times 80$$

$$\rightarrow a = -4.4$$

therefore an equation of the line of best fit is
y=–4.4+2.4925x

**Example 6**:

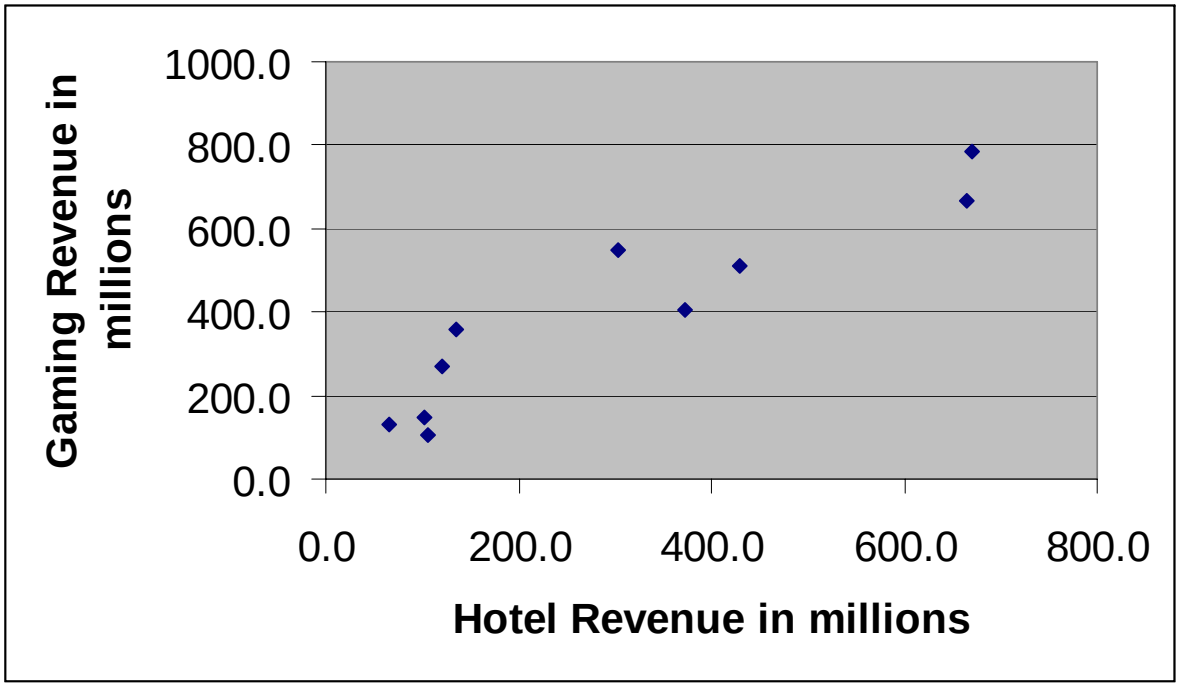
**The following data is from Cornell Hotel and Restaurant Administration Quarterly**, **October 1997 for 10 Los Vegas Casino Hotels**.

| Company | Hotel Revenue ($M) | Gaming Revenue ($M) |
|---|---|---|
| Boyd Gaming | 303.5 | 548.2 |
| Circus Circus Enerprises | 664.8 | 664.8 |
| Grand Casinos | 121.0 | 270.7 |
| Hilton Corp. Gaming Div | 429.6 | 511.0 |
| MGM Grand, Inc. | 373.1 | 404.7 |
| Mirage Resorts | 670.9 | 782.8 |
| Primadonna Resorts | 66.4 | 130.7 |
| Rio Hotel & Casino | 105.8 | 105.5 |
| Sahara Gaming | 102.4 | 148.7 |
| Station Casinos | 135.8 | 358.5 |

**Let us examine the association between the hotel revenue and the gaming revenue with the hotel revenue as the explanatory variable and the gaming revenue as response. Following show a descriptive statistics of these two variables. H and G denote the hotel and the gaming revenues respectively.**

```
Variable          N       Mean     Median     TrMean      StDev    SE Mean
H                10      297.3      219.7      279.5      231.7       73.3
G                10      392.6      381.6      379.7      233.6       73.9
correlation of H and G = 0.931
```

**First step to do is to make a scatter plot of the two variables. A scatter plot is**




**This shows a positive association between the two variables. If we want to find the straight line that represents this association the best, namely the regression line, the following is one way of doing that.**

**Equation is** $\hat{y} = a + bx$ **, we use the notation y for observed value and** $\hat{y}$ **for the value predicted by the line. where**

**b** $= \mathbf{r}\frac{s_y}{s_x}$ **a** $= \overline{y} - \mathbf{b}\overline{x}$

**r: correlation**

$\overline{x}, \overline{y}$ **are the means of the x and the y values respectively.**

$s_x, s_y$ **are the standard deviations of the x and y values respectively.**

**Therefore for this example, we have**

$\mathbf{b} = \mathbf{0.931}\frac{233.6}{231.7} \cong \mathbf{0.9386}$

$\mathbf{a} = \mathbf{392.6} - \mathbf{0.9386}(297.3) \cong \mathbf{113.5542}$

**Equation is** $\hat{y} = 113.5542 + 0.9386x$

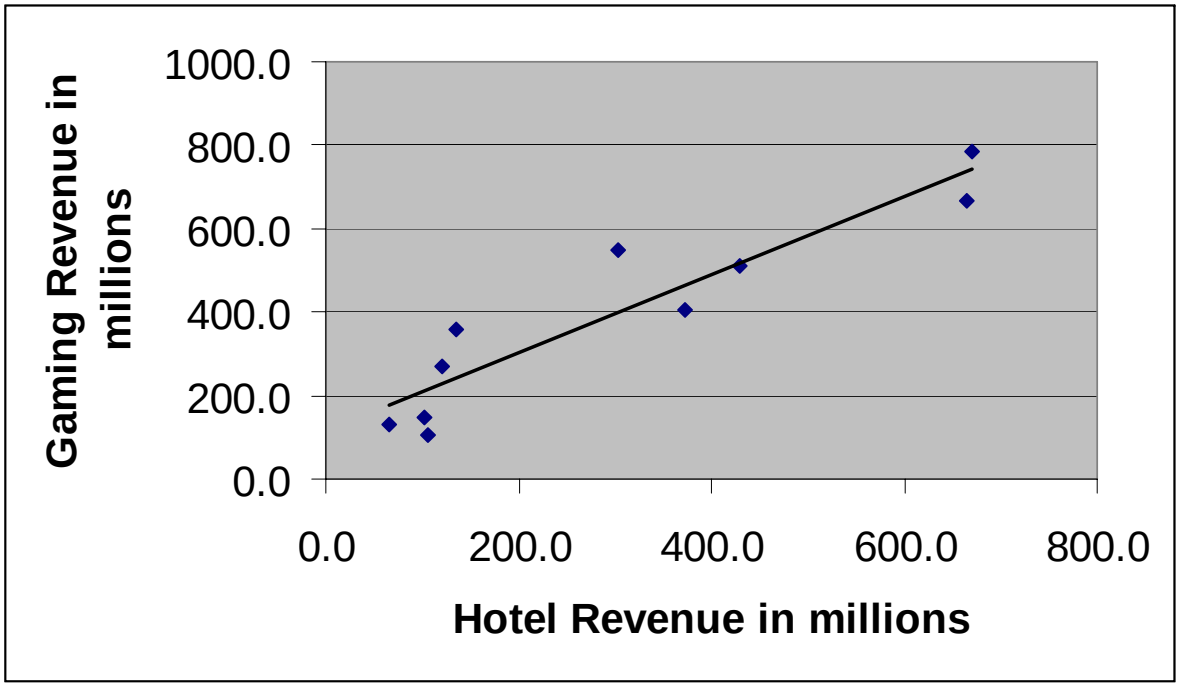
**We use this line to predict the values of response for a given value of the explanatory variable**.

**For example**:

**If x= 200**, $\hat{y} = 113.5542 + 0.9386(200) = 301.2742$

**If x=600**, $\hat{y} = 113.5542 + 0.9386(600) = 676.7142$

**We may use the pair (200,301.2742) and (600, 676.7142) to graph the line in the scatter plot**.




**Note that when we go according to the line we understand how predicted y (the predicted value) varies or in other words this variation is explained by the line.**

**This value is numerically given by** $r^2 = 0.931^2$
**or** $0.866761$

**that is approximately 86.68% of the variation in y is explained by this line.**

**For a particular point in the data set, the residual is the difference between the observed value and the predicted value.**

**For example, for the value (670.9,782.8) we may note that**
**the predicted value is** $\hat{y} = 113.5542 + 0.9386(670.9) = 743.26094$

**Therefore the residual for this point is** $782.8 - 743.26 = 39.54$

**It is clear from the graph that the point (373.1,404.7) for the MGM has the largest residual. These numbers become more interesting when who have the entire context in front of you.**

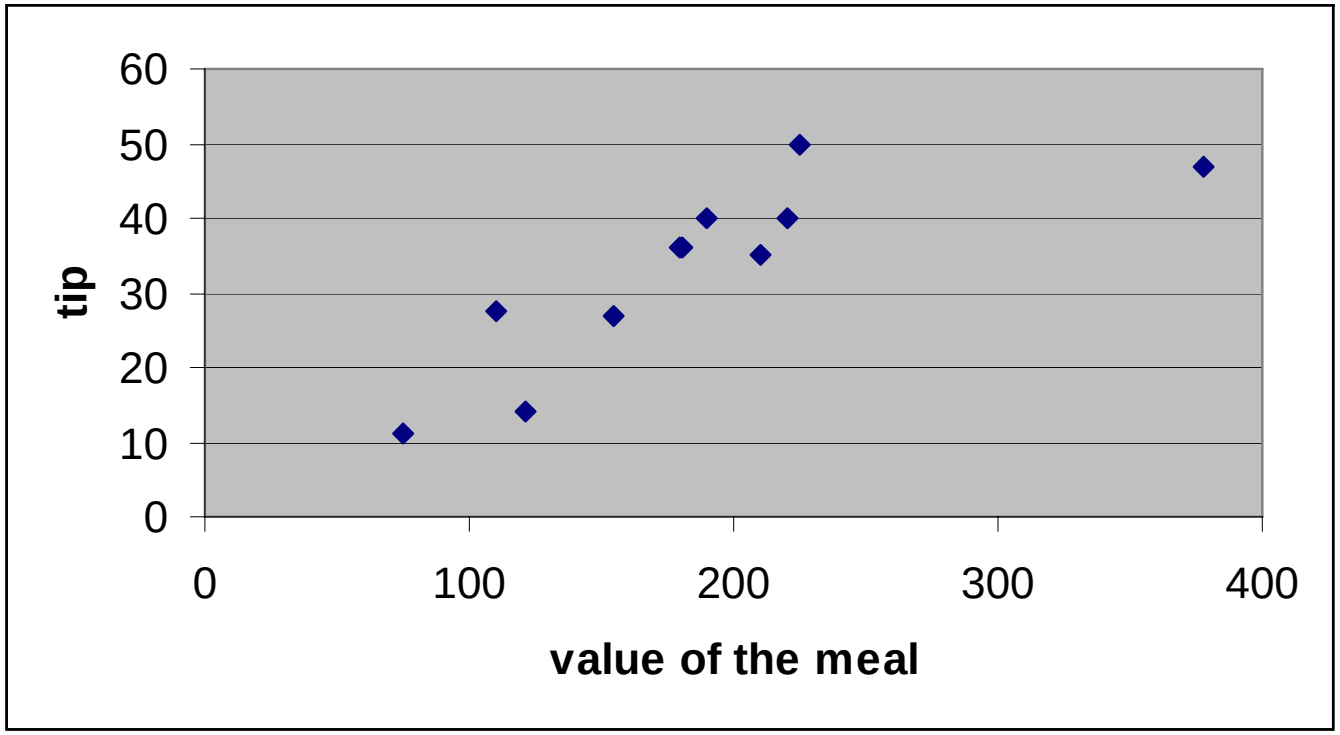
---

**Let us look at the following example to see the (visual) identification of an influential observation.**

**Example 7.**
**Suppose that the following data shows the relation between the value of meal and tip left at an upscale restaurant. Both the amounts are in US dollars.**

| meal | tip |
|---|---|
| 181 | 36 |
| 220 | 40 |
| 110 | 27.5 |
| 190 | 40 |
| 210 | 35 |
| 75 | 11.25 |
| 180 | 36 |
| 155 | 27 |
| 378 | 47 |
| 121 | 14 |
| 225 | 50 |

**a scatter plot is**




**with meal as explanatory and tip as response. The two values are positively associated and have the following summary statistics**

| Variable | N | Mean | Median | TrMean | StDev | SE Mean |
|---|---|---|---|---|---|---|
| meal | 11 | 185.9 | 181.0 | 176.9 | 79.7 | 24.0 |
| tip | 11 | 33.07 | 36.00 | 33.61 | 12.27 | 3.70 |

correlation of meal and tip = 0.807

**Remember that equation of the line of best fit or the regression line is given by**

$\hat{y} = \mathbf{a} + \mathbf{bx}$

$\mathbf{b} = \mathbf{r}\frac{s_y}{s_x}$ $\mathbf{a} = \overline{y} - \mathbf{b}\overline{x}$

**r**: **correlation**

$\overline{x}, \overline{y}$ **are the means of the x and the y values respectively**.

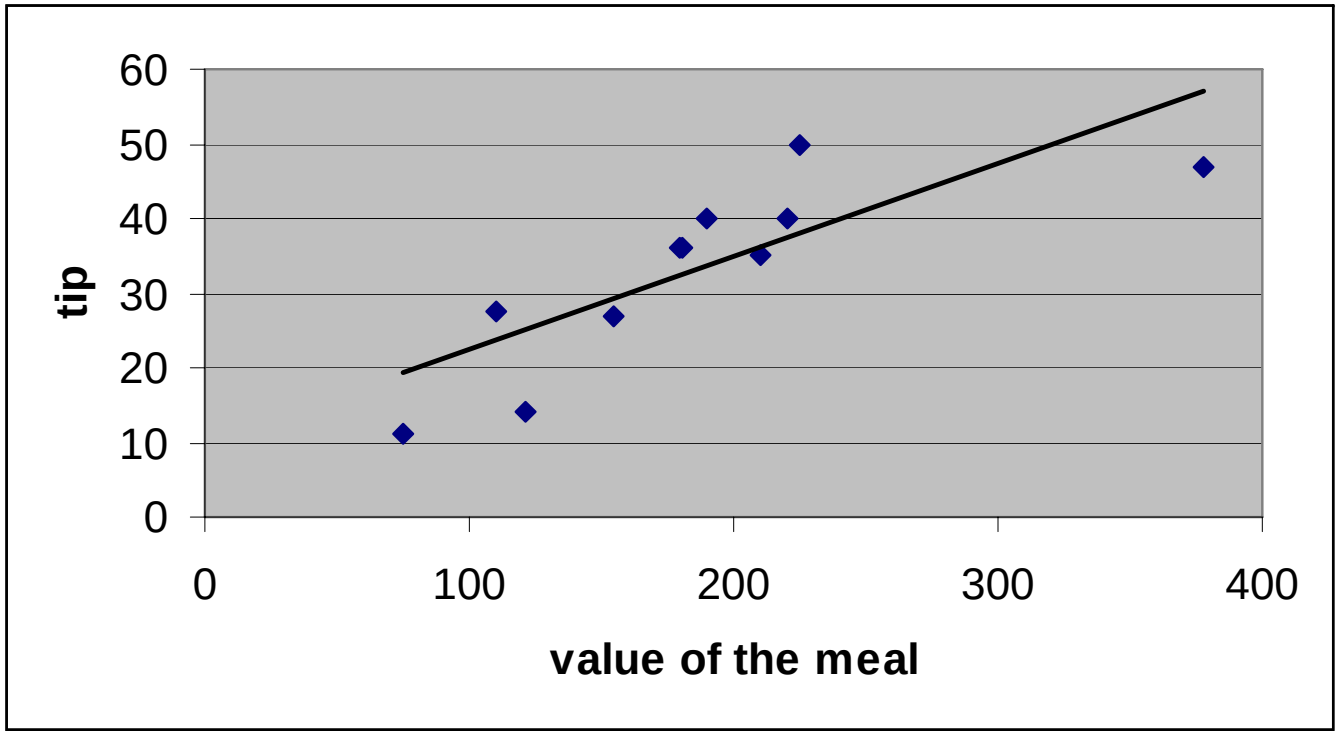
**For the line of best fit**, **we compute**

**b**$= .807\frac{12.27}{79.7} \cong 0.1242$

**a**$= 33.07 - 0.1242 \times 185.9 = 9.9812$

**Equation**: $\hat{y} = 9.9812 + 0.1242x$

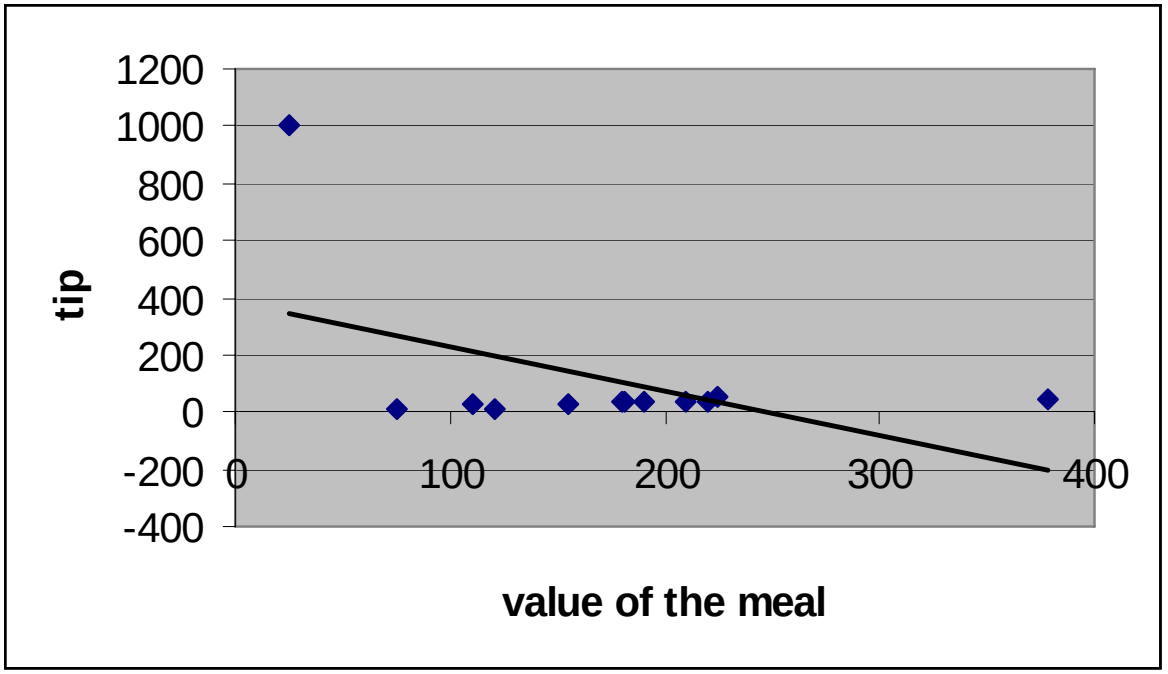
**Plot the equation in the scatter plot**




**Note that the point (378,47) is pulling the line towards it. Such a point is called an influential observation. There are many diagnostic techniques for identifying the influential observations, packages like MINITAB identify the points of heavy leverage or influence in the output.**

**Try recalculating the regression line by removing the above point, you shall a remarkable shift it in the line towards the other points.**

**To see how much effect can an outlier have on the regression line, let us put a point (25,5000) in the data. Note that**




**A point that has such a heavy leverage on the regression line is called an influential observation**.

Usually, influential observations are outliers in the x-direction.