# EIN 5332

# Correlation

## Chapter 23 Sections 23.1-4

Karen E. Schmahl Ph.D., P.E.

Univariate Data – during sampling, one variable is collected

Bivariate Data – during sampling, two variables are collected with each observation

- Ordered pairs of data

  For every observation i, there will be an x and a y
- Is there a relationship between x and y?

# Bivariate Data

Examples

- % additive in material, material strength
- Cost of living, average salaries in area
- Exercise amount, weight loss
- Temperature, energy usage

# Scatter Plot or Diagram

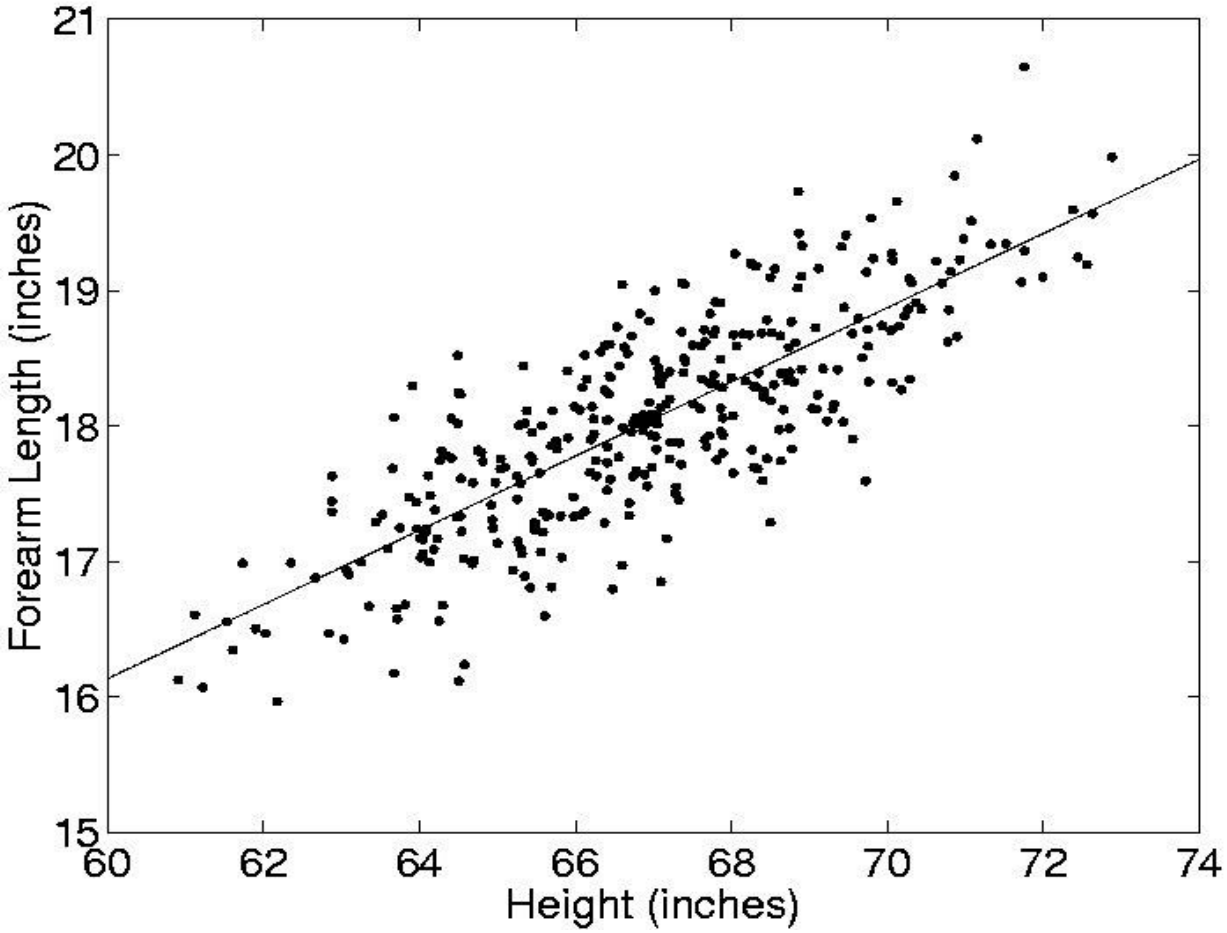
A graph of plotted points that show the relationship between ordered pairs of data.

- Used to identify any potential relationship between the variables.
- Distribution of data on the diagram often indicates what type of relationship may exist.

# 5.19 Scatter Diagram (Plot of Two Variables)

- A scatter diagram (plot) assesses the relationship between two variables
  - 50 to 100 pairs of samples should be plotted
  - the independent variable is on the x-axis while the dependent variable is on the y-axis.
- The correlation and regression techniques can be used to test the statistical significance of relationships. (Ch. 23)

# 5.19 Scatter Diagram (Plot of Two Variables)



# 5.19 Scatter Diagram (Plot of Two Variables)

- A scatter diagram relationship does not predict a true cause-and-effect relationship.
- The plot only shows the strength of the relationship between two variables

# Correlation Analysis

## Pearson Correlation Coefficient – r

- Reflects the degree of linear relationship between two variables.

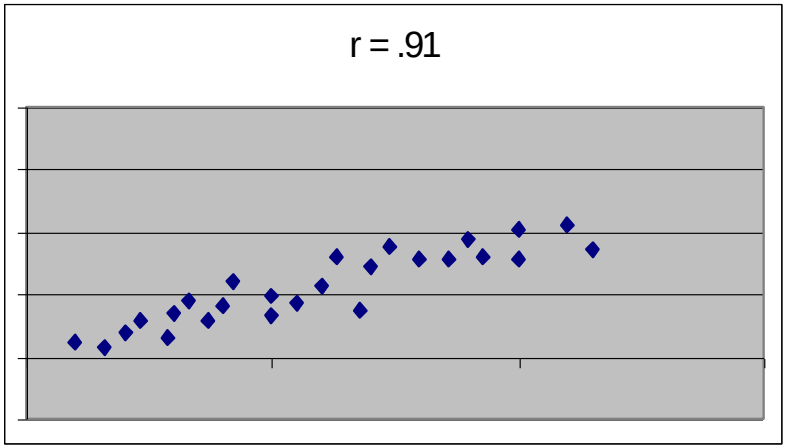
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n} y_i - \bar{y})^2}}$$

# Correlation Analysis

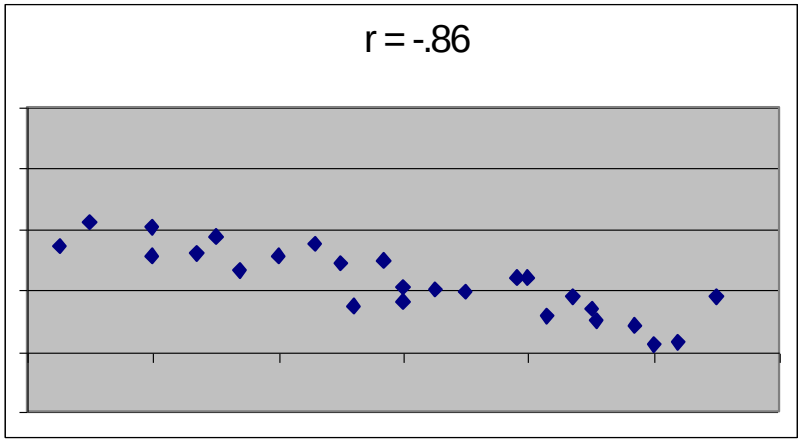
## Pearson Correlation Coefficient – r

- Reflects the degree of linear relationship between two variables.
- Ranges from -1 to +1
- The closer to | 1 | the stronger the relationship
- Positive correlation - both variables increase or decrease together,
- Negative correlation - as one variable increases, the other decreases.

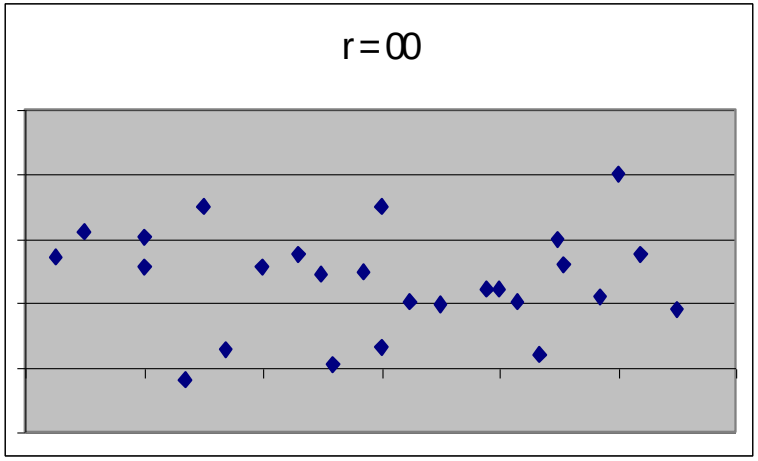
## Positive Correlation



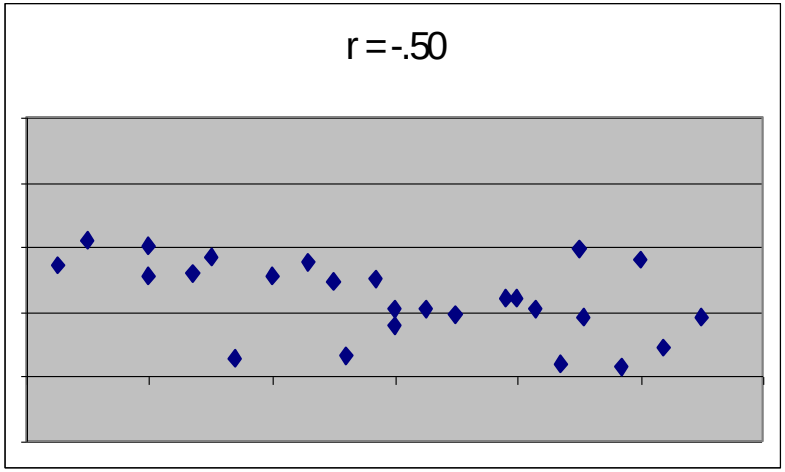

## Negative Correlation




## No Correlation




## ? Correlation



# Examples

Which is the most likely correlation to occur ?

Weight of a car (x) vs. miles per gallon (y)

A. Positive B. Negative C. No correlation

Miami summer outside temperature (x) vs. power level generated at the power plant

A. Positive B. Negative C. No correlation

Number of alcoholic drinks (x) vs. score on a dexterity test (y)

A. Positive B. Negative C. No correlation

# Correlation Analysis

Pearson Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n} y_i - \bar{y})^2}}$$

- Based on deviations from the means
- r is a ratio – unitless

# Correlation Analysis

## Pearson Correlation Coefficient – equivalent formula

$$r = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

Set up data in corresponding pairs in table.

Calculate elements of formula: $\Sigma x$, $\Sigma x^2$, $\Sigma y$, $\Sigma y^2$, $\Sigma xy$

Plug into formula

In an experiment to determine corrosion rates, steel plates were placed in a solution of 10% hydrochloric acid for various periods of time (in hours) and the weight loss (in mg) was measured. The data collected is in the box below. How strong is the relationship between the time and weight loss?

| Time | Loss | | | |
|---|---|---|---|---|
| x | y | xy | x^2 | y^2 |
| 2 | 0.7 | 1.4 | 4 | 0.49 |
| 4 | 6.9 | 27.6 | 16 | 47.61 |
| 6 | 7.2 | 43.2 | 36 | 51.84 |
| 8 | 11.9 | 95.2 | 64 | 141.61 |
| 10 | 16.2 | 162 | 100 | 262.44 |
| 12 | 22.5 | 270 | 144 | 506.25 |

| ∑x | ∑ y | ∑ xy | ∑ x^2 | ∑ y^2 |
|---|---|---|---|---|
| 42 | 65.4 | 599.4 | 364 | 1010.24 |

$$r = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

Plug the sums below into the formula to calculate the correlation coefficient.

| Time x | Loss y | xy | x^2 | y^2 |
|---|---|---|---|---|
| 2 | 0.7 | 1.4 | 4 | 0.49 |
| 4 | 6.9 | 27.6 | 16 | 47.61 |
| 6 | 7.2 | 43.2 | 36 | 51.84 |
| 8 | 11.9 | 95.2 | 64 | 141.61 |
| 10 | 16.2 | 162 | 100 | 262.44 |
| 12 | 22.5 | 270 | 144 | 506.25 |
| ∑x | ∑ y | ∑ xy | ∑ x^2 | ∑ y^2 |
| 42 | 65.4 | 599.4 | 364 | 1010.24 |

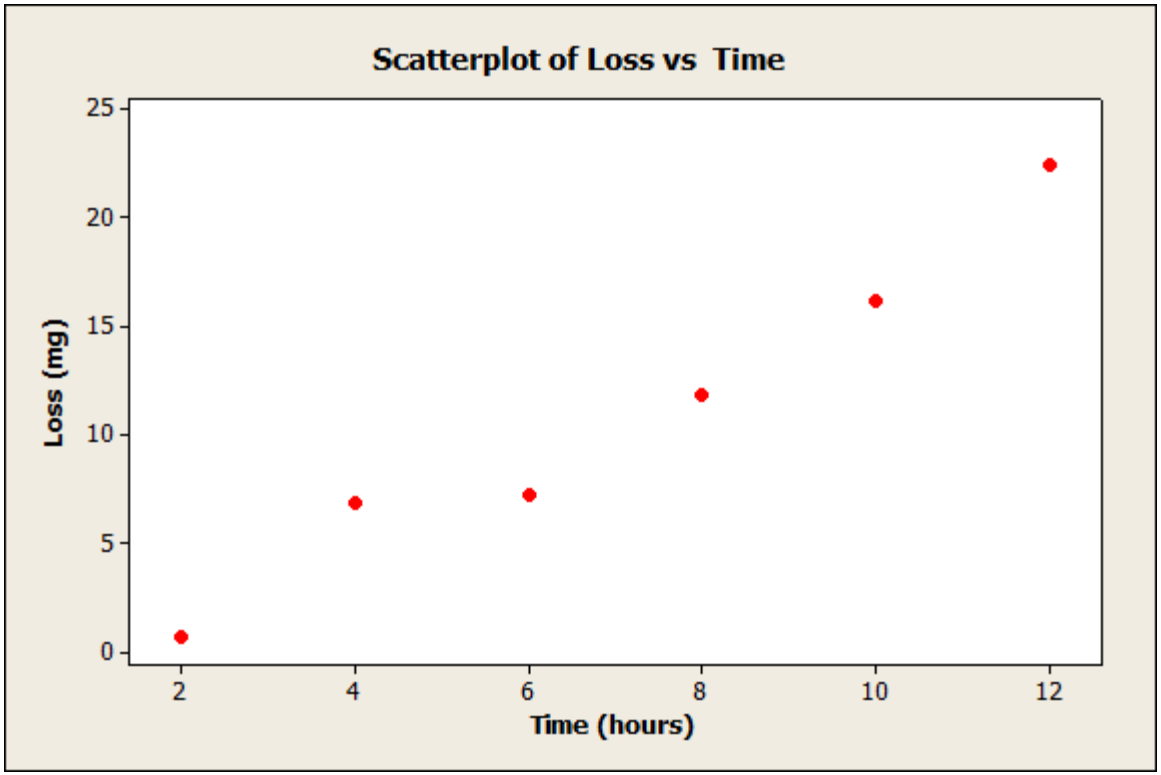
The correlation coefficient r is

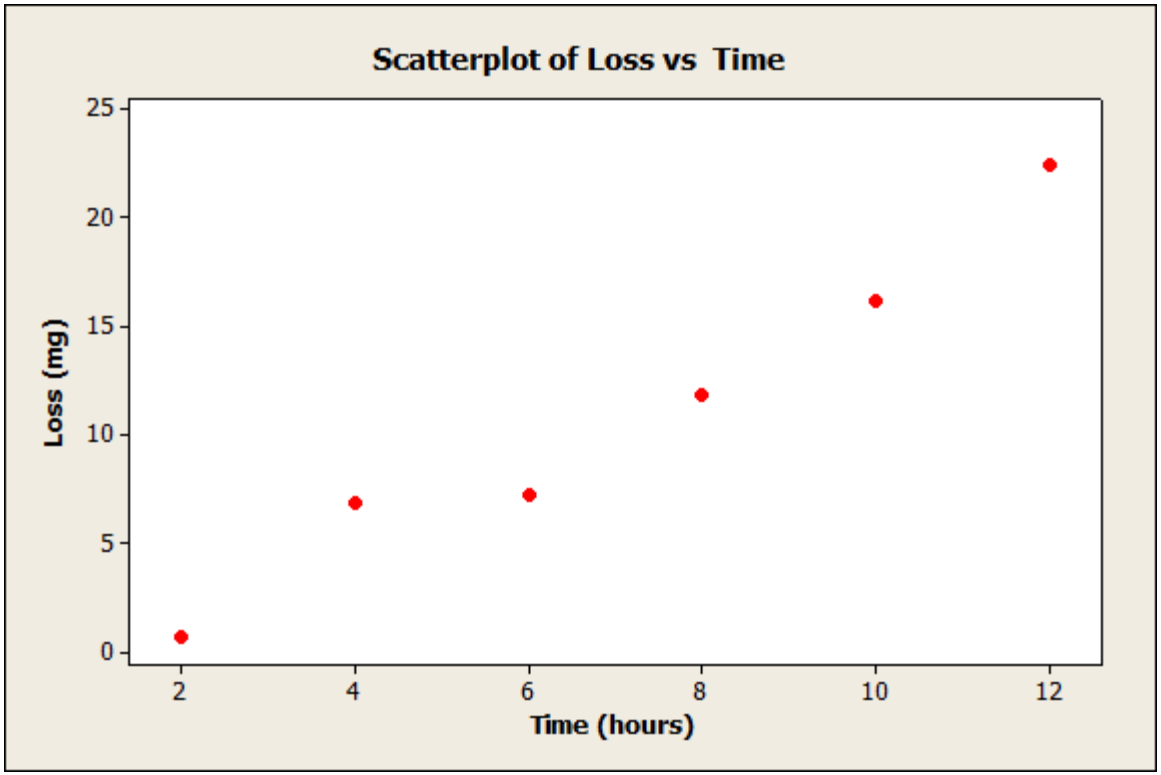
A. +0.98

B. +0.87

C. -0.65

D. +1.05


Scatterplot of Loss vs Time
Loss (mg)
25
20
15
10
5
0
2
4
6
8
10
12
Time (hours)




Which of the following statements is true?

A. Time is the dependent variable and loss is the independent variable.

B. Time is the independent variable and loss is the dependent variable .

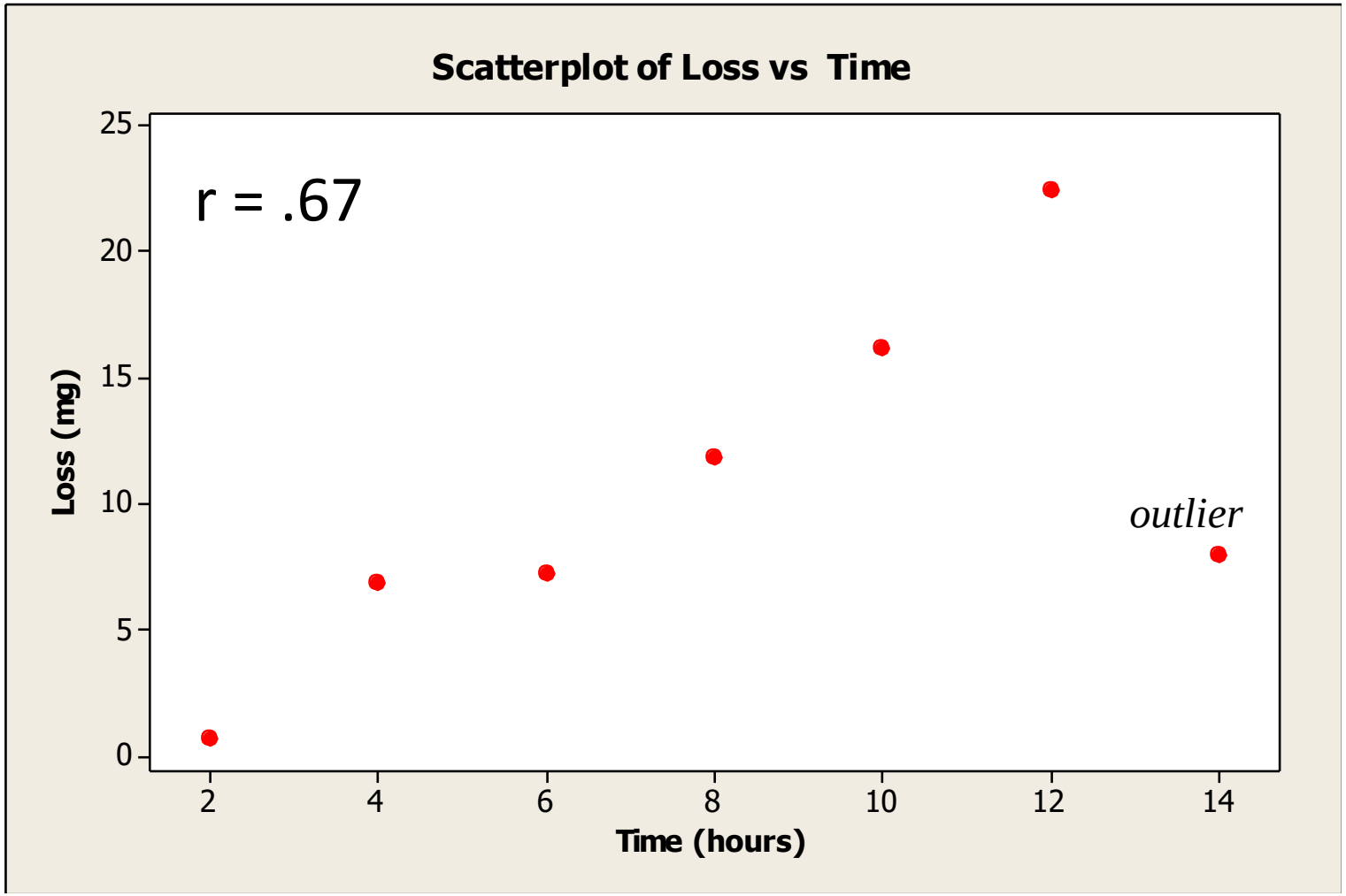
# Correlation - Cautions

- Watch for the impact of outliers

- Correlation coefficient is only for linear relationships

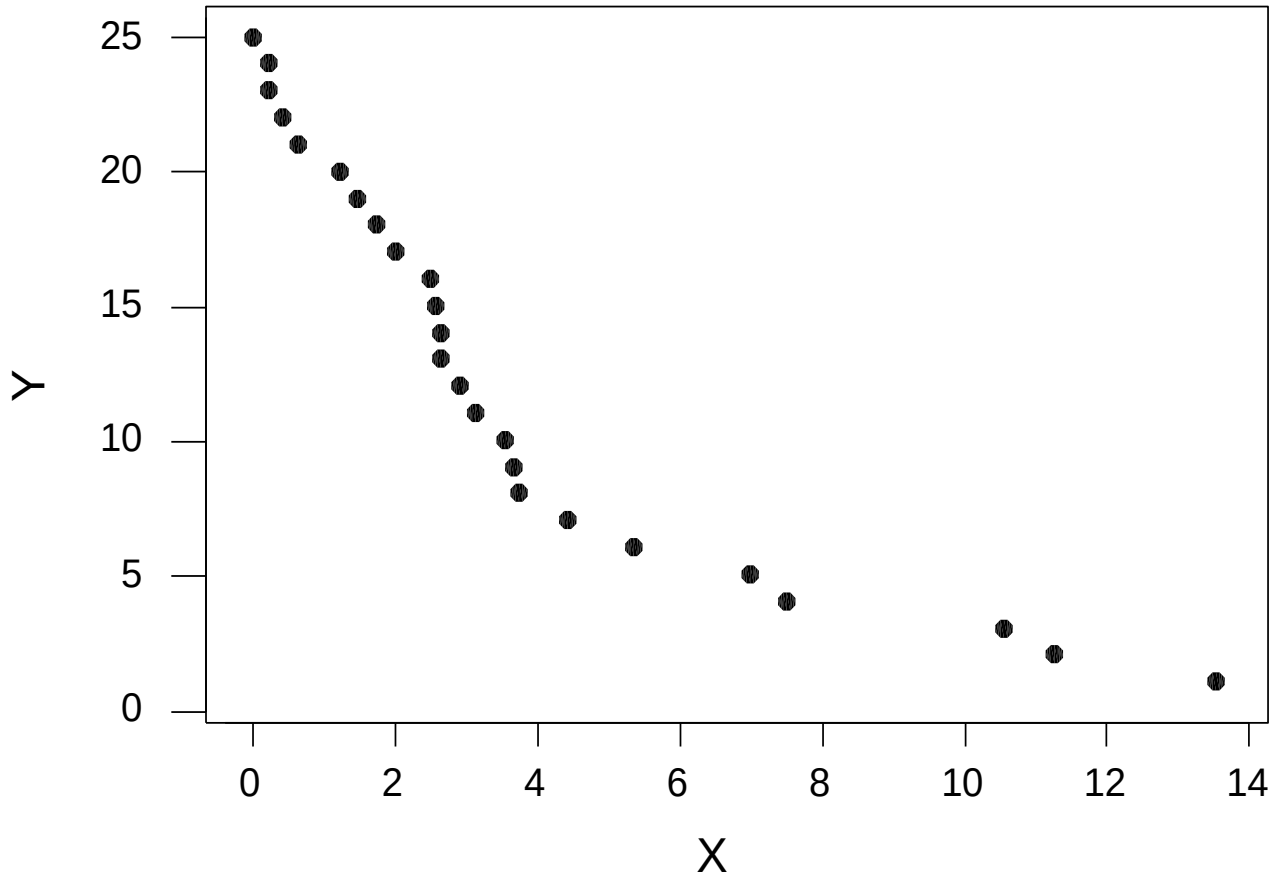
- Relationship may exist for only limited conditions or for a subset of the data

# Correlation - Cautions



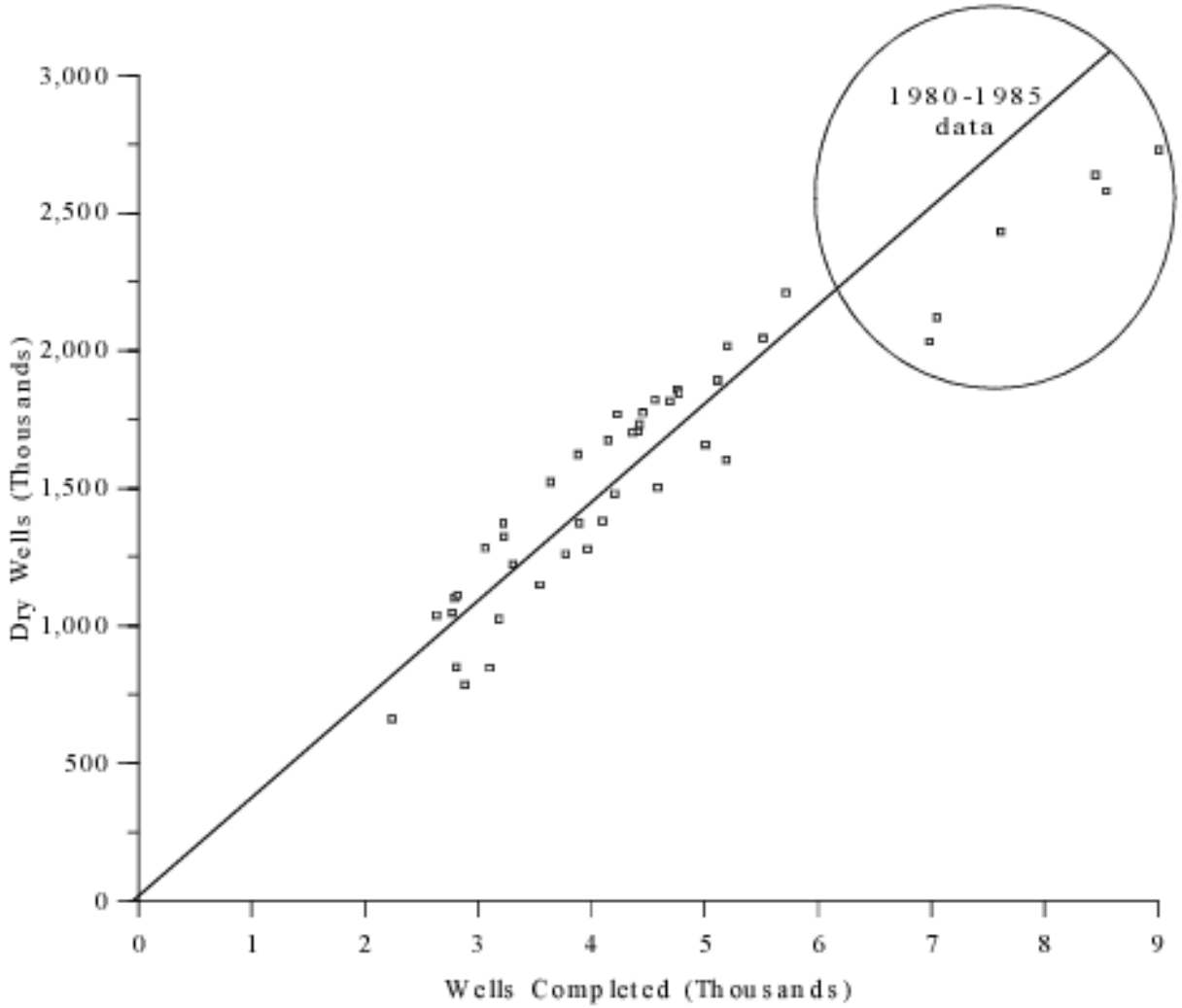

Watch for the impact of outliers

# Correlation - Cautions




Correlation coefficient is for linear relationships

## Oil and Gas Exploratory and Development Wells, 1949 to 1992




http://www.eia.doe.gov/neic/graphs/scatter.htm

# Correlation vs Causation

- Correlation tells us if the values of two variables are statistically related

- The variable themselves may or may not be related. We could have
  - *X causes Y*
  - *Y causes X*
  - *The X* and *Y* relationship is caused by something else

Video - Correlation versus Cause and Effect-(10 minutes)

## Confounding

Third variable is correlated to both variables of interest, but there is no cause and effect between the variables of interest.

## When can we imply causation?

When controlled experiments are performed to verify the causes.

http://onlinestatbook.com/stat_sim/reg_by_eye/index.html

Go to the link above to access a simulation about the correlation coefficient. You can use the simulation to get a better feel for what the data would look like for different values of the correlation coefficient.

After you click on the link, you may have to wait for the "begin" button to appear. Press begin and a new window will open with the simulation.

A scatter plot is shown. Guess the r value, selecting one of the values off to the right.

To check you guess, press "show r". How did you do?

To try another data set, press "New Data"

# Related Assignments

Please see  Blackboard for related assignments