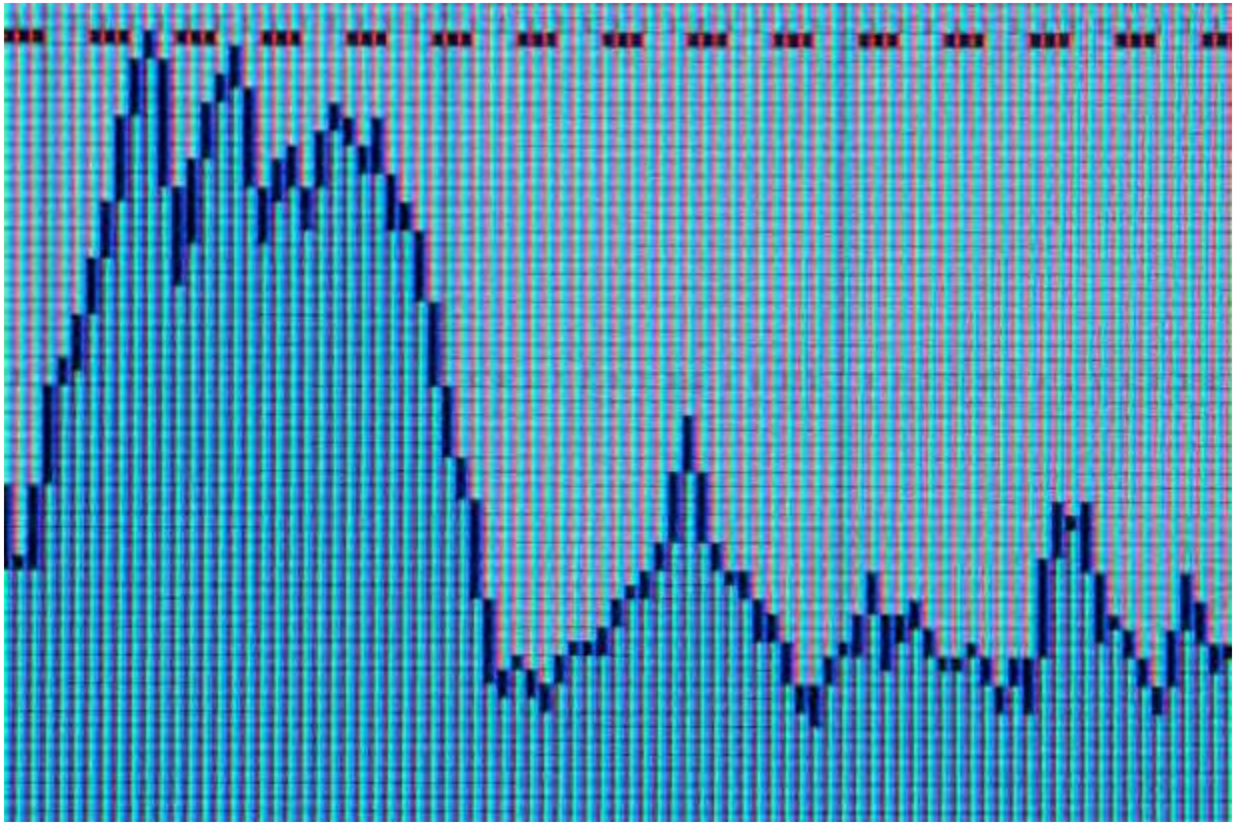
# Introductory Biostatistics
# PBIO 504

## Correlation
## and
## Linear Regression

**Textbook Chapters 17-18**

# Correlation



Here we see a relationship across time. It can be Dow Jones index or an exchange rate, etc.

# Introduction

- When we are dealing with two continuous variables, such as children's age and height, there are several ways to look at the data and determine their statistical relationship.
- This relationship might or might not be cause-and-effect, and **correlation** is a valid approach to explore if there is a **linear relationship**.
- When the causal relationship is known or strongly suspected, we can use **regression** to predict levels of one variable, given the values of the other variable.

# Correlation is about Linear Relationships

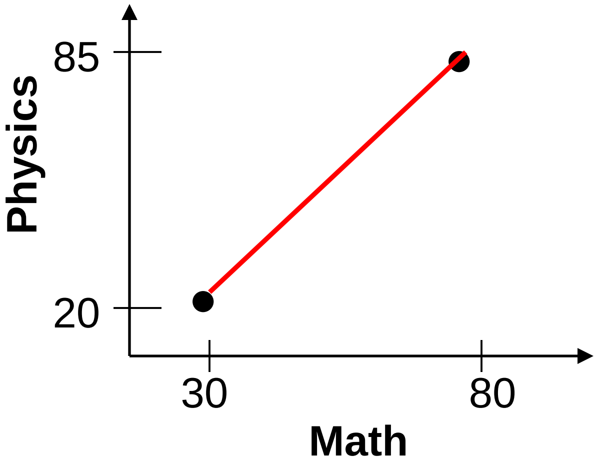
Consider the following two sets of data

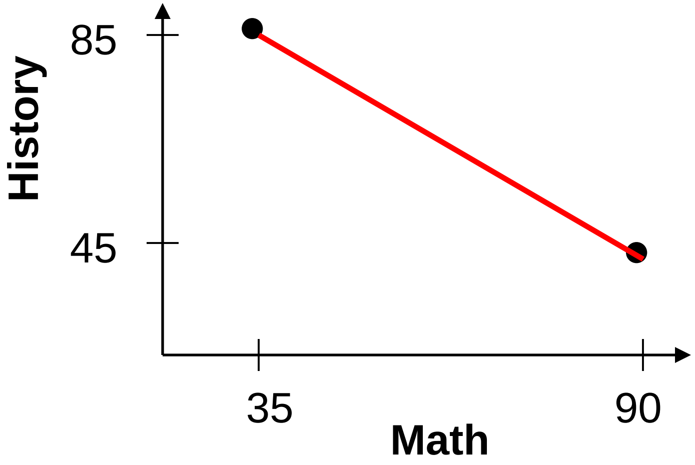
| Student | Math | Physics |
|---|---|---|
| 1 | 30 | 20 |
| 2 | 80 | 85 |

Data set A

| Student | Math | History |
|---|---|---|
| 1 | 35 | 85 |
| 2 | 90 | 45 |

Data set B





What is the relationship between math and physics grades? Between math and history grades?

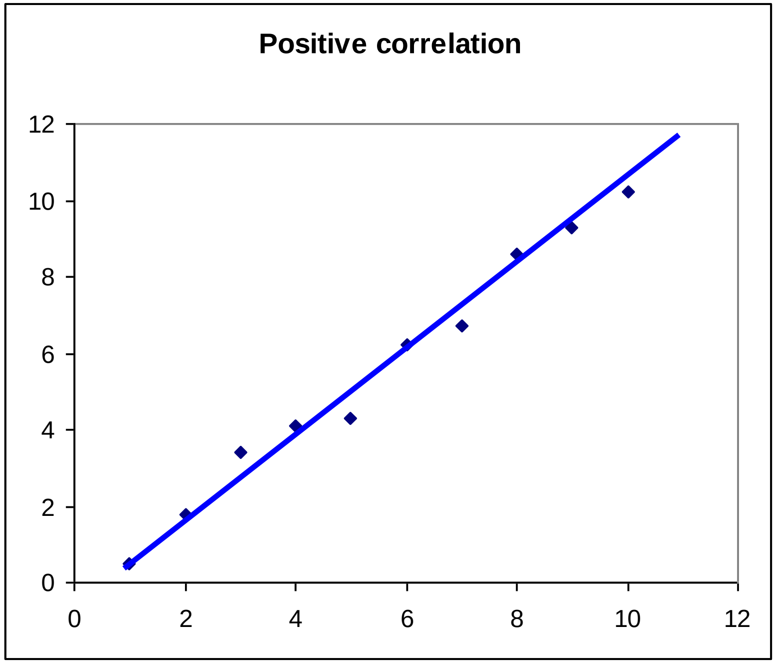
# Correlation

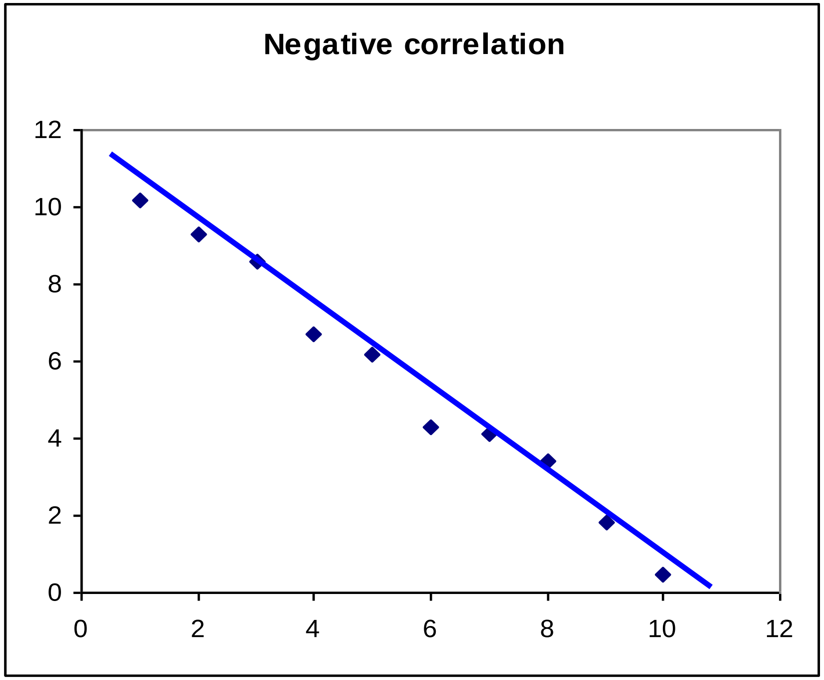
- **Definition:**
  - The degree to which two continuous variables are linearly related;
  - Study of existence, magnitude and direction of the relationship between two variables; or
  - The quantification of the degree to which two variables are linearly related
- Correlation (rho, or $\rho$) ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation)
- A correlation of 0 means that there is no linear relationship between the two variables
- Perfect correlation means that knowing one variable allows perfect knowledge of the other variable

$\rho$ is used for the population, while $r$ is used for the sample.

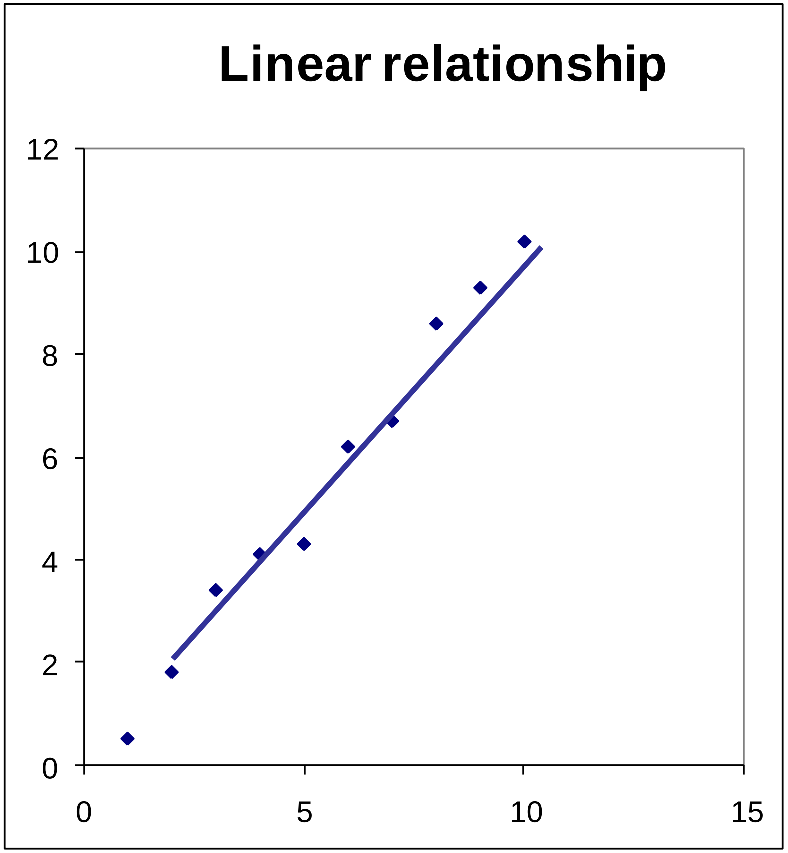
# Types of Correlation

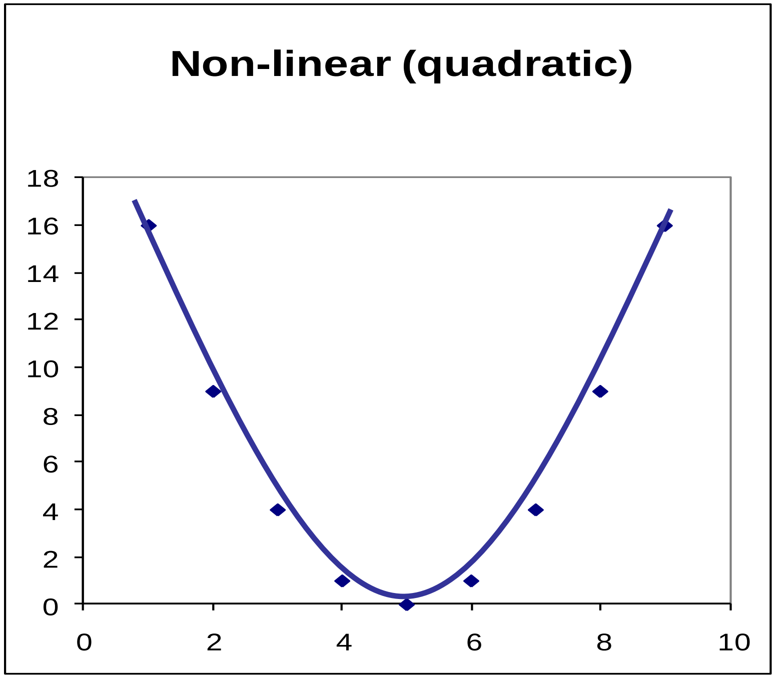
- **Positive Correlation:** Two variables change in the same direction – if one increases the other also increases, or if one decreases, the other also decreases
- **Negative Correlation:** Two variables change in the opposite direction - if one increases, the other decreases and vice versa





# Linear vs. Nonlinear Relationships

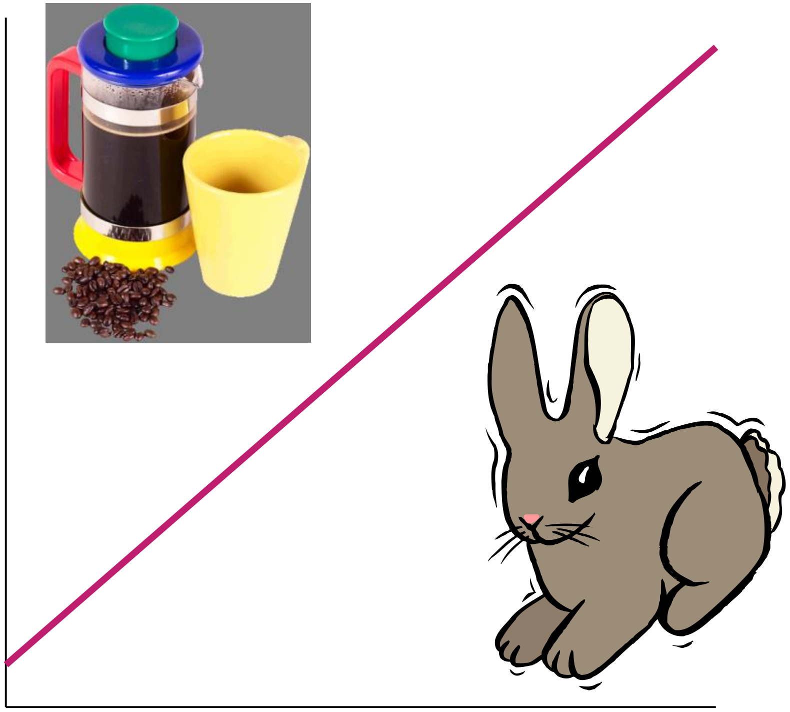




**Correlation can be used here, ……… but not here!**

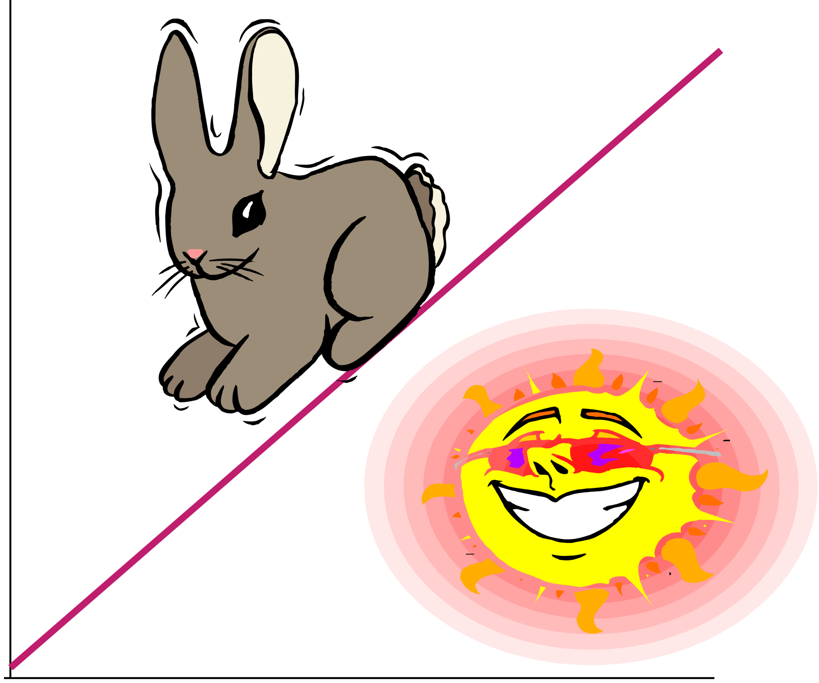
# Correlation does not imply Causation

Price of coffee



Number of hares per acre in Alaska

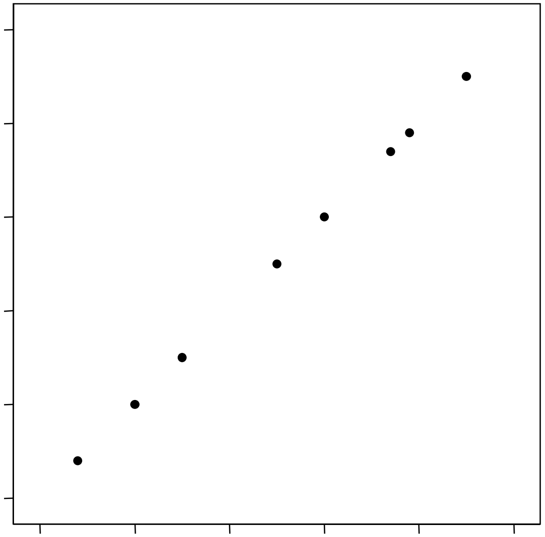
No. of hares/acre in Alaska



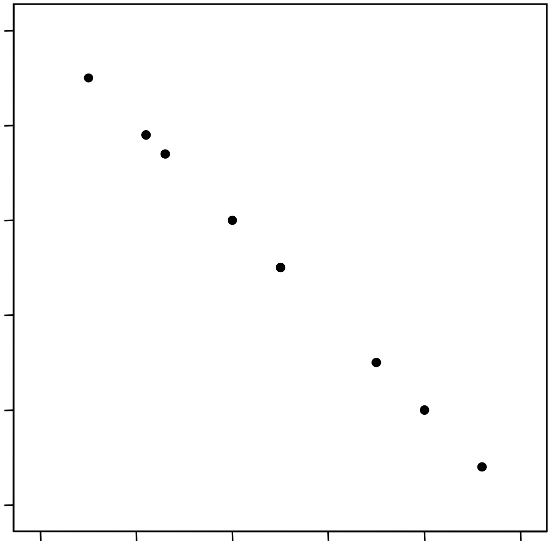
Number of solar flares, March-July

# Degrees of Correlation

- **Perfect correlation** (data points exactly on the line)
  - Two variables changes in the same direction and in the same proportion
  - The correlation between the two variables is perfect positive (*r=1*)  or it is a perfect negative (r=-1)

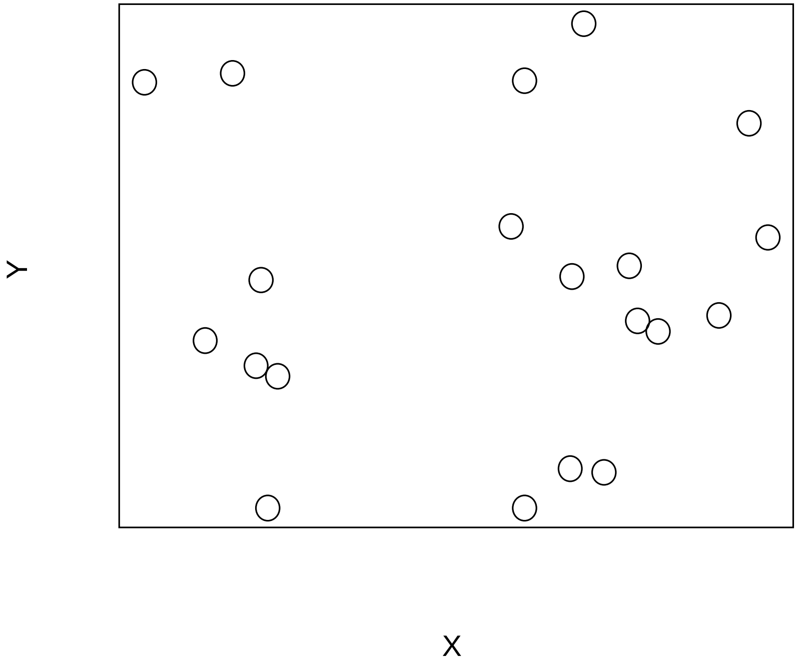


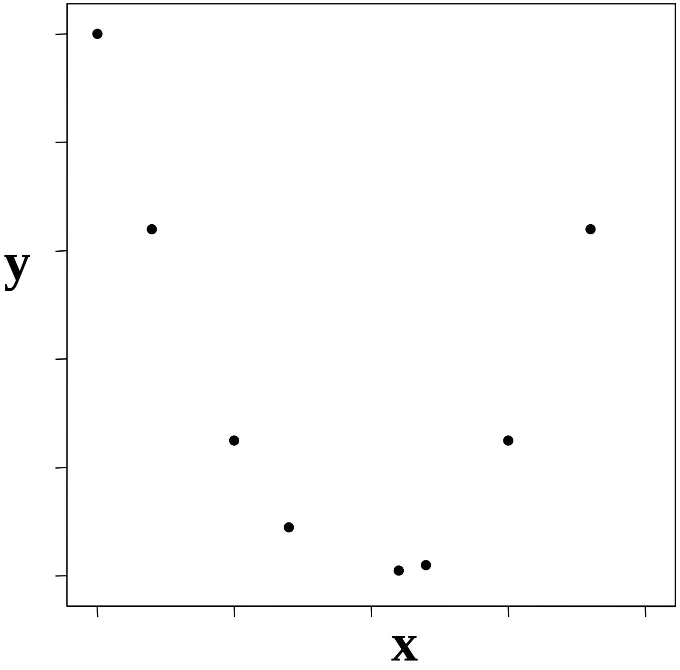
r= 1



r= -1

- **Absence of correlation**
  - Two variables exhibit no relationship, or a change in one variable does not lead to a change in the other variable
  - We say that there is no correlation between the two variables (*r*= *0*)





- Limited degrees of correlation
  - Two variables are not perfectly correlated
  - It may be positive, negative or zero but lies within the limits ±1

| Degree | Positive | Negative |
|---|---|---|
| Absence of Correlation | 0 | 0 |
| Perfect Correlation | 1 | -1 |
| High Degree | 0.75 to 1 | -1 to - 0.75 |
| Moderate Degree | 0.25 to 0.75 | - 0.75 to -0.25 |
| Low Degree | 0 to 0.25 | - 0.25 to 0 |

Some consider 0.33 to 0.66 as cut-off points

# Determining Correlation

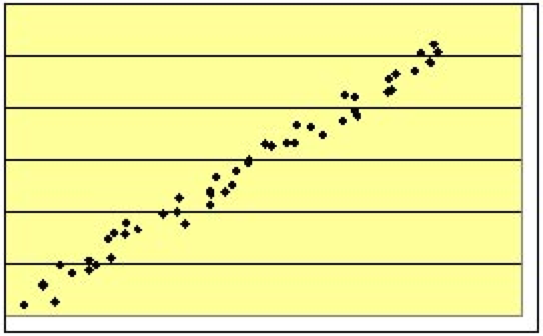
- Methods
  - See Scatter Plot
  - Pearson’s correlation coefficient
  - Spearman’s Rank-correlation coefficient: a non-parametric method that will be discussed later in the semester; see chapter 17.3

# Determining Correlation – Scatter Plot

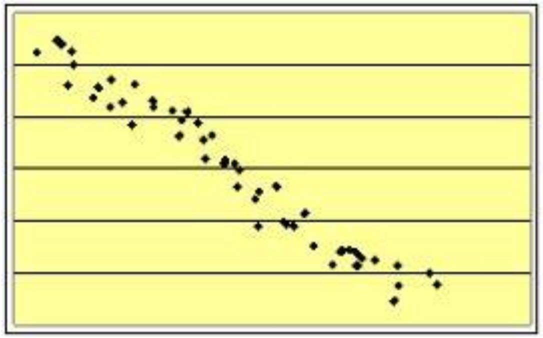
- Values of the two variables are graphed - one along the horizontal axis (x-axis) and the other along the vertical axis (y-axis)
- By plotting the data, we see points (dots) on the graph which are generally scattered, and hence the name ‘Scatter Plot’
- The manner in which these points are scattered suggest the degree and the direction of correlation, and they also help us decide if the dots follow a linear pattern or not.

# Scatter Plots

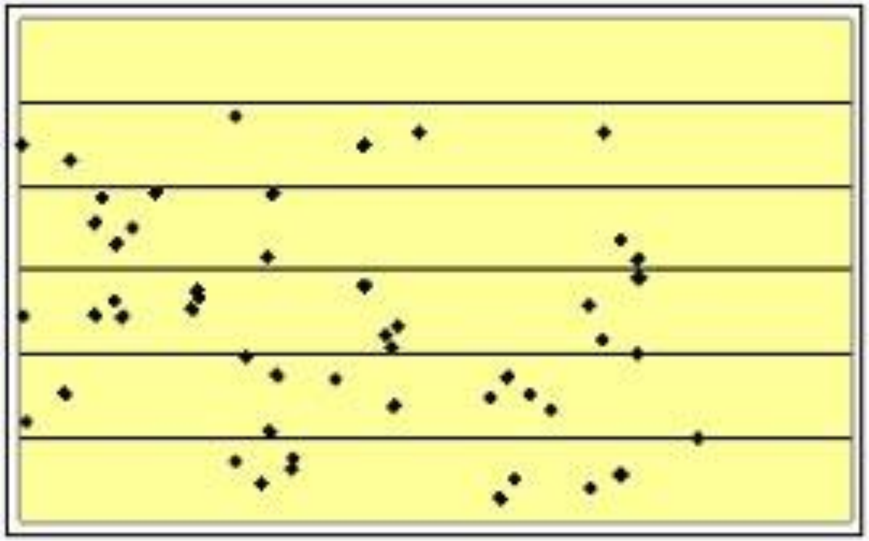
Positive, negative or no correlation?



**Positive**



**Negative**

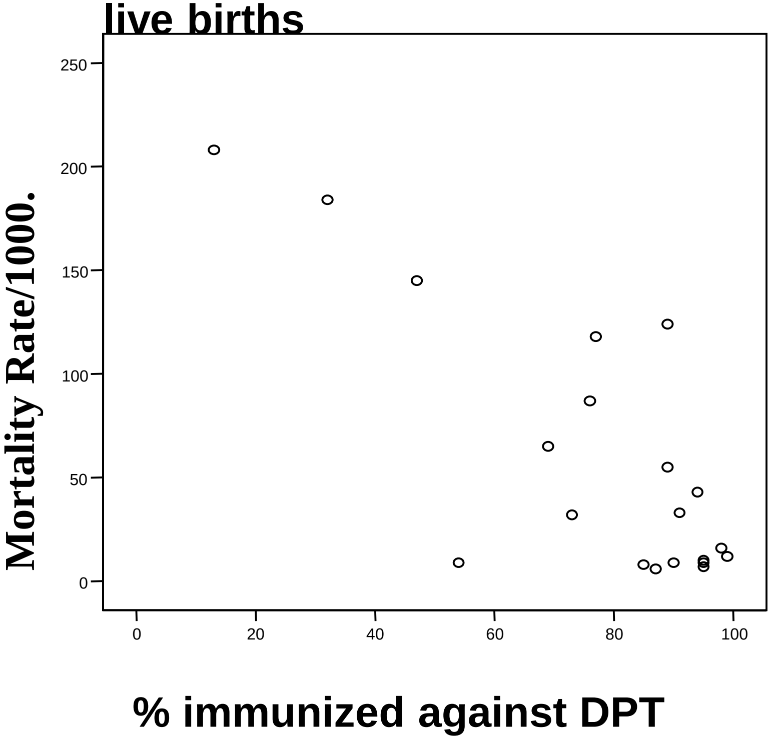


**No correlation**

# Example from your Text Book (pp 398-99)

Percentage of children immunized against DPT and under-five mortality rate for 20 countries, 1992

Under-five mortality rate per 1000 live births



| Nation | Percentage Immunized | Mortality Rate per 1000 Live Births |
|---|---|---|
| Bolivia | 77 | 118 |
| Brazil | 69 | 65 |
| Cambodia | 32 | 184 |
| Canada | 85 | 8 |
| China | 94 | 43 |
| Czech Republic | 99 | 12 |
| Egypt | 89 | 55 |
| Ethiopia | 13 | 208 |
| Finland | 95 | 7 |
| France | 95 | 9 |
| Greece | 54 | 9 |
| India | 89 | 124 |
| Italy | 95 | 10 |
| Japan | 87 | 6 |
| Mexico | 91 | 33 |
| Poland | 98 | 16 |
| Russian Federation | 73 | 32 |
| Senegal | 47 | 145 |
| Turkey | 76 | 87 |
| United Kingdom | 90 | 9 |

# Pearson’s Correlation Coefficient

- Let the correlation between two continuous variables *X* and *Y* be denoted by *ρ* (rho)
- The sample estimate of ***ρ*** is given by **r**

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

- The correlation coefficient has no units of measurement (that is, it is dimensionless)

# Mathematical Example

- Using the mortality data above
- $\bar{x} = 77.4\%,$ $\bar{y} = 59.0$ per 1000 live births

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - 77.4)^2 = 10630.8,$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - 59.0)^2 = 77498.$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - 77.4)(y_i - 59.0) = -22706,$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)}} = \frac{-22706}{\sqrt{(10630.8)(77498)}} = -0.79$$

# Interpretation of the result

- There is a strong linear relationship between the percentage of children immunized against DPT in a specified country and its under-five mortality rate
- Since *r* is negative (*r* = *-0.79*), mortality rate decreases in magnitude as percentage of immunization increases
- This is sometimes called an "inverse relationship" or a "negative correlation" since the coefficient has a negative sign.
- Note: If the correlation did not have a negative sign, we would call it a "positive correlation" or "positive relationship"

# Hypothesis Testing

- To determine whether any (significant) correlation exists between the random variable X and Y, we test the following hypothesis
- $H_0$: $\rho=0$ versus $H_A$: $\rho\neq0$
- We use the test statistic t (where n= #observations):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- *t* has a t-distribution with *n-2* degrees of freedom ($t_{n-2}$), having some similarities to the paired t-test we studied earlier.

# Example of Hypothesis Testing

- Using the mortality data above, *r = -0.79, n=20*

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.79\sqrt{20-2}}{\sqrt{1-(-0.79)^2}}$$

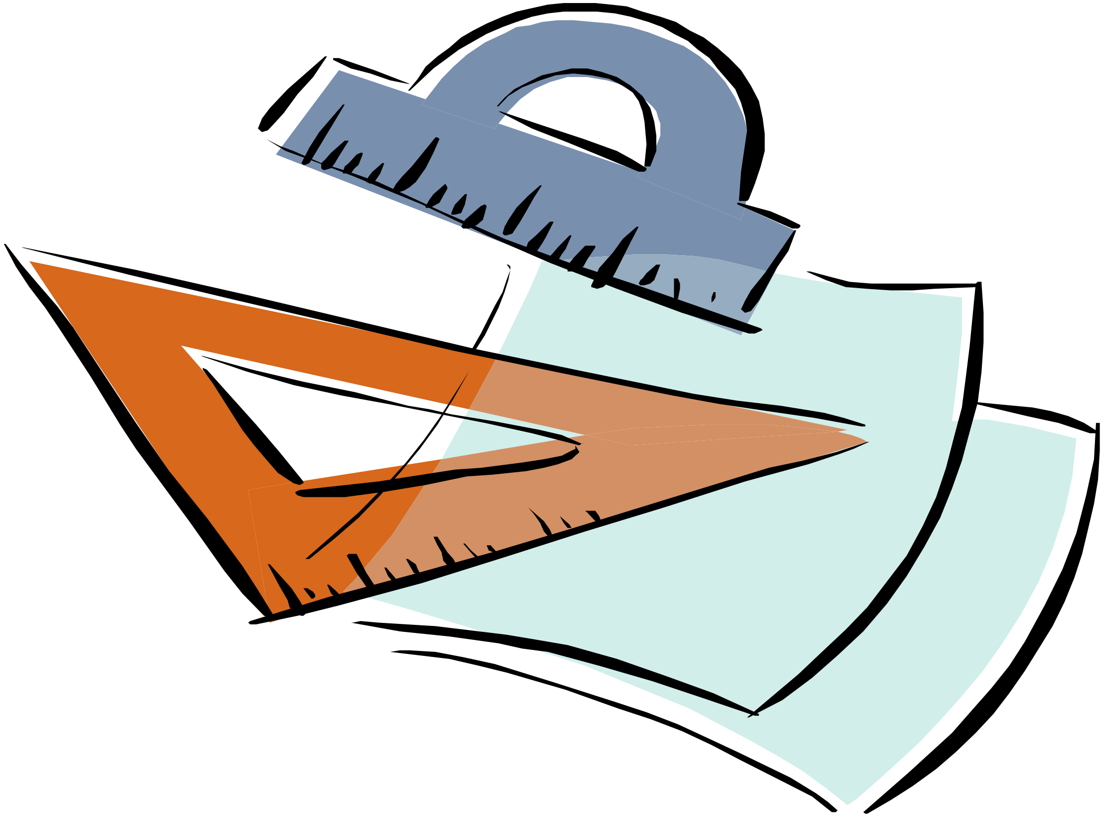
*= -5.47*

- Calculate its p-value: p<0.0001
- Reject the null hypothesis that ***ρ***=0 at 0.05 level
- Conclusion: There is a significant correlation (linear relationship) between immunization and under-five mortality

# Assumptions of Correlation

- The relationship between the two variables is linear
- Both variables are normally distributed
- There are no outliers – scores that fall outside the range of the rest of the scores on the scatter plot. The test statistic is highly sensitive to extreme values, which can yield misleading results.

# LINEAR REGRESSION



# Linear Regression

- GOAL: To find a linear equation relating a dependent (or outcome) variable, Y, to an independent (or predictor variable), X
- Stated another way, we wish to measure the magnitude, direction, and strength of the linear relationship between Y and X
- Requirement: both variables must be continuous

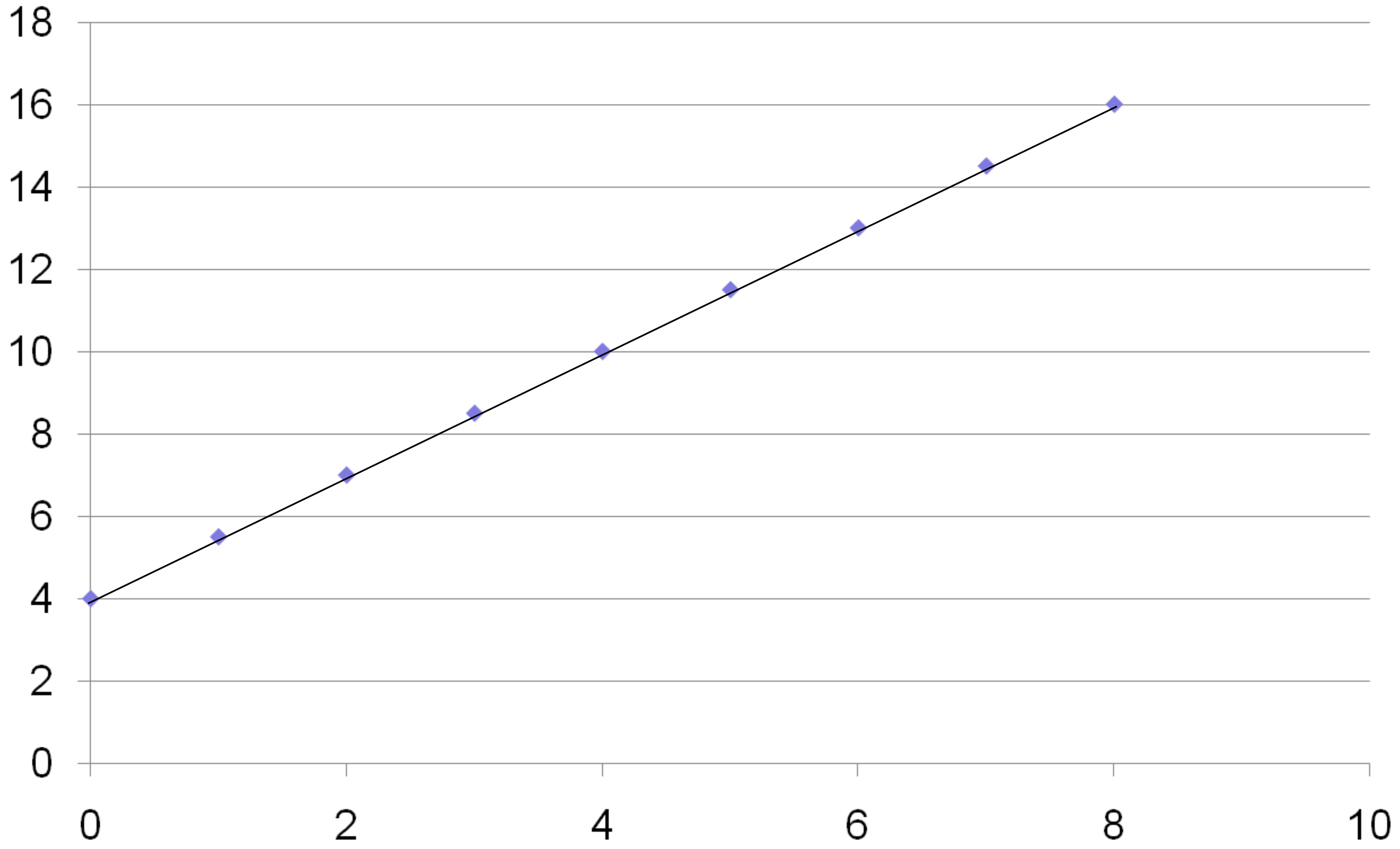
# Overview of the Regression Line

- The equation of a line is given by $y = a + Bx$
- $B$ is the slope

  $a$ is the $y$-intercept
- A definition of $B$ is that for every one unit increase in $x$, there is an $B$ unit increase in $y$
- A definition of $a$ is the value of $y$ when $x$ is equal to zero

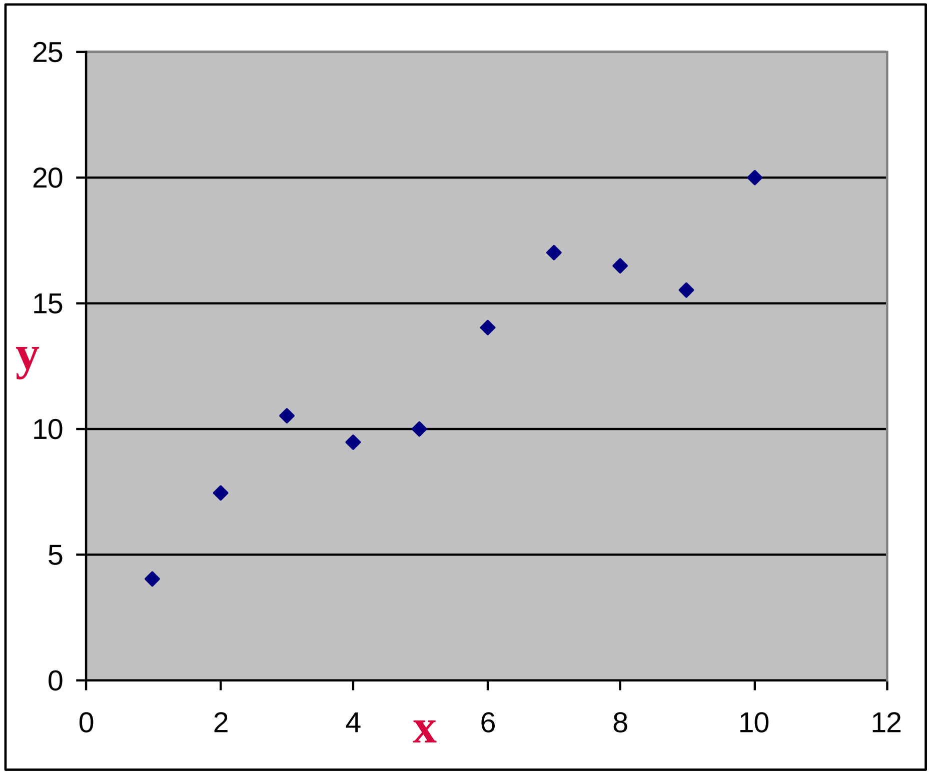
# Example of a line equation:

$$y = 4 + 1.5*x$$



# Procedure – start with scatter plot

- Look at the data in the scatter plot
- Does there seem to be a correlation (linear relationship) in the data?
- Is the data perfectly linear?
- Could we fit a line to this data?



# Simple Linear Regression

- **Y** is the dependent variable
- **X** is the independent variable
- We try to predict Y from X
- Called “simple” because there is only one independent variable **X**
- If there are several independent variables, it’s called multiple linear regression (not covered in this course)

# Simple Linear Regression Method

- Linear regression tries to find the best line to fit the data
- One of the assumptions is that the relationship between the predictor and the outcome is linear
- We call this the population regression line:

$$E(Y|x) = \beta_0 + \beta_1 x$$

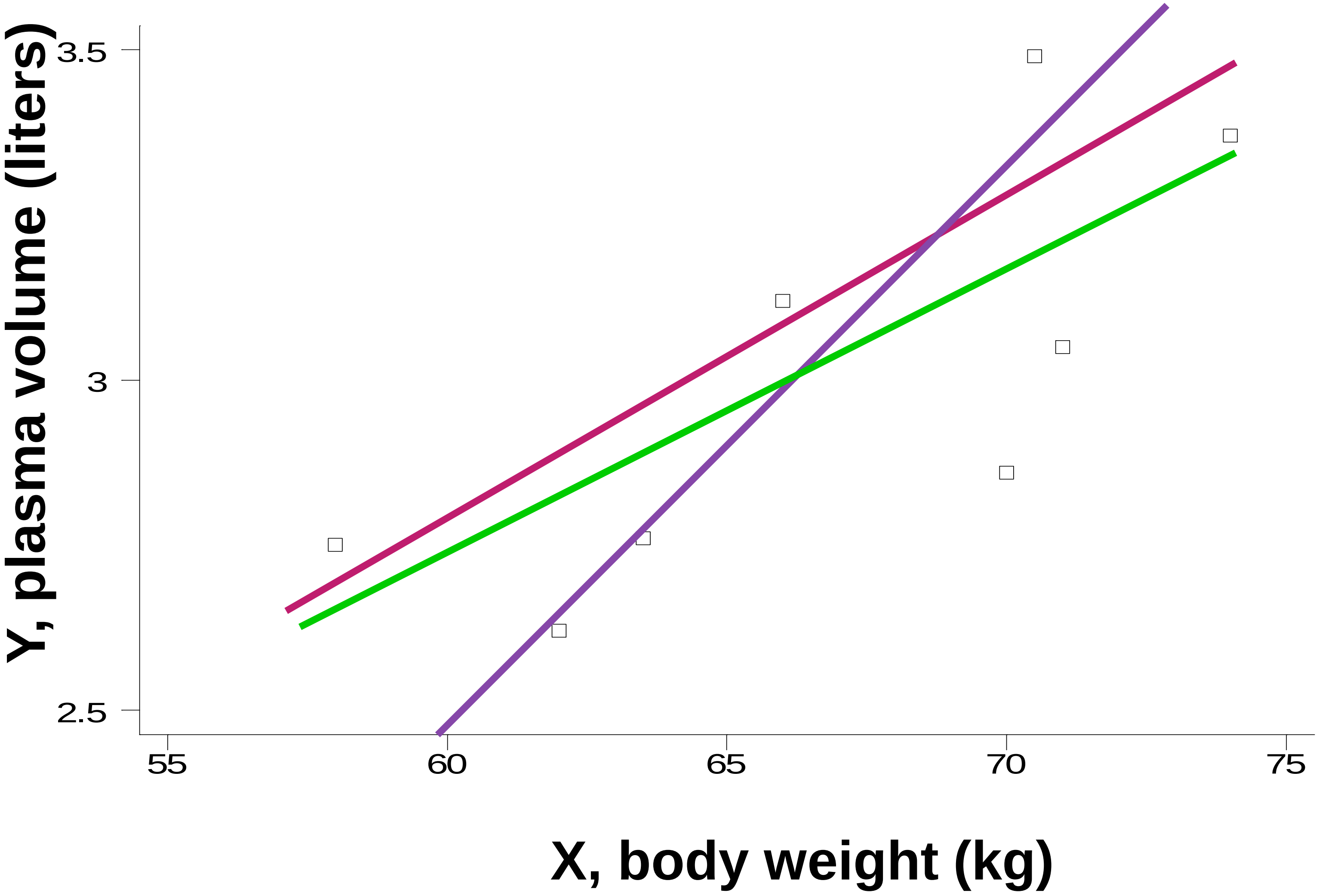
- This equation says that the expected (E) or predicted mean of Y given a specific value of X=x is defined by the β parameters
- There are two keys to understanding the equation
  - The equation is conditional on the value of X=x
  - The coefficients act exactly like the slope and y-intercept from the simple equation of a line (where is $\beta_0$ the intercept and $\beta_1$ is the slope)

# Linear Regression Example

Relationship between body weight (X) and plasma volume (Y)

| Subject | Body Weight (Kg) | Plasma Volume (L) |
|---|---|---|
| 1 | 58.0 | 2.75 |
| 2 | 70.0 | 2.86 |
| 3 | 74.0 | 3.37 |
| 4 | 63.5 | 2.76 |
| 5 | 62.0 | 2.62 |
| 6 | 70.5 | 3.49 |
| 7 | 71.0 | 3.05 |
| 8 | 66.0 | 3.12 |

# Drawing by hand: which line is right?



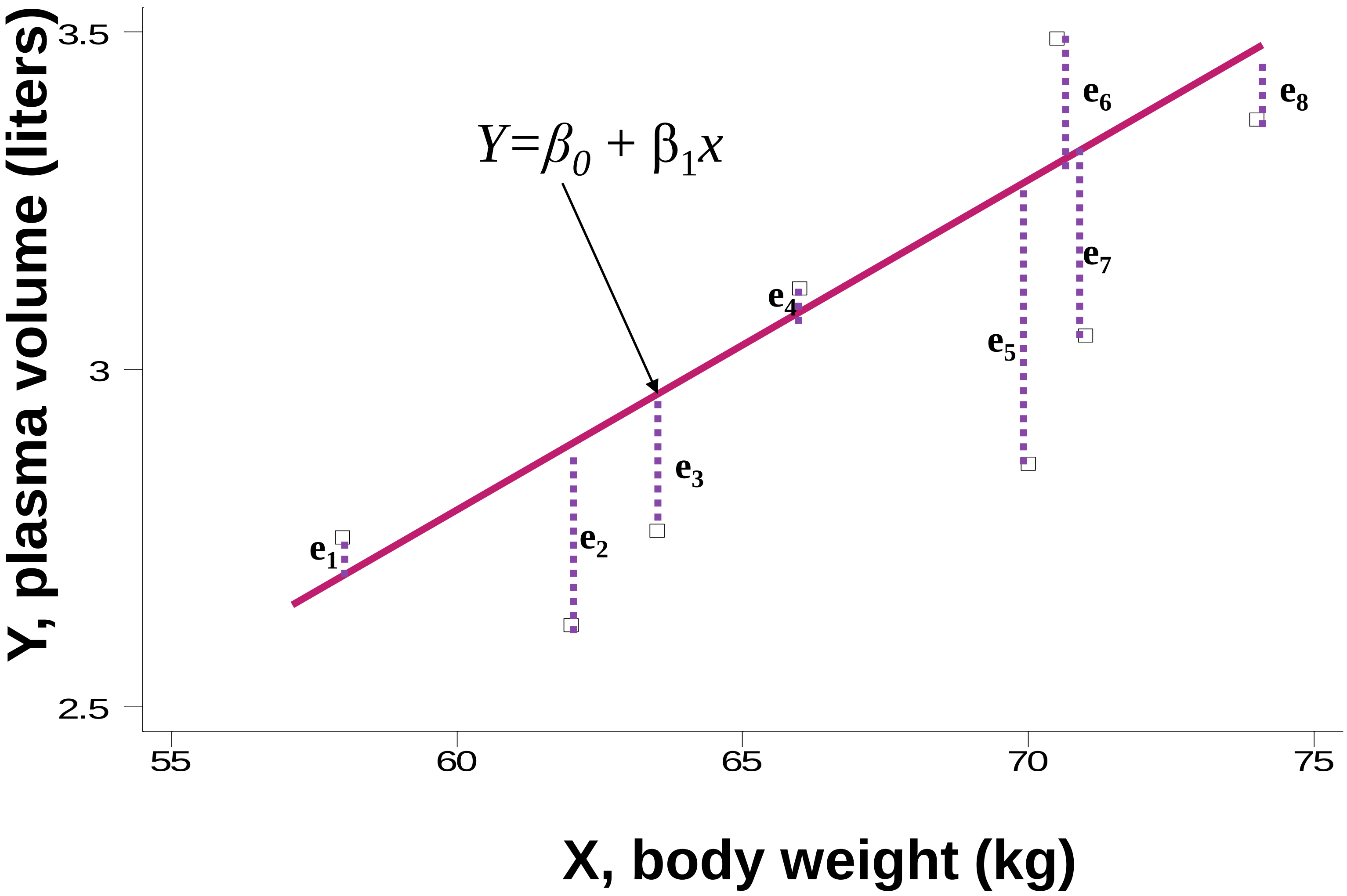
# How Do We Choose the “Right” Line?

- The linear regression line is the line which gets “closest” to all of the points
- How do we measure closeness to more than one point?

$$\text{minimize} \sum_{i=1}^{n} (\mathrm{y_i} - \mathrm{point\_on\_line_i})^2$$

The difference between an observed Y and a predicted Y is called a residual or error (e). We want to minimize the sum of the squares of the residuals.

# Regression "errors" (e)



# Least squares method

- The method employed to find the best line is called least squares
- This method find the values of $\beta_0$ and $\beta_1$ that minimize the squared vertical distance from the line to each of the points
- This is the same as minimizing the sum of the $e_i^2$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_1\right)\right)^2$$

# The Slope

- The slope $\beta_1$ is the expected change in Y corresponding to a unit increase in X
  - $\beta_1 = 0$ No linear relationship btw Y and X
  - $\beta_1 > 0$ Positive correlation (as X increases Y tends to increase)
  - $\beta_1 < 0$ Negative correlation (as X increases Y tends to decrease)

Note: Slope=0 same as zero corr.

# Plasma Example

- Suppose we used simple linear regression to estimate a line relating plasma volume to body weight from 8 data points. The regression equation is:

Plasma volume = 0.0857 + 0.0436 * weight

or (*y = 0.0857 + 0.0436*x*)

- Estimate of the intercept is 0.0857
- Estimate of the slope is 0.0436
  - Interpretation: For each kilogram increase in body weight, we expect plasma volume to increase by 0.0436 liters

# Plasma Example (cont.)

- Measurement of plasma volume is very time consuming
- Body weight easy to measure. Use equation and body weight to estimate plasma volume
  - If we know an individual’s weight, we can estimate his plasma volume
  - Estimate the plasma volume for a 60kg man
  - $y = 0.0857 + 0.0436 * x$
  - For this individual, $x = 60$

  → $y = 0.0857 + 0.0436 * 60 = 2.7$ liters

# Plasma Example (cont.)

- If we calculate the correlation coefficient for X and Y, we get r = 0.759, which means high positive correlation.

- **Only in this specific case** of simple linear regression, the coefficient of determination **$R^2 = r^2$** $= 0.759^2 = 0.576$ (*not true otherwise*)

- This means that 57.6% of the variation in plasma volume is explained or accounted for by body weight

# Hypothesis Testing

- $H_0$: $\beta_1=0$ (the slope is zero)
- $H_a$: $\beta_1 \neq 0$
- We need to calculate a test statistic based on our sample and compare to appropriate distribution to get a p-value

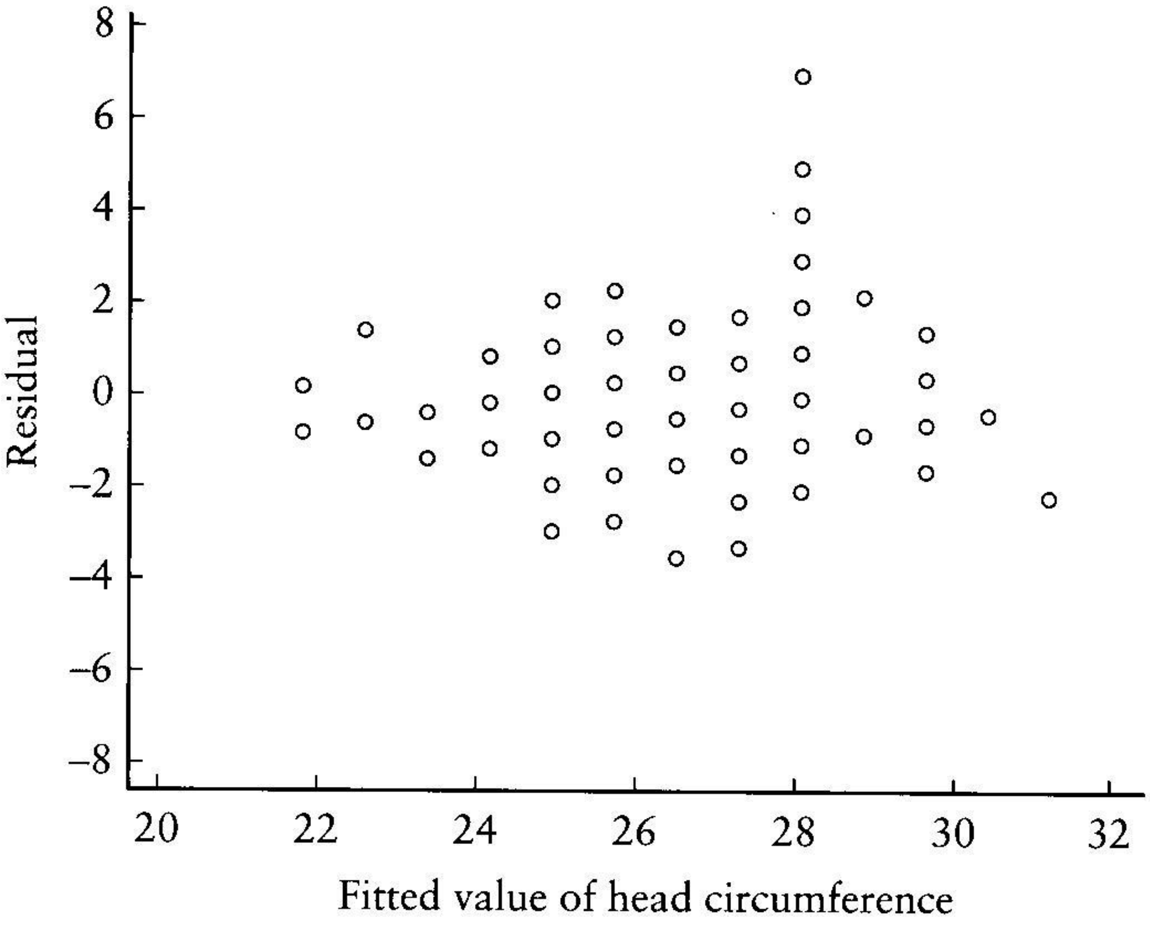
$$t = \frac{\text{estimated\_slope}}{\text{SE(estimated\_slope)}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.0436}{0.0153} = 2.85$$

- p-value = 0.03 based on t-distribution with ***n-2*** degrees of freedom, so we conclude that the slope is not zero (reject null hypothesis), and therefore plasma volume is positively correlated (linear relationship) with body weight

# Assumptions of linear regression

- **Linearity and Distribution**
  - The relationship between the outcome and the predictors can be described by a linear relationship
  - The errors, $e_i$, are normally distributed
- **Homoscedasticity of the errors**
  - The errors, $e_i$, have the same variance
- **Independence**
  - All of the data points are independent

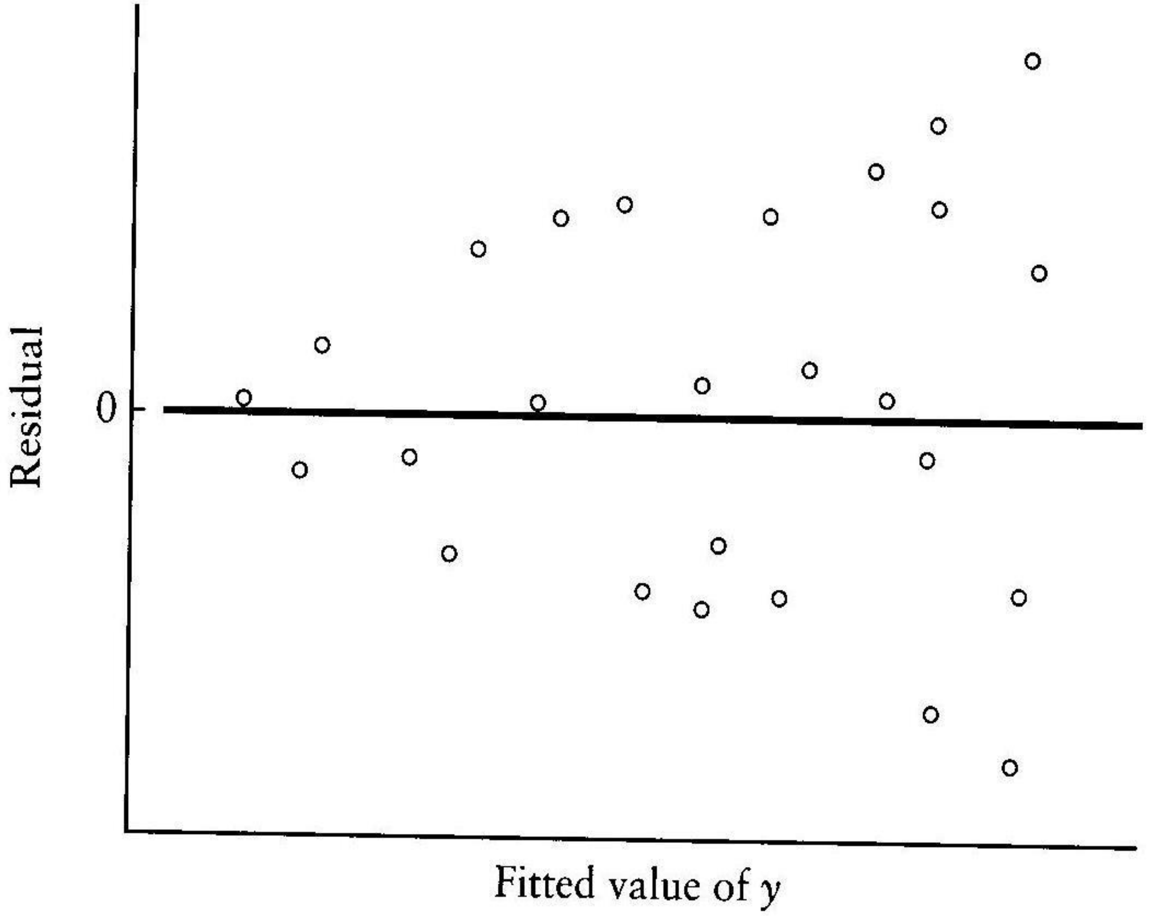
# Example in which Homoscedastic Errors Assumption is Met



Note that the variability of the residuals (errors) is random, i.e. they do not increase as the fitted value of the dependent variable increases

Text, page 435

# Example in which the Homoscedastic Errors Assumption is Violated



Note that the variability of the residuals (errors) is increasing as the fitted value of the dependent variable increases

Text, page 435

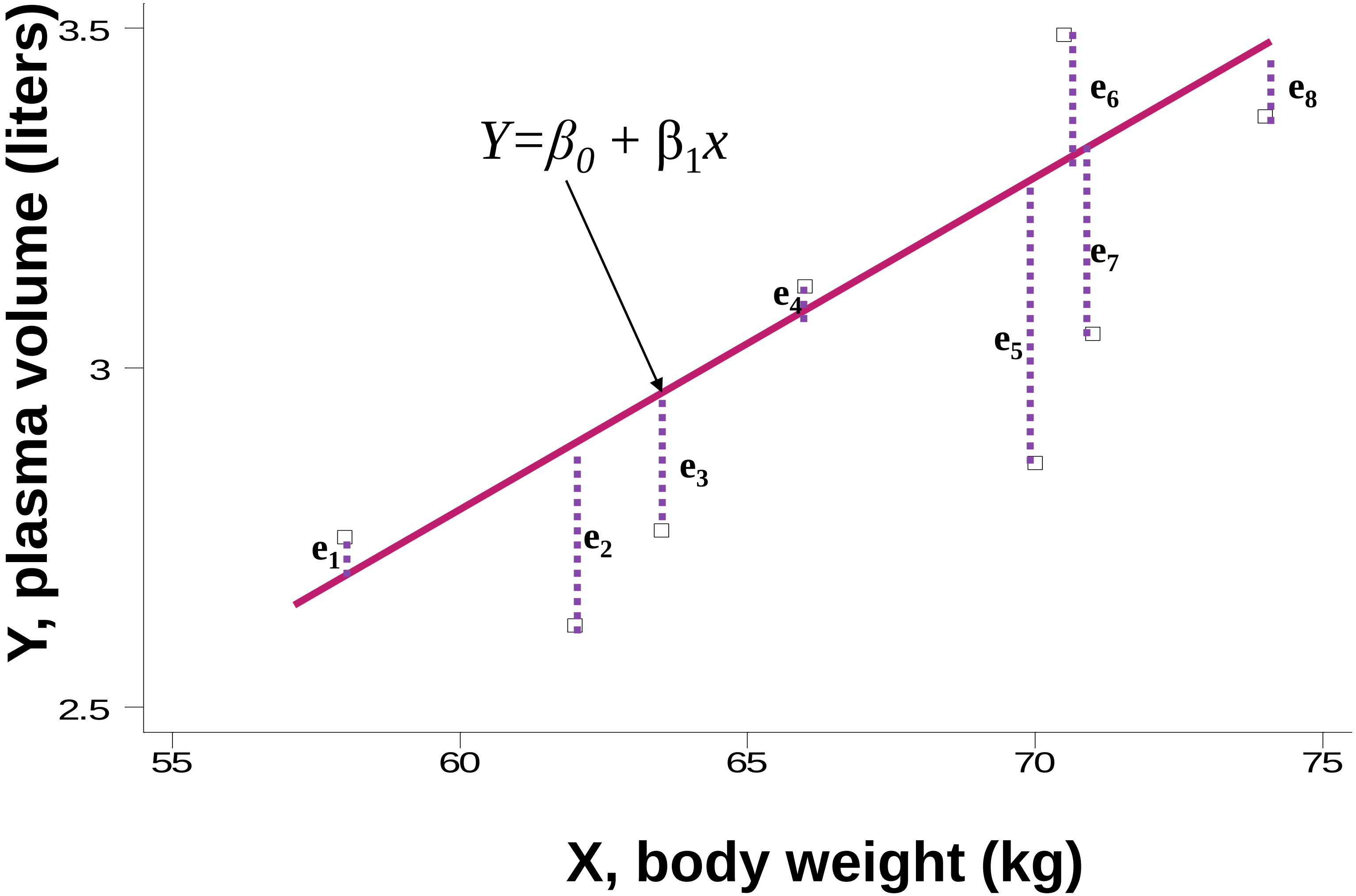
Note: For residuals we do not want to see patterns

# Some Notes about Simple Linear Regression

- Model predicts Y from X
- Relationship between Y and X is a straight line (linear)—start with scatter plot
- Can only use equation for predicting Y values based on X values inside range of original X values
  - ***Beware of extrapolation!*** (next slide)
- Pairs of data points (**x**, **y**) are independent
- **$r^2$** – same as $R^2$ , % variation in Y explained by X

# Perils of Extrapolation: in this example, can we be sure that the line continues to be linear, out of the data range?

Note: the lowest weight was 58 kg, so we could not predict for 55 kg for ex.



# Slope vs. Correlation Coefficient

- Both indicate the direction of the linear relationship (positive or negative)

- Regression provides a straight line equation, where the slope is the expected change in **Y** per unit increase in X

- Correlation does not provide a straight line equation, it indicates how close to linearity is the relationship (association) between var.

# Correlation and Regression In STATA

# Correlation in STATA

- **correlate** provides the Pearson r

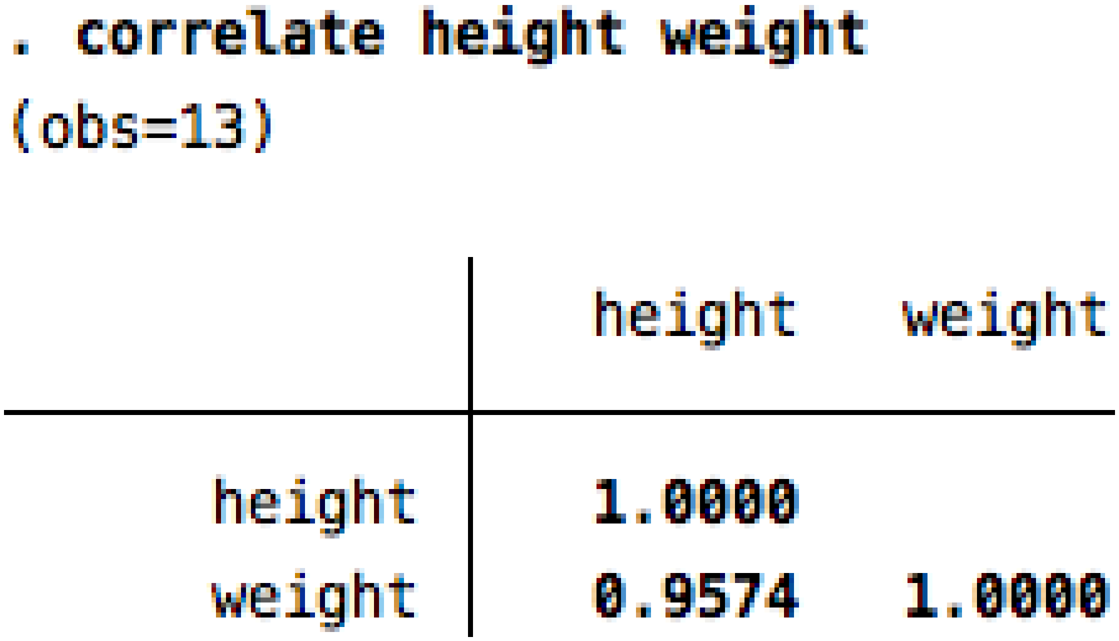
- Syntax:

```
correlate {first variable} {second variable}
```

- Example:

```
correlate height weight
```

# Example of Correlate Output



```
. correlate height weight
(obs=13)

             |   height   weight
-------------+------------------
      height |   1.0000
      weight |   0.9574   1.0000
```

**Strong corr btw Ht and Wt, 0.9574 very close to one.**

# Regression in STATA

- **Regress** provides estimates of the slope and intercept, their p-values, the $R^2$ of the model, and other information.
- Syntax:

**regress dependent_var independent_var**

- Example (in which weight was in pounds and height in inches):

**regress weight height**

Note: For regression it is important to decide which is the dependent variable and which one is the independent variable depending on the research question.
For correlation it does not matter which variable is listed first.

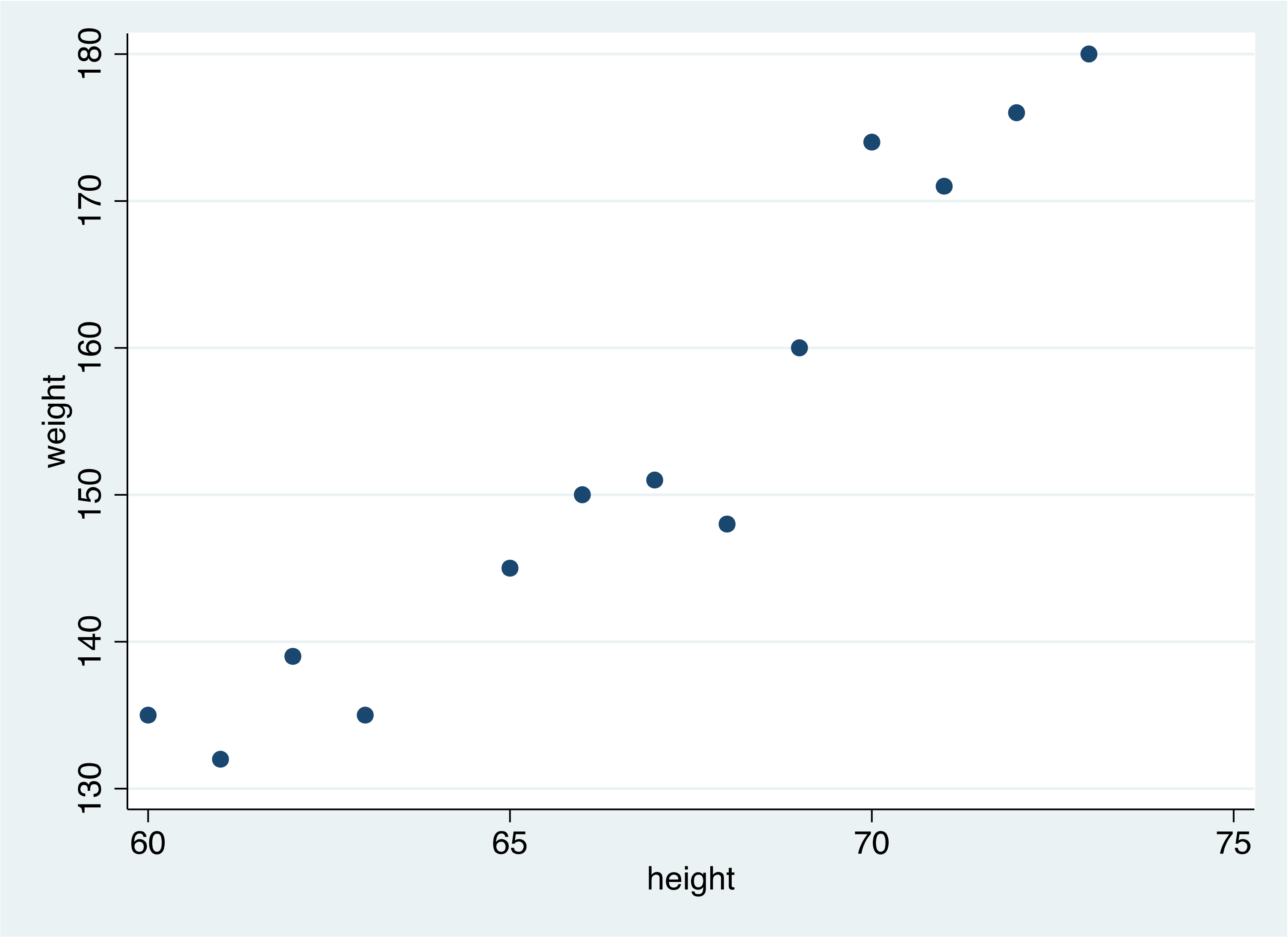
# The data set for this example

| | |
|---|---|
| 60 | 135 |
| 61 | 132 |
| 62 | 139 |
| 63 | 135 |
| 65 | 145 |
| 66 | 150 |
| 67 | 151 |
| 68 | 148 |
| 69 | 160 |
| 70 | 174 |
| 71 | 171 |
| 72 | 176 |
| 73 | 180 |

Note that 13 people were measured: the first number is the height (inches) and the second is the weight (pounds)

# Scatter Plot of Ht and Wt

## (created in STATA: *scatter weight height*)

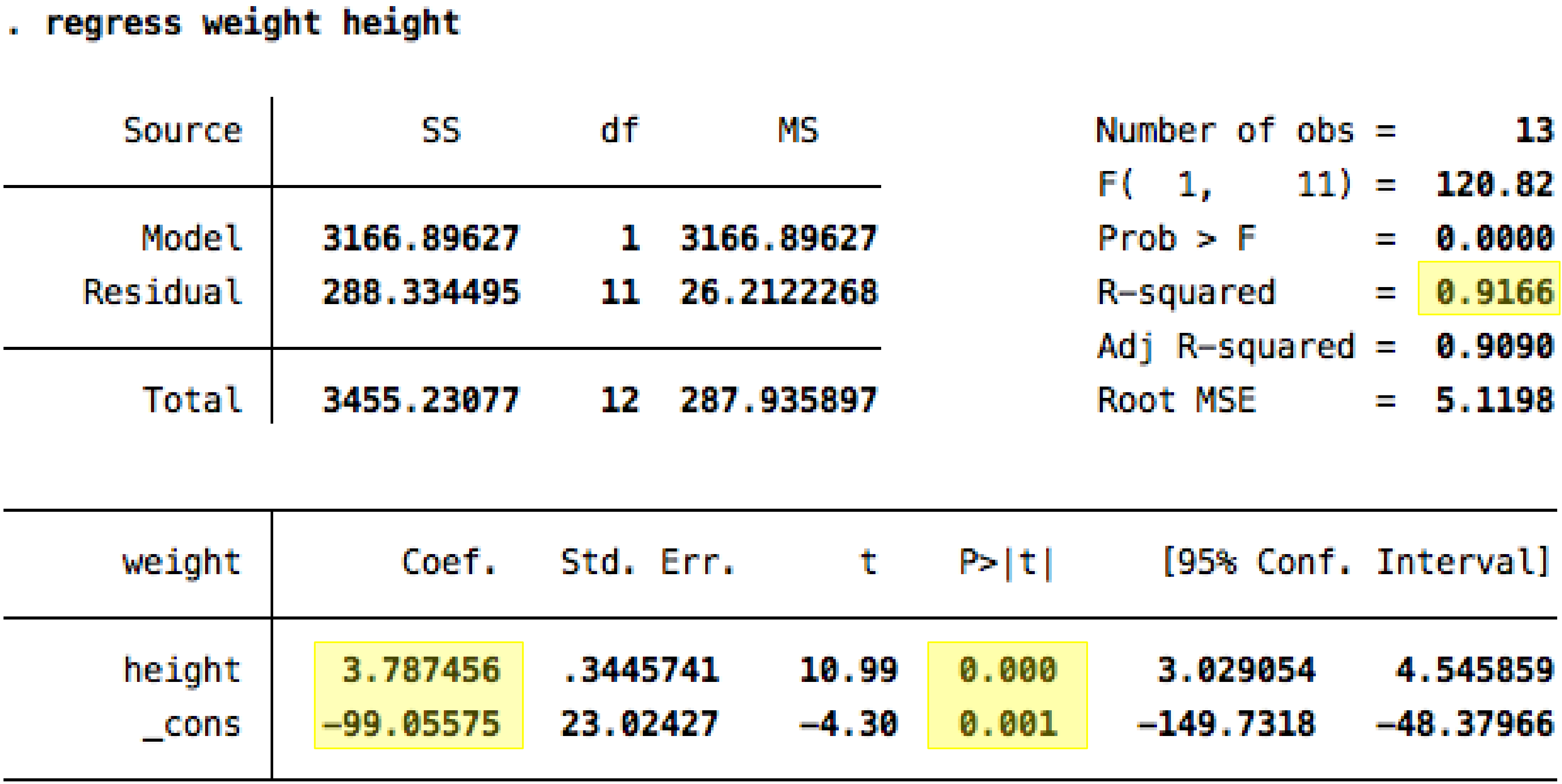


From the previous scatter plot, is it appropriate to proceed to linear regression?

Are all the assumptions true?

How can we tell?

# Example of Regress Output



```
. regress weight height

      Source |       SS       df       MS              Number of obs =      13
-------------+------------------------------           F(  1,    11) =  120.82
       Model |  3166.89627     1  3166.89627           Prob > F      =  0.0000
    Residual |  288.334495    11  26.2122268           R-squared     =  0.9166
-------------+------------------------------           Adj R-squared =  0.9090
       Total |  3455.23077    12  287.935897           Root MSE      =  5.1198

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   3.787456   .3445741    10.99   0.000     3.029054    4.545859
       _cons |  -99.05575   23.02427    -4.30   0.001    -149.7318   -48.37966
------------------------------------------------------------------------------
```

# Interpretation of the Output

- **R-squared=0.9166**, meaning that ~92% of the variability in weight in this sample is explained by height.
- **Intercept= -99.0558** (The intercept is not always meaningful as it may correspond to a point out off range. Here no meaning bc there is no zero height.)
- **Slope=3.788**, meaning that, for one inch increase in height, the weight increases by ~3.8 pounds.
- **P<0.0001** for the test for the slope=0, meaning that we can be extremely confident in rejecting the null hypothesis that the slope=0 (and also zero corr).

# Prediction based on the Regression Equation

- Results indicate weight = -99 + 3.8*Height
- This information can be used to predict weights given different heights.
- For example, a person who is 6 feet tall (72 inches) is predicted to weigh 174.6 lbs ( -99 + 3.8*72 = 174.6).
- Suppose we wanted to predict the weight for a person who was 7 feet tall: would we want to do this? No, it only predicts weight for height values within the range.