Regression Relationships

Nouman M Durrani

Introduction

- Regression analysis is the most widely used method for the analysis of dependence
  - that is, for examining the relationship between a set of independent variables (X's) and a single dependent variable (Y).
- Regression (in general) is a linear combination of independent variables that corresponds as closely as possible to the dependent variable.

$$Y=b_0+b_1x_1+b_2x_2 + \ldots+b_nx_n$$

Regression versus Classification

- Classification: the output variable takes class labels

- Regression: the output variable takes continuous values

Examples

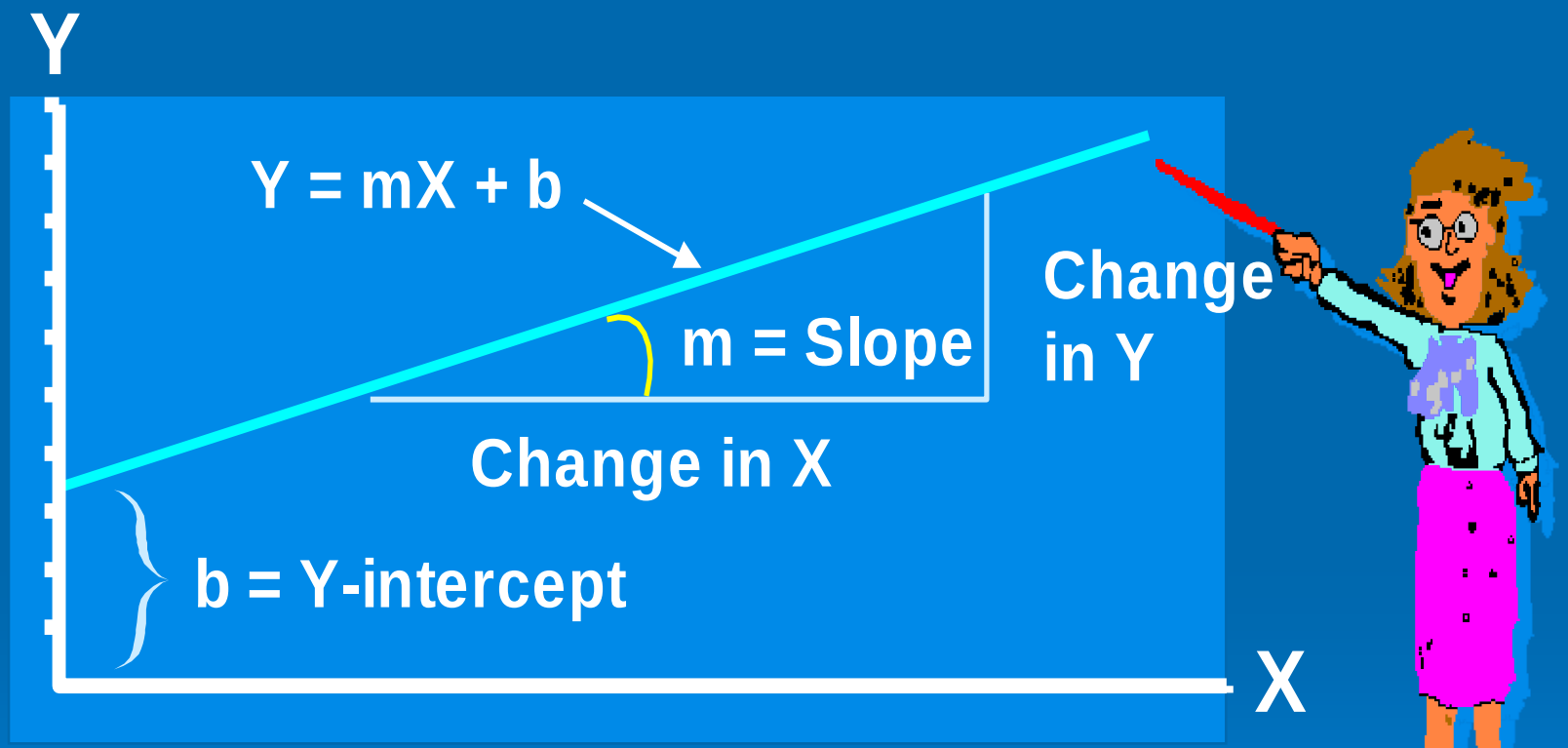
- Predicting House Value
  - Actual Price: £100,000
  - Predicted 1: £99,950 (Very Good Prediction)
  - Predicted 1: £50,000 (Very Bad Prediction)

- Predicting Car Premium
  - Using Location, Age, History etc

Regression Techniques

- Linear Regression
- Ridge Regression
- Lasso Regression
- And many more

Linear Equations



Simple Linear Regression

- The Simple Linear Regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A sequence of random variables is IID if and only if the following two conditions are satisfied:
1. the terms of the sequence are mutually independent;
2. they all have the same probability distribution.

where y_i is the response of the i[th] observation

β_0 is the y-intercept

β_1 is the slope

x_i is the value of the predictor variable for the ith observation

$\varepsilon_i \sim \text{iid } \text{Normal}(0, \sigma^2)$ is the random error

$i = 1, \ldots, n$

Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

Random Error

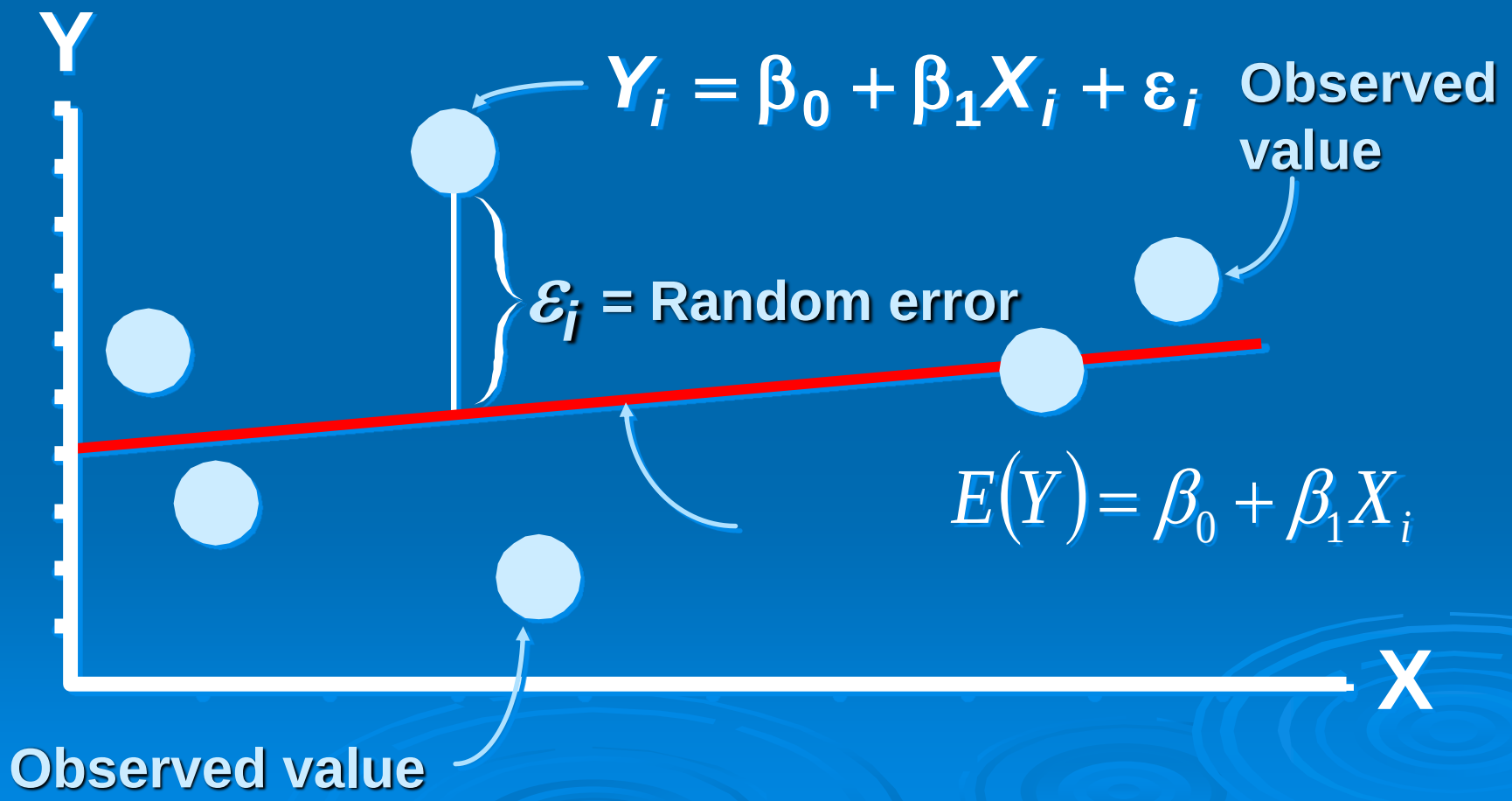
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable (e.g., CD+ c.)

Independent (Explanatory) Variable (e.g., Years s. serocon.)

Population Linear Regression Model

Try to find a line that will fit all my points, based on the Least square method

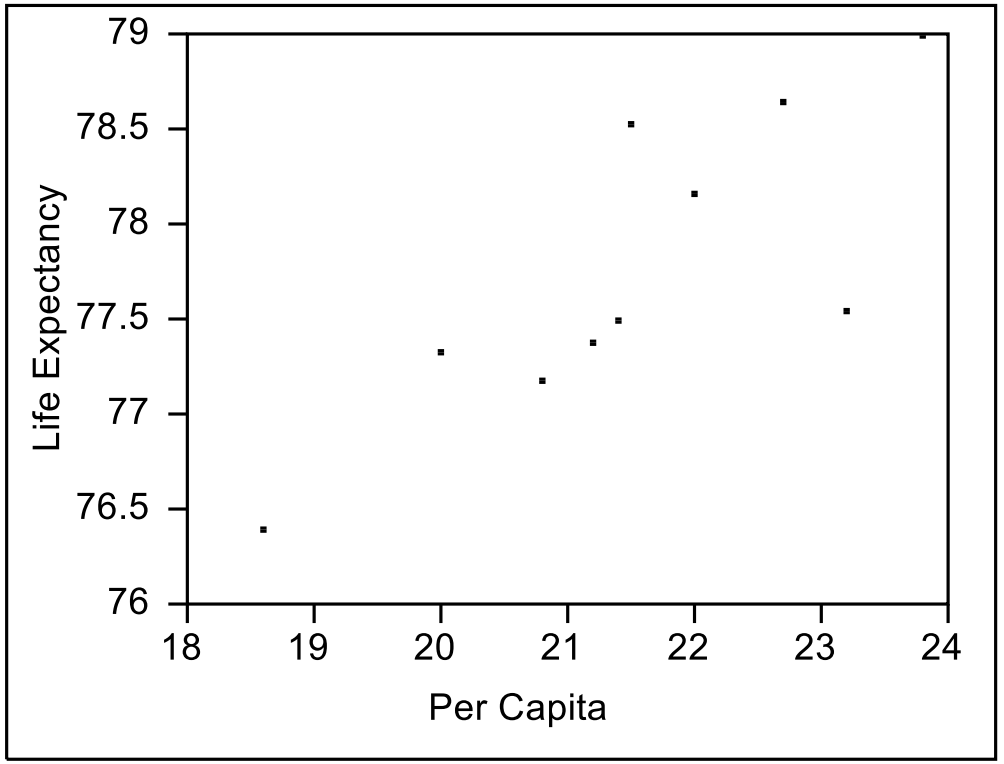


Simple Linear Regression

- Simple Linear Regression (SLR) is used to model the relationship between two continuous variables.

- The variable on the x-axis is often called the explanatory or predictor variable.

- The variable on the y-axis is called the response variable.

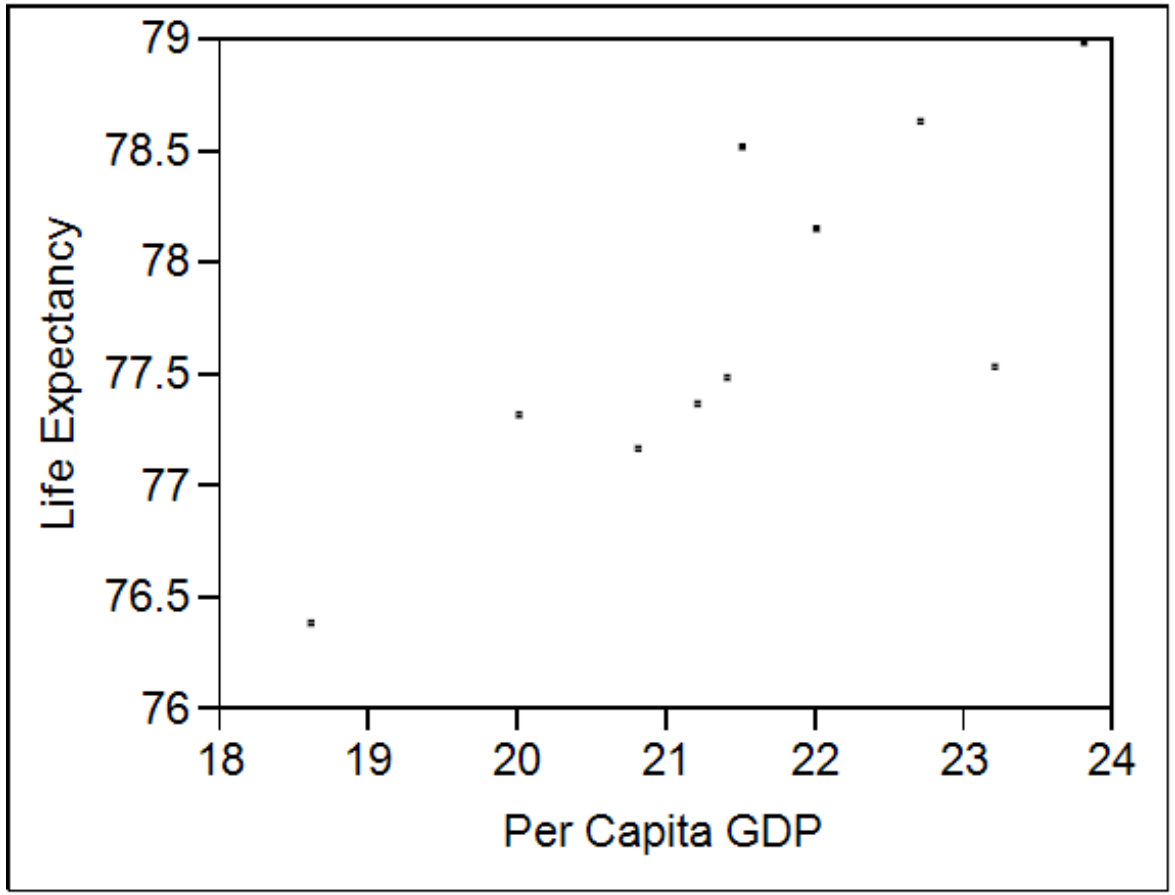


Sullivan (pg. 193)

- Scatterplots are used to graphically examine the relationship between two quantitative variables.

Simple Linear Regression

Can we describe the behavior between the two variables with a linear equation?

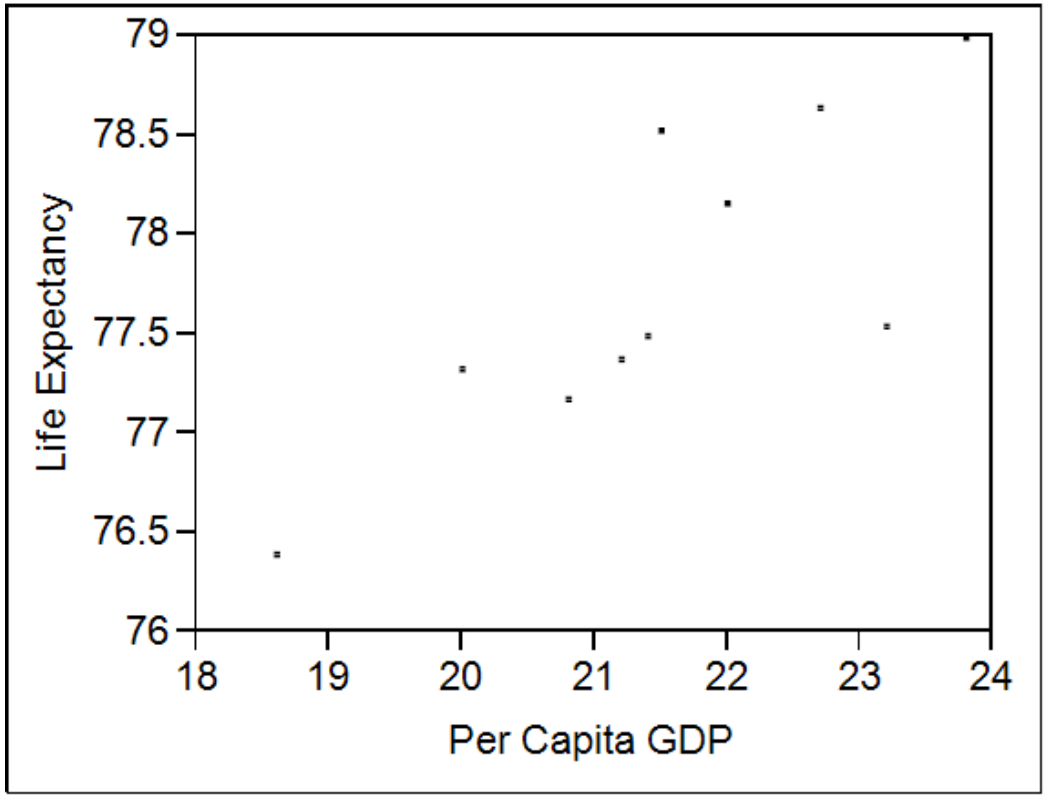


- The variable on the x-axis is often called the explanatory or predictor variable.
- The variable on the y-axis is called the response variable.

Simple Linear Regression

- Objectives of Simple Linear Regression
  - Determine the significance of the predictor variable in explaining variability in the response variable.
    - (i.e. Is per capita GDP useful in explaining the variability in life expectancy?)
  - Predict values of the response variable for given values of the explanatory variable.
    - (i.e. if we know the per capita GDP can we predict life expectancy?)

- Note: The predictor variable does not necessarily cause the response.



Regression Model Purposes

Regression models are used for purposes of description, inference and prediction.

1. Description:

- How can we describe the relationship between the dependent variable and the independent variables?
- How strong is the relationship captured by the model?

2. Inference:

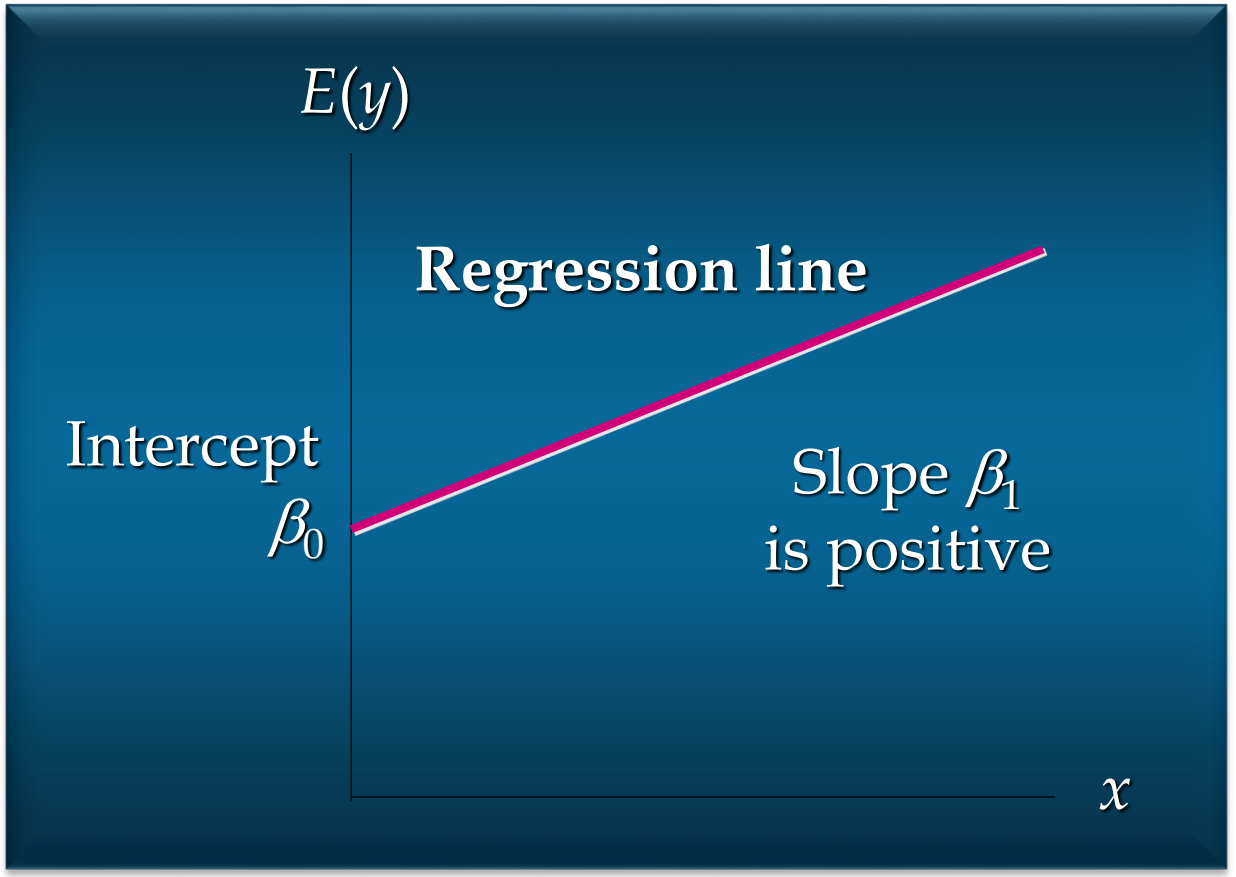
- Is the relationship described by the model statistically significant (i.e., is this level of association between the fitted values and the actual values likely to be the result of chance alone?)
- Which independent variables are most important?

3. Prediction:

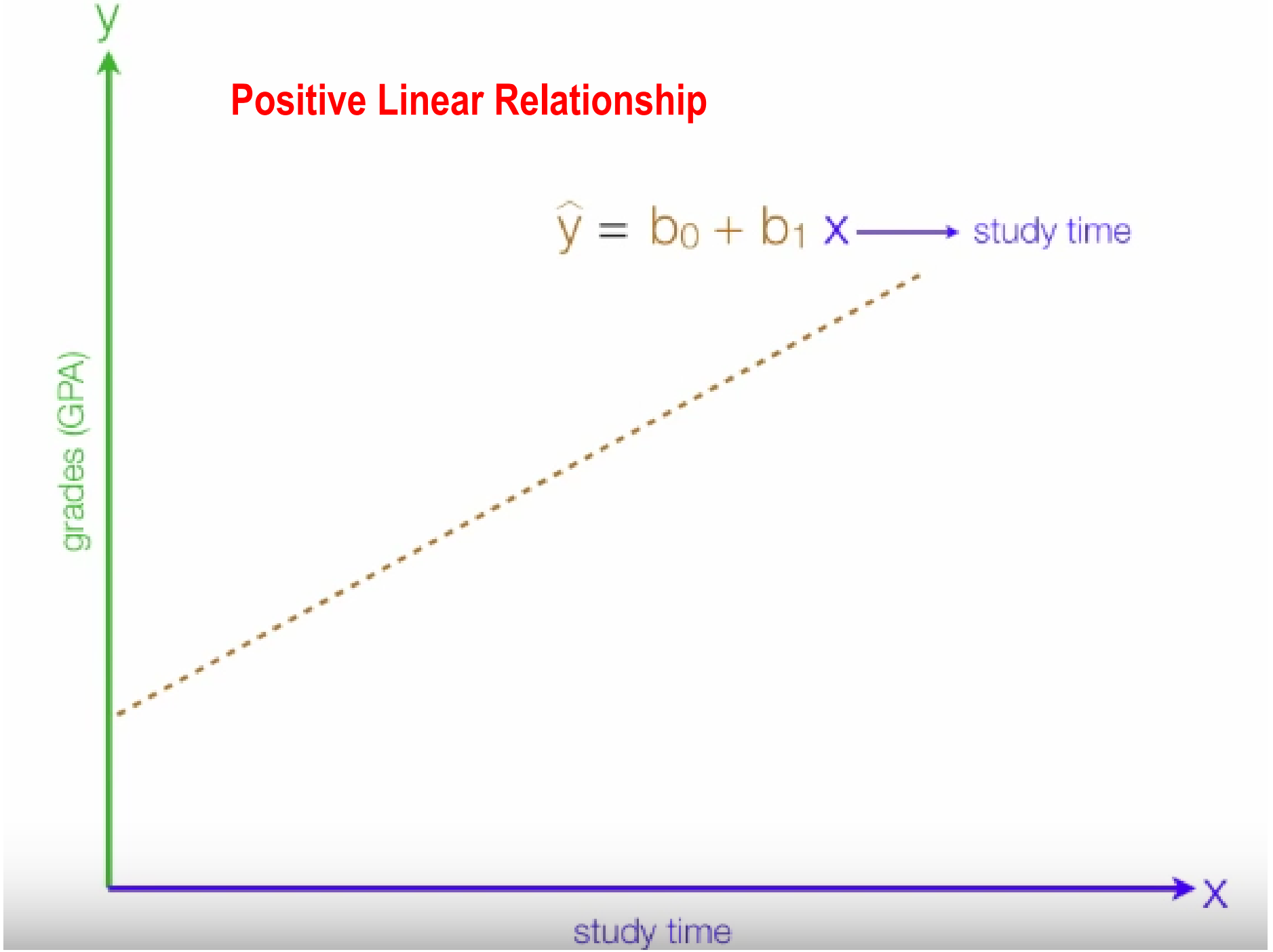
- How well does the model generalize the observations outside the sample?

Simple Linear Regression Equation

Positive Linear Relationship

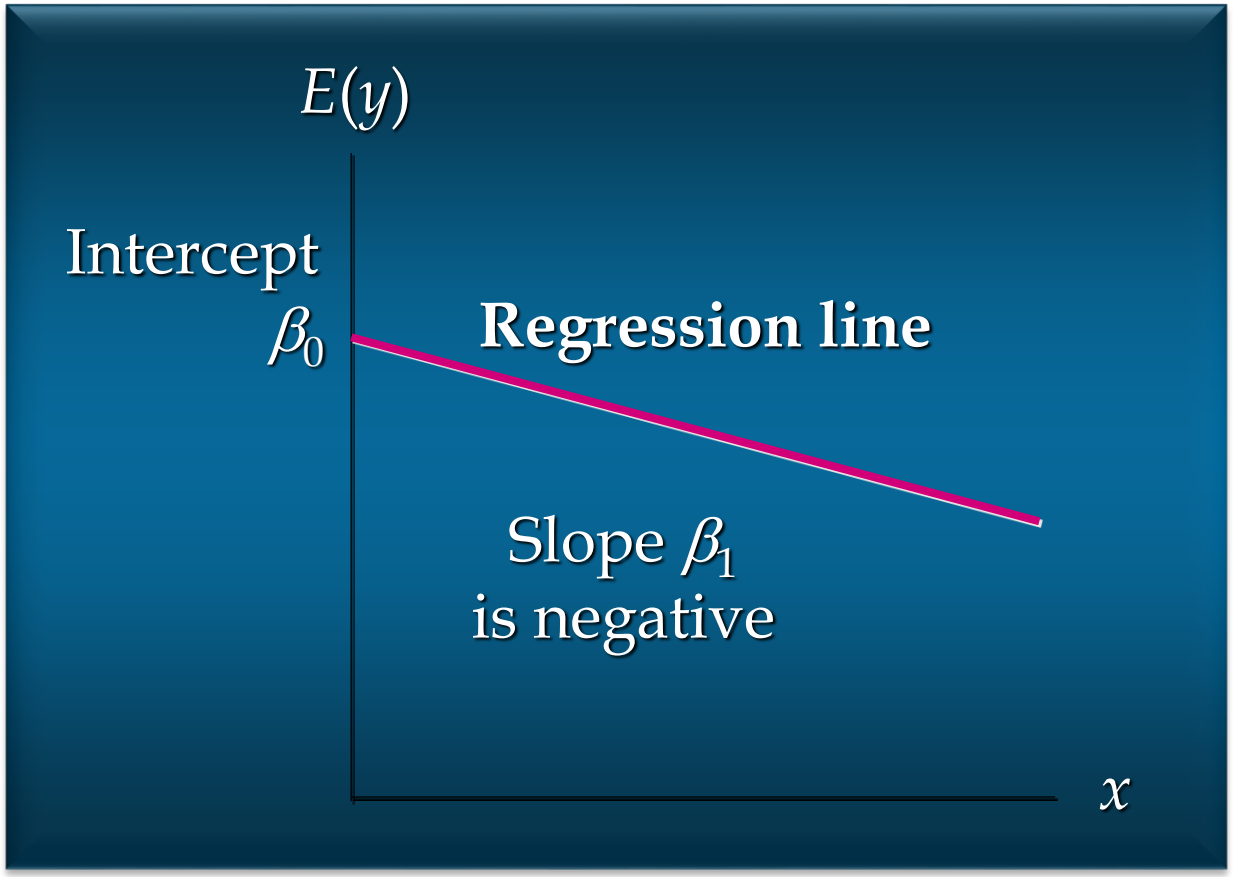


Positive Linear Relationship

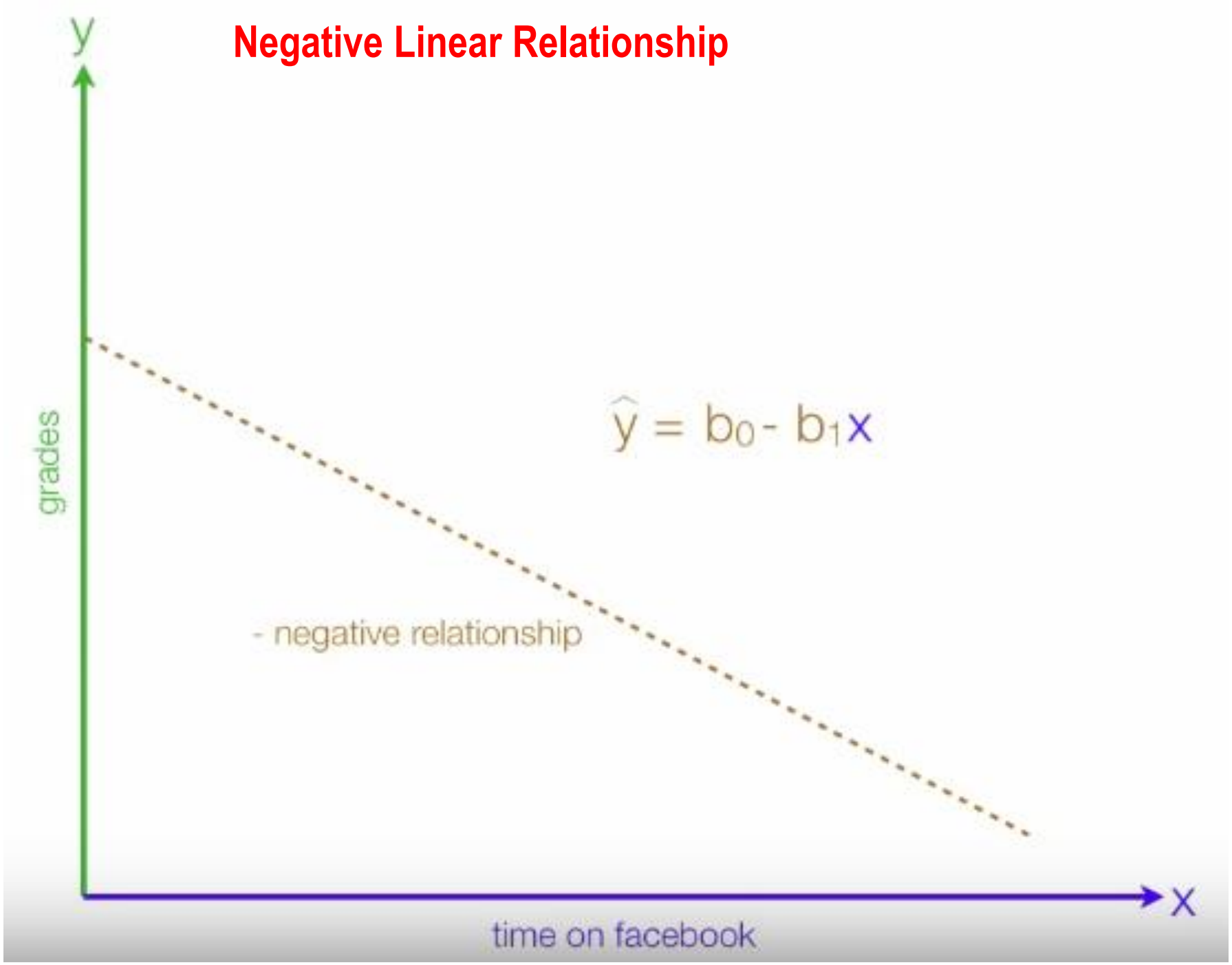


Simple Linear Regression Equation

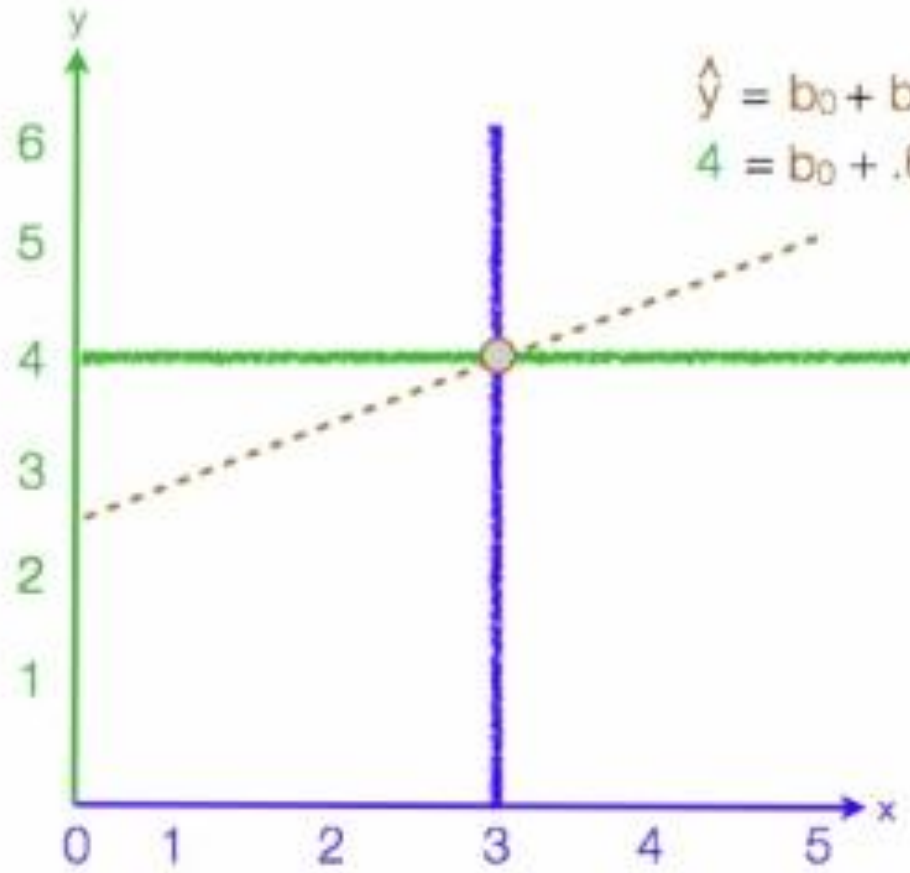
Negative Linear Relationship



Negative Linear Relationship



Simple Linear Regression



$\hat{y} = b_0 + b_1 x$

$4 = b_0 + .6(3)$

$b_0 = 2.2$

$b_1 = .6$

$\hat{y} = 2.2 + .6x$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
				10	6
mean	3	4			

$4 = b_0 + .6(3)$

$4 = b_0 + 1.8$

$-1.8 \quad -1.8$

$2.2 = b_0$

$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$

Estimating Parameters: Least Squares Method

Least Squares

- 1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

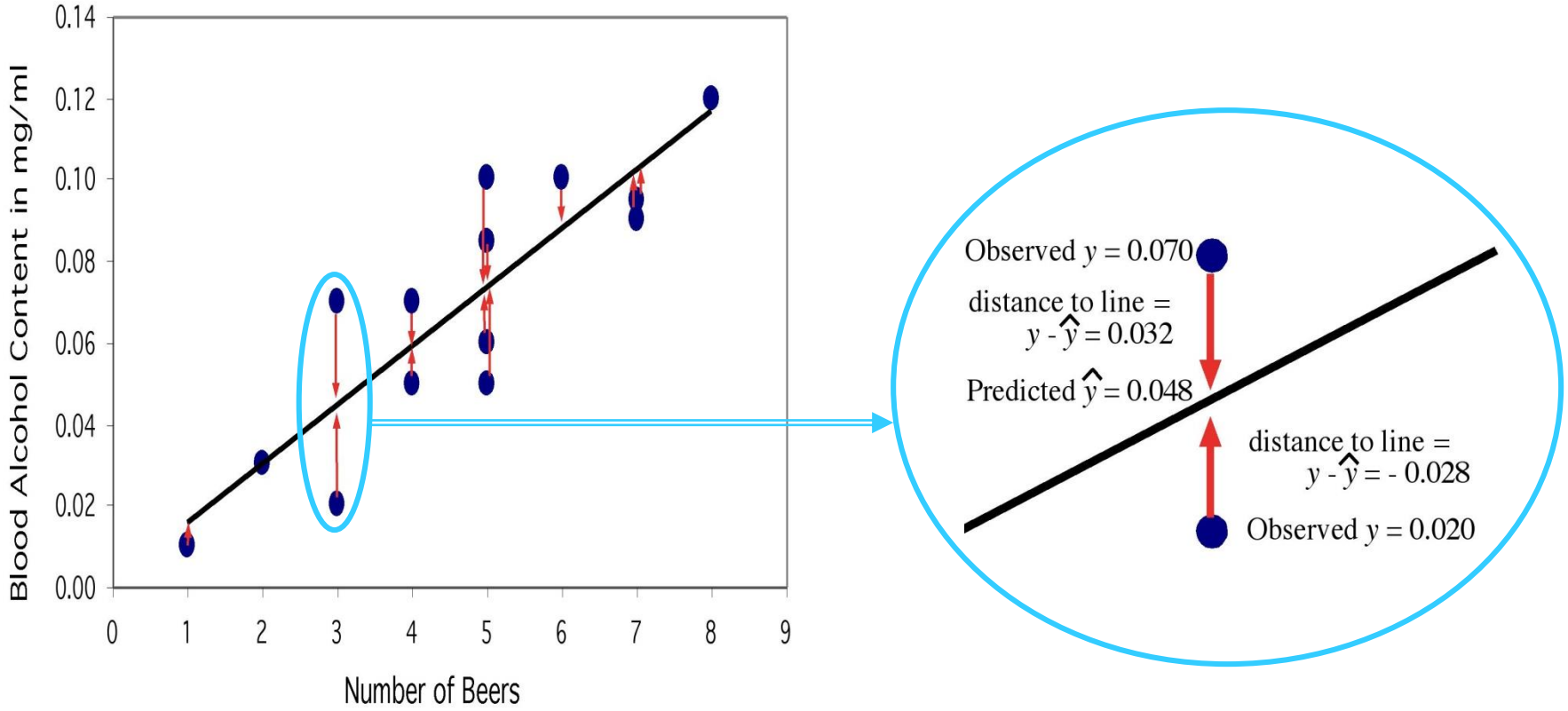
$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

The principle of least squares estimates the parameters B0 and B1 by minimizing the sum of squares of the difference between the observations and the line in the scatter diagram. Such an idea is viewed from different perspectives. When the vertical difference between the observations and the line in the scatter diagram is considered, and its sum of squares is minimized to obtain the estimates of B0 and B1, the method is known as direct regression.

The least-squares regression line

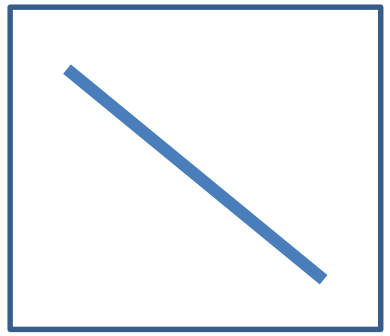
The **least-squares regression line** is the unique line such that the sum of the **vertical distances** between the data points and the line is zero, and the sum of the squared vertical distances is the smallest possible.



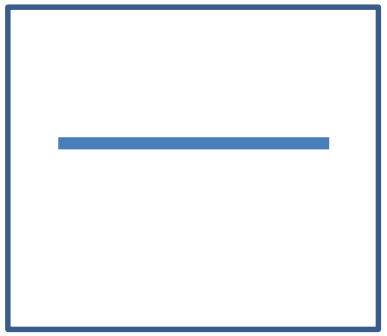
Notation

$\hat{y}$ is the predicted *y* value on the regression line

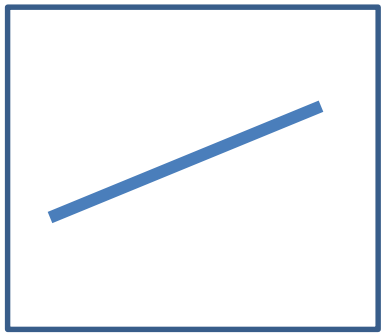
$$\hat{y} = \text{intercept} + \text{slope } x \qquad \hat{y} = a + bx$$



slope < 0



slope = 0



slope > 0

Not all calculators/software use this convention. Other notations include:

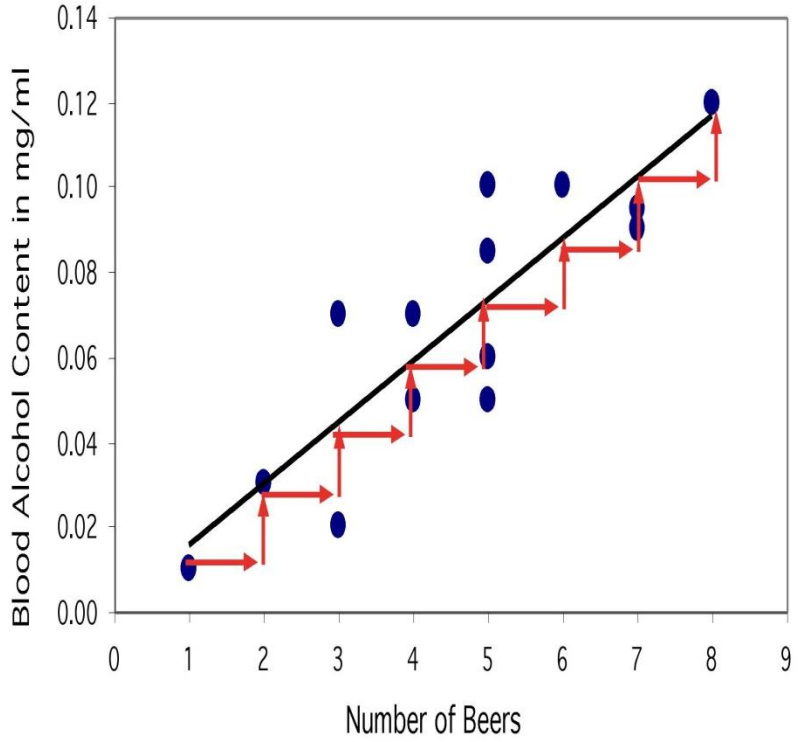
$$\hat{y} = ax + b$$

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = \text{variable\_name } x + \text{constant}$$

Interpretation

The **slope** of the regression line describes how much we expect *y* to change, on average, for every unit change in *x*.



The **intercept** is a necessary mathematical descriptor of the regression line. It does not describe a specific property of the data.

Finding the least-squares regression line

The **slope of the regression line, *b*,** equals: $b = r\frac{s_y}{s_x}$

r is the correlation coefficient between x and y
s_y is the standard deviation of the response variable y
s_x is the standard deviation of the explanatory variable x

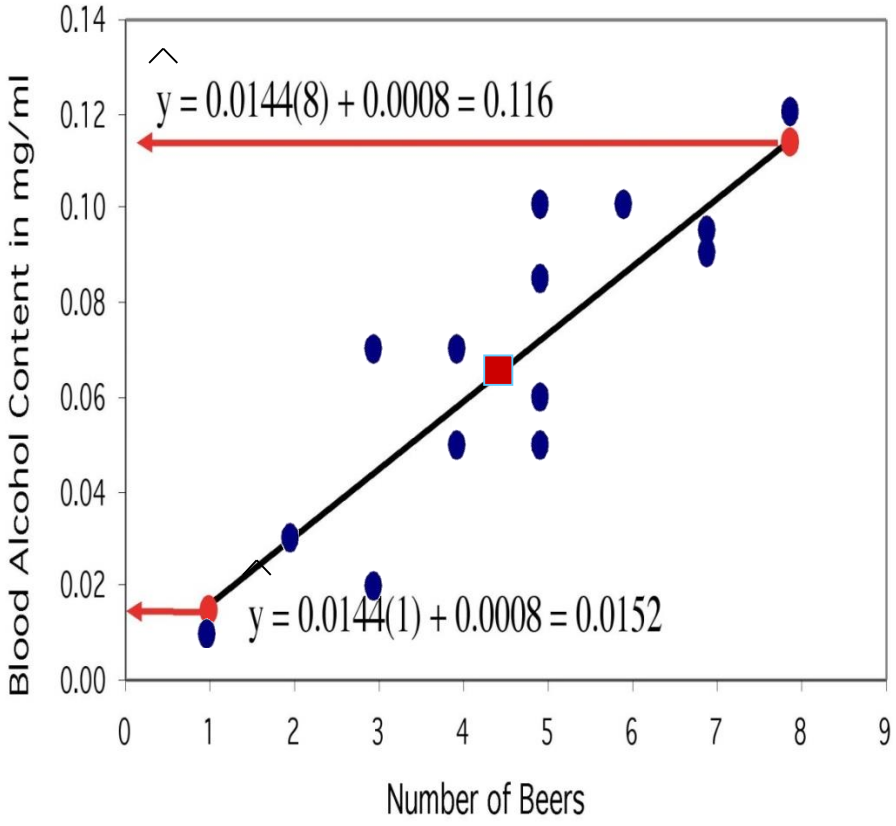
The **intercept, *a*,** equals: $a = \bar{y} - b\bar{x}$

$\bar{x}$ and $\bar{y}$ are the respective means of the x and y variables

Plotting the least-square regression line

Use the regression equation to find the value of y for two distinct values of x, and draw the line that goes through those two points.

Hint: The regression line always passes through the mean of x and y.

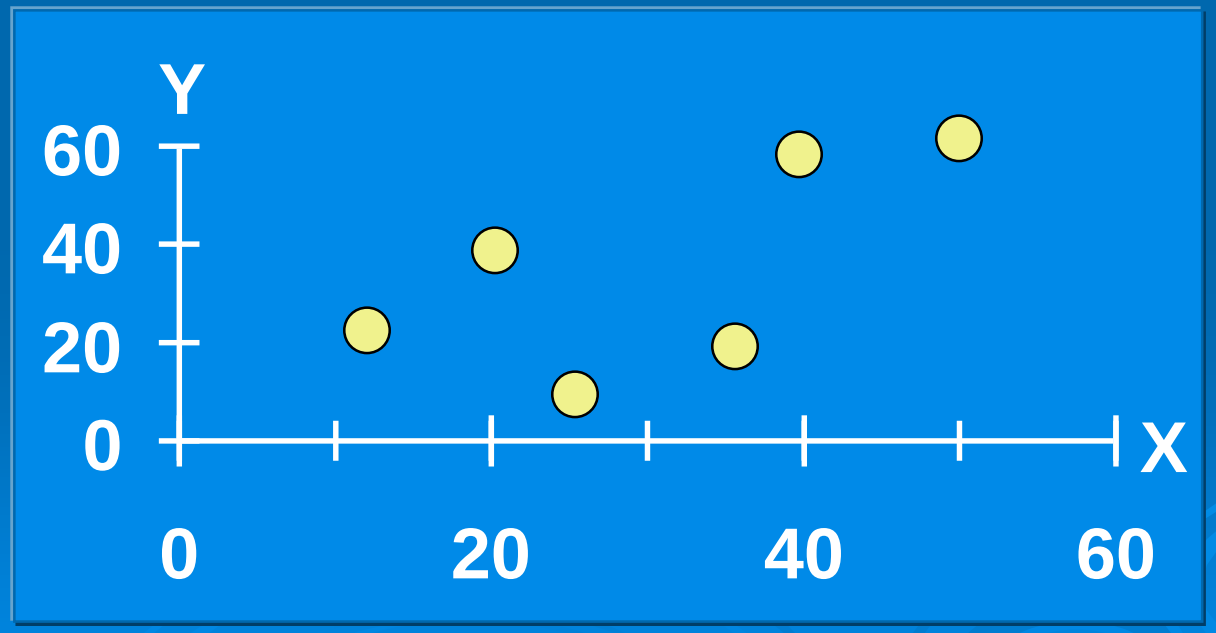


The points used for drawing the regression line are derived from the equation.

They are NOT actual points from the data set (except by pure coincidence).

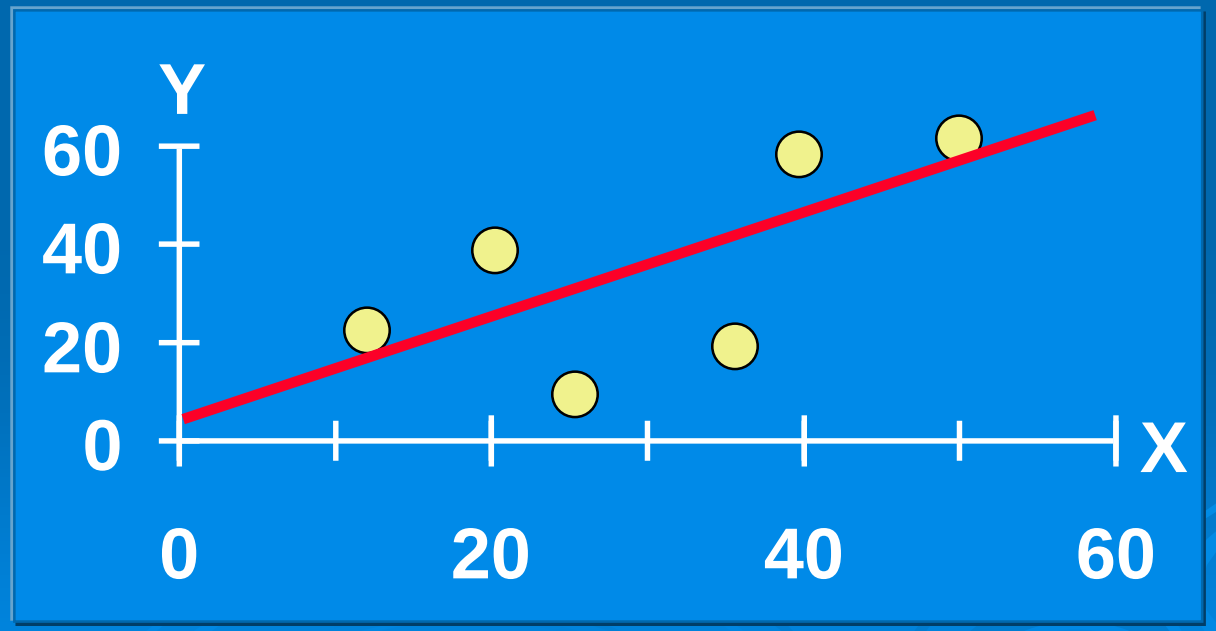
Scatter plot

- 1. Plot of All (X_i, Y_i) Pairs
- 2. Suggests How Well Model Will Fit



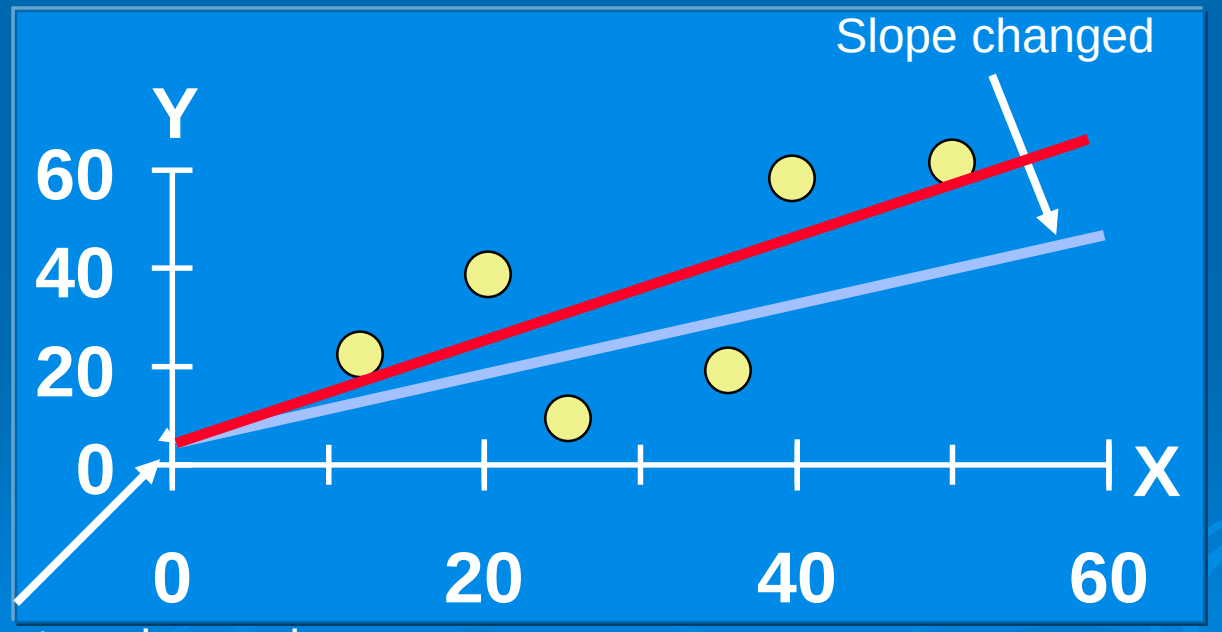
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



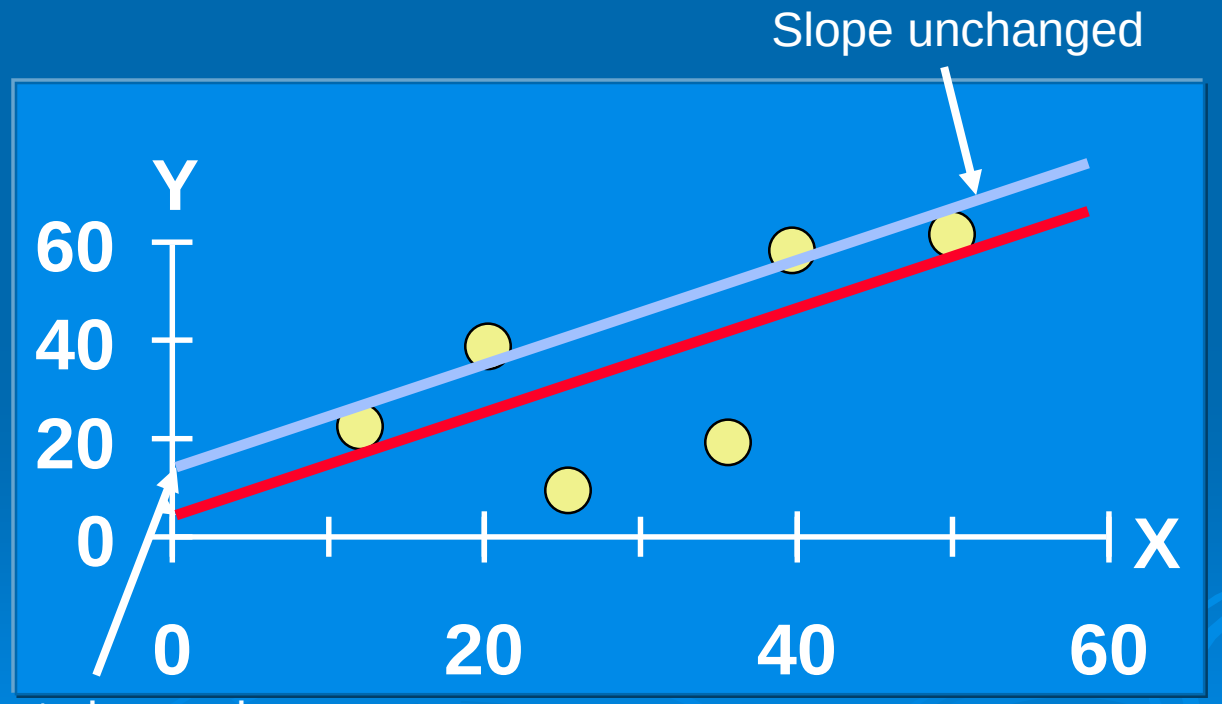
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



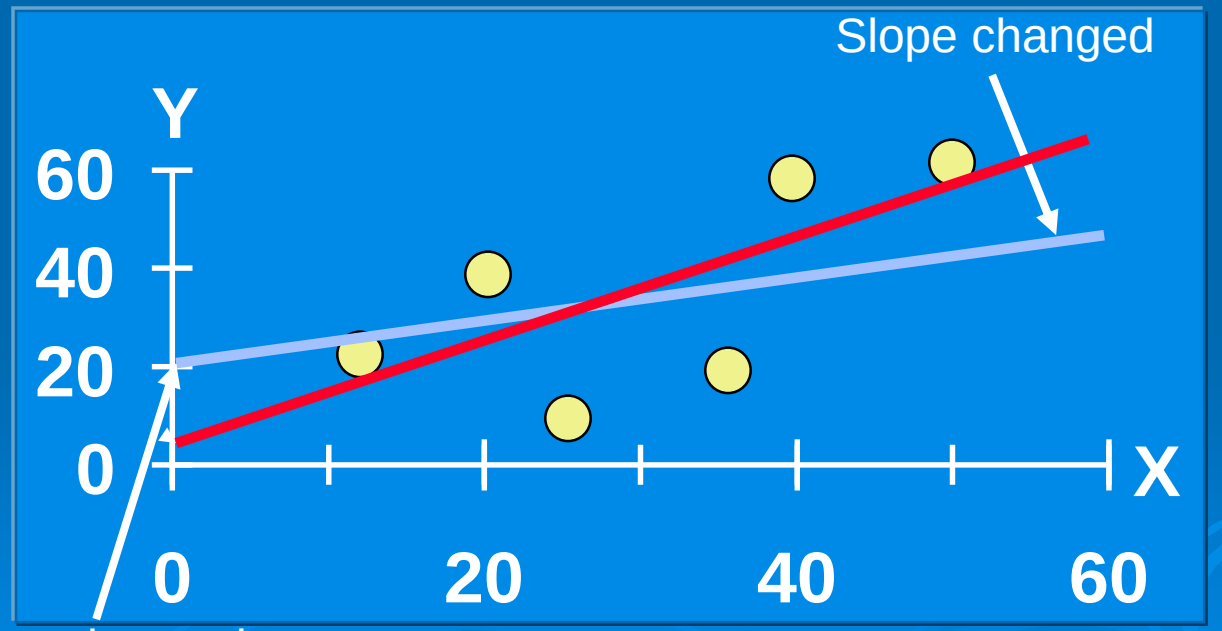
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Least Squares

- The "least squares" method is used to determine the line of best fit for a set of data.
- It also providing a visual demonstration of the relationship between the data points.
- Each point of data represents the relationship between a known independent variable and an unknown dependent variable.

Least Squares

- 1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values are a Minimum. *But* Positive Differences Off-Set Negative ones. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Least Squares

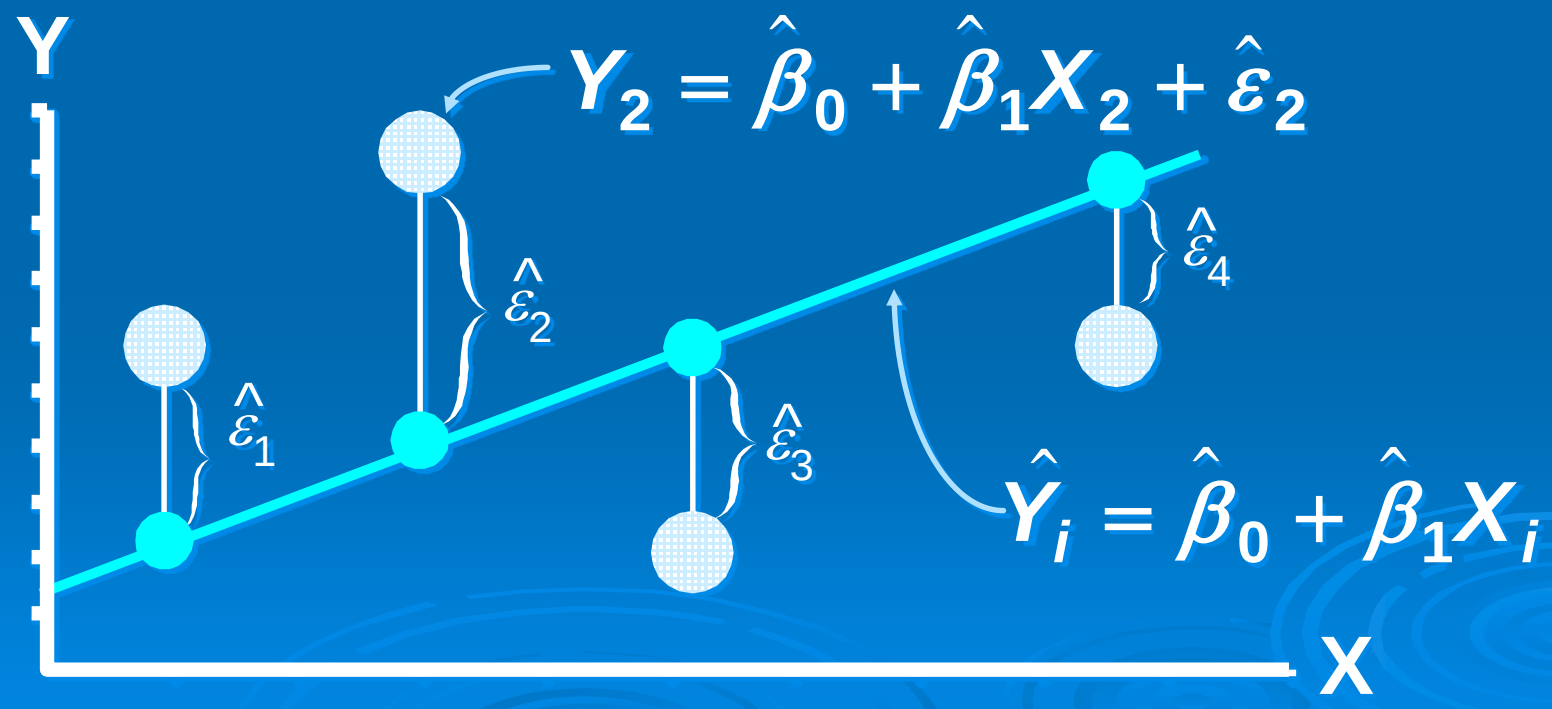
- 1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i\right)^2}{\partial \beta_0}$$

$$= -2\left(n\bar{y} - n\beta_0 - n\beta_1 \bar{x}\right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1}$$

$$= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$= -2 \sum x_i (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i)$$

$$\beta_1 \sum x_i (x_i - \bar{x}) = \sum x_i (y_i - \bar{y})$$

$$\beta_1 \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Ssxy is the “sum of squares” for each pair of observations x and y and SSxx is the “sum of squares” for each x observation.

Computation Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Interpretation of Coefficients

- 1. Slope ($\hat{\beta}_1$)
  - Estimated *Y* Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in *X*
    - If $\hat{\beta}_1 = 2$, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*

Interpretation of Coefficients

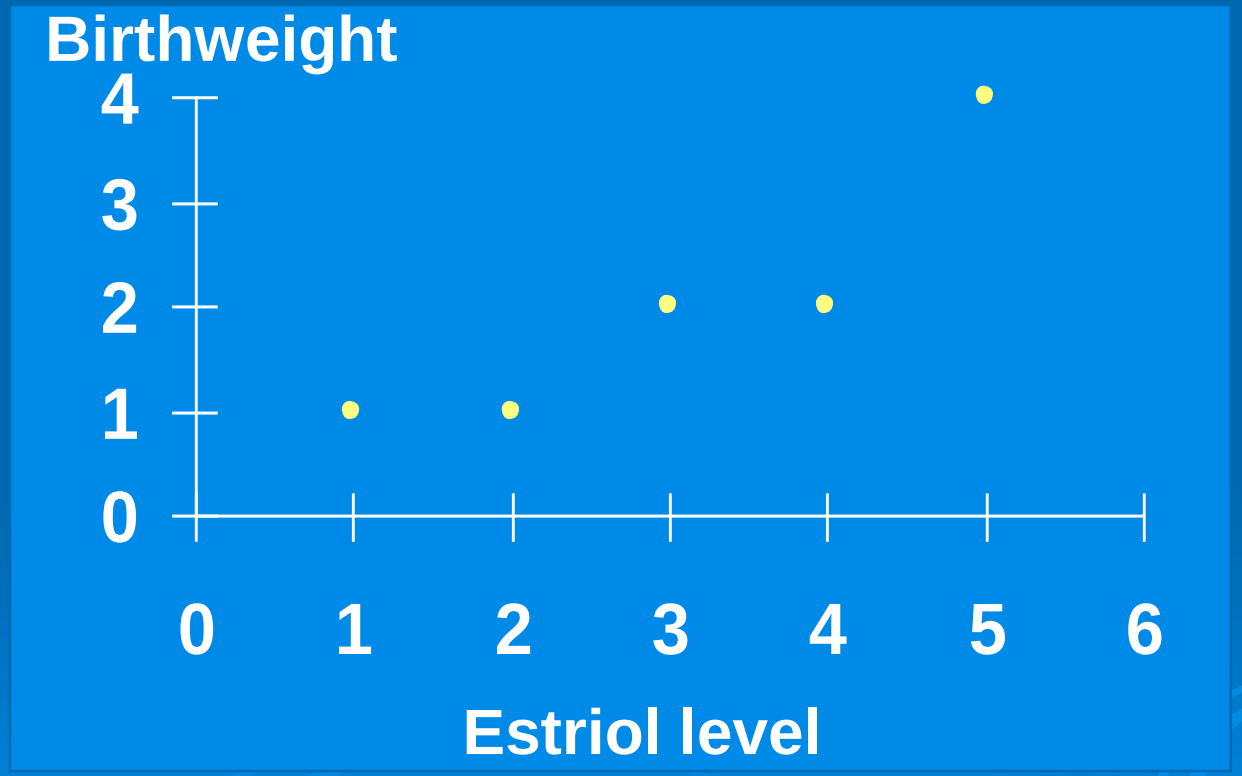
- 1. Slope ($\hat{\beta}_1$)
  - Estimated *Y* Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in *X*
    - If $\hat{\beta}_1 = 2$, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*
- 2. Y-Intercept ($\hat{\beta}_0$)
  - Average Value of *Y* When *X* = 0
    - If $\hat{\beta}_0 = 4$, then Average *Y* Is Expected to Be 4 When *X* Is 0

Parameter Estimation Example

- Obstetrics: What is the **relationship** between Mother's Estriol level & Birthweight using the following data?

Estriol (mg/24h)	**Birthweight (g/1000)**
1	**1**
2	**1**
3	**2**
4	**2**
5	**4**

Scatterplot
Birthweight vs. Estriol level



Parameter Estimation Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

Parameter Estimation Solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = 0.70$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2 - (0.70)(3) = -0.10$$

Coefficient Interpretation Solution

Coefficient Interpretation Solution

- 1. Slope ($\hat{\beta}_1$)
  - Birthweight (Y) Is Expected to Increase by .7 Units for Each 1 unit Increase in Estriol (X)

Coefficient Interpretation Solution

- 1. Slope ($\hat{\beta}_1$)
  - Birthweight (Y) Is Expected to Increase by .7 Units for Each 1 unit Increase in Estriol (X)
- 2. Intercept ($\hat{\beta}_0$)
  - Average Birthweight (Y) Is -.10 Units When Estriol level (X) Is 0
    - Difficult to explain
    - The birthweight should always be positive

Simple Linear Regression

- Example: Reed Auto Sales

Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.

Simple Linear Regression

- Example: Reed Auto Sales

Number of TV Ads (x)	Number of Cars Sold (y)
1	14
3	24
2	18
1	17
3	27
$\Sigma x = 10$	$\Sigma y = 100$
$\bar{x} = 2$	$\bar{y} = 20$

Estimated Regression Equation

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{20}{4} = 5$$

- y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x} = 20 - 5(2) = 10$$

- Estimated Regression Equation

$$\hat{y} = 10 + 5x$$

Coefficient of Determination

- Relationship Among SST, SSR, SSE

SST = SSR + SSE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Coefficient of Determination

- The coefficient of determination is:

  - $r^2 = SSR/SST$

where:

SSR = sum of squares due to regression

SST = total sum of squares

r^2 gives information about the goodness of fit of a model.
In regression, the r^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points.
An r^2 of 1 indicates that the regression line perfectly fits the data.

Coefficient of Determination

- r^2 = SSR/SST = 100/114 = .8772

- The regression relationship is very strong; 87.72% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

Correlation coefficient

- The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0.
- A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.

- For example, It determine the level of correlation between the price of crude oil and the stock price of an oil-producing company, such as Exxon Mobil Corporation. Since oil companies earn greater profits as oil prices rise, the correlation between the two variables is highly positive.

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$$

where:

b_1 = the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$

Sample Correlation Coefficient

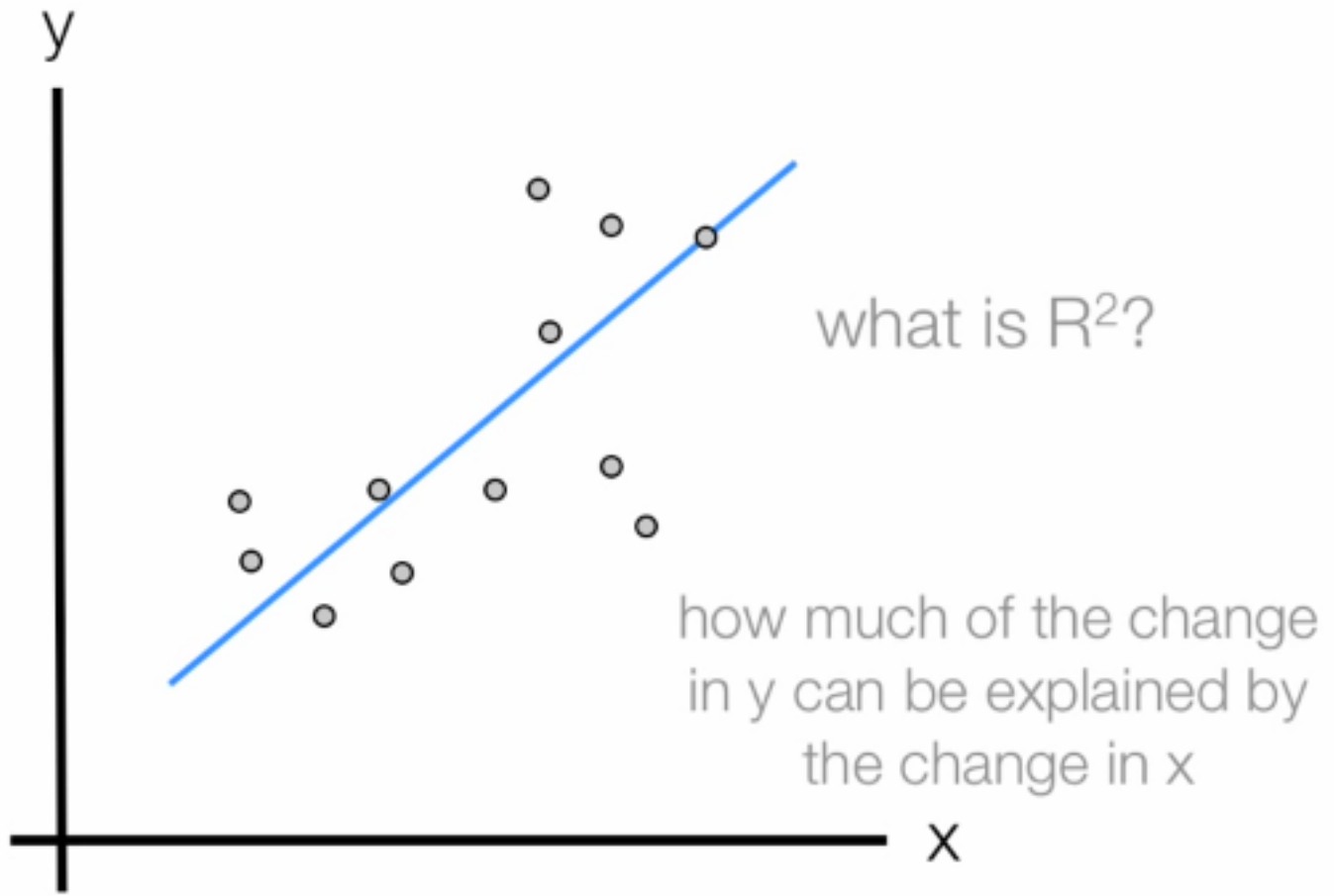
- $r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$
- The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is "+".
  - $r_{xy} = +\sqrt{.8772}$
  - $r_{xy} =$ +.9366

when you see r,
think relationship!

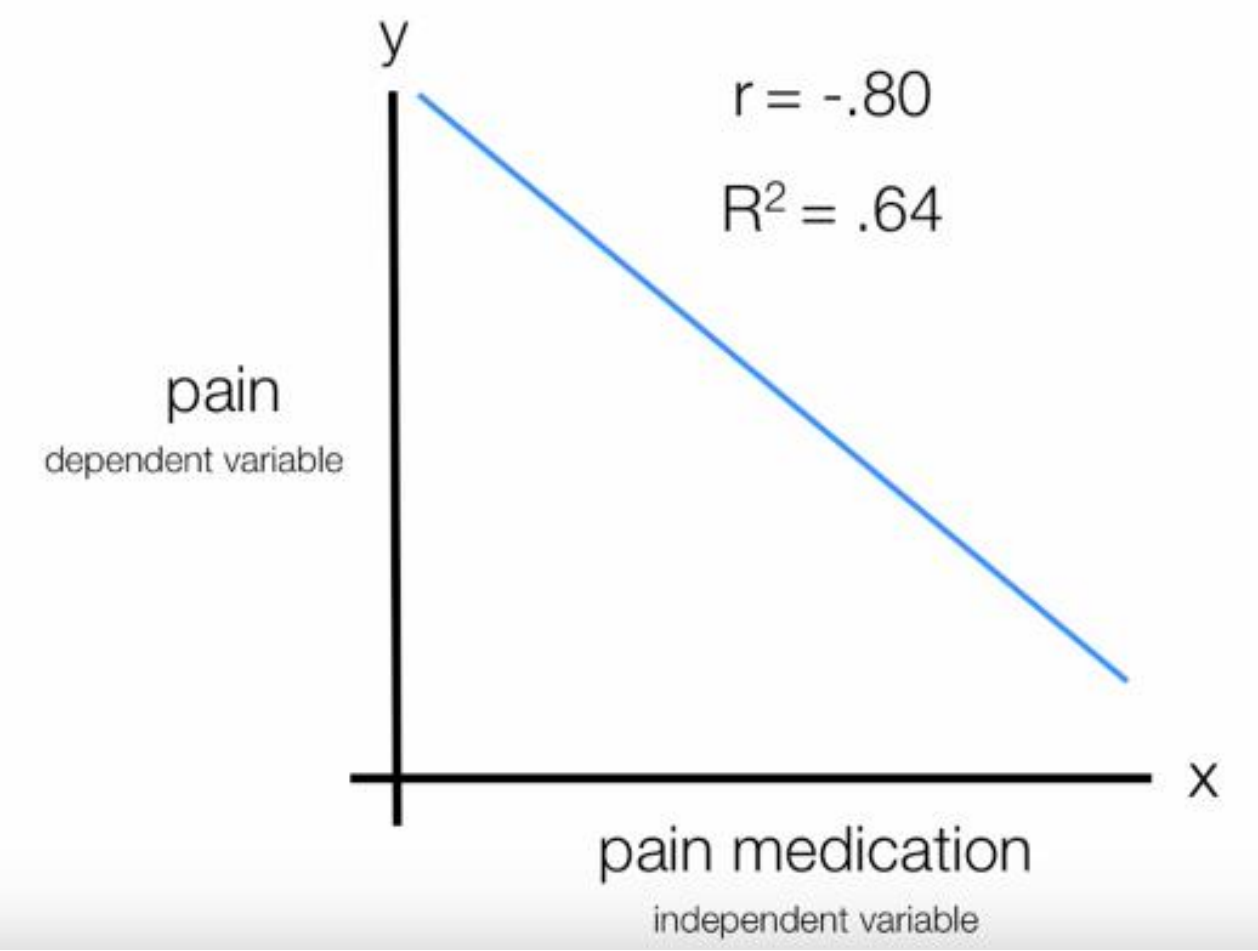
r = 1, perfect relationship

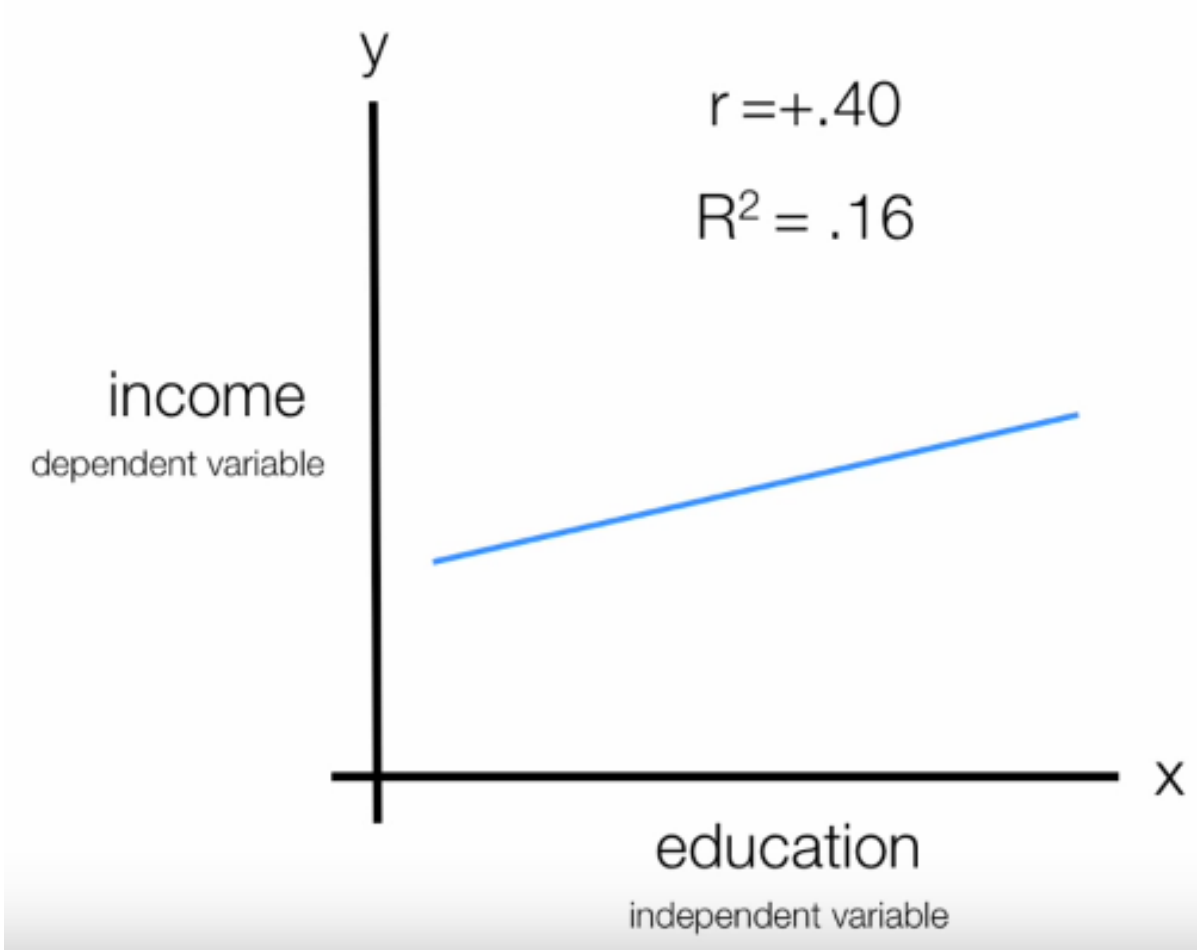
r = 0, no relationship

r = -1, negative relationship



when x changes, y changes

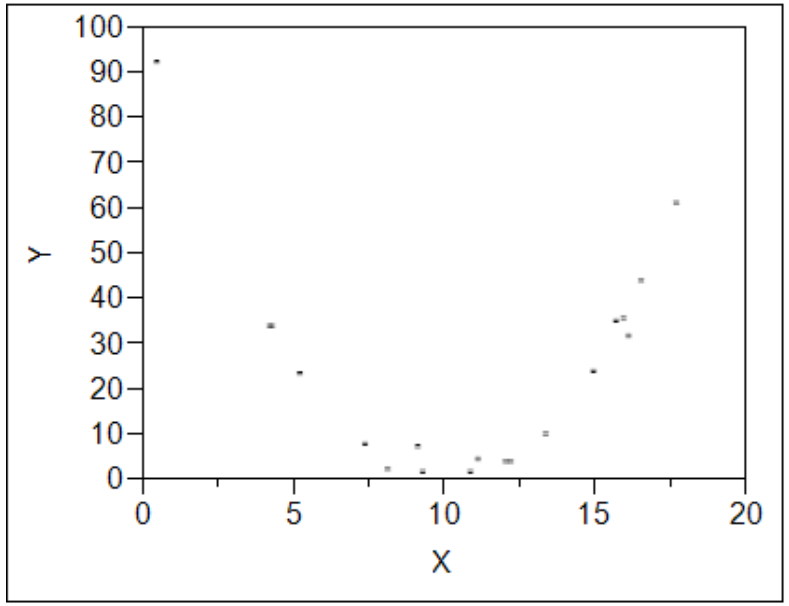


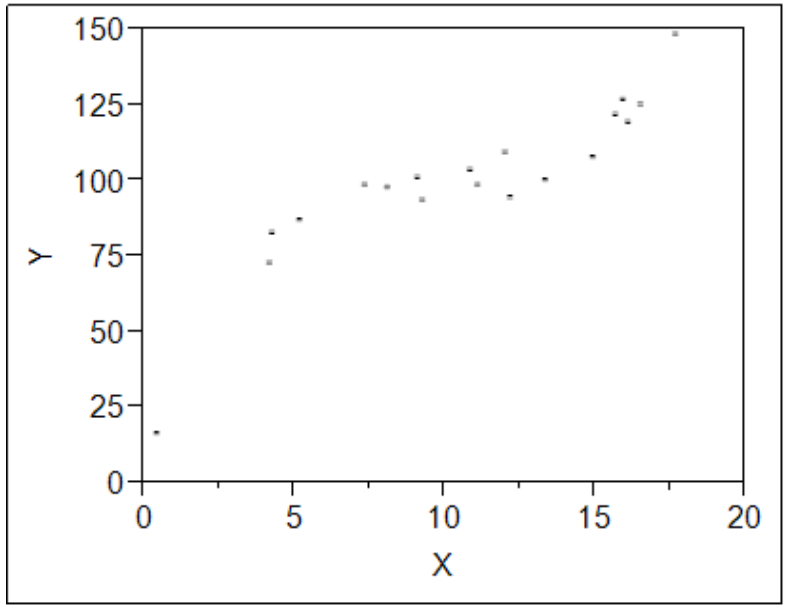


R^2 shows the intensity or how strong is the relationship.

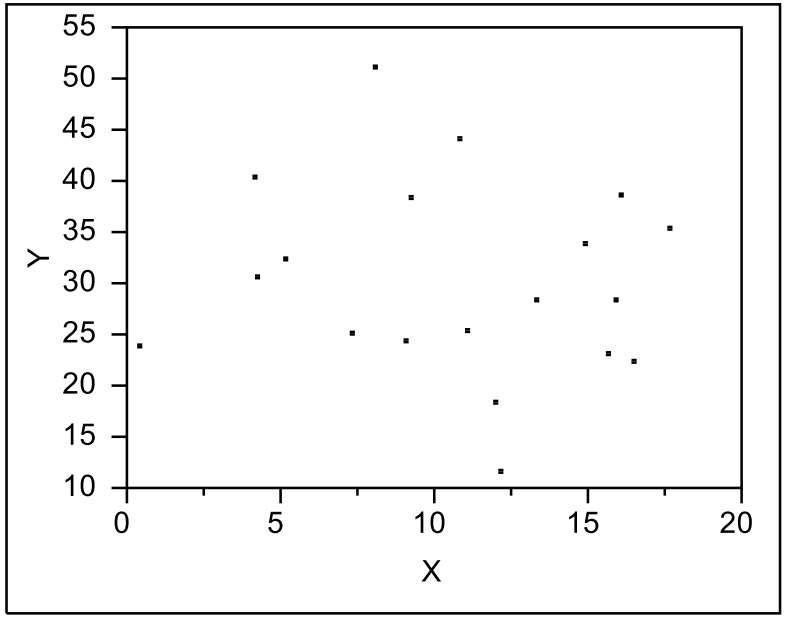
Types of Relationships Between Two Continuous Variables

- Curved Relationship





- No Relationship



	x	y	x^2	y^2	xy
	-1	-1	1	1	1
	1	2	1	4	2
	2	3	4	9	6
	4	3	16	9	12
	6	5	36	25	30
	7	8	49	64	56
Σ	19	20	107	112	107

$\Sigma x, \Sigma y, \Sigma x^2, \Sigma y^2, \Sigma xy$

$n = 6$

$y = \tilde{a}x + \tilde{b}$

$y \approx 0.9324x + 0.3808$

$$\hat{a} = \frac{(\Sigma x)(\Sigma y) - n\Sigma xy}{(\Sigma x)^2 - n\Sigma x^2} = \frac{(19)(20) - 6(107)}{(19)^2 - 6(107)} = \frac{-262}{-281} \approx 0.9324$$

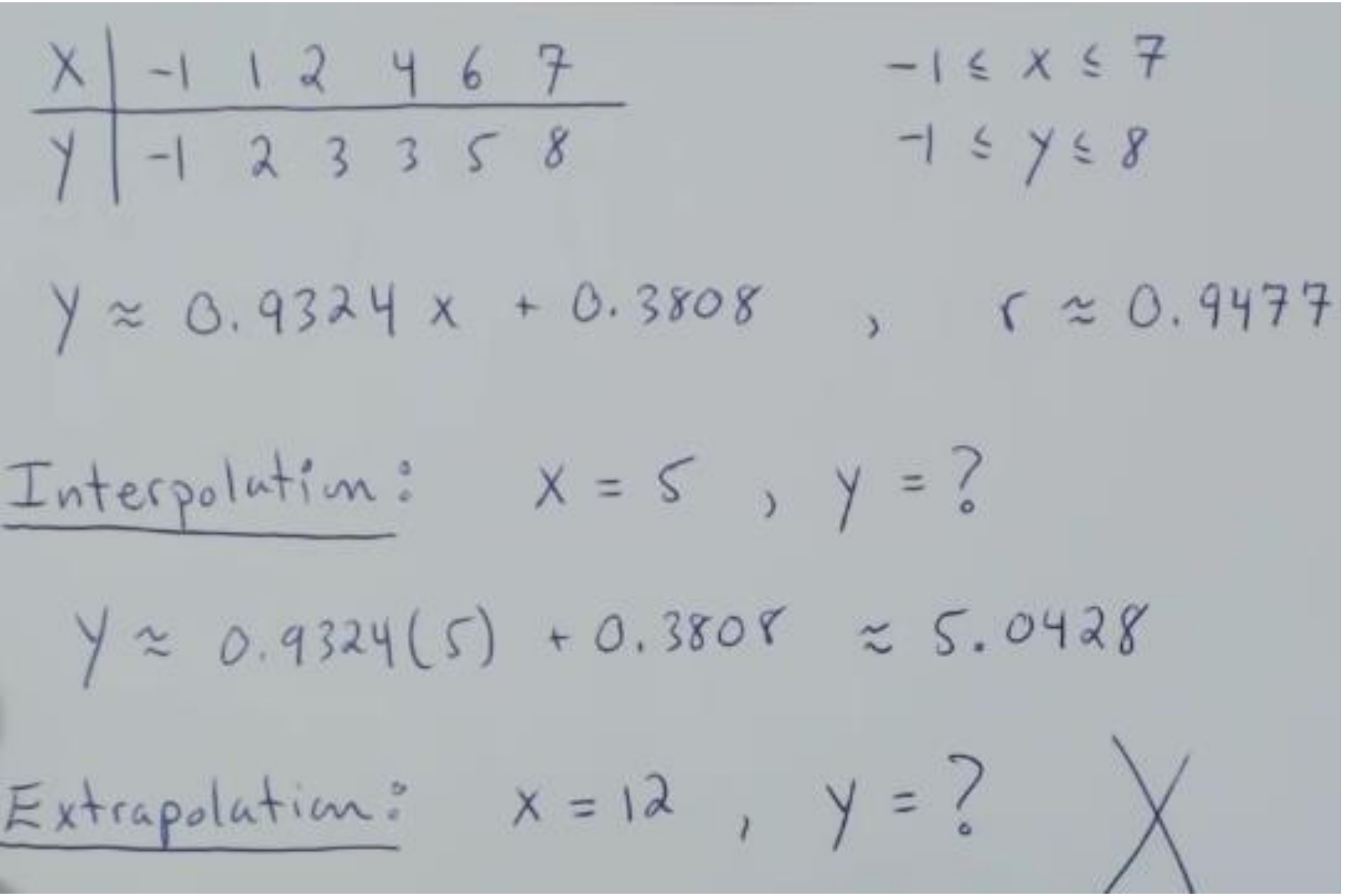
$$\hat{b} = \frac{(\Sigma x)(\Sigma xy) - (\Sigma y)(\Sigma x^2)}{(\Sigma x)^2 - n\Sigma x^2} = \frac{(19)(107) - (20)(107)}{(19)^2 - 6(107)} = \frac{-107}{-281} \approx 0.3808$$

$$s_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1}} = \sqrt{\frac{107 - \frac{1}{6}(19)^2}{5}} \approx 3.0605$$

$$s_y = \sqrt{\frac{112 - \frac{1}{6}(20)^2}{5}} \approx 3.0111$$

$$r = \left(\frac{s_x}{s_y}\right)\hat{a} \approx 0.9477 \quad (\approx 1)$$

$$r = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{(n-1)\, s_x\, s_y} \approx 0.9477$$



Extrapolation is an estimation of a value based on extending a known sequence of values or facts beyond the area that is certainly known.
Extrapolate is to infer something that is not explicitly stated from existing information.

Interpolation is an estimation of a value within two known values in a sequence of values.
Polynomial interpolation is a method of estimating values between known data points.

When graphical data contains a gap, but data is available on either side of the gap or at a few specific points within the gap, interpolation allows us to estimate the values within the gap.