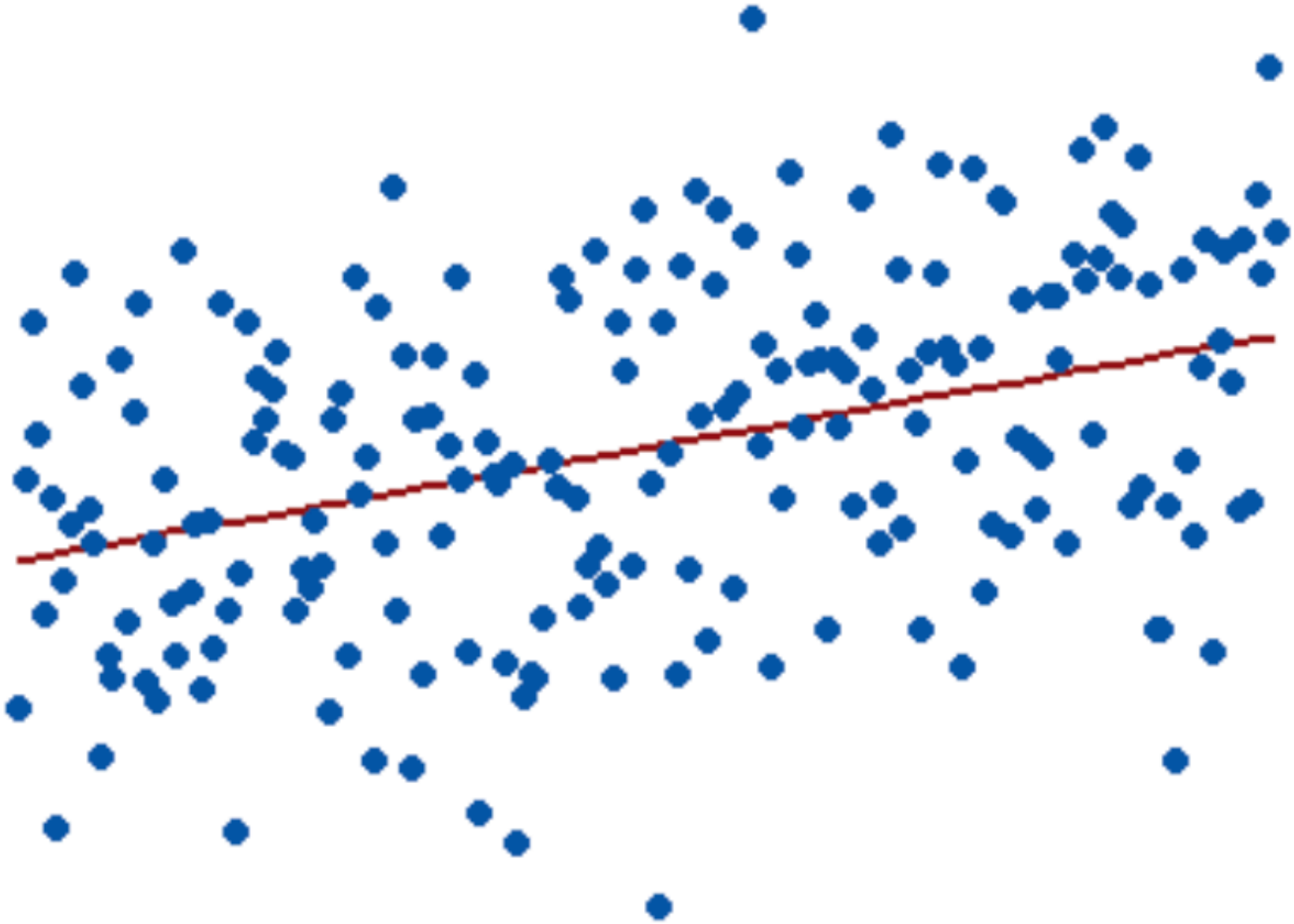

regression
variables
analysis
independent
model
parameters
beta
data
assumptions
values
dependent
Regression
Jump
linear
used
Statistical
function
unknown
methods
squares
least
may
edit

# Regression Analysis

- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression
- Panel Data Regression
- Non-linear Regression



Try to fit a linear equation to data

# Simple Linear


- The **simple linear regression** is used to predict a quantitative outcome y on the basis of one single predictor variable x.
- The goal is to build a mathematical model (an equation) that defines y as a function of the x variable
- it's possible to use it for predicting future outcome on the basis of new x values.

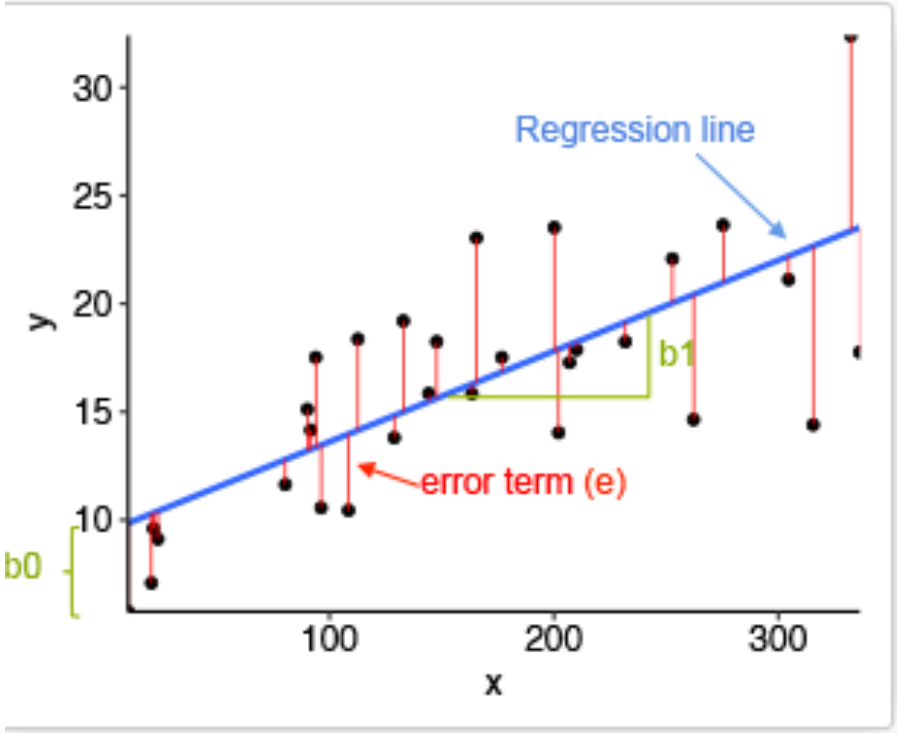
http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/

# Simple Linear Regression

# $y = b0 + b1*x + e$

- x is the predictor or independent variable
- y is the response or dependent variable
- b0 and b1 are the regression beta coefficients
- b0 is the intercept of the regression line; the predicted value when x = 0.
- b1 is the slope of the regression line.
- e is the error term (also known as the residual error), the part of y that can be explained by the regression model

# Simple Linear Regression




- The best-fit regression line is in blue
- The intercept (b0) and the slope (b1) are shown in green
- The error terms (e) are represented by vertical red lines

# Simple Linear Regression

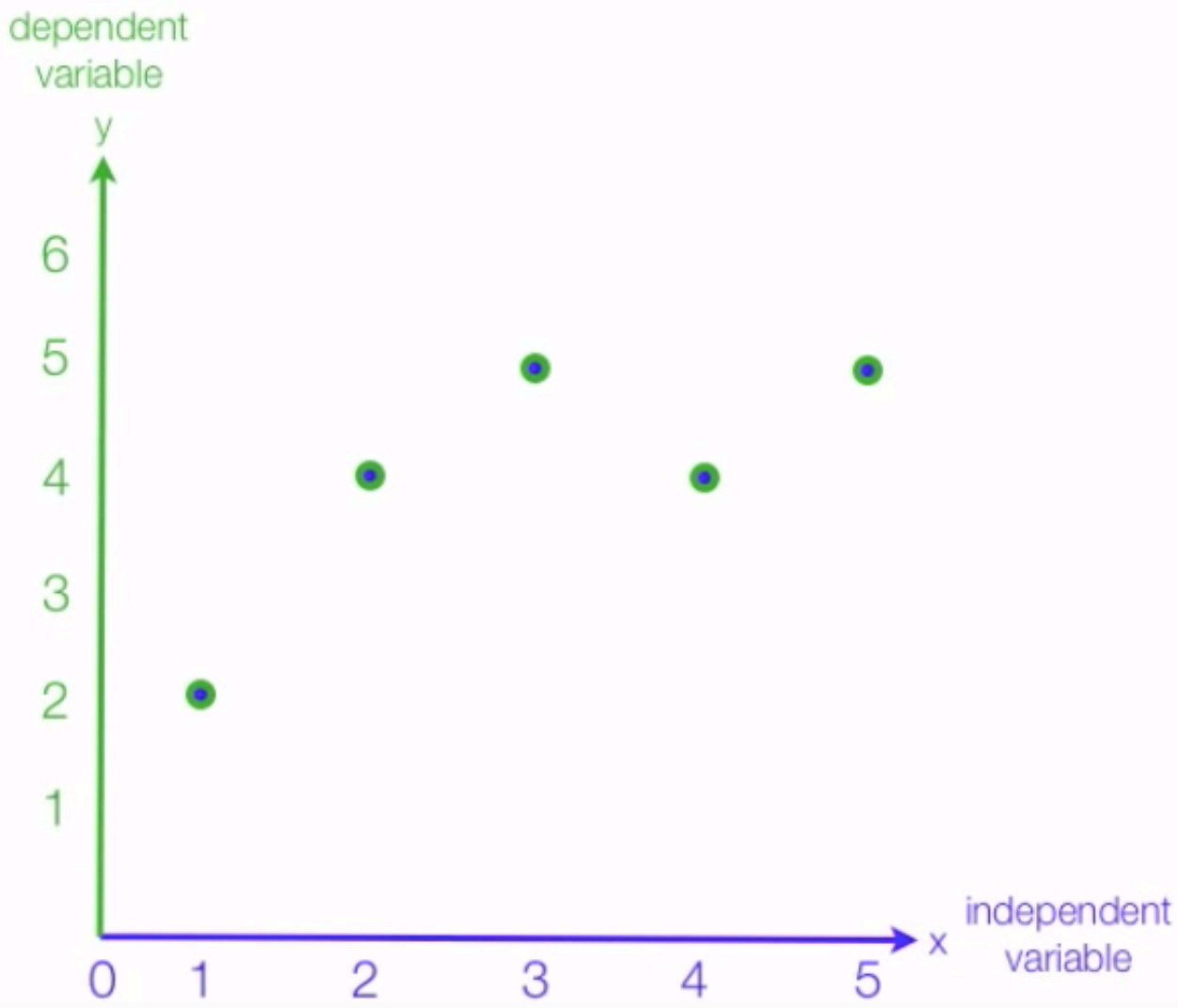
- Not all the data points fall exactly on the fitted regression line
- Some of the points are above the blue curve and some are below it
- Overall, the residual errors (e) have approximately mean zero.
- The sum of the squares of the residual errors are called the **Residual Sum of Squares or RSS**
- The average variation of points around the fitted regression line is called the **Residual Standard Error (RSE)**
- This is one the metrics used to evaluate the overall quality of the fitted regression model: The lower the RSE, the better it is

# Simple Linear Regression

- Since the mean error term is zero, the outcome variable y can be approximately estimated as follow: **y ~ b0 + b1*x**
- The beta coefficients (b0 and b1) are determined so that the RSS is as minimal as possible. This is an optimization problem
- This method of determining the beta coefficients is technically called **least squares regression** or **ordinary least squares (OLS) regression**.
- Once, the beta coefficients are calculated, a **t-test** is performed to check whether or not these coefficients are significantly different from zero
- A non-zero beta coefficients means that there is a significant relationship between the predictors (x) and the outcome variable (y).

# Tutorial on Simple Linear Regression

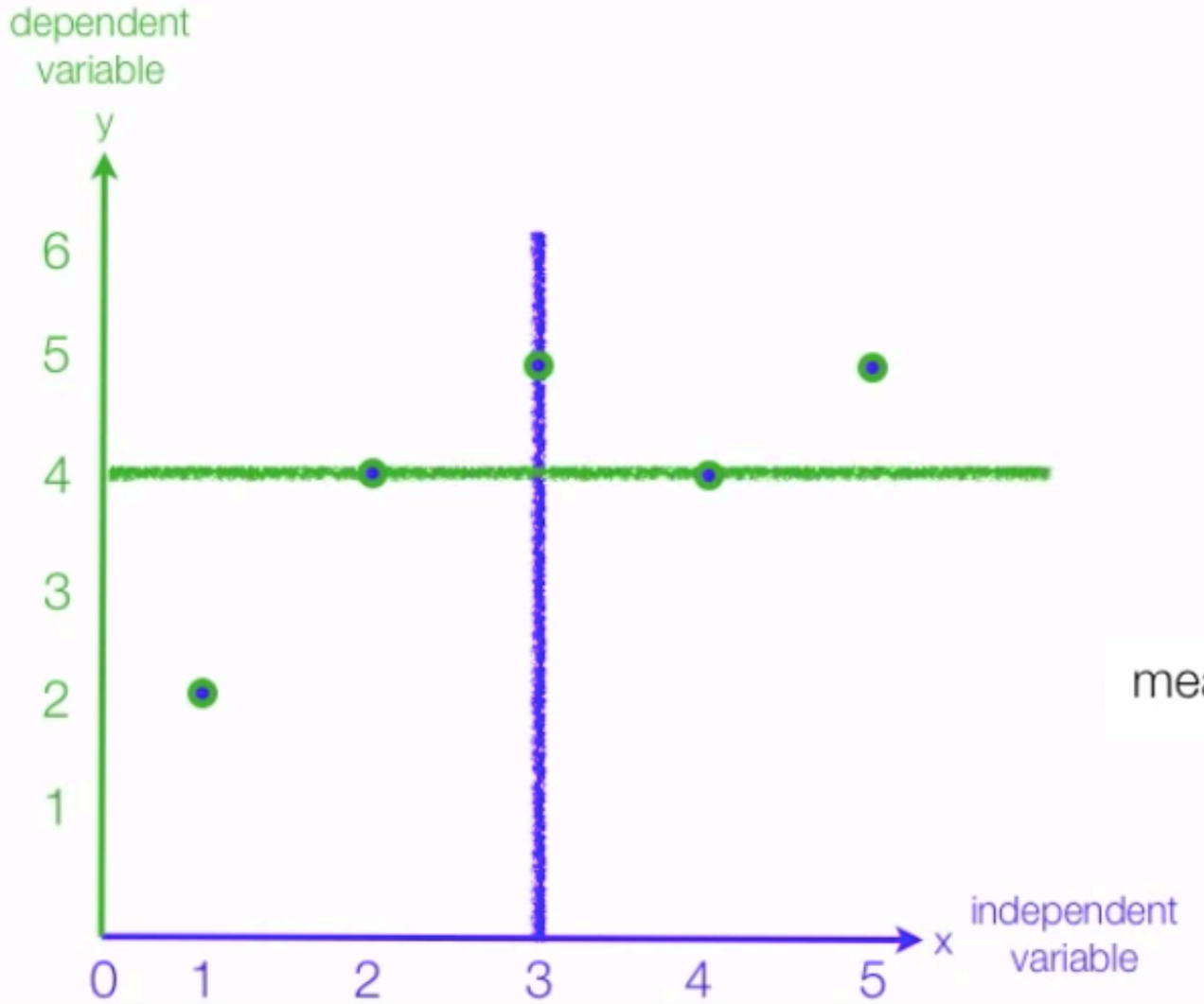
# Simple Linear Regression Analysis




| independent variable | dependent variable |
|---|---|
| x | y |
| 1 | 2 |
| 2 | 4 |
| 3 | 5 |
| 4 | 4 |
| 5 | 5 |

https://www.youtube.com/watch?v=JvS2triCgOY

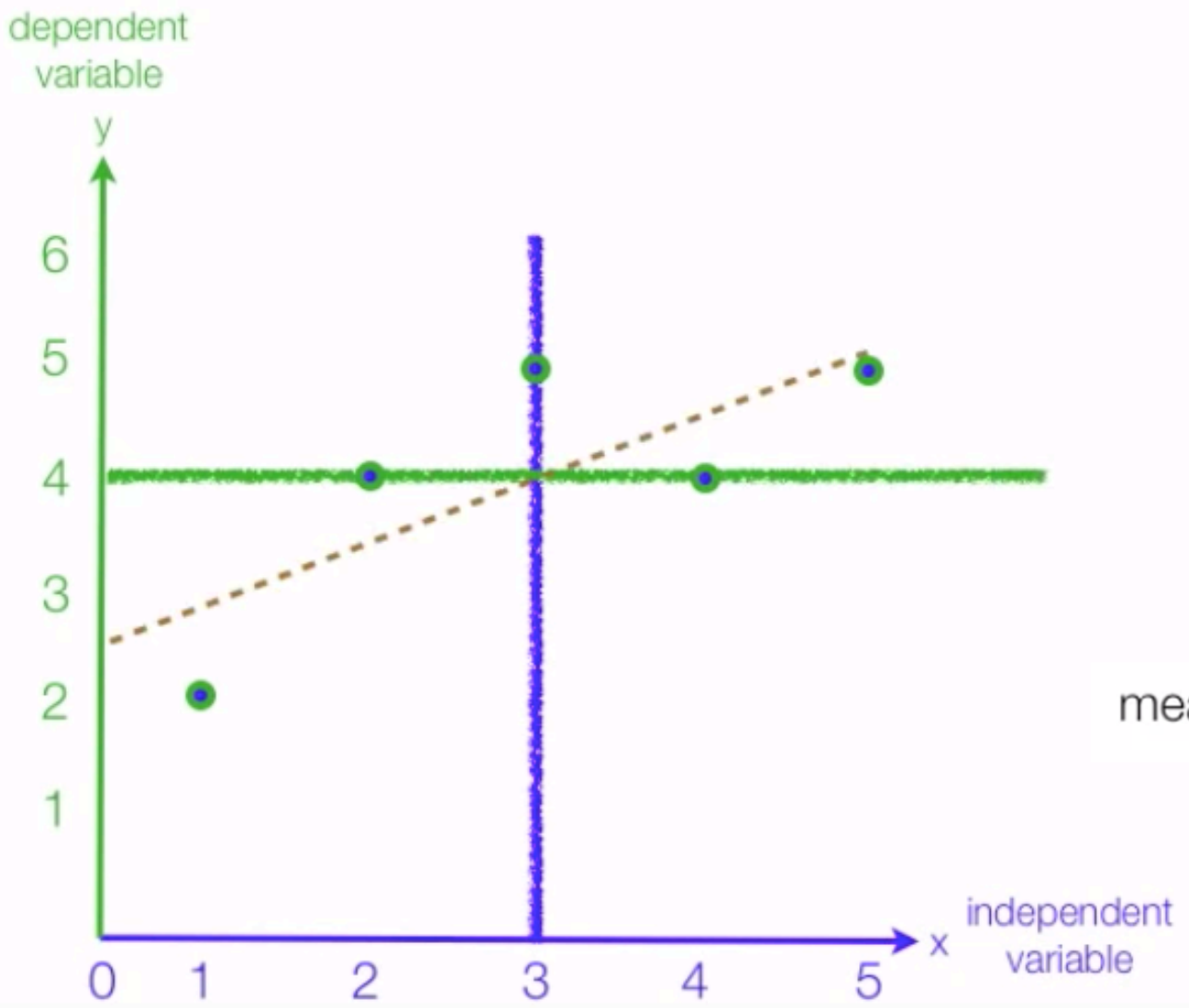
# Simple Linear Regression Analysis




| independent variable | dependent variable |
|---|---|
| x | y |
| 1 | 2 |
| 2 | 4 |
| 3 | 5 |
| 4 | 4 |
| 5 | 5 |

mean $\bar{x} = 3$ $\bar{y} = 4$

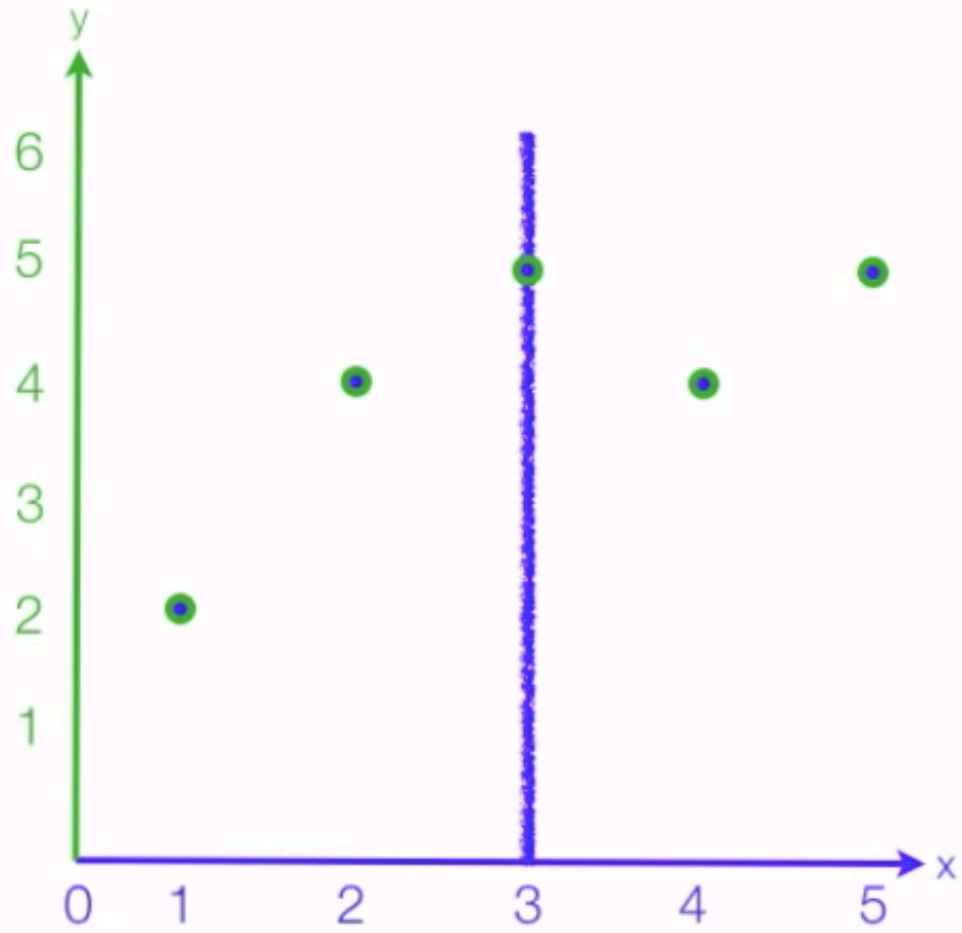
# Simple Linear Regression Analysis




| independent variable | dependent variable |
|---|---|
| x | y |
| 1 | 2 |
| 2 | 4 |
| 3 | 5 |
| 4 | 4 |
| 5 | 5 |

mean $\bar{x} = 3$ $\bar{y} = 4$

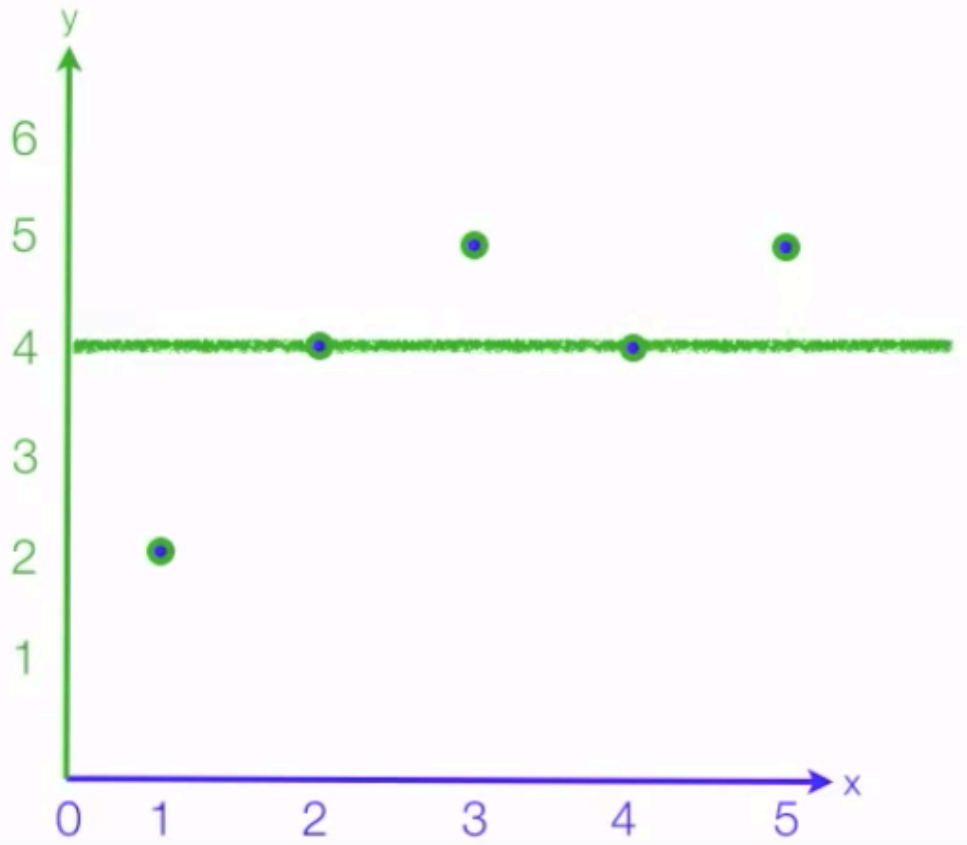
# Simple Linear Regression Analysis




| x | y | $x - \bar{x}$ |
|---|---|---|
| 1 | 2 | -2 |
| 2 | 4 | -1 |
| 3 | 5 | 0 |
| 4 | 4 | 1 |
| 5 | 5 | 2 |

mear $\bar{x} = 3$ $\bar{y} = 4$

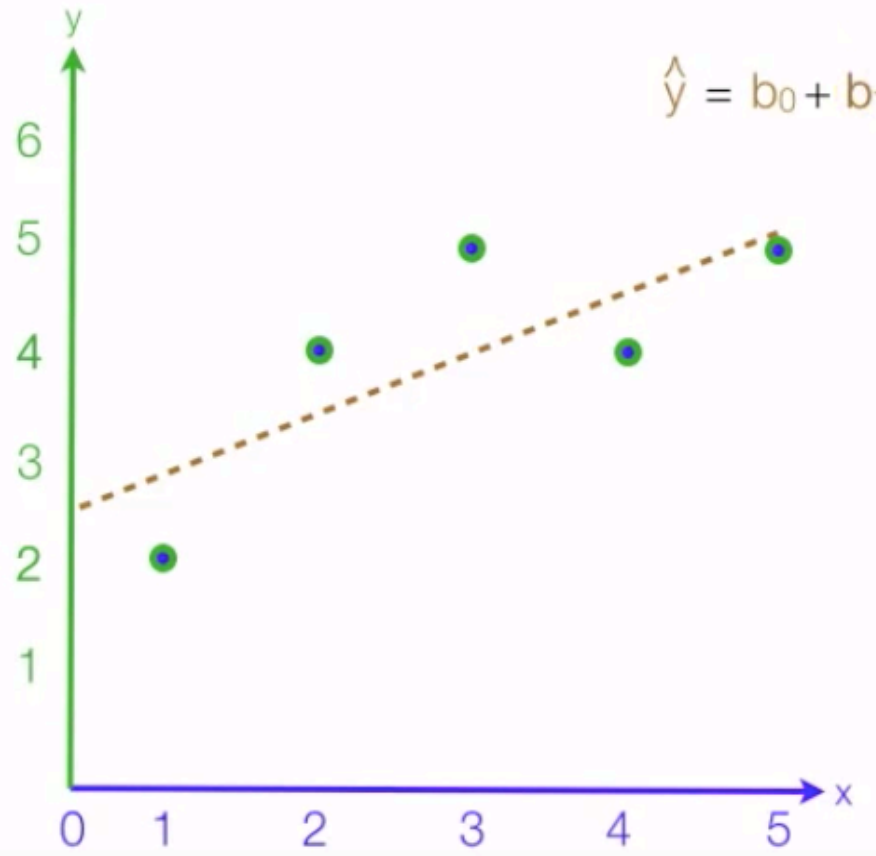
# Simple Linear Regression Analysis




| x | y | $x - \bar{x}$ | $y - \bar{y}$ |
|---|---|---|---|
| 1 | 2 | -2 | -2 |
| 2 | 4 | -1 | 0 |
| 3 | 5 | 0 | 1 |
| 4 | 4 | 1 | 0 |
| 5 | 5 | 2 | 1 |

mean $\bar{x} = 3$ $\bar{y} = 4$

# Simple Linear Regression Analysis



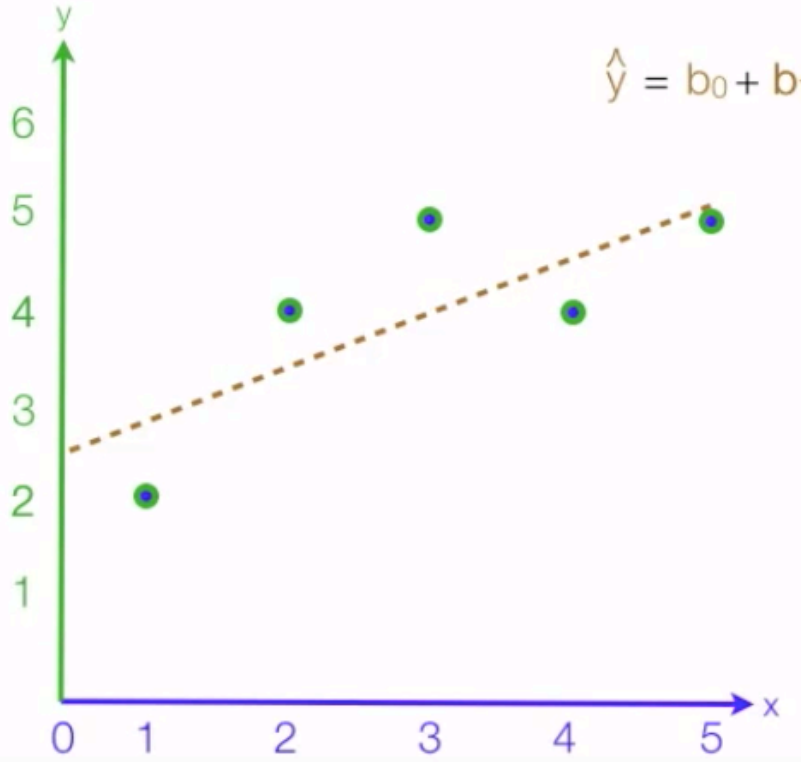

$$\hat{y} = b_0 + b_1 x$$

| x | y | $x - \bar{x}$ | $y - \bar{y}$ |
|---|---|---|---|
| 1 | 2 | -2 | -2 |
| 2 | 4 | -1 | 0 |
| 3 | 5 | 0 | 1 |
| 4 | 4 | 1 | 0 |
| 5 | 5 | 2 | 1 |

mean $\bar{x} = 3$ $\bar{y} = 4$

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

# Simple Linear Regression Analysis



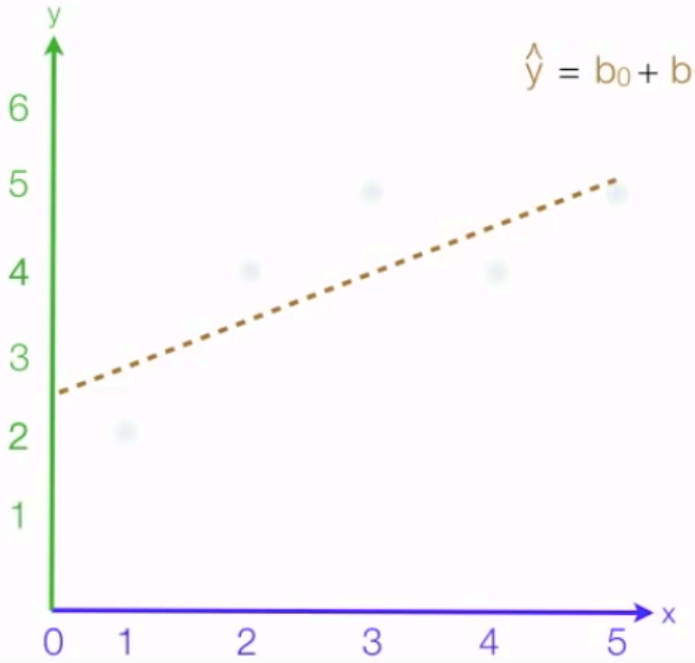

$$\hat{y} = b_0 + b_1 x$$

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

mean $\bar{x} = 3$ $\bar{y} = 4$

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

# Simple Linear Regression Analysis




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

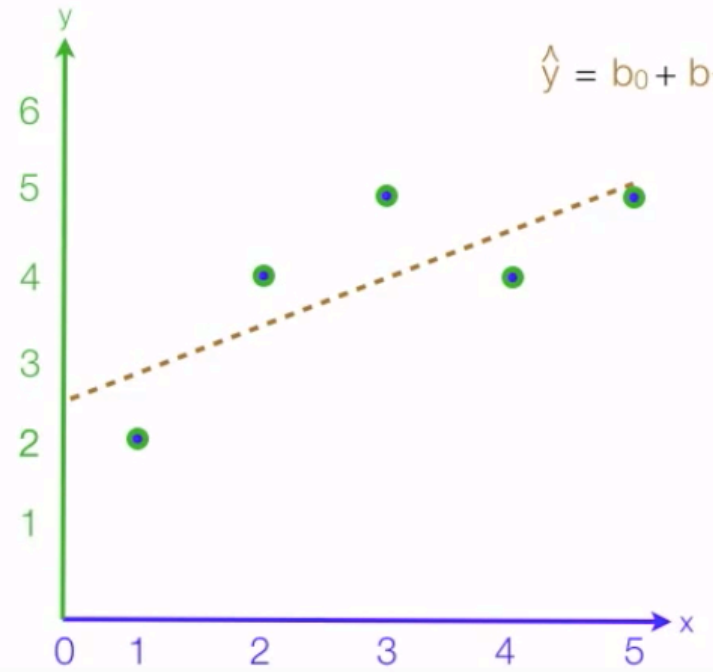
| x | y | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|---|
| 1 | 2 | -2 | -2 | 4 | 4 |
| 2 | 4 | -1 | 0 | 1 | 0 |
| 3 | 5 | 0 | 1 | 0 | 0 |
| 4 | 4 | 1 | 0 | 1 | 0 |
| 5 | 5 | 2 | 1 | 4 | 2 |
| | | | | 10 | 6 |

mean $\bar{x} = 3$ $\bar{y} = 4$

$$b_1 = \frac{6}{10} = .6$$

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

# Simple Linear Regression Analysis



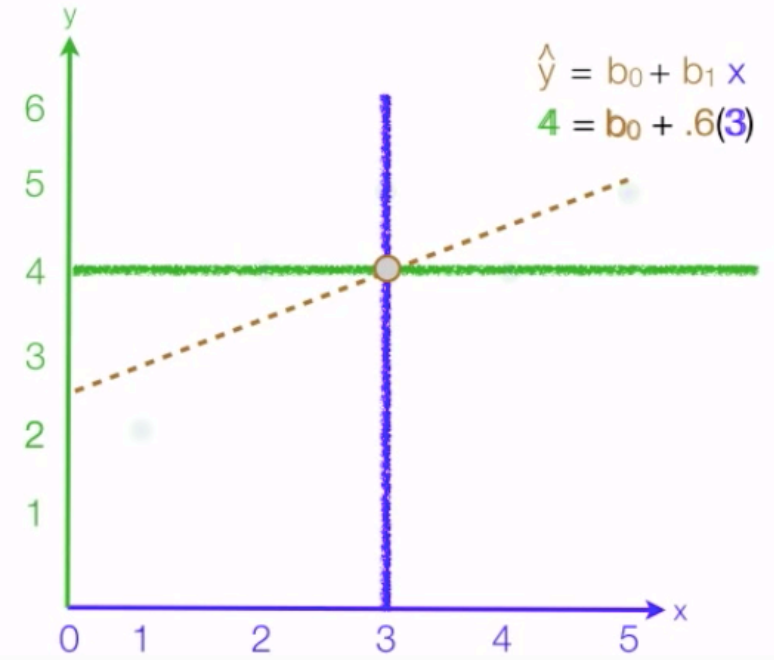

| x | y | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|---|
| 1 | 2 | -2 | -2 | 4 | 4 |
| 2 | 4 | -1 | 0 | 1 | 0 |
| 3 | 5 | 0 | 1 | 0 | 0 |
| 4 | 4 | 1 | 0 | 1 | 0 |
| 5 | 5 | 2 | 1 | 4 | 2 |
| | | | | 10 | 6 |

mean $\bar{x} = 3$ $\bar{y} = 4$

$$b_1 = \frac{6}{10} = .6$$

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

# Simple Linear Regression Analysis

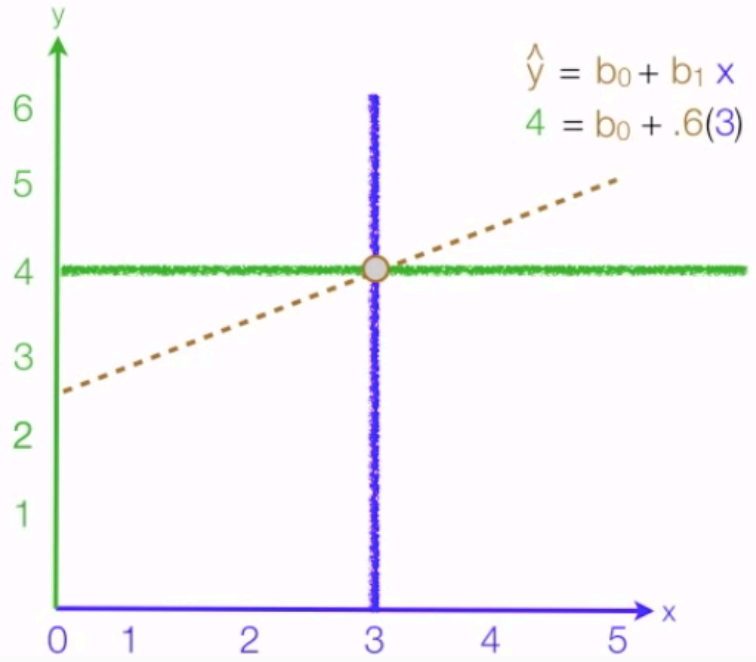



| x | y | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|---|
| 1 | 2 | -2 | -2 | 4 | 4 |
| 2 | 4 | -1 | 0 | 1 | 0 |
| 3 | 5 | 0 | 1 | 0 | 0 |
| 4 | 4 | 1 | 0 | 1 | 0 |
| 5 | 5 | 2 | 1 | 4 | 2 |
| | | | | 10 | 6 |

mean $\bar{x} = 3$ $\bar{y} = 4$

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

# Simple Linear Regression Analysis




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

$4 = b_0 + .6(3)$

| 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 $\bar{x} = 3$ $\bar{y} = 4$

$4 = b_0 + .6(3)$

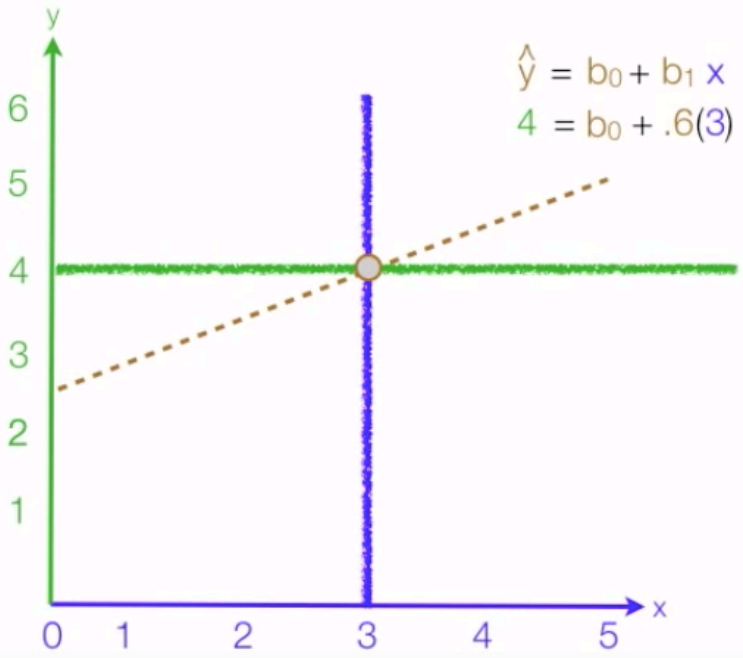
$4 = b_0 + 1.8$

$-1.8 \quad -1.8$

$2.2 = b_0$

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

# Simple Linear Regression Analysis




$b_0 = 2.2$

$b_1 = .6$

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

| 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 $\bar{x} = 3$ $\bar{y} = 4$

$4 = b_0 + .6(3)$

$4 = b_0 + 1.8$

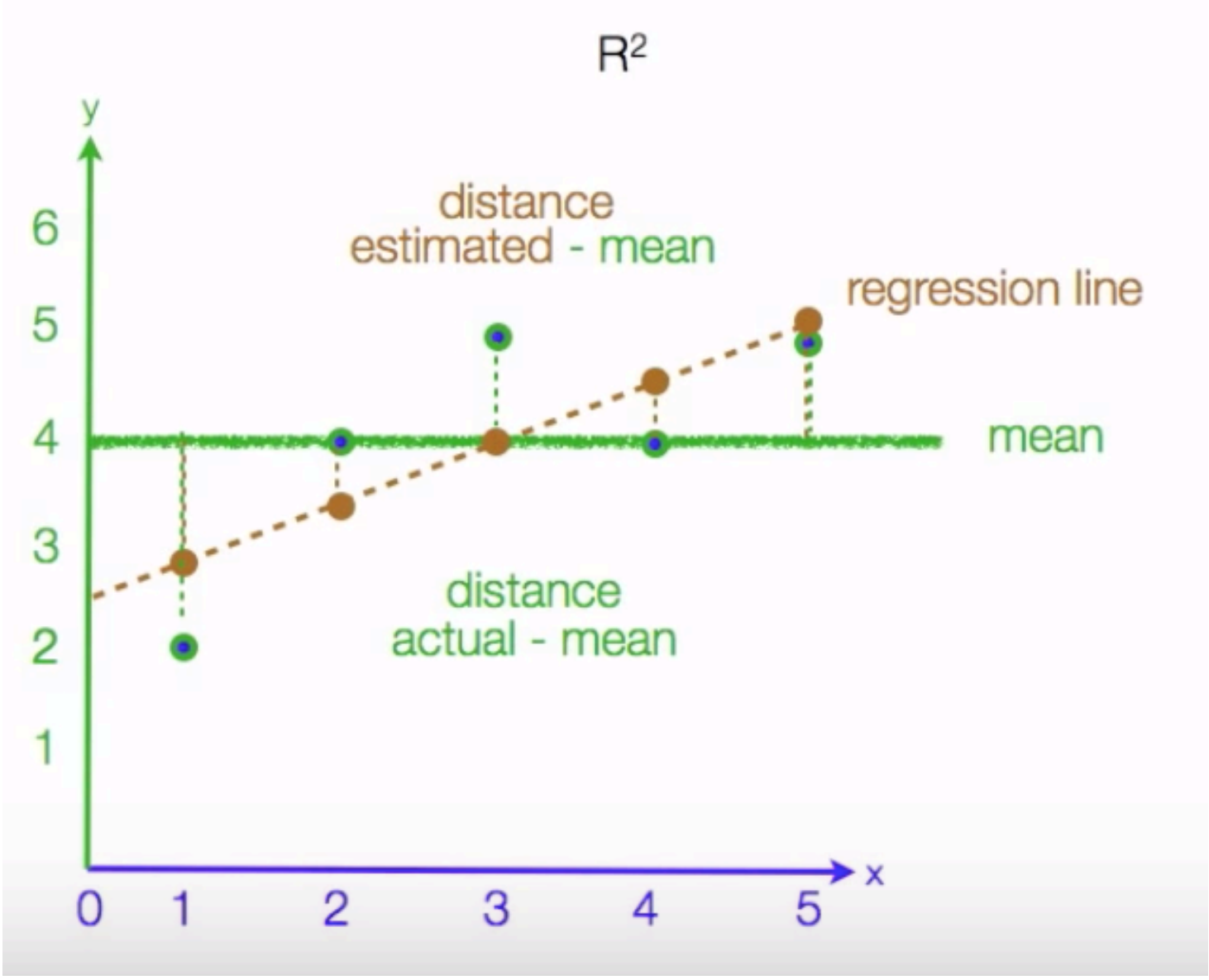
$-1.8 \quad -1.8$

$2.2 = b_0$

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

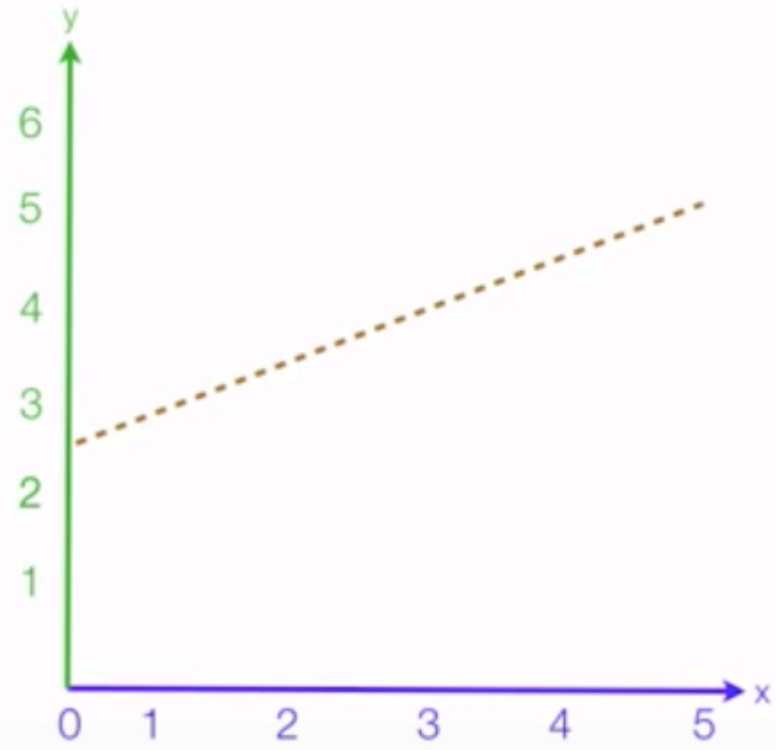
# What is R Squared?

## A measure if Fitness



# R Squared

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$



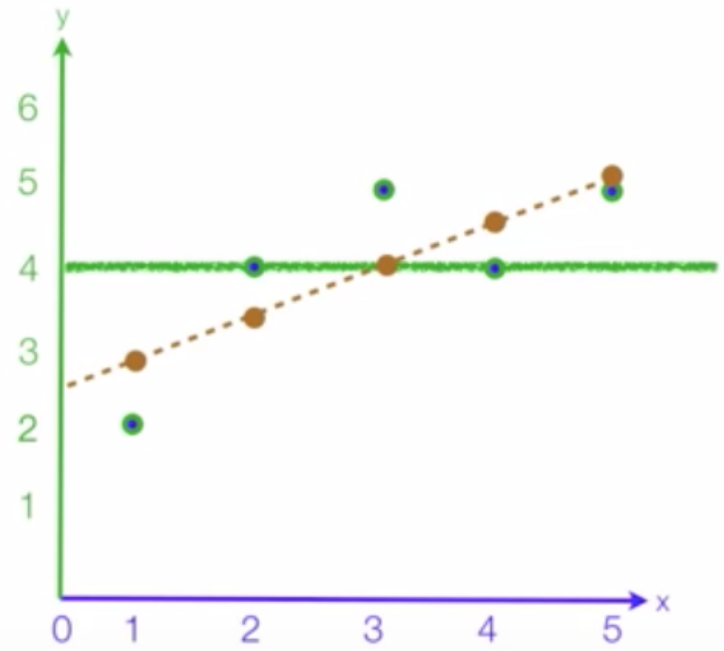

$\hat{y} = 2.2 + .6x$ estimated values

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

mean $\bar{x} = 3$ $\bar{y} = 4$

# R Squared

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$



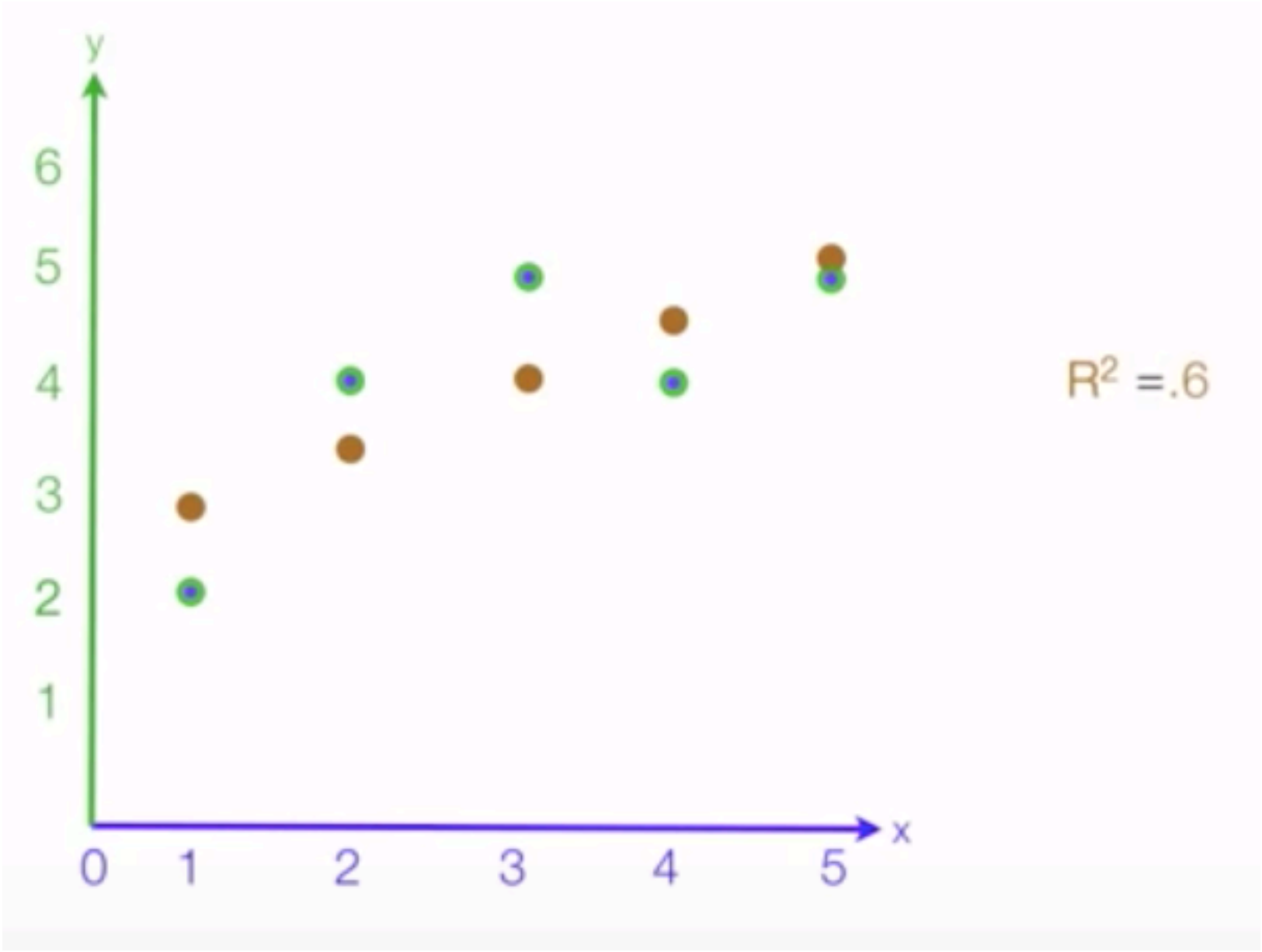

$$\hat{y} = 2.2 + .6x$$

| x | y | $y - \bar{y}$ | $(y - \bar{y})^2$ | $\hat{y}$ | $\hat{y} - \bar{y}$ | $(\hat{y} - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 2 | -2 | 4 | 2.8 | -1.2 | 1.44 |
| 2 | 4 | 0 | 0 | 3.4 | -.6 | .36 |
| 3 | 5 | 1 | 1 | 4 | 0 | 0 |
| 4 | 4 | 0 | 0 | 4.6 | .6 | .36 |
| 5 | 5 | 1 | 1 | 5.2 | 1.2 | 1.44 |
| | | | 6 | | | 3.6 |

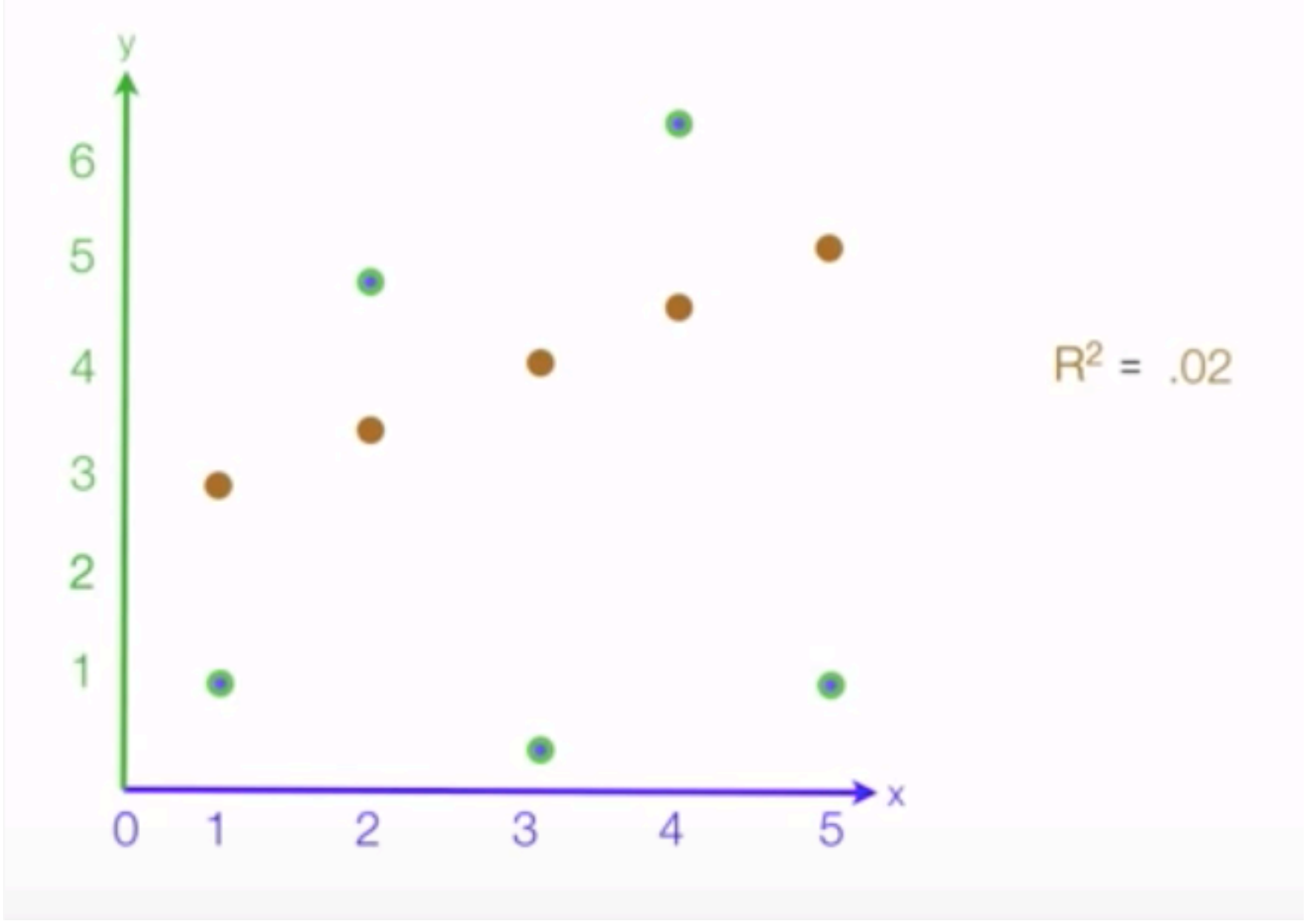
mean $\bar{x} = 3$ $\bar{y} = 4$

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{3.6}{6} = .6$$

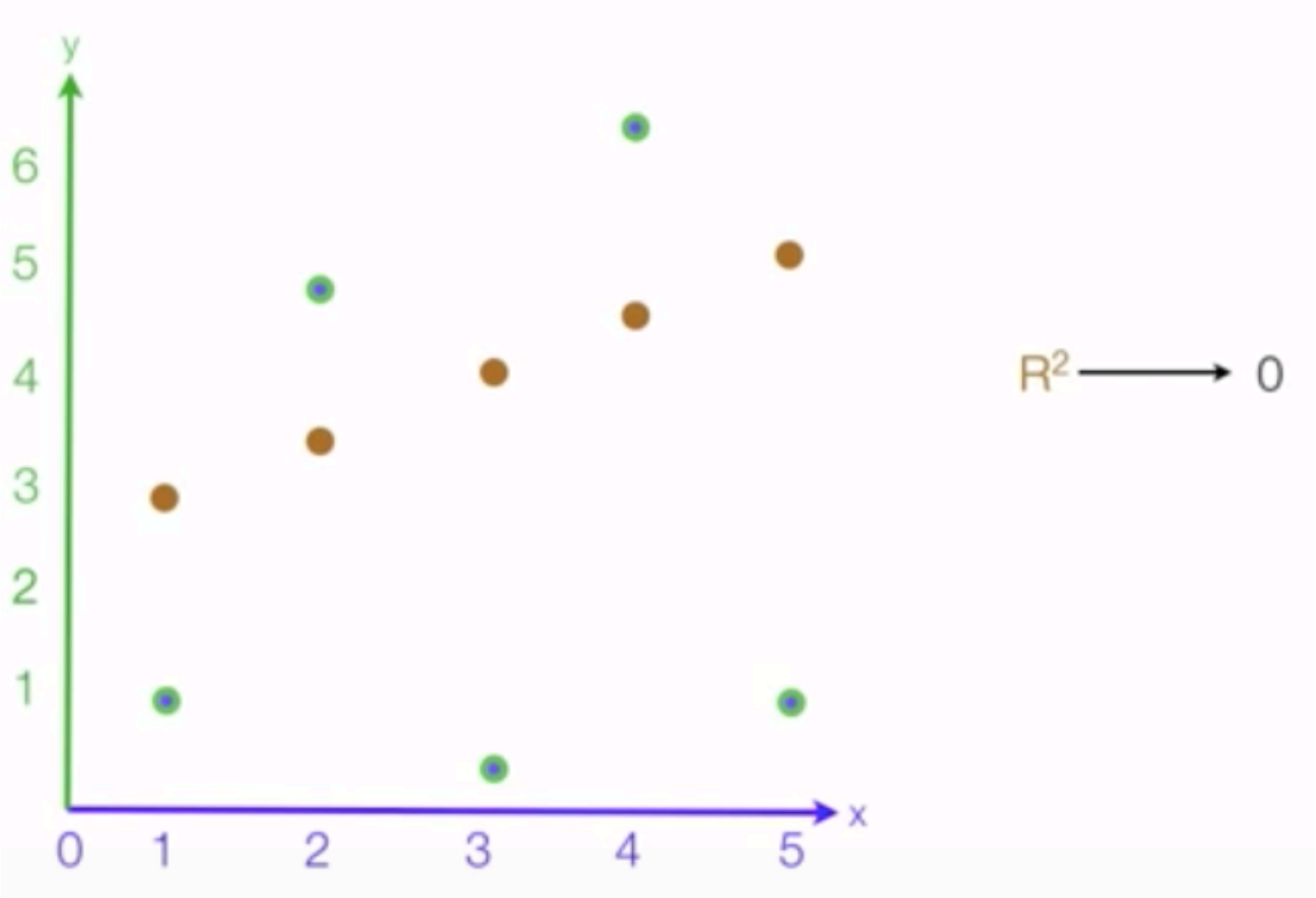
# R Squared



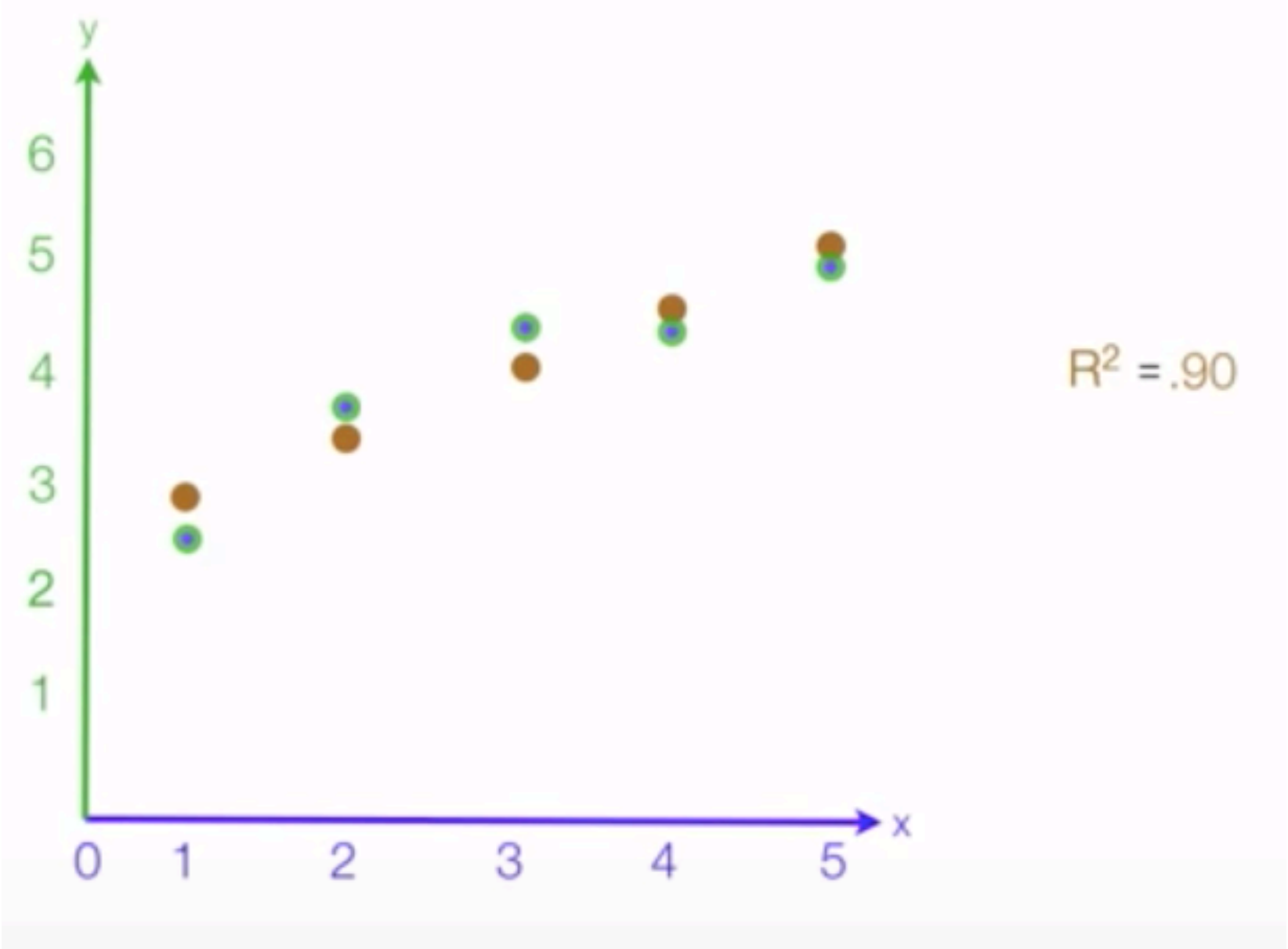
# R Squared



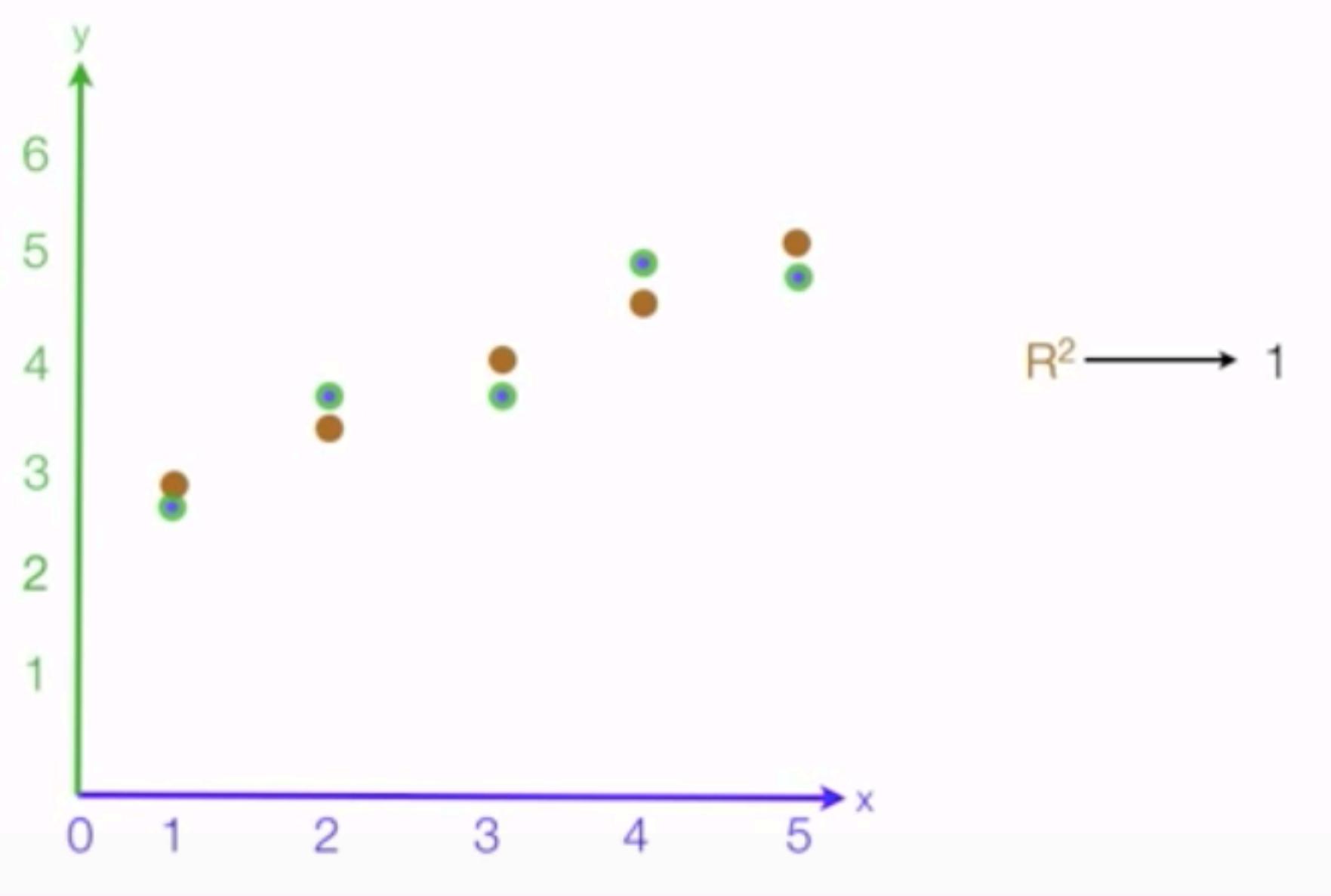
# R Squared



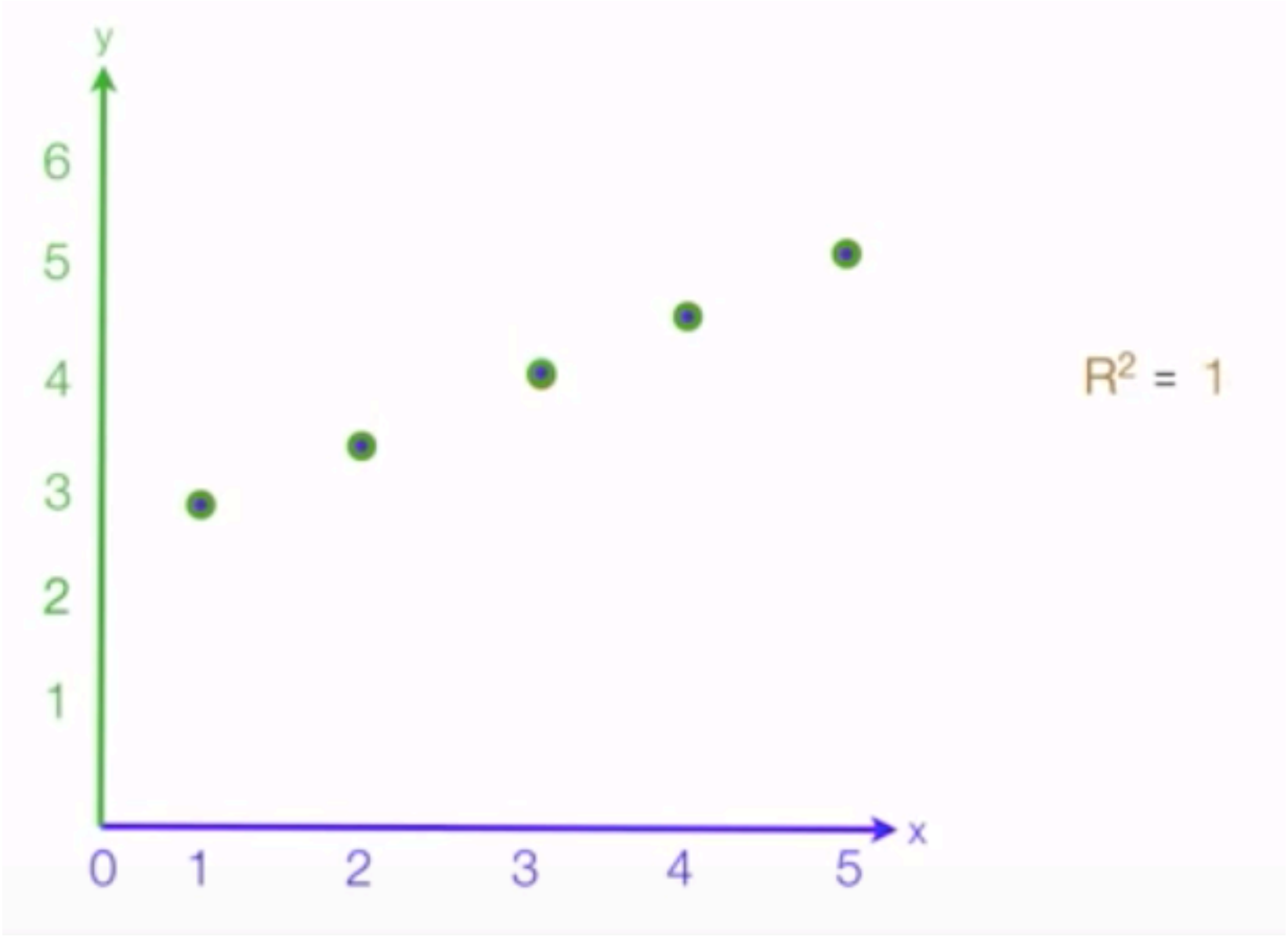
# R Squared



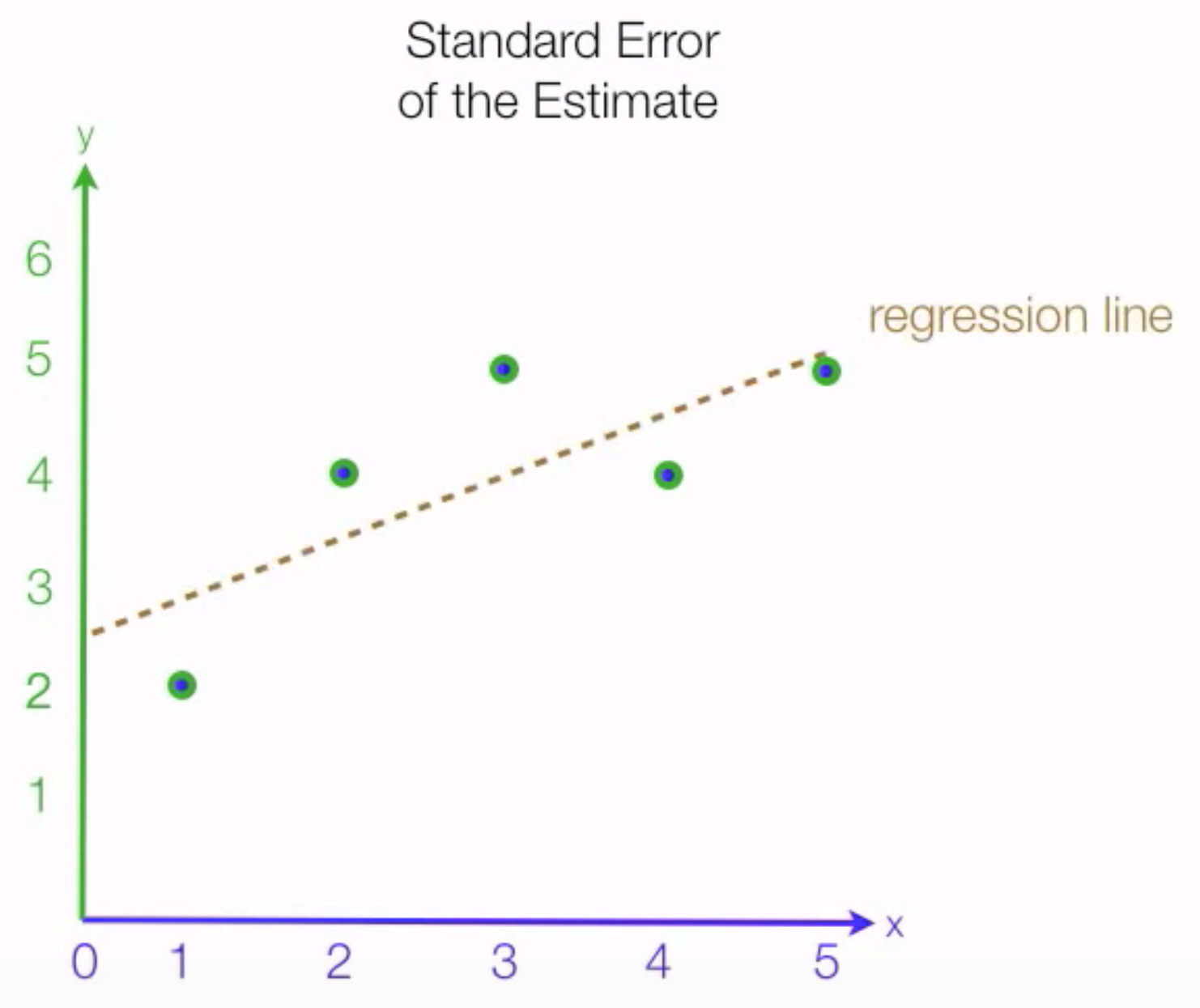
# R Squared



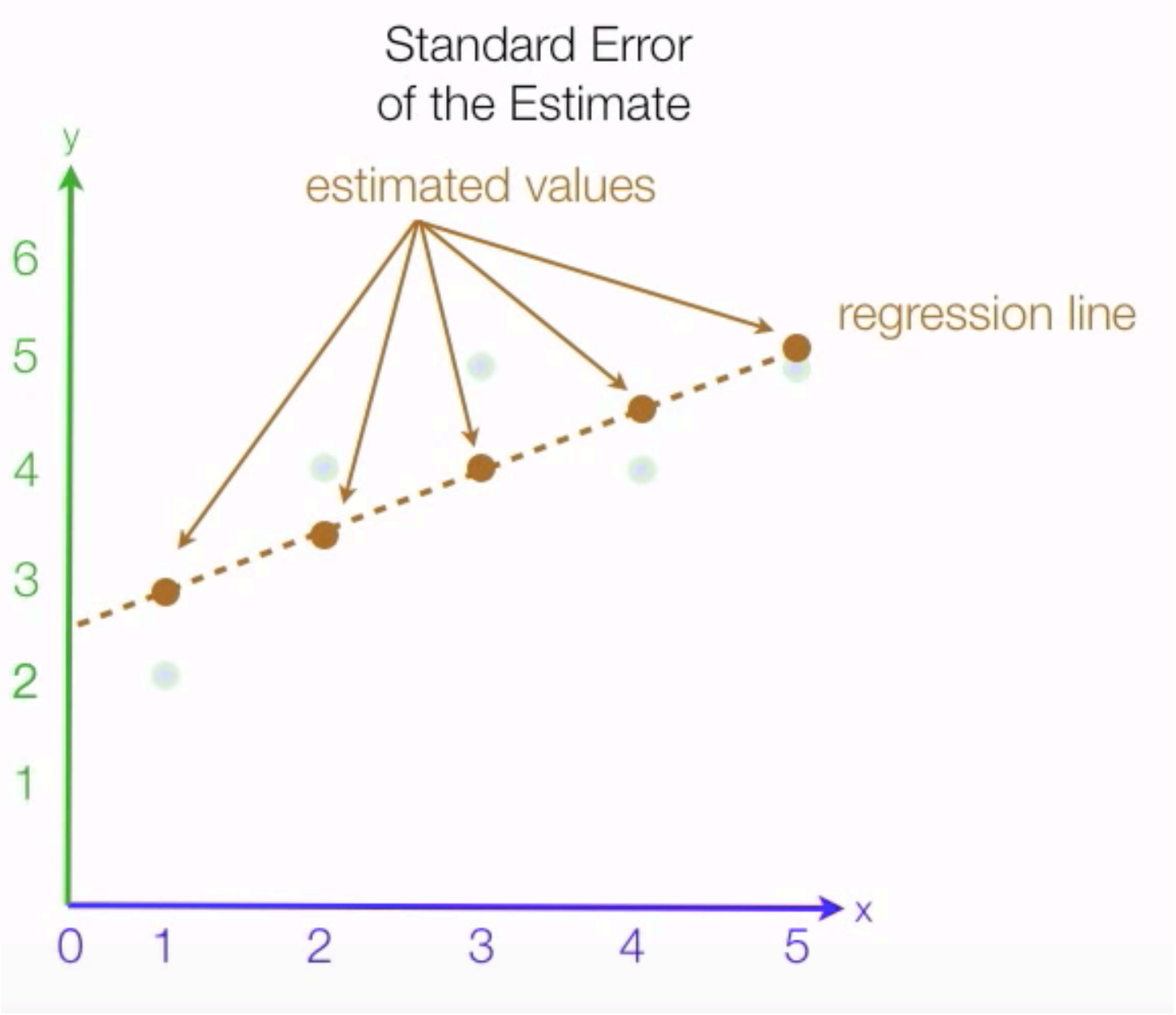
# R Squared



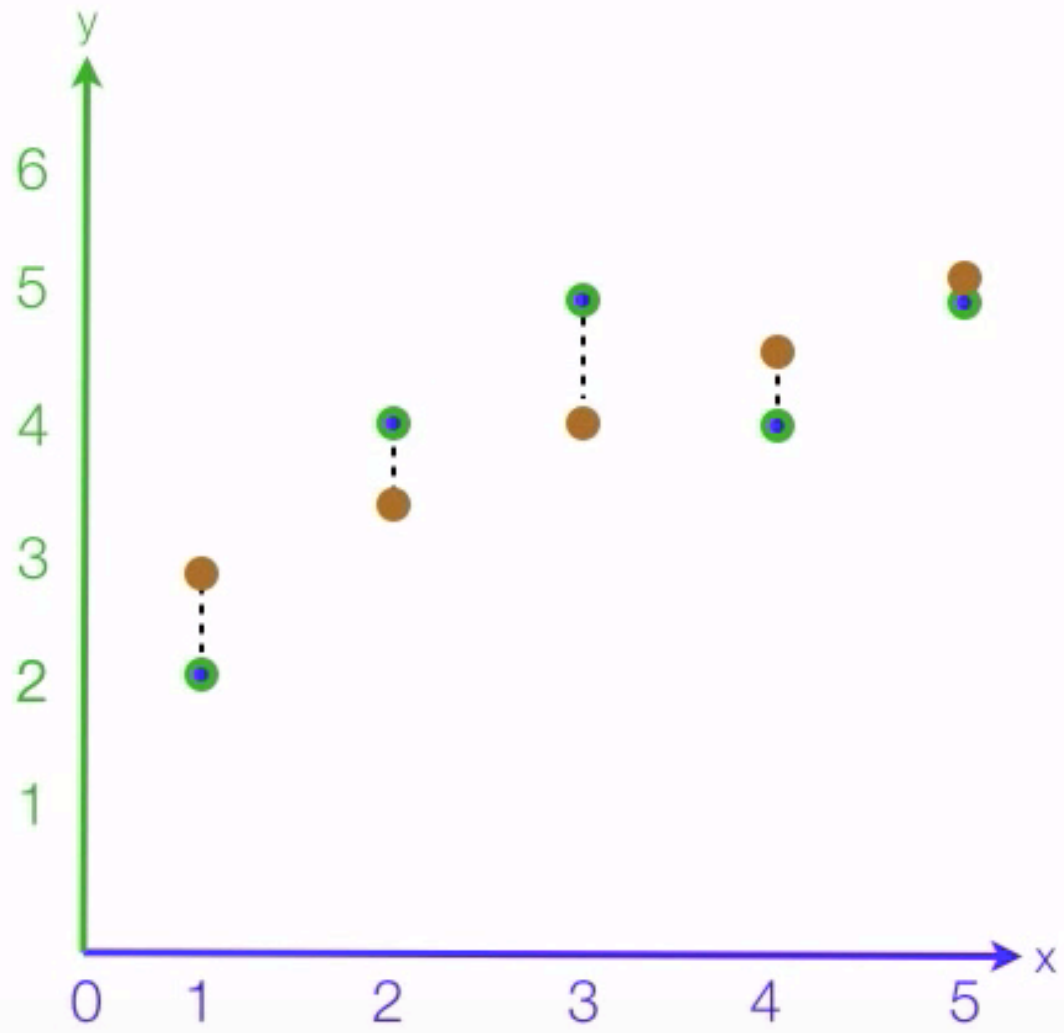
# Mean Squared Error



# Mean Squared Error

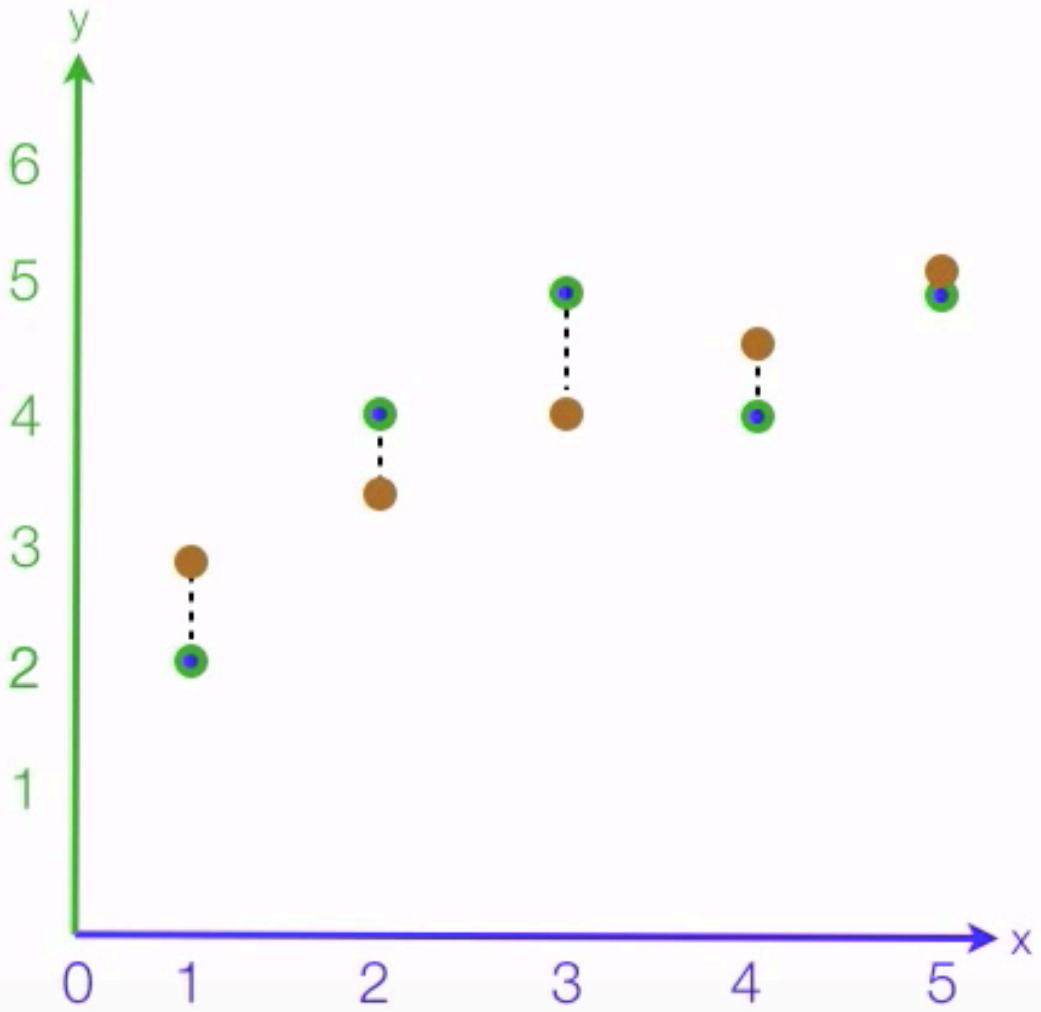


# Mean Squared Error




| | actual | estimated |
|---|---|---|
| x | y | $\hat{y}$ |
| 1 | 2 | 2.8 |
| 2 | 4 | 3.4 |
| 3 | 5 | 4 |
| 4 | 4 | 4.6 |
| 5 | 5 | 5.2 |

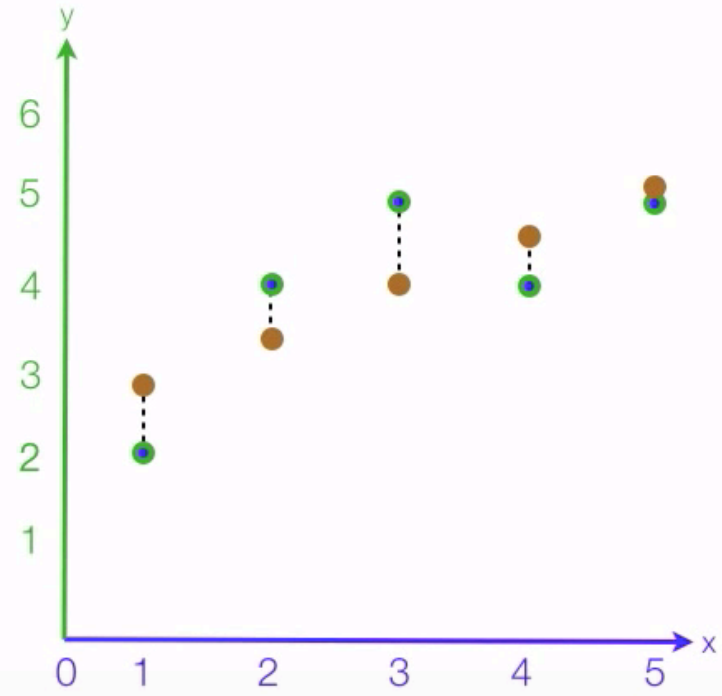
# Mean Squared Error




| x | y | $\hat{y}$ |
|---|---|---|
| 1 | 2 | 2.8 |
| 2 | 4 | 3.4 |
| 3 | 5 | 4 |
| 4 | 4 | 4.6 |
| 5 | 5 | 5.2 |

$$\text{Standard Error of the Estimate} = \sqrt{\frac{\Sigma(\hat{y} - y)^2}{n - 2}}$$

# Mean Squared Error

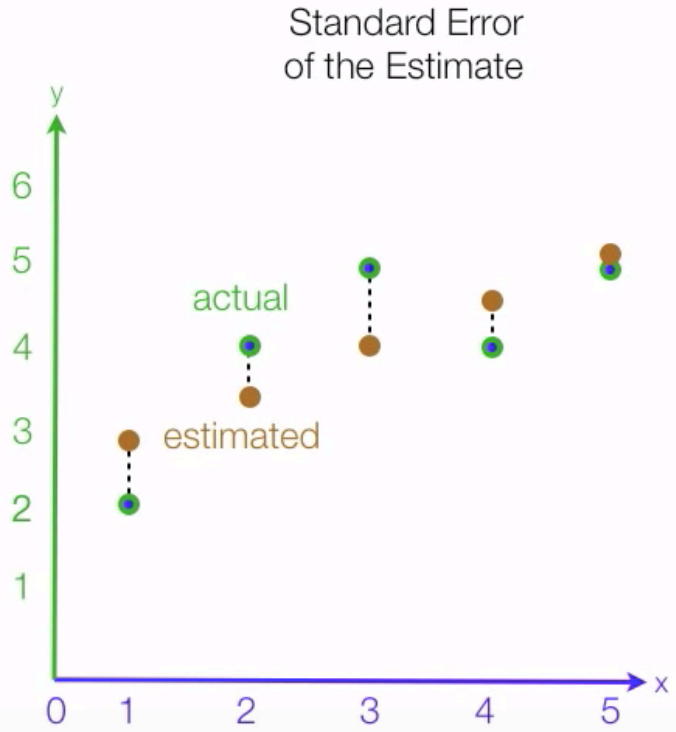


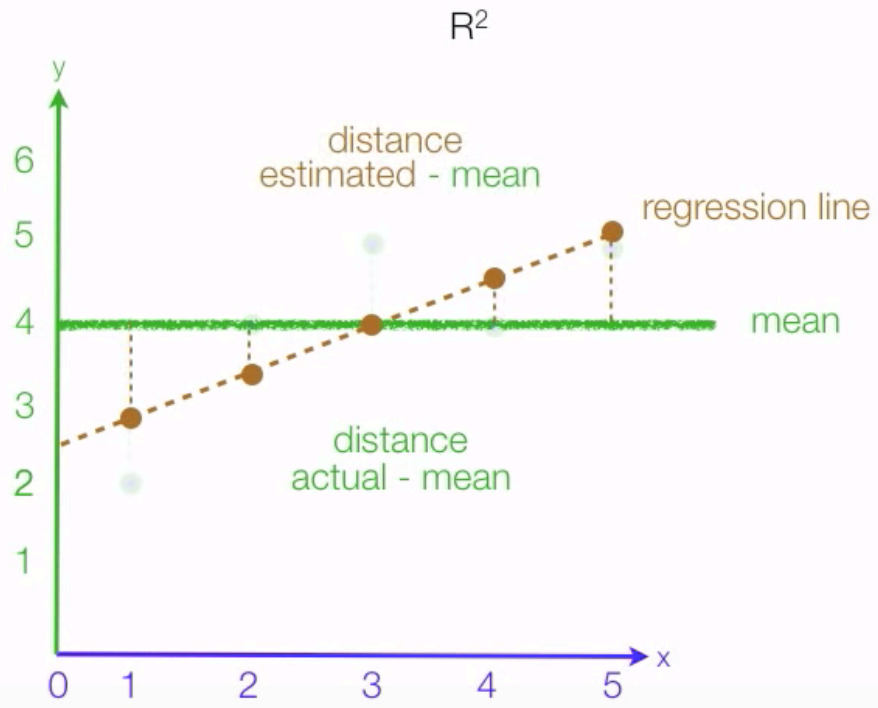

| x | y | $\hat{y}$ | $\hat{y}$ - y | ($\hat{y}$ - y)$^2$ |
|---|---|---|---|---|
| 1 | 2 | 2.8 | .8 | .64 |
| 2 | 4 | 3.4 | -.6 | .36 |
| 3 | 5 | 4 | -1 | 1 |
| 4 | 4 | 4.6 | .6 | .36 |
| 5 | 5 | 5.2 | .2 | .04 |
| | | | | 2.4 |

Standard Error of the Estimate $= \sqrt{\frac{\Sigma(\hat{y} - y)^2}{n - 2}} = \sqrt{\frac{2.4}{5 - 2}} = \sqrt{\frac{2.4}{3}} = \sqrt{.8} = .89$

# Mean Squared Error And $R^2$:

## Measures of Fitness



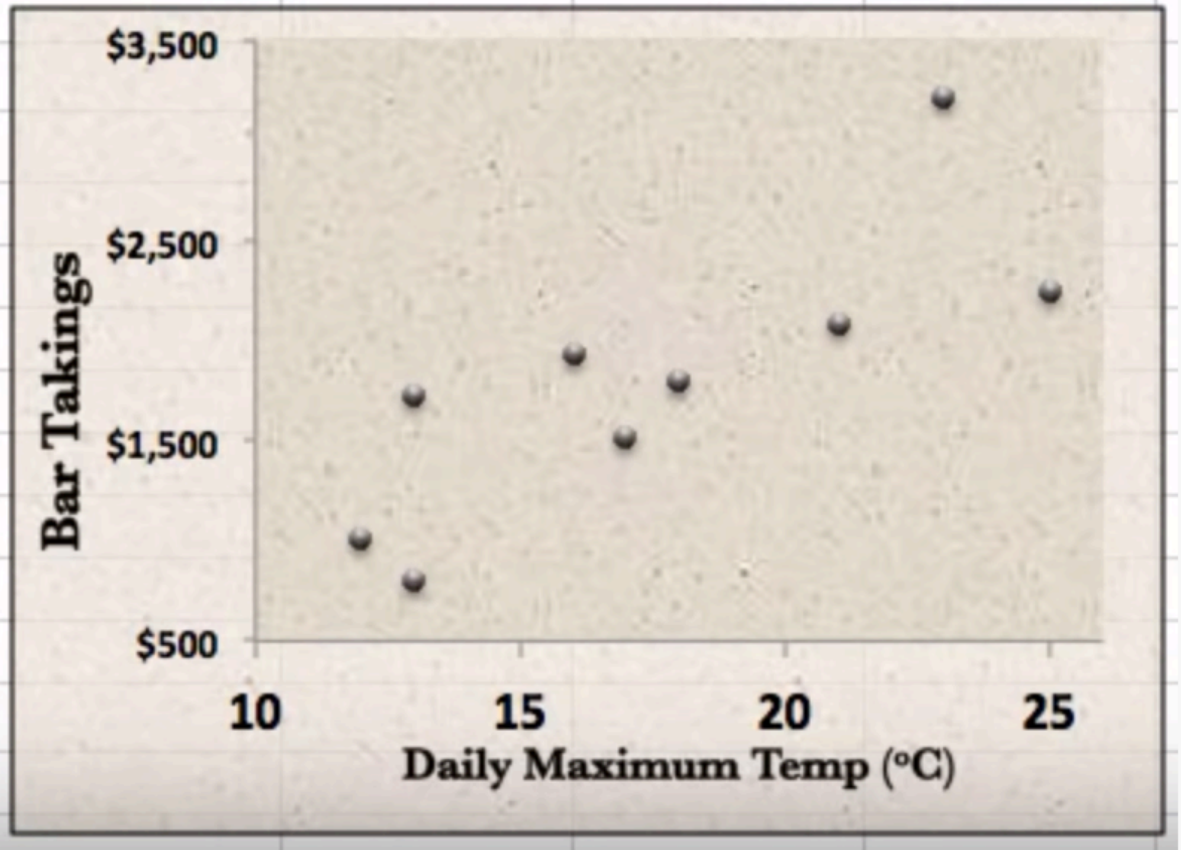



# End of Tutorial on Simple Linear Regression

# An Example with British Pubs

# An Example: British Pubs

| Day | Takings | Temp (°C) |
|---|---|---|
| 3-Jun | $3,213 | 23 |
| 10-Jun | $2,089 | 21 |
| 17-Jun | $2,253 | 25 |
| 24-Jun | $1,801 | 18 |
| 1-Jul | $801 | 13 |
| 8-Jul | $1,934 | 16 |
| 15-Jul | $1,720 | 13 |
| 22-Jul | $1,514 | 17 |
| 29-Jul | $1,0 | 12 |



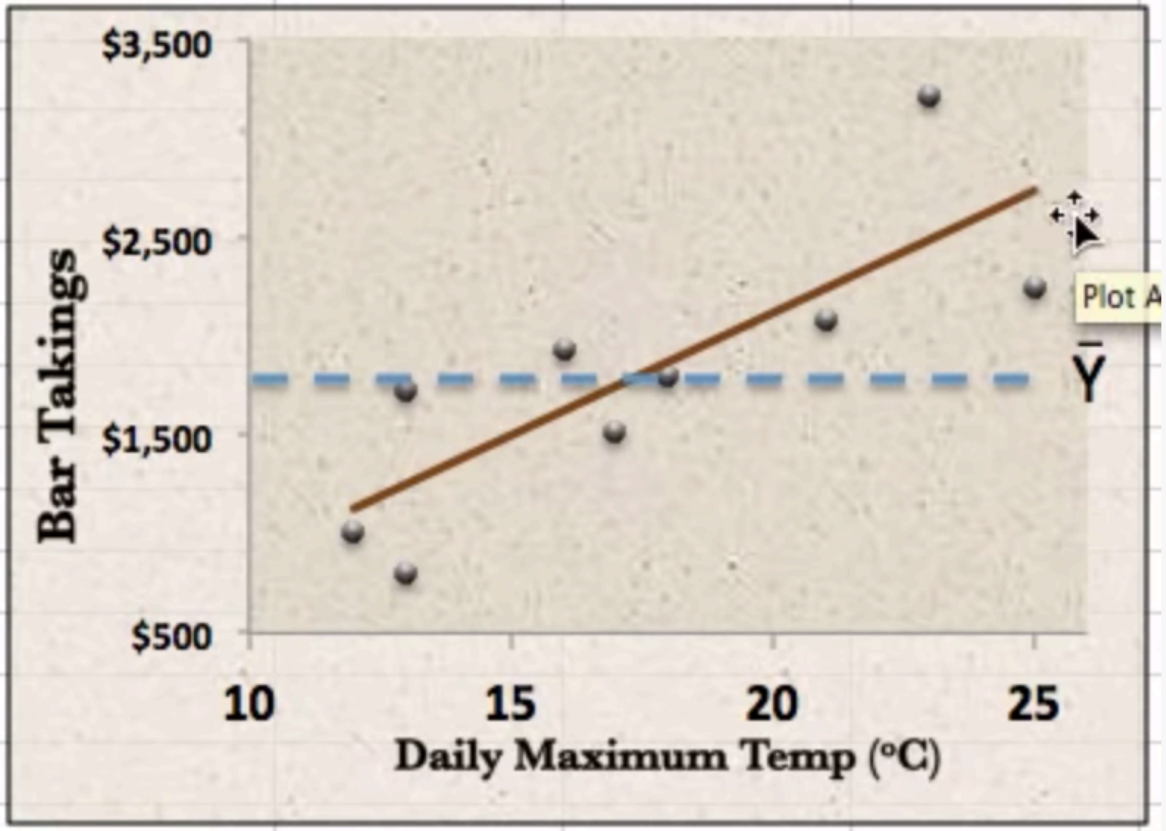

https://www.youtube.com/watch?v=aq8VU5KLmkY

# An Example

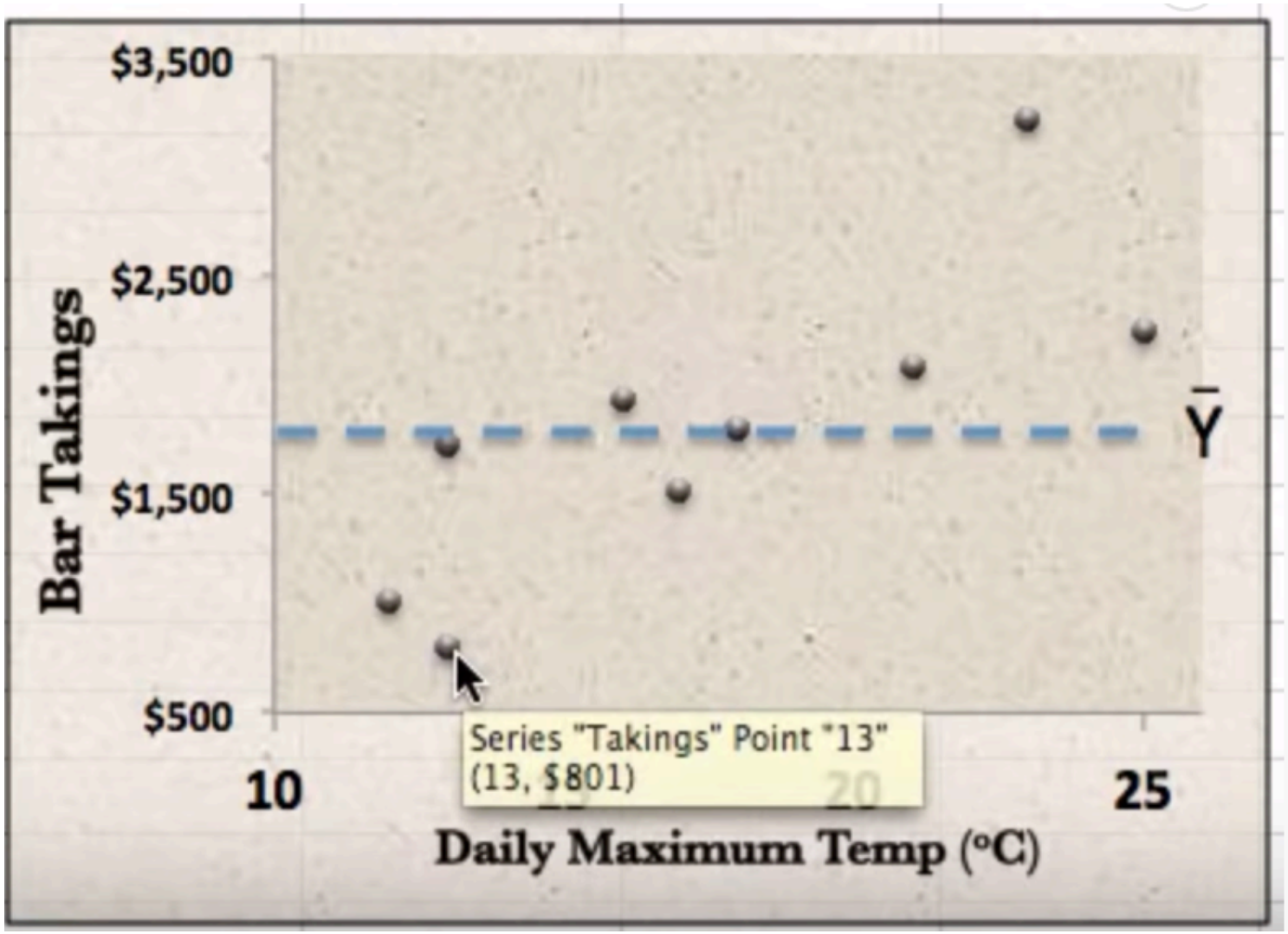
| Day | Takings | Temp (°C) |
|---|---|---|
| 3-Jun | $3,213 | 23 |
| 10-Jun | $2,089 | 21 |
| 17-Jun | $2,253 | 25 |
| 24-Jun | $1,801 | 18 |
| 1-Jul | $801 | 13 |
| 8-Jul | $1,934 | 16 |
| 15-Jul | $1,720 | 13 |
| 22-Jul | $1,514 | 17 |
| 29-Jul | $1,017 | 12 |

SAMPLE REGRESSION LINE

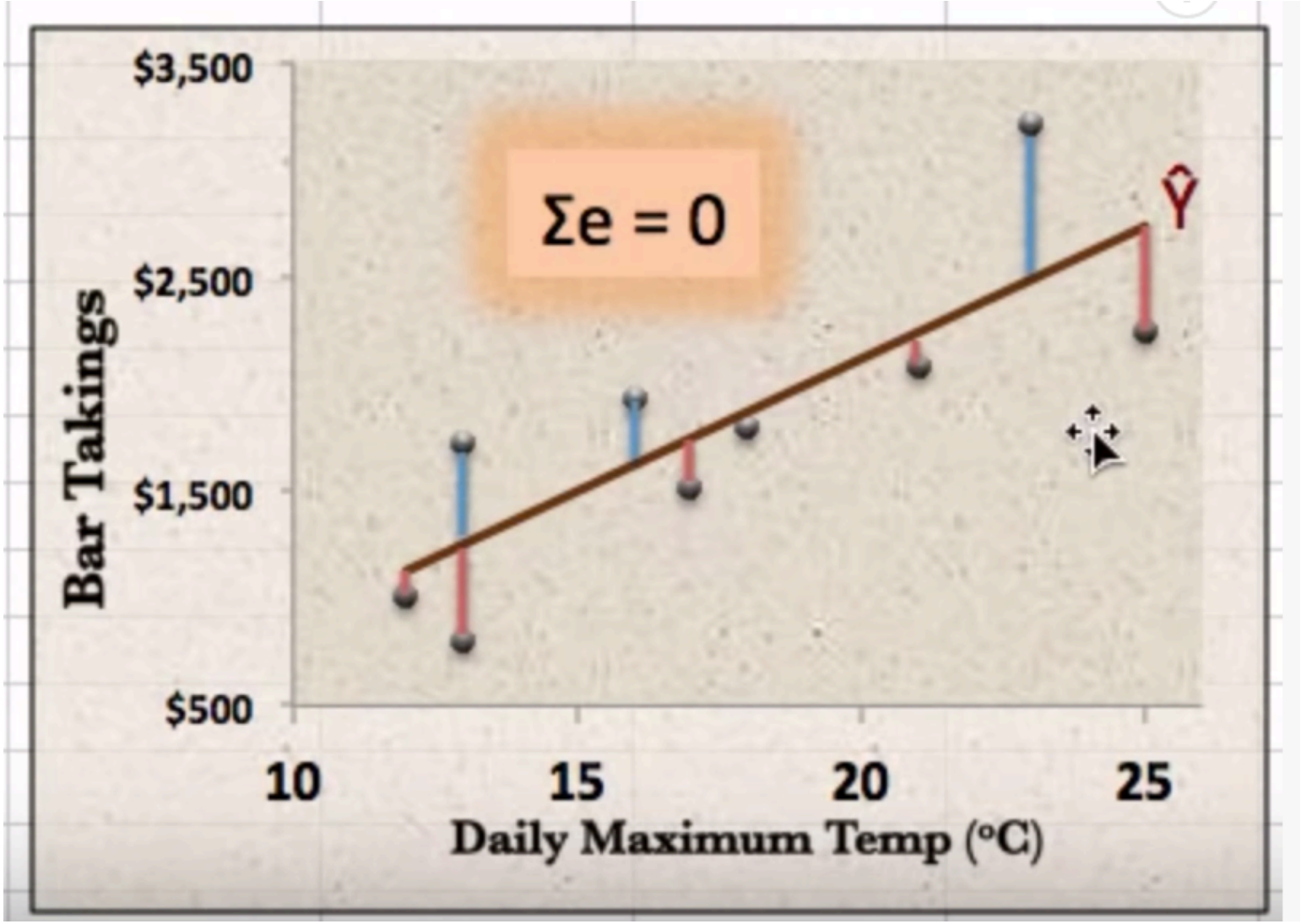
$\hat{Y} = -353.11 + 123.54X$



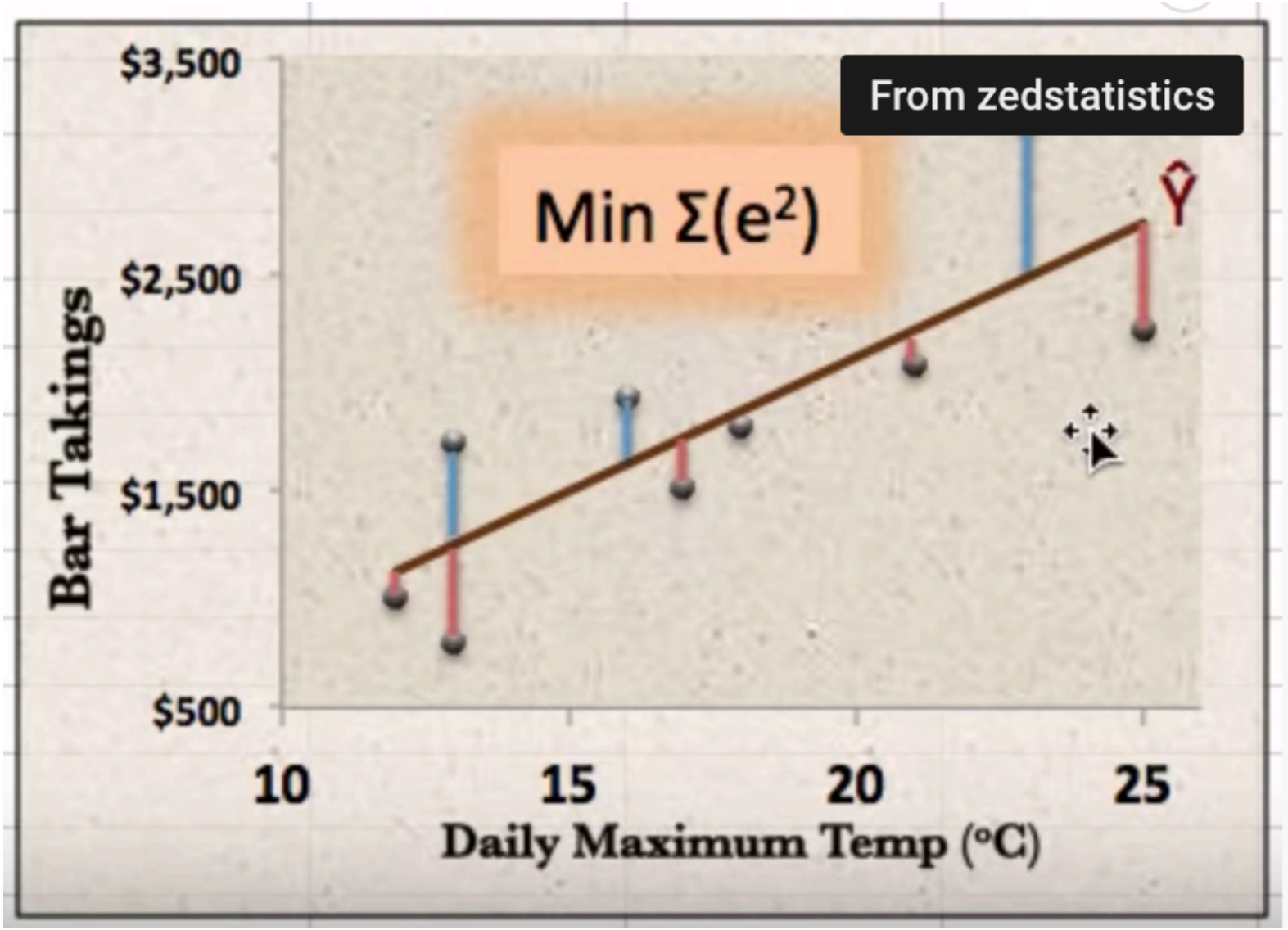
# An Example


$3,500
$2,500
$1,500
$500
Bar Takings
$\bar{Y}$
Series "Takings" Point "13"
(13, $801)
10
25
Daily Maximum Temp (°C)

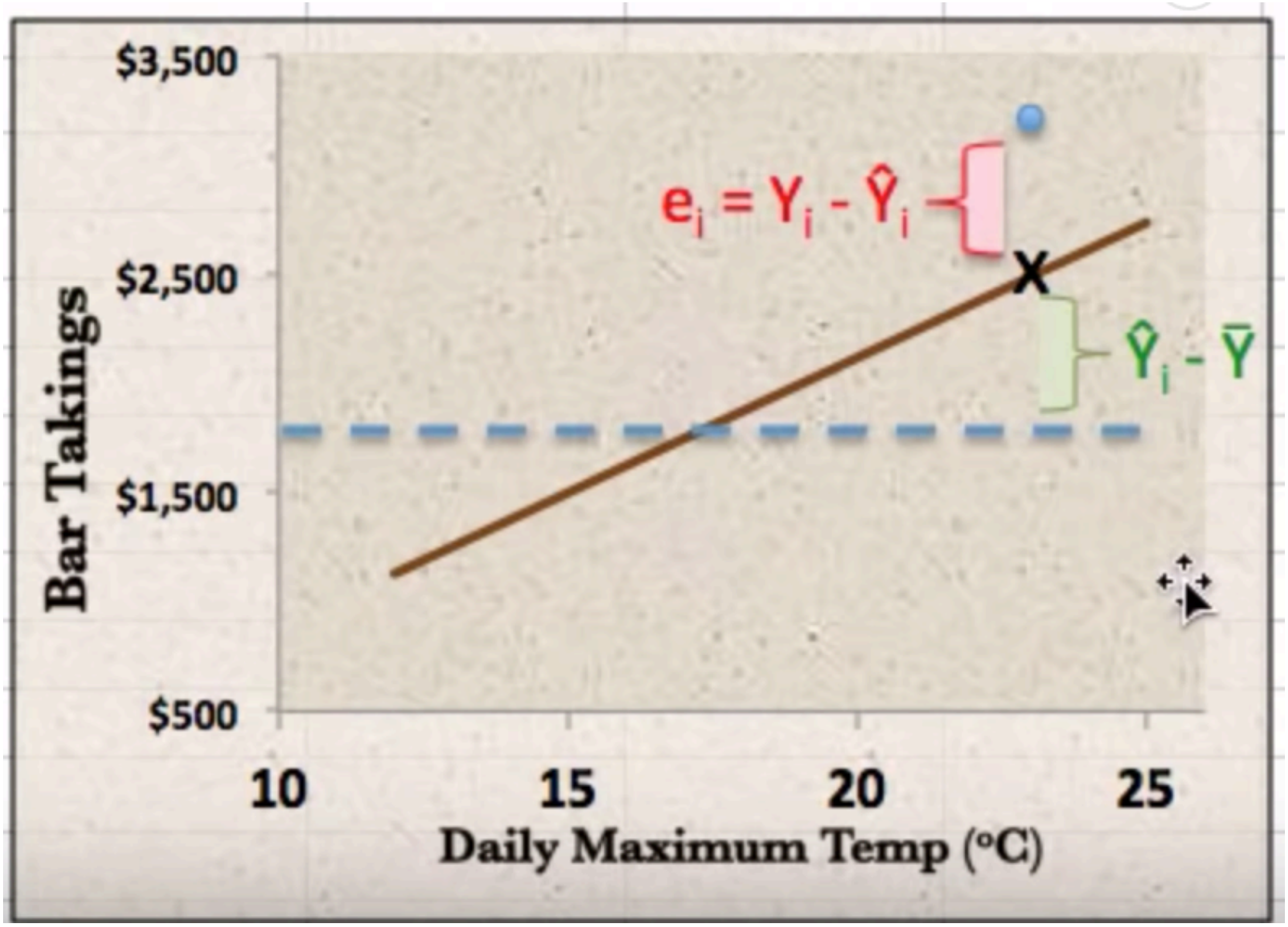
# An Example


$3,500
$2,500
$1,500
$500
Bar Takings
Σe = 0
Ŷ
10
15
20
25
Daily Maximum Temp (°C)

# An Example


From zedstatistics
Min Σ(e²)
Ŷ
$3,500
$2,500
$1,500
$500
Bar Takings
10
15
20
25
Daily Maximum Temp (°C)

# An Example

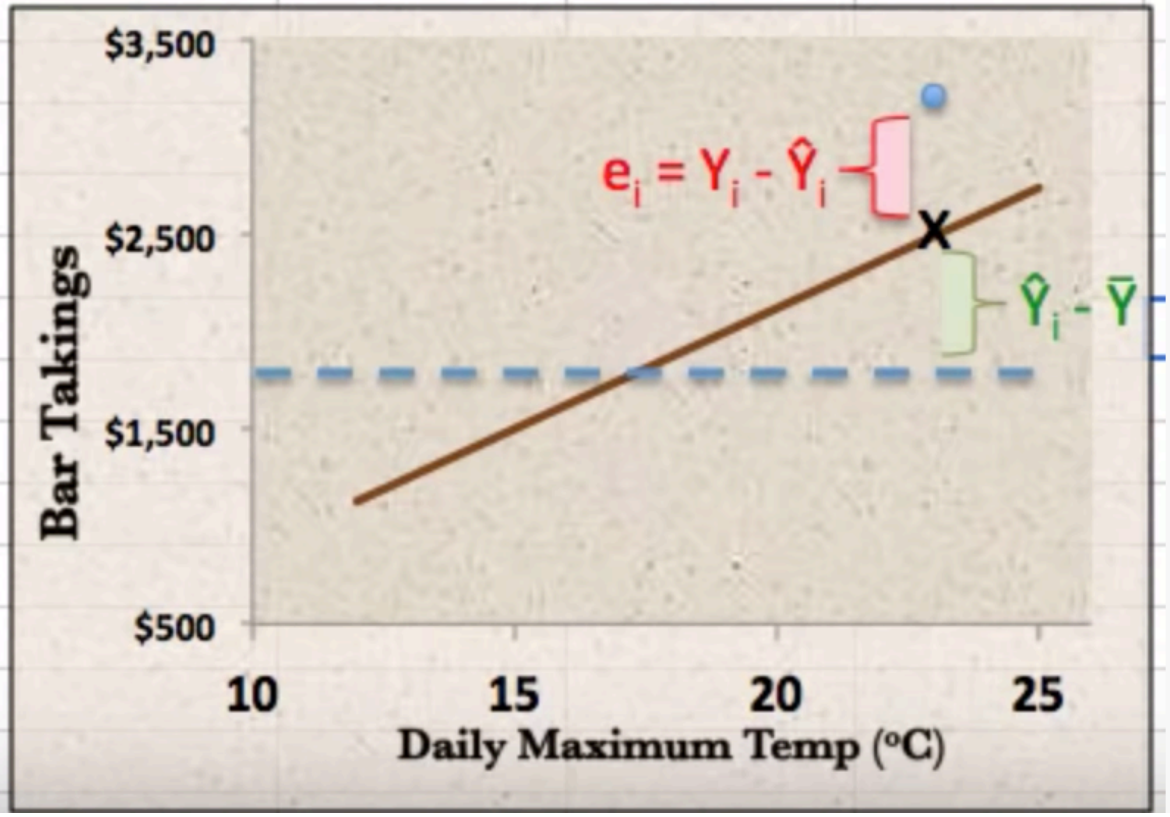

$3,500
$2,500
$1,500
$500
Bar Takings
10
15
20
25
Daily Maximum Temp (°C)
$e_i = Y_i - \hat{Y}_i$
$\hat{Y}_i - \bar{Y}$

# An Example

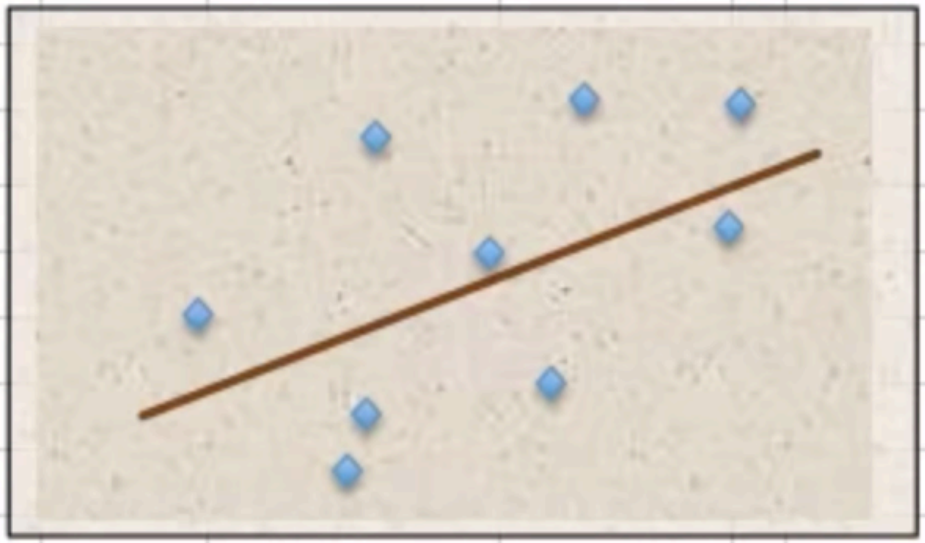
$$SSR = \Sigma( \hat{Y}_i - \bar{Y} )^2$$
$$SSE = \Sigma( Y_i - \hat{Y}_i )^2$$

$$SST = SSR + SSE$$
$$SST = \Sigma(Y_i - \bar{Y})^2$$

$$R^2 = SSR/SST$$

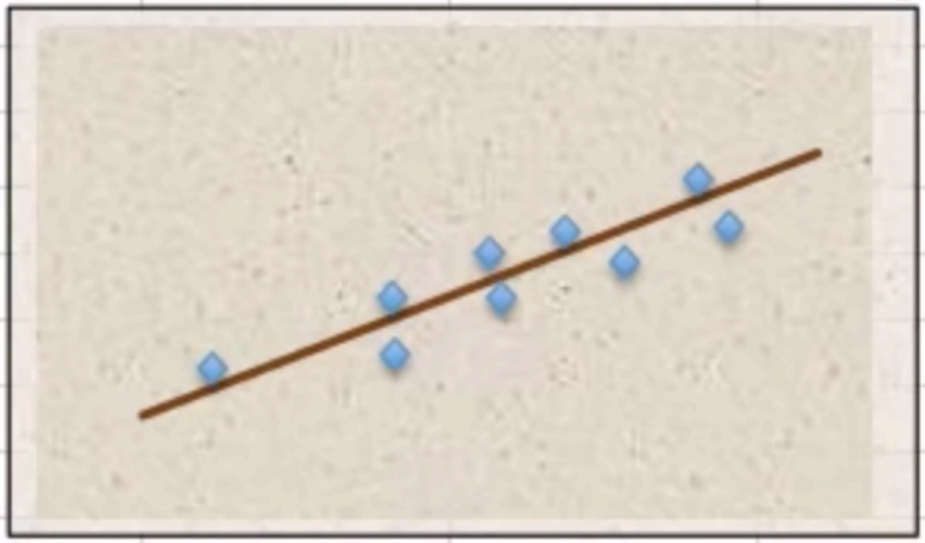


# An Example



High SSE

Low $R^2$

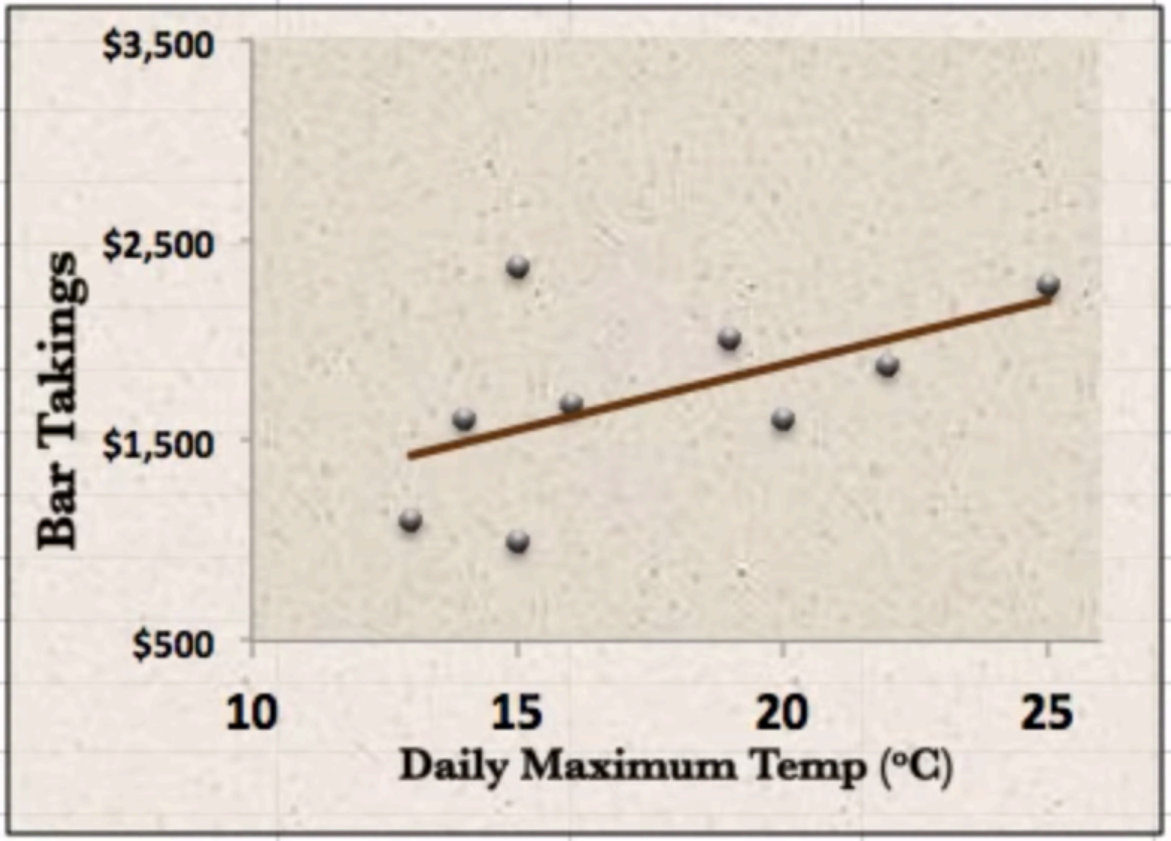


Low SSE

High $R^2$

# An Example: British Pubs

| Day | Takings | Temp (°C) |
|---|---|---|
| 5-Aug | $1,602 | 14 |
| 12-Aug | $1,688 | 16 |
| 19-Aug | $2,017 | 19 |
| 26-Aug | $1,100 | 13 |
| 2-Sep | $1,609 | 20 |
| 9-Sep | $1,880 | 22 |
| 16-Sep | $997 | 15 |
| 23-Sep | $2,366 | 15 |
| 30-Sep | $2,280 | 25 |




SAMPLE REGRESSION LINE

$\hat{Y} = 586.03 + 64.558X$

# An Example

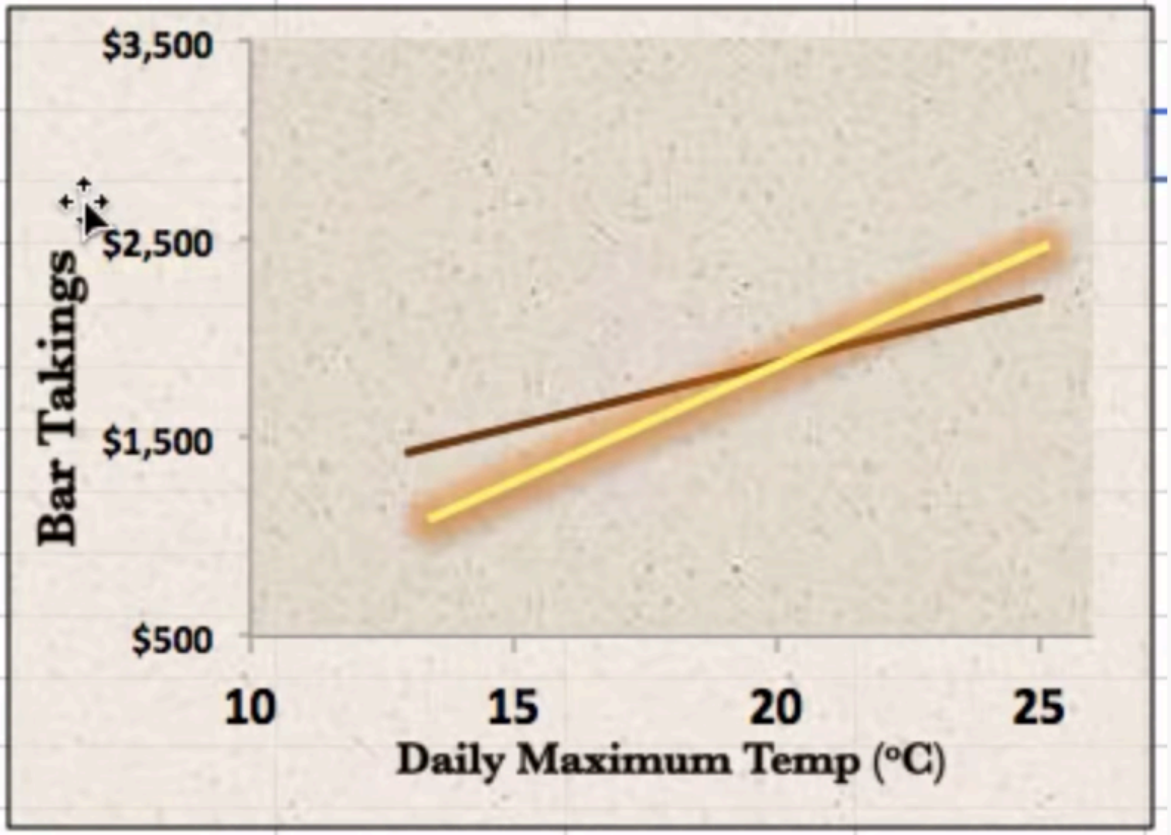
POPULATION REGRESSION FUNCTION

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

| Day | Takings | Temp (°C) |
|---|---|---|
| 5-Aug | $1,602 | 14 |
| 12-Aug | $1,688 | 16 |
| 19-Aug | $2,017 | 19 |
| 26-Aug | $1,100 | 13 |
| 2-Sep | $1,609 | 20 |
| 9-Sep | $1,880 | 22 |
| 16-Sep | $997 | 15 |
| 23-Sep | $2,366 | 15 |
| 30-Sep | $2,280 | 25 |

SAMPLE REGRESSION LINE

$$\hat{Y} = 586.03 + 64.558X$$

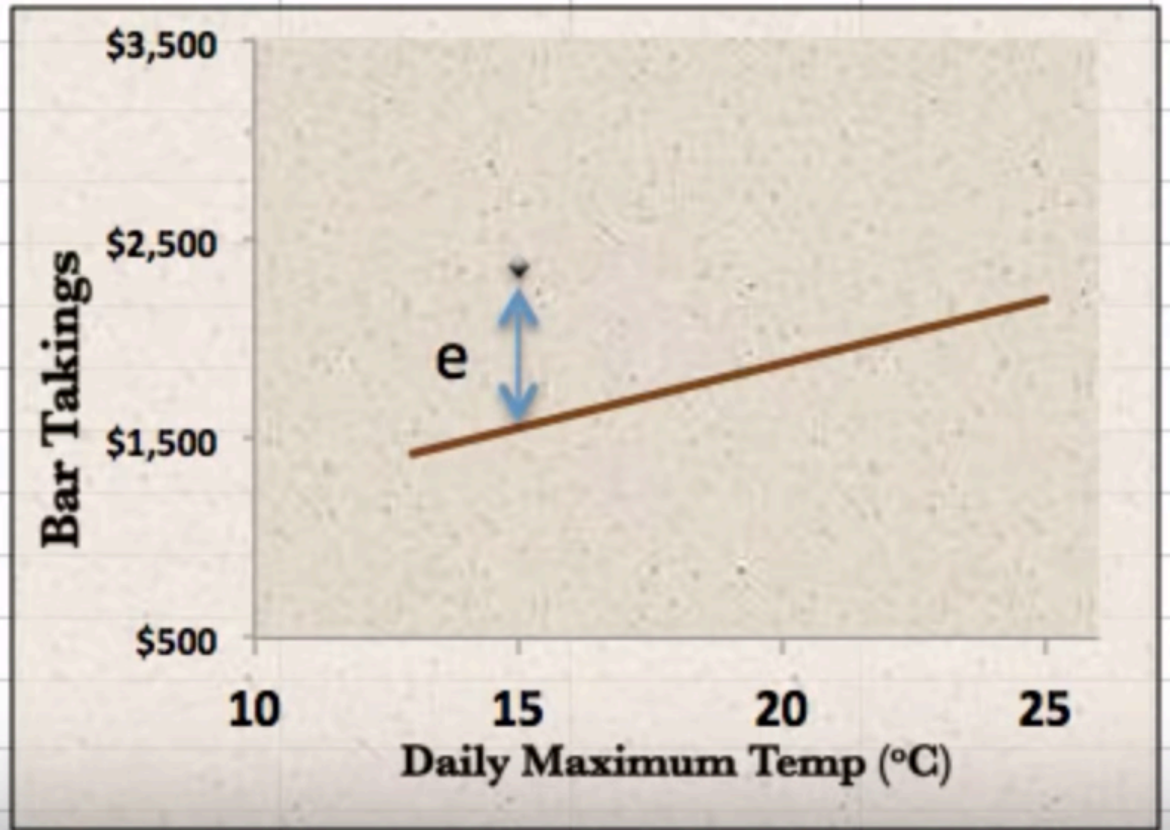


# An Example

POPULATION REGRESSION FUNCTION

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

| Day | Takings | Temp (°C) |
|---|---|---|
| 5-Aug | $1,602 | 14 |
| 12-Aug | $1,688 | 16 |
| 19-Aug | $2,017 | 19 |
| 26-Aug | $1,100 | 13 |
| 2-Sep | $1,609 | 20 |
| 9-Sep | $1,880 | 22 |
| 16-Sep | $997 | 15 |
| 23-Sep | $2,366 | 15 |
| 30-Sep | $2,280 | 25 |

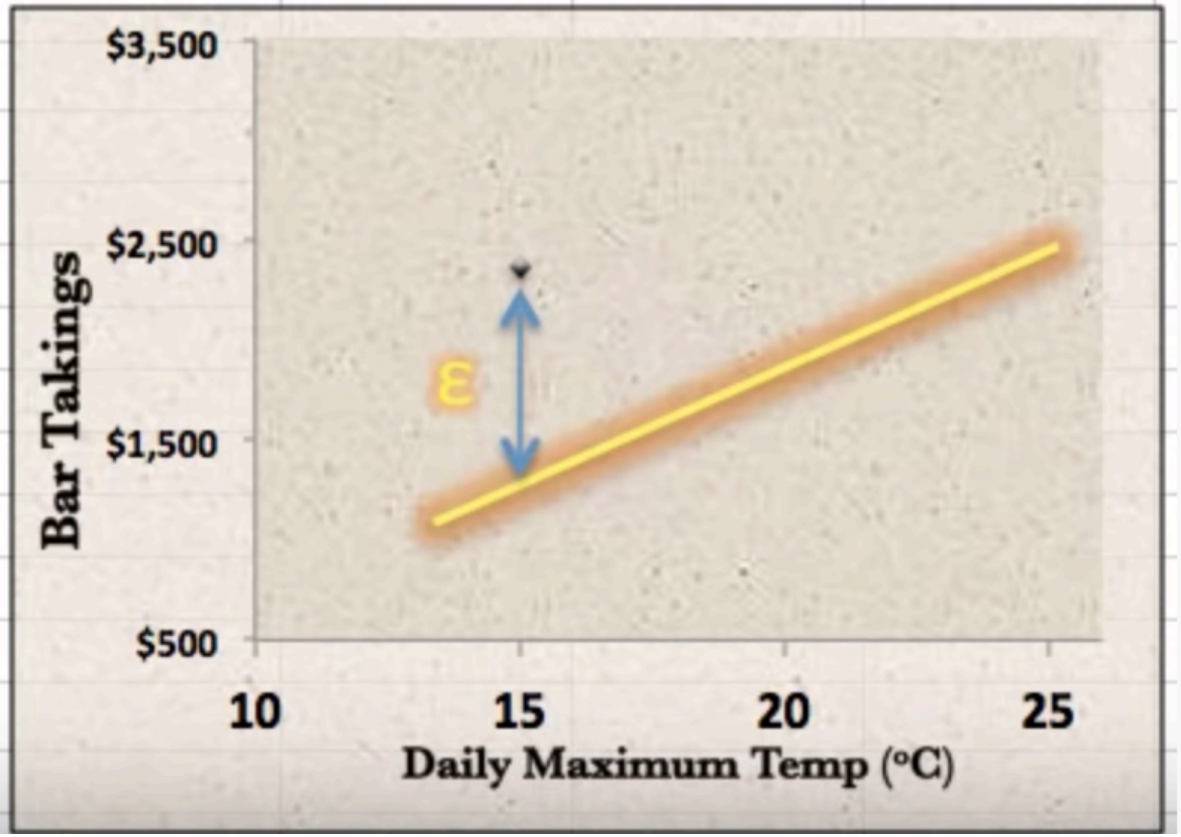


# An Example

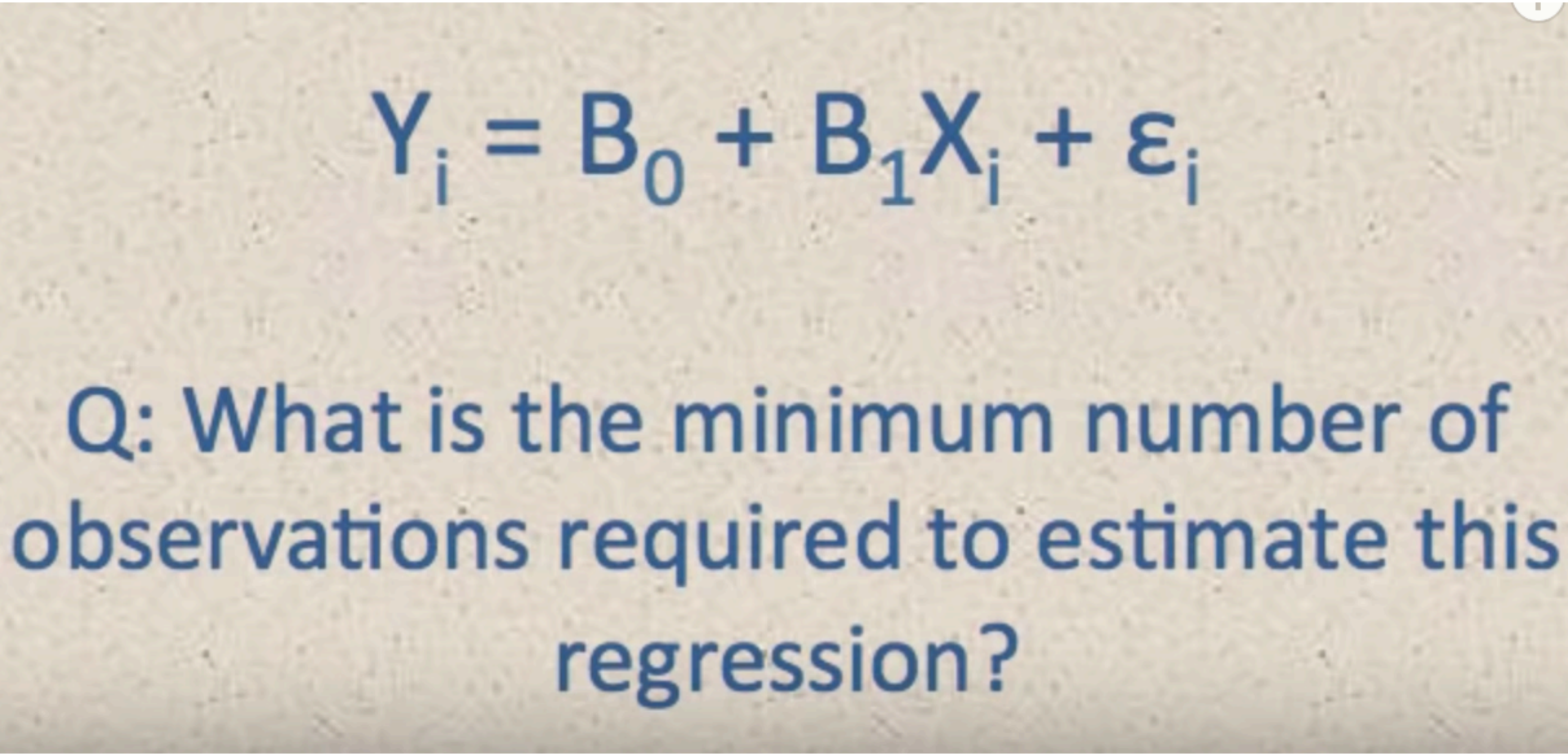
POPULATION REGRESSION FUNCTION

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

| Day | Takings | Temp (°C) |
|---|---|---|
| 5-Aug | $1,602 | 14 |
| 12-Aug | $1,688 | 16 |
| 19-Aug | $2,017 | 19 |
| 26-Aug | $1,100 | 13 |
| 2-Sep | $1,609 | 20 |
| 9-Sep | $1,880 | 22 |
| 16-Sep | $997 | 15 |
| 23-Sep | $2,366 | 15 |
| 30-Sep | $2,280 | 25 |

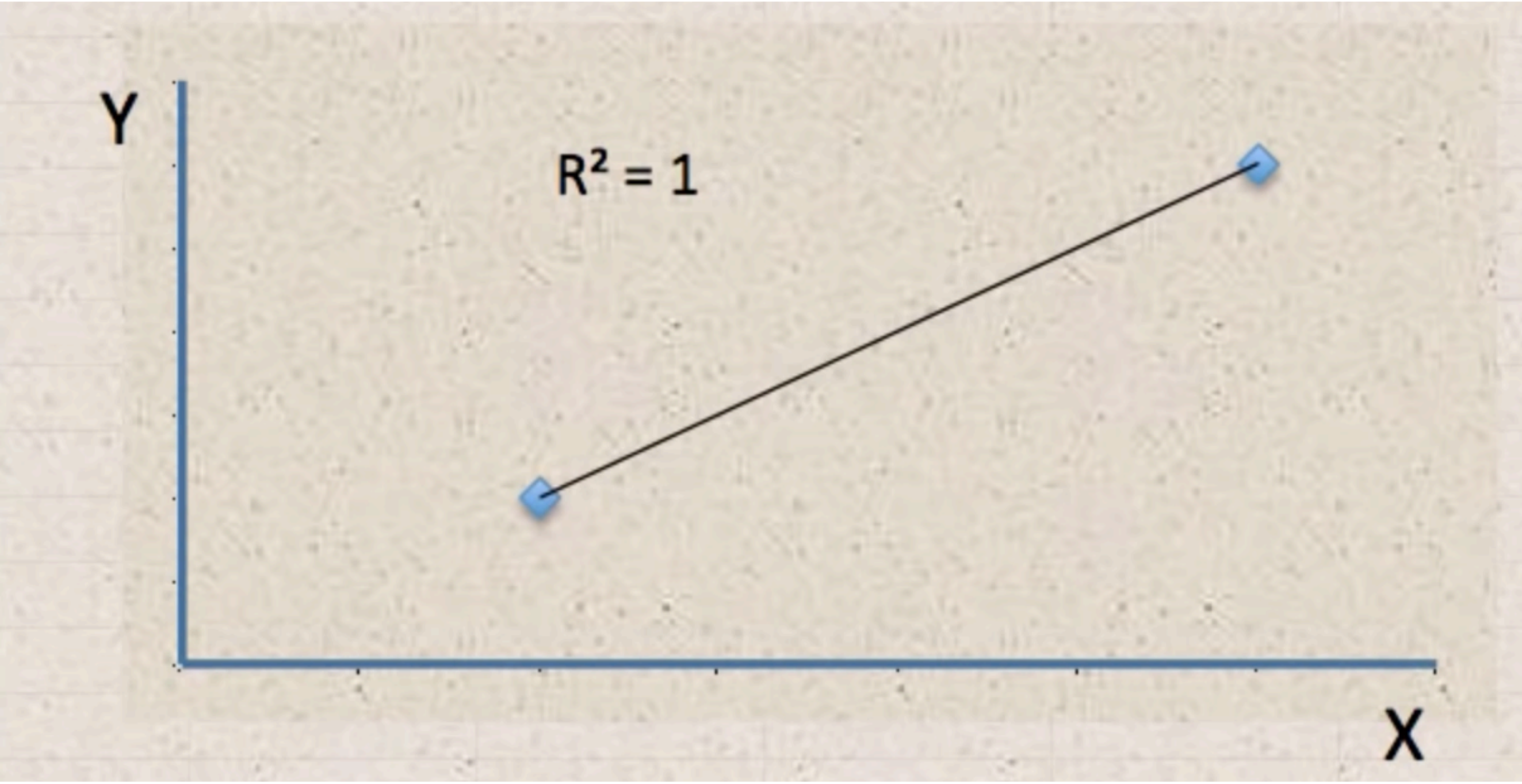


# Degrees of Freedom

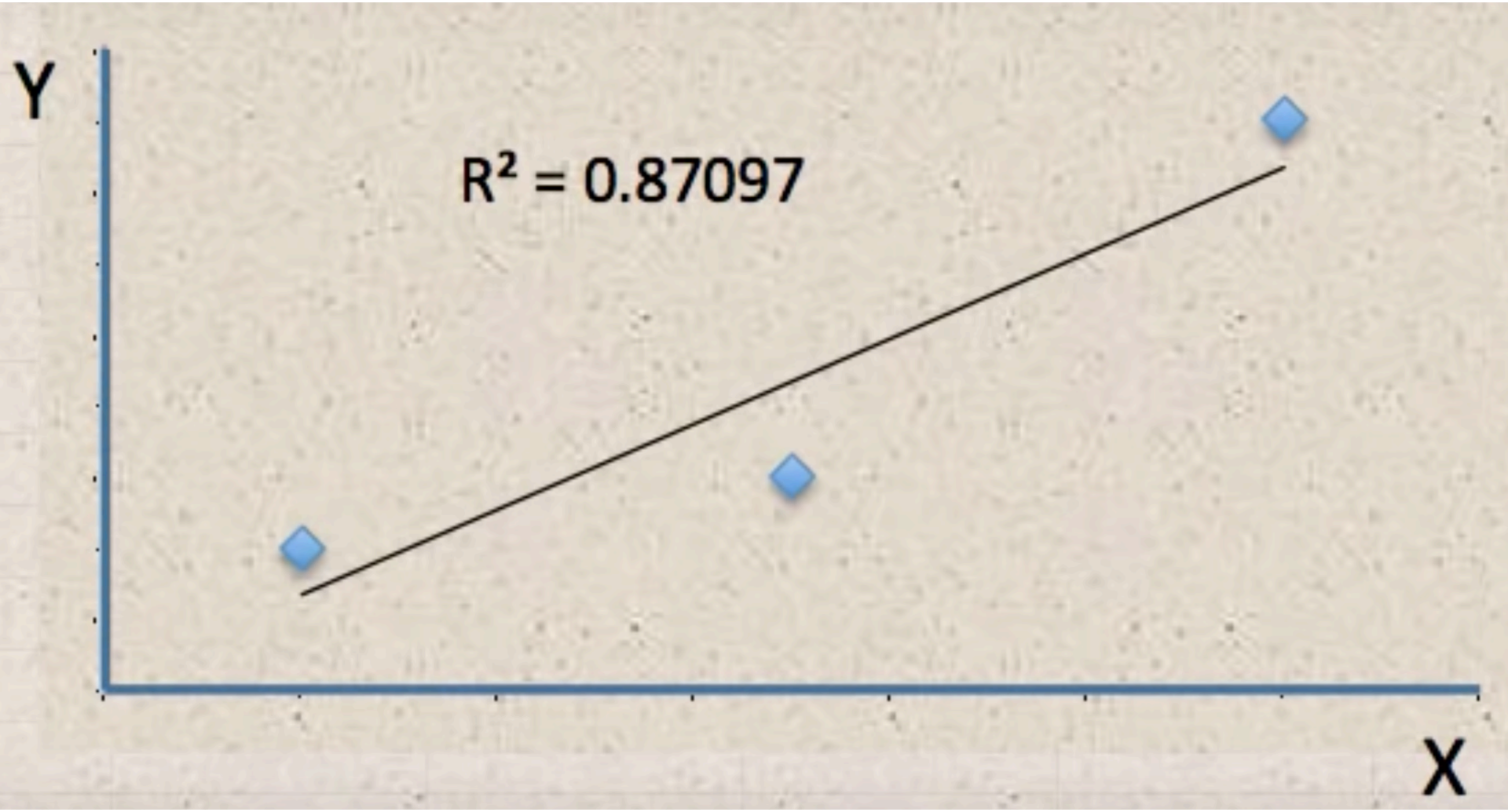



https://www.youtube.com/watch?v=4otEcA3gjLk

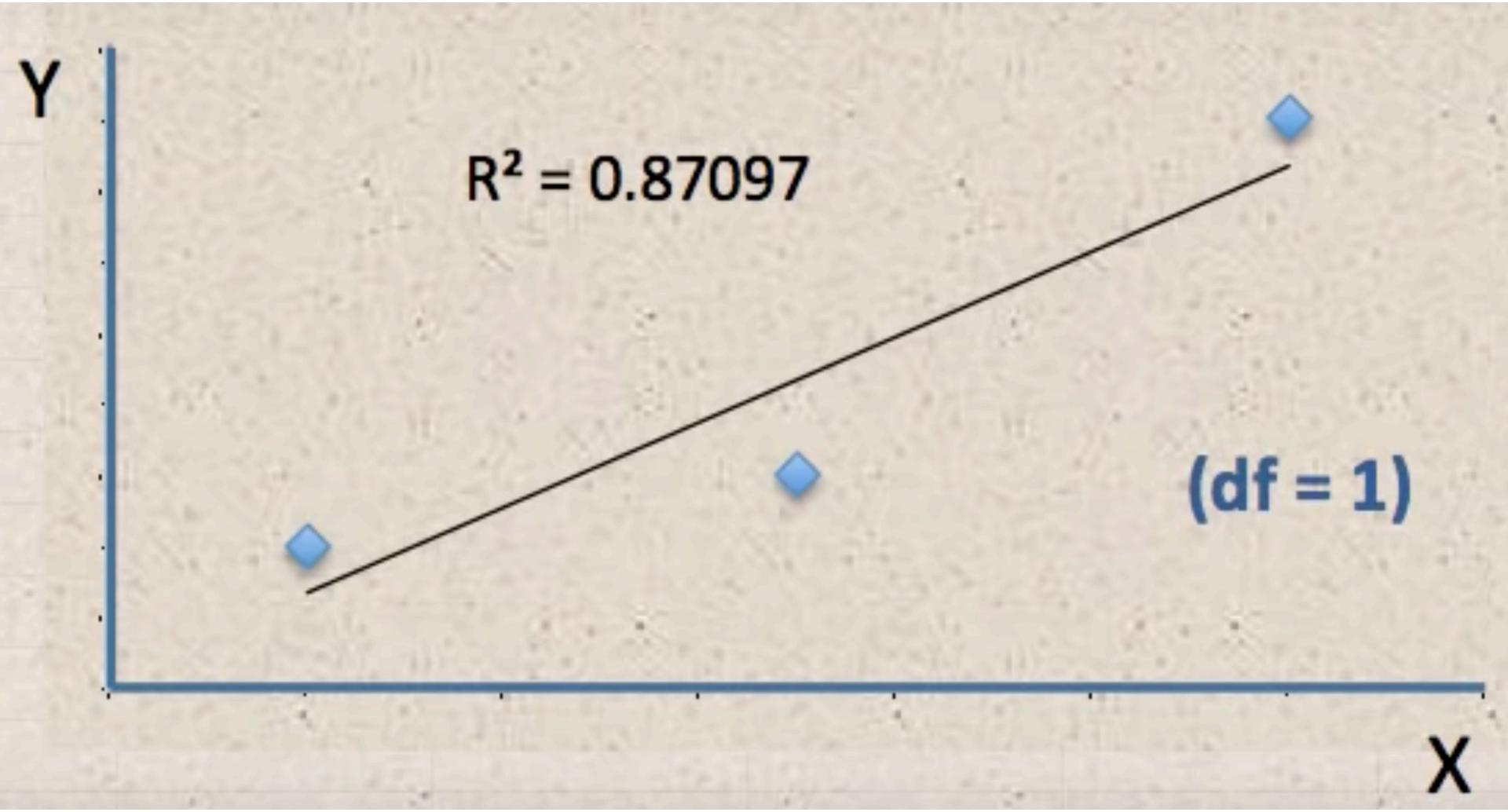
# Degrees of Freedom



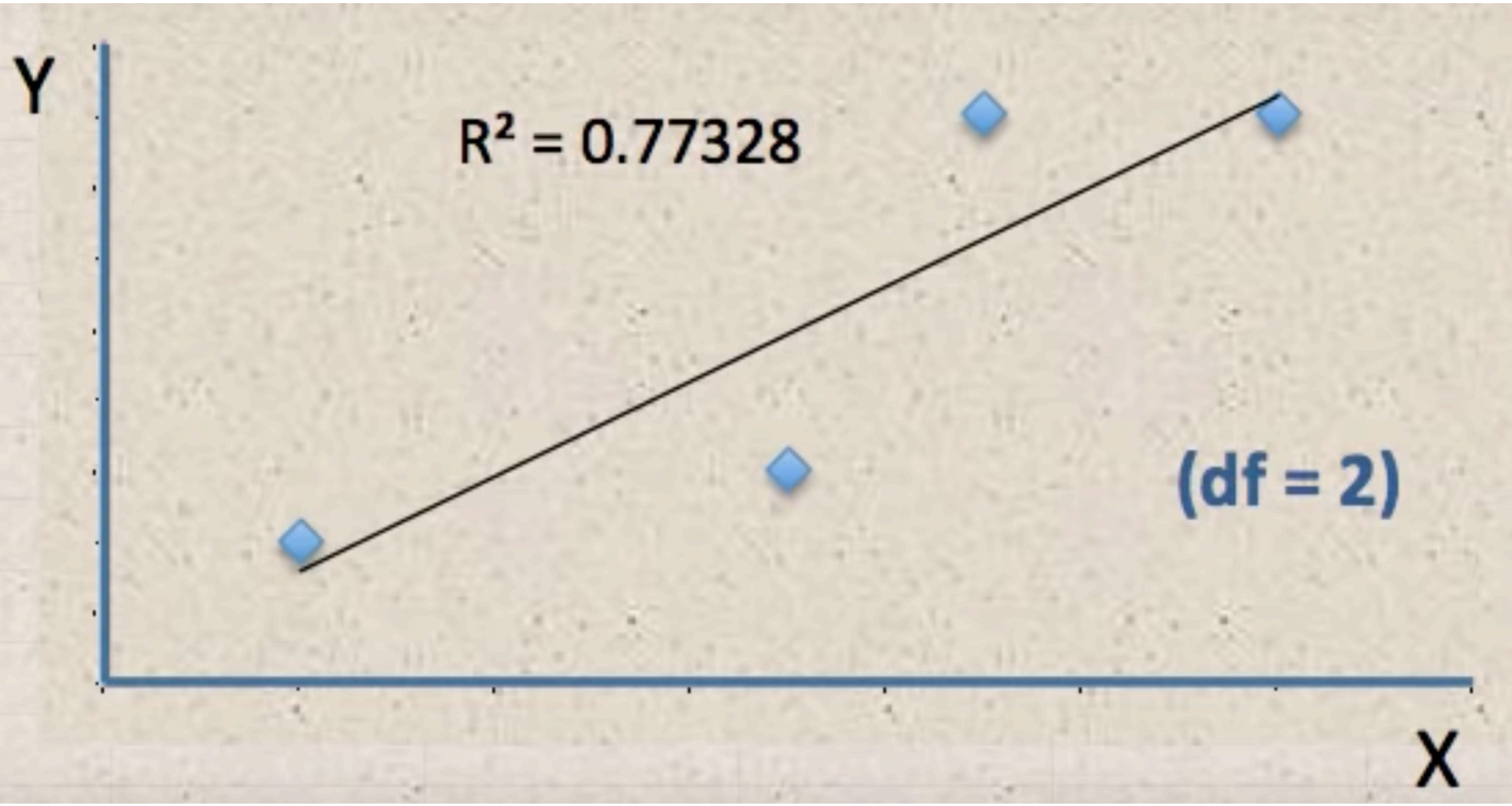
# Degrees of Freedom



# Degrees of Freedom



# Degrees of Freedom

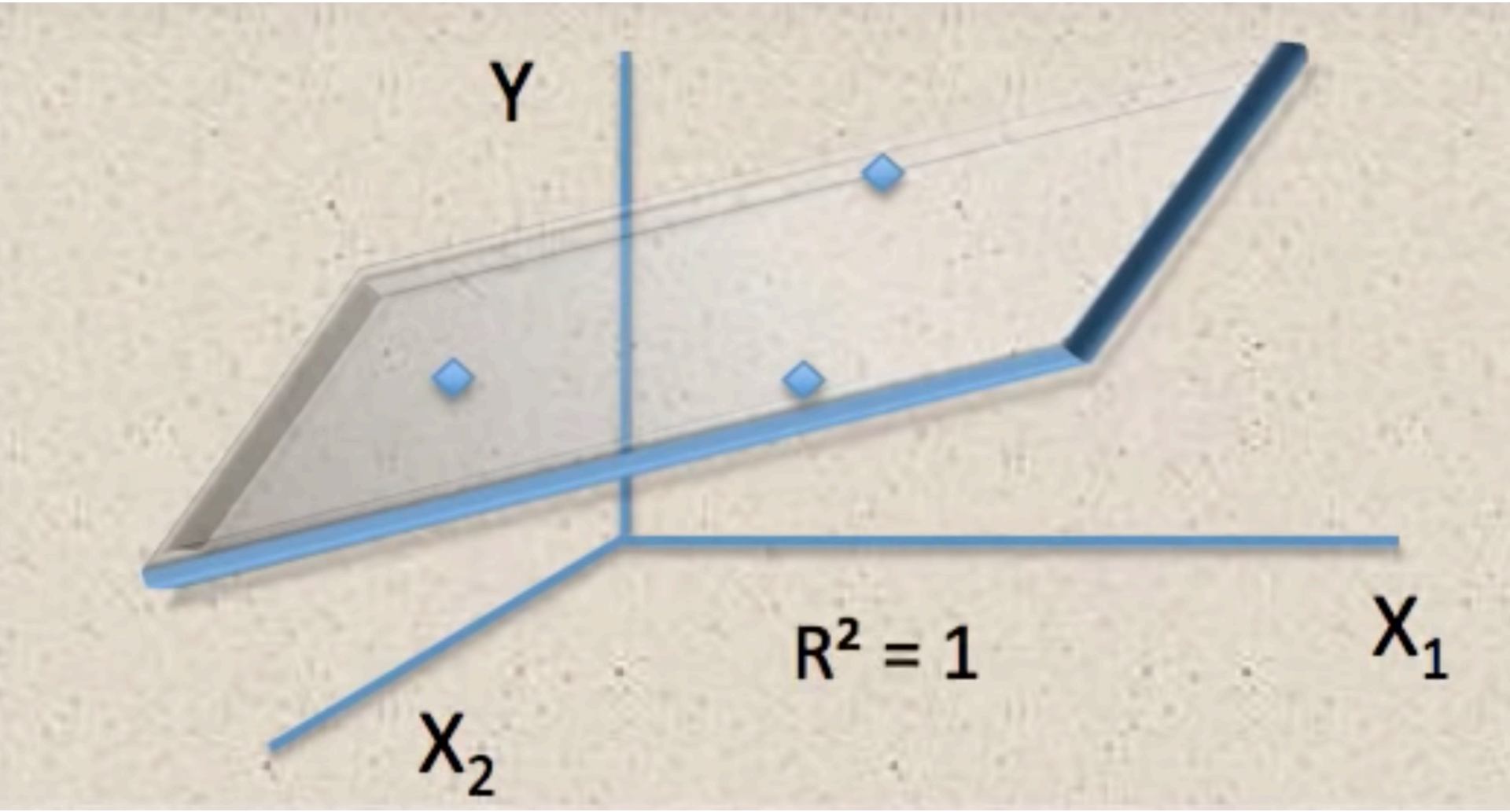


# Degrees of Freedom

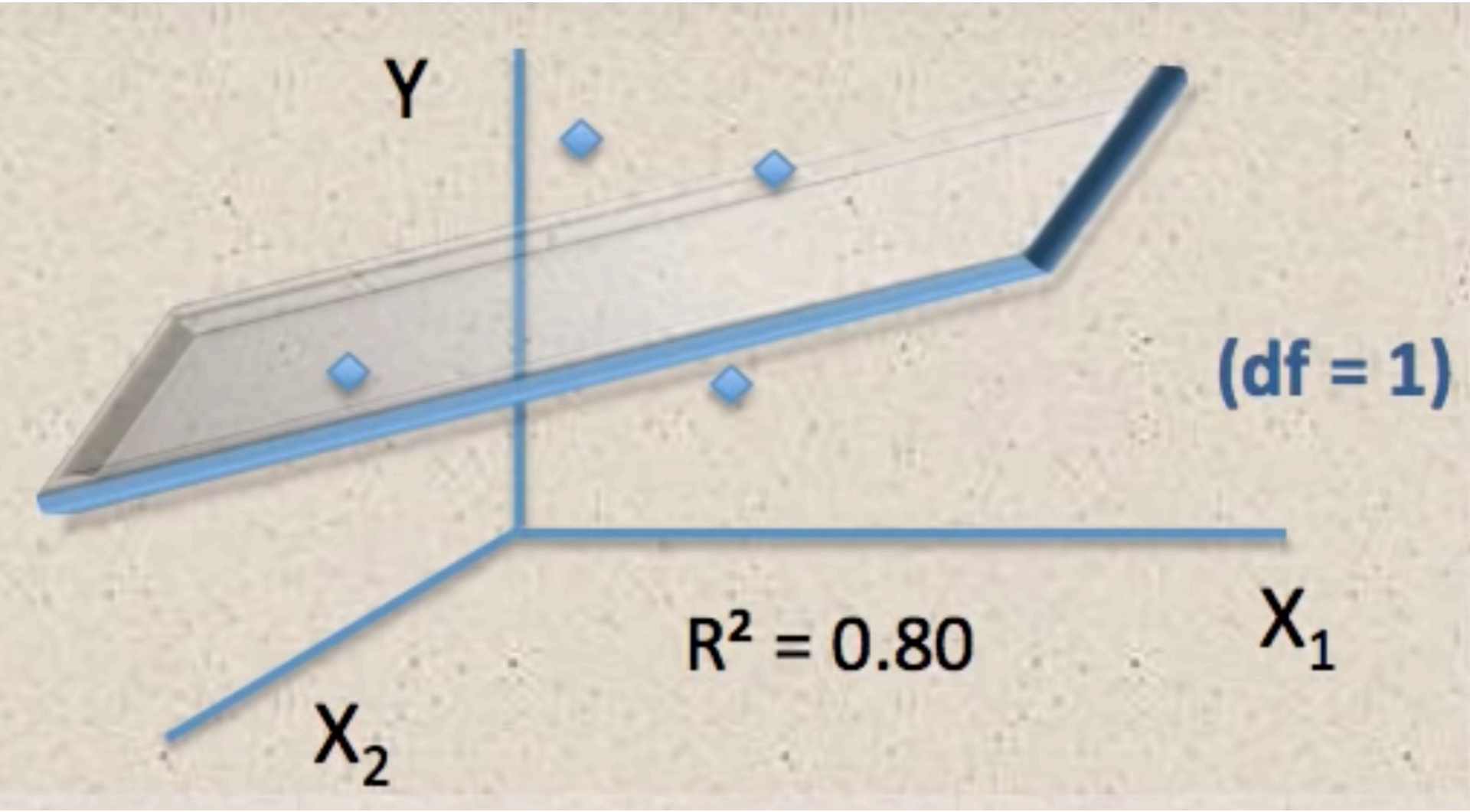
$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \varepsilon_i$$

Q: What is the minimum number of observations required to estimate this regression?

# Degrees of Freedom



# Degrees of Freedom



# Degrees of Freedom

$$df = n-k-1$$

k= the number of explanatory (X) variables

# Degrees of Freedom

Question: How does degrees of freedom relate to R-squared?

As df decreases,
(ie. more variables added to a given model)
R-squared will ONLY increase

# Adjusted $R^2$

$$\bar{R}^2 = \begin{cases} 1 - (1 - R^2)\dfrac{n-1}{n-k-1} \\ \textbf{OR} \\ 1 - \left(\dfrac{SSE}{SST}\right)\dfrac{n-1}{n-k-1} \end{cases}$$

# Adjusted $R^2$

as k increases, Adj $R^2$ will tend to decrease, reflecting the reduced power in the model.

# Adjusted $R^2$

| number of observations, n | number of variables, k | $R^2$ | Adj-$R^2$ |
|---|---|---|---|
| 25 | 4 | 0.71 | 0.6520 |
| 25 | 5 | 0.76 | 0.6968 |
| 25 | 6 | 0.78 | 0.7067 |
| 25 | 7 | 0.79 | 0.7035 |
| | | | |
| 10 | 4 | 0.71 | 0.4780 |
| 10 | 5 | 0.76 | 0.4600 |
| 10 | 6 | 0.78 | 0.3400 |
| 10 | 7 | 0.79 | 0.0550 |

# End of Example with British Pubs

# Business Problem

# Business Problem

- A company needs to build a predictive model for estimating sales of one of its main line of products
- The company holds an annual sales budget and needs to know how best invest this money in order to maximize product sales
- Options for advertising budget include purchasing ads in either Youtube, Facebook, or  newspaper
- The company holds historical records of 200 campaigns

# Business Problem

Best Media for Marketing Campaign

| Period | Youtube | Facebook | Newspaper | Sales |
|---|---|---|---|---|
| 1 | 276.1 | 45.4 | 83.0 | 26.5 |
| 2 | 53.4 | 47.2 | 54.1 | 12.5 |
| 3 | 20.6 | 55.1 | 83.2 | 11.2 |
| 4 | 181.8 | 49.6 | 70.2 | 22.2 |

# Business Problem

- What are future sales if we invest marketing budget in Youtube?

# Hands-on Exercise with R

- > install.packages("datarium")
- > data("marketing", package = "datarium")
- > head(marketing, 4)

- youtube facebook newspaper sales
- 1 276.12 45.36 83.04 26.52
- 2 53.40 47.16 54.12 12.48
- 3 20.64 55.08 83.16 11.16
- 4 181.80 49.56 70.20 22.20

```
> cor(marketing$sales, marketing$youtube)
[1] 0.7822244
```

se

- The correlation coefficient measures the level of the association between two variables x and y
- Its value ranges between -1 (perfect negative correlation: when x increases, y decreases) and +1 (perfect positive correlation: when x increases, y increases).
- A value closer to 0 suggests a weak relationship between the variables.
- A low correlation ($-0.2 < x < 0.2$) probably suggests that much of variation of the outcome variable (y) is not explained by the predictor (x). In such case, we should probably look for better predictor variables.
- In our example, the correlation coefficient is large enough, so we can continue by building a linear model of y as a function of x.

- **sales = b0 + b1 * youtube**

```
> model <- lm(sales ~ youtube, data = marketing)
> model

Call:
lm(formula = sales ~ youtube, data = marketing)

Coefficients:
(Intercept)     youtube
   8.43911      0.04754
```

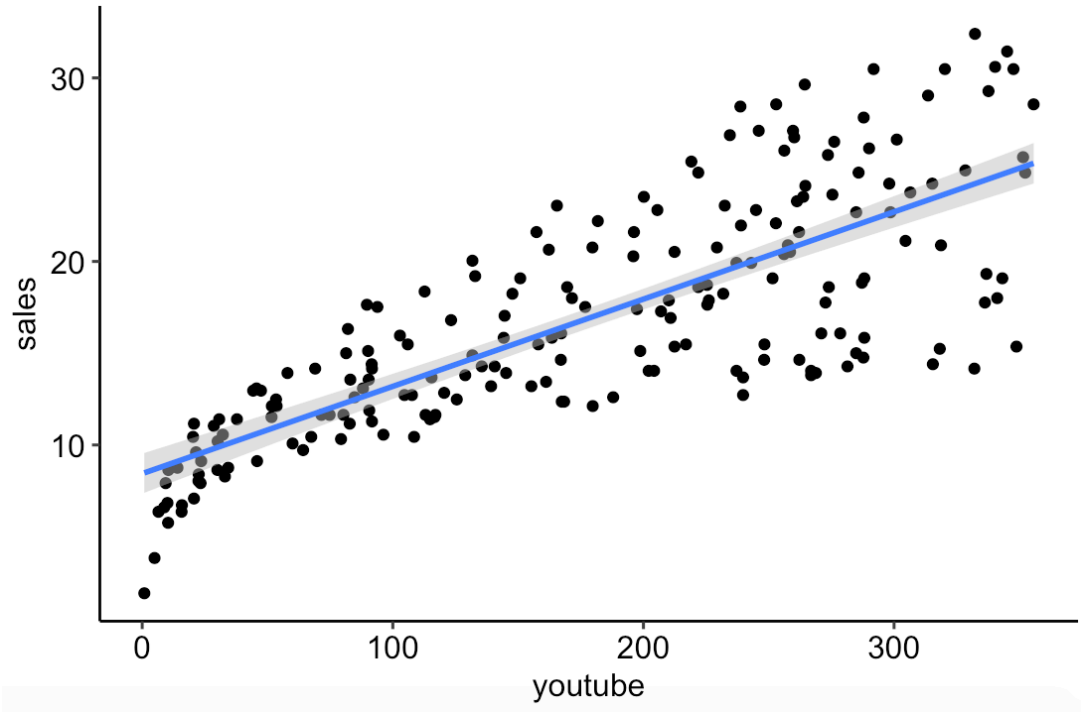
# Interpretation

- The estimated regression line equation can be written as follow:

  **sales = 8.44 + 0.048*youtube**

- The intercept (b0) is 8.44. It can be interpreted as the predicted sales unit for a zero youtube advertising budget. Recall that, we are operating in units of thousand dollars. This means that, for a youtube advertising budget equal zero, we can expect a sale of 8.44 *1000 = 8440 dollars.
- The regression beta coefficient for the variable youtube (b1), also known as the slope, is 0.048
- This means that, for a youtube advertising budget equal to 1000 dollars, we can expect an increase of 48 units (0.048*1000) in sales
- That is, sales = 8.44 + 0.048*1000 = 56.44 units. As we are operating in units of thousand dollars, this represents a sale of 56440 dollars.

```
> ggplot(marketing,
aes(youtube, sales)) +
+ geom_point() +
+ stat_smooth(method
= lm)
`geom_smooth()` using
formula 'y ~ x'
```



# Model Assessment

Before using this formula to predict future sales, you should make sure that this model is statistically significant, that is:

There is a statistically significant relationship between the predictor and the outcome variables

The model that we built fits very well the data in our hand.

# > summary(model)

```
Call:
lm(formula = sales ~ youtube, data = marketing)

Residuals:
    Min      1Q  Median      3Q     Max
-10.059  -2.353  -0.228   2.483   8.650

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.43911    0.54941    15.4   <2e-16 ***
youtube      0.04754    0.00269    17.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.91 on 198 degrees of freedom
Multiple R-squared:  0.612,  Adjusted R-squared:  0.61
F-statistic:  312 on 1 and 198 DF,  p-value: <2e-16
```

# Model Assessment

- The summary outputs shows 6 components, including:
- **Call**. Shows the function call used to compute the regression model.
- **Residuals**. Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value.
- **Coefficients**. Shows the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.
- **Residual standard error** (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits to our data.

# Coefficient Significance

- The coefficients table, in the model statistical summary, shows:

- the estimates of the beta coefficients
- the standard errors (SE), which defines the accuracy of beta coefficients. For a given beta coefficient, the SE reflects how the coefficient varies under repeated sampling. It can be used to compute the confidence intervals and the t-statistic.
- the t-statistic and the associated p-value, which defines the statistical significance of the beta coefficients.
- ## Estimate Std. Error t value Pr(>|t|)
- ## (Intercept) 8.4391 0.54941 15.4 1.41e-35
- ## youtube 0.0475 0.00269 17.7 1.47e-42

# t-statistic and p-values

- For a given predictor, the t-statistic and its associated p-value test whether or not there is a statistically significant relationship between a given predictor and the outcome variable, that is whether or not the beta coefficient of the predictor is significantly different from zero.
- The statistical hypotheses are as follow:

- **Null hypothesis (H0): the coefficients are equal to zero (i.e., no relationship between x and y)**
- **Alternative Hypothesis (Ha): the coefficients are not equal to zero (i.e., there is some relationship between x and y)**

# t-statistic and p-values

- Mathematically, for a given beta coefficient (b), the t-test is computed as t = (b - 0)/SE(b), where SE(b) is the standard error of the coefficient b. The t-statistic measures the number of standard deviations that b is away from 0. Thus a large t-statistic will produce a small p-value.
- The higher the t-statistic (and the lower the p-value), the more significant the predictor. The symbols to the right visually specifies the level of significance. The line below the table shows the definition of these symbols; one star means $0.01 < p < 0.05$. The more the stars beside the variable's p-value, the more significant the variable.
- A statistically significant coefficient indicates that there is an association between the predictor (x) and the outcome (y) variable.

# t-statistic and p-values

- In our example, both the p-values for the intercept and the predictor variable are highly significant, so we can reject the null hypothesis and accept the alternative hypothesis, which means that there is a significant association between the predictor and the outcome variables.

# Standard errors and confidence intervals

- The standard error measures the variability/accuracy of the beta coefficients. It can be used to compute the confidence intervals of the coefficients.
- For example, the 95% confidence interval for the coefficient b1 is defined as b1 +/- 2*SE(b1), where:
- the lower limits of b1 = b1 - 2*SE(b1) = 0.047 - 2*0.00269 = 0.042
- the upper limits of b1 = b1 + 2*SE(b1) = 0.047 + 2*0.00269 = 0.052
- That is, there is approximately a 95% chance that the interval [0.042, 0.052] will contain the true value of b1. Similarly, the 95% confidence interval for b0 can be computed as b0 +/- 2*SE(b0).

# confint(model)

- ## 2.5 % 97.5 %
- ## (Intercept) 7.3557 9.5226
- ## youtube 0.0422 0.0528

# Model accuracy

- Once you identified that, at least, one predictor variable is significantly associated to the outcome, you should continue the diagnostic by checking how well the model fits the data. This process is also referred to as the *goodness-of-fit*
- The overall quality of the linear regression fit can be assessed using the following three quantities, displayed in the model summary:
- The Residual Standard Error (RSE).
- The R-squared (R2)
- F-statistic
- ## rse r.squared f.statistic p.value
- ## 1 3.91 0.612 312 1.47e-42

# **Residual standard error** (RSE)

- The RSE (also known as the model sigma) is the residual variation, representing the average variation of the observation points around the fitted regression line. This is the standard deviation of residual errors
- RSE provides an absolute measure of patterns in the data that can't be explained by the model. When comparing two models, the model with the small RSE is a good indication that this model fits the best the data
- Dividing the RSE by the average value of the outcome variable will give you the prediction error rate, which should be as small as possible
- In our example, RSE = 3.91, meaning that the observed sales values deviate from the true regression line by approximately 3.9 units in average
- Whether or not an RSE of 3.9 units is an acceptable prediction error is subjective and depends on the problem context. However, we can calculate the percentage error. In our data set, the mean value of sales is 16.827, and so the percentage error is 3.9/16.827 = 23%.
- sigma(model)*100/mean(marketing$sales)
- [1] 23.2

# R-squared and Adjusted R-squared:

- The R-squared (R2) ranges from 0 to 1 and represents the proportion of information (i.e. variation) in the data that can be explained by the model. The adjusted R-squared adjusts for the degrees of freedom.
- The R2 measures, how well the model fits the data. For a simple linear regression, R2 is the square of the Pearson correlation coefficient.
- A high value of R2 is a good indication. However, as the value of R2 tends to increase when more predictors are added in the model, such as in multiple linear regression model, you should mainly consider the adjusted R-squared, which is a penalized R2 for a higher number of predictors.
- An (adjusted) R2 that is close to 1 indicates that a large proportion of the variability in the outcome has been explained by the regression model.
- A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

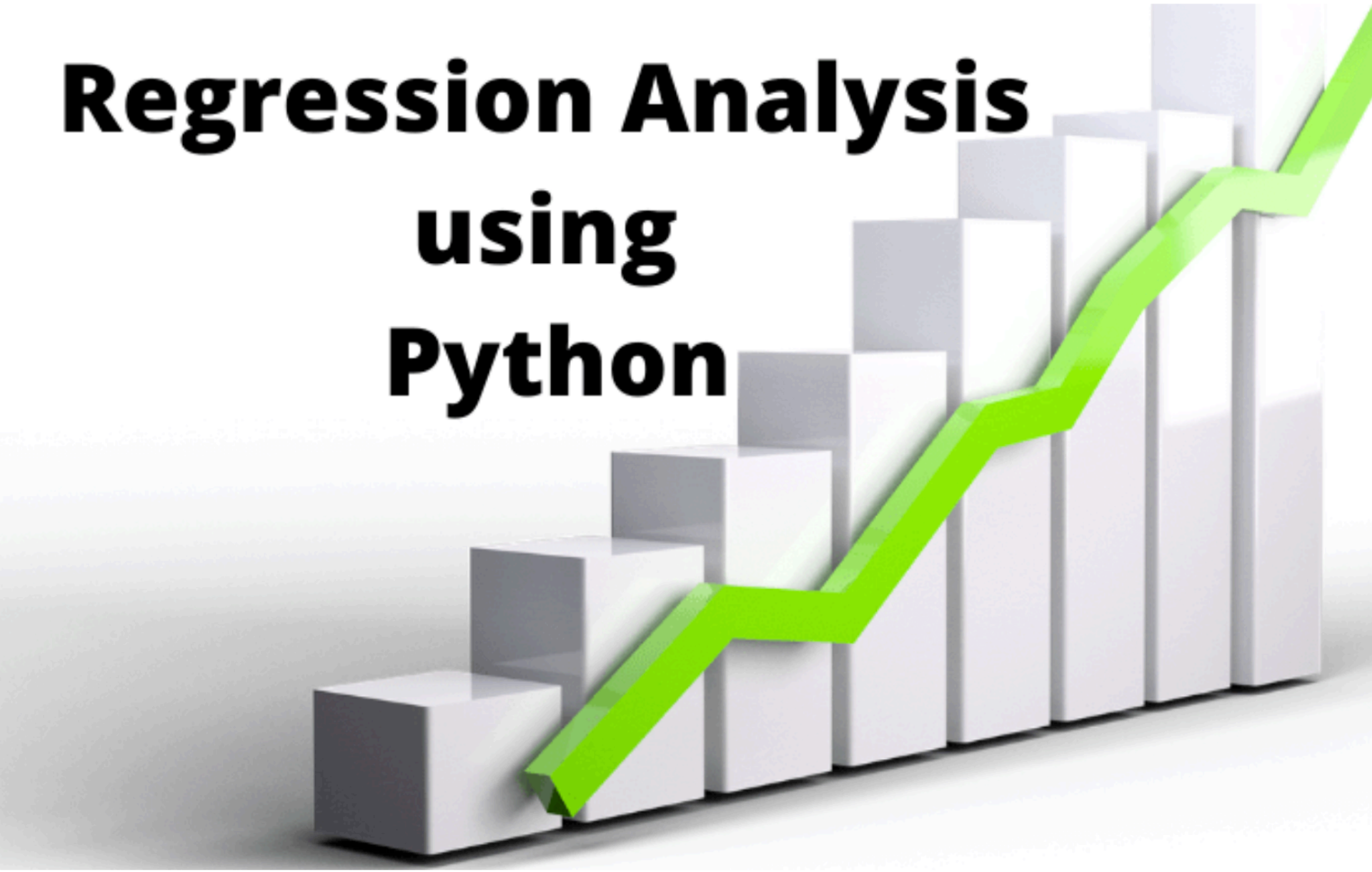
# F-statistic

- The F-statistic gives the overall significance of the model. It assess whether at least one predictor variable has a non-zero coefficient
- In a simple linear regression, this test is not really interesting since it just duplicates the information in given by the t-test, available in the coefficient table. In fact, the F test is identical to the square of the t test: 312.1 = (17.67)^2. This is true in any model with 1 degree of freedom
- The F-statistic becomes more important once we start using multiple predictors as in multiple linear regression
- A large F-statistic will correspond to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 312.14 producing a p-value of 1.46e-42, which is highly significant

# Summary

- After computing a regression model, a first step is to check whether, at least, one predictor is significantly associated with outcome variables.
- If one or more predictors are significant, the second step is to assess how well the model fits the data by inspecting the Residuals Standard Error (RSE), the R2 value and the F-statistics. These metrics give the overall quality of the model:
  - RSE: Closer to zero the better
  - R-Squared: Higher the better
  - F-statistic: Higher the better

# Regression Analysis using Python



# Practice with the Following Hands-on Exercises

- https://realpython.com/linear-regression-in-python/

# Francisco J. Cantú-Ortiz, PhD

Professor of Computer Science and Artificial Intelligence
Tecnológico de Monterrey
Enago-Academy Advisor for Strategic Alliances

E-mail: fcantu@tec.mx, fjcantor@gmail.com
Cel: +52 81 1050 8294, SNI-2 CVU: 9804
Personal Page: http://semtech.mty.itesm.mx/fcantu/
Facebook: fcantu; Twitter: @fjcantor; Skype: fjcantor
Orcid: 0000-0002-2015-0562
Scopus ID:6701563520
Researcher ID: B-8457-2009
https://www.researchgate.net/profile/Francisco_Cantu-Ortiz
https://scholar.google.com.mx/citations?hl=es&user=45-uuK4AAAAJ
https://itesm.academia.edu/FranciscoJavierCantuOrtiz
Ave. Eugenio Garza Sada No. 2501, Monterrey N.L., C.P. 64849, México