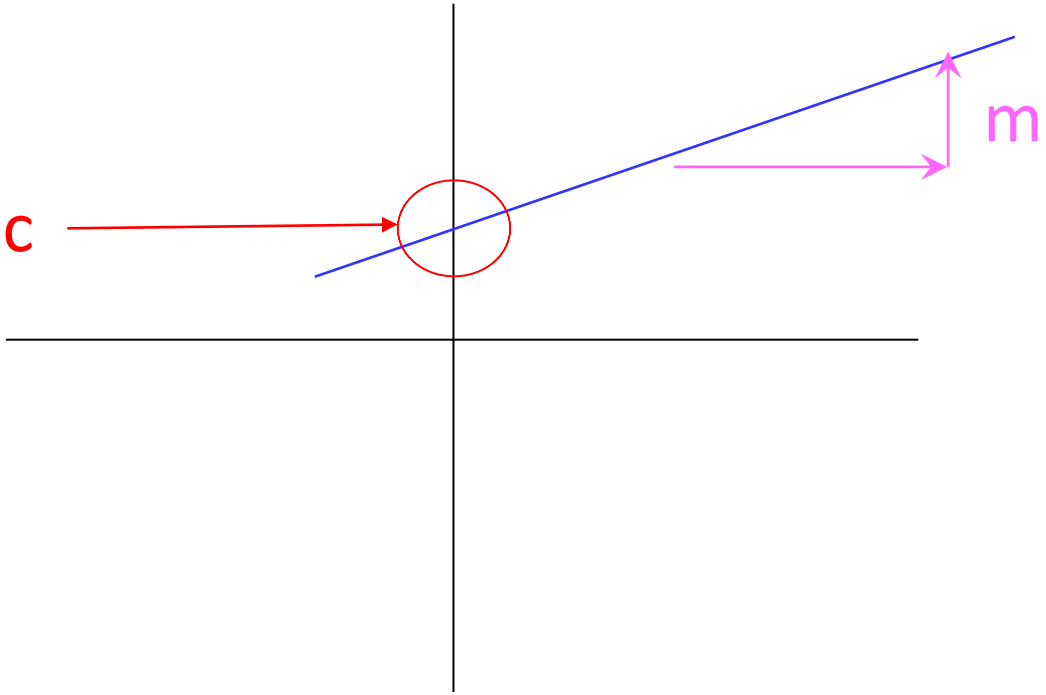
Linear Regression

What is “Linear”?

- Remember this?
- y = mx + c



What’s Slope?

- A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

Prediction

- If you know something about X, this knowledge helps you predict something about Y

Regression

- Regression is a statistical method used to describe the nature of the relationship between variables—that is, positive or negative, linear or nonlinear.
- Regression answers the following questions;
    1. What type of relationship exists?
    2. What kind of predictions can be made from the relationship?

Regression

1. What type of relationship exists?

- There are two types of relationships: simple and multiple.

- In a simple relationship, there are two variables:
  - **independent variable** (predictor variable) and a
  - **dependent variable** (response variable).

- In a multiple relationship, there are two or more independent variables that are used to predict one dependent variable.

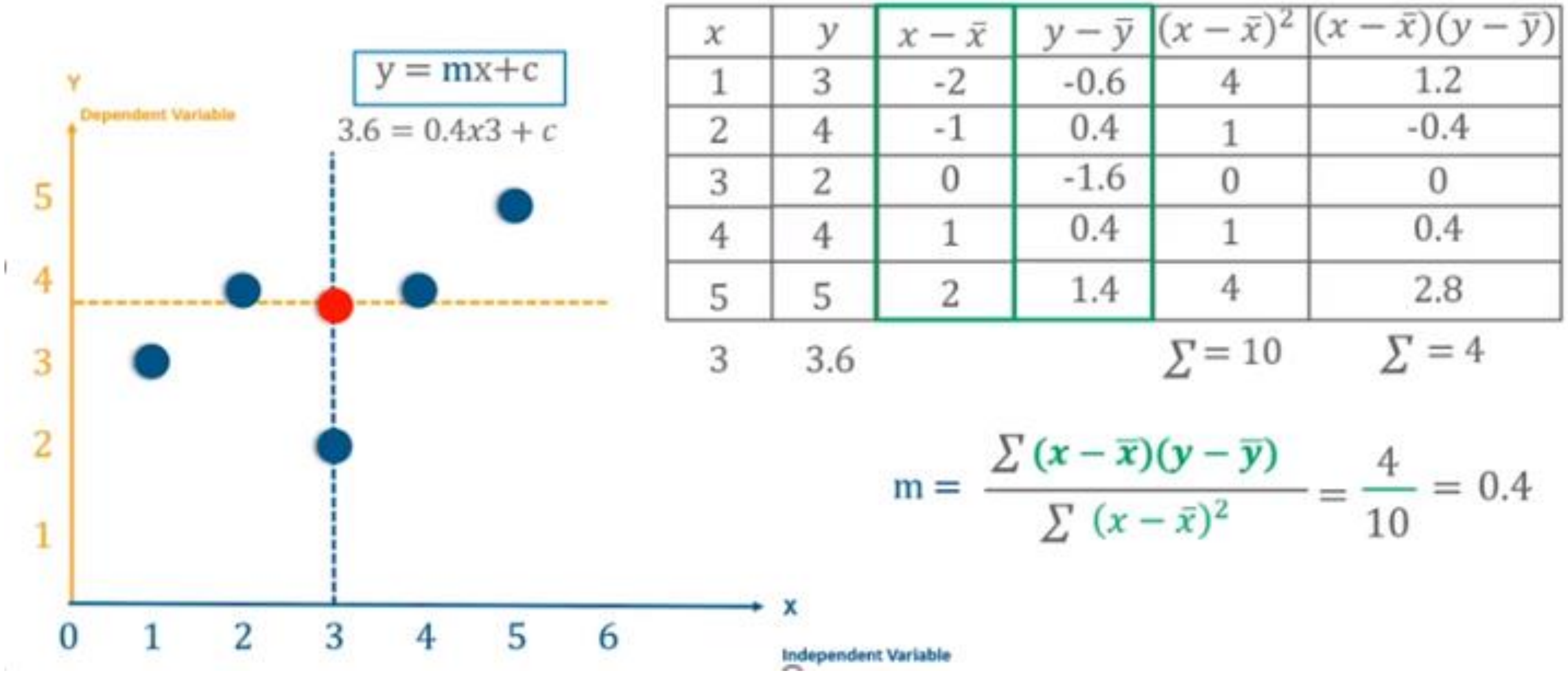
Regression

2. What kind of predictions can be made from the relationship?

- Predictions are made in all areas and daily. Examples include
    - weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price predictions, and sports predictions.
- Some predictions are more accurate than others, due to the strength of the relationship
- That is, the stronger the relationship is between variables, the more accurate the prediction is.

Linear Regression

Ordinary Least Squares (OLS)



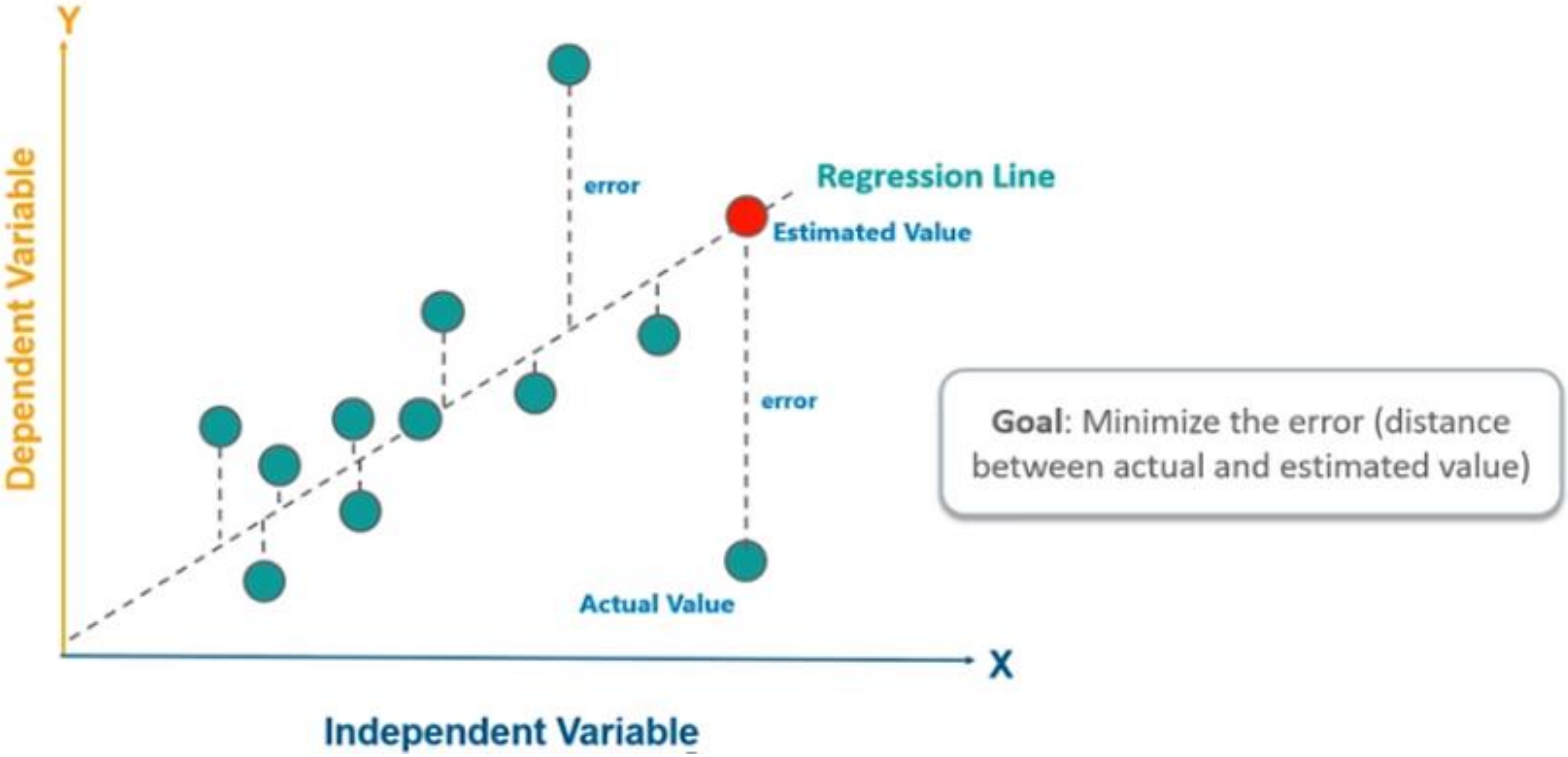
x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
3	3.6			$\sum = 10$	$\sum = 4$

$$m = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{4}{10} = 0.4$$

Ordinary least squares - line of best fit

- regression line passes through
- calculation of slope and intercept
- interpretation of slope & intercept

OLS



Goal: Minimize the error (distance between actual and estimated value)

OLS

$y_i = mx_i + c + \varepsilon_i$

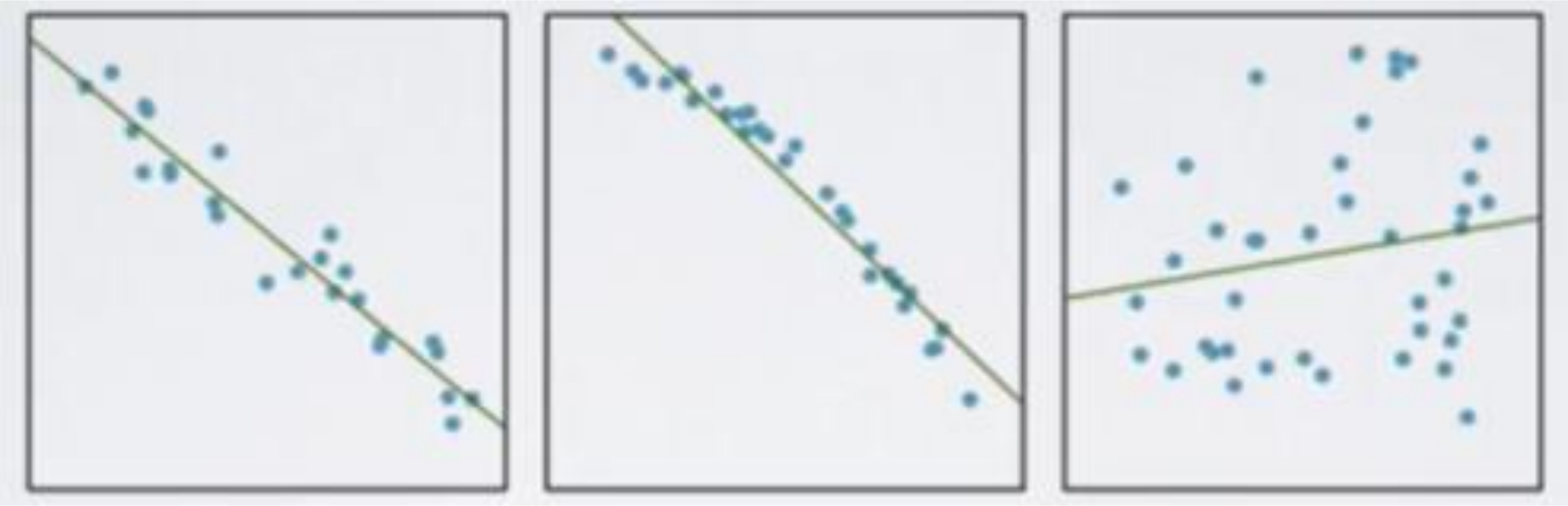
- You can consider the OLS as a strategy to obtain, from your model, a 'straight line' which is as close as possible to your data points
- Even though OLS is not the only optimization strategy, it is the most popular for this kind of tasks, since the outputs of the regression (that are, coefficients) are unbiased estimators of the real values of **m** and **c**
- The Gauss-Markov theorem states that OLS produces estimates that are better than estimates from all other linear model estimation methods when some assumptions hold true
- As long as your model satisfies the OLS assumptions for linear regression, you can rest easy knowing that you're getting the best possible estimates

Assumptions of OLS

- Linear Relationship between the features (X) and target (Y)
- Little or no Multicollinearity between the features
- The error term has a constant variance (no heteroscedasticity)
- Normal distribution of error terms

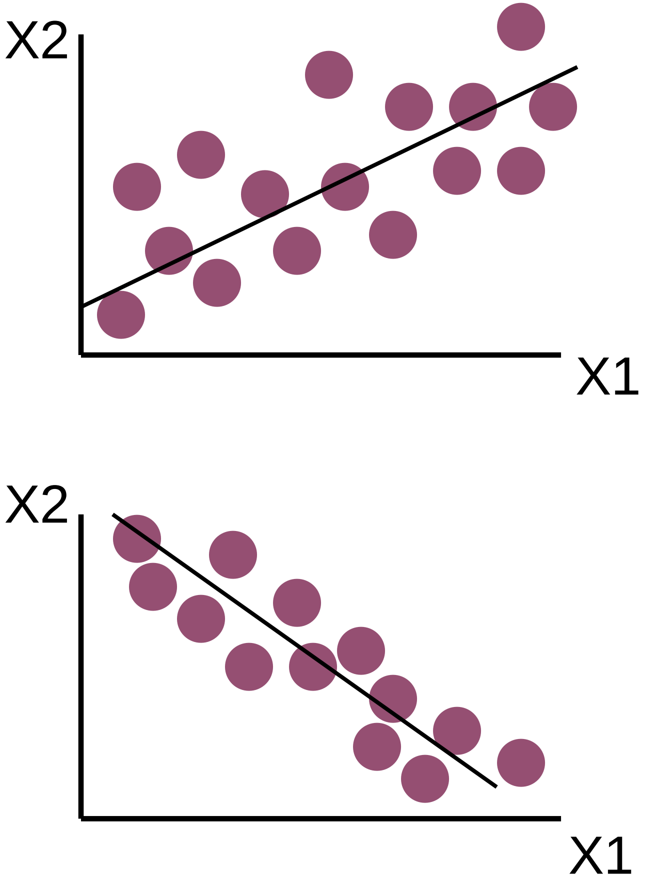
Linear Relationship between the features and target

- Relationship between the features (x) and target (y)
- We can check this using a scatter plot (straight line trend)

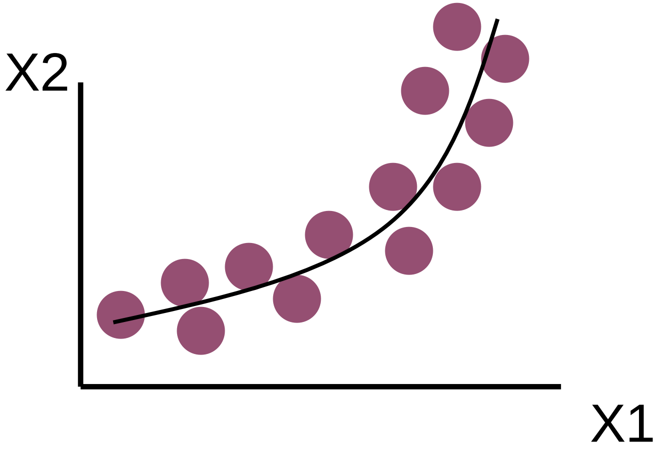


Little or no Multicollinearity between the features

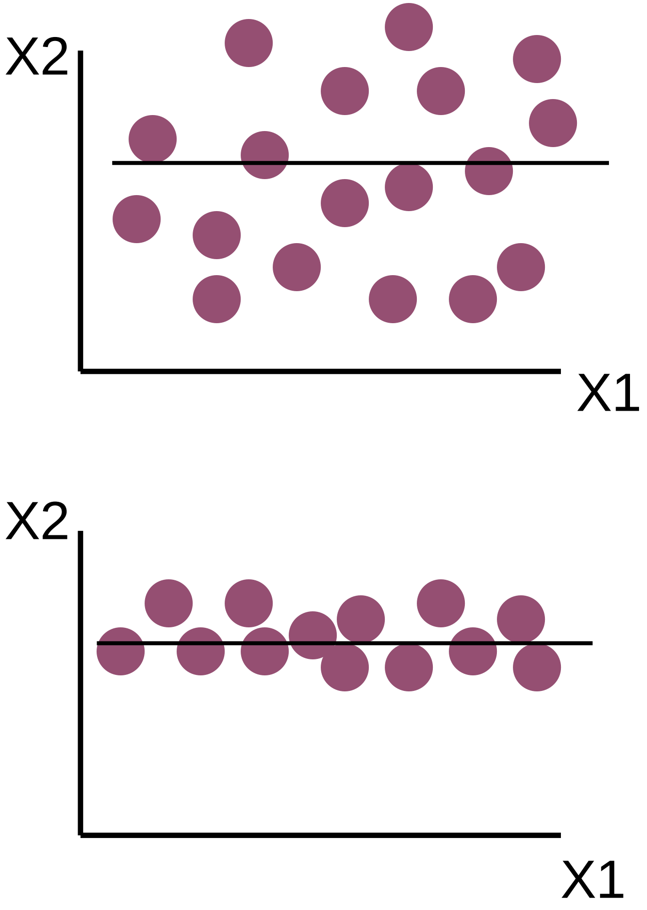
Linear relationships



Curvilinear relationships

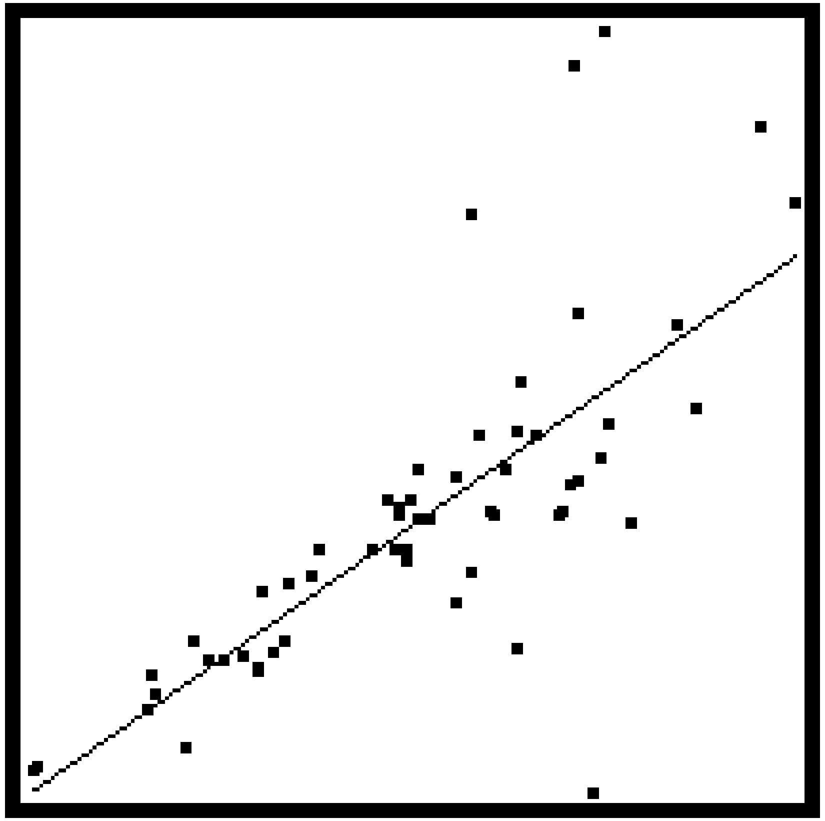


No relationship



Homoscedasticity v/s Heteroscedasticity

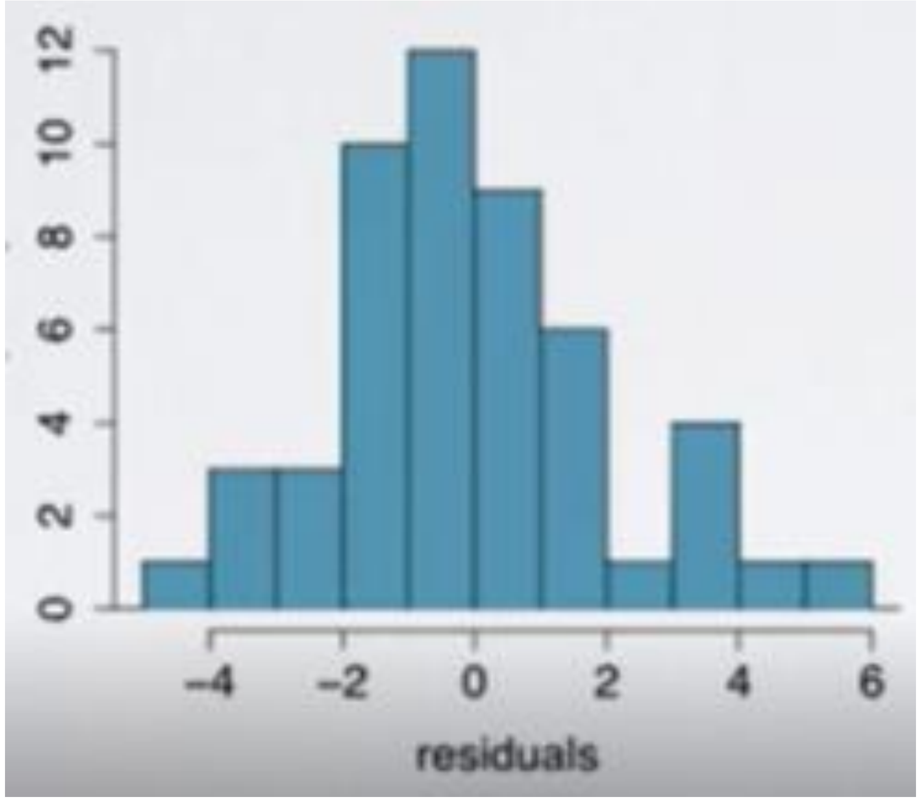
- Homoscedasticity means that the variance around the regression line is the same for all values of the predictor variable (X)
- This plot shows a violation of this assumption. For the lower values on the X-axis, the points are all very near the regression line
- For the higher values on the X-axis, there is much more variability around the regression line.



Normal distribution of error terms

$y_i = mx_i + c + \varepsilon_i$

- Residuals (ε_i) should nearly normally distributed and centered around 0
- Check using a histogram of residuals



$y_i = \alpha + \beta x_i + \varepsilon_i$

Evaluation Metrics in Regression Models (Cont.)

Mean Absolute Error (MAE) is the mean of the absolute value of the errors

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

Mean Squared Error (MSE) is the mean of the squared error

Root Mean Squared Error (RMSE) is the square root of the mean squared error

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

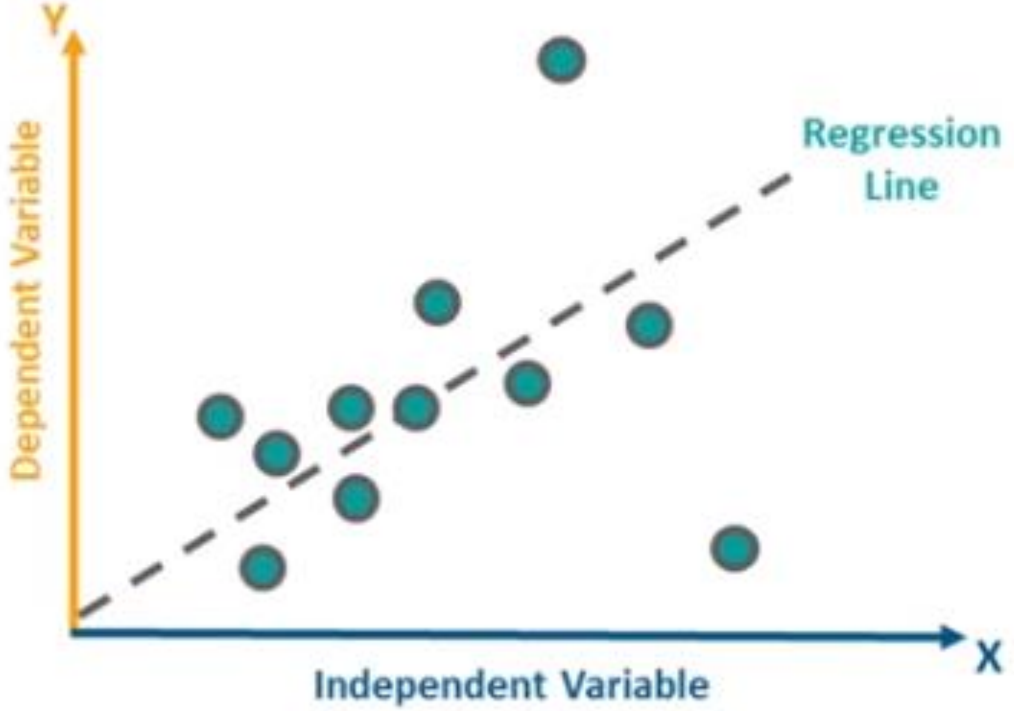
Where,

$\hat{y}$ $-$ *predicted value of y*

[illegible]

Properties of Regression Line

- Minimizes the sum of squared differences between observed values (**y**) and predicted values ($\mathbf{y_p}$)
- Passes through the mean of the X(**x**) and Y values (**y**)
- **Regression constant** (**c**) is equal to the y intercept of the regression line
- **Regression coefficient** (**m**) is the average change in the dependent variable (**y**) for a 1-unit change in the independent variable (**x**). It is the slope of the regression line

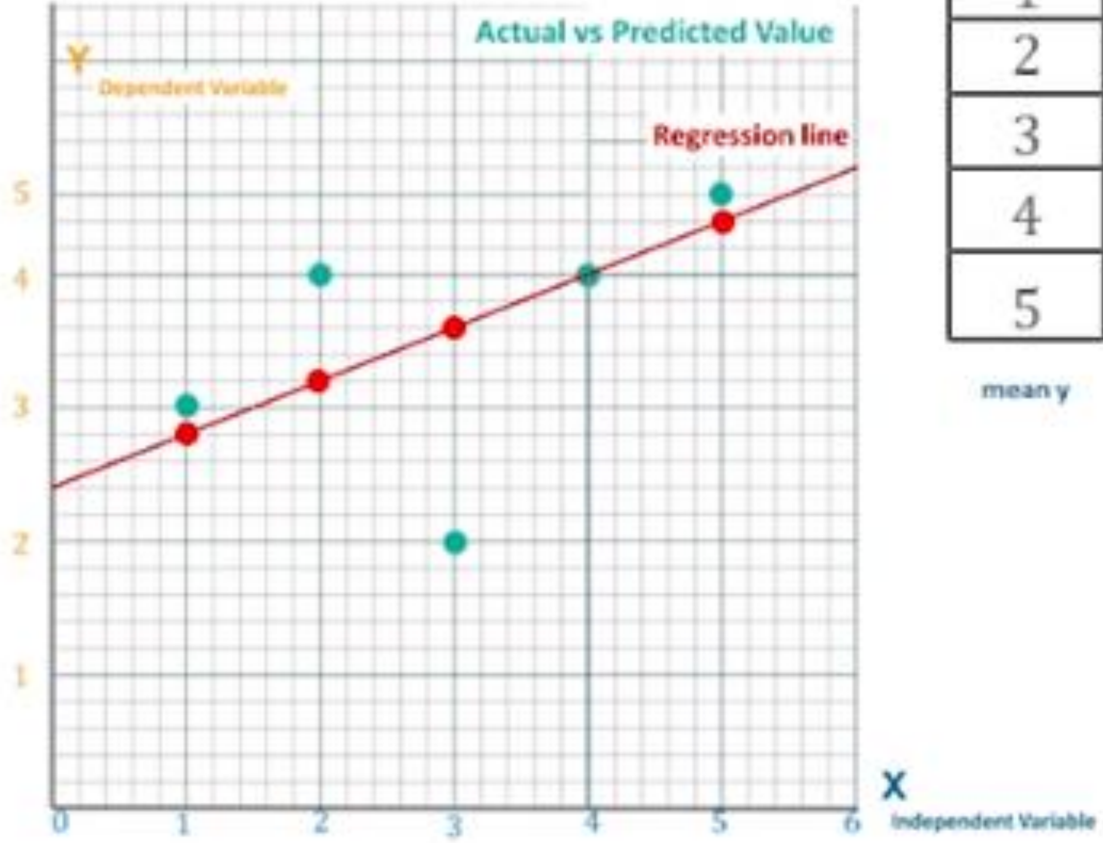


Checking The Goodness Of Fit: R-square

- **R-squared** value is a statistical measure of how close the data are to the fitted regression line
- Also known as **coefficient of determination** or the **coefficient of multiple determination**
  - **$R^2 = 0$** : Dependent variable cannot be predicted from the independent variable
  - **$R^2 = 1$:** Dependent variable can be predicted without error from the independent variable
  - **R^2** between 0 and 1 indicates the extent to which the dependent variable is predictable
  - **$R^2 = 0.20$** → 20 percent of the variance in *Y* is predictable from *X*

R^2

Calculation Of R^2(Cont.)

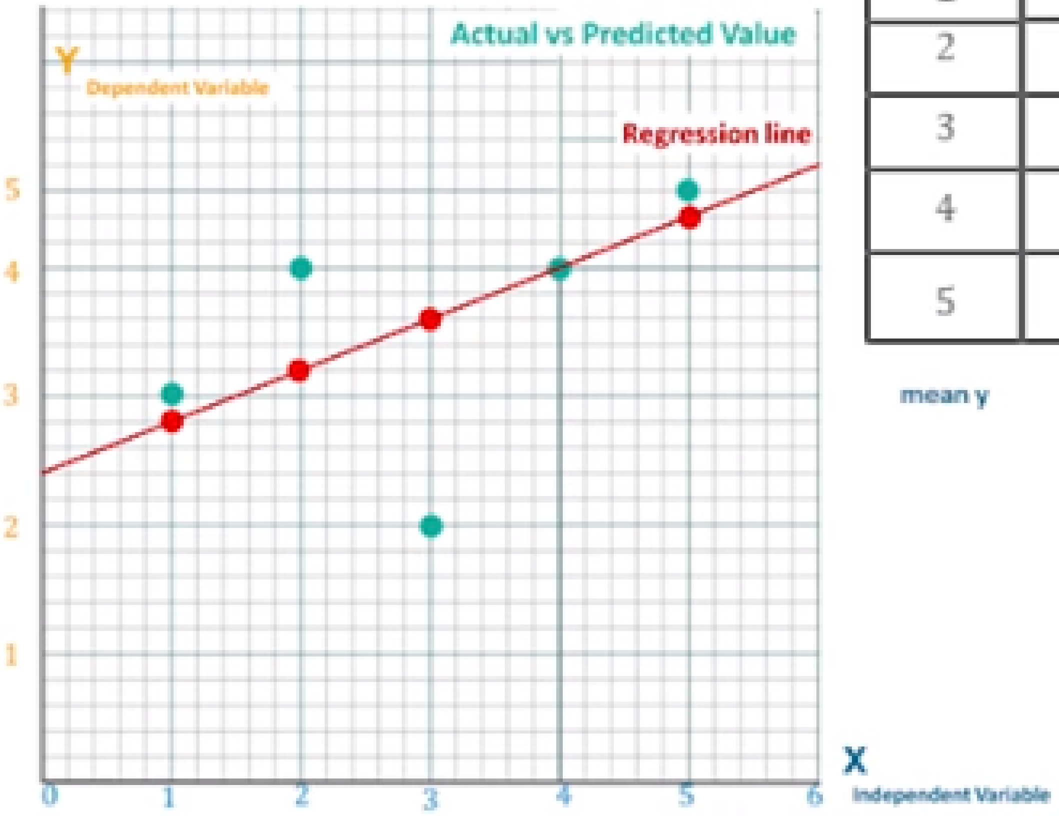


x	y	$y-\bar{y}$	$(y-\bar{y})^2$	y_p	$(y_p - \bar{y})$
1	3	– 0.6	0.36	2.8	-0.8
2	4	0.4	0.16	3.2	-0.4
3	2	−1.6	2.56	3.6	0
4	4	0.4	0.16	4.0	0.4
5	5	1.4	1.96	4.4	0.8

mean y 3.6

$$R^2 = \frac{\Sigma \ (y_p - \bar{y})^2}{\Sigma \ (y - \bar{y})^2}$$

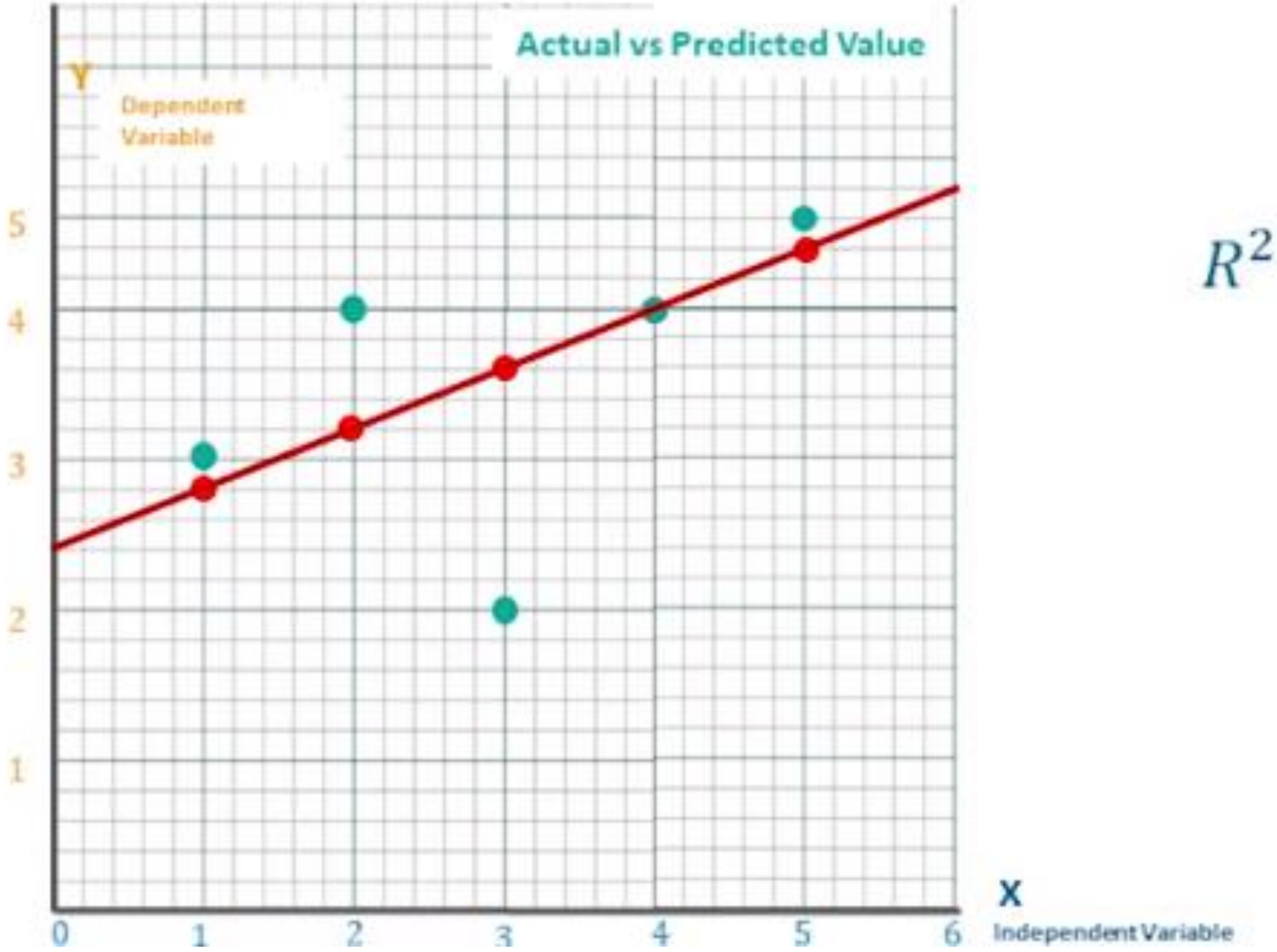
Calculation Of R^2 (Cont.)



x	y	$y-\bar{y}$	$(y-\bar{y})^2$	y_p	$(y_p-\bar{y})$	$(y_p-\bar{y})^2$
1	3	– 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	–1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		Σ 5.2			Σ 1.6

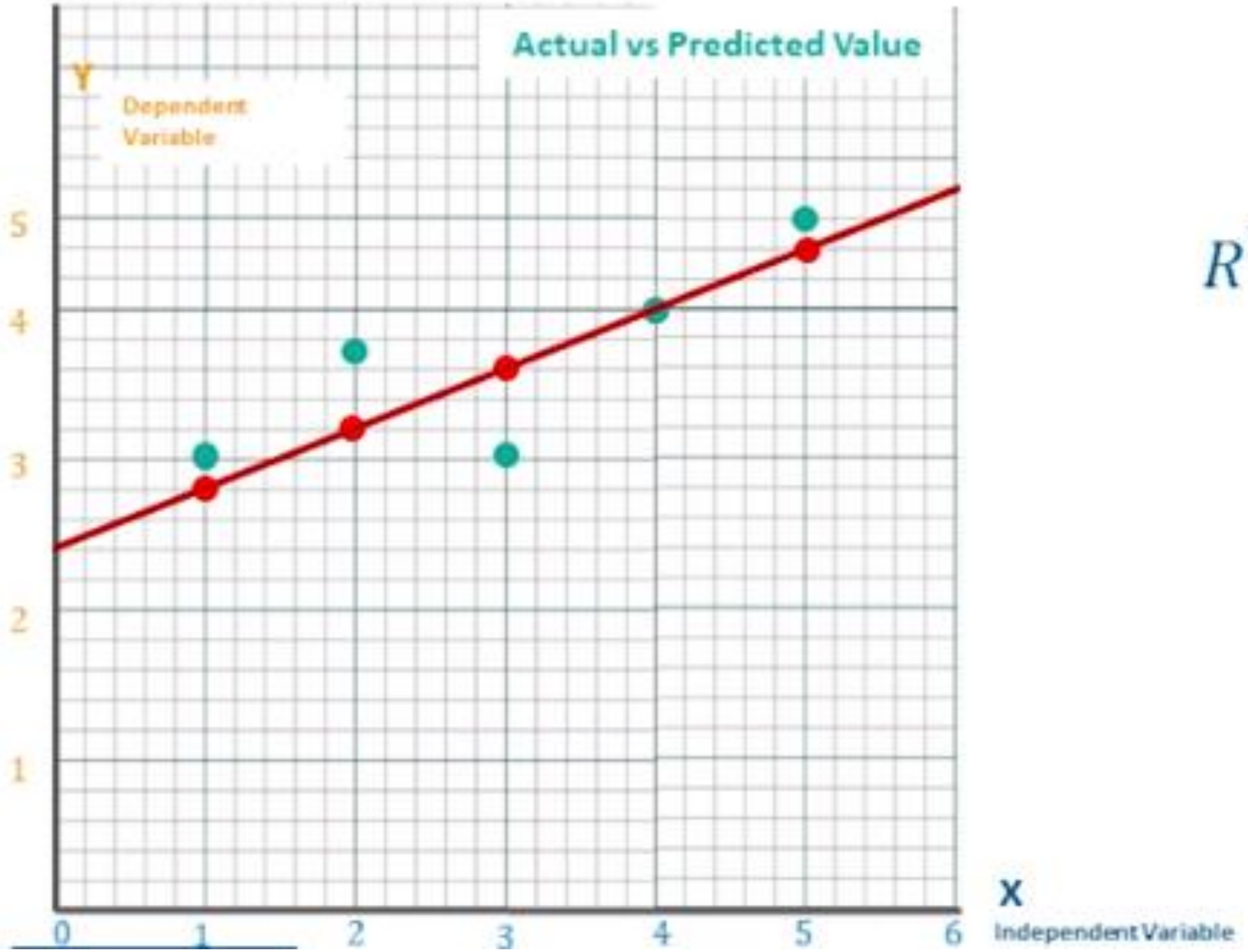
$$R^2 = \frac{1.6}{5.2} = \frac{\Sigma\ (y_p-\bar{y})^2}{\Sigma\ (y-\bar{y})^2}$$

Calculation Of R^2(Cont.)



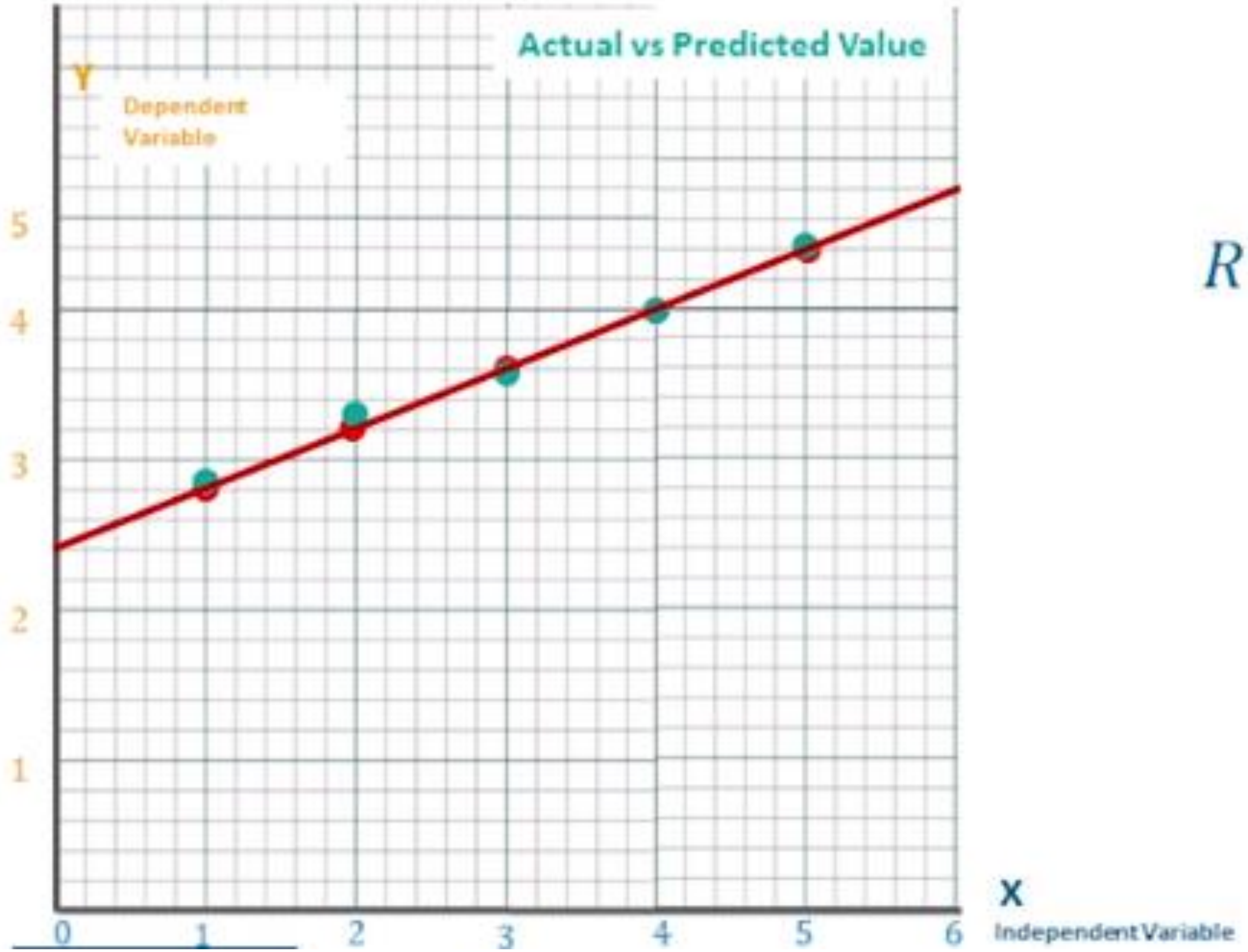
$$R^2 \approx 0.3$$

Calculation Of R^2 (Cont.)



$$R^2 \approx 0.7$$

Calculation Of R^2(Cont.)



$$R^2 \approx 0.9$$