# LINEAR REGRESSION

Machine learning

Mr. Devendra D. Patil
Assistant Professor
Department of Statistics
Vivekanand College, Kolhapur

# WHAT IS LINEAR REGRESSION?

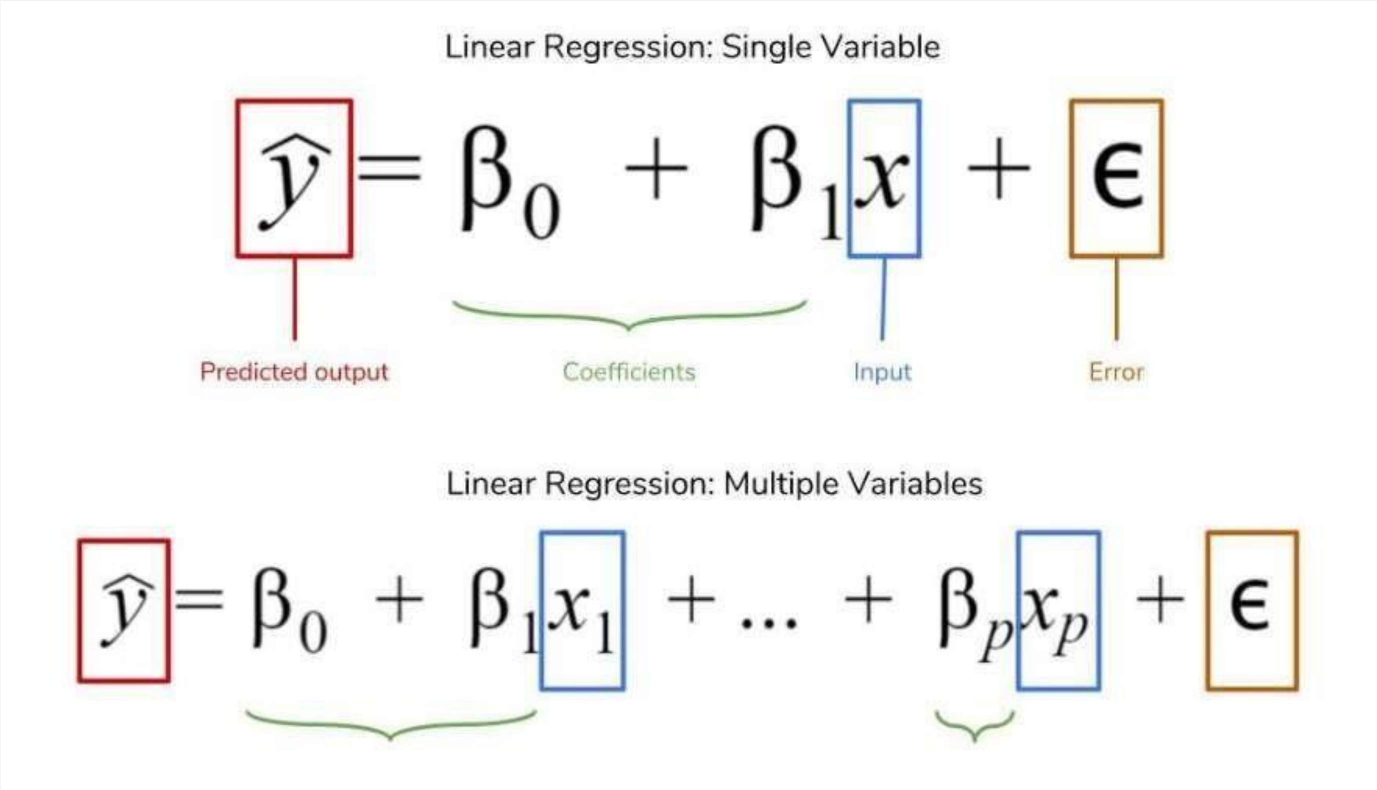
A linear regression is a data plot that graphs the linear relationship between an independent and a dependent variable(s).

It is typically used to visually show the strength of the relationship, and the dispersion of results.

E.g. to see test the strength of the relationship between amount of ice cream eaten and obesity.

Take the independent variable, the amount of ice cream, and relate it to the dependent variable, obesity, to see if there was a relationship.

# MATHEMATICAL FORM



# WHAT DO WE USE LINEAR REGRESSION FOR?

The overall idea of linear regression is to examine 2 things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

- Which variables in particular are significant predictors of the outcome variable and in what way do they –*indicated by the magnitude and sign of the beta estimates*– impact the outcome variable?

# KEY POINTS …

- When selecting the model for the analysis, an important consideration is model fitting.
- Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as $R^2$).
- overfitting can occur by adding too many variables to the model, which reduces model generalizability.
- A simple model is usually preferable to a more complex model.
- Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

# SIMPLE LINEAR REGRESSION

A college bookstore must order books 2 months before each semester starts. They believe that the number of books that will be sold for any particular course is related to the number of students registered for the course when the books are ordered.

They would like to develop a linear regression equation to help plan how many books to order.

From past records, the bookstore obtains the number of students registered, X, and the number of books actually sold for a course, y for 12 different semesters.

| Semester | No of students | Books |
|---|---|---|
| 1 | 36 | 31 |
| 2 | 28 | 29 |
| 3 | 35 | 34 |
| 4 | 39 | 35 |
| 5 | 30 | 29 |
| 6 | 30 | 30 |
| 7 | 31 | 30 |
| 8 | 38 | 38 |
| 9 | 36 | 34 |
| 10 | 38 | 33 |
| 11 | 29 | 29 |
| 12 | 26 | 26 |

# WHAT IS THE ERROR TERM?

- An error term is a variable in a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables.

- The error term is also known as the residual, disturbance, or remainder term.

# WHAT IS THE ERROR TERM?

- Within a linear regression model tracking a stock's price over time, the error term is the difference between the expected price at a particular time and the price that was actually observed.
- In instances where the price is exactly what was anticipated at a particular time, the price will fall on the trend line and the error term will be zero.
- *Points that do not fall directly on the trend line exhibit the fact that the dependent variable, in this case, the price, is influenced by more than just the independent variable, representing the passage of time.*
- *The error term stands for any influence being exerted on the price variable, such as changes in market sentiment.*

# ERROR CALCULATION

The residual is calculated as: $r_i = y_i - \hat{y}$

where

$r_i$ = residual value

$y_i$ = observed value for a given x value

$\hat{y}$ = predicted value for a given x value

- The magnitude of a typical residual can give us a sense of generally how close our estimates are.
- The smaller the residual standard deviation, the closer is the fit to the data.
- In effect, the smaller the residual standard deviation is compared to the sample standard deviation, the more predictive, or adequate, the model is.

linear equation is $\hat{y} = 1x + 2$, the residual for each observation can be found.

For the first set, the actual y value is 1, but the predicted y value given by the equation is $\hat{y} = 1(1) + 2 = 3$. The residual value is, therefore, 1 – 3 = -2, a negative residual value

| X | Y | $\hat{y}$ | error | $error^2$ |
|---|---|---|---|---|
| 1 | 1 | 3 | -2 | 4 |
| 2 | 4 | 4 | 0 | 0 |
| 3 | 6 | 5 | 1 | 1 |
| 4 | 7 | 6 | 1 | 1 |

Sum of squared residuals: 6

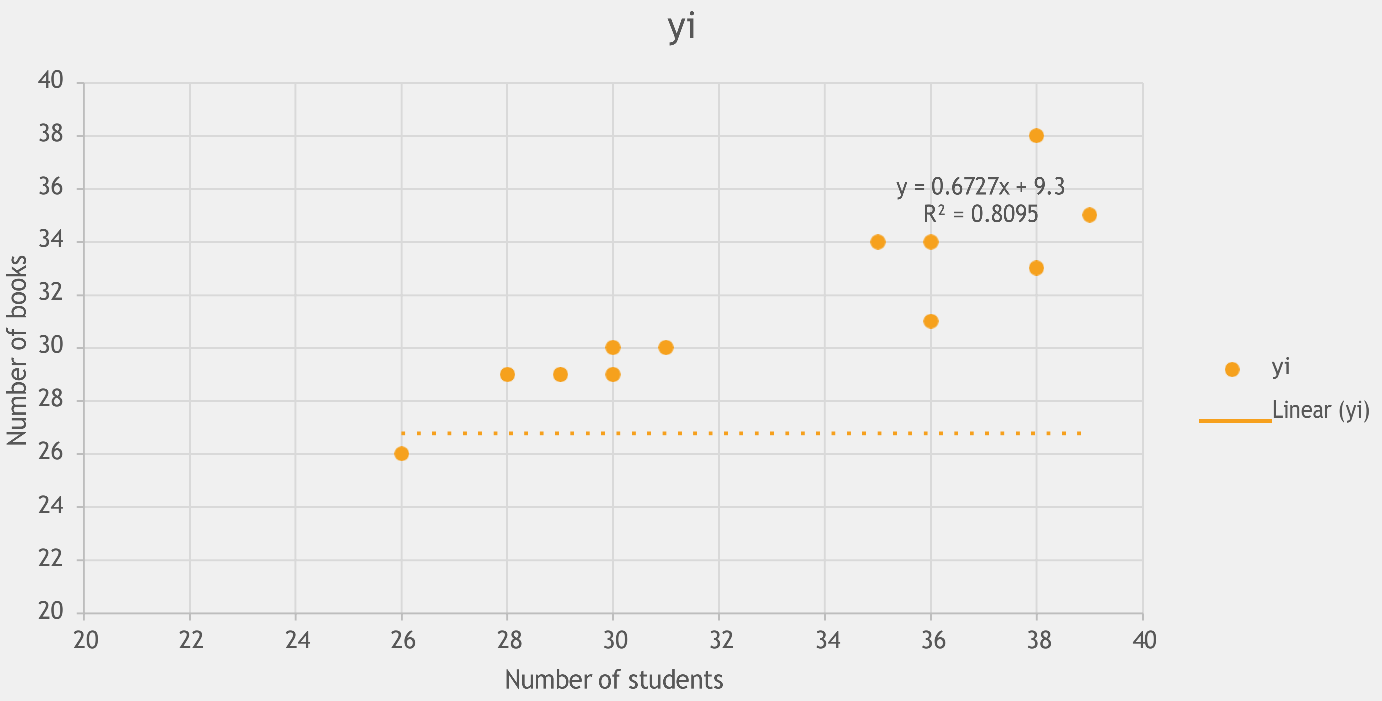
Number of residuals less 1: 4 - 1 = 3

Residual standard deviation: $\sqrt{(6/3)} = \sqrt{2} \approx 1.4142$

Slide no. 9

# REGRESSION LINE

Obtain a scatter plot of the number of books sold versus the number of registered students.

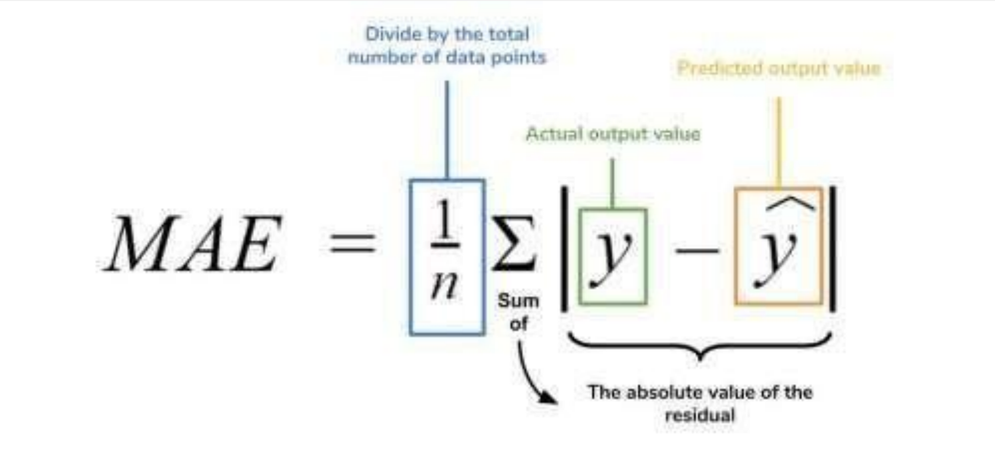


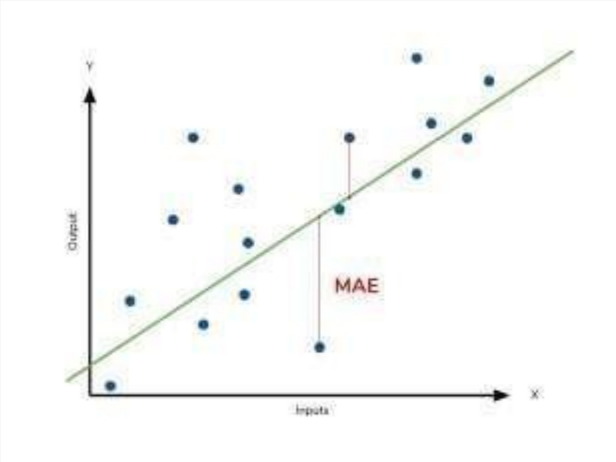
# MEAN ABSOLUTE ERROR

- Calculate the residual for every data point, taking only the absolute value of each so that negative and positive residuals do not cancel out.

- Take the average of all these residuals.

- Effectively, MAE describes the typical magnitude of the residuals.



# R-SQUARED

- Is a goodness-of-fit measure for linear regression models.
- Indicates the % of the variance in the dependent variable that the independent variables explain collectively.
- R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.
    - the $R^2$ is always going to be between -∞ and 1
- *Small R-squared values are not always a problem, and high R-squared values are not necessarily good!*

# ADJUSTED R2

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

$R^2$ = sample R-square

$p$ = Number of predictors

$N$ = Total sample size.

Thank You!