Linear Regression

Content

- Introduction to Regression
- Definition of Linear Regression
- Lines of Regression
- Estimation of Coefficient
- Interpretation

Regression: Introduction

Correlation

$r = 0.80$

y x

$y = f(x)$

Tells you how "values in Y" change as a function of changes in "value of X"

1. Use one independent variable at a time to predict *Y* (Simple Linear Regression)

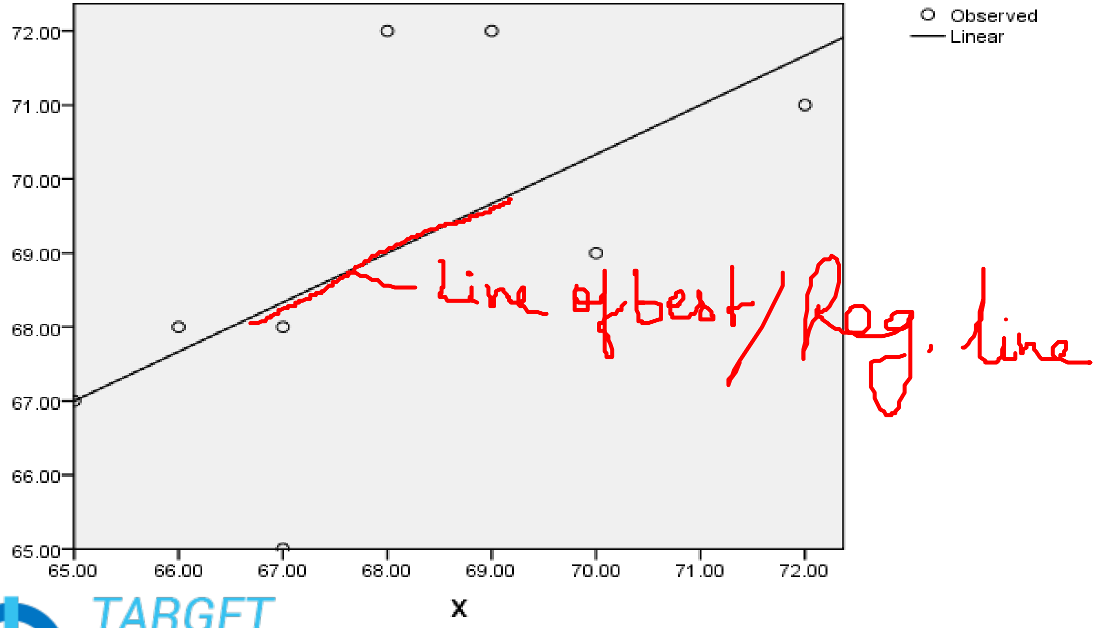
2. Assess the combined effects of the independent variables on *Y* (Multiple/Multi-variable Linear Regression)

y = const.

TARGET ORTHO

Definition

- Regression Analysis is a **mathematical measure of the average relationship** between two or more variables in terms of the original units of the data.



$$Y = a + bX$$

$$\text{where } a = \text{the } Y \text{ intercept}$$
$$b = \text{the slope}$$

TARGET ORTHO

Figure 01: Linear Regression Line

Line of Regression

- Line of "Best Fit"

- Obtained by Principal of "Least Squares"

Principal of Least Squares

- Line of best fit is obtained by Minimizing the sum of squares of residuals.
- Residual = Observed value – Predicted Value

$= Y - \hat{Y}$

Y X

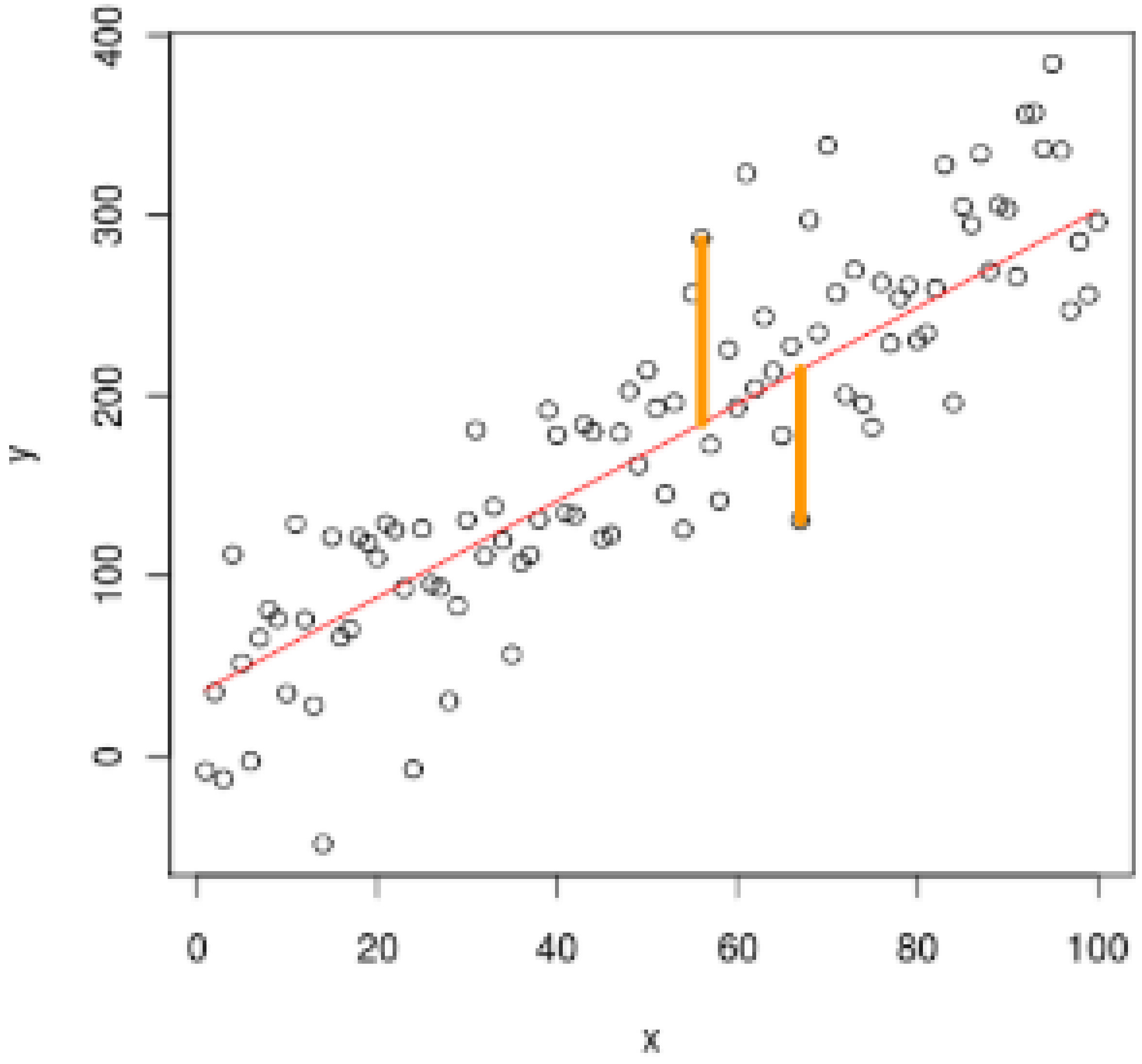


Figure 02: Demonstration of Residuals

Two Lines of Regression

- Y on X
- X on Y

$c = \alpha$ $\quad y = \alpha + \beta x$

$m = \beta$ $\quad y = mx + c$ — Line

$SBP = 110 + 0.5\ Age$ —— (1)

$Age = 40$

	y		$\hat{y}$	
	SBP	Age	SBP	Residue
2 001	125	40	130	−5
	122	30	125	−3

$SBP = 110 + 0.5 \times 40$

$= 130$

$Age = 30$

$SBP = 110 + 0.5 \times 30$

$= 125$

Two Lines of Regression

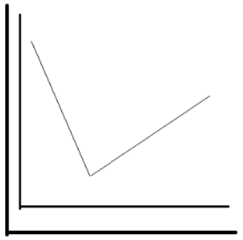
- Two lines of regression are **not reversible** or interchangeable

- **Basis & assumptions** for deriving these equations are quite **different**

- Regression line of **Y on X** is obtained by **minimizing the sum of squares of residuals parallel to Y- axis**.

Properties

- r = ±1, both the lines of regression coincide
- Correlation coefficient is the geometric mean of two regression coefficients.
- If one regression coefficient is greater than 1, than another should be less than 1.
- Regression coefficient are independent of change of origin but not of scale.



Two Components of Linear Regression: Coefficient

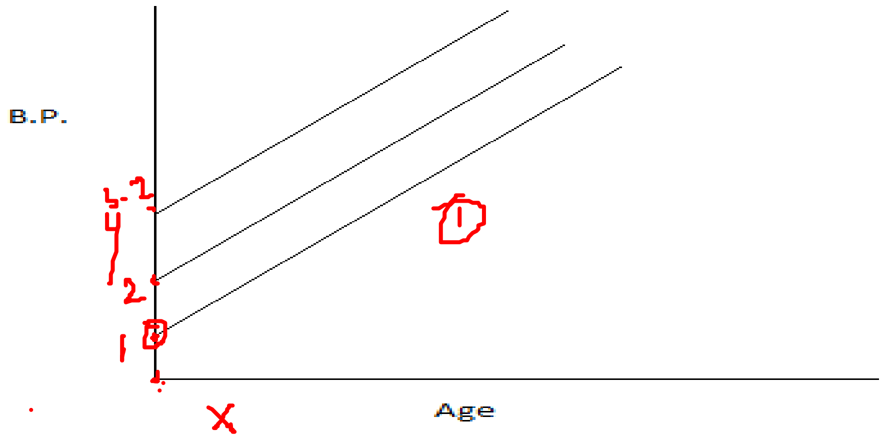
- Also known as “Slope”
- Theoretically, $- \infty \leq b \leq + \infty$

Two Component of Linear Regression: Y intercept

$y = mx + c$ or $y = \alpha + \beta x$

$y = \alpha$

- Minimum Value of Y
- Value of Y at X=0



Interpret Slope & Intercept

Y = 90 - 2X — ①

$Y = \alpha + \beta X$ — ②

Where Y = Final exam score

$\alpha = 90 \quad \beta = -2$

X = Number of absences from class during the semester

$Y = \alpha + \beta X$

$Y = 90 - 2X$

Coefficient of Determination (R^2)

- It tells us amount of explained variation over total variation
- R^2 =MSS/TSS = 1-RSS/TSS
- Where, MSS=Model Sum of Squares (or Explained Sum of Squares)
    - TSS= Total Sum of Squares
    - RSS= Residuals Sum of Squares

$R^2 = 0.80$

$0 \leq R^2 \leq 1$

Adjusted R^2

- Adjusted R-squared is a statistical measure that is used to assess the goodness-of-fit of a regression model, while adjusting for the number of predictors in the model.

- Unlike the regular R-squared, which can misleadingly increase with the addition of irrelevant predictors, adjusted R-squared provides a more accurate reflection of the model's explanatory power when multiple predictors are involved.

- It's particularly useful for comparing models with different numbers of independent variables.

Formula

- Adjusted $R^2 = 1 - [(1-R^2)(n-1)/(n-k-1)]$

- Where:
- - R^2 is the regular R-squared
- - n is the number of observations
- - k is the number of predictors

Key Aspects

- Penalizes for excessive predictors
- Useful for comparing models with different numbers of independent variables
- Does not account for every aspect of model quality

Conclusion

- Adjusted R-squared is essential for evaluating the performance of regression models, particularly when comparing models with a different number of explanatory variables. It helps to ensure that the model is both succinct and informative.

References

- John O. Rawlings, Sastry G. Pantula, David A. Dickey. Applied Regression Analysis: A Research Tool. Second Edition. Springer

- Daryl S. Paulson. Handbook of Regression and Modeling Applications for the Clinical and Pharmaceutical Industries. Chapman & Hall/ CRC Biostatistics Series

- a visual guide to CRISP-DM methodology (http://www.crisp-dm.org/download.htm)

- https;//www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis assumptions-plots-solutions/

- http://www.statisticshowto.com/durbin-watson-test-coefficient/

- David M. Levine, David F. Stephan, Kathryn A. Szabat. STATISTICS FOR MANAGERS USING MICROSOFT EXCEL, 8th Edition. Pearson Publication.

Lets Connect!

draanchalawasthi@gmail.com

https://www.youtube.com/c/sscrindia

+91 750.625.0403