Topic for the class-Regression

Unit _3 : Title-Descriptive statistics

Date & Time : 5.9.24 10.00 AM – 10.50 AM

Dr. Bhramaramba Ravi

Professor

Department of Computer Science and Engineering

GITAM School of Technology (GST)

Visakhapatnam – 530045

Email: **bravi@gitam.edu**

Unit3-syllabus

- **UNIT 3 Descriptive statistics 9 hours, P - 2 hours**
- Measures of Central Tendency – Measures of Variation – Quartiles and Percentiles –

  Moments – Skewness and Kurtosis. Exploratory Data Analytics Descriptive Statistics – Mean,

  Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA, Random variable, Variance, covariance, and correlation- Linear transformations of random variables, Regression.
- https://www.coursera.org/learn/data-visualization-r

Regression

- Models can be built to predict continuous data values (*regression models*) or categorical data (*classification models*).
- Simple methods to generate these models include *linear regression*, *logistic regression*, *classification and regression trees*.

Linear regression

- how to generate *linear* models to describe a relationship between one or more independent variables and a single response variable.
- For example, we could build a linear regression model to predict cholesterol levels using data about a patient's age.
- This model will likely be a poor predictor of cholesterol levels; however, incorporating more information, such as body mass index (BMI) may result in a model that provides a better prediction of cholesterol levels.
- Using a single independent variable is referred to as *simple linear regression*, whereas using
- more than one independent variable is referred to as *multiple linear regression*.
- Although these models do not make causal inferences, they are useful for understanding how a set of independent variables is associated with a response variable

Fitting a simple linear regression model

A simple linear regression model can be generated where there is a linear relationship between two variables. For example, Figure 6.5 shows the relationship between the independent variable *Age* and the response variable *Blood fat content*. The diagram shows a high degree of correlation between the two variables. As variable *Age* increases, response variable *Blood fat content* increases proportionally. A straight line, representing a linear model, can be drawn through the center of the points.

This straight line can be described using the formula

$$y = b_0 + b_1 x$$

where b_0 is the point of intersection with the y-axis and b_1 is the slope of the line, which is shown graphically in Figure 6.6. The simple linear regression model is usually shown with an error term; however, it is not included here to simplify the example.

Fitting a simple linear regression model

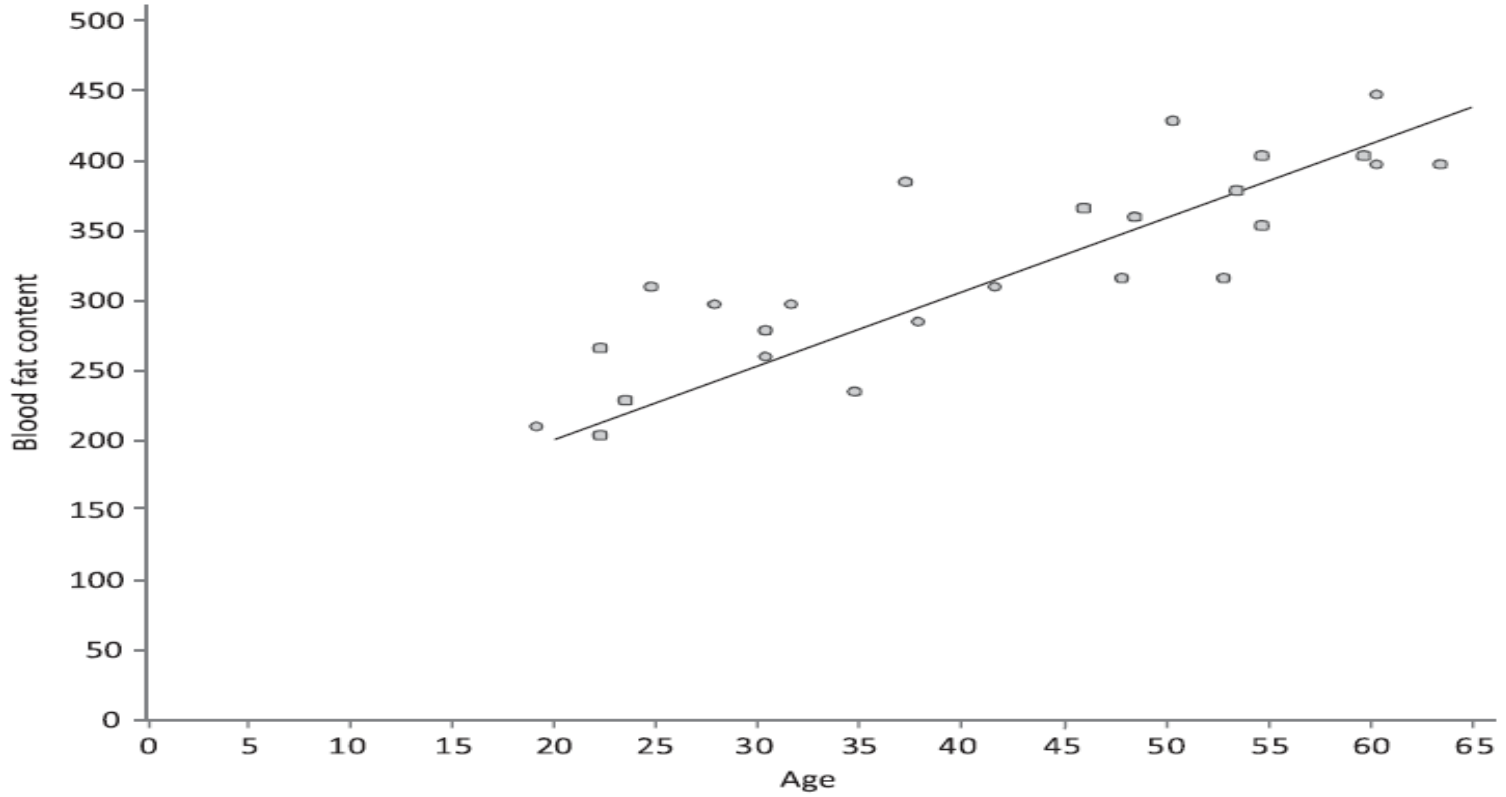

FIGURE 6.5 A straight line drawn through the relationship between variables *Age* and *Blood fat content*.

Fitting a multiple linear regression model

In most practical situations, a simple linear regression is not sufficient because the models will need more than one independent variable. The general form for a multiple linear regression equation is a linear function of the independent variables:

$$y = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{pi} + e_i$$

where the response variable (y) is shown with p independent variables (x-variables), b_0 is a constant value, k is the number of coefficients of the independent variables, and e_i refers to an error term measuring the unexplained variation or noise in the linear relationship.

Logistic regression

- The multiple linear regression approach can only be used to make predictions when the response variable is continuous.
- It cannot be used when the response variable is categorical
- Logistic regression is a popular approach to building models where the response variable is usually binary (dichotomous).
- For example, the response variable could indicate whether a consumer purchases a product (1 if they purchase and 0 if they do not) or whether a candidate drug is potent (1 if the candidate drug is potent and 0 if it is not).
- Logistic regression provides a flexible and easy-to-interpret method for building models from binary data.

THANK YOU