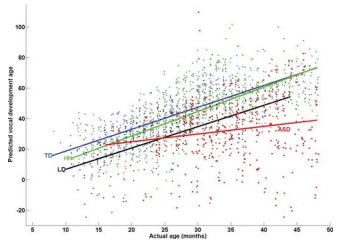
Multiple Linear Regression

Tushar B. Kute,
http://tusharkute.com



Multiple Linear Regression

- Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:
  - How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
  - The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Multiple Linear Regression – Ex.

- You are a public health researcher interested in social factors that influence heart disease.
- You survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.
- Because you have two independent variables and one dependent variable, and all your variables are quantitative, you can use multiple linear regression to analyze the relationship between them.

Assumptions

- Multiple linear regression makes all of the same assumptions as simple linear regression:
- Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
- Independence of observations: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.

Assumptions

- In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model.
- If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.
- Normality: The data follows a normal distribution.
- Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

How to perform?

- The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

  - y = the predicted value of the dependent variable
  - B_0 = the y-intercept (value of y when all other parameters are set to 0)
  - B_1X_1= the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
  - … = do the same for however many independent variables you are testing
  - B_nX_n = the regression coefficient of the last independent variable
  - e = model error (a.k.a. how much variation there is in our estimate of y)

How to perform?

- To find the best-fit line for each independent variable, multiple linear regression calculates three things:
  - The regression coefficients that lead to the smallest overall model error.
  - The t-statistic of the overall model.
  - The associated p-value (how likely it is that the t-statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).
- It then calculates the t-statistic and p-value for each regression coefficient in the model.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com