LINEAR REGRESSION

30 october,2022 by: Piyush rawat

Regression analysis:

The technique for using data to identify relationships among variables and use these relationships to make further predictions. Linear dependence means constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns). Examples:

- We have to predict selling price of a house with given details: For each house, we have complete information about its size, the number of bedrooms, bathrooms, total rooms, the corresponding property tax, etc., and also the price at which the house was eventually sold. A linear model with such information would be:
  Selling price =
  $\beta 0 + \beta 1$ (sq. ft.) + $\beta 2$ (no. bedrooms) +$\beta 3$ (no. bath) + $\beta 4$ (no. acres) +$\beta 5$ (taxes) + error
  where:
  where, $\beta 1$ represents the increase in selling price for each additional square foot of area: it is the marginal cost of additional area. Similarly, $\beta 2$ and $\beta 3$ are the marginal costs of additional bedrooms and bathrooms, and so on. The error reflects the fact that two houses with exactly the same characteristics need not sell for exactly the same price.
- Most economic forecasts are based on regression models. Consider the problem of predicting growth of the economy in the next quarter. A linear model for predicting growth would be:
  Next qtr. growth = $\beta 0 + \beta 1$ (last qtr. growth) + $\beta 2$ (this qtr. growth) +$\beta 3$ (index value) + $\beta 4$ (factory orders) +$\beta 5$ (inventory levels) + error
  Here all the coefficients would be derived from past data.
- A regression model specifies a relation between a dependent variable Y and certain explanatory variables X1..., XK. A linear model sets $Y=\beta 0 + \beta 1X1 + \cdot\cdot + \beta K\ XK + e$. Where e is the error term.

Simple Linear Regression:

1) It's a model with just one explanatory variable. I.e.: $Y = \beta 0 + \beta 1X + e$.
   where X is the variability in variable Y. Example: A corporation is concerned about maintaining parity in salary levels of purchasing managers across different divisions. As a rough guide, it determines that purchasing managers responsible for similar budgets in different divisions should have similar compensation. The scatter plot for this data would include straight line fit to the data. The slope of this line determines the marginal increase in salary with increase in budget responsibility for employees.

2) Regression uses the least squares criterion also, ie: to minimize the error. There might be error in the predicted and the actual line so, the least squares criterion chooses β0 and β1 to minimize the sum of squared errors.
3) Slope will be given by: $\hat{\beta}_1 = Cov[X, Y] / StdDev[X]$.
4) In short regression picks the line which minimizes the sum of squared errors. This choice is reported through the estimated values of β0 and β1.

Understanding broader aspects:

- ANOVA is stands for analysis of variance. DF stands for degrees of freedom, SS for sum of squares, and MS for mean square. The mean squares are just the sum of squares divided by the degrees of freedom: MS = SS/DF.
- The Total SS means the total variability in the salary levels. The Regression SS is the explained variation. The Error SS is the unexplained variation. This reflects differences in salary levels that cannot be attributed to differences in budget responsibilities. The explained and unexplained variation sum to the Total SS.
- The Quantity coefficient of determination or R-square defines how much of variability has been explained. It's given by:
  R^2 = Explained variability/Total variability
  = SSR/SST
  Thus, high R-square value defines strong linear relation between two variables.
- T stat or test statistics is defined as the ratio of estimated slope to std err, which if have high ratio value gives small p-value which is considered to be good. [$t = \hat{\beta}_1 - \beta_1 / s\hat{\beta}_1$].

Multiple Regression:

- So, in general we always have more than one explanatory variables for getting a meaningful model of regression, where comes the part of multiple regression. The general multiple linear regression model with K explanatory variables is
  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + e$,
  Our data consists of observations (Y1, X11..., XK1) (Y2, X12..., XK2) ⋯ (Yn, X1n,...,XKn).
- Sometimes data might be redundant like having different units of same data which can be termed as multicollinearity, which can lead to poor results so this must look over by the user.
- As we know adding more variables can cause more variation and hence the slope so it must be corrected with formula: adjusted $R2 = 1 - [SSE/(n - K - 1)]/[SST/(n - 1)]$.