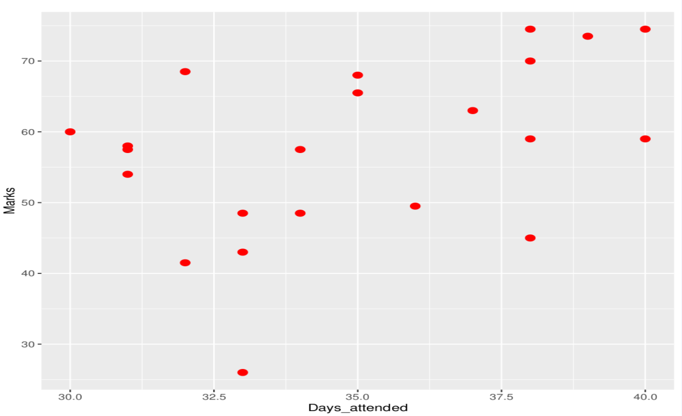
Statistical Inference (MA862)

Lecture Slides

Topic 5: Linear Regression

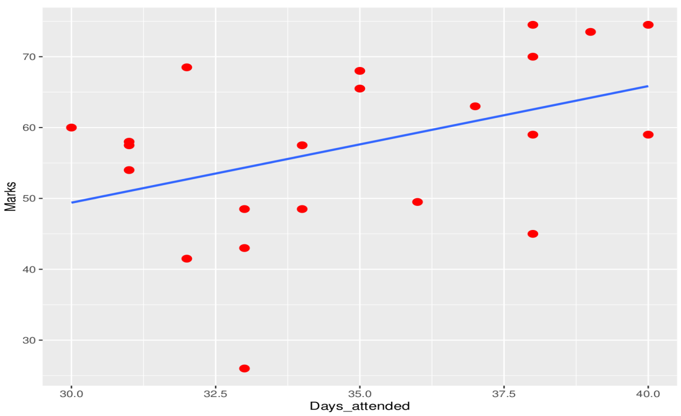
Regression

- Question: What is the impact of attending classes on students' final marks?
- Let's start with a real data from IITG which you can feel about it!!



Regression

- Question: What is the impact of attending classes on students' final marks?
- Let's start with a real data from IITG which you can feel about it!!



Linear Regressions

- We have one particular variable that we are interested in understanding or modeling, such as sales of a particular product, sale price of a home, or voting preference of a particular voter. This variable is called the target, response, or dependent variable, and is usually represented by y.
- We have a set of p other variables that we think might be useful in predicting or modeling the target variable (for *e.g.* the price of the product, the competitor's price, and so on; or the lot size, number of bedrooms, number of bathrooms of the home, and so on; or the gender, age, income, party membership of the voter, and so on). These are called the predicting, independent variables, or features and are usually represented by $x_1, x_2, \ldots, x_p$.

Linear Regressions

- Thus, we have

$$y = f(\boldsymbol{x}; \boldsymbol{\beta}) + \varepsilon,$$

  for some real valued function f, where $\boldsymbol{x}$ is vector of predictors, $\boldsymbol{\beta}$ is the vector of parameters, and ε is error.
- If f is linear in the parameters vector $\boldsymbol{\beta}$, then the regression is called linear regression.

Examples and Role of Transformations

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ is a linear model.
- $y = \beta_0 + \beta_1 x + \beta_2 x^2$ is a linear model, because it is linear in β (even though not in x).
- $y = \beta_0 + \beta_1 x^{\beta_2}$ is a non-linear model, as it is not linear in β.
- $y = \beta_0 x^{\beta_1}$ is not a linear model, but $\ln y = \ln \beta_0 + \beta_1 \ln x$ is.
- $y = \frac{e^{\beta x}}{1+e^{\beta x}}$, where $y \in (0,1)$
- $y = \frac{1}{\beta_0+\beta_1 x_1+\beta_2 x_2}$.

Main use of Linear Regressions

Typically, a regression analysis is used for one (or more) of three purposes:

1. modeling the relationship between x and y;
2. prediction of the target variable (forecasting);
3. testing of hypotheses.

Simple Linear Regression

- Just one predictor x, i.e. $p = 1$.
- The model for the simple linear regression is given by

$$y = \beta_0 + \beta_1 x + \epsilon,$$

  where y is the outcome variable (random), x is the independent/predictor variable (non-random) and ϵ is the random error term. β_0 (intercept) and β_1 (slope) are model parameters (unknown constants).
- Equivalently, the model can be written for $i = 1, 2, \cdots, n$ number of observations $(x_1, y_1), \cdots, (x_n, y_n)$ as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n.$$

- How do you interpret β_0 (intercept) and β_1 (slope)?

Least Squares Estimation

- Goal: To estimate β_0, β_1 by minimizing error in some sense (*e.g.* squared error)
- One reasonable way is to use the principle of Least Squares, *i.e.* minimize the objective function

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  with respect to β_0, β_1.
- Differentiate $Q(\beta_0, \beta_1)$ with respect to β_0, β_1 and equate the partial derivatives to zero to get the estimates $\hat{\beta}_0, \hat{\beta}_1$.
- The resulting equations are called normal equations:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \text{ and } \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Least Squares Estimation

- The solution is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = S_{xy}/S_{xx}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \ \text{ and } \ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimator (LSE) of β_0 and β_1, respectively.