Statistical Inference and Multivariate Analysis (MA324)

Lecture Slides
Lecture 25

Linear Regression

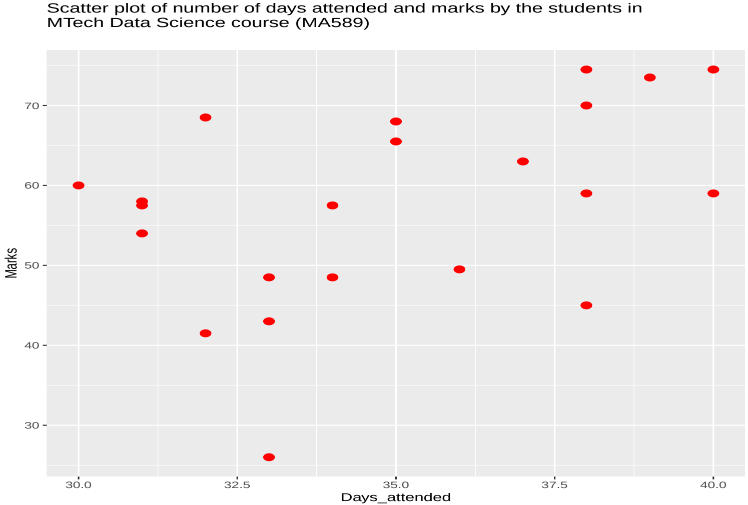
Indian Institute of Technology Guwahati

Jan-May 2023

Research Question: What is the impact of attending classes on students' final marks

Research 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!!



- We will do a regression analysis to the above data to find the answer. Is using regression appropriate here?

- **Strong Advice:** As a Data Scientist/Statistician, you have to **spent a lot of your time** and effort to **pre-process/clean the raw data** to make it **analysis ready**. It is part of your job to clean the data: so learn it quickly!!

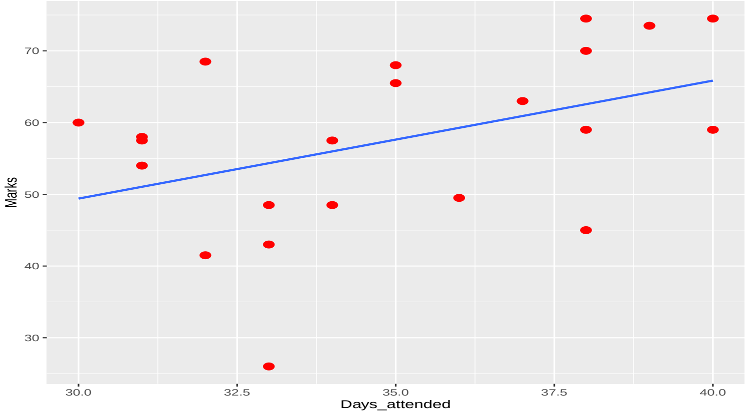
- Here is the output after fitting a regression line in R:

Fitting Regression line

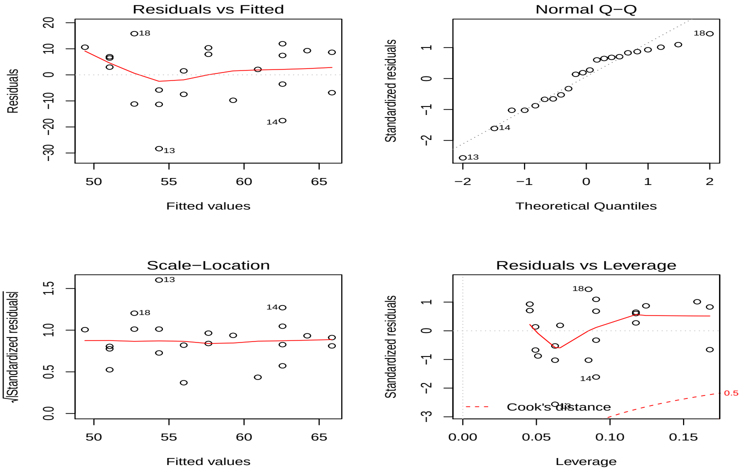
Scatter plot of number of days attended and marks by the

students in MTech Data Science course (MA589):

with fitted regression line



Residual Analysis



Linear Models: Simple and Multiple Linear Regressions

The regression framework can be characterized in the following way[1]:

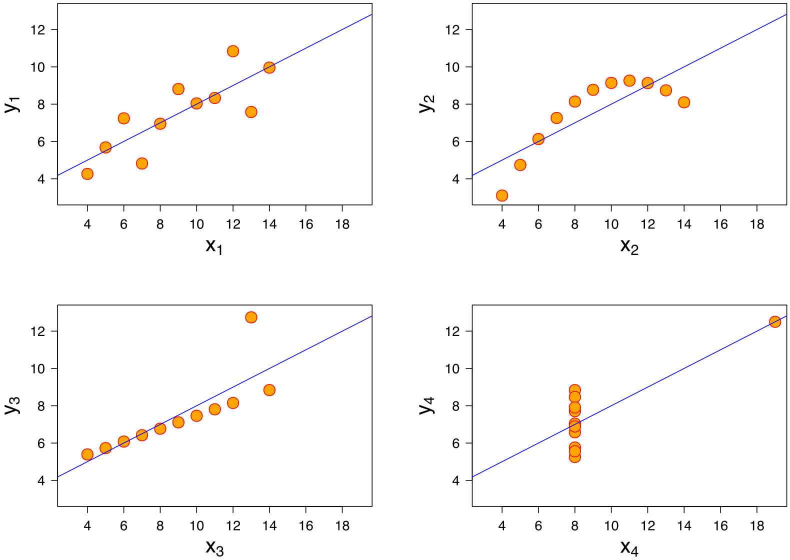
- We have one particular variable that we are **interested in understanding or modeling**, such as sales of a particular product, sale price of a home, or voting preference of a particular voter. This variable is called the **target, response, or dependent variable, and is usually represented** by y.
- We have a set of p **other variables** that we think might be **useful in predicting or modeling the target variable** (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, or independent variables, and are usually represented** by $x_1, x_2, \cdots, x_p$.

[1] Handbook of Regression Analysis. By Samprit Chatterjee and Jeffrey S. Simonoff.

- 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. and **testing of hypotheses**.

Importance of graphing data before analyzing it



- Which one of the above do you think has **highest correlation and ideal for linear regression?**

[2] The graph is taken from the Wikipedia

- Which one of the above do you think has **highest correlation and ideal for linear regression?**

- In **all the four graphs**: mean of x = 9 (with variance 11); mean of y = 7.50 (with variance 4.1) ; correlation between x and y = 0.816

- Fitted linear regression in each cases: $y = 3 + 0.5x$

- In 1973, **Anscombe** demonstrated the **importance** of **graphing data**[2] before analyzing it and the **effect of outliers** on statistical properties

[2] The graph is taken from the Wikipedia

Linear Models: Meaning of Linearity and Transformation

What does “Linear” mean?

- A linear model is **linear in the parameters β, but not necessarily in the x's,** e.g.

  * $y = \beta_0 + \beta_1 x + \beta_2 x^2$ 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,** because it is not linear in β.

Transformation

- Clearly $y = f(x) = \beta_0 x^{\beta_1}$ is **not a linear model,** but

$$\ln f(x) = \ln \beta_0 + \beta_1 \ln x$$

If we let $f^* = \ln f, \beta_0^* = \ln \beta_0, \beta_1^* = \beta_1, x^* = \ln x$, we have

$$f^* = \beta_0^* + \beta_1^* x^*,$$

which is **a linear model.**

Transformation

- Thus, although linear models seem to be simple and restrictive, they **can actually be quite flexible by transformation of the response and the predictors.**

- Linear models are **not just straight lines, they can be curved.** Can you give an examples?

Transform the following **non-linear models to linear models:**

- $y = \frac{e^{\beta x}}{1+e^{\beta x}}$, where $y \in (0,1)$.

- $y = \frac{1}{\beta_0+\beta_1 x_1+\beta_2 x_2}$.
- $y = e^{e^{\beta x}} - 1$, where $y > 0$.