Prediction Using Regression Methods

Contents

Basics

- Regression is concerned with determining a relation between a single numeric dependent variable and one or more numeric independent variables; given, a set of observed values of the set of independent variables and the corresponding values of the output variable.
- As the name implies, the dependent variable **depends on** the value of the independent variable *or* variables.
- The single numeric dependent variable is a real continuous variable.

Contd…

- **Examples:**

  1. Let us say, we want to have a system that **can predict the price of a used car (y)**.
     - **Inputs** are, the car attributes like
       - model **(x_1)**
       - type **(x_2)**
       - year **(x_3)**
       - engine capacity **(x_4)**
       - mileage **(x_5)** , and
       - other information that we believe affect a car's worth **(x_i)** .
     - The **output** is the price of the car.

Contd…

2. Consider the **navigation of a mobile robot**, say an autonomous car.
   - The **output** is the angle by which the steering wheel should be turned at each time, to advance without hitting obstacles and deviating from the route.
   - **Inputs** are provided by sensors on the car like a video camera, GPS and so forth.

Contd…

3. The **selling price of a house (y)** can depend on
   - the desirability of the location **(x_1)**
   - the number of bedrooms **(x_2)**
   - the number of bathrooms **(x_3)**
   - the year the house was built **(x_4)**
   - the square footage of the plot **(x_5)**, and
   - a number of other factors **(x_i)**.

Contd…

4. The **height of a child (y)** can depend on
   - the height of the mother **(x_1)**
   - the height of the father **(x_2)**
   - nutrition **(x_3)**, and
   - various environmental factors **(x_i)**.

Contd…

- There are **several forms of regression** such as
  - **Simple Linear Regression**
  - **Multiple Regression**
  - **Polynomial Regression**
  - **Logistic Regression**

Contd…

- We focus on the most basic linear regression methods – **those that use straight lines**.
  - When there is only one independent variable, it is known as **simple linear regression**.
  - In the case of two or more independent variables, this is known as **multiple linear regression**, or simply **multiple regression**.

Simple Linear Regression

- This simplest form of regression follows a straight line.
- Straight lines can be defined using the slope – intercept form **y = mx + c** (OR, **y = a + bx**, OR **y = α + βx**).
  - y: dependent variable
  - x: independent variable
  - m (or, b) : slope *(gradient)*
  - c(or, a) : intercept

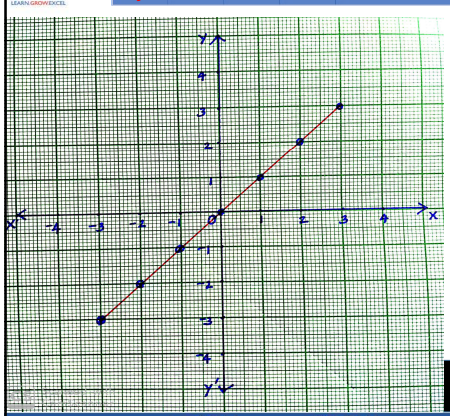
Contd…

- **m** – value specifies **how much the line rises** for each increase in x.
- **c** – value specifies the point **where the line crosses**, or intercepts, **the y – axis**.

Contd…

- **Problem:**
- Construct the straight lines corresponding to
  a) y = x
  b) y = -1.x
  c) y = 0.5x + 2

x	-3	-2	-1	0	1	2	3
y	-3	-2	-1	0	1	2	3

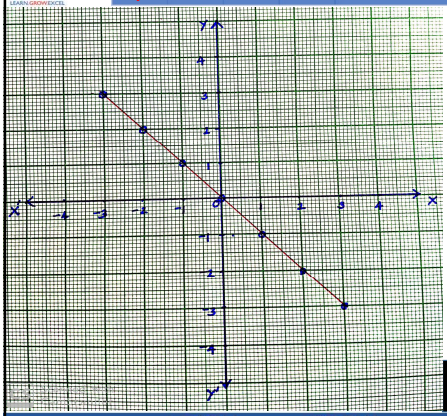


m=1 c=0

$y = 1.x + 0$

LEARN . GROW . EXCEL GOPEEKRISHNAN R

x	-3	-2	-1	0	1	2	3
y	3	2	1	0	-1	-2	-3

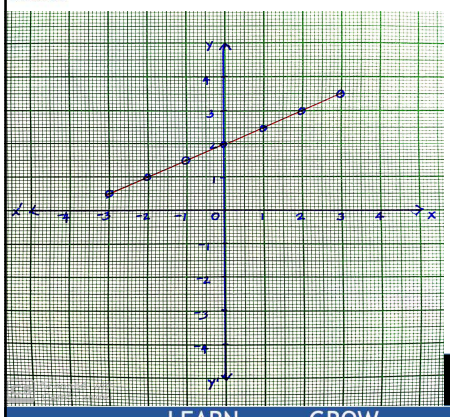


m= -1 c=0

$y = -1.x + 0$

LEARN . GROW . EXCEL GOPEEKRISHNAN R

x	-3	-2	-1	0	1	2	3
y	0.5	1	1.5	2	2.5	3.0	3.5



m=0.5 c=2

$y = 0.5.x + 2$

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- Regression equations **model** given data using a similar slope – intercept format.
- The ML algorithm here…
  - **identifies best possible optimal** values of ***c*** and ***m*** so that, the specified line is best able to relate the supplied *x* values to the value of *y*.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

What do we do here?

1. We, select the data set (past data==training data).
2. We, Select the feature (X) to be plotted along X – axis and the feature (Y) to be plotted along Y – axis.
3. We, draw the scatter plot for the past data (**actual data values** like [x_1, y_1], [x_2, y_2], [x_3, y_3]…)
4. We, compute the best optimal values for α and β using the given past data.
5. We, now have $y = \alpha + \beta.x$
6. We, find all **predicted values** of y (like $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$, …) corresponding to each x value (like x_1, x_2, x_3, …) substituted into the equation in step-5.

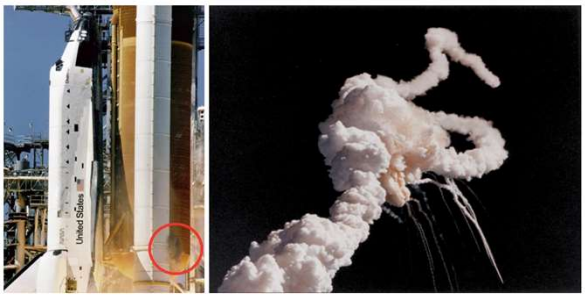
LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

7. Now, for any new unseen 'x' value, the corresponding 'y' value can be found out.
   - This is because α, β, and **x** are there…**just plug them** - in $y = \alpha + \beta x$). **This is** prediction.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

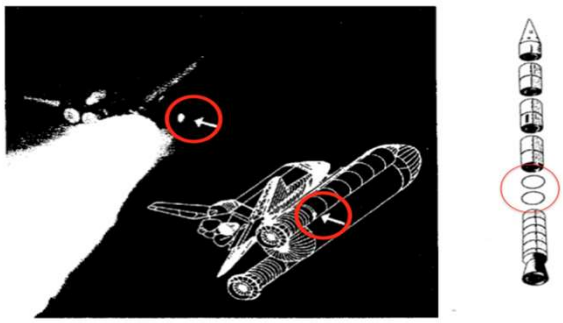
Example:



- Crash that caused the disintegration of the Challenger space shuttle by Nasa on **January 28, 1986**.
- Seven crew members on board were killed.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…



LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

LEARN . GROW . EXCEL

Contd…

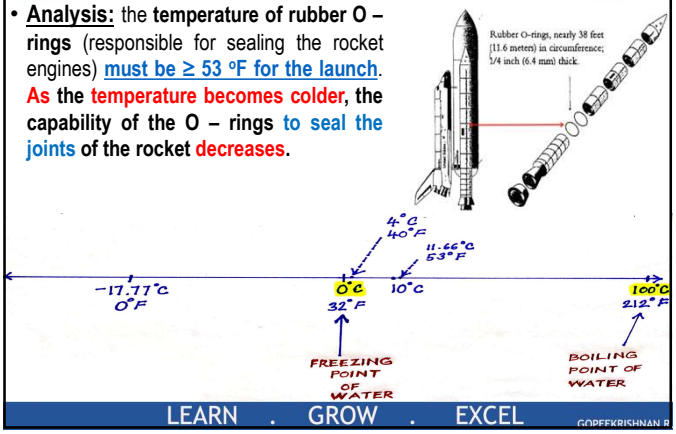
- **Reason:** *launch temperature. The weather on the launch day was unusually cold i.e., below freezing (i.e., **below 0^0C**).*

Fahrenheit	Celcius	Meaning
0 °F	-17.77 °C	
25 °F	-3.88 °C	
30 °F	-1.11 °C	
32 °F	0 °C	Freezing Point of Water
212 °F	100 °C	Boiling Point of Water

LEARN . GROW . EXCEL

Contd…

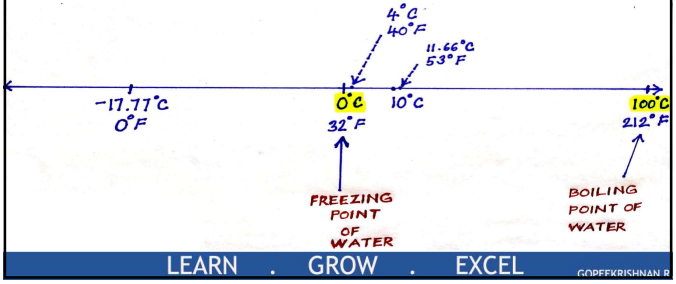
- **Analysis:** the **temperature of rubber O – rings** (responsible for sealing the rocket engines) **must be ≥ 53 °F for the launch**. **As the temperature becomes colder, the capability of the O – rings to seal the joints of the rocket decreases.**



LEARN . GROW . EXCEL

Contd…

- The engineers **never** tested **(earlier)** the withstand capacity of rubber O – rings ***below* 40 °F (= 4 °C).**
- Note that the temperature on the launch day of Challenger was unusually cold and below freezing point i.e., **below 32 °F (= 0 °C).**



LEARN . GROW . EXCEL

Contd…

- **Rocket engineers knew that cold temperatures** could make
  - **the components like rubber O-Rings more brittle** and,
  - **such components could not be used for sealing rocket parts properly**.
- AND….they knew further that…launches at such **cold temperatures could result in dangerous fuel leakage**.

Contd…

- But they needed data to support this claim.

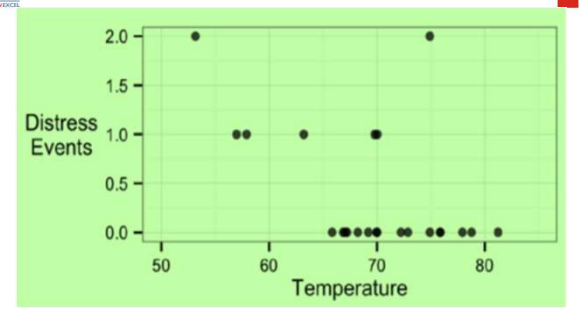
Contd…

- They prepared a **REGRESSION MODEL**….with
  - **Temperature** along **X – Axis** and
  - **Distress Events** along **Y – Axis**.
- Prepared a Scatter Plot.

Contd…

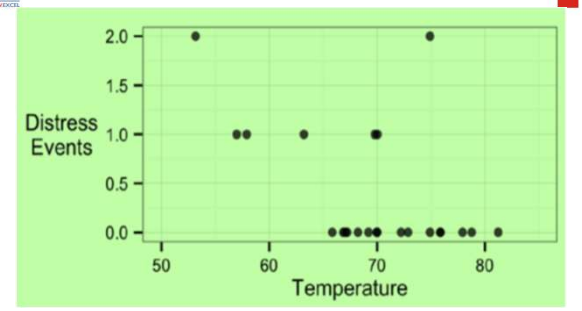
- **Past data** (training data) is used for plotting…
- This past data … is data about **23 previous** (before the Challenger launch on January 28, 1986) **space shuttle launches**.
  - Temperatures of these 23 launches along the X – Axis.
  - Number of Distress Events along the Y – Axis.
- This launch i.e., Challenger space shuttle launch has to be considered as the 24th launch (unseen test data).

Contd…



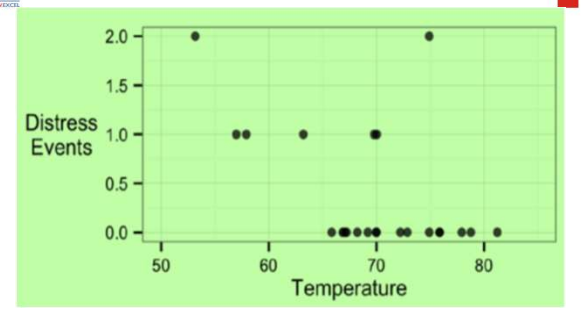
- Number of O-Ring Distress Events observed for the previous 23 space shuttle launches against temperatures are shown here.

Contd…



- On examining the plot…there is an apparent trend…
- Launches occurring at **higher temperatures** tend to have **fewer O-Ring Distress Events**.

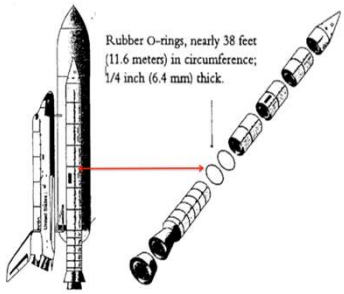
Contd…



- Additionally, the coldest temperature had **two** (**2.0** in the Y – Axis) Distress Events.
- This 2.0 **was also** reached at a higher temperature (almost **74^0 F**) too.

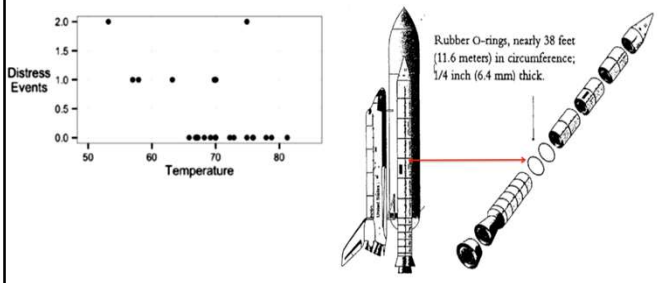
Contd…

- A rocket has 6 primary O – Rings.
- Therefore, upto 6.0 distresses can occur per flight.



Contd…

- A rocket can usually survive with **at the most** 1.0 distress points.
- But, each additional distress points increase failures.



Contd…

- With these much of information, the decision to launch the space shuttle at a temperature less than the freezing point **was embarrassing**.
- But… **we have to prove this using regression…**
- That is, at a temperature below the freezing point, NASA should had postponed its decision to launch its 24th mission on January 28, 1986.

Contd…

- We know, a simple linear regression model defines a relationship between a dependent variable and a single independent predictor variable using a line defined by an equation of the following form:

$$y = \alpha + \beta x$$

- α describes where the line crosses (intercepts) the Y – axis.
- β is the slope that describes the expected reduction in the number of O – ring failures for each degree increase in the temperature.

Contd…

- Assume, we know α and β respectively for this example.
- **i.e., α= 3.70 and, β= -0.048.**
- Therefore, our **regression model** is

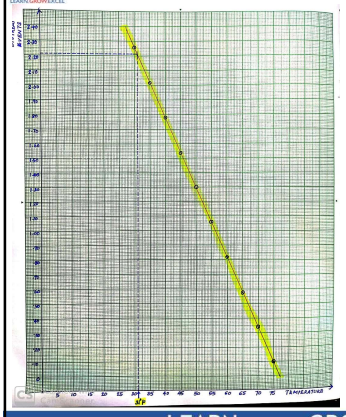
$$y = 3.70 - 0.048x$$

- **Note:** *We will discuss soon, how α and β values are estimated.*

Contd…

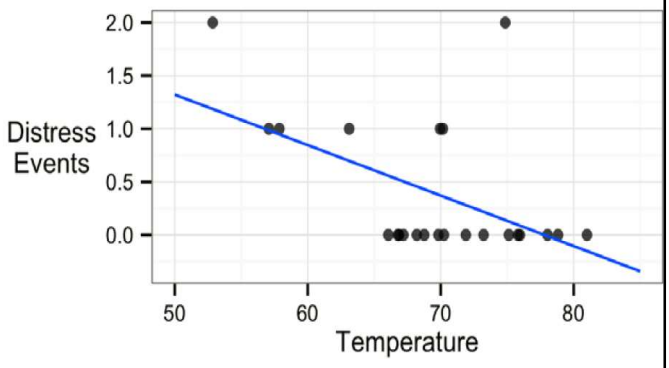
- Now, we draw the scatter plot **(assume, the training data is already in the plot)**.
  - Along the X – Axis, we have the Temperature.
  - Along the Y – Axis, we have the Distress Points.
- Substitute each temperature value (x) to the above formulation (**y= 3.70 – 0.048x**) to get the predicted distress points (y).
- Now we connect all the new predicted y values, to get the **regression straight line**.
- The **straight line contains all the predicted values only…the training data are scattered throughout the plot.**

x	30⁰F	35⁰F	40⁰F	45⁰F	50⁰F	55⁰F	60⁰F	65⁰F	70⁰F	75⁰F
y	2.26	2.02	1.78	1.54	1.30	1.06	0.82	0.52	0.34	0.10



- **From the plot…**
  - At **60 ⁰F**, we predict **0.82 Distress Events**.
  - At **70 ⁰F**, we predict around **0.3 Distress Events**.
  - Now, at **31 ⁰F** (or, **0⁰ C**), we would expect about **3.70 – 0.048*31 = 2.21 Distress Events.**
- **This means, the Challenger launch was nearly 3 times more risky than the typical launch at 60 ⁰F, and over 8 times more risky than a launch at 70 ⁰F.**

Contd…



Contd…

- From the scatter plot, it is clear that, **the line does not pass through each data point exactly**.
- Instead, it cuts through the data somewhat evenly.

Ordinary Least Square Estimation

- To find the optimal values for α and β **(i.e., to find the best straight line among the number of possible ones),** an estimation method known **as Ordinary Least Squares (OLS)** is used.

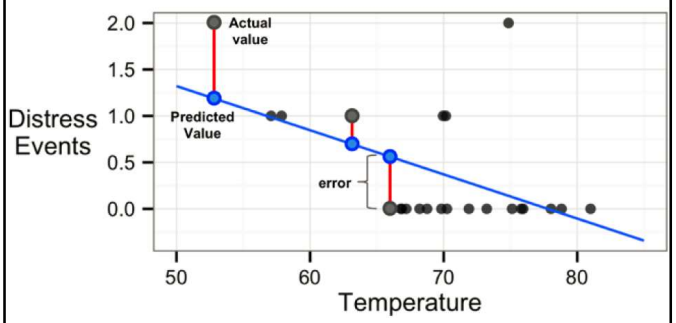
LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- In OLS Regression, the slope (β) and intercept (α) are chosen **in such a way that** they **minimize the sum of the squared errors**.
- This error is the vertical distance between
  - the **predicted y value** and,
  - the **actual y value**.
- These errors are known as **residuals.**
- This is illustrated in the following diagram.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…



LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- Let $\hat{y}_i$ be the **predicted value** of y_i.
- Then **the sum of the squares of errors** is given by

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

- So, we are required to find the values of α and β such that **E is minimum**.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- By using proper values found out for α and β, we can minimize the distance between any value on the scatter plot (past data) and, any value on the regression line (predicted value).
- There can be a number of such lines **possible…but the one that minimizes the errors between all such points is the best regression line that is selected**… **depends only on good values of α and β.**

Finding α and β

- The **means of x, y** are given by

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

Contd…

- The **variance of x** is given by

$$\text{Var}(x) = \frac{1}{n-1}\sum(x_i - \bar{x}_i)^2$$

Contd…

- The **co-variance of x *and* y**, denoted by **Cov(x,y)** is defined as

$$\text{Cov}(x,y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

Contd...

- It can be shown that the optimal values of α and β can be computed using

$$\beta = \frac{Cov(x, y)}{Var(x)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

Example

- **Obtain** the **linear regression model** for the following data set.

x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

Contd...

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x}).(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \alpha = \bar{y} - \beta . \bar{x}$$

Contd...

$$\bar{y} = \frac{1.00 + 2.00 + 1.30 + 3.75 + 2.25}{5} = \frac{10.3}{5} = 2.06$$

$$\bar{x} = \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3.00$$

Contd...

$$\beta = \frac{(1-3.00).(1-2.06)+(2-3.00).(2-2.06)+(3-3.00).(1.30-2.06)+(4-3.00).(3.75-2.06)+(5-3.00).(2.25-2.06)}{(1-3.00)^2+(2-3.00)^2+(3-3.00)^2+(4-3.00)^2+(5-3.00)^2}$$

$$= \frac{-2\times-1.06+-1\times-.06+0\times-0.76+1\times1.69+2\times.19}{(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2}$$

$$= \frac{2.12+.06+1.69+.38}{4+1+1+4} = \frac{4.25}{10} = 0.425$$

Contd...

$$\alpha = \bar{y} - \beta.\bar{x} = 2.06 - 0.425\times 3.00$$

$$= 2.06 - 1.275$$

$$= 0.785$$

Contd...

∴ The linear regression model for the given data set is

$$y = 0.785 + 0.425x$$

Correlation

- The **correlation** between two variables (usually between x and y) **is a number** that **indicates** **how closely (strongly) their relationship follows a straight line**.
- When we just say correlation, we just refer to **Pearson's Correlation Coefficient.**

Contd…

- **The value found out by this coefficient always ranges between: -1 and +1**
  - The **extreme values of correlation** (**-1** and **+1**) indicate a **perfect linear relationship**.
  - A **correlation value close to 0** indicates **the absence of a linear relationship**.

Contd…

- The following formula defines **Pearson's Correlation**:

$$\rho_{x,y} = \mathrm{Corr}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}$$

- Using this formula, we can calculate the correlation between the launch temperature (x) and the number of O – ring distress events (y).

Contd…

- Computationally, the correlation gave us a result of **-0.5111264**.

-1.....-0.511.....-0.5.....0.....+0.5.....+1

- As the correlation value of **-0.5111264** is very near to **-1**, we can judge that, **there is a moderately strong negative linear association** between temperature (x) and distress points (y).

Contd…

- Thus, for the NASA engineers studying the O – ring data, this **would had been a very clear indicator that a low temperature launch could be problematic**.

Multiple Linear Regression

- **Most real – world analyses have more than** one independent variable.
- In such instances, we would be using **multiple linear regression**…for most numeric prediction tasks.

Contd…

- **Example:**
  - The selling price of a house (y) can depend on
    - the desirability of the location (x_1),
    - the number of bedrooms (x_2),
    - the number of bathrooms (x_3),
    - the year the house was built (x_4),
    - the square footage of the plot (x_5) and,
    - a number of other factors.

Contd…

- This is an extension of simple linear regression.
- In both the cases, the goal is similar - **find value of beta coefficients β_i that minimizes the prediction error of a linear equation.**
- Key difference (between SLR and MLR): in MLR, **there are additional terms** for additional independent variables.

Contd…

- MLR equations generally follow the following equation:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

- **where:**
  - α is the intercept term
  - β_i are the COEFFICIENTS
  - x_i are the actual values
  - ε the error term

Contd…

- **Understanding This Equation:**
  - For each independent variable (feature), a coefficient (β) is given.
  - That is, each independent variable will have a separate estimated effect on the value of y.
  - Usually, α is represented as β_0.
  - Therefore, the equation can be rewritten as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

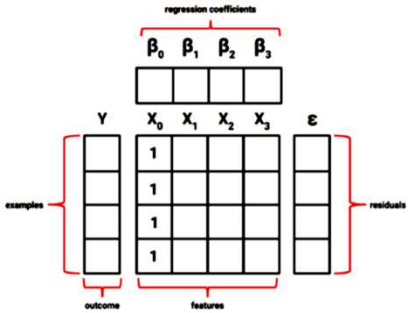
Contd…

- Though we know the intercept β_0 is independent on any of the x variables, we could write β_0 as $\beta_0.x_0$, where x_0 is a constant with the value 1.

Contd…

- Thus, we can re-write the MLR equation as

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$



Contd…

- The above representation can be described in a **matrix form** as:

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

- where:
  - Y: is a vector, a row for every example.
  - X: is a matrix of independent variables.
  - β: is a vector of regression coefficients.
  - ε: is another vector of residuals.
- Goal is:
  - Solve for β.

Contd…

- Best estimate of β can be computed as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

- where:
  - T: indicates the transpose of the given matrix.
  - $()^{-1}$: indicates the matrix inverse.

Strengths & Weaknesses of MLR

STRENGTHS	WEAKNESSES
• Most common approach for modeling numeric data.	• Makes strong assumptions about the data.
• Can be adapted to model any modeling task.	• The model's form must be specified by the user in advance.
• Provides estimates of both the strength and size of the relationships among features and the outcome.	• Does not handle missing data.
	• Only works with numerical data; so, categorical; data requires extra processing.
	• Requires some knowledge of Statistics to understand the model.

References

1. Machine Learning with R, Second Edition, Brett Lantz, PACKT Publishing.