Introduction to Statistical Methods

Regression Analysis

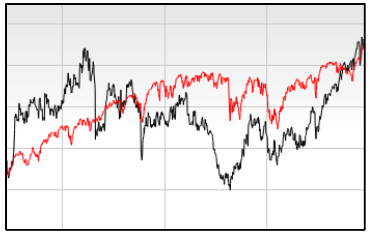
Revision-2.0

Prof Vineet Garg

BITS Pilani Work Integrated Learning Programmes (WILP)
Bangalore Professional Development Center

Introduction

- In statistics, **Regression Analysis** is a set of statistical processes for estimating the relationships between a dependent variable (the outcome variable) and one or more independent variables (also called predictors, covariates or features).
- Regression Analysis helps to formulate a **supervised model** that describes the data and can be used to forecast.
- In this module, few types of regression analysis will be reviewed.
- **Example**: Researchers are interested to find out if the stock price of two companies rise and fall in the related manner. What are the measures available in statistics to estimate the relationship between these two variables? Let us understand that first.



- **Covariance**: The covariance (s_{XY}) of two variables X and Y measures how the two are linearly related. A positive covariance would indicate a positive linear relationship, a negative covariance would indicate the opposite and a 0 covariance would indicate uncorrelated.

$$S_{xy} = \frac{1}{(N-1)} \sum_{k=1}^{N} (x_k - \bar{x})(y_k - \bar{y})$$

Did you notice the resemblance?

$$Variance\ (s_x^2) = \frac{1}{(N-1)} \sum_{k=1}^{N} (x_k - \bar{x})^2$$

- **Correlation**: The correlation of two variables X and Y equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related between with a value between -1 and 1. This is also called **Pearson's Coefficient of Correlation** named after the English statistician Karl Pearson who developed it. It is often denoted by **r**.

$$r = \frac{S_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})\ /(N-1)}{\sqrt{\sum_{k=1}^{N}(x_k - \bar{x})^2\ /(N-1)}\sqrt{\sum_{k=1}^{N}(y_k - \bar{y})^2\ /(N-1)}} = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{N}(x_k - \bar{x})^2}\ \sqrt{\sum_{k=1}^{N}(y_k - \bar{y})^2}}$$

Examples

$X = \{-3, 6, 0, 3, -6\}$ and $Y = \{1, -2, 0, -1, 2\}$

Mean of X = 0, mean of Y = 0

$S_{XY} = \{(-3-0)(1-0) + (6-0)(-2-0) + (0-0)(0-0) + (3-0)(-1-0) + (-6-0)(2-0)\} / 4$

$= (-3-12+0-3-12)/4 = (-30)/4 = \mathbf{-7.5}$

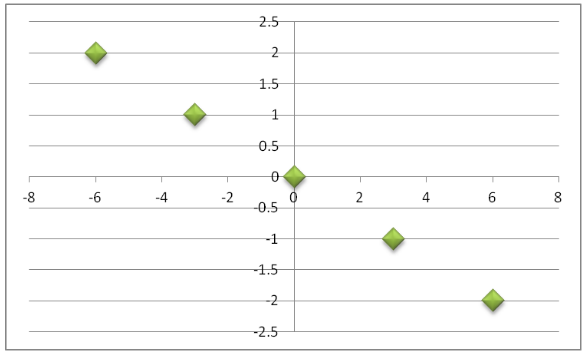
(Negative sign of covariance; X decreases ⇒ Y increases)

$\sigma_X = 4.74$

$\sigma_Y = 1.58$

Correlation (X, Y) = -7.5 / (4.74 x 1.58) = **-1**

Strong Negative Correlation



$X = \{3, 6, 0, 3, 6\}$ and $Y = \{1, 2, 0, 1, 2\}$

Mean of X = 3.6, mean of Y = 1.2

$S_{XY} = \{(3-3.6)(1-1.2) + (6-3.6)(2-1.2) + (0-3.6)(0-1.2) + (3-3.6)(1-1.2) + (6-3.6)(2-1.2)\} / 5$

$= (0.12+1.92+4.32+0.12+1.92)/4 = 8.4/4 = \mathbf{2.1}$

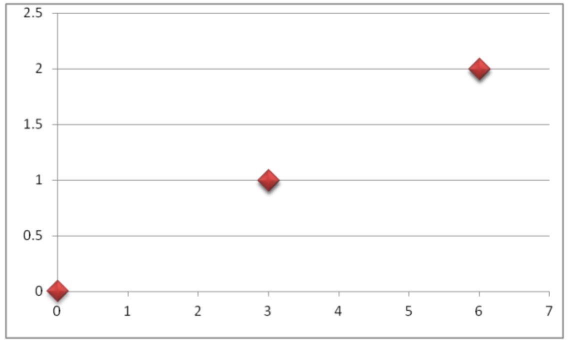
(Positive sign of covariance; X increases ⇒ Y increases)

$\sigma_X = 2.50$

$\sigma_Y = 0.84$

Correlation (X, Y) = 2.1 / (2.50 x 0.84) = **1**

Strong Positive Correlation



Exercise

1. Find out the covariance and correlation of the vectors X = {-3, -2, -1, 0, 1, 2, 3} and Y = {9, 4, 1, 0, 1, 4, 9} and compare the answer with their approximate scatter plot.

   (Answer Covariance and Correlation = 0, Parabolic)

2. Monthly average Net Asset Value (NAV) of a mutual fund and the amount invested by an investor in that fund for few months are provided below. Mathematically establish if the investor is considering the NAV of the last month for the investment in the current month.

   (Answer: No, Covariance = 0)

Months (2019)	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**June**
NAV (₹)	44	43	42	45	43	-
Investment (₹)	-	4900	6500	7700	7900	5500

Simple Linear Regression

- The most elementary regression model is called the **simple linear regression** or bi-variate regression involving two variables in which one variable is predicted by the another variable.
- In linear regression, the variable to be predicted is called the **dependent variable** and is designated as y. The predictor is called the **independent variable** and is designated as x.
- The first step in the linear regression is to have some insight into the data by visualization techniques (like scatter plot) to check if an imaginary line can be drawn that passes through most of the points.
- The next step is to find out the equation of the line itself.
- The equation of a line is given by:

$$ax + by + k = 0$$

$$\text{or, } y = (-a/b)x + (-k/b)$$

Where (-a/b) is the slope of the line and (-k/b) is the Y-axis intercept of the line.

- a, b and k are constants. Any coordinate on this line will satisfy the equation.
- Above equation can be simply written as:

$$y = mx + c$$

Where m is the slope and c is the intercept on the Y-axis.

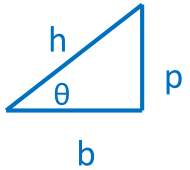
- In statistics, the regression line is written in a slightly different form:

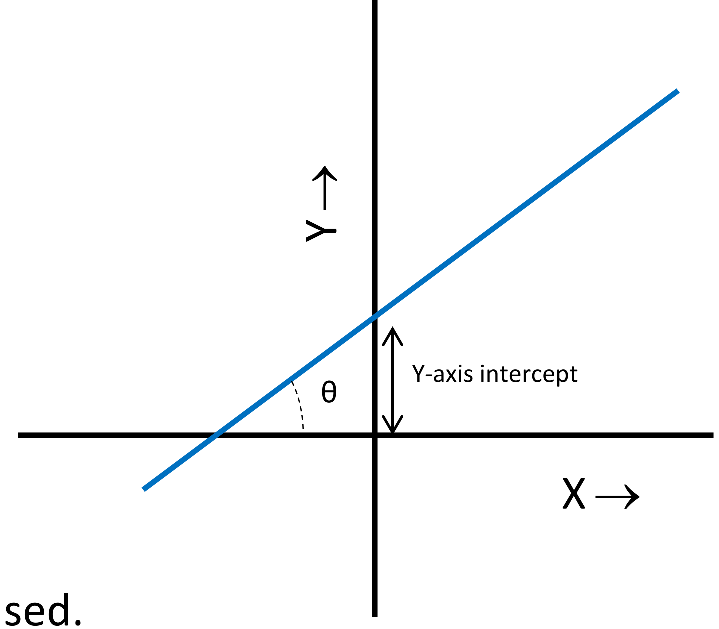
$$\hat{y} = \beta_0 + \beta_1 x$$

Where, $\hat{y}$ = predicted value of y, β_0 = the intercept on the Y – axis, β_1 = the slope

- In statistics, most of the time, we deal with samples, so the **β terms** are denoted by corresponding **b terms** (b_0 and b_1) indicating that sample is being used.

sin θ = p / h
cos θ = b / h
tan θ = p / b





Example

- The corresponding equation of line for the sample is given by:

$$\hat{y} = b_0 + b_1 x$$

- The values of b0 and b1 are calculated by:

$$b_1 = \frac{Covariance\ of\ \mathrm{x}\ and\ \mathrm{y}}{Variance\ of\ \mathrm{x}} = \frac{S_{xy}}{S_{xx}}$$

$$\text{then, } b_0 = \bar{y} - b_1.\bar{x}$$

Example: Form the regression equation for the given data.

$\bar{x} = 3.5$

$\bar{y} = 30.17$

$S_{xy} = 5.9$

$S_{xx} = 3.5$

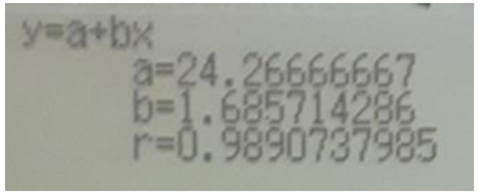
$b_1 = 5.9\ /3.5 = 1.69$

$b_0 = 30.17\ - (1.69 \text{ x } 3.5) = 24.26$

So regression line is: $\boldsymbol{\hat{y} = 24.26 + 1.69.x}$

It can be used to forecast the value of y for x = 7 as 36.09

x **(independent)**	**y** **(dependent)**
1	26
2	28
3	29
4	31
5	32
6	35



```
y=a+bx
        a=24.26666667
        b=1.685714286
        r=0.9890737985
```

Leverage your scientific calculator to verify the solution.
Menu → Statistics → y = a + b.x

Exercise

1. Develop the simple regression model for the following data, considering y as the dependent variable.

	1	2	3	4	5	6	7
x	140	119	103	91	65	29	24
y	25	29	46	70	88	112	128

(Answer: $\hat{y}$ = 144.414 – 0.898x)

2. The count of beds and full time support staff needed in a hospital are given in the table below. Develop a simple regression model, considering staff count (y) as the dependent variable.

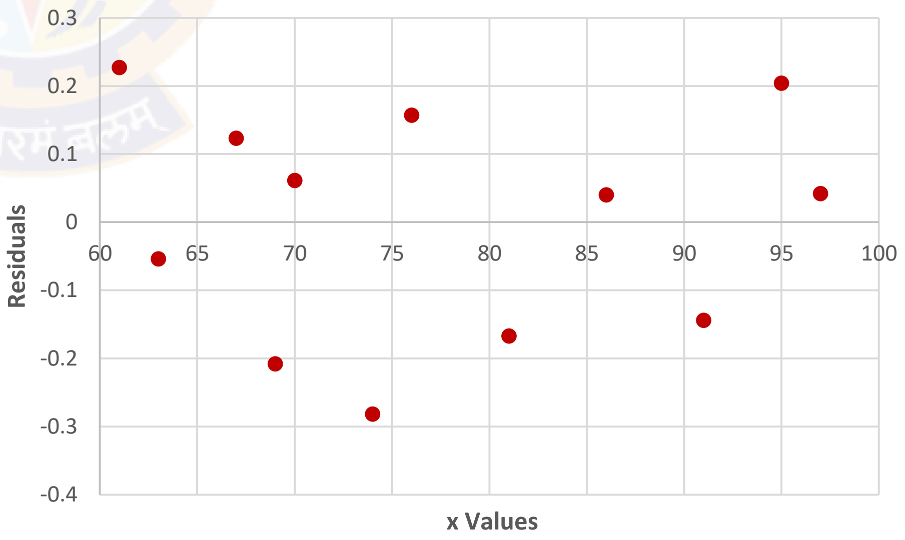
	1	2	3	4	5	6	7	8	9	10	11	12
Beds	23	29	29	35	42	46	50	54	64	66	76	78
Staff Count	69	95	102	118	126	125	138	178	156	184	176	225

(Answer: $\hat{y}$ = 30.888 + 2.232x)

Verification: Residual Analysis

- How do we ensure that the line equation that is obtained truly represents the model for the bivariate data?
- The historical data can be used for verification performing the **back-fitting**. The predicted $\hat{y}$ values are compared with the actual y values. The difference is called the **residual** (y - $\hat{y}$). This is also called the **Error-terms** when the entire population is in the context.
- For the data shown in the table, the simple regression model can be found as: $\hat{\mathbf{y}}$ **= 1.57 + 0.0407x** and residuals are calculated (establish the equation and verify the result yourself).
- The sum of all these residuals will be always 0 (ignoring the minor rounding off differences).
- The residuals can be plotted against x values as shown.

#	x	y	$\hat{y}$	$y - \hat{y}$
1	61	4.28	4.053	0.227
2	63	4.08	4.134	-0.054
3	67	4.42	4.297	0.123
4	69	4.17	4.378	-0.208
5	70	4.48	4.419	0.061
6	74	4.3	4.582	-0.282
7	76	4.82	4.663	0.157
8	81	4.7	4.867	-0.167
9	86	5.11	5.07	0.04
10	91	5.13	5.274	-0.144
11	95	5.64	5.436	0.204
12	97	5.56	5.518	0.042



What does it reveal?

Simple Regression: Assumptions

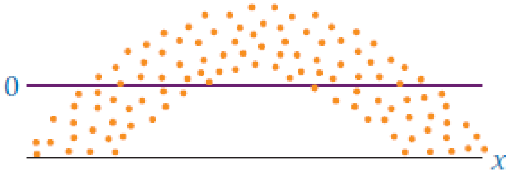
The plot of <u>residuals for large samples</u> or <u>error-terms for the population</u>, w.r.t. the independent variable x, must exhibit the following properties:

i. **The model has to be linear**: in the shown figure (i) the error terms are negative for low and high values of x and are positive for middle values of x, so it is not linear.

ii. **The error terms need to have constant variances (Homoscedasticity)**: in the shown figure (ii), the error terms have large variances for the small values of x and small variances for the large values of x, so the model is heteroscedastic.

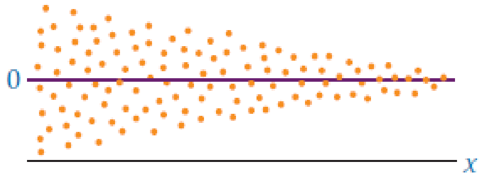
iii. **The error terms are to be independent:** in the shown figure (iii), the error terms themselves are following a trend. They are not independent and the model in not perfect.

iv. **The error terms are to be normally distributed**: the distribution of the error terms should follow a normal distribution. We will review it in the next slide.



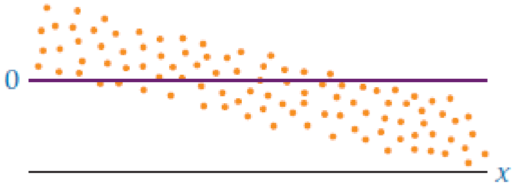
✗

<u>(i) Non-linear Model</u>



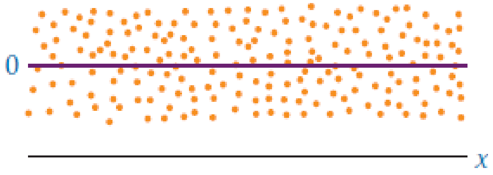
✗

<u>(ii) Variable Variance of the Error-Terms</u>



✗

<u>(iii) Inter-dependent Error-Terms</u>

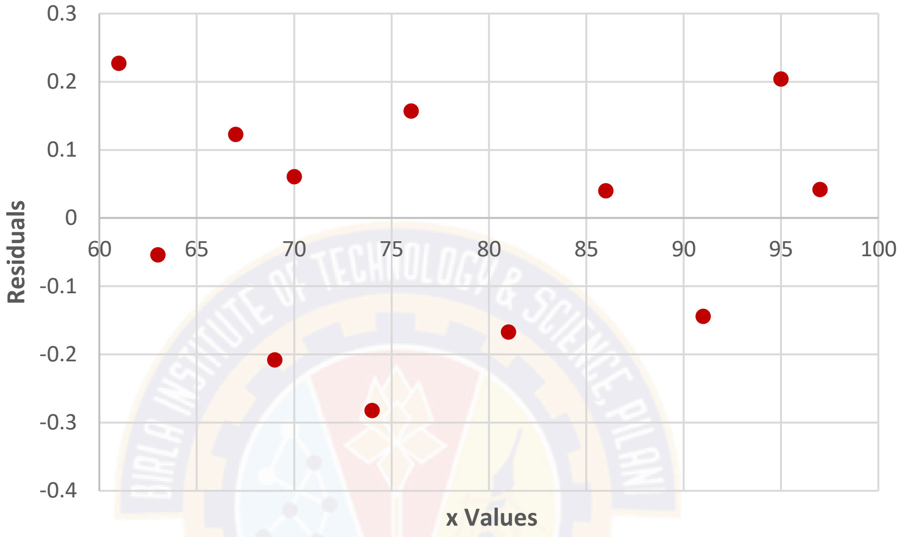


✓

<u>Healthy Error-Terms Plot</u>

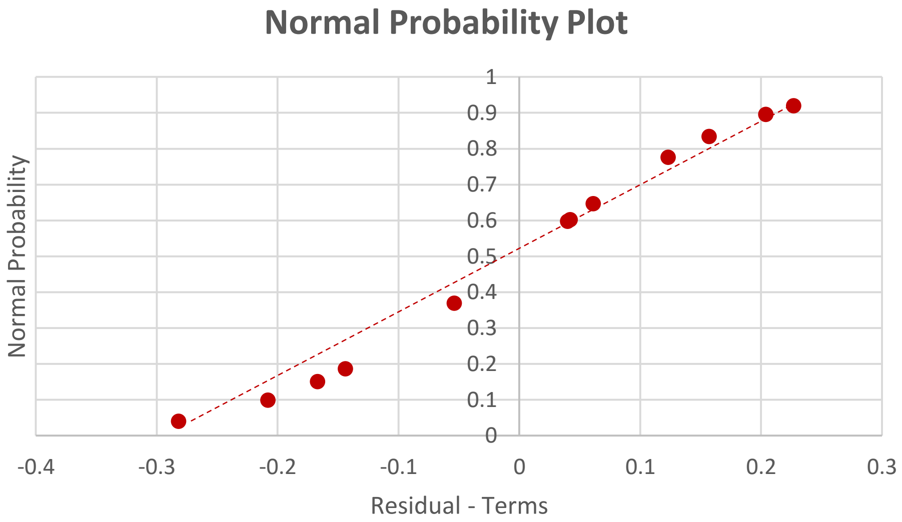
Example

#	x	y	$\hat{y}$	$y-\hat{y}$
1	61	4.28	4.053	0.227
2	63	4.08	4.134	-0.054
3	67	4.42	4.297	0.123
4	69	4.17	4.378	-0.208
5	70	4.48	4.419	0.061
6	74	4.3	4.582	-0.282
7	76	4.82	4.663	0.157
8	81	4.7	4.867	-0.167
9	86	5.11	5.07	0.04
10	91	5.13	5.274	-0.144
11	95	5.64	5.436	0.204
12	97	5.56	5.518	0.042



Sorted $(y-\hat{y})$	Normal Probability
-0.282	0.04076632
-0.208	0.099458648
-0.167	0.151195841
-0.144	0.18694529
-0.054	0.369517435
0.04	0.597801455
0.042	0.602574974
0.061	0.647066259
0.123	0.776518432
0.157	0.834116644
0.204	0.89633793
0.227	0.919701463

- The given data does not have sufficient data points to check the linearity, variance and independence of the residual-terms. But their distribution can be verified; if it is normal.
- The normal probabilities of the residual terms $(y-\hat{y})$ are calculated using the following steps:
  - Sort the residual terms $(y-\hat{y})$.
  - Identify the mean and standard deviation of $(y-\hat{y})$ set.
  - Perform the z-transformation for each term.
  - Find out the Normal probability from -∞ to the z value for all terms.
- Plot the Residual-Terms vs. Probabilities.
- Near straight line indicates that points are in Normal Distribution. ***Why?***



Error of the Estimate

- We have reviewed that the sum of residuals will always be 0. So how do we compare two regression models on the same dataset?
- **Sum of Squared Error (SSE)** which is $(y - \hat{y})^2$ is a useful measure. As a matter of fact, the terms b_0 and b_1 are calculated using calculus by minimizing the SSE.
- Standard deviation of the error is called **Standard Error of the Estimate (s_e)**. When dealing with n samples, the denominator is used as (n-2) because *slope* and *intercept* are the two parameters being approximated for the population.

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

- If residual terms are normally distributed, then the values of these terms will follow the μ ± k.σ rule, where μ = 0 and σ = s_e.
- 99.7 % of the residual terms will be within the (μ ± 3.σ) or (± $3s_e$) range.
- Any residual term beyond the select range of k can be considered an outlier.

$$\hat{y} = 1.57 + 0.0407x$$

#	x	y	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
1	61	4.28	4.053	0.227	0.051529
2	63	4.08	4.134	-0.054	0.002916
3	67	4.42	4.297	0.123	0.015129
4	69	4.17	4.378	-0.208	0.043264
5	70	4.48	4.419	0.061	0.003721
6	74	4.3	4.582	-0.282	0.079524
7	76	4.82	4.663	0.157	0.024649
8	81	4.7	4.867	-0.167	0.027889
9	86	5.11	5.07	0.04	0.0016
10	91	5.13	5.274	-0.144	0.020736
11	95	5.64	5.436	0.204	0.041616
12	97	5.56	5.518	0.042	0.001764
Sum				**-0.001**	**0.314337**

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{0.314337}{12-2}} = 0.1773$$

Coefficient of Determination (R^2)

- The square of coefficient of correlation (r) is called the **Coefficient of Determination (R^2).**
- An R^2 of 1 means perfect prediction of y by x and that 100% of the variability of y accounted for is by x.
- An R^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x.
- For the given table, r = 0.9482 and R^2 = 0.899. It means the regression model captures for the 89.9% of variability of y and 10.1% variability is unexplained.
- Critical applications (e.g. space, nuclear reaction, human safety etc.) might need a high coefficient of determination of the regression model for the forecasting.

Example: for a critical mission a space agency has a bi-variate data gathered for x and y. Can it use the simple regression model to forecast the values of y from x?

The coefficient of correlation (r) is = -0.98591

The coefficient of determination (R^2) is = $(-0.98591)^2$ = 0.972018

A high value of R^2 (97.20%) indicates that the regression model would capture a high degree of variability of y with x and it would be a very good model.

#	x	y
1	61	4.28
2	63	4.08
3	67	4.42
4	69	4.17
5	70	4.48
6	74	4.3
7	76	4.82
8	81	4.7
9	86	5.11
10	91	5.13
11	95	5.64
12	97	5.56

r = 0.9482 and R^2 = 0.899

x	y
140	25
119	29
103	46
91	70
65	88
29	112
24	128

Hypothesis Testing for Slope

- For the given data, if the mean of y is calculated and it comes as $\bar{y}$ = 4.7242. This is called the mean model.
- Note that it is mean model $\bar{y}$ and not $\hat{y}$. The $\hat{y}$ model has to be calculated separately with linear regression technique (shown on the top of the table).
- Why the mean model ($\bar{y}$ = 4.7242) cannot be used as the predicted value for any value of x?
- This mean model basically represents a horizontal line of 0 slope.
- Does the regression model that is established offer anything more than the $\bar{y}$ model?
- Because the slope of the $\bar{y}$ model is zero, one way to determine whether the regression line adds significant predictability is to test if its slope is different from zero.
- Therefore, the requirement boils down to test the hypothesis:

  H_0: $\beta_1 = 0$ mean model is good enough

  H_A: $\beta_1 \neq 0$ no mean model is not good
- This is a two-tailed test. Left-end (negative relationship) and right-end (positive relationship) one-tailed hypothesis testing can also be performed.
- The hypothesis testing will involve t-statistics.

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = 1.57 + 0.0407x$$

#	x	y	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
1	61	4.28	4.053	0.227	0.051529
2	63	4.08	4.134	-0.054	0.002916
3	67	4.42	4.297	0.123	0.015129
4	69	4.17	4.378	-0.208	0.043264
5	70	4.48	4.419	0.061	0.003721
6	74	4.3	4.582	-0.282	0.079524
7	76	4.82	4.663	0.157	0.024649
8	81	4.7	4.867	-0.167	0.027889
9	86	5.11	5.07	0.04	0.0016
10	91	5.13	5.274	-0.144	0.020736
11	95	5.64	5.436	0.204	0.041616
12	97	5.56	5.518	0.042	0.001764
Sum				**-0.001**	**0.314337**

$$t = \frac{b_1 - \beta_1}{s_b/\sqrt{n}}$$

$$s_b = \frac{s_e}{\sqrt{S_{xx}}}, \ s_e = \sqrt{\frac{SSE}{(n-2)}}$$

S_{xx} is variance for x.

Example

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = 1.57 + 0.0407x$$

#	x	y	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
1	61	4.28	4.053	0.227	0.051529
2	63	4.08	4.134	-0.054	0.002916
3	67	4.42	4.297	0.123	0.015129
4	69	4.17	4.378	-0.208	0.043264
5	70	4.48	4.419	0.061	0.003721
6	74	4.3	4.582	-0.282	0.079524
7	76	4.82	4.663	0.157	0.024649
8	81	4.7	4.867	-0.167	0.027889
9	86	5.11	5.07	0.04	0.0016
10	91	5.13	5.274	-0.144	0.020736
11	95	5.64	5.436	0.204	0.041616
12	97	5.56	5.518	0.042	0.001764
Sum				**-0.001**	**0.314337**

For the 95% confidence level, test the hypothesis:

H_0: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

From the hypothesis it is a two-tailed test.
Given that:
1-α = 0.95, so α = 0.05 and α/2 = 0.025
t-critical value for α/2 and given degree of freedom (v = n-2 =10) = **2.228**
Hypothesis value $\beta_1 = 0$
From the regression model:
b1 = 0.0407

$$s_e = \sqrt{\frac{\text{SSE}}{(n-2)}} = \sqrt{\frac{0.314337}{(12-2)}} = 0.1773$$

$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \frac{0.1773}{\sqrt{153.55}} = 0.01431$$

$$t - \text{statistic} = \frac{b_1 - \beta_1}{s_b/\sqrt{n}} = \frac{0.0407 - 0}{0.01431/\sqrt{12}} = \mathbf{9.852}$$

Since *t-statistic* > *t-critical* value, so null hypothesis is rejected. It means….
Regression model is providing predictive information more than $\bar{y}$ model.

$$t = \frac{b_1 - \beta_1}{s_b/\sqrt{n}}$$

$$s_b = \frac{s_e}{\sqrt{S_{xx}}}, \; s_e = \sqrt{\frac{\text{SSE}}{(n-2)}}$$

Exercise

In the previous exercise, the regression model was prepared to predict the staff-count based on the beds in a hospital. Using the confidence level as 99%, test the claim if the model has a positive slope.

	1	2	3	4	5	6	7	8	9	10	11	12
Beds	23	29	29	35	42	46	50	54	64	66	76	78
Staff Count	69	95	102	118	126	125	138	178	156	184	176	225

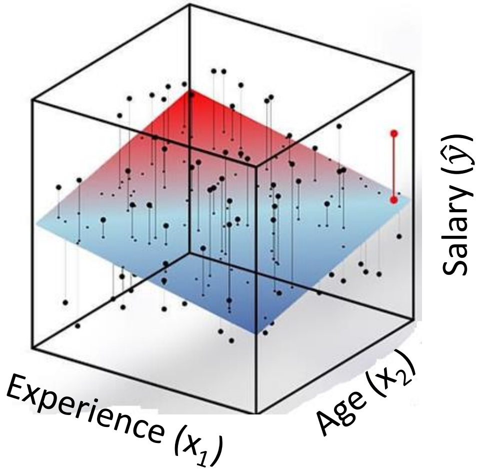
(Answer: Right-end tailed test, t-statistic = 8.84, claim is accepted; regression model has better predictability than $\bar{y}$ model)

Multiple Regression

- In simple regression, we have reviewed, that it is a bivariate linear regression in which one dependent variable y is predicted by one independent variable x.
- If there are two or more independent variables (x_1, x_2 etc.), it is called the **Multiple Regression**. Its model is represented by the following expression:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Here, k is the count of independent variables, β_0 is the **regression coefficient** and β_i coefficient is the **partial regression coefficient** for its corresponding independent variable (x_i).
- $\hat{y}$ that is an dependent variable and it is also called **response variable** in multiple regression terminology.
- The partial regression coefficient of an independent variable represents the increase that will occur in the value of response variable from a one-unit increase in that independent variable if all other variables are held constant. With this definition since all independent variables contribute to predict the response variable partially, their coefficients are called partial regression coefficients.
- If the independent variables are correlated with each other, it is called the problem of **collinearity** or **multicollinearity**.



Analogous to the simple regression, the requirement in the multiple regression is to fit a *plane* or a *surface*.

Multiple Regression: Solving the Equation

- From the multiple regression equation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- The error term (ε) can be written as:

$$\varepsilon = y - \hat{y}$$
$$\varepsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

- The Sum of Squared Error (SSE) for the independent variables:

$$\sum \varepsilon^2 = \sum [y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)]^2$$

- Partially differentiating the equation for different regression coefficients (β terms) and set them to 0 for minimizing the SSE:

$$\frac{\partial SSE}{\partial \beta_0} = -2.\sum [y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)] = 0$$
$$\frac{\partial SSE}{\partial \beta_1} = -2.\sum x_1 [y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)] = 0$$

……………………………

$$\frac{\partial SSE}{\partial \beta_k} = -2.\sum x_k [y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)] = 0$$

- Re-arranging the terms, these equations can be written as:

$$n\beta_0 + \beta_1 \sum x_1 + \beta_2 \sum x_2 + \cdots + \beta_k \sum x_k = \sum y$$
$$\beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_2 x_1 + \cdots + \beta_k \sum x_k x_1 = \sum y x_1$$

……………………………….

$$\beta_0 \sum x_k + \beta_1 \sum x_1 x_k + \beta_2 \sum x_2 x_k + \cdots + \beta_k \sum x_k^2 = \sum y x_k$$

- This set of equations can be written in the matrix form as:

$$\begin{bmatrix} n & \sum x_1 & \sum x_2 & \dots & \sum x_k \\ \sum x_1 & \sum x_1^2 & \sum x_2 x_1 & \dots & \sum x_k x_1 \\ \dots & \dots & \dots & & \dots \\ \sum x_k & \sum x_1 x_k & \sum x_2 x_k & \dots & \sum x_k^2 \end{bmatrix} . \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum y x_1 \\ \sum y x_2 \\ \dots \\ \sum y x_k \end{bmatrix}$$

- For the sample data, β terms are represented by b terms and the matrix equation can be solved to get the b terms.

Example: develop the regression model for the given data.

$$\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} . \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum yx \end{bmatrix}$$

x	y
1	26
2	28
3	29
4	31
5	32
6	35

$$\begin{bmatrix} 6 & 21 \\ 21 & 91 \end{bmatrix} . \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 181 \\ 663 \end{bmatrix}$$

Or, $\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 6 & 21 \\ 21 & 91 \end{bmatrix}^{-1} . \begin{bmatrix} 181 \\ 663 \end{bmatrix} = \begin{bmatrix} \mathbf{24.266} \\ \mathbf{1.6857} \end{bmatrix}$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\underbrace{ad-bc}_{\text{determinant}}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Check the link for the inverse of bigger square matrices. You can also calculate it using the calculator (e.g. x^{-1} button).

Example

Develop the regression model for the given data.

$$\begin{bmatrix} n & \sum x_1 & \sum x_2 & \cdots & \sum x_k \\ \sum x_1 & \sum x_1^2 & \sum x_2 x_1 & \cdots & \sum x_k x_1 \\ \cdots & \cdots & \cdots & & \cdots \\ \sum x_k & \sum x_1 x_k & \sum x_2 x_k & \cdots & \sum x_k^2 \end{bmatrix} . \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum y x_1 \\ \sum y x_2 \\ \cdots \\ \sum y x_k \end{bmatrix}$$

$$\begin{bmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_2 x_1 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{bmatrix} . \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum y x_1 \\ \sum y x_2 \end{bmatrix}$$

$$\begin{bmatrix} 10 & 1850 & 74 \\ 1850 & 358146 & 13756 \\ 74 & 13756 & 630 \end{bmatrix} . \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 301 \\ 55022 \\ 2383 \end{bmatrix}$$

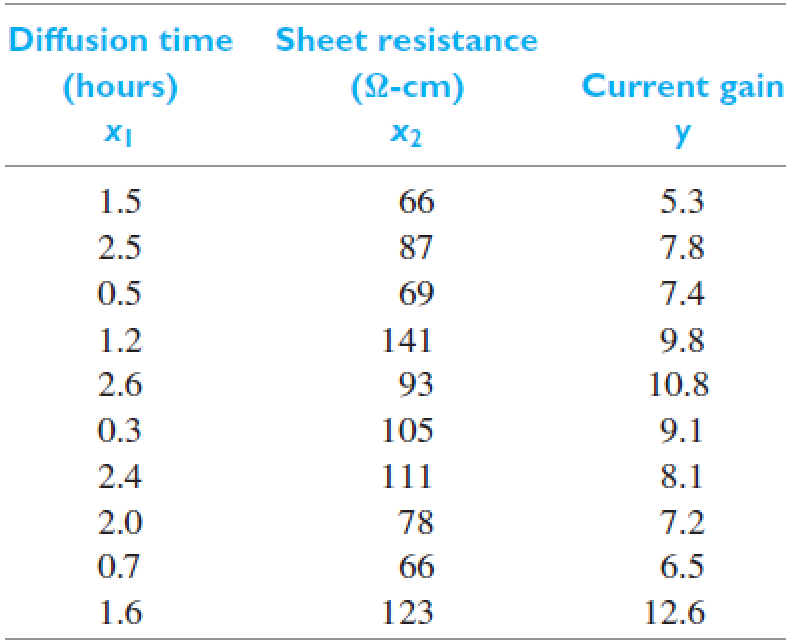
$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 10 & 1850 & 74 \\ 1850 & 358146 & 13756 \\ 74 & 13756 & 630 \end{bmatrix}^{-1} . \begin{bmatrix} 301 \\ 55022 \\ 2383 \end{bmatrix} = \begin{bmatrix} \mathbf{25.028} \\ \mathbf{-0.049} \\ \mathbf{1.9281} \end{bmatrix}$$

$\hat{y}$ = 25.028 - 0.049.x_1 + 1.9281.x_2

#	y	x1	x2
1	12	174	3
2	18	281	9
3	31	189	4
4	28	202	8
5	52	149	9
6	47	188	12
7	38	215	5
8	22	150	11
9	36	167	8
10	17	135	5

Exercise

Fit a regression plane for the following sample data to determine the relationship between processing variables (x_1 and x_2) and the current gain (y) of a transistor in the integrated circuit.



Diffusion time (hours) x_1	Sheet resistance (Ω-cm) x_2	Current gain y
1.5	66	5.3
2.5	87	7.8
0.5	69	7.4
1.2	141	9.8
2.6	93	10.8
0.3	105	9.1
2.4	111	8.1
2.0	78	7.2
0.7	66	6.5
1.6	123	12.6

(Answer: $\hat{y} = 2.266 + 0.225.x_1 + 0.0623.x_2$)

Metrics of Multiple Regression

- Let us say for some data, the actual value of y and predicted value of y ($\hat{y}$) is calculated after forming the regression model for k=2 independent variables.
- The difference between y and $\hat{y}$ provide the **residual terms**.
- The residual terms can be used to verify the assumptions of regression (like it was done for simple regression): linearity, homoscedasticity, independence and normal distribution.
- The sum of $(y - \hat{y})^2$ is called the **Sum of Squared Error (SSE).**
- Standard deviation of the error is called **Standard Error of the Estimate (s_e)**. When dealing with samples, the denominator is used as (n-k-1) for k independent variables.

$$s_e = \sqrt{\frac{\mathrm{SSE}}{n-k-1}}$$

- If residual terms are normally distributed, then the values of these terms will follow the μ ± k.σ rule, where μ = 0 and σ = s_e. 99.7 % of the residual terms will be within the (μ ± 3.σ) or (± $3s_e$) range.
- The proportion of variation of the dependent variable (y), accounted for by the independent variables in the regression model is called the **Coefficient of Multiple Determination (R^2)**. It is defined by:

$$R^2 = 1 - \frac{SSE}{n.S_{YY}}$$

#	y	$\hat{y}$	y-$\hat{y}$	$(y-\hat{y})^2$
1	63.000	62.499	0.501	0.251
2	65.100	71.485	-6.385	40.768
3	69.900	71.568	-1.668	2.782
4	76.800	78.639	-1.839	3.382
5	73.900	74.097	-0.197	0.039
6	77.900	75.653	2.247	5.049
7	74.900	83.012	-8.112	65.805
8	78.000	105.699	-27.699	767.235
9	79.000	68.523	10.477	109.768
10	83.400	81.139	2.261	5.112
11	79.500	88.282	-8.782	77.124
12	83.900	91.335	-7.435	55.279
13	79.700	85.618	-5.918	35.023
14	84.500	95.391	-10.891	118.614
15	96.000	85.456	10.544	111.176
16	109.500	102.774	6.726	45.239
17	102.500	97.665	4.835	23.377
18	121.000	107.183	13.817	190.909
19	104.900	109.352	-4.452	19.820
20	128.000	111.230	16.770	281.233
21	129.000	105.061	23.939	573.076
22	117.900	134.415	-16.515	272.745
23	140.000	132.430	7.570	57.305

2861.110

$$s_e = \sqrt{\frac{\mathrm{SSE}}{n-k-1}} \sqrt{\frac{2861.110}{23-2-1}} = 11.96$$

$$R^2 = 1 - \frac{SSE}{n.S_{YY}} = 1 - \frac{2861.11}{23 * 480.467} = 0.7411$$

Nonlinear Regression: Exponential Model

- In the table, the sales and advertising cost for seven companies are provided. The requirement is to fit a regression model to predict the sales (y) based on the spent money on advertising (x).
- When the sales is plotted against advertising expenditure, the plot reveals no linear relationship between the variables. The closest resemblance is with the exponential curve.
- Let us say the exponential relationship is represented by $\hat{y} = \alpha_0 \alpha_1^x$. Where $\alpha_0$and α_1 are constants.
- The relationship can be transformed taking log on the both the sides.

$$\log \hat{y} = \log \alpha_0 + \mathrm{x}.\log \alpha_1 \text{ --------- } -(i)$$

- Alternatively, it can be written as:

$$\widehat{y'} = \beta_0 + \beta_1 \mathrm{x}; \quad \text{where } \beta_0 = \log \alpha_0 \text{ and } \beta_1 = \log \alpha_1$$

- The sales values are transformed taking log on base-10 and advertising costs are retained as it is (as per the transformed form).
- Using the transformed data, the simple regression model is established as:

$$\widehat{y'} = 2.9003 + 0.4751.\mathrm{x} \text{ --- } -(ii)$$

- Comparing (i) and (ii):

$$\log \alpha_0 = 2.9003 \text{ and } \log \alpha_1 = 0.4751$$

- Taking the anti-logs,

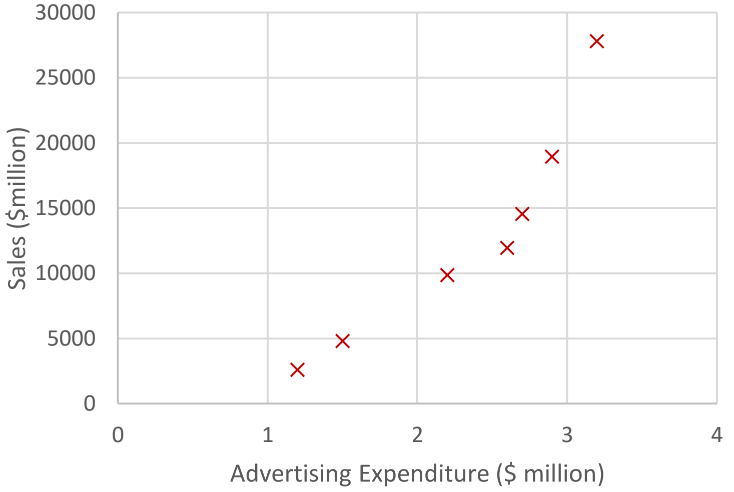
$$\alpha_0 = \text{antilog}_{10}(2.9003) = 794.88$$

$$\alpha_1 = \text{antilog}_{10}(0.4751) = 2.99$$

- So, the exponential relationship can be represented as:

$$\hat{y} = \mathbf{794.88.(2.99)^x}$$

#	Sales ($ million): y	Advertising Expenditure ($ million): x
1	2580	1.2
2	11942	2.6
3	9845	2.2
4	27800	3.2
5	18926	2.9
6	4800	1.5
7	14550	2.7



#	log of Sales on base 10: y'	Advertising Expenditure ($ million): x
1	3.4116	1.2
2	4.0771	2.6
3	3.9932	2.2
4	4.4440	3.2
5	4.2771	2.9
6	3.6812	1.5
7	4.1629	2.7

Exercise

The relationship between y and x for the given data is given by $\hat{\boldsymbol{y}} = \boldsymbol{\beta_0 . x^{\beta_1}}$. Find the values of $\beta_0$and β_1.

#	y	x
1	1.20	450
2	9.00	20200
3	4.50	9060
4	3.20	3500
5	13.00	75600
6	0.60	175
7	1.80	800
8	2.70	2100

[Answer: $\hat{y} = (0.055857).\, \mathrm{x}^{0.49606}$]

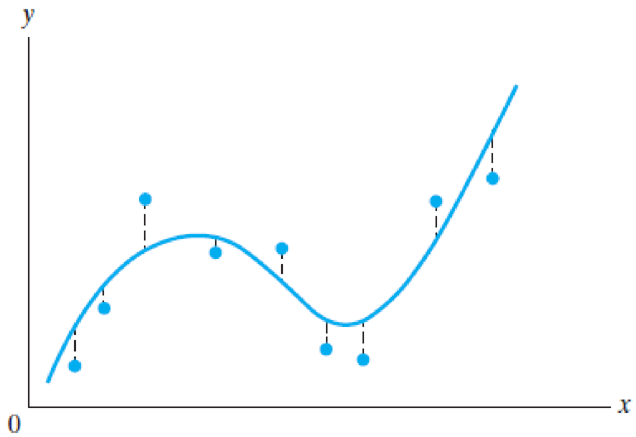
Nonlinear Regression: Polynomial / Curvilinear Model

- If there is no clear indication about the shape of the plot, then a polynomial is attempted through the following regression equation:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

- Using the method of partially differentiating the equation w.r.t. the individual regression coefficients and equating them to 0 to minimize the SSE, the regression coefficients can be calculated using the following matrix equation:

$$\begin{bmatrix} n & \sum x & \sum x^2 & \cdots & \sum x^p \\ \sum x & \sum x^2 & \sum x^3 & \cdots & \sum x^{p+1} \\ \cdots & \cdots & \cdots & & \cdots \\ \sum x^p & \sum x^{p+1} & \sum x^{p+2} & \cdots & \sum x^{2p} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \\ \sum x^2 y \\ \cdots \\ \sum x^p y \end{bmatrix}$$



Example: For the given table, fit a second degree polynomial.

x	y
0	12.00
1	10.50
2	10.00
3	8.00
4	7.00
5	8.00
6	7.50
7	8.50
8	9.00

$$\begin{bmatrix} n & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \\ \sum x^2 y \end{bmatrix}$$

$$\begin{bmatrix} 9 & 36 & 204 \\ 36 & 204 & 1296 \\ 204 & 1296 & 8772 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 80.50 \\ 299.00 \\ 1697.00 \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 9 & 36 & 204 \\ 36 & 204 & 1296 \\ 204 & 1296 & 8772 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 80.50 \\ 299.00 \\ 1697.00 \end{bmatrix} = \begin{bmatrix} 12.184 \\ -1.846 \\ 0.1829 \end{bmatrix}$$

$$\boldsymbol{\hat{y} = 12.184 - 1.846x + 0.1829x^2}$$

How Degree of Polynomial Decided?

- The procedure of identification of degree consists of first fitting a straight line as well as a second-degree polynomial and testing the null hypothesis $\beta_2 = 0$ in $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$.
- That means to test if nothing is gained by including the quadratic term in $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$
- If this null hypothesis can be rejected, then a third-degree polynomial is fitted and the hypothesis $\beta_3 = 0$ is tested.
- It means to test if nothing is gained by including the cubic term in $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
- This procedure is continued until the null hypothesis $\beta_i = 0$ is rejected in two successive steps and that means there is no apparent advantage in carrying the extra terms.

Exercise

The following data shows the y = glucose concentration (g/L) and x = fermentation time (days) for a particular blend of malt liquor. Fit a second degree polynomial.

x	y
1	74
2	54
3	52
4	51
5	52
6	53
7	58
8	71

(Answer: $\hat{y} = 84.482 - 15.875.x + 1.7679.x^2$)

Linear Regression: Gradient Descent Approach

- The objective of Linear Regression is to form a model that learns the training dataset (x, y).
- The model would in the following form, where $\hat{y}$ is the forecasted or predicted value.

$$\hat{y} = b_0 + b_1 x$$

- The Sum of Squared Error (SSE) loss for all records of training dataset is given by:

$$Loss(L) = \Sigma(y - \hat{y}\,)^2$$

- In Gradient Descent, the objective is learn the values of b_0 and b_1 starting from their some initial values such that when the learning is complete the loss is minimum.
- For one record of the training dataset, the gradient of Loss (L) w.r.t. b_1:

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial \hat{y}} . \frac{\partial \hat{y}}{\partial b_1}$$

- Where:

$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial (y - \hat{y}\,)^2}{\partial \hat{y}} = -2(y - \hat{y}) = 2(\hat{y} - y)$$

- And:

$$\frac{\partial \hat{y}}{\partial b_1} = \frac{\partial (b_0 + b_1 x\,)}{\partial b_1} = x$$

- So:

$$\frac{\partial L}{\partial b_1} = 2(\hat{y} - y).x$$

- Similarly:

$$\frac{\partial L}{\partial b_0} = 2(\hat{y} - y)$$

- For the Gradient Descent Learning:

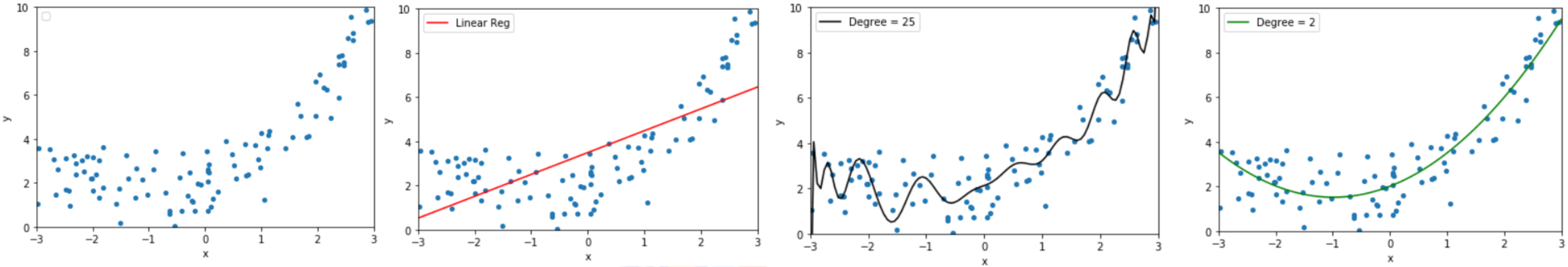
$$b_0 = b_0 - \eta . \sum \frac{\partial L}{\partial b_0}$$

- And also:

$$b_1 = b_1 - \eta . \sum \frac{\partial L}{\partial b_1}$$

- Where, η is the learning rate to control the abrupt movements while coming down to the minimum loss. Its value is kept low (< 1).
- The values of b_0 and $b_1$are updated after each epoch (iteration) until loss has reached to a minimum level or a desired count of epochs are executed.
- The final values of b_0 and b_1 will represent the formed model.
- Since 2 is a constant factor, it can be ignored from the partial derivative terms.
- Gradient Descent can be applied to more complex models, where the concepts of maxima/minima cannot be applied.

Under and Overfitting: General Concept



- The plot shows 100 non-linear data points generated based on a quadratic equation with Gaussian noise.
- Taking the data points as training data, a linear regression model is attempted to fit the dataset. It is obvious that the model is **underfitting** it, as it tries to superficially capture the trend.
- A degree 25 polynomial is attempted to fit the training data. The model is **overfitting** it as it tries to learn as many training data points as possible.
- A degree 2 polynomial (quadratic) is attempted to fit the training data. The model is properly **fitting** the trend because the training data itself is generated using the quadratic equation.
- In the practical situations, the source of the data will not be known, so how the complexity of the model would be decided? How it will be identified if the model is under or overfitting the training data?
- One way is to perform the cross-validation. If the model is performing well on the training data but poor on the validation data, it shows that the model is overfitting. If it performs poor on both, it shows that the model is underfitting.
- In **Machine Learning (ML)**, many times it is desirable to overfit the training data than underfit and then use some approach so that the model can perform better with test data also. One such approach is **regularization**. There are several regularizations techniques like **Lasso (L1)**, **Ridge (L2)** and **Elastic Net** Regularization.

Thank You

Appendix

(The information is provided for reference. Not part of evaluation components)

Regularization: Intuitive Understanding

- Certain data points are given (blue colour).
- Using these data points, the linear regression model will be in the following form (brown colour):

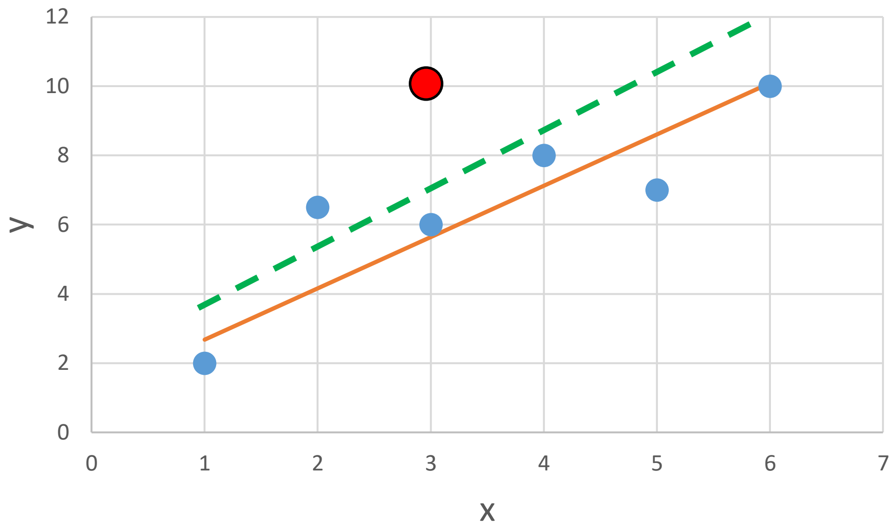
$$\hat{y} = b_0 + b_1 x$$

- If this model is perfect, the loss will be minimum for this model.
- Let us say $Loss(L) = (y - \hat{y})^2$
- Let us say there is one more point in this dataset (red colour) that needs to be considered for the model.
- If this new point is considered and if the model $\hat{y} = b_0 + b_1 x$ is kept as it is the loss will be no more minimum because $(y - \hat{y})^2$ for this new data point will be added in the total loss.
- Here the model formation means adjusting the values b_0 and b_1 so that the loss is minimum.
- If an additional term in the loss function is added like $\alpha. {b_1}^2$or $\alpha. |b_1|$, there is more work to do to make the loss minimum. Here α is a regularization hyperparameter to control the impact.
- Or in other words, b_0 and $\mathrm{b}_1$will be penalized to make the loss minimum. Like for b_1:

$$Loss(L_{reg}) = L + \alpha. {b_1}^2 \quad or$$

$$Loss(L_{reg}) = L + \alpha. |b_1|$$

- The new model formed with this new loss function will be more average in nature (**green dashed line**), probably avoiding many points but still the loss will be minimum for all the data points; including the new point also. We can also say model is avoiding overfitting and becoming more generic.



Regularization: Ridge and Lasso

- Regularization is a technique to reduce the overfitting. It is achieved by constraining the parameters of the model.
- **Tikhonov Regularization** (**L2 Regularization** or **Ridge Regression**) uses the following new loss (L_{reg}) function:

$$L_{reg} = L + \alpha.\frac{1}{2}.\sum_{i=1}^{n} b_i^2$$

- $\frac{1}{2}.\sum_{i=1}^{n} b_i^2$ is half of the square of the L2 norm of the parameters in the model (e.g. b_0, b_1, b_2 etc.)
- Here α (sometime termed as λ also) is a hyperparameter to control the degree of regularization. Increasing value of α avoids overfitting.
- **Example**: Let us say in a model if there are two parameters $b_0\ and\ b_1$, then gradient of loss w.r.t. b_0

$$\frac{\partial L_{reg}}{\partial b_0} = \frac{\partial L}{\partial b_0} + \alpha.\frac{1}{2}.\frac{\partial(b_0^2 + b_1^2)}{\partial b_0} = \frac{\partial L}{\partial b_0} + \alpha.b_0$$

- For the gradient descent through L2 Regularization (similar for b_1):

$$b_0 = b_0 - \eta.(\frac{\partial L}{\partial b_0} + \alpha.b_0)$$

- **L1 Regularization or LASSO Regression** (Least Absolute Shrinkage and Selection Operator Regression) uses the L1 norm of the weights:

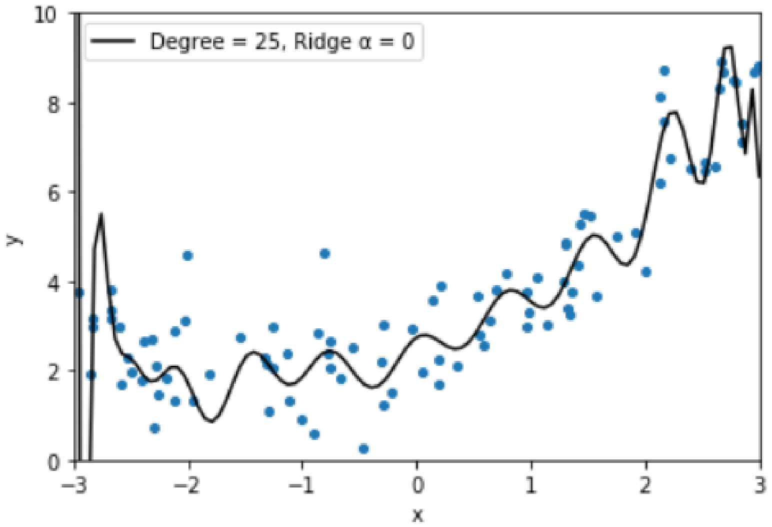
$$L_{reg} = L + \alpha.\sum_{i=1}^{n} |b_i|$$

- For the gradient descent through L1 Regularization:

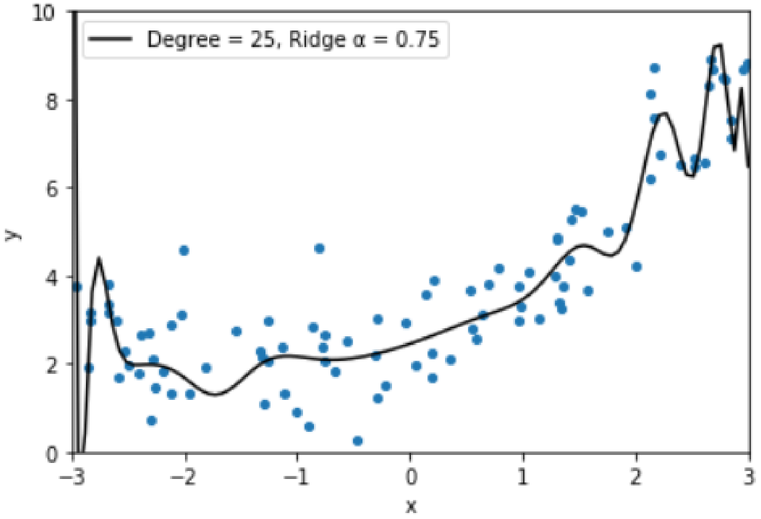
$$b_0 = b_0 - \eta.(\frac{\partial L}{\partial b_0} + \alpha)$$

- There is a Ridge and Lasso combination method which can also be used: ElasticNet Regularization.

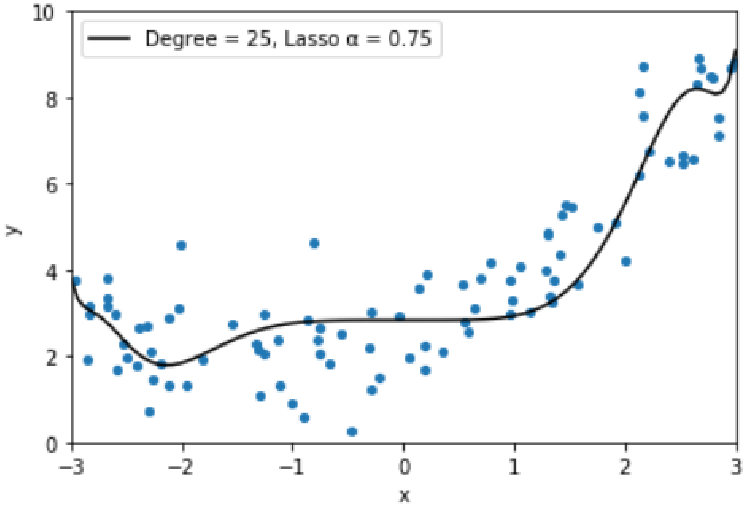
Example: Ridge and LASSO Regression



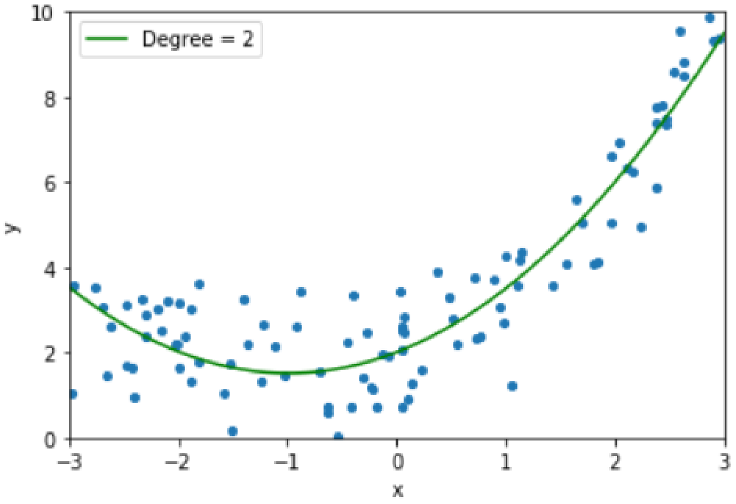
Model with Ridge (L2) Regularization with α = 0



Model with Ridge (L2) Regularization with α = 0.75



Model with LASSO (L1) Regularization with α = 0.75



Best Fit

Regularization helps to obtain a model that is closer to the best fit!