DATA ANALYTICS

Unit 2:Correlation Analysis

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Correlation Analysis

Mamatha H R
Department of Computer Science and Engineering

Introduction to Correlation

- *Correlation is a statistical measure of an association relationship that exists between two random variables.*
- *Correlation is not necessarily a causal relationship.*
- *Correlation is important in analytics since it helps to identify variables that may be used in the model building and also useful for identifying issues such as multi-collinearity that can destabilize regression-based models.*

Introduction to Correlation

- Correlation is a measure of the strength and direction of relationship that exists between two random variables and is measured using correlation coefficient.

- In others words, correlation is a measure of association between two variables. Correlation can assist the data scientists to choose the variables for model building that is used for solving an analytics problem

DATA ANALYTICS

In today's class…

- Correlation between two continuous random variables (ratio or interval scale)
- Correlation between two ordinal variables
- Correlation between a continuous random variable and a dichotomous (binary) random variable
- Correlation between two binary random variables

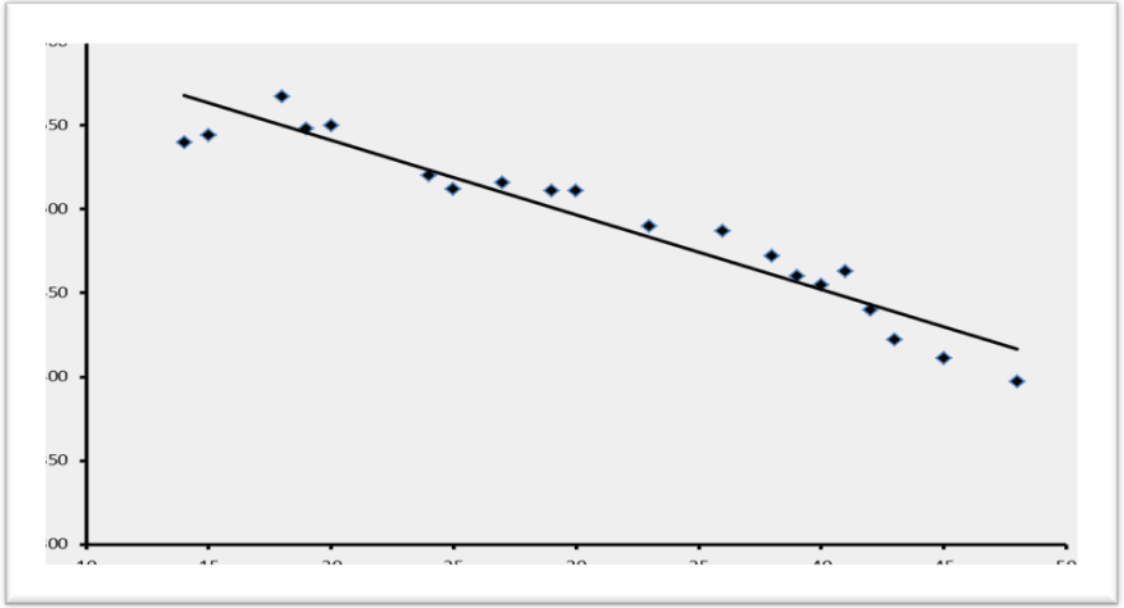
Pearson Correlation Coefficient

Pearson product moment correlation (in short Pearson correlation) is used for measuring the strength and direction of the linear relationship between two continuous random variables *X* and *Y*.

Data on age and average call duration (in seconds)

Age	14	15	18	19	20	24	25	27	29	30
Call Duration	540	544	567	548	550	520	512	516	511	511
Age	33	36	38	39	40	41	42	43	45	48
Call Duration	490	487	472	460	455	463	440	422	411	397

In the figure, we can see that the average call duration (Y) decreases as the age of the customer (X) increases. We can measure the strength of the linear association relationship using a numerical measure called correlation coefficient.



Association relationship between age and average call duration

Calculation of Pearson Product Moment Correlation Coefficient

- Pearson product moment correlation is used when we are interested in finding linear relationship between two continuous random variables (that is, the variable should be either of ratio or interval scale).

- The range of two variables can be different, thus we need to standardize the variables which can be used for measuring the correlation between two variables.

Calculation of Pearson Product Moment Correlation Coefficient

- Let X_i be different values of the variable X and Y_i be different values of Y. Then the standardized values of X and Y are given by

$$Z_X = \left(\frac{X_i - \bar{X}}{\sigma_X} \right) \qquad Z_Y = \left(\frac{Y_i - \bar{Y}}{\sigma_Y} \right)$$

- Where $\bar{X}$ and $\bar{Y}$ are mean values of random variables X and Y; σ_X and σ_Y are the corresponding standard deviations. The Pearson's correlation coefficient is given by

$$r = \frac{\sum_{i=1}^{n} Z_X Z_Y}{n} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right) \times \left(Y_i - \bar{Y} \right)}{n \sigma_X \sigma_Y}$$

Where n is the number of cases in the sample. The formula in above Eq.is also frequently used to account for the degrees of freedom and recommended when the standard deviation is calculated from sample

Calculation of Pearson Product Moment Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\times\left(Y_i - \bar{Y}\right)}{(n-1)S_X S_Y}$$

However, here we will be using the formula in above Eq.For large samples, the correlation coefficients calculated using Eqs. will converge.

Where S_X and S_Y are the standard deviation of random variables X and Y calculate from the sample. We can note the following properties from Eq.

$$r = \frac{\sum_{i=1}^{n} Z_X Z_Y}{n} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\times\left(Y_i - \bar{Y}\right)}{n\sigma_X \sigma_Y}$$

Properties

- Whenever the value of X_i is greater than mean and if the corresponding value of Y_i is also greater than mean, then the numerator in equation will be positive.

- Whenever the value of X_i is lesser than mean and if the corresponding value of Y_i is also lesser than mean, then the numerator in equation will be positive.

- Whenever the value of X_i is lesser than mean (or greater than mean) and the corresponding value of Y_i is greater than mean (or lesser than mean), then the numerator in equation will be negative.

It is possible that we may have combinations of three cases listed above in a data set. Thus the numerator in Eq.is likely to be positive, negative, or zero.

The value of Pearson's correlation coefficient lies between –1 and +1. Equation is mathematically equivalent to below Eqs.

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \times \sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where Cov(*X*,*Y*) is the covariance between random variables *X* and *Y* and is given by

$$\mathrm{Cov}(X, Y) = E\left(\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right)$$

Properties of Pearson Correlation Coefficient

- The value of correlation coefficient lies between –1 and +1. High absolute value of r, $|r|$, indicates strong relationship between the two variables.
- Positive value of r indicates positive correlation (as value of X increases, the value of Y also increases) and negative value of r indicates negative correlation (as the value of X increases, the value of Y decreases).
- The sign of correlation coefficient is same as the sign of covariance between the two random variables.

Properties of Pearson Correlation Coefficient

- Assume that the value of Pearson correlation coefficient between X and Y is r. Let Z_1 and Z_2 be the linear combinations of X and Y ($Z_1 = A + BX$ and $Z_2 = C + DY$). Then the correlation coefficient between Z_1 and Z_2 will be r when the signs of B and D are same (both are positive or negative) and $-r$ when the signs of B and D are opposite.
- Mathematically, square of correlation coefficient is equal to the co-efficient of determination (R^2) of the linear regression model, that is $r^2 = R^2$.
- Pearson correlation coefficient value may be zero even when there is a strong non-linear relationship between variables X and Y (Reed, 1917). Thus low correlation coefficient value cannot be taken as an evidence of no relationship.

Example

The average share prices of two companies over the past 12 months are shown in Table . Calculate the Pearson correlation coefficient.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

The average values are $\bar{X} = 292.9717$ and $\bar{Y} = 229.8292$

The following equation is used for calculating the correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

DATA ANALYTICS

Calculation of correlation coefficient

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
274.58	219.50	–18.39	–10.33	189.97	338.25	106.6917
287.96	242.92	–5.01	13.09	–65.61	25.12	171.3699
290.35	245.90	–2.62	16.07	–42.13	6.87	258.2717
320.07	256.80	27.10	26.97	730.86	734.32	727.4259
317.40	240.60	24.43	10.77	263.11	596.74	116.0109
319.53	245.23	26.56	15.40	409.02	705.35	237.1857
301.52	232.09	8.55	2.26	19.33	73.07	5.111367
271.75	222.65	–21.22	–7.18	152.35	450.36	51.54043
323.65	231.74	30.68	1.91	58.62	941.16	3.651284
259.80	214.43	–33.17	–15.40	510.82	1100.36	237.1343
263.02	201.86	–29.95	–27.97	837.72	897.10	782.2743
286.03	204.23	–6.94	–25.60	177.70	48.19	655.3173
Sum				3241.77	5916.89	3351.98

From Table , we have

$$\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y}) = 3241.77$$

$$\sum_{i=1}^{12}(X_i - \bar{X})^2 = 5916.89$$

Correlation coefficient r = $\sum_{i=1}^{12}(Y_i - \bar{Y})^2 = 3351.98$

$$\frac{3241.77}{\sqrt{5916.89} \times \sqrt{3351.98}} = 0.7279$$

Spurious Correlation

One of the major problem with correlation is the possibility of spurious correlation between two random variables which in many cases is caused due to some other latent variable (hidden variable) that influences both variables for which the correlation is calculated.

Following are few examples of spurious correlation between two random variables:

Crime rate versus ice cream sale: It has been reported that the sale of ice cream and crime rates are positively correlated (Levitt and Dubner, 2009). Obviously, ice cream is not driving the crime rate. In this case the hidden variable is the temperature (summer increasing the ice cream sale) and also increasing crime (people on vacation and locked houses becomes easy target).

Spurious Correlation

Doctors and deaths: Number of doctors is positively correlated with number of deaths in villages, that is, as the number of doctors increases, the deaths also increase. We can be sure that doctors are not causing the deaths to increase (Young, 2001).

Divorce rate in Maine and per capita consumption of margarine: The divorce rate in Maine was highly correlated with per capita consumption of margarine (based on data between 1999 and 2009.The correlation was 0.9926 (Source: tylervigen.com).

Correlation coefficient: points to ponder

1. Does correlation mean two variables are related?

2. If there is no correlation between two variables, can we conclude they are not related?

Hypothesis Test for Correlation Coefficient

For any two sets of data the Pearson correlation coefficient is most likely to give a value other than zero. Many thumb rules exist to group the correlation value as no correlation, low correlation, medium correlation, and high correlation (Monroe and Stuit, 1933).

Let ρ be the population correlation coefficient. The null and alternative hypotheses are given by

H_0:	**$\rho = 0$ (there is no correlation between two random variables)**
H_A:	$\rho \neq 0$ (there is a correlation between two random variables)

- The sampling distribution of correlation coefficient *r* follows an approximate *t*-distribution with (*n* – 2) degrees of freedom (*df*) where *n* is the number of cases in the sample for calculating the correlation coefficient.
- Two degrees of freedom are lost since we estimate two mean values from the data. The mean of the sampling distribution is ρ and the corresponding standard deviation is (Ezekiel, 1941) $\sqrt{\frac{1-r^2}{n-2}}$
- The *t*-statistic for null hypothesis is given by $t_{\alpha,n-2} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$
- When the null hypothesis is $\rho = 0$, the test statistic in above Eq. becomes $t_{\alpha,n-2} = r\sqrt{\frac{n-2}{1-r^2}}$

https://archive.org/details/in.ernet.dli.2015.137861/page/n51/mode/2up (Page 487)

DATA ANALYTICS

Example

The average share prices of two companies over the past 12 months are shown in Table. conduct the following two hypothesis tests at α = 0.05:

(a) The correlation between share prices of two companies is zero.

(b) The correlation between share prices of two companies is at least 0.5.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

PES UNIVERSITY ONLINE

Solution

(a) The null and alternative hypotheses are:

H_0: $\rho = 0$

H_A: $\rho \neq 0$

The corresponding *t*-statistic is $t = r\sqrt{\frac{n-2}{1-r^2}} = 0.7279\sqrt{\frac{12-2}{1-0.7279^2}} = 3.3569$

Note that this is a two-tailed test and the critical *t*-value at α = 0.05 and *df* = 10 is 2.2281

Since the calculated *t*-statistic is higher than the critical *t*-value, we reject the null hypothesis and conclude that there is a significant correlation between share prices of two companies.

The corresponding *p*-value is 0.0072

Solution

(b) The null and alternative hypotheses are given by

$H_0: \rho \leq 0.5$

$H_A: \rho > 0.5$

The corresponding *t*-statistic is $t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7279-0.5}{0.2168} = 1.05$

This is a right-tailed test and the corresponding *t*-critical value is 1.8124

The calculated *t*-value is less than the critical value of *t*, and thus we retain the null hypothesis and conclude that the correlation between share prices of two companies is less than 0.5.

The corresponding *p*-value is 0.1592 .

Spearman Rank Correlation

- Pearson correlation is appropriate when the random variables involved are both from either ratio scale or interval scale.
- When both random variables are of ordinal scale, we use Spearman rank correlation (also known as Spearman's rho denoted by ρ_s).
- The Spearman rank correlation, r_s, estimated from a sample is given by (Yule and Kendall 1937, Woodbury, 1940)

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}$$

where D_i = difference in the rank of case i under variables X and Y (that is $X_i - Y_i$).

- The sampling distribution of Spearman correlation r_s also follows an approximate t-distribution with mean ρ_s and standard deviation with $n - 2$ degrees of freedom

$$\sqrt{\frac{1 - r_s^2}{n - 2}}$$

Example

Ranking of 12 countries under corruption and Gini Index (wealth discrimination) are shown in Table. Calculate the Spearman correlation and test the hypothesis that the correlation is at least 0.2 at $\alpha = 0.02$.

Countries	1	2	3	4	5	6	7	8	9	10	11	12
Corruption	1	4	12	2	5	8	11	7	10	3	6	9
Gini Index	2	3	9	5	4	6	10	7	8	1	11	12

Solution

The Spearman rank correlation calculations are shown in Table

Country	Corruption Rank (X_i)	Gini Index (Y_i)	$D = X_i - Y_i$	D^2
1	1	2	–1	1
2	4	3	1	1
3	12	9	3	9
4	2	5	–3	9
5	5	4	1	1
6	8	6	2	4
7	11	10	1	1
8	7	7	0	0
9	10	8	2	4
10	3	1	2	4
11	6	11	–5	25
12	9	12	–3	9
			$\sum_{i=1}^{12} D_i^2$	68

Solution

The Spearman rank correlation is given by

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6 \times 68}{12(12^2-1)} = 0.7622$$

The null and alternative hypotheses are

H_0: $\rho_s < 0.2$

H_A: $\rho_s \geq 0.2$

The corresponding *t*-statistic is

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1-r_s^2}{n-2}}} = \frac{0.7622 - 0.2}{0.2046} = 2.74$$

The one-tailed *t*-critical value for $\alpha = 0.02$ and $df = 10$ is 2.35.

Since the calculated *t*-statistic value is more than the *t*-critical value, we reject the null hypothesis and conclude that Spearman rank correlation between two countries is at least 0.2.

Point Bi-Serial Correlation

Point bi-serial correlation is used when we are interested in finding correlation between a continuous random variable and a dichotomous (binary) random variable.

Point Bi-Serial Correlation

- Assume that the random variable X is a continuous random variable and Y is a dichotomous random variable. Then the following steps are used for calculating the correlation between these two variables:
  1. Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1. Then we group the data into two subsets such that in one group the value of Y is 0 and in another group the value of Y is 1
  2. Calculate the mean values of two groups: Let $\bar{x}_0$ and $\bar{x}_1$ be the mean values of groups with $Y = 0$ and $Y = 1$, respectively.
  3. Let n_0 and n_1 be the number of cases in a group with $Y = 0$ and $Y = 1$, respectively, and S_X be the standard deviation of the random variable X.

Point Bi-Serial Correlation

The point bi-serial correlation is given by (Pearson, 1909 and Soper, 1914)

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_X} \sqrt{\frac{n_0 n_1}{n(n-1)}}$$

where n is the total number of cases in the sample and S_X is the standard deviation of X estimated from sample and is given by

$$S_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}$$

Example

Ms Sandra Ruth, data scientist at Airmobile, is interested in finding the correlation between the average call duration and gender. Table provides the average call duration (measured in seconds) and gender of 30 customers of Airmobile. In Table, male is coded as 0 and Female is coded as 1. Calculate the point bi-serial correlation.

Gender	1	1	0	0	0	1	0	1	1	0
Call Duration	448	335	210	382	407	231	359	287	288	347
Gender	1	1	1	1	1	0	0	1	0	0
Call Duration	408	382	303	201	447	439	383	277	279	213
Gender	1	1	0	1	1	0	1	0	1	0
Call Duration	383	355	362	401	331	421	367	437	326	351

Solution

From the data, we can calculate the following values:

$$\bar{X} = 345.33,\ \bar{X}_0 = 353.07,\ \bar{X}_1 = 339.4118,\ S_X = 71.7189,\ n_0 = 13,\ n_1 = 17$$

Bi-serial correlation is given by

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_X}\sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{339.4118 - 353.07}{71.7189}\sqrt{\frac{13\times 17}{30(29)}} = -0.0960$$

There is very low negative correlation between gender and call duration.

The Phi-Coefficient

Karl Pearson recommended the use of the Phi-coefficient when both variables are binary for calculating the association relationship (Cramer, 1946). Let *X* and *Y* be two random variables both taking binary values (that is, *X* takes values 0 or 1 and similarly *Y* also takes values either 0 or 1). One can create a contingency table as shown in Table below.

	Y = 0	Y = 1	Total
X = 0	N_{00}	N_{01}	$N_{X0} = N_{00} + N_{01}$
X = 1	N_{10}	N_{11}	$N_{X1} = N_{10} + N_{11}$
Total	$N_{Y0} = N_{00} + N_{10}$	$N_{Y1} = N_{01} + N_{11}$	

Also, for contingency tables for presence or absence of two categorical variables

	Drink Coffee	Don't drink Coffee
Drink Tea	N_{11}	N_{10}
Don't drink Tea	N_{01}	N_{00}

The Phi-Coefficient

In the contingency table (Table in previous slide)

- N_{00} = Number of cases in the sample such that $X = 0$ and $Y = 0$
- N_{01} = Number of cases in the sample such that $X = 0$ and $Y = 1$
- N_{10} = Number of cases in the sample such that $X = 1$ and $Y = 0$
- N_{11} = Number of cases in the sample such that $X = 1$ and $Y = 1$
- N_{X0} = Number of cases in the sample such that $X = 0$
- N_{X1} = Number of cases in the sample such that $X = 1$
- N_{Y0} = Number of cases in the sample such that $Y = 0$
- N_{Y1} = Number of cases in the sample such that $Y = 1$
- The Phi-coefficient is given by

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}}$$

Example

Joy Finance (JF) is a company that provides gold loans (in which gold is used as guarantee against the loan). Mr Georgekutty, Managing Director of JF, collected data to understand the relationship between loan default status (variable *Y*) and the marital status of the customer (variable *X*). Data is collected on past 40 loans and is shown in Table 8.8. Calculate the Phi-coefficient. In Table , *Y* = 0 implies non-defaulter, *Y* = 1 is defaulter, *X* = 0 is single, and *X* = 1 is married.

X	1	0	1	0	0	0	0	0	1	0
Y	0	1	0	1	0	0	0	1	1	1
X	0	1	1	0	0	1	0	0	0	1
Y	0	1	1	1	0	0	1	1	0	0
X	1	0	0	0	1	1	1	0	0	1
Y	0	0	0	1	0	0	0	0	0	0
X	1	0	0	0	1	0	1	0	1	1
Y	1	0	0	1	1	0	1	1	0	1

Solution

The contingency table for the data shown in above Table is given in below Table

		Y		Total
		0	1	
X	0	13	10	23
	1	10	7	17
Total		23	17	40

From Table , we have N00 = 13, N01 = 10, N10 =10, N11 = 7, NX0 = 23, NX1 = 17, NY0 = 23, and NY1 = 17

The Phi-coefficient is given by

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}} = \frac{7\times13 - 10\times10}{\sqrt{23\times17\times23\times17}} = -0.0230$$

Since the Phi-coefficient is very small, we can conclude that there is not much association between the marital status and loan default.

Summary

(a) Correlation is a measure of strength and direction of linear relationship between two random variables. It can be used only when the relationship is linear.

(b) Correlation captures only association relation and not a causal relation.

(c) Pearson product moment correlation is used when two random variables are continuous. In the case of two ordinal variables the appropriate correlation is Spearman rank correlation.

Summary

(d) Bi-serial correlation is used when correlation is calculated between one continuous and one binary random variable. Phi-coefficient is used for calculating correlation between two binary random variables.

(e) Correlation is an important measure and can be used for feature selection while building regression models.

(f) One of the drawbacks of correlation is the spurious correlations, it is possible that two variables with no explainable relationship may have high correlation coefficient.

Exercise

- Mention and explain the different correlation coefficients.
- For each of the correlation coefficient, find out an application and explore how it is used in that application.

The journey so far…

- Unit 1: Exploratory Data Analysis + Visualization
  - Data acquisition and framing questions given data
  - Taking stock of data: missing values or NA, outliers, anomalies (incorrect data, inconsistent data, incomplete data, noisy data, etc.)
  - Data cleaning and pre-processing
  - Data integration (removal of redundany)
    - Dimensionality reduction (wavelets, PCA, feature subset selection, feature creation)
    - Data reduction (sampling, binning/ histograms, discretization, clustering, etc.)
  - Data transformation (z-score, range normalization, unit norm, etc.)
- Unit 2: Correlation analysis
- Coming up next week – Unit 2: Regression

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

DATA ANALYTICS

Unit 2:Introduction to Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Introduction to Regression

Mamatha H R

Department of Computer Science and Engineering

Causation

If there is a significant linear correlation between two variables, then one of five situations can be true.

- There is a direct cause and effect relationship
- There is a reverse cause and effect relationship
- The relationship may be caused by a third variable
- The relationship may be caused by complex interactions of several variables
- The relationship may be coincidental

What is Regression?

- Regression is a tool for finding existence of an association relationship between a dependent variable (Y) and one or more independent variables (X_1, X_2, …, X_n) in a study.
- The relationship can be linear or non-linear.
- A dependent variable (response variable) "measures an outcome of a study (also called outcome variable)".
- An independent variable (**explanatory variable**) "explains changes in a response variable".
- Regression often set values of explanatory variable to see how it affects response variable (predict response variable)

Regression - Definition

A statistical technique that attempts to determine the existence of a possible relationship between one **dependent variable** (usually denoted by Y) and a collection of Independent variables.

Regression is used for generating a new hypothesis and

for validating a hypothesis

Regression

Regression is a **supervised learning algorithm** under Machine Learning terminology
An important tool in **Predictive Analytics**

Terms dependent and independent do not necessarily imply a causal relationship between two variables.

Regression is not designed to capture causality.

Purpose of regression is to predict the value of dependent variable given the value(s) independent variable(s)

DATA ANALYTICS

Regression Nomenclature

Dependent Variable	Independent Variable
Explained Variable	Explanatory variable
Regressand	Regressor
Predictand	Predictor
Endogenous Variable	Exogenous Variable
Controlled Variable	Control Variable
Target Variable	Stimulus Variable
Response Variable	
Feature	Outcome Variable

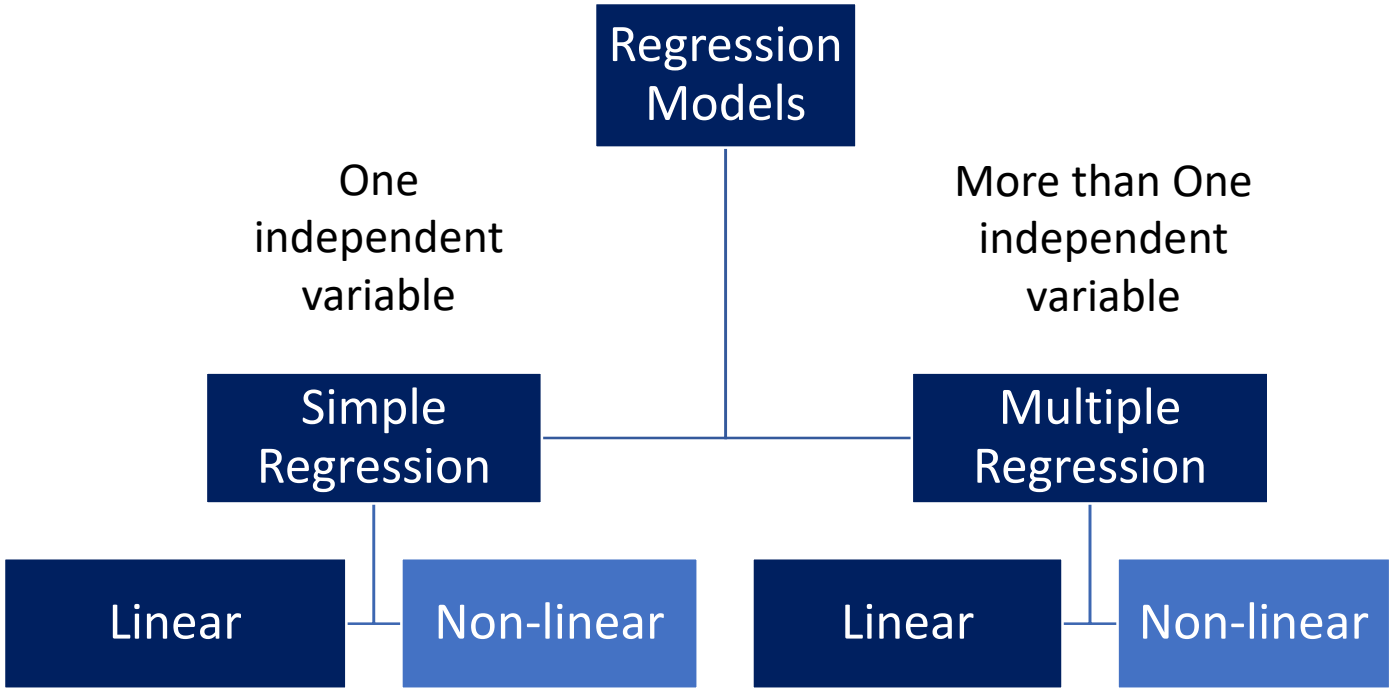
DATA ANALYTICS

Where is it used?

- ✓ Every functional area of management uses regression.
- ✓ Finance: CAPM, Non-performing assets, probability of default, Chance of bankruptcy, credit risk.
- ✓ Marketing: Sales, market share, customer satisfaction, customer churn, customer retention, customer life time value.
- ✓ Operations: Inventory, productivity, efficiency.
- ✓ HR – Job satisfaction, attrition.

DATA ANALYTICS

Types of Regression



Types of Regression

- Simple linear regression – refers to a regression model between two variables.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

- Multiple linear regression – refers to a regression model on more than one independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- Nonlinear regression.

$$Y = \beta_0 + \frac{1}{\beta_1 + \beta_2 X_1} + X_2^{\beta_3} + \varepsilon$$

DATA ANALYTICS

Unit 2:Linear Regression

Mamatha H R, Gowri Srinivasa
Department of Computer Science and Engineering

Linear Regression

- Linear regression stands for a linear relationship between the dependent variable and regression coefficients.

- The following equation will be treated as linear as far as regression is concerned.

$$Y = \beta_1 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_2^2$$

Simple Linear Regression Model Building

A simple linear regression model is developed to understand how the value of a KPI is associated with changes in the values of an independent variable.

Some examples are as follows:

1. A hospital may be interested in finding how the total treatment cost of a patient varies with the body weight of the patient.
2. E-commerce companies such as Amazon, Bigbasket and Flipkart would like to understand the number of customer visits to their portal and the revenue.
3. Retailers such as Walmart, Target, Reliance Retail, Hyper City, etc. would be interested in understanding the impact of price cut promotions on the revenue of their private labels (store brands or house brands).
4. Original equipment manufacturers (OEMs) would like to know the impact of duration of warranty on the profit.

Assumptions

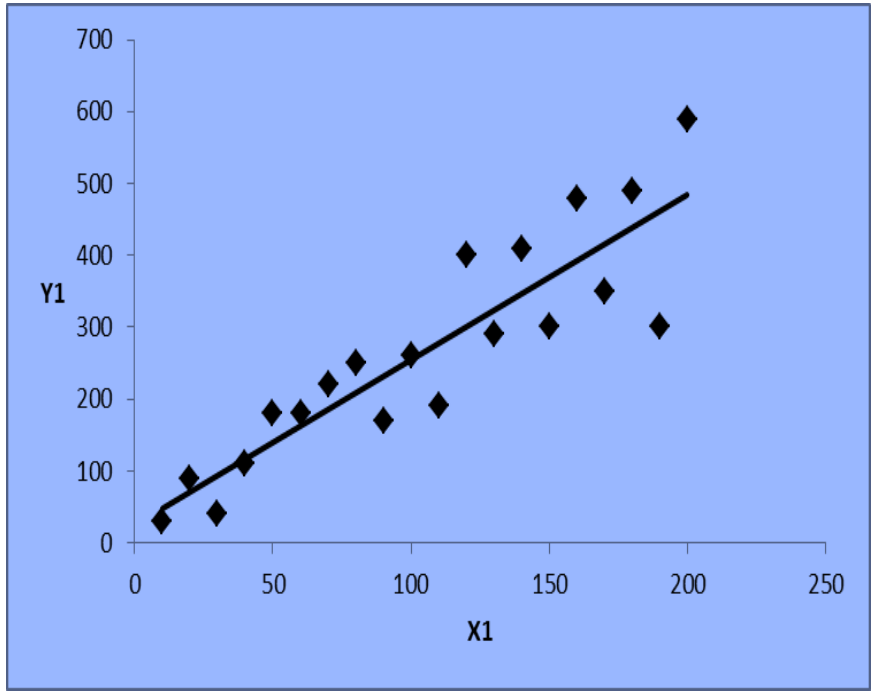
The method of least squares gives the best equation under the assumptions stated below (Harter 1974, 1975):

- The regression model is linear in regression parameters.
- The explanatory variable, *X*, is assumed to be non-stochastic (i.e., *X* is deterministic).
- The conditional expected value of the residuals, $E(\varepsilon_i / X_i)$, is zero.
- In case of time series data, residuals are uncorrelated, that is, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
- The residuals, ε_i, follow a normal distribution.
- The variance of the residuals, Var(εi|Xi), is constant for all values of Xi. When the variance of the residuals is constant for different values of Xi, it is called homoscedasticity. A non-constant variance of residuals is called heteroscedasticity

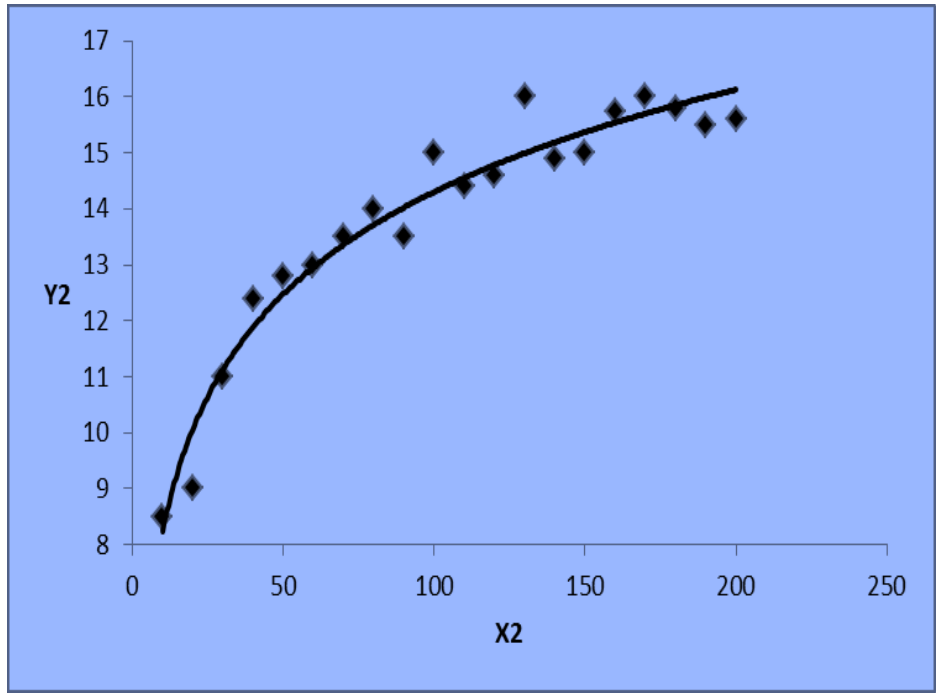
Define the Functional Form of Relationship

For better predictive ability (model accuracy) it is important to specify the correct functional form between the dependent variable and the independent variable. Scatter plots may assist the modeller to define the right functional form.

Linear relationship between X_1 and Y_1



Log-linear relationship between X_2 and Y_2.



Linear-Log

Log-Log

Estimate the Regression Parameters

Once the functional form is specified, the next step is to estimate the regression parameters. The method of **Ordinary Least Squares (OLS)** is used to estimate the regression parameters.

OLS fits regression line through a set of data points such that the sum of the squared distances between the actual observations in the sample and the regression line is minimized (i.e., SSE is minimized).

OLS provides the Best Linear Unbiased Estimate (BLUE).

That is, as sample size $\rightarrow \infty$, the coefficient estimates converge to the population parameter(s).

https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/#:~:text=The%20Gauss%2DMarkov%20theorem%20states%20that%20if%20your%20linear%20regression,of%20all%20possible%20linear%20estimators.

Estimation of Parameters using Ordinary Least Squares

Given a set of dependent variable values (Y_i) and the corresponding independent variable values (X_i), each subject to a random error (ε_i), one has to find the best equation to represent the relationship between the dependent and independent variables.

OLS Estimation

In ordinary least squares, the objective is find the optimal values of β_0 and β_1 that will minimize the **Sum of Squared Errors** (SSE) given in below Eq:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To find the optimal values of β_0 and β_1 that will minimize SSE, we have to equate the partial derivative of SSE with respect to β_0 and β_1 to zero.

$$\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n} -2(Y_i - \beta_0 - \beta_1 X_i) = 2\left(n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} -2X_i(Y_i - \beta_0 - \beta_1 X_i) = -2\sum_{i=1}^{n}\left(X_i Y_i - \beta_0 X_i - \beta_1 X_i^2\right) = 0$$

OLS Estimation

Solving the system of equations for β0 and β1, we get the estimated values as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i y_i - x_i \bar{Y}\right)}{\sum_{i=1}^{n}\left(x_i^2 - x_i \bar{x}\right)} = \frac{\sum_{i=1}^{n} X_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i (X_i - \bar{X})}$$

Perform Regression Model Diagnostics

Regression is often misused since many times the modeller fails to perform necessary diagnostics tests before applying the model.

Before it can be applied it is necessary that the model created is validated for all model assumptions including the definition of the function form.

If the model assumptions are violated, then the modeller has to use some remedial measure; it is also possible that there is no association relationship between the variables at all.

Validate the Model using the Validation Data Set

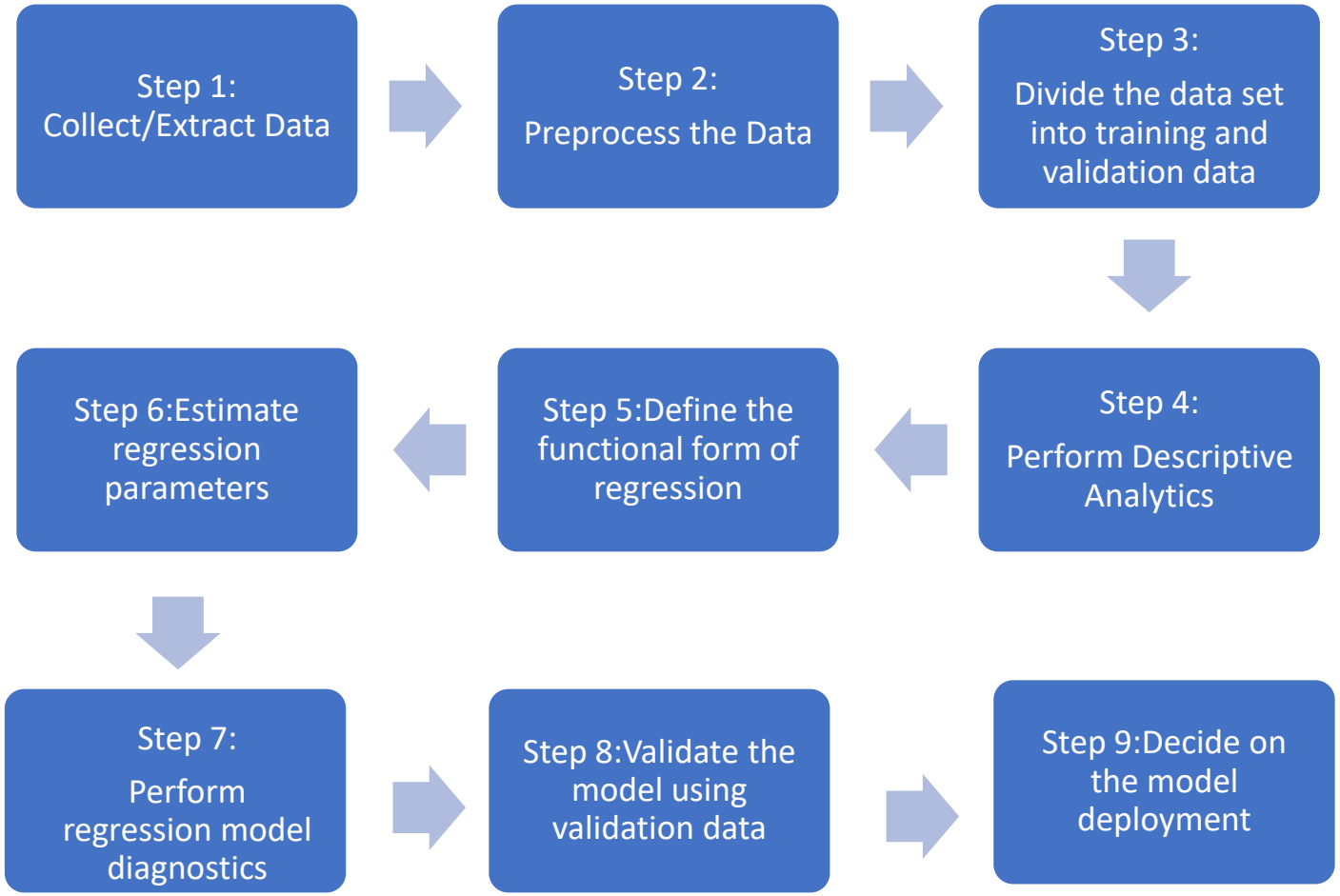
A major concern in analytics is over-fitting, that is, the model may perform very well in the training data set but may perform badly in validation data set. It is important to ensure that the model performance is consistent in the validation data set as was in the training data set. In fact, the model may be cross-validated using multiple training and test data sets.

Decide on the Model Deployment

The final step in the regression model is to generate actionable items and business rules that can be used by the organization

DATA ANALYTICS

Framework for SLR model development



Example :

Salary of Graduating MBA Students versus Their Percentage Marks in Grade 10

Table in next slide provides the salary of 50 graduating MBA students of a Business School in 2016 and their corresponding percentage marks in grade 10 . Develop a linear regression model by estimating the model parameters.

DATA ANALYTICS

Salary of MBA students versus their grade 10 marks

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
1	62	270000	26	64.6	250000
2	76.33	200000	27	50	180000
3	72	240000	28	74	218000
4	60	250000	29	58	360000
5	61	180000	30	67	150000
6	55	300000	31	75	250000
7	70	260000	32	60	200000
8	68	235000	33	55	300000
9	82.8	425000	34	78	330000
10	59	240000	35	50.08	265000
11	58	250000	36	56	340000
12	60	180000	37	68	177600
13	66	428000	38	52	236000
14	83	450000	39	54	265000
15	68	300000	40	52	200000
16	37.33	240000	41	76	393000
17	79	252000	42	64.8	360000
18	68.4	280000	43	74.4	300000
19	70	231000	44	74.5	250000
20	59	224000	45	73.5	360000
21	63	120000	46	57.58	180000
22	50	260000	47	68	180000
23	69	300000	48	69	270000
24	52	120000	49	66	240000
25	49	120000	50	60.8	300000

Solution

Using Eqs., the estimated values of β_0 and β_1 are given by

$$\hat{\beta}_0 = 61555.3553 \text{ and } \hat{\beta}_1 = 3076.1774$$

The corresponding regression equation is given by

$$\hat{Y}_i = 61555.3553 + 3076.1774 X_i$$

Where $\hat{Y}_i$ is the predicted value of Y for a given value of X_i .

The equation can be interpreted as follows: for every one percentage increase in grade 10 marks, the salary of the MBA students will increase at the rate of 3076.1774 on an average. The notations

Interpretation of Simple Linear Regression Coefficients

- Interpretation of regression coefficients is important for understanding the relationship between the response variable and the explanatory variable and the impact of change in the values of explanatory variables on the response variable.

- The interpretation will depend on the functional form of the relationship between the response and the explanatory variables.

- Interpretation of β_0 and β_1 in $Y = \beta_0 + \beta_1 X$

When the functional form is $Y = \beta_0 + \beta_1 X$, the value of $\beta_0 = E(Y/X=0)$.

$\beta_1 = \frac{\partial Y}{\partial X}$ that is β_1 is the change in the value of Y for the unit change in the value of X, where $\frac{\partial Y}{\partial X}$ is the partial derivative of Y with respect to X.

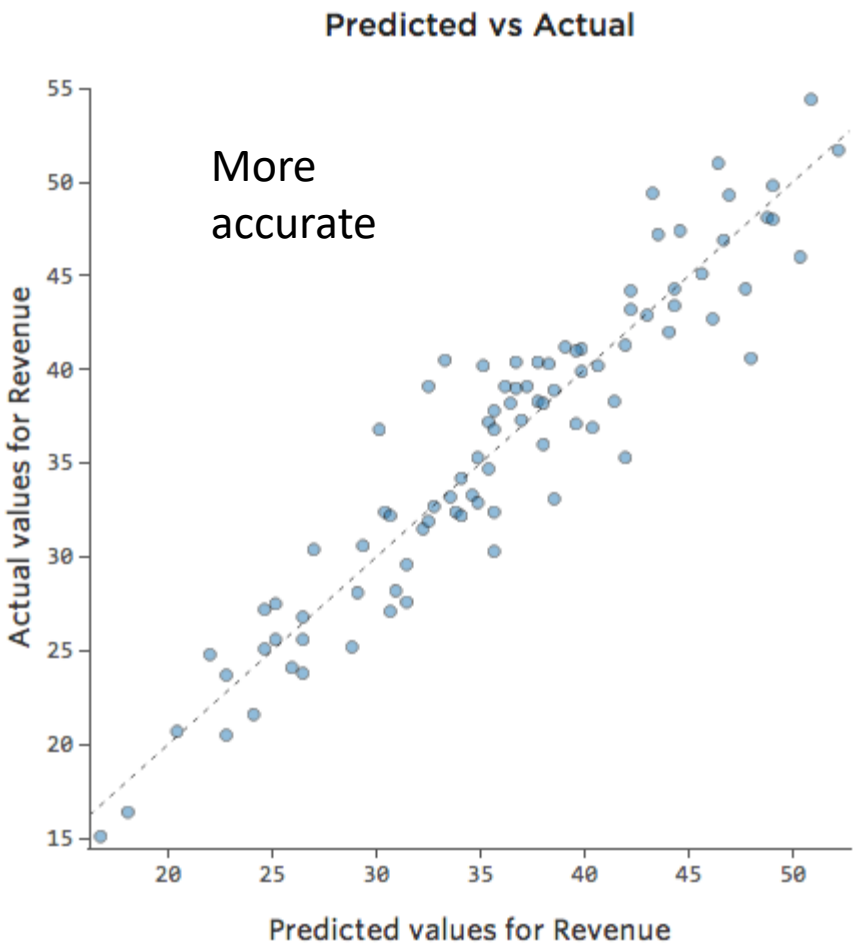
Validation of the Simple Linear Regression Model

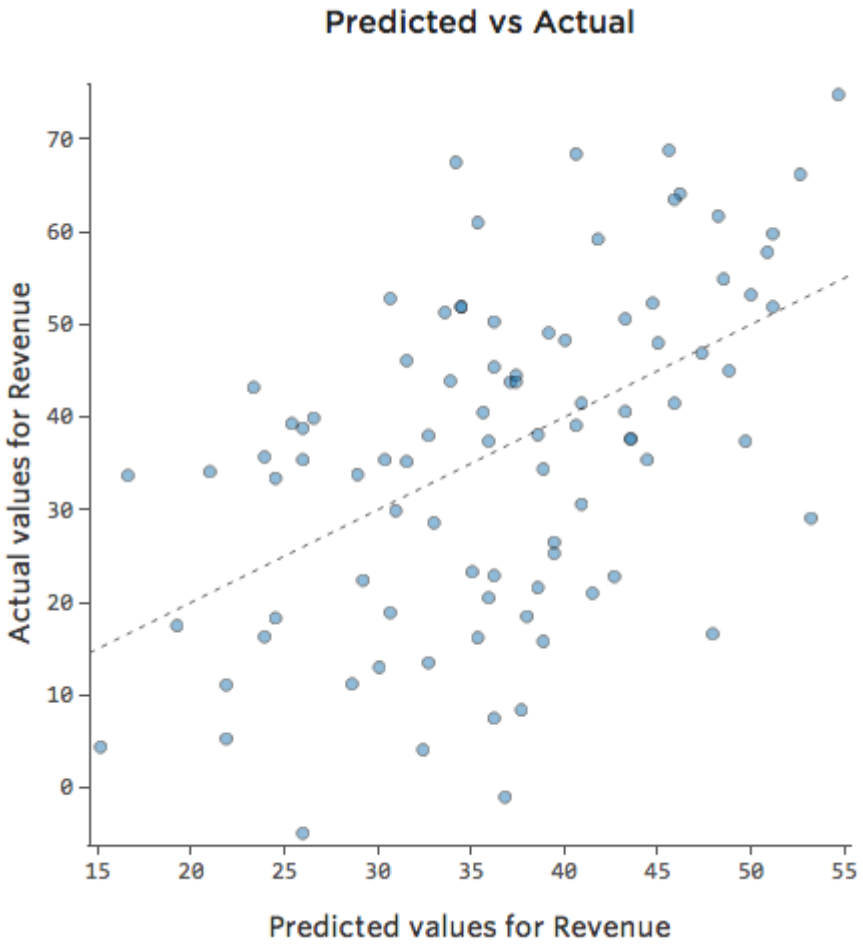
It is important to validate the regression model to ensure its validity and goodness of fit before it can be used for practical applications. The following measures are used to validate the simple linear regression models:

- Prediction accuracy
- Residual analysis to validate the regression model assumptions.
- Co-efficient of determination (*R*-square).
- Hypothesis test for the regression coefficient
- Analysis of Variance for overall model validity (relevant more for multiple linear regression).
- Outlier analysis.

The above measures and tests are essential, but not exhaustive.

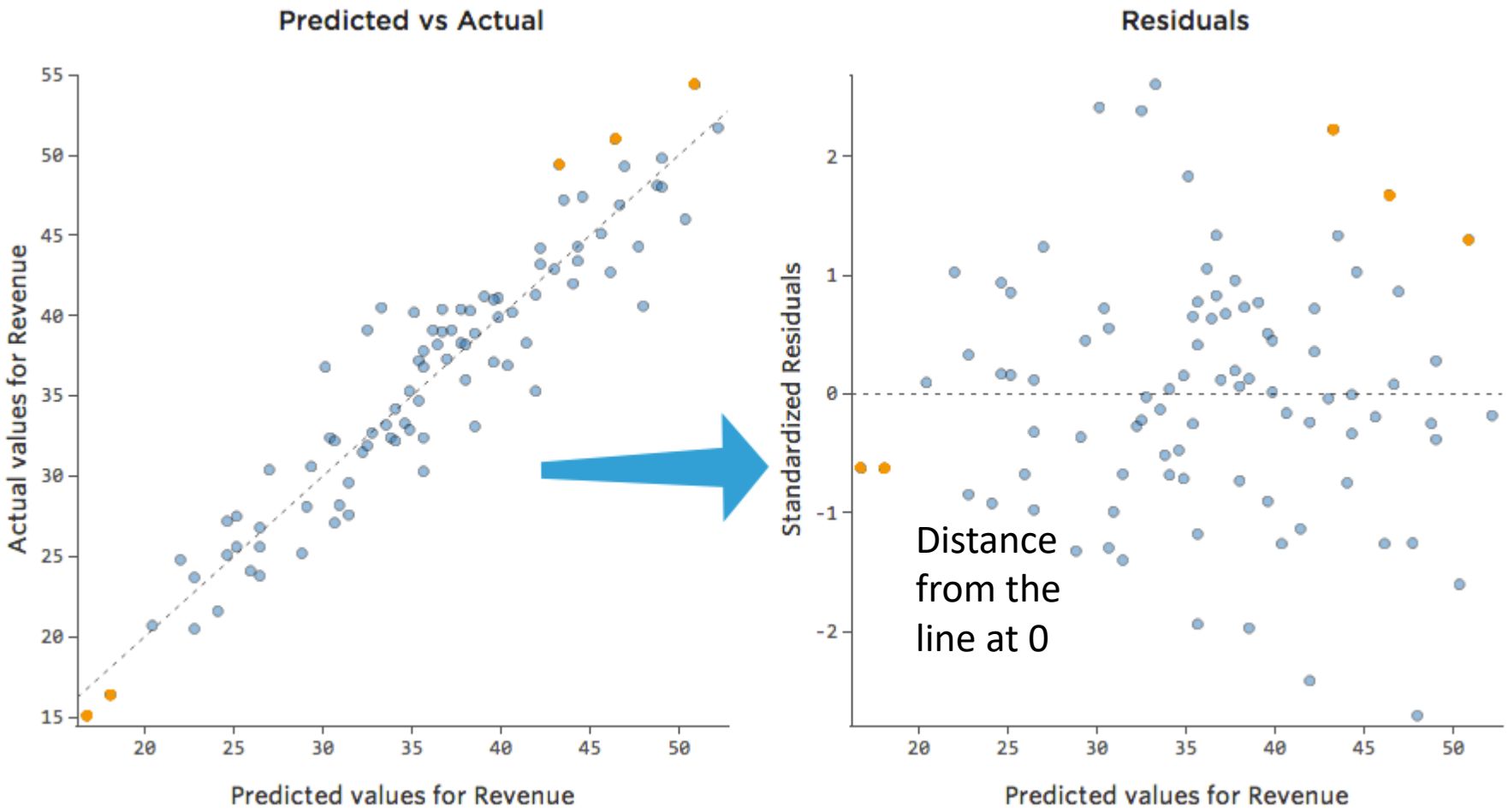
Evaluating Regression Models





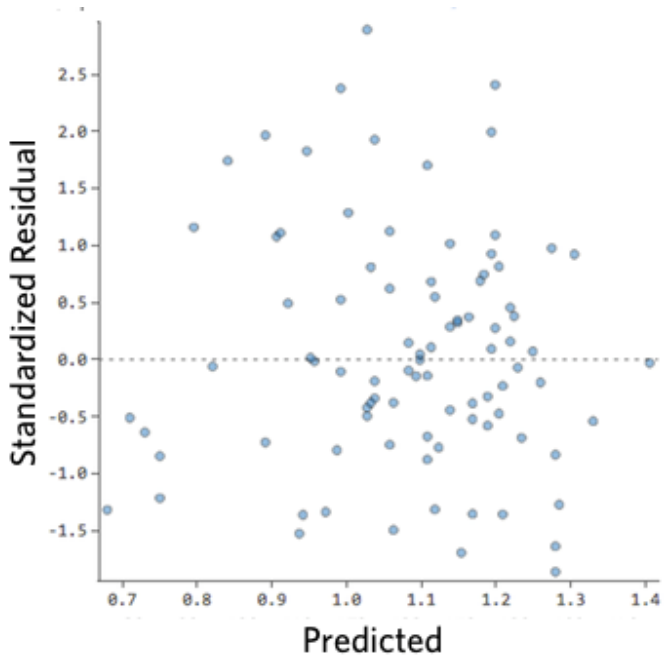
http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

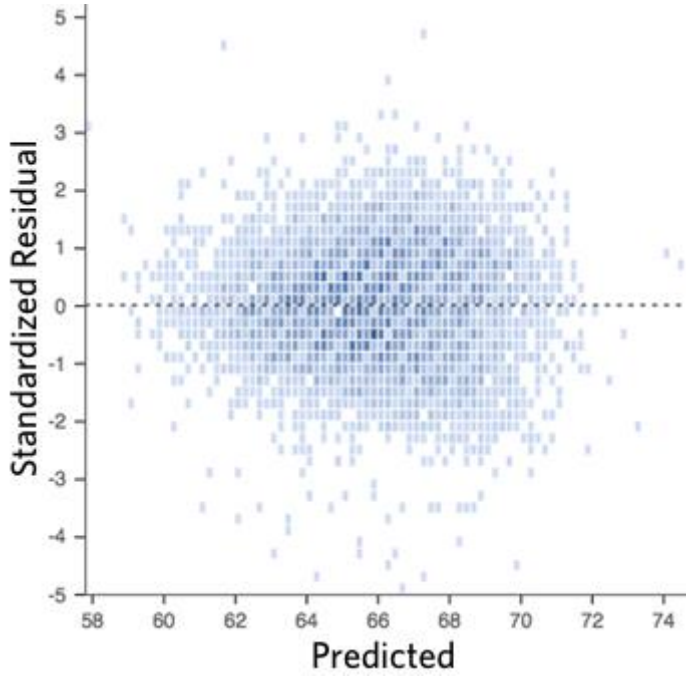
How bad is the prediction?
Study the 'Residuals' (residual plot)



http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

Residual plots that are ‘good’





Randomly distributed → tending to cluster in the center
Distribution is random (i.e., there is no ‘pattern’)
Clustered around the lower single digits of the vertical axis

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

Test of Homoscedasticity

An important assumption of regression model is that the residuals have constant variance (homoscedasticity) across different values of the explanatory variable (X).

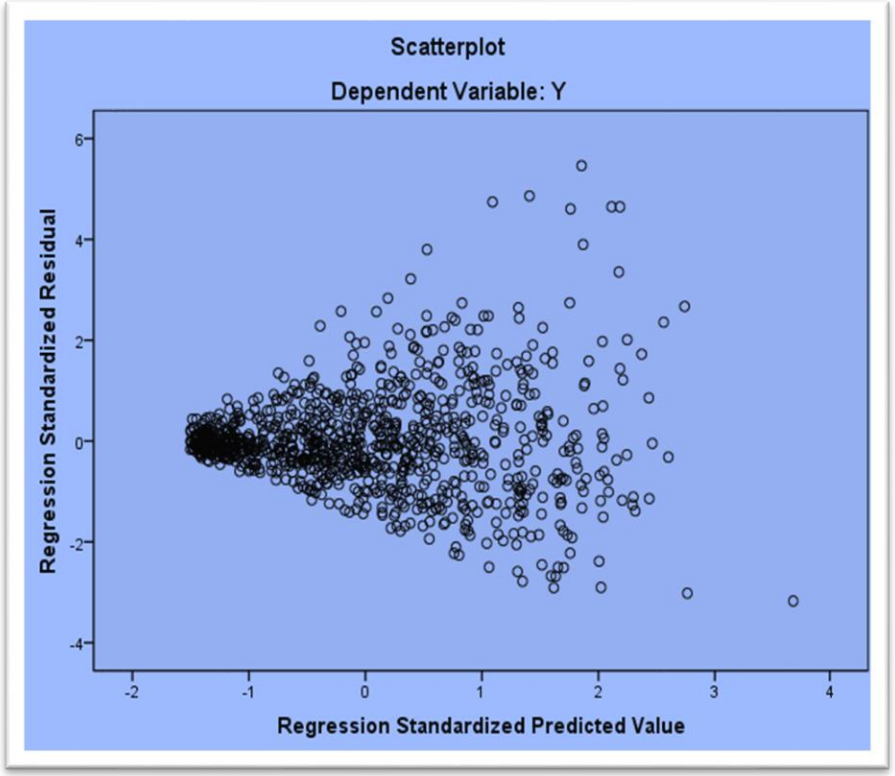
That is, the variance of residuals is assumed to be independent of variable X. Failure to meet this assumption will result in unreliability of the hypothesis tests.

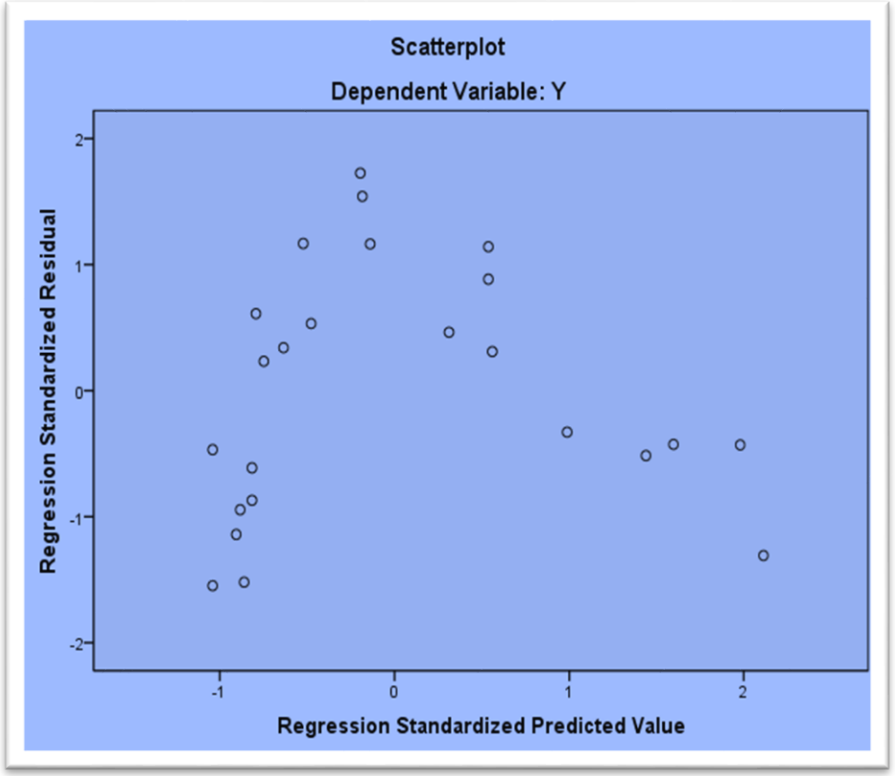
Testing the Functional Form of Regression Model

Any pattern in the residual plot would indicate incorrect specification (misspecification) of the model.

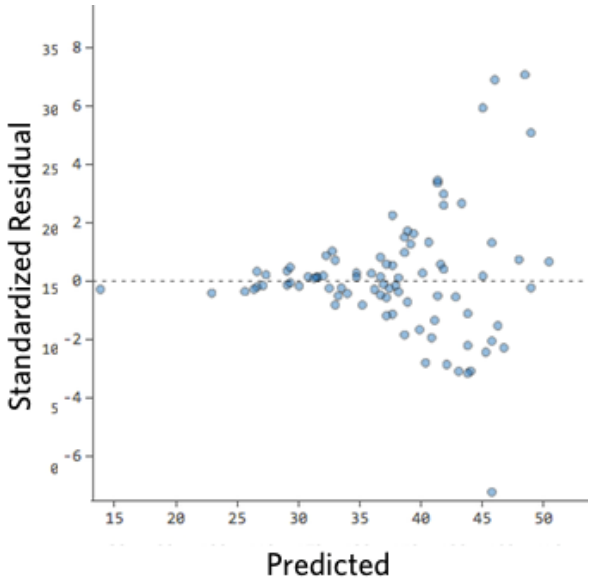
Testing the Functional Form of Regression Model

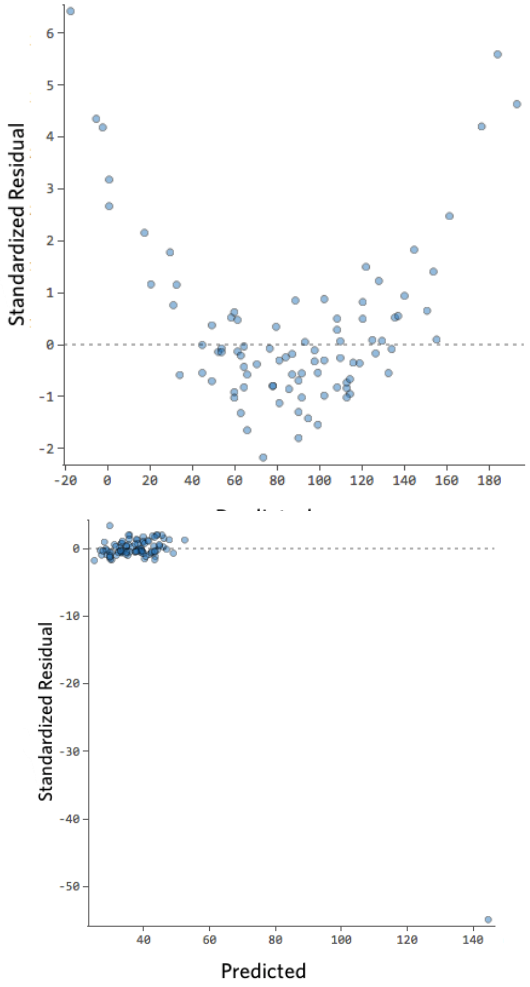
Any pattern in the residual plot would indicate incorrect specification (misspecification) of the model.

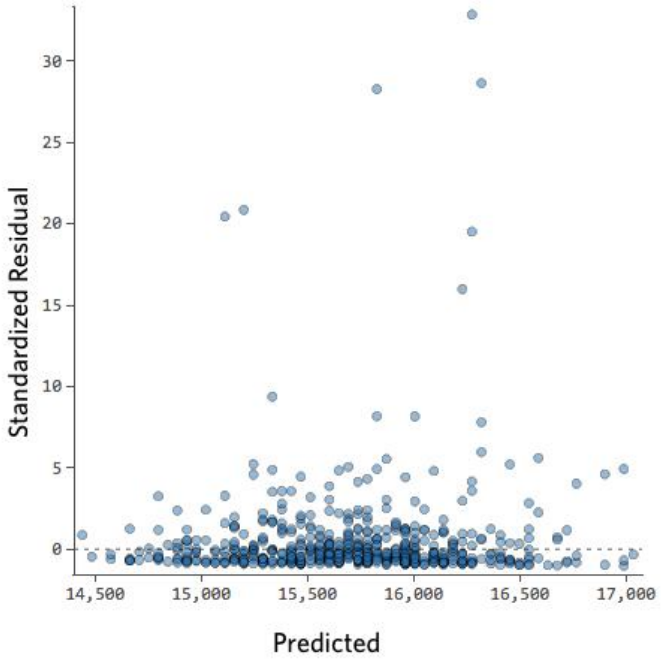




Problems with residual plots







These plots aren't evenly distributed vertically, or they have an outlier, or they have a clear shape to them.
If you can detect a clear pattern or trend in your residuals, then your model has room for improvement.

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

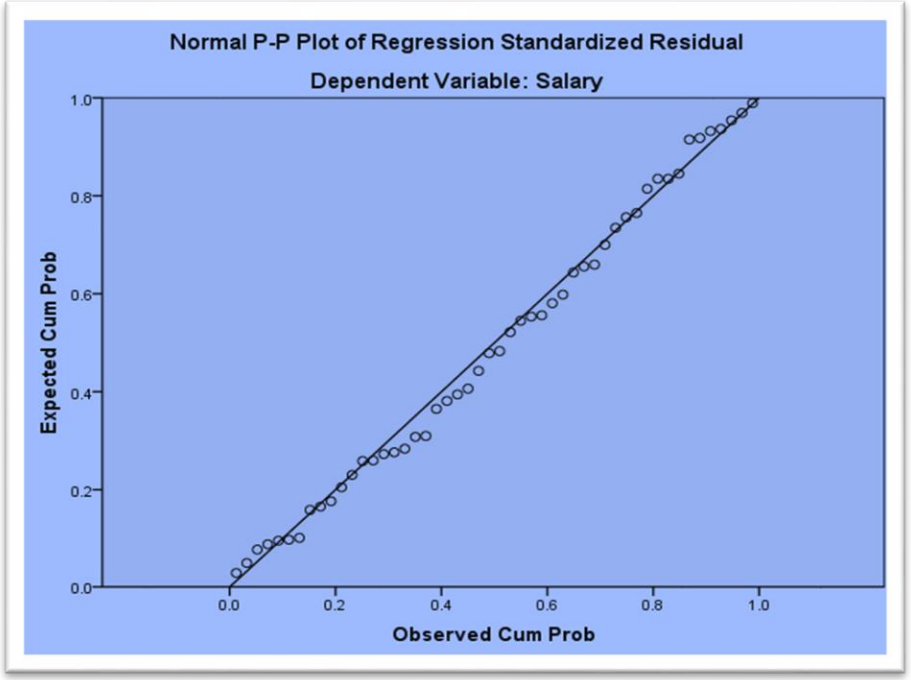
Residual Analysis

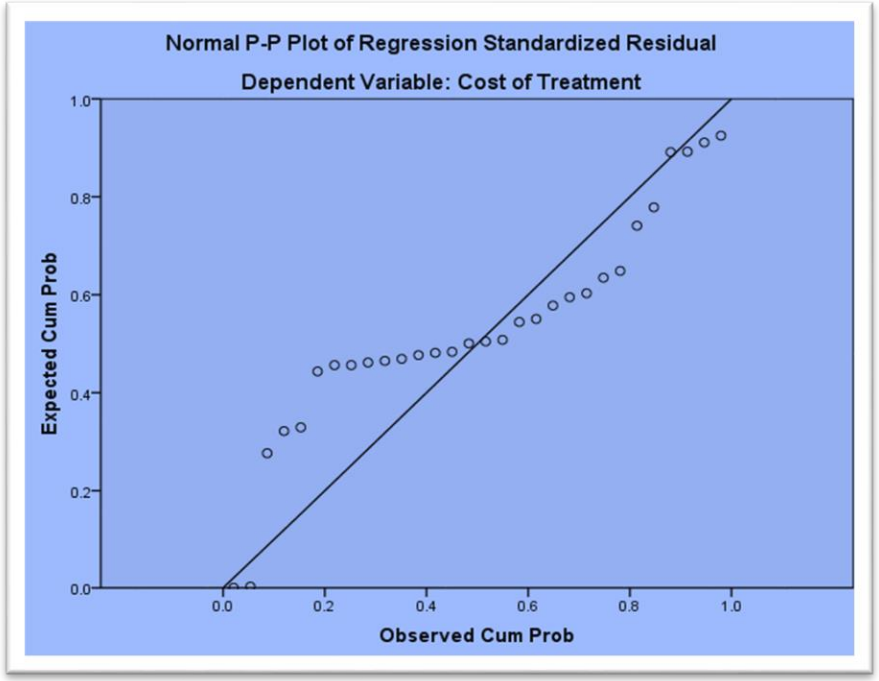
Residual (error) analysis is important to check whether the assumptions of regression models have been satisfied. It is performed to check the following:

- The residuals $(Y_i - \hat{Y}_i)$ are normally distributed.
- The variance of residual is constant (homoscedasticity).
- The functional form of regression is correctly specified.
- If there are any outliers

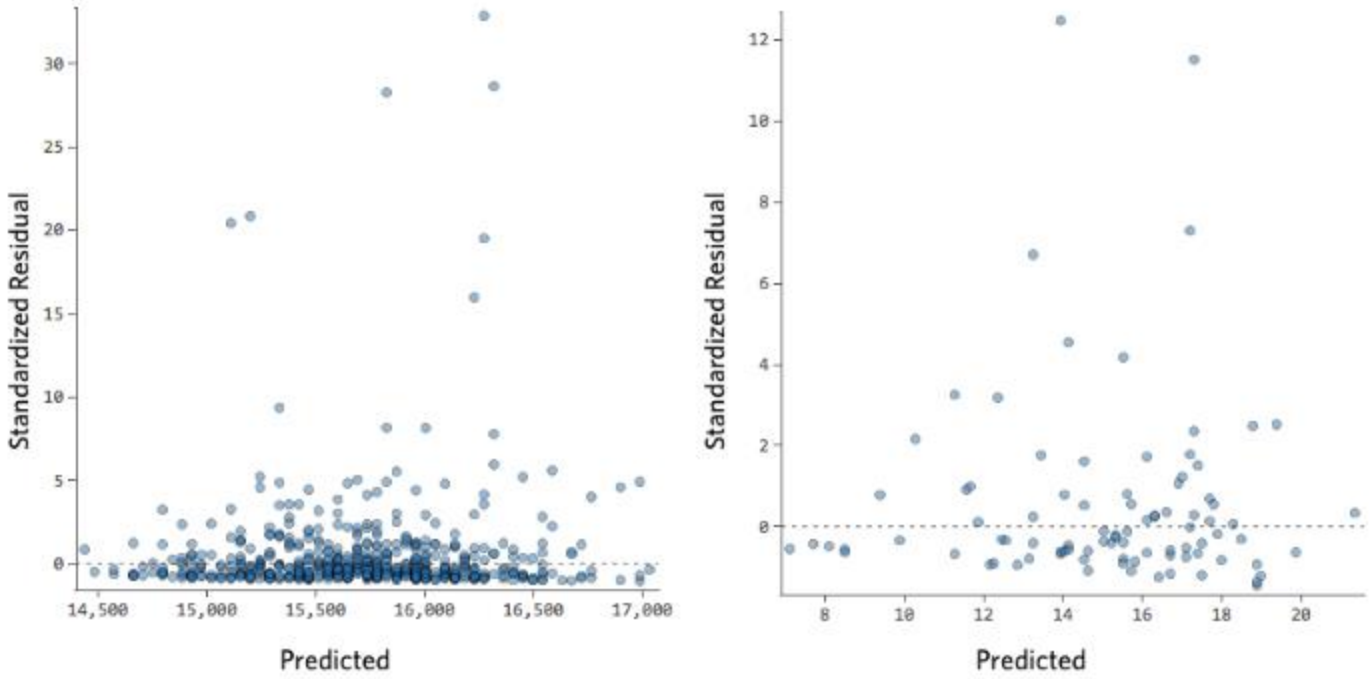
Checking for Normal Distribution of Residuals $(Y_i - \hat{Y}_i)$

- The easiest technique to check whether the residuals follow normal distribution is to use the P-P plot (Probability-Probability plot).
- The P-P plot compares the cumulative distribution function of two probability distributions against each other





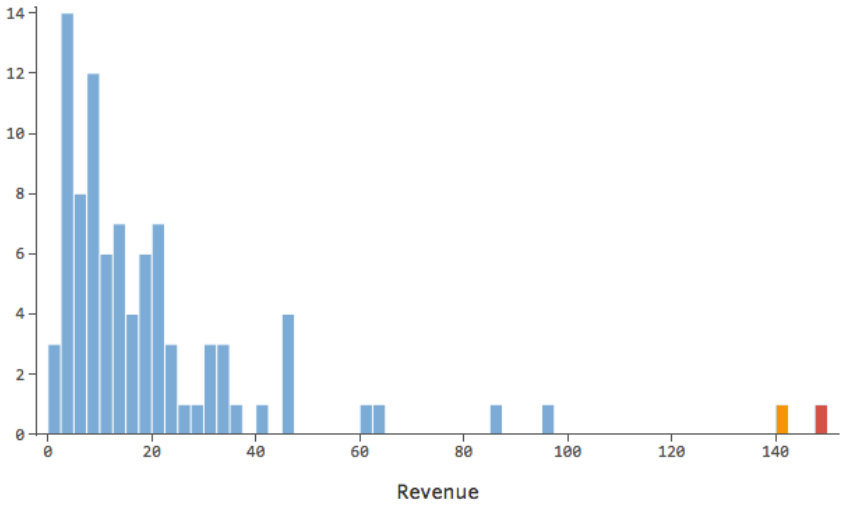
Y-axis unbalanced

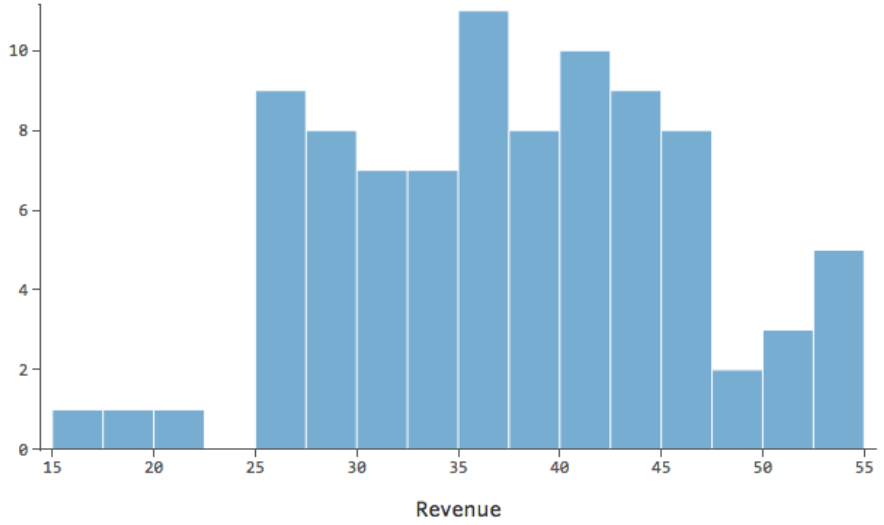


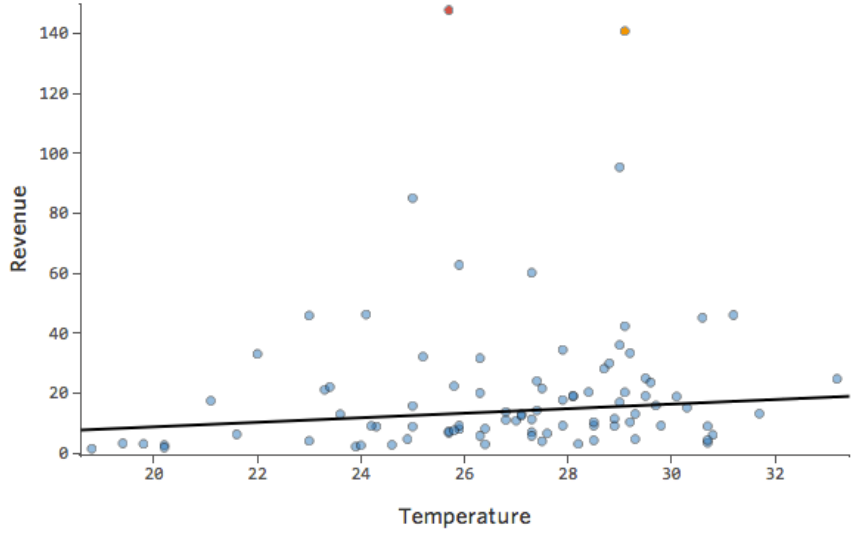
- Lemonade stalls have usually low revenues
- Once in a while, there is a high revenue

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

Y-axis unbalanced





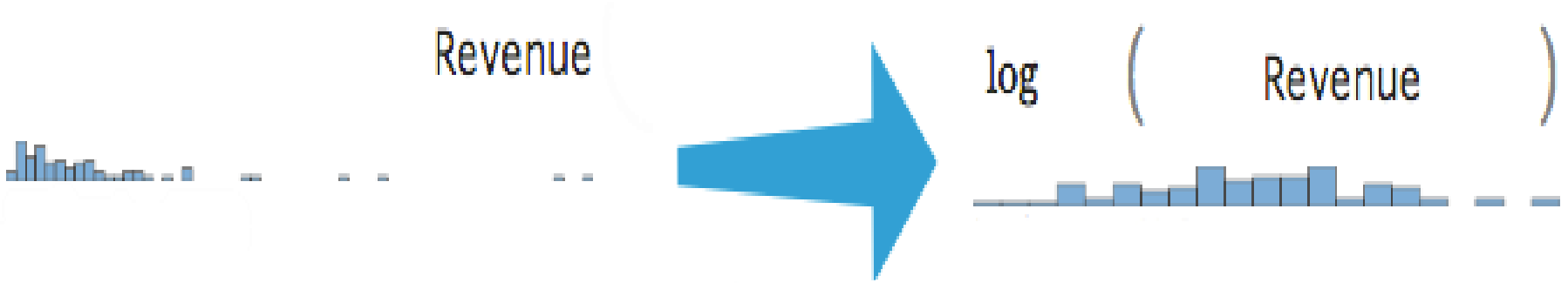


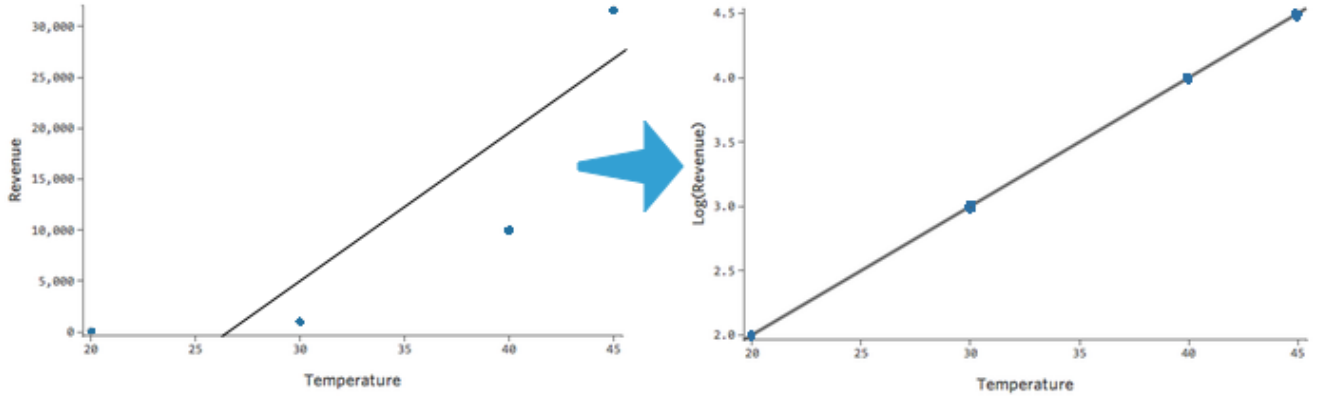
- We see an unbalanced plot

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

How do we fix this?

- Transform the variable…





Revenue = 1,120 * Temperature + 27,132

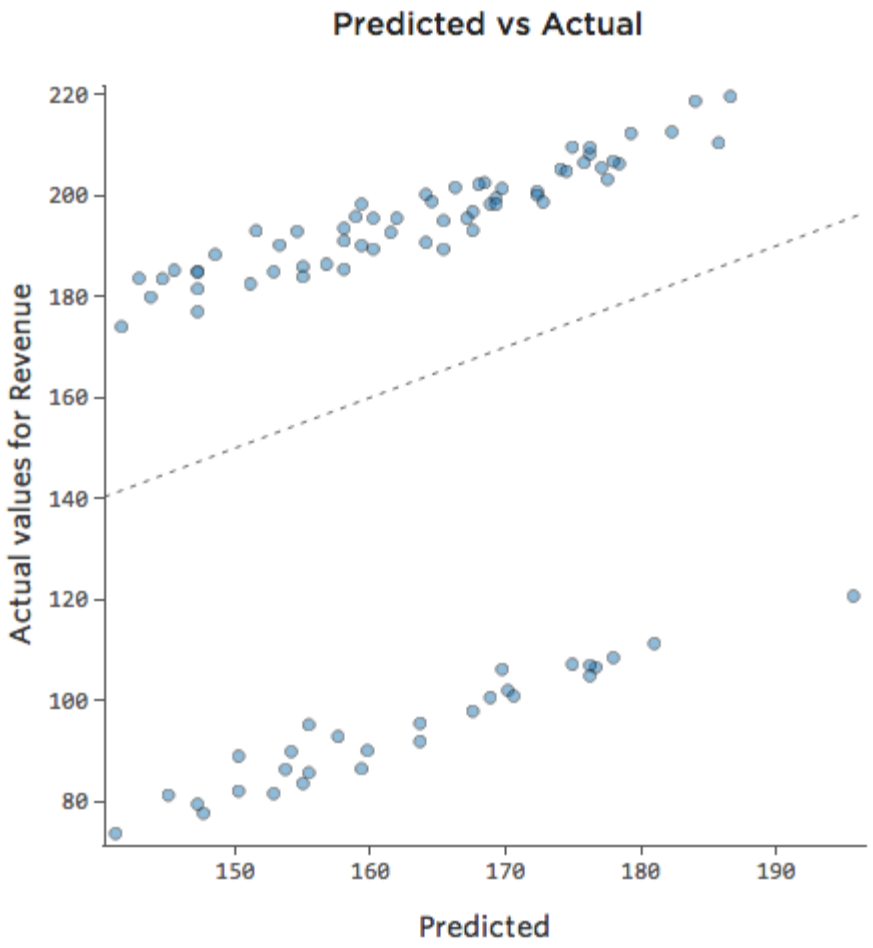
r-squared = 0.69

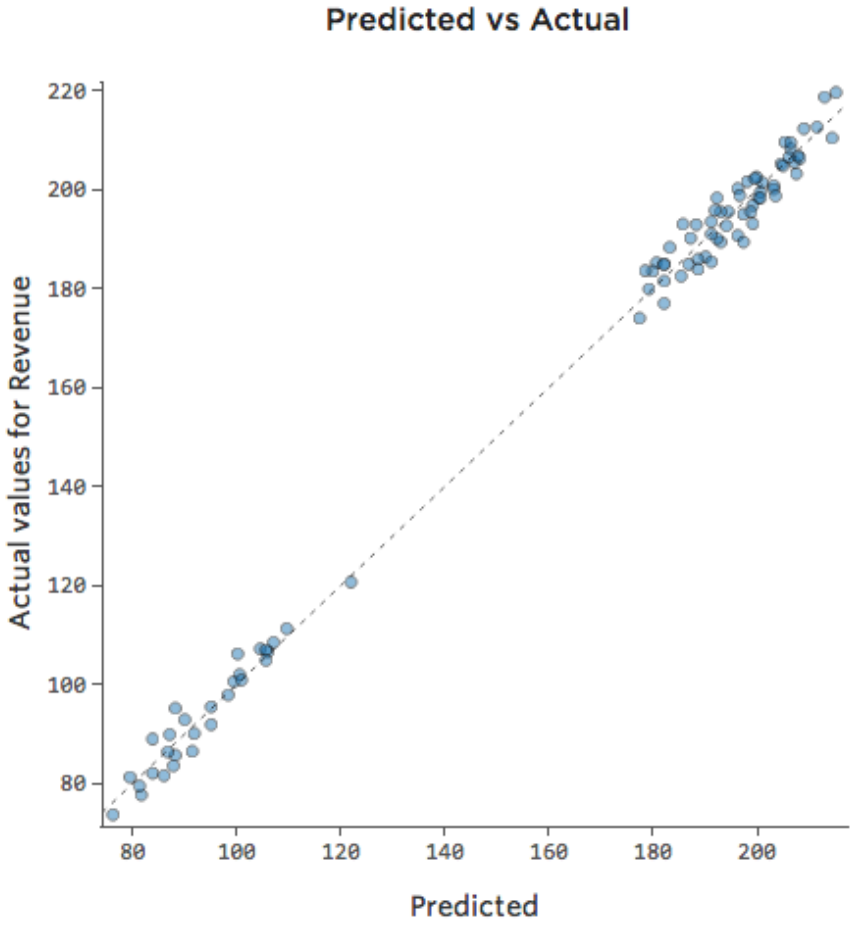
Log(Revenue) = 0.1 * Temperature

r-squared = 1

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

Adding a new variable

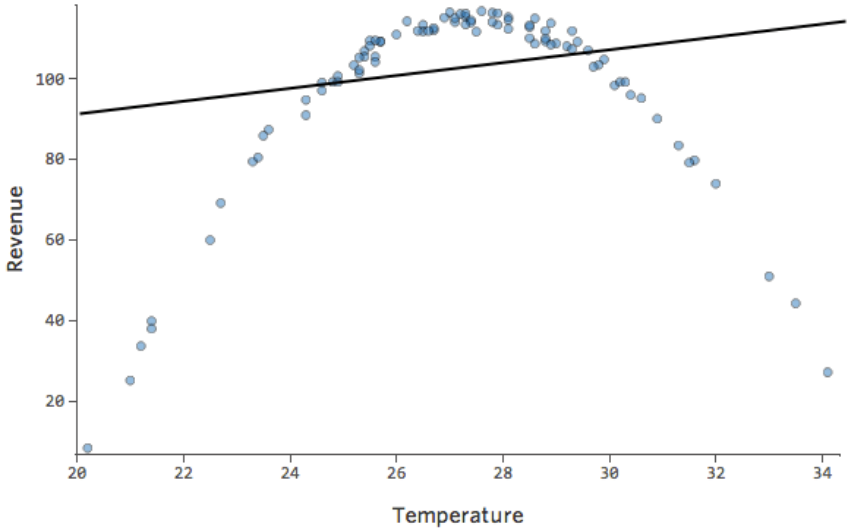


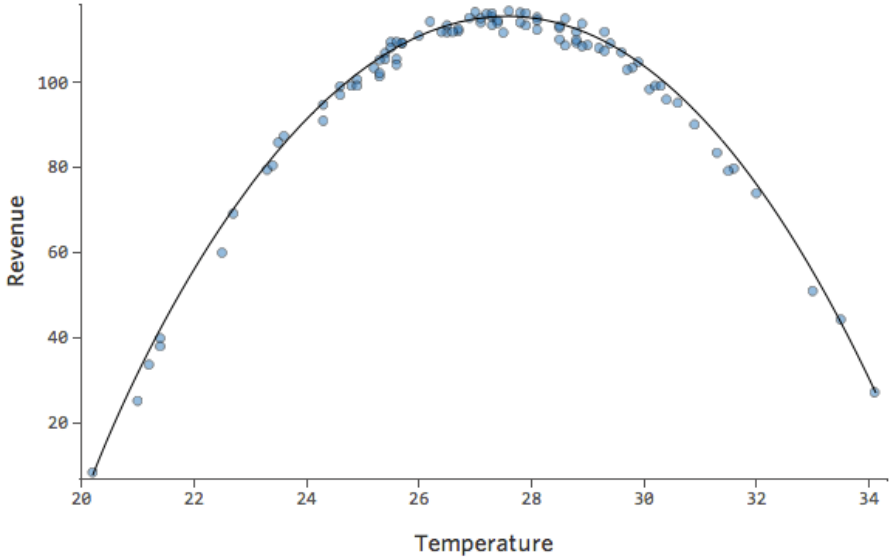


- Including ‘weekend’ in the model (like changing ‘date’ to ‘day’)

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

Fixing non-linearity





$y = 1.7x + 51$

-> A terrible model

$y = -2x^2 + 111x - 1408$

-> A much better model!

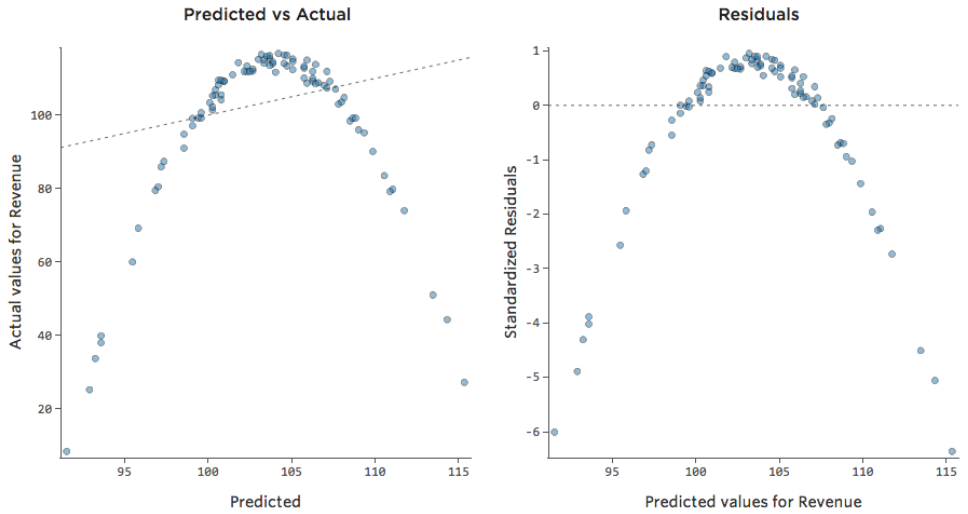
- Or… resort to non-linear models!

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

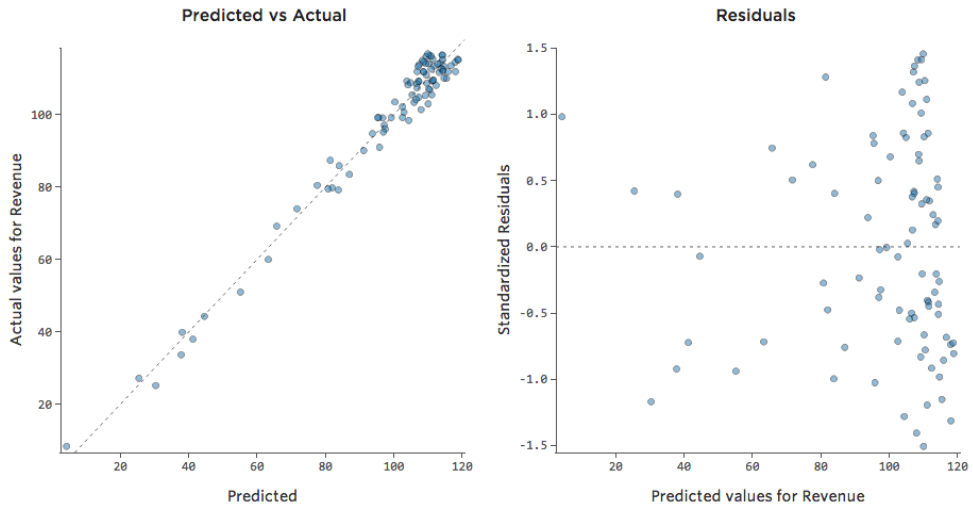
Comparing diagnostic plots for linear vs non-linear models when the relationship between variable is nonlinear

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header

- Linear model



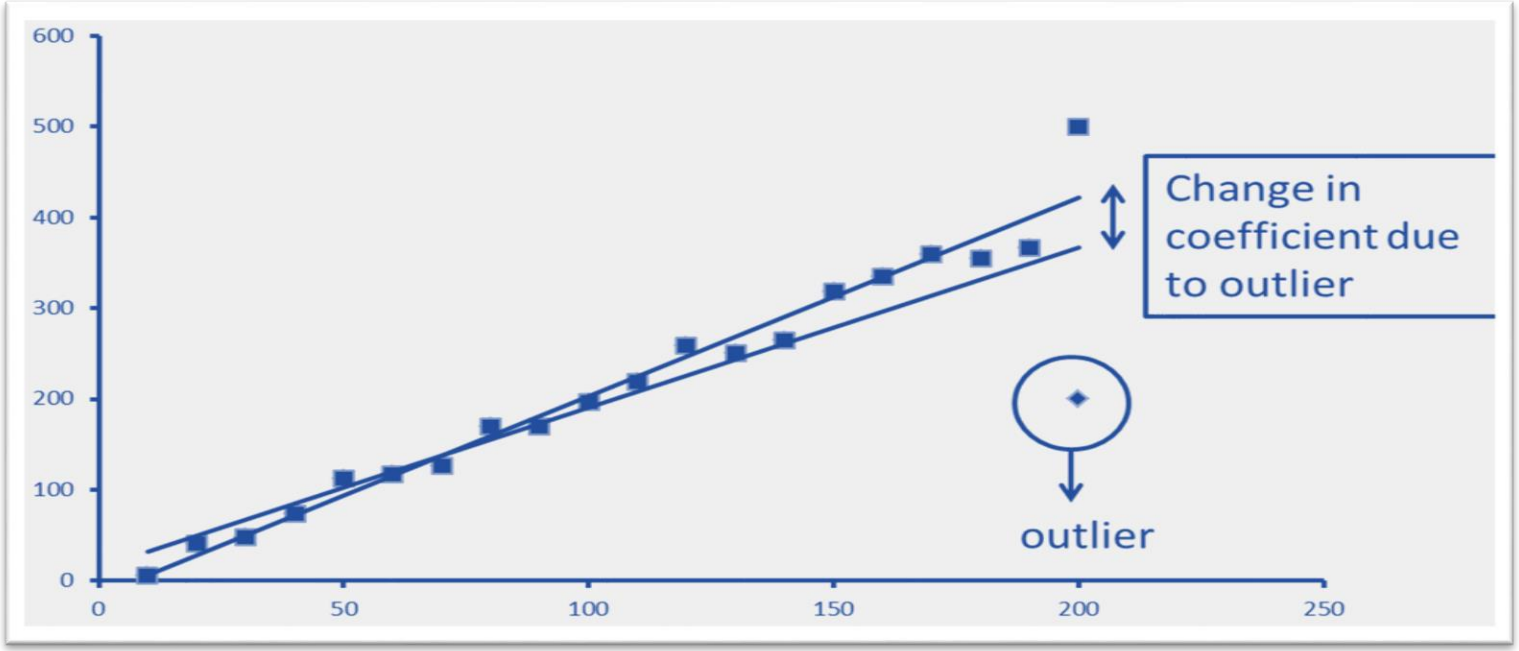
- Non-linear model



DATA ANALYTICS

Outlier Analysis

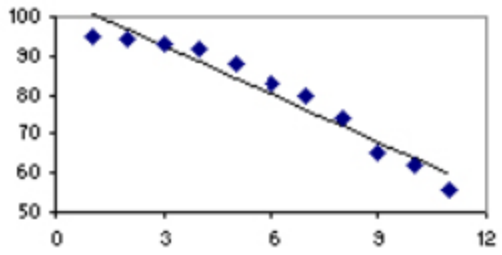
- Outliers are observations whose values show a large deviation from mean value
- Presence of an outlier can have significant influence on values of regression coefficients. Thus, it is important to identify the existence of outliers in the data



What effect do outliers have on the regression model?

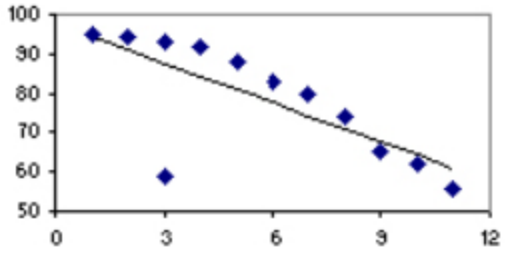
- Influential points: outliers that change the slope of the regression line

Without Outlier



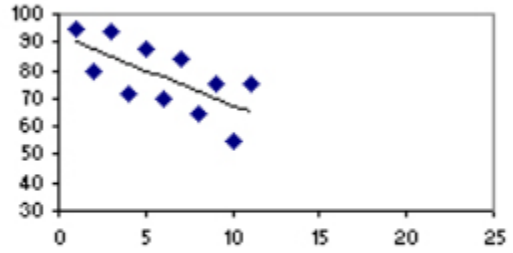
Regression equation: $\hat{y} = 104.78 - 4.10x$
Coefficient of determination: $R^2 = 0.94$

With Outlier



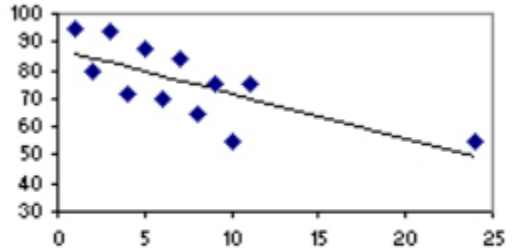
Regression equation: $\hat{y} = 97.51 - 3.32x$
Coefficient of determination: $R^2 = 0.55$

Without Outlier



Regression equation: $\hat{y} = 92.54 - 2.5x$
Slope: $b_0 = -2.5$
Coefficient of determination: $R^2 = 0.46$

With Outlier



Regression equation: $\hat{y} = 87.59 - 1.6x$
Slope: $b_0 = -1.6$
Coefficient of determination: $R^2 = 0.52$

https://stattrek.com/regression/influential-points.aspx?Tutorial=ap

Regression Vs Correlation

- Both quantify "strength of relationship", between two continuous variables.
- Correlation: commutative, assumes both variables to be random variables
- Regression: What is the change in Y for a unit change in X?
  - X is fixed with no error (i.e., deterministic)

https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/

DATA ANALYTICS

Correlation or Regression?

- You have two measuring systems and you want to see how well they agree with each other. So you measure the same 20 parts with each measuring system.
  - Correlation (Variables are interchangeable)

- You want to predict blood pressure for different doses of a drug.
  - Regression (Drug doses are fixed + Prediction)

- A clinical trial has multiple endpoints and you want to know which pair of endpoints has the strongest linear relationship.
  - Correlation (scatterplot or correlation matrix would identify the linear relationship)

- You want to know how much the response (Y) changes for every one unit increase in (X).
  - Regression (slope in a regression analysis gives this)

https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017 (Ch. 9.1-9.4)

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

DATA ANALYTICS

Unit 2:Linear Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Linear Regression Contd.,

Mamatha H R
Department of Computer Science and Engineering

Coefficient of Determination (R-Square or R2)

- The co-efficient of determination (or *R*-square or R^2) measures the percentage of variation in *Y* explained by the model ($\beta_0 + \beta_1 X$).
- The simple linear regression model can be broken into explained variation and unexplained variation as shown in

$$\underbrace{Y_i}_{\text{Variation in Y}} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Variation in Y explained by the model}} + \underbrace{\varepsilon_i}_{\text{Variation in Y not explained by the model}}$$

In absence of the predictive model for Y_i, the users will use the mean value of Y_i. Thus, the total variation is measured as the difference between Y_i and mean value of Y_i

Description of total variation, explained variation and unexplained variation

Variation Type	Measure	Description
Total Variation (SST)	$(Y_i - \bar{Y})^2$	Total variation is the difference between the actual value and the mean value.
Variation explained by the model (SSR)	$(\hat{Y}_i - \bar{Y})^2$	Variation explained by the model is the difference between the estimated value of Y_i and the mean value of Y
Variation not explained by model (SSE)	$(Y_i - \hat{Y}_i)^2$	Variation not explained by the model is the difference between the actual value and the predicted value of Y_i (error in prediction)

The relationship between the total variation, explained variation and the unexplained variation is given as follows:

$$\underbrace{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}_{SST}=\underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}_{SSR}+\underbrace{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}_{SSE}$$

where SST is the sum of squares of total variation, SSR is the sum of squares of variation explained by the regression model and SSE is the sum of squares of errors or unexplained variation.

Coefficient of Determination or R-Square

The coefficient of determination (R^2) is given by

$$\text{Coefficient of determination} = R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SSR}{SST} = \frac{\left(\hat{Y}_i - \bar{Y}\right)^2}{\left(Y_i - \bar{Y}\right)^2}$$

Since SSR = SST – SSE, the above Eq. can be written as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\left(\hat{Y}_i - Y_i\right)^2}{\left(Y_i - \bar{Y}\right)^2}$$

Coefficient of Determination or R-Square

Thus, R^2 is the proportion of variation in response variable Y explained by the regression model. Coefficient of determination (R^2) has the following properties:

- The value of R^2 lies between 0 and 1.
- Higher value of R^2 implies better fit, but one should be aware of spurious regression.
- Mathematically, the square of correlation coefficient is equal to coefficient of determination (i.e., $r^2 = R^2$).
- We do not put any minimum threshold for R^2; higher value of R^2 implies better fit. However, a minimum value of R^2 for a given significance value α can be derived using the relationship between the F-statistic and R^2

Spurious Regression

Number of Facebook users and the number of people who died of helium poisoning in UK

Year	Number of Facebook users in millions (X)	Number of people who died of helium poisoning in UK (Y)
2004	1	2
2005	6	2
2006	12	2
2007	58	2
2008	145	11
2009	360	21
2010	608	31
2011	845	40

DATA ANALYTICS

Facebook users versus helium poisoning in UK

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.996442					
R Square	0.992896					
Standard Error	1.69286					
Observations	9					
ANOVA						
		SS	MS	F	Significance F	
Regression	1	2803.94	2803.94	978.4229	8.82E-09	
Residual	7	20.06042	2.865775			
Total	8	2824				
	Coefficients	Standard Error	t-stat	P-value	Lower 95%	Upper 95%
Intercept	1.9967	0.76169	2.62143	0.034338	0.195607	3.79783
FB	0.0465	0.00149	31.27975	8.82E-09	0.043074	0.050119

The *R*-square value for regression model between the number of deaths due to helium poisoning in UK and the number of Facebook users is 0.9928. That is, 99.28% variation in the number of deaths due to helium poisoning in UK is explained by the number of Facebook users.

The regression model is given as $Y = 1.9967 + 0.0465 X$

Hypothesis Test for Regression Co-efficient (t-Test)

- The regression co-efficient (β_1) captures the existence of a linear relationship between the response variable and the explanatory variable.
- If β_1 = 0, we can conclude that there is no statistically significant linear relationship between the two variables.

  ➢ The estimate of β_1 using OLS is given by

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{\bar{X}\sum_{i=1}^{n}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Contd.,

Above eq. can be written as follows:

$$\beta_1 = \frac{\sum_{i=1}^{n} K_i Y_i}{\sum_{i=1}^{n} K_i^2} \text{ where } K_i = (X_i - \bar{X})$$

That is, the value of β_1 is a function of Y_i (K_i is a constant since X_i is assumed to be non-stochastic)

The standard error of β_1 is given by

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{(X_i - \bar{X})^2}}$$

Contd.,

In above Eq. Se is the standard error of estimate (or standard error of the residuals) that measures the accuracy of prediction and is given by

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-2}}$$

The denominator in above Eq. is (n – 2) since β_0 and β_1 are estimated from the sample in estimating Y_i and thus two degrees of freedom are lost. The standard error of $\hat{\beta}_1$ can be written as

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{(X_i - \bar{X})^2}} = \frac{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \Big/ n-2}}{\sqrt{(X_i - \bar{X})^2}}$$

Contd.,

The null and alternative hypotheses for the SLR model can be stated as follows:

H_0: There is no relationship between *X* and *Y*

H_A: There is a relationship between *X* and *Y*

- $\beta_1 = 0$ would imply that there is no linear relationship between the response variable *Y* and the explanatory variable *X*. Thus, the null and alternative hypotheses can be restated as follows:

  $H_0: \beta_1 = 0$

  $H_A: \beta_1 \neq 0$

- The corresponding *t*-statistic is given as

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_e(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{S_e(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)}$$

Test for Overall Model: Analysis of Variance (F-test)

How good is the overall 'fit'?

The null and alternative hypothesis for *F*-test is given by

H_0: There is no statistically significant relationship between *Y* and any of the explanatory variables (i.e., all regression coefficients are zero).

H_A: Not all regression coefficients are zero

- Alternatively:

H_0: All regression coefficients are equal to zero

H_A: Not all regression coefficients are equal to zero

- The *F*-statistic is given by

$$F = \frac{MSR}{MSE} = \frac{MSR/1}{MSE/n-2}$$

Z-Score

Z-score is the standardized distance of an observation from its mean value. For the predicted value of the dependent variable *Y*, the Z-score is given by

$$Z = \left(\frac{\hat{Y}_i - \bar{Y}}{\sigma_Y} \right)$$

Where and are, respectively, the mean and the standard deviation of dependent variable estimated from the sample data.

Mahalanobis Distance

Mahalanobis distance is the distance between specific values of the independent variable (X_i) to the centroid of all observations of the explanatory variable. Distances value of more than chi-square critical value (with degrees of freedom is equal to the number of explanatory variables) is classified as outliers.

Cook's Distance

Cook's distance measures how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters. Cook's distance for simple linear regression is given by

$$D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1)MSE}$$

where D_i is the Cook's distance measure for i^{th} observation,

$\hat{Y}_{j(i)}$ is the predicted value of j^{th} observation including i^{th} observation,

$\hat{Y}_j$ is the predicted value of j^{th} observation after excluding i^{th} observation from the sample, MSE is the Mean–Squared–Error.

k is the number of regression coefficients

Leverage Value

Leverage value of an observation measures the influence of that observation on the overall fit of the regression function. Leverage value for an observation in SLR is given by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Leverage value of more than $2/n$ or $3/n$ is treated as highly influential observation. In Eq. the first term $(1/n)$ will tend to zero for large value of n.

DFFit and DFBeta

- DFFit is the change in the predicted value of Y_i when case *i* is removed from the data set. DFBeta is the change in the regression coefficient values when an observation *i* is removed from the data.

Sum of Squared Errors (SSE)

- The sum of squared errors SSE output is 5226.19.
- To do the best fit of line intercept, we need to apply a linear regression model to reduce the SSE value at minimum as possible.
- To identify a slope intercept, we use the equation

$$y = mx + b,$$

'm' is the slope
'x' → independent variables
'b' is intercept

Ordinary Least Squares (OLS) Method

We will use Ordinary Least Squares method to find the best line intercept (b) slope (m)

To use OLS method, we apply the below formula to find the equation

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b = \bar{y} - m * \bar{x}$$

$x = independent\ variables$

$\bar{x} = average\ of independent\ variables$

$y = dependent\ variables$

$\bar{y} = average\ of dependent\ variables$

Ordinary Least Squares (OLS) Method

We need to calculate slope ‘m’ and line intercept ‘b’.

Years of Experience x_i	Salary (in 1000$) y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2	15	-5.56	-30.44	169.24	30.91
3	28	-4.56	-17.44	79.53	20.79
5	42	-2.56	-3.44	8.81	6.55
13	64	5.44	18.56	100.97	29.59
8	50	0.44	4.56	2.01	0.19
16	90	8.44	44.56	376.09	71.23
11	58	3.44	12.56	43.21	11.83
1	8	-6.56	-37.44	245.61	43.03
9	54	1.44	8.56	12.33	2.07
$\bar{x}$ =7.56	$\bar{y}$ = 45.44			∑ = 1037.8	∑ = 216.19

OLS method calculations

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b = \bar{y} - m * \bar{x}$$

m = 1037.8 / 216.19

m = 4.80

b = 45.44 - 4.80 * 7.56 = 9.15

Hence, y = mx + b → 4.80x + 9.15

y = 4.80x + 9.15

Calculate SSE

Let us calculate SSE again by using our output equation.

SSE calculations again after OLS method implementation

Years of Experience x	Salary (in 1000$) y	$\bar{y}$ = mx + b $\bar{y}$ = 4.79x + 9.18	Error $y_i - \bar{y}$	Error²
2	15	18.76	-3.76	14.14
3	28	23.55	4.45	19.80
5	42	33.13	8.87	78.68
13	64	71.45	-7.45	55.50
8	50	47.5	2.5	6.25
16	90	85.82	4.18	17.47
11	58	61.87	-3.87	14.98
1	8	13.97	-5.97	35.64
9	54	52.29	1.71	2.92
				SSE = 245.38

Sum of Squared Error got reduced significantly from 5226.19 to 245.38.

OLS

- Ordinary Least Square method looks simple and computation is easy.
- OLS method will work for both univariate dataset which is single independent variables and single dependent variables and multi-variate dataset containing a single independent variables set and multiple dependent variables sets.
- Ordinary least squares (OLS) is a non-iterative method Ordinary Least Squares solution is the analytical solution and this solution is not scalable.
- Applying this to complex and non-linear algorithms like Support Vector Machine will not be feasible.
- So we will find the numerical approximation of this solution by iterative method — which would be close to (but not exactly equal to) the OLS solution — which gave us the exact solution.

Gradient Descent Algorithm

Gradient descent algorithm's main objective is to minimize the cost function. It is one of the best optimization algorithms to minimize errors (difference of actual value and predicted value).

Let's represent the hypothesis h, which is function or a learning algorithm.

$$h_\theta = \theta_0 + \theta_1 x$$

The goal is similar like the OLS operation that we did to find out a best fit of intercept line 'y' in the slope 'm'. Using Gradient descent algorithm also, we will figure out a minimal cost function by applying various parameters for theta 0 and theta 1 and see the slope intercept until it reaches convergence.

Gradient Descent Algorithm

Analogy:

Imagine a valley and a person with no sense of direction who wants to get to the bottom of the valley. He goes down the slope and takes large steps when the slope is steep and small steps when the slope is less steep. He decides his next position based on his current position and stops when he gets to the bottom of the valley which was his goal.

Gradient Descent Algorithm

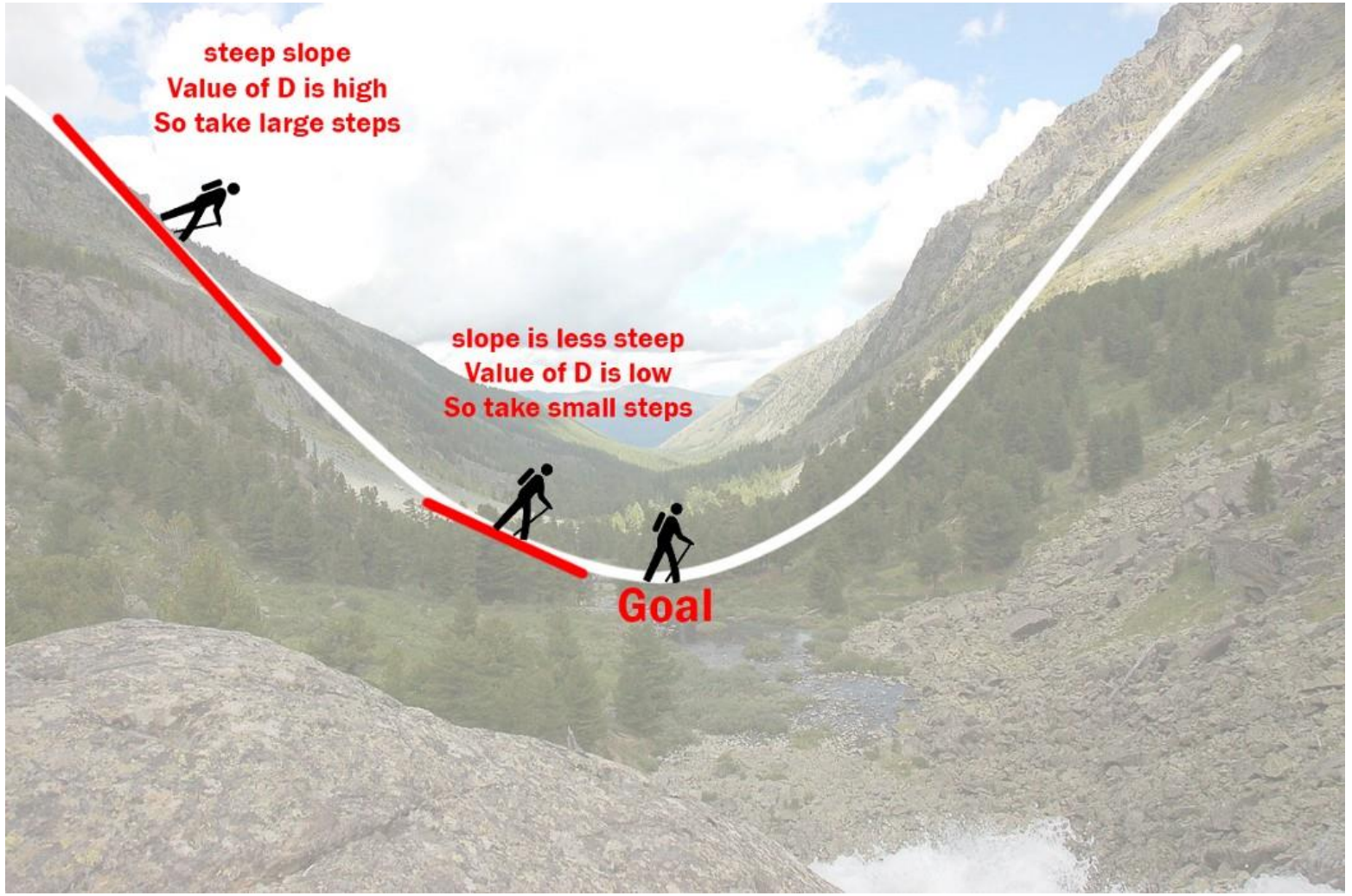


Illustration of how the gradient descent algorithm works

Gradient Descent Algorithm

Let's try applying gradient descent to m and c and approach it step by step:
Initially let m = 0 and c = 0. Let L be our learning rate. This controls how much the value of m changes with each step. L could be a small value like 0.0001 for good accuracy.

Calculate the partial derivative of the loss function with respect to m, and plug in the current values of x, y, m and c in it to obtain the derivative value D.

$$D_m = \frac{1}{n}\sum_{i=0}^{n} 2(y_i - (mx_i + c))(-x_i)$$

$$D_m = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - \bar{y}_i)$$

D_m is the value of the partial derivative with respect to m.

Gradient Descent Algorithm

Find the partial derivative with respect to c, Dc :

$$D_c = \frac{-2}{n} \sum_{i=0}^{n} (y_i - \bar{y}_i)$$

Now we update the current value of m and c using the following equation:

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$

We repeat this process until our loss function is a very small value or ideally 0 (which means 0 error or 100% accuracy). The value of m and c that we are left with now will be the optimum values.

Gradient Descent Algorithm

In our analogy,

- m can be considered the current position of the person.
- D is equivalent to the steepness of the slope
- L can be the speed with which he moves.
- Now the new value of m that we calculate using the above equation will be his next position, and L×D will be the size of the steps he will take.
- When the slope is more steep (D is more) he takes longer steps and when it is less steep (D is less), he takes smaller steps.
- Finally he arrives at the bottom of the valley which corresponds to our loss = 0.

Exercise

- Explore more on the Gradient Descent Algorithm- Derivation part

- List the various linear and non-linear algorithms where the Gradient Descent is used.

- Implement both OLS and Gradient descent on a dataset of your choice and list the value of SSE in both the cases.

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

DATA ANALYTICS

Unit 2:Multiple Linear Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Multiple Linear Regression

Mamatha H R

Department of Computer Science and Engineering

Multiple Linear Regression

- Multiple linear regression means linear in regression parameters (beta values). The following are examples of multiple linear regression:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_2^2 \ldots + \beta_k x_k + \varepsilon$$

An important task in multiple regression is to estimate the beta values (β_1, β_2, β_3 etc…)

Regression: Matrix Representation

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{k1} \\ 1 & x_{12} & x_{22} & x_{k2} \\ & & & \\ & & & \\ 1 & x_{1n} & x_{2n} & x_{kn} \end{bmatrix} \bullet \begin{bmatrix} \beta_0 \\ \beta_1 \\ \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

Ordinary Least Squares Estimation for Multiple Linear Regression

The assumptions that are made in multiple linear regression model are as follows:

- The regression model is linear in parameter.
- The explanatory variable, X, is assumed to be non-stochastic (that is, X is deterministic).
- The conditional expected value of the residuals, $E(\varepsilon_i/X_i)$, is zero.
- In a time series data, residuals are uncorrelated, that is, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

MLR Assumptions

- The residuals, ε_i, follow a normal distribution.

- The variance of the residuals, Var(ε_i/X_i), is constant for all values of X_i. When the variance of the residuals is constant for different values of X_i, it is called **homoscedasticity**. A non-constant variance of residuals is called **heteroscedasticity**.

- There is no high correlation between independent variables in the model (called **multi-collinearity**). Multi-collinearity can destabilize the model and can result in incorrect estimation of the regression parameters.

The regression coefficients $\hat{\boldsymbol{\beta}}$ is given by

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

The estimated values of response variable are

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

In above Eq. the predicted value of dependent variable $\hat{Y}_i$ is a linear function of Y_i. Equation can be written as follows:

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}$ is called the **hat matrix**, also known as the **influence matrix**, since it describes the influence of each observation on the predicted values of response variable.

Hat matrix plays a crucial role in identifying the outliers and influential observations in the sample.

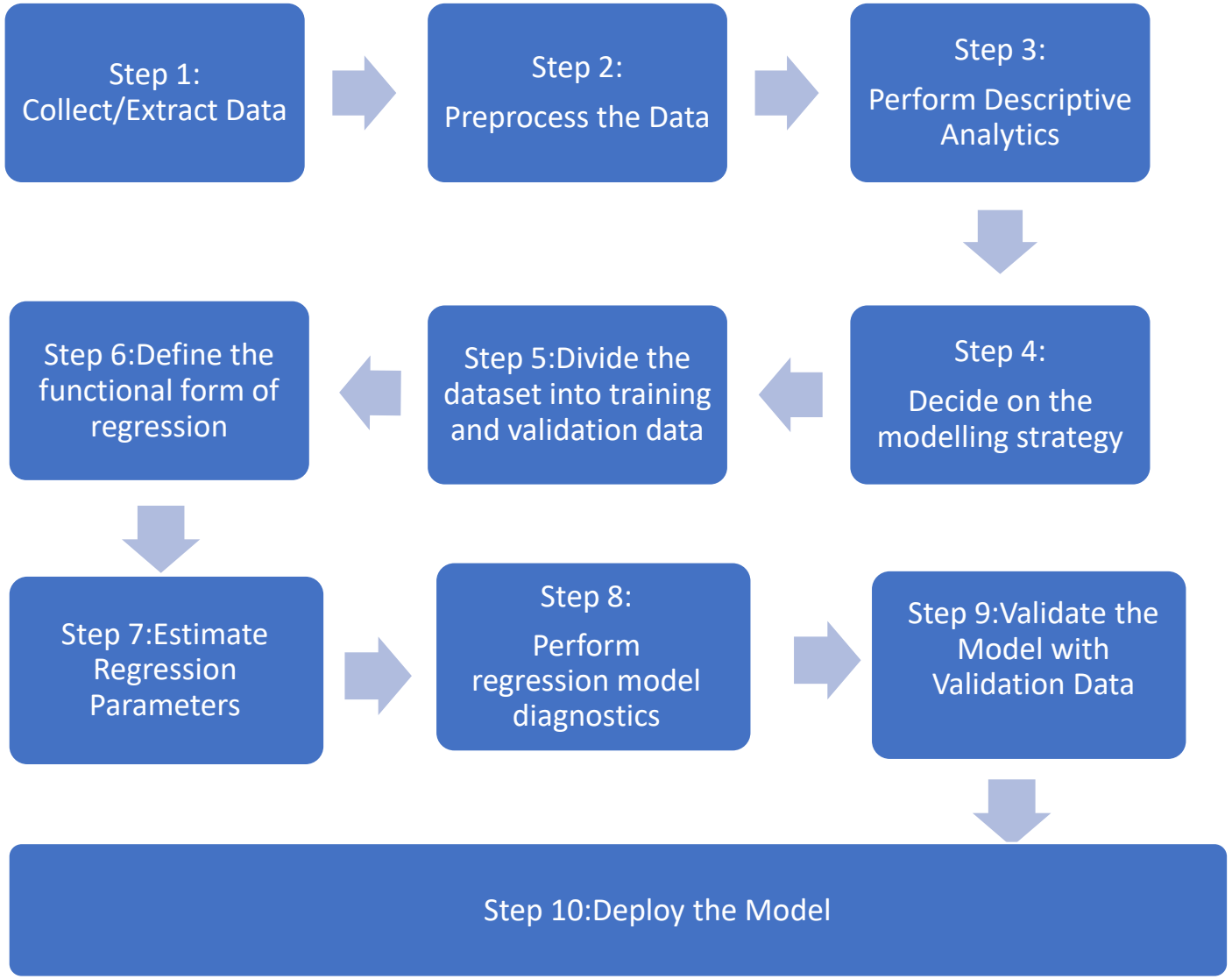
Multiple Linear Regression Model Building

A few examples of MLR are as follows:

- The treatment cost of a cardiac patient may depend on factors such as age, past medical history, body weight, blood pressure, and so on.
- Salary of MBA students at the time of graduation may depend on factors such as their academic performance, prior work experience, communication skills, and so on.
- Market share of a brand may depend on factors such as price, promotion expenses, competitors' price, etc.

DATA ANALYTICS

Framework for building multiple linear regression (MLR)



Modelling Strategy

- When the number of variables runs into several hundreds, building regression models can get complicated due to multi-collinearity as well as computational complexity since estimation of regression parameters involves matrix inversion (Hat Matrix).

- The data scientist may also use specific variable selection approaches such as Forward Selection, Backward Elimination or Stepwise Regression

Define the Functional Form

Most data scientists may start with a linear relationship between the dependent and the independent variables. However, the functional form may be changed if there is a lack of fit.

Estimate Regression Parameters

- Once the functional form is specified, the next step is to estimate the partial regression coefficients using the method of **Ordinary Least Squares** (OLS).
- OLS is used to fit a polygon through a set of data points, such that the sum of the squared distances between the actual observations in the sample and the regression equation is minimized.
- OLS provides the **Best Linear Unbiased Estimate** (BLUE), that is,

Where $\hat{\beta}$ is the population parameter and $E\left[\beta - \hat{\beta}\right] = 0$ is the estimated parameter value from the sample

Perform Regression Model Diagnostics

F-test is used for checking the overall significance of the model whereas t-tests are used to check the significance of the individual variables. Presence of multi-collinearity can be checked through measures such as **Variance Inflation Factor** (VIF).

Validate the Model using Validation Data

The measures that can be used for validating the model in the validation data are as follows:

- R^2 or Adjusted R^2
- Mean absolute percentage error, $\sum_{i=1}^{K} \frac{1}{K} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\%$ where K is the number of cases in the validation data.
- Root Mean Square Error (RMSE), $\sqrt{\sum_{i=1}^{K} \frac{1}{n} \left(Y_i - \hat{Y}_i \right)^2}$

Part (Semi-Partial) Correlation and Regression Model Building

The increase in the coefficient of determination, R^2, when a new variable is added is given by the square of the semi-partial correlation of the newly added variable with dependent variable Y.

Consider a regression model with two independent variables (say X_1 and X_2). The model can be written as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

Partial Correlation

- Partial correlation is the correlation between the response variable Y and the explanatory variable X_1 when influence of X_2 is removed from both Y and X_1 (in other words, when X_2 is kept constant).

- Alternatively, partial correlation is the correlation between residualized response and residualized explanatory variables.

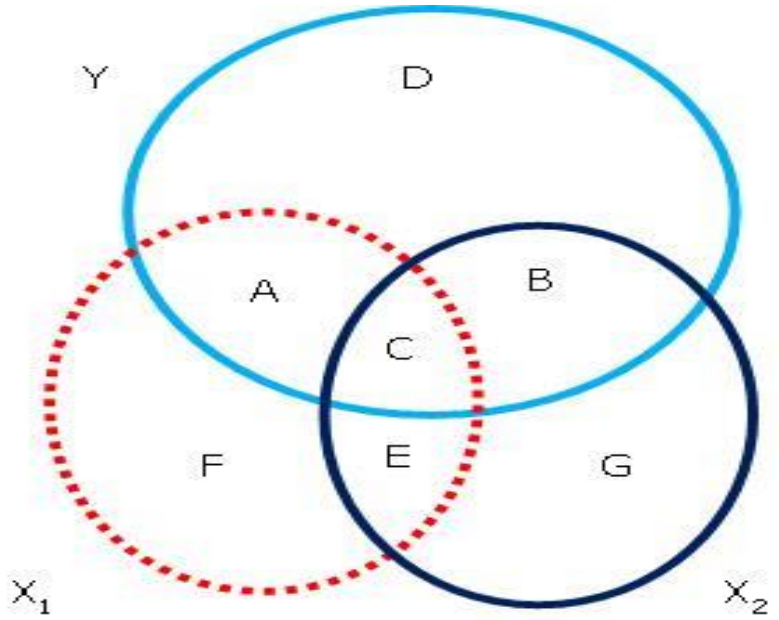
Let r_{YX_1,X_2} denote the partial correlation between Y and X_1 when X_2 is kept constant. Then r_{YX_1,X_2} is given by

$$r_{YX_1,X_2} = \frac{r_{YX_1} - r_{YX_2} \times r_{X_1X_2}}{\sqrt{(1-r_{YX_2}^2) \times (1-r_{X_1X_2}^2)}}$$

Semi-Partial Correlation (or Part Correlation)

- Consider a regression model between a response variable Y and two independent variables X_1 and X_2.
- The semi-partial (or part correlation) between a response variable Y and independent variable X_1 measures the relationship between Y and X_1 when the influence of X_2 is removed from only X_1 but not from Y.
- It is equivalent to removing portions C and E from X_1 in the Venn diagram shown in Figure
- Semi-partial sr_{YX_1,X_2} correlation between Y and X_1, when influence of X2 is removed from X1 is given by

$$sr_{YX_1,X_2} = \frac{r_{YX_1} - r_{YX_1} r_{YX_2}}{\sqrt{(1 - r^2_{X_1X_2})}}$$



Semi-partial (part) correlation plays an important role in regression model building.

The increase in R-square (coefficient of determination), when a new variable is added into the model, is given by the square of the semi-partial correlation.

Example:

The cumulative television rating points (*CTRP*) of a television program, money spent on promotion (denoted as *P*), and the advertisement revenue (in Indian rupees denoted as *R*) generated over one-month period for 38 different television programs is provided in Table 10.1. Develop a multiple regression model to understand the relationship between the advertisement revenue (R) generated as response variable and promotions (*P*) and *CTRP* as

DATA ANALYTICS

Example:

Serial	CTRP	P	R	Serial	CTRP	P	R
1	133	111600	1197576	20	156	104400	1326360
2	111	104400	1053648	21	119	136800	1162596
3	129	97200	1124172	22	125	115200	1195116
4	117	79200	987144	23	130	115200	1134768
5	130	126000	1283616	24	123	151200	1269024
6	154	108000	1295100	25	128	97200	1118688
7	149	147600	1407444	26	97	122400	904776
8	90	104400	922416	27	124	208800	1357644
9	118	169200	1272012	28	138	93600	1027308
10	131	75600	1064856	29	137	115200	1181976
11	141	133200	1269960	30	129	118800	1221636
12	119	133200	1064760	31	97	129600	1060452
13	115	176400	1207488	32	133	100800	1229028
14	102	180000	1186284	33	145	147600	1406196
15	129	133200	1231464	34	149	126000	1293936
16	144	147600	1296708	35	122	108000	1056384
17	153	133400	1329648	36	139	104400	1415316

Example:

The MLR model is given by

$$R(\text{Advertisement Revenue}) = \beta_0 + \beta_1 \times \text{CTRP} + \beta_2 \times \text{P}$$

The regression coefficients can be estimated using OLS estimation. The SPSS output for the above regression model is provided in tables

Model Summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.912[a]	0.832	0.822	57548.382

Example:

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Constant	41008.840	90958.920		0.451	0.655
	CTRP	5931.850	576.622	0.732	10.287	0.000
	P	3.136	0.303	0.736	10.344	0.000

The regression model after estimation of the parameters is given by

$$R = 41008.84 + 5931.850\ CTRP + 3.136\ P$$

For every one unit increase in CTRP, the revenue increases by 5931.850 when the variable promotion is kept constant, and for one unit increase in promotion the revenue increases by 3.136 when CTRP is kept constant. Note that television-rating point is likely to change when the amount spent on promotion is changed.

Standardized Regression Co-efficient

- A regression model can be built on standardized dependent variable and standardized independent variables, the resulting regression coefficients are then known as **standardized regression coefficients**.
- The standardized regression coefficient can also be calculated using the following formula:

$$\text{Standardized Beta} = \hat{\beta} \times \left(\frac{S_{X_i}}{S_Y} \right)$$

- Where S_{X_i} is the standard deviation of the explanatory variable Xi and SY is the standard deviation of the response variable Y.

Regression Models with Qualitative Variables

- In MLR, many predictor variables are likely to be qualitative or categorical variables. Since the scale is not a ratio or interval for categorical variables, we cannot include them directly in the model, since its inclusion directly will result in model misspecification. We have to pre-process the categorical variables using dummy variables for building a regression model.

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

Ch. 9.1-9.7.5 (both inclusive), Ch. 10.1-10.6 (both inclusive)

Coming up next week…

- Regression models with qualitative variables
- Validation and diagnostics for MLR
- Variable selection methods
  - Forward, backward and stepwise selection
  - Ridge and lasso regression
- Multivariate and nonlinear regression
- Classification using Logistic regression

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

DATA ANALYTICS

Unit 2:Multiple Linear Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Multiple Linear Regression

Mamatha H R

Department of Computer Science and Engineering

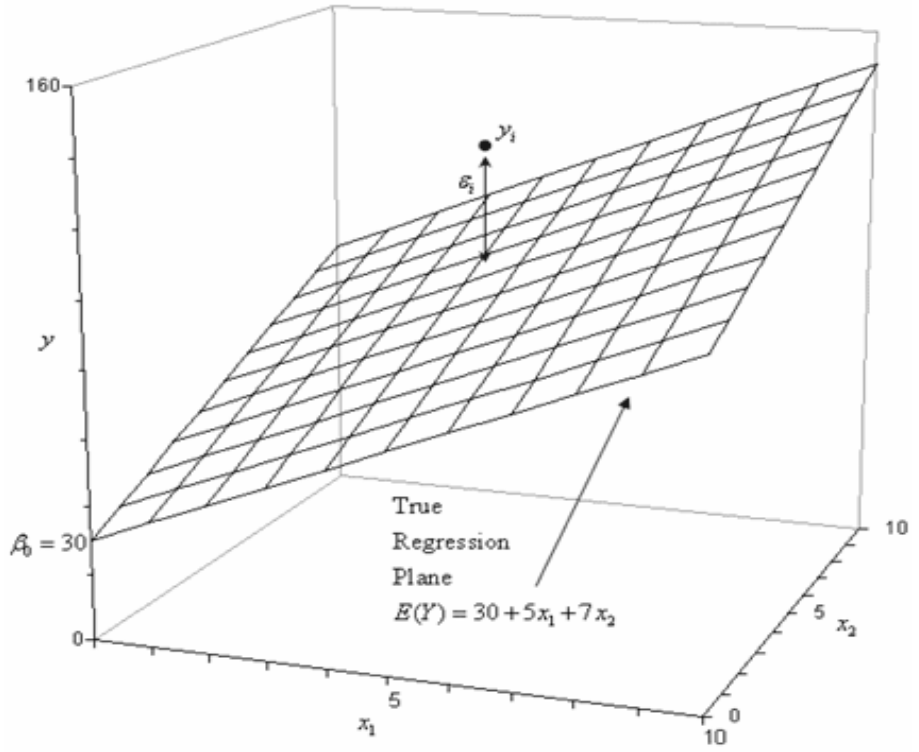
DATA ANALYTICS

Multiple Linear Regression - Hyperplane

The functional form of MLR is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

The model in Equation is called a response surface (hyperplane) that can be complex depending on the functional form of the relationship.



Example:

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Constant	41008.840	90958.920		0.451	0.655
	CTRP	5931.850	576.622	0.732	10.287	0.000
	P	3.136	0.303	0.736	10.344	0.000

The regression model after estimation of the parameters is given by

$$R = 41008.84 + 5931.850\ CTRP + 3.136\ P \quad \text{(Eq.10.22)}$$

For every one unit increase in CTRP, the revenue increases by 5931.850 when the variable promotion is kept constant, and for one unit increase in promotion the revenue increases by 3.136 when CTRP is kept constant. Note that television-rating point is likely to change when the amount spent on promotion is changed.

Partial Regression Coefficients

Modeling of multiple regression model and the mathematics behind partial regression coefficient.

$$R = \alpha_0 + \alpha_1 \times CTRP + \varepsilon_1$$

Note that, ε_1 is the variation in R (revenue generated through advertisement) not explained by $CTRP$.

$$P = \delta_0 + \delta_1 \times CTRP + \varepsilon_2$$

Here ε_2 is the variation in P not explained by $CTRP$.

$$\varepsilon_1 = \eta_0 + \eta_1 \times \varepsilon_2 + \varepsilon_3$$

The third model is between ε_1 (variation in advertisement revenue not explained by $CTRP$) and ε_2 (variation in promotion expenditure not explained by $CTRP$)

DATA ANALYTICS

Interpretation of MLR coefficients – Partial Regression Coefficient- Summary

TABLE 10.10 Model development in multiple linear regression

Model	Estimated Parameters Values	Model Interpretation
$R = \beta_0 + \beta_1 \times CTRP + \beta_2 \times P + \varepsilon$	$R = 41008.84 + 5931.85\ CTRP + 3.136 \times P$	Variation in R explained by $CTRP$ and P
$R = \alpha_0 + \alpha_1 \times CTRP + \varepsilon_1$	$R = 625763.106 + 4569.214 \times CTRP$	Variation in R explained by $CTRP$
$P = \delta_0 + \delta_1 \times CTRP + \varepsilon_2$	$P = 186456.659 - 434.495 \times CTRP$	Variation in P explained by $CTRP$
$\varepsilon_1 = \eta_0 + \eta_1 \times \varepsilon_2$	$\varepsilon_1 = 3.136 \times \varepsilon_2$ (The value of η_1 is same as that of β_2)	ε_1 is variation in R not explained by $CTRP$ ε_2 is variation in P not explained by $CTRP$

- That is, every new variable added to the model is partialled out from other independent variables and regressed with the partialled out dependent variable.

- The partial regression coefficient provides the change in the response variable for a unit change in the explanatory variable, when all other explanatory variables are kept constant or controlled.

- For example, for every one unit increase in CTRP, the revenue increases by 5931.84 provided the promotion expenses are kept constant.
- Similarly when the promotion is increased by one unit, the revenue increases by 3.136 provided CTRP is kept constant.
- However, in practice, it may not be possible to control a variable in many situations.

Standardized Regression Co-efficient

- A regression model can be built on standardized dependent variable and standardized independent variables, the resulting regression coefficients are then known as **standardized regression coefficients**.
- The standardized regression coefficient can also be calculated using the following formula:

$$\text{Standardized Beta} = \hat{\beta} \times \left(\frac{S_{X_i}}{S_Y} \right)$$

- Where S_{X_i} is the standard deviation of the explanatory variable X_i and S_Y is the standard deviation of the response variable Y.

S_Y = 136527.88, S_{CTRP} = 16.85, S_P = 32052.62

Standardized regression coefficient for

CTRP = 5931.85 * (16.85/136527.88) = 0.732

Standardized regression coefficient for

P = 3.136 * (32052.62/136527.88) = 0.736

- For one standard deviation change in the explanatory variable, the standard regression coefficient captures the number of standard deviations by which the response variable will change.

Regression Models with Qualitative Variables

- In MLR, many predictor variables are likely to be qualitative or categorical variables. Since the scale is not a ratio or interval for categorical variables, we cannot include them directly in the model, since its inclusion directly will result in model misspecification. We have to pre-process the categorical variables using dummy variables for building a regression model.

DATA ANALYTICS

Example:

The data in Table provides salary and educational qualifications of 30 randomly chosen people in Bangalore. Build a regression model to establish the relationship between salary earned and their educational qualifications.

S. No.	Education	Salary	S. No.	Education	Salary	S. No.	Education	Salary
1	1	9800	11	2	17200	21	3	21000
2	1	10200	12	2	17600	22	3	19400
3	1	14200	13	2	17650	23	3	18800
4	1	21000	14	2	19600	24	3	21000
5	1	16500	15	2	16700	25	4	6500
6	1	19210	16	2	16700	26	4	7200
7	1	9700	17	2	17500	27	4	7700
8	1	11000	18	2	15000	28	4	5600
9	1	7800	19	3	18500	29	4	8000
10	1	8800	20	3	19700	30	4	9300

Solution

Note that, if we build a model $Y = \beta_0 + \beta_1 \times$ Education, it will be incorrect. We have to use 3 dummy variables since there are 4 categories for educational qualification. Data in Table 10.12 has to be pre-processed using 3 dummy variables (HS, UG and PG) as shown in Table.

Pre-processed data (sample)

Observation	Education	Pre-processed data			Salary
		High School (HS)	Under-Graduate (UG)	Post-Graduate (PG)	
1	1	1	0	0	9800
11	2	0	1	0	17200
19	3	0	0	1	18500
27	4	0	0	0	7700

Example:

The corresponding regression model is as follows:

$$Y = \beta_0 + \beta_1 \times HS + \beta_2 \times UG + \beta_3 \times PG$$

where HS, UG, and PG are the dummy variables corresponding to the categories high school, under-graduate, and post-graduate, respectively.

The fourth category (none) for which we did not create an explicit dummy variable is called the **base category**. In Eq, when $HS = UG = PG = 0$, the value of Y is β_0, which corresponds to the education category, “none”.

The SPSS output for the regression model in Eq. using the data in above Table is shown in Table in next slide.

Example:

Table 10.14 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t-value	p-value
		B	Std. Error	Beta		
1	(Constant)	7383.333	1184.793		6.232	0.000
	High-School (HS)	5437.667	1498.658	0.505	3.628	0.001
	Under-Graduate (UG)	9860.417	1567.334	0.858	6.291	0.000
	Post-Graduate (PG)	12350.000	1675.550	0.972	7.371	0.000

The corresponding regression equation is given by

$$Y = 7383.33 + 5437.667 \times HS + 9860.417 \times UG + 12350.00 \times PG$$

Note that in Table 10.4, all the dummy variables are statistically significant $\alpha = 0.01$, since p-values are less than 0.01.

Interpretation of Regression Coefficients of Categorical Variables

In regression model with categorical variables, the regression coefficient corresponding to a specific category represents the change in the value of *Y* from the base category value (β_0).

Interaction Variables in Regression Models

- Interaction variables are basically inclusion of variables in the regression model that are a product of two independent variables (such as $X_1 X_2$).

- Usually the interaction variables are between a continuous and a categorical variable.

- The inclusion of interaction variables enables the data scientists to check the existence of conditional relationship between the dependent variable and two independent variables.

DATA ANALYTICS

Example:

The data provides salary, gender, and work experience (WE) of 30 workers in a firm. In Table gender = 1 denotes female and 0 denotes male and WE is the work experience in number of years. Build a regression model by including an interaction variable between gender and work experience.

S. No.	Gender	WE	Salary	S. No.	Gender	WE	Salary
1	1	2	6800	16	0	2	22100
2	1	3	8700	17	0	1	20200
3	1	1	9700	18	0	1	17700
4	1	3	9500	19	0	6	34700
5	1	4	10100	20	0	7	38600
6	1	6	9800	21	0	7	39900
7	0	2	14500	22	0	7	38300
8	0	3	19100	23	0	3	26900
9	0	4	18600	24	0	4	31800
10	0	2	14200	25	1	5	8000
11	0	4	28000	26	1	5	8700
12	0	3	25700	27	1	3	6200
13	0	1	20350	28	1	3	4100
14	0	4	30400	29	1	2	5000
15	0	1	19400	30	1	1	4800

Solution

Let the regression model be:

$$Y = \beta_0 + \beta_1 \times \text{Gender} + \beta_2 \times \text{WE} + \beta_3 \times \text{Gender} \times \text{WE}$$

The SPSS output for the regression model including interaction variable is given in Table

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	13443.895	1539.893		8.730	0.000
	Gender	–7757.751	2717.884	–0.348	–2.854	0.008
	WE	3523.547	383.643	0.603	9.184	0.000
	Gender*WE	–2913.908	744.214	–0.487	–3.915	0.001

Example:

The regression equation is given by

$$Y = 13442.895 - 7757.75\ \text{Gender} + 3523.547\ \text{WE} - 2913.908\ \text{Gender} \times \text{WE}$$

Equation can be written as

- For Female (Gender = 1)

$$Y = 13442.895 - 7757.75 + (3523.547 - 2913.908)\ \text{WE}$$

- For Male (Gender = 0)

$$Y = 13442.895 + 3523.547\ \text{WE}$$

That is, the change in salary for female when WE increases by one year is 609.639 and for male is 3523.547. That is the salary for male workers is increasing at a higher rate compared female workers. Interaction variables are an important class of derived variables in regression model building.

Perform Regression Model Diagnostics

F-test is used for checking the overall significance of the model whereas t-tests are used to check the significance of the individual variables. Presence of multi-collinearity can be checked through measures such as **Variance Inflation Factor** (VIF).

Validate the Model using Validation Data

The measures that can be used for validating the model in the validation data are as follows:

- R^2 or Adjusted R^2
- Mean absolute percentage error, $\sum_{i=1}^{K} \frac{1}{K} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\%$ where K is the number of cases in the validation data.
- Root Mean Square Error (RMSE), $\sqrt{\sum_{i=1}^{K} \frac{1}{n} \left(Y_i - \hat{Y}_i \right)^2}$

Validation of Multiple Regression Model (Adjusted R-square)

The following measures and tests are carried out to validate a multiple linear regression model:

- Coefficient of multiple determination (*R*-Square)

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}$$

- *SSE* is the sum of squares of errors and *SST* is the sum of squares of total deviation. In case of MLR, *SSE* will decrease as the number of explanatory variables increases, and *SST* remains constant.
- To counter this, R2 value is adjusted by normalizing both SSE and SST with the corresponding degrees of freedom. The adjusted R-square is given by
- Adjusted *R*-Square, which can be used to judge the overall fitness of the model.

$$\text{Adjusted R - Square} = 1 - \frac{\text{SSE}/(n - k - 1)}{\text{SST}/(n - 1)}$$

Statistical Significance of Individual Variables in MLR – t-test

Checking the statistical significance of individual variables is achieved through *t*-test. Note that the estimate of regression coefficient is given by Eq:

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

This means the estimated value of regression coefficient is a linear function of the response variable. Since we assume that the residuals follow normal distribution, Y follows a normal distribution and the estimate of regression coefficient also follows a normal distribution. Since the standard deviation of the regression coefficient is estimated from the sample, we use a t-test.

The null and alternative hypotheses in the case of individual independent variable and the dependent variable *Y* is given, respectively, by

- H_0: There is no relationship between independent variable X_i and dependent variable *Y*
- H_A: There is a relationship between independent variable X_i and dependent variable *Y*

Alternatively,

- H_0: $\beta_i = 0$
- H_A: $\beta_i \neq 0$

The corresponding test statistic is given by

$$t = \frac{\hat{\beta}_i - 0}{S_e(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

- *F*-test to check the statistical significance of the overall model at a given significance level (α) or at (1 – α)100% confidence level.

- Conduct a residual analysis to check whether the normality, homoscedasticity assumptions have been satisfied. Also, check for any pattern in the residual plots to check for correct model specification.

- Check for presence of multi-collinearity (strong correlation between independent variables) that can destabilize the regression model.

- Check for auto-correlation in case of time-series data.

Validation of Overall Regression Model – F-test

Analysis of Variance (ANOVA) is used to validate the overall regression model. If there are *k* independent variables in the model, then the null and the alternative hypotheses are, respectively, given by

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_1: Not all β's are zero.

F-statistic is given by: $F = MSR/MSE$

Partial F-Test

The objective of the partial *F*-test is to check where the additional variables (X_{r+1}, X_{r+2}, …, X_k) in the full model are statistically significant.

The corresponding partial *F*-test has the following null and alternative hypotheses:

- H_0: $\beta_{r+1} = \beta_{r+2} = \ldots = \beta_k = 0$
- H_1: Not all β_{r+1}, β_{r+2}, …, β_k are zero
- The partial *F*-test statistic is given by

$$\text{Partial F} = \left(\frac{(\text{SSE}_\text{R} - SSE_F)/(k-r)}{MSE_F} \right)$$

Impact of Multicollinearity

- The sign of the regression coefficient may be different, that is, instead of negative value for regression coefficient, we may have a positive regression coefficient and vice versa.

- Adding/removing a variable or even an observation may result in large variation in regression coefficient estimates.

Variance Inflation Factor (VIF)

Variance inflation factor (VIF) measures the magnitude of multi-collinearity. Let us consider a regression model with two explanatory variables defined as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

To find whether there is multi-collinearity, we develop a regression model between the two explanatory variables as follows:

$$X_1 = \alpha_0 + \alpha_1 X_2$$

Variance inflation factor (*VIF*) is then given by:

$$VIF = \frac{1}{1 - R_{12}^2}$$

The value $1 - R_{12}^2$ is called the tolerance

$\sqrt{VIF}$ is the value by which the t-statistic is deflated. So, the actual t-value is given by

$$t_{actual} = \left(\frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)} \right) \times \sqrt{VIF}$$

Remedies for Handling Multi-Collinearity

- When there are many variables in the data, the data scientists can use **Principle Component Analysis** (PCA) to avoid multi-collinearity.

- PCA will create orthogonal components and thus remove potential multi-collinearity. In the recent years, authors use advanced regression models such as **Ridge regression** and **LASSO regression** to handle multi-collinearity.

Auto-Correlation

Auto-correlation is the correlation between successive error terms in a time-series data. Consider a time-series model as defined below:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

Durbin-Watson Test for Auto-Correlation

Durbin–Watson is a hypothesis test to check the existence of auto-correlation (Durbin and Watson, 1950, . Let ρ be the correlation between error terms (ε_t, ε_{t-1}). The null and alternative hypotheses are stated below:

H0: $\rho = 0$

H1: $\rho \neq 0$

The Durbin–Watson statistic, D, for correlation between errors of one lag is given by

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \cong 2\left(1 - \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}\right)$$

The Durbin–Watson test has two critical values, D_L and D_U. The inference of the test can be made based on the following conditions:

- If $D < D_L$, then the errors are positively correlated.
- If $D > D_L$, then there is no evidence for positive auto-correlation.
- If $D_L < D < D_U$, the Durbin–Watson test is inconclusive.
- If $(4 – D) < D_L$, then errors are negatively correlated.
- If $(4 – D) > D_U$, there is no evidence for negative auto-correlation.
- If $D_L < (4 – D) < D_U$, the test is inconclusive.

Distance Measures and Outliers Diagnostics

The following distance measures are used for diagnosing the outliers and influential observations in MLR model.

- Mahalanobis Distance
  - Overcomes drawbacks of Euclindean distance
  - Helps find outliers
- Cook's Distance
  - Measures change in regression parameters
  - How does y change when a sample is left out?
- Leverage Values
  - Influence of an observation on the overall fit
- DFFIT and DFBETA Values
  - DFFIT: difference in fitted value when an observation is removed
  - SDFFit: standardized DFFit
  - DFBeta: change in regression coefficients when an observation is removed
  - DFFBeta: Standardized DFBeta

$$D_M(X_i) = \sqrt{(X_i - \mu_i)^T S^{-1} (X_i - \mu_i)}$$

$$D_i = \frac{\left(\hat{\mathbf{Y}}_\mathbf{j} - \hat{\mathbf{Y}}_{\mathbf{j(i)}}\right)^\mathbf{T}\left(\hat{\mathbf{Y}}_\mathbf{j} - \hat{\mathbf{Y}}_{\mathbf{j(i)}}\right)}{(k+1) \times MSE}$$

$$h_i = [\mathbf{H_{ii}}] = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$$

$$DFFIT = \hat{y}_i - \hat{y}_{i(i)}$$

$$SDFFIT = \frac{\hat{y}_i - \hat{y}_{i(i)}}{S_e(i)\sqrt{h_i}}$$

$$DFBETA_i(j) = \hat{\beta}_j - \hat{\beta}_{j(i)}$$

$$SDFBETA_i(j) = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{S_e(\hat{\beta}_{j(i)})}$$

Variable Selection in Regression Model Building (Forward, Backward, and Stepwise Regression)

Forward Selection

The following steps are used in building regression model using forward selection method.

Step 1: Start with no variables in the model. Calculate the correlation between dependent and all independent variables.

Step 2: Develop simple linear regression model by adding the variable for which the correlation coefficient is highest with the dependent variable (say variable X_i). (A variable can be added only when the corresponding *p*-value is less than the value α.) Let the model be $Y = \beta_0 + \beta_1 X_i$. Create a new model $Y = \alpha_0 + \alpha_1 X_i + \alpha_2 X_j$ $(j \neq i)$, there will be (k-1) such models. Conduct a partial-F test to check whether the variable X_j is statistically significant at α.

Step 3: Add the variable X_j from step 2 with smallest *p*-value based on partial *F*-test if the *p*-value is less than the significance α.

Step 4: Repeat step 3 till the smallest *p*-value based on partial *F*-test is greater than α or all variables are exhausted.

Backward Elimination Procedure

Step 1: Assume that the data has "*n*" explanatory variables. We start with a multiple regression model with all *n* variables.

That is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$. We call this full model.

Step 2: Remove one variable at a time repeatedly from the model in step 1 and create a reduced model (say model 2), there will be k such models. Perform a partial *F*-test between the models in step 1 and step 2.

Step 3: Remove the variable with largest p-value (based on partial F-test) if the p-value is greater than the significance α (or the F-value is less than the critical F-value).

Step 4 : Repeat the procedure till the p-value becomes less than α or there are no variables in the model for which the p-value is greater than α based on partial F-test.

PES UNIVERSITY ONLINE

Stepwise Regression

- Stepwise regression is a combination of forward selection and backward elimination procedure

- In this case, we set the entering criteria (α) for a new variable to enter the model based on the smallest *p*-value of the partial *F*-test and removal criteria (β) for a variable to be removed from the model if the *p*-value exceeds a pre-defined value based on the partial *F*-test ($\alpha < \beta$).

Avoiding Overfitting - Mallows's Cp

Mallows's C_p (Mallows, 1973) is used to select the best regression model by incorporating the right number of explanatory variables in the model. Mallow's C_p is given by

$$C_p = \left(\frac{SSE_p}{MSE_{full}} \right) - (n - 2p)$$

where SSE_p is the sum of squared errors with p parameters in the model (including constant), MSE_{full} is the mean squared error with all variables in the model, n is the number of observations, p is the number of parameters in the regression model including constant.

DATA ANALYTICS

Answers to some questions asked after class…

- How do we calculate VIF for more than two variables?
  - VIF for a variable can be computed as the ratio of the overall model variance to the variance of the model that includes only that independent variable OR VIF for a variable can be computed by treating that variable as a dependent variable and all others as independent variables (similar to the earlier example of a model for Promotions = $\alpha_0 + \alpha_1$CTRP + e_2)
  - High VIF for a variable ⇒ that independent variable is highly correlated with (one or more) variables in the overall model
- How does computing VIF tell us anything about multicollinearity?
  - VIF tells us how the behaviour (variance) of one independent variable is influenced (or inflated) by its interaction (or correlation) with other independent variables
  - VIF tells us how predictable (i.e., linearly dependent) an independent variable is w.r.t. other independent variables
- How is Cook's distance related to DFFit?
  - Cook's distance and DFFit are conceptually similar (difference in 'fit' (dependent variable or y_hat) with and without each sample point)

    (DFBeta is the difference (DF) in betas (each regression coefficient) with and without a sample)
- Durbin-Watson test for autocorrelation
  - When shift length (or window size) = 1 (i.e., first order autocorrelation), the statistic could range from 0 to 4:
    - D=2 implies no autocorrelation ($D_L = D_U = 2$)
    - 0 to <2 is positive autocorrelation (common in time series data)
    - >2 to 4 is negative autocorrelation (less common in time series data)
  - A **rule of thumb** is that test statistic values in the range of 1.5 to 2.5 are relatively normal. Values outside of this range could be cause for concern. It has been suggested that values less than 1 or more than 3 are a definite cause for concern.

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017 (Ch 10.1-10.19.1)

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

DATA ANALYTICS

Unit 2: Regression – Multiple and Multivaraiate

Mamatha.H.R

Department of Computer Science and Engineering

DATA ANALYTICS

Transformations

Transformation is a process of deriving new dependent and/or independent variables to identify the correct functional form of the regression model

Transformation in MLR is used to address the following issues:

- Poor fit (low R^2 value).
- Patten in residual analysis indicating potential non-linear relationship between the dependent and independent variable

  For example, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is used for developing the model

  instead of $\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, resulting in a clear pattern in residual plot
- Residuals do not follow a normal distribution
- Residuals are not homoscedastic

Example

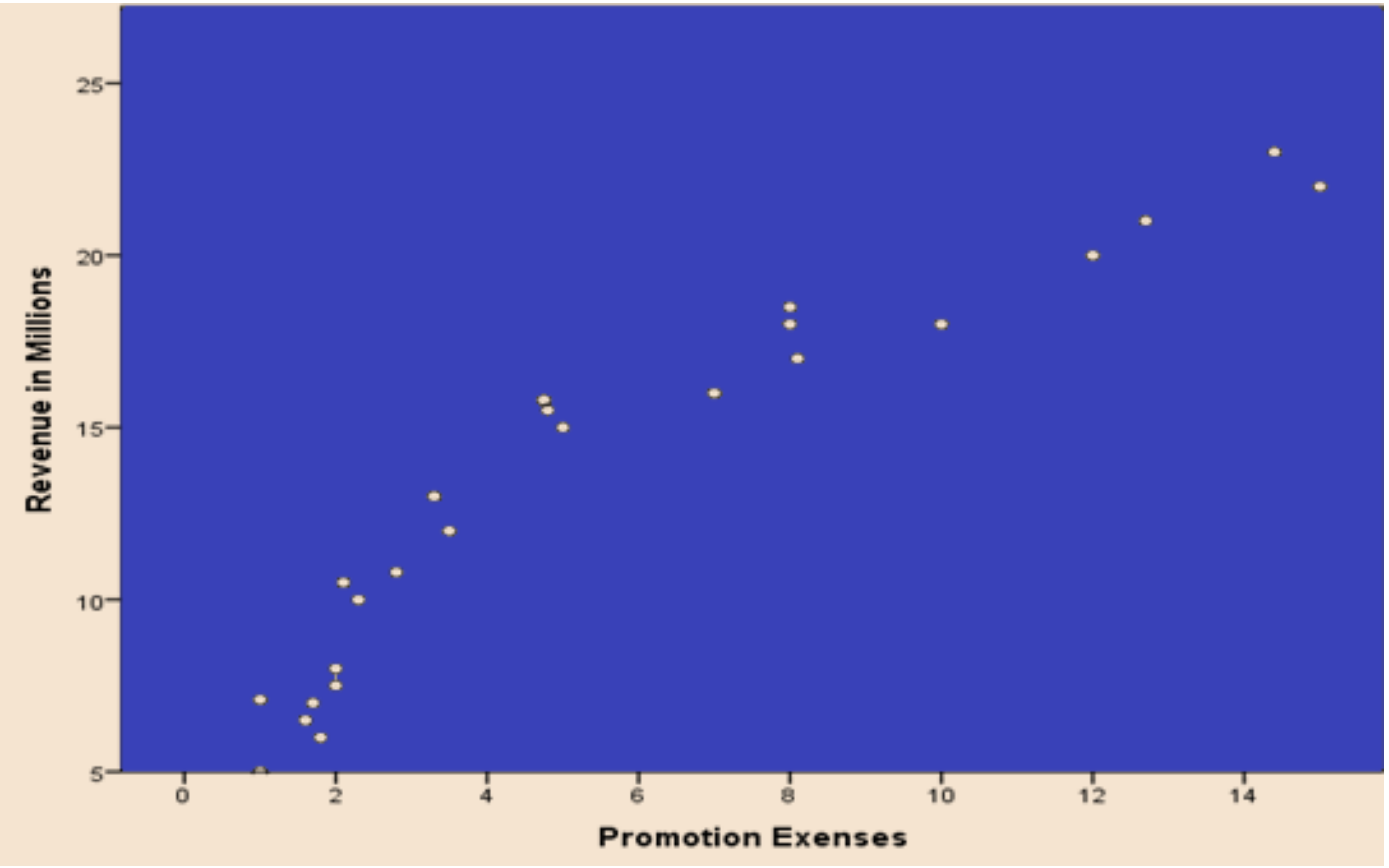
Table shows the data on revenue generated (in million of rupees) from a product and the promotion expenses (in million of rupees). Develop a regression model

S. No.	Revenue in Millions	Promotion Expenses	S. No.	Revenue in Millions	Promotion Expenses
1	5	1	13	16	7
2	6	1.8	14	17	8.1
3	6.5	1.6	15	18	8
4	7	1.7	16	18	10
5	7.5	2	17	18.5	8
6	8	2	18	21	12.7
7	10	2.3	19	20	12
8	10.8	2.8	20	22	15
9	12	3.5	21	23	14.4
10	13	3.3	22	7.1	1
11	15.5	4.8	23	10.5	2.1
12	15	5	24	15.8	4.75

Motivating Transformations

Let *Y* = Revenue Generated and *X* = Promotion Expenses

The scatter plot between Y and X for the data in Table is shown in Figure. It is clear from the scatter plot that the relationship between X and Y is not linear; it looks more like a logarithmic function.

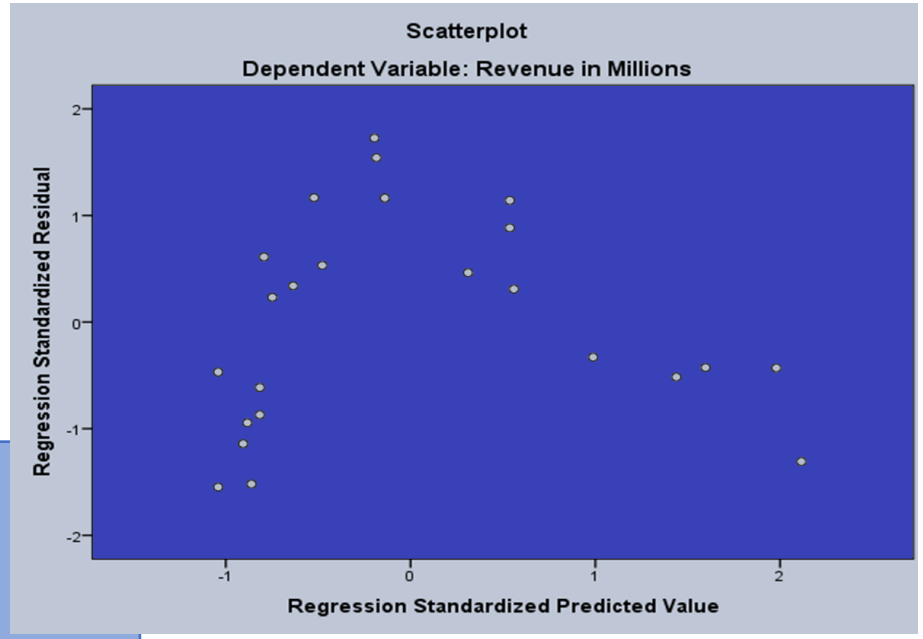


Pre-Transformation: Pattern in Residuals

Consider the function $Y = \beta_0 + \beta_1 X$. The output for this regression is shown below. There is a clear increasing and decreasing pattern in Figure indicating non-linear relationship between *X* and *Y*.

Model Summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.940	0.883	0.878	1.946



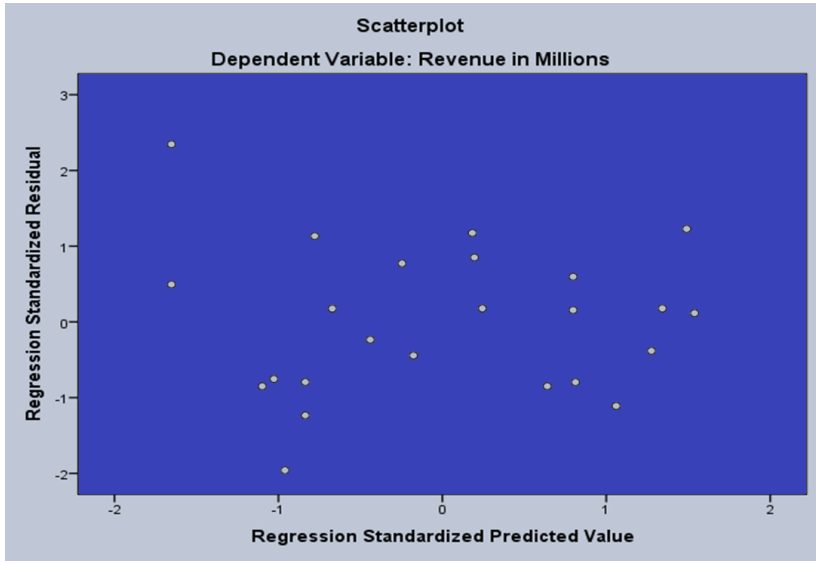
Coefficients

		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	T	Sig.
1	(Constant)	6.831	0.650		10.516	0.000
	Promotion Expenses	1.181	0.091	0.940	12.911	0.000

Post Transformation: No Pattern in Residuals

Since there is a pattern in the residual plot, we cannot accept the linear model ($Y = \beta_0 + \beta_1 X$).

Next we try the model $Y = \beta_0 + \beta_1 \ln(X)$. The SPSS output for $Y = \beta_0 + \beta_1 \ln(X)$ and the residual plot are shown.



Model Summary

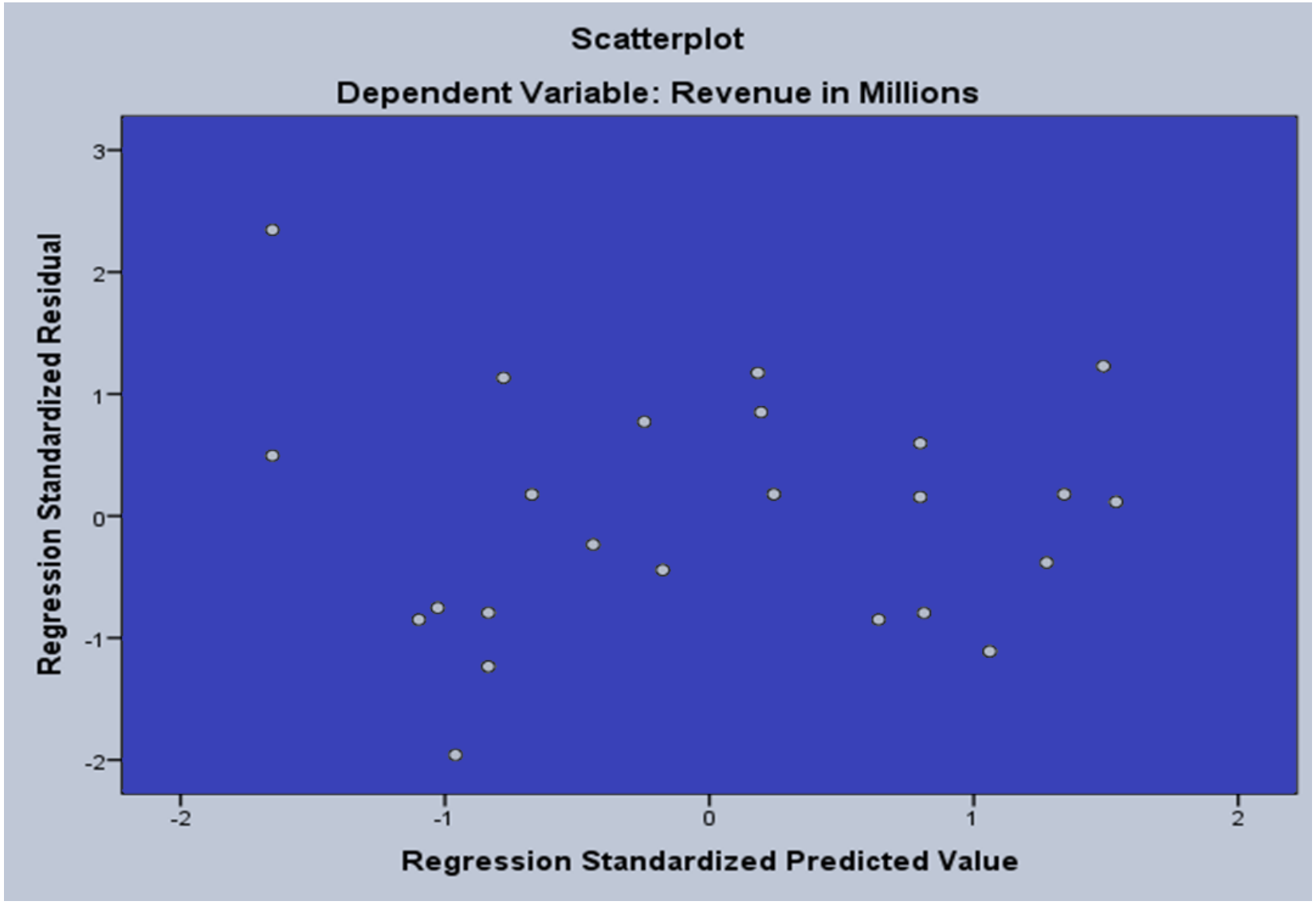
Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.980	0.960	0.959	1.134

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.439	0.454		9.771	0.000
	ln (X)	6.436	0.279	0.980	23.095	0.000

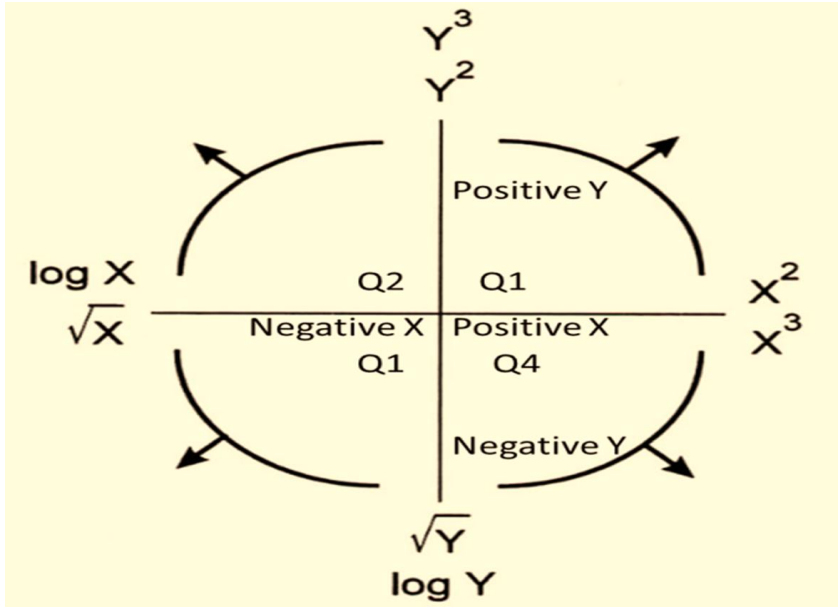
Note that for the model $Y = \beta_0 + \beta_1 \ln(X)$, the R^2-value is 0.96 whereas the R^2-value for the model $Y = \beta_0 + \beta_1 X$ is 0.883. Most important, there is no obvious pattern in the residual plot of the model $Y = \beta_0 + \beta_1 \ln(X)$. The model $Y = \beta_0 + \beta_1 \ln(X)$ is preferred over the model $Y = \beta_0 + \beta_1 X$.

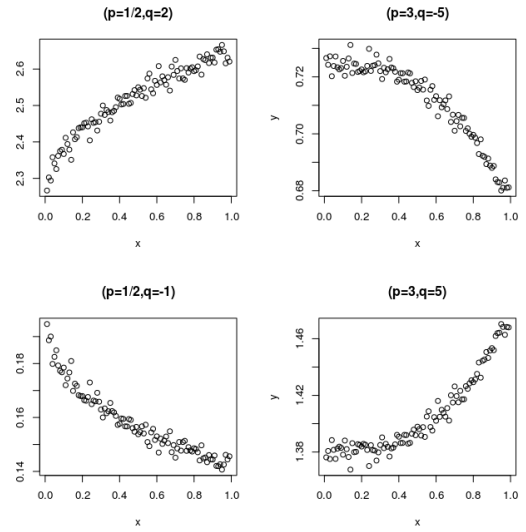
Residual plot for the model Y = β_0 + $\beta_1$ln(X).



Tukey and Mosteller's Bulging Rule for Transformation

- An easier way of identifying an appropriate transformation was provided by Mosteller and Tukey (1977), popularly known as Tukey's Bulging Rule.
- To apply Tukey's Bulging Rule we need to look at the pattern in the scatter plot between the dependent and independent variable.





Shape of Scatter Plot	Suggested Transformation for X	Suggested Transformation for Y
Q1 (*X* and *Y* positive)	X^p where $p > 1$ (e.g. X^2, X^3, etc.)	Y^q where $q > 1$ (e.g. Y^2, Y^3, etc.)
Q2 (*X* negative and *Y* positive)	X^p where $p < 1$ (e.g., $\ln(X), \sqrt{X}$, etc.)	Y^q where $q > 1$ (e.g. Y^2 and Y^3 etc)
Q3 (Both *X* and *Y* negative)	X^p where $p < 1$ (e.g. $\ln(X), \sqrt{X}$, etc.)	Y^q where $q < 1$ (e.g. $\ln(Y), \sqrt{Y}$, etc.)
Q4 (*X* positive and *Y* negative)	X^p where $p > 1$ (e.g. X^2, X^3, etc.)	Y^q where $q < 1$ (e.g. $\ln(Y), \sqrt{Y}$, etc.)

https://freakonometrics.hypotheses.org/14967

DATA ANALYTICS

MvLR Model: Scalar Form

Multivariate Regression (MvLR): Predict multiple dependent variables using multiple independent variables

The multivariate (multiple) linear regression model has the form

$$y_{ik} = b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij} + e_{ik}$$

for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$ where

- $y_{ik} \in \mathbb{R}$ is the k-th real-valued response for the i-th observation
- $b_{0k} \in \mathbb{R}$ is the regression intercept for k-th response
- $b_{jk} \in \mathbb{R}$ is the j-th predictor's regression slope for k-th response
- $x_{ij} \in \mathbb{R}$ is the j-th predictor for the i-th observation
- $(e_{i1}, \ldots, e_{im}) \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \mathbf{\Sigma})$ is a multivariate Gaussian error vector

MvLR Model: Assumptions

The fundamental assumptions of the MLR model are:

1. Relationship between X_j and Y_k is linear (given other predictors)
2. x_{ij} and y_{ik} are observed random variables (known constants)
3. $(e_{i1}, \ldots, e_{im}) \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \mathbf{\Sigma})$ is an unobserved random vector
4. $\mathbf{b}_k = (b_{0k}, b_{1k}, \ldots, b_{pk})'$ for $k \in \{1, \ldots, m\}$ are unknown constants
5. $(y_{ik}|x_{i1}, \ldots, x_{ip}) \sim \mathrm{N}(b_{0k} + \sum_{j=1}^{p} b_{jk}x_{ij}, \sigma_{kk})$ for each $k \in \{1, \ldots, m\}$
   note: homogeneity of variance for each response

Note: b_{jk} is expected increase in Y_k for 1-unit increase in X_j with all other predictor variables held constant

MLR Model: Matrix Form

The multivariate multiple linear regression model has the form

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where

- $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ response matrix
  - $\mathbf{y}_k = (y_{1k}, \ldots, y_{nk})' \in \mathbb{R}^n$ is k-th response vector $(n \times 1)$
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times (p+1)}$ is the $n \times (p+1)$ design matrix
  - $\mathbf{1}_n$ is an $n \times 1$ vector of ones
  - $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})' \in \mathbb{R}^n$ is j-th predictor vector $(n \times 1)$
- $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m] \in \mathbb{R}^{(p+1) \times m}$ is $(p+1) \times m$ matrix of coefficients
  - $\mathbf{b}_k = (b_{0k}, b_{1k}, \ldots, b_{pk})' \in \mathbb{R}^{p+1}$ is k-th coefficient vector $(p+1 \times 1)$
- $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ error matrix
  - $\mathbf{e}_k = (e_{1k}, \ldots, e_{nk})' \in \mathbb{R}^n$ is k-th error vector $(n \times 1)$

MLR Model: Matrix Form

Matrix form writes MLR model for all *nm* points simultaneously

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ y_{31} & \cdots & y_{3m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_{01} & \cdots & b_{0m} \\ b_{11} & \cdots & b_{1m} \\ b_{21} & \cdots & b_{2m} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ e_{21} & \cdots & e_{2m} \\ e_{31} & \cdots & e_{3m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

Fitted Values and Residuals

SCALAR FORM:

Fitted values are given by

$$\hat{y}_{ik} = \hat{b}_{0k} + \sum_{j=1}^{p} \hat{b}_{jk} x_{ij}$$

and residuals are given by

$$\hat{e}_{ik} = y_{ik} - \hat{y}_{ik}$$

MATRIX FORM:

Fitted values are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

and residuals are given by

$$\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Hat Matrix

Note that we can write the fitted values as

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\mathbf{B}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix.

$\mathbf{H}$ is a symmetric and idempotent matrix: $\mathbf{H}\mathbf{H} = \mathbf{H}$

$\mathbf{H}$ projects $\mathbf{y}_k$ onto the column space of $\mathbf{X}$ for $k \in \{1, \ldots, m\}$.

PES UNIVERSITY ONLINE

DATA ANALYTICS

Unit 2: Other forms of regression
Ridge, lasso and polynomial
Nonlinear regression

Mamatha H R, Gowri Srinivasa
Department of Computer Science and Engineering

Bias-Variance Trade-Off in Multiple Regression

The simple linear regression model, in which you aim at predicting n observations of the response variable, Y, with a linear combination of m predictor variables, X, and a normally distributed error term with variance σ^2:

$$Y = X\beta + \varepsilon,$$

$$\varepsilon \sim N(0, \sigma^2).$$

The true parameters, β, are not known we have to estimate them from the sample. In the Ordinary Least Squares (OLS) approach, we estimate them as β^ in such a way, that the sum of squares of residuals is as small as possible.

Bias-Variance Trade-Off in Multiple Regression

In other words, we minimize the following loss function:

$$L_{OLS}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 = ||y - X\hat{\beta}||^2$$

In order to obtain the infamous OLS parameter estimates,

$$\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y).$$

DATA ANALYTICS

Bias-Variance Trade-Off in Multiple Regression

In statistics, there are two critical characteristics of estimators to be considered: the bias and the variance.

The bias is the difference between the true population parameter and the expected estimator: $Bias(\hat{\beta}_{OLS}) = E(\hat{\beta}_{OLS}) - \beta.$

It measures the accuracy of the estimates. Variance, on the other hand, measures the spread, or uncertainty, in these estimates. It is given by

$$Var(\hat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1},$$

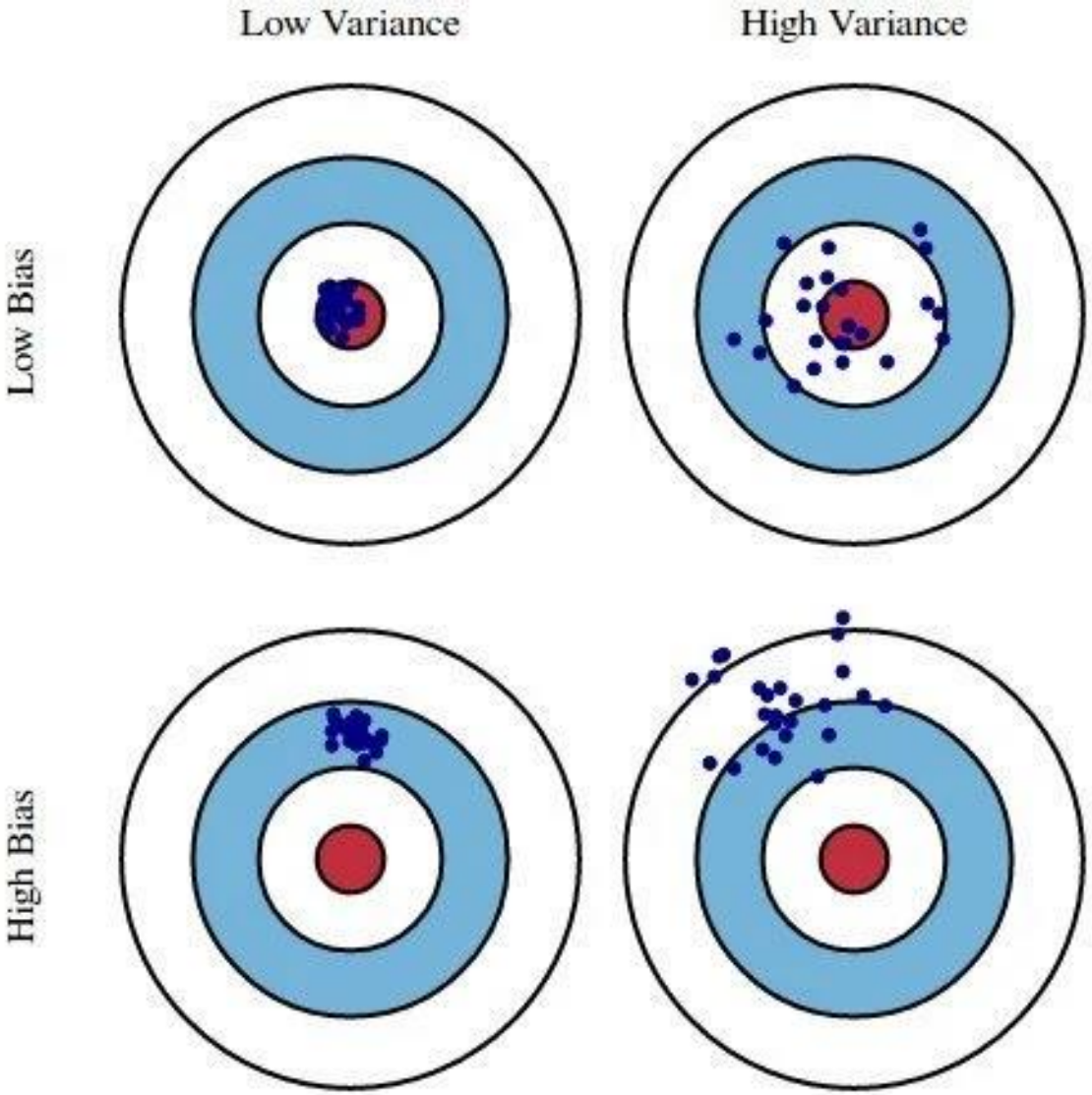
where the unknown error variance σ^2 can be estimated from the residuals as

$$\hat{\sigma}^2 = \frac{e'e}{n-m},$$

$$e = y - X\hat{\beta}.$$

DATA ANALYTICS

Illustration of Bias and Variance



Imagine the bull's-eye is the true population parameter that we are estimating,
β, and the shots at it are the values of our estimates resulting from four different estimators - low bias and variance, high bias and variance, and the combinations.

Bias-Variance Trade-Off in Multiple Regression

Both the bias and the variance are desired to be low, as large values result in poor predictions from the model.

- In fact, the model's error can be decomposed into three parts:
  - error resulting from a large variance,
  - error resulting from significant bias,
  - and the remainder - the unexplainable part.

$$E(e) = (E(X\hat{\beta}) - X\beta)^2 + E(X\hat{\beta} - E(X\hat{\beta}))^2 + \sigma^2 = Bias^2 + Variance + \sigma^2$$

Bias-Variance Trade-Off in Multiple Regression

The OLS estimator has the desired property of being unbiased. However, it can have a huge variance.

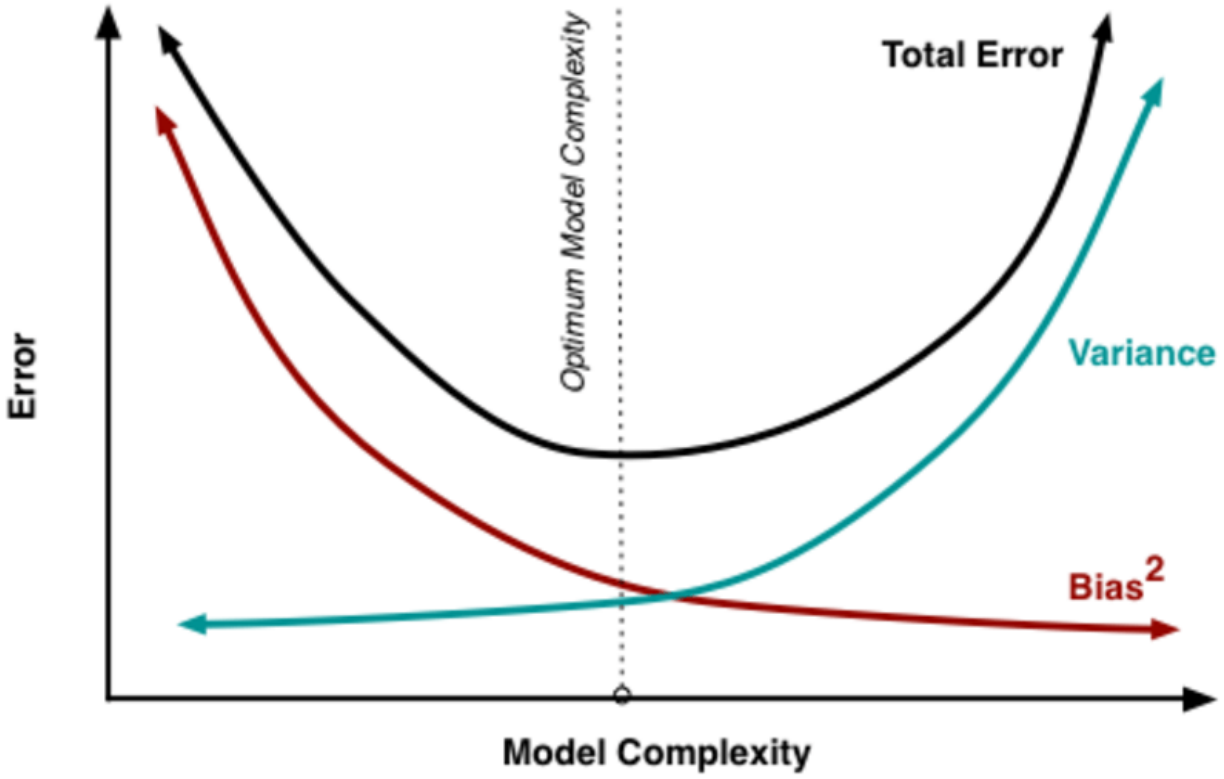
Specifically, this happens when:

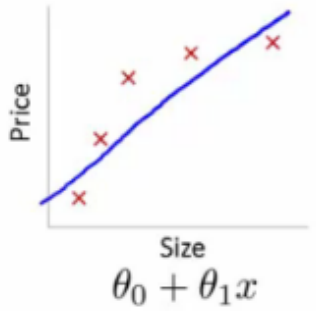
- The predictor variables are highly correlated with each other
- There are many predictors.
- If m approaches n, the variance approaches infinity $\hat{\sigma}^2 = \frac{e'e}{n-m}$,

- The general solution to this is:
  reduce variance at the cost of introducing some bias
- This approach is called regularization and is almost always beneficial for the predictive performance of the model

DATA ANALYTICS

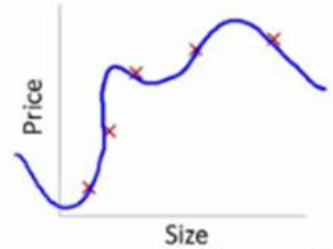
Bias-Variance Trade-Off in Multiple Regression





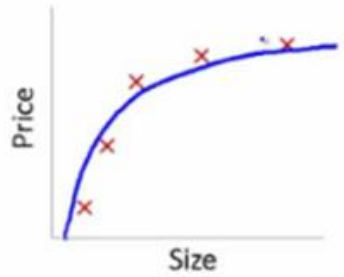
$\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

What do Lasso and Ridge Regression do?

To decrease the model complexity, that is the number of predictors.

We could use the forward or backward selection for this, but that way we would not be able to tell anything about the removed variables' effect on the response.

Removing predictors from the model can be seen as settings their coefficients to zero (Lasso).

Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way (Ridge).

This way, we decrease model complexity while keeping all variables in the model.

DATA ANALYTICS

Lasso Regression

- LASSO: Least absolute shrikange and selection
- Assumptions same as linear regression, normality not assumed
- Uses L_1 norm or the 'absolute value' of coefficients scaled by shrinkage
- λ is a tunable parameter
- Lasso tends to zero out smaller (unimportant) coefficients (and helps with feature selection)

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

Limitations:

• In small-n-large-p dataset the LASSO selects at most n variables before it saturates.

• If there are grouped variables (highly correlated between each other) LASSO tends to select one variable from each group ignoring the others

Ridge Regression

- Assumptions same as linear regression, normality not assumed
- A shrinkage term is added to the objective (SSE) function
- λ is a tunable parameter; penalizes flexibility of the model
- We shrink the estimated association of each variable
- $\lambda=0$ has no effect and as $\lambda\to\infty$ and ridge regression coefficient estimates approach 0

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

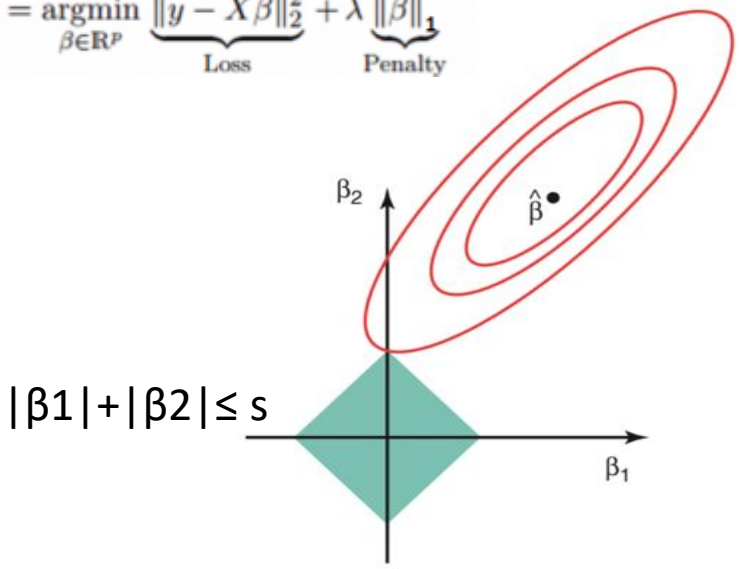
- Coefficients produced by OLS are scale invariant but that is not the case with Ridge Regression, so we must remember to scale the input

$$\tilde{x}_{ij}=\frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2}}$$

DATA ANALYTICS

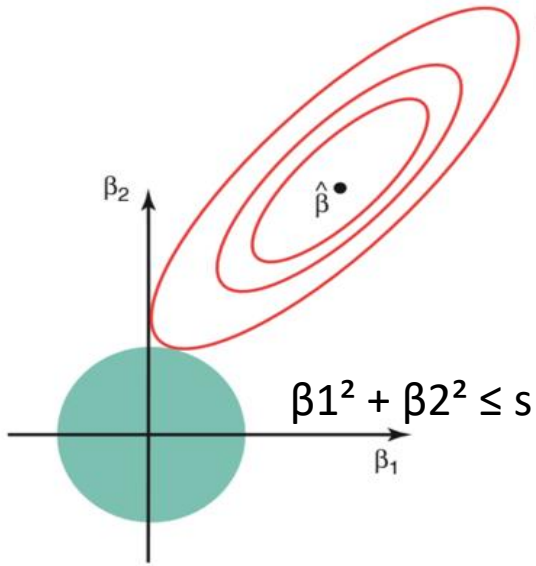
Lasso Vs Ridge

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$



$|\beta 1| + |\beta 2| \leq s$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$



$\beta 1^2 + \beta 2^2 \leq s$

- Correlated variables have similar weights using Ridge and whereas one is high and the other(s) nearly zero with Lasso
- Interpretability: Lasso zeros out unimportant coefficients and hence performs feature selection whereas ridge gives a small weight but includes them all
- Both achieve reduction in variance without increase in bias

https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

RIDGE REGRESSION

- Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated).
- In this phenomenon, one predicted value in multiple regression models is linearly predicted with others to attain a certain level of accuracy.
- The concept multicollinearity occurs when there are high correlations between more than two predicted variables.
- In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value.
- By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Choice of Regularization Parameter

Question : how much bias are we willing to accept in order to decrease the variance? Or: what is the optimal value for λ?

- Choose λ such that some information criterion, e.g., AIC or BIC, is the smallest. approach emphasizes the model's fit to the data
- This approach boils down to estimating the model with many different values for λ and choosing the one that minimizes the Akaike or Bayesian Information Criterion:

$$AIC_{ridge} = nlog(e'e) + 2df_{ridge},$$
$$BIC_{ridge} = nlog(e'e) + 2df_{ridge}log(n),$$

where df_{ridge} is the number of degrees of freedom

- A more machine learning-like approach is to perform cross-validation and select the value of λ that minimizes the cross-validated sum of squared residuals.

ElasticNet Regression: Combining L1 and L2 norms

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

- Inherits Ridge's stability under rotation
- It encourages group effect in case of highly correlated variables
- There are no limitations on the number of selected variables
- It can suffer with double shrinkage

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

References

https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-

https://www.statisticshowto.com/lasso-regression

https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

Polynomial Function: Definition

Reminder: a polynomial function has the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$$
$$= \sum_{j=0}^{n} a_j x^j$$

where $a_j \in \mathbb{R}$ are the coefficients and x is the indeterminate (variable).

Note: x^j is the j-th order polynomial term

- x is first order term, x^2 is second order term, etc.
- The degree of a polynomial is the highest order term

Model Form (scalar)

The polynomial regression model has the form

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_i^j + e_i$$

for $i \in \{1, \ldots, n\}$ where

- $y_i \in \mathbb{R}$ is the real-valued response for the i-th observation
- $b_0 \in \mathbb{R}$ is the regression intercept
- $b_j \in \mathbb{R}$ is the regression slope for the j-th degree polynomial
- $x_i \in \mathbb{R}$ is the predictor for the i-th observation
- $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term

Model Form (matrix)

The polynomial regression model has the form

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ 1 & x_3 & x_3^2 & \cdots & x_3^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix}$$

Note that this is still a linear model, even though we have polynomial terms in the design matrix.

PR Model Assumptions (scalar)

The fundamental assumptions of the PR model are:

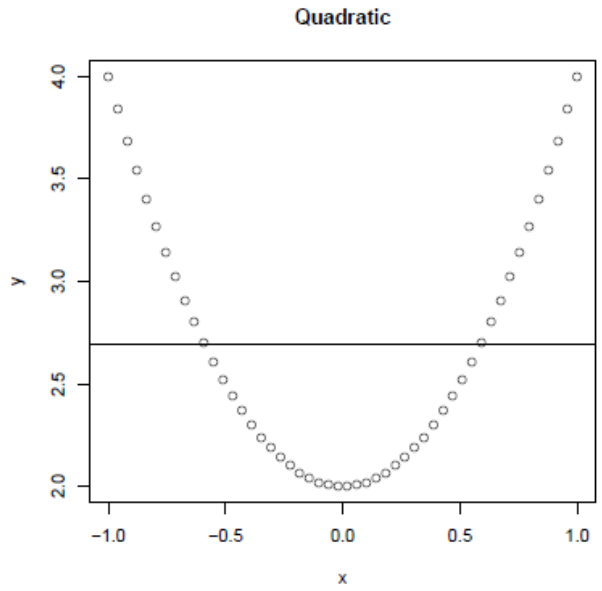
1. Relationship between X and Y is polynomial
2. x_i and y_i are observed random variables (known constants)
3. $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is an unobserved random variable
4. $b_0, b_1, \ldots, b_p$ are unknown constants
5. $(y_i|x_i) \overset{\text{ind}}{\sim} \mathrm{N}(b_0 + \sum_{j=1}^{p} b_j x_i^j, \sigma^2)$
   note: homogeneity of variance

Note: focus is estimation of the polynomial curve.

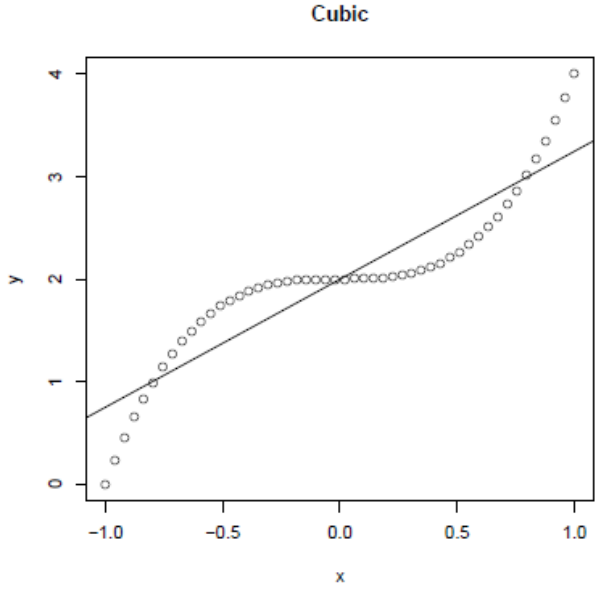
DATA ANALYTICS

Polynomial Function: Simple Regression

```
> x=seq(-1,1,length=50)
> y=2+2*(x^2)
> plot(x,y,main="Quadratic")
> qmod=lm(y~x)
> abline(qmod)
```




```
> x=seq(-1,1,length=50)
> y=2+2*(x^3)
> plot(x,y,main="Cubic")
> cmod=lm(y~x)
> abline(cmod)
```



Polynomial Regression: Properties

Some important properties of the PR model include:

1. Need $n > p$ to fit the polynomial regression model
2. Setting $p = 1$ produces simple linear regression
3. Setting $p = 2$ is quadratic polynomial regression
4. Setting $p = 3$ is cubic polynomial regression
5. Rarely set $p > 3$; use cubic spline instead

Multicollinearity: Problem

Note that x_i, x_i^2, x_i^3, etc. can be highly correlated with one another, which introduces multicollinearity problem.

```
> set.seed(123)
> x = runif(100)*2
> X = cbind(x, xsq=x^2, xcu=x^3)
> cor(X)
          x       xsq       xcu
x   1.0000000 0.9703084 0.9210726
xsq 0.9703084 1.0000000 0.9866033
xcu 0.9210726 0.9866033 1.0000000
```

Multicollinearity: Partial Solution

You could mean-center the x_i terms to reduce multicollinearity.

```
> set.seed(123)
> x = runif(100)*2
> x = x - mean(x)
> X = cbind(x, xsq=x^2, xcu=x^3)
> cor(X)
                x         xsq         xcu
x   1.00000000 0.03854803 0.91479660
xsq 0.03854803 1.00000000 0.04400704
xcu 0.91479660 0.04400704 1.00000000
```

But this doesn't fully solve our problem. . .

Orthogonal Polynomials: Definition

To deal with multicollinearity, define the set of variables

$$z_0 = a_0$$
$$z_1 = a_1 + b_1 x$$
$$z_2 = a_2 + b_2 x + c_2 x^2$$
$$z_3 = a_3 + b_3 x + c_3 x^2 + d_3 x^3$$

where the coefficients are chosen so that $z_j' z_k = 0$ for all $j \neq k$.

The transformed z_j variables are called orthogonal polynomials.

How to detect non linearity ?

1. Theory – in many sciences we have theories about nonlinear relations within some phenomenon.

2. Scatterplot – when looking at the plot you can see that data points are not linear or even not nearly linear.

3. Seasonality in data – i.e. in agriculture, building industry we often have seasonality within the data.

4. Estimated model does not fit the data well or does not fit it at all; the estimated β's are not significant – this might suggest nonlinearity.

5. Can often do incremental F tests

NON LINEAR REGRESSION

Central idea of non-linear regression: same as linear regression, just with non-linear features

Two ways to construct non-linear features

1. Explicitly (construct actual feature vector)
2. Implicitly (using kernels)

E.g. $\phi(x_i) = \begin{bmatrix} x_i^2 \\ x_i \\ 1 \end{bmatrix}$

DATA ANALYTICS

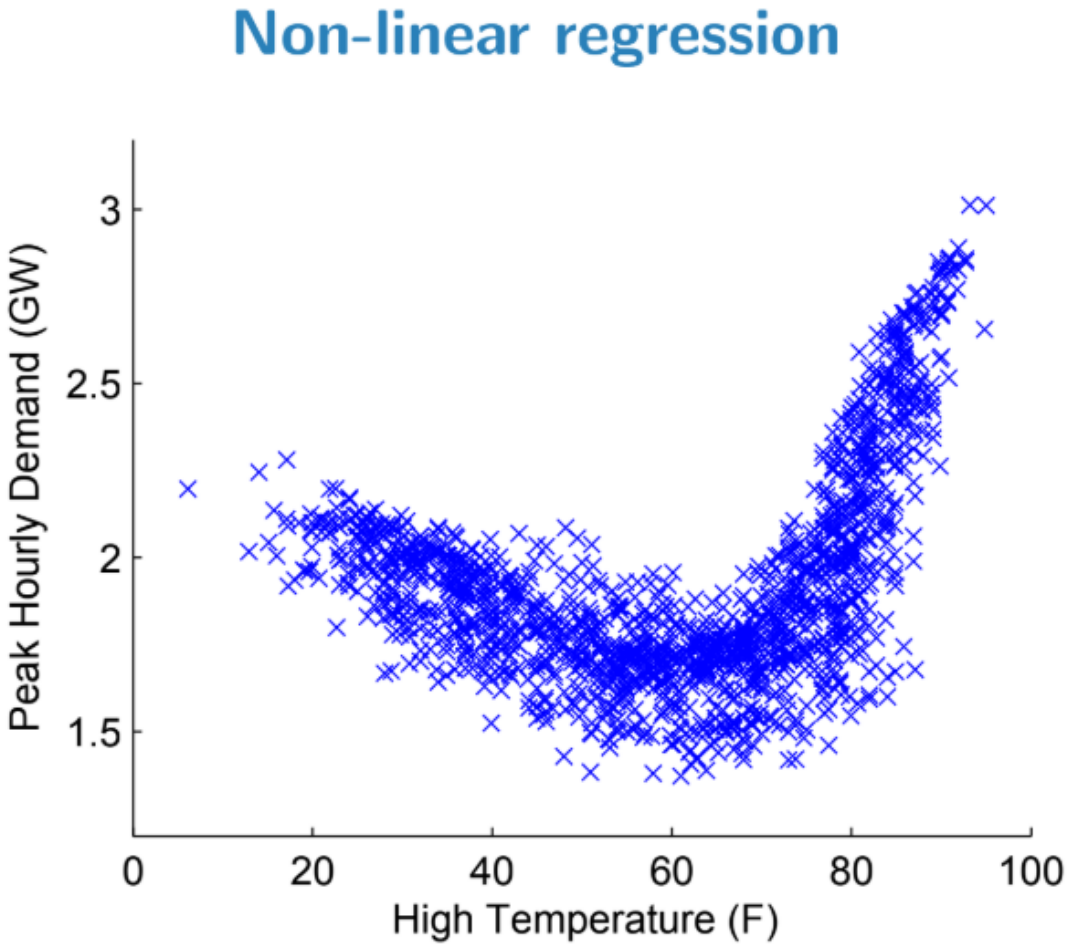
NON LINEAR REGRESSION

Some popular nonlinear regression models:

1. Exponential model: $(y = ae^{bx})$
2. Power model: $(y = ax^{b})$
3. Saturation growth model: $\left(y = \frac{ax}{b+x}\right)$

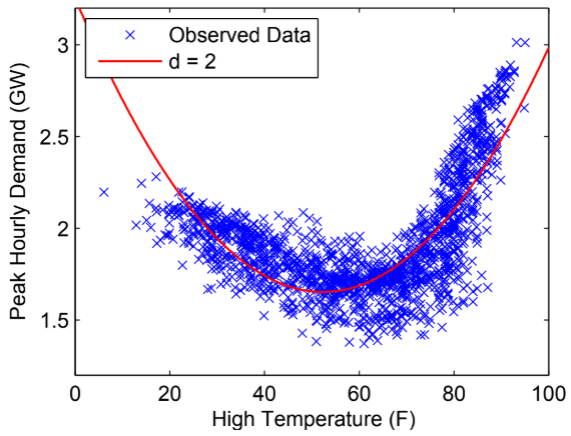
DATA ANALYTICS

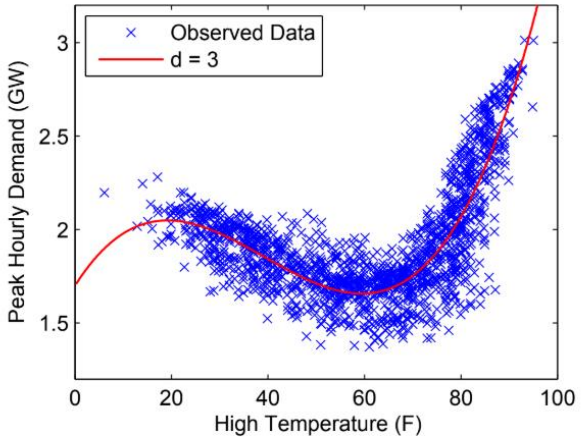
NON LINEAR REGRESSION

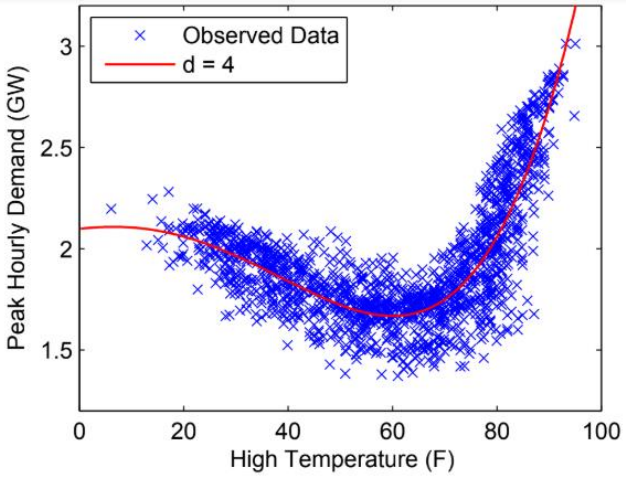


DATA ANALYTICS

NON LINEAR REGRESSION







Degree 4 polynomial

NON LINEAR REGRESSION

Why use non-linear regression?

Transformation is necessary to obtain variance homogeneity, but transformation destroys linearity.

Linearity does not fit, and the transformation seems to destroy other parts of the model assumptions, e.g. the assumption of variance homogeneity.

Theoretical knowledge indicates that the proper relation is intrinsically non-linear.

Interest is in functions of the parameters that do not enter linearly in the model

DATA ANALYTICS

Some questions asked after class

- **Is there proof we are reducing total error if we add bias?**

- While there is no 'proof' other than our empirical understanding of the total error (which is convex and) comprises the bias, variance and 'unexplained' residuals (or 'noise)

  We can rephrase this question as: is there anyway to detect high bias or high variance? The answer is 'yes' and there are ways of remedying these problems as well:

- **Detection of high variance (Regime #1)**

  1.Training error is much lower than test error

  2.Training error is lower than some desired tolerance 't'

  3.Test error is above 't'

  Remedies

  a. Add more training data

  b. Reduce model complexity -- complex models are prone to high variance

  c. Bagging (will be taught in the course on Machine Intelligence)

- **Detection of high bias in the model (Regime #2)**

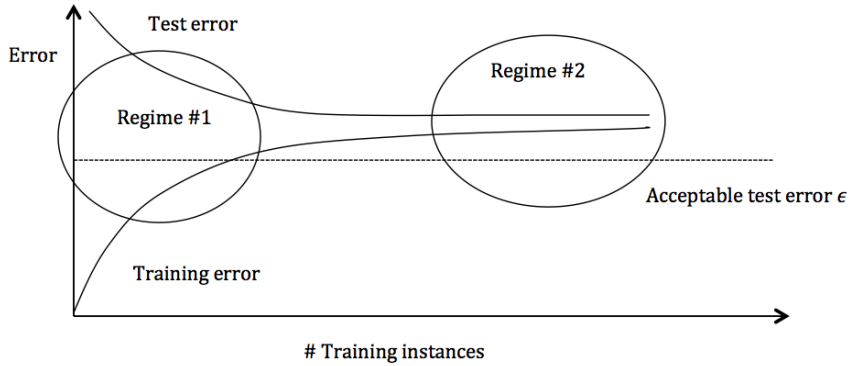
  1. Training error is higher than a desired tolerance 't'

  Remedies

  a. Use more complex model (e.g. kernelize, use non-linear models)

  b. Add features

  c. Boosting (will be covered later in the course)



https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html

- **How do we know the error function is indeed convex?**

  The error function (SSE) is a quadratic function in one variable and hence, convex

https://towardsdatascience.com/understanding-convexity-why-gradient-descent-works-for-linear-regression-aaf763308708

https://www.princeton.edu/~aaa/Public/Teaching/ORF523/S16/ORF523_S16_Lec7_gh.pdf

DATA ANALYTICS

Some questions asked after class

- **Can we minimize both bias and variance using Ridge/ Lasso**
- Please note: Both bias and variance cannot be arbitrarily minimized
- We are *adding* bias to minimize the overall error (i.e., reduce the high variance)

- **Can the shrinkage parameter (lambda) be positive, negative or zero?**
- In theory yes; though let us see what each option for the shrinkage value would mean:
- Zero: same as simple linear regression (bias would be low (or zero) $\Rightarrow$ variance is very high)
- Positive: we are reducing the bias; if the overall equation is to be minimized, then larger the λ, the closer β's would get to zero
- Negative: if lambda is negative, our loss can be reduced indefinitely; it means (some of the) β's$\rightarrow\infty$
- So, in practice, lambda is range limited (i.e., we insist on $\lambda>0$)

References

Additional references (for the interested student)

http://users.stat.umn.edu/~helwig/notes/polyint-Notes.pdf

http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

DATA ANALYTICS

Unit 2:Logistic Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Logistic Regression

Mamatha H R, Gowri Srinivasa
Department of Computer Science and Engineering

DATA ANALYTICS

Classification Problems

- Classification is an important category of problems in which the decision maker would like to classify the case/entity/customers into two or more groups

- Examples of Classification Problems:

  - Customer profiling (customer segmentation)
  - Customer Churn
  - Credit Classification (low, high and medium risk)
  - Employee attrition
  - Fraud (classification of transaction to fraud/no-fraud)
  - Stress levels
  - Text Classification (Sentiment Analysis)
  - Outcome of any binomial and multinomial experiment

DATA ANALYTICS

ODDS and ODDS RATIO

We will consider a data-set that tells us about depending on the gender, whether a customer will purchase a product or not

Gender	Purchase	
	Yes	No
Female	159	106
Male	121	125

Odds, which describes the ratio of success to ratio of failure

Considering females group,

- we see that probability that a female will purchase (success) the product is = 159/265 (yes/total number of females).
- Probability of failure (no purchase) for female is 106/265.
- In this case the odds is defined as (159/265)/(106/265) = 1.5.
- Higher the odds, better is the chance for success. Range of odds can be any number between [0 , ∞].

DATA ANALYTICS

ODDS and ODDS RATIO

Odds ratio, is the ratio of odds.

Considering the example, Odds ratio, represents which group (male/female) has better odds of success, and it's given by calculating the ratio of odds for each group.

$$odds = \frac{\pi}{1-\pi}$$

So odds ratio for females = odds of successful purchase by female / odds of successful purchase by male = (159/106)/(121/125).

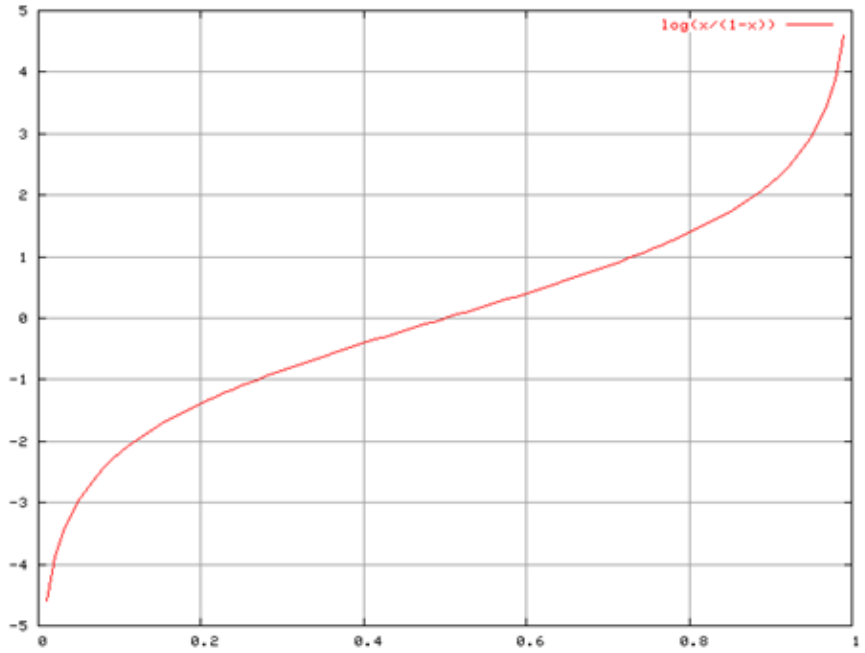
$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}$$

Odds ratio for males will be the reciprocal of the above number.

Odds ratio can vary between 0 to positive infinity, log (odds ratio) will vary between [-∞, ∞]

Logit Function

- The logit function is the logarithmic transformation of the logistic function. It is defined as the natural logarithm of odds.
- Logit of a variable π (with value between 0 and 1) is given by:



$$Logit(\pi) = \ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x$$

Logistic Transformation

- The logistic regression model is given by:

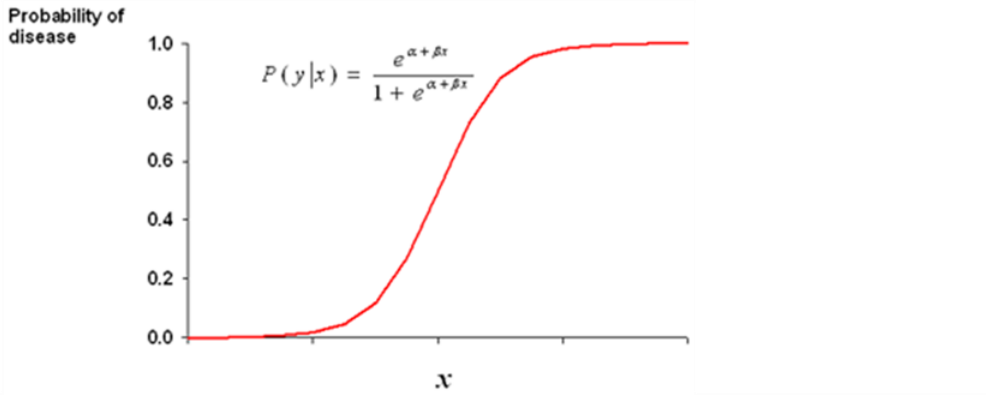
$$\pi_i = \frac{e^{(\beta_0+\beta_1 X_i)}}{1+e^{(\beta_0+\beta_1 X_i)}}$$

$$\frac{\pi_i}{1-\pi_i} = e^{(\beta_0+\beta_1 X_i)}$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

Function with linear properties

$$P(Y=1 \mid X=x) = \pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$



β = 0 implies that P(Y|x) is same for each value of x
β > 0 implies that P(Y|x) is increases as the value of x increases
β < 0 implies that P(Y|x) is decreases as the value of x increases

Odds Ratio for Binary Logistic Regression

$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}$$

If OR = 2, then the event is twice likely to occur when X = 1 compared to X = 0.

Odds ratio approximates the relative risk.

Parameter Estimation in Logistic Regression (Maximum Likelihood Estimate)

Likelihood function for Binary Logistic Function

- Probability density function for binary logistic regression is given by:

$$P(Y=1|Z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)=\pi(Z)=\frac{e^Z}{(1+e^Z)} \qquad (11.7)$$

The probability (likelihood) function of binary logistic regression for specific observation Y_i (Y_i=0 or 1) is given by

$$P(Y_i)=\pi(Z)^{Y_i}(1-\pi(Z))^{1-Y_i} \qquad (11.8)$$

DATA ANALYTICS

Estimation of parameters

Assume that the data set has n observations, $Y_1, Y_2, \ldots, Y_n$. The likelihood function, which is a joint probability, $L(Y_1, Y_2, \ldots, Y_n)$ for a specific Z_i $(= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi})$ is given by

$$L = P(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} \pi(Z_i)^{Y_i} [1 - \pi(Z_i)]^{1-Y_i} \tag{11.9}$$

The log-likelihood function is given by

$$\ln(L) = LL = \sum_{i=1}^{n} Y_i \ln[\pi(Z_i)] + \sum_{i=1}^{n} (1 - Y_i)[\ln(1 - \pi(Z_i))] \tag{11.10}$$

For mathematical simplicity, assume that $Z_i = \beta_0 + \beta_1 X_i$. Equation (11.10) can be written as

$$LL(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \ln[1 + \exp(\beta_0 + \beta_1 X_i)] \tag{11.11}$$

Taking partial derivatives with respect to β_0 and β_1 and equating them to zero, we get the following first-order conditions (Hosmer and Lemeshow, 2000; Kleinbaum and Klein, 2011):

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_0} = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = 0 \tag{11.12}$$

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_1} = \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \frac{X_i \exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = 0 \tag{11.13}$$

Limitations of MLE

- Maximum likelihood estimator may not be unique or may not exist.

- No closed form solution $\Rightarrow$ use iterative procedure to estimate the parameter values.

Interpretation of LR coefficients

- β_1 is the change in log-odds ratio for unit change in the explanatory variable.

- β_1 is the change in odds ratio by a factor exp(β_1).

$$\beta_1 = \ln\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))}\right) = \text{Change in ln odds ratio}$$

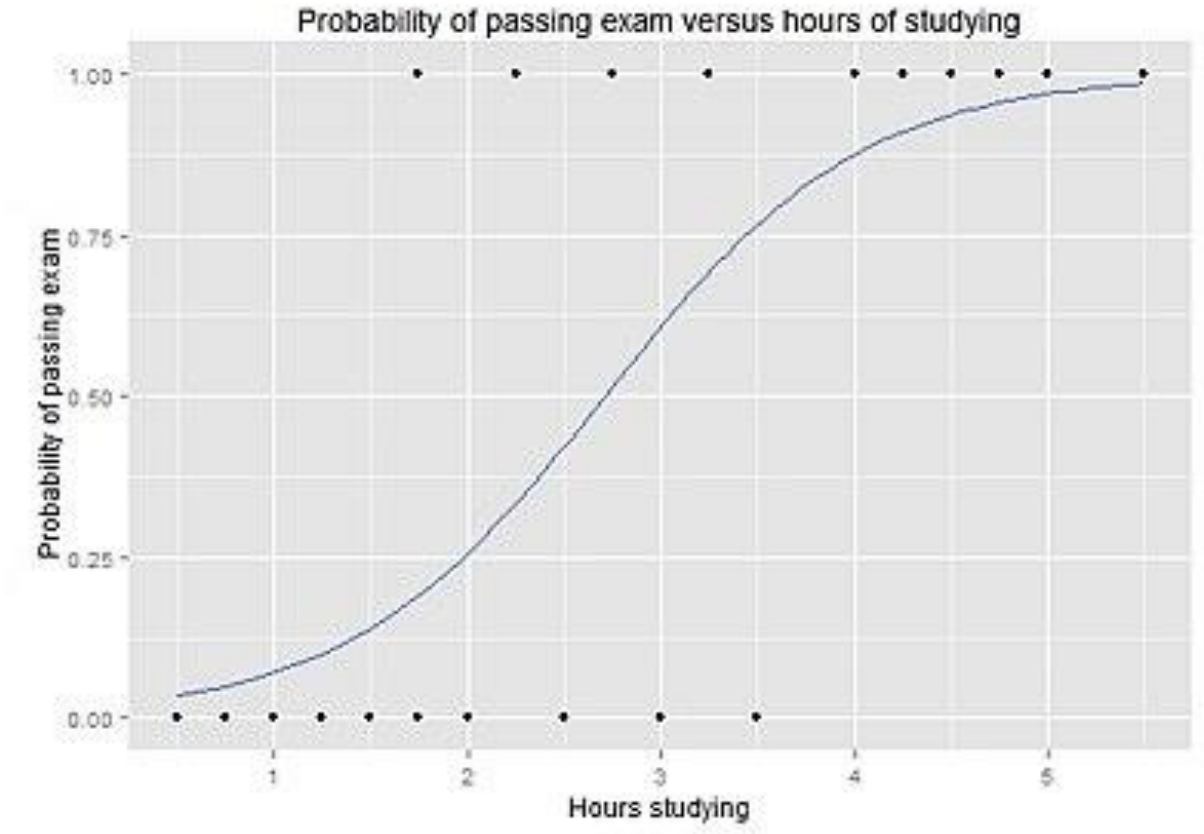
$$e^{\beta_1} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))} = \text{Change in odds ratio}$$

Detailed example: https://www.youtube.com/watch?v=Rj3pVkjAwY0

Example

Hours of study	Passing exam		
	Log-odds	Odds	Probability
1	−2.57	0.076 ≈ 1:13.1	0.07
2	−1.07	0.34 ≈ 1:2.91	0.26
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97

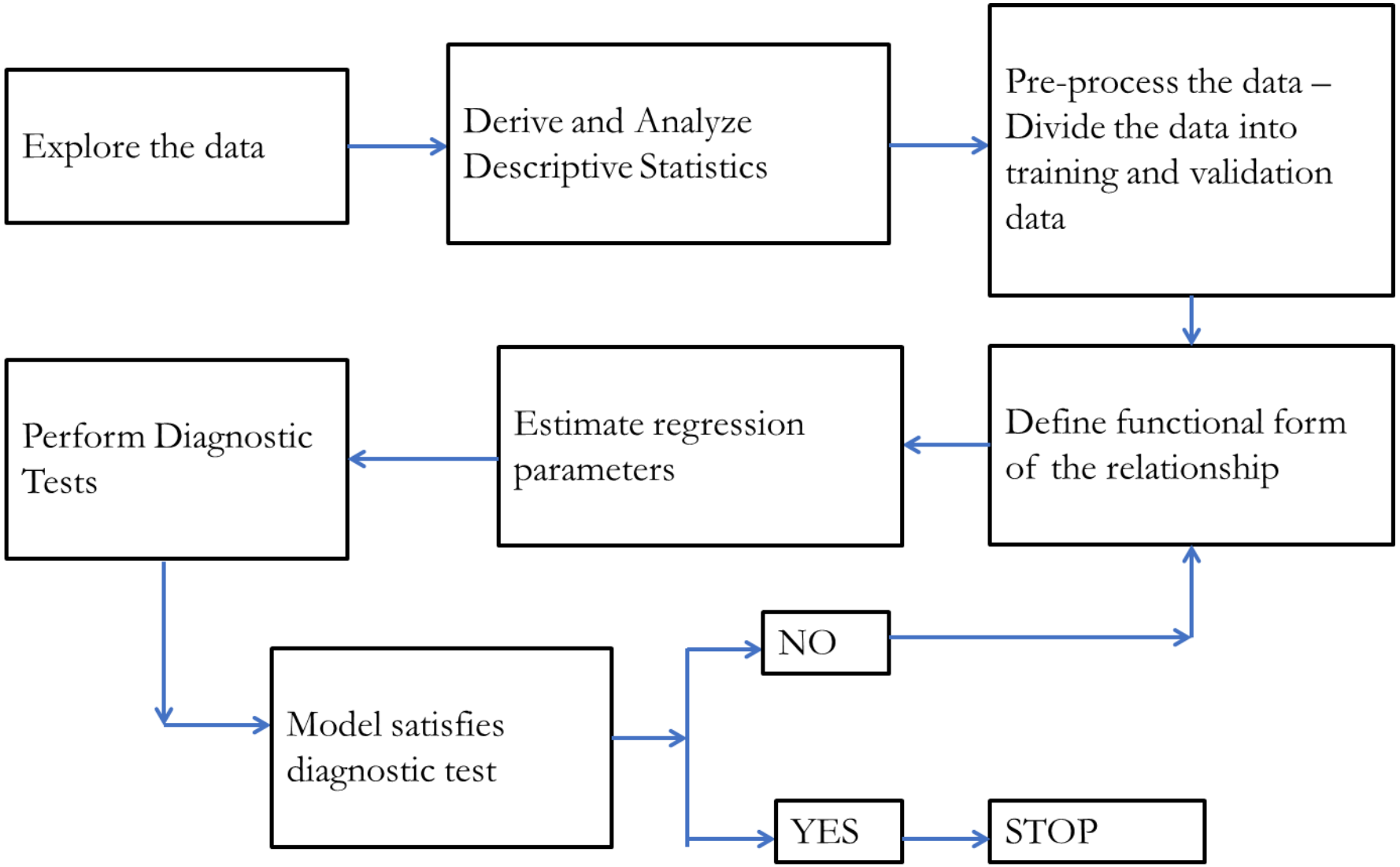
One hr of study increases log odds of passing by 1.5046



	Coefficient	Std.Error	P-value (Wald)
Intercept	−4.0777	1.7610	0.0206
Hours	1.5046	0.6287	0.0167

https://en.wikipedia.org/wiki/Logistic_regression

Logistic Regression Model Development



Wald's test

Wald's test is used for checking statistical significance of individual predictor variables (equivalent to *t*-test in MLR model). The null and alternative hypotheses for Wald's test are:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

Wald's test statistic is given by

$$W = \left[\frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)} \right]^2$$

R^2 in Logistic Regression

- In linear regression R^2 is the proportion of variation explained by the regression model.

- It is not possible to develop a R^2 type measure for Logistic Regression since the variance of the error term is not constant.

- Many Pseudo R^2 values are used in Logistic Regression. Pseudo R2 is an indicator of strength of relationship.

R^2 in Logistic Regression

- **R-squared is a measure of improvement from null model to fitted model** - The denominator of the ratio can be thought of as the sum of squared errors from the null model--a model predicting the dependent variable without any independent variables.

- In the null model, each *y* value is predicted to be the mean of the *y* values.

Pseudo R^2

It is not possible to calculate R^2 as in the case of continuous dependent variable in a logistic regression model.

However, many pseudo R^2 values are used which compare the intercept-only model to the model with independent variables.

Cox and Snell R^2

Cox and Snell R^2 is given by

$$R^2 = 1 - \left\{ \frac{L(\text{Intercept only model})}{L(\text{Full Model})} \right\}^{2/N}$$

Based on Log-likelihood ratio

$$R^2 = 1 - \left(\frac{LL(\text{Null Model})}{LL(\text{ Model})} \right)^{2/n}$$

Null Model : Model without predictors

Full Model: Model with predictors

n is the number of observations

DATA ANALYTICS

Unit 2:Confusion matrices and Metrics

Mamatha H R

Department of Computer Science and Engineering

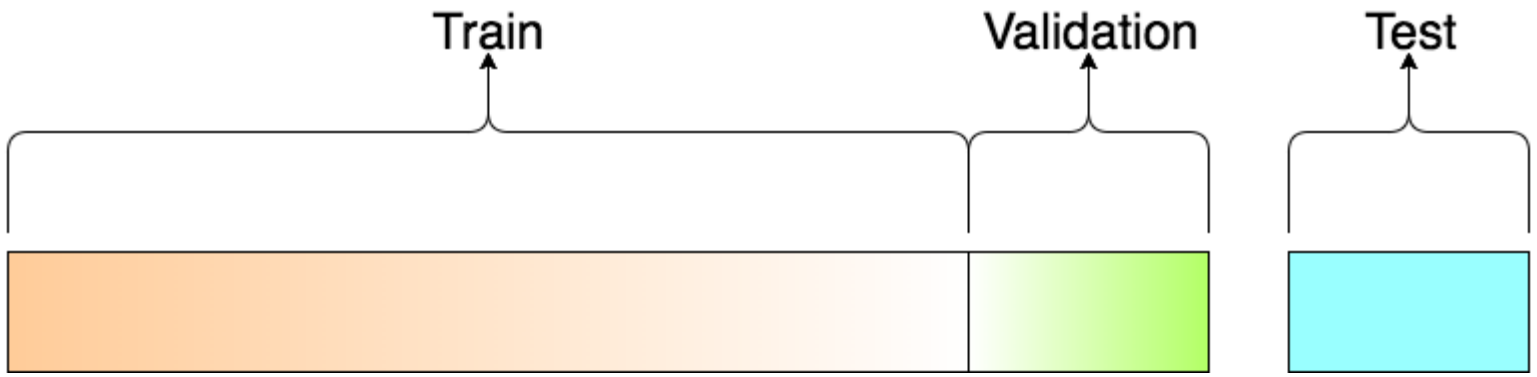
Train, Validation and Test Sets

Training Dataset: The sample of data used to fit the model.
The actual dataset that we use to train the model .The model sees and learns from this data.

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Dataset Split Ratio



A visualization of the splits

Data Split depends on 2 things.

First, the total number of samples in your data

second, on the actual model you are training.

Dataset Split Ratio

Some models need substantial data to train upon, so in this case you would optimize for the larger training sets.

Models with very few hyperparameters will be easy to validate and tune, so you can probably reduce the size of your validation set,

but if your model has many hyperparameters, you would want to have a large validation set as well.

Also, if you happen to have a model with no hyperparameters or ones that cannot be easily tuned, you probably don’t need a validation set too!

Confusion matrix

The confusion matrix is a metric that is often used to measure the performance of a classification algorithm.

In binary classifiers as with the spam filtering example, in which each email can be either spam or not spam.

The confusion matrix will be of the following form:

	Predicted: Real Email	Predicted: Spam Email
Actual: Real Email	True Negatives (TN)	False Positives (FP)
Actual: Spam Email	False Negatives (FN)	True Positives (TP)

Confusion matrix

The predicted classes are represented in the columns of the matrix, whereas the actual classes are in the rows of the matrix.

We then have four cases:

True positives (TP): the cases for which the classifier predicted 'spam' and the emails were actually spam.

True negatives (TN): the cases for which the classifier predicted 'not spam' and the emails were actually real.

False positives (FP): the cases for which the classifier predicted 'spam' but the emails were actually real.

False negatives (FN): the cases for which the classifier predicted 'not spam' but the emails were actually spam.

Classification Performance Metrics

Accuracy: Out of all the classes, how much we predicted correctly

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The ability of the model to correctly classify positives and negatives are called sensitivity and specificity

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

Recall or
True positive rate

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}}$$

True negative rate

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

$$\text{F} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017 (Ch. 11.1-11.4, 11.6-11.6.2)

https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning

https://online.stat.psu.edu/stat504/node/216/

https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-f384152a33d1

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

DATA ANALYTICS

Unit 2:Confusion matrices and Metrics

Mamatha H R

Department of Computer Science and Engineering

Confusion matrix

The confusion matrix is a metric that is often used to measure the performance of a classification algorithm.

In binary classifiers as with the spam filtering example, in which each email can be either spam or not spam.

The confusion matrix will be of the following form:

	Predicted: Real Email	Predicted: Spam Email
Actual: Real Email	True Negatives (TN)	False Positives (FP)
Actual: Spam Email	False Negatives (FN)	True Positives (TP)

Classification Performance Metrics

Accuracy: Out of all the classes, how much we predicted correctly

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The ability of the model to correctly classify positives and negatives are called sensitivity and specificity

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}}$$

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

$$\text{F} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion matrix to compare two classifiers

- Which is the better classification model (wrt Class A)?

		Predicted A	Predicted A’
Actual	A	60	34
	A’	1	12

Model 1

Recall(A) = 60/94
Precision(A) = 60/61
F1_score(A) = 2RP/(R+P)
= 0.774
Accuracy(Model1)
= 0.673

Recall(A’)=12/13
Precision(A’)=12/34
F1_score(A’)=0.51

		Predicted A	Predicted A’
Actual	A	90	4
	A’	8	5

Model 2

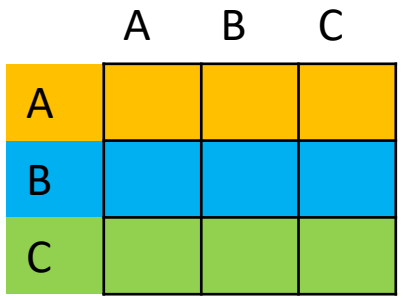
Recall(A) = 90/94
Precision(A) = 90/98
F1_score(A) = 2RP/(R+P)
= 0.937
Accuracy(Model2)
= 0.888

Recall(A’)=5/13
Precision(A’)=5/9
F1_score(A’)=0.45

DATA ANALYTICS

Confusion matrices for multiple classes

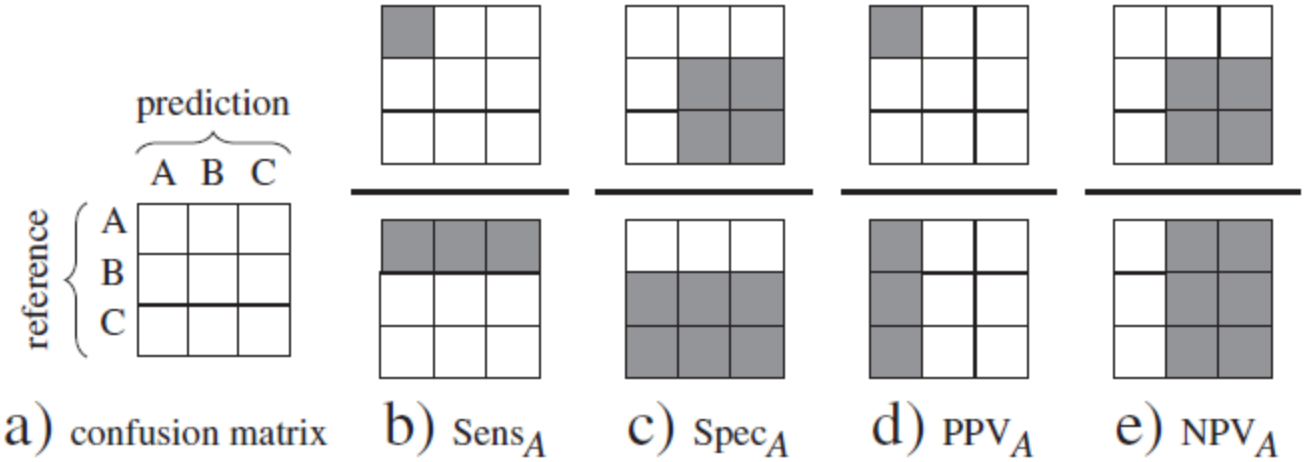
- What is Recall(A)?
- What is Specificity(B)?
- What is Precision(C)?
- What is the average accuracy of this model?



		Predicted class: Cat	Dog	Rabbit
Actual class:	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

		Predicted class: Cat	NotCat
Actual class:	Cat	5	3
	NotCat	2	17

		Predicted class: Dog	NotDog
Actual class:	Dog	3	3
	NotDog	5	16



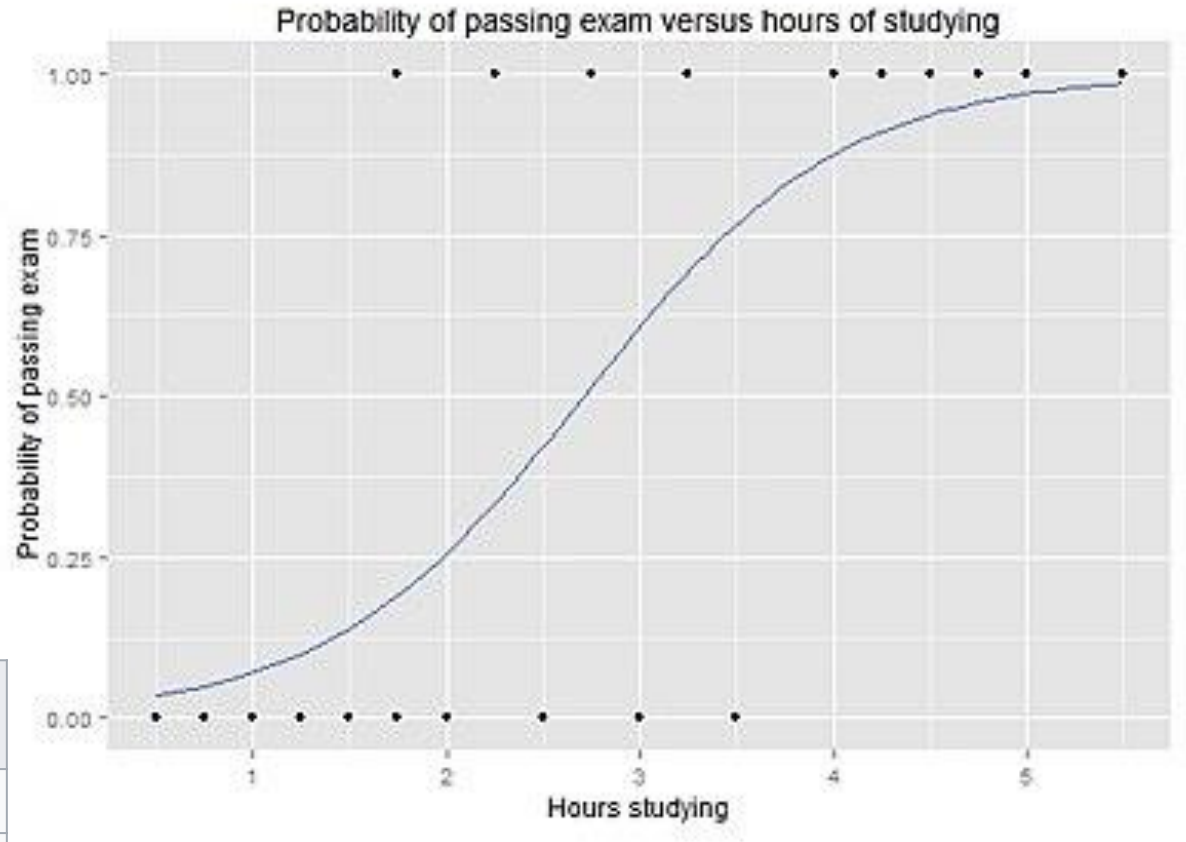
https://stats.stackexchange.com/questions/91044/how-to-calculate-precision-and-recall-in-a-3-x-3-confusion-matrix

Revisiting the example

Hours of study	Passing exam		
	Log-odds	Odds	Probability
1	−2.57	0.076 ≈ 1:13.1	0.07
2	−1.07	0.34 ≈ 1:2.91	0.26
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97

One hr of study increases log odds of passing by 1.5046

	Coefficient	Std.Error	P-value (Wald)
Intercept	−4.0777	1.7610	0.0206
Hours	1.5046	0.6287	0.0167



https://en.wikipedia.org/wiki/Logistic_regression

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

DATA ANALYTICS

Concordant and Discordant Pairs

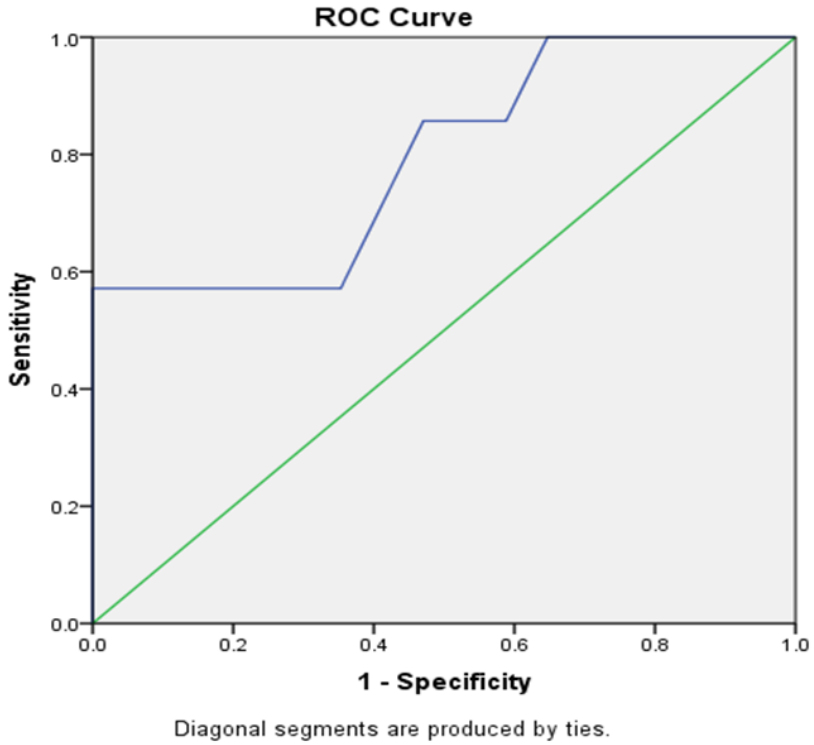
- Discordant Pairs. A pair of positive and negative observations for which the model has no cut-off probability to classify both of them correctly are called discordant pairs.
- Concordant Pairs. A pair of positive and negative observations for which the model has a cut-off probability to classify both of them correctly are called concordant pairs.
- Divide the dataset into positives (y=1) and negatives (y=0).
- For a randomly chosen positive and negative, if the probability of positive (obtained using logistic regression model) is greater than probability of negative then such pairs are called concordant pairs.
- For a randomly chosen positive and negative, if the probability of positive is less than probability of negative then such pairs are called discordant pairs.
- Area under the ROC curve is the proportion of concordant pairs in the dataset.

Hours of study	Passing exam	
	Probability	Label
1	0.070	0
2	0.260	0
3	0.610	0
4	0.870	1
5	0.950	1
6	0.970	1
7	0.980	0

(1,5): concordant pair
(4,7): discordant pair

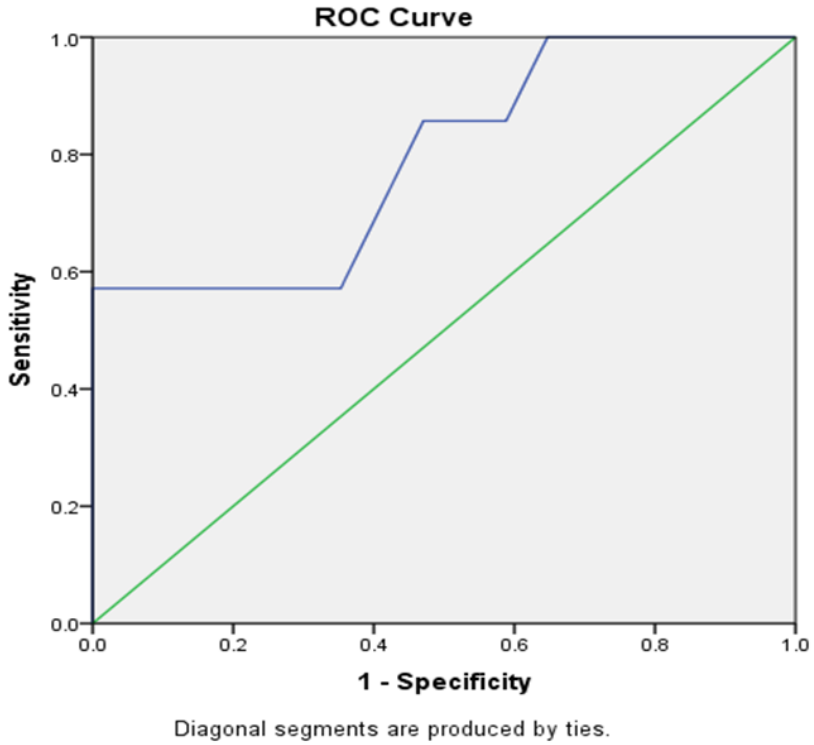
Receiver Operating Characteristics (ROC) Curve

- ROC curve is a plot between sensitivity (true positive rate) in the vertical axis and 1 – specificity (false positive rate) in the horizontal axis.
- The higher the area under the ROC curve, the better the prediction ability.



Receiver Operating Characteristics (ROC) Curve

- ROC curve is a plot between sensitivity (true positive rate) in the vertical axis and 1 – specificity (false positive rate) in the horizontal axis.
- The higher the area under the ROC curve, the better the prediction ability.



Area Under ROC Curve (AUC)

- Area under the ROC (AUC) curve is interpreted as the probability that the model will rank a randomly chosen positive higher than randomly chosen negative.

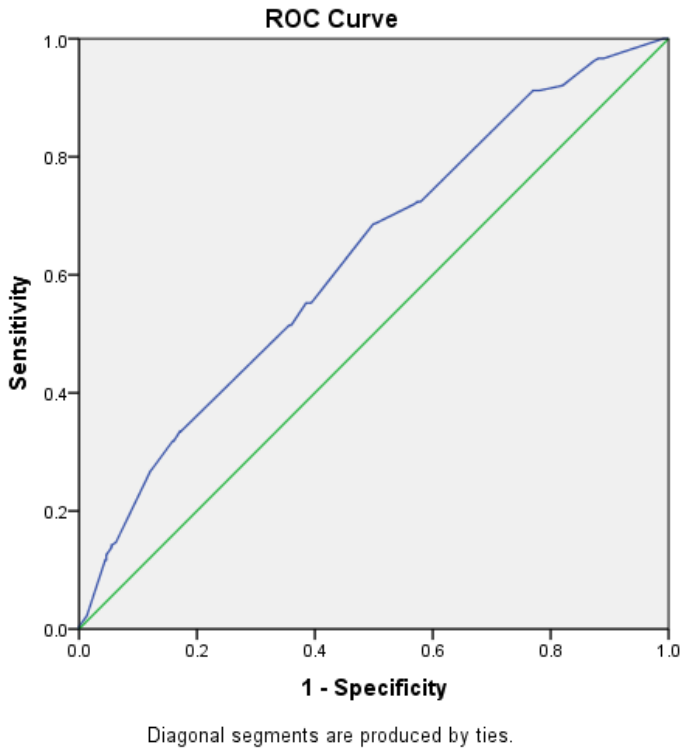
- If n_1 is the number of positives (1s) and n_2 is the number of negatives (0s), then the area under the ROC curve is the proportion of cases in all possible combinations of (n_1, n_2) such that n_1 will have higher probability than n_2.

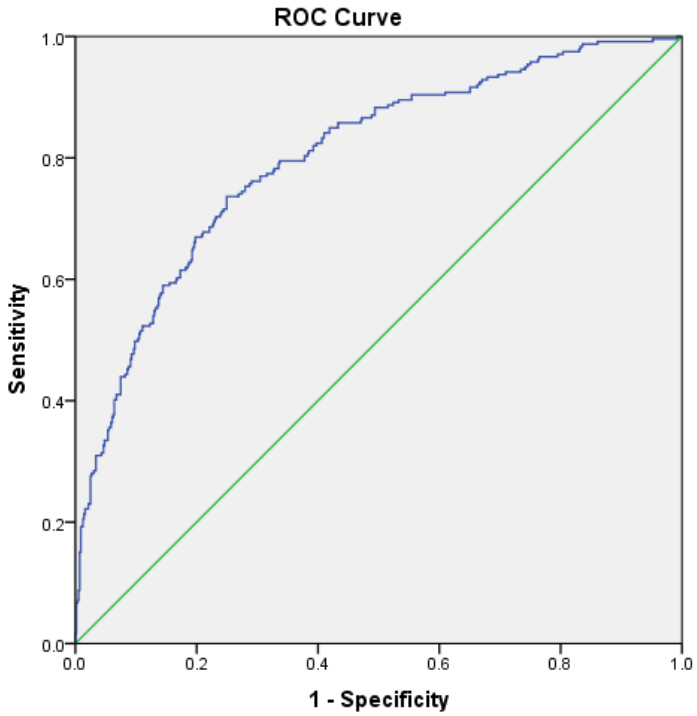
AUC = P (Random Positive Observation) > P(Random Negative Observation)

Area Under the ROC Curve (AUC) is a measure of the ability of the logistic regression model to discriminate positives and negatives correctly.

DATA ANALYTICS

Area Under ROC Curve (AUC)





AUC = 0.629

AUC = 0.801

ROC Curve

- General rule for acceptance of the model:

- If the area under ROC is:

  - 0.5 $\Rightarrow$ No discrimination

  - 0.7 $\leq$ ROC area < 0.8 $\Rightarrow$ Acceptable discrimination

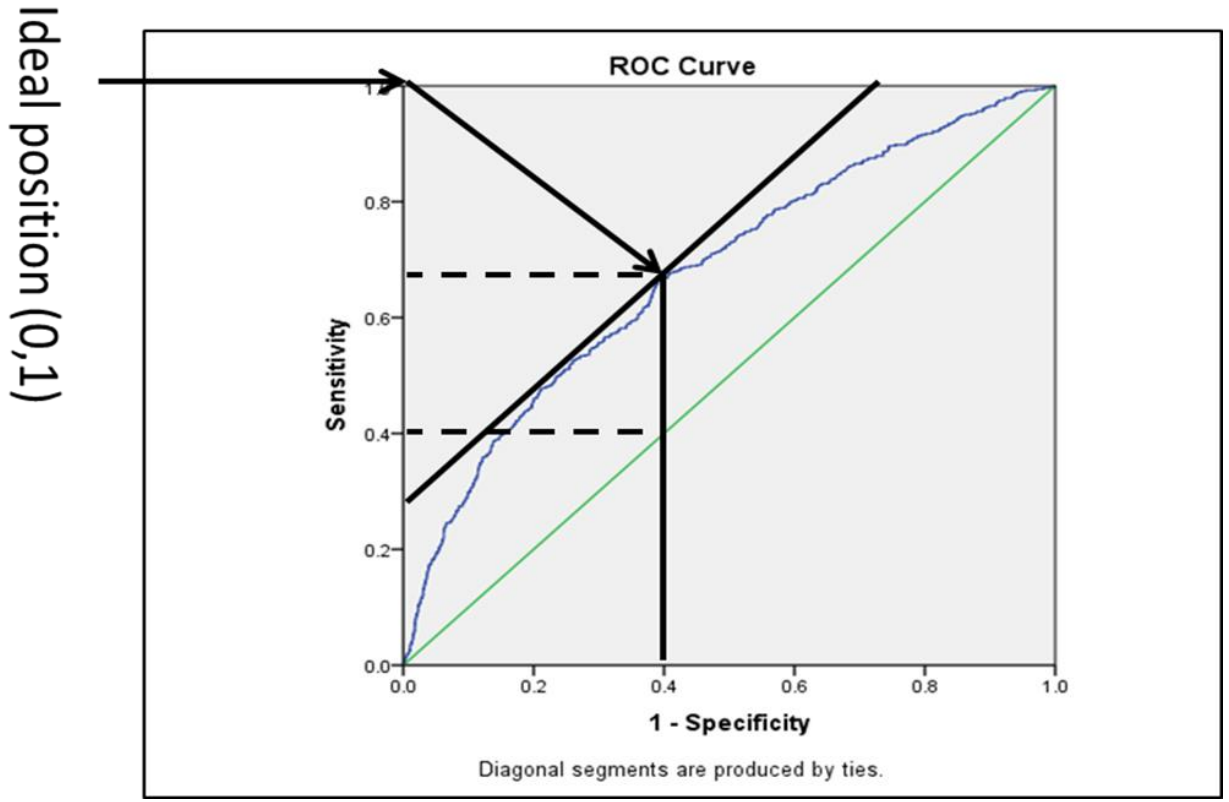
  - 0.8 $\leq$ ROC area < 0.9 $\Rightarrow$ Excellent discrimination

  - ROC area $\geq$ 0.9 $\Rightarrow$ Outstanding discrimination

Youden's Index for Optimal Cut-Off Probability

Youden's Index (1950) is a classification cut-off probability, for which the following function is maximized (also known as J statistic):

Youden's Index = J Statistic = $\underset{P}{\text{Max}}$[Sensitivity(p) + Specificity(p) – 1]



Cost-Based Cut-Off Probability

In cost-based approach, we assign penalty cost for misclassification of positives and negatives. Assume that cost of misclassifying negative (0) as positive (1) is C_{01} and cost of misclassifying positive (1) as negative (0) is C_{10} as shown in Table

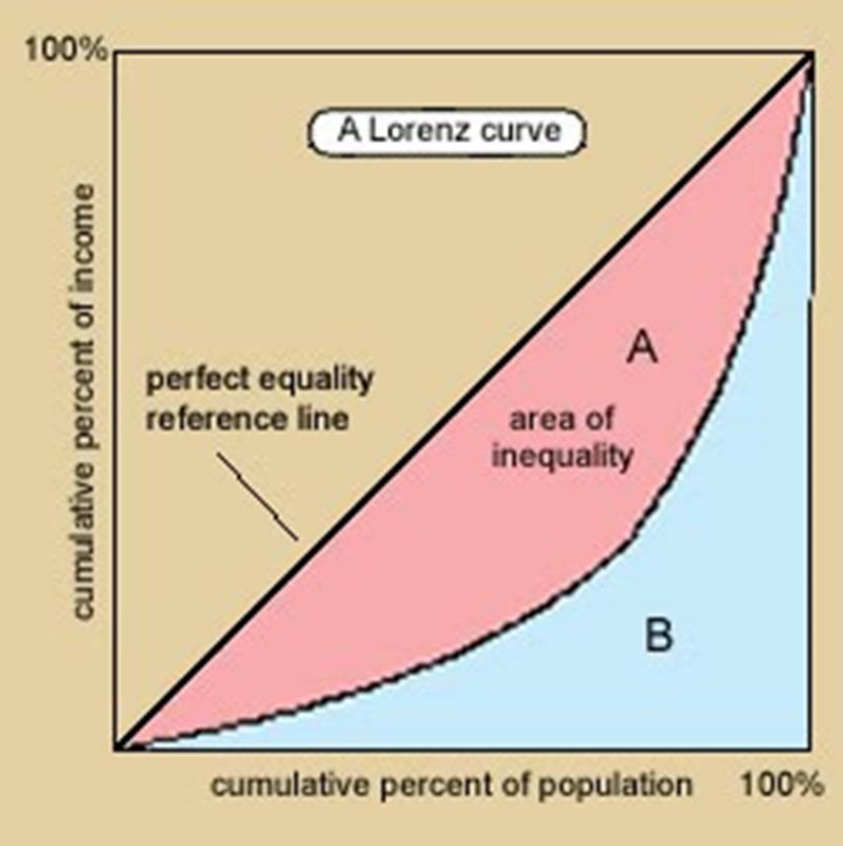
Observed	Classified	
	0	I
0	---	C_{01}
I	C_{10}	---

The optimal cut-off probability is the one which minimizes the total penalty cost and is given by

$$\min_{p} \left[C_{01}P_{01} + C_{10}P_{10} \right]$$

DATA ANALYTICS

Lorenz Curve



Gini Index is a statistical measure of dispersion

Gini Coefficient = A / (A+B)

Gini Coefficient = 2 AUC - 1

AUC = Area Under the ROC Curve

Gini Coefficient

- Gini coefficient measures individual impact of the an explanatory variable.

- Gini coefficient = 2 AUC – 1

- AUC = Area under the ROC Curve

Questions asked after the class

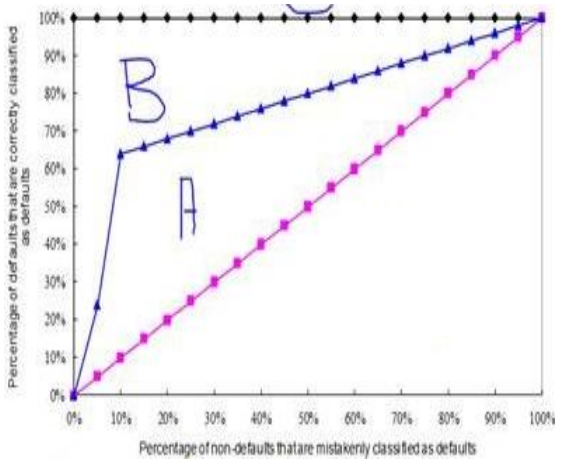
- Can discordant pairs be thought of as outliers?
  - Indeed, discordancy tests are used to detect outliers and one or both the points in the discordant pair could be outliers.
  - However, we must also be aware there are other possibilities:

  (a) The parameters (coefficients) could be better estimated

  (b) The current model (logistic regression) is not suitable for modeling the data on hand (some preprocessing may be required before we can model the data using Logistic Regression or we could explore alternatives)

- Why is Gini coefficient = (2 AUC – 1)?
  - (AUC = Area under the ROC Curve)
  - Gini = A/(A+B)
    - A = area under the curve and the diagonal
    - B = area under the perfect model and diagonal
    - Gini (in RoC) = A/(A+B) = A/0.5 = 2A
    - AUC for this case = A + ½
    - AUC = Gini/2 + ½

    ⇒ Gini = 2AUC-1



https://stats.stackexchange.com/questions/342329/gini-and-auc-relationship

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017 (Ch. 11.1-11.4, 11.6.5, 11.7.2-11.7.3)

https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning

https://online.stat.psu.edu/stat504/node/216/

https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-f384152a33d1

DATA ANALYTICS

Unit 3: Time Series Analysis

Jyothi R.
Department of Computer Science and
Engineering

DATA ANALYTICS

Unit 3: Introduction to Time Series Data

Jyothi R., Gowri Srinivasa

Department of Computer Science and Engineering

DATA ANALYTICS

INTRODUCTION TO FORECASTING

- Forecasting - important and frequently addressed problems in analytics

- Inaccurate forecasting has a significant impact

- For example
  - non-availability of product → customer dissatisfaction
  - too much inventory → erodes the organization's profit

- Necessary to forecast the demand for a product and service as accurately as possible.

- Every organization prepares long-range and short-range planning
  - forecasting demand for product and service is an important input
    for both long-range and short-range planning

- Budget allocation, manpower, warehouse capacity, machine resource planning, etc., based on forecast of demand for a product

INTRODUCTION TO FORECASTING

- Forecasting can be very challenging with stock keeping units (SKUs) running into several millions.

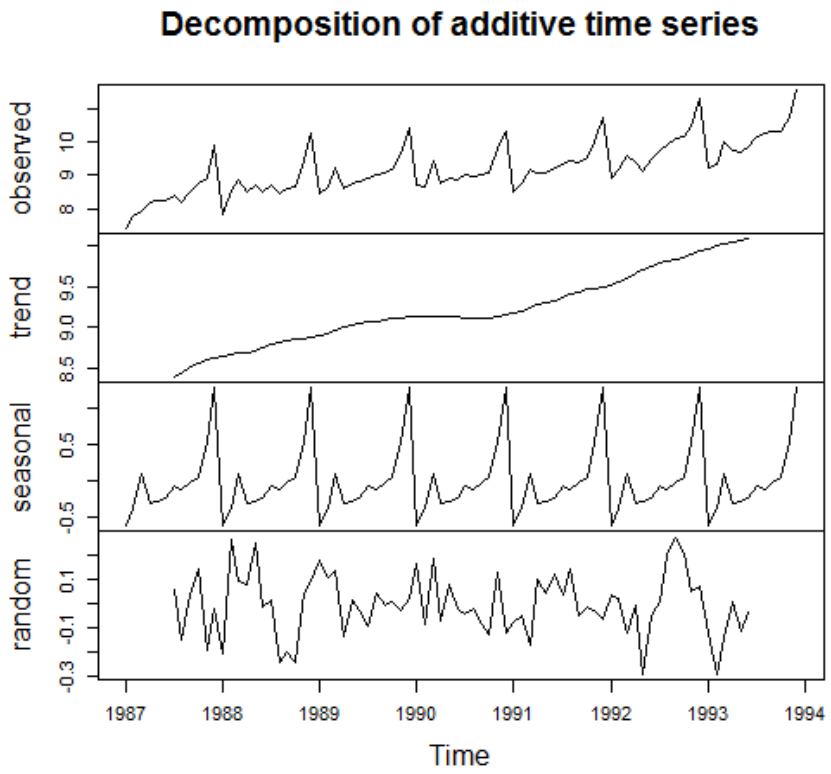
1. Boeing 747-400 has more than 6 million parts and several thousand unique parts (Hill, 2011). Forecasting demand for spare parts is important since non-availability of mission critical parts can result in aircraft on ground (AOG) which can be very expensive for airlines.

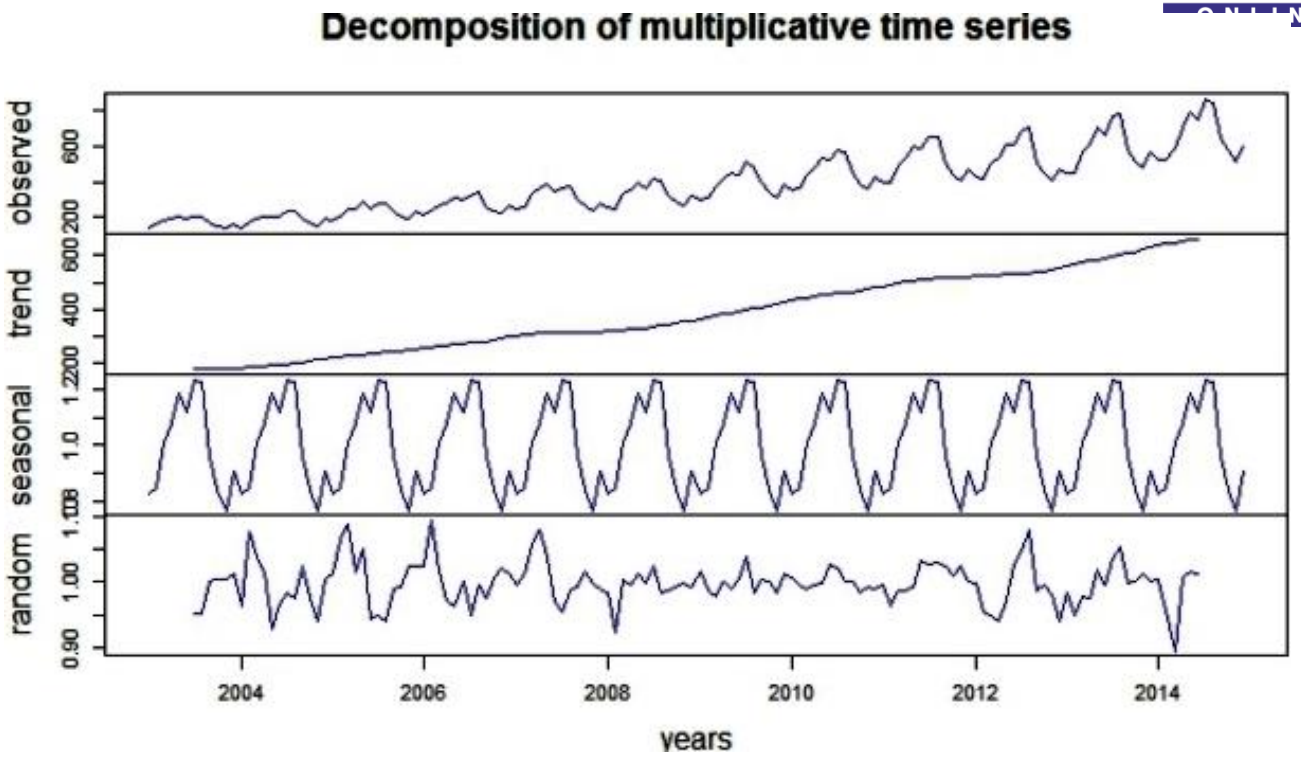
2. Amazon.com sells more than 350 million products through its E-commerce portal. Amazon itself sells about 13 million SKUs and has more (about 2 million) retailers selling their products through Amazon (Ali, 2017).

3. Walmart sells more than 142,000 products through their supercenters. Being a brick-and-mortar retail store, Walmart has to maintain stock for each and every product sold and predict demand for the products as accurately as possible.

DATA ANALYTICS

Additive and Multiplicative Time Series Data





$$Y_t = T_t + S_t + C_t + I_t$$

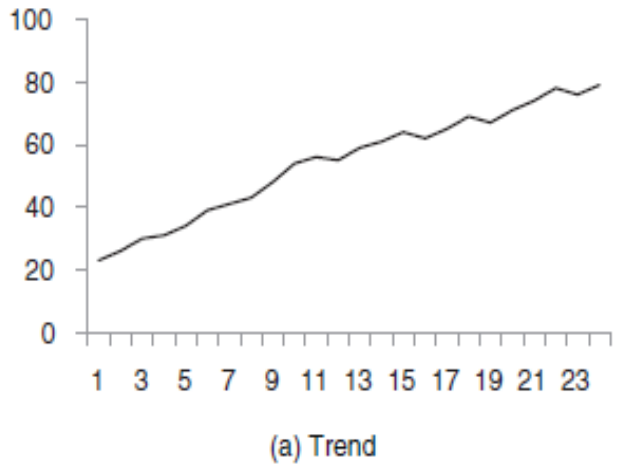
$$Y_t = T_t \times S_t \times C_t \times I_t$$

DATA ANALYTICS

COMPONENTS OF TIME-SERIES DATA

From a forecasting perspective, a time-series data can be broken into the following components

1. **Trend Component (T_t):** Trend is the consistent long-term upward or downward movement of the data over a period of time.



(a) Trend

https://school.stockcharts.com/doku.php?id=chart_analysis:trend_lines

DATA ANALYTICS

COMPONENTS OF TIME-SERIES DATA Contd.

2. Seasonal Component (S_t): Seasonal component is the repetitive upward or downward movement (or fluctuations) from the trend that occurs within a calendar year such as seasons, quarters, months, days of the week, etc.

- The upward or downward fluctuation may be caused due to festivals, customs within a society, school holidays, business practices within the market such as ‘end of season sale’, and so on.

- For example, in India demand for many products surge during the festival months of October - December.

- Seasonal fluctuation occurs at fixed intervals (such as months, quarters) known as periodicity of seasonal variation and repeats over time.

Seasonal Component (S_t): Contd.

The seasonal component consists of effects that are reasonably stable with respect to timing, direction and magnitude. It arises from systematic, calendar related influences such as:

- **Natural Conditions:** Weather fluctuations that are representative of the season (uncharacteristic weather patterns such as snow in summer would be considered irregular influences)

- **Business and Administrative procedures:**
  Start and end of the school term

- **Social and Cultural behavior:**
  Christmas

Seasonal Component (S_t):

It also includes calendar related systematic effects that are not stable in their annual timing or are caused by variations in the calendar from year to year, such as:

- **Trading Day Effects**

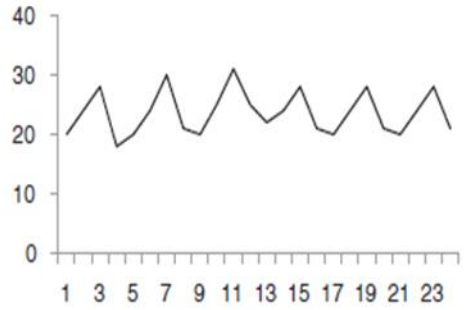
the number of occurrences of each of the day of the week in a given month will differ from year to year

- There were 4 weekends in March in 2000, but 5 weekends in March of 2002

- **Moving Holiday Effects**

holidays which occur each year, but whose exact timing shifts

- Diwali, Easter, Ramadan



(b) Seasonality (fixed periodicity)

Identifying seasonal components:

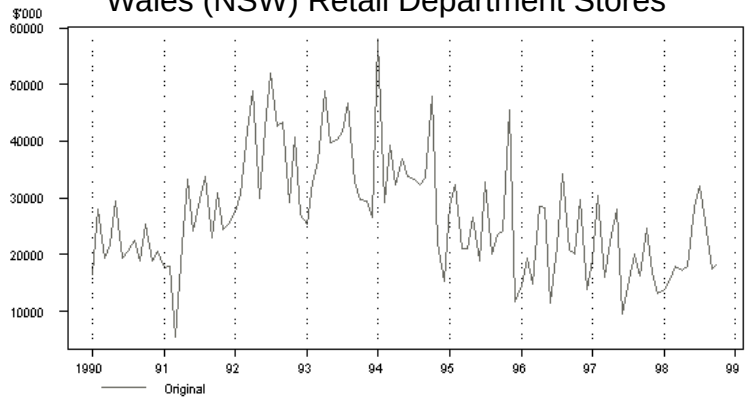
Regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend

COMPONENTS OF TIME-SERIES DATA

Seasonal Component (S_t):

- Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend.
- In this example, the magnitude of the seasonal component increases over time, as does the trend.

Monthly Retail Sales in New South Wales (NSW) Retail Department Stores



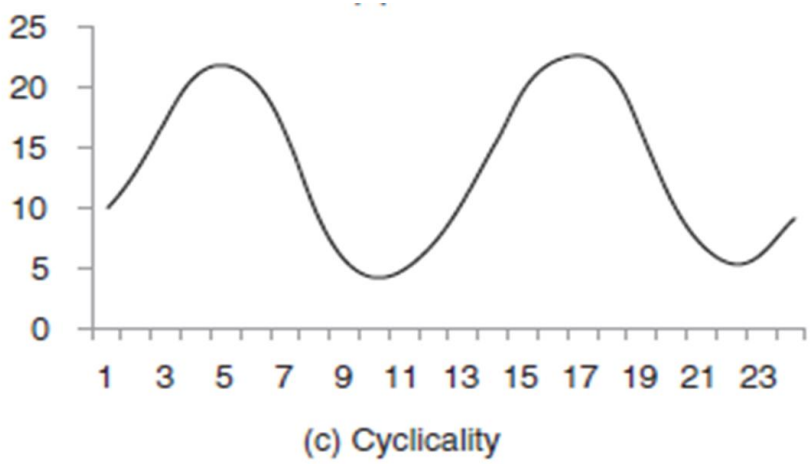
Obvious large seasonal increase in December retail sales in New South Wales due to Christmas shopping

DATA ANALYTICS

COMPONENTS OF TIME-SERIES DATA

3. Cyclical Component (C_t): Cyclical component is fluctuation around the trend line that happens due to macro-economic changes such as recession, unemployment, etc.

- Cyclical fluctuations have repetitive patterns with a time between repetitions of more than a year



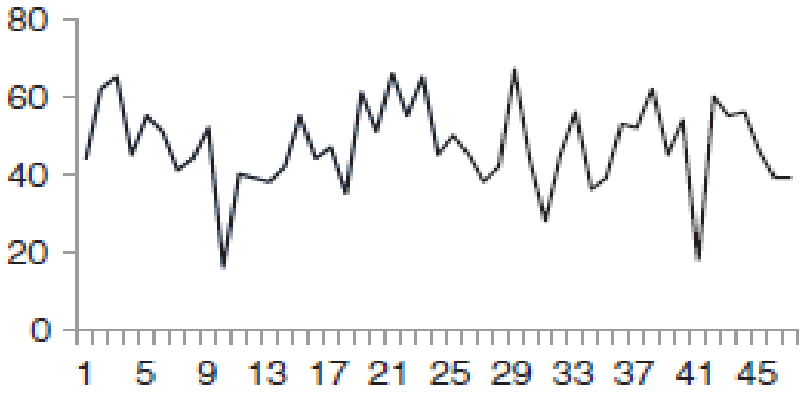
(c) Cyclicality

- A major difference between seasonal fluctuation and cyclical fluctuation is that seasonal fluctuation occurs at fixed period within a calendar year, whereas cyclical fluctuations have random time between fluctuations.
- That is, periodicity of seasonal fluctuations is constant, whereas the periodicity of cyclical fluctuations is not constant.

DATA ANALYTICS

COMPONENTS OF TIME-SERIES DATA contd.

4. Irregular Component (I_t): Irregular component is the white noise or random uncorrelated changes that follow a normal distribution with mean value of 0 and constant variance.



(d) Irregular

- What remains after the seasonal and trend components of a time series have been estimated and removed.
- It results from short term fluctuations in the series which are neither systematic nor predictable

DATA ANALYTICS

Additive and Multiplicative Time Series Revisited

- The additive time-series model is given by

$$Y_t = T_t + S_t + C_t + I_t$$

- The additive models assume that the seasonal and cyclical components are independent of the trend component.
- Additive models are not very common since in many cases the seasonal component may not be independent of trend.
- The additive model is appropriate if the seasonal component remains constant about the level (or mean) and does not vary with the level of the series.

- The multiplicative time-series model is given by

$$Y_t = T_t \times S_t \times C_t \times I_t$$

- Multiplicative models are more common and are a better fit for many data sets.

- In many cases, we will use the form

$$Y_t = T_t \times S_t$$

- To estimate the cyclical component we will need a large data set.

- The multiplicative model is more appropriate, if seasonal variation is correlated with level (local mean).

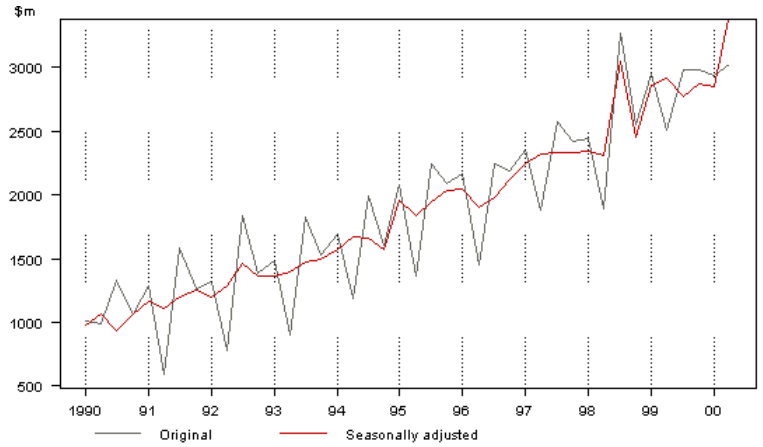
DATA ANALYTICS

Additive and Multiplicative Time Series Revisited

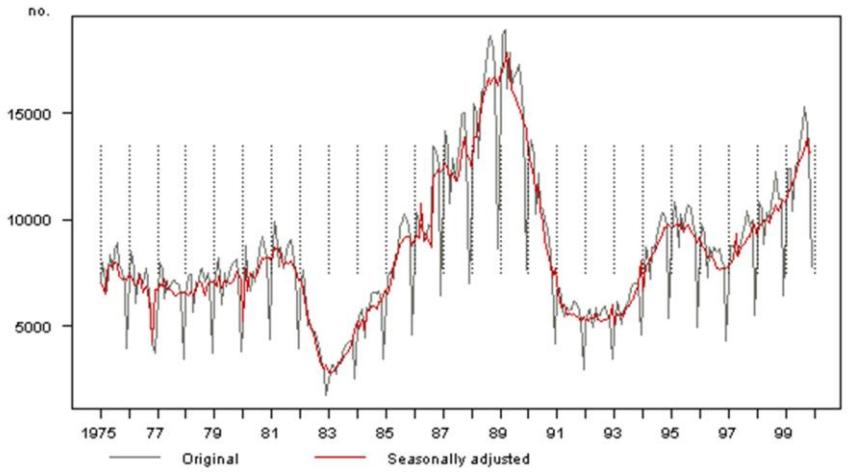
- The additive time-series model is given by $Y_t = T_t + S_t + C_t + I_t$

- The multiplicative time-series model is given by $Y_t = T_t \times S_t \times C_t \times I_t$

General Government and Other Current Transfers to Other Sectors



Monthly NSW ANZ Job Advertisements



The underlying level of the series fluctuates but the magnitude of the seasonal spikes remain approximately stable

The trend has the same units as the original series, but the seasonal and irregular components are unitless factors, distributed around 1

Decomposition of Time Series Data - Additive

- Decomposition models are typically additive or multiplicative, but can also take other forms such as pseudo-additive.

Additive Decomposition

In some time series, the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls. In such cases, an additive model is appropriate.

In the additive model, the observed time series (O_t) is considered to be the sum of three independent components: the seasonal S_t, the trend T_t and the irregular I_t.

Observed series = Trend + Seasonal + Irregular

$$O_t = T_t + S_t + I_t$$

Seasonally adjusted series = Observed-Seasonal

$$SA_t = O_t - \hat{S}_t$$

$$= T_t + I_t$$

COMPONENTS OF TIME-SERIES DATA contd.

- **Multiplicative Decomposition**

  In many time series, the amplitude of both the seasonal and irregular variations increase as the level of the trend rises. In this situation, a multiplicative model is usually appropriate.

  In the multiplicative model, the original time series is expressed as the product of trend, seasonal and irregular components.

- Observed series = Trend x Seasonal x Irregular

$$O_t = T_t + S_t + I_t$$

$$\text{Seasonally Adjusted series} = \text{Observed} \div \text{Seasonal}$$
$$= \text{Trend} \times \text{Irregular}$$

$$SA_t = O_t / \hat{S}_t$$
$$= T_t \times I_t$$

COMPONENTS OF TIME-SERIES DATA contd.

Pseudo-Additive Decomposition:

- The multiplicative model cannot be used when the original time series contains very small or zero values
- This is because it is not possible to divide a number by zero
- In these cases, a pseudo additive model combining the elements of both the additive and multiplicative models is used
- This model assumes that seasonal and irregular variations are both dependent on the level of the trend but independent of each other.

The original data can be expressed in the following form:

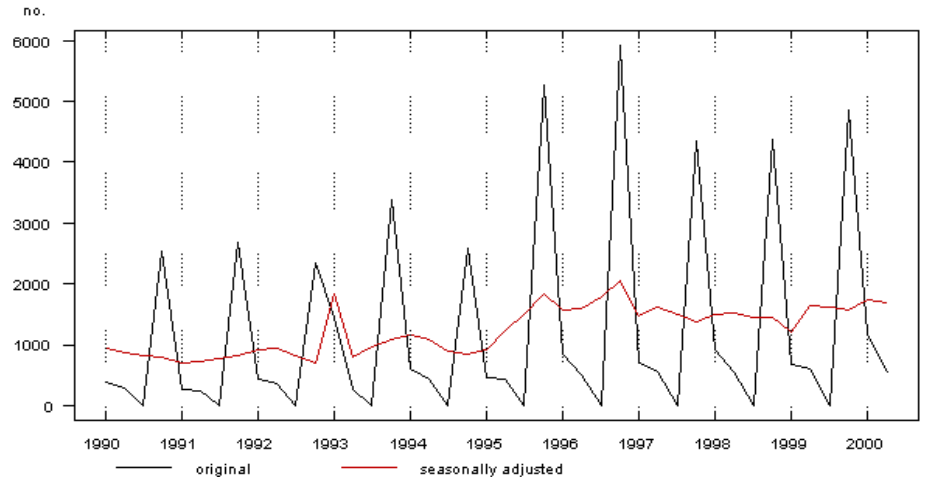
$$O_t = T_t + T_t \times (S_t - 1) + T_t \times (I_t - 1)$$
$$= T_t \times (S_t + I_t - 1)$$

- Both the seasonal factor S_t and the irregular factor I_t centered around one
- We need to subtract one from S_t and I_t to ensure that the terms $T_t \times (S_t - 1)$ and $T_t \times (I_t - 1)$ are centered around zero.
- These terms can be interpreted as the additive seasonal and additive irregular components respectively; the original data O_t will be centered around the trend values T_t .

COMPONENTS OF TIME-SERIES DATA contd.

- An example of series that requires a pseudo-additive decomposition model is shown below.
- This model is used as cereal crops are only produced during certain months, with crop production being virtually zero for one quarter each year.

Quarterly Gross Value for the Production of Cereal Crops



This model is used as cereal crops are only produced during certain months, with crop production being virtually zero for one quarter each year.

References

Text Book:

"Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017 (Chapter 13.1-13.2)

Additional reference and image courtesy:

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

THANK YOU

Dr. Mamatha H R
Professor, Department of Computer Science
mamathahr@pes.edu

+91 80 2672 1983 Extn 834

Ms. Jyothi R.
Assistant Professor, Department of Computer Science
jyothir@pes.edu