LINEAR REGRESSION

Mr.Gangadhar Immadi

immadi.gangadhar@gmail.com

9986789040

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)
- It is a Supervised Learning Technique
- Regression model establishes existence of association between two variables, but not causation.

BCCI Bans Girlfriends and Wives

Girlfriends and wives create such a "distraction" that Indian batsmen can't make runs, bowlers fail to take wickets and fielders drop simple catches.

Regression is not designed to capture causal relationship

Interesting Hypotheses

- Good looking couples are more likely to have girl child(ren)!
- Married people are more happier than singles!!!
- Vegetarians miss fewer flights.
- Black cars have more chance of involving in an accident than white cars in moon light.
- Women use camera phone more than men.
- Left handed men earn more money!
- Smokers are better sales people.
- Those who whistle at workplace are more efficient.
- Hospital treatment cost prediction
- Salary and CGPA

- Impact of NREGA
- Prediction of purchase by customer based on website visits
- Relationship between Service Level Agreement & Service cost
- Salary prediction based on Key Performance Indicators

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 new hypothesis and for validating a hypothesis
- Terms dependent and independent does not necessarily imply a causal relationship between two variables.

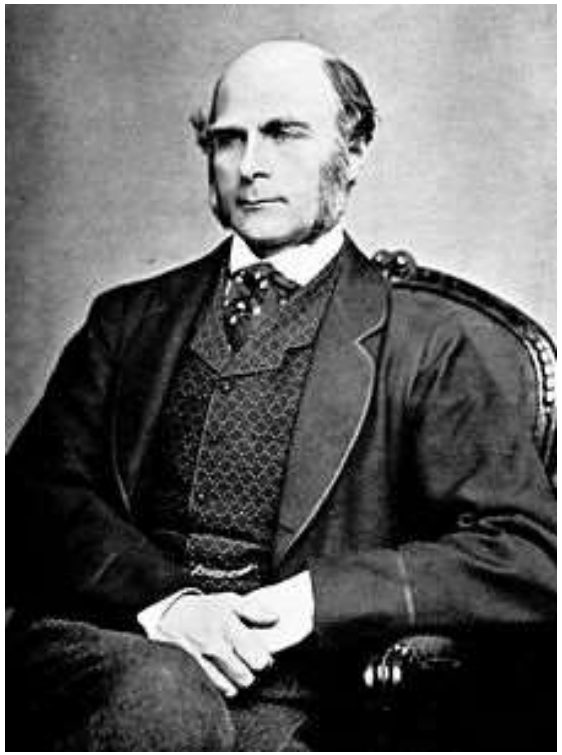
- Regression is not designed to capture causality.

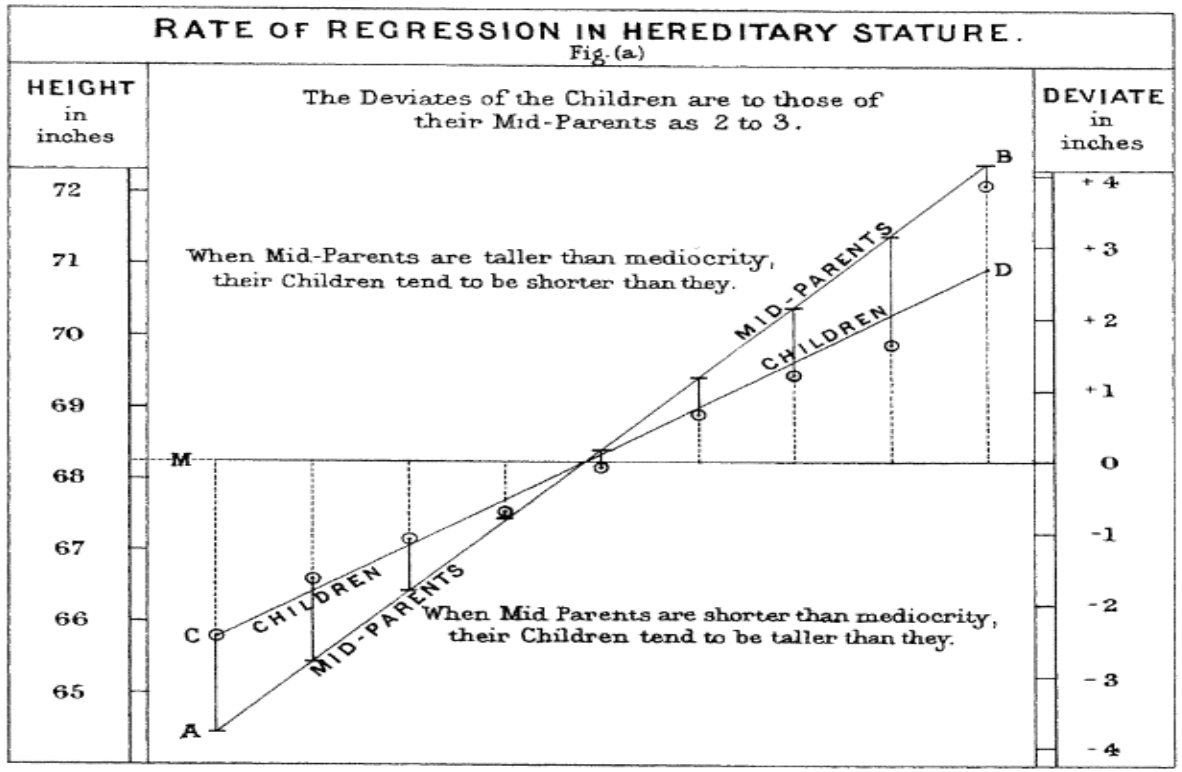
- Purpose of regression is to predict the value of dependent variable given the value(s) independent variable(s)

Regression History

- Francis Galton was the first to apply regression.
- Claimed that height of children of tall parents "regress towards mean of that generation".
- Modern regression analysis is developed by R A Fisher.

Ref: F Galton, "Regression towards mediocrity in hereditary stature", *Nature*, Vol. 15, 246-263, 1886

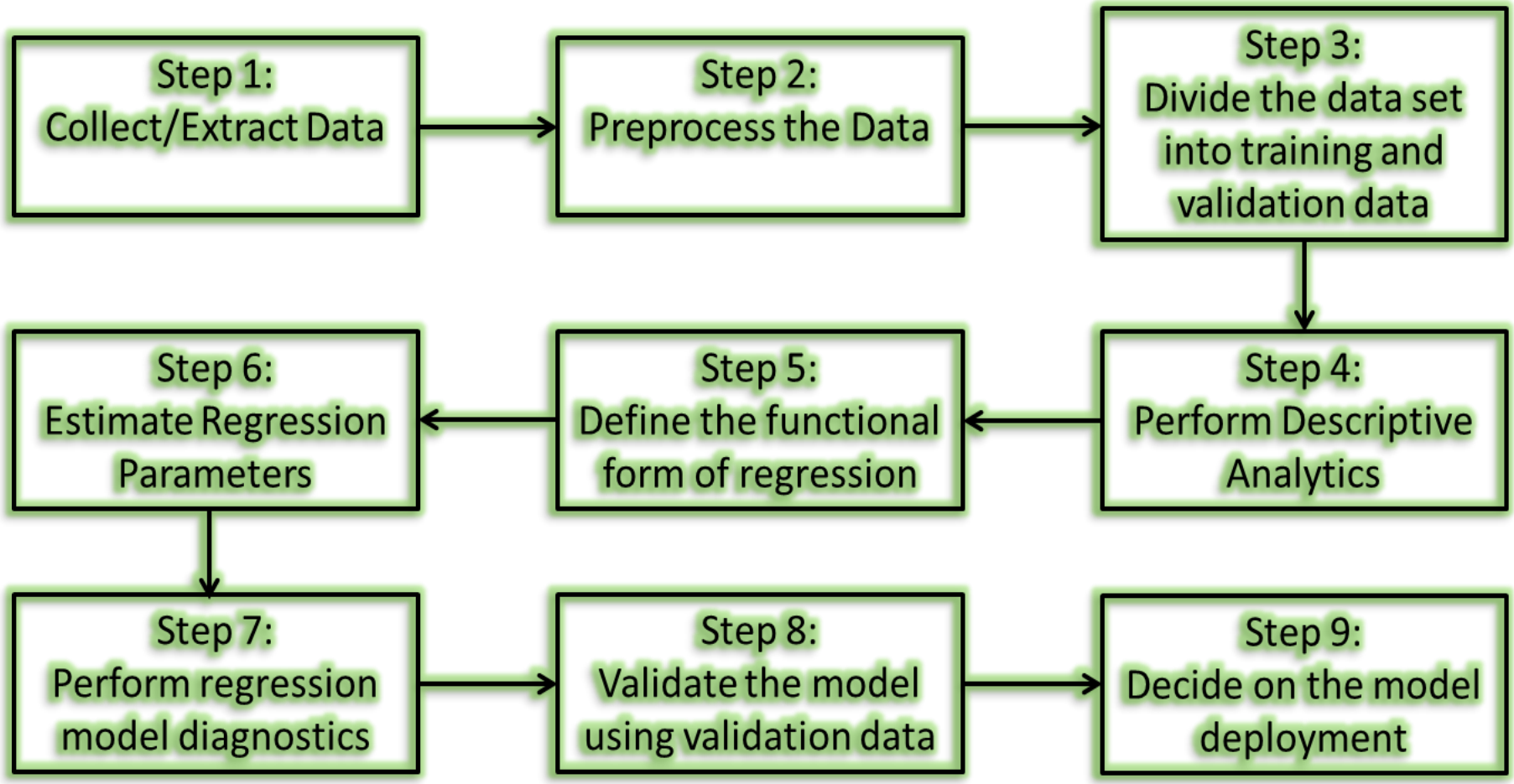




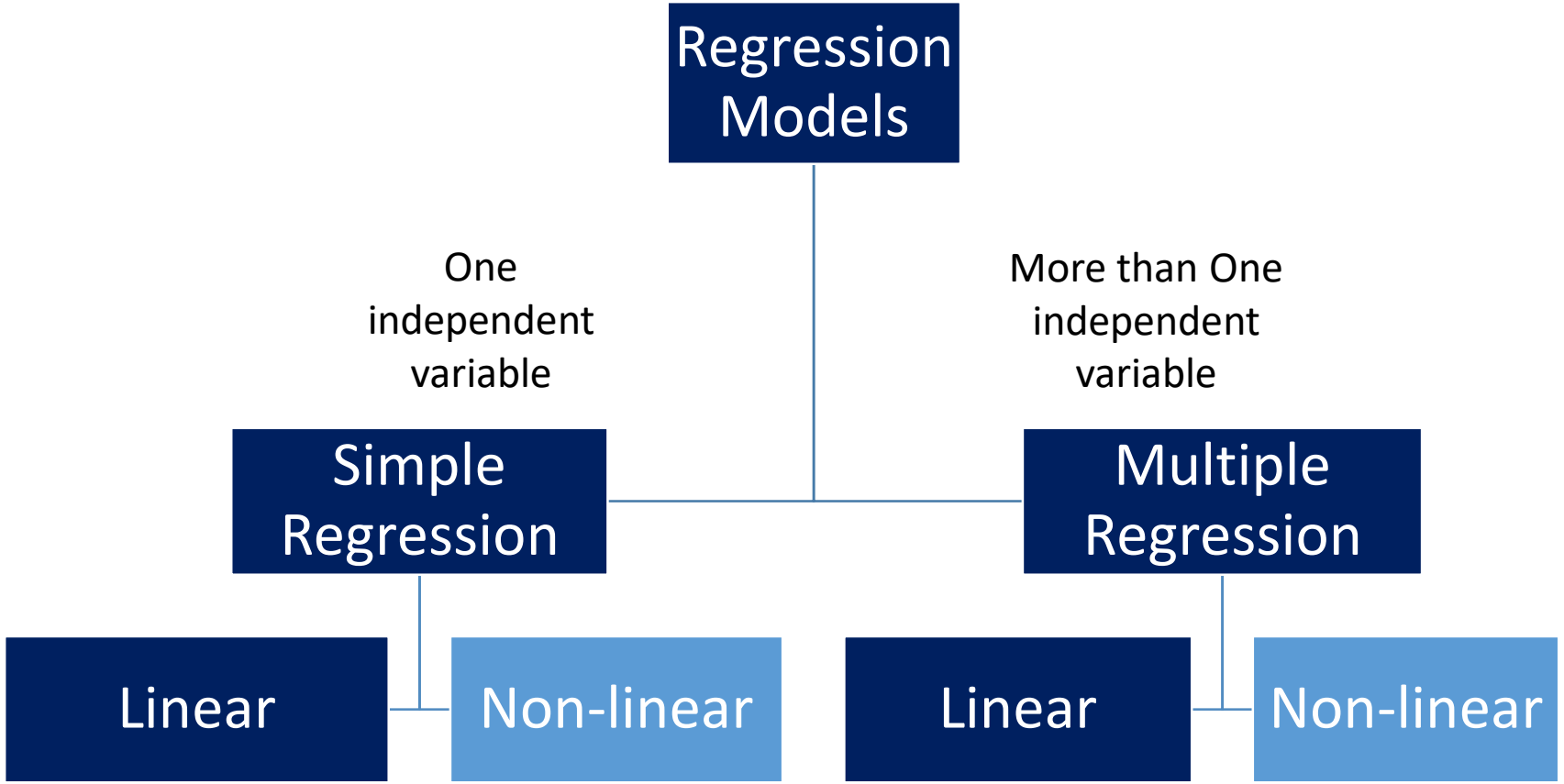
Where is it used?

- ✓ 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.

Framework for SLR model development



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.

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

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

OLS Estimation

In ordinary least squares, the objective is find the optimal values of B_0 and B_1 that will minimize the **Sum of Squared Errors** (SSE) given in belwo 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.

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

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

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

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

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

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

$$\hat{\beta}_1 = \frac{COV(X,Y)}{VAR(X)}$$

Example

- For the following dataset, build a linear regression model and predict the number of minutes taken when number of units serviced is 45

# Units Serviced	# Minutes Taken	xi-xbar	yi-ybar	(xi-xbar)*yi-ybar	(xi-xbar)**2
2	28	-4.4166	-74.4166	328.6683556	19.50635556
3	49	-3.4166	-53.4166	182.5031556	11.67315556
4	65	-2.4166	-37.4166	90.42095556	5.83995556
4	75	-2.4166	-27.4166	66.25495556	5.83995556
5	89	-1.4166	-13.4166	19.00595556	2.00675556
6	98	-0.4166	-4.4166	1.83995556	0.17355556
7	110	0.5834	7.5834	4.42415556	0.34035556
8	120	1.5834	17.5834	27.84155556	2.50715556
9	140	2.5834	37.5834	97.09295556	6.67395556
9	145	2.5834	42.5834	110.0099556	6.67395556
10	150	3.5834	47.5834	170.5103556	12.84075556
10	160	3.5834	57.5834	206.3443556	12.84075556
6.416667	102.4166			1304.916667	86.91666672

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.

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

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

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

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:

- 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).
- Residual analysis to validate the regression model assumptions.
- Outlier analysis.

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

Coefficient of Determination (R-Square or R^2)

- 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

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

Description of total variation, explained variation and unexplained variation

Variation Type	Measure	Description
Total Variation (SST)	$Y_i - \bar{Y}$	Total variation is the difference between the actual value and the mean value.
Variation explained by the model	$\hat{Y}_i - \bar{Y}$	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	$Y_i - \hat{Y}_i$	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:

$$\underbrace{Y_i - \bar{Y}}_{\text{Total Variation in Y}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Variation in Y explained by the model}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Variation in Y not explained by the model}}$$

It can be proved mathematically that sum of squares of total variation is equal to sum of squares of explained variation plus sum of squares of unexplained variation

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

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

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

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

Example

- For the following dataset, build a linear regression model and predict the number of minutes taken when number of units serviced is 45

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

# Units Serviced	# Minutes Taken	yi-ybar	Yi (EST)	Yi-Yi(EST)	(Yi-Yi(EST))**2	(Yi-Ybar)**2
2	28	-74.4166	36.1224	-8.1224	65.97338176	5537.830356
3	49	-53.4166	51.1324	-2.1324	4.54712976	2853.333156
4	65	-37.4166	66.1424	-1.1424	1.30507776	1400.001956
4	75	-27.4166	66.1424	8.8576	78.45707776	751.6699556
5	89	-13.4166	81.1524	7.8476	61.58482576	180.0051556
6	98	-4.4166	96.1624	1.8376	3.37677376	19.50635556
7	110	7.5834	111.1724	-1.1724	1.37452176	57.50795556
8	120	17.5834	126.1824	-6.1824	38.22206976	309.1759556
9	140	37.5834	141.1924	-1.1924	1.42181776	1412.511956
9	145	42.5834	141.1924	3.8076	14.49781776	1813.345956
10	150	47.5834	156.2024	-6.2024	38.46976576	2264.179956
10	160	57.5834	156.2024	3.7976	14.42176576	3315.847956
6.416667	102.4166				323.6520251	19914.91667

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
2012	1056	51

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

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 standard error of β_1 is given by

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

In above Eq. S_e 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 can be written as

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

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)

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

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

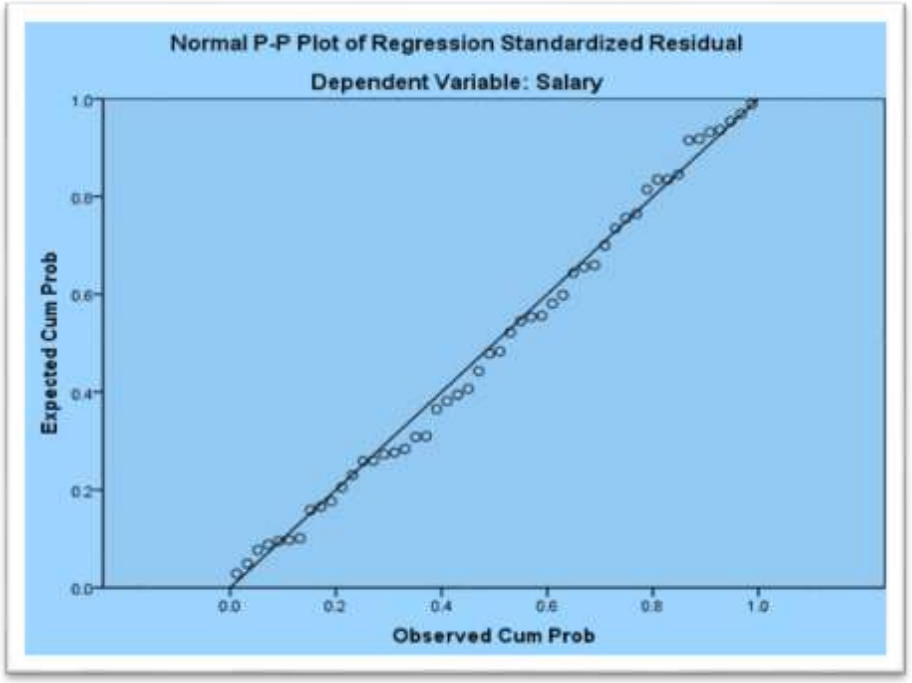
Residual Analysis

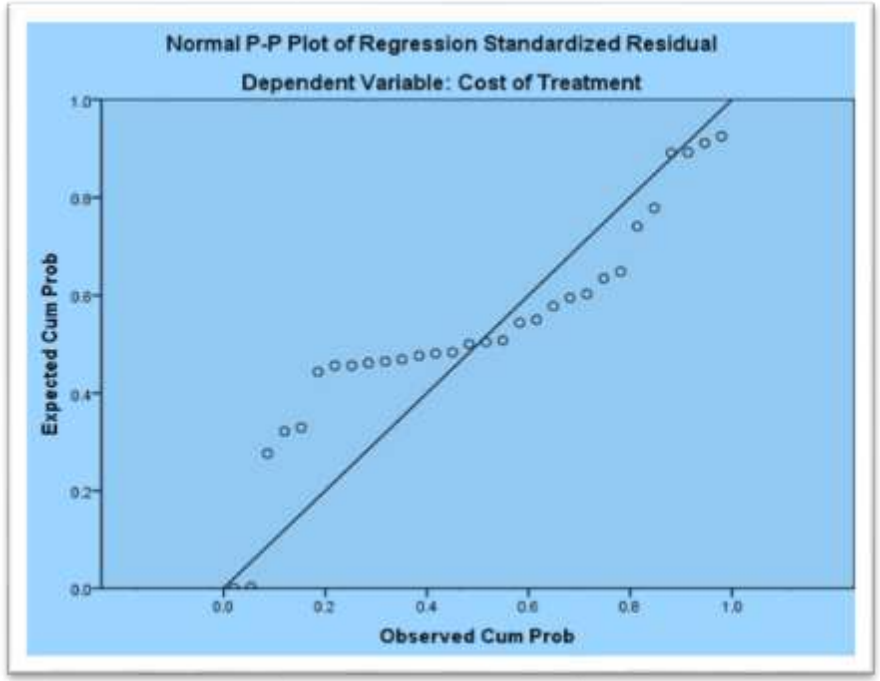
Residual (error) analysis is important to check whether the assumptions of regression models have been satisfied. It is performed to check the following:

- 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





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*).

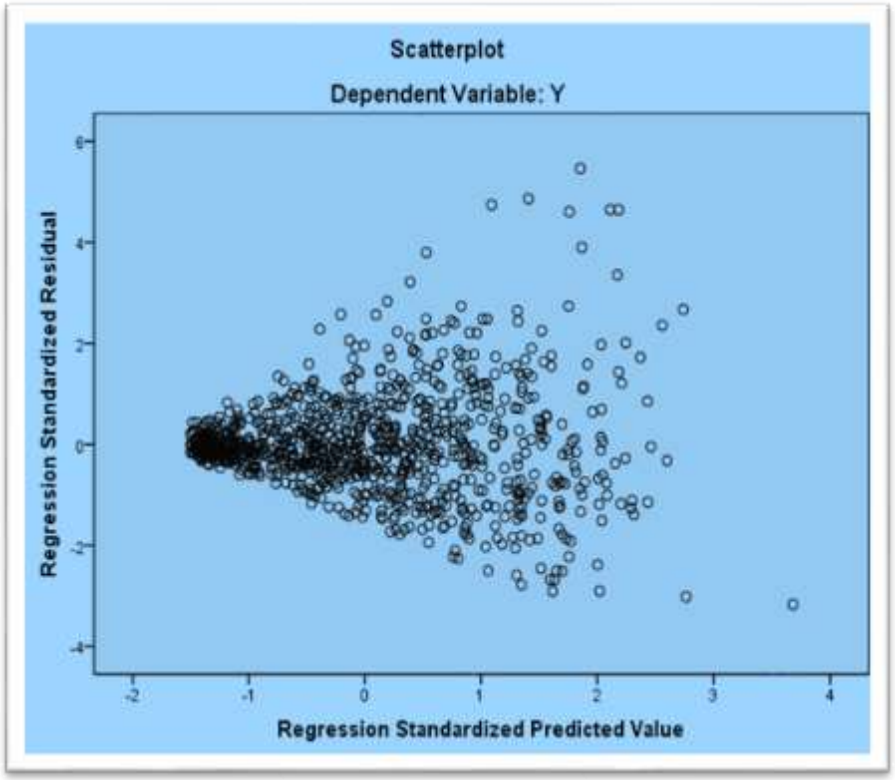
That is, the variance of residuals is assumed to be independent of variable *X*. Failure to meet this assumption will result in unreliability of the hypothesis tests.

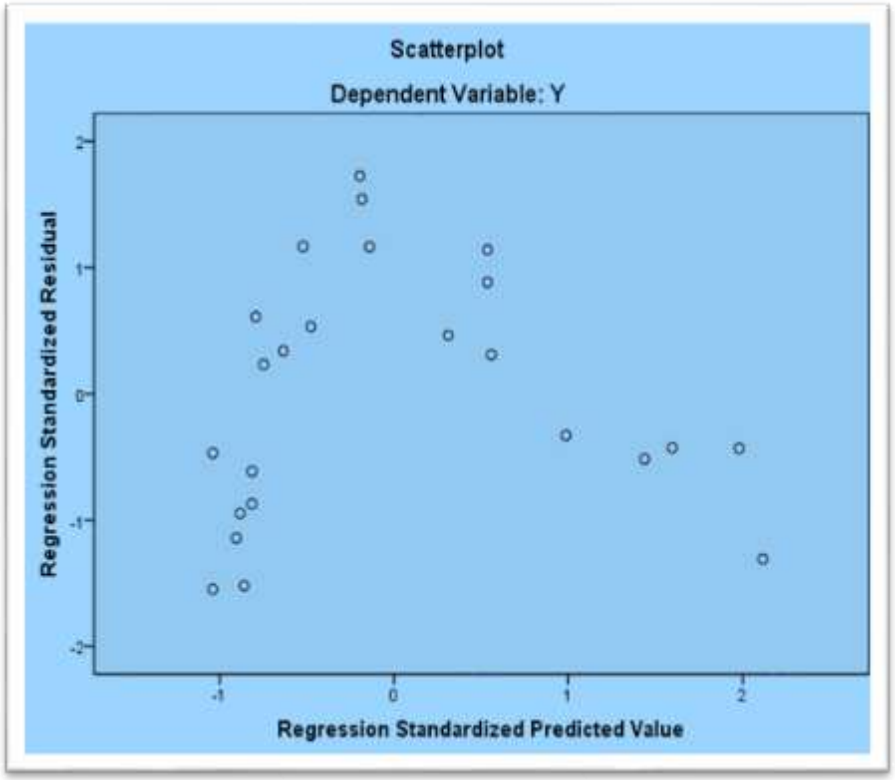
Testing the Functional Form of Regression Model

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

Testing the Functional Form of Regression Model

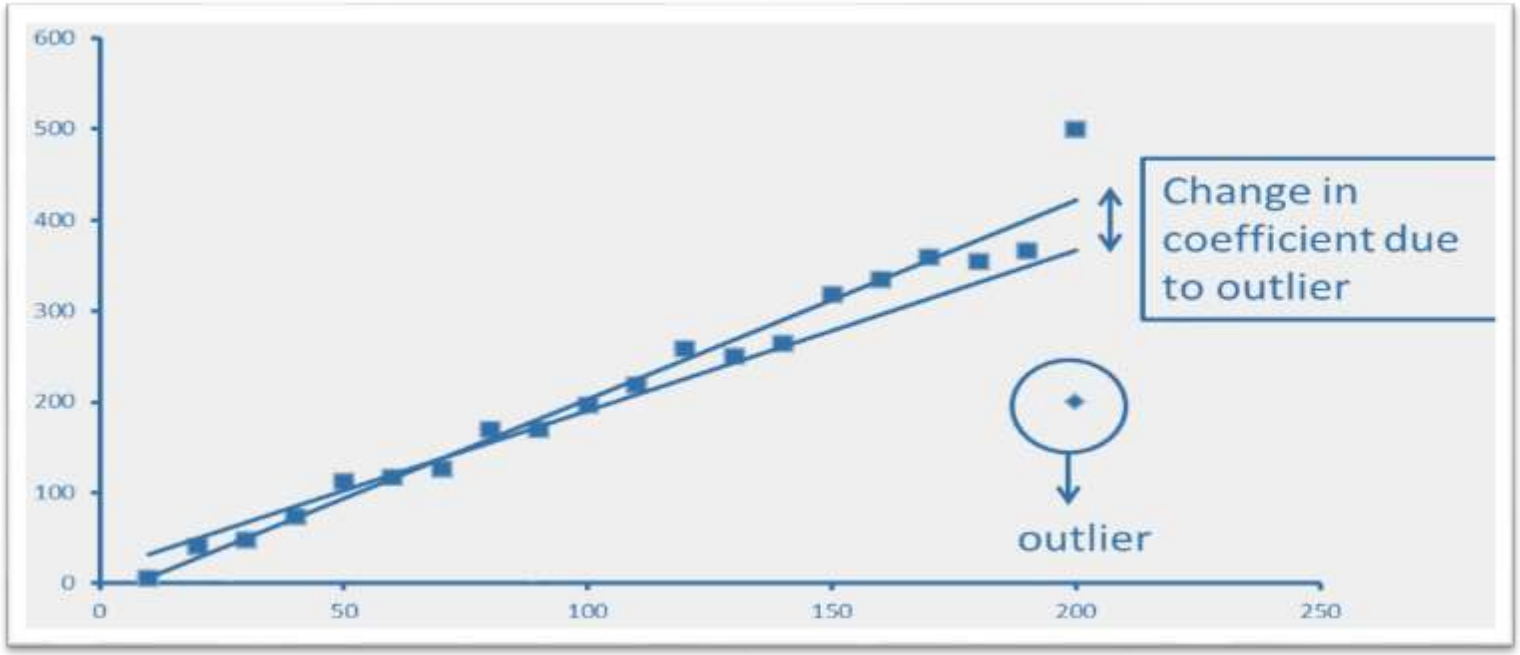
Any pattern in the residual plot would indicate incorrect specification (misspecification) of the model.





Outlier Analysis

- Outliers are observations whose values show a large deviation from mean value, that is ( ) large
- 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



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.

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

ASSUMPTIONS

- The regression model is linear in parameter.
- In a time series data, residuals are uncorrelated, that is, Cov(ε_i, ε_j) = 0 for all $i \neq j$.
- 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 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 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.

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.

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 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

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

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.

Example

The data in below table 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. Discuss the insights based on the regression output.

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

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.

Validation of Multiple Regression Model

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

- Coefficient of multiple determination (*R*-Square) and Adjusted *R*-Square, which can be used to judge the overall fitness of the model.

- *t*-test to check the existence of statistically significant relationship between the response variable and individual explanatory variable at a given significance level (α) or at (1 – α)100% confidence level.

- *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.

Co-efficient of Multiple Determination (*R*-Square) and Adjusted *R*-Square

As in the case of simple linear regression, *R*-square measures the proportion of variation in the dependent variable explained by the model. The co-efficient of multiple determination (*R*-Square or R^2) is given by

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

- *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

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

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:

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

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

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

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

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

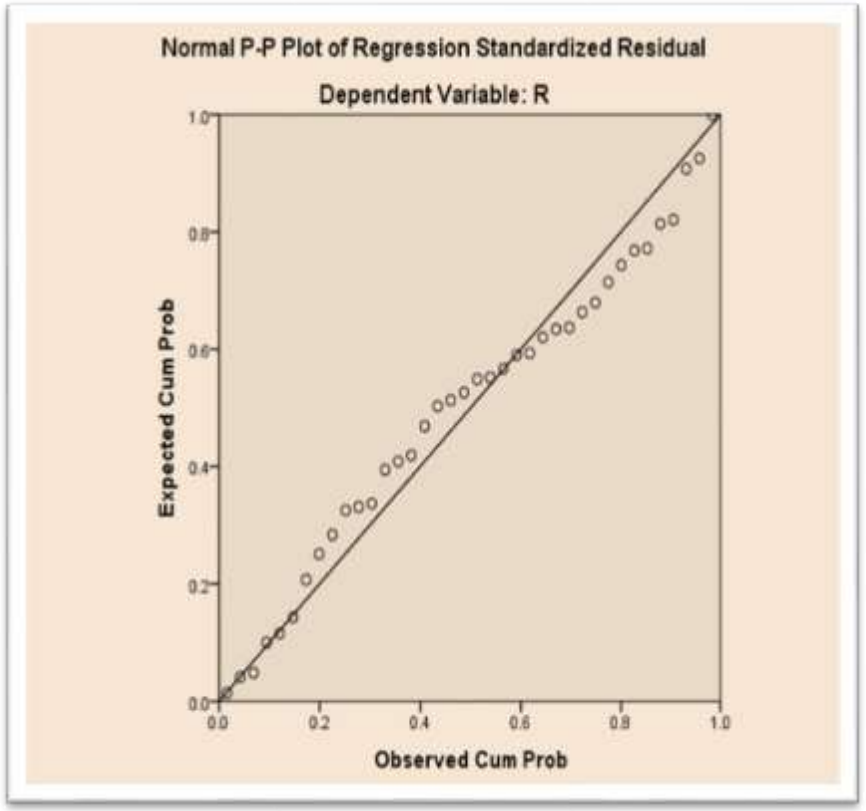
H_1: Not all βs are zero.

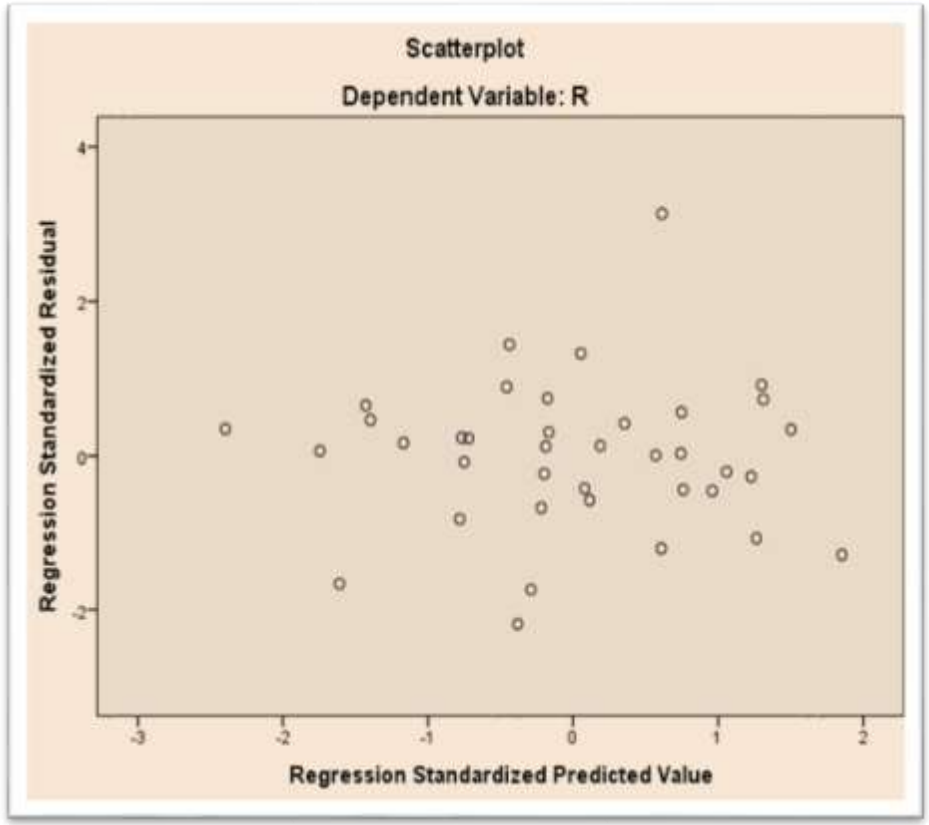
F-statistic is given by:

$$F = MSR/MSE$$

Residual Analysis in Multiple Linear Regression

Residual analysis is important for checking assumptions about normal distribution of residuals, homoscedasticity, and the functional form of a regression model.





Multi-Collinearity and Variance Inflation Factor

Multi-collinearity can have the following impact on the model:

- The standard error of estimate of a regression coefficient may be inflated, and may result in retaining of null hypothesis in *t*-test, resulting in rejection of a statistically significant explanatory variable.
- The *t*-statistic value is $\left(\hat{\beta} / S_e(\hat{\beta})\right)$
- If is inflated, then the *t*-value will be underestimated resulting in high *p*-value that may result in failing to reject the null hypothesis.
- Thus, it is possible that a statistically significant explanatory variable may be labelled as statistically insignificant due to the presence of multi-collinearity.

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.