



# 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**

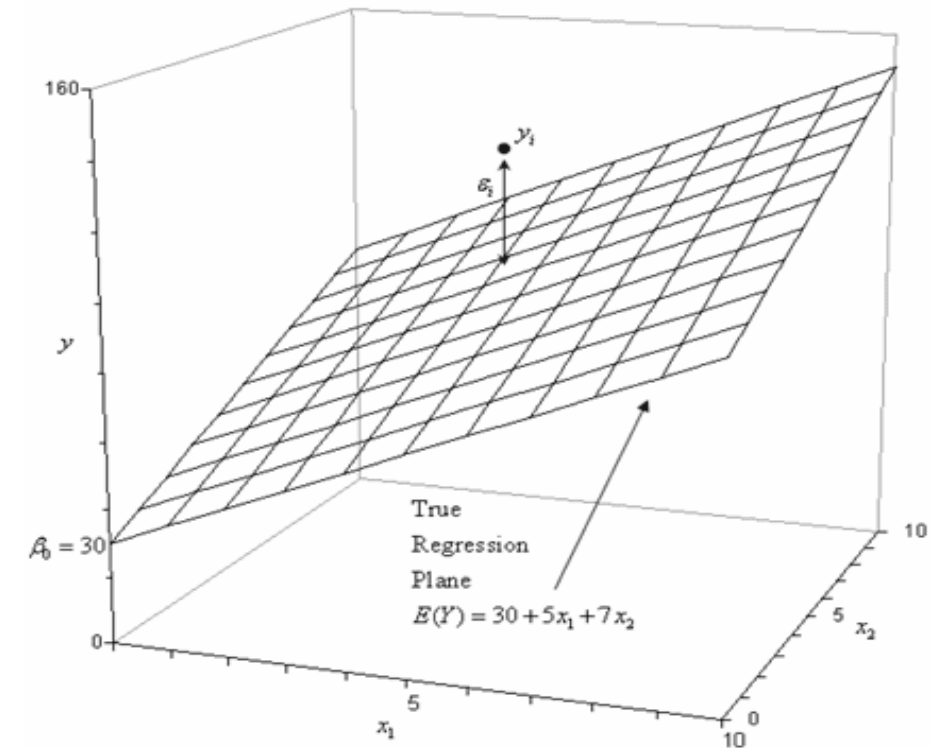
Department of Computer Science and Engineering



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

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

*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,  $\varepsilon_1$  is the variation in  $R$  (revenue generated through advertisement) not explained by  $CTRP$ .

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

Here  $\varepsilon_2$  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  $\varepsilon_1$  (variation in advertisement revenue not explained by  $CTRP$ ) and  $\varepsilon_2$  (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
$R = \beta_0 + \beta_1 \times CTRP + \beta_2 \times P + \varepsilon$	$R = 41008.84 + 5931.85 \times CTRP + 3.136 \times P$	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 $\eta_1$ is same as that of $\beta_2$ )	$\varepsilon_1$ is variation in $R$ not explained by $CTRP$ $\varepsilon_2$ 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.

- 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  $Y = \beta_0 + \beta_1 \times \text{Education}$ , 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  $\beta_0$ , 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.

# DATA ANALYTICS

## 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 ( $\beta_0$ ).

### 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

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



## DATA ANALYTICS

### 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.

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, *R*<sup>2</sup> 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.

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

## 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

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

- *F*-test to check the statistical significance of the overall model at a given significance level ( $\alpha$ ) or at  $(1 - \alpha)100\%$  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

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

$H_1$ : Not all  $\beta$ '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 ( $X_{r+1}, X_{r+2}, \dots, X_k$ ) in the full model are statistically significant.

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

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

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

- 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) 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

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

- 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 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  $\rho$  be the correlation between error terms  $(\varepsilon_t, \varepsilon_{t-1})$ . The null and alternative hypotheses are stated below:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

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  $D > D_U$ , 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.

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

- Mahalanobis Distance

- Overcomes drawbacks of Euclidean distance
- Helps find outliers

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

- Cook's Distance

- Measures change in regression parameters
- How does y change when a sample is left out?

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

- Leverage Values

- Influence of an observation on the overall fit

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

- 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

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

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

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

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

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

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  $\alpha$ .) 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  $\alpha$ .

**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  $\alpha$ .

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

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

That is  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$ . 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  $\alpha$  (or the  $F$ -value is less than the critical  $F$ -value).

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



- Stepwise regression is a combination of forward selection and backward elimination procedure
- In this case, we set the entering criteria ( $\alpha$ ) for a new variable to enter the model based on the smallest  $p$ -value of the partial  $F$ -test and removal criteria ( $\beta$ ) 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 $C_p$

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. Mallows'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.

## 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 =  $\alpha_0 + \alpha_1 \text{CTRP} + e_2$ )
  - High VIF for a variable  $\Rightarrow$  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  $y_{\text{hat}}$ ) 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.

### **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](mailto:mamathahr@pes.edu)**

+91 80 2672 1983 Extn 834