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.

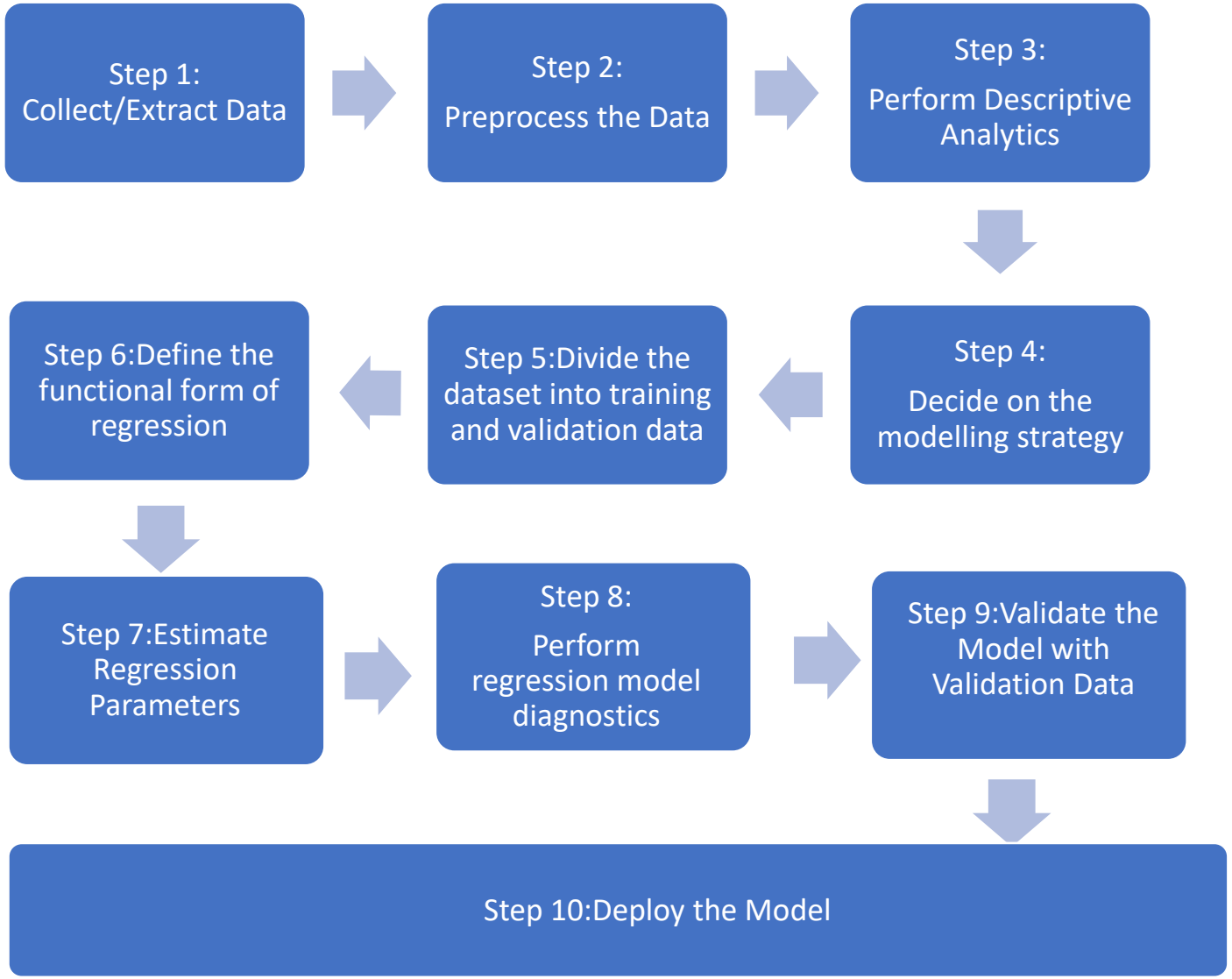
Multiple Linear Regression Model Building

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.

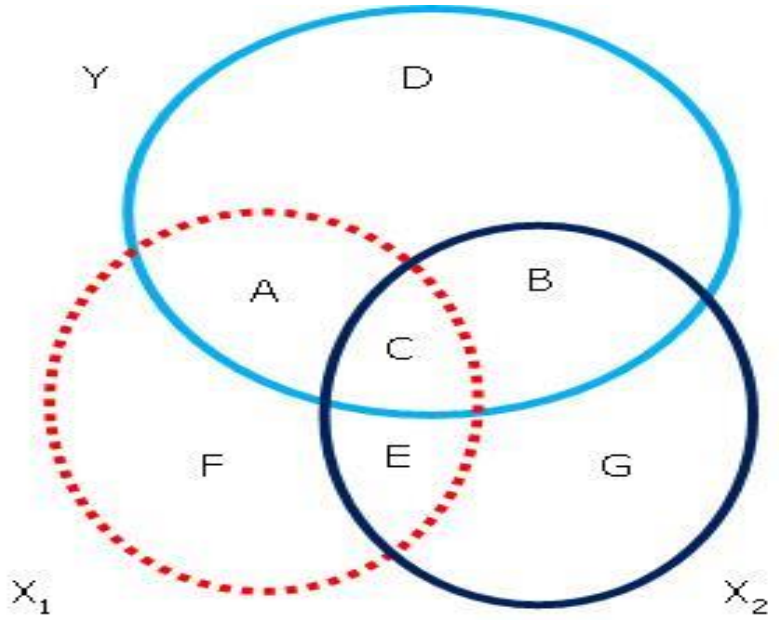
Let r_{YX_1,X_2} denote the partial correlation between Y and X_1 when X_2 is kept constant. Then r_{YX_1,X_2} is given by

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

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	122400	1320648	36	120	194400	1415316
18	96	158400	1102704	37	128	176400	1338060
19	104	165600	1184316	38	117	172800	1457400

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.

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 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

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 = 1)

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

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.

DATA ANALYTICS

- *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.

DATA ANALYTICS

Exercise

- To be done

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