DATA ANALYTICS

Unit 2:Multiple Linear Regression-2

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

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), $\sqrt{\sum_{i=1}^{K} \frac{1}{n} \left(Y_i - \hat{Y}_i \right)^2}$

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).

DATA ANALYTICS

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. 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 = 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 (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.

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.

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

DATA ANALYTICS

Example

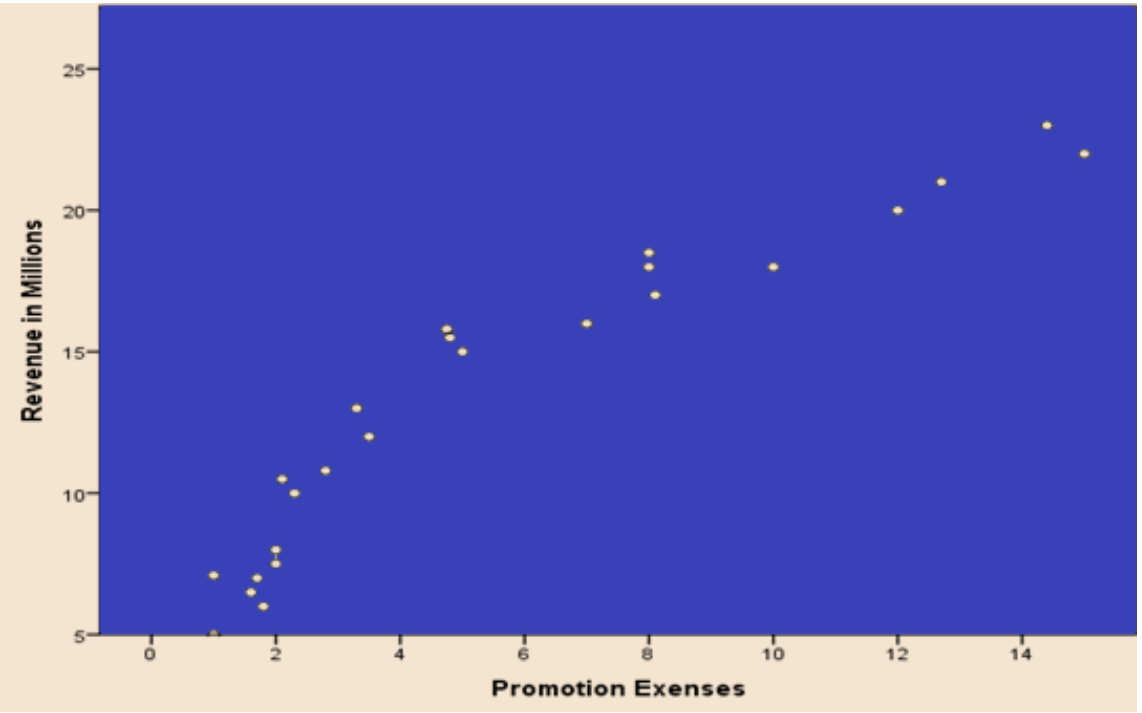
Table shows the data on revenue generated (in million of rupees) from a product and the promotion expenses (in million of rupees). Develop a regression model

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

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.



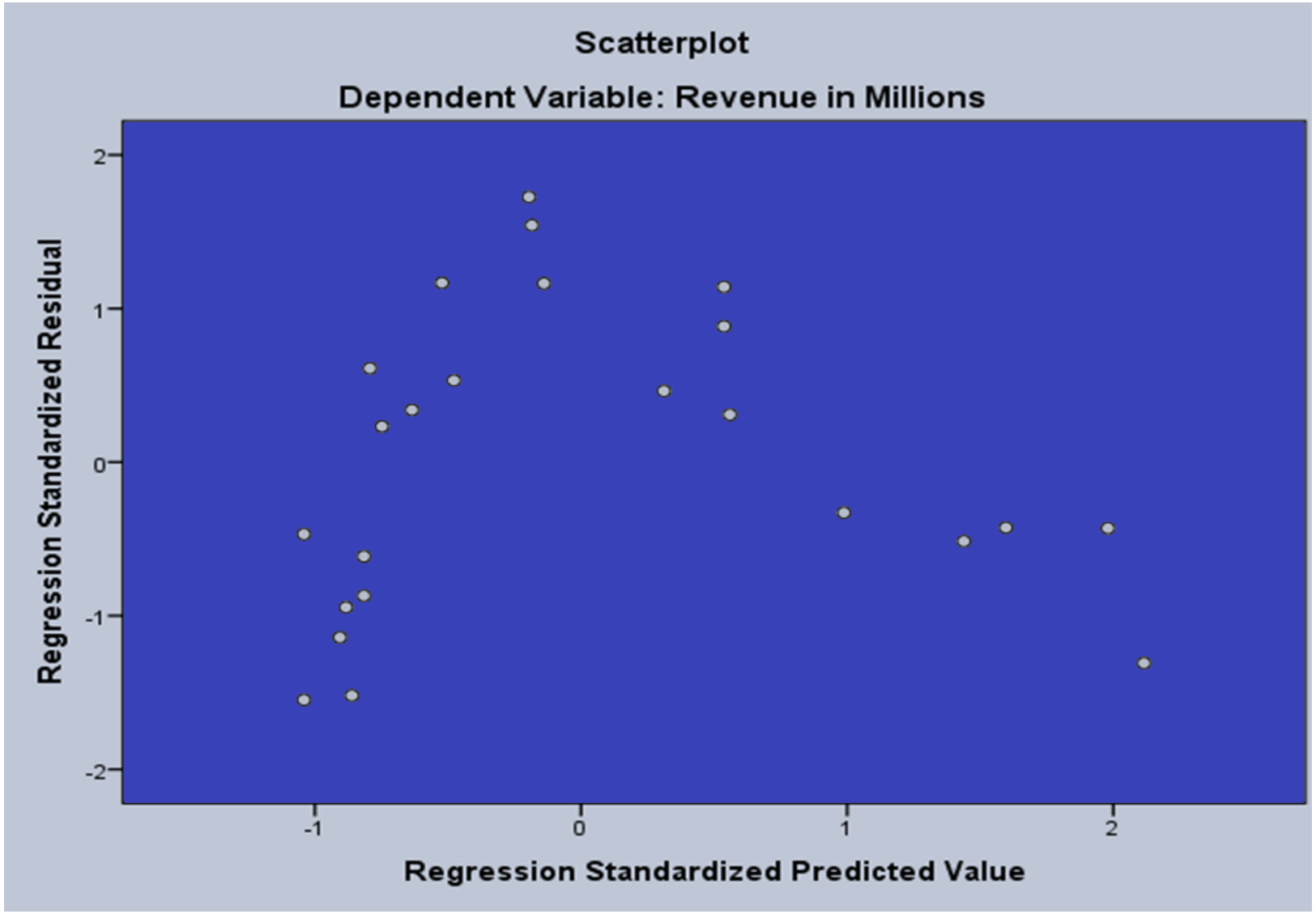
Consider the function $Y = \beta_0 + \beta_1 X$. The output for this regression is shown in below tables and in Figure . There is a clear increasing and decreasing pattern in Figure indicating non-linear relationship between *X* and *Y*.

Model Summary

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

Coefficients

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


Scatterplot
Dependent Variable: Revenue in Millions
Regression Standardized Residual
2
1
0
-1
-2
-1
0
1
2
Regression Standardized Predicted Value

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 $Y = \beta_0 + \beta_1 \ln(X)$ is shown in Tables 10.31 and 10.32 and the residual plot is shown in Figure 10.11.

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 $Y = \beta_0 + \beta_1 X$.

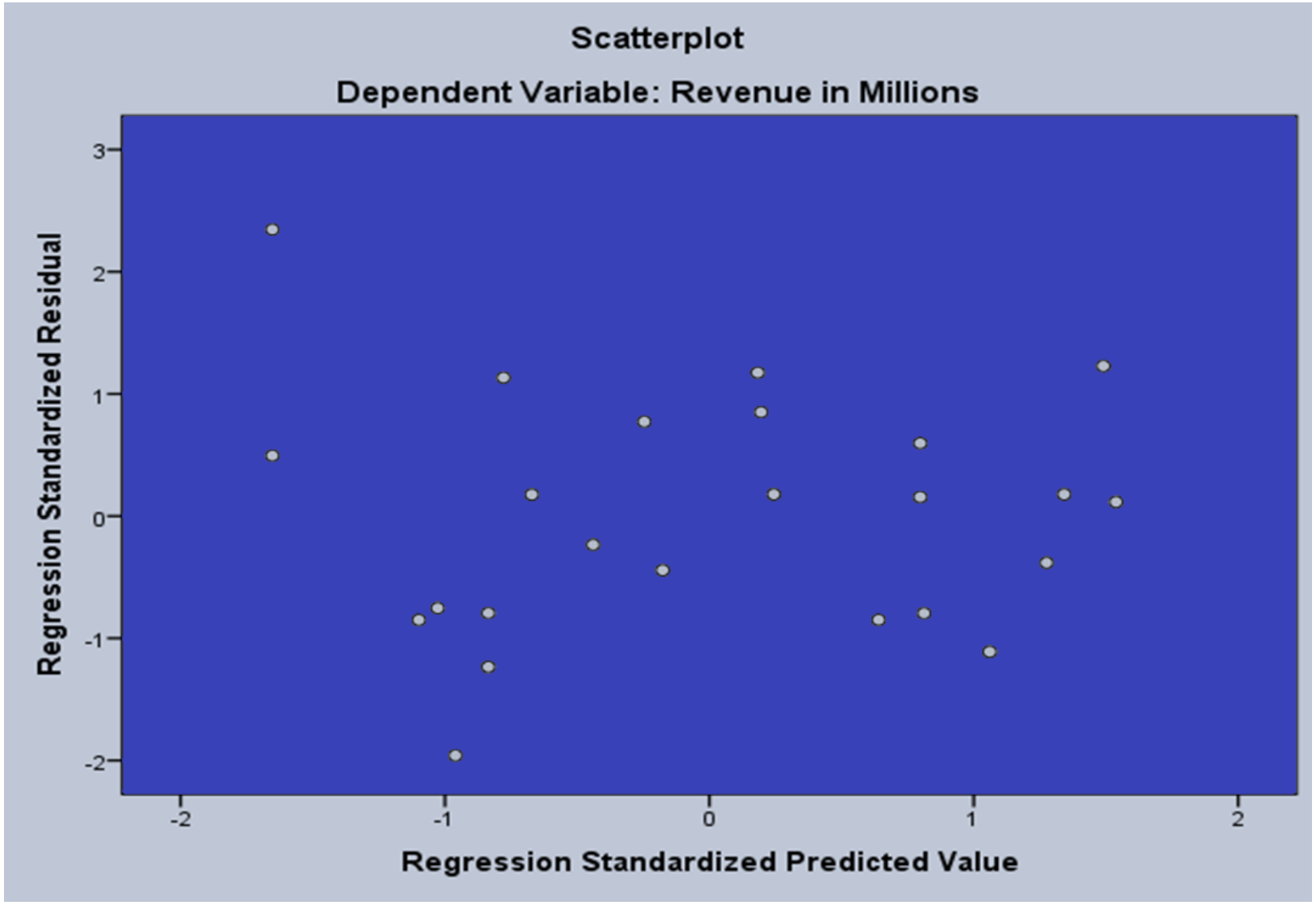
Model Summary

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

Coefficients

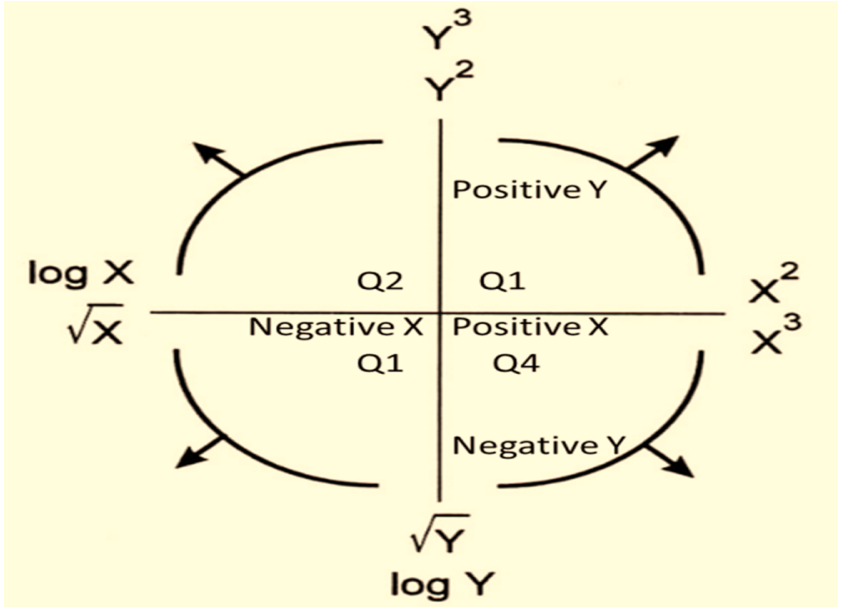
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.439	0.454		9.771	0.000
	ln (X)	6.436	0.279	0.980	23.095	0.000

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.)

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834