DATA ANALYTICS

Unit 2:Multiple Linear Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Multiple Linear Regression Contd.,

Mamatha H R

Department of Computer Science and Engineering

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

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

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:

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

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

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

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

$$H_0: \updownarrow_1 = \updownarrow_2 = \updownarrow_3 = \ldots = \updownarrow_k = 0$$

H_1: Not all s are zero.

F-statistic is given by:

$$F = MSR/MSE$$

Validation of Portions of a MLR Model – 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.

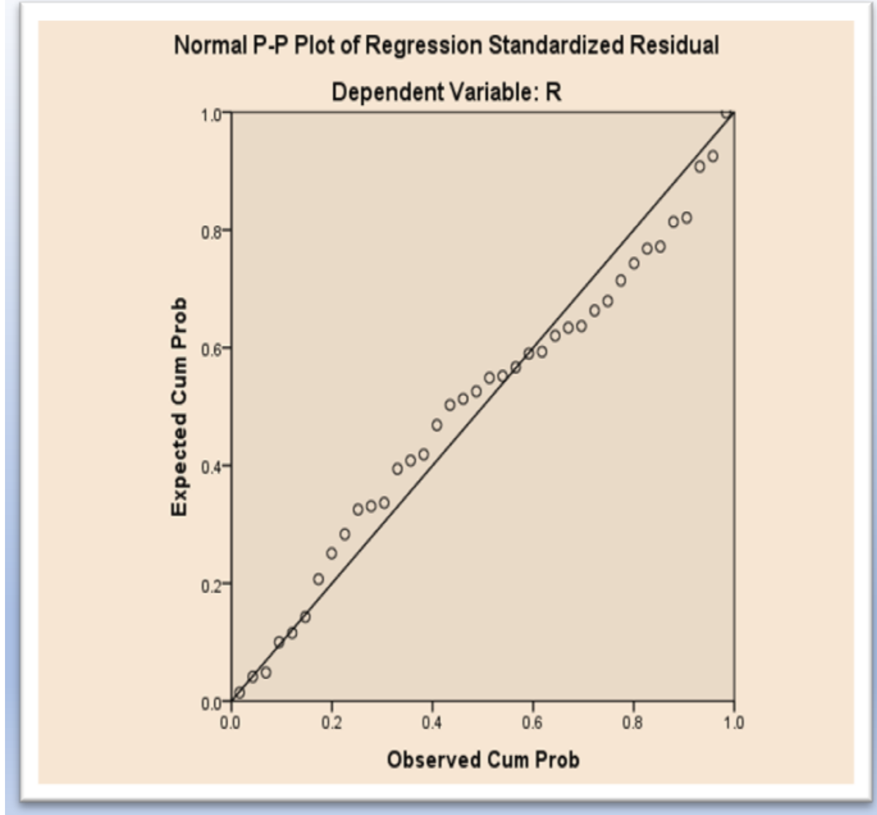
The corresponding partial *F*-test has the following null and alternative hypotheses:

- 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

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

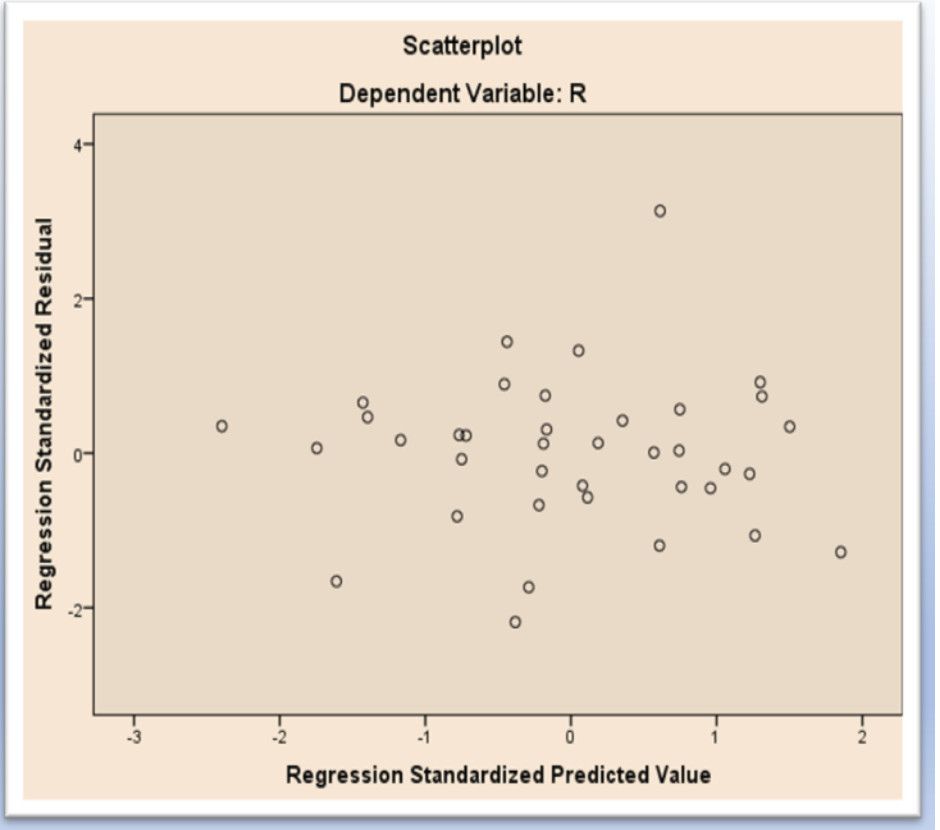
Residual Analysis in Multiple Linear Regression

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



Normal P-P Plot of Regression Standardized Residual

Dependent Variable: R



Scatterplot

Dependent Variable: R

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
- 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:

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2}$$

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

$$\boxed{X_1 = \alpha_0 + \alpha_1 X_2}$$

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

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

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

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

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:

$$\boxed{Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t}$$

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

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

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

❑ Cook’s Distance

❑ Leverage Values

❑ DFFIT and DFBETA Values

Mahalanobis Distance

- Mahalanobis distance (1936) is a distance between a specific observation and the centroid of all observations of the predictor variables.
- Mahalanobis distance overcomes the drawbacks of Euclidian distance while measuring distances between multivariate data.
- Mathematically, Mahalanobis distance, DM, is given by (Warrant et al. 2011)

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

Cook’s Distance

- Cook’s distance (Cook, 1977) measures the change in the regression parameters and thus how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters.

- Cook’s distance for multiple linear regression is given by (Bingham 1977, Chatterjee and Hadi 1986)

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

Leverage Value (or Hat Value)

- Leverage value of an observation measures the influence of that observation on the overall fit of the regression function and is related to the Mahalanobis distance
- Leverage point hi is nothing but the i^{th} diagonal element of the hat matrix,
- Leverage value for an observation in MLR is given by

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

$$\mathbf{H = X(X^T X)^{-1} X^T}$$

DFFIT and SDFFIT

DFFIT measures the difference in the fitted value of an observation when that particular observation is removed from the model building. DFFIT is given by

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

where, $\widehat{\boldsymbol{Y}_i}$ is the predicted value of i^{th} observation including i^{th} observation, $\hat{Y}_{i(i)}$ is the predicted value of i^{th} observation after excluding i^{th} observation from the sample.

The standardized DFFIT (SDFFIT) is given by (Belsley *et al.* 1980, Ryan 1990)

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

$S_e(i)$ is the standard error of estimate of the model after removing ith observation and h_i is the ith diagonal element in the hat matrix. The threshold for DFFIT is defined using **Standardized DFFIT** (SDFFIT). The value of SDFFIT should be less than

$$2\sqrt{(k+1)/n}$$

DFBETA and SDFBETA

DFBETA measures the change in the regression coefficient when an observation "*i*" is excluded from the model building. DFBETA is given by

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

where DFBETA$_i$(j)is the change in the regression coefficient for independent variable j when observation i is excluded.

The standardized DFBETA value (SDFBETA) for observation i is given by (Belsley et al 1980, Ryan 1990)

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

SDFBETAi(j) is the standardized DFBETA value for variable j after removing observation i and $S_e(\hat{\beta}_{j(i)})$ is the standard error of $\hat{\beta}_j$ after removing observation 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). Note that 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

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

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.

Example

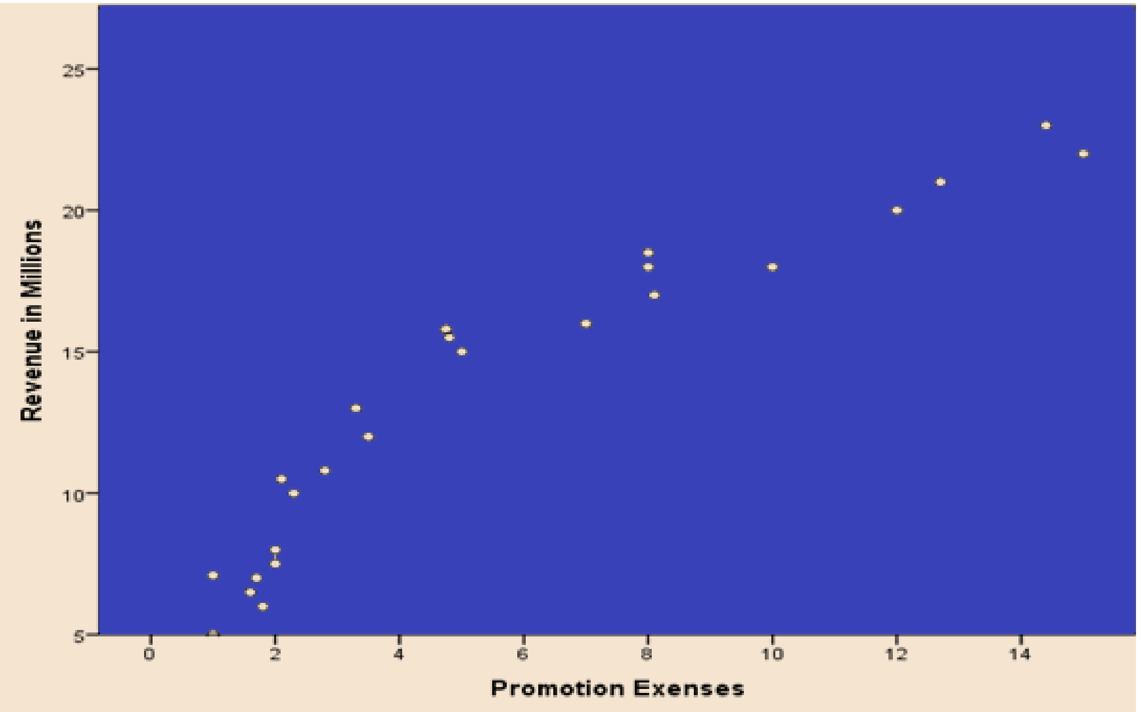
Table shows the data on revenue generated (in million of rupees) from a product and the promotion expenses (in million of rupees). Develop an appropriate regression model

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.



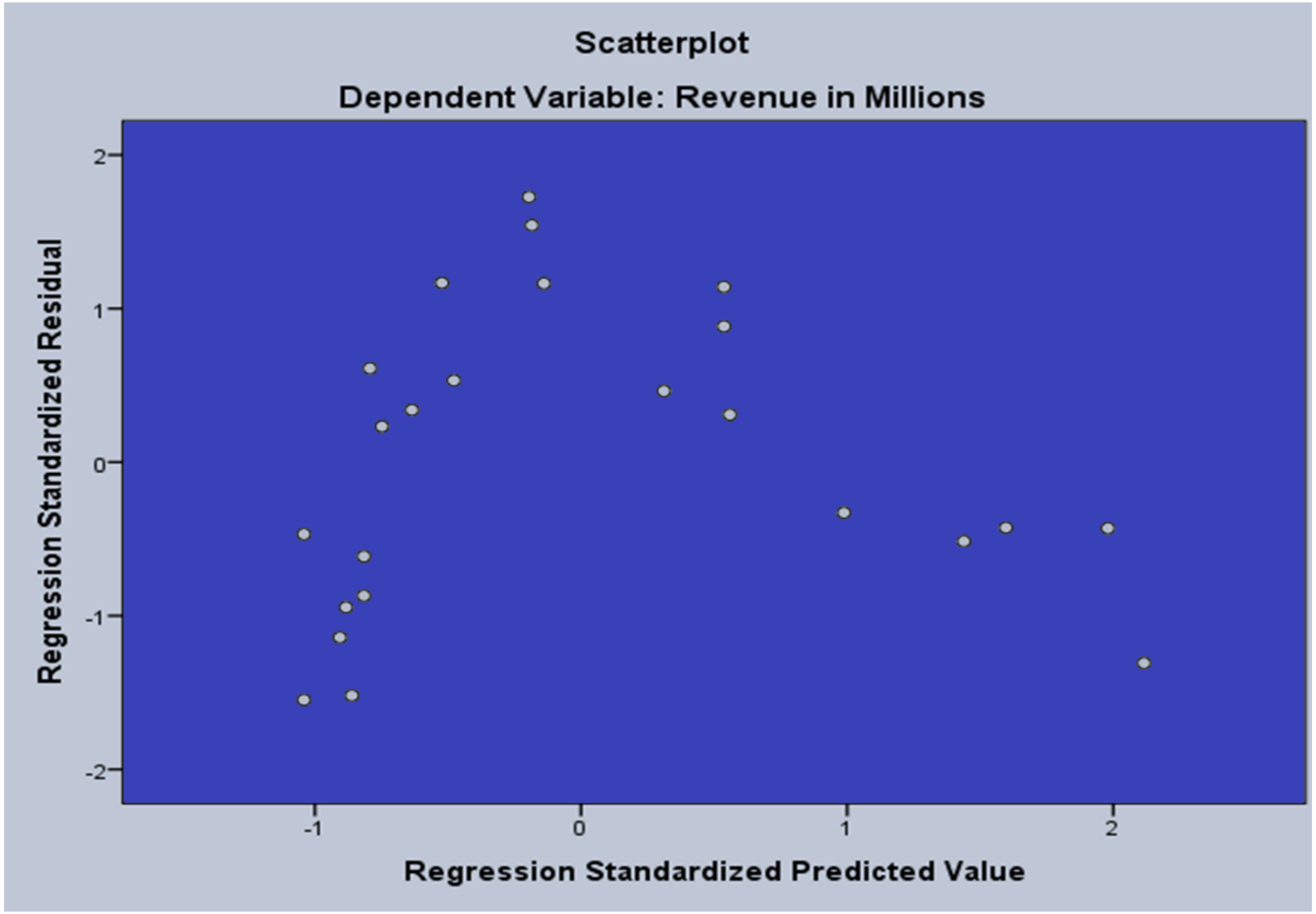
Consider the function $Y = \beta_0 + \beta_1 X$. The output for this regression is shown in below tables and in Figure . There is a clear increasing and decreasing pattern in Figure indicating non-linear relationship between X and Y.

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
Regression Standardized Predicted Value
2
1
0
-1
-2
-1
0
1
2

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 $Y = \beta_0 + \beta_1 X$.

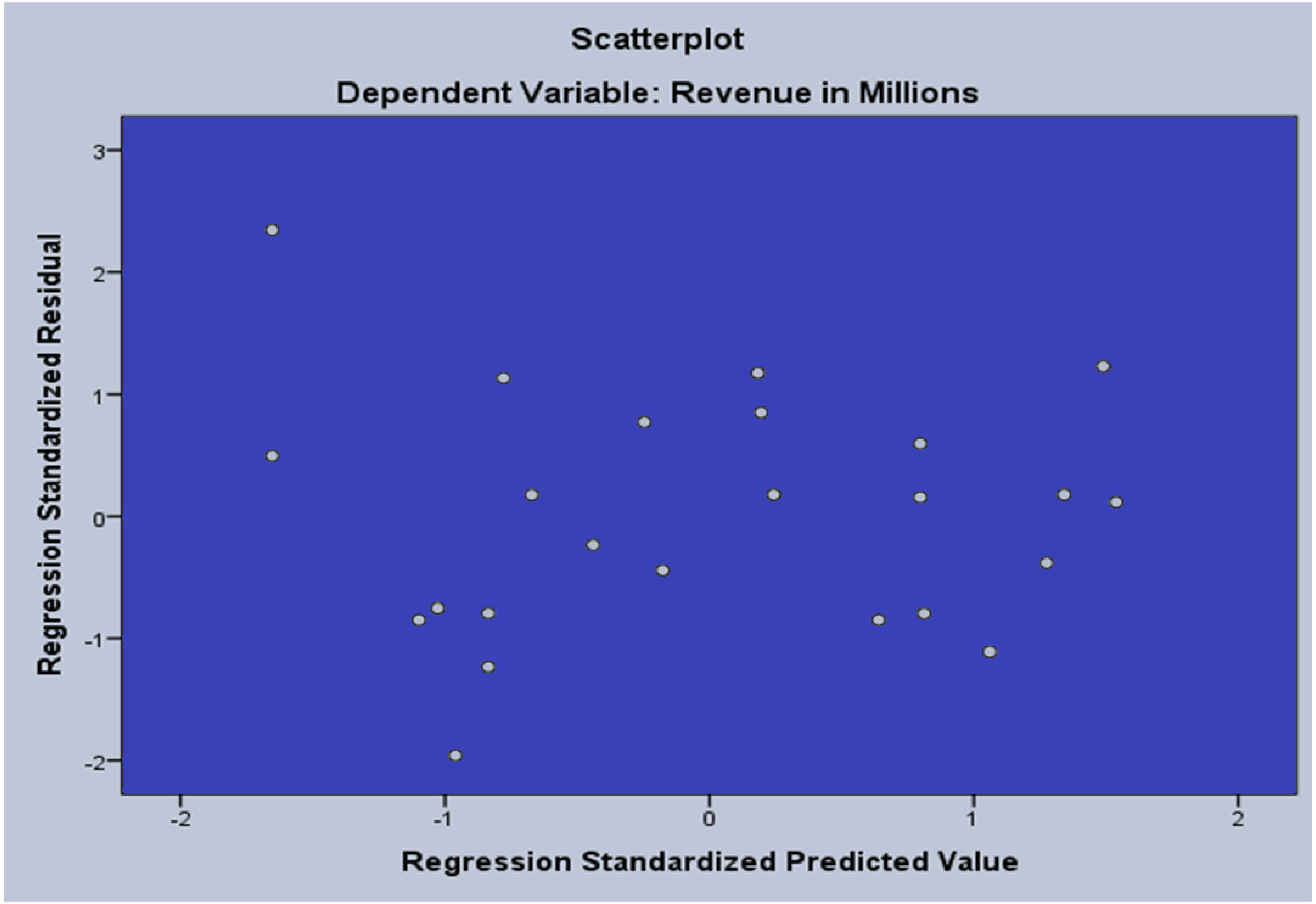
Model Summary

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 = $\beta_0 + \beta_1$ln(X).

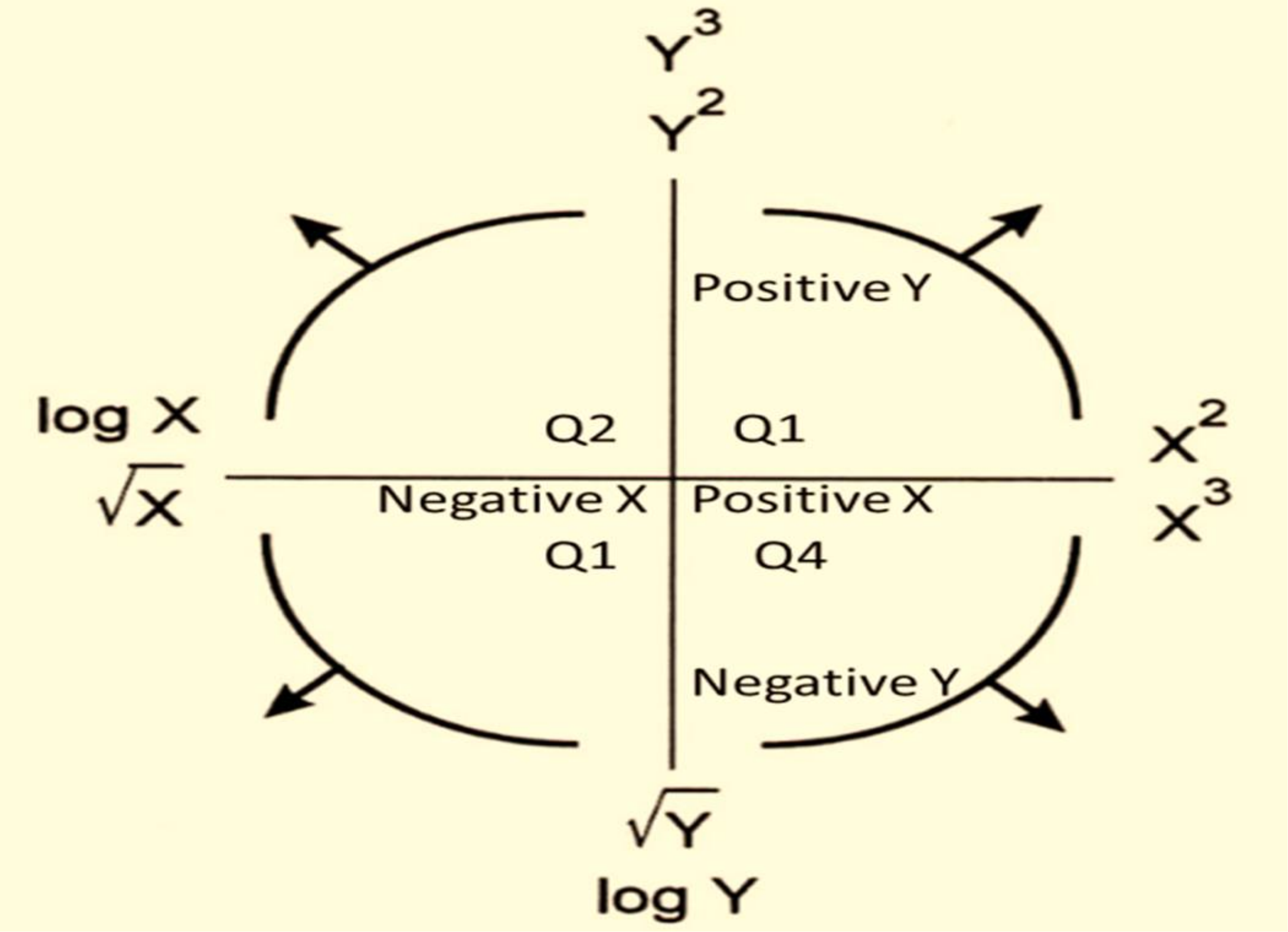


Tukey and Mosteller’s Bulging Rule for Transformation

- 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.

Tukey’s Bulging Rule (adopted from Tukey and Mosteller, 1977).



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