DATA ANALYTICS

Unit 2:Linear Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Linear Regression
Contd.,

Mamatha H R
Department of Computer Science and Engineering

Coefficient of Determination (R-Square or R2)

- 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

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

In absence of the predictive model for Y_i, the users will use the mean value of Y_i. Thus, the total variation is measured as the difference between Y_i and mean value of Y_i (i.e., Y_i -).

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

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

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

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

The *R*-square value for regression model between the number of deaths due to helium poisoning in UK and the number of Facebook users is 0.9928. That is, 99.28% variation in the number of deaths due to helium poisoning in UK is explained by the number of Facebook users.

The regression model is given as $Y = 1.9967 + 0.0465\,X$

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 estimate of β_1 using OLS is given by

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{\bar{X}\sum_{i=1}^{n}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Contd.,

Above eq. can be written as follows:

$$\beta_1 = \frac{\sum_{i=1}^{n} K_i Y_i}{\sum_{i=1}^{n} K_i^2} \text{ where } K_i = (X_i - \bar{X})$$

That is, the value of β_1 is a function of Y_i (K_i is a constant since X_i is assumed to be non-stochastic)

The standard error of β_1 is given by

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

Contd.,

In above Eq. Se 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 $\hat{\beta}_1$ can be written as

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

Contd.,

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

$$\boxed{t = \frac{\hat{\beta}_1 - \beta_1}{S_e(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{S_e(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)}}$$

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

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

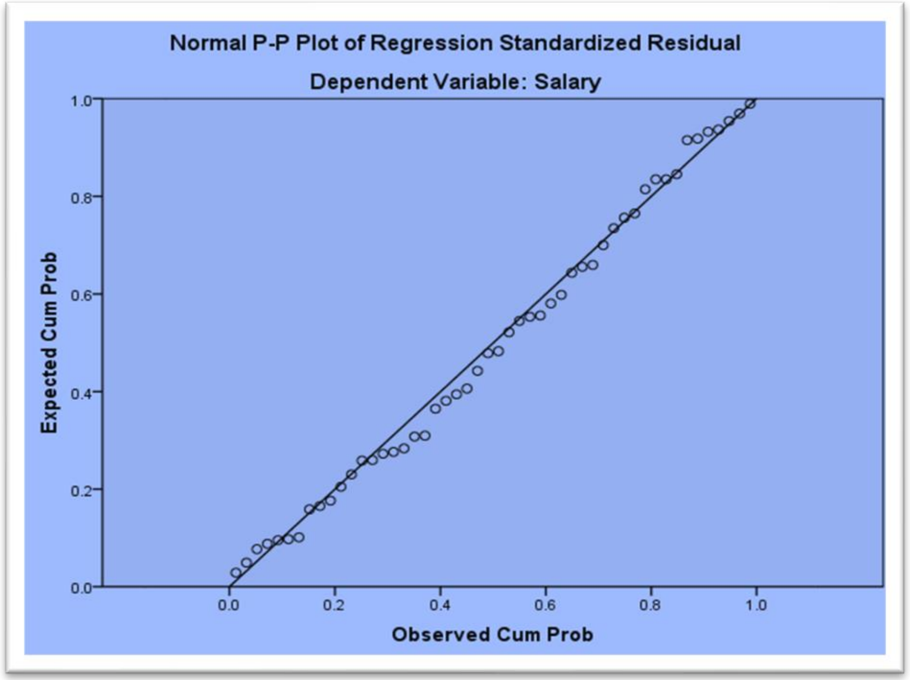
Residual Analysis

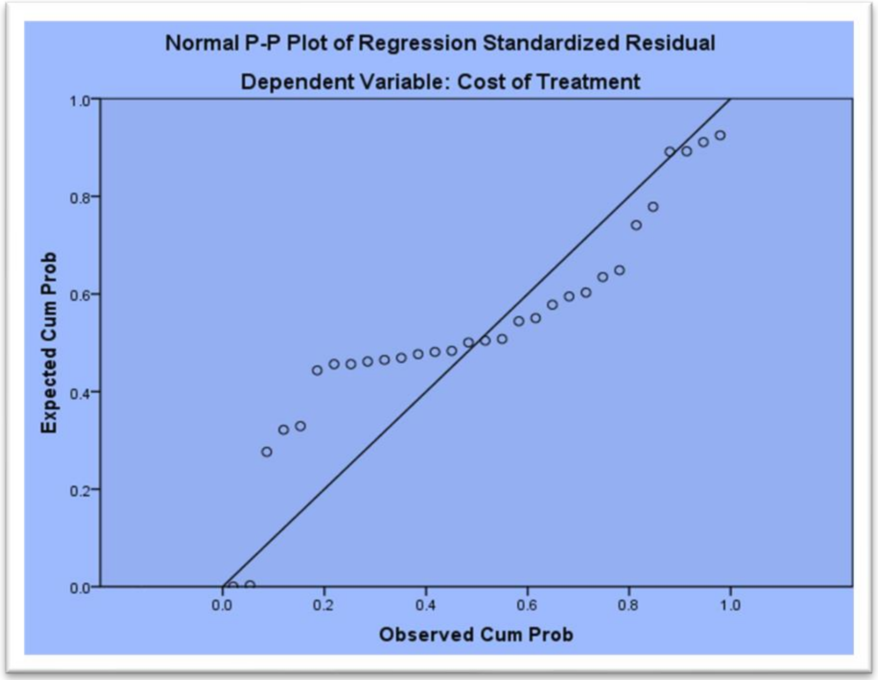
Residual (error) analysis is important to check whether the assumptions of regression models have been satisfied. It is performed to check the following:

- 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).

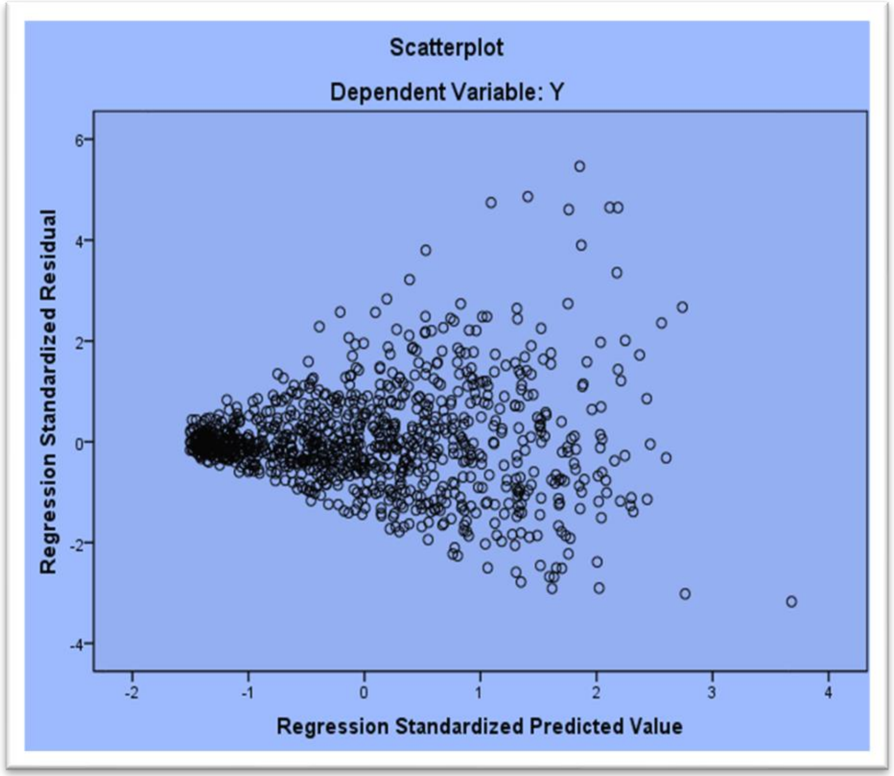
That is, the variance of residuals is assumed to be independent of variable X. Failure to meet this assumption will result in unreliability of the hypothesis tests.

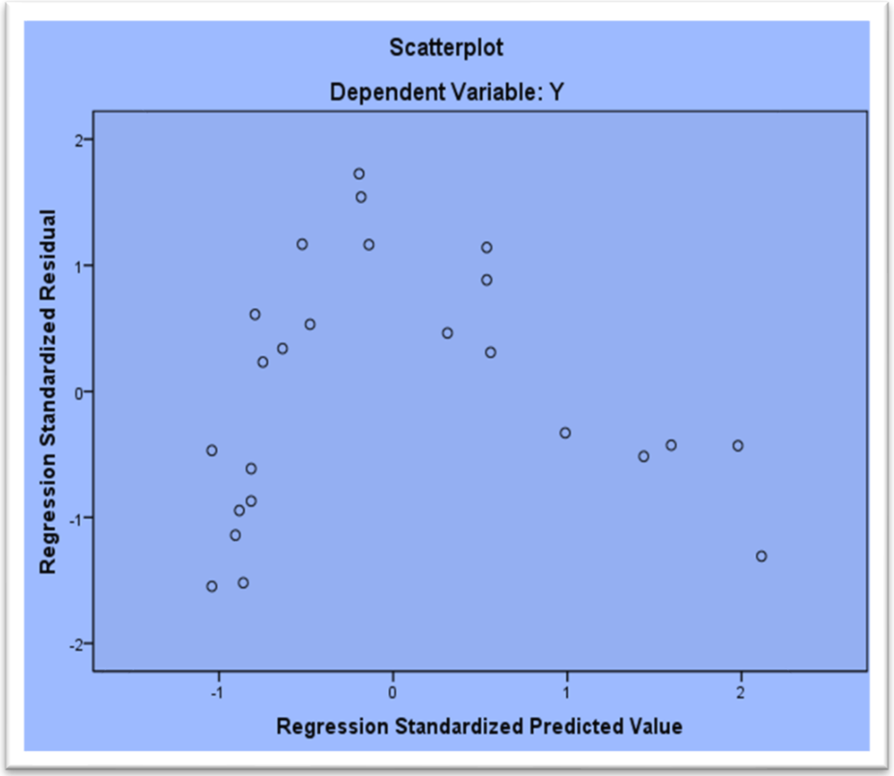
Testing the Functional Form of Regression Model

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

Testing the Functional Form of Regression Model

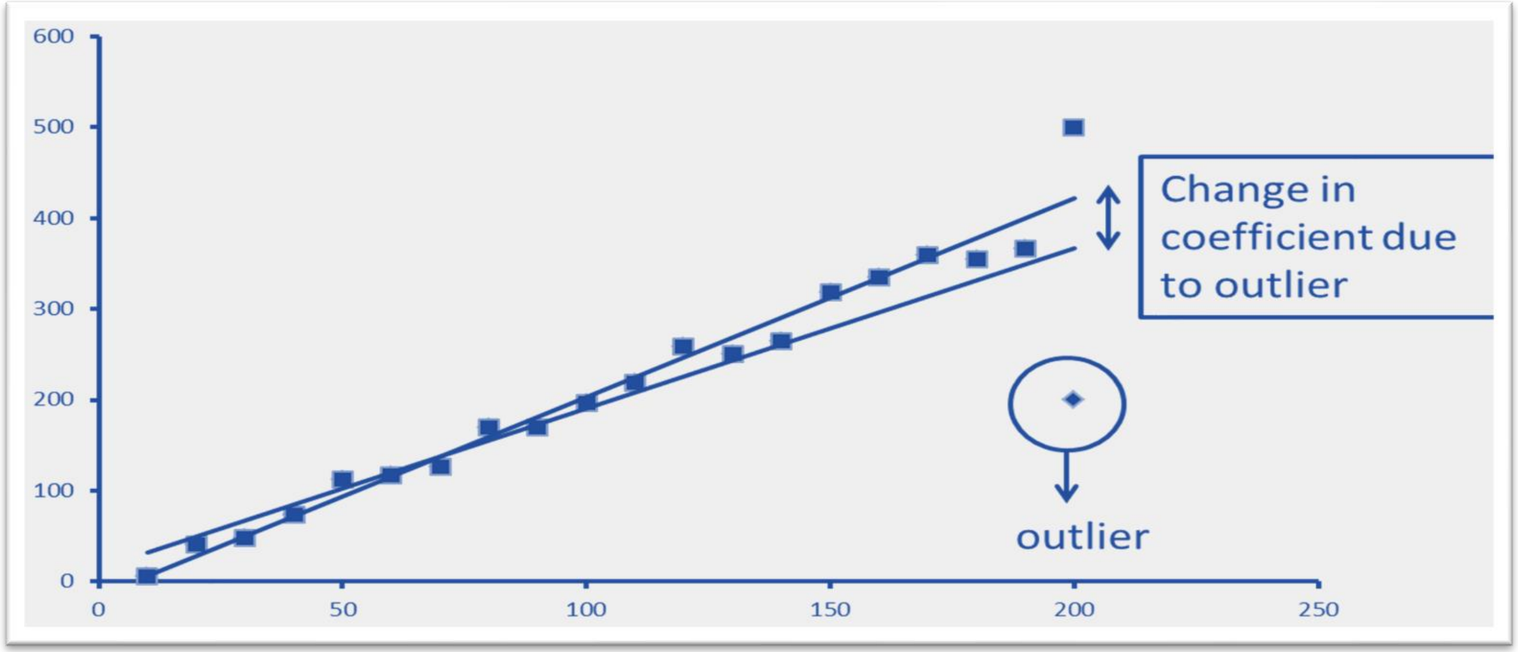
Any pattern in the residual plot would indicate incorrect specification (misspecification) of the model.





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 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.

Mahalanobis Distance

Mahalanobis distance is the distance between specific values of the independent variable (X_i) to the centroid of all observations of the explanatory variable. Distances value of more than chi-square critical value (with degrees of freedom is equal to the number of explanatory variables) is classified as outliers.

Cook's Distance

Cook's distance measures 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 simple linear regression is given by

$$D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(i)})^2}{MSE}$$

where D_i is the Cook's distance measure for i^{th} observation,

$\hat{Y}_{j(i)}$ is the predicted value of j^{th} observation including i^{th} observation,

$\hat{Y}_j$ is the predicted value of j^{th} observation after excluding i^{th} observation from the sample, MSE is the Mean–Squared–Error.

Leverage Value

Leverage value of an observation measures the influence of that observation on the overall fit of the regression function. Leverage value for an observation in SLR is given by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Leverage value of more than 2/*n* or 3/*n* is treated as highly influential observation. In Eq. the first term (1/*n*) will tend to zero for large value of *n*.

DFFit and DFBeta

- DFFit is the change in the predicted value of Y_i when case *i* is removed from the data set. DFBeta is the change in the regression coefficient values when an observation *i* is removed from the data.

Confidence Interval for Regression coefficients β0 and β1

The standard error of estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$S_e(\hat{\beta}_0) = \frac{S_e \times \sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{n \times SS_X}}$$

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{SS_X}}$$

where

$$S_e = \sqrt{\frac{\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

Where S_e is the standard error of residuals and SSX = $\sum_{i=1}^{n}(X_i - \bar{X})^2$

The interval estimate or (1-α)100% confidence interval for $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\hat{\beta}_1 \mp t_{\alpha/2,n-2} S_e(\hat{\beta}_1)$$

$$\hat{\beta}_0 \mp t_{\alpha/2,n-2} S_e(\hat{\beta}_0)$$

Confidence Interval for the Expected Value of Y for a Given X

- Since the point estimates are subjected to higher levels of error, due to uncertainties around estimation of parameters and natural variation in the data around the predicted line, the user would like to know the interval estimate or the **confidence interval** for the conditional expected value.
- The confidence interval of the expected value of Y_i for a given value of X_i is given by

$$\hat{Y}_i \pm t_{\alpha/2, n-2} \times S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- Where the term $S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$ is the standard error of E(Y|X).

Prediction Interval for the Value of Y for a Given X

The prediction interval of Y_i for a given value of X_i is given by

$$\hat{Y}_i \pm t_{\alpha/2,n-2} \times S_e \times \sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

where the term, $S_e \times \sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$ is the standard error of Yi for a given Xi value

Contd.,

For large *n*, the confidence interval of *E*(*Y*/*X*) will converge to

$$\hat{Y}_i \pm t_{\alpha/2,n-2} \times S_e$$

This is because, as n $\rightarrow \infty$, the term

$$\sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$ converges to 1

Exercise

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