Assumptions of Linear Regression

Major Assumptions

- Normality
- Linearity
- Homoscedasticity
- Multicollinearity

Assumption of Normality

- Normality of the error term (The error term is normally distributed)

Diagnosis

Histogram, steam and leaf display, box plot or normal probability plot of residulas

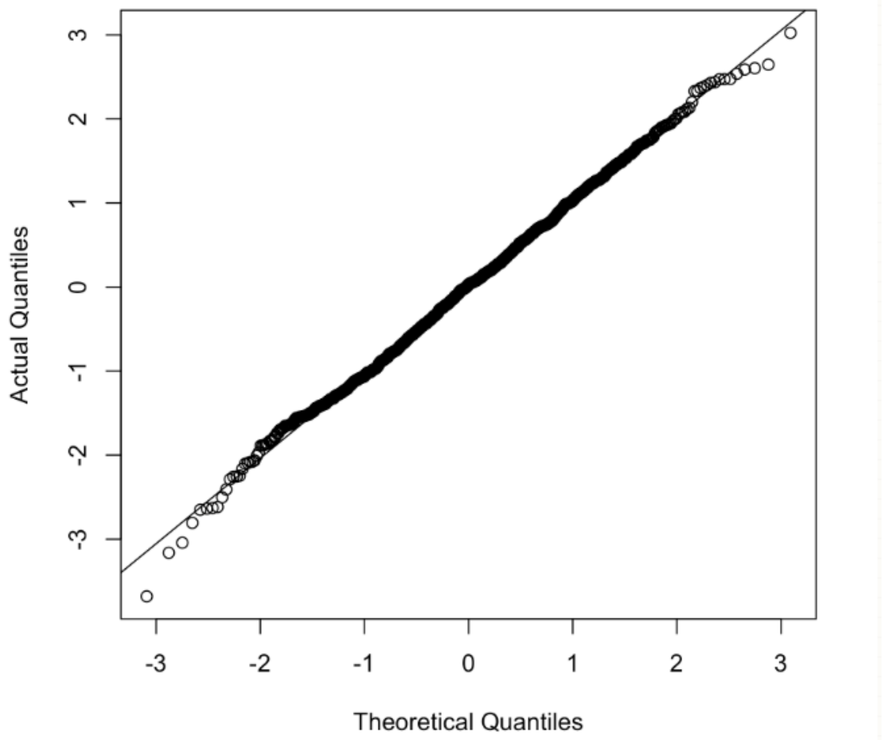


Figure 03: Normality Plot

Non-Normality : What to do ?

- If residuals are slightly depart from normality, no need to do any thing

- If residuals are very far from normality, then we can use various transformations

	X	
Y	X	$\log X$
Y	*linear* $\hat{Y}_i = \alpha + \beta X_i$	*linear-log* $\hat{Y}_i = \alpha + \beta \log X_i$
$\log Y$	*log-linear* $\log \hat{Y}_i = \alpha + \beta X_i$	*log-log* $\log \hat{Y}_i = \alpha + \beta \log X_i$

Assumption of Linearity

- The dependent variable retains a linear relationship with the independent variables

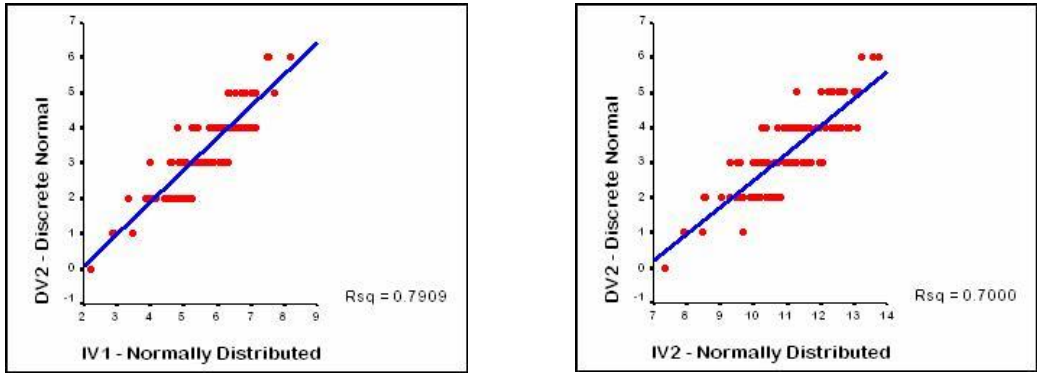


Figure 4: Linear Relationship

Homogeneity of Variance

- It is assumed that variance of error term is the same across all values of the independent variable

- Plot the standardized residuals against the predicted values. There should be equal spread.

Figure 5: Satisfactory Residual Plot

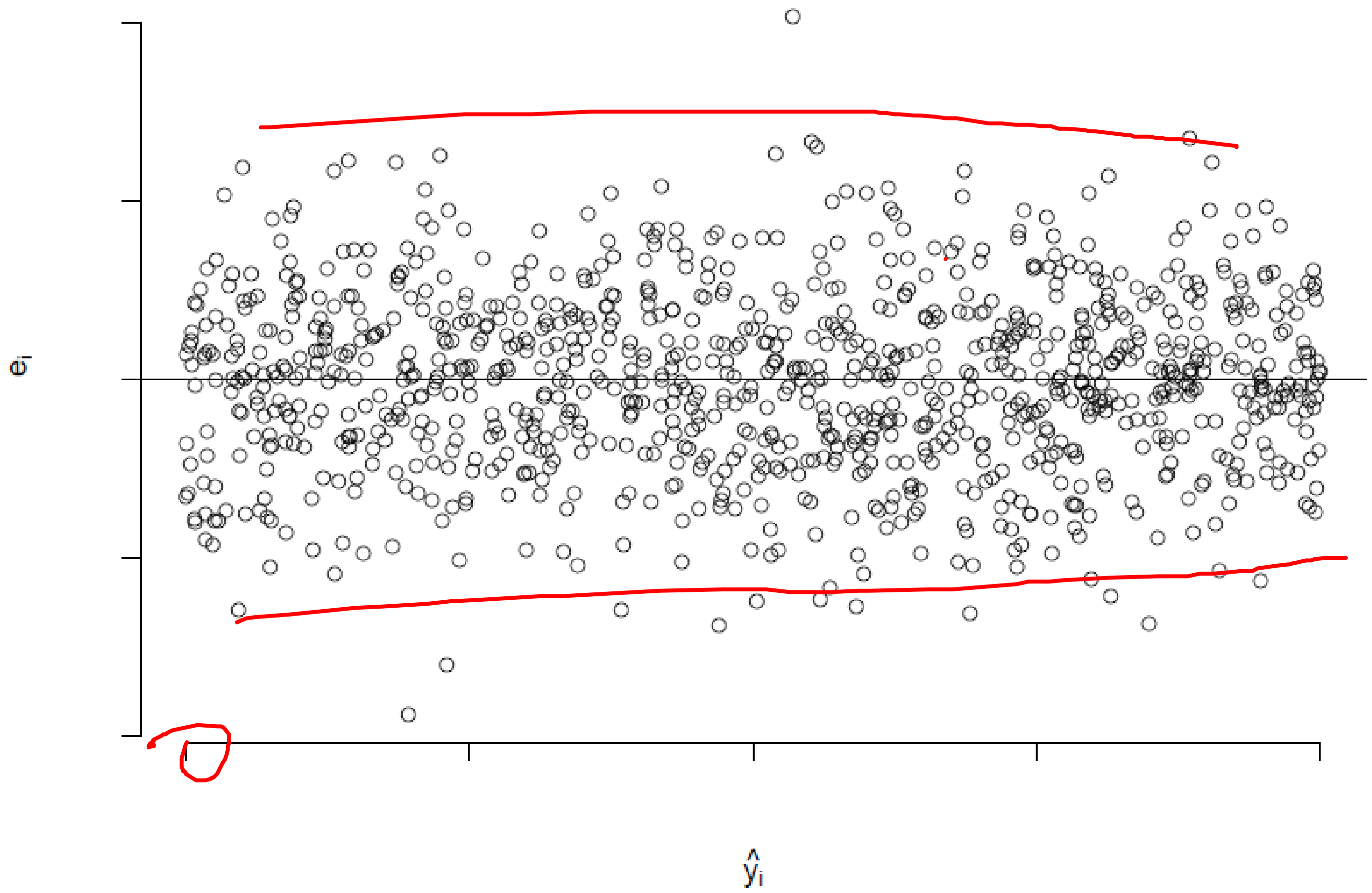


Figure 6: Non Constant Variance

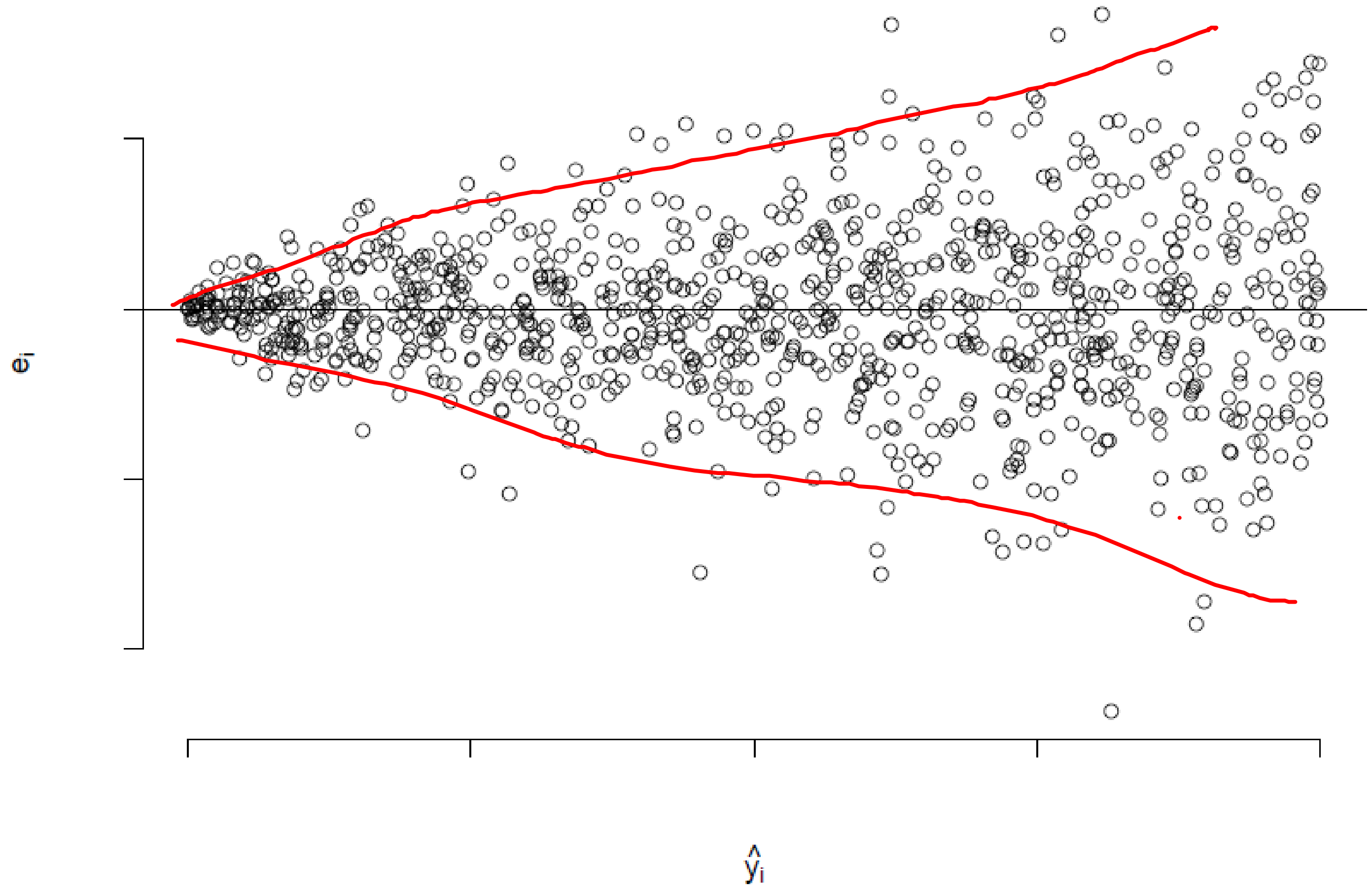
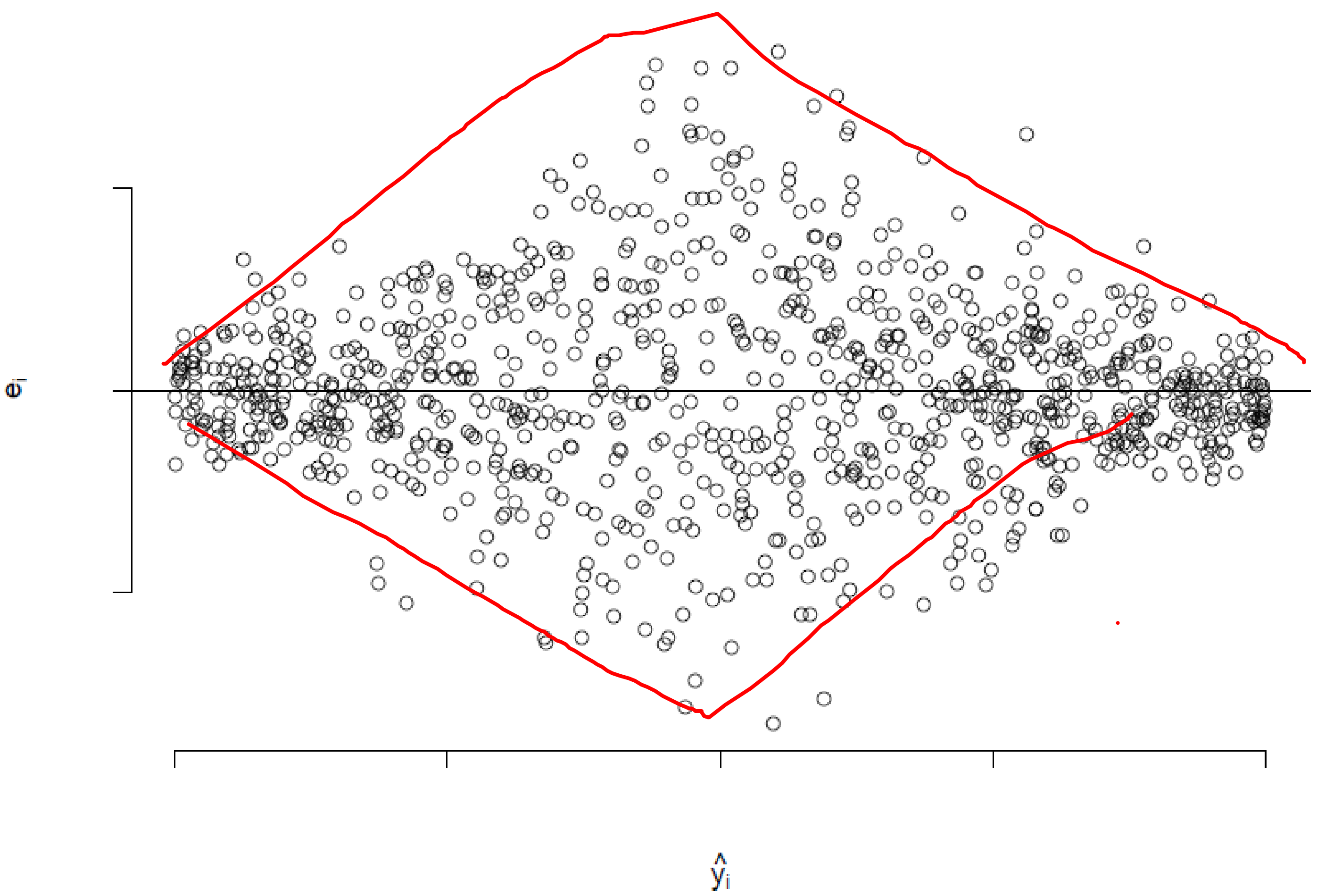


Figure 7: Non Constant Variance



Adverse effect of Heteroscedasticity

- The variance of error terms is used in computing *t*-tests of β coefficients. If this variance is not constant, then ***t*-tests are not healthy** (not efficient, i.e.: the probability of type 2 error is higher)

- However, the coefficients are unbiased. Therefore heteroscedasticity is **not a '*fatal illness*'**

- Check by **White test** or similar tests.

Solution

- Use heteroscedasticity-adjusted *t*-statistics and *p*-values

- Use Data Transformation

Multi Collinearity

- Strong relationship among explanatory variables.

- *Example:*

$$X_3 = 2X_1 + 5X_2$$

Multi Collinearity

- WHR= **waist** circumference / **hip** circumference.
- BMI=weight/height^2
- DV:SBP   IV: WHR
- DV:SBP   IV: BMI

TARGET ORTHO

Adverse effects of Multicollinearity

- Variances of regression coefficients are inflated

- Regression coefficients may be different from their true values**, even signs**

- **Adding or removing** variables produces large changes in **coefficients**. (inconsistency)

- In some cases, the *F* ratio may be **significant, R^2** may be **very high** despite the **all *t* ratios are insignificant** (suggesting no significant relationship)

Multi collinearity: Detection

- The analysis exhibits the signs of multicollinearity — such as, estimates of the coefficients vary from model to model

- The *t*-tests for each of the individual slopes are non-significant ($P > 0.05$), but the overall *F*-test for testing all of the slopes are simultaneously 0 is significant ($P < 0.05$)

- The correlations among pairs of predictor variables are large

- Looking at correlations only among *pairs* of predictors, however, is limiting. It is possible that the pairwise correlations are small, and yet a linear dependence exists among three or even more variables ($X_3 = 2X_1 + 5X_2$)

Variance Inflation Factor (VIF)

variances — of the estimated coefficients are inflated when multi collinearity exists. So, the variance inflation factor for the estimated coefficient b_k —denoted VIF_k —is just the factor by which the variance is inflated

Interpretation

- VIFs exceeding 5 warrant further investigation
- VIFs exceeding 10 are signs of serious multi collinearity

Tolerance Factor

$$\mathrm{VIF} = \frac{1}{\mathrm{tolerance}}$$

Interpretation

- Tolerance of less than 0.20 warrant further investigation
- Tolerance of less than 0.10 are signs of serious multi collinearity

Multi collinearity : What to do?

- Drop a collinear variable from the regression
- Combine collinear variables (e.g. use their sum as one variable)

Independence of Errors

Autocorrelation

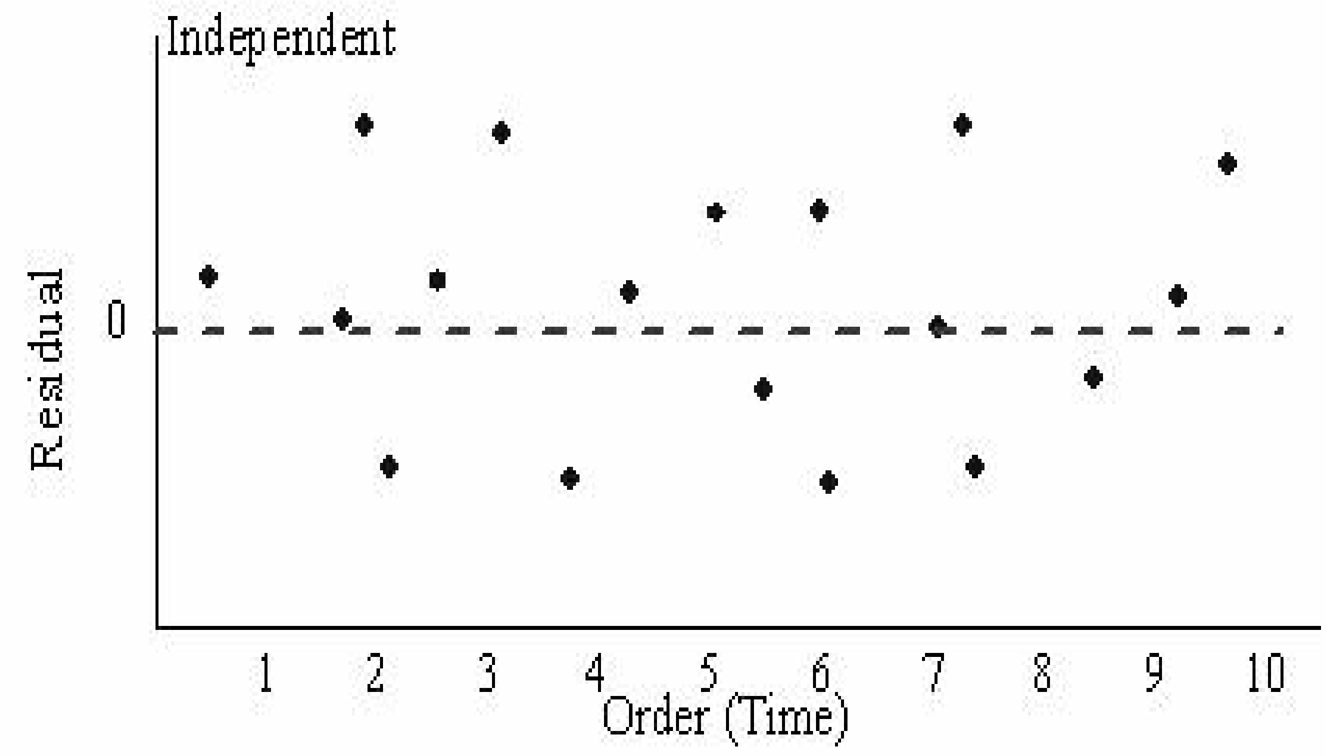


Figure 8: Independence of Error

Independence of Errors

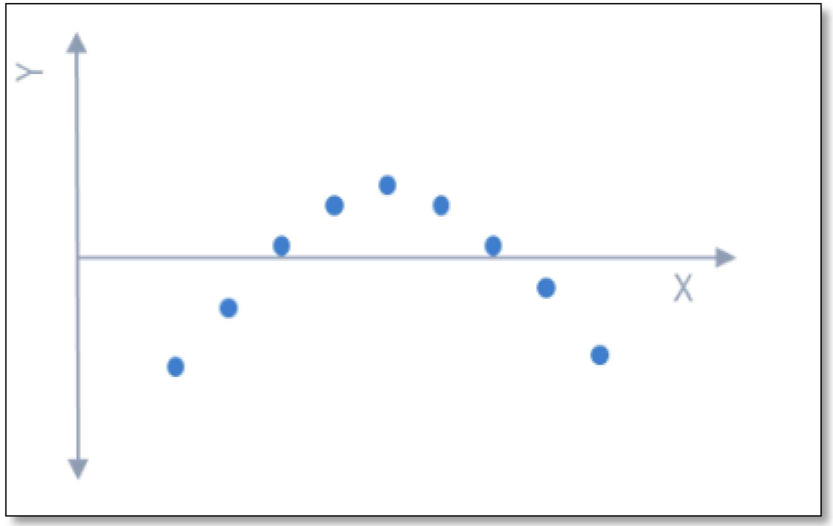


Figure 9: Presence of Autocorrelation

Independence of Errors: Detection

- Residual plots

  (Figure number 8 & 9)

- Durbin-Watso

$$\frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Where e_t=residual at the time period t

Independence of Errors: Detection

H_0 = No first order autocorrelation.

H_1 = first order correlation exists.

- DW Statistic = 2 → No Autocorrelation

- A **rule of thumb** is that test statistic values in the range of 1.5 to 2.5 are relatively normal

Summary Regression

- Relationship between two variables is **functional dependence** of one on the other
- **Magnitude of one variable** (DV) is assumed to be determined by **a function** of the **magnitude of the second variable**(IV)
- The **reverse may not true** (Ex.: DV: Blood Pressure; IV: Age)
- The term **dependent ≠ Cause & Effect** Relationship

References

- John O. Rawlings, Sastry G. Pantula, David A. Dickey. Applied Regression Analysis: A Research Tool. Second Edition. Springer

- Daryl S. Paulson. Handbook of Regression and Modeling Applications for the Clinical and Pharmaceutical Industries. Chapman & Hall/ CRC Biostatistics Series

- a visual guide to CRISP-DM methodology (http://www.crisp-dm.org/download.htm)

- https;//www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis assumptions-plots-solutions/

- http://www.statisticshowto.com/durbin-watson-test-coefficient/

- David M. Levine, David F. Stephan, Kathryn A. Szabat. STATISTICS FOR MANAGERS USING MICROSOFT EXCEL, 8th Edition. Pearson Publication.

Lets Connect!

draanchalawasthi@gmail.com

https://www.youtube.com/c/sscrindia

+91 750.625.0403

TARGET ORTHO