Multiple Linear Regression

Normality and

Homoscedasticity Assumptions

Contents

1. The Assumptions of Normality and Homoscedasticity
2. Residual v/s Predicted Plot in R
3. Q-Q plot of residuals
4. Shapiro Wilk test to assess Normality of Errors
5. Absence of Normality – Remedial Measure
6. Box cox Transformation in R

Normality and Homoscedasticity

- The errors in Multiple Linear Regression are assumed to follow Normal Distribution.

- If Normality of Errors is not true then statistical tests and associated P values based on F and t distribution are not reliable.

- **Homoscedasticity** describes a situation in which variance of error term is same across all values of the independent variables.

- In the absence of Homoscedasticity (Or presence of Heteroscedasticity) the standard errors of parameter estimates are incorrect.

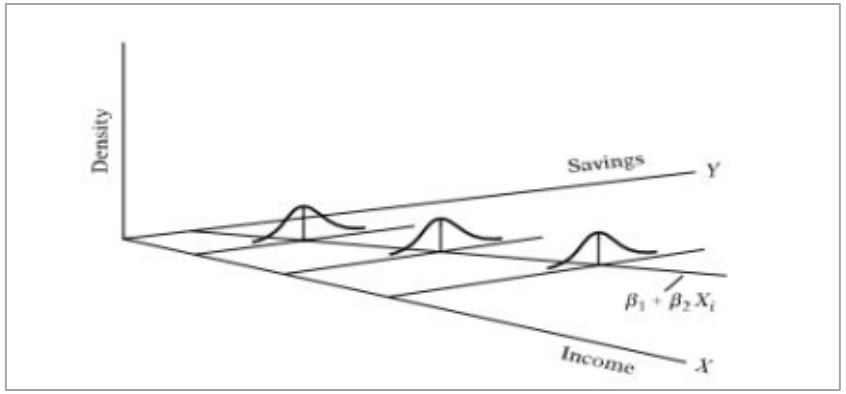
Assumption of Homoscedasticity

- Variance of error term must be constant across the independent variables (defined by X values)

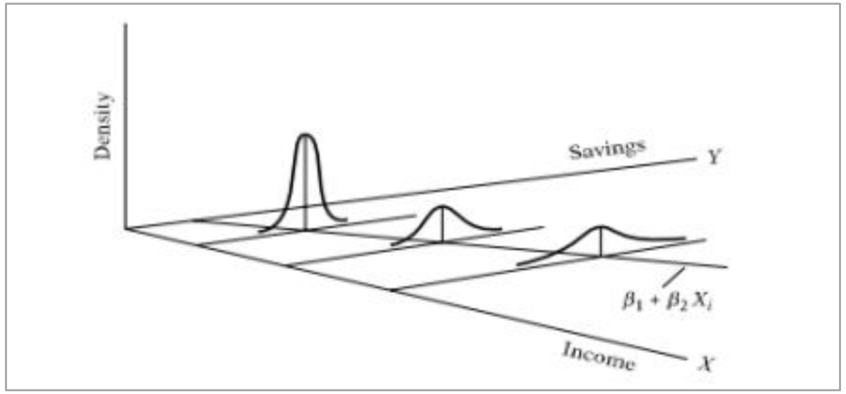
$$V\left({}^{e_i}/_{x_i}\right) = \sigma^2 \ \textit{indicates homoscedasticity}$$

$$V\left({}^{e_i}/_{x_i}\right) = \sigma_i^2 \ \textit{indicates heteroscedasticity}$$

Homoscedastic Errors

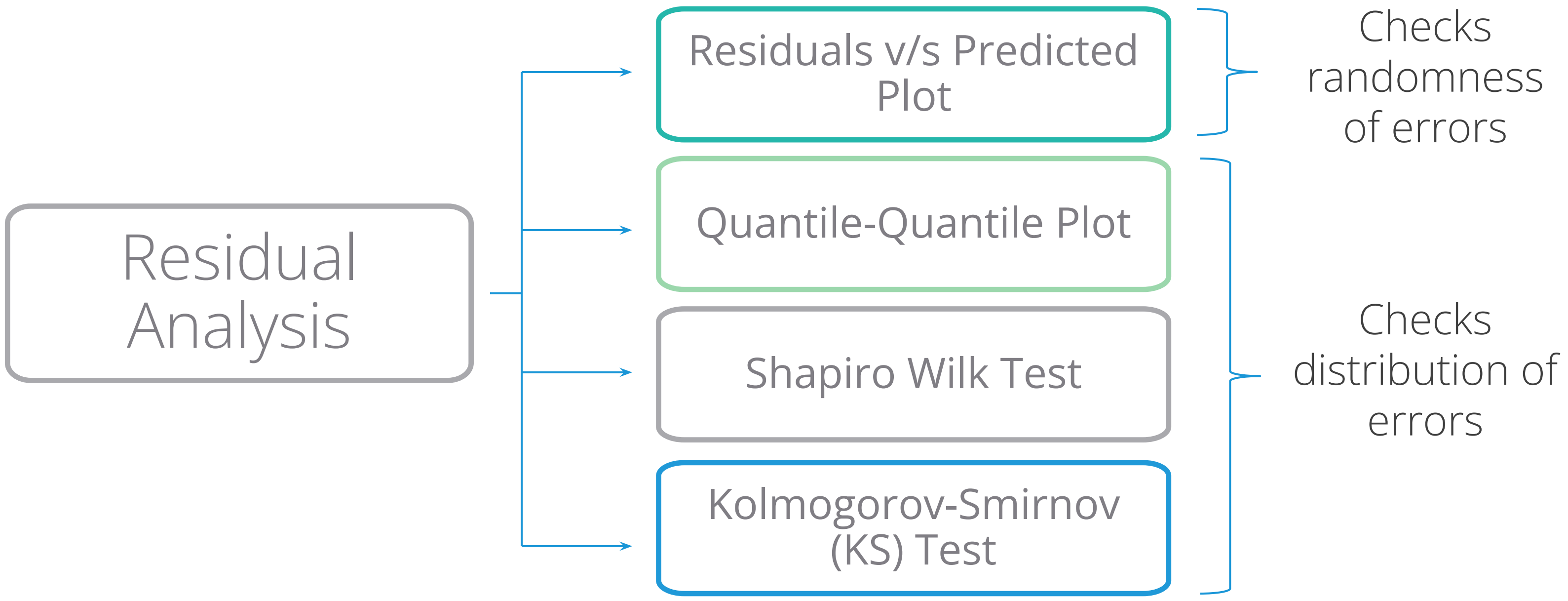


Heteroscedastic Errors



Residual Analysis

Observed Value – Predicted value = Residual



Residual Analysis
Residuals v/s Predicted Plot
Quantile-Quantile Plot
Shapiro Wilk Test
Kolmogorov-Smirnov (KS) Test
Checks randomness of errors
Checks distribution of errors

Residual Analysis for Performance Index Data

Continuing with the “Performance Index “ data,

- Model job performance index (jpi) based on aptitude score (aptitude), test of language (tol), technical knowledge (technical) and general information (general)
- Get the fitted values and thus the residuals.
- Analyse the distribution of residuals

Residual v/s Predicted Plot in R

```
#Importing the Data, Fitting Linear Model and Calculate Fitted Values
and Residuals
perindex<-read.csv("Performance Index.csv",header=TRUE)
jpimodel<-lm(jpi~aptitude+tol+technical+general, data=perindex)
perindex$pred<-fitted(jpimodel)
perindex$resi<-residuals(jpimodel)
```

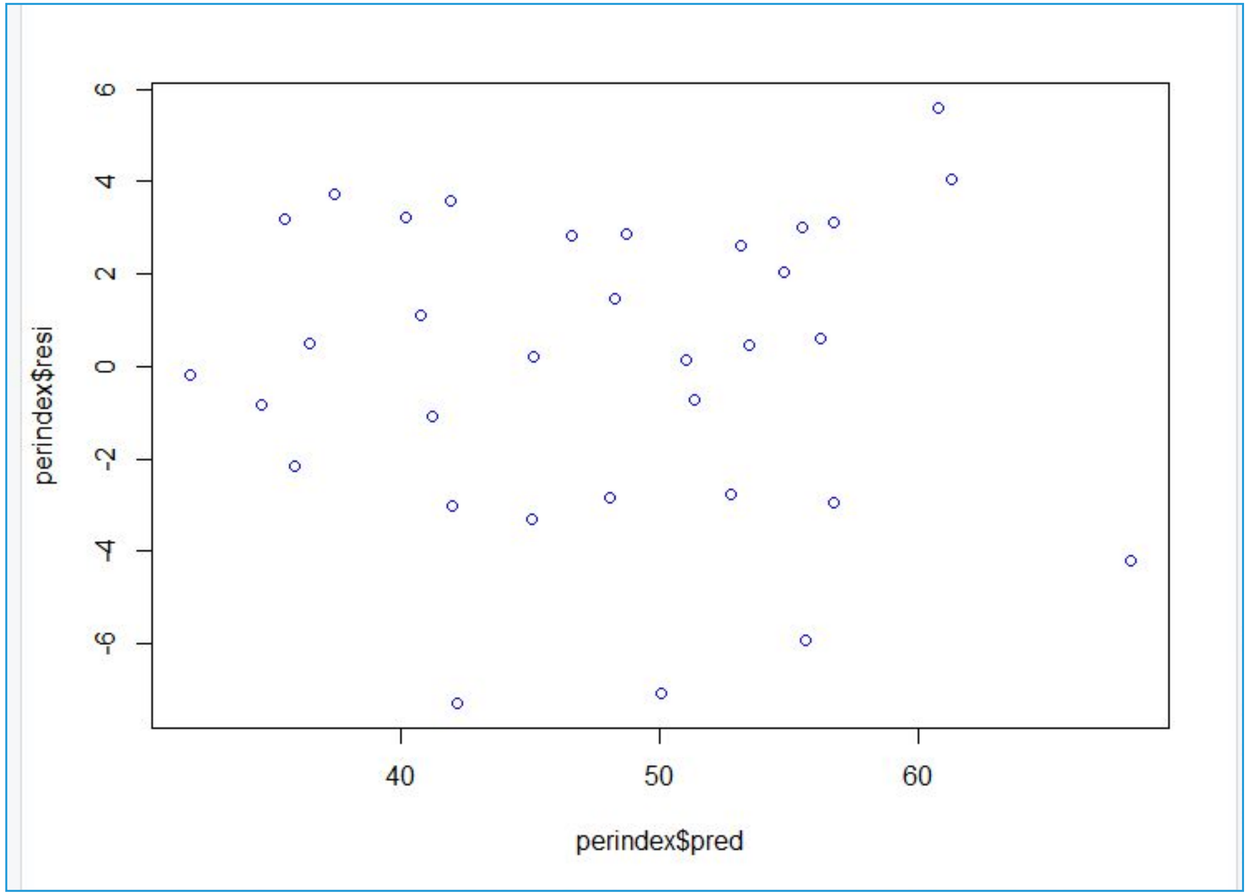
- *lm() fits a linear regression.*
- *fitted() and residuals() fetch fitted values and residuals respectively.*

```
#Residuals v/s Predicted Plot
plot(perindex$pred,perindex$resi,col="blue")
```

plot() is used to plot predicted values against residuals.

Residual v/s Predicted Plot in R

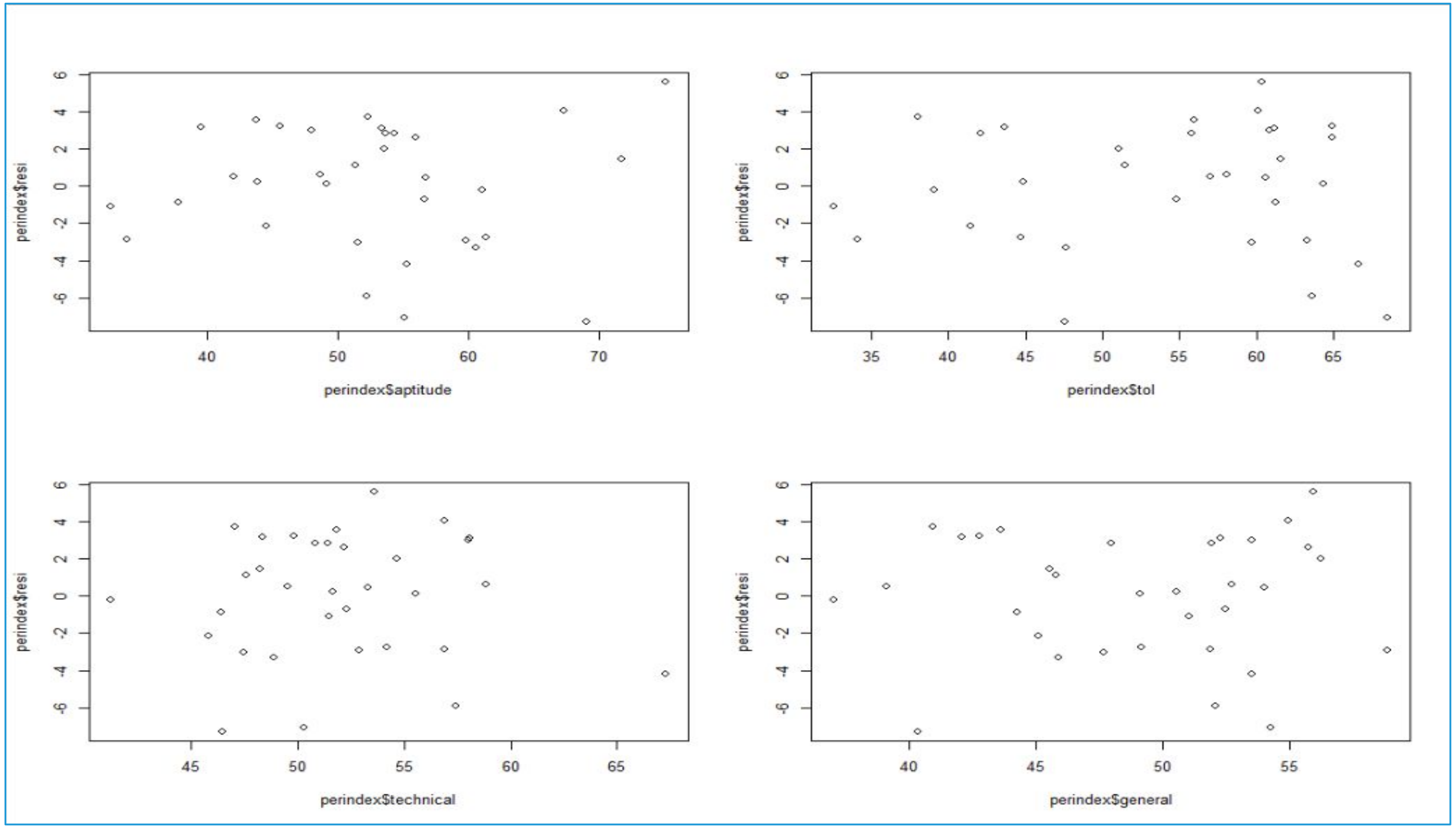
```
# Output
```



Interpretation:

- *Residuals in our model are randomly distributed which indicates presence of Homoscedasticity*

Residual v/s Independent variables Plot in R

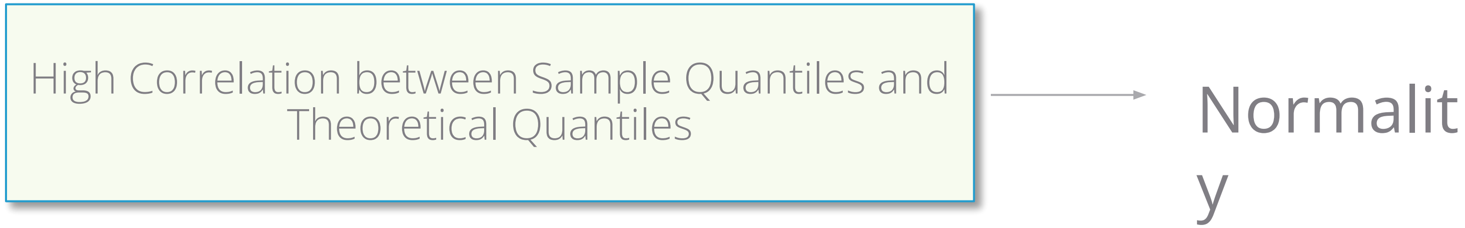


Interpretation:

- *Residuals in our model are randomly distributed which indicates presence of Homoscedasticity*

QQ Plot in R

- **The Quantile-Quantile (QQ) Plot** is a powerful graphical tool for assessing normality.
- Quantiles are calculated using sample data and plotted against expected quantiles under Normal distribution.



High Correlation between Sample Quantiles and Theoretical Quantiles

Normality

- If the data are truly sampled from a Gaussian (Normal) distribution, the **QQ plot will be linear.**

QQ Plot in R

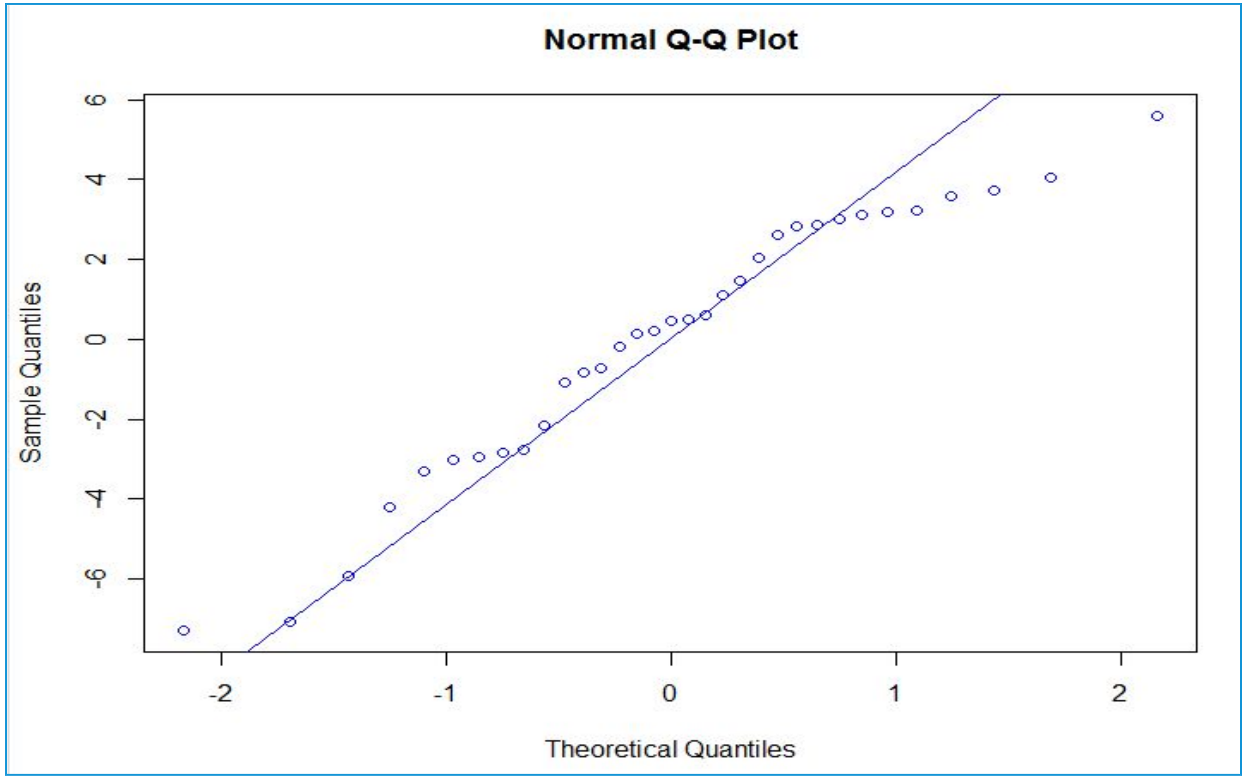
```
#QQ Plot
qqnorm(perindex$resi,col="blue")

qqline(perindex$resi,col="blue")
```

- *qqnorm() produces a plot with theoretical quantiles on x axis against the sample quantiles on y axis.*
- *Column for which normality is being tested is specified in the first argument.*
- *qqline() adds a line which passes through the first and third quartiles.*

QQ Plot in R

```
# Output
```



Interpretation:

- *Most of these points are close to the line except few values indicating no serious deviation from Normality.*

Shapiro Wilk Test

Objective	To **correlate**, sample ordered values with expected Normal scores in order **to test normality of the sample**

Null Hypothesis (H_0): Sample is drawn from Normal Population

Alternate Hypothesis (H_1): Not H_0

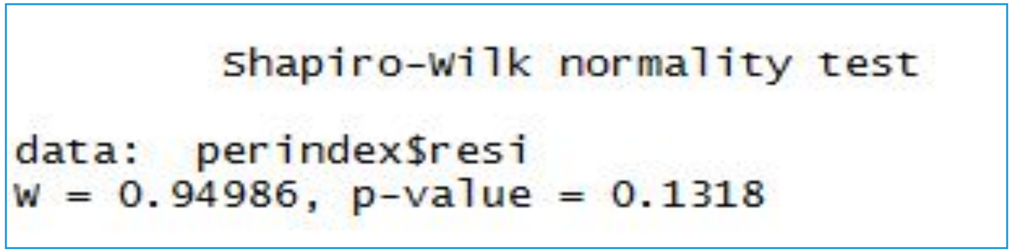
Test Statistic	
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Shapiro Wilk Test in R

```
# Shapiro Wilk Test
shapiro.test(perindex$resi)
```

shapiro.test() from basic stats package, returns correlation coefficient w and p-value.

```
# Output
```



```
        Shapiro-Wilk normality test

data:  perindex$resi
W = 0.94986, p-value = 0.1318
```

Interpretation:

- *p-value>0.05, Do not reject H_0. Normality can be assumed.*

Absence of Normality – Remedial Measure

Mathematical Transformation of the dependent variable is used as a remedial measure in case of serious departure from Normality.

Typically Log Transformation is used. However, there is general transformation called as Box Cox Transformation given as :

- Box Cox transformation

$$
\begin{aligned}
Y^* &= \frac{Y^\lambda - 1}{\lambda} \qquad \lambda \neq 0 \\
&= \log Y \qquad\quad \lambda = 0
\end{aligned}
$$

Where Y is the response variable

- R can automatically detect the optimum λ using **`boxcox()`** in package MASS

Quick Recap

This session explained in detail **normality of errors**. Here’s a quick recap:

Normality Assumption	• Error terms should be normally distributed
Homoscedasticity	• Errors should have constant variance across X values
Residual v/s Predicted Plot	• Ideally should be randomly distributed
QQ Plot	• Used to check if errors follow Normal distribution
Shapiro Wilk Test	• Test for Normality assessment of errors
Box Cox Transformation	• Transforming non normal response to normal