Lecture 14:

Multiple Regression (2): Analysis of Residuals and Diagnostic Checks

Two aspects that we consider when judging the suitability of a statistical model to describe real-life variables are (i) How good the model fits the data (ii) How closely the estimated model satisfies the underlying assumptions. There are certain assumptions underlying the multiple regression for valid inference from the model.
(i): The multiple linear regression equation is a suitable description of the relationship between the response and the predictors i.e. the model's mean function is correct.
(ii): The variance of response over all possible predictor variables is constant.
(iii): The distribution of response is normal and
(iv): The observations are independent.

Analysis of residuals [e = $y - \hat{y}$] of the estimated regression helps in assessing these assumptions.
Assessment of the independence assumption is very difficult in most cases and generally requires that we have additional information about the data, such as the time sequence in which the data were collected. We will not consider methods for checking the independence assumption.

Residual plots can be used as diagnostic tools for assessing the validity of the regression model assumptions. In order to check for the suitability of the regression equation, we plot the residuals versus the predicted values of the response variable (e versus $\hat{y}$), and the residuals versus each of the predictor variables (e versus x_i, for i 1, 2, . . . , k). If the regression equation is suitable, each of these plots should show the residuals roughly centred and symmetric about the horizontal axis. Deviations from this expected pattern indicate that the multiple linear regression equation may not be suitable.

The following Figure 1 (left) plots the scatter diagram along with the fitted linear regression. The right panel plots the residual e against the fitted value $\hat{y}$. There appears to be no systematic pattern and the model appears to be a satisfactory description of the relations between y and x variable. Note: we can also plot residual against each x variable.

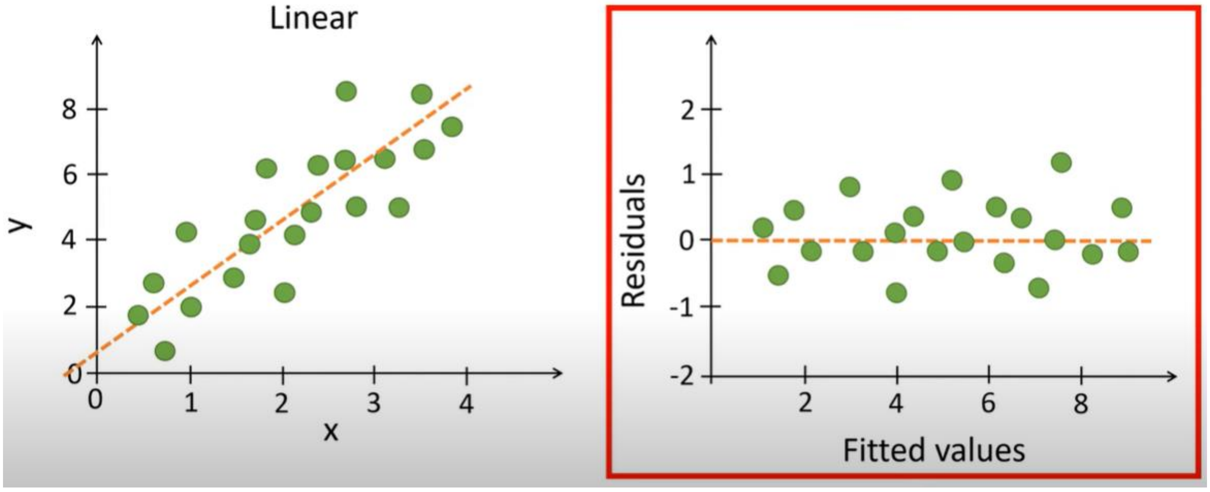


Fig 1:

The Fig 2 (left) below indicates a nonlinear pattern of data but a wrong model (linear) was fitted and the resulting residual plot (right) clearly indicates the non-linear pattern indicating the fitted model is not suitable.

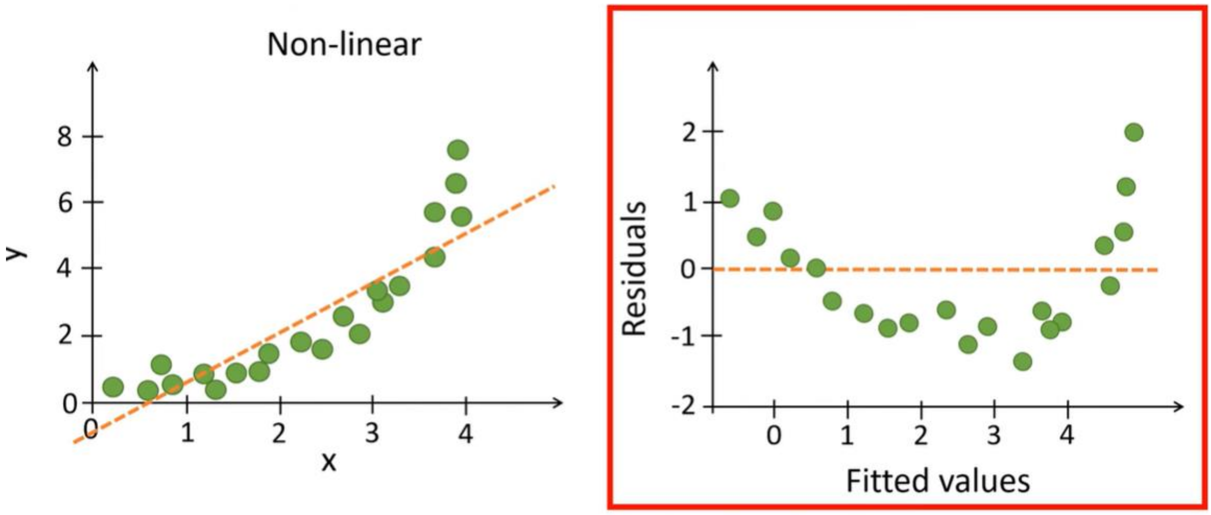
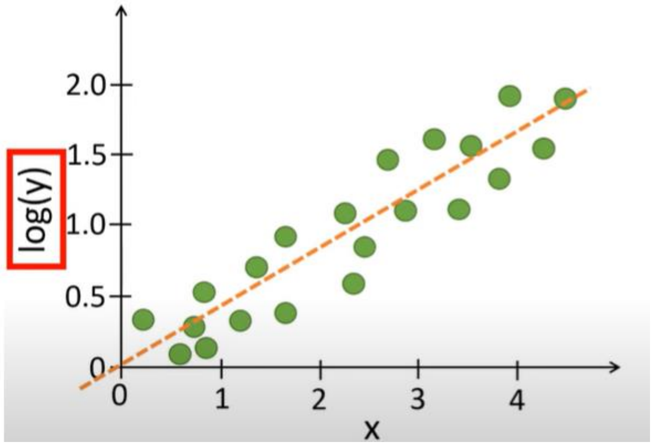


Fig 2:

Sometimes estimating the model on transformed data e.g. log of y instead of corrects the assumption failure as indicated by the following Fig 3.



The constant standard deviation assumption can be checked using the same residual plots as those used to check for the suitability of the regression equation. If the standard deviation is constant across all predictors variable values, the plot of residuals versus predicted values and the plots of residuals versus predictor variables are expected to exhibit roughly constant variation as the predicted values or the predictor variable values change. If the variation in the residuals changes as the predicted values change or as the values of one (or more) of the predictor variables change, the assumption of constant standard deviation comes into question.

The following Fig 4 indicates that the variance of the residual appears to be constant. No issue with the fitted model.

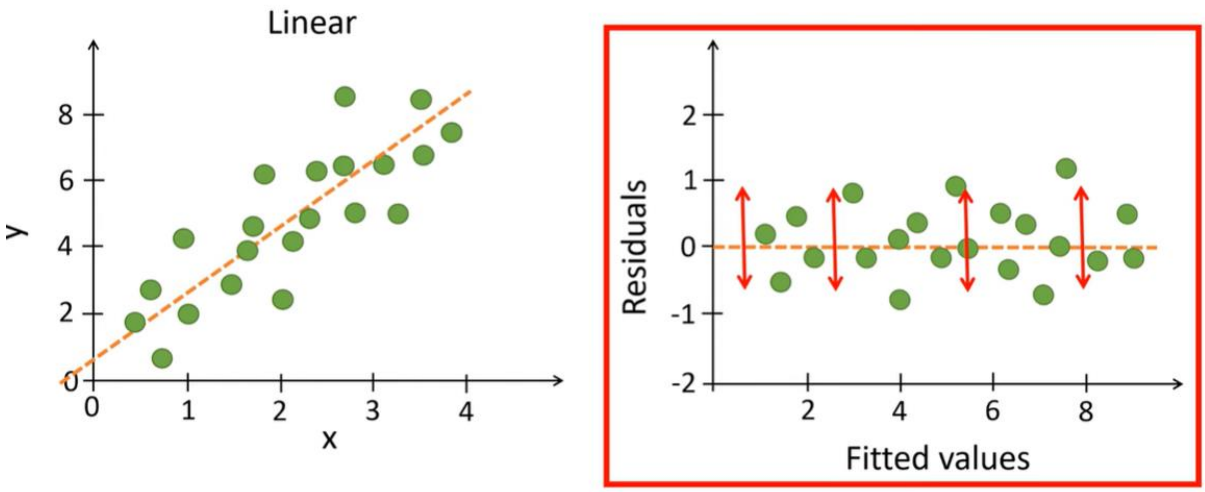


Fig 4:

On the other hand, the following Figure 5 (left) indicates that variance of error appears to increase with increase in x values so is not constant. The corresponding residual plot (right) clearly indicates the increasing variance hence failure of assumption # 2.

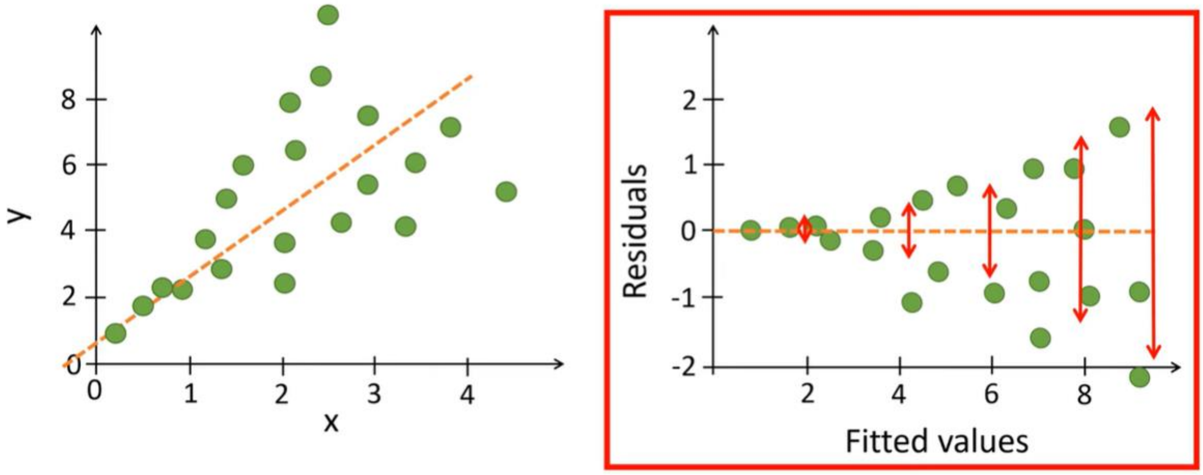


Fig 5:

Figure 6 indicates a similar issue with decreasing variance indicating failure of constant variance assumption.

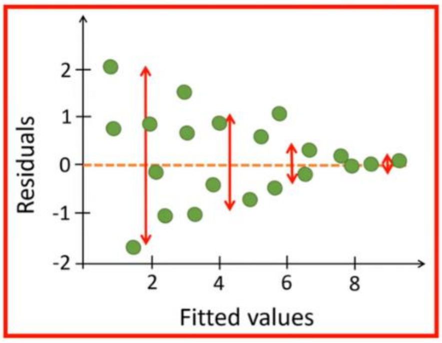


Fig 6:

An unequal variance will result in p-values associated with the estimated parameters in the regression model that are smaller compared to if there is equal variance. This will result in increased risk of type I error i.e., showing that variable in the regression is significant while it is in fact not significant.

A normal probability plot of the residuals may be used to assess the normality assumption. Such a plot should be roughly linear.

Departures from linearity indicate possible nonnormal populations. The following figure 7 indicates the properly fitted linear model and the right panel shows that the normal quantile-quantile (qq plot) appears to be roughly linear indicating no issue with the normality assumption.

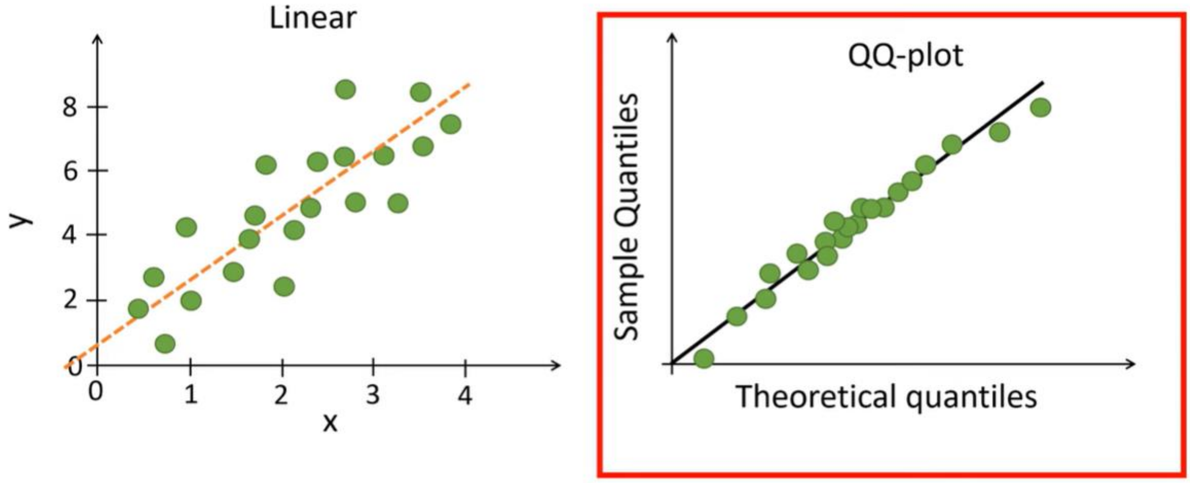


Fig 7:

The departure from normality can manifest itself in several forms. Figure 8 indicates a pattern in the scatter plot where positive values are more scattered (left) resulting in positively skewed residual distribution (middle) and showing up as non-normal, (in particular, a concave up or u shaped qq plot).

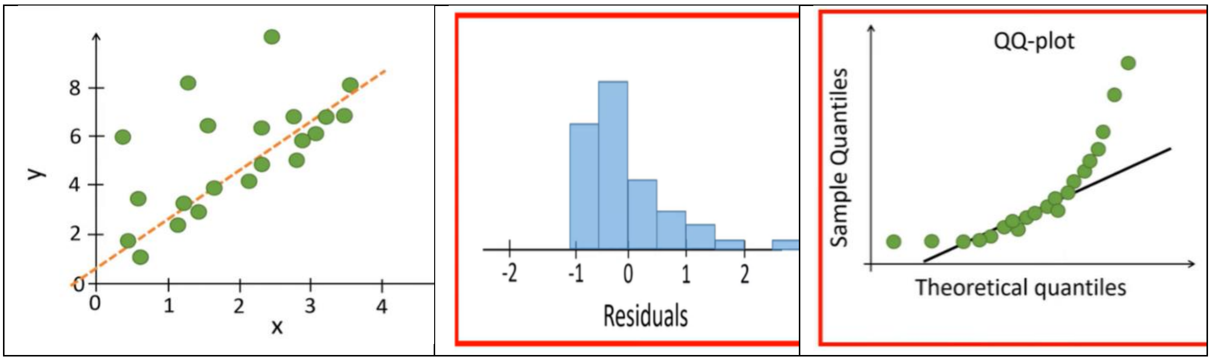


Fig 8:

A concave down (reverse u shaped) qq plot indicates a negative skewness.

Non normality may be due to omitting important variables or due to presence of outliers as indicated in the following figure 9.

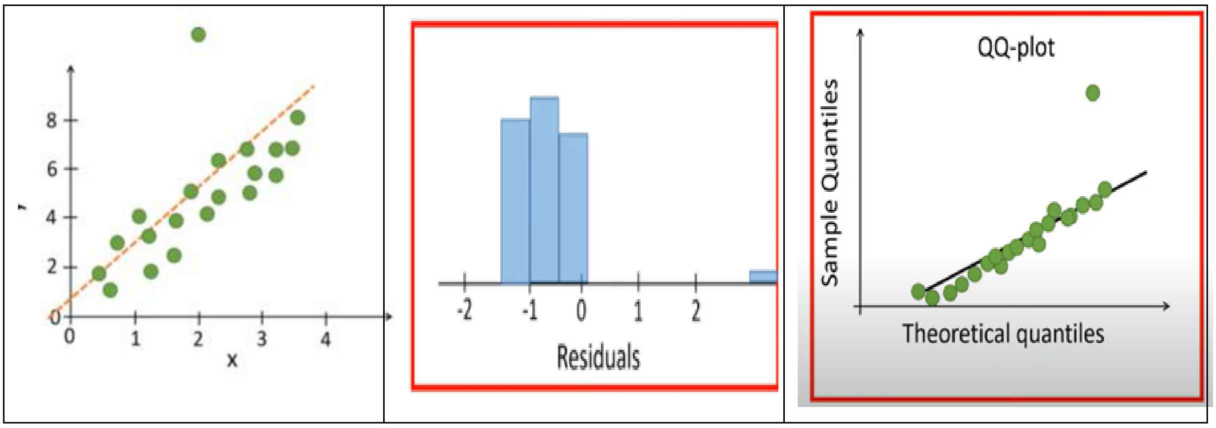
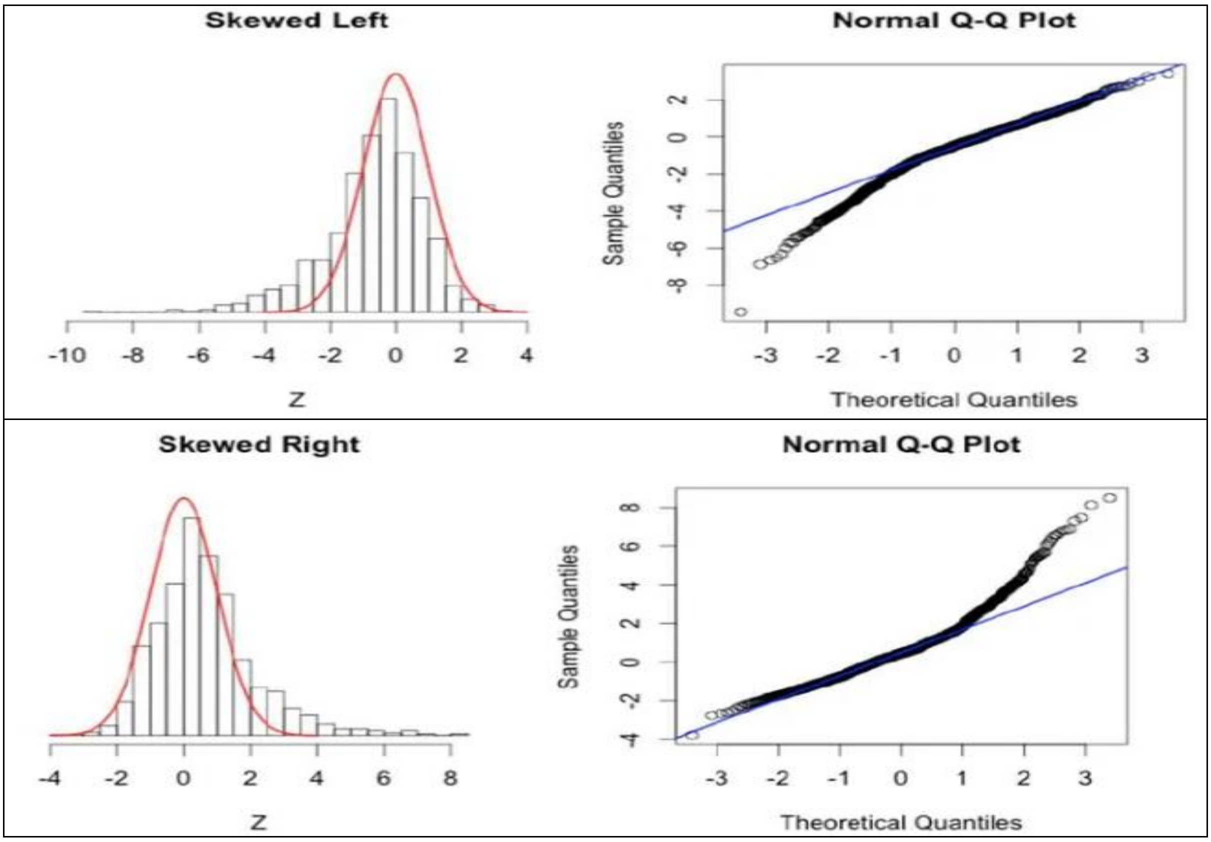


Fig 9:

Normality of the variables can also be tested using formal tests e.g. the Shapiro-Wilk test which has null hypothesis of normality. The following (Fig 10) are further examples of how deviations from normality are reflected in the qq plot.



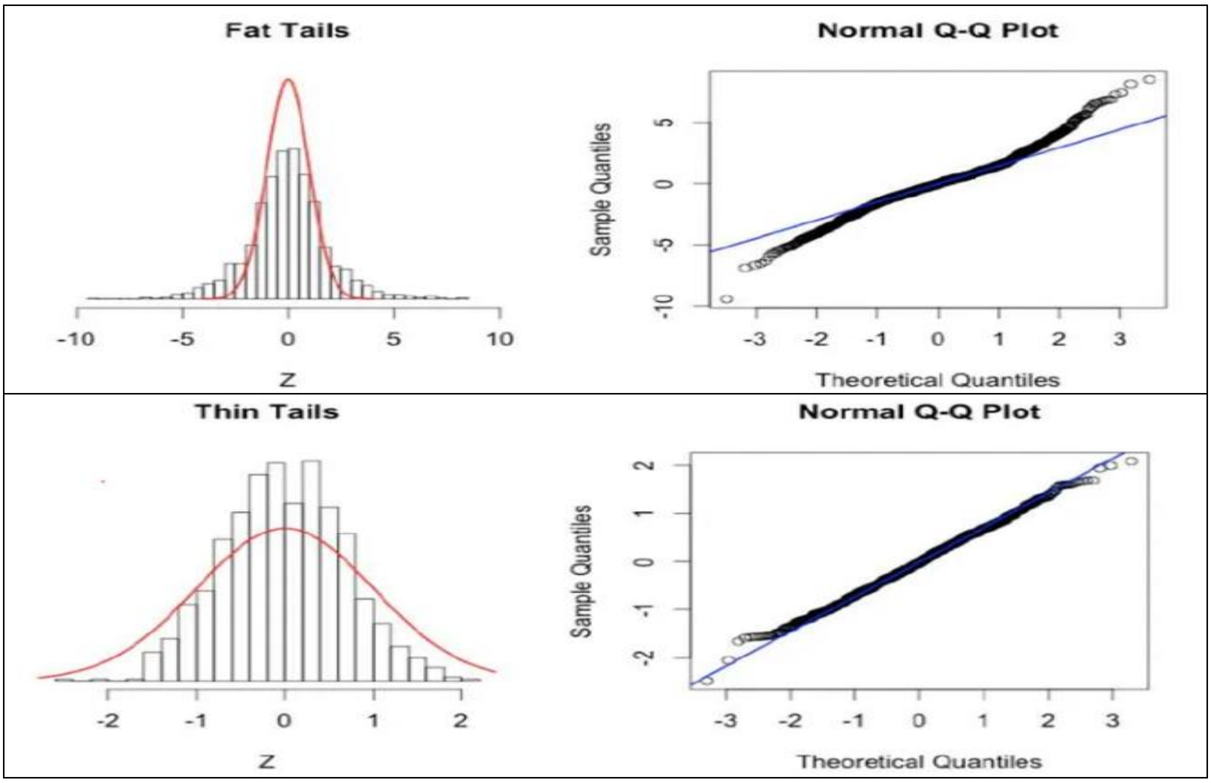


Fig 10:

To check for possible outlier data points, we look for residuals that are large in magnitude. The determination of whether a residual is large is made by obtaining the number of standard deviations a residual is from its mean of 0.

The standard deviation of each residual is approximated by the standard error of the estimate (also known as standard error of regression) and is given by:

$$se = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-(k+1)}}$$

As a rough rule, we consider any data point with a residual whose absolute value is larger than $2s_e$ as a potential outlier that should be investigated further. An easier way is to calculate the standardized residual defined as

$$Standardized\ residual = \frac{residual}{se}$$

An observation with a standardized residual greater than 2 in absolute value indicates a potential outlier.

Just as an outlier indicates an extreme observation in the vertical direction, the leverage of the observation measures extreme cases in the horizontal direction. An observation with a high leverage can potentially disturb the regression slope estimate. An observation with a high outlier and a large leverage can be potentially an influential observation in the sense that removal of this observation can bring big changes in the regression coefficients.

A measure of potentially influential observation is the Cook's distance given by:

$$Cook's\ Distance_i = \frac{\sum_{j=1}^{n}[\hat{y}_j - \hat{y}_j(i)]^2}{(k+1)\ MSE}$$

Where $\hat{y}_j$ indicated fitted value and $\hat{y}_j(i)$ indicates fitted value when *ith* observation deleted from the data set. MSE is square of the standard error of regression.

Normally a Cook's distance greater than 0.5 or 1 indicates a potential influential observation. Sometimes the following cut off is also used.

$$cutt\ off = \frac{4}{n-k-1}$$

The R software's plot(model) command provides four diagnostic plots which are useful in assessing assumption of the model. Here model is the name of the regression model object estimated using the 'lm' command. A typical model plot is as follows (Fig 11): These four plots in one graph window can be obtained as:

```
par(mfrow = c(2,2))
plot(model)
```

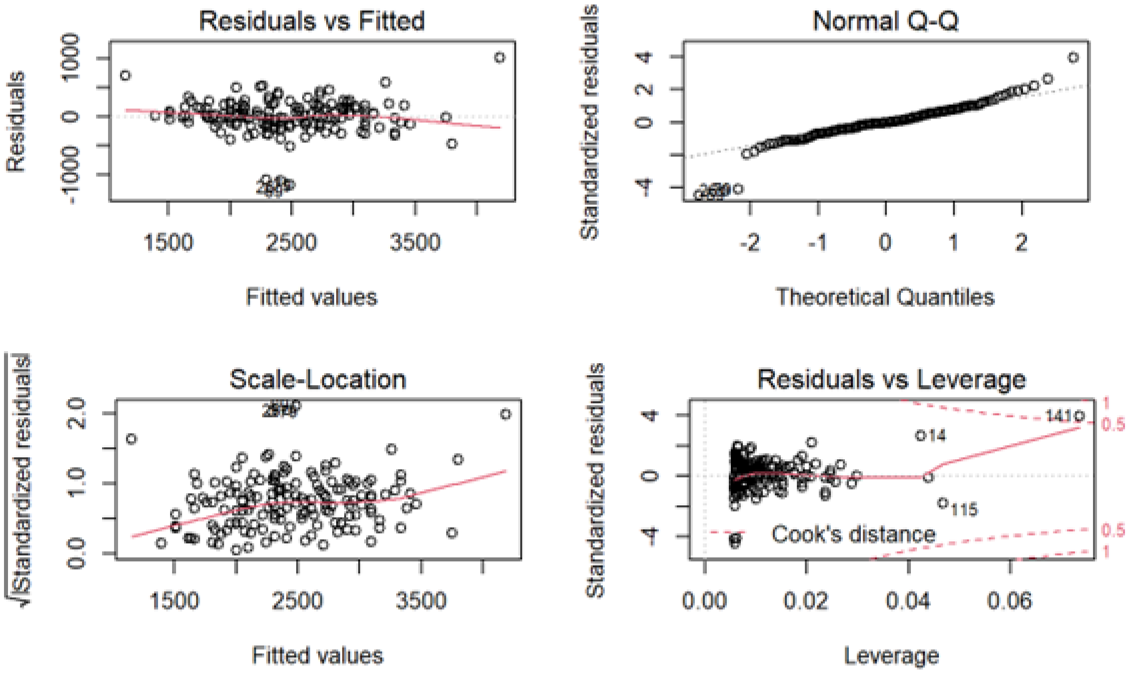


Fig 11:

The two plots in the first column are used to assess linearity (correct specification of mean function), constant variance, and any outliers etc. The top right graph is the qq plot of normality and the bottom right is the leverage plot with an indication of high cook's distance indicated by dashed lines.

Example: Consider the data of 176 real estate property sales prices in a city as related to the value of land, value of improvement (building structure etc.) and the city district where the property is located. There are 4 districts. The three quantitative variables are measured in thousands of dollars. The number of properties in the four area are as follows:

```
      CHEVAL   DAVISISLES HUNTERSGREEN     HYDEPARK
          44           42           56           34
```

The scatter plot matrix of the variables is given as follows (Fig 12). The plot indicates quite a strong positive relationship between sales price and value of land as well as of sales price with value of improvement. From the scatter plot of sales vs land it is evident that in their bivariate relationship the constant error variance assumption may not be met.

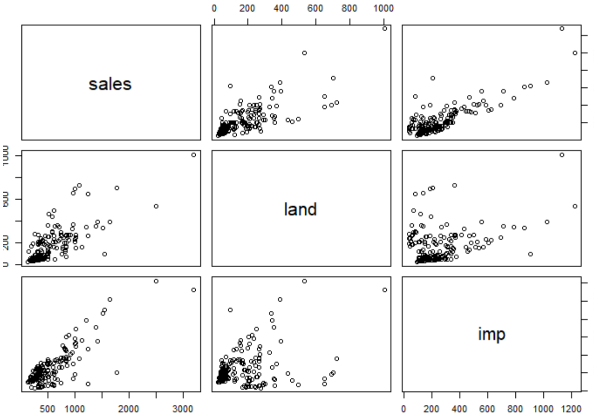


Fig 12:

The following plots (Fig 13) also show the marginal distribution (histogram) of individual variables as well as the bivariate relationship. All the three variables appear to be highly positively skewed.

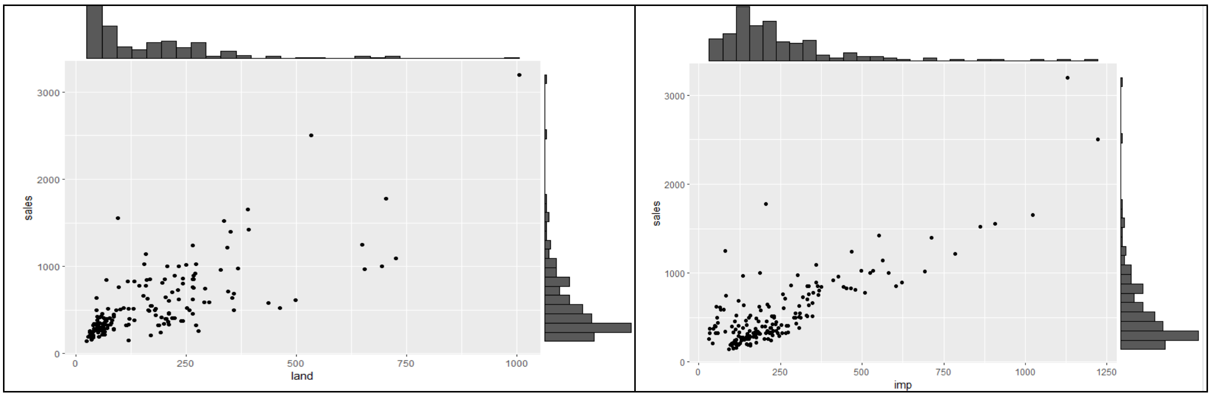
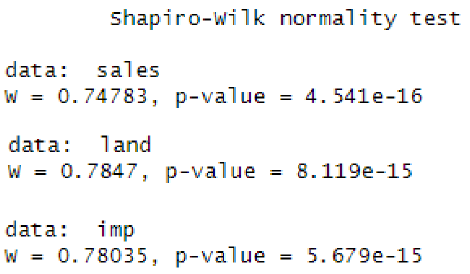


Fig 13:



```
        Shapiro-Wilk normality test

data:  sales
W = 0.74783, p-value = 4.541e-16

data:  land
W = 0.7847, p-value = 8.119e-15

data:  imp
W = 0.78035, p-value = 5.679e-15
```

The Shapiro Wilk W test (a test with normality under the null hypothesis) has small p-values for all the three variables. This substantiates the visual impression of non-normality of the three variables.

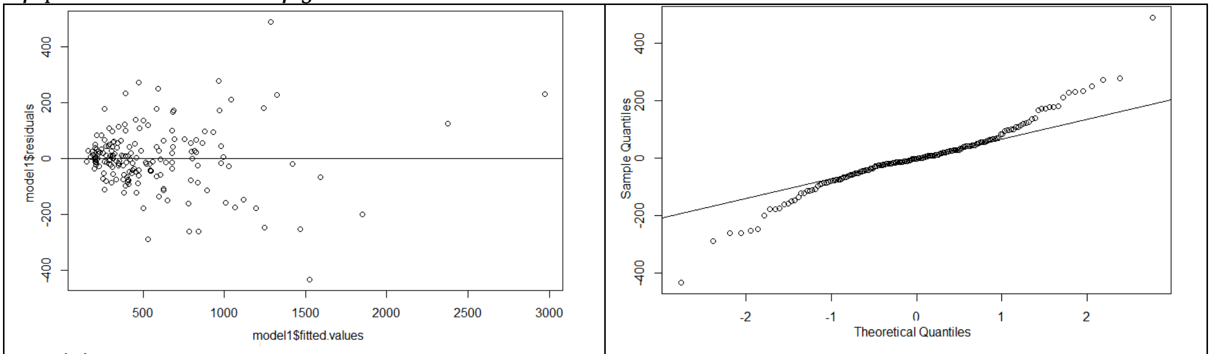
We may consider some potential models to describe the relationship between the sales price and its determinants. These models are respectively lin-lin, log-lin, lin-log and log-log in sales and value of land and improvement. The three dummy variables of three locations are also included.

Model1:$Sales = \beta_0 + \beta_1 Land + \beta_2 Imp + \beta_3$ DAVISISLES $+\beta_4$ HUNTERSGREEN + $+\beta_5$ HYDEPARK+ e

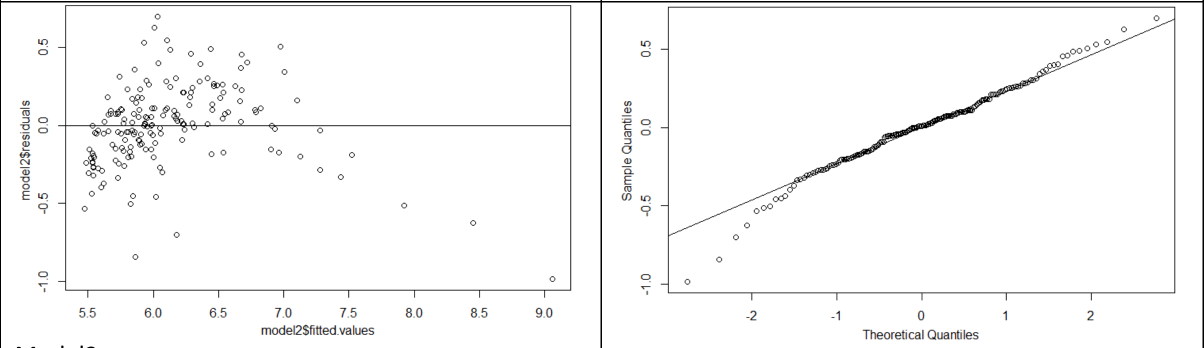
Model2: $\log(Sales) = \beta_0 + \beta_1 Land + \beta_2 Imp + \beta_3$ DAVISISLES $+\beta_4$ HUNTERSGREEN + $+\beta_5$ HYDEPARK $+e$

Model3: Sales $= \beta_0 + \beta_1 \log(Land) + \beta_2 \log(Imp) + \beta_3$ DAVISISLES $+\beta_4$ HUNTERSGREEN + $+\beta_5$ HYDEPARK $+e$

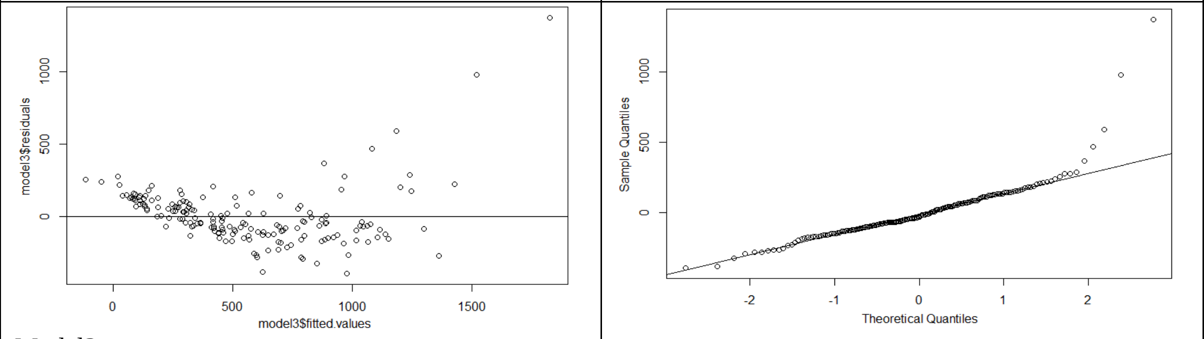
Model4: $log(Sales) = \beta_0 + \beta_1 \log(Land) + \beta_2 \log(Imp) + \beta_3$ DAVISISLES + $+\beta_4$ HUNTERSGREEN $+\beta_5$HYDEPARK $+e$



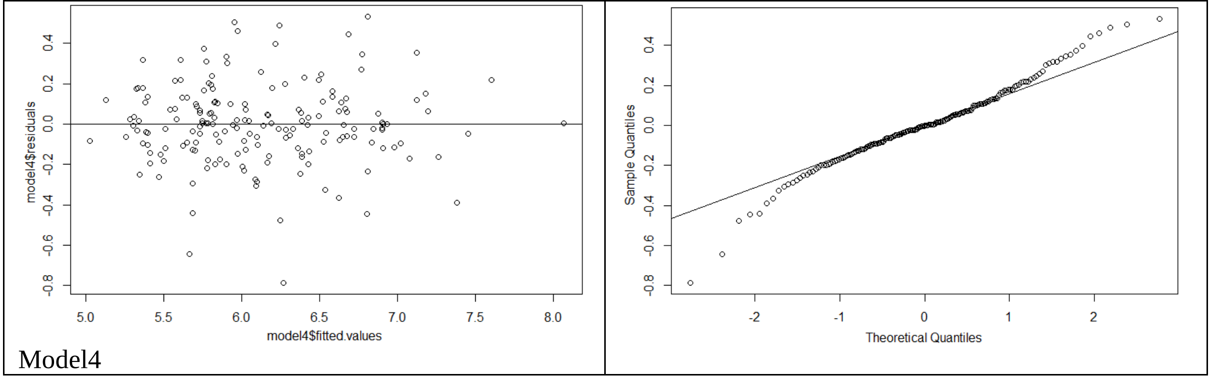
Model1



Model2



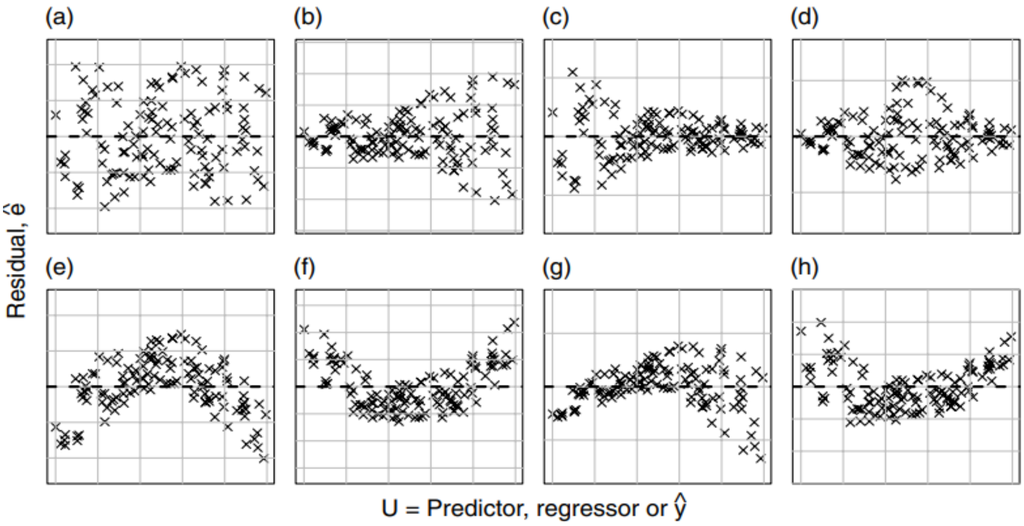
Model3



Model4

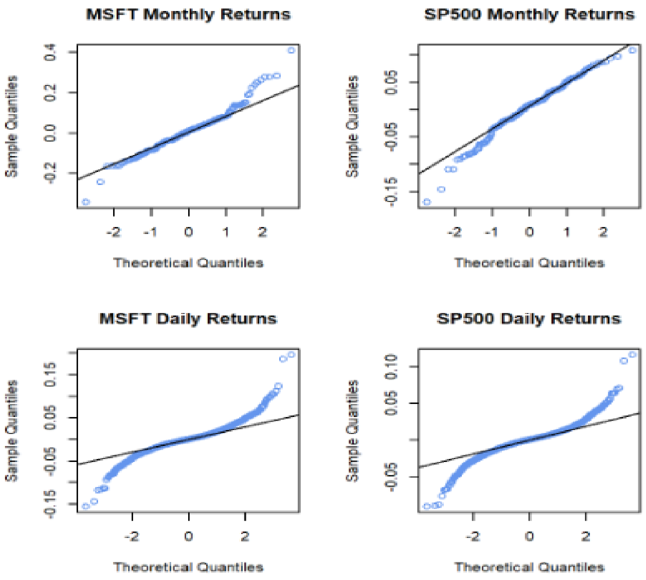
The residual plot for model1 indicates non constant variance. Model 2 and 3 indicates incorrect men function specification as well as problem with the constant variance, The residual plot of model4 appears to be the most satisfactory as it indicates a correctly specified mean function with assumption of constant variance also appears to be met. Some deviation from normality is evident from the tails of the residual distribution. However, it is the case that with a large sample size (176 in our case), non-normality is not considered a serious violation and the inference can be justified asymptotically. Also, no observation in case of model 4 has the cook distance greater than 0.5.

Ex1: Assess any problem with the models given the following residual plots from (a) to (h):



[Sol: a) good b to d non-constant variance, e and f incorrect mean function (non-linearity), g and h (both non constant variance as well as incorrect mean function]

Ex2: The following qq plots are for distribution of an individual stock (Microsoft) and the index returns for both the monthly and daily data. Comment on the normality of the returns.



[Sol: Both the MSFT and SP500 daily returns are heavy tailed and depart from normality. Both monthly returns are more closely normal distributed than daily returns. However, the MSFT monthly returns have some evidence of heavy tails, while slight negative skewness is observed in the SP500 returns.]