Chapter 4
Model Adequacy Checking

The fitting of the linear regression model, estimation of parameters testing of hypothesis properties of the estimator, is based on the following major assumptions:

1. The relationship between the study variable and explanatory variables is linear, at least approximately.
2. The error term has zero mean.
3. The error term has a constant variance.
4. The errors are uncorrelated.
5. The errors are normally distributed.

The validity of these assumptions is needed for the results to be meaningful. If these assumptions are violated, the result can be incorrect and may have serious consequences. If these departures are small, the final result may not be changed significantly. But if the deviations are large, the model obtained may become unstable in the sense that a different sample could lead to an entirely different model with opposite conclusions. So such underlying assumptions have to be verified before attempting to regression modeling. Such information is not available from the summary statistic such as t-statistic, F-statistic or coefficient of determination.

One crucial point to keep in mind is that these assumptions are for the population, and we work only with a sample. So the main issue is to make a decision about the population on the basis of a sample of data.

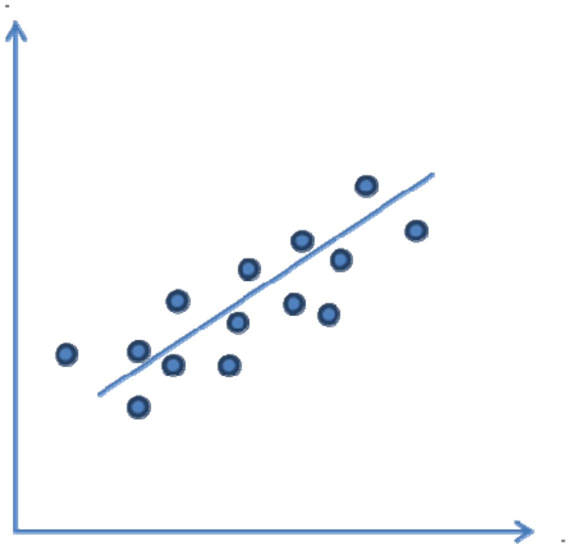
Several diagnostic methods to check the violation of regression assumption are based on the study of model residuals with the help of various types of graphics.

Checking of the linear relationship between study and explanatory variables

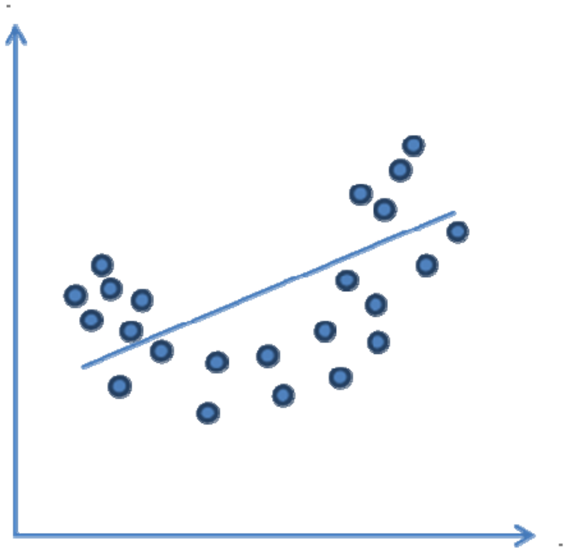
1. Case of one explanatory variable

If there is only one explanatory variable in the model, then it is easy to check the existence of the linear relationship between y and X by scatter diagram of the available data.

If the scatter diagram shows a linear trend, it indicates that the relationship between y and X is linear. If the pattern is not linear, then it suggests that the relationship between y and X is nonlinear. For example, the following figure indicates a linear trend



whereas the following graph suggests a nonlinear trend:



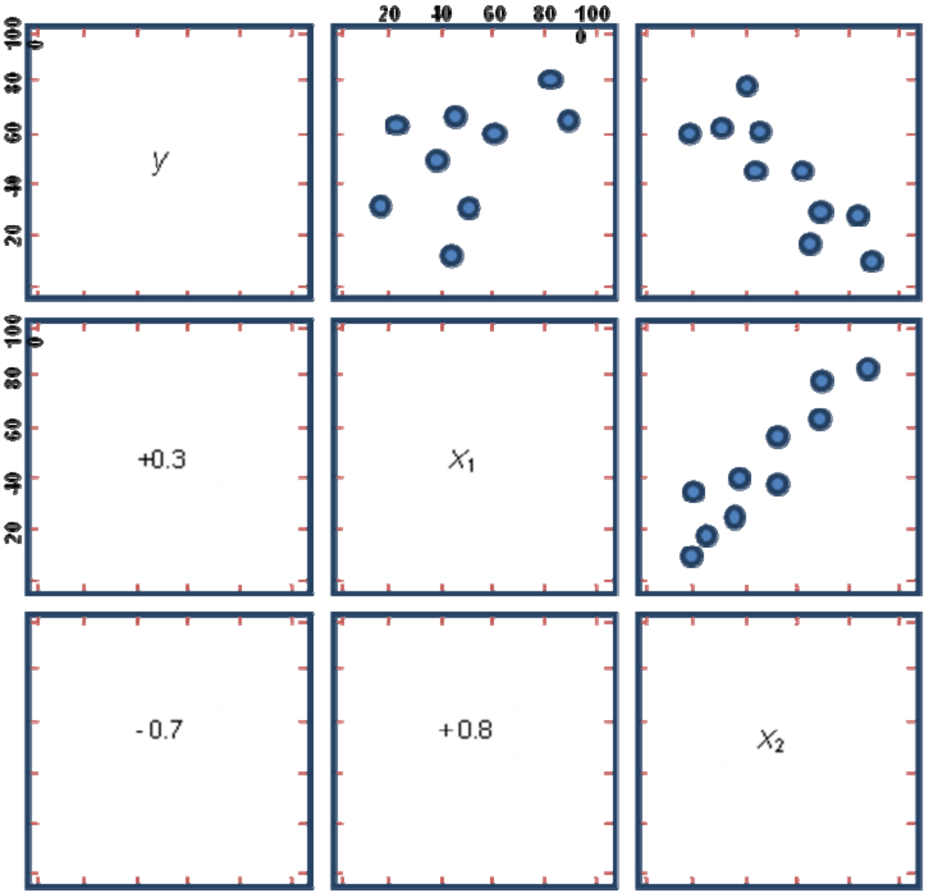
2. Case of more than one explanatory variables

To check the assumption of linearity between the study variable and the explanatory variables, the **scatter plot matrix** of the data can be used. A scatterplot matrix is a two-dimensional array of two-dimension plots where each form contains a scatter diagram except for the diagonal. Thus, each scenario sheds some light on the relationship between a pair of variables. It gives more information than the correlation coefficient between each pair of variables because it provides a sense of linearity or nonlinearity of the relationship and some awareness of how the individual data points are arranged over the region. It is a scatter diagram of $(y \text{ versus } X_1),\ (y \text{ versus } X_2),\ \ldots,\ (y \text{ versus } X_k)$.

Another option to present the scatterplot is

- display the scatterplots in the upper triangular part of the plot matrix.
- Mention the corresponding correlation coefficients in the lower triangular part of the matrix.

Suppose there are only two explanatory variables and the model is $y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$ then the scatterplot matrix looks like as follows.



Such an arrangement helps in examining of plot and corresponding correlation coefficient together. The pairwise correlation coefficient should always be interpreted in conjunction with the corresponding scatter plots because

- the correlation coefficient measures only the linear relationship and
- the correlation coefficient is non-robust, i.e., one or two observations can substantially influence its value in the data.

The presence of linear patterns is reassuring, but the absence of such patterns does not imply that the linear model is incorrect. Most of the statistical software provides the option for creating the scatterplot matrix. The view of all the plots indicates that a multiple linear regression model may provide a reasonable fit to the data.

It is to be kept in mind that we get only the information on pairs of variables through the scatterplot of $(y \text{ versus } X_1),\ (y \text{ versus } X_2),\ \ldots,\ (y \text{ versus } X_k)$ whereas the assumption of linearity is between y and jointly with ($X_1, X_2,.., X_k$).

If some of the explanatory variables are themselves interrelated, then these scatter diagrams can be misleading. Some other methods of sorting out the relationships between several explanatory variables and a study variable are used.

Residual analysis

The **residual** is defined as the difference between the observed and fitted value of study variable. The i^{th} residual is defined as

$$e_i = y_i \sim \hat{y}_i = y_i - \hat{y}_i,\ i = 1, 2, ..., n$$

where y_i is an observation and $\hat{y}_i$ is the corresponding fitted value.

Residual can be viewed as the deviation between the data and the fit. So it is also a measure of the variability in the response variable that is not explained by the regression model.

Residuals can be thought of as the observed values of the model errors. So it can be expected that if there is any departure from the assumptions on random errors, then it should be shown up by the residual. The analysis of residuals help in finding the model inadequacies.

Assuming that the OLSE estimates the regression coefficients in the model $y = X\beta + \varepsilon$, we find that:

- Residuals have zero mean as

$$\begin{aligned} E(e_i) &= E(y_i - \hat{y}_i) \\ &= E(X_i\beta + \varepsilon_i - X_i b) \\ &= X_i\beta + 0 - X_i\beta \\ &= 0 \end{aligned}$$

- An approximate average variance of residuals is estimated by

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-k} = \frac{\sum_{i=1}^{n} e_i^2}{n-k} = \frac{SS_{res}}{n-k} = MS_{res}.$$

- Residuals are not independent as the n residuals have only $n-k$ degrees of freedom. The nonindependence of the residuals has little effect on their use for model adequacy checking as long as n is not small relative to k.

Methods for scaling residuals

Sometimes it is easier to work with scaled residuals. We discuss four ways for scaling the residuals.

1. Standardized residuals:

The residuals are standardized based on the concept of residual minus its mean and divided by its standard deviation. Since $E(e_i)=0$ and MS_{res} estimates the approximate average variance, so logically the scaling of residual is

$$d_i = \frac{e_i}{\sqrt{MS_{res}}},\ i = 1,2,...,n$$

is called as standardized residual for which

$$E(d_i) = 0$$
$$Var(d_i) \approx 1.$$

So a large value of $d_i (> 3,$ say) potentially indicates an outlier.

2. Studentized residuals

The standardized residuals use the approximate variance of e_i as MS_{res}. The studentized residuals use the exact variance of e_i.

We first find the variance of e_i.

In the model $y = X\beta + \varepsilon$, the OLSE of β is $b = (X'X)^{-1}X'y$ and the residual vector is

$$\begin{aligned} e &= y - \hat{y} \\ &= y - Xb \\ &= y - Hy \\ &= (I-H)y \quad \text{where } H = X(X'X)^{-1}X' \\ &= (I-H)(X\beta + \varepsilon) \\ &= X\beta - HX\beta + (I-H)\varepsilon \\ &= X\beta - X\beta + (I-H)\varepsilon \\ &= (I-H)\varepsilon \\ &= \bar{H}\varepsilon \end{aligned}$$

Thus $e = \bar{H}y = \bar{H}\varepsilon$, so the residuals are the same linear transformation of y and ε.

The covariance matrix of residuals is

$$\begin{aligned} V(e) &= V(\bar{H}\varepsilon) \\ &= \bar{H}V(\varepsilon)\bar{H} \\ &= \sigma^2\bar{H} \\ &= \sigma^2(I-H) \end{aligned}$$

and $V(\varepsilon) = \sigma^2 I.$

The matrix $(I-H)$ is symmetric and idempotent but generally not diagonal. So residuals have different variances, and they are correlated.

If h_{ii} is the i^{th} diagonal element of hat matrix H and h_{ij} is the $(i, j)^{th}$ element of H, then

$$\begin{aligned} Var(e_i) &= \sigma^2(1-h_{ii}) \\ Cov(e_i, e_j) &= -\sigma^2 h_{ij}. \end{aligned}$$

Since $0 \le h_{ii} \le 1,$ so if MS_{res} is used to estimate the $Var(e_i)$ then

$$\begin{aligned} \widehat{Var}(e_i) &= \hat{\sigma}^2(1-h_{ii}) \\ &= MS_{res}(1-h_{ii}) \\ \Rightarrow MS_{res} & \text{ overestimates the } Var(e_i). \end{aligned}$$

Now we discuss that h_{ii} is a measure of location of the i^{th} point in x-space.

Regression variable hull (RVH):

It is the smallest convex set containing all the original data $x_i = (x_{i1}, x_{i2}, ..., x_{ik}),\ i = 1, 2, ..., n.$

The h_{ii} depend on the Euclidian distance of x_i from the centroid and on the density of the points in RVH.

In general, if a point has the largest value of h_{ii}, say $h_{\max}$, then it will lie on the boundary of the RVH in a region of the x-space. In such a region, where the density of the observations is relatively low. The set of points x (not necessarily the data points used to fit the model) that satisfy

$$x'(X'X)^{-1}x \le h_{\max}$$

is an ellipsoid enclosing all points inside the RVH. So the location of a point, say, $x_0 = (x_{01}, x_{02}, ..., x_{0k}),$ relative to RVH is rejected by

$$h_{00} = x_0'(X'X)^{-1}x_0.$$

Points for which $h_{00} > h_{\max}$ are outside the ellipsoid containing RVH. If $h_{00} < h_{\max}$ then the point is inside the RVH. Generally, a smaller the value of h_{00} indicates that the point x_0 lies closer to the centroid of the x-space.

Since h_{ii} is a measure of location of the i^{th} point in x-space, the variance of e_i depends on where the point x_i lies. If h_{ii} is small, then $Var(e_i)$ is larger, which indicates a poorer fit. So the points near the centre of the x-space have poorer least-squares fit than the residuals at more remote locations. Violation of model assumptions are more likely at distant points, and these violations may be hard to detect from the inspection of ordinary residuals e_i (or the standardized residuals d_i) because their residuals will usually be smaller.

So a logical procedure is to examine the studentized residuals of the form

$$r_i = \frac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}$$

in place of e_i (or d_i). For r_i,

$$E(r_i) = 0$$
$$Var(r_i) = 1$$

regardless of the location of x_i when the form of the model is correct.

In many situations, the variance of residuals stabilizes (particularly in large data sets), and there may be little difference between d_i and r_i. In such cases d_i and r_i often convey equivalent information.

However, since any point with a

- large residual and
- large h_{ii}

is potentially highly influential on the least-squares fit, so examination of r_i is generally recommended.

If there is only one explanatory variable then

$$r_i = \frac{e_i}{\sqrt{MS_{res}\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{s_{xx}}\right)\right]}}, \; i = 1, 2, ..., n \,.$$

- When x_i is close to the midpoint of x-data, i.e., $x_i - \bar{x}$ is small then estimated standard deviation of e_i is large.
- Conversely, when x_i is near the extreme ends of the range of x-data, then $x_i - \bar{x}$ is large and estimated standard deviation of e_i is small.
- When n is really large, the effect of $(x_i - \bar{x})^2$ is relatively small. So in big data sets, r_i may not differ dramatically from d_i.

PRESS residuals:

The PRESS residuals are defined as $(y_i - \hat{y}_{(i)})$ where $\hat{y}_{(i)}$ is the fitted value of the i^{th} response based of all the observation except the i^{th} one.

Reason: If y_i is really unusual, then the regression model based on all the observations may be overly influenced by this observation. This could produce a $\hat{y}_i$ that is very similar to y_i and consequently e_i will be small. So it will be challenging to detect any outlier.

If y_i is deleted, then $\hat{y}_{(i)}$ cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier.

Procedure

- Delete the i^{th} observation,
- Fit the regression model to remaining $(n-1)$ observations,
- Calculate the predicted value of y_i corresponding to the deleted observation.
- The corresponding prediction error $e_{(i)} = y_i - y_{(i)}$
- Calculate $e_{(i)}$ for each $i = 1, 2, ..., n$.

These prediction errors are called **PRESS residuals** because they are used in computing the prediction error sum of squares. They are also called as **deleted residuals**.

Now we establish a relationship between e_i and $e_{(i)}$.

Relation between e_i and $e_{(i)}$

Let $b_{(i)}$ be the vector of regression coefficients estimated by withholding the i^{th} observations. Then

$$b_{(i)} = \left(X'_{(i)} X_{(i)}\right)^{-1} X'_{(i)} y_{(i)}$$

where $X_{(i)}$ is the X-matrix without the vector of i^{th} observation and $y_{(i)}$ is the y-vector without the i^{th} observation. Then

$$\begin{aligned} e_{(i)} &= y_i - \hat{y}_{(i)} \\ &= y_i - x_i \hat{b}_{(i)} \\ &= y_i - x_i (X'_{(i)} X_{(i)})^{-1} X'_{(i)} y_{(i)} \end{aligned}$$

We use the following result in further analysis.

Result: If $X'X$ is a $k \times k$ matrix and x be its i^{th} row vector then $(X'X - x'x)$ denotes the $X'X -$ matrix with the i^{th} row withheld. Then

$$\left[X'X - x'x\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x'x(X'X)^{-1}}{1 - x(X'X)^{-1}x'}.$$

Using this result, we can write

$$\left[X_{(i)}^{'}X_{(i)}\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i^{'}x_i(X'X)^{-1}}{1-h_{ii}}$$

where $h_{ii} = x_i(X'X)^{-1}x_i^{'}$.

Then

$$\begin{aligned}
e_{(i)} &= y_i - x_i\left(X_{(i)}^{'}X_{(i)}\right)^{-1}X_{(i)}^{'}y_{(i)} \\
&= y_i - x_i\left[(X'X)^{-1} + \frac{(X'X)^{-1}x_i^{'}x_i(X'X)^{-1}}{1-h_{ii}}\right]X_{(i)}^{'}y_{(i)} \\
&= y_i - x_i(X'X)^{-1}X_{(i)}^{'}y_{(i)} - \frac{x_i(X'X)^{-1}x_i^{'}x_i(X'X)^{-1}X_{(i)}^{'}y_{(i)}}{1-h_{ii}} \\
&= y_i - x_i(X'X)^{-1}X_{(i)}^{'}y_{(i)} - \frac{h_{ii}x_i(X'X)^{-1}X_{(i)}^{'}y_{(i)}}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - (1-h_{ii})x_i(X'X)^{-1}X_{(i)}^{'}y_{(i)} - h_{ii}x_i(X'X)^{-1}X_{(i)}^{'}y_{(i)}}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}X_{(i)}^{'}y_{(i)}}{1-h_{ii}}.
\end{aligned}$$

Using $X'y = X_{(i)}^{'}y_{(i)} + x_i^{'}y_i$ (as x_i is $1 \times k$ vector), we can write

$$\begin{aligned}
e_{(i)} &= \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}(X'y - x_i^{'}y_i)}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}X'y + x_i(X'X)^{-1}x_i^{'}y_i}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - x_ib + h_{ii}y_i}{1-h_{ii}} \\
&= \frac{y_i - x_ib}{1-h_{ii}} \\
&= \frac{e_i}{1-h_{ii}}.
\end{aligned}$$

Looking at the relationship between e_i and $e_{(i)}$, it is clear that calculating the PRESS residuals does not require fitting in different regressions. The $e_{(i)}'s$ are just the ordinary residuals weighted according to the diagonal elements h_{ii} of H. It is possible to calculate the PRESS residuals from the residuals of a single least-squares fit to all n observations.

Residuals associated with points for which h_{ii} is large will have large PRESS residuals. Such points will generally be **high influence** points.

The large difference between ordinary residual and PRESS residuals indicate a point where the model fits the data well, and a model without that point **predicts** poorly.

Now

$$\begin{aligned}Var(e_{(i)}) &= Var\left(\frac{e_i}{1-h_{ii}}\right)\\ &= \frac{1}{(1-h_{ii})^2}Var(e_i)\\ &= \frac{1}{(1-h_{ii})^2}(1-h_{ii})\sigma^2\\ &= \frac{\sigma^2}{1-h_{ii}}.\end{aligned}$$

The **standardized PRESS residual** is

$$\frac{e_{(i)}}{\sqrt{Var(e_{(i)})}} = \frac{\left(\frac{e_i}{1-h_{ii}}\right)}{\sqrt{\frac{\sigma^2}{(1-h_{ii})}}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

which is same as the Studentized residuals.

4. R-student

The studentized residual r_i is often considered as an outlier diagnostic and MS_{res} is used as an estimate of σ^2 in computing r_i. This is referred to as **internal scaling** of the residuals because MS_{res} is an internally generated estimate of σ^2 obtained from the fitting the model to all n observation .

Another approach is to use an estimate of σ^2 based on a data set with i^{th} observation removed, say $s_{(i)}^2$.

First, we derive an expression for $s_{(i)}^2$. Using the identity

$$\left[X_{(i)}^{'}X_{(i)}\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i^{'}x_i(X'X)^{-1}}{1-h_{ii}}.$$

Post multiply both sides by $(X'y - x_i^{'}y_i)$, we get

$$b_{(i)} = b - (X'X)^{-1}x_i^{'}y_i + \frac{(X'X)^{-1}x_i^{'}x_i(X'X)^{-1}(X'y - x_i^{'}y_i)}{1-h_{ii}}$$

$$b - b_{(i)} = (X'X)^{-1}x_i^{'}y_i - \frac{(X'X)^{-1}x_i^{'}x_i\left[b - (X'X)^{-1}x_i^{'}y_i\right]}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})(X'X)^{-1}x_i^{'}y_i - (X'X)^{-1}x_i^{'}x_ib + (X'X)^{-1}x_i^{'}h_{ii}y_i}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i^{'}\left[y_i - x_ib\right]}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i^{'}e}{1-h_{ii}}$$

$$b_{(i)} = b - \frac{(X'X)^{-1}x_i^{'}e}{1-h_{ii}}.$$

Now consider

$$(n-k-1)s_{(i)}^2 = \sum_{j\neq i=1}^{n}(y_j - x_jb_{(i)})^2$$

$$= \sum_{j=1}^{n}\left[y_j - x_jb + \frac{x_j(X'X)x_i^{'}e_i}{1-h_{ii}}\right]^2 - \left(y_i - x_ib + \frac{h_{ii}e_i}{1-h_{ii}}\right)^2$$

$$= \sum_{j=1}^{n}\left[e_j + \frac{h_{ij}e_i}{1-h_{ii}}\right]^2 - \frac{e_i^2}{(1-h_{ii})^2}$$

$$= \sum_{j=1}^{n}e_j^2 + \frac{2e_i}{1-h_{ii}}\sum_{j=1}^{n}e_jh_{ij} + \frac{e_i^2}{(1-h_{ii})^2}\sum_{j=1}^{n}h_{ij}^2 - \frac{e_i}{(1-h_{ii})}$$

$$= \sum_{j=1}^{n}e_j^2 + \frac{h_{ii}e_i^2}{(1-h_{ii})^2} - \frac{e_i^2}{(1-h_{ii})^2}$$

$$= \sum_{j=1}^{n}e_j^2 - \frac{e_i^2}{1-h_{ii}} \quad \text{(using } Hy = H\hat{y},\ \sum_{j=1}^{n}e_jh_{ij} = 0, \sum_{j=1}^{n}h_{ij}^2 = h_{ij} \text{ as } H \text{ is idempotent)}$$

$$= (n-k)MS_{res} - \frac{e_i^2}{1-h_{ii}}$$

Thus

$$s_{(i)}^2 = \frac{1}{n-k-1}\left[(n-k)MS_{res} - \frac{e_i^2}{1-h_{ii}}\right].$$

This estimate of σ^2 is used instead of MS_{res} to produce an **externally studentized residual**, usually called ***R*-student** given by

$$t_i = \frac{e_i}{\sqrt{s_{(i)}^2(1-h_{ii})}}, \; i = 1, 2, ..., n.$$

In many situations, t_i will differ little with r_i. However, if the i^{th} observation is influential, then $s_{(i)}^2$ can differ significantly from MS_{res} and the $R-$student statistic will be more sensitive to this point.

Under the usual regression assumption, t follows a t-distribution with $(n-k-1)$ degrees of freedom.

Residual plots

The graphical analysis of residuals is a very effective way to investigate the adequacy of the fit of a regression model and to check the underlying assumptions. Various types of graphics can be examined for different assumptions, and these graphics are generated by regression software. It is better to plot the original residuals as well as scaled residuals. Typically, the studentized residuals are plotted as they have constant variance.

Normal probability plot

The assumption of normality of disturbances is very much needed for the validity of the results for testing of hypothesis, confidence intervals and prediction intervals. Small departures from normality may not affect the model significantly, but gross nonnormality is more dangerous. The normal probability plots help in verifying the assumption of normal distribution. If errors are coming from a distribution with thicker and heavier tails than normal, then the least-squares fit may be sensitive to a small set of data. Heavy tailed error distribution often generates outliers that “pull” the least-squares too much in their direction. In such cases, other estimation techniques like robust regression methods should be considered.

The normal probability plots is a plot of the **ordered standardized residuals** versus the so-called normal scores. The normal scores are the cumulative probability

$$P_i = \frac{\left(i - \frac{1}{2}\right)}{n}, \; i = 1, 2, ..., n.$$

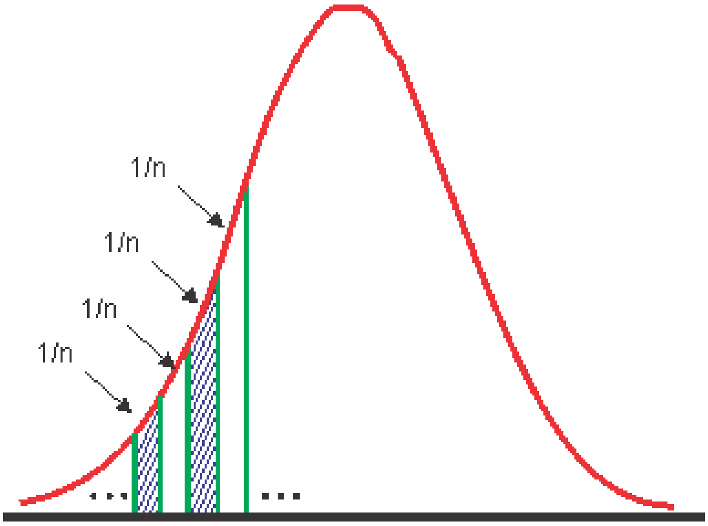
If the residuals $e_1, e_2, ..., e_n$ are ordered and ranked in increasing order as

$$e_{[1]} < e_{[2]} < ... < e_{[n]},$$

then the $e_{[i]}$'s are plotted against P_i and the plot is called normal probability plot. If the residuals are normally distributed, then the ordered residuals should be approximately the same as the ordered normal scores. So the resulting points should lie around the straight line with an intercept zero and a slope of one (these are the mean and standard distributions of standardized residuals).

The rationales behind plotting $e_{[i]}$ against $P_i = \dfrac{\left(i - \frac{1}{2}\right)}{n}$ is as follows:

- Divide the whole unit area under the normal curve into n equal areas.
- We have a sample of size n data sets.
- We might "except" that one observations lies is each section, so marked out.
- The first section has one point, so the cumulative probability is $P_1 = 1/n$. Second section has one point, so cumulative probability up to second section is $P_2 = (1/n) + (1/n) = 2/n$ and so on.
- Then i^{th} ordered residual observation is plotted against the cumulative area to the middle of i^{th} section, which is $\dfrac{\left(i - \frac{1}{2}\right)}{n}$.
- The factor ½ is used for end correction as all the observations scattered inside the stripe are assumed to be concentrated at the midpoint of the stripe.



Different software uses a different criterion. For example, BMDP uses

$$P_i = \frac{i - \frac{1}{3}}{n + \frac{1}{3}}$$

which produces detrended normal probability plots from which slope is removed.

Minitab uses $P_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$ and converts to a normal score.

Such differences are not crucial in real use.

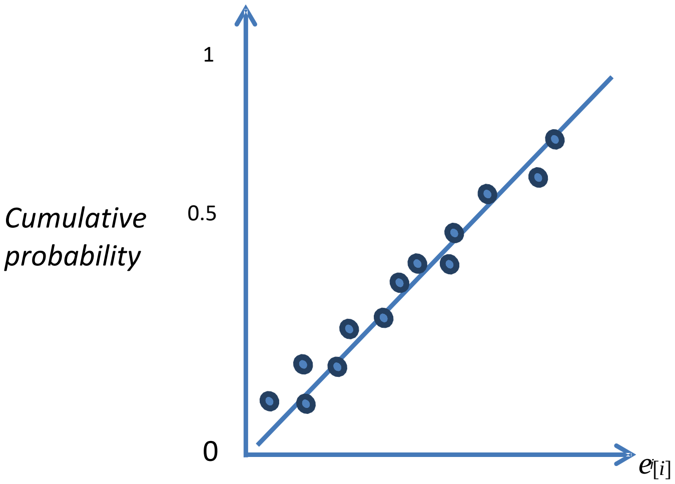
The straight line is usually determined visually with emphasis on the central values rather than the extremes. A substantial departure from a straight line indicates that the distribution is not normal.

Sometimes the normal probability plots are constructed by plotting the ranked residuals $e_{[i]}$ against the expected normal value $\Phi^{-1}\left[\frac{\left(i - \frac{1}{2}\right)}{n}\right]$ where Φ denotes the standard normal cumulative distribution. This follows from the fact that

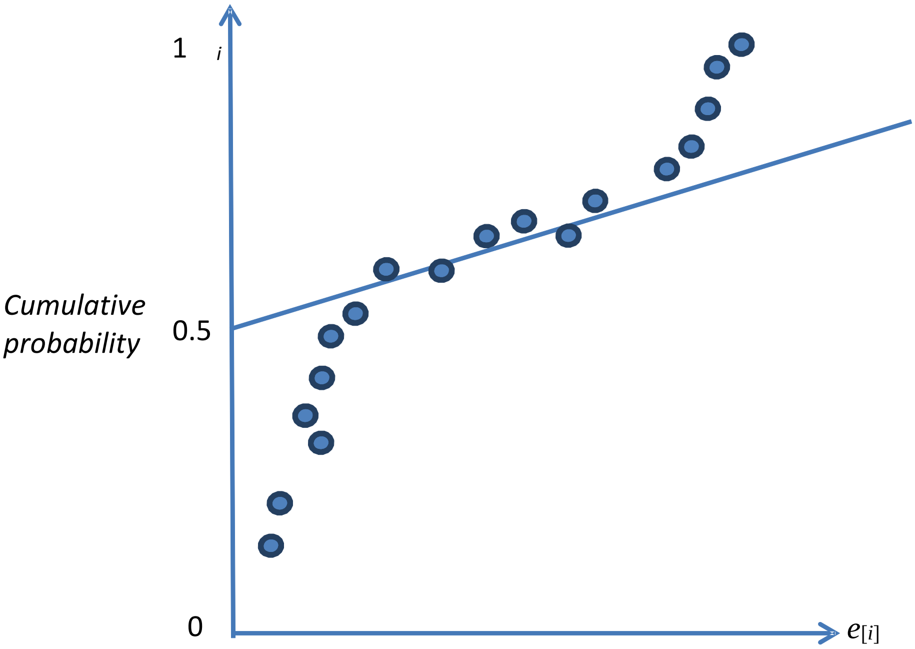
$$E\left[e_{[i]}\right] \approx \Phi^{-1}\left[\frac{\left(i - \frac{1}{2}\right)}{n}\right].$$

Various interpretations of the graphic patterns are as follows.

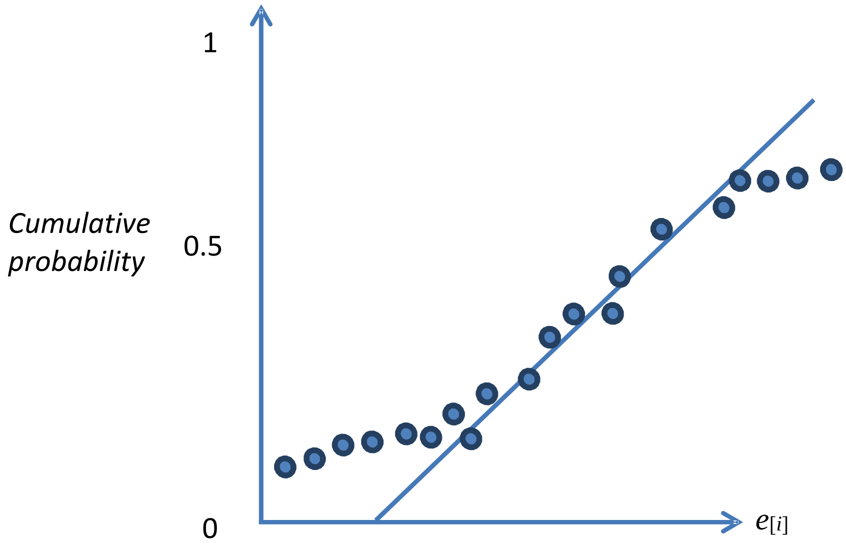
(a) This figure has an ideal normal probability plot. Points lie approximately on the straight line and indicate that the underlying distribution is normal.



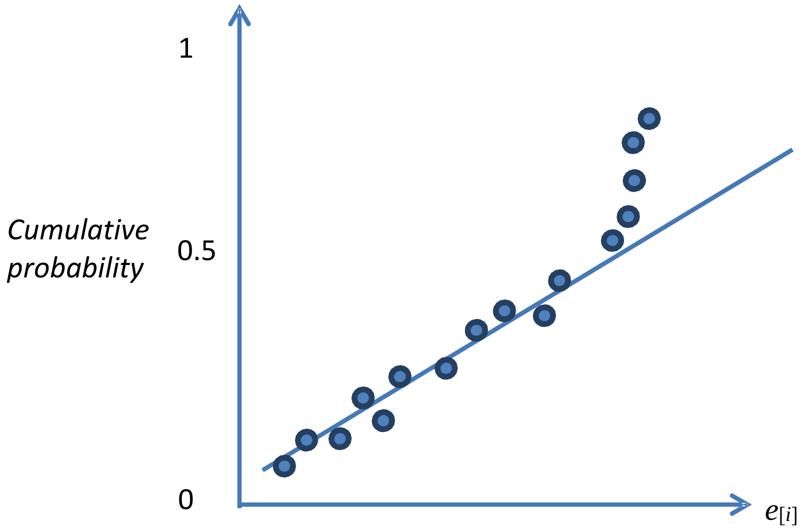
(b) This figure has sharp upward and downward curves at both extremes. This indicates that the underlying distribution is heavy-tailed, i.e., the tails of the underlying distribution are thicker than the tails of normal distribution.



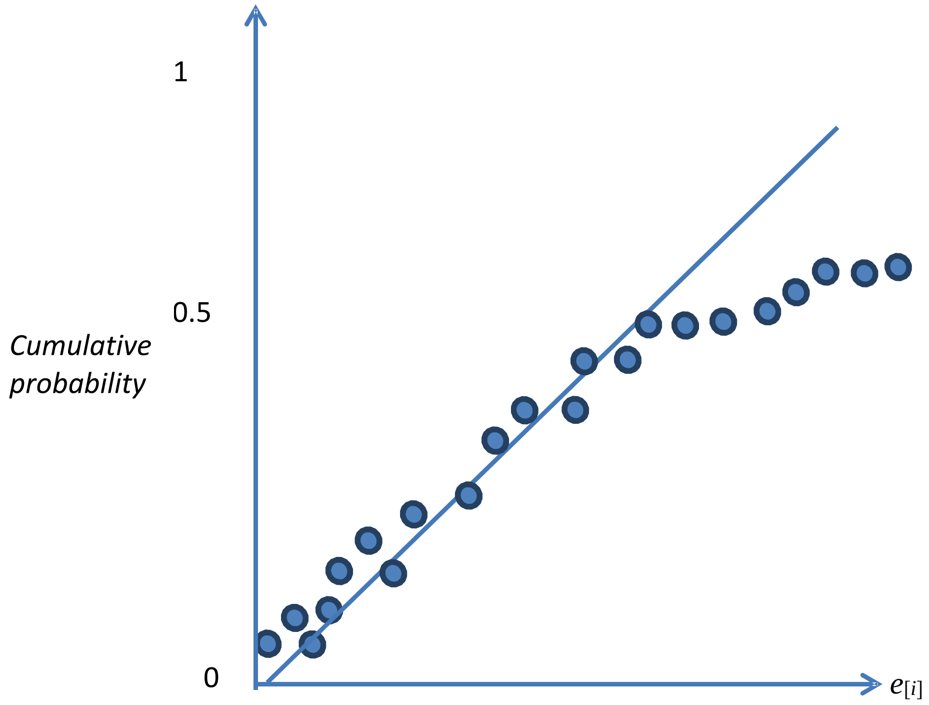
(c) This figure has flattening at the extremes for the curves. This indicates that the underlying distribution is light-tailed, i.e., the tails of the underlying distribution are thinner than the tails of normal distribution.



(d) This figure has a sharp change in the direction of the trend in an upward direction from the mid. This indicates that the underlying distribution is positively skewed.



(e) This figure has a sharp change in the direction of the trend in the downward direction from the mid. This indicates that the underlying distribution is negatively skewed.



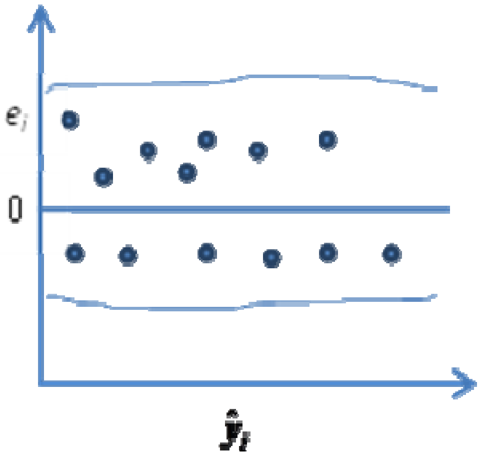
Some experience and expertise are required to interpret the normal probability plots because the samples taken from a normal distribution will not plot precisely as a straight line.

- Small sample sizes $(n \leq 16)$ often produce normal probability plots that deviate substantially from linearity.
- Larger sample sizes $(n \geq 32)$ produce plots which are much better behaved.
- Usually, about $n = 20$ is required to produce stable and easily interpretable normal probability plots.
- If residuals are not from a random sample, normal probability plots often exhibit no unusual behaviour even if the disturbances (ε_i) are not normally distributed. Such residuals are often remnants of a parametric estimation process and are linear combinations of the model errors (ε_i).
- Thus fitting the parameters tends to destroy the evidence of nonnormality in the residuals. Consequently, we can not rely on the normal probability plots to detect the departures from normality.
- Commonly seen defect found is normal probability plots is the occurrence of one or two large residuals. Sometimes, this is an indication that the corresponding observations are outliers.

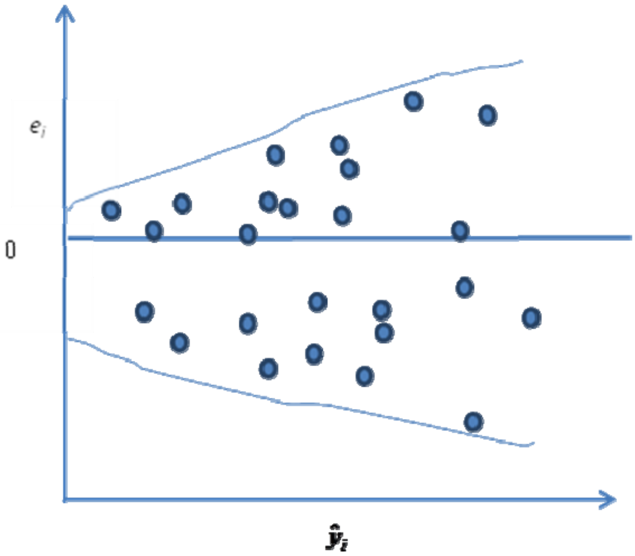
Plots of residuals against the fitted value

A plot of residuals (e_i) or any of the scaled residuals $(d_i, r_i$ or $t_i)$ versus the corresponding fitted values $\hat{y}_i$ is helpful in detecting several common types of model inadequacies. Following types of plots of $\hat{y}_i$ versus e_i have particular interpretations:

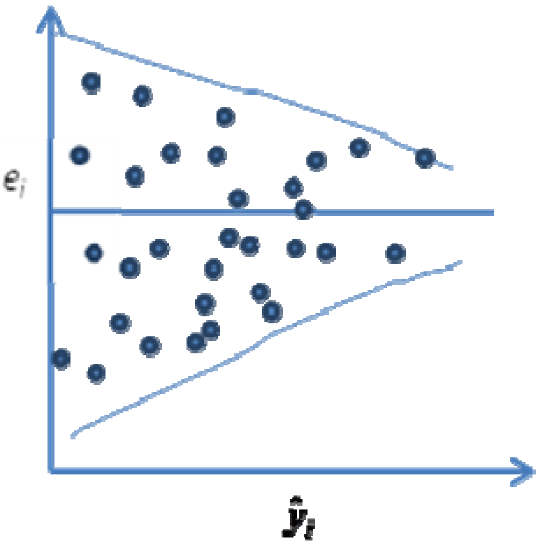
(a) If the plot is such that the residuals can be contained in a **horizontal band** fashion (and residual fluctuates more or less in a random manner inside the band), then there are no visible model defects.



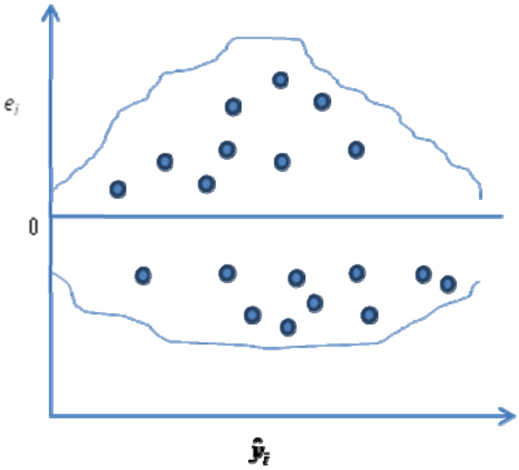
(b) It plot is such that the residuals can be contained in an **outward opening funnel** then such pattern indicates that the variance of errors is not constant, but it is an increasing function of y.



(c) If plots are such that the residuals can be accommodated in an inward opening funnel, then such a pattern indicates that the variance of errors is not constant, but it is a decreasing function of y.

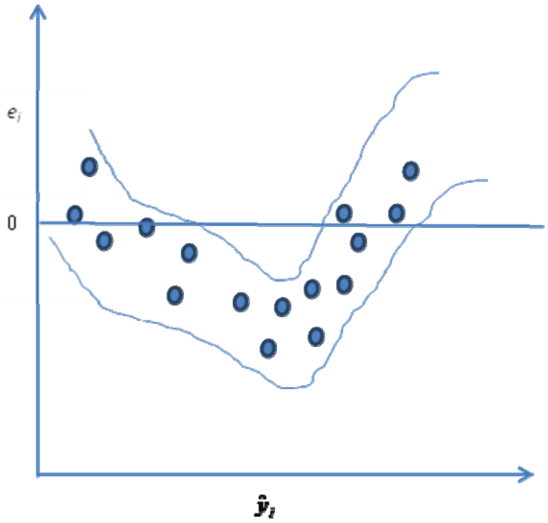


(d) If the plot is such that the residuals can be accommodated inside a double bow, then such a pattern indicates that the variance of errors is not constant but y is a proportion between 0 and 1. The y then may have a Binomial distribution. The variance of a Binomial proportion near 0.5 is higher as compared to near-zero or 1. So the assumed relationship between y and X's is nonlinear.



The usual approach to deal with such inequality of variances is to apply a suitable transformation to either the explanatory variables or the study variable or use the method of weighted least squares. In practice, transformations on study variable are generally employed to stabilize the variance.

(e) If the plot is such that the residuals are contained inside a curved plot, then it indicates nonlinearity. The assumed relationship between y and X's is non-linear. This could also mean that some other explanatory variables are needed in the model. For example, a squared error term may be necessary. Transformations on explanatory variables and/or study variable may also be helpful in these cases.



Note: A plot of residuals against $\hat{y}_i$ may also reveal one or more unusually large residuals. These points are potential outliers. Large residuals that occur at the extreme $\hat{y}_i$ values could also indicate that either the variance is not constant or the true relationship between y and X is nonlinear. These possibilities should be investigated before the points are considered outliers.

Plots of residuals against explanatory variable

Plotting of residuals against the corresponding values of each explanatory variable can also be helpful. We proceed as follows

- Consider the residual on $Y-$axis and values of the j^{th} explanatory variable x_{ij}'s, $(i=1,2,...,n)$ on $X-$axis. This is the same way as we have plotted the residuals against $\hat{y}_i$. In place of $\hat{y}_i$'s, now we consider x_{ij}'s.
- Interpretation of the plots is the same as in the case of plots of residuals versus $\hat{y}_i$. This is as follows. If all the residuals are contained in
  - a horizontal band, and the residuals fluctuate more or less in a random fashion within this band, then it is desirable, and there are no visible model defects.
  - an outward opening funnel shape or inward opening funnel shape, then it indicates that the variance is nonconstant.

- a double bow pattern or nonlinear pattern then it indicates the assumed relationship between y and x_j is not correct. The possibilities like y may be a proportion, a higher-ordered term is X_j (e.g. X_j^2) are needed, or a transformation is required to be considered in such a case.

Note 1: In the case of simple linear regression, it is not necessary to plot residuals versus $\hat{y}_i$ and explanatory variable. The reason is that the fitted values $\hat{y}_i$ are linear combinations of the values of the explanatory variable X_i, so the plots would only differ is the scale for the abscissa ($X -$axis).

Note 2: It is also helpful to plot the residuals against explanatory variables that are **not currently** is the model, but which could potentially be included. Any structure in the plot of residuals versus an omitted variable indicates that incorporation of that variable could improve the model.

Note 3: Plotting residuals versus explanatory variable is not always the most effective way to reveal whether a curvature effect (or a transformation) is required for that variable in the model. Partial regression plots are more effective in investigating the relationship between the study variable and explanatory variables.

Plots of residuals in time sequence

If the time sequence in which the data were collected is known, then the residuals can be plotted against the time order. We proceed as follows:

- Consider the residuals on Y-axis and time order on $X -$axis. This is the same way as we have plotted the residuals against $\hat{y}_i$. In place of $\hat{y}_i$, just use the time order.
- Interpretation of the plots is the same as in the case of plots of residuals versus $\hat{y}_i$. This is as follows.

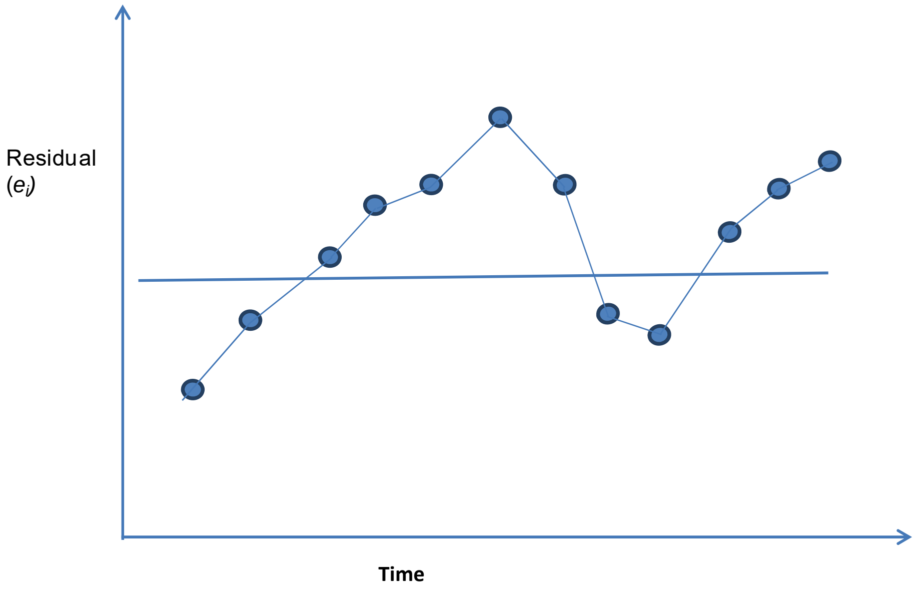
If all the residuals are contained in

- a horizontal band, and the residuals fluctuate more or less in a random fashion within this band, then it is desirable and indicates that there are no obvious model deflects.
- An outward opening funnel shape or inward opening funnel shape, then it indicates that the variance is not constant but changing with time.
- Double bow pattern or nonlinear pattern, then it indicates that the assumed relationship is nonlinear. In such a case, the linear or quadratic terms in time should be added to the model.

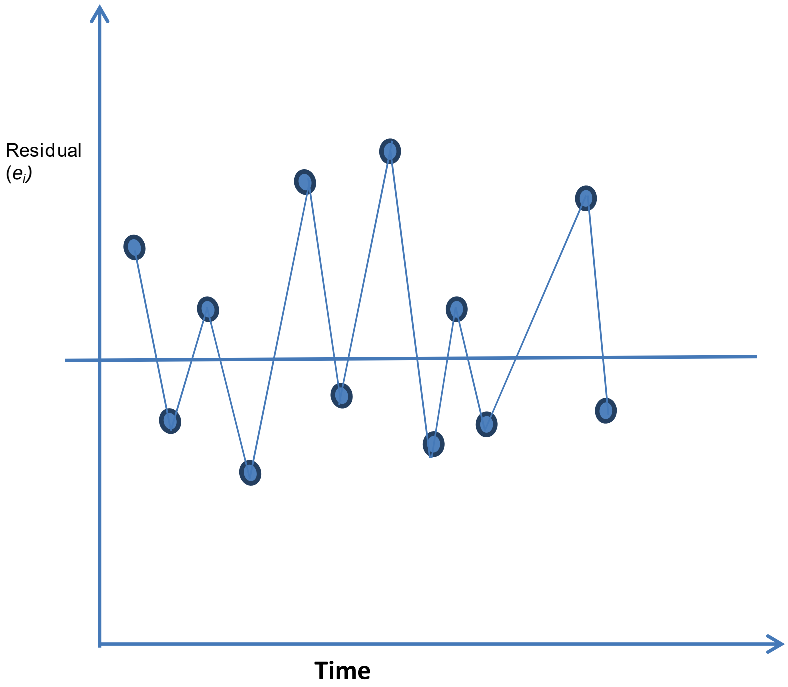
The time sequence plot of residuals may indicate that the errors at one time period are correlated with those at other time periods. The correlation between model errors at different time periods is called **autocorrelation.**

If we have a plot like following, then it indicates the presence of autocorrelation.

Following type of figure indicates the presence of positive autocorrelation



Following type of figure indicates the presence of negative autocorrelation



The methods to detect the autocorrelation and to deal with the time-dependent data are available under time series analysis. Some measures are discussed further in the module on autocorrelation.

Partial regression and partial residual plots

Partial regression plot (also called as **added variable** plot or **adjusted variable plot**) is a variation of the plot of residuals versus the predictor. It helps better to study the marginal relationship of an explanatory variable given the other variables that are in the model. A limitation of the plot of residuals versus an explanatory variable is that it may not completely show the correct or complete marginal effect of an explanatory variable given the other explanatory variables in the model. The partial regression plot helps in evaluating whether the relationship between study and explanatory variables is correctly specified. They provide information about the marginal usefulness of a variable that is not currently in the model.

In partial regression plot

- Regress y on all the explanatory variable except the j^{th} explanatory variables X_j and obtain the residuals $e\left[y / X_{(j)}\right]$, say where $X_{(j)}$ denotes the X -matrix with X_j removed.
- Regress X_j on all other explanatory variables and obtain the residuals $e\left[X_j / X_{(j)}\right]$, say
- Plot both these residuals against $e\left[X_j / X_{(j)}\right]$.

These plots provide information about the nature of the marginal relationship for j^{th} explanatory variable X_j under consideration.

If X_j enters into the model linearly, them the partial regression plot should show a linear relationship, i.e., the partial residuals will fall along a straight line with a nonzero scope.

See how:

Consider the model

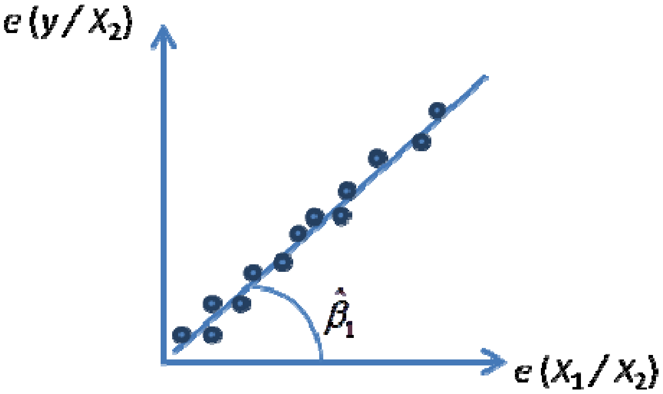
$$\begin{aligned} y &= X\beta + \varepsilon \\ &= X_{(j)}\beta_{(j)} + X_j\beta_j + \varepsilon \end{aligned}$$

then residual is $e = (I - H)$ where $H = X(X'X)^{-1}X'$ and $H_{(j)} = X_{(j)}(X_{(j)}^{'}, X_{(j)}^{'})^{-1}X_{(j)}^{'}$ is the H-matrix based on $X_{(j)}$. Premultiply $y = X\beta + \varepsilon$ by $(I - H_{(j)})$ and noting that $(I - H_{(j)})X_{(j)} = 0,$ we have

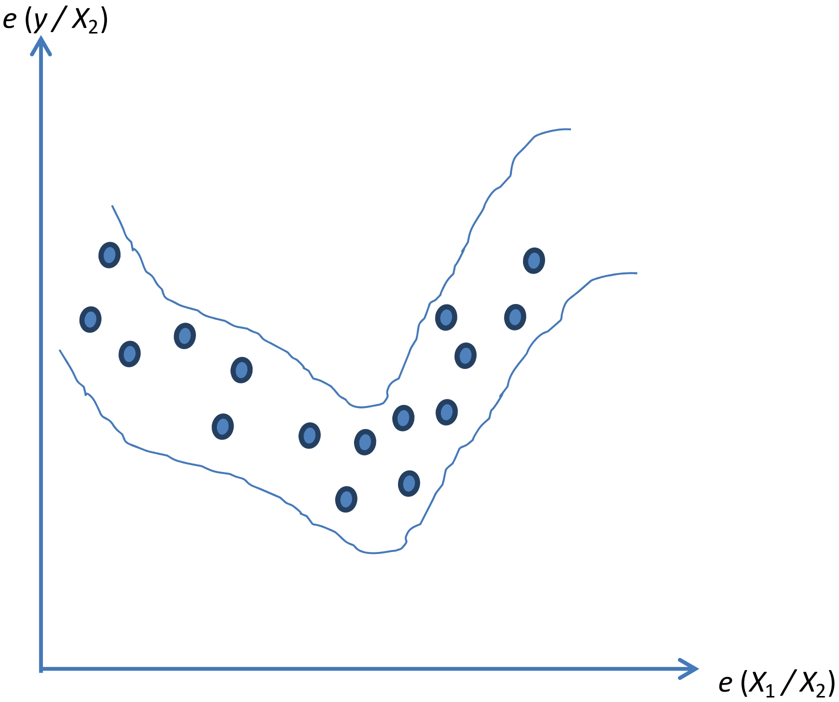
$$\begin{aligned} (I - H_{(j)})y &= (I - H_{(j)})X_{(j)}\beta + \beta_j(I - H_{(j)})X_j + (I - H_{(j)})\varepsilon \\ &= 0 + \beta_j(I - H_{(j)})X_j + (I - H_{(j)})\varepsilon \\ e\left[y / X_{(j)}\right] &= \beta_j\, e\left[X_j / X_{(j)}\right] + \varepsilon^* \end{aligned}$$

where $\varepsilon^* = (I - H_{(j)})\varepsilon.$

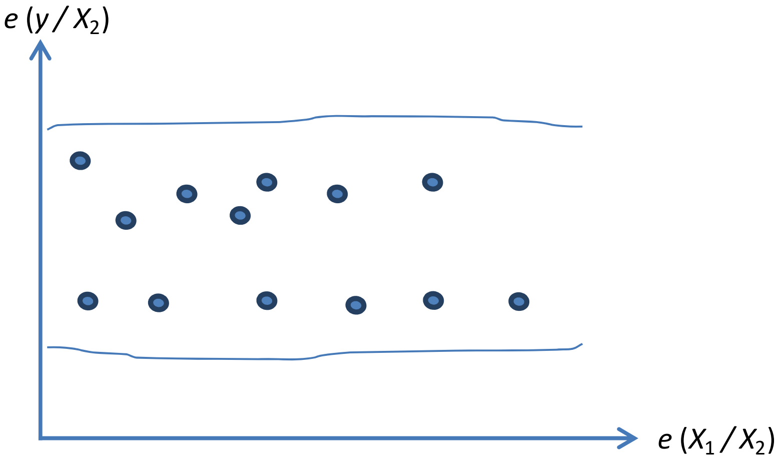
This suggests that a partial regression plot which is a plot between $e\left[y / X_{(j)}\right]$ and $e\left[X_j / X_{(j)}\right]$ (like between y and X) should have a slope β_j. Thus if X_j linearly enters the regression, the partial regression plot should show linear relationship passing through the origin. For example, like



If the partial regression plot shows a curvilinear band, then higher-order terms in X_j or a transformation may be helpful.



If X_j is a candidate variable which is considered for inclusion in the model, then a horizontal band on the regression plot indicates that there is no additional useful information in X_j for predicting y. This indicates that β_j is nearly zero.



Example: Consider a model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon .$$

We want to know about the nature of marginal relationship for X_1 and also want to know whether the relationship between y and X_1 is correctly specified or not ?

To obtain the partial regression plot.

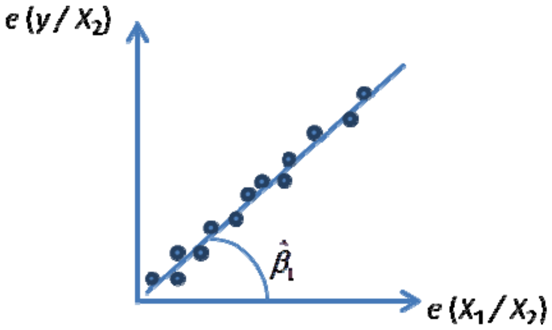
- Regress y on X_2 and obtain the fitted values and residuals

  $$\hat{y}_i(X_2) = \hat{\theta}_0 + \hat{\theta}_1 x_{i2}$$
  $$e_i(y/X_2) = y_i - \hat{y}_i(X_2),\ i = 1,2,...,n.$$

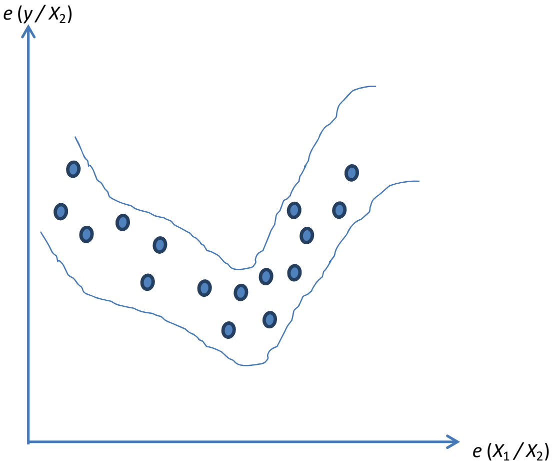
- Regress X_1 on X_2 and find the residuals

  $$\hat{X}_{i1}(X_2) = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}$$
  $$e_i(X_1/X_2) = x_{i1} - \hat{X}_{i1}(X_2),\ \ i = 1,2,...,n.$$

- Plot $e_i(y/X_2)$ against the X_1 residuals $e_i(X_1/X_2)$.
- If X_1 enters into the model linearly, then the plot will look like as follows:

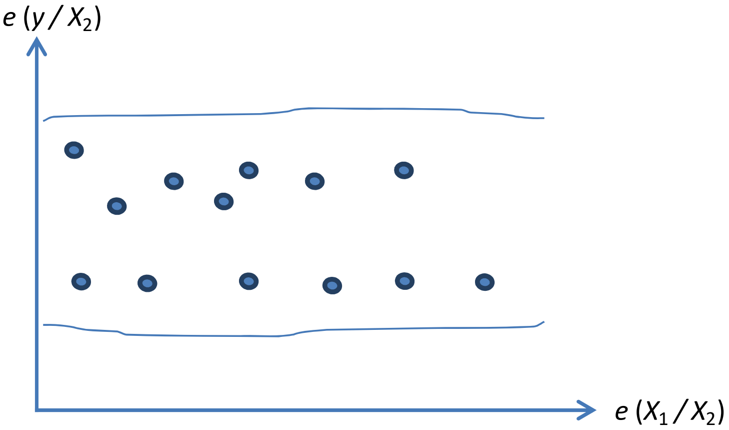


- The slope of this line is the regression coefficient of X_1 in the multiple linear regression model.



- If the partial regression plot shows a curvilinear band, then higher-order terms in X_1 or a transformation X_1 may be helpful.

- If X_1 is a candidate variable which is considered for inclusion in the model, then a horizontal band on the regression plot indicates that there is no additional useful information for predicting y.



Some comments on partial regression plots:

1. Partial regression plots need to be used with caution as they only suggest a possible relationship between study and explanatory variables. The plots may not give information about the proper form of the relationship of several variables that are already in the model are incorrectly specified.

   Some alternative forms of relationship between study and explanatory variables should also be examined with several transformations.

   Residual plots for these models should also be examined to identify the best relationship or transformation.

2. Partial regression plots will not, in general, detect the interaction effect among the regressors.
3 Partial regression plots are affected by the existence of the exact relationship among explanatory variables (Problem of multicollinearity) and the information about the relationship between study, and explanatory variables may be incorrect.

   In such cases, it is better to construct a scatter plot of explanatory variables like X_i verus X_j. If they are highly correlated, multicollinearity is introduced, and properties of estimators like ordinary least squares of regression coefficients are disturbed.

Partial residual

A residual plot closely related to the partial regression plot in the partial residual plot. It is designed to show the relationship between the study and explanatory variables.

Suppose the model has k explanatory variable and we are interested in the j^{th} explanatory variable X_j. Then $X = (X_{(j)}, X_j)$ where $X_{(j)}$ is the $X-$matrix with X_j removed. The model is

$$\begin{aligned} y &= X\beta + \varepsilon \\ &= X_{(j)}\beta_{(j)} + X_j\beta_j + \varepsilon \end{aligned}$$

where $\beta_{(j)}$ is the vector of all $\beta_1, \beta_2, ..., \beta_k$ except β_j. The fitted model is

$$\hat{y} = X_{(j)}\hat{\beta}_{(j)} + X_j\hat{\beta}_j + e$$

or $\hat{y} - X_{(j)}\hat{\beta}_{(j)} = X_j\hat{\beta}_j + e$

where e is the residual based on all k explanatory variables.

Then partial residual for X_j $(j = 1, 2, .., k)$ is given by

$$\hat{y} - X_{(j)}\hat{\beta}_{(j)} = X_j\hat{\beta}_j + e$$

or $e(y / X_j) = e + \hat{\beta}_j X_j$

or $e_i^*(y / X_j) = e_i + \hat{\beta}_j x_{ij},\ i = 1, 2, ..., n$.

Partial residuals plots

A residual plot closely related to the partial regression plot in the partial residual plot. It is designed to show the relationship between the study and explanatory variables.

Suppose the model has k explanatory variables $X_1, X_2, ..., X_k$. The partial residuals for X_j are defined as

$$e_i^*(y/X_j) = e_i + \hat{\beta}_j x_{ij},\ i = 1, 2, ..., n$$

where e_i are the residuals from the model containing all the k explanatory variables and $\hat{\beta}_j$ is the estimate of the j^{th} regression coefficient.

When $e_i^*(y/X_j)$ are plotted against x_{ij}, the resulting display has a slope $\hat{\beta}_j$. The interpretation of the partial residual plot is very similar to that of the partial regression plot.

Statistical tests on residuals

We may apply certain statistical tests to the residuals to obtain a quantitative measure of some of the model inadequacies. They are not widely used. In many applications, residual plots are more informative than the corresponding tests. However, some residual plots do require some skill and experience to interpret. In such cases, the statistical tests may prove useful.

The PRESS statistic

The PRESS residuals are defined as

$$e_{(i)} = y_i - \hat{y}_{(i)},\ i = 1, 2, ..., n$$

where $\hat{y}_{(i)}$ is the predicted value of the i^{th} observed study variable based on a model fit to the remaining $(n-1)$ points. The large residuals are useful in identifying those observations where the model does not fit well or the observations for which the model is likely to provide poor predictions for future values.

The prediction sum of squares is defined as the sum of squared PRESS residuals and is called as PRESS statistic as

$$PRESS = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)} \right]^2 = \sum_{i=1}^{n} \left[\frac{e_i}{1 - h_{ii}} \right]^2$$

The PRESS statistic is a measure of how well a regression model will perform in predicting new data. So this is also a measure of model quality. A model with a small value of PRESS is desirable. This can also be used for comparing regression models.

R^2 for prediction based on PRESS

The PRESS statistic can be used to compute an R^2-like statistic for prediction, say

$$R^2_{\text{prediction}} = 1 - \frac{PRESS}{SS_T}$$

where SS_T is the total sum of squares. This statistic gives some indication of the predictive capability of the regression model. For example, if $R^2 = 0.89$, then it indicates that the model is expected to explain about 89% of the variability in predicting new observations.

Detection and treatment of outliers

- An outlier is an extreme observation.
- Residuals that are considerably larger in absolute value than the others say, 3 or 4 times of standard deviation from the mean indicate potential outliers in y-space. This idea is derived from the 3-sigma or 4-sigma limits.
- Depending on their location, outliers can have moderate to severe effects on the regression model.
- Outliers may indicate a model failure for these points.
- Residual plots against $\hat{y}_i$ and normal probability plots help in identifying outliers. Examination of scaled residuals, e.g., studentized and *R*-student residuals are more helpful as they have mean zero and variance one.
- Outliers can also occur in explanatory variables in X-space. They can also affect the regression results.
- Sometimes outliers are "bad" values occurring as a result of unusual but explainable events. For example, faulty measurements, incorrect recording of data, failure of measuring instrument etc.
- Bad values need to be discarded but should have strong nonstatistical evidence that the outlier is a bad value before it is discarded. Discarding bad values is desirable because least-squares pull the fitted equation toward the outlier.
- Sometimes outlier is an unusual but perfectly plausible observation. If such observations are deleted, then it may give a false impression of improvement in the fit of the equation.
- Sometimes the outlier is more critical than the rest of the data because it may control many key model properties.
- The effect of outliers on the regression model may be checked by dropping these points and refitting the regression equation.
- The value of t-statistic, F-statistic, R^2 and residual mean square may be sensitive to outliers.

An outlier test based on *R*-student

A common way to model an outlier is the mean shift outlier model.

Suppose we fit a model

$$y = X\beta + \varepsilon$$

when the true model is

$$y = X\beta + \delta + \varepsilon$$

where δ is a $n \times 1$ vector of zeros except for the u^{th} observation which has a value δ_u. Thus

$$\delta = (0, 0, ..., 0, \delta_u, 0, ..., 0)$$

Assume $\varepsilon \sim N(0, \sigma^2 I)$ for both the models we fit. Our objective is to find an appropriate statistic for testing $H_0 : \delta_u = 0$ verus $H_0 : \delta_u \neq 0$. This procedure assumes that we are specifically interested is u^{th} observation, i.e., that we have a priori information that the u^{th} observation may be an outlier.

First, we find an appropriate estimate of δ_u. Consider u^{th} residual as its estimate. The $n \times 1$ residual vector is

$$e = \left[I - H\right] y = \left[I - X(X'X)^{-1} X'\right] y.$$

Then

$$\begin{aligned}
E(e) &= \bar{H} y \\
&= \bar{H} E(y) \\
&= \bar{H}(X\beta + \delta) \\
&= \bar{H} X\beta + \bar{H}\delta \\
&= 0 + \left[I - H\right]\delta \\
&= \left[I - X(X'X)^{-1} X\right]\delta.
\end{aligned}$$

Thus $E(e_u) = (1 - h_{uu})\delta_u$

$$\Rightarrow \hat{\delta}_u = \frac{e_u}{1 - h_{uu}}$$

is an unbiased estimator of δ_u where h_{uu} is the u^{th} diagonal element of H.

It may be observed that $\hat{\delta}_u$ is simply the u^{th} PRESS residual. Further, the covariance matrix of e is

$$\begin{aligned}
V(e) &= V\left[(I - H)y\right] \\
&= (I - H)V(y)(I - H) \\
&= \sigma^2 (I - H)
\end{aligned}$$

$$Var(e_u) = (1 - h_{uu})\sigma^2.$$

So

$$\begin{aligned}Var(\hat{\delta}_u) &= Var\left(\frac{e_u}{1-h_{uu}}\right)\\ &= \frac{(1-h_{uu})\sigma^2}{(1-h_{uu})^2}\\ &= \frac{\sigma^2}{1-h_{uu}}.\end{aligned}$$

Also e is a linear combination of normally distributed y. So e is also normally distributed. Thus $\hat{\delta}_u$ is also normally distributed.

Consequently, under $H_0 : \delta_u = 0$,

$$\frac{\left(\frac{e_u}{1-h_{uu}}\right)}{\left(\frac{\sigma}{\sqrt{1-h_{uu}}}\right)} = \frac{e_u}{\sigma\sqrt{1-h_{uu}}} \sim N(0,1).$$

The quantity $\frac{e_u}{\sigma\sqrt{1-h_{uu}}}$ is simply an example of studentized residual. Since σ^2 is unknown and $\frac{MS_{res}}{\sigma^2}$ is a Chi-square random variable, so a candidate test statistic is

$$\frac{e_u}{\sqrt{MS_{res}(1-h_{uu})}}$$

which follows a *t*-distribution if $e = \left[I-H\right]y$ and $SS_{res} = y'(I-H)y$ are independent. Since

$$\left[I-H\right]\sigma^2 I\left[I-H\right] = \sigma^2(I-H) \neq 0,$$

so e and SS_{res} are not actually independent.

We already have developed $S_{(1)}^2$ which is related to the residual mean square in a regression model with i^{th} observation withheld given by

$$S_{(i)}^2 = \frac{(n-k)MS_{res} - \frac{e_i^2}{1-h_{ii}}}{n-k-1}.$$

This estimate of σ^2 is independent of e_u by the basic independence assumption on random errors. So σ^2 can be replaced by $s_{(u)}^2$ and an appropriate test statistic for the mean shift outlier model is

$$\frac{e_u}{s_{(u)}\sqrt{1-h_{uu}}}$$

which is the externally studentized residual or R-student.

Under $H_0 : \delta_u = 0$, $\quad \dfrac{e_u}{s_{(u)}\sqrt{1-h_{uu}}} \sim t(n-k-1)$

and under $H_0 : \delta_u \neq 0$, $\quad \dfrac{e_u}{s_{(u)}\sqrt{1-h_{uu}}} \sim$ noncentral $t\left[(n-k-1, \gamma\right]$

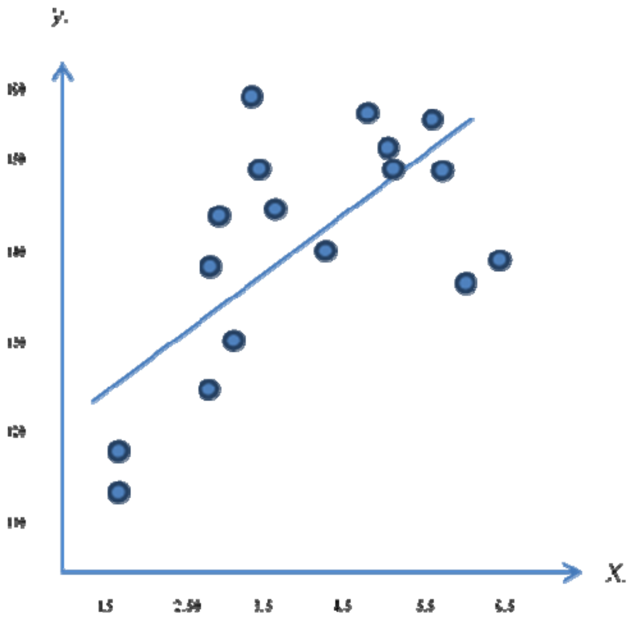
with noncentrality parameter

$$\gamma = \frac{\delta_u}{\sigma / (\sqrt{1-h_{uu}})} = \frac{\delta_i \sqrt{1-h_{uu}}}{\sigma}.$$

Note that the power of this test depends on h_{uu}. If we fit an intercept to our model, then $\frac{1}{n} \leq h_{uu} \leq 1$. So maximum power occurs when $h_{uu} = \frac{1}{n}$, i.e., at the center of the data cloud is terms of the X's. As $h_{uu} \to 1$, the power goes to 0. In other words, this test has less ability to detect outliers at the high leverage data points (Note that the concept of leverage point is discussed in later sections).

Test for lack of fit of a regression model

This test for lack of fit of a regression model is based on the assumptions of normality, independence and constant variance which are satisfied. Only the first order or straight-line character of the relationship is in doubt. For example, the data in the following scatter plot where the indication is there that straight line fit is not very satisfactory.



The test procedure determines if there is systematic curvature is present. The test requires to replicate observations on y for at least one level of x, and they should be true replications and not just the duplicate readings or measurement of y.

The true replications consist of running n_i separate experiments at $x = x_i$ and observe y. It is not just running a single experiment at $x = x_i$ and measuring y n_i times in which the information only on the variability of the method of measuring y is obtained. These replicated observations are used to get a model-independent estimate of σ^2.

Suppose we have n_i observations on y at the i^{th} level of $x_i, i = 1, 2, ..., m.$ Let y_{ij} be the j^{th} observation on y at x_i, $i = 1, 2, ..., m, j = 1, 2, ..., n_i$; $n = \sum_{i=1}^{m} n_i$ is the total number of observations.

Consider the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

Let $\bar{y}_i$ be the mean of n_i observations on x_i. Then the $(i, j)^{th}$ residual is

$$(y_{ij} - \hat{y}_i) = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

$$\begin{tikzcd}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 \arrow[d] & = & \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 \arrow[d] & + & \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2 \arrow[d] & \text{(obtained by squaring and summing over } i \text{ and } j\text{)} \\
SS_{res} \arrow[d] & = & SS_{PE} \arrow[d] & + & SS_{LOF} \arrow[d] & \\
\boxed{\text{Residual sum of squares}} & & \boxed{\text{Sum of squares due to pure error}} \arrow[d] & & \boxed{\text{Sum of squares due to lack of fit}} \arrow[d] & \\
& & \boxed{\text{Measures pure error}} & & \boxed{\text{Measures lack of fit}} &
\end{tikzcd}$$

If assumption of constant variance is satisfied, then SS_{PE} is a **model independent measure of pure error** because only the variability of $y's$ at each x level is used to compute SS_{PE}.

Since there are $(n_i - 1)$ degrees of freedom for pure error at each level of x_i, the number of degrees of freedom associated with SS_{PE} is $\sum_{i=1}^{m}(n_i - 1) = n - m$. SS_{LOF} is a weighted sum of squared deviations between $\bar{y}_i$ at each level of x and corresponding fitted value.

If $\hat{y}_i$ are close to $\bar{y}_i$, then there is a strong indication that the regression function is linear.
If $\hat{y}_i$ deviate considerably from $\bar{y}_i$ then it is likely that the regression function is not linear. The degrees of freedom associated with SS_{LOF} is $m-2$ because there are m levels of x and two degrees of freedom are lost because two parameters must be estimated to obtain $\bar{y}_i$.

Computationally,

$$SS_{LOF} = SS_{res} - SS_{PE}.$$

The test statistic for lack of fit is

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)}$$
$$= \frac{MS_{LOF}}{MS_{PE}}$$

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^{n} n_i\left[E(y_i) - \beta_0 - \beta_1 x_i\right]^2}{(m-2)}.$$

If true regression is linear, then $E(y_i) = \beta_0 + \beta_1 x_i$ and $E(MS_{LOF}) = \sigma^2$.
If true regression is nonlinear, then $E(y_i) \neq \beta_0 + \beta_1 x_i$ and $E(MS_{LOF}) > \sigma^2$.
If true regression function is linear, then

$$F_0 \sim F(m-2, n-m).$$

So to test for lack of fit, compute F_0 and conclude that regression function is not linear if $F_0 > F_\alpha(m-2, n-m)$ at α level of significance.

If we conclude that regression function is not linear then the tentative model must be abandoned and we attempt to find a more appropriate model.

If $F_0 < F_\alpha(m-2, n-m)$ then there is no strong evidence of lack of fit. They MS_{PE} and MS_{LOF} are often combined to estimate σ^2.

If F ratio for lack of fit is not significant and $H_0 : \beta_1 = 0$ is rejected, then this does not guarantee that model will be satisfactory for prediction. It is suggested that the F-ratio must be at least four or five times the $F_\alpha(m-2, n-m)$ if the regression model is to be useful for prediction.

A simple measure of potential prediction performance is found by comparing the range of fitted values, i.e., $\left(\hat{y}_{\max} - \hat{y}_{\min}\right)$ to their average standard error. Regardless of the term of the model, the average variance of the fitted values is

$$\overline{Var(\hat{y})} = \frac{1}{n}\sum_{i=1}^{n} Var(\hat{y}_i) = \frac{k\sigma^2}{n}$$

where k is the number of parameters is the model.

In general, the model is not likely to be satisfactory predictor unless the range of $\hat{y}_i$ is large relative to the estimated standard error $\sqrt{\frac{k\hat{\sigma}^2}{n}}$ where $\hat{\sigma}^2$ is a model-independent estimate of error variance.

Estimation of pure error from near-neighbours:

In test of lack of fit

$$SS_{res} = SS_{PE} + SS_{LOF}$$

SS_{PE} is computed using responses at repeat observations at some level of x. This is a model-independent estimate of σ^2.

This general principle can be applied to any regression model.

Calculation of SS_{PE} requires repeat observations on the response y at the same set of levels on the explanatory variables $x_1, x_2 ..., x_{k,}$, i.e., some of the rows of X-matrix must be same.

In practice, repeat observations do not often occur in multiple regression, and the procedure of lack of fit is not often useful.

A method to obtain a model-independent estimate of error when there are no exact repeat points are the procedures which search for those points in x-space that are near-neighbours.

This is the sets of observations that have been taken with near-identical levels of $x_1, x_2, ..., x_k$. The response y_i from such near-neighbours can be considered as repeat points and used to obtain an estimate of pure error.

As a measure of the distance between any two points, $x_1, x_2, ..., x_k$ and $x_{i'1}, x_{i'2}, ..., x_{i'k}$, use the weighted sum of squared distance (WSSD)

$$D_{ii'}^2 = \sum_{j=1}^{k} \left[\frac{\hat{\beta}_j (x_{ij} - x_{i'j})}{\sqrt{MS_{res}}} \right]^2$$

The pairs of points with small values of $D_{ii'}^2$ are "near neighbours" , i.e., they are relatively close together in x-space. Pairs of points for which $D_{ii'}^2$ is large (e.g., $D_{ii'}^2 >> 1$) are widely separated is x-space. The residuals at two points with a small value of $D_{ii'}^2$ can be used to obtain an estimate of pure error.

The estimate is obtained from the range of residuals at the points i and i', say

$$E_i = |e_i - e_{i'}|.$$

There is a relationship between the range of a sample from a normal population and the population standard deviation. For example, for sample size = 2, this relationship is

$$\sigma^2 \equiv \frac{E}{1.128} = 0.886E.$$

The quantity σ^2 so obtained is an estimate of the standard deviation of pure error.

An efficient algorithm may be used to compute this estimate as as follows:

- First arrange the data points $x_{ii}, ..., x_{ik}$ in order of increasing $\hat{y}_i$.
- Note that points with different values of $\hat{y}_i$ cannot be near neighbour but those with similar values of $\hat{y}_i$ could be neighbours (or they could be near the same contour of constant $\hat{y}$ but for apart in some x-coordinates).

Then

1. Compute the values of $D_{ii'}^2$ for all $(n-1)$ pairs of points with adjacent values of $\hat{y}$. Repeat this calculation for the pairs of points separated by one, two and three intermediate $\hat{y}$ values. This will produce $(4n-10)$ values of $D_{ii'}^2$.
2. Arrange the $(4n-10)$ values of $D_{ii'}^2$ found is step 1. Let $E_u, u = 1, 2, ..., (4n-10)$ be the range of the residuals at these points.
3. For the first m values of E_u, calculate an estimate of the standard deviation of pure error as

$$\hat{\sigma} = \frac{0.886}{m} \sum_{u=1}^{m} E_u.$$

Note that $\hat{\sigma}$ is based on the average range of the residuals associated with the m smallest values of $D_{ii'}^2, m$ must be chosen after inspecting the values of $D_{ii'}^2$. One should not include values of E_u is the calculation for which the weighted sum of squared distance is too large.