LINEAR REGRESSION ANALYSIS

MODULE – IV

Lecture - 17

Model Adequacy Checking

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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 much but gross nonnormality is more serious.

The normal probability plots help in verifying the assumption of normal distribution.

If the 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 **ordered standardized residuals** versus **normal scores**. The normal scores are the cumulative probability defined as

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

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

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

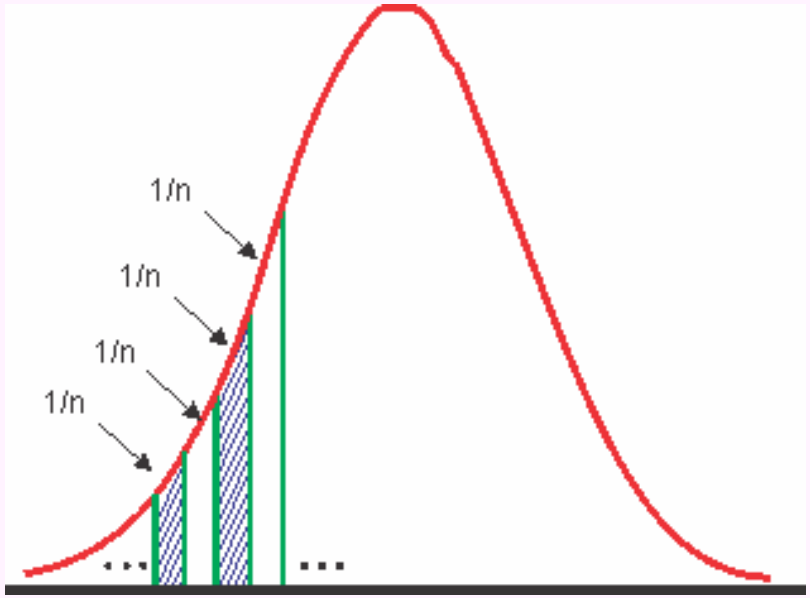
then the $e_{[i]}$'s are plotted against P_i and the plot is called normal probability plot.

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 approximately on 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 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.
- First section has one point, so cumulative probability is $P_i = 1/n$. Second section has one point, so cumulative probability upto 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 1/2 is used for end correction as all the observations which are scattered inside the stripe are assumed to be concentrated at the mid point of the stripe.



Different software use 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 important in real use.

The straight line is usually determined visually with emphasis on the central values rather than the extremes. 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.

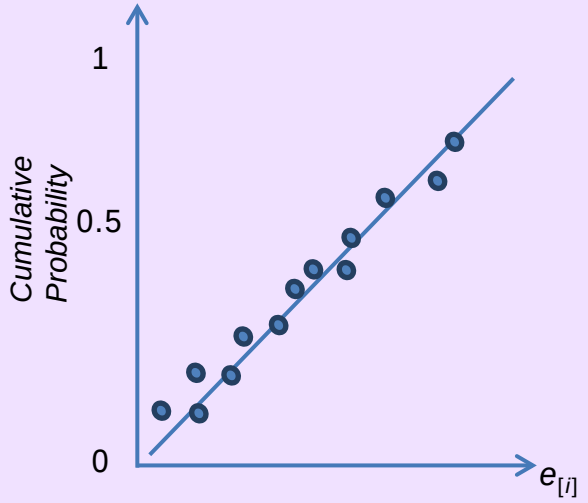
This follows from the fact that

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

Various interpretations to the graphic patterns is as follows.

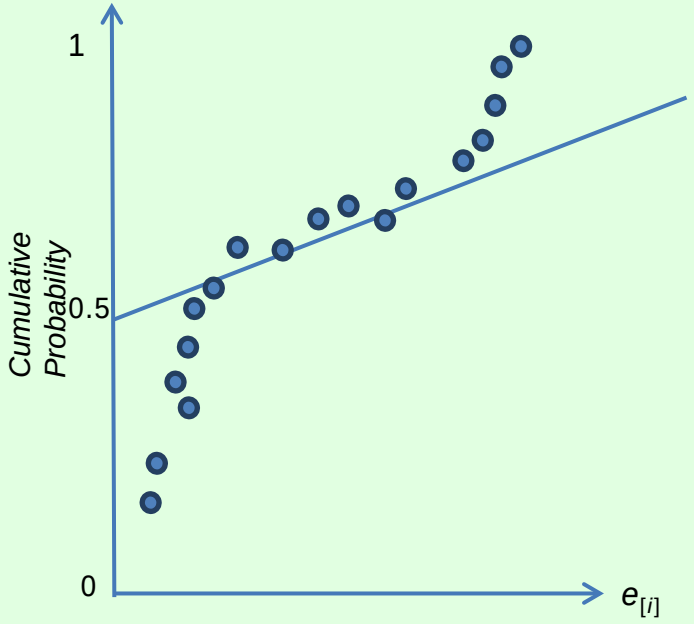
a) This is an ideal normal probability plot.

The points lie approximately on the straight line and indicate that the underlying distribution is normal.

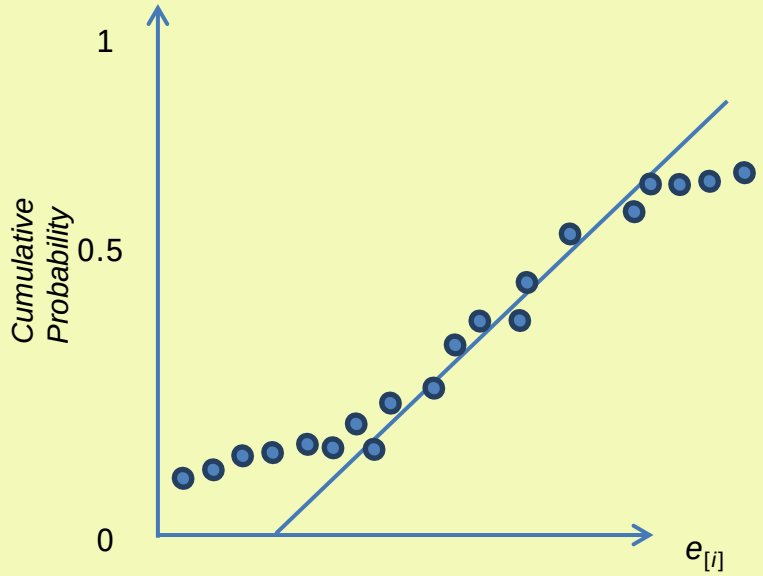


b) The figure has sharp upward and downward curves at both extremes.

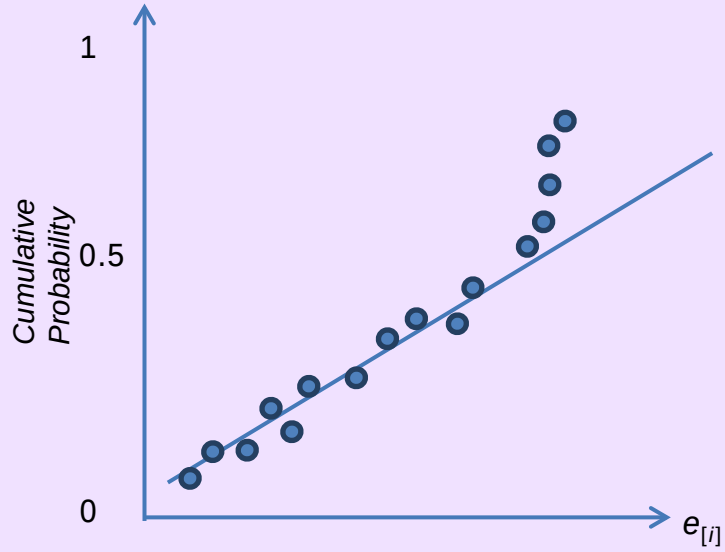
This indicates that the underlying distribution is heavy tailed, i.e., the tails of underlying distribution are thicker than the tails of normal distribution.



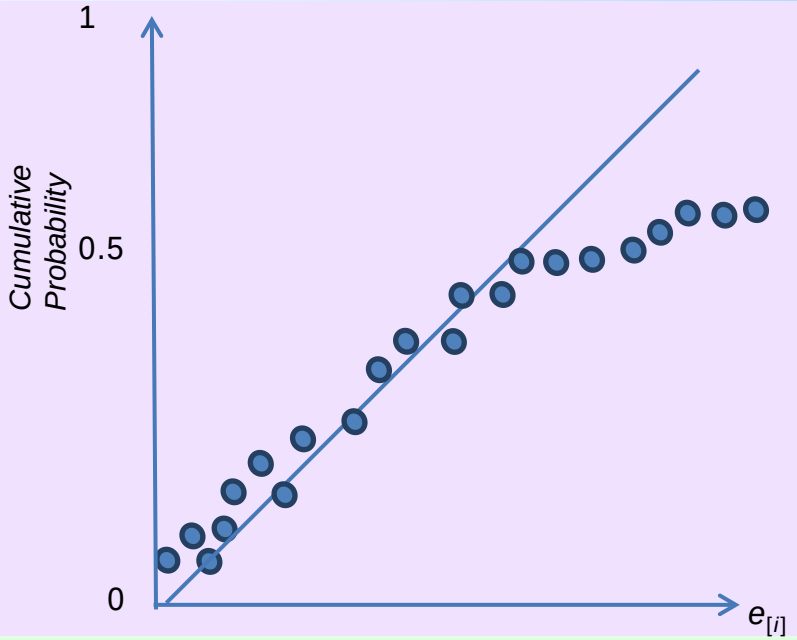
c) The figure has flattening at the extremes for the curves.
This indicates that the underlying distribution is light tailed, i.e., the tails of the underlying distribution are thinner than the tails of normal distribution.



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



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



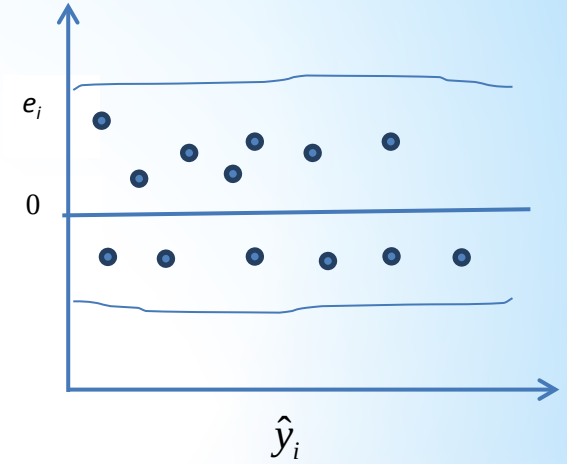
Some experience and expertise is required to interpret the normal probability plots because the samples taken from a normal distribution will not plot exactly as a straight line.

- Small sample sizes $(n \le 16)$ often produce normal probability plots that deviate substantially from linearity.
- Larger sample sizes $(n \ge 32)$ produces 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 the 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 and consequently, we can not rely on the normal probability plots to detect the departures from normality.
- Commonly seen defect found in normal probability plots is the occurrence of one or two large residuals. Sometimes, this is an indication that the corresponding observations are outliers.

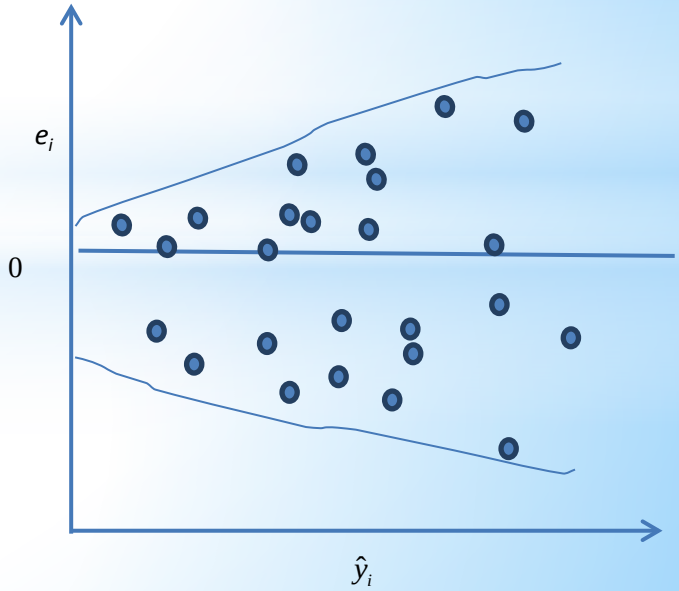
Plots of residuals against the fitted value

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

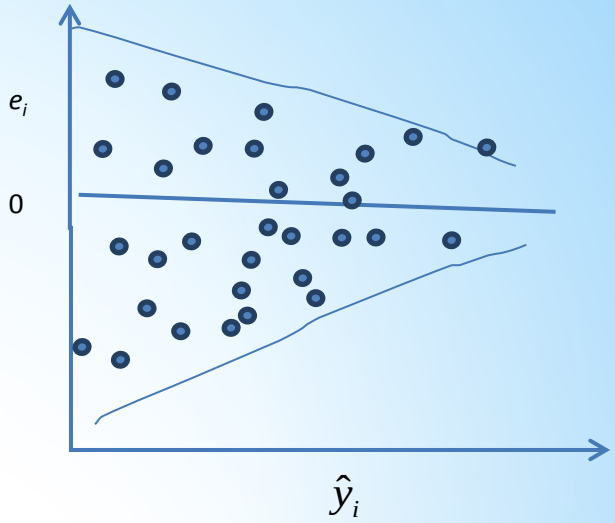
(a) If plot is such that the residuals can be contained in a **horizontal band** (and residual fluctuates is more or less in a random fashion inside the band) , then there are no obvious model defects.



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

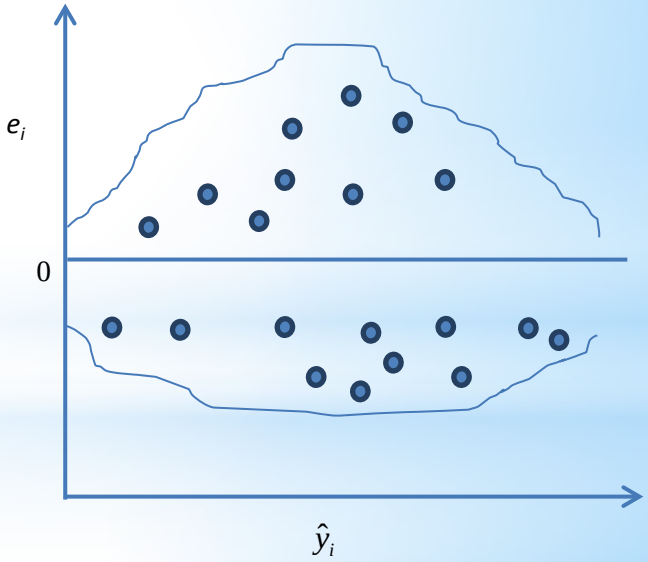


(c) If plots is such that the residuals can be accommodated in an inward opening funnel, then such pattern indicates that the variance of errors is not constant but it is a decreasing function of *y*.

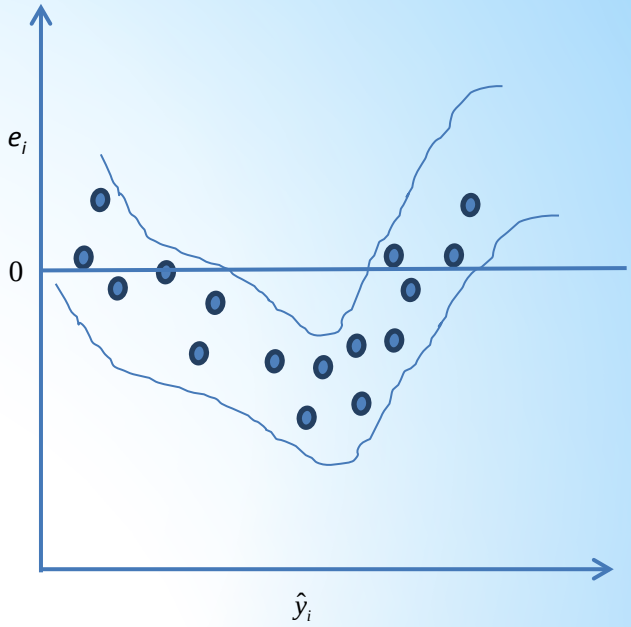


(d) If plot is such that the residuals can be accommodated inside a double bow, then such pattern indicates that the variance of errors is not constant but *y* is a proportion between 0 and 1. The *y* then may have a Binomial distribution. The variance of a Binomial proportion near the value 0.5 is greater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

Usual approach to deal with such inequality of variances is to apply a suitable transformation to either the explanatory variables, 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 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 is 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 residuals on *Y* - axis and values of 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 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 fluctuates more or less in a random fashion within this band, then it is desirable and there are no obvious model defects.

- an outward opening funnel shape or inward opening funnel shape indicates that the variance is nonconstant.

- a double bow pattern or nonlinear pattern indicates the assumed relationship between *y* and X_j is not correct. The possibilities like *y* may be a proportion, higher ordered term is X_j (e.g. X_j^2) are needed or a transformation is needed are 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 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 in the model. 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.