LINEAR REGRESSION ANALYSIS

MODULE – IV

Lecture - 18

Model Adequacy Checking

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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 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.

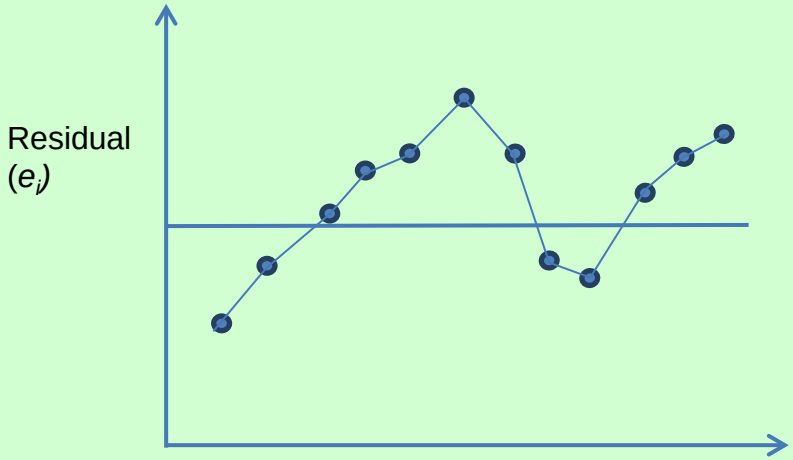
- Outward opening or inward opening funnel indicates that the variance is not constant but changing with time.

- Double bow pattern or nonlinear pattern 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.**

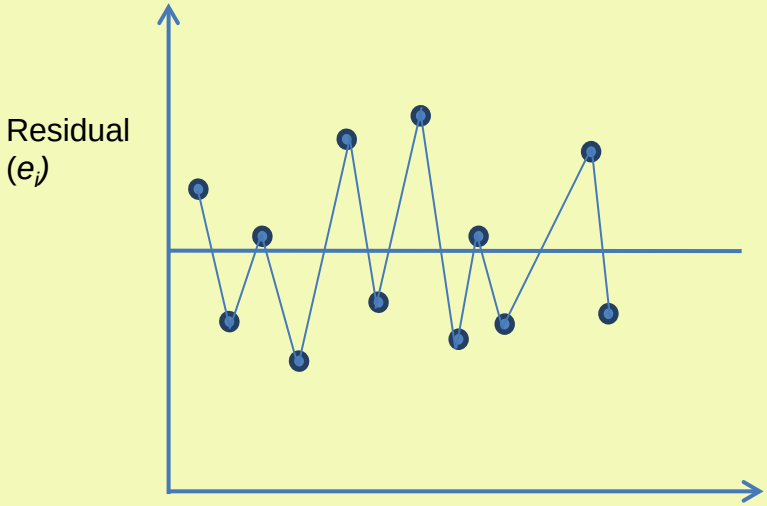
If we have a plot like following, then it indicates the presence of autocorrelation.

Following type of figure indicates the presence of positive autocorrelation:



Presence of positive autocorrelation

Following type of figure indicates the presence of negative autocorrelation:



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 in helpful in evaluating whether the relationship between study and explanatory variables is correctly specified.

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

These plots provide the 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 slope.

See how:

Consider the model $y = X\beta + \varepsilon$

$= X_{(j)}\beta_{(j)} + X_j\beta_j + \varepsilon$

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

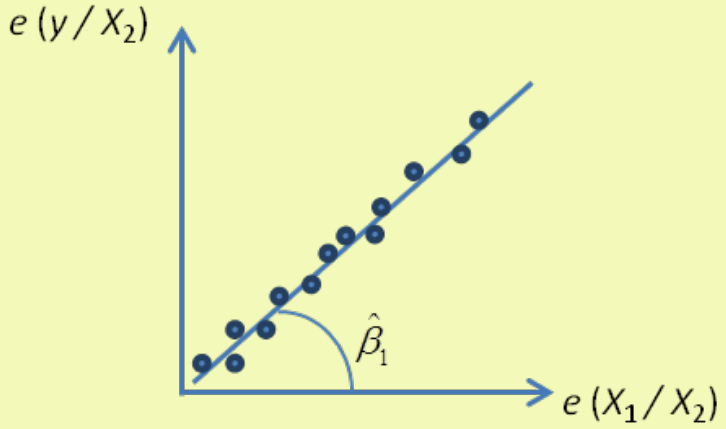
$$(I - H_{(j)})y = (I - H_{(j)})X_{(j)}\beta + \beta_j(I - H_{(j)})X_j + (I - H_{(j)})\varepsilon$$

$$= 0 + \beta_j(I - H_{(j)})X_j + (I - H_{(j)})\varepsilon$$

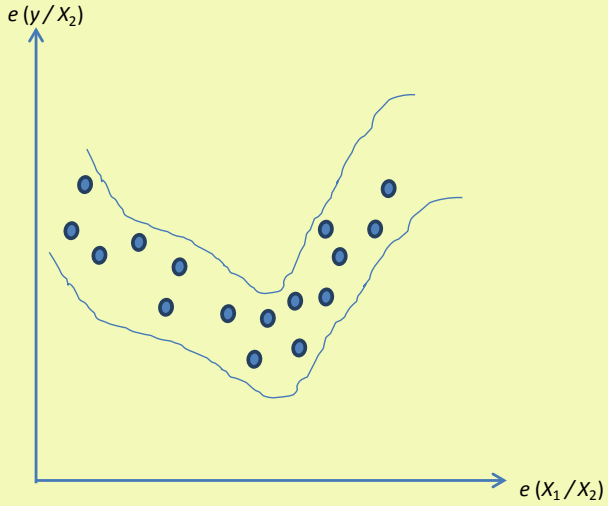
$$e\left[y / X_{(j)}\right] = \beta_j\, e\left[X_j / X_{(j)}\right] + \varepsilon^*$$

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

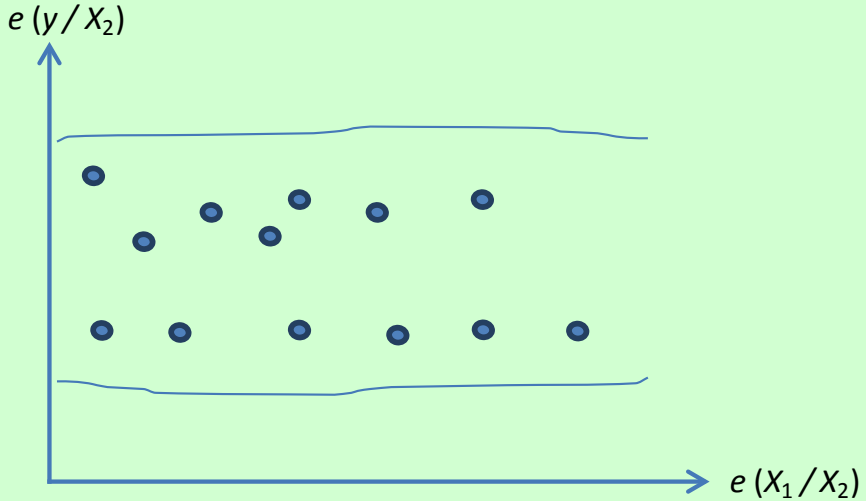
This suggests that a partial regression plot which is a plot between $e\left[y / X_{(j)}\right]$ and $e\left[X_j / X_{(j)}\right]$ (like between *y* and *X*) should have slope β_j. Thus if X_j enters the regression in a linear fashion, the partial regression plot should show linear relationship passing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

$$\hat{y}_i(X_2) = \hat{\theta}_0 + \hat{\theta}_1 x_{i2}$$

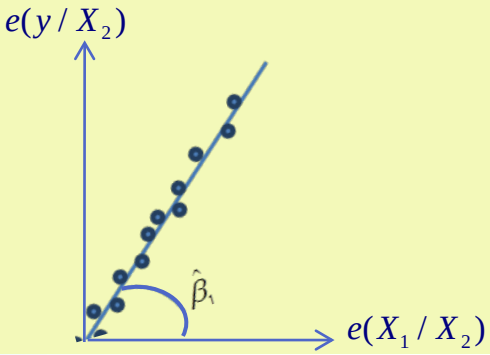
$$e_i(y / X_2) = y_i - \hat{y}_i(X_2),\ i = 1, 2, ..., n.$$

- Regress X_1 on X_2 and find the residuals

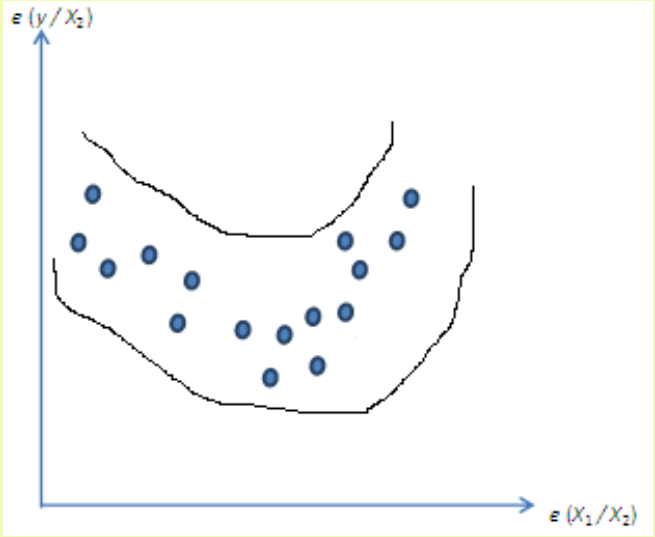
$$\hat{X}_{i1}(X_2) = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}$$

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

- 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:

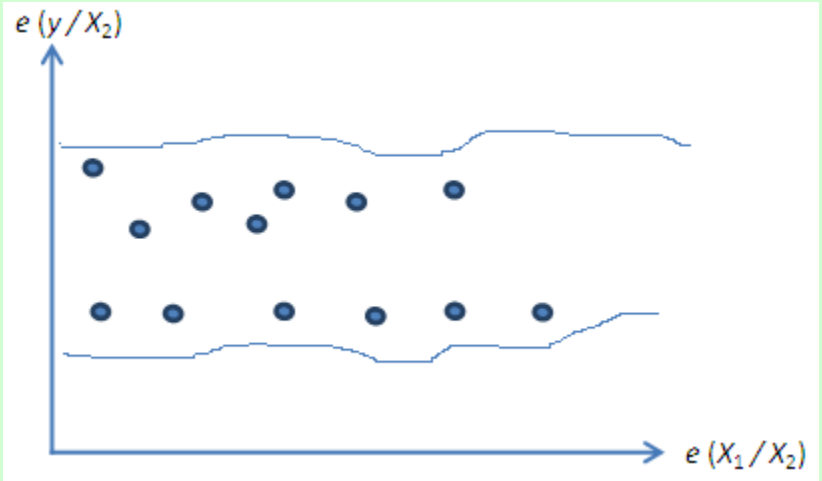


- The slope of this line is the regression coefficient of X_1 in the multiple linear regression model.



- 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 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 but not correctly 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 interaction effect among the regressors.

3. Partial regression plots are affected by the existence of exact relationship among explanatory variables (termed as 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 regression

A residual plot closely related to the partial regression plot is 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 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

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

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

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

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

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

When $e_i^*(y / X_j)$ are plotted against x_{ij}, the resulting display has slope $\hat{\beta}_j$. The interpretation of 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 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 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 “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 fit of equation.
- Sometimes the outlier is more important 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.