LINEAR REGRESSION ANALYSIS

MODULE – IV

Lecture - 15

Model Adequacy Checking

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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

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

The validity of these assumption 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 and conclusions may not be affected much.

But if the departures are large, the model obtained may become unstable in the sense that a different sample could lead to an entirely different model with different 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 important 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 take a decision about the population on the basis of a sample of data.

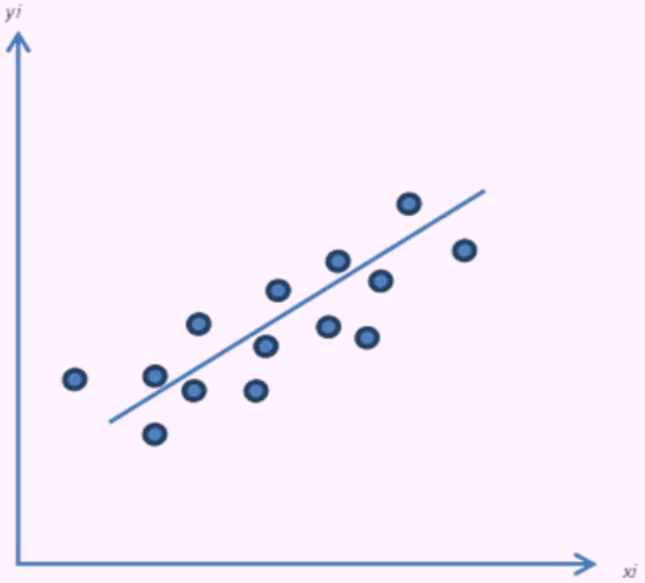
Several diagnostic methods to check the violation of regression assumption are based on the study of model residuals with the help of various types of graphics.

Checking of linear relationship between study and explanatory variables

Case of one explanatory variable

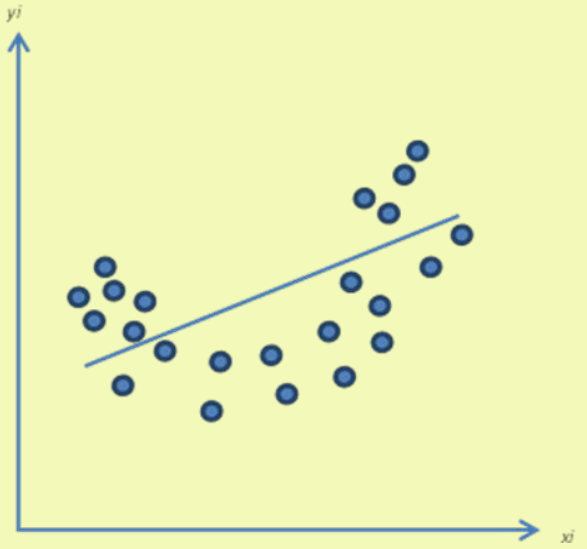
If there is only one explanatory variable in the model, then it is easy to check the existence of linear relationship between y and X by scatter diagram of the available data.

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



Linear trend

The following figure indicates a nonlinear trend:



Nonlinear trend

2. Case of more than one explanatory variables

To check the assumption of linearity between study variable and explanatory variables, the **scatter plot matrix** of the data can be used.

A scatter plot matrix is a two dimensional array of two dimension plots where each form contains a scatter diagram except for the diagonal.
Thus, each plot 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 gives a sense of linearity or nonlinearity of the relationship and some awareness of how the individual data points are arranged over the region.

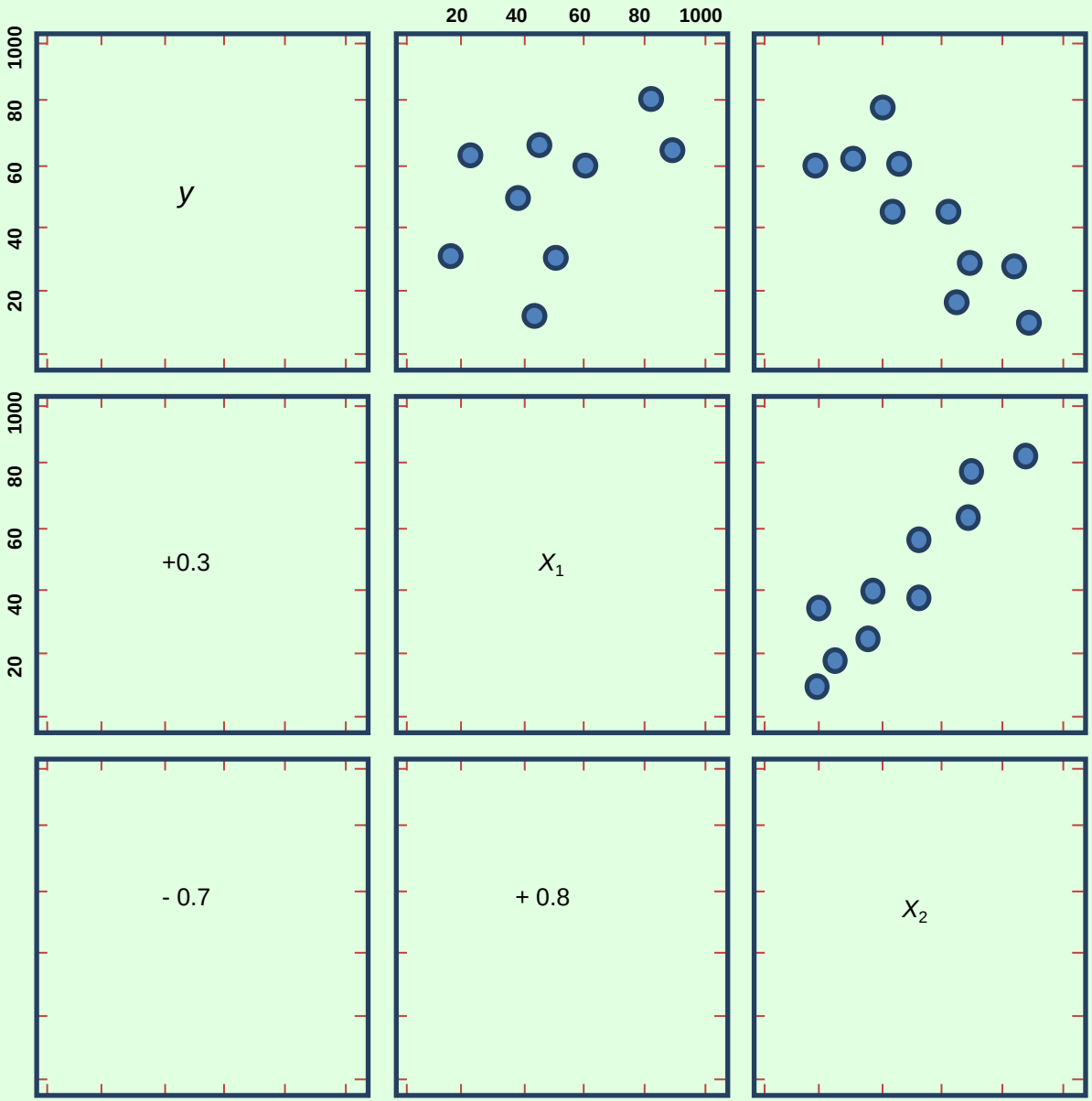
It is a scatter diagram of (y versus X_1), (y versus X_2),…, (y versus X_k).

Another option to present the

-- scatter plots in the upper triangular part of 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 scatter plot matrix looks like as follows:



Scatter plot matrix for two variables case

Such 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., its value can be substantially influenced by one or two observations in the data.

The presence of linear patterns is reassuring but absence of such patterns does not imply that linear model is incorrect.

Most of the statistical software provide the option for creating the scatter plot matrix. The view of all the plots provides an indication that a multiple linear regression model may provide a reasonable fit to the data.

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

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

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

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

We consider it as $e_i = y_i - \hat{y}_i,\ i = 1, 2, ..., n$

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 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. Analysis of residual helps is finding the model inadequacies.

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

- Residuals have zero mean as

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

- 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 methods 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 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

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

$\Rightarrow MS_{res}$ overestimates the $Var(e_i)$.

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