LINEAR REGRESSION ANALYSIS

MODULE – XII

Lecture - 36

Polynomial Regression Models

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Test of significance

To test the significance of highest order term, we test the null hypothesis

$$H_0 : \alpha_k = 0.$$

This hypothesis is equivalent to $H_0 : \beta_k = 0$ in polynomial regression model.

We would use

$$F_0 = \frac{SS_{reg}(\alpha_k)}{SS_{res}(k)/(n-k-1)}$$

$$= \frac{\hat{\alpha}_k \sum_{i=1}^{n} P_k(x_i) y_i}{SS_{res}(k)/(n-k-1)}$$

$$\sim F(1, n-k+1) \text{ under } H_0.$$

If order of the model is changed to $(k+r)$, we need to compute only r new coefficients. The remaining coefficients $\hat{\alpha}_0, \hat{\alpha}_1, ..., \hat{\alpha}_k$ do not change due to orthogonality property of polynomials. Thus the sequential fitting of the model is computationally easy.

When X_i are equally spaced, the tables of orthogonal polynomials are available and the orthogonal polynomials can be easily constructed.

First 7 orthogonal polynomials are as follows:

Let d be the spacing between levels of x and $\{\lambda_j\}$ be the constants chosen so that polynomials will have integer values. The tables are available.

$$P_0(x_i) = 1$$

$$P_1(x_i) = \lambda_1 \left[\frac{x_i - \bar{x}}{d} \right]$$

$$P_2(x_i) = \lambda_2 \left[\left(\frac{(x_i - \bar{x})}{d} \right)^2 - \left(\frac{n^2 - 1}{12} \right) \right]$$

$$P_3(x_i) = \lambda_3 \left[\left(\frac{(x_i - \bar{x})}{d} \right)^3 - \left(\frac{x_i - \bar{x}}{d} \right) \left(\frac{3n^2 - 7}{20} \right) \right]$$

$$P_4(x_i) = \lambda_4 \left[\left(\frac{(x_i - \bar{x})}{d} \right)^4 - \left(\frac{x_i - \bar{x}}{d} \right)^2 \left(\frac{3n^2 - 13}{14} \right) + \frac{3(n^2 - 1)(n^2 - 9)}{560} \right]$$

$$P_5(x_i) = \lambda_5 \left[\left(\frac{(x_i - \bar{x})}{d} \right)^5 - \frac{5}{18} \left(n^2 - 7 \right) \left(\frac{x_i - \bar{x}}{d} \right)^3 + \frac{1}{1008} \left(15n^4 - 230n^2 + 407 \right) \left(\frac{x_i - \bar{x}}{d} \right) \right]$$

$$P_6(x_i) = \lambda_6 \left[\left(\frac{(x_i - \bar{x})}{d} \right)^6 - \frac{5}{44} \left(3n^2 - 31 \right) \left(\frac{x_i - \bar{x}}{d} \right)^4 + \frac{1}{176} \left(5n^4 - 110n^2 + 329 \right) \right) \left(\frac{x_i - \bar{x}}{d} \right)^2$$

$$- \frac{5}{14784} (n^2 - 1)(n^2 - 9)(n^2 - 25) \Bigg]$$

An example of the table for $n = 5$ is as follows:

x_i	P_1	P_2	P_3	P_4
1	-2	2	-1	1
2	-1	-1	- 2	-4
⋮	⋮	⋮	⋮	⋮
5				
$\sum_{i=1}^{n}\{P_j(x_i)\}^2$	10	14	10	70
λ	1	1	$\frac{5}{6}$	$\frac{35}{12}$

The orthogonal polynomials can also be constructed when *x's* are not equally spaced.

Piecewise polynomial (Splines)

Sometimes it is exhibited in the data that a lower order polynomial does not provide a good fit. A possible solution in such situation is to increase the order of the polynomial but it may always not work. The higher order polynomial may not improve the fit significantly. Such situations can be analyzed through residuals, e.g., the residual sum of square may not stabilize or the residual plots fail to explain the unexplained structure. One possible reason for such happening is that the response function has different behavior in different ranges of independent variables. This type of problems can be overcome by fitting an appropriate function in different ranges of explanatory variable. So polynomial will be fitted into pieces. The spline function can be used for such fitting of polynomial in pieces.

Splines and knots

The piecewise polynomials are called splines. The joint points of such pieces are called as knots. If polynomial is of order *k*, then the spline is a continuous function with (*k* - 1) continuous derivatives. For this, the function values and first (*k* - 1) derivatives agree at the knots.

Cubic splines:

For example, consider a cubic spline with *h* knots. Suppose the knots are $t_1 < t_2 < ... < t_h$ and cubic spline has continuous first and second derivatives at these knots. This can be expressed as

$$E(y) = S(x) = \sum_{j=0}^{3} \beta_{oj} x^j + \sum_{i=1}^{h} \beta_i (x - t_i)_+^3$$

where

$$(x - t_i)_+ = \begin{cases} x - t_i & \text{if } x - t_i > 0 \\ 0 & \text{if } x - t_i \le 0. \end{cases}$$

It is assumed that the position of knots are known. Under this assumption, this model can be fitted using the usual fitting methods of regression analysis like least squares principal.

In case, the knot positions are unknown, then they can be considered as unknown parameters which can be estimated. But in such situation, the model becomes non-linear and methods of non-linear regression can be used.

Issue of number and position of knots

It is not so simple to know the number and position of knots in a given set of data. It is tried to keep the number of knots as minimum as possible and each segment should have minimum four or five data points. There should not be more than one extreme point and one point of inflexion in each segment. If such points are to be accommodated, then it is suggested to keep the extreme point in the center of segment and point of inflexion near the knots.

It is also possible to fit the polynomials of different orders in each segment and to impose different continuity restrictions at the knots. Suppose it is to be accomplished in a cubic spline model. If all $(h+1)$ pieces of polynomial are cubic, then a cubic spline model without continuity restrictions is

$$E(y) = S(x) = \sum_{j=0}^{3} \beta_{oj} x^j + \sum_{i=1}^{h} \sum_{j=0}^{3} \beta_i (x - t_i)_+^j$$

where

$$(x - t_i)_+^0 = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \le 0. \end{cases}$$

If the term $\beta_{ij}(x - t_i)_+^j$ is in the model, then j^{th} derivative of $S(x)$ at t_i is discontinuous.

If the term $\beta_{ij}(x - t_i)_+^j$ is not in the model, then j^{th} derivative of $S(x)$ is continuous at t_i.

So the model is fitted better when required continuity restrictions are fewer because then more parameters will be included in the model.

If more continuity restrictions are needed, then it indicates that the model is not well fitted but the finally fitted curve will be smoother. The test of hypothesis in multiple regression model can be used to determine the order of polynomial segments and continuity restrictions.

Example

Suppose there is only one knot at t in a cubic spline without continuity restrictions given by

$$E(y) = S(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \beta_{10}(x-t)_+^0 + \beta_{11}(x-t)_+^1 + \beta_{12}(x-t)_+^2 + \beta_{13}(x-t)_+^3 .$$

The term involving β_{10}, β_{11} and β_{12} are present in the model, so $S(x)$, its first derivative $S'(x)$ and second derivative $S''(x)$ are not necessarily continuous at t. Next question arises is how to judge the quality of fit. This can be done by test of hypothesis as follows:

$H_0 : \beta_{10} = 0$ tests the continuity of $S(x)$

$H_0 : \beta_{10} = \beta_{11} = 0$ tests the continuity of $S(x)$ and $S'(x)$

$H_0 : \beta_{10} = \beta_{11} = \beta_{12} = 0$ tests the continuity of $S(x), S'(x)$ and $S''(x)$.

The test $H_0 : \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = 0$ indicates that cubic spline fits data better than a single cubic polynomial over the range of explanatory variable *x*.

This approach is not satisfactory if the knots are large in number as this makes *X'X* ill-conditioned. This problem is solved by using cubic *B*- spline which are defined as

$$B_i(x) = \sum_{j=i-4}^{i} \left[\frac{(x-t_j)_+^3}{\prod_{\substack{m=i-4 \\ m \neq j}}^{i} (t_j - t_m)} \right], \quad i = 1, 2, ..., h+4$$

$$E(y) = S(x) = \sum_{i=1}^{h+4} \gamma_i B_i(x)$$

where $\gamma_i's \ (i = 1, 2, ..., h+4)$ are parameters to be estimated. There are eight more knots - $t_{-3} < t_{-2} < t_{-1} < t_0$ and $t_{h+1} < t_{h+2} < t_{h+3} < t_{h+4}$. Choose $t_0 = x_{\min}$, $t_{h+1} = x_{\max}$ and other knots arbitrarily.

Polynomial models in two or more variables

The techniques of fitting of polynomial model in one variable can be extended to fitting of polynomial models in two or more variables.

A second order polynomial is more used in practice and its model is specified by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon.$$

This is also termed as **response surface**. The methodology of response surface is used to fit such models and helps in designing an experiment.