15 Polynomial Regression

Applied regression analysis and other multivariable methods

Yi Zhou

May 16, 2016

Preview

Polynomial models

- Only one basic independent variable to be considered
- The special cases of the general multiple regression model
  - the second-order (prarabola) polynomial models
  - the higher-order polynomial models
  - orthogonal polynomials

Polynomial models

Mathematical model, a polynomial of order k in x:

$$y = c_0 + c_1 x + c_2 x^2 + \cdots + c_k x^k$$

- c's and k are constants
- k=1, the simple polynomial (namely, the straight line)
- k=2, the second-order polynomial (namely, the parabola)

Statistical model, a parabolic model/quadratic model

- $$\mu_{Y|X} = \beta_0 + \beta_1 X + \beta_2 X^2$$
- $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + E$$

Least-square procedure for fitting a prabola

The parameters are chosen so as to minimize the sum of squares of deviations (SSE)

- It's not necessary to present the precise formulas
- Computer program

ANOVA table for second-order polynomial regression

- Variables-added-in-order tests
  - aids in choosing the most parsinomious yet relevant model possible
- Natural variable orderings, either from the largest to the smallest power of the predictor or vice versa
- Variables-added-last test should be avoided with polynomial models

The inference associated with second-order polynomial regression

Basic inferential questions

- Is the overall regression significant?
- Does the second-order model provide significantly more predictive power?
- Is it necessary to add high-order terms?

Test for overall regression and strength of the overall parabolic relationship

H_0: There is no significant overall regression using X and X^2

- The overall F test
- Degree of freedom
- R^2: the proportionate reduction in the errorsum of squares obtained by using X and X^2

Test for the addition of the X^2 term to the model

H_0: The addition of the X^2 term to the straight-line model does not significantly improve the predition of Y over and above that achieved by the straight line model itself

- Partial F test

Testing for accuracy of the second-order model

- Lack-of-fit test

Example requiring a second-order model

- ANOVA table
- Other factors taken into consideration
  - the R^2-value for the parabolic model is very high
  - the incrase of R^2 is not very large
  - the scatter diagram
  - the simpler model is preferable

Fitting and testing higher-order model

How large an order of polynomial model depends on

- the problem being studied and the amount and type of data being collected
- the number of bendsin the polynomial curve
- the quantity of data

Lack-of-fit tests

The classic LOF tests evaluates a model more complex than one under primary consideration

- only if there are replicate observations
- n total observations
- d X's are distinct
- $r = n - d$ replicates

A classic LOF test compares the fit of a polynomial of order $d - 1 = n - r - 1$
ANOVA for the classic LOF test

- $SSE = SS_{PE} + SS_{LOF}$ (pure-error sum of square, LOF sum of square)
- multiple partial F test: $F = \frac{MS_{LOF}}{MS_{PE}}$

Orthogonal polynomial

`Natural polynomial` vs. `orthogonal polynomial`

- Basic motivation for using orthogonal polynomial: to avoid the serious collinearity
- The orthoginal polynomial are pairwise uncorrelated

Two desirable properties:

- the orthogonal polynomial variables contain exactly the same information as the simple polynomial variables
- the orthogonal polynomial variables are uncorrelated with each other

The partial F test of $H_0 : \beta_j^* = 0$ for the orthogonal polynomial model is equivalent to the partial F test of $H_0 : \beta_j = 0$ in the reduced natural polynomial model

Strategies for choosing a polynomial model

Model selection procedures (see Chap 16):

- Forward-selection model-building strategy:
  - can produce misleading results
  - test for the importance of a candidate predictor
  - can lead to underfitting the data
- Backward-elimination strategy:
  - may overfit the data

It is important to interatively conduct the residual analysis

- A plot of jackknife residuals against X
- The need for a higher-order model often appears as a nonliner trend in the residuals