Chapter 12 Polynomial Models and Curvilinear Models

12.1 Polynomial Models

- One predictor variable

Regression model: $Y = f(X) + \varepsilon$.

SLR model: $Y = f(X) + \varepsilon$ if $f(X) = \beta_0 + \beta_1 X$.

What can we do if $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2$ (a quadratic function of X)?

Further, what if the form of $f(X)$ is unknown? Under some conditions on $f(X)$, by Taylor expansion

$$f(X) = f(0) + \frac{f'(0)}{1!}X + \frac{f''(0)}{2!}X^2 + \cdots + \frac{f^{(k)}(0)}{k!}X^k + \frac{f^{(k+1)}(\xi)}{(k+1)!}X^{k+1}$$
$$\approx f(0) + \frac{f'(0)}{1!}X + \frac{f''(0)}{2!}X^2 + \cdots + \frac{f^{(k)}(0)}{k!}X^k \text{ for } |X| \le c < 1$$

Thus, we can fit the model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \varepsilon \text{ for } |X| \le c < 1.$$

Notes: (1) Any smooth function can be approximated by a polynomial.
(2) A linear model implies "linear" in the parameters rather than the predictor variables.

Q: Given data, how to determine the order *k*?
A: Start with a SLR model and then increase the polynomial order one by one until the polynomial after which the adjusted R^2 will decrease.

Ex. 12.1 A research on how the yield *Y* (kg/hectare) of a crop varies with the time between flowering and harvesting *X* (days) results in the following data.

X	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46
Y	2508	2518	3304	3423	3057	3190	3500	3883	3823	3646	3708	3333	3517	3214	3103	2776

(1) Is it adequate to fit the data by a straight line? Is a linear relationship between *Y* and *X* significant at $\alpha = 0.05$?
(2) Does adding the X^2 term to the model significantly improve the model?
(3) Whether is adding the X^3 term to the second order model warranted? Use $\alpha = 0.05$.

- k (≥ 2) predictor variables

A full second-order (quadratic) model for $k = 2$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \varepsilon$$

A full third-order model for k =2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \beta_{111} X_1^3 + \beta_{112} X_1^2 X_2 + \beta_{122} X_1 X_2^2 + \beta_{222} X_2^3 + \varepsilon$$

Note: The coefficient of $X_1^2 X_2$ is labeled as β_{112}.

A full second-order model for k = 3:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \varepsilon$$

Higher-order models for k predictor variables $X_1, \cdots, X_k$ can be obtained similarly.

Q: What is the interpretation of a cross term?

Definition 12.1 (Interaction) If the change in Y as X_i changes by 1 unit also depends on $X_j, i \neq j$, then X_i and X_j are said to *interact*.

Notes: (1) Interactions are given by cross-terms. (2) Can have $X_i^2 X_j$ or $X_i X_j^2$, $i \neq j$.

Ex. Consider the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \varepsilon$.
Let $Y' = \beta_0 + \beta_1 (X_1 + 1) + \beta_2 X_2 + \beta_{11} (X_1 + 1)^2 + \beta_{12} (X_1 + 1) X_2 + \beta_{22} X_2^2 + \varepsilon$. Then,

$$\begin{aligned} Y' - Y &= [\beta_0 + \beta_1 (X_1 + 1) + \beta_2 X_2 + \beta_{11} (X_1 + 1)^2 + \beta_{12} (X_1 + 1) X_2 + \beta_{22} X_2^2 + \varepsilon] \\ &\quad - [\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \varepsilon] \\ &= \beta_1 + \beta_{11} (2X_1 + 1) + \beta_{12} X_2 \end{aligned}$$

— depends not only on X_1 but also on X_2 as long as $\beta_{12} \neq 0$, and thus X_1 and X_2 interact.

- Fitting an appropriate polynomial model

To fit an appropriate first-order linear regression model, we have the backward elimination procedure. However, this procedure cannot be used for polynomial models because of the special relationship among some terms in the model.

For polynomial models, we have an important criterion.

Origin Shift Criterion: A reduced model is considered *sensible* or *well formulated* if any shift in predictor variables ($X_i = Z_i + a_i, i = 1, \cdots, k$) produces a model of unchanged form in the new variables $Z_i, Z_2, \cdots, Z_k$.

Ex. 12.2 (1) Is the model $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_{111} X^3 + \varepsilon$ well formulated?
(2) Is the model $Y = \beta_0 + \beta_1 X_1 + \beta_{12} X_1 X_2 + \varepsilon$ well formulated?

Rule: According to the origin shift criterion, only the highest-order terms can be deleted at first. If a higher-order term is in the model, all lower-order terms related to it must also be in the model, no matter whether they are significant or not.

Note: (1) According to the rule, β_0 cannot be eliminated in any circumstances.
(2) An appropriate polynomial model can be found by the backward elimination procedure with the rule considered.

Ex. 12.3 For the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \varepsilon$, we can eliminate any second-order term.
(i) If $\beta_{11} \neq 0$, $\beta_{12} = \beta_{22} = 0$, X_1 must be in the model. However, X_2 can be eliminated.
(ii) If $\beta_{22} \neq 0$, $\beta_{11} = \beta_{12} = 0$, X_2 must be in the model. However, X_1 can be eliminated.
(iii) If $\beta_{12} \neq 0$, $\beta_{11} = \beta_{22} = 0$, both X_1 and X_2 must be in the model.

- Important tests under the rule

Under the rule, we can test

(1) Whether some or all highest-order terms can be eliminated simultaneously.
(2) Whether a predictor variable is necessary.
(3) For all other tests, we need to check whether the reduced model is well formulated.

Ex. 12.4 For the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \varepsilon$, we can test
(I) H_0: $\beta_{11} = \beta_{12} = 0$ vs. H_a: at least one of them is not 0.
(II) To see whether X_1 is necessary, test
H_0: $\beta_1 = \beta_{11} = \beta_{12} = 0$ vs. H_a: at least one of them is not 0.

Q: Are the reduced models well formulated? Check them by yourself!

Ex. 12.5 For the steam plant data, we have seen that an appropriate model is $Y = \beta_0 + \beta_1 X_1 + \beta_7 X + \varepsilon$.

(1) Test to determine whether it is necessary to add the second-order terms to the model. Use $\alpha =$ 0.05.

(2) Fit an appropriate model for Y on X_1 and X_7 with the second-order terms considered. Use $\alpha =$ 0.05.

12.5 Curvilinear Models

- Transformations of the predictor variables (Curvilinear regression)

More generally, we can have the models with functions of the predictor variables, called *curvilinear* models. Polynomial models are special cases of the curvilinear models.

Original model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$.

General curvilinear model:

$$Y = \alpha_0 + \alpha_1 Z_1(X_1, \cdots, X_{p-1}) + \cdots + \alpha_k Z_k(X_1, \cdots, X_{p-1}) + \varepsilon,$$

where $Z_i(X_1, \cdots, X_{p-1})$ is a known function of $X_1, \cdots, X_{p-1}, i = 1, \cdots, k$.

Note: k is not necessarily equal to $p - 1$.

Ex. 12.6 Original model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Which of the following models are curvilinear models?

(1) $Y = \alpha_0 + \alpha_1 \sqrt{X_1} + \alpha_2 \ln(X_2) + \varepsilon$.
(2) $y = \alpha_0 + \alpha_1 X_1 \sqrt{X_2} + \alpha_2 e^{X_1 - X_2/2} + \alpha_3 \sin(X_1) + \varepsilon$
(3) $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 e^{\alpha_3 X_2} + \varepsilon$

Note: Essentially, curvilinear models are the models with transformations on predictor variables. The transformation of X_i is determined by the relationship between *Y* and X_i, which can be found by: (1) the scatter plot of *Y* vs. X_i or (2) the residual plot of *e* vs. X_i.