Regression Models for Quantitative and Qualitative Predictors

Yang Feng

Two types of predictors

- Quantitative. (e.g., multiple linear regression)
- Qualitative. (e.g., indicator variables)

This Class:

- Polynomial Regression Models
- Interaction Regression Models

Polynomial Regression Models

- When the true curvilinear response function is indeed a polynomial function.
- When polynomial function is a good approximation to the true function.

One-predictor variable-second order

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \epsilon_i$$

where

$$x_i = X_i - \bar{X}$$

- X is centered due to the possible high correlation between X and X^2.
- Regression function: $E\{Y\} = \beta_0 + \beta_1 x + \beta_{11} x^2$, *quadratic response function*
- β_0 is the mean response when $x = 0$, *i.e.*, $X = \bar{X}$.
- β_1 is called the linear effect.
- β_{11} is called the quadratic effect.

One Predictor Variable-Third Order

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \epsilon_i$$

where

$$x_i = X_i - \bar{X}$$

One Predictor Variable-Higher Orders

- Employed with special caution.
- Tends to overfit
- Poor prediction

Two Predictors-Second Order

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \epsilon_i$$

where

$$x_{i1} = X_{i1} - \bar{X}_1, x_{i2} = X_{i2} - \bar{X}_2$$

- The coefficient β_{12} is called the interaction effect coefficient.
- More on interaction later.
- Three Predictors- Second Order is similar.

Implementation of Polynomial Regression Models

- Fitting—Very easy, just use the least squares for multiple linear regressions since they can all be seen as a multiple regression.
- Determine the order—Very important step!

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \epsilon_i$$

  Naturally, we want to test whether or not $\beta_{111} = 0$, or whether or not both $\beta_{11} = 0$ and $\beta_{111} = 0$.
  How to do the test?

Extra Sum of Squares

Decomposition SSR into $SSR(x)$, $SSR(x^2|x)$ and $SSR(x^3|x, x^2)$.

- Test whether $\beta_{111} = 0$: use $SSR(x^3|x, x^2)$.
- Test whether both $\beta_{11} = 0$ and $\beta_{111} = 0$: use $SSR(x^2, x^3|x)$.

Time for a real example!

Further Comments on Polynomial Regression

- There are drawbacks.
  1. Sometimes polynomial models are more expensive in degrees of freedom than alternative nonlinear models or linear models with transformed variables.
  2. Serious multicollinearity may be present even when the variables are centered
- An alternative to using centered variables is to use *orthogonal polynomials*.

Interaction Regression Models

- Additive effects:

$$E\{Y\} = f_1(X_1) + f_2(X_2) + \cdots + f_{p-1}(X_{p-1})$$

- General effects with interactions. Example:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- This cross-product term $\beta_3 X_1 X_2$ is called an interaction term.

Interpretation of Regression Models with Interactions

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- The change in mean response with a unit increase in X_1 when X_2 is held constant is

$$\beta_1 + \beta_3 X_2$$

- Similarly, a unit increase in X_2 when X_1 is constant:

$$\beta_2 + \beta_3 X_1$$

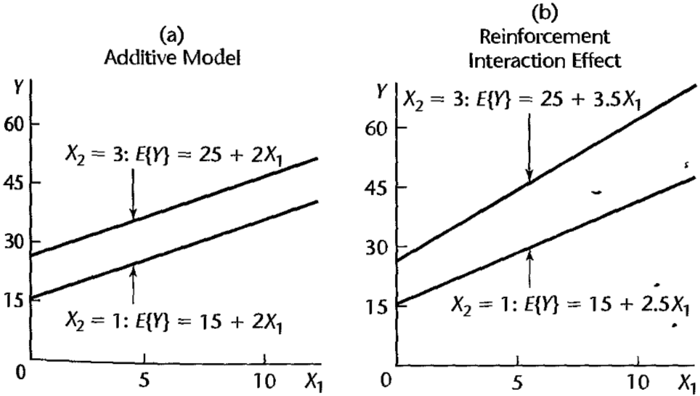
First type of interaction

First, suppose β_1 and β_2 are positive.

- Reinforcement (synergistic) type: $\beta_3 > 0$

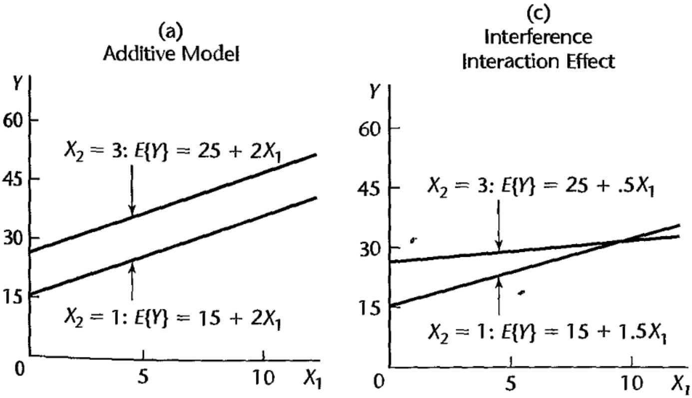
$$E\{Y\} = 10 + 2X_1 + 5X_2 + .5X_1X_2$$

- Conditional Effects Plot:



- Interference (antagonistic) type: $\beta_3 < 0$

$$E\{Y\} = 10 + 2X_1 + 5X_2 - .5X_1X_2$$



Implementation of Interaction Regression Models

- Center the predictor variables to avoid the *high multicollinearities*

$$x_{ik} = X_{ik} - \bar{X}_k$$

- Using prior knowledge to reduce the number of interactions. If we have 8 predictors, then we have 28 pairwise terms in total. For p predictors, the number is $p(p-1)/2$.

Implementation of Interaction Regression Models

- Center the predictor variables to avoid the *high multicollinearities*

$$x_{ik} = X_{ik} - \bar{X}_k$$

- Using prior knowledge to reduce the number of interactions. If we have 8 predictors, then we have 28 pairwise terms in total. For p predictors, the number is $p(p-1)/2$.

Now a real example...

Qualitative Predictors

Examples:

- Gender (male or female)
- Purchase status (yes or no)
- Disability status (not disabled, partly disabled, fully disabled)

A study of innovation in insurance industry

- Objective: related the speed with which a particular insurance innovation is adopted (Y) to the size of the insurance firm (X_1) and the type of the firm.
- Response Y: quantitative, continuous
- Predictor X_1: quantitative,
- Second predictor: type of firm, stock companies and mutual companies.

Qualitative Predictor with Two Classes

Suppose

$$X_2 = \begin{cases} 1, & \text{if stock company;} \\ 0, & \text{otherwise.} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{if mutual company;} \\ 0, & \text{otherwise.} \end{cases}$$

Then, we have the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

Design Matrix

Suppose, we have $n = 4$ observations, the first two being stock firms, the second two be mutual firms. Then

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{pmatrix}$$

Design Matrix

Suppose, we have $n = 4$ observations, the first two being stock firms, the second two be mutual firms. Then

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{pmatrix}$$

- Observation: first column is equal to the sum of the X_2 and X_3 columns, linear dependent...
- Solution: A qualitative variables with c classes will be represented by $c - 1$ indicator variables, each taking on the values 0 and 1.

Interpretation

Now, we drop the X_3 from the regression model:

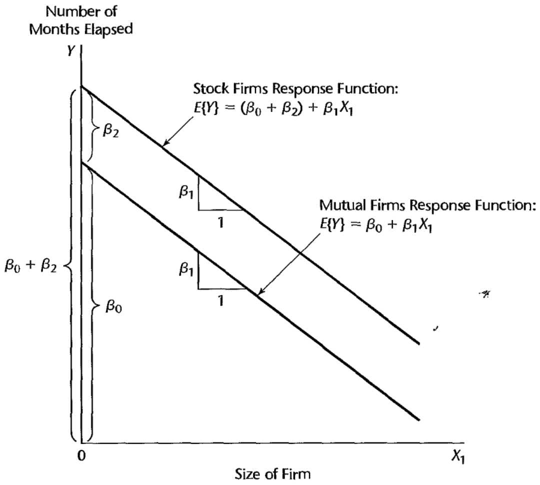
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

where

$$X_1 = \text{size of the firm}$$

$$X_2 = \begin{cases} 1, & \text{if stock company;} \\ 0, & \text{otherwise.} \end{cases}$$

Interpretation(Cont')



More than Two Classes

- Regression of tool wear (Y) on tool speed (X_1) and tool model (four classes M_1, M_2, M_3, M_4).
- 4 classes $\rightarrow$ 3 indicator variables
- Define

$$X_2 = \begin{cases} 1, & \text{if tool model } M_1; \\ 0, & \text{otherwise.} \end{cases}$$

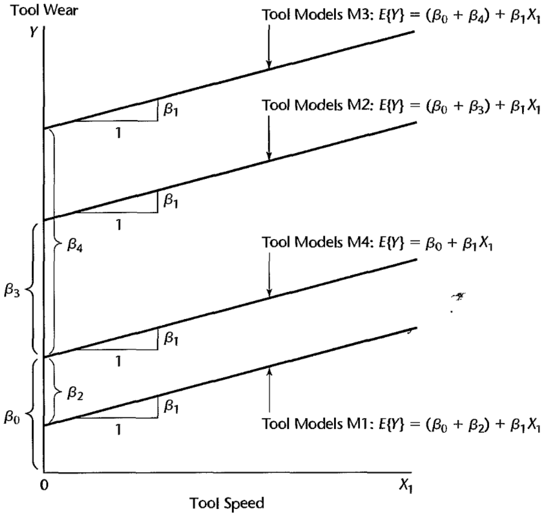
$$X_3 = \begin{cases} 1, & \text{if tool model } M_2; \\ 0, & \text{otherwise.} \end{cases}$$

$$X_4 = \begin{cases} 1, & \text{if tool model } M_3; \\ 0, & \text{otherwise.} \end{cases}$$

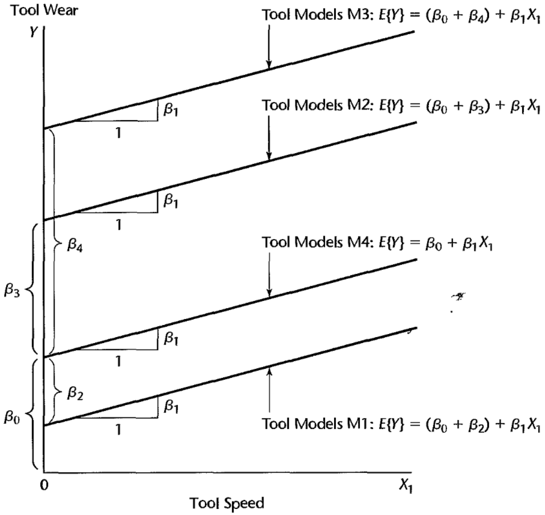
- Then, we have the following first-order regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i$$

Interpretation



Interpretation



Some Considerations in Using Indicator Variables

- An alternative: *allocated codes*.
- For example, the predictor variable "frequency of product use" has three classes: frequent user, occasional user, nonuser. We can use a single X_1 variable to denote it as follows:

$$X_1 = \begin{cases} 3, & \text{Frequent User;} \\ 2, & \text{Occasional User;} \\ 1, & \text{Nonuser.} \end{cases}$$

- Then, we have the regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

Difficulties with allocated codes

- The mean response with the regression function will be:

Class	$E\{Y\}$
Frequent User	$\beta_0 + 3\beta_1$
Occasional User	$\beta_0 + 2\beta_1$
Nonuser	$\beta_0 + \beta_1$

- Key implication:

$$\begin{aligned} & E\{Y|\text{frequent user}\} - E\{Y|\text{occasional user}\} \\ = & E\{Y|\text{occasional user}\} - E\{Y|\text{nonuser}\} \end{aligned}$$

- Using indicator variables doesn't have this restriction since it has one more variable to denote them.

Other Codings for Indicator Variables

- For the stock company and mutual company data:

$$X_2 = \begin{cases} 1, & \text{if stock company;} \\ -1, & \text{if mutual company.} \end{cases}$$

- Another alternative: use indicator variable for each of the c classes and drop the intercept term:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

where

$$X_1 = \text{size of the firm}$$

$$X_2 = \begin{cases} 1, & \text{if stock company;} \\ 0, & \text{otherwise.} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{if mutual company;} \\ 0, & \text{otherwise.} \end{cases}$$

Interactions between Quantitative and Qualitative Variables

- Almost the same as the regular interactions
- Read Chapter 8.5 and 8.6 after class

Comparison of Two or More Regression Functions

Three examples:

- A company operates two productions lines for making soap bars. For each line, the relationship between the speed of the line and the amount of scrap for the day was studied.
- An economist is studying the relationship between amount of savings and level of income for middle-income families from urban and rural areas, based on independent samples from the two populations.
- Two instruments were constructed for a company to identical specifications to measure pressure in an industrial process.

Soap Production Lines Example

- Y: scrap, X_1: line speed. X_2: code for production line.
- Interaction model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$$

where

$$X_{i1} = \text{line speed}$$

$$X_{i2} = \begin{cases} 1, & \text{if production line 1;} \\ 0, & \text{if production line 2.} \end{cases}$$

$$i = 1, 2, \cdots, 27$$