

# Multiple Linear Regression III

Sara López-Pintado

*Department of Statistics*

*Rutgers University*

Fall, 2005

## Main Topics:

- Polynomial regression
- Interaction regression models
- Qualitative variables

- Polynomial regression models

- For example: Second order model with a predictor variable

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

where  $x_i = x_i - \bar{x}$ .

- Usually the variables are centered
- Careful with extrapolating with these models
- The sign of  $\beta_2$  determines the shape of the curves

- For example: One predictor third order:

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

- Example: Two predictors-second order

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

- Fit the regression model in the usual way
- For testing use a Hierarchical approach to fitting:

$$SSR(x, x^2, x^3) = SSR(x) + SSR(x^2|x) + SSR(x^3|x, x^2)$$

Drawbacks of polynomial regression models:

- Can be expensive in degrees of freedom
- Problem with multicollinearity even after centering the data
- Careful over fitting the data with high order polynomials.

## - Interaction Regression Model

A response function with  $p - 1$  predictors will contain additive effects if it can be written as:

$$E[y] = f_1(x_1) + f_2(x_2) + \dots + f(x_{p-1})$$

Example: the effects of  $x_1$  and  $x_2$  are additive

$$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$$

but the following model is not additive:

$$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Careful because now  $\beta_1$  and  $\beta_2$  don't have the same interpretation as before: the effect on  $y$  of the predictor variables are dependent on the level of the other variables

- The mean increase in the response variable for a unit increase in  $x_1$  given  $x_2$  is constant is:

$$\beta_1 + \beta_3 x_2$$

- Similarly, the change in mean response for a unit increase in  $x_2$  given  $x_1$  constant is

$$\beta_2 + \beta_3 x_1$$

If there are many predictor variables in a regression model the possible number of interaction increases a lot. Ways of selecting these interaction terms:

- Plotting the residual of an additive model versus each interaction term
- Only use those interaction terms involving predictor variables that are important in the model

- Qualitative independent variables

How to include qualitative variables in a regression model?

- Example: An economist wants to relate the speed with which a particular insurance innovation is adopted ( $y$ ) with the size of the firm ( $x_1$ ) and the type of firm ( $x_2$ ).

$y$  measures number of months between first firm adopted the innovation and the given firm adopted it

$x_1$  (quantitative) measures the amount of total assets of a firm

$x_2$  (qualitative) two classes: stock companies, and mutual companies

To include qualitative variables in the model we use dummy variables: a qualitative variable with  $c$  classes will be represented by  $c - 1$  indicator functions, each taking values 0 and 1.

$$x_2 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

$\beta_2$  is the differential effect of type of firm (how much higher or lower the mean response line is for the classes coded 1 than for the class coded 0, for any given  $x_1$ .)

- Model containing interaction effects

In previous example we could include the possibility of interaction effects between size of company and type of company

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i,$$

The response function for this regression model is

$$E[y] = \beta_0 + \beta_1 x_1$$

for mutual firms

$$E[y] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1$$

for stock firms

- More complex models

Consider the regression model with a qualitative variable with more than two classes:  $y$  productivity and  $x_1$  quantitative variable (investment) and a qualitative variable: 4 types of machines (A, B, C, and D)

$$x_2 = \begin{cases} 1 & \text{Machine A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{Machine B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{Machine C} \\ 0 & \text{otherwise} \end{cases}$$

A first-order regression model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i,$$

Machine A

$$E[Y] = \beta_0 + \beta_2 + \beta_1 x_1,$$

Machine B

$$E[Y] = \beta_0 + \beta_3 + \beta_1 x_1,$$

Machine C

$$E[Y] = \beta_0 + \beta_4 + \beta_1 x_1,$$

Machine D

$$E[Y] = \beta_0 + \beta_1 x_1,$$

First order regression model with interaction

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} x_{i2} + \beta_6 x_{i1} x_{i3} + \beta_7 x_{i1} x_{i4} + \varepsilon_i,$$

- More than one qualitative predictor variable

$y$   $\longrightarrow$  advertising expenditure

$x_1$   $\longrightarrow$  sales

$x_2$   $\longrightarrow$  type of firm (two classes: incorporated and sales management)

$x_3$   $\longrightarrow$  quality of sales (two classes: high and low)

$$x_2 = \begin{cases} 1 & \text{if firm incorporated} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if quality of sales management high} \\ 0 & \text{otherwise} \end{cases}$$

- First order model with no interaction

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon$$

same slope for all the type of firm-quality of sales combination

- First-order model with certain interactions added

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \varepsilon$$

For example: If type of firm is incorporation and quality of sales high:

$$E[y] = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5) x_1$$

- Comparison of two or more regression functions (soap production lines example)

$y$   $\longrightarrow$  amount of soap

$x_1$   $\longrightarrow$  line speed

$x_2$   $\longrightarrow$  code for the production line: two classes (1 if production line 1 and 0 if production line 2)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

To test if the regression lines for the two production lines are the same:

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_1 : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$$

- Piecewise linear regression: sometimes the regression of  $y$  on  $x$  follows a particular linear relation in some range of  $x$ , but follows a different relation elsewhere

Example:

$y$   $\longrightarrow$  unit cost

$x_1$   $\longrightarrow$  lot size

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 500) x_{i2} + \varepsilon_i$$

where

$$x_{i1} = \text{lot size}$$
$$x_{i2} = \begin{cases} 1 & \text{if } x_{i1} > 500 \\ 0 & \text{otherwise} \end{cases}$$

The mean response function is:

If  $x_1 \leq 500$  then  $x_2 = 0$

$$E[y] = \beta_0 + \beta_1 x_1$$

If  $x_1 > 500$  then  $x_2 = 1$

$$E[y] = \beta_0 - \beta_2 500 + (\beta_1 + \beta_2)x_1$$

We can extend this model to more than two piecewise regression lines, regression functions with discontinuity,...