9.2 More complex models

- Qualitative independent variables
- Interaction Model
- Polynomial Regression Models
- Summary of First-order and Second-order Models
- Coefficients of Partial Determination

Qualitative Predictors

定性的

- To quantify qualitative predictors, we use indicator variables (dummy variables).
- An indicator variable is a categorical explanatory variable with two levels:
  - yes or no, on or off, male or female
  - coded as 0 or 1
- If more than two levels, the number of indicator variables needed is (number of levels - 1)

Indicator-Variable Example (with 2 Levels)

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Let:

Y = pie sales

X_1 = price

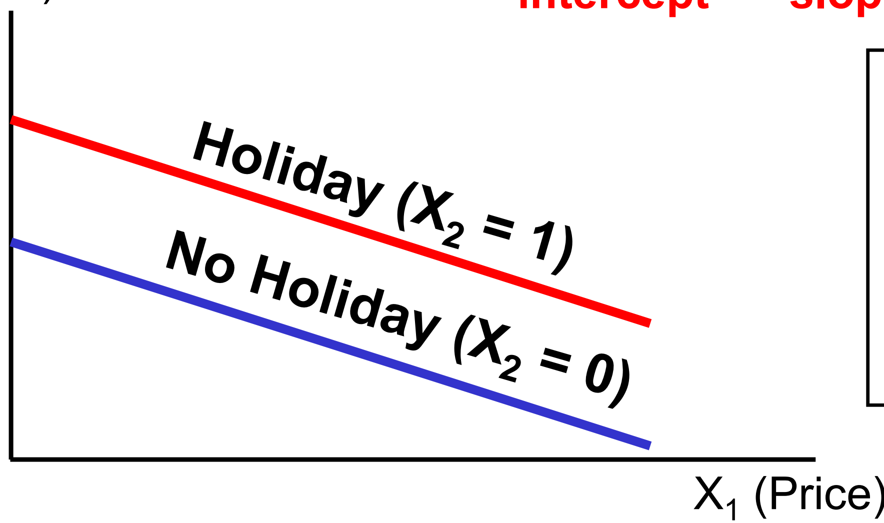
X_2 = holiday (X_2 = 1 if a holiday occurred during the week)
(X_2 = 0 if there was no holiday that week)

Indicator-Variable Example (with 2 Levels)

$$\hat{Y} = b_0 + b_1 X_1 + b_2(1) = (b_0 + b_2) + b_1 X_1$$ **Holiday**

$$\hat{Y} = b_0 + b_1 X_1 + b_2(0) = b_0 + b_1 X_1$$ **No Holiday**

Different intercept **Same slope**



If H_0: $\beta_2 = 0$ is rejected, then “Holiday” has a significant effect on pie sales

Interpreting the Indicator Variable Coefficient (with 2 Levels)

Example: $\text{Sales} = 300 - 30(\text{Price}) + 15(\text{Holiday})$

Sales: number of pies sold per week
Price: pie price in $

$$\text{Holiday:} \begin{cases} 1 & \text{If a holiday occurred during the week} \\ 0 & \text{If no holiday occurred} \end{cases}$$

b_2 = 15: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

Indicator-Variable Models (more than 2 Levels)

- The number of dummy variables is **one less than the number of levels**
- Example:

  Y = house price ; X_1 = square feet

- If style of the house is also thought to matter:

  Style = $\underbrace{\text{ranch, split level, condo}}$

Three levels, so two dummy variables are needed

Indicator-Variable Models (more than 2 Levels)

- Example: Let “condo” be the default category, and let X_2 and X_3 be used for the other two categories:

$$Y = \text{house price}$$
$$X_1 = \text{square feet}$$
$$X_2 = 1 \text{ if ranch, } 0 \text{ otherwise}$$
$$X_3 = 1 \text{ if split level, } 0 \text{ otherwise}$$

The multiple regression equation is:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Interpreting the indicator Variable Coefficients (with 3 Levels)

Remark 9.2

Consider the regression equation:

$$\hat{Y} = 20.43 + 0.045X_1 + 23.53X_2 + 18.84X_3$$

For a condo: $X_2 = X_3 = 0$

$$\hat{Y} = 20.43 + 0.045X_1$$

For a ranch: $X_2 = 1$; $X_3 = 0$

$$\hat{Y} = 20.43 + 0.045X_1 + 23.53$$

With the same square feet, a ranch will have an estimated average price of 23.53 thousand dollars more than a condo

For a split level: $X_2 = 0$; $X_3 = 1$

$$\hat{Y} = 20.43 + 0.045X_1 + 18.84$$

With the same square feet, a split-level will have an estimated average price of 18.84 thousand dollars more than a condo.

Remark 9.2

ranch split-level

dummy variable

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + \varepsilon_i$$

$$\hat{y} = 20.43 + 0.045 X_1 + 23.53 X_2 + 18.84 X_3$$

Difference between ranch and cando $= \hat{b}_2 = 23.53$

split level and cando $= \hat{b}_3 = 18.84$

Ranch and split level: $= \hat{b}_2 - \hat{b}_3 = 4.69$

An alternative model:

$$X_2 = \begin{cases} 0 & \text{cando} \\ 1 & \text{split level} \\ 2 & \text{Ranch} \end{cases}$$

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

diff between spilt level and Cando $= \beta_2$

Ranch and split level $= \beta_2$

Ranch and Cando $= \beta_2$

→ meaningless

Interaction Regression Models

- Hypothesizes interaction between pairs of X variables
  - Response to one X variable may vary at different levels of another X variable

(Interaction terms)

- Contains two-way cross product terms

for example: $X_2 = \begin{cases} 1 \\ 0 \end{cases}$

  - $$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$
    $$= b_0 + b_1 X_1 + b_2 X_2 + b_3 (X_1 X_2)$$

$X_2 = 1 \Rightarrow b_0 + b_2 + (b_1 + b_3) X_1$

$X_2 = 0 \Rightarrow b_0 + b_1 X_1$

$X_2 = 0 \Rightarrow \beta_0 + \beta_1 X_1$

Interaction Regression Models

Example: 3 predictor variables

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

$$= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_6 X_{i6} + \varepsilon_i$$

hypothesis: $\beta_4 = \beta_5 = \beta_6 = 0.$

$$\Rightarrow \underset{\sim}{\beta} = \begin{pmatrix} \underset{\sim}{\beta_1} \\ \underset{\sim}{\beta_3} \end{pmatrix}$$

$H_0: \underset{\sim}{\beta_3} = 0$

Effect of Interaction

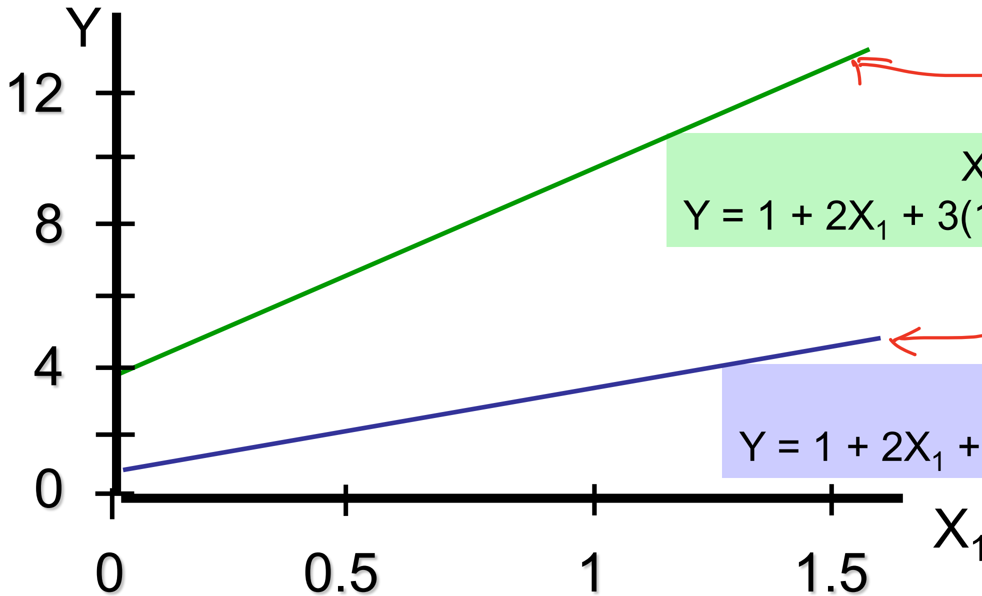
- Given: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

- Without interaction term, effect of X_1 on Y is measured by β_1
- With interaction term, effect of X_1 on Y is measured by $\beta_1 + \beta_3 X_2$
- Effect changes as X_2 changes

Effect of Interaction

Suppose X_2 is a dummy variable and the estimated regression equation is

$$\hat{Y} = 1 + 2X_1 + 3X_2 + 4X_1X_2$$



$X_2 = 1$:
$Y = 1 + 2X_1 + 3(1) + 4X_1(1) = 4 + 6X_1$

$X_2 = 0$:
$Y = 1 + 2X_1 + 3(0) + 4X_1(0) = 1 + 2X_1$

Slopes are different if the effect of X_1 on Y depends on X_2 value

Significance of Interaction Term

- Can perform a partial F-test for the contribution of a variable to see if the addition of an interaction term improves the model

- Multiple interaction terms can be included
  - Use a partial F-test for the simultaneous contribution of multiple variables to the model

Polynomial Regression Models

When are polynomial regression models being used?

- When the true curvilinear response function is indeed a polynomial function
- When the true curvilinear response function is unknown (or complex) but a polynomial function is a good approximation to the true function.

Polynomial Regression Models

Example: 1 predictor variable, second order

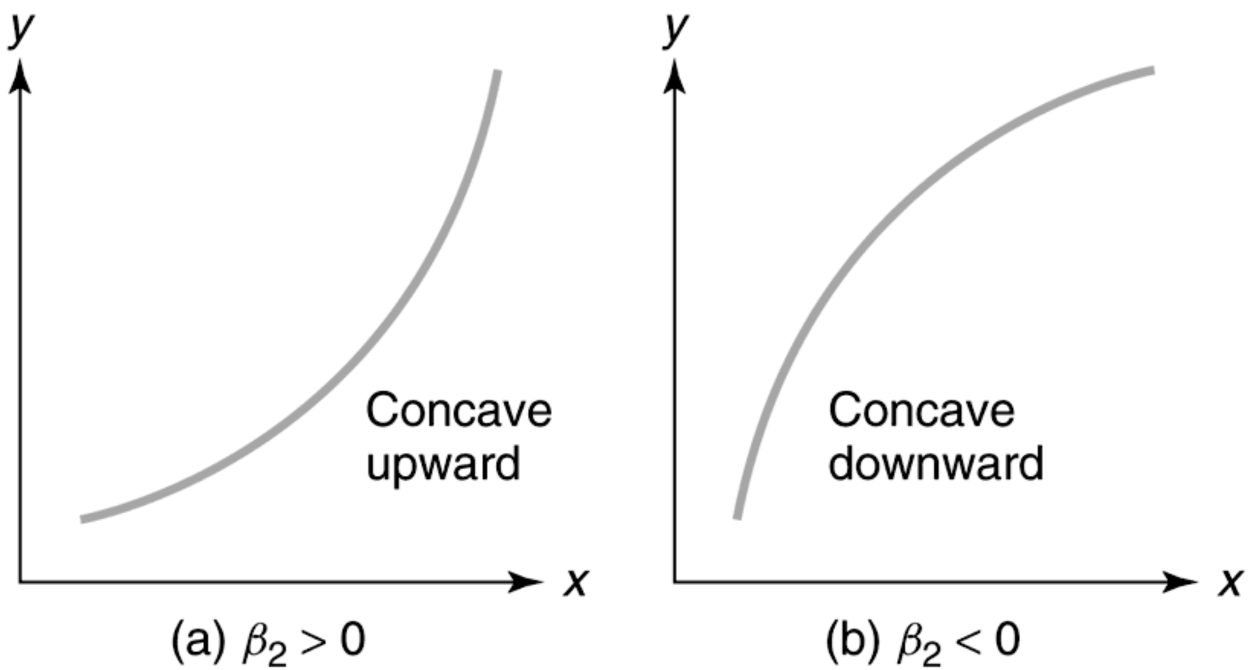
$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where $= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$ where $X_{i2} = X_{i1}^2$

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

The reason for using a centered predictor variable in the polynomial regression model is that X and X2 often will be highly correlated. Centering the predictor variable often reduces the multicollinearity substantially, and tends to avoid computational difficulties.

Graphs for two quadratic models



Polynomial Regression Models

Example: 2 predictor variables, second order

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

where

$$x_{i1} = X_{i1} - \overline{X}_1$$

$$x_{i2} = X_{i2} - \overline{X}_2$$

Coefficients of partial determination

$$R^2_{Yj.(\text{all variables except } j)} = \frac{SSR(X_j \mid \text{all variables except } j)}{SSE(\text{all variables except } j)}$$

- Measures the proportion of variation in the dependent variable that is explained by X_j while controlling for (holding constant) the other explanatory variables
- Coefficients of partial correlation

Remark 9.3 Coefficients of partial determination

Full model: $\underset{\sim}{y} = (\underset{\sim}{X_1}\ \underset{\sim}{X_2}) \begin{pmatrix} \underset{\sim}{\beta_1} \\ \underset{\sim}{\beta_2} \end{pmatrix} + \underset{\sim}{\varepsilon}$

Reduced model: $\underset{\sim}{y} = \underset{\sim}{X_1}\underset{\sim}{\beta_1} + \underset{\sim}{\varepsilon}$ (Full model + H_0: $\underset{\sim}{\beta_2} = 0$)

Coefficient of partial determination (for X_i) $= \dfrac{SSE_{reduced} - SSE_{full}}{SSE_{reduced}}$ $\in (0,1)$

$> SSE_{full}$

Special case: Reduced model = $\begin{pmatrix} \text{Full model} \\ + H_0: \underset{\sim}{\beta_2} = 0 \end{pmatrix}$

$y \sim X_1 + X_2 + \cdots + X_5$

$lm(X_1 \sim X_2 + X_3 + X_4 + X_5) \rightarrow R_1^2$

$lm(X_2 \sim X_1 + X_3 + X_4 + X_5) \rightarrow R_2^2$