# Ch8. Diagnostics and model building

## 8.1 Model validation and diagnostics–(i) Residuals

The model is

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

Residuals are defined

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{y} - \hat{\mathbf{Y}} = (\boldsymbol{I} - \mathbf{H})\boldsymbol{y} = (\boldsymbol{I} - \mathbf{H})\boldsymbol{\epsilon}$$

where $\boldsymbol{x}$ is $n \times (p+1)$ and the hat matrix (projection matrix) is

$$\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

In terms of the elements $h_{ij}$ of $\mathbf{H}$,

$$\hat{\epsilon}_i = \epsilon_i - \sum_{j=1}^{n} h_{ij}\epsilon_j, \quad i = 1, 2, ..., n.$$

# Model validation and diagnostics–(i) Residuals

## Properties

1. $E(\hat{\boldsymbol{\epsilon}}) = \mathbf{0}$ (residual mean is the same as the error mean)
2. $\text{Cov}(\hat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (residuals are NOT independent)
3. $\text{Cov}(\hat{\boldsymbol{\epsilon}}, \boldsymbol{y}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (residuals correlated with the observations)
4. $\text{Cov}(\hat{\boldsymbol{\epsilon}}, \hat{\mathbf{Y}}) = \mathbf{0}$ (residuals uncorrelated with the predicted values)
5. $\bar{\hat{\epsilon}} = \sum_{i=1}^{n} \hat{\epsilon}_i / n = \hat{\boldsymbol{\epsilon}}' \mathbf{1}/n = 0$
6. $\hat{\boldsymbol{\epsilon}}' \boldsymbol{y} = SSE$
7. $\hat{\boldsymbol{\epsilon}}' \hat{\mathbf{Y}} = 0$
8. $\hat{\boldsymbol{\epsilon}}' \mathbf{X} = \mathbf{0}'$ ($\hat{\boldsymbol{\epsilon}}'$ is orthogonal to each column of $\mathbf{X}$)

# 8.1 Model validation and diagnostics–(i) Residuals

## Model in centered form

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \\ &= \alpha + \beta_1 (x_{i1} - \bar{x}_1) + \cdots + \beta_2 (x_{ip} - \bar{x}_p) + \epsilon_i \end{aligned}$$

In matrix form, the centered model is

$$\boldsymbol{y} = (\mathbf{1}, \boldsymbol{X}_c) \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, ..., \beta_p)'$, and $\boldsymbol{X}_c = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{X}_1$.
It can be shown from previous notes that the least squares estimators of the parameters are

$$\hat{\alpha} = \bar{y}, \quad \hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{y}.$$

# 8.1 Model validation and diagnostics–(i) Residuals

Hence the predicted value is

$$\hat{\boldsymbol{y}} = \bar{y}\mathbf{1} + \boldsymbol{X_c}(\boldsymbol{X_c}'\boldsymbol{X_c})^{-1}\boldsymbol{X_c}'\boldsymbol{y} = \left(\frac{1}{n}\mathbf{1}'\boldsymbol{y}\right)\mathbf{1} + \boldsymbol{H}_c\boldsymbol{y} = \left(\frac{1}{n}\boldsymbol{J} + \boldsymbol{H}_c\right)\boldsymbol{y}$$

Since $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, hence

$$\boldsymbol{H} = \frac{1}{n}\boldsymbol{J} + \boldsymbol{H}_c = \frac{1}{n}\boldsymbol{J} + \boldsymbol{X_c}(\boldsymbol{X_c}'\boldsymbol{X_c})^{-1}\boldsymbol{X_c}'.$$

# 8.1 Model validation and diagnostics–(i) Residuals

Hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \{h_{ij}\}$ (Let $\boldsymbol{x}$ be a matrix with full column rank and with $\mathbf{1}$ as its first column)

**Properties**

1. $1/n \leq h_{ii} \leq 1$ for $i = 1, 2, ..., n$.
2. $-.5 \leq h_{ij} \leq .5$ for all $j \neq i$.
3. $h_{ii} = 1/n + (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'(\boldsymbol{x}_c'\boldsymbol{x}_c)^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)$, where $\mathbf{x}_{1i}' = (x_{i1}, x_{i2}, ..., x_{ik})$, $\bar{\mathbf{x}}_1' = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_k)$, and $(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'$ is the $i$th row of the centered matrix $\boldsymbol{x}_c$.
4. $\text{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p + 1$.

# 8.1 Model validation and diagnostics–(ii) Residual Analysis

1. Variance of the residuals is not constant
   - Studentized residual
   - Studentized deleted residual (externally studentized residual)
2. Deleted residuals
3. Press (prediction sum of squares)

# 8.1 Model validation and diagnostics–(iii) Influential Observations

1. Leverage $h_{ii}$
2. Cook's distance

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \boldsymbol{X}' \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(p+1)\hat{\sigma}^2}$$

# Model validation and diagnostics

**Example.** Ch8-Example-Diagnostics

# 8.2 Multicollinearity

**Multicollinearity** in regression refers to the case when one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

## Detection

- Significant correlations between pairs of independent variables in the model;
- Nonsignificant t-tests for all (or nearly all) the individual $\beta$ parameters when the F-test for overall model adequacy $H_0 : \beta_1 = \beta_2 = \ = \beta_p = 0$ is significant;
- Opposite signs (from what is expected) in the estimated parameters;

# 8.2 Multicollinearity – Detection

- A variance inflation factor (VIF) for a $\beta$ parameter greater than 10, where

$$(VIF)_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p,$$

and $R_j^2$ is the multiple coefficient of determination for the model

$$\mathrm{E}(x_j) = \alpha_0 + \alpha_1 x_1 + \ + \alpha_{j-1} x_{i-1} + \alpha_{j+1} x_{j+1} + \ + \alpha_p x_p.$$

- Let the ordered eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ be $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r$ where $r = p + 1$.
  - Condition number: $\kappa = \sqrt{\lambda_1/\lambda_r}$;
  - Condition indexes: $\sqrt{\lambda_1/\lambda_j}$ where $j = 2, \ldots, r$;
  - Multicollinear problem: Condition number or indexes $\geq 30$.
  - Note in some references, condition index is also called condition number.

# 8.2 Multicollinearity – example

**Example.** Ch8-Example-Multicollinearity

# 8.3 Variable selection

- Goal is to develop a model with the best set of independent variables
  - Easier to interpret if unimportant variables are removed,
  - Lower probability of collinearity.
- Stepwise regression procedure: Provide evaluation of alternative models as variables are added (or withdrawn).
- Best-subset approach: Try all combinations and select the best using various criteria, such as the highest adjusted $R^2$.
- Other methods: LASSO, Ridge, Elastic Net

# 8.3 Variable selection – example

**Example.** Ch8-Example-Variable-selection

# 8.4 More complex models

- Qualitative independent variables;
- Interaction Model;
- Polynomial Regression Models;
- Summary of First-order and Second-order Models;
- Coefficients of Partial Determination;
- Other regression models.

# 8.4 More complex models – Qualitative independent variables

- 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)

# 8.4 Qualitative independent variables: 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.

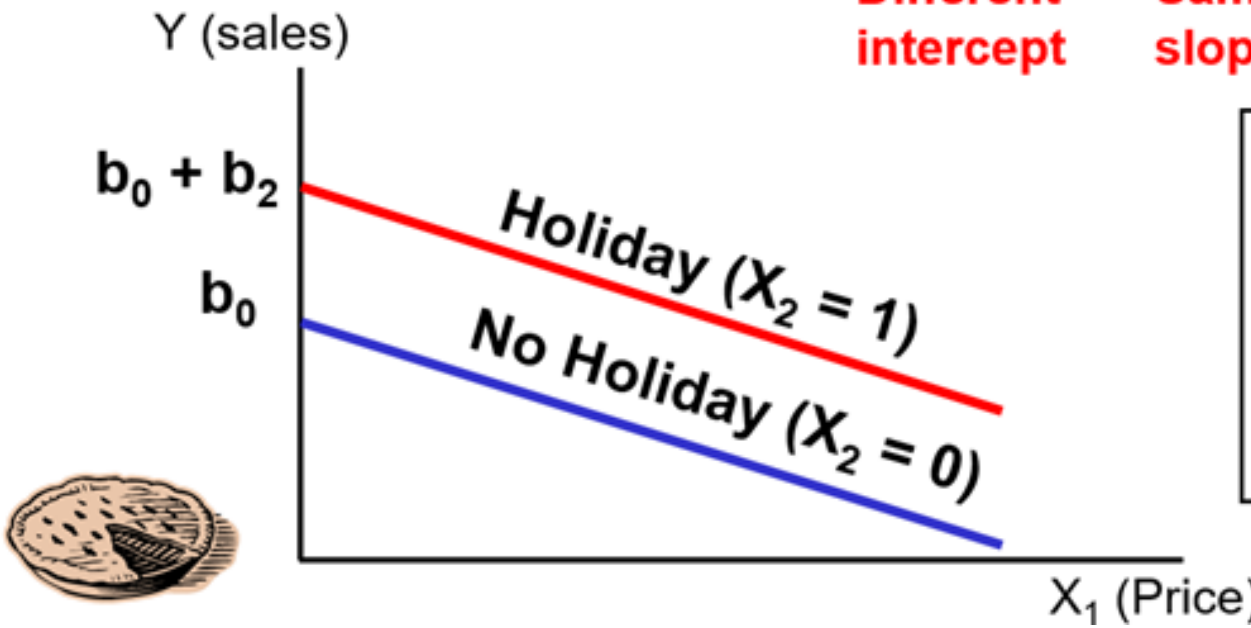
# 8.4 Qualitative independent variables:
# Indicator-Variable Example (with 2 Levels)

$$\hat{Y} = b_0 + b_1X_1 + b_2(1) = (b_0 + b_2) + b_1X_1 \quad \text{Holiday}$$

$$\hat{Y} = b_0 + b_1X_1 + b_2(0) = b_0 + b_1X_1 \quad \text{No Holiday}$$

Different intercept

Same slope




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

4

# 8.4 Qualitative independent variables: Interpreting the Indicator-Variable Example (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

# 8.4 Qualitative independent variables:
# Indicator-Variable Example (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 = ranch, split level, condo

Three levels, so two dummy variables are needed

6

# 8.4 Qualitative independent variables: Indicator-Variable Example (more than 2 Levels)

Let 'condo' be the default category, and let $x_2$ and $x_3$ be used for the other two categories:

- $Y$ = house price;
- $X_1$ = square feet;
- $X_2$ = 1 if ranch, 0 otherwise;
- $X_3$ = 1 if split level, 0 otherwise

The multiple regression equation is:

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

# 8.4 Qualitative independent variables: Interpreting the indicator Variable Coefficients (with 3 Levels)

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.

# 8.4 Interaction Regression Models

- Hypothesizes interaction between pairs of $X$ variables
  - Response to one $X$ variable may vary at different levels of another $X$ variable
- Contains two-way cross product terms

$$\begin{aligned} \hat{Y} &= b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 \\ &= b_0 + b_1 X_1 + b_2 X_2 + b_3 (X_1 X_2) \end{aligned}$$

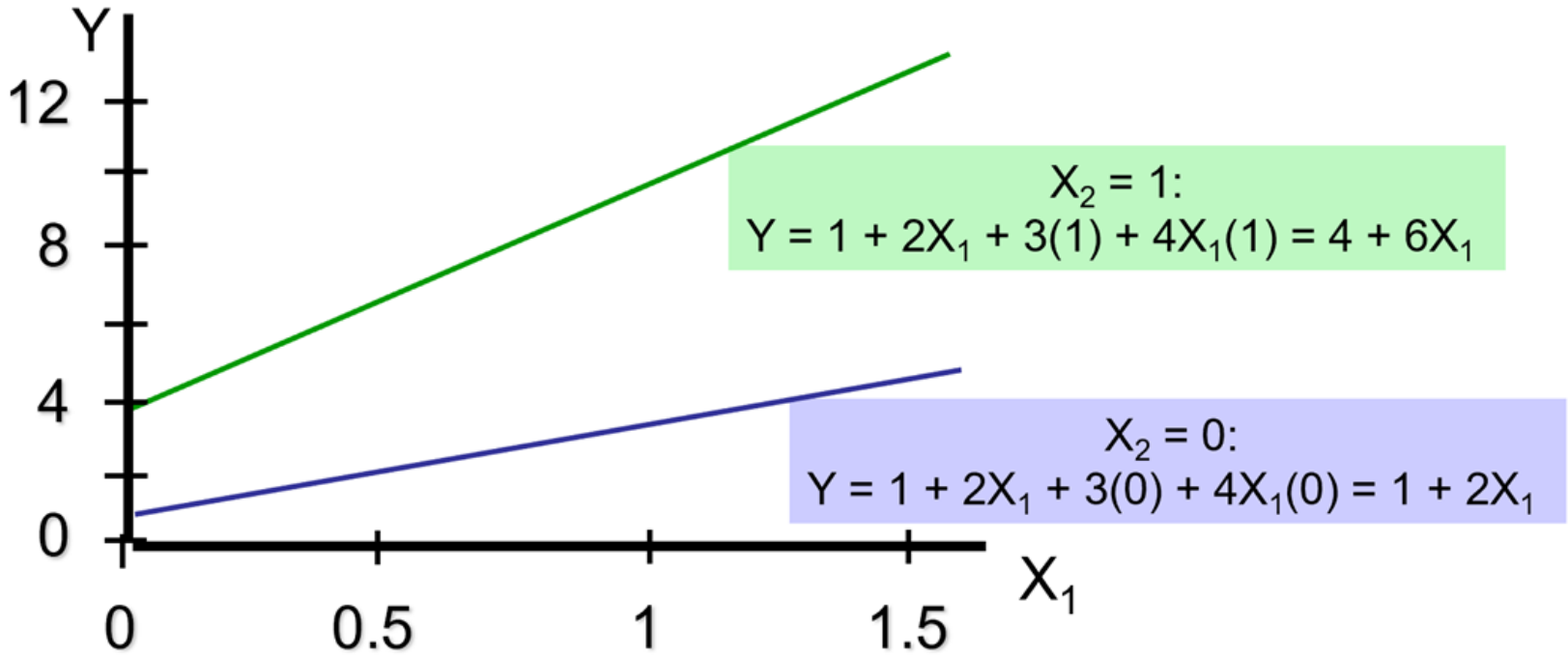
# 8.4 Interaction Regression Models

- Given $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 (X_1 X_2)$
- Without interaction term, effect of $X_1$ on $Y$ is measured by $\beta_1$;
- With interaction term, effect of $X_1$ on $Y$ is measured by $\beta_1 + \beta_3 X_2$;
- Effect changes as $X_2$ changes.

# 8.4 Interaction Regression Models

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

# 8.4 Interaction Regression Models: 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.

# 8.4 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.

# 8.4 Polynomial regression models

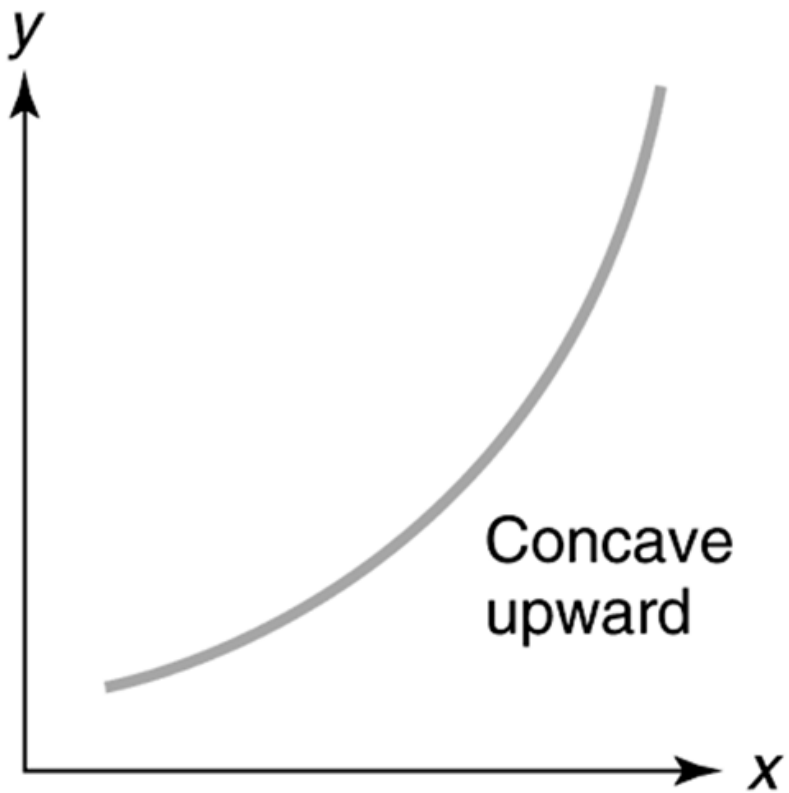
**Example** Predictive variable, second order

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

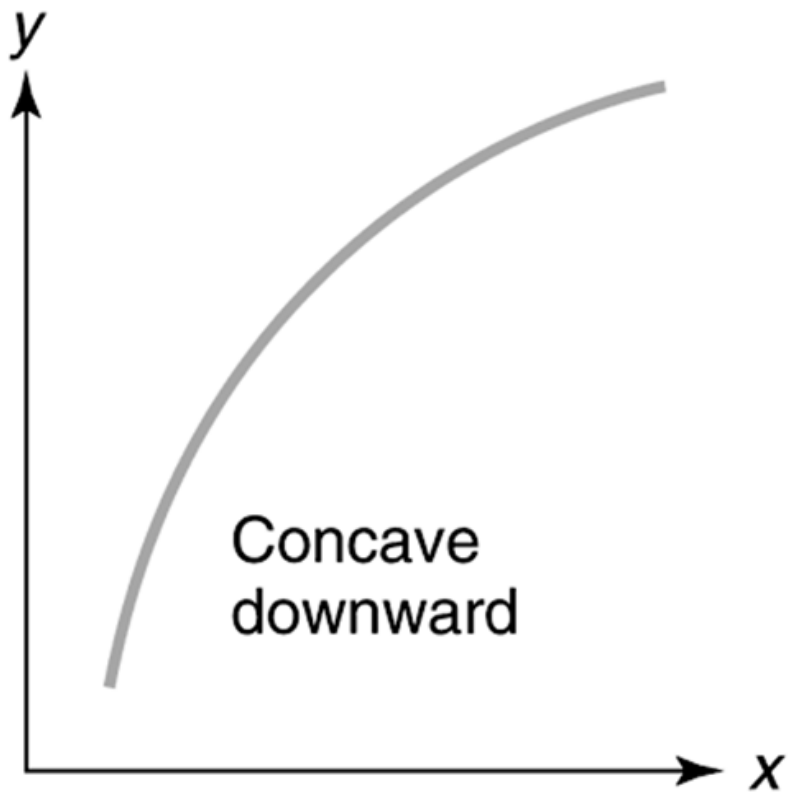
where $x_i = X_i - \bar{X}$.

- The reason for using a centered predictor variable in the polynomial regression model is that $X$ and $X^2$ often will be highly correlated.
- Centering the predictor variable often reduces the multicollinearity substantially, and tends to avoid computational difficulties.

# 8.4 Polynomial regression models: graphs for two quadratic models




(a) $\beta_2 > 0$




(b) $\beta_2 < 0$

# 8.4 Coefficients of partial determination

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

- 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.