Multiple Linear Regression

- Dependent variable (y) depends on p independent variables x_j, $j = 1, 2, \ldots, p$
- General linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

- For ith observation

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \epsilon_i$$

- Objective: Using n observations, estimate regression coefficients

Multiple Linear Regression

- Approach similar to simple regression

  Minimize the sum of squares of the errors

- Vector and matrix notations

$$\mathbf{y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{1,1} - \bar{x}_1 & x_{2,1} - \bar{x}_2 & \cdots & x_{p,1} - \bar{x}_p \\ x_{1,2} - \bar{x}_1 & x_{2,2} - \bar{x}_2 & \cdots & x_{p,2} - \bar{x}_p \\ \vdots & \vdots & \cdots & \vdots \\ x_{1,n} - \bar{x}_1 & x_{2,n} - \bar{x}_2 & \cdots & x_{p,n} - \bar{x}_p \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

- The linear model in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \; E(\boldsymbol{\epsilon}) = \mathbf{0}, \; Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

- SSE

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Multiple Linear Regression

- Minimization of the SSE leads to the normal equations

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

- Assumption: $(\mathbf{X}^T\mathbf{X})$ is of full rank p (invertible)
- The coefficients vector

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y};\ \ \beta_0 = \bar{y} - \bar{\mathbf{x}}^T\hat{\boldsymbol{\beta}}$$

- The properties of the estimators

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$
$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE)

Multiple Linear Regression

- Estimate of the error variance

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-p-1}$$

where *(n-p-1)* is the degrees of freedom (df)

- 1-α confidence intervals for $\beta_j, j = 0, 1, \ldots, p$

$$\beta_j \in [\hat{\beta}_j - t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_j), \hat{\beta}_j + t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_j)]$$

$t_{(n-p-1,\, \alpha/2)}$ is the $(1 - \alpha/2)$ percentile point of the *t*-distribution with *(n-p-1) df*

$$s.e.(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}$$

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$$

Multiple Linear Regression

- Multiple correlation coefficient

$$Cor(y,\hat{y}) = \frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2 \sum(\hat{y}_i-\bar{\hat{y}})^2}}$$

- The coefficient of determination R^2

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$$

- Adjusted R-squared, $R_a{}^2$

$$R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Multiple Linear Regression

- Fitted model is adequate or can be reduced further?
  - Test significance of individual coefficient $\widehat{\beta}$
  - A general unified test on the full model (FM) vs the reduced model (RM)
- Hypothesis testing

  H_0: *Reduced model is adequate*

  H_1: *Full model is adequate*

Multiple Linear Regression

- Testing two models: RM with k parameters
- F-statistic

$$F_o = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

Degrees of freedom

  - Note that SSE(RM) ≥ SSE(FM)

- For α-significance level: Reject H_o if

$$F_o \geq F_{(p+1-k, n-p-1; \alpha)}$$

where F-statistic for the given dfs from the table

Multiple Linear Regression

Menu pricing in Restaurants of NYC

y : Price of dinner

x_1: Customer rating of the food (Food)

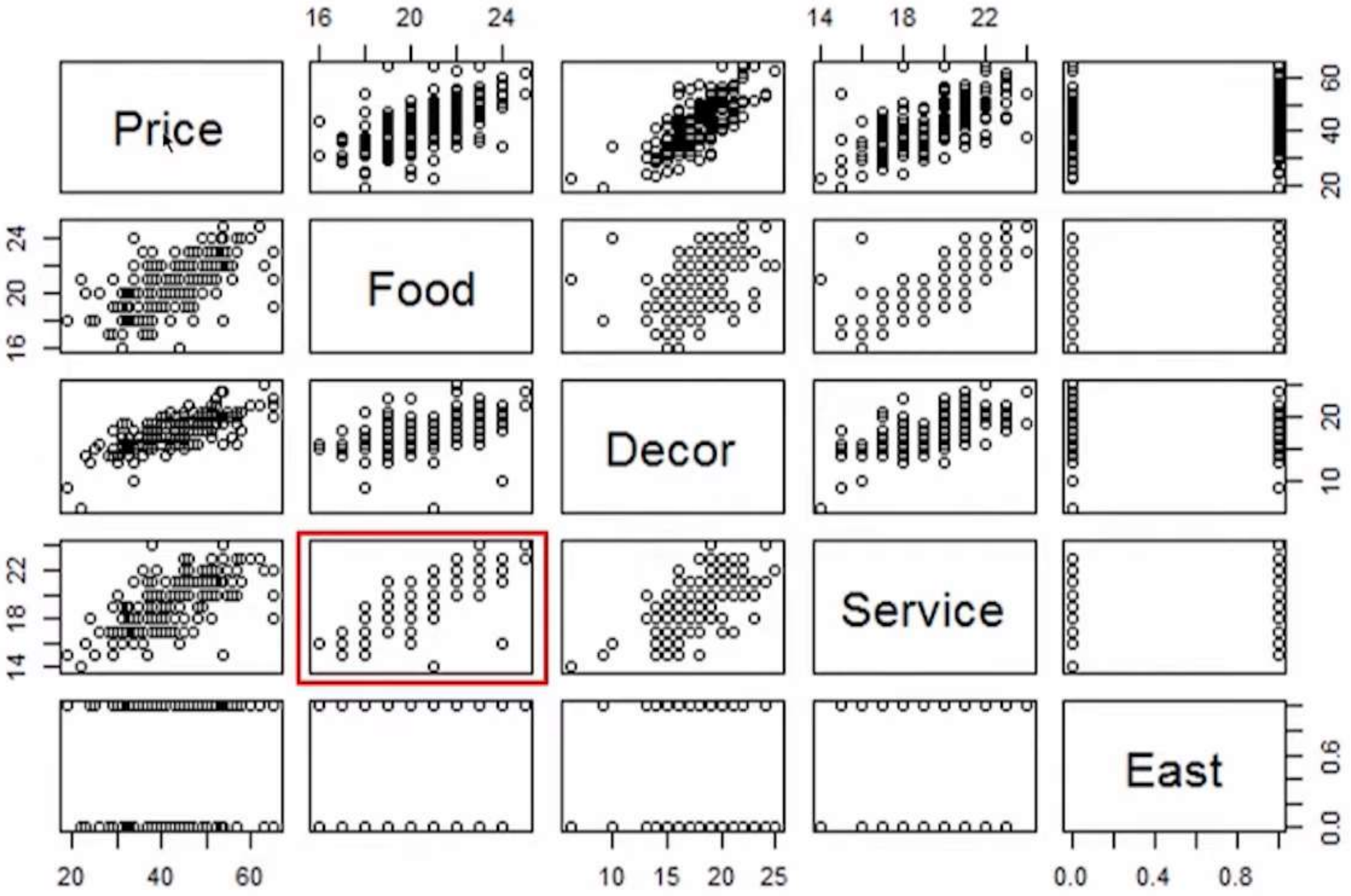
x_2: Customer rating of the décor (Décor)

x_3: Customer rating of the service (Service)

x_4: If the restaurant is east or west (East)

Objective: Build a model

Multiple Linear Regression



Multiple Linear Regression

Regression output from R

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.023800   4.708359  -5.102 9.24e-07 ***
Food          1.538120   0.368951   4.169 4.96e-05 ***
Decor         1.910087   0.217005   8.802 1.87e-15 ***
Service      -0.002727   0.396232  -0.007   0.9945
East          2.068050   0.946739   2.184   0.0304 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.738 on 163 degrees of freedom
Multiple R-squared:  0.6279, Adjusted R-squared:  0.6187
F-statistic: 68.76 on 4 and 163 DF,  p-value: < 2.2e-16
```

$$\hat{y}_i = -24.024 + 1.538x_1 + 1.910x_2 - \boxed{0.003x_3} + 2.068x_4$$

Remove x_3

Multiple Linear Regression

Regression output from R without Service variable

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279, Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

$$\hat{y}_i = -24.027 + 1.536x_1 + 1.910x_2 + 2.067x_4$$

Caution: Removing several predictors may have a dramatic effect on the coefficients in the reduced model

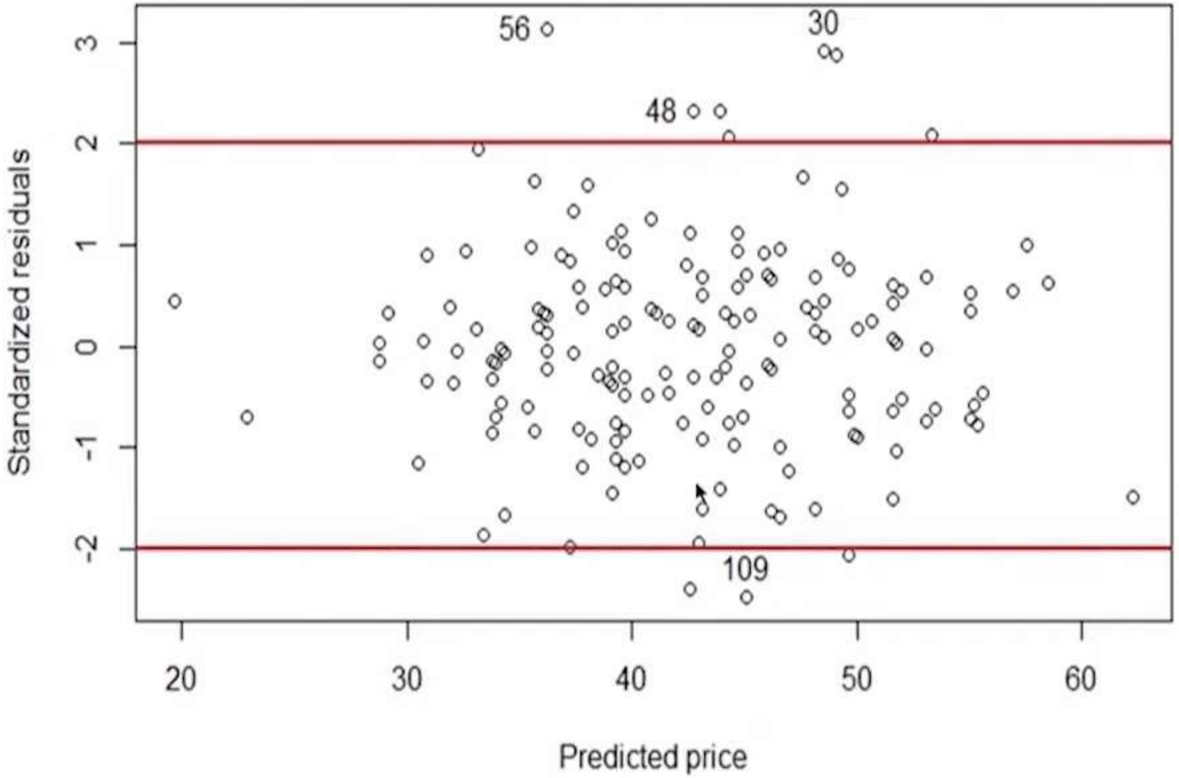
Multiple Linear Regression: Diagnostics

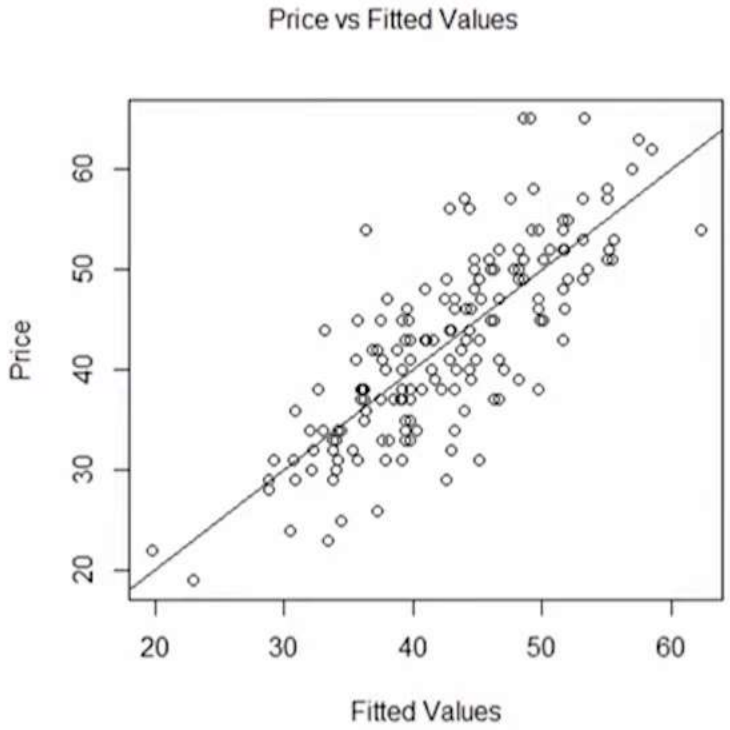
- Residual plots: Standardized residuals for assessing
  - Linear vs nonlinear model
  - Normality of the errors
  - Homoscedastic vs heteroscedastic errors

Similar to Simple regression

Multiple Linear Regression: Testing for linearity

- Residuals plot: standardized residuals vs fitted values





No Pattern: Based on this and other measures (R2, F-test) we can conclude that a linear model is acceptable