AMA3602

Applied Linear Models

Department of Applied Mathematics

Lecturer : Dr. Catherine Liu

Lecture Time: 17:30-18:20, Mon. &. 12:30-13:20, Tues.

01-02/2024

Chapter **2**

Multiple Linear Regression

Reference: Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). Chapter 3: Multiple Linear Regression. *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

MULTIPLE LINEAR REGRESSION

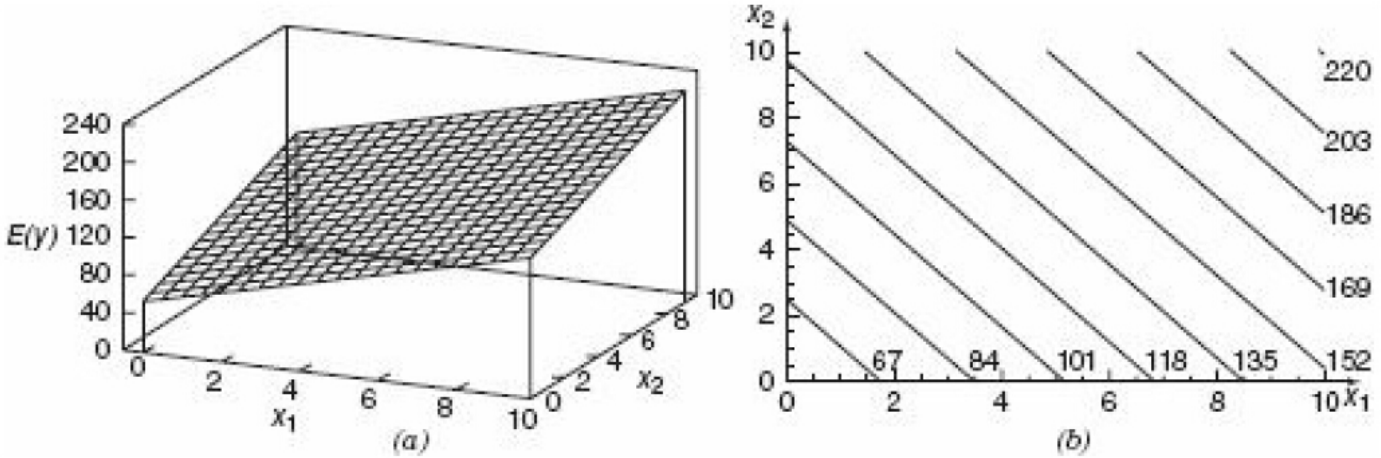
Multiple Regression Models

Multiple linear regression (MLR) model

- **Multiple regression model**: A regression model that involves more than one regressor variable.
- Example: A **multiple linear regression model** with two regressor:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where y denotes the yield, x_1 denotes the temperature, and x_2 denotes the catalyst concentration.
The term **linear** is used because the aforementioned equation is a linear function of the unknown parameters β_0, β_1, and β_2.



- $E(y) = 50 + 10x_1 + 7x_2$, where we assume $E(\varepsilon) = 0$;
- Figure (a)
  - ⋆ The parameter β_0 is the intercept of the regression plane. If the range of the data includes $x_1 = x_2 = 0$, then β_0 is the mean of y when $x_1 = x_2 = 0$. Otherwise β_0 has no physical interpretation;
  - ⋆ The parameter β_1 indicates the expected change in response (y) per unit change in x_1 when x_2 is held constant;
  - ⋆ β_2 measures the expected change in y per unit change in x_2 when x_1 is held constant.
- Figure (b) shows a **contour plot** of the regression model, that is, lines of constant expected response $E(y)$ as a function of x_1 and x_2. Notice that the contour lines in this plot are parallel straight lines.

General Form

- Sample MLR Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$, $i = 1, \ldots, n$.
- Assumption: $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$, $i = 1, \ldots, n$.
- In matrix notation, the model is given by:

$$y_{n\times 1} = X_{n\times p} \cdot \beta_{p\times 1} + \varepsilon_{n\times 1}$$

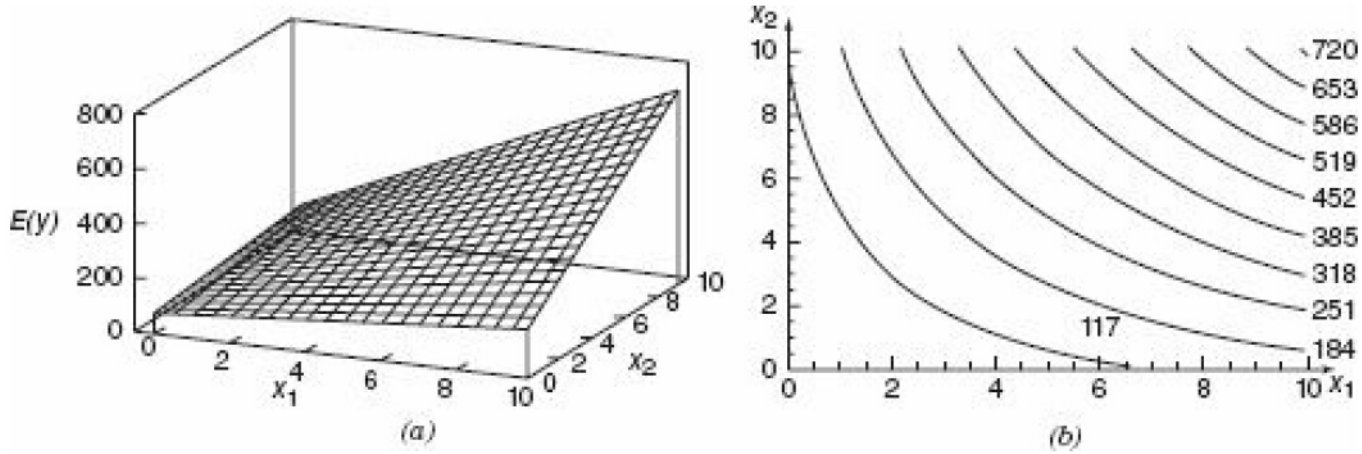
  where $p = k + 1$.
- Assumption: $\mathrm{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I_{n\times n}$.

Remark:

- The parameters β_j, $j = 0, 1, \ldots, k$ are called the **regression coefficients**.
- This model describes a hyperplane in the k-dimensional space of the regressor variables x_j.
- The parameters β_j represents the expected change in the response y per unit change in x_j when all of the remaining regressor variables x_i $(i \neq j)$ are held constant. For this reason, the parameters β_j, $j = 0, 1, \ldots, k$ are often called **partial regression coefficients**.
- MLR models are often used as **empirical models** or **approximating functions**. That is, the true functional relationship between y and x_1, x_2,..., x_k is unknown, but over certain ranges of the regressor variables the linear regression model is an adequate approximation to the true unknown function.

Models that are more complex in structure may often still be analyzed by multiple linear regression techniques

1. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$
2. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$



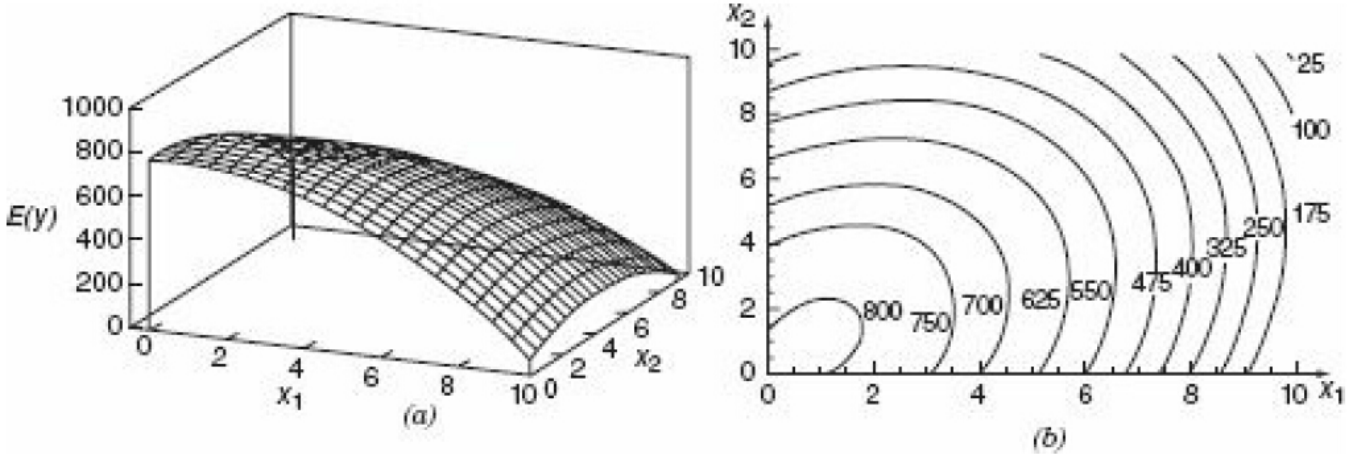
- $E(y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$, where we assume $E(\varepsilon) = 0$;
- Although this model is a linear regression model, the shape of the surface that is generated by the model is not linear.
- In general, **any regression model that is linear in the parameters** (the β's) **is a linear regression model, regardless of the shape of the surface that is generates**.
- Figure provides a nice graphical interpretation of an interaction. Generally, interaction implies that the effect produces by changing one variable (x_1) depends on the level of the other variable (x_2)

Consider the **second-order model with interaction**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

If we let $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1 x_2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$, and $\beta_5 = \beta_{12}$, then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$



- $E(y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1 x_2$, where we assume $E(\varepsilon) = 0$;
- These plots indicate that the expected change in y when x_1 is changed by one unit is a function of both x_1 and x_2.
- The quadratic and interaction terms in this model produce a mound-shaped function.
- Depending on the values of the regression coefficients, the second-order model with interaction is capable of assuming a wide variety of shapes; thus, it is a very flexible regression model.

Matrix Approach to SLR & MLR Analysis

⋆ Sample SLR Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i = (1\ x_{i1}) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \varepsilon_i, \quad i = 1, \cdots, n$$

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times 2} \boldsymbol{\beta}_{2\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

- $\mathbf{y}_{n\times 1}$: response vector
- $\mathbf{X}_{n\times 2}$: design matrix
- $\boldsymbol{\beta}_{2\times 1}$: coefficient vector
- $\boldsymbol{\varepsilon}_{n\times 1}$: model error vector

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1\ x_{11} \\ 1\ x_{21} \\ \vdots \\ 1\ x_{n1} \end{pmatrix}_{n\times 2} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

⋆ Sample MLR Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$= (1\ x_{i1}\ x_{i2} \cdots x_{ik}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \varepsilon_i, \quad i = 1, \cdots, n$$

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p} \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \text{where } p = k+1$$

- $\mathbf{y}_{n\times 1}$: response vector
- $\mathbf{X}_{n\times p}$: design matrix
- $\boldsymbol{\beta}_{p\times 1}$: coefficient vector
- $\boldsymbol{\varepsilon}_{n\times 1}$: model error vector

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1\ x_{11}\ x_{12} \cdots x_{1k} \\ 1\ x_{21}\ x_{22} \cdots x_{2k} \\ \vdots \\ 1\ x_{n1}\ x_{n2} \cdots x_{nk} \end{pmatrix}_{n\times p} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{p\times 1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

⋆ Assumptions on ε for SLR

$$\mathrm{E}(\boldsymbol{\varepsilon}_{n\times 1}) = \begin{pmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_n) \end{pmatrix}_{n\times 1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n\times 1} = \mathbf{0}_{n\times 1}$$

$$\mathrm{Var}(\boldsymbol{\varepsilon}_{n\times 1}) = \begin{pmatrix} \mathrm{Cov}(\varepsilon_1,\varepsilon_1) & \mathrm{Cov}(\varepsilon_1,\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_1,\varepsilon_n) \\ \mathrm{Cov}(\varepsilon_2,\varepsilon_1) & \mathrm{Cov}(\varepsilon_2,\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_2,\varepsilon_n) \\ & & \vdots & \\ \mathrm{Cov}(\varepsilon_n,\varepsilon_1) & \mathrm{Cov}(\varepsilon_n,\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_n,\varepsilon_n) \end{pmatrix}_{n\times n}$$

$$= \begin{pmatrix} \sigma^2 & \cdots & 0 \\ & \vdots & \\ 0 & \cdots & \sigma^2 \end{pmatrix}_{n\times n} = \sigma^2 \mathbf{I}_{n\times n}$$

⋆ Assumptions on ε for MLR

$$\mathrm{E}(\boldsymbol{\varepsilon}_{n\times 1}) = \begin{pmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_n) \end{pmatrix}_{n\times 1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n\times 1} = \mathbf{0}_{n\times 1}$$

$$\mathrm{Var}(\boldsymbol{\varepsilon}_{n\times 1}) = \begin{pmatrix} \mathrm{Cov}(\varepsilon_1,\varepsilon_1) & \mathrm{Cov}(\varepsilon_1,\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_1,\varepsilon_n) \\ \mathrm{Cov}(\varepsilon_2,\varepsilon_1) & \mathrm{Cov}(\varepsilon_2,\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_2,\varepsilon_n) \\ & & \vdots & \\ \mathrm{Cov}(\varepsilon_n,\varepsilon_1) & \mathrm{Cov}(\varepsilon_n,\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_n,\varepsilon_n) \end{pmatrix}_{n\times n}$$

$$= \begin{pmatrix} \sigma^2 & \cdots & 0 \\ & \vdots & \\ 0 & \cdots & \sigma^2 \end{pmatrix}_{n\times n} = \sigma^2 \mathbf{I}_{n\times n}$$

⋆ Least-squares Estimation for SLR

$$\mathbf{X}_{2\times n}^T \mathbf{X}_{n\times 2} \hat{\beta}_{2\times 1} = \mathbf{X}_{2\times n}^T \mathbf{y}_{n\times 1}$$

$$\hat{\beta}_{2\times 1} = (\mathbf{X}^T\mathbf{X})_{2\times 2}^{-1} \mathbf{X}_{2\times n}^T \mathbf{y}_{n\times 1}$$

(provided that the inverse matrix exists.)

⋆ Fitted Values

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times 2}\hat{\beta}_{2\times 1}$$

⋆ Hat Matrix

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times 2} (\mathbf{X}^T\mathbf{X})_{2\times 2}^{-1} \mathbf{X}_{2\times n}^T \mathbf{y}_{n\times 1}$$

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{H}_{n\times n} \mathbf{y}_{n\times 1}$$

$$\mathbf{H}_{n\times n} = \mathbf{X}_{n\times 2} (\mathbf{X}^T\mathbf{X})_{2\times 2}^{-1} \mathbf{X}_{2\times n}^T$$

- Remark: $\mathbf{H}$ is idempotent since $\mathbf{H}\,\mathbf{H} = \mathbf{H}$.

⋆ Residuals

$$\hat{\mathbf{e}}_{n\times 1} = \mathbf{y}_{n\times 1} - \hat{\mathbf{y}}_{n\times 1}$$

$$= \mathbf{y}_{n\times 1} - \mathbf{X}_{n\times 2}\hat{\beta}_{2\times 1}$$

⋆ Least-squares Estimation for MLR

$$\mathbf{X}_{p\times n}^T \mathbf{X}_{n\times p} \hat{\beta}_{p\times 1} = \mathbf{X}_{p\times n}^T \mathbf{y}_{n\times 1}$$

$$\hat{\beta}_{p\times 1} = (\mathbf{X}^T\mathbf{X})_{p\times p}^{-1} \mathbf{X}_{p\times n}^T \mathbf{y}_{n\times 1}$$

(provided that the inverse matrix exists.)

⋆ Fitted Values

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times p}\hat{\beta}_{p\times 1}$$

⋆ Hat Matrix

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times p} (\mathbf{X}^T\mathbf{X})_{p\times p}^{-1} \mathbf{X}_{p\times n}^T \mathbf{y}_{n\times 1}$$

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{H}_{n\times n} \mathbf{y}_{n\times 1}$$

$$\mathbf{H}_{n\times n} = \mathbf{X}_{n\times p} (\mathbf{X}^T\mathbf{X})_{p\times p}^{-1} \mathbf{X}_{p\times n}^T$$

- Remark: $\mathbf{H}$ is idempotent since $\mathbf{H}\,\mathbf{H} = \mathbf{H}$.

⋆ Residuals

$$\hat{\mathbf{e}}_{n\times 1} = \mathbf{y}_{n\times 1} - \hat{\mathbf{y}}_{n\times 1}$$

$$= \mathbf{y}_{n\times 1} - \mathbf{X}_{n\times p}\hat{\beta}_{p\times 1}$$

SLR $\Rightarrow$ MLR

k = 2 $\Rightarrow$ k = p (p $>$ 2)

Least-Squares Estimation

Observation, i	Response, y	Regressors			
		x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

- Sample regression model:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n. \ (n > k) \end{aligned}$$

- Assumptions: $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$, $i, j = 1, \ldots, n.$

- The least-squares function is

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

- The **least-squares normal equations**:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik} y_i$$

- The solution to the normal equations will be the **least-squares estimators** $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

In **matrix** notation

- $\mathbf{y} = \mathbf{X}\beta + \varepsilon$,
  where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- In general, $\mathbf{y}$ is an $n \times 1$ vector of the observations, $\mathbf{X}$ is an $n \times p$ matrix of the levels of the regressor variables, β is a $p \times 1$ vector of the regression coefficients, and ε is an $n \times 1$ vector of random errors.

- To find the **vector of least-squares estimators**, $\hat{\beta}$ that minimizes:

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- Normal equations

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

  $\Rightarrow$ The **least-squares normal equations**:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

- The **least-squares estimator** of $\boldsymbol{\beta}$, provided that the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$ exists, is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Remark: The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ will always exist if the regressors are **linearly independent**, that is, if no column of the $\mathbf{X}$ matrix is a linear combination of the other columns.

- From the least-squares normal equations, we obtain

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

Remark:

- In this display we see that $\mathbf{X}'\mathbf{X}$ is a $p \times p$ symmetric matrix and $\mathbf{X}'\mathbf{y}$ is a $p \times 1$ column vector;
- Note the special structure of the $\mathbf{X}'\mathbf{X}$ matrix. The **diagonal elements** of $\mathbf{X}'\mathbf{X}$ are the sums of squares of the elements in the columns of $\mathbf{X}$, and the **off-diagonal** elements are the sums of cross products of the elements in the columns of $\mathbf{X}$;
- **The elements of $\mathbf{X}'\mathbf{y}$** are the sums of cross products of the columns of $\mathbf{X}$ and the observations y_i.

- The fitted regression model corresponding to the levels of the regressor variables $\mathbf{x}' = [1, x_1, x_2, \ldots, x_k]$ is

$$\hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$$

- The vector of fitted values $\hat{y}_i$ corresponding to the observed values y_i is

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

  where the $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the **hat matrix**. It maps the vector of observed values into a vector of fitted values.The hat matrix and its properties play a central role in regression analysis.

- The different between the observed value y_i and the corresponding fitted value $\hat{y}_i$ is the residual $e_i = y_i - \hat{y}_i$. The n residuals may be conveniently written in matrix notation is

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

  There are several other ways to express the vector of residuals $\mathbf{e}$ that will prove useful, including

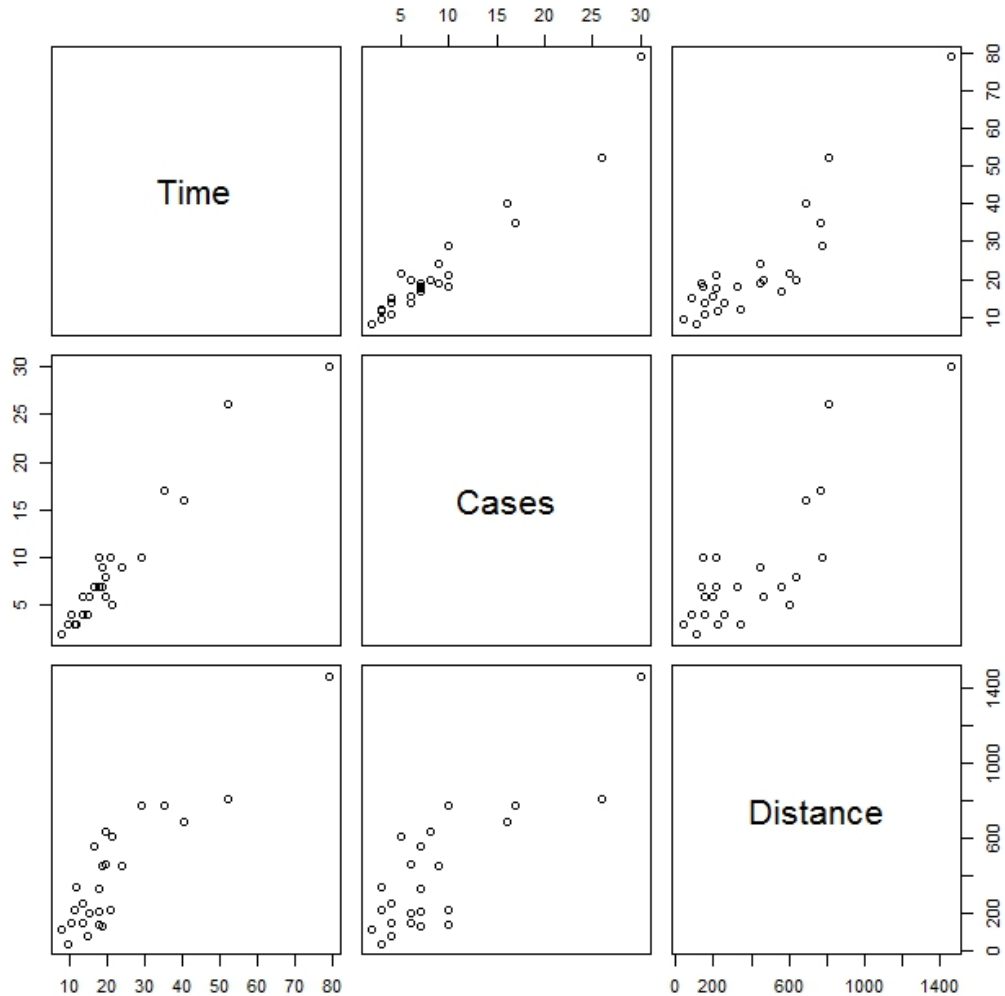
$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Example: the delivery time data

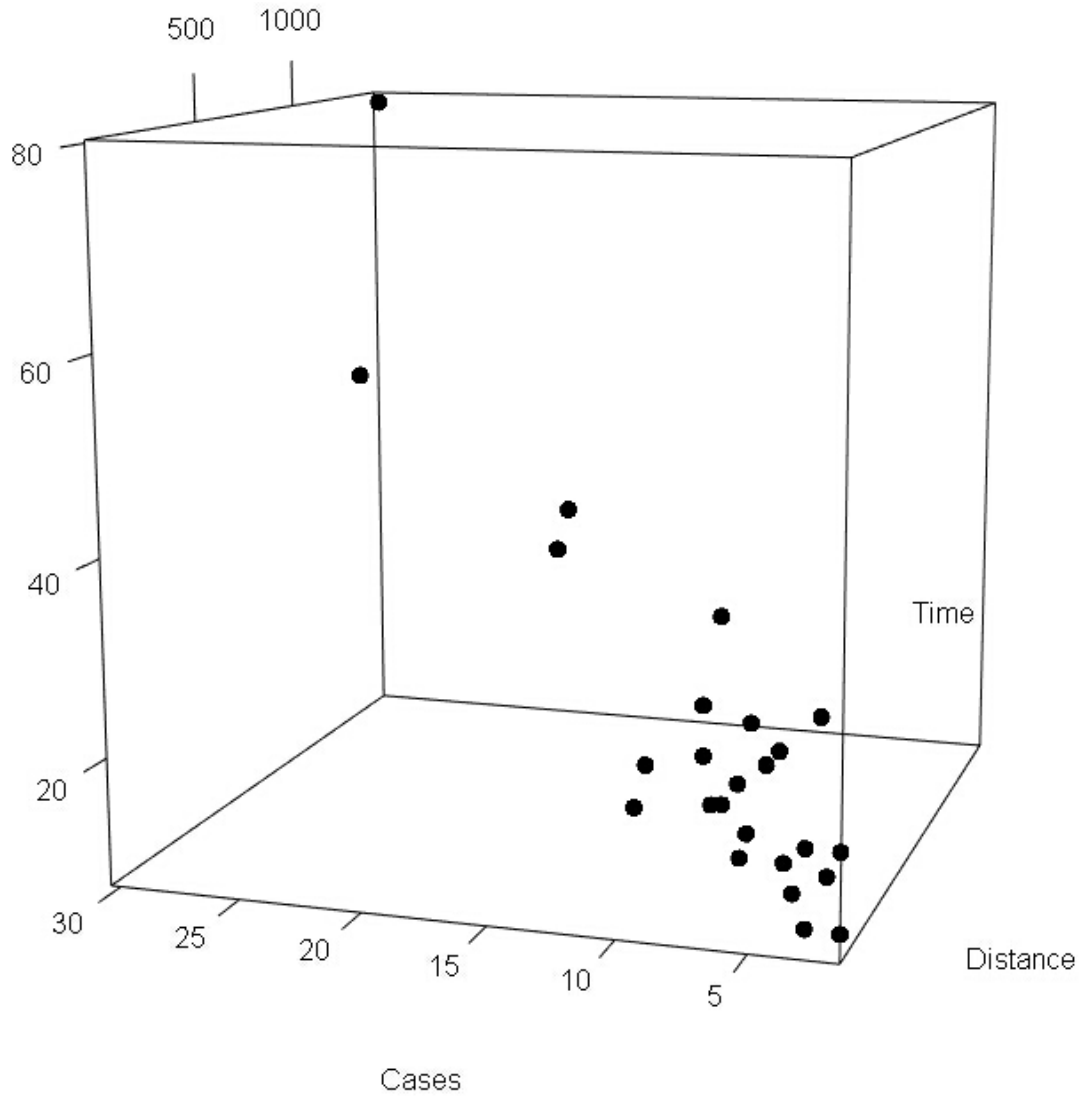
- A soft drink bottler is analyzing the vending machine service routes in his distribution system, and he is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet.
- y: the delivery time;
  x_1: the No. of cases of product stocked;
  x_2: the distance walked by the route driver.

Observation No., i	Delivery Time, y_i (psi)	No. of Cases, x_1	Distance, x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

```
> setwd("C:/Users/user/Desktop/AMA514_2016/AMA514_2016_new/Data")
> E3.1<-read.table("data_example_3.1.txt", header=T)
> #E3.1
> pairs(~ Time + Cases + Distance, data=E3.1)
```



```
> library(rgl)
> plot3d(E3.1$Distance, E3.1$Cases, E3.1$Time, xlab="Distance", ylab="Cases", zlab="Time", size=10)
```



```
> ## Simplify notation
> y<-E3.1$Time; x1<-E3.1$Cases; x2<-E3.1$Distance
> ## a. Fit a multiple linear regression
> res<-lm(y~x1+x2)
> summary(res)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
x1          1.615907   0.170735   9.464 3.25e-09 ***
x2          0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

- The least-squares fit is $\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2$

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	16.68	21.7081	−5.0281
2	11.50	10.3536	1.1464
3	12.03	12.0798	−0.0498
4	14.88	9.9556	4.9244
5	13.75	14.1944	−0.4444
6	18.11	18.3996	−0.2896
7	8.00	7.1554	0.8446
8	17.83	16.6734	1.1566
9	79.24	71.8203	7.4197
10	21.50	19.1236	2.3764
11	40.33	38.0925	2.2375
12	21.00	21.5930	−0.5930
13	13.50	12.4730	1.0270
14	19.75	18.6825	1.0675
15	24.00	23.3288	0.6712
16	29.00	29.6629	−0.6629
17	15.35	14.9136	0.4364
18	19.00	15.5514	3.4486
19	9.50	7.7068	1.7932
20	35.10	40.8880	−5.7880
21	17.90	20.5142	−2.6142
22	52.32	56.0065	−3.6865
23	18.75	23.3576	−4.6076
24	19.83	24.4028	−4.5728
25	10.75	10.9626	−0.2126

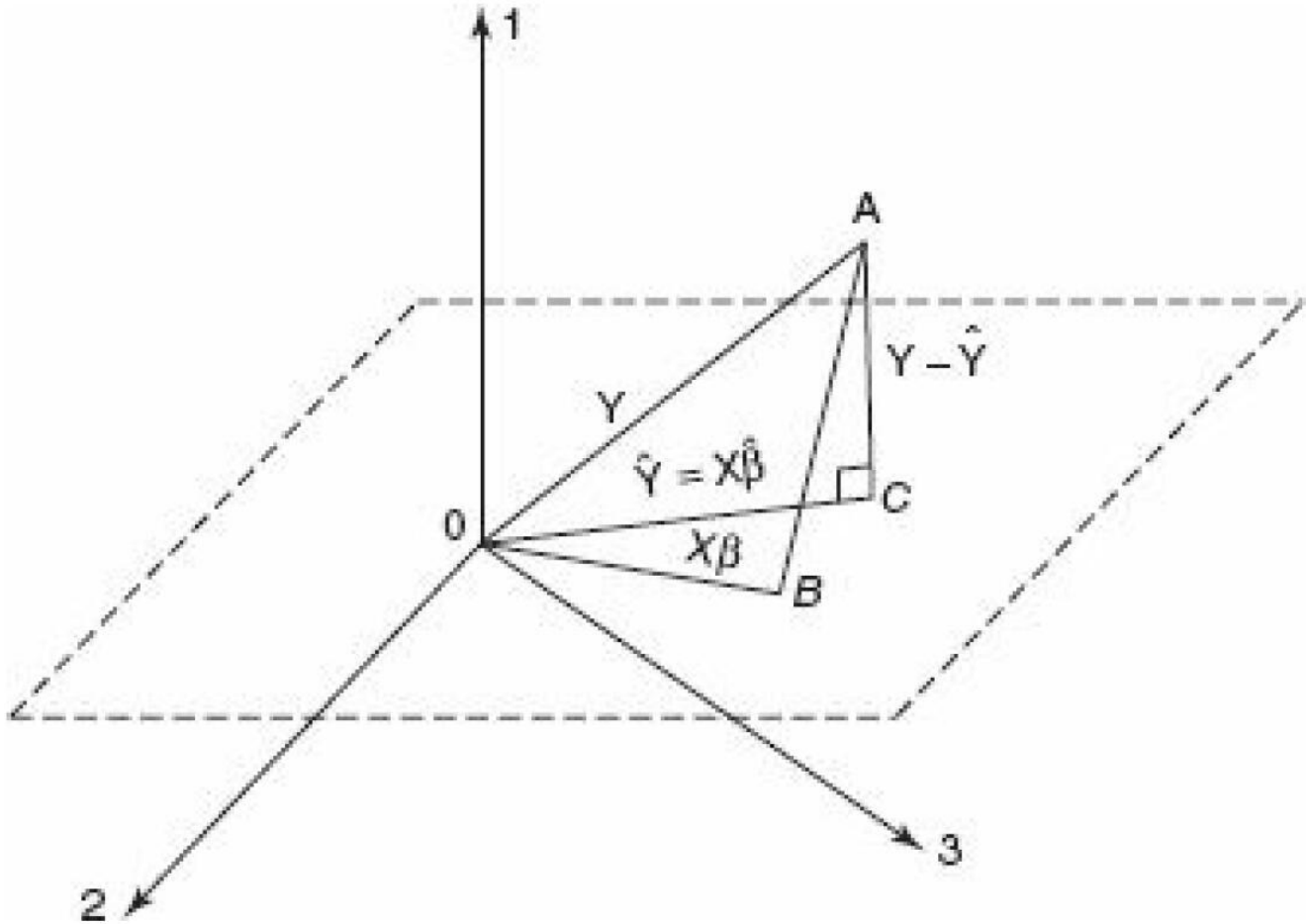
Remark: This table shows the observations y_i along with the corresponding fitted values $\hat{y}_i$ and the residuals e_i from this model.

An intuitive **geometrical interpretation of least squares** is sometimes helpful.

- We may think of the vector of observations $\mathbf{y}' = [y_1, y_2, \ldots, y_n]$ as defining a vector from the origin to the point A in the following figure. Note that $y_1, y_2, \ldots, y_n$ form the coordinates of an n-dimensional sample space.
- The $\mathbf{X}$ matrix consists of $p(n \times 1)$ column vectors, for example, $1, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Each of these columns defines a vector from the origin in the sample space. These p vectors form a p-dimensional subspace called the **estimation space**. The estimation space for $p = 2$ is shown in the following figure. We may represent any point in this subspace by a linear combination of the vectors $1, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Thus, any point in the estimation space is of the form $\mathbf{X}\boldsymbol{\beta}$. Let the vector $\mathbf{X}\boldsymbol{\beta}$ determine the point B. The squared distance from B to A is just

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- Therefore, minimizing the squared distance of point A defined by the observation vector $\mathbf{y}$ to the estimation space requires finding the point in the estimation space that is closest to A. The squared distance is a minimum when the point in the estimation space is the foot of the line from A normal (or perpendicular) to the estimation space. This is point C in the figure. This point is defined by the vector $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ Therefore, since $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is perpendicular to the estimation space, we may write $\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$ or $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ which we recognize as the least-squares normal equations.


1
A
Y − Ŷ
Y
Ŷ = Xβ̂
C
0
Xβ
B
3
2

Properties of the least-squares estimators

- Gauss-Markov theorem:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] \\ &= \boldsymbol{\beta} \end{aligned}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$.
Thus, $\hat{\boldsymbol{\beta}}$ is an **unbiased estimator** of $\boldsymbol{\beta}$ if the model is correct.

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- LSE of $\hat{\boldsymbol{\beta}}$ is best linear unbiased estimators (BLUE) of $\boldsymbol{\beta}$.
- Remark: if we let $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$, then

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}, \quad \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$$

Estimation of σ^2

- $SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$
- Substituting $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$

$$\begin{aligned} SS_{Res} &= (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\beta} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\beta}'\mathbf{X}'\mathbf{y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \end{aligned}$$

- Since $\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$,

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$$

- **Residual mean squares**: $MS_{Res} = \frac{SS_{Res}}{n-p}$.
- Remark: an **unbiased estimator** of σ^2 is given by: $\hat{\sigma}^2 = MS_{Res}$. As noted in the simple linear regression case, this estimator of σ^2 is **model dependent**.

```
> anova(res)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 5382.4  5382.4 506.619 < 2.2e-16 ***
x2         1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $\hat{\sigma}^2 = \frac{SS_{Res}}{n-p} = \frac{233.7}{22} = 10.6$

```
> ### Example_2.9
> setwd("C:/Users/user/Desktop/AMA514_2016/R_textbook/Chapter1_Jan2/Examples/Example2.9")
> E2.9<-read.table("data-ex-2-9.txt", header=T)
> res<-lm(y~x, data=E2.9)
> ## b. ANOVA
> anova(res)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1 5382.4  5382.4  307.85 8.22e-15 ***
Residuals 23  402.1    17.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

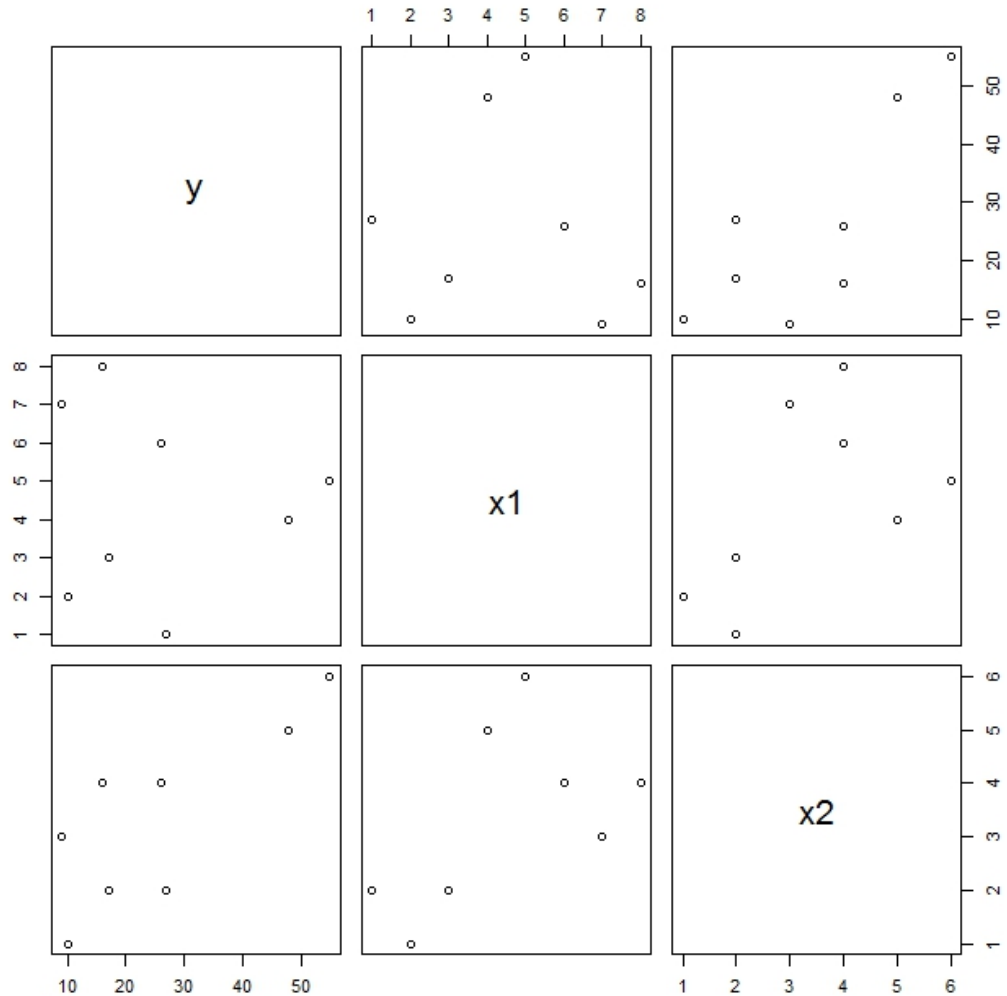
Remark:

- The above figure displays a least-square fit to the delivery time data using only one regressor, cases (x_1). The residual mean square for this model is 17.5, which is considerably larger than the result obtained above for the two regressor model. Which estimate is "correct"?
- Both estimates are in a sense correct, but they depend heavily on the choice of model. Perhaps a better question is which **model** is correct?
- Since σ^2 is the variance of the errors (the unexplained noise about the regression line), we would usually prefer a model with a small residual mean square to a model with a large one.

Inadequacy of Scatter Diagrams in Multiple Regression

```
> ### Figure_3.7
> setwd("C:/Users/user/Desktop/AMA514_2016/AMA514_2016_new/Data")
> Fg3.7<-read.table("Figure_3.7.txt", header=T)
> pairs(~ y + x1
```

y	x_1	x_2
10	2	1
17	3	2
48	4	5
27	1	2
55	5	6
26	6	4
9	7	3
16	8	4



```
> ## a. Fit a multiple linear regression
> res<-lm(y~x1+x2, data=Fg3.7)
> summary(res)

Call:
lm(formula = y ~ x1 + x2, data = Fg3.7)

Residuals:
         1          2          3          4          5          6          7          8
-5.638e-16 -1.765e-15 -6.076e-15  3.196e-15  4.453e-15 -4.379e-16  1.661e-15 -4.668e-16

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) 8.000e+00  3.462e-15  2.311e+15   <2e-16 ***
x1          -5.000e+00  7.013e-16 -7.129e+15   <2e-16 ***
x2           1.200e+01  1.019e-15  1.177e+16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.836e-15 on 5 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.974e+31 on 2 and 5 DF,  p-value: < 2.2e-16
```

- $y = 8 - 5x_1 + 12x_2$

Remark:

- This example illustrates that constructing scatter diagrams of y versus x_j $(j = 1, 2, \ldots, k)$ can be misleading, even in the case of only two regressors operating in a perfectly additive fashion with no noise.
- A more realistic regression situation with several regressors and error in the y's would confuse the situation even further. If there is only one (or a few) dominant regressor, or if the regressors operate nearly independently, the matrix of scatterplots is most useful. However, when several important regressors are themselves interrelated, then these scatter diagrams can be very misleading.

Maximum-likelihood estimation

- Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- Assumption: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- $f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right)$
- $L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)$
- Since $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, then

$$L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

$$\Rightarrow \qquad \ln L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = -\tfrac{n}{2}\ln(2\pi) - n\ln(\sigma) - \tfrac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- It is clear that for a fixed value of σ the log-likelihood is maximized when the term $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is minimized.
- Therefore, **the maximum-likelihood estimator of β under normal errors is equivalent to the least-squares estimator** $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
  The maximum-likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n}$$

Hypothesis Testing in MLR

- Test problems and methods:
  1. Test for significance of regression - ANOVA
  2. Tests on individual regression coefficients - (partial) t-test
  3. Tests on subsets of coefficient - partial F-test
- **MLR model**: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$, $i = 1, \ldots, n$.
  $\Leftrightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- Assumption: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

Test problem 1: Test for significance of regression

- It is an overall or global test of model adequacy, and to determine if there is a linear relationship between the response y and any of the regressor variables $x_1, x_2, \ldots, x_k$
- Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0, \text{for at least one } j, j = 1, \ldots, k.$$

- The test procedure is a generalization of the analysis of variance used in SLR. The total sum of squares SS_T is partitioned into a sum of squares due to regression, SS_R, and a residual sum of squares, SS_{Res}. Thus, $SS_T = SS_R + SS_{Res}$

$$SS_R = \hat{\beta}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \text{ and } SS_R/\sigma^2 \overset{H_0}{\sim} \chi^2_k$$

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}, \text{ and } SS_{Res}/\sigma^2 \overset{H_0}{\sim} \chi^2_{n-k-1}$$

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \text{ and } SS_R/\sigma^2 \overset{H_0}{\sim} \chi^2_{n-1}$$

- Under H_0:

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} \sim F_{k,n-k-1},\ SS_R \perp SS_{Res}$$

- $E(MS_{Res}) = \sigma^2$, $E(MS_R) = \sigma^2 + \frac{\beta^{*\prime}\mathbf{X}_c'\mathbf{X}_c\beta^*}{k\sigma^2}$
  where $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_k)'$ and $\mathbf{X}_c$ is the "**centered**" model matrix given by

$$\mathbf{X}_c = \begin{bmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1k} - \overline{x}_k \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{2k} - \overline{x}_k \\ \vdots & \vdots & & \vdots \\ x_{i1} - \overline{x}_1 & x_{i2} - \overline{x}_2 & \cdots & x_{ik} - \overline{x}_k \\ \vdots & \vdots & & \vdots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & \cdots & x_{nk} - \overline{x}_k \end{bmatrix}$$

- If at least one $\beta_j \neq 0$, then F_0 follows a noncentral F distribution with k and $n-k-1$ degrees of freedom and a noncentrality parameter of $\frac{\beta^{*\prime}\mathbf{X}_c'\mathbf{X}_c\beta^*}{\sigma^2}$
- Reject H_0 if $F_0 > F_{\alpha,k,n-k-1}$.

ANOVA

ANOVA for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	$MS_R = SS_R / k$	MS_R / MS_{Res}
Residuals	SS_{Res}	$n-k-1$	$MS_{Res} = SS_{Res} / (n-k-1)$	
Total	SS_T	$n-1$		

Example: the delivery time data

- A soft drink bottler is analyzing the vending machine service routes in his distribution system, and he is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet.
- y: the delivery time;
  x_1: the No. of cases of product stocked;
  x_2: the distance walked by the route driver.

Observation No., i	Delivery Time, y_i (psi)	No. of Cases, x_1	Distance, x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

```
> setwd("C:/Users/YANG Jin/Desktop/AMA514_2016_new/Data")
> E3.3<-read.table("data_example_3.1.txt", header=T)
> #E3.3
> ## Simplify notation
> y<-E3.3$y; x1<-E3.3$x1; x2<-E3.3$x2
> n<-25
> I<-rep(1, times=n)
> X<-cbind(I,x1,x2)
> beta<-solve(t(X)%*%X)%*%t(X)%*%y        ## \hat beta
>
> t(y)%*%y-(sum(y))^2/n                   ## SS_T
          [,1]
[1,] 5784.543
> t(beta)%*%t(X)%*%y-(sum(y))^2/n         ## SS_R
          [,1]
[1,] 5550.811
> t(y)%*%y-t(beta)%*%t(X)%*%y             ## SS_Res
          [,1]
[1,] 233.7317

> ## ANOVA
> A<-cbind(x1,x2)
> res<-lm(y~A)
> anova(res)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          2 5550.8 2775.41  261.24 4.687e-16 ***
Residuals 22  233.7   10.62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- To test $H_0 : \beta_1 = \beta_2 = 0$ vs. $H_1 : \beta_i \neq 0,\ i = 1, 2$, we calculate the statistic

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{2775.41}{10.62} = 261.24$$

- Since the p-value is very small, we conclude that delivery time is related to delivery volume and/or distance. However, this does not necessarily imply that the relationship found is an appropriate one for predicting delivery time as a function of volume and distance.
- Further tests of model adequacy are required.

R^2 and Adjusted R^2

```
> ## R^2 and Adjusted R^2
> summary(res)$r.squared
[1] 0.9595937
> summary(res)$adj.r.squared
[1] 0.9559205
```

- Two other ways to assess the overall adequacy of the model are R^2 and the adjusted R^2, denoted R^2_{Adj}.
- In general, R^2 never decreases when a regressor is added to the model, regardless of the value of the contribution of that variable. Therefore, it is difficult to judge whether an increase in R^2 is really telling us anything important.
- Some regression model builders prefer to use an **adjusted R^2** statistic, defined as

$$R^2_{\text{Adj}} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}$$

Since $SS_{Res}/(n-p)$ is the residual mean square and $SS_T/(n-1)$ is constant regardless of how many variables are in the model, R^2_{Adj} will only increase on adding a variable to model if the addition of the variable reduces the residual mean square.

Remark: In subsequent chapters, when we discuss model building and variable selection, it is frequently helpful to have a procedure that can guard against overfitting the model, that is, adding terms that are unnecessary. The adjusted R^2 penalizes us for adding terms that are not helpful, so it is very useful in evaluating and comparing candidate regression models.

Test problem 2: Tests on individual regression coefficients

- Reason:
  - ⋆ Adding a variable to a regression model always causes SS_R to increase and SS_{Res} to decrease;
  - ⋆ The addition of a regressor also increases the variance of $\hat{\boldsymbol{y}}$;
  - ⋆ Adding an unimportant regressor may increase the MS_{Res}, which may decrease the usefulness of the model.
- Individual regression coefficients
  - ⋆ Hypothesis: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$, $j = 1, \ldots, k$
  - ⋆ Under H_0,

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1}$$

    where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$.
  - ⋆ Reject H_0 if $|t_0| > t_{\alpha/2, n-k-1}$

Remark: This is really a **partial** or **marginal** test because the regression coefficient $\hat{\beta}_j$ depends on all of the other regressor variables x_i $(i \neq j)$ that are in the model. Thus, this is **a test of the contribution of x_j given the other regressors in the model**.

Example: the delivery time data

- Suppose we wish to assess the value of the regressor variable x_2 given that the regressor x_1 is in the model.

```
> res<-lm(y~x1+x2)
> summary(res)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
x1          1.615907   0.170735   9.464 3.25e-09 ***
x2          0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

```
> ## ANOVA
> A<-cbind(x1,x2)
> res<-lm(y~A)
> anova(res)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          2 5550.8 2775.41  261.24 4.687e-16 ***
Residuals 22  233.7   10.62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> solve(t(X)%*%X)
                  I            x1            x2
I   1.132152e-01 -4.448593e-03 -8.367257e-05
x1 -4.448593e-03  2.743783e-03 -4.785709e-05
x2 -8.367257e-05 -4.785709e-05  1.228745e-06
```

- Hypotheses: $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$
- Under H_0,

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{0.0144}{\sqrt{(10.62)(0.000001228)}} = 3.981$$

- Since $t_{0.025,22} = 2.074$, we reject H_0 and conclude that the regressor x_2 contributes significantly to the model given that x_1 is also in the model.

Test problem 3: Tests on subsets of coefficients

Aim: To investigate the contribution of a subset ($\boldsymbol{\beta}_2$ below) of the regressor variables to the MLR model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

- Consider the regression model with k regressors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$, $\boldsymbol{\varepsilon}$ is $n \times 1$, and $p = k + 1$.

- We would like to determine if some subset of $r < k$ regressors contributes significantly to the regression model. Let the vector of regression coefficients be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \hline \boldsymbol{\beta}_2 \end{bmatrix}$$

where $\boldsymbol{\beta}_1$ is $(p - r) \times 1$ and $\boldsymbol{\beta}_2$ is $r \times 1$.

- The **full model**:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

where the $n \times (p - r)$ matrix $\boldsymbol{X}_1$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_1$ and the $n \times r$ matrix $\boldsymbol{X}_2$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_2$.

- For the full model, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, then

$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ (p degrees of freedom) and $MS_{Res} = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n-p}$

- To find the contribution of the terms of β_2 to the regression, fit the model assuming that the null hypothesis $H_0 : \beta_2 = 0$ is true. We wish to test the hypotheses:

$$H_0 : \beta_2 = 0 \text{ vs. } H_1 : \beta_2 \neq 0$$

- Under H_0, the **reduced model** is $\mathbf{y} = \boldsymbol{X}_1 \beta_1 + \varepsilon$
- $SS_R(\beta_1) = \hat{\beta}_1' \mathbf{X}_1' \mathbf{y}$ $(df = p - r)$
- The regression sum of squares due to β_2 given that β_1 is already in the model is

$$SS_R(\beta_2 | \beta_1) = SS_R(\beta) - SS_R(\beta_1)$$

with $df = p - (p - r) = r$. This sum of squares is called the **extra sum of squares due to β_2** because it measures the increase in the regression sum of squares that results from adding the regressors $x_{k-r+1}, x_{k-r+2}, \ldots, x_k$ to a model that already contains $x_1, x_2, \ldots, x_{k-r}$.

- **Partial F-test**: $SS_R(\beta_2|\beta_1)$ is independent of MS_{Res}, and Under H_0,

$$F_0 = \frac{SS_R(\beta_2|\beta_1)/r}{MS_{Res}} \sim F_{r,n-p}$$

- Reject H_0 if $F_0 > F_{\alpha,r,n-p}$, concluding that at least one of the parameters in β_2 is not zero, and consequently at least one of the regressors $x_{k-r+1}, x_{k-r+2}, \ldots, x_k$ in $\boldsymbol{X}_2$ contribute significantly to the regression model.

Remark:

- Partial F-test measures the contribution of the regressors in $\boldsymbol{X}_2$ given that the other regressors in $\boldsymbol{X}_1$ are in the model;
- Partial F-test plays a major role in model building, that is, in searching for the best set of regressors to use in the model;
- If $\beta_2 \neq 0$, that is, under H_1 then F_0 follows a noncentral F distribution with a noncentrality parameter of

$$\lambda = \frac{1}{\sigma^2}\beta_2'\boldsymbol{X}_2'[I - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1']\boldsymbol{X}_2\beta_2$$

  This result is quite important. If there is multicollinearity in the data, there are situations where β_2 is markedly nonzero, but this test actually has almost no power (ability to indicate this difference) because of a near-collinear relationship between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. In this situation, λ is nearly zero even though β_2 is truly important. This relationship also points out that the maximal power for this test occurs when $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, are orthogonal, i.e. $\boldsymbol{X}_2'\boldsymbol{X}_1 = \boldsymbol{0}$.

- Consider the model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
- $SS_R(\beta_1|\beta_0, \beta_2, \beta_3)$, $SS_R(\beta_2|\beta_0, \beta_1, \beta_3)$, $SS_R(\beta_3|\beta_0, \beta_1, \beta_2)$ are single-degree-of-freedom sums of squares that measure the contribution of each regressor x_j, $j = 1, 2, 3$, to the model given that all of the other regressors were already in the model. That is, we are assessing the value of adding x_j to a model that did not include this regressor.
- In general, we could find $SS_R = (\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)$, $1 \leq j \leq k$, which is the increase in the regression sum of squares due to adding x_j to a model that already contains $\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k$. Some find it helpful to think of this as measuring the **contribution of x_j as if it were the last variable added to the model**.

Remark:

- The extra-sum-of-squares method can be used to test hypotheses about any subset of regressor variables that seems reasonable for the particular problem under analysis. Sometimes we find that there is a natural hierarchy or ordering in the regressors, and this forms the basis of a test.
- When we think of adding regressors one at a time to a model and examining the contribution of the regressor added at each step given all regressors added previously, we can partition the regression sum of squares into marginal single-degree-of-freedom components.

Example: the delivery time data

```
> t(beta)%*%t(X)%*%y-(sum(y))^2/n          ## SS_R(\beta_1,\beta_2|\beta_0)
          [,1]
[1,] 5550.811
> S_xy<-sum(y*(x1-mean(x1)))               ## S_xy
> S_xy
[1] 2473.344
> S_xx<-sum((x1-mean(x1))^2)               ## S_xx
> beta_1<-S_xy/S_xx                        ## \hat beta_1
> beta_1
[1] 2.176167
> beta_1*S_xy                              ## SS_R(\beta_1|\beta_0)
[1] 5382.409
```

- Suppose that we wish to investigate the contribution of the variable distance x_2 to the model.
- Hypotheses: $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$
- Since $SS_R(\beta_1, \beta_2|\beta_0) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 5550.811$ and $SS_R(\beta_1|\beta_0) = \hat{\beta}_1 S_{xy} = (2.1762)(2473.344) = 5382.409$, then
$$SS_R(\beta_2|\beta_1, \beta_0) = 5550.811 - 5382.409 = 168.402$$
- $F_0 = \frac{SS_R(\beta_2|\beta_1,\beta_0)/1}{MS_{Res}} = \frac{168.402/1}{10.62} = 15.85$
- Since $F_{0.05,1,22} = 4.30$, we reject H_0 and conclude that distance x_2 contributes significantly to the model.

Remark:Since this partial F test involves a single variable, it is equivalent to the t test.

Special cases of orthogonal columns in X

- Consider the model
$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}\end{aligned}$$
- The extra-sum-of-squares method allows us to measure the effect of the regressors in $\boldsymbol{X}_2$ conditional on those in $\boldsymbol{X}_1$ by computing $SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$. However, if the columns in $\boldsymbol{X}_1$ are **orthogonal** to the columns in $\boldsymbol{X}_2$, we can determine a sum of squares due to $\boldsymbol{\beta}_2$ that is free of any dependence on the regressors in $\boldsymbol{X}_1$.
- The normal equations are
$$\begin{bmatrix}\boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1'\mathbf{y} \\ \boldsymbol{X}_2'\mathbf{y}\end{bmatrix}$$
- If the columns of $\boldsymbol{X}_1$ are orthogonal to the columns in $\boldsymbol{X}_2$, i.e. $\boldsymbol{X}_1'\boldsymbol{X}_2 = \mathbf{0}$ and $\boldsymbol{X}_2'\boldsymbol{X}_1 = \mathbf{0}$. Then the normal equations become
$$\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\mathbf{y}, \quad \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\mathbf{y}$$
with solution
$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\mathbf{y}, \quad \hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\mathbf{y}$$

Remark: The least-squares estimator of $\boldsymbol{\beta}_1$ is $\hat{\boldsymbol{\beta}}_1$ regardless of whether or not $\boldsymbol{X}_2$ is in the model, and the least-squares estimator of $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_2$ regardless of whether or not $\boldsymbol{X}_1$ is in the model.

- The regression sum of squares for the full model is

$$\begin{aligned} SS_R(\boldsymbol{\beta}) &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= [\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2'] \begin{bmatrix} \boldsymbol{X}_1'\mathbf{y} \\ \boldsymbol{X}_2'\mathbf{y} \end{bmatrix} \\ &= \hat{\boldsymbol{\beta}}_1'\boldsymbol{X}_1'\mathbf{y} + \hat{\boldsymbol{\beta}}_2'\boldsymbol{X}_2'\mathbf{y} \\ &= \mathbf{y}'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\mathbf{y} + \mathbf{y}'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\mathbf{y} \\ &\equiv SS_R(\boldsymbol{\beta}_1) + SS_R(\boldsymbol{\beta}_2) \end{aligned}$$

- Therefore,

$$\begin{aligned} SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) &= SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_2) \equiv SS_R(\boldsymbol{\beta}_1) \\ SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) &= SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1) \equiv SS_R(\boldsymbol{\beta}_2) \end{aligned}$$

- Consequently, $SS_R(\boldsymbol{\beta}_1)$ measures the contribution of the regressors in $\boldsymbol{X}_1$ to the model unconditionally, and $SS_R(\boldsymbol{\beta}_2)$ measures the contribution of the regressors in $\boldsymbol{X}_2$ to the model unconditionally. Because we can unambiguously determine the effect of each regressor when the regressors are orthogonal, data collection experiments are often designed to have orthogonal variables.

- As an example of a regression model with orthogonal regressors, consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where the $\boldsymbol{X}$ matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

The levels of the regressors correspond to the 2^3 factorial design. It is easy to see that the columns of $\boldsymbol{X}$ are orthogonal. Thus, $SS_R(\beta_j), j = 1, 2, 3$, measures the contribution of the regressor x_j to the model regardless of whether any of the other regressors are included in the fit.

Testing the general linear hypothesis

- Hypotheses: $H_0 : \boldsymbol{T}\boldsymbol{\beta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{T}\boldsymbol{\beta} \neq \mathbf{0}$
  where $\boldsymbol{T}$ is an $m \times p$ matrix of constants, such that only r of the m equations in $\boldsymbol{T}\boldsymbol{\beta} = 0$ are independent.
- The full model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, with $\begin{cases} \text{LSE } \hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ SS_{Res}(FM) = \boldsymbol{y}'\boldsymbol{y} - \hat{\beta}'\boldsymbol{X}'\boldsymbol{y}, \ (df = n - p) \end{cases}$
- The reduced model: $\boldsymbol{y} = \boldsymbol{Z}\gamma + \varepsilon$, where $\boldsymbol{Z}$ is an $n \times (p - r)$ matrix and γ is a $(p - r) \times 1$ vector of unknown regression coefficients with
  $\begin{cases} \text{LSE } \hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ SS_{Res}(RM) = \boldsymbol{y}'\boldsymbol{y} - \hat{\gamma}'\boldsymbol{Z}'\boldsymbol{y}, \ (df = n - p + r) \end{cases}$

  Remark: $SS_{Res}(RM) \geq SS_{Res}(FM)$.
- The sum of squares due to the hypothesis $H_0 : \boldsymbol{T}\boldsymbol{\beta} = \mathbf{0}$:
  $SS_H = SS_{Res}(RM) - SS_{Res}(FM)$, with $df = n - p + r - (n - p) = r$.
  The test statistic is
  $$F_0 = \frac{SS_H / r}{SS_{Res}(FM)/(n - p)}$$
- Reject H_0 if $F_0 > F_{\alpha, r, n-p}$.

Examples

- Example I: Consider the model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
  - ⋆ We wish to test $H_0 : \beta_1 = \beta_3$. This hypothesis may be stated as $H_0 : \boldsymbol{T}\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{T} = [0, 1, 0, -1]$.
  - ⋆ There is only one equation in $\boldsymbol{T}\boldsymbol{\beta} = 0$, namely, $\beta_1 - \beta_3 = 0$
  - ⋆ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
    $= \beta_0 + \beta_1 (x_1 + x_3) + \beta_2 x_2 + \varepsilon$
    $= \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \varepsilon$
    where $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1 (= \beta_3)$, $z_1 = x_1 + x_3$, $\gamma_2 = \beta_2$, and $z_2 = x_2$.
  - ⋆ $SS_{Res}(RM)$ has $n - 4 + 1 = n - 3$ degrees of freedom by fitting the reduced model. The sum of squares due to hypothesis $SS_H = SS_{Res}(RM) - SS_{Res}(FM)$ has $n - 3 - (n - 4) = 1$ degree of freedom.
  - ⋆ $F_0 = \frac{SS_H/1}{SS_{Res}(FM)/(n-4)}$ or $t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_3}{\text{se}(\hat{\beta}_1 - \hat{\beta}_3)} = \frac{\hat{\beta}_1 - \hat{\beta}_3}{\sqrt{\hat{\sigma}^2 (C_{11} + C_{33} - 2C_{13})}}$

- Example II: Consider the model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
  - ⋆ We wish to test $H_0 : \beta_1 = \beta_3,\ \beta_2 = 0$. This hypothesis may be stated as $H_0 : \boldsymbol{T}\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{T} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$.
  - ⋆ There are two equations in $\boldsymbol{T}\boldsymbol{\beta} = 0$, namely, $\beta_1 - \beta_3 = 0$ and $\beta_2 = 0$
  - ⋆ $y = \beta_0 + \beta_1 x_1 + \beta_1 x_3 + \varepsilon$
    $= \beta_0 + \beta_1 (x_1 + x_3) + \varepsilon$
    $= \gamma_0 + \gamma_1 z_1 \varepsilon$
    where $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1 (= \beta_3)$, $z_1 = x_1 + x_3$.
  - ⋆ $SS_{Res}(RM)$ has $n-2$ degrees of freedom, so $SS_H = SS_{Res}(RM) - SS_{Res}(FM)$ has $n - 2 - (n-4) = 2$ degrees of freedom.
  - ⋆ $F_0 = \frac{SS_H/2}{SS_{Res}(FM)/(n-4)}$.

Remark:

- The test statistic for the general linear hypothesis has another form:

$$F_0 = \frac{\hat{\boldsymbol{\beta}}' \boldsymbol{T}' [\boldsymbol{T}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{T}']^{-1}\boldsymbol{T}\hat{\boldsymbol{\beta}}/r}{SS_{Res}(FM)/(n-p)}$$

- There is a slight extension of the general linear hypothesis that is occasionally useful.
  - ⋆ Hypotheses: $H_0 : \boldsymbol{T}\boldsymbol{\beta} = \boldsymbol{c}$ vs. $H_1 : \boldsymbol{T}\boldsymbol{\beta} \neq \boldsymbol{c}$
  - ⋆ The test statistic:

$$F_0 = \frac{(\boldsymbol{T}\hat{\boldsymbol{\beta}} - \boldsymbol{c})' [\boldsymbol{T}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{T}']^{-1}(\boldsymbol{T}\hat{\boldsymbol{\beta}} - \boldsymbol{c})/r}{SS_{Res}(FM)/(n-p)}$$

  - ⋆ Reject H_0 if $F_0 > F_{\alpha,r,n-p}$
- Finally, if the hypothesis $H_0 : \boldsymbol{T}\boldsymbol{\beta} = 0$ (or $H_0 : \boldsymbol{T}\boldsymbol{\beta} = \boldsymbol{c}$) cannot be rejected, then it may be reasonable to estimate $\boldsymbol{\beta}$ subject to the constraint imposed by the null hypothesis. It is unlikely that the usual least-squares estimator will automatically satisfy the constraint. In such cases a **constrained least-squares estimator** may be useful.

Confidence interval on the regression coefficients

- **MLR model**: Sample MLR Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$, $i = 1, \ldots, n$.
- Assumption: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.
- $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}$, $j = 0, 1, \ldots, k$. C_{jj} is the jth diagonal element of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix.
- **A $100(1-\alpha)$ percent confidence interval (CI) for the regression coefficient** β_j, $j = 0, 1, \ldots, k$ is given by

$$\hat{\beta}_j - t_{\alpha/2, n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

where $\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$.

Confidence interval on the regression coefficient for the delivery time data

- Method I

```
> res<-lm(y~x1+x2)
> summary(res)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
x1          1.615907   0.170735   9.464 3.25e-09 ***
x2          0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

```
> ## ANOVA
> A<-cbind(x1,x2)
> res<-lm(y~A)
> anova(res)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          2 5550.8 2775.41  261.24 4.687e-16 ***
Residuals 22  233.7   10.62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> solve(t(X)%*%X)
                  I            x1            x2
I   1.132152e-01 -4.448593e-03 -8.367257e-05
x1 -4.448593e-03  2.743783e-03 -4.785709e-05
x2 -8.367257e-05 -4.785709e-05  1.228745e-06
> qt(0.975, df=22)
[1] 2.073873
```

A 95% confidence interval for β_1 is

$$\hat{\beta}_1 - t_{0.025,22}\sqrt{\hat{\sigma}^2 C_{11}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,22}\sqrt{\hat{\sigma}^2 C_{11}}$$

$$1.6159 - (2.0739)\sqrt{(10.62)(0.0027438)} \leq \beta \leq 1.6159 + (2.0739)\sqrt{(10.62)(0.0027438)}$$

- Method II

```
> confint(res,level=0.95)
                2.5 %     97.5 %
(Intercept) 0.066751987 4.61571030
x1          1.261824662 1.96998976
x2          0.006891745 0.02187791
```

A 95% confidence interval for β_1 is

$$1.2618 \leq \beta_1 \leq 1.9700$$

Confidence interval on the mean response

- Define $\boldsymbol{x}_0 = [1, x_{01}, x_{02}, \ldots, x_{0k}]'$
  The fitted value at this point is $\hat{y}_0 = \boldsymbol{x}_0 \hat{\boldsymbol{\beta}}$
- $E(\hat{y}_0) = \boldsymbol{x}_0' \boldsymbol{\beta} = E(y|\boldsymbol{x}_0)$, $Var(\hat{y}_0) = \hat{\sigma}^2 \boldsymbol{x}_0' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_0$
- A $100(1-\alpha)$ percent confidence interval (CI) on the mean response at the point $x = x_0$ is given by

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_0} \leq E(y|\boldsymbol{x}_0) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_0}$$

Remark:

- The interval width is a minimum for $x_0 = \overline{x}$ and widens as $|x_0 - \overline{x}|$ increases;
- We would expect our best estimates of y to be made at x values the center of the data and the precision of estimation to deteriorate as we move to the boundary of the x space.

Confidence interval on $E(y|\boldsymbol{x}_0)$ for the delivery time data

A 95% confidence interval on the mean delivery time for an outlet requiring $x_1 = 8$ cases and where the distance $x_2 = 275$ feet. Therefore, $\boldsymbol{x}_0 = [1, 8, 275]'$

- Method I

```
> ## ANOVA
> A<-cbind(x1,x2)
> res<-lm(y~A)
> anova(res)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          2 5550.8 2775.41  261.24 4.687e-16 ***
Residuals 22  233.7   10.62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> x0<-c(1,8,275)
> y0=x0%*%beta                          ## \hat y_0
> t(x0)%*%(solve(t(X)%*%X))%*%x0        ## x0'(X'X)^{-1}x0
            [,1]
[1,] 0.05397255
> x0<-c(1,8,275)
> y0=x0%*%beta                          ## \hat y_0
> y0
         [,1]
[1,] 19.22432
> t(x0)%*%(solve(t(X)%*%X))%*%x0        ## x0'(X'X)^{-1}x0
            [,1]
[1,] 0.05397255
> qt(0.975, df=22)
[1] 2.073873
```

A 95% confidence interval on the mean delivery time at this point is

$$\hat{\mu}_{y|x_0} - t_{0.025,22}\sqrt{\hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0} \leq E(y|\boldsymbol{x}_0) \leq \hat{\mu}_{y|x_0} + t_{0.025,22}\sqrt{\hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

$$19.2242 - 2.0739\sqrt{(10.62)(0.0540)} \leq E(y|\boldsymbol{x}_0) \leq 19.2242 + 2.0739\sqrt{(10.62)(0.0540)}$$

- Method II

```
> ## 95% CI for the mean value by x1=8, x2=275
> predict(res,data.frame(x1=8,x2=275),interval='confidence',level=0.95)
       fit     lwr      upr
1 19.22432 17.6539 20.79474
```

A 95% confidence interval on the mean delivery time at this point is

$$17.6539 \leq E(y|\boldsymbol{x}_0) \leq 20.7947$$

Remark:

- The length of the CI on the mean response is a useful measure of the quality of the regression model.

```
> res1<-lm(y~x1)
> predict(res1,data.frame(x1=8),interval='confidence',level=0.95)
       fit      lwr      upr
1 20.73011 18.98918 22.47104
```

- If we consider the simple linear regression model with x_1 cases as the only regressor, the 95% CI on the mean delivery time with $x_1 = 8$ cases is $(18.9892, 22.4710)$. The length of this interval is $22.4704 - 18.9892 = 3.4812$ minutes. Clearly, adding cases to the model has improved the precision of estimation.

```
> predict(res,data.frame(x1=16,x2=688),interval='confidence',level=0.95)
       fit      lwr      upr
1 38.09251 36.10864 40.07638
> predict(res1,data.frame(x1=16),interval='confidence',level=0.95)
       fit      lwr      upr
1 38.13945 35.60104 40.67785
```

- The change in the length of the interval depends on the location of the point in the x space.Consider the point $x_1 = 16$ cases and $x_2 = 688$ feet. The 95% CI for the multiple regression model is $(36.1086, 40.0764)$ with length 3.9678 minutes, and for the simple linear regression model the 95% CI at $x_1 = 16$ cases is $(35.6010, 40.6779)$ with length 5.0769 minutes. The improvement from the multiple regression model is even better at this point.
- Generally, the further the point is from the centroid of the x space, the greater the difference will be in the lengths of the two CIs.

Simultaneous confidence intervals on regression coefficients

A set of confidence or prediction intervals that are all true simultaneously with probability $1-\alpha$ are called **simultaneous** or **joint confidence** or **joint prediction intervals**.

- A joint confidence region for the multiple regression model parameter β.
  - $\star$ $\frac{(\hat{\beta}-\beta)'\boldsymbol{X}'\boldsymbol{X}(\hat{\beta}-\beta)}{pMS_{Res}} \sim F_{p,n-p}$
  - $\star$ This implies that $P\left\{\frac{(\hat{\beta}-\beta)'\boldsymbol{X}'\boldsymbol{X}(\hat{\beta}-\beta)}{pMS_{Res}} \leq F_{\alpha,p,n-p}\right\} = 1-\alpha$
  - $\star$ A $100(1-\alpha)$ percent **joint confidence region** for all of the parameters in β is $\frac{(\hat{\beta}-\beta)'\boldsymbol{X}'\boldsymbol{X}(\hat{\beta}-\beta)}{pMS_{Res}} \leq F_{\alpha,p,n-p}$

  Remark: This inequality describes an elliptically shaped region. Construction of this joint confidence region is relatively straightforward for simple linear regression $(p=2)$. It is more difficult for $p=3$ and would require special three-dimensional graphics software.

Simultaneous confidence intervals on regression coefficients for the rocket propellant data

- A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal housing.
- y: the shear strength of the bond;
  x: the age of propellant.

Observation, i	Shear Strength, y_i (psi)	Age of Propellant, x_i (weeks)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

- For the case of simple linear regression, we can show that

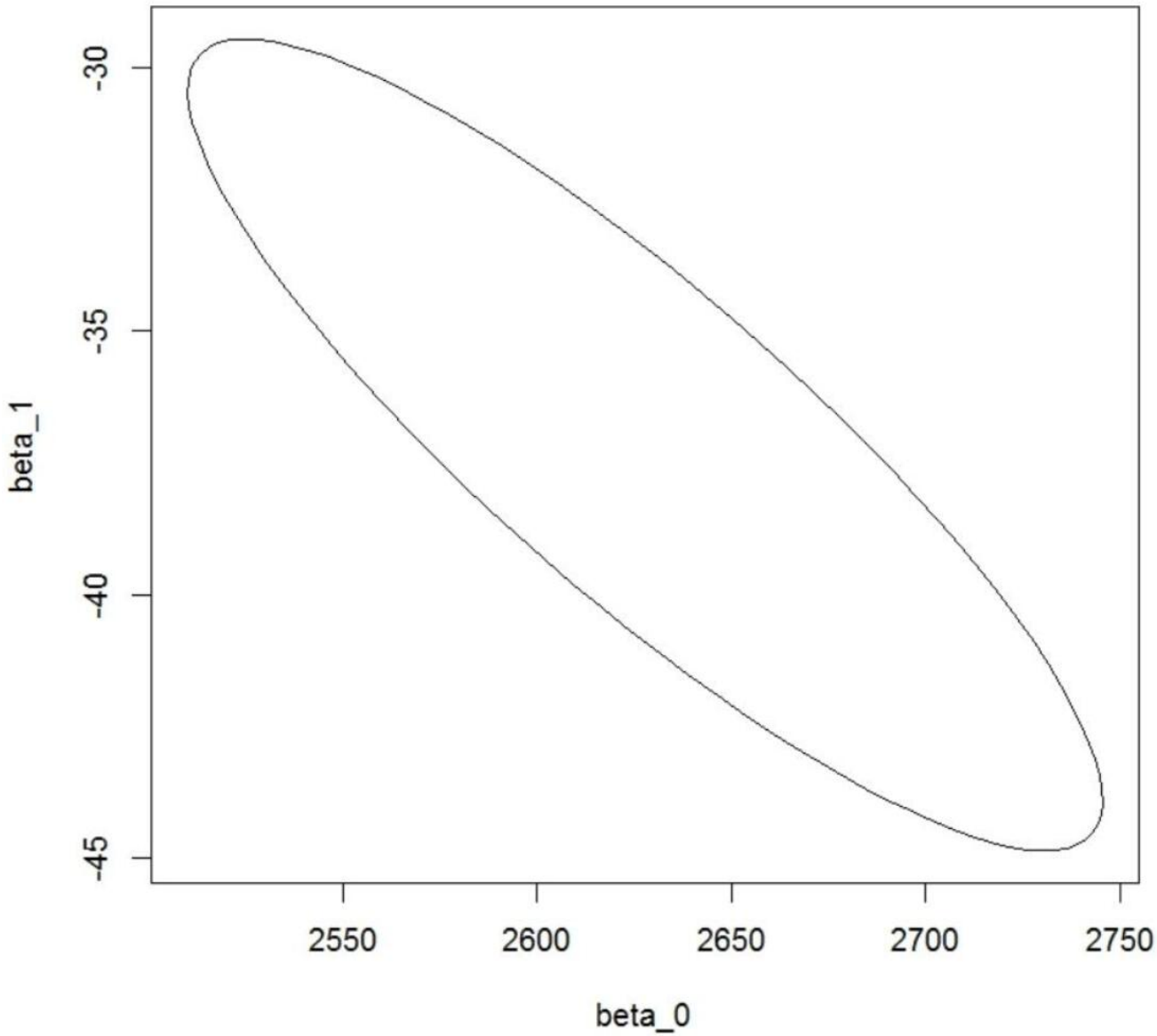
$$\frac{n(\hat{\beta}_0-\beta_0)^2+2\sum_{i=1}^{n}x_i(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)+\sum_{i=1}^{n}x_i^2(\hat{\beta}_1-\beta_1)^2}{2MS_{Res}} \leq F_{\alpha,2,n-2}$$

```
> setwd("C:/Users/YANG Jin/Desktop/AMA514_2016_new/Data")
> E3.10<-read.table("data_example_2.1.txt", header=T)
> ## Simplify notation
> y<-E3.10$y; x<-E3.10$x; n<-20
> res<-lm(y~x)
> summary(res)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

```
> sum(x)
[1] 267.25
> sum(x^2)
[1] 4677.688
> anova(res)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1527483 1527483  165.38 1.643e-10 ***
Residuals 18  166255    9236
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> qf(0.95,df1=2,df2=18)
[1] 3.554557
```

- $[20(2627.822-\beta_0)^2+2(267.25)(2627.822-\beta_0)(-37.154-\beta_1)+(4677.688)(-37.154-\beta_1)^2]/[2(9236)]=3.5546$

```
> library(ellipse)
> B<-ellipse(res, which = c(1, 2), level = 0.95)
> plot(B, type="l", xlab="beta_0", ylab="beta_1")
```



Remark:

- The tilt of the ellipse is a function of the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$, which is $-\overline{x}\sigma^2/S_{xx}$. A positive covariance implies that errors in the point estimates of β_0 and β_1 are likely to be in the same direction, while a negative covariance indicates that these errors are likely to be in opposite directions;
- In our example, $\overline{x}$ is positive, so $Cov(\hat{\beta}_0, \hat{\beta}_1)$ is negative. Thus, if the estimate of the slope is too steep (β_1 is overestimated), the estimate of the intercept is likely to be too small (β_0 is underestimated).
- The elongation of the region depends on the relative sizes of the variances of β_0 and β_1. Generally, if the ellipse is elongated in the β_0 direction (for example), this implies that β_0 is not estimated as precisely as β_1.

Another general approach for obtaining simultaneous interval estimates of the parameters

These CIs may be constructed by using

$$\beta_j \pm \Delta se(\hat{\beta}_j)$$

Several methods may be used to choose Δ.

- **Bonferroni method**
  - $\Delta = t_{\alpha/2,n-p} \Rightarrow \beta_j \pm t_{\alpha/2,n-p} se(\hat{\beta}_j)$
    The probability is at least $1-\alpha$ that all intervals are correct.
  - Notice that the **Bonferroni confidence intervals** look somewhat like the ordinary one-at-a-time CIs based on the t distribution, except that each Bonferroni interval has a confidence coefficient $1-\alpha/p$ instead of $1-\alpha$.

```
> setwd("C:/Users/YANG Jin/Desktop/AMA514_2016_new/Data")
> E3.10<-read.table("data_example_2.1.txt", header=T)
> ## Simplify notation
> y<-E3.10$y; x<-E3.10$x; n<-20
> res<-lm(y~x)
> summary(res)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
> qt(0.9875,df=18)
[1] 2.445006
```

$$\hat{\beta}_0 - t_{0.0125,18}\, se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{0.0125,18}\, se(\hat{\beta}_0)$$

$$2627.822 - (2.4450)(44.184) \leq \beta_0 \leq 2627.822 + (2.4450)(44.184)$$

$$2519.792 \leq \beta_0 \leq 2735.852$$

$$\hat{\beta}_1 - t_{0.0125,18}\, se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{0.0125,18}\, se(\hat{\beta}_1)$$

$$-37.154 - (2.4450)(2.889) \leq \beta_1 \leq -37.154 + (2.4450)(2.889)$$

$$-44.218 \leq \beta_1 \leq -30.090$$

We conclude with 90% confidence that this procedure leads to correct interval estimates for both parameters. Remark: The confidence ellipse is always a more efficient procedure than the Bonferroni method because the volume of the ellipse is always less than the volume of the space covered by the Bonferroni intervals. However, the Bonferroni intervals are easier to construct.

- **Scheffé's-method**: $\Delta = 2(F_{\alpha,p,n-p})^{1/2}$
- **Maximum modulus t procedure**: $Delta = u_{\alpha,p-n-p}$, where $u_{\alpha,p-n-p}$ is the upper α-tail point of the distnbution of the maximum absolute value of two independent student t random variables each based on $n-2$ degrees of freedom.

Remark: An obvious way to compare these three techniques is in terms of the lengths of the CIs they generate. Generally the Bonferroni intervals are shorter than the Scheffé intervals and the maximum modulus t intervals are shorter than the Bonferroni intervals.

Prediction of new observation

- Define $\boldsymbol{x}_0 = [1, x_{01}, x_{02}, \ldots, x_{0k}]'$
  The fitted value at this point is $\hat{y}_0 = \boldsymbol{x}_0 \hat{\boldsymbol{\beta}}$
- A $100(1-\alpha)$ percent prediction interval for this future observation is

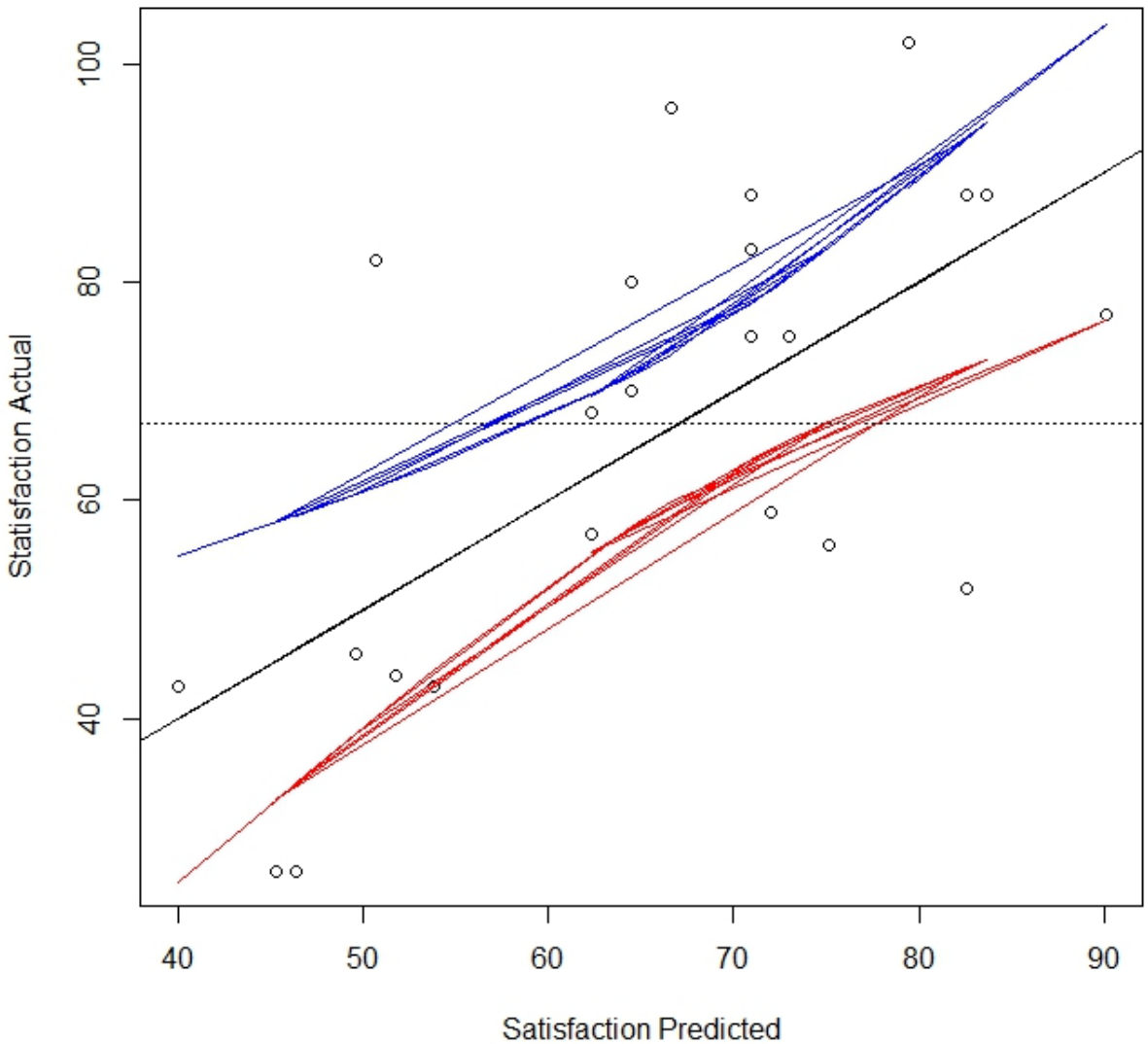
$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\ \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\ \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)}$$

- Example: the delivery time data
  A 95% ponfidence interval on the mean delivery time at the point $\boldsymbol{x}_0' = [1, 8, 275]$ is

$$19.2242 - 2.0739\sqrt{(10.62)(1+0.0540)} \leq y_0 \leq 19.2242 + 2.0739\sqrt{(10.62)(1+0.0540)}$$

$$12.28 \leq y_0 \leq 26.16$$

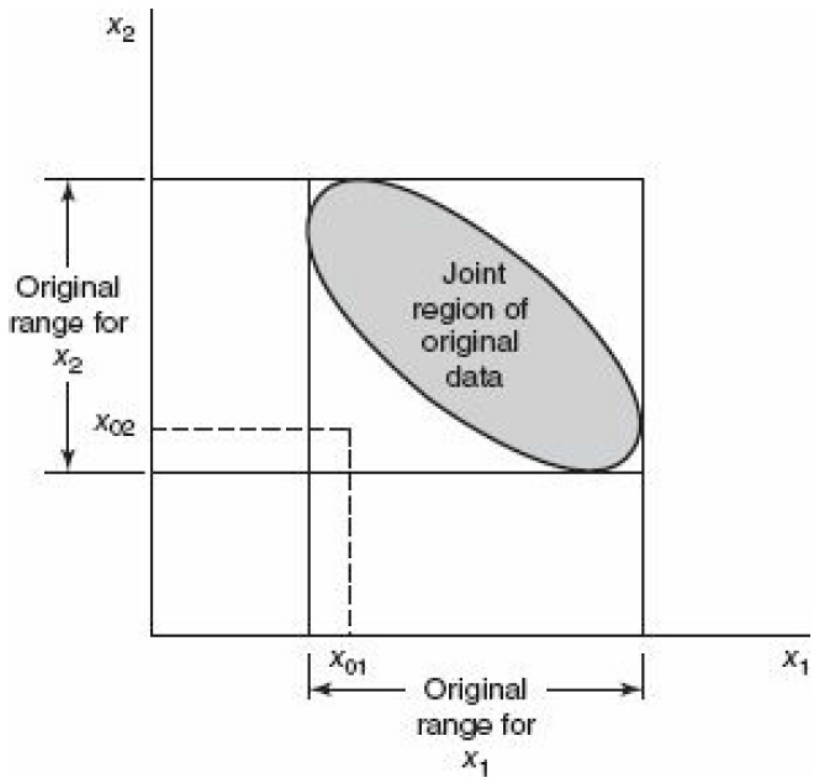
```
> ### ANOVA
> y<-Q1$y; x1<-Q1$x1; x2<-Q1$x2
> n<-25
> I<-rep(1, times=n)
> X<-cbind(I,x1,x2)
> res1<-lm(y~X)
> anova(res1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          2 8768.8  4384.4  46.772 1.193e-08 ***
Residuals 22 2062.3    93.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> ## Actual by predicted plot
> plot(fitted(result), Sec2.7$y, xlab="Satisfaction Predicted", ylab="Statisfaction Actual")
> res<-lm(Sec2.7$y~fitted(result))
> abline(res)
> abline(h=67, lty=3, lwd=1)
> lines(
+     x   = fitted(result),
+     y   = predict(res, Sec2.7, interval = "confidence", level=0.95)[ , "lwr" ],
+     col = "red")
> lines(
+     x   = fitted(result),
+     y   = predict(res, Sec2.7, interval = "confidence", level=0.95)[ , "upr" ],
+     col = "blue")
> lines(
+     x   = fitted(result),
+     y   = predict(res, Sec2.7, interval = "confidence", level=0.95)[ , "fit" ],
+     col = "black")
```


Statisfaction Actual
100
80
60
40
40
50
60
70
80
90
Satisfaction Predicted

- In the multiple linear regression model we notice that the plot of actual versus predicted response is much improved when compared to the plot for the simple linear regression model.
- Furthermore, the model is significant and both variables, age and severity, contribute significantly to the model. The R^2 has increased from 0.4266 to 0.8096. The mean square error in the multiple linear regression model is 93.7, considerably smaller than the mean square error in the simple linear regression model, which was 270.0. The large reduction in mean square error indicates that the two-variable model is much more effective in explaining the variability in the data than the original simple linear regression model.
- This reduction in the mean square error is a quantitative measure of the improvement we qualitatively observed in the plot of actual response versus the predicted response when the predictor age was added to the model.
- Finally, the response is predicted with better precision in the multiple linear model. Adding an important predictor to a regression model can often result in a much better fitting model with a smaller standard error and as a consequence narrow confidence intervals on the mean response and narrower prediction intervals.

Hidden extrapolation in multiple regression

- In predicting new responses and in estimating the mean response at a given point $x_{01}, x_{02}, \ldots, x_{0k}$ one must be careful about **extrapolating** beyond the region containing the original observations.
- It is very possible that a model that fits well in the region of the original data will perform poorly outside that region. In multiple regression it is easy to inadvertently extrapolate, since the levels of the regressors $(x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, 2, \ldots, n$, jointly define the region containing the data.

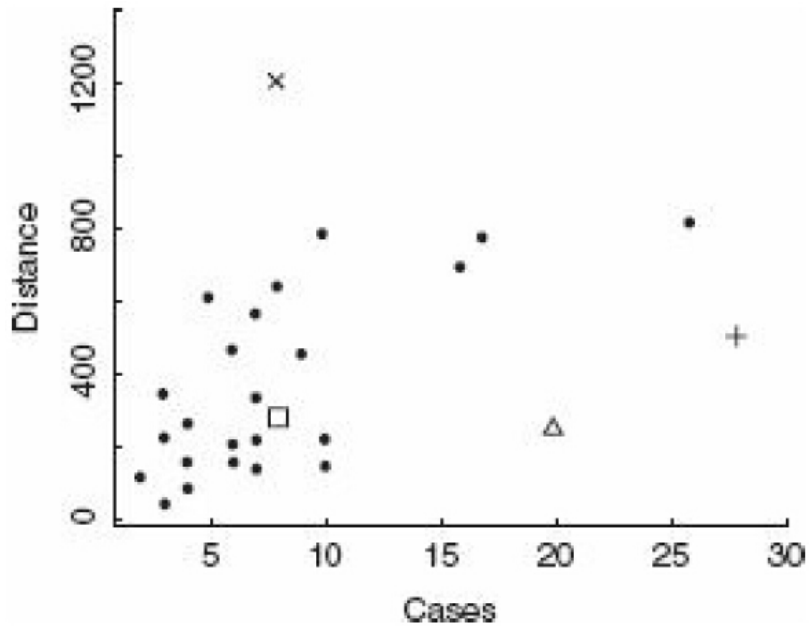


- Since simply comparing the levels of the x's for a new data point with the ranges of the original x's will not always detect a hidden extrapolation, it would be helpful to have a formal procedure to do so.
- Define the smallest convex set containing all of the original n data points $(x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, 2, \ldots, n$, as the **regressor variable hull** (RVH).
- If a point $x_{01}, x_{02}, \ldots, x_{0k}$ lies inside or on the boundary of the RVH, then prediction or estimation involves interpolation, while if this point lies outside the RVH, extrapolation is required.
- The diagonal elements h_{ii} of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ are useful in detecting **hidden extrapolation**.The values of h_{ii} depend both on the Euclidean distance of the point xi from the centroid and on the density of the points in the RVH.
- In general, the point that has the largest value of h_{ii}, say $h_{\max}$.
  - $h_{\max}$ will lie on the boundary of the RVH in a region of the x space where the density of the observations is relatively low.
  - The set of points $\boldsymbol{x}$ (not necessarily data points used to fit the model) that satisfy $\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} \leq h_{\max}$ is an ellipsoid enclosing all points inside the RVH.
  - Thus, if we are interested in prediction or estimation at the point $\boldsymbol{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, the location of that point relative to the RVH is reflected by $h_{00} = \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0'$
    - If $h_{00} > h_{\max}$, then the points are outside the ellipsoid enclosing the RVH and are erxtrapolation points;
    - if $h_{00} < h_{\max}$, then the point is inside the ellipsoid and possibly inside the RVH and would be considered an interpolation point because it is close to the cloud of points used to fit the model.

Hidden extrapolation-the delivery time data

Observation No., i	No. of Cases, x_1	Distance, x_2	h_{ii}
1	7	560	0.10180
2	3	220	0.07070
3	3	340	0.09874
4	4	80	0.08538
5	6	150	0.07501
6	7	330	0.04287
7	2	110	0.08180
8	7	210	0.06373
9	30	1460	0.49829
10	5	605	0.19630
11	16	688	0.08613
12	10	215	0.11366
13	4	255	0.06113
14	6	462	0.07824
15	9	448	0.04111
16	10	776	0.16594
17	6	200	0.05943
18	7	132	0.09626
19	3	36	0.09645
20	17	770	0.10169
21	10	140	0.16528
22	26	810	0.39158
23	9	450	0.04126
24	8	635	0.12061
25	4	150	0.03334

Point	Symbols in Figure	x_{10}	x_{20}	h_{00}
a	$\square$	8	275	0.05346
b	$\triangle$	20	250	0.58917
c	$+$	28	500	0.89874
d	$\times$	8	1200	0.86736



Remark: In Figure, point a, for which $h_{00} = 0.05346$, is an interpolation point since $h_{00} = 0.05346 < h_{\max} = 0.49829$. The remaining points b, c, and d are all extrapolation points, since their values of h_{00} exceed $h_{\max}$.

Standard regression coefficients

- Generally the units of the regression coefficient β_j are units of y/units of x_j. For this reason, it is sometimes helpful to work with scaled regressor and response variables that produce **dimensionless regression coefficients**. These dimensionless coefficients are usually called **standardized regression coefficients**.
- Two popular scaling techniques: unit normal scaling and unit length scaling.
- Unit normal scaling
  - $$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k$$

    where $s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1}$ is the sample variance of regressor x_j.
  - $$y_i^* = \frac{y_i - \overline{y}}{s_y}, \quad i = 1, 2, \ldots, n$$

    where $s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$ is the sample variance of the response. All of the scaled regressors and the scaled responses have sample mean equal to zero and sample variance equal to 1.
  - Use the new variables, the regression model becomes

    $$y_i^* = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

  - Centering the regressor and response variables by subtracting $\overline{x}_j$ and $\overline{y}$ removes the intercept from the model. The least-squares estimator of $\boldsymbol{b}$ is $\hat{\boldsymbol{b}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}^*$

- Unit length scaling
  - $$w_{ij} = \frac{x_{ij} - \overline{x}_j}{s_{jj}^{1/2}}, \ i = 1, 2, \ldots, n, \ j = 1, 2, \ldots, k$$

    where $s_{jj} = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$ is the corrected sum of squares for regressor x_j. In this scaling, each new regressor w_j has mean $\overline{w}_j = 0$ and length $\sqrt{\sum_{i=1}^{n} (w_{ij} - \overline{w}_j)^2} = 1$
  - $$y_i^0 = \frac{y_i - \overline{y}}{SS_T^{1/2}}, \ i = 1, 2, \ldots, n$$
  - In terms of these variables, the regression model is

    $$y_i^0 = b_1 w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \varepsilon_i, \ i = 1, 2, \ldots, n$$
  - The vector of least-squares regression coefficients is

    $$\hat{\boldsymbol{b}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{y}^0$$

    where $\boldsymbol{W}'\boldsymbol{W}$ is in the form of a **correlation matrix**, i.e.

    $$\boldsymbol{W}'\boldsymbol{W} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & \cdots & r_{2k} \\ r_{13} & r_{23} & 1 & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{bmatrix}$$

    where $r_{ij} = \frac{\sum_{u=1}^{n} (x_{ui} - \overline{x}_i)(x_{uj} - \overline{x}_j)}{(SS_{ii} SS_{jj})^{1/2}} = \frac{S_{ij}}{(SS_{ii} SS_{jj})^{1/2}}$ is the simple correlation between regressors x_i and x_j.

- Unit length scaling (Cont'd)
  - ⋆ Similarly,

$$\boldsymbol{W}'\boldsymbol{y}^0 = [r_{1y}, r_{2y}, r_{3y}, \cdots r_{ky}]'$$

where $r_{jy} = \frac{\sum_{u=1}^{n}(x_{uj}-\overline{x}_j)(y_u-\overline{y})}{(SS_{jj}SS_T)^{1/2}} = \frac{S_{jy}}{(SS_{jj}SS_T)^{1/2}}$ is the simple correlation between regressors x_j and the response y.

  - ⋆ If unit normal scaling is used, the $\boldsymbol{Z}'\boldsymbol{Z}$ matrix is closely related to $\boldsymbol{W}'\boldsymbol{W}$. In fact,

$$\boldsymbol{Z}'\boldsymbol{Z} = (n-1)\boldsymbol{W}'\boldsymbol{W}$$

Remark: Consequently, the estimates of the regression coefficients $\hat{\boldsymbol{b}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}^*$ and $\hat{\boldsymbol{b}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{y}^0$ are identical. That is, it does not matter which scaling we use; they both produce the same set of dimensionless regression coefficients $\hat{\boldsymbol{b}}$.

- The The regression coefficients are usually called **standardized regression coefficients**. The relationship between the original and standardized regression coefficients is

$$\hat{\beta}_j = \hat{b}_j\left(\frac{SS_T}{S_{jj}}\right)^2,\ j = 1, 2, \ldots, k,\ \text{and}\ \hat{\beta}_0 = \overline{y} - \sum_{j=1}^{k}\hat{\beta}_j\overline{x}_j$$

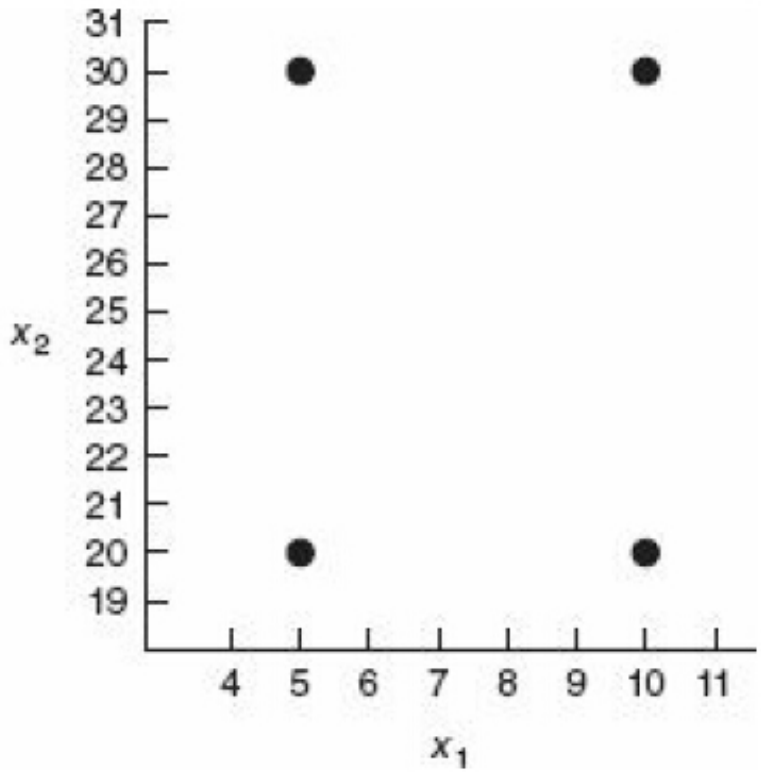
Remark:
  - ⋆ In interpreting standardized regression coefficients, we must remember that they are still partial regression coefficients;
  - ⋆ Furthermore, the b_j are affected by the range of values for the regressor variables.
  - ⋆ Consequently, it may be dangerous to use the magnitude of the as a measure of the relative importance of regressor x_j.

Multicollinearity

Regression models are used for a wide variety of applications. A serious problem that may dramatically impact the usefulness of a regression model is **multicollinearity**, or **near-linear dependence** among the regression variables.

- Multicollinearity implies near-linear dependence among the regressors. The regressors are the columns of the $\boldsymbol{X}$ matrix, so clearly an exact linear dependence would result in a **singular** $\boldsymbol{X}'\boldsymbol{X}$. The presence of near-linear dependencies can dramatically impact the ability to estimate regression coefficients.

x_1	x_2
5	20
10	20
5	30
10	30
5	20
10	20
5	30
10	30



- Suppose we use the unit length scaling fore the data in the aforementioned table, so that the $\boldsymbol{X}'\boldsymbol{X}$ matrix (here is called $\boldsymbol{W}'\boldsymbol{W}$) will be in the form of a correlation matrix
- This results in $\boldsymbol{W}'\boldsymbol{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $(\boldsymbol{W}'\boldsymbol{W})^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

  For the soft drink delivery time data, we showed that

  $\boldsymbol{W}'\boldsymbol{W} = \begin{bmatrix} 1.00000 & 0.8248215 \\ 0.8248215 & 1.00000 \end{bmatrix}$ and $(\boldsymbol{W}'\boldsymbol{W})^{-1} = \begin{bmatrix} 3.11841 & -2.57023 \\ -2.57023 & 3.11841 \end{bmatrix}$
- For the hypothetical data in the aforementioned table,

$$\frac{Var(\hat{b}_1)}{\sigma^2} = \frac{Var(\hat{b}_2)}{\sigma^2} = 1$$

  While for the soft drink delivery time data

$$\frac{Var(\hat{b}_1)}{\sigma^2} = \frac{Var(\hat{b}_2)}{\sigma^2} = 3.11841$$

  - ⋆ In the soft drink delivery time data the variances of the regression coefficients are **inflated** because of the multicollinearity.
  - ⋆ This multicollinearity is evident from the nonzero off-diagonal elements in $\boldsymbol{W}'\boldsymbol{W}$. These off-diagonal elements are usually called simple correlations between the regressors, although the term correlation may not be appropriate unless the x's are random variables
  - ⋆ The off-diagonals do provide a measure of linear dependency between regressors. Thus, multicollinearity can seriously affect the precision with which regression coefficients are estimated.

- The main diagonal elements of the inverse of the $\boldsymbol{X}'\boldsymbol{X}$ matrix in correlation form [$(\boldsymbol{W}'\boldsymbol{W})^{-1}$ above] are often called **variance inflation factors (VIFs)**, and they are an important multicollinearity diagnostic.
- For the soft drink data, $VIF_1 = VIF_2 = 3.11841$, while for the above table, $VIF_1 = VIF_2 = 1$ implying that the two regressors x_1 and x_2 are **orthogonal**.
- $VIF_k = \frac{1}{1-R_j^2}$, where R_j^2 os the coefficient of multiple determination obtained from regressing x_j on the other regressor variables.
- Regression models fit to data by the method of least squares when strong multicollinearity is present are notoriously poor prediction equations, and the values of the regression coefficients are often very sensitive to the data in the particular sample collected.
- The diagnosis and treatment of multicollinearity is an important aspect of regression modeling.

Why do regression coefficients have the wrong sign?

- When using multiple regression, occasionally we find an apparent contradiction of intuition or theory when one or more of the regression coefficients seem to have the **wrong sign**
- Regression coefficients may have the wrong sign for the following reasons:
  - ⋆ The range of some of the regressors is too small.
  - ⋆ Important regressors have not been included in the model.
  - ⋆ Multicollinearity is present.
  - ⋆ Computational errors have been made.

Polynomial Regression Models

The linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is a general model for fitting any relationship that is linear in the unknown parameters β. This includes the important class of **polynomial regression models**.

- The second-order polynomial in one variable:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

- The second-order polynomial in two variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

- In general, the kth-order polynomial model in one variable is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

  If we set $x^j = x_j$, $j = 1, 2, \cdots, k$, then it becomes a multiple linear regression model in the k regressors $x_1, \cdots, x_k$. Thus, a polynomial model of order k may be fitted using the techniques studied previously.

- Polynomials are widely used in situations where the response is curvilinear, as even complex nonlinear relationships can be adequately modeled by polynomials over reasonably small ranges of the x's.