- Write the general equation for a **multiple** linear regression model.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik} + \varepsilon_i \quad i = 1, 2, 3 \cdots n.$$

- Write the least-squares equations for a multiple linear regression in matrix form.

$$\hat{\beta} = (X'X)^{-1} X'Y \qquad X'Y = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix} \qquad X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

- State the test statistic and confidence interval formulas for a linear function of parameters in multiple linear regression.

$$H_0 : a'\beta = (a'\beta)_0$$

$$T = \frac{a'\hat{\beta} - (a'\beta)_0}{S\sqrt{a'(X'X)^{-1}a}} \qquad \text{CI}: \ a'\beta \pm t_{\alpha/2} S \sqrt{a'(X'X)^{-1}a}$$

- Describe the general process of testing the hypothesis that $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

Test for $a'\beta$

① $H_0 : a'\beta = (a'\beta)_0$

Chose 1 test from below

$$H_a \begin{cases} a'\beta > (a'\beta)_0 \\ a'\beta < (a'\beta)_0 \\ a'\beta \neq (a'\beta)_0 \end{cases}$$

② Perform T test

$$T = \frac{a'\hat{\beta} - (a'\beta)_0}{S\sqrt{a'(X'X)^{-1}a}}$$

Reject H_0 if $\begin{cases} t > t_\alpha \\ t < -t_\alpha \\ |t| > t_{\alpha/2} \end{cases}$ $df : (n-(k+1))$

1. Consider the general linear model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$, where $E[\epsilon] = 0$ and $V(\epsilon) = \sigma^2$. Notice that $\hat{\beta}_1 = \mathbf{a}'\hat{\beta}$, where the vector $\mathbf{a}$ is defined by $a_j = 1$ if $j = i$ and $a_j = 0$ if $j \neq i$.

   Use this to verify that $E[\hat{\beta}_1] = \beta_i$ and $V(\hat{\beta}_i) = c_{ii}\sigma^2$, where c_{ii} is the element in row i and column i of $(\mathbf{X}'\mathbf{X})^{-1}$.

$$
\begin{aligned}
E(\hat{\beta}_1) = E(a'\hat{\beta}) &= a' E((X'X)^{-1}X'Y) \\
&= a'[(X'X)^{-1}X' E(Y)] \\
&= a'[(X'X)^{-1} X'X\hat{\beta}] \\
&= a'[I\hat{\beta}] \\
&= a'\hat{\beta} \Rightarrow \text{unbiased Estimator.}
\end{aligned}
$$

$$Var(\hat{\beta}) = (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} = \sigma^2 (X'X)^{-1}$$

$$Var(\hat{\beta}_j) = \sigma^2 [X_j' x_j - x_j' X_{-j}(X_{-j}' X_{-j})^{-1} X_{-j}' x_j]^{-1}$$

$$
\begin{aligned}
A &= X_1' X_1 \\
B &= X_1' X_{-1} \\
C &= X_{-1} X_1 \\
D &= X_{-1}' X_{-1}
\end{aligned}
\qquad
X'X = \begin{bmatrix} A & B \\ C & D \end{bmatrix}
$$

$$(X'X)^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & \cdots \\ \cdots & \cdots \end{bmatrix}$$

$$Var(\hat{\beta}) = C_{ii}\sigma^2$$

$\square$

2. A real estate agent's computer data listed the selling price Y (in thousands of dollars), the living area x_1 (in hundreds of square feet), the number of floors x_2, number of bedrooms x_3, and number of bathrooms x_4 for newly listed condominiums. The multiple regression model $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ was fit to the data obtained by randomly selecting 40 condos currently on the market.

   (a) If $R^2 = 0.802$, is there sufficient evidence to conclude that at least one of the independent variables contributes significant information for the prediction of selling price?

   (b) If $S_{yy} = 15530.6$, what is SSE?

   (c) The realtor theorizes that square footage, x_1, is the most important predictor variable, and that the other variables can be left out without losing much prediction information. A simple linear regression of selling price vs. square footage was fit using the same 40 condos, and its SSE was 1553. Can the other independent variables, $x_2, x_3, and x_4$ be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ $\quad H_A: \beta_1 | \beta_2 | \beta_3 | \beta_4 \neq 0.$

a) $F = \frac{n-(k+1)}{k}\left(\frac{R^2}{1-R^2}\right)$ where n = # of Data Points, k = # of indep. Var

$= \frac{40-(4+1)}{4}\left(\frac{0.802}{1-0.802}\right)$

$= 35.442.$

When $\alpha = 0.05$

$F_{0.05}\ v_1 = 4,\ v_2 = 35 = 2.641$

Since $35.442 > 2.641$

$\Rightarrow$ Reject null hypothesis $\Rightarrow$ At least one indep Var contribute Sigificant Contribution.

b) $SSE =$ $\qquad R^2 = \frac{S_{yy} - SSE}{S_{yy}}$

$S_{yy} = 15530.6$ $\qquad -(R^2 \cdot S_{yy} - S_{yy}) = SSE$

$\Rightarrow -(0.802 \cdot 15530.6 - 15530.6)$

$= 3075.0588$

(c) The realtor theorizes that square footage, x_1, is the most important predictor variable, and that the other variables can be left out without losing much prediction information. A simple linear regression of selling price vs. square footage was fit using the same 40 condos, and its SSE was 1553. Can the other independent variables, $x_2, x_3, and x_4$ be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

$H_0 = X_2 = X_3 = X_4 = 0$ H_A: X_2, X_3, X_4 is not 0.

use F test $\dfrac{(1553 - 3075)/(4-1)}{1553/(40-5)} = -11.43$

$F_{\alpha=0.05,\ \nu_1=3,\ \nu_2=35} = 3.23 > -11.43$

Ho Accepted.

$\Rightarrow$ other indep. Var can be dropped

3. A response Y is a function of three independent variables x_1, x_2, and x_3 that are related as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

(a) Fit this model to the $n = 7$ data points shown in the accompanying table.

y	x_1	x_2	x_3
-1	-3	5	-1
0	-2	0	1
0	-1	-3	1
1	0	-4	0
2	1	-3	-1
3	2	0	-1
3	3	5	1

$$X'Y = \begin{bmatrix} 10 \\ 14 \\ 10 \\ 3 \end{bmatrix}$$

From R lm method

$\beta_0 = 1.42857$

$\beta_1 = 0.500$

$\beta_2 = 0.11905$

$\beta_3 = -0.500$

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{7} & 0 & 0 & 0 \\ 0 & \frac{1}{28} & 0 & 0 \\ 0 & 0 & \frac{1}{84} & 0 \\ 0 & 0 & 0 & \frac{1}{6} \end{bmatrix}$$

$\hat{Y} = 1.42857 + 0.5X_1 + 0.1195X_2 - 0.5X_3$

(b) Predict Y when $x_1 = 1$, $x_2 = -3$, $x_3 = -1$. Compare the result with the observed data in row 5 of the table. Why are these values not equal?

2 1 -3 -1

$\hat{Y} = 1.42857 + 0.5 \cdot 1 + 0.1195 \cdot (-3) + 0.5$

$= 2.07007 \approx 2.$

There is Error in Linear model

(c) Do the data present sufficient evidence to indicate that x_3 contributes information for the prediction of Y? Test the hypothesis $H_0 : \beta_3 = 0$, using $\alpha = 0.05$.)

$$\text{T-test} = \frac{\hat{\beta}_3 - \mu_0}{S\sqrt{C_{11}}}$$

$$\frac{1}{6} \cdot (-3)$$

$$\| $$

$$(X'X)^{-1} X'Y$$

$$SSE_R = Y'Y - \hat{\beta} X'Y = \sum_{i=1}^{7} Y_i^2 - (-0.5) X'Y$$

$$= 24 - (-0.5)(-3)$$

$$= 22.5$$

$$S^2 = \frac{SSE}{n-4} = \frac{22.5}{7-4} = \frac{22.5}{3} = 7.5 \Rightarrow S = \sqrt{7.5}$$

$$SSE_C = Y'Y - \hat{\beta} X'Y = 24 - 1.55(-3)$$

$$= 28.6$$

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{(SSE_C)/(n-[k+1])} = \frac{(22.5 - 28.6)(4-3)}{28.6/(7-(4+1))}$$

$$= -.43$$

$F_{\alpha=0.05}$, $\nu_1 = 1$, $\nu_2 = 2$ $\Rightarrow$ 18.51 $\Rightarrow$ $F < F_\alpha$, H_0 Accepted.