

Fitting Regression Models With More Than One Regressor

Reference: Chapter 13 of Devore's 7th Edition

S. Maghsoodloo

Multiple linear regression (MLR, or MLREG) is the generalization of SLR (Simple Linear Regression) where the response, y , is modeled as a function of 2 or more regressor (or independent or explanatory) variables. Further, polynomial regression (PREG) is also a special case of MLR, as will be illustrated later, and therefore, PREG will not be discussed specifically herein. Consider the example 13.16 on pp. 541-542 of Devore, taken from the article "Applying Stepwise Multiple Regression Analysis to the Reaction of Formaldehyde with Cotton Cellulose" (Textile Research J., 1984: 157-165), where there are $k = 4$ regressor (or predictor) variables, and the dependent variable y represents durable press rating (a quantitative measure of resistance to wrinkle). The FLCs (Factor Level Combinations), \mathbf{X} , which is the 30×5 design matrix, and the 30×1 response vector \mathbf{Y} are reproduced below in Table 29 for your convenience, where $k = 4$ independent variables (or df)

Table 29. (The 30×5 design matrix \mathbf{X} and 30×1 vector-response \mathbf{Y} for the Example 13.16 of Devore)

\mathbf{X} (is the 30×5 design matrix)						\mathbf{Y} (is the 30×1 vector)	\mathbf{X} (continued)						\mathbf{Y}
FLC _{<i>i</i>}	X ₀	X ₁	X ₂	X ₃	X ₄	Response y _{<i>i</i>}	FLC _{<i>i</i>}	X ₀	X ₁	X ₂	X ₃	X ₄	y _{<i>i</i>}
1	1	8	4	100	1	1.4	16	1	4	10	160	5	4.6
2	1	2	4	180	7	2.2	17	1	4	13	100	7	4.3
3	1	7	4	180	1	4.6	18	1	10	10	120	7	4.9
4	1	10	7	120	5	4.9	19	1	5	4	100	1	1.7
5	1	7	4	180	5	4.6	20	1	8	13	140	1	4.6
6	1	7	7	180	1	4.7	21	1	10	1	180	1	2.6
7	1	7	13	140	1	4.6	22	1	2	13	140	1	3.1
8	1	5	4	160	7	4.5	23	1	6	13	180	7	4.7
9	1	4	7	140	3	4.8	24	1	7	1	120	7	2.5
10	1	5	1	100	7	1.4	25	1	5	13	140	1	4.5
11	1	8	10	140	3	4.7	26	1	8	1	160	7	2.1
12	1	2	4	100	3	1.6	27	1	4	1	180	7	1.8
13	1	4	10	180	3	4.5	28	1	6	1	160	1	1.5
14	1	6	7	120	7	4.7	29	1	4	1	100	1	1.3
15	1	10	13	180	3	4.8	30	1	7	10	100	7	4.6

The MLR model for the example in the above Table 29 is given by

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i = \mu_i + \epsilon_i, \quad (57)$$

where x_1 = HCHO (Formaldehyde Concentration), x_2 = Catalyst Ratio, x_3 = Curing Temperature, and x_4 = Curing Time, β_j 's ($j = 0, 1, 2, 3, k = 4$) are parameters (i.e., unknown constants), \mathbf{x}_0 is a 30×1 vector whose value is always equal to 1 for all $i = 1, 2, \dots, n = 30$ FLCs for this example, and $\mu_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$. For instance, the value of y_5 is modeled as $y_5 = 1 \times \beta_0 + 7\beta_1 + 4\beta_2 + 180\beta_3 + 5\beta_4 + \epsilon_5$, where ϵ_i 's ($i = 1, 2, 3, \dots, n = 30$) are assumed $\text{NID}(0, \sigma_\epsilon^2)$.

Henceforth, for convenience we will use the symbol σ^2 for the error variance σ_ϵ^2 . The reader must be cognizant of the fact that in classical regression theory, it is assumed that the design variables x_j ($j = 1, 2, \dots, k$) are fixed, i.e., the levels of all independent variables are selected without error (not at random) by the experimenter and hence $V(x_j) \equiv 0$ for all $j = 1, 2, \dots, k$, where $k = 4$ for Table 29. Only the classical regression is covered in this course, i.e., only y_i 's and ϵ_i 's in model (57) are random variables, while β_j 's and x_j 's are not rv's, and as a result y_i 's are $\text{NID}(\sum_{j=0}^k \beta_j x_{ij}, \sigma_\epsilon^2)$.

Exercise 102. Show that in the case of classical regression, $E(y_i) = \mu_i =$

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik}$, and $V(y_i) = \sigma^2$ for all $i = 1$ to n .

Our objective, just like in SLREG, is to estimate the $k+1$ ($= 5$ for our example) parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ in such a manner that the least squares function (LSF)

$$L(\beta_j) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})^2$$

is minimized. In order to minimize the above $L(\beta_j)$ wrt the parameters β_j ($j = 0, 1, 2, 3, 4$), the partial derivatives of the LSF, $\partial L / \partial \beta_j$, must be required to equal zero for all j . This generally leads to a system of $(k+1)$ least squares normal equations (LSNEs), which must be solved simultaneously for the $k+1$ unknowns $\hat{\beta}_j$ ($j = 0, 1, 2, 3, k = 4$). The $k+1$ ($= 5$ for Table 29) partial derivatives are provided below

$$\partial L / \partial \beta_0 = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4}) (-1) \quad (58a)$$

$$\partial L / \partial \beta_1 = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4}) (-x_{i1}) \quad (58b)$$

$$\partial L / \partial \beta_2 = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4}) (-x_{i2}) \quad (58c)$$

$$\partial L / \partial \beta_3 = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4}) (-x_{i3}) \quad (58d)$$

$$\partial L / \partial \beta_4 = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4}) (-x_{i4}) \quad (58e)$$

The RHS of Eq. (58a) when set equal to zero leads to the 1st LS normal equation as

$$\sum_{i=1}^n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \hat{\beta}_3 \sum_{i=1}^n x_{i3} + \hat{\beta}_4 \sum_{i=1}^n x_{i4} = \sum_{i=1}^{30} y_i$$

Dividing both sides of the above equation by the number of FLCs, n, results in

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_3 \bar{x}_3 + \hat{\beta}_4 \bar{x}_4 = \bar{y} \quad (59a)$$

$$\rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 - \hat{\beta}_3 \bar{x}_3 - \hat{\beta}_4 \bar{x}_4$$

The RHS's of Equations (58 b, c, d, & e) when set equal to zero give rise to the other 4 LS normal equations, respectively.

$$\hat{\beta}_0 \sum x_1 + \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2 + \hat{\beta}_3 \sum x_1 x_3 + \hat{\beta}_4 \sum x_1 x_4 = \sum x_1 y \quad (59b)$$

$$\hat{\beta}_0 \sum x_2 + \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2 + \hat{\beta}_3 \sum x_2 x_3 + \hat{\beta}_4 \sum x_2 x_4 = \sum x_2 y \quad (59c)$$

$$\hat{\beta}_0 \sum x_3 + \hat{\beta}_1 \sum x_1 x_3 + \hat{\beta}_2 \sum x_2 x_3 + \hat{\beta}_3 \sum x_3^2 + \hat{\beta}_4 \sum x_3 x_4 = \sum x_3 y \quad (59d)$$

$$\hat{\beta}_0 \sum x_4 + \hat{\beta}_1 \sum x_1 x_4 + \hat{\beta}_2 \sum x_2 x_4 + \hat{\beta}_3 \sum x_3 x_4 + \hat{\beta}_4 \sum x_4^2 = \sum x_4 y \quad (59e)$$

In equations (59), we have removed the index i from all summations only for convenience, i.e., all summations range from $i = 1$ to $i = n$, where $n = 30$ for this Example. Using the data in Table 29, we obtain the following raw (uncorrected) statistics: $n = 30$,

$$\begin{aligned} \sum x_1 &= 182, \quad \sum x_2 = 204, \quad \sum x_3 = 4280, \quad \sum x_4 = 118, \quad \bar{x}_1 = 6.06667, \quad \bar{x}_2 = 6.80, \\ \bar{x}_3 &= 142.66667, \quad \bar{x}_4 = 3.93333, \quad \sum x_1^2 = 1266, \quad \sum x_1 x_2 = 1253, \quad \sum x_1 x_3 = 26160, \\ \sum x_1 x_4 &= 706, \quad \sum x_2^2 = 1998, \quad \sum x_2 x_3 = 29180, \quad \sum x_2 x_4 = 766, \quad \sum x_3^2 = 639200, \\ \sum x_3 x_4 &= 16720, \quad \sum x_4^2 = 670, \quad \sum y_i = 106.80, \quad \sum x_1 y = 678.50, \quad \sum x_2 y = 860.1, \\ \sum x_3 y &= 15594.00, \quad \sum_{i=1}^{30} x_{i4} y_i = 430.20, \quad \text{USS} = \sum_{i=1}^{30} y_i^2 = 437.080, \quad \text{and } \bar{y} = 3.56. \end{aligned}$$

Substituting the 20 pertinent statistics out of the above 26 into equations (59) yields a set of 5 LS normal equations with 5 unknowns, listed below, for Table 29.

$$\left\{ \begin{array}{l} 30\hat{\beta}_0 + 182\hat{\beta}_1 + 204\hat{\beta}_2 + 4280\hat{\beta}_3 + 118\hat{\beta}_4 = 106.8 \\ 182\hat{\beta}_0 + 1266\hat{\beta}_1 + 1253\hat{\beta}_2 + 26160\hat{\beta}_3 + 706\hat{\beta}_4 = 678.5 \\ 204\hat{\beta}_0 + 1253\hat{\beta}_1 + 1998\hat{\beta}_2 + 29180\hat{\beta}_3 + 766\hat{\beta}_4 = 860.1 \\ 4280\hat{\beta}_0 + 26160\hat{\beta}_1 + 29180\hat{\beta}_2 + 639200\hat{\beta}_3 + 16720\hat{\beta}_4 = 15594 \\ 118\hat{\beta}_0 + 706\hat{\beta}_1 + 766\hat{\beta}_2 + 16720\hat{\beta}_3 + 670\hat{\beta}_4 = 430.2 \end{array} \right. \quad (60a)$$

Eqs. (60a) show that $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$, where $\mathbf{X}'\mathbf{Y} = \mathbf{X}^T\mathbf{Y}$ is the 5×1 vector on the RHS of (60a), and $\mathbf{X}'\mathbf{X} = \mathbf{X}^T\mathbf{X}$ is a 5×5 matrix of coefficients given below.

$$\mathbf{A} = (\mathbf{X}'\mathbf{X}) = (\mathbf{X}^T\mathbf{X}) = \begin{bmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \\ 30 & 182 & 204 & 4280 & 118 \\ 182 & 1266 & 1253 & 26160 & 706 \\ 204 & 1253 & 1998 & 29180 & 766 \\ 4280 & 26160 & 29180 & 639200 & 16720 \\ 118 & 706 & 766 & 16720 & 670 \end{bmatrix} \quad (60b)$$

One way to solve the above system of 5 equations with 5 unknowns of Eq. (60a) is to use Cramer's Rule. Accordingly, we define 5 other matrices \mathbf{A}_j ($j = 0, 1, 2, 3, 4$) as follows:

The 5×5 matrix of coefficients in equation (60b) and the matrix \mathbf{A}_j are identical

except for its j^{th} column, which is $\mathbf{COL}_j = \mathbf{X}'\mathbf{Y} = [106.8 \ 678.5 \ 860.1 \ 15594 \ 430.2]^T$, i.e., for $k = 4$, $\mathbf{COL}_j = [\sum y \ \sum x_1y \ \sum x_2y \ \sum x_3y \ \sum x_4y]^T$, and the $n \times (k+1) = 30 \times 5$ design matrix \mathbf{X} is given in Table 29. Then, by Cramer's Rule, $\hat{\beta}_j = \det(\mathbf{A}_j) / \det(\mathbf{A})$ for $j = 0, 1, 2, 3, 4$. For example, the matrix \mathbf{A}_2 for the data of Table 29 is given by

$$\mathbf{A}_2 = \begin{bmatrix} n & \sum x_1 & \sum y & \sum x_3 & \sum x_4 \\ \sum x_1 & \sum x_1^2 & \sum x_1y & \sum x_1x_3 & \sum x_1x_4 \\ \sum x_2 & \sum x_1x_2 & \sum x_2y & \sum x_2x_3 & \sum x_2x_4 \\ \sum x_3 & \sum x_1x_3 & \sum x_3y & \sum x_3^2 & \sum x_3x_4 \\ \sum x_4 & \sum x_1x_4 & \sum x_4y & \sum x_3x_4 & \sum x_4^2 \end{bmatrix},$$

or

$$\mathbf{A}_2 = \begin{bmatrix} 30 & 182 & 106.8 & 4280 & 118 \\ 182 & 1266 & 678.5 & 26160 & 706 \\ 204 & 1253 & 860.1 & 29180 & 766 \\ 4280 & 26160 & 15594 & 639200 & 16720 \\ 118 & 706 & 430.2 & 16720 & 670 \end{bmatrix}.$$

The matrix $\mathbf{A} = \mathbf{X}'\mathbf{X} = \mathbf{X}^T\mathbf{X}$ is always symmetrical, while the matrices \mathbf{A}_j ($j = 0, 1, 2, \dots, k = 4$) are not, in general, symmetric. Excel (or Matlab) computations give $\det(\mathbf{A}) = 17.0147010528 \times 10^{12}$, $\det(\mathbf{A}_0) = -15.521016533 \times 10^{12}$, $\det(\mathbf{A}_1) = 2.734711393 \times 10^{12}$, $\det(\mathbf{A}_2) = 3.73954437984 \times 10^{12}$, $\det(\mathbf{A}_3) = 0.1909996304 \times 10^{12}$, and $\det(\mathbf{A}_4) = 1.73506460544 \times 10^{12}$. Hence, $\hat{\beta}_0 = \det(\mathbf{A}_0) / \det(\mathbf{A}) = -0.912212$, $\hat{\beta}_1 = \det(\mathbf{A}_1) / \det(\mathbf{A}) = 0.1607264$, $\hat{\beta}_2 = 0.2197831$, $\hat{\beta}_3 = 0.0112256$, and $\hat{\beta}_4 = 0.1019744$. These 5 estimates of β_j 's give rise to the following fitted MLREG model:

$$\hat{y}_i = -0.91221x_{i0} + 0.16073x_{i1} + 0.21978x_{i2} + 0.011226x_{i3} + 0.10197x_{i4}. \quad (61)$$

Notice that the coefficients of the above regression model are in complete agreement with those of Devore's in the Table at the bottom of his page 541. Further, at 1st glance, the regressor variable x_2 in equation (61) seems to have the largest impact on the response variable y because its coefficient 0.21978 is the largest in absolute value. The true (or net, or partial) statistical influence of the 4 independent variables x_j ($j = 1, 2, 3, 4$) on the dependent variable y , which represents resistance to wrinkle, will be determined mostly by $t_{n-k-1} = \hat{\beta}_j / se(\hat{\beta}_j)$, $j = 1, 2, 3, 4$.

We now develop a general matrix algebra approach for obtaining the coefficient estimates in a MLR (or MLREG) model. The symbol ' or ^T will denote matrix transpose and large bolded capital letter is used to represent a matrix. The necessary matrices, including the design matrix \mathbf{X} that is composed of n FLCs, are again defined below.

For our Example, the dimension of the vector \mathbf{Y} is 30×1 , that of matrix \mathbf{X} is 30×5 , $\mathbf{X}' = \mathbf{X}^T$ is 5×30 , that of vector $\boldsymbol{\beta}$ is 5×1 , and $\boldsymbol{\epsilon}$ is a 30×1 vector; clearly, $\mathbf{A} = \mathbf{X}'\mathbf{X} = \mathbf{X}^T\mathbf{X}$ is a 5×5 symmetric matrix.

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{matrix} & \mathbf{x}_0 & \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_k \\ \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \end{matrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}. \quad (62a)$$

First, we rewrite the MLR model (57), which is valid only for the i^{th} observation y_i , for all the n values of \mathbf{Y} in matrix form using (62a).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (62b)$$

In order to obtain the least-squares estimate of the 5×1 vector $\boldsymbol{\beta}$, we first use the fact that

the LSF in matrix form is given by

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n=30} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{Y}) + \boldsymbol{\beta}'\mathbf{A}\boldsymbol{\beta} \end{aligned}$$

Second, we take the partial derivative of $L(\boldsymbol{\beta})$ wrt the vector $\boldsymbol{\beta}$ and will require $\partial L(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ to equal to the zero vector, $\mathbf{0} = [0 \ 0 \ 0 \ 0 \ 0]^\top$, in order to minimize the LSF with respect to all the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ ($k = 4$ for our example). In the following matrix development, bear in mind that $\mathbf{Y}'\mathbf{Y}$ ($= \sum_{i=1}^{n=30} y_i^2 =$ the USS) is independent of the 5×1 column vector $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \beta_2 \ \beta_3 \ \beta_4]^\top$.

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \partial L / \partial \beta_0 \\ \partial L / \partial \beta_1 \\ \cdot \\ \partial L / \partial \beta_k \end{bmatrix} = 0 - 2(\mathbf{X}'\mathbf{Y}) + 2(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} \xrightarrow{\text{Set to}} \mathbf{0} \quad (63)$$

Eq. (63) yields the heterogeneous system $-(\mathbf{X}'\mathbf{Y}) + (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{0}$ of 5 ($= k+1$) equations with 5 unknowns whose solutions can now easily be obtained by first transposing $[-(\mathbf{X}'\mathbf{Y})]$ to the RHS and then multiplying both sides by the symmetric $(k+1) \times (k+1) = 5 \times 5$ matrix $\mathbf{C} = \mathbf{A}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$, i.e., $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{A}\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Y}) \rightarrow (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \rightarrow$

$$\text{Or} \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) = \mathbf{A}^{-1}(\mathbf{X}'\mathbf{Y}) = \mathbf{C}(\mathbf{X}'\mathbf{Y}), \quad (64)$$

where, again, the $(k+1) \times (k+1)$ matrix $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of $\mathbf{A} = \mathbf{A}^{-1}$. Further,

like $(\mathbf{X}'\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})$, the matrix \mathbf{C} is also square and symmetrical. Applying Eq. (64)

to the data of Table 29, we obtain

$$\hat{\beta} = \mathbf{C}(\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) = \mathbf{A}^{-1}(\mathbf{X}'\mathbf{Y})$$

$$= \begin{bmatrix} 1.09527 & -0.03213 & -0.01123 & -0.004844 & -0.02533 \\ & 0.006256 & -0.000138 & -4.1223 \times 10^{-5} & 0.000253 \\ & & 0.001658 & -2.3269 \times 10^{-6} & 0.000285 \\ & & & 3.53376 \times 10^{-5} & 1.73 \times 10^{-5} \\ & & & & 0.00493 \end{bmatrix} \times \begin{bmatrix} 106.8 \\ 678.5 \\ 860.1 \\ 15594 \\ 430.2 \end{bmatrix}$$

$\hat{\beta} =$ \mathbf{C} \times $(\mathbf{X}'\mathbf{Y})$

$$\hat{\beta} = \begin{bmatrix} -0.91221 \\ 0.16073 \\ 0.21978 \\ 0.01123 \\ 0.10197 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix}. \quad (\text{Note that for the above example, } \mathbf{C} = \mathbf{A}^{-1} \text{ is a } 5 \times 5 \text{ matrix, while } \mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{Y} \text{ is a } 5 \times 1 \text{ vector.})$$

The components of the above 5×1 vector $\hat{\beta}$ are identical to the previous LS estimates using Cramer's Rule given in Eq. (61) on page 177.

Exercise 103. Use Excel to verify that $\mathbf{X}'\mathbf{X} = \mathbf{A}$, which is given in Eq. (60b) for this Example, and on the same sheet verify the elements of the above 5×1 vector estimator $\hat{\beta}$ in the 2 different methods outlined so far.

Excel Procedure For Matrix Algebra

Step 1: Enter the elements of the $n \times (k+1)$ design matrix \mathbf{X} on a spreadsheet and highlight (HL) the entire array \rightarrow Name Range \rightarrow Name \rightarrow \mathbf{X} \rightarrow ok.

For Table 29, You now have a 30×5 design matrix \mathbf{X} .

Step 2: HL (highlight) a 5×30 area → = Transpose(\mathbf{X}); do not use the Excel Transpose function because it does not work all the times. Then Shift-Ctrl-Enter. HL this 5×30 area that contains $\mathbf{X}' = \mathbf{X}^T$ → Right-click → Name Range → Name: XT and ok.

Step 3: HL a $(k+1) \times (k+1) = 5 \times 5$ area → = MMULT(XT, X) →

Shift -Ctrl -Enter → HL this 5×5 area again → Name Range → Name: \mathbf{A} → ok.

Step 4: HL another $(k+1) \times (k+1) = 5 \times 5$ area → = MINVERSE(\mathbf{A}) → Shift-Ctrl-Enter → You now have the AINV → HL the AINV matrix → Name Range → Assign name AINV → to this area. Note that AINV = \mathbf{C} ; it seems that Excel does not allow naming any variable as \mathbf{C} ? Repeat the above steps until the Betahat = $\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}'\mathbf{Y}) = \mathbf{C}(\mathbf{X}'\mathbf{Y})$ column vector is obtained. Compare your answer against the model (61) at the bottom of p. 177.

Residuals in MLREG

Recall that by definition a model residual $e_i = y_i - \hat{y}_i$, where \hat{y}_i is given

by the model (61) on page 177. Equation (62b) shows that $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ (because the best predictor of the vector $\boldsymbol{\epsilon}$ is the $\mathbf{0}$ vector), and therefore, the fitted vector $\hat{\mathbf{Y}}$ for all the $n = 30$ observations of the Example 13.16 of Devore on his p. 541 is given by

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}_{30} \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{C}(\mathbf{X}'\mathbf{Y}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y}, \text{ or}$$

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = (\mathbf{X}\mathbf{C}\mathbf{X}')\mathbf{Y} = (\mathbf{X}\mathbf{C}\mathbf{X}^T)\mathbf{Y} \rightarrow \mathbf{H} = \mathbf{X}\mathbf{C}\mathbf{X}^T$$

The $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}\mathbf{C}\mathbf{X}^T$ is called the Hat matrix because it projects the response vector \mathbf{Y} onto the vector $\hat{\mathbf{Y}}$ (or **Y-hat**) through the matrix equation $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. The $n \times 1$ residual vector, therefore, is given by $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, and as a result,

$$SS_{\text{Residuals}} = \sum_{i=1}^n e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}'\mathbf{Y} - 2\hat{\mathbf{Y}}'\mathbf{Y} + \hat{\mathbf{Y}}'\hat{\mathbf{Y}} \quad (65)$$

However, $\hat{\mathbf{Y}}'\mathbf{Y} = (\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{Y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{Y}) = \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$, where we have made use of Eq. (64) which shows that $\mathbf{C}^{-1}\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$. Note that $\hat{\mathbf{Y}}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$,

implies that $\sum_{i=1}^n \hat{y}_i y_i = \sum_{i=1}^n \hat{y}_i^2$. Substituting $\hat{\mathbf{Y}}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ into Eq. (65) results in $SS_{\text{Residuals}} =$

$$\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2 = \left(\sum_{i=1}^n y_i^2 - CF\right) - \left(\sum_{i=1}^n \hat{y}_i^2 - CF\right). \text{ Hence,}$$

$$SS_{\text{Residuals}} = SS(\text{Error}) = SS(\text{Unexplained}) = SS_T - SS(\text{Explained}), \quad (66)$$

where $SS(\text{Explained}) = SS(\text{Model}) = SS(\text{Regression}) = \sum_{i=1}^n \hat{y}_i^2 - CF$. Equation (66) is similar to

ANOVA where $SS(\text{Total}) = SS(\text{Model}) + SS(\text{Error}) = SS(\text{Explained}) + SS(\text{Unexplained})$ for all statistical models.

Definition. A matrix, \mathbf{B} , is said to be idempotent iff $\mathbf{B}^2 = \mathbf{B}$. For example, the identity matrix, \mathbf{I}_n , is idempotent.

Exercise 104. Show that the three $n \times n$ matrices \mathbf{H} , \mathbf{I}_n , and $\mathbf{I}_n - \mathbf{H}$ are all symmetrical and idempotent, where \mathbf{I}_n is an $n \times n$ identity matrix.

Exercise 105. Use Eq. (59a) to show that the model

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4}$, having 4 regressors, is also given by

$$\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_{i1} - \bar{x}_1) + \hat{\beta}_2 (x_{i2} - \bar{x}_2) + \hat{\beta}_3 (x_{i3} - \bar{x}_3) + \hat{\beta}_4 (x_{i4} - \bar{x}_4).$$

Then, prove that $\sum_{i=1}^n e_i = \sum (y_i - \hat{y}_i) = 0$, which implies that $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$. Therefore, the

CF for SS(Regression), which is $(\sum_{i=1}^n \hat{y}_i)^2 / n$, is also equal to $(\sum_{i=1}^n y_i)^2 / n$. As a result, show

$$\text{that } SS(\text{Reg}) = SS(\text{Explained}) = \sum_{i=1}^n \hat{y}_i^2 - CF = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

We now develop matrix formulas for $SS(\text{Total}) = SS_T = S_{yy}$, $SS(\text{Model}) = SS(\text{Reg})$, and $SS(\text{Residuals}) = SS(\text{Unexplained})$. To this end, let \mathbf{I}_n be the $n \times n$ identity matrix and the vector $\mathbf{1}$

be an $n \times 1$ column vector every element of which is equal to 1, i.e., $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$, $\mathbf{1}' = [1 \quad 1 \dots 1]$,

where $\mathbf{1}'$ is an $1 \times n$ row vector, and $\mathbf{J} = (\mathbf{1}\mathbf{1}')/n$ is an $n \times n$ matrix all of whose elements are equal to $1/n$. From the above matrix definitions, we deduce that

$$CF = \left(\sum_{i=1}^n y_i \right)^2 / n = (\mathbf{1}'\mathbf{Y})^2 / n = (\mathbf{Y}'\mathbf{1})(\mathbf{1}'\mathbf{Y}) / n = \mathbf{Y}'[(\mathbf{1}\mathbf{1}')/n]\mathbf{Y} = \mathbf{Y}'\mathbf{J}\mathbf{Y}. \quad (67a)$$

$$SS_T = \mathbf{Y}'\mathbf{Y} - CF = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{J})\mathbf{Y} \quad (67b)$$

$$SS(\text{Reg}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} - \mathbf{Y}'\mathbf{J}\mathbf{Y} = (\mathbf{H}\mathbf{Y})'(\mathbf{H}\mathbf{Y}) - \mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'(\mathbf{H}'\mathbf{H})\mathbf{Y} - \mathbf{Y}'\mathbf{J}\mathbf{Y} \rightarrow$$

$$SS_{\text{Model}} = \mathbf{Y}'\mathbf{H}\mathbf{Y} - \mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'(\mathbf{H} - \mathbf{J})\mathbf{Y}, \quad (67c)$$

where we have made use of the fact that the $n \times n$ matrix \mathbf{H} is idempotent. Thus,

$$SS(\text{RES}) = SS(\text{Total}) - SS(\text{Model}) = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}. \quad (67d)$$

Exercise 106. Prove that $SS(\text{Regression}) = \mathbf{Y}'(\mathbf{H} - \mathbf{J})\mathbf{Y} = \sum_{j=1}^k \hat{\beta}_j S_{jy}$, where $S_{jy} =$

$$\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y}) = \sum_{i=1}^n (x_{ij} - \bar{x}_j)y_i = \sum_{i=1}^n x_{ij} y_i - (\sum_{i=1}^n x_{ij})(\sum_{i=1}^n y_i) / n = \sum_{i=1}^n x_{ij} y_i -$$

$\bar{x}_j(\sum_{i=1}^n y_i)$. Hint: $SS(\text{Reg}) = \mathbf{Y}'\mathbf{H}\mathbf{Y} - CF = \mathbf{Y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} - CF = \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{Y}) - CF$. Further, show

that for our example, $S_{1y} = 30.58$, $S_{2y} = 133.86$, $S_{3y} = 357.20$, and $S_{4y} = 10.12$.

Example 46. We now use the equations (67), developed above, to obtain the $SS_T = SS(\text{Total}) = USS - CF = 437.08 - (106.80^2)/30 = 56.872$ (with $30 - 1 = 29$ df). Next, we compute

$$SS_{\text{Reg}} = \mathbf{Y}'(\mathbf{H} - \mathbf{J})\mathbf{Y} = \sum_{j=1}^4 \hat{\beta}_j S_{jy} = (0.1607264)(30.58) + 0.2197831(133.86) + 0.0112256$$

$$(357.20) + 0.1019774(10.12) = 39.37694 = SS(\text{Model}) \text{ with } 4 \text{ df, where } S_{1y} = \sum_{i=1}^n (x_{i1} - \bar{x}_1)y_i =$$

$$\sum_{i=1}^n x_{i1} y_i - \bar{x}_1 \sum_{i=1}^n y_i = 678.50 - 6.066667(106.80) = 30.580, \text{ etc. Therefore, } SS_{\text{RES}} = SS(\text{Total}) -$$

$SS(\text{Model}) = 56.8720 - 39.3769 = 17.4951$ (with $29 - 4 = 25$ df). The exact P -value for

$F_0(\text{Model}) = 14.0672$ is $\hat{\alpha} = 0.00000385$.

In order to develop CIs (confidence intervals) and conduct tests of hypotheses on the parameter vector $\boldsymbol{\beta}$, we need to show that if \mathbf{B} is any $p \times n$ constant matrix and \mathbf{Y} is an $n \times 1$ random vector, then the $\text{COV}(\mathbf{B}\mathbf{Y}) = \mathbf{B}\text{COV}(\mathbf{Y})\mathbf{B}'$, where $E(\mathbf{Y}) = \boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \dots \quad \mu_n]'$ is an $n \times 1$ parameter vector of the n population means, where $\mu_1 = E(y | \text{at FLC}_1)$, $\mu_2 = E(y | \text{at FLC}_2)$, etc.

Proof. By definition, $\text{COV}(\mathbf{B}\mathbf{Y}) = E[(\mathbf{B}\mathbf{Y} - \mathbf{B}\boldsymbol{\mu})(\mathbf{B}\mathbf{Y} - \mathbf{B}\boldsymbol{\mu})^T] =$

$$E[(\mathbf{B}\mathbf{Y} - \mathbf{B}\boldsymbol{\mu})(\mathbf{Y}'\mathbf{B}' - \boldsymbol{\mu}'\mathbf{B}')] = E[\mathbf{B}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y}' - \boldsymbol{\mu}')\mathbf{B}'] = \mathbf{B}E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y}' - \boldsymbol{\mu}')]\mathbf{B}'$$

$$= \mathbf{BCOV}(\mathbf{Y})\mathbf{B}' = \mathbf{BCOV}(\mathbf{Y})\mathbf{B}^T.$$

Exercise 107. Use the above covariance property, and the fact that under the regression model (67b) the $\mathbf{COV}(\mathbf{Y}) = \mathbf{I}_n \sigma_\epsilon^2$, to show that (a) $\mathbf{COV}(\hat{\mathbf{Y}}) = \mathbf{H} \sigma_\epsilon^2$, (b) $\mathbf{COV}(\mathbf{e}) = (\mathbf{I}_n - \mathbf{H}) \times \sigma_\epsilon^2$, where \mathbf{e} is the $n \times 1$ residual vector, and (c) $\mathbf{COV}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma_\epsilon^2 = \mathbf{C} \sigma^2$, where the diagonal elements of the symmetric $(k+1) \times (k+1)$ matrix $\mathbf{COV}(\hat{\boldsymbol{\beta}})$ give the $V(\hat{\beta}_j)$, $j = 0, 1, 2, \dots, k$ and its off-diagonal elements give the $\text{COV}(\hat{\beta}_j, \hat{\beta}_r)$ for $r = 0, 1, 2, \dots, k \neq j$.

Part (b) of Exercise 107 shows that the $V(e_i) = (1 - h_{ii}) \sigma_\epsilon^2$, and therefore, the Studentized residuals are given by

$$r_i = e_i / \sqrt{(1 - h_{ii}) \times \text{MS}(\text{RES})}, \quad i = 1, 2, \dots, n. \quad (67e)$$

where h_{ii} is the amount of leverage exerted by y_i on \hat{y}_i . The diagonal elements of \mathbf{H} for Table 29 are $h_{11} = 0.190347$, $h_{22} = 0.248039$, $h_{33} = 0.141299$, ..., $h_{30,30} = 0.172067$. Any FLC with large h_{ii} , and consequently, with large r_i (say with absolute value of r_i greater than 2), is highly influential on the least squares fit. Minitab provides an option that lists the vectors $\hat{\mathbf{Y}}$, the $se(\hat{y}_i) = \sqrt{h_{ii} \text{MS}(\text{RES})}$, the vector \mathbf{e} , and Studentized Residuals r_i . Notice that Minitab uses the designation of SRES for the Studentized residuals r_i . For our Example 46 (or Table 29), the largest r_i in absolute value is $r_9 = 2.04$, which implies that the FLC number 9, $\mathbf{FLC}_9 = [1 \quad 4 \quad 7 \quad 140 \quad 3]^T$, has the highest influence on the regression coefficients while $r_{24} = 0.01$ implies that the $\mathbf{FLC}_{24} = [1 \quad 7 \quad 1 \quad 120 \quad 7]^T$ has almost no impact on $\hat{\beta}_j$ ($j = 0, 1, 2, 3, 4$). As a matter of fact, I removed the $\mathbf{FLC}_{24} = [1 \quad 7 \quad 1 \quad 120 \quad 7]^T$ from the design matrix \mathbf{X} of Table 29 and used Minitab to obtain the following fitted model:

$$\hat{y}_{(24)} = -0.9130x_0 + 0.16066x_1 + 0.21985x_2 + 0.01123x_3 + 0.10187x_4,$$

which is almost identical to the previous regression model \hat{y}_i given in equation (61).

However, if the $\mathbf{FLC}_9 = [1 \quad 4 \quad 7 \quad 140 \quad 3]^T$ is removed from the design matrix \mathbf{X} , the resulting regression model is $\hat{y}_{(9)} = -1.14930x_0 + 0.18387x_1 + 0.21915x_2 + 0.01127x_3 + 0.11102x_4$, whose coefficients are not as close as $\hat{y}_{(24)}$ to the model (61).

Definition. The trace of a square matrix is simply the sum of its diagonal elements.

For example, the Trace (\mathbf{I}_n) = n, the $\text{Tr} \begin{bmatrix} 5 & 9 \\ 4 & -5 \end{bmatrix} = 0$, $\text{Tr}(\mathbf{A}_2) = 642026.10$ (see page 177),

$\text{Tr}(12) = 12$ and the $\text{Tr}(\mathbf{X}^T\mathbf{X}) = n + \sum_{i=1}^k x_{ii}^2$.

Exercise 108. (a) Let \mathbf{A} and \mathbf{B} be 2 compatible matrices (not necessarily square); show that $\text{Tr}(\mathbf{A}\times\mathbf{B}) = \text{Tr}(\mathbf{B}\times\mathbf{A})$ iff $\mathbf{A}\times\mathbf{B}$ is square; further, if \mathbf{A} and \mathbf{B} are square matrices $\text{Tr}(\mathbf{A} + \mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B})$, $\text{Tr}(c\mathbf{A}) = c\text{Tr}(\mathbf{A})$ for any scalar constant c , and $\text{Tr}(\mathbf{B}^{-1}\mathbf{A}\mathbf{B}) = \text{Tr}(\mathbf{A})$. (b) Use these trace properties of a matrix to prove that $E(\text{SS}_{\text{Residuals}}) = (n - k - 1)\sigma_{\epsilon}^2$ and hence, an unbiased estimator of σ_{ϵ}^2 is $\hat{\sigma}_{\epsilon}^2 = \text{MS}_{\text{RES}} = (\text{SS}_{\text{Residuals}})/(n - k - 1)$ for all classical regression models. As an application of the above stated properties for Trace of a matrix, it follows that the $\text{Trace}(\mathbf{H}) = \text{Trace}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{Trace}[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] = \text{Trace}(\mathbf{I}_{k+1}) = k+1$. Thus, any $h_{ii} > 2(k+1)/n$ is considered significantly larger than its expected value of $(k+1)/n$, and hence, highly influential.

CONFIDENCE INTERVALS FOR β_j ($j = 0, 1, 2, 3, \dots, k$)

The main assumption in MLR is that y_i 's ($i = 1, 2, \dots, n$) are $\text{NID}(\mu_i, \sigma_{\epsilon}^2)$. Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) = \mathbf{C}(\mathbf{X}'\mathbf{Y})$, every $\hat{\beta}_j$ is a linear combination of y_i 's ($i = 1, 2, \dots, n$), resulting in the normality of each $\hat{\beta}_j$ with $E(\hat{\beta}_j) = \beta_j$. Further, from Exercise 107 above the $se(\hat{\beta}_j) = (C_{jj}\text{MS}_{\text{Residuals}})^{1/2}$, where C_{jj} is the diagonal element of the matrix $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ pertaining to β_j , $j =$

0, 1, 2, ..., k. Note that C_{00} is the element in the first row and 1st column of \mathbf{C} . Therefore, a 95% t-CI for the parameter β_j is given by

$$\hat{\beta}_j \pm t_{0.025, n-k-1} \times se(\hat{\beta}_j). \quad (68)$$

If the interval in equation (68) excludes 0, then the null hypothesis $H_0: \beta_j = 0$ must be rejected at the 5% LOS for any $j = 0, 1, 2, \dots, k$. This, in turn, will imply that the regressor variable x_j ($j = 1, 2, \dots, k$) has a statistically significant impact on the response variable y at the 5% level.

Further, under the null hypothesis $H_0: \beta_j = 0$, the statistic $\hat{\beta}_j / se(\hat{\beta}_j)$ has a W. S. Gosset's Student t-distribution with $(n - 1 - k)$ *df*. The reader must be cognizant of the fact that $\hat{\beta}_0$ adds no contribution to the regression of y on x_j 's, $j = 1, 2, \dots, k$.

Example 46 Continued. (c) We now use equation (68) in order to obtain the 95% CI for β_1 . From $\mathbf{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma_{\epsilon}^2 = \mathbf{C} \sigma_{\epsilon}^2$, we deduce that the $se(\hat{\beta}_1) = (C_{11} MS_{RES})^{1/2} = (0.006256 \times 0.6998)^{1/2} = 0.06617 \rightarrow HCIL = t_{0.025, 25} \times se(\hat{\beta}_1) = 2.060 \times 0.06617 = 0.1363 \rightarrow 0.16073 \pm 0.1363 \rightarrow 0.02442 \leq \beta_1 \leq 0.29703$; since this 95% CI excludes zero, the null hypothesis $H_0: \beta_1 = 0$ must be rejected at the 5% LOS. Thus, the effect of x_1 on Y is statistically significant at the 5% level.

Further, if we wish to directly test $H_0: \beta_1 = 0$, W/O obtaining a CI, we may compute the statistic $t_0 = \hat{\beta}_1 / se(\hat{\beta}_1) = 0.16073 / 0.06617 = 2.429$ and compare it against the threshold value $t_{0.025, 25} = 2.060$.

Exercise 109. Obtain the 95% CI's for β_2 , β_3 and β_4 of the Example 46 above and use them to test the null hypotheses $H_0: \beta_2 = 0$, $H_0: \beta_3 = 0$ and $H_0: \beta_4 = 0$ at $\alpha = 0.05$. Further, compute all the three t statistics and compare your results against the 2.5 percentage point of t, making statistical inferences.

CONFIDENCE INTERVAL FOR THE MEAN RESPONSE μ_0

Let $\mathbf{X}_0 = [1 \ x_{01} \ x_{02} \ \dots \ x_{0k}]^T$ be a specified **FLC** (within the range of the **X** factor space) so that $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \dots + \hat{\beta}_k x_{0k} = \hat{\beta}'\mathbf{x}_0 = \mathbf{x}_0'\hat{\beta}$ is an unbiased estimator of $E(y | \mathbf{x}_0) = \mu_0 = \mathbf{x}_0'\boldsymbol{\beta} = \beta_0 x_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_k x_{0k} = \boldsymbol{\beta}'\mathbf{x}_0$. The reader should not confuse the $(k+1) \times 1$ vector \mathbf{X}_0 with the $n \times 1$ vector \mathbf{x}_0 of Table 29. In order to obtain the $se(\hat{y}_0)$, we must 1st compute the $V(\hat{y}_0)$.

$$\begin{aligned} V(\hat{y}_0) &= V(\mathbf{x}_0'\hat{\beta}) = E[(\mathbf{x}_0'\hat{\beta} - \mathbf{x}_0'\boldsymbol{\beta})(\mathbf{x}_0'\hat{\beta} - \mathbf{x}_0'\boldsymbol{\beta})'] = E[\mathbf{x}_0'(\hat{\beta} - \boldsymbol{\beta})(\hat{\beta} - \boldsymbol{\beta})'\mathbf{x}_0] \\ &= \mathbf{x}_0' E[(\hat{\beta} - \boldsymbol{\beta})(\hat{\beta} - \boldsymbol{\beta})'] \mathbf{x}_0 = \mathbf{x}_0' \text{COV}(\hat{\beta}) \mathbf{x}_0 = \mathbf{x}_0' (\mathbf{C} \sigma_\epsilon^2) \mathbf{x}_0 \rightarrow \end{aligned}$$

$$V(\hat{y}_0) = (\mathbf{x}_0' \mathbf{C} \mathbf{x}_0) \sigma_\epsilon^2 \quad (69a)$$

Eq. (69a) clearly shows that the

$$se(\hat{y}_0) = [(\mathbf{x}_0' \mathbf{C} \mathbf{x}_0) \times MS_{RES}]^{1/2}. \quad (69b)$$

Therefore, the 95% CI for the $\mu_0 = E(y | \text{at } \mathbf{X}_0)$, the mean of y at \mathbf{X}_0 , is given by

$$\mathbf{x}_0'\hat{\beta} - t_{0.025, n-k-1} \times se(\hat{y}_0) \leq \mu_0 \leq \mathbf{x}_0'\hat{\beta} + t_{0.025, n-k-1} \times se(\hat{y}_0).$$

Example 46 Continued (d). The objective is to estimate the mean of response y at \mathbf{X}_0

= **FLC**₅ = [1 7 4 180 5]^T, and then, obtain the 95% CI for $\mu_0 = \mu_5 = \beta_0 + 7\beta_1 + 4\beta_2$

$$+ 180\beta_3 + 5\beta_4. \rightarrow \hat{y}_0 = \mathbf{x}_0'\hat{\beta} = [1 \ 7 \ 4 \ 180 \ 5] \times \begin{bmatrix} -0.91221 \\ 0.16073 \\ 0.21978 \\ 0.011226 \\ 0.10197 \end{bmatrix} = 3.62248 \rightarrow \mathbf{x}_0' \mathbf{C} \mathbf{x}_0 =$$

$$0.1051521, se(\hat{y}_0) = [0.1051521 \times 0.6998]^{1/2} = 0.271267 \rightarrow HCIL = 0.55881 \rightarrow \mathbf{x}'_0 \hat{\boldsymbol{\beta}} - 0.55881 \leq \mu_0 \leq 3.62248 + 0.55881 \rightarrow 3.06367 \leq \mu_{y|\mathbf{x}'_0=[1 \ 7 \ 4 \ 180 \ 5]} \leq 4.18129.$$

THE PREDICTION INTERVAL FOR THE AVERAGE OF N FUTURE OBSERVATIONS AT \mathbf{x}_0

Let \bar{y}_0 be the average of $N \geq 1$ future observations at an \mathbf{x}_0 (that was not necessarily used

in design matrix \mathbf{X}), i.e., $\bar{y}_0 = \sum_{r=1}^N y_{r0} / N$. Since a point forecast of \bar{y}_0 is $\mathbf{x}'_0 \hat{\boldsymbol{\beta}}$, then the

forecast error $\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}$ is normally distributed with $E(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}) = E(\mathbf{x}'_0 \boldsymbol{\beta} + \bar{\epsilon}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}) = 0$

and $V(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}) = V(\bar{y}_0) + \mathbf{x}'_0 \text{COV}(\hat{\boldsymbol{\beta}}) \mathbf{x}_0 = (\frac{1}{N} + \mathbf{x}'_0 \mathbf{C} \mathbf{x}_0) \times \sigma_{\epsilon}^2$. Therefore, the statistic

$[(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}) - 0] / se(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}})$ has the Student t-distribution with $n - k - 1$ *df* and as a result,

the 95% PI for the future \bar{y}_0 is $\mathbf{x}'_0 \hat{\boldsymbol{\beta}} \pm t_{.025, n-k-1} \times se(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}})$, i.e.,

$$\mathbf{x}'_0 \hat{\boldsymbol{\beta}} - t_{.025, n-k-1} \times se(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}) \leq \bar{y}_0 \leq \mathbf{x}'_0 \hat{\boldsymbol{\beta}} + t_{.025, n-k-1} \times se(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}), \quad (70)$$

where $se(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}) = \sqrt{(\frac{1}{N} + \mathbf{x}'_0 \mathbf{C} \mathbf{x}_0) MS_{RES}}$,

and the common value of $N=1$ observation in the future. The above PI (Prediction Interval) has a 95% probability to actually contain a future \bar{y}_0 .

Example 46 Continued (e). Suppose we intend to make $N = 3$ future observations at $\mathbf{x}_0 = [1 \ 6 \ 8 \ 150 \ 4]'$. Note that this \mathbf{x}_0 is not a **FLC** from the design matrix \mathbf{X} . We wish to obtain an interval that has a Pr of 0.95 to contain the average of 3 future observations \bar{y}_0 made at \mathbf{x}_0 . From Eq. (70), the $se(\bar{y}_0 - \mathbf{x}'_0 \hat{\boldsymbol{\beta}}) = 0.5096 \rightarrow HPIL = 1.0498 \rightarrow PI = 3.902144 \pm 1.0498 \rightarrow 2.85238 \leq \bar{y}_0 \leq 4.95191$; this last prediction interval has a 95% Pr of containing a future \bar{y}_0 based on $N = 3$ observations. Note that a PI for \bar{y}_0 is always wider than the

corresponding CI for μ_0 . This is due to the fact that a PI contains 2 sources of error (from the model and from repeatability in the future), while a CI contains only model error. Minitab provides PIs only for a single future observation (i.e., $N=1$), which for this example is $2.147 \leq y_0 \leq 5.657$. This last interval has 95% Pr of containing the random variable y_0 .

Exercise 110. Obtain a 95% PI for a single future observation to be made at the 5×1 vector $\mathbf{X}_0 = \mathbf{FLC}_5 = [1 \quad 7 \quad 4 \quad 180 \quad 5]^T$ and compare the length of your PI against the corresponding CI obtained on page 194.

THE NET (OR PARTIAL) CONTRIBUTION OF ONE OR MORE REGRESSOR VARIABLE(S)

For the sake of illustration, suppose the following MLR model has been fitted to a data of size n .

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 \quad (71)$$

Then the explained variation in y due to the 5 regressors in model (71) is given by

$$SS_{\text{Reg}}(x_1, x_2, x_3, x_4, x_5) = \sum_{j=1}^5 \hat{\beta}_j S_{jy} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (72)$$

W/O loss of generality, we consider the net (or partial) contributions of the independent variables x_2 and x_5 to the total explained SS_{Model} in equation (72). In order to compute this net contribution, designated by $SS_{\text{Reg}}(x_2, x_5 | x_1, x_3, x_4)$, we must 1st regress the response y on the independent variables x_1, x_3, x_4 , using the same n data points, which will lead to a new regression model such as:

$$\hat{y} = b_0 + b_1 x_1 + b_3 x_3 + b_4 x_4 \quad (73)$$

The coefficients b_j ($j = 1, 3, 4$) in y -arc of Eq. (73) are, in general, different from $\hat{\beta}_j$ ($j = 1, 3, 4$) of equation (71) unless the matrix $\mathbf{A} = \mathbf{X}'\mathbf{X} = \mathbf{X}^T\mathbf{X}$ is diagonal, in which case the FLCs: $[1 \quad x_{i1}$

$x_{i2} \dots x_{ik}]$, $i = 1, 2, \dots, n$, form an orthogonal design. The net contribution of x_2 and x_5 is defined as

$$SS_{\text{Reg}}(x_2, x_5 | x_1, x_3, x_4) = SS_{\text{Reg}}(x_1, x_2, x_3, x_4, x_5) - SS_{\text{Reg}}(x_1, x_3, x_4) =$$

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 - \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{j=1}^5 \hat{\beta}_j S_{jY} - \sum_{j=1,3,4} b_j S_{jY} = \sum_{i=1}^n \hat{y}_i^2 - \sum_{i=1}^n \hat{y}_i^2 \quad (74)$$

The last relationship in (74) follows from the fact that $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n \hat{y}_i$. The partial F-

statistic for testing $H_0: \beta_2 = \beta_5 = 0$ is given by

$$F_0 = MS_{\text{Reg}}(x_2, x_5 | x_1, x_3, x_4) / MS_{\text{RES}} = \frac{SS_{\text{Reg}}(x_2, x_5 | x_1, x_3, x_4) / 2}{MS_{\text{RES}}}, \quad (75)$$

where MS_{RES} is computed under model (71).

Finally, it can be proven, using the Gram-Schmidt orthogonalization procedure, that for any MLREG model the coefficient of the last independent variable, namely $\hat{\beta}_k$, is the same for both the original non-orthogonal (or oblique) model and its corresponding orthogonal representation. Since it is quite arbitrary as to which of the x 's we would designate as x_k , this leads to the net (or partial) contribution of any single regressor x_r as

$$\delta_r^2 = \hat{\beta}_r^2 / C_{rr} = \sum_{j=1}^k \hat{\beta}_j S_{jY} - \sum_{j'r} b_j S_{jY}, \quad r = 1, 2, \dots, k. \quad (76)$$

Therefore, from equation (76) the statistic for testing $H_0: \beta_r = 0$ is $F_0 = \delta_r^2 / MS_{\text{RES}}$, which

has an F distribution with $v_1 = 1$ and $v_2 = n - k - 1$ *df*. This last F_0 statistic is generally referred to as the partial F, because it tests the significance of the net contribution of x_r to the overall regression. Note that C_{rr} is the $(r+1)$ th diagonal element of the matrix $\mathbf{C} = \mathbf{A}^{-1} = (\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}$, $r = 0, 1, 2, \dots, k$, where the element in the 1st row and column pertaining to $\hat{\beta}_0$, is the

element in the first row and column of the matrix $\mathbf{C} = \mathbf{A}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$, and $C_{k+1,k+1}$ pertains to x_k , $k > 0$. Further, $F_0 = (t_0)^2 = [\hat{\beta}_r / se(\hat{\beta}_r)]^2$.

Example 46 Continued (f).

To obtain the net contribution of x_1 to the overall regression SS of the model (61) given on page 182, we regress y on the variables x_2 , x_3 and x_4 , which results in the following model:

$$\hat{y} = -0.086743 + 0.22332 x_2 + 0.012285 x_3 + 0.095486 x_4 \rightarrow$$

$SS_{\text{Reg}}(x_2, x_3, x_4) = 35.247613$; thus, $\delta_1^2 = 39.37694 - 35.247613 = 4.129324$. On the other hand, we can also compute δ_1^2 from $\hat{\beta}_1^2 / C_{11} = (0.160726)^2 / 0.006256 = 4.12930$; the discrepancy in the 5th decimal place is strictly due to rounding error. Note that if we form the partial F statistic for testing $H_0: \beta_1 = 0$, we obtain $F_0 = \delta_1^2 / MS_{\text{RES}} = 4.1293 / 0.6998 = 5.901$, which is consistent with the Minitab output on my website because the value of $t_0^2(x_1) = (2.429)^2 = 5.901$ of Minitab is the same as the value of the partial F statistic $F_0(x_1) = 5.901$.

Exercise 111. (a) For the regression model of the Example 46, test the null hypothesis $H_0: \beta_1 = \beta_4 = 0$. (b) Conduct the partial F test for the variable x_2 .

SEQUENTIAL SUM OF SQUARES

Minitab provides Seq. SS each with 1 *df*. In order to obtain the Seq. SS's, 1st the total regression of y on x_1 must be obtained. For the data of Table 29, this leads to $\hat{y}_i = 2.41388 + 0.18892x_{i1}$, whose $SS(\text{Reg}) = \text{Seq. SS}(x_1) = 0.18892 \times 30.580 = 5.7772$, which agrees with Minitab's output on my website. Second, in order to obtain the Seq. SS due to x_2 , we regress y on x_1 and x_2 , resulting in $\hat{y}_i = 1.07655 + 0.168475x_{i1} + 0.21491x_{i2}$, whose $SS_{\text{Reg}}(x_1, x_2) = 0.168475 \times 30.58 + 0.21491 \times 133.86 = 33.91982$. Therefore, the $\text{Seq. SS}(x_2) = 33.91982 - 5.7772 = 28.1426$, which also agrees with the Minitab's output.

Exercise112. Verify that $\text{Seq. SS}(x_3) = 3.348$ and $\text{Seq. SS}(x_4) = 2.1094$.

In general, the sequential SS 's will not equal to δ_j^2 unless the design matrix is orthogonal (i.e., the matrix $\mathbf{A} = \mathbf{X}'\mathbf{X} = \mathbf{X}^T\mathbf{X}$ is diagonal). For the 4-regressor model of Example 13.16 on pp. 541-542 of Devore, $\delta_1^2 = 4.12930$, $\delta_2^2 = 29.13507$, $\delta_3^2 = 3.56598$, $\delta_4^2 = 2.10942$ so that $\sum_{j=1}^4 \delta_j^2 = 38.93977 < SS_{\text{Reg}}(x_1, x_2, x_3, x_4) = 39.37694$.

MODEL BUILDING PROBLEMS IN MLREG

Let p be the maximum possible number of regressor variables ($p \geq k$) that are candidates for inclusion in the MLR model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \dots + \beta_p x_p + \epsilon. \quad (77)$$

It is generally inefficient to include all the p independent variables (or regressors) in the model (77), rather the objective should be to identify a subset of size k from the p candidate regressors that satisfies the following 3 conditions:

(1) The value of the multiple determination coefficient $R_k^2 = SS_{\text{Reg}}(x_1, x_2, \dots, x_k) / SS(\text{Total})$ exceeds at least, say 75%, i.e., the explained variation in the response y by the k regressors (if at all possible) is at least 75%.

(2) Since SS_{Reg} always increases (albeit perhaps very slightly) as more regressor variables are added to the model, it is best to adjust the value of R_k^2 in the above condition (1) to account for this arbitrary increase by measuring the % explained variation in y (from x_1, x_2, \dots, x_k) using the adjusted R_k^2 defined as:

$$\text{Adj } R_k^2 = \bar{R}_k^2 = \frac{(n-1)R_k^2 - k}{n-1-k}. \quad (78)$$

The value of \bar{R}_k^2 in (78), unlike R_k^2 , may actually decrease as k increases toward p because in general, $n \gg k$. The set of k regressors out of the p independent variables should be selected in such a manner that has the maximum (or close to maximum) \bar{R}_k^2 among the ${}_p C_k = p!/[k!(p-k)!]$ possible sets.

(3) The value of the C_p statistic

$$C_p = [SS_k(\text{RES}) / MS_p(\text{RES})] - n + 2(k+1)$$

should not exceed k but should be minimal and less than k . This is due to the fact that C_p is an estimator of the total expected standardized Mean Square Error, $E \sum_{i=1}^n (\hat{y}_i - \mu_i)^2 / \sigma_{\epsilon}^2$, and hence, the regressors x_1, x_2, \dots, x_k ($k < p$) must be selected in such a manner that minimize C_k relative to k . Note that $E(C_p) \cong k$.

There are several model building procedures in regression that generally, but not always, lead to the same “most parsimonious” regression model having the same set of k predictors. These are Stepwise Regression, Forward Selection, and Backward Elimination.

The significance level (α_{in}) used to judge the contribution of the i^{th} regressor to the overall regression varies from software to software. SAS sometimes uses $\alpha_{in} = 0.50$ and sometimes uses $\alpha_{in} = 0.15$. Minitab recommends $\alpha_{in} = \alpha_{out} = 0.15$, and personally I believe α should not exceed the range 0.20-0.25 significance level. The four most common model building procedures are FORWARD Selection, Stepwise Regression, Backward Elimination, and Best Subsets in Minitab (or MAXR in SAS) procedure. FORWARD Selection starts with the best regressor, i.e., the one with the largest R_{Model}^2 , then finds the next best one to add to what

exists, the next best, etc. Stepwise Regression is similar to FORWARD except that there is an extra step in which all variables in the new model are checked to see if they remain significant at the α_{in} level. Backward Elimination starts with all p regressors in the model, then drops the least significant one, then the next, and the next, etc. Best Subsets (or SAS's MAXR) procedure is a rather long and tedious procedure, but basically finds the best (i.e., with the largest R^2 and Smallest C_p statistic) one-variable regression model, then the best 2-variable model, then the best 3-variable model, and so on through the best k -variable model. The user must decide, from the output, which model is the best and most parsimonious. Minitab has only two of the above four procedures (Stepwise Regression and Best Subset Regression). I recommend the following 5 criteria for selecting one out of the k regression models of Minitab's Best Subsets Procedure.

- (i)** The $F_0(\text{Model})$ should be nearly largest (or the P -value for testing the model significance should be nearly the smallest) amongst all possible regression models.
- (ii)** Both R_k^2 and specially the value of \bar{R}_k^2 must be the largest or nearly so.
- (iii)** The value of C_p statistic should be less than k and its value relative to k should be minimum. Regression models for which $C_p \gg k$ exhibit too much bias (leading to poor predictions).
- (iv)** The selected model with k regressors should have all coefficients significant at levels, say, 25% (if possible 15%) or less. The user must be able to assess both statistical and practical significance of a regressor (this is why $\alpha = 0.05$ is too small).
- (v)** Finally, the experimenter must leave sufficient df ($\nu_2 \geq 6$) for $MS_{RES} = MS(\text{Error})$ so that the partial F tests will have sufficient power to reject $H_0: \beta_j = 0$ at the α_{in} LOS. Note that the sampling distribution of F_0 is not stable when $\nu_2 < 6$. You will be using Minitab's Best Subsets Procedure near the end of STAT 3611 Lab.