INSY 7310 Reference: Chapter 10 of Montgomery (7e) Maghsoodloo

Fitting Regression Models

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) 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 page 601 of Jay L. Devore's 6th edition, 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 wrinkle resistance). The data matrix, **X** having 30 factor level combinations **FLCs**, is provided in Table 1 atop the next page. The MLR model for this example is

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \in_i , \qquad (1)$$

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), and $\mathbf{x_0}$ is a 30×1 vector whose value is always equal to 1 for all rows i = 1, 2, …, n = 30. For example, the value of y_5 is modeled as $y_5 = \beta_0 + 7\beta_1 + 4\beta_2 + 180\beta_3 + 5\beta_4 + \in_5$, where $\in_i$'s (i = 1, 2, 3, …, n = 30) are assumed NID(0, $\sigma^2_\in$). Henceforth, for convenience we will use the symbol σ^2 for the error variance $\sigma^2_\in$; further, the index i runs over different **FLCs** from 1 to n, while the index j over the regressors. The reader must be cognizant of the fact that in classical regression theory, it is assumed that the regressor variables x_j (j = 1, 2, …, 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, …, k. Only the classical regression is covered in this course, i.e., only y_i's and $\in_i$'s in model (1) are random variables (rvs), while β_j's and x_j's are not rvs. Further, regression analysis can analyze data from both planned and unplanned experiments, while ANOVA of factorials pertains to planned experiments.

Exercise 1. Show that in the case of classical regression, $E(y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} +$

Table 1 (The Data from Textile Research J., 1984:157-165; k = 4 *df* = 4 *degrees of freedom*)

X is a 30×5 Design Matrix Y is a 30×1 Vector Response Each FLC is a 5×1 Vector

FLC	x_0	x_1	x_2	x_3	x_4	Response y_i	FLC	x_0	x_1	x_2	x_3	x_4	y_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

$\beta_3 x_{i3} + \ldots + \beta_k x_{ik}$, and $V(y_i) = \sigma^2_{\in} = \sigma^2$ for all i = 1 to n.

Our objective, just like in SLREG, is to estimate the k + 1 (= 5) parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, where for our present example k = 4 independent variables (i.e., the regression model has 4 degrees of freedom = *df*), in such a manner that the least squares function (LSF)

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

is minimized. In order to minimize $L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = L(\beta_j)$ wrt (with respect to) the parameters β_j (j = 0, 1, 2, 3, ..., k = 4), the partial derivatives of the LSF, $\partial L/\partial\beta_j$, must be required to be zero for all j. Further, $\partial^2 L / \partial\beta_j^2$ must exceed zero for all j. The 1st partial derivatives when set to zero generally lead 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). The k + 1 (= 5) partial derivatives are provided below (note that the index i always runs over rows of the design matrix **X** and j over the regressors):

$$\partial L / \partial\beta_0 = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})(-1) \tag{3a}$$

$$\partial L / \partial\beta_1 = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})(-x_{i1}) \tag{3b}$$

$$\partial L / \partial\beta_2 = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})(-x_{i2}) \tag{3c}$$

$$\partial L / \partial\beta_3 = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})(-x_{i3}) \tag{3d}$$

$$\partial L / \partial\beta_4 = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})(-x_{i4}) \tag{3e}$$

Note that Eqs. (3) show that $\partial^2 L / \partial\beta_j^2 > 0$ for all j and hence the 5×1 solution vector

$\hat{\beta} = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2 \quad \hat{\beta}_3 \quad \hat{\beta}_4]^T$ minimizes $L(\beta_j) = \sum_{i=1}^{n} \in_i^2$. The RHS (Right Hand Side) of Eq. (3a) when set equal to zero leads to the 1st LS (least squares) normal equation as

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum x_2 + \hat{\beta}_3 \sum x_3 + \hat{\beta}_4 \sum_{i=1}^{n} x_{i4} = \sum_{i=1}^{30} y_i$$

Or: $$\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_3 \bar{x}_3 + \hat{\beta}_4 \bar{x}_4 = \bar{y} \tag{4a}$$

The RHS's of Eqs. (3 b, c, d, & e) when set equal to zero give rise to the other 4 normal equations, respectively.

$$\hat{\beta}_0 \sum x_1 + \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2 + \hat{\beta}_3 \sum x_1 x_3 + \hat{\beta}_4 \sum x_1 x_4 = \sum (x_1 y) \qquad (4b)$$

$$\hat{\beta}_0 \sum x_2 + \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2 + \hat{\beta}_3 \sum x_2 x_3 + \hat{\beta}_4 \sum x_2 x_4 = \sum x_2 y \qquad (4c)$$

$$\hat{\beta}_0 \sum x_3 + \hat{\beta}_1 \sum x_1 x_3 + \hat{\beta}_2 \sum x_2 x_3 + \hat{\beta}_3 \sum x_3^2 + \hat{\beta}_4 \sum x_3 x_4 = \sum x_3 y \qquad (4d)$$

$$\hat{\beta}_0 \sum x_4 + \hat{\beta}_1 \sum x_1 x_4 + \hat{\beta}_2 \sum x_2 x_4 + \hat{\beta}_3 \sum x_3 x_4 + \hat{\beta}_4 \sum x_4^2 = \sum x_4 y \qquad (4e)$$

In Eqs. (4) 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 **FLCs** for this Example. Using the data in Table 1, we obtain the summary statistics:

$$n = 30,\ \sum x_1 = 182,\ \sum x_2 = 204,\ \sum x_3 = 4280,\ \sum x_4 = 118,,\ \bar{x}_1 = 6.06667,\ \bar{x}_2 = 6.80$$

$$\bar{x}_3 = 142.66667,\ \bar{x}_4 = 3.93333,\ \sum x_1^2 = 1266,\ \sum x_1 x_2 = 1253,\ \sum x_1 x_3 = 26160$$

$$\sum x_1 x_4 = 706,\ \sum x_2^2 = 1998,\ \sum x_2 x_3 = 29180,\ \sum x_2 x_4 = 766,\ \sum x_3^2 = 639200,$$

$$\sum x_3 x_4 = 16720,\ \sum x_4^2 = 670,\ \sum y_i = 106.80,\ \sum x_1 y = 678.50,\ \sum x_2 y = 860.1$$

$$\sum x_3 y = 15594.00,\ \sum_{i=1}^{30} x_{i4} y_i = 430.20,\ \text{USS} = \sum_{i=1}^{30} y_i^2 = 437.080,\ \text{and}\ \bar{y} = 3.56.$$

Substituting the 20 pertinent of the above 26 statistics into Eqs. (4) yields the following set of 5 normal equations with 5 unknowns for the data of Table 1.

$$\begin{cases} 30\hat{\beta}_0 + 182\hat{\beta}_1 + 204\hat{\beta}_2 + 4280\hat{\beta}_3 + 118\hat{\beta}_4 = 106.8 \\ 182\hat{\beta}_0 + 1266\hat{\beta}_1 + 1253\hat{\beta}_2 + 26160\hat{\beta}_3 + 706\hat{\beta}_4 = 678.5 \\ 204\hat{\beta}_0 + 1253\hat{\beta}_1 + 1998\hat{\beta}_2 + 29180\hat{\beta}_3 + 766\hat{\beta}_4 = 860.1 \\ 4280\hat{\beta}_0 + 26160\hat{\beta}_1 + 29180\hat{\beta}_2 + 639200\hat{\beta}_3 + 16720\hat{\beta}_4 = 15594 \\ 118\hat{\beta}_0 + 706\hat{\beta}_1 + 766\hat{\beta}_2 + 16720\hat{\beta}_3 + 670\hat{\beta}_4 = 430.2 \end{cases} \qquad (5)$$

One way to solve the above system of 5 equations with 5 unknowns is to use Cramer's Rule (see my website for further details). Accordingly, we define the 6 matrices **A** and $\mathbf{A}_j$ (j = 0, 1, 2, 3, 4) as follows:

$$\mathbf{A} = (\mathbf{X'X}) = (\mathbf{X^T X}) = \begin{bmatrix} 30 & 182 & 204 & 4280 & 118 \\ 182 & 1266 & 1253 & 26160 & 706 \\ 204 & 1253 & 1998 & 29180 & 766 \\ 4280 & 26160 & 29180 & 639200 & 16720 \\ 118 & 706 & 766 & 16720 & 670 \end{bmatrix},$$ and the matrix $\mathbf{A_j}$

is identical to the above matrix **A = (X′X)** except for its j^{th} column, which is $\mathbf{COL_j} = \begin{bmatrix} 106.8 \\ 678.5 \\ 860.1 \\ 15594 \\ 430.2 \end{bmatrix} =$

$\begin{bmatrix} \sum y \\ \sum (x_1 y) \\ \sum (x_2 y) \\ \sum (x_3 y) \\ \sum (x_4 y) \end{bmatrix}$ for j = 0, 1, 2, 3, 4, and the n×(k+1) = 30×5 design matrix **X** is given in Table 1.

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 1 is given by $\mathbf{A_2} = \begin{bmatrix} n & \sum x_1 & \sum y & \sum x_3 & \sum x_4 \\ \sum x_1 & \sum x_1^2 & \sum x_1 y & \sum x_1 x_3 & \sum x_1 x_4 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2 y & \sum x_2 x_3 & \sum x_2 x_4 \\ \sum x_3 & \sum x_1 x_3 & \sum x_3 y & \sum x_3^2 & \sum x_3 x_4 \\ \sum x_4 & \sum x_1 x_4 & \sum x_4 y & \sum x_3 x_4 & \sum x_4^2 \end{bmatrix}$

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

The (information) matrix $\mathbf{X^T X} = \mathbf{X'X}$ is always symmetrical while the matrices $\mathbf{A_j}$ **(j = 0, 1, 2,…, k = 4)** are not in general symmetric. Excel (or Matlab) computations give det(**A**) = $|\mathbf{A}|$ = 17.014701×10^{12}, det($\mathbf{A_0}$) = $|\mathbf{A_0}|$ = −15.521016533×10^{12}, det($\mathbf{A_1}$) = 2.734711393×10^{12}, det($\mathbf{A_2}$) = 3.73954437984×10^{12}, det($\mathbf{A_3}$) = 0.1909996304×10^{12} , and det($\mathbf{A_4}$) = 1.73506460544×10^{12}.

Hence, $\hat{\beta}_0 = \det(\mathbf{A_0}) / \det(\mathbf{A}) = -0.9122121$, $\hat{\beta}_1 = \det(\mathbf{A_1}) / \det(\mathbf{A}) = 0.1607264$, $\hat{\beta}_2 = 0.2197831$, $\hat{\beta}_3 = 0.0112256$, and $\hat{\beta}_4 = |\mathbf{A_4}| / |\mathbf{A}| = 0.1019744$. The above 5 estimates of β_j's give rise to the following fitted MLR (Multiple LINEAR Regression) model:

$$\hat{y}_i = -0.9122121 x_0 + 0.1607264 x_{i1} + 0.2197831 x_{i2} + 0.0112256 x_{i3} + 0.1019744 x_{i4} \quad (6)$$

Note that, at 1st glance, the regressor variable x_2 in Eq. (6) seems to have the largest impact on the response variable Y because its coefficient 0.2197831 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 y will be determined by $t_{n-k-1} = \hat{\beta}_j / se(\hat{\beta}_j)$, $j = 1, 2, 3, 4 = k$.

We now develop a general matrix algebra approach for obtaining the coefficient estimates in a MLR model. The symbols ′ and ^{T} will denote matrix transpose and the bolded font capital letter is used to represent a matrix. The necessary matrices, including the design matrix **X** that is composed of n planned or unplanned **FLCs**, are defined below:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2k} \\ . & . & . & . & . & . & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & x_{n1} & x_{n2} & & & & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}, \text{ and } \boldsymbol{\in} = \begin{bmatrix} \in_1 \\ \in_2 \\ . \\ . \\ . \\ \in_n \end{bmatrix}. \quad (7a)$$

For the data of Table 1, the dimension of the vector **Y** is 30×1, that of matrix **X** is 30×5, that of vector $\boldsymbol{\beta}$ is 5×1, and $\boldsymbol{\in}$ is a 30×1 vector; clearly, (**X′X**) = **A**. First, we rewrite the MLR (1), which is valid only for the i[th] observation y_i, for all the n **FLCs** in matrix form using Eq. (7a).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\in} \quad (7b)$$

In order to obtain the least-squares estimate of the vector $\boldsymbol{\beta}$, we first use the fact that the LSF in matrix form is given by

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n=30} \in_i^2 = \in' \in = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}$$

$$= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{Y}) + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{Y}) + \boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}. \qquad (8)$$

Second, we take the partial derivative of $L(\boldsymbol{\beta})$ wrt the vector $\boldsymbol{\beta}$ and will require that the $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ equals the **zero vector** in order to minimize the LSF with respect to all the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ (k = 4 for our example). In the following matrix development bear in mind that $\mathbf{Y}'\mathbf{Y} = \mathbf{Y}^T\mathbf{Y}$ (= $\sum_{i=1}^{n=30} y_i^2$ = the USS) is independent of the column vector $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4]'$, where USS stands for Uncorrected Sum of Squares.

$$\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \begin{bmatrix} \partial L/\partial\beta_0 \\ \partial L/\partial\beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \partial L/\partial\beta_k \end{bmatrix} = 0 - 2(\mathbf{X}'\mathbf{Y}) + 2(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} \xrightarrow{\text{set equal to}} \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \qquad (9a)$$

Further, Eq. (9a) shows that $\partial L^2(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^2 = \mathbf{diag}(\mathbf{X}^T\mathbf{X}) = 2[n \quad \sum_{i=1}^{n} x_{i1}^2 \quad \sum_{i=1}^{n} x_{i2}^2 \quad \ldots \quad \sum_{i=1}^{n} x_{ik}^2]^T$, which exceed the zero vector and hence the solution $\hat{\boldsymbol{\beta}}$ minimizes $L(\boldsymbol{\beta})$. Eq. (8) now yields the heterogeneous system $\rightarrow \mathbf{0} = (\mathbf{X}'\mathbf{Y}) - (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Y}) - \mathbf{A}\hat{\boldsymbol{\beta}}$ of 5 (= k+1) equations with 5 unknowns whose solutions can now easily be obtained by transposing $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}$ to the LHS and multiplying both sides by the inverse (or reciprocal) of the symmetric (k+1)×(k+1) = 5×5 matrix $\mathbf{A} = \mathbf{X}'\mathbf{X}$. That is, $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Y}) \rightarrow (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$

Or: $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}(\mathbf{X'Y}) = \mathbf{A^{-1}}(\mathbf{X'Y}) = \mathbf{C}(\mathbf{X'Y})$, (9b)

where the (k+1)×(k+1) = 5×5 matrix $\mathbf{C} = (\mathbf{X'X})^{-1}$ = the inverse or reciprocal of $\mathbf{A} = \mathbf{A^{-1}}$. Further, like ($\mathbf{X'X}$), the matrix $\mathbf{C}$ must also be square and symmetrical. Applying Eq. (9b) to the data of Table 1, we obtain

$$\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X'Y}) = \mathbf{(X'X)^{-1}(X'Y)} = \mathbf{A^{-1}(X^{T}Y)}$$

$$= \begin{bmatrix} 1.09527 & -0.03213 & -0.01123 & -0.004844 & -0.02533 \\ & 0.006256 & -0.000138 & -4.1223\times10^{-5} & 0.000253 \\ & & 0.001658 & -2.3269\times10^{-6} & 0.000285 \\ & & & 3.53376\times10^{-5} & 1.73\times10^{-5} \\ & & & & 0.00493 \end{bmatrix} \times \begin{bmatrix} 106.80 \\ 678.5 \\ 860.1 \\ 15594.0 \\ 430.20 \end{bmatrix}$$

$$= \begin{bmatrix} -0.91221 \\ 0.16073 \\ 0.21978 \\ 0.01123 \\ 0.10197 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \hat{\boldsymbol{\beta}},$$ which is identical to the previous LS estimates of Eq. (6) on page 6 of my notes using Cramer's Rule.

Exercise 2. Use Excel to verify that $\mathbf{X'X} = \mathbf{X^{T}X} = \mathbf{A}$, which is given atop page 5 for this Example, and on the same spreadsheet verify the elements of the above 5×1 vector estimator $\hat{\boldsymbol{\beta}}$ in the 2 different methods outlined so far.

Residuals in MLREG

Recall that by definition a model residual is defined as $e_i = y_i - \hat{y}_i$, where for the data of

Table 1, $\hat{y}_i$ is given by the model (6) on page 6 of my notes. Eq. (7b) shows that $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ (because the best predictor of the vector $\in$ is the 0 vector), and therefore, the fitted vector $\hat{\mathbf{Y}}$ for all the n = 30 observations is given by $\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ . \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X(C)X'Y} = \mathbf{X(X'X)^{-1}(X'Y)}$

$= \mathbf{[X(X'X)^{-1}X']Y = (XCX^T)Y = HY.}$

The n×n (30×30 for Table 1) matrix $\mathbf{H} = \mathbf{X(X'X)^{-1}X'} = \mathbf{XCX' = XCX^T}$ is called the Hat matrix because it projects the vector **Y** onto the vector $\hat{\mathbf{Y}}$ (or Y-Hat) through the matrix equation $\hat{\mathbf{Y}} = \mathbf{HY}$. The n×1 residual vector, therefore, is given by $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, and as a result

$$SS_{Residuals} = \sum_{i=1}^{n} e_i^2 = \mathbf{e'e} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y'Y} - 2\hat{\mathbf{Y}}'\mathbf{Y} + \hat{\mathbf{Y}}'\hat{\mathbf{Y}}. \quad (10)$$

However, $\hat{\mathbf{Y}}^{\mathbf{T}}\mathbf{Y} = (\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{Y} = \hat{\boldsymbol{\beta}}'(\mathbf{X'Y}) = \hat{\boldsymbol{\beta}}'(\mathbf{X'X})\hat{\boldsymbol{\beta}} = (\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$, where we have made use of Eq. (9b) which shows that $(\mathbf{X'X})\hat{\boldsymbol{\beta}} = (\mathbf{X'Y})$. 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 the Eq. (10) results in $SS_{Residuals}$ =

$\mathbf{Y'Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} \hat{y}_i^2 = (\sum_{i=1}^{n} y_i^2 - CF) - (\sum_{i=1}^{n} \hat{y}_i^2 - CF)$, which shows that $SS_{Residuals}$ =

SS(Error) = SS(Unexplained) = SS(Total) – SS(Regression) = SS_T – SS(Model), where

SS_T = CSS = $\sum_{i=1}^{n} y_i^2 - CF$, and SS (Model) = SS(Regression) = SS(Explained)= $\sum_{i=1}^{n} \hat{y}_i^2 - CF$. Note that this last equation is similar to ANOVA where SS(Total) = SS(Model) + SS(Error) = SS(Explained) + SS(Unexplained) for all statistical models.

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

Exercise 3. Show that the three n×n matrices, **H** , $\mathbf{I_n}$, and $\mathbf{I_n - H}$ are all symmetrical and idempotent, where $\mathbf{I_n}$ is an n×n identity matrix. Further, show that except for the identity matrix $\mathbf{I_n}$, an idempotent matrix cannot have an inverse.

Exercise 4. Use Eq. (4a) to show that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4}$ is also given by $\hat{y}_i = \overline{y} + \hat{\beta}_1 (x_{i1} - \overline{x}_1) + \hat{\beta}_2 (x_{i2} - \overline{x}_2) + \hat{\beta}_3 (x_{i3} - \overline{x}_3) + \hat{\beta}_4 (x_{i4} - \overline{x}_4)$, which represents the corrected form of the fitted model. As a result, prove that $\sum_{i=1}^{n} e_i = \sum (y_i - \hat{y}_i) \equiv 0$, which implies that $\sum_{i=1}^{n} \hat{y}_i \equiv \sum_{i=1}^{n} y_i$. Therefore, the CF for SS(Regression), which is $(\sum_{i=1}^{n} \hat{y}_i)^2 / n$, is also equal to $(\sum_{i=1}^{n} y_i)^2 / n$. Then, show that SS(Regression) = SS(Model) = SS(Explained) = $\sum_{i=1}^{n} \hat{y}_i^2$ – CF = $\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$.

We now develop matrix formulas for SS_T = SS(Total) = S_{yy}, SS(Model) = SS(Reg), and SS(Residuals) = SS(Unexplained). This will be helpful if one wishes to use Matlab in order to obtain different SS's in regression. To this end, let $\mathbf{I_n}$ be the n×n identity matrix and the vector $\mathbf{1}$ be an n×1 column vector every element of which is equal to 1, i.e., $\mathbf{1}$ =[1 1 … 1]′, where $\mathbf{1'}$ is an 1×n row vector, and the matrix **J** = $(\mathbf{1} \times \mathbf{1'})/n$ is an n×n matrix all of whose elements are equal to 1/n. From these matrix definitions we deduce that

$$CF = (\mathbf{Y'1})^2/n = (\mathbf{Y'1})(\mathbf{1'Y})/n = \mathbf{Y'}[(\mathbf{11'})/n\,]\,\mathbf{Y} = \mathbf{Y'JY}. \qquad (11a)$$

$$SS_T = CSS = \mathbf{Y'Y} - CF = \mathbf{Y'Y} - \mathbf{Y'JY} = \mathbf{Y'}(\mathbf{I_n} - \mathbf{J}\,)\mathbf{Y} \qquad (11b)$$

$$SS(Reg) = SS_{Model} = \hat{\mathbf{Y}}'\,\hat{\mathbf{Y}} - \mathbf{Y'JY} = (\mathbf{HY})'(\mathbf{HY}) - \mathbf{Y'JY} = \mathbf{Y'(H'H)Y} - \mathbf{Y'JY}$$

$$= \mathbf{Y'HY} - \mathbf{Y'JY} = \mathbf{Y'(H - J)Y}\,, \qquad (11c)$$

where we have made use of the fact that the n×n hat matrix **H** is idempotent. Thus,

SS(RES) = SS(Total) − SS(Reg) = $\mathbf{Y'(I_n - H)Y}$.

(11d)

Exercise 5. Prove that SS(Regression) = $\mathbf{Y'(H - J)Y} = \sum_{j=1}^{k} \hat{\beta}_j S_{jy}$, where S_{jy} =

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

Hint: SS(Reg) = $\mathbf{Y'HY}$ − CF = $\mathbf{Y'[X(X'X)^{-1}X']Y}$ − CF = $\hat{\boldsymbol{\beta}}'(\mathbf{X'Y})$ − CF. Further, show that for our example given in Table 1, S_{1y} = 30.58, S_{2y} = 133.86, S_{3y} = 357.20, and S_{4y} = 10.12.

Example 1. We now use the Eqs. (11), which were developed above, to obtain SS_T = SS(Total) = USS – CF = 437.08 –(106.80^2)/30 = 56.8720 (with 30 − 1 = 29 *df*). The Minitab output posted on my website verifies this value of this Total SS. Next we compute SS_{Reg} =

$\mathbf{Y'(H - J)Y} = \sum_{j=1}^{4} \hat{\beta}_j S_{jy}$ = (0.1607264)(30.58) +0.2197831 (133.86) + 0.0112256 (357.20) + 0.1019774 ×(10.12) = 39.37694 = SS(Model) with 4 *df* because there are 4 regressors (or independent variables), where $S_{1y} = \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i = \sum_{i=1}^{n} x_{i1}y_i - \bar{x}_1 \sum_{i=}^{30} y_i$ = 678.50 – 6.066667(106.80) = 30.580, etc. Therefore, SS_{RES} = SS(Total) – SS(Model) = 56.8720 − 39.3769 = 17.4951; these SS's are also consistent with the Minitab output. You should review the ANOVA Table provided by Minitab, which does not provide more than 3 decimals for *P-values.* The exact *P-value* for F_0(Model) = 14.0672 is *p* = 0.00000385. The other statistics provided by Minitab will be derived and discussed later in these notes.

In order to develop CIs (confidence intervals) and conduct tests of hypotheses on the vector parameter $\boldsymbol{\beta}$, we need to show that if **B** is any p×n constant matrix and **Y** is an n×1 random vector response, then the COV(**BY**) = **B**COV(**Y**)**B′**, and E(**Y**) = $\boldsymbol{\mu}$ = [μ_1 μ_2 … μ_n]′ is an n×1 parameter vector of the n population means, where n stands for the number of **FLCs** in

the design matrix **X**. **Proof.** By definition, COV(**BY**) = E[(**BY** − **Bμ**)(**BY** − **Bμ**)′] = E[(**BY** − **Bμ**)(**Y′B′** − **μ′B′**)] = E[**B**(**Y** − **μ**)(**Y′** − **μ′**)**B′**)] = **B**E[(**Y** − **μ**)(**Y′** − **μ′**)]**B′**= **B**COV(**Y**)**B′** .

Exercise 6. Use the above covariance property, and the fact that under a regression model similar to Eq. (1) the COV(**Y**) = $\mathbf{I_n}\sigma_{\epsilon}^2$, to show that (a) COV($\hat{\mathbf{Y}}$) = $\mathbf{H}\sigma_{\epsilon}^2$, (b) COV(**e**) = ($\mathbf{I_n}$–**H**) σ_{ϵ}^2, where **e** is the n×1 residual vector, and (c) COV($\hat{\boldsymbol{\beta}}$) = $(\mathbf{X'X})^{-1}\sigma_{\epsilon}^2$ = $\mathbf{C}\sigma^2$, where the diagonal elements of the symmetric (k+1)×(k+1) matrix COV($\hat{\boldsymbol{\beta}}$) give the V($\hat{\beta}_j$), j = 0, 1, 2,…, k and its off-diagonal elements give the COV($\hat{\beta}_j$, $\hat{\beta}_r$) for r = 0, 1, 2,…, k ≠ j.

From part (a) of the above exercise *se*($\hat{y}_i$) = $\sqrt{h_{ii} \times \mathrm{MS(RES)}}$, and part (b) shows that the V(e_i) =$(1 - h_{ii})\sigma_{\epsilon}^2$, resulting in the Studentized residuals given by

$$r_i = e_i / \sqrt{(1-h_{ii})\,\mathrm{MS_{RES}}} \quad , \quad i = 1, 2, \ldots, n. \qquad (12)$$

Any **FLC** with large h_{ii} and consequently with large r_i (say with absolute value, $|r_i|$, greater than 2) is highly influential on the least squares fit. Minitab provides an option that lists the 3 vectors $\hat{\mathbf{Y}}$, , **e**, and Studentized Residuals **r**. Note that Minitab uses the designation Standardized Residuals instead of Studentized residuals, which is not precise, and Montgomery uses the designation r_i for the i^{th} Studentized residual. Further, most authors (and SAS) refer to r_i in Eq. (12) also as the Student(ized) residuals. For our Example 2, the largest r_i in absolute value is r_9 = 2.04, which implies that the **FLC** number 9, [1 4 7 140 3]^T, has the highest influence on the regression coefficients, while r_{24} = 0.01 implies that the **FLC** [1 7 1 120 7]^T has almost no impact on $\hat{\beta}_j$ (j = 0, 1, 2, 3, 4). As a matter of fact, I removed **FLC** 24 from the design matrix **X** of Table 1 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 full regression model $\hat{y}_i$ given in Eq. (6). However, if the **FLC$_9$** = [1 4 7 140 3]^T is removed from the design matrix **X**, the resulting regression model $\hat{y}_{(9)}$ =

$-1.14930x_0 + 0.18387x_1 + 0.21915x_2 + 0.01127x_3 + 0.11102x_4$, is clearly different from the model (6) on page 6 of my notes.

Definition. The trace of a (square) matrix is simply the sum of its diagonal elements. For example, the Trace (I_n) = n, the $\text{Tr}\begin{bmatrix} 5 & 9 \\ 4 & -5 \end{bmatrix} = 0$, the Tr (12) = 12, the $\text{Tr}\begin{bmatrix} 7 & 9 & 12 \\ 15 & -5 & -8 \\ 13 & 2 & 9 \end{bmatrix} =$ 11, and the trace of the magic(4) matrix $\begin{bmatrix} 16 & 3 & 2 & 13 \\ 5 & 10 & 11 & 8 \\ 9 & 6 & 7 & 12 \\ 4 & 15 & 14 & 1 \end{bmatrix}$ is 34. Note that the det(magic) for an even integer n > 2 seems to be zero and the entries in the magic(n) matrix range from 1, 2, 3, 4, … to n^2. Further, it can be shown that the trace of magic(n) = Tr(magic(n)) = $(n^3+n)/2$.

Bonus HW1. (a) Let **A** and **B** be 2 compatible matrices (not necessarily square) such that **AB** and **BA** are square matrices; show that Tr(**AB**) = Tr(**BA**); further, Tr(**A** + **B**)= Tr(**A**) + Tr(**B**), Tr(b**A**) = bTr(**A**) for any scalar constant b, and Tr ($\mathbf{B^{-1}AB}$) = Tr(**A**). (b) Use these properties of the trace of a matrix to prove that $E(SS_{RES}) = (n - k - 1)\sigma_{\in}^2$ and hence an unbiased estimator of $\sigma_{\in}^2$ is $\sigma_{\in}^2 = MS_{RES} = (SS_{Residuals})/(n - k - 1)$ for all classical regression models.

CONFIDENCE INTERVALS FOR β_j (j = 0, 1, 2, 3, …, k)

The main assumption in MLR is that y_i's (i = 1, 2, …, n) are NID(μ_i, $\sigma_{\in}^2$), where $\mu_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$. Because $\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}(X'Y) = (CX')Y}$, every $\hat{\beta}_j$ is a linear combination of y_i's (i = 1, 2, …, n) resulting in the normality of each $\hat{\beta}_j$ with $E(\hat{\beta}_j) = \beta_j$. Further, from **Exercise 6** above the $se(\hat{\beta}_j) = (C_{jj}MS_{Residuals})^{1/2}$, where C_{jj} is the (j+1)th diagonal element of the matrix $\mathbf{C} = \mathbf{(X'X)^{-1}}$, j = 0, 1, 2, ..., k. Therefore, a 95% CI for the parameter β_j is given by

$$\hat{\beta}_j \pm t_{0.025;n-k-1} \times se(\hat{\beta}_j). \qquad (13)$$

If the interval in Eq. (13) excludes 0, then the null hypothesis H_0: $\beta_j = 0$ must be rejected at the nominal 5% LOS (Level of Significance) for any j = 0, 1, 2, …, k. This, in turn, will imply that the regressor variable x_j, j =1, 2, …, k, has a statistically significant impact on the (mean of) 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 the (Gosset) Student's t-distribution with $n - 1 - k = n - p$ *df*, where p = k+1.

Example 1 Continued. (c) We now use Eq. (13) in order to obtain the 95% CI for β_1. From $COV(\hat{\boldsymbol{\beta}}) = \mathbf{(X'X)^{-1}}\sigma_{\epsilon}^2 = \mathbf{C}\sigma_{\epsilon}^2$, we deduce that the $se(\hat{\beta}_1) = (C_{11}MS_{RES})^{1/2} = (0.006256 \times 0.6998)^{1/2} = 0.066166 \rightarrow$ HCIL = $t_{0.025;25} \times se(\hat{\beta}_1) = 2.05954 \times 0.06617 = 0.136272 \rightarrow$ the CI for β_1 is : $0.16073 \pm 0.136272 \rightarrow 0.024455 \leq \beta_1 \leq 0.29700$; since this 95% CI excludes zero, the null hypothesis $H_0 : \beta_1 = 0$ must be rejected at the 5% LOS. Note that this result is consistent with that of the Minitab's output because the *P-value* that Minitab lists for the regressor x_1 is $\hat{\alpha} = p = 0.023 < \alpha = 0.05$. Thus, the effect (or impact) of x_1 on Y is 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 $t_0 = (\hat{\beta}_1 - 0)/se(\hat{\beta}_1) = 0.16073 / 0.066166 = 2.4291$ and compare against the threshold value of $t_{0.025;25} = 2.05954$, which agrees with Minitab's output to 2 decimals.

Exercise 7. Obtain the 95% CI's for β_2, β_3 and β_4 of Example 2 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 Minitab output.

CONFIDENCE INTERVAL FOR THE MEAN RESPONSE $E(Y \mid \mathbf{x_0}) = \mu_0$

Let $\mathbf{x_0} = [1 \quad x_{01} \quad x_{02} \ \ldots \ x_{0k}]' = [1 \quad x_{01} \quad x_{02} \ \ldots \ x_{0k}]^T$ be a specified **FLC** (within the range of the **X** factor space) so that $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \ldots + \hat{\beta}_k x_{0k} = \hat{\boldsymbol{\beta}}' \mathbf{x_0} = \mathbf{x_0}' \hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\mu_0 = \mathbf{x_0}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \ldots + \beta_k x_{0k} = \beta' x_0$. In order to obtain the $se(\hat{y}_0)$, we must 1st compute the $V(\hat{y}_0)$ as shown below.

$$V(\hat{y}_0) = V(\mathbf{x_0'}\hat{\boldsymbol{\beta}}) = E[(\mathbf{x_0'}\hat{\boldsymbol{\beta}} - \mathbf{x_0'}\boldsymbol{\beta})(\mathbf{x_0'}\hat{\boldsymbol{\beta}} - \mathbf{x_0'}\boldsymbol{\beta})^T] = E[\mathbf{x_0'}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{x_0}]$$

$$= \mathbf{x}_0' E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']\mathbf{x}_0 = \mathbf{x}_0' COV(\hat{\boldsymbol{\beta}})\mathbf{x}_0 = \mathbf{x}_0'(\mathbf{C}\sigma_{\epsilon}^2)\mathbf{x}_0 = (\mathbf{x}_0'\mathbf{C}\mathbf{x}_0)\sigma_{\epsilon}^2$$

(14a)

Eq. (14a) clearly shows that the

$$se(\hat{y}_0) = [(\mathbf{x}_0'\mathbf{C}\mathbf{x}_0)MS_{RES}]^{1/2} \quad . \tag{14b}$$

Therefore the 95% CI for the $E(Y \mid x_0) = \mu_{y|x_0}$, the mean of y at $\mathbf{x_0}$, is given by

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} - t_{0.025;n-k-1} \times se(\hat{y}_0) \leq \mu_0 \leq \mathbf{x}_0'\hat{\boldsymbol{\beta}} + t_{0.025;n-k-1} \times se(\hat{y}_0). \tag{14c}$$

Example 1 Continued. Estimate the mean of response y at $\mathbf{x_0} = \mathbf{FLC_5} = [1 \quad 7 \quad 4 \quad 180 \quad 5]'$ and obtain the 95% CI for $\mu_5 = \beta_0 + 7\beta_1 + 4\beta_2 + 180\beta_3 + 5\beta_4$. Note that this specified $\mathbf{x_0}$ is actually the **FLC** number 5 in our design matrix **X.** $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = [1 \quad 7 \quad 4 \quad 180 \quad 5] \times$

$$\begin{bmatrix} -0.91221 \\ 0.16073 \\ 0.21978 \\ 0.011226 \\ 0.10197 \end{bmatrix} = 3.62248;$$

thus the 5th residual is $e_5 = 4.6 - 3.62248 = 0.97752$. Further, $\mathbf{x}_0'\mathbf{C}\mathbf{x}_0 =$ 0.1051521 and from (14b), $se(\hat{y}_0) = [0.1051521 \times 0.6998]^{1/2} = 0.271267 \rightarrow$ HCIL = 0.558685 $\rightarrow$

$\mathbf{x}_0'\hat{\boldsymbol{\beta}} - 0.558685 \leq \mu_5 \leq 3.622479 + 0.558685 \rightarrow 3.063794 \leq \mu_5 \leq 4.181164 \rightarrow$ CIL = 4.1811635 – 3.0637945 = 1.117369.

THE PREDICTION INTERVAL (PI) 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 **X**), where $\bar{y}_0 = \sum_{r=1}^{N} y_{r0}/N$. The default value of N is always 1. 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($\bar{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}$) = 0 and V($\bar{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}$) = V($\bar{y}_0$) + $\mathbf{x}_0'$ COV($\hat{\boldsymbol{\beta}}$)$\mathbf{x_0}$ = $(\frac{1}{N} + \mathbf{x}_0'\mathbf{C}\,\mathbf{x_0})\sigma_{\epsilon}^2$. Therefore, the 95% PI for the future $\bar{y}_0$ is

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm t_{0.025;n-k-1} \times se(\bar{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}), \quad \text{or}$$

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} - t_{.025;n-k-1} \times se(\bar{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}) \le \bar{y}_0 \le \mathbf{x}_0'\hat{\boldsymbol{\beta}} - t_{.025;n-p} \times se(\bar{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}), \qquad (15a)$$

where $se(\bar{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \sqrt{(1/N + \mathbf{x}_0'\mathbf{C}\,\mathbf{x}_0) \times MS_{RES}}$ and p = k+1. (15b)

The PI in Eq. (15a) has a 95% probability to actually contain the future rv $\bar{y}_0$.

Example 1 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 **X**, but this specified $\mathbf{x_0}$ is within the range of our factor space. We wish to obtain an interval that has a Pr of 0.95 to contain the average of the N = 3 future observations, $\bar{y}_0$, made at $\mathbf{x_0}$. From Eq. (15b), the $se(\bar{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}})$ = 0.509595 $\rightarrow$ HPIL = 1.04953 $\rightarrow$ PI = 3.902144 $\pm$ 1.04953 $\rightarrow$ 2.852614 $\le \bar{y}_0 \le$ 4.951674; this last prediction interval has 95% Pr of containing the future rv $\bar{y}_0$ based on N = 3 observations made at $\mathbf{x_0}$ =[1 6 8 150 4]′, i.e., the Pr[2.852614 $\le \bar{y}_0 \le$ 4.951674] = 0.95. Note that the length of a PI for the rv $\bar{y}_0$ is always wider than the corresponding CI for the parameter μ_0 because a PI always has two sources of error (one from the fitted model and the other from future observations).

Exercise 8. Obtain a 95% PI for a single future observation to be made at $\mathbf{x_0}$ = $\mathbf{FLC_5}$ = [1 7 4 180 5]′ and compare the length of your PI against the corresponding CIL = 1.117369 atop page 16.

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.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 \qquad (16a)$$

Then, the explained variation (with 5 *df*) due to model (16a) is given by

$$SS_{Reg}(\text{due to } x_1, x_2, x_3, x_4, x_5) = \sum_{j=1}^{5} \hat{\beta}_j S_{jy} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 . \qquad (16b)$$

W/O loss of generality we consider the net (or partial) contributions of both variables x_2 and x_5 to the total explained SS in the model (16a). In order to compute this net contribution $SS_{Reg}(x_2, x_5 \mid x_1, x_3, x_4)$, we must 1st regress the response y on the variables x_1, x_3 , x_4 using the same n data points, which leads to a new regression model such as:

$$\widehat{y} = b_0 + b_1 x_1 + b_3 x_3 + b_4 x_4 \qquad (17)$$

The coefficients b_j (j = 1, 3, 4) in Eq. (17) are, in general, different from $\hat{\beta}_j$ (j = 1, 3, 4) of Eq. (16a) unless the (information) matrix **A** = **X'X** is diagonal in which case the **FLCs** [1 x_{i1} x_{i2} … x_{ik}], i = 1, 2,…, n, form an orthogonal design. The net contribution of x_2 and x_5 is defined as

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

$$= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \quad - \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = \sum_{j=1}^{5} \hat{\beta}_j S_{jy} - \sum_{j=1,3,4} b_j S_{jy} \rightarrow$$

$$SS_{Reg}(x_2, x_5 \mid x_1, x_3, x_4) = \sum_{i=1}^{n} \hat{y}_i^2 - \sum_{i=1}^{n} \widehat{y}_i^2 \qquad (18)$$

This last relationship in (18) 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 F statistic for testing $H_0 : \beta_2 = \beta_5 = 0$ is given by $F_0 = MS_{Reg}(x_2, x_5 \mid x_1, x_3, x_4)/MS_{RES} = \frac{SS_{Reg}(x_2, x_5 \mid x_1, x_3, x_4)/2}{MS_{RES}}$, where MS_{RES} is computed under the model (16a).

Finally, it can be proven, using the Gram-Schmidt orthogonalization procedure, that for any MLR model the coefficient of the last variable, namely $\hat{\beta}_k$ [for the model 16 (a) the value of k = 5], is the same for both the original non-orthogonal (or oblique) model and its orthogonal representation; the proof is very long and tedious. Since it is 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

$$\delta_r^2 = \hat{\beta}_r^2 / C_{r,r} = \sum_{j=1}^{k} \hat{\beta}_j S_{jy} - \sum_{j \neq r}^{k} b_j S_{jy} , \; r = 1, 2, \ldots, k. \qquad (19)$$

In fact it can be shown that the value of C_{kk} is the same for the (**X'X**) matrix and its corresponding orthogonal representation, where C_{kk} is the last diagonal element of **C**.

Therefore, from Eq. (19) the statistic for testing $H_0 : \beta_r = 0$ is $F_0 = \delta_r^2 / MS_{RES}$, which has an F distribution with $\nu_1 = 1$ and $\nu_2 = n - k - 1$ *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 **C** = $\mathbf{A^{-1}}$ = $\mathbf{(X'X)^{-1}}$, r = 0, 1, 2, ..., k where C_{00} pertains to $\hat{\beta}_0$, i.e., C_{11} is actually the element in the 2nd row and 2nd column of the matrix **C** = $\mathbf{(XTX)^{-1}}$. Further, $F_0 = (t_o)^2 = [\hat{\beta}_r / se(\hat{\beta}_r)]^2$.

Example 1 Continued (f). In order to obtain the net contribution of x_1 to the overall regression SS of the Eq. (6) model on page 6, we regress y on the independent variables x_2, x_3, and x_4, which results in the following model: $\hat{y}$ = – 0.086743 + 0.22332x_2 + 0.012285x_3 +

$0.095486x_4 \rightarrow SS_{Reg}(x_2, x_3, x_4) = 35.247613$; 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 $\beta_1 = 0$, we obtain $F_0 = \delta_1^2 / MS_{RES} = 4.1293 / 0.6998 = 5.901$, which is consistent with the Minitab's output because the value of $t_0^2 = (2.429)^2$ for the variable x_1 of Minitab is the same as the value of the partial F statistic $F_0 = 5.901$.

Exercise 9. (a) For the regression model of Example 1 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 these SS's, 1st the total regression of y on x_1 must be obtained. For Table 1 this leads to $\hat{y}_i = 2.41388 + 0.18892x_{i1}$ whose SS(Reg) = $0.18892 \times 30.580 = 5.7772$, which agrees with Minitab's output.

Second, in order to obtain the Seq SS due to x_2, we must regress y on x_1 and x_2 resulting in

$$\tilde{y}_i = 1.07655 + 0.168475x_{i1} + 0.21491x_{i2}$$

whose $SS_{Reg}(x_1, x_2) = 0.168475 \times 30.58 + 0.21491 \times 133.86 = 33.91982$. Therefore, the Seq $SS(x_2) = 33.91982 - 5.7772 = 28.1426$, which also agrees with the Minitab's output. In order to obtain the Seq $SS(x_3)$, 1st the regression of y on x_1, x_2 and x_3 must be obtained and second the corresponding $SS_{Reg}(x_1, x_2, x_3)$ must computed. Then, Seq $SS(x_3) = SS_{Reg}(x_1, x_2, x_3) - 33.91982$.

Exercise 10. Verify that Seq $SS(x_3) = 3.348$ and Seq $SS(x_4) = 2.1094$. Further, show that

$\sum_{j=1}^{k} \text{Seq SS}(x_j) \equiv SS_{Model}$ and hence $\delta_k^2 =$ Seq $SS(x_k)$, where x_k is last regressor.

PRESS RESIDUALS

For all statistical models, an ordinary residual is defined as the difference between the actual observation y_i and the value of y predicted from the model, i.e., $e_i = y_i - \hat{y}_i$. As an example, for the data of Table 1, the actual observed value of y at the first **FLC** is $y_1 = 1.4$ and the fitted y value is given (from Eq. 6) by

$\hat{y}_1 = -0.91221\times 1 + 0.16073\times 8 + 0.21978\times 4 + 0.011226\times 100 + 0.10197\times 1 = 2.4773 \rightarrow$

$e_1 = 1.4 - 2.4773 = -1.0773$. The prediction residual (or PRESS residual), $e_{(i)}$, is a measure of how well the model will predict the i^{th} observation y_i if the i^{th} observation is removed from the model and a new regression model is obtained with the remaining $(n-1)$ data points. For example, recall that if we remove the 9^{th} FLC from the model for Table 1, we will obtain the following regression model: $\hat{y}_{(9)} = -1.14930x_0 + 0.18387x_1 + 0.21915x_2 + 0.01127x_3 + 0.11102x_4$, whose SS(Total) now has 28 *df*. Inserting $\mathbf{FLC_9} = [1 \quad 4 \quad 7 \quad 140 \quad 3]^T$ into the above model yields $\hat{y}_{(9)} = 3.0311 \rightarrow e_{(9)} = y_9 - \hat{y}_{(9)} = 4.8 - 3.0311 = 1.76891$.

Since the value $e_{(9)}$ is large, then the regression model W/O the **FLC** number 9 is inadequate in predicting y_9. Because $\mathbf{FLC_9}$ is a very influential point, the full model (with all n = 30 points) should do better in predicting y_9. In fact, you may verify that $e_9 = y_9 - \hat{y}_9 = 4.8 - 3.147 = 1.65332 < e_{(9)}$.

Clearly, it is not practical to obtain n different regression models in order to compute the PRESS residuals $e_{(i)}$, i = 1, 2, 3, ..., n. Fortunately, it turns out that each $e_{(i)}$ can be computed, W/O resorting to further regression modeling, from

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \qquad (20)$$

where h_{ii} is the i^{th} diagonal element of the hat matrix $\mathbf{H} = \mathbf{XCX'} = \mathbf{XCX^T}$. To verify Eq. (20), I used Minitab to obtain $h_{99} = 0.065281$, which also agrees with my Excel file Example13.16. Substitution into (20) yields $e_{(9)} = e_9/(1 - h_{99}) = 1.65332/0.93472 = 1.76879$, which except for rounding error, is identical to the value of $e_{(9)}$ obtained atop this page.

It can easily be argued that $0 \le h_{ii} < 1$ and hence $e_{(i)} \ge e_i$ for all i =1, 2, ..., n. In fact, the Trace(**H**) = $\sum_{i=1}^{n} h_{ii}$ = Tr[$\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$] = Tr[$\mathbf{X'X}(\mathbf{X'X})^{-1}$] = Tr($\mathbf{I_{k+1}}$) = k + 1 = Rank (**X**) ≤ n. Note that although this development does not guarantee that each $h_{ii} < 1$ even if we have shown that their sum $\sum_{i=1}^{n} h_{ii} \le n$, but the fact that $V(e_i) = (1 - h_{ii})\,\sigma_{\in}^2 > 0$ guarantees that each h_{ii} must be less than 1 and the value of any h_{ii} cannot equal to 1 because the variance of no rv in the universe can equal to zero. The Rank(X) = k+1 because the design matrix **X** has exactly k+1 columns which, in general, should be independent; otherwise, Rank(**X**) < k+1 when some the regressors are significantly correlated. The prediction residual sum of squares (PRESS) is defined below.

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$$

Note that PRESS $\ge \sum_{i=1}^{n} e_i^2$ = SS(Residuals). Clearly, the smaller the value of PRESS is, the better the predictability of the model for the **FLCs** already in the model. Therefore, as a measure of overall model predictability we define

$$R_{Pred}^2 = 1 - (\text{PRESS}/SS_T) \le R_{Model}^2 . \qquad \text{(see Eq. 10.51 on p. 411 of Montgomery)}$$

MODEL BUILDING PROBLEMS IN MLR

Let q be the maximum possible number of regressor variables (q ≥ k) that are candidates for inclusion in the MLR model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \ldots + \beta_q x_q + \in . \qquad (21)$$

It is generally inefficient to include all the q independent variables in the model (21), rather the objective should be to identify a subset of size k from the q (> k) candidate regressors that satisfies the following 3 conditions:

(1) The value of the multiple determination coefficient R_k^2 = $SS_{Reg}(x_1, x_2, \ldots, x_k)/SS(Total)$ exceeds at least, say 70%, i.e., the explained variation in the mean of response y by the k < q regressors (if at all possible) is at least 70%. The explained percent of 70% is somewhat arbitrarily selected by me; if it is possible to obtain a model with $R_k^2 \geq 80\%$ (similar to Pareto Principle of 80/20), the regression modeler should attempt to do so only by adding regressors that contribute significantly to SS_{Reg}.

(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 using the adjusted R_k^2 as defined below:

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

The value of above $\bar{R}_k^2$, unlike R_k^2, may actually decrease as k increases toward q because in general n > k. The set of k regressors out of the q independent variables should be selected in such a manner that has the maximum (or close to maximum) $\bar{R}_k^2$ among the ${}_qC_k = {}_q(\text{Choose})_k$ = q!/[k!(q – k)!] possible sets.

(3) The value of the C_p statistic

$$C_p = [SS_k(RES)/MS_q(RES)] - n + 2(k+1)$$

must not exceed k but should be fairly close to but less than k. This is due to the fact that C_p is an estimator of the total standardized Mean Square Error, given by $E\sum_{i=1}^{n}(\hat{y}_i - \mu_i)^2/\sigma_\epsilon^2$, and hence the regressors $x_1, x_2, \ldots, x_k$ (k < q) must be selected in such a manner that minimize C_p relative to k. Note that we could easily refer to the above C_p statistic as C_k because it refers to a model containing k regressors, but statistical software's generally refer to it as the C_p or C-p statistic. Further, $E(C_p) = E(C_k) \cong k$.

There are several model building procedures in regression that generally lead to the same "most parsimonious" (i.e., the least number of regressors) 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 $\alpha_{in} = 0.15$. Minitab recommends $\alpha_{in} = \alpha_{out} = 0.15$, and personally I believe α should not exceed the range 0.20-0.25 significance level. Further, the experimenter must use his/her judgment to distinguish between statistical and practical significance.

The most 4 common model building procedures are FORWARD Selection, Stepwise Regression, Backward Elimination, and Best Subsets (or MAXR) procedures. FORWARD Selection starts with the best regressor, i.e., the one with the largest R^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 q regressors in the model, then drops the least significant one, then the next, and the next, etc. Minitab's Best Subsets (or SAS's MAXR) procedure is a rather long and tedious procedure, but it basically finds the best (i.e., with the largest R^2) one-variable regression model, then the best 2-regressor model, then the best 3-regressor 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, BSR). I recommend the following 3 criteria for selecting one out of the k regression models of MAXR or Minitab's BSR. (i) The F_0(Model) should be nearly largest (or the *P-value* for testing the model significance should be nearly the smallest) amongst all the k 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 $C_p > k$ exhibit too much bias. Further, the selected model with k or less regressors should have all coefficients significant at most, say, at the 20% level, and Lack-of-Fit

(in case of repeat observations at a **FLC**) should not be significant at the 5% level. Finally, the experimenter must leave sufficient *df* ($\nu_2 \geq 6$) for MS_{RES} = MS(Error) so that the F-test 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 $\nu_2 < 5$. By now you should deduce that all ANOVA models are special cases of regression models, where ANOVA models pertain to planned experiments, while regression procedures can analyze data from all experiments.

Testing for Lack of Fit (Section 10.8 of Montgomery)

The regression model given in Eq. (6) cannot be tested for LOF (Lack of Fit) because there are no repeat (or replications) at any of the 30 **FLCs**. In experiments where some FLCs are repeated, then the corresponding regression model can be tested for LOF. I have discussed this subject-matter extensively in my SLREG document (pp.156-161 from my STAT3610 notes) and is listed on my website. It is required that you download this document and review it ASAP.

CORRELATION

Regression is generally applicable if the regressor (or independent) variable on the RHS of the model can be controlled by the experimenter so that V(X) = 0. In studies where both variables X and Y have to be measured from the same sampling unit, i.e., the (x, y) pair form a bivariate random vector, then it is best to conduct a correlation analysis than regressing y on x. To this end, let $[X_1 \quad X_2]^T$ be a 2×1 bivariate vector; then the population total correlation coefficient (or partial correlation of order zero) between X_1 and X_2 is defined as

$$\rho_{12} = \frac{\mathrm{COV}(X_1, X_2)}{\sigma_1 \times \sigma_2} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \rho,$$

where $\sigma_{12} = E[(X_1 - \mu_1)\times(X_2 - \mu_2)] = E(X_1X_2) - E(X_1)\times E(X_2)$ is the covariance between the random variable pair X_1 and X_2. It can be proven that $-1 \le \rho \le 1$, or $|\rho| \le 1$, which you will be asked to do in the following Bonus Exercise.

Bonus 2. Prove that $|\rho| \le 1$ by expanding the $V(aX_1 + bX_2)$, where a and b are arbitrary real constants of your choice.

When $\rho = \pm 1$, the two random variables X_1 and X_2 are said to be perfectly correlated. If X_1 and X_2 are independent rvs, then $\rho = 0$, but $\rho = 0$ does not always imply that X_1 and X_2 are independent. However, if X_1 and X_2 have the joint bivariate normal density function, then $\rho = 0$ does guarantee that X_1 and X_2 are independent.

In practice, the value of the population correlation coefficient, ρ_{12}, is unknown and has to be estimated from a random sample of size n pairs. The sample point estimate of ρ is given by

$$\hat{\rho} = r = \frac{\hat{\sigma}_{12}}{S_1 S_2} = \frac{S_{12}/(n-1)}{S_1 S_2}, \qquad (22a)$$

where the numerator of r is the sample covariance $\hat{\sigma}_{12} = S_{12}/(n-1) =$

$\frac{1}{n-1}\sum_{i=1}^{n}[(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)]$; equation (22a) can also be written as

$$r = \hat{\rho} = \frac{S_{12}}{\sqrt{S_{11} S_{22}}} = \frac{\sum_{i=1}^{n}\left[(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)\right]}{\sqrt{\left[\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right]\times\left[\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2\right]}} \qquad (22b)$$

Example 2. The following data give the final averages of 15 randomly selected ISE students in Engineering Statistics (X_1) and Operations Research (X_2 = OR).

X_1: 86% 75 63 64 92 58 78 90 85 77 69 82 84 94 76%

X_2: 77% 85 70 57 83 69 76 82 95 87 62 86 83 85 88%

The sample statistics are: $\sum_{i=1}^{15} x_{i1} = 1173$, $\bar{x}_1 = 78.20$, $\sum_{i=1}^{15} x_{i2} = 1185$, $\bar{x}_2 =$

79.00, $\sum_{i=1}^{15} x_{i1}^2 = 93405$, $\sum_{i=1}^{15} x_{i2}^2 =$ = 95145, $\sum_{i=1}^{15} x_{i1}x_{i2} = 93755$, $S_{11} = 1676.40$,

$S_{22} = 1530$, $S_{12} = 1088$, $S_1 = 10.9427$, $S_2 = 10.4540$, $\hat{\sigma}_{12} = 77.7143$, $r = \hat{\sigma}_{12}/(S_1S_2) = 0.67935$. Note that if x_2 is regressed on x_1 and the resulting model R^2 is computed, then $r = \sqrt{R^2_{\text{Model}}}$.

TEST OF HYPOTHESIS ABOUT ρ

There are two different tests that can be conducted on the population parameter ρ: (1) H_0: ρ = 0, (2) H_0: ρ = ρ_0, where $\rho_0 \neq 0$.

(1) Testing H_0: ρ = 0 versus one of the 3 alternatives H_1: ρ ≠ 0, or H_1: ρ < 0, or H_1 : ρ > 0.

Recall that statistical inference on a parameter is conducted using the sampling distribution (SMD) of the point estimator. The point estimator of ρ is the sample correlation coefficient r. It can be shown that the null SMD of the statistic

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (23)$$

follows a Student's t-distribution with $(n - 2)$ *df*. For the example 2 above, the most appropriate alternative is H_1: ρ > 0. Therefore, the critical region for testing H_0: ρ = 0 at the LOS α = 0.05 is (1.771, ∞). The value of our test statistic from (23) is $t_0 = (0.67935\sqrt{13})/\sqrt{1-0.46152}$ = 3.3379, which easily exceeds $t_{0.05,13}$ = 1.771, leading to the rejection of zero correlation between X_1= Engr Stat and X_2 = OR. The *P-value* (or the Pr level) of the test is $\hat{\alpha}$ = $P(T_{13} \geq 3.33794) = 0.002672$, which is less than 0.05 as expected because H_0 was rejected at the 5% level of significance. If the assumption of joint bivariate normal density for the 2×1 vector $[X_1 \quad X_2]^T$ is indeed tenable, then the rejection of H_0: ρ = 0 implies that X_1 and X_2 are not independent; otherwise, the rejection of H_0 implies that X_1 and X_2 are linearly related.

Exercise 11(a). Use results from regression to show that the statistic $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ has a Student's t sampling distribution. Hint: First refer to an ANOVA table for the regression of X_2 on X_1 and then use the fact that $F_{1,\nu_2} = t^2_{\nu_2}$.

(2) Testing H_0: $\rho = 0.50$ versus the alternative H_1: $\rho \neq 0.50$ at the LOS $\alpha = 0.05$.

The t_0 given in equation (23) cannot be used to test H_0: $\rho = 0.50$ because the expression for t_0 is free of ρ. However, Sir R. A. Fisher (1921, *Metron* 1, No.4, 1) found the remarkable logarithmic transformation

$$\mathcal{Z} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \text{arctanh}(r) = \tanh^{-1}(r)$$

whose sampling distribution (SMD) approaches normality very much faster (perhaps as much as 10 times faster) than that of r. That is, $\mathcal{Z} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ is approximately normal with $E(\mathcal{Z}\,|\,\rho) \cong \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \tanh^{-1}(\rho)$ and $V(\mathcal{Z}) \cong 1/(n-3)$. For the Example 2 above, the approximate SMD of $\mathcal{Z}$ is depicted in Figure 1, where under the null hypothesis $E(\mathcal{Z}\,|\,\rho = 0.50) \cong \frac{1}{2}\ln\left(\frac{1+0.50}{1-0.50}\right) = 0.54931 = \mu_{\mathcal{Z}}$ and $V(\mathcal{Z}) \cong 1/(15-3) = 0.083333$. The acceptance interval for testing the 2-sided H_0: $\rho = 0.50$ consists of values of $\mathcal{Z}$ given by $(A_L, A_u) = (\mathcal{Z}_L, \mathcal{Z}_u) = (-0.0165, 1.1151)$. The value of the test statistic is $\mathcal{Z} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.67935}{1-0.67935}\right) =$ arctanh(0.67935) = 0.82791, which is well inside the AI = (– 0.0165, 1.1151) and hence we cannot reject H_0: $\rho = 0.50$. The *P-value* of the test is given by $\hat{\alpha} = 2\times\Pr(\mathcal{Z} \geq 0.82791) = 2\times\Pr[Z \sim N(0,1) \geq 0.9651) = 0.33450$. Therefore, we cannot deduce that the data provide sufficient evidence for $\rho \neq 0.50$ but they do provide sufficient evidence that $\rho > 0$. Note that Kendall and Stuart (Vol. 1, 2nd edition, p. 391) give a better approximation for

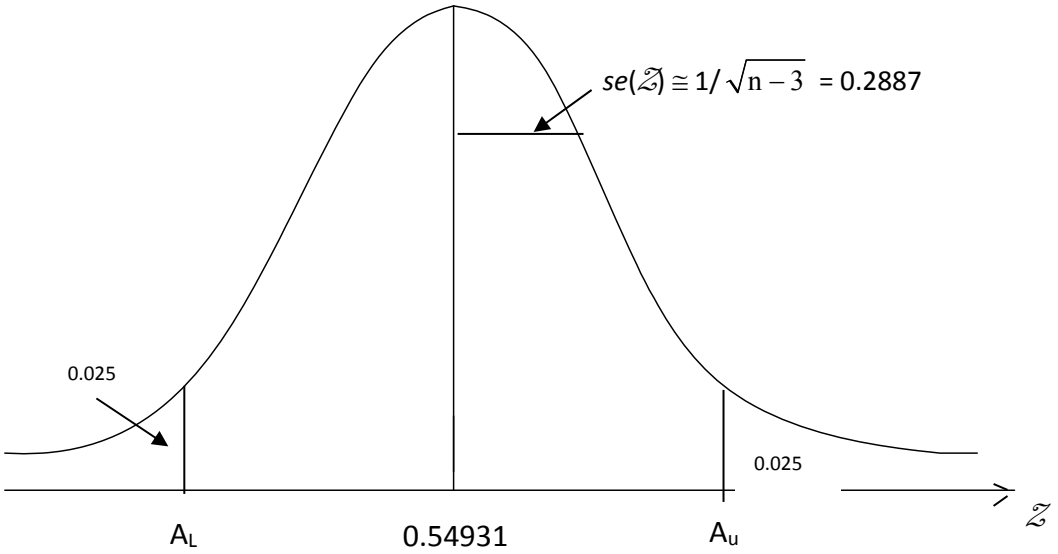

Figure 1. The Approximate SMD of $\mathcal{Z}$ under H_0

V[$\mathcal{Z} = \frac{1}{2}\ln(\frac{1+r}{1-r})$] as $\frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2}$. For small ρ and n > 11, this last approximation is almost equal to 1/(n–3).

OBTAINING a 95% CI FOR ρ

Again, we have to make use of the fact that $\mathcal{Z} = \frac{1}{2}\ln(\frac{1+r}{1-r})$ = arctanh(r) = $\tanh^{-1}(r)$ ~ N[$\frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$, 1/(n –3)] = N[$\tanh^{-1}(\rho)$, 1/(n –3)], as depicted in Figure 2, which clearly shows that

$$\Pr[\mu_{\mathcal{Z}} - 1.96\times se(\mathcal{Z}) \le \frac{1}{2}\ln(\frac{1+r}{1-r}) \le \mu_{\mathcal{Z}} + 1.96\times se(\mathcal{Z})] = 0.95 \qquad (24)$$

After several algebraic steps, Eq. (24) will show that in general ρ_L = tanh[arctanh(r) –

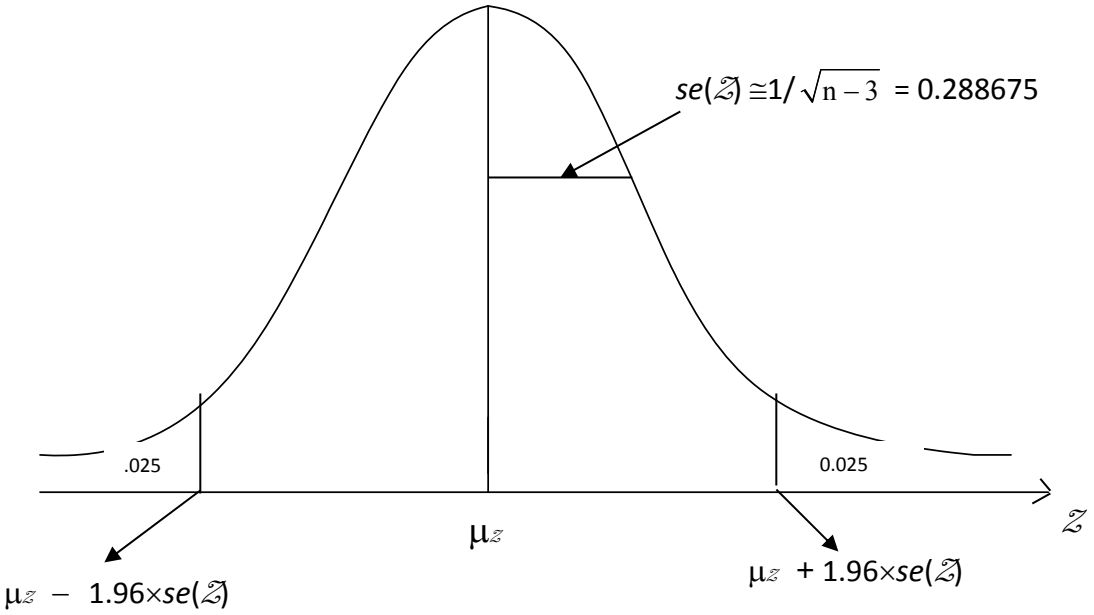

Figure 2. The Approximate SMD of $\mathcal{Z}$

$Z_{\alpha/2}/\sqrt{n-3}\,]$ = tanh$\left(\frac{1}{2}\ln\frac{1+r}{1-r} - Z_{\alpha/2}/\sqrt{n-3}\right)$, where tanh(r) = $\frac{e^{r}-e^{-r}}{e^{r}+e^{-r}}$ and $Z_{0.025}$ = 1.959964 (≅ 1.96). For our example, the 95% lower confidence limit is ρ_L = tanh(0.26210) = 0.25626. Further, ρ_U = tanh[arctanh(r) + $Z_{0.025}/\sqrt{n-3}\,]$ so that the 95% upper confidence limit for our example is ρ_u = tanh(1.39371) = 0.88398, i.e., $0.25626 \le \rho \le 0.88398$. As expected, this last CI does include the hypothesized value of $\rho_0 \equiv 0.50$ because the null hypothesis $H_0 : \rho = 0.50$ could not be rejected at the 5% level.

A closer look at the above correlation analysis between X_1 = Engr Stat & X_2 = OR average grades at semester's end reveals that the zero-order correlation r_{12} is not an exact measure of true linear relationship between X_1 and X_2 because there are several other variables that impact student overall avenges in any one course, i.e., the correlation between Engr Stat and OR may be purely incidental because both variables are highly correlated with other variables listed below. In our example, both X_1 = Engr Stat & X_2 = OR are impacted by X_3 = Average study time per week, X_4 = Student's IQ, X_5 = Socio-Economic conditions, X_6 = Student's diet & health, Student's amount of interest in the course, etc. If we cannot sample

students with identical IQ, Average study time per week, the same Socio-Economic conditions, etc, then we have to make use of partial (or net) correlation of order larger than zero. That is, a true interdependence between X_1 and X_2 must be evaluated while variables X_3, X_4, X_5,… are held constant, or their effects X_3, X_4, X_5,… are removed from both X_1 and X_2. This leads us to the following definition.

Partial CORRELATION (PC)

The sample PC coefficient (SPCC) of order (or rank) 1 between X_1 and X_2 , removing the impact of a 3rd variable X_3 on both X_1 and X_2, is defined as

$$r_{12.3} = \frac{-C_{12}}{(C_{11}C_{22})^{1/2}}, \qquad (25a)$$

where C_{ij} is the cofactor of the (i, j)th element in the correlation matrix

$$\mathbf{r} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}, \qquad (25b)$$

which is always square and symmetrical so that $r_{ij} = r_{ji}$ for all $i \neq j$, and for obvious reason $r_{ii} \equiv 1$ for all diagonal elements. In Eq. (25a), subscripts 1 & 2 are *primary* and subscript 3 is said to be *secondary*. The cofactor C_{ij} of a matrix is given by $C_{ij} = (-1)^{i+j} \times M_{ij}$, where M_{ij} is called the *minor* of submatrix with both the ith row and jth column removed from the original square matrix. A minor, M_{ij}, is the determinant of the resulting square submatrix with both the ith row and jth column removed from the original matrix. For the correlation matrix, **r** in (25a), the *minors* and cofactors are as follows. $C_{11} = M_{11} = \det\begin{bmatrix} 1 & r_{23} \\ r_{32} & 1 \end{bmatrix} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$ (due to symmetry); $C_{21} = -M_{21} = -\det\begin{bmatrix} r_{12} & r_{13} \\ r_{32} & 1 \end{bmatrix} = -\begin{vmatrix} r_{12} & r_{13} \\ r_{32} & 1 \end{vmatrix} = -r_{12} + r_{13}r_{32} = -r_{21} + r_{31}r_{23} = C_{12}$ (due to

symmetry); $C_{31} = M_{31} = \det\begin{bmatrix} r_{12} & r_{13} \\ 1 & r_{23} \end{bmatrix} = \begin{vmatrix} r_{12} & r_{13} \\ 1 & r_{23} \end{vmatrix} = r_{21}r_{23} - r_{13} = r_{21}r_{32} - r_{31} = M_{13}$ (due to

symmetry), etc. Note that the Laplace expansion of the det(**r**), in terms of its column 1, gives the det(**r**) = |**r**| = $1\times C_{11} + r_{21}\times C_{21} + r_{31}\times C_{31}$. For the 3×3 matrix **r** in (25b) there are 5 other Laplace expansions of the |**r**|. Substituting the above cofactors into Eq. (25a) results in

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{[(1-r_{13}^2)(1-r_{23}^2)]^{1/2}} \tag{26}$$

The sample PCC of order 2 between X_1 and X_2, conditioned on X_3 and X_4 held fixed (or held constant, i.e., removing the impacts of both X_3 and X_4 on X_1 and X_2), is given by

$$r_{12.34} = \frac{-C_{12}}{(C_{11}C_{22})^{1/2}}, \tag{27a}$$

where C_{ij} now is the cofactor in the symmetric correlation matrix

$$\mathbf{r} = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{bmatrix}. \tag{27b}$$

Again, note that $r_{12.34}$ gives the net (or partial) correlation between X_1 and X_2, while removing the impacts of X_3 and X_4 on both X_1 & X_2. For the correlation matrix in Eq. (27b), we have $C_{12} = -M_{12} = -\det\begin{bmatrix} r_{21} & r_{23} & r_{24} \\ r_{31} & 1 & r_{34} \\ r_{41} & r_{43} & 1 \end{bmatrix} = -\begin{vmatrix} r_{21} & r_{23} & r_{24} \\ r_{31} & 1 & r_{34} \\ r_{41} & r_{43} & 1 \end{vmatrix} = -r_{21}(1-r_{34}^2) + r_{31}(r_{23} - r_{24}r_{34}) - r_{41}(r_{23}r_{34} - r_{24})$.

In the case of multivariate normal underlying population, Kendall & Stuart (1972) give the variance of $r_{12.345\ldots}$ as

$$V(r_{12.345\ldots}) = \frac{1}{n}(1-\rho_{12.345\ldots}^2)^2 \tag{28a}$$

Thus, under the null hypothesis H_0: $\rho^2_{12.345\ldots} = 0$, when n is moderately large, say n > 15, the approximate sample variance reduces to 1/n with *se* $\cong 1/\sqrt{n}$. If $\rho^2_{12.345\ldots}$ is not specified under H_0, then the variance in (28a) reduces approximately to

$$v(r_{12.345\ldots}) \cong \frac{1}{n}(1 - r^2_{12.345\ldots})^2 \qquad (28b)$$

It has been shown in statistical literature that the SMD of $(r_{12.34\ldots}\sqrt{n-2-nsv}) / (1 - r^2_{12.345\ldots})^{1/2}$ is that of Student's t with *df* = n−2−nsv, where nsv is the number of secondary variables. This last statistic can be used to test only $\rho^2_{12.345\ldots} = 0$, and no other values of $\rho^2_{12.345\ldots}$.

Before providing an example of SPCC computation, it should be noted that there is another interesting way to compute the sample PCC $r_{12.345\ldots}$. First, regress the primary variable X_1 on all the secondary variables X_3, X_4, X_5, … and obtain the residuals e_{11}, e_{21}, …, e_{n1}. Second, regress X_2 on all secondarys X_3, X_4, X_5, … and obtain the residuals e_{12}, e_{22}, …, e_{n2}. Then,

$$r_{12.345\ldots} = \frac{\sum_{i=1}^{n}(e_{i1} \times e_{i2})}{\sqrt{(\sum e^2_{i1}) \times (\sum e^2_{i2})}} \qquad (29)$$

Example 3. The following data give the overall averages in Engr Stat & OR of 15 ISE students with their corresponding average number of hours spent per week on both courses, while we are assuming that each student roughly apportioned his/her time almost equal amount to each course.

X_1: 86% 75 63 64 92 58 78 90 85 77 69 82 84 94 76%

X_2: 77% 85 70 57 83 69 76 82 95 87 62 86 83 85 88%

X_3: 9.5 8.7 7.2 7.2 10.3 4.1 7.7 11.8 10.9 7.0 3.8 9.3 8.5 10.9 7.6 hrs/week

Our objective is to compute the value of $r_{12.3}$ and test its statistical significance at the 5% level. As in **Example 2** on page 26 of these notes, r_{12} = 0.679351, and similar calculations show that

r_{13} = 0.851637, and r_{23} = 0.6518645. Inserting these zero-order correlation coefficients into Eq. (26), we obtain $r_{12.3} = (r_{12} - r_{13}r_{23}) / [(1-r_{13}^2)(1-r_{23}^2)]^{1/2}$ = 0.31247642. Thus, when we remove the impact of average study-time, X_3, from both the Engr Stat and OR average grades, the net correlation has been reduced to 0.31248 from the total correlation of 0.679351. Using the fact that the SMD of $(r_{12.34\ldots}\sqrt{n-2-nsv}) / (1-r_{12.345\ldots}^2)^{1/2}$ is that of a Student's t with 12 *df*, we obtain t_o = 1.1395105. Therefore, the *P-value* of the test is equal to $\hat{\alpha}$ = 2Pr($t_{12} \geq t_0$) = 0.276726, which is not at all significant at the 5% level, so that we cannot reject $H_0 : \rho_{12.3} = 0$ at significant levels as high as 25%. This implies that the correlation between Engr STAT averages with that of OR may have been incidental from a statistical standpoint. Furthermore, I used MS Excel and regressed X_1 on X_3 and found that $\sum_{i=1}^{15} e_{i1}^2$ = 460.531677; then, regressed X_2 on X_3 and obtained $\sum_{i=1}^{15}(e_{i1} \times e_{i2})$ = 198.9088493 and $\sum_{i=1}^{15} e_{i2}^2$ =879.861281. The use of Eq. (29) now yields $r_{12.3}$ = 198.9088493/ $\sqrt{(460.5316769)\times(879.8612809)}$ = 0.31247642, as before. Note that the secondary variable X_3 somewhat masks another variable, namely the student's amount of interest in any one course; e.g., student number 1 has X_3 = 9.5 hours, which may equal to 5.5 hrs/week on Engr Stat and only 4 hrs/week on OR. However, in the middle of page 33 I made the assumption that each student had almost equal-interest (another secondary variable) in Engr Stat & OR, notwithstanding the fact that both courses are highly quantitative.

Exercise 11(b). The IQ's of the 15 students of Example 3, in the same order as the data atop page 33, are

X_4: 120 118 115 114 123 119 117 126 121 118 115 120 121 119 120

Compute the value of 2nd-order SPCC, $r_{12.34}$, using both formulas (27 a &b) and (29). **(c)** Test its statistical significance at the nominal level of 5%.