

1. The data shown below relate to a study on the reaction of formaldehyde with cotton cellulose, and were given by Devore (1991). The data consist of measurements on three predictor variables and the response Y with $n = 30$ cases. The variables are:

X_1 = HCHO (formaldehyde) concentration,

X_2 = catalyst ratio,

X_3 = curing temperature,

Y = durable press rating, a quantitative measure of wrinkle resistance.

X_1	X_2	X_3	Y	X_1	X_2	X_3	Y
8	4	100	3.4	4	10	160	4.6
2	4	180	3.2	4	13	100	4.3
7	4	180	4.6	10	10	120	4.9
10	7	120	4.9	5	4	100	2.9
7	4	180	4.6	8	13	140	4.6
7	7	180	4.7	10	1	180	3.6
7	13	140	4.6	2	13	140	3.1
5	4	160	4.5	6	13	180	4.7
4	7	140	4.8	7	1	120	3.4
5	1	100	2.4	5	13	140	4.5
8	10	140	4.7	8	1	160	3.1
2	4	100	2.6	4	1	180	2.8
4	10	180	4.5	6	1	160	2.5
6	7	120	4.7	4	1	100	2.3
10	13	180	4.8	7	10	100	4.6

These data are stored in the file: /u/r/e/reinsel/stat333/press-rating.dat

- i) Provide simple scatter plots of Y versus X_1 , X_2 , and X_3 , respectively, and make any comments that seem suitable.

- ii) Fit a full second-order polynomial (response) model to the data, of the general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \varepsilon$$

Follow by performing a complete analysis, leading to selection of a ‘reasonable’ final reduced model (one that satisfies the origin shift criterion, e.g., see pp. 267–268 in text-book). A ‘complete’ analysis should provide details and include formal justifications (e.g., consideration of F-tests, R^2 , adjusted- R^2 , and S^2 values) for the selection of a final reduced model, and should include the usual plots and examination of (various types of) residuals from the final fitted model for checking assumptions and adequacy of the model.

- iii) As an additional exercise for these data, for your final fitted model obtain and examine both the standardized and studentized residuals, and find the Cook statistic value corresponding to each data case. Check to assess whether any of the cases seem ‘unusual’ in terms of outlier behavior or extreme influence, and discuss.

2. i) For the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \dots, n$, show that the i th diagonal of the ‘hat’ matrix \mathbf{H} , $h_{ii} \equiv \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$ takes the form $h_{ii} = (1/n) + (X_i - \bar{X})^2/S_{xx}$, where $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$. (Hint: It is convenient to express the model in ‘centered’ form

as $Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \varepsilon_i$, so that $\mathbf{X}_i = [1, X_i - \bar{X}]'$ and $\mathbf{X}'\mathbf{X} = \text{Diag}\{n, S_{xx}\}$.)

ii) For the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$, $i = 1, \dots, n$, determine a condition or circumstance under which the h_{ii} values will take the explicit form $h_{ii} = (1/n) + (X_{i1} - \bar{X}_1)^2/S_{x_1x_1} + (X_{i2} - \bar{X}_2)^2/S_{x_2x_2}$, where $S_{x_2x_2} = \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2$. Verify your result under the stated condition. (Again, consider the ‘centered’ form $Y_i = \beta_0^* + \beta_1 (X_{i1} - \bar{X}_1) + \beta_2 (X_{i2} - \bar{X}_2) + \varepsilon_i$, and consider circumstances involving orthogonality.)

iii) When the special results for the model in (ii) do hold, also give simple expressions for the LSE b_2 of β_2 in the model, for $\text{Var}(b_2)$, and for $\text{SSR}(b_2|b_0, b_1)$, involving $S_{x_2x_2}$.

3. Suppose a response variable Y is fitted by LS using the straight line model $E(Y) = \beta_0^* + \beta_1^* X$, based on a sample of $n = 7$ observations with X -values equal to $-5, -3, -1, 0, 1, 3, 5$. However, it is feared that there may be some additional quadratic effect and that Y may actually follow the quadratic model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$.

i) Under the assumption that the quadratic model was actually the true model, determine the biases of the LS estimates b_0 and b_1 in estimating β_0 and β_1 when the straight line model is estimated. [Note: You need to calculate the ‘bias’ matrix (column vector in this case), $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$, where $\mathbf{X}_1 = [1, \mathbf{X}]$ with $\mathbf{X} = (-5, -3, -1, 0, 1, 3, 5)'$, and $\mathbf{X}_2 = (25, 9, 1, 0, 1, 9, 25)'$ is the column of X^2 -values.]

ii) Use the result from Problem 4(ii) to find the expected value of the MSE_2 from fitting the ‘reduced’ straight line model $E(Y) = \beta_0^* + \beta_1^* X$, when the quadratic model is actually the true model. [Note: You need to find the values of $\tilde{\mathbf{X}}_2 = \mathbf{X}_2 - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$.]

4. Consider the linear regression model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ is $n \times p$ and \mathbf{X}_1 is $n \times p_1$, $p_1 < p$. The ‘extra’ regression sum of squares due to inclusion of the \mathbf{X}_2 terms, after the \mathbf{X}_1 terms, is

$$\text{SSR}(\mathbf{b}_2 | \mathbf{b}_1) = S_1 - S_2 = \mathbf{b}_2'\tilde{\mathbf{X}}_2'\mathbf{Y} = \mathbf{Y}'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\mathbf{Y} \equiv \mathbf{Y}'\tilde{\mathbf{H}}_2\mathbf{Y},$$

say, where $\mathbf{b}_2 = (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\mathbf{Y}$, $\tilde{\mathbf{H}}_2 = \tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'$, and $\tilde{\mathbf{X}}_2 = \mathbf{X}_2 - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = (\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2$. $\text{SSR}(\mathbf{b}_2 | \mathbf{b}_1)$ is also the ‘hypothesis’ sum of squares for testing $H_0: \boldsymbol{\beta}_2 = \mathbf{0}$.

Also recall the useful result that if \mathbf{Y} is a $n \times 1$ random vector with mean vector $\boldsymbol{\mu} = E(\mathbf{Y})$ and covariance matrix $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{Y})$, and \mathbf{A} is a $n \times n$ symmetric matrix of constants, then the random variable $Q = \mathbf{Y}'\mathbf{A}\mathbf{Y}$ (a quadratic form in \mathbf{Y}) has mean or expected value equal to

$$E(Q) = E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu},$$

where $\text{tr}(\mathbf{B})$ denotes the *trace* of a matrix \mathbf{B} , the sum of its diagonal elements.

i) Use the result above to determine the *expected value* of $\text{SSR}(\mathbf{b}_2 | \mathbf{b}_1)$, i.e., determine $E[\text{SSR}(\mathbf{b}_2 | \mathbf{b}_1)]$; hence also give the expected value of the mean square $\text{MSR}(\mathbf{b}_2 | \mathbf{b}_1) = \text{SSR}(\mathbf{b}_2 | \mathbf{b}_1)/(p - p_1)$. Express your results in simplest terms by noting that $\mathbf{X}_1'\tilde{\mathbf{X}}_2 = \mathbf{0}$ and also $\mathbf{X}_2'\tilde{\mathbf{X}}_2 = \mathbf{X}_2'(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2 = \mathbf{X}_2'(\mathbf{I} - \mathbf{H}_1)(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2 \equiv \tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2$, showing in particular that the expected values do not involve the parameters $\boldsymbol{\beta}_1$, only the parameters $\boldsymbol{\beta}_2$ and σ^2 .

ii) When the ‘reduced’ model $E(\mathbf{Y}) = \mathbf{X}_1\boldsymbol{\beta}_1^*$ is estimated by LS, we have the LS estimate $\mathbf{b}_1^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$ and we know that the residual SS is

$$\text{SSE}_2 = \mathbf{Y}'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}_1^*\mathbf{X}_1'\mathbf{Y} = \text{SSE}_1 + \text{SSR}(\mathbf{b}_2 | \mathbf{b}_1),$$

where SSE_1 is the residual SS from fitting the ‘full’ model. Using the known fact (e.g., see Problem 2 of Assignment 4) that $E[\text{SSE}_1] = (n - p)\sigma^2$ and the results from (i), determine $E[\text{SSE}_2]$ and hence also $E[\text{MSE}_2]$, where $\text{MSE}_2 = \text{SSE}_2/(n - p_1)$, under the assumption that the ‘full’ model is the true model, i.e., $\boldsymbol{\beta}_2 \neq \mathbf{0}$.