**Stat-440/640 Regression and Time Series Analysis**
**Fall 2018**

Exam 1B

Name ________________________________

show work to ensure full credit

1. Soluble dietary fiber (SDF) can provide health benefits by lowering blood cholesterol and glucose levels. The article "Effects of Twin-Screw Extrusion on Soluble Dietary Fiber and Physicochemical Properties of Soybean Residue" (Food Chemistry, 2013: 884–889) reported the data SDF.txt on $y$ = SDF content (%) in soybean residue and the three predictors extrusion temperature ($x_1$, in $^0C$), feed moisture ($x_2$, in %), and screw speed ($x_3$, in rpm) of a twin-screw extrusion process.

   Answer the following questions using the data Soluble Dietary Fiber (SDF) posted on D2L under the name SDF.txt.

   (a) (10 pts) Find the fitted least squares linear regression model (with the regression coefficients reported to five decimals).

   (b) (10 pts) Is there very convincing evidence for including that at least one of the second-order predictors is providing useful information over and above what is provided by the three first-order predictors? Justify the answer by fitting an appropriate regression model.

(c) (15 pts) Test if the linear regression model (found in (b)) is significant at $\alpha = 0.05$ level of significance. Carry out the test step-by-step by formulating the appropriate null and alternative hypotheses, ANOVA table, test statistic, critical value, conclusion etc.

(d) (15 pts) Should the interaction predictors be included in the model? Fit an appropriate model and justify this using the $t$- test. Which independent variables are significantly related to $y$.

(e) (10 pts) Find both $R^2$ and $\bar{R}^2$. What does $R^2$ tell us about our model in part (a) and part(b)?

(f) (15 pts) Construct a 95% prediction interval on the SDF content for a given soybean residue when $x_1 = 26$, $x_2 = 100$, $x_3 = 180$ using model in part(b) and part (d). Should you include interaction predictors in the model?

2. (5 bonus pts) It is more convenient to deal with multiple regression models if they are expressed in matrix notation. This allows a very compact display of the model, data, and results. In matrix notation, the model is given by

$$y = X\beta + \epsilon \text{ , where}$$

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

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The least-squares estimator for the parameter vector $\beta$ is given by $\hat{\beta} = (X^T X)^{-1} X^T y$ and the fitted regression model is $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$ , where $H = X(X^T X)^{-1} X^T$.

Show that $SS_T$, $SS_{reg}$, and $SS_E$ can be expressed i terms of of the following quadratic forms:

$SS_T = y^T(I - \frac{1}{n}J)y$ , $SS_{reg} = y^T(H - \frac{1}{n}J)y$, and $SS_E = y^T(I - H)y$,
where $I$ is the identity matrix, and $J$ is a matrix of one's with appropriate sizes.