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

Exam 1B

Name Justin Hood

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).

We consider the model where, $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

We compute,

$\beta_0 = 2.73000$
$\beta_1 = 0.07925$
$\beta_2 = 0.03288$
$\beta_3 = 0.01431$

$\Rightarrow y \sim 2.73000 + 0.07925 x_1 + 0.03288 x_2 + 0.01431 x_3$

(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.

Second order Model:
$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2$
where,
$\beta_0 = -1.197 \times 10^2$
$\beta_1 = 1.499$
$\beta_2 = 6.392 \times 10^{-1}$
$\beta_3 = 7.478 \times 10^{-1}$
$\beta_4 = -2.700 \times 10^{-2}$
$\beta_5 = -2.756 \times 10^{-3}$
$\beta_6 = -2.037 \times 10^{-3}$

We first consider the model, $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2$
The p-value for the tests $H_0: \beta_i = 0$; $H_A: \beta_i \neq 0$ are then,

$P(\beta_4) = .000107 < .05$
$P(\beta_5) = 1.50 \times 10^{-6} < .05$
$P(\beta_6) = 2.21 \times 10^{-5} < .05$

With this model, all terms are significant, so we could keep this model as our second-order approximation.

We might also consider adding the non-linear terms in one at a time, and testing for significance in the variables. These results follow,

$P(y \sim \text{Linear} + x_1^2) = .0855 > .05$
$P(y \sim \text{Linear} + x_2^2) = .00358 < .05$ ✓
$P(y \sim \text{Linear} + x_3^2) = .0388 < .05$ ✓

Consider now $y \sim \text{Linear} + x_2^2 + x_3^2$
$P(\beta_{x_2^2}) = .000414 < .05$ ✓
$P(\beta_{x_3^2}) = .00336 < .05$ ✓
Include these.

Then we may consider adding x_1^2 into the model, arriving at the significance from above. Thus, we use the model in the left margin.

(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.

Our model in (b) has 7 Beta values, => K=6 for this model, we then compute, the explained and unexplained variance. The F value is then,

$n = 17$

$n - (k+1) = 10$

$$F = \frac{\text{Exp Var}/k}{\text{unex Var}/(n-(k+1))} = 53.74276$$

with 6 and 10 degrees of freedom. So, we shall test,

$H_0: B_1 = B_2 = \dots = B_6 = 0$ vs. $H_A: \exists B_i \neq 0$ $\alpha = .05$

we find F^* with $k/n-(k+1)$ deg. Free to be $F^* = 3.22$

So,

$F_{model} \gg F^* \iff p < \alpha$

So, we may reject H_0 in favor of our alternative model.

Interaction Model Coef:

$\beta_0 = -1.320 \times 10^2$
$\beta_1 = 1.699 \times 10^0$
$\beta_2 = 7.506 \times 10^{-1}$
$\beta_3 = 8.159 \times 10^{-1}$
$\beta_4 = -2.760 \times 10^{-2}$
$\beta_5 = -2.756 \times 10^{-3}$
$\beta_6 = -2.037 \times 10^{-3}$
$\beta_7 = -6.187 \times 10^{-4}$

(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.

First, we consider the model, $y \sim \text{Second Order} + \beta_7 x_1 x_2 + \beta_8 x_2 x_3 + \beta_9 x_1 x_3$

we then perform the test, $H_0: B_i = 0$; $H_A: B_i \neq 0$ $i = 7, 8, 9$, $\alpha = .05$

with $t = \frac{b_i}{s_{b_i}}$. These results follow:

$P(\beta_7) = P(x_1 x_2 \in \text{Model}) = .3515 > \alpha$
$P(\beta_8) = P(x_2 x_3 \in \text{Model}) = .0257 < \alpha$ ✓
$P(\beta_9) = P(x_1 x_3 \in \text{Model}) = .5158 > \alpha$

By this approach, we see that the only significant interaction is that between x_2 & x_3, or the moisture & screw speed.

we now consider adding these interactions one-at-a-time and in pairs. The same test as above, and the results follow

$P(y \sim \text{Second} + x_1 x_2) = .4655 > \alpha$
$P(y \sim \text{second} + x_2 x_3) = .0173 < \alpha$ ✓
$P(y \sim \text{Second} + x_1 x_3) = .6197 > \alpha$

we find the only single interaction that is significant to be $x_2 x_3$ as before.

$y \sim \text{Second} + x_1 x_2 + x_2 x_3$	$y \sim \text{Second} + x_2 x_3 + x_1 x_3$
$P(x_1 x_2) = .3318 > \alpha$	$P(x_2 x_3) = .0224 < \alpha$ ✓
$P(x_2 x_3) = .0192 < \alpha$ ✓	$P(x_1 x_3) = .5129 > \alpha$
$y \sim \text{Second} + x_1 x_2 + x_1 x_3$	Again, the only significant interaction is $x_2 x_3$. Thus, our final model will be,
$P(x_1 x_2) = .4864 > \alpha$	
$P(x_2 x_3) = .6303 > \alpha$	

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_2 x_3$$

$R^2 = \frac{Exp}{Tot}$

$\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right)$

$V_{Tot} = \sum (y-\bar{y})^2$

$= 16.7982$

(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)?

We consider our three models as follows:

Linear (a)	Quadratic (b)	Interaction (d)
$V_{Exp} = \sum (\hat{y}-\bar{y})^2 = 5.3701$	$V_{Exp} = \sum(\hat{y}-\bar{y})^2 = 16.2929$	$V_{Exp} = \sum(\hat{y}-\bar{y})^2 = 16.53795$
$R^2 = \frac{5.3701}{16.7982} = .3197$	$R^2 = \frac{16.2929}{16.7982} = .9699$	$R^2 = \frac{16.53795}{16.7982} = .9845$
$\bar{R}^2 = .1627$	$\bar{R}^2 = .95187$	$\bar{R}^2 = .9725$

Looking at this analysis, we see that for (a) and (b) the R^2 value is very different. Because R^2 is a measure of closeness to the model and its true y-values, we see that (b) is the better of the 2 models.

(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?

First, we consider, $Range(x_1) = [25, 35]$, $Range(x_2) = [90, 130]$, $Range(x_3) = [160, 200]$

So, our test values are within the experimental region.

Next, we consider the model from (b)

$$\hat{y}(26, 100, 180) = -119.7 + 1.695(26) + .639(100) + .748(180) - .027(676) - .0027(10000) - .00204(32400)$$

$$= 11.16662$$

Our interval is then,

$$\left[\hat{y} \pm t^{10}_{.025}\, s\sqrt{1+Dist}\right] \qquad t^{10}_{.025} = 2.228, \quad s\sqrt{1+Dist} = S_{Error}$$

Then, we find

$$[10.6086, 11.7247] = I_b$$

Consider now, model (d)

$$\hat{y}(26, 100, 180) = -132 + 1.699(26) + .751(100) + .815(180) - .027(676) - .0028(10000) - .00204(32400) - .000619(18000)$$

$$= 11.16662$$

Our interval is then,

$$\left[\hat{y} \pm t^{9}_{.025}\, s\sqrt{1+Dist}\right] \qquad t^{9}_{.025} = 2.262 \qquad s\sqrt{1+Dist} = S_{Error}$$

Then,

$$[10.73803, 11.59522] = I_d$$

We note that I_b is wider than I_d, so we conclude that the interaction predictor model is worth using, since we know these intervals contain the true value at a 95% level, so smaller intervals are helpful to accuracy.

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

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

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

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

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

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.

Consider,

$$SS_T = \sum_i (y_i - \bar y)^2 = \sum (y_i^2 - 2\bar y y_i + \bar y^2) = \sum_i y_i^2 - 2\bar y \sum_i y_i + \bar y^2 \sum_i 1$$

$$= \sum_i y_i^2 - 2\bar y \sum_i y_i + n\frac{(\sum y)^2}{n^2}$$

$$= \sum_i y_i^2 - 2\bar y \sum_i y_i + \sum y (\bar y)$$

$$= \sum_i y_i^2 - \bar y \sum_i y_i$$

and

$$y^T(I - \tfrac{1}{n}J)y = \underbrace{y^T I y}_{\alpha} - \frac{1}{n}\underbrace{y^T J y}_{\beta} = \alpha - \frac{1}{n}\beta$$

$$\alpha = \sum y_i^2$$

$$\beta = \sum_i \left(y_i \sum_j y_j\right) \Rightarrow \frac{1}{n}\beta = \sum_i y_i \bar y = \bar y \sum_i y_i$$

So,

$$\alpha - \frac{1}{n}\beta = \sum_i y_i^2 - \bar y \sum y_i = SS_T$$ As desired.