Data Analysis Project

Course	STA1001
Student Name	Lu Gan
Student Number	1003119608
Submission Date	2018. 06. 14

5.4 Question 3

(a)

Based on the information introduced in this question. We can know the response variable is quality. The end of harvest is a predictor variable. There is also a dummy variable, unwanted rain.

5.10 provides the first model. It is also a full model which includes the influence of the dummy variable to both the slope and intercept of the regression line. It has an interaction term.

$$Quality = \beta_0 + \beta_1 End\ of\ Harvest + \beta_2 Rain + \beta_3 End\ of\ Harvest * Rain + e$$

From the first part of the regression output from R, we can get the coefficients and other statistics of this full model. The p value of the interaction term is 0.0120 (less than 0.05). It shows the interaction term is statistically significant.

From another aspect, if $\beta_3 = 0$, we can get the reduced model.

$$Quality = \beta_0 + \beta_1 End\ of\ Harvest + \beta_2 Rain + e$$

Then in the analysis of variance table for the full model and the reduced model, the F value for full model is 0.01203 showing evidence that $\beta_3 \neq 0$ and the interaction term is statistically significant.

(b)

(i)

If there is no unwanted rain at harvest, the dummy variable is 0 so the model becomes the following form.

$$Quality = \beta_0 + \beta_1 End\ of\ Harvest + e$$

If the quality rating is decreased by 1 point, the $\beta_1 End\ of\ Harvest$ is also decreased by 1. From the regression output, the coefficient of end of harvest is -0.03145. Thus,

the number of days of delay is

$$\frac{-1}{-0.03145} \approx 31.8$$

It takes about 32 days delay to the end of harvest to decrease the quality rating by 1 point when there is no unwanted rain at harvest.

(ii)

If there is unwanted rain at harvest, the dummy variable is 1 so the model is as follows.

$$Quality = \beta_0 + \beta_2 + (\beta_1 + \beta_3)End\ of\ Harvest + e$$

Similarly, because the estimated values of β_1 and β_3 are -0.03145 and -0.08314. The number of days of delay is

$$\frac{-1}{-0.03145 - 0.08314} \approx 8.7$$

It takes about 9 days delay to the end of harvest to decrease the quality rating by 1 point when there is unwanted rain at harvest.

6.7 Question 4

(a)

In this question, the response variable is Y=Krafft Point (KPOINT), the predictors include RA (x_1), HEAT (x_2), VTINV (x_3) and DIPINV (x_4). The first model takes each predictor into consideration.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

The model in 6.38 is not a valid model.

Figure 6.58 contains plots to show the relations among predictors. We can observe that the variable HEAT does not appear to be linearly related with other three predictors. One of the assumptions to distinguish valid model does not hold. The model may not

be a valid model.

Figure 6.60 gives plots of standardized residuals against each predictor. The plots involve RA and VTINV show curved rather than random scatter patterns. Thus, model 6.38 does not appear to be a valid model for the data.

(b)

As stated in part (a), we can learn that the model is invalid. Firstly, the plots of standardized residuals against RA and VTINV shows curved patterns but not random scatters. Besides, figure 6.59 shows the diagnostic plots from model 6.38. In normal Q-Q plot, the curve is not a straight line but a quadratic one showing evidence of non-normality of the data. The plot of standardized residuals against fitted values are indicative of nonconstant error variance. The two features with residuals for model checking are not satisfied. Thus, in this case that model fit to the data is invalid, but we cannot say anything about what part of the model is misspecified.

From the output of R, the overall F-test for model 6.38 is highly statistically significant and the only estimated regression coefficient that is not statistically significant is VTINV. And three of the variance inflation factors exceeds 5 showing that the associated regression coefficients are poorly estimated due to multicollinearity. We can drop the VTINV predictor from the first model and consider the reduced model.

If a linear model is invalid to describe the relationships between the response variable and predictors. We can consider non-linear models or make transformations of some predictors. After transformations, we can observe the matrix scatter plots again to confirm if they have collinearity and use plots of residuals against transformed predictors to see if they have random patterns. If yes, a linear model can be fitted to the transformed data.