Assignment 5-6

JARED YU

1. Do Problem 7.6, (a), (b), (d), (e), page 255 of Textbook

The carbonation level of a soft drink beverage is affected by the temperature of the product and the filler operating pressure. Twelve observations were obtained, and the resulting data are shown below (not shown here).

a. Fit a second-order polynomial.

Ans:

The dataset contains two independent variables, Temperature (x_1) and Pressure (x_2) along with the dependent variable Carbonation (y). The second-order polynomial model for such a dataset can be expressed as follows,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon. \quad (1)$$

The coefficients β_{11}, β_{22}, and β_{12} correspond with the squared and interaction terms of the polynomial model. Using R, a dataset was first constructed from the original dataset to include the three additional variables (i.e., x_1^2, x_2^2, and $x_1 x_2$) in addition to a vector of 1's (this can also be thought of as "x_0" or $\mathbf{1}_n$ which is a vector of n 1's).

From this matrix that includes the original terms in addition to the second-order terms, a multiple linear regression model was fit. This was done by calculating $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, where $\mathbf{X}$ is the new data matrix that includes the second-order terms and the ones vector while $\mathbf{y}$ is the dependent variable. The resulting vector of estimated coefficients is as follows,

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 = 3{,}025.318750 \\ \hat{\beta}_1 = -194.272901 \\ \hat{\beta}_2 = -6.050668 \\ \hat{\beta}_{11} = 3.625875 \\ \hat{\beta}_{22} = 1.154250 \\ \hat{\beta}_{12} = -1.331710 \end{bmatrix}.$$

b. Test for significance of regression.

Ans:

To test for the for the significance of regression in equation (1), we would have the following hypothesis test,

$$H_0: \beta_1 = \beta_2 = \beta_{11} = \beta_{22} = \beta_{12} = 0 \text{ vs.}$$
$$H_1: \beta_j \neq 0 \text{ or } \beta_{jk} \neq 0 \text{ for at least one } j \in \{1,2\} \text{ and } k \in \{1,2\}.$$

The above hypothesis test is checking if all the coefficients (excluding the intercept) are equal to zero. The alternative hypothesis looks slightly different from normal due to the second index, but nothing special is implied. The test statistic for this hypothesis test is as follows,

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}},$$

where

$$SS_R = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n}, SS_{Res} = \mathbf{y}'\mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n}, n = 12, k = 5.$$

Performing this test in RStudio leads to a test statistic of $F_0 = 177.1677$. The resulting p-value is 1.9832×10^{-6}, therefore the decision is to reject the null hypothesis at the $\alpha = 0.01$ significance level. The conclusion then is the evidence suggests that at least one of the coefficients $\beta_1, \beta_2, \cdots, \beta_{12}$ is nonzero.

c. (skip)
d. Does the interaction term contribute significantly to the model?

Ans:
To test if the interaction term contributes significantly to the model, it is possible to do a test on the corresponding individual regression coefficient. The hypothesis test is as follows,

$$H_0: \beta_{12} = 0 \; vs. \; H_1: \beta_{12} \neq 0.$$

The test statistic for such a hypothesis test is as follows,

$$t_0 = \frac{\hat{\beta}_{12}}{\sqrt{\hat{\sigma}^2 C_{55}}},$$

where C_{55} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ that corresponds with $\hat{\beta}_{12}$. Performing this test in RStudio leads to a test statistic of $t_0 = -1.4860$. The resulting p-value is 0.1878, therefore the decision is that we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level. The conclusion then is that there is insufficient evidence to reject the claim that $\beta_{12} = 0$. In other words, there is not enough evidence to say that the interaction term contributes significantly to the model.

e. Do the second-order terms contribute significantly to the model?

Ans:
To test if the second-order terms have a significant contribution given the first-order terms, it is possible to do a partial F test. The hypothesis test is as follows,

$$H_0: \boldsymbol{\beta}_2 = 0 \text{ vs. } H_1: \boldsymbol{\beta}_2 \neq 0,$$

where $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \hline \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \hline \beta_2 \\ \beta_{11} \\ \beta_{22} \\ \beta_{12} \end{bmatrix}$.

The test statistic for such a hypothesis test is as follows,

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/r}{MS_{Res}},$$

where $SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1)$, and r is the length of $\boldsymbol{\beta}_2$ which in this case is 3. Performing this test in RStudio leads to a test statistic of $F_0 = 5.0578$. The resulting p-value is 0.0442, therefore the decision is to reject the null hypothesis at the $\alpha = 0.05$ significance level. The conclusion then is that there is sufficient evidence to reject the claim that $\boldsymbol{\beta}_2 = 0$. In other words, there is enough evidence to say that the second-order terms contribute significantly to the model.

2. Do Problem 8.9, page 281 of Textbook

Suppose that a one-way analysis of variance involves four treatments but that a different number of observations (e.g., n_i) has been taken under each treatment. Assuming that $n_1 = 3$, $n_2 = 2$, $n_3 = 4$, and $n_4 = 3$, write down the **y** vector and **X** matrix for analyzing these data as a multiple regression model. Are any complications introduced by the unbalanced nature of these data?

Ans:
The **y** vector and $\mathbf{X}'$ matrix for the given one-way analysis of variance model would appear as follows,

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \\ y_{41} \\ y_{42} \\ y_{43} \end{bmatrix}, \qquad \mathbf{X}' = \begin{matrix} x_1 & x_2 & x_3 & x_4 \end{matrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The $\mathbf{y}$ vector is a 12×1 vector containing all the y_{ij} terms. The $\mathbf{X}'$ matrix is the 12×4 matrix where each row corresponds to a set of binary values that are 1 if the row belongs to treatment i and 0 otherwise. However, to set the $\mathbf{X}'$ matrix such that it fits the multiple linear regression model some changes need to be made. Namely, we look at three indicator variables instead of four which are defined as follows:

$$x_1 = \begin{cases} 1 & \text{if the observation is from treatment 1} \\ 0 & \text{otherwise} \end{cases}$$
$$x_2 = \begin{cases} 1 & \text{if the observation is from treatment 2} \\ 0 & \text{otherwise} \end{cases}$$
$$x_3 = \begin{cases} 1 & \text{if the observation is from treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

Then, the matrix $\mathbf{X}$ can be written as follows,

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

In the one-way analysis of variance, the formulas for the different sum of squares to perform the test factor in already the possibility of there being varying sizes for n_i. Therefore, having unbalanced group sizes alone does not cause complications, since it does not violate any of the assumptions inherent in one-way analysis of variance.

3. Do problem 8.11, (a), (b), (c), (d), page 282 of Textbook

Montgomery [2009] presents an experiment concerning the tensile strength of synthetic fiber used to make cloth for men's shirts: The strength is thought to be affected by the percentage of cotton in the fiber. The data are shown below (not shown).

a. Write down the **y** vector and **X** matrix for the corresponding regression model.

Ans:

The **y** vector can be written as follows,

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \\ y_{34} \\ y_{41} \\ y_{42} \\ y_{43} \\ y_{44} \\ y_{45} \\ y_{51} \\ y_{52} \\ y_{53} \\ y_{54} \\ y_{55} \end{bmatrix} = \begin{bmatrix} 7 \\ 7 \\ 15 \\ 11 \\ 9 \\ 12 \\ 17 \\ 12 \\ 18 \\ 18 \\ 14 \\ 18 \\ 18 \\ 19 \\ 19 \\ 19 \\ 25 \\ 22 \\ 19 \\ 23 \\ 7 \\ 10 \\ 11 \\ 15 \\ 11 \end{bmatrix}.$$

As seen in the previous some changes need to be made first before the **X** matrix can be expressed in the regression format. Let us use four indicator variables to represent the five treatment levels in **X**. These are defined as follows:

$$x_1 = \begin{cases} 1 & \text{if the observation has 15\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the observation has 20\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if the observation has 25\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if the observation has 30\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

Then, the matrix **X** can be written as follows,

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

In the above matrix **X**, the first column is a column of all 1's, which is used for the intercept term in the regression model. The second to fifth column correspond with $x_1, \cdots, x_4$ respectively.

b. Find the least-squares estimates of the model parameters.

Ans:

Similar to Problem 1 part a), a multiple linear regression model was fit to the data seen in Problem 3 part a). This leads to the following $\widehat{\boldsymbol{\beta}}$,

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 = 10.8 \\ \hat{\beta}_1 = -1 \\ \hat{\beta}_2 = 4.6 \\ \hat{\beta}_3 = 6.8 \\ \hat{\beta}_4 = 10.8 \end{bmatrix}.$$

This is based on the following model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

It should be noted that this is a first-order model and so it excludes second-order terms such as interaction terms.

c. Find a point estimate of the difference in mean strength between 15% and 25% cotton.

Ans:

$$E(y|x_1 = 1, x_2 = x_3 = x_4 = 0) - E(y|x_3 = 1, x_1 = x_2 = x_4 = 0)$$
$$\rightarrow (\beta_0 + \beta_1) - (\beta_0 + \beta_3) = \beta_1 - \beta_3$$

We can estimate this by using $\hat{\beta}_1 - \hat{\beta}_3$, which equates to $\boxed{-7.8}$. So, it is estimated that the difference in mean strength between 15% and 25% cotton is -7.8, which implies that the mean strength for 25% cotton is higher than for 15% cotton.

d. Test the hypothesis that the mean tensile strength is the same for all five cotton percentages.

Ans:

Since we are considering the mean tensile strength, we can start by looking at the single-factor fixed-effects analysis of variance and regression. The analysis-of-variance model can be seen as follows,

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, \cdots, 5, \qquad j = 1, \cdots, 25.$$

Here, μ is the grand mean, τ_i represents the effect of the ith treatment, and ε_{ij} is an $NID(0, \sigma^2)$ error component. To test the hypothesis that all means are equal, we can look at the following hypothesis test,

$$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0 \text{ vs. } H_1: \text{at least one } \tau_i \neq 0 \text{ for } i = 1, \cdots, 5.$$

A corresponding regression model can be seen as follows,

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \varepsilon_{ij}, \qquad i = 1, \cdots, 5, \qquad j = 1, \cdots, 25.$$

Here, the x_{ij} terms are the equivalent to part a), except for the addition of the j index to indicate the observation that the variable associates with. This can be seen as follows:

$$x_{1j} = \begin{cases} 1 & \text{if observation } j \text{ has 15\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2j} = \begin{cases} 1 & \text{if observation } j \text{ has 20\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{3j} = \begin{cases} 1 & \text{if observation } j \text{ has 25\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{4j} = \begin{cases} 1 & \text{if observation } j \text{ has 30\% cotton} \\ 0 & \text{otherwise} \end{cases}$$

Given that the null hypothesis is true, it would lead to the following amongst the regression model parameters,

$$\beta_0 = \mu, \qquad \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0,$$

so, the reduced model would be expressed as,

$$y_{ij} = \beta_0 + \varepsilon_{ij}.$$

To perform such a test, we can look back to what was done previously where we do a test for the significance of regression across all the coefficients. In other words, we would have the following hypothesis test,

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \text{ vs.}$$
$$H_1: \beta_j \neq 0 \text{ for at least one } j = 1, \cdots, 4.$$

The above hypothesis test is checking if all the coefficients (excluding the intercept) are equal to zero. The test statistic for this hypothesis test is as follows,

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}},$$

which has been similarly previously calculated in the first problem. The resulting test statistic is $F_0 = 14.7568$, which has a corresponding p-value of 9.1279×10^{-6}. The decision rule then is to reject the null hypothesis at the $\alpha = 0.01$ significance level. The conclusion then is that there is enough evidence to reject the claim that the mean tensile strength is the same for all five cotton percentages.

4. Problem 8.12, (a), (b), (c), page 282-283 of Textbook

Two-Way Analysis of Variance. Suppose that two different sets of treatments are of interest. Let y_{ijk} be the kth observation level i of the first treatment type and level j of the second treatment type. The two-way analysis-of-variance model is

$$y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk} \quad i = 1,2,\cdots,a, \qquad j = 1,2,\cdots,b, \qquad k = 1,2,\cdots,n$$

where τ_i is the effect of level i of the first reatment type, γ_j is the effect of level j of the second treatment type, $(\tau\gamma)_{ij}$ is an interaction effect between the two treatment types, and ε_{ijk} is an $NID(0,\sigma^2)$ random-error component.

a. For the case of $a = b = n = 2$, write down a regression model that corresponds to the two-way analysis of variance.

Ans:

Here, we have a regression model of the form,

$$y_{ijk} = \beta_0 + \alpha_i x_{ijk,1} + \beta_j x_{ijk,2} + \beta_{ij} x_{ijk,1} x_{ijk,2} + \varepsilon_{ijk}, \qquad i = 1,2, \qquad j = 1,2, \qquad k = 1,2,$$

where

$$x_{ijk,1} = \begin{cases} 1 & \text{if the observation } k \text{ is from treatment } \tau_1 \\ -1 & \text{if the observation } k \text{ is from treatment } \tau_2, \end{cases}$$

$$x_{ijk,2} = \begin{cases} 1 & \text{if the observation } k \text{ is from treatment } \gamma_1 \\ -1 & \text{if the observation } k \text{ is from treatment } \gamma_2. \end{cases}$$

This form of the model is going beyond the first order terms by including the interaction between x_1 and x_2 with β_3.

The reasoning for the 1, -1 encoding is due to some discussions within reference [1]. Although at this stage it is possible to use the binary 1, 0 combination, using this method provides for some convenience in part c). That will be elaborated on later.

b. What are the **y** vector and **X** matrix for this regression model?

Ans:

$$\mathbf{y} = \begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

c. Discuss how the regression model could be used to test the hypotheses $H_0: \tau_1 = \tau_2 = 0$ (treatment type 1 means are equal), $H_0: \gamma_1 = \gamma_2 = 0$ (treatment type 2 means are equal), and $H_0: (\tau\gamma)_{11} = (\tau\gamma)_{12} = (\tau\gamma)_{22} = 0$ (no interaction between treatment types).

Ans:

The expected value of the regression model in part a) can be seen as follows,

$$E(Y_{ijk}) = \mu_{..} + (\mu_{i.} - \mu_{..})x_{ijk,1} + (\mu_{.j} - \mu_{..})x_{ijk,2} + (\mu_{ij} - \mu_{i.} - \mu_{.1} - \mu_{..})x_{ijk,1}x_{ijk,2}.$$

Now, given the special 1, -1 encoding seen in part a), it is possible to see that the following holds:

$$E(Y_{111}) = \mu_{..} + (\mu_{1.} - \mu_{..})(1) + (\mu_{.1} - \mu_{..})(1) + (\mu_{11} - \mu_{1.} - \mu_{.1} - \mu_{..})(1)(1) = \mu_{11}$$

$$E(Y_{211}) = \mu_{..} + (\mu_{1.} - \mu_{..})(-1) + (\mu_{.1} - \mu_{..})(1) + (\mu_{11} - \mu_{1.} - \mu_{.1} - \mu_{..})(-1)(1)$$
$$= 2\mu_{.1} - \mu_{11} = 2\left(\frac{\mu_{11} + \mu_{21}}{2}\right) - \mu_{11} = \mu_{21}$$
$$E(Y_{121}) = \mu_{..} + (\mu_{1.} - \mu_{..})(1) + (\mu_{.1} - \mu_{..})(-1) + (\mu_{11} - \mu_{1.} - \mu_{.1} - \mu_{..})(1)(-1)$$
$$= 2\mu_{1.} - \mu_{11} = 2\left(\frac{\mu_{11} + \mu_{12}}{2}\right) - \mu_{11} = \mu_{12}$$
$$E(Y_{221}) = \mu_{..} + (\mu_{1.} - \mu_{..})(-1) + (\mu_{.1} - \mu_{..})(-1) + (\mu_{11} - \mu_{1.} - \mu_{.1} - \mu_{..})(-1)(-1)$$
$$= 2\mu_{..} - 2\mu_{1.} - 2\mu_{.1} + \mu_{11}$$
$$= 4\left(\frac{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}}{4}\right) - 2\left(\frac{\mu_{11} + \mu_{12}}{2}\right) - 2\left(\frac{\mu_{11} + \mu_{21}}{2}\right) + \mu_{11}$$
$$= (\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}) - \mu_{12} - \mu_{11} - \mu_{21} = \mu_{22}$$

The above is showing how for a special cause of α_1, β_1, and $(\alpha\beta)_{11}$. From these expectations, they can be related to other combinations of α_i, β_j, and $(\alpha\beta)_{ij}$ for various ijk combinations. The first hypothesis test, $\tau_1 = \tau_2 = 0$, is essentially asking us to check $\alpha_1 = \alpha_2 = 0$. The second asks us $\gamma_1 = \gamma_2 = 0$ which corresponds to $\beta_1 = \beta_2 = 0$. The last is asking us to check $(\tau\gamma)_{11} = (\tau\gamma)_{12} = (\tau\gamma)_{22} = 0$ which corresponds to $(\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{22} = 0$. So, we can see that the hypotheses tests below will follow.

To test $H_0: \tau_1 = \tau_2 = 0$, we can check if $H_0: \alpha_1 = \alpha_2 = 0$ vs. $H_1: \alpha_i \neq 0, i = 1,2$.
To test $H_0: \gamma_1 = \gamma_2 = 0$, we can check if $H_0: \beta_1 = \beta_2 = 0$ vs. $H_1: \beta_j \neq 0, j = 1,2$.
To test $H_0: (\tau\gamma)_{11} = (\tau\gamma)_{12} = (\tau\gamma)_{22} = 0$, we can check if $H_0: (\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{22} = 0$ vs. $H_1: (\alpha\beta)_{ij} \neq 0$, for at least one (ij) in $\{11, 12, 22\}$.

5. Use any math/stat software (e.g., www.numbergenerator.org/randomnumbergenerator) of your choice to find a random number generator to randomly select 12 rows of data table used in Problem 8.16 (page 283) of Textbook and then perform an analysis of your generated data and discuss your results. State the assumptions for your analysis.

Smith et al. [1992] discuss a study of the ozone layer over the Antarctic. These scientists developed a measure of the degree to which oceanic phytoplankton production is inhibited by exposure to ultraviolet radiation (UVB). The response is INHIBIT. The regressors are UVB and SURFACE, which is depth below the ocean's surface from which the sample was taken.

The data follow (not shown).

Perform an analysis of these data. Discuss your results.

Ans:
The chosen rows are 1, 2, 3, 4, 5, 7, 11, 12, 13, 14, 15, 16. The data itself includes two regressors and one response variable. Amongst the regressors, one is numeric and the other is categorical. The response variable itself is numeric. Therefore, it would make sense to try and fit a regression type model that can handle both the numeric and categorical features. This model can be seen below,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon,$$

where x_1 corresponds with UVB, x_2 corresponds with SURFACE, and ε is $NID(0, \sigma^2)$. The x_2 term is an indicator variable that can be seen as follows,

$$x_2 = \begin{cases} 1 & \text{if the observation is DEEP} \\ 0 & \text{if the observation is SURFACE.} \end{cases}$$

A first step will be to do a test for the significance of regression. The hypothesis test is as follows,

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_1: \beta_j \neq 0 \text{ for at least one } j.$$

As seen before, the test statistic for such a hypothesis test is an F_0 statistic which follows a $F_{k,n-k-1}$ distribution (here, $k = 3$ and $n = 12$). Here, $F_0 = 9.1194$, which leads to a p-value of 0.0058. The decision rule then is to reject the null hypothesis at the $\alpha = 0.01$ significance level. The conclusion then is that there is enough evidence to reject the claim that the none of the regressors contribute significantly to the model.

Next, we can check the interaction term to see if it's significant. To test if the interaction term contributes significantly to the model, it is possible to do a test on the corresponding individual regression coefficient. This has been shown before and so the same process will be repeated. The hypothesis test is as follows,

$$H_0: \beta_3 = 0 \; vs. \; H_1: \beta_3 \neq 0.$$

Performing this test in RStudio leads to a test statistic of $t_0 = 2.2396$. The resulting p-value is 0.0555, therefore the decision is that we reject the null hypothesis at the $\alpha = 0.1$ significance level. The conclusion then is that there is enough evidence to reject the claim that $\beta_3 = 0$. In other words, there is enough evidence to say that the interaction term contributes significantly to the model. (*Note: It is worth mentioning however that the hypothesis failed to reject the null at* $\alpha = 0.05$.) A possibility is to drop the interaction term and refit the model with just β_1 and β_2, however looking at the p-value it is close to the traditional 0.05 level. Therefore, such a choice will not be made in this case.

From the previous test, it seems that it could be worth keeping the interaction term in the model. We can then examine the individual test statistics for β_1 and β_2 to see how well they contribute to the model individually. Therefore, next these two hypotheses tests will be conducted,

$$H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0 \text{ and } H_0: \beta_2 = 0 \text{ vs. } H_1: \beta_2 \neq 0.$$

For β_1 and β_2 the corresponding test statistics based on the individual regression coefficients like in the previous paragraph are $t_0 = 1.2453$ and -0.2702 respectively. The corresponding p-values then are 0.2482 and 0.7939 respectively. Given these two results, the decision rule would be to fail to reject the null hypotheses that either regression coefficients are significant to the model. A conclusion then is that there is not enough evidence to reject the claim that $\beta_1 = 0$ and in the other case $\beta_2 = 0$.

These results are interesting, since they indicate that the individual coefficients alone are not making a significant impact to the model, however their interaction does have a significant impact. However, it may not make sense to drop the individual coefficients while at the same time keeping their interaction.

After performing the data analysis, it seems then that the combination of the oceanic phytoplankton production's exposure to ultraviolet radiation along with the depth below the ocean's surface has a significant impact on the inhibiting response variable. However, each of

the factors alone do not have any significant impact, and therefore they likely need to be examined together for additional insights to be made.

Reference

[1] Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin.

Code Appendix

```
library(MPV) # Load the textbook data

### Problem 7.6
orig_df <- MPV::p7.6
n <- nrow(orig_df)
ones <- rep(1, n)
y <- orig_df[,1]
x1 <- orig_df[,2]
x2 <- orig_df[,3]

# part (a) Fit a second-order polynomial.
df_a <- cbind(ones, x1, x2, x1^2, x2^2, x1*x2)

beta_hat_calc <- function(X, y) {
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  return(beta_hat)
}
H_calc <- function(X) {
  H <- X %*% solve(t(X) %*% X) %*% t(X)
  return(H)
}
y_hat_calc <- function(H, y) {
  y_hat <- H %*% y
  return(y_hat)
}
e_calc <- function(y, y_hat) {
  e <- y - y_hat
  return(e)
}

beta_hat_a <- beta_hat_calc(X = df_a, y = y)
H_a <- H_calc(X = df_a)
y_hat_a <- y_hat_calc(H = H_a, y = y)
e_a <-  e_calc(y = y, y_hat = y_hat_a)
t(e_a) %*% e_a # 2.302162

# part (b) Test for significance of regression.
SS_Res_calc <- function(y, beta_hat, X) {
  SS_Res <- (t(y) %*% y) - (t(beta_hat) %*% t(X) %*% y)
  return(SS_Res)
}
SS_T_calc <- function(y) {
  n <- length(y)
  SS_T <- (t(y) %*% y) - ((sum(y)^2) / n)
  return(SS_T)
}
SS_R_calc <- function(beta_hat, X, y) {
  n <- length(y)
  SS_R <- (t(beta_hat) %*% t(X) %*% y) - ((sum(y)^2) / n)
  return(SS_R)
}

SS_Res_a <- SS_Res_calc(beta_hat = beta_hat_a, X = df_a, y = y)
SS_T_a <- SS_T_calc(y = y)
SS_R_a <- SS_R_calc(beta_hat = beta_hat_a, X = df_a, y = y)
```

```
SS_Res_a == SS_T_a - SS_R_a
SS_Res_a; SS_T_a;  SS_R_a # 2.302142, 342.1899, 339.8878
k <- 5; n - k - 1; n - 1
MS_R_a <- SS_R_a / k # 67.97755
MS_Res_a <- SS_Res_a / (n - k - 1) # 0.3836904
F_a <- MS_R_a / MS_Res_a # 177.1677

alpha <- 0.1
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))
pf(q = F_a, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)

# part (d) Does the interaction term contribute significantly to the model?
C_mat_calc <- function(X) {
  C_mat <- solve(t(X) %*% X)
  return(C_mat)
}

sigma_hat_squared <- MS_Res_a
C_mat_a <- C_mat_calc(X = df_a)
t_a <- (beta_hat_a[6] / sqrt(sigma_hat_squared * C_mat_a[6,6]))

alpha <- 0.05
qt(p = (1 - alpha / 2), df = (n - k - 1))

# Reference: https://stats.stackexchange.com/questions/45153/manually-calculating-p-value-from-t-value-in-t-test
2 * pt(q = abs(t_a), df = (n - k - 1), lower.tail = FALSE)

# part (e) Do the second-order terms contribute significantly to the model?
df_a1 <- df_a[,c(1,2,3)]
beta_hat_red <- beta_hat_calc(X = df_a1, y = y)
SS_R_b1_a <- SS_R_calc(beta_hat = beta_hat_red, X = df_a1, y = y)
SS_R_b2_given_b1_a <- SS_R_a - SS_R_b1_a

r <- 3
F_a1 <- (SS_R_b2_given_b1_a / r) / MS_Res_a # 5.057906
alpha <- 0.05
qf(p = (1 - alpha), df1 = r, df2 = (n - k - 1))
pf(q = F_a1, df1 = r, df2 = (n - k - 1), lower.tail = FALSE)

### Problem 8.11
orig_df <- MPV::p8.11
n <- nrow(orig_df)
y <- orig_df[,1]
Xs <- orig_df[,2]
ones <- rep(1, n)

indicator_ones <- rep(1, 5)
indicator_zeros <- rep(0, 5)
x1 <- c(indicator_ones, rep(indicator_zeros, 4))
x2 <- c(rep(indicator_zeros, 1), indicator_ones, rep(indicator_zeros, 3))
x3 <- c(rep(indicator_zeros, 2), indicator_ones, rep(indicator_zeros, 2))
x4 <- c(rep(indicator_zeros, 3), indicator_ones, rep(indicator_zeros, 1))

X <- cbind(ones, x1, x2, x3, x4)

# part (b)  Find the least-squares estimates of the model parameters.
beta_hat <- beta_hat_calc(X = X, y = y)
H <- H_calc(X = X)
y_hat <- y_hat_calc(H = H, y = y)
e <-  e_calc(y = y, y_hat = y_hat)
t(e) %*% e # 161.2

# part (c)  Find a point estimate of the difference in mean strength between 15% and 25% cotton.
beta_hat[2] - beta_hat[4] # -7.8

# part (d)  Test the hypothesis that the mean tensile strength is the same for all five cotton percentage
```

```r
s.
k <- 4
SS_R_full <- SS_R_calc(beta_hat = beta_hat, X = X, y = y)
MS_R_full <- SS_R_full / k # 118.94
SS_Res <- SS_Res_calc(beta_hat = beta_hat, X = X, y = y)
MS_Res <- SS_Res / (n - k - 1) # 8.06
F_0 <- MS_R_full / MS_Res # 14.75682
pf(q = F_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)

### Problem 8.16
Location <- seq(1, 17)
INHIBIT <- c(0.00, 1.00, 6.00, 7.00, 7.00, 7.00, 9.00, 9.50, 10.00, 11.00, 12.50, 14.00, 20.00, 21.00, 25.
00, 39.00, 59.00)
UVB <- c(0.00, 0.00, 0.01, 0.01, 0.02, 0.03, 0.04, 0.01, 0.00, 0.03, 0.03, 0.01, 0.03, 0.04, 0.02, 0.03,
0.03)
SURFACE <- c('Deep', 'Deep', 'Deep', 'Surface', 'Surface', 'Surface', 'Surface', 'Deep', 'Deep', 'Surface
', 'Surface', 'Deep', 'Deep', 'Surface', 'Deep', 'Deep', 'Deep')
orig_df <- data.frame(Location, INHIBIT, UVB, SURFACE)

set.seed(1); chosen_rows <- sample(Location, 12)
chosen_rows <- sort(chosen_rows)
n <- length(chosen_rows)
ones <- rep(1, n)
df <- orig_df[chosen_rows,]
y <- df$INHIBIT
x1 <- df$UVB

# Reference: https://stackoverflow.com/questions/40780088/r-code-categorical-variable-to-1-and-0
df$SURFACE <- as.factor(df$SURFACE)
df$is_Deep <- as.numeric(df$SURFACE)
df[df$is_Deep == 2,]$is_Deep <- 0
x2 <- df$is_Deep
X <- cbind(ones, x1, x2, x1*x2)

# Test for significance of regression
k <- 3
beta_hat <- beta_hat_calc(X = X, y = y)
SS_R_full <- SS_R_calc(beta_hat = beta_hat, X = X, y = y)
MS_R_full <- SS_R_full / k
SS_Res <- SS_Res_calc(beta_hat = beta_hat, X = X, y = y)
MS_Res <- SS_Res / (n - k - 1)
F_0 <- MS_R_full / MS_Res # 9.119374
pf(q = F_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)

# Test for interaction term significance
sigma_hat_squared <- MS_Res
C_mat_a <- C_mat_calc(X = X)
t_0 <- (beta_hat[4] / sqrt(sigma_hat_squared * C_mat_a[4,4]))

alpha <- 0.05
qt(p = (1 - alpha / 2), df = (n - k - 1))

2 * pt(q = abs(t_0), df = (n - k - 1), lower.tail = FALSE)

# Test b1 and b2
t_0_b1 <- (beta_hat[2] / sqrt(sigma_hat_squared * C_mat_a[2,2]))
t_0_b2 <- (beta_hat[3] / sqrt(sigma_hat_squared * C_mat_a[3,3]))

2 * pt(q = abs(t_0_b1), df = (n - k - 1), lower.tail = FALSE)
2 * pt(q = abs(t_0_b2), df = (n - k - 1), lower.tail = FALSE)
```