Exploration 5.1: Biodiesel

With an increasing concern worldwide for environmental protection and conservation, alternative energy sources are receiving much attention. Biodiesel, derived from vegetable oils, is one such alternative fuel. Vegetable-based fuels are biodegradable, non-toxic, and significantly reduce pollution. A recent study investigated the process of using a catalyst (a chemical) to convert the triglycerides in sunflower oil to methyl ester (ME), the fuel source in biodiesel. Two factors were studied: the temperature at which the chemical reaction was performed (degrees C) and the amount of catalyst present (as a concentration, weight%).

STEP 1: Ask a research question.

Various combinations of temperature and catalyst concentration were used in the study, with the order of the reactions determined randomly. Each chemical reaction was allowed to run for 4 minutes, after which the yield (%) of methyl ester was determined. The goal of the study is to identify the combination of temperature and catalyst concentration that provides the maximum conversion of sunflower oil to methyl ester (Vicente, G. et al., 1998, "Application of the factorial design of experiments and response surface methodology to optimize biodiesel production." *International Journal of Industrial Crops and Products*, pp. 29–35.)

STEP 2: Design the study and collect data.

1. Consider the variables of *temperature* and *catalyst concentration*. Would you classify the underlying variables as quantitative or categorical? Are there any advantages to treating them either way?

2. Open the ***Biodiesel*** data file. How many unique combinations of temperature and catalyst concentration were run in this experiment? Were any of these combinations run more than once? If so, which one(s) and how many times?

With only 3–5 possible values for each variable, we could consider a *factorial design* that measures the response variable at each possible combination of the explanatory variable values. Such a design is certainly possible with quantitative variables, but these researchers instead used a "circular" design, only making repeat observations at the "center point" of the design.

3. Why is it useful that some of the treatments have more than one observation? Why do you think they chose to only replicate at the center value rather than say a few more runs at 45

°C and 1.71% concentration? (*Hint*: What key advantage will result from their design? See the graphs in the next question.)

STEP 3: Explore the data.

Even though the researchers only took measurements at a few values, both *temperature* and *concentration of catalyst* can be treated as quantitative and it would be appropriate to examine scatterplots of the different associations to look for trends.

Use statistical software to produce a ***matrix scatterplot*** summarizing the association between each explanatory variable and the response, and between the two explanatory variables.

4. Briefly describe the association between yield % and each explanatory variable. Does each association appear to be approximately linear? Do you think either explanatory-response association will be statistically significant? What temperature appears to maximize yield? What catalyst concentration appears to maximize yield?

5. Are the two explanatory variables linearly related to each other? Explain.

Another useful graph with two quantitative explanatory variables is a ***three-dimensional scatterplot***.

Use statistical software to create a 3D scatterplot for these data. That is, create a 3D scatterplot that plots temperature on the *x*-axis, catalyst concentration on the *z*-axis and yield on the *y*-axis.

6. Based on this graph, to maximize yield, what combinations of temperature and catalyst concentration appear to maximize the yield?

STEP 4: Draw Inferences beyond the data.

The two-variable (least-squares regression) model estimates the coefficients for both variables simultaneously by fitting a ***plane*** through the heights of the responses at each treatment (like placing a piece of cardboard through the middle of the heights of responses). This plane or *response surface* is fit over the entire "explanatory variable region" (all possible explanatory variable value combinations), assuming the same relationships hold across all possible explanatory variable values, not just the ones we observed in the study (also known as *interpolation*). Just like some observations are above a regression line and some are below, some of the observed responses are above the plane (positive residuals) and some are below the plane (negative residuals). The best fitting plane minimizes the sum of the squared residuals.

Key Idea: A two-variable regression model with quantitative variables fits a ***plane*** (think of a piece of cardboard) through the response variable values in order to minimize the sum of the squared residuals from that surface of the plane. Some observations will fall above the plane (positive residuals) and some will fall below the plane (negative residuals).

Use statistical software to create a multiple regression model using both temperature and concentration and then display the resulting "response surface." This is sometimes called an "additive model" where we are adding the contributions of each explanatory variable to the same model.

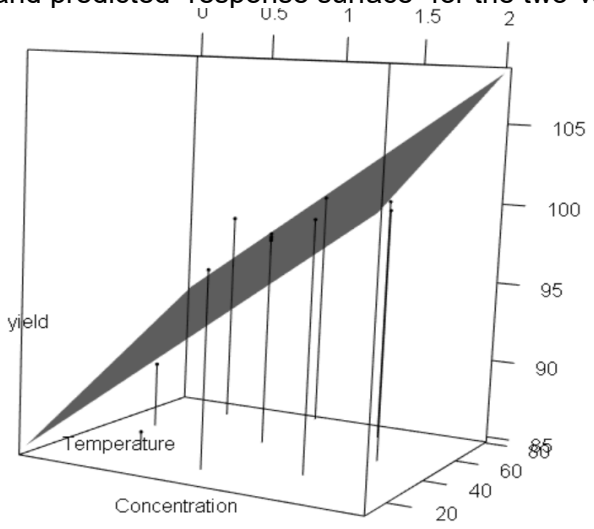
7. How do the regression coefficients in the two-variable model compare to the coefficients from the two one-variable analyses? (If you aren't sure, run those analyses.) Do the slope coefficients change? Why or why not? Are the two-variable model R^2 and *SSmodel* values the sums of the corresponding values from the one-variable analyses?

Key Idea: The lack of linear association in the explanatory variables ensures that the slope coefficients in the two-variable model will be the same as in the two one-variable prediction models and that the variation in the response variable that is explained by each explanatory variable can be separated into distinct components.

8. Does the response surface graph of the two variable model help you see the temperature and catalyst concentration that will maximize the yield? Does the response surface of predicted values seem to match up well with the observations in the 3D scatterplot? (See also Figure 1)

Include a copy of the multiple regression model output.

Figure 1: Observed data and predicted "response surface" for the two-variable linear model



9. Write out the full multiple regression model prediction equation that you have found, using good statistical notation. Then use this equation to determine the prediction equation between *yield* and *concentration* when temperature = 20 ^{0}C. Also find the prediction equation between *yield* and *concentration* when temperature is 60 ^{0}C? (*Hint*: Write out the two (simplified) equations.)

Full equation:

Temp = 20^0:

Temp = 60^0:

10. Based on what you saw in the previous question, write an interpretation of the slope coefficient of *concentration* in the multiple regression model. (*Hint*: What do you say about *temperature*?)

Key idea: A model where we predict a response variable using two explanatory variables fits a "parallel lines" model where, regardless of the value of x_2, each equation between y and x_1 has the same slope. In other words, this model does not allow for any interaction between the two explanatory variables.

Variable Importance

11. Which variable, temperature or concentration, seems more "important" in predicting yield? How are you deciding?

We normally think of the slope coefficient as telling us about "impact"—how quickly the response variable changes with a one-unit increase in the explanatory variable. However, it is problematic to compare a one-degree temperature increase to a one percentage point change in catalyst concentration because these variables are on much different scales. The range of temperatures in the study is 50°C, whereas concentrations range from 0.20 to 1.71%. One way to answer the question of which variable is more important in predicting yield is to compare the R^2 values. Another approach, which has some additional benefits, is to first ***standardize*** each variable, and then use these standardized variables in the two-variable model instead.

Definition: To ***standardize*** a variable we subtract the mean and divide by the standard deviation.

Use a spreadsheet package and/or statistical software to standardize both explanatory variables and fit a new multiple regression model with these standardized variables.

12. How, if at all, has the model changed when using the standardized variables? How has the relationship, if at all, between temperature and concentration changed? Provide a one-sentence interpretation of the intercept. Would you consider the intercept meaningful in this context? Provide a one-sentence interpretation of the coefficient of temperature. Which variable, temperature or concentration, seems to have a larger impact on the yield in terms of standard deviation units?

Key Idea: Some researchers would feel more comfortable comparing slopes when both explanatory variables are in standard deviation units. This is assuming that a one standard deviation change is of comparable cost or interest for the two explanatory variables.

Key Idea: ***Standardizing*** quantitative variables has several advantages including providing a more meaningful intercept (guaranteed to correspond to the means of the explanatory variables rather than involving extrapolation or nonsensical values) and providing more comparability of the magnitude of the slope coefficients.

Validity Conditions

Key Idea: The ***validity conditions*** for a two-variable model with quantitative variables are the same as we have seen before:

- Each explanatory variable is linearly related to the response variable,
- the observations are independent,
- the residuals follow a symmetric, bell-shaped distribution,
- the variability in the residuals is constant for each combination of explanatory variable values.

Include a copy of the Residuals vs. Predicted plot. If your software allows, color code the residuals vs. predicted plot by one of the explanatory variables.

[Type here]

13. Does the *Residuals* vs. *Predicted* plot indicate any problems with our model? If so, describe the pattern revealed in this plot.

A simple way to create a more "flexible" model is to include interactions. To allow for an interaction between two quantitative explanatory variables, we multiply the two explanatory variable columns together and include that product as a variable in the model.

In your data set, create a new variable (i.e., column) that is the product of the temperature and catalyst concentration variables and examine the scatterplot of this new variable vs. *temperature.*

14. From the graph, describe the behavior of the association between the interaction variable and the temperature variable.

Even with a balanced design such as this one, it is highly likely that the product term, representing the interaction will be linearly related to one or both of the variables multiplied together to form the interaction. It turns out one way to control for this is to use the standardized variables instead.

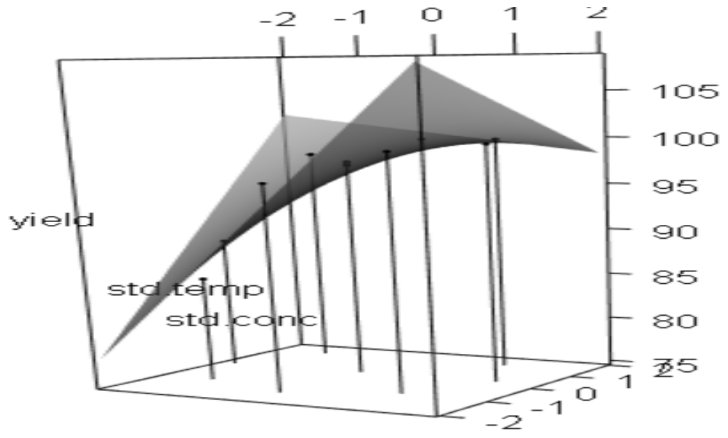
In your data set, create a new variable (i.e., column) that is the product of the standardized temperature and standardized catalyst concentration variables and examine the scatterplot of this variable vs. standardized *temperature and standardized catalyst concentration.*

15. Is the standardized interaction variable linearly related to either of the standardized explanatory variables?

Key Idea: Another advantage to standardizing explanatory variables is that the interaction term (product of the standardized variables) will not be linearly related to either standardized variable involved in the interaction.

Use statistical software to create a multiple regression model using the standardized variables and the interaction of the standardized variables and then display the resulting response surface. Compare to Figure 2.

Figure 2: The predicted response surface from the multiple regression with interaction model using the standardized variables



Key Idea: An interaction between two quantitative variables allows "bends" in the response surface of predicted values. The interaction can sometimes account for any curvature seen in the residual plot from the additive model (without the interaction term).

16. How many degrees of freedom are used to include the interaction term in the model? How do the sum of square values for the two standardized explanatory variables change between this model and the one with the standardized variables but no interaction? (Why?) Is the interaction statistically significant?

17. Suppose we hold temperature at a standardized value of 0 (i.e., 45^0C) what is the resulting prediction equation between *yield* and standardized *concentration*? (Show your work.) Repeat holding standardized temperature at 1 (i.e., about 28^0C). How does the slope between yield and standardized concentration change between the two regression lines?

Full equation:

Standardized Temp = 0:

Standardized Temp = 1:

Change in slope:

If your software allows, display the fitted model of *yield* vs. *concentration* for these two values of *temperature*.

Another way to visualize this interaction is to use an interaction plot where we choose two values (one low, one high) for one of the variables and look at the conditional equations.

STEP 5: Formulate conclusions.

18. Based on the sign of the interaction coefficient (and the surface plot and the equations), summarize the nature of the interaction in these data.

Remember that the goal of this study is to maximize the conversion of sunflower oil to methyl ester. Once the response surface is fit across all the explanatory variable combinations, the optimal combination of temperature and concentration may not be the same setting you would find if you optimized each variable separately, and may not even be one of the combinations that was actually observed in the experiment.

19. Based on this two-variable interaction model, what combination(s) of (standardized) temperature and concentration would you suggest using for the chemical process in order to maximize the yield of methyl ester? What combinations of (unstandardized) temperature and concentration do these correspond to?

Use software to calculate the 95% prediction interval for the amount of methyl ester yield at this combination of temperature and concentrations.

20. Interpret the interval in context.

STEP 6: Look back and ahead.

21. Examine the Residuals vs. Predicted Plot. Has including the interaction sufficiently dealt with the curvature in the data?

22. What next steps would you recommend in this research?