

When asked to explain something, or to provide an interpretation for a quantity, provide an explanation that could be understood by someone who does not have formal training in statistical methods. Keep your explanations brief.

1. Choosing how many and which model terms to put on the right-hand side of a regression equation is a somewhat subjective process. Choosing the model with the largest  $R^2$  is not a good idea because adding a variable to a model will never decrease  $R^2$ . Why is it not always a good idea to choose the model with the smallest value of  $S$ ?
  
2. Consider the antique clock auction data discussed in class. In this example the highest bid price for the clocks (the response) was to be modeled as a function of the clock's age and the number of bidders (the  $X$ s). In a two-variable problem like this we might be concerned with both "correlation between the  $X$ s" and "interaction between the  $X$ s." They are *not* the same.
  - (a) Draw a simple graph or plot to illustrate the meaning of "correlation between the  $X$ s" in the above problem. Make sure to label your axes.
  
  - (b) Draw a simple graph or plot to illustrate the meaning of "interaction between the  $X$ s" in the above problem. Make sure to label your axes.
  
3. In some applications it is possible to control the  $X$ s. This allows the use of "designed experiments." A popular method is to use a rectangular array (e.g., the square in the point of sale/media advertising example. What some advantages of such a design?

4. A study was conducted to build a model that can be used to predict  $Y$ , the number of paid visitors to a pool-water park in a resort district of New Hampshire. Let

$$X_1 = \begin{cases} 1 & \text{for a weekday} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{for a sunny day} \\ 0 & \text{otherwise} \end{cases} \quad X_3 = \text{Degrees F}$$

The following three models were fit to the data from 30 consecutive days in the summer of 1992:

$$\text{Model 1 } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

$$\text{Model 2 } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \epsilon$$

$$\text{Model 3 } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_1 X_2 + \epsilon$$

giving the following results

Model	$SS_{YY}$	$SSE$
1	1762646	77212
2	1762646	76727
3	1762646	76292

- (a) Briefly explain why the  $SS_{YY}$  values are the same for all three of these models and why the  $SSE$  values differ.

- (b) In Model 2, what is the expected increase in paid visitors for an additional degree F in temperature?

- (c) Test the null hypothesis that  $\beta_4 = 0$  in Model 2, using  $\alpha = .05$ . In *practical* terms, what is your conclusion?

- (d) Briefly explain practical interpretation of Model 2, relative to Model 1.

(e) Briefly explain practical interpretation of Model 3, relative to Model 1.

(f) Briefly explain the practical interpretation of  $\beta_4$  in Model 2.

(g) For Model 2, Briefly explain whether you would expect  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ , and  $\beta_4$  to be positive or negative and, for each, give the reason why.

5. Under the assumptions of the standard regression model,  $\sigma$ , the standard deviation of the residuals, does not depend on the  $X$ s. We compute the “Error” degrees of freedom as the total number of observations minus the number of parameters in the regression surface that need to be estimated. Briefly explain why this might be considered “the effective sample size for estimating  $\sigma$ ?”
6. Consider the computer software training example from Lab 7, where the effect of Training Method and Previous Experience on time to complete a set of tasks was investigated. The presence of interaction between Training Method and Previous Experience could have important implications for decision making. Explain briefly.

7. Data were collected on the finished floor size of homes (in square feet), the unfinished floor size (e.g., basement size, also in square feet), and the sales price, in thousands of dollars. The correlation matrix for these data, based on a sample of 25 homes, was:

	Price	Finished	Unfinished
Price	1	.85	.75
Finished	.85	1	.91
Unfinished	.75	.91	1

The following three models were fit to the data:

$$\text{Model 1} \quad \text{Price} = \beta_0 + \beta_1 \text{Finished} + \beta_2 \text{Unfinished} + \epsilon$$

$$\text{Model 2} \quad \text{Price} = \beta_0 + \beta_1 \text{Finished} + \epsilon$$

$$\text{Model 3} \quad \text{Price} = \beta_0 + \beta_2 \text{Unfinished} + \epsilon$$

- (a) The  $t$ -ratios for  $\beta_1$  and  $\beta_2$  in Model 1 were both less than 1. The  $t$ -ratios for  $\beta_1$  and  $\beta_2$  in Models 2 and 3, respectively, were both statistically significant. Give an intuitive explanation for this.
- (b) Could regression analysis computer output for models 1, 2, or 3 for this problem be used for predicting house prices? Briefly explain the conditions under which such predictions can be made.
- (c) Could regression analysis computer output for models 1, 2, or 3 for this problem be used to estimate the price of an additional foot of unfinished floor space? Briefly explain why or why not?

8. Consider the following two alternative models relating sales to media and point of sale advertising  $X_1$  and  $X_2$ , respectively. All variables are in units of thousands of dollars.

$$\text{Model 1 } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$\text{Model 2 } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \epsilon$$

- (a) In Model 1, what is the expected increase in  $Y$  for an additional dollar of Media advertising?
- (b) In Model 2, what is the expected increase in  $Y$  for an additional dollar of Media advertising?
- (c) Briefly explain, and give an equation, to show how you could use regression analysis computer output to test for the need for the more complicated Model 2.
- (d) Suppose that the evidence in support of Model 2 was significant at the 10% level of significance, but not at the 5% level of significance. What would be a good argument for presenting the results of the study in terms of Model 1?



