

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.

1. How many different levels of x are needed to fit the model $E(y) = \beta_0 + \beta_1x + \beta_2x^2$? How large of a sample is needed to, in addition, estimate σ^2 ? Draw a picture to help you explain.

2. Multicollinearity can make interpretation of regression coefficients difficult.

(a) Briefly explain *why* multicollinearity can make interpretation of regression coefficients difficult.

(b) Briefly explain how proper use of a designed experiment can reduce or eliminate multicollinearity.

(c) Briefly explain why it is *not* always possible to use a designed experiment to avoid multicollinearity.

(d) *List* some of the symptoms of multicollinearity.

-
-
-

3. The National Science Foundation conducted a survey to evaluate the salaries of Men and Women working in professional scientific jobs. The survey included twenty randomly selected individuals in each category (for a total sample size of 160 scientists). A summary of the data, given below, shows the mean in each category.

Discipline	Gender	
	Female	Male
Physics	42.1	49.1
Mathematics	36.2	43.3
Biology	34.3	41.4
Medicine	36.6	46.9

The following two models were fit to the data:

Model	Terms	SSE
1	Gender and Discipline	2230
2	Gender and Discipline with interaction $\text{Discipline} \times \text{Gender}$	2150

- (a) Plot the data in a manner that will allow you to assess whether there might be interaction between gender and discipline. What do you conclude from the plot?
- (b) What would be the *practical* interpretation if there were interaction in this example?
- (c) Write down the dummy variable regression model for Model 1 and Model 2. Carefully define all dummy variables.
- (d) Do a statistical test to see if there is strong evidence of interaction or not. Use $\alpha = .05$. State your conclusion.

4. Data were collected to build a model that could be used to predict the number of paying visitors to a pool/water park in a resort district in New Hampshire. The data were available for 30 consecutive days during the summer of 1997. Let

$$x_1 = \begin{cases} 1 & \text{Saturday or Sunday} \\ 0 & \text{weekday} \end{cases} \quad x_2 = \begin{cases} 1 & \text{sunny day} \\ 0 & \text{not a sunny day} \end{cases} \quad x_3 = \text{Temperature in degrees F}$$

The following three models were fit to the data

$$\text{Model 1: } E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \quad \text{SSE} = 77212$$

$$\text{Model 2: } E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3, \quad \text{SSE} = 76729$$

$$\text{Model 2: } E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_1 x_2, \quad \text{SSE} = 76292$$

$$SS_{yy} = \sum_1^{32} (y_i - \bar{y}_i)^2 = 1762646 \text{ for all of the models.}$$

- (a) Provide a *practical* description of each of these models (i.e., find an explanation that describes the differences in practical terms with out using words like “interaction”).

- Model 1:

- Model 2:

- Model 3:

- (b) Test the null hypothesis that $\beta_4 = 0$ versus the alternative $\beta_4 \neq 0$ in Model 2, using $\alpha = .05$.

- (c) Explain the *practical* interpretation of the hypothesis $\beta_4 \neq 0$ in Model 2.

- (d) For Model 2 explain whether you would expect the parameters β_1 , β_2 , β_3 , and β_4 to be positive or negative. For each, briefly explain why.

- β_1 :

- β_2 :

- β_3 :

- β_4 :

- (e) For Model 2, the predicted number of paying customers on a sunny Wednesday is 195. Find an approximate 95% prediction interval for the number of paying customers on a sunny Wednesday.