

You must show all of your work

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. Complete concise explanations are preferred.

1. An industrial psychologist conducted an experiment to study the relationship between worker productivity and pay incentive for two different manufacturing plants: one that was a union plant and another where there was no union. The manufacturing plants manufacture castings. Productivity, in terms of number of castings produced, was the response variable (Y). Nine workers were selected at random from each plant, and divided up equally to receive \$0.20, \$0.30, and \$0.40 bonus for each casting produced. Let the bonus amount be denoted by x_2 . Also, let $x_1 = 0$ for the union plant and $x_1 = 1$ for the non-union plant.

Two models were fit to the data

$$\text{Model 1: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

$$\text{Model 2: } y = \beta_0 + \beta_1 x_1 + \epsilon$$

where $\epsilon \sim \text{nid}(0, \sigma^2)$. The following tables gives a summary of the results:

	Model 1	Model 2
β_0	1365	1389
β_1	6.21	6.23
β_2	47.8	
β_3	.033	
$SSE = \sum_{i=1}^{18} (y_i - \hat{y}_i)^2$	23349	34056
$SS_{yy} = \sum_{i=1}^{18} (y_i - \bar{y})^2$	80681	80681

- (a) Construct an analysis of variance table for Model 1.

- (b) Use a partial F test to compare Model 1 and Model 2. Use $\alpha = .05$. What is your conclusion?

- (c) What is the *practical interpretation* or conclusion of the hypothesis tested in part 1b?

- (d) What is an estimate of σ for model 1? Explain the *practical interpretation* of σ .
- (e) How many degrees of freedom are associated with the estimate in part 1d? Explain the *practical interpretation* of degrees of freedom, in the context of this problem.
- (f) Explain how the results of this experiment might have differed if 90 individuals had been selected from each plant, instead of only 9 individuals.

2. An analyst has fit a simple regression between x and y for a data set with 100 observations using the model

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

where $\epsilon \sim \text{nid}(0, \sigma^2)$. The value of $R^2 = .99$, all plots suggested an excellent fit for the straight-line model, and testing the null hypothesis $H_0 : \beta_1 = 0$ resulted in a t -ratio of 97.3.

- (a) Approximately, what is the p -value for the hypothesis $H_0 : \beta_1 = 0$? *Briefly explain.*
- (b) In spite of the large t -ratio, explain why *might* it be a mistake to claim that x causes y ?
- (c) What advice would you have for the analyst if he/she really wanted to test if x causes y ?

3. Explain the difference between influence and leverage. Draw a simple picture to help explain.

4. The inventory manager for a company has collected data on inventory for a line of home furnishings during each month for the past 5 years (total of 60 observations), in an effort to build a forecasting model that might allow cost savings through better control of inventory levels. Explanatory variables used included housing starts for the state in the past two months (x_1 and x_2) and local mortgage interest rates in the past two months (x_3 and x_4). A multiple regression model was fit to the data. The Durbin-Watson statistic for the computer output was 1.37. What can you conclude from this?

5. A study was conducted to investigate the relationship between sales on a given day for a department store and the amount spent on radio advertising on the previous evening. The advertising values were varied according to a pre-specified randomized pattern over a period of 50 days. A simple regression model was fit to these data. A careful examination of the residuals revealed that in most cases (about 66%) a positive residual was followed by another positive residual and that negative residuals were, in most cases (again about 66%) followed by another negative residual.
- (a) Which of the standard regression assumptions was probably violated in this example? Explain.
 - (b) What might have caused such a violation of the assumption?
 - (c) If all of the usual assumptions of simple regression hold (note that they probably do *not* in the example described above), what would be the probability that the residual day t will be positive, given that the residual on day $t - 1$ is positive?
 - (d) How can one check to see if the residuals described above could have been caused by randomness alone instead of violation of some regression assumption?
6. Outliers are of concern to analysts. If an outlier is caused by a data-recording error, the error should be fixed if possible. Otherwise it may be best to discard the observation. In other cases, outliers can teach us something.
- (a) Explain how one can determine whether an outlier is influential or not.
 - (b) Give an example of when an outlier, known not to be an error, should still be omitted from the model fitting process.
 - (c) Explain why an outlier might be extremely important to an analyst.
7. You have been given a computer file with relevant data and have been asked to find a model relating the selling price of houses Y to the assessed value of the land (X_1) and the assessed value of the improvements on the land (X_2). What are the first two steps that you would take in the data analysis/model development task?