

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.

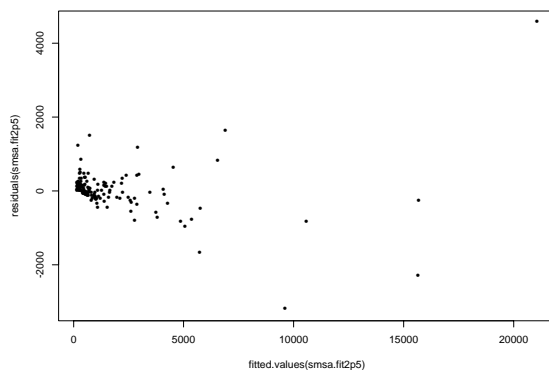
1. Why it is dangerous to use a regression model to predict a value of the response that is outside of the range of the explanatory variables. Is it ever reasonable to make such a prediction? Explain.
  
2. In a regression model, one is often interested in whether there is interaction or not. Assume that the response of construction job profitability is related to two explanatory variables: the engineer in charge ( $x_1 = 0$  for Jones and  $x_1 = 1$  for Smith) and the state in which the job was completed ( $x_2 = 0$  for Illinois and  $x_2 = 1$  for Iowa).
  - (a) Write down the interaction and no-interaction models and briefly explain the *practical* difference between these two different models.
  
  - (b) Draw two pictures that might help explain the difference in part (a).
  
  - (c) Briefly explain the *practical* interpretation of the parameters in the no-interaction model.
  
  - (d) If the construction company finds that there is strong interaction between  $x_1$  and  $x_2$ , how might they use this information to improve job profitability? Give two possibilities.

3. Suppose that the *interaction model* in problem ?? needs to be extended to allow for job size ( $x_3$ ) as an explanatory variable and, because of curvature in the data an  $x_3^2$  term is also to be included.

(a) Briefly explain the difference in the model's *practical* interpretation if all possible interaction terms with  $x_3$  and  $x_3^2$  are included versus no interaction terms with  $x_3$  and  $x_3^2$  are included. To aid in making your explanation, write down some appropriate models.

(b) Make a *list* of the steps that you would follow to test to see if the data provide evidence that the interaction terms with  $x_3$  and  $x_3^2$  are needed or not. The list should contain an appropriate formula that would be used for the test and explain how to use this formula.

4. The following figure is a plot of residuals versus fitted values for the model `smsa.fit2p5 <- lm(Physicians ~ Population )` for the SMSA data. What does this plot tell you?



5. 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.
  
  - (e) Briefly explain why it is possible that serious multicollinearity will not be detected just by looking at a pairs plot and/or a correlation matrix. What is a good statistic for detecting multicollinearity?
6. Briefly explain the difference between leverage and influence. Use pictures to aid in your explanation.

7. A regression model has been developed to predict Sales (measured in units of thousands of dollars). The data available for fitting this model were collected on 50 successive weeks. In order to simplify the model, a square root transformation was used. The resulting model was

$$y^* = \sqrt{\text{sales}} = 5 + 2x + \epsilon$$

where  $x$  is the amount of money spent on advertising in the previous week (units of thousands of dollars). In terms of  $\sqrt{\text{sales}}$ , the variability in the data from the fitted model appears to be constant. A prediction for sales for  $x = 2$  is desired. The computer output gave  $S = 1.1$  and for  $x = 2$ ,  $S_{\hat{y}^*} = .05$ . Find a 95% prediction interval for sales for the next week.

8. Refer to problem ???. There was concern that because the data were collected over time, that the independence assumption might not be valid.

(a) Explain why the independence assumption might not be valid in a problem like this.

(b) If the Durbin-Watson statistic for this problem were 1.25, what would you conclude about the assumption of independence?