Post-Live Session Homework: Unit 11

Assignments

On this upload page you will submit unit 11 homework for grading.

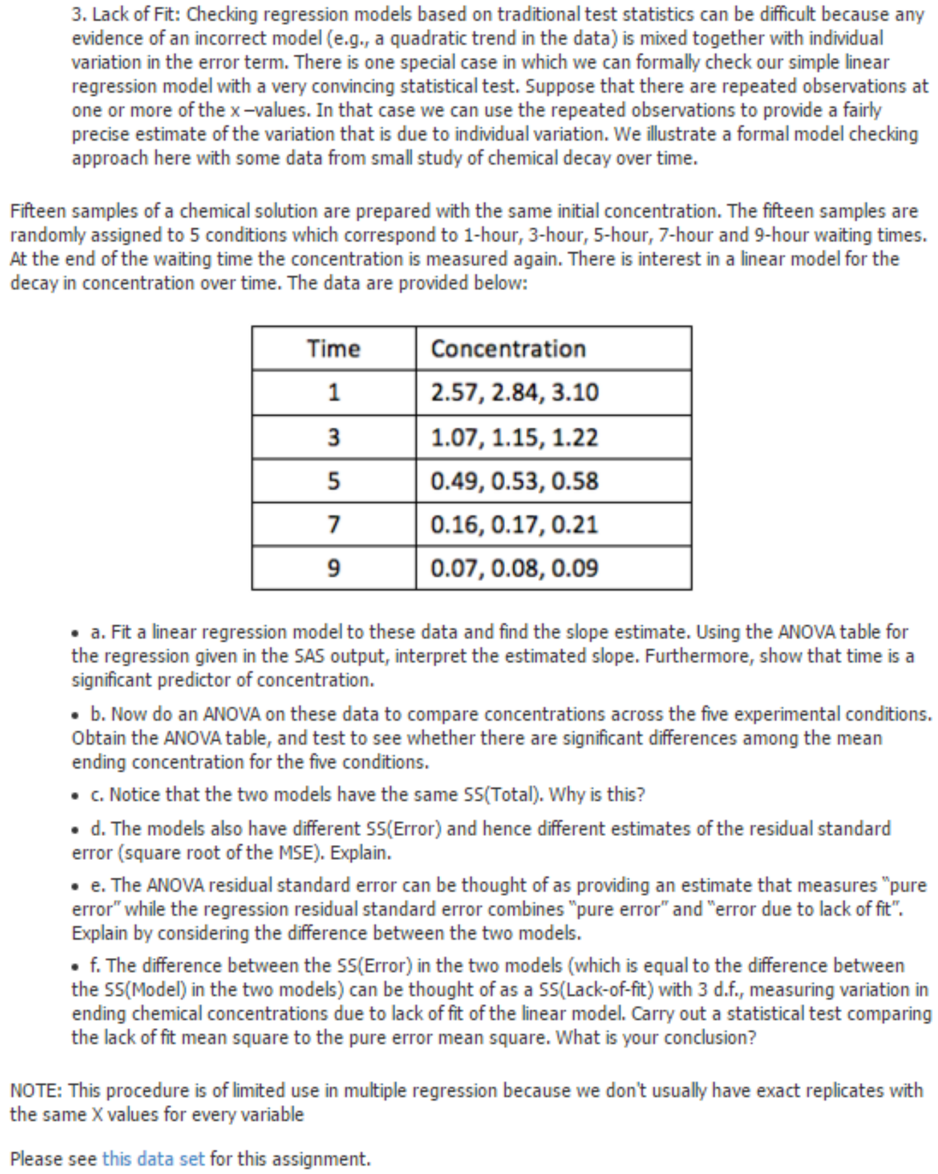
Regression Diagnostics and Model Refinement

1. What determines the crime rate within a state? The Statistical Abstract of the United States contains data on the crime rate in each state plus additional variables (see file crime2005.csv). The variables for this data set are VI = violent crime rate (number of violent crimes per 100,000 population), VI2 = violent crime rate (number of violent crimes per 10,000 population), MU = murder rate, ME = percent in metropolitan areas, WH = percent white, HS = percent high school graduates, PO = percent below the poverty level. We are going to examine the relationship between VI and ME, the percent of individuals living in metropolitan areas.

- a. Obtain a scatterplot of the data, making sure to place the correct variable on the correct axis (you are not told which one is which for a reason!) Describe the relationship in the plot (form, direction, and strength).
- b. What is the linear regression equation for the data?
- c. Is the slope significantly different from 0? Use output to support your answer.
- d. Give a 95% confidence interval for the slope.

2. Diagnostics: With the exception of the scatter plot, we did not do any exploratory data analysis on this data set. In order to determine whether our hypothesis tests and predictions can be trusted, we need to examine regression diagnostics. In regression (as in most models we will use from this point) assumptions are examined by looking at the residuals for the model.

- a. What are the assumptions of the linear regression model?
- b. What should a plot of the residuals versus the predicted values look like if the assumptions of linearity and constant variance are not violated?
- c. What should a QQplot for the residual look like if the residuals are Normally distributed?
- d. What would we see in a plot of the residuals versus the predicted values if the assumption of constant variance were violated?
- e. Are there any high leverage values in this data set? If so, which cases? Explain your reasoning.



3. Lack of Fit: Checking regression models based on traditional test statistics can be difficult because any evidence of an incorrect model (e.g., a quadratic trend in the data) is mixed together with individual variation in the error term. There is one special case in which we can formally check our simple linear regression model with a very convincing statistical test. Suppose that there are repeated observations at one or more of the x –values. In that case we can use the repeated observations to provide a fairly precise estimate of the variation that is due to individual variation. We illustrate a formal model checking approach here with some data from small study of chemical decay over time.

Fifteen samples of a chemical solution are prepared with the same initial concentration. The fifteen samples are randomly assigned to 5 conditions which correspond to 1-hour, 3-hour, 5-hour, 7-hour and 9-hour waiting times. At the end of the waiting time the concentration is measured again. There is interest in a linear model for the decay in concentration over time. The data are provided below:

Time	Concentration
1	2.57, 2.84, 3.10
3	1.07, 1.15, 1.22
5	0.49, 0.53, 0.58
7	0.16, 0.17, 0.21
9	0.07, 0.08, 0.09

- a. Fit a linear regression model to these data and find the slope estimate. Using the ANOVA table for the regression given in the SAS output, interpret the estimated slope. Furthermore, show that time is a significant predictor of concentration.
- b. Now do an ANOVA on these data to compare concentrations across the five experimental conditions. Obtain the ANOVA table, and test to see whether there are significant differences among the mean ending concentration for the five conditions.
- c. Notice that the two models have the same SS(Total). Why is this?
- d. The models also have different SS(Error) and hence different estimates of the residual standard error (square root of the MSE). Explain.
- e. The ANOVA residual standard error can be thought of as providing an estimate that measures "pure error" while the regression residual standard error combines "pure error" and "error due to lack of fit". Explain by considering the difference between the two models.
- f. The difference between the SS(Error) in the two models (which is equal to the difference between the SS(Model) in the two models) can be thought of as a SS(Lack-of-fit) with 3 d.f., measuring variation in ending chemical concentrations due to lack of fit of the linear model. Carry out a statistical test comparing the lack of fit mean square to the pure error mean square. What is your conclusion?

NOTE: This procedure is of limited use in multiple regression because we don't usually have exact replicates with the same X values for every variable

Please see this data set for this assignment.