Homework Assignment 5

Group homework (up to four members per group)
due in class on Tuesday, 03/18/2014

STA 371G, Statistics and Modeling, Spring 2014

Problem 1 (5 points)

Suppose we are modeling house price as depending on house size, the number of bedrooms in the house and the number of bathrooms in the house. Price is measured in thousands of dollars and size is measured in thousands of square feet.

Suppose our model is:

$$P = 20 + 50\,\text{size} + 10\,\text{nbed} + 15\,\text{nbath} + \epsilon, \quad \epsilon \sim N(0, 10^2).$$

(a) Suppose you know that a house has size =1.6, nbed = 3, and nbath =2.

What is the distribution of its price given the values for size, nbed, and nbath.

(hint: it is normal with mean = ?? and variance = ??)

(b) Given the values for the explanatory variables from part (a), give the 95% predictive interval for the price of the house.

(c) Suppose you know that a house has size =2.6, nbed = 4, and nbath =3. Give the 95% predictive interval for the price of the house.

(d) In our model the slope for the variable nbath is 15. What are the units of this number?

(e) What are the units of the intercept 20? What are the units of the the error standard deviation 10?

Problem 2 (5 points)

The data for this question is in the file **Profits.csv**, which can be found in the course website.

There are 18 observations.
Each observation corresponds to a project developed by a firm.
y = Profit: profit on the project in thousands of dollars.
x1= RD: expenditure on research and development for the project in thousands of dollars.
x2=Risk: a measure of risk assigned to the project at the outset.

We want to see how profit on a project relates to research and development expenditure and "risk".

(a) Plot profit vs. each of the two x variables. That is, do two plots y vs. x1 and y vs x2. You can't really understand the full three-dimensional relationship from these two plots, but it is still a good idea to look at them. Does it seem like the y is related to the x's?

(b) Suppose all you knew was risk=7. Run the simple linear regression of profit on risk and get the 68% plug-in predictive interval for profit.

(c) Suppose a project has risk=7 and research and development = 76. Give the 95% plug-in predictive interval for the profit on the project. Compare that to the correct, predictive interval (using the predict function in R).

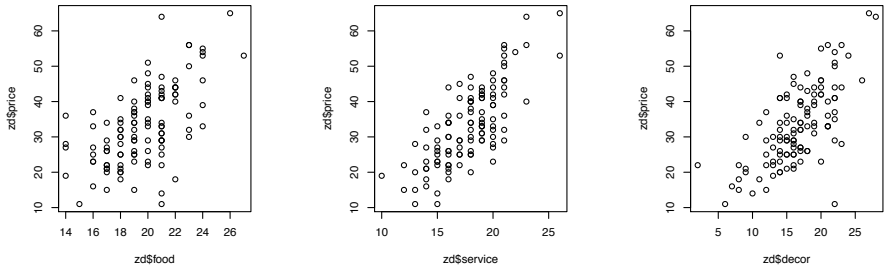
(d) How does the size of your interval in (c) compare with the size of your interval in (b)? What does this tell us about our variables?

Problem 3 (10 points)

The data for this question is in the file **zagat.xls**, which can be found in the course website. The data is from the Zagat restaurant guide. There are 114 observations and each observation corresponds to a restaurant.
There are 4 variables:
price: the price of a typical meal
food: the zagat rating for the quality of food.
service: the zagat rating for the quality of service.
decor: the zagat rating for the quality of the decor.

We want to see how the price of a meal relates the quality characteristics of the restaurant experience as measured by the variables food, service, and decor.

(a) Plot price vs. each of the three x's (shown below). Does it seem like our y (price) is related to the x's (food, service, and decor) ?



(b) Suppose a restaurant has food = 18, service=14, and decor=16. Run the regression of price on food, decor, and service and give the 95% predictive interval for the price of a meal.

(c) What is the interpretation of the coefficient estimate for the explanatory variable food in the multiple regression from part (b) ?

(d) Suppose you were to regress price on the one variable food in a simple linear regression? What would be the interpretation of the slope? Plot food vs. service. Is there a relationship? Does it make sense? What is your prediction for how the estimated

coefficient for the variable food in the regression of price on food will compare to the estimated coefficient for food in the regression of price on food, service, and decor? Run the simple linear regression of price on food and see if you are right! Why are the coefficients different in the two regressions?

(e) Suppose I asked you to use the multiple regression results to predict the price of a meal at a restaurant with food = 20, service = 3, and decor =17. How would you feel about it?

Problem 4: Baseball (5 points)

Using our baseball data (**RunsPerGame.xls**), regress R/G on a binary variable for league membership (League = 0 if National and League = 1 if American) and OBP.

$$R/G = \beta_0 + \beta_1 League + \beta_2 OBP + \epsilon$$

(a) Based on the model assumptions, what is the expected value of R/G given OBP for teams in the American League? How about the National League?

(b) Interpret β_0, β_1 and β_2.

(c) After running the regression and obtaining the results, can you conclude with 95% probability that the marginal effect of OBP on R/G (after taking into account the League effect) is positive?

(d) Test the hypothesis that $\beta_1 = 0$ (with 99% probability). What do you conclude?

Problem 5 (10 points)

Read the case "Orion Bus Industries: Contract Bidding Strategy" in the course packet. Orion Bus Industries wants to develop a method for determining how to bid on specific bus contracts to maximize expected profits. In order to do this, it needs to develop a model of winning bids that takes into account such factors as the number of buses in the contract, the estimated cost of the buses and the type of bus (e.g. length, type of fuel used, etc.). The data set is available in the course website. This data set only includes the bus contracts from Exhibit 1 in the case where Orion did not win the contract. This eliminates 28 of the 69 observations and leaves a sample of size n = 41 observations.

(a) Run a regression of $WinningBid$ against $NumberOfBusesInContract$, $OrionsEstimatedCost$, $Length$, $Diesel$ and $HighFloor$, ie, the following regression model:

$$WinningBid_i = \beta_0 + \beta_1 NumberOfBusesInContract_i + \beta_2 OrionsEstimatedCost_i + \beta_3 Length_i + \beta_4 Diesel_i + \beta_5 HighFloor_i + \epsilon_i$$

What is the estimated regression model? How would you interpret the estimated coefficient associated with the dummy variable Diesel?

(b) What is the estimate of σ^2 in the model in part (a)?

The city of Louisville, Kentucky is putting out a contract for bid for five 30-foot, low-floor, diesel-fuelled buses. Orion estimates their cost to manufacture these buses to be $234,229 per bus.

(c) Using the model in part (a), what is the distribution representing the uncertainty about the amount of the winning bid per bus for this contract? In particular, what are the mean and standard deviation of the distribution?

(d) Given the distribution in part (c), what is the probability that Orion wins the contract if it bids $240,000 per bus? If it wins the contract, what is its profit per bus per bus?

(e) What is the probability that Orion loses the contract if it bids $240,000 per bus? If it loses the contract, what is its profit per bus? (You do not need to take into account the cost of putting the bid together when determining the profit for a lost contract.)

(f) Why is there uncertainty about the profit per bus that Orion will obtain if it bids $240,000 per bus? What is the probability distribution representing this uncertainty? In particular, what is the mean of the distribution (i.e. what is the expected profit per bus if it bids $240,000 per bus)?

We now want to develop an Excel spreadsheet (or program in R) that will allow ExpectedProfit to be plotted against different possible bid amounts (i.e. $240,000; $241,000; ...; $260,000). The maximum of this graph will give Orion the bid amount that will maximize expected profit.

(g) Using the plot, what should Orion bid if it wants to maximize expected profit per bus?