Homework Assignment 4

Problem 1

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

For this problem us the data is the file **Profits.csv**.

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 a project has risk=7 and research and development = 76. Give the 95% predictive interval for the profit on the project.

(c) Suppose all you knew was risk=7. Run the simple linear regression of profit on risk and get the 95% predictive interval for profit.

(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

The data for this question is in the file **zagat.xls** . 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. 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 single 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 confident would you feel about it?

Problem 7: Housing Price Structure

The file **MidCity.xls**, available on the class website, contains data on 128 recent sales of houses in a town. For each sale, the file shows the neighborhood in which the house is located, the number of offers made on the house, the square footage, whether the house is made out of brick, the number of bathrooms, the number of bedrooms, and the selling price. Neighborhoods 1 and 2 are more traditional whereas 3 is a more modern, newer and more prestigious part of town. Use regression models to estimate the pricing structure of houses in this town. Consider, in particular, the following questions and be specific in your answers:

1. Is there a premium for brick houses everything else being equal?
2. Is there a premium for houses in neighborhood 3, all else being equal?
3. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

Problem 8: What causes what??

Listen to this podcast:
`http://www.npr.org/blogs/money/2013/04/23/178635250/episode-453-what-causes-what`

1. Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city)
2. How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below.
3. Why did they have to control for METRO ridership? What was that trying to capture?

EFFECT OF POLICE ON CRIME

TABLE 2

TOTAL DAILY CRIME DECREASES ON HIGH-ALERT DAYS

	(1)	(2)
High Alert	−7.316*	−6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

Figure 1: The dependent variable is the daily total number of crimes in D.C. This table present the estimated coefficients and their standard errors in parenthesis. The first column refers to a model where the only variable used in the High Alert dummy whereas the model in column (2) controls form the METRO ridership. * refers to a significant coeficient at the 5% level, ** at the 1% level.