Homework Assignment
Section 4

Tengyuan Liang
Business Statistics
Booth School of Business

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:

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

(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% plug-in 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 68% plug-in 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 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

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.

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

1. Based on the model assumptions, what is the expected value of R/G given OBP for teams in the AL? How about the NL?
2. Interpret β_0, β_1 and β_2.
3. 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?
4. Test the hypothesis that $\beta_1 = 0$ (with 99% probability). What do you conclude?

Problem 5

Read the case "Orion Bus Industries: Contract Bidding Strategy" (Search it... you should be able to find it! May cost you a few bucks...). 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 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?

Problem 6: Beauty Pays!

Professor Daniel Hamermesh from UT's economics department has been studying the impact of beauty in labor income (yes, this is serious research!!).

First, watch the following video:
`http://thedailyshow.cc.com/videos/37su2t/ugly-people-prejudice`

It turns out this is indeed serious research and Dr. Hamermesh has demonstrated the effect of beauty into income in a variety of different situations. Here's an example: in the paper *"Beauty in the Classroom"* they showed that *"...instructors who are viewed as better looking receive higher instructional ratings"* leading to a direct impact in the salaries in the long run.

By now, you should know that this is a hard effect to measure. Not only one has to work hard to figure out a way to measure "beauty" objectively (well, the video said it all!) but one also needs to *"adjust for many other determinants"* (gender, lower division class, native language, tenure track status).

So, Dr. Hamermesh was kind enough to share the data for this paper with us. It is available in our class website in the file **"BeautyData.csv"**. In the file you will find, for a number of UT classes, course ratings, a relative measure of beauty for the instructors, and other potentially relevant variables.

1. Using the data, estimate the effect of "beauty" into course ratings. Make sure to think about the potential many *"other determinants"*. Describe your analysis and your conclusions.

2. In his paper, Dr. Hamermesh has the following sentence: *"Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible"*. Using the concepts we have talked about so far, what does he mean by that?

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:

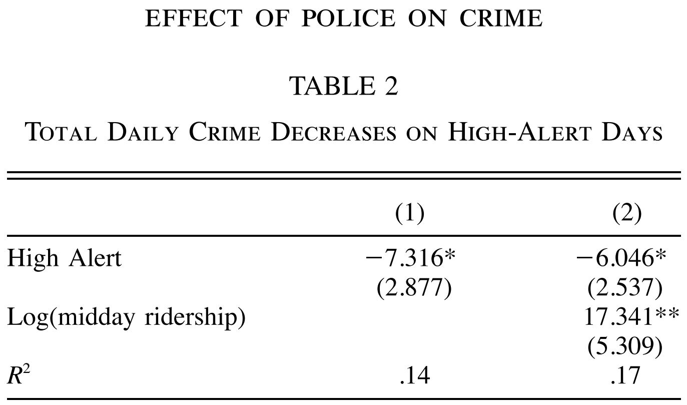
1. Is there a premium for brick houses everything else being equal?
2. Is there a premium for houses in neighborhood 3?
3. Is there an extra premium for brick houses in neighborhood 3?
4. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

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?
4. In the next page, I am showing you "Table 4" from the research paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?



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.

TABLE 4

Reduction in Crime on High-Alert Days: Concentration on the National Mall

	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert × District 1	−2.621**	−2.621*	−2.621*
	(.044)	(1.19)	(1.225)
High Alert × Other Districts	−.571	−.571	−.571
	(.455)	(.366)	(.364)
Log(midday ridership)	2.477*	2.477**	2.477**
	(.364)	(.522)	(.527)
Constant	−11.058**	−11.058	−11.058^{+}
	(4.211)	(5.87)	(5.923)

Figure 2: The dependent variable is the daily total number of crimes in D.C. District 1 refers to a dummy variable associated with crime incidents in the first police district area. This table present the estimated coefficients and their standard errors in parenthesis.* refers to a significant coeficient at the 5% level, ** at the 1% level.

Problem 9: Don't Take Your Vitamins

Read the following article:

`http://fivethirtyeight.com/features/dont-take-your-vitamins/`

List a few ideas/concepts that you have learned so far in this class that helps you understand this article.