Intro to Machine Learning
Take Home Exam

Carlos M. Carvalho
Texas MSBA
McCombs School of Business

Due: July 31st at 10pm (CT)

Book Problems

Chapter 2: #10
Chapter 3: #15
Chapter 6: #9 and #11
Chapter 8: #8 and #11
Chapter 10: #7

Problem 1: 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 2: 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 and answer the following questions:

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 3: 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 5: Final Project

Describe your contribution to the final group project (1/2 page max).