Exam Questions

8. Hypothesis test (20 points)

A recent report asserted that people who worked more hours also tended to be fatter (among those in certain occupations). (The paper doesn't give precise numbers so I'll make them up ??? don't bother with Google.) The paper did much more econometric analysis of course. Nevertheless, suppose that, of the 7219 women working non-strenuous occupations, 23% are working more than 40 hours/week. Of those women in non-strenuous occupations working more than 40 hours/week, 27.3% were obese; of those women in non-strenuous occupations working less than 40 hours/week, 24.6% were obese. There were also 714 women in strenuous occupations with 21% working more than 40 hours/week. Of the women in strenuous occupations working more than 40 hours/week, 28.1% were obese while 37.4% were obese among those working fewer hours. Does it seem likely that overtime makes certain groups more likely to be obese? *J Abramowitz, "Working Hours, Body Mass Index, and Health Status: A Time Use Analysis"*

Regression Analysis from given data (no datasets in R required)

Statistics using Datasets (R required)

1. Taxi data OLS, kNN, HT, CI (30 points)

Using a subsample of the taxi data, I find that on weekends there were 193750 rides paid with credit cards and 187694 rides paid with cash.

a. Find a 90% confidence interval for the fraction of rides paid in cash.
b. On weekdays there were 582335 rides paid with a card and 509798 paid in cash. What is a 90% confidence interval for the fraction paid in cash now?
c. Are these proportions statistically significantly different? Explain and calculate t-stat and p-value.
d. What are some possible explanations? What data would you want to consider additionally? I'm not (yet) asking for data just an explanation of your thought process.
e. Using that data (the smaller sample is on Blackboard) can you construct a knn estimate of which fares are likely to tip more than 15%? OLS estimates of tip amount?

2. ATUS data using CI, OLS, kNN (30 points)

ATUS records the numbers of minutes in a typical day that people spend on various activities. ACT_WORK is the number of minutes spent working; I'll define more than 420 minutes (7 hours) as fulltime.

number	Fulltime	Parttime
Educ HS diploma	8219	11879
Educ some college	6412	9989
Educ Bachelor	8065	12521

a. Conditional on the individual having a HS diploma, what fraction are working fulltime? What is a 95% Confidence Interval?
b. Conditional on the individual working part-time, what fraction have a Bachelor's degree? What is the 95% Confidence Interval?
c. Using the data on Blackboard, which is a subset of the ATUS data that selects people who are employed, can you find additional important factors in explaining the time spent at work? Explain with some OLS models and/or knn.

3. CEX data using OLS, tables (30 points)

Consider the CEX data; estimate some models to explain APPARPQ, expenditure on apparel in the previous quarter (includes MENBOYPQ, WOMGRLPQ, and FOOTWRPQ - expenditure on apparel for Men and boys; Women and girl; footwear). How important iseducational attainment on this expenditure category?

a. What are conditional mean expenditure on apparel for different educational levels? What about conditional means for those who spent a non-zero amount?
b. Can you estimate some interesting OLS models and discuss the important variables in explaining apparel expenditure? Explain.
c. Can you estimate further useful models? Explain

4. PUMS data using hypothesis tests (20 points)

Consider the PUMS data for people in NY, that we've been using in class. For now restrict attention to just working people (explain how you might define that).

a. Do a statistical test of the difference in average age between working people in the Bronx vs working people in Brooklyn. What is the 95% confidence interval for the difference in means?
b. What if you were using the Age data but regularized so that the min is zero and max is one [recall my function,

```
normalize <- function(X_in) {
  min_X_in <- min(X_in,na.rm = TRUE)
  max_X_in <- max(X_in,na.rm = TRUE)

  (X_in - min_X_in) / abs(max_X_in - min_X_in)
}
```

]. Would the statistical test come out the same? Why or why not?

5. PUMS using proportions, tables (25 points)

I used the PUMS data to look at wages and commute type, getting this table for people in the City: (you can answer parts a-c without R)

	bus	Car	Subway
Wage below $25,000	1501	2394	3704
Wage above $75,000	385	1825	2194

a. Given that someone takes the bus to work, what is the probability that they're making wages above $75,000?
b. Given that someone takes the subway to work, what is the probability that they make wages below $25,000?
c. Given that someone has wage above $75,000, what is the probability that they drive a car to work?
d. Using the PUMS data, can you narrow this further - what are the socioeconomics of bus/subway in the various boroughs? What is the wealthiest PUMA area and how do the people living there tend to commute?

6. CEX data using plots, tables, OLS kNN (25 points)

Use the CEX data that I provided and consider the fraction spent on entertainment, ENTERTPQ/TOTEXPPQ.

a. Find some descriptive statistics about this fraction, for some subgroups. Tell me something interesting about this data. Are there sub-categories that explain some of the variation?

b. Create a histogram and/or density plot. What do these reveal?
c. Estimate a linear regression and discuss what this shows.
d. Estimate a k-nn classification to predict which households are in the lowest 25% in terms of entertainment spending. Discuss what variables are important in classifying.

7. ATUS data. 25pts using kNN, OLS, HT

Using the ATUS data, describe the time spent working (ACT_WORK). Do people with more education work more or less hours than people with less education? What other factors are important? You should choose a variety of methods (perhaps including comparison of means, linear regression, nearest neighbor) that demonstrate your econometric virtuosity. Carefully specify the statistical tests that you perform, including the null hypothesis and test statistics including t-stat and p-value.

9. CEX data (20 points)

I used the CEX data to look at the fraction of spending going to health insurance. I get the following table, grouped by education of the reference person: %Insurance No HS HS diploma Some college, no degree Assoc degree Bach degree Adv degree less than 10% 467 1385 1191 615 1181 521 11% - 20% 82 231 157 71 122 58 21% - 30% 21 65 27 10 32 7 more than 30% 8 18 14 1 3 2 a. Conditional on the reference person having a college degree (Associate's, Bachelor's or Advanced), what fraction devote more than 20% of spending to health insurance? b. Conditional on the reference person having less than a college degree, what fraction spend more than 20% on health insurance? c. Is this difference statistically significant? d. What is the overall share (in this sample) of people with any college degree? What share of people spending more than 20% is made up of people with any college degree? e. Are those break points (+/- 20%; any degree) reasonable? Can you suggest better? Explain. f. What problems might there be, with the classification and analysis here? Can you do better with the CEX data?

10. !!!

(25 points) After the Nobel Prize awards to Fama, Hansen, and Shiller, we look at predictability of stock returns, using data on stocks in the S&P500. There are some days where many of these company's shares have negative returns; other days where many have positive. In 2012, more than 70% of the companies had positive returns on about 25% of the days; on another 25% of the days fewer than 30% had "up??? returns. On the days following"70% up??? days, the average return was .06 percent, with standard deviation of 1.72; on days following "30% up,??? the average return was .10 percent, with standard deviation of 1.66. There were 65 days of 70% or more up; there were 59 days of 30% or fewer up. xxi. (1 pt) What is the difference in means? xxii. (2 pts) What is the standard error of the difference in means? xxiii. (2 pts) Is this difference statistically significant? What is the p-value? Explain. xxiv. (20 pts) Using the data given on Blackboard, specify more hypotheses about stock behavior and test these.

11. !!!

11. (30 points) With the NSA spying revelations, we return to questions of whether there is wage discrimination

against people with ancestry from the Middle East or North Africa (MENA). I've created program in SPSS syntax and R that you can run, which will define MENA_ANC if the person's ancestry is from MENA (except Israel) or MENA_BPL if the person's birthplace is MENA. You should consider whether there are differences in wages and incomes between people from the MENA or others; of course one decision to make is who is a relevant comparison group. Calculate averages between groups, considering also things like education; which are statistically significant? Explain in detail.

12. !!!

(20 points) Use the ATUS data (available from Blackboard) on the time that people spend in different activities. o. Among households with kids, what is the average time spent on activities related to kids? p. Among households with kids, how much time to men and women spend on activities related to kids? Form a hypothesis test for whether there is a statistically significant difference between the time that men and women spend with kids. What is the p-value for the hypothesis of no difference? What is a 95% confidence interval for the difference in time? q. Why do you think that we would find these results? Explain (perhaps with some further empirical results from the same data set).

13. !!!

64. Using the ATUS dataset that we've been using in class, form a comparison of the mean amount of TV time watched by two groups of people (you can define your own groups, based on any of race, ethnicity, gender, age, education, income, or other of your choice).

a. What are the means for each group? What is the average difference?
b. What is the standard deviation of each mean? What is the standard error of each mean?
c. What is a 95% confidence interval for each mean?
d. Is the difference statistically significant?

14. !!!

Use the CPS dataset (available from Blackboard) to do a regression. Explain why your dependent variable might be caused by your independent variable(s). What additional variables (that are in the dataset) might be included? Why did you exclude those? Next examine the regression coefficients. Which ones are significant? Do the signs match what would be predicted by theory? Are the magnitudes reasonable? (Note your answer should be a well-written few paragraphs, not just terse answers to the above questions. No SPSS output dumps either!)

15. !!!

For the ATUS dataset, use "Analyze Descriptive Statistics Crosstabs" to create a joint probability table showing the fractions of males/females about the amount of time spent on the computer vs watching TV (if either or both are above average). Find and interpret the joint probabilities and marginal probabilities. Do this for age groups as well.

6. Regression Analysis using Datasets (R Required)

Problem 1 !!!

(20 points) Use the Fed SCF 2010 data (available from Blackboard). This is the Survey of Consumer Finances, which is not representative (without using the weights, which you need not do for now) ??? it intentionally oversamples rich people to find out about their finances. Concentrate for now on the variable "SAVING" (about the 100th variable in the list) which is the amount that people have in their savings accounts. f. Test the null hypothesis that there is no difference between people who are older or younger than 65. What is the p-value for this test?

Problem 2!!!

(25 points) Use the ATUS data (available from Blackboard) on the time that people spend in different activities. Construct a linear regression explaining the time that people spend on enjoyable activities (t_enjoy which includes most of the T12 items). Restrict the data to include only those people spending a non-zero amount of time on such activities. g. What are likely to be some of the most important determinants of time

spent on enjoyable activities? Which of these are in the ATUS data? Should the person's wage be included (do you think income or substitution effect would dominate)? What are some important determinants, that you could imagine a survey measuring, that are not in the ATUS data? You might find descriptive statistics for the included variables. h. Carefully specify and estimate a linear regression. What are the statistically significant coefficients? Which explanatory variables are most important? Are there surprises? Discuss your results. (You might want to estimate more models or create additional variables.)

Problem !!!

(25 points) Use the PUMS data (available from Blackboard) on the residents of NYC. Consider the time (in minutes) spent by people to travel to work; this variable has name JWMNP. r. How many men and women answered this question? What variables do you think would be relevant, in trying to explain the variation in commuting times? s. Form a linear regression with the dependent variable, “JWMNP Travel Time to Work,” and relevant independent variables. t. Which independent variables have coefficients that are statistically significantly different from zero?

Problem !!!

Use the SPSS dataset, atus_tv from Blackboard, which is a subset of the American Time Use survey. This time we want to find out which factors are important in explaining whether people spend time watching TV. There are a wide number of possible factors that influence this choice. a. What fraction of the sample spend any time watching TV? Can you find sub-groups that are significantly different? b. Estimate a regression model that incorporates the important factors that influence TV viewing. Incorporate at least one nonlinear or interaction term. Show the SPSS output. Explain which variables are significant (if any). Give a short explanation of the important results.

Problem !!!

41. Estimate the following regression:: S&P100 returns = ???0 + ???1(lag S&P100 returns) + ???2(lag interest rates) + ?? using the dataset, financials.sav. Explain which coefficients (if any) are significant and interpret them.

Problem 6. OLS !!!

17. (20 points) I will consider a simple question of the relation of employment to production ??? relevant both for questions of “jobless recovery” and worker productivity. In the R dataset, “macro_data1.Rdata”, I give monthly data for the US on payroll (total nonfarm), the unemployment rate, and an index of industrial production for the period from February 1948 to August 2014. There is also a dummy variable for when the US was in a recession (as defined by NBER). The dataset has both the level of each of these (denoted lvl_) and log difference (denoted ld_), where ????????*????(????) = ????????????(????????????????*(????)) ??? ????????????(????????????_????(???? ??? 1)). You can use the command, load(“macro_data1.RData”), to get the data in. I estimate the following regression for the period from 2000-date: ?????????????????????????????????? ???????????????????????????? = .000739 + .0512 ?? ?????????????????????????????????????? ??? .00270??????????????????????????????????? The intercept coefficient has standard error of .00011, the slope coefficient on percent change in production has standard error or .0161, and the Recession dummy has standard error of .0003. The R-squared is 0.4943.

a. What is the t-statistic for the slope coefficient? What is its p-value? (Carefully specify the null hypothesis.) What is a 95% confidence interval for the slope coefficient?
b. Suppose that next month (not a recession month), the percent change in production is 0.004 ??? what would the regression predict is the percent change in payroll?
c. How would you critique this regression? What might be improved?
d. Can you find some other interesting results from the data given? Explain.