7022MK PhD ECONOMETRICS I (fall 2017)

Modified Final Exam (Take Home) (December 2, 2017)

Extended Deadline: Dec, 7th 2017, 11.00 p.m. (originally Dec. 3rd)

Readme: *General Instructions:*

Modifications:

- The new deadline: Thu, Dec. 7th 2017. 11pm
- **You can "unselect" one or two exercises worth less than 40 points.** You will be graded on the remaining total, that is you will not loose any points by unselecting these exercises. **State at the beginning of your exam, which exercises were unselected.**
- Question A3 (The job-market paper task), is reduced to the first three papers you would have selected.

Advice:

- This exam consists of four parts.
  - **Part I: Reading Assignment [50pts]** Those are easy points, but read the paper.
  - **Part II: Theoretical Excercises [100pts]** consists of imagined problems. **Theory questions** are more heavily based on theory, and
  - **Part III: Applied questions and Real Problems [100pts]** ask you to relate what you learned to an empirical setting. It may partly have aspects of coding on paper, or feature real problems taken from research.
  - **Part IV: Learning and Coding Exercises.[150pts]** You have to choose 3 out of the 4 Problems.

  Answer each part separately.
- As a rule of thumb, each question gets increasingly difficult. This is true within Questions and within parts. I encourage you to move on to the next question/part if you get stuck.
- Depending on your type consider starting with the Applied Problems. Do make sure to skim over all questions before beginning with your answers.
- Be concise and direct in your answers. *Do not waste time with information that is not relevant for answering the questions.*
- Only interpretation questions are allowed. Ask others before you ask me.
- You may interact with fellow students if you would like to do so, but you have to state in the beginning of your answer, who you interacted with on which problem, and you have to provide your own solution.

- 4 days will almost certainly be too little time. Generally short answers are better. 2 (maybe 4) sentences is a good guideline for making one subpoint or giving one example. Try to do the easy tasks on all exercises, but move on once you get stuck (to make sure to do all the easy tasks).
- Even if you answer concisely and progress quickly, you will likely not have enough time, because it's likely that the exam is too long. Don't worry about it – it is too long for everybody.
- Good Luck for the Exam!

Excercises:

Reading: .

Question R1: Goldfarb Tucker Guide [50pts]

- What are the differences in the Etiquettes for DiD and IV. Based on the key assumption in the identification strategies, try to explain why the etiquette is different for the two methods.
- What are the differences in the Etiquettes for RD and IV. Based on the key assumption in the identification strategies, try to explain why the etiquette is different for the two methods.

Review/Revision (Theory) Excercises [100 pts. total]: .

Question T1: Definitions and Estimators [33pts]

1.1.) [10pts] Linear Probability model:

Suppose you have data on working behavior from 2002 and credit ratings data from 2008 and you want to predict whether a person was working in 2008 or not, based on their 2002 characteristics. y_i can take two values 0 and 1, you consider the linear regression model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_0 + u_i \qquad i = 1, \dots, n$$

with $\boldsymbol{\beta}_0$ belonging to a K-dimensional parameter space and

$$y_i = \begin{cases} 0 & \text{if } \textit{person was working in } 2008 \\ 1 & \text{if } \textit{person was retired/unemployed in } 2008 \end{cases}$$

a. Assume you wanted to use "linear regression" to estimate $\boldsymbol{\beta}_0$.

- How do you call such a dependent (0/1) variable?;
- How is the above model called (Regressing linearly with a dummy as dependent variable)?
- Use method of moments **OR** minimize the mean squared error to derive the estimator. [3pts]
- How would you interpret any of the estimated coefficients $\boldsymbol{\beta}_{0,k}$?
- Bonus: Can you imagine what is generally considered to be major shortcoming of this model [2 pts]

(Hint: part a is meant to be straight forward)

Question 1.2[13pts]: Advertising Beer Again

Recall the exercise were I asked you to discuss with an advertisement company whether the effect of advertisement beer in a stadium can be inferred from comparing the beer consumption in the stadium to the beer consumption in a Publix nearby.

- Before we start: Reconsider the plain comparison of beer consumption in the stadium (with advertisement) and beer consumption in Publix (no advertisement).
  - Write down the definition of the ATE we want to measure. Explain in which sense the ATE is unobservable (*Hint: think of counterfactual observations in the ATE.*
  - Write down what you can observe when you compare outcomes of a treated and a comparison group. (assuming that the sample averages are equivalent with population averages in the subgroups you may use conditional expectations.)
  - Show that these two expressions are generally different, by formally highlighting that the observed quantity consists of the treatment effect and the selection bias (i.e. decompose the observed quantity into treatment effect and selection bias).
  - Take this to the example of advertising beer in a stadium and explain, always referring to your formal definition, what the selection bias is in this case.

After long debates with your rather stubborn advertising company, they decided to run a cross check on whether advertisement for the beer works in publix stores, as it would work in a stadium. As a result they run a campaign for your beer in publix, which is based on special offer vouchers. There are two vouchers green and blue. Blue vouchers are distributed in the football stadium and green vouchers are sent to randomly selected participants in the publix loyalty program. Either voucher gives a discount of 25% when purchasing any quantity of beer in publix. We assume away that it might be possible to trade or pass the voucher (even though this is likely an absurd or a very high tech assumption).

- What is a Local Average Treatment Effect (LATE)? Provide a formal definition. Add verbal explanations.
- Can the voucher design above (green or blue alone) estimate a LATE? Under which what assumptions can you estimate the LATE?
- For both voucher colors, characterize the always takers, never takers, defiers and compliers.
- Assume you obtain two different LATEs for blue and green vouchers, how would you check if the difference is statistically significant? Under which condition would you attribute economic importance to that difference?
- If the effects differed, where would you expect to see the larger effect, and what would you infer from such heterogeneous effect?
- Under which formal condition/assumption would the LATE that you estimated through vouchers directly give you the ATE? Which of the two vouchers do you think comes closer to satisfying these assumptions? Justify your choice.
- Given your answer above, what necessary condition would you require for the estimated effect of green and blue vouchers, to not reject that these assumptions are not satisfied.
- Why is the setup of the advertisement company insufficient to figure out the ATE? Suggest a design that could maybe achieve this purpose?

.)1.3 [10pts]: Panel/DiD Problem

Note: This excercise is based on material to be covered on Nov. 14th. Try to get as far as you can, before the class and prepare to finish only that after next Monday.
For T = 2 consider the standard unoberved effects model:

$$y_{it} = x_{it}\beta + c_i + u_{it}; \quad t = 1, 2$$

where c_i is a time-invariant unobserved factor that is correlated with x. Let $\hat{\beta}_{FE}$ and $\hat{\beta}_{FD}$ denote the fixed effects and first difference estimators, respectively.

1. Show that the FE and FD estimates are numerically identical.

2. Show that the error variance estimates from the FE and FD methods are numerically identical.

3. Under what assumptions are these estimators consistent? State the assumptions and show consistency (sufficient for one-dimensional x_{it}, if you so prefer...)

Question T2: IV Problems and Method of Moments [17pts]

Take the model

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + u_i \\ \mathrm{E}(\mathbf{x}_i u_i) &= \mathbf{0} \\ \boldsymbol{\Omega} &= \mathrm{E}(\mathbf{x}_i \mathbf{x}_i' u_i^2) \end{aligned}$$

1. Find the method of moment estimators for $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}})$ for $(\boldsymbol{\beta}, \boldsymbol{\Omega})$. (Make sure that you notice the difference between the population moments and the sample moments).

2. In this model are $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}})$ efficient estimators of $(\boldsymbol{\beta}, \boldsymbol{\Omega})$?, justify your answer.

3. Quantify the bias if you use $\mathrm{E}(\mathbf{x}_i u_i) = \mathbf{0}$ when, in fact,

$$\mathrm{E}(\mathbf{x}_i u_i) = \kappa \neq 0$$

4. Assume you have one Instrument Z with

$$\mathrm{E}(\mathbf{Z}_i u_i) = \mathbf{0} \; and \; Cov(Z, x) \neq 0$$

   Use the method of moments to derive the IV Estimator (aka “2SLS”).

5. Running IV is implemented by estimating a first stage regression of the endogenous variable on all exogeneous variables and the instruments, and a second stage regression where y is regressed on the exogeneous (x-)Variables and the estimated endogenous x. Where in the above derivation of the IV Estimator can you see the first-stage and where the second stage regression.

6. (*Challenging*) Derive the expected bias of the 2SLS estimator. Show that it is 0 under the assumptions in question 4.

7. Assume now, the instrument is "weak and dusty", i.e. Cov(Z,x) is almost 0 and Cov(Z,u) is very small, but positive. Show the "weak instrument bias" under the new assumptions.

8. Analyze for any given κ, HOW "weak and dusty" your instrument can maximally be, to justify using IV as an alternative to OLS. (If you did not manage to do 3 and 4 you can use the formulas on the slides.)

Question T3: The unknown UK [17pts]

Consider the following demeaned[1] true model.

$$\begin{aligned} & y = x_1\beta_1 + UK\tau + \epsilon \\ \textit{with} & \\ & \mathrm{E}(\epsilon|\mathbf{x}) = 0 \\ & \mathrm{Cov}(UK, x_1) = -0.4 \\ & \sigma_{UK} = \sigma_{x_1} = 1.5 \end{aligned}$$

Questions:
1.) [6pts] The variable UK is unobservable and hence unknown to you. But you know its statistical association with x_1. Now you are asked to run a regression of y on x_1. Is that a problem that introduces bias, or is it a useful approach? If yes, why, if no, why not? (Argue with approx. 3-5 sentences...)

2.) [5pts] Proceeding you are looking at estimating the truncated regression

$$y_i = x_{1,i}\beta_1 + u_i$$

a.) Write down, what you know or assume about the residual u_i and about $\mathrm{E}(u|\mathbf{x})$. Depending on your answer in (1.), either write down and discuss the assumptions that guarantee an unbiased estimator, or, if you expect a bias, quantify it. [4pts]

b.) Do you have enough information to say what the result of regressing x_1 on UK would be? If yes, quantify, if no, say which additional information you would need. [2pts]

3.) [3pts] Derive the estimator for β_1 in the truncated model by using one of the methods we have seen in class (e.g. using the method of moments OR finding $\tilde{\beta_1}$, the β that minimizes the sum of the squared residuals.) Which estimator do you find?
Hint: If you think this helps you to save time, you may quote results from the previous excercises.

4.) [2pts] Now derive a formal expression for the expected bias.

5.) [1pts] Depending on your answer in 2a.) provide a proof that the expected bias = 0 (i.e. estimator is unbiased) under your assumptions or formally show that the bias is non-zero, and confirm your quantification.

Question T4: Linear and Non-Linear GMM [33pts]

Hint: this exercise has three subparts: If you get stuck with one subpart, it is worth trying the next one, until you get stuck again,...

[1] the constant can be ignored

2.1) [16pts.] Consider the the linear regression model as a special case of GMM. Given

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

. For $i = 1, \ldots, n$. Define the following moment equations

$$\begin{aligned} \mathbf{g}_1(\mathbf{w}_i, \boldsymbol{\beta}) &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})\,\mathbf{x}_i \\ \mathbf{g}_2(\mathbf{w}_i, \boldsymbol{\beta}) &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})\,\mathbf{h}(\mathbf{x}_i) \end{aligned}$$

where $\mathbf{w}_i = \begin{pmatrix} y_i \\ \mathbf{x}_i \end{pmatrix}$, $i = 1, \ldots, n$, $\mathbf{x}_i$ is a $K \times 1$ vector and $\mathbf{h}(\cdot)$ is a $P \times 1$ vector-valued function with $P \geq K$.

a. [5pts] Suppose that $E[u_i|\mathbf{x}_i] = 0$. Show that solving for $\hat{\boldsymbol{\beta}}$ from

$$\sum_{i=1}^{n} g_1(\mathbf{w}_i, \boldsymbol{\beta}) = 0$$

is equivalent to obtaining an estimate for $\boldsymbol{\beta}$ from

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 .$$

What will you get if you estimate GMM based on $\mathbf{g}_1(\mathbf{w}_i, \boldsymbol{\beta})$? USe your derivations from above to defend your answer.

b.) [3pts] Can you think of an example function to use in $\mathbf{h}(\cdot)$? (*Hint: If not, simply move on.*)

c. [2pts] Under the above conditions, show that

$$E[\mathbf{g}_2(\mathbf{w}_i, \boldsymbol{\beta})] = 0.$$

(*Hint: also if you cannot show this, assume it is correct, given that* $E[u_i|\mathbf{x}_i] = 0$. *and move on.*)

d.) [3pts] Provide the definition of the GMM-Estimator.

e.) [3pts] Propose a GMM estimator for $\boldsymbol{\beta}_0$ based on $\mathbf{g}_2(\mathbf{w}_i, \boldsymbol{\beta})$.

2.2) [13pts] Instrumental Variables

a.) [4pts] You are actually sceptical about your assumption that $E[u_i|\mathbf{x}_i] = 0.$. Provide 4 reasons, why this assumption is frequently violated in Econometric applications.

b.) [4pts] You have access to a set of instrumental variables stacked in a vector $\mathbf{z}_i$ and you can augment your dataset (still in the linear model) so that you now have $\mathbf{w}_i = \begin{pmatrix} y_i \\ \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix}$, $i = 1, \ldots, n$, $\mathbf{x}_i$ is a $K \times 1$ vector, $\mathbf{z}_i$ is a $L \times 1$ vector and $\mathbf{h}(\cdot)$ is a $P \times 1$ vector-valued function with $P \geq K$ and $L \geq K$. Provide a new suitable moment equation $\mathbf{g}_3(\mathbf{w}_i, \boldsymbol{\beta})$ that allows you consistently estimate $\boldsymbol{\beta}$ if you use the new variables $\mathbf{z}_i$

c.) [3pts] What other assumptions are required for consistent estimation of $\boldsymbol{\beta}$?

d.) [2pts] Propose a GMM estimator based on $\mathbf{g}_3(\mathbf{w}_i, \boldsymbol{\beta})$ Choose the number of instruments freely.

2.3) [5pts.] Non-linear GMM model Rather than doubting the exogeneity of $\mathbf{x}_i$ you would prefer to relax the linearity assumption. In a paper you see the more general model that they use non-linear GMM

$$y_i = m(\mathbf{x}_i'\boldsymbol{\beta}) + u_i$$

for $i = 1, \ldots, n$. and the authors define the following moment equations

$$\begin{aligned} \mathbf{g}_1(\mathbf{w}_i, \boldsymbol{\beta}) &= (y_i - m(\mathbf{x}_i'\boldsymbol{\beta}))\frac{\partial m(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ \mathbf{g}_2(\mathbf{w}_i, \boldsymbol{\beta}) &= (y_i - m(\mathbf{x}_i'\boldsymbol{\beta}))\,\mathbf{h}(\mathbf{x}_i) \end{aligned}$$

where $\mathbf{w}_i = \begin{pmatrix} \mathbf{x}_i \\ y_i \end{pmatrix}$, $i = 1, \ldots, n$, $\mathbf{x}_i$ is $K \times 1$ vector, and $\mathbf{h}(\cdot)$ is $P \times 1$ vector-valued function with $P \geq K$.

a. [2pts] Can you provide linear model above (Part 1) as a special case of the model in this part, i.e. can you find a function $m(\mathbf{x}_i'\boldsymbol{\beta})$ that gives the model in part 1?

b. [1pts] Suppose that $E[u_i|\mathbf{x}_i] = 0$. Show that solving for $\hat{\boldsymbol{\beta}}$ from

$$\sum_{i=1}^{n} g_1(\mathbf{w}_i, \boldsymbol{\beta}) = 0$$

is equivalent to obtaining an estimate for $\boldsymbol{\beta}$ from

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i'\boldsymbol{\beta}))^2 .$$

c. [1pts] Under the above conditions, show that

$$E[\mathbf{g}_2(\mathbf{w}_i, \boldsymbol{\beta})] = 0.$$

d. [1pts] Propose an optimal(efficient) GMM estimator for $\boldsymbol{\beta}_0$ based on $\mathbf{g}_2(\mathbf{w}_i, \boldsymbol{\beta})$.

Applied and Real Problem:

[100 pts total]

Question A1 Squirrelnomics - IV [14pts]

Recall your favourite computer-squirellogist, nephew/niece or other younger friend, who had become very interested in the effect of squirrel populations on nut tree populations. The worry is that very large numbers of squirrels are very bad for the tree populations.

S/he knows python pretty well and managed to scrape data that contains the number of squirrels per square mile and the number of relevant trees and plants (pecan, walnut, hazelnut, almond, pistachio, acorn, cashew, chestnut, hickory nut, pine, and macadamia trees) both. His data was evaluated at the end of 2014 and covers all of the U.s., Russia, China, India, South Korea, Croatia, Colombia, Ukraine and Austria.

Questions:

You like you young scientist-minded friend a lot and want to help finding the true effect. So you consider an IV-strategy.

i [2pts.] Given a valid Z, provide the formal definition of the IV-Estimator.

ii [2pts.] Describe which steps you would take in the estimation procedure (bullet points are good)

iii [2pts.] Which 2 conditions would a valid instrument Z have to satisfy.

iv [2pts.] Which conditions can be tested and how?

v [3pts.] Now, back to squirrelnomics, think of

- (i) a good instrument (Which conditions are satisfied for good instruments?),
- (ii) and a bad instrument (Which conditions are not satisfied for bad instruments?)
- (iii) for both candidates discuss (verbally but briefly!) why they succeed or fail (referring to your conditions above).

vi [3pts.] Finally, considering the bad IV, show formally what happens if you use it.

Question A2: (Identification, DiD and RD) Online "Markets" [36pts]

You are asked to help estimate the effect of online word of mouth (tripadvisor-like reviews) for doctors and hospitals on the quality of the health system. You could get data on health outcomes in either Canada or Portugal. The data would be very expensive, but they would contain administrative (health insurance) data on all doctors and all health related services that the insured used (medications, treatments by doctors, hospital expenses,...). Moreover, you have access to the full data from the largest three platforms for online reviews in each country. (all reviews, for all doctors and hospitals; on the review level, you know the rating, when the review was published, and the self-declared age and gender of the reviewer; from these you can generate the current average ratings of any given doctor etc.)

Since you could get variation over time you want to study how increasing availability of online reviews affects doctors' and hospitals' performance. You can get reasonably objective performance measures, by measuring how well patients were (or at least how long they were healthy/lived) after any given treatment. (indicators such as avg. survival rate for critical operations; average time of survival after critical operation; time away from treatment after treatment, etc.

a.) [18pts.] Identification:
In what follows, I use "health professional" ("HP") for both doctors and hospitals.

- A first idea is taking a cross section and regressing the number of available reviews on an HPs performance. Comment.
- After some data gathering a panel would be available. Write down a suitable panel specification. Is that better? Why or why not? Is it good enough? Why or why not?
- Let's get systematic: Provide a structured procedure to reach a research design. Which elements does it have?
- Provide the ideal experiment that would be suitable to identify the causal effect of interest?
- Discuss, which threats to identification you see?
- Suggest a design that circumvents as many as possible (ideally all).
- What will be the critical (identifying) assumptions of your design, what will be the remaining key weaknesses? (i.e. assumptions that are not satisfied etc.)
- The question is underresearched and potentially extremely important. However the data collection is quite costly and the analysis will take a lot of time. Given your suggestions above, would you recommend moving forward with this research?

b.) [18pts.] Difference in Differences or Regression Discontinuity?

- Write down the estimation equation of a Difference in Difference design.
- Write down the estimation equation of a Regression Discontinuity design.
- Suggest the best (real or pseudo-) Diff in Diff or RD that you can come up with. Do you think it would be feasible to set up a Difference in Difference in this context? Why or Why not? If you picked RD, discuss whether it will be sharp or fuzzy and what that implies...
- You are now worried about the following: As users gain access to the internet and write reviews, doctors also get access to the internet and they might use it for consultation and learning (Web, Databases, other Doctors via mail).
  - Why could that be a problem? Under which assumptions would you not have to worry? Provide a formalization of your reasoning (show the bias/absence of bias)
  - Do you think the DiD or RD you suggested would work? Under which (key, not regularity) assumptions? Refer to you answer above to argue this formally.

  - If your approach does not work, can you think of a way to improve the design? If it does, what would you have to show to readers of your paper to convince them. Discuss by referring to all points of the DiD/RD Etiquette.

Question A3: Job Market Papers [50pts]

In the data folder, you will find a folder called "Job Market Papers," which contains 10 papers. *(Note: this folder goes online tomorrow, on Nov 30th!)* Note that these are real Job Market Papers by this year's candidates. (So this is roughly the level that you will have to achieve in 3-4 years from now). Pick **three** (previously four) papers according to the following rule:

- Based on the first letter of your first name:
  - evaluate the letter's position in the alphabet, and use that number.
  - Of that number use single digit for choosing Paper 1. (0 gives paper number 10)
  - Use the sum of the 2 digits to choose paper number 2.
  - If you end up with the same number use paper number 10-x
  - if that still is the same paper or it does not make sense, use the next highest (x+1).
- Repeat the procedure, based on the first letter of your family name (if both names start in the same letter take the second letter):
  - pick the third paper based on the decimal digit of the position your family name's first letter has in the alphabet.
  - BONUS [10pts.]: you may pick a fourth paper (e.g. if you already went through 4 papers) based on the sum of the digits. **Note that the fourth paper is no longer assigned.**
  - whenever you get to a paper you already have picked, choose the subsequent paper, until you arrive at one that you have not yet seen.
- Example: Michael Kummer starts in M (position 13) and K (position 11)
  1. decimal digit of 13 is 1 → Paper 1 is paper number 1
  2. sum of digits in 13 gives 4 → Paper 2 is paper number 4
  3. decimal digit in 11 is 1 → Paper 3 would be number 1 again, so instead pick 9
  4. sum of digits of 11 is 2 → Paper 4 is paper number 2
- so, my papers would be 1, 4, 9, and, optionally, 2.
- Note: if the first letter is among the first 10, e.g. you have D in position 04, that gives paper 10 and 4.

Tasks:

1. State your two letter numbers, and which papers you are allocated.

2. Download the papers and for each paper answer the following questions:
    (a) WHAT is their main question?
    (b) WHY do they analyze this problem, and why should economists care?
    (c) HOW do they analyze the problem? (Which methods do they use?)
    (d) SO WHAT do they find?

    Rank the papers according to the following criteria:

    - could you find all the information in the abstract, did you have to browse conclusions, or did you have to dig into the introduction or even deeper?
    - Which paper do you find most compelling based on their main question and research strategy? (Rank them, and provide your reasons.)
3. For each paper answer the following question regarding their identification approaches:
    - What's their identification strategy?
    - What are the key assumptions in that identification strategy?
    - Taking this to their setting, what do you think are their most important identifying assumptions? Can you write the key assumptions not in terms of X and Y, but in terms of their dependent and independent variables?
    - How credible is this assumption, can you think of a counterexample, or not?
4. Look at the robustness checks.
    - Do they use any alternative identification strategies? Which?
    - Do they address your main concern? How?
5. Now rank the papers for how convincing you consider their main identification strategy.
6. Would you invite your number 1 for a job talk?

A little bit of background: "The jobmarket" in Econ begins in early fall every year and culminates in the ASSA meetings which are typically held right after new year. Every candidate applies with their CV, their Job market paper, typically a second research paper, and a minimum of three letters of reference. Ot access some profiles of Job Market Candidates, and their papers, you can start here: by googling or then, e.g., here for a page of candidates from one program. To give you an idea, we had a relatively broad search this year and received a high 3-digit number of applications.

Practice Excercises: .

Pick 3 Exercises worth 150 points. Note 1: You may use a different language than R with a 10 per cent discount. Provide the corresponding commands and let me know the language you are using. Note 2: If you do not know one thing move on.

Question P1a: Job training in R [20pts]

You are hired as a data analyst for a major online dating website. They are planning to improve their matching algorithm so they ask you to investigate which factors are related with extra marital affairs.
(1) (5 points) Define the population of interest. Describe how you would collect the data for analysis.
(2) (15 points, 2.5 each) After your data collection procedure you were able to assemble the following dataset: "affairs.RData." Load the data into R and answer the following questions.
(a) Generate a nice summary statistics table with the main variables. (b) How many men and women are in your sample? (c) How many people have kids? (d) What's the average age of your respondents? What's the average education of women? (e) Create a bar plot of the number of people by attitudes towards religion. (f) Create a bar plot of the number of people by happiness in marriage.

Question P1b: Job training in R [30pts]

Imagine you had data on a job training experiment for a group of men. (let the filename be: "jtrain2.RData"). Men could enter the program starting in January 1976 through about mid 1977. The program ended in December 1977. The idea is to test whether participation in the job training program had an effect on unemployment probabilities 1978. The variable "train" is the job training indicator. The variable "unem78" is the indicator of unemployment in 1978.

Programming Questions

(1) Clean your R working environment (how do you do this?)

(2) Set your R working directory and load the ggplot2 library, and
(3) load the "jtrain2.RData" dataset into R.

(4) Describe the data:

- Which strategy (and associated commands) can you use to check if the data was read correctly and which variables you have?
- Generate a table with summary statistics (*Hint: use stargazer*
- Which command can you use to figure out how many people in the sample participated in the job training program?
- How to detect outliers?

Table 1: Regression Results - Question P1b

```
##
## ===============================================
## Dependent variable:
## ---------------------------
##                           unem78
## -----------------------------------------------
## train                     -0.132**
##                           (0.055)
##
## Constant                  0.384***
##                           (0.035)
##
## -----------------------------------------------
## Observations              300
## R2                        0.019
## Adjusted R2               0.016
## Residual Std. Error       0.467 (df = 298)
## F Statistic               5.804** (df = 1; 298)
## ===============================================
## Note: *p<0.1; **p<0.05; ***p<0.01
```

- How can you analyze the distribution of $unem78$, age and $educ$?

(5) Assume you found the following variables:

- unem78 ; train ; unem74 ; unem75 ; age ; educ ; black ; hisp ; married

Run a "linear regression" with the following model:

$$unem78 = \beta_0 + \beta_1 train + \beta_2 unem74 + \beta_3 unem75 + \beta_4 age + \beta_5 educ + \beta_6 black + \beta_7 hisp + \beta_8 married$$

(6) Assume your output is the one in table 1: Interpret the coefficient for the train variable. Did you obtain the same result?

(7) BONUS: (5 points) TRY to run a "probit regression" with the following model:

$$unem78 = \beta_0 + \beta_1 train + \beta_2 unem74 + \beta_3 unem75 + \beta_4 age + \beta_5 educ + \beta_6 black + \beta_7 hisp + \beta_8 married$$

Most likely you do not know how to do that, which strategy would you use to find out, given that R is such a widely used tool?

Question P2a: Mock Monte Carlo: Asymptotics vs. Endogeneity [25pts]

1.) Generate a 5-dimensional multivariate normal (X1, X2, X3, Z1, Z2). 5,000 observations, with bilateral correlations as follows:
Correlations:

	x1	x2	x3	Z1	Z2
x1	1.0	0.2	0.1	0.35	0.0
x2	0.2	1.0	0.0	0.4	0.0
x3	0.1	0.0	1.0	0.0	0.4
Z1	0.35	0.4	0.0	1.0	0.6
Z2	0.0	0.0	0.4	0.6	1.0

Add a vector of ones (or any other constant), and in addition generate 3 error terms.

- a. ε_1 should follow a normal around 0 with variance 1
- b. ε_2 should follow a negative binomial NB(2, .6)
- c. $\varepsilon_3 = 0.15x1 + \varepsilon_1$

Finally generate three dependent variables:

$$y_i = 0.2x_1 + 2x_2 + 0.7x_3 + \varepsilon_i$$

...where i takes the values 1, 2 and 3.

- a. Compute the sample covariances for all pairs and the sample variances (including the 3 y variables) and show the result.
- b. Now generate 3 more samples with 50, 500, and 100,000 observations. Report the three new sample variance-covariance matrices.

2.) Next, use the smallest sample, and run a regression for all three variables.

- Which coefficient $beta_3$ do you expect regression 1, which one do you find, and why?
- Which coefficient $beta_3$ do you expect regression 2, which one do you find, and why?
- Which coefficient $beta_3$ do you expect regression 3, which one do you find, and why?

3.) Now increase the sample size gradually, and, for all three sample sizes, run an OLS regression for all three variables.

- as sample size increases, how does coefficient $beta_3$ change in regression 1, compared to your first estimation. How much further/closer is your result to what you expect to find, and why is that so?
- as sample size increases, how does coefficient $beta_3$ change in regression 2, compared to your first estimation. How much further/closer is your result to what you expect to find, and why is that so?
- as sample size increases, how does coefficient $beta_3$ change in regression 3, compared to your first estimation. How much further/closer is your result to what you expect to find, and why is that so?

4.) Is there any fundamentally biased coefficient in one of the three regressions? Which is it, and can you suggest a different strategy to estimate the coefficient correctly? Sketch or do.

Question P2b: IV Excercise [25pts]

IV Problems

Consider again the omitted variable model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \gamma q + \nu$$

where q represents the omitted variable and $E[\nu|\mathbf{x}, q] = 0$. ν has zero mean and is uncorrelated with $x_1; \ . \ . \ . \ ; x_K$ and q. The unobservable q is thought to be correlated with at least some of the x_j. The solution that would usually follow is to put q in the error term, and then to find instruments for any element of x that is correlated with q.

Assume without loss of generality that $E(q) = 0$. You have a single indicator of q, written as $q1 = \delta_1 q + a_1, \ \delta_1 \neq 0$, where a_1 has zero mean and is uncorrelated with each of x_j , q, and ν. In addition, $z_1; z_2; ...; z_M$ is a set of variables that are (1) redundant in the structural equation from above and (2) uncorrelated with a_1.

1. Suggest an IV method for consistently estimating the β_j . Be sure to discuss what is needed for identification.
2. If the equation above is a log(wage) equation, q is ability, q_1 is IQ or some other test score, and $z_1; ...; z_M$ are family background variables, such as parents' education and number of siblings, describe the economic assumptions needed for consistency of the the IV procedure in part a.
3. Carry out this procedure using the data in wage_educ.csv. Include among the explanatory variables exper, tenure, educ, married, south, urban, and black. First use IQ as q_1 and then KWW. Include in the z_h the variables meduc, feduc, and sibs. Discuss the results.

Question P3: Panel Excercise [50pts]

Thailand's Million Baht Village Fund Program is among the largest scale government micro finance initiatives of its kind. This intervention injected potential funds into 77,000 heterogeneous Thai villages. Each transfer of 1 million baht (about $24,000) was used to form an independent village bank for lending within the village. Every village, whether poor or wealthy, urban or rural, was eligible, and all villages in our data did indeed receive the funds.

Goal: Our objective is to measure the impact of village funds (*vfst*) on the number of new short term loans by households (*newst*) and on households' total consumption level (*tc*).

Data: The data set includes the following variables:

caseid: household identifier
year: year
village: village id
newst: total amount of new short term credit
tc: total household consumption
vfst: village funds available for lending
madult: male adults in the household
fadult: female adults in the household
kids: children in the household
maleh: indicator variable (=1 if head of household is male)
farm: indicator variable (=1 if farmer)
ageh: age of head of household
age2h: age of head of household
educh: years of education of head of household

Questions:

1. *Theoretical Question*: Assuming that the amount of initial village funds (the amount of available funds in each village before receiving the 1 million bahts) was randomly assigned. Write down, in equation form, the model that allows us to estimate the impact of village funds on i) new short term credit, ii) total consumption.

2. *Theoretical Question*: Assuming the same context described in question 1, explain how you would assess whether village funds have a different impact on female lead households relative to male lead households. Write down your model in equation form.

3. *Theoretical Question*: Comment on the plausibility of the assumption provided in 1 (that initial village funds are random). If this assumption fails, what impact will it have on your estimates of the effect of village funds on i) new short term credit and ii) total consumption?

4. *Theoretical Question*: Now assume that initial village funds are not random and instead are based on the size of the local population, number of firms, etc. What estimation procedure would you suggest using in order to assess the impact of village funds on households' new short term credit and total consumption? Carefully justify your choice. How would this procedure avoid the problems you described in question 3?

5. *Practical Question*: Now estimate the model you proposed in question 1 using i) Ordinary Least Squares (OLS), ii) First Differences (FD), iii) and Fixed Effects (FE). Do not forget to include a time trend in all your models. Interpret your results relating the obtained beta coefficients of each model with your responses to the previous questions.

6. *Theoretical Question*: Why is it important to include a time trend?

7. *Practical Question*: Test the assumption of serial correlation in the first differences model.

Question P4: DiD Excercise [50pts]

Texting Drivers

In order to reduce the number of fatal driving accidents caused by sending or receiving text messages many states in the U.S. have banned texting while driving. (Notice that this restriction is not the same as forbidding the use of any type of handheld device while driving, it only bans text messaging while driving.)

Goal: Your goal is to determine whether there was a reduction in fatal accidents following the state bans on text messaging.

Data: The data in 'txtbans.RData' comes from the Fatality Analysis Reporting System (FARS) of the NHTSA, an American census for all motor vehicle crash fatalities. It includes monthly information on fatal accidents. This data is merged with data on the enactment of text message bans and some state demographics.

The variables in the dataset are:

state: state id
time: month id
after: indicator variable (=1 if after treatment)
treated: indicator variable (=1 if text ban was implemented in that state)
laccsvobyinhab: log of (accsvobyinhab+1)
txmsban: indicator variable (=1 if text message ban is in place)
callban: indicator variable (=1 if cell-phone calls ban is in place)
bantime: id of the month when the ban was implemented
pop: state population
lpop: log of state population
unemp: state unemployment rate
lunemp: log of state unemployment rate
permale: percentage of males in the state
rgastax: gas (fuel) tax.
lrgastax: log of gas (fuel) tax
accsvobyinhab: number of fatal single-vehicle accidents with a sole occupant crashing into a non vehicular object per 1000 inhabitants.

Questions:

1. Explore the data set. How is this case different from the panel cases we have seen until now?
2. First, keep only the observations from April to run a simple 2-period diff in diff.
   - Define the control and treatment group and write down your regression model in equation form.
   - Create a table with the means of the variables of interest for the treatment and control groups before and after treatment.
   - Use the table you created to calculate the difference-in-differences estimator over all periods.

- As seen in class, run a simple difference-in-differences regression model in order to obtain the same estimator as in 2).
- Interpret all the coefficients in the model. Was the text message ban effective in preventing car accidents?
- On top of the standard linear regression model assumptions, what is the additional assumption needed for difference-in-differences to estimate the average treatment effect?

3. For turning to the multi-period Diff-in-Diff, start again from the full dataset:

   - Using R's plotting capabilities, explore whether or not the assumption in part 2 is reasonable. Justify your choice of plots.
   - For running a multi-period diff in diff, return to the original dataset and create a new variable that measures "months since ban" (negative before the ban, set the "placebo-bantime" for untreated units in the middle of the dataset or (even better) randomly)
   - Plot the treated and untreated units, but use time since ban as new time. Do your conclusions regarding the assumption in part 2 change?
   - Create period-specific time dummies (1 month to treatment, month of treatment (months to treatment == 0)) , 1 month since treatment, 2 months since treatment, 3 months since treatment, etc.). Moreover create crossterms that capture month-specific treatment effects "(treated)*(1 month since treatment), (treated)*(2 months since treatment), (treated)*(3 months since treatment),... – **Write down the corresponding regression model.**
   - Now estimate a multi-period difference-in-differences. Compare the estimated effects with the figure you plotted.
   - Give examples of what could be credible violations of the assumption in the previous part that could help explain the results observed.

4. The variable *callban*: indicates whether a state has a law in place forbidding cell-phone calls while driving. What percentage of the states that implemented the text ban also had a call ban in place? How can the existence of a ban on cell-phone calls affect the enforcement of the text ban?

5. (*Hard (Bonus) Question*) There is a function called "linearHypothesis" in the package "car." Find out on your own how this function works. Try to test the assumption in part 2 using this function. *Hint:*: The function is similar to running the model that interacted the treated variable and the time period dummies. Think about the interpretation of these coefficients.