Quantitative Empirical Methods Exam

Yale Department of Political Science, January 2015

You have seven hours to complete the exam. This exam consists of two parts. Provide complete answers to **both** sections.

Back up your assertions with algebra wherever appropriate, and remember to show your work. Good answers will provide a direct answer that illustrates an understanding of the question, and calculations or statistical arguments to validate the answer. Where applicable, exceptional answers will include all of these *as well as* proofs that are technically complete, including formally articulating sufficient assumptions and regularity conditions.

Part 1 (Short Answer Section) consists of ten short answer questions. The only aids permitted for Part 1 are (i) one page of double-sided notes, (ii) a calculator, and (iii) a word processor on one of the Statlab computers to write up your answers (you may also write up answers using pencil/pen and paper). *Advice*: Note there are multiple correct answers to some questions, and we encourage you to give the most complete (but still succinct) solution possible. Do not leave sub-parts of questions unanswered.

After handing in your answers for Part 1 of the exam, you may begin Part 2 (though feel free to look ahead). You may hand in Part 1 whenever you wish, but we recommend spending no longer than five hours on Part 1.

Part 2 (Computer Assisted Section) will involve using statistical software to answer one longer exercise with six associated questions. A complete answer to Part 2 will include code and output, as well as your written answers. *Advice*: We recommend that you explain what you are trying to do in comments. Even if you are not able to execute your program correctly, you can receive partial credit for explaining clearly what you wanted to do and why.

For Part 2, you are permitted (i) unrestricted use of your own computer with access to the internet or (ii) use of a Statlab computer with access to the internet. The only restriction for Part 2 is that you may not interact with anyone, online or otherwise.

For Part 2 (Computer Assisted Portion) of the exam, please turn in a hard copy of your code to Colleen, and also email a digital copy of the code to colleen.amaro@yale.edu.

1 Short Answer Section

1. Assume that the conditional expectation function of Y given X and Z,

$$\mathrm{E}\left[Y|X,Z\right] = 42 + 50XZ + 37X^5 + 50Z^2.$$

What is the "marginal effect" of Z on Y (i.e., instantaneous change in the conditional expectation function) when $X = 5$ and $Z = 1$?

2. In an observational study of the effects of attending Catholic school, the central dependent variable of interest is a binary variable, Y, which indicates whether or not the student graduated from high school. The treatment variable, T, indicates Catholic school attendance. Suppose that you know the joint distribution (Y, T).

Suppose that the joint probability mass function of (Y, T),

$$f(y,t) = \begin{cases} A & : y = 0, t = 0 \\ B & : y = 1, t = 0 \\ C & : y = 0, t = 1 \\ 1 - A - B - C & : y = 1, t = 1 \\ 0 & : otherwise. \end{cases}$$

Assume the stable unit treatment value assumption (SUTVA), i.e., $Y = Y_1 T + Y_0(1 - T)$. Answer all questions in terms of A, B, and C.

   (a) Make no further assumptions. Place sharp bounds on the ATE, $\mathrm{E}\left[Y_1 - Y_0\right]$. What is the width of these bounds?

   (b) Assume that Catholic school does not prevent any student from graduating who would have otherwise done so. Place sharp bounds on the ATE. What is the width of these bounds?

   (c) Assume that Catholic school was randomly assigned. Place sharp bounds on the ATE. What is the width of these bounds?

3. Suppose that the data generating process is $Y_i = a + bX_i + u_i$, with $\mathrm{E}\left[u_i|X_i\right] = 0$. However, the researcher observes only values of Y_i when $Y_i \geq 0$. Suppose that the researcher drops all observations of (Y_i, X_i) with missing values on Y_i, and performs an ordinary least squares (OLS) regression on the remaining values of Y_i on X_i (and a constant). Will the estimated slope from this regression generally be consistent for b? Why or why not?

4. You observe the following empirical cumulative distribution function for n draws from a random variable X:

$$\hat{F}(x) = \begin{cases} 0 & : x < 0 \\ 0.4 & : 0 \leq x < 1 \\ 0.8 & : 1 \leq x < 2 \\ 1 & \quad x \geq 2 \end{cases}$$

(a) What is the sample mean of X?

(b) What is the (plug-in) sample variance of X? (I.e., do not correct for $n/(n-1)$.)

(c) Assume $n = 100$. Estimate a normal approximation-based 95% confidence interval for $\mathrm{E}[X]$.

(d) Assume $n = 100$. Estimate a normal approximation-based 95% confidence interval for $\Pr(X \leq 1.457)$

5. (a) Loosely speaking, what is the central limit theorem?

   (b) Provide a formal statement of the central limit theorem.

   (c) Provide a proof sketch for the central limit theorem. [Note: it is *not* expected that you will solve this.]

6. Suppose that you have n observations i.i.d. (Y, D), where both Y and D are continuous random variables with finite variance. Articulate a set of (nontrivial) conditions under which the following expression: $\widehat{\mathrm{Cov}}\,(Y, D)/\widehat{\mathrm{Var}}\,(D)$, well approximates a causal effect (of D on Y). $\widehat{\mathrm{Cov}}\,(.)$ denotes the sample covariance and $\widehat{\mathrm{Var}}\,(D)$ denotes the sample variance of D.

7. Define nonparametric, semiparametric, and parametric models. Give examples of each.

8. Assume you have regressor matrix X and outcome vector $\mathbf{Y}$. Assume X contains a constant.

   (a) Under what conditions is the OLS solution from the regression of $\mathbf{Y}$ on X uniquely defined? (I.e., there exists one and only one possible solution that minimizes the sum of squared residuals.) Consider an OLS solution $\hat{\beta}$.

   (b) Give an example of data for which the OLS solution would not be uniquely defined.

   (c) Assuming that there exists a uniquely defined OLS solution, write down a closed-form expression for $\hat{\beta}$ using matrix algebra.

   (d) Assuming that there exists a uniquely defined OLS solution, what is the mean of the elements of $\mathbf{Y} - X\hat{\beta}$?

9. A researcher randomly assigns 500 of 1000 names to a treatment that consists of a phone call inviting the subject to turn out to vote in a coming election. Once randomization is done, the researcher realizes that 600 names in the list appear once and 200 were duplicated and appear twice. Before the phone calls are made, duplicate phone calls are discarded, so that no one receives more than one phone call. (I.e., anyone that was assigned to at least one phone call receives treatment.)

   (a) What is the probability of assignment to the treatment group among those whose names appear once and among those who appear twice in the original list?

   (b) Given the manner in which treatment was assigned, propose an unbiased (or approximately unbiased) estimator of the average treatment effect. You may assume the stable unit treatment value assumption (SUTVA).

10. For the analysis of longitudinal data (i.e., TSCS or panel data), there is debate in the social sciences about the use of fixed effects, random effects, and pooled regression for estimating causal effects. What are fixed effects regression, random effects regression, and pooled regression? Summarize and critically assess the arguments made about these types of estimators.

2 Computer Assisted Portion

People sometimes discuss how "regression adjustment" can cause efficiency losses in estimating average causal effects for randomized experiments. Some researchers suggest using the difference-in-means estimator instead of OLS adjusting for covariates, especially when n is small. We want you to conduct some simulations to examine these claims in a simplified setting.

Consider a joint distribution of potential outcomes and a single covariate $(Y(0), Y(1), X) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0, 0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.3 & 0.3 \\ 0.3 & 10 & 5 \\ 0.3 & 5 & 10 \end{pmatrix}.$$

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Your inferential target is $\mathrm{E}\left[Y(1) - Y(0)\right]$, or the "population average treatment effect." Note that, in this example, $\mathrm{E}\left[Y(1) - Y(0)\right] = 0$.

You conduct a randomized experiment by independently drawing n units from the joint distribution $(Y(0), Y(1), X)$ and randomly assigning np of these units to treatment $Z = 1$ and the remaining $n(1-p)$ to control $Z = 0$. You then observe the triple (Y_i, Z_i, X_i) for all n units, where $Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$.

Consider the following two estimators of $\mathrm{E}\left[Y(1) - Y(0)\right]$:

- $\hat{\beta}^A$, the estimate of β^A from fitting the model $Y_i = \alpha^A + \beta^A Z_i + \epsilon_i$ with OLS. Note that $\hat{\beta}^A$ is logically equivalent to the difference-in-means between outcomes in treatment vs. control.
- $\hat{\beta}^B$, the estimate of β^B from fitting the model $Y_i = \alpha^B + \beta^B Z_i + \gamma^B X_i + \epsilon_i$ with OLS.

We want you to assess the operating characteristics of $\hat{\beta}^A$ and $\hat{\beta}^B$ using simulations. Use at least 2500 simulations to calculate your answers. Because your inferential target is the population average treatment effect, note that you should *not* condition on having sampled a particular set of n units from $(Y(0), Y(1), X)$.

11. Suppose that n = 10 and p = 0.2. What is the root-mean-squared-error (RMSE) of $\hat{\beta}^A$? What is the RMSE of $\hat{\beta}^B$?
12. Suppose that n = 10 and p = 0.5. What is the RMSE of $\hat{\beta}^A$? What is the RMSE of $\hat{\beta}^B$?
13. Suppose that n = 1000 and p = 0.2. What is the RMSE of $\hat{\beta}^A$? What is the RMSE of $\hat{\beta}^B$?
14. Suppose that n = 1000 and p = 0.5. What is the RMSE of $\hat{\beta}^A$? What is the RMSE of $\hat{\beta}^B$?
15. Discuss your findings.