Quantitative Empirical Methods Exam

Yale Department of Political Science, January 2018

You have seven hours to complete the exam. This exam consists of three parts.

Back up your assertions with mathematics where appropriate and 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. Questions will not be weighted equally, and a holistic score will be assigned to the exam, and thus it is important to demonstrate your understanding of the material to the best of your ability.

Part 1 (Short Answer Section) consists of seven short answer questions. *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. Starred subquestions are especially difficult.

Part 2 (Essay Section) contains a recent, well-regarded empirical article. We will ask you to offer an evaluation of its methodological approach and presentation of results. In particular, we will advise you to pay particular attention to the identification conditions (either explicit or implicit), the associated estimation strategy, and possible threats to inference. Your response may be anywhere from 500 to 1500 words.

The only aids permitted for Parts 1 and 2 are (i) one page of double-sided notes, (ii) a word processor on one of the Statlab computers to write up your answers (you may also write up your answers to Part 1 using pencil/pen and paper). After handing in your answers for Parts 1 and 2 of the exam, you may begin Part 3 (though feel free to look ahead). You may hand in Parts 1 and 2 whenever you wish, but we recommend spending no longer than five hours on Parts 1 and 2.

Part 3 (Computer Assisted Section) will involve using statistical software to answer one longer exercise. A complete answer to Part 3 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 3, 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 3 is that you may not interact with anyone, online or otherwise. For Part 3 (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

There are a total of seven questions in the Short Answer Section. Starred (*) subquestions are especially difficult, and may be skipped.

1. Prove or disprove the following claims.

   (a) If $\operatorname{Cov}[X, Y] = 0$, then $\operatorname{Var}[X] = \operatorname{Var}[Y] = 0$.

   (b) If $\operatorname{Var}[X] = \operatorname{Var}[Y] = 0$, then $\operatorname{Cov}[X, Y] = 0$.

   (c) If $\operatorname{E}[X] = \operatorname{E}[Y] = 0$, then $\operatorname{E}[XY] = 0$.

   (d) If $\operatorname{E}[XY] = 0$ and $\operatorname{E}[X] = 0$, then $\operatorname{Cov}[X, Y] = 0$.

   (e) If $\rho[X, Y] = 0$ and $\rho[X, Z] = 0$, then $\rho[Y, Z] = 0$.

   (f) If both $\operatorname{Var}[X] > 0$ and $\operatorname{Var}[Y] > 0$, then $\operatorname{Var}[XY] > 0$.

2. Suppose that you collected 100 observations that are i.i.d. (X, Y). Of these,

   - 40 observations take on the value $(X = 0, Y = 1)$.
   - 20 observations take on the value $(X = 0, Y = 0)$.
   - 20 observations take on the value $(X = 1, Y = 1)$.
   - 20 observations take on the value $(X = 1, Y = 0)$.

   Estimate a 95% confidence interval for $\operatorname{E}[Y|X = 1] - \operatorname{E}[Y|X = 0]$ under a normal approximation.

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

$$\operatorname{E}[Y|X, Z] = 1 + 3XZ + X^2.$$

   Further assume that X and Z are independent and each distributed $U(0, 2)$.

   (a) What is the marginal effect of X on Y when $X = 1$ and $Z = 2$?

   (b) What is the marginal effect of Z on Y when $X = 0$ and $Z = 2$?

   (c) What is the marginal effect of X on Y when both X and Z are at their means?

   (d) What is the average marginal effect of X on Y? (*)

4. Assume the following data generating process for some outcome y_i:

$$y_i = \beta x_i + \epsilon_i$$

where ϵ_i is i.i.d. $N(0,1)$, and x_i is *non-random*. We observe n draws from this DGP, $(y_1, x_1), ..., (y_n, x_n)$. Consider three estimators of β:

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

$$\hat{\beta'} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

$$\hat{\beta''} = \sum_{i=1}^n \frac{y_i}{x_i}$$

(a) Prove that all three are unbiased estimators of β.

(b) Derive the variance of each.

(c) Prove that $\hat{\beta}$ is the MLE for β. (*)

5. For random variables X and Y with $\text{Var}\left[X\right] > 0$, prove that the minimum MSE linear predictor of Y given X is $g(X) = \alpha + \beta X$, where

$$\alpha = \text{E}\left[Y\right] - \frac{\text{Cov}\left[X,Y\right]}{\text{Var}\left[X\right]}\text{E}\left[X\right],$$

$$\beta = \frac{\text{Cov}\left[X,Y\right]}{\text{Var}\left[X\right]}.$$

6. Consider the following study. Z indicates assignment to treatment, D indicates receipt of treatment, and Y is a binary outcome. Table entries are the number of subjects of each type. Assume that observations are i.i.d. (Z, D, Y).

	$Z=1$	$Z=0$
$D=1, Y=1$	170	50
$D=0, Y=1$	20	100
$D=1, Y=0$	30	30
$D=0, Y=0$	80	120

(a) Enumerate the assumptions for the LATE (CACE) theorem.

(b) What evidence (if any) does the table provide for each assumption?

(c) Estimate the LATE (CACE) using the Wald IV estimator.

(d) Articulate a set of assumptions under which the LATE (CACE) equals the ATE.

(e) Using the assumptions of the LATE (CACE) theorem, compute sharp bounds on the ATE. (*)

7. A paper includes the following regression table, computed using ordinary least squares and robust standard errors. It reports the results from a regression conducted on a crosssection of 192 countries. The Polity IV Score is the DV; logged foreign aid (in current USD) and a dummy variable for whether or not the country is English speaking are the regressors. You do not need to add, subtract, multiply, or divide: e.g., "$0.533 - 5 \times (1.444 + 3.023)$" would be an acceptable form of answer.

Dependent Variable: $PolityIVScore$			
Specification:	(1)	(2)	(3)
$ForeignAid$	0.437 (0.115)	1.391 (0.983)	0.652 (0.139)
$(ForeignAid)^2$		-0.025 (0.028)	
$(ForeignAid)^3$		0.004 (0.008)	
$SpeaksEnglish$	4.650 (0.292)	4.617 (0.307)	21.851 (3.111)
$(SpeaksEnglish) \times$ $(ForeignAid)$			-0.884 (0.167)
Intercept	-5.562 (2.296)	-14.42 (8.584)	-9.828 (2.762)
n	192	192	192

(a) Under specification (1), what is the regression estimate of $\mathrm{E}\left[PolityIVScore|ForeignAid = 10, SpeaksEnglish = 1\right]$?

(b) Under specification (2), what is the regression estimate of $\mathrm{E}\left[PolityIVScore|ForeignAid = 10, SpeaksEnglish = 1\right]$?

(c) Under specification (3), what is the regression estimate of $\mathrm{E}\left[PolityIVScore|ForeignAid = 10, SpeaksEnglish = 1\right]$?

(d) Under specification (1), estimate a 95% confidence interval for $\partial\mathrm{E}\left[PolityIVScore|ForeignAid, SpeaksEnglish\right])/\partial(ForeignAid)$ under a normal approximation.

(e) Under specification (2), compute the regression estimate of $\partial(\mathrm{E}\left[PolityIVScore|ForeignAid, SpeaksEnglish\right])/\partial(ForeignAid)$ when $ForeignAid =$ 10.

(f) Under specification (3), compute the regression estimate of $\partial(\mathrm{E}\left[PolityIVScore|ForeignAid, SpeaksEnglish\right])/\partial(ForeignAid)$ when $SpeaksEnglish$ = 0.

(g) Under specification (3), is the coefficient on $(SpeaksEnglish) \times (ForeignAid)$ statistically significant at the $p < 0.05$ level (using the standard normal approximation-based significance test)? How do you know? What does it mean for this coefficient to be statistically significant?

2 Essay section

Read the article attached to your exam. Offer a critical evaluation of its methodological approach and presentation of results. Note: "critical" does not imply that you should only criticize — it is recommended that you give credit to the authors when their arguments are convincing and/or novel with respect to standard practice. Your response may be anywhere from 500 to 1500 words.

We advise you to pay particular attention to the identification conditions (either explicit or implicit), the associated estimation strategy, and possible threats to inference. Justify each of your claims and, where applicable, suggest ways in which this line of research might be improved. (We do not expect you to have special expertise in the topic area, but we do expect you to bring to bear your general analytical skills as a political scientist).

A special request: in your answer, please incorporate a path diagram (i.e., a directed graph, though it need not meet all of the requirements of a DAG) representing the causal theory or theories of the causal relationships under consideration in the assigned article. This may be hand-drawn. For which edges in the path diagram does the article provide an estimate?

Article: Jens Hainmueller, Barbara Hofmann, Gerhard Krug, and Katja Wolf. 2016. Do Lower Caseloads Improve the Performance of Public Employment Services? New Evidence from German Employment Offices. *Scandinavian Journal of Economics*. 118(4): 941–974.

3 Computer Assisted Portion

In this section, you will compare the performance of three predictive regression estimators using k-fold cross validation. (Be sure to set a seed!)

There is a dataset available at
`https://www.dropbox.com/s/c5vh6oq8vj1xnp7/exam_data.csv?dl=1`

This dataset consists of 3 columns (Y, $X1$, $X2$) and 500 rows corresponding to observations. The goal is to predict Y using $X1$ and $X2$.

We would like to predict Y given $X1$ and $X2$ using:

1. OLS: Ordinary Least Squares Linear Regression
2. Logit: Logistic Regression, and
3. BART: Bayesian Additive Regression Trees, using the `R` package `dbarts`'s default settings.

(For OLS and Logit, include $X1$ and $X2$ linearly — i.e., no interactions, polynomials, etc. Similarly, do not transform or otherwise preprocess the data for any of the estimators.)

Using $k = 20$ folds, we could like you to answer:

1. Which estimator (OLS, Logit, or BART) has the lowest k-fold cross validation estimate of RMSE?
2. Briefly explain your analysis and results. What, if anything, can you infer about the data from your results?

Note: We expect no existing familiarity with k-fold cross validation or BART, and you are encouraged to consult online resources, including Wikipedia, `R` documentation, and example code.