Political Methodology Comprehensive Examination, May 2021
Department of Political Science, George Washington University

Instructions: *Read all questions before answering any of them. When you use substantive examples in your answers, we strongly prefer to see examples from political science. Answer all questions in part I. Answer 3 questions in part II (but note the game theory question and full special topics question each count for two questions). If you are completing the exam at GW, feel free to hand-write answers or parts of answers in a blue book, but carefully label them and note that you are using the blue book in your typed document. If you are completing the exam at home, you can similarly include photos/scans of hand-written material. Good luck!*

Part I

1. A colleague of yours explains they ran a survey experiment last year with 540 respondents and didn't find any significant results. However, because they are certain their theory is right, they reran the experiment with another 1,000 respondents and then another 1,000 and then another 1,000, until finally they were able to report some significant findings. You mention some concerns with this procedure, but your colleague points out they now simply have the most information on which to base their conclusions. What are some of the issues with what your colleague did? Is this procedure, repeated enough times, likely to return significant results and why?

2. You're running a simple regression with a continuous dependent variable and a single continuous explanatory variable. In a cold sweat, you realize a typo added an extra 10 to one of your Y values, but you can't figure out which one. Colleague A tells you it's no big deal—all of the Y values are just averaged when running OLS, so it doesn't matter which observation was affected and it should be easy to adjust. Colleague B disagrees—it matters which observation is affected and it will be difficult to figure out the correct slope coefficient. Who's right and why? If A is correct, how do you adjust? If B is correct, how will the effect vary?

3. Who voted for the Nazis? In a recent study, researchers attempted to answer this question by analyzing aggregate election data from the 1932 German election during the Weimar Republic. The goal of analysis was to investigate which types of voters (based on their occupation category) cast ballots for the Nazis. The table below shows the variable names and descriptions. Each observation in the analysis represents a German precinct.

Name	Description
`shareself`	Proportion of self-employed potential voters
`shareblue`	Proportion of blue-collar potential voters
`sharewhite`	Proportion of white-collar potential voters
`sharedomestic`	Proportion of domestically employed potential voters
`shareunemployed`	Proportion of unemployed potential voters
`nvoter`	Number of eligible voters
`nazivote`	Number of votes for Nazis

One hypothesis says that the Nazis received much support from blue-collar workers. The output below is from a linear regression where the overall Nazi vote share is regressed on the proportion of blue-collar voters.

```
##
## Call:
## lm(formula = sharenazis ~ shareblue, data = nazis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30151 -0.07133 -0.00092  0.06986  0.33037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39558    0.01661  23.812   <2e-16 ***
## shareblue    0.06518    0.05220   1.249    0.212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.108 on 679 degrees of freedom
## Multiple R-squared:  0.002291,   Adjusted R-squared:  0.0008218
## F-statistic: 1.559 on 1 and 679 DF,  p-value: 0.2122
```

a) Compute the 95% confidence interval of the estimated slope coefficient and provide a substantive interpretation of these quantities.
b) When used to make conclusions about how individuals vote, what does this model assume with regards to the Nazis' expected vote share among blue-collar voters and the proportion of blue-collar voters in the precincts?

In order to investigate which occupation categories were more likely to cast ballots for the Nazis, the authors estimated a linear regression where the overall Nazi vote share is regressed on the proportion of each type of occupation. The output is presented below.

```
##
## Call:
## lm(formula = sharenazis ~ -1 + shareself + shareblue + sharewhite +
##     sharedomestic + shareunemployed, data = nazis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28271 -0.06847 -0.00055  0.06790  0.32369
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## shareself       1.11426    0.16677   6.681 4.95e-11 ***
## shareblue       0.54038    0.03848  14.042  < 2e-16 ***
## sharewhite      0.28509    0.07501   3.801 0.000157 ***
## sharedomestic   0.05221    0.09120   0.572 0.567181
## shareunemployed -0.02816    0.07014  -0.401 0.688202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1024 on 676 degrees of freedom
## Multiple R-squared:  0.9435, Adjusted R-squared:  0.9431
## F-statistic:  2259 on 5 and 676 DF,  p-value: < 2.2e-16
```

c) What does the F-statistic tell you?
d) Note that this model does not contain an intercept. What are the implications of omitting the intercept in this case? Assuming all potential voters fall into one of these five categories, what would the results be if an intercept was included?
e) Provide a substantive interpretation of the R-squared and the Adjusted R-squared.
f) Compute the 95% confidence interval of each significant coefficient, and provide a substantive interpretation of each coefficient and its confidence interval.
g) What assumptions are necessary to permit the interpretation of these coefficients?
h) Why do the results for `shareblue` differ across the two regressions? Without making additional modeling assumptions, what can we learn about blue-collar support for the Nazis?
i) A reviewer points out that some of the coefficients do not quite make sense because the estimates are outside of the range. What could be causing this?
j) Another reviewer asks whether interaction terms might be appropriate here. Is this valid? Give an example of an interaction term that could be used and explain how it might be interpreted.

Part II

1. Suppose you are interested in assessing whether incarceration causally affects political participation in the United States. To minimize concerns about selection bias and measurement error, you decide to use administrative records. By merging sentencing data to voter records, you expect to compare formerly incarcerated individuals to individuals convicted of crimes but not incarcerated. In other words, your empirical strategy is based on comparing the participatory patterns of these two groups of convicts. This would allow you to hold constant the fact that everyone in this sample is convicted of a crime. How would you estimate the effect of serving time in prison on voting for multiple elections? What are the required assumptions for this design to be valid? Discuss potential limitations and threats to the validity of this design.

2. Synthetic control and generalized synthetic control are two alternatives to the difference-in-differences model. What specific assumptions or problems are these alternatives aimed at addressing? Give some specific examples where they would be superior to difference-in-differences. What do you regard as the remaining limitations or major assumptions that the alternatives cannot help with?

3. Imagine you are in the audience in a job talk where the job candidate is advancing a theory linking economic inequality to increased political support for illiberal political parties. To provide evidence for the argument, the candidate presents qualitative case study evidence from two countries, which were selected because they are countries where support for illiberal parties has grown substantially in recent years. A professor in the audience argues that the research design is flawed because the candidate has "selected on the dependent variable." What are the merits of this critique? How should the job candidate respond? What arguments could they make and evidence could they provide in defense of their research strategy?

4. Because of attrition or survey item non-response, experimental (or survey-based panel) studies often have missing outcomes information for some subjects/units. What are the potential consequences of this situation? What would you do to assess the severity of the attrition

problem? What strategies can you use to account for this issue in your analyses and what are their strengths and limitations?

5. You have used non-probability sampling to draw a survey sample. Your goal is to make inferences about the national population of adults. The left column in the table below shows the percentage of the target population in different age and education categories (from a population census). The right column shows the percentage of your sample that falls into each age and education category. What issues or biases might arise when trying to make an inference about the population from this sample? How specifically would you use weighting to make your sample "look more like" the target population? Be specific about the weights you would construct and the precise steps that you would take.

	Population (census)	**Your Sample**
AGE		
18-35	28.8	19.0
36-50	21.3	20.9
51-64	29.8	31.8
65+	20.1	28.4
EDUCATION		
No High School	6.8	2.8
High School grad	30.6	19.7
Some college	23.0	15.7
2-year	10.6	11.3
4-year	18.7	28.6
Post-grad	10.4	21.9

Special Topics

Note: You have the option of answering the first three parts (a-c) only, which will count as one question. If you answer all parts, it will count as two questions.

A student is examining what country-level factors predicted a Biden win in the 2020 election. She performs a lasso regression in R and estimates the optimal parameter λ by 10-fold cross-validation. She uses the following code:

```
> library(glmnet)
> X = as.matrix(data[,-1])
> Y = as.vector(data[,1])
> lambdas = 10^seq(3, -2, by = -.1)
> cv.out = cv.glmnet(X,Y,nfolds=10,alpha = 1)
> cv.out$lambda.min
[1] 0.01
> min(cv.out$cvm)
[1] 0.10
```

(a) Why does the student use cross-validation?

(b) The student concludes that the test MSE of lasso regression with $\lambda = 0.01$ is approximately 0.10. Explain the problem with this estimate and suggest a better way to estimate the test MSE.

(c) The student finds the model shrinks coefficient estimates about the share of population that is African-American or Hispanic to zero, but keeps variables about county-level median income (-0.06) and college graduation rates (0.13). What does this tell us about the correlates of a Biden win?

(d) After fitting the model on the dataset, the student predicts how well the model fits the data. She assesses the model's accuracy by transforming the predicted values to 1 if the Pr(Biden Win) > 0.5 and zero otherwise. She gets the following confusion matrix:

```
> confusionMatrix(ifelse(lasso.pred > 0.5, "1", "0"), data$bidenwon)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2441  312
         1   62  295
```

(i) What is the overall accuracy of the model? What is the NIR?

(ii) What is the sensitivity and specificity of the model?

(iii) Is this a good model? If so, how do you know? If not, how could the student improve the model?

Game Theory: Counts as two questions

Consider an activist who wants a dictator to implement a political reform. The activist comes in three types: Radical, Moderate, and Quiet. The dictator's prior beliefs over these types are given by q_R, q_M, and $q_Q = 1 - q_R - q_M$. The order of the game is as follows:

1. The activist chooses to protest or not at cost c.
2. The dictator implements the reform or not.
3. The activist chooses to launch a revolution or not, at cost d to both players and with likelihood of success p.

The payoffs are such that Radical types will revolt no matter what. Quiet types will never revolt, but prefer getting the reform. Moderate types will revolt if and only if the reform is *not* granted (i.e., the reform satisfies them). Implementing the reform costs the dictator 1, with $d > 1$. The dictator also gets benefit W from ruling and 0 otherwise. If a revolution is attempted, the activist's payoff does not depend on whether the reform was granted (since they'll either be in charge or in jail), but assume the dictator still loses 1 by granting the reform.

(a) What is the total payoff to the dictator if they do not reform and face revolt? What is the total payoff to the dictator if they reform and avoid revolt?

(b) After seeing step 1, the dictator will update their beliefs on the type they are facing. Call the updated beliefs in step 2 q'_R, q'_M, and q'_Q. For what set of updated beliefs will the dictator implement the reform in step 2?

(c) What are the conditions for each type of activist to protest in step 1?

(d) Using (b) and (c), under what conditions is there a separating equilibrium? (This includes cases where two of the three types overlap, but the third does something different.)

(e) In the separating equilibrium, what is the probability that reform occurs (assuming the dictator's initial beliefs give the correct probabilities of each type)? What is the probability of revolt?

(f) How does the structure of signaling in step 1 and/or payoffs for the activist types need to change to get an equilibrium that is maximally beneficial for the dictator?