Political Methodology Comprehensive Examination, May 2016
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. Question GT in part II counts as two questions. After the exam turn in a blinded empirical paper to Taylor demonstrating your ability to use statistical models* OR *schedule an oral exam. Good luck!*

Part I

1. After showing a scatterplot exhibiting a positive linear correlation between a pretest X and a posttest Y ($r \approx .65$), a presenter comments that the lowest decile observations on the X variable increased more on average than the highest decile observations on the X variable. He proceeds to give an elaborate explanation as to why this is substantively important. Can you propose a simpler statistical explanation for his "finding"? Explain. [*Hint:* you can answer this in two sentences. Be brief.]

2. To investigate the effects of international agreements on repression, an author models repression as a function of three types of international agreements (and other variables; this is from a published article). The three types of agreements were human rights agreements (HRAs), "soft" preferential trade agreements (PTAsoft), and "hard" preferential trade agreements (PTAhard), each lagged one year. The estimated coefficients and standard errors on these three variables are as follows: $\hat{\beta}_{HRA} = .082, SE = .07$; $\hat{\beta}_{PTAsoft} = -.265, SE = .16$; $\hat{\beta}_{PTAhard} = -.255, SE = .12$. Given these estimates, the author concludes the following: "State commitments to comply with HRAs and soft PTAs do not systematically lead to decreasing repressive behavior in the following year. Quite the contrary, HRA and PTAsoft have no significant effects on the likelihood of repression. When PTAs contract hard human rights standards, by contrast, member states are systematically more likely to decrease repression." Do you agree with the conclusions reached by the author? Why or why not? Are there any additional tests you would like to perform? Are there additional substantive claims to be made about the results? [As in #1, keep your answer short. Focus on the interpretation of the estimates.]

3. After the game theory question in part II, you will find a section labelled **Fundamentals (Part I, question 3)**. It contains variable definitions, Stata output, and a series of questions. Answer those questions as concisely as possible.

Part II

1. An article (this is based on a real paper) seeks to test the effect of regime type (democracy vs. autocracy) on child health. It does this using a survey of women in developing countries, with individuals as the unit of analysis. Unfortunately, the survey was not completed every year in each country. Table 1 below shows the country, year, regime type, and number of women (in each country-year) making up the sample. The empirical model uses OLS to predict a continuous measure of child health, with the country's regime type (a dummy variable for democracy) as the main variable of interest. To account for country heterogeneity, country fixed effects are included. You can ignore the role of other control variables.

Describe the advantages and limitations of this empirical design. If the authors claim this design is superior to the typical country-year setup because it has more than 1 million data points, what would your response be?

Table 1: Data structure for Part II, question 1

	Kenya	**Nigeria**	**Cameroon**	**Mali**	**Swaziland**	**India**	**Nepal**	**Honduras**	**Burma**
Year	**1999**	**2003**	**1998**	**1999**	**1997**	**2000**	**1999**	**2000**	**2001**
Democracy	0	0	0	1	0	1	0	1	0
N	101,897	45,675	78,889	49,992	75,678	120,381	15,459	83,893	45,349
Year	**2004**	**2007**		**2005**	**2001**	**2005**			**2006**
Democracy	1	0		1	0	1			0
N	89,785	56,901		34,542	78,943	118,093			83,451
Year					**2004**				
Democracy					0				
N					81,980				

What is an equivalent way to set up this model and get the same results regarding regime type? How could this design be improved?

2. Prior to estimating a logit model with a dichotomous dependent variable Y, you discover that one of your dummy independent variables D takes on the value of 1 for all observations where $Y = 1$. Will that cause a problem in estimating the effect of D on Y? If not, why not? If so, what can you do about the problem?

3. Briefly explain the homophily principle and its implications for the analysis of political networks. How does the principle complicate causal inference in network data? What solutions are available for distinguishing the effects of homophily from the effects of network structures? If you illustrate with examples, use examples of political phenomena.

4. You want to analyze some data with the following properties. The dependent variable can be considered a count variable, and a large proportion of the observations are zeroes. The main independent variable of interest is likely to be endogenous Explain the steps you would take in selecting a model. Can you deal with the count variable, the excess zeroes, and the endogeneity all in a single model? If not, what will guide your model choice? What are some of the tradeoffs or shortcomings from choosing your recommended model? Would there be auxiliary analyses that you might present in an appendix? How would you decide what belongs in the main text of your report versus in an appendix?

5. Suppose a mysterious stranger gives you a dataset. She wants you to figure out if the first variable in the dataset is caused by the second. Other variables are potentially useful as controls or in estimation. However, you're not told what any of these variables are and the variable names are inscrutable. What are the limits to doing causal inference in this situation? Are there any conclusions one can draw (perhaps conditional on assumptions) or causal inference techniques that could possibly apply?

GT: counts as two questions

Consider an activist who wants a dictator to implement some 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. Implementing the reform costs the dictator 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) Call the updated beliefs of the dictator 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? 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?

Fundamentals (Part I, question 3)

For this question, use the OLS and logistic regression output below. The data are from the 2016 American National Election Study pilot survey. The observations are Democratic identifiers (including Democratic-leaning independents). "R" denotes survey (R)espondents.

The variables used below include:
ClintonFT: Feeling thermometer for Hillary Clinton, scale of 0 to 100
sup_hil_not_sanders: Vote intention in Democratic primary 1 if Clinton, 0 if Sanders
Gender Discrmin: Whether or not the R feels they have personally experienced a lot of gender discrimination, 0 if no, 1 if yes
Local_terror_worry: Whether or not the R worries a lot about a terrorist event occurring in their local area, 0 if no, 1 if yes
Minwage: R's opinion of the minimum wage on a 4 point scale (1=should be raised, 2=kept the same, 3=lowered, 4=eliminated)
Getahead= How much opportunity R sees in America today for the average person to get ahead (1=none, 2=a little, 3=a moderate amount, 4= a lot, 5=a great deal)
Femoff_issues: R's assessment of how much female elected officials are likely to focus on issues that mainly affect women (1=a great deal more likely to focus on women, 2=moderately more likely to focus on women, 3=a little more likely to focus on women, 4=no more likely to focus on men ot women, 5=a little more likely to focus on men, 6= moderately more likely to focus on men, 7=a great deal more likely to focus on men)
Follow: how much R follows politics on a 4 point scale (1=Most of the time, 2=some of the time, 3=only now and then, 4=hardly at all)
Women: 1 for women Rs, 0 for men Rs
Black: 1 for Rs that chose black as their race, 0 otherwise
Hispanic: 1 for Rs that chose Hispanic as their race, 0 otherwise
otherRace: 1 for non-white/non-black/non-hispanic Rs (Asian, mixed, other), 0 otherwise

Source	SS	df	MS		
				Number of obs	= 553
				#4 F(10, 542)	**= 5.82**
#6 Model	**39065.828**	10	3906.5828	Prob > F	= 0.0000
Residual	363866.917	542	671.341175	R-squared	= 0.0970
				#5 Adj R-squared	**= 0.0803**
Total	402932.745	552	729.950625	Root MSE	= 25.91

ClintonFT	**#1 Coef.**	**#2 Std. Err.**	t	P>|t|	**#3[95% Conf.**	**Interval]**
women	1.315898	2.249748	0.58	0.559	-3.103397	5.735192
Gender Discrimin	-.5011863	3.290931	-0.15	0.879	-6.965728	5.963355
local_terror_worry	2.523558	2.58036	0.98	0.329	-2.545173	7.59229
minwage	-3.785915	2.150771	-1.76	0.079	-8.010782	.438953
getahead	4.845829	1.198763	4.04	0.000	2.491038	7.20062
femoff_issues	.4545297	.3883526	1.17	0.242	-.3083309	1.21739
follow	-3.555783	1.314837	-2.70	0.007	-6.138584	-.9729816
black	13.9804	3.035169	4.61	0.000	8.018263	19.94254
hispanic	-1.475287	3.735297	-0.39	0.693	-8.81272	5.862146
otherRace	.9659694	4.508633	0.21	0.830	-7.890566	9.822505
_cons	60.70036	5.029369	12.07	0.000	50.82092	70.5798

Logistic regression

Log likelihood = -284.61308

Number of obs	=	497
#7 LR chi2(10)	**=**	**114.92**
Prob > chi2	=	0.0000
Pseudo R2	=	0.1680

sup_hil_not_sanders	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
women	.2426956	.2077314	1.17	0.243	-.1644505	.6498418
self_disc_gender_LOT	-.2257896	.3209315	-0.70	0.482	-.8548038	.4032246
loc_terror_worry	1.246239	.2654065	4.70	0.000	.7260513	1.766426
minwage	.6168494	.269273	2.29	0.022	.0890841	1.144615
getahead	.4929081	.122362	4.03	0.000	.2530829	.7327333
femoff_issues	-.0452006	.0360809	-1.25	0.210	-.115918	.0255167
follow	-.064548	.1269065	-0.51	0.611	-.3132802	.1841842
black	1.66255	.3172365	5.24	0.000	1.040778	2.284323
hispanic	.5752656	.3592221	1.60	0.109	-.1287969	1.279328
otherRace	-.0170922	.389938	-0.04	0.965	-.7813567	.7471722
_cons	-2.051892	.5113486	-4.01	0.000	-3.054117	-1.049667

.

a. (For each emboldened item in the output (1-7), give a mathematical formula for how it is computed and **briefly** explain its meaning/interpretation, being sure to note what, if any, population parameter it is meant to estimate. (Note that the items are numbered—use that numbering for the order of your answers!)

b. Now suppose that not all of the Gauss-Markov (CRLM) assumptions hold. In particular, the data are characterized by heteroskedasticity, and it is a function of the Xs. Nevertheless, you estimated your model with OLS. For each of the six emboldened items on the OLS regression output, explain the implications. Be sure to note whether or not you would

expect a different value (as compared to the value you'd expect if the Gauss-Markov assumptions held) and if so, where that change in value would come from. Also be sure to mention statistical properties (bias, consistency, etc.), where applicable, when estimating via OLS under these data conditions.

c. Suppose you have reason to believe (a theory!) that the heteroskedasticity you worried about in part b was a function of race. You conduct multiple formal statistical tests for heteroskedasticity and find support for it (i.e., you reject the null of homoscedasticity). What would support for that theory mean and what would be your next step?

d. Using the OLS regression model, how would you test the hypothesis that "race doesn't matter to Americans' evaluations of Hillary Clinton"? If you can test the hypothesis from this output alone, do so (set-up/report/interpret). If you cannot, explain why not and what else you'd need to know. [Go back to assuming the Gauss-Markov assumptions hold for this question.]

e. All else equal, what is the expected difference in Clinton ratings between black women and white men?

f. You present the results of these models at a panel about Election 2016. One audience member gets up and says "Well, obviously your model of voter preference is much better than your model of Clinton evaluations. Look at those p-values and [pseudo] R-squareds!" What is the audience member talking about and is he justified? Are there any cautions or lessons you might want to share with this audience member?

g. Another audience member wants to dismiss your models entirely because you failed to account for voters' attitudes about government policies to address gender discrimination. "That must be important to Clinton voters," he argues. You reply to the critic that you're not concerned about that, because previous research has shown unified support for such policies among Democrats. Explain the problem the audience member was claiming you had, and how your response was addressing it.

h. Another audience member asks whether you "considered the argument that the men who are Sanders supporters ("Bernie Bros") are far more likely to value his position on the minimum wage? That these guys are especially drawn to the economic argument?" Did your model do that? If so, report and discuss the relevant results. If not, write down an amended model that would consider the argument the audience member raised. Discuss what information from that model (including any necessary tests) you would use to answer the audience member's question.

i. One more audience member comments, "I have concerns about your dependent variable. I just don't think that people are capable of reporting to you exactly where their feelings about Clinton fall on a 0 to 100 thermometer scale." You reply, "I agree to some extent – the evidence suggests that they really only have a general placement on the scale, and that they simply guess at exactly which number to report." You don't seem too worried. Explain.

j. Yet another grumpy audience member gets up… "I hope you don't think you can dismiss gender as important here. I know your dummy variable is insignificant. But the effect of

gender isn't simply about women and men categorically disagreeing on Clinton. It's buried in their attitudinal and experiential differences. Men are less likely to experience discrimination, are less supportive of the minimum wage, and are much more likely to believe Americans can still get ahead with hard work. And ***that*** matters." Wow. Explain what this audience member is arguing and how you could shed empirical light on her argument. Be sure to note what you can say with just the output here and what other information you would need and how you would use it.

k. One of the reporters at this panel is actually impressed by your presentation and would like you to comment on how attitudes about the minimum wage will affect the general election in November. Would you feel comfortable commenting? Why or why not? If so, what kind of comment would you make?