Methods Exam, 2013

The time for the exam is eight hours. You have from 8:00 a.m. to 5:00 p.m. The exam is open note and open book. You are not allowed to collaborate.

Math Foundations:

1.

Using two different methods, find the inverse of the following matrix.

$$\begin{bmatrix} 1 & 2 & 1 \\ 5 & 7 & 0 \\ -1 & -1 & 0 \end{bmatrix}$$

2. In words, explain the conceptual differences between an antiderivative and a definite integral.

3.

For each function, find its derivative and antiderivative.

a)

$$f(x) = x$$

b)

$$f(x) = e^{-x}$$

c)

$$f(x) = -\frac{1}{15}$$

d)

$$f(x) = x^{-1}$$

e)

$$f(x) = 2x^{\frac{3}{2}}$$

f)

$$f(x) = 4\sqrt{\frac{x}{7}}$$

g)

$$f(x) = -x^2 + 2x + 10x^3$$

Probability & Statistics Questions:

1. Consider a loaded five-sided die that comes up 1 with probability $\frac{1}{2}$, 2 with probability $\frac{1}{5}$, 3 with probability $\frac{1}{5}$, 4 with probability $\frac{1}{20}$, and 5 with probability $\frac{1}{20}$.

a. What is the expected value of a roll of the die?

b. What are the variance and standard deviation of a roll of the die?

2. Suppose $Cov(X, Y) = 4$ and $Q = 12 + 3X$. What is $Cov(Q, Y)$?

3. Consider the joint probability density function for two random variables X and Y: $f(x, y) = x + y$ for $0 < x < 1$ and $0 < y < 1$ and $f(x, y) = 0$ otherwise. Are X and Y independent? Show your work.

4. Let U_i be a sequence of Uniform random variables with support between 0 and 2. Each of the U_i are independent of one another. Define $Y_i = 0$ if $0 < u_i \leq 1$, $Y_i = 1$ if $1 < u_i \leq \frac{3}{2}$, and $Y_i = 2$ if $\frac{3}{2} < u_i \leq 2$. Consider the sum of 50 realizations of Y_i, $\sum_{i=1}^{50} Y_i$. Use the central limit theorem to find an approximation for the probability that $\sum_{i=1}^{50} Y_i$ is greater than 35. Show your work.

5. Suppose a missile detector has 5 percent false positives and 8 percent false negatives. That is, 5 percent of the time the detector will indicate that a missile has launched when no missile has launched and 8 percent of the time it will indicate that no missile has launched when a missile has launched. Suppose that 95 percent of the time no missile is launched. Conditional on observing the detector indicate a missile launch, what is the probability that a missile was actually launched?

6. a. You observe a sample of n independent observations, $X_1, X_2, ...X_n$ from a Uniform distribution with support $[0, \theta]$. θ is an unknown parameter that you would like to estimate. Derive a Method of Moments estimator for θ.

b. Now you observe a sample of n observations, $Y_1, Y_2, ...Y_n$ from a Uniform distribution with support $[-\gamma, \gamma]$. γ is an unknown parameter that you would like to estimate. Derive a Method of Moments estimator for γ.

Core Material:

QUESTION 1

Consider the OLS regression output from Stata below. The observations are American adults in the 2004 National Election Study survey.

```
. reg  gwbush01 PID ideology bushwar

      Source |       SS       df       MS              Number of obs =     973
-------------+------------------------------           F(  3,   969) =  292.74
       Model |  51.5823226     3  17.1941075           Prob > F      =  0.0000
    Residual |  56.9138034   969  .058734575           R-squared     =  0.4754
-------------+------------------------------           Adj R-squared =  0.4738
       Total |  108.496126   972  .111621529           Root MSE      =  .24235

------------------------------------------------------------------------------
    gwbush01 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PID |   .6596368   .0223505    29.51   0.000     .6157759    .7034977
    ideology |   .0325418   .0181637     1.79   0.074    -.0031031    .0681866
     bushwar |   .0038069   .0199563     0.19   0.849    -.0353557    .0429695
       _cons |   .2103008   .0227522     9.24   0.000     .1656515      .25495
------------------------------------------------------------------------------
```

gwbush01 = Feeling thermometer for President George W. Bush, scaled to run from 0 (very cool affect/dislike) to 1 (very warm affect/like), with .5 meaning "neutral"

PID = Seven point party identification scale, running from 0 for strong Democrats to 1 for strong Republicans (categories are strong Republican, weak Republican, Independent that leans Republican, Independent, Independent that leans Democrat, weak Democrat, strong Democrat)

ideology = three point variable, where 0=liberal, .5=moderate, and 1=conservative

bushwar = four point variable indicating the degree of support for Bush's handling of the war in Iraq, where 0 is strongly disapprove, .33 is disapprove somewhat, .67 is approve somewhat, and 1 is strongly approve

a. Using this regression model, how would you test the hypothesis that "the war effort doesn't matter to Americans' evaluations of George W. Bush"? 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. [Assume CLRM holds for this question.]

b. All else equal, what is the expected difference in Bush ratings between weak Republicans and weak Democrats? Between moderates and conservatives? How confident are you in those predicted differences (and how do you know)?

c. You present the results from this regression at a conference and an audience member wants to dismiss them "because you forgot the economy." You reply to the critic that you're not concerned about the economy, because at the point at which the survey was conducted, the economy was booming and sentiment about the economy was high among all sorts of Americans. Explain the problem the audience member was claiming you had, and how your response was addressing it. Then explain how you would test the argument between you

and the audience member (what would you run, how would you decide on whose model is right, etc.?)

d. Another audience member asks whether you "considered the argument that political information or awareness should have caused a divide between partisans? That Democrats and Republicans really would have reacted differently to more information about Bush's handling of the war effort?" Did this regression 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 not immediately reported in Stata regression output) you would use to answer the audience member's question.

e. One more audience comment. "I have concerns about your dependent variable. I just don't believe that people are capable of reporting to you exactly where their feelings about Bush 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 sense of their placement on the scale, and that they simply guess at *exactly* which number to report." You don't seem too worried. Explain.

QUESTION 2

Explain the linearity assumption of the CLRM. Are there ways in which it is not as restrictive as it might seem?

QUESTION 3

What are the consequences of measurement error for OLS? Discuss possible solutions.

Advanced Material:

QUESTION 1 – Choose 2 of the following 4 questions and provide a CONCISE, ½ to 1 page response.

1. Brambor, Thomas, William Clark, & Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses" *Political Analysis*. 14: 63-82.

   In a multiplicative interaction model, the constitutive terms refer to the elements that constitute the interaction term(s). Briefly explain why an analyst should nearly always include all of the constitutive terms when specifying an interaction model. Some scholars argue that they omit constitutive terms to reduce multicollinearity issues. What do you make of this argument?

2. King, Gary, Michael Tomz, & Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" *American Journal of Political Science* 44: 341-355.

   In this article, King and his co-authors discuss how one might use statistical simulation methods to calculate quantities of interest and obtain more substantive information from non-linear empirical models. Explain exactly what these substantive quantities of interest might be, how the proposed simulation methods work, and what assumptions are being made. You may find it helpful to discuss these issues with specific reference to a particular model such as a probit/logit, count model, duration model etc..

3. Box-Steffensmeier, Janet M. & Christopher J. W. Zorn. 2001. "Duration Models and Proportional Hazards in Political Science" *American Journal of Political Science* 45: 972-988.

   Explain the source of the proportional hazards assumption in some duration models and what it means substantively. How do you identify violations of the proportional hazards assumption? What steps can you take to deal with violations of the proportional hazards assumption? Explain substantively why these 'solutions' work.

4. King, Gary and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis*. 9 (2):137-163.

   Assume two researchers come to you with datasets collected to study whether increased education leads to a greater chance of an individual switching party registration. The first researcher conducts a random sample of 2000 registered voters, and finds that only 3 percent of these report recently switching party. The second researcher conducts a stratified sample (using say voter registration files), sampling 500 party switchers, and then 500 observations from the population of non-switchers. Explain in each case what would be the problem of running a naive model predicting party switching (dependent) as a function of education (independent) without any corrections to the parameter estimates. Explain in each case an appropriate correction. After corrections, which dataset would you expect to have estimates with lower mean squared error?

**** This section has choices****
Directions – Choose 2 of the following # Questions

QUESTION 1
In event history analysis, often spells or episodes are left-censored, right-censored, or have repeated transitions. Describe the problems such spells create and where political research is likely to encounter them. Then evaluate solutions to these problems.

QUESTION 2
Internet surveys have become common. What advantages, if any, do Internet surveys have over other survey modes? Summarize any research you know about how similar the results of Internet surveys are to surveys taken through other modes. Discuss the limitations of Internet surveys from the Total Survey Error perspective. (In answering this question, be sure to indicate what type of Internet surveys you are critiquing.)

QUESTION 3
Experimental research is lauded for its superior "internal validity." What does this phrase mean, and how does internal validity differ from external validity? How is internal validity achieved? What is a true experiment anyway, and how does it differ from other varieties of experimental research/research designs (quasi-experiments, simulations)? What are some of the major threats to internal validity in social science research generally? Does improved internal validity trade off with other desirable properties of social scientific research? Please consider the relationship between the experiment itself (esp. the actual treatment) and the theoretical concept of interest.

QUESTION 4
Suppose that you are analyzing voting data with measures at both the individual and county level. Write the separate level-1 and level-2 equations as well as the combined multi-level model. What assumptions are necessary to estimate this model? How would you interpret a random intercept and random coefficient model? Besides a multi-level model, what other models could you estimate? Why is a multi-level model superior?

QUESTION 5
There are several different statistical estimation approaches available, including least-squares, maximum-likelihood, and Bayesian. Compare their assumptions and relative advantages and disadvantages. Under what circumstances would the approaches you choose give identical results? Which one is considered more general and why?

QUESTION 6
Explain the concept of the bootstrap. What does the bootstrap simulate? When can the bootstrap fail?

QUESTION 7

7. Find the solution of the difference equation for the given initial condition. Graph the solution sequence with time on the horizontal axis. Characterize the limiting behavior of the sequence.

$Y_{t+1} - 3Y_t = .10$ $\quad\quad$ $Y_0 = 1$

$4Y_{t+1} - Y_t = 2$ $\quad\quad$ $Y_0 = 4$

$Y_{t+1} + Y_t = 0$ $\quad\quad$ $Y_0 = -10$

$3Y_{t+1} + Y_t = 1$ $\quad\quad$ $Y_0 = 1$

$7Y_{t+1} + 7Y_t = 3$ $\quad\quad$ $Y_0 = 5$

Why are nonlinear difference equation important in the study of time series analysis?