PS699: Problem Set 1 -- Matrix Algebra and Calculus

For this and all subsequent problem sets: write your answer neatly on a fresh sheet of paper. Where assignments on the computer are given, strive to present the information necessary to answer the questions as neatly and clearly as possible for your GSI (guidelines will usually be given).

This assignment and several subsequent ones use the data set, "Political Structure in Developed Democracies," created for this class from the Political Data Handbook of OECD Countries by Jan-Erik Lane, David McKay, and Kenneth Newton. Your GSI will e-mail you on how to access the data. The Excel file also contains the definitions of the variables in a table below the data.

1) In your spreadsheet, create the data matrix of Scandinavian countries' as the rows/observations (in this order: Denmark, Finland, Iceland, Norway, Sweden) and the following as the columns/variables (in this order) Lower-house seats (LSEATS), Proportionality Index (LPROP), Number of Governments Postwar (NGOVPW), and Social Security Spending as a % of total government expenditure (SSSPEND). Call this matrix **A**. (You can do this problem by hand without too much difficulty, but you'll want to learn how to manipulate data matrices in a spreadsheet, so we are going to do it therein.)

[In the interest of conserving paper, and as good practice for the future (in terms of presentation of data and work), try to neatly arrange parts a-e on one page and parts f-h on no more than two pages.]

a) Print the data matrix **A**. Make sure it is labeled so an outsider would know what it is.

b) What are the dimensions of **A**? (write on that printout clearly & prominently, label it clearly)

c) What is its transpose, **A′**? what are the dimensions of **A′**? (transpose **A** in the spreadsheet & print it, clearly labeled; write dimensions clearly on that printout.) (Transpose is a "Paste Special" in Excel)

d) In matrix **A**, what are the positional coordinates of the number of governments in Sweden's postwar history? of LPROP in Denmark? (Highlight or circle these data & label their coordinates)

e) Using your spreadsheet's "column average" function to calculate the row vector of means, **m′**.(print out **m′**, label it clearly).

f) Use the spreadsheet to calculate **A**-**m**, call that **B**. (print **B**, label it clearly)

g) Use the spreadsheet to calculate (1/n)(**B′B**); call the result **V**. Neatly print out a page containing **B′, B, B′B**, and **V**. Write out the steps (including spreadsheet formulae) you employed to obtain the result on that page. Make sure each matrix is clearly labeled on the printout.

(Matrix multiplication in Excel is a pain as far as I can tell. Use the Lotus 1-2-3 mime. Go to "Help" menu, then "Lotus 1-2-3 Help..." menu, then "Data", then "Matrix", then "Multiply". It prompts for a first array: put one matrix in there; and for a second array: put the other in there; and for a destination: select a cell where you want the top-left corner of the resulting matrix. If you insist on the Bill Gates garbage, the command is something like "=mmult(A6..B10,C6..G8)" or something like that. Where A6..B10 is the range of your first matrix and C6..G8 is the range of your second. Put the formula in your top-left cell and copy it down. Something like that. Or try (for older excel versions): (1) highlight the range where the result will go (the whole and exact range), then type =mmult(range 1,range 2) where range 1 is the first matrix's range and range 2 is the second. Then hit CTRL+SHIFT+ENTER. Use Lotus or the Lotus commands, they're much easier.)

h) Do you recognize any statistical meanings for the elements of the solution to g? What are they?

2) In your spreadsheet, create two vectors, one equal to all 23 countries' observations on the average duration of governments in the postwar era, DGOVPW, and the other equal to the average number of parties in government in the postwar era, NPGOVPW. Subtract the mean of DGOVPW across these countries from every observation on DGOVPW; call the resulting vector **y**. Do the same for NPGOVPW, and call the resulting vector **x**. Use the scalar formula derived in class to determine the b coefficient on the **x** that minimizes **y** - b(**x**).

On one piece of paper:

a) Present a graph where you plot **y** against **x**. Put the line given by **y** = b**x** on this graph too. Label the graph appropriately.

b) Write the formula you used to obtain b.

c) Give a substantive interpretation of b.

d) Now create another vector, $\mathbf{x}_2$, equal to a (23 x 1) column of ones. Create a matrix **X**, by putting the column of ones left of the column NPGOVPW (without the means subtracted). Using the multivariate formula given in class, find the coefficient vector **b**, which minimizes the distance from (23x1) vector **y**, given by DGOVPW (without the mean subtracted), to a line given by **Xb**. Print out $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ (showing the vector **b**, then the "=", then $(\mathbf{X'X})^{-1}$, then $\mathbf{X'y}$. Notice anything about the answers to a)-c) and d)? Why did that occur?

3) In the notes, I proved the commutative and associative properties of matrix addition.
   a) Now prove the distributive property of transposition over addition: $(\mathbf{A}+\mathbf{B})' = \mathbf{A}' + \mathbf{B}'$.
   b) Using these three properties, prove that $[(\mathbf{A}+\mathbf{B})+\mathbf{C}]' = \mathbf{C}' + \mathbf{A}' + \mathbf{B}'$.
   c) Does all this mean that we can add matrices in any order we like, transposing them as directed whenever?

4) Prove $(ABC)' = C'B'A'$
   a) First show that the expression on the right-hand side is conformable for multiplication if the expression on the left-hand side is.
   b) Now do the proof (Hint: first prove that $(\mathbf{AB})' = \mathbf{B'A'}$; then use that result.)

5) At the end of section 2.4.7, Greene points out that the angle between $\mathbf{x}b$ and $\mathbf{y}$, θ, could be considered a measure of how far $\mathbf{x}b$ is from $\mathbf{y}$. Its cosine varies from 0-1, 0 when $\mathbf{x}b$ and $\mathbf{y}$ are orthogonal (perpendicular) and 1 when $\mathbf{x}b$ and $\mathbf{y}$ are the same line segment. Its cosine is given by $\cos\theta = (\mathbf{x}b'\mathbf{y}) / (\|\mathbf{x}b\| \cdot \|\mathbf{y}\|)$.
   a) Substitute the formula for $\|\mathbf{v}\|$ into this equation and rewrite it.
   b) Does the right-hand side of this look similar to any regression quantity you recognize? What? (Hint: call $\mathbf{x}b=\hat{\mathbf{y}}$)

6)

$$C = \begin{bmatrix} 1 & 3 \\ 5 & 6 \end{bmatrix}, \quad E = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 2 & 1 \\ 1 & 3 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 1 & 4 \\ 3 & 1 & 5 \end{bmatrix}$$

   a) what is CE'
   b) what is C^{-1}
   c) what is $(E'E)^{-1}$
   d) what is F^{-1}

7) In each of the following cases, how much does $y=f(x,\cdot)$ change for a marginal increase in x (*i.e.*, take the partial derivative $\partial y/\partial x$. Unless otherwise specified, no other variables besides y is a function of x. Also, e here is the natural.)

a) $y = b_0 + b_1x + b_2z + \varepsilon$ b) $y = e^{a+bx}$ c) $y = wxz$

d) $y = b_0 + b_1x + b_2z + b_3xz + \varepsilon$ e) $y = ln(72x^{3e})$ f) $y = [e^{a+bx}]/[wxz]$

8) In each of the following, solve the requested integral:

$$a) \int_{-\infty}^{1} e^{a+bx}dx \qquad b) \int_{a}^{c} (b_0+b_1x+b_2x^2)dx$$

PS 699: Problem Set 2 – Probability

1) The variance of a Bernoulli random variable is $\pi(1-\pi)$ where π is the probability of a successful Bernoulli trial (*i.e.*, probability that X=1).

a) first, using the formula for the p.f. of a Bernoulli random variable, and the formula for a variance, prove that V(X) = $\pi(1-\pi)$.

b) now, at what π is the variance of X highest? Even if you know the answer already, prove it by showing your maximizing V(X) with respect to π, *i.e.* $\text{Max}_{\pi}\ \pi(1-\pi)$.

2) Dice have six sides, all equally likely to "come up" (*i.e.*, appear on the top face when the dice are rolled. If you roll two dice and call their sum X,

a) draw (using excel, or by hand--carefully and as clearly as possible if the latter) the p.f. of X, *i.e.* draw f(X).

b) from what distribution would you say the random variable X is drawn?

c) what is the mean of X?

d) what is the variance of X?

e) based on your answer to these questions, especially b, give a general formula for the mean and variance of this type of probability distribution. (General here means that if I met another random variable like this one, and if I know the parameters of its probability function, I can use the formulae you give to find mean and variance.)

3) You may use the statistical tables from Greene or the probability functions in Stata or Excel or any other program as you wish to solve the following:

a) X is a normal random variable with mean 3 and variance 16. What is the probability that any draw, x, on this variable is greater than 7? Less than 1? Between 7 and 1? What is the 95% confidence interval for draws x? (That is, what symmetric range around the mean contains 95% of the probability. Drawing pictures of the distribution and shading the area you want may help if you're having trouble with this problem.)

b) Redo 3a only now X is a t with 10 degrees of freedom.

c) Redo 3a only now X is a Chi-Squared with 3 degrees of freedom. The last part, the confidence interval, is optional.

(4) *Jensen's Inequality*: Suppose X is a random variable equal to 1 with probability 1/3, 2 with probability 1/3, and 3 with probability 1/3. What is

(a) E(X)? (b) $[E(X)]^2$? (c) $E(X^2)$? (d) $ln[E(X)]$? (e) $E[ln(X)]$?

(f) Compare (b) to (c) (*i.e.*, is b larger or smaller than a?) and (d) to (e). Can you guess what it is about the functions $f(x)=x^2$ and $f(x)=ln(x)$ that makes these comparisons differ?

(5) The following table gives probabilities for a discrete bivariate distribution: f(x,y).

X \ Y	y = 1	y = 2	y = 3	y = 4
x = 1	.1	0	.1	.15
x = 2	.1	.1	0	0
x = 3	0	0	0	.1
x = 4	.1	.25	0	0

What is the probability...

a) x is 4 given that y is 2? b) y is 1 given that x is 4?

What is the conditional probability function...

c) f(x|y=1)? d) f(y|x=1)? e) What outcome for y makes x=1 most likely?

f) What is the probability (x<3 and y<3)? g) What is the probability x=4?

h) What is the probability that y=2?

(6) Suppose we have two independent, standard-normal-distribution random-variables Y and X.

a) do I need to tell you what kind of independence Y and X exhibit before you can do the rest of this problem, or will any of the three in the class notes suffice? Is this generally the case?

What is the probability that...

b) x>1 and y>1? c) x>1 or y>1?

What is the probability that...

d) x>1 given that y>1? e) y>1 given that x>1?

(7) You run a regression in which you estimate: $\mathbf{y} = b_0 + b_1\mathbf{x} + b_2\mathbf{z} + b_3\mathbf{x}\cdot\mathbf{z} + \mathbf{e}$

Stata spits out that $b_0 = 3.5$, $b_1 = 2$, $b_2 = 1$ $b_3 = -3$, assuming that x and z both vary from 0-1, graph the following in a spreadsheet (label appropriately and print):

a) $\partial y/\partial x$ on the "y" axis against z on the "x" axis

b) $\partial y/\partial z$ on the "y" axis against x on the "x" axis

c) in terms of the regression, can you tell me (in words) what these two lines mean?

d) Generically, what is the variance of $\partial y/\partial x$? (The trick here is to remember that, in the context of regression parameters, the **b's** are the random variables and the **x's** are constant ("across repeated samples"). Remembering that, just use the formula for V(a+bX+cY) given in class notes, remembering that in that formula the random variables are represented by X and Y. You have to be careful to get the notation lined up right with the substance. This, too, is a useful and important exercise.)

e) Generically, what is the variance of $\partial y/\partial z$?

f) suppose, finally, that Stata tells you $V[(b_1,b_2,b_3)']$ was given by the matrix:

$$V\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \sum_b = \begin{matrix} +0.256 & -0.027 & -0.309 \\ -0.027 & +1.296 & -1.624 \\ -0.309 & -1.624 & +2.652 \end{matrix}$$

g) graph $\partial y/\partial x \pm 1$ std. dev. (here's where using a spread sheet formula will help a *lot*)

h) graph $\partial y/\partial z \pm 1$ std. dev.

i) give your best guess as to why $\partial y/\partial x \pm 1$ std. dev. and $\partial y/\partial z \pm 1$ std. dev. are shaped like they are?

Spreadsheet help: create a column of numbers counting by, say, 0.1, from 0 - 1 (which as I told you is the range of x and z). Using the formulae you found in (a) and (b), use these columns of x and z to create two more columns, $\partial y/\partial x$ and $\partial y/\partial z$. Spreadsheets can use formulae in the form of, say, 4.7 + 3*(A3) where (A3) is the cell where the value that you want to multiply by 3 is located. Write the appropriate equation in this format in for the first row of (contrived) data. Now copy that formula down to the last row in which you have data. That's it; you're done. You can look at how the spreadsheet automatically updated the formulae as it copied down just by selecting any one of the cells in this new column. Once you have formulae for $V(\partial y/\partial x)$ and $V(\partial y/\partial z)$ you will need to use them similar to create columns of data equal $\partial y/\partial x \pm 1$ std. dev. and $\partial y/\partial z \pm 1$ std. dev.

PS 699: Problem Set 3

1. A recent article in the News/Free Press (true story) reported on a study concluding that there was no evidence of a link between a woman having had an abortion and her risk of breast cancer. (There were medical theories/hunches that suggested the possibility of a link.) 1.5 million women were in the study altogether. The concluding paragraph of the article states "Among the 280,965 women in the study who had 370,715 abortions, 1,338 cases of breast cancer were diagnosed by 1992. In the remaining group of women who had not had an abortion, 8,908 cases of breast cancer occurred, according to the study."

Test the hypothesis that the mean rate of cancer occurrence is the same regardless of having had an abortion or not. In particular...

a) What assumption(s), if any, do you need to make to test the hypothesis?
b) At what p-level could you just reject the hypothesis? (show your work; make whatever assumptions necessary, indicating which)
c) Does the conclusion of "no evidence of any difference" seem warranted to you given this evidence?

2. In our data base of characteristics of developed democracies, take the sample of countries with both upper and lower chambers of their legislatures. Assume that these available observations are a random sample from the universe of possible developed democracies with bicameralism (two legislative chambers). Test the hypothesis that the mean number of seats in each house is equal. In particular...

a) What assumption(s), if any, would you make before proceeding?
b) At what p-level could you just reject the hypothesis? (show your work)
c) On this basis, what would you conclude regarding their equality or inequality?
d) Would you be comfortable concluding that, rather than equal, one chamber, lower or upper, has higher mean number of seats? Which (if either) would you conclude has higher mean number of seats and on the basis of what test?

3. In each of the following, assume the available observations in our data set represent a random sample from the population of possible developed democracies. (You may use a spreadsheet or any useful computer program you like.) For each variable mentioned...

a) give an estimate from the sample of its mean in the population,
b) give a 90% confidence interval for the sample mean
c) test the hypothesis that the mean is μ (I'll give μ in parentheses after each variable); report a p-level, don't merely say reject / don't reject
d) test the hypothesis that the variance is σ^2 (I'll give σ^2 in parentheses after each variable); report a p-level, don't merely say reject / don't reject
e) state whatever distributional assumptions, if any, you are making in answering a-d

3i) Lower House Seats (LSEATS, μ=300, σ^2=31,000)
3ii) Lower House Proportionality Index (LPROP, μ=90, σ^2=25)
3iii) Number of Secondary Government Units (SECGOV, μ=35, σ^2=250)
3iv) Average Number of Parties in Government in Postwar Era (NPGOVPW, μ=1.5, σ^2=1)
3v) Number of Governments in the Postwar Era (NGOVPW, μ=22.5, σ^2=100)

4. a) What are the sample covariance and correlation between NPGOVPW and NGOVPW?

4. b) Test the hypothesis that the variance of the Vanhanen Index of Democracy is greater in 1970 than in 1960 (VDEM60 and VDEM70).

[Kmenta, 2nd ed. section 5-2 might be helpful for 3 & 4 if you're having difficulty]

5. a) Show that sample variance measured by , $s_a^2 = (1/n)\sum_i(X_i-\bar{x})^2$ is...

i) biased
ii) asymptotically unbiased (unbiased in the limit as n goes to ∞) and consistent. It may help you to know that the variance

of s^2 across repeated samples is $2\sigma^4/n$.
iii) and find its M.S.E.

5. b) Show that sample variance measured by , $s^2 = (1/(n-1))\sum_i(X_i-\bar{x})^2$ is...

i) unbiased
ii) Since it is unbiased, it is asymptotically unbiased; is it consistent? It may help you to know that the variance of s^2 across repeated samples is $2\sigma^4/(n-1)$.
iii) and find its M.S.E.

5. c) Given these results, do you think S^2 or S_a^2 is *better*? Why? On what might your answer depend?

6) The Bernoulli distribution is given by:

$f_{bern}(y_i|\pi_i) = Pr(Y=y_i|\pi_i) = \pi_i^{yi}(1-\pi_i)^{1-yi}$. (This is a Bernoulli, so each y_i is 0 or 1.)

Assume you have simple random sample of N bernoulli random variables with parameter π_i, and that each

$\pi_i=(1+e^{-xi\cdot\beta})^{-1}$.

a) What is the joint distribution of the y_i's? *I.e.*, what is $f(y_1, y_2, \ldots, y_n \mid x)$?

b) This joint distribution is the likelihood function of our data on y given our data on x and our assumption of how x relates to π_i. What, then, is the log-likelihood; i.e., what is $ln[f(\cdot)]$?

c) Take the derivative of the log-likelihood with respect to β and set it equal to zero? *I.e.*, set $\partial[ln f(\cdot)]/\partial\beta = 0$. β^* cannot be analytically found as a function of x and y, but simplify the implicit function given by $\partial[ln f(\cdot)]/\partial\beta = 0$ as much as possible.

You just set up a logit estimation.

7) You run two regressions. In one, you estimate $y=b_0+b_1X_1+b_2X_2+b_3X_3+e$; in the other, you estimate $y=a_0+a_1X_1+\xi$.

a) If the regression output told you that the log-likelihood of the 1st model was 17.3 and, of the 2nd, was 15.4; how might you test the hypothesis that the 2nd model was an insignificant restriction on the 1st?

b) Suppose, instead, your statistical package gives you that b_2=.05 and b_3=.025 and that the variance-covariance of those two coefficients is:

$$V(B) = \begin{matrix} .02 & -.01 \\ -.01 & .04 \end{matrix}$$

Now how might you test the hypothesis that the 2nd model was an insignificant restriction on the 1st?

c) Suppose the t-stat on b_2 was 1.2 and on b_3 it was 0.5; can you tell if the 2nd model was an insignificant restriction on the 1st based only on this info? How? or Why not?

PS699: Problem Set 4

In this assignment, in any statistics package you like, and using our data set, you will regress the lower-house proportionality index (LPROP) on the (country's average) lower-house electoral district magnitude (LMAG) for all available data points.

Background info: The proportionality index provides a summary measure for the entire set of parties in a country of the discrepancy between percentage votes for each party and that party's percentage of seats. If these percentages match up perfectly for all parties, the proportionality index is at its maximum of 100. An electoral district's magnitude is the number of lower house representatives (MP's in most cases) elected in that district. One representative per House district in the US. In the Netherlands, there is only one electoral district: all MP's are elected from a single, national, electoral district. LMAG is the average district magnitude for the country's entire set of electoral districts. Germany is a strange case which elects to the lower house from a two-tier system with separate effective electoral districts for the two tiers. The first tier is 1 MP per district and there are about 260 such districts. The remaining 260 seats are filled so as to bring the seat proportions in the country into line with parties' shares of the national vote.

1. Your hypothesis is, obviously, that district magnitude and vote-seat proportionality are related somehow. State specifically a hypothesis of this sort testable by the linear regression suggested above.

2. Given your theory and the goal of testing this hypothesis, make a decision about how to code LMAG for Germany. Briefly (very briefly, a paragraph should do for our purposes) describe that decision and defend it.

3. Regress LPROP on LMAG and present your results (print out your stats package results).

4. Do you believe that LMAG and LPROP are absolutely related (i.e., each 1-unit change in LMAG produces a β-unit change in LPROP? Or do you think perhaps that percentage changes LMAG produce absolute changes in LPROP (i.e., each 1% change in LMAG produces a β-unit change in LPROP)? Or that absolute changes in LMAG produce percentage changes in LPROP (i.e., each 1-unit change in LMAG produces a β% change in LPROP)? Or that percentage changes in LMAG produce percentage changes in LPROP (i.e., each 1% change in LMAG produces a β% change in LPROP)? Pick one (defend it in at most a paragraph).

5. Run that regression and present your results (print out your stats package results).

6. Using these results (from 5), give a **substantive description** (i.e., what does it mean in terms of your theory here) of each of the following and give a formula and/or describe formally how they are calculated **mathematically**.

a) The coefficient, b, on LMAG or *ln*(LMAG) (whichever you used).
b) The standard error of that coefficient
c) The t-statistic and associated confidence interval (95%) and p-level.
d) R^2 and the estimate of σ_ε^2 (the "Standard Error of the Estimate" or "...of the Regression," or whatever your stats package calls it)
e) The F-Statistic for the regression and its associated p-level.

PS 699: Problem Set 5--Multivariate Regression:

I. We may have concluded last time that the proportionality of electoral/parliamentary outcomes (the congruence between the distribution of votes across parties and the distribution of parliamentary seats across parties) as measured by LPROP was (linearly) positively related to the natural log of LMAG (*i.e.*, absolute values of LPROP relate linearly to proportionate values of LMAG). (Whether you concluded this or not, we are concluding it now.) Now, suppose we consider two extensions of this simple proposition: a) that LPROP is similarly (proportionately) related to the number of electoral districts and b) that it is simply (linear-absolutely) related to the electoral system (whether the system is proportional representation or plurality/majority). Theoretically, we consider (a) likely because disproportionality happens when some parties are disproportionately favored by some vote distribution in a district given the set of electoral rules and others are disadvantaged. We might argue, then, greater number of electoral districts might make it less likely that the same party (parties) were so advantaged in every district. Thus, more districts might improve the proportionality of results of national elections to parliaments because advantages in one district might be balanced by disadvantages in another. (b) is considered likely because proportional representation, since it produces more proportional outcomes at the district level than plurality or majority (practically by definition), ought also to produce more proportionate results at the national level.

A. Suggest a linear multivariate regression model which you might use to empirically evaluate these theoretical propositions.

B. Estimate this regression by OLS (print out your results). (Assume for these purposes that we have come to consensus that Germany should be coded LMAG=260, LEFORM=0, and no German dummy variable is to be included.)

C. Interpret your results for the individual coefficients. *I.e.*, what do your coefficients mean? and report somehow on your uncertainty about these coefficients (*i.e.*, make some substantively meaningful statement about their standard errors, statistical significance, &/or confidence).

D. Test the hypothesis that neither of the new variables added significant explanatory power to our model from last week.

E. Test the hypothesis that the whole model fails to offer significantly different explanatory power than merely knowing the sample mean of proportionality for this set of countries.

II. A political economist suggests to you a theory regarding how governments make social-security-spending policy. Specifically, s/he says:

"Of course, social-security spending (as a share of GDP) responds to unemployment and to the age distribution of the economy/polity. After all, social security spending is more or less exactly that: spending on unemployment insurance and old-age 'insurance'. However, aside from that, I would also argue that social-security spending is used as a political tool. For example, suppose that coalition governments find it harder to stay together the more parties there are in the coalition. Suppose further that social-security spending helps them maintain the coalition because it buys votes for all parties in government and is thus one few things they can all agree on. Suppose finally that smaller coalitions and single-party governments are able to agree on more things and thus can focus on other spending priorities. Then, social-security spending should also be positively related to the number of parties in government. Another example, I bet the more people vote the more the poor are represented so the more the government will spend on social security, again, *ceteris paribus*."

Leaving aside for the moment the various dates the data in our data base were collected (*i.e.*, assume all data were collected and refer to the appropriate period), and leaving aside that the above arguments are woefully incompletely spelled out in their logic, I suggest estimating the following CNLR model by OLS regression as a means of empirically evaluating the above theorist's claims:

$$SSPENDG = \beta_c + \beta_a AGE + \beta_u UE + \beta_n NPGOVPW + \beta_v VPART + \varepsilon$$

A) Estimate this model by OLS (print your results).

B) a) Interpret the estimates of the coefficients β_n and β_v substantively.
b) Evaluate the statistical evidence for or against the theorist's hypothesis regarding coalition size and SS spending.
c) Evaluate the statistical evidence for or against the theorist's hypothesis regarding voter participation and SS spending.
d) Evaluate the statistical evidence for or against the theorist's hypothesis that SS spending is used as a political tool using the evidence available from this regression.

C) Test the hypothesis that, the effect on SS spending of 1% more elderly, *ceteris paribus*, is equal to the effect on SS spending of 1% more unemployed, *ceteris paribus*.

D) Suppose some country experiences an increase of 1% of its population that is both elderly and officially unemployed (this is impossible in most countries, but possible in some). What is your estimate of the change in SS spending in that country as a result? Give a 90% confidence interval for that estimate.

E) (More Difficult) Find the 95% *joint* confidence area for your estimates of β_n and β_v.

PS699, Problem Set 6: More Hypothesis Testing

Consider the argument that the duration of (postwar, democratic) governments is a (linear) function of two characteristics of the sitting governments: the parliamentary support of the parties in that government and the number of parties in that government, *i.e.*:

Duration of Gov'ts = f(parliamentary support, # parties in government, stochastic component)

Specifically, the argument is that duration is an increasing function of parliamentary support and a decreasing function of the number of parties in government.

A. Using the data from developed democracies in the postwar period, suggest a simple linear model (*i.e.*, don't transform anything into logs or anything) which you could estimate to evaluate the evidence for or against this argument.

B. Estimate this model by OLS and report on the results as they relate to the arguments. Specifically:
1) what have you estimated the effect of each suggested independent variable on duration, controlling for the other, to be[1]
2) report on your certainty regarding this estimate in any substantively revealing manner

C. A critic of your model (its yours not mine now that it has a critic) suggests that, while your general argument is correct as far as it goes, it leaves out a tremendously important (in critic's view) factor: party discipline. Specifically, the critic suggests that the US, Japan, France, Italy, Canada, Greece, Spain, and Portugal have low party discipline and the rest have high party discipline. "Party discipline," the windbag continues, "is really what explains government duration: more party discipline produces more durable governments and that's all there is to it."
1) use the critic's own definition of high and low party discipline countries to define a nominal variable distinguishing those two groups of countries.[2] Add this new variable to the set of independent variables in your regression from B and re-estimate. (For future reference, add this variable to your data set and call it PD).
2) report on your results as in B.
3) regarding the critic's claim(s) *vis-à-vis* party discipline and government duration, what would you conclude?

D) The critic returns and says: "Wait! I didn't mean that party discipline, measured so crudely, would explain all the differences in government durations across developed democracies by itself nor really that it was a factor explaining duration like your others. I meant that the impact of the other variables, such as parliamentary support and number of parties in government, on government duration was different depending on whether the county's parties had much party discipline or not." Never mind that this is not what the windbag said, go ahead and:
1) Suggest a way to evaluate empirically the claim that the effects on government duration of parliamentary support and of the number of parties is different in PD and non-PD countries.[3]
2) Carry out that test and report on it.[4]

E) "No, wait! Here's what I meant to say..." [Cheeze Wiz! This guy's annoying, huh?] "...I meant to say that we should consider your model, where government duration is a function of parliamentary support and the number of parties in government, and my model, where its a function of party discipline and whether it's effectively a presidential system or not [where did that come from?!], as alternatives. Mine's better."

---

[1] As always, we want to state this in substantive, as well as numeric terms, *i.e.* we don't simply say "the coefficient on x is b," we say something substantively meaningful like "*ceteris paribus*, a 1% increase in parliamentary support for government produces an *x* month increase [decrease] in government duration."

[2] I believe this division is about right, but the groupings could easily be slightly erroneous as a matter of actual empirical fact. I wouldn't use this division in any actual research without further exploration.

[3] Assume the critic means to say that the mean duration controlling for parliamentary support and for the number of parties in government would also be different in PD *versus* non-PD countries.

[4] There may be more than one test that could evaluate this claim; use the one involving separate estimation of the same regression equations in different samples.

1) Suggest a way to evaluate empirically this claim.[5]
2) Carry out that test and report on it.

F) After thinking about the critic's pestering and your findings, you consider the following set of propositions:

* government duration is a function of parliamentary support, the number of parties in government, and party discipline,
* the effect of parliamentary support on government duration depends upon the number of parties in government and whether there is party discipline or not,
* the effect of the number of parties in government on government duration depends upon parliamentary support and whether there is party discipline or not, and
* the effect of party discipline on the duration of governments depends upon the amount of parliamentary support and the number of parties in government.

1) Suggest a model capable of shedding empirical light on all of these propositions.
2) Estimate this model by OLS (print out the results).

(In the following questions, "what does the evidence say about..." means "conduct a test corresponding to the statement and report on it.")

3) What does the evidence from this regression say about your (entire) first proposition?
4) What does the evidence say about the proposition that government duration is a function of parliamentary support? ...of the number of parties in government? ...of party discipline?
5) What does the evidence say about the proposition that the effect of party discipline on government duration depends on the amount of parliamentary support? ...depends on the number of parties in government?
6) What does the evidence say about the proposition that the effect of parliamentary support on government duration depends on the presence or absence of party discipline? ...depends on the number of parties in government?
7) What does the evidence say about the proposition that the effect of the number of parties in government on government duration depends on the amount of parliamentary support? ...depends on party discipline?
8) How many unique hypotheses are encompassed in your last three propositions?

[5] Again, there may be more than one way to do this; one way we covered is most appropriate though.

Problem Set 7: More Interactions and Omitted Variables

I. INTERACTIONS: Consider a fully interactive linear model of the proposition that Government Duration is a function of the number of parties in government, parliamentary support for government, and party discipline. I mean by fully interactive that you wish to consider hypotheses that the effects of each variable on government duration depends on each of the others.

$DGOVPW=\beta_0+\beta_1 PSUPGPW+\beta_2 NPGOVPW+\beta_3 PD+\beta_4 PD*PS+\beta_5 PD*NP+\beta_6 NP*PS+\varepsilon$

A. Estimate this model using our data set.
B. What is the estimated effect of parliamentary support on government duration?
C. What is the estimated effect of the number of parties in government on government duration?
D. What is the estimated effect of party discipline on government duration?

Using any graphics program you like (a spreadsheet is convenient for this)...

F. Create a graph of the effect of parliamentary support on government duration, as a function of the number of parties in government, when party discipline is high. Create another graph of the same thing, only when party discipline is low. On both of these graphs, graph a 90% confidence interval around the effect line as well.
G. Graph the effect of the number of parties in government on government duration, as a function of parliamentary support for the government, when party discipline is high. Create another graph of the same thing, only when party discipline is low. On both of these graphs, have a 90% confidence interval around the effect line graphed as well.
H. Create a table of the effects of party discipline at various levels of parliamentary support and number of parties in government. In particular, parliamentary support goes from about 40 to about 80 (%) in our sample (this is called the sample range of the variable) and the number of parties goes from 1 to about 4. The mean of parliamentary support is about 57 and the mean number of parties is about 2. So, create a 3x3 table of the effects of party discipline (measured by the movement from a low PD country to a high one) on government duration at the various combinations of these values for the other variables. Each cell entry in the table should include the effect of PD at that combination of PS and NP, the standard error of that effect, and a t-test that the effect at that point is zero.

II. OMITTED VARIABLES: Go back to model from Problem Set 5:

(1) $SSPENDG = \beta_c + \beta_a AGE + \beta_u UE + \beta_n NPGOVPW + \beta_v VPART + \varepsilon$.

Suppose this were the true model, but you instead estimate (*i.e.*, you didn't know the true model until a critic came along and told you that you left something out):

(2) $SSPENDG = \beta_c^* + \beta_a^* AGE + \beta_u^* UE + \beta_n^* NPGOVPW + \varepsilon$

A. What is β_n^* in terms of (a) coefficient(s) of model (1) and perhaps some other regression? Run the regressions necessary to estimate b_n^* (*i.e.* the estimate of β_n^*), whatever coefficient(s) you needed from model (1) to answer the first part of this question, and whatever other thing(s) you may have needed. Show that b_n^* is indeed equal to the estimated analog of what you said β_n^* was equal to.
B. Under what conditions, if any, would b_n^*, the estimate of β_n^*, be an unbiased estimate of β_n, the effect of NP on SS spending controlling for voter participation, age, and unemployment? *I.e.*, when if ever would $E(b_n^*) = \beta_n$?
C. Describe the conditions under which the s.e.(b_n^*) be less than, greater than, or equal to the s.e.(b_n)? Under what conditions would b_n^* be efficient (have minimum variance)?

Suppose, now, (1) is still the truth but you estimate: $SSPENDG = b_c + b_a AGE + b_u UE + b_n NP + b_v VP + b_p PD + \varepsilon$.

D. What is the expected value of b_p, $E(b_p)$? What about the other coefficients?

E. Are your coefficient estimates efficient? Why or why not?

PS699 Problem Set 8: Some Regression Diagnostics

Let's return to one of our more successful sets of hypotheses, that the duration of governments in postwar democracies is a linear function of the number of parties in government, parliamentary support for the government, and party discipline. Re-estimate that as I have done here:

```
fit dgovpw npgovpw psupgpw PD
```

```
  Source |       SS       df       MS                 Number of obs =      23
---------+------------------------------               F(  3,    19) =    4.10
   Model |  806.117215     3  268.705738               Prob > F      =  0.0211
Residual |  1244.90002    19  65.5210536               R-squared     =  0.3930
---------+------------------------------               Adj R-squared =  0.2972
   Total |  2051.01723    22  93.2280561               Root MSE      =  8.0945

------------------------------------------------------------------------------
  dgovpw |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 npgovpw |  -3.520692   2.001078     -1.759   0.095      -7.708997    .6676123
 psupgpw |   .4338356   .2105621      2.060   0.053      -.0068759    .8745471
      PD |   9.242491   3.638877      2.540   0.020       1.626235    16.85875
   _cons |    1.24442   11.07908      0.112   0.912      -21.94436     24.4332
---------+--------------------------------------------------------------------
```

For each of the diagnostic plots and statistics below, print out the results and answer the questions.
Answers might well be simply "no" or "nothing in particular".

```
. avplots , s([ctry])
```

1. This plots the residuals from regressing Y on all X but one against the residuals from regressing that one on the others.
   a1) Does anything you see in these three graphs alarm you as indicating a possible violation of the CNLRM assumptions?
   a2) What would you expect to see if all was well on that score?
   b1) Does anything you see in these three graphs alarm you as indicating potential inordinate influence of some country(ies) on the results?
   b2) What would you expect to see if all was well on this score?
   c) Anything else you think worth noting about this?

```
. cprplot npgovpw , s([ctry])

. cprplot psupgpw , s([ctry])

. cprplot PD , s([ctry])
```

2. These three plot the residuals from regressing Y on all X's plus the line given by b*x against that x.
   a1) Does anything you see in these three graphs alarm you as indicating a possible violation of the CNLRM assumptions?
   a2) What would you expect to see if all was well on that score?
   b1) Does anything you see in these three graphs alarm you as indicating potential inordinate influence of some country(ies) on the results?
   b2) What would you expect to see if all was well on this score?
   c) Anything else you think worth noting about this?

```
. rvfplot , s([ctry])
```

3. This plots your residuals versus fitted values.
   a1) Does anything you see in this graph alarm you as indicating a possible violation of the CNLRM assumptions?
   a2) What would you expect to see if all was well on that score?
   b1) Does anything you see in this graph alarm you as indicating potential inordinate influence of some country(ies) on the results?
   b2) What would you expect to see if all was well on this score?
   c) Anything else you think worth noting about this?

```
. rvpplot PD , s([ctry])

. rvpplot psupgpw , s([ctry])

. rvpplot npgovpw , s([ctry])
```

4. These three plot your residuals versus each of the X variables.

a1) Does anything you see in these three graphs alarm you as indicating a possible violation of the CNLRM assumptions?
a2) What would you expect to see if all was well on that score?
b1) Does anything you see in these three graphs alarm you as indicating potential inordinate influence of some country(ies) on the results?
b2) What would you expect to see if all was well on this score?
c) Anything else you think worth noting about this?

```
. dfbeta
DFnpgovp:  DFbeta(npgovpw)
DFpsupgp:  DFbeta(psupgpw)
DFPD:      DFbeta(PD)

. fpredict cooksd , c
```

5. Print out and examine these three vectors of DFBetas and the Cook's distance summary score.
a1) Does anything you see in these four vectors alarm you as indicating a possible violation of the CNLRM assumptions?
a2) What would you expect to see if all was well on that score?
b1) Does anything you see in these four vectors alarm you as indicating potential inordinate influence of some country(ies) on the results?
b2) What would you expect to see if all was well on this score?
c) Anything else you think worth noting about this?

```
. vif

Variable |      VIF      1/VIF
---------+----------------------
 npgovpw |      1.24    0.807147
 psupgpw |      1.23    0.810485
      PD |      1.05    0.948403
---------+----------------------
Mean VIF |      1.18
```

5. Here are the variance inflation factors.
a1) Does anything you see in these values alarm you as indicating a possible violation of the CNLRM assumptions?
a2) What would you expect to see if all was well on that score?
b1) Does anything you see in these three values alarm you as indicating potential inordinate influence of some country(ies) on the results?
b2) What would you expect to see if all was well on this score?
c) Anything else you think worth noting about this?

```
. lvr2plot , s([ctry])
```

6. This graph plots leverage (potential influence) against normalized squared residuals.
a1) Does anything you see in these values alarm you as indicating a possible violation of the CNLRM assumptions?
a2) What would you expect to see if all was well on that score?
b1) Does anything you see in these three values alarm you as indicating potential inordinate influence of some country(ies) on the results?
b2) What would you expect to see if all was well on this score?
c) Anything else you think worth noting about this?

```
. fpredict StudRes , rstu
```

7. Using these studentized residuals, can you:
a) Figure out a regression to run testing the claim that variance is higher among Westminster (UK, Canada, Australia, and New Zealand) systems? What would you do? (Don't have to do it.)
b) Figure out a regression to run testing the claim that residuals in neighboring countries were correlated? What would you do? (Don't have to do it)
c) In which (may be none, one, or several) of the above plots using regular residuals do you think Stata would have done better to use studentized residuals?

PS 699 Problem Set 9: Heteroskedasticity and Serial Correlation

You have a new data set in the lab called MES_699.dta. The data were given to me by Harold Clark of the University of North Texas. The data are part of those given to him by Mackuen, Erikson, or Stimson (MES) from their article:

MacKuen, Michael B., Erikson, Robert S., and Stimson, James A., "Peasants or bankers? the American electorate and the U.S. economy," The American Political Science Review 86:597-611 September 1992.

MES wish to ask whether presidential approval responds to the economy in either or some combination of four ways: are people retrospective and/or prospective (do they reward a president for past economic performance and/or for expected future performance)? and are they personal and/or sociotropic in these economic evaluations (are they concerned with their own economic situation and/or with their perception of the national economy)?

This assignment is not really to be taken as a comment one way or another on their questions, their answers, or their research design. Nor should what we do here be viewed as sufficient to answer these questions. Nor should the sequence of questions or the questions themselves be taken as a guide to a research program (it has many extraneous and questionable steps as such). It's an exercise.

(1) Taking the above questions as nestable ("putable" in one model), you consider first estimating the following model (the data are quarterly: one observation per quarter (three month period) from Eisenhower to Bush):

$$APP = \beta_0 + \beta_{pe}\, PEXP + \beta_{pa}\, PAGO + \beta_{be}\, BFUT + \beta_{ba}\, BAGO + \beta_v\, VIET + \beta_e\, EVENTS + \varepsilon$$

where APP is the % of people surveyed who approve of the president's handling of his job
PEXP is % of people who expect their financial situation to approve for them in the near future
PAGO is % of people who reported that their financial situation was good in the near past
BFUT is % of people who expect the economy to improve over the near term
BAGO is % of people who reported that the economy was doing well in the near past
VIET is the number of war dead in vietnam (in thousands that year-to-date)
EVENTS is basically a dummy variable coded 1 for the occurrence of a high-profile media-covered event involving the president (see the article for more details on all these variables).

(a) Estimate this model by OLS (in Stata, using) "reg" or "fit" command. Print it out.
(b) Report on your coefficient estimates (*i.e.*, interpret them ("a such-and-such change in such-and-such is estimated to produce a such-and-such change in approval"), and be sure to give some indication of the certainty of those estimates (standard error, p-level, confidence interval...).

(2) It occurs to you now, that Approval may be more variant at some times than at others. You decide that perhaps you ought to test for heteroskedasticity. You decide on White's General Test (which is to regress squared residuals on all your X, your $x_i x_j$, and your X^2 terms.) Stata: you can use *predict err , re* to generate a variable, err, equal to your residuals. Stata: you can use *gen varname=formula* to generate any other variables needed for the test.
(a) Print the test regression output.
(b) Report the results of the White test (the statistic, the p-level, your conclusion)

(3) Based on these results, you decide you had better "be safe" and use White's robust standard errors instead of the usual OLS ones.
(a) Re-estimate using White's and printout the new results
(b) Comment on any noticeable changes: point them out and just briefly indicate what the change(s) suggest(s) to you regarding the structure of your error variances and X's.
(c) Assume the true residuals from this regression are drawn from independent distributions that have variance which possibly differs from observation to observation.
(i) What are the properties of these coefficient estimates (bias? consistency? efficiency?)
(ii) What about these standard errors?

(4) Referring back to White's test, suppose you decided to view it or something like it as constructive. In particular, suppose you take these results as possibly indicating some heteroskedasticity as a function of BAGO, PAGO, $BAGO^2$, and $PAGO^2$ (in a real project, you'd want to have a theory which predicted such a thing).
(a) Report on a Glesjer's test regressing $ln(err^2)$ on these four variables (and a constant).
(b) Use the predictions from this regression as estimates of the natural log of the variance of each observation. Use those predictions to generate a weighting variable (call it *wt*) by which you could conduct FWLS (re-)analysis of the model. (Hint: the weight involves an estimate of the variance of the observation, *not* the log of the variance. Be sure to transform back to the right scale using e^x.) Graph the weighting variable over time. (In Stata, *graph wt timeindx , s([date])* will do.)

(5) Re-estimate the model of (1) by FWLS using this weighting series from #4. (in Stata, after the regression model type *[aweight=wt]*.
(a) Present and report on the results (interpret coefficients and give some statement about their estimated certainty (s.e., T, confidence, whatever)).

(b) R^2 and M.S.E. have changed. Why? What, if anything, should be made of this?
(c) Highlight and comment on any other noticeable changes in the results.
(d) If your model of the error-variance structure is right...
    (i) what are the properties of your coefficient estimates?
    (ii) what about your standard errors?

(6) After all this, it suddenly occurs to you that it's possible that an observation in one quarter is not necessarily independent of the previous observation. What does the Durbin-Watson statistic from model 1 say? (i.e., what is it and what conclusion does it suggest, check the table in Greene to get d^{upper} and d^{lower}. (In Stata you have to estimate it again, this time using *regdw*, to get the DW).

(7) Uh-Oh! You decide you'd better use Newey-West standard errors (and you consider any lag beyond 5 periods to be negligible) because you no longer trust your OLS standard errors. Printout the results for the estimation with the Newey-West Coefficient Variance-Covariance Matrix. In Stata: the command is *newey* put model here *, lag(5)*.
(a) Comment on any noticeable changes from model 1. *E.g.*, have your p-levels changed much? What does this tell you about your previous assumption of independent observations?
(b) If all lags beyond 5 show no correlation, and the rest of your assumptions hold:
    (i) what are the properties of your coefficient estimates?
    (ii) what about your standard errors?

[In all of the following models, you need to include the dummies (all of them) for the 1st quarter of an administration. This prevents any effects from previous administrations from "bleeding" into the next ones.]

(8) You read on in your handy econometrics test and it says: estimate a model where $\varepsilon_t = \rho\ \varepsilon_{t-1} + \gamma_t$. Re-estimate model 1, only this time using a procedure which estimates $\varepsilon_t = \rho\ \varepsilon_{t-1}\ \gamma_t$. Several could do (Cochrane-Orcutt, Prais-Winston, Hildreth-Lu, ...); use Cochrane-Orcutt here (in stata: the command is *corc* put model here *, t(timeindx)*.
(a) What is the estimated coefficient on last period's residual in this model?
(b) Comment on any noticeable changes in the results from model 1. (Be sure to mention the change in DW statistics from model without to model with AR(1).)
(c) In particular, what do you think of your previously held independence assumption now? What about OLS's consistency and unbiasedness in the face of correlated residuals? Is that reassuring given the changes from question 1 to here? What about your Newey-West results? Do they seem adequate now?
(d) If this AR(1) process is the right process generating your residuals:
    (i) what are the properties of your coefficient estimates?
    (ii) what about your standard errors?
(e) Does the DW statistic give you much confidence that this is the right residual process?

(9) Generate a variable equal to lagged approval (*gen app=app[_n-1]*). Using this variable, reconsider the temporal process generating your data. In particular, estimate a model where the current approval is a function of previous approval (and the rest of model 1) and the residuals from this model are not correlated.
(a) Estimate this model by OLS an print out the results.
(b) Comment on any noticeable changes from model 1.
(c) Using an applicable test, report on whether the residuals from this regression continue to exhibit first-order correlation.
(d) Assuming the conclusions of this test are correct:
    (i) what are the properties of your coefficient estimates?
    (ii) what about your standard errors?
(e) Which model seems better now, (8) or (9)?

(10) Re-estimate model 9, but this time include the set of dummy variables for each president (dde, jfk, *etc.*). Why do you suppose your coefficient on lagged approval changed the way it did? [Hint: suppose there were fixed differences from president to president, but you didn't control for that. What would OLS try to do with the coefficient on lagged approval to make up for this omitted variable bias?]

(11) Using the estimates from the model in 10 and a spreadsheet, graph the response of approval to a temporary and to a permanent 1 point increase in BFUT (that's two response lines).

(12) Use your statistics package to graph residuals over time (*graph residuals timeindx, s([date])*.) Notice any outliers? Anything happen in that quarter that might explain this? (hint: Bush is in White House, troops were involved, oil was involved...) (That should have been in the events series, don't you think?)

PS 699, Problem Set 10: Binary Dependent-Variable Models

This problem set uses Paul Huth's "Extended Deterrence" data set. The data code the particulars of 58 historical cases of attempted extended deterrence. Loosely, attempted extended deterrence is when one country publicly vows to defend another in the face of a threat from a third so as to get the threatener to back down. The threatener is also called the attacker, the extended-deterrence attempter is also called the defender. The threatened party is called the protégé. A pair of archetypical examples would be the US attempts to deter North Korea from threatening South Korea (that's failure in the 1950s and success in the 1980s). US is defender, North Korea is threatener or attacker, South Korea is protégé.

Specifically, we will be using the variables:

det_suc : (dependent variable) coded 1 if attempted extended deterrence succeeds
imbalfor : the immediate balance of forces (*i.e.*, relative military might in the area, defender relative to attacker)
stbalfor : the short-term balance of forces (*i.e.*, relative military might of defender, in area or not)
ltbalfor : the long-term balance of forces (*i.e.*, a measure of relative military might defender could eventually bring to bear if confrontation went to protracted war--*e.g.*, economic capacity)
nuke : coded 1 if defender has nuclear capacity, 0 otherwise
alliance : coded 1 if there is an explicit military alliance between defender and protégé
arms_xfr : measures arms imports by protégé from defender
for_trde : measures protégé's share of defender's foreign trade

1. Using logit, estimate a model predicting the probability of deterrence success (given deterrence attempted) as a function of the other 7 variables. Print out your results.

2. Which independent variables make extended-deterrence success more likely? Which make it less likely? For which is it relatively uncertain? Summarize these answers in a bullet outline; example bullet-entry:

- An immediate balance-of-forces favoring the defender increases the likelihood of extended-deterrence success, *ceteris paribus*. This effect is among the substantively largest of the independent variables and is the most statistically significant as well.

3. Test the hypothesis that short-term balance-of-forces and long-term balance-of-forces adds nothing to the model's predictive power.

4. Drop long-term balance-of-forces from the model and re-estimate. Print it out.

5. In the model just estimated, how much does acquiring nuclear capacity increase a defender's probability of succeeding at attempted deterrence?
   a) Answer first using the derivative method. Leave your answer as a formula.
   b) Evaluate that formula at the sample mean of all independent variables.
   c) Fixing other variables at their sample means, answer using the difference method

6. Fixing all other independent variables to their means, offer a graph showing the predicted probability of deterrence success (y-axis) as the protégé's share of defender's trade goes from 0 to 10 (%, approximately the sample range of for_trde) (x-axis).

7. Re-estimate the model of #4 by Probit. Do any of the hypothesis tests on the individual coefficients change noticeably? The coefficients change, but do the estimated effects? To explore the latter, re-answer 5a, 5b, and 5c for this probit model.