Sociology 505 Winter 2023: Problem Set #1 Due date: 1/20/2023 5 PM

Use short answers (you might need 4/5 sentences or a small paragraph but not much more text for a decent answer). Be concise and to the point. You will probably find it convenient to use Rmarkdown, but you can also cut from your output and paste into a Word document (or some other text format) and then add typed comments; we don't need to see the full output from R, Stata or other statistical programs. When you do have "calculations" to perform showing some of your work is useful. (You can show calculations either typed out or neatly written down).

The problems cover introductory materials (roughly through week 2, part of week 3) from Class and from various readings. The set also introduces some computer applications; the data required for these problems are on the class website. Note you may have to edit some of these data files to work the problems correctly. You may also need to get some consulting, but generally these files and problems associated with executing the analysis should be covered in class. Feel free to e-mail me/TA with questions.

Data for the problems are in csv format with row #1 being variable names. Each problem should give you enough information to define what the variables are (e.g. their definition, metric etc).

Problem Set 1: A review of bivariate models/equations and some introductory multivariate equations:

Problem 1.1, 1.2, and 1.3 make you think about the properties of these linear models with normally distributed stochastic component… and exercise some programming skills.

Problem #1.1. (Note: Variable names appear in parentheses.) Using the data from a sample of US school districts/counties (see file on course website **districtdata2023**) execute a model that has the districts' general economic well-being as captured by the income level in the county (INCOME $'s per capita) predicting investment in education (EXPEND $ per student) (i.e. there is a positive association between economic well-being of the population and willing to invest in the future—schooling of the young).

Part 1) Describe the linear relationship between EXPEND and INCOME (where EXPEND is predicted by INCOME)

a) describe/report details about the univariate distributions of the two variables (EXPEND and INCOME); include at least on measure of central tendency and one measure of variability; produce a bivariate plot of association of EXPEND and INCOME.

b) estimate and describe the bivariate linear model relation between the two variables using ML estimation (EXPEND is the outcome variable).

c) calculate conditional mean of expenditures based on the model if the per capita income is 25000 or if it is 35000

d) look at the data and calculate the raw response residual/deviation for CountyID #73600 and CountyID # 41500. What does the deviation tell you about school expenditure in County 73600 and 41500 with respect to the estimated equation?

e) calculate the median value of income and run the same model from (b) but select only those cases that are below the median value and again for only those observed cases above the median. What do these two results suggest about the overall equation or model you calculated on the full sample---e.g. how appropriate the systematic part of the model is based on this information? In looking at your plot from part(a) above describe why your results from this exercise are or are not surprising.

f) do the same exercise but select if EXPEND is less than EXPEND's median of 2800? Compare differences or similarities in the f's results to (a) and to (e). Explain any differences you might observe.

g) estimate a model in which you predict "expend" simply by a best fitting constant and describe this model; what features of your model and information from " b " above makes you think your model in "b" is better than your model in "g" (or is it???). (Just some intuition as to why you think "a" or "b" is better).

Problem #1.2) Run a multiple regression using the same data and model predicting EXPEND but add the variable % college graduates in the population (COLGRAD--% of population age 25-39 with college degree---unit of measure is "in percentage points").

a) describe the relations of COLGRAD and INCOME to EXPEND from the results of this new model.

b) explain why there is a difference between the effect of income on expenditure in Part 1 "b" above and what you see now in Problem 1.2.

c) standardize the two partial derivatives for the two independent variables and interpret their effects

d) THOUGHT QUESTION: If there were basically no change in the magnitude of the coefficient for income what would you know or conclude about the relation between INCOME and COLGRAD?
e) THOUGHT or DO QUESTION: If we calculated INCOME as deviation from its mean (i.e. observed income minus mean of income) for each case what would change in your model in "1.2 a" (if anything) and why?
f) calculate/describe the variance/covariance matrix for the 3 variables in this problem; also the standardized variance/covariance matrix for the same 3 variables.

Problem #1.3
Using the AdolMentalHealth2021 dataset (a random sample of adolescent youth from a NW metropolitan area)--- estimate this multiple regression equation (using maximum likelihood estimation)

HOPELESS=f(AGE,CPBONDS,CABONDS,FDISTRES)
Where age is in years, cpbonds and cabonds are in counts of people, and family distress is a set of items in a Likert scale from 0 to 6 where 0 is no/low distress and 6 is high distress (note: you can talk about "hopeless units" and "distress units" as continuous scales). Hopeless is a continuous scale ranges from 10 – 30 and is the weighted average of a set of Likert items; higher scores = more hopelessness; one can think of the scale as "hopeless units".

Hopelessness scale = linear additive model with independent variables AGE, # of positive supportive peer friends (CPBONDS), # of nonfamily adults(e.g. teachers-- CABONDS) who are seen as supportive, and family distress 10pt Likert scale (FDISTRES))

a) estimate and interpret the point estimate results from the above multivariate model. Make sure you evaluate the range of magnitude of the marginal effects for each variable and discuss your sense of the likely direction of each effect.

b) comment on whether "positive" peers (cpbonds) or number of supportive adults (cabonds) appears more important to an individual's level of hopelessness.

c) What is the partial derivative of change in hopelessness with respect to #of positive adult bonds in the 4 variable model.

d) compare a model with only age and family distress (call that Model 1) to the model with these two variables plus the two variables counting supportive people (peers and adults) in the model (call that Model 2)… discuss intuitively how you might consider that adding peer information is important to understanding/predicting hopelessness… (i.e. what might be evidence that suggests improvement in the model) (I am looking for both a statistical assessment and also an intuitive feel)

e) THOUGHT QUESTION: How might you test a model that says the absolute effect of positive peers and supportive adults have the same effect on hopelessness. Is this a legitimate question? (Hint: use algebra and set their effects to be the same in magnitude). **Bonus: if you estimate such a model (kudos if you get it correct and no harm for trying if you get it wrong).**

f) THOUGHT QUESTION: Would it be reasonable to do the same in comparing the effects of FAMILY DISTRESS and number of POSITIVE PEERS? Yes, No, Maybe… and why.

Consultation #1.1: This problem is to be accomplished independently. (Please hand this section in separate file from the above Problem Set #1.

One of your professors is trying to get a piece of research published in the Journal of Education Effects (a new online journal). He has asked you to read the paper and respond to some summaries he has written to you. In the tables below the professor reports the means and the variance/covariance and correlation matrix (correlations above the diagonal/covariances below…variances on the diagonal are in **bold**). The data are from a random sample of cross-national urban places of size 250,000+ population (N=198); each variable is defined in Table 2. The professor wants to demonstrate that number of non-state sponsored NGO's (averaged over a 3-year period) that provide social and health service is a function of the social capital of the population (based on the level of education in the urban place), the general democratic political climate, and the personal income of the city. He really thinks that personal/social capital (education) and level of democratization positively drives the level of non-profit primacy in non-state sponsored social and health services (he comes from a theoretical framework that states providing their population services is a function liberalization of society).

In his draft paper and some summary queries to you, the professor makes a variety of statements. He would like you to check/review these statements. Some are simply describing the univariate descriptive statistics (see means/s.d.'s in the table). Some are from the bivariate relations. For example, he makes some statements to the effect that the amount of non-governmental agencies in the urban place has strong positive bivariate relations with each of the variables (see the correlation table he produced). In the end, his major model is to include all 5 independent variables in a model to see their association with NGOs while controlling for the other variables.

He summarizes the following points in the manuscript (see below). As a friendly reviewer of his research you want to be thorough and look at his tables and results from the model. Point to anything you might see as a problem in the tables or in his discussions. You have worked with the professor before and know he's a big thinker but is prone to skip over details, makes mistakes when he cuts/pastes from output to documents; he's a historian by trade and is not the best quantitative researcher under the stars. (Do note the study design and measures used are well respected indicators of the key concepts the professor is using so the issue is really what does he do with the data and statistics he generates/the models he executes).

If you find a need to correct something be explicit in your comments, show him specifically the value you are discussing; if necessary a quick hand calculation or show a formula to indicate what he should do will be useful.

(YOU should have answered 6 questions below…A,B, C1,2,3, and 4)

a) Review the descriptive and correlation/covariances tables provided and point to any suspect or inconsistent values; suggest what may be wrong with the values. (You don't have to recalculate though you may find it useful to do some easy quick calculations --- if something is totally unclear, you can just point out things that he should check out --- i.e. you may not have enough information to state how to fix something but you tell him to double check his results and what might he look for/expect).

b) Your colleague made a big point that all the independent variables he includes have bivariate relations (see table 2) that are all positively related to INGO and that the tertiary enrollment and income per capita seem to be the strongest effects. Since education of the populous is his pet variable he goes into some length describing that as the percentage of individuals enrolled in college/university goes up by one percentage point (say from 60% to 61%) the number of NGO's increase by about 2/3 (r=.658) and the this appears slightly weaker for primary enrollment where the change in NGO per change in percentage points enrollment in primary education is only .558.

c) From his multi-variable model in Table 3, he goes on to make a few more points. (Make sure his points are accurate---he paraphrased these points to you below---provide alternative phrasing if need be! Or point out some better way to get at what he wants to say).

C1- First, he's pleased that his two pet education variables seem to be related to the outcome and that it is clear that tertiary enrollment is a stronger factor (greater than 4 times stronger) than primary education completion (13.3 vs 2.77).

C2-Second, he notes his democracy variable is equal to zero given the 95% confidence interval contains zero, but this confuses him since the correlation with INGO is so high.

C3-Additionally, he is somewhat surprised and discusses the fact that personal income per capita has the strongest effect (Z=9.27) even though the unstandardized coefficient or marginal effect for tertiary education is larger (13.3 vs .074).

C4-Finally, he points to the log likelihood of -1398 and says the model is an appropriate model.

Table 1: Means and standard deviations:

stats	ingo	wbpri	tert	democ	bincpc	popsize
mean	616.702	81.13131	18.89192	651	2349.682	1292.374
sd	472.965	840.56101	11.584	257.1768	2910.477	837.2149

Table 2: Variances are on the diagonal in BOLD, bivariate correlations above the diagonal, and covariances below the diagonal (N=198)

ingo	**223696**	.5579	0.6584	.3832	0.6746	-1.1209
wbpri	7649.73	**840.561**	-.6498	0.4560	0.3239	-0.1733
tert	3607.37	218.231	**134.189**	0.4071	0.4818	0.0597
democ	46612.7	3399.74	1212.72	**66139.9**	0.2619	-0.3245
bincpc	928614	27328.4	16242.7	195997	**8470876**	0.0154
popsize	-47854	-4206.81	578.623	-69864.5	37466.7	**700929**

Measurement –

1. INGO the avg number of non-profit non-government service providers over the previous 3-year period
2. WBPRI the percentage of population between 20-29 who finished elementary school (thru grade 8)
3. TERT the percentage of individuals enrolled in university of population aged 20-29
4. DEMOC a political science established interval scale (range 100 to 1000); higher score more democratic
5. BINCPC average dollars of yearly income per work forced age person
6. POPSIZE – total population in 1000's (e.g. 1292.3 is 1,292,300)

```
Generalized linear models                         No. of obs      =        198
Optimization     : ML                             Residual df     =        192
                                                  Scale parameter =   82678.23
Deviance         =   15874220.2                   (1/df) Deviance =   82678.23

Log likelihood   = -1398.851885
```

TABLE 3:

ingo	Coef.	Std. Err.	Z-vale	[95% Conf.	Interval]
wbpri	2.770221	.9862266	2.81	.8372523	4.70319
tert	13.34946	2.627325	5.08	8.199999	18.49892
democ	.028838	.0964693	0.30	-.1602384	.2179144
bincpc	.0747049	.0080609	9.27	.0589058	.090504
popsize	-.0637849	.0270591	-2.36	-.1168197	-.01075
_cons	27.88112	89.38647	0.31	-147.3131	203.0754