Psychology 308c Midterm Review

Dale Berger

The midterm will focus on topics covered in Howell's Chapters 9 and 15 through 15.9.

Recognition and understanding the meaning of concepts are important, but be sure you can work in the opposite direction as well. That is, you may be asked to analyze and evaluate a situation, and you will need to determine what statistical concepts apply.

Be sure to review your notes, the text, handouts, reserve readings, and homework assignments.

1. What assumptions are made for a test of statistical significance of a sample correlation? How can you judge whether each assumption has been violated? How might you deal with a violation of each assumption?

2. When might you use Fisher's r' transformation? Why? How is the shape of the sampling distribution of r related to the population value of rho?

3. What factors affect the size of a sample correlation? How might a sample correlation coefficient fail to capture the relationship between two variables? Identify specific problems and consider how you would diagnose them and deal with them.

4. What are partial and semipartial correlations? How do you calculate and interpret each?

5. What affects the accuracy of an estimate of Y_i for an individual case? In the equation for the standard error of the estimate for an individual case, explain the logic of why that term matters and why it affects accuracy of prediction in the way it does.

6. What is meant by the standard error of estimate? What is the mean square residual?

7. What is multicollinearity? What is tolerance? How are they related? How do you calculate each? Why are they important in applied regression?

8. How does multiple regression capitalize on chance? How can you deal with this capitalization when you test and report results from a regression analysis? Consider adjusted R squared and Wilkinson's tables.

9. How can you assess the 'importance' of each of several predictor variables?

Psychology 308c Chapter 15 Review

Dale Berger

Sources:

Class notes.
Howell, Chapter 15, except Section 15.15 (logistic regression)
(Cohen, Cohen, West, & Aiken: Chapters 5, 6, 7, and parts of 8)

Study Questions (Give a try before consulting answers):

1. If $r_{Y2} = (r_{Y1})(r_{12})$, then $r_{Y2.1}$, $r_{Y(2.1)}$, ß$_2$, and B_2 are all equal to zero. Give an intuitive explanation (not just a formula) to account for this fact.

2. What is suppression? Describe how you could detect suppression in an analysis and how it might affect your interpretations and strategy for analysis.

3. What is tolerance? Describe how tolerance is calculated, how it is interpreted, and why it is important.

4. Evaluate the arguments: "A variable that is a code for 'missingness' is qualitatively different from scale information on the original variable."

Answers:

1. The product of the two correlations, $(r_{Y1})(r_{12})$, indicates the amount of overlap between variables Y and X2 that one can explain by the overlap of each with variable X1. If the actual correlation r_{Y2} differs substantially from this product, this indicates some special relationship between Y and X2 that cannot be accounted for by the overlap of these two variables with X1.

2. In general, suppression often means that one predictor is more closely related to the "error" portion of another predictor variable than to the "predictive" portion of that variable. "Suppression" is only a label for this special relationship among two or more predictors and a criterion variable. It is necessary to look closely at the nature of the relationship to understand and describe it.

3. Tolerance is a measure of the portion of variance in a variable that cannot be predicted from the other predictor variables in a regression equation. If tolerance is very small, then the variable overlaps substantially with other predictor variables, and it probably does not have much unique predictive information. If tolerance is very low for a variable (i.e., overlap is high), it is arbitrary whether we assign regression weights to that variable or instead assign weights to the other variables that can predict it. Consequently, the standard error for the weight is inflated, and we have low statistical power for detecting the contribution of a variable with low tolerance. When one variable has low tolerance, often one or more other variables also have low tolerance.

4. 'Missingness' is qualitatively different from scale information. If someone fails to divulge their salary, the missingness variable probably says more about their personality than about their salary.

1. Describe briefly how each of the following concepts is important in multiple regression analysis:

   multicollinearity
   dummy coding (indicator coding)
   missing data
   linearity
   partial correlation; semipartial correlation
   nonorthogonality
   homoscedasticity
   standard error of estimate
   etc.

2. You have access to data from 500 adults on a survey of attitudes toward government social programs. The variables include the following:

Variable	Description and code
ATT	Total score on a 30-point scale of support for government social programs; 99=missing (about 10% are missing).
AID	Current government payments to respondent. 1=no current aid; 2=food stamps only; 3=aid program other than food stamps; 4=food stamps plus other aid program; 9=missing (about 2% are missing).
AGE	Respondent's age, range is 21 to 94; 99=missing (about 20% missing).
CITY	City of residence. 1=New York; 2=Chicago; 3=Los Angeles (no missing data)

   Your task is to examine how attitude (ATT) can be explained by AID, AGE, and CITY. You expect that support for government social programs is strongest for younger people and about the same lower level for middle-aged and older people. You also expect that support is much stronger for people who receive food stamps than for those who do not, and you don't expect much effect of receiving aid other than food stamps. You would like to compare attitudes in Los Angeles to attitudes in each of the other two cities. You wish to control for age effects before you test for AID effects, and control for both age and AID when you test for CITY effects.

a) Prepare a table in which you describe in detail the names and coding for each new variable that you create.

b) Discuss your strategy for analysis. In particular, specify the order of entry of the variables, give a rationale for the order, and discuss how you plan to interpret results along the way.

3. Decide whether each of the following statements is true of false, and give a brief rationale for your answer. Your rationale is more important than the T-F answer.

   a) If X and Y are not related in the population, then the population correlation coefficient must equal zero.

   b) If the distributions of X and Y are both very skewed, it is not possible that $r_{xy} = 1.0$.

   c) If every X score is multiplied by 2, the value of r_{xy} is quadrupled.

   d) The maximum value possible for a standardized regression coefficient beta (β) is 1.0.

   e) The procedure recommended by Cohen and Cohen for dealing with missing data (create a 'missingness' variable, etc.) provides a larger N/k ratio than does listwise deletion.

   f) If one is interested in comparing each of three treatment groups with a control group, it would be appropriate to use dummy coding for the treatment groups.

4. Your old buddy Bumble came to see you in an unusually depressed mood. He told you that he was very interested in the correlation between two concepts where another researcher had reported $p < .001$, but a test on his data produced $p = .4817$. What information would you need to help Bumble understand how this could happen? Explain your reasoning for each issue.

5. The next day Bumble was back, looking much happier. "I have been analyzing some of my data, and I found a correlation with p=.000! I have found the impossible! I think I will be famous!" What information would you like to help you decide how much enthusiasm his results deserve?

Use the information on the attached printout to answer the following questions. You will need to do some hand calculations.

6. What is the variable RE1? Describe it in English. Why is it useful for this regression analysis?

7. The correlation between POLIT and RE2 is .202. Test this correlation for statistical significance, and explain in English what it means.

8. Is the strength of religious belief related to POLIT after the effects of AGE, NUMCHILD, and AGXNUM have been removed? (Use $\alpha = .05$)

9. Examine tolerance for AGXNUM in Models 1 and 2, and explain in simple English.

10. Use the observed data to estimate population parameters. How many cases would you need if you desired an 90% chance of detecting a significant ($p<.05$) effect of religion at the point it was entered?

11. If NUMCHILD were entered into the analysis on the first step and AGE on the second step, what would the R squared added be for AGE? Is this statistically significant?

```
TITLE  "COMMAND FILE FOR ICPSR.DAT: ICPSR.SPS".
DATA LIST FILE="A:BERGER1/ICPSR.DAT"
  /IDNUM 1-4  AGECAT 6  SEX 8  RACE 10  LEVEL 12  NUMCHILD 14  EDUC 16-17
   RELIG 19  POLIT 21  INCOME 23-24  WORK 26  HAPPY 28  HEALTH 30  TAXES 32
   COURTS 34  POT 36  CAPITAL 38  GUNS 40  TRUST 42  VOCAB 44-45  AGE 47-48.
VARIABLE LABELS
   IDNUM    "RESPONDENT'S IDENTIFICATION NUMBER"
   AGECAT   "CATEGORIZED AGE"
   SEX      "SEX"
   RACE     "RACE"
   LEVEL    "SOCIAL CLASS"
   NUMCHILD "NUMBER OF CHILDREN"
   EDUC     "EDUCATION"
   RELIG    "STRENGTH OF RELIGIOUS BELIEF"
   POLIT    "POLITICAL CONSERVATISM"
   INCOME   "FAMILY INCOME"
   WORK     "WORK SATISFACTION"
   HAPPY    "OVERALL HAPPINESS"
   HEALTH   "SELF REPORT OF HEALTH"
   TAXES    "ARE FEDERAL INCOME TAXES TOO HIGH?"
   COURTS   "ARE THE COURTS TOO HARSH?"
   POT      "SHOULD MARIJUANA BE LEGALIZED?"
   CAPITAL  "DO YOU FAVOR THE DEATH PENALTY?"
   GUNS     "DO YOU FAVOR REQUIRING GUN PERMITS?"
   TRUST    "CAN PEOPLE BE TRUSTED?"
   VOCAB    "SCORE ON A 10 ITEM VOCABULARY TEST"
   AGE      "AGE".
VALUE LABELS
   AGECAT   1 "1-29"  2 "30-44"  3 "45-60"  4 "61+"  9 "MISSING"
  /SEX      1 "MALE"  2 "FEMALE"
  /RACE     1 "WHITE"  2 "BLACK"  3 "OTHER"
  /LEVEL    1 "LOWER"  2 "WORKING"  3 "MIDDLE"  4 "UPPER"  9 "MISSING"
  /NUMCHILD 8 "8+"  9 "MISSING"
  /EDUC     8 "GRADE SCHOOL"  12 "HIGH SCHOOL"  16 "COLLEGE"
            20 "8+ YEARS OF COLLEGE"  98 "DON'T   KNOW"  99 "MISSING"
  /RELIG    0 "NONE"  1 "STRONG"  2 "MODERATE"  3 "SOMEWHAT"  9 "MISSING"
  /POLIT    1 "EXTREME LIBERAL"  7 "EXTREME CONSERV"  8 "MISSING"  9 "MISSING"
  /INCOME   1 "UNDER 2K"  2 "2K-2999"  8 "8K-9999"  9 "10K  -  14999"
            10 "15K  -  19999"  11 "20K  -  24999"  12 "25K+"  13 "REFUSED"
            98 "DON'T KNOW"  99 "MISSING"
  /WORK     0 "NO JOB"  1 "VERY SATISFIED"  4 "VERY DISSATISFIED"
            8 "DON'T KNOW"  9 "MISSING"
  /HAPPY    1 "VERY HAPPY"  2 "PRETTY HAPPY"  3 "NOT TOO HAPPY"
            8 "DON'T KNOW"  9 "MISSING"
  /HEALTH   1 "EXCELLENT"  2 "GOOD"  3 "FAIR"  4 "POOR"
            8 "DON'T KNOW"  9 "MISSING"
  /TAXES    1 "TOO HIGH"  2 "ABOUT RIGHT"  3 "TOO LOW"  4 "PAYS NONE"
            8 "DON'T KNOW"  9 "MISSING"
  /COURTS   1 "TOO HARSH"  2 "NOT HARSH ENOUGH"  3 "ABOUT RIGHT"
            8 "DON'T KNOW"  9 "MISSING"
  /POT      1 "YES"  2 "NO"  8 "DON'T KNOW"  9 "MISSING"
  /CAPITAL  1 "YES"  2 "NO"  8 "DON'T KNOW"  9 "MISSING"
  /GUNS     1 "YES"  2 "NO"  8 "DON'T KNOW"  9 "MISSING"
  /TRUST    1 "YES"  2 "NO"  3 "IT DEPENDS"  8 "DON'T KNOW"  9 "MISSING"
  /VOCAB    99 "REFUSED"
  /AGE      89 "89+" 99 "REFUSED".
MISSING VALUES  AGECAT TO LEVEL, RELIG, POLIT, WORK TO TRUST (8,9)
   /NUMCHILD (9)   /EDUC, VOCAB, AGE (98,99)
   /INCOME (13,98,99).
RECODE RELIG (0=1)(1,2,3=0)(ELSE=SYSMISS) INTO RE1.
RECODE RELIG (1=1)(0,2,3=0)(ELSE=SYSMISS) INTO RE2.
RECODE RELIG (2=1)(0,1,3=0)(ELSE=SYSMISS) INTO RE3.
COMPUTE AGXNUM = AGE * NUMCHILD.
REGRESSION
  DESCRIPTIVES = DEFAULTS
 /VARIABLES = (COLLECT)
 /STATISTICS = DEFAULT, CHA, F
 /DEPENDENT = POLIT
 /ENT AGE  /ENT NUMCHILD  /ENT AGXNUM  /ENT RE1 RE2 RE3.
```

Descriptive Statistics

	Mean	Std. Deviation	N
POLITICAL CONSERVATISM	3.85	1.30	255
AGE	45.03	18.49	255
NUMBER OF CHILDREN	2.09	1.94	255
AGXNUM	105.3882	117.2614	255
RE1	6.7E-02	.2499	255
RE2	.3765	.4855	255
RE3	.4000	.4909	255

Correlations

		POLITICAL CONSERVATISM	AGE	NUMBER OF CHILDREN	AGXNUM	RE1	RE2	RE3
Pearson Correlation	POLITICAL CONSERVATISM	1.000	.212	.212	.220	-.200	.202	-.135
	AGE	.212	1.000	.320	.511	-.251	.327	-.156
	NUMBER OF CHILDREN	.212	.320	1.000	.925	-.198	.149	-.102
	AGXNUM	.220	.511	.925	1.000	-.182	.210	-.142
	RE1	-.200	-.251	-.198	-.182	1.000	-.208	-.218
	RE2	.202	.327	.149	.210	-.208	1.000	-.634
	RE3	-.135	-.156	-.102	-.142	-.218	-.634	1.000

Model Summary[a,b]

Model	Variables Entered	Variables Removed	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	AGE[c,d]	.	.212	.045	.041	1.27	.045	11.915	1	253	.001
2	NUMBER OF CHILDREN[e,d]	.	.261	.068	.061	1.26	.023	6.231	1	252	.013
3	AGXNUM[f,d]	.	.263	.069	.058	1.26	.001	.299	1	251	.585
4	RE3, RE1, RE2[g,d]	.	.325	.106	.084	1.24	.037	3.392	3	248	.019

a. Dependent Variable: POLITICAL CONSERVATISM

b. Method: Enter

c. Independent Variables: (Constant), AGE

d. All requested variables entered.

e. Independent Variables: (Constant), AGE, NUMBER OF CHILDREN

f. Independent Variables: (Constant), AGE, NUMBER OF CHILDREN, AGXNUM

g. Independent Variables: (Constant), AGE, NUMBER OF CHILDREN, AGXNUM, RE3, RE1, RE2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19.175	1	19.175	11.915	.001[b]
	Residual	407.162	253	1.609		
	Total	426.337	254			
2	Regression	28.999	2	14.500	9.196	.000[c]
	Residual	397.338	252	1.577		
	Total	426.337	254			
3	Regression	29.471	3	9.824	6.213	.000[d]
	Residual	396.866	251	1.581		
	Total	426.337	254			
4	Regression	45.114	6	7.519	4.891	.000[e]
	Residual	381.223	248	1.537		
	Total	426.337	254			

a. Dependent Variable: POLITICAL CONSERVATISM

b. Independent Variables: (Constant), AGE

c. Independent Variables: (Constant), AGE, NUMBER OF CHILDREN

d. Independent Variables: (Constant), AGE, NUMBER OF CHILDREN, AGXNUM

e. Independent Variables: (Constant), AGE, NUMBER OF CHILDREN, AGXNUM, RE3, RE1, RE2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	F	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.182	.210		230.583	.000
	AGE	1.5E-02	.004	.212	11.915	.001
2	(Constant)	3.120	.209		223.254	.000
	AGE	1.1E-02	.004	.161	6.278	.013
	NUMBER OF CHILDREN	.107	.043	.160	6.231	.013
3	(Constant)	3.036	.260		136.825	.000
	AGE	1.3E-02	.006	.187	5.449	.020
	NUMBER OF CHILDREN	.168	.120	.252	1.954	.163
	AGXNUM	-1.2E-03	.002	-.109	.299	.585
4	(Constant)	3.457	.327		111.423	.000
	AGE	7.7E-03	.006	.109	1.730	.190
	NUMBER OF CHILDREN	.116	.121	.174	.916	.340
	AGXNUM	-5.7E-04	.002	-.052	.068	.794
	RE1	-.848	.374	-.164	5.131	.024
	RE2	.119	.238	.045	.250	.618
	RE3	-.303	.233	-.115	1.697	.194

a. Dependent Variable: POLITICAL CONSERVATISM

Excluded Variables[a]

Model		Beta In	F	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	NUMBER OF CHILDREN	.160[b]	6.231	.013	.155	.898
	AGXNUM	.151[b]	4.541	.034	.133	.739
	RE1	-.157[b]	6.235	.013	-.155	.937
	RE2	.149[b]	5.324	.022	.144	.893
	RE3	-.104[b]	2.836	.093	-.105	.976
2	AGXNUM	-.109[c]	.299	.585	-.034	9.358E-02
	RE1	-.139[c]	4.886	.028	-.138	.921
	RE2	.141[c]	4.870	.028	.138	.891
	RE3	-.096[c]	2.440	.120	-.098	.973
3	RE1	-.137[d]	4.607	.033	-.135	.899
	RE2	.142[d]	4.902	.028	.139	.890
	RE3	-.098[d]	2.542	.112	-.100	.969

a. Dependent Variable: POLITICAL CONSERVATISM

b. Independent Variables in the Model: (Constant), AGE

c. Independent Variables in the Model: (Constant), AGE, NUMBER OF CHILDREN

d. Independent Variables in the Model: (Constant), AGE, NUMBER OF CHILDREN, AGXNUM

1. Check the index of Howell, other statistics text books, and use the Internet. You should be familiar with most of the terms. Focus on why each concept is important, not just on definitions.

2. This is a complex problem that could be addressed in many ways. The method described here has some advantages, but other approaches have special advantages of their own. You should be clear what your goals are as you select a particular approach.

a) AID is a nominal variable with four levels. It is necessary to recode it into three variables, each with one degree of freedom, if we wish to capture all of the variance in the original variable. We can consider the four conditions to be the four cells of a 2x2 ANOVA design. Thus, the first recoded variable could be the contrast of no food stamps vs. food stamps (just for ourselves, we can give it the SPSS-ese name of AIDFS), the second recoded variable could be a contrast of no other aid vs. some other aid (AIDO), and the third recoded variable would then be the interaction of the first two recoded variables (AIDFSXO). [Don't you love SPSS-ese?]

Original AID variable	AIDFS	AIDO	AIDFSXO
1 = no aid	-1	-1	1
2 = food stamps only	1	-1	-1
3 = other aid only	-1	1	-1
4 = food stamps plus other	1	1	1

The SPSS syntax to accomplish this is

```
RECODE AID (1,3 = -1)(2,4 = 1)(ELSE = SYSMIS) INTO AIDFS.
RECODE AID (1,2 = -1)(2,3 = 1)(ELSE = SYSMIS) INTO AIDO.
COMPUTE AIDFSXO = AIDFS * AIDO.
```

AGE has a substantial problem with missing data. To capture the information in 'missingness' and to avoid losing about 100 cases, we need to create a 'missingness' variable, AGEMISS. We also wish to test for a curvilinear relationship between AGE and ATT, so we need to create a quadratic component for AGE. We will not look for more complex polynomial components unless there is an apparent pattern in the residuals. The SPSS syntax is

```
RECODE AGE (99 = 1)(ELSE = 0) INTO AGEMISS.
COMPUTE AGESQ = AGE * AGE.
```

CITY is a nominal variable with three levels (df=2), and it must be recoded into two variables. If we wish to use Los Angeles as a reference city for comparisons, we could use effects coding.

Original CITY variable	NYVLA	CHVLA
1 = New York	1	0
2 = Chicago	0	1
3 = Los Angeles	-1	-1

The SPSS syntax for this is

```
RECODE CITY (1=1)(2=0)(3=-1) INTO NYVLA.
RECODE CITY (1=0)(2=1)(3=-1) INTO CHVLA.
```

b) The order of entry of variables into the regression analysis will depend on the goals of the researcher. If we assume that we wish to control for age before we test the effects of AID, and we wish to control for age and AID when we test for CITY effects, the following SPSS commands would be reasonable.

```
MISSING VALUES   ATT (99) AID (9).
REGRESSION
  VARIABLES = COLLECT
 /STATISTICS = DEFAULTS CHA
 /DEPENDENT = ATT
 /METHOD = ENTER AGEMISS  /ENT AGE  /ENT AGESQ
 /ENT AIDO  /ENT AIDFS  /ENT AIDFSXO
 /ENT NYVLA CHVLA.
```

There are seven /ENT commands, so this analysis will fit seven hierarchical models.

Note that we must not define 99 as missing on AGE, because we have defined a missingness variable and we wish to keep all of the cases on age. This may seem peculiar, but when we enter AGEMISS first, we remove variance associated with the distinction between AGE=99 and AGE equal to any other value. The order of entry into the regression analysis is crucial. AGEMISS must be entered prior to the two variables AGE and AGESQ, or the contribution of AGE and AGESQ would reflect the arbitrary values assigned for missing data. When AGEMISS is entered first, the test of R squared added by AGE on the second step is a test of the correlation between age and ATT for those cases with data on both variables. The test of R squared added by AGESQ in the third model is a test of the quadratic component of age in predicting ATT. The test of the beta weight for AGESQ in Model 3 is equivalent, a test of the quadratic component beyond the linear component. Note that the test of beta for AGE in Model 3 is not useful - it would test the contribution of the linear component beyond the quadratic.

We hypothesized a difference in ATT for people who receive food stamps compared to those who do not receive food stamps. There are several possible tests. First, we could examine the simple correlation between ATT and AIDFS. This measures the relationship between the two variables, ignoring all other variables. The test of statistical significance on the correlation is equivalent to an independent t-test comparing the two AIDFS groups on ATT.

If we wish to control for age when we test for the effects of AIDFS, we can examine the table of Excluded Variables for Model 3. This table shows an F test and what the beta value would be for AIDFS if it were entered next (after the three AGE variables). Also shown is the partial correlation between AIDFS and ATT with the three AGE variables partialled out. These statistics address the question of whether people who receive food stamps differ from people of the same age who do not receive foodstamps regarding attitude toward government social programs.

We could also test whether AIDFS contributes beyond the three age variables and AIDO. This test is given in Model 5, where the three AGE variables and AIDO have been entered. In Model 6, we also have a test of the interaction between the two main effects of AID, with the contribution of IDFSXO beyond the main effects of AID and age. It is not useful to test the partial contributions of the two main effects of AID (AIDFS and AIDO) beyond their interaction variable.

Finally, we can examine differences between the cities. The Excluded Variables in Model 6 will show tests for NYVLA and CHVLA. The test for the partial correlation and 'beta in' for NYVLA is equivalent to an independent samples t-test comparing ATT of respondents in New York to respondents in Los Angeles, controlling for the age and AID variables. CHVLA provides a similar test for Chicago vs. Los Angeles.

When both NYVLA and CHVLA are entered into Model 7, the tests of the beta weights for these variables has a different interpretation. When both CITY variables are in the model, a test of the unique contribution of NYVLA can be interpreted as a test of New York compared to the other cities pooled (because CHVLA has removed variance attributable to differences between Chicago and Los Angeles), controlling for the AID category and for the linear, quadratic, and missingness effects of age.

3. a) T If X and Y are not related, then the linear relationship in the population must be zero.

   b) F If X and Y are skewed in exactly the same way, their correlation can be perfect.

   c) F A correlation is a standardized coefficient that is independent of linear transformations of the form **X' = a + bX**, except that the sign on the correlation will change if **b** is negative. The correlation between height and weight is the same whether height is measured in inches or centimeters, and weight in pounds or kilograms.

   d) F When beta falls outside of the range from zero to the correlation, we say there is 'suppression.' See the smog example from class notes.

   e) F Not necessarily, because the Cohen and Cohen procedure adds a new "missingness" variable for each original variable with missing data. It is possible for N/k to be larger or smaller depending on how many missing cases are saved.

   f) T The test of significance for a beta weight when both dummy variables are in the model can be interpreted as a test of the difference between the group coded 1 and the reference group that is coded 0 for both dummy variables.

4. You will need to check carefully the procedures used by both Bumble and the other researcher (his friend, Bimble?).

   Did Bumble use the same measures as the other researcher?
   (If he had a different measure, he might not get the same results.)

   Did the two researchers use the same sampling procedure from the same population?
   (If they didn't, then the results come from different populations, which may have different effects.)

4. (continued)

   Were the sample sizes comparable? (If Bumble had a very small sample, he may not have much power, and so is unlikely to be able to detect an effect even if it is there. If the other researcher had a very large sample, even a small true effect could produce p<.001.)

   Check the plot of raw data for outliers, etc. for both researchers' data if possible. (If there are outliers in either data set, the statistical models and tests of significance may be unstable and not replicable.)

   Were the same statistical models tested? (If different procedures were used, the test results may differ. For example, if one researcher tested a simple correlation while the other researcher tested a partial correlation, the results of the statistical tests may differ.)

5. First, p=.000 in SPSS means p<.0005, not p = zero. Again, there are many possible explanations for a result of p<.0005, only some of which are interesting or important. Be sure to include an explanation of how Bumble could find p<.0005 but not have an interesting finding.

   Check his measures to see if there may be a trivial explanation. (The correlation between height measured in inches and height measured in centimeters will be near 1 and likely will produce p<.0005, but that is not very interesting. Two questions asking the same thing will probably be highly correlated, but no one will be surprised.)

   Check a plot of the data for outliers. (Extreme scores, errors, failure to omit missing data codes, etc. could greatly inflate the calculated sample correlation.)

   How large was Bumble's sample? (If he has a huge sample, even a trivially small effect could be statistically significant.)

   What is the effect size? (An effect size with a confidence interval could be very useful for determining how much enthusiasm the test of statistical significance deserves.)

   How many variables did he consider? (If Bumble began with 100 variables, he would have 100*99/2 = 4950 pairwise correlations to choose from. If he picked the biggest one to test simply because it was the biggest, he is likely to find a highly significant statistical test that may not replicate.)

   The main advice is to get close to the data, including the design, data collection, and analysis, and be sure to look at plots and summary descriptive statistics.

SPSS output questions - Problems 6 - 11. You may need to search through the syntax file and the output, and even do some hand calculations to answer some of these questions.

6. Check the coding on RELIG and the recoding into RE1. RE1 is equal to 1 when the strength of religious belief is "none" and RE1 is equal to zero for all other responses (except missing). Thus the dummy variable RE1 is a contrast between no religious belief vs. all other levels of strength of religious belief from "somewhat" to "strong."

   A set of (g-1) dummy variables can be used to capture all of the information in a categorical variable with (g) levels. In this example, the order of the categories in the original RELIG variable is nonordinal, so simple correlations with RELIG would be **very misleading**. Our set of three dummy variables captures all of the information in the original four-category RELIG variable, including nonlinear relationships.

   An alternative procedure, which may be better, is to recode RELIG into a single new variable (e.g., RELIGX) by changing the order to be ordinal (e.g., 0=none; 1=somewhat; 2=moderate; 3=strong). RELIGX would allow us to assess the linear relationship of the four levels of religious belief strength with POLIT, using only a single degree of freedom. However, if the relationship is strongly nonlinear (e.g., a big difference in POLIT for those with "strong" religious beliefs compared to all others), then the dummy coding may be more efficient, because much of the variance in POLIT could be captured by a single dummy variable (e.g., RE2).

7. The Correlations table shows that the correlation between RE2 and Political Conservatism is .202. If you test this by hand using the t-test formula, you will find $t=3.28$, $df=253$, $p<.01$. The coding for RE2 assigns a larger value for someone with strong religious beliefs than for other respondents, so the significant positive correlation indicates that, on average, people with "strong" religious beliefs are politically more conservative than people who do not have strong religious beliefs.

8. To assess the effects of the strength of religious belief, we need to consider the effect of all three dummy variables at once (RE1, RE2, and RE3). The three variables were added as a set in Model 4. The Model Summary table shows that these three variables together increased R squared by .037, which is statistically significant according the associated F test, $F(3, 248) = 3.392$, $p=.019$.

9. Collinearity for a specified variable is the R squared value obtained when predicting that specified variable from the set of other predictor variables already in the model. Tolerance is simply (1 - collinearity). Collinearity is the proportion of variance of a potential predictor that can be explained by the predictors in the model, while tolerance is the proportion of variance of a potential predictor that cannot be predicted. Note that tolerance and collinearity have nothing to do with the criterion variable, only with the predictor variables.

9. (continued)

The only predictor in Model 1 is AGE. Thus, the collinearity for AGXNUM if it were to be entered next is the proportion of variance in AGXNUM that can be explained by AGE. From the table of Correlations, we see that the correlation between AGE and AGXNUM is .511. This means that $(.511)^2 = .261$ is the collinearity of AGXNUM with AGE, and so the tolerance for AGXNUM if it were to be entered into Model 1 is (1 - .261) = .739. We can find this number in the table of Excluded Variables as the entry for tolerance of AGXNUM in Model 1.

Model 2 uses both AGE and NUMCHILD as predictors, so collinearity of AGXNUM for Model 2 is the multiple R squared obtained by predicting AGXNUM from AGE and NUMCHILD. This multiple R squared could be calculated by hand to give .906. Thus, the tolerance for AGXNUM in Model 2 is (1 - .906) = .094 which is 9.4 E-02 in scientific notation (see Excluded Variables).

The tolerance for AGXNUM is low for Model 2 because it overlaps considerably with the other predictors that are in the model (i.e., AGE and NUMCHILD). The unique part of AGXNUM, however, is quite important theoretically. It is the unique interaction component of AGE and NUMCHILD.

10. We can use the Model Summary table to find the R^2 for Model 3 just before the three RELIG variables were entered, and the R^2 for Model 4 just after the three variables were entered. The R^2 increased from .069 to .106. From this we calculate the effect size $f^2 = [R^2_{Y.AB} - R^2_{Y.A}]/[1 - R^2_{Y.AB}] = (.106 - .069)/(1 - .106) = .0414$. This effect size is just a bit larger than .02 which Cohen and Cohen consider to be small, and much below .15 which is considered to be a medium effect size. Next we go to Cohen and Cohen's Table E.2 (in the handout packet) to find L=14.17 for k_B=3 and power=.90. We calculate $n^* = [L/f^2] + k_A + k_B + 1 = (14.17/.0414) + 3 + 3 + 1 = 349$. Thus, we estimate that we need at least 349 cases to have the desired level of power. You could also use G*Power.

11. r^2 for NUMCHILD and POLIT = $(.212)^2 = .045$, R^2=.068 for Model 2, so R^2 added by AGE = .023. The test of R^2 added is equal to a test of beta for AGE in Model 2, $F_{1,252} = 6.278$, p=.013.

Bonus power question:

A well-designed study produced an effect that was statistically significant at exactly p=.050. You plan to replicate the study as closely as you can, using the same measures and the same size sample selected randomly from the same population. What is the probability that you will observe a statistically significant effect with p<.05? (i.e., What is the power of your proposed study?)

Answer to bonus power question:

The power of your study is only 50%. All else being unchanged, you are just as likely to obtain a smaller effect as you are to obtain a larger effect than that which was reported earlier. If you wish to have power = .90, you will need to use more cases than the previous study.

Psychology 308c: Applied Multiple Regression Sample Questions for Final

Dale Berger

1. Indicate whether each statement is true or false, and give a brief rationale for your answer. Your rationale is worth more than the T-F answer.

T F a) If X and Y are not related in the population, then ρ_{xy} must equal zero.

T F b) If the distributions of X and Y are both positively skewed, $r_{xy} < 1.00$.

T F c) If a scattergram for X and Y shows a quadratic relationship, $r_{xy} < 1.00$.

T F d) If every X score is multiplied by two, the value of r_{xy} is quadrupled.

T F e) The difference between an actual Y score and the predicted Y score from multiple regression is the standard error of estimate.

T F f) It is not possible for a standardized regression coefficient β to exceed 1.00.

T F g) If X and Y are perfectly correlated, the standard error of estimate is zero.

T F h) If the standard error of estimate equals the standard deviation of Y, then $r_{xy} = 1.00$.

T F i) The procedure recommended by Cohen and Cohen for dealing with missing data provides a larger N/k ratio than does listwise deletion.

T F j) If X and Y have a bivariate normal relationship in the population with a standard error of estimate equal to 2.0, and the predicted value of Y = 10 for X = 20, then for the set of cases where X=20, about 68% of the actual Y scores fall between 8 and 12.

T F k) Variables X and Y have a bivariate normal relationship with $\rho_{xy} = .40$. A random sample of 100 cases was drawn, and the 50 cases for which X fell below the 25^{th} percentile or above the 75^{th} percentile were omitted. The r_{xy} based on the remaining 50 cases is likely to be greater than .40.

2. Bumble, who is personnel director for Beasley Industries, has an interesting way to predict job success for applicants. Instead of using regression analysis to generate a prediction equation, Bumble simply adds together years of education, the score on a relevant aptitude test, and years of experience, and he hires the people who have the greatest total scores. Evaluate his procedure.

Refer to the SPSS syntax file and the printout on the following pages to answer Questions 3 to 7. You may need to do your own calculations for some questions. Be sure to show your work.

3. Do males and females differ on vocabulary? Describe and test using $\alpha = .05$.
   a) Test ignoring the effects of education.
   b) Test controlling for the effects of education (hold education constant).

4. Is there an interaction between sex and education in predicting vocabulary? Use $\alpha = .05$.

5. Describe and interpret the effect of RACE1, both
   a) ignoring other variables, and
   b) controlling for education, sex, and the interaction of education by sex.

6. Use the observed statistics as estimates of the population parameters. Estimate the power of the test for the joint effect of the two race variables when the effects of sex, education, and their interaction have already been entered into the model. (Use alpha = .05.)

SPSS syntax file:

```
TITLE  'MULTIPLE REGRESSION FOR SAMPLE FINAL EXAM: MRFINRV2.SPS'.
DATA LIST FILE="C:\My Documents\DATA\BERGER1\icpsr.dat"
 /SEX 8  RACE 10  EDUC 16-17  VOCAB 44-45.
VARIABLE LABELS
  SEX      "Sex"
 /RACE     "Race"
 /EDUC     "Education"
 /VOCAB    "Vocabulary".
VALUE LABELS
  SEX      1 "MALE"  2 "FEMALE"
 /RACE     1 "WHITE"  2 "BLACK"  3 "OTHER"
 /EDUC     8 "GRADE SCHOOL"  12 "HIGH SCHOOL"  16 "COLLEGE"
             20 "8+ YEARS OF COLLEGE"  98 "DON'T   KNOW"  99 "MISSING"
 /VOCAB    99 "REFUSED".
MISSING VALUES  EDUC,VOCAB (98,99).

RECODE RACE (1=1)(2,3=0) INTO RACE1.
RECODE RACE (2=1)(1,3=0) INTO RACE2.
COMPUTE  EDXSEX = EDUC * SEX.

REGRESSION
 /DESC=MEAN, STDDEV, CORR
 /VARIABLES = (COLLECT)
 /STAT=DEF,CHA
 /DEPENDENT = VOCAB
 /ENT EDUC /ENT SEX /ENT EDXSEX /ENT RACE1 RACE2.
```

Descriptive Statistics

	Mean	Std. Deviation	N
Vocabulary	6.05	2.29	266
Education	12.08	3.11	266
Sex	1.57	.50	266
EDXSEX	18.7519	7.4021	266
RACE1	.9098	.2870	266
RACE2	8.271E-02	.2760	266

Correlations

		Vocabulary	Education	Sex	EDXSEX	RACE1	RACE2
Pearson Correlation	Vocabulary	1.000	.511	-.053	.293	.266	-.240
	Education	.511	1.000	-.123	.549	.076	-.105
	Sex	-.053	-.123	1.000	.736	.070	-.069
	EDXSEX	.293	.549	.736	1.000	.098	-.110
	RACE1	.266	.076	.070	.098	1.000	-.953
	RACE2	-.240	-.105	-.069	-.110	-.953	1.000

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.511[a]	.261	.258	1.97	.261	93.171	1	264	.000
2	.511[b]	.261	.255	1.97	.000	.036	1	263	.849
3	.511[c]	.261	.253	1.98	.000	.161	1	262	.688
4	.568[d]	.323	.310	1.90	.061	11.794	2	260	.000

a. Predictors: (Constant), Education

b. Predictors: (Constant), Education, Sex

c. Predictors: (Constant), Education, Sex, EDXSEX

d. Predictors: (Constant), Education, Sex, EDXSEX, RACE1, RACE2

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	361.357	1	361.357	93.171	.000[a]
	Residual	1023.906	264	3.878		
	Total	1385.263	265			
2	Regression	361.499	2	180.750	46.434	.000[b]
	Residual	1023.764	263	3.893		
	Total	1385.263	265			
3	Regression	362.129	3	120.710	30.911	.000[c]
	Residual	1023.134	262	3.905		
	Total	1385.263	265			
4	Regression	447.231	5	89.446	24.792	.000[d]
	Residual	938.032	260	3.608		
	Total	1385.263	265			

a. Predictors: (Constant), Education

b. Predictors: (Constant), Education, Sex

c. Predictors: (Constant), Education, Sex, EDXSEX

d. Predictors: (Constant), Education, Sex, EDXSEX, RACE1, RACE2

e. Dependent Variable: Vocabulary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.513	.486		3.117	.002
	Education	.376	.039	.511	9.653	.000
2	(Constant)	1.428	.659		2.168	.031
	Education	.377	.039	.512	9.585	.000
	Sex	4.699E-02	.246	.010	.191	.849
3	(Constant)	2.032	1.642		1.238	.217
	Education	.328	.128	.445	2.549	.011
	Sex	-.340	.993	-.074	-.342	.733
	EDXSEX	3.176E-02	.079	.103	.402	.688
4	(Constant)	-1.859	2.188		-.850	.396
	Education	.321	.125	.436	2.573	.011
	Sex	-.413	.961	-.090	-.430	.668
	EDXSEX	3.163E-02	.076	.102	.414	.679
	RACE1	4.253	1.357	.534	3.133	.002
	RACE2	2.649	1.418	.320	1.868	.063

a. Dependent Variable: Vocabulary

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	Sex	.010[a]	.191	.849	.012	.985
	EDXSEX	.018[a]	.285	.776	.018	.699
	RACE1	.228[a]	4.457	.000	.265	.994
	RACE2	-.189[a]	-3.628	.000	-.218	.989
2	EDXSEX	.103[b]	.402	.688	.025	4.303E-02
	RACE1	.229[b]	4.447	.000	.265	.988
	RACE2	-.189[b]	-3.617	.000	-.218	.982
3	RACE1	.230[c]	4.462	.000	.266	.986
	RACE2	-.192[c]	-3.650	.000	-.220	.978

a. Predictors in the Model: (Constant), Education

b. Predictors in the Model: (Constant), Education, Sex

c. Predictors in the Model: (Constant), Education, Sex, EDXSEX

d. Dependent Variable: Vocabulary

HINTS: (try them yourself first)

1. T F T F F F T F F T F (Be able to explain your reasoning)

2. Consider standard deviations.

3. a) you can do a t-test on r by hand; b) consider the second model
   Note: One student interpreted the correlation between Sex and Vocabulary (r = -.053) as follows: "Vocabulary decreases when sex increases." While that may be true, it is not a good description of the results in this data set.

4. Consider R square change.

5. Check the command syntax carefully, then look at (a) r and (b) 'Beta in' for excluded variables.

6. Use G*Power with an effect size f^2 = .about .09 (you should use the exact, more precise value).