Packet of Class Materials for

Applied Multiple Regression

CGU Psychology 308c
Spring 2014

(Additional documents and readings are available on Sakai – see p. 3.)

Instructor: Dale Berger
Teaching Associates: Nic Barreto, Maggie Burkhart, Val Dubon, Aly Lopez, Stephen Weltz

Syllabus

Multiple Correlation and Regression — Psychology 308c
Dale Berger — Spring 2014

Course Description:

This 2-unitcourse introduces the logic and application of correlation and multiple regression models with the goal of helping the student to develop knowledge and skills needed to use regression analyses appropriately (diagnostics, applications, interpretations, presentations). We use material from Chapters 9, 10, and 15 of Howell (8th ed.), plus supplementary material, especially from Cumming (2013) and Cohen et al. (2003). SPSS is used extensively and Excel applications are introduced.

Required Text:

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont: Wadsworth

Supplementary Sources:

Cumming, G. (2013). The New Statistics: Why and how. *Psychological Science*. Advance online publication. doi: 10.1177/0956797613504966 (on Sakai under Resources)

Cohen, J., Cohen, P., West, S. G., & Aiken, L.S. (2003). *Applied multiple regression/ correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. [CCWA: THE reference book for regression in the social sciences.]

Stevens, J. P. (2007). *Intermediate Statistics* (3rd ed.). Routledge. [This inexpensive paperback is accessible, filled with examples and useful advice. ISBN: 978-0-8058-5466-4]

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon. [This is an excellent resource for common multivariate methods.]

Materials on reserve on Sakai, Howell's http://www.uvm.edu/~dhowell/fundamentals8/ plus other online resources you may find (See what you can find – tell me about good resources).

Professor:

Dale Berger 909-621-8084 (SBOS office); 607-3714 (direct line)
Office Hours in ACB101: Most Tuesdays 1:00-3:00
Also by appointment and by email: dale.berger@cgu.edu

Teaching Associates: 909-621-8084

Aly	Albertina.Lopez@cgu.edu	Monday	4:00 -- 5:50	ACB 119
Val	Valeska.Dubon@cgu.edu	Tuesday	4:00 -- 5:50	ACB 208
Maggie	Margaret.Burkhart@cgu.edu	Wednesday	11:00 -- 12:50	ACB 119
Nic	Nicolas.Barreto@cgu.edu	Wednesday	4:00 -- 5:50	ACB 208
Stephen	Stephen.Weltz@cgu.edu	Thursday	4:00 -- 5:50	ACB 208

Class Meetings and Labs:

Lecture (Dale) Tuesday and Thursday, 9:00-10:50am, Burkle 16
Lab (TAs) Burkle 16 Tuesday and Thursday, 8:00-9:00am (AL & SW) and 11:00-11:50 (NB, MB, VD)

Additional Sources:

Our Sakai website will provide essential online support for the course. All enrolled students are automatically included in the roster for this website. You may access the login at **https://sakai.claremont.edu:8443/portal/login .** Let me know if you have difficulty. Also, the Internet is a wonderful resource. A good entry point is **http:/wise.cgu.edu**.

SPSS Instruction:

Howell provides a very nice introduction to SPSS. You can access the manual here: http://www.uvm.edu/~dhowell/fundamentals7/SPSSManual/SPSSLongerManual/SPSSLongerManual.html Howell includes data sets and a glossary of statistical terms and support for our text. I encourage you to explore http://www.uvm.edu/~dhowell/StatPages/StatHomePage.html

UCLA provides extensive statistical support, including tutorials and demonstrations of SPSS at http://www.ats.ucla.edu/stat/spss/ If you find other useful resources, let me know.

Student Learning Outcomes:

By the end of this course, students will be able to do the following:

1. Translate verbal descriptions of various regression designs into SPSS designs.
2. Assess the appropriateness of regression analysis, considering assumptions and goals
2. Conduct regression analysis using SPSS.
3. Interpret SPSS output.
4. Describe findings for sophisticated and for lay audiences (the APA and the PTA).

Homework:

There are three exercise sets (each worth 20 points) and one project (worth 30 points). Late homework will be penalized one point per day up to five points. Please be neat with your homework: Use a word processor for text where possible, identify problems by number, show your work, circle or highlight your answers, and present answers for computer problems on a separate sheet. You may work in teams of 2 or 3 people on homework, though you may work individually if you prefer. If you work in a team, it is critically important that every member of the team be engaged in all phases of the project. Each member should be able to do all aspects of the project, including all computer applications. If you receive a score less than 17, you may resubmit within a week to potentially improve your score to a maximum of 17.

Study Partners:

We encourage each student to work with a study partner or two. Working with your study partner on homework exercises and study questions will help you learn how to explain and present statistical information to others. We will help you find a study partner if you wish.

Examinations:

The midterm examination will be on Tuesday, February 18, covering Chapter 9 and 15.1 through 15.9 plus handouts. The final examination on Tuesday, March 11 will be comprehensive, but emphasize material covered after the midterm. Both exams are **9:00-11:50,** in-class, open-book, open-notes, calculator but no computer. We will review the midterm exam during class on Thursday February 20 at 9:00 and review the final exam on Thursday March 13 at 9:00.

Grading:

The course grade will be based on total points: three homework exercise sets (60), midterm examination (60), project (30), and final examination (100).

CGU Policy on Grading Standards:

Letter Grade	Grade Points	Description	Learning Outcome
A	4.0	Complete mastery of course material and additional insight beyond course material	Insightful
B	3.0	Complete mastery of course material	Proficient
C	2.0	Gaps in mastery of the course material; not at level expected by the program	Developing
U	0.0	Unsatisfactory	Ineffective

CGU Policy on Scientific and Professional Ethics:

The work you do in this course must be your own. Feel free to build on, react to, criticize, and analyze the ideas of others but, when you do, make it known whose ideas you are working with. You must explicitly acknowledge when your work builds on someone else's ideas, including ideas of classmates, professors, and authors you read. If you ever have questions about drawing the line between others' work and your own, ask the course professor who will give you guidance. Exams must be completed independently. Any collaboration on answers to exams, unless expressly permitted, may result in an automatic failing grade and possible expulsion from the Program.

Multiple Correlation and Regression Schedule
Psychology 308c Dale Berger Spring 2014

This schedule should be considered tentative because it may be modified as the course progresses. It will be helpful to at least skim through relevant or assigned readings before class. Detailed study can be done after the class covering the material. Focus on class handouts and Howell, and use the supplementary sources for additional help.

Weekly topics and sources

1. Introduction, history of regression, regression vs. correlation, covariance, calculation of r, regression equation, proportion of variance explained, standard error of estimate, tests of significance for r and b, accuracy of estimation of an individual case, SPSS applications. Howell, Chapter 9. (CCWA: Chapter 1 and Chapter 2 through Section 2.7.)

2. Sampling distribution for r, Fisher's r to r' transform, confidence intervals for rho, testing whether rho has a value other than zero, testing for a difference between two independent dependent correlation coefficients, pooling estimates of rho, comparing two dependent correlation coefficients, factors that affect correlation (restriction of range, pooled populations, shape of the distributions, reliability), robustness, power analysis for tests of r, determination of sample size, assumptions for statistical tests. Howell: Chapter 9. (CCWA Chapter 2.)

3. Multiple R, partial correlation, semipartial correlation, multiple correlation with two predictors, beta, b, R^2 added, tests of significance, shrunken or adjusted R^2, adjustment for selection of predictors from a larger set, Wilkinson's tables for stepwise analyses, suppression, tolerance, multicollinearity, SPSS REGRESSION. Howell: Chapter 15 through Section 15.9. (CCWA: Chapters 3 and 4.)

--------- Midterm exam covering Chapter 9 and 15.1 through 15.9 plus handouts ---------

4. Sets of predictors, power analysis for sets of predictors, "importance" of a predictor, diagnostics (residual analysis, outliers, leverage, influential data points), transformations. Howell: Chapter 15 through Section 15.12. (CCWA: Chapter 5.)

5. Nominal scales, coding (dummy, effects, contrast coding), ANOVA with multiple regression, nonorthogonal designs, interactions, mediation, moderation. Howell: Chapter 15 through Section 15.14. (CCWA: Chapter 6.)

6. Power functions, Durbin-Watson statistic, strategies for analysis, missing data, analysis of covariance. Howell: Chapter 15 through Section 15.14. (CCWA: Chapters 7, 8, and 9; 10 and 11 have relevant supplemental material); handouts.

Important Dates:

1/30	Homework 1 due	3/6	Homework 4 due
2/11	Homework 2 due	3/11	Final exam
2/18	Midterm exam	3/13	Review final exam
2/27	Homework 3 due		

Psychology 308 Dale Berger **Errors in Howell (8th ed.)**

Our textbook is Howell, D. C. (2013), *Statistical Methods for Psychology* (8th ed.), Belmont, CA: Wadsworth. Howell is good about posting errors that people find in his books. You can access this at http://www.uvm.edu/~dhowell/methods8/Errata/Errata_for_Methods_8th_ed.html I strongly recommend that you go to this site and make the corrections in your book.

Howell recently posted error that students from last year and I found. Here are a couple of the most important ones, but check his errata site for others.

p. 293, Section 9.16: The formulas for computing power for tests of Pearson's r should use Fisher's transformed value rather than untransformed values of r and rho. Thus, the first formula should be $d = \rho'_1 - \rho'_0$ rather than $d = \rho_1 - \rho_0$; the next formula should be $\delta = d\sqrt{N-1} = \rho'_1\sqrt{N-1}$

Thus, the example will produce δ = 2.17 rather than 2.10, and power is estimated to be .59 rather than .56. Computation of the N needed for power = .80 gives N=83 rather than N=88. The correction doesn't matter much here, but it is more important with larger values of r or rho.

G*Power can be used to solve this problem – it gives the answer of N=82.

The formula for d would be clearer if parentheses were used (thanks to Nic for pointing this out): $d = (\rho'_1 - \rho'_0) = (\rho'_1 - 0) = \rho'_1$ when testing the null hypothesis that population ρ = 0.

p. 527, first paragraph in Sample Sizes, end of line 4: this should be an R squared, not R. Howell pointed out this error in an earlier edition, but the error is still in the 8th edition but no note in the errata section. This error is especially important because the elegant little formula is useful for building intuitive expectations of chance effect sizes.

This section would be a good place to remind readers that tests of the unique contribution of an individual variable can require very large samples when predictors are correlated with each other (collinearity is high), as described by Maxwell (2000).

Please tell me if you find any additional errors - Let the hunt begin!

Topics:

Pearson product moment correlation coefficient
Tests of statistical significance; assumptions; violations of assumptions
Introduction to SPSS and Excel for regression

Sources:

Howell, Chapter 9
Class notes and handouts.
Havlicek & Peterson, (1977). *Psychological Bulletin, 84,* 373-377. (read carefully)
(This classic paper is on Sakai)
CCWA: Chapter 1 and Chapter 2 through Section 2.7

Exercises:

The goal of these exercises is to gain an understanding of how a correlation coefficient and a regression equation are computed, and how these statistics are affected by extreme scores. We also will practice explaining the findings to a nontechnical audience. You will be asked to compute various statistics by hand, with Excel, and with SPSS.

If you are working with a group, my strong recommendation is that you each do the problems individually, including the SPSS analyses, think about the interpretations, and then meet to discuss and collaborate on writing up the report which will include only one set of answers to the three problems.

A teacher was interested in using age of students to predict "self-actualization." He collected the following data from a class at a community college:

Student Number	1	2	3	4	5	6	7	8	9	10	11
Age (X)	17	17	18	18	19	19	19	20	21	22	58
Self-Actualization (Y)	60	43	57	50	58	63	65	71	75	89	45

1. Enter these data into SPSS for analysis. First, be Bumble and complete Problem 1 (See the MRC01 handout for a step-by-step example of simple regression in SPSS.) Attach your annotated output (label where you found answers to the questions).

a) What is the correlation between age and self-actualization?
b) Calculate a t-test for the correlation. Use the formula on p. 280 of Howell.
c) Find the *p* value for your t-test. (You can use StatWISE or find it in the output)
d) What is the regression equation for predicting self-actualization from age?
e) Use the equation to predict self-actualization for the teacher who is age 30.

f) Use SPSS to generate a scatterplot of the data.
g) Assess whether a linear regression model is appropriate, whether the assumptions for the tests of statistical significance are met.

2. Next, let's not be Bumble. The first thing WE would have done with a new set of data is to take a close look at the data, as Bumble finally did in 1f and 1g.
   We see that one case is an extreme outlier.
   Remove Student #11 who is age 58 with self-actualization of 45.
   Redo Problem 1, all parts. [Not to be Bumble, we would do parts 1f and 1g first!]
   Attach annotated output.

3. Write up a brief (1-2 pages?) summary of your findings and interpretations for Exercises 1 and 2.
   a) Begin with a one paragraph abstract or executive summary for the teacher.
   b) In the body of the summary, include relevant numeric information and tests, along with explanations. There is no need to present invalid statistics in great detail, though you should explain why you believe certain statistics are not valid.
   c) Include a discussion of the effects of the extreme score on the computed statistics and on the validity of the statistical tests of significance. Consider the findings of Havlicek and Peterson.
   d) Include a discussion of issues in predicting self-actualization for the teacher (age 30).

4. Apply Excel to the first 10 cases from Problem 1– attach your Excel worksheet with name/date. **Bold** and label the output where you find the answers:

   a) Use the regression equation from Problem 2d to find the predicted Y_i value for each X_i.

   b) Use Excel to calculate $SS_{TOT} = \sum(y_i - \bar{y})^2$; $SS_{REG} = \sum(\hat{y}_i - \bar{y})^2$; $SS_{ERR} = \sum(y_i - \hat{y}_i)^2$; and find these numbers in your SPSS output, circle, and describe.

   c) Verify that $SS_{TOT} = SS_{REG} + SS_{ERR}$. Note that SS_{TOT} is the numerator of a common formula for computing an estimate of the variance for Y. Part of the variance can be predicted from X using SS_{REG}, while part cannot be predicted from X (SS_{ERR}).

   d) What proportion of SS_{TOT} for Y cannot be predicted with X? What proportion of SS_{TOT} for Y can be predicted with X? Compare this proportion to r squared. Interpret.

Study Questions: Practice explaining these to your skeptical study partner(s)

Can Pearson's r be calculated for variables that are not distributed normally?
Will SPSS compute a correlation between two nominal variables?
How would you interpret such a correlation?
How might the test of significance be affected when assumptions are violated?
Under what circumstances would you expect an outlier to be most consequential?
Could you test a correlation between two dichotomous variables? How?

Hints and Answers (Give these problems a good try before you look at the answers):

You need to create a SPSS data file. One way is to enter the student number, X, and Y in three successive columns, with data from each student on a separate line. You can use variable labels like **student**, **X**, and **Y**.

For part d): Click on Statistics; Regression; Linear; specify the regression model. See the handout SPSS Point-and-Click Example: MRC01.

1. a) r = -.291
   b) t(9) = -.914, ns [$t_{9,.05/2} = 2.262$] (You can calculate by hand)
   c) *p*=.385 (two-tailed)
   d) $Y_i' = 68.935 + (-.332)*X_i$ (Use Analyze, Regression, Linear, …)
   e) $Y_i' = 68.935 + (-.332)*30 = 58.975$, or about 59.0
   f) See MRH1
   g) Consider the population for which you have data

2. You need to exclude data for Student 11. Use the 'filter' under Data; Select Cases, and enter **student** < 11. These commands will exclude cases with **student** ID greater than 10.

   a) r = .914
   b) t(8) = 6.379, *p*<.001, [$t_{8,.05/2} = 2.306$]
   c) SPSS reports *p*=.000; you should report *p*<.001
   d) $Y_i' = -75.442 + (7.292)*X_i$
   e) $Y_i' = -75.442 + (7.292)*30 = 143.318$!
   f) and g) Consider the population for which you have data

3. Although the results of Havlicek and Peterson show that the tests of statistical significance for r are quite robust, this example illustrates that a really extreme outlier can wreak havoc, especially in a small sample. Also, predictions are risky for cases that fall outside of the range of the observed data – such predictions are based on strong assumptions that cannot be validated from the data (linearity beyond the range of observed data).

4. It is important to develop your skills with Excel - Excel is especially handy for repetitive applications of formulas. You can copy and paste data from SPSS into Excel and vice versa. You can make separate columns for X, Y, predicted Y, deviations of Y from the mean of Y, deviations of predicted Y from the mean of Y, and deviations of Y from the predicted Y. Then you can make columns of the squares of each of the last three terms. Be sure to document your worksheet well – what is in each column, what the purpose of the sheet is, date, and your name(s).

4. a) See handout MRH2. Enter data first, then apply the formula $\hat{Y}_i = -75.442 + 7.292\ X_i$

   b) Apply the formula to produce $\hat{Y}_1 = 55.82$, and by subtraction $(Y_1 - \hat{Y}_1) = 50 - 55.82 = -5.82$. Follow the same procedure to generate all ten predicted values and their deviations from observed. $SS_{TOT} = 1526.9$, $SS_{REG} = 1276.0$, and $SS_{ERR} = 250.9$.

   c) It checks out: 1275.9 + 250.9 = 1526.9 (within rounding error).

   d) The proportion of Y variance that cannot be predicted from X is 250.9/1526.9 = .1644. Thus, the proportion that can be predicted is 1 - .1644 = .8356 or 1275.9/1526.9. If we take the square root, we get .914 which is the correlation between X and Y. That is, r^2 = the proportion of variance in Y that can be predicted from X. This relationship is a key concept in regression analysis. Print your Excel worksheet

Thoughts to Guide Responses to Study Questions:

One can calculate a Pearson's r between any two numerical variables, even if they are not normally distributed. A computer program may report a value for r to six or more decimal places, and never worry about whether this is appropriate. (Even a computer program will choke, however, if one of the "variables" is a constant!)

When it comes time to interpret a correlation, you need to examine the distributions and know what the original scores represent. The statistic r is not a good descriptor of a relationship if the relationship is not linear. If you have a nominal variable with more than two levels (where the numbers are used as labels for different groups), a big red flag should wave at you. Stop! The simple correlation (that SPSS will compute, if you ask it to) will not be very useful. The variance of a nominal variable with more than two categories is not meaningful, so it doesn't help to know how much of the variance of a nominal variable can be predicted from another variable. You can recode a nominal variable with k levels into (k-1) dichotomous 'dummy' variables, and then use multiple regression procedures to analyze the relationship between the original nominal variable (as represented by the k-1 dummy variables) and other measures.

The test of statistical significance for correlations assumes that in the population the residuals from the predicted values of the dependent variable (Y) are distributed normally, with homogeneous variance for all levels of the independent predictor variable (X). Note that the X variable does not need to be continuous. It can even be a dichotomous variable, such as gender. The study by Havlicek and Peterson shows that the correlation coefficients and corresponding tests of statistical significance are very robust with respect to modest departures from normal residuals. However, with large departures the t-test may be compromised. Extreme outliers can distort correlations substantially. With two dichotomous variables, the Pearson correlation is equal to the phi coefficient, and an appropriate test of the null hypothesis that there is zero correlation between two dichotomous variables is the chi-square test of independence, not t.

Topics:
Confidence intervals, reliability, pooling correlation estimates, power

Sources:
Class notes and Sakai materials
Howell Chapter 9

Exercises: Interpret all findings in nontechnical language

1. Use the ten data points from Problem 2 of Exercise Set 1:

   a) Find a 95% confidence interval for the population correlation, ρ. Interpret.
   b) Why is this interval not symmetric around the sample correlation?
   c) Predict Y_i for $X_i = 21$.
   d) Find a 95% confidence interval for Y_i when $X_i = 21$. Interpret.
   e) Predict Y_i for $X_i = 58$.
   f) Find a 95% confidence interval for Y_i when $X_i = 58$.
   g) Why is the confidence interval wider when $X_i = 58$ than when $X_i = 21$? Explain in intuitive terms.

2. Suppose you are interested in the relationship between attitudes toward reducing air pollution (X), and behaviors that reduce air pollution (Y). With a sample of n=19 you found r_{xy}=.31. The reliability of your measure of X is .30 and the reliability of Y is .40.

   a) Estimate r_{xy} for perfectly reliable measures of X and Y (if they could be found).
   b) Test H_o: $\rho_{xy} = 0$. (Hint: Do not use the r you estimated in Exercise 2a.)

3. Suppose that three experiments with exactly the same two variables produced r_{xy}= .40 with n=18, r_{xy}=.35 with n=23, and r_{xy}=.52 with n=15.

   a) Can the correlations be considered homogeneous?
   b) Assuming that the answer to a) is "yes," pool the three estimates of ρ_{xy}.
   c) Test H_o: $\rho_{xy} = 0$, using the pooled estimate from b).

4. From one sample of n=87 cases, the following correlations were observed between three variables: $r_{xy} = .40$, $r_{xz} = .30$, and $r_{yz} = .70$. Test H_o: $\rho_{xy} = \rho_{xz}$. [Hint: Dependent!] You may use StatWISE. From http://wise.cgu.edu , go to WISE Stuff, scroll down to Excel Downloads, select StatWISE. The program replaces common statistical tables.

5. How many cases should you sample if you would like a 90% chance of getting a significant correlation (two-tailed, α=.01) if the population correlation is .30?

6. Give a very brief description of the data set you plan to use for your final project (see attached description). Include a description of the central research question, the source of the data set, number of cases, and description of the variables that you will use.

Study Questions (Practice explaining these to someone. No need to turn them in.)

7. What is the sampling distribution of r_{xy}? Discuss how the sampling distribution of r_{xy} is influenced by the sample size and the population correlation, ρ_{xy}. Give an explanation that your friend Bumble can understand.

8. List the factors that influence the accuracy of a regression estimate of Y_i, and explain how each factor influences the accuracy.

9. In class we discussed several characteristics of X and Y that affect the size of r_{xy} (restriction of range, combining data from disparate populations, shape of the distributions, reliability). For each, give a description of the nature of the effect and an intuitive explanation of why it occurs (i.e., explain the effect to a first year statistics student).

10. What is the appropriate way to test for a correlation between two dichotomous variables? (Hint: You could represent these data in a 2x2 table. How do you test for a relationship between two dichotomous variables in a 2x2 table?)

Hints:

1. See Section 9.9 in Howell; also the handout on Formulae for MR/C may be useful.

   More hints for d and f: You need to use the $s'_{Y.X}$ formula in Section 9.9.

   The standard error of estimate is an unbiased population estimate that uses (n-2) in the denominator.

   The unbiased standard error of estimate = $\sqrt{\frac{(SS_{TOT})(1-r^2)}{(n-2)}}$

5. Pay close attention to alpha and desired power. There is error in Howell not noted in his errata web page but included in the summary of errors provided in this packet. You need to use the Fisher transformed value for rho rather than rho when calculating delta.

Answers:

1. a. prob [.67 < ρ_{xy} < .98] = 95% (Consult notes and text for interpretation.)

   b. The sampling distribution for r_{xy} is not symmetrical. Sample correlations cannot be larger than 1 or smaller than -1. With a large population correlation (e.g., .80), sample correlations cannot be much larger than the population value (limit is 1.0) though they could be much lower.

   c. $\hat{Y}_i$ for $X_i = 21$ is 77.69

   d. prob [63.14 < Y_i < 92.24] = 95% (Consult notes and text for interpretation.)

   e. $\hat{Y}_i$ for $X_i = 58$ is 347.46

   f. prob [243.66 < Y_i < 451.26] = 95%

   g. Because of sampling error, the estimate of the slope of the regression line may be off by a bit. The consequences of a small error in the slope of the regression line is magnified as one moves farther from the mean. Thus, errors in estimates of Y are larger for points farther from the mean on X. Also, note that the confidence interval assumes that the relationship is linear for all values of X, and that errors are normally distributed about the regression line, with equal variance for all values of X. When a value for X is taken far from the range where data have been observed, there is no way to check whether these last assumptions are valid. If they are not valid, the estimates can be very inaccurate.

2. a. $r_{x^*y^*} = .896$
   b. $t = 1.34$; compare to $t_{17, .05/2} = 2.11$

3. a. $\chi^2 = .338$; compare to tabled $\chi^2_{2, .05} = 5.99$
   b. pooled r = .41
   c. With t, use df = (18-2) + (23-2) + (15-2) = 50.
   With z, use Σ(N-3) = 47 for calculating the error term. z = 3.00

4. t (84) = 1.29 from StatWISE; the t-test in Howell gives the same value with df = N-3 = 84.

5. See Section 9.16 but use rho prime rather than rho in the formula to define delta; if you use delta of 3.85 with the corrected formula, n is about 156. Procedures in CCWA give n=153, Cohen's power program gives n=157, G*Power gives n=154.

6. Recommendation: Plan a project that focuses on conceptual issues rather than on the statistics. Use the statistical analysis as a tool to describe your data, test your hypotheses, and provide support for your interpretations.

Topics:

Multiple regression with multiple predictors, SPSS REGRESSION
Partial correlation, choice of predictors
Multicollinearity, suppression, alpha inflation with stepwise regression

Sources (see resources online in Sakai for our course)**:**

Class notes. Howell, Chapter 15 to Section 15.8; CCWA: Chapters 3 and 4
Wilkinson, L. (1979). *Psychological Bulletin, 86,* 168-174.
Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 213-217.
Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499-510.
Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 4,* 434-468.

Exercises:

1. Your task is to conduct and interpret a hierarchical regression analysis where you have two sets of predictors, including a categorical variable. The data set is available on Sakai in a SPSS Windows file BANK.SAV. All summaries should be in nontechnical language.

   For the clerical staff only, we are interested in whether the variables ethnic group and gender (Set B: MINORITY and SEX) are related to current salary after the effects of educational level and seniority on the job (Set A: EDLEVEL and TIME) have been removed as covariates. Limit your analyses to clerical staff. Dummy code MINORITY and SEX.

   a. Check the shape of the distribution of SALNOW. Generate a histogram, and find the skew and kurtosis.

   b. Compute LNSALNOW = LN(SALNOW), and check the shape of the distribution of LNSALNOW. Generate a histogram, the skew, and kurtosis.

   c. Use hierarchical multiple regression analysis to test the statistical significance of the contributions of each set, Sets A and B, to predicting LNSALNOW. Provide a table in which you report simple r for each predictor, R^2 added, with F, df, and p for the R^2 added for each set, and B and SE(B) for each variable in the final model. You will need to create your table, and not just cut and paste SPSS tables.

   d. In no more than three pages, summarize and interpret your findings for the entire problem. (Include a summary of the research question, a brief description of the design and sample, check of assumptions, interpretation of findings in the table, including B and beta values and p-values. Your description should include information on the size and direction of effects.) Attach annotated output, indicating where you found useful information, especially information included in your summary.

2. a. Use a t-test for independent groups to test for a difference in LNSALNOW for males and females. Report the means, t-test value with df, and the p value.
   b. Compute a correlation between gender and LNSALNOW. Report the r, a t-test of statistical significance with df and p-value.
   c. Compare your findings and explain why these results are the same.

3. Suppose you used SPSS stepwise regression, and allowed the program to select the 6 best predictors out of a set of 20 possible predictors (N=50). You found $R^2 = .530$.

   a. Test this R^2 for significance with the procedure SPSS uses (hand calculation).

   b. Test R^2 with a conservative test by assuming k=20 predictors (hand calculation).

   c. Test R^2 using the tables provided by Wilkinson (in your packet) (compare to table value).

   d. Evaluate the merits of each of these three procedures. Which would you use? Why?

   e. Suppose the correlations were all zero in the population. How big would you expect the sample multiple correlation squared to be if you used 20 predictors and N=50?

4. Assume that in the population $R^2_{Y.AB} = .40$ and $R^2_{Y.A} = .30$, and that $k_A = k_B = 2$:

   a. Calculate the sample size necessary to have an 80% chance of rejecting the null hypothesis that R^2 added by Set B is zero in the population (use $\alpha = .01$).

   b. When might such calculations of sample size be useful?

Study Questions (not to be turned in; practice with a study partner):

5. Describe each of the following concepts and explain how they may be used and why they are important: partial correlation, semipartial correlation, multiple correlation, R squared added, beta weight, B weight, tolerance, power analysis for R, shrinkage formula, centering

6. In what circumstances would you test a group of variables together as a set, as opposed to testing each variable separately?

7. Bumble replicated a study by using the same criterion variable and the same five predictor variables. Some of the regression coefficients in his equation were very different from the coefficients in the equation the other researcher reported. Bumble concluded that the other researcher screwed up. Identify and discuss factors that might contribute to differences in the equations between the published study and Bumble's replication.

8. When you have a high degree of 'multicollinearity,' the size and tests of statistical significance for regression coefficients may be misleading. Why? How can you detect problems, and how do you deal with them?

9. Evaluate and elaborate the following statement:

   The size and significance of the contribution of any variable (or set of variables) can be manipulated by changing the order in which the variable(s) is added to a hierarchical analysis. Thus, one should always use stepwise analyses and allow the program to determine the order of entry for predictors.

Answers:

1. a. You can click **Data; Select cases; If condition; If; jobcat=1; continue; filtered; OK. Statistics; Summarize; Explore;** Select **Salnow to dependent; Statistics; Descriptives; Continue; Plots; Histogram; Continue; OK.** The histogram shows substantial skew and kurtosis, with skew = 1.292 and kurtosis = 2.725.

   b. After a log transformation, the salary distribution is much closer to normal, with skew = .380 and kurtosis = .175.

   c. For Set A, R^2 added = .338, F=57.23, df=2, 224, and p<.0001.
   For Set B, R^2 added = .118, F=24.14, df=2, 222, and p<.0001.

2. Compare assumptions and the null hypotheses for the two tests.

3. a. SPSS would give F=8.08, p<.001

   b. F=1.635, p>.05 compared to tabled value of $F_{20, 29, .05} = 1.94$

   c. Observed R^2 of .530 exceeds Wilkinson's tabled value of .43 (with m=20, k=6, N=50, α=.05), so p<.05.

   d. The SPSS significance test is much too liberal, as it does not consider the size of the pool from which the variables were selected. The conservative test is too conservative, in that it treats the observed R^2 as if it used information from all 20 predictors. The Wilkinson test is the appropriate test, because it takes into account the total number of potential predictors that were considered.

   e. .408. Work backwards from the 'shrinkage' formula, assuming population R = 0, or use the corrected formula from Howell. [Chance $R^2 = p/(N-1) = k/(n-1)$.]

4. See your class notes on estimating sample size. You may use the power table from Cohen and Cohen or G*Power. It could be instructive to use both. If you use G*Power, include a screen shot (Gadwin is a nifty free program for capturing portions your screen).
   $f^2 = .16667$; $n^* = 89$.

Please read these directions and notes carefully.

This can be a joint project with no more than two others. The challenge is to apply your own regression analysis to a data set of your choice. Include a check on assumptions (distributions, residuals), at least one interaction term, at least two steps in a hierarchical analysis, and a summary of your results with a brief interpretation. Attach relevant printouts and syntax files as appendices, but tables, figures, and appropriate statistics should be incorporated into body of the paper. Include a figure that displays the interaction, even if it is not statistically significant.

The analyses that you do will depend on your choice of research questions and your data. As an example, you might choose to predict a dependent variable using three sets of predictors, where one or more of the predictors is categorical (requiring dummy coding), and the third set is an interaction between two variables included in the first two sets. Your interest may be in the contribution of the second set beyond the first set, and in the size and direction of the relationships with individual predictors. Possibly missing data coding could be involved. The handouts in your packet, as well as the review materials, may be helpful.

It is best if you use your own data set, but if you get stuck let us know early and we can help find a data set. Many data sets are available through SPSS and you can find others on the Internet, perhaps through links on http://wise.cgu.edu . Pose your own research questions, and design analyses to address the questions. Use a diagram with circles and arrows to represent your conceptual model. Your Y variable should be reasonably normally distributed, not time series.

You don't need to report all the measures of skew and kurtosis in the text, though you should examine your variables carefully to make sure there are no serious problems for your analysis.

Interpret your results in nontechnical language. Rather than say "Sex was negatively correlated with the Peabody Scale" it is better to say "Women had a higher mean score than men on the Peabody, 74 vs. 67, respectively." Don't use SPSSese for names of variables.

Generally, you need to reformat tables from SPSS to attain APA style. Be sure to use clear labels, with footnotes to elaborate as necessary. Include information on size and direction of effects, along with statistical significance and a description of importance.

When you interpret results, be careful to distinguish the tests for the overall model and tests for R square added. It often is useful to include simple r, R square added, and final beta in the same table, to facilitate your discussion and interpretation. Size and direction of effects should be clear.

When testing an interaction, be sure that you interpret the R square added by the interaction term(s) only after the main effects are in the model, and do not interpret the final beta for the main effects after you have entered the interaction term. Consider centering variables, especially when modeling an interaction with a continuous variable where zero is not meaningful.

Please follow these guidelines for your paper.

The paper should be in APA format, double-spaced, but with tables and graphs incorporated into the text, not at the end. You may be able to use some SPSS output directly, though usually it will be better to edit the tables and figures to make them more appropriate for your paper.

Abstract (< 1 p)
* provide an overview and summary

Introduction (1-3pp)
* explain the research question, why it is interesting, etc.
* provide a graphical representation of your conceptual model (circles and arrows)

Method (1-3pp)
* describe your sample - who, where, why, how many (cite your data source so one can replicate)
* describe your measures, show histograms of continuous variables
* describe assumption checks, any transformations, etc.
* explain how you coded any categorical variable, how you dealt with missing data
* describe the logic of your analysis, including a description and test for an interaction.

[Note: This is more detail than you usually put into a paper for publication, but we'd like you to include all the steps to demonstrate your mastery of the whole process.]

Results (3-6pp)
* present your results, using APA-style tables and graphs in the text. You will need to edit or redo SPSS output.
* include a plot of the interaction, even if it is not statistically significant. You can use Excel, and you probably will need to edit the figure (fix labels, colors, markers, etc.)
* report relevant statistics following APA style (including df or N, etc.). Be sure to include effect size measures.

Discussion (1-2pp)
* interpret results
* discuss practical implications
* discuss any limitations
* any suggestions for future research?

*** Attach relevant printouts that include your syntax (not included in the page count).**

Length: probably about 7 to 12 pages, depending on how many figures, etc. you use.

Let us know if you have any questions or if you'd like to check on any issues along the way.

It is important to follow the guidelines. Projects should include an abstract, a graphical representation of your theoretical model, formal citation of the source of your data, description of measures (including histograms), description of how assumptions were checked, and explanation of coding of a categorical variable and an interaction. Results should include tables and graphs based on the output, but generally not just pasted in from SPSS output, and the description should include information on size and direction of effects, including interactions. The SPSS syntax and printout should be attached.

Main effects MUST be entered before interactions. The test of interaction is the test for R square added by the interaction term(s). After the interaction is in the model, do not interpret the B or beta weights for the main effects. It is usually appropriate to 'center' continuous variables that are used in an interaction.

When you have a categorical variable like ethnicity, it does not make sense to report the mean. Even if you later comment that the mean is uninterpretable, there is no reason to report the mean. If there are five groups, you need four dummy variables to capture the original variable. Also, you need four separate terms, entered on one step, to capture an interaction with that variable. Check frequencies for categorical variables; pool or delete groups with very small *n*.

Watch your sample size carefully in each analysis, to make sure you understand why cases are missing. If you have only a few missing data, you may choose to omit those cases. If you choose to 'plug' with mean substitution, it would be prudent to construct a 'missingness' variable (coded 1 where X is missing, and 0 for cases where X is present). Check how the missingness variable is related to other variables, which may allow you to better understand the reasons for missing data. In a regression analysis, the missingness variable should be entered before or on the same step as the 'plugged' variable. Preferred methods for dealing with missing data are multiple imputation or maximum likelihood estimation.

The rationale for order of entry should not necessarily be based on what you think will be the biggest or most important effects. You may choose to enter 'nuisance' variables first, for statistical control. Consider whether it is theoretically more interesting to talk about the effects of B with A removed, or the effects of A with B removed. If there is a causal flow, it makes most sense to enter variables in order of causal flow. If some variables are established predictors or common sense controls, they probably should be entered first, to allow you to test whether your new variables add anything useful. If some variables are inexpensive to obtain, perhaps they should be used first, to test if more expensive variables contribute additionally to the prediction.

You may not need to transform your variables. Try a sensitivity analysis, whereby you compare results for transformed and untransformed data. If transformation doesn't matter, perhaps you can avoid the complication. You can mention in a footnote that you also analyzed transformed data and your findings were materially the same (if that is true). Residuals should be normally distributed, but the predictor variables don't need to be. Watch out for outliers, as they can carry disproportionate weight.

Avoid SPSSese. Your reader doesn't know what DUM1, EDXSEX, or Q1.42 mean.

It is not correct to say something like "Males were correlated with high assertiveness." Sex is a constant if you limit the description to males. You could say "Males were significantly more assertive than females" and give the mean and SD for each group.

Be careful with B vs. beta. The B weights are regression weights for unstandardized variables. They are hard to compare, because they depend on the standard deviations of the variables. Beta weights are standardized, so they are easier to compare. Beta weights for main effects can be compared to the simple r values (simple bivariate correlations), but only before interaction terms are entered. The overlap of interaction terms and main effects is greatly reduced if the main effects are first 'centered.' B and beta coefficients are much more stable and easier to interpret in the presence of an interaction with centered variables.

When you interpret interactions, it is useful to supply supplementary information. If ethnicity and gender interact in predicting income, then you should report the sex difference in income for the different ethnic groups. If gender and a continuous variable like education interact in predicting income, then you can compute the correlation between education and income for males and females separately, report, and discuss.

Table 1: Hierarchical regression of Income in $1000s on Sex and Ethnicity (*N* = 432)

Step	Predictor	r	R^2 Change	Final B	SE(B)	Beta
1.	Sex (M=0; F=1)	-.139**	.019**	-8.49*	3.15	-.098*
	(*N* female = 203)					
2.	Ethnicity (df = 3, 428)		.117**			
	White (*N* = 177)	.204**		12.33*	5.91	.181**
	Black (*N* = 92)	-.144**		-11.08*	5.88	-.081*
	Hispanic (*N* = 121)	-.042		-6.77	4.92	-.033
	(Asian; *N* = 42)	(.019)				
3.	Sex by Ethnicity (df = 3, 425)		.095**			
	Sex by White			-1.09	3.85	
	Sex by Black			11.43**	4.08	
	Sex by Hispanic			-7.89*	3.11	
	Constant			33.86**	2.99	

*p < .05; **p < .01. Note: Cumulative R^2 = .231, p < .01; adjusted R^2 = .218. Beta values are from the second step, with Sex and Ethnicity in the model, but no interaction. No dummy variable is included in the regression model for the reference group (Asian).

{If you show B weights, include the constant. In the text, describe the size and direction of main effects and the interaction components if statistically significant or theoretically interesting. With continuous predictors, centering is generally recommended. Reporting beta weights from Step 2 is unconventional, but it provides useful information, especially if variables are uncentered.}