EDRM612 Syllabus/Objectives

Spring Semester, 2003

Introduction

1. **Identify the independent and dependent variables in a research problem.**

The independent variable is the variable that is used to explain, cause, or predict the dependent variable. The dependent variable is the effect, or the variable to be predicted. In causal terms, the independent variable causes the dependent variable.

2. **Know how the statistical procedures in this course (ANOVA, ANCOVA, Correlation, and Multiple Regression) are related to other statistical procedures (Chi square, Discriminant Analysis, t-test, MANOVA, and Canonical Analysis).**

The statistical procedures listed in the table below are classified by the number and type of dependent and independent variables analyzed. You will not be tested over the characteristics of procedures not covered in EDRM612.

Analysis type	Independent	Dependent
Chi square	1 Categorical	1 Categorical
Discriminant Analysis	1+ Interval	1 Categorical
t-test	1 Categorical (2 groups)	1 Interval
ANOVA	1+ Categorical	1 Interval
ANCOVA	1+ Categorical 1+ Interval	1 Interval
Correlation	1 Interval	1 Interval
Multiple Regression	1+ Interval	1 Interval
MANOVA	1+ Categorical	2+ Interval
Canonical Analysis	2+ Interval	2+ Interval

3. **Know the type of data required to do the statistical procedures of this course (correlation, multiple regression, ANOVA, ANCOVA).**

In order to do a correlation analysis you must have two variables in which the data consists of matched or paired cases. The two paired variables are usually referred to as X and Y. For correlation analysis either variable can be designated as X or Y. For multiple regression you have one dependent variable (Y) and one or more independent variables (X). Both X and Y are analyzed as if they were interval or better data (there are procedures that can transform other variables). ANOVA analyzes one interval (or better) dependent variable (the same type as multiple regression), but the independent variable is treated as if it were a categorical variable. ANCOVA combines characteristics of ANOVA and multiple regression. One interval (or better) dependent variable is used (the same as ANOVA and multiple regression), but two types of independent variables are used–one or more categorical independent variables (the same as ANOVA) and one or more interval (or better) independent variables (the same as multiple regression).

Simple Analysis of Variance

4. **Know the relationship between the terms factor, level, treatment, group, and variable.**

The independent variables in ANOVA are called factors. The values of the factors are called levels, treatments, or groups. The term treatment is usually reserved for an experimentally assigned group and level is sometimes reserved for variables that are quantitative in nature (even though analyzed as groups).

5. **Know the main assumptions of ANOVA.**

The main assumptions of ANOVA are:

a. interval data on X (the dependent variable)
b. normal distribution on X for the population from which each group was selected
c. equal variance on X for the population from which each group was selected
d. observations are independent of each other (this is almost always satisfied)

6. **Know how to identify violations of the assumptions of ANOVA.**

Assumption a is determined by the type of variable used. The other assumptions depend on the data used. Scatterplots of the data can identify assumptions b and c. The Levene's test of homogeneity of variance is also used to test assumption c.

7. **Know what to do when the assumptions of ANOVA are violated.**

ANOVA is fairly robust for violations of assumptions b and c. If there are severe violations with these, use a nonparametric test. Assumption a must be satisfied.

8. **Know the meaning of the sums of squares in the ANOVA table in terms of deviation scores.**

$SS_{Between}$ = the sum of the squared deviations between the group means and the grand mean.
SS_{Within} = the sum of the squared deviations between the individual scores and the group means.
SS_{Total} = the sum of the squared deviations between the individual scores and the grand mean.

9. **Know the relationship between explained, unexplained, and total variation and sum of squares for between, within, and total.**

Explained variation = Between sum of squares.
Unexplained variation = Within sum of squares.
Total variation = Total sum of squares

10. **Know the relationship between each mean square in an ANOVA table and a variance.**

Mean squares are equal to variances. For example, MS within can be called error variance.

11. **Know the mean squares used to compute an F ratio.**

F = MS Between/ MS Within.

12. **Know how to use error bar charts to evaluate the differences between group means.**

Error bar charts indicate the mean of each group along with a confidence interval around the mean. If there is overlap between the confidence intervals in the bar charts, the differences are unlikely to be significant.

13. **Know the meaning of fixed and random effects.**

Fixed effects are those where the groups studied are the only groups to which the results are to be applied. Random effects are those where the groups are a sample of those of interest. Most research using ANOVA deals with fixed effects. These are the only problems we will deal with in EDRM612.

14. **Know the meaning of small, medium, and large effect sizes in ANOVA and how they are computed.**

Effect sizes in ANOVA refer to the differences between the means. They are frequently interpreted in either of two ways: in standard deviation units (z scores) or an eta squared value. When different analyses are compared or combined (e.g., in meta analysis) each difference of means is converted to a z score. Eta squared summarizes differences between all means being compared in one analysis. Conventional standards for interpreting z score effect sizes are $.2\sigma$ (.2 standard deviation difference between two means) for a small effect, $.5\sigma$ for a medium effect, and $.8\sigma$ for a large effect. Eta squared cutoff guidelines are .01, .06, and .14.

15. **Given a data set with more than two groups of one independent variable, use SPSS to test for the significance of the difference between the means and interpret the results correctly (One-way ANOVA).**

16. Know when tests of multiple comparisons are appropriate and why they are needed.

When more than two means are being compared the initial test between all means is called an omnibus test. Tests of multiple comparisons are helpful to compare pairs or sets of these multiple means. Tests of multiple comparisons are needed to compensate for inflated alpha rates done when many tests are being conducted.

17. Know the meaning of contrast or familywise (experimentwise) alpha rates.

If there were really no differences between groups, using a contrast alpha rate of .05 for each test would result in finding a significant difference in 5% of the contrasts. Using a familywise alpha rate of .05 would result in a significant difference 5% of the time you did an analysis (ignoring the number of comparisons you were making).

18. Know the meaning of post-hoc and a priori tests.

Post-hoc tests are done after finding a significant F in an omnibus test. Tests specified before the omnibus F test is done are called "a priori" tests.

19. Know the issues involved in determining which test of multiple comparisons is best to use.

Different tests vary in their power, how they control for familywise Type I error, their appropriateness with unequal variances or sample sizes among groups, and their sensitivity to the number of tests to run (all or a subset), or whether the comparisons are specified in advance.

20. Given SPSS output from a one-way ANOVA, interpret the results of a multiple comparison test.

Results are reported either as p values (significance) for every pair of means or indicating homogeneous subsets of means which are not significantly different from each other.

Factorial Analysis of Variance

21. Know the meaning of the term "factorial design."

A factorial design is a design that includes two or more factors. Factorial designs are frequently referred to by the number of factors, such as a two-way design, three-way design, etc. They are also referred to by the number of categories in each factor, such as 2x4 or 3x2x5 designs.

22. Know the meaning of row, column, layer, simple, and main effects.

The factors in a two-way design are usually called rows and columns and in a three-way design they are called rows, columns, and layers. The term "effect" is a general term referring to the difference in means between rows, between columns, and between layers in a factorial design. The row effect is the difference between the row means. If the difference between the means is large you would have a large row effect. Similarly you might have a large column or layer effect in a three-way design. Row, column, and layer effects together are also called "Main" effects. Simple effects are effects of each independent variable at only one level of the other independent variables. You might have a row effect (difference between rows) at column one. In contrast, other components of the ANOVA model covered later (interaction and covariate effects) might be of "Secondary" importance.

23. Know the meaning of two-way and three-way interaction.

A two-way interaction is when the row effect is not consistent (not the same) over the columns and the column effect is not consistent over the rows. For example, if the overall difference between two rows is 10 points, interaction would occur if the difference between the rows was not 10 points within each of the columns.

A three-way interaction is when the two-way interaction is not consistent over the third factor. For example, the overall two-way interaction may be that the row effect is three times as large in column one as in column two overall, but it is not consistent for each of the layers of the third variable. The difference in row effect between the two columns may be four times as much at one layer but only two times as much in the other layer.

24. Know the meaning of ordinal and disordinal interaction.

Ordinal interaction occurs in a two-way interaction when the order of the categories from highest to lowest is maintained across all levels of the other category even though the differences between the categories is not consistent.

Disordinal interaction occurs when the inconsistency results in a different ordering of the categories at each level of the other variable. Interaction is expressed graphically as non-parallel lines. The lines cross in disordinal interaction but do not cross in ordinal interaction.

25. Know how to use multiple line charts to determine interaction.
Interaction effects in a multiple line chart are indicated by non-parallel lines.

26. Know the components of a two-way ANOVA table and the meaning of sums of squares for rows, columns, and interaction in terms of cell, marginal, and expected means.

SS_{Row}	weighted sum of the squared differences between	the row means and the grand mean.
SS_{Column}	weighted sum of the squared differences between	the column means and the grand mean.
$SS_{Interaction}$	weighted sum of the squared differences between	the cell means and the expected cell means.

The expected cell means are those that would exist if each dimension (row, column, layer, etc.) was consistent across the other dimensions.

27. Know the advantage of factorial designs over separate one-way designs.
Row and column sums of squares (and mean squares) would be identical in a factorial design and separate one-way designs with the same data (in the most common type of analysis). The error term, however, will usually be smaller in a factorial design, therefore resulting in a larger F. The factorial design also allows you to study the interaction of the separate effects. This is not possible in separate one-way designs.

28. Given a data set, be able to complete an analysis of a two-way ANOVA using SPSS and interpret the results.

29. Given a printout of a two-way ANOVA from SPSS, be able to interpret the main effect and interaction results.
Means are given in two places: a Descriptive Statistics table gives the actual or weighted means. An Estimated Marginal Means table gives unweighted means. The unweighted means are an estimate of what the means would be if the groups had been equal in size. It is important to select the appropriate means for your interpretation The default in SPSS is to use unweighted means (Type III Sum of Squares)..

30. Be able to complete an ANOVA table given summary data (number of subjects, sum of squares, and number of levels for each factor.

31. Know the meaning of and how to interpret ANOVA results using Type I, II, and III Sum of Squares.
Type I SS adjusts each effect for those listed before it in the list of effects. It would be used if there is a hierarchy of cause and effect factors being hypothesized. Type II SS adjusts the main effects for each other (not the interaction) and the interaction effect for both main effects. It is equivalent to the regression approach. Type III SS evaluates differences between the unweighted means. Type III is the default and most common in SPSS. It assumes that differences in sample sizes are a result of random events which is what would normally occur in an experiment where the experimental and control groups would have equal n's at the beginning of the experiment. Type IV will not be covered in this class. It is useful with missing cells (cells with no subjects).

32. Know the meaning of tests of simple main effects and when to conduct them and pairwise comparisons in a factorial design.
Tests of simple main effects are one-way F tests for individual levels of one or more of the independent variables using the overall error term. Pairwise multiple comparisons can be done on either the marginal means or cells means as in a one-way design. Simple main effects are conducted after a significant interaction has been found. In many cases the interaction can be interpreted satisfactorily by inspection of the means and not using simple main effects tests. The SPSS procedures for simple main effects and pairwise comparisons for factorial designs require using complicated lmatrix syntax commands. This is necessary to ensure that the correct error term is used. An alternative procedure would be to do a one-way F test using SPSS but compute the F ratios by hand using the overall error term from the two-way analysis.

33. Know how a repeated measures design differs from one-way and factorial ANOVA designs.

In a repeated measures design, the groups being compared are composed of the same subjects. Because of this, the error term for the difference between the means needs to be computed differently to account for the similarity that is expected when measuring the same persons. In effect, the difference between subjects across all variables (groups) is removed from the error term which results in a more powerful test. Many designs include both a between-subjects factor and a within-subjects (repeated measures) factor.

Analysis of Covariance

34. Know the characteristics (purposes) of Analysis of Covariance.

ANCOVA is a design where the scores on the dependent variable are adjusted on the basis of another variable, called the covariate. The purpose is to remove differences between the groups due to the covariate. The resulting interpretation is made on the adjusted means. Any ANOVA design can be extended to an ANCOVA design. The covariate is the variable that is controlled for and the ANOVA factors are categorical variables. The covariate is usually a quantitative variable that is not grouped when used in the ANCOVA design.

35. Given a data set, complete an analysis of a one-way ANCOVA design using SPSS and interpret the results.

36. Given a one-way ANCOVA output from SPSS, correctly interpret the F values obtained.

37. Know the effect of and value of ANCOVA.

ANCOVA is valuable since it removes variation due to the covariate (an extraneous factor). The effect will usually be to lower the error term (MS_{Within}) which increases the power of the test. The difference between the adjusted means, however, may be greater or smaller than the difference between the observed means. The effect of the covariate may have masked a real difference between the groups or caused the observed difference. Therefore planning to remove variation due to the covariate may increase or decrease the probability of finding a significant difference.

38. Know problems associated with using ANCOVA.

When the covariate is correlated with one or more of the independent variables, it is necessary to determine whether removing the variation predicted by the covariate might also be removing some of the effect of the independent variable that also causes the variation of the covariate. In effect, you want to remove variation in the dependent and independent variables caused by the covariate but not remove variation simply correlated with it or variation in the covariate caused by the independent variable.

39. Know the assumptions of ANCOVA.

In addition to the normal assumptions of ANOVA, ANCOVA has the additional assumption of homogeneity of regression slopes.

40. Be able to recognize violations of assumptions of ANCOVA from a scatterplot matrix.

Scatterplots of Y and the covariate for each of the groups (cells) can help identify violation of the homogeneity of regression assumption.

Simple Correlation/Regression

41. Identify types and degrees of relationships from scatterplots.

A zero correlation will have a scatterplot in which the average value of each Y score remains the same (has the same height) as you move from left to right. A perfect correlation will have all points lying in the same non-horizontal straight line. A positive correlation has the points higher on the right side of the scatterplot and a negative correlation has the points higher on the left side. The use of the SPSS "sunflowers" option is necessary when multiple cases occupy the same position on the scatterplot.

42. Know the assumptions necessary for correct use of correlation and regression and the relevance of each to correct interpretation.

Correlation/regression analysis has many assumptions including:

a. the Y (dependent variable) values are normally distributed at each X (independent variable) value for the population

b. the Y values have equal variances at each X value for the population (homoscedasticity)

c. a linear relationship exists between X (or transformed X) and Y for the population

If assumptions a and b are violated, the error in prediction will not be consistent for each X value. If the relationship is not linear, the correlation coefficient will underestimate the true relationship.

43. Know what to do about violations of these assumptions.

Regression is fairly robust to violations of assumptions a and b. If c is violated, the data should be modified before analysis.

44. Recognize violations of assumptions from scatterplots.

Non-normal distributions frequently occur when there is a limiting upper or lower boundary to the Y value at certain X values, causing the Y distribution at those values to be skewed away from the limiting boundary. In this case regression will consistently over or underpredict at extreme values.

Unequal variances are most commonly noticed as a fan-shaped distribution where the Y values are more closely clustered around the regression line at one end than they are at the other end. In this case there is more error in prediction where there is large variance.

A non-linear relationship exists when the means of Y at each X value are not on a straight line. The most common non-linear distribution are U-shaped or J-shaped distributions.

In examining scatterplots for violations of assumptions, unless you find a non-random variation from the patterns expected you should assume that the population characteristics do not violate the assumptions.

45. Know how to use a lowess smooth (locally weighted regression scatterplot smoothing) to evaluate linear relationships.

A lowess smooth in SPSS approximates, with an irregular curve, the actual relationship in the sample data. If there is a recognizable non-linear curve that makes theoretical sense, the linear assumption can be considered to be violated.

46. Know the meaning of a correlation coefficient in terms of the sign (+ or -) and the numerical value.

A positive correlation signifies that as one variable goes up, the other goes up. A negative correlation signifies that as one variable goes up, the other goes down. A positive or negative 1.00 is perfect correlation. No correlation is shown by a 0.00 correlation. A negative correlation is just as strong as a positive correlation when they have the same absolute value.

47. Know the usefulness of verbal labels such as high and low in interpreting a correlation coefficient.

The terms "high" and "low" as applied to a correlation coefficient are only meaningful in the context of the type of situation and the size of the correlation coefficients to which the given coefficient is being compared. A coefficient of .70 would be a low reliability coefficient for a standardized test (.90 is common) but a high validity coefficient for that test to predict grades (.50 is common).

48. Know the relationship between correlation and causality.

Finding a large (or significant) correlation does not imply a causal relationship between the variables. It only is one bit of evidence that can be used to determine causality. In order to have a cause and effect relationship, there must be a correlation between the X and Y variables (perhaps with other things being controlled). It is a necessary condition of causality, but not sufficient.

49. Know the effect on the correlation coefficient of changing each data value by adding, subtracting, multiplying, or dividing a constant.

Adding, subtracting, multiplying, or dividing a constant to all of the numbers of either or both of the variables does not change the correlation coefficient. The correlation coefficient describes the relationship between the scores in standardized or z-score form.

50. Know the effect on the correlation coefficient of unreliable measures, a truncated range, or a curvilinear relationship.

To the extent that either of the variables are unreliable (have measurement error) the correlation coefficient will be spuriously low (underestimate the true relationship between the variables). Scores with a large component of "randomness" cannot correlate highly with anything,.

To the extent that either of the variables has a restricted range (not the full range of the population of interest), the correlation will be spuriously low (underestimate the true relationship between the variables). This is because the prediction error (lack of perfect correlation) will be a larger proportion of the total variance in the restricted range.

If the relationship between X and Y is not a linear relationship, the correlation coefficient reported using a regression program will underestimate the true relationship. Regression programs report the linear relationship between the variables even if that is not the best or true relationship between the variables.

51. Know the limitations of using the regression equation to estimate the relationship between X and Y.

The value of a regression equation is determined by the size of the related correlation coefficient, not by the size of the regression coeffients ("a" and/or "b"). The "a" and "b" values are influenced by both the relationship and the mean and standard deviations of the variables.

52. Know the meaning of the least squares criterion for the regression line.

The regression line is the straight line that minimizes the squared deviations between the line and each Y value.

53. Know the meaning of "a" (b_0) and "b" (b_1) in a regression equation.

"a" in a regression equation is called the constant (algebraic and regression term) and the Y-intercept (geometric term). It is called the constant because it is the value that is added to the prediction of each score no matter what the X value is. It is called the Y-intercept because this is the Y value where the regression line crosses (or intercepts) the Y axis.

"b" is called the slope (geometric term) and the regression coefficient (algebraic and regression term). It is called the slope because it is the amount of change in the height of the prediction line (vertical axis– Y) that occurs for every change in the horizontal axis (X) of one unit. It is called the regression coefficient because it is the number that is multiplied by the X value when predicting.

54. Be able to plot a regression equation on a scatterplot by hand and using the scatterplot option in SPSS.

The SPSS sunflower option is needed to display the density of points in the plot when there are a large number of subjects resulting in many data points falling at the same place on the plot.

55. Know the relationship between Y and Y'.

Y values are observed values while Y' are values predicted from a regression equation from an observed or hypothesized X value. Y values are actual data points scattered across the plot while Y' values are points on the regression line which normally do not correspond to real data.

56. Given a regression equation, know how to calculate Y'.

57. Know the meaning of a residual.

A residual is Y-Y'. It is that part of the score that remains after prediction. If the residual is zero, the score is predicted with no error. A residual is sometimes called error signifying that it is the extent to which the predicted value is incorrect for each case.

58. For each of the following questions, be able to answer them given a SPSS printout or use SPSS to analyze raw data to answer them:

a. What is the regression equation?

b. **Predict Y, given X**

c. **As X increases by 1 raw score unit, Y goes up by how many units?**

59. Given a correlation coefficient, know the amount of variance in Y which is predicted by X.

r^2 is the percent of variance of one variable (either X or Y) that can be accounted for (predicted, explained) by the other variable.

60. Know the relationship between r and b in a raw score regression equation and between r and ß when the equation is in z score form.

There is no way to predict what b will be given r unless r=.00 in which case b=0. Both the degree of relationship and the standard deviations of X and Y influence the value of b. Given constant X and Y standard deviations, as r increases, b increases. When there is one predictor, r=ß. With more than one predictor r seldom if ever equals β.

61. Know the meaning of regression to the mean.

Regression to the mean occurs whenever prediction takes place with less than perfect prediction. The predicted Y values are closer to the mean of Y (in terms of z scores) than the X values are to the X mean. With zero prediction there is complete regression to the mean and the Y mean is predicted for everyone. With perfect prediction there is no regression to the mean and the predicted Y score is the same distance from the mean of Y (in z scores) as the X value is from the X mean. When $r_{XY} = .00$, for any z_X, $z_{Y'} = 0.00$. When $r_{XY} = 1.00$, $z_{Y'} = z_X$.

62. Know the definition of ß and the relationship between b and ß.

β is the value by which the z_X is multiplied to predict $z_{Y'}$.. It is called the standardized regression coefficient or beta. ß and b can be converted one to the other by multiplying each by a ratio of the standard deviations of X and Y. ß is multiplied by σ_Y/σ_X to find b. b is multiplied by σ_X/σ_Y to find ß.

63. Given a correlation coefficient and a z score for X, predict the z score for Y.

The predicted z score for Y is found by multiplying the z score for X by the correlation coefficient (which equals ß with one predictor). $z_{Y'} = rz_X$

64. Know the advantages/disadvantages of listwise, pairwise, and mean substitution procedures for generating a correlation matrix for regression analysis.

Since the listwise procedure only uses cases with complete data on all variables specified, this method will frequently throw away the data from many subjects that have incomplete data. It is possible to throw away almost all of your subjects even if there is relatively little missing data. All you would need is for each person to be missing one score.

The pairwise procedure uses all of the data but may result in a correlation matrix that cannot be used for further analysis (cannot be inverted) since each of the correlation coefficients is not necessarily based on the same subjects. This is particularly true if much data is missing.

The mean substitution procedure uses all of the existing data and replaces missing data with a mean value. Since data is introduced that do not really exist and the data are the same for all people with missing values, the resulting correlations would be expected to be lower than would really exist if the data were complete, but it is a legitimate compromise if the other two methods cannot be used to satisfaction.

65. For each of the following questions, be able to answer them given a SPSS printout or use SPSS to analyze raw data to answer them:

a. **Is r^2 significant?**

b. **How accurate is the prediction?**

c. **As X goes up by 1 z score unit, Y goes up by how many z score units?**

d. **Is the regression coefficient significant?**

66. Know the relationship between and meaning of sum of squares for total, regression, and residual, in terms of deviation scores.

SS_{Total} = the sum of the squared deviations between each Y score and the Y mean.
$SS_{Regression}$ = the sum of the squared deviations between each predicted Y score (Y') and the Y mean.

$SS_{Residual}$ = the sum of the squared deviations between each Y score and the predicted Y score (Y').

67. Know the relationship between standard error of estimate and:
a. standard deviation of residuals
b. standard deviation of Y
c. proportions under the normal curve

The standard error of estimate can be thought of as the same thing as the standard deviation of the residuals.

The maximum value for the standard error of estimate is the standard deviation of Y. It has this value when r=.00. As r gets larger, the standard error of estimate gets progressively smaller until when r=1.00, the standard error of estimate is 0.

The proportions under the normal curve can be used to interpret the standard error of estimate, Y raw scores and predicted Y scores in the same way that the standard deviation of Y is used with Y raw scores and the Y mean. For example, if the assumptions for regression are met, 68% of the Y raw scores are within one standard error of estimate of the predicted Y score for each X value, and 95% are within two standard errors, etc. Another way of stating this would be to say that 68% of the predicted Y scores are within one standard error of estimate of the raw Y scores.

68. Be able to compute r^2, σ_Y, and the standard error of estimate given an ANOVA table.

r^2 = SS Regression / SS Total.

σ_Y is found by taking the square root of the total variance for Y (SS Total/df Total).

The standard error of estimate is found by taking the square root of the error variance (MS residual).

69. Be able to complete an ANOVA table for Regression (Source, SS, df, MS, and F) given summary data (number of subjects, number of groups, and two sums of squares).

70. Be able to test a regression coefficient (b or ß) for significance.

The p value associated with the regression coefficient is based on the the degrees of freedom for the error term in the ANOVA table (Residual).

71. Know the effect of increasing N on r, F and tests of significance of r^2 and the regression coefficient.

As N gets larger, r does not change in a consistent manner. F will get larger (due to a greater $df_{residual}$ and consequently smaller $MS_{residual}$), t will get larger (smaller standard error), so therefore both F and t are more likely to be significant.

72. Know the relationship between significance and importance (statistical and practical significance) of r^2.

Statistical significance of r^2 depends on two things: the size of r^2 and N. An r^2 of .00001 could be significant if the sample size were large enough. Importance depends on two things: the size of r^2 and the significance of r^2. If the r^2 is not significant, it is not important, no matter how large it is. If it is too small, it is not important, whether or not it is significant. The determination of what is "too small" is a subjective decision made by each researcher. R^2=.10 may be too small to be important in some situations and large enough to be important in other situations. A significant r^2 may or may not be important.

Multiple Regression

73. Know the number of dimensions needed to visualize a regression equation with more than one independent variable.

A regression equation with one predictor can be represented by a one dimensional straight line in two dimensions. An equation with two predictors can be represented by a two dimensional plane in three dimensions. With three predictors the regression equation is a three dimensional object in four dimensional space. The number of dimensions needed to represent the regression equation is one more than the number of predictors. Since most multiple regression models have more than two predictors, visual devices are seldom used to represent regression equations.

74. Know how residuals are visualized with more than one predictor.

With one predictor a residual is a straight line between the Y value of a point and the Y' value of the regression line at the specified X value. With two predictors, a residual is a straight line between the point and the plane in three dimensional space. With more than two predictors geometrical interpretations are difficult or impossible. Residuals are usually treated algebraically rather than geometrically.

75. Know a multiple correlation as a correlation between Y and Y'.

With regression, Y and Y' are usually different, with the difference between them called a residual. With one predictor Y' is really a transformed X score and the correlation between X and Y can be thought of as the same thing as the correlation between Y' and Y.

With more than one predictor the X's combine together to form Y' and the multiple correlation is mathematically the same as the simple correlation between two variables, Y and Y'. In this case Y' is the linear combination of the X values that best correlates with Y. The regression equation describes how the linear combination is done.

76. Recognize how the least squares criterion is applied to multiple regression.

With one variable the regression line minimizes the squared residuals (the distances from each point or score to the regression line). With more than one variable the regression line minimizes the same thing, the squared residuals, with the squared residuals now considered as the distance from the point represented by all of the values in multidimensional space to the spatial object that is described by the regression equation.

77. Recognize the difference between r and R.

The symbol for the multiple correlation coefficient is R. It cannot be negative, even with one predictor. This is because there will never be a negative correlation between Y and Y'. If there is a negative correlation between X and Y, the regression equation will take this into account with a negative regression coefficient. With one predictor the absolute values of R and r will be the same. As additional variables are added to regression models, the relationship between Y and Y' will either stay the same (no relationship between the new variables and Y) or increase (a positive or negative relationship between the new variables and Y), so R can only get larger–it will never become negative even if most of the X's have a negative relationship with Y.

78. Recognize how b's and ß's differ between one and two or more predictor models for the same predictors.

A b or ß with more than one predictor is sometimes called a partial regression coefficient to distinguish it from the b or ß for a one-predictor model. In a one-predictor model b or ß can be used to describe the relationship of a variable all by itself–how Y would be expected to change if X changed. With two or more predictors, the b's and ß's describe how a change in Y would be expected to occur if the X related to the b or ß changed, but assuming that the other X's in the equation remained constant. In effect it is interpreted as saying that if the other predictors were controlled (not allowed to vary), a change in X would have this given predicted change in Y. In a one-predictor model there is no assumption about other variables not changing.

The b and ß values usually change as more variables are included in a regression model. In general b and ß get smaller if the variables added to the model have a correlation with the variables already included in the model. The higher the correlation between the variables, the more change you would expect.

The b and ß values are only interpretable if you specify exactly what variables are included in the total equation. With different combinations of variables, you will get different b's and ß's for the same set of data.

Extremely unusual coefficients can occur when the predictors are highly correlated. This is explained later under the topic of multicollinearity.

79. Recognize how adding variables to a regression model changes R^2.

As variables are added to a regression model, R can only remain the same or increase. Adding variables cannot reduce the predictive value of the equation. This is another reason why negative R values are not possible. The range of values possible for R is .00 to 1.00. R will remain constant if the additional variables have no relationship to the residuals based on the variables previously in the equation. Later in the course we will cover the relationship between the residuals and additional variables in more detail as we study partial and semipartial correlations and incremental R^2.

80. Recognize the effect on the standard error of estimate of additional variables.

In most cases, the standard error of estimate (the standard deviation of the residuals) will get smaller as additional variables are added to a regression equation. It may get larger if the additional variable does not increase the R^2 value more than a chance amount.

81. Recognize the factors influencing the significance of regression coefficients.

The two main factors influencing the significance of a regression coefficient are the correlation of the variable with Y and the correlation of the variable with the variables already in the equation. If the new variable is highly correlated with any of the other variables in the equation either with the variable alone or with a combination of variables, the variable is not likely to make a contribution to the model and thus will not be significant.

82. Recognize that R^2 values are difficult to predict from simple correlation coefficients.

The value of many variables together in predicting Y depends to a great extent to how they are correlated together. If two predictors are highly correlated together this can be seen in the correlation matrix. If combinations of variables are correlated together, this is extremely difficult or impossible to identify by inspection. What you must do is actually run tests to find these relationships. For example you might find that three variables have relatively low correlations with each other but two of them combined might have a high correlation with the third one. Tolerance as explained in the next objective will do this task for us.

83. Recognize the meaning of "tolerance" in regression analysis and how it is used to assist interpretation of regression coefficients.

Tolerance values are measures of the extent to which the independent variables in a regression model are correlated with the other variables in the model. The tolerance values reported are equal to one minus the R^2 value of the variable in question with all of the other predictors in the regression model.

A high tolerance value means little intercorrelation of that variable with the other predictors. Its regression coefficient can be expected to be relatively stable as the other variables are added or removed from the model. High tolerance is good.

A low tolerance value means that variable is highly correlated with one of the other variables or some combination of them. Its regression coefficient can be expected to be very unstable as other variables are added or removed from the model and is not likely to be a significant predictor in the model since it explains much of the same variance as the other predictors. Even if it was a good predictor by itself, it would not be a significant predictor in the presence of the other variables with which it is intercorrelated. Low tolerance is bad.

84. Recognize the meaning of the components of a covariance and correlation matrix.

An intermediate step in regression analysis is a covariance matrix or a correlation matrix. A correlation matrix is composed of simple correlation coefficients. A covariance matrix includes all of the information of a correlation matrix but also includes variability information and is composed of variances and covariances. Correlations and betas can be generated from a correlation matrix. Computation of raw score regression coefficients (b) and the standard error of estimate require the standard deviation of the variables which is found in the covariance matrix. Correlation matrices found in journal articles can be used to re-analyze the data but not all statistics are possible.

85. Recognize the meaning of the Sig. values reported on computer printouts.

Sig. in computer printouts stand for “significance.” Values reported under the headings of “Sig.” are “p” values or probability values. A “Sig.” value of .032 means that in the condition that there was no true relationship in the population, the relationship found in the data would have a 3.2% probability of occurring. If the probability reported is below .05, it would be significant at the .05 level.

86. Recognize what information to provide when asked to test a regression coefficient or regression model for significance.

When a regression model is tested for significance, the R^2 is tested. The F from the ANOVA table is used to determine the significance. The following three things should be reported:

a. The R^2 value

b. The associated p value

c. A declaration of whether the R^2 is significant or not.

When a regression coefficient is tested for significance, the t is used to determine the significance. The following three things should be reported:

a. The t value
b. The p value reported on the printout
c. A declaration of whether the regression coefficient (or predictor in the equation) is significant or not.

87. Know the difference between explanation and prediction using multiple regression.

Explanation deals with cause and effect relationships. These interpretations are to be avoided with multiple regression unless multiple regression is used in conjunction with causal modeling which uses techniques not covered in this class, or includes other data used with making causal inferences. Explanation has as its ultimate goal theory building.

Most common uses of multiple regression deal with prediction. The purpose of prediction is to predict a score on the dependent variable with as little error as possible, using the variables that can be expected to be stable in predicting in many different situations.

88. Know how shrinkage effects multiple regression results.

When a prediction equation is used with new data (which is its intended use), the percent of variation which will be explained will usually not be as high as that reported by the R^2 with the original data. The reduction in R^2 which would result from correlating the new Y values with the predicted Y values is called shrinkage.

Shrinkage occurs because the R^2 with the original sample takes the error which exists in the sample and predicts it as if it were true variance. When a new sample is used, that error is not repeated but the regression equation tries to predict it and a lower R^2 results.

89. Know the factors affecting shrinkage.

Shrinkage depends largely on three factors: the sample size, the number of variables in the equation, and whether all available variables are used. Various sources recommend that you should have 10, 30, 100, 200, or 400 subjects per independent variable to achieve a stable R^2. They all agree that you need many subjects (observations) per variable to achieve any kind of stability in the R^2. When selecting variables from a larger pool of variables (such as in stepwise regression), more shrinkage is expected and a larger N should be used.

90. Know the meaning of "Adjusted R-Square" reported in computer printouts.

"Adjusted R-Square" is an attempt to estimate shrinkage by the use of a formula. It will always be lower than the reported "Multiple R-Square". The formula adjusts the R^2 value downward based on N and k (sample size and number of predictors). It does not take into consideration the capitalizing on chance that occurs when selecting a model from a large set of variables (stepwise regression).

91. Know the meaning and proper use of the term multicollinearity.

Multicollinearity means that the independent variables are highly correlated or not orthogonal. This intercorrelation may be between any two variables or sets of variables. The term is usually only used in situations when a very high intercorrelation exists between the independent variables in the data set. Data exhibiting multicollinearity is sometimes called collinear data.

92. Know the reasons for collinear data.

Collinear data may be the result of measuring variables that are naturally related, using poor measurement of underlying factors, or sample deficiencies.

93. Know how multicollinearity affects statistical inference and prediction.

When multicollinearity exists, regression coefficients become unstable in sign and magnitude. The standard errors of the coefficients become large, causing low t values.

94. Know how to detect multicollinearity.

Multicollinearity may be detected by:

a. high zero-order coefficients between X and the other X's
b. high R^2 between an independent variable and all other independent variables (low tolerance)

c. implausible signs of b and ß
d. unstable coefficients as variables are added or deleted to the model
e. unstable coefficients if the data is re-analyzed with more or less good data (not including removing outliers)
f. high overall F (or R^2) but low t values for all variables in the model
g. a low t value for an expected important variable

95. Know what to do when data is collinear.

When you have collinear data you may do one of the following:

a. choose better variables that are less correlated with each other
b. remove variables from the model that are highly correlated with each other
c. use alternative procedures other than OLS–Ridge regression uses biased coefficients that result in a slightly lower R^2 but give stable and more meaningful coefficients (a debatable technique)–Principle components regression conducts a regression analysis using underlying factors that are uncorrelated as the independent variables
d. use the equation you have but do not try to interpret the regression coefficients (do not give importance to which variable is the best in the model) or give importance to which variables happen to be in the model or not in the model

Hierarchical Regression

96. Recognize how statistical control is achieved through partial and semipartial correlation.

In experiments, control is exerted through physical control and random sampling. In ex post facto and correlation studies, statistical control is the only type of control possible. One way to achieve this control is through the use of partial and/or semipartial correlations.

Partial and semipartial correlations are possible when you have at least three variables and you wish to study the relationship between two of the variables, controlled for one or more other variables. You have one dependent variable, one independent variable, and one or more controlled variables. The influence of the controlled variables is removed from one or both of the other variables.

97. Recognize the meaning of partial and semipartial correlation coefficients in terms of the variables controlled for and residuals being correlated.

In a partial correlation, the influence of the controlled variable is removed from both the independent and dependent variables. In a semipartial correlation, the influence of the controlled variable is removed from only the independent variable. When the influence of more than one controlled variable is removed, the correlations are called higher-order partial and semipartial correlations. When no variables are controlled the correlations are called zero-order correlations (the normal correlation coefficient).

A partial correlation can be thought of as a correlation between two sets of residuals. The set of residuals remaining after the influence of the controlled variables has been removed from the independent variable is correlated with the set of residuals remaining after the influence of the controlled variables has been removed from the dependent variable.

A semipartial correlation can be thought of as a correlation between one set of residuals and one set of raw scores. The set of residuals remaining after the influence of the controlled variables has been removed from the independent variable is correlated with the set of raw scores of the dependent variable.

98. Recognize the correct notation and terminology for partial and semipartial correlations with one or more variables controlled.

$r_{Y1.2}$ is the partial correlation of Y with 1, controlled for 2 or the partial correlation of Y with 1 with the effect of 2 removed from both Y and 1.

$r_{Y(1.2)}$ is the semipartial correlation of Y with 1, controlled for 2 or the semipartial correlation of Y with 1, with the effect of 2 removed from 1 only. The parentheses indicate that the effect of 2 is only removed from 1.

99. Recognize the meaning of squared partial and semipartial correlation coefficients.

The squared partial correlation coefficient is equal to the percent of the variance of the dependent variable remaining after the influence of the controlled variables has been removed or accounted for, that is explained by the independent variable.

The squared semipartial correlation coefficient is equal to the percent of the original total variance of the dependent variable explained by the independent variable, that has not already been accounted for by the controlled variables.

100. Recognize the value of squared partial and semipartial correlation coefficients.

The squared semipartial coefficient indicates how much additional information a variable will add to a regression model (how much higher R^2 will be).

The squared partial coefficient indicates how much of the remaining unexplained variance can be accounted for by the new variable.

In most situations, the squared semipartial coefficient is used. This is usually called the incremental R^2, referring to the R^2 that is an increment (or added on) to the R^2 without the variable.

The squared semipartial is more important in evaluating the importance of the variance accounted for. The squared partial coefficient is more important in evaluating the significance (p value) of the variable. For example, if the R^2 without a variable is .9999, the most one additional variable can account for is .01%. If an additional variable accounted for this .01%, the squared semipartial coefficient would be .0001 which would seem to be too small to be important. However, this .01% accounts for all (100%) of the remaining variance. The partial correlation would be 1.00 indicating a perfect prediction of the remaining variance which would rarely occur by chance, therefore being highly significant (low p value).

The partial coefficient of 1.00 would indicate good (significant) prediction but the semipartial coefficient squared (.0001) would indicate small (not very important) predictive value.

In most cases the semipartial coefficient is the most relevant. For example when you use test scores to predict gpa, the semipartial coefficient of one test score controlled for other tests is probably controlling for the common elements between the tests (including intelligence). The partial coefficient between a test and gpa, controlled for the other tests would be the coefficient you would find if the factor of intelligence was removed from the test, but also from gpa. Obviously you could never (nor would you want to) keep intelligence from affecting gpa, so the partial correlation would describe something that could never exist. Another reason why partial correlations are usually not reported is that the tests of significance (p values) for the semipartial and partial correlations are identical.

101. Recognize the relationship between the squared multiple correlation, squared zero-order correlations, and squared semipartial correlations.

If there is a high correlation between the independent variables, the semipartial coefficients of each variable with the dependent variable will be small compared to the comparable zero-order coefficients. If there is no correlation between the independent variables, the zero-order and semipartial coefficients are the same.

A multiple R^2 is the sum of the squared zero-order correlations of each X with Y only when the independent variables are not correlated with each other. In most cases the R^2 is the sum of one squared zero-order correlation and many squared semipartial correlations such as:

$$R^2_{Y.123} = r_{Y1}^2 + r_{Y(2.1)}^2 + r_{Y(3.12)}^2$$

102. Recognize the effect of intercorrelations between the independent variables on the contribution of additional variables in the multiple regression equation.

The greater the amount of intercorrelation between the independent variables, the smaller the semipartial coefficients will be and the smaller the contribution of succeeding variables would be when added to a prediction equation.

103. Know the effect of order of entry in evaluating the importance of variables in predicting Y.

A variable may be very important as the only predictor of Y (a high zero-order coefficient), but not important as the second predictor if the predictor already in the model predicts much of the same variance as the new variable. In this case the new variable would have a low semipartial coefficient with the 1st predictor controlled for. To evaluate the importance of a predictor you should report both the value of the variable alone and in one or more combinations of the variable with other relevant variables.

104. Know how and why regression coefficients will change as additional variables are added to the equation.

To the extent that the semipartial coefficients for a variable changes as more and more variables are controlled for, the regression coefficients will also change. The greater the degree of intercorrelation, the greater the amount of change in the regression coefficients. In most cases the regression coefficients of variables already in the model get smaller as additional variables are added as the common variance they share is divided between the variables.

105. Know the meaning and interpretation of suppression.

Suppression is indicated in a model when either of two things occur: 1) the sign of a variable's zero-order correlation is different from the sign of its regression coefficients and its partial and semipartial correlations, and 2) a variable's standardized regression coefficient (beta) is much larger than its zero-order correlation. Under the condition of suppression a variable functions in the model primarily to suppress some of the variance of one or more other variables in the model to allow them to be better predictors. A typical situation is to find a pair of variables exhibiting suppression: one with a larger than expected beta (larger than the zero-order correlation) and the other variable having a large beta with an unexpected sign (opposite that of the zero-order correlation).

106. Know the meaning of incremental R^2

An incremental R^2 is found by determining the amount of variance a variable or set of variables accounts for in addition to another variable or set of variables. It is found by subtracting the R^2 of the original (smaller) set of variables from the R^2 of the final (larger) set.

107. Know the meaning of full model, restricted (reduced) model, hierarchical, and simultaneous regression.

The variables included in the DEFINITIONS of full and restricted models are the independent variables in the models. The dependent variable remains the same in all models, so is not usually specified. The variables which are in the initial small model are considered to be the restricted or reduced model. These are the variables that are controlled for. The variance of these variables is removed before testing for the increment due to the added variables. The variables to be added are included in the full model in addition to the variables in the restricted model. The full model includes all independent variables in the final model.

Hierarchical regression is when one or more variables are added to other variables. The hierarchy is that some variables are entered first into the model prior to other models. The order is frequently due to the earlier variables being more important, more theoretically based, cheaper, or a desire to control for these variables.

The standard way of using regression is to include all of the variables to be considered in the model at the same time. This is called simultaneous regression.

108. Know the relationship between an incremental R^2 and a semipartial correlation squared.

When only one variable is added to a restricted model, the incremental R^2 is equal to a squared semipartial correlation coefficient. When more than one variable is added, the comparable coefficient is sometimes called a multiple semipartial coefficient.

109. Know appropriate full and restricted models when given situations testing for the significance of:

a. a zero-order correlation

The full model contains one X and the restricted model contains no Xs.

b. a multiple correlation

The full model contains all of the Xs and the restricted model contains no Xs.

c. adding one or more predictors to one or more initial predictors (incremental R^2).

The restricted model contains the initial predictors and the full model contains the initial and the added predictors.

110. Know the usefulness of the t ratio in testing for the significance of incremental R^2.

The t values reported for a variable in the computer printout are testing the incremental R^2 for that variable when added to all of the other variables used in the model being reported in that equation.

111. For each of the following questions, be able to answer them given SPSS results or use SPSS to analyze raw data to answer them:

a. Is the difference between a full and restricted model significant and/or important?

b. **Is the addition of a given predictor or set of predictors to an existing predictor or set of predictors significant and/or important?**
c. **Is the removal of a given predictor from a model significant and/or important? Is the addition of a given predictor to the other variables in a model significant and/or important? (These tests are the same)**

112. Know the major problem involved when many good predictors are available to use in a regression equation.

When many good predictors are available to use in a regression equation, it is seldom advisable to use more than a few. After the first four or five are included in a model, addition of more variables usually does not increase the R^2 by a significant or important amount. The sum effect of adding variables after the first few usually results in instability in the coefficients and shrinkage with cross validation. In other words the model with more predictors is not as good as a smaller model. A major problem in regression, then, is how to choose how many predictors to use and which ones to include.

113. Know the factors to include in determining which variables to include in a regression equation.

The final selection of variables will usually depend on a mix of 1) the cost of measuring and including the variable, 2) the theoretical importance of the variable (you expect the variable to be more stable), 3) the significance of adding the variable, 4) the importance of the incremental R^2 when the variable is added, and 5) the value of other statistical criteria that may be used (such as C_p).

Stepwise Regression

114. Know the procedures, advantages and disadvantages for the all possible, forward, backward, and stepwise procedures for selecting variables for a regression equation.

The all possible method is by far the most time consuming method of selecting variables and is rarely used. This method takes every possible combination of variables and picks the one with the largest R^2 for each number of predictors. The models with the largest R^2 for each number can be compared for cost, theory, significance, and importance to choose the best model. Unless there is a small number of predictors, this is not a practical method. BMDP and SAS use a mathematical algorithm that approximates this method very efficiently (best subsets regression). If forward and backward stepwise methods are both used together the resulting model will usually be the same as that found by using the best subsets model.

The forward method begins with no variables in the model and adds variables one at a time until a nonsignificant addition occurs, in each case adding the variable that would give the highest incremental R^2 (the highest partial or semipartial correlation with the dependent variable controlled for the variables already in the model). It keeps variables in the model that may not be necessary as other variables are added.

The backward method begins with all variables in the model and deletes variables one at a time until a significant deletion occurs. The variable deleted is the one which would result in the smallest incremental R^2 (the smallest partial or semipartial correlation with the dependent variable controlled for the other variables in the model). If the number of predictors is very large, it takes many steps and much computational time to get down to a good model. Since all variables begin in the model, variables that only work together are considered. The backward method tends to give models that are slightly larger than the stepwise method (for the same R^2). For this reason it is best to use it in conjunction with the stepwise method.

The stepwise method is a variation of the forward method but includes a feature of the backward method. It begins with no variables in the model and adds variables one at a time like the forward method but then after the variable has been added, an attempt is made to delete the variables that were previously in the model. If one or more of them were significant predictors in the previous model but are no longer significant, the one with the smallest incremental R^2 will be removed. This would occur if the variable added in combination with the other predictors makes a variable no longer needed. The variable would be needed without the last variable added. Models including variables that only predict in combination (when taken together) will not be found using this method. For this reason it is not a good method to use by itself. It should always be used in combination with the backward method.

115. Be able to answer the following questions based on a SPSS printout.

a. **Is the addition of a predictor at a given step significant?**
b. **Would a significant loss occur with the removal of a predictor in a succeeding step?**

c. **Is the complete regression model at a given step significant?**
d. **What is the regression equation at a given step?**
e. **Using the regression equation, what is a specific predicted value?**
f. **How accurate is the regression equation at a given step?**
g. **What is the meaning of the numbers labeled as: correl, part cor, partial correlation, PIN, POUT, and tolerance.**
h. **What information is provided by the change in regression coefficients from step to step?**

116. Know how to add or remove sets of variables using SPSS.

Successive /STEPWISE (or BACKWARD or FORWARD) subcommands can be used for subsets of variables. All variables in each subset that meet the entering/removing criteria are entered or removed before moving to the next subset. At any step variables can be forced in using the ENTER subcommand or forced out using the REMOVE subcommand.

117. Know how to add variables in a hierarchical sequence.

Variables that are prioritized in an a priori sequence can be added using the subcommands in the order the variables are to be entered.

118. Be able to evaluate regression models using SPSS by entering variables:

a. **simultaneously**
b. **hierarchical—adding variables or sets of variables**
c. **stepwise (forward)**
d. **backward (stepwise)**

119. Know the dangers in using stepwise regression.

Stepwise regression is a powerful technique that lures the researcher into letting the computer do the selection. If many variables are included in the initial set of variables, there is likely to be a lot of shrinkage and variables that make no theoretical sense may be included in the model. If there is high intercorrelation between the predictors, you may not get the best variables in the model.

120. Know the factors to consider in selecting appropriate PIN and POUT values for model selection.

If PIN values in the forward or stepwise solutions are too small (.01) the automatic cutoff may occur before important variables would be added that might meet the criteria, but only in the presence of other variables. This makes it more likely that variables acting together as good predictors will be missed.

If PIN values in the forward or stepwise solutions are very high the model may get too large before the automatic cutoff is reached. This is not usually a problem as subjective judgment can be used to select a model other than the last one computed.

Therefore a large PIN (.05-.20) is preferred for forward or stepwise methods.

The opposite is true for backward methods. If the POUT is too large, not enough variables will be removed. A small POUT (.001-.01) is preferred.

121. Know the preciseness of the F values reported with variable selection programs.

Whenever multiple models are being considered and variables are selected according to a statistical criterion, the probability levels for the t's and F's are not accurate. They should only used as approximations. In general the probabilities of chance occurrence are higher than the p values associated with the t or F would indicate.

122. Know the criteria to use in selecting a "best" model using SPSS.

The "best" model should be selected using the following criteria:

a. high adjusted R^2 for the model
b. high contribution to R^2 (incremental R^2) for all predictors in the model
c. all theoretically important variables in the model
d. high t values for all predictors in the model
e. variables included in the model also appear in other good models

f. coefficients are stable in other models
g. assumptions of regression are met
h. the model is consistent with models found using other methods (Stepwise/Backward)

Dummy Variables

123. Know how to construct and use dummy variables in regression to use categorical variables as predictors.

A dummy variable consists of 1's for subjects in one group and 0's for subjects in all of the other groups. To account for the variability of a categorical variable, you need to set up one less variable than the number of categories in the variable.

The dummy variables constructed must be used in combination to test for the significance of the categorical variable. The test of significance for the categorical variable is the test of the incremental R^2 for the dummy variables together in addition to the other predictors. The coefficients and the test of significance for the individual dummy variables are usually not of interest.

124. Be able to set up dummy variables in regression and test the significance of categorical variables.

Relationship between ANOVA and Regression

125. Know the relationship between ANOVA and multiple regression in the following areas:

a. Type of data used

Both ANOVA and MR use quantitative dependent variables. Both ANOVA and MR can use categorical independent variables. Only MR uses quantitative independent variables. Quantitative independent variables must be converted to categorical (nominal) variables to be used by ANOVA.

b. Type of design/philosophy of research

ANOVA is more closely associated with experimental designs in which variables are controlled by the experimenter. MR is more closely associated with ex post facto designs with naturally occurring data where control is done statistically.

c. number of independent variables/ease of implementation

ANOVA seldom uses more than 5 independent variables, with 2 or 3 most common. It becomes very difficult to interpret with more than three independent variables. MR frequently uses many more than that. MR is only slightly more complex as you add independent variables.

d. explained and unexplained variance

ANOVA deals with between groups and within groups variance while MR deals with regression and residual variance. They are equivalent terms for explained and unexplained variance.

e. null hypotheses tested/focus of study

ANOVA studies differences between group means (null hypothesis of equal means) while MR studies relationships between variables (null hypothesis of zero correlation coefficient).

f. equivalence of results

If the same data is used with ANOVA and MR, the same ANOVA table will be generated and the tests of significance (F and p values) will be identical.

g. type of results communicated

ANOVA results include a test of significance (F) and importance (means) while MR results include an F and R^2.

h. popularity

ANOVA and MR each account for about 20-40% of the statistical techniques used in research in the behavioral sciences and education. No other statistical technique comes close to this.