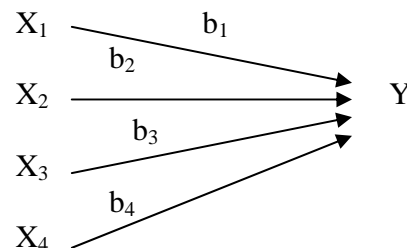


Neuendorf
Multiple Regression

The Model



Assumptions:

1. Multivariate normal distributions
 - a- For individual IVs, or pairs of IVs, look at scatterplots
2. Linearity (i.e., linear relationships)
 - a- For individual IVs, check scatterplots and/or theory
 - b- For entire prediction/equation (i.e., multiple IVs), check residual plot
 - c- See Hair Ch. 2 and additional COM 531 handout for data transform ideas
3. No extreme multicollinearity (intercorrelations among IVs)
 - a- Check a correlation matrix among IVs (provided under the Descriptives option in the Linear Regression procedure in SPSS). . . values above about .80 are problematic
 - b- Check tolerances (unique IV variance proportions--e.g., a TOL of .80 indicates that 20% of that variable's variance is shared by other IVs) and VIFs (variance inflation factors, $1/\text{TOL}$). . . must request these from Linear Regression. . . you want high tolerances (p. 230 of Hair indicates that a .10 or higher is OK) and low VIFs
 - c- Inspect condition index and regression coefficient variance-decomposition matrix. . . this gives the *multivariate* picture with regard to multicollinearity (TOL and VIF both assess multicollinearity one IV at a time). . . the process of using condition indexes to assess multicollinearity is well-described in the 5th edition of Hair, which I will copy and make available to the class (it's been dropped in the 6th edition).

- d- Problem? High multicollinearity leads to unstable partial coefficients. Solution? Put variables in scale(s), drop some IVs, or include all as a block and ignore partials.
4. Homoscedasticity of residuals
- a- Visual inspection--look at residuals plot (To generate this type of residuals plot, you must ask for Plot within Linear Regression in SPSS, then specify *ZRESID as Y and *ZPRED as X; this will give you a graph like that in Figure 4.10 of Hair.)
 - b- Statistical test--the SPSS procedure EXPLORE has the Levene test for homogeneity of variance, but this would require the somewhat involved process of saving residuals and predicted values of Y (*ZRESID and *ZPRED), and then conducting analyses on them
 - c- Data transformations due to heteroscedasticity of residuals? See Hair p. 207 and p. 253 for suggestions
5. Residuals (errors of prediction) should be random (independent) and normal
- a- Visual inspection--see residual plots from SPSS procedure Linear Regression (click on "Histogram" and "Normal Probability Plot"; see Hair figures 4.5 and 4.12 to compare)
 - b- Statistical tests--the SPSS procedure EXPLORE can give the normal probability plot for residuals, and provide Kolmogorov-Smirnov and Shapiro-Wilk statistics that test for deviations from normality; again, this requires saving your residuals and then running *those* through EXPLORE

Decisions to make:

1. Entry of Variables
- a- Forced Simultaneous (SPSS calls it "enter")
 - b- Forced Hierarchical (SPSS refers to "blocks" using "enter")
 - c- Stepwise Forward ("forward")
 - d- Backward Elimination ("backward")
 - e- Combination of Stepwise Forward and Backward Elimination ("stepwise"--common)
 - f- Combination of Hierarchical and Stepwise/Backward

2. Dummy or Effects Coding? (Also, see separate Neuendorf handout on Dummy/Effects coding)
 - a- Need $c-1$ dummies, where c =# of values on the categorical IV. . . if you use c rather than $c-1$, you'll have perfect multicollinearity
 - b- For Dummy, each partial regression coefficient (b or β) and its corresponding F test assess the difference between the "1" group and the referent/comparison (all "0") group
 - c- For Effect, each partial regression coefficient (b or β) and its corresponding F test assess the difference between the "1" group and the average of all other groups

3. Interactions?
 - a- See Hair p. 423 for examples of the notion of interactions. . . Linear Regression doesn't discriminate between ordinal and disordinal interactions
 - b- To include an interaction, use the multiplicative term (e.g., $X_4 = X_{Z1} * X_{Z2}$, where X_4 stands for the interaction). Generally, we wish to standardize the two IVs; note that they are shown in "Z" form.
 - c- Can test for sig. contribution by F of $R^2_{Y.1234} - R^2_{Y.123}$, where $X_4 = X_{Z1} * X_{Z2}$. . . that is, X_4 becomes just like any other contributor
 - d- A sig. interaction does not tell the whole story, though. . . need to take "representative values" of IVs and see DV values to find the pattern

4. Repeated Measures Design?
 - a- Control for subject ID (dummy coded--you need $n-1$ dummy variables!)

Statistics:

Other than the various statistical tests of assumptions described in the first section of this handout, the important statistics are few. . . a "parsimonious" view:

1. Multiple squared correlations (R^2 's)--indicate the proportion of the variance of Y that is explained by a set of IVs. This may be incremental (as in a single step of a stepwise model or a hierarchical model) or total (the variance explained by the final, full regression model). An incremental R^2 is referred to by SPSS as " R^2 change". (Careful--the output gives you the "total" R^2 after each step in a model; this is neither the incremental R^2 nor the final, total R^2 --it's the total up to that point.) There is also an "adjusted R^2 " reported, which reduces the "inflation" that occurs with a large number of

IVs (see separate handout on adjusted R^2). Each R^2 is tested with an F test.

2. Partial regression coefficients--unstandardized (b 's) and standardized (β 's) coefficients indicate the unique contribution of each IV. Each is a partial slope--the change in Y for a unit change in X, controlling for the other X's in the equation. The significance of each partial regression coefficient is tested with an F, which is the same for unstandardized and standardized.
3. Standard errors and confidence intervals for the prediction and for the partial regression coefficients—the SEE (standard error of the estimate) is the standard deviation of the residuals, from which one might calculate a confidence interval for any given case's predicted value of Y. The SE of each β allows one to establish a CI (confidence interval) around the coefficient.

5/09