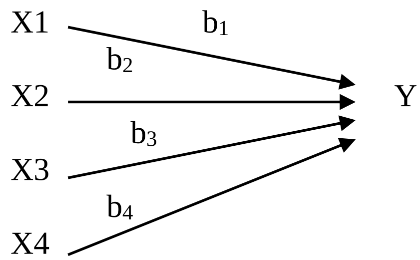
Neuendorf

Multiple Regression

<u>The Model</u>



<u>Assumptions</u>:

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 <u>theory</u>

   b- For entire prediction/equation (i.e., multiple IVs), check residual plot

   c- See Hair et al. Ch. 2 and additional COM 631 handouts for data transform ideas

3. No extreme multicollinearity (intercorrelations among IVs)

   a- Check a correlation matrix among IVs (provided under the Statistics → 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 80% is unique to that variable/not shared) and VIFs (variance inflation factors, 1/TOL). . . must request these from Linear Regression. . . you want higher tolerances (p. 204 of Hair et al. indicates that a .10 or higher is generally OK) and lower 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 et al. (but dropped from later editions), which is available on our class web site.

d- Problem? High multicollinearity leads to unstable partial coefficients. Solutions? (a) Put variables in scale(s), (b) drop some IVs that are redundant, or (c) include all as a block and ignore the partials (for that block, look only at R^2 change and not the betas).

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 et al.)

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 both the residuals and the predicted values of Y (*ZRESID and *ZPRED), and then conducting analyses on them

c- Data transformations due to heteroscedasticity of residuals? See Hair et al. chapter 2 and the COM 631 handout on transformations for suggestions

5. Residuals (errors of prediction) should be random (independent) and normal

a- Visual inspection--see residual plots from SPSS procedure Linear Regression (Under "Plots," click on "Histogram" and "Normal Probability Plot"; see Hair et al. Figure 4.5 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

<u>Decisions to make:</u>

1. Entry of Variables

   a- Forced Simultaneous (SPSS calls it “enter”)

   b- Forced Hierarchical (SPSS refers to “blocks” using “enter”)

   c- Stepwise Forward (SPSS: “forward”)

   d- Backward Elimination (SPSS: “backward”)

   e- Mixed Stepwise Forward and Backward Elimination (SPSS: “stepwise”–the most common)

   f- Combinations of the above, block by block

2. Dummy or Effect(s) Coding? (Also, see separate Neuendorf handout on Dummy/Effects coding)

   a- Need c-1 dummies, where c = # of values/categories on a nominal IV. . . if you use c rather then c-1, you'll have perfect multicollinearity in that groups of variables and SPSS will not be happy

   b- If there is a single dummy for a nominal IV (e.g., FEMALE, where 0=male and 1=female), then the partial regression coefficient (b or β) and its corresponding F test assess the difference between the "1" group and the “0” group (e.g., the difference between female and male)

   c- For Dummy coding where there are two or more dummies representing the nominal variable, 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(s) coding where there are two or more effects codes representing the nominal variable, 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 et al. p. 377 for examples of the notion of interactions. . . Linear Regression doesn't discriminate between ordinal and disordinal interactions

   b- To include an interaction, use a multiplicative term (e.g., $X4=X1_c*X2_c$, where X4 here stands for the interaction term for X1 and X2). Generally, we wish to "center" the two independent variables when their interaction is calculated; note that they bear the subscript "c". Centering involves simply subtracting a variable's mean from all the sample's scores on that variable (e.g., COMPUTE $X1_c = X1 - 4.32$).

c- Can test for sig. contribution by F of the interaction's R^2 change: $R^2_{Y.1234} - R^2_{Y.123}$, where $X4=X1_c*X2_c$. . . that is, X4 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!)

<u>Statistics:</u>

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^2s)--indicate the proportion of the variance of Y that is explained by a set of IVs. This may be incremental (as in an R^2 change for a single step of a stepwise model or a block in 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) and standardized (β, beta) 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 Xs in the equation. The significance of each partial regression coefficient is tested with an F, which will be 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. And, the SE of each β allows one to establish a CI (confidence interval) around the β coefficient.