Chapter 5a:

Multiple Regression Analysis

Multiple Regression is a statistical technique for estimating the relationship between a dependent variable and two or more independent (or predictor) variables.

How does multiple regression work (consider how bivariate regression works)?

That is, what does it do?

1. Multiple regression uses the ordinary least squares solution (as does bi-variate regression).

That is, it describes a line for which the (sum of squared) differences between the predicted and the actual values of the dependent variable are at a minimum.

Or, in still more technical words, the regression model can be thought of as representing the "function" that minimizes the sum of the squared errors.

$$Y_{pred} = a + b_1X_1 + B_2X_2 \ldots + B_nX_n$$

2. Multiple regression produces a model that identifies the best weighted combination of independent variables to predict the dependent (or criterion) variable.

$$Y_{pred} = a + b_1X_1 + B_2X_2 \ldots + B_nX_n$$

How might
"specification error"
come into play here?

If there are variables left out of the equation that have substantial affect on the dependent variable, then the weights (b and beta) assigned to the independent variables that are included will be substantially affected.

It is because of specification error that we want to review the literature and examine existing theories BEFORE creating our regression model.

In sum, multiple regression allows us to:

--estimate the relative importance of several hypothesized predictors and

--assess the contribution of the combined variables to change the dependent variable.

Indices are often constructed so that a group of ordinal variables can be treated as a single Interval-Ratio variable (SPSS uses the compute tab to add variables together to create an index).

Cronbach's alpha is used to assess whether the variables to be added are measuring the same concept. It measures the internal reliability of the items and ideally is .70 or higher.

The lower the alpha score the more likely that the variables are not measuring the same concept and so should not be added together (analyze, scale, reliability analysis).

The Regression Equation

$Y_{pred} = a + b_1X_1 + B_2X_2 \ldots + B_nX_n$

Y_{pred} = dependent variable or the variable to be predicted.

X = the independent or predictor variables

a = "raw score equations" include a constant or Y intercept

b = b weights; or partial regression coefficients. They show the relative contribution of their independent variable on the dependent variable when controlling for the effects of the other predictors

Using the Regression Equation to predict success of applicants to graduate school

If we ran a regression equation with the dependent variable being the GPA of all current graduate students and the predictor variables being undergraduate GPA score and GRE score, what would the resulting regression analysis tell us?

Suppose the resulting regression equation (explaining college GPA) was:

Y_{pred} = .22 + (.5)(HS-GPA) + (.001)(GRE)

What would this tell us?

Suppose we want to predict the success score of one applicant, Erin, based on her HS-GPA of 3.80 and her GRE score of 1350.

How would we do this?

HS-GPA of 3.80 and GRE score of 1350.

Regression Equation Predicting College GPA

C-GPA$_{pred}$ = .22 + (.5)(HS-GPA) + (.001)(GRE)

Erin's Regression Equation

C-GPA$_{pred}$ = .22 + (.5)(3.80) + (.001)(1350)

C-GPA$_{pred}$ = .22 + 1.9 + 1.35

C-GPA = 3.47 (What does this tell us?)

The variate in multiple regression

The variate is the combination of variables on the right side of the regression equation

$Y_{pred} = a + b_1X_1 + B_2X_2 \ldots + B_nX_n$

The variate is the predicted Y score (not the actual)

The variate is sometimes viewed as representing a "latent" variable

How do you decide which variables to include in the variate?

One approach is to examine past literature and theory and from this to develop a "theoretical" variate.

This is sometimes referred to as the "standard" (simultaneous) regression method

A second approach is to examine statistics that show the effects of each variable both within and out of the equation. The "statistical variate" is built based on those variables showing the most effect. These are sometimes called "Forward and Backward Stepwise Regression"

Lets work through an example of the standard multiple regression method. Lets suppose we have decided that we want to study job commitment and be able to predict whether an employee is going to be committed to her/his job.

What are our first steps?

Lets suppose we examine the literature, and more specifically the theory, explaining job commitment. We find:

Commitment = empowerment + job satisfaction. + management support

Lets further suppose we have a set of data for nursing home employees that uses a Likert Scale (i.e., statements that allow for strongly agree to strongly disagree). What are our next steps?

Commitment = empowerment + job satisfaction. + management support

Do a reliability test for each set below and then, if the alpha scores are acceptable, create index variables

Commitment = 07, 37, 61r
Empowerment = 27, 54, 76, 86, 90
Job Satisfaction (already created)
Management Support = 19, 43, 82

What's next?

Commitment = empowerment + job satisfaction. + management support

Perform the regression analysis and interpret the data (analyze, regression, linear, identify dependent and independent variables).

What do each of the following tell us?

R^2, adjusted R^2, constant, b coefficient, beta, F-test, t-test

Break

It is interesting to note that for a multiple regression "R^2" (the coefficient of multiple determination) is used rather than "r" (Pearson's correlation coefficient) to assess the strength of this more complex relationship (as compared to a bi-variate correlation)

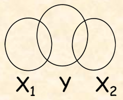
The adjusted R^2 adjusts for the inflation in R^2 caused by the number of variables in the equation. As the sample size increases above 20 cases per variable, adjustment is less needed (and vice versa).

Interpretation of the *b coefficients*

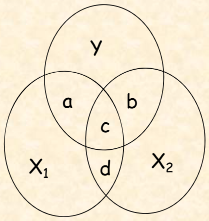
A multiple regression coefficient (or b coefficient) measures the amount of increase or decrease in the dependent variable for a one-unit difference in the independent variable, controlling for the other independent variable(s) in the equation.

Interpretation of the *b coefficients*

Ideally, the independent variables are uncorrelated. Consequently, controlling for one of them will not affect the relationship between the other independent variable and the dependent variable.



(X_1 and X_2 are uncorrelated)



Demonstration of two independent variables that are somewhat correlated with each other and with a dependent variable.

In sum, if the two independent variables are uncorrelated, we can uniquely partition the amount of variance in Y due to X_1 and X_2 and bias is avoided.

Small intercorrelations between the independent variables will not greatly biased the b coefficients.

However, large intercorrelations will biased the b coefficients and for this reason other mathematical procedures are needed (we will be covering interaction affects and multicollinearity in more depth)

Interpreting the Standard Regression Coefficients

All Beta coefficients are in z-score form and thus can be compared with each other. That is, since the independent variables are now in the same metric, we can determine their relative ability to predict the dependent variable.

Consequently, the independent variable with the largest beta weight can be viewed as having the largest impact on the dependent variable.

However, one must be cautious to suggest that one variable has "twice" the effect of another due to various problems such as shared correlation.

Betas cannot be compared or generalized across different samples. This is because betas are "sensitive" to correlations with other predictors.

Solution: consider using other statistics in addition to the betas such as structure coefficients and squared semipartial correlations (these will be covered in class)

Commitment = empowerment + job satisfaction. + management support

How much of the variance in job commitment is accounted for?

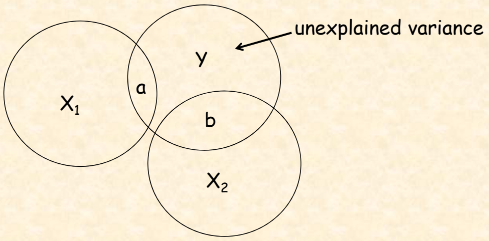
Break

Calculating a standard regression equation:

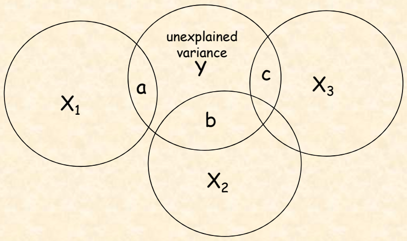
While we may put all three variables into the regression equation at the same time, the regression analysis considers each separately.

More specifically, each variable is entered into the regression equation after the others have already been entered so that the unique (additional) contribution of the variable can be calculated.

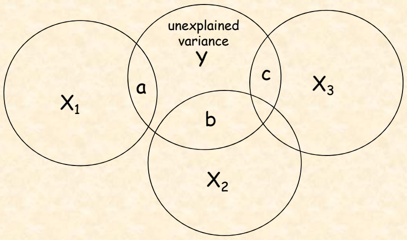
Here is a pictorial of a regression equation showing how X_1 and X_2 account for some of the variance of Y ("a" and "b"). The "unexplained" variance is also identified.



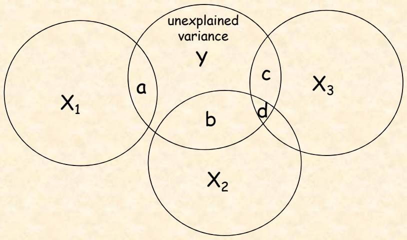
If we include X_3 into the regression equation, the analysis will determine what remaining portion of Y's variance (the unexplained or residual variance) can be explained. This would be "c".



Thus, in evaluating the contribution of X_3, the predictor currently under consideration, it is the residual variance of Y that X_3 must target after statistically removing (i.e., controlling for, partialing out) the effects of X_1 and X_2.

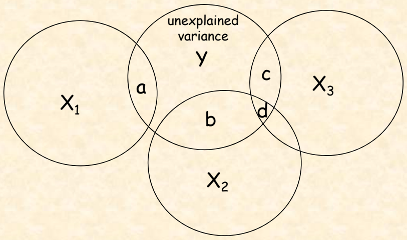


This effect of a single predictor after controlling for others is referred to as a partial correlation (for X_3 this would be c + d). However, the residual uniquely explained and unshared by any other predictors is referred to as the squared semipartial correlation (for X_3 this would be "c" squared)

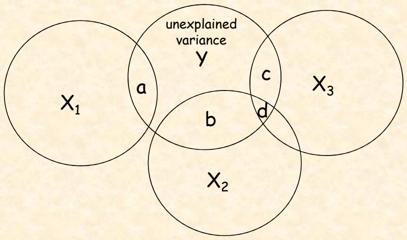


The semipartial correlation is reported in SPSS output as the "part correlation" and must be squared to get the squared semipartial correlation.

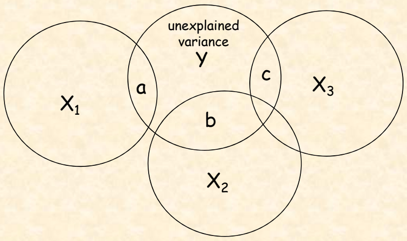
The SSC scores do not add up to equal R^2 because they do not include the shared variance (overlap or correlation) between the independent variables.



It is also interesting to note that, in this pictorial that shows how X_1 , X_2 , and X_3 account for some of the variance of Y, "a" , "b" , "c" and "d", added together, represent the R^2 . The "explained" or "accounted for" variance.



It's also interesting to note that X_2 and X_3 are correlated. When two predictors are correlated it will affect their beta weights. Therefore we will be revisiting this situation.



Repeating the process for each predictor

As previously noted, each predictor is taken in turn. That is, all other predictors are first placed in the equation and then the predictor of interest is entered. This allows us to determine the unique (additional) contribution of the predictor variable.

By repeating the procedure for each predictor we can calculate the unique contribution of each.

The Structure Coefficient

This is the bivariate correlation between (1) a particular independent variable and (2) the predicted scores which can also be thought of in terms of the variate since the variate creates the predicted score.

This is conceptually similar to factor analysis where a correlation is computed between each independent variable and a "factor" (or latent variable).

The Structure Coefficient

SPSS does not provide the structure coefficients. They can be easily calculated by dividing the Pearson correlation between the given predictor and the actual (measured) dependent variable and R the multiple correlation. Or:

$$\frac{\text{Pearson Corr for predictor \& dep. var}}{\text{multiple correlation (R)}}$$

The Structure Coefficient

The larger the structure coefficient the better the predictor reflects the construct underlying the variate.

Independent variables can be compared as to how well they reflect the underlying construct.

A Comparison of Structure Coefficients and Beta weights

Beta weights take into account the predictor's correlation with the other predictors in the analysis, structure coefficients do not.

Beta weights can exceed the range of ± 1 while structure coefficients can not.

Both can be used to calculate the relative contributions of predictors. Statisticians disagree on the value of structure coefficients.

The F and t tests

The F-test is used as a general indicator of the probability that any of the predictor variables contribute to the variance in the dependent variable within the population.

The null hypothesis is that the predictors' weights are all effectively equal to zero. That is, none of the predictors contribute to the variance in the dependent variable in the population.

The F and t tests

t-tests are used to test the significance of each predictor in the equation.

The null hypothesis is that a predictor's weight is effectively equal to zero when the effects of the other predictors are taken into account. That is, it does not contribute to the variance in the dependent variable within the population.

What do we mean by significance?

Thus, one question to ask is:

Does the independent variable contribute to the R^2 when controlling for other independent variables in the regression equation?

The same question is asked for each b coefficient.

In still other words, does the variable's contribution to the slope (regression line) result in more error being reduced than when the variable is not considered.

Null Hypotheses: $b_{yx1} = 0$

In sum, the *t* test is used to determine the significance of the b coefficient and is calculated separately for each.

$$t = \frac{b_{yx}}{s_b} = \frac{\text{sample } b_{yx}}{\text{standard error for } b_{yx}}$$

Using the T-test to determine the significance of the "a" intercept

The question to ask is:

When the independent variables are equal to zero, does the "a" intercept contribute to reducing the error when predicting the dependent variable?

A significant "a" intercept (or constant) indicates that any reduction found in the sample can be expected to exist in the population

Testing the "a" intercept

The *t* test is again used.

The specific "*t*" formula is very complex and fortunately is calculated by statistical software packages.

Using the t-test for bivariate (zero-order) correlation coefficients

Zero-order (bivariate) correlation coefficients are often used to compare correlations between all possible pairs of variables in a regression equation.

For each pair, a special form of the t-test is used to determine its significance. The null hypothesis is that the correlation is unrelated (correlation = 0).

Bivariate correlation coefficients can be obtained through SPSS' "correlation" command.

Comparing "Nested" Regression Equations
(also sometimes referred to as hierarchial linear models)

Nested regression equations are a series of two or more regression equations where independent variables are successively added to an equation to observe changes in the predictors' relationship to the dependent variable (also referred to as block-entry regression method).

That is, does adding a unique set of independent variables significantly increase the R^2.

When comparing the R^2 of an original set of variables to the R^2 after additional variables have been included, the researcher is able to identify the unique variation explained by the additional set of variables.

Any co-variation between the original set of variables and the new variables will be attributed to the original variables.

Example of “Nested” Regression Equations

With nurse aide job satisfaction as the dependent variable, the researcher could run 3 regression equations, each adding additional independent variables:

Regression 1: age, marital status, # of children in household (demographic variables)

Regression 2: age, marital status, # of children in household (demographic variables); rating of organization, time for providing care, adequate staffing (organizational variables)

Example of “Nested” Regression Equations

Regression 3: age, marital statis, # of children in household (demographic variables); rating of organization, time for providing care, adequate staffing (organizational variables); autonomy, competence, impact of work (empowerment)

Determining the significance of adding “Nested” Regression Equations

A special formula of the F-test is used to compare the R^2 of the first equation to the R^2 of the second to determine whether the differences between the two can be generalized to the population.

$$F = \frac{(R^2_2 - R^2_1) / ((K_2 - K_1)}{(1 - R^2_2) / (N - K_2 - 1)}$$

--Subscripts attached to R2 and K indicate whether values come from the first (less-inclusive) equation or the second.

--K indicates the number of independent variables in the equation.

Determining the significance of adding “Nested” Regression Equations

The purpose of the F test is to determine whether additional variation in the dependent variable is explained by adding the additional variables.

For the F test to be significant, the difference in R2 must be large relative to the number of independent variables added to the second equation.

A similar idea is to conduct a Significance Test in order to Compare Two Regression Equations

The researcher may want to determine whether the independent variables affect the dependent variable the same for two different groups (say men and women)

For example, in 1995 a Ph.D. student (Dr. Leslie Stanley-Stevens) wanted to know if a set of independent variables had the same effect on job satisfaction for men as for women (she’s now an associate professor at Tarelton State University).

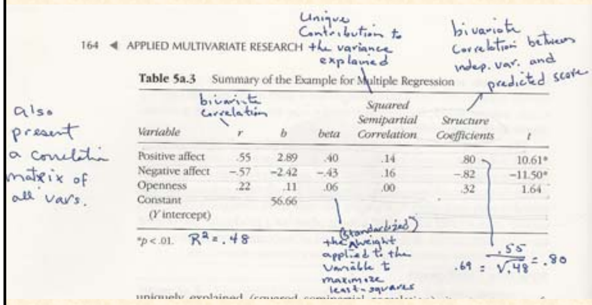
Comparing two regression equations can be accomplished with the correlation difference test.

This is a statistical test to determine whether two correlation coefficients differ.

In short it transforms both the sample correlations to Z scores, applies the test, and the result is a Z statistic that can be checked for its significance level.

If the Z statistic is significant, then the difference found between the two correlations is not due to chance but to the fact that the two groups are affected differently by the independent variables.

Presentation of data using statistical procedures learned in class



164 ◀ APPLIED MULTIVARIATE RESEARCH

Unique Contribution to the variance explained

bivariate Correlation between indep. var. and predicted score

Table 5a.3 Summary of the Example for Multiple Regression

also present a correlation matrix of all vars.

bivariate correlation

Variable	*r*	*b*	*beta*	*Squared Semipartial Correlation*	*Structure Coefficients*	*t*
Positive affect	.55	2.89	.40	.14	.80	10.61*
Negative affect	−.57	−2.42	−.43	.16	−.82	−11.50*
Openness	.22	.11	.06	.00	.32	1.64
Constant (*Y* intercept)		56.66				

*$p < .01$. $R^2 = .48$

(standardized) the weight applied to the variable to maximize least-squares

$.69 = \frac{.55}{\sqrt{.48}} = .80$

uniquely explained (squared semipartial correlation

Also note mistakes in the text (as far as I can determine):

P.161, "The slanted-line areas…" should be "The dark gray" areas…

P.166, "The second numerical column…" should be "The first numerical column…"

P.167, "an increase of 2.89 points on the positive affect measure is, on average, associated with a 1-point gain in self-esteem." This should be reversed since self-esteem is the dependent variable (one unit change in the independent variable produces a quantified change in the dependent variable not vice versa).

DRABBLE by Kevin Fagan
DAD, YOU'D BETTER SIT DOWN! I'VE GOT BIG NEWS!
I GOT A JOB AT ALL-MART! I STAND AT THE DOOR AND GREET PEOPLE AS THEY ENTER THE STORE!
CAN YOU BELIEVE IT?! WHAT A SWEET JOB!!
I WAS SO EXCITED, I COULD HARDLY SLEEP DURING MATH CLASS!
CONGRATULATIONS SON!

Suppressor Variables

Variables that are NOT highly correlated with the dependent variable but are able to heighten the numerical effect of ANOTHER predictor on the dependent variable.

In other words, when a suppressor variable is included in a regression, the suppressor variable will appear largely unrelated to the dependent variable while the effect of ANOTHER predictor variable will increase substantially.

How Does it Work?

Suppressor Variables are difficult to conceptualize. Please read the Meyers text for further explanation (p.182)

How Does it Work?

By virtue of its correlation with a predictor variable, it accounts for (statistically controls for or negates) that portion of the predictor variable not related to the dependent variable and thus makes the predictor variable a better predictor than it would be in the absence of the suppressor variable.

Indications of suppressor variables include:

- the correlation between it and the dependent variable is substantially smaller than its beta weight
- its Pearson correlation with the dep. variable and its beta have different signs
- it may have a near zero correlation with the dependent variable but yet is a significant predictor
- it may have little or no correlation with the dependent variable but is correlated with one or more of the predictors

"Complete" and "Intrinsic" Linear Regression

Linear regression can be viewed from two perspectives: a "completely" linear model and an "intrinsically" linear model:

1. a completely linear model exists when the level of measurement for all the variables of interest are linear

   Consequently, the weighted variables can be added together to obtain the predicted score

"Complete" and "Intrinsic" Linear Regression

2. a less than completely linear model or "intrinsically" linear model exists when there is a need to alter or transform one or more of the variables (e.g., squared; logged).

   In these regression models we can still add the weighted variables together to obtain the predicted score.

When thinking about intrinsically linear models, what are some situations where we would want to alter or transform the dependent variable?

What about the independent variable?

Examples of transforming the dependent variable

1. if the dependent variable is dichotomous, we can do a logit transformation. This will cause us to interpret the dependent variable in terms of proportions

2. If heteroscedasticity is a problem with our data, we can do a natural logarithm transformation to solve this problem (we will be examining "assumptions" we make about the data when doing regression analyses; one of these assumptions is that there is no heteroscedasticity).

Examples of transforming the independent variable

1. if the independent variable has a curvilinear relationship with the dependent variable (also referred to as polynomial regression), the variable can be squared, cubed, or whatever is called for.

   A "quadratic function" has one "bend" in the curve (also referred to as a second-order polynomial) and is squared while a "cubic function" has two bends (also referred to as a third-order polynomial) and is cubed.

Examples of transforming the independent variable

2. Using a log transformation of the independent variable is appropriate if the independent and dependent variables increase together up to a point and then the effects of the independent variable no longer have the same increasing effect on the dependent variable (the effect "levels off" at some point).

Examples of transforming the independent variable

3. "Dummy" variables are used if the level of measurement of the independent variable is nominal. The nominal variable is "transformed" into a set of "dummy" variables where each value of the nominal variable becomes a separate "dummy" variable.

   We will be covering Dummy Variables in more detail later.

Examples of transforming the independent variable

4. Interaction terms. Regression analysis "assumes" that the independent variables are unrelated (independence assumption). However, suppose that the relationship between X_1 and the dependent variable differs for different levels of X_2.

   For example, the effect of age on income differs depending on the level of education the person has.

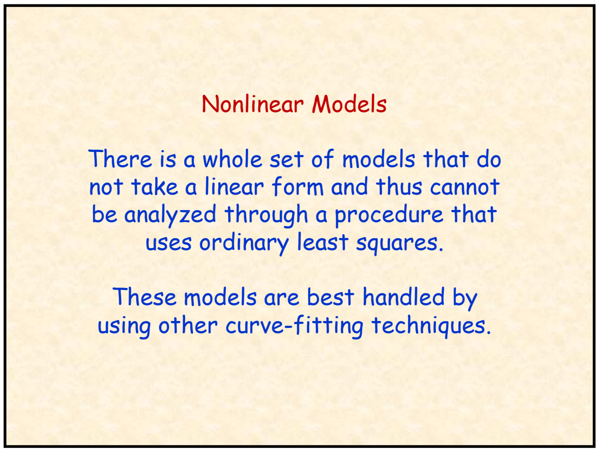
Examples of transforming the independent variable

In this example there is an "interaction" between age and education, with education being referred to as a "moderator" variable (also called a "conditional" variable or a "specification" variable)

Examples of transforming the independent variable

While the variables are not "transformed" in the same sense as squaring or logging a variable, a new interaction term is introduced into the regression equation to account for the inter-relationship between the two independent variables.

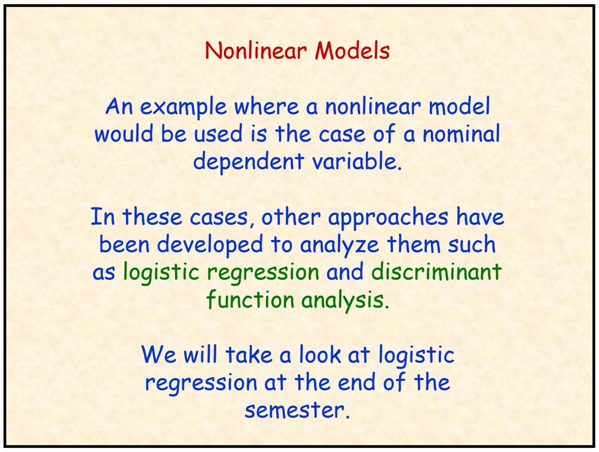
We will be examining interaction effects and how to identify and interpret them.



Nonlinear Models

There is a whole set of models that do not take a linear form and thus cannot be analyzed through a procedure that uses ordinary least squares.

These models are best handled by using other curve-fitting techniques.

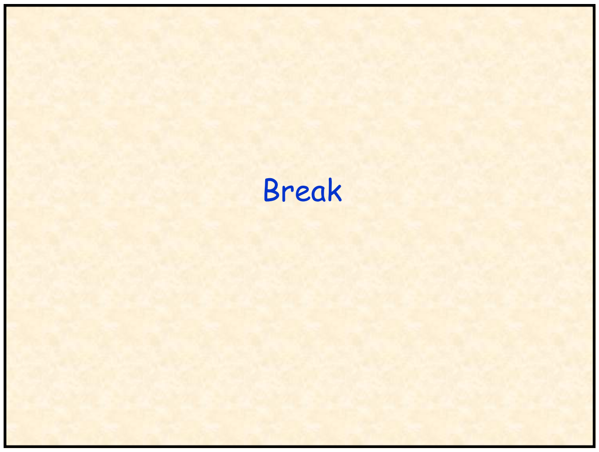


Nonlinear Models

An example where a nonlinear model would be used is the case of a nominal dependent variable.

In these cases, other approaches have been developed to analyze them such as logistic regression and discriminant function analysis.

We will take a look at logistic regression at the end of the semester.



Break