The University of Georgia

Variance Partitioning

November 8, 2007

ERSH 8320

Today's Lecture

- Muddying the waters of regression.
- What not to do when considering the relative importance of variables in a regression model.
- What you can say about certain techniques we have discussed.
- Discussion of the limitations of regression analysis.

Explanation

- Last several classes we focused on the predictive aspects of multiple regression analysis.
- Explanation is the ultimate goal of science.
- Part of explanation is determining the relative importance of variables.
- Statistical techniques important in explanation:
  - Variance partitioning.
  - Analysis of effects.

R^2 Partitioning

- From the lecture on partial correlation, recall that the R^2 for a multiple regression model can be decomposed into a set of squared semi-partial correlations ($R^2_{y.1234}$ represents the R^2 for a regression with Y as the dependent variable, and X_1, X_2, X_3, and X_4 as independent variables):

$$R^2_{y.1234} = r^2_{y1} + r^2_{y(2.1)} + r^2_{y(3.12)} + r^2_{y(3.123)}$$

- Rearranging the ordering of the variables will produce the same value of R^2:

$$R^2_{y.1234} = r^2_{y4} + r^2_{y(1.4)} + r^2_{y(2.41)} + r^2_{y(3.412)}$$

- What does change, however, is the proportion of variance incremented by adding an additional variable.

Order of Entered Variables

- Recall the Erie house data, our most familiar data set...
- We have data for 27 houses sold in the mid 1970's in Erie, Pennsylvania:
  - X_1: Current taxes (local, school, and county) $\div$ 100 (dollars).
  - X_2: Number of bathrooms.
  - X_4: Living space $\div$ 1000 (square feet).
  - Y: Actual sale price $\div$ 1000 (dollars).

Order of Entered Variables

- First, fit the following sequence of models:
  1. $Y = a + b_1X_1$
  2. $Y = a + b_1X_1 + b_2X_2$
  3. $Y = a + b_1X_1 + b_2X_2 + b_4X_4$
- The R^2 for each model:
  1. 0.838 - X_1 by itself.
  2. 0.856 - X_2 adds 0.018 to model with X_1.
  3. 0.935 - X_4 adds 0.074 to model with X_1 and X_2.

Order of Entered Variables

- Now, fit the following sequence of models:
  1. $Y = a + b_4 X_4$
  2. $Y = a + b_4 X_4 + b_2 X_2$
  3. $Y = a + b_4 X_4 + b_2 X_2 + b_1 X_1$
- The R^2 for each model:
  1. 0.863 - X_4 by itself.
  2. 0.903 - X_2 adds 0.040 to model with X_4.
  3. 0.935 - X_1 adds 0.074 to model with X_4 and X_2.

Order of Entered Variables

- Finally, fit the following sequence of models:
  1. $Y = a + b_2X_2$
  2. $Y = a + b_2X_2 + b_1X_1$
  3. $Y = a + b_2X_2 + b_1X_1 + b_4X_4$
- The R^2 for each model:
  1. 0.854 - X_2 by itself.
  2. 0.902 - X_1 adds 0.048 to model with X_2.
  3. 0.935 - X_4 adds 0.033 to model with X_2 and X_1.

Which Variable is Most Important?

- From the three sets of regressions we found that:
  - The R^2 for the full model (with all three variables) did not change as a function of how the variables were entered.
  - The incremental variance accounted for changed as a function of when the variable was entered into the model.
- Based on these three models, which variable would you say does the best at explaining the sale price of a home?
  - Number of bathrooms?
  - Current taxes?
  - Living space?

Limitations of R^2

- Realizing that the increase in R^2 is a function of the other variables already in the model, one can come to understand several things about using R^2 to determine the relative importance of variables.

- R^2 sample specific.
  - Even if all regression effects are the same, R^2 can be reduced.

- R^2 changes from sample to sample because of the variability of a given sample on:
  - The variables under study.
  - The variables not under study.
  - Errors in the measurement of the dependent variable (Y).

Limitations of R^2 for Variance Partitioning

- So, R^2 is not all that it seems to be.
- A quote against variance partitioning:

> "It would be better to simply concede that the notion of 'independent contribution to variance' has no meaning when predictor variables are intercorrelated" (Darlington, 1968, p. 179)

Purposes of Variance Partitioning

- If the order of entry of a variable is important in incremental partitioning, how does one determine the order?

- To answer this question, distinguish between two purposes of such an analysis:

  1. To study the effect of an independent variable (or multiple independent variables) on the dependent variable after having controlled or other variables.

  2. To study the relative effects of a set of independent variables on the dependent variable.

Purposes of Variance Partitioning

- Purpose #1 equates to that of control.

  > "Such an analysis [for control purposes] is not intended to provide information about the relative importance of variables, but rather about the effect of a variable(s) after having controlled for another variable(s)." (Pedhazur, 1997, p. 245)

- The purpose of today's class is to demonstrate why incremental partitioning of variance is not valid for determining the relative importance of variables.

- Incremental partitioning of variance can be used for studying the effects of a set of variables while controlling for a set of variables.

Control

- Deciding upon which variables are in need of controlling for is not arbitrary.

- Unless dictated clearly by theory, there is often no way of knowing which variables to control for in an analysis.

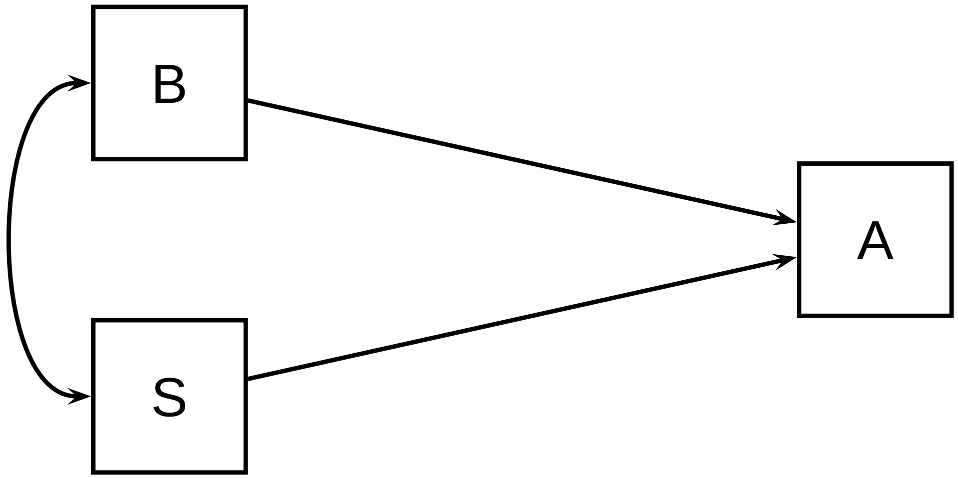
- Imagine the example given by Pedhazur (p. 245), of a study with three variables (exclusively):

  B - Student background characteristics.
  S - School quality.
  A - Academic achievement.

Definitions

- In this example, causality is depicted by what is called a path diagram.
- Path diagrams are commonly used in analyses using Structural Equation Modeling (SEM) techniques.
- Time permitting, we will briefly discuss this topic in a future lecture.
- Therefore, to understand this example, two terms are in need of definition:
  - Exogenous variable - a variable whose variability is assumed to be determined by causes outside of the causal model under consideration.
  - Endogenous variable - a variable whose variability is to be explained by exogenous and other endogenous variables in the causal model.

Example Model #1



- Exogenous variables - B and S.
- Endogenous variable - A.
- Curved line with arrowheads at both ends - correlated variables (correlation is assumed and not modeled - always the case with correlated exogenous variables).
- Straight line - causal path.

Example Model #1

- Often in non-experimental or observational research, exogenous variables are correlated.
- Typically, these variables are used as input into a regression analysis.
- The correlation between exogenous variables is taken as fact, and is not being explained by such a model.
- Because such exogenous variables are typically correlated, inferences about the relative effects of either are impossible to ascertain.
- Incremental variance partitioning cannot be used to describe the relative effects of each variable.

Example Model #2

- Exogenous variable - B.
- Endogenous variables - S and A.

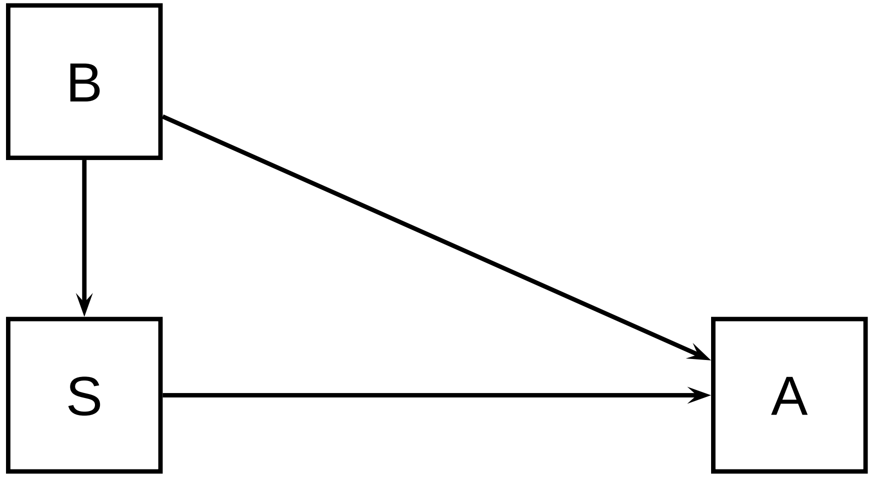
Example Model #2

- In this situation, the hypothesized model has B (background) having a direct effect on S (school quality).
- B does not have a direct effect on A (achievement).
- S mediates the effects of B on A.
- Because of the direct effect, this model allows for the study of the effect of S after controlling for B:

$$R^2_{A.BS} - R^2_{A.B}$$

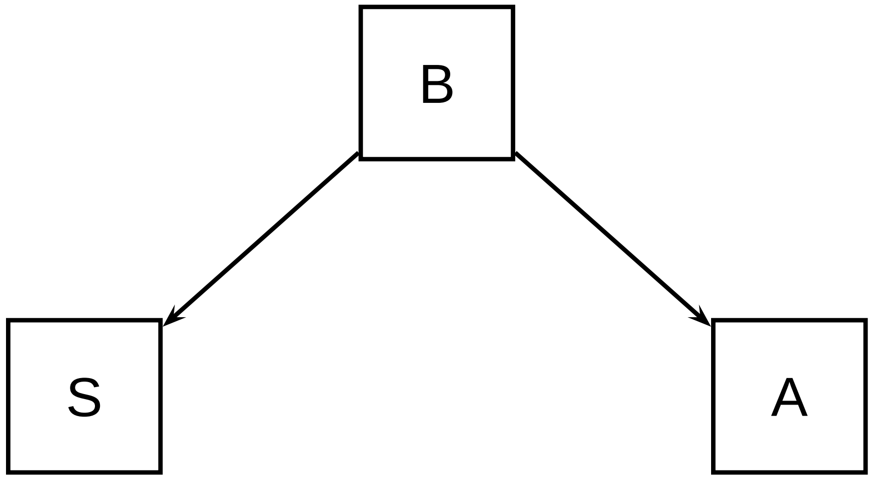
- You may recall this as the squared semipartial correlation between A and S, with the effects of B removed from S, or $r^2_{A(S.B)}$.

Example Model #3



- Exogenous variable - B.
- Endogenous variables - S and A.
- S mediates the effects of B on A.
- B additionally has a direct on A.
- Think path analysis.

Example Model #4

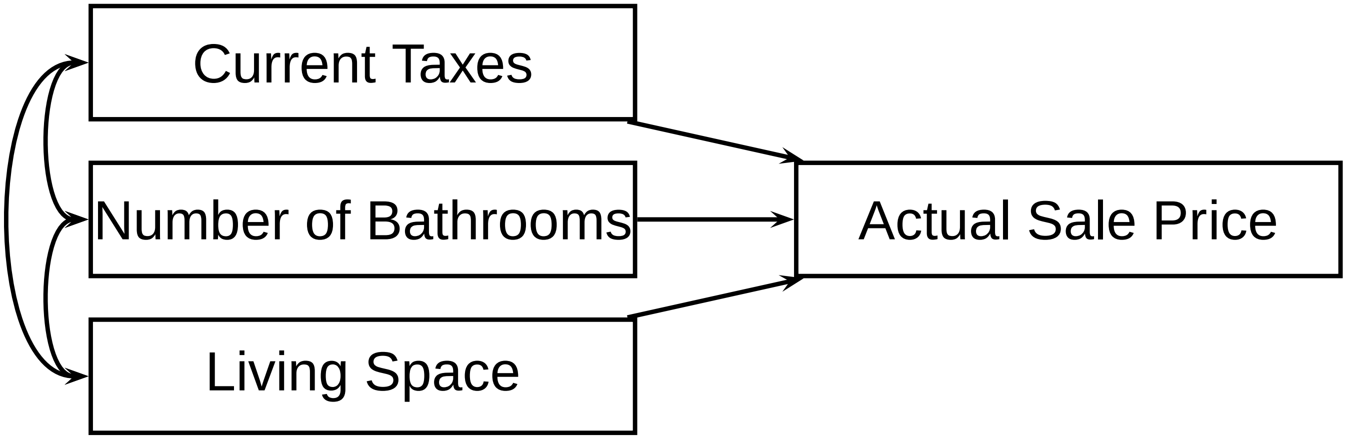


- Exogenous variable - B.
- Endogenous variables - S and A.
- Here, controlling for B would equate to calculating the partial correlation between S and A with the effect of B removed from both S and A, or $r_{SA.B}$.

Erie Example

- Of the four models, which best represents our well known Erie house analysis?
- Ask yourself the following questions:
  - Which variables are exogenous?
  - Which variables are endogenous?
  - What are the direct effects?
  - Are the exogenous variables correlated?
- Multiple regression analyses are fairly straight in the answers to these questions.
- Only experimental research can sometimes control the correlation of exogenous variables through experimental design and control.

Erie House Diagram



- Exogenous variables - Current Taxes, Number of Bathrooms, Living Space.
- Endogenous variables - Actual Sale Price.
- Because of the correlated exogenous variables, describing the relative importance of variables by incremental variance is not possible.

Commonality Analysis

- Commonality analysis is a method of partitioning the variance to find the proportion of variance accounted for in a dependent variable by unique portions of each of the independent variables.
- The method of commonality breaks down to analyzing each of the possible regression models that could be built to predict the dependent variable.
- The unique portion of each independent variable is equivalent to the proportion of variance incremented when it is the last variable entered in an analysis.
- This, literally, is the squared semipartial correlation between the dependent variable and the independent variable in question, partialing out all other independent variables from the variable in question.

Commonality Analysis

- For each variable in a regression analysis, the uniqueness can be found.

- For each variable combination (two-way, three-way, etc...) the amount of variance in the dependent variable accounted for by the variable combination is called the communality.

- These values are all attainable from running multiple multiple regression analyses.

Commonality Analysis

- To find the communality of a set of variables, or to find the uniqueness of a variable, there is an easy schematic formula to use.

  1. For each variable of interest, use $(1 - X_a)$, where a symbolizes the variable of interest.
  2. For all other variables use X_b.
  3. Multiply all together, and then multiply by -1.
  4. Expand the multiple, the resulting factor is the combination of model $R^2_{Y.ab}$ to produce the communality or uniqueness.

Commonality Analysis

- For our Erie example, with X_1, X_2, and X_3, the uniqueness of X_1 is found by:

$$-(1 - X_1)X_2X_4 = -(X_2X_4 - X_1X_2X_4) = X_1X_2X_4 - X_2X_4$$

- So the uniqueness $(U(1))$ is given by (note, this is the squared semipartial correlation):

$$U(1) = R^2_{Y.124} - R^2_{Y.24}$$

- The communality of X_1 and X_2 is found by:

$$-(1 - X_1)(1 - X_2)X_4 = X_2X_4 + X_1X_4 - X_4 - X_1X_2X - 4$$

- So the communality $(C(12))$ is found by:

$$C(12) = R^2_{Y.24} + R^2_{Y.14} - R^2_{Y.4} - R^2_{Y.124}$$

Commonality Analysis

	X_1	X_2	X_4	Y
X_1	1.000			
X_2	0.876	1.000		
X_4	0.832	0.901	1.000	
Y	0.915	0.924	0.929	1.000

- $R^2_{Y.12} = 0.902$
- $R^2_{Y.14} = 0.928$
- $R^2_{Y.24} = 0.903$
- $R^2_{Y.124} = 0.935$

Commonality Analysis

	X_1	X_2	X_4
$U(1)$	0.032		
$U(2)$		0.007	
$U(4)$			0.033
$C(12)$	0.033	0.033	
$C(14)$	0.016		0.016
$C(24)$		0.058	0.058
$C(124)$	0.756	0.756	0.756
$\hat{\rho}$	0.915	0.924	0.929
$\hat{\rho}^2$	0.837	0.853	0.863

Uniqueness Comments

- When the independent variables are uncorrelated, the uniqueness of each is equal to the squared zero-order correlation between it and the dependent variable.

- When the independent variables are uncorrelated, all commonalities are zero.

Commonality Analysis

- Commonality analysis again cannot answer the relative importance of the variables.
- Such analyses are useful for predictive research, but may not be too helpful in explanation.
- Commonality analysis, regardless of theoretical model, yields exactly the same results
- Commonalities can have negative signs (suppressor variables).
- Exponential increase in commonalities accompanies linear increase in independent variables.

Final Thought

- Variance partitioning is good for describing model effects with other variables controlled statistically.

- Variance partitioning is not good for determining

- Pedhazur provides many examples of published research reports where incremental variance is used inappropriately. This is suggested reading.

Next Time

- More on explanation.
- Chapter 10 - Analysis of Effects.
- More questions, not many more answers.