Analysis of Effects

November 13, 2005

ERSH 8320

Today's Lecture

- Muddying the waters of regression.
- What to consider when investigating the relative importance of variables in a regression model.
- Factors playing into interpretation of regression coefficients.

Analysis of Regression Coefficients

- Typically, one can ascertain valuable information directly from the coefficients (or parameters) of a regression model.
- In interpreting such coefficients, however, one must recognize that their values may not be accurate due to a variety of reasons:

  1. Model specification errors.
  2. Errors in measurement.
  3. Collinearity.

- Additionally, when should one consider interpreting standardized coefficients?

Regression Effects and Research Type

- In experimental research, one has control over the levels of the independent variables, in addition to having random assignment.
- Because of this control and randomization, regression coefficients can be interpreted confidently.
  - The coefficient of an independent variable represents the change in the dependent variable caused by an increase in the independent variable.
- Variables that may be related to the dependent variable but are not manipulated are incorporated into the random error term due to randomization in experimental research.

Regression Effects and Research Type

- In quasi-experimental research, one manipulates levels of the independent variables, but does not randomly assign observations to differing conditions.
- In nonexperimental (or observational) research, one cannot manipulate levels of the independent variables, and one does not randomly assign observations to differing conditions.
- Lack of randomization means that variables not under study can systematically influence the dependent variable, potentially biasing the results of the analysis.

> "To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)" (Box, 1966; p. 629).

Proxy Variables

- Often in nonexperimental research, variables are proxies for target variables in the regression equation.
- Such proxies are used because the target variable of interest is difficult to measure.
- Care should be taken in interpretation of such proxies so that they are interpreted as the variables they truly are, and not the variable they represent.

Intercorrelated Variables

- Another regression idea that is often not plausible in nonexperimental research is that of the holding variables constant while other variables change.
- If variables are highly correlated, holding some values constant while changing others does not seem realistic.
- Having correlated variables implies that as one variable changes, so do the others in a systematic way.
- Even though this is the case, you often hear of regression parameters as the change in the dependent variable when the specific parameter's independent variable changes, while all other independent variables are held constant.

Specification Errors

- Specification errors are errors where the wrong model was used to analyze the data.
- Types of specification errors include:
  1. Omission of relevant variables.
  2. Inclusion of irrelevant variables.
  3. Nonlinear relationships between X and Y.
  4. Nonadditive relationships between X and Y.

Omission of Relevant Variables

- If you run a regression and omit relevant variables, your estimates will be biased.
- The term “relevant variables” refers to variables that are correlated with other variables in the equation (either X or Y).
- If omitted variables are uncorrelated with X yet are correlated with Y, the standard errors of your model will be biased (upward).
- If omitted variables are uncorrelated with X *and* uncorrelated with Y, then no bias occurs.

Inclusion of Irrelevant Variables

- So you are concerned with omitting relevant variables and so you add every possible variable to your regression model, including the kitchen sink.

- Good news: By including irrelevant variables you do not bias your other regression coefficients.

- Bad news: By including irrelevant variables, you do reduce the number of degrees of freedom for the entire model.

- If your sample is small to begin with, this can result in a dramatic increase in the standard error of the estimate for each of your regression coefficients.

Nonlinearity

- Nonlinearity of a regression model occurs when nonlinear relationships exist between the predictor variables and Y.

- The multiple regression model posits a linear relationship between Y and each of the X variables:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + e$$

- If the true relationship is something like:

$$Y = a + b_1X_1 + b_2X_1^2 + \ldots + b_kX_1^k + e$$

  then remedies exist that can be accomplished by techniques that will be taught after spring break.

- Otherwise...recall the lecture on transformations...or try nonlinear regression techniques (driven by theory).

Nonadditivity

- Nonadditivity of variables in the regression model is another type of specification error.
- One type of nonadditivity is that of interacting predictor variables:

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + e$$

- Interaction terms will be covered in more detail in a later lecture.

Detecting Specification Errors

- Detection of specification errors can be difficult.
- Often the easiest type of specification errors to detect are those made by inclusion of irrelevant errors.
- After that, one can sometimes detect if a nonlinear relationship occurs by inspection of residual plots (see diagnostics lecture).
- Omitting relevant variables is often the most difficult error to detect, particularly if you have not measured variables that have an outside influence upon your model.

Measurement Errors

- Measurement errors have been classified into three different groups:
  - Conceptual.
  - Consistent.
  - Random.

- Errors in measurement have been discussed briefly, but this discussion always ends with something like “beyond our scope.”

Measurement Errors: Conceptual

- Conceptual errors are committed when proxy variables are used to represent a construct important in a regression.
- Because using variables related to the construct are not perfectly measures of a construct, the measurement of that construct is taken in error.
- This problem can sometimes be overcome by modeling multiple indicators (multiple proxies) of constructs using advanced techniques such as Structural Equation Modeling.
- Conceptual measurement errors lead to biased regression parameters and standard errors.

Measurement Errors: Consistent

- Consistent measurement errors are errors that occur systematically across sets of variables.
- For example, measurement of sensitive information can bring about socially desirable responses in survey items.
- Furthermore, systematic errors in instruments, research settings, interviewers, and so on can be the cause.
- Consistent measurement errors lead to biased regression parameters and standard errors.

Measurement Errors: Random

- Random errors are the result of random processes.
- Such errors equate to inflating the random variability in the variables contained in a regression model.
- Variability that is not systematically explained by the model (as a linear function of the predictor variables) ends up increasing the SS_{res} of a model.
- As a result, standard errors are inflated.
- R^2 is reduced.
- A deflation of effects is also present (downward bias in regression coefficients).

Collinearity

- Collinearity (a.k.a. multicollinearity) occurs when variables under study are highly related.
- A literal (geometric) meaning of collinearity is given by what happens when two variables fall upon the same line.
- Having variables on the same line leads to having a singular $\mathbf{X}'\mathbf{X}$ matrix, meaning $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist.
- Recall:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$var(\mathbf{b}) = \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{SS_{res}}{N-k-1}(\mathbf{X}'\mathbf{X})^{-1}$$

Collinearity

- Collinearity is very much a matter of degree.
- Complete linear dependence rarely happens (unless some variables are produced by additive components of others).
- Collinearity is often hard to detect.
- Recall MATLAB stating certain matrices were singular to a given tolerance...the same type of ideas hold in this case as well.
- Big biases in both the point estimate of the regression coefficients and the standard errors occur due to collinearity.

Collinearity Diagnostics: VIF

- The variance inflation factor, or VIF, is a measure of the collinearity of a given predictor variable.
- For a variable i, the VIF is calculated by computing the R^2 from a regression with i and all other predictor variables.

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Larger values of VIF indicate more collinearity.
- The larger the VIF, the larger the standard error of the regression coefficient for variable i.

Collinearity Diagnostics: Tolerance

- The tolerance is another measure of the collinearity for a given variable i.
- Like the VIF, the tolerance is calculated by computing the R^2 from a regression with i and all other predictor variables.

$$\text{Tolerance}_i = 1 - R_i^2 = \frac{1}{VIF_i}$$

- Smaller values of tolerance indicate more collinearity.
- The smaller the tolerance, the larger the standard error of the regression coefficient for variable i.
- Tolerance and VIF can be calculated in SPSS by checking “collinearity diagnostics” from the Statistics box under Analyze...Regression...Linear.

Collinearity Diagnostics: Condition Indices

- Condition indices are computed from finding the eigenvalues from the correlation matrix of the variables in the study.

- The condition number gives an overall summary of collinearity from an analysis (larger indicates more collinearity):

$$CN = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

- The condition index gives an estimate of collinearity for each successive eigenvalue:

$$CI_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

Scaling and Centering Variables

- Scaling and centering variables can affect the values obtained in the condition indices.
- Some suggest normalizing the variables (dividing each X by $\sqrt{\sum X^2}$).
- Others suggest that centering a variable can help (subtracting the mean of each variable from itself).
- Centering a variable does not necessarily reduce collinearity, although it can mask the effects of collinearity.
- Centering is useful, however, for things like nonlinear regression functions (powers of X - this is sometimes referred to as nonessential collinearity).

Collinearity Remedies

- Consider dropping suspect variables from an analysis.
  - But that could lead to specification errors.
- Reduce the number of variables in your analysis by calculating principal components.
- Using multiple indicators to measure latent factors in more advanced topics (SEM).

Standardized Coefficients

- Standardized regression coefficients are computed by:

$$\beta_j = b_j \frac{s_j}{s_y}$$

- The size of β reflects not only the effect of the relationship, but also the variability in the model.
- Because the variability of the model is also reflected by the unstandardized coefficients, these can differ widely from different samples.
- There is much debate about the effectiveness of relying on standardized coefficients.

Interpretability of Types of Coefficients

- Unstandardized coefficients reflect the scale upon which Y was measured, often leading to easier interpretation of effects.
- Pedhazur suggests reporting both standardized and unstandardized coefficients, allowing for readers to interpret both values.
- There is agreement that unstandardized coefficients should be used when comparing regression equations across groups.
- We will discuss statistical techniques for doing this after spring break.

Hierarchical v. Simultaneous Analyses

- Hierarchical analysis (for the current chapter and lecture) will be considered the analyses presented in the previous lecture: an analysis that proceeds by adding variables to the model sequentially.
- Simultaneous analyses are considered analyses that have the model formulated well in advance, and have all variables entered into the model at one time.

Final Thought

- Using regression coefficients can be an effective method for determining the relative importance of predictor variables.

- In nonexperimental research, however, literal evaluation of the effects can be misleading.

- Regression effects can be inaccurate due to the problems we covered today.

- Be sure to make inferences accordingly, using caution as you proceed.

Overview
Analysis of Effects
Specification Errors
Measurement Errors
Collinearity
Standardized Coefficients
Hierarchical Analyses
Wrapping Up
- Final Thought
- Next Class

Next Time

- No class Thursday...or the following week.
- Chapter 11, 12, and 14 (on November 27).