Regression Diagnosis:

Examining Assumptions, Identifying Violations, and Exploring Solutions

Hsueh-Sheng Wu

CFDR Workshop Series

Summer 2010

BGSU

Center for Family and Demographic Research

Outline

- What is regression?
- A closer look at the assumptions in regression analysis
  - What does each of these assumptions mean?
  - What happens if these assumptions are violated?
  - How to detect the violations of the assumptions?
  - What to do if these assumptions are violated?
- Other considerations in using regression analysis
- Conclusions

What Is Regression?

Regression is used to study the relation between a single dependent variable and one or more independent variables. In regression, the dependent variable *y* is a linear function of the *x*'s, plus a random disturbance ε.

$$y = a + b_1x_1 + b_2x_2 + \varepsilon$$

a is the intercept

y is the dependent variable

x_1 and x_2 *are independent variables*

$b_1$and b_2 *are regression coefficients*

ε represents the combined effects of all the causes of y that are not included in the equation, but can influence the relations between x's and y

Five Assumptions of Regression

1. Linearity
   - y is a linear function of the x’s
2. Mean independence
   - the mean of the disturbance term is always 0 and does not depend on the value of x’s
3. Homoscedasticity
   - The variance of ε does not depend on the x’s
4. Uncorrelated disturbances
   - The value of ε for any individual in the sample is not correlated with the value of ε for any other individuals
5. Normal disturbance
   - ε has a normal distribution

1. Linearity

What it means?

- The dependent variable *y* is a linear function of the *x*'s
- Possible causes of violating this assumption:
  - Inaccurate specification of the regression models
  - Influential observations

What are the consequences?

- Biased estimates of a, b1, b2
- Inaccurate prediction of y

1. Linearity (continued)

How to detect the violation?

- Plot y against x
- Plot residuals against x
- Plot residuals against y_{hat}
- F test for lack of fit

Solutions:

- Re-specify the model by mathematically transforming x's. e.g., for a curvilinear relation, you can square the x's.
  - log transform
  - exponentiation is the use of the inverse of a logarithm, as in $x' = \varepsilon^x$
  - polynomial transformation is the use of powers of the variable, as in $x' = x^2$, $x' = x^3$, x' = SQRT(x). We use this approach often in multiple regression.
  - rescale the x variable into a dummy (dichotomous) variable
- Restrict the range of x
- Identify the influential cases and examine whether they should be included in the sample

2. Mean Independence

What it means?

- The mean of the disturbance term is always 0 and does not depend on the value of x’s.
- Possible causes of violating this assumption:
  - omitted x variables: if any of the omitted variables is associated with the x’s.
  - reverse causation: if y influence x’s, then ε is associated with the x’s.
  - measurement error in the x: x includes not only x but also something else. This something else will get into ε.

What are the consequences?

- Biased estimates of a, b1, b2
- Inaccurate prediction of Y

2. Mean Independence (continued)

How to detect the violation?

- Without additional data, there is no easy way to detect the violation of this assumption.

Solutions:

- Use of past literatures to justify your model
- Use experimental design to collect your data, which not only support the mean independence assumption, but also avoid reverse causation
- If you use survey design and have measures of relevant variables that have not been included in the model, you can include these variables in the model to reduce the possibility of violating this assumption
- Use simultaneous equations to model reciprocal relations between x’s and y
- Choose measures with high reliability or include measurement models in regression analysis

3. Homoscedasticity

What it means?

- Homoscedasticity means that the variance of ε is the same across all levels of x’s.
- Possible causes of violating this assumption.
  - Improvement in data collection techniques: During the course of data collection, the interviewers are getting better and less likely to commit error in collecting data.
  - Learning: Respondents are less likely to have errors in answering the same questions when being interviewed in the follow-up survey than in the baseline survey.
  - Outliers

What are the consequences?

- Inefficiency: observations with larger disturbance variance contain less information than observations with smaller disturbance variance. but OLS weights them equally.
- Bias in standard errors can leads to incorrect conclusions.

3. Homoscedasticity (continued)

How to detect the violation?

- Plot residuals against X
- Plot residuals against Y_{hat}
- Breusch-Pagan/Cook-Weisberg Test for linear forms of heteroskedasticity
- White's General Test for non-linear forms of Heteroskedasticity

Solutions:

- Re-specify the model or transform the dependent variable
- Use robust standard errors
- Use weighted least squares only if you know what weights to use

4. Uncorrelated Disturbances

What it means?

- The disturbance variables for any two individuals must be uncorrelated.
- Possible causes of violating this assumption
  - Sample design: simple random sampling is not likely to cause this problem, but a cluster sampling is.
  - The selection of unit of analysis, e.g., the couple
  - The use of panel data

What are the consequences?

- Inefficient estimates
- Downward bias in estimated standard errors, which means that there will be a tendency to conclude that relations exist when they really don’t.

4. Uncorrelated Disturbances (cont.)

How to detect the violation?

- Calculate the residuals for all respondents and then examine correlations between the residuals of suspected groups of respondents
- Intra-class correlation

Solutions:

- Include the cluster variables into the models as a control
- Use regression with robust standard errors
- Use generalized least squares

5. Normality

What it means:

- The disturbance term ε need to be normally distributed, but x's and y do not.
  - Positive Skewness
  - Negative Skewness
  - Positive Kurtosis
  - Negative Kurtosis
- Possible causes of violation of this assumption
  - The true distribution of the variable, e.g., some variables follow a binomial or poisson distribution.
  - Measurement artifacts
  - Inadequate sample

What are the consequences?

- When the sample is extremely small (e.g., below 100), the violation of this assumption leads to inaccurate estimates of confidence intervals and p-values. As the sample gets larger, the central limit theorem suggested that we can get pretty accurate confidence intervals and p-values.

5. Normality (continued)

How to detect the violation:

- Graphic methods: Stem-and-leaf plot, (skeletal) box plot, dot plot, histogram
- Formal statistical tests: look at the statistics for Skewness and Kurtosis.

Solutions:

- Using larger samples
- Using conservative p-values (e.g., using 0.01 rather than 0.05)

Other Considerations

Avoid the multicolinearity among independent variables

- If two independent variables are perfectly correlated with each other, the regression model may not run.
- If two independent variables are only highly, but not perfectly correlated with each other, the regression model may run, but it is hard to interpret the meaning of regression coefficients.
- Use Variance Inflation Factor (VIF). a VIF is greater than 2.5 indicates a possible multicollinearity problem
- Possible solutions:
  - If these independent variables measure the same concept, you can delete one or more of these variables from the model, combine them into an index, or estimate a latent variable model.
  - If these independent variables measure different concepts, you need to (1) get a larger sample, hoping to increate the sample variability and then reduce multicolinearity of the independent variables or (2) stratify the sample and hope that the multicolinearity problem may disappear.

Other Consideration (cont.)

The influence of outliers

- Outliers could significantly the regression coefficients and their standard errors
- Look at the distribution of each variable
- Standardized the scores of each variable. Any score greater than 2.5 is a candidate for an outlier.
- Formal tests: Cook's D, DFITS, DFBETAS
- Solutions:
  - Run regression with or without outliers. If no differences were found, there is no need to do anything. If differences were found, you need to consider remove of a outlier or outliers.
  - Use Formal tests to identify influential cases and decided whether you want to keep them or not.

Conclusions

- When using regression analysis, researchers should be aware of whether its five assumptions have been met.
  - The violations of linearity and having mean independence created biased regression coefficients
  - the violations of homoscedasticity and uncorrelated disturbances lead to inaccurate confidence intervals and p-values for regression coefficients.
  - The violation of normality of the disturbance term is less severe than the other four assumptions, especially if you have large samples.
- Multicolinearity of independent variables and outliers are two additional issues needed to be considered in conducting regression analysis.
- Understanding the assumptions of regression analysis helps better understand other techniques, such as Logistic Regression, Structural Equation Models or Hierarchical Linear Model.
- If you have any questions about running regression analysis, CFDR provides programming support. Please feel free to contact Hsueh-Sheng Wu @ 372-3119 or wuh@bgsu.edu.