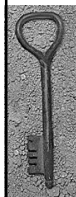




Linear Regression: Assumptions and Issues



Review: Bivariate regression

- ◆ Regression coefficient formulas:

$$Y = a + bX + e$$

$$b = \frac{s_{YX}}{s_X^2}$$

$$a = \bar{Y} - b\bar{X}$$

- Q: What is the interpretation of a regression slope, intercept?



Review: R-Square

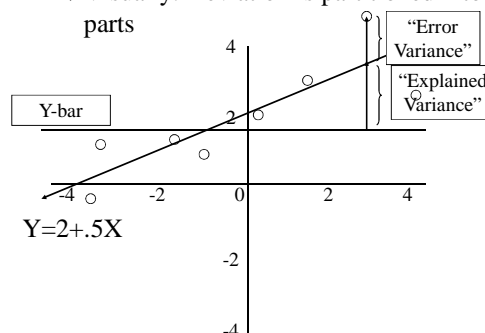
- ◆ The R-Square statistic indicates how well the regression line “explains” variation in Y
- ◆ It is based on partitioning variance into:
 - 1. Explained (“regression”) variance
 - The portion of deviation from Y-bar accounted for by the regression line
 - 2. Unexplained (“error”) variance
 - The portion of deviation from Y-bar that is “error”

- ◆ Formula:

$$R_{YX}^2 = \frac{SS_{REGRESSION}}{SS_{TOTAL}} = \frac{s_{YX}^2}{s_X^2 s_Y^2}$$

Review: R-Square

- ◆ Visually: Deviation is partitioned into two parts



Review: Correlation Coefficient

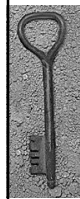
- ◆ R-square = the square of the r
- ◆ r is a measure of linear association
- ◆ r ranges from –1 to 1
 - 0= no linear association
 - 1 = perfect positive linear association
 - -1 = perfect negative linear association
- ◆ R-square: ranges from 0 to 1



Review: Multivariate Regression

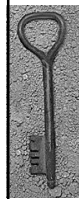
$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k + e$$

- ◆ b_i , “partial slopes”: the average change in Y associated with one unit change in X_i , when the other independent variables are held constant
- ◆ R-square: share of variation in Y explained by all independent variables
- ◆ Standardized coefficients allow us to compare the relative importance of variables
- ◆ Dummy variables
- ◆ Interactions between variables



Review: Model Selection

- ◆ 1) Look for increase in Adjusted R-Square
- ◆ 2) Conduct a F-test of two R-square
- ◆ 3) Automatic model selection
 - Backward, forward, stepwise
- ◆ Use theories to guide your model building



Regression Assumptions

- ◆ 1. Large, random sample
 - For more independent variables, larger N is needed
- ◆ 2. No measurement error
 - All variables are accurately measured
 - Unfortunately, error is common in measures
 - Survey questions can be biased
 - People give erroneous responses (or lie)
 - Aggregate statistics (e.g., GDP) can be inaccurate
 - This assumption is often violated to some extent
 - We do the best we can:
 - Design surveys well, use best available data
 - There are advanced methods for dealing with measurement error



Regression Assumptions

- ◆ 1. Large, random sample
- ◆ 2. No measurement error
- ◆ 3. No specification error
 - Specification error = “wrong model”
 - 1. Function form: linear, additive relationship
 - 2. Variables: no relevant independent variables are excluded; no irrelevant variables are included

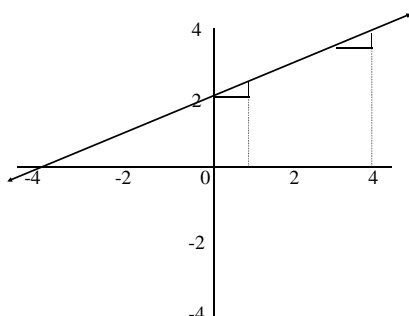


Assumptions: Specification Errors

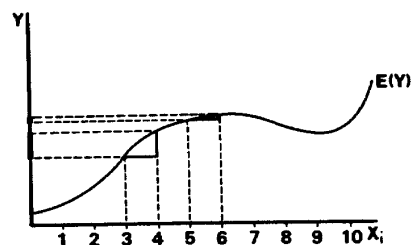
- ◆ 1. Function form: Linearity, additivity
 - Linearity: the change in Y associated with a unit change in X_1 is the same regardless of the level of X_1 .

Linearity

- ◆ Change in Y is the same for X at all levels



Nonlinearity



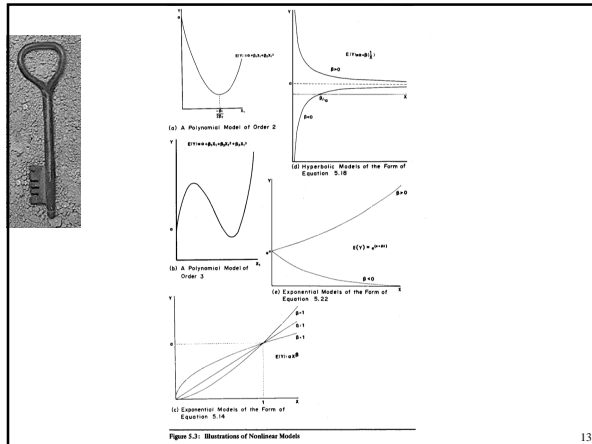


Figure 5.3: Illustrations of Nonlinear Models

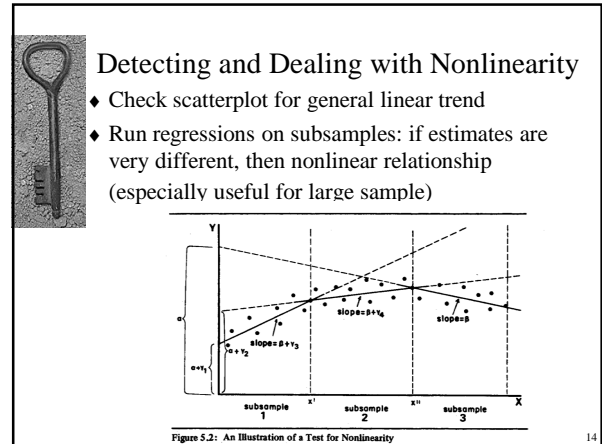


Figure 5.2: An Illustration of a Test for Nonlinearity

Detecting and Dealing with Nonlinearity

- ◆ Check scatterplot for general linear trend
- ◆ Run regressions on subsamples
- ◆ Apply nonlinear models:
 - Polynomial model: $Y = a + b_1X_1 + b_2X_1^2 + b_3X_1^3 + e$
 - Exponential model: $Y = aX^be$
- Often can be converted to linear models
 - Polynomial model: $X_2=X_1^2, X_3=X_1^3$
 - Exponential model: Log transformation: $\text{Log}(Y) = \text{Log}(a) + b\text{Log}(X) + \text{Log}(e)$

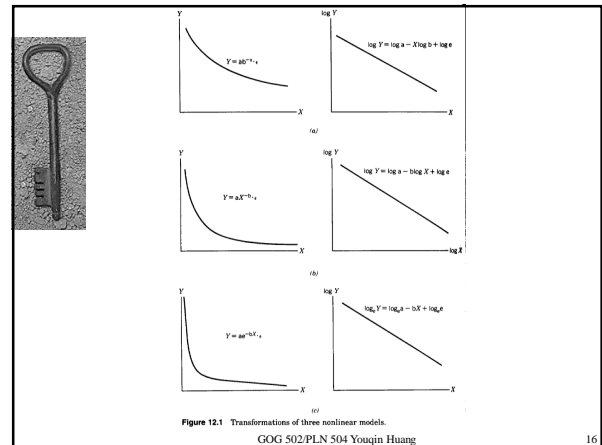
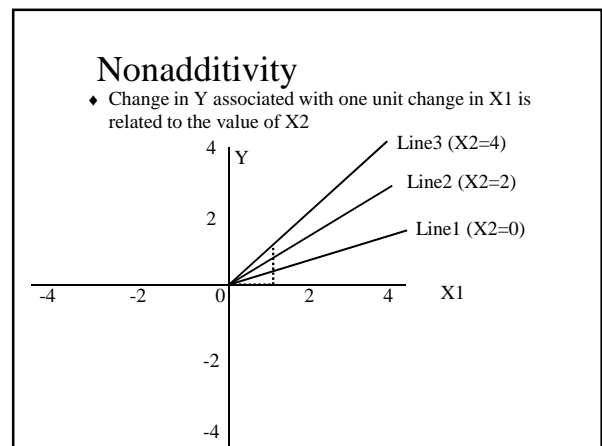


Figure 12.1 Transformations of three nonlinear models.

Assumptions: Specification Errors

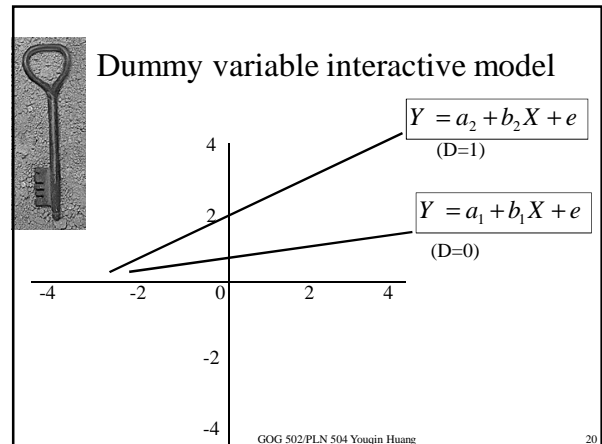
- ◆ 1. Function form: Linearity, additivity
 - Linearity: the change in Y associated with a unit change in X_1 is the same regardless of the level of X_1 .
 - Additivity: the amount of change in Y associated with a unit change in X_1 is the same, regardless of values of the other X_s in the model



Dealing With Nonadditivity

- ◆ Dummy variable interactive model:
 - When D=0: $Y = a_1 + b_1X + e$
 - When D=1: $Y = a_2 + b_2X + e$
 - OR: $Y = a + b_1X + b_2D + b_3X * D + e$
 - Example: urban vs. rural; male vs. female
 - Different intercepts, different slopes

GOG 502/PLN 504 Youqin Huang 19



Dealing With Nonadditivity

- ◆ Dummy variable interactive model:
 - When D=0: $Y = a_1 + b_1X + e$
 - When D=1: $Y = a_2 + b_2X + e$
 - OR: $Y = a + b_1X + b_2D + b_3X * D + e$
 - Example: urban vs. rural; male vs. female
- ◆ Multiplicative model: $Y = a + b_1X_1 + b_2X_2 + b_3X_1 * X_2 + e$
- ◆ Nonlinear interactive model: $Y = aX_1^{b_1} X_2^{b_2} e$

GOG 502/PLN 504 Youqin Huang 21

Assumptions: Specification Errors

- ◆ 1) Correct function form
- ◆ 2) Correct variables: no relevant independent variables are excluded; no irrelevant variables are included
- ◆ Leave relevant variables out:
 - True model: $Y = a + b_1X_1 + b_2X_2 + e$
 - You specify: $Y = a + b_1X_1 + e$
- ◆ If X_1 and X_2 are correlated:
 - X_1 is correlated with the error term
 - $e = b_2X_2 + e$; OLS estimate will be biased
 - b1 will be biased: includes effect of X_2
- ◆ If X_1 and X_2 are uncorrelated:
 - b1 estimate is unaffected
 - Standard error for X_1 will be smaller, more likely to be significant

$$S_{b1} = \sqrt{\frac{\sum(Y_j - \hat{Y}_j)^2 / (n-3)}{\sum(X_{1j} - \bar{X}_1)^2 (1 - r_{X_1, X_2}^2)}}$$

GOG 502/PLN 504 Youqin Huang 22

Assumptions: Specification Errors

- ◆ Including irreverent variables
 - True model: $Y = a + b_1X_1 + e$
 - You specify: $Y = a + b_1X_1 + b_2X_2 + e$
- ◆ If X_1 and X_2 are uncorrelated:
 - b2 is close to zero, will not be significant
 - Estimation for b1 is unbiased
- ◆ If X_1 and X_2 are correlated:
 - Estimation for b1 is not biased
 - But with larger standard errors, inefficient estimation

$$S_{b1} = \sqrt{\frac{\sum(Y_j - \hat{Y}_j)^2 / (n-3)}{\sum(X_{1j} - \bar{X}_1)^2 (1 - r_{X_1, X_2}^2)}}$$

GOG 502/PLN 504 Youqin Huang 23

Regression Assumptions

- ◆ 1. Large, random sample
- ◆ 2. No measurement error
- ◆ 3. No specification error
 - Model specification is difficult – it is hard to be certain that all relevant variables are included
 - Use theory and previous research as a guide
 - Don't leave irrelevant variables in the model
 - A low R-square is a hint: much of the variation in Y has not been explained

GOG 502/PLN 504 Youqin Huang 24



Regression Assumptions

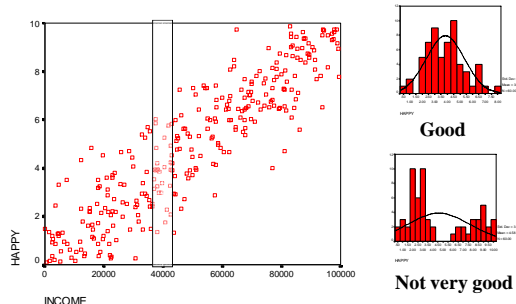
- ◆ 1. Large, random sample
- ◆ 2. No measurement error
- ◆ 3. No specification error
- ◆ 4. Normality:
 - Y_i is normally distributed for every outcome of X in the population -- “conditional normality”
 - Ex: happy (Y) vs. income (X)
 - Suppose we look only at a sub-sample: $X = 40,000$
 - Is a histogram of happy approximately normal?
 - What about for people with $X = 60,000, 100,000$?
 - If **all** are roughly normal, the assumption is met

GOG 502/PLN 504 Youqin Huang

25

Regression Assumptions: Normality

Examine sub-samples at different values of X .
Make histograms and check for normality.



Regression Assumptions

- ◆ 1. Large, random sample
- ◆ 2. No measurement error
- ◆ 3. No specification error
- ◆ 4. Normality:
 - Y_i is normally distributed for every outcome of X in the population, also called “conditional normality”
 - Error (e) is normally distributed with expected value of zero
 - Errors shouldn't be systematically positive or negative
 - Error is uncorrelated with predictors in the equation (X_i 's)

GOG 502/PLN 504 Youqin Huang

27

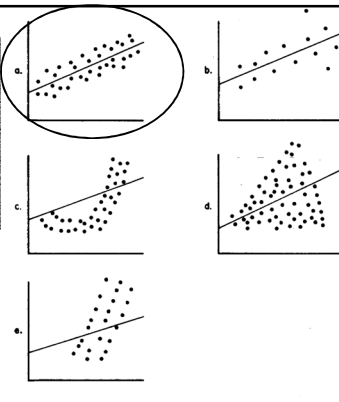


Figure 8a-e: Some Possible Patterns for Residuals

GOG 502/PLN 504 Youqin Huang

28



Regression Assumptions

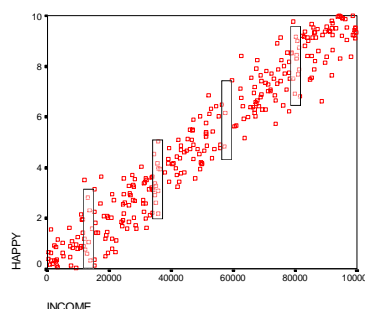
- ◆ 5. “Homoskedasticity:”
 - The variances of errors are identical at different values of X
 - Versus “heteroskedasticity”, where errors vary with X

GOG 502/PLN 504 Youqin Huang

29

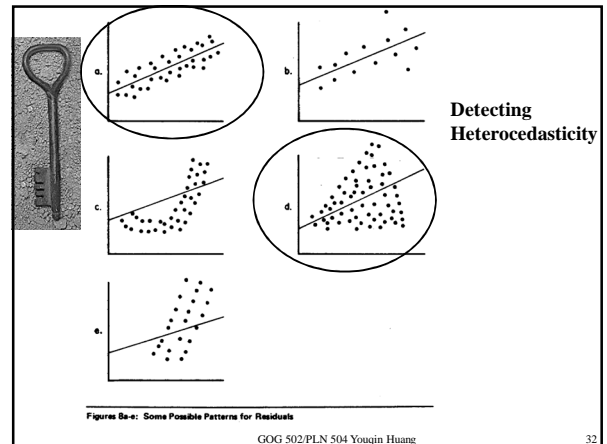
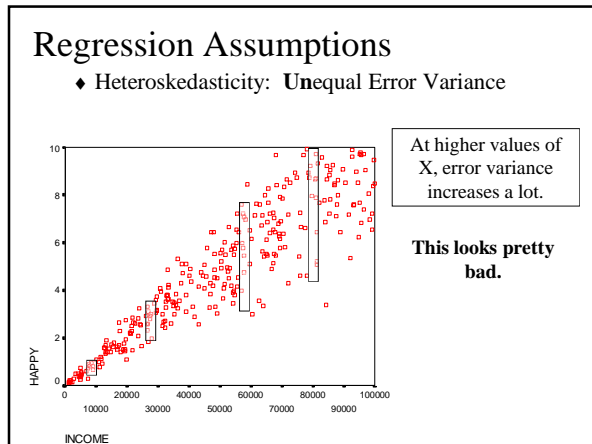
Regression Assumptions

- ◆ Homoskedasticity: Equal Error Variance



Examine error at different values of X .
Is it roughly equal?

Here, things look pretty good.



Regression Assumptions

- ◆ “Heteroskedasticity”
 - Estimation is unbiased, but not efficient
 - A result of interaction between X and other variable not in the model → appropriate model specification
 - Generalized Least Squares (GLS) regression
 - Can yield BLUE estimators when heteroskedasticity is present
 - OLS: minimize SSE $\sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N e_i^2$
 - vs. GLS: minimized a weighted SSE $\sum_{i=1}^N \frac{1}{s^2_{e_i}} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N \frac{e_i^2}{s^2_{e_i}}$
 - Observations with larger errors are given a smaller weight

GOG 502/PLN 504 Youqin Huang 33

Regression Assumptions

- ◆ 1. Large, random sample
- ◆ 2. No measurement error
- ◆ 3. No specification error
- ◆ 4. Normality
- ◆ 5. Homoskedasticity
- ◆ 6. No autocorrelation
 - The errors for different values of X are not correlated
 - It is common for variables to be characterized by correlations between adjacent values in space and time
 - Two contexts, two subfields of statistical analysis :
 - Serial correlation: time-series data, e.g. GNP each year
 - Spatial autocorrelation: spatial data, spatial analysis
 - The first law of geography: things closer to each other are more similar

GOG 502/PLN 504 Youqin Huang 34

Regression Assumptions

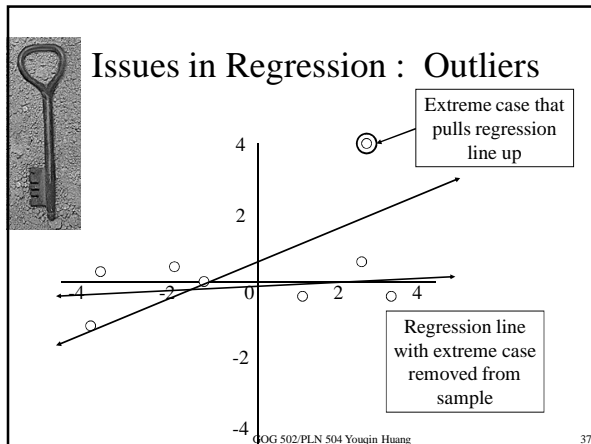
- ◆ Usually, not all assumptions are met perfectly
- ◆ Substantial departure from assumptions means you must qualify your conclusions
- ◆ Overall, regression is robust to violations of assumptions
 - It often gives fairly reasonable results, even when assumptions aren't perfectly met
- ◆ Various modifications of regression can handle situations where assumptions aren't met
- ◆ But, there are also further diagnostics to help ensure that results are meaningful...
 - e.g., dealing with outliers that may affect results

GOG 502/PLN 504 Youqin Huang 35

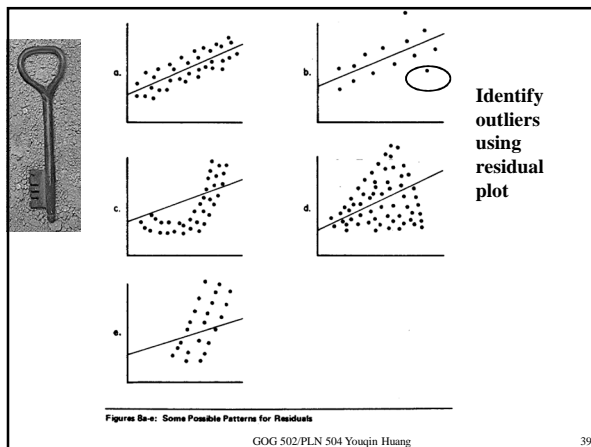
Issues in Regression #1: Outliers

- ◆ Even if regression assumptions are met, slope estimates can have problems
- ◆ Example: Outliers
 - Errors in coding or data entry
 - Highly unusual cases
 - Or, sometimes they reflect important “real” variation
- ◆ Even a few outliers can dramatically change estimates of the slope (b)

GOG 502/PLN 504 Youqin Huang 36



- ### Strategy for Dealing with Outliers
- ◆ 1. Identify them
 - Look at scatterplots for extreme values
 - Compute diagnostic statistics to identify outliers (descriptive statistics, residual plot)
- GOG 502/PLN 504 Youqin Huang 38



- ### Strategy for Dealing with Outliers
- ◆ 1. Identify them
 - ◆ 2. Depending on the circumstances:
 - A) Drop cases from sample and re-do regression
 - Especially for coding errors, very extreme outliers
 - Or if there is a theoretical reason to drop cases
 - Lose information, smaller sample
 - B) Keep the outliers if there is no good reason to drop them. It is a judgment call.
 - C) Report two regressions, with and without outliers
 - Have to explain two sets of results, may be inconsistent
 - D) Transform the variable
 - Interpretation is less straightforward
- GOG 502/PLN 504 Youqin Huang 40

Issues #2: Multicollinearity

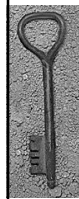
- ◆ High correlation between independent variables
- ◆ Effects on coefficients and standard error

$$b_1 = \left(\frac{s_Y}{s_{X_1}} \right) \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{1 - r_{X_1X_2}^2}$$

$$\hat{\sigma}_{b_1} = \frac{1}{\sqrt{1 - r_{X_1X_2}^2}} \frac{\hat{\sigma}}{\sqrt{\sum (X_1 - X_2)^2}}$$

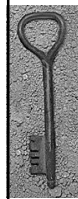
GOG 502/PLN 504 Youqin Huang 41

- ### Issues #2: Multicollinearity
- ◆ High correlation between independent variables
 - ◆ Effects on coefficients and standard error
 - Inflate coefficients and s.e.
 - ◆ Detecting multicollinearity:
 - Coefficients of existing variables change significantly with the addition of a new variable
 - Correlation matrix (rule of thumb: $|r| > 0.8$)
- GOG 502/PLN 504 Youqin Huang 42



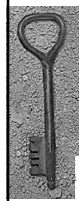
Issues: Multicollinearity

- ◆ Strategies:
 - Remove variables: if X1 and X2 are highly correlated, keep only one of them
 - Create a summary index: several highly correlated indicators measuring a common feature.
 - Socioeconomic status: a indicator summarizing the joint effect of education, income, occupation
 - Factor analysis



Issues #3: Data Aggregation

- ◆ Multiple levels of analysis
- ◆ It is incorrect to assume that relationships existing at one level of analysis will necessarily demonstrate the same strength at another level
- ◆ Three types of erroneous inferences:
 - Individualistic fallacy: impute macrolevel relationships from microlevel relationships
 - Cross-level fallacies: make inferences from one subpopulation to another at the same level of analysis
 - Ecological fallacy: make inferences from higher to lower levels of analysis
- ◆ Aggregation reduces variation, thus increases r

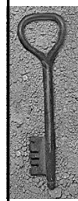


Issues: Data Aggregation

- ◆ Income=a + b*education
- ◆ A survey of 952 households in LA
- ◆ Also collected information at tract level and two governmental groupings.

TABLE 11.9 Correlation and Slope Coefficients Derived from Household and Aggregated Data

Data Set	Method of Data Generalization	r	r ²	b
952 units: households	Not applicable	.4028	.1623	857.60
1556 units: census tracts	Tract mean	.6434	.4140	2413.64
134 units: Welfare Planning Council (groups of tracts)	Group mean	.7606	.5785	2808.21
35 units: Regional Planning Commission (groups of tracts)	Group mean	.8503	.7230	3103.62



Issues #4: Missing Data

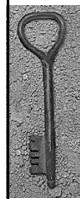
- ◆ Replace missing value with mean
- ◆ Exclude case listwise
- ◆ Exclude case pairwise
- ◆ If missing is coded “-9”, “-99”, be careful when conducting your analysis

The image shows the SPSS software interface. The main window displays a data view with columns for 'Race', 'White', 'Hispanic', 'Asian', 'Other', 'Income', and 'Education'. A dialog box titled 'Linear Regression: Options' is open, showing settings for missing data handling. The 'Missing' section is checked, and the 'Use probability of F' option is selected. The 'Exclude cases listwise' option is also checked. The 'Display' section is set to 'None'.



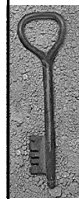
Issues #5: Models and “Causality”

- ◆ People often use statistics to support theories or claims regarding causality
 - They hope to “explain” some phenomena
 - What factors make kids drop out of school
 - Whether or not discrimination leads to wage differences
 - What factors make corporations earn higher profits
- ◆ Statistics provide information about association
- ◆ Always remember: Association (e.g., correlation) is not causation!
 - Association can be spurious



Issues #5: Models and “Causality”

- ◆ Multivariate models can estimate “partial” relationships
 - i.e., associations **controlling for** other variables
 - We can assess each variable’s correlation over and above other variables
- ◆ Multivariate variables provide some capacity to identify “spurious” relationships
 - Often, spurious correlations disappear once other variables are introduced into a multivariate model



Issues #5: Models and “Causality”

- ◆ Question: If we “control for” every possible spurious relationship, can we identify true causal relationships among variables?
 - Can we conclude: “poverty causes crime”?
- ◆ Answer: No, not really
 - 1. First of all, we can never include **all possible** relevant variables into a single model
 - 2. Often, causality can run in the opposite direction



Issues #5: Models and “Causality”

- ◆ However: Carefully executed multivariate analyses are one of the best ways to provide support for arguments and theories
 - Even though they do not necessarily “prove causality”
- ◆ Good models require (at a minimum):
 1. Unbiased samples
 2. Careful measurement of phenomena
 3. Careful application of statistical methods
 - Assumptions met, relevant control variables included, etc
 4. Acknowledgement of limitations of data/methods
 - Only then can we start drawing tentative conclusions!



Models and Causality: Advice

- ◆ 1. Stay “close” to your data
 - Always spend a lot of time looking at raw data, simple descriptive statistics
 - You’ll catch errors and get a sense of relationships among variables
- ◆ 2. Learn to develop multivariate models
 - Explore different variables
 - Learn how control variables work
 - Learn to tell when your model is blowing up
 - Do common-sense reality checks
- ◆ 3. Don’t over-interpret! Be humble, cautious



Summary

- ◆ Regression assumptions
 1. Large, random sample
 2. No measurement error
 3. No specification error
 4. Normality
 5. Homoskedasticity
 6. No autocorrelation
- ◆ Issues:
 - Outliers
 - Multicollinearity
 - Aggregation
 - Missing values
 - Association vs. causality