SOS3003 **Applied data analysis for social science** Lecture note 11-2010 Some last words

Erling Berge
Department of sociology and political science
NTNU

Term paper

- At the beginning of the class we were talking about the requirements of the paper and a bit about how to write it
- I hope you have studied it carefully:

http://www.sv.ntnu.no/iss/Erling.Berge/2010%20SOS3003%20SemOppgKravEN201001.pdf

- **Deadline for paper: 10 May**
- **Delivery by e-mail to <ISSInnlevering@svt.ntnu.no>**

Term paper

- The term paper shall be an independent work demonstrating how multiple regression can be used to analyze a social science problem. The paper should be written as a journal article, but with more detailed documentation of data and analysis, for example by means of appendices.
- Based on information about the dependent variable a short theoretical discussion of possible causal mechanisms explaining some of the variation in the dependent variable is presented. This leads up to a model formulation and operationalisation of possible causal variables taken from the data set. If missing data on one or more variables causes one or more cases to be dropped from the analysis, the selection problem must be discussed.
- By means of multiple regression (OLS or Logistic) the model should be estimated and the results discussed in relation to the initial theoretical discussion
- More details will be available in a separate paper

Step 1: Dependent variable

- Investigate the distribution of cases on the dependent variable
  - Think about what mechanisms may generate high or low variable values for particular cases
  - Make a list of such mechanisms
  - Can you find information on these mechanisms in the data?
  - Make suitable variables of those you find

Step 2 Types of research problems

- The dependent variable will usually be either an
  - Indicator of obtained status of some kind (education level, marriage status)
  - Indicator of activity of some kind (work, industry, leisure activity, voting behaviour)
  - Indicator of strength of attitude or belief of some kind (political preferences, trust, type of entertainment)
- The problem of modelling the variation is different for the different types of variables
  - They will have different causal structures

Step 3 Types of causal mechanisms I

- Structural causation
  - Social structure does have causal impacts that are not well understood. In a framework of methodological individualism one may say that it limits and orders the options that actors can choose from. Hence, variables such as age, sex, and place of living can be used as proxies for poorly understood causal factors.
  - Budget constraints (time and income constraints) have the same character. They limit and orders the options that actors can choose from. However, they enter the model more through the way the dependent variable is constructed, and the kind of link function (linear or logistic) used to mediate between observations and dependent variable.

Step 3 Types of causal mechanisms II

- Individual causation
  - Preferences (norms, values, attitudes) may be difficult to observe in detail but are assumed to be present
  - Resources (income/ capital, education/ human capital, access to networks/ social capital) are usually measured extensively even if unevenly. Here there are budget constraints
  - Perception of opportunities will often depend on position in social structure
  - Beliefs about resources and opportunities are important. They may be based on both fact and fiction

Step 4 Explanatory variables

- All kinds of explanatory variables are allowed
- Make a list of conceivable variables
- Look for direct or indirect indicators for the variables. Approximations are allowed
- Construct new variables where variation and codes match as well as possible the intended indicator
- Then the first model can be estimated

Elaborating the model

1. Is the distribution of the residual normal?
2. If no:
   i. Curvilinearity? If yes, fix the problem and go back to 1. Else
   ii. Missing variables? (correlates with both y- and x-variables)
   iii. Heteroscedasticity? If yes, fix the problem and go back to 1. (Fixing this may entail transformation to symmetry.)
   iv. If tests are trustworthy remove obviously irrelevant variables and go back to 1
3. If yes: you have a first estimate of your model
4. Consider how it may be improved!

Basic sources of error

- Errors in theory / model
  - Model specification: valid conclusions require a correct (true) model
- Errors in the sample
  - Selection bias
- Measurement problems
  - Missing cases and measurement errors
  - Validity og reliability
- Multiple comparisons
  - Conclusions are valid only for the sample

Serious errors from the term papers of last fall

- Lack of understanding of variables and measurement scales
  - Relation to measurement units
  - Relation to correlations among variables
  - Relation to dummy coding
- Lack of understanding of measurement units
  - Relation to interpretation of results

Test for Curvilinear Relations

- Testing for curvilinearity in age
  - Set age squared = age2
- Remember:
  - Age is one substance variable that may be represented either by one technical variable or by two technical variables (somewhat like one variable being represented by different ways of coding)
    - Substance variable Age is represented by age
    - Substance variable Age is represented by age + age2

Testing

- Model 0
  - (some variables)
- Model 1
  - (some variables) + age
- Model 2
  - (some variables) + age + age2
- In model 1 the impact of Age is tested by the t-test and the corresponding p-value (there is no difference between the substance variable and its technical representation)

Testing 2

- In model 1 the test may conclude that Age does not contribute to the model. If so we go to model 2
- In model 2 the testing of the impact of the substance variable Age (represented by age and age2) is done by an F-test of Model 2 against Model 0
- The F-test may conclude that Age does not contribute to the model. Then we drop both age and age2.
- The F-test may conclude that Age (represented by age and age2) contributes significantly to the model. Then we keep both age and age2

Testing 3

- In model 1 the test may conclude that Age does contribute to the model. If so we may still go to Model 2
- If either the t-test of model 1, or the F-test of model 2, or both show that Age contributes significantly to the model, there are several possibilities
  - T-test significant, F-test not significant: drop age2, keep age
  - T-test significant, F-test significant, p-value of age is unchanged or higher (compared to model 1) while p-value of age2 is clearly insignificant: drop age2, keep age

Testing 4

- (continued)
  - T-test significant, F-test significant, p-value of age improves (compared to model 1): keep age2 no matter what p-value for age2 is
  - T-test significant, F-test significant, p-value of age shows no significance (compared to model 1) while p-value of age2 shows clear significance: keep age2 no matter what p-value for age is
  - T-test significant, F-test significant, p-value of both age and age2 show no significance but are fairly close. Then the F-test decides. Keep age2.
- And remember: age2 never appears alone, always with age

Sample size in logistic regression

Literature cited:

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. London: Sage.
- Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 49 (12):1373-1379.

Large sample properties

- The good properties of ML estimates of binary logistic regression models are large sample properties that obtain as sample size goes towards infinity. Hence
- A sample needs to be “large enough”
- What “large enough” means is not clear
- What happens when you have too small a sample is largely unknown
- Long (1997) puts 100 cases as an absolute lower bound

Calculation of lower bounds

- A lower bound of 100 must be adjust according to number of variables in the model and the distribution of cases on the dependent variable.
- Peduzzi et al. (1996) suggest:
- Let p be the smallest of the proportions of negative or positive cases in the population and k the number of covariates (the number of independent variables), then the minimum number of cases to include is:
- N = 10 k / p
- If the resulting number is less than 100 you should increase it to 100
- Or you may say that the maximum number of variables you can include in the model will be
- k = N*p/10

Causal analysis I

- Experiment
  - Randomized causal impacts (”treatment”) provide precise causal conclusions about effects (”response”) if there is significant differences in the mean response (effect)
  - Experiments can be impossible to achieve due to
    - Practical conditions
    - Economic constraints
    - Ethical judgements
- Instead one tries to obtain quasi-experiments
  - Using for example regression analysis

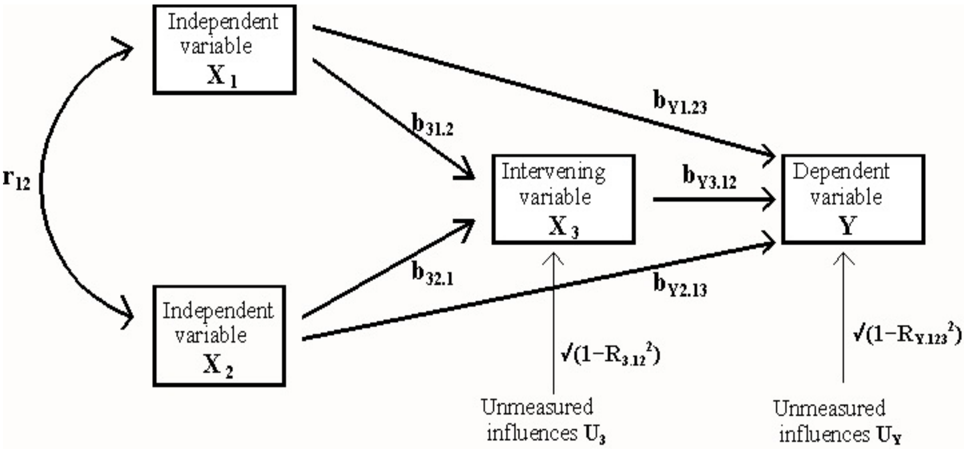
Causal modelling II

- “path analysis” or “structural equations modelling” go back to the 60ies
- Jöerskog and Sörbom: LISREL
  - Use maximum likelihood to estimate model parameters maximising fit to the variance-covariance matrix
  - Commonly available in statistical packages
    - Covariance structural modelling
    - Structural equation modelling
    - Full information maximum likelihood estimation

Low-Tech approach

- Uses OLS to do simple versions of the structural equations models
- The key assumption is the causal ordering of variables. In survey data this ordering is supplied by theory
- The causal diagram visualize the order of causation:
  - Causality flows from left to right
  - Intervening variables give rise to indirect effects
  - “reverse causation” creates problems

Path coefficients
Figure 3


Independent variable X_1
Independent variable X_2
r_{12}
$b_{31.2}$
$b_{32.1}$
Intervening variable X_3
$b_{Y1.23}$
$b_{Y3.12}$
$b_{Y2.13}$
Dependent variable Y
$\sqrt{(1-R_{3.12}^2)}$
$\sqrt{(1-R_{Y.123}^2)}$
Unmeasured influences U_3
Unmeasured influences U_Y

Some elements in figure 3

$b_{31.2}$, $b_{32.1}$	Standardized regression coefficients (“beta weight”) from the regression of X_3 on X_1 controlled for X_2 and from the regression of X_3 on X_2 controlled for X_1
$R_{3.12}^2$	Coefficient of determination (R^2) from the regression of X_3 on X_1 and X_2
$\sqrt{\{1-R_{3.12}^2\}}$	The error term from the regression of X_3 on X_1 and X_2

The structural model of figure 3

- $\hat{Y} = b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}X_3$
- $\hat{X}_3 = b_{31.2}X_1 + b_{32.1}X_2$

- In structural equations variables and coefficients are standardized
- That means that variables have an average of 0 and a standard deviation of 1 and that coefficients vary between -1 and +1

Direct, Indirect and Total Effects

- *Direct effects* are the path coefficients linking two variables without any intervening variable
- *Indirect effects* equal the product of coefficients along any series of causal paths that link one variable to another
- *Total effects* equal the sum of all direct and indirect effects linking two variables

Indirect effects as products of path coefficients

- $\hat{Y} = b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}X_3$
- $\hat{X}_3 = b_{31.2}X_1 + b_{32.1}X_2$
- Means that we have
- $\hat{Y} = b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}X_3$
- $= b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}(b_{31.2}X_1 + b_{32.1}X_2)$
- $= b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}b_{31.2}X_1 + b_{Y3.12}b_{32.1}X_2$
- $= (b_{Y1.23} + b_{Y3.12}b_{31.2})X_1 + (b_{Y2.13} + b_{Y3.12}b_{32.1})X_2$

- Compare compound coefficients to the diagram

Path Coefficients = Direct effects

X_1 to Y: $\mathbf{b_{Y1.23}}$	standardized regression coefficient of Y on X1, controlling for X2 and X3
X_2 to Y: $\mathbf{b_{Y2.13}}$	standardized regression coefficient of Y on X2, controlling for X1 and X3
X_3 to Y: $\mathbf{b_{Y3.12}}$	standardized regression coefficient of Y on X3, controlling for X1 and X2
X_1 to X_3: $\mathbf{b_{31.2}}$	standardized regression coefficient of X3 on X1, controlling for X2
X_2 to X_3: $\mathbf{b_{32.1}}$	standardized regression coefficient of X3 on X2, controlling for X1

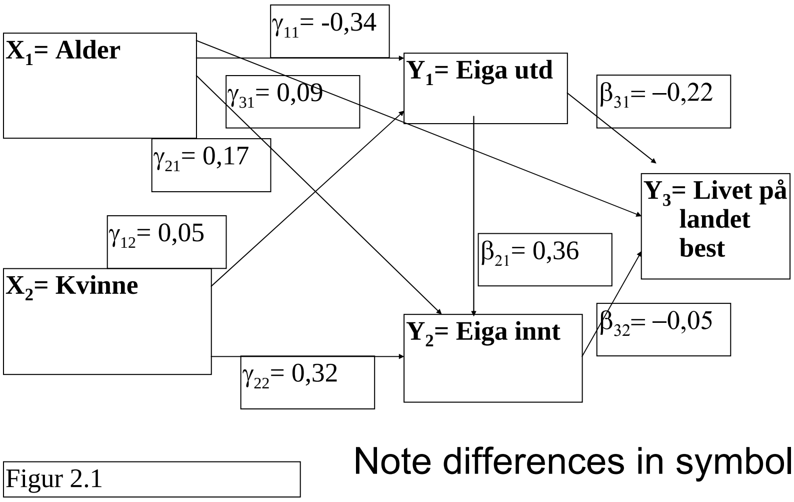
Indirect and total effects

Indirect effects	
X_1 to Y, through X_3:	$\mathbf{b_{31.2} \times b_{Y3.12}}$
X_2 to Y, through X_3:	$\mathbf{b_{32.1} \times b_{Y3.12}}$
Total effects	
X_1 to Y:	$\mathbf{b_{Y1.23} + (b_{31.2} \times b_{Y3.12})}$
X_2 to Y:	$\mathbf{b_{Y2.13} + (b_{32.1} \times b_{Y3.12})}$

Adding to multiple regressions

- We learn something new if the indirect effects are large enough to have substantial interest
- More than two steps of causation tends to become very weak
  - 0.3*0.3*0.3 = 0.027
  - 0.3 standard deviation change in causal variables leads to a 0.027 standard deviation change in the dependent variable

Example of a path diagram



Note differences in symbols

Comment to the figure above

- The β coefficients go from one Y variable to another
- The γ coefficients go from one X variable a Y variable
- The coefficient indexing indicates which variables they link. The first index tells the dependent variable. The second index tells the independent variable
- The coefficients are standardized (OLS) regression coefficients (“beta weights”)

The structural model of the example

- $\hat{Y}_3 = \gamma_{31}X_1 + \gamma_{32}X_2 + \beta_{31}Y_1 + \beta_{32}Y_2$
- $\hat{Y}_2 = \gamma_{21}X_1 + \gamma_{22}X_2 + \beta_{21}Y_1$
- $\hat{Y}_1 = \gamma_{11}X_1 + \gamma_{12}X_2$
- $\hat{Y}_3 = 0.09X_1 - 0.22Y_1 - 0.05Y_2$
- $\hat{Y}_2 = 0.17X_1 + 0.32X_2 + 0.36Y_1$
- $\hat{Y}_1 = -0.34X_1 + 0.17X_2$

Direct and indirect effects on “Livet på landet best” from age

- Direct effect: $\gamma_{31} = 0.09$
- Indirect effect by way of “Eiga utd” and “Eiga innt”
- $\beta_{31} * \gamma_{11} + \beta_{32} * \beta_{21} * \gamma_{11} + \beta_{32} * \gamma_{21}$
- $(-0.22)*(-0.34)+(-0.05)*(0.36)*(-0.34)+(-0.05)*(0.17)$
- $0.22*0.34 + 0.05*0.36*0.34 - 0.05*0.17$
- 0.0748 + 0.00612 – 0.0085 = 0.07242
- Total effect = 0.09 + 0.07242 = 0.16242
- Increasing age by 1 st. dev. leads to an increase of 0.16 st.dev. in the strength of support for “Livet på landet best”

Variables and measurement in structural models

- All interval scale variables used in multiple regression (including non-linear transformed variables and interaction terms) can be included in structural equations models
- But interpretation becomes tricky when variables are complex. Conditional effect plots are very useful
- Robust, quantile, logit, and probit regression should not be used
- Categorical variables should not be used as intervening variables
- Scales or index variables can be used as usual in OLS regression

Concluding on structural equations modelling

- Including factors from factor analysis as explanatory variables make it possible to approximate a LISREL type analysis
- If assumptions are true LISREL will perform a much better and provides more comprehensive estimation, but too often assumptions are not true. Then the low-tech approach has access to the large toolkit of OLS regression for diagnostics and exploratory methods testing basic assumptions and discovering unusual data points
- Simple diagnostic work sometimes yields the most unexpected, interesting and replicable findings from our research

Principal components and factor analysis

- Principal components and factor analysis are both methods for data reduction
- They seek underlying dimensions that are able to account for the pattern of variation among a set of observed variables
- Principal components analysis is a transformation of the observed data where the idea is to explain as much as possible of the observed variation with a minimum number of components

Factor analysis

- Estimates coefficients on - and variable values of - unobserved variables (Factors) to explain the co-variation among an observed set of variables
- The assumption is that a small set of the unobserved factors are able to explain most of the co-variation
- Hence factor analysis can be used for data reduction. Many variables can be replaced by a few factors

Factor analysis

- $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \dots + \ell_{kj}F_j + \dots + \ell_{kJ}F_J + u_k$
  - $k = 1, 2, 3, \dots, K$
- Symbols
  - K observed variables, Z_k ; k=1, 2, 3, … , K
  - J unobserved factors, F_j ; j=1, 2, 3, … , J where J<K
  - For each variable there is a unique error term, u_k, also called unique factors while the F factors are called common factors
  - For each factor there is a **standardized** regression coefficient, ℓ_{kj}, also called factor loading; k refers to variable no, j refers to factor no. An index denoting case no has been omitted here.

Correlation of factors

- Factors my be correlated or uncorrelated
  - Uncorrelated: they are then called **orthogonal**
  - Correlated: they are then called **oblique**
- Factors may be rotated
  - Oblique rotations create correlated factors
  - Orthogonal rotations create uncorrelated factors

Principal components

- Represents a simple transformation of variables. There are as many principal components as there are variables
- Principal components are uncorrelated
- $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \dots + \ell_{kj}F_j + \dots + \ell_{kK}F_K$
- If the last few principal components explain little variation we can retain J<K components. Thus Principal Components also can be used to reduce data.
- $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \dots + \ell_{kj}F_j + \dots + \ell_{kJ}F_J + v_k$

  where J<K and

  the residual v_k has small variance and consist of the discarded principal components

Principal components vs factor analysis

- Principal components analysis attempts to explain the observed variation of the variables
- Factor analysis attempts to explain their inter-correlations
- Use principal components to generate a composite variable that reproduce the maximum variance of observed variables
- Use factor analysis to model relationships between observed variables and unobserved latent variables and to obtain estimates of latent variable values
- The choice between the two is often blurred, to some degree it is a matter of taste

The number of principal components

- K variables yield K principal components
- If the first few components account for most of the variation, we can concentrate on them and discard the remaining
- The eigenvalues of the standardized correlation matrix provides a guide here
- Components are ranked according to eigenvalues
- A principal component with an eigenvalue λ<1 accounts for less variance than a single variable
- Thus we discard components with eigenvalues below 1
- Another criterion for keeping components is that each component should have substantive meaning

Eigenvalues and explained variance

- In a covariance matrix the sum of eigenvalues equals the sum of variances.
- In a correlation matrix this = K (the number of variables) since each standardized variable has a variance of 1
- Thus the sum of eigenvalues of the principal components
- $\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_K = K$ and
- λ_j / K = proportion of variance explained by component no j

How many factor should we retain?

- In principal component analysis factors with eigenvalues above 1 is recommended
- In principal factor analysis factors with eigenvalues above 0 is recommended
- Procedure:
  - Extract initial factors or components
  - Rotate to simple structure
  - Decide on how many factors to retain
  - Obtain and use scores for the retained factors, ignoring discarded factors

Factor scores

- Both principal components and factor analysis may be used to compute composite scores called factor scores
- Recall that variables and factors are assumed to be related like
  - $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \ldots + \ell_{kj}F_j + \ldots + \ell_{kK}F_K$
- Then it is possible to find values c_{ij} making
  - $\hat{F}_j = c_{1j}Z_1 + c_{2j}Z_2 + \ldots + c_{kj}Z_j + \ldots + c_{Kj}Z_K$
- The coefficients c_{ij} are the factor score coefficients. They come from the regression of the factor F_j on the variables

Rotation to simple structure

- The idea is to transform (rotate) the factors so that the loadings on each components make it easier to interpret the meaning of the component
- If the loading are close either to 1 or -1 on one factor and close to 0 on all others the structure is simpler to interpret: we rotate to "simple structure". The rotated factors fit data equally well but are simpler to interpret
- Rotations may be
  - Orthogonal (rotation method typically: varimax)
  - Oblique (rotation method typically: oblimin, promax)

Why rotate?

- Underlying unobserved dimensions may in theory be seen as correlated
- Allowing correlated factors may provide even simpler structure than uncorrelated factors, thus easier to interpret
- All rotations fit data equally well
- Hence the one chosen depends on a series of choices done by the analyst
- Try different methods to see if results differ

Concluding (1)

- Principal components
  - transformation of the data, not model based. Appropriate if goal is to compactly express most of the variance of k variables. Minor components (perhaps all except the first) may be discarded and viewed as a residual.
- Factor analysis
  - Estimates parameters of a measurement model with latent (unobserved) variables.

Concluding (2)

- Rotation
  - If we retain more than one factor rotation simplifies structure and improves interpretability
    - Orthogonal rotation (varimax) maximum polarization given uncorrelated factors
    - Oblique rotation (oblimin, promax) further polarization by permitting interfactor correlations. The results may be more interpretable and more realistic than uncorrelated factors
- Scores
  - Factor scores can be calculated for use in graphs and further analysis, based on rotated or unrotated factors and principal components