On writing term papers for SOS3003

Erling Berge

Seminar II

- **Requirements**
- **On writing term paper**
- **Choice of dependent variable**

Preliminary

- It is a goal that everybody have different variables
- Sources for the variables might be
  - Your own data
  - Data used in a previous term paper for SOS3003
  - If this is not relevant
  - Variables will be available from the European Social Survey

Purpose

- The perm paper is a part of the formal examination results and will be evaluated. The mark for the paper will have a weight of 0.6 in the final grade of the course.
- 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.

Dependent variable

- Based on theory the variation in an appropriately chosen dependent variable shall be explained.
- The dependent and independent variables can either be taken from a data set prepared for the class or they may be taken from data collected in other ways for example for your own masters thesis.
- Your own data will have to satisfy some minimum requirements securing that a valid multiple regression. Hence, use of your own data has to be approved.

Requirements for a dependent variable

- The variable must vary !!!!
- OLS regression needs interval (or ratio) scale
- Logistic regression require a dichotomy (exactly 2 values)

Formalities

- Title page
- Preface
- Abstract (100-200 words)
- The main body of the paper (10.000 words)
- References
- Appendices
  - Binding: no binding
  - Deadline: November 23

Requirements (1)

- Theory
- Model formulation – operationalisation
  - Descriptive statistics – transformations?
  - One nominal scale dependent variable with 3 or more categories
  - Interaction term
  - Curvilinearity
  - Conditional effect plot
- Missing data and selection problems?

Requirements (2)

- Multicollinearity
- The impact of outliers and influential cases
- The model specification has to be evaluated.
- In OLS regression heteroskedasticity
- In OLS regression autocorrelation
- In OLS regression the distribution of the residual
- In LOGIT regression the problem of discrimination

Advice on the main body (1)

The main body of the paper might be structured like this

1. An introduction of 1/3 - 1 pages stating the research question and describing why the research question is interesting and/ or valuable.
2. A short discussion of theory relevant to the research question (1 – 3 pages)
3. A short summary/mention of previous research relevant to the question (1 – 2 pages).
4. A description of main hypotheses (1 – 2) pages.

Advice on the main body (2)

5. A description of the data set, the dependent variable (in some detail) and the independent variables (in much less detail)(1-2 pages).
6. A description of the analysis results based on the basic beginning model, tests for more complicated effects, eliminated variables, and the final model (7 – 13 pages).
7. A short conclusions section, summarizing the most important findings. This should be about one page, longer than the conclusions statements in the abstract, which should be only 1-3 sentences (in the abstract)

More on variables

- Finding a dependent variable
- Variables and variation
- Measurement theory and measurement level
- Coding and recoding

On finding a dependent variable

- Is the topic the variable speaks to interesting?
- Is there sufficient variation among people on this variable? Make a frequency distribution.
- Find out the number of missing cases. There should not be ”too many missing” (less than 10%?)
- If the variable is unsuitable for OLS maybe it can be recoded to a dichotomy for use in a logistic regression

Scales

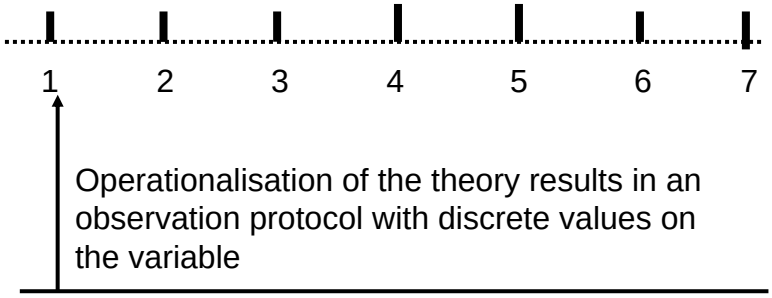
Scales	Nominal	Ordinal	Interval	Ratio
nominal	**groups**			
ordinal	groups	+ **ranks**		
interval	groups	+ ranks	+ **distance**	
ratio	groups	+ ranks	+ distance	+ **absolute zero**
examples	Municipality	Strength of attitude to EU	Temperature in C^0	Age Temperature in K^0

Ordinary variables

- Very many variables in sociology and political science are actually ordinal scales
- But with some assumptions satisfied, and for the purposes here, they can be treated as interval scales. The assumptions are
  - The number of categories is large enough (more than 5)
  - The observations are distributed across (almost) all categories. There must be a sufficient number of persons outside the 2-3 modal categories
  - It is reasonable to assume that in reality the scale is at least interval (continuous with distance measure)

Measuring variables

Our observations can in practice distinguish among 7 different values only



Theory may assume that in reality there is a continuous scale

Typically: direction and strength of opinions or emotions

Dichotomous variables

- Has 2 values or 2 codes and can be used in all kinds of regressions as independent variables
- All variables can be recoded to have only 2 values
- If the 2 codes are 0 and 1 the interpretation of their effect when they are used as independent variables is much easier than if other codes are used (e.g. 1 and 2)
- The number of categories in the smalles category must be “large enough”

On the addition of new variables

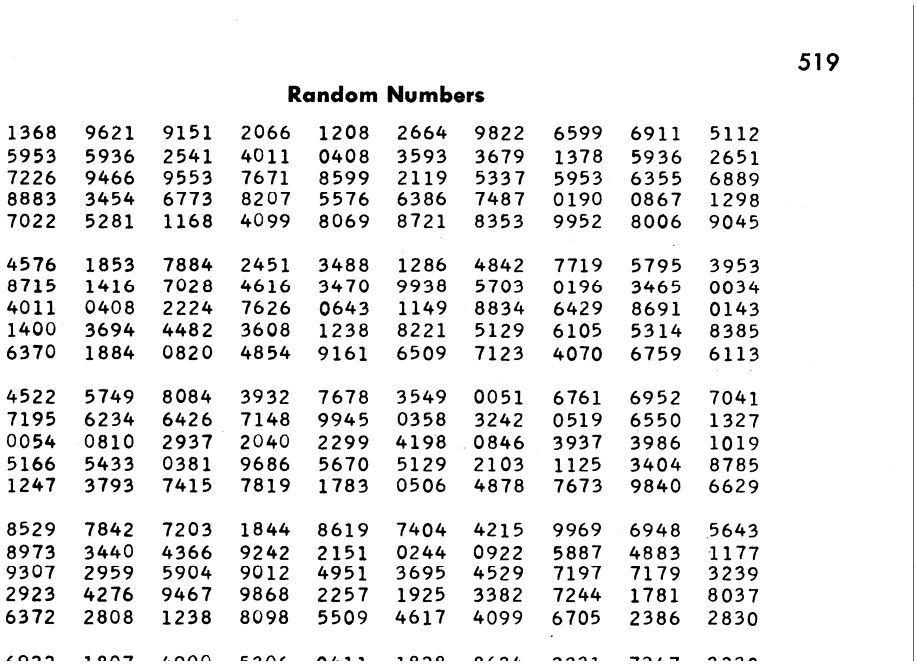
- It is not common that existing theory will give precise prescriptions for what variables to include in a model. Usually there is an element of trial and error in developing a model
- When new variables are added to a model several things happen
  - The explanatory force increase: R^2 increase, but will the increase be significant?
  - The coefficient of the regression shows the effect on y. Is this effect significantly different from 0?
  - If the coefficient is significantly different from 0, is it also so big that it is of substantial interest?
  - Spurious coefficients can decline. Do the new variable change the interpretation of the effect of the other variables?

Parsimony

- Parsimony is what might be called an aesthetic criterion of a good model. We want to explain as much as possible of the variation in y by means of as few variables as possible
- The adjusted coefficient of determination, Adjusted R^2, is based on parsimony in the sense that it takes into consideration the complexity of the data relative to the complexity of the model by the difference between n and K

  (n-K is the degrees of freedom in the residual,

  n = number of observations, K = number of estimated parameters)



519

Random Numbers

1368	9621	9151	2066	1208	2664	9822	6599	6911	5112
5953	5936	2541	4011	0408	3593	3679	1378	5936	2651
7226	9466	9553	7671	8599	2119	5337	5953	6355	6889
8883	3454	6773	8207	5576	6386	7487	0190	0867	1298
7022	5281	1168	4099	8069	8721	8353	9952	8006	9045
4576	1853	7884	2451	3488	1286	4842	7719	5795	3953
8715	1416	7028	4616	3470	9938	5703	0196	3465	0034
4011	0408	2224	7626	0643	1149	8834	6429	8691	0143
1400	3694	4482	3608	1238	8221	5129	6105	5314	8385
6370	1884	0820	4854	9161	6509	7123	4070	6759	6113
4522	5749	8084	3932	7678	3549	0051	6761	6952	7041
7195	6234	6426	7148	9945	0358	3242	0519	6550	1327
0054	0810	2937	2040	2299	4198	0846	3937	3986	1019
5166	5433	0381	9686	5670	5129	2103	1125	3404	8785
1247	3793	7415	7819	1783	0506	4878	7673	9840	6629
8529	7842	7203	1844	8619	7404	4215	9969	6948	5643
8973	3440	4366	9242	2151	0244	0922	5887	4883	1177
9307	2959	5904	9012	4951	3695	4529	7197	7179	3239
2923	4276	9467	9868	2257	1925	3382	7244	1781	8037
6372	2808	1238	8098	5509	4617	4099	6705	2386	2830