On writing term paper for SOS3003

Erling Berge

Purpose

- The term paper is a part of the formal examination results and will be evaluated. The mark for the paper will have a weight of 0.51 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.

Preliminary

- It is a goal that everybody have different variables
- Sources for the variables might be
  - Your own data [recommended] for example collected in other ways for example for your own masters thesis.
  - If you in an earlier SOS3003 class has started on a paper you may continue here with the same variable
  - For the rest of you relevant variables will be available from the European Social Survey and other surveys available from NSD

Dependent variable

- Based on theory the variation in an appropriately chosen dependent variable shall be explained.
- A dependent variable will have to satisfy some minimum requirements securing that a valid multiple regression is possible. Hence, use of your own data has to be approved.

Requirements for a dependent variable

- The variable must vary !!!!
- The requirement is that a dependent variable in OLS regression has to be interval (or ratio) scale. In principle it has to be able to take any value between minus infinity and plus infinity.
- Deviations from this may cause problems. Most of them can be overcome

- Logistic regression require a dichotomy (exactly 2 values)

  (but can easily be extended to multiple categories)

A note on the dependent variable in OLS regression:

- It is not, I repeat NOT, most emphatically **NOT** required that a dependent variable in OLS regression shall have any particular distribution such as a normal distribution
- In some other types of models this is different. Maximum likelihood factor analysis for example assumes a multivariate normal distribution
- When normal distributions are introduced, they are assumed in order to be able to do tests

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. Do you have at least 7 different variable values?
- 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

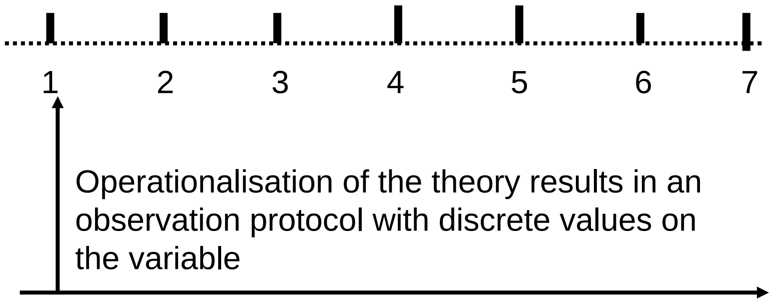
Scales	Nominal	Ordinal	Interval	Ratio
nominal	**groups**			
ordinal	groups	+ **ranks**		
interval	groups	+ ranks	+ **distance**	
ratio	groups	+ ranks	+ distance	+ **absolute zero**
examples	Municipality	Strength of attitude to EU	Temperature in C^0	Age Temperature in K^0

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



Operationalisation of the theory results in an observation protocol with discrete values on the variable

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. But think carefully before doing that. Reducing the variation in the data needs a good reason
- 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 cases in the smallest 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 structure of a paper

- **Title page**
- **Preface**
- **Abstract**
- **The main body of the paper**
- **References**
- **Appendices**

Title page

There has to be

- **TITLE**
- **CANDIDATE NUMBER**
  - You may add other things as you find appropriate on a title page
- The paper is to be sent by e-mail to **<ISSInnlevering@svt.ntnu.no>**

Preface

- In a preface the appropriate acknowledgments are presented.
- If the data used have been collected by SSB (Statistics Norway), SSB should be acknowledged and absolved of responsibility for the interpretations, for example by saying:

Acknowledgement of data source

- “(Some of the) Data used in the present publication have been taken from (…name of the data set…). Data in anonymous form were made available through the Norwegian Social Science Data Service (NSD). Data were originally collected by Statistics Norway. Neither Statistics Norway, nor the Norwegian Social Science Data Service has any responsibility for the analysis or interpretations presented here.”
- For municipal or county data, or data from other sources than SSB and NSD, the formula should be adapted appropriately.

… eller på norsk

- "(En del av) de data som er benyttet i denne publikasjonen er hentet fra .... undersøkelsen (årstal). Data i anonymisert form er stilt til disposisjon gjennom Norsk samfunnsvitenskapelig datatjeneste (NSD). Innsamling og tilrettelegging av data ble opprinnelig utført av Statistisk Sentralbyrå. Hverken Statistisk Sentralbyrå eller NSD er ansvarlige for analysen av data eller de tolkninger som er gjort her."
- For kommune eller fylkesdata eller data frå andre kjelder må formularet tilpassast. For kommunedata skriv ein t.d. ikkje «i anonymisert form».
- Skriv du på nynorsk går det sikkert greitt å omsetje formularet ☺

Abstract

- An abstract or summary will comprise 100 – 200 words summarizing very briefly the data source and the main findings. The abstract should be formatted separately, with spacing of 1.0

The main body of the paper

- The abstract, the main body of the paper, and references are limited to a maximum of **10.000** **words** as counted by Microsoft Word

References and Appendices

- The list of references should follow some accepted standard.
- One or more appendices will includes tables, figures (graphs), and a little explanatory text to describe aspects of the data relevant to variable transformations, scale construction, tests for additional variable transformations, tests for violations of statistical assumptions, and the examination of influential outliers .
- There is no size limitation for the appendices. However, in most cases about 5 – 15 pages, as appropriate to the particular analyses presented, will be sufficient

Requirements (1)

1. Based on descriptive statistics for the variables included in the model, their distributions shall be investigated and possible transformations considered. But do not start with transformations. Transformations should be used if their use will improve the analysis (i.e. if there are theoretical reasons to believe that the marginal relationship between explanatory variable and dependent variable is curvilinear (see pt 4 below) or if use of transformations will make tests more trustworthy (i.e. the residual is closer to a normal distribution).

Requirements (2)

2. The model must contain at least one nominal scale variable with **more than** 2 categories. This is done to include use of dummycoded multinomial variables.
3. Possible interaction among variables shall be considered and at least one interaction term has to be tested.
4. Possible curvilinear relationships have to be considered and at least one curvilinear relationship has to be tested.
5. At least one “conditional effect plot” should be presented and interpreted.

Requirements (3)
The following problems have to be discussed

6. Multicollinearity has to be considered
7. The impact of outliers and influential cases has to be considered
8. The model specification has to be evaluated.
9. In OLS regression heteroscedasticity has to be considered
10. In OLS regression autocorrelation has to be considered
11. In OLS regression the distribution of the residual has to be considered
12. In LOGIT regression the problem of discrimination has to be considered

Requirements (4)

- For all models important and relevant tests of significance, parameters, coefficients of determination, and sometimes confidence intervals, have to be report and correctly interpreted.
- At the end the results should be discussed in relation to the original problem.

Advice

- The detailed requirements presented in points 1-12 cannot be used as a blueprint for writing the paper. Not all of this will be expected to be found in the body of the paper. In the text much may be briefly mentioned while the documentation will be in the appendices.

Advice on the main body (1)

The main body of the paper might be structured like this

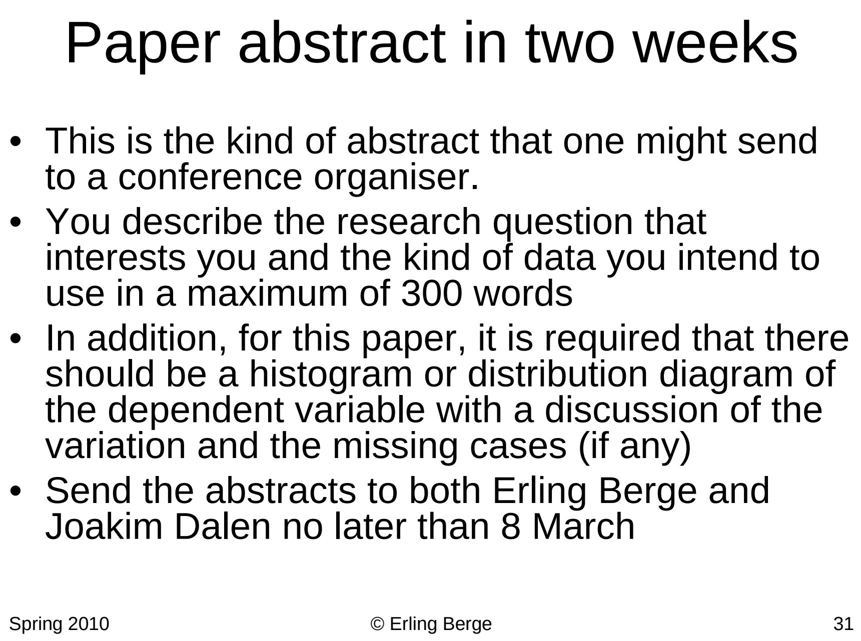
- An introduction of 1/3 - 1 pages stating the research question and describing why the research question is interesting and/ or valuable.
- A short discussion of theory relevant to the research question (1 – 3 pages)
- A short summary/mention of previous research relevant to the question (1 – 2 pages).
- 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, that is much longer than the conclusions statements in the abstract, which should be only 1-3 sentences (in the abstract)

Advice on what is important

- The most important part of the work is the model specification.
- The tools for diagnosis are important to improve the model specification (e.g. interaction terms, curve elements, variable transformations) and to clarify problems of interpretation (e.g. distribution of the residual, impact on dependent variable when the relationship is curvilinear)
- The technical requirements are there to improve model specification and interpretation



Paper abstract in two weeks

- This is the kind of abstract that one might send to a conference organiser.
- You describe the research question that interests you and the kind of data you intend to use in a maximum of 300 words
- In addition, for this paper, it is required that there should be a histogram or distribution diagram of the dependent variable with a discussion of the variation and the missing cases (if any)
- Send the abstracts to both Erling Berge and Joakim Dalen no later than 8 March