REQUIREMENTS FOR THE TERM PAPER FOR
SVSOS3003 "Applied statistical data analysis for the social sciences"

A. 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 0.54 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.

B. SUGGESTED PROCEDURE

Based on sociological or political science theory a problem discussion should explain the reasons for analysing the variation in an appropriately chosen dependent variable. The dependent and independent variables can either be taken from a data set prepared for the class (see separate presentation) 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 can be performed. Hence, use of your own data has to be approved. The requirement is basically that the dependent variable either is a measurement scale variable with sufficient variation for ordinary least squares regression or a nominal scale variable for logistic regression.

C. FORMALITY

Title page

The first page of the paper shall at a minimum contain Kandidat no[1] and a title indicating which dependent variable is investigated.

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

> "(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.

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

The list of references should follow some accepted standard.

[1] Starting this fall a student is given a separate number for identification for each examination. This is the "Kandidat" number.

Appendices
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 formal size requirement for the appendices. However, in most cases about 5 – 15 pages, as appropriate to the particular analyses presented, will be sufficient.

Deadline for the paper
For the spring term the deadline is May 10 and for the fall term it is November 23. The paper is delivered in .doc format by e-mail to ISSInnlevering@svt.ntnu.no. If this fails, send it to the department office in 3 copies. For e-mail delivery the deadline is 24:00 on the date. For mail delivery at the office the deadline is end of office hours, 15:45.

Binding
The paper does not need more binding than staples.

D. REQUIREMENTS OF THE PAPER
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

In the specification of a first regression model for estimation the following elements has to be included:

1. Based on descriptive statistics for the variables included in the model, their distributions shall be investigated and possible transformations considered. 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).
2. The model must contain at least one nominal scale variable with more than 2 categories.
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.

Based on the results from the first model estimated, it is expected that tests for whether the data conform to the assumptions required for the model are reported on. This means that 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 heteroskedasticity 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

In most cases this will lead to testing more than one model. 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.

OBSERVE
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.

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 clarify problems of interpretation (e.g. distribution of the residual, impact on dependent variable when the relationship is curvilinear).

ADVICE ON FORMAT AND REQUIREMENTS

Note: There is a distinction between the paper requirements, and the advice here about what improves a paper. The list of analyses which must be included and the paper length considerations are requirements. Descriptions of sections, section length, and formats are recommendations.
The basic model for this paper is that it should have the same form as a serious research article for an academic journal, under length constraints insisted upon by the editor. The course paper is different from that format in one respect, and that is that the course paper includes a much longer appendix documenting additional data descriptions, analyses, and checks for problems than would be included in a published journal article.

The analysis project upon which the paper is based: The papers are to be based on regression analysis, using either OLS or logistic regression. The analyses should study the effects of a set of independent variables on a single dependent variable. (Some students may study two to four dependent variables within the context of a structural equation model; but they should concentrate their detailed analyses relevant to various requirements listed below on a single dependent variable.) The data set and dependent variable used by each student must be approved by the instructor. Every student in a given term must choose a unique combination of dependent variable(s) and data set. Most students are encouraged to choose variables from the European Social Survey (ESS), because this provides many possible variables, is easily available for the class, and is familiar to the instructors. However, with permission, they may use other data sets instead. The data do need to be based on approximately random, representative samples of cases, of a reasonable size for the use of the regression statistics taught.

If a student took SOS3003 earlier, turned in a paper for that course and either did not complete the course, or received an unsatisfactory grade, they may use the same paper (hopefully improved) as their paper for this term; and they have a priority in using the variable used previously. To get a final grade for the course this term, they also must take the school exam again in addition to turning in a course paper.

Structure of the paper

Length requirement has been changed from number of pages to number of words since number of words now is the most common way for journals to limit the size of articles. The length of 10.000 words is in the upper range of what different journals specify and should be taken seriously. Graders will be asked to take it seriously. They are under no obligation to grade on the basis of pages beyond the 10.000 words, and are instructed to count exceeding page length as negatively as a paper not being complete enough. A journal editor would return your paper for rewriting if you exceeded his/her instructions regarding length.

Advice on various sections of the paper:

Abstract

It looks nice in italics and/or a smaller font than the rest, so as to be about 1/3 – 1/2 page.

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.
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)

References

Using reference software and selecting a standard for presenting the references is good practice for writing a master thesis. A well developed standard with explicit advice on all kinds of documents is the Chicago Manual of Style Online. Use the style developed for science and social science. In the EndNote software it is called Chicago 15th B. The online edition is also the 15th edition of this manual.

However, the main requirement is to follow some standard consistently.

COMMENTS ON VARIOUS SECTIONS OF THE PAPER

The theory and previous research section*:* Everybody asks how important it is and how long it should be. Different instructors may have different opinions on this, and sensors (graders) can also. Most of us think that this paper is very demanding, and it is really tough to write a good theory section when you are forced to take an available variable in competition with all the other students. For this reason, most of us do not want to put undue emphasis on the theory section, and it should not be more than 1 – 3 pages. Nevertheless, we are pushed to grade papers in a competitive model, and there is no question that everything else being equal, the quality of the theory section has an influence, especially related to the highest grades.

Note that the "theory" does not have to be based mostly on literature and can and should include your own ideas and hypotheses. Also remember that without theory the analysis has no meaning at all.
The same things can be said about the discussion of previous research. It does not have the highest importance, but is still influential. For those using the ESS data, it is a little early for there to be a huge mass of published articles available, but there is a lot, especially if you can find conference papers, and much of it is relevant to many papers published based on the World Values Surveys , the International Social Survey Program (ISSP), and other similar projects. If you can find such papers, they will lead you to theory.

Your hypotheses: Much of the statistical literature approaches issues as if you only had one hypothesis. This is a model for hard (physical and medical – "real-fag") science research, which is not always so relevant for social science research these days, when we can tear through hypotheses in minutes of analysis based on secondary analysis. You should be guided first by your theoretical hypotheses, while you will often test additional hypotheses quickly, as your data teach you and lead you on. No one will want to read a null and alternative hypothesis for all possible relationships. So, what you should do is to settle on one to four main beginning hypotheses to describe before the analysis section. These may be described in terms of very specific null and alternative hypotheses. If you discover new relationships during analysis, you can describe them there. There are likely to be a number of variables which you think have possible or likely effects on your dependent variable, but are not part of your main theoretical hypotheses. You can label these "control variables," and comment on the reason for including them, and their likely effects, as briefly as possible.

Tables and figures to include in the text: This depends partly upon how much space you have within the 10.000 words, but the test is whether these would be included in the text of a professional article. Such articles often include a table describing all of the variables included in the various models, including their number of valid cases, maximum and minimum values, means, and standard deviations. Generally, you should include such a table, though it can be omitted if you have more important things for the text, but in such cases it should be in the appendix. In general, you do not need to present figures (graphs) for the univariate distributions of the independent variables, even in the appendix. You might present such a figure or figures for the dependent variable in the appendix, if you think that its distribution was problematic, or involved a transformation, and needed some discussion of how to deal with it.
In the text you definitely should include tables showing the usual statistics for coefficients, and tests of significance for coefficients for your most important models. How many such models are represented in tables in the text depends on your topic, data, and analyses, but usually there should be 2 – 4 such models. Sometimes the coefficients for more than one model can be included in a single table. Other times, separate tables may be best.
The statistics for tests of significance of overall models and contrasts between models, and coefficients of determination or their analogues, can be reported in the text, or as part of these larger tables (they do not require tables of their own). Conditional effect plots might be included in either the text or the appendix. Graphs showing non-linear effects might also be in either the text or the appendix. The results of your analyses testing assumptions, looking for influential cases, testing scaling, or looking for non-linearities should definitely be stated in your text (with references to where these might be included in the appendix), but the actual graphs and tables of statistics for these should usually be in the appendix.

What belongs in the appendix? The appendix is the place for graphs and tables of statistics related to tests of whether the assumptions underlying the models are met, and whether influential cases of outliers are problematic. The appendix also may include statistics for additional models which were tested, and led to the development of the final model. It might also include small matrices or tables related to scale construction. You should number or label your graphs and figures, so that you can refer to them clearly by page and label in the text of your paper. You definitely should NOT thoughtlessly include masses of SPSS output on univariate variable distributions, large bivariate correlation matrices, etc.

Mistakes you should definitely avoid: *You definitely should not include passages of general explanations of various statistics in an abstract way unrelated to your specific research analyses. There should not be any sections which are simple restatements or summaries of textbook explanations of statistics. You should show that you understand the statistics by the way you use them, and the way you interpret them relative to your specific research questions and results.*
Take great care in your assumptions about the direction of causality between variables, especially subjective attitude variables measured at the same time. If such assumptions are implicit in your model, they should be explicitly stated and justified in your discussion. If such assumptions are questionable, you should mention this. Such considerations might lead you to try different models, including and excluding the relationships with debatable causal relations.
Be sure to describe each of your variables very clearly - the exact meanings of your dummy categories and the omitted reference category, any variable transformations, the wording of variable questions for attitude items, what high versus low values mean, and the items included in any scale. If this takes space, some of these details can be placed in the appendix.

Features of the very best quality papers: The analysis requirements listed above (such as test at least one interaction effect, one set of dummy variables for a variable with at least three categories, on non-linear term, etc.) are minimum requirements. Graders are likely to be impressed with more extensive analyses in addition to this minimum, especially if they are related to theory and/or yield substantive findings. Special attention to technical details of data exploration is important, as is a good discussion of the relation between theoretical ideas and your data and analyses. Unusually ambitious choices of data or analysis techniques are appreciated; but, you should be very careful not to "get in over your head," with analysis problems which might take time and space which would end up hurting the final overall paper. It is very difficult in an exercise with these kinds of constraints, but signs of *creativity* in variable choice and formation, hypotheses, and analyses, are appreciated indeed.
You will receive only one overall grade for the course. The paper is weighted 0.6 and the school exam weighted 0.4 in the graders' evaluation of the overall grade.