

## **REQUIREMENTS FOR THE TERM PAPER FOR SVSOS3003 “Applied statistical data analysis for the social sciences”**

### **A. 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 0.6 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.

### **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 Student number 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.

#### **References**

The list of references should follow some accepted standard.

#### **Deadline for the paper**

For the spring term the deadline is May 15 and for the fall term it is November 30. The paper is delivered at the department office in 2 copies.

#### **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.

Based on the results from the first model estimated the following problems have to be discussed

5. In OLS regression heteroskedasticity has to be considered
6. In OLS regression autocorrelation has to be considered
7. In OLS regression the distribution of the residual has to be considered
8. In LOGIT regression the problem of discrimination has to be considered
9. Multicollinearity has to be considered
10. The impact of outliers and influential cases has to be considered
11. The model specification has to be evaluated.
12. In the discussion and interpretation of results conditional effect plots has to be used.

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. 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).