ENGLISH WITH GRADER’S INSTRUCTIONS

EXAMINATION FOR SOS3003
Applied Social Statistics, Fall 2011

Academic contact during the examination: Albert Andrew Simkus
Telephone: 99 53 21 74
Examination date and time: 15 December 2011 09:00 – 15:00
Examination duration: 6 hours Study points: 15
Number of pages in bokmål, excluding appendix/attachments: 6 Number of pages in English, excluding appendix/attachments 6
Number of pages in appendix including attachments: 14
Internal Grading date: 11 January 2012
Permitted helpful materials in the examination: Calculator

NO books or notes of any kind permitted.

ENGLISH

General Instructions: For each set of questions, the approximate weight given the answers for that section toward the total grade, and the approximate length of good answers for that section are given in parentheses before each set of questions. These are guidelines so that you can tell approximately how long an answer is expected for each part, and you should write neither too much nor too little. The quality of the answers is much more important than the length of the answers. The grading of the exam will be based on the overall quality of the examination, and particularly good answers for some questions may or may not balance out particularly bad answers. Weight will also be given to the consideration that – regardless of question length - some questions are intentionally rather easy, while others are more difficult. If Norwegian is your native language, it is still wise to read the English-language version of questions as well, to be extra certain you understand the question details clearly.

You must write clearly! Clearly label and number the question you are answering, and keep your answer pages in the order of the questions!
QUESTION BLOCK I

(THE TOTAL WEIGHT FOR THIS BLOCK OF QUESTIONS IS APPROXIMATELY 30 PERCENT, NO QUESTION SHOULD REQUIRE MORE THAN ¼ - ¾ PAGE TO ANSWER).

Part 1: 1.1.1 – 1.1.5

FOR THE FOLLOWING FIVE QUESTIONS, YOU SHOULD NAME THE STATISTIC AND/OR SYMBOL THAT IS DESCRIBED IN THE QUESTION, OR ANSWER WITH A SHORT SENTENCE.

1.1.1 In a univariate distribution, which statistic is equal to the average squared deviation of the individual values a variable from the mean for that variable?

   The variance

1.1.2 In OLS regression, if you wanted to evaluate two models for the same dependent variable, in terms of the variance explained by the models, and one model included several more independent variables than the other, what statistics should you compare?

   The adjusted R-squares

1.1.3 In logistic regression, what statistic is equal to the square root of the Wald test?

   The t – test

1.1.4 What is the difference between a standard deviation and a standard error?

   The standard error is the standard deviation of a sampling distribution.

1.1.5 In OLS regression, exactly what does the unstandardized regression coefficient tell you?

   The average number of units the dependent variable is expected to change with one unit of increase in the given independent variable, controlling for the effects of the other independent variables.

Part 2 1.2.1 – 1.2.10

1.2.1 You can generally test the statistical significance of individual regression coefficients using a t – test. However, in a few situations, you should test the significance of adding blocks of two or more independent variables to the equation. Briefly list the situations where you should do this, and give the names for such tests in OLS and logistic regression (the names of the test statistics).
You would do this when you are testing the overall significance of the improvement of the model when adding a set of dummy variables representing a single theoretical variable; you would also do this when testing the significance of adding a polynomial squared term for curvilinearity, and when testing adding interaction effects, especially in logistic regression. The test for OLS is the F-test for change in R-square; the test for logistic regression is the test for the Likelihood-ratio Chi-square (change in -2LL).

1.2.2 Which measure of “influential cases” is best to use when your analysis concentrates on one particular hypothesis? List standards for evaluating the seriousness of such influence, and justify the one you would prefer.

One or more DFBETAS, because they are specific to specific regression coefficients.

Criteria: 1) If the absolute value of the DFBETA is > 2
2) A sample-size adjusted cutoff is and absolute value > 2/sqrt n
3) DFBETAS < Q1-3IQR or DFBETAS > Q3+3IQR
4) Gaps between outliers and the remaining cases

Any reasonable reason for preference should be satisfactory. I emphasize the first criteria, since it is the real indicator of how serious the effect is, plus maybe number 4 to look for errors in data. Since the students are not allowed to use books or notes, they should be given full credit for suggesting as few as two of these four.

1.2.3 In OLS multiple regression, what is an advantage of comparing standardized (BETA) coefficients within a single model? Are these coefficients also helpful in comparing models for different samples or populations?

Betas can be useful in comparing the relative explanatory importance of variables measured on different metrics within the same regression with the same sample or population. It is incorrect to compare Betas between different models and samples, since they are a function of the standard deviations of variables and these can differ across samples.

1.2.4 Can multicollinearity be a problem in logistic regression? If so, how can you test for it with SPSS?

Yes, it can be a problem in logistic regression as well as in OLS regression. You can test for it by running the problem as an OLS regression and examining the tolerance statistics. The regression coefficients would incorrect in this case, but the tolerance statistics are completely based on the intercorrelations among the INDEPENDENT variables, and are not affected by the problem of having a dichotomous dependent variable.

1.2.5 Is heteroskedasticity a serious problem to check for in logistic regression? If so, how would you test for it?

No, its not an issue at all! There is necessarily a huge amount of heteroskedasticity, but the model is handled differently than in OLS. You would never test for it.
1.2.6 List five words or expressions which can describe the shapes of problematic univariate frequency distributions for OLS residuals.

Skew, Kurtosis, Heavy Tails, Assymetry, bi- or multi-modal. (Other non-normal distributions (uniform, U-shaped, etc.)

1.2.7 List two situations where independent variables may have worrisomely low values of “tolerance,” yet it is reasonable to keep the variables in the model.

When one variable is a function of the other, such as a squared term for a polynomial for curvilinearity.

When including an interaction effect including the same variables in the model as main effects.

1.2.8 What are two reasons why you can’t use OLS regression with a dichotomous (binary) dependent variable?

1. Because OLS regression could give predicted values below 0 or above 1, which are impossible. Because there would necessarily be huge problems with heteroskedasticity.

1.2.9 Imagine that you look at a set of statistics estimated from a multiple regression model. In the SPSS output, you notice that the unstandardized regression coefficient for the variable age squared is 0.000. But the probability for the t-test indicates that the coefficient is statistically significant. At the same time, the coefficient for a dummy variable for gender (being male) is .001 and is not statistically significant. Is this possible, or an obvious mistake? Very briefly explain.

It is very possible. The variable for age-squared will have very high scores, and the regression coefficient indicates the change in the dependent variable with only one-unit of change in the independent variable. This being the case, the coefficient can easily be so small that the regression coefficient is rounded off to zero. You can mouse-click on the coefficient in the SPSS output several times and see its value to a far higher level of precision. Depending on how extreme the problem is, you might want to change your variable for age to age/10 or age/100.

1.2.10 If you believe an independent variable measured on the ratio level has non-linear effects, you might add a variable for that variable squared, or you might break the continuous variable into a series of dummy variables of equal intervals. What would be the
advantages and disadvantages of either approach?

Using a series of dummy variables might reveal and capture non-linear patterns which do not match the curvilinear function of the polynomial. For example, maybe the effect of education is best captured by a simple difference between those with university educations versus everyone else. Or, maybe there are just three steps of differences, between primary, secondary, and university education. Age might have similar, step-like effects, rather than being smoothly curvilinear throughout its range.

On the other hand, if there is a smooth curvilinear pattern, using dummies is inefficient, and does not provide as powerful a test, since it uses up more degrees-of-freedom and makes it less likely the change in R-square will be statistically significant.

QUESTION BLOCKS II AND III: DATA DESCRIPTION

The data for blocks II and III of the exam are all from a survey of residents of Macedonia in 2003.
In this survey, ethnic Albanians were over-sampled to provide lower standard errors for estimates for the Albanians. Members of smaller ethnic minorities are omitted from these analyses, which concentrate on ethnic Macedonian-Albanian differences. The independent variables for both blocks are:

MAK- Ethnic Macedonian, with ethnic Macedonians coded “1” and ethnic Albanians coded “0.”
MALE – A dummy variable with males coded as “1”, and females coded as “0.”
AGE - The respondent’s age, in single years, from age 21 – 80.
AGESQ – The variable AGE squared.
EDYEARS – The respondent’s number of years of formal education completed.
EDYEARSSQ-The respondent’s number of years of formal education squared.

RURAL4 – A four-point scale of how rural the respondent’s place of residence was (the capital city of each country was coded “1”, the most rural small villages and isolated farms were coded “4”, and there are scaled categories between).

There are also variables for interaction effects between ethnicity and age and between ethnicity and education: MAxAge and MAxEDyrs.

There are also sets of dummy categories for age in ten-year age groups, levels of completed education, and size of place of residence. The omitted (reference) category is always the lowest group, denoted by an “x” in this list.

ED_0_7x,
ED_8_10
ED_11
ED_12_13
ED_14_15
ED_16_25
AGE21_30x
AGE31_40
AGE41_50
AGE51_60
AGE61_70
AGE71_80

Capitol
City
Town
Village

QUESTION BLOCK II: OLS REGRESSION
A simple microscale based on the average response of respondents to three Likert-format attitude items was formed as a measure of ethnic “exclusionism” or “intolerance. The three items were:

1. “Nationally mixed marriages must be more unstable than others.”
2. “Men (people) can feel completely safe only when the majority belong to their nation (nationality).
3. “Among nations (nationalities) it is possible to create cooperation, but not full trust.”

The range of possible responses ranged from 1 – 5, with high values indicating ethnic exclusionism/intolerance.

This scale is the dependent variable for the analyses in this second block of questions.

At the end of this examination are nine pages of SPSS output for Question Block II based on a series of TWELVE (9+3) OLS regression models with this scale as the dependent variable. Be sure to consider all 12 models, even though they are presented in two sets.

**NB!** In the succession of models in these tables, specific variables may be removed, as well as added, to facilitate various tests of changes in the models. Be sure to examine the tables of coefficients carefully to see exactly which variables are being added or deleted!

Based on what you observe in the tables and figures for these pages of output, answer the following question.

2.1 In approximately 3 – 7 written pages, describe the most important conclusions you can draw about the determinants of this scale. Include statements about the best models and most important variables, the directions, magnitudes, linearities or forms, and meanings of the effects of specific variables, and assessments of possible violations of regression assumptions and other problems. Pay attention to both statistical significance and substantive importance.

This question is intentionally being asked in a very open non-specific form. You will be evaluated on the basis of how well you recognize, organize, and state the most important findings and issues. We want you to recognize the most important questions and answer them. As a guide, you might think about the organization of the analysis and results sections of your term paper.

**QUESTION BLOCK II: OLS REGRESSION**

*What the analyses show*
1. The first thing we note is that the beginning basic model for the simple main effects treated as variates and dichotomies (Model 1) shows Macedonian ethnicity, years of education, and rural residence to have statistically significant effects while age and gender do not. The effect of ethnicity is by far the largest, followed by education, as indicated by their standardized regression coefficients (BETAS) of -.274 and -.134, respectively. The effect of rural residence is barely significant at the .05 level and its regression coefficients are relatively weak. Substantively, ethnic Macedonians show much lower average levels of ethnic intolerance than do ethnic Albanians, and years of education have a negative effect on intolerance. Rural residence has a positive, though weak, effect on intolerance. The R-square for the model as a whole is statistically significant, but substantively moderate for this kind of dependent variable.

2. The second model adds a variable for age-squared, creating a polynomial for curvilinearity in the effect of age. It is just barely statistically significant (p=.035) and results in an improvement of only .002 in R-square. In model 3, the variable for R-square is removed, and in model 4, a term for years of education squared is added to test for curvilinearity in the effect of years of education. The improvement in R-square is .003 and statistically significant (p=.003). In model 5, the variable for age-squared is added back to the model, but with the effect for education squared already in the model, the improvement in R-square is negligible (.001) and not statistically significant (p=.075). These tests indicate that adding a squared term for curvilinearity for the educational effect improves the model more than does adding a squared term for the age effect. And, once the squared term for education is in the model, the squared term for age does not make a significant improvement. We conclude that there is a significant negative effect for age, which shows barely significant curvilinearity, but that net of education, age has no significant effect, neither linear nor curvilinear.

3. In model 6, the model is reduced to the strong negative effect of Macedonian ethnicity plus the insignificant effect of gender. Model 7 adds five dummy variables for the effect of six ten-year intervals of age. Why might someone do this? Because it is possible at this point that the weak effects of age are simply due to the fact that we are not catching the right steps or functions. There could be historic cohort differences, or non-curvilinear age effects. But neither Model 7 nor later ones indicate that this is true. The differences across age/cohort categories are inconsistent, before or after controls for other variables. Model 8 adds dummy variables for educational categories, and here we see a very different story. Increasing education has increasingly negative effects, and they are monotonic if not perfectly linear. Model nine adds dummy variables for size of place of residence. These too are monotonic if not perfectly linear. Models 6 through 9 lead us to conclude that the effects of age/birth cohort are weak, inconsistent, and due to neither simple curvilinear functions nor to step effects across age groups or age cohorts. On the other hand, the effects of years of education, and to a lesser extent, ruralness of residence, are quite evident and consistent in the analyses based on dummy variables.

4. The three final models reflect the following logic: OK, age/cohort does not seem to have any consistent or important effects. Education and ruralness of residence evidently do. Maybe we should accept simple and efficient measures of education, rural residence, and maybe even age, so as to be able to look for interaction effects in a practical and efficient way. So we go back to accepting years of education, age, and ruralness of residence as continuous variates,
and test for interaction effects, in models 2 and 3 of Table 3. What we discover is that if we accept treating the main independent variables as continuous variables, there are significant interaction effects between being ethnic Macedonian and age, as well as between being ethnic Macedonian and years of education. According to the coefficients, the ways these work is that the effect of age on making one more ethnically intolerant is greater for ethnic Macedonians than for ethnic Albanians; and, the effect of education on making one more ethnically tolerant is greater for ethnic Macedonians than for ethnic Albanians.

Which is the “best” model?

So, in the end, which is the best model? This problem has intentionally been developed in such a way that the final answer is not easy; and, in fact more than one model can be considered “best” from different points of view. Furthermore, the sequence of tests and alternative model details requires understanding the implications of the variables being added and withdrawn. The best possible answers should consider the statistical tests, adjusted R-squares, parsimony, and substantive significance of the choices among models. The quality of the argument should be counted as much as the final decision. The following is an example of a kind of argument which might be made. Others are possible, and the average student is not expected to be able to provide a discussion on quite this level of detail.

Based on the sequence of statistical tests, the adjusted R-squares, and other considerations, Models 4 or 9 in Table 1, or Model 3 in Table 3 are all justifiable contenders for “best model.” Model 5, Table 1 has a slightly higher adjusted R-square than Model 4, and builds upon Model 4, but the F-test for improvement from Model 4 to 5 is not significant.

The adjusted R-square is highest for Model 9, Table 1 (.124) and this could be an argument for preferring that model. However, the adjusted R-square for Model 9 is only .001 higher than Model 3, Table 3, which uses many less parameters. Model 9 probably achieves its slightly higher adjusted R-square than that for Model 4, Table 1 based on picking up small non-linearities (or even non-curvilinearities) in the relationships. We might consider that the large number of dummy variables is just picking up a very little more random variation. The use of the adjusted R-square for comparisons should correct for this, but we might suspect that it might not adjust quite precisely enough for differences as small as .001.

One could argue that Model 4, Table 1 has an adjusted R2 which is only slightly lower than that for Model 9 (.121 versus .124), yet is far more parsimonious. If simplicity and parsimony are a value, one might prefer Model 4 over Model 9 amongst the choices in Table 1.

However, Model 3 of Table 3 has an adjusted R-square of .123. In terms of parameters, it is much more parsimonious than Model 9, Table 1, and almost as parsimonious as Model 4, Table 1. The sequence of significance tests, adjusted R-squares, and concern about parsimony leads us to focus on Models 4 and 9, Table 1, and Model 3 Table 3. We note that the difference in adjusted R-square between the simplest model, Model 1, Tables 1 and 3, and Model 9, Table 1, is only .006. Among the top three contenders for best model, the biggest difference is .003. There are a number of arguments to console us about worrying about such small differences, but they are not part of this examination question. The important thing is that there is still another criterion for judging models – their substantive meanings.
The substantive difference between the three possible best models is that Model 4 Table 1 includes a parsimonious curvilinear effect for education, Model 9, Table 1 is not parsimonious but picks up irregularities in the forms of the relationships, and Model 3, Table 3 leaves out non-linear effects but adds two interaction effects among independent variables. There are no significance tests included for differences among these three models, but their substantive differences are clear. Which is more important, the precise functional form of the effects of education and place of residence, and how efficiently these can be represented, or the interaction effects between ethnicity and education and age?

For me, the interaction effects in Model 3, Table 3 are substantively most interesting and might have been a priori hypotheses. The model indicates that the effects of age (positive) and education (negative) are greater for Macedonians (who have lower average scores on intolerance) than for ethnic Albanians (who have higher average scores). To me, this substantive finding is more important and interesting than the trivial advantage of Model 9 having an adjusted R-square larger by .001 due to either chance or being more sensitive to non-linearities. Thus, on the basis of significance tests, adjusted R-squares, parsimony and substantive theoretical interest, I would prefer Model 3, Table 3.

BUT, the above discussion is an orientation for graders about the details of the analyses, and things to look for in the very best of possible answers. It is not a statement of what is expected for students who show the required level of competence. Competent students should show basic understanding of the results, reach a justifiable conclusion, and give some justification for it. The best students should struggle with the issues above to various degrees.

Problems with one of the models, the data, and their relationships

What about problems with the model and its fit to these data? Obviously, we could not evaluate these without choosing a model, and the first part of this question asks students to choose a model. So the diagnostic figures which are presented had to be for an arbitrary model, and they are figures based on the most basic model, Model 1, which fits the data only slightly less well than the most preferable models.

Figures 1 and 2 display the distributions of the residuals for Model 1, Tables 1 and 4. These should have a normal distribution. The figures show that the distribution is not perfectly normal, perhaps slightly negatively skewed, with problems at the tails, but not too seriously different from a normal distribution.

Figure 3, showing a scatterplot of the absolute values of the residuals by the unstandardized values of the dependent variable is, of course, an indicator of versus homoscedasticity versus heteroscedasticity. Of course, it would be difficult to evaluate changes in the variance of the residuals across the horizontal dimension based on the data points alone, but the degree to which the regression fit or Loess fit is horizontal, is an indicator. From this a reasonable conclusion is that some heteroscedasticity is apparent but not to too serious.

Figures 4 – 10 show box plots of measures of influence for individual cases. Students should recognize the reason for producing such figures for these statistics. They should recognize that
these statistics show that no individual cases make important differences for the model parameters. However, for finding data errors and limiting the effects of extreme influential cases, it is wise to check out the most deviant cases. These are surely cases 1,410, 2208, 1511, 362, and 171 in the model overall. It would be interesting to see student arguments about influential cases relative to specific independent variables, as indicated by the DFBETAs. But the expectation here is that students show recognition of what the figures are about, a basic ability to evaluate the impact of the most influential cases, and the ability to check out the most suspicious cases, even if their impact is minimal.

**QUESTION BLOCK III: LOGISTIC REGRESSION**

(The total weight for this section is approximately 35 percent)

A dichotomous variable has been computed from a three-variable scale of traditionalism regarding gender roles. Respondents with the most traditional gender role values were given codes of “1” while those with more liberal attitudes were given codes of “0.” A series of three binary logistic regression models were estimated for predicting this dependent variable. The results for these models are included in the tables and figures for Block III in the appendix. Some of these models include variables for the interaction between ethnicity and age (MAxAge) and ethnicity and years of education (MAxEDyrs).

3.1. Which model is the best model?

For this block of the questions, the answer is easy: The interactions between ethnicity and age and ethnicity and education were both significant, and there was no experimenting with functional forms of the main effects of age, education, and residence. The best model is the last one, that in Block 3. For this model, the Nagelkerke R-square was .172, not too bad for these kinds of variables and models, and a dichotomous treatment of the dependent variable.

3.2. Based on the series of models, describe the statistical significance, direction, and importance of the various effects.

In terms of having traditional attitudes towards gender roles: Males are more traditional than females; age does not show a significant difference (! net of education; education is by far most important and as should be expected, has a negative effect on gender role traditionalism; rural residence is significantly positively associated with gender role traditionalism. Interactions between ethnicity and both age and education are statistically significant: age had a stronger positive effect on traditionalism for ethnic Macedonians than for ethnic Albanians, and the negative effect of education was greater for ethnic Macedonians than for ethnic Albanians.

3.3. Based on the third model estimated, calculate the predicted logit, odds, and
probability of having traditional values for two individuals. Both individuals are males, 40 years old, with 12 years of education, and a value of 2 for the variable for “RURAL4,” the variable for the population size of the place of residence, BUT one individual is an ethnic Macedonian while the other is an ethnic Albanian.

Macedonian:

Logit  
-.972

Odds  
.378

Probability  
.274

Albanian:

Logit  
.030

Odds  
1.030

Probability  
.507

3.4. Which cases are the most influential cases for the third model, as observed in Figure 11? How important are these cases? Which other measures of influence might you want to examine, and why?

The most influential cases for the model overall, based on Cook’s D, are cases 1495, 1825, 1990, 431, 403, 58, 171, and 953, perhaps more. Due to the large sample size, none of these have large impacts on the model coefficients: But they do point out suspicious cases to be checked for data error.

One might also want to check the DFBETAs because these would indicate influential cases with regard to specific independent variables and their logistic regression coefficients.