

**Intermediate Social Statistics
Hilary Term 2008**

Final Exam

Raymond Duch and Tom Snijders

Oxford University

www.raymond Duch.com

Instructions

The final exam consists of the following six questions. Data for all of the questions are available on the class web site at www.raymond duch.com. You should answer the data analysis questions in a professional format similar to a published article. The data results should be summarized in properly formatted tables and graphs where appropriate. Note: the Stata commands you used to estimate models should be included in an **appendix** rather than in the text of your responses to the questions. You are expected to work independently on both the computer work and the write-up for this exam – any violation of this requirement will result in a failing grade. The exam is due Friday, Fourth Week of Trinity Term 2008 and should be turned into Dave Armstrong.

Question 1

- Give the basic assumptions of the Mokken model for nonparametric scaling of binary questionnaire items (you may formulate the assumptions verbally, but it is also allowed to do it with mathematical formulae).
- Describe the assumption of double homogeneity in this model.
- Give the definition of Loevinger's homogeneity coefficient for binary items in a Mokken scale.
- What are the thresholds for Loevinger's coefficient to indicate useful, medium, and good scalability?

Question2

Suppose that a number of questionnaire items are available that purport to measure the same underlying attitude.

- What determines in practice your choice between a Factor Analysis treatment and a Mokken Scale Analysis treatment of the data, to establish the scalability of these items in one common scale?
- What is the main underlying mathematical argument?

Question 3

Let $p = \Pr(Y=1)$

Where Y is a coin flip and

$Y = 1$ for a head 0 for a tail

Coin flips are independent and the probability of observing a head or a tail is the same at each trial.

1. What is the likelihood for the i th toss?
2. What is the likelihood function for p ?
3. Demonstrate how we would find the estimate of p that maximizes this likelihood function. You will need to provide all of the steps involved in finding this maximum employing the analytic method.

Question 4

The following log likelihood is for which statistical model? Explain in detail the modeling problems for which this statistical model would be most appropriate. Provide an explanation of how to interpret the coefficients in this estimator.

$$\ln L = \sum_{y_i=1}^N y_i \ln \Phi(\mathbf{x}_i \beta) + \sum_{y_i=0}^N (1 - y_i) \ln [1 - \Phi(\mathbf{x}_i \beta)] \quad (1)$$

Question 5

We have estimated an ordered logit model to explore how characteristics of a target state impact the success of economic sanctions imposed by the United States against that state. The dependent variable is a coding of the success of the sanctions from failed to successful. The table reports the logic coefficients from this model. The argument is that sanctions will be more effective when imposed on countries in which citizens have more freedom of expression and where the press is more free. Also, we think that economic sanctions will be more successful when imposed on countries that are heavily dependent on trade with the US.

Data: 125 sanction episodes in which the US imposed economic sanctions on a target country.

Dependent Variable: coded 1 for a failure, 2 for partial failure, 3 for partial success, 4 for success

Independent variables:

Freedom = scale of 0 to 10 with 10 being most free and 0 least free (mean = 3, sd = 2)

Press = scale from -10 to 10 with -10 indicating the least press freedom and 10 the most press freedom (mean = -3, sd=4)

Trade = (value of exports to the US from target +value of imports from the US from target)/(GDP of target) All denominated in 1980 dollars. (mean = .1, sd=.03)

Table 1: Results

	Coefficient
Freedom	0.23
Press	0.22
Trade	5.61
Cut point 1	-1.10
Cut point 2	-0.03
Cut point 3	1.40

1. What is the stochastic component of this model?
2. What is the systematic component of this model?

3. What is the difference in the expected probability of a sanction being a failure, a partial failure, a partial success, and a success when Trade increases from .1 to .15 and all other variables are held at their mean?
4. Although you don't have the values for the 125 cases for the dependent variable, you can estimate from the results what percentage of cases in the raw data fell into each of the 4 categories of the dependent variable, what are these percentages?
5. Comment on the strengths and weaknesses of this model in this situation

Question 6

In 2004, the European Election Study asked citizens of Great Britain a number of questions including the subset in the dataset `GB04.dta`. The variables in the dataset are:

labvote Whether the respondent voted for the Labour Party 1=labour Party vote; 2=vote for other parties.

age Respondent's age

urban Size of area in which respondent lives (1=rural, 2=small or middle size town, 3=large town)

lrself Left-right Self-Placement (1=left, 10=right)

hhincome Household income in pounds sterling

socclass Self-reported social class (1=working, 2=low-middle, 3=middle, 4=up-middle, 5=upper)

econ How is economy now compared to 12 months ago (1=a lot better, 5=a lot worse)

immig Whether respondent thinks immigration is the most important issue facing government (1=yes, 0=no)

1. Estimate a logit model that includes all of the variables mentioned above (with labvote as the dependent variable). Interpret your results. Make sure to justify decisions to either keep variables like class, urban and econ as continuous variables or treat them as categorical. Do your results depend on these decisions?
2. How well does this model fit the data? What sorts of things would you want to look at to find out?

3. What values would the person most likely to vote labour have on the independent variables in the model? What about the person least likely to vote for labour? Estimate the probability of voting labour for each of these individuals. Are they what you expected?
4. All else equal, which of the following people would the labour party rather see turnout to vote:
 - (a) a middle-class person from an urban area with a left-right self-placement of 5
 - (b) a working-class person from a rural area with a left-right self-placement of 2
5. Create a graph showing how the probability of voting labour changes as left-right self-placement changes when immigration is the respondent's most important issue versus when immigration is not the most important issue, holding all other values constant at their median values.
6. Which of the variables mentioned above has the biggest effect on labour vote? What are the things you need to keep in mind when trying to answer this question?

Question 7

In 2006, the US General Social Survey asked a battery of questions about how scientific individuals thought various physical and social sciences were. Here, we are asking you to model peoples views on sociology. The variables in the dataset `gss06.dta` are as follows:

socsci How scientific is sociology (1=very, 2=pretty, 3=not very, 4=not at all)

knwsci Respondent's level of knowledge about science and technology (1=very informed, 2=somewhat informed, 3=neither informed nor uninformed, 4=somewhat uninformed, 5=very uninformed)

age Respondent's age

polviews Think of self as liberal or conservative (1 extremely liberal, 7 extremely conservative)

income Total family income

sex Respondent's sex (1=male, 2=female)

1. Estimate an ordered logit model that tests the following hypotheses. Include all variables in the model at the same time.
 - I know science and this ain't science: people who know science very well think sociology is less scientific than those who do not know science all that well.
 - Liberals don't know science: those squishy liberals who believe in global warming think everything is scientific so the more liberal you are, the more scientific you should think sociology is.
 - Men are from Mars: men are more likely to think things are scientific than are women, so men should be more likely to think sociology is scientific.
 - Mo' money: people with higher incomes will think that sociology is less scientific.
 - Age of innocence: younger people will be more likely to think that sociology is scientific than older people.
2. Interpret the results from the above model with respect to the hypotheses above.
3. Test the hypothesis that the coefficients on age and income are simultaneously zero.
4. Given what you know about the model coefficients what is the profile of the person who would be most likely to say sociology is very scientific? What is the profile of the person who would be most likely to say that sociology is not at all scientific? Estimate the probability of being in each of the four categories for each of these profiles (you

can use Stata to help do this). Present these findings. How reasonable do you think these predictions are? Does it make sense for us to make predictions for these profiles? Why or why not?

5. Generate a graph that shows how predictions change as a function of political views holding all other variables constant at their median values.
6. How well do you think this model helps us understand why people have the feelings they do about the “scientificness” of sociology? Why?

Question 8

In the dataset `bumps2.dta` are data compiled from the Oxford Summer Eights races from 2005. For those unfamiliar with the Pimm's-induced fury of splashing and cheering that is the "bumps," I'll explain. Ever year, there is a series of races where crews start at the same time, in a staggered fashion and the goal for each team is to bump the boat in front of them. The data represent some characteristics of each college along with the number of net bumps given by the college. If the college was a net loser, it scores zero. If a college bumped the boat in front of it twice and was bumped once, it scores a 1. The variables in the dataset are as follows:

college Name of the college

founded The year the college was founded

ugrads The number of undergrads in the college

grads Number of graduate students in the college

totstu $ugrads + grads$

endow Total college endowment in thousands of pounds.

crews The number of crews fielded by the college

bumps Number of bumps given by each college

1. Just looking at the bumps variable, without running a statistical model, is the bumps variable over-dispersed? How do you know? What does this mean about the relative utility of the Poisson model versus the Negative Binomial model for these data?
2. Estimate a Poisson model for the bumps variable and test the following hypotheses separately. For each hypothesis, interpret the results and the extent to which the hypothesis is supported by the data. If the hypothesis is not supported, can you provide an explanation why?
 - Experience matters - so colleges founded earlier should have more bumps.
 - Spirit matters - Graduate students are too nerdy to care about such things, so the percentage of the college's student population that is undergraduates should have a positive effect on bumps.
 - Size matters - so colleges with bigger overall student populations should have an easier time fielding a quality team.
 - Money matters - Schools with bigger endowments should be able to pay for better gear so schools with more money should have more bumps.

- Team size matters - colleges that field more crews should have higher numbers of bumps.
3. Estimate a poisson that includes all of these variables together. Interpret the results.
 4. Estimate a negative binomial regression. How do these results differ from the previous ones?
 5. Generate a graph that shows, on average, how well each model fits the data.

Question 9

In the dataset `civwardurat2.dta`, you are given data on the duration of civil wars and some basic characteristics about those wars. The variables are as follows:

warname Name of War

warst First year of war

interven Whether there was outside intervention in the war (1=yes, 0=no)

wardays Duration of civil war in days

westhem Whether war occurred in the Western Hemisphere (1=yes, 0=no)

europa Whether war occurred in Europe (1=yes, 0=no)

africa Whether war occurred in Africa (1=yes, 0=no)

midwest Whether war occurred in the Middle East (1=yes, 0=no)

asia Whether war occurred in Asia (1=yes, 0=no)

oceania Whether war occurred in Oceania (1=yes, 0=no)

type War type (1=Fight for control of central government, 0=Fight over local control)

1. Present a graph that shows what the empirical survival curve looks like (i.e., the one we would get if we don't want to make distributional assumptions about the survival times)?
2. Now, estimate an exponential survival model without covariates and present the survival curve from that model. How does this compare to the one from the previous model?
3. Estimate a survival model to test the hypothesis that more recent wars last longer than those from the more distant past. How would you interpret the regression results? Is the Exponential distribution appropriate here or should you use the Weibull distribution? Use this type of model (either Exponential or Weibull) for the remainder of the analysis.
4. Estimate a survival model to test the hypothesis that wars in which other countries intervene are shorter in duration. Interpret your findings. Have Stata draw the survival curve for wars where individuals intervene versus those where no one intervenes.
5. Estimate a survival model that allows you to test for regional differences in civil war duration. Interpret the findings. Do you find that there are any regional differences?

6. Estimate a survival model that allows you to test whether wars over central government control are longer than those concerning local control. Interpret your findings. What is the expected duration of central government control wars versus those concerning local control?
7. Finally, estimate a model that includes the type of war, intervention and the year the war started. How do these results differ from the ones generate above. What is the expected duration of a war that start in 1850 had outside intervention and was over local control versus a war that started in 1970 without outside intervention and concerned control of the central government?