SOS3003
Examination question 3

Fall 2009

Erling Berge

FALL 2009 QUESTIONS

- The questions use data from Malawi collected during field work in 2007. The data come from long interviews and questionnaire forms collected from 270 households plus some additional informers. The data also comprise trust game data from 267 pairs of players. In the present questions we use data from the trust game. More on the sample and variables is presented below.

QUESTION 3

- Among the questions about trust there was one simple binary question: “Generally speaking, do you think most people can be trusted or that they cannot be trusted?” Those who answered “yes, most people can be trusted” were coded 1 on our dependent variable “Trust”, and those who did not were coded 0. There was one missing.
- To investigate the correlation between “Trust” and the underlying dimensions of trust investigated in the previous question, we will run a logistic regression with Trust as dependent variable. Three models were estimated. The results are presented in the tables for question 3.
- **a) Determine if sex contributes significantly to the model of Trust. Find a 95% confidence interval for the direct effect of sex in model 3**

a) Does sex contribute significantly to the model of Trust?

- Three models of Trust have been estimated. In model 2 Sex appears alone in addition to the mistrust indexes that were introduced first in model 1. The Wald statistic of sex is 0.379 in this model has a p-value of 0.538. **Sex cannot be said to contribute to this model.**
- The same message follows from the ChiSquare statistic for the difference between model 2 and 1. This is given as 0.382 in the omnibus test of model coefficients (step 1, block) with a sig. level of 0.536.

a) Sex in model 3

- Sex also appears in model 3 as an interaction term for the mistrust indexes. The p-value of Sex drops a bit and one of the interaction terms is clearly significant. To test if the group in total contributes to the model we need to use a likelihood ratio test comparing model 3 to model 1

a) Model 3 against model 1

- If we compare one big model with K parameters to one small model with H fewer parameters (the big model has H more variables) the test statistic

$$\chi^2_{H} = -2\{\log_e \mathcal{L}_{K\text{-}H} - \log_e \mathcal{L}_{K}\}$$

$$= -2\log_e \mathcal{L}_{K\text{-}H} - \{-2\log_e \mathcal{L}_{K}\}$$

- will follow a ChiSquare distribution with H degrees of freedom.

a) Likelihood ratio test

- If the ChiSquare is large it would seem unlikely that the null hypothesis of no contribution from the H new variables is true.
- In this test we will find the following statistics useful:

Model	-2 Log likelihood	K
0	139.987	1
1	114.647	5=1+4
2	114.265	6=5+1
3	100.413	10=6+4

a) Concluding on sex

- Comparing model 3 with model 1 we have H=5 and K=10. Hence the
- χ^2_5 = 114.647 - 100.413 = 14.234
- In the ChiSquare distribution with 5 degrees of freedom the critical value for 0.05 level of significance is 11.07. If the null hypothesis is true, finding a value of 14,234 or larger has less probability than 0.05. We do not believe the null hypothesis is true in this case and will instead believe that Sex and the interaction terms do have an impact on the general probability of trusting people.

Confidence interval for sex in model 3

- A 95% confidence interval for the direct effect of Sex in model 3 can be found if we assume the sample is large enough that the distribution of Wald statistic follows a ChiSquare distribution. Then t = $b_{sex}/SE_{b_{sex}}$ = SQRT(Wald) follows the normal distribution.
- Large enough must here mean at least above 100 observations. Model 3 has 10 parameters, i.e. K=10, and n=102. It is thus a borderline case. But let us here assume this to be large enough

Confidence intervals

- Then a (1-α) confidence interval for the population parameter β_k from a model with K parameters estimated on n cases is found as

$$b_{Sex} - \tau_\alpha * SE_{b_{Sex}} < \beta_{Sex} < b_{Sex} + \tau_\alpha * SE_{b_{Sex}}$$

- where τ_α is the critical value from the Normal distribution. The critical values of the normal distribution do not depend on sample size or degrees of freedom. In the table of the Normal distribution and 0.05 level of significance we find that the critical value is 1.96.

Finding the confidence interval

- With b_{sex} = -0.396 and $SE_{b_{sex}}$ = 0.514 we find
- $-0.396 – 1.96* 0.514 < \beta_{Sex} < -0.396 + 1.96*0.514$
- $-1.40344 < \beta_{Sex} < 0.61144$
- Since the interval includes zero we understand that the null hypothesis of no direct effect of Sex, is true with a probability of at least 0.95

Question 3

- **b) Write up the equation that will produce the probability for saying “yes, most people can be trusted” as function of mistrust to modern authorities (MistMA244) in a conditional effect plot that will minimize predicted probabilities for women, also likewise write up the equation that will maximize predicted probabilities for men.**

3 b)

Variable	B	To minimize probability for women take variable value	To maximize probability for men take variable value
MistLoca244	-1.032	MAX	MIN
MistOuts244	-1.770	MAX	MIN
MistTA244	1.397	MIN	MAX
MistMA244	-.049	MistMA244	MistMA244
Sex	-.396	SEX=0	SEX=1
SexMistLo244	-.180	0	MIN
SexMistOut244	.992	0	MAX
SexMistTA244	-2.010	0	MIN
SexMistMA244	-.206	0	MistMA244
Constant	-.049		

3 b)

- The equation for the Logit will be
- L = -0.049 – 1.032*MistLoca244 - 1.770*MistOuts244 +

  1.397*MistTA244 –

  0.049*MistMA244 –

  0.396*Sex –

  0.180SexMistLo244 + 0.992*SexMistOut244 –

  2.010*SexMistTA244 –

  0.206*SexMistMA244

3 b) Variable information

	N	Minimum	Maximum	Mean	Std. Deviation
Mistrust of locals 244 cases = MistLoca244 [a]	103	-2.25542	2.83231	-.0395262	.94261378
Mistrust of outsiders 244 cases = MistOuts244 [a]	103	-2.59681	2.71402	.0654729	.98419632
Mistrust of traditional authorities 244 cases = MistTA244 [a]	103	-2.33551	3.09894	.0444267	1.05185159
Mistrust of modern authorities 244 cases = MistMA244 [a]	103	-3.20264	2.31673	-.0050544	1.01545742

3b) Logits for women and men

- For women the equation for the logit is

  $L = -0.049 -1.032*2.83231 -1.770*2.71402 +1.397*(-2.33551) -0.049*\text{MistMA244}$

- For men the equation for the logit will be

  $L = -0.049 -1.032*(-2.25542) -1.770*(-2.59681) +1.397*3.09894 -0.396 -0.180*(-2.25542) +0.992*(2.71402) -2.010*(-2.33551) - 0.255*\text{MistMA244}$

- To find the conditional probabilities we insert the logit into the equation

  $\Pr(Y_i = 1) = 1/(1 + \exp[-L_i])$

Question 3 c

- **c) For model 3 discuss possible deviation from the assumptions necessary for obtaining trustworthy parameter estimates.**

- A logistic model can be estimated by the maximum likelihood method, and valid inferences can be made if the following assumptions are met:

3 c) Assumptions

- The model is correctly specified, i.e.:
  - All conditional probabilities for Y=1 are logistic functions of the x-variables (this means the logit is linear in its parameters)
  - There are no irrelevant variables included in the model
  - There are no relevant variables excluded from the model
- All independent variables have been measured without errors
- All cases are independent

3 c) Assumptions

- In addition it should be observed that the method also require
  - No perfect multicollinearity
  - No perfect discrimination
- And that the precision of the estimates are affected by
  - High degree of multicollinearity
  - High degree of discrimination
  - Small sample

3 c)

- If the assumptions are met, the estimates of the parameters will be unbiased, efficient (minimum variance) and normally distributed. The likelihood ratio test can be used and in large samples b_k/ SE_{b_k} will asymptotically follow a normal distribution.
  - We cannot test if all relevant variables have been included.
  - We cannot test if variables have been measured without errors.
  - We cannot test if all cases are independent.

3 c)

- It is possible to test if the logit is linear in its variables. But there is not presented sufficient information here.
- From the p-values for the coefficients of model 3 we see that MistMA possibly is an irrelevant variable.
- There is some degree of multicollinearity due to the introduced interaction terms, but not to a degree that affects our conclusions here. The same may probably be the case for discrimination, but we know even less of this.

3 c)

- The most important problem is probably the small sample. With 102 cases and 10 parameters to estimate we are operating close to the lower boundary according to the literature. Hamilton (page 225) advices that n-K >100, but if the distribution of Y is skewed it might be necessary with a considerably larger sample.

3 c) Sample size (added April 2010)

- Sample size calculation for logistic regression appears to be a complex problem. However, the simulation study by Peduzzi et al. (1996) [1] suggests the following guideline for a minimum number of cases to include in a study.
- Let p be the smallest of the proportions of negative or positive cases in the population and k the number of covariates (the number of independent variables), then the minimum number of cases to include is:
- N = 10 k / p   or 10*9/ 0.463 = 194
- The model is estimated on 122 cases.

[1] Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 49 (12):1373-1379.

3 c) Other problems

- Problems of multicollinearity and discrimination are also basically problems caused by too small samples.
- One frequent consequence of small samples is influential cases. In the present case this can be investigated.
- The analog to Cook’s influence statistic picks out 6 cases outside 1.5 IQR from the mean. The leverage statistic finds 5 cases in this way

3 c) Influence table (continues)

- Looking at case 1011 that is on both lists plus the case that is highest on either list (excluding 1011) we find one man and one woman with some influence.

Case no	1011	3017	3036
Generosity	-60	-40	-40
MatriMatri	0	0	0
MatriPatri	0	0	0
PatriPatri	1	1	1
OtherMarri	0	0	0
OwnMattr	0	0	1
OwnRadio	1	1	1
Sex	1	1	0
Age	24	29	32
North	1	1	1
Central	0	0	0
South	0	0	0

3 c) Influence table (end)

Case no	1011	3017	3036
MistOuts266	-2.4306	2.2644	-0.1260
MistLoca266	1.0947	-2.1736	0.3341
MistOuts244	-1.8930	2.7022	0.0330
MistTA244	-1.0851	-0.9336	-1.3187
MistLoca244	1.4884	-1.6699	0.8303
MistMA244	-1.5302	-2.0398	0.1697
PREDPROB	0.5692	0.6384	0.0565
PREDGROUP	1	1	0
COOKsINFLU	0.6237	0.2781	1.2577
LEVERAGE	0.3206	0.3293	0.0700
RESIPROB	-0.5692	0.3615	0.9434
RESILOGIT	-2.3212	1.5662	17.6854
RESIstand	-1.5745	1.1567	2.4856
RESInorm	-1.1494	0.7525	4.0847
DEVIANCE	-1.2977	0.9473	2.3969