SOS3003
Examination questions

Spring 2004

Erling Berge

Question 1 (OLS-regression, weight 0,5)

- In a study of the legal system the impact of having victims of crime in the family on trust in the legal system has been investigated in a multivariate framework by controlling for the impact of other variables. This is done by OLS regression. The dependent variable is "TRUST IN THE LEGAL SYSTEM". The variable reports the respondent's opinion on a scale from 0 = "no trust at all in the legal system" to 10 = "complete trust in the legal system". Five control variables are introduced sequentially. Some results from this analysis are included in appendix tables for question 1.

a) Discuss the relation between "having victims of crime in the family" and "trust in the legal system" as revealed by these regression results
b) Compute a confidence interval for the regression coefficient of "having victims of crime in the family" with a significance level of 0.01. Test if "employment status" makes a significant contribution to the model
c) Construct a conditional effect plot of the impact of country in model 6
d) Formulate the complete model estimated
e) Discuss the degree to which the assumptions of an OLS regression have been met
f) Discuss any indications of problems related to multicollinearity and influential cases

Dependent Variable: Trust in the legal system	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6
(Constant)	5,568	6,014	5,425	5,493	5,020	5,579
Victim of crime in IP's family	**-,187**	**-,099**	**-,142**	**-,157**	**-,199**	**-,187**
Safe after dark		-,431	-,144	-,131	-,136	-,146
Unsafe after dark		-,978	-,504	-,474	-,445	-,457
Very unsafe after dark		-1,803	-1,225	-1,166	-1,088	-1,121
Spain			-,812	-,789	-,714	-,685
Sweden			,858	,852	,913	,908
Norway			1,071	1,055	1,049	1,044
Selfempl				-,025	-,007	,005
Notempl				-,187	-,049	-,191
Education in years					-,003	,013
Education in years squared					,003	,002
Age in years						-,034
Age in years squared						,000089

a) Discussion I

- In the last and most comprehensive model we see that on the 10 point scale of trust to the legal system, the trust declines with 0.187 points if the person has a victim of crime in the family after control for the impact of feeling safe walking in the neighbourhood after dark, home country, employment status, education, and age. The effect is significantly different from 0 at level 0.01

a) Discussion II

- From model 1 we see the bivariate relation between the two variables is -0.187 exactly as in model 6 where 5 other variables has been controlled for. In model 2 the coefficient is up to -0.099 after the introduction of the variable “feeling safe walking in the neighbourhood after dark” and it is down to -0.199 in model 5. However, in model 2 it is not significantly different from 0 at test level 0.05

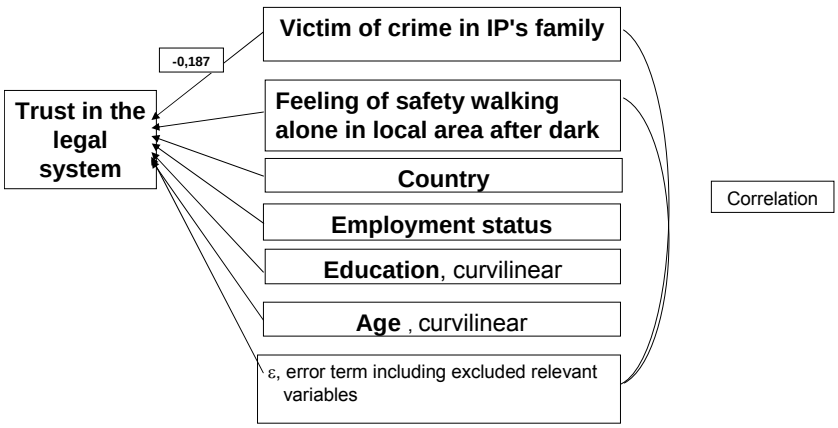
a) Discussion III

- In model 1 all variables except “crime victim in the family” are excluded from the model. If excluded variables correlate with included variables at the same time as they have an impact on the dependent variable, they are relevant for the model and need to be included. The relatively large change in the regression coefficient (from -0.187 to -0.099) after control for “feeling safe …” shows that either there is a correlation between the two variables (“felling safe …” and “victim of crime …”) or that they both correlate in some way with excluded variables.

a) Discussion IV

- The tolerance of “victim of crime …” is all the time high, and even in model 6 it is over 0.96. This indicates that it is not the correlation between “victim of crime …” and any of he other variables in the model that is the reason for the changes in effect. The most reasonable explanation is that there are excluded variables not included in any of the 6 models.

Conceptual model



b) Compute a confidence interval for the regression coefficient of "having victims of crime in the family" with a significance level of 0.01. Test if "employment status" makes a significant contribution to the model

- $b_{\text{Victim of crime}} - SE_{\text{Victim of crime}} * t_{1\%} < \beta_{\text{Victim of crime}} < b_{\text{Victim of crime}} + SE_{\text{Victim of crime}} * t_{1\%}$
- $-0{,}187 - 0{,}062 * 2{,}576 < \beta_{\text{Victim of crime}} < -0{,}187 + 0{,}062 * 2{,}576$
- $-0{,}187 - 0{,}159712 < \beta_{\text{Victim of crime}} < -0{,}187 + 0{,}159712$
- $-0{,}347 < \beta_{\text{Victim of crime}} < -0{,}0273$

b) Test if "employment status" makes a significant contribution to the model

$$F^{H}_{n-K} = \frac{\frac{RSS_{[K-H]} - RSS_{[K]}}{H}}{\frac{RSS_{[K]}}{n-K}}$$

Model		Sum of Squares	df
1	Regression	49,511	1
	Residual	44253,295	7389
	Total	44302,806	7390
2	Regression	1708,098	4
	Residual	42594,709	7386
	Total	44302,806	7390
3	Regression	5198,063	7
	Residual	39104,744	7383
	Total	44302,806	7390
4	Regression	5254,469	9
	Residual	39048,337	7381
	Total	44302,806	7390
5	Regression	5680,968	11
	Residual	38621,838	7379
	Total	44302,806	7390
6	Regression	5782,884	13
	Residual	38519,922	7377
	Total	44302,806	7390

c) Construct a conditional effect plot of the impact of country in model 6

- Predicted Y = 5,579 -0,187*Victim of crime in IP's family -0,146*Safe after dark -0,457*Unsafe after dark -1,121*Very unsafe after dark -0,685*Spain +0,908*Sweden +1,044*Norway +0,005*selfempl -0,191*notempl +0,013*Education in years +0,002*Education in years squared -0,034*Age in years +0,000089*Age in years squared

c) Construct a conditional effect plot of the impact of country in model 6

- Predicted Y = **5,579 -0,685*Spain +0,908*Sweden +1,044*Norway** + const,

where

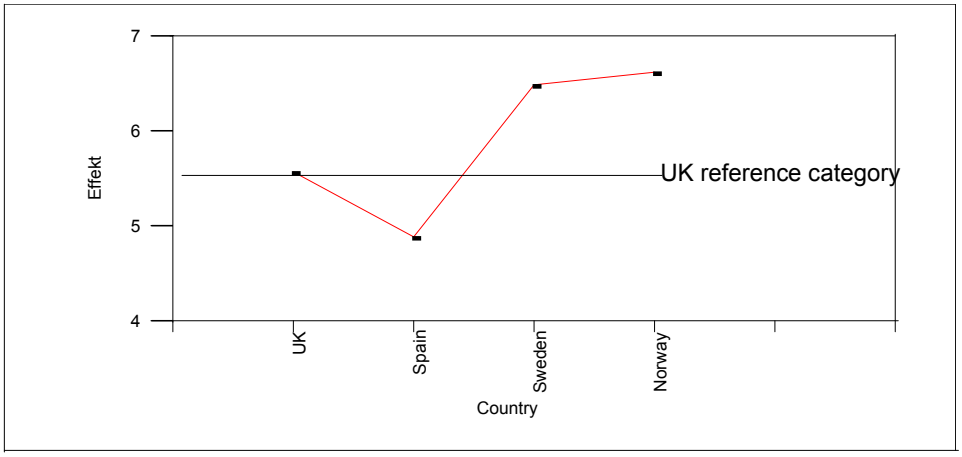
- const = -0,187*Victim of crime in IP's family -0,146*Safe after dark -0,457*Unsafe after dark -1,121*Very unsafe after dark +0,005*selfempl -0,191*notempl +0,013*Education in years +0,002*Education in years squared -0,034*Age in years +0,000089*Age in years squared

c) Construct a conditional effect plot of the impact of country in model 6

Impact of country on
" **Trust in the legal system** "

- Living in UK effect = const +5,579
- Living in ES effect = const +4,894
- Living in SE effect = const +6,487
- Living in NO effect = const +6,623

c) Construct a conditional effect plot of the impact of country in model 6



d) Formulate the complete model estimated

To define a model there are three types of elements that need to be considered:

- Definitions of the elements of the model (variables, error term, population and sample)
- Definitions of the relationships among the elements of the model (the equation linking variables and error term, the sampling procedure linking sample and population, theories and time sequences of events and observations linking causes and effects)
- Definitions of the assumptions that have to be met in order to use a particular method (such as the OLS method for linear regression) for estimating the model (model specification, distribution and properties of the error term)

At a minimum the formulation will include **variable definitions, the formula linking variables and error term, and the assumptions needed to make valid inferences** from the estimates of a particular procedure.

d)

- The data has been collected in 2002 by ESS in Great Britain, Spain, Sweden and Norway as simple random samples.
- We want to explain variation in trust to the legal system (Y) in the 4 countries.
- It assumed that there is a linear or curvilinear relation between the dependent variable and the independent variables defined above

d)

- All models in question 1 are regression models of the form
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} +...+ \beta_{13} x_{i13} + \varepsilon_i$
- where ”i” runs over the population. If we let k=0, 1, 2, 3, … ,13, β_k will be the unknown parameters showing how many measurement units of y will be added to y per unit increase in X_k
- ”ε_i” is the error term, a variable that comprises all relevant factors not observed as well as random noise in the measurement of y. The 13 x-variables are defined in the model 6 table and the section of variable definitions

d)

Y	Trust in the legal system
X_1	**Victim of crime in IP's family**
	Feeling of safety walking alone in local area after dark
X_2	Safe after dark
X_3	Unsafe after dark
X_4	Very unsafe after dark
	Country
X_5	Spain
X_6	Sweden
X_7	Norway
	Employment status
X_8	selfempl
X_9	notempl
	Education
X_{10}	Education in years
X_{11}	Education in years squared
	Age
X_{12}	Age in years
X_{13}	Age in years squared

d)

- An OLS (ordinary least squares) estimate of the model parameters defined above can be found as the b-values of
- $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + ... + b_{13} x_{i13}$
- that minimizes the sum of squared residuals,

$$RSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} e_i^2$$

(For "$\hat{y}_i$" read "estimated" or "predicted" value of y_i or just "y-hat")

d)

- In the four countries 7816 persons have been interviewed
- There are missing information for many persons on may variables leaving after listwise deletion 7391 cases for this analysis
- There are no reasons to believe that the missing cases are anything but random in relation to Y

Spain	1729
United Kingdom	2052
Norway	2036
Sweden	1999
Total	**7816**

d)

- OLS estimates will be unbiased and efficient with a known sampling distribution if the following assumptions are true:
- I: The model is correct, that is
  - All relevant variables are included
  - No irrelevant variables are included
  - The model is linear in the parameters

d)

- II: The Gauss-Markov requirements for "Best Linear Unbiased Estimates" (BLUE)
  - Fixed x-values (no random component in their measurement)
  - The error terms have an expected value of 0 for all cases "i"
    - $E(e_i) = 0$ for all "i"
  - The error terms have constant variance for all cases "i" (homoscedasticity) for all "i"
    - $var(e_i) = \sigma^2$ for all "i"
  - The error terms do not correlate with each other across cases (no autocorrelation) for all "i" ≠ "j"
  - $cov(e_i , e_j) = 0$ for all "i" ≠ "j"

d)

- III: The error terms are normally distributed
  - The error terms are normally distributed (and with the same variance) for all cases for all "i"
    - $\varepsilon_i \sim N(0, \sigma^2)$ for all "i"

- Inferences from a sample to a population can be obtained with a known confidence if the estimates come from a simple random sample from the population of interest.

1 e) Discuss the degree to which the assumptions of an OLS regression have been met

Some of the stated assumptions cannot be tested. In particular we cannot test if

- All relevant variables are included
- Variables are without measurement error
- The error term in reality has mean 0 and variance 1

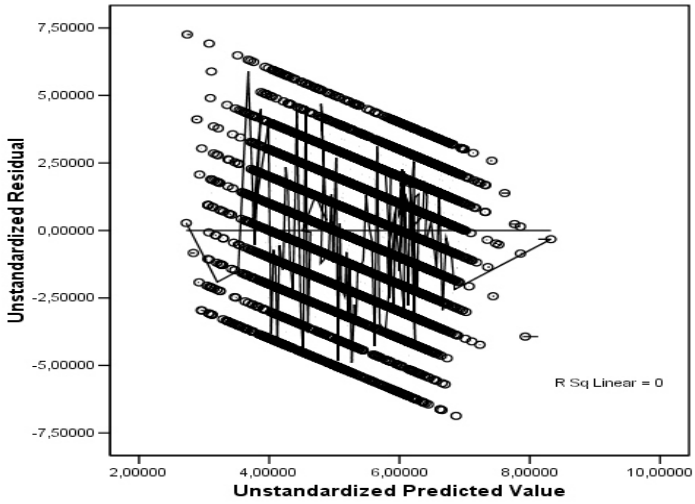
We can test if

- irrelevant variables have been included in the model
- the model is curvilinear in the included variables
- there is heteroscedasticity and/ or autocorrelation
- the error term is normally distributed

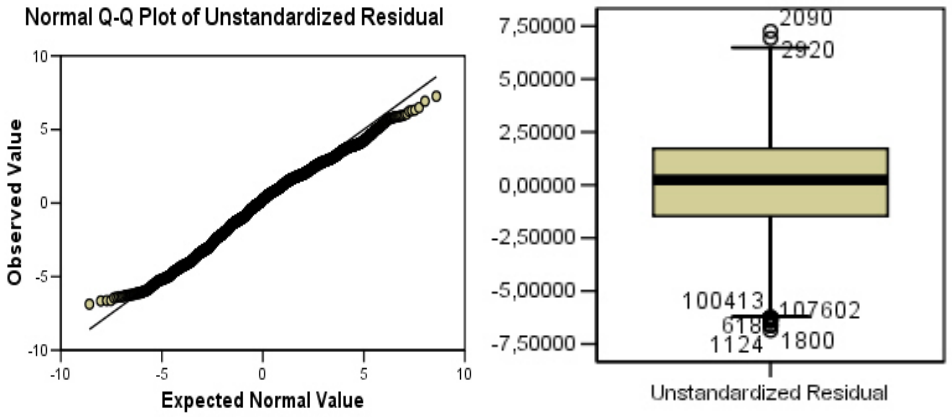
1 e) Model specification: curvilinearity

- In the table for 1a we see that both education and age are included as second degree polynomials. We see that both elements on both variables are significantly different from 0 at test level 0.05. Then further testing is unnecessary
- No irrelevant variables
- No autocorrelation

e) Heteroscedasticity?
<the line is made in “Chart Editor” by choosing ”Add Chart Element” and ”Interpolation Line” and then ”Spline”>



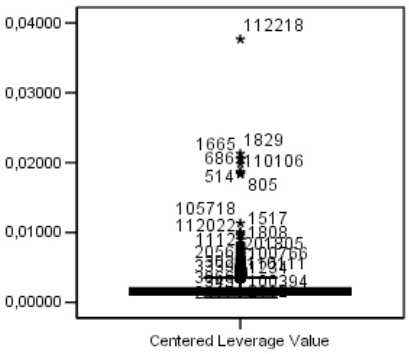
1 e) Normally distributed residual?



1 f) Discuss any indications of problems related to multicollinearity and influential cases

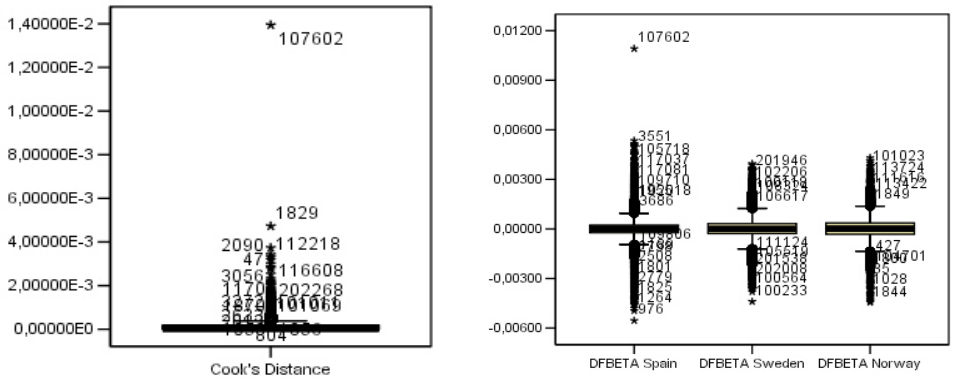
Model 6	variables	tolerance
	Victim of crime in IP's family	,965
	Safe after dark	,722
	Unsafe after dark	,718
	Very unsafe after dark	,809
	Spain	,617
	Sweden	,642
	Norway	,620
	selfempl	,928
	notempl	,610
	Education in years	,054
	Education in years squared	,059
	Age in years	,030
	Age in years squared	,027

1 f) Influential cases: leverage



- Leverage
- High potential for influence do not always give large influence on the regression results

f) Influential cases: DFBETAS



•DFBETAS

f) One case with large influence?

- In box plots of Cook's distance, DFBETAS, and centered leverage, there are 2 cases with values clearly larger than the rest. This is case no 107602 with large values for both Cook's D and DFBETAS for the variables Spain and Education in years and 112218 with high value on centered leverage. The high leverage do not translate into real influence

f) 3 cases with a residual larger than 3 standard deviations away from the mean

Case Number /id no	Std. Residual	Trust in the legal system	Predicted Value	Residual
854 /2090	3,029	10	3,08	6,923
1188 /2920	3,177	10	2,74	7,259
2330 /107602	-3,006	0	6,87	-6,868

Case 107602 is the only case that ought to be investigated further by running regressions with and without this case