EKSAMENSOPPGÅVER SVSOS3003
Vår 2010
FRAMLEGG TIL LØYSING

Erling Berge
Institutt for sosiologi og statsvitenskap
Norges Teknisk Naturvitskapelege Universitet

«Bruksanvisning»

Når ein går i gang med å løyse oppgåver må ein ha i minnet at oppgåvene ofte er problematiske i høve til modellbygginga sitt krav om at modellen må vere fundert på den best tilgjengelege teorien. Mangelen på teoretisk fundament for oppgåvene kan forsvarast ut frå to perspektiv. Det avgjerande er rett og slett mangelen på tid og høvelege data for å lage eksamensoppgåver av den «realistiske» typen det i eit slikt høve er tale om. Men tar ein for gitt at oppgåvene sjeldan kan seiast å vere teoretisk velfundert, gir jo dette studentane lettare gode poeng i arbeidet med å vurdere modellane kritisk ut frå spesifikasjonskravet.

Når ein studerer framlegga til løysingar er det viktig å vere klar over at det som er presentert ikkje er nokon fasit. Dei fleste oppgåvene kan løysast på mange måtar. Dei tekniske sidene av oppgåvene er sjølvsagt eintydige. Men i dei mange vurderingane (som t.d. «Er fordelinga av denne residualen tilstrekkeleg nær normalfordelinga til at vi kan tru på testane?») er det nett vurderingane og argumentasjonen som er det sentrale.

På eksamen er tida knapp. Svært få rekk i eksamenssituasjonen å gjere grundig arbeid på heile oppgåvesettet. I arbeidet med dette løysingsframlegget har det vore gjort meir arbeid enn det ein ventar å finne til eksamen. Somme stader er det teke med meir detaljar i utrekningar og tilleggsstoff som kan vere relevant, men ikkje nødvendig. Men det er ikkje gjort like grundig alle stader.

Det må takast atterhald om feil og lite gjennomtenkte vurderingar. Underteikna har like stor kapasitet til å gjere feil som andre. Kritisk lesing av studentar er den beste kvalitetskontroll ein kan ønskje seg. Den som finn feil eller som meiner andre vurderingar vil vere betre, er hermed oppfordra til å seie frå
(t.d. på e-mail: <Erling.Berge@svt.ntnu.no>)

Permitted helpful materials in the examination:

Allison, Paul D. 2002. *Missing data.* No 136 Quantitative Applications in the Social Sciences. London: Sage**.**

Hamilton, Lawrence C. 1992. *Regression with graphics.* Belmont: Duxbury.

Hamilton, Lawrence C. 2008. A Low-Tech Guide to Causal Modelling. http://pubpages.unh.edu/~lch/causal2.pdf , Handout, 12 pages.

Berge, Erling. 2010. *SOS3003 Applied data analysis for social science: Collected lecture notes spring 2010.* Handout, 293 pages.

Norwegian-English / English-Norwegian dictionary

Calculator

Both textbooks and handouts may contain written notes.

It is NOT ALLOWED to bring examinations and grader's advice for previous years.

QUESTIONS 1 and 2

Questions 1 and 2 use data from Malawi collected during field work in 2007. The data come from long interviews and questionnaire forms collected from 270 households plus 13 key informers. More on the sample and variables is presented below.

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

In this question we explore some determinants of the size of a trust index constructed by principal components analysis from 16 variables expressing strength of trust in various institutions and groups of persons. From the 16 variables 4 components were identified and rotated by varimax to simple structure. They were interpreted to indicate 1) Trust in people outside the village, 2) Trust in traditional authorities, 3) Trust in people within the village, and 4) Trust in modern institutions.

This question uses as the dependent variable the index called "Trust in traditional authorities". It will for short be called "Trust". Without necessarily implying any causal structure, the expressed trust in traditional authorities is supposed to vary with behaviour in areas where compliance with traditional authorities can be observed. This includes taking care of churchyards and contributions to unpaid public work projects. Differences in culture and influence from urban living are controlled for by the regional location of the households. Structural determinants such as sex and age will be tested out together with indicators of wealth (owning mattress and owning radio). In the tables for question 1 seven nested models of *Trust* have been estimated.

a) Describe the impact of sex and age on the determination of trust in traditional authorities. Find a 99% confidence interval for the impact of *Participated in graveyard clearing project over the last 12 months* on *Trust* in model 1.

b) Present the formula for the F-test of the contribution of regional location of households on *Trust*, and find the quantities needed to perform the test. Find the critical value in the table of the F-distribution to secure a probability of 0.10 or less for doing a type I error. Discuss briefly how trustworthy the t-tests and F-tests are in this model.

c) Outline briefly the problem of influential cases. Based on the tables attached to this question what can be said about influential cases in this particular study?

d) The index of *Trust in traditional authorities* has 39 missing observations. Outline briefly the general problem of biased samples. Discuss in more detail the possibilities for having a biased sample in this particular study of *Trust*.

a)

Describe the impact of sex and age on the determination of trust in traditional authorities. Find a 99% confidence interval for the impact of Participated in graveyard clearing project over the last 12 months *on* Trust *in model 1.*

Sex and age, both linear and curvilinear, and their interactions are included successively in models 5-7. See excerpts from the tables below. The short answer to the question of the impact of sex and age on the level of the trust index is that there is no impact. The p-values (Sig.) shows that there is no linear or curvilinear effect of age, no effect of sex, and no effect from interactions of sex and age.

In model 6 and even more so in model 7 we see that the level of multicollinearity is high and will affect the precision of the t-test (confidence intervals will be wider). The F-change statistic for each new model shows also that the new variables do not contribute significantly to the model at the 5% level. But to be on the safe side we can do an F-test of Model 7 against model 4:

$$F_{n-K}^{H} = \frac{\dfrac{RSS_{[K-H]} - RSS_{[K]}}{H}}{\dfrac{RSS_{[K]}}{n-K}}$$

From the ANOVA table we find n=236, K=16, H=5

$RSS_{\text{model 4}}$ = 189.943

$RSS_{\text{model 7}}$ =187.944

From this we find F^{5}_{220}= 0.3998/0.8543 = 0.4680 while the α=0.05 critical value in the F-distribution with 5 and more than 120 degrees of freedom is 2.21. We cannot reject the null hypothesis of no impact from sex, curvilinear age and their interaction.

In discussing the impact of curvilinear age and interactions of sex and age be aware that one cannot infer anything from single coefficients. In model 7 for example the impact of sex involves both the coefficient for sex itself and the coefficients of interaction of sex and age and sex and age squared.

Models 5-7

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
5	Sex of respondent	.134	.143	.067	.941	.348	.720	1.388
	Age of respondent	.000	.004	.000	.006	.995	.874	1.144
6	Sex of respondent	.160	.342	.080	.470	.639	.126	7.909
	Age of respondent	-.001	.012	-.015	-.077	.939	.099	10.075
	Interaction of sex and age	.001	.008	.019	.084	.933	.070	14.211
7	Sex of respondent	-.192	.924	-.096	-.208	.835	.017	57.773
	Age of respondent	.049	.067	.804	.724	.470	.003	336.975
	Interaction of sex and age	-.017	.044	-.492	-.380	.704	.002	457.242
	Age of respondent squared	-.001	.001	-.819	-.740	.460	.003	333.646
	Interaction of age and age squared	.000	.000	.434	.388	.698	.003	340.907

Model 1 consists of only 2 variables. One variable tells if the household has contributed to graveyard clearing project over the last 12 months, the other how many days. We note that there is a degree of multicollinearity between these two. This will cause the confidence intervals to be a bit larger than otherwise. The test of the whole model suggests that the two variables contribute to the explanation of the variance in trust. The fact that *Number of days worked on graveyard clearing* is not significant alone may perhaps be due to multicollinearity. We see in model 4 that the level of multicollinearity is affected by the region dummies. The impact of number of days drops a bit and the variance of the participation dummy increases a bit. Based on the evidence here it is difficult to determine if *Number of days worked on graveyard clearing* is a relevant variable or not.

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.158	.083		1.908	.058		
	L8.k. Participated in graveyard clearing project over the last 12 months	-.560	.198	-.276	-2.828	.005	.432	2.315
	L8k Number of days worked on graveyard cleaning last 12 months (missing=0)	.149	.105	.138	1.414	.159	.432	2.315

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

$$b_k - t_\alpha * SE_{b_k} < \beta_k < b_k + t_\alpha * SE_{b_k}$$

where t_α is the critical value from the t-distribution with n-K degrees of freedom in a two tailed test with α level of significance.

We want to find a 99% confidence interval for the impact of “Participated in graveyard clearing project over the last 12 months” on Trust in model 1. Since (1-α) = 0.99, α = 0.01. From table A4.1 in Hamilton (1992) and with n – K = 236 – 3 = 233 degrees of freedom and α = 0.01, we find $t_{0.01}$ = 2.576. Let v = *L8.k. Participated in graveyard clearing project over the last 12 months*

From the parameter estimate of b_v = -0.56 and SE_{b_v} = 0.198 we find the confidence interval to be
$-0.56 - 2.576*0.198 < \beta_v < -0.56 + 2.576*0.198$
$-1.070048 < \beta_v < -0.049952$

This means that if the household contributes to graveyard clearing during the last year the trust in traditional authority declines between 0.05 and 1.071 index points. On the index the observed minimum is -3.10 and observed maximum is 2.34.

b)
Present the formula for the F-test of the contribution of regional location of households on Trust*, and find the quantities needed to perform the test. Find the critical value in the table of the F-distribution to secure a probability of 0.10 or less for doing a type I error. Discuss briefly how trustworthy the t-tests and F-tests are in this model.*

T-tests of single dummy variables tests if the effect of the dummy is different from the reference category. It does not say anything about the compound contribution of the group of dummies that comprise the substantive variable. To test the simultaneous contribution of a group of variables to the explanation of the variance of a dependent variable we use the F-test to compare two models, one model without the group of variables and then the same model with the group included as explanatory variables.

To answer this part of the question correctly the candidate has to provide the formula and determine the quantities the formula needs. Using the F-change statistic can just be seen as a check that it has been done correctly.

The F-statistic:

$$F_{n-K}^{H} = \frac{\dfrac{RSS_{[K-H]} - RSS_{[K]}}{H}}{\dfrac{RSS_{[K]}}{n-K}}$$

follows a F-distribution with H and n-K degrees of freedom if it is true that the H extra variables included in the big model have no effect (if H_0 “No impact of the new variables” is true) and the assumptions of OLS regression are met. In this formula the $RSS_{[K]}$ is the sum of squares of the residuals of the big model with K parameters (or K-1 variables) and $RSS_{[K-H]}$ is the sum of squared residuals in the small model where the H new variables are not included. We reject the null-hypothesis that the H new variables do not have an impact with level of significance α if F^{H}_{n-K} is larger than the critical value for level of significance α in the table of the F-distribution with H and n-K degrees of freedom.

The dummies of the region variable are included for the first time in model 4. There are 5 dummies with region Phalombe excluded as reference category. This means that H=5. We have already seen that n=236 and in model 4 K=11, hence n-K = 236 – 11 = 225.

	ANOVA	Sum of Squares	df	Mean Square	F	Sig.
3	Regression	17.210	5	3.442	3.667	.003(c)
	Residual	215.873	230	.939		
	Total	233.083	235			
4	Regression	43.140	10	4.314	5.110	.000(d)
	Residual	189.943	225	.844		
	Total	233.083	235			

From the ANOVA table we find that
RSS[K] = 189.943 and
RSS[K-H] = 215.873

The F^{5}_{225} = [(215.873 - 189.943)/5] / [189.943/225] =

25.93/(5*0.84419) = 6.143.

From table A4.2 in Hamilton (1992) we find that the critical value of 1.85 for 5 and more than 120 degrees of freedom will provide a test level of 0.1. We found F^{5}_{225} = 6.143 which is very unlikely given a true null hypothesis. We conclude that the regional location of the household is a significant part of the explanation for the variation in the trust index.

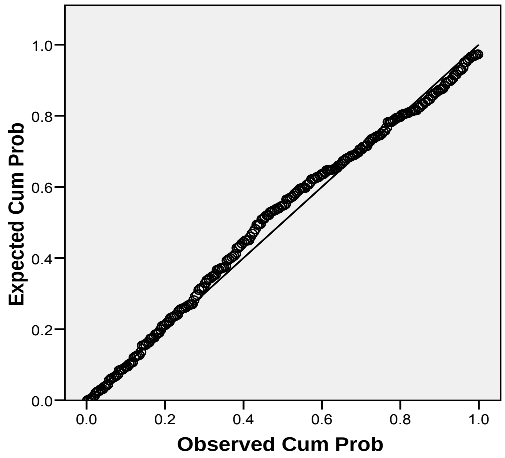
A test of our computations is found in the change statistics in the Model Summary table where the addition to model 4 provides a F-change value of 6.143 just as we determined here.

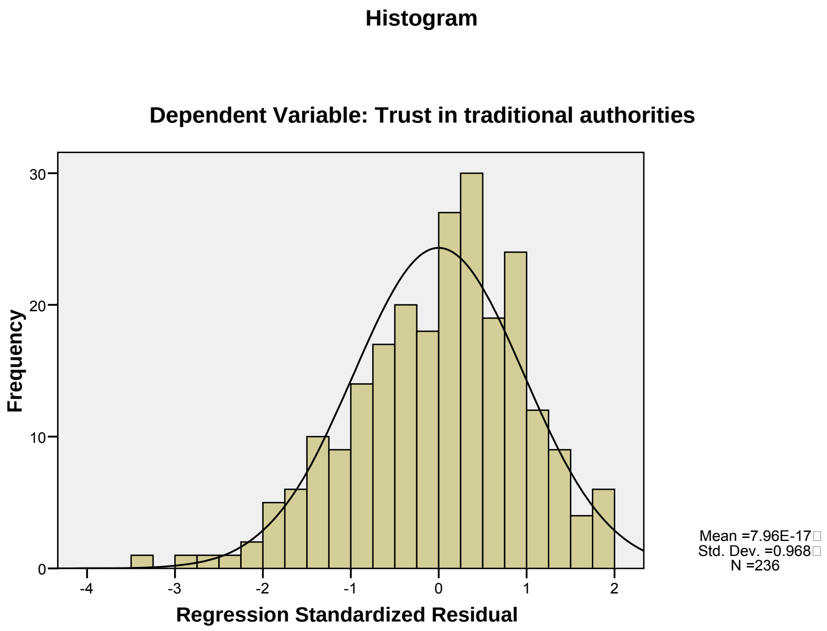
The validity of t- and F-tests is affected by excluded relevant variables, non-linear relationships, measurement errors in the variables, heteroscedasticity, autocorrelation, correlation between x-variable and error term, and non-normal distribution of the error term. To determine the validity of the tests, the distribution of the residual is the most important statistic to inspect. It can be seen as a general purpose tool to determine the validity of F-tests and t-tests.

In exploratory models like the ones presented here, it cannot be taken for granted that residuals are normally distributed. It must be determined empirically. We do not have information on the residuals for model 4, but for model 7 we have the following normal probability plot and histogram.

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Trust in traditional authorities





We see there are no large residuals and some more observations than ideal between 0 and 1. The normal probability plot also suggests small deviations from the normal distribution. On the whole it seems reasonable to conclude that the approximation to a normal distribution is acceptable. And since the variables entered in models 5-7 are clearly irrelevant and do not contribute to the explanations of the model, it seems fair to assume the residuals of model 4 will be distributed more or less like in the present figure. Based on this it seems reasonable to conclude that the F-tests and t-tests are trustworthy.

c)

Outline briefly the problem of influential cases. Based on the tables attached to this question what can be said about influential cases in this particular study?

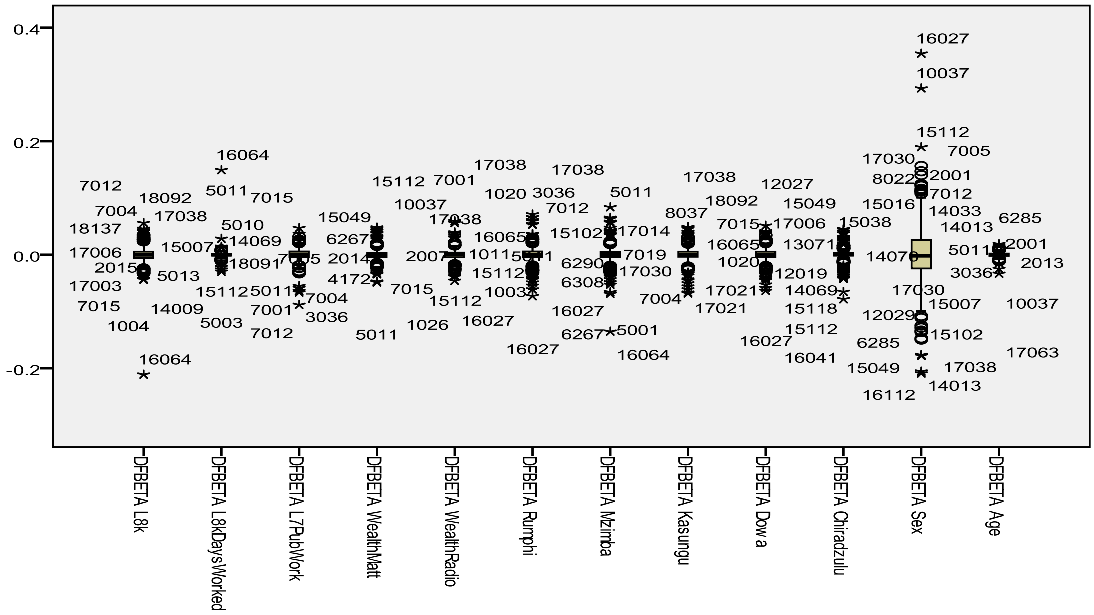
A case has influence on the regression results if its deletion substantially changes the result. A particular case may be suspected of having such influence on the regression results if it appears as an outlier either in terms of y-value of in some x-variable value. But in multivariate cases this is not easy to detect. Influence is often due to particular combinations of variable values.

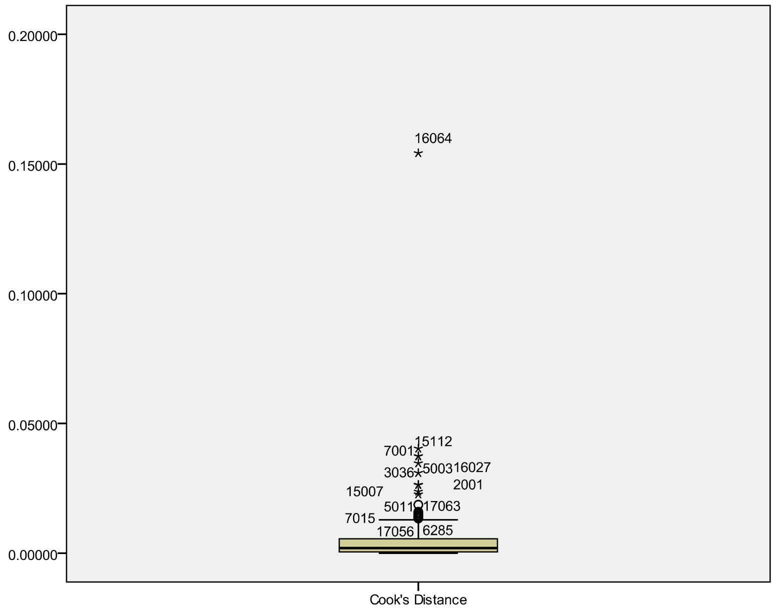
Sometimes a variable value is an error introduced in the data manipulations giving the case a value higher or lower than the others by an order of magnitude. Such cases should be removed. But most influential cases have valid variable values.

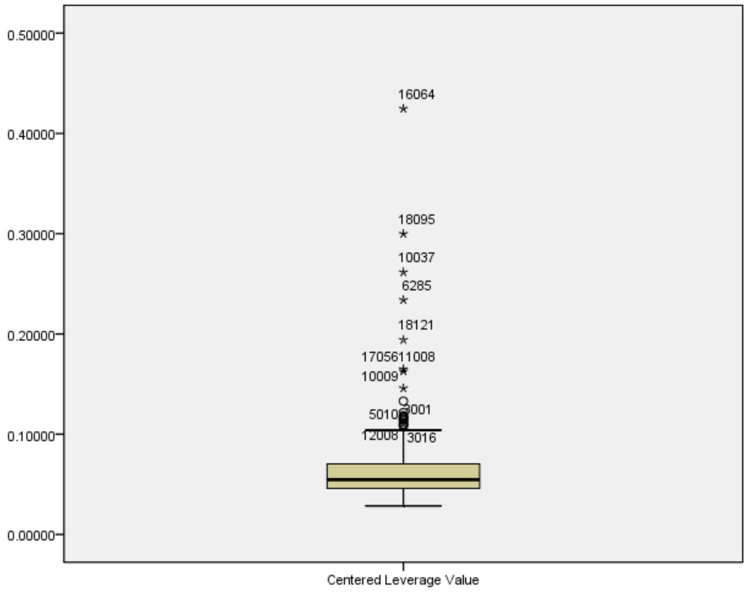
Sometimes a case with influence may indicate that there is a missing variable in the model. But more often it is simply due to too few cases in the data. In small samples single cases will often have large influence.

If influential cases with valid variable values are found, we may first consider possible excluded variables that might make the case less unusual, or if more cases can be obtained. If no such variable is available and a larger sample is impossible, we should report regression results both with the case included and with the case excluded.

The sample used for question 1 is small. We should expect influential cases. In the tables for the question we find box plots of influence statistics such as leverage, Cook’s D and DFBETAS. They are presented below:







Cook’s D indicates that case 16064 is the overall most influential case in the data. The leverage value says that it is influential due to some combination of values on x-variables, not due to an extreme value on y.

The criterion for identifying a high value on leverage given by Hamilton (1992) is h > 0.5. To assess the centered leverage values produced by SPSS we need to add the mean of h, which is equal to K/n where K is number of parameters in the model. In model 7 the mean of h is 16/236 = 0.068. Even without an exact value of the centred h value of case 16064, we see that it is less than 0.5. But it is definitely in the risky region between 0.2 and 0.5. In the table with case observations we see that h = 0.42 + 0.068 = 0.488.

In the box plots of DFBETAS case 16064 appears to have relatively high values for the variables L8k (participated in graveyard clearing), L8kDaysWorked (number of days worked on graveyard clearing), and the dummy for the district Mzimba. And the and case appears at the top of clear gaps.

In addition to this case we see that many DFBETAS for the variable sex are high and at the top of a gap for cases 10037 and 16027. The one thing in common for these cases is age. They are either young (16027 and 16064) or old (10037). Other than that nothing much can be concluded. The DFBETAS suggest that for variables other than sex the b_k’s change by up to 0.2 standard errors (standard deviations). To see how much this actually represents in terms of model estimate requires a re-estimation of a model where the cases 10037, 16027, and 16064 are dropped one by one.

The plot of leverage values lists case 18095 as the second highest on potential for influence. But it does not appear as influential in plots of DFBETAS or Cook’s D. Inspection of variable values for 18095 reveals a value of 60 in L8kDaysWorked even though the variable definition lists variable values from 0 to 7 as valid. Clearly this value is wrong. Probably it should have been 6. But since the case do not affect the regression results for either the whole model or the parameter for the variable it is not influential.

Variable values on selected cases

HHQIDNO	10037	14013	16027	16064	16112	18095
Trust	0.50	-1.40	1.73	1.04	0.98	-.08
Rumphi	0	0	0	0	0	0
Mzimba	0	0	0	0	0	0
Kasungu	0	0	0	0	0	0
Dowa	1	0	0	0	0	0
Chiradzulu	0	1	0	0	0	0
Phalombe	0	0	1	1	1	1
Sex	1	1	1	1	0	0
Age	86	22	18	24	18	29
WealthMatt	0	0	0	1	0	0
WealthRadio	1	1	0	1	1	0
L8kDaysWorked	1	1	0	7	0	60
L7PubWork	1	1	1	1	1	1
LEVERAGE (h)	0.26	0.07	0.10	.42	0.08	.30

d)

The index of Trust in traditional authorities *has 39 missing observations. Outline briefly the general problem of biased samples. Discuss in more detail the possibilities for having a biased sample in this particular study of* Trust.

Biased samples can be seen as the result of missing observations. If missing observations on the dependent variable y is not random, then the sample for this particular variable will be biased. Truncated, selected or censored data on x-variables will not cause problems for inferences in regression studies. But if it occurs for a y-variable estimated coefficients will be biased. The general advice is to construct a model that may predict who will have data missing and who will not have data missing on the particular dependent variable. This can be used to correct the biased estimates.

To investigate if we have a biased sample here we need to see if we can find any indications that those who have the value missing on the trust index are different from the total sample. In the tables of variable definitions for question 1 we see that there are 39 (or 13.8%) missing observations on the trust index. This is a bit high to be overlooked. In the variable definitions we find the distribution of those missing for the explanatory variables. In the distributions of the 39 missing cases we note that 4 cases is about 10%. Before starting to worry we should probably look for a 10% or larger difference between the distribution of the missing and the distribution on the variable in the sample. We note the deviations for the following variables

- District: Mzimba and Dowa are overrepresented and Phalombe is underrepresented among those missing. This will strengthen the rural and matrilineal presence in the sample analysed.
- Owning radio: those who do not own radios are more often missing increasing the presence of radio owners in the sample analyzed
- Participation in unpaid public work during the last 12 months: those who participated are overrepresented among the missing and hence underrepresented among those studied in the regression.
- Participated in graveyard clearing: here those who participated are overrepresented among the missing and hence underrepresented in the sample studied.

In this we may see a pattern: “modern” people from urban districts are less likely to participate in public work, graveyard clearing, but more likely to own a radio. The sample studied is most likely biased towards modern urban attitudes. How much this bias affects the estimated coefficients is difficult to judge without constructing a selection model to correct the estimated coefficients.

QUESTION 2 (Logistic regression, weight 0.4)
The people of Malawi practice several different forms of lineage systems. The basic distinction starts with determining if children belong to the husband’s or the wife’s lineage (patrilineal or matrilineal). This distinction is then qualified by the location of the married couple in the husband’s or the wife’s village of origin or elsewhere (patrilocal, matrilocal, or other locations). In the marriage customs of the various lineage systems payment for the bride is practiced to varying degrees. One particular type of payment is called lobola. The dependent variable in this study, *Bridepayment*, records if lobola has been paid by the household.
In the attachment for question 2 two models of payment for the bride has been estimated. Model 1 is a regression of *Bridepayment* on family type, sex, and age. Model 2 is a regression of *Bridepayment* on regional location of the households.

a) Determine, in the regression of *Bridepayment* on family type, sex, and age, if family type is related to the probability of paying for the bride.. Discuss the differences between the various family types in the propensity to use bridal payment. Use the model without control for sex and age.
b) Determine, in the regression of *Bridepayment* on family type, sex, and age, the effect of sex controlling for the effect of age and family type. Discuss the impact of age and how it affects the estimate of the effect of sex controlling for the effect of family type.
c) Determine, in the regression of *Bridepayment* on family type, sex, and age, by means of the likelihood ratio test if the interaction between sex and age contributes significantly to the model at a level of significance of 0.10.
d) Determine, in the regression of *Bridepayment* on regional location of the household, the degree to which the assumptions of a logistic regression have been fulfilled.

In commenting on the impact of various variables on Bridepayment it might be useful to remember that the answer to this question of Bridepayment is a statement of fact, not a question of like or dislike.

a)
Determine, in the regression of Bridepayment on family type, sex, and age, if family type is related to the probability of paying for the bride.. Discuss the differences between the various family types in the propensity to use bridal payment. Use the model without control for sex and age.

In model 1 of the regression of Bridepayment on family type, sex and age the family type is the first variable to enter as an explanatory variable. Blocks 2 to 5 add sex, age, age as curvilinear element, and finally interaction of sex and age.

Family type is a dummy coded variable with Chikamwini (matrilineal and matrilocal) as reference category and 3 included categories (Chitengwa (matrilineal and patrilocal), Patripatri (patrilineal and patrilocal), and OtherMarriageS).

One may answer the first question by noting that the omnibus test of the model coefficients provides a chi-square value of 142.215 with 3 degrees of freedom. With a 5% test level we find that all chi-square values above 7.815 will lead to rejection of the null hypothesis of zero explanatory force of the variable Family type.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	142.215	3	.000
	Block	142.215	3	.000
	Model	142.215	3	.000

A more standard way of determining if Family type contributes to the explanation of the variation in Bridepayment is to perform a Likelihood ratio test. In this test we compare two nested models where the big model has the 3 categories of the family variable and the small model do not. Otherwise the two models are identical. Since Family type is the first variable to enter in Block 1 the only common element in this case is the constant first entered in Block 0.

If we assume that the H extra variables in the big model do not contribute to the explanation of the dependent variable and compare the big model with K parameters to the small model without the H extra variables the test statistic

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

will follow a ChiSquare distribution with H degrees of freedom. 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 summary	Block 1 Big model	Block 0 Small model
-2 Log likelihood	189.024	331.239
	K=4	H=3

Hence we find
χ^2_3 = 331.239 - 189.024 = 142.215 (exactly as in the omnibus test, of course).

Choosing a test level of 5% we see that in table A4.3 in Hamilton (1992) a chisquare value larger than 7.815 has a probability of less than 0.05 if the null hypothesis is true. We have to conclude that the hypothesis of no contribution of the 3 family type variables probably is false.

In block 1 we find the following estimates of the impact of family type:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Chitengwa	3.308	.524	39.888	1	.000	27.337
	PatriPatri	4.301	.485	78.513	1	.000	73.800
	OtherMarriageS	3.036	.557	29.667	1	.000	20.829
	Constant	-2.092	.335	38.950	1	.000	.123

a. Variable(s) entered on step 1: Chitengwa, PatriPatri, OtherMarriageS.

Interpreting the results of a dummy coded variable depends on the excluded category. The excluded category in the table defining family type is called Chikamwini. We see that the odds for having paid for the bride is 73.8 times larger in the patrilineal/ patrilocal households than in the chikamwini households, and even in the Chitengwa households, and other marriage systems the odds are 27 times and 20 times higher.

One may also study the propensity to pay lobola in conditional effect plots of estimated probabilities. To construct such plots we use Pr(Y_i = 1) = 1/(1 + exp[-L_i]). In this formula
L_i = -2.092 + 3.036* OtherMarriageS + 4.301* PatriPatri + 3.308*Chitengwa

We see for example that L_i = -2.092 for the Chikamwini households. This means that Pr(Y_i = yes) = 1/(1 + exp[2.092]) = 0.110 for Chilkamwini households, while Pr(Y_i = yes) = 1/(1 + exp[-2.209]) = 0.901 for patrilineal and patrilocal households. Chitengwa and OtherMarrigeS are in between.

b)
Determine, in the regression of Bridepayment on family type, sex, and age, the effect of sex controlling for the effect of age and family type. Discuss the impact of age and how it affects the estimate of the effect of sex controlling for the effect of family type.

In Block 3 we find an estimate of the impact of sex when family type and age are controlled for. The coefficient of 0.845 tells us that men have a logit that is 0.845 higher than women after control for family type and age. The Wald test tells that this difference is significantly different from 0 with a test level of 0.05. The probability of finding a Wald statistic of 4.414 or larger given that there is no impact of Sex is 0.036. However, according to the Wald statistic age does not contribute significantly to explaining Bridepayment in this version of the model. This is confirmed by the LogLikelihood Ratio test of age based on the difference between Block 3 and Block 2. The test statistic is from a chisquare distribution with 1 degree of freedom. With a test level of 5% the test statistic has to be larger than 3.841 before we discard the null hypothesis of no contribution from age to the model of Bridepayment. We find χ^2_1 = 184.596 – 182.955 = 1.641,

Block 3: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Chitengwa	2.824	.554	26.009	1	.000	16.851
	PatriPatri	4.037	.494	66.708	1	.000	56.629
	OtherMarriageS	2.905	.567	26.223	1	.000	18.272
	Sex	.845	.402	4.414	1	.036	2.327
	Age	.015	.012	1.625	1	.202	1.015
	Constant	-2.968	.611	23.602	1	.000	.051

a. Variable(s) entered on step 1: Sex, Age.

We find Sex, Family type, and Age also in blocks 4 and 5 where age as curvilinear element and interactions with sex are explored.

In model 4 age squared (age2) is introduced. The p-value (Sig) of the Wald statistic is high for both age and age2, though it is slightly lower for age compared to block 3. Sex is still significant in this block. But age as a curvilinear variable clearly is not contributing to the explanation of Bridepayment. This is confirmed by the LogLikelihood ratio test of block 4 against block 2. We find χ^2_2 = 184.596 – 181.699 = 2.897, clearly below the critical value of 5.991 at a 5% test level.

In block 5, when we add the interaction terms for age and sex, the p-values increase for both age terms, and the interaction terms also have too high p-values. The conclusion is clear: age does not contribute to the explanation of Bridepayment neither as a linear nor as a curvilinear element. Neither does it interact with sex. And introducing interaction terms make also sex insignificant.

Block 5: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Chitengwa	2.817	.557	25.527	1	.000	16.720
	PatriPatri	4.067	.504	65.106	1	.000	58.392
	OtherMarriageS	2.786	.576	23.402	1	.000	16.223
	Sex	3.517	2.903	1.468	1	.226	33.689
	Age	-.069	.202	.117	1	.733	.933
	Age2	.000	.002	.055	1	.815	1.000
	SexAge	.107	.135	.620	1	.431	1.112
	SexAge2	-.001	.001	.372	1	.542	.999
	Constant	-5.827	2.193	7.057	1	.008	.003

a. Variable(s) entered on step 1: SexAge, SexAge2.

It was noted that the logit for males are 0.845 higher than for females. From this we see that males (sex=1) have 2.3 times higher odds than women of having said that there was paid lobola.

Again, based on Pr(Y = 1) = 1/(1 + exp[-L]), and with the block 3 estimate of the logit we find L = -2.968+2.824Chitengwa +4.037PatriPatri +2.905OtherMarriageS +0.845Sex +0.015Age.

With this we can produce conditional effect plots to explore the relationship of sex to the probability of saying yes on the variable Bridepayment.

c)

Determine, in the regression of Bridepayment on family type, sex, and age, by means of the likelihood ratio test if the interaction between sex and age contributes significantly to the model at a level of significance of 0.10.

Due to multicollinearity a likelihood ratio test of the contribution of the interaction terms of sex and age needs to involve either the age variable or the sex variable in addition to the interaction terms. The variables in Model 1 are entered in the different blocks like this:

Variable	**Enters first in**
Chitengwa	Block 1
PatriPatri	Block 1
OtherMarriageS	Block 1
Sex	Block 2
Age	Block 3
Age2	Block 4
SexAge	Block 5
SexAge2	Block 5

From inspecting model 3 we know that sex contributes significantly to the model while age does not. It is a general experience that curvilinear age and sex and age interactions may change this.

A likelihood ration test of age and the interaction of age and sex will in this case have to be based on a comparison of block 5 and block 2. We see that H=4, K=9 and χ^2_3 = 184.596 - 180.118 = 4.478. With the test level of 10% the chi-square is much less than the critical value of 7.779. We conclude that age as a curvilinear element and its interaction with sex do not contribute to the explanation of Bridepayment.

It is known that multicollinearity is a problem in logistic regression. It is also known that technical variables such as curvilinear and interaction terms will correlate with their defining variables. This means in practice that the causal impact of a variable and its correlated terms will be distributed more or less randomly among the various parts of the group of correlated terms. To test only the interaction terms of sex and curvilinear age will not be a fair test of their combined impact. At least one of the two variables sex and age needs to be included in the test. Given that age is curvilinear there will be high correlation between the two age terms, age and age^2. Multicollinearity will be further exacerbated by the interaction terms. The most reasonable test of age and interaction of age and sex is to compare block 5 to block 2.

As long as this test rejects the null hypothesis of no impact from the group of variables we can safely drop both age and interaction of age and sex from our model. If the test had come out confirming a positive contribution to explaining the variation in *Bridepayment* one would need further investigations to determine if both curvilinear age and interaction between age and sex would be needed in the model. Comparing block 4 to block 2 will determine if curvilinear age is a correct specification of the model. If block 5 is a significant contribution to the model on its own there is no problem. But as noted this is not in general a fair test of the contribution from the interaction terms. To see if there may be an interaction between age and sex we will have to perform the same test as above comparing block 5 to block 2. Assuming this to be positive we need to track the changes in p-values for the age and sex variables to determine if the interaction terms contribute. If any of the p-values of sex, age and age2 in general improve we should keep the interaction terms even if they alone do not show up as significant.

d)

Determine, in the regression of Bridepayment on regional location of the household, the degree to which the assumptions of a logistic regression have been fulfilled.

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

- 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

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

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_{bk} 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.

The estimate of model 2 has only regional location as explanatory variable for Bridepayment with the region of Phalombe as reference category. This means that it contains only 5 dummy coded regional location variables. Hence the logit is linear in the parameters. Also the model summary with -2LogLikelihood of 140.807 assures us that the variable District is relevant for the model. But due to the study of Model 1 of *Bridepayment* we should also suspect that relevant variables are excluded. A relevant variable in Model 1 is relevant for this model if it correlates with the region variable. We do not have any measure of such correlations.

Since we have an estimate of the model we cannot have perfect multicollinearity or perfect discrimination. So except for excluded relevant variables the assumptions of logistic regression are met. But inspecting the estimate of the impact of regional location on Bridepayment it becomes clear that something must be very wrong in this model.

SPSS notes in the model summary table that "Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found." The estimated model comes out like this

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Rumphi	22.908	6129.380	.000	1	.997	8.885E9
	Mzimba	23.231	6129.380	.000	1	.997	1.228E10
	Kasungu	22.812	6129.380	.000	1	.997	8.077E9
	Dowa	22.861	6129.380	.000	1	.997	8.481E9
	Chiradzulu	.000	8829.279	.000	1	1.000	1.000
	Constant	-21.203	6129.380	.000	1	.997	.000

a. Variable(s) entered on step 1: Rumphi, Mzimba, Kasungu, Dowa, Chiradzulu.

Something is wrong! Problems of this kind may be caused by a high degree of multicollinearity or discrimination. But since the model consists of only one explanatory variable included as 5 dummy coded regions we cannot have any degree of multicollinearity. The problem we see is called discrimination.

Going back to the variable definition tables we see a table of "**Bridal payment according to sex, district location and family type**". Here we see that in the districts Phalombe and Chiradzulu all households report that lobola has not been paid. No households in Chiradzulu and Phalombe said "yes, bridal payment has been made"! And for the other four districts we see that the number of households saying "no, bridal payment has not been made" is very small. While we do not have perfect discrimination we do have a high degree of discrimination, causing very high values on the estimates of the standard error of the estimates leading to insurmountable problems for inference in the model. The fact that Phalombe was chosen as reference category exacerbates the problems.

Bridal payment according to location of household		Bridepayment has been paid		Total
		no	yes	
District of household location is Rumphi	0 no	102	110	212
	1 yes	6	33	39
District of household location is Mzimba	0 no	103	105	208
	1 yes	5	38	43
District of household location is Kasungu	0 no	102	113	215
	1 yes	6	30	36
District of household location is Dowa	0 no	100	101	201
	1 yes	8	42	50
District of household location is Chiradzulu	0 no	68	143	211
	1 yes	40	0	40
District of household location is Phalombe	0 no	65	143	208
	1 yes	43	0	43

Since Phalombe is the reference category all computation of marginal effects breaks down. If one of the other districts had been chosen as reference category, for example Rumphi, we would have gotten reasonable estimates for the 3 districts Mzimba, Kasungu, and Dowa. But Chiradzulu and Phalombe would still be impossible to estimate, and we will be warned that "Estimation terminated at iteration number 20 because maximum iterations have been reached. Final solution cannot be found."

Using Rumphi as reference category produces the following estimates. Look at the size of the estimated standard errors for Chiradzulu and Phalombe.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Mzimba	.323	.651	.247	1	.619	1.382
	Kasungu	-.095	.630	.023	1	.880	.909
	Dowa	-.047	.588	.006	1	.937	.955
	Chiradzulu	-22.908	6355.067	.000	1	.997	.000
	Phalombe	-22.908	6129.370	.000	1	.997	.000
	Constant	1.705	.444	14.754	1	.000	5.500

a. Variable(s) entered on step 1: Mzimba, Kasungu, Dowa, Chiradzulu, Phalombe.

This kind of problem may in some cases be related to small sample size. The model is estimated on 251 cases with a split of 143 yes and 108 no on the question of payment of lobola. Thus the sample is above the minimum threshold recommended. But usually increasing the sample might alleviate a problem of discrimination as found here. But it is unlikely to do so in this case since *not paying lobola* is a defining characteristic of matrilineal cultures. And since there is a very clear regional separation of matrilineal cultures, the regional variable will be a proxy for family system. Thus rural village households from the matrilineal regions will never say yes to having paid lobola. If urban areas were included we might find migrant households with other family systems where lobola may be paid (but not necessarily).

QUESTION 3 (Structural equations, weight 0,1)
In a study of attitudes towards rural society people in a random sample of the Norwegian population were asked to express how strongly they agreed or disagreed to the proposition: “Life in the countryside is more satisfying than life in towns.” This is the dependent variable, called *Livet på landet best*, in a structural model with 2 intermediate variables, income (*E.inntekt*) and education (*E.utdanning*); and 2 independent variables, age (*Alder*) and a dummy indicating if the respondent is a woman (*Kvinne*). Several variants of the structural equations have been estimated and are presented in the attachment for question 3.

a) Draw a path diagram of the relations of the structural model. Find the best estimates of the path coefficients to indicate the strength of the relations in the diagram and write the coefficients into the diagram.
b) Determine the size of the direct impact of *Kvinne* on *Livet på landet best*. Determine the size of the indirect effect of *Kvinne* on *Livet på landet best*.

a)
Draw a path diagram of the relations of the structural model. Find the best estimates of the path coefficients to indicate the strength of the relations in the diagram and write the coefficients into the diagram.

The question may be solved either based on the notation used by Hamilton (2008) or by the notation used in lecture 11/2010.

First using the notation from lecture 11/2010:

Defining
Y_3=Livet på landet best,
Y_2=E.inntekt,
Y_1=E.utd,
X_2=Kvinne,
X_1=Alder,
and assuming that the variables are standardized z-score variables and that the three regression equations satisfies the requirements for OLS regression, the fully specified structural model can be written:

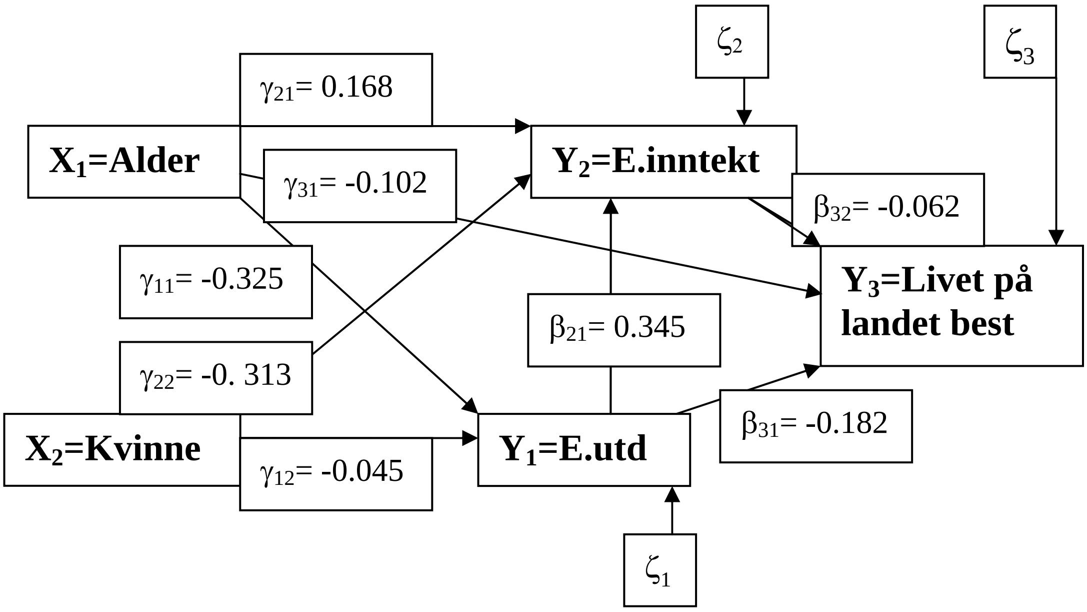
$$Y_1 = \gamma_{12}X_2 + \gamma_{11}X_1 + \zeta_1$$
$$Y_2 = \beta_{21}Y_1 + \gamma_{22}X_2 + \gamma_{21}X_1 + \zeta_2$$
$$Y_3 = \beta_{32}Y_2 + \beta_{31}Y_1 + \gamma_{32}X_2 + \gamma_{31}X_1 + \zeta_3$$

These equations are estimated in models 7, 4, and 1. Inspecting the estimates of β and γ in model 1 we see that γ_{32} is not significantly different from 0. Hence, the real relations between Y_3 and the explanatory variables are estimated by

$Y_3= \quad \beta_{32}Y_2 + \beta_{31} Y_1 + \gamma_{31}X_1 + \zeta_3$

in model 3

The path diagram will look like the following:



The ζ-variables represent the residuals, the unexplained variation of the dependent variable.

Second using the notation from Hamilton (2008):

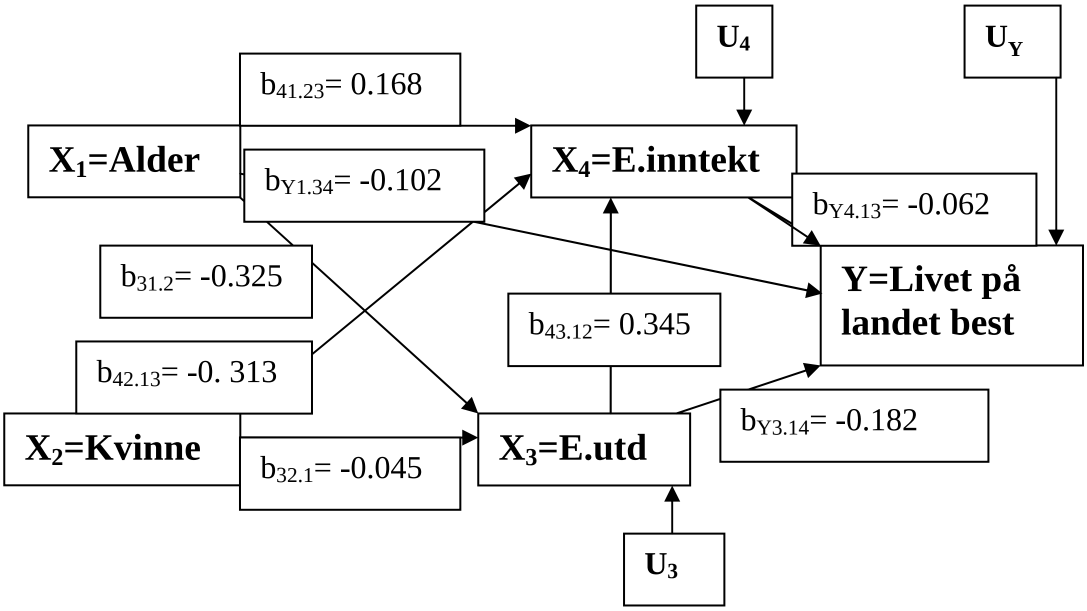
Defining
Y =Livet på landet best,
X_4=E.inntekt,
X_3=E.utd,
X_2=Kvinne,
X_1=Alder.
and assuming that the variables are standardized z-score variables and that the three regression equations satisfies the requirements for OLS regression, the fully specified structural model can be written:

$X_3= \qquad b_{32.1}X_2 + b_{31.2}X_1 + U_3$
$X_4= \qquad b_{43.12}X_3 + b_{42.13}X_2 + b_{41.23}X_1 + U_4$

$Y = b_{Y4.123}X_4 + b_{Y3.124}X_3 + b_{Y2.134}X_2 + b_{Y1.234}X_1 + U_Y$

These equations are estimated in models 7, 4, and 1. Inspecting the estimates of the b_{Y*} coefficients in model 1 we see that $b_{Y2.134}$ is not significantly different from 0. Hence, the real direct relations between Y and the explanatory variables are estimated by
$Y = b_{Y4.13}X_4 + b_{Y3.14}X_3 + b_{Y1.34}X_1 + U_Y$
in model 3

The path diagram will look like the following:



The U-variables represent the residuals, the unexplained variation of the dependent variable.

b)

Determine the size of the direct impact of Kvinne on Livet på landet best.
Determine the size of the indirect effect of Kvinne on Livet på landet best.

As seen above the direct impact of Kvinne on Y is not significantly different from zero. It is in the diagram above absent or it can be set to 0.

The indirect impact of Kvinne has three paths: one by way of E.utd directly to Y, one by way of E.inntekt directly to Y, and lastly one by way of E.utd and E.inntekt to Y.

The 3 indirect paths are

$b_{32.1}$ * $b_{Y3.14}$ =	-0.045 * -0.182	0.008190
$b_{42.13}$ * $b_{Y4.13}$ =	-0. 313 * -0.062	0.019406
$b_{32.1}$ * $b_{43.12}$ * $b_{Y4.13}$ =	-0.045 * 0.345 * -0.062	0.000963
Sum indirect effects of Kvinne		0.028559

Comment: with no direct effect of Kvinne and with an indirect effect of less than 0.03 one may conclude that Kvinne has no substantial impact on the strength of opinion on “livet på landet best”.