EKSAMENSOPPGÅVER SVSOS3003
Haust 2009
FRAMLEGG TIL LØYSING

Erling Berge
Institutt for sosiologi og statsvitenskap
Norges Teknisk Naturvitskapelege Universitet

«Bruksanvisning»

Når ein går igang 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 lesning 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>)

ENGLISH

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

In this question we explore the propensity to be generous to people within your own community when you do not know the identity of the person you show generosity. It was determined that trust (as measured by answering "yes, most people can be trusted" to the question "Generally speaking, do you think most people can be trusted or that they cannot be trusted?") did not show any realionship with level of generosity. Instead three other types of factors were considered: general personal characteristics, indicators of wealth, and indicators of culture.

a) Describe the impact of "Mattress owned" on Generosity as it is estimated by model 4. Find a 95% confidence interval for the impact.

b) Determine if the interaction between "Sex of respondent" and the variables "Mattress owned" and "Radio owned" contribute significantly to the explanation of variance in the dependent variable. Use a 0.05 level of significance for the test and state explicitly the hypothesis that is being tested.

c) Present the assumptions that need to be satisfied if the estimates and tests of the 9 models are to be trustworthy. Determine if the tables presented give any reason to doubt that the model assumptions are satisfied

d) Based on the tables presented what can you say about the factors affecting level of generosity? Discuss in particular the impact of sex and age.

a) Describe the impact of "Mattress owned" on Generosity as it is estimated by model 4. Find a 95% confidence interval for the impact.

The dependent variable for models 1-9 is Generosity, the amount of Kwacha returned over or below what is defined as a fair share of the profit from the initial investment in a trust game. The variable "Mattress owned" is an indicator of relative wealth in communities where many sleep directly on the floor. It takes the value of 1 if the respondent owns a mattress, zero otherwise.

Model 4 tells us that a person owning a mattress returns 19.2 Kwacha more than a fair division of the profit controlling for differences between sexes and age groups, as well as ownership of radio. It is a bit puzzling that ownership of radio has a large negative impact, meaning that if you own both mattress and radio you are not nearly as generous as if you own only mattress.
In OLS regressions estimates of the model parameters, b_k , are known to follow a t-distribution if the estimates come from a simple random samples, and the null hypothesis of zero value of the population parameter is true.

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

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

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 find in model 4 that $b_{\text{Mattress owned}} = 19.268$, $SE_{b(\text{Mattress owned})} = 8.398$, n = 116, and K = 8. Hence n-K = 108, and since the table of the t-distribution in Hamilton (page 350) for $\alpha = 0.05$ gives us critical values for 60 (t=2) and 120 degrees of freedom (t=1.98), we see that for df=108 $1.98 < t_\alpha < 2.00$. Since df=108 is closer to 120 than to 60, one may here interpolate using the conservative value of 1.99. Normally one will choose to use the value of 2, but also 1.98 will be acceptable.

This means that the 0.95 confidence interval will be
$19.268 - 8.398*1.99 < \beta_k < 19.268 + 8.398*1.99$
$19.268 - 16{,}71202 < \beta_k < 19.268 + 16{,}71202$

$2{,}55598 < \beta_k < 35{,}98002$
$2{,}55 < \beta_k < 35{,}98$

b) Determine if the interaction between "Sex of respondent" and the variables "Mattress owned" and "Radio owned" contribute significantly to the explanation of variance in the dependent variable. Use a 0.05 level of significance for the test and state explicitly the hypothesis that is being tested.

Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
4	(Constant)	32.819	29.539		1.111	.269		
	Sex of respondent	-119.871	46.696	-1.749	-2.567	.012	.017	59.219
	Age of respondent	-2.065	1.483	-1.048	-1.393	.167	.014	72.309
	Age squared	.020	.016	.976	1.277	.204	.013	74.545
	Sex * Age	5.089	2.213	3.691	2.299	.023	.003	328.752
	Sex * Age squared	-.050	.023	-2.322	-2.136	.035	.007	150.811
	Mattress owned	19.268	8.398	.220	2.294	.024	.849	1.178
	Radio owned	-14.420	7.194	-.201	-2.004	.048	.778	1.285
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
5	(Constant)	31.475	30.019		1.049	.297		
	Sex of respondent	-117.719	47.343	-1.717	-2.486	.014	.017	59.854
	Age of respondent	-1.994	1.533	-1.012	-1.300	.196	.013	75.964
	Age squared	.020	.016	.946	1.197	.234	.013	78.327
	Sex * Age	4.879	2.333	3.539	2.092	.039	.003	359.166
	Sex * Age squared	-.047	.025	-2.206	-1.907	.059	.006	167.850
	Mattress owned	22.310	11.337	.255	1.968	.052	.474	2.111
	Radio owned	-16.126	9.540	-.225	-1.690	.094	.450	2.222
	Sex * Own mattress	-6.975	17.062	-.057	-.409	.684	.408	2.450
	Sex * Own radio	4.130	14.694	.058	.281	.779	.188	5.328

We want to determine if the two variables (interaction terms) "Sex*Own mattress" and "Sex*Own radio" contribute significantly to the model of Generosity. We want to test
H_0: $\beta_{\text{Sex* Own mattress}} = 0$ and $\beta_{\text{Sex* Own radio}} = 0$ against the alternative
H_A: $\beta_{\text{Sex* Own mattress}} \neq 0$ and $\beta_{\text{Sex* Own radio}} \neq 0$

In testing if interactions between sex and indicators of wealth contribute to the model, we inspect models 4 and 5. This is where they appear for the first time. Due to multicollinearity tests of single coefficients cannot be trusted.

Sex is also involved in all models 1-3 including interactions with age. This results in a very high degree of multicollinearity in the models from 3 on. But the variance inflation factor (VIF=1/tolerance) of Sex does not increase very much from model 4 to 5. Hence the test of the contribution of the interaction terms with the wealth indicators can leave out sex alone.

In the change statistics of the Model summary table, the value of the F-statistic for the contributions of Sex*Own mattress and Sex*Own radio to the model is 0.098 with 2 and 106 degrees of freedom. The probability of finding this low or lower values of the

F-statistic, given that the population values of the model parameters for these two variables are zero, is 0.907 ("Sig.F-change" column of the table). But the tolerances, particularly for Sex*Own radio, are low.
Inspecting the coefficients of Own Mattress and Own radio in the two models we see that the p-values of the two increases significantly from model 4 to 5.
Since both interaction terms do not contribute significantly to the model, and since the p-values in the tests of the Own mattress and Own radio variables increase, the F-test of the two interaction terms should be considered to be valid.

In general a test statistic (in this case F) is constructed assuming the null hypothesis of no impact of the tested variables is true. The hypothesis we want to test here is then:
H_0: In model 5 $\beta_{\text{Sex*Own mattress}}$ and $\beta_{\text{Sex*Own radio}}$ are both equal to 0

Then we have to compare model 4 and 5. The F-statistic:

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

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.

In this case we have that H=2, n=116, K=10, and from table A4.2 in Hamilton we see that the critical value for F^{2}_{106} with level of significance 0.05 is approximately 3.07. We conclude that the two interaction terms do not contribute to the model specification if we find that the computed value F^{2}_{106} is less than the critical value 3.07 (table value of F^{2}_{120} for α=0.05) of assuring a test level of 0.05. This F-value has already been computed in the Model summary table and is there given as 0.098, far below the critical value.

Alternatively:
An alternative avenue for finding the F-value, is to look up the residual sums of squares in the ANOVA table and compute the value according to the formula given above. To compute the F-value we need to find RSS(K) = 115080.630, RSS(K-H) = 115292.517, H=2, n=116, K=10, and RSS(K) / (n-K) = 1085,666. Then it follows that (RSS(K-H) - RSS(K))/H = (115292.517 - 115080.630)/2 = 211,887/2 = 105,9435 and {(RSS(K-H) - RSS(K))/H} / {RSS(K) / (n-K)}= 105,9435/ 1085,666 = 0,09758. The F-value we compute this way is of course exactly the same as the one reported by SPSS in the Model summary table for the test of changes in model. The conclusion is that the two interaction terms do not contribute to the model specification. They are irrelevant variables and should be removed.

c) Present the assumptions that need to be satisfied if the estimates and tests of the 9 models are to be trustworthy. Determine if the tables presented give any reason to doubt that the model assumptions are satisfied

All models in question 1 are regression models of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_{19} X_{19i} + \varepsilon_i .$$

where ”i” runs over the household population of 18 Malawian willages. If we let k=0, 1, 2, 3, … ,19, 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 19 x-variables are defined in the model 9 table and the section of variable definitions

An OLS (ordinary least squares) estimate of the model parameters defined above can be found as the b-values of $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + ... + b_{13} x_{13i}$ that minimizes the sum of squared residuals,

$$RSS = \Sigma_i(y_i - \hat{y}_i)^2 = \Sigma_i e_i^2$$

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

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

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"
  - $E(\varepsilon_i) = 0$ for all "i"
- The error terms have constant variance for all cases "i" (homoscedasticity) for all "i"
  - $var(\varepsilon_i) = \sigma^2$ for all "i"
- The error terms do not correlate with each other across cases (no autocorrelation) for all "i" ≠ "j"
  - $cov(\varepsilon_i, \varepsilon_j) = 0$ for all "i" ≠ "j"

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.

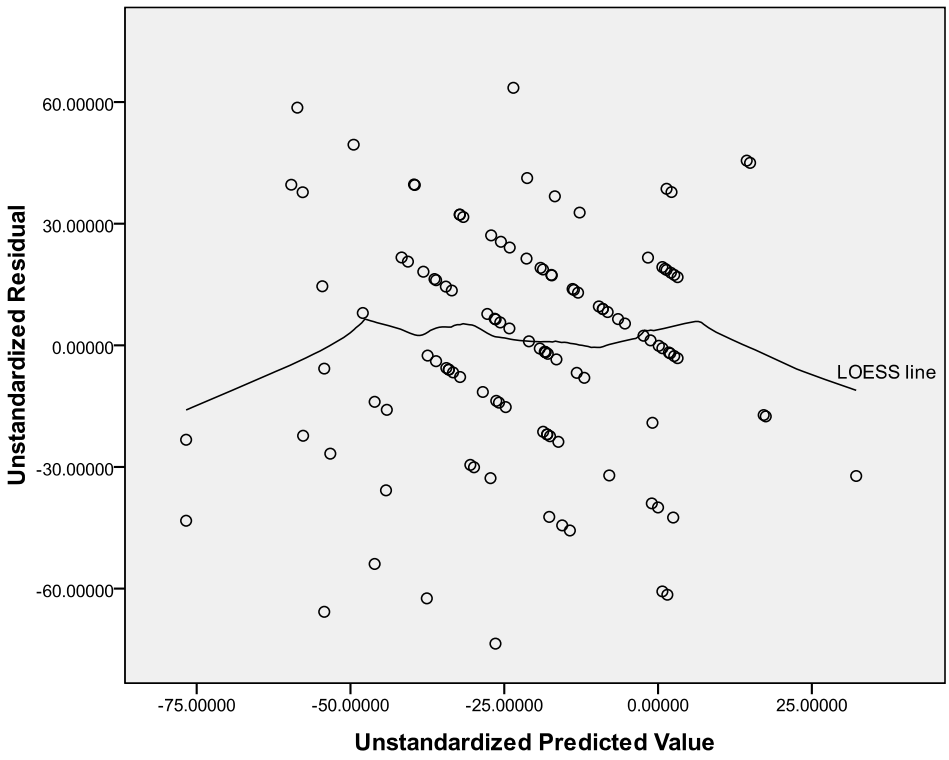
Some of the stated assumptions cannot be tested. In particular we cannot test if

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

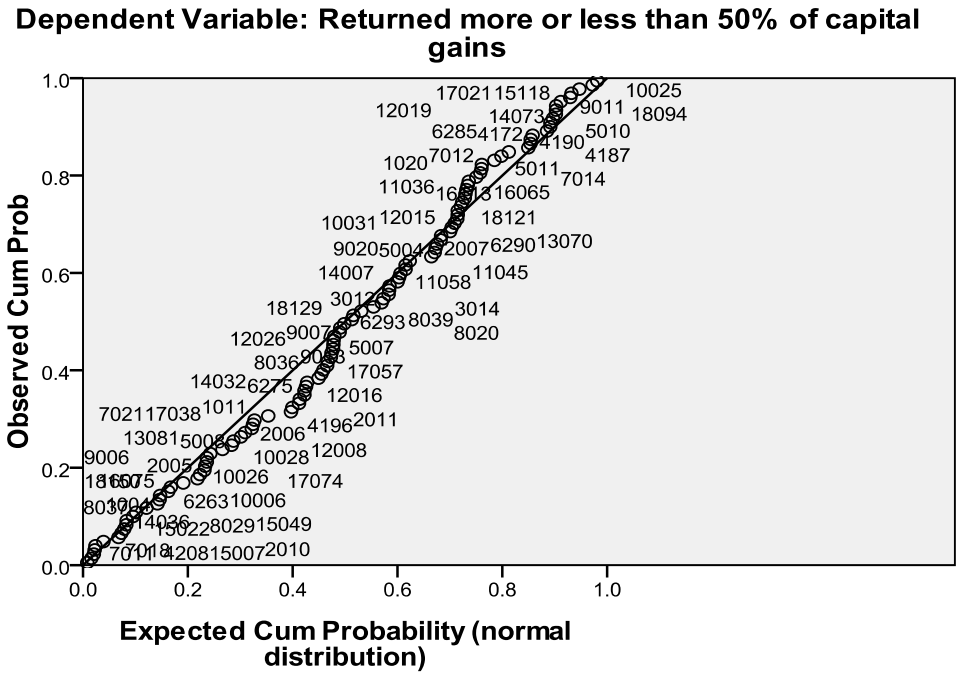
We can test if

1. irrelevant variables have been included in the model
2. the model is curvilinear in the included variables
3. there is heteroscedasticity and/ or autocorrelation
4. the error term is normally distributed

Discussion of the assumptions in relation to model 9.
As concluded in point b) above there are irrelevant variables in the model. The consequence of irrelevant variables is that variances are larger than they otherwise would be, making confidence intervals wider and estimates less precise.

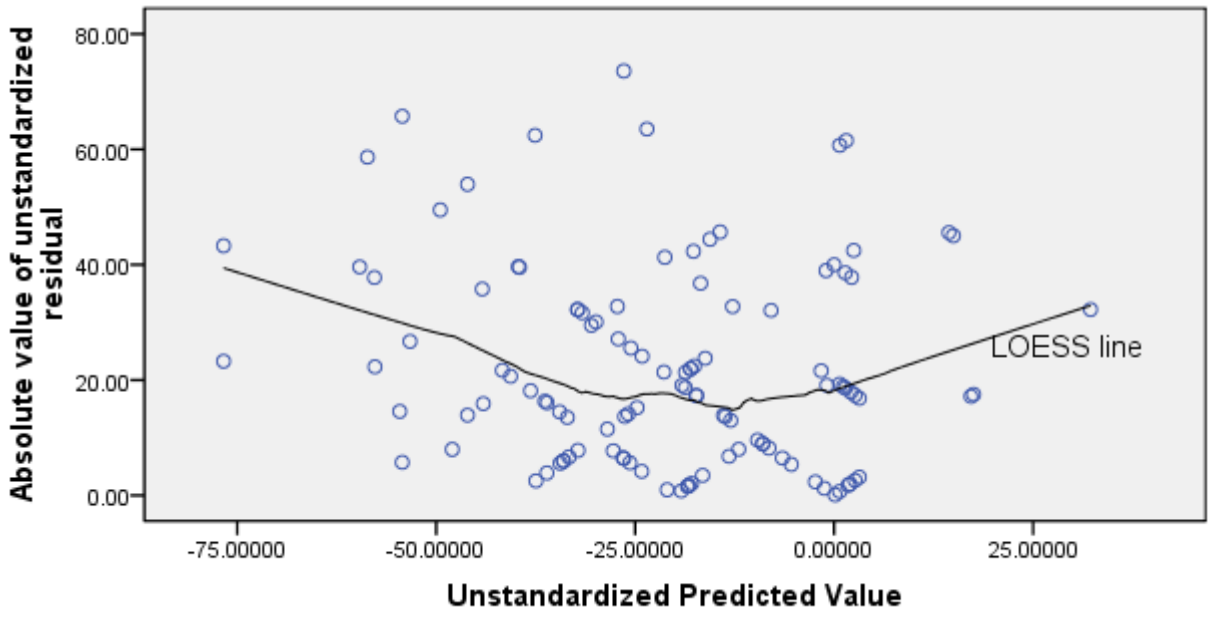


Normal P-P Plot of Regression Standardized Residual



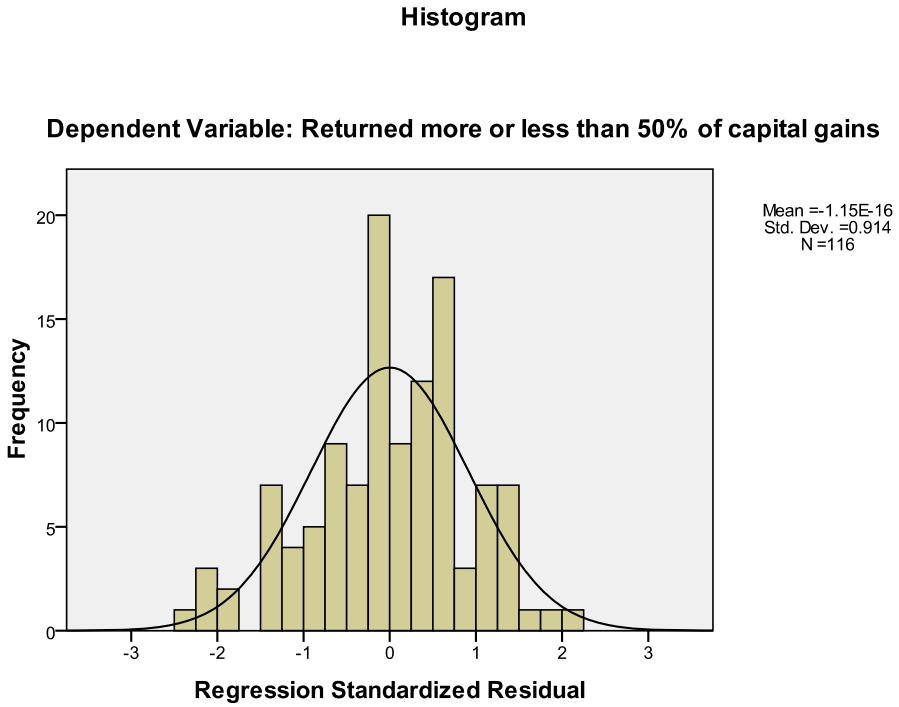
All variables except Age are binary coded. Hence only age can be curvilinearly related to the dependent variable Generosity. And already in model 3 it is established that age is curvilinearly related to Generosity in interaction with sex. In the models 6-9 this relationship seems to be disappearing. But fluctuations in the p-values for sex and age (Sig.) may be assumed to be related to the introduction of several interaction terms with sex that perhaps might be irrelevant. A model without the interaction terms between sex and Own mattress, Own radio and Region, possibly also marriage system, should be estimated before judgment is passed on sex and age in a larger model than model 3.

In judging the degree of heteroscedasticity we look at the plot of predicted values against the residuals. The LOESS line is for central parts of the scatter plot fairly level. This suggests that the degree of heteroscedasticity is low and probably introduced through limited variation both on the dependent and the independent variables. The fluctuations of the LOESS line are mirrored in the deviations from the diagonal of the Normal Probability plot. Another possible reason for the curved LOESS line may be influential cases at the extremes of the predicted values scale. Taking a closer look at the LOESS line in the scatter plot of the absolute value of the unstandardized residual we see 2 cases with ca -75 as predicted values and one with ca +35. In the box plot of the standardized predicted value 2 cases, 8037 and 10006, appear as extreme outliers. These are both women, 48 years old, and have the same values also on all other variables included in the model except the dependent. Together they have high influence. The conclusion h ere is that the sample probably is too small. One should probably also consider to report results both with these two and without them.



To judge autocorrelation we need to think about the possible causes of such correlations. Regions and districts are purposively selected. Sorting data according to geographical proximity might reveal any autocorrelation due to this. Assuming this has been done before the computation of the Durbin-Watson statistic of 1.92 (see the model summary table) one may conclude that there probably is not any autocorrelation in the data. Hamilton's table A4.4 gives for samples of 100 and models of 5 variables an upper limit of 1.78. Model 9 has 19 variables and is estimated on 116 cases. So at most the test would be inconclusive, but probably we could reject the hypothesis of autocorrelation.

To evaluate the requirement of normally distributed residuals we inspect the distribution of the residual in the histogram of the residual. There are some deviations but they do not seem to be systematic in relation to the distribution. However, in a sample of only 116 cases also this distribution ought to alert us to the possibility of influential observations. "No influential case" is not a requirement per se, but their presence may destroy the normal distribution of the error term.



There are several statistics we can inspect to evaluate the possible presence of influential cases. One basic statistic is the leverage, h. SPSS reports the centered leverage, that is the leverage minus the mean. The sample mean of the leverage is K/n, or in this case 20/116 = 0.172. The maximum of the centered leverage is 0.57. Thus the absolute value of the maximum is 0.742. This is above the 0.5 where Hamilton advices

us to avoid the case. Looking at the boxplot of the h statistic we find 4 cases outside the 1.5IQR distance from the median. They are 2006, 6285, 7014, and 17038.

In the boxplot of the standardized residual and the standardized predicted value we find large values for the cases 4208, 8037, and 10006.

Another general indicator is Cook's D statistic. Inspecting the box plot of Cook's D statistic we see that there are 8 cases with values more than 1.5*IQR from the median. Three of these are among those with large h. Looking similarly at the box plots of the standardized residual and standardized predicted value we find the cases 4208, 6285, 7014, 7018, 8029, 8037, 10025, and 17038.

Looking also at the box plots for the DFBETAS, we see numerous cases with values exceeding 1.5IQR from the median. Looking for gaps in the distribution we find the cases: 2006, 4208, 6285, 7014, 7021, 8037, 10025, 15118, 17038, and 18094. They are distributed across the variables as follows:

Variable group	Cases with highest values DFBETAS (single case or groups)				
Sex	6285	7014			
Age	-				
Wealth indicators	4208	15118	18094		
Region dummies	7014	7021	8037	10025	17038
Marriage system dummies	2006	7014	17038		

Five cases are found as potentially influential by only one statistic, the cases 7018, 8029, 10006, 15118, and 18094. Case 10006 has the largest predicted value we shall look at this together with the cases 2006, 4208, 6285, 7014, 8037, 10025, and 17038.

Variables	Case 2006	Case 4208	Case 6285	Case 7014	Case 8037	Case 10006	Case 10025	Case 17038
MatriMatri	0	0	0	0	0	0	0	0
MatriPatri	1	0	0	1	0	0	0	1
PatriPatri	0	1	0	0	0	0	0	0
OtherMarri	0	0	1	0	1	1	1	0
Generosity	-40	-100	0	0	-120	-100	0	0
OwnMattr	0	1	1	0	0	0	0	0
OwnRadio	1	1	0	1	1	1	0	0
Sex	0	1	1	0	0	0	0	0
Age	52	27	85	25	48	48	28	47
North	1	1	1	0	0	0	0	0
Central	0	0	0	1	1	1	1	0
South	0	0	0	0	0	0	0	1

It is not obvious why these are influential cases. We see that 3 of the 6 most non-generous players are included here, and none of the generous. Two of these are middle aged women living in the central districts.

The large DFBETAS for Sex on cases 6285 and 7014 must be due to the particular combination of values for these cases. 6285 is an old man living in the north and with 0 Generosity. 7014 is a young woman living in the Central region also with 0 Generosity. The influence we see is probably a consequence of few cases rather than any other kind of problems.

d) Based on the tables presented what can you say about the factors affecting level of generosity? Discuss in particular the impact of sex, age, and wealth in the models 1-4.

There are estimates of 9 models all nested hierarchically so that all previous models are contained in the last.The dependent variable is "Returned more or less than 50% of capital gains".

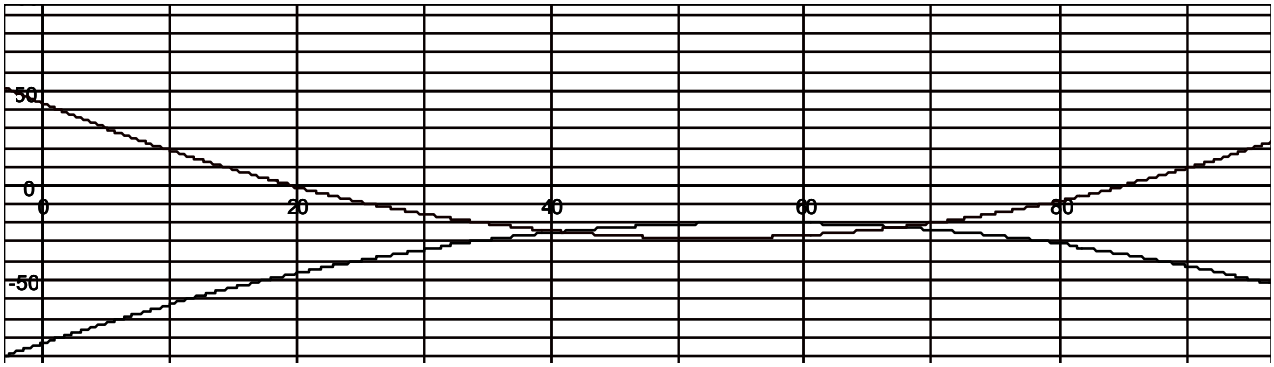
In the first 3 models the only variables involved were sex and age:

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-17.432	8.513		-2.048	.043		
	Sex of respondent	-12.328	6.384	-.180	-1.931	.056	.988	1.013
	Age of respondent	.055	.184	.028	.300	.765	.988	1.013
2	(Constant)	-.841	23.650		-.036	.972		
	Sex of respondent	-11.923	6.419	-.174	-1.857	.066	.981	1.020
	Age of respondent	-.788	1.136	-.400	-.694	.489	.026	38.636
	Age squared	.009	.012	.433	.752	.454	.026	38.515
3	(Constant)	43.248	29.905		1.446	.151		
	Sex of respondent	-126.694	47.557	-1.848	-2.664	.009	.017	58.641
	Age of respondent	-2.722	1.465	-1.382	-1.858	.066	.015	67.360
	Age squared	.026	.016	1.271	1.683	.095	.014	69.473
	Sex * Age	4.979	2.257	3.611	2.206	.029	.003	326.372
	Sex * Age squared	-.046	.024	-2.128	-1.925	.057	.007	148.803

It is remarkable that Sex alone is not quite significant at 5% level, and Age is far from being significant alone. Age as curvilinear variable seems to do better than as a linear variable, but is still far from being significant. But introducing the interaction between Sex and Age as a curvilinear variable makes the group clearly significant at 5% level and even more remarkable the least significant single element, Age squared has a p-value of 0.095 despite a very high degree of multicollinearity.

A conditional effect plot of age and sex might be interesting to inspect. Plotting the relationship as determined in model 3 we find that
Y=43.248 -126.694Sex -2.722Age +0.026Age*Age +4.979Sex*Age -0.046Sex*Age*Age
will provide 2 curves showing how generosity varies by age for men and women.

Such a curve is presented below:



$$y=43.248 -126.694\times 1 -2.722\times x +0.026\times x^2 +4.979\times x\times 1 -0.046\times x^2\times 1$$
$$y=43.248 -126.694\times 0 -2.722\times x +0.026\times x^2 +4.979\times x\times 0 -0.046\times x^2\times 0$$

Convex curve = men
Concave curve = women

From this we see that both young and old women are more generous than men, while men and women between about 30 and 80 years of age are about equal in generosity. However, we should also note that there minimum observed age is 15 and maximum is 85, and that there probably are very few cases below 20 and above 80. Hence a figure like this will exaggerate the differences between the sexes. Extrapolation from the observed range of a variable is not advisable.

In the models 4-9 the other explanatory factors are added:

Explanatory factors	**Variables**	**Results**
In models 4 and 5 the indicators of wealth and their interactions with Sex are introduced	Mattress owned Radio owned Sex * Own mattress Sex * Own radio	In model 4 the wealth indicators are significant. In model 6 they become marginally insignificant while the interaction terms clearly are irrelevant variables
In models 6 and 7 the indicators of regional cultures and their interaction with Sex are introduced. The reference category here is the Central region	North region South region Sex*North Sex*South	In model 6 we see that the region variable is significant while model 7 shows that the interaction terms are irrelevant variables
Then finally in models 8 and 9 the indicators of marriage system and their interactions with Sex are introduced	Matrilineal and matrilocal Patrilineal and paralegal Other marriage patterns Sex * Matrilineal and matrilocal Sex * Patrilineal and patrilocal Sex * Other marriage patterns	In model 8 we find that the marriage system variable do not contribute to the model by themselves, however, model 9 shows that their interaction with Sex contributes significantly

In model 4 the wealth indicators "Own mattress" and Own radio" are added.

The simple reasoning behind wealth as and explanation for generosity is that the relatively wealthier would be more generous towards their fellows. The reasoning may be too simple. The two wealth indicators work in different directions in relation to generosity, and they do so consistently across all models. Why this should be so is not obvious. One might want to rethink the reasoning behind their interpretation as wealth indicators.

The last model estimated is model 9. To facilitate the discussion we drop the irrelevant interaction terms without re-estimating the model.

Model		B	Std. Error	t	Sig.		VIF
9	(Constant)	-16.644	38.993	-.427	.670		
	Sex of respondent	-98.703	54.634	-1.807	.074		92.589
	Age of respondent	.235	1.552	.151	.880		90.380
	Age squared	-.002	.016	-.129	.897		90.816
	Sex * Age	3.228	2.323	1.389	.168		413.649
	Sex * Age squared	-.031	.024	-1.257	.212		189.957
	Mattress owned	32.334	11.899	2.717	.008		2.701
	Radio owned	-19.555	9.052	-2.160	.033		2.324
	North region	3.847	16.893	.228	.820		7.802
	South region	27.801	17.082	1.627	.107		8.276
	Matrilineal and matrilocal	-14.310	21.658	-.661	.510		13.707
	Patrilineal and patrilocal	-6.344	22.706	-.279	.781		14.095
	Other marriage patterns	-46.917	21.813	-2.151	.034		5.876
	Sex * Matrilineal and matrilocal	35.404	26.953	1.314	.192		6.453
	Sex * Patrilineal and patrilocal	29.770	26.353	1.130	.261		14.573
	Sex * Other marriage patterns	75.449	27.447	2.749	.007		3.856

The regional variables may indicate differences in culture as well as correlate with differences in the research teams collecting data in the 3 regions. The north and central region are not so different. But living in the southern region clearly increases generosity compared to living in the central region. Now, it is also the case that the southern region basically is matrilineal and matrilocal while the north basically is patrilineal and patrilocal and the central region mixed but perhaps leaning towards the patrilineal values. This means that there may be inter-correlations between region and marriage system further complicating the interpretation of the coefficients.
The fact that the interaction terms are significant where the marriage system alone is not, speaks to the reasonable suspicion that being man in a matrilineal culture is very different from being man in a patrilineal culture. But further interpretation depends on re-estimating the model with fewer variables, and more attention to the limited number of cases.

OPPGÅVE 2

In the same study as used above eight questions were asked about mistrust of people, and 14 about mistrust of institutions. It was assumed that there was at least one underlying trust-dimension responsible for the pattern of responses. To explore this question a principal component analysis was performed. In the analysis 16 of the 22 questions about mistrust were used.

a) Discuss the number of underlying dimensions and their meaning as far as attached tables allow.

The principal component analysis within the factor analysis framework allows the construction of indexes. It does not require many assumptions. It is sufficient that the variables can be used to compute Pearson correlations. It is fairly common to use as a device for detecting underlying attitude dimensions in a series of attitude questions. In this case we have a series of questions about trust and suspects that there are in reality a small number of more basic personality traits that shape the patern of responses from each individual.

The principal component analysis transforms K variables into K components in a way that maximizes the amount of explained total variance in the first component, then finds a component orthogonal to this explaining a maximum of the remaining variance. In this way K components are extracted. We want to determine how many to retain as indexes representing the original variables.

The usual approach is to look at the eigenvalues of the components. The sum of eigenvalues add up to the number of standardized variables where each variable has a variance of 1. Hence components with eigenvalues less than 1 explain less variance than one variable. It does not seem fruitful to keep components contributing that little to the explanation of the total variance.

In the table below we see that only the 4 first components have eigenvalues above 1.

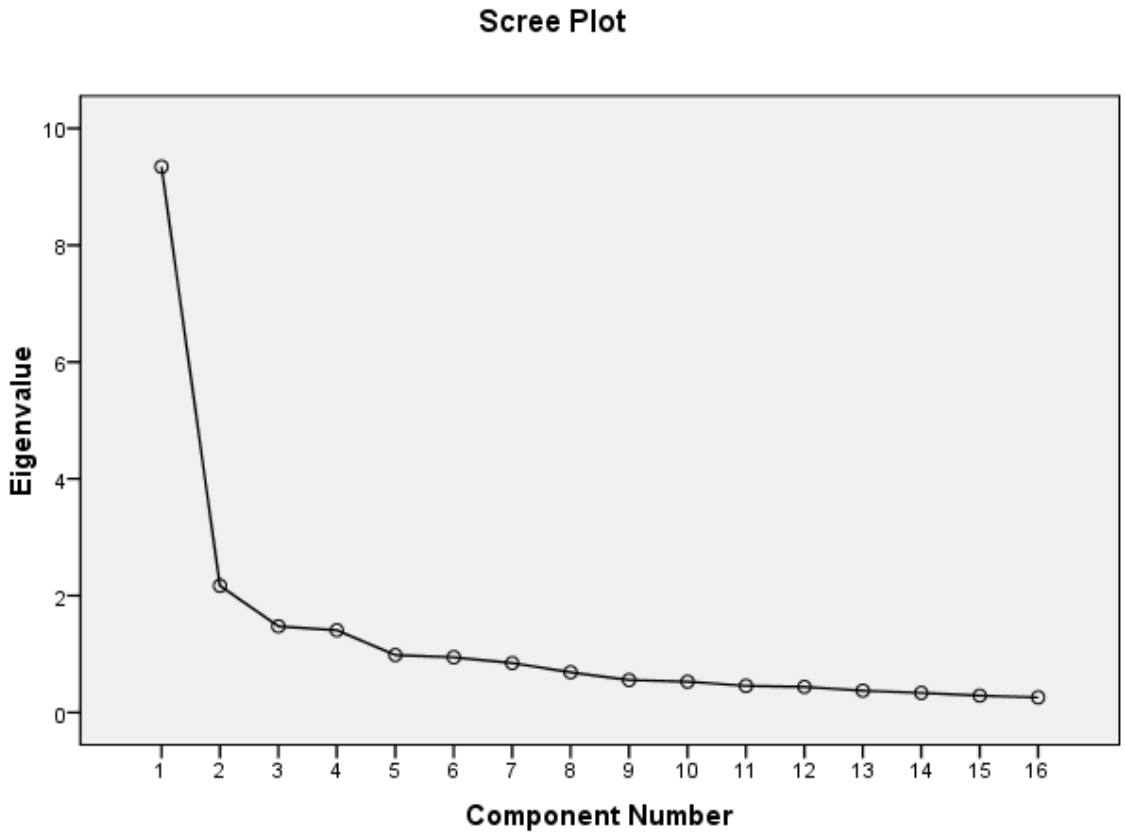
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.545	40.905	40.905	6.545	40.905	40.905	2.937	18.355	18.355
2	1.617	10.108	51.013	1.617	10.108	51.013	2.738	17.112	35.467
3	1.332	8.323	59.335	1.332	8.323	59.335	2.534	15.837	51.304
4	1.131	7.066	66.402	1.131	7.066	66.402	2.416	15.097	66.402

5	.859	5.369	71.771						
6	.743	4.642	76.413						
7	.709	4.430	80.843						
8	.613	3.832	84.675						
9	.502	3.140	87.815						
10	.473	2.957	90.772						
11	.375	2.345	93.117						
12	.273	1.708	94.825						
13	.256	1.601	96.425						
14	.236	1.473	97.898						
15	.206	1.289	99.188						
16	.130	.812	100.000						

Extraction Method: Principal Component Analysis.

In the scree plot this corresponds to a levelling off in the eigenvlues after component no 4. The conclusion is that one at a maximum may retain 4 components to explain 66.4% of the variance of the original 16 variables.



To justify 4 components they have to provide some substantial information. To find the meaning of the four components it is usually helpful to rotate them to simple

structure. This is done in the varimax procedure. The rotated component matrix provides the most easily interpreted link between dimensions 8components) and variables. The coefficient on each factor tells how much that factor affects the value of a variable. The higher the coefficient the more it affects the size of the variable. These coefficients are in factor analysis called factor loadings. Taking their square provides a correlation between the factor and the variable. To determine which variables correlate highly with each factor we take a look at the coefficients larger than 0.5.

Rotated Component Matrix[a]

	Rescaled Component			
	1	2	3	4
M2.d. Mistrust in Traditional Authorities	.185	.853	.142	.133
M2.e. Mistrust in group village headmen	.248	.866	.155	.145
M2.f. Mistrust in village headmen	.163	.764	.333	.247
M2.j. Mistrust in police	.293	.425	.050	.597
M2.k. Mistrust in traders	.215	.074	.130	.868
M2.l. Mistrust in teachers	.099	.437	.408	.465
M2.m.Mistrust in school administrators	.121	.372	.426	.546
M2.n. Mistrust in religious leaders	.109	.321	.616	.222
M3.a. Mistrust in family members	.205	.172	.654	.031
M3.b. Mistrust in relatives	.222	.014	.803	.122
M3.c. Mistrust in people in own village	.496	.181	.542	.211
M3.d. Mistrust in people outside the village	.690	.098	.167	.188
M3.e. Mistrust in people of same ethnic group	.808	.233	.252	.043
M3.f. Mistrust in people outside ethnic group	.814	.151	.221	.149
M3.g. Mistrust in people from same church/mosque	.408	.237	.511	.100
M3.h. Mistrust in people not from same church/mosque	.796	.154	.146	.178

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. The component matrix is also called factor loading matrix

The first factor correlates mostly with indicators of mistrust in people from outside the local village, the second correlates with indicators of mistrust in traditional authorities, the third factor correlates with indicators of mistrust in local people and religious leaders, and the forth factor correlates with indicators of mistrust in modern authorities.

Mistrust in teachers is somewhat ambiguous in this picture by correlating moderately with 3 factors, but mostly with modern authorities. It may seem to correspond to the kind of social reality they live in. They are a bit of every group except outsiders.

All four components provide a meaningful interpretation and will be useful in subsequent studies.

OPPGÅVE 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.

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.

But 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.

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

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

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 varialbles 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

Comparing model 3 with model 1 we have H=5 and K=10. Hence the
$\chi^2_5 = 114.647 - 100.413 = 14.234$

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 and impact on the general probability of trusting people.

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 = Sqrt(Wald)/SE_{Sex} 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,

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. With b_{Sex} = -0.396 and $SE_{b(Sex)}$ = 0.514 we find
-0.396 – 1.96* 0.514 < β_{Sex} < -0.396 + 1.96*0.514
-1.40344 < β_{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 0.95.

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.

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		

	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

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

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*MistMA244
[If model 2 is used for women: L = -0.159 -1.014*2.83231 -0.848*2.71402 +0.117*(-2.33551) -0.037*MistMA244 (=-5.691 if also MistMA244 gets the value to minimize)]
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*MistMA244

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

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

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:

- 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.

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.

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. J. Scott Long (1997. Regression Models for Categorical and Limited Dependent Variables. London: Sage; page 53-54) adds the advice that there ought to be at least 10 observations per parameter estimated, and concurs that considerably more is needed if the dependent variable is skewed. In model 3 there are 10 parameters estimated, but the dependent variable is not particularly skewed with the smallest category comprising 46.3% of the cases. The 20 missing cases may of course affect this. But no information is available on that. 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 1.55IQR from the mean: 3036, 5011, 5007, 1011, 4208, and 12008. The leverage statistic adds 1011, 3017, 16006, 2014, and 6252. Only 1011 is high on both so this case and 3017, and 3036 might be inspected.

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
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