Rita Kurban

Professor Aslim

SS154

26 November 2018

Homework 3

Question 1 - Extramarital Affairs

Model specification

All variables seem to be reasonable for the model, that's why I didn't leave out any of them. However, there might be an issue of omitted variables since the dataset is not exhaustive. For example, we don't have any information on the persons' feelings. If being deeply in love decreases one's probability to cheat on their partner, the model would never be able to account for that. Authors that studied extramarital affairs indicated other relevant factors such as growing up in urban areas *(Smith, 2012)*, women's culture and upbringing, including affairs of their parents *(Allen et al., 2005.)*

I decided to build a linear model that doesn't include any interaction, quadratic, and cubic terms to ensure the interpretability of model outcomes. I treated all the variables as continuous, including occupation since it is coded as ordinal, using the reverse Hollingshead classification scale *(Fair, 1978.)*

Collinearity is another issue that might threaten the validity of the model since it results in inflated standard errors. The correlation matrix below suggests that there is a medium level of correlation among some of the variables. For example, the age is correlated with the number of

years married (0.8941.) However, excluding these variables from the dataset might increase the omitted variable bias while their effect on the SE doesn't play such a significant role in this case since I'm not performing hypothesis testing. That's why I decided to keep them.[1]

	v1	v2	v3	v4	v5	v6	v7	v8
v1	1.0000							
v2	-0.1111	1.0000						
v3	-0.1290	0.8941	1.0000					
v4	-0.1292	0.6739	0.7728	1.0000				
v5	0.0788	0.1366	0.1327	0.1418	1.0000			
v6	0.0799	0.0280	-0.1091	-0.1419	0.0322	1.0000		
v7	0.0395	0.1061	0.0418	-0.0151	0.0357	0.3823	1.0000	
v8	0.0277	0.1626	0.1281	0.0867	0.0041	0.1839	0.2012	1.0000

Table 1. Correlation Matrix

a) Probit and Logit Models

The estimates of both probit and logit are statistically significant for all variables except *v4* and *v8*. Since probit and logit have a non-linear functional form where linear predictor appears inside the nonlinear function, I calculated the margins to interpret the magnitude of the coefficients in the tables below.

[1] **#specification:** I figured out what model specification is applicable in this situation by looking at several factors such as collinearity, omitted variables, and functional form. Later in the paper, I also performed the VIF test, checked the goodness of fit, and argued why the Poisson model is more appropriate for the count data compared to the linear regression.

```
. quietly logit $ylist $xlist

. margins, dydx(*)

Average marginal effects                        Number of obs     =      6,366
Model VCE    : OIM

Expression   : Pr(A), predict()
dy/dx w.r.t. : v1 v2 v3 v4 v5 v6 v7 v8
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
v1	-.1307846	.0048635	-26.89	0.000	-.1403169	-.1212523
v2	-.011047	.0018606	-5.94	0.000	-.0146938	-.0074003
v3	.0200929	.0019457	10.33	0.000	.0162793	.0239064
v4	-.0007731	.0057737	-0.13	0.893	-.0120893	.0105431
v5	-.0685161	.0061621	-11.12	0.000	-.0805935	-.0564386
v6	-.0071627	.0028226	-2.54	0.011	-.012695	-.0016304
v7	.0292639	.0061686	4.74	0.000	.0171736	.0413543
v8	.0022648	.0041867	0.54	0.589	-.0059409	.0104705

Table 2. Logit margins

```
. quietly probit $ylist $xlist

. margins, dydx(*)

Average marginal effects                        Number of obs     =      6,366
Model VCE    : OIM

Expression   : Pr(A), predict()
dy/dx w.r.t. : v1 v2 v3 v4 v5 v6 v7 v8
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
v1	-.1320021	.0049242	-26.81	0.000	-.1416532	-.1223509
v2	-.0109044	.0018455	-5.91	0.000	-.0145215	-.0072873
v3	.0202055	.0019405	10.41	0.000	.0164022	.0240089
v4	-.001213	.0057824	-0.21	0.834	-.0125462	.0101203
v5	-.0686007	.0061467	-11.16	0.000	-.0806481	-.0565533
v6	-.0073076	.0028101	-2.60	0.009	-.0128152	-.0017999
v7	.0293697	.006158	4.77	0.000	.0173002	.0414392
v8	.0020283	.0041519	0.49	0.625	-.0061092	.0101658

Table 3. Probit margins

Model Comparison & Outcome Interpretation[2]

There are no substantial differences in the outcomes of the two models. Therefore, both logit and probit are valid models to use in this case. The reason for that is that probit assumes a normal distribution of errors. Since our sample size is pretty big (6366 obs.,) the central limit theorem holds. Therefore, the difference between probit and logic can only be indicated in rare edge cases.

Rating of the marriage is negatively correlated with extramarital affairs (around -0.13 for both models.) The sign makes sense since people who have happy marriages wouldn't cheat on their partners. The coefficients of both probit and logit can be interpreted as percent changes (a point rise in marriage rating decreases the probability of affair by 15%.) Age is also negatively correlated with affairs (-0.1) which might mean that young people are more likely to have extramarital relationships. The number of years married (0.2) is positively correlated since the probability of cheating is higher when people live together longer. Being religious (-0.07) and high level of education (-0.007) make people more faithful. As expected, the coefficient in front of the occupation is positive (0.03) which indicates that having a high social status has a **negative** effect on the number of affairs (because of the reverse ordering.)

b) *Ordered Probit Model*

Ordered probit is used in cases where a categorical variable has an inherent order. In this model, the intercepts that I will identify are the cut-off points that decide which category an

[2]**#correlation, #regression:** I estimated the effect of control variables on extramarital affairs, interpreted the correlation coefficients and provided a detailed explanation on why the estimates are plausible and reliable or why they are not. This regression indicates that there is a correlation between the variables. However, it cannot prove any causal relationships since we don"t have a control group and a treatment variable. I also touch upon the difference between statistical and practical significance in the ordered probit model where most of the effects are very small in magnitude.

observation falls into. The interpretation of the coefficients and marginals are the same as in standard probit and logit models. I used a similar model specification since the model is very similar to the previous two. However, the dependent variable (marriage ranking) is susceptible to biases because of the subjective nature of self-reported rankings. Some people might feel cultural pressure to report their marriage to be more successful than it's actually is, or, instead, report a very poor marriage ranking after a minor quarrel with their partner.

```
. oprobit v1 v2 v3 v4 v5 v6 v7 v8

Iteration 0:   log likelihood = -7926.4872
Iteration 1:   log likelihood = -7820.1782
Iteration 2:   log likelihood = -7820.1602
Iteration 3:   log likelihood = -7820.1602

Ordered probit regression                       Number of obs     =      6,366
                                                LR chi2(7)        =     212.65
                                                Prob > chi2       =     0.0000
Log likelihood = -7820.1602                     Pseudo R2         =     0.0134
```

v1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
v2	-.0051292	.0047036	-1.09	0.275	-.0143481	.0040896
v3	-.0078737	.005051	-1.56	0.119	-.0177735	.0020262
v4	-.0569631	.015174	-3.75	0.000	-.0867037	-.0272226
v5	.1309812	.0160991	8.14	0.000	.0994276	.1625348
v6	.0275244	.0072503	3.80	0.000	.0133142	.0417347
v7	.0138844	.0161007	0.86	0.388	-.0176724	.0454412
v8	.0280724	.0106927	2.63	0.009	.0071151	.0490297
/cut1	-1.639317	.1315933			-1.897235	-1.381398
/cut2	-.940552	.1273708			-1.190194	-.6909099
/cut3	-.2011641	.1264895			-.449079	.0467507
/cut4	.7674965	.1267434			.5190839	1.015909

Table 4. Ordered probit output

The regression indicates that age, number of years married, and occupation are not statistically significant while the rest of the variables are significant at 99%.

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
v2						
_predict						
1	.0001803	.0001659	1.09	0.277	-.0001448	.0005054
2	.0004787	.0004396	1.09	0.276	-.0003829	.0013402
3	.000867	.0007955	1.09	0.276	-.0006922	.0024262
4	.0004787	.0004396	1.09	0.276	-.000383	.0013404
5	-.0020047	.0018383	-1.09	0.275	-.0056077	.0015984
v3						
_predict						
1	.0002768	.0001791	1.55	0.122	-.0000743	.0006278
2	.0007348	.0004723	1.56	0.120	-.0001909	.0016605
3	.0013309	.0008546	1.56	0.119	-.0003441	.0030059
4	.0007348	.0004731	1.55	0.120	-.0001925	.0016622
5	-.0030773	.0019741	-1.56	0.119	-.0069465	.0007919
v4						
_predict						
1	.0020023	.0005588	3.58	0.000	.0009071	.0030975
2	.0053158	.0014345	3.71	0.000	.0025043	.0081274
3	.0096287	.0025798	3.73	0.000	.0045724	.014685
4	.0053163	.0014449	3.68	0.000	.0024843	.0081483
5	-.0222631	.0059307	-3.75	0.000	-.033887	-.0106393
v5						
_predict						
1	-.0046041	.0006865	-6.71	0.000	-.0059496	-.0032586
2	-.0122232	.0015892	-7.69	0.000	-.0153379	-.0091085
3	-.0221402	.0027906	-7.93	0.000	-.0276096	-.0166709
4	-.0122244	.0016447	-7.43	0.000	-.0154479	-.0090008
5	.0511919	.0062935	8.13	0.000	.0388569	.0635269
v6						
_predict						
1	-.0009675	.0002678	-3.61	0.000	-.0014924	-.0004426
2	-.0025686	.0006853	-3.75	0.000	-.0039117	-.0012255
3	-.0046526	.0012322	-3.78	0.000	-.0070676	-.0022376
4	-.0025688	.0006912	-3.72	0.000	-.0039236	-.001214
5	.0107575	.0028339	3.80	0.000	.0052032	.0163117
v7						
_predict						
1	-.000488	.0005671	-0.86	0.389	-.0015995	.0006234
2	-.0012957	.0015039	-0.86	0.389	-.0042432	.0016518
3	-.0023469	.0027227	-0.86	0.389	-.0076833	.0029895
4	-.0012958	.001504	-0.86	0.389	-.0042436	.001652
5	.0054265	.0062927	0.86	0.388	-.006907	.0177599
v8						
_predict						
1	-.0009868	.0003849	-2.56	0.010	-.0017411	-.0002325
2	-.0026197	.001004	-2.61	0.009	-.0045875	-.000652
3	-.0047452	.0018125	-2.62	0.009	-.0082977	-.0011927
4	-.00262	.001008	-2.60	0.009	-.0045956	-.0006443
5	.0109717	.0041791	2.63	0.009	.0027807	.0191626

Table 5. Ordered probit margins

Robustness Check

The table shows the cut-offs for each category: -1.64, -.94, -0.20, and 0.77. Any predicted outcome variable below -1.64 predicts a marriage ranking of 1, while anything above 0.77 predicts a marriage ranking of 5. To check for robustness, I ran the regression without insignificant variables and got estimates that have the same signs but different magnitudes:

/cut1	-1.538252
/cut2	-.8424062
/cut3	-.1049438
/cut4	.8622145

Table 6. Estimates for the reduced model

I used AIC to find out which model specification is better and came to the conclusion that the first model (AIC = 15662.32), even though it includes insignificant variables, has a lower AIC and is, therefore, more reliable than the second model (AIC = 15674.73). Omitted variables can be a reason why the second model is not 100% robust to the addition of independent variables. For example, excluding a person's age and occupation is not a clever decision since they definitely play an important role in the regression. This example shows that despite being insignificant, some variables are still important in the regression and cannot be omitted. At the same time, it proves that the first model might also be susceptible to omitted variable bias since the addition of new independent variables could have further changed the estimates.

Results

The marginal mean for variables in the table is the predicted mean of the dependent variable for each of the individual outcomes. For example, *v4* indicates that each additional child increases women's likelihood to report the marriage ranking of 1 by 0.2% and the outcome of 5

by 0.5%.[3] The sign and the magnitude of all the statistically significant variables make sense. However, none of the estimates has a magnitude that is higher than |0.05|. Thus, the practical significance of these variables is questionable.

Question 2 - Hospital Visits

Functional Form

The following analysis uses the Poisson regression to examine the predictors for the number of hospital visits which is a count variable. The data has 23 independent variables. However, it might still be not enough to eliminate omitted variable bias since many of these variables include demographic characteristics: gender, age, marital status, etc. But not that many variables account for the actual causes of going to the hospital, for example, chronic diseases and pregnancy.

A big number of independent variables increases the risk of multicollinearity, which, as stated above, can inflate the standard errors. I decided to remove some of the variables that represent similar things. For example, we have two variables for being handicapped: the degree and a dummy. I only keep the continuous variable. The household income *(hhninc)* is correlated with the occupation dummy variables *(bluec, whitec, self)*. Therefore, I'll only use the variable working that indicates whether the person is employed. A similar situation occurs with the educational dummies *(haupts, reals, fachs, abitur)* all of which are well represented by the *educ* variable that indicates the years of education. I tested the multicollinearity among the remaining variables and saw a decrease in the mean VIF score from 29.64 to 6.34. Here is the correlation

[3] **#appliedeconometrics:** I implemented the Ordered Probit Model, specified its functional form, compared different models and chose the best one. I also interpreted the coefficients of the model and performed a robustness check that made it possible for me to identify the importance of variables that seemed to be insignificant.

matrix of the remaining variables, none of the variables have a correlation of more than 57% (it seems like all government officials have public insurance):

	working	bluec	whitec	self	beamt	docvis	hospvis	public	addon
working	1.0000								
bluec	0.3533	1.0000							
whitec	0.4210	-0.3706	1.0000						
self	0.1483	-0.1437	-0.1667	1.0000					
beamt	0.1927	-0.1610	-0.1855	-0.0726	1.0000				
docvis	-0.1238	-0.0302	-0.0429	-0.0486	-0.0457	1.0000			
hospvis	-0.0254	-0.0073	-0.0125	-0.0072	-0.0117	0.1361	1.0000		
public	-0.0774	0.1930	0.1353	-0.1223	-0.5709	0.0629	0.0102	1.0000	
addon	0.0127	-0.0510	0.0341	0.0346	0.0037	-0.0014	0.0055	0.0497	1.0000

Table 7. Correlation matrix

	Delta-method					
	dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0008425	.0004045	-2.08	0.037	-.0016354	-.0000497
hsat	-.0204589	.0021398	-9.56	0.000	-.0246527	-.016265
handper	.0006182	.00018	3.43	0.001	.0002653	.0009711
educ	-.0054521	.0019137	-2.85	0.004	-.0092029	-.0017013
hhninc	6.23e-06	1.90e-06	3.29	0.001	2.52e-06	9.95e-06
working	-.0044872	.0087524	-0.51	0.608	-.0216417	.0126672
docvis	.0035447	.000446	7.95	0.000	.0026706	.0044187
addon	.0340134	.0211638	1.61	0.108	-.0074669	.0754937

Table 8. Poisson Margins

The margins indicate that there is a negative effect of age, health satisfaction, education, and employment on the number of hospital visits. Their magnitude is not that high, mostly in the decimal points and statistically significant except employment and insurance. Doctor visits, insurance, income, and the degree of handicap have a positive correlation with hospital visits and have a similar magnitude *(Table 8.)* Some of the variables have unpredicted sign and magnitude, for example, age. Even though its magnitude is so low, it seems like the sign should be positive

because it's quite apparent that old people have more health issues and need to visit hospitals more often.

Regression Model

Before jumping to the discussion of the possible reasons for the problems mentioned above, let's compare the output of the OLS model and see why the Poisson model is more appropriate for the count data.

hospvis	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0016841	.0005165	-3.26	0.001	-.0026964	-.0006717
hsat	-.0232905	.0036651	-6.35	0.000	-.0304743	-.0161067
handper	.0013861	.0004146	3.34	0.001	.0005735	.0021988
educ	-.0054149	.001869	-2.90	0.004	-.0090781	-.0017516
hhninc	7.88e-06	2.86e-06	2.76	0.006	2.28e-06	.0000135
working	-.0069345	.010855	-0.64	0.523	-.0282108	.0143418
docvis	.0170734	.0035225	4.85	0.000	.0101692	.0239776
addon	.0410029	.0325852	1.26	0.208	-.0228657	.1048716
_cons	.3430048	.0429309	7.99	0.000	.2588581	.4271516

Table 9. Regression Outputs

The magnitude and the signs of the effects are almost identical for linear regression estimates and the Poisson model which indicates that, in some cases, the regression can estimate the magnitude pretty well despite the count outcome variable. Nonetheless, Poisson regression is more appropriate than linear regression for the count data. The estimated coefficients of the maximum-likelihood Poisson estimator don't depend on the assumption that $E(y_j) = Var(y_j)$, so even if the assumption is violated, the estimates of the coefficients $b_0, b_1, ..., b_k$ are unaffected

which cannot be said about the linear regression model. Apart form that, standard errors and confidence intervals in the regression model are pretty far away from the Poisson model, therefore, invalid. The reason for that is that they are by default calculated assuming normality for the outcome which doesn't hold in case of the count outcome variable. Many distributions of count data are positively skewed with many observations in the data set having a value of 0. The high number of 0's in the data set prevents the transformation of a skewed distribution into a normal one.

Goodness of Fit and Overdispersion

To check the goodness of fit, I used *estat gof* function. I conclude that the model poorly fits the data because the goodness-of-fit chi-squared test is statistically significant (1.0).

```
estat gof

      Deviance goodness-of-fit =  19414.24
      Prob > chi2(27312)       =    1.0000

      Pearson goodness-of-fit  =  116156.9
      Prob > chi2(27312)       =    0.0000
```

Table 10. Goodness of fit

Excess variation is a common problem that arises when applying the Poisson regression model. It means that the variance is larger than expected (larger than the mean). There are four sources of excess variation. If it comes from an incorrect functional form or omitted variables, those are systematic errors that can be solved. If the excess variation comes from a measurement error or dependence within counts, there is overdispersion, and a solution might be to turn to the negative binomial distribution instead. I tried to specify the functional form as precisely as

possible. However, there might still be an issue of omitted variables as stated above. Therefore, overdispersion can definitely be a problem. However, there is no clear way of checking for that since there is no 100% guarantee that the functional form is correct and no variables are omitted. To conclude, the model has clear evidence of excess variation, but not overdispersion.

References:

Allen, E., Atkins, D., Baucom, D., Snyder, D., Gordon, K., & Glass, S. (2005). Intrapersonal, interpersonal, and contextual factors in engaging in and responding to extramarital involvement. Clinical Psychology: Science and Practice, 12(2), 101-130. doi:10.1093/clipsy.bpi014

Fair, R. C. (1978). A theory of extramarital affairs. Journal of Political Economy, 86(1), 45-61.

Smith, I. (2012). Reinterpreting the economics of extramarital affairs. Review of Economics of the Household, 10(3), 319-343.

Stata Code:

```
*1
clear import delimited /Users/ritakurban/Downloads/TableF17-2.csv
gen A = yrb
replace A = 1 if yrb>0

global ylist A
global xlist v1 v2 v3 v4 v5 v6 v7 v8

describe $ylist $xlist
summarize $ylist $xlist
corr

* Probit model
probit $ylist $xlist

* Logit model
logit $ylist $xlist


* Marginal effects (at the mean and average marginal effect)

quietly logit $ylist $xlist
margins, dydx(*)

quietly probit $ylist $xlist
margins, dydx(*)


* Ordered probit model coefficients
oprobit v1 v2 v3 v4 v5 v6 v7 v8
estat ic
oprobit v1 v4 v5 v6 v8
estat ic
margins, dydx(*) atmeans

*2
clear
import delimited /Users/ritakurban/Downloads/q2.csv, delimiter(";")

poisson hospvis id female year age hsat handdum handper hhninc hhkids educ married haupts
reals fachhs abitur univ working bluec whitec self beamt docvis public addon
vif, uncentered

poisson hospvis age hsat handper educ hhninc working docvis addon, vce(robust)
vif, uncentered
margins, dydx(*) atmeans
estat gof

regress hospvis age hsat handper educ hhninc working docvis addon, vce(robust)
```