Second Assignment

Malgorzata, Kevin, Johannes

09th of February, 2019

Microeconometrics 2018/2019

Assignment 2

Group Members:

- Johannes Wagner
  Humboldt-Universität Berlin - ID 598797 - Msc Statistics
  wagnejoh@hu-berlin.de
- Malgorzata Paulina Olesiewicz
  Humboldt-Universität Berlin - ID 598939 - Msc Statistics
  malgorzata.paulina.olesiewicz@student.hu-berlin.de
- Kevin Hoppe
  Humboldt-Universität Berlin - ID 598247 - Msc Statistics
  Kevin.Hoppe1@web.de

Approximate individual contributions:

Task	Johannes	Malgorzata	Kevin
1	1/3 theory + programming	1/3 theory	1/3 theory
2	1/3 theory	1/3 theory	1/3 theory
3	1/3 theory	1/3 theory + programming	1/3 theory
4	1/3 theory	1/3 theory	1/3 theory+ programming

Task 1: Choosing the variables

The choice of variables should be grounded on both theory and empirical evidence. Our theoretical perspective is guided by the assumption that higher social, cultural and economical capital leads to a lower probability of rejection. First, we chose three variables representing relevant social characteristics given above theory that are stored in the variables "education", "marital status", and "wealth". We then used forward stepwise selection to identify those covariates with the highest empirical relevance for the dependent variable "rejection". From the resulting list of nine empirically highly relevant variables, we picked four that are also theoretically relevant, thus obtaining a total of seven regressors.

To create the logit model (table 1) we have chosen following variables with regard to our theory:

Indicator	Variable	Scale	Expected effect
Education (cultural capital)	sch	[0,1]	-
Marital Status (social capital)	married	[0,1]	-
Wealth in Dollar (economic capital)	netw	[-7919, 28023]	-
Private mortgage insurance (economic capital)	inson	[0,1]	-
Former bankruptcy (economic capital)	pubrec	[0,1]	+
Ethnic majority (social & cultural capital)	white	[0,1]	-
Ratio of obligations to income (economic capital)	obrat	[0, 95]	+

Given our seven explanatory variables, we can now estimate the conditional probability of a rejection. In our sample, the estimated probability of rejection for an individual with average characteristics is $\mathrm{P}\left(Y = 1 | X = \overline{\mathbf{x}}\right) = 9.04\%$.

Task 2: Results interpretation

Just looking at our estimated coefficients we can not say which one has the biggest effect on our dependent variable. That is because we can not directly compare the magnitude of coefficients from explanatory variables with different scales. For interpretation, we usually want to use coefficients as marginal probability effects given a little change in the explanatory variable. Since the effect of a "unit change" can mean quite different effects given the scale of the variable, you cannot compare variables with different scales. For example, given our variables, a unit change in the binary variable "pubrec" represents the difference between just two options (Yes/No) while a unit change in "obrat" just represents one step on a scale of many options. Furthermore, coefficients are random variables and have to be tested for them being significantly different from zero and from each other. Also, before making any interpretations, the coefficients should be transformed into marginal probability effects.

Task 3: Sensitivity and Specificity

The notion of sensitivity and specificity is used to describe how accurately a model predicts the binary outcomes. Sensitivity is the fraction of correctly predicted positive ($Y = 1$) outcomes and specificity describes the fraction of correctly predicted negative ($Y = 0$) outcomes. We aim to find

Table 1: Logit Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7905235	0.3839173	-7.2685528	0.0000000
sch	-0.1264278	0.1794267	-0.7046211	0.4810460
married	-0.4227597	0.1584850	-2.6675060	0.0076417
netw	0.0000635	0.0000622	1.0221728	0.3066991
inson	4.8774009	0.7489776	6.5120781	0.0000000
pubrec	1.8356245	0.2081772	8.8176070	0.0000000
white	-1.2235890	0.1703907	-7.1810767	0.0000000
obrat	0.0517525	0.0088016	5.8798722	0.0000000

an optimal equilibrium between the two fractions, which allows us the best possible simultaneous prediction of both outcomes.

In our prediction, we use c as a threshold above which the outcome is predicted to be positive. Consequently, any result of our logit model which will be equal or below the threshold will be precited to be negative.

The most logical starting point for the binary response prediction model is $c = 0.5$

Threshold	Sensitivity	Specificity
0.5	0.2418033	0.9885189

For $c = 0.5$, we observe that almost all (99%) negative responses but only 24% of the positive responses are predicted correctly. Since a threshold c that is closer to 1 results in a higher specificity (the likelihood that we will predict negative responses increases) we will decrease the threshold to $c = 0.3$.

Threshold	Sensitivity	Specificity
0.3	0.3647541	0.9615385

We observe some improvement in the accuracy of predicted positives (sensitivity now at 36%) while keeping the specificity at a high level (now 96%). We are going in the right direction. To find the optimal prediction threshold c we used the "InformationValue" package.

Optimal Cut off	Sensitivity	Specificity
0.1176232	0.6516393	0.8254879

The optimal cut-off level is $c = 0.118$ where 65% of positive outcomes and 83% of negative outcomes are predicted correctly.

Taks4: Multinomial Logit

The difference in the coefficient estimates is zero up to at least the fourth decimal place:

(Intercept)	-1.7e-05
educ	-2.6e-05
marry	5.0e-06
insur	5.8e-05
netw	0.0e+00
bankr	-4.0e-06
white	-3.0e-06
oblig	1.0e-06

This is because the multinomial logit model (MNL) reduces to the binomial logit model in case of a binomial dependent variable as can be seen in the formulas. In the MNL, the probability π_{ij} of individual i choosing alternative j is given by:

$$\pi_{ij_{multinomial}} = \frac{\exp(x_i'\beta_j)}{\sum_{r=1}^{J} \exp x_i'\beta_j}.$$

Compare this to the binomial logit model, where the probability π_i of individual i picking alternative $j = 1$ is given by:

$$\pi_{i_{binomial}} = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

In the MNL, due to identification constraints, β_1 is fixed at 0. This establishes $j = 1$ as the reference category. For the remaining $J-1$ categories, β_j coefficients are estimated. If the dependent variable has only two categories, i.e. $J = 2$, this means that only one β and one π_i need to be calculated (for the one category that is not the reference category) so the index j in π_{ij} and β_j can be dropped. Since the constraint for $\beta_1 = 0$ means that $\exp(x_i'\beta_1)$ evaluates to 1, this reduces the denominator in the MNL to $1 + \exp(x_i'\beta_2)$. After dropping the now obsolete index of the coefficient vector, the two formulas given above are equal.