Second Assignment

Malgorzata, Kevin, Johannes

09th of February, 2019

Microeconometrics 2018/2019

Assesment 2

Group Members:

Johannes Wagner, ID: 598797, Msc Statistics, wagnejoh@hu-berlin.de

Malgorzata Paulina Olesiewicz, ID:598939, Msc Statistics, malgorzata.paulina.olesiewicz@student.hu-berlin.de

Kevin Hope, ID: 598247, Msc Statistics, Kevin.Hoppe1@web.de

Approximate individual contributions:

Task	Johannes	Malgorzata	Kevin
1	33 % theory	33 % theory	33% theory
2	100% theory	-	-
3	-	100% theory + progr	-
4	??	??	100% progr + ??

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

Task 1 : Chosing the variables for the model

The choice of variables should be grounded on both theory and empirical evidence. Our theoretocal perspective is guided by the assumption that higher social, cultural and economical capital leads to a lower probability of receiving a loan rejection. Therefore we choose three variables, which represent relevant social characteristics given our theory: marriage, networth and educational level.
Our next step was to look at a simple linear regression model for all variables of the data set and run a forward stepwise selection process on them. We obtained nine highly relevant variables and choose another four variables from this list, with respect to their suitability to our theory.

To create the logit model we have chosen followig variables:

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]	-
Appr. private insurance (economic capital)	inson	[0,1]	-
Experience of bunkruptcy (economic capital)	pubrec	[0,1]	+
Ethnical majority (social & cultural capital)	white	[0,1]	-
Ration of obligations vs. income (ecomoic capital)	obrat	[0, 95]	+

Results interpretation

You can not directly compare the magnitude of the coefficients of explanatory variables with different scales, since the effect of a "unit change" can mean quite different effects given the scale of the variable. However, in case the variables have the same scale or are standardized, it is possible to compare the size of their coefficients since the logit function is strictly monotoniously increasing. Before interpreting the magnitude, you should be aware that the coefficients are random representations of the true parameter and the confidence level of two or more parameters being different needs to be tested. Also coefficients need to be transformed into probabilities before they can be interpreted in a meaningfull way.

Sensitivity and Specificity

The notion of sensitivity and specificity is used to describe how accurately model predicts the binary outcomes. Sensitivity describes the fraction of correctly predicted positive (Y=1) outcomes and specificity describes the fraction of correctly predicted negative (Y=0) outcomes. Our aim is to find an optimal equilibrium between the two fractions, which would allow us simultaneously the best possible prediction of both outcomes.

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

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

```
table(dataSelect$reject)
```

```
##
##    0    1
## 1742  244
```

```
predicted=predict(model,dataSelect, type="response")
sen_1=sensitivity(dataSelect$reject,predicted,threshold = 0.5)
spec_1=specificity(dataSelect$reject,predicted,threshold = 0.5)
print(sen_1)
```

```
## [1] 0.2418033
```

```
print(spec_1)
```

```
## [1] 0.9885189
```

With threshold c=0.5 we can observe very high fraction of negative responds being correctly predicted (98%) but only 29% positive responds have been predicted correctly.Since the closer the threshold to 1 the higher specificity (the likelihood that we will predict negative respond increases) we will decrease the threshold to 0.3.

```
sen_2=sensitivity(dataSelect$reject,predicted,threshold = 0.3)
spec_2=specificity(dataSelect$reject,predicted,threshold = 0.3)
print(sen_2)
```

```
## [1] 0.3647541
```

```
print(spec_2)
```

```
## [1] 0.9615385
```

We can observe some improvment in prediction of positive outcomes to 47% and prediction ofnegative respond has still been very accurate - 95%. We are going in the right direction.To find out optimal cut off threshold for the prediction we have used the "InformationValue" package.

```
data_2=data.frame(dataSelect$reject, predicted)
data_noNA=data_2[complete.cases(data_2), ]###getting rid off N/A in prediction

a= optimalCutoff(actuals = data_noNA$data.reject, predictedScores =data_noNA$predicted,optimise
```

```
sen_3=sensitivity(data$reject,predicted,threshold = a$optimalCutoff)
```

```
## Warning in actual_dir == 1 & predicted_dir == 1: longer object length is
## not a multiple of shorter object length
```

```
spec_3=specificity(data$reject,predicted,threshold =a$optimalCutoff)
```

```
## Warning in actual_dir != 1 & predicted_dir != 1: longer object length is
## not a multiple of shorter object length
```

```
print(a$optimalCutoff)
```

```
## [1] NA
```

```
print(sen_3)
```

```
## [1] 0
```

```
print(spec_3)
```

```
## [1] 0
```

The optimal cut off level is c=0.088 where 76% of positive respons is being predicted correctly and 72% of negative respons is being predicted correctly.

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.