University of Tartu, Estonian Center of Industrial Mathematics, The European Consortium for Mathematics in Industry, COST Action TD1409 Mathematics for industry network

4 - 8 February 2019, Tartu, Estonia

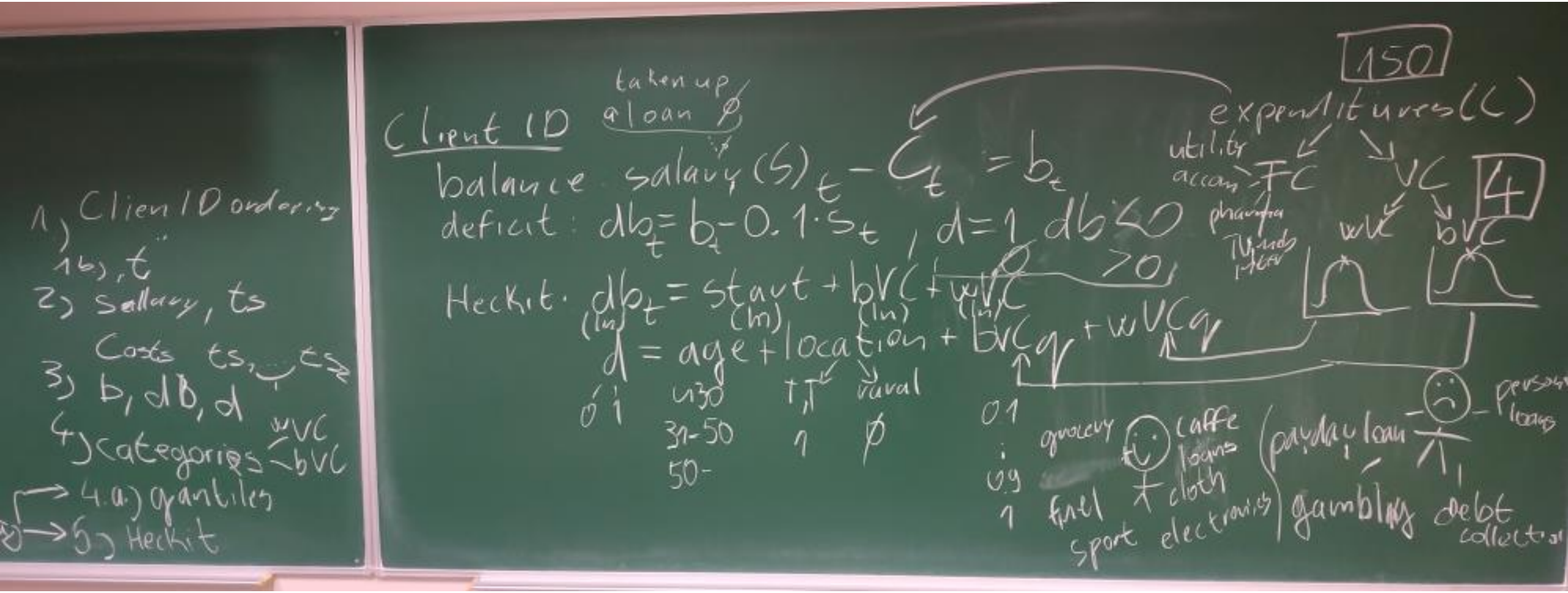
Estonian Study Group with Industry – ESGI 151

Problem number 4 – LHV Bank fully automated personal finance tool

4a. Drivers of financial balance

Dr. habil. Gábor Dávid KISS
associate professor
University of Szeged
Faculty of Economics and Business Administration
Hungary

Original concept



Data and Model

Software:
SQL, Matlab, Gretl

Dataset (2018):

- CLIENT_ID (N=500);
- GENDER (F/M);
- AGE (u30, 31-50, a50);
- ADDRESS (T-T, rural);
- STARTING_BALANCE (€);
- CATEGORY (N=150);
- SUM (€)

Model

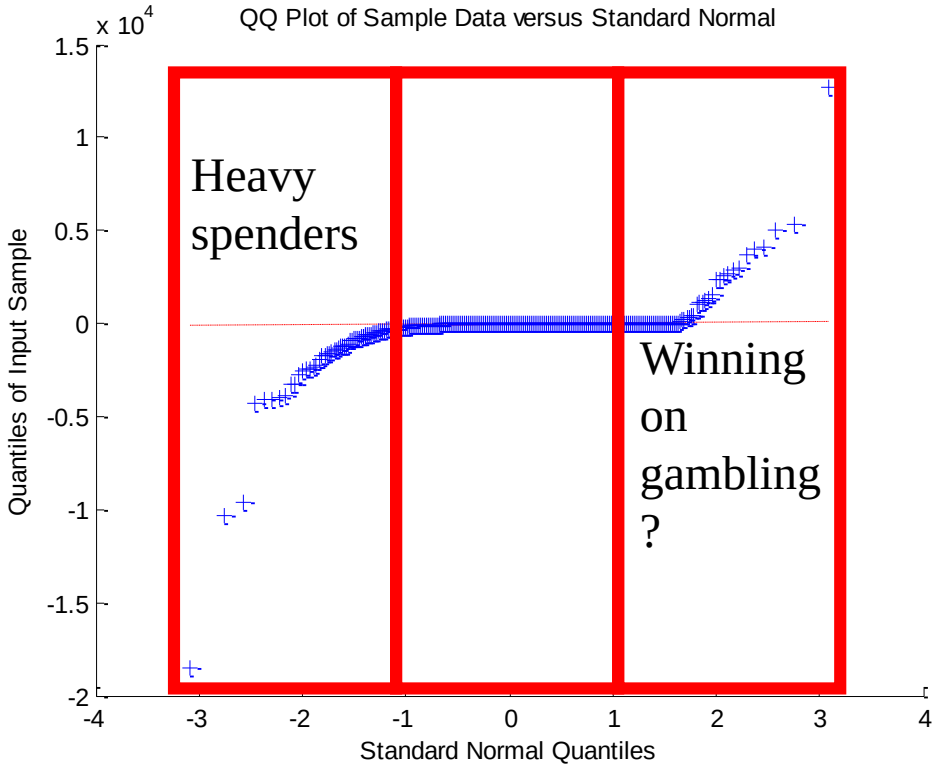
- Balance=Salary – Costs (€, dummy)
- Costs=FC+VC
- VC=whitelisted VC + blacklisted VC
- whitelisted VC (€, 33-66-99%)
  - Groceries, Theatre, Cafes, Fuel, Sport, Books, Household products, Clothing
- Blacklisted VC (€, 33-66-99%)
  - Gambling, Debt collection, Payday loan, Personal loan

Steps

1. Random sampling (N=500) from clients (Client ID) – SQL server
2. String-to-value conversion – Matlab (*165.5 seconds*)
    1. Gender (0-1), Address (0-1), Age (1-2-3), w_cat (0-1), b_cat (0-1),
    2. salary day, sums of salary, w VC, b VC and total costs
    3. New ID and time codes for clients and their transactions (panel regression req.)
    4. Determining 33% spending quantiles for w/b VC
    5. Determining balance and state of deficit
    6. Panel construction
        1. Ln of € spending
        2. Variables: ln_start,ln_w_sum,ln_b_sum,ln_salary_sum,ln_balance,ln_db, deficit_d,w_cat_d,b_cat_d,w_VCq,b_VCq,age,address,gender,ID,Time
3. Heckit model fitting – Gretl

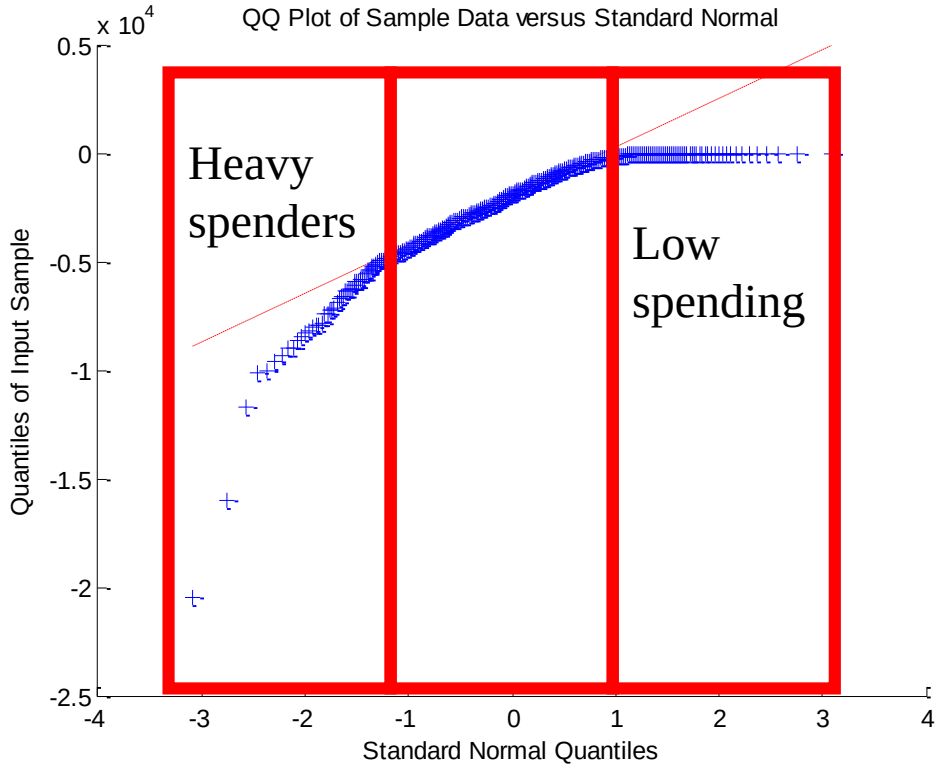
33-33-33% quantiles

QQ plot of blacklisted VC



-1 0 1

QQ plot of whitelisted VC



-1 0 1

Model

- Heckit panel (two-step) regression:
  - The first step is the selection mechanism that sorts and explains the state of balance (Z* = 1 cases)

Selection mechanism (step 1):

$$Z_i^* = w_i'\gamma + u_i;\ Z_i = 1 \text{ if } Z^* > 0, \text{ otherwise 0; where } Z_i = 1 \text{ denotes extreme returns,} \quad (7)$$

$$\text{Prob}(Z_i = 1 | w_i) = \phi(w_i'\gamma) \text{ and } \text{Prob}(Z_i = 0 | w_i) = 1 - \phi(w_i'\gamma). \quad (8)$$

  - The second step is a regression model which explains the drivers of the budgetary balance

Regression model (2nd step):

$$Y_i = X_i'\beta + \varepsilon_i \quad Z_i = 1 \text{ , where } Y_i \text{ is the dynamic conditional correlation.} \quad (9)$$

  - The error terms: uncorrelated, no specification error, lambda significant
- Applied Heckit model:
  - Regression: ln_balance = const. + βln_start + βln_bVC + βln_wVC +ε *labda*
  - Selection eq.: balance_dummy=const. + βage + βaddress + βgender + βwVCq + βbVCq+u
    - Assumptions: $\beta<0$, $\beta>0$, $\beta\sim0$

Greene, W.H. (2003): *Econometric Analysis*. Prentice Hall. Pearson. New Jersey (p. 784, (20-22))
Wooldridge, J.M. (2012): *Introductory Econometrics: A Modern Approach*. Cengage Learning, Mason.

Im, Pesaran and Shin (2003) test of unit root

	first 500	second 500	third 500	forth 500
tbar:	-88.7787	-87.6416	-88.0869	-88.0621
Wbar:	-378.7174	-373.7820	-375.7149	-374.9341
Wbar_pvalue:	**0**	**0**	**0**	**0**
Zbar:	-388.5131	-383.4496	-385.4327	-385.3220
Zbar_pvalue:	**0**	**0**	**0**	**0**
critical:	[-2.3263-1.6449-1.2816]	[-2.3263-1.6449-1.2816]	[-2.3263-1.6449-1.2816]	[-2.3263-1.6449-1.2816]
tbar_DF:	-162.2975	-161.3663	-160.0640	-161.1429
Zbar_DF:	-715.8993	-711.7524	-705.9531	-710.7577
Zbar_DF_pvalue:	**0**	**0**	**0**	**0**
pi:	[14x1double]	[14x1double]	[14x1double]	[14x1double]
tstats_ind:	[14x1double]	[14x1double]	[14x1double]	[14x1double]
pmax:	2	2	2	2
Ti:	[14x1double]	[14x1double]	[14x1double]	[14x1double]
model:	'model with intercept'	'model with intercept'	'model with intercept'	'model with intercept'

No unit root

Im, Pesaran and Shin (2003): Testing for Unit Root in Heterogeneous Panels. *Journal of Econometrics*, 115, 53-74

Results

Two-step Heckit								
	first 500		second 500		third 500		forth 500	
	coeff.	p-value	coeff.	p-value	coeff.	p-value	coeff.	p-value
Dependent variable: ln_db								
const	-0,9349	0,0000***	0,0039	1,92e-040***	-0,9259	0,0000***	-0,9100	0,0000***
ln_start	-0,0254	7,29e-079***	-0,0070	9,27e-016***	-0,0623	1,72e-071***	-0,0701	0,0000***
ln_b_sum	-0,0155	0,1023	-0,0106	0,0002***	-0,0161	0,5287	-0,0561	1,92e-07***
ln_w_sum	0,0857	1,88e-127***	-0,0131	1,45e-078***	0,0748	5,84e-017***	0,0745	1,05e-069***
lambda	**0,2170**	**0,0000*** **	**0,0352**	**1,75e-060*** **	**0,5339**	**2,14e-077*** **	**0,2554**	**4,76e-138*** **
Selection equation (deficit_d)								
const	1,4499	0,0000***	1,5410	0,0000***	2,1928	0,0000***	2,1806	0,0000***
age	-0,0887	8,56e-042***	-0,0073	0,2406	-0,2075	7,34e-157***	-0,1871	1,80e-156***
address	0,2270	2,10e-144***	-0,0548	2,53e-011***	0,0190	0,0630*	-0,1295	3,72e-042***
gender	0,0437	7,00e-07***	-0,0572	5,55e-013***	-0,1465	1,53e-047***	-0,2259	7,83e-132***
w_VCq	-0,5015	0,0000***	-0,1265	9,55e-125***	-0,2657	0,0000***	-0,3009	0,0000***
b_VCq	-0,2921	0,0000***	-0,1870	2,33e-168***	-0,0949	7,90e-029***	0,0195	0,0148**
Total observations:	257564		264327		250769		256649	
Censored observations:	13503 (5,2%)		16009 (6,1%)		8271 (3,3%)		11146 (4,3%)	

Interpretation

Initial surplus →deficit
Blacklisted VC → deficit
Whitelisted VC → sufficit
Lambda: significant

Older ages →sufficit
Metropolitan area → ?
Males → sufficit (?)
Heavy VC spending → d
(whitelisted > blacklisted)

Two-step Heckit								
	first 500		second 500		third 500		forth 500	
	coeff.	p-value	coeff.	p-value	coeff.	p-value	coeff.	p-value
Dependent variable: ln_db								
const	-0,9349	0,0000***	0,0039	1,92e-040***	-0,9259	0,0000***	-0,9100	0,0000***
ln_start	-0,0254	7,29e-079***	-0,0070	9,27e-016***	-0,0623	1,72e-071***	-0,0701	0,0000***
ln_b_sum	-0,0155	0,1023	-0,0106	0,0002***	-0,0161	0,5287	-0,0561	1,92e-07***
ln_w_sum	0,0857	1,88e-127***	-0,0131	1,45e-078***	0,0748	5,84e-017***	0,0745	1,05e-069***
lambda	**0,2170**	**0,0000*** **	**0,0352**	**1,75e-060*** **	**0,5339**	**2,14e-077*** **	**0,2554**	**4,76e-138*** **
Selection equation (deficit_d)								
const	1,4499	0,0000***	1,5410	0,0000***	2,1928	0,0000***	2,1806	0,0000***
age	-0,0887	8,56e-042***	-0,0073	0,2406	-0,2075	7,34e-157***	-0,1871	1,80e-156***
address	0,2270	2,10e-144***	-0,0548	2,53e-011***	0,0190	0,0630*	-0,1295	3,72e-042***
gender	0,0437	7,00e-07***	-0,0572	5,55e-013***	-0,1465	1,53e-047***	-0,2259	7,83e-132***
w_VCq	-0,5015	0,0000***	-0,1265	9,55e-125***	-0,2657	0,0000***	-0,3009	0,0000***
b_VCq	-0,2921	0,0000***	-0,1870	2,33e-168***	-0,0949	7,90e-029***	0,0195	0,0148**
Total observations:	257564		264327		250769		256649	
Censored observations:	13503 (5,2%)		16009 (6,1%)		8271 (3,3%)		11146 (4,3%)	

Discussion

- Theoretical model: simple but works
- Theoretical model can be sophisticated
  - More VC categories
  - Deeper regional breakdown
  - Cash-preference included
- Quantile panel regression
- Additional sample tests (increasing sample size – RAM)