Johan Löfgren
PMU/MIÖ

Logistisk regression av bortfall

GGS 2021

2021-09-08

Inledning

Logistisk regression används då vi vill undersöka samband mellan en dikotom responsvariabel Yoch en eller fler förklarande X-variabler. I vårt fall är responsvariabeln olika uppdelningar av urvalsmängden och de förklarande variablerna är registervariabler från Befolkningsregistret och Utbildningsregistret. Här låter vi X-variablerna vara dummy-variabler som indikatorer för respektive kategori av registervariabeln.

Modeller

Urvalet delas upp i kategorierna *ej svarande (I)* och *svarande (II)*. Vi skapar en regressionsmodell där responsvariabeln definieras som

$$Y = \begin{cases} 1, & om\ ej\ svarande\ (I) \\ 0, & om\ svarande\ (II) \end{cases}$$

Notera att övertäckningen tas bort från modellen. Ej svarande utgörs av olika kategorier av bortfall, exempelvis:

- Ej inkommen
- Vägrare
- Förhindrad medverkan
- Skyddad identitet
- Saknar adress
- Fel person svarat

Ej inkommen utgör den absoluta majoriteten av bortfallet (ca 95 procent), så därför gör inga separata analyser för olika delar av bortfallet. Övriga kategorier av bortfall är för små, för att några intressanta analyser ska kunna göras.

Förklarande variabler

Följande registervariabler har använts som förklarande variabler:

Tabell 1 Förklarande variabler

Variabel (benämning)	Kategorier (koder)
Kön	1 = Man, 2= Kvinna
Ålder	(Ålder: år) 1 = 18 – 24 2 = 25 – 34 3 = 35 – 44 4 = 45 – 54 5 = 55 – 59 Dummy-variabler för respektive kategori
Födelseland	1 = Födda i Sverige 2 = Födda i Norden 3 = Födda i övriga Europa 4 = Födda i övriga världen Dummy-variabler för respektive kategori
Utbildningsnivå	1 = Låg (förgymnasial, uppgift saknas) 2 = Medel (gymnasial) 3 = Hög (eftergymnasial) Dummy-variabler för respektive kategori
Stockholm	1 = Folkbokförd i Stockholms län 2 = Folkbokförd i övriga Sverige Dummy-variabler för respektive kategori

Resultat

I detta avsnitt redovisas resultaten av SAS-körningarna med regressionsmodellen Y. Körningarna har begränsats till det nationella urvalet.

För att tolka modellen används oddskvoten som antar värden större än 0. Värden större än 1 innebär att oddset, dvs att sannolikheten för det inträffade, ökar för den aktuella gruppen när övriga grupper är konstanta. Värden mindre än 1 motsvarar ett minskat odds eller sannolikhet.

Bortfallet fördelat på förklarande variabler

I tabell 2 visas antal och andel bortfall i det nationella urvalet fördelat på de förklarande variablerna. Det totala bortfallet är 73,0 procent.

Bortfallsandelen minskar med ökad utbildningsnivå och med ökande ålder. Det är ett större bortfall för män än för kvinnor. Det förekommer också relativt stor skillnad i bortfallsandel för de olika födelselandsgrupperna. Stockholms län har i stort sett samma bortfallsandel som övriga Sverige.

Tabell 2 Antal och andel svar/bortfall fördelat på förklarande variabler

	Totalt bortfall		
	Svar	**Bortfall**	**Total**
Kön			
Kvinnor	4 462	9 993	14 455
	30,9	69,1	100
Män	3 620	11 894	15 514
	23,3	76,7	100
Åldersgrupp			
18-24	878	3 502	4 380
	20,1	79,9	100
25-34	1 575	6 104	7 679
	20,5	79,5	100
35-44	1 934	5 240	7 174
	27,0	73,0	100
45-54	2 332	4 799	7 131
	32,7	67,3	100
55-59	1 363	2 242	3 605
	37,8	62,2	100
Födelseland			
Sverige	6 872	15 408	22 280
	30,8	69,2	100
Övriga Norden	131	302	433
	30,3	69,7	100
Övriga Europa	487	1 997	2 484
	19,6	80,4	100
Övriga Världen	592	4 180	4 772
	12,4	87,6	100
Utbildningsnivå			
Låg	674	4 471	5 145
	13,1	86,9	100
Mellan	2 765	9 859	12 624
	21,9	78,1	100
Hög	4 643	7 557	12 200
	38,1	61,9	100
Stockholm			
Stockholms län	2 062	5 350	7 412
	27,8	72,2	100
Övriga Sverige	6 020	16 537	22 557
	26,7	73,3	100
Total	8 082	21 887	29 969
	27,0	73,0	100

SAS-utskrift för statistisk modell

Logistisk regression, Johan Löfgren, SCB 2021-09-07

The LOGISTIC Procedure

Model Information	
Data Set	WORK.GGS_ANALYS
Response Variable	Totalt bortfall
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	29969
Number of Observations Used	29969

Response Profile		
Ordered Value	**Totalt bortfall**	**Total Frequency**
1	1	21887
2	0	8082

Probability modeled is Totalt bortfall=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	34942.196	32201.243
SC	34950.504	32300.938
-2 Log L	34940.196	32177.243

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	2767.4804	11	<.0001
Score	2597.3467	11	<.0001
Wald	2361.8035	11	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.0745	0.0381	796.3129	<.0001
Man	1	0.2804	0.0275	104.2543	<.0001
18-24 år	1	0.0256	0.0502	0.2609	0.6095
25-34 år	1	0.3910	0.0403	94.0669	<.0001
45-54 år	1	-0.2699	0.0384	49.5042	<.0001
55-59 år	1	-0.5495	0.0457	144.3641	<.0001
Född i Övr Norden	1	0.0995	0.1098	0.8213	0.3648
Född i Övr Europa	1	0.6190	0.0546	128.5602	<.0001
Född i Övr världen	1	1.0627	0.0480	490.4610	<.0001
Stockholms län	1	-0.0522	0.0318	2.6927	0.1008
Låg utbildning	1	0.3705	0.0491	56.9130	<.0001
Hög utbildning	1	-0.8448	0.0301	790.0569	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Man	1.324	1.254	1.397
18-24 år	1.026	0.930	1.132
25-34 år	1.478	1.366	1.600
45-54 år	0.763	0.708	0.823
55-59 år	0.577	0.528	0.631
Född i Övr Norden	1.105	0.891	1.370
Född i Övr Europa	1.857	1.669	2.067
Född i Övr världen	2.894	2.634	3.179
Stockholms län	0.949	0.892	1.010
Låg utbildning	1.448	1.316	1.595
Hög utbildning	0.430	0.405	0.456

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	68.4	**Somers' D**	0.385
Percent Discordant	29.9	**Gamma**	0.392
Percent Tied	1.8	**Tau-a**	0.152
Pairs	176890734	**c**	0.693

Av oddskvoterna att döma är sannolikheten för bortfall större för män än för kvinnor.

Sannolikheten för bortfall minskar med ökande ålder (åldersgrupp) i jämförelse med referensgruppen 35-44 år.

Sannolikheten för bortfall ökar med minskande ålder i jämförelse med referensgruppen 35-44 år. Den yngsta åldersgruppen har dock ingen signifikant skillnad i sannolikhet för bortfall, jämfört med referensgruppen.

Sannolikheten för bortfall är hög för de som är födda i övriga världen. Även födda i övriga Europa har en hög sannolikhet för bortfall jämfört med referensgruppen (födda i Sverige).

Låg utbildning har en högre sannolikhet för bortfall jämfört med referensgruppen.