ODHADOVANIE VÝSLEDKOV VOLIEB

JOZEF JANOVSKÝ (273898)

1. ZDROJ DÁT

Používame verejne dostupné dáta, ktoré sú súčasťou výskumného projektu „The Record of American Democracy, 1984-1990“[1], agregujúceho množstvo informácií o americkej spoločnosti a politike. Dáta v ňom pochádzajú jednak zo sčítania ľudu z roku 1990 a doplnené sú výsledkami volieb rôznych druhov v období rokov 1984-1990. Tieto údaje sú dostupné na úrovni oblastí využitých práve pri danom sčítaní ľudu. Štát, na ktorom analýzu prevedieme si vyberáme podľa toho, aby mal čo najviac týchto oblastí, čo vydeľuje Pennsylvániu, ktorá ich má 2674. Na základe údajov zo sčítania ľudu sa budeme snažiť pomocou modelu logistickej regresie odhadnúť či, v danej oblasti získali v prezidentských voľbách v roku 1988 viac hlasov republikáni alebo demokrati. Následne overíme možnosť využitia tohto modelu pri predikcii výsledkov ďalších v datasete zahrnutých volieb.

2. POPIS DÁT

Z datasetu sme vybrali 10 premenných a doplnili ich o tri skonštruované premenné (získané transformáciami premenných obsiahnutých v datasete), ktoré dohromady predstavujú vysvetľujúce premenné. Vysvetľovanou premennou je „víťazstvo“ republikánov v danej oblasti. „Predvýber“ vysvetľovacích premenných z pôvodného počtu 3360 premenných v datasete (vrátane dodatočne skonštruovaných) prebiehal na základe korelácií s vysvetľovanou premennou a výsledkov parciálneho stepwise logisticko-regresného modelovania[2], a to tak aby sme vzali do úvahy štatisticky najvýznamejšie najvýznamnejšie premenné.

URBANperc1 - Percento obyvateľstva danej oblasti, žijúceho v meste.[3]
HOUSSIperc1 - Percento domácností poberajúce dávky sociálneho zabezpečenia.[4]
HOUINCave – Priemerný príjem domácnosti v danej oblasti v $ za rok 1989.[5]
P0280030 – Počet obyvateľov starších než 65 rokov, ktorí doma rozprávajú iným jazykom než angličtinou, španielčinou, či ázijskými alebo pacifickými jazykmi a zároveň vôbec neovládajú angličtinu.
P0310014 – Počet obyvateľov starších než 5 rokov, ktorí doma hovoria juhoslovanským jazykom.
P0340033 – Počet obyvateľov, ktorí majú viacerých predkov z Juhoslávie.
P0350033 – Počet obyvateľov, ktorí majú jedného predka z Juhoslávie.

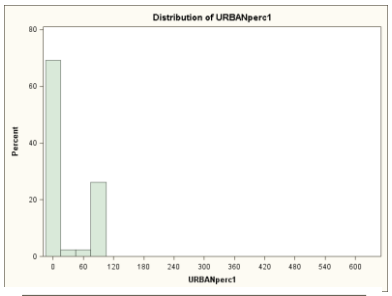
[1] Gary King; Bradley Palmquist; Greg Adams; Micah Altman; Kenneth Benoit; Claudine Gay; Jeffrey B. Lewis; Russ Mayer; and Eric Reinhardt. 1997. ``The Record of American Democracy, 1984-1990,'' Harvard University, Cambridge, MA [producer], Ann Arbor, MI: ICPSR [distributor].
[2] Máme tým na mysli, že sme povolili prítomnosť približne 800 vysvetľujúcich premenných v modeli a spustili stepwise procedúru. Tento postup sme zvolili z praktických technických výpočtových dôvodov (pri pustení procedúry na všetkých 3360 premenných počítač zamrzával na príliš dlhú dobu).
[3] V reči premenných datasetu ide o: P0060001/P0010001.
[4] V reči premenných datasetu: P0940001/P0050001.
[5] V reči premenných datasetu: P0980001/P0050001.

P1000001 **–** Celkový príjem farmárov samo-zamestnávateľov za rok 1989 v $.
P124B012 **–** Počet rodín s príjmom pod úrovňou chudoby, v ktorých je hlavou domácnosti matka bielej pleti bez manžela, bez detí mladších než 18 rokov.
P087A019 **–** Počet domácností, v ktorých je jej hlavou 35 až 44-ročný človek bielej pleti, zarábajúci menej než 5000$ ročne.
P1270007 **–** Počet domácností nad úrovňou chudoby, v ktorých je jej hlavou 65 až 74-ročný muž bez manželky.
P1270023 **–** Počet domácností pod úrovňou chudoby, v ktorých je jej hlavou 65 až 74-ročná žena bez manžela.
H025A001 **–** Medián roku stavby bytovej jednotky.
rep84, rep86, rep88, rep90 **–** Binárna premenná nadobúdajúca hodnotu 1 ak v príslušných voľbách do Snemovne reprezentantov získali republikáni viac hlasov v danej oblasti a hodnotu 0, ak získali menej hlasov než demokrati.
rep84p, rep88p **-** Binárna premenná nadobúdajúca hodnotu 1 ak v príslušných prezidentských voľbách získali republikáni viac hlasov v danej oblasti a hodnotu 0, ak získali menej hlasov než demokrati.

3. DESKRIPTÍVNA ANALÝZA „SUROVÝCH“ DÁT

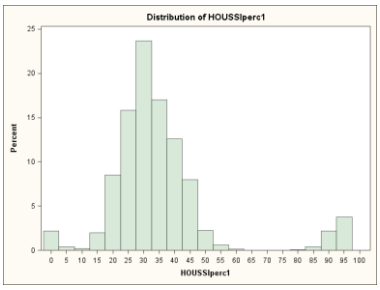
Variable	Mean	Std Dev	Minimum	Maximum	N	N Miss	Median
URBANperc1	28.5	45.9	0.0	642.2	2581	93	0.0
HOUSSIperc1	35.3	17.7	0.0	100.0	2581	93	31.9
HOUINCave	21928.0	11467.1	0.0	126561.6	2581	93	21201.4
P0280030	75.9	196.3	0.0	5079.0	2585	89	0.0
P0310014	53.6	166.8	0.0	4500.0	2585	89	0.0
P0340033	66.3	176.1	0.0	3600.0	2585	89	0.0
P0350033	84.0	441.5	0.0	18900.0	2585	89	0.0
P1000001	207001.8	505334.8	-151474.0	11053324.0	2585	89	38165.0
P124B012	79.1	146.4	0.0	1701.0	2585	89	0.0
P087A019	131.6	205.4	0.0	2700.0	2585	89	60.0
P1270007	93.8	172.6	0.0	2965.0	2585	89	0.0
P1270023	32.8	104.2	0.0	2178.0	2585	89	0.0
H025A001	1953.8	67.8	0.0	1987.0	2585	89	1958.0
rep84	0.6	0.5	0.0	1.0	2586	88	1.0
rep86	0.6	0.5	0.0	1.0	2591	83	1.0
rep88	0.6	0.5	0.0	1.0	2598	76	1.0
rep90	0.7	0.5	0.0	1.0	2317	357	1.0
rep84p	0.8	0.4	0.0	1.0	2595	79	1.0
rep88p	0.7	0.5	0.0	1.0	2603	71	1.0

Po všeobecnom prehľade základných vlastností všetkých premenných, uvedených v tabuľke, sa teraz na ne pozrieme jednotlivo. Ako vyplýva zo zavedenia vysvetľujúcich premenných, každá z nich je intervalového typu. Deskriptívna analýza dát pritom poukázala aj na to, že je možné, že sú dáta nejakým spôsobom systematicky zaťažené technickou chybou. Existuje v nich totiž príliš veľa násobkov čísla 30, a to pri prakticky všetkých premenných. Podľa priloženej dokumentácie však neexistuje dôvod na takýto charakter dát. Kompletne prerábať celý dataset však presahuje naše možnosti a aj vzhľadom na ilustratívny charakter tejto analýzy pristupujeme k dátam tak, že ide o vedomú konštrukciu a v ďalšom sa zameriame na klasické ošetrenie kvality dát, odstránením extrémnych hodnôt, chýbajúcich hodnôt a pod.



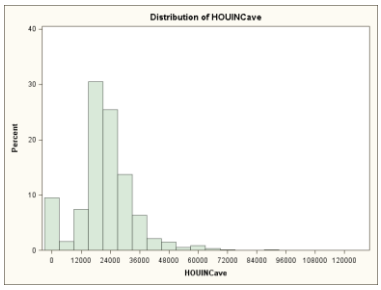
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0.00000	1742	249	99.8728	1
2	1.14924	1	250	100.0000	588
3	1.28076	1	251	139.4988	1
4	2.31990	1	252	493.6015	1
5	2.75041	1	253	642.2018	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	93	3.48	100.00



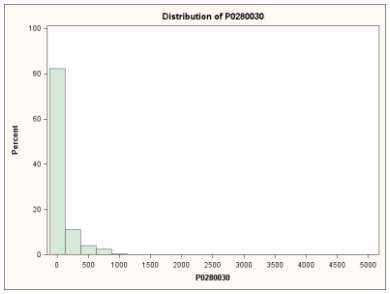
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0.000000	4	2431	96.2567	1
2	0.731645	1	2432	96.3039	1
3	0.818640	1	2433	96.3455	1
4	0.906736	1	2434	96.7222	1
5	0.953778	1	2435	100.0000	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	93	3.48	100.00



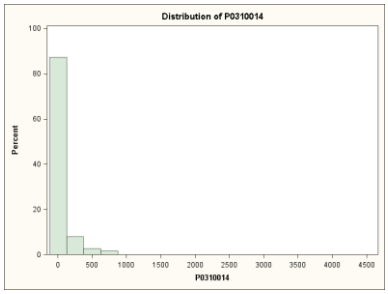
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0.0000	1	2576	71431.9	1
2	83.3889	1	2577	79199.3	1
3	100.2042	1	2578	90250.9	1
4	157.8333	1	2579	92685.2	1
5	185.4095	1	2580	126561.6	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	93	3.48	100.00



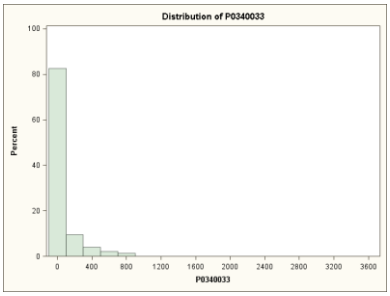
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1798	79	870	8
2	30	19	80	900	5
3	31	3	81	987	1
4	32	6	82	2700	1
5	33	4	83	5079	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



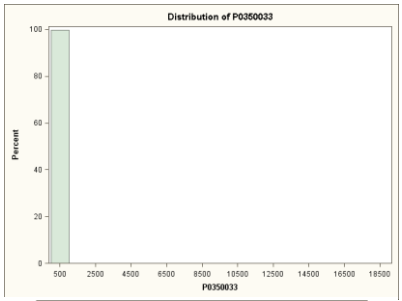
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	2028	83	840	6
2	30	6	84	870	1
3	31	3	85	900	2
4	32	2	86	1800	1
5	33	4	87	4500	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



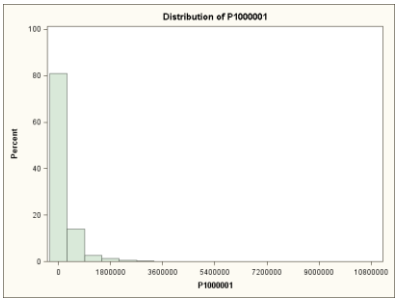
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1915	86	810	7
2	30	9	87	840	2
3	31	5	88	900	3
4	32	2	89	1800	3
5	33	1	90	3600	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



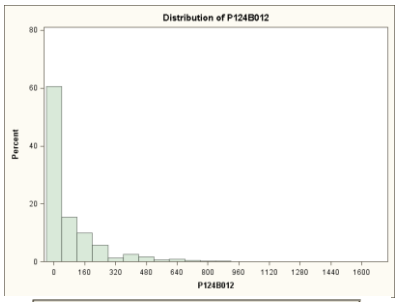
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1797	116	1800	1
2	30	8	117	2700	2
3	31	5	118	4500	1
4	32	5	119	7200	1
5	33	7	120	18900	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



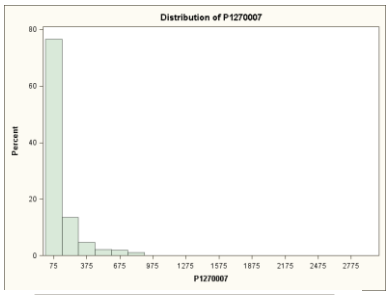
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	-151474	1	1813	4116621	1
2	-135000	1	1814	5015025	1
3	-112469	1	1815	5285254	1
4	-99248	1	1816	8072304	1
5	-89991	1	1817	11053324	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



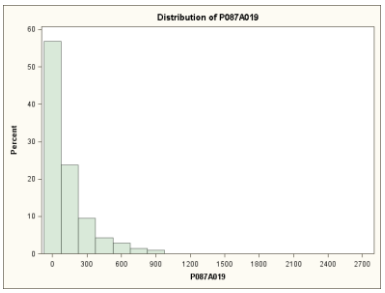
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1501	59	810	6
2	30	46	60	840	2
3	32	5	61	870	3
4	33	1	62	900	3
5	34	3	63	1701	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



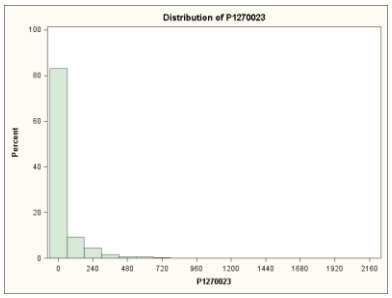
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1050	84	900	9
2	30	33	85	976	1
3	31	3	86	1800	2
4	32	8	87	2609	1
5	33	1	88	2700	2

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



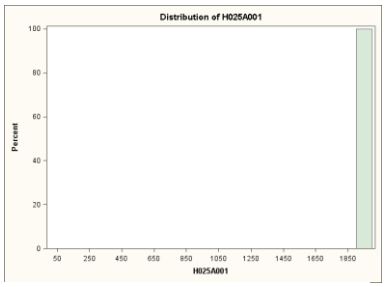
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1420	68	810	10
2	30	23	69	829	1
3	31	4	70	870	6
4	32	1	71	1800	1
5	33	3	72	2965	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



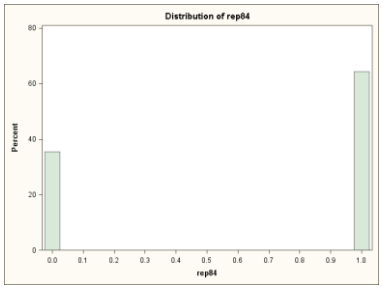
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	2118	32	720	2
2	30	23	33	780	1
3	32	1	34	810	3
4	36	2	35	900	1
5	38	1	36	2178	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



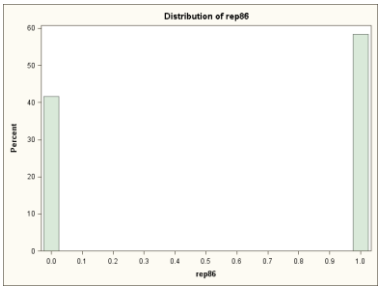
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	3	46	1983	4
2	1939	590	47	1984	5
3	1940	23	48	1985	2
4	1941	48	49	1986	1
5	1942	19	50	1987	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	89	3.33	100.00



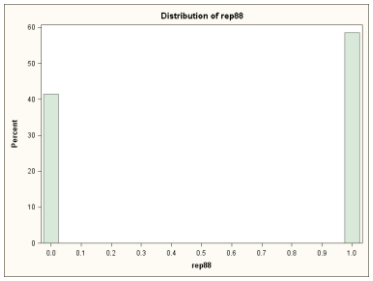
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	919	1	0	919
2	1	1667	2	1	1667

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	88	3.29	100.00



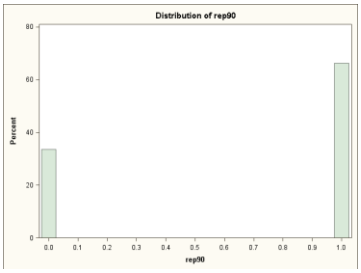
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1077	1	0	1077
2	1	1514	2	1	1514

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	83	3.10	100.00



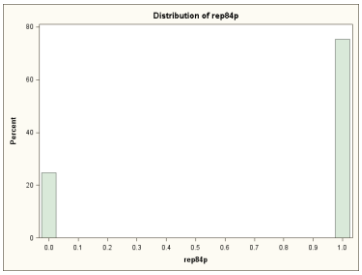
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	1078	1	0	1078
2	1	1520	2	1	1520

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	76	2.84	100.00



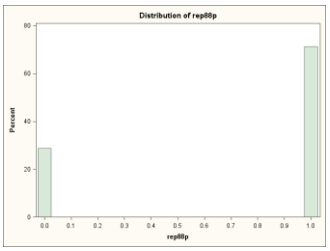
Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	780	1	0	780
2	1	1537	2	1	1537

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	357	13.35	100.00



Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	641	1	0	641
2	1	1954	2	1	1954

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
	79	2.95	100.00



Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	748	1	0	748
2	1	1855	2	1	1855

Missing Values			
Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
	71	2.66	100.00

4. TRANSFORMÁCIA DÁT

V prvom rade sme zmenili mierku u premenných, ktoré boli výrazne odlišné v hodnotách od ostatných a potom sme ošetrili extrémne a chýbajúce hodnoty:

- Premennú HOUINCave sme delíme stami a premennú P1000001 tisícmi.
- V premennej URBANperc1 dosahujú tri prípady hodnoty vyše 100, takže ich nahradíme hodnotou 100.
- V premennej HOUINCave priradíme 5 najvyšším hodnotám hodnotu 6. najvyššieho.
- V premennej P0280030 priradíme trom najvyšším pozorovaniam hodnotu 900.
- V premennej P0310014 priradíme dvom najvyšším pozorovaniam hodnotu 900.
- V premennej P0340033 priradíme štyrom najvyšším pozorovaniam hodnotu 900.
- V premennej P0350033 priradíme siedmim najvyšším pozorovaniam hodnotu 900.
- V premennej P1000001 priradíme dvom najvyšším pozorovaniam hodnotu 3. najvyššej.
- V premennej P124B012 priradíme pozorovaniu s najvyššou hodnotou hodnotu 900.
- V premennej P087A019 priradíme 6 pozorovaniam s najvyššou hodnotou hodnotu 900.
- V premennej P1270007 priradíme 2 pozorovaniam s najvyššou hodnotou hodnotu 870.
- V premennej P1270023 priradíme pozorovaniu s najvyššou hodnotou hodnotu 900.
- V premennej H025A001 priradíme 3 pozorovaniam s hodnotou 0 hodnotu 1939 (rok).
- Odstraňujeme tých 89 pozorovaní, ktoré obsahovali chýbajúce hodnoty pre všetky vysvetľujúce premenné.
- Zvyšné chýbajúce hodnoty vo vysvetľovaných premenných dopĺňame priemermi.

Dostávame tak tabuľku:

Variable	Mean	Std Dev	Minimum	Maximum	N	N Miss	Median
URBANperc1	28.2	43.3	0.0	100.0	2585	0	0.0
HOUSSIperc1	35.3	17.7	0.0	100.0	2585	0	31.9
HOUINCave	218.9	112.1	0.0	707.1	2585	0	212.1
P0280030	73.5	163.2	0.0	900.0	2585	0	0.0
P0310014	51.8	139.8	0.0	900.0	2585	0	0.0
P0340033	64.2	154.1	0.0	900.0	2585	0	0.0
P0350033	71.4	160.0	0.0	900.0	2585	0	0.0
P1000001	203.7	453.7	-151.5	5285.3	2585	0	38.2
P124B012	78.8	143.8	0.0	900.0	2585	0	0.0
P087A019	128.8	183.5	0.0	900.0	2585	0	60.0
P1270007	92.7	161.1	0.0	870.0	2585	0	0.0
P1270023	32.3	96.8	0.0	900.0	2585	0	0.0
H025A001	1956.1	12.8	1939.0	1987.0	2585	0	1958.0
rep84	0.6	0.5	0.0	1.0	2563	22	1.0
rep86	0.6	0.5	0.0	1.0	2561	24	1.0
rep88	0.6	0.5	0.0	1.0	2558	27	1.0
rep90	0.7	0.5	0.0	1.0	2288	297	1.0
rep84p	0.8	0.4	0.0	1.0	2567	18	1.0
rep88p	0.7	0.5	0.0	1.0	2552	33	1.0

Pokračujeme kategorizáciou premenných, a to tak, aby boli počty pozorovaní v nich približne vyrovnané a aby sme docielili vysokú "information value" danej premennej (a vhodné hodnoty „weights of evidence"). Ak je to možné, do kategórií delíme po deciloch, v prípade, že by z týchto dôvodov nebolo takého rozdelenie možné, delíme do menšieho počtu skupín. Ďalej uvádzame tabuľky pre premenné po kategorizácii. Vysvetľujeme pritom premennú rep88p, teda „víťazstvo", či „porážku" republikánskeho kandidáta v prezidentských voľbách 1988.

Medium predictivity (IV = 0.1321, 2 groups)

Attributes of URBANperc1		Total		Democrat		Republican		Measures	
Group of URBANperc1	URBANperc1	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	1742	67.4%	424	55.2%	1318	72.5%	75.7%	-27.30
1	1.1492443325 - 100	843	32.6%	344	44.8%	499	27.5%	59.2%	48.92
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Overpredictive (IV = 0.7539, 10 groups)

Attributes of HOUSSIperc1		Total		Democrat		Republican		Measures	
Group of HOUSSIperc1	HOUSSIperc1	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0 - 21.174004193	258	10.0%	55	5.9%	203	12.3%	78.7%	-72.78
1	21.204819277 - 24.87745098	259	10.0%	42	4.5%	217	13.1%	83.8%	-106.42
2	24.886052871 - 27.636643572	258	10.0%	63	6.8%	195	11.8%	75.6%	-55.18
3	27.653631285 - 29.865771812	259	10.0%	75	8.1%	184	11.1%	71.0%	-31.94
4	29.868228404 - 31.899641577	258	10.0%	83	8.9%	175	10.6%	67.8%	-16.79
5	31.901840491 - 34.393063584	259	10.0%	95	10.2%	164	9.9%	63.3%	3.21
6	34.395750332 - 37.5	260	10.1%	110	11.8%	150	9.1%	57.7%	26.79
7	37.508283632 - 41.206896552	257	9.9%	113	12.2%	144	8.7%	56.0%	33.56
8	41.263157895 - 46.916625555	259	10.0%	169	18.2%	90	5.4%	34.7%	120.81
9	46.959459459 - 100	258	10.0%	124	13.3%	134	8.1%	51.9%	50.05
		2585	**100.0%**	**929**	**100.0%**	**1656**	**100.0%**	**64.1%**	.

Overpredictive (IV = 0.5386, 10 groups)

Attributes of HOUINCave		Total		Democrat		Republican		Measures	
Group of HOUINCave	HOUINCave	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0 - 35.409783316	258	10.0%	111	14.5%	147	8.1%	57.0%	58.02
1	36.130554707 - 155.07543478	259	10.0%	146	19.0%	113	6.2%	43.6%	111.74
2	155.29767241 - 179.09220199	258	10.0%	110	14.3%	148	8.1%	57.4%	56.44
3	179.15354839 - 195.00162712	259	10.0%	94	12.2%	165	9.1%	63.7%	29.85
4	195.02225 - 212.06009288	258	10.0%	81	10.5%	177	9.7%	68.6%	7.95
5	212.06037199 - 231.1350177	259	10.0%	73	9.5%	186	10.2%	71.8%	-7.41
6	231.18693431 - 255.62896257	259	10.0%	45	5.9%	214	11.8%	82.6%	-69.82
7	255.66244949 - 289.46469586	258	10.0%	49	6.4%	209	11.5%	81.0%	-58.94
8	289.56674056 - 342.37201422	259	10.0%	41	5.3%	218	12.0%	84.2%	-80.98
9	342.72412402 - 707.08332429	258	10.0%	18	2.3%	240	13.2%	93.0%	-172.91
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Medium predictivity (IV = 0.1079, 2 groups)

Attributes of P0280030		Total		Democrat		Republican		Measures	
Group of P0280030	P0280030	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	1798	69.6%	451	58.7%	1347	74.1%	74.9%	-23.30
1	30 - 900	787	30.4%	317	41.3%	470	25.9%	59.7%	46.73
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Strong predictivity (IV = 0.3386, 2 groups)

Attributes of P0310014		Total		Democrat		Republican		Measures	
Group of P0310014	P0310014	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	2028	78.5%	468	60.9%	1560	85.9%	76.9%	-34.28
1	30 - 900	557	21.5%	300	39.1%	257	14.1%	46.1%	101.59
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Medium predictivity (IV = 0.1962, 2 groups)

Attributes of P0340033		Total		Democrat		Republican		Measures	
Group of P0340033	P0340033	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	1915	74.1%	461	60.0%	1454	80.0%	75.9%	-28.75
1	30 - 900	670	25.9%	307	40.0%	363	20.0%	54.2%	69.36
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Strong predictivity (IV = 0.3071, 2 groups)

Attributes of P0350033		Total		Democrat		Republican		Measures	
Group of P0350033	P0350033	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	1797	69.5%	393	51.2%	1404	77.3%	78.1%	-41.21
1	30 - 900	788	30.5%	375	48.8%	413	22.7%	52.4%	76.46
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Strong predictivity (IV = 0.3843, 5 groups)

Attributes of P1000001		Total		Democrat		Republican		Measures	
Group of P1000001	P1000001	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	-151.474 - 0	678	26.2%	300	39.1%	378	20.8%	55.8%	63.00
1	0.032 - 11.272	355	13.7%	138	18.0%	217	11.9%	61.1%	40.85
2	11.4 - 87.751	518	20.0%	164	21.4%	354	19.5%	68.3%	9.17
3	87.903 - 283.656	517	20.0%	105	13.7%	412	22.7%	79.7%	-50.59
4	284.212 - 5285.254	517	20.0%	61	7.9%	456	25.1%	88.2%	-115.05
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Weak predictivity (IV = 0.0962, 2 groups)

Attributes of P124B012		Total		Democrat		Republican		Measures	
Group of P124B012	P124B012	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	1501	58.1%	363	47.3%	1138	62.6%	75.8%	-28.15
1	30 - 900	1084	41.9%	405	52.7%	679	37.4%	62.6%	34.44
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Weak predictivity (IV = 0.0986, 4 groups)

Attributes of P087A019		Total		Democrat		Republican		Measures	
Group of P087A019	P087A019	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	1050	40.6%	258	33.6%	792	43.6%	75.4%	-26.04
1	30 - 142	673	26.0%	205	26.7%	468	25.8%	69.5%	3.57
2	150 - 900	862	33.3%	305	39.7%	557	30.7%	64.6%	25.89
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Weak predictivity (IV = 0.0840, 2 groups)

Attributes of P1270007		Total		Democrat		Republican		Measures	
Group of P1270007	P1270007	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	1420	54.9%	344	44.8%	1076	59.2%	75.8%	-27.92
1	30 - 870	1165	45.1%	424	55.2%	741	40.8%	63.6%	30.29
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Weak predictivity (IV = 0.0786, 2 groups)

Attributes of P1270023		Total		Democrat		Republican		Measures	
Group of P1270023	P1270023	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	0	2118	81.9%	569	74.1%	1549	85.3%	73.1%	-14.03
1	30 - 900	467	18.1%	199	25.9%	268	14.7%	57.4%	56.35
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Strong predictivity (IV = 0.4536, 5 groups)

Attributes of H025A001		Total		Democrat		Republican		Measures	
Group of H025A001	H025A001	N	%	N of Democrat	% of Democrat	N of Republican	% of Republican	Republican rate	WOE
0	1939	593	22.9%	273	35.5%	320	17.6%	54.0%	70.23
1	1940 - 1952	444	17.2%	179	23.3%	265	14.6%	59.7%	46.88
2	1953 - 1962	512	19.8%	169	22.0%	343	18.9%	67.0%	15.33
3	1963 - 1968	491	19.0%	75	9.8%	416	22.9%	84.7%	-85.20
4	1969 - 1987	545	21.1%	72	9.4%	473	26.0%	86.8%	-102.13
		2585	**100.0%**	**768**	**100.0%**	**1817**	**100.0%**	**70.3%**	.

Zhrnieme, že 4 premenné vykázali slabú prediktivitu, 3 stredne silnú, 4 silnú a 2 príliš silnú. Najmä premenná HOUSSIperc1 je kvôli svojej nezvyčajne vysokej hodnote IV „podozrivá“. Keďže tá je však jednoducho konštruovaná ako percento domácností, poberajúcich nejakú formu sociálnych dávok, autor nevidí v čom by mohlo dôjsť k chybe pri jej konštrukcii. Zároveň je logické, že tu závislosť existuje. Voličmi Republikánov sú typicky dobre zabezpečení členovia vyššej vrstvy, ktorí sú menej závislí na sociálnom zabezpečení. Takto je možné vysvetliť i mierne „overpredictive“ závislosť pri premennej HOUINCave.

5. MODELY LOGISTICKEJ REGRESIE

Na kategorizovaných dátach teraz vybudujeme logisticko-regresné modely. Najprv budujeme model metódou stepwise s požadovanou významnosťou premennej na vstupe rovnou 0,05 a na udržanie sa v modeli rovnou 0,01. Do modelu (Model 1) zahrňujeme konštantu a do výberu povoľujeme každú z vysvetľujúcich premenných. Výsledky, dosiahnuté po 8 krokoch, prinášajú nasledujúce tabuľky a graf.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.2458	0.1778	570.5281	<.0001
rank_HOUSSIperc1	1	0.1357	0.0242	31.3822	<.0001
rank_P0310014	1	0.8033	0.1327	36.6238	<.0001
rank_P0340033	1	0.7817	0.1323	34.9353	<.0001
rank_P0350033	1	0.7378	0.1295	32.4565	<.0001
rank_P1270007	1	0.3171	0.1109	8.1800	0.0042
rank_HOUINCave	1	0.1922	0.0243	62.7338	<.0001
rank_P1000001	1	0.2583	0.0410	39.6839	<.0001
rank_H025A001	1	0.1218	0.0435	7.8449	0.0051

V modeli zostalo 8 premenných, poradie kategórií v každej z nich sme tiež upravili tak aby sme dostali kladné odhady koeficientov (preto ich mená začínajú „rank_“). Korelácia žiadnej

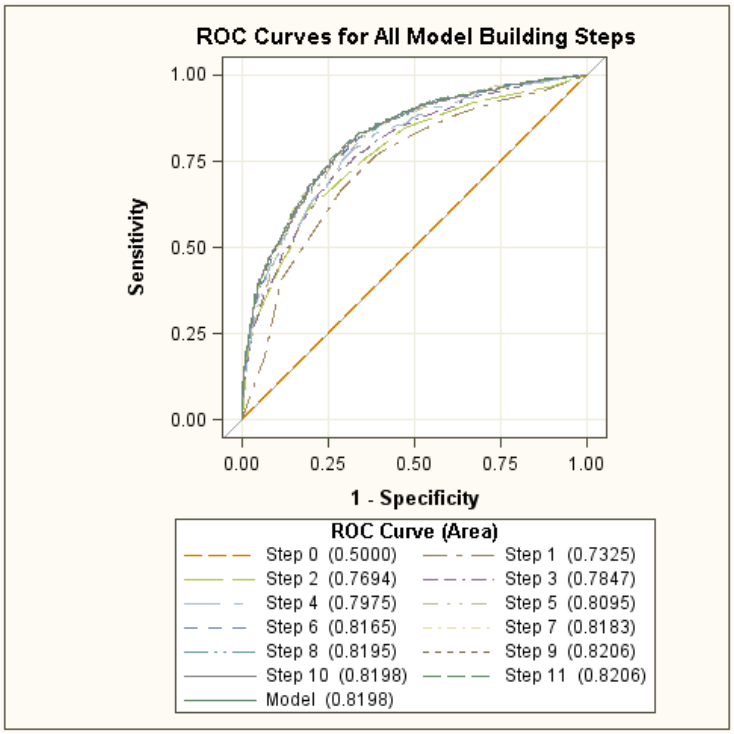
dvojice z nich pritom v absolútnej hodnote nedosahuje 0,5. Udávame aj odhady parametrov pomocou Waldovho testu.

Wald Confidence Interval for Parameters			
Parameter	Estimate	95% Confidence Limits	
Intercept	-4.2458	-4.5942	-3.8974
rank_HOUSSIperc1	0.1357	0.0882	0.1832
rank_P0310014	0.8033	0.5431	1.0634
rank_P0340033	0.7817	0.5225	1.0409
rank_P0350033	0.7378	0.4840	0.9916
rank_P1270007	0.3171	0.0998	0.5344
rank_HOUINCave	0.1922	0.1447	0.2398
rank_P1000001	0.2583	0.1779	0.3387
rank_H025A001	0.1218	0.0366	0.2070

Vzhľadom na silné vzťahy prediktivity jednotlivých premenných nie je prekvapením ani vysoká výpovedná schopnosť modelu, zrejmá z vysokého počtu konkordantných párov oproti diskonkordantným, ako i z následného vysokého Sommersovho D a obdobných štatistík.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	82.0	Somers' D	0.640
Percent Discordant	18.0	Gamma	0.640
Percent Tied	0.0	Tau-a	0.262
Pairs	1335495	c	0.820

Takto vyzerajú ROC krivky:



Zaujímalo nás tiež ako sa model zmení, keď neumožníme prítomnosť dvoch „podozrivých premnných“, teda HOUSSIperc1 a HOUINCave. V šiestom kroku procedúra stepwise terminuje s týmito výsledkami:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.1892	0.1373	539.7966	<.0001
rank_P0310014	1	0.8705	0.1260	47.7312	<.0001
rank_P0340033	1	0.5622	0.1231	20.8623	<.0001
rank_P0350033	1	0.6298	0.1219	26.6804	<.0001
rank_P124B012	1	0.3069	0.1017	9.1130	0.0025
rank_P1000001	1	0.3413	0.0388	77.5088	<.0001
rank_H025A001	1	0.3391	0.0393	74.4072	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	77.9	Somers' D	0.565
Percent Discordant	21.4	Gamma	0.570
Percent Tied	0.7	Tau-a	0.232
Pairs	1335495	c	0.783

Vidíme, že aj odobratí premenných si model stále udržuje vysokú výpovednú hodnotu. Došlo pritom len k ďalšej zmene čo do zastúpenia premenných v modeli.

Nakoniec otestujeme Model 1 aj na ostatných voľbách, ku ktorým máme k dispozícii dáta. Pre stručnosť sa obmedzíme len na uvedenie tabuľku odhadov koeficientov a kvality modelu (ako v predošlom prípade).

Voľby do Snemovne 1984

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7534	0.1331	428.0790	<.0001
rank_HOUSSIperc1	1	0.0549	0.0216	6.4612	0.0110
rank_P0310014	1	0.6732	0.1213	30.8060	<.0001
rank_P0340033	1	0.5447	0.1175	21.4742	<.0001
rank_P0350033	1	0.6809	0.1155	34.7867	<.0001
rank_P1270007	1	0.3164	0.0978	10.4640	0.0012
rank_HOUINCave	1	0.1606	0.0211	57.8990	<.0001
rank_P1000001	1	0.1816	0.0358	25.7564	<.0001
rank_H025A001	1	0.0279	0.0388	0.5180	0.4717

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.8	Somers' D	0.495
Percent Discordant	25.2	Gamma	0.496
Percent Tied	0.0	Tau-a	0.227
Pairs	1501992	c	0.748

Voľby do Snemovne R. 1986

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9798	0.1162	290.4782	<.0001
rank_HOUSSIperc1	1	0.1041	0.0208	25.0665	<.0001
rank_P0310014	1	0.4862	0.1188	16.7621	<.0001
rank_P0340033	1	0.5177	0.1124	21.2064	<.0001
rank_P0350033	1	0.4514	0.1107	16.6232	<.0001
rank_P1270007	1	0.1823	0.0919	3.9363	0.0473
rank_HOUINCave	1	0.0608	0.0197	9.5584	0.0020
rank_P1000001	1	0.1222	0.0335	13.3019	0.0003
rank_H025A001	1	0.0725	0.0366	3.9191	0.0477

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.3	Somers' D	0.407
Percent Discordant	29.6	Gamma	0.407
Percent Tied	0.0	Tau-a	0.198
Pairs	1593240	c	0.704

Voľby do Snemovne R. 1988

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9011	0.1153	271.7663	<.0001
rank_HOUSSIperc1	1	0.0999	0.0209	22.9558	<.0001
rank_P0310014	1	0.4443	0.1186	14.0341	0.0002
rank_P0340033	1	0.5275	0.1121	22.1509	<.0001
rank_P0350033	1	0.4954	0.1103	20.1752	<.0001
rank_P1270007	1	0.1621	0.0917	3.1246	0.0771
rank_HOUINCave	1	0.0329	0.0197	2.7967	0.0945
rank_P1000001	1	0.1604	0.0334	23.0301	<.0001
rank_H025A001	1	0.0677	0.0365	3.4353	0.0638

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.2	Somers' D	0.405
Percent Discordant	29.8	Gamma	0.405
Percent Tied	0.0	Tau-a	0.197
Pairs	1589185	c	0.702

Voľby do Snemovne R. 1990

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7311	0.1444	357.7172	<.0001
rank_HOUSSIperc1	1	0.0973	0.0237	16.8540	<.0001
rank_P0310014	1	0.7136	0.1269	31.6057	<.0001
rank_P0340033	1	0.4347	0.1249	12.1201	0.0005
rank_P0350033	1	0.5978	0.1233	23.4976	<.0001
rank_P1270007	1	0.3734	0.1030	13.1312	0.0003
rank_HOUINCave	1	0.0464	0.0230	4.0838	0.0433
rank_P1000001	1	0.2055	0.0382	28.9725	<.0001
rank_H025A001	1	0.0984	0.0416	5.5942	0.0180

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.0	Somers' D	0.481
Percent Discordant	26.0	Gamma	0.481
Percent Tied	0.0	Tau-a	0.215
Pairs	1168860	c	0.740

Prezidentské voľby 1984

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.4461	0.1874	563.0703	<.0001
rank_HOUSSIperc1	1	0.1305	0.0249	27.3653	<.0001
rank_P0310014	1	0.6953	0.1356	26.2949	<.0001
rank_P0340033	1	0.5095	0.1364	13.9527	0.0002
rank_P0350033	1	0.8791	0.1340	43.0436	<.0001
rank_P1270007	1	0.1328	0.1153	1.3269	0.2494
rank_HOUINCave	1	0.2130	0.0254	70.4225	<.0001
rank_P1000001	1	0.2343	0.0425	30.4050	<.0001
rank_H025A001	1	0.1365	0.0449	9.2389	0.0024

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	81.7	Somers' D	0.635
Percent Discordant	18.2	Gamma	0.635
Percent Tied	0.0	Tau-a	0.236
Pairs	1224222	c	0.818

Najdôležitejšia skutočnosť plynúca z tohto porovnania je, že model vyvinutý pre prezidentské voľby je o poznanie menej vhodný pre odhadovanie výsledkov volieb do Snemovne reprezentantov. Naopak, výsledky prezidentských volieb z roku 1984 sa mu podarilo odhadnúť takmer rovnako presne ako tých v roku 1988. Na základe porovnania si tiež môžeme všimnúť, ktoré premenné nie sú robustné vo svojej štatistickej významnosti na hladine významnosti 0,01, najčastejšie sa jedná o premenné „rank_P1270007“, „rank_HOUINCave“ a „rank_H025A001“.