Valgnatt – prognosemodell

versjon 1.02
JHEL/ 2003-05-15

Beregningskjerne

Det finnes en populasjon med i alt N enheter. Enhetene deles i H strata (merk at strata kan bli definert avhengig av utvalget). Når prognoser skal beregnes foreligger det et utvalg S av enheter. Data faller i to kategorier:

$$\mathbf{Y}_i, i \in S$$

som er innrapporterte stemmetall for alle partier/partigrupper,

$$\mathbf{X}_{1i},\ldots,\mathbf{X}_{Ji}\ ,\ i = 1,\ldots,N$$

som er stemmetall fra forhåndsstemmer og tidligere valg: variable som er kjent for alle enheter i populasjonen. $\mathbf{Y}_i, \mathbf{X}_{ji}$ er vektorer med dimensjon K, lik antall partier og partigrupper.

Mao:

X: kjente data (forh.stemmer, tidligere valg). Det er J slike vektorer pr krets
Y: ny-innrapporterte stemmetall
i: indeks for kretser. Det er N i alt
j: identifiserer forh.stemmer/tidligere valg, det er J muligheter i alt
k: indeks for parti, i alt er det K (dimensjonen til X-ene,Y)

Stemmeandelene som svarer til hver komponent beregnes for hver i. Dvs

$$p_i^k = \frac{Y_i^k}{\sum_k Y_i^k}, \quad q_{ji}^k = \frac{X_{ji}^k}{\sum_k X_{ji}^k}, k = 1,\ldots,K,$$

samt tilsvarende andeler for hele populasjonen, for de variablene som er kjent der:

$$q_j^k = \frac{\sum_{i=1}^{N} X_{ji}^k}{\sum_{i=1}^{N}\sum_k X_{ji}^k}.$$

I samplet beregnes stemmeandelene

$$p_S^k = \frac{\sum_{i\in S} Y_i^k}{\sum_{i\in S}\sum_k Y_i^k}$$

og

$$q_{jS}^{k} = \frac{\sum_{i \in S} X_{ji}^{k}}{\sum_{i \in S} \sum_{k} X_{ji}^{k}}, \quad j = 1,\ldots,J.$$

Stemmeandelene for hver enhet utgjør vektorene

$$\mathbf{P}_i = \left(p_i^1,\ldots,p_i^K\right), \quad \mathbf{Q}_{ji} = \left(q_{ji}^1,\ldots,q_{ji}^K\right),\ j = 1,\ldots,J.$$

Beregningene faller i tre trinn:

- estimering av regresjonskoeffisienter
- prognostisering av andeler i populasjonen
- usikkerhetsberegning

Estimering av koeffisienter

Til stemmeandelene tilpasses en lineær regresjonsmodell

$$\mathbf{P}_i = \alpha\mathbf{1} + \beta_1\mathbf{Q}_{1i} + \cdots + \beta_J\mathbf{Q}_{Ji} + \mathbf{e}_i\ ,\ i \in S, \quad (*)$$

på følgende måte: alle vektorene fra enhetene i utvalget S blir stablet opp i "lange" vektorer:

$$\widetilde{\mathbf{P}}' = (\mathbf{P}_1',\ldots,\mathbf{P}_{n_S}')', \quad \widetilde{\mathbf{Q}}_j' = (\mathbf{Q}_{j1}',\ldots,\mathbf{Q}_{jn_S}')',\ j = 1,\ldots,J\ ,$$

der n_S betegner utvalgsstørrelsen. I tilpasningen skal det være mulig å velge ut hvilke komponenter som skal være med, og det skal være mulig å foreta en vekting. Deretter estimeres koeffisientene ved minste kvadraters metode på de lange vektorene vha QR-dekomposisjon. Det må tas behørig hensyn til numerisk stabilitet i denne beregningen, bl a ved dobbel presisjon.

Beregningen skjer slik:
Alle ligningene i (*) stables som linjeblokker i et stort m-kv. problem:

$$\mathbf{U} = \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}\ ,$$

med i alt $n = n_S K$ linjer og **b** har dimensjon $p = J + 1$.

Ved Householder-transformasjoner bestemmes Householder QR-dekomposisjonen $\mathbf{Z} = \mathbf{QR}$, der Q er ortonormal, R øvretriangulær (obs at denne Q ikke må forveksles med de andre Q-ene i resten av notatet). Q, R er blokkdelte slik:

$$\mathbf{Q} = \left(\mathbf{Q_0}\ \mathbf{Q_1}\right), \mathbf{R} = \begin{pmatrix} \mathbf{R_0} \\ \mathbf{0} \end{pmatrix},$$

med R_0 øvretriangulær av dimensjon $p \times p$. Q beregnes ikke eksplisitt.
Da fremkommer løsningen slik:

1) beregn

$$\mathbf{Q}'\mathbf{U} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix}$$

2) løs likningssystemet

$$\mathbf{R}_0\mathbf{b} = \mathbf{c}$$

3) finn kovariansmatrisen mm til estimatene ved

$$\tilde{\sigma}^2 = \frac{1}{n-p}\|\mathbf{d}\|^2, \quad \mathbf{K} = \mathbf{R}_0'\mathbf{R}_0.$$

De fleste implementeringer gjør 1) og 2) i ett metodekall.

Det finnes mye god og gratis Fortran, C og C++-kode. Mulige kilder til Javakode er:

- Colt: http://hoschek.home.cern.ch/hoschek/colt/V1.0.3/doc/index.html Ser seriøst ut!
- JAMA http://math.nist.gov/javanumerics/jama/
- JLAPACK http://www.cs.utk.edu/f2j/
- Java LINPACK http://www1.fpl.fs.fed.us/linear_algebra.html
- kommersiell prog.vare: http://www.vni.com/products/imsl/jmsl.html Selges i Norge av Telemarksgruppen (samme som har S+)

Etter minste kvadraters metode foreligger en estimert koeffisientvektor $\tilde{\boldsymbol{\beta}} = (\tilde{\alpha}, \tilde{\beta}_1, \ldots)$ med estimert kovariansmatrise $\tilde{\sigma}^2\mathbf{K}^{-1}$. Estimatet $\tilde{\boldsymbol{\beta}}$ trekkes inn mot en apriorimiddelverdi $\boldsymbol{\beta}_0$ iht formelen

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}_0 + \left(\mathbf{I} + \frac{\tilde{\sigma}^2}{\tau_0^2}\mathbf{K}^{-1}\right)^{-1}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$

der τ_0^2 er apriorivariansen til betaene.

I første versjon kan minste kvadraters metode utgå, og aprioriverdiene benyttes slik de er.

Beregning av prognoser for hele populasjonen

Uten stratifisering beregnes prognoser på følgende måte:

Det beregnes en prediksjon for stemmeandelene for utvalget iht den estimerte modellen:

$$p_S^{*k} = \beta_1^* q_{1S}^k + \cdots + \beta_J^* q_{JS}^k,$$

dvs uten konstantleddet.

Differansen mellom prediksjon og observert andel i utvalget beregnes og blåses opp til en prognose for hele populasjonen:

$$\hat{p}^k = p_S^k - p_S^{*k} + \beta_1^* q_1^k + \cdots + \beta_J^* q_J^k. \quad (**)$$

Disse estimatene kan ligge utenfor intervallet (0,1). Endelige prognoser fremkommer ved å trunkere til [0,1].

Stratifisering

Estimering av koeffisienter gjøres på bakgrunn av hele datamaterialet under ett, dvs at det samme koeffisientsettet brukes innen alle strata. For hvert stratum h beregnes så en prognose $\hat{p}_h^k$ for prosentandeler innen stratumet iht (**) ovenfor (stratum brukes som populasjon). Deretter veies de stratumvise estimatene sammen til et estimat for hele populasjonen:

$$\hat{p}^k = \sum_h v_h \hat{p}_h^k .$$

Stratumvektene $v_1, \ldots, v_H$ er gitt ved

$$v_h = \frac{M_h N_h / m_h}{\sum_h M_h N_h / m_h},$$

der m_h er totalt antall stemmeberettigede i utvalget som også faller i stratum h, M_h er totalt antall stemmeberettigede i stratum h, N_h er totalt antall stemmer i utvalget fra stratum h.

Prognose for enkelte enheter

Proporsjonen p_i^k for en enhet i som ikke er med i utvalget kan prognostiseres ved

$$\hat{p}_i^k = p_S^k - p_S^{*k} + \beta_1^* q_{1i}^k + \cdots + \beta_J^* q_{Ji}^k .$$

(Dette forutsetter at samplet har en viss størrelse.) Alle variable må være kjent for den enheten som skal prognostiseres, også forhåndsstemmer dersom disse inngår i modellen.

Korreksjon for ulike datagrunnlag

I praksis vil man måtte ta høyde for at kretsene faller i to grupper ved prognosetidspunktet:

A) ingen forhåndsstemmer foreligger
B) forhåndsstemmer foreligger

For gruppe A estimeres ingen enkeltresultater (evt byresultater).

Andre prognoser kan beregnes ved å la gruppe A utgjøre et eget stratum, og bruke en annen modell, bare basert på tidligere valg, innen dette stratumet. Det er også mulig å lage en modell

basert på splitting av alle kategorier in to, med og uten forhåndsstemmer. Hver delkategori estimeres separat og summeres i etterhånd.

Dersom A er moderat, dvs under ca 75% av populasjonen, er det enklest å gå fram på følgende vis:

1. for alle kretser i gruppe A settes alle absolutte og relative forhåndsstemmetall til 0 (dvs at q_{ji}^k settes til 0 før minste kvadraters metode)
2. strata slås evt sammen slik at utvalget S inneholder minimum 2 komplette kretser (gruppe B) i hvert stratum
3. prognoser beregnes akkurat som om alle data var komplette

Dersom utvalget inneholder for få komplette kretser må en alternativ modell brukes.

Beregning av usikkerhet

Denne tar utgangspunkt i den estimerte residualvariansen $\tilde{\sigma}^2$ fra regresjonsanalysen. La

$$w_i = \frac{\sum_k Y_i^k}{\sum_{i,k} Y_i^k}$$

betegne andelen stemmer (i populasjonen) fra enhet i . Disse størrelsene kan estimeres ved f eks

$$\hat{w}_i = \frac{M_i}{\sum_{i,k} M_i} ,$$

der M_i er antall stemmeberettigede i kommunen/kretsen. Sett $\hat{w}_S = \sum_{i \in S} \hat{w}_i$. For en estimert andel (i hele populasjonen) $\hat{p}^k$ estimeres variansen til å være

$$\text{estvar}\left(\hat{p}^k\right) = K\tilde{\sigma}^2 \frac{\sum_{i \in S} \hat{w}_i^2}{\hat{w}_S^2}\left(1 - \hat{w}_S\right) ,$$

der K er en empirisk justeringsfaktor som skal ta høyde for at utvalget ikke er tilfeldig.

Basert på historiske valg ser det ut til at antallet regioner bør bestemme korreksjonsfaktoren. Rimelige verdier er:

1. for fylkes- og byprognoser: $K = 2$
2. for landsprognoser med 3 eller færre strata: $K = 3$
3. for landsprognoser med 4 strata (dvs sentrale kommuner): $K = 1$

Med disse verdiene tyder de historiske data på at variansen sjelden blir sterkt underestimert.

Som prediksjonsintervall brukes et visst antall standardavvik:

$$\left[\hat{p}^k - c\sqrt{est\,\mathrm{var}\left(\hat{p}^k\right)},\ \hat{p}^k + c\sqrt{est\,\mathrm{var}\left(\hat{p}^k\right)}\right] .$$

For stratifisert utvalg brukes for enkelhets skyld samme formel.

Marginer ved mandatfordeling

For å prognostisere marginer ved mandatfordeling, dvs antall stemmer i det endelige resultatet som skal til for å vinne eller miste et mandat, brukes den prognostiserte prosentfordelingen multiplisert med et grovestimat for det endelige antallet avgitte stemmer. Grovestimatet fremkommer som antall stemmeberettigete totalt, multiplisert med fremmøteprosenten i utvalget.

Bruk av metoden

Organisering av beregningene

Det må være funksjonalitet i implementeringen av modellen for fleksibelt å kunne spesifisere

- hvilke partier som skal med
- hvilke X-variable som skal med

Typisk vil det være et forholdsvis lite antall partier: A, H, FRP, KRF, SP, SV, V som blir brukt i estimeringen av koeffisientene. De samme koeffisientene blir deretter brukt på alle småpartiene.

Ved utprøvingen viser det seg at forhåndsstemmer samt resultatene fra to eller tre foregående valg egner seg best som X-variable. Metoden bør videre kjøres med data på kretsnivå.

Ved estimering for flere domener (dvs fylker, hele landet, byer) som det ønskes prognose for, skjer beregningen i to trinn:
1) På grunnlag av alle foreliggende data (dvs fra hele landet) estimeres koeffisienter
2) For hvert domene (der det foreligger valgtingsstemmer) beregnes en prognose iflg (**), dvs at hvert domene etter tur betraktes som populasjonen

Ved stortingsvalg må metoden kjøres for hvert fylke, for Oslo og for landet som helhet. Mandatfordeling av hhv fylkes- og utjevningsmandater må skje på grunnlag av prosentfordelinger, ikke absolutte stemmetall.

Det forutsettes også at resultatene fra de tidligere valg er kjent på kretsnivå og kodet korrekt mht gjeldende partinavn, kommunenummer og kretsnummer. Ved evt endringer må historiske data regnes om, imputeres osv.

Hvordan strata bestemmes

Strata bestemmes ut fra utvalget S. Utgangspunktet er regioner og sentralitet. Regioner er definert slik:

- region 1: fylkene 1,...,8
- region 2: fylkene 9,...,15
- region 3: fylkene 16,...,20

Som sentrale kretser regnes

1. kretser i de fire byene
2. kommuner med andel tettbygd minst 80% samt minst 5000 stemmeberettigede og fylke 1-16

Som strata brukes i utgangspunktet et stratum for sentrale kretser samt de tre regionene (unntatt sentrale kretser), dersom utvalget er stort nok til at hvert stratum inneholder minst to komplette kretser. Dersom så ikke er tilfelle, slås strata sammen i følgende rekkefølge, inntil kriteriet om to komplette kretser pr stratum er tilfredsstilt:

1. sentralt stratum fordeles på regionene
2. region 1 og 2 slås sammen og/eller region 2 og 3 slås sammen
3. alle regionene slås sammen

For estimering i fylker er det bare aktuelt med to strata: fylke og sentrale kretser (i fylket).

Når resultater kan publiseres

Det må være avgitt stemmer i utvalget (for det aktuelle domenet), også for forhåndsstemmer og tidligere valg.

Dersom dekningen av forhåndsstemmer er tilstrekkelig svak, brukes en modell basert utelukkende på tre foregående valg. Dette skjer dersom minst én av følgende betingelser inntreffer:

- færre enn 25% av kretsene i den aktuelle populasjonen har innrapportert forhåndsstemmer
- den forventede forhåndsstemmeandelen (relativt til stemmetallet for siste valg) er under 2% i mer enn 25% av kretsene (som har rapportert) i den aktuelle populasjonen. (Dette inkluderer kretser med eksplisitt innrapportert 0 forhåndsstemmer etter 1. opptelling.)
- utvalget har færre enn to komplette kretser (i domenet)

Det kreves alltid at $Kn_S \geq J + 1$. Som regel vil dette kravet være oppfylt, men det er tenkelig at prognoser ikke kan beregnes, selv for modeller uten forhåndsstemmer.

Dersom det er mulig å beregne en landsprognose, sjekkes om det estimerte standardavviket for stemmeandelene er under en fastsatt grense, med foreslått verdi 1,5 prosentenhet. I så fall publiseres prognosen.

For mindre populasjoner (byer og fylker) må usikkerhetskravet settes høyere, foreslått verdi er 2,5 prosentenheter.

Hvis imidlertid andelen innrapporterte kretser er tilstrekkelig stort, publiseres prognosen uansett estimert usikkerhet. Det foreslås at grensen settes til 20% for landsprognoser og 40% for by- og fylkesprognoser.

Det kan være naturlig å tillate publisering av alle øvrige prognoser (som det er mulig å beregne iht de andre kriteriene ovenfor) såfremt en landsprognose kan publiseres. Dette gjelder stortingsmandatfordeling og prognoser for individuelle kommuner(utenom bykommunene). For stortingsmandater brukes estimater for hver region (strata 1-3) dersom det aktuelle fylkesresultatet ikke er tilgig.

Andre forhold

De estimerte andelene vil ikke nødvendigvis summere seg til 1. Det forutsetter at alle $\hat{p}^k \in [0,1]$. Hvis andelen Andre er stor nok, kan den evt justeres for å gi sum 1, men vil da trolig være et dårligere estimat.

Aprioriverdier og varians for koeffisientene bestemmes skjønnsmessig på grunnlag av estimering på historiske data for hele populasjonen. Det foreligger egnede verdisett, men tuning vil forekomme.

Det må være mulig å lese ut estimerte koeffisienter osv fra metoden mens den kjører i driftsmiljøet.

Dersom alle forhåndsstemmer i en by legges i egne tekniske kretser, må disse stemmene innrapporteres som forhåndsstemmer. Både tekniske kretser og de tilhørende ordinære kretsene (som har 0 forhåndsstemmer) må behandles som ufullstendig rapporterte kretser.

Testing

Skjer ved sammenlikning med R-kode på utvalgte eksempler.

Vedlegg

Aprioriverdier

modell	**forh. stemmer**	**tidl. valg (regnet bakover)**			**standardavvik**
		1.	2.	3.	τ_0
med forh. stemmer	0,5	0,25	0,15	0,05	0,05
uten forh. stemmer	-	0,5	0,25	0,15	0,05

Sentrale kommuner

Iht 2002-inndeling av kommuner (i alt 53 kommuner). Bykommunene er inkludert:

```
  101  104  105  106  124  136  213  214  215  216  217  219  220  228
  230  231  233  235  301  403  501  602  604  625  627  701  702  704
  706  709  722  805  806  906 1001 1018 1102 1103 1106 1121 1122 1124
 1127 1149 1201 1221 1228 1502 1503 1504 1531 1601 1663
```