Methods for Small-Area-Estimation using Non-Representative Samples

Roberto Cerina

Nuffield College, University of Oxford

▶ The material for this workshop can be found in the following git-hub repository:
https://github.com/robertocerinaprojects **⧉**;

▶ you can contact me on
roberto.cerina@nuffield.ox.ac.uk for further information and material on the topic.

Introduction to Small-Area Estimation

To obtain reliable estimates for a quantity of interest at a desired small-area level;

  i. small-area examples include geographic/ political areas (e.g. municipalities; parliamentary constituencies; geographic regions etc.);

  ii. quantities of interest can be continuous measurements (e.g. height, weight, ideology scores.), including uncertainty measures, such as probabilities or odds (e.g. voting for a given party; contracting a disease etc.).
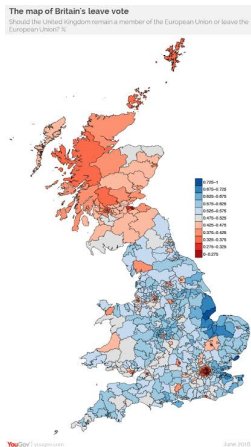
Figure: https://yougov.co.uk/topics/politics/
articles-reports/2016/06/21/yougov-referendum-model.

Nationally Representative Samples

Samples (without replacement) of the quantity of interest, where the probability of inclusion (poi) of each possible outcome is equal to the relative frequency of that outcome in the population.

▶ Typical pre-election survey has roughly $N = 1,000$ to $2,000$;

▶ the population of interest is the US voting population;

▶ the small-areas of interest are the 51 states (including the District of Columbia);

▶ X is a dichotomous random variable such that $X = 1$ if an individual is a registered Republican; $X = 0$ otherwise.

- We could assume each state-level sample is independent;
- Estimate 51 independent state-level means by taking the state-level averages.;
- Assume the observations are distributed as follows:

$$X_{si} \sim \text{Bern}(p_s)$$

- The maximum-likelihood estimator for π_s is the state-level mean:

$$\hat{p}_s = \frac{\sum_i^n x_{si}}{n_s};$$

- The variance for this estimator which meets the Cramer Rao Lower Bound corresponds to:

$$\text{Var}(\hat{p}_s) = \frac{\hat{p}_s(1 - \hat{p}_s)}{n_s}$$

UNIVERSITY OF
OXFORD

▶ Asymptotically, the distribution of the estimator can be characterized as follows:

$$\hat{p}_s \sim N(p_s, \text{Var}(\hat{p}_s))$$

▶ a 95% confidence interval over the population mean can then be obtained:

$$\hat{p}_s \pm 1.96\sqrt{\frac{\hat{p}_s(1 - \hat{p}_s)}{n_s}};$$

▶ $1.96\sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n_s}}$ is referred to as the *margin of error*.

Note that we can find the minimum sample size required to
meet a given level of confidence as follows:
For a given margin of error $m.o.e. <= m$, we can solve for n_s:

$$1.96\sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n_s}} \leq m; \tag{1}$$

$$n_s \geq \left(\frac{1.96}{m}\right)^2 \hat{p}_s(1-\hat{p}_s); \tag{2}$$

- Note that the sample size still depends on a value $\hat{p}_s$ which we don't know a-priori;
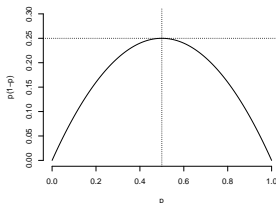- set $\hat{p} = 0.5$, since it corresponds to a situation of hypothetical maximum-uncertainty.

Figure: Relationship between magnitude of proportion and variance of estimator. A sample size fit to estimate a proportion $\hat{p} = 0.5$ will also be powered to estimate $\hat{p} = 0.25$, whilst the opposite is not true.

► Statistical Considerations: Acceptable m.o.e. for a proportion ranges from 0.01 to 0.03 in typical voting-intention surveys, corresponding minimum sample size is ranges just over 1000 units to just under 10,000, where the minimum sample size increases exponentially as the m.o.e. decreases;

► Practical Considerations: cost of telephone poll usually between 1 and 5 USD per respondent, and each state should be polled on multiple occasions over a campaign; each wave can cost up to 50,000 USD;

► maybe we can use prior information - for instance only focus on what we know to be the 10 to 15 swing-states - still prohibitive for most research outfits.

▶ Big News Network in the UK conducted a survey of its subscribers prior to the 2016 UK Brexit referendum;

▶ their users are not representative of the population at large;

▶ they are interested in estimating Brexit voting intentions at the constituency level;

▶ analyzing this data after the fact, we can test the performance of various techniques for area estimation.

| p_T_nextGE	nextGE_V	T_pastGE	V_pastGE	brexit_V	p_T_brexit	gender	age	region	ethnicity	edu	tenure	pcon
1.00	conservative	1	conservative	leave	10.00	1. Male	2. 25-34	4. East Midlands	1. White	2. Level 1	4. Rents (with or without housing benefit)	E14001025
1.00	ukip	1	ukip	leave	10.00	2. Female	3. 35-44	9. South West	1. White		2. Owns with a mortgage or loan	E14001031
0.50		0		leave	7.00	2. Female	2. 25-34	6. East of England	1. White	2. Level 1	2. Owns with a mortgage or loan	E14001000
	pc	1			8.00	2. Female	6. 65+	10. Wales	1. White		4. Rents (with or without housing benefit)	W07000045
	labour	1	labour	remain	10.00	2. Female	3. 35-44	7. London	2. Black	2. Level 1	4. Rents (with or without housing benefit)	E14000615
	labour	1	labour	remain	10.00	2. Female	3. 35-44	10. Wales	1. White	2. Level 1	2. Owns with a mortgage or loan	W07000071
1.00	conservative	1	libdem	remain	10.00	1. Male	2. 25-34	7. London	2. Black	5. Level 4	1. Owns outright	E14001060
1.00		1	conservative		10.00	1. Male	4. 45-54	7. London	2. Black	5. Level 4	4. Rents (with or without housing benefit)	E14000804
	pc	1	pc	remain	10.00	1. Male	6. 65+	10. Wales	1. White	3. Level 2	1. Owns outright	W07000041
1.00		1	libdem		10.00	2. Female	5. 55-64	11. Scotland	1. White	4. Level 3	1. Owns outright	S14000026
	labour	1	labour	leave	10.00	1. Male	6. 65+	11. Scotland	1. White	2. Level 1	1. Owns outright	S14000046
	snp	1	snp	remain	10.00	1. Male	4. 45-54	11. Scotland	1. White	2. Level 1	2. Owns with a mortgage or loan	S14000049
0.70	labour	1		leave	10.00	1. Male	3. 35-44	3. Yorkshire and the Humber	1. White	2. Level 1	4. Rents (with or without housing benefit)	E14000645
1.00	ukip	1	ukip	leave		1. Male	1. 18-24	1. North East	1. White	5. Level 4	2. Owns with a mortgage or loan	E14000832
	novote	0		novote		1. Male	5. 55-64	2. North West	1. White		1. Owns outright	E14000958
1.00	conservative	1	conservative	remain	10.00	2. Female	5. 55-64	8. South East	1. White	3. Level 2	1. Owns outright	E14000818
	labour	1	labour	remain	10.00	2. Female	4. 45-54	11. Scotland	1. White	4. Level 3	4. Rents (with or without housing benefit)	S14000059
0.90	conservative	1	conservative			2. Female		5. West Midlands	1. White	5. Level 4	1. Owns outright	E14000722
		1				2. Female		7. London				E14000555
1.00	labour	1	other	remain	10.00	1. Male	2. 25-34	2. North West		5. Level 4		E14001054
	novote	1	conservative	remain	10.00	1. Male	5. 55-64	7. London		5. Level 4		E14000770
1.00	conservative	1	conservative	leave	10.00	2. Female	6. 65+	8. South East	1. White	3. Level 2	1. Owns outright	E14000874
0.90		1	libdem	remain			3. 35-44			4. Other		E14000983
0.70	labour	0			8.00	2. Female	3. 35-44	10. Wales	1. White	2. Level 4	2. Owns with a mortgage or loan	W07000049
1.00	ukip	1	conservative			2. Female		8. South East	1. White	6. Other	1. Owns outright	E14000997
0.10		1	conservative	leave	1.00	1. Male	6. 65+	9. South East	1. White	6. Other	1. Owns outright	E14000879
1.00	snp	1	snp		10.00	2. Female	5. 55-64	11. Scotland	1. White	3. Level 2	2. Owns with a mortgage or loan	S14000022
1.00	labour	1	labour	remain	10.00	1. Male	4. 45-54	7. London	2. Black	5. Level 4	4. Rents (with or without housing benefit)	E14000721
	labour	1	labour	leave	10.00	2. Female	3. 35-44	10. Wales	1. White	2. Level 1	4. Rents (with or without housing benefit)	W07000077
1.00	conservative	1	conservative	leave	10.00	2. Female	6. 65+	7. London	1. White	5. Level 4	4. Rents (with or without housing benefit)	E14000673

UNIVERSITY OF
OXFORD
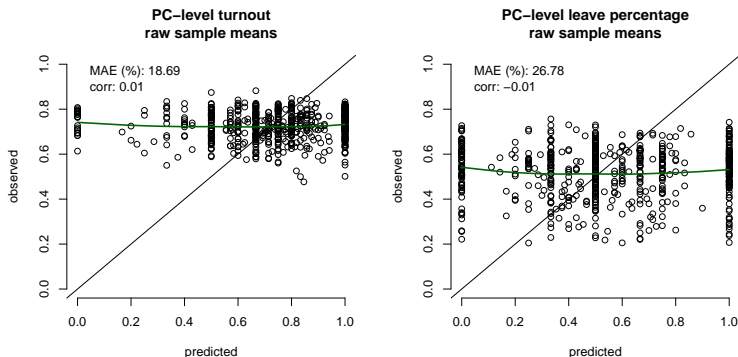
Figure: Predictive performance of raw sample means for each of the areas of interest.

- ▶ unacceptable error levels: elections are decided by few percentage points, a prediction which is over 20 points off on average is good to nobody;

- ▶ the turnout prediction looks somewhat better on average, but the low correlation suggests it does only so because of lower variance in the true turnout levels.

Regression + Post-Stratification

📄

- by assuming observations are fully independent, we are wasting information;
- a better assumption is 'conditional independence';
- within the constituencies, and across constituencies, voters will be similar according to a number of characteristics;

- ▶ Proposition: break areas of interest into mutually-exclusive sub-categories defined by socio-demographic characteristics;
- ▶ e.g. for each constituency, define the the following groups of interest: 'male-age> 65-graduate'; 'female-ag> 65-graduate'; etc.
- ▶ groups are mutually exclusive - summing over their shares in the population should give 1;
- ▶ we refer to these as 'cells' (you'll understand why);
- ▶ leverage correlations between the socio-demographic characteristics that define these cells, and response variable, to predict response probability for each category;
- ▶ aggregate these categories at the cell-level according to their population size;

UNIVERSITY OF
OXFORD

$$\hat{p}_c = \frac{\sum_g \hat{\pi}_{c,g} \times N_{c,g}}{\sum_g N_{c,g}}$$

▶ Challenges:
  i. need to gather reliable values for N_g;
  ii. need to produce estimates for $\hat{\pi}_{c,g}$

- An initial set of counts can be found in census cross-tabs;
- examples from the UK:
  https://www.nomisweb.co.uk/census/2011/lc5102ew

pcon_name	pcon	N	age	edu
Battersea	E14000549	26067	25-34	Level4andabove
Bermondsey and Old Southwark	E14000553	23261	25-34	Level4andabove
Hampstead and Kilburn	E14000727	22998	25-34	Level4andabove
Vauxhall	E14001008	22715	25-34	Level4andabove
Sheffield Central	E14000919	22659	16-24	Level3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Sheffield, Hallam	E14000922	73	25-34	Apprenticeship
Wirral South	E14001043	72	16-24	Other
Neath	W07000069	70	16-24	Other
Finchley and Golders Green	E14000703	67	16-24	Apprenticeship
Wirral West	E14001044	54	16-24	Other
Sefton Central	E14000916	54	16-24	Other

- There are highly precise (as they are based on the entire census) but severely limited, as they only offer 2 or 3 variables in depth;
- most applications will need a finer definition;
- this can be obtained by large representative surveys, or individual-level census sub-samples;
- both these sources are at the individual-level, and hence require us to do some of the work to turn individual responses into useable counts.

State	Zone	Gender	Age	Rurality	...	Caste
(19) West Bengal 19	(04) Eastern 04	(1) Male 1	(54-64]	(0) rural 0	...	(2) Forward/General (except Brahmin) 2
(19) West Bengal 19	(04) Eastern 04	(1) Male 1	(44-54]	(0) rural 0	...	(2) Forward/General (except Brahmin) 2
(33) Tamil Nadu 33	(06) Southern 06	(2) Female 2	(24-34]	(1) urban 1	...	(3) Other Backward Castes (OBC) 3
(10) Bihar 10	(02) North-Central 02	(1) Male 1	(17-24]	(0) rural 0	...	(3) Other Backward Castes (OBC) 3
(27) Maharashtra 27	(05) Western 05	(2) Female 2	(34-44]	(1) urban 1	...	(2) Forward/General (except Brahmin) 2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
(08) Rajasthan 08	(01) North 01	(1) Male 1	(17-24]	(1) urban 1	...	(5) Scheduled Tribes (ST) 5
(32) Kerala 32	(06) Southern 06	(2) Female 2	(34-44]	(0) rural 0	...	(2) Forward/General (except Brahmin) 2
(10) Bihar 10	(02) North-Central 02	(1) Male 1	(34-44]	(0) rural 0	...	(1) Brahmin 1
(24) Gujarat 24	(05) Western 05	(2) Female 2	(24-34]	(1) urban 1	...	(5) Scheduled Tribes (ST) 5
(32) Kerala 32	(06) Southern 06	(1) Male 1	(34-44]	(0) rural 0	...	(3) Other Backward Castes (OBC) 3

Table: A random sample of individuals from the *India Human Development Survey*[DVN31] (IHDS) a nationally representative survey of 135,986 voting-age individuals from 42,152 households, conducted between November 2011 and October 2012.

UNIVERSITY OF
OXFORD

i. restricted in size due to privacy concerns - may be good for 'shallow' cells, but not for 'deep' ones (by shallow we mean say the distribution of 'Ethnicity' or 'Ethnicity by Age' in the population; a deep cell would be 'Ethnicity by Age by Education by Constituency';

ii. might not be at the level desired - i.e. in UK, individual-level sample comes at the 'Local Authority (LA)' level - whilst we are usually interested in 'Parliamentary Constituency' level. This adds a matching step, which usually adds uncertainty to the frame;

iii. more generally, stratification frames can be out-dated, as census and other large samples are seldom conducted, and even more rarely made publicly available.

Crossing Boundaries: UK Example

WD16CD	WD16NM	PCON16CD	PCON16NM	LAD16CD	LAD16NM	FID
E05010888	Bishopston and Ashley Down	E14000602	Bristol West	E06000023	Bristol, City of	2001
E05010889	Bishopsworth	E14000601	Bristol South	E06000023	Bristol, City of	2002
E05009474	Newton Farm	E14000743	Hereford and South Herefordshire	E06000019	Herefordshire, County of	2003
E05010890	Brislington East	E14000599	Bristol East	E06000023	Bristol, City of	2004
E05010891	Brislington West	E14000601	Bristol South	E06000023	Bristol, City of	2005
E05008820	Stenson	E14000935	South Derbyshire	E07000039	South Derbyshire	4995
E05003153	Parkside	E14000543	Barrow and Furness	E07000027	Barrow-in-Furness	4996
E05003366	Wirksworth	E14000664	Derbyshire Dales	E07000035	Derbyshire Dales	4997
E05003513	Castle	E14000996	Tiverton and Honiton	E07000042	Mid Devon	4998
E05010638	Limestone Peak	E14000748	High Peak	E07000037	High Peak	4999
E05003204	Seascale	E14000647	Copeland	E07000029	Copeland	5000

Table: Example of boundary types from the UK 2016 map. Stratification frame data will come from one of these and our challenge is to match to the level of interest. Notice that this is a many-to-many match: multiple LAs will lie within PCs, and multiple PCs will line in multiple LAs - uncertainty is inevitable.

UNIVERSITY OF OXFORD

► Could estimate these by calculating the grup level means for each group g in constituency c independently;

► Problems:

   i. we don't have representatives for every group in our sample, so we can't produce $\hat{\pi}_{c,g}$ for all the groups we need;

  ii. many groups in our samples will have extremely low-n;

 iii. simple group-level means will tend to over-fit the data;

 iv. groups are NOT independent! Though they might be conditionally independent...

Regression Models for Cell-Response Estimation

- ▶ Can instead use regression to produce conditional means;
- ▶ usually interested in probability for categorical responses (in Brexit ref. we have choices $j = \{\text{Leave}, \text{Remain}, \text{Stay Home}\}$);
- ▶ will need categorical regression and link function - (multinomial) logit usually the most appropriate choice;

$$\boldsymbol{y}_i \sim \text{Multinomial}(\pi_{i,j}, ..., \pi_{i,J}, n = 1)$$

$$\pi_{i,j} = \frac{\exp(\omega_{i,j})}{\sum_j \exp(\omega_{i,j})}$$

$$\omega_{i,j} = \log(\mu_{i,j})$$

$$\mu_{i,j} = \sum_k x_{i,k}\beta_{k,j}$$

- model is not identifiable - because probabilities must sum to 1, we only have $J - 1$ separate parameters to estimate;
- without this restriction, there are many possible solutions for coefficients of interest;
- choose category $j = J$ for $j = 1, ..., J$ as baseline, and set coefficients to zero, and estimate $\beta_{k,j}$ for $k = 1, ..., K$ covariates and $j = 1, ..., (J-1)$ choices;
- $\beta_{k,J} = 0 \ \forall \ k$;
- interpretation of coefficients will be relative to this baseline category;

- ▶ note that this is akin to specifying $J - 1$ independent binomial logit regressions - Independence of Irrelevant Alternatives (IIA) assumption.

- ▶ further note the interpretation of multinomial logit coefficients is not straight forward: $\exp(\beta_{k,j})$ represents the amount by which to multiply the relative risk $\frac{\Pr(y_i = j)}{\Pr(y_i = J)}$ when variable x_k increases by 1 unit;

- ▶ for the purposes of our excercise however, the value of the coefficients is not of interest - interpretation is only useful to flag potential errors in coding/dataset;

- ▶ what we care about is the cell-level predictions.

OXFORD

- ▶ for each individual $i = 1, .., N$ in our survey we have access to K characteristics, e.g. Area, Age, Ethnicity, Education Level etc.;

- ▶ these are categorical, individual level characteristics;

- ▶ define a groups $g = 1, ..., G$ to uniquely identify every mutually-exclusive and exhaustive combination of the K characteristics;

- ▶ so if $K = 3$ (Area, Age, Education), then $g = 1$ idicates the voting groups England - 18 to 36 - Low-Education; $g = 2$ stands for England - 18 to 36 - High-Education, etc.;

- ▶ for each of these groups we want to find π_g^*;

UNIVERSITY OF
OXFORD

► note that group *g* will represent a unique row in the stratification frame.

pcon_name	pcon	N	age	edu
Battersea	E14000549	26067	25-34	Level4andabove
Bermondsey and Old Southwark	E14000553	23261	25-34	Level4andabove
Hampstead and Kilburn	E14000727	22998	25-34	Level4andabove
Vauxhall	E14001008	22715	25-34	Level4andabove
Sheffield Central	E14000919	22659	16-24	Level3

▶ Assume a simple model of the following form, for $j = 1, ..., (J-1)$, where the notation $g[i]$ indicates that each observation in our survey can be assigned to a given group g:

$$\mu_{g[i],j} = \beta_j + \sum_a \alpha_{a,j} x_{a,i}^{age} + \sum_e \eta_{e,j} x_{e,i}^{edu} + \sum_r \rho_{r,j} x_{r,i}^{reg} + \sum_c \gamma_{c,j} x_{c,i}^{area}$$

▶ all covariates are dummy variables corresponding to the respective age, education and area categories;

▶ after estimating the regression coefficients, *trained* on the survey data, we can easily fit the model to the stratification frame, and obtain a prediction for each g in the frame.

UNIVERSITY OF
OXFORD

- ▶ point-estimates for many interesting versions of this model (high dimensions, non-standard parameter specification, etc.) cannot be analytically calculated;
- ▶ a flexible and reliable way to estimate multinomial regression coefficients and their uncertainties is to do so via Bayesian computational methods, such as the Gibbs sampler[GSC+13];

▶ software such as JAGS[P$^+$03] or STAN[CGH$^+$17] can be used to flexibly specify and estimate complex Bayesian model;

▶ recall Bayes theorem: $\Pr(Z|Y) = \frac{\Pr(Y|Z) \times \Pr(Z)}{\Pr(Y)}$;

▶ the typical structure of a Bayesian model for a parameter Z requires a prior distribution $\Pr(Z)$, a likelihood for data that the parameter is assumed to have generated - and hence can tell us about its plausible values, $\Pr(Y|Z)$;

▶ the constant of proportionality $\Pr(Y)$ is not explicitly calculated, due to computationally challenging integration, and is rather approximated via computational methods;

▶ the result of Bayesian modeling is the posterior distribution of a parameter, $\Pr(Z|Y)$, i.e. the suitable space of values occupied by the parameter in the parameter space, given it gave rise to observations Y;

UNIVERSITY OF
OXFORD

- ▶ Bayesian methods require computationally intensive - and often intractable - integration to estimate posterior distributions and predictive distributions;

- ▶ Monte Carlo Markov Chain (MCMC) algorithms, such as the Gibbs Sampler[CG92], Metropolis-Hastings[CG95] or Hamiltonian Monte Carlo[Bet17] can help us find posteriors without the need for integration;

- one potential problem which is often encountered is speed - these models tend to be slow to *converge*, especially when the number of categories increase;

- convergence for practical prupuses is measured as the value of the Gelman-Rubin statistic $\hat{R}$, which we want to be ideally 1, but values below 1.1 are considered good enough to make reliable inference;

- potential issues that lead to slow convergence include multicollinearity; sheer size of the data; poor prior specification; poor initial value choice; inefficient sampling etc.

- in many practical applications, INLA[RMC09], a software based on variational-inference methods, can be faster, though sometimes less precise.

- ▶ we begin with a prior specification comparable to a fixed-effects model in a non-Bayesian regression setting;
- ▶ note that we use the JAGS parametrization for the normal distribution, such that $x \sim N(\mu, \tau)$, where μ represents the mean parameter, and $\tau = \frac{1}{\sigma^2}$ is the precision parameter - the inverse of the variance:

$$\beta_j \sim N(0, 0.01); \qquad (3)$$
$$\alpha_{a,j} \sim N(0, 0.01); \qquad (4)$$
$$\eta_{e,j} \sim N(0, 0.01); \qquad (5)$$
$$\rho_{r,j} \sim N(0, 0.01); \qquad (6)$$
$$\gamma_{c,j} \sim N(0, 0.01) \qquad (7)$$

UNIVERSITY OF OXFORD

► the specification is non-informative, as 0.01 corresponds to a standard deviation of 10, which is quite large on the logit scale - hence this prior is akin to a uniform distribution;

► the parameters are completely independent a-priori - there is no *borrowing of strength.*

OXFORD

```
for(i in 1:N){
    y[i] ~ dcat(pi[i,1:N.j])
    for(j in 1:N.j){
    pi[i,j] <- omega[i,j]/sum(omega[i,1:N.j])
    omega[i,j] <- exp(mu[i,j])
    mu[i,j] <- beta[j] +
               alpha[age_id[i],j] +
               eta[edu_id[i],j] +
               rho[region_id[i],j] +
               gamma[pcon_id[i],j]
} }
}
```

```
beta[1] <- 0
for(j in 2:N.j){
beta[j] ~ dnorm(0,0.1)
}

for(a in 1:N.age ){
alpha[a,1] <- 0
for(j in 2:N.j){
alpha[a,j] ~ dnorm(0,0.1)
} }

for(r in 1:N.region ){
rho[r,1] <- 0
for(j in 2:N.j){
rho[r,j]~ dnorm(0,0.1)
} }

for(e in 1:N.edu ){
eta[e,1] <- 0
for(j in 2:N.j){
eta[e,j]~ dnorm(0,0.1)
} }

for(c in 1:N.pcon){
gamma[c,1] <- 0
for(j in 2:N.j){
gamma[c,j]~ dnorm(0,0.1)
} }
```

► the multinomial model can be difficult to converge due to sampling issues;

► it can be shown that the likelihood function of the Multinomial model is equivalent to that of a Poisson model with an individual-level intercept (with a flat prior) to enforce a sum-to-1 in a stochastic way (see [**?**] for details on the JAGS implementation; [LGR17] for details on the likelihood equivalence);

► there are two practical changes we need to make the transformation: i) turn the categorical vector $y_i = j \ \forall \ j = 1, ..., J$ into a dummy-matrix $y_{i,j} = 0, 1$; ii) include individual level intercept λ_i in the model with a flat prior.

$$\boldsymbol{y}_{i,j} \sim \text{Poisson}(\omega_{i,j})$$
$$\omega_{i,j} = \log(\mu_{i,j})$$
$$\mu_{i,j} = \lambda_i + \sum_k x_{i,k}\beta_{k,j}$$
$$\lambda_i \sim N(0, 0.01)$$

▶ note: you still need the softmax link function to obtain probability-level predictions:

$$\pi_{i,j} = \frac{\exp(\omega_{i,j})}{\sum_j \exp(\omega_{i,j})}$$

```
    pi[i,j] <- omega[i,j]/sum(omega[i,1:N.j])

brexit[i,j] ~ dpois(omega[i,j])
omega[i,j] <- exp(mu[i,j])
   mu[i,j] <- lambda[i] +
              beta[j] +
              alpha[age_id[i],j] +
              eta[edu_id[i],j] +
              rho[region_id[i],j] +
              gamma[pcon_id[i],j]
```

UNIVERSITY OF
OXFORD

▶ we stratify the predicted proportion of individuals who say they would turn out by area (1- the proportion of 'stayed home') as follows:

$$\hat{p}_c^T = \frac{\sum_g \hat{\pi}_{c,g}^T \times N_{c,g}}{\sum_g N_{c,g}}$$

▶ we stratify the predicted proportion of individuals who say they would vote leave as follows:

$$\hat{p}_c^L = \frac{\sum_g \hat{\pi}_{c,g}^L \times (N_{c,g} \times \hat{\pi}_{c,g}^T)}{\sum_g (N_{c,g} \times \hat{\pi}_{c,g}^T)}$$
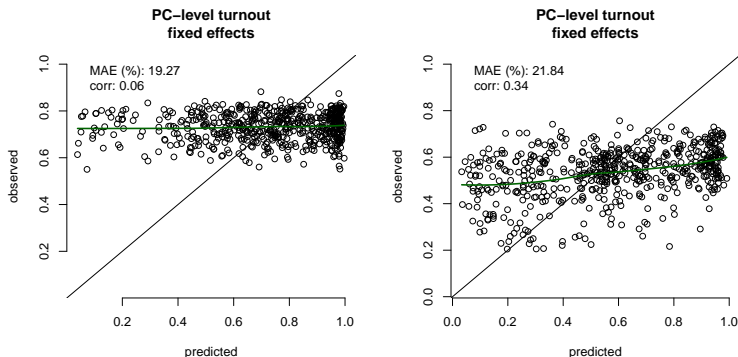
Figure: Predictive performance of fixed-effects multinomial-poisson model, after stratification.

- minor improvements, still very poor predictive power;
- it's clear from the above that the model is over-fitting: too many predictions stuck at extreme values (close to zero or 1) - we need to apply some 'regularization';
- we still cannot really predict the constituencies for which we have no data in the survey - these are imputed according to the prior in the fixed effects model, but this is non-informative, so there is no real information there;

- ▶ we want to obtain 'regularized' coefficients to curb over-fitting;
- ▶ this involves estimating coefficients by reducing the weight of extreme or low-evidence (low-n) observations;
- ▶ one common and intuitive way of doing this is via *multilevel* modeling;

▶ the typical multilevel model will be some variation of the following form:

$$y_j \sim N(\alpha_j, \sigma_y^2) \tag{8}$$
$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \tag{9}$$

UNIVERSITY OF
OXFORD

- the multilevel estimator for the group-level mean α_j (where a group here stands for, say, an age-group, or an education group) can be described as follows:

$$\hat{\alpha}_j = \lambda\mu_a + (1 - \lambda)y_j, \tag{10}$$

- where $\bar{y}_j$ represents the response for group-level j, and μ_a is the global mean across all groups;
- $\lambda = 1 - \frac{\sigma_a^2}{\sigma_a^2 + \sigma_y^2}$ is the 'pooling factor', defined by the between-groups variance and the observation level variance, deciding the degree of shrinkage toward the global mean;

UNIVERSITY OF
OXFORD

- as the local mean is pulled toward the global mean, we obtain a phenomena called 'shrinkage';
- shrinkage refers to the impact that the shared variance prior has on the regression coefficients; namely, it will pull these toward the 'global mean' for the set of groups which share the variance parameter;
- shrunk coefficients are subject to some degree of L2 regularization;
- the prior on the group-level variance parameter plays a key role in defining the amount of shrinkage: too high and it will shrink the coefficients entirely toward the global mean; too small and it will have no effect;
- we assign a non-informative prior and allow for whatever amount of shrinkage is best supported by the data.

▶ no reason to fit a random effect on the intercept in our application - shrinkage does not have significant impacts for groups lower or equal to 3;

▶ standard deviations for each set of groups given independent, non-informative priors:

$$\beta_j \sim N(0, 0.01); \tag{11}$$

$$\alpha_{a,j} \sim N(0, \tau_{l=1}); \tag{12}$$

$$\eta_{a,j} \sim N(0, \tau_{l=2}); \tag{13}$$

$$\gamma_{a,j} \sim N(0, \tau_{l=3}); \tag{14}$$

$$\tau_l = \frac{1}{\sigma_l^2} \tag{15}$$

$$\sigma_l \sim \text{Uniform}(0, 5) \tag{16}$$

UNIVERSITY OF
OXFORD

```
beta_star[1] <- 0
for(j in 2:N.j){
beta_star[j] ~ dnorm(0,0.01)
}

for(i in 1:N){
lambda[i] ~ dnorm(0,0.01)
}

for(a in 1:N.age ){
alpha[a,1] <- 0
for(j in 2:N.j){
alpha[a,j] ~ dnorm(0,tau[1])
}
for(j in 1:N.j){
alpha_star[a,j] <- alpha[a,j]*aux[1]
} }
```

▶ note that each random effect is specified with an auxiliary nuisance paramter, such that:

$$\alpha^* = \alpha \times \text{aux}_1 \qquad (17)$$

$$\alpha \sim N(0, \tau_{l=1}) \qquad (18)$$

$$\text{aux}_1 \sim N(0, 0.01) \qquad (19)$$

▶ it has been shown empirically[GVDHB08] that the auxiliary parameter improves the convergence capacity of the model by: a) increasing randomness and hence reducing dependency on past values in the markov chain; b) triggering a 'jumping' capacity in the posterior distribution of α^* which breaks the vicious circle leading to estimates of $\sigma_{l=1}$ getting stuck near zero.

UNIVERSITY OF
OXFORD

```
for(l in 1:4){
tau[l] <- pow(sigma[l],-2)
sigma[l] ~ dunif(0,5)
}

for(aux_id in 1:4){
aux[aux_id] ~ dnorm(0,0.01)
}
```

UNIVERSITY OF
OXFORD
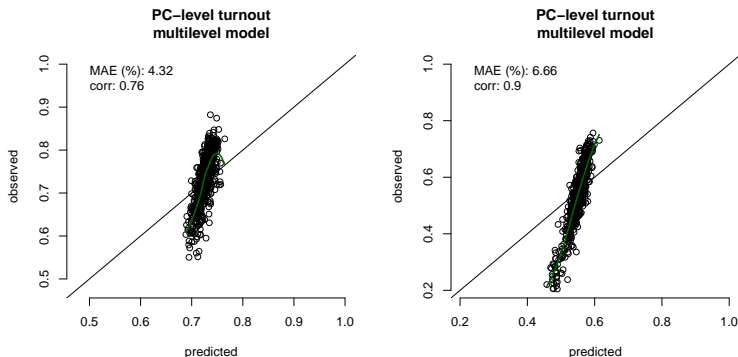
Figure: Predictive performance of multilevel, multinomial-poisson model, after stratification.
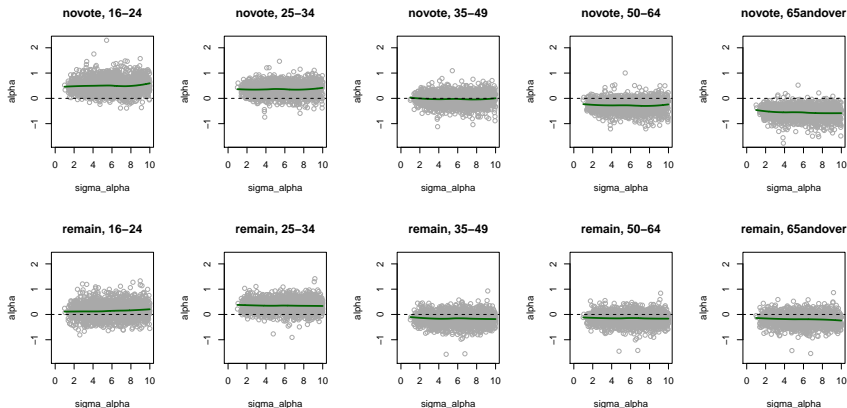
Figure: Shrinkage effect of the variance parameter on choice-age random effects; no support for σ close to zero.

- the correlation is very high now - meaning we have the order of the observations roughly right;
- MAE is also much lower, but still a bit too high to be considered reliable;
- we seem to be under-estimating the variance around the mean prediction, a direct result of implemented shrinkage;
- need more information on constituencies to lower the degree of pooling - shrinkage may be doing too much here;

$$\mu_{g[i],j} = \beta_j + \sum_a \alpha_{a,j} x_{a,i}^{age} + \sum_e \eta_{e,j} x_{e,i}^{edu} + \sum_r \rho_{r,j} x_{r,i}^{reg} + \sum_c \gamma_{c,j} x_{c,i}^{area}$$
$$+ \sum_j \phi_{k,j} x_{c[i],k}^{area_p red}$$

- ▶ the areal predictor may help reduce the bias brought in by the random effects estimates;
- ▶ it also captures direct contextual effects i.e. how much are you more or less likely to vote leave given you live in a student-constituency? probably alot less than average, independently of which constituency;
- ▶ note that there is a high risk of over-fitting: studies have shown [LP09, LP13] including more predictors than necessary lowers estimates' accuracy.

- another question is which variables to include: in absence of application-specific rationale, you might opt for a variable-selection procedure such as Lasso or Spike-and-Slab;
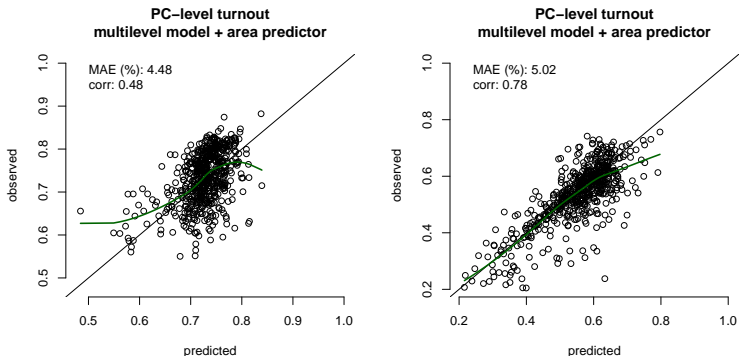
Figure: Predictive performance of multilevel model augmented via the areal predictor.

Discussion

- ▶ you should now know how to obtain reliable area-level estimates from non-representative data;
- ▶ you should know how to estimate multinomial choice models using computational Bayesian methods, along with the Poisson transform;
- ▶ you should understand the importance of regularization/shrinkage, as applied by multilevel modeling, for predictive performance;
- ▶ you should know how to use JAGS to fit a variety of models, manipulate priors and speed up MCMC convergence;
- ▶ you should understand the importance of the linear predictor at the area-level.

UNIVERSITY OF
OXFORD

► i. how to choose elements of the linear predictor/ which random effects, under large surveys and extensive stratification frames? Can implement large-scale regularization (Ridge, Lasso, Spike and Slab - see[GKRT18]) and be agnostic about model structure;

► ii. why limit the linear predictor to the area-level? Some evidence it does not make a difference in US public opinion[LP09], but could be important in other applications!

- ▶ iii. we looked at a very limited stratification frame; this would need to be 'expanded' if possible; to expand frames we need training data to be fit to them;
- ▶ so for in stance if we wanted to include 2015 voting intention as an Individual-level predictor, we would have to ensure we could stratify by it, and hence we would have to know the joint counts 'age-edu-pcon-2015vote';
- ▶ these are not available, so we need to impute them;
- ▶ this methodology is currently subject of (my) research (ask me about it if interested).

In-Class Practice

📄

▶ We will now go through the R code on the Github, make sure you all have the relevant packages and understand every step;

▶ Your Task: fit a multilevel model which includes interactions between age and education on JAGS (how-to will be explained during the tutorial):

    i. Produce performance plot;

   ii. Produce a variance-parameter-shrinkage plot for the interacted random effects;

  iii. pick one constituency; produce histograms for turnout and percentage remain (Hint: you will need to use the sims.list object from your JAGS model output, and run a prediction model on each simulation.)

UNIVERSITY OF
OXFORD

📄 Michael Betancourt, *A conceptual introduction to hamiltonian monte carlo*, arXiv preprint arXiv:1701.02434 (2017).

📄 George Casella and Edward I George, *Explaining the gibbs sampler*, The American Statistician **46** (1992), no. 3, 167–174.

📄 Siddhartha Chib and Edward Greenberg, *Understanding the metropolis-hastings algorithm*, The american statistician **49** (1995), no. 4, 327–335.

📄 Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell, *Stan: A probabilistic programming language*, Journal of statistical software **76** (2017), no. 1.

UNIVERSITY OF
OXFORD

📄 Sonalde Desai, Reeve Vanneman, and National Council of Applied Economic Research, *India Human Development Survey-II (IHDS-II) 2011-12*, ICPSR36151-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor] (2015-07-31).

📄 Max Goplerud, Shiro Kuriwaki, Marc Ratkovic, and Dustin Tingley, *Sparse multilevel regression (and poststratification (smrp))."*, Unpublished manuscript, Harvard University (2018).

📄 Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin, *Bayesian data analysis*, Chapman and Hall/CRC, 2013.

UNIVERSITY OF
OXFORD

📄 Andrew Gelman, David A Van Dyk, Zaiying Huang, and John W Boscardin, *Using redundant parameterizations to fit hierarchical models*, Journal of Computational and Graphical Statistics **17** (2008), no. 1, 95–122.

📄 Jarod YL Lee, Peter J Green, and Louise M Ryan, *On the" poisson trick" and its extensions for fitting multinomial regression models*, arXiv preprint arXiv:1707.08538 (2017).

📄 Jeffrey R Lax and Justin H Phillips, *How should we estimate public opinion in the states?*, American Journal of Political Science **53** (2009), no. 1, 107–121.

📄 _____, *How should we estimate sub-national opinion using mrp? preliminary findings and recommendations*, annual meeting of the Midwest Political Science Association, Chicago, 2013.

📄 Martyn Plummer et al., *Jags: A program for analysis of bayesian graphical models using gibbs sampling*, Proceedings of the 3rd international workshop on distributed statistical computing, vol. 124, Vienna, Austria., 2003.

📄 Håvard Rue, Sara Martino, and Nicolas Chopin, *Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations*, Journal of the royal statistical society: Series b (statistical methodology) **71** (2009), no. 2, 319–392.