Notes for Statistics Lab 52568 - 2020/21

Or Zuk

1 Israel Election Dataset - Analysis and Methods

1.1 Introduction

We describe here the dataset for elections for the 23nd Knesset in Israel (March 2020), with votes per city. (**Remark:** by city here we mean any 'yeshuv' which can be a city, village, kibutz etc.).

We also describe the data analysis and tools used to answer different questions about the data.

1.1.1 Notations:

- Our dataset is a matrix $\boldsymbol{N} \in \mathbb{R}_{\boldsymbol{C} \times \boldsymbol{K}}$ where K is the number of parties and C is the number of cities. n_{ij} is the number of voters for party j in city i. In addition, we have the following:
- Let $n_{i\bullet} = \sum_{j=1}^{K} n_{ij}$ be the total number of legal votes ('kolot ksherim') in city i. $\tilde{n}_{i\bullet}$ is the total number of eligible voters in city i ('baalei zhut bhira'). From these, we can calculate the voting turnout at city i: $v_i = \frac{n_{i\bullet}}{\tilde{n}_{i\bullet}}$. (In the data file you are given both $\tilde{n}_{i\bullet}$ and the voting turnout, but you need to re-calculate v_i without 'kolot psulim').
- Similarly the number of total votes for party j is $n_{\bullet j} = \sum_{i=1}^{C} n_{ij}$.
- Let $n = \sum_{i=1}^{C} n_{i\bullet} = \sum_{i=1}^{C} \sum_{j=1}^{K} n_{ij}$ be the total votes across all cities ('kolot ksherim'). Similarly, let $\tilde{n} = \sum_{i=1}^{C} \tilde{n}_{i\bullet}$ be the total number of eligible votes in Israel ('baalei zhut bhira').
- We also define $\tilde{n}_{ij}$ - the total number of supporters for party j in city i - that is, how many individual would have voted for party j if all eligible voters in city i were forced to vote. In contrast to the previous quantities, this quantity is not observed and cannot be computed directly from the data.
- Let also z_i is the number of bad votes 'psulim' in city i.

1.2 Computing Parties Vote Share

1. The fraction of votes for party j in the elections is $p_j \equiv \frac{n_{\bullet j}}{n}$. The vector $p = (p_1, ..., p_K)$ represents the share of votes for each party, such that s_i, the number of seats in the parlament for each party, is approximately $s_i \approx 120 p_i$ (the exact relationship is much more complicated, and includes rounding, thresholding small parties due to 'ahuz hasima', the Badder-Offer law etc.).

2. Since the elections are meant to represent the opinions of all citizens in the country, voting turnout may be an issue as it can distort the actual preferences of the citizens - that is, if the turnout for the potential voters of party i is much larger than the turnout of the potential voters of party v_j, then the share of votes p_i may be much higher for the first party compared to p_j for the second, even if in the general population the situation is reversed.

3. A natural question which we would like to answer is: can we infer from the elections results the actual preferences in the population? A followup question is: if every citizen in the coutry would have voted, would we see a significantly different result in the elections?

4. Denote by $\tilde{n}_{\bullet j}$ the (unknown) total number of votes for party j if every citizen actually voted. Similarly, denote by q_j the (unknown) share of votes for the party in this situation, given by $q_j \equiv \frac{\tilde{n}_{\bullet j}}{n}$.

5. Our goal will be to esitmate the q_j values from the election results.

1.3 A Statistical Model for Voting

We assume that each person in Israel decides in advance which party he/she prefers. Then, on election day, people from city i who prefer party j vote

with probability v_{ij}. Therefore, the number of actual voters for party j in village i is $n_{ij} \sim Binom(\tilde{n}_{ij}, v_{ij})$. Both $\tilde{n}_{ij}$ and v_{ij} are unknown parameters, and we will try to estimate them from the data in order to make a correction for the total number of votes. The problem is that the number of unknown parameters is $\approx 2K \times C$ which is on the order of the data size, and we have no hope of estimating the parameters reliably. Therefore, we need to make additional assumptions in order to estimate parameters.

1.4 Simulation Study

Our goal is to estimate the unknown q_j values (partie's proportion in the popoulation) from the observed p_j values (parties proportion in the election). For the real data, we don't know how good will our estimates be.

However, we can make different assumptions on the voting probabilities of individuals, and evaluate the performance of different corrections under these assymptions in a simulation study. The high-level description for a simulation study is as follows:

1. Choose values for the real numbers of voters $\tilde{n}_{ij}$ and voting probabilities v_{ij} , and compute the parties proportions q_j from the $\tilde{n}_{ij}$ values.
2. Simulate (many times) the observed number of voters in the election n_{ij} using $n_{ij} \sim Binom(\tilde{n}_{ij}, v_{ij})$
3. Apply a correction (see next section) to get estimators $\hat{\tilde{n}}_{ij}$ and subsequentially estimators $\hat{q}_j$ for the population proportions
4. Compare the true values q_j to the estimated values $\hat{q}_j$: Compute the empirical bias, variance and mean-suared error of the estimators $\hat{q}_j$.

Will use **parameter tying -** that is, assume that the value of different parameters is the same. We can suggest the following options:
1. Constant voting turnout per city: $v_{ij} = v_i$. (C parameters in total).
2. Constant voting turnout per party: $v_{ij} = u_j$. (K parameters in total).
3. An additive/multiplicative model: $v_{ij} = u_j + v_i$ or $v_{ij} = u_j v_i$. This model will have $K + C$ parameters, still far less than the data size ($K \times C$).

Note: in all of these models we do parameter tying for the v_{ij} parameters. We don't consider here what to do with the $\tilde{n}_{ij}$ parameters. This will be clearer when we estimate the parameters.

1.5 Estimating total votes

We propose here different estimators for the votes distribution if everybody voted. The estimators differ in their assumptions, computation and statistical properties.

1. We can first do the following simple correction: if in city i the voting turnout was v_i, this means that every vote actually counted in this city reprsents not one but v_i^{-1} votes from the populatio of the city. We can thus give weights to the votes in each city. We get the following esitmator for the votes in a city: First, compute the v_i values:

$$v_i = \frac{n_{i\bullet}}{\tilde{n}_{i\bullet}} \tag{1}$$

Then, we can use the v_i values to compute the correction:

$$\hat{\tilde{n}}_{ij} = \frac{n_{ij}}{v_i} \tag{2}$$

$$\hat{\tilde{n}}_{\bullet j} = \sum_{i=1}^{C} \hat{\tilde{n}}_{ij} = \sum_{i=1}^{C} n_{ij} v_i^{-1} \tag{3}$$

$$\hat{q}_j = \frac{\hat{\tilde{n}}_{\bullet j}}{\sum_{k=1}^{K} \tilde{n}_{\bullet k}} = \frac{\hat{\tilde{n}}_{\bullet j}}{\tilde{n}} = \frac{\sum_{i=1}^{C} n_{ij} v_i^{-1}}{\sum_{i=1}^{C} \sum_{j=1}^{C} n_{ij} v_i^{-1}} \tag{4}$$

This estimator adjusts the voting in each city according to the voting turnout. We used this estimator in class, and the results were shown in lab 2.

A main problem with this adjustment is that it assumes that all voters in a city are equally likely to vote. But what if the voters of a certain party are more/less likely to vote?

2. To develop the next estimator, we will assume that the voter turnout for each **party** is a **constant** (rather than for each **city**). Let u_j be the voter turnout for party j. Then we have:

$$\tilde{n}_{\bullet j} = n_{\bullet j} u_j^{-1} \tag{5}$$

and therefore, if we knew the u_j values, we could use the estimator:

$$\hat{q}_j = \frac{n_{\bullet j} u_j^{-1}}{\sum_{k=1}^{K} n_{\bullet k} u_k^{-1}} \tag{6}$$

We next need to estimate the u_js from the data. After we do so, we can just plug in the estimators $\hat{u}_j^{-1}$ into the above equation to get:

$$\hat{q}_j = \frac{n_{\bullet j} \hat{u}_j^{-1}}{\sum_{k=1}^{K} n_{\bullet k} \hat{u}_k^{-1}}. \tag{7}$$

How would we estimate the parties voting turnout? The problem is that we cannot repeat the computation we did before for v_i. If we take Equation (1), the analogous equation for u_j would be:

$$u_i = \frac{n_{\bullet j}}{\tilde{n}_{\bullet j}} \tag{8}$$

But, in contrast to the observable $\tilde{n}_{i \bullet}$ (total number of eligible voters in city i), we don't know $\tilde{n}_{\bullet j}$ (total potential number of voters for party j). Instead, we will develop a different method for estimating the u_j values and for computing the correction, described next.

The idea is conceptualy simple: if in cities where a party is strong we see higher voting turnouts, then the voting turnout for the voters of this parties is high (and the same for lower turnouts indicating a lower turnout for the party). To translate this idea into mathematical formulation, we would like the cities voting turnout v_i to be explained by the parties voting turnouts u_j. That is:

$$v_i \approx \frac{n_{i\bullet}}{\sum_{j=1}^{K} n_{ij} u_j^{-1}}. \tag{9}$$

Since $v_i = \frac{n_{i\bullet}}{\tilde{n}_{i\bullet}}$, we can require $\tilde{n}_{i\bullet} \approx \sum_{j=1}^{K} n_{ij} u_j^{-1}$ for each city i. Summing over the cities, we can formulate a least-squares problem:

$$(\hat{u}_1^{-1}, ..., \hat{u}_K^{-1}) = argmin_{u_1^{-1},..,u_K^{-1}} \sum_{i=1}^{n} (\sum_{j=1}^{K} n_{ij} u_j^{-1} - \tilde{n}_{i\bullet})^2 \tag{10}$$

That is, the inverse turnout parameters u_j^{-1} can be obtained as the least squares solution of a linear regression problem with design matrix N and outcome vector $y = \tilde{n_{|\bullet}}$ where $\tilde{n_{|\bullet}} = (\tilde{n_{1\bullet}}, ..., \tilde{n_{C\bullet}})$. The least-squares solution is therefore:

$$\hat{u}^{-1} = [N^T N]^{-1} N^T \tilde{n_{|\bullet}} \tag{11}$$

and our estimator for q is

$$\hat{q}_j = \frac{\sum_{i=1}^{C} n_{ij} \hat{u}_j^{-1}}{\sum_{i=1}^{C} \sum_{j'=1}^{K} n_{ij'} \hat{u}_{j'}^{-1}}. \tag{12}$$

We can write the estimator in a vector form using matrix and vector operations as follows:

$$\hat{q} = \frac{\mathbf{1}_C^T N diag([N^T N]^{-1} N^T \tilde{n_{|\bullet}})}{\left\| \mathbf{1}_C^T N diag([N^T N]^{-1} N^T \tilde{n_{|\bullet}}) \right\|_1}. \tag{13}$$

where $\mathbf{1}_C^T$ is a constant row vector of ones of length C, $diag(v)$ for a vector v is a diagonal square matrix with the values of v on the diagonal, and $|| \cdot ||_1$ is the L_1-norm: $||x||_1 = \sum_i |x_i|$. The estimator we get here is different from the one in Eq. (4).

Question: How does it perform with respect to the actual election results compared to the previous estimator? we can run it on the real data and look if the results make sense.

Question: Do you see any problems with this estimator? what are the issues that are problematic and/or can be improved?

Alternatively, we may not want to fit the actual number of voters in each city, but rather the proportion how voted in each city

$$(\hat{u}_1^{-1}, ..., \hat{u}_K^{-1}) = argmin_{u_1^{-1},..,u_K^{-1}} \sum_{i=1}^{C} (\sum_{j=1}^{K} \frac{n_{ij}}{n_{i\bullet}} u_j^{-1} - \frac{\tilde{n}_{i\bullet}}{n_{i\bullet}})^2 \tag{14}$$

3. Ideally, we would like to know the parameters v_{ij} representing the voting turnout for the voters of party j at city i. Then, if we have good estimators $\hat{v}_{ij}^{-1}$we can simply use the estimated votes $\hat{\tilde{n}}_{ij} = n_{ij}\hat{v}_{ij}^{-1}$ to estimate q. The problem is that there are too many such parameters $(K \times C)$, which are equal to the number of observation. In fact, any division of the votes for people who didn't vote between the parties will give a valid v_{ij}- for example, assuming that all the voters who didn't vote in the entire country would have voted for the Pirates party.

   Our first estimator essentially assumed that turnout is constant across parties, that is: $v_{ij} = v_i$. The next estimator assumed that turnout is constant across cities (for the same part), that is: $v_{ij} = u_j$ - that is, the turnout for a party varies between cities but in the same way for all parties. We can offer a richer model which combines the two. Let u be the parties turnout vector and let v be a cities turnout vector. We can assume for example an additive model $v_{ij} = v_i + u_j$ or a multiplicative model $v_{ij} = v_i u_j$. This gives a model with $K+C$ parameters which we can try to estimate with the $C \times K$ observations. However, since the linear regression form in Eq. (10) is in terms of the u_j^{-1}(which stand for the v_{ij}^{-1}), it is mathematically more convenient to use the parameterization: $v_{ij}^{-1} = v_i^{-1} + u_j^{-1}$,or $v_{ij} = \frac{1}{1/v_i + 1/u_j}$ i.e. $v_{ij}$is half the harmonic mean of v_i and u_j.

   We can again use least-squares for estimation.

$$\tilde{n_{i\bullet}} \approx \sum_{j=1}^{K} n_{ij}(v_i^{-1} + u_j^{-1}) \tag{15}$$

   and we get the following least-squares optimization problem:

$$(\hat{u}^{-1}, \hat{v}^{-1}) = argmin_{u^{-1},v^{-1}} \sum_{i=1}^{C} (\sum_{j=1}^{K} n_{ij}(v_i^{-1} + u_j^{-1}) - \tilde{n_{i\bullet}})^2 \tag{16}$$

   While this is a convenient linear regression problem, there is a problem here: we have more unknowns $(K + C)$ than equations (C). To overcome this problem, we can use regularization (e.g. Ridge regression), or formulate a different optimization problem and different estimators. It is not clear how best to solve this problem.

1.6 Principal Component Analysis

1.7 Sampling: Using polls (midgam) to predict the elections' outcome

Here, we show how to obain election results from sampling. The goal is to use a subset of the votes and calculate based on this subset an estimator $\hat{p_j}$for the true votes proportion $p_j \equiv \frac{n_{\bullet j}}{n}$ for each party j. There are many sampling methods and we will study a few simple ones. Ideally, we would have liked to sample m random individuals out of the n total votes, where the assumption is that we sample individuals with uniform probability and without replacement. That is, a sample of m out of n individuals is in our poll (midgam). Such sampling is what pollsters try to achieve when they conduct polls, e.g. when using online or telephone polls, although in practice there may be many biases to such polling strategies (e.g. some individuals may not be available / not answer the calls, or lie when asked to which party they vote). Nevertheless, we provide a simple mathematical analysis for such polling, assuming that random uniform sampling is performed.

Let X_{ij} be a Bernoulli random variable set to 1 if individual i voted for party j in the poll, and 0 otherwise. Let $I \subset 1,..,n$ with $|I| = m$ be the random set of individuals belonging to the poll. Then, the simplest estimate for the voting share of party j is the proportion of voters for this party in the poll, that is:

$$\hat{p_j} = \frac{1}{m} \sum_{i \in I} X_{ij}. \tag{17}$$

By linearity of expectation, we have:

$$E[\hat{p_j}] = \frac{1}{m} \sum_{i \in I} E[X_{ij}] = \frac{1}{m} \sum_{i \in I} p_j = p_j. \tag{18}$$

Thus, the estimator $\hat{p_j}$ is unbiased. To calculate the variance of this estimator we need to consider the pairwise covariances between the X_{ij} r.vs. However, we can make the a simplifying approximation, by assuming that the individuals are sampled **with replacement.** This assumption, while incorrect, simplifies the calculations and provides an excellent approximation when $m \ll n$ because the probability that the same individual will be sampled twice is negligible. We get under this assumption indpdendence between the X_{ij}'s and therefore:

$$Var[\hat{p_j}] \approx \frac{1}{m^2} \sum_{i \in I} Var[X_{ij}] = \frac{1}{m^2} \sum_{i \in I} p_j(1-p_j) = \frac{p_j(1-p_j)}{m}. \tag{19}$$

1.7.1 Sampling Ballots

At the night of the elections, it is customary by T.V. stations to conduct exit polls, where voters are asked to specify their votes in the actual elections. This method avoids or at least reduces several of the biases that arise in telephone polls - the individuals asked are only those that actually voted in the elections, and moreover, they are more likely to report correctly their vote in the poll. A disadvatange of these polls is that it is not practical to sample individuals in this manner, and instead a few ballots are sampled, and all voters in these ballots are asked to report their vote. We consider a set $I \subset 1,..,C$ of $b = |I|$ ballots, sampled randomly and uniformly among all C ballots in the country.

The proportion of votes for party j in these ballots is used as an estimator for p_j:

$$\hat{p_j} = \frac{\sum_{i \in I} n_{ij}}{\sum_{i \in I} \sum_{j'=1}^{K} n_{ij'}}. \tag{20}$$

In similar to random sampling of individuals, it is also possible to show that the estimator $\hat{p_j}$ above is unbiased, and it is possible to derive an approximate formula for the variance. We omit these derivations here, and will calculate the variance empirically using simulations.

1.7.2 Startified Sampling

When we perform random sampling, it is possible that certain areas and populations are not represented in the sample. To reduce this risk, it is possible to design the sample such that we ensure that different groups are represented. This is called stratified sampling ("midgam schavot"). It is deriable to divide to homogenous groups, and sample ballots from each such group, i.e. to minimize the diversity within a the groups, and maximize the differences between groups. Formally, consider a stratified sample with L strata ("layers"), $\bigcup_{l=1}^{L} A_l = 1, .., C$ and $A_l \bigcap A_r = \emptyset$, $\forall l \neq r$. We divide the b ballots of our poll into $b_1..., b_L$ ballots where b_l ballots are used for strata A_l, with $\sum_{i=1}^{L} b_i = b$. The set of ballots is thus $I = \bigcup_{i=1}^{L} I_l$ with $I_l \subset A_l$ and $|I_l| = b_l$. We next compute the frequency of each party in each strata:

$$\hat{p_j^{(l)}} = \frac{\sum_{i \in I_l} n_{ij}}{\sum_{i \in I_l} \sum_{j'=1}^{K} n_{ij'}}. \tag{21}$$

Next, our estimator is obtained as a weighted average of the strata estimates:

$$\hat{p_j} = \frac{\sum_{l=1}^{L} N_l \hat{p_j^{(l)}}}{\sum_{l=1}^{L} N_l}, \tag{22}$$

where N_l is the total number of eligible votes in strata l, given by: $N_l = \sum_{i \in A_l} \sum_{j=1}^{K} n_{ij}$.

The above estimator can be shown to be unbiased, in similar to the random sampling estimator. If the strata A_l and the number of ballots in each strata b_l are chosen carefully (for example we can consider $L = 10$ socio economic "eshkolot" since we know that voting patters are different between them), this estimator can have a reduced variance, compared to the random sampling estimator. However, a bad choice can actually increase the variance of the estimator.

1.7.3 Shrinkage using previous results

We can combine the information obtained from the sample with other prior information to improve the accuracy of our estimator. For example, if we know that a certain party usually gets around 10% of votes, but in a small poll we have conducted this party got 15% of the votes, we may suspect that this high results occured by chance or due to unknown biases, and reduce our estimate for this party from 15% to a lower figure. To formalize this intuition, let r_j be the total proportion of votes to party j in previous elections (e.g. in the September 2019 elections in Israel), and let $p_j$be the (unknown) proportion in the upcoming elections (e.g. suppose that we wanted to predict the 2020 elections results in advance). Let $\hat{p_j}$ be our estimator from the poll (for example, a random poll of ballots), and let $0 \leq \alpha \leq 1$ be a scalar. Then, we can define the following estimator:

$$\hat{p_j}(\alpha) = \alpha \hat{p_j} + (1-\alpha) r_j. \tag{23}$$

Suppose that $\hat{p_j}$ is an unbiased estimator for p_j with variance σ^2. Then, we can compute the bias and variance of the combined estimator $\hat{p_j}(\alpha)$:

$$E[\hat{p_j}(\alpha)] - p_j = (1-\alpha)(r_j - p_j). \tag{24}$$

$$Var[\hat{p_j}(\alpha)] = \alpha^2 \sigma^2. \tag{25}$$

and the mean-squared error is:

$$MSE[\hat{p_j}(\alpha)] = (1-\alpha)^2 (r_j - p_j)^2 + \alpha^2 \sigma^2. \tag{26}$$

Our goal is to minimize the MSE, i.e. the sum of the variance and the squared-bias. As we increase α we reduce the bias of the estimator, but increase the variance. This is called bias-variance tradeoff.

It is possible to find the optimal α minimizing the MSE as a function of σ^2, r_j and p_j. In general this optimal α will be smaller than 1 - that is, using the information from the previous elections can reduce our mean-squared error, although it introduces bias, because it reduces the variance.