Swing Voters’ Vote Choice Prediction Using Multilevel Logit Model to Improve Election Survey Accuracy

D Irvani1,2, K Sadik2*, A Kurnia2, A Saefuddin2, Erfiani2

1 Saiful Mujani Research and Consulting, Indonesia
2 Department of Statistics, IPB University, Indonesia

*E-mail: kusmans@apps.ipb.ac

Abstract. Public opinion surveys are often used to predict an election result. However, the predictions are not always accurate due to many factors. The presence of swing voters at the time of survey is one of the sources of the inaccuracy. On the other hand, election surveys are also often conducted by using multi-stage random sampling method so that ordinary models such as logit model generally do not provide satisfactory results. The data, hence, is complex and may be approached by multilevel models. The study is conducted to assess the extent to which a prediction of swing voters’ vote choice through a multilevel logit model can improve survey accuracy. The data used in this study was the result of a survey conducted using stratified multistage random sampling method two weeks before the 2019 presidential election. The model with 15 predictors and random effects for villages and neighborhood providing 96.3% accuracy and AUC reached 99.1% in the validation process. Based on the final model, the swing voters in this survey were predicted to vote more for Candidate B (10.4%) than Candidate A (7.5%). The direction of the swing voters’ support different from the loyal voters who prefer Candidate A (49.1%) than Candidate B (33.0%). The prediction of swing voters’ vote choice using multilevel logit model significantly improved the survey accuracy. Before the swing voters’ support was predicted the absolute deviation between the survey result and the election result was quite large, around 6.4%-11.5%. After swing voters’ support was predicted the absolute difference shrank to 1.1%.

1. Introduction
Public opinion surveys are often used to predict an election result. However, the predictions are not always accurate due to many factors. The presence of swing voters at the time of survey is one of the sources of the inaccuracy. Hence, predicting election result will be related to predicting the swing voters’ vote choice.

On the other hand, election surveys are also often conducted by using multi-stage random sampling method. The method is generally applied for efficiency or due to lack of complete sampling frame. The data has a complex structure so that ordinary models such as logit model generally do not provide satisfactory results. The complex data may be approached by multilevel model [1].

Multilevel models have been used in various studies. Austin et al. [2] depicts that the models are increasingly being used in health studies. Multilevel models are also popular in voting behavior studies. Vermonte [13] used a multilevel multinominal logit model to analyze factors influenceed party choice. Gelman et al [4] used a multilevel model to analyze the existence of swing voters. Meanwhile, Kiewiet de Jonge et al [7] used the model to predict voter turn out and vote choice.
This study is conducted to assess the extent to which a prediction of swing voters' vote choice through a multilevel logit model can improve survey accuracy. The data used in this study was the result of a survey conducted using stratified multistage random sampling method two weeks before the 2019 presidential election. The result of this study is expected to be useful for academic and practical purposes.

2. Literature Review

2.1. Multilevel Logit Model

The multilevel logit model (for a two-level case) has equation formula [1]

\[
g[E(y_{ij} | u_i)] = \text{logit}[P(y_{ij} = 1 | u_i)] = x_i \beta + z_{ij} u_i, \tag{1}
\]

where \(y_{ij}\) is the binary response for subject \(j\) in cluster \(i\). \(\beta\) is the fixed effects for explanatory variables \(x_i\), \(\{u_i\}\) are random effects for cluster \(i\) assumed to have a particular probability distribution, usually assumed to be independent of \(N(0, \sigma_u^2)\). \(x_{ij}\) and \(z_{ij}\) are row vectors of known values of explanatory variables for subject \(j\) in cluster \(i\). Estimation of parameters can be done using the marginal likelihood method. For observations of \(y\) and random effects \(u\), let \(f(y | u; \beta)\) be the probability function for \(y\) given \(u\), and \(f(u; \Sigma_u)\) is the normal density function for \(u\). Then the marginal likelihood function is

\[
\ell(\beta, \Sigma_u; y) = f(y; \beta, \Sigma_u) = \int f(y | u; \beta) f(u; \Sigma_u) \, du. \tag{2}
\]

In particular, the marginal likelihood function in equation (2) for a two-level logistic-normal random intercept model is

\[
\ell(\beta, \sigma_u^2; y) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{j=1}^{d} \left[ \frac{\exp(x_{ijk} \beta + u_i)}{1 + \exp(x_{ijk} \beta + u_i)} \right]^{y_{ij}} \frac{1}{1 + \exp(x_{ijk} \beta + u_i)} f(u_i; \sigma_u^2) du_i. \tag{3}
\]

**Marginal likelihood** \(\ell(\beta, \Sigma_u; y)\) can be approximated numerically via several methods: Gauss-Hermite quadrature, Monte Carlo, and Laplace [1,6].

2.2. Stratified-Multistage Random Sampling

Stratified-multistage random sampling is a complex sampling method which combines stratified sampling and cluster sampling techniques, and sample selection is carried out gradually. Suppose a population \(U\) is divided into \(H\) subpopulations or strata, and \(U_h\) consists of \(N_{1h}\) primary sampling units (PSU). Suppose that \(PSU_i\) in stratum \(h\) has \(N_{2hi}\) secondary sampling units (SSU), \(SSU_j\) in \(PSU_i\) in stratum \(h\) has \(N_{3hij}\) tertiary sampling units (TSU), and \(TSU_k\) in \(SSU_j\) in \(PSU_i\) in stratum \(h\) has \(N_{4hijk}\) ultimate sampling units (USU). From this population structure, then sample selection is carried out gradually. First, \(n_{1h}\) PSU’s are selected randomly in each stratum \(h\). Second, \(n_{2hi}\) SSU’s are selected randomly in each selected \(PSU_i\). Third, \(n_{3hij}\) TSU’s are selected randomly in each selected \(SSU_j\). Fourth, \(n_{4hijk}\) USU’s are selected randomly in each selected \(TSU_k\).

Suppose \(y_{hijkl}\) is the value of a variable \(y\) on \(PSU_i\) in \(TSU_k\) in \(SSU_j\) in \(PSU_i\) in stratum \(h\), then the unbiased estimator for the population total is [10]

\[
\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_{1h}} \sum_{j=1}^{n_{2hi}} \sum_{k=1}^{n_{3hij}} \sum_{l=1}^{n_{4hijk}} w_{hijkl} Y_{hijkl}, \tag{4}
\]

where \(w_{hijkl} = \frac{n_{1h}}{n_{1h}} \times \frac{n_{2hi}}{n_{2hi}} \times \frac{n_{3hij}}{n_{3hij}} \times \frac{n_{4hijk}}{n_{4hijk}}\). The unbiased estimator for the variance of the population total estimate is
\[
\hat{\mu} = \frac{\hat{Y}}{M},
\]

where \(M = \sum_{h=1}^{H} \sum_{i=1}^{n_{1h}} \sum_{j=1}^{n_{2hi}} \sum_{k=1}^{n_{3hij}} \sum_{l=1}^{n_{4hijkl}} w_{hijkl} \) [11]. The variance of (6) can be estimated by linearization method [8]:

\[
\hat{\sigma}^2 = \hat{\mu}^2 \left( \hat{\varphi} - 2 \hat{\varphi}(\hat{Y}) + \hat{M}^{-2} \hat{\varphi}(M) - 2(\hat{Y} \hat{M})^{-1} \hat{\varphi}(\hat{Y}, \hat{M}) \right),
\]

where \(\hat{\varphi}(\hat{M})\) is the estimated variance for the population size estimate and \(\hat{\varphi}(\hat{Y}, \hat{M})\) is the covariance between \(\hat{\mu}\) dan \(\hat{M}\). Statistics \(\hat{\varphi}(\hat{M})\) and \(\hat{\varphi}(\hat{Y}, \hat{M})\) can be obtained based on (5) by modifying the \(s^2_{1h}\), \(s^2_{2hi}\) dan \(s^2_{hijkl}\) according to calculation objectives.

The confidence interval \((1-\alpha)\) for \(\mu\) is

\[
CI \ (1 - \alpha) = \hat{\mu} \pm z_{\alpha/2} SE,
\]

where \(z_{\alpha/2}\) is the standard normal statistic, \(\alpha \in (0,1)\) is the level of significance desired, and \(SE = \sqrt{\hat{\sigma}^2(\hat{\mu})}\) is the estimated standard error for \(\hat{\mu}\).

### 3. Method

#### 3.1. Data

The data used in this study was the result of a national survey conducted two weeks ahead of the 2019 presidential election with a total of 2,285 respondents. Respondents were selected using stratified multistage random sampling method with the following procedure. First, the population is stratified by province and rural-urban status, and sample size in each stratum is allocated proportionally to the population size. Respondents were selected gradually. First, villages (PSU) in each stratum were selected with simple random sampling without replacement. Second, 4 RT/neighborhood (SSU) were selected randomly without replacement in each selected village. Third, 2 families (TSU) were selected at random without replacement in each selected RT. Forth, 1 family member (SSU) in each selected family who has the right to vote was selected at random as respondent.
3.2. Variables
The response variable in this study is vote choice (1=Candidate A, 0=Candidate B). The explanatory variables include demographics (gender, rural-urban, age, education, ethnicity, religion, region), score of the possibility of choosing a candidate, assessment of candidate’s personal quality, evaluation of the incumbent's performance, assessment of candidate's performance in presidential debate, party choice, support for candidates in a regional election, attitudes towards negative issues about candidates, and intensity of following social media (Table 1).

Table 1. Variables and The Explanation

| No | Variables                                      | Description/Code                                      |
|----|-----------------------------------------------|-------------------------------------------------------|
| 1  | Presidential-vice presidential vote           | 1=Candidate A, 0=Candidate B                           |
| 2  | Gender                                        | 1=Male, 0=Female                                      |
| 3  | Rural-Urban                                   | 1=Rural, 0=Urban                                      |
| 4  | Age                                           | In years                                              |
| 5  | Education                                    | Scale 1-10: 1=Never went to school, 10=Bachelor’s degree or higher |
| 6  | Ethnicity                                     | 1=Javanese, 0=Others                                  |
| 7  | Religion                                      | 1=Islam, 0=Others                                     |
| 8  | Region                                        | 1=Central Java-DIY, 0=Others                          |
| 9  | Possibility Score of Voting for a Candidate   | Scale -10 to +10: -10=definitely vote for Candidate B, +10=definitely vote for Candidate A |
| 10 | Personal Quality                              | A composite index of four likeability items of Candidate: (1) A1, (2) A2 (3) B1, (4) B2. Each item is measured on a 3-point scale: -1 = Dislike, 0 = Neutral, 1 = Like. The index is formed by the formula: Personal Quality = (A1 + A2) - (B1 + B2), resulting in an index on a scale of -4 to +4, where -4 means likes B1 and B2 but dislikes A1 and A2, and + 4 means likes A1 and A2 but dislikes B1 and B2. |
| 11 | Incumbent's performance                       | Scale 1-5: 1=Highly dissatisfied, 5=Highly satisfied. |
| 12 | Assessment of the presidential debate         | Scale 1-3: 1=Candidate B won the debate, 2=Neutral, 3=Candidate A won the debate. |
| 13 | Party choice                                  | 1=Vote for party which support Candidate A, 0=Vote for party which support Candidate B/Others. |
| 14 | Support for gubernatorial candidate in DKI election | 1=Support Candidate X, 0=Support Candidate Y/Neutral |
| 15 | Candidate Negative Issues                     | A composite index of 5 items of attitudes towards negative issues about Candidate A: (1) PKI, (2) PRC accomplice, (3) Anti-Islam, (4) Authoritarian, (5) abusing one's position, where each item is measured by a scale 3 points, 1 = disagree, 2 = neutral, 3 = agree. The five items are averaged to produce an index on a scale of 1-3, where 1 = disagree with all negative opinions about Candidate A, 3 = agree with all negative issues about Candidate A. |
| 16 | Social Media Access                          | A composite index of 7 items about intensity in accessing social media: (1) Facebook, (2) Twitter, (3) Instagram, (4) Youtube, (5) BBM, (6) WhatsApp, (7) Line; where each item is measured on a scale of 1-5 (1 = never, 5 = every day or most days). The seven items are averaged to form an
3.3. Model Specification
The multilevel logit model used in this study is a three-level logit model
\[
\text{logit}(P(y_{ijk} = 1|u_i, v_{ij}) = x_{ijk} \beta + u_i + v_{ij}, \tag{9}
\]
where \(P(y_{ijk} = 1|u_i, v_{ij})\) is the probability of choosing Candidate A for person \(k\) in RT \(j\) and village \(i\) given \(u_i\) and \(v_{ij}\), \(u_i\) is the random effect for village \(i\) and assumed that \(\{u_i\}\) are to be independent from a \(N(0, \sigma_u^2)\), \(v_{ij}\) is the random effect for RT \(j\) in the village \(i\) and \(\{v_{ij}\}\) are assumed to be independent from \(N(0, \sigma_v^2)\), \(\beta\) are the fixed effects for explanatory variables \(x\).

3.4. Data Analysis
The data analysis was divided into two parts, cross validation and prediction, as follows.

1. 10-folds cross-validation. The loyal voters data was divided randomly into 10 groups, where one group as testing data and the other 9 groups became training data. The model (9) the was estimated using the training data, and the model was used to predict the testing data using various cutoff. This procedure was repeated (10 times) using different testing and training data groups to predict entire loyal voters data. The performance of prediction was assessed based on the accuracy = (true positive + true negative) / sample size, sensitivity = true positive / (true positive + false negative), specificity = true positive / (true negative + false positive), and the area under the ROC (curve plots true positive rates and false positive rates) or the AUC (area under the curve). A cutoff point which produced the most balanced sensitivity and specificity was determined as the optimum cutoff, which is one of the common ways for determining the cut off [5].

2. Prediction. The Model (9) was estimated using all loyal voters data and the model was used to predict the responses in the swing voters data using the optimum cutoff. The proportion of the total electability of the candidates was estimated using (6) and the confidence interval was estimated using (8). Finally, result of prediction was compared with the official result.

The data analysis was performed using R software, especially \textit{lme4} package by Bates et al [3] and \textit{survey} package by Lumley [9].

4. Result and Discussion

4.1. Preliminary Analysis
The survey result two weeks before election showed that Candidate A received 49.1% of popular votes and Candidate B 33.0%. Meanwhile, around 17.9% of voters were not able to confirm their choice (swing voters). Even though Candidate A was superior to Candidate B when this survey was conducted, the support obtained was still below 50%. The survey result didn’t provide a clear picture of the likely outcome of the election.

With the idea that political choice may related to demographic factors (sociological approach), as a first step, a simple analysis was carried out by comparing the demographic profiles of supporters of both candidates and swing voters. The descriptive analysis result showed that there were differences in demographic characteristics between the Candidates A supporters and Candidates B supporters, especially in terms of education, religion, region, and rural-urban status. Compared to the Candidate B supporters, the Candidate A supporters had a lower level of education, a lower proportion of Muslim voters, a larger proportion of Javanese, and a larger proportion of living in Central Java-DIY and rural area (Table 2). If the profiles of the two supporters are compared to the swing voters, it can be seen that the swing voters’ characteristics were relatively closer to those of Candidate B supporters. Therefore, intuitively, more swing voters likely to vote for Candidate B than Candidate A.
Table 2. Voters Demographic Profiles

|                | Gender (Male) | Rural-Urban (Rural) | Ethnicity (Javanese) | Religion (Moslem) | Region (Central Java-DIY) | Age | Education |
|----------------|---------------|---------------------|----------------------|-------------------|----------------------------|-----|-----------|
| Candidate A Voters | 0.502         | 0.561               | 0.525                | 0.824             | 0.237                      | 44.189 | 4.682    |
| Candidate B Voters | 0.500         | 0.407               | 0.253                | 0.998             | 0.070                      | 42.339 | 5.482    |
| Swing Voters     | 0.467         | 0.488               | 0.429                | 0.956             | 0.118                      | 42.756 | 5.216    |
| Absolute difference between of Candidate A Voters and Swing Voters | 0.035         | 0.072               | 0.097                | 0.132             | 0.119                      | 1.433 | 0.534    |
| Absolute difference between Candidate B Voters and Swing Voters | 0.033         | 0.082               | 0.175                | 0.042             | 0.048                      | 0.417 | 0.266    |

4.2. Cross-Validation and The Optimum Cutoff
The cross-validation analysis shows that the prediction with 15 predictors and random effects for villages and neighborhood providing the AUC 99.1%. The optimum cutoff is 0.64. By using this cutoff, the prediction on testing data provides 96.3% accuracy (Figure 1).

Figure 1 (a) Cross-validated ROC and AUC, (b) Plot for cross-validated sensitivity and specificity against cutoff.
4.3. The Final Model and Swing Voters’ Vote Choice Prediction

The final model was carried out based on all loyal voters data (n = 1883) and the model was used to predict the swing voter choice (n = 402). From this model, the random effects for villages and RT’s have estimated variance of $\hat{\sigma}_u^2 = 0.4758$ and $\hat{\sigma}_v^2 = 1.203 \times 10^{-8}$ consecutively. The variation of villages has an important role in explaining response variation, while the variation of RT’s does not have a large effect (Table 3).

The next step is to predict the swing voters’ vote choice using the final model. The model predicted that the swing voters would heavily vote for Candidate B more than Candidate A. Of a total of 17.9% of swing voters, Candidate B was predicted to get an additional 10.4% of votes while Candidate A’s votes would only increase by 7.5%. The direction of the swing voters’ support different from the loyal voters who generally preferred Candidate A (49.1%) than Candidate B (33.0%), as suspected in the comparative voter profile analysis in the previous section.

The difference in support between loyal and swing voters also shows that an election survey is at risk of bias when it is unable to gather information from a particular group (swing voters). This finding confirms the previous study by Shirani-Mehr et al. [12] that many election surveys actually had larger total errors than expected due to non-sampling error problem.

Table 3. The Multilevel Logit Model

|                     | $\beta$ | $SE(\beta)$ | Pr(>|z|) |
|---------------------|---------|-------------|----------|
| Fixed Effects       |         |             |          |
| (Intercept)         | -0.93075| 1.82858     | 0.610754 |
| Gender              | 0.02965 | 0.31766     | 0.925630 |
| Rural-Urban         | 0.73888 | 0.36269     | 0.041631 |
| Age                 | -0.0016 | 0.01421     | 0.910323 |
| Education           | -0.10505| 0.07688     | 0.171825 |
| Ethnicity           | 1.04768 | 0.37580     | 0.005306 |
| Religion            | -3.21774| 1.20105     | 0.007382 |
| Region              | 0.22429 | 0.61555     | 0.715576 |
| Possibility Score of Voting for a Candidate | 0.57596 | 0.05648 | < 2e-16 |
| Personal Quality    | 0.52898 | 0.14863     | 0.000372 |
| Incumbent's performance | 0.74692 | 0.17188 | 1.39e-05 |
| Presidential Debate | 1.38195 | 0.35793     | 0.000113 |
| Party choice        | 0.88584 | 0.32180     | 0.005910 |
| Support for gubernatorial candidate in DKI election | 0.27023 | 0.44950 | 0.547728 |
| Candidate Negative Issues | -1.48653 | 0.40700 | 0.000260 |
| Social Media Access | -0.09175| 0.26181     | 0.726002 |
| Random Effects      |         |             |          |
| Estimated variance of random effects for villages $\{u_i\}$: $\hat{\sigma}_u^2 = 0.4758$ |
| Estimated variance of random effects for RT $\{v_{ij}\}$: $\hat{\sigma}_v^2 = 1.203 \times 10^{-8}$ |

$n = 1883$

4.4. Prediction of Total Electability and Survey Accuracy

This study has an opportunity to investigate the model performance empirically by comparing the prediction result with the election result. After the swing voters’ support was predicted the total votes for candidate A would be 56.6% and Candidate B 43.4%. This prediction turned out to be close to the
election result which were held about two weeks after the survey: Candidate A 55.5% and Candidate B 44.5% (Table 4).

Before the swing voters’ support was predicted, the survey result seemed quite far from the election results: the absolute difference between the survey and the election result on Candidate A's electability was around 6.4%, while the absolute difference on Candidate B's electability was around 11.5%. After swing voters’ support was predicted the difference between the survey and the election narrowed considerably became ± 1.1%. This is an empirical evidence that a predictive model of swing voters’ vote choice using multilevel logit can improve survey accuracy.

Table 4. Comparison of Survey and Election Result. The number in parentheses is 95% confidence interval.

|                  | Survey Result               | Election Result |
|------------------|-----------------------------|-----------------|
|                  | Without Swing Voters’ Vote  |                 |
|                  | Choice Prediction           |                 |
| Candidate A      | 49.1% (46.4%-51.9%)         | 55.5%           |
| Candidate B      | 33.0% (30.2%-35.9%)         | 44.5%           |
| Swing Voters     | 17.9% (15.8%-20.0%)         |                 |
| Total            | 100.0%                      | 100.0%          |

5. Conclusion

The multilevel logit model with 15 predictors and random effects for villages and neighborhoods providing 96.3% prediction accuracy and the area under the ROC curve reached 99.1% in cross-validation process. Based on the final model, the swing voters were predicted to vote more for Candidate B (10.4%) than Candidate A (7.5%). The prediction of swing voters’ vote choice using multilevel logit model significantly improved the survey accuracy. Before the swing voters’ support was predicted, the absolute deviation between the survey result and the election result was quite large, around 6.4%-11.5%. After swing voters’ support was predicted, the absolute difference shrunk to 1.1%.

In this study the standard error of candidate’s electability was estimated using design-based method by assuming that the swing voters predicted response values as actual observations. This approach certainly has limitations due to the fact that these are a set of predicted values and hence contain prediction errors. This case may pave the way for further research.

References

[1] Agresti A 2015 *Foundations of Linear and Generalized Linear Models* (Hoboken: Wiley)

[2] Austin PC and Merlo J 2016 Intermediate and advanced topics in multilevel logistic regression analysis *Statistics in Medicine*, 36(20): 3257-3277

[3] Bates D, Maechler M, Bolker B, & Walker S 2015 Fitting Linear Mixed-Effects Models Using
Gelman A, Goel S, Rivers D, & Rothschild D 2016 The Mythical Swing Voter Quarterly Journal of Political Science 11(1), 103-130

Habibzadeh F, Habibzadeh P, & Yadollahie M 2016 On determining the most appropriate test cut-off value: the case of tests with continuous results Biochimia Medica; 26(3): 297–307

Handayani D, Notodiputro K A, Sadik K, & Kurnia A 2017 A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM) AIP Conference Proceedings, 1827, 020033

Kiewiet de Jonge CP, Langer G, & Sinozich S 2018 Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression with Poststratification (MRP) Public Opinion Quarterly 82(3), 419–446

Lehtonen R and Pahkinen E 2004 Practical Methods for Design and Analysis of Complex Surveys, 2nd Ed (Chichester: Wiley)

Lehman T 2019 survey: analysis of complex survey samples R package version 3.35-1

Preston J 2009 Rescaled Bootstrap for Stratified Multistage Sampling Survey methodology 35(2): 227-234

Scheaffer R L, Mendenhall W, Ott L, & Gerrow K G 2012 Elementary Survey Sampling, 7th ed. (Boston: Books/Cole)

Shirani-Mehr H, Rothschild D, Goel S, & Gelman A 2018 Disentangling Bias and Variance in Election Polls Journal of the American Statistical Association, 113, 607–614

Vermonte P J 2014 People, places and voting behavior: The case of Indonesia’s 2004 legislative elections (Northern Illinois University, PhD dissertation).