An Update To The Washington Post Election Night Model

John Cherian, Lenny Bronner

December 2020

In October, we published a write-up of the model that we built for the general election on Nov. 3. In the nearly two months since, we've updated it and made improvements that we'd like to share.

As a quick reminder, our model uses vote counts as they are reported on election night to forecast turnout, Democratic votes and Republican votes. Additionally, we quantify the uncertainty around those predictions. We apply conformalized quantile regression to forecast the percentage change in votes between the previous election results and the true results for unobserved counties. The covariates used for this prediction include demographic features such as the proportion of registered voters that are Black, Hispanic and Asian, the proportion that are over the age of 65 and under the age of 30, median household income and the proportion with a Bachelor degree or more. Our prediction then consists of the 0.5th quantile from this regression. The "fuzzy bars" on the forecast page that quantify our uncertainty around this prediction are computed by applying a correction (i.e., the "conformalization" in conformalized quantile regression) to the 0.05 and 0.95th quantiles predicted by this model. To get a better sense of our method, please take a look at our original write-up.

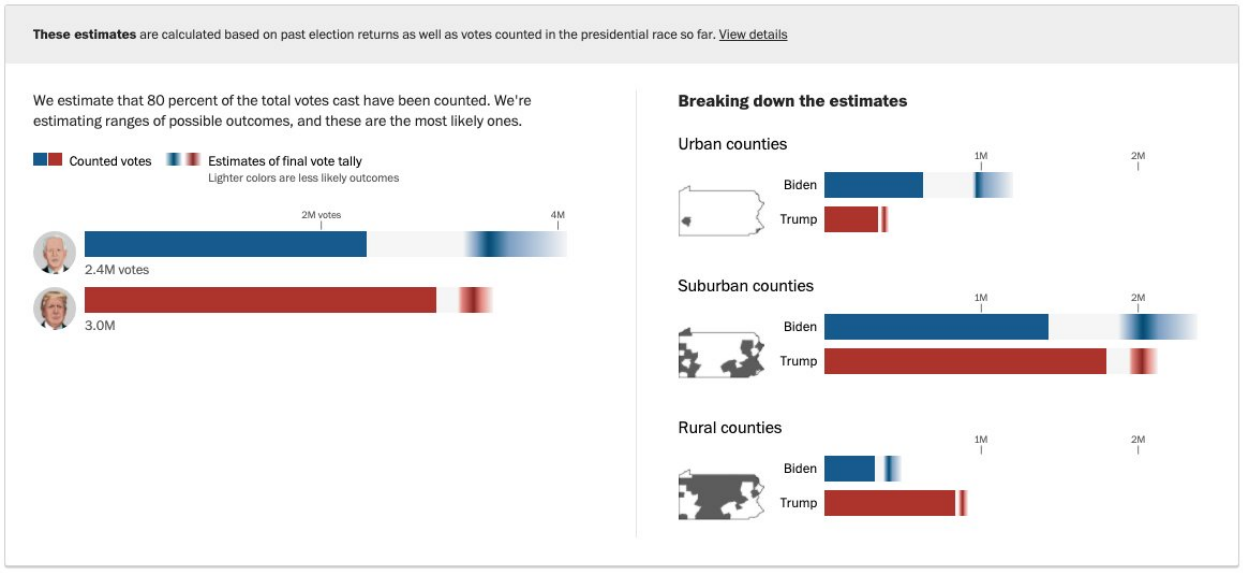


Figure 1: Model visualization for Pennsylvania on the morning of Nov. 4.

1 Precinct Modeling

In the general election, we regressed county results against the previously mentioned covariates. For the Georgia run-offs, we will fit our model using precinct results. Our model can only be updated when a

geography fully reports, i.e., a county or a precinct records and then reports all of its votes. By forecasting precincts rather than counties, we will be able to update our model more rapidly as there are many more precincts and they are much smaller than counties. Thus, we reach a critical mass of finished precincts before we would reach the same number of finished counties.

Precincts also help us solve a problem with final reporting. Since precincts are either fully reporting their results or not reporting results at all, we will have a better sense of how many geographies are completed and thus can be used in the model. At many points during November's general, we had problems knowing which counties had actually finished counting all their votes.

In order to generalize our model, we will refer to the smallest geographic entity for which we produce predictions as *units*. Depending on the granularity of reported data in future elections, units will refer to *either* counties or precincts.

2 Conformal Aggregations

Our second major update is a change in how we aggregate unit-level prediction intervals from our model to produce a state-level forecast. This model improvement would not be possible without the invaluable contributions of Lihua Lei, who both conceived of and developed the parametric model that we applied to the non-conformity scores.

2.1 Previous Work

As we emphasized in the original write-up, the only assumption we make is that the data points, and, therefore, the non-conformity scores, are exchangeable.[1] Our lack of distributional assumptions ensures that our model is robust, but it complicates the task of producing state-level prediction intervals.

In our initial write-up, we discussed the problem of "aggregating" prediction intervals, and ultimately presented a straightforward, but conservative solution. Namely, we computed a state prediction interval by summing the lower and upper bounds of all relevant unit prediction intervals. Under the assumption that the unit-level errors of our forecast are jointly Gaussian, applying this method is equivalent to assuming perfect correlation between the unit-level errors; this ultimately yields a state-level prediction interval whose width grows *linearly* in the number of units that are aggregated. By contrast, if we assumed that the unit-level errors were independent, the state-level prediction interval would only grow on the order of the square root of the number of aggregated units. Luckily, in our general election model, units corresponded to counties, so our conservative aggregation procedure was not particularly cumbersome. The high true correlation between county-level forecast errors as well as the limited number of counties in any given state ensured that our prediction intervals did not become overly wide.

For our precinct model, the conservative aggregation strategy of summing unit interval bounds no longer suffices. Over 2,000 precincts will report votes in the Georgia run-offs, so summing thousands of precinct-level bounds is a bad idea. Doing so results in state-level prediction intervals whose width exceeds 10% of the total votes cast even after over half the precincts have reported! While the errors in our precinct forecasts are still far from independent, an aggregation strategy that assumes perfect correlation is too conservative.

2.2 Parametric Conformalization

After splitting the observed units into training and calibration sets, $\mathcal{T}$ and $\mathcal{C}$, respectively, recall that we compute 90% unit prediction intervals as follows:

1. Fit $\hat{r}_{0.95}(x_i)$ and $\hat{r}_{0.05}(x_i)$ using $(x_i, r_i) \in \mathcal{T}$
2. Compute non-conformity scores $E_i = \max(r_i - \hat{r}_{0.95}(x_i), \hat{r}_{0.05}(x_i) - r_i)$ for $(x_i, r_i) \in \mathcal{C}$

[1] Assuming exchangeability is equivalent to assuming that the likelihood of the data is invariant to permutations of the data indices (i.e. all orderings of the data are equally likely). The assumption that the data points are independent and identically distributed (i.i.d.) implies exchangeability, but not vice-versa.

3. Estimate conformal correction: $C \approx$ 90-th percentile of the E_i

4. Output $(\hat{r}_{0.05}(x_j) - C, \hat{r}_{0.95}(x_j) + C)$ given x_j for which r_j is unobserved

Exchangeability implies that an unseen non-conformity score is equally likely to fall in any place amongst the E_i of the calibration set. This implies that unseen non-conformity scores are only larger than C with probability 10%. So, "correcting" the prediction interval by C ensures that the interval coverage is 90%.[2] Our exchangeability assumption is thus sufficient to bound an unseen E_j and thereby produce informative unit-level prediction intervals. However, it does not help us upper bound the value of a sum of unseen non-conformity scores, $\sum_{j=1}^{k} E_j$. To generate narrower aggregate intervals, we must make additional assumptions.

Let us denote the two expressions whose maximum is defined to be E_i as follows:

$$L_i := \hat{r}_{0.05}(x_i) - r_i$$
$$U_i := r_i - \hat{r}_{0.95}(x_i)$$

We now assume the following model for L_i and U_i.

$$L_i \sim \mathcal{N}(\mu_l, \sigma_l^2); \quad \text{Cor}(L_i, L_j) = \rho_l$$
$$U_i \sim \mathcal{N}(\mu_u, \sigma_u^2); \quad \text{Cor}(U_i, U_j) = \rho_u$$

To be clear, we are not assuming in this model that our forecast error is Gaussian; instead we are claiming that the upper and lower non-conformity scores are Gaussian with distinct parameterizations. The resulting prediction intervals will still be asymmetric, but we replace our assumption of exchangeability for L_i and U_i with a stronger assumption that allows us to perform inference on $\sum_{j=1}^{k} L_j$ and $\sum_{j=1}^{k} U_j$.[3]

Using this working model, we can derive an alternative method for constructing aggregated prediction intervals with marginal coverage.[4] Let S denote the set of units belonging to a particular state and w_i denote the previous result in unit i (i.e., the normalizing constant in the forecasted residual), and as before $\mathcal{C}$ is the calibration set. Then, using some elbow grease, we can show that:

$$\frac{\sum_{i\in\mathcal{C}} w_i L_i}{\sum_{i\in\mathcal{C}} w_i} - \frac{\sum_{j\in S} w_j L_j}{\sum_{j\in S} w_j} \sim \mathcal{N}\left(0, (1-\rho_l)\sigma_l^2\gamma^2\right)$$
$$\text{where } \gamma = \frac{\sum_{i\in\mathcal{C}} w_i^2}{\left(\sum_{i\in\mathcal{C}} w_i\right)^2} + \frac{\sum_{j\in S} w_j^2}{\left(\sum_{j\in S} w_j\right)^2}$$

Letting z_α denote the α-quantile of a standard normal, this implies that:

$$P\left(\sum_{j\in S} w_j L_j > \sum_{j\in S} w_j \left(\frac{\sum_{i\in\mathcal{C}} w_i L_i}{\sum_{i\in\mathcal{C}} w_i} - z_\alpha\sqrt{1-\rho_l}\sigma_l\gamma\right)\right) = 1-\alpha \tag{1}$$

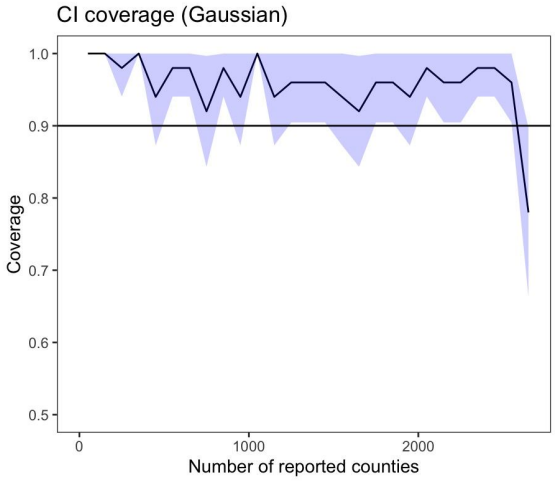
Luckily, we can compute this lower bound on the non-conformity score! Note that after selecting a calibration set, we observe $\frac{\sum_{i\in\mathcal{C}} w_i L_i}{\sum_{i\in\mathcal{C}} w_i}$. And even though $(1-\rho_l)\sigma_l^2$ is unknown, the sample variance of the observed L_i is an unbiased estimator for this quantity. Since we are relying on a noisy estimate of this parameter, we must be careful before claiming that the computed lower bound holds with probability $1-\alpha$. To account for the error in our parameter estimate, we first use the non-parametric bootstrap to compute a confidence interval for $(1-\rho_l)\sigma_l^2$ and take its $(1-\alpha/2)$-th percentile. Then, we replace z_α with $z_{\alpha/2}$ in Eq. 1 and use the union bound to justify that the probability defined in Eq. 1 is at least $1-\alpha$. Note that a similar argument can be used to derive an upper bound for $\sum_{j\in S} w_j U_j$.

[2] You might also think of this procedure as inverting a rank test.

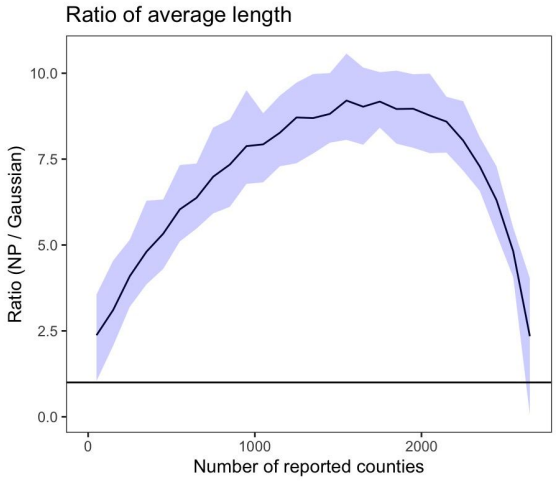
[3] It is certainly not true that U_i and L_i are truly normally distributed. "All models are wrong..."

[4] Recall that $1-\alpha$ marginal coverage means that the intervals cover the true prediction with probability $1-\alpha$ over random calibration *and* unobserved units.

But what is the point of assuming a model that produces a narrow bound if the model does not hold as a representation of the underlying data? In practice, the distribution of the non-conformity scores appears slightly skewed. But, luckily, the bound derived above is robust across many simulated elections.[5] In Figures 2 and 3, we show that this parametric conformalization of the aggregate intervals produces valid predictions intervals that are several times smaller than the nonparametric approach to aggregation that we had described in the previous write-up.



(a) State coverage. Error bars are computed using 50 simulated elections.



(b) Ratio of prediction interval length between nonparametric (NP) general election model and Gaussian Georgia model. Error bars are computed using 50 simulated elections.

Figure 2: State aggregation.

On election night, we will construct multiple parametric models for the non-conformity score that are tailored to the particular aggregation we are interested in. Given a particular state or precinct category (e.g., rural, urban, suburban), our model will use units from the calibration set that are more similar to the units we are aggregating on. This is crucially important when we compute aggregated prediction intervals over sums of units that are more highly correlated than the whole set of precincts might suggest.[6]

3 Acknowledgments

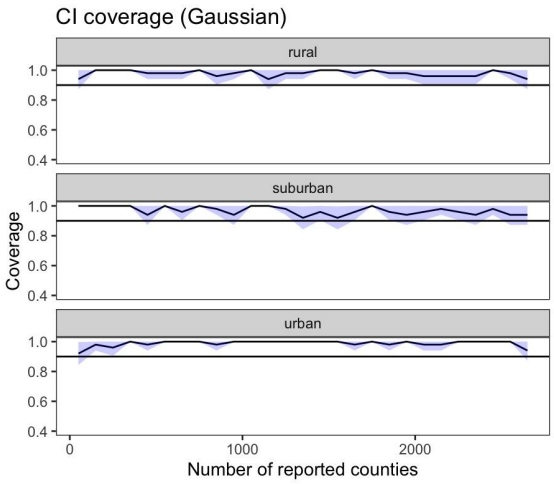
Please reach out with any suggestions or if you spot any mistakes. We'd love to hear from you.

Special thanks to Emmanuel Candès for helpful advice and discussions and Anthony Pesce for providing precinct level data for Georgia.

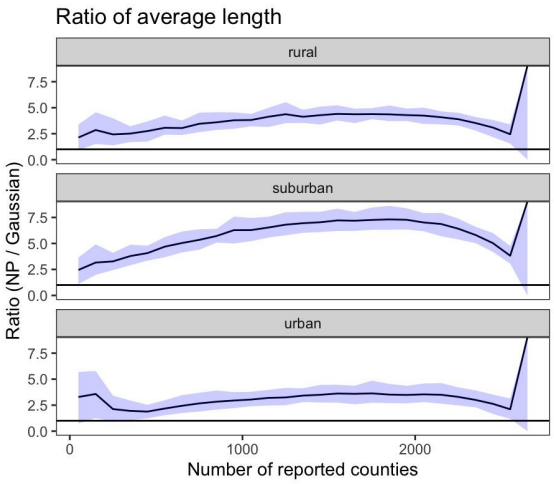
The original model was built in cooperation with Decision Desk HQ/0ptimus Analytics.

[5] Because this procedure amounts to inverting a z-test, this observation is related to the robustness of the two-sample z/t-test, which controls Type I error (analogous to the probability of violating this bound) even for fairly skewed, non-normal data.

[6] See Romano et al. (2019) "With Malice Towards None: Assessing Uncertainty via Equalized Coverage." and Tibshirani et al. (2019) "Conformal Prediction Under Covariate Shift." for related work in nonparametric conformal inference.



(a) Category coverage. Error bars are computed using 50 simulated elections.



(b) Ratio of prediction interval length between nonparametric (NP) general election model and Gaussian Georgia model. Error bars are computed using 50 simulated elections.

Figure 3: Category aggregations.