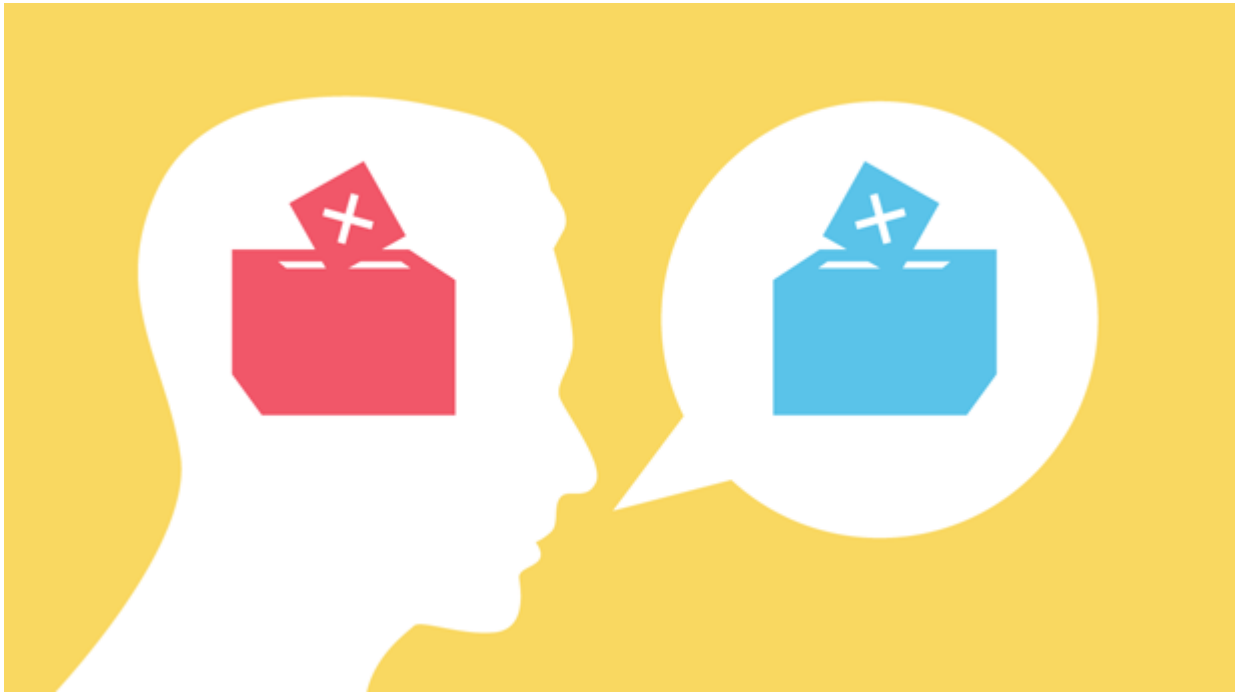
The Devil in the Polling Data

The same problem that caused the 2007 financial crisis also tripped up the polling data ahead of this year's presidential election.

By Pradeep Mutalik



The devil in the data that left election forecasters with egg on their face this week has a familiar name — it's the same villain that tripped up the banks that financed subprime mortgages back in 2007, causing the financial crisis. Its name is "correlated error."

Prediction models can make very accurate forecasts based on many not-so-accurate data points, but they depend on a crucial assumption — that the data points are all independent. In election forecasting, the data points are polls, which are clearly imperfect. Every individual poll has a relatively large margin of error amounting to several percentage points, sometimes favoring one candidate, sometimes the other, all skewed by hundreds of small things — the specific respondents chosen, the means of contact, the phrasing of questions, the representation of voter demographics and so on. These errors can be magically smoothed out by poll aggregation, giving a much more accurate mean polling number — provided the errors in individual polls were all due to different causes, and were therefore independent and uncorrelated. We saw this magic in the accurate predictions made by forecasters like Nate Silver and Sam Wang in the 2012 elections.

Abstractions navigates promising ideas in science and mathematics. Journey with us and join the conversation.

But this year we saw something different: Almost all the swing state polls overscored Clinton's numbers by two to six percent. This error is called "systematic" or "correlated error." Since it affected most or all polls, it was probably caused by some common disrupting factor or factors that were outside the well-established and hitherto reliable poll methodology itself. It was this correlated error that completely threw off the prediction models. Likewise, leading up to the 2007 crisis, financial institutions misjudged the probability of massive subprime loan defaults because they failed to realize that the chances of individual defaults were correlated, not independent.

What could have caused this correlated error to skew all the polls in 2016? That's a subject that pollsters are trying hard to research and pinpoint right now. It will take several months for their findings to be released. But here are two possible speculative causes that can explain this perfectly. I have to give credit for these to Michael Moore, who back in July wrote an amazingly prescient article in which he predicted exactly how Trump would win in excruciating detail. Both of these reasons are ultimately related to the well-documented enthusiasm gap in this election just as it was in the Brexit vote, where a similarly large polling error took place.

1) Emotional voters: All of us are familiar with the situation where our minds incline one way and our hearts tug another. Answering a poll is a boring intellectual exercise, while casting a ballot in the solitude of a voting booth is an empowering, emotional one. It is easy to imagine somewhat conflicted voters who answered "Clinton" to a pollster but in a fit of emotion cast their vote for Trump. If a small, but consistent proportion of Trump voters acted this way, it would have affected all polls and given them all the same correlated error.

2) Depressed voters: Most pollsters try to determine how likely a respondent is to vote and factor that in their final numbers. If there were sizable numbers of Clinton voters who told pollsters that they fully intended to vote but on election day did not find the will or enthusiasm to actually go cast a ballot, that could also explain some of the correlated error. As Moore put it back in August, "If

people could vote from their sofa via their Xbox or remote control, Hillary would win in a landslide."

Other factors like the inability to contact rural voters have been proposed, but it seems to me that good pollsters should have been able to overcome those kinds of problems.

So even the best of the pollsters have a lot to learn. How about the modelers?

I think modelers need to make some changes too.

Consider a hypothetical state that had numbers similar to Michigan this year. The raw polls showed about a 3.5 percent edge for Clinton. I've tried to reverse engineer two simple models with predictions and behavior similar to the FiveThirtyEight and PEC models using the same kinds of tools they used. Imagine that Model 1 predicted a 70 percent probability of Clinton winning and Model 2 predicted a 99 percent probability. Here is how these predictions would have to be modified in the presence of systematic correlated polling error:

With correlated error of:	0%	1%	2%	3%	4%
Probability of Clinton win:					
Model 1	70%	65%	59%	53%	47%
Model 2	99%	95%	84%	63%	37%

The actual correlated error for Michigan turned out to be four percentage points. If Model 1 had known and taken into account this magnitude of correlated error, its prediction of Clinton winning would have changed from 70 percent to just 47 percent, and Model 2's prediction would have changed from 99 percent to 37 percent. Both models would have predicted a Trump win in this hypothetical scenario. What's interesting is how large the swings in the probabilities are with very small changes in the correlated error.

Some readers here defended Nate Silver's forecast, which had the probability of Clinton winning at 71.4 percent, on the grounds that it should not surprise anyone that about a one in three chance materialized. Technically, that is correct. I also agree that Silver's model had some built-in defense against correlated error, while the other models had much less or none. But remember how large the swings in probabilities were in the models above. The modelers knew about the Brexit fiasco, which had a correlated error of four points, in an election with a similar "enthusiasm gap." As I argued last month, it is extremely misleading to state such a potentially fragile probability to one decimal place: It implies that you are confident about the accuracy of the prediction to the precision stated. Most people are not deeply familiar with the technical details of a probability and tend to think of it as a "score" of the race. They are easily misled by the falsely stated precision. As I recommended then, probabilistic election forecasts should be dispensed with altogether and replaced with the seven-point qualitative scale already in wide use. If probabilities have to be stated, they should include a hedging statement that shows how much they would change in the presence of, say, a two or four percent correlated error as "margins of error." If forecasters had done this, the potentially large error swings would have discouraged people from taking the forecasts as gospel truth. It would have saved the entire field of election forecasting from public embarrassment.

Hopefully, further research will identify the causes of correlated polling errors and find ways to detect them, and the modelers will build on the lessons learned from this humbling experience.