Learning in Investment Decisions: Evidence from Prediction Markets and Polls

David S. Lee* Princeton University and NBER

Enrico Moretti UC Berkeley and NBER

December 2008

Corresponding Author: David S. Lee

Session Title: Polls, Votes, and Elections

Session Chair: JOHN DINARDO, University of Michigan

Discussants: MEREILLE JACOBSON, University of California-Irvine
HEATHER ROYER, Case Western Reserve University

* David S. Lee: Industrial Relations Section, Firestone Library, Princeton, NJ 08544-2098, Phone: (609) 258 9548, Fax: (609) 258 2907, Email: davidlee@princeton.edu. Enrico Moretti: Department of Economics, 637 Evans Hall #3880, Berkeley, CA 94720-3880, Phone: (510) 642 6649, Fax: (510) 643 7042, Email: moretti@econ.berkeley.edu.
We thank Wifag Adnan and Fanyin Zheng for research assistance, and Jesse Rothstein for helpful discussions.

Investment decisions are almost always made under uncertainty. Over time, learning occurs, as the flow of information on the costs and benefits of an investment decision reduces its uncertainty (Dixit ad Pindyck, 1994). Political prediction markets offer an ideal setting to test how investors learn, and how learning affects market prices. Prediction markets consist of contracts that pay 1 dollar in the event of a particular candidate winning an election. These contracts are bought and sold in the marketplace, with the idea that those with the best predictive information have incentives to enter this market, and that the equilibrium price will reflect the marginal investor's subjective probability of a particular candidate winning.[1] One appealing feature of prediction markets for tests of investor learning is the availability of high frequency data on popular opinion polls. Unlike for the case of most other markets, opinion polls allow an observer to exactly quantify the flow of information that investors receive over time on the underlying value of the asset. Furthermore, unlike in many other markets, political prediction markets have a fixed date where the true value of the asset is revealed (i.e. election day).

In this paper, we explore how polls and prediction markets interact in the context of the 2008 U.S. Presidential election. We begin by presenting some evidence on the relative predictive power of polls and prediction markers. If almost all of the information that is relevant for predicting electoral outcomes is not captured in polling, then there is little reason to believe that prediction market prices should co-move with contemporaneous polling. If, at the other extreme,

[1] There is a growing literature on prediction markets. See for example See Wolfers and Zitzewitz (2004, 2006, 2007, 2008); Arrow, Forsythe, Gorham, Hahn, Hanson, Ledyard, Levmore, Litan, Milgrom, Nelson, Neumann, Ottaviani, Schelling, Shiller, Smith, Snowberg, Sunstein, Tetlock, Tetlock, Varian, Wolfers and Zitzewitz (2008); Cowgill, Wolfers and Zitzewitz (2008), Wolfers and Zitzewitz (2007); Gurkaynak and Wolfers (2005); Servan-Schreiber, Pennock, Wolfers and Galebach (2004).

there is no useful information beyond what is already summarized by the current polls, then market prices should *react* to new polling information in a particular way. Using both a random walk and a simple autoregressive model, we find that the latter view appears more consistent with the data. Rather than anticipating significant changes in voter sentiment, the market price appears to be reacting to the release of the polling information.

We then outline and test a more formal model of investor learning. In the model, investors have a prior on the probability of victory of each candidate, and in each period they update this probability after receiving a noisy signal in the form of a poll. This Bayesian model indicates that the market price should be a function of the prior and each of the available signals, with weights reflecting their relative precision. It also indicates that more precise polls (i.e. polls with larger sample size) and earlier polls should have more effect on market prices, everything else constant. The empirical evidence is generally, although not completely, supportive of the predictions of the Bayesian model.

I. Background and Data Description

We begin by describing the broad time series patterns found in both the polling and prediction market data. We use polling data available from the web site www.pollster.com, which collects polling numbers on presidential preference from a few dozen polling organizations.[2]

To provide a visual impression of the time series pattern of the polling data, we first expand the dataset so that each observation is at the day-polling organization level (i.e. we assign the polling result of a three-day poll to three separate observations). Then, for each candidate, we

[2] They list 541 separate polls conducted by 46 distinct polling entities ranging from January 2, 2007 to Election Day, 2008. The two organizations with the most number of polls during this period were Rasmussen (91) and Gallup (68). Typically, a survey will ask respondents if the election were held today, which candidate they would support, and responses would be collected over a number of days; a typical sample window is 3 days.

regress the percentage for a candidate on a full set of day dummy variables and dummy variables for the polling organization, where the latter dummy variables are meant to account for a polling organization effect, accounting for time-invariant differences in adjusting for likely turnout, for example. The Obama minus McCain difference in the day coefficients (we choose Gallup as the reference organization) constitutes the "raw data" for our initial exploration of the data. Figure 1 plots a 3-day moving average of this series between June 3, 2008 and Election Day.[3] The figure also plots the Obama minus McCain difference in the price of a contract that pays 1 dollar if the candidate is elected president.

Figure 1 shows a broadly similar time-series pattern in both series, with a rise and then fall between the beginning of June and the end of August, a precipitous decline in the first weeks of September, and then a sharp turnaround and sustained gain for Obama starting in mid-September. For reference, we show a selection of five dates for what might be considered significant events during the course of the campaign.[4] Arguably, prior to each of these events (with the exception of the first) the event or how it would affect the two campaigns was unanticipated or unknown, but immediately after the fact, it was widely believed that the second event helped and the third through fifth events hurt the McCain campaign.

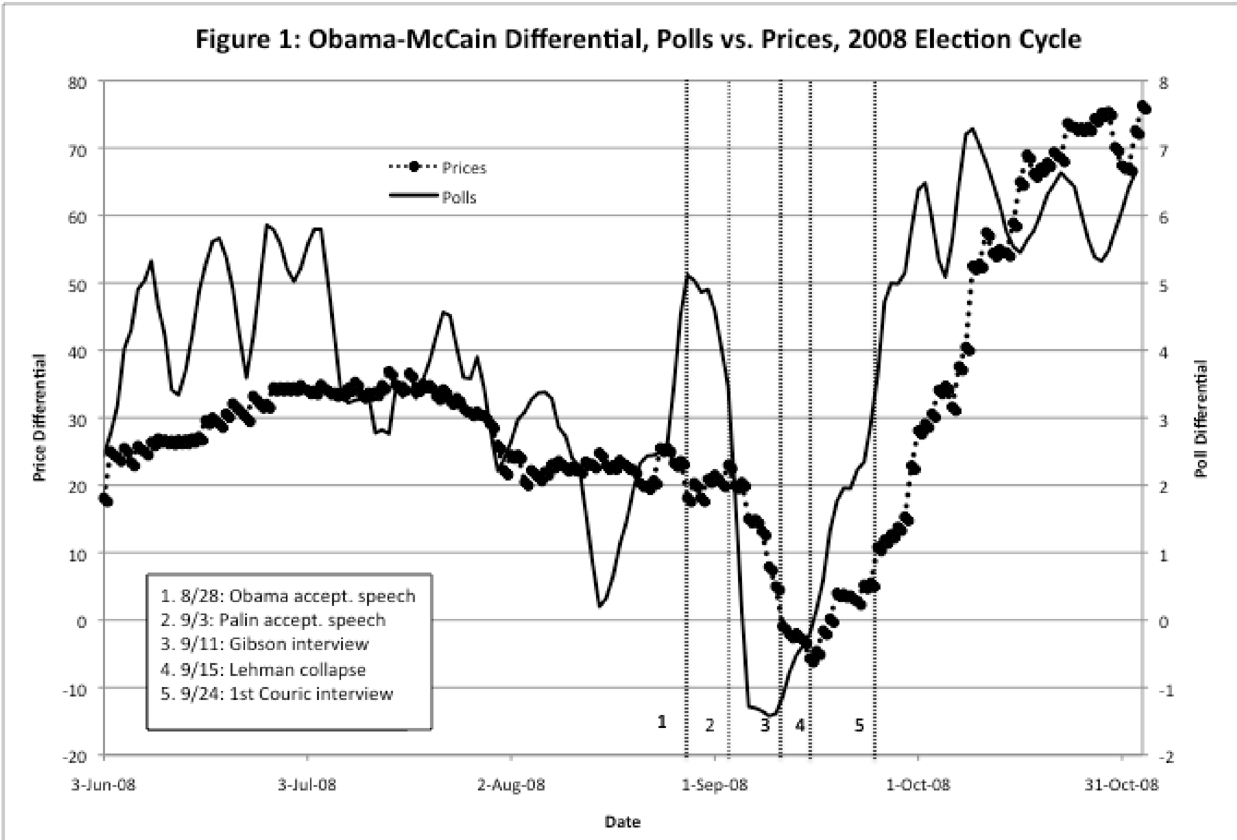
The figure shows no discontinuous drop in response to any of these events nor at any time throughout the June to November period. Indeed, it appears that the price series falls and rises no more quickly than the poll series, and is generally smoother throughout the period. In addition, it appears that the sharp downturn (early September) and rebound (mid-September)

---

[3] June 3, 2008 was the last day of the Democratic Primaries. Around this time, there had emerged a widespread belief that Obama was to be the Democratic nominee. The Iowa Electronic Markets had Obama's nomination trading at around 93.

[4] August 28th (Obama's acceptance of the nomination, followed the next day by the announcement of Sarah Palin as running mate), September 3rd (Palin's acceptance speech followed the next day by McCain's), September 11th (Charles Gibson's Palin interview, which was viewed as unfavorable towards Palin), September 15th (the collapse of Lehman Brothers, and soon after, McCain's statement that the fundamentals of the economy were strong), and September 24th (Katie Couric's 1st Palin interview, which also received unfavorable reviews).

occur in the polls several days *before* a similar pattern emerges in the price data. That is, rather than *anticipating* these significant turnarounds, the price series appears to be *reacting* to the

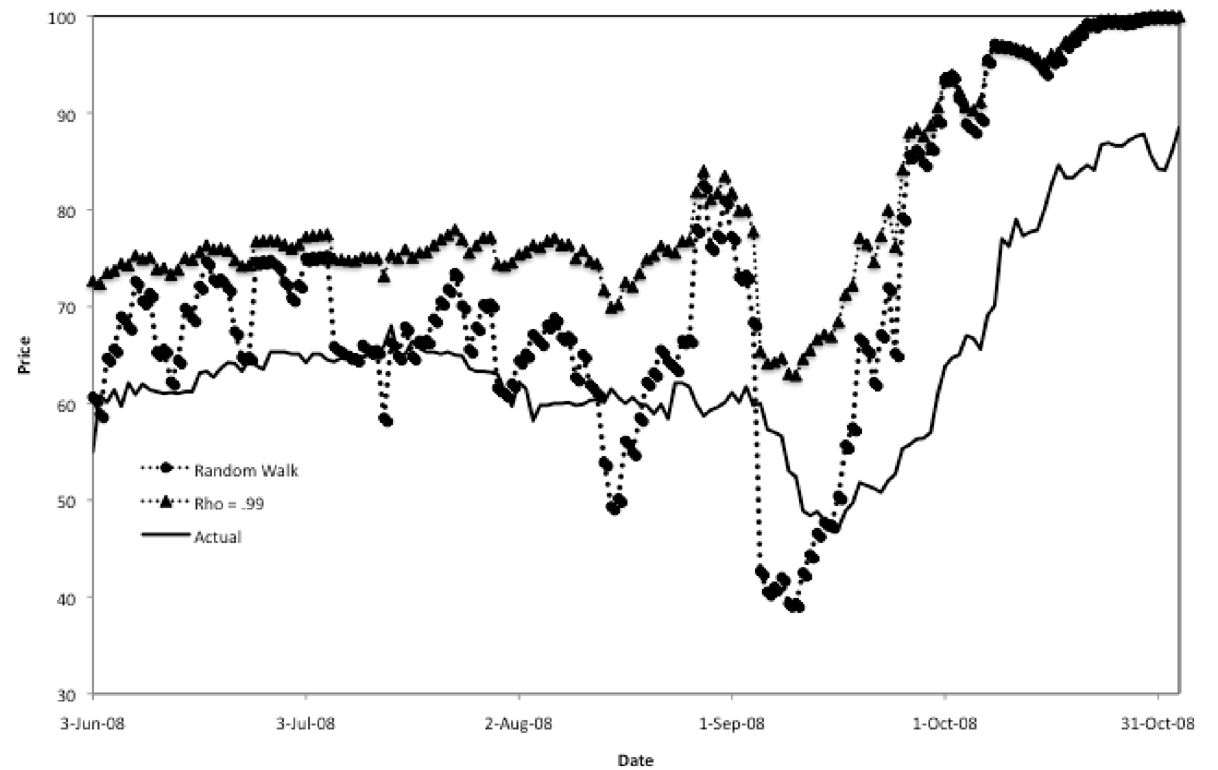


release of the polling information.

Before proceeding to our main analysis, we provide a somewhat crude way of exploring this notion. Suppose it was believed that the Obama-McCain polling difference y_t was a random walk process with error term e_t distributed as a normal with variance σ^2. Then at time t the forecast of y_T is simply y_t, today's current poll. That forecast has a standard deviation $v = ((T-t)\sigma^2)^{1/2}$; as the date of the election nears, the contemporaneous poll becomes an increasingly better predictor of the actual Obama-McCain difference on Election Day. We take the forecasted probability of an Obama victory at time t to be $Pr[y_T > 0 \mid y_t] = Pr[y_t + e_{t+1} + \ldots + e_T > 0 \mid y_t] =$

$\Phi(y_t / v)$.[5] This resulting probability could be viewed as the reservation price for someone who

Figure 2: Obama Victory Market Prices, Actual vs. Predicted



adopted the random walk model.

The solid circles in Figure 2 represent this predicted price series.[6] This predicted price follows the same general pattern as the actual price of an Obama victory (solid line), but in comparison it is much more variable over time. Indeed, in this perspective, one would expect the price to be sensitive to daily fluctuations in the polls.

An alternative approach is to depart slightly from the random walk model, and assume that the poll difference follows $y_t = \alpha + u_t$, where u_t is an AR(1) with autoregressive parameter $\rho = 0.99$. In this model, new innovations to the polling differences do persist, but eventually die

[5] This must necessarily be viewed as an illustrative approximation, since the polls give forecasts of the popular vote as opposed to the Electoral College count, which determines the Presidency.

[6] We estimate σ^2 using first differences in the polling difference, but treat it as known in producing the forecasted probability.

out over time. Producing forecasts and forecast standard errors, analogously to the random walk case, we illustrate the resulting predicted price series as the solid triangles in Figure 2. In comparison to the random walk case, predicted prices in the first half of the period evolve in a much smoother way, owing to the fact that when the Election Day is very far away, the forecast is closer to the overall mean, and new innovations to the series play little role in the forecasted probability of winning. As the election nears, the forecast is closer to the contemporaneous value of y_t, and new innovations play a larger role.

Overall, while the prices predicted by this autoregressive model do not exactly match, the time pattern is quite similar to the actual price series. The patterns we see in the data, at least in this presidential election cycle, are suggestive that rather than *anticipating* how the polling numbers might react to current events, prices are using changes in polls to update their forecast of the final election outcome.

We consider it instructive to explore the empirical relevance of the extreme version of this view: that all relevant new information for predicting the outcome of the election is entirely contained by polling information.

II. New Information, Learning and Bayesian Updating

In this Section we present a simple model of investor learning, and some empirical evidence on the predictions of the model. We begin in sub-section A by proposing a version of the Normal Learning Model that is useful to characterize how investors seeking to determine the probability of victory of two opposing candidates may incorporate the flow of information that they receive from public opinion polls. Sub-section B presents three empirical tests of the predictions of the model.

A. Conceptual Framework

Suppose that there are only two candidates running for president. Let Δ be the true difference between the number of popular votes for the Democratic candidate and the Republican candidate. Before the election, Δ is unknown and is assumed to be normal with mean δ and precision *h*:

(1) $$\Delta \sim N(\delta, 1/h)$$

We assume that one poll becomes exogenously available in each period *t*, (*t*=1, 2, … ,T; where T is election day). Investors receive the following noisy signal from the public opinion polls:

(2) $$Z_t = \Delta + e_t$$

where we assume that e_t is a normally distributed, unbiased, idiosyncratic noise with precision k_t

(3) $$e_t \sim N(0, 1/k_t)$$

Since the polls have finite sample, they contain noise. The variable e_t represents this small-sample noise, and we assume that its variance depends on the poll sample size, N_t. Specifically, the precision of a given poll is a function of the square root of its sample size: $k_{ts} = (N_{ts})^{1/2}$, so that larger polls have larger precision.[7]

In period 1, only one poll, Z_1, is available. It is possible to show that the conditional distribution of the difference between the number of votes for the Democratic candidate and the Republican candidate given the signal is $f(\Delta \mid Z_1)$ is the normal $N(m_1, v_1)$, where the expected conditional difference between the number of votes for the Democratic candidate and the Republican candidate and the conditional variance are, respectively

(4) $$m_1 \equiv E_1[\Delta \mid Z_1] = w_1 \delta + (1-w_1) Z_1$$

$$v_1 \equiv Var[\Delta \mid Z_1] = 1/(k_1+h).$$

[7] A census of the entire voting population, would have infinite precision ($k_{ts} = \infty$) and would therefore reveal the true Δ_s.

and $w_1 = h/(k_1+h)$. The market price in period t=1 of a prediction market security that pays \$1 in case of victory of the Democratic candidate, P_1, is equal to the conditional probability calculated in period 1 that the democratic candidate will win the election. Given equation 4, this conditional probability is

$$(5) \qquad P_1 \equiv \text{Prob}(\Delta > 0 \mid Z_1) = 1 - F[\, - [w_1\, \delta + (1-w_1)\, Z_1\,]\, (k_1+h)^{1/2}\,]$$

where F[ ] is the standard normal cumulative distribution function.

Equation 5 indicates that the price for the Democratic candidate security is a function of the weighted average of the prior, δ, and the signal, Z_1. The weight w_1 reflects the relative precision of the prior relative to the signal. A more precise poll generates a more precise signal (large k_1), and therefore more weight is put on the signal and less on the prior (small w). When the signal is less precise, more weight is put on the prior (large w). In the extreme, if the signal had infinite precision (as in the case of a poll of infinite sample size), all the weight would be on the signal, and the prior would receive no weight.

A second interesting results is that not only a marginal shift in the public opinion polls in favor of the democratic candidate will result in an increase in the market price, but such price increase will be larger the more precise the signal (i.e. the larger the poll sample size):

$$(6) \qquad d\, P_1 / d\, Z_1\, d\, k_1 = f(-m_1/(v_1^{1/2}))\ [h / (k_1+h)^{3/2}] > 0$$

This makes intuitive sense: a larger poll contains more information on the true electoral gap between the democrat and republican candidates and therefore generates a more precise signal. A more precise signal shifts the market price more than a less precise signal, everything else constant, because it leads to more updating.

In each subsequent period, a new poll becomes available. Iterating the Normal Learning Model, it is possible to show that after *t* periods, the conditional distribution of the difference between the number of votes for the Democratic candidate and the Republican candidate given the prior and the t signals $Z_1, Z_2, \ldots, Z_t$, $f(\Delta_s \mid Z_{1s}, Z_{2s}, \ldots, Z_{ts})$, is the normal $N(m_t, v_t)$ where

(7) $$m_t \equiv E_t[\Delta \mid Z_1, Z_2, \ldots, Z_t] = [h/(\sum_j^t k_j + h)]\ \delta + \sum_j^t [k_j/(\sum_j^t k_j + h)]\ Z_j$$

and

(8) $$v_t = \sum_j^t k_j + h.$$

As before, the conditional mean m_t is a weighted average of prior and each of the t signals $Z_1, Z_2, \ldots, Z_t$ with weights reflecting each element relative precision. The market price at time *t* is therefore

(9) $$P_t \equiv 1 - F[-\{[h/(\sum_j^t k_j + h)]\ \delta + \sum_j^t [k_j/(\sum_j^t k_j + h)]\ Z_j\} / \{\sum_j^t k_j + h\}^{1/2}]$$

In this setting, uncertainty declines in each period. Equation (9) indicates that the marginal amount of information provided by each subsequent poll is smaller and smaller so that the effect of a poll on market price declines over time:

(10) $$d\,P_t / d\,Z_t\, d\,t < 0$$

Consider two identical signals (equal realization and equal precision). If one occurs at time 1 and the other one at time 10, the former will move the market price more than the latter. The intuition is that in period 10 more is already known about Δ, and therefore the marginal effect on an additional piece of information is smaller. In the limit, after infinite number of periods, the true Δ is revealed.

B. Empirical Evidence

We now test whether this simple model of Bayesian updating is generally consistent with some broad features of the data. We use data on the price of Obama and McCain winner-take-all securities from Intrade matched with data on polls. We use data for the period from January 1st 2008 to the day before the elections, aggregated at the weekly level.[8]

Table 1

	(1)	(2)	(3)	(4)
Polls_t	4.901	2.055	.011	4.902
	(1.292)	(.782)	(.559)	(.991)
$\text{Polls}_t * N_t$		4.941	.495	
		(1.059)	(.179)	
Polls_t *(Weeks to Elections)$_t$				-.137
				(.028)
Polls_{t-1}			2.437	3.544
			(.913)	(.766)
$\text{Polls}_{t-1} * N_{t-1}$			.345	
			(.158)	
Test: Coeff on Polls_t =0 and Coeff on $\text{Polls}_t * N_t$ =0 (p-value)			4.22	
			(.022)	
Test: Coeff on Polls_{t-1} =0 and Coeff on $\text{Polls}_{t-1} * N_{t-1}$ =0 (p-value)			15.14	
			(.000)	
Test: Coeff on Polls_t =0 and Coeff on Polls_t * Weeks to Elections_t =0 (p-value)				26.87
				(.000)

Notes: Newey-West standard errors in parenthesis. The error structure is assumed to be heteroskedastic and autocorrelated up to 3 lags. The dependent variable is the difference in the price of an Obama and McCain winner-take-all security from Tradesport (scale 0 to 100). "Polls_t" is the average difference in polls between Obama and McCains in week t (scale 0 to 100). "N_t" is the aggregate sample size of polls in week t (in thousands of respondents). "(Weeks to elections) $_t$" is the number of weeks left before election day. Sample size is 45 weeks.

In column 1 of Table 1, we report the estimate of a regression where the dependent variable is the difference in the price of an Obama victory security and a McCain victory security in a given week and the independent variable is the average difference in polls between Obama and McCain for that week. Given the time series structure of the data, we report Newey-West standard errors in parenthesis, where we assume the error structure to be heteroskedastic and autocorrelated up to 3 lags.

[8] Our data report the date when a poll started and the date when a poll ended, but not the exact date when it was released. To match polls to prices, we assume that it takes 2 days after a poll is completed to be released.

The coefficient in column 1 indicates that a 1 percentage point increase in the relative support for Obama is associated with a 4.9 cents increase in the relative price of the Obama victory security. Note that this coefficient is not expected to be equal to 1 for two reasons. First, the dependent variable is the price of a winner-take-all security, not a security that reflects popular vote. The relationship between popular vote as measured by the polls and probability of winning is non-linear. A linear regression is necessarily an approximation that holds only for marginal changes in polls.[9] Second, if the Normal Learning Model is correct, the model in column 1 omits an important variable, and is therefore likely to be biased. Equations 5 and 9 indicate that the market price in any given period should be a function not just of the polls in *that period*, but also of the polls in the *previous periods* (as well as the prior).

In column 2, we report the estimate of a model that includes both the difference in polls between Obama and McCain in week *t*, and the average difference in polls between Obama and McCain in weeks *t-1* and *t-2*. Consistent with the simple Bayesian updating model in sub-section A, both current and lagged polls affect current market price. Models that include average difference in polls between Obama and McCain in weeks *t-1*, *t-2*, *t-3* or weeks *t-1*, *t-2*, *t-3* and *t-4* yield similar estimates.

A second testable prediction of the Bayesian model is that polls with larger sample size should affect prices more than polls with smaller sample size, holding constant the poll outcome (equation 6). In other words, take two polls that predict the same a margin of victory for the Democratic candidate. If the first poll has larger sample size than the second poll, the first poll should result in a larger price increase than the second poll. Consistent with this prediction, column 3 indicates that the interaction between poll outcome and poll size, N_t, (measured in

[9] The Iowa Electronic Market provides securities based on the popular vote. However, these securities start trading only when the candidates receive their party nomination, which is about three months before the elections. The sample size is therefore too small to be useful.

thousands of respondents) is positive. This is true both for the current poll and for the lagged polls. At the bottom of the table we show that the main effect of polls and its interaction with sample size are jointly statistically significant. Since the average aggregate poll size in a week is 2,518 the effect of a one percentage point increase in relative current polls raises market price on average by 1.257 percentage points (1.257 = .011 + .495*2.518).

A third prediction of the model is that uncertainty should decline over time, so that the marginal effect on prices of the latest signal should becomes smaller over time (equation 10). In column 4, we include the interaction of current polls outcomes and the number of weeks left before the election. Contrary to the prediction of the model, the coefficient is negative, suggesting that polls closer to the election have more of an impact on prices that earlier polls.

References

Dixit, Anivash and Robert Pindyck, "Investment Under Uncertainty", Princeton University Press (1994).

Arrow, Forsythe, Gorham, Hahn, Hanson, Ledyard, Levmore, Litan, Milgrom, Nelson, Neumann, Ottaviani, Schelling, Shiller, Smith, Snowberg, Sunstein, Tetlock, Tetlock, Varian, Wolfers and Zitzewitz "The Promise of Prediction Markets" *Science*, 320, p.877, May 16 2008.

Bo Cowgill, Justin Wolfers and Eric Zitzewitz Using Prediction Markets to Track Information Flows: Evidence from Google (2008).

Justin Wolfers and Eric Zitzewitz Prediction Markets in Theory and Practice, The New Palgrave Dictionary of Economics, 2nd ed. Larry Blume and Steve Durlauf (eds.)

Refet Gurkaynak and Justin Wolfers Macroeconomic Derivatives: An Initial Analysis of Market-Based Macro Forecasts, Uncertainty and Risk NBER International Seminar on Macroeconomics, 2005.

Justin Wolfers and Eric Zitzewitz Interpreting Prediction Market Prices as Probabilities NBER Working Paper #12200

Justin Wolfers and Eric Zitzewitz Five Open Questions About Prediction Markets, Information Markets: A New Way of Making Decisions in the Public and Private Sectors, AEI-Brookings Press, eds: Robert Hahn and Paul Tetlock.

Justin Wolfers and Eric Zitzewitz Prediction Markets Journal of Economic Perspectives, 18(2), Spring 2004.

Emile Servan-Schreiber, David Pennock, Justin Wolfers and Brian Galebach Prediction Markets: Does Money Matter? Electronic Markets, 14(3), September 2004.