USING PRIMARY ELECTIONS TO CONTROL FOR SELECTION OF U.S. CONGRESSIONAL CANDIDATES

KYLE KRETSCHMAN

Abstract. Primary elections provide an explicit selection criterion for general election candidates. This paper estimates how candidate characteristics determine U.S. House of Representative general election vote shares using data from the primary elections to control for selection bias. I quantify the impact of these factors by developing a two stage discrete choice model. The first stage models a primary election as a political party’s decision of which candidate to nominate for the general election. The second stage represents the general election and uses individual voter maximization to derive the predicted vote share for each candidate. The two stages are econometrically linked by the correlation of the unobserved candidate characteristics in each stage.

1. Introduction

Primary elections are an integral part of the election process in the United States. The primary elections whittle down all the potential candidates to only the most serious, and hopefully, qualified candidates. In many U.S. elections, the primary can be more important

Date: August 19, 2010.
email: kyle.kretschman@mail.utexas.edu.

I would like to thank Eugenio Miravete for the idea for this paper. I would also like to thank Richard Dusansky, Jason Abrevaya, and Brian Roberts for all of their comments and support. All remaining errors are my own.

1
than the general election because the electorate greatly favors one party over the other. It seems obviously important to factor in the primary elections when you attempt to estimate the determinants of general elections.

The majority of the empirical election literature has ignored primary elections when estimating functions that predict voting outcomes. For a long time, the empirical literature focused on reduced form estimates to quantify the impact of candidate characteristics such as expenditures and incumbency on the general election vote share. A selection of these are Jacobsen (1978), Green & Krasno (1988), Levitt (1994), and Gerber (1998) (see the introduction of Gerber (2004) for a thorough summary of the reduced form literature). All of this work has focused on the general election results and have not incorporated primary elections. The prior literature that models primary elections and general elections has mainly been theoretical work that focuses on the individual candidates choosing their platform. See Owen & Grofman (2006), Callander & Wilkie (2007), Adams & Merrill (2008), and Agranov (2009).

So far, these models struggle to produce estimable equations and link the theory to observed data since the main decision in the models is econometrically unobservable. Recently, empirical election literature has started to use discrete choice methods to provide empirical results that are based on individual optimization, Rekkas (2007) and Kretschman & Mastronardi

\footnote{For clarification, the terms “percentage of the vote” and “vote share” have been used interchangeably in the past. However, in this paper, vote share will refer to the number of votes a candidate receives divided by the number of registered voters. This is not equal to the percentage of votes received, which equals the number of votes received divided by the total number of votes cast. The difference arises because this model allows for a registered voter to choose not to vote.}
This work is based on adapting the empirical discrete choice methods to model elections as the result of utility maximizing behavior from voters. This type of model moves the empirical election literature away from estimating reduced form correlations and instead estimates preference parameters for the voters. Both papers use the seminal work of Berry (1994) and Berry, Levinsohn and Pakes (1995) (hereafter BLP) to estimate the impact of candidate characteristics on vote shares.

The purpose of this paper is to develop a method to incorporate primary election data to control for the selection bias in general election estimates. The candidates in a general election are explicitly selected for the election by primary elections that occur earlier in the year. Therefore, the population of general election candidates is a selected sample. This paper develops a method to deal with this issue. An election cycle broken down into two stages with a political party choosing the optimal candidate in the first stage and registered voters voting for the utility maximizing candidate in the second stage. The two stages are linked in two ways. By backward induction, the political party’s objective will be to choose the candidate that has the best chance of winning the general election. The second link arises from the fact that both the political party and the individual voters have a valuation of the unobserved candidate characteristics. Though these two valuations are different, they are correlated and the correlation is a parameter to be estimated in the model. This method allows the general election estimates to test and control for the existence of selection bias.

This paper contributes to the election literature in many ways. Similar to Rekkas (2007) and Kretschman & Mastronardi (2010), instead of postulating reduced form relationships
between vote shares and candidate characteristics, the estimable equations are derived from individual agents maximizing an objective function. Therefore, the estimates from this method are preference parameters for the agents making the decisions rather than correlation coefficients. This paper advances beyond these two papers by developing a method that shows a direct link between a theory of how candidates are selected in the primary and how this affects the general election vote shares. Neither of these papers dealt with how candidates in the general election were selected.²

This method is an advancement beyond the two stage theoretical work because it provides a theoretical model that can directly be estimate with the available data. The only reliable outcome variable in primary elections is who is chosen as the winner. The participation rates are incredibly low in primary elections and information on the size of the electorate for each election is very fuzzy. This model only needs to know who won the primary election. Instead of postulating how candidates behave, the main assumption is that the political party’s pick the candidate who provides the best chance of winning the general election. While this method does abstract away from why a candidate wins a primary, the method uses the available information to identify the the influence of primary elections on general election results.

The model is estimated using the 2006 U.S. House of Representative election data and a unique supplement of primary election expenditures. While producing similar results to previous work, the preliminary results do show evidence of selection bias. The estimation

²This might not be an important issue in Rekkas (2007) because she analyzes Canadian elections.
produces a positive, significant correlation between the political party’s unobserved valuation in the primary election and the voter’s valuation of the candidate’s unobserved characteristics. This means that there exists positive selection in the primaries. This provides empirical evidence pick the candidate that has the best chance of winning the general election.

The organization for the remainder of this paper is as follows. Section 2 presents the two stage discrete choice model for candidate selection and voting. Section 3 describes the multistep estimation method for recovering the preference parameters of the model while controlling for the endogeneity of expenditures. Section 4 explains the unique data set gathered for this analysis. Section 5 provides and interprets the empirical results. Section 6 concludes.

2. TWO STAGE DISCRETE CHOICE MODEL

This paper models a U.S Congressional election cycle as a two stage game. The first stage represents the primary election and the decision makers are the political parties. In this stage, each political party nominates a candidate to run in the general election. The party has to choose from a discrete set of candidates for its nomination. The objective is to pick the candidate that has the highest chance of winning the general election.

The second stage represents the general election. In this stage, the decision makers are registered voters. Each voter in the congressional district has to make the discrete choice between the candidates in the general election. Even though the voting decision is not
observed at the individual level, a predicted vote share share function for each candidate is
derived from this representative agent model similar to BLP.

The stages of this model are connected in two ways. Through backward induction, the
political parties objective is to choose the candidate with the highest expected vote share in
the second stage. The political parties are uncertain about the general election vote share
for each primary candidate since they do not know how much more money the candidate will
spend before the general election. Also, the party’s valuation of the unobserved characteristics
of the candidate will not be the same as the valuation by a registered voter in a general
election. However, the unobservable valuations should be correlated and this correlation
links the two stages theoretically and econometrically.

2.1. 2nd Stage: General Elections. The standard way to solve a discrete time game is
through backward induction. Therefore, the second stage is presented first to show how the
results are used as the objective function for the first stage.

2.1.1. Utility of a Registered Voter. Assume that there exists a representative utility-maximizing
agent in each congressional district.\(^3\) This agent is a registered voter and is facing a discrete
voting decision: vote for the Democratic Party candidate, vote for the Republican Party
candidate, or not vote for one of the two candidates from the dominant parties. The third
option is the outside option that includes the choice of voting for a third party candidate or

\(^3\)There is obviously many dimensions of voter heterogeneity such as the distribution of income within each
congressional district. I ignore voter heterogeneity, i.e random coefficients, at this to keep the model simple.
Including random coefficients is a straightforward extension the will be done at a later date.
abstaining from the election. I define the outside option this way because every representative to the U.S. House has been a member of the Democratic or Republican party during past decade. Assume each agent wants to maximize his utility from voting.

\[
\max_{j \in \{D, R, \text{NotVote}\}} U_{mnj} = V(X_{mj}) + e_{mnj}
\]

The Democratic Party is denoted by the subscript D, the Republican Party denoted by the subscript R. For this decision problem, the \(m\) subscript denotes the market which is defined as a congressional district, the \(n\) subscript denotes the representative agent and the \(j\) subscript denotes the different voting options: Democratic, Republican, and Not Vote. For example, \(U_{mnR}\) is the utility that a registered voter in congressional district \(m\) receives from voting for the Republican candidate. It is important to note that this utility function is defined as the utility of voting for a specific candidate and not the utility from having that candidate be your representative. Therefore, \(V(\cdot)\) is the indirect utility function from the vector of observable characteristics, \(X_{mj}\), and \(e_{mnj}\) is the unobserved utility that equates \(V(\cdot)\) and the actual utility of each individual. I assume \(V(\cdot)\) is linear in its arguments.

In this stage, there are two sources of unobserved utility. One is the utility from the candidate characteristics that are observed by the agent but unobserved by the econometrician, denoted by \(\xi_{mj}\). This can be thought of as the unobserved candidate quality that differs for each candidate in each market. The second source of unobserved utility is the idiosyncratic utility shock, denoted by \(e_{mnj}\). This shock represents the unobserved “taste” for each
candidate. Therefore, $\epsilon_{mnj} \equiv \xi_{mj} + \epsilon_{mnj}$. With these assumptions, rewrite equation (1) as:

$$\max_{j \in \{D, R, NotVote\}} U_{mnj} = \beta X_{mj} + (\xi_{mj} + \epsilon_{mnj})$$

2.1.2. Predicted Vote Share Function. While $U_{mnj}$ is not observable, using the fact that the agent is a utility maximizer and assuming a distribution on $\epsilon_{mnj}$ allows for the derivation of the probability of voting for each of the candidates. Define the mean utility level for each candidate in each market as:

$$\delta_{mj} \equiv \beta X_{mj} + \xi_{mj}$$

(3)

$\therefore$ Equation 2 is simplified to $U_{mnj} = \delta_{mj} + \epsilon_{mnj}$

(4)

$\delta_{mj}$ collects all the candidate characteristics and coefficients within a district. Simplifying the utility this way shows that the mean utility level of each candidate, $\delta_{mj}$ and the taste shock completely determine the voting choice for each agent, therefore, the probability of each agent voting for each candidate. From here on, I drop the $m$ subscript to make the equations more concise.

A discrete choice model can only identify relative utility levels. I redefine $\epsilon_{nj}$ as the utility shock difference with the outside option shock, $(\epsilon_{nj} - \epsilon_{nNotVote})$, and normalize the mean utility of the outside option to zero, $(\delta_{nNotVote} \equiv 0)$. Therefore, the probability of voting for candidate $j$ can be found as:

$$P_{nj} = \int_{B_{nj}} dP(\epsilon_{nj}, \epsilon_{nk}) \text{ where } B_{nj} = \{\epsilon_{nj}, \epsilon_{nk} | U_{nj} > U_{nk} \forall k \neq j\}$$

(5)

$\therefore B_{nj} = \{(\epsilon_{nj}, \epsilon_{nk}) | \delta_{j} + \epsilon_{nj} > 0; \delta_{j} + \epsilon_{nj} > \delta_{k} + \epsilon_{nk}\}$

(6)
Assume that all $\epsilon_{nj}$ are independently distributed type 1 extreme value. This assumption leads to the well known result that the predicted probability voting for candidate $j$ has the logit form.\(^4\)

\[
P_j = \frac{e^{\delta_j}}{1 + \sum_{j' \in \{D,R\}} e^{\delta_{j'}}} \quad \forall j \in \{D, R\}
\]

By the law of large numbers, the predicted vote share for each candidate is equal to the probability that the representative agent votes for candidate $j$.

\[
\hat{s}_j(\beta, \xi; X) = \frac{e^{\delta_j}}{1 + \sum_{j' \in \{D,R\}} e^{\delta_{j'}}} \quad \forall j \in \{D, R\}
\]

Without any agent heterogeneity, there is a closed form solution for $\delta_j$ if we equate the predicted vote share function with the observed vote share for each candidate. Define $s_j$ as the observed vote share of candidate $j$ and $s_N$ is the vote share of the outside option. Some straightforward algebra produces:

\[
\ln(s_j) - \ln(s_N) = \delta_j
\]

\[
\therefore, \ln(s_j) - \ln(s_N) = \beta X_j + \xi_j
\]

\(^4\)See Train (2003) for the algebra.
2.1.3. Endogeneity of Candidate Expenditures. \( \xi_j \) is defined as all the characteristics of the candidate that are unobserved by the econometrician that affect the mean utility level of the candidate. Some of these characteristics are charisma, public speaking ability, physical appearance and an incumbent’s performance while in office. These characteristics not only affect the number of votes a candidate receives, but also affect the amount of contributions a candidate raises. A candidate with greater contributions can expend more on the congressional race. Therefore, the unobserved characteristics are correlated with the candidate’s expenditures and the estimation procedure will account for the endogeneity of candidate campaign expenditures by using 2SLS since this stage produces a linear estimation equation.

2.2. 1st Stage: Primary Elections.

2.2.1. Political Party’s Discrete Choice. By backward induction, the political parties know that to have the highest chance of winning the election, they must choose the candidate that will have highest value of \( \delta_j \). However, the parties do not know \( \delta_j \) for certain. The primary elections take place before the general election so the observable characteristics are different for the primaries. Also, the parties cannot perfectly predict the valuation of the unobservable characteristics in the general election. The parties have there own valuation that is imperfectly correlated with \( \xi_j \). Therefore, the decision problem for each party, \( p \), is to choose the candidate with the highest expected \( \delta_j \) given the information available at the time of the primary and the parties own valuation of the candidates unobservable characteristics.
The party also has the outside option of choosing no candidate to nominate for the general election.

\[(11) \max_{c \in \{1, \ldots, C_p, NoC\}} E(\delta_j | X_c, \psi_c) \]

The subscript \(c\) denotes the different candidates in the primary election and \(C_p\) is the total number of candidates in each election which varies by party and congressional district. Let \(X_c\) be the observable candidate characteristics at the time of the primary election and \(\psi_c\) be the party valuation of the unobservable candidate characteristics. By equation 4 the maximization problem becomes:

\[(12) \max_{c \in \{1, \ldots, C_p, NoC\}} E(\beta X_j + \xi_j | X_c, \psi_c) \]

Some of the information in \(X_j\) is not available at the time of the primary. Specifically, the political party can only see how much money the candidate has spent up until the primary. Also, the political party will have different preference parameters for the candidate characteristics. Therefore, rewrite the maximization problem as

\[(13) \max_{c \in \{1, \ldots, C_p, NoC\}} \beta X_c + E(\xi_j | X_c, \psi_c) \]

I now decompose \(\xi_j\) into the linear projection, \(\xi_j = \gamma \psi_c + \tilde{\xi}_j\), where \(E(\tilde{\xi}_j | \psi_c) = 0\) by the definition of a linear projections. This is the same as assuming a linear correlation between
At this point, it is important to address the endogeneity of the candidate expenditures in the primary since this stage does not produce a linear model. The preprimary expenditures are endogenous with $\psi_c$ for all the same reasons that the general election expenditures are endogenous to $\xi_j$. In addition to those reasons, the party’s unobserved valuation of the candidate includes the expectation of how much the candidate will spend in the general election. To deal with the endogeneity, decompose the unobserved party valuation into the part that is correlated with preprimary expenditures and the part that is not, $\psi_c = \rho v_c + \hat{\psi}_c$ where $v_c$ is the residual of the reduced form regression of the preprimary expenditures on a set of exogenous instruments. The allows for control function approach from Petrin & Train (2010) to control for endogeneity in this stage. Finally, for identification purposes, rescale the optimization problem by dividing by the coefficient $\gamma$ so that the optimization problem becomes

$$\max_{c \in \{1, \ldots, C_p, NoC\}} \frac{\beta}{\gamma} \hat{X}_c + \rho v_c + \hat{\psi}_c$$

The solution to this optimization problem is to pick the candidate with the highest objective value. This solution allows for the derivation of the probability that each candidate is nominated to the general election, $P_c$. Since this stage cannot identify both $\beta$ and $\gamma$ and for conciseness, define $\tilde{\beta} \hat{X}_c \equiv \frac{\beta}{\gamma} \hat{X}_c + \rho v_c$. Again, similar to the first stage, normalize the
the mean utility of the outside option to 0, $\tilde{\beta} \tilde{X}_{N\alpha C} = 0$. Then, define $\psi^*_c$ as the difference between the valuation for candidate $c$ and the outside option, $\psi^*_c = \tilde{\psi}_c - \tilde{\psi}_{N\alpha C}$.

\begin{align}
P_c &= \text{Prob}(\tilde{\beta} \tilde{X}_c + \psi^*_c > 0, \tilde{\beta} \tilde{X}_c + \psi^*_c > \tilde{\beta} \tilde{X}_k + \psi^*_k \quad \forall k \neq c) \\
&= \text{Prob}(\psi^*_c > -\tilde{\beta} \tilde{X}_c, \psi^*_c + \tilde{\beta} \tilde{X}_c - \tilde{\beta} \tilde{X}_k > \psi^*_k \quad \forall k \neq c)
\end{align}

Unlike the first stage, I must make a distributional assumption about the unobserved candidate characteristics $\tilde{\psi}_c$. Assume that all the $\tilde{\psi}_c$ are independently distributed type 1 extreme value. This leads to the standard multinomial logit result.

\begin{equation}
P_c = \frac{e^{\tilde{\beta} \tilde{X}_c}}{1 + \sum_{c' \in \{1, \ldots, C_p\}} e^{\tilde{\beta} \tilde{X}_{c'}}} \quad \forall c \in \{1, \ldots, C_p\}
\end{equation}

with $\tilde{\beta} \tilde{X}_c \equiv \frac{\tilde{\beta}}{\gamma} \tilde{X}_c + \rho v$

3. Estimation

3.1. 1st Stage Estimation. The first stage of the model is estimated using the control function method of Petrin & Train (2010) to control for endogeneity of expenditures. The method is a two step procedure where the first step regresses the endogenous variable on a set of instruments and the exogenous regressors. Then, the residual for each observation, $\tilde{v}_c$ is used as an additional regressor in a maximum likelihood estimation.

The likelihood function for this stage is defined as:
\[
\ell = \prod_{m=1}^{M} \prod_{p=1}^{2} \prod_{c=1}^{C_{mp}} P_{c}^{i(c)}
\]

with
\[
P_{c} = \frac{e^{\tilde{\beta} \tilde{X}_{c}}}{1 + \sum_{c' \in \{1, \ldots, C_{mp}\}} e^{\tilde{\beta} \tilde{X}_{c'}}}
\]
and
\[
\tilde{\beta} \tilde{X}_{c} \equiv \frac{\hat{\gamma}}{\gamma} \tilde{X}_{c} + \rho \tilde{v}_{c}
\]

where \(i(c)\) is an indicator function

\[
i(c) = \begin{cases} 
1 & : \text{Party } p \text{ in market } m \text{ nominates candidate } c \\
0 & : \text{Otherwise}
\end{cases}
\]

Estimates for \(\tilde{\beta}\) are obtained by standard Maximum Likelihood Estimation of \(\ln(\ell)\).

3.2. **2nd Stage Estimation.** The estimation equation for the second stage results from equation 10. Estimation of the parameter vector \(\beta\) is done using linear methods. However, estimating this equation would not account for the correlation between \(\xi_{j}\) and \(\psi^{*}_{c}\). In a small abuse of notation, let the subscript \(c^{*}\) represent the optimal choice of candidate in the first stage, so that \(c^{*} = j\) in the second stage since only the optimal candidates are selected in the first stage. This means that \(\psi^{*}_{j}\) references the valuation of the candidate unobserved characteristics by the political party in the first stage. Recall that \(\xi_{j} = \gamma \psi^{*}_{j} + \tilde{\xi}_{j}\) where \(E(\tilde{\xi}_{j} | \psi^{*}_{j}) = 0\). This decomposes the unobserved characteristics in the second stage into a part correlated with the unobserved characteristics from the first stage and the rest of the
characteristics that are not. Now the estimation equation becomes

\[ \ln(s_j) - \ln(s_N) = \beta X_j + \gamma \psi^*_j + \tilde{\xi}_j \]

Since \( \psi^*_j \) is unobservable, I need \( \hat{\psi}_j \equiv E(\psi^*_j | \psi^*_j > -\tilde{\beta} \bar{X}_j, \psi^*_j + \tilde{\beta} \bar{X}_j - \tilde{\beta} \bar{X}_c > \psi^*_c \quad \forall c \neq j) \) to estimate this equation. This means that I need to find the expected value of the unobserved candidate characteristics in the first stage for candidate \( j \) given the selection rule that he won the primary election. This expected value can be found for each candidate based on the results of the first stage estimation and the distributional assumption.

\[ E(\psi^*_j | \psi^*_j > -\tilde{\beta} \bar{X}_j, \psi^*_j + \tilde{\beta} \bar{X}_j - \tilde{\beta} \bar{X}_c > \psi^*_c \quad \forall c \neq j) = \]

For conciseness, redefine: \( s \equiv \psi^*_j, V_j \equiv \tilde{\beta} \bar{X}_j, \ & V_k \equiv \tilde{\beta} \bar{X}_c \)

\[ (22) \quad \frac{1}{1 - F(-\tilde{\beta} \bar{X}_j)} \int_{-\tilde{\beta} \bar{X}_j}^{\infty} \psi^*_j \prod_{c \neq j} F(\psi^*_j + \tilde{\beta} \bar{X}_j - \tilde{\beta} \bar{X}_c) f(\psi^*_j) d\psi^*_j \]

\[ (23) \quad = \frac{1}{1 - e^{-V_j}} \int_{-V_j}^{\infty} s \prod_{c \neq j} \frac{e^{s+V_j-V_c}}{1 + e^{s+V_j-V_c}} \frac{e^s}{(1 + e^s)^2} ds \]

\[ (24) \quad = (1 + e^{-V_j}) \int_{-V_j}^{\infty} s \prod_{c} \frac{e^{s+V_j-V_c}}{1 + e^{s+V_j-V_c}} ds \]

If there is only one candidate in primary, the closed form solution is:

\[ (25) \quad V_j e^{-V_j} + (1 + e^{-V_j}) \ln(1 + e^{-V_j}) \]
Otherwise, it is straightforward to approximate the single dimensional integral in equation (25). With \( \hat{\psi}_j \) for every general election candidate, the estimation of the second stage model can be done using 2SLS on:

\[
\ln(s_j) - \ln(s_N) = \beta X_j + \gamma \hat{\psi}_j + \tilde{\xi}_j
\]  

3.3. **Standard Errors.** The estimation of the preference parameters in this model is done in four distinct steps. Theoretically, it would be possible to derive the standard error formula after each step. However, this generally becomes difficult for a two step estimation process and would be incredibly difficult in this model because of the four steps and the inclusion of an approximated regressor. Therefore, the standard errors are found through a bootstrap process with 1000 repetitions. The resampling for the bootstrap must be done at the general election level to maintain consistency of the sample. Each bootstrapped sample will have equal number of general elections but will have a different number of candidates.

4. **Data**

4.1. **Overview.** For this analysis, a unique data set was compiled from a variety of sources. All of the election results and candidate expenditure data is provided by the Federal Election Commission (FEC). Biographical information on the candidates comes from VoteSmart.org. Voter registration statistics and the date for the primary elections come from each state’s board of elections. This paper currently uses data from the 2006 election cycle.

---

5The instruments used for this analysis will be discussed in the data section
The FEC provides summary files that document each candidate’s expenditures during the two year election cycle. The amounts reported in this file provide the total amounts spent for the whole election but they do not provide the timing of the expenditures. Of most importance to this analysis, these files do not provide the total amount spent before the primary election. This value must be determined by looking at the individual candidate’s expenditure report filings.

Each candidate is required to file quarterly reports in addition to prepimary reports and pregeneral election reports. In each of these reports, every expenditure reported to the FEC are dated. Each expenditure is also supposed to indicate if the expenditure was for the primary election or the general election. However, this denotation is not mandatory and is missing for the majority of candidates. Therefore, it is necessary to make some assumptions to determine the primary expenditure variable for each candidate.

If the candidate won the primary election, prepimary expenditures are defined as the total amount spent during the election until the day of the election. This value can be found by looking at the prepimary expenditure report and the next quarterly expenditure report. The prepimary report is filed 20 days before the primary election and provides the total amount spent up until the day of the report. This number is then added with all expenditures from the next quarterly report that are dated before or on the day of the primary. If the candidate did not win the primary, it is assumed that all the candidates expenditures during the election cycle were used for the primary election. The primary losers prepimary expenditures are the amount reported in the election summary file.
Biographical information for candidates is provided by VoteSmart.org. Every election, this organization sends out a survey to every candidate requesting biographical information. The survey includes the standard questions of birthdate, gender, and marital status. The survey also asks for information on educational and occupational history. Due to the limited response rates for some of the questions, this paper will only use the gender information at this time.

To estimate the second stage of the model, it is necessary to have the total number of registered voters in each congressional district at the time of the general election. The majority of states either provide this information on their website or the numbers can be obtained by contacting the board of elections. However, without this information, the state must be excluded from the sample. The most common reasons that a state did not have the total number of registered voters by congressional district is that they do not have the data aggregated at all or the data is aggregated at the county level and the congressional districts do not follow county boundaries. I also exclude the congressional districts in Connecticut, Louisiana, and Utah because they do not have the standard procedural process from primary and general elections. Given these exclusion, I am able to use the congressional districts from 35 states.\footnote{The states included in the sample: AK, AR, AZ, CA, CO, DE, FL, ID, IN, IA, KY, ME, MD, MN, MT, NE, NV, NH, NJ, NM, NY, ND, OH, OK, OR, PA, RI, SC, SD, TX, VT, VA, WA, WV, WY.}

4.2. Control Variables. The candidate and election characteristics that are used as control variables are very similar in the first and second stages. Both stages use a dummy variable
for incumbency, denoted *Incumbent*, party specific dummy variables, *dDem* and *dRep*, and a gender dummy variable, *dFemale*. These variables vary by candidate and do not change from the primary election to the general election. State dummy variables are also included in both stages, but these are of limited interest individually and will be excluded from the summary statistics and estimation results.

As mentioned earlier, the covariate that changes greatly between the two stages is the amount the candidate has spent. Since primary elections occur months before the general election, the primary expenditures are much less than the general election expenditures. The actual control variables used in the estimation is the natural log of one plus the candidate expenditures. The value is also interacted with the incumbency dummy to allow for different marginal effects of expenditures for incumbents and challengers.

Table 1 shows the summary statistics for the primary elections. This table shows that the vast majority of candidates are male and the average amount spent per candidate is close to a quarter of a million dollars. Also, this sample averages about 1.6 candidates per primary with the maximum number of candidates in a single primary being 8 candidates.

### Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full Sample</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Expenditures ($)</td>
<td>240,949</td>
<td>333,208</td>
<td>0</td>
<td>2,916,921</td>
<td></td>
</tr>
<tr>
<td>Incumbent</td>
<td>.3093</td>
<td>.4625</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>dFemale</td>
<td>.1604</td>
<td>.3672</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total # of Candidates</td>
<td></td>
<td>873</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total # of Primary Elections</td>
<td></td>
<td>555</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 provides the summary statistics for the general elections. The average amount spent but the candidates is over $900,000 and almost half the candidates are incumbents. Also, the majority of general elections in this sample have a Democrat and Republican running against each other. 294 total elections and 555 candidates means that 33 candidates ran unopposed in the general election. The table also shows that the average candidate has a .25 vote share but there is significant variation in this number. This average share per candidate implies about a 50% voter turnout rate for the 2006 general election. This is fairly standard percentage for a U.S. midterm election.

4.3. **Instruments.** Hausman (1996) type instruments that were popularized by Nevo (2000) are constructed to deal with the endogeneity of candidate expenditures. The congressional districts provide a distinct boundary for the elections within each state. For both stages, two instruments are created using these boundaries. The first is the average expenditure by candidates in the same party and in the same state. The second is the average expenditure...
by candidates in the opposing party and in the same state.\textsuperscript{7} Both of these instruments exclude the district that is being instrumented when calculating the averages. This ensures that these expenditures are not correlated with the vote share for the district that is being instrumented.

In the first stage, the actual instruments used are the natural log of the average primary expenditures in the rest of the state by candidates in the same party and the natural log of the average primary expenditures in the rest of the state by candidates in the opposing party. These two instruments are then interacted with the incumbency dummy variable to create four instruments for the first stage.\textsuperscript{8} In the second stage, a similar set of four instruments is constructed except the average is now over general election candidates only. Appendix A provides the reduced form regressions of the endogenous variables on the instruments and the exogenous variables to show the correlation between the instruments and the endogenous variables.

5. Empirical Results

5.1. 1st Stage Results.

5.1.1. Coefficients. The estimated preference parameters from the first stage are presented in Table 3. The magnitudes of these coefficients are not nearly as important as the significance in this stage. Most of the coefficients have the expected sign. The coefficients on

\textsuperscript{7}These instruments make it necessary to exclude an state that only has one congressional district.

\textsuperscript{8}The interaction is done because expenditure is interacted in with the incumbency dummy variable in the estimation.
\( \ln(PrimaryExpend) \), \( \ln(PrimaryExpend) \times Incumbent \) and \( Incumbent \) are all positive and significant. That implies that an increase in any of these variables leads to a higher probability of being nominated for the general election. Surprisingly, the female dummy has a negative, significant coefficient which implies that a woman is less likely to be nominated by the political party than a man. Since the value of the \( SenateRace \) dummy variable does not vary by the different candidates within a primary, the positive significant value implies that a political party is more likely to nominate a candidate versus choosing the outside option of not nominating any candidate for the general election when a Senate Race is happening concurrently.

\( \hat{v}_1 \) and \( \hat{v}_2 \) are the additional parameters used to control for the endogeneity of expenditures. Recall that these parameters are the residuals from the reduced form regressions of the endogenous variables on the instruments and all the exogenous variables. While the coefficient on \( \hat{v}_2 \) is not significant itself, the coefficients are jointly negative significant and this confirms that negative endogeneity bias would occur without including these parameters.

5.1.2. Predicted \( \hat{\psi} \) Values. The main function of the first stage coefficients is using the coefficients to find the predicted value of \( \hat{\psi} \). Recall that \( \hat{\psi} \) is the expected value of the party valuation of the unobserved candidate characteristics and the formula for this value is given by equation (25). Figure 1 provides a histogram of these values for all the primary winners.\(^9\)

\(^9\)It is impossible to calculate \( \hat{\psi} \) for the loser's of the primary election. Since they were not chosen in the primary, you cannot place any restrictions on the valuation of their unobserved characteristics.
Table 3.

First Stage Parameters

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(PrimaryExpend)</td>
<td>4.6515***</td>
<td>(0.5886)</td>
</tr>
<tr>
<td>ln(PrimaryExpend) * Incumbent</td>
<td>4.8857***</td>
<td>(0.9878)</td>
</tr>
<tr>
<td>Incumbent</td>
<td>0.3802***</td>
<td>(0.0769)</td>
</tr>
<tr>
<td>dFemale</td>
<td>-3.2037***</td>
<td>(1.0946)</td>
</tr>
<tr>
<td>SenateRace</td>
<td>1.157***</td>
<td>(0.3750)</td>
</tr>
<tr>
<td>dDem</td>
<td>0.4909</td>
<td>(0.4142)</td>
</tr>
<tr>
<td>dRep</td>
<td>0.8361</td>
<td>(0.5862)</td>
</tr>
<tr>
<td>( \hat{v}_1 )</td>
<td>-4.1161***</td>
<td>(0.5612)</td>
</tr>
<tr>
<td>( \hat{v}_2 )</td>
<td>-0.1351</td>
<td>(0.2738)</td>
</tr>
</tbody>
</table>

Observations: 849

Bootstrap standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Coefficients for the state fixed effects are omitted.

The histogram shows that the majority of primary elections result in an expected \( \hat{\psi} \) of about 0. The unrestricted expected value of \( \psi \) is 0, so this means that the primary election does not provide any additional information for these candidates. This occurs because the value of the observed candidate characteristic is very large and therefore does not place any restriction on the values of the unobserved characteristics. As an example, if an incumbent candidate spends a lot of money on the primary election, these observed characteristics explain why the candidate was nominated for the primary, it is impossible to infer any additional information on his unobserved characteristics.
While the majority of candidates have a $\hat{\psi}$ close to zero, 101 of the 530 general election candidates that have a $\hat{\psi} > .05$ and all of these candidates are not incumbent candidates. This highlights the fact the empirical gain from the primary process is to have a measure of candidate quality for non incumbents to include in the second stage estimation.

5.2. 2nd Stage Results. Table 4 reports the estimated coefficients from three different specifications of the second stage model to show the importance of the selection process. The first column estimates the second stage estimation equation, equation (26), without...
controlling for the endogeneity of expenditure. The second column control does control for endogeneity but does not control for selection. The third columns estimates the model controlling for selection and endogeneity.

Column (1) shows that even without controlling for endogeneity, this method produces results that qualitatively similar to the accepted results from prior literature. A challenger has a higher marginal return to expenditure and an incumbent has a significant inherent advantage. Additionally, the positive significant coefficient on SenateRace implies that a concurrent Senate election increase the overall voter participation in the election. Most

![Table 4. Second Stage Parameters](image)

Table 4. Second Stage Parameters

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>Selection, No Endogeneity</th>
<th>Endogeneity, No Selection</th>
<th>Selection and Endogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>(\ln(\text{GeneralExpend}))</td>
<td>0.0716***</td>
<td>0.1059***</td>
<td>0.0946**</td>
</tr>
<tr>
<td></td>
<td>(0.00085)</td>
<td>(0.0334)</td>
<td>(0.0412)</td>
</tr>
<tr>
<td>(\ln(\text{GeneralExpend}) \ast \text{Incumbent})</td>
<td>-0.0403*</td>
<td>-0.0712</td>
<td>-0.0195</td>
</tr>
<tr>
<td></td>
<td>(0.0221)</td>
<td>(0.1923)</td>
<td>(0.2286)</td>
</tr>
<tr>
<td>(\text{Incumbent})</td>
<td>0.9495***</td>
<td>1.2023</td>
<td>0.5675</td>
</tr>
<tr>
<td></td>
<td>(0.3034)</td>
<td>(2.5483)</td>
<td>(3.004)</td>
</tr>
<tr>
<td>(d\text{Female})</td>
<td>0.0035</td>
<td>-0.0064</td>
<td>-0.0074</td>
</tr>
<tr>
<td></td>
<td>(0.0325)</td>
<td>(0.0351)</td>
<td>(0.0379)</td>
</tr>
<tr>
<td>(\text{SenateRace})</td>
<td>0.2512***</td>
<td>0.2543***</td>
<td>0.2789***</td>
</tr>
<tr>
<td></td>
<td>(0.0375)</td>
<td>(0.0859)</td>
<td>(0.0428)</td>
</tr>
<tr>
<td>(d\text{Dem})</td>
<td>-1.9602***</td>
<td>-2.3221***</td>
<td>-2.2293***</td>
</tr>
<tr>
<td></td>
<td>(0.1121)</td>
<td>(0.4356)</td>
<td>(0.5064)</td>
</tr>
<tr>
<td>(d\text{Rep})</td>
<td>-2.1018***</td>
<td>-2.4167***</td>
<td>-2.3426***</td>
</tr>
<tr>
<td></td>
<td>(0.1073)</td>
<td>(0.3623)</td>
<td>(0.4165)</td>
</tr>
<tr>
<td>(\hat{\psi})</td>
<td>0.3441***</td>
<td>0.2610</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0913)</td>
<td>(0.1708)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>539</td>
<td>539</td>
<td>539</td>
</tr>
</tbody>
</table>

Bootstrap standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Coefficients for the state fixed effects are omitted.
importantly in this specification, the coefficient on $\hat{\psi}$ provides a test of the existence of selection bias. The significant coefficient is a rejection of the null hypothesis of no selection bias. The positive magnitude implies that the political party’s unobserved valuation is positively correlated with the voter’s unobserved valuation. This is evidence that the political party chooses the candidate that it believes will have the best probability of winning the election.

Columns (2) and (3) show the effect of controlling for selection when the estimates also control for the endogenous expenditures. These results are much less precise, therefore, are preliminary until another year of data can be incorporated to increase the precision. However, these show the beginning of the importance controlling for selection. In column (3), the selection coefficient is not significantly measured, but is incredibly close. The t-value for the coefficient 1.53. The inclusion of the selection term reduces the magnitude of both coefficients associated with the incumbent challenger. The marginal gain from an increase in expenditure is much closer for challengers and incumbents in column (3) compared to column (2), and incumbents have a much lower inherent advantage.

5.2.1. Expenditure Elasticities. The best way to interpret the magnitude of the expenditure coefficients is through the elasticities. The expenditure elasticities that show the changes in predicted vote shares for each of the voting options from an increase in expenditure. The elasticities calculated using the estimated coefficients from specification (3) are presented in Table 5. This table reports the average own- and cross-expenditure elasticities. These elasticities were calculated for every congressional district and the average elasticity across
Table 5.

Elasticities: Specification 3

<table>
<thead>
<tr>
<th></th>
<th>Challenger</th>
<th>Incumbent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Democrat</td>
<td>Republican</td>
</tr>
<tr>
<td>$\Delta \hat{s}_D$</td>
<td>0.0746%</td>
<td>-0.0309%</td>
</tr>
<tr>
<td>$\Delta \hat{s}_R$</td>
<td>-0.0263%</td>
<td>0.0797%</td>
</tr>
<tr>
<td>$\Delta \hat{s}_{Not\text{Vote}}$</td>
<td>-0.0483%</td>
<td>-0.0488%</td>
</tr>
</tbody>
</table>

Notes: Values are the average elasticities from all the districts.

all the districts in the sample are reported. The values show the average percentage change in predicted vote share for each option if a candidate would increase his expenditure by 1%.

These elasticities quantify an very important fact. Candidates can increase their vote share by increasing expenditures and the increase in votes has two different sources. The increase comes not only from stealing votes from the other candidate, but also from convincing voters to vote for them who would have otherwise abstained. This can be seen from the negative percentage change in the share of registered voters will would choose to not vote. In these estimates, the share of voters who would abstain decreases by a larger percentage than the share of voters who would switch for whom they vote for all types of candidates.

These elasticities also provide a calculation of the cost of an additional vote. The average vote share in this sample is .2495 and the average candidate spent about $920,000. This table shows that, by spending 1% more, an Democratic challenger can increase his vote share by 0.0746%. Therefore, an additional $9,200 expenditure increases his vote share to .2497.
With an average of about 400,000 registered voters in each district, the difference in vote shares is equal to 74 votes. This implies that it cost about $124 per new vote. This is a preliminary calculation but shows another one more advantage of this methodology.

6. Conclusion

7. Appendix

Appendix A. Reduced Form Regression on Expenditures

This table shows that the instruments are highly correlated with the endogenous expenditures, the first condition that must be satisfied for a valid instrument set. The first two columns regress the first stage endogenous variables, $ln(\text{primexpend})$ and $ln(\text{primexpend}) \times \text{incumb}$ on the instrument set and the exogenous variables. Columns 3 and 4 show the results of the reduced form regression for $ln(\text{genexpend})$ and $ln(\text{genexpend}) \times \text{incumb}$. The F test at the bottom of each column test the joint hypothesis that all the instruments are insignificant. This hypothesis is rejected at .001 level for all four regressions.
### Reduced Form Regressions

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnrestprimavgownparty</td>
<td>-2.376***</td>
<td>-0.297***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.597)</td>
<td>(0.0521)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lnrestprimavgoppparty</td>
<td>-3.256***</td>
<td>-0.256***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.569)</td>
<td>(0.0484)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lnrestprimavgownpartyincumb</td>
<td>-1.256***</td>
<td>0.209***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.351)</td>
<td>(0.0534)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lnrestprimavgopppartyincumb</td>
<td>-0.488</td>
<td>0.0400</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.385)</td>
<td>(0.0544)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lnrestgenavgownparty</td>
<td></td>
<td></td>
<td>-2.449***</td>
<td>-0.352***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.639)</td>
<td>(0.0608)</td>
</tr>
<tr>
<td>lnrestgenavgoppparty</td>
<td></td>
<td></td>
<td>-3.204***</td>
<td>-0.311***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.597)</td>
<td>(0.0555)</td>
</tr>
<tr>
<td>lnrestgenavgownpartyincumb</td>
<td></td>
<td></td>
<td>-0.435</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.494)</td>
<td>(0.0936)</td>
</tr>
<tr>
<td>lnrestgenavgopppartyincumb</td>
<td></td>
<td></td>
<td>0.455</td>
<td>-0.00491</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.417)</td>
<td>(0.0552)</td>
</tr>
<tr>
<td>incumbent</td>
<td>26.54***</td>
<td>9.784***</td>
<td>5.037</td>
<td>12.20***</td>
</tr>
<tr>
<td></td>
<td>(6.295)</td>
<td>(0.991)</td>
<td>(8.279)</td>
<td>(1.415)</td>
</tr>
<tr>
<td>ddem</td>
<td>75.36***</td>
<td>6.745***</td>
<td>85.21***</td>
<td>9.097***</td>
</tr>
<tr>
<td></td>
<td>(13.25)</td>
<td>(1.219)</td>
<td>(15.11)</td>
<td>(1.585)</td>
</tr>
<tr>
<td>drep</td>
<td>74.22***</td>
<td>6.771***</td>
<td>84.13***</td>
<td>9.190***</td>
</tr>
<tr>
<td></td>
<td>(13.24)</td>
<td>(1.222)</td>
<td>(15.09)</td>
<td>(1.588)</td>
</tr>
<tr>
<td>senatorace</td>
<td>-5.092*</td>
<td>-0.744***</td>
<td>-5.767**</td>
<td>-1.232***</td>
</tr>
<tr>
<td></td>
<td>(3.071)</td>
<td>(0.210)</td>
<td>(2.925)</td>
<td>(0.306)</td>
</tr>
<tr>
<td>dfemale</td>
<td>0.931**</td>
<td>-0.0397</td>
<td>1.032***</td>
<td>-0.00323</td>
</tr>
<tr>
<td></td>
<td>(0.361)</td>
<td>(0.0337)</td>
<td>(0.386)</td>
<td>(0.0337)</td>
</tr>
<tr>
<td>Observations</td>
<td>849</td>
<td>849</td>
<td>849</td>
<td>849</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.863</td>
<td>0.998</td>
<td>0.868</td>
<td>0.998</td>
</tr>
<tr>
<td>F Test: F =</td>
<td>12.33</td>
<td>10.76</td>
<td>8.727</td>
<td>8.555</td>
</tr>
<tr>
<td>p =</td>
<td>9.68e-10</td>
<td>1.68e-08</td>
<td>6.71e-07</td>
<td>9.16e-07</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Coefficients for the state fixed effects are omitted.
References


