A Causal Reanalysis of "A Natural Experiment in Proposal Power and Electoral Success"

Spencer Braun, Samuel Wong

Fall 2020

1 Introduction

The 2014 paper "A Natural Experiment in Proposal Power and Electoral Success" by Loewen et al.[1] studies the relationship between the power to propose legislation and re-election for Canadian Members of Parliament. Canada's legislative branch is comprised of the Senate and the House of Commons. While the Senate members are appointed by the Governor General (the head of the legislative branch) via the recommendation of the Prime Minister (head of the executive branch), the House of Commons members are elected by the people of Canada. Of these House of Commons members, some are selected by the Prime Minister to be cabinet members, while the remaining are non-cabinet members.

Members of the cabinet have the right to propose legislation by direction of the Prime Minister. This contrasts the non-cabinet members, who since 2004 had have a lottery system that determines who gets a chance to introduce legislation. The authors focus on the non-cabinet members proposal power as an opportunity to present a causal relationship, as the lottery system creates a "natural" randomized experiment. The paper concludes that those in the governing party who are endowed with this proposal power receive an increase in their share of the vote in the next election.

2 Study Overview

2.1 Study Design

The population of this study was comprised of legislative and electoral records of incumbents serving in the 38th (2004-2006) and 39th (2006-2008) Canadian Parliaments. The total number of units in the 38th Parliament were 206, while the total number of units in the 39th Parliament were 198. Let $Z_i \in \{0, 1\}$ denote whether or not unit i was given the opportunity to propose legislation. As previously mentioned, the lottery system lends itself to be a "natural" randomized experiment, thus the assignment vector $\vec{Z}$ can be viewed as a realization of an experimental design η, where η is a CRD(N_1, N). Loewen et al. consider those assigned the power to propose the treatment group, while those without this power are considered controls. Finally, it is noted that effects should be interpreted as "intent to treat" effects, as members can (and do) win the power to propose but fail to introduce any legislation during a parliamentary session.

2.2 Outcomes

The outcome variable was the percentage of votes obtained in the re-election, referred to as "vote share" in the original analysis. This continuous outcome, denoted Y_i for unit i, was used to measure the efficacy of the power to propose on electoral success. Using the potential outcomes framework, $Y_i(1)$ and $Y_i(0)$ denote the outcomes of possessing and not possessing the power to propose legislation, respectively.

2.3 Causal Estimand and Key Assumptions

The causal estimand is the average treatment effect of the power to propose on electoral vote share:

$$\tau^{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0) = \bar{Y}(1) - \bar{Y}(0)$$

For the assignment, we assume SUTVA. Formally, this means that $Y_i(\vec{Z}) = Y_i(Z_i)$ for $i = 1, ..., N$. This is a reasonable assumption as whether or not one member wins the lottery should not affect the outcome of the percentage of votes for a different member during re-election. Embedded in the SUTVA assumption also is the assumption that there are no hidden versions of the treatment. This makes sense as well, as the members only have two distinct options. They either "get the chance to propose" or "do not get the chance to propose".

As this is a "natural" randomized experiment, we view the design η as a CRD(N_1, N). The CRD(N_1, N) fulfills the following 4 properties:

- Probabilistic: we have that $0 < P(Z_i|\vec{X}, \vec{Y}(0), \vec{Y}(1)) < 1$. That is, that there is a non-zero probability of getting some assignment into a group
- Known assignment mechanism: the probability is known
- Individualistic: An assignment for a unit does not depend on other units. $P(Z_i|\vec{X}, \vec{Y}(0), \vec{Y}(1)) = P(Z_i|\vec{X}_i, Y_i(0), Y_i(1))$
- Unconfoundedness: The assignment does not depend on the outcome. $P(\vec{Z}|\vec{X}, \vec{Y}(0), \vec{Y}(1)) = P(\vec{Z}|\vec{X})$

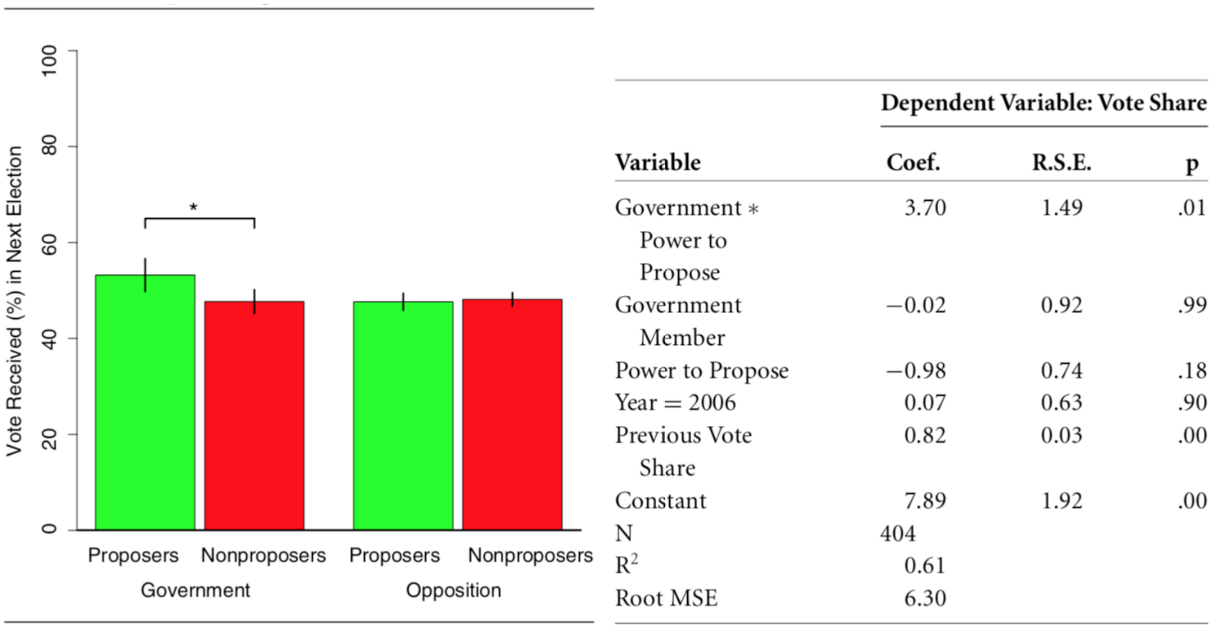
Thus, we are able to perform appropriate Neymanian and Fisherian analysis with these aforementioned assumptions.

3 Replication and Limitations of Original Analysis

In the original analysis, the authors found that the opportunity to introduce legislation increased the vote share of the governing party candidates by 5.26 percentage points (p = 0.01, two-tailed). For the opposing party candidates, this opportunity to introduce legislation had no effect (p = 0.57, two-tailed). The authors hypothesize that only governing members receive an electoral advantage as they have increased power to dictate the legislative agenda and are therefore more likely to be viewed as successful by constituents. Additionally, while opposition members also receive the power to *propose* legislation, since the governing party constitutes the largest partisan bloc in the chamber, it is significantly harder for opposition members to pass their legislation. Thus governing members receive the preponderance of attention in the original analysis.

In order to check for imbalance in covariates, the authors then conducted a linear regression analysis that controlled for the election year and the candidate vote share in the previous election. They concluded that the government incumbents experienced a 2.73 percentage point increase (95% CI = [0.29, 5.17]) in when given the opportunity to propose. This finding reinforced the previous conclusion of the effect sign, though including even this small number of covariates cut the effect size in half. Figure 1 (with these results) is taken from the authors of the original paper.

The original analysis has limitations when assessing the ATE. This paper draws its significant conclusion from this main effect, and claims that the dataset is balanced due to this regression analysis that takes into account two covariates. However, it does not take into account many of other covariates that could be affecting the main result. Later on the paper, the authors perform separate t-testing on one covariate at a time to bolster their claims of random assignment, yet they do not include a causal analysis on multiple covariates at one time, nor matching, in order to validate their claims.



	Dependent Variable: Vote Share		
Variable	Coef.	R.S.E.	p
Government ∗ Power to Propose	3.70	1.49	.01
Government Member	−0.02	0.92	.99
Power to Propose	−0.98	0.74	.18
Year = 2006	0.07	0.63	.90
Previous Vote Share	0.82	0.03	.00
Constant	7.89	1.92	.00
N	404		
R^2	0.61		
Root MSE	6.30		

Figure 1: Plot demonstrating result from conditional t-test (left) and regression analysis output (right) taken from "A Natural Experiment in Proposal Power and Electoral Success." Loewen et al. use the t-test to argue that members with the "Power to Propose" in the governing party receive a statistically significant increase in vote share in their next election. The regression analysis is meant to demonstrate this result is robust to some covariate control.

4 Exploratory Analysis

In our analysis, we wanted to make use of other covariates that may help balance the two groups, as well as replicate the existing analysis to make important changes where we saw necessary. Therefore, in order to perform this causal analysis, we first performed exploratory analysis.

When re-running the analysis from the original authors both in the provided Stata code as well as in R, we matched all of the results in both the main paper as well as the supplemental analysis except for the mean vote received percentage. We believe that the original authors had a typo in their main effect result.

Their reported main effect in the paper was a mean vote percentage difference of 5.26 points, whereas our replication of this analysis resulted in a difference of 5.52 points. We believe our analysis is correct as the provided bar chart from the paper itself (that we were able to replicate) has aggregate values which match the mean vote percentage that we obtained.

Additionally, the original analysis relies on many separate χ^2 tests to check for consistent covariate imbalances between treatment and control groups. This is a fine first step but fails to consider potential interactions between covariates that might explain a portion of the reported causal effect. A more complete treatment through a causal inference framework could reduce the potential mediators that may underlie this relationship.

5 Analysis

5.1 Adaptation of Original Analysis

The main results of the original analysis stem from two statistical tests to estimate the relationship between electoral vote share and the power to propose. The first is a t-test comparing the difference in the mean electoral vote share between members with and without the power to propose, split by governing or opposition party. The second is a regression to estimate this effect while controlling for a small number of covariates, including election year and vote share in the previous election.

While the original study found a null result when party was ignored, it was not obvious *a priori* that a causal reanalysis should condition on party affiliation. Looking at the covariate balance between treatment and control groups (Figure 2), there are significant imbalances including in party ("gov") as well as media mentions, a potential confounder for the vote share received in election.

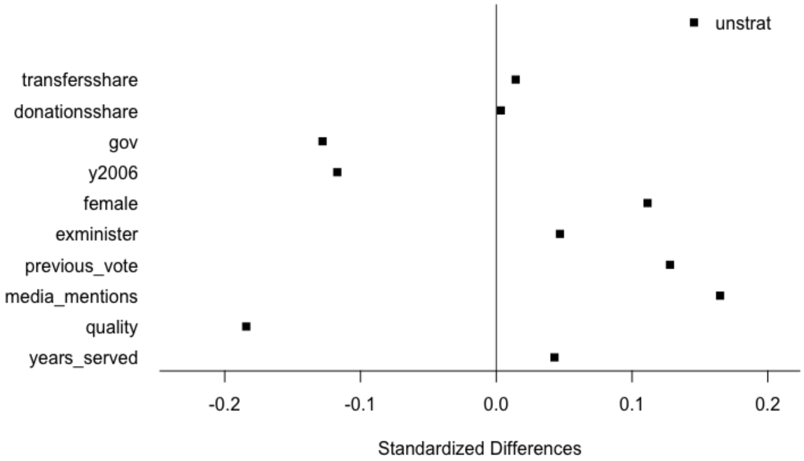


Figure 2: Standardized differences in covariates between treatment and control groups when governing and opposition members are considered together.

If instead we limit to the governing party (Figure 3), those imbalances mostly disappear but are replaced by large differences in counts of members who were ex-ministers and the vote share received in the prior election. Both comparisons were considered for analysis, but the focus fell on members of the governing party to maintain consistency with the results of the Loewen et al. Given that the original analysis found a null result without conditioning on governing party, our primary interest is whether the significant result of increased vote share for governing members holds under a rigorous causal framework.

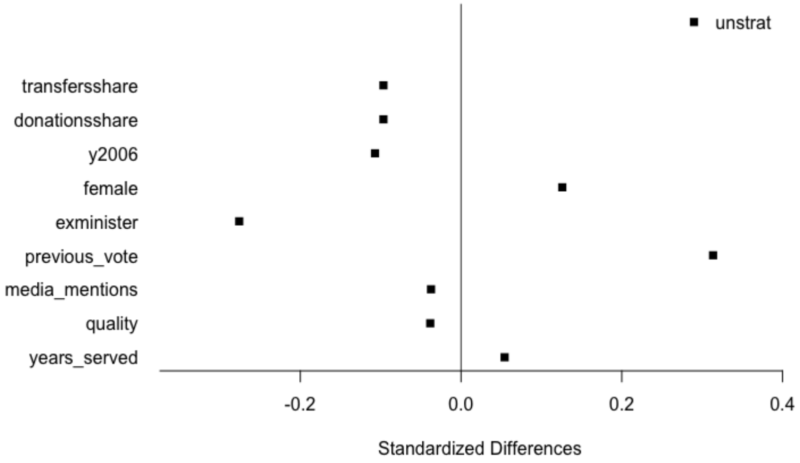


Figure 3: Standardized differences in covariates between treatment and control groups, limiting to members in the governing party.

5.2 Propensity Score Estimation

As noted above, while the authors conducted a number of randomization tests after taking the simple difference in means, they made no effort to achieve covariate balance between the treated and control units. While there were many covariates available in the dataset to choose from, some suffered from large numbers of missing values while others could not be identified due to limited documentation.

In order to achieve balance, we turned to estimating propensity scores, defined as $\pi(X_i) = P(Z_i = 1|X_i)$, the probability of being in the treatment group given the covariates X_i. The goal was to use a balancing score (in this case the propensity score), such that Z_i was independent of X_i conditional on this balancing score.

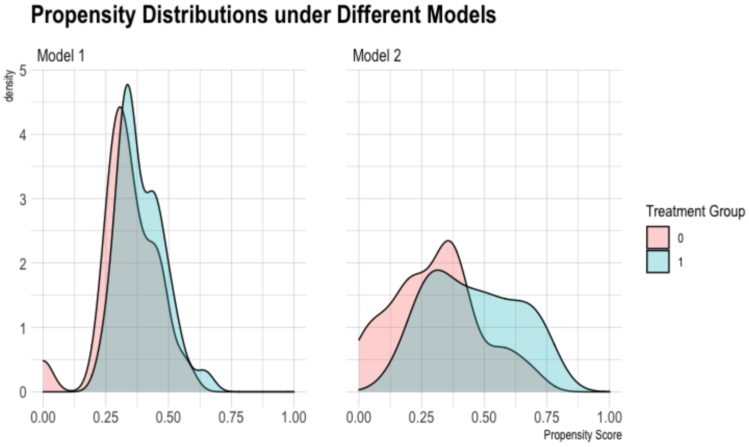


Figure 4: Propensity score distributions under different logit models. Model 1 (left) generated propensity scores with much more overlap in distributions between treatment and control groups. Model 2 (right), a larger model considering more covariates, achieved significantly worse overlap.

To estimate $\hat{\pi}$, we used a logistic regression, which has the form $logit(\pi_i) = \beta^T X_i + \epsilon_i$. Different models were attemped in order to achieve balanced propensity scores between treatment and control, and we found that a simpler model considering campaign resources, election year, governing party, gender, prior vote share, media mentions, candidate quality, and years served yielded the best results (Figure 4).

Adding covariates tended to separate the densities of scores between assignments and led to more extreme values that could increase variance in the IPW estimators employed. As noted in Imbens and Rubin, 2015 [2], the goal of estimating propensity scores is to improve the covariate balance between treatment and control more than creating a fully predictive model. Since the first, more sparse model better achieves that goal, its propensity scores were used in the analysis.

5.3 Matching Analysis

We decided to performing matching analysis in order to produce a more accurate causal picture of the results. In optimal matching, the objective is the minimize a certain distance metric d over all the matches. The goal is to obtain M^{opt}, where it is defined as

$$M^{opt} = argmin_M \sum_{i \in I_1} \sum_{j \in M(i)} d(\vec{X_i}, \vec{X_j})$$

The distance metric that was used for this analysis was the Mahalanobis distance, defined as

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T \hat{\Sigma}^{-1} (X_i - X_j)}$$

Both optimal pair and non-pairwise matching were considered. However, across the two elections, the governing party had 35 candidates with and 63 without the power to propose. Clearly, only a 1:1 exact matching was possible as there were not enough controls to perform multiple matches. A simple optimal match reduced covariate imbalance, particularly in ex-minister, measures of campaign resources, and election year but did cause balance in quality and years in parliament to worsen. Notably, while the imbalance in propensity scores decreased, the standardized difference remained high compared to the covariates.

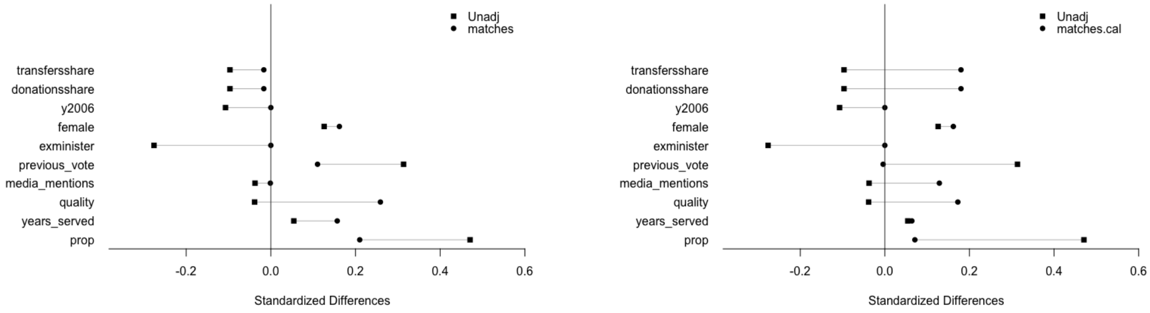


Figure 5: Standardized differences between treatment and control groups before and after matching. A simple 1:1 exact matching (left) increased balance in many covariates, but left propensity scores relatively imbalanced. Adding a caliper to the 1:1 matching (right) improved the propensity balance while sacrificing that of some other covariates.

In order to reduce propensity score imbalance, we applied constraints to the problem, namely a caliper. In this case, we solve M^{opt} with the constraint that $|\hat{\pi}(X_i) - \hat{\pi}(X_j)| \leq c * SD(\hat{\pi}(X))$, here with $c = 0.1$. This produced a matching with nearly balanced propensity scores, but sacrificed the improved balance in some covariates like measures of campaign resources, "transfershare" and "donationshare." Despite these limitations, the caliper produced higher quality matches in covariates likely to cause confounding, such as quality, years served, and previous vote share.

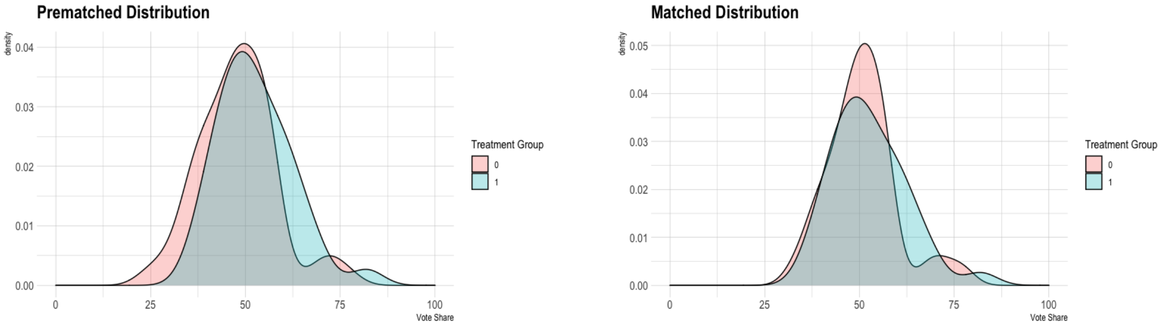


Figure 6: Distributions of outcomes (vote share) in treatment and control groups. The distribution without matching (left) showed significant advantages for the treatment share, while after matching (right), the relationship became less clear cut.

Bringing in the outcome variable, it is clear that the matching process shifted the relative distributions of vote share between treatment and control groups (Figure 6). While the conclusion that the power to propose increases vote share for members of the governing party is quite clear from the unmatched density plot, the relationship is murkier after matching. Statistical testing is clearly necessary to quantify the new relationship.

A Fisher Randomization Test (FRT) was conducted to detect statistical significance in the vote share difference between treated and control paired units. The Fisherian analysis does not require asymptotic assumptions. In the FRT, we compute our test statistic $T(Z^{obs}, Y^{obs})$. After simulating a large number of samples from the underlying null distribution T, we calculated our p-value as $P_\eta(T(Z, Y^{obs}) \geq T^{obs}|H_0) = \sum_Z \mathbb{1}\{T(Z, Y^{obs}) \geq T^{obs}\}P_\eta(Z)$.

Matching produced an observed difference in means $\hat{\tau}^{DIM} = 1.88$, considerably smaller than the 5.26 point difference reported in Loewen et al. or the 5.52 points found using their reanalysis code. The FRT also failed to find statistical significance at the $\alpha = 0.05$ level for this difference, producing a p-value of 0.088 for the one-tailed hypothesis $H1 : Y(1) > Y(0)$.

Therefore, while matching found some effect that at least agrees in sign with the results presented in the original analysis, once covariates are more closely balanced, the significance of these results disappears.

5.4 Sensitivity Analysis

While the FRT produced a p-value above what many might consider significant, 0.08 is still close to more classic inference thresholds like $\alpha = 0.05$. With more data, this finding could become significant, and the estimated effect size is large enough to have impact on the outcome of elections. We must acknowledge that to produce these results, we made strong assumptions around ignorability that could fail to hold in the real world. Thus it is important to consider not just the found effect size but also how robust the findings are to violations of these assumptions.

We assumed that within a matched pair, the probability of treatment of unit 1, π_1, was equal to that of unit 2, π_2. Sensitivity analysis considers potential deviations from this strong assumption; if we allow the odds ratio between units for the probability of treatment to vary, what is the maximum p-value we might find under these new conditions.

More formally, we define our odds ratio as $\nu_k = \frac{odds_{k1}}{odds_{k2}}$. We perform sensitivity analysis to find $\frac{1}{\Gamma} \leq \nu_k \leq \Gamma$; for all k, for $\Gamma \geq 1$. For the p-values that we obtain, for the values for each Γ, we calculate

$$M(\Gamma) = max_{\vec{\nu} \in [\frac{1}{\Gamma}, \Gamma]^k}(pval_{\vec{\nu}})$$

We used the senm function from sensitivitymult package in R to compute $M(\Gamma)$ under loosening restrictions on the odds ratio of paired units.

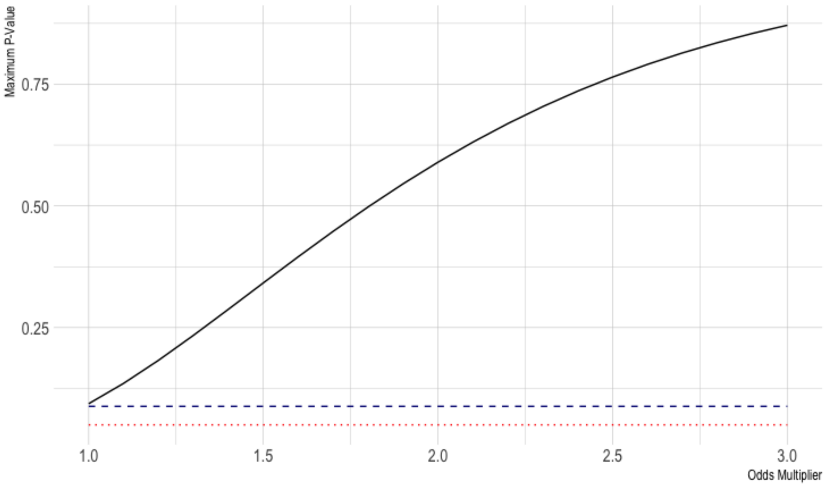


Figure 7: Maximum p-value under varying levels of Γ, an odds multiplier parameter allowing deviations from equal treatment odds ratio of paired units. Deviations from ignorability assumptions led to quickly declining significance.

Small deviations from 1, corresponding to equal probability of treatment, maintained p-value near 0.08. Quite quickly though, as the odds ratio becomes imbalanced, the p-values surpass any realistic threshold of significance. This finding is a clear indication that even the modest effect found by matching relies heavily on the ignorability assumptions imposed.

5.5 Propensity Score Subclassification

Subclassication is a technique that allows for more classical Neymanian statistical tests under normal distribution assumptions. In an ideal case, we divide data along propensity score in order to view the causal effects through the lens of a stratified randomized experiment. Here the propensity scores require coarsening to be stratified, as they follow a continuous distribution.

More formally, we have K different propensity scores and we can stratify based on these scores. We will have K strata, each with units with that propensity score. More formally, we define $I_k = \{i : S_i = k\}$, where S_i is the strata. Then $N(k) = |I_k|; N_1(k) = \sum_{i=1}^{n} \mathbb{1}\{S_i = k\}Z_i; N_0(k) = \sum_{i=1}^{n} \mathbb{1}\{S_i = k\}(1 - Z_i)$

We claim the proposition that for all $i \in I_k$ and $\pi(X_i) = \pi_k$, that $\vec{Z}_{I_k}|_{\sum_{i\in I_k} Z_i = N_1^{obs}(k)} \sim CRD(N_1^{obs}(k), N(k))$. In other words, the assignment for each strata is a CRD.

Secondly, we claim the proposition that $\vec{Z}|_{\sum_{i\in I_k} Z_i = N_1^{obs}(k)} \sim SCRD(\vec{N_1}^{obs}, \vec{N})$. In other words, the assignment across the whole experiment is SCRD.

Due to these two propositions, we can view this as an SCRD and use the same approaches as in a randomized experiment. For example, we can use the estimator $\hat{\tau}^{SCRD} = \sum_{k=1}^{K} \frac{N(k)}{N}\hat{\tau}_k$, where $\hat{\tau}_k$ is the DIM estimator within stratum k. From before, we know that $E_\eta[\hat{\tau}^{SCRD}] = \tau$, so the estimator is unbiased.

Treatment and control units were divided into 5 relatively balanced groups based on propensity score quantiles. The overall effect size is calculated by finding the causal effect within each stratum and taking a weighted sum to form the estimator $\hat{\tau}^{strat} = \sum_{k=1}^{K} \frac{N(k)}{N}\hat{\tau}_k$. The calculated effect size is $\hat{\tau}^{strat} = 2.65$, noticeably larger

Quantile	Treated Units	Control Units	Difference in Means	Variance
1	3	17	3.01	0.75
2	5	14	0.57	0.25
3	11	9	4.11	0.39
4	8	11	1.27	0.39
5	8	12	4.11	1.31

Table 1: Statistics by stratum used in estimator calculations for subclassification by propensity score.

than the causal effect found under the matching approach. However the variance of the estimator is quite large as well at $Var(\hat{\tau}^{strat}) = 3.08$. Therefore when a Neymanian confidence interval is calculated using a normal approximation, the confidence interval $(-0.79, 6.09)$ includes the possibility of a null effect. Thus subclassification comes to the same conclusion; there is not a significant effect of power to propose on vote share once we account for covariates.

5.6 Inverse Propensity Weighted Estimators

An alternative to a matching approach, leaving some units excluded from the analysis, is to consider weighting outcomes instead. A popular methodology is to inversely weight outcomes by a measure of the likelihood a unit is chosen for treatment, thereby eliminating the selection effects that may be present in an unadjusted difference in means. Here, the estimated propensity scores serve as a convenient measure of the likelihood of treatment, hence the name Inverse Propensity Weighted (IPW) Estimators. Additionally, it is important to note that while matching focused on a finite sample analysis, the IPW estimators consider effects from the superpopulation perspective.

$$\text{Horvitz-Thompson Estimator : } \hat{\tau}^{HT} = \frac{1}{N}\sum_i \frac{Z_i}{\pi(X_i)}Y_i - \frac{1}{N}\sum_i \frac{1-Z_i}{1-\pi(X_i)}Y_i$$

The Horvitz-Thompson estimator is an unbiased IPW for τ^{ATE}, but can suffer from high variance for poorly behaved propensity scores. Given the observed distribution in scores was concentrated toward the middle of the range $[0, 1]$, it seemed worthwhile to apply Horvitz-Thompson to the difference in vote share received.

$$\text{Hajek Estimator : } \hat{\tau}^{H} = \left[\sum_{i=1}^{N} \frac{Z_i}{\pi(X_i)}\right]^{-1} \sum_{i=1}^{N} \frac{Z_i}{\pi(X_i)}Y_i - \left[\sum_{i=1}^{N} \frac{(1-Z_i)}{(1-\pi(X_i))}\right]^{-1} \sum_{i=1}^{N} \frac{(1-Z_i)}{1-\pi(X_i)}Y_i$$

Alternatively, the Hajek estimator substitutes some bias for reduced variance. Its improvements are most felt when propensity scores are not necessarily centrally distributed on $[0, 1]$.

Estimator	Estimated $\hat{\tau}$	Standard Error	P-Value
Horvitz-Thompson	2.33	2.005	0.248
Hajek	2.33	2.001	0.248

Table 2: Estimated effect sizes calculated using Horvitz-Thompson and Hajek estimators.

Table 2 shows the results of the IPW estimation of the effect size. In both cases, a causal effect was found close to that which was found in the original paper's regression analysis; however, neither IPW estimate produces a statistically significant result. This is consistent with our findings under matching, where we found that the paper's effect size may be exaggerated or its significance inflated.

6 Discussion

The original analysis claimed that members of the Canadian House of Commons with the power to propose legislation and were members of the governing party received a higher vote share in elections that those without this power. Its reliance on a randomization mechanism, a lottery to determine who received the power the propose, was not unreasonable and led to quite simple statistical techniques to gauge effect. While the authors reported both a t-test and a regression, their emphasis on the smaller effect size found under regression may indicate that they believed that their ability to view this natural experiment as a pure randomized control trial may be lacking.

Under a variety of causal estimation techniques, the result fails to hold significance even for reduced effect sizes. Ultimately, this difference may indicate that there are additional factors at play in the relationship between legislative power and electoral outcomes. While the lottery mechanism seems mostly random, constituents may form opinions about representatives in ways that interact with their legislative victories.

Additionally, the analysis may be complicated by the fact that the assignment is intent to treat. While the power to propose is randomly assigned, the act of introducing legislation is not. It is hard to imagine a causal connection between simply being able to propose a bill and electoral success, and therefore we may see effect sizes obscured by those in the treatment group who failed to partake in the actual treatment of introducing legislation.

Ultimately, while the original analysis under Loewen et al. was not especially problematic, a more thorough causal framework could have made the paper stronger. First, while randomization checks were certainly warranted, the self-contained nature of individual χ^2 tests limited the ability to understand how covariates interact. Matching and propensity score methods offered a much more robust methodology for finding differences between treatment groups. Second, the authors should have considered the sensitivity of their analysis to assumptions, given that the effect under study arose from a natural experiment and there are numerous paths for representatives and constituents to build relationships. Finally, it is important to note the variation in effect size and significance across the causal methods used. They relied on different assumptions and considered different populations, and some expertise may be required to disentangle how these apply to the workings of the Canadian House of Commons. An explicit discussion of key assumptions and reasons why their validity seems reasonable would strengthen the conclusions of the paper.

References

[1] P. J. Loewen, R. Koop, J. Settle, and J. H. Fowler, "A natural experiment in proposal power and electoral success," *American Journal of Political Science*, vol. 58, no. 1, pp. 189–196, 2014.

[2] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. USA: Cambridge University Press, 2015.

[3] P. Rosenbaum, *Design of Observational Studies*. 01 2010.

[4] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Dec. 2008.