Causal Inference on Level of Education and Attitude Toward Abortion

Yunfei Yu - 1005976918

December 19, 2021

Abstract

Debate on whether to ban abortion has been one of the major issues concerning women's rights over the past years. Demographic summaries on the topic provide evidence that a higher level of education is associated with a more pro-choice attitude towards abortion. Given this insight, this analysis aim at determining the causal relationship between the level of education and attitude towards *banning* abortion. We hypothesize that completing a higher level of education causes people to hold a more negative attitude towards banning abortion. By performing propensity score matching techniques on the 2011 Canadian Election Survey data, we were able to make causal inferences on the observation data. Logistic regression and two-sample t-test results support the hypothesis. The log odds of respondents who completed post-secondary education (treatment group) to *support* banning abortion is on average 0.45 lower than respondents who did not complete post-secondary education (control group). In addition, the observed difference of 6.78% between the proportions of respondents who *oppose* banning abortion in the treatment and control groups is significant at a 0.05 level of significance. In both treatment and control groups, the majority of respondents (82.89% and 76.11%, respectively) oppose banning abortion. Thus, we conclude that people who complete a relatively higher level of education have greater odds to disapprove of banning abortion; and a higher proportion of people who complete a higher level of education hold a negative attitude towards banning abortion.

Keywords

Causal Inference, Survey, Propensity Score Matching, Abortion Attitudes, Education.

Introduction

Over the past few years, the pro-life versus pro-choice debate regarding abortion has been one of the major issues concerning women's rights. On the one hand, pro-life activists argue that anti-abortion is a step towards preserving human lives (BBC, 2019). On the other hand, people who are pro-choice argue that women have the right to maintain autonomy over their bodies (BBC, 2019). While both sides seem reasonable, one would be curious about the factors that cause individuals engaging in the debate to take their stance. In other words, what factors influence or manipulate different people's beliefs on abortion? Intuitively, gender differences should be a major factor influencing attitudes on abortion. One would expect women to take a more pro-choice stance than men. Yet, previous research has shown that men and women hold similar overall attitudes on abortion; gender difference is a moderate predictor only when religiosity is controlled (Barkan, 2014). Thus, we would like to identify what factors, other than gender and religiosity, cause **differences in abortion attitudes**.

According to demographic summaries of the debate between 2018 and 2021 (Gallup, 2021), people who received a higher level of **education** have a stronger tendency towards "pro-choice" compared to those who received a lower level of education. Hence, there is evidence that level of education is associated with abortion attitudes, but whether the association is significant and whether the level of education *causes* differences in abortion attitudes requires further statistical analysis. Based on the demographic summaries and based on intuition, we hypothesize that people who completed a higher level of education would hold a more pro-choice attitude towards abortion than people with a relatively lower level of education.

To perform statistical analysis on the causal relationship between education and attitudes toward abortion, we use the Canadian Election Survey 2011 (CES2011) data which includes 2231 responses to the question "Should abortion be banned?" (*R Documentation*) in addition to demographic characteristics such as the province of living and gender. To match the hypothesis with the information given by the data, we alter the hypothesis to that **a higher level of education cause people to hold a more negative attitude towards banning abortion**, or in other words, answer *No* to the question "Should abortion be banned?". We hope that the results from this analysis provide insight on the topic and facilitate further research in social sciences and women's rights-related fields.

To test the hypothesis, we make causal inferences on the CES2011 data. If the data were to be experimental, respondents would be randomly assigned to the *treatment* (completed post-secondary education) group or the *control*/untreated (did not complete post-secondary education) group. In this case, the completion of post-secondary education is the *treatment*. We can thus conclude on causation because randomized allocation to treatment or control group balance *covariates*, which are variables that take the same value regardless of the treatment, and thus the only difference is caused by the treatment (Taback, 2019). However, the CES2011 data used in this analysis is observational and is not randomized. To conclude on causation, covariates for both treated and untreated groups are balanced via the propensity score matching method. *Propensity score* of each respondent is the probability that the respondent completed post-secondary education conditioned on all the covariates such as province and place of living, gender and religiosity (Taback, 2019). This also implies that respondents with similar propensity scores tend to have similar distributions of observed covariates (Taback, 2019). Thus, by matching propensity scores, we are able to identify groups of respondents that are similar before treatment and attribute differences between the two groups to the treatment.

Finally, we would like to provide a brief overview of the report. The CES2011 data will be introduced in detail in the following **Data** section. The **Data** section also includes the data collection and cleaning process. Then, the **Methods** section provides a detailed explanation of statistical methods used in the analysis, including propensity score matching, two-sample t-test and logistic regression. The analysis outcomes are showcased and interpreted in the **Results** section. Finally, the **Conclusion** section discusses the results and explains the limitations of the analysis. An overview of the actual data, supplementary figures and tables, and an ethics statement are included in the **Appendix**.

Data

The data used for causal inference in this report is a subset of the 2011 Canadian National Election Study (CES2011) (*R Documentation*), extracted from the `r` package `carData`. The data is chosen specifically because the subset record answers that reflect respondents' attitude toward whether abortion should be banned as well as respondents' level of education, thus suitable for answering the research question. The subset contains 9 columns or variables that record answers to different survey questions of 2231 respondents or observations. The 9 variables include demographic variables such as a respondent's province of living, gender and level of education, and specifically, includes a respondent's yes/no answer to the survey question "Should abortion be banned?" (*R Documentation*) which reflects the respondent's attitude towards abortion.

Data Collection Process

Since the data is a subset of CES2011, the collection process is documented in the CES2011 technical documentation. The respondents to the CES2011 was selected based on a two-stage sampling process from the desired population, which is all Canadian citizens who are at least 18 years of age, reside in one of the ten Canadian provinces and speak at least one of the Canadian official languages, that is, English or French (Northrup, 2012). The first stage of the sampling process involves the selection of households from a sampling frame of a list of Canadian telephone numbers (Northrup, 2012). Although the ideal full list of all residential telephone numbers is inaccessible, a list of most numbers is constructed via telephone books and commercially available lists of numbers (Northrup, 2012). A random sample from the list is generated by a computer (Northrup, 2012). Then, the second stage of the sampling process involves the selection of a respondent from a selected household (Northrup, 2012). To be eligible for the CES, the respondent must be a Canadian citizen at least 18 years of age (Northrup, 2012). And, if more than one person in the household is available for the CES, the person with the next birthday is selected (Northrup, 2012).

The survey consists of a telephone survey and a web-based survey. Calls were made to selected respondents during the day and the evening on both weekdays and weekends (Northrup, 2012). To increase the response rate and thus to make the sample more representative of the target population, respondents who initially refused to respond to the survey were contacted a second time (Northrup, 2012). The web-based survey is sent to the email address provided by the respondents and respondents who initially refused to provide an email address were also asked for the second time (Northrup, 2012).

Due to the nature of the data collection process, the data is subject to foreseeable limitations such as a relatively low response rate. The response rate of the CES2011 was 41% (Northrup, 2012). That is, more than half of the selected respondents did not complete the survey.

Data Cleaning Process

As the data is extracted directly from an `r` package, it is already in a well-organized format that is readable by `r`. That is, the extracted data has no duplicated column names or row names, and has no missing values. Thus, only a simple cleaning had to be done to facilitate propensity score matching and logistic regression, which are the main methods of analysis used in this report.

Specifically, the variable that indicates a respondent's level of education, which contains multiple levels, is converted into a binary variable. Whether a respondent completes post-secondary education served as the threshold when creating the binary variable where respondents who completed post-secondary education are assigned to the category "post-secondary" (`PS`) and respondents who did not or have not yet completed post-secondary education are assigned to the category "below post-secondary" (`belowPS`). Thus, the category "post-secondary" encapsulates respondents who received a Bachelor's degree, a Graduate diploma or attended a college; and the category "below post-secondary" consists of respondents who received a High-School diploma, did not receive a High-School diploma or received some post-secondary education (*R Documentation*). The conversion of a multilevel variable for education to a binary variable is done by creating a new variable using the `mutate` function from the `tidyverse` package in `r` on the data. The input to the `mutate` function is `ed_lev = ifelse(#condition, yes = "PS", no = "belowPs")` where *ed_lev* is the name of the new variable indicating level of education, and the condition is whether a respondent's level of education is

`bachelors` (Bachelor's degree), `college`, or `higher` (Graduate diploma). The `ifelse` function evaluates whether the condition holds for an observation, if so then the category is set to `PS`, otherwise the category is set to `belowPS`. In addition, to prepare for the logistic regression, an indicator variable that parallels the binary variable for education is created. The indicator variable *post_s* is created via `mutate` and `ifelse` where `ifelse` evaluates whether the level of education of an observation is "post-secondary". If so, the indicator variable is evaluated to 1 and otherwise is evaluated to 0.

Then, the variables for a) importance of religion to a respondent and b) a respondent's attitude toward banning abortion are renamed to better represent the corresponding survey question. The variable that corresponds to the question "Would you say that religion is very important, somewhat important, not very important, or not important at all?" (*R Documentation*) is named "importance" and is renamed as "religiosity". The variable that corresponds to the question "Should abortion be banned?" (*R Documentation*) is named "abortion" and is renamed as "ban_abortion". The renaming of variables is done using the function `rename` from the `tidyverse` package in `r`. The command `rename(new_variable_name = "old_variable_name")` serves the purpose.

Finally, variables that are not important to the statistical analysis are excluded from the cleaned data. These variables include provincial population and weight of sample size to the whole population since intuitively, a respondent's attitude towards banning abortion is unlikely to be affected by the size of the population nor by the weight of the sample size. The household id is kept in the cleaned data to provide a unique identifier to each observation.

The cleaned CES data is stored into a new data frame called `CES_new` which contains 8 variables on 2231 observations. Variable names, types and descriptions are displayed in *Table 1*:

Variable Name	Type	Description
id	Numerical	Household ID number
province	Categorical	Respondent's province of residence
gender	Categorical	Respondent's gender with binary categories `Male` and `Female`
religiosity	Categorical	Importance of religion to a respondent with levels `not`,`notvery`, `somewhat`, and `very`
urban	Categorical	Respondent's place of residence with categories `rural` and `urban`
ed_lev	Categorical	Respondent's level of education with binary categories `belowPS` indicating that a respondent does not or has not yet completed post-secondary education and `PS` indicating that a respondent completed post-secondary education and either attended a college, received a Bachelor's degree or received a Graduate diploma
post_s	Categorical	A binary variable that parallels *ed_lev*, indicating whether a respondent completed post-secondary education with 0 corresponding to `belowPS` and 1 corresponding to `PS`, created for the use of logistic regression

Variable Name	Type	Description
ban_abortion	Categorical	Respondent's answer to the survey question "Should abortion be banned?", with binary categories `Yes` and `No`

Table 1: Variable Descriptions

See **Appendix A2: Materials** for a glimpse of the data.

Data Summary

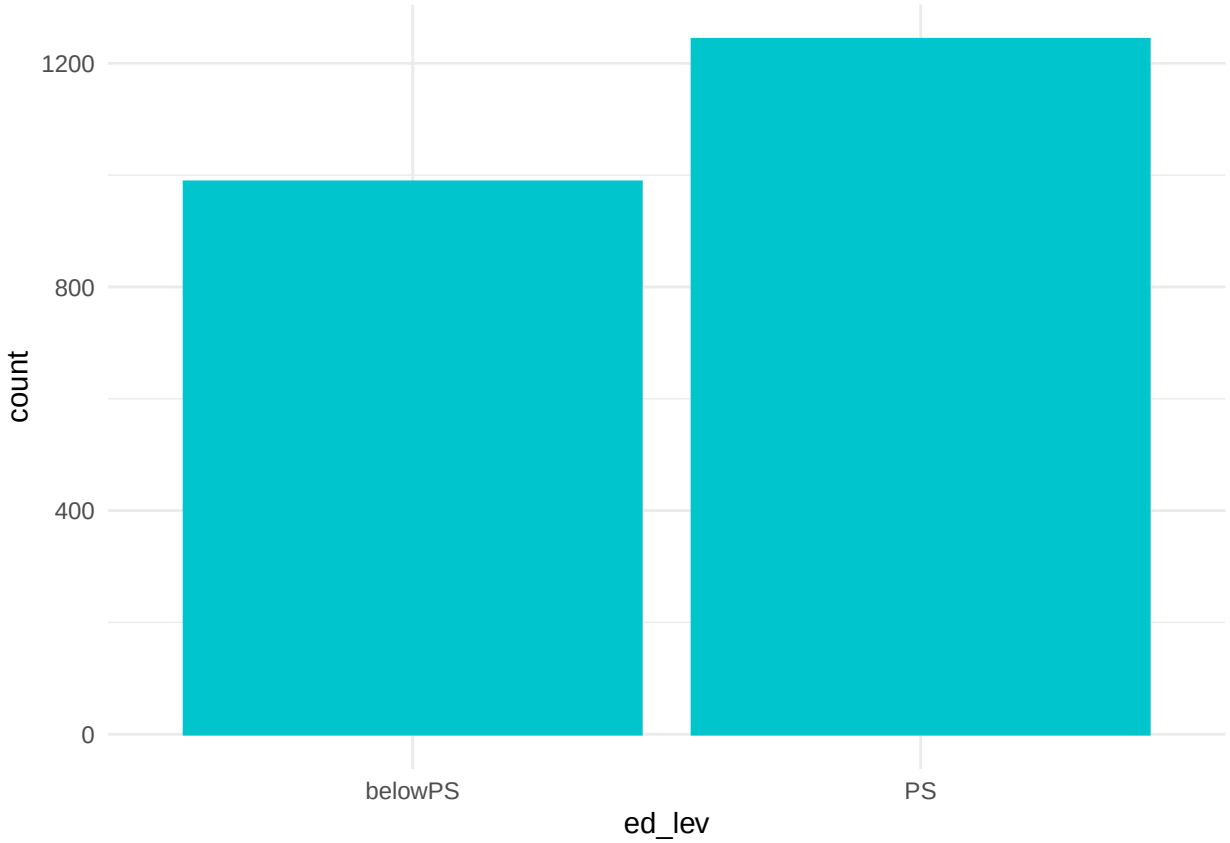

Figure 1: Distribution by Level of Education

Figure 1 demonstrates the distribution of 2231 observations over 2 categories of education. As shown by the bar plot, 988 out of 2231 respondents did not or had not yet completed post-secondary education. That is, they either received or have not received a High-School diploma or received some post-secondary education. 1243 respondents have completed post-secondary education and received either a Bachelor's degree or a Graduate diploma or completed studies in a college.

Table 2 provide a numerical summary of a) the number of respondents that fall in each category of education level and b) the number and proportion of respondents within each group that supports banning abortion (i.e. answering "yes" to the question "Should abortion be banned?") and that does not support banning abortion (i.e. answering "no" to the question "Should abortion be banned?"). Out of 988 respondents who did not complete post-secondary education, 236 respondents supported banning abortion, and 752 respondents opposed banning abortion. Compared to the 1243 respondents who completed post-secondary education, out of which 177 respondents supported banning abortion, and 1066 respondents opposed banning abortion. In

both groups, the majority of respondents (76.11% and 85.76%) did not support banning abortion and less than a quarter (23.89% and 14.23%) supported banning abortion. However, a slightly higher proportion of respondents who completed post-secondary education (85.76%) opposed banning abortion than respondents who did not complete post-secondary education (76.11%). The difference in proportions seems to support the expectation that people who completed a higher level of education would hold a more negative attitude towards banning abortion. But whether the difference is statistically significant needs further analysis.

Education Category	Number in Total	Number Pro-Banning	Number Anti-Banning	% Pro-Banning	% Anti-Banning
Not completed post-secondary education	988	236	752	23.89	76.11
Completed post-secondary education	1243	177	1066	14.24	85.76

Table 2: Number and proportion of respondents support/do not support banning abortion

Figure 2 is a graphical representation of the numerical summaries displayed in *Table 2*. The bar plot on the left corresponds to respondents who did not complete post-secondary education. The bar plot on the right corresponds to the group of respondents who completed post-secondary education.

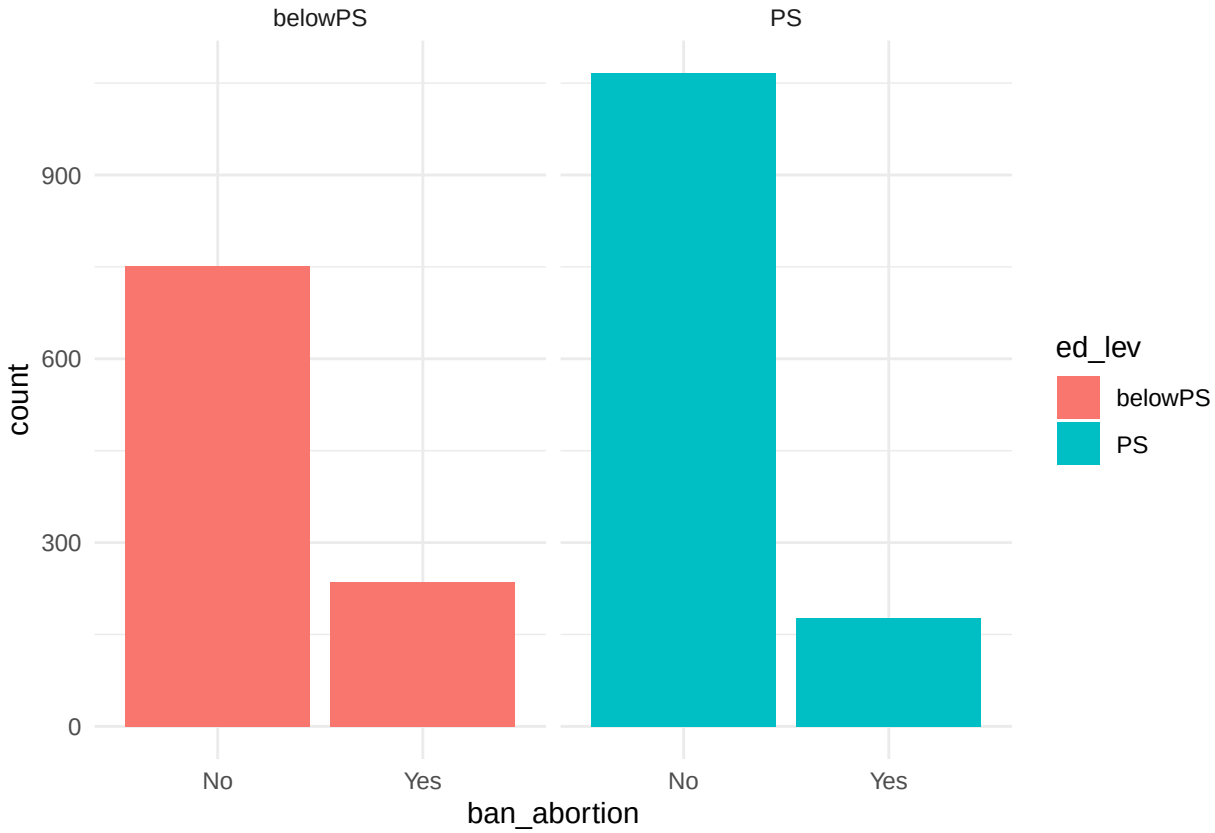

Figure 2: Attitude Towards Abortion by Level of Education

Plots showcasing distributions of 2231 respondents over gender, province, religiosity and place of living are included in **Appendix: Supplementary Plots**. Female and male respondents are relatively equally

represented in the sample, with approximately 200 more female respondents than male respondents. The majority of the respondents reside in Ontario and Quebec, followed by British Columbia. The majority of respondents (over 1500) live in an urban area, and less than 500 respondents live in a rural area. As for the importance of religion, the categories "somewhat important" and "not very important" see most and least respondents, respectively. The remaining respondents are distributed roughly equally into the categories "very important" and "not at all important".

The dataset is reduced to 1976 observations after performing propensity score matching. The $1976 = 988 \times 2$ observations are resulted from matching 1243 respondents who completed post-secondary education to 988 respondents who did not complete post-secondary education, and 255 respondents are unmatched. A more detailed description of matched data is included in the **Methods** section. A glimpse of the matched data is also included in **Appendix A2: Materials**.

All analysis for this report was programmed using `R version 4.0.4`.

Methods

To answer the research question "Does a higher level of education *cause* a more negative attitude towards banning abortion?", causal inference will be performed on the `CES_new` data. In this case, completing post-secondary education can be seen as the treatment. Firstly, propensity score matching is carried out to identify groups of respondents that are similar in terms of the province of living, gender, religiosity and place of living (urban/rural). These covariates are included in the propensity model because a) previous research (Barkan, 2014) on abortion found that gender and religiosity act interactively to affect attitude towards abortion and b) intuitively, province and place of living can be correlated with education and people's political attitude. Thus, these covariates need to be controlled for us to be able to attribute the difference in attitude towards abortion to education alone. However, since 1243 respondents completed post-secondary education (treated) but only 988 respondents who had not completed post-secondary education (untreated), it is not possible to pair every treated observation with an untreated observation. Thus, there will be a data reduction and the size of the matched pairs will be restricted by the number of untreated observations.

Propensity score of a single observation can be estimated by the logistic regression model:

$$log(\frac{P(T_i = 1)}{P(T_i = 0)}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4}$$

(Taback, 2019) where T_i is an indicator of whether the i^{th} observation is treated (i.e. $T_i = 1$ indicates respondent i completed post-secondary education and otherwise $T_i = 0$). X_k for $k \in 1, 2, 3, 4$ are covariates corresponding to province of living, gender, religiosity and place of living, respectively. The parameters β_k for $k \in 1, 2, 3, 4$ represents the change in log odds result from shifting to the current level of X_k and β_0 is the intercept. To obtain estimated propensity scores, the `predict` function is called on the logistic regression model (propensity score model) with parameter `type` set to `"response"`. The vector of propensity scores is then concatenated to the `CES_new` data by mutating a new variable whose values are the propensity score of each observation. Then, the `CES_new` data is arranged in ascending order of the propensity scores using `arrange` function such that the observation with smallest propensity of being treated is at the top. Ordering the data by propensity score is an intermediate step that facilitates visual assessment of any possible pairs. In addition, the indicator variable for completion of post-secondary education with categories 0 and 1 is convert to integer values 0 and 1 using `as.integer` function.

When performing propensity score matching, a vector of matched indices is obtained via the `matching` function from the `arm` package (Hill & Su) in `r`. The input vectors to the `matching` function are the indicator variable and the estimated propensity scores. The parameter `replace` of the `matching` function is set to `FALSE` by default such that each untreated observation is paired to only one treated observation. The vector of matched indices is then concatenated to the `CES_new` data using `r` function `cbind`. Finally, the matched data is created by filtering observations that are paired with another observation. This is equivalent to filtering out observations whose matched indices are 0 or NA using the `filter` function. Variables inherit from the `matching` function, including `match.ind`, `pairs`, and `cnts`, as well as the treatment indicator variable, are

excluded from the matched data using the `select` function. Thus, the matched data contain $2 \times 988 = 1976$ observations and is stored into a data frame called `CES_matched`.

After the respondents are matched for propensity scores, a two-sample t-test is carried out to test whether the difference in proportions of respondents who oppose banning abortion in the two groups is statistically significant. The null hypothesis is that the *difference in proportions is 0* and the alternative hypothesis is that the *difference in proportions is not 0*. An indicator variable for whether a respondent opposes banning abortion with integer values 1 and 0 is created. The indicator variable evaluates to 1 if a respondent answers that abortion should *not* be banned and 0 otherwise. The creation of the indicator is done via the `mutate` function and the `ifelse` function. In addition, the matched data is separated into two groups, both of size 988, with one group being only the respondents who did not complete post-secondary education and the other being only the respondents who completed post-secondary education. The separation of treatment and control groups is done by filtering observations with the corresponding category of education using the `filter` function. Hence, the proportion of respondents who oppose banning abortion is the mean of the indicator (i.e proportion of 1's) for each group.

In the two-sample t-test, we assumed that the two groups a) are independent of each other and b) have equal variances. Intuitively, the independence assumption is met because the two groups consist of different respondents. And since the distribution of abortion attitudes for the two groups are similar (see **Results** section), the equal variance assumption also holds.

Finally, regression analysis is performed on the `CES_matched` data to assess whether the level of education is a significant predictor of the attitude towards banning abortion. Since a respondent's answer to whether abortion should be banned is binary (yes/no), logistic regression would be appropriate. The response is a respondent's attitude towards banning abortion and the predictors are a respondent's education and other covariates. Covariates including province and place of living, gender and religiosity of a respondent are also included in the logistic regression model because, in addition to education, we are also interested in identifying other variables that influence abortion attitudes. The theoretical logistic regression model for the i^{th} observation is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i}$$

where p is the probability that a respondent believes abortion *should* be banned and thus $1 - p$ is the probability that a respondent believes abortion *should not* be banned. X_k for $k \in 1, 2, 3, 4, 5$ corresponds to the province of living, gender, the importance of religion, education and place of living of the i^{th} respondent. The parameters β_k for $k \in 1, 2, 3, 4, 5$ represent the change in log odds by switching from the base level of X_k to the current level and β_0 is the intercept. Significant predictors are identified based on p-values based on the idea that predictors with smaller p-value tend to have greater significance.

Results

After propensity score matching, respondents in both education categories share similar distributions of covariates. Graphical representation showcasing the distribution of covariates are included in **Appendix: Supplementary Plots**.

A numerical summary of the number and proportion of respondents in each education category who support or oppose banning abortion after being matched for propensity score is displayed in *Table 3*. The matched data contain 988 respondents who completed post-secondary education (treatment group) and 988 respondents who did not complete post-secondary education (control group). The outcome for the control group stays the same as before propensity score matching (PSM) since every respondent in the control group is matched with a respondent in the treatment group. Thus the control group after PSM is the same as before PSM. The outcome for the treatment group after PSM, although different in exact numbers and proportions due to a size reduction after matching, retains a similar pattern as before PSM. Out of 988 respondents in the treatment group, 17.11% supported banning abortion which is 2.87% higher than before PSM (14.24%); 82.89% opposed banning abortion, compared to 85.76% before PSM.

Figure 3 is a graphical representation of outcomes displayed in *Table 3*. In both the treatment group and

the control group, the majority of respondents opposed banning abortion, which is consistent with the outcome before PSM. Also consistent with the previous outcome, a higher proportion of respondents in the treatment group (82.89%) opposed banning abortion than in the control group (76.11%). But, the difference in proportion after PSM is not as significant as before PSM (6.78% and 9.65%, respectively).

Education Category	Number in Total	Number Pro-Banning	Number Anti-Banning	% Pro-Banning	% Anti-Banning
Not completed post-secondary education	988	236	752	23.89	76.11
Completed post-secondary education	988	169	819	17.11	82.89

Table 3: Number and proportion of respondents support/do not support banning abortion after PSM

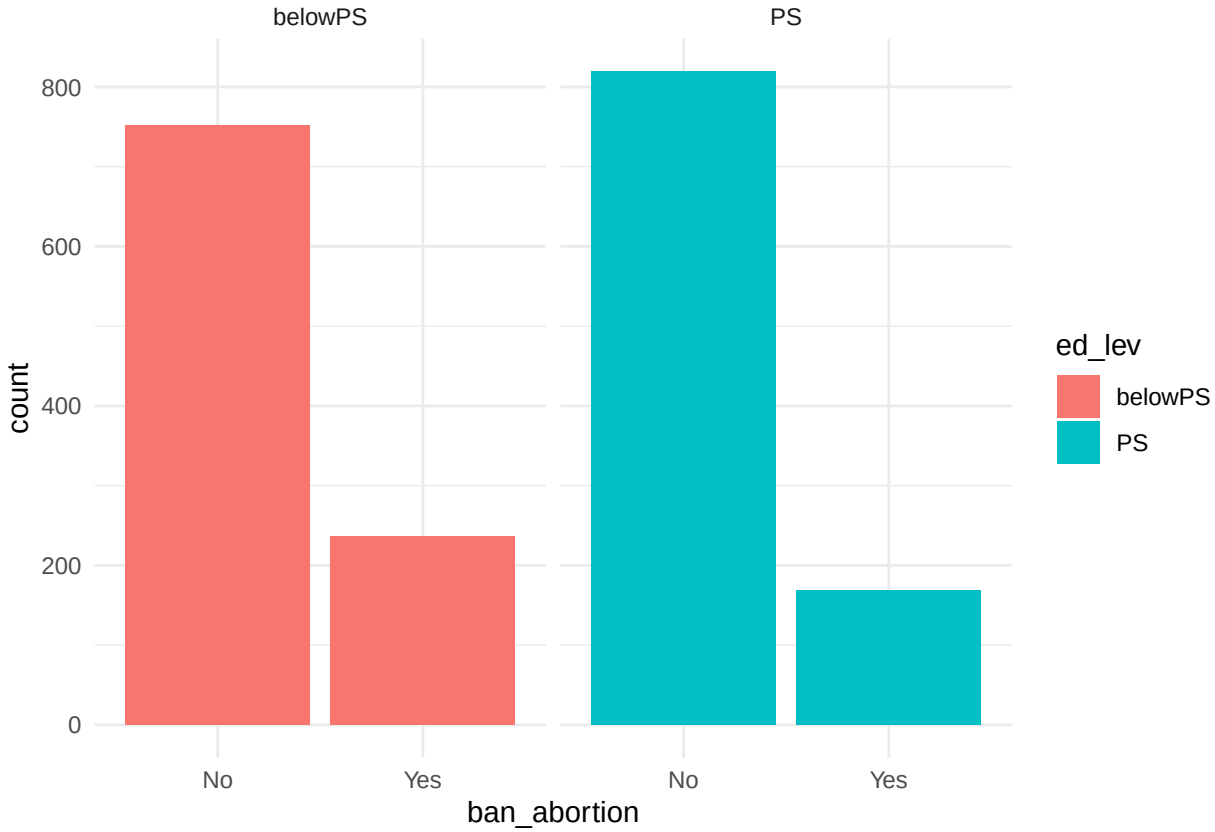

Figure 3: Attitude Towards Abortion by Education After PSM

The two-sample t-test at 0.05 level of significance returns a p-value of 1.85×10^{-4}. The p-value indicates there is a probability of 1.85×10^{-4} of observing a difference as or more extreme than 6.78%, providing strong evidence *against* the null hypothesis, which hypothesize that the proportions of respondents oppose banning abortion in the treatment and control groups are the same. The 95% confidence interval is (0.0323, 0.1033), meaning that if the sample were to be recollected and proportions recalculated infinitely many times, the proportion of treated respondents that oppose banning abortion would be between 3.23% to 10.33% higher than that of untreated respondents.

The results of logistic regression analysis are displayed in *Table 4*. Categories that are not listed in *Table 4*, namely, residing in *Alberta*, being a *female* respondent, attributing *no importance* to religion, *not completing* post-secondary education and living in *rural* area are treated as the base level in logistic regression. Thus, all results from the regression analysis are in comparison to these categories. The logistic regression shows that the treatment condition is a strong predictor of attitudes toward banning abortion (p-value=0.0004). Completing post-secondary education is associated with, on average, a 0.45 reduction in the log odds of answering *Yes* to the question "Should abortion be banned?".

Based on p-values, several categories of the covariates also strongly predict abortion attitudes. The log odds of supporting banning abortion of respondents who are British Columbia residents (p-value=0.0131), Ontario residents (p-value=0.0530) and Quebec residents (p-value=0.0197) is, on average, 0.87, 0.56 and 0.71 smaller than that of respondents who reside in Alberta, respectively. Male respondents (p-value=0.0044) have an average log odds of supporting banning abortion that is 0.37 higher than female respondents. In addition, respondents who think of religion as somewhat important (p-value=1.13e-05) and very important (p-value<2e-16) have an average log odds of supporting banning abortion that is 1.19 and 2.93 higher than respondents who attribute no importance to religion, respectively.

Predictor	Estimate	P-value
(Intercept)	-2.47625	1.05e-10
Province = BC	-0.87089	0.013113
Province = MB	0.35586	0.317572
Province = NB	0.13431	0.733144
Province = NL	-0.30731	0.454342
Province = NS	-0.05909	0.880075
Province = ON	-0.55532	0.052969
Province = PE	-0.07107	0.849049
Province = QC	-0.70521	0.019730
Province = SK	-0.12061	0.742909
Gender = Male	0.37138	0.004382
Religiosity = notvery	0.53507	0.121801
Religiosity = somewhat	1.19381	1.13e-05
Religiosity = very	2.92907	< 2e-16
Education = PS	-0.45396	0.000359
Place of living = urban	-0.19678	0.174550

Table 4: Logistic Regression Results

Conclusions

Concerning the pro-choice versus pro-life debate on abortion, we are interested in identifying factors that influence one's attitude towards banning abortion. Demographic summaries on the topic show that a higher level of education is associated with a more pro-choice attitude towards abortion. In other words, people who completed a higher level of education seem to hold a more negative attitude toward banning abortion. In this report, we investigated through a statistical lens the causal relationship between people's education and their attitudes toward banning abortion. We hypothesized that people who complete a relatively higher level of education hold a more negative attitude towards banning abortion. In other words, we expected that a higher level of education would cause a pro-choice attitude towards abortion. To be able to infer causation on observational CES2011 survey data, 2231 respondents are matched for their propensity of completing post-secondary education (i.e. being treated) based on their provinces and places of living, gender, and perceived importance of religion. The estimated propensity score of each respondent is obtained through logistic regression on the covariates. Each of the 988 untreated respondents is paired to a treated respondent, resulting in 1976 respondents after PSM. PSM ensures the similarity of treated and untreated respondents in

terms of covariates, thus any differences in attitude towards abortion between the two groups of respondents are explained solely by the treatment.

To test the initial hypothesis, a two-sample t-test and then a logistic regression analysis is performed on the matched data. From numerical and graphical summaries of the matched data, a larger proportion of respondents in the treatment group opposed banning abortion than in the control group. Thus, a two-sample t-test was carried out to test whether this difference in the proportions is statistically significant. The null hypothesis is that the proportions in two groups are the same, versus the alternative hypothesis which claims the difference between two proportions. At 0.05 level of significance, the two-sample t-test rejects the null hypothesis. A 95% confidence interval also suggests that the proportion of treated respondents who oppose banning abortion would be between 3.23% and 10.33% higher than that of untreated respondents 95% of the time. Thus, the results from the two-sample t-test confirm our expectation that respondents who completed post-secondary education are more likely to hold a negative attitude toward banning abortion.

In addition, a logistic regression model is fit on the matched data where the response variable is respondents' yes/no answers to the survey questions "should abortion be banned?", and the predictors are the treatment and the covariates. In doing so, we aim to a) estimate the effect size of the level of education on attitudes toward banning abortion and b) determine whether the covariates predict attitudes toward banning abortion. The results from the logistic regression suggest that completing post-secondary education is associated with a smaller log odds of supporting banning abortion. In other words, respondents who completed post-secondary education have a larger odds of opposing banning abortion, which supports our hypothesis. In addition to the treatment, living in British Columbia, Ontario and Quebec associate with a smaller log odds of supporting banning abortion; and perceiving religion as somewhat important or very important and identifying as male is associated with a higher log odds of supporting banning abortion.

To conclude, results from a two-sample t-test and logistic regression on matched data support our hypothesis that a higher level of education predicts a more negative attitude towards banning abortion. CES respondents who completed post-secondary education are more likely to answer *No* to the question "should abortion be banned?" and the proportion of respondents who completed post-secondary education and oppose banning abortion is statistically significantly higher than that of respondents who did not or have not yet completed post-secondary education. Since covariates are controlled through propensity score matching, we can attribute the difference in attitudes toward abortion between treated and untreated respondents to the treatment. Hence, we conclude that a higher level of education *causes* a more negative attitude towards banning abortion.

Limitations

In this analysis, a causal inference was performed on observational data. The specific approach that we took was to use propensity score matching to mimic the effect of randomization as in experimental trials so that the treatment group and control group are similar in terms of covariates. The causal inference relies on the definition of the propensity score, which is the probability that an observation is treated given all *observed* covariates and this implies that observations with similar propensity scores have similar distributions of observed covariates (Taback, 2019). However, this also implies that observations in the treatment and control groups may differ in *unobserved* covariates that are not controlled for via PSM. Thus, the difference in attitudes toward banning abortion between respondents who completed post-secondary education and those who did not complete post-secondary education can be due to unobserved covariates rather than their level of education. Covariates that have the potential to influence the results include a respondent's age and family members' attitudes towards banning abortion. Intuitively, we expect that older adults might hold a more conservative attitude towards abortion and thus are less likely to oppose banning abortion; or that a respondent whose parents or spouse strongly oppose or support banning abortion would hold a similar attitude as their family members.

Furthermore, PSM is subject to data reduction. Since the sizes of the treatment group and control group are different (1243 and 988, respectively), the size of matched data is restricted by the group with a smaller size because each observation in the control group pairs with only one observation in the treatment group and vice versa. Thus, the dataset is reduced from 2231 observations to 1976 observations after PSM. The information on the remaining 255 respondents (11.43% of the data) is lost. The inevitable data reduction can

also influence the reproducibility of the results. If the exact analysis is to be performed again with different data, the reproduced results can be biased if the number of observations in the treatment and control groups is significantly different.

Lastly, CES2011 data on attitudes towards abortion is used because of its openness and accessibility. The dataset is built into `r` package `carData` and thus is conveniently loaded into `r` and can be analyzed with relatively little cleaning which saves both time and effort. The downside, however, is that the collection of data was done ten years ago relative to the time this analysis is performed. Hence, though providing some insight, the results from this analysis might not be generalizable to today's society. Considering progress in sexual education, female rights movements and changes in related policies over the past ten years, the results from the analysis on CES2011 data may or may not be representative in 2021.

Next Steps

Future studies and analyses on the topic may target controlling for the limitations mentioned in the previous section. Researchers can perform statistical analysis on a more recent dataset or design and conduct a new survey on attitudes towards abortion. To avoid waste of data associated with PSM, other causal inference methods such as regression discontinuity design, propensity score stratification (Taback, 2019) and difference in difference method may be appropriate depending on the survey design. In addition, the design of the survey or choice of data may consider including more demographic variables that take the same value regardless of whether a respondent is treated (i.e. covariates) to reduce bias in causal inference.

Discussion

In conclusion, this analysis investigates the causal relationship between the level of education and attitudes towards abortion. Results suggest that people who complete a relatively higher level of education are more likely to disapprove of banning abortion. A higher proportion of CES2011 survey respondents who completed post-secondary education think that abortion should not be banned, and a respondent in that group also had a greater odds to oppose banning abortion. Overall, these results are in line with our hypothesis. The most insightful result, however, was that regardless of whether post-secondary education is completed or not, the majority of respondents do believe that abortion should remain legally accessible. To end this report, we hope that the presented results can shed light on future works on the topic and in related fields.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
4. BBC. (2019, June 14). *What's going on in the fight over US abortion rights?* BBC News. Retrieved December 3, 2021, from https://www.bbc.com/news/world-us-canada-47940659.
5. Barkan, S. E. (2014). *Gender and abortion attitudes: Religiosity as a suppressor variable.* Public Opinion Quarterly, 78(4), 940–950. https://doi.org/10.1093/poq/nfu047
6. Gallup. (2021, November 20). *"Pro-choice" or "Pro-life," 2018-2021 demographic tables.* Gallup.com. Retrieved December 19, 2021, from https://news.gallup.com/poll/244709/pro-choice-pro-life-2018-demographic-tables.aspx
7. R Documentation. (n.d.). *2011 Canadian National Election Study, With Attitude Toward Abortion.* R. Retrieved December 14, 2021, from https://vincentarelbundock.github.io/Rdatasets/doc/carData/CES11.html.
8. Northrup, D. (2012, June). *The 2011 Canadian Election Survey, Technical Documentation.* Toronto; Institute for Social Research, York University.
9. Taback, N. (2019). *Design of Experiments and Observational Studies.*
10. Hill, J., & Su, Y.-S. (n.d.). *Single Nearest Neighborhood Matching.* Retrieved December 16, 2021.

Appendix

A1: Ethics Statement

In constructing this report, we are aware of and are consciously avoiding publication bias. All statistical results from logistic regression and two-sample t-test are presented in the exact values as in the regression or hypothesis test r output. Values of regression estimates, p-values and confidence interval boundaries have not been manipulated intentionally for obtaining a more statistically significant result. The form and values of CES2011 data extracted from r package carData are preserved and not manipulated except for minor cleaning.

In addition, background information, data information, statistical methods and concepts that are used or presented in this report are based on evidence from reliable sources including peer-reviewed journals, news articles, CES2011 documentation, r documentation and materials used for university-level statistical courses. All ideas and evidence from these sources are properly paraphrased and cited in APA format including in-text citations and a bibliography at the end of the report. No ideas from external sources are used without citation.

A2: Materials

The first 6 rows of the cleaned data `CES_new` are displayed in the following table:

id	province	gender	religiosity	urban	ed_lev	post_s	ban_abortion
2851	BC	Female	somewhat	urban	belowPS	0	No
521	QC	Male	not	urban	PS	1	No
2118	QC	Male	somewhat	urban	PS	1	Yes
1815	NL	Female	very	urban	belowPS	0	No
1799	ON	Male	not	rural	PS	1	No
1103	ON	Female	not	urban	PS	1	No

The first 6 rows of the matched data `CES_matched` are displayed in the following table:

id	province	gender	religiosity	urban	ed_lev	ban_abortion	propensity_post_s
3007	NS	Male	very	rural	belowPS	Yes	0.305831
1830	NS	Male	very	rural	belowPS	Yes	0.305831
2888	NS	Male	very	rural	belowPS	Yes	0.305831
563	NS	Male	very	rural	belowPS	No	0.305831
1641	NS	Male	very	rural	belowPS	No	0.305831
1837	NS	Male	very	rural	belowPS	No	0.305831

Supplementary Plots

Figure 4: Distribution by Gender, Province, Religiosity and Place of Living

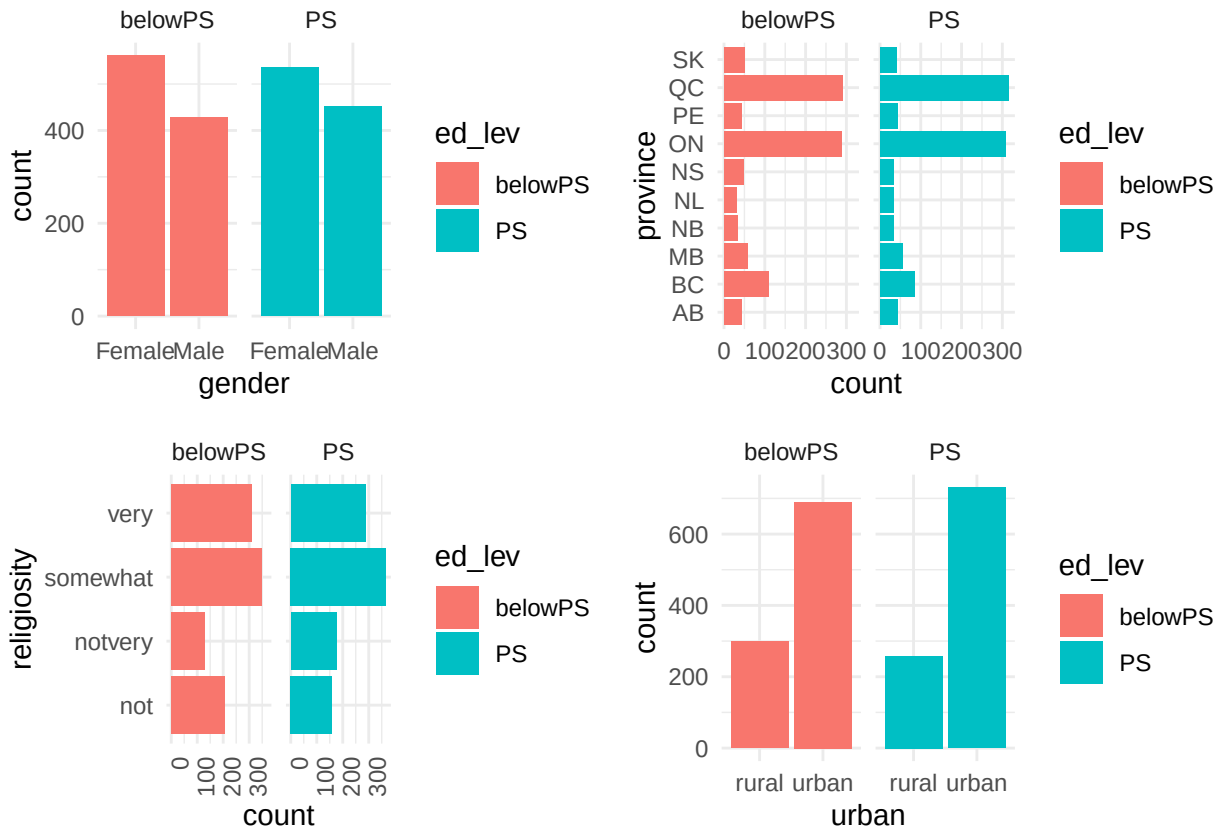

Figure 5: Distribution of Covariates after PSM