Paper AS-216-2023

D-I-D the policy have an impact?

Difference-in-difference methods applied to survey data in SAS®

Melanie Dove, University of California, Davis

ABSTRACT

PROC SURVEYLOGISTIC was used to conduct a difference-in-difference (DID) analysis to compare how youth e-cigarette use changed pre-policy to post-policy and between students exposed to a flavored tobacco sales restriction relative to unexposed students. To obtain the DID odds ratio, an interaction term between the year and policy group was included in the model. This paper will explore the options needed to perform this type of logistic regression modelling.

INTRODUCTION

It can be challenging to examine the impact of a policy, as a randomized control trial usually cannot be conducted due to ethical or feasibility concerns or the policies may already exist. Difference-in-difference (DID) methods can be used to examine the impact of policies or programs by comparing changes in outcomes over time between a population exposed to a policy and a population that is not (Caniglia 2020, Wing 2018). These methods are commonly used with survey data, as survey data is usually collected on a regular basis over time. In this paper, we will describe how we used DID methods to examine the impact of a policy to ban flavors from tobacco products on youth e-cigarette use in SAS®, using data from the California Healthy Kids Survey. Additional details about the methods and results of this analysis are available in the published paper (Dove 2023).

DESCRIPTION OF THE DATA AND VARIABLES

CALIFORNIA HEALTHY KIDS SURVEY

The California Healthy Kids Survey (CHKS) is a comprehensive whole child, school climate, and youth risk behavior data collection service available to all California local education agencies, and is funded by the California Department of Education. Students attending middle and high schools (grades 7, 9 and 11) in California completed an in-person survey during the academic school year from September to June. Although most schools participate once every two years, schools are on different two-year cycles so that data is collected every year. Participation is voluntary, confidential, and passive parental consent was obtained.

MAIN VARIABLES

The three main variables used in a DID analysis are the outcome, policy (or intervention) group, and time. An interaction term between policy and time is used to obtain the DID estimate, as shown in the following equation:

$$Y = ß_0 + ß_1 Policy + ß_2 Time + ß_3 Policy*Time$$

In this example, current (past month) e-cigarette use is the outcome variable (Y). Exposure to a flavored tobacco sales restriction is the policy variable, and year is the time variable. For the year variable, 2019/20 is compared to 2017/18, as the policies were implemented in 2018/19. All variable categories are coded as 0/1, with 0 being the referent group. The names and categories of the three variables are listed below:

- Year: 0=2017/2018, 1=2019/2020
- Policy: 0=no policy, 1=policy
- Ecig: 0=no, 1=yes

DESCRIPTIVE ANALYSIS USING PROC SURVEYFREQ

To adjust for the fact that students are clustered within schools, SAS® survey procedures are used to analyze the data. The CLUSTER statement is included with the variable for each individual school (variable name = school). The CHKS data were not stratified or weighted, so we do not include the STRATA or WEIGHT statements. More information on how to analyze survey data in SAS® is available in a separate paper (Dove 2020).

PROC SURVEYFREQ is used to estimate the percent of students who used e-cigarettes in each year and in each policy group, using the following code:

```
proc surveyfreq data=chks;
cluster school;
tables year* policy *ecig/ row cl;
run;
```

In the TABLES statement, the row percent is requested using the option 'row' and the 95% confidence intervals around the row percent are requested using the option 'cl'. Below is a table of the results:

Policy	Pre-policy (2017/18) Percent (95% CI)	Post-policy (2019/20) Percent (95% CI)
Yes	10.5 (7.7, 13.3)	11.1 (9.2, 13.1)
No	12.8 (11.2, 14.5)	11.4 (10.2, 12.7)

Table 1. The percent (95% CI) of high school students who used e-cigarettes by year and policy group

As shown in Table 1, e-cigarette use did not substantially change from pre- to post-policy for students with a policy (10.5% to 11.1%), and slightly decreased among students without a policy (12.8% to 11.4%). The next section will cover how to statistically test if there are pre- to post-policy differences in e-cigarette use within and between these two policy groups.

DID ANALYSIS USING PROC SURVEYLOGISTIC

PROC SURVEYLOGISTIC is used to obtain the following three estimates:

1. The odds ratio comparing the pre- to post-policy odds of e-cigarette use for students *without* a policy
2. The odds ratio comparing the pre- to post-policy odds of e-cigarette use for students *with* a policy
3. The DID odds ratio comparing the odds ratios from #2 to #1, indicating if there is a pre- to post-policy difference in e-cigarette use in students exposed to a policy relative to unexposed students.

More specifically, in #3 we are testing the following null and alternative hypothesis in this example:

H_0: There is *no* association between flavored tobacco sales restrictions and e-cigarette use

H_1: There is an association between flavored tobacco sales restrictions and e-cigarette use

If the p-value is <0.05 than we will reject the null hypothesis and conclude that there is a statistically significant association between flavored tobacco sales restrictions and e-cigarette use.

The odds ratio provides an estimate of the magnitude of the association between the exposure and outcome. Odds ratios greater than 1 indicate that there is a positive association between the exposure and outcome and odds ratios less than 1 indicate that there is a negative association. An odds ratio equal to 1 suggests that there is no association between the exposure and outcome. The odds ratio can be obtained by exponentiating the regression coefficients (labeled 'Estimates' in the Displays) from a logistic

regression model. The following SAS code requests the odds ratios from #1-3 above:

```
proc surveylogistic data=chks;
 cluster school;
 model ecig (event='1')= year policy  policy*year/ expb;
 estimate '2017.2018 vs. 2019.2020 no policy' year 1/exp cl e;
 estimate '2017.2018 vs. 2019.2020 policy' year 1 policy*year 1/exp cl e;
 estimate 'DID odds ratio' policy*year 1/exp cl e;
run;e
```

The MODEL statement includes the outcome (ecig), year, and policy variables and their interaction (policy*year). The '(event='1')' option is used to model the probability that students are using e-cigarettes (ecig=1). The 'expb' option requests the odds ratio, as shown in the Exp(Est) column in Display 1 below.

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Exp(Est)
Intercept	-1.9149	0.0736	-26.02	<.0001	0.147
year	-0.1325	0.0539	-2.46	0.0162	0.876
policy	-0.2259	0.1660	-1.36	0.1776	0.798
year*policy	0.1974	0.1316	1.50	0.1378	1.218

NOTE: The degrees of freedom for the t tests is 78.

Display 1. Output from the SURVEYLOGISTIC procedure

Because there is an interaction term in the model, the regression coefficients (labeled 'Estimates') and odds ratios (labeled Exp(Est)) in Display 1 cannot be directly interpreted. This is because the interaction term specifies that there are two different pre- to post- policy odds ratios – one for students with a policy and one for students without a policy. Below is a description for how to obtain these two odds ratios, as well as the DID odds ratio.

1. PRE- TO POST POLICY ODDS RATIO FOR STUDENTS WITHOUT A POLICY

The '**year'** regression coefficient in Display 1 (-0.1325) compares the odds of e-cigarette use in 2019/2020 to the odds of e-cigarette use in 2017/2018 for students in the policy referent group (policy =0, those without a policy). The odds ratio can be obtained by exponentiating the 'year' regression coefficient: exp (-0.1325)=0.876, also shown in the Exp(Est) column.

ESTIMATE statements provide an easier way of obtaining odds ratios and 95% confidence intervals. The first ESTIMATE statement in the code above (labeled 2017.2018 vs 2019.2020 no policy) requests this specific odds ratio and 95% confidence interval. Because the calculation only includes the regression coefficient from the 'year' variable, only 'year' was included in the statement after the label. The 'exp' and 'cl' options request the odds ratio and 95% confidence limits. The 'e' option requests the Estimate Coefficients table that shows which variable's regression coefficients are included in the odds ratio calculation. This provides a way to check that the code for your ESTIMATE statements is correct.

The output from the first ESTIMATE statement is shown below in Display 2. As shown in the Estimate Coefficients table, only the regression coefficient from the 'year' variable is included in the odds ratio calculation. The odds ratio is shown in the column labeled 'Exponentiated', and the 95% confidence intervals are shown in the next 2 columns. The results indicate that there is a significant pre- to post-

policy decrease in the odds of using e-cigarettes among students without a policy (OR=0.88, p=0.016).

Estimate Coefficients	
Parameter	Row1
Intercept: ecig=0	
year	1
policy	
year * policy	

Estimate											
Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
2017.2018 vs. 2019.2020 no policy	-0.1325	0.05390	78	-2.46	0.0162	0.05	-0.2399	-0.02522	0.8759	0.7867	0.9751

Display 2. Output from the first ESTIMATE statement: pre- to post-policy odds ratio for students without a policy

2. PRE- TO POST-POLICY ODDS RATIO FOR STUDENTS WITH A POLICY

The regression coefficients for **year** and **year*policy** (in Display 1) are used to get the pre- to post-policy odds ratio for students with a policy: exp (-0.1325+ 0.1974) = 1.067. This odds ratio is not shown in Display 1. The second ESTIMATE statement (labeled '2017.2018 vs. 2019.2020 policy') requests this odds ratio and the 'year' variable and interaction term (year*policy) are included in the statement. The Estimate Coefficients table below in Display 3 confirms that these two variables are used to calculate the odds ratio. The results indicate that there is not a significant difference in the pre- to post-policy odds of e-cigarette use among students with a policy (OR=1.07, p=0.59).

Estimate Coefficients	
Parameter	Row1
Intercept: ecig=0	
year	1
policy	
year * policy	1

Estimate											
Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
2017.2018 vs. 2019.2020 policy	0.06487	0.1201	78	0.54	0.5907	0.05	-0.1742	0.3040	1.0670	0.8401	1.3552

Display 3. Output from the second ESTIMATE statement: pre- to post-policy odds ratio for students with a policy

3. DID ODDS RATIO

The regression coefficient for **year*policy** in Display 1 is used to get the DID odds ratio comparing the pre- to post-policy change in e-cigarette use for students exposed to a policy relative to students not exposed to a policy: exp(0.1974) = 1.218. The two odds ratios calculated in the previous sections can also be compared to obtain the DID odds ratio (1.067/ 0.8759 = 1.218). The third ESTIMATE statement (labeled 'DID odds ratio') requests this odds ratio and the interaction term is included after the label. From the result listed in the 'Exponentiated' column in Display 4 below, there is not a significant difference in the pre- to post-policy odds of e-cigarette use in students with a policy relative to those without a policy (OR=1.22, p=0.14). Since the p-value is >=0.05, the null hypothesis (from the beginning of this section) is accepted, and we conclude that there is no association between flavored tobacco sales restrictions and

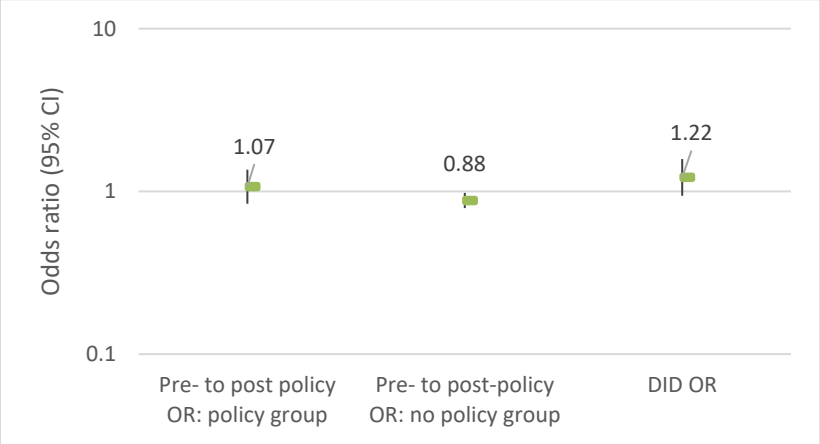
e-cigarette use.

Estimate Coefficients	
Parameter	Row1
Intercept: ecig=0	
year	
policy	
year * policy	1

Estimate											
Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
DID odds ratio	0.1974	0.1316	78	1.50	0.1378	0.05	-0.06467	0.4595	1.2182	0.9374	1.5832

Display 4. Output from the third ESTIMATE statement: DID odds ratio

In the figure in Display 5, the pre- to post-policy odds ratios for students with (1.07) and without a policy (0.88) are shown. The DID odds ratio (1.22) indicates that these two odds ratios are not statistically different, suggesting that there is not an impact of the policy on e-cigarette use.



Display 5. Pre- to post-policy odds ratios for students with and without a policy and the DID odds ratio

OPTIONS

Adding covariates to the model

One of the advantages of DID methods is that they account for bias from potential confounding in two ways. First, they account for characteristics that differ between the two policy groups and do not change over time, such as race/ethnicity, gender, or education. Second, they account for characteristics that change over time, but impact the outcome in the same way for the two policy groups. For example, a policy that increases the tax on e-cigarettes may impact e-cigarette use the same for the two policy groups. The DID model allows for the adjustment of covariates on the model statement, as shown below with variables for income and population density:

```
model ecig (event='1')= year policy income popdensity policy*year/ expb;
```

Policy variable with more than two categories

DID methods can be used if the policy variable has more than two categories. In this new example, the policy variable (called policy2) has three categories of strong policy, medium policy, and no policy, coded as follows:

- Policy: 0=no policy, 1=medium policy, 2=strong policy

In this example, the first three ESTIMATE statements calculate the pre- to post-policy odds ratios for each of the three policy groups, and the next two ESTIMATE statements calculate the DID odds ratios, one for the medium policy group and one for the strong policy (the no policy group is the referent). The numbers after the policy2*year interaction term change to account for the extra policy category. The numbers after the year variable remain the same, as this variable did not change. A CLASS statement is also added so that each policy group will be compared to the referent, specified as the first or no policy group:

```
proc surveylogistic data=chks;
 cluster school;
 class policy2 (ref=first)/param=ref;
 model ecig (event='1')= year policy2 policy2*year/expb;
 estimate '2017.2018 vs. 2019.2020 no policy' year 1 /exp cl e;
 estimate '2017.2018 vs. 2019.2020 medium policy' year 1 policy2*year 1/exp
 cl e;
 estimate '2017.2018 vs. 2019.2020 strong policy' year 1 policy2*year 0 1/exp
 cl e;
 estimate 'DID odds ratio medium vs. none' policy2*year 1/exp cl e;
 estimate 'DID odds ratio strong vs. none' policy2*year 0 1/exp cl e;
run;
```

In addition to a policy variable with more than two categories, DID methods can be implemented in SAS® for a year variable with more than two categories (i.e., more time before and/or after policy implementation) and staggered policy implementation dates. These topics will be explored in a future paper.

ASSUMPTIONS

For the DID method to provide unbiased estimates of the impact of a policy, certain assumptions must be met. One of the main assumptions is referred to as the parallel trends assumption and specifies that the trends in the outcome in those with and without a policy must be similar in the absence of the policy (or before the policy). This assumption can be assessed by graphically examining the pre-policy trends in the outcome in those with and without the policy, statistically testing whether the pre-policy trends differ in those with and without the policy by including an interaction term in a model (i.e., time*policy), or conducting an event-study (Maclean 2020). If there is not any pre-policy data, it is difficult to assess this assumption.

CONCLUSION

DID methods can be implemented in SAS® to determine the impact of a policy. By including an interaction term between time and policy in the model and using ESTIMATE statements, DID odds ratios can be obtained. These DID odds ratio provide the magnitude of the association between a health policy and outcome, provided certain assumptions are met. In our example, we did not find an association between the policy (flavored tobacco sales restrictions) and outcome (youth e-cigarette use).

REFERENCES

Caniglia EC, Murray EJ. Difference-in-difference in the time of cholera: a gentle introduction for epidemiologists. *Curr Epidemiol Rep* 2020; 7: 203–11.

Dove MS and Heck K. 2020. "Survey Data Analysis Made Easy with SAS®". Proceedings of the SAS Global Forum 2020 Conference. Virtual, SAS Institute Inc. Available at https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/4635-2020.pdf

Dove MS, Gee K, Tong EK. Flavored tobacco sales restrictions and teen e-cigarette use: Quasi-experimental evidence from California. *Nicotine and Tobacco Research* 2023; 25: 127-134.

Maclean JC, Halpern MT, Hill SC, Pesko MF. The effect of Medicaid expansion on prescriptions for breast cancer hormonal therapy medication. *Health Serv Res*. 2020; 55: 399-410.

Wing C, Simon KI, Bello-Gomez RA. Designing difference in differences studies: Best practices for public health policy research. *Annual Review of Public Health* 2018; 39: 453-469.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Melanie Dove
University of California, Davis
mdove@ucdavis.edu