Proposition 99 Revisited

Jared Amani Greathouse
jgreathouse3@student.gsu.edu
Policy Data Analysis
Andrew Young School of Policy Studies, Georgia State University

October 8, 2024

1 Introduction and Background

Despite gains in education and awareness, tobacco smoking remains a serious public health problem across the world. Recent estimates suggest that there are around, as of 2020, 880,000 deaths from *secondhand smoke* every year (Yousuf et al., 2020). The same paper cites statistics saying that one billion people continue to smoke globally, with 6 million deaths resulting from tobacco and its related causes every year As a result, since the late 20th century, governments have placed restrictions on tobacco smoking such as public smoking bans, workplace restrictions, and so on. Naturally, this subject remains a popular topic in empirical social science (Flor et al., 2021), since having good estimates of the effect of our anti-tobacco policies is critical to shaping public health policy, tax policy, and other areas of interest. This paper however is not concerned with recent developments, as interesting as those are. Instead, I revisit the original anti-tobacco policy passed by California in 1988, Proposition 99 (P99). My goal is to estimate the counterfactual for California's cigarette sales per capita, or how the cigarette sales trends would look if P99 were never passed.

P99 was the first comprehensive anti-tobacco policy enacted in any state in the United States. It raised the excise taxes on tobacco by 25 cents per pack, which produced a massive windfall in state and local government to market against tobacco usage. As Abadie et al. (2010) note, P99 enabled many indoor clean air ordinances to be passed at local levels. These ordinances were concerned with banning smoking in workplaces, restaurants (by customers), in movie theaters, and in other areas where secondhand smoke was very likely to affect others. Other states soon passed similar policies, such as Massachusetts in 1993 and Arizona in 1994. Despite the success of the results of Abadie et al. (2010), who generally show that P99 was successful in reducing tobacco consumption, this paper re-examines the issue using an updated difference-in-differences (DID) method. This paper offers a useful contribution to the literature on public health, since it uses an updated methodology to inform policymakers and the public on the effects of public health interventions designed to reduce tobacco use. Since lives are at stake with the implementation (or not) of these policies, having the best evidence applied to these questions is a worthwhile task to see if our policies to improve health are sufficient to some acceptable margin.

2 Data

I use the dataset of Abadie et al. (2010). They collect annual data on the average price of tobacco as well as the consumption of cigarettes per year. They define the price of tobacco in cents and the consumption of tobacco as the number of packs bought per the state's population. The reason they use the rate per capita is because it provides a normalized metric by which to compare states with. If we used simply the raw number of cigarettes purchased, California would dominate all other states because of population, making it impossible to meaningfully compare California to other states. Therefore, the per capita rate allows for us to make better between-state comparisons.

I consider a panel dataset of $N + 1$ states, where each state is indexed by i. My unit of time, indexed by t, is the year from 1970 to 2000. My outcome y_{it}, as in Abadie et al. (2010), is the annual per capita packs purchased across all states. The goal of this paper is to see how P99 affected the cigarette consumption trends of California. In this case, $i = 1$ is the state of California (the treated unit or state that did the intervention we are studying) and $i \in \{2 \dots 39\}$ represents the control group of states I consider (or, the ones that never did any such intervention in the study period). Abadie et al. (2010) mention they drop Massachusetts, Arizona, Oregon, Florida, Alaska, Hawaii, Maryland, Michigan, New Jersey, New York, Washington, as well as Washington DC from the control group of states. So as to measure the impact of P99, instead of contaminating the effect size with the effects of the respective policies of other states, I also omit these states from my study. P99 was passed in 1988. However, it did not go into effect until 1989. So, T_0 represents the year of the treatment, 1989, giving us 12 post-intervention periods and 19 pre-intervention periods. Thus, we have a dummy variable $d_{it} \in \{0, 1\}$, which is equal to 1 from 1989 to 2000 if the state is California, else 0.

3 Methods

Our goal is to estimate the counterfactual, or y_{1t}^0. These are the values of tobacco consumption we'd see if P99 was not enacted. The difficulty however is that we can see units as treated or not, $y_{it} = y_{1t}^1 d_{it} + (1 - d_{it})\, y_{1t}^0$. We can't see California both treated under P99 and not at the same time. We must estimate the untreated values, then.

Our statistic of interest is the average of the post-intervention differences between the treated unit and the counterfactual predictions, or $\frac{1}{T_2}\sum_{t=T_0}^{T}(y_{1t}^1 - y_{1t}^0)$, also called the average treatment effect on the treated (ATT).

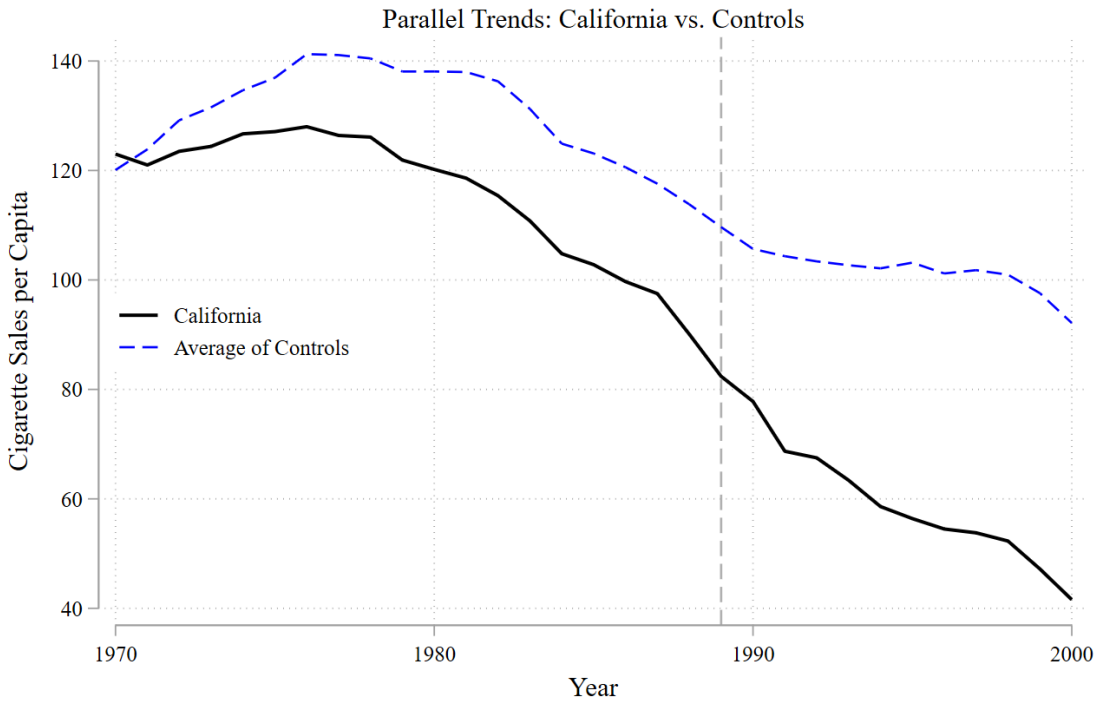

Figure 1: Validating Parallel Trends

I employ the DID method to estimate the causal impact of P99 on tobacco consumption rates. The DID method uses simple linear regression to construct the counterfactual (or, the untreated, post-intervention tobacco consumption values) for California. It proceeds in this manner:

$$y_{1t} = \alpha + \bar{y}_{it} \quad \forall t \in \mathcal{T}_1 \tag{1}$$

where y_{1t} represents the smoking rates for California in the pre-intervention period, $\bar{y}_{it}$ represents the average of control outcomes in the pre-1989 period, α represents the average difference of between the control group and the treated unit's outcomes, and $\mathcal{T}_1$ represents the set of pre-intervention periods. The control group mean is fixed, so we are effectively only estimating the systemic differences across time between the treated unit and the average of controls. More precisely, we are using linear

regression to estimate the value for α which minimizes the prediction errors between the outcomes of our treated unit and the control group's outcomes.

The key identifying assumption of the DID method is the parallel trends assumption. Or, that absent the treatment, the difference between the treated unit and average of controls would be constant. More precisely, since we cannot observe the counterfactual in $t \in \mathcal{T}_2$, we are concerned with the validity of parallel trends in the pre-intervention period. I use a plot to illustrate this. Figure 1 plots the outcome trajectories of California versus the average of 38 controls. We can see that the trends are not exactly parallel. The control group average of tobacco smoking rose consistently until around 1976, whereas California's pre-intervention trends were either very flat or declining. This suggests that using all 38 control units may not be a good control group for California. Specifically, it suggests that other state-specific factors, unobserved to us, are driving the consumption of tobacco to rise at a faster rate than that of California.

Ideally we want the trends of the treated and control units to look as similar as possible in the pre-intervention period. The DID model shifts the average of controls up or down via the α term– because we employ linear regression, DID is the prediction line which best fits the treated unit's outcomes in the pre-intervention period, given our control group. For our purposes, a better fitting DID model in the pre-intervention period means that the trends between the treatment unit and control group average are more parallel to each other.

4 Results

As a first result, let's see how the DID model actually performs in terms of parallel trends using all 38 controls. Figure 2 plots our results. We can see that our suspicions were validated. The DID prediction line does *not* fit California very well in the pre-intervention period. It misses the first few years of California's observed values (trending upwards, as I said above). It misses, crucially, the observed values of California from 1977 to 1988, the decade before the policy was enacted. As a result, we can say that the actual effect of P99 is a lot less than what is implied by the DID model here. The ATT DID estimates is -27.34911, with a 95% Confidence Interval of $[-33.022, -21.675]$. This is an extremely large effect size. Furthermore, the Confidence Interval is very wide, suggesting the true effect of the ATT lies between

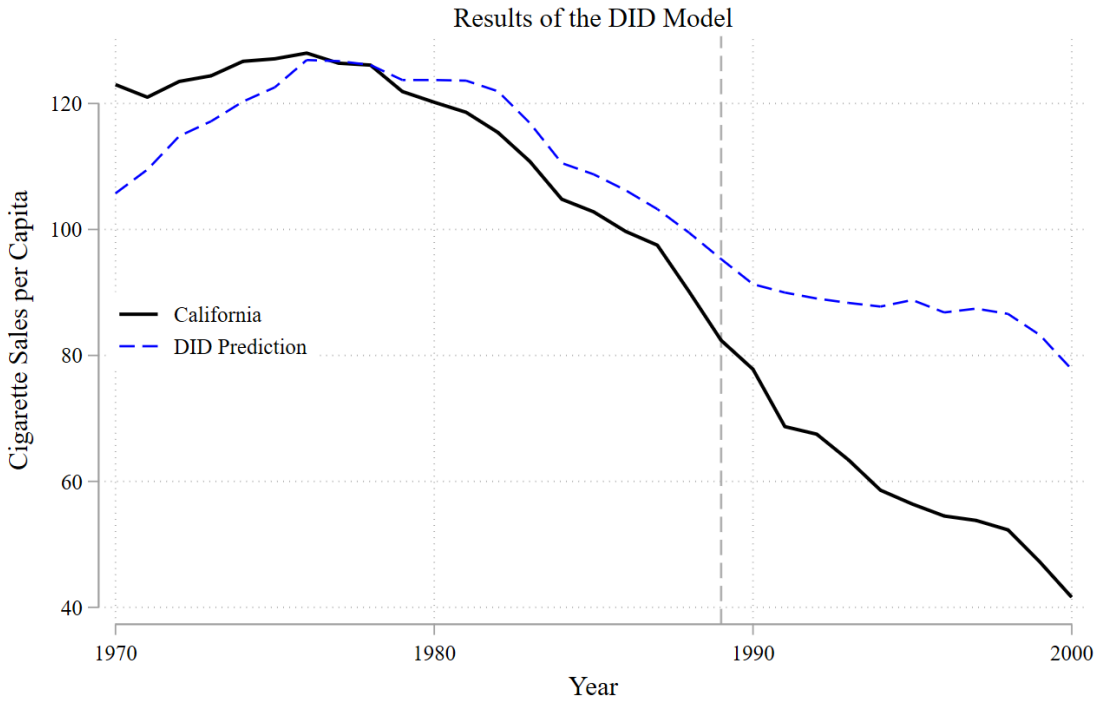

Figure 2: Difference-in-Differences

a decrease of 33 or 21. The t statistic is $|-9.76|$ (much higher than the 1.96 threshold for statistical significance) and the p-value is well below 0. However, I must emphasize that these DID results *are not to be trusted*. Given the apparent violation of parallel trends, the key identifying assumption of DID, we should refine our analysis to ensure that we are comparing the best units to California as possible.

To remedy the apparent parallel trends violation, we can consider one of two methods. Firstly, we can consider a conditional parallel trends assumption, where we add additional covariates to the DID model and presume that parallel trends holds conditional on those covariates. On the other hand, we may consider a simple modification of the DID method. Using the Stata code from Greathouse et al. (2024), I implement the Forward DID method by Li (2024). Interested readers may consult Li (2024) for details on this method, or Greathouse et al. (2024). Basically, the method uses a machine-learning technique called forward-variable selection to choose the control group for DID, then estimates the ATT with only that control group. Specifically, Forward DID chooses Montana, Colorado, Nevada, and Connecticut as

the ideal control group for California in the pre-treatment period. Let's see how the average of these controls compares to the trajectory of California. Figure 3 plots the average of these controls versus the trajectory of California. We can see that this line,

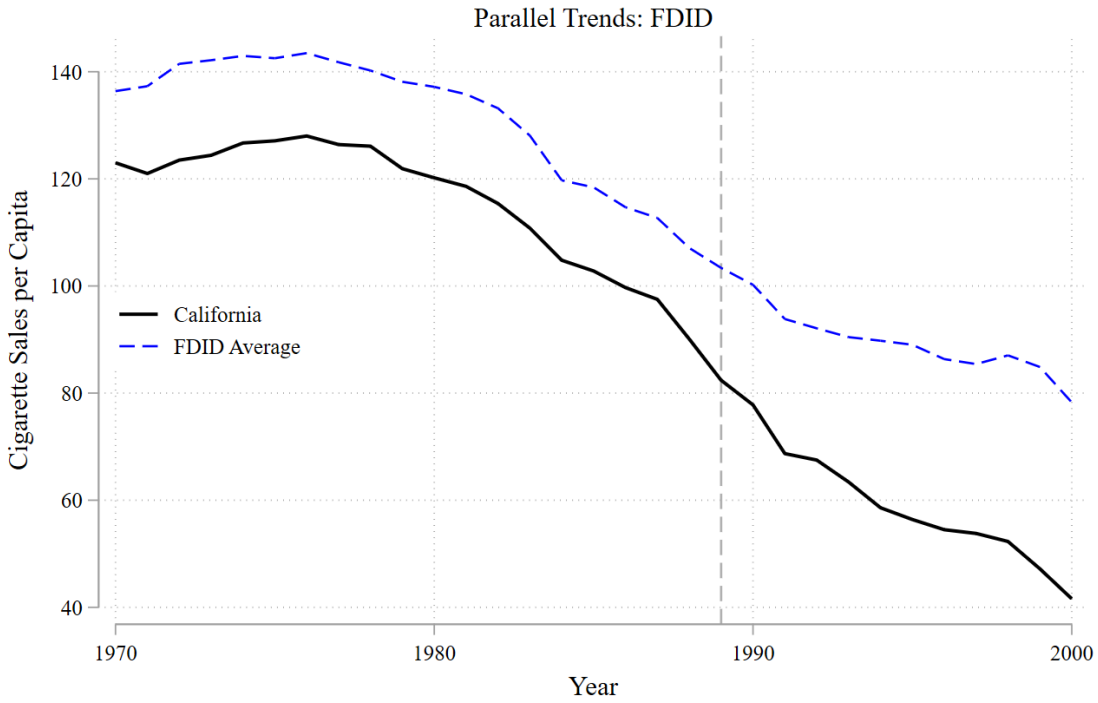

Figure 3: Forward Difference-in-Differences

the blue FDID line, is much more parallel to the trajectory of California in the pre-treatment period. The two lines clearly have the same (or a very very similar) trend to California. This suggests parallel trends holds, since the α term (our intercept from the DID model) is much more likely to fit well to California with the improved control group, instead of the situation where we use all controls. Now let's plot the results of the FDID counterfactual in Figure 4.

We can see that we have very good pre-intervention fit. Since the difference is so minimal, this suggests that the average of the differences between California and these controls would be constant (or approximately constant) absent P99. This means that our Forward DID analysis is successful in that the basic identifying assumptions of the model holds.

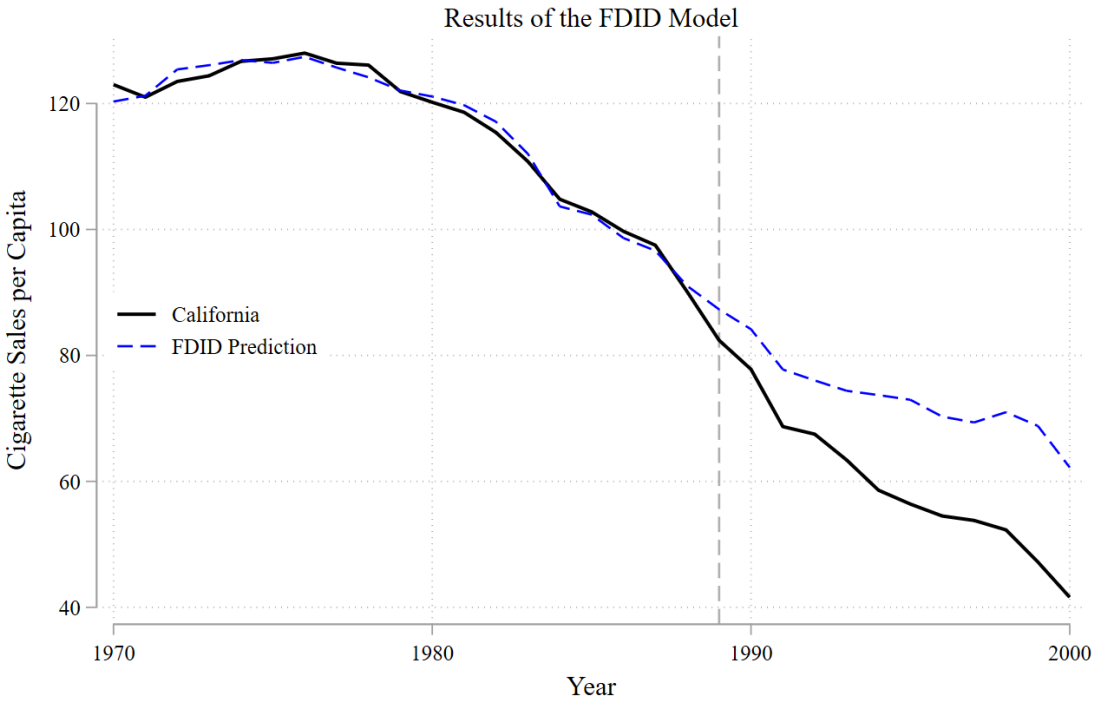

Figure 4: Forward Difference-in-Differences Predictions

Now we can consider the treatment effects. The ATT here, with the better control group, is roughly -13.64. This is **half** of the original DID treatment effect average! We can see that by using a good control group that provides more credence to the underlying model assumptions, we get much different results. The t-statistic and p-values are, respectively, 29.66 and much less than 0. The 95% Confidence Interval is $[-14.54861, -12.74481]$, which is much tighter than the original DID estimate. One may notice that both the original DID result and this one are both "statistically significant" in the sense that the t-statistic is much greater than 1.96 and the p-value is much less than 0.05. What this means from a statistical perspective is that we *cannot simply rely on* p-values, confidence intervals, and other simple inferential statistics to judge whether an analysis is valid or not. This has been advised time and

again by statisticians (Greenland et al., 2016).[1] For this course, I expect a similar standard of attention to detail/thought.

5 Discussion and Conclusion

From a practical angle however, we still maintain a similar analysis as above. The ATT suggests a clear reduction of the per capita smoking rate in California as a result of P99. However, this effect size is just much smaller than the DID model would suggest. The central takeaway here is that anti-tobacco policies, when comprehensive and widespread as P99 was, can be quite effective in reducing the rate of tobacco consumption. This paper revisited the Proposition 99 example of a state-wide anti-tobacco intervention, finding that the policy reduced consumption compared to what it would have been otherwise.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Flor, L. S., Reitsma, M. B., Gupta, V., Ng, M., & Gakidou, E. (2021). The effects of tobacco control policies on global smoking prevalence. *Nature Medicine*, *27*(2), 239–243.

Greathouse, J., Coupet, J., & Sevigny, E. (2024, August). Greed is good: Estimating forward difference-in-differences in stata. https://jgreathouse9.github.io/publications/FDIDSJ.pdf

[1] The traditional model of research is that your statistical model spits out a result at you (in the form of a p-value) and you must live with the result it gives you. If the p-value is less than 0.05, it is called "statistically significant" (when we use the 95% confidence interval). However, as I've shown here, a result can be statistically significant but practically garbage if the assumptions of the underlying model, parallel trends in this case, do not hold. For this course, I expect the validity of a research design to be the main thing you consider for if a result is plausible. Of course, DID is only one such model out of the innumerable models that exist in statistics, but this principle follows us regardless of the question, be it measuring association or casual inference.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337–350.

Li, K. T. (2024). Frontiers: A simple forward difference-in-differences method. *Marketing Science*, *43*(2), 239–468. https://doi.org/10.1287/mksc.2022.0212

Yousuf, H., Hofstra, M., Tijssen, J., Leenen, B., Lindemans, J. W., van Rossum, A., Narula, J., & Hofstra, L. (2020). Estimated worldwide mortality attributed to secondhand tobacco smoke exposure, 1990-2016. *JAMA network open*, *3*(3), e201177–e201177.