Causal Effect of COVID Lockdowns on China's Air Quality

Jerome Freudenberg and David Tang

December 6, 2021

Abstract

We apply two difference-in-differences models to investigate whether COVID lockdowns had a causal effect on air quality in Chinese cities at the beginning of 2020. We find no effect using a traditional two-period approach on a subset of the data, but we find one when we incorporate multiple time periods on the whole data. Nevertheless, our abilities to make any conclusions are limited by our nuisance function model fit.

1 Introduction

COVID-19 induced lockdown measures have been hypothesized to reduce urban pollution and benefit the environment. Since lockdown measures were applied at different times across cities, we can apply a staggered difference-in-differences model to estimate the causal effects of lockdown on air quality. This was done by He et al. (2020), which examined the short-term impact of lockdown and COVID prevention measures on urban Air Quality Index (AQI) and particulate matter (PM) data in China using a linear two way fixed effects model [4]. They found that both AQI and particulate matter concentrations were significantly reduced in locked-down cities.

In this analysis, we relax the parametric assumption and estimate the average treatment affect with a doubly robust estimator using their dataset. We first apply this to a standard two period context by considering only cities that were treated on a single day (i.e. went into lockdown) or not at all. We then extend our analysis to consider the full set of treated cities using a shared nuisance function estimation approach. In the former analysis, using only a small subset of the cities, we find no evidence of a short-term causal effect for COVID lockdown on AQI and PM. In the latter, predicated on a few simplifying assumptions, we find that lockdowns significantly decrease both AQI and PM concentration. However, our ability to fit the nuisance functions in this approach was limited based on the available data.

2 Linear Model

The authors in the original paper use the following linear model to identify and estimate the causal effect of city lockdowns on air quality:

$$Y_{it} = \beta C_{it} + X_{it}\alpha + \mu_i + \pi_t + \varepsilon_{it}$$

where C_{it} is the indicator for city i being in lockdown at time t, X_{it} is a set of weather covariates, μ_i is the fixed effect for city i, and π_t is the fixed effect for time t. This standard model is known in the literature as the two way fixed effects model (TWFE).

In addition to obvious concerns with using a parametric model for both identification and estimation, this model comes with another caveat in the form of a complicated interpretation of the estimated treatment effect that arises from heterogeneity in treatment adoption date across cities [1]. Since the adoption of the treatment varies across units, the estimated treatment effect β does not have a simple interpretation. Specifically, β has been shown to be a weighted average of various treatment effects with the undesirable property of possibly having negative weights [3].

We seek to estimate the average treatment effect with a non-parametric model that is suitable for the scenario described with staggered treatment effects.

3 Group Average Treatment Effect

One standard way of estimating the treatment effects in this staggered adoption scenario is to

define a group-time average treatment effect. Following [2], the group average treatment effect is defined as

$$ATT(g,t) = \mathbb{E}[Y_t(1) - Y_t(0)|G_g = 1]$$

In this equation, $Y_t(1)$ and $Y_t(0)$ are the potential outcomes at time t and G_g is an indicator for whether or not a unit first adopted the treatment in period g. Defining the group time average treatment effect allows us to stratify the units by when they were first treated to control for the staggered adoption times. Such a definition does not preclude treatment effects that change depending on adoption time. Moreover, this parameter allows for dynamic treatment effects, i.e., treatment effects that vary in the post periods.

After estimating group time average treatment effects for all groups and all times, we can aggregate the estimates with some desirable weighting, giving us more flexibility in summarizing the effect compared to the TWFE model.

In theory, this definition of group average treatment effects is nice because it can be identified with few assumptions outside of conditional parallel trends and irreversiblity of treatment, mitigating the strong parametric assumptions in TWFE. In practice, however, estimating group average treatment effects requires estimating generalized propensity scores

$$p_g(X) := \mathbb{P}(G_g = 1|X, G_g + C = 1)$$

where C is the indicator variable that is 1 if a unit is never treated. The naive approach to this would be to subset the data for each group and estimate separate generalized propensity scores for each subset. This approach, however, fails when the groups are small. In our data, there are several days for which only one or two cities entered into lockdown, i.e., there were many days g such that the number of cities for which $G_g = 1$ was very small. The histogram in figure 1 shows the distribution of lockdown start dates in our data.

This data limitation makes the statistical problem of estimating $p_g(X)$ naively with separate estimators impossible. To address this

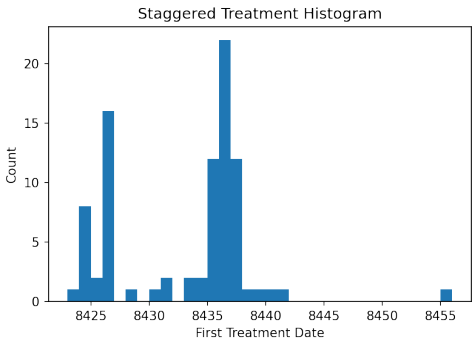

Figure 1: The distribution of lockdown dates

data limitation, we need an estimation procedure that allows us to share data when estimating the nuisance functions (see Shared Estimation).

4 Assumptions and Identification

We can identify the group time average treatment effect under relatively mild conditions compared to the parametric identification with the TWFE model. The two substantive assumptions in addition to the standard overlap and sampling conditions are:

1. Conditional Parallel Trends: For all t and g such that $t \geqslant g$, we have that

   $$\begin{aligned}&\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] \\ =&\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, C = 1]\end{aligned}$$

   This is the standard conditional parallel trends assumption stated for each treatment group g.

2. Irreversibility of the treatment: for all times t,

   $$A_t = 1 \implies A_{t+1} = 1$$

   where A_t denotes whether or not the subject has been treated by time t.

Under these assumptions, the group time average treatment effect can be identified with the same identification argument in standard nonparametric difference-in-differences applied to each group. This allows us to use the usual double machine learning estimator for estimating group average treatment effects.

5 Estimation in One Period

Before moving on to other nuisance function estimation procedures to handle data limitation problems, we estimated the group average treatment effect for the cities that started treatment on Feburary 5th, 2020. The rational for choosing this date is that 23 cities simultaneously entered into lockdown on this date, giving us a large enough group size to estimate the nuisance functions needed in the group average treatment effect.

As we are only considering one treatment date, we will simplify matters to only consider one pre-period and one post-period by averaging the data in all pre and post treatment periods. This strategy does not estimate any heterogeneity in the post period treatment effect, but it should still be valid for estimating an average effect in the post periods.

With these simplifications, the model reduces to the standard non-parametric two period diff-in-diff with a subset of the data. As such, we estimate the average treatment effect on the treated with the standard doubly robust machine learning estimator.

5.1 Model Fitting

As stated before, one significant advantage of the non-parametric identification result is that we can use flexible machine learning estimation techniques to fit the nuisance functions. In table 2, we compare several different outcome and treatment models for cities that entered lockdown on February 5th, 2020.

In all models, we used a set of weather covariates that included precipitation, snow, temperature, and temperature squared as well as a set of city characteristics that included population, GDP, number of firms, miles of highway, wastewater emissions, SO_2 emissions, and dust emissions.

For each outcome model, we calculated its cross validated mean squared error using 5 folds, and for each treatment assignment model, we did the same with its cross entropy.

The table in figure 2 shows that the random forest models with max depth 10 and the XGBoost models have the best fit for the outcome model and the treatment model. In particular, it is worth noting that the linear models do not fit the data particularly well, again demonstrating the value of defining the causal estimand in a non-parametric fashion.

	outcome	model	Q mse	Q baseline	g ce	g baseline
0	AQI	Lin./Log. Reg	576.20	699.17	0.45	0.36
1	AQI	RF (depth 10)	389.38	699.17	0.23	0.36
2	AQI	RF (depth 3)	418.21	699.17	0.26	0.36
3	AQI	XGBoost	355.49	699.17	0.26	0.36
4	PM	Lin./Log. Reg	368.31	475.68	0.45	0.36
5	PM	RF (depth 10)	264.22	475.68	0.23	0.36
6	PM	RF (depth 3)	283.60	475.68	0.26	0.36
7	PM	XGBoost	260.88	475.68	0.26	0.36

Figure 2: Outcome and treatment model fit diagnostics

5.2 Overlap Condition

We need the overlap condition to be satisfied in order to identify the average treatment effect for cities entering into lockdown on February 25th. Specifically, we need the propensity scores to be strictly less than one. The histograms in figure 3 show the propensity scores estimated using the various machine learning models.

We see that the overlap condition is satisfied when we estimate the propensity scores using random forests or logistic regression suggesting that the model is in theory identifiable (the overlap condition is stated in terms of population parameters and not their finite sample estimates). In practice, the extreme propensity scores estimated by XGBoost would inflate the variance of the causal estimate. Importantly, the overlap condition is not blatantly violated, suggesting that we are able to identify and estimate a causal effect.

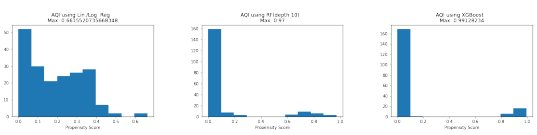

Figure 3: Estimated propensity scores. Left column is logistic regression, middle column is RF, and right column is XGBoost.

5.3 Conditional Parallel Trends

Following the suggestion in [2], we can test conditional parallel trends non-parametrically in the pre-period with the following null hypothesis:

$$H_0 : \mathbb{E}[Y_t - Y_{t-1} | X, G_g = 1] - \mathbb{E}[Y_t - Y_{t-1} | X, C = 1] = 0$$

for all pre-periods t such that $t < g$. Specifically, this tests that the trends in group g do not differ from the trends in the never treated group in all pre-periods before g. Although this does not guarantee that conditional parallel trends holds in the post period, this can at least provide a sanity check for the plausibility of our assumption.

We can estimate this parameter using the random forest outcome function estimated above and construct confidence intervals with a two sample t-test.

The plot in figure 4 shows 95 percent confidence intervals for the estimated conditional expectations of air quality index trends and particulate matter trends in the 4 weeks leading up to the lockdown (we aggregate by week to limit any day to day heterogeneity).

We see that the conditional pre-period trends in the treated group and the control group are roughly parallel. Assuming parallel trends in the post-period does not seem entirely unreasonable.

5.4 Results

We can plug our estimated nuisance functions into the standard double machine learning estimator to estimate the average treatment effect for cities that entered into lockdown on February 5th, 2020. The results using the various nuisance functions are displayed in table 5.

We get the same qualitative conclusion regardless of the machine learning estimator used for the nuisance functions. In each case, we see that the causal estimate is statistically insignificant. Unsurprisingly the standard errors obtained using the XGBoost nuisance functions are significantly larger than the standard errors returned by the other two estimation techniques. This relates to the point mentioned above regarding the extreme propensity scores that the XGBoost classifier predicts.

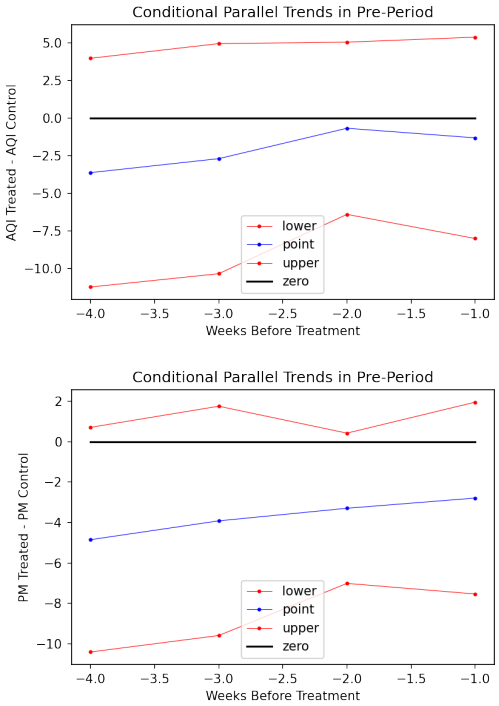

Figure 4: Pre-period parallel trends in AQI and PM

	outcome	model	estimate	p/m
0	AQI	Lin./Log. Reg	-9.55	15.03
1	AQI	RF (depth 10)	0.67	10.23
2	AQI	XGBoost	-3.94	34.30
3	PM	Lin./Log. Reg	-7.35	12.24
4	PM	RF (depth 10)	0.76	8.83
5	PM	XGBoost	-10.86	38.14

Figure 5: Causal estimates

The large standard errors indicate that our sample size is not large enough to detect a causal effect. While some of this data limitation is due to the inherent nature of the problem, we can attempt to mitigate this lack of statistical power using more sophisticated estimation techniques that are able to share data when estimating group average treatment effects.

6 Shared Estimation

While we are able to estimate a group average treatment effect for cities entering lockdown on February 5th, 2020, this estimation approach is not optimal.

First, this approach is unable to capture any heterogeneity in treatment timing (we are only estimating the group time average treatment effect for one group). We would hope that this treatment effect is representative of the treatment effect for all groups, but in reality, this may not be the case. It is not hard to envision a scenario in which the later lockdowns had different effects from the earlier lockdowns due to people's evolving understanding of the virus.

Second, this approach ignores most of the data, causing us to lose statistical power and giving relatively large standard errors. For these reasons, we turn to an estimation procedure designed to estimate group average treatment effects with limited data.

The issue with estimating the group average treatment effect naively is that the nuisance parameters are group specific and can only be estimated with observations from a particular group. However, it is not unreasonable to imagine that the nuisance functions for the different groups are similar. Instead of learning entirely different nuisance functions for each group, we can estimate a single nuisance function treating the time as an additional parameter.

To formalize this, let us assume that

$$(Y_t - Y_{t-1}, G_t, X) \sim P_t$$

where P_t is a distribution depending on time and G_t is the same variable defined above for the group average treatment effect. For the sake of simplicity, we will only focus on contrasts between the period directly proceeding the lockdown and the period directly following the lockdown.

In this case, we can estimate the nuisance functions $p(X, t)$ and $Q(G, X, t)$ with the hope that the time dependence of p and Q is relatively simple to model. The naive estimator for the group average treatment effect falls out as a special case where we allow a completely different nuisance function for each treatment group.

We can also consider the other extreme case where we allow $P_t = P$ for all t. In this case, we would be imposing the assumption that the treatment effect and impact of covariates are invariant with respect to time. As mentioned above, such an assumption may not be appropriate for our problem of estimating the effect of COVID lockdowns given the significant uncertainty surrounding the virus in the earlier stages of the pandemic.

This estimation procedure will only work well when the time dependence of the effects is relatively easy to learn. If the relationship is too complicated, we would not be able to get good fitted values for the samples, leading back to the same problem that we encountered with the naive estimator.

After estimating $\hat{p}$ and $\hat{Q}$ with general machine learning techniques, we can use the standard doubly robust machine learning estimator to obtain a point estimate $\hat{\tau}_g$ and variance $\hat{\sigma}_g^2$ for $ATT(g, g)$ for every group g with more than 1 unit. We can then combine these estimates with the inverse variance weighting to get a single point estimate and variance.

$$\hat{\tau} = \frac{\sum_g \hat{\tau}_g / \hat{\sigma}_g^2}{\sum_g 1/\hat{\sigma}_g^2} \text{ and } \hat{\sigma}^2 = \frac{1}{\sum_g 1/\hat{\sigma}_g^2}$$

Unlike the implicit weighting used in the linear model, the inverse variance weighted average has the guarantee of positive weights and a nice interpretation as the minimum variance weighting.

6.1 Model Fitting

Table 6 displays model fitting diagnostics using various machine learning methods to estimate the nuisance functions in the procedure described in this section.

Compared to the baseline model and the linear model, the random forest and XGBoost estimators have reasonable predictive power in estimating the outcome model. However, none of the machine learning models are particularly predictive of the treatment assignment. This

	outcome	model	Q mse	Q baseline	g ce	g baseline
0	AQI	Lin./Log. Reg	1246.10	1267.46	0.04	0.03
1	AQI	RF (depth 10)	990.59	1267.46	0.03	0.03
2	AQI	RF (depth 3)	1177.31	1267.46	0.03	0.03
3	AQI	XGBoost	1049.65	1267.46	0.03	0.03
4	PM	Lin./Log. Reg	690.80	702.45	0.04	0.03
5	PM	RF (depth 10)	517.46	702.45	0.03	0.03
6	PM	RF (depth 3)	650.15	702.45	0.03	0.03
7	PM	XGBoost	533.26	702.45	0.03	0.03

Figure 6: Shared estimation model fit diagnostics

	outcome	model	Q mse	Q baseline	g ce	g baseline
0	AQI	Lin./Log. Reg	1323.91	1327.08	0.04	0.03
1	AQI	RF (depth 10)	1246.48	1327.08	0.03	0.03
2	AQI	RF (depth 3)	1320.27	1327.08	0.03	0.03
3	AQI	XGBoost	1469.49	1327.08	0.03	0.03
4	PM	Lin./Log. Reg	494.12	491.21	0.04	0.03
5	PM	RF (depth 10)	555.82	491.21	0.03	0.03
6	PM	RF (depth 3)	526.51	491.21	0.03	0.03
7	PM	XGBoost	646.77	491.21	0.03	0.03

Figure 7: Model fit on treated

is because the treatment, in this modelling scenario, is very rare.

More generally, the models are not able to fit the observations from the treated units very well. Although the machine learning models perform better than baseline on the full test set, they do not improve upon the baseline outcome prediction for treated units. This is shown in table 7 where we compare the model MSE on treated units in the test set against the baseline MSE from predicting the outcome from the outcome conditional solely on treated status.

The poor predictive accuracy of the outcome and treatment models are cause for concern: the asymptotic results for the double machine learning estimator rely on consistent estimators for the nuisance functions. There is no immediately obvious way for how to increase the predictive performances of our models given our limited set of covariates and small sample size. Nonetheless, we will proceed with causal estimation, keeping in mind that our conclusions may be biased by the poor estimation of the nuisance functions.

	outcome	model	estimate	p/m
0	AQI	Lin./Log. Reg	-7.85	3.64
1	AQI	RF (depth 10)	-8.08	2.85
2	AQI	XGBoost	-13.67	1.21
3	PM	Lin./Log. Reg	-5.48	1.57
4	PM	RF (depth 10)	-5.62	2.08
5	PM	XGBoost	-6.37	1.88

Figure 8: Inverse weighted average treatment effect on the treated

	day	estimate	p/m
0	20200124	-28.711168	9.444318
1	20200125	42.160968	37.211422
2	20200126	15.619417	20.002507
3	20200131	0.376498	12.608754
4	20200202	-29.458210	7.067823
5	20200203	-54.786224	65.860538
6	20200204	-2.472994	8.172015
7	20200205	-3.228286	5.156871
8	20200206	0.461201	5.782714

Figure 9: Average treatment effect on AQI by lockdown day

6.2 Results

The inverse weighted average of the group average treatment effects is reported in table 8. We also report the individual group specific treatment effects estimated with the random forest regressor for the various treatment dates in table 9.

While we get significant estimates in several treatment groups and an overall significant estimate of the causal effect, we emphasize that our results may be biased due to the poor outcome and treatment model fits.

7 Conclusion

In this project, we set up a non-parametric framework for estimating the causal effect of lockdowns on air quality index and particulate matter concentration in China. We point out that the original estimate from [4] does not have a clear interpretation due to the staggered adoption of treatment across units and the implicit weighting used in the parametric linear model.

When estimating a group average treatment effect for cities entering into lockdown on February 5th, we do not find a statistically significant effect.

While we were unable to estimate the nuisance functions using shared estimation technique well enough to draw reliable conclusions, we believe that this approach has the potential to be an interesting way of estimating group average treatment effects in staggered difference in difference designs. With the increasing availability and scope of COVID-19 data compared to when the dataset was collected, performing this analysis on a more recent dataset would be an interesting and plausible direction for future research.

8 Code Availability

Code/data analysis done for this project are available at `https://github.com/JFreud/lockdown-air-quality`.

9 Acknowledgements

We thank Professor Veitch for his very helpful input on modeling treatment across multiple time periods.

References

[1] S. Athey and G. W. Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 2021.

[2] B. Callaway and P. H. C. Sant'Anna. Difference-in-differences with multiple time periods. *SSRN*, 12 2020.

[3] C. de Chaisemartin and X. D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96, September 2020.

[4] G. He, Y. Pan, and T. Tanaka. The short-term impacts of covid-19 lockdown on urban air pollution in china. *Nature Sustainability*, 3:1005–1011, 12 2020.