The impact of air pollution on birthweight: evidence from grouped quantile regression

Martina Pons[1] [ID]

Abstract

Estimates of the average effect of pollution on birthweight might not provide a complete picture if more vulnerable infants are disproportionately more affected. To address this, I focus on the distributional effect of particulate matter pollution ($PM_{2.5}$) on birthweight. To estimate the impact, this paper uses grouped quantile regression, a methodology developed by Chetverikov et al. (Econometrica 84(2): 809–833, 2016), which allows estimating the impact of a group-level treatment on an individual-level outcome when there are group-level unobservables. The analysis reveals nonhomogeneous effects indicating that pollution disproportionately affects infants in the lower tail of the conditional distribution, whereas average effects suggest only minimal and not economically significant impact of pollution on birthweight. The findings are also consistent across different specifications.

Keywords Air pollution · Birthweight · Infant health · Quantile regression

JEL Classification I18 · Q52 · Q53

1 Introduction

Public awareness about pollution has been increasing over the last decades, and concerns about the possible adverse effects on health have been rising. A growing number of studies have taken on the challenge of estimating the potential welfare gain of pollution reduction. To facilitate the computation of lifetime exposure to pollution,

✉ Martina Pons
  martina.pons@vwi.unibe.ch

1   Department of Economics, University of Bern, Schanzeneckstrasse 1, 3001 Bern, Switzerland

⚛ Springer

many focused on infant health. In the economic literature, there are two prominent measures of infant health, namely birthweight and mortality. This paper concentrates on the former.

Birthweight is not only an indicator of neonatal health and is negatively correlated with infant mortality, but also low birthweight causes substantial medical costs (see Almond et al. 2005). Estimating the impact of pollution on health presents itself with some difficulties as there might be factors that influence pollution exposure, which are also correlated with health so that pollution is not randomly assigned.[1] Economic status, degree of urbanization, race, education, and some behaviors such as smoking are some examples of confounding variables. For instance, Currie (2011) documents that Hispanic and African-American mothers as well as smokers are disproportionately more likely to live close to hazardous emission sites, while the opposite is true for more educated mothers. Similarly, Currie et al. (2009) point out that in urban areas, where there is also a higher share of highly educated people, the pollution concentration is generally higher. Since it is unlikely that we can control for all confounding factors, studies often rely on fixed effects and instrumental variables to isolate the causal effect (see, for example, Chay and Greenstone 2003a, b; Currie et al. 2009; Knittel et al. 2016).

A large body of the literature has focused on the impact of pollution on birth outcomes and mortality (see, for example, Currie and Walker 2011; Currie et al. 2009; Currie and Neidell 2005; Knittel et al. 2016; Chay and Greenstone 2003a, b). Some studies have also suggested that the effect might be larger in the lower tail. For instance, Knittel et al. (2016) document that the impact of traffic on infant mortality is nearly entirely absorbed by premature or low birthweight infants and Currie et al. (2009) document that the effect of carbon monoxide (CO) concentration on the incidence of low birthweight was greater than the impact on average birthweight. At the same time, existing studies using quantile regression methods find that the impact of birth inputs and other variables, including mother's race and education, varies across the distribution (see, for example, Abrevaya 2001; Abrevaya and Dahl 2008; Koenker and Hallock 2001; Chernozhukov and Fernández-Val 2011). These results, taken together, suggest that the effect of pollution might be larger among more vulnerable infants. If this is the case, estimates of the average effects of pollution on birthweight might underestimate the potential welfare gain of reducing pollution. This is because the marginal welfare gain of additional weight is higher at the bottom of the distribution. In other words, a decline in birthweight in the lower tail of the distribution causes more economic and welfare costs compared with an equivalent decrease in the upper tail. For this reason, in this paper, I use 1999–2004 birth certificate data from the USA to study the distributional effect of particulate matter pollution ($PM_{2.5}$) on the conditional distribution of birthweight.[2]

Despite the exact biological mechanism through which exposure to pollution may affect birth outcomes is not yet wholly understood (Gehring et al. 2014), reasonable explanations include the effects of pollution on fetal growth and prematurity (Slama et al. 2008). For example, particulate matter pollution could impact the mother's

---

[1] For a more complete review of these challenges, see Graff Zivin and Neidell (2013).

[2] $PM_{2.5}$ stands for particulate matter or fine particles with a diameter of 2.5 micrometers or less.

immune system and increase the risk of infections, possibly leading to preterm birth (Slama et al. 2008). If pollution is associated with decreased birthweight due to shorter gestation, one might expect a negative effect, especially in the lower tail of the distribution. By contrast, the impact going through the fetal growth channel is ambiguous. On the one hand, if all mothers are affected equally, higher pollution would cause a downward shift in the distribution. On the other hand, if some mothers are more susceptible to pollution than others, the effect might not be homogeneous.

To estimate the effect on the distribution, I use an estimator proposed by Chetverikov et al. (2016), which they refer to as *grouped IV quantile regression*. As this paper does not use instrumental variables, I refer to the estimator as grouped quantile regression. To the best of my knowledge, neither this model nor standard quantile regression has been used to estimate the effect of pollution on birthweight. This paper builds on the work of Chay and Greenstone (2003b), who also study the effect of pollution in the USA. However, they focus on the 1980–1982 period and concentrate on total suspended particulates (TSPs) pollution, which is also a measure of suspended particulate matter, but of much larger size.[3]

This work contributes to the existing literature studying the impact of pollution on birthweight by providing estimates of the magnitude of the impact on weight on different parts of the conditional distribution, and not only average effects on weight or impacts on the incidence of low birthweight. Quantile regression enables to better estimate the costs of pollution, also considering the different economic implications of a decrease in birthweight at the bottom of the distribution. In contrast to other studies using quantile regression, this paper focuses on an endogenous treatment that varies only at the group-level and not on individual-level birth inputs.

The results suggest a negative effect of $PM_{2.5}$ on the lower tail of the weight distribution and no effect on the upper tail. According to my preferred specification, an increase in $PM_{2.5}$ by one standard deviation is associated with a decrease in birthweight by 28.0 g in the 5th percentile of the distribution and 20.5 g in the 1st decile. By looking at average effects, I do not find an impact of pollution on birthweight.

The remainder of the paper is structured as follows. Section 2 introduces the empirical approach, including a short overview of the estimator. Chapter 3 presents the data, while Sect. 4 focuses on the results. Following this, Sect. 5 concludes.

2 Empirical approach

The grouped quantile regression estimator developed by Chetverikov et al. (2016) is particularly suited for estimating the effect of a group-level treatment on an individual-level outcome. For the estimation, variables are divided into two categories, namely group level and individual level, and groups are defined as month of birth and county combinations. In this case, pollution is a group-level treatment as all individuals born in a given county and month are assumed to be exposed to the same level of pollution. Other group-level variables are the county-level controls, while details about the single births are individual-level or within-group covariates. For reason of clarity, in the

---

[3] TSPs include all suspended particles smaller than 40 micrometers.

remainder of the paper, when I refer to month, I mean the month and year combination unless otherwise specified.

2.1 Model

The model assumes that the conditional quantile function of birthweight of individual i born in county c and month m, denoted bw_{icm}, is a function of individual-level covariates z_{icm}, particulate matter pollution $PM_{2.5,cm}$, group-level controls x_{cm}, as well as group-level unobservables $\eta_{cm}(u)$:

$$Q_{bw_{icm}|z_{icm},PM_{2.5,cm},x_{cm},\eta_{cm}}(u) = z'_{icm}\delta_{cm}(u) + PM_{2.5,cm}\beta(u)$$
$$+x'_{cm}\theta(u) + \eta_{cm}(u), \quad u \in U, \qquad (1)$$

where U contains all the quantiles of interest. Individual-level control variables in z_{icm} include mother's age and education, the average number of daily cigarettes smoked during gestation, as well as dummies for Hispanic and black mothers, and the sex of the infant. The vector $\delta_{cm}(u)$ contains the parameters of the individual-level variables and is allowed to vary across groups. The pollutant of interest $PM_{2.5,cm}$ is the average concentration of particulate matter with a diameter smaller or equal to 2.5 micrometers per cubic meter during pregnancy, and $\beta(u)$ is the parameter of interest. The vector x_{cm} contains average temperature and precipitations during gestation as well as their squares and county-level average personal income. Similarly to Chay and Greenstone (2003a), I also include controls for different county-level transfers such as unemployment benefits, public assistance medical care benefits, food stamps, personal current transfer receipts, and income maintenance benefits. Finally, $\eta_{cm}(u)$ is the group-level error term. The county-level income variables are included to control for the socioeconomic characteristics of the counties, whereas weather variables are included as these are likely to affect pollution and may have an impact on health (Graff Zivin and Neidell 2013).

Gestation length is an important determinant of birthweight and may be itself a consequence of pollution exposure. Consistent with this hypothesis, Currie et al. (2009) provide some evidence suggesting that higher pollution leads to a reduction in the duration of the pregnancy. For this reason, I do not control for gestational age as it is an outcome and, thus, endogenous.

Although model (1) controls for potential confounding factors such as education, smoking habits, and race, there might be other variables influencing pollution exposure as well as birthweight, such that $PM_{2.5,cm}$ is endogenous. To deal with this endogeneity, Chay and Greenstone (2003a) use fixed effects and IV and find similar results with both estimation strategies. Due to the difficulty of finding a proper instrument, this paper relies on two types of fixed effects. More precisely, I assume that η_{cm} is additively separable in month and county effects. The month of birth fixed effects accounts for time-varying unobservables, and the county-level fixed effects account for unobserved but time-invariant county-level factors. Before presenting the model with the fixed effects, it is useful to discuss the estimator. Hence, in the follow-

ing subsections, I first present the estimator and then continue with the fixed effects model.

2.2 Grouped quantile regression

Grouped quantile regression is performed in two stages. The first stage consists of regressing the dependent variable, birthweight, on individual-level covariates for each group, using quantile regression. For each group and quantile, the resulting constant is saved, and the data are collapsed at the group level. In the second stage, the constant from the first stage is regressed on the variable of interest, $PM_{2.5}$, using one observation per group by OLS. Therefore, the observations are weighted by the size of the groups. Additionally, the second stage regression might include group-level controls and fixed effects. Differently, the first stage uses only data for one group at the time, where the unit of observations are individuals. Thus, there is no time variation in the single first-stage regressions. A different way to see it is to consider the groups as subsamples, which are used independently from each other in the first stage, whereas in the second stage, the units of observations are counties, which are observed over time.[4]

When there are no individual-level covariates, and there is no group-level heterogeneity, grouped quantile regression reduces to the minimum distance estimator proposed by Chamberlain (1994). Chetverikov et al. (2016) point out that grouped quantile regression has several advantages. First, its bias disappears fast as the sample size increases. Second, grouped quantile regression remains consistent in the presence of group-level unobserved heterogeneity, which biases traditional quantile regression as well as quantile regression with fixed effects.[5] Third, inference in the second stage is easy to compute, and standard errors can be estimated without having to take into account the first stage. Fourth, the computation is faster than for alternative estimators. For example, computing the fixed effects quantile regression estimators of Koenker (2004) or Kato et al. (2012) is not feasible because of the large number of observations and fixed effects. Canay (2011) proposed an alternative estimator that can be computed within a reasonable amount of time. However, this method is too restrictive, as it assumes that the fixed effects are constant across the distribution. While this estimator comes with attractive features, there are a few caveats that should be acknowledged. First, it is questionable whether the first stage can be ignored when computing the standard errors. This problem might be more severe if the groups are small since Chetverikov et al. (2016) assume that the number of observations per group increases faster than the number of groups, so that the first-stage error does not matter. Second, the estimated constant changes substantially when the regressors are reparameterized. Thus, by using only the first stage constant in the second stage, the estimator might not be invariant to reparameterizations of the individual-level variables. Addressing these issues is left for future work.

---

[4] For additional information about the estimator as well as the asymptotic theory, see Chetverikov et al. (2016).

[5] Group-level unobservables, even when uncorrelated with the treatment, cause problems similar to left-hand side measurement errors (see, for example, Hausman 2001; Hausman et al. 2019). While measurement errors in the dependent variable do not bias OLS linear regression estimates, such errors are a source of bias in quantile regression.

2.3 Model with fixed effects

To include the fixed effects, let $\eta_{cm}(u) = \lambda_m(u) + \gamma_c(u) + \epsilon_{cm}(u)$, so that the model becomes

$$
\begin{aligned}
Q_{bw_{icm}|z_{icm},PM_{2.5,cm},x_{cm},\lambda_m,\gamma_c,\epsilon_{cm}}(u) = {}& z'_{icm}\delta_{cm}(u) + PM_{2.5,cm}\beta(u) + x'_{cm}\theta(u) \\
& + \lambda_m(u) + \gamma_c(u) + \epsilon_{cm}(u), \quad u \in U \qquad (2)
\end{aligned}
$$

where the unobserved time-invariant county effects are captured by $\gamma_c(u)$, the month effects by $\lambda_m(u)$, and $\epsilon_{cm}(u)$ is the group-level error term. Group-level variables and the fixed effects enter only the second stage regression, which is estimated by OLS. In this case, like in mean regression, the county fixed effects capture all time-invariant county-level factors, and the month fixed effects capture time-varying unobservables that are the same across counties. Thus, the identifying assumption is that, conditional on the fixed effects, pollution is as good as randomly assigned. A threat to identification could arise if, for example, wealthier households, who have on average heavier children, decide to move to less polluted areas. Currie et al. (2009) tested a similar hypothesis and did not find substantial evidence that pollution was an important motivating factor. If, after controlling for the fixed effects, pollution is not exogenous, the remaining endogeneity might be small.

Birthweight of infants born in the same county is likely to be correlated as mothers in a given county are exposed to the same set of doctors, hospitals, and other living circumstances. Furthermore, since I observe counties over time, the error terms are serially correlated. For instance, a highly polluted county today is likely to be polluted also next month. Since the second stage is performed with one observation per group, to take these issues into account, the standard errors are clustered at the county level.

3 Data

To estimate the effect of pollution on birthweight, I merge longitudinal natality data with environmental, income, and transfer data. Natality data are taken from the Vital Statistics of the National Center for Health Statistics. The natality datasets contain information on the birth certificates of every birth in the USA for a given calendar year, including mother's residence state and county, birth month, some socio-demographic characteristics of the mother and father, as well as information about the pregnancy and the birth. Due to new restrictions, the publicly available datasets include geographic indicators only until 2004. Since for the pollutant of interest, the number of measurements is substantially lower for the years before 1999, this study focuses on the 1999–2004 period.

The natality datasets contain approximately four million observations for each year; however, many were not included in the final dataset. First, I remove the observations for which the mother's residence is abroad or is in the outlying areas of the USA. Second, as the dataset does not contain information for counties with less than 100,000 inhabitants, I could not match these observations with environmental data. These

observations, as well as those for which birthweight, mother's age, education, race, Hispanic origins, or the number of cigarettes smoked during the pregnancy are missing, are dropped. Furthermore, consistent with previous research, multiple births are excluded.

Information on tobacco use during pregnancy is not available for California, and thus, controlling for smoking during pregnancy leads to a considerable loss of observations. Given the importance of this variable, also considering its variation across counties and time, it is included in the analysis.[6]

Monthly county-level data on temperature and average precipitation are from the Center for Disease Control and Prevention, and daily pollution data are from the United States Environmental Protection Agency. Precipitations are measured in millimeters, whereas the daily average concentration of particulate matter is measured in micrograms per cubic meter. There are two variables for temperature, one containing average maximum temperature and the other average minimum temperature in degree Celsius. So, temperature is calculated as the average of the two. For each birth month and county combination, I created variables for pollution exposure, temperature, and precipitation during pregnancy. These variables are prone to measurement errors for different reasons. First, as the exact date of birth is missing, exposure is calculated as the average pollution in the month of birth and the previous eight months, regardless of whether the infant was born on the first or last day of the month. Second, not all pregnancies last nine months. Third, as pointed out by Greenstone and Gayer (2009), actual pollution exposure is not observed, and pollution levels can vary substantially within a county. Fourth, people can influence their exposure to pollution by spending a different amount of time outdoors. Therefore, I can only approximate exposure. On the other hand, this approximation enables the use of grouped quantile regression as pollution becomes a group-level treatment.

Particulate matter of the different kinds comes from various sources and is divided into primary and secondary sources. The primary includes combustion processes, traffic, construction as well as power plants (Larssen and Hagen 1997). However, the Environmental Protection Agency (2019) argues that most of it derives from secondary sources, namely chemical reactions of other pollutants in the air, which are emitted by traffic, electricity generation, or industry. Particulate matter exposure can have effects on lung- and heart-related diseases, and pregnant women might also be more susceptible to particulate matter (Kelly and Fussell 2012). The finest $PM_{2.5}$ is particularly health-damaging since, due to their small size, some particles can enter into the lungs and may find a way into the bloodstream (Kelly and Fussell 2012).

The Bureau of Economic Analysis provides county-level data on population, income, and transfers. These include per capita personal income, public assistance medical care benefits, food stamps (SNAP), personal current transfer, unemployment benefits, and income maintenance benefits. With the population data, I transform the transfer variables in per capita term. As only yearly data are available, I construct variables for income and transfer during pregnancy for each month of birth as the weighted

---

[6] The results are similar when keeping the observations with missing values for cigarettes and including an interaction term between a dummy for California and the month fixed effects.
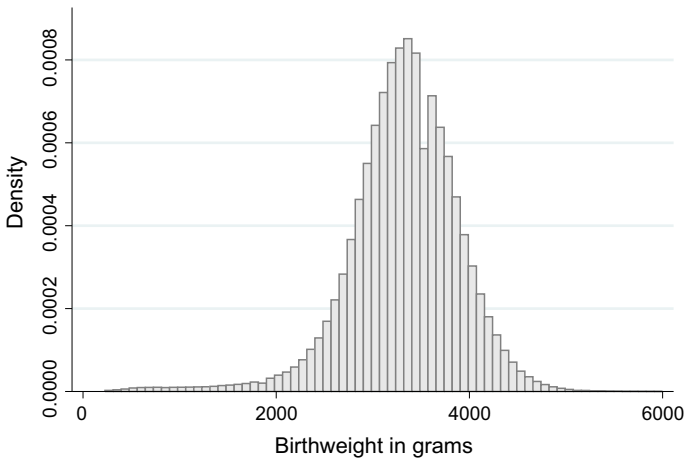
Fig. 1 Distribution of birthweight

average of the number of months of gestation in the current and previous year.[7] To assign income, transfers, and environmental variables to infants born at the beginning of 1999, I also use data for the year 1998. So, these data cover the 1998–2004 period.

Some observations from the birth dataset could not be matched with environmental or income data. Therefore, these observations are not included in the final dataset. Groups with less than 200 births are removed to ensure that there are enough observations to estimate the first stage. The final dataset includes 9,016,342 individual observations divided into 13,239 groups, with, on average, 681 observations each. This comprises data for 272 different counties and 43 different states, for which there are, on average, data for 47 months out of 72. It is important to note that given the considerable number of observations that have been excluded, the sample may not be representative of the USA. In the final dataset, the mothers of two-third of the remaining 9 million infants reside in counties with a population of over half a million people and only 10% in counties with a population between 100,000 and 250,000 inhabitants. Nonetheless, over half of the mothers reside in small cities with a population below 100,000 inhabitants. Therefore, it seems that the dataset represents only urban and suburban areas.

Figure 1 shows the distribution of birthweight, while Table 1 shows descriptive statistics for different levels of pollution. Columns 1 to 4 provide the values for the four quartiles of $PM_{2.5}$, while the last column gives the summary statistics for the whole sample. About one-fifth of the mothers are black, while over 20% have Hispanic origins. Compared with the 2000 census data, the share of black and Hispanic mothers is substantially larger than in the population.[8] This discrepancy may be due to the different fertility rates across women or due to the loss of observations.

---

[7] For example, for an infant born in February 2002, income during gestation is calculated as $\frac{2}{9} income_{2002} + \frac{7}{9} income_{2001}$, since seven-ninth of the pregnancy was in the year 2001.

[8] According to the 2000 census, these groups represent 12% and 13% of the population, respectively (United States Census Bureau 2001).

Table 1 Summary statistics

| | PM$_{2.5}$ Level | | | | |
	(1) 1st quartile	(2) 2nd quartile	(3) 3rd quartile	(4) 4th quartile	(5) Total
Individual-Level Variables					
Birthweight					
Mean	3330.3	3336.9	3317.7	3301.3	3321.6
Standard Deviation	556.0	571.0	582.4	592.6	575.8
5th percentile	2440	2438	2381	2353	2410
10th percentile	2704	2693	2665	2637	2680
25th percentile	3033	3033	3005	3005	3020
50th percentile	3345	3373	3345	3340	3345
75th percentile	3685	3686	3685	3661	3685
90th percentile	3969	3997	3983	3969	3979
% Female	0.488	0.488	0.488	0.489	0.488
% Black	0.102	0.169	0.248	0.311	0.208
% Hispanic	0.299	0.234	0.201	0.163	0.224
Mother's age	27.35	27.63	27.79	27.43	27.55
Mother's education	12.88	12.98	13.07	13.01	12.99
Cigarettes	0.701	0.678	0.667	0.867	0.728
Group-level variables					
PM$_{2.5}$					
Mean	9.344	12.16	14.04	16.71	13.06
Standard Deviation	1.237	0.569	0.595	1.306	2.863
Temperature	16.33	14.76	13.26	12.63	14.25
Precipitations	2.317	2.797	2.933	2.752	2.700
Per capita personal income	32142.3	34514.2	34555.1	34919.1	34032.6
Public medical benefits	675.0	801.2	1077.2	1039.7	898.3
SNAP	61.63	58.92	74.37	82.05	69.24
Personal current transfers	3888.4	4004.6	4440.2	4478.2	4202.8
Unemployment benefits	124.7	132.3	146.6	129.8	133.4
Income maintenance benefits	365.3	390.4	480.0	514.8	437.6
N	2,254,309	2,253,995	2,255,235	2,252,803	9,016,342

The descriptive statistics show that infants exposed to a higher level of pollution also differ in other characteristics. For example, mothers of children exposed to a higher level of PM$_{2.5}$ during pregnancy are less likely to be Hispanic as well as more likely to be black and to smoke a higher number of cigarettes than mothers of children in lower quartiles of pollution. Furthermore, there seems to be a negative correlation between birthweight and PM$_{2.5}$, which looks stronger lower percentiles of birthweight. It is also noteworthy that the 5th percentile is the only one that falls under the definition of

Table 2 OLS estimates of the effect of $PM_{2.5}$ on birthweight

| | Birthweight in grams | | | | | | |
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$PM_{2.5}$	-10.38^{***}	-0.923	0.0481	4.751	-0.267	-4.364	-4.167
	(0.605)	(4.166)	(4.450)	(4.146)	(3.220)	(3.594)	(3.593)
Individual controls		x	x	x	x	x	x
Weather controls			x	x	x	x	x
Income controls				x	x	x	
Transfers controls				x	x	x	
County fixed effects					x	x	x
72 months fixed effects						x	x

The regression estimates give the effect of an increase in $PM_{2.5}$ exposure during gestation by one standard deviation on birthweight. The dependent variable is birthweight in grams. Results are based on a sample of 9,016,342 individual births divided into 13,239 groups. Regressions are weighted for the number of births in each group. Heteroscedasticity robust standard errors in parentheses are clustered at the county level. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

low birthweight so that most of the percentiles in the lower tail of the distribution are already above this threshold.[9]

The weather variables are also correlated with pollution as $PM_{2.5}$ exposure is higher in colder climates. Similarly, income and transfer variables vary substantially across the quartiles of $PM_{2.5}$ exposure. For instance, average personal income, as well as various transfer variables, is larger in higher quantiles of pollution exposure. This could indicate higher inequality in more polluted areas.

4 Results

This section presents first the results of the average effect of pollution on birthweight, and then the impact on the distribution using grouped quantile regression. To facilitate the interpretation of the coefficients, the independent variable of interest is normalized to have a mean of zero and a standard deviation of one.

4.1 Mean estimates

Table 2 shows the results of OLS regressions. To allow δ_{cm} to vary across groups, the estimation is performed in two stages simply using OLS instead of quantile regression in the first stage. In Column 1, average group-level birthweight is regressed on $PM_{2.5}$ without control variables. The coefficient is negative; however, it is most likely a severely biased estimate of the causal effect. Column 2 includes individual-level covariates. The estimate becomes smaller in absolute value and is not statistically significant. Column 3 includes weather controls and Column 4 also income and transfers controls. Although of small magnitude, the coefficients are perversely signed. As

---

[9] Low birthweight is defined as birthweight below 2500 g.

shown before, several variables that might influence birthweight are correlated with pollution. As this is the case for observed variables, it is plausible that pollution is correlated with some unobserved factors that affect birthweight. To correct for this bias, I add county and month fixed effects.[10] With the county fixed effects (Column 5), the point estimate is positive but small and with no economic significance. Column 6 also includes month fixed effects. The coefficient has the expected sign, but it remains of small magnitude with no economic implication. This last estimate suggests that an increase in pollution exposure by one standard deviation leads to a decrease in birthweight by 4.4 g. Column 7 includes the fixed effects but not income and transfers controls. The coefficient changes only slightly, indicating that most of the explanatory power of these variables is captured by the fixed effects. Besides being statistically insignificant, the average estimates do not suggest any economically meaningful effect of $PM_{2.5}$ on birthweight.[11]

4.2 Distributional effects

This subsection presents the distributional effect estimated using grouped quantile regression for the set of quantiles $u \in \{0.05, 0.1, \ldots, 0.95\}$. Table A.2 in Supplementary Appendix illustrates the estimated coefficients for all specifications. The results of the first stage are not displayed. Instead, kernel densities of the distribution of $\hat{\alpha}_{cm,1}(u)$ are included in Figure A.1 in Supplementary Appendix.

Figure 2 shows the results using the main specification. $PM_{2.5}$ seems to have a negative effect on birthweight, mostly in the lower tail of the conditional distribution. According to these estimates, an increase in $PM_{2.5}$ by one standard deviation is associated with a decrease in birthweight by 28.0 g in the 5th percentile and by 20.5 g in the 10th percentile. The effect in the 15th and 20th percentiles is more modest, and an increase in $PM_{2.5}$ by one standard deviation is associated with a decrease in birthweight by 10.8 and 9.3 g, respectively. The estimates are statistically significant at conventional levels only up to the 2nd decile. In higher quantiles, the coefficients become smaller and remain negative up to the 45th percentile. In the higher half of the distribution, the estimates are close to zero and are neither economically nor statistically significant.[12]

The confidence intervals are large, mostly considering the large sample size. Plausible reasons are the measurement error as well as the susceptibility to attenuation bias of the fixed effects estimator. Furthermore, it could be that grouped quantile regression loses a large amount of information by keeping only the constant in the second stage.

To assess the sensitivity of the results, I estimate other specifications with different control variables and a slight variation of the fixed effects. Figure 4 shows the results without control variables in the second stage. In this specification, the estimates are

---

[10] Recall that months are defined as year and month combination. To give an illustration, six years of data imply that there are 72 months effects.

[11] With the full set of controls and fixed effects, the estimates are close to zero and not statically significant at conventional levels also when using conventional OLS.

[12] The results are robust when the second stage regressions are not weighted by the number of observations in each group. Further, the results are similar, both, when clustering at the state level as well as when using two-way clustering at the month and county level.
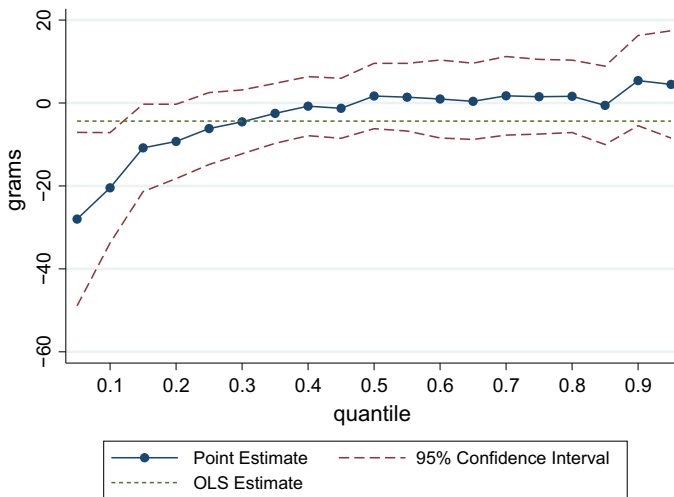
Fig. 2 Effect of $PM_{2.5}$ on the conditional distribution of birthweight. The figure illustrates the grouped quantile regression estimates of the effect of an increase in $PM_{2.5}$ exposure during gestation by one standard deviation on the conditional distribution of birthweight. The dashed green line shows the OLS point estimate. The dependent variable is birthweight in grams. Results are based on a sample of 9,016,342 individual births divided into 13,239 groups. Results are from the second stage using month and county fixed effects. Individual controls are included in the first stage, while the second stage includes weather, income, and transfers controls. Regressions are weighted for the number of births in each group. Heteroscedasticity robust standard errors clustered at the county level are used to construct the 95% confidence interval.

shifted upwards; however, without many differences. Figure 5 illustrates the results of the regressions with only weather controls in the second stage. These results are almost identical to those including income and transfers controls. Figure 6 shows the result of the second stage including a different set of fixed effects. As in the other regressions, the model uses county fixed effects; however, instead of including fixed effects for every single month and year combination, I use fixed effects for the month of the year and the year itself. The results are close to those in Fig. 2. Nonetheless, in the higher half of the distribution, the coefficients are closer to zero, and a few more quantiles are negative.

The findings suggest that there is a negative effect of $PM_{2.5}$ pollution only in the lower tail of the conditional distribution. While this paper does not study the incidence of prematurity, the larger impact on pollution in the lower tail of the distribution might suggest that the prematurity channel may be important as suggested in Currie et al. (2009) and Currie and Walker (2011). Nonetheless, without additional analysis, further conclusions about the causal chain are not possible.

Although I focus on a different pollutant, my results are consistent with previous studies. As Chay and Greenstone (2003b), I find only minimal effects of pollution with no economic implication on average birthweight. Moreover, consistent with Chay and Greenstone (2003b), Currie et al. (2009), and Knittel et al. (2016), my results suggest that the negative impact of pollution is larger among more vulnerable infants. It should, however, be noted that none of these studies focuses on $PM_{2.5}$, but on PM_{10} or TSPs,
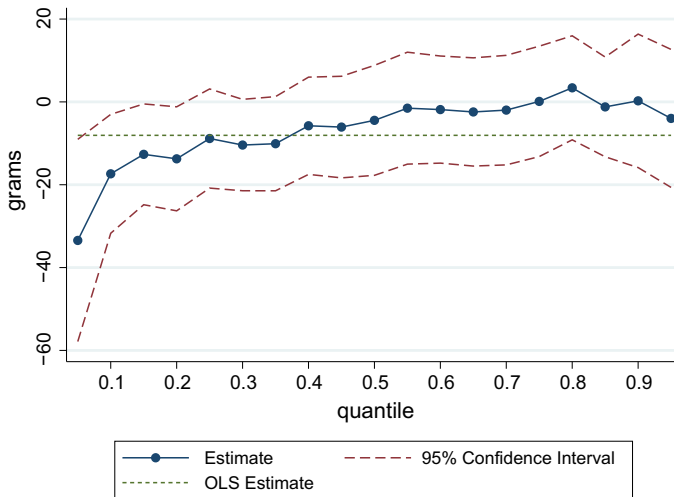
Fig. 3 Effect of PM_{10} on the conditional distribution of birthweight. The figure illustrates the grouped quantile regression estimates of the effect of an increase in PM_{10} exposure during gestation by one standard deviation on the conditional distribution of birthweight. The dashed green line shows the OLS point estimate. The dependent variable is birthweight in grams. Results are based on a sample of 7,793,245 individual births divided into 10,603 groups. Results are from the second stage using month and county fixed effects. Individual controls are included in the first stage, while the second stage includes weather, income, and transfers controls. Regressions are weighted for the number of births in each group. Heteroscedasticity robust standard errors clustered at the county level are used to construct the 95% confidence interval.

which have some similarities, or CO, which differs substantially in its characteristics and in how it might affect health.

Two additional points regarding the results should be noted. First, if pollution were to affect stillbirth or miscarriage, this analysis would be based on a potentially selected sample of living births (Chay and Greenstone 2003a, b). The bias caused by censoring might also vary across the distribution. If pollution affects miscarriage disproportionately among infants that would have had poor birth outcomes, then this potential selection bias would be larger in the lower tail. In this case, the estimates should be regarded as lower bounds. A second concern with the results is the measurement error associated with the approximation used to assign pollution to women. A short discussion of this issue is included in Supplementary Appendix, and future research should explore this further.

4.3 Alternative pollutant

To assess the plausibility of the findings, I use the same dataset to estimate the effect of PM_{10} on birthweight. PM_{10} is defined as the concentration of particulate matter smaller or equal to 10 micrometers per cubic meter during pregnancy. As for this pollutant there are different missing values, these results rely on a partly different sample with 7,793,245 individual births and 10,603 groups. The summary statistics of this sample are in Supplementary Appendix in Table A.2.

Despite the similarity between these contaminants, they have a relatively low correlation (0.1197), and from the summary statistics, it is clear that PM_{10} is assigned differently compared to $PM_{2.5}$. For instance, there is only a minimal correlation between PM_{10} and the variable black. Moreover, the relationships between PM_{10} and Hispanic, cigarettes, or mother's education appear to go in the opposite direction compared to $PM_{2.5}$. The same is true also for weather variables. Additionally, average personal income and most of the transfer variables are lower in higher quartile of PM_{10} exposure, which is the opposite pattern compared to $PM_{2.5}$. For these reasons, PM_{10} allows estimating the effect of a quite similar pollutant, whose exposure appears confounded in a distinct way by observable variables.

Figure 3 shows the result of the regression estimating the effect of PM_{10} on the conditional distribution of birthweight. As before, pollution is normalized to have a mean of zero and a standard deviation of one. Further supporting the previous findings, the results are similar to those of $PM_{2.5}$ and suggest a higher effect of pollution in the lower tail of the conditional distribution. An increase in pollution by one standard deviation is associated with a decline in birthweight by 33.5 g in the 5th and by 17.4 g in the 10th conditional percentile, whereas the decline in the 15th and 20th percentiles is of 12.7 and 13.8 g, respectively. In higher quantiles, the coefficients are close to zero.

5 Conclusion

This paper uses grouped quantile regression to estimate the effect of pollution on the conditional distribution of birthweight. The magnitude of the effect varies substantially across the distribution, suggesting not only that it is crucial to consider the distributional effects, but also that estimates of the average impact might underestimate the adverse effect of pollution. I find a negative effect of $PM_{2.5}$ only in the lower tail of the distribution. The estimates suggest that an increase in $PM_{2.5}$ by one standard deviation leads to a reduction of birthweight by 28.0, 20.5, and 10.8 g in the 5th, 10th, and 15th percentiles. These results are also consistent across different specifications.

To further assess the robustness of my findings, I estimate the same model for PM_{10}, a pollutant that appears assigned in a distinct way compared to $PM_{2.5}$. Supporting the results, I find a similar pattern, with the strongest effect in the lower tail of the distribution and no effect in the higher tail.

When interpreting these results, a few caveats should be considered. In this analysis, I estimated the effect of pollution on the conditional distribution of birthweight. Given the greater interest for the unconditionally small infants, it would be relevant to estimate the impact on the unconditional quantiles. Further, this paper analyzes the effect of pollution during the nine months before birth. However, the effect of pollution on birthweight might vary depending on the stage of the pregnancy. Future research should explore different definitions of the treatment.
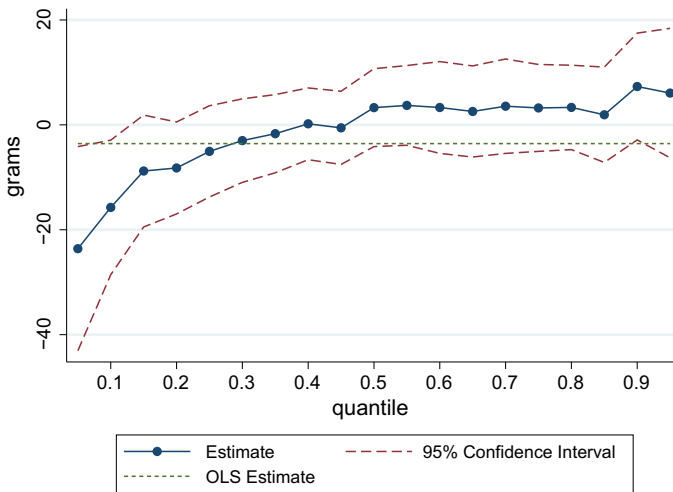
Declaration

A Appendix

Fig. 4 Effect of $PM_{2.5}$ on the conditional distribution of birthweight: Specification 2. The figure illustrates the grouped quantile regression estimates of the effect of an increase in $PM_{2.5}$ exposure during gestation by one standard deviation on the conditional distribution of birthweight. The dashed green line shows the OLS point estimate. The dependent variable is birthweight in grams. Results are based on a sample of 9,016,342 individual births divided into 13,239 groups. Results are from the second stage using month and county fixed effects. Individual controls are included in the first stage, while no control variables are included in the second stage. Regressions are weighted for the number of births in each group. Heteroscedasticity robust standard errors clustered at the county level are used to construct the 95% confidence interval.
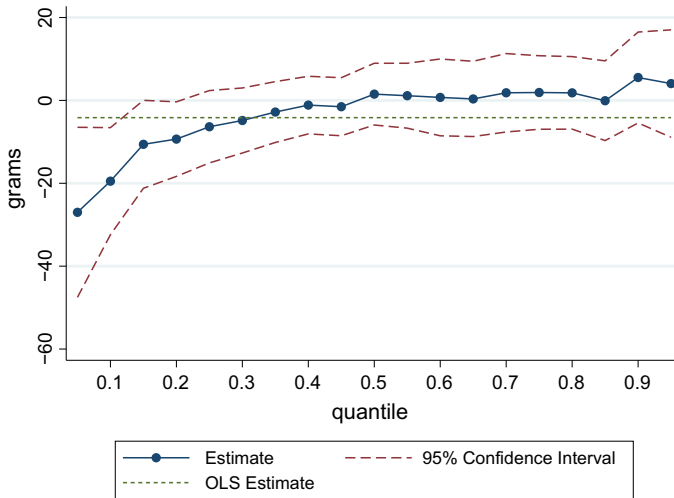
Fig. 5 Effect of $PM_{2.5}$ on the conditional distribution of birthweight: Specification 3. The figure illustrates the grouped quantile regression estimates of the effect of an increase in $PM_{2.5}$ exposure during gestation by one standard deviation on the conditional distribution of birthweight. The dashed green line shows the OLS point estimate. The dependent variable is birthweight in grams. Results are based on a sample of 9,016,342 individual births divided into 13,239 groups. Results are from the second stage using month and county fixed effects. Individual controls are included in the first stage, while the second stage includes weather controls. Regressions are weighted for the number of births in each group. Heteroscedasticity robust standard errors clustered at the county level are used to construct the 95% confidence interval.
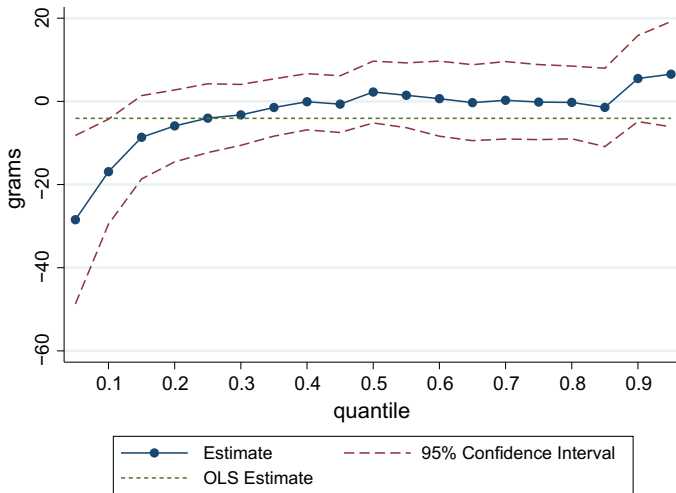
Fig. 6 Effect of PM$_{2.5}$ on the conditional distribution of birthweight: Specification 4. The figure illustrates the grouped quantile regression estimates of the effect of an increase in PM$_{2.5}$ exposure during gestation by one standard deviation on the conditional distribution of birthweight. The dashed green line shows the OLS point estimate. Results are based on a sample of 9,016,342 individual births divided into 13,239 groups. Results are from the second stage using county fixed effects as well as year and month fixed effects. In this regression, month fixed effects can be seen as a dummy for each of the 12 months of the year. Individual controls are included in the first stage, while the second stage includes weather, income, and transfers controls. Regressions are weighted for the number of births in each group. Heteroscedasticity robust standard errors clustered at the county level are used to construct the 95% confidence interval.

References

Abrevaya J (2001) The effects of demographics and maternal behavior on the distribution of birth outcomes. Empir Econ 26(1):247–257

Abrevaya J, Dahl CM (2008) The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. J Bus Econ Stat 26(4):379–397

Almond D, Chay KY, Lee DS (2005) The costs of low birth weight. Q J Econ 120(3):1031–1083

Canay IA (2011) A simple approach to quantile regression for panel data. Econom J 14(3):368–386

Chamberlain G (1994) Quantile regression, censoring, and the structure of wages. Adv Econ 1:171–209

Chay KY, Greenstone M (2003a) Air quality, infant mortality, and the clean air act of 1970. NBER Working Paper (w10053)

Chay KY, Greenstone M (2003b) The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. Q J Econ 118(3):1121–1167

Chernozhukov V, Fernández-Val I (2011) Inference for extremal conditional quantile models, with an application to market and birthweight risks. Rev Econ Stud 78(2):559–589

Chetverikov D, Larsen B, Palmer C (2016) Iv quantile regression for group-level treatments, with an application to the distributional effects of trade. Econometrica 84(2):809–833

Currie J (2011) Inequality at birth: some causes and consequences. Am Econ Rev 101(3):1–22

Currie J, Neidell M (2005) Air pollution and infant health: what can we learn from california's recent experience? Q J Econ 120(3):1003–1030

Currie J, Walker R (2011) Traffic congestion and infant health: evidence from e-zpass. Am Econ J Appl Econ 3(1):65–90

Currie J, Neidell M, Schmieder JF (2009) Air pollution and infant health: lessons from New Jersey. J Health Econ 28(3):688–703

Environmental Protection Agency (2019) Particulate Matter (PM) Pollution. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics. Accessed 23 April 2019

Gehring U, Tamburic L, Sbihi H, Davies HW, Brauer M (2014) Impact of noise and air pollution on pregnancy outcomes. Epidemiology 25(3):351–358

Graff Zivin J, Neidell M (2013) Environment, health, and human capital. J Econ Lit 51(3):689–730

Greenstone M, Gayer T (2009) Quasi-experimental and experimental approaches to environmental economics. J Environ Econ Manag 57(1):21–44

Hausman J (2001) Mismeasured variables in econometric analysis: problems from the right and problems from the left. J Econ Perspect 15(4):57–67

Hausman J, Liu H, Luo Y, Palmer C (2019) Errors in the Dependent Variable of Quantile Regression Models. No. w25819. National Bureau of Economic Research

Kato K, Galvao AF, Montes-Rojas GV (2012) Asymptotics for panel quantile regression models with individual effects. J Econom 170(1):76–91

Kelly FJ, Fussell JC (2012) Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter. Atmos Environ 60:504–526

Knittel CR, Miller DL, Sanders NJ (2016) Caution, drivers! children present: traffic, pollution, and infant health. Rev Econ Stat 98(2):350–366

Koenker R (2004) Quantile regression for longitudinal data. J Multivar Anal 91(1):74–89

Koenker R, Hallock KF (2001) Quantile regression. J Econ Perspect 15(4):143–156

Larssen S, Hagen LO (1997) Air Quality in Europe, 1993: A Pilot Report. European Environment Agency

Slama R, Darrow L, Parker J, Woodruff TJ, Strickland M, Nieuwenhuijsen M, Glinianaia S, Hoggatt KJ, Kannan S, Hurley F, Kalinka J, Šrám R, Brauer M, Wilhelm M, Henrich J, Ritz B (2008) Meeting report: atmospheric pollution and human reproduction. Environ Health Perspect 116(6):791–798

United States Census Bureau (2001) Population by Age, Sex, Race, and Hispanic or Latino Origin for the United States: 2000. https://www.census.gov/data/tables/2000/dec/phc-t-09.html. Accessed 24 April 2019