Luke Burstein, Andrew Howe, Percy Pham (Group 3)

ECON 298: Econometrics

Group Replication Project: Lead and Mortality

I. PURPOSE

The purpose of this report is to replicate the results of the paper *Lead and Mortality* written by Karen Clay, Werner Troesken, and Michael Haines using Stata, discuss the replication results, and briefly introduce possible extensions to the phenomena in the study

a. Motivation

In *Lead and Morality (2014)*, the authors analyze how waterborne lead exposure affected the rate of infant mortality between 1900-1920 in urban centers of the USA. During this period, almost 23% of the American populous had exposure to lead through the water systems in their cities. The amount of lead that an individual in one of these cities was exposed to was the result of the chemical properties of their city's respective water source. For example, areas with water having higher acidity (lower pH) had lead leach into their water systems. The authors choose to look at infant mortality because it is a strong and measurable proxy for general population health. Lead exposure during the prenatal stage and infancy is more likely to result in death. Adult exposure more often results in sickness making it harder to quantify. Until the 1970s, high concentrations of lead in water was thought to be mostly safe. While the negative health effects of lead exposure are evident today, very little retroactively research has been done on leads impact during the early 20th century on a national scale. The research done in this paper is important for two primary reasons. One, it generally adds to the existing literature on the adverse health effects lead has for the population. While lead piping in cities is no longer a major issue today, this work could help prompt policy aimed at limiting other sources of lead exposure. For example, lead-

based paint is still a widespread issue in the United States. Second, the findings are a valuable extension of literature that explains the declining rates of infant mortality rates in the United States over the 20th century.

b. Data and Methods

The study looks at data from 172 cities, home to nearly 84% of America's urban population. Infant mortality statistics were collected every 5 years from the majority of these cities. Data on water characteristics was collected from a number of different sources including but not limited to IPUMs, U.S. Geological Survey, or local resources. The most important of these measurements was the level of acidity or pH in a city's water. For the analysis, cities with only lead pipes were divided into groups based on their pH level quartile and compared to average infant mortality in those areas. As part of this analysis, the authors factor in other variables such as changes in milk quality, which would influence infant mortality. The author's further divide the sample by city size. The authors then use regression analysis to find out the correlation between variables.

c. Results and Conclusions

The authors found that cities with lead-pipes had higher infant mortality than those without. Furthermore, regression analysis shows that lower pH in cities with lead piping was associated with higher mortality in 1900. Additionally, the results seem to indicate that lead levels in water increased quickly in cities with pH levels below 7.3. Introducing a "time" variable to the regression evidences this causality further. Between 1900-1920, the author expected that the corrosiveness of lead piping would diminish, improvements would be made in water treatment, and lead piping would be phased out by safer alternatives. As a result, if lead was a causal factor, infant mortality rates would be expected to decline. The results corroborate this hypothesis. The authors observed

child mortality rates between 1900-1920 falling faster in cities with lead pipes than in those without. Overall, the study concludes that between 1900 and 1920 water leeching into the water supply was directly related to increases in the rate of infant mortality. The level of pH was an important variable in this relationship. Cities with lead pipes and lower levels of water acidity had significantly lower levels of infant mortality than those cities with lead pipes and higher levels of water acidity.

Our group will attempt to replicate these results to identify the determinants of cities with lead pipes in 1900 and the effects of lead, pH, and hardness on infant mortality in 1900.

II. Replication Data and Methods

The data provided for replication is only a subset of the original data used in the original study. The replication data contains data of 172 U.S cities in 1900, instead of data of cities from 1900 to 1920 like the original data. Therefore, it is impossible to replicate the regression analysis used to identify the effects of lead and pH on infant mortality from 1900-1920. Also, the replication data does not have data for noninfant rate and percent white; therefore, it is impossible to replicate the regression analysis to identify the effects of lead and pH on noninfant mortality in 1900 and percent white cannot be included in other replications.

The data for lead pipes is not divided the same way as the original data – in the original data, it is divided into 3 categories – lead only, mixed lead, and no lead, whereas in the replication data, it is divided into 2 categories only – no lead and have lead pipes (combining both mixed lead and lead only together). The replication data has data for the name of the cities; however, it is not divided into different regions (e.g. Mid-Atlantic, East North Central, New England) so that the effect of those variables can be easily analyzed in the replicated regression.

The rest of the replication data is similar to the original data, and there are 172 observations in the replication data, the same as in the original data. The table below shows the summary statistics of the replication data (excluded year, city, state, and age):

Variable	Mean	Min	Max	Label
hardness	113.436	2	445	Water hardness index
ph	7.322674	5.7	8.9	Water pH
infrate	.396194	.1097561	.8444657	Infant mortality rate (deaths per 100 in population)
typhoid_rate	.0410789	0	.1442786	Typhoid death rate
np_tub_rate	.0204759	0	.0574163	Non-pulmonary tuberculosis death rate
mom_rate	.1989602	.1308901	.323741	Fraction of population who are women of child-bearing age
population	1087.962	85	34355.54	City population (in 100s)
precipitation	3.328629	.9546053	4.754057	Average precipitation in state
temperature	49.32409	40.81126	70.62281	Average temperature in state
lead	.6802326	0	1	Indicator =1 if city had lead pipes
foreign_share	.2169771	0	.5822222	Fraction of population who are foreign born

III. Replication Results

The replication results address the research question of whether waterborne lead exposure influences infant mortality. A linear regression of the natural log of (pH-5.675) on infant mortality rate based on data of 172 U.S cities in 1900 is shown in the following figure. It shows that pH, which is negatively associated with lead concentration in water, is also negatively associated with child morality. The variable ln(pH-5.675) is used to ensure the results are evaluated within the sample range and the data is normalized. For example, a pH of 6.675, the 25th percentile, would results in an ln(pH-5675) of 0.

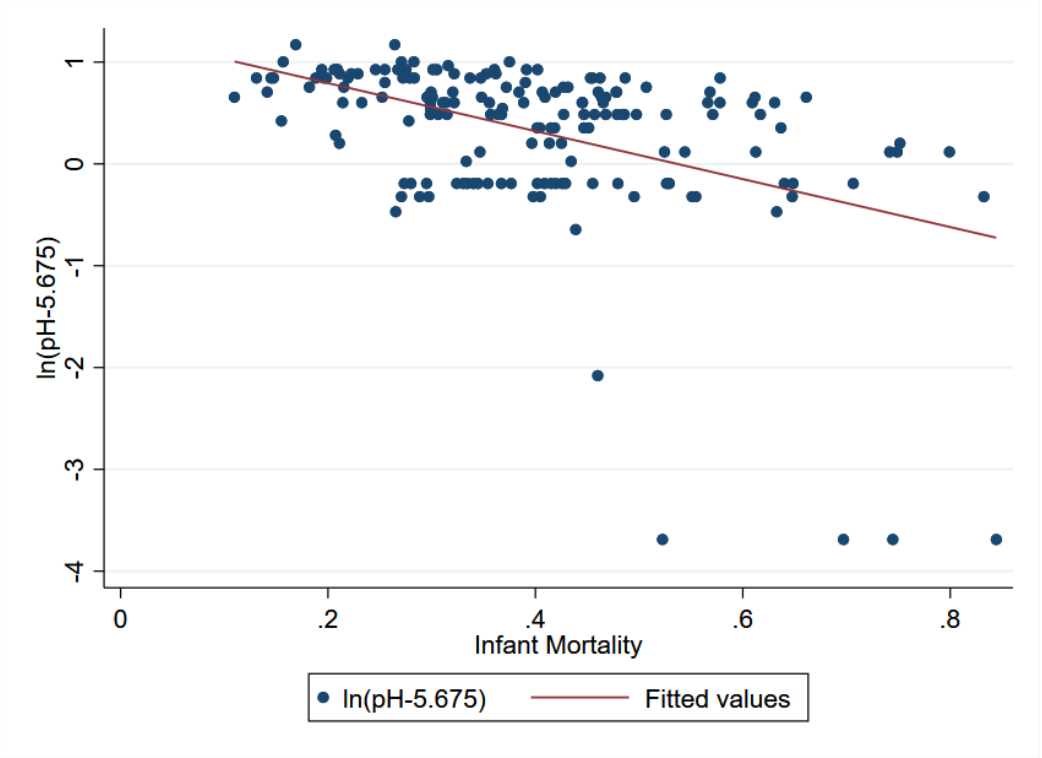

Figure 1 – A scatterplot for Infant Mortality Rate and ln(pH-5.675) of the cities

The replication regression analysis is summarized in the tables below. The tables are labeled to loosely match the labels of the tables in the original study

VARIABLES	(1) Lead	(2) Lead and pH ≤ 7.3
ln(pH - 5.675)	-0.132** (0.0589)	
ln(hardness)	0.147*** (0.0405)	0.00426 (0.0632)
Population	1.85e-05*** (6.46e-06)	8.73e-05** (3.66e-05)
Precipitation	-0.0636 (0.0712)	-0.716*** (0.236)
Temperature	0.00457 (0.00866)	0.0668*** (0.0120)
Typhoid_rate	2.263** (0.987)	-4.010* (2.170)
Observations	172	66
R-squared	0.119	0.223

Table 3 – Determinants of a City Having Lead Pipes in 1900
Note: Robust standard errors in parentheses. A constant was estimated but not reported.
*Statistically significant at *** p<0.01, ** p<0.05, * p<0.1*

VARIABLES	(1) Infant Mortality	(2) Infant Mortality	(3) Infant Mortality	(4) Infant Mortality
pH - 6.675				-0.0579 (0.0517)
Lead	0.0483** (0.0231)	0.0480** (0.0198)	0.0440** (0.0191)	0.0680* (0.0366)
Lead × (pH – 6.675)				-0.0384 (0.0594)
1(pH > 7.3)				0.0485 (0.0863)
Lead × 1(pH > 7.3)				-0.0200 (0.0982)
Typhoid			1.317*** (0.354)	1.313*** (0.384)
Nonpulmonary tuberculosis			1.653** (0.748)	1.797** (0.784)
ln(pH – 5.675)	-0.0592*** (0.0205)	-0.0262** (0.0132)	-0.0246* (0.0140)	
Lead × ln(pH – 5.675)	-0.0433* (0.0242)	-0.0375** (0.0161)	-0.0398** (0.0168)	
Demographics and other controls	No	Yes	Yes	Yes
Observations	172	172	172	172
R-squared	0.231	0.423	0.487	0.484

Table 4 – Effects of Lead and pH on Infant Mortality in 1900

*Note: Robust standard errors in parentheses. A constant was estimated but not reported. Demographic and other controls are fraction foreign born, fraction white, fraction women ages 20 to 40, state temperature, and state precipitation. Statistically significant at *** p<0.01, ** p<0.05, * p<0.1*

VARIABLES	(1) Infant Mortality Lead Only and Mixed-Lead	(2) Infant Mortality Lead Only and Mixed-Lead	(3) Infant Mortality Lead Only and Mixed-Lead
Hardness			-0.000629** (0.000243)
pH – 6.675			-0.170** (0.0653)
Hardness × (pH – 6.675)			0.000399 (0.000334)
1(pH > 7.3) × (pH – 6.675)			-0.0310 (0.0821)
Hardness × 1(pH > 7.3)			-2.92e-05 (0.000291)
Lead × 1(pH > 7.3)			-0.104 (0.0710)
ln(pH – 5.675)	-0.103*** (0.0129)	-0.0576*** (0.0193)	
ln(hardness)		-0.0445** (0.0177)	
Observations	117	117	117
R-squared	0.290	0.335	0.387

Table 5 – Effects of Lead and pH, and Hardness on Infant Mortality in 1900
Note: Robust standard errors in parentheses. A constant was estimated but not reported.
*Statistically significant at *** p<0.01, ** p<0.05, * p<0.1*

In *Table 3*, unlike in the original study, in column 1, the coefficients on ln(pH – 5.675), ln(hardness) and typhoid are statistically significant. The coefficient on population remains statistically significant, which suggests population affects the choice of using lead pipes. However, in column 2, the results are much different compared to column 1 as R-squared are larger, which suggests that the model in column 2 is a better fit, as we exclude ln(pH – 5.675) from the model and only include cities with pH ≤ 7.3. Coefficient for ln(hardness) becomes not statistically significant, while coefficients for precipitation and temperature become significant. The coefficient on population and typhoid remains significant. In sum, the replication model is more comparable to the original when only cities with pH ≤ 7.3 are analyzed, as city size and

temperature, and neither pH nor hardness, influence the choice of pipes; however, the model suggests that the water quality (typhoid) and precipitation can affect the choices. These differences can be partially attributed to the fact that city population isn't divided into quartile and the region is not included in the data to add them to the regression analysis.

In *Table 4*, it is notable that the inclusion of demographics and other controls significantly improved R-squared, which is the explanatory power of the model: R-squared jumps from 0.231 in column 1 to over 0.4 in the next three columns. The inclusion of nonpulmonary tuberculosis and typhoid in the model of column 3 also slightly improve the fit of the regressions as R-squared increases from 0.423 in column 2 to 0.487 in column 3. This trend is similar to what the original report observes. Also, the coefficients for lead, typhoid and lead × ln(pH – 5.675) remains statistically significant, which suggests that those variables affect infant mortality. However, there are still differences from the original report's regression as ln(pH – 5.675) and nonpulmonary tuberculosis becomes significant. But overall, the regression table here fits most of what the original paper reports that lead pipes and low pH were associated with higher city-level infant mortality, even though there are differences which can be blamed to the fact that many coefficients cannot be replicated, and lead variable is divided differently.

In *Table 5,* clearly shows that the addition of the other water characteristic variables such as hardness help improve the models' fitness. The R-squared observed increased from 29% up to 38.7%, which is a much more drastic increase compared to the original regression. However, the R-squared of this model is much lower than that in the original report (which is more than 60%). In general, this table lines up with most of the regression results in the original paper that the use of lead service pipes was related to infant mortality as the coefficients of ln(pH - 5.675), pH - 6.675, and hardness still remains statistically significant. There are still differences such as the

coefficient of Hardness × (pH – 6.675) and 1(pH > 7.3) × (pH – 6.675) is not statistically significant because of the difference in how categories of pipes are divided, but those differences do not affect much to the overall conclusion of the tables.

To sum up, the replication results corroborate the results shown in the original report by Clay et al. While the replication data from this experiment could not completely match the data from the study, the replication data supports most of the same conclusions that the authors make. The lack of variables that could be included in the regression (geographical region) or how different the data is divided (categories of pipes) may affect variable bias and decrease the validity of the results but it still holds most of the points that the authors discuss in the original report.

IV. Potential Extension

This study explored the phenomenon of infant mortality rate being affected by water pH and lead exposure. Examining the other effects of lead exposure could be a potential extension to the data already studied. For example, the authors mention that morbidity and IQ are related to lead exposure, and so access to data related to those variables would expand the scope of this study. The phenomena investigated by the study could also be extended by looking into high school graduation rates, further educational attainment, and how populations affected differently by lead exposure compare. There are also ways in which the data in the present study could have been extended to improve our understanding of the causal relationship between lead exposure and infant mortality rate.

Future studies could attempt to eliminate omitted variable bias by including other factors which may affect infant mortality rate other than those listed. This study placed a heavy focus on the effects of water acidity in lead pipes on infant mortality rate. Some of the variables involved in the study were the pH of the water, noninfant rate, percent lead only, percent white, and percent

women ages 20-40. However, factors such as income, education, or socioeconomic status altogether may affect the quality of water, therefore leading to bias that has not been accounted for in the study. Since these factors were ignored, the results from the study may be incomplete. A dataset that included more variables about the populations involved could yield more robust and sound results. Improving our understanding of the effects of lead exposure on issues such as infant mortality rate, IQ, and morbidity is important to increasing the quality of life and safety of populations. Therefore, future studies should consider some of these possible extensions to yield informative results regarding the effects of lead exposure on human life.