Statistics I Final Project

Rebecca Wagner, Adam Xu

Fall 2024

1 Find a research paper on a topic of interest that uses OLS or probit/logit/LPM.

1. **Provide a complete citation to this work. Please include a link (or send the paper) if possible. Discuss the key finding of the paper. Do the authors provide replication data? If so, provide a link/citation to data.**

   Fletcher, J. M. (2015). New evidence of the effects of education on health in the US: Compulsory schooling laws revisited. Social Science & Medicine, 127, 101-107. doi:10.1016/j.socscimed.2014.09.052. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6180294/

   The paper investigates the causal effect of education on health outcomes in the United States by revisiting the use of compulsory schooling laws as an instrumental variable. It utilizes a larget dataset than prior studies. Key findings include:

   (a) Health Improvement: Each additional year of education is associated with improved self-reported health outcomes, reductions in poor health indicators, and lower mortality rates.

   (b) Instrumental Variable (IV) Effects: The use of IVs significantly enhances the estimated impact of education on health compared to ordinary least squares (OLS) estimates, suggesting stronger causal links.

   (c) Subgroup Effects: The health benefits of education are more pronounced among individuals with lower educational attainment, who are more likely to be influenced by compulsory schooling laws.

   (d) Policy Implications: The study highlights the potential for educational policies to serve as tools for improving public health. However, the study also acknowledges limitations, including the representativeness of the data and the precision of the estimates, particularly due to sample size.

   The authors do not explicitly provide replication data or a direct link to the dataset used in the study.

2. **What is the analytical question? (If there are multiple analyses/questions addressed in the paper, focus on a single important question)**.

   The primary analytical question addressed in the paper is

   > *What is the causal effect of education on health outcomes in the United States, as estimated through the use of compulsory schooling laws as an instrumental variable?*

   This question is central to the paper's goal of understanding whether increases in educational attainment, driven by external policy changes (compulsory schooling laws), lead to measurable improvements in various health outcomes, including self-reported health, poor health indicators, and mortality rates.

3. **Why is it important to answer this question?**

   This question is important because it examines whether policies aimed at increasing educational attainment can have broader societal benefits, particularly in improving public health. By addressing potential issues of reverse causality and omitted variable bias using instrumental

variable analysis, the paper aims to provide robust causal evidence that supports the linkage between education and health.

4. **Write out a statistical model and pay special attention to your subscripts, as these indicate your unit of analysis (e.g., individuals at a given point in time, national data over time, or perhaps individual data over time). Be sure to define the dependent variable and the key independent variable (or, in some cases, variables).**

   The author explicitly defines their model as:

$$health_{isa} = \beta_0 + \beta_1 schooling_{isa} + \beta_2 X_{isa} + \lambda_s + \eta_a + \epsilon_{isa}$$

   Where the terms are:

   $\boldsymbol{health_{isa}}$ The dependent variable represents a health outcome (e.g., self-reported health, poor health indicators, or mortality) for individual i, born in state s, and cohort a.

   $\boldsymbol{schooling_{isa}}$ The key independent variable represents the predicted years of schooling from the first stage, reflecting the variation driven by compulsory schooling laws.

   $\boldsymbol{X_{isa}}$ Covariates include individual characteristics such as race, gender, marital status, and other socio-demographic variables.

   $\boldsymbol{\lambda_s}$ Fixed effects for state of birth.

   $\boldsymbol{\eta_a}$ Fixed effects for year of birth.

5. **Does the model use any specific functional form (e.g., probit, log)? If yes, explain why.**

   No, the primary model used in the paper does not explicitly rely on a specific functional form like probit or log. Instead, the analysis is based on linear regression models in both stages of the instrumental variable approach. Because the primary health outcomes in the paper (e.g., self-reported health on a 5-point scale or mortality rates) are treated as continuous variables, which are well-suited for linear regression models. While mortality could be modeled using a logit or probit approach, the paper focuses on linear regression for simplicity and comprability with previous studies.

6. **Which sources of endogeneity are the authors able to control for with additional control variables?**

   The authors address multiple potential sources of endogeneity in the relationship between education and health outcomes by including additional control variables in their models. These sources include:

   Omitted Variable Bias The authors control for socio-demographic variables X_{isa}such as race, gender, and marital status. These factors may simultaneously influence both education (e.g., access to schooling) and health outcomes (e.g., disparities in healthcare access).

   State-Level Effects By including state-of-birth fixed effects λ_s, the model accounts for unobserved factors specific to an individual's state of birth, such as differences in state education policies, economic conditions, and healthcare systems.

   Cohort Effects Birth-year fixed effects η_a control for generational differences, such as changes in compulsory schooling laws, public health improvements, or economic conditions over time, which might affect both education levels and health outcomes.

   Measurement Error Using compulsory schooling laws as an instrumental variable helps mitigate the measurement error in self-reported education levels by relying on policy-driven variation in schooling.

   While these control variables cannot address all potential sources of endogeneity, they significantly reduce biases

7. **Which sources of endogeneity are the authors unable to control for? Explain the implications for your analysis (only work through one such (important) example in detail). Use the conditions for omitted variable bias and our discussion of the implications of omitted variable bias to discuss whether the particular omitted factor will bias results upward or downward. If the paper is a RCT, explain a source of bias that the RCT is designed to avoid.**

   Families value of education If families moved from states with stricter compulsory schooling laws to states with more lenient laws (or vice versa) based on their preferences for education or health opportunities, this migration behavior introduces omitted variable bias. For example: Families valuing education highly may migrate to states with better schooling policies, resulting in children receiving more education and potentially better health outcomes. Conversely, families with lower health status may prioritize healthcare access over education quality, introducing a different bias.

   Direction of Bias If migration is non-random and correlates positively with both education and health outcomes (e.g., families prioritizing education and health simultaneously), the model may overestimate the causal effect of education on health (upward bias). Conversely, if migration is driven by factors such as economic hardship or poor health conditions, which correlate negatively with both education and health, the bias could underestimate the effect (downward bias).

   Economic Factors The authors are unable to know the economic status of their participants, which may also impact their health outcomes.

8. **Could measurement error in key independent variables cause endogeneity? Explain why or why not.**

   Measurement error in the key independent variable $schooling_{isa}$ could cause endogeneity.

   Due to the available data, the authors utilize state of birth rather than state of residence to classify compulsory schooling laws. Participants in the surveys may have moved states between birth and schooling at age 14, resulting in incorrect state laws.

   If participants move states randomly with respect to compulsory schooling laws, then the error is not systematic and should not have much impact on results other than to decrease precision.

   If participants move states *in response* to compulsory schooling laws, the error is systematic and will likely cause bias through endogeneity.

   The authors address this as a limitation in their study.

9. **Are any of the included variables possibly "post-treatment" variables?**

   It is unlikely that any of the included variables are post treatment variables. For this to be the case, an independent covariate would need to be causally affected by the treatment variable $schooling_{isa}$. Covariants include gender, race, and martial status.

   One may argue that marital status *may* be affected by compulsory schooling laws, if students are more likely to meet a future spouse at school. We argue this effect, if any, would be minimal.

10. **What factors affect the precision of the results? Is the test statistically powerful? What are the implications of your answer to this question for how you will interpret your results?**

    Several factor affect the precision of the results.

    Sample Size A small sample size can cause the results to be imprecise. The authors write explicitly that "Results appear underpowered, suggesting that further use of this methodology may require even larger and potentially unattainable sample sizes in the US"

    Multicolinearity If independent variables vary together, this affects our ability to identify individual results. We are not concerned with multicolinearity in this study.

    Measurement Error If variables are measured with error, this causes estimates to be less precise. As previously discussed, there is potential for measurement error in the independent variable $schooling_{isa}$.

The implication is that we need to approach these results with some caution.

Particularly with reference to the sample size, we know that the results of this study are somewhat underpowered. This would affect our ability to identify relationships between variables, and we may not identify them when they are there.

Also, the potential for measurement error of the independent variable means that we are less confident in our estimates

Therefore, we may be skeptical of any causal claims based on this study.

11. **Discuss, as appropriate, autocorrelation, heteroscedasticity, and multicollinearity and what the authors do (or should do) about them.**

    There is no time series data in this paper, so there is no concern with autocorrelation.

    There is potential for errors to be similar in states, due to similar external factors (economic conditions, etc). For this reason, the authors use clustered standard errors to by state. They also use robust standard errors to account for additional heteroscedasticity.

12. **Have the authors considered heterogeneous treatment effects? (That is, have they considered whether the effect of the key independent variable(s) differs across subgroups?) Describe how they do or could address this possibility.**

    The authors have considered heterogeneous treatment effects in a few ways. First, they include fixed effect variables in their model for both state of birth and year of birth.

    Additionally, they evaluate their model for both the entire sample, and for low income participants. They find that the results are much stronger for low income participants.

13. **Are the results generalizable?**

    There are two concerns when it comes to generalizability

    Whether the sample is similar to nationally representative samples" The authors compare their data (NIA/AARP) to a broader NIH sample that is assumed to be widely representative. They find that, while their sample is more white and more educated than the NIH data, heath indicators are similar.

    Whether the sample who gave social security number information is a selected sample The authors only use participants of the survey who provided a SSN in order to identify birth year, state, and mortality. When they run the model on the whole sample and compare to the model with SSNs, they find similar results.

    In general, the authors do believe their results to be generalizable.

2 Discuss a new research design: Develop a hypothesis to test on a question related to the above article. This hypothesis must be different than the hypothesis tested in the paper discussed in part 1. (For example, if the paper covers the effect of income on high school graduation in Mexico, you cannot simply propose research on the effect of income on high school graduation in the U.S.) If you are in doubt, be sure to see me.

1. **What is the analytical question?**

   Does/How does the political party of the governor predict the magnitude of compulsory school laws in that state.

2. **Why is it important to answer this question?**

   From the original paper, we may argue that educational policies have an affect on health outcomes, which means that understanding what states have implemented what policies and why is also a factor in overall health outcomes.

3. **Write out a statistical model and pay special attention to your subscripts, as these indicate your unit of analysis (e.g., individuals at a given point in time, national data over time, or perhaps individual data over time). Be sure to define the dependent variable and the key independent variable (or, in some cases, variables).**

   $SchoolingIndex_{S,T} = \beta_0 + \beta_1 GovernorsParty_{S,T} + \beta_2 GovernorsParty_{S,T-1}$
   $+ \beta_3 StateSenateParty_{S,T} + \beta_3 StateHouseParty_{S,T} + \beta_i X_{S,T}$

   Where the terms are:

   $\boldsymbol{SchoolingIndex_{S,T}}$ An index of the strength of compensatory schooling laws in one state for one governers term

   $\boldsymbol{GovenorsParty_{S,T}}$ A dummy variable indicating the political party of the governor (0 - democrat, 1 - republican) in one state for a term

   $\boldsymbol{GovernorsParty_{S,T-1}}$ A lagged dummy variable indicating the political party of the governor in the state during the previous term, accounting for carrying over of policies.

   $\boldsymbol{SenateParty_{S,T}}$ Continuous variable representing the proportion of senate seats held by democrats in the state during the governors term

   $\boldsymbol{StateHouseParty_{S,T}}$ Continuous variable representing the proportion of house seats held by democrats in the state during the governors term

   $\boldsymbol{\beta_i X_{S,T}}$ Our controling variables, as described below.

4. **What, in your opinion, is the single most important statistical issue that you need to address? Why?**

   We believe that statistical power will be the biggest issue in this analysis. As our unit of analysis is states over time, we are limited to the 50 states by the length of time, which is low enough to cause concern of low power.

   We would also be concerned with multicolinearity, because the governor, house, and senate are often related in their political leaning. This will cause a lack of precision.

5. **Does the model use any specific functional form (e.g., probit, log)? If yes, explain why.**

   No, our model does not use any specific functional form. We are not utilizing percentages (logs) as the scale of our data points should all be similar, and conceptually percent changes do not necessarily make the most sense.

   Also, we are not using and probit or logit because out independent variable is not dichotomous.

6. **Which sources of endogeneity would you expect to be able to control for with control variables?**

   We expect to be able to control for factors such as

   School Quality Depending on how often standardize testing is given in states, we may be able to include a measure of school quality for each governors term in their state.

   Economic Factors We should be able to control for economic differences across states and terms with a measure of state GPD per capita

   Demographic Factors We can also control for racial makeup in states, which may impact the voting/political tendancies of the state.

7. **Which sources of endogeneity would you expect to be unable to control for? Explain the implications for your analysis. Use the conditions for omitted variable bias and our discussion of the implications of omitted variable bias to discuss whether the particular omitted factor will bias results upward or downward. If you propose an RCT, explain a source of bias that the RCT will avoid. (That is, for an RCT paper, explain a source of endogeneity that would occur if an RCT were not used.)**

State Public Opinion Public opinions across the state on schooling initiatives would be difficult to measure, and may have a direct impact on compulsatory schooling laws through ballot initiatives, etc. It would also have an effect on political variables through the election of officials. In this way, it is correlated with our independent variable and affects our dependent variable.

Direction of Bias If public attitudes are aligned with governing political party, then they will be positively correlated with political variables and the schooling index, making the reported effect of the party's governance stronger than it actually is. On the other hand, if public attitudes are negatively related with governing political party, then the affect of the partys governance will be reported as weaker than it actually is.

8. **Could measurement error in the key independent variable cause endogeneity? Explain why or why not.**

   No, we would not likely see measurement error in the key independent variable of governor's political party, as this is clearly defined during election cycles.

9. **Are any of the included variables possibly "post-treatment" variables?**

   Some of the included variables could be post-treatment variables. For example, economic measures and school quality measures might be influenced by the key independent variable of Governor's political party. Including these variables in the model would control for part of the effect of political leaning on schooling laws, potentially biasing the estimates downward and obscuring the true causal relationship. To avoid this, these variables should either be excluded or carefully considered when interpreting the results.

10. **What factors affect the precision of the results? Is the test statistically powerful? What are the implications of your answer to this question for how you will interpret your results?**

    Some factors that may affect the precision of the results are

    Sample Size A limited number of U.S. states reduces precision

    Measurement Error Inaccuracies in the measure of schooling index may reduce precision.

    Variability in the Data If there is little variation in political leaning or schooling laws, it reduces statistical power.

    Multicollinearity High correlation between political variables (e.g., governor party and legislature party) inflates standard errors.

    If the results are imprecise (large standard errors), we must interpret insignificant findings cautiously. They might reflect limited power rather than the absence of a true effect. If precision is high and the results are significant, we can more confidently infer a relationship between state political leaning and compulsory schooling laws. Addressing precision issues, such as improving variable measurement, increasing sample size, or correcting for serial correlation, is critical for robust interpretation.

11. **Discuss, as appropriate, autocorrelation, heteroscedasticity, and multicollinearity and what the authors do (or should do) about them.**

    Autocorrelation With data across time, there is a high likelyhood for autocorrelation across points. Include a lagged dependent variable in the model to account for the correlation of errors over time.

    Heteroscedasticity Likely, error terms will not be consistent between states. Use heteroscedasticity-robust standard errors (e.g., White standard errors) to correct for non-constant variance.

    Multicollinearity There is a chance for multicolinearity between political variables. Use an auxiliary regression (e.g., Variance Inflation Factor, VIF) to test for relationships between independent variables. If multicollinearity is detected, consider removing or combining highly correlated variables

12. **Do you expect the treatment effects to be heterogeneous? (That is, will the effect of the key independent variable differ across sub-groups?) Describe how to address this possibility.**

    Yes, I expect the treatment effects to be heterogeneous across sub-groups. The effect of state political leaning on the strictness of compulsory schooling laws may differ based on economic conditions, regional characteristics (e.g., South vs. Northeast), or historical trends in education policy. For instance, states with stronger economies may implement stricter laws regardless of political leaning, while poorer states may show greater variability. To address this possibility, we can include interaction terms in the model. For example, interact political variables (e.g., governor party) with economic indicators or regional dummies to test for subgroup differences. Alternatively, we can run the regression on separate subsamples (e.g., states with high vs. low GDP) to identify varying effects. Comparing the results across sub-groups will help us understand how political leaning interacts with other factors to influence compulsory schooling laws.

13. **Are the results generalizable?**

    The results may have limited generalizability due to the specific context of the analysis. First, the unit of analysis is U.S. states, and the findings are tied to the American political and institutional framework. Therefore, the results may not directly apply to other countries with different political systems, education policies, and economic conditions. Second, the generalizability within the U.S. could also be limited if there are significant regional differences (e.g., political culture, economic disparities) or if the sample periodreflects unique historical circumstances. To improve generalizability, robustness checks across different time periods and sub-groups (e.g., by region or state economic status) should be conducted.