Econ 771: Health Economics II – Module 1 Empirical Exercise

Due — 5pm, September 23rd

Name: Amy Lim

1 Overview

The purpose of this assignment is to learn how to use panel data and causal inference, specifically difference-in-difference and event studies. We will be using "traditional" and "modern" empirical strategies to estimate the expansion of Medicaid on uncompensated care in hospitals. For this assignment, I used STATA, but moving forward, I will be learning and transitioning to R.

2 Data

The analysis depends on three data sources: Hospital Cost Reporting Information System (HCRIS), Provider of Service (POS), and selected insurance data from the American Community Survey (ACS). Data are available online, and for this assignment, I downloaded all of the data from our class OneDrive folder.

2.1 Hospital Cost Reporting Information System (HCRIS)

Medicare-certified providers are required to submit cost reports to CMS. Since data are submitted by providers, there exists some reporting inconsistencies. From this data set, we will be utilizing payment and revenue variables to construct the outcome variable: uncompensated care.

2.2 Provider of Service (POS)

POS is gathered by CMS and updated quarterly to provide information on characteristics such as: provider certification, termination, accreditation, ownership, name, location and other aspects organized by CMS provider number. This data set helps us identify hospital type, in particular, whether they are a private not-for-profit or a private for-profit hospital for our analysis.

2.3 Insurance Data from ACS

This data set comes from ACS, but I will be using version of this data is complied by the Kaiser Family Foundation (KFF) called Medicaid Expansion. KFF documented the year each state expanded medicaid or if they have never expanded. This data will help us identify our treatment.

3 Data Cleaning

Using the described data, we can merge them to answer the assigned questions. After downloading the data from OneDrive, I had to use RStudio to create text files that could be read into any statistical software. Upon careful inspection, I started with the POS because it was the largest.

I limit the POS data set to only include private for-profit and private not-for profit hospital. Since we are working with 17 years of data in our assignment (2003-2019), I assumed that hospitals could switch types. In other words, hospitals could be for-profit in one year and change their status in subsequent years and vice-versa. After cleaning the POS data, I move on and clean HCRIS. Since HCRIS is submitted by providers to CMS, there are errors in variables of interest such as state. For example, one hospital in Arizona submitted their data with "Arizona" as the state variable and not as the state abbreviation as specified by CMS. These inconsistencies were accounted for with some manual data cleaning. Once the appropriate updates were made, I linked POS and HCRIS by their CMS provider number, state, and year. Finally, I merge the Medicaid Expansion data after creating indicators of when the states expanded medicaid for our analysis. This final merge nice because the expansion data only includes the 50 states plus the District of Columbia and all US territories should be dropped in this data set.

The data set now contains all the identifiers for us to do the analysis. We have a total of $56,104$ observations.

4 Questions

*Question 8 will be answered in R. The package was too difficult to understand in Stata.

1. Provide and discuss a table of simple summary statistics showing the mean, standard deviation, min, and max of hospital total revenues and uncompensated care over time. Revenue is calculated by using the given variable tot_pat_care from the inial data cleaning steps with R code porvided by Ian. Uncompensated care is calculated by adding the given variables tot_uncomp_care_charges, bad_debt, and uncomp_care then subtracting tot_uncomp_care_partial_pmts.

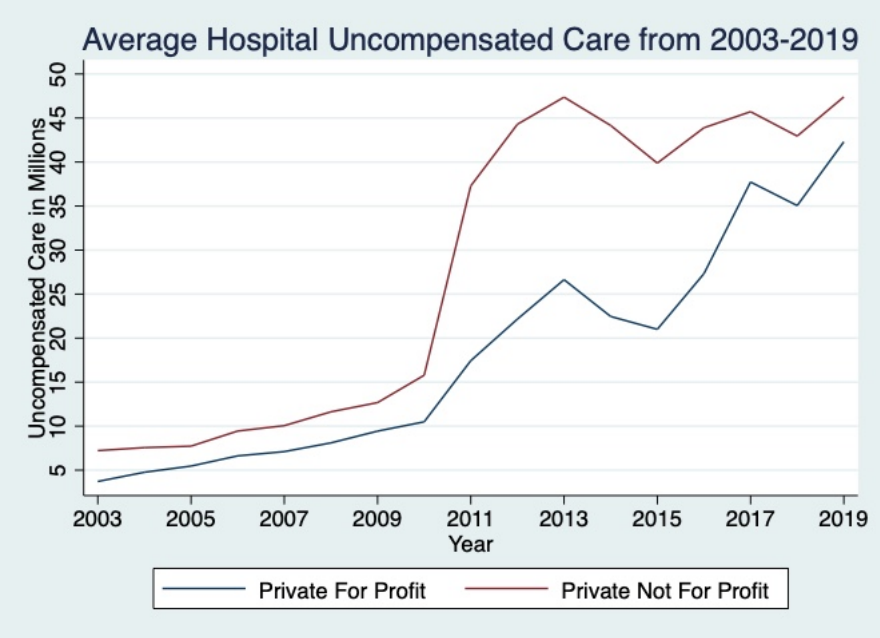
Table 1: Summary Statistics for Revenue

Year	Sum	Mean	SD	Min	Max	N
2003	481,953	187.90	335.27	0	4,723	2,565
2004	542,115	204.03	366.18	0	5,526	2,657
2005	631,736	223.78	393.93	0	6,164	2,823
2006	751,066	256.51	472.87	-0	6,718	2,928
2007	819,546	277.53	510.32	0	7,126	2,953
2008	913,552	306.87	557.52	0	7,743	2,977
2009	1,037,183	336.42	622.97	0	9,139	3,083
2010	1,181,750	365.30	695.48	0	10,185	3,235
2011	1,314,255	390.91	746.25	0	10,572	3,362
2012	1,467,234	423.69	811.67	0	11,865	3,463
2013	1,605,932	453.78	879.76	0	12,752	3,539
2014	1,771,522	490.73	950.94	0	13,376	3,610
2015	1,937,473	528.64	1,005.73	0	14,144	3,665
2016	2,144,186	570.41	1,094.84	0	15,619	3,759
2017	2,337,703	613.41	1,201.93	0	16,863	3,811
2018	2,565,352	672.09	1,337.54	0	18,677	3,817
2019	2,839,151	736.10	1,488.99	0	22,001	3,857
Total	24,341,709	433.87	914.58	-0	22,001	56,104

Table 2: Summary Statistics for Uncompensated Care

Year	Sum	Mean	SD	Min	Max	N
2003	14,517	5.66	17.90	0	304	2,565
2004	16,765	6.31	23.98	0	820	2,657
2005	18,867	6.68	25.76	0	939	2,823
2006	23,931	8.17	33.34	0	1,075	2,928
2007	25,914	8.78	29.64	0	736	2,953
2008	30,168	10.13	33.96	0	993	2,977
2009	34,938	11.33	31.35	0	584	3,083
2010	44,004	13.60	59.00	0	2,794	3,235
2011	97,690	29.06	73.20	-7	1,390	3,362
2012	122,009	35.23	93.17	-7	2,773	3,463
2013	138,210	39.05	90.43	0	1,556	3,539
2014	128,378	35.56	88.06	-3	1,727	3,610
2015	118,553	32.35	86.79	-0	2,132	3,665
2016	140,100	37.27	99.58	0	2,330	3,759
2017	161,981	42.50	109.06	-0	1,699	3,811
2018	151,934	39.80	104.53	-0	2,247	3,817
2019	175,100	45.40	132.39	-195	3,039	3,857
Total	1,443,058	25.72	80.57	-195	3,039	56,104

2. Create a figure showing the mean hospital uncompensated care from 2003 to 2019. Show this trend separately by hospital ownership type (private not for profit and private for profit).



3. Using a simple DD identification strategy, estimate the effect of Medicaid expansion on hospital uncompensated care using a traditional two-way fixed effects (TWFE) estimation:

$$y_{it} = \alpha_i + \gamma_t + \delta D_{it} + \epsilon_{it} \tag{1}$$

where $D_{it} = 1(E_i \leq t)$ in Equation (1) is an indicator set to 1 when a hospital is in a state that expanded as of year t or earlier, γ_t denotes time fixed effects,α_i denotes hospital fixed effects, and y_{it} denotes the hospital's amount of uncompensated care in year t. Present four estimates from this estimation in a table: one based on the full sample (regardless of treatment timing); one when limiting to the 2014 treatment group (with never treated as the control group); one when limiting to the 2015 treatment group (with never treated as the control group); and one when limiting to the 2016 treatment group (with never treated as the control group). Briefly explain any differences.

Table 3: Twoway Fixed Effects

	(1) Full Sample	(2) 2014	(3) 2015	(4) 2016
Treatment	-28635025.8***	-31934865.0***	-30540727.3***	-41084619.9***
	(2362750.9)	(2752682.1)	(2892737.9)	(3250255.8)
Constant	2572156.5**	2763023.2**	1293938.1	1529514.2
	(1220579.6)	(1387664.5)	(2107883.3)	(2109954.4)
Observations	56,104	47,282	27,166	25974
R^2	0.093	0.102	0.103	0.108

Standard errors in parentheses

Dependent variable: Hospital Uncompensated Care.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

From Table 3, we can see that the expansion of Medicaid has a negative effect on uncompensated care. More specifically, the average treatment effect of Medicaid expansion reduces the amount of uncompensated care for a hospital. Furthermore, the largest reductions occur for hospitals who are early adopters.

4. Estimate an "event study" version of the specification in part 3:

$$y_{it} = \alpha_i + \gamma_i + \sum_{\tau<-1} D_{it}^{\tau}\delta_{\tau} + \sum_{\tau>=0} D_{it}^{\tau}\delta_{\tau} + \epsilon_{it} \tag{2}$$

where $D_{it}^{\tau} = 1(t - E_i = \tau)$ in Equation (2) is essentially an interaction between the treatment dummy and a relative time dummy. In this notation and context, τ denotes years relative to Medicaid expansion, so that $\tau = -1$ denotes the year before a state expanded Medicaid, $\tau = 0$ denotes the year of expansion, etc. Estimate with two different samples: one based on the full sample and one based only on those that expanded in 2014 (with never treated as the control group).

Table 4: Event Study

	(1) Full Sample	(2) 2014
F3event	20946394.6***	27186718.1***
	(2830285.6)	(3449253.1)
F2event	16405546.3***	25890460.7***
	(1580278.2)	(2572226.6)
L0event	6078895.1***	14262802.9***
	(1333313.4)	(2113951.4)
L1event	-3553920.8***	6782626.0***
	(1324033.8)	(2236048.2)
L2event	-10197492.7***	-3214465.6
	(1475032.1)	(2321196.6)
L3event	-16777839.5***	-14050842.5***
	(1387788.7)	(2090304.9)
L4event	-16166085.6***	-13641397.8***
	(1276954.7)	(1674991.2)
Constant	20591121.1***	20728878.9***
	(919215.2)	(1010689.9)
Observations	55664	46915
r2	0.642	0.650

Standard errors in parentheses

Dependent variable: Hospital Uncompensated Care.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5. Sun and Abraham (SA) show that the δ_τ coefficients in Equation (2) can be written as a non-convex average of all other group-time specific average treatment effects. They propose an interaction weighted specification:

$$y_{it} = \alpha_i + \gamma_i + \sum_e \sum_{\tau \neq -1} (D_{it}^\tau \times 1(E_i = e))\delta_{e,\tau} + \epsilon_{it} \tag{3}$$

Re-estimate your event study using the SA specification in Equation (3). Show your results for $\hat{\delta}_{e,\tau}$ in a Table, focusing on states with $E_i = 2014$, $E_i = 2015$, and $E_i = 2016$.

Table 5: Sun and Abraham 2020

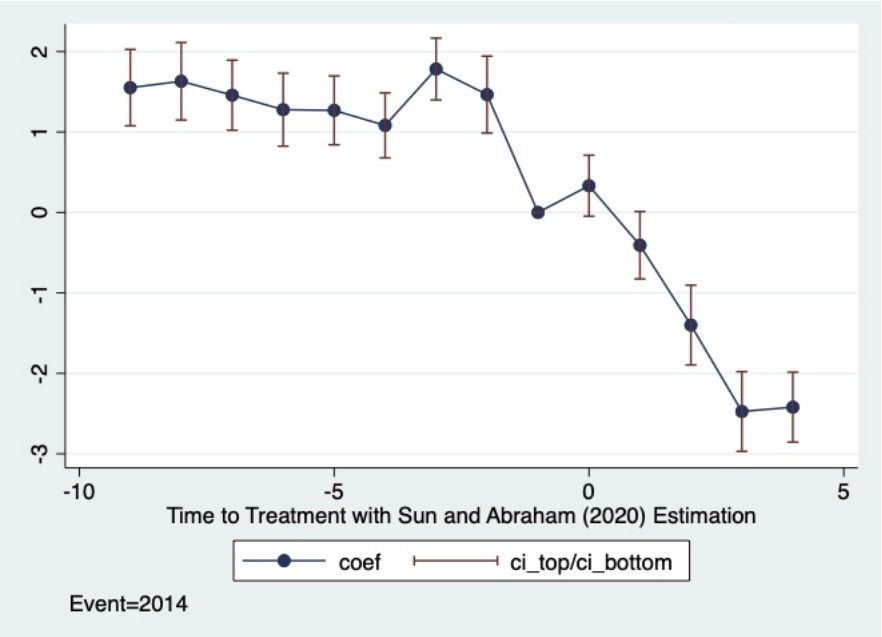
	(1) 2014	(2) 2015	(3) 2016
__00000V	15516571.7***	2975525.5	2894290.6**
	(2422846.5)	(2199162.4)	(1223864.2)
__00000Z	16308490.6***	2266161.4	647647.8
	(2456909.0)	(2077032.3)	(1417107.0)
__000013	14579929.4***	606255.5	-199361.6
	(2219016.5)	(2102063.9)	(1360326.9)
__000017	12776377.2***	459069.7	-1882224.0
	(2312766.5)	(1979845.9)	(1350303.4)
__00001B	12689830.8***	9988585.9	-13194822.0***
	(2185180.2)	(11202221.3)	(2046081.0)
__00001F	10822009.9***	-8105960.5**	-18602008.3***
	(2061714.2)	(4000347.0)	(2886161.4)
__00001J	17830578.6***	-13368962.8***	-22285936.4***
	(1958473.9)	(3538014.0)	(2797497.7)
__00001N	14652225.2***	-26693412.4***	-19068399.0***
	(2439189.1)	(3527175.6)	(3248855.0)
__00001R	3323159.3*	-37440308.5***	-29485102.4***
	(1932983.7)	(4017723.6)	(4110753.0)
__00001V	-4073720.9*	-38223175.7***	-52134224.1***
	(2136422.6)	(3874486.8)	(3914785.3)
__00001Z	-13997062.7***	-44857416.7***	-50229645.6***
	(2528705.0)	(4751015.3)	(4361749.1)
__000023	-24737843.4***	34199675.3***	-61525516.2***
	(2527839.3)	(305843.2)	(5004653.4)
Constant	26670474.2***	34199675.3***	33623473.1***
	(451757.0)	(305843.2)	(166893.9)
Observations	46915	26918	25735
r2	0.649	0.650	0.662

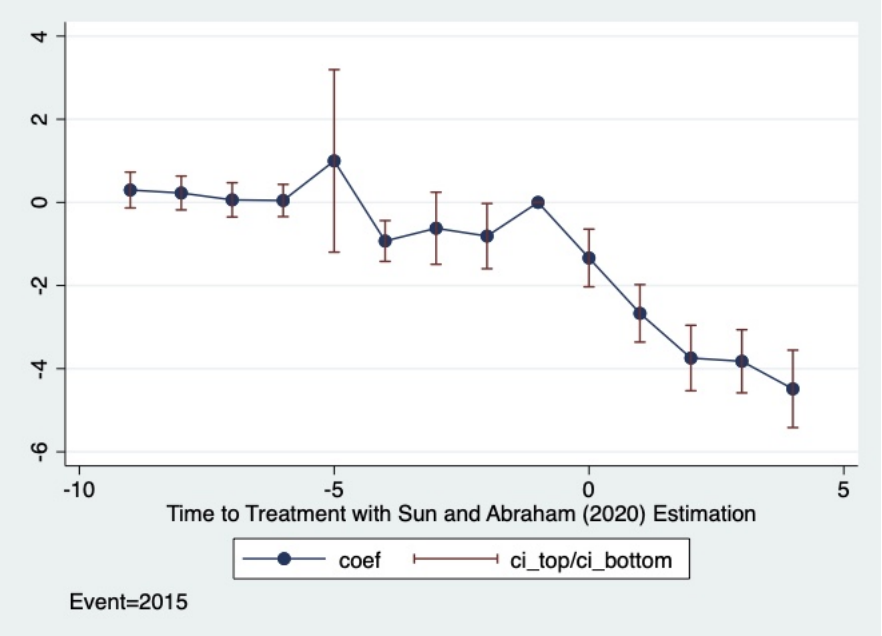
Standard errors in parentheses

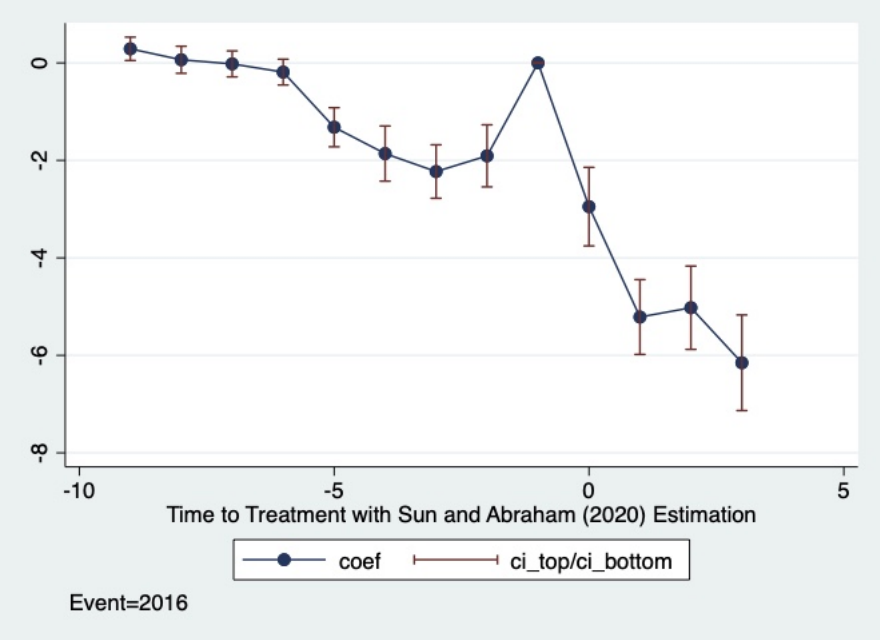
Dependent variable: Hospital Uncompensated Care.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

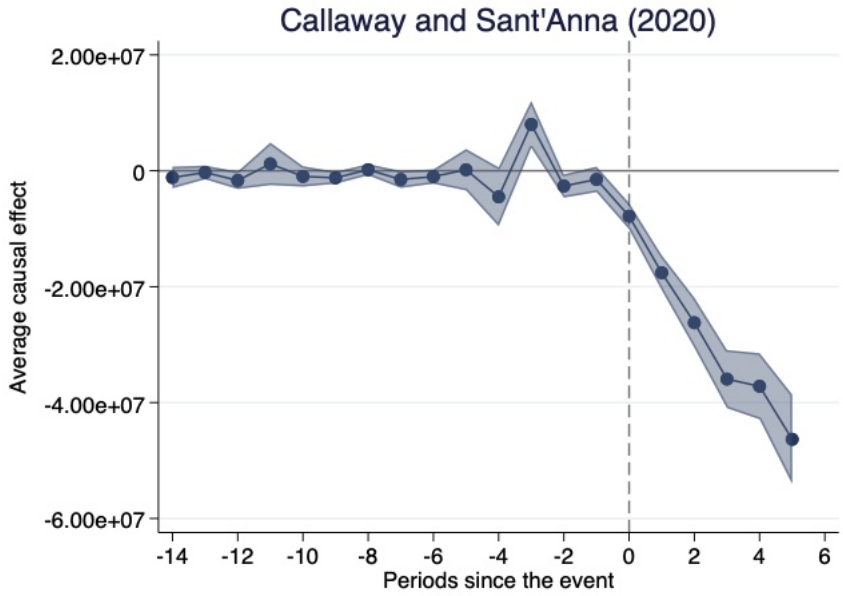
6. Present an event study graph based on the results in part 5. Hint: you can do this automatically in R with the fixest package (using the sunab syntax for interactions), or with eventstudyinteract in Stata. These packages help to avoid mistakes compared to doing the tables/figures manually and also help to get the standard errors correct.







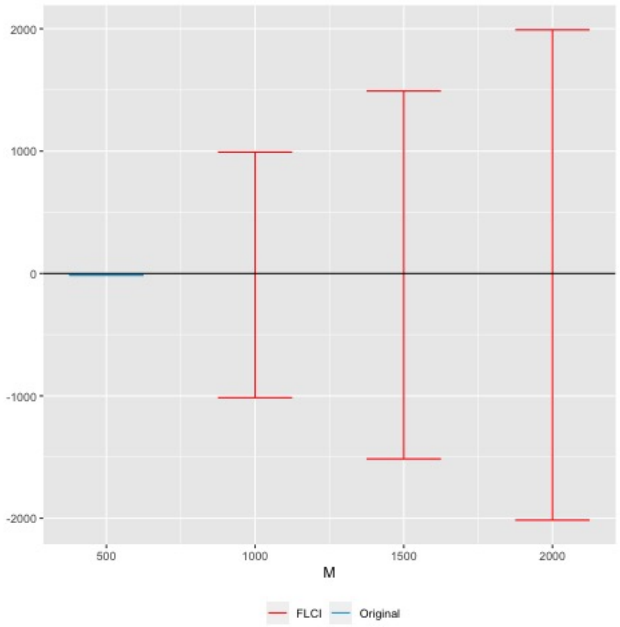
7. Callaway and Sant'Anna (CS) offer a non-parametric solution that effectively calculates a set of group-time specific differences, $ATT(g,t) = E[y_{it}(g)y_{it}(\infty)|G_i = g]$, where g reflects treatment timing and t denotes time. They show that under the standard DD assumptions of parallel trends and no anticipation, $ATT(g,t) = E[y_{it}y_{i,g1}|G_i = g]E[y_{it} - y_{i,g1}|G_i = \infty]$, so that $\hat{ATT}(g,t)$ is directly estimable from sample analogs. CS also propose aggregations of $\hat{ATT}(g,t)$ to form an overall ATT or a time-specific ATT (e.g., $ATTs$ for τ periods before/after treatment). With this framework in mind, provide an alternative event study using the CS estimator. Hint: check out the did package in R or the csdid package in Stata.



8. Rambachan and Roth (RR) show that traditional tests of parallel pre-trends may be underpowered, and they provide an alternative estimator that essentially bounds the treatment effects by the size of an assumed violation in parallel trends. One such bound RR propose is to limit the post-treatment violation of parallel trends to be no worse than some multiple of the pre-treatment violation of parallel trends. Assuming linear trends, such a relative violation is reflected by:

$$\Delta(\bar{M}) = \left\{ \delta : \forall t \geq 0, |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq \tilde{M} \times \max_{i<0} |(\delta_{s+1} - \delta_s) - (\delta_s - \delta_{s-1})| \right\} \tag{4}$$

The authors also propose a similar approach with what they call "smoothness restrictions," in which violations in trends changes no more than M between periods. The only difference is that one restriction is imposed relative to observed trends, and one restriction is imposed using specific values. Using the Yonesto iD package in R or stata, present a sensitivity plot of your CS ATT estimates using smoothness restrictions, with assumed violations of size $M \in \{500, 1000, 1500, 2000\}$. Check out the Github repo here for some help in combining the ronestoiD package with CS estimates. Note that you'll need to edit the function in that repo in order to use pre-specifled smoothness restrictions. You can do that by simply adding Mrec-wvee in the ereateSensitivityResulta function for typewnoothinens .



9. Discuss your findings and compare estimates from different estimators (e.g., are your results sensitive to different specifications or estimators? Are your results sensitive to violation of parallel trends assumptions?).

The results are the same across estimators and our results suggest that Medicaid

expansion is beneficial for hospitals in reducing uncompensated care. Furthermore, the earlier they adopt, the greater the magnitude of savings. Expanding Medicaid can be costly, but it also means that some of the uninsured patients will transition to insured and reduce the burden of uncompensated care for hospitals.

10. Reflect on this assignment. What did you find most challenging? What did you find most surprising?

    Given the wealth of data and characteristics in our merged data set, it would be interesting to look at how Medicaid expansion could effect other hospital finance outcomes. This assignment was much harder than I anticipated. The coding was extremely difficult on stata and i will be transitioning to RA for future assignments. Moreover, I will read each question carefully much earlier.