Assignment2

Jung Jae Kim

Questions

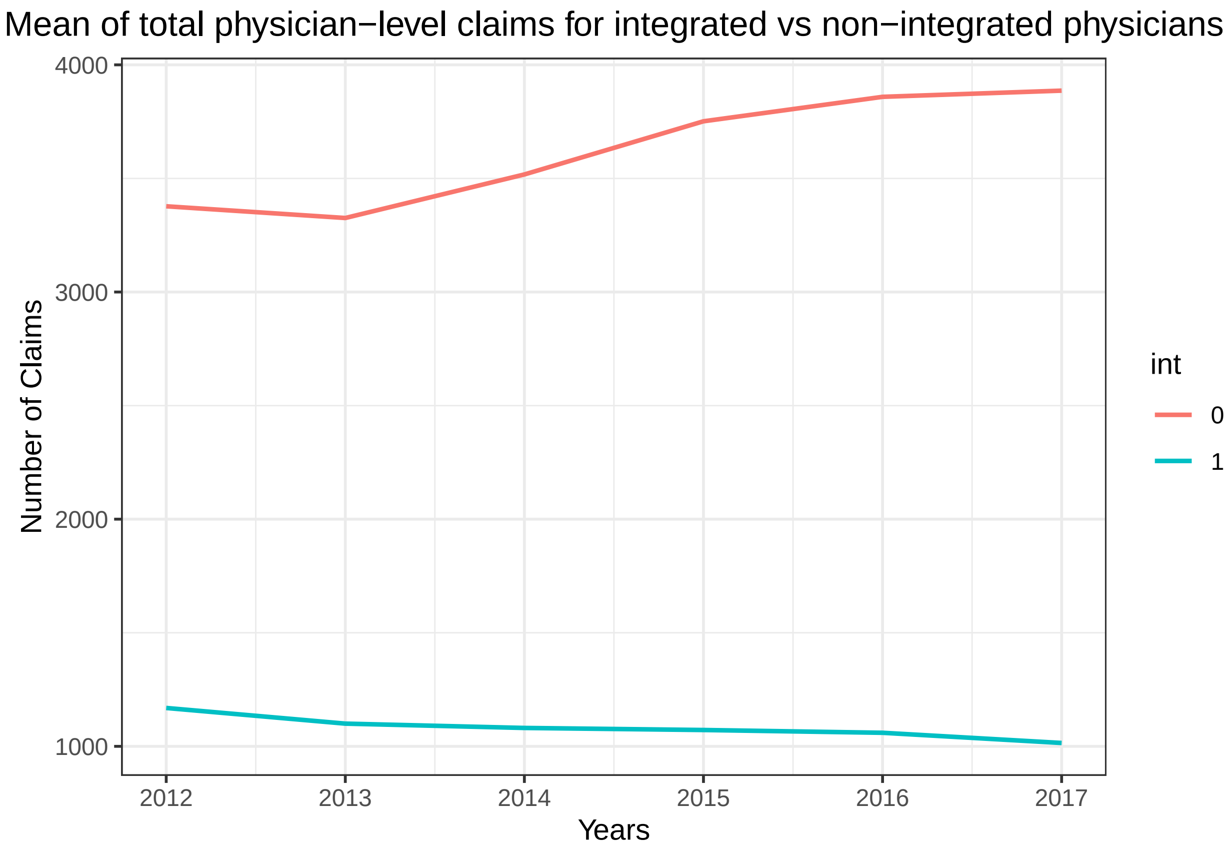
1. Provide and discuss a table of simple summary statistics showing the mean, standard deviation, min, and max of total physician-level Medicare spending, claims, and patients. Use the Medicare utilization and payment data to calculate total spending, claims, and patients at the physician level. You can do this using the average Medicare allowed amt * bene_day_srvc_cnt (or Medicare allowed amt * line_srvc_cnt) for spending, bene_day_srvc_cnt or the line_srvc_cnt for claims, and bene_unique_cnt for patients. The patient counts will include some overlap since the data are by service, but that's OK for our purposes.

Names	Mean	Std.Dev.	Min	Max
Spending	137743.609	280251.913	0.93	26288558
Claims	2699.641	13038.636	4.00	5750425
Patients	1013.114	1906.225	11.00	724713

2.Form a proxy for integration using the ratio:

$$INT_{it} = \mathbf{1}\left(\frac{HOPD_{it}}{HOPD_{it} + OFFICE_{it} + ASC_{it}} \geq 0.75\right), \quad (1)$$

where $HOPD_{it}$ reflects the total number of claims in which physician i bills in a hospital outpatient setting, $OFFICE_{it}$ is the total number of claims billed to an office setting, and ASC_{it} is the total number of claims billed to an ambulatory surgery center. As reflected in Equation (1), you can assume that any physician with at least 75% of claims billed in an outpatient setting is integrated with a hospital. Using this 75% threshold, plot the mean of total physician-level claims for integrated versus non-integrated physicians over time.



3.Estimate the relationship between integration on (log) total physician claims using OLS, with the following specification:

$$y_{it} = \delta INT_{it} + \beta x_{it} + \gamma_i + \gamma_t + \varepsilon_{it}, \tag{2}$$

where INT_{it} is defined in Equation (1), x_{it} captures time-varying physician characteristics, and γ_i and γ_t denote physician and time fixed effects. Please focus on physician's that weren't yet integrated as of 2012, that way we have some pre-integration data for everyone. Impose this restriction for the remaining questions. Feel free to experiment with different covariates in x_{it} or simply omit that term and only include the fixed effects.

	mod.fe
Dependent Var.:	log_claim
int	-0.4015*** (0.0058)
Fixed-Effects:	——————-
npi	Yes
Year	Yes
________________________	______________________________
S.E.: Clustered	by: npi
Observations	2,070,312
R2	0.90040
Within R2	0.01153

4.How much should we be “worried” about endogeneity here? Extending the work of @altonji2005, @oster2019 derives the expression

$$\delta^* \approx \hat{\delta}_{D,x_1} - \rho \times \left[\hat{\delta}_D - \hat{\delta}_{D,x_1}\right] \times \frac{R^2_{max} - R^2_{D,x_1}}{R^2_{D,x_1} - R^2_D} \xrightarrow{p} \delta, \tag{3}$$

where x_1 captures our observable covariates (or fixed effects in our case); δ denotes the treatment effect of interest; $\hat{\delta}_{D,x_1}$ denotes the coefficient on D from a regression of y on D and x_1; R^2_{D,x_1} denotes the R^2 from that regression; $\hat{\delta}_D$ denotes the coefficient on D from a regression of y on D only; R^2_D reflects the R^2 from that regression; R^2_{max} denotes an unobserved “maximum” R^2 from a regression of y on D, observed covariates x_1, and some unobserved covariates x_2; and ρ denotes the degree of selection on observed variables relative to unobserved variables. One approach that Oster suggests is to consider a range of R^2_{max} and ρ to bound the estimated treatment effect, where the bounds are given by $\left[\hat{\delta}_{D,x_1}, \delta^*(R^2_{max}, \rho)\right]$. Construct these bounds based on all combinations of $\rho \in (0, .5, 1, 1.5, 2)$ and $R^2_{max} \in (0.5, 0.6, 0.7, 0.8, 0.9, 1)$ and present your results in a table. What do your results say about the extent to which selection on observables could be problematic here? Hint: you can also look into `psacalc` in `Stata` or `robomit` in `R` for implementation of @oster2019 in `Stata` or `R`, respectively.

	R2max=0.95	R2max=1
rho=0	[-0.401548,-0.401548]	[-0.401548,-0.401548]
rho=0.5	[-0.401548,-0.382146]	[-0.401548,-0.362588]
rho=1	[-0.401548,-0.362743]	[-0.401548,-0.323628]
rho=1.5	[-0.401548,-0.343341]	[-0.401548,-0.284668]
rho=2	[-0.401548,-0.323939]	[-0.401548,-0.245708]

The selection on unobservable could not be problematic here since the bound is not too big.

5. Construct the change in Medicare payments achievable for an integrated versus non-integrated physician practice due to the 2010 update to the physician fee schedule, ΔP_{it}. Use this as an instrument for INT_{it} in a 2SLS estimator following the same specification as in Equation (2). Present your results along with those of your “first stage” and “reduced form”.

	first_stage	second_stage
Dependent Var.:	int	log_claim
practice_rev_change	1.42e-5*** (8.05e-7)	
fitted_int		-4.719*** (0.1513)
Fixed-Effects:	—————	—————
npi	Yes	Yes
year	Yes	Yes
————————	————————	————————
S.E.: Clustered	by: npi	by: npi
Observations	1,730,565	1,730,565
R2	0.81286	0.91111
Within R2	0.00276	0.00460

6.Assess the "need" for IV by implementing a Durbin-Wu-Hausman test with an augmented regression. Do this by first estimating the regression, $INT_{it} = \lambda\Delta P_{it}+\beta x_{it}+\gamma_i+\gamma_t+\varepsilon_{it}$, take the residual $\hat{\nu} = INT_{it}-\hat{INT}_{it}$, and run the regression

$$y_{it} = \delta INT_{it} + \beta x_{it} + \gamma_i + \gamma_t + \kappa\hat{\nu} + \varepsilon_{it}.$$

Discuss your results for $\hat{\kappa}$.

	durbin_wu_hausman
Dependent Var.:	log_claim
int	-4.719*** (0.1494)
resid	4.343*** (0.1496)
Fixed-Effects:	————————
npi	Yes
year	Yes
————————————	————————————————
S.E.: Clustered	by: npi
Observations	1,730,565
R2	0.91205
Within R2	0.01511

Since the coefficien of residuals is significantly different from 0, we need IV.

7.Now let's pay attention to potential issues of weak instruments. As we discussed in class, one issue with weak instruments is that our typical critical values (say, 1.96 for a 95% confidence interval) from the equation of interest (sometimes called the structural equation) are too low in the presence of a weak first-stage. These issues are presented very clearly and more formally in the Andrews, Stock, and Sun (2019) survey article. For this question, you will consider two forms of inference in the presence of weak instruments. Hint: Check out the `ivmodel` package in R or the `ivreg2` command in Stata for help getting the AR Wald statistic. - Present the results of a test of the null, $H_0 : \delta = 0$, using the Anderson-Rubin Wald statistic. Do your conclusions from this test differ from a traditional t-test following 2SLS estimation of Equation (2)? - Going back to your 2SLS results…inflate your 2SLS standard errors to form the tF adjusted standard error, following Table 3 in @lee2021. Repeat the test of the null, $H_0 : \delta = 0$, using standard critical values and the tF adjusted standard error.

```
[1] "Anderson-Rubin confidence interval is  [-4.8831,-4.5622]"
```

```
[1] "AtF adjusted confidence interval is  [-4.8369,-4.6014]"
```

Since the F-statistics of the first stage regression is sufficiently large, the results are similar.

8.Following the Borusyak and Hull (2021) working paper (BH), we can consider our instrument as a function of some exogenous policy shocks and some possibly endogenous physician characteristics, $\Delta P_{it} = f\left(g_{pt}; z_{ipt}\right)$, where g_{pt} captures overall payment shocks for procedure p at time t, and z_{ip} denotes a physician's quantity of different procedures at baseline. We can implement the BH re-centering approach as follows: - Consider hypothetical price changes over a set of possible counterfactuals by assuming that the counterfactuals consist of different allocations of the observed relative price changes. For example, take the vector of all relative price changes, reallocate this vector randomly, and assign new hypothetical relative price changes. Do this 100 times. This isn't "all" possible counterfactuals by any means, but it will be fine for our purposes. - Construct the expected revenue change over all possible realizations from previously, $\mu_{it} = E[\Delta P_{it}] = \sum_{s=1}^{100} \sum_p g_{pt}^s z_{ip}$. - Re-estimate Equation (2) by 2SLS when instrumenting for INT_{it} with $\tilde{\Delta} P_{it} = \Delta P_{it} - \mu_{it}$. Intuitively, this re-centering should isolate variation in the instrument that is only due to the policy and remove variation in our instrument that is due to physician practice styles (the latter of which is not a great instrument).

	iv_fixed_bh
Dependent Var.:	log_claim
int	-0.7152*** (0.1563)
Fixed-Effects:	————————-
npi	Yes
year	Yes
______________	_______________________
S.E.: Clustered	by: npi
Observations	1,730,565
R2	0.91099
Within R2	0.00322

9. Discuss your findings and compare estimates from different estimators.

Answers: Although the effect of integration on claims is consistently negative, there is a huge difference in terms of magnitude of estimates from FE and IV regression results. Moreover, BH result indicates that we need to use instruments with only exogenous policy shocks.

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

Answers: It was very helpful to conduct several assumption check for IV model. It would be very interesting if I had a weak instrument, which is not the case in this exercise, and did same exercise.