Econ 771 – Module 2 Empirical Exercise

October 14

Name: Amy Lim

1 Overview

The purpose of this assignment is to learn about instrumental variables and issues that arise in applied microeconomics. Working with three data sources, we will use instrumental variables to estimate how changes in the CMS physician fee schedule and physician integration effects on claims, which is a bill that healthcare providers submit to a patient's insurance provider. For this assignment, I am still using Stata, but I will fully commit to R starting assignment 3.

2 Data

The analysis depends on three data sources: The Medicare Data on Provider Practice and Specialty (MD-PPAS), Medicare Utilization and Payment Data, and Physician Fee Schedule 2010 update (PFS) . Data are available online, and for this assignment, I downloaded all of the data from our class OneDrive folder.

2.1 The Medicare Data on Provider Practice and Specialty (MD-PPAS)

MD-PPAS is a data source that assigns Medicare providers to medical practices and elaborates on the CMS provider specialty classification system. The MD-PPAS Data will provide us a way to measure physician integration by using NPI (physician level) and link to the other two sources to get data on claims and spending.

2.2 Medicare Utilization and Payment Data (PUF)

The PUF dataset provides information on use, payments, submitted charges and beneficiary demographic and health characteristics organized by NPI. Although this data is quite rich, we will only use the data on claims and spending. Combined with MD-PPAS, we can contstruct our intial instruments.

2.3 Physician Fee Schedule 2010 update (PFS)

Thanks to Ian for sharing this data and saving us the pain of constructing the data ourselves. The raw data comes from CMS and this data gives us the ability to measure prices changes in procedures (for claims and spending) so we can later use as an instrumental variable.

3 Data Cleaning

Using the described data, we can merge them to answer the assigned questions. After downloading the data from OneDrive, I started to restrict the data to the specifications of the assignment. The data sets are massive compared to our previous assignment, however, there are not as many errors and therefore cleaner to use. I started with keeping data from 2012-2017 for MD-PPAS and PUF, then I merged the two by NPI. The greatest challenge was working with PUF because it was incredibly large for the purposes of our assignment. In order to make it smaller, I collapsed the data with the variables that we needed for the analysis by summing across NPI and hospital code that we will need to merge from MD-PPAS (NPI) and PFS (hcpcs). After successfully merging MD-PPAS and PUF, I manipulate the PFS data as described in question 5 to have the fees for 2012-2017.

4 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. The patient counts will include some overlap since the data are by service, but that's OK for our purposes.

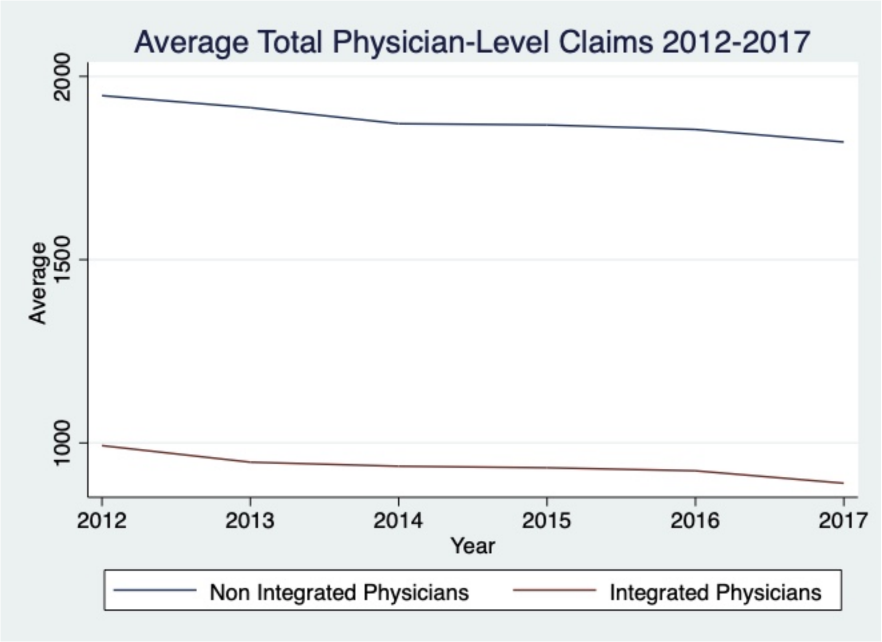
Table 1: Summary Statistics for Medicare Spending, Claims, and Patients at the Physician Level

Year	Mean	SD	Min	Max	N
2012					
Spending	4,473,796.57	26,513,732.93	3.67	13,996,507,136.00	479,832.00
Claims	1,639.30	5,219.24	11.00	2,931,786.00	479,832.00
Patients	1,043.92	2,084.56	11.00	724,713.00	479,832.00
2013					
Spending	4,348,209.32	30,655,793.67	5.36	17,884,876,800.00	486,256.00
Claims	1,592.06	3,014.03	11.00	1,059,386.00	486,256.00
Patients	1,023.22	1,870.65	11.00	721,303.00	486,256.00
2014					
Spending	4,212,294.70	23,934,934.17	2.54	12,262,518,784.00	490,655.00
Claims	1,547.42	2,809.02	11.00	756,682.00	490,655.00
Patients	1,003.08	1,751.14	11.00	545,218.00	490,655.00
2015					
Spending	4,205,056.73	27,159,519.35	0.93	15,415,290,880.00	494,255.00
Claims	1,535.78	2,939.39	11.00	868,554.00	494,255.00
Patients	1,008.46	1,861.28	11.00	605,766.00	494,255.00
2016					
Spending	4,227,895.08	28,327,483.19	1.72	16,213,949,440.00	498,319.00
Claims	1,517.30	2,944.99	11.00	954,028.00	498,319.00
Patients	1,003.24	1,878.09	11.00	635,666.00	498,319.00
2017					
Spending	4,163,748.07	29,995,929.54	1.47	17,472,516,096.00	500,010.00
Claims	1,473.82	3,018.41	11.00	989,010.00	500,010.00
Patients	978.94	1,926.53	11.00	656,227.00	500,010.00
Total					
Spending	4,270,439.84	27,863,260.28	0.93	17,884,876,800.00	2,949,327.00
Claims	1,550.21	3,420.95	11.00	2,931,786.00	2,949,327.00
Patients	1,009.88	1,897.42	11.00	724,713.00	2,949,327.00

2. Form a proxy for integration using the ratio:

$$INT_{it} = \mathbb{1}\left(\frac{HOPD_{it}}{HOPD_{it} + OFFICE_{it} + ASC_{it}} \geq 0.75\right) \tag{1}$$

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 total physician claims using OLS, with the following specification:

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

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.

Table 2: Physician Integration on Logged Total Physician Claims

	(1)
Integration	-0.179***
	(0.00392)
Average Submitted Charge Amount	0.00000793***
	(0.000000706)
Average Medicare Allowed Amount	0.000181***
	(0.00000534)
Constant	6.265***
	(0.00643)
Observations	2,749,830
R^2	0.902

Standard errors in parentheses

Dependent variable: Log Total Physician Claims

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

4. How much should we be "worried" about endogeneity here? Extending the work of Altonji, Elder, and Taber (2005), Oster (2019) 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; δ denotes the treatment effect of interest;$\hat{\delta_{D,x_i}}$ 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

$$\delta \times \frac{\text{Cov}\left(W_1, X\right)}{\text{Var}\left(W_1\right)} = \frac{\text{Cov}\left(W_2, X\right)}{\text{Var}\left(W_2\right)} \tag{4}$$

Construct the value in Equation (3) 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 Oster (2019) in Stata or R, respectively.

Table 3: Oster Bounds

R^2_{max}	$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 1.5$	$\rho = 2$
0.5	-0.179	0.6132811	1.226562	1.839843	2.453124
0.6	-0.179	0.4607262	0.9214523	1.382179	1.842905
0.7	-0.179	0.3081712	0.6163425	0.9245137	1.232685
0.8	-0.179	0.1556163	0.3112326	0.4668489	0.6224653
0.9	-0.179	0.0030614	0.0061228	0.0091842	0.0122456
1	-0.179	-0.1494935	-0.2989871	-0.4484806	-0.5979741

5. Construct the total change in Medicare payments achievable for an integrated versus non-integrated physician 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".

Table 4: IV Regression Results

	(1) First-Stage	(2) Reduced Form	(3) IV
Integration	-0.0958***		2.910***
	(0.00389)		(0.0479)
Average Submitted Charge Amount	0.00000351***	0.00000300***	-0.0000608***
	(0.000000716)	(0.000000729)	(0.00000123)
Average Medicare Allowed Amount	0.000204***	0.000207***	0.000691***
	(0.00000597)	(0.00000609)	(0.00000783)
Physician Fee Schedule		0.0000357***	
		(0.00000117)	
Constant	6.122***	6.092***	4.792***
	(0.00664)	(0.00629)	(0.0181)
Observations	2,682,845	2,677,869	2,724,029
R^2	0.902	0.903	.

Standard errors in parentheses

Dependent variable: Log Total Physician Claims

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

6. Assess the "need" for IV by implementing a Durbin-Wu-Hausman test with an

augmented regression. Do this by first estimating the regression, $y_{it} = \delta INT_{it} + \beta x_{it} + \gamma_i + \gamma_t + \epsilon_{it}$, take the residual $\nu = INT_{it} - I\hat{N}T_{it}$, and run the regression

$$y_{it} = \delta INT_{it} + \beta x_{it} + \gamma_i + \gamma_t + \kappa nu + \hat{} \epsilon_{it}(5)$$

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

Table 5: Wu-Hausman Test Results

	(1)
Integration	-0.217***
	(0.00163)
Average Submitted Charge Amount	-0.0000383***
	(0.000000314)
Average Medicare Allowed Amount	0.000534***
	(0.00000203)
Linear prediction	1.135***
	(0.0230)
Constant	5.524***
	(0.00858)
Observations	2,872,624
R^2	0.221

Standard errors in parentheses

Dependent variable: Log Total Physician Claims

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

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:

   - 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 Lee et al. (2021). Repeat

the test of the null, $H_0 : \delta = 0$, using standard critical values and the tF adjusted standard error.

Table 6: **regression table**

	(1)
Integration	2.910***
	(0.0479)
Average Submitted Charge Amount	-0.0000608***
	(0.00000123)
Average Medicare Allowed Amount	0.000691***
	(0.00000783)
Constant	4.792***
	(0.0181)
test	6288.304

Standard errors in parentheses

Dependent variable: Log Total Physician Claims

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

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_{ipt} 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, $\mu = E[\Delta P_{it}] = \sum_{s=1}^{100} \sum_p g_{pt}^s z_{ip}$
   - Re-estimate Equation (2) by 2SLS when instrumenting for INT_{it} with $\Delta P_{it} = \tilde{\Delta} P_{it} - \mu_{it}$. 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).

I tried really hard to do this. My loop is still running... sorry. I'm using Stata, and I used a loop to run shufflevar to generate the the simulation. see the code for my sad attempt. I plan on working with Rachel to see how she ran it, but it feels wrong to copy her code and say I did it.

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

   The results indicate that that integration has a negative effect on claims. I think this means that increasing physician integration decreases the number claims, which means that spending on care would go down. This could indicate that the physician integration could yield cost savings. I will be honest, I am still having trouble understanding all of the interpretations.

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

    Again, I have under estimated the amount of time it would take me to merge the data. In this assignment, I learned how to compromise with the data and be more creative with collapsing them before merging and merging within loops. I had some trouble visualizing the loop for number 4, but I love how these assignments keep pushing me to write more efficient code as well. It was also a much harder to make the IV than I thought. I don't know how many I did correctly, but I found myself hand writing some of these equations on paper to better understand how the instrumental variable worked.