# Assignment 2

## Wonjun Choi

## October 14, 2022

1. Table 1 presents the summary statistics of total spending, the number of claims, and the number of patients for each physicians over the period over 2012 to 2017.

|  | Mean | S.D. | Min | Max |
|---|---|---|---|---|
| spending | 675748.86 | 1456660.04 | 5.94 | 79587207.68 |
| claim | 13244.02 | 61967.26 | 11.00 | 10453052.00 |
| patient | 4970.18 | 9080.18 | 11.00 | 1991234.00 |

Table 1: Summary statistics

2. The upper line is the trend of unintegrated physicians, and the bottom line is of integrated physicians. The total number of claims of unintegrated physicians begins to increase after the price shock is fully implemented, while it remains in the similar level for integrated physicians. The figure is rescaled to be put in the right place.
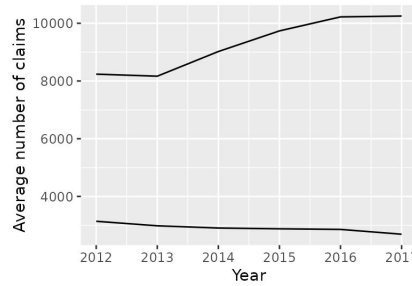


Figure 1: Trend of total number of claims by integration status

3. Table 2 contains the OLS result of regressing the total number of claims on integration status. I added physician level and year level fixed effects in the regression. The standard error was clustered in physician level. The result suggests that the integrated physicians makes 0.4% less claims than the unintegrated physicians.

|  | claims |
| --- | --- |
|  | (1) |
| INT | -0.4015*** |
|  | (0.0058) |
|  |  |
| Observations | 2,070,312 |
| $R^2$ | 0.90040 |
| Within $R^2$ | 0.01153 |
|  |  |
| npi fixed effects | ✓ |
| Year fixed effects | ✓ |

Table 2: OLS

4. If the treatment variable $INT$ is correlated with unobserved variables, the coefficient can be biased. We assume hypothetical correlation and see how the coefficient can change in several scenarios following Altonji et al. (2005) and Oster (2019). We consider the following construction.

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

where $x_1$ is a control variables (fixed effects) and $\hat{\delta}_X$ and $R^2_X$ indicates the treatment effect and $R^2$ in the regression when a vector $X$ is used as a regressor. $R^2_{max}$ is a hypothetical goodness-of-fit $R^2$ in a scenario where we were able to observe unobserved variables. $R^2_{max} = 1$ indicates the situation where we can observe every variables that determines the outcome. $\rho$ is the hypothetical correlation between the treatment variable and the unobservables. I think I included the rownames in R code, but it somehow disappeared.

Table 3 displays the hypothetical 95% confidence intervals in each scenarios with $\rho \in \{-1, -0.5, 0, 0.5, 1\}$ and $R^2_{max} \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. Since the

| $R^2_{max} = 0.5$ | $R^2_{max} = 0.6$ | $R^2_{max} = 0.7$ | $R^2_{max} = 0.8$ | $R^2_{max} = 0.9$ | $R^2_{max} = 1$ |
|---|---|---|---|---|---|
| NA | NA | NA | NA | NA | [-0.4,-0.4] |
| NA | NA | NA | NA | NA | [-0.4,-0.36] |
| NA | NA | NA | NA | NA | [-0.4,-0.32] |
| NA | NA | NA | NA | NA | [-0.4,-0.28] |
| NA | NA | NA | NA | NA | [-0.4,-0.25] |

Table 3: Altonji, Elder, and Taber (2005)

original $R^2_{D,x_1}$ was already 0.91, the results of $R^2_{max} \leq 0.9$ is omitted. The treatment effect was significantly negative in every scenarios although the magnitude varied.

5. The treatment variable $INT$ can be endogenous if the number of claims is under consideration in the decision of whether to merge. To estimate the true treatment effect, I use an exogenous price change as an instrument of the treatment.

| | INT | log_claims | |
|---|---|---|---|
| | (1) | (2) | (3) |
| PriceChange | $1.42 \times 10^{-5}$*** | $-6.72 \times 10^{-5}$*** | |
| | $(8.05 \times 10^{-7})$ | $(2.15 \times 10^{-6})$ | |
| INThat | | | -4.719*** |
| | | | (0.1513) |
| | | | |
| Observations | 1,730,565 | 1,730,565 | 1,730,565 |
| $R^2$ | 0.81286 | 0.91111 | 0.91111 |
| Within $R^2$ | 0.00276 | 0.00460 | 0.00460 |
| | | | |
| npi fixed effects | ✓ | ✓ | ✓ |
| Year fixed effects | ✓ | ✓ | ✓ |

Table 4: 2SLS

Table 4 presents the 2SLS results. The treatment effect coefficient was -4.72.

6. $INT$ can reflect some effect from the unobserved variables. We isolate the treatment effect from the other effects by implementing Durbin-Wu-Hausman

test with an augmented regression

$$y_{it} = \delta INT_{it} + \gamma_i + \gamma_t + \kappa\hat{\nu} + \epsilon_{it}$$

where $\hat{\nu} = INT_{it} - I\hat{N}T_{it}$ is a orthogonalized treatment variable that captures the effect of $INT$ other than the treatment effect.

|  | log_claims |
| --- | --- |
|  | (1) |
| INT | -4.719*** |
|  | (0.1494) |
| INTres | 4.343*** |
|  | (0.1496) |
|  |  |
| Observations | 1,730,565 |
| R$^2$ | 0.91205 |
| Within R$^2$ | 0.01511 |
|  |  |
| npi fixed effects | ✓ |
| Year fixed effects | ✓ |

Table 5: DWH test

Table 5 displays the result of DWH test. From the coefficient of $INTres$, we can see the treatment variable indeed captures some other effects if it is not instrumented. The result implies that integration actually decreases the number of claims, but there are some systematic difference in integrated and unintegrated physicians (I am not sure).

7. If the instrument variable does not sufficiently explains the treatment variable, the IV results might be unreliable. We suggest 2 confidence intervals that takes the weak instruemnt problem into account.

|  | lower | upper |
| --- | --- | --- |
| Anderson-Rubin | -4.88 | -4.56 |
| Lee | -4.84 | -4.60 |

Table 6: Confidence intervals

4

Table 6 shows the confidence intervals. They were not that different from the previous 2SLS estimator. Actually, the first stage F statistic was over 4000, so in our exercise the instrument has a sufficient explanation power.

8. Our instrument variable is a function of an exogenous price shock and physician characteristics. However, physician characteristics can be possibly endogenous. To eliminate the possible endogeneity, Borusyak and Hull (2021) working paper suggested recentering the instrument variable with simulated pseudo instruments. See Question 8 of HW2 for detail.

|  | log_claims |
| --- | --- |
|  | (1) |
| INThat | -0.7152*** |
|  | (0.1601) |
|  |  |
| Observations | 1,730,565 |
| $R^2$ | 0.91071 |
| Within $R^2$ | $8.09 \times 10^{-5}$ |
|  |  |
| npi fixed effects | ✓ |
| Year fixed effects | ✓ |

Table 7: Borusyak and Hull (2021)

Table 7 shows the estimation results with the recentered instrument. The treatment effect was significantly smaller than the original 2SLS estimates.

9. In this exercise, we did only one main estimation (2SLS). All of the other things were auxiliary analysis. The results consistently indicates integration and the number of claims has a negative relationship, but the magnitude varies especially in Borusyak and Hull's approach suggesting that the instrument might be able to polished more.

10. I think I could have learn more about practical issues of IV, which was so fun, if I started the assignment earlier. I thought I would be able to start earlier, but I was ignorant of the fact that the semester usually gets busier as time goes by. Also, the data work is quite overwhelming. I can't imagine doing this without the guidance code. But I think I learned a lot, anyway. Researchers are great people.