

Tarea 3: ENIGH Salud Alimentaria

Inferencia Causal

Mateo De La Roche Macías

Oregon tiene actualmente una lista de espera de 90,000 personas de bajos ingresos que buscan acceder a Medicaid, pero sólo dispone de 30,000 cupos disponibles. Adicionalmente, el estado carece de datos concretos para verificar si el acceso a Medicaid realmente tiene un impacto positivo en la salud autoreportada de los beneficiarios.

Vamos a realizar un análisis sobre los efectos del acceso al seguro de salud en la salud autoreportada. Los datos que se utilizan en el análisis fueron obtenidos mediante una lotería de selección aleatoria para asignar los cupos disponibles de Medicaid.

El estado de Oregon llevó a cabo la lotería, resultando en la selección aleatoria de 29,834 individuos de una lista inicial depurada de casi 75,000 solicitantes (los otros 15,000 solicitantes no eran elegibles para el programa).

Para evaluar el impacto del programa, se organizó una serie de encuestas en tres momentos clave: antes de la implementación del programa, a los seis meses (submuestra de 5000 individuos) y a los doce meses

needmet med X_m , ,health gen bin X_m , baddays phys X_m y baddays ment X_m

Balances

a) Verificación de aleatoriedad al asignar los grupos de tratamiento y control en datos basales.

Tabla de balance para para datos antes de la implementación.

En los datos tenemos como variables basales, año de nacimiento, género, ingresos del hogar, raza, educación.

```
# A tibble: 16 x 4
  variables1      Media_control1 Media_trat1 p_value1
  <chr>          <dbl>          <dbl>    <dbl>
```

1	needmet_med_0m	0.510	0.563	1.37e-13
2	hhinc_cat_0m	6.32	6.69	2.72e- 8
3	baddays_ment_0m	11.4	10.6	1.24e- 6
4	baddays_phys_0m	9.48	8.92	3.61e- 4
5	health_gen_bin_0m	0.393	0.369	5.89e- 4
6	birthyear_list	1966.	1966.	1.98e- 3
7	female_list	0.594	0.574	4.20e- 3
8	race_white_0m	0.843	0.833	5.47e- 2
9	race_asian_0m	0.0314	0.0362	6.57e- 2
10	race_hisp_0m	0.0939	0.101	7.74e- 2
11	edu_0m	2.28	2.25	7.86e- 2
12	race_amerindian_0m	0.0602	0.0546	9.06e- 2
13	person_id	37411.	37240.	5.79e- 1
14	race_other_qn_0m	0.0734	0.0752	6.33e- 1
15	race_pacific_0m	0.00868	0.00917	7.12e- 1
16	race_black_0m	0.0337	0.0335	9.29e- 1

La tabla anterior plantea como hipótesis nula que no hay diferencia entre los grupos de tratamiento y control. Aunque la mayoría de las variables no parecen ser diferentes, el p-value en las variables `birthyear_list`, `female_list` y `hhinc_cat_0m` son menores a 0.05 lo cual nos permite concluir con un nivel de significancia del 0.05 que hubo un sesgo en la asignación de tratamiento. Parece ser que el año de nacimiento se utilizó como variable para asignar el tratamiento y que hubo un sesgo significativo otorgándole mayor tratamiento médico a las mujeres. Lo más preocupante es que el promedio de ingresos entre ambos grupos es significativamente muy diferente por lo que podría haber habido alguna especie de sesgo a la población con ingresos menores.

b) Verificación de representatividad en submuestra y después de 12 meses

Tabla de balance para para datos después de 6 meses.

```
# A tibble: 17 x 4
```

variables1	Media_control1	Media_trat1	p_value1
<chr>	<dbl>	<dbl>	<dbl>
1 ins_any_6m	0.285	0.533	7.81e-63
2 needmet_med_6m	0.567	0.627	4.57e- 5
3 baddays_ment_6m	11.9	10.5	6.00e- 5
4 baddays_phys_6m	10.2	8.94	7.82e- 5
5 health_gen_bin_6m	0.465	0.407	9.83e- 5
6 race_amerindian_6m	0.0555	0.0450	1.07e- 1
7 race_pacific_6m	0.00583	0.0103	1.12e- 1

8	race_asian_6m	0.0418	0.0519	1.20e- 1
9	hhinc_cat_6m	7.06	7.27	1.55e- 1
10	birthyear_list	1965.	1966.	2.68e- 1
11	race_hisp_6m	0.0946	0.0877	4.30e- 1
12	person_id	37320.	36944.	5.63e- 1
13	race_other_qn_6m	0.0737	0.0701	6.44e- 1
14	female_list	0.575	0.580	7.20e- 1
15	race_black_6m	0.0260	0.0245	7.44e- 1
16	race_white_6m	0.849	0.847	8.43e- 1
17	edu_6m	2.31	2.31	9.92e- 1

Submuestra de 6 Meses: Hay diferencias significativas en varias variables, incluidas baddays_ment_6m, baddays_phys_6m, health_gen_bin_6m, ins_any_6m y needmet_med_6m. Esto sugiere que la submuestra de 6 meses puede no ser representativa del balance inicial, y podría haber sesgo en la asignación de tratamiento en esta submuestra. En especial se nota una gran diferencia en la variable *ins_any_6m*, la cual indica si tenía algún tipo de seguro, se puede ver como muchas personas del tratamiento si tenían seguro en comparación con las del grupo control.

Tabla de balance para para datos después de 12 meses.

```
# A tibble: 17 x 4
```

	variables1	Media_control1	Media_trat1	p_value1
	<chr>	<dbl>	<dbl>	<dbl>
1	ins_any_12m	0.320	0.505	2.68e-142
2	needmet_med_12m	0.637	0.715	2.51e- 29
3	health_gen_bin_12m	0.439	0.389	8.53e- 12
4	baddays_ment_12m	11.3	10.4	1.87e- 7
5	baddays_phys_12m	9.62	8.91	1.02e- 5
6	hhinc_cat_12m	6.75	7.03	9.08e- 5
7	race_black_12m	0.0358	0.0297	2.29e- 2
8	race_amerindian_12m	0.0636	0.0573	7.88e- 2
9	race_hisp_12m	0.0973	0.104	1.16e- 1
10	female_list	0.593	0.583	1.43e- 1
11	race_white_12m	0.847	0.840	1.78e- 1
12	race_asian_12m	0.0312	0.0344	2.27e- 1
13	edu_12m	2.33	2.32	4.33e- 1
14	race_other_qn_12m	0.0814	0.0846	4.41e- 1
15	person_id	37432.	37231.	5.33e- 1
16	race_pacific_12m	0.00841	0.00924	5.54e- 1
17	birthyear_list	1966.	1966.	8.26e- 1

Aquí también se puede observar un sesgo, sin embargo, se observa un sesgo parecido al que se observa en la submuestra tomada a los 6 meses, esto quiere decir que lo más probable es que los encuestadores estén introduciendo alguna especie de sesgo al momento de encuestar.

c) Efecto en representatividad

El uso de la muestra de seis meses puede afectar la representatividad de los resultados debido a los sesgos observados. Sin embargo, dado que estos sesgos también están presentes en la muestra de doce meses, esto sugiere que el diseño del estudio y los métodos de recolección de datos están introduciendo estos sesgos de manera sistemática. Es esencial ajustar los análisis subsecuentes para estas variables sesgadas para obtener estimaciones precisas del impacto del tratamiento.

Efecto de tratamiento simple

a) Definición de efectos

- **Efecto Promedio del Tratamiento (ATE):** El ATE es la diferencia promedio en los resultados entre aquellos que recibieron el tratamiento (acceso a Medicaid) y aquellos que no lo recibieron, en la población general. En este contexto, el ATE mide el impacto promedio de tener acceso a Medicaid en la salud general autoreportada de los beneficiarios, independientemente de si realmente usaron el servicio o no.
- **Efecto Promedio en los Tratados (TOT):** El TOT es la diferencia promedio en los resultados entre aquellos que recibieron el tratamiento (acceso a Medicaid) y aquellos que no lo recibieron, pero solo entre los individuos que realmente utilizaron el tratamiento. En este contexto, el TOT mide el impacto promedio de tener acceso a Medicaid en la salud general autoreportada de los beneficiarios que realmente utilizaron los servicios de Medicaid.

En este caso el *policy maker* se debería de fijar en el ATE ya que no solo le importan los beneficios propios del programa, si no también el beneficio que esté tiene sobre la población total (beneficio más externalidades). El TOT le interesaría más a los médicos o a Medicaid para tener el efecto de su programa.

b) Evaluación del efecto de tratamiento utilizando el estimador de Neyman

Prueba de Hipótesis

- H_0 : El tratamiento (ganar la lotería) no tiene efecto sobre los beneficiarios ($ATE = 0$)
- H_1 : El tratamiento (ganar la lotería) tiene un efecto en la salud autoreportada de los beneficiarios $ATE \neq 0$

Definición de ATE :

$$ATE = \mathbb{E}[Y_i^T] - \mathbb{E}[Y_i^C]$$

Donde:

- Y_i^T : es el resultado potencial para el individuo i si recibe el tratamiento (pertenece a T)
- Y_i^C : es el resultado potencial para el individuo i si no recibe el tratamiento (pertenece a C)

\$ATE

[1] -0.05026374

\$SE

[1] 0.007354982

\$t_stat

[1] -6.833972

\$p_value

[1] 8.525632e-12

Efecto Promedio del Tratamiento (ATE):

- El ATE estimado es -0.0503. Esto significa que, en promedio, el grupo de tratamiento (los que ganaron la lotería de Medicaid) tiene una probabilidad 5.03% menor de reportar buena salud general (health_gen_bin_12m) en comparación con el grupo de control (los que no ganaron la lotería de Medicaid). Este resultado sugiere que el tratamiento tuvo un efecto negativo en la salud general autoreportada de los beneficiarios.

Nota: El signo negativo es por como está construida la definición del ATE, no debería interpretarse como un efecto negativo.

Error Estándar (SE):

- El error estándar del ATE es 0.00735. Por lo tanto, la estimación es precisa.

Significancia:

- El P-value es de 8.53e-12, por lo cual podemos aceptar con un nivel de significancia de 0.001 (o hasta menor) que el efecto del tratamiento (ganar la lotería de Medicaid) sobre la salud general autoreportada doce meses después es estadísticamente significativo.

c) Estimación OLS simple (sin controles) errores heterocedásticos

Ecuación estimada:

```
health_gen_bin_12m = 0.4390846 + -0.05026374 * treatment
```

Resultados:

t test of coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4390846  0.0052567  83.529 < 2.2e-16 ***
treatment    -0.0502637  0.0073550  -6.834 8.526e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El coeficiente del tratamiento representa el cambio en la probabilidad de reportar buena salud general asociado con ganar la lotería de Medicaid. La estimación es -0.0503, lo que indica que, en promedio, ganar la lotería de Medicaid reduce la probabilidad de reportar buena salud general en un 5.03%. Este resultado es estadísticamente significativo al nivel del 0.001. Esto concuerda con lo obtenido en la prueba de Neyman.

d) Estimación OLS simple (con controles) errores heterocedásticos

t test of coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.02363549  0.56726618  26.4843 < 2.2e-16 ***
treatment    -0.04513575  0.00711337  -6.3452 2.275e-10 ***
birthyear_list -0.00735573  0.00028872 -25.4769 < 2.2e-16 ***
female_list   -0.00989841  0.00723057  -1.3690 0.171
hhinc_cat_12m -0.01731397  0.00069785 -24.8105 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los controles se eligieron basándonos en las variables que tenían sesgos al hacer las tablas de balance. Al utilizarlas como control podemos reducir el efecto que tienen y obtener un mejor *ATE*.

Al controlar por las variables sesgadas, podemos ver que el coeficiente del tratamiento disminuye, esto significa que que ganar la lotería de Medicaid reduce la probabilidad de reportar buena salud general en un 4.51% (significativo con un significancia muy baja menor 0.001)

Para los coeficientes de las variables control:

- Cada año adicional de edad está asociado con una reducción de 0.74% en la probabilidad de reportar buena salud general.
- Ser mujer no tiene un efecto significativo en la probabilidad de reportar buena salud general. (nota al profesor, si fuera un reporte quitaría las variables no significativas y volvería a correr la regresión solo con controles significativos pero para no llenar de regresiones el inciso lo dejaré con los iniciales)
- Un aumento en la categoría de ingresos del hogar está asociado con una reducción del 1.73% en la probabilidad de reportar buena salud general.

e) Agregando el efecto de seis meses después como control

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	15.15168558	1.13264104	13.3773	< 2.2e-16 ***
treatment	-0.05109877	0.01440705	-3.5468	0.0003938 ***
birthyear_list	-0.00739741	0.00057695	-12.8216	< 2.2e-16 ***
female_list	-0.02614606	0.01415146	-1.8476	0.0647253 .
hhinc_cat_6m	-0.01878106	0.00133046	-14.1162	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Es curioso observar que en el análisis a seis meses se ve un impacto marginalmente significativo en ser mujer, aunque no es altamente significativo (p-valor = 0.0647). Esto confirma que la forma en la que se realizó la submuestra podría no haber sido correcta. Sin embargo, el efecto del tratamiento a los seis meses (-5.11%) es consistente y no es muy diferente del efecto observado a los doce meses (-4.51%), lo que sugiere una robustez en la influencia del tratamiento sobre la salud general autoreportada.

f) Reporte de resultados

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, May 19, 2024 - 2:16:52 AM

Table 1

	<i>Dependent variable:</i>		
	Health General (12m)		Health General (6m)
	(1)	(2)	(3)
Treatment	−0.050*** (0.007)	−0.045*** (0.007)	−0.051*** (0.014)
Year of Birth		−0.007*** (0.0003)	−0.007*** (0.001)
Female		−0.010 (0.007)	−0.026* (0.014)
Household Income		−0.017*** (0.001)	
hhinc_cat_6m			−0.019*** (0.001)
Constant	0.439*** (0.005)	15.024*** (0.574)	15.152*** (1.140)
Observations	17,895	17,895	4,673
R ²	0.003	0.069	0.078
Adjusted R ²	0.003	0.069	0.078
Residual Std. Error	0.492 (df = 17893)	0.475 (df = 17890)	0.477 (df = 4668)
F Statistic	46.709*** (df = 1; 17893)	331.095*** (df = 4; 17890)	99.377*** (df = 4; 4668)

Note:

*p<0.1; **p<0.05; ***p<0.01

El tamaño del efecto encontrado (-4.51% a los 12 meses y -5.11% a los 6 meses) es relativamente pequeño pero estadísticamente significativo. Esto puede considerarse un efecto moderado en el contexto de intervenciones de salud pública, donde cambios en la percepción de salud general pueden reflejar una combinación de múltiples factores. Con los controles, la R^2 mejora de forma significativa, lo que quiere decir que si explica mejor los datos.

Primeras diferencias

a) Análisis de primeras diferencias (datos basales y doce meses después)

Call:

```
lm(formula = diff_health_gen_bin ~ treatment_0m, data = merged_data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.04467	-0.04467	-0.02235	-0.02235	0.97765

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.044675	0.007203	6.202	5.79e-10 ***
treatment_0m	-0.022324	0.010188	-2.191	0.0285 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.51 on 10023 degrees of freedom

Multiple R-squared: 0.0004788, Adjusted R-squared: 0.0003791

F-statistic: 4.801 on 1 and 10023 DF, p-value: 0.02846

El coeficiente del tratamiento nos indica que recibir el tratamiento esta relacionado con una mejor en la salud autoreportada de un -2.23% en comparación al grupo control (con significancia del 0.05).

b) Tabla de 2x2

	treatment	mean_health_0m	mean_health_12m
1	0	0.3929373	0.4390846
2	1	0.3691688	0.3888208

[1] -0.02649517

Cuando utilizamos diferencias en diferencias simple para calcular el efecto del tratamiento, obtenemos un valor de -2.65%, el cual es mayor que cuando calculamos el efecto utilizando el modelo de primeras diferencias, que es -2.23%. Esto se debe a que, al utilizar el modelo de primeras diferencias, nos fijamos solo en el efecto de las personas que estaban presentes en ambos periodos. En el modelo de diferencias en diferencias simple, se incluyen todas las observaciones disponibles, independientemente de si los individuos están presentes en ambos periodos. Es posible que en este segundo modelo tengamos más personas que recibieron el tratamiento en el grupo de doce meses después, lo que hace que el efecto real se perciba mayor.

c) Utilizando MCO.

$$\Delta\text{Health Gen Bin} = \beta_0 + \beta_1 \text{Treatment} + \epsilon$$

Call:

```
lm(formula = diff_health_gen_bin ~ treatment_0m, data = merged_data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.04467	-0.04467	-0.02235	-0.02235	0.97765

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.044675	0.007203	6.202	5.79e-10 ***
treatment_0m	-0.022324	0.010188	-2.191	0.0285 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.51 on 10023 degrees of freedom

Multiple R-squared: 0.0004788, Adjusted R-squared: 0.0003791

F-statistic: 4.801 on 1 and 10023 DF, p-value: 0.02846

Se cumple el mismo resultado (-2.23%).

d) Modelo de efectos fijos

Primeras diferencias

a) Estimación de primeras diferencias

Oneway (individual) effect First-Difference Model

Call:

```
plm(formula = health_gen_bin ~ treatment, data = pdata, model = "fd")
```

Unbalanced Panel: n = 28270, T = 1-3, N = 42291

Observations used in estimation: 14021

Residuals:

Min.	1st Qu.	Median	3rd Qu.	Max.
-1.023251	-0.023251	-0.023251	-0.023251	0.976749

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	0.0232508	0.0041896	5.5497	2.914e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 3450.4

Residual Sum of Squares: 3450.4

R-Squared: -0.0021968

Adj. R-Squared: -0.0021968

El intercepto representa la diferencia promedio en la salud general autoreportada entre los tres periodos. Hay un aumento de 2.33% en la probabilidad de reportar buena salud general.

b) Tabla 3x2

A tibble: 6 x 3

	time	treatment	mean_health
	<dbl>	<int>	<dbl>
1	0	0	0.393
2	0	1	0.369
3	6	0	0.465
4	6	1	0.407
5	12	0	0.439

6	12	1	0.389
---	----	---	-------

1. DID 6 meses:

$$\begin{aligned} DID_{6m} &= (\text{mean_health}_{6m,\text{treatment}} - \text{mean_health}_{0m,\text{treatment}}) \\ &\quad - (\text{mean_health}_{6m,\text{control}} - \text{mean_health}_{0m,\text{control}}) \end{aligned}$$

$$\begin{aligned} DID_{6m} &= (0.407 - 0.369) \\ &\quad - (0.465 - 0.393) \\ &= 0.038 - 0.072 \\ &= -0.034 \end{aligned}$$

2. DID 12 meses:

$$\begin{aligned} DID_{12m} &= (\text{mean_health}_{12m,\text{treatment}} - \text{mean_health}_{0m,\text{treatment}}) \\ &\quad - (\text{mean_health}_{12m,\text{control}} - \text{mean_health}_{0m,\text{control}}) \end{aligned}$$

$$\begin{aligned} DID_{12m} &= (0.389 - 0.369) \\ &\quad - (0.439 - 0.393) \\ &= 0.020 - 0.046 \\ &= -0.026 \end{aligned}$$

c) MCO