Pontificia Universidad Católica de Chile

Facultad de Agronomía e Ingeniería Forestal

Problema 1

Experimento Aleatorio

Curso: Evaluación de Impacto de Políticas Ambientales - AGE3802

Profesor: Rodrigo Arriagada

Diogo Valente Polónia

03/11/2022

Problema 1: Experimento Aleatorio

El objetivo de esta tarea es evaluar los efectos de la implementación, en las plantas de la empresa química EarthFriends, de nuevos protocolos de Sistemas de Manejo Ambiental (SMA) en sus niveles de polución. El análisis fue realizado con recurso a Stata y sigue anexado a este informe lo *.do file* creado.

1. Verificación de la asignación aleatoria a los grupos de tratamiento y control

Según la ley de los grandes números, a través de una asignación aleatoria de un dado programa, para muestras suficientemente grandes, todas las variables observables o no observables se distribuyen igualmente entre los grupos de control y tratamiento. Esto significa que no hay diferencias estadísticamente significativas entre los grupos, excepto la característica de tratamiento, es decir, que es posible inferir que las diferencias de productividad entre los grupos es causada solo por el tratamiento. Así, es importante evaluar si existen diferencias estadísticamente significativas entre las mismas variables observables de los distintos grupos, porque si no, entonces no se puede garantizar que la asignación aleatoria fué efectiva. Para esto, realizamos una prueba *t* para las medias de distribución de estas variables para los distintos grupos, con las siguientes hipótesis:

- H0: No hay diferencias entre las medias de los dos grupos
- H1: Hay diferencias entre las medias de los dos grupos

En el caso de las variables rechacen la hipótesis nula, i.e. valor p < alfa (0.05), habrá pruebas estadísticamente significativas de que la media de estas variables es diferente entre los grupos, por lo que no se podría comparar los grupos de tratamiento y control, sin controlar estas variables.

El *test t* fué hecho en Stata, con recurso a lo comando “ttest”, agrupando las observaciones por valor de *assignment* y sus resultados se pueden ver en la siguiente tabla (con interpretación para un nivel de significancia de 5%):

Variable	**P-Value** (Pr(|T| > |t|))	**Interpretación**
pre_pollution	0.1765	No rechaza H0
age	0.1765	No rechaza H0
employees	0.2885	No rechaza H0

Tabla 1 -Resultados del test t *entre grupos de control y tratamiento*

Problema 1: Experimento Aleatorio

Como no hay variables que rechazan la hipótesis nula (valor p < alfa), no hay pruebas estadísticamente significativas de que la media de estas variables sea diferente entre los grupos de control (plantas status quo) y tratamiento (plantas con nuevos protocolos), por lo que no se puede concluir que la asignación de tratamiento no fué aleatoria. Los valores p son significativos a 5%, 10% y 15%, lo que permite una gran confianza estadística.

2. Verificación de la asignación aleatoria a los grupos con distintos protocolos

Siguiendo lo mismo pensamiento de la situación anterior, se realizó una prueba *t* solo dentro de las plantas tratadas (agregand el comando "if protocol != 0" al "ttest"), entre el grupo de las plantas asignadas al protocolo de implementación a nivel de planta (*protcol = 1*) y el grupo de las asignadas al protocolo de implementación corporativa (*protcol = 2)*. Se presentan los resultados en la Tabla 2 (con interpretación para un nivel de significancia de 5%).

Variable	**P-Value** (Pr(|T| > |t|))	**Interpretación**
pre_pollution	0.0386	Rechaza H0
age	0.0386	Rechaza H0
employees	0.1322	No rechaza H0

Tabla 2 -Resultados del test t *entre grupos con diferentes tratamientos*

Hay evidencia estadística de un sesgo de selección en la asignación del tipo de protocolo, ya que hay variables que rechazan la hipótesis nula, en que el valor p < alfa (pre_pollution y age). Así hay pruebas estadísticamente significativas de que la media de estas variables es diferente entre el grupo con implementación a nível de planta y el grupo con standard, por lo que no se pueden comparar los grupos sin controlar por estas variables.

Esto significa que los grupos no son comparables, es decir, que no son estadísticamente iguales. En términos prácticos, significa que el impacto en los resultados en la polución entre los tipos de implementación puede no deberse únicamente al tipo de implementación, si no que a diferencias inherentes entre los grupos, lo que hace que la evaluación no sea válida internamente. Para superar este problema, podemos controlar las distintas variables observadas incluyéndose en el modelo de regresión, de modo que sus efectos no se confundan con el efecto del tratamiento, lo que nos permite confiar más en los resultados. Sin embargo, para aumentar la solidez de la evaluación, habría que hacer una nueva selección aleatoria, hasta que no se produzcan diferencias.

Problema 1: Experimento Aleatorio

3. Valores estimados de la polución con sus respectivos intervalos de confianza al 95%

Siendo Y la variable de resultado (*outcome*), entonces Y deberá ser la variable *post_pollution*, y la variable de tratamiento T deberá ser la variable *assignment*, que indica se la planta fué asignada un nuevo protocolo de SMA. Para calcular los valores esperados se utilizó el comando "ci means" en Stata y sus resultados se presentan en la Tabla 3.

	Valor Esperado	**Intervalo al 95%**	**Amplitude Intervalo**	**Numero Observaciones**
E(Y)	4383.195	[3675.697; 5090.694]	1408.996	60
E(Y|T=1)	3813.044	[3033.840; 4592.247]	1558.408	40
E(Y|T=0)	5523.499	[4115.223; 6931.776]	2816.552	20

Tabla 3 -Resultados de la media de la post_pollution con intervalo de confianza al 95%

De estos resultados se puede concluir lo siguiente:

1. Se detectan diferencias entre los niveles de polución entre los grupos de tratamiento y control, con una disminución de 1710.455 (5523.559 - 3813.044) del nível de polución. O sea, se detecta un impacto (β) del tratamiento:

$$\beta = E(Y|T = 1) - E(Y|T = 0) = -1710.455$$

2. Este impacto es significativo porque lo valor esperado E(Y|T=1) no está contenido en el intervalo de confianza de E(Y|T=0), y vice-versa. Sin embargo, como los intervalos de confianza se intersectan, la significancia estadística puede no ser muy elevada lo que puede incitar a controlar las variables observadas para aumentar la significancia;
3. Las amplitudes de los intervalos de confianza aumentan, reflectando la disminución del tamaño muestral de cada uno de los grupos

4. Estimación del impacto con regresión lineal usando la variable assignment

Haciendo la regresión a través de lo comando "reg" en Stata, sin controlar por ninguna variable, se obtienen los resultados de la Figura 1.

Como se esperaba, obtenemos un impacto estimado (visto a través del coeficiente de la variable assignment) de -1710.456, lo mismo calculado anteriormente. Además, este coeficiente, para un nivel de significancia de 5%, es estadísticamente significativo, ya que el

valor p (2.1%) es menor que el nivel de significancia (5%). Con esta regresión se podría concluir, entonces, que la implementación de nuevos protocolos de Sistemas de Manejo Ambiental (SMA) disminuye los niveles de polución en 1710.456 valores. Esta conclusión es válida porque no se identificaron diferencias significativas entre los grupos y no hay evidencia de que los grupos no fueran aleatorios.

Source	SS	df	MS	Number of obs	=	60
				F(1, 58)	=	5.61
Model	39008779.6	1	39008779.6	Prob > F	=	0.0212
Residual	403541345	58	6957609.39	R-squared	=	0.0881
				Adj R-squared	=	0.0724
Total	442550124	59	7500849.56	Root MSE	=	2637.7

post_pollu~n	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
assignment	-1710.456	722.3716	-2.37	0.021	-3156.439	-264.4718
_cons	5523.499	589.8139	9.36	0.000	4342.858	6704.14

Figura 1 - Resultados de la regresión lineal entre la variable post_pollution *y* assignment *sin controlar por ninguna otra variable*

5. Estimación del impacto controlando por las variables observables

La Figura 2 presenta los resultados de la regresión lineal controlada por las variables *age* y *employees*, que, como ya se había confirmado, se consideran igualmente distribuidas entre los grupos de tratamiento y control (no hay evidencia estadística para contrariarlo). Por lo tanto, no se espera un impacto significativo de estas variables en los valores *post_pollution*, sólo una reducción del error de estimación y un consecuente aumento de la significancia.

De hecho, el control de la regresión por las variables *age* y *employees* (edad de la planta y número de empleados) permite aumentar la significancia de la estimación: el *valor p* disminuir de 2.1% para 1.3% y el error de la estimación (*_cons*) disminuir de 5523.499, os sea el valor de E(Y|T=0), para 4827.919.

Se concluye con mayor grado de significancia que la implementación de los nuevos protocolos de Sistemas de Manejo Ambiental (SMA) disminuye los niveles de polución en 1894.012 valores, un pequeño aumento del impacto en relación al calculado antes.

Los coeficientes de las variables age y employees no presentan un valor significativo pues su *valor p* es mucho mayor que 5% (15% y 97%, respectivamente), por lo que no hay evidencia que estas variables contribuyen significativamente para los valores de polución de las plantas,

pero como su introducción disminuye el error, se puede concluir que estas variables tienen una relación con el error.

Por todo esto se concluye que el resultado no es sustancialmente distinto de lo anterior pero es más significativo.

Source	SS	df	MS	Number of obs	=	60
				F(3, 56)	=	2.57
Model	53534026.7	3	17844675.6	Prob > F	=	0.0634
Residual	389016097	56	6946716.03	R-squared	=	0.1210
				Adj R-squared	=	0.0739
Total	442550124	59	7500849.56	Root MSE	=	2635.7

post_pollu~n	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
assignment	-1894.012	739.8779	-2.56	0.013	-3376.166	-411.8585
age	141.8294	98.08312	1.45	0.154	-54.65471	338.3135
employees	-.1963911	4.959705	-0.04	0.969	-10.13188	9.739093
_cons	4827.919	1108.012	4.36	0.000	2608.304	7047.535

Figura 2 - Resultados de la regresión lineal entre la variable post_pollution *y* assignment *controlando por* age *y* employees