Business Applications of Data Analytics

Juan José Zunino – 20F1150

Introducción

El objetivo de este informe es presentar y analizar la técnica de *propensity score matching,* las ventajas que presenta a la hora de realizar inferencia estadística y sus posibles campos de aplicación.

Motivación

Un estudio observacional es una investigación empírica cuyo objetivo es explicar la relación entre causa/efecto en casos donde no es posible realizar una experimentación controlada, en el sentido de asignar aleatoriamente gente a distintos procesos (Cochran, 1995).

Para desarrollar y explicar la técnica estadística, se hará uso de datos sobre el otorgamiento de prestamos. El objetivo es definir si existe *una diferencia estadísticamente significativa* de los otorgamientos de créditos entre personas que se graduaron y las que no. Esto debe ser considerado un estudio observacional, ya que es *casi* imposible experimentar sobre el otorgamiento aleatorio de prestamos entre individuos graduados y no graduados, ya que de ser realizado podría significar una posible pérdida para la entidad financiera.

En primera instancia, se realizará una prueba estadística conocida como *t-test* para evaluar si el data set se encuentra balanceado. Luego, se aplicará la técnica de *propensity score matching*, evaluando sus resultados y comparando con los obtenidos anteriormente. Por último, se realizará un *caliper matching* y se comparará los resultados obtenidos.

Sesgo en estudios observacionales

Los estudios observacionales pueden estar sesgado, es decir, que consistentemente se favorezca a las personas graduadas a la hora de ser otorgadas un préstamo respecto de las personas no graduadas, y viceversa. Para demostrar esto, realizamos un test de medias a todas las covariables observadas menos la de interés (i.e., otorgamiento del préstamo).

X: $Graduado$ ($Grupo\ de\ control$)
Y: $No\ es\ graduado$ (Grupo de tratamiento)

$$H_0 : \mu_X - \mu_Y = 0$$
$$H_1 : \mu_X - \mu_Y \neq 0$$

Si el p-valor asociado al estadístico es menor/igual a 0.1, decimos que hay evidencia estadísticamente suficiente para rechazar la hipótesis nula, a un nivel de confianza del 90%.

	Covariable	P-valor
1	MarriedYes	0.9713
2	Dependents1	0.755
3	Self_EmployedYes	0.9111
4	ApplicantIncome	0
5	CoapplicantIncome	0.0128
6	LoanAmount	0
7	Loan_Amount_Term	0.0694
8	Credit_History1	0.253
9	Property_AreaSemiurban	0.4001
10	Property_AreaUrban	0.5667

De esta manera podemos concluir que 4 de 10 covariables presentan evidencia suficiente de que están desbalanceadas

Propensity score matching

Propensity score matching es una técnica estadística que intenta estimar el efecto de un tratamiento, una política, u otra intervención controlando por las covariables que predicen el recibimiento del tratamiento. Al realizar un pareamiento por puntaje de propensión, se espera que, al balancear las variables observables, se balanceen las variables no observables, y así poder eliminar cualquier *overt y hidden bias*. Como es complejo realizar emparejamientos individuales, los individuos son emparejados en base a su *propensity score*, que es la probabilidad de que un individuo reciba el tratamiento dadas las covariables observadas. De esta manera, generamos un set de matches, al cual le realizamos el mismo test de diferencias de media para corroborar que los covariables observadas hayan sido balanceadas:

	Covariable	P-Valor
1	MarriedYes	0.7633
2	Dependents1	0.8857
3	Self_EmployedYes	0.4297
4	ApplicantIncome	0.7027
5	CoapplicantIncome	0.99
6	LoanAmount	0.4671
7	Loan_Amount_Term	0.4306
8	Credit_History1	0.8559
9	Property_AreaSemiurban	1
10	Property_AreaUrban	0.8756

Podemos observar, que para un nivel de confianza del 90%, no tenemos evidencia suficiente para rechazar la hipótesis nula. De esta manera, podemos corroborar como nuestro data set quedo balanceado

Una vez balanceado el data set, realizamos el *test de Welch* para corroborar si la media de los individuos que reciben préstamo del grupo de control es igual a la media de los individuos que reciben el préstamo del grupo de tratamiento:

```
        Welch Two Sample t-test

data:  treated.y and control.y
t = -0.59934, df = 191.86, p-value = 0.5497
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1769458  0.0944716
sample estimates:
mean of x mean of y
0.6288660 0.6701031
```

Concluimos que no hay evidencia suficiente para afirmar que la media de los individuos que recibe el préstamo del grupo de control es diferente de la media de los individuos que recibe el préstamo del grupo de tratamiento.

Propensity score caliper matching

Otra manera de emparejar individuos es con el uso de *caliper,* el cual genera emparejamientos si la distancia (*Mahalanobis* en este caso) de propensity score es menor a un valor determinado, 0.2 para nuestro estudio. El objetivo es obtener un punto medio entre un buen balanceo de las características observables y obtener buenos emparejamientos individuales.

```
                       stand.diff.before stand.diff.after
GenderMale                   0.153439641      0.111831366
MarriedYes                   0.004102293     -0.043045978
Dependents1                  0.035620950      0.020733775
Self_EmployedYes            -0.012672364     -0.119839887
ApplicantIncome             -0.393669042      0.026520660
CoapplicantIncome           -0.219957317      0.001519452
LoanAmount                  -0.509595308      0.069987248
Loan_Amount_Term            -0.228163158     -0.122614243
Credit_History1             -0.135441891      0.028040132
Property_AreaSemiurban      -0.095379824      0.000000000
Property_AreaUrban          -0.064805935     -0.022360967
```

Se puede observar que con el uso de caliper, de todas formas, logramos balancear las covariables observadas. Una vez realizado el emparejamiento corrigiendo por el caliper, realizamos el *test de Welch:*

```
        Welch Two Sample t-test

data:  treated.y and control.y
t = -2.3867, df = 187.31, p-value = 0.018
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.28245658 -0.02682177
sample estimates:
mean of x mean of y
0.6288660 0.7835052
```

En este caso, concluimos que hay evidencia suficiente para rechazar la hipótesis nula que plantea que la media de los individuos que reciben un préstamo del grupo de control y el de tratamiento es igual

Conclusión

A partir de los test estadísticos que realizamos, específicamente el del caliper, determinados que hay evidencia suficiente para rechazar la hipótesis de que la media de prestamos que son otorgados al grupo de control es igual a la media de prestamos otorgados al grupo de tratamiento.

A la hora de realizar inferencia estadística en estudios observacionales, es importante conocer y aplicar adecuadamente técnicas como *propensity score matching,* ya que la misma permite corregir el sesgo observable (*overt bias*), y al hacer eso, se espera que también se elimine el sesgo no observable (*hidden bias*). De esta, manera podemos obtener coeficientes más robustos, que mejoran la calidad de la inferencia.

Aclaración: Todos los gráficos fueron generados con el script matching.R