Universitat
de les Illes Balears

Causal inference and heterogeneous treatment effects

Joan Gabriel Salom Pons

Memoria del Trabajo de Fin de Máster

Máster Universitario en Análisis de datos masivos en economía y empresa
(Técnicas y aplicaciones a la gestión económica y empresarial)
de la

UNIVERSITAT DE LES ILLES BALEARS

Curso Académico 2017-2018

2018-09-06

Tutor del Trabajo Dr. Jan Olof William Nilsson

Causal inference and heterogeneous treatment effects

Joan Gabriel Salom Pons
Tutor: Dr. William Nilsson

Treball de fi de Màster Universitari en Analisi de Dades Massives en Economia i Empresa (MADM)

Universitat de les Illes Balears
07122 Palma de Mallorca
joan.salom1@estudiant.uib.cat

Resumen

En el presente trabajo de fin de máster, se busca realizar una introducción a la inferencia causal, tanto desde un enfoque teórico como práctico, poniendo principal énfasis en los avances que se han producido en dicha disciplina a raíz de la introducción del *machine learning*. Para ello, se introducirá inicialmente el marco teórico y técnico convencional de la inferencia causal, suministrando análisis y ejemplos prácticos, para continuar con la exposición de las nuevas técnicas y ventajas suministradas con la inclusión del *machine learning*.

Abstract

The main objective of the present master's thesis consists in the introduction of the causal inference discipline from both theorical and practical perspective, paying special attention to the advances achieved in causal inference after the introduction of machine learning. First, theorical and practical framework of conventional assumptions and causal inference methods will be introduced, and following that introduction, the advantages of the introduction of machine learning in the field of the causal inference will be assessed.

Palabras clave: inferencia causal, árboles causales, heterogeneidad, efectos heterogéneos del tratamiento, *outcomes* potenciales, balanceo, ignorabilidad, tratamiento, *causal MARS*, *propensity scores*.

1. Introducción

Cuando se habla de inferencia causal, se está hablando de una disciplina que lleva siglos de gestación, pero que se ha desarrollado como disciplina en las últimas décadas. Aquello que hace distinta a la inferencia causal de la inferencia estadística convencional es que la segunda, en base a las distribuciones de las variables de los elementos de una muestra, estima modelos y parámetros en base a los cuales deduce una relación entre variables, mientras que en la inferencia causal se busca realizar inferencia de una causa hacia una variable de estudio a partir de un planteamiento de contrafactualidad. Se plantea, pues, qué impacto hubiera habido en la variable dependiente objeto de estudio de no haberse dado el valor observado en la variable causal, en comparación a lo que sí se ha dado y observado.

Hay dos grandes figuras a la hora de comentar aportaciones formales a la disciplina de la inferencia causal: Rubin (que a partir de los 60 desarrolla un trabajo previo introducido por Neyman en 1923), que desarrolla e introduce el modelo de *outcomes* Potenciales, y Pearl, que en la década de los noventa introduce los Modelos Causales Estructurales (SCM por sus siglas en inglés).

Se pondrá el foco en este primer apartado introductorio en el modelo de *outcomes* potenciales de Rubin, puesto que es el marco teórico del cual parten los modelos de *machine learning* para la inferencia causal que se abordan en el trabajo.

1.1. Neyman y Rubin: modelos causales y outcomes potenciales

En 1923, Neyman plantea un concepto novedoso para su época, que acuña como "rendimiento potencial" en un artículo en el que valora cómo estimar el rendimiento de un terreno de cultivo.

La gran novedad contenida en su trabajo es que habla de rendimiento potencial, pero no real, dado que cada parcela sólo puede ser cultivada con una variedad de cultivo concreta. Planteando un sistema de asignación aleatoria de los cultivos (a través de una urna) demuestra que se puede obtener una estimación insesgada del valor del terreno partir de los datos observados.

Rubin [9] desarrolla durante la década de los 70 las ideas planteadas por Neyman, y contribuye a formalizar lo que hoy en día se conoce como modelos de *outcome* potenciales. Los modelos de outcomes potenciales se basan en la idea de que, los individuos sujetos a un tratamiento1 determinado, al que llamaremos A (que para simplificar, supondremos que puede tomar dos valores a = 1 ó a = 0), pueden manifestar distintos *outcomes* potenciales Ya, en función del valor que tome el tratamiento. En un mundo ideal, de observarse todos los *outcomes* potenciales, calcular un efecto causal del tratamiento podría ser tan sencillo como calcular las diferencias de valores que toman los *outcomes* potenciales bajo las distintas versiones del tratamiento y promediarlas. Con un tratamiento A binario dónde A toma los valores 0 ó 1, el efecto causal del tratamiento para la población objeto de estudio podría definirse como (1) E(Y1 – Y0). No obstante, los *outcomes* potenciales de los individuos no son observados, dado que el individuo sólo puede estar sujeto a una versión del tratamiento a la vez.

De tomarse únicamente los valores observados para cada individuo (en control y en tratamiento) y se tratase de estimar el efecto causal directamente sobre los datos, se estaría computando como efecto causal (2) E(Y | A = 1) – E(Y| A=0), que no tiene por qué coincidir necesariamente con (1), dado que en (1) se estaría calculando el efecto causal utilizando los *outcomes* potenciales de todos los individuos de la muestra, mientras que en (1) se estarían utilizando solamente los *outcomes* observados para individuos condicionando a que han recibido (A=1) o no han recibido (A=0) el tratamiento.

No obstante, como Rubin señala, se puede conseguir un buen estimador del efecto causal a partir de los datos observados, partiendo de la aleatoriedad en la asignación del tratamiento.

[1] Inicialmente, llamaremos tratamiento a una intervención bien definida por el investigador y que resulta accionable por el mismo.

Si bien el propio Rubin plantea que es difícil la aleatoriedad de por sí en el diseño de los experimentos, ésta puede simularse controlando las variables de confusión más correlacionadas con el *outcome* y la asignación del tratamiento.

Así pues, los modelos de *outcomes* potenciales tienen como principal premisa, la de aproximar a partir de los datos observados en los individuos bajo el experimento, los *outcomes* potenciales y estimar así los efectos causales. Para hacer dichas estimaciones, una serie de supuestos son requeridos:

- SUTVA: que establece que el *outcome* potencial de un individuo debería permanecer inalterado ante cambios en la asignación del tratamiento en otros individuos
- Consistencia: establece que el *outcome* potencial de un individuo es igual al *outcome* observado para el mismo valor del tratamiento: E(Ya) = E(Y|A=a)
- Positividad: establece que para cualquier valor de las X de los individuos debe haber una mínima probabilidad de recibir cada una de las posibles variaciones del tratamiento P(A=a|X=x) > 0.
- Ignorabilidad: un supuesto clave que establece que la asignación del tratamiento, condicionando por las variables X de las que se dispone debe ser independiente de los *outcomes* potenciales.

Dicho supuesto de ignorabilidad requiere de una explicación más detallada, puesto que es uno de los supuestos clave y que fácilmente resulta violado en experimentos no aleatorizados, requiriendo de esfuerzos para tratar los datos de tal forma que se pueda considerar aceptable tal supuesto.

1.2. Modelo de *outcomes* potenciales desde una aproximación práctica

Para reforzar la exposición de los conceptos previamente expuestos, se propone un análisis práctico de un dataset de acceso público. El análisis se realiza a través del programa estadístico R, y los paquetes utilizados para el análisis son Tableone, MatchIt y Matching.

El ejemplo en este caso práctico usa datos que están a disposición del público, acerca de un estudio realizado por Lalonde (1986), en el que estudia el posible impacto de un programa de formación laboral, sobre la renta de los beneficiados por el mismo. La variable objeto de estudio es la diferencia de renta en función de la percepción de la formación.

Las características de las variables se detallan en el anexo.

De haberse realizado un reparto aleatorio de la formación entre los individuos, el investigador podría proceder directamente a analizar las diferencias en las variables dependientes observadas, entre aquellos individuos en la muestra de control y la de tratamiento, haciendo los contrastes pertinentes (por ejemplo, un *test-t*, como el presentado en la figura 1:

```
        One Sample t-test

data:  sample_assuming_random_diff
t = -3.5631, df = 184, p-value = 0.0004668
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -4432.379 -1273.148
sample estimates:
mean of x
-2852.764
```

Figura 1: elaboración propia

Aparentemente, la formación habría sido perjudicial para los receptores de ésta, al observarse una caída significativa en las rentas percibidas tras la formación. Sin embargo, lo que realmente ocurre, es que se ha producido una violación del supuesto de ignorabilidad, dado que la asignación de la formación (el tratamiento) no es aleatoria ni independiente de las otras variables recogidas en el dataset y que afectan al outcome, y el efecto causal estimado en el cuadro 1 está viciado por el impacto de dichas variables en la renta.

Es fácil pensar que la formación irá enfocada a individuos con “circunstancias” más desfavorables o en situaciones de más pobreza, lo que es lo mismo que decir que la formación se asignará con más probabilidad a aquellos individuos que presenten una serie de características (valores de variables) que estén asociadas (correlacionadas) con niveles más bajos de renta, lo cual implica que en la muestra de tratamiento los valores de las X (y en consecuencia del outcome) son distintos (previsiblemente más bajos) a los de la muestra de control antes de dar ningún tratamiento, impidiendo así que la medición de las diferencias de los outcomes en una muestra y otra sea un buen estimador (insesgado) del efecto causal promedio.

Toda esta explicación es más fácil de ver en la figura:

	Stratified by treat		
	0	1	SMD
n	429	185	
age (mean (sd))	28.03 (10.79)	25.82 (7.16)	0.242
educ (mean (sd))	10.24 (2.86)	10.35 (2.01)	0.045
black (mean (sd))	0.20 (0.40)	0.84 (0.36)	1.668
hispan (mean (sd))	0.14 (0.35)	0.06 (0.24)	0.277
married (mean (sd))	0.51 (0.50)	0.19 (0.39)	0.719
nodegree (mean (sd))	0.60 (0.49)	0.71 (0.46)	0.235
re74 (mean (sd))	5619.24 (6788.75)	2095.57 (4886.62)	0.596
re75 (mean (sd))	2466.48 (3292.00)	1532.06 (3219.25)	0.287

Figura 2: Elaboración propia

La distribución de los receptores de la formación revela que los niveles de estudios son menores, predominan las personas de color, solteras y con un salario en periodos previos al tratamiento mucho menor.

Para poder obtener una correcta medición del efecto causal de la formación en la renta, el investigador debe controlar las variables de confusión que están dificultando la correcta apreciación del efecto causal. Para ello, es habitual recurrir a métodos de macheo, para procurar que las distribuciones de dichas variables sean lo más homogéneas posibles entre ambas muestras. Los métodos y criterios utilizados para lograr un macheo eficaz pueden diferir y se abordarán en mayor o menor medida durante el trabajo, pero la idea fundamental que gira a su alrededor es la de lograr que el supuesto de ignorabilidad del tratamiento pase de ser un supuesto quebrantado por el propio diseño del experimento a un supuesto plausible, a través de la agrupación de individuos “similares” (puede ser porque presentan unos valores de sus X similares, o unas probabilidades de recibir el tratamiento casi idénticas), para que así, el hecho de haber recibido o no el tratamiento para uno u otro individuo deje de estar ligado a sus características particulares y pase a ser una cuestión de aleatoriedad.

Hay diversos métodos de macheo para lograr simular un experimento aleatorizado a partir de los datos de un experimento que no lo es. Uno de los métodos, cuyo cuadro se presenta a continuación, consiste en machear una a una las observaciones de la muestra bajo tratamiento con su “mejor pareja posible” en la muestra de control, usando la distancia de *mahalanobis* y generando las parejas uniendo a las observaciones de tratamiento con aquellas de control con una distancia menor. Dicho método de macheo se conoce como *greedy matching*, y al hacerse por parejas, de modo *one-to-one*.

La figura 3 recoge una submuestra de los datos tras aplicar un método de emparejamiento uno a uno basado en la minimización de la distancia de *mahalanobis* de las características de los individuos

en tratamiento con los de control. Se eligen las mejores parejas posibles para todos los individuos en la muestra de tratamiento.

	Stratified by treat		
	0	1	SMD
n	185	185	
age (mean (sd))	24.21 (9.55)	25.82 (7.16)	0.190
educ (mean (sd))	10.23 (2.37)	10.35 (2.01)	0.052
black (mean (sd))	0.43 (0.50)	0.84 (0.36)	0.943
hispan (mean (sd))	0.06 (0.24)	0.06 (0.24)	<0.001
married (mean (sd))	0.20 (0.40)	0.19 (0.39)	0.027
nodegree (mean (sd))	0.69 (0.46)	0.71 (0.46)	0.035
re74 (mean (sd))	2681.77 (4754.79)	2095.57 (4886.62)	0.122
re75 (mean (sd))	1523.69 (2810.24)	1532.06 (3219.25)	0.003

Figura 3: Elaboración propia

Como puede apreciarse, el emparejamiento ha generado que los individuos seleccionados de entre la muestra de control para emparejarse con los de la muestra bajo tratamiento tengan una distribución de las variables muy similar a las de los individuos en tratamiento.

El investigador podría probar con otro método de emparejamiento, basado en la *propensity score* [8] de los individuos del experimento. El criterio de emparejamiento basado en emparejar a aquellos individuos con probabilidades de recibir el tratamiento se suele denominar como *balancing score*. En este método, primero se estima, mediante el método apropiado para el tipo de tratamiento (en este caso, al ser binario, una regresión logística podría ser utilizada), la probabilidad de recibir el mismo para cada individuo en base a las variables de confusión recogidas. Adicionalmente, los métodos de emparejamiento suelen ir acompañados del establecimiento de un criterio de mínimos, conocido con un *caliper*, y se descartan aquellas observaciones de tratamiento que no encuentran una pareja en la muestra de control mínimamente similar. De forma que, del mismo modo que los métodos de emparejamiento refuerzan y hacen plausible el supuesto de ignorabilidad del tratamiento, el *caliper* refuerza el de positividad.

Tras el emparejamiento, el efecto promedio del tratamiento difiere notablemente del reportado en la figura 1, sobre los datos sin emparejar:

```
        One Sample t-test

data:  y_diffs
t = 1.3601, df = 110, p-value = 0.1766
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -534.2272 2872.0436
sample estimates:
mean of x
 1168.908
```

Figura 4: elaboración propia.

El curso de formación parece tener un impacto positivo en lugar de negativo, con un efecto promedio de 1168 u.m. Sin embargo, dicha diferencia no resulta significativa.

1.3. Efectos del tratamiento heterogéneos, significado, semántica y soluciones

En el apartado previo se ha mencionado como factor importante en el análisis causal el control de las variables de confusión. A través de los ejemplos, se ha visto como el control de las variables de confusión, (por ejemplo, a través de emparejamiento por distancia de mahalanobis o por la *propensity score*) permite obtener estimaciones insesgadas del efecto causal del tratamiento.

No obstante, en ocasiones la obtención de dicho efecto causal a través de métricas agregadas, como el efecto promedio del tratamiento, puede encerrar información relevante acerca del impacto real del tratamiento entre los pacientes y llevar al investigador a error. En el siguiente ejemplo se intenta plasmar la problemática de los efectos heterogéneos del tratamiento y la importancia de analizarlos debidamente para conocer el impacto de este.

Para el siguiente ejemplo, se han simulado una serie de variables X, una asignación aleatoria del tratamiento y un ourtcome que es función de las variables generadas y del tratamiento. Simula un estudio en el que se quiere analizar el impacto de unas supuestas sesiones de formación en individuos que están empleados en trabajos no cualificados, y se busca medir el impacto de las sesiones de formación en el salario que perciben los individuos tras participar en dicho experimento.

Las variables simuladas son:

- Age: número entero, generado a partir de redondear una normal N(26, 3.5)
- Gender: variable binaria, 1 representa mujer, 0 hombre, generadas con equiprobabilidad.
- Wg_past: variable obtenida una N(18000, 3500).
- degree: variable binaria, 1 indica estudios de grado y superiores, 0 ausencia de estudios superiores, generada con equiprobabilidad.
- treated: binaria que recoge el recibimiento o no del tratamiento: se ha generado aleatoriamente y con equiprobabilidad, por lo que el principio de ignorabilidad se respeta ya desde el propio diseño del experimento.

Wg_after_treatment: representa el outcome tras el experimento, se ha generado de la siguiente manera:

(6) $wg\_after\_treatment_i = 3500 + 0.8*wg\_past_i - 200*gender_i - 1300*degree_i + 2500*treated_i + 1500*treated_i*degree_i + e_i$

siendo e N(0, 200)

De la ecuación 6, hay un componente que indica heterogeneidad en el efecto del tratamiento sobre el salario. El salario es función únicamente de las variables explicativas escogidas (no se ha omitido ninguna variable relevante a la hora de recopilar los datos) y el tratamiento afecta directamente a la renta con un efecto marginal constante e igual a 2500 para todos los individuos excepto para un subconjunto concreto, el de los graduados, donde el efecto del tratamiento asciende hasta 4000 gracias a la interacción con coeficiente 1500 recogida en la recta de regresión.

El procedimiento descrito en el apartado 1.1 del trabajo facilita la obtención de un efecto causal promedio del tratamiento. De acuerdo con lo descrito, el investigador podría estudiar la distribución de los regresores entre la muestra de control y la de tratamiento, y observaría que son sumamente parejas (dado que la asignación del tratamiento ha sido aleatoria). Confiando en que no debería sufrir sesgos por variables de confusión, podría proceder a testear el impacto del tratamiento y reportar el efecto causal promedio del mismo. Dicho efecto estimado se reporta en la figura 5:

```
        Paired t-test

data:  all_sample_treated and all_sample_control
t = 52.375, df = 493, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3207.065 3457.063
sample estimates:
mean of the differences
              3332.064
```

Figura 5: elaboración propia

De acuerdo con el test computado, el efecto causal de las políticas de formación sobre el salario de los beneficiados es positivo, significativo y está comprendido entre 3200 y 3457 u.m. Sin embargo, se está infravalorando el impacto de la formación sobre los graduados, a la vez que sobrevalorando el impacto de dichas políticas de formación sobre el resto de la muestra.

De sospechar la existencia de dicha heterogeneidad del efecto del tratamiento en función la disposición o no de un título universitario por parte del beneficiado, podría estratificarse la muestra en base a dicha variable y estimar los efectos causales del tratamiento en ambas muestras por separado.

Para poder haber estimado correctamente el cálculo del efecto causal para uno y otro subconjunto, idealmente el investigador debería haber presupuesto que el efecto causal diferiría en función de la posesión o no de estudios superiores, y ya a priori se debería haber diseñado el experimento teniendo eso en mente y preparando los datos para la estratificación.

No resulta difícil pensar por qué motivos la existencia de efectos causales heterogéneos de los tratamientos es un reto al que los investigadores deben hacer frente, ni tampoco intuir que la estratificación de la muestra no resulta una solución viable cuando el número de variables a controlar es elevado o el efecto heterogéneo va ligado a variables continuas en los pacientes. La identificación y estimación de la existencia de efectos heterogéneos puede ser clave en experimentos y ensayos clínicos, en los que, por ejemplo, un medicamento puede tener un impacto positivo para algunos pacientes, pero negativo en otros grupos, y la no identificación de la heterogeneidad llevaría a reportar un efecto causal promedio para toda la muestra que no sería representativo de los distintos efectos que se dan entre los distintos individuos del estudio.

Sin embargo, grandes avances en la estimación de efectos causales heterogéneos se han dado en épocas cercanas a la actualidad, consistiendo en la aplicación de métodos propios del *machine learning*, junto con teorías y supuestos ya existentes, como los del modelo de outcomes potenciales de Rubin.

2. Machine Learning e Inferencia Causal

El *machine learning* es una disciplina de la inteligencia artificial enfocada a que los ordenadores aprendan de los datos. El término fue acuñado en 1959 por Arthur Samuel, y la disciplina ha ido creciendo en importancia y relevancia tanto en el ámbito académico como el profesional. En los campos bajo el machine learning, destaca el aprendizaje supervisado, en el que el algoritmo trata de predecir con la máxima precisión posible una variable dependiente objeto de estudio. Para ello, se utilizan una serie de variables explicativas y se entrenan y validan distintos algoritmos.

En el aprendizaje supervisado, hay una serie de algoritmos que han ganado popularidad: los árboles

de decisión, y algoritmos derivados como *random forest* y *boosting*. Dichos algoritmos, aunque en ocasiones resultan útiles para afrontar determinados problemas, y agradan al proporcionar una visualización de las relaciones entre los regresores y la variable dependiente, no resultan adecuados para los problemas de la inferencia causal en su diseño original (Neville 1998).

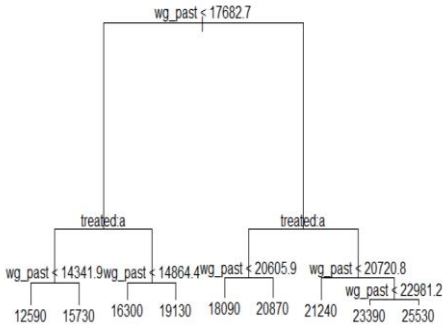

Figura 6: Elaboración propia

2.1. Árboles de decisión y árboles causales

En los últimos años, diversos autores (Li et al [6], Athey e Imbens [2], Powers et al. [7] entre otros) han realizado el esfuerzo de incorporar y adaptar algoritmos de *machine learning* para que sean aplicables en la inferencia causal, resultando en modelos capaces de estimar el efecto causal del tratamiento en entornos donde dicho efecto es heterogéneo.

Para los datos expuestos en el apartado 1.2.1 se ha estimado un árbol de regresión incluyendo como variables explicativas las diferentes características de los individuos y la variable binaria que indica si han recibido o no la formación. Los resultados se presentan en la figura 6.

El árbol de clasificación empieza incluyendo como primer corte el salario percibido con anterioridad al tratamiento. A posteriori, las dos regiones generadas se dividen en función de la recepción o no de la formación, y nuevamente, en función del salario anterior al tratamiento. Se observa como en cada hoja que desciende de haber recibido el tratamiento, el salario es mayor que en aquellas en las que no se ha recibido. No obstante, dicho árbol no resulta apropiado para estimar efectos causales dado que no controla los efectos de las variables de confusión (de haberlas habido), y tampoco detecta la heterogeneidad del tratamiento, ya que no busca dividir las regiones en función de encontrar los efectos más heterogéneos del tratamiento, sino los valores promedio más heterogéneos en el outcome, pudiendo generar hojas en las que solo se localicen principalmente observaciones de tratamiento o de control, generando una hoja con unos valores homogéneos del outcome pero sin una lectura a nivel de efecto causal.

A continuación, se presenta el análisis utilizando el algoritmo de árbol causal de Athey & Imbens [2] en su modalidad "honesta":

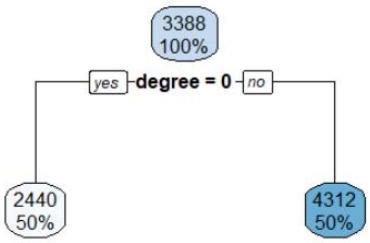

Figura 7: elaboración propia

Dicho árbol presenta el efecto promedio estimado del tratamiento para el conjunto de individuos, así como cortes en las variables en las que se localiza la heterogeneidad en dicho tratamiento. Se presenta en este árbol causal como punto de corte la posesión o no de título universitario y se computa el efecto causal del a formación en el caso de no poseer el título (2440 u.m.) y en el caso de poseerlo (4312 u.m.), así como el efecto causal promedio para la muestra observada (3388 um).

Athey e Imbens [2] adaptan en su trabajo el algoritmo de árboles de decisión a la inferencia causal. Una diferencia evidente entre una y otra modalidad del árbol es que las hojas devuelven el efecto del tratamiento estimado (como diferencia de los outcomes promedio con y sin tratamiento en la hoja) en lugar del outcome promedio.

Otra modificación que añaden a los árboles de clasificación para convertirlos en árboles causales es la opción de estimar los árboles de modo honesto. La modalidad honesta consiste en realizar una partición en la muestra de entreno del árbol, usando una fracción para construir las particiones y la otra, independiente de la primera, para calcular el efecto estimado del tratamiento en las hojas resultantes de las particiones anteriores. De este modo, los autores se aseguran de que haya insesgadez en la estimación del efecto causal en las hojas, al haberse usado una muestra independiente a la que construyó el árbol para estimar los resultados de las hojas, a costa de sacrificar observaciones de entreno en la construcción del árbol. Se produce un trade-off varianza sesgo en comparación al criterio

adaptativo habitualmente utilizado, que puede llevar a ganancias en el ajuste.

Otra modificación adicional en relación con los árboles de decisión realizada por los autores radica en la aplicación de distintos criterios a la hora de generar particiones en el árbol causal. Proponen distintos criterios de partición en su trabajo, planteando la división basada en t-test sobre las diferencias de efecto causal estimado entre las dos potenciales hojas, y la extrapolación de la minimización del *RMSE* (con ligeras modificaciones, como su estimación modificada bajo el criterio de honestidad, y la maximización del negativo del *RMSE* esperado bajo honestidad) a árboles de outcomes transformados y de ajuste en las hojas a partir de una predicción con una función de regresión en lugar de una partición.

Para el árbol presentado en el cuadro once, se ha recurrido a la estimación mediante modalidad honesta, y utilizando un input adicional que exige un tamaño mínimo de observaciones de control y tratamiento en cada hoja. Junto con el árbol causal, Athey et al. [3] desarrollan para R en el paquete *causalTree* el árbol causal previamente presentado, así como su ensamblado en *causalForest* y *propensityForest*. *causalForest* ensambla los árboles causales de un modo equivalente al que los *randomForest* agrupan árboles de decisión: eligiendo un subconjunto de regresores aleatorio en cada corte de cada árbol, construyendo los distintos árboles y promediando los outcomes predichos en sus hojas para devolver el outcome estimado para cada observación dados los valores de sus regresores. En el caso del causalForest, la predicción consiste en un efecto causal estimado del tratamiento para cada individuo en función de los valores de sus regresores.

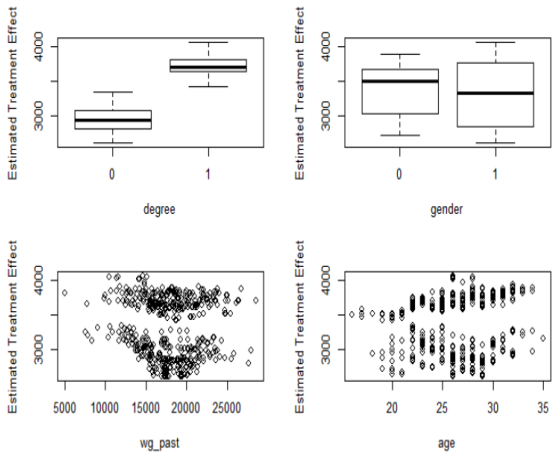

Figura 8: elaboración propia

En la figura 8 se grafican los valores predichos para el efecto causal del tratamiento para los datos simulados para el apartado 1.3 del trabajo y los valores para cada variable del individuo. Se aprecia como la heterogeneidad en la estimación del efecto causal de la política se da con la variable *degree*. Para ninguna de las otras variables se aprecia un patrón claro de heterogeneidad en el efecto del tratamiento estimado. El bosque causal parece haber capturado la heterogeneidad del tratamiento en función del grado, si bien las estimaciones muestran cierto grado de error.

Los propios autores han desarrollado numerosos paquetes estadísticos para R que permiten una rápida implementación de los algoritmos de árboles causales y causal Forest. Uno de los paquetes más reconocidos es grf, que desarrolla una generalización de los algoritmos de árboles con implantaciones específicas a la inferencia causal. A continuación, se presentan la estimación del efecto causal en función de la variable que causa heterogeneidad, junto con la tabla que presenta la importancia de cada una de las variables en la construcción de los árboles en un causalForest con 400 árboles sobre los datos simulados.

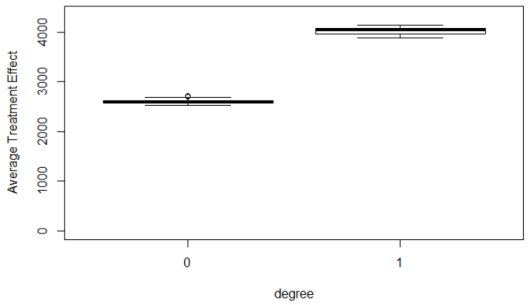

Figura 9: elaboración propia

Variable Names <fctr>	variable Importance <dbl>
age	0.13847032
gender	0.05947697
degree	0.59120486
wg_past	0.21084786

Figura 10

Como se puede apreciar, la estimación del efecto causal es precisa, capturando la heterogeneidad del efecto de la posesión del grado en el impacto de la política formativa e identificando de la importancia de dicha variable presentándola como la más

recurrida en la construcción de los árboles del bosque

Tanto el causalTree como el causalForest expuestos previamente parten del supuesto de ignorabilidad, que, en muchas ocasiones, no tiene porqué sostenerse. Por esas razones, desarrollan el propensityForest. El propensityForest construye el árbol usando las variables que mejor predicen la recepción del tratamiento, y a continuación predice el outcome en las hojas generadas, permitiendo así contemplar y controlar los casos en la que la asignación del tratamiento no es aleatoria y el principio de ignorabilidad se vulnera, pero quedando incompleto como instrumento para la inferencia causal al no tener en cuenta la posible heterogeneidad en el tratamiento a la hora de construir los árboles.

2.2. Métodos para efectos heterogéneos en experimentos no aleatorizados

Los métodos expuestos en el apartado 2.1 parten del supuesto de que el experimento está aleatorizado, supuesto que, en muchos casos, no se da en experimentos reales al no poderse asignar los tratamientos de forma aleatoria por una u otra cuestión. Los *propensityForest* de Athey et al. [3] lidian con dicho problema, pero no con la heterogeneidad en el efecto del tratamiento. Powers et al. [8] proponen una serie de métodos capaces de lidiar con efectos de tratamiento heterogéneos en contextos de no aleatoriedad en la asignación del tratamiento, siempre que se controlen todas las variables que influyen en su asignación. Los métodos que proponen se explican a continuación:

2.2.1 Pollinated Transformed Outcome (PTO) Forests

El punto de partida de este algoritmo radica en obtener una estimación de la probabilidad de recibir el tratamiento (propensity score) para cada individuo, y transformar su outcome y_i a Z_i, siguiendo Z_i la siguiente expresión:

$$\textbf{(6) } Z_i = T_i * \frac{Yi}{\pi(Xi)} + (1 - T_i) * \frac{-Yi}{1-\pi(Xi)},$$

resultando Z_i en un estimador insesgado del efecto del tratamiento, τ. A continuación, se realizan entre 2 y 3 pasos:

- Se estima Z_i a partir de las X disponibles para cada observación vía *random forest*, construyendo un bosque F.
- Con el bosque F, se pulen las estimaciones en cada hoja sustituyendo Z_i por Yi, por separado para individuos bajo tratamiento y bajo control. Es decir, con el árbol ya construido en el paso 1, se sustituye el *outcome* a calcular en las hojas, de Z a Y, por separado para tratados y no tratados. De la diferencia de *outcomes* obtenidos en el paso 2 se obtiene la estimación de τ.
- (opcionalmente) se construye un *random forest* para predecir la predicción de τ obtenida en el paso dos, dotando de mayor interpretabilidad (vía importancia de las variables en el árbol) al modelo.

El paso 2 ayuda a reducir la varianza en las estimaciones del efecto del tratamiento, dado que se sustituye Z por Y. Cabe señalar que Z contiene la *propensity score* en el denominador, cuyo valor puede tender a 0 o a 1, causando que el valor de Z sea altamente volátil. Reemplazando Z por Y en el paso 2 se logra suavizar dicho problema.

2.2.2 Causal Boosting

Se propone una adaptación del método de boosting para la inferencia causal. Principalmente, se utiliza como instrumento de aprendizaje un árbol causal modificado para devolver, en lugar del efecto del tratamiento esperado para cada hoja, los dos pares de medias condicionadas del outcome para los tratados y no tratados respectivamente. Salvando el particular instrumento de clasificación débil, el resto del algoritmo funciona de la manera convencional en el boosting: se empieza con una función vacía f(x) =0, que resulta en errores Yi – f(x), y posteriormente se van incorporando nuevas funciones ajustadas por un parámetro de aprendizaje sobre los residuos, ajustando en cada etapa con la adición de un nuevo clasificador débil y el reajuste de los residuos. Los modelos generados van de 1 hasta K, que son el número de clasificadores débiles que se deciden incluir para construir el modelo de boosting, y se propone una adaptación de validación cruzada para seleccionar el mejor.

La manera en que se ajusta causal boosting a aquellos casos en las que el tratamiento no se puede considerar de asignación aleatorizada es mediante la estimación de la probabilidad de recibir el tratamiento y el ajuste del efecto del tratamiento estimado en cada hoja en función de la cantidad de gente que pertenece a los distintos “estratos” o rangos de probabilidad de recibir el tratamiento. La probabilidad de recibir el tratamiento puede ser estimada mediante algoritmos de aprendizaje supervisado diverso, como la regresión logística. Powers et al [8] proponen utilizar *propensity forests*, que vienen a ser *random forest* para

clasificación, pero que, en lugar de devolver la clase estimada por mayoría, devuelven las probabilidades estimadas.

Una vez se tienen las probabilidades, se decide hacer una división en diferentes estratos S, de igual longitud, por ejemplo, si S= 4, s1 = [0, 0.25), s2 = [0.25, 0.5), s3 = [0.5, 0.75) y s4 = [0.75, 1]. Para construir el árbol, se generan una serie de hojas l, con una cantidad de gente nl en cada hoja. Se trata entonces de ponderar el efecto del tratamiento (ATE) estimado en la hoja en función del efecto del tratamiento para cada estrato de probabilidad y el peso de cada estrato en la cantidad de gente nl que hay en toda la hoja.

$$(8)\ t=\frac{\sum_{s=1}^{S} nsl\ (\bar{Y}1sl-\bar{Y}0sl)}{\sum_{s=1}^{S} nsl}$$

De manera similar pueden obtenerse las varianzas en las hojas, y el criterio propuesto para decidir entre las opciones de corte se basa en un contraste sobre las diferencias de ATE estimadas en cada candidata a hoja.

2.2.3 Causal MARS

MARS (*multi-adaptative regression splines*) es un método de regresión no paramétrico flexible y que no está sujeto a restrictivos supuestos, como puede ser el caso de las regresiones lineales estimadas por MCO. El algoritmo añade al intercepto, una serie de funciones base: identifica variables y puntos de corte del tipo $Bi(x_i) = \max(0, xi - C)$ y $Bi(x_i) = \max(0, C - xi)$, siendo c el punto de corte. El modelo de MARS resultante consiste en el intercepto B_0 y una serie de adiciones del tipo $B_iB(x_i)$, siendo B_i el coeficiente asignado a la función base $Bi(x_i)$.

De manera más formal, la ecuación puede escribirse:

$$(12)\ y_i = F(X_i) = B_0 + \sum_{i=1}^{n} Bi * Bi(Xi)$$

Scott Powers et al. proponen adaptar MARS a la inferencia causal. La adaptación que realizan del MARS para encontrar la heterogeneidad en el efecto del tratamiento la realizan estimando por separado un MARS para los individuos bajo tratamiento y otro para los individuos en control, y para estimar los coeficientes, compara la caída del error de entrenamiento en el caso de incluir las funciones base con distintos coeficientes para la muestra de tratamiento y la de control con la caída en el caso de incluir las dos funciones base con el mismo coeficiente. Los coeficientes se estiman por MCO y el criterio de *backward regression* puede utilizarse para determinar el tamaño óptimo de funciones base. Para reducir la varianza del Causal MARS, se propone hacer *boostrap* a la muestra, estimar un *causal MARS* para cada submuestra de *boostrap* y promediar los resultados. El método también puede adaptarse a experimentos donde no se de la ignorabilidad, estimando para cada estrato de probabilidad de recibir el tratamiento un *causal MARS* con la misma función base pero distintos coeficientes y ponderando el criterio de optimización por los distintos estratos.

2.2.4 Generalized Random Forest con Variables Instrumentales

Hay otra aportación de Athey, Tibshirani y Wager [2] que resulta relevante en la inferencia causal, como es el Random Forest generalizado que permite la inclusión de variables instrumentales. El uso de las variables instrumentales está muy extendido en el campo de la econometría. Se aplica en casos en los que se considera que el tratamiento depende de variables de confusión que escapan del control del investigador y que potencialmente pueden sesgar los análisis. En ese caso, se recurre a usar variables instrumentales, variables que estén fuertemente correlacionadas con el tratamiento y sean independientes del outcome a estudiar, afectando al outcome únicamente de forma indirecta a través del efecto que el instrumento tiene en el tratamiento..

Los algoritmos descritos en 2.2.1,2.2.2 y 2.2.3 parten del supuesto de ignorabilidad: el tratamiento se supone plenamente independiente o independiente del outcome una vez condicionado por las variables disponibles[2]. Por tanto, sufren la amenaza de la omisión de variables de confusión del análisis, que, al quedar sin controlar, pueden generar sesgo en las estimaciones.

Por ello, la posibilidad de disponer de un método que pueda estimar la heterogeneidad en el tratamiento y además ser apropiado en estos casos, como lo sería *generalized random forest* con variables instrumentales permite abarcar de forma satisfactoria muchos más problemas de inferencia causal.

2.3. The SPRINT Data Analysis Challenge

Los métodos para capturar la posible heterogeneidad del tratamiento propuestos por Powers et al. [8] se ponen a prueba sobre datos reales en un reto propuesto en 2016 por la New England Journal of Medicine. En dicho reto, se facilitaron datos acerca de un experimento aleatorizado consistente en una intervención médica

[2] Es decir, el investigador dispone de todas las variables que afectan al tratamiento y puede controlarlas, estimar con precisión la *propensity score* y simular con éxito la aleatorización del tratamiento.

más intensiva sobre la presión sanguínea de los pacientes, y se propuso estimar el efecto de la intervención sobre el riesgo de tener problemas cardíacos de diversa índole.

Los dos métodos que los autores usan para estimar el efecto de la intervención sobre el riesgo de enfermedades cardíacas son *causal MARS* y causal Boosting. Los autores observan que los dos métodos llevan a estimaciones muy diversas del impacto del tratamiento en el riesgo. *Causal MARS* parece detectar una heterogeneidad en el efecto de la intervención médica más consistente de lo que detecta *causal Boosting*.

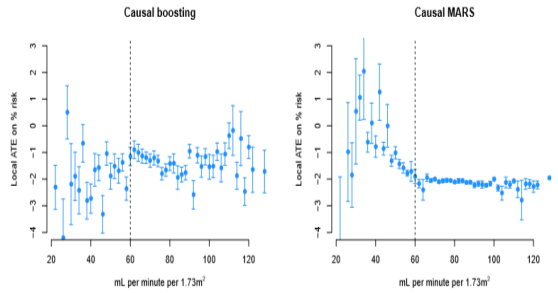

Figura 10: Some methods for heterogeneous treatment effect estimation in high-dimensions. Scott Powers et al. (2017)

Causal MARS encuentra, tanto en la muestra de entreno como validación, un efecto de -0.2 sobre el riesgo para pacientes que no estén crónicamente enfermos[3] . El efecto medido según *Causal Boosting* no es tan consistente.

3. Casos prácticos

3.1. Efectos del tabaco en el peso de los niños

En su artículo, Abrevaya [1] estudia el efecto de fumar durante el embarazo en el peso de los niños. Se aplican sobre los datos del artículo[4] los métodos de *PTO Forests* y *Causal MARS*. Primeramente, se procede a estimar la propensity score para cada individuo, dado que los datos no parecen balanceados.

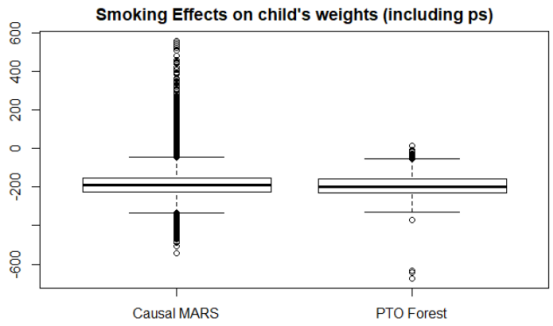

Figura 11: elaboración propia

Ambos métodos reportan un impacto de una reducción de 200 gramos en el peso del niño, sin detectarse heterogeneidad aparente en dicho efecto. No obstante, cabe señalar que la estimación de la *propensity score* no presenta un buen ajuste, revelando que, probablemente, hay variables de confusión sin controlar, pudiendo generar sesgo en las estimaciones obtenidas[5].

3.2. Gap salarial en el mercado malasio

En esta aplicación práctica se plantea estimar el impacto que tiene pertenecer al género femenino en el salario (en escala logarítmica) que se percibe en el mercado de trabajo en Malasia. Del dataset original se realiza un filtrado de datos, seleccionando solo a los individuos para los cuales se tiene registrado un salario. Se define una variable dummy para recoger el sexo femenino. Para ello, se toma como referencia la dummy *men* y se asigna a women un 0 cuando men toma valor de 1 y viceversa. Se hace así para, en el caso de necesitar balancear los datos, se dispongan de más observaciones en control que en tratamiento, dato que el porcentaje de hombres en la muestra ya filtrada era mayor.

Se aplican cuatro métodos sobre la muestra: *Causal Forest*, *Causal MARS*, *Causal Boosting* y *PTO Forests*. La distribución del efecto causal estimado por cada método es similar, siendo el efecto promedio estimado de alrededor de -0.3 en el salario por el hecho de ser mujer. Además de estimar el efecto promedio, se detectan tres variables que causan heterogeneidad, siendo la más relevante[6] la siguiente:

[3] Ml per minuto y 1.73m^2 menor a 60 indica enfermedad.
[4] Filtrando únicamente el primer nacimiento para cada madre en el panel de datos.

[5] Sin embargo, el efecto de -200 no es tan distinto de otras estimaciones referenciadas en el paper.
[6] La experiencia potencial y la edad causan efectos heterogéneos en el gap, medido en valor absoluto. No obstante, a medida que aumenta la experiencia y la edad, también lo hace el salario en valor absoluto, con lo que parece coherente que la edad sea un factor que afecte al gap salarial, al menos en valor absoluto.

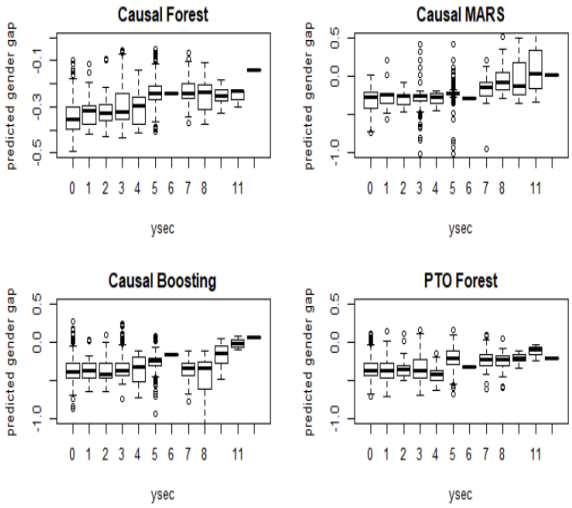

Figura 12: elaboración propia

De forma aparentemente consistente, parece que los años de educación tras la educación primaria (*ysec*)[7] contribuyen a reducir el gap. Para mujeres sin estudios más allá de la educación primaria, el gap parece incluso superior al -0.3. A medida que los estudios superiores crecen, el gap tiende a 0. Esta detección de heterogeneidad podría servir de recomendación para que, de quererse reducir la desigualdad que sufren las mujeres en el mercado laboral malasio, se abogase por políticas que incentivasen a las mujeres a seguir estudiando una vez superada la educación primaria.

3.3. El experimento de seguros de salud en Oregón

Este experimento [5] puede considerarse un caso donde urge aplicar métodos de variables instrumentales. Para individuos con bajos niveles de renta, se hace una lotería y los seleccionados tienen derecho a pedir un seguro de salud. Se busca estimar el impacto de no poseer seguro sobre el grado de enfermedad mental del individuo.

Aproximamos el problema de dos vías: primeramente, considerando que no tener un seguro de salud puede ser función de múltiples variables no observadas, se usa la lotería como instrumento y se estima por *instrumental forest* el efecto de la falta de seguro sobre la (falta de) salud mental. La segunda vía se basa en estimar las *propensity scores*, y, estratificando por ellas, aplicar *causal MARS* y causal Boosting. Se compara el efecto estimado por instrumental forest y causal MARS en la figura siguiente:

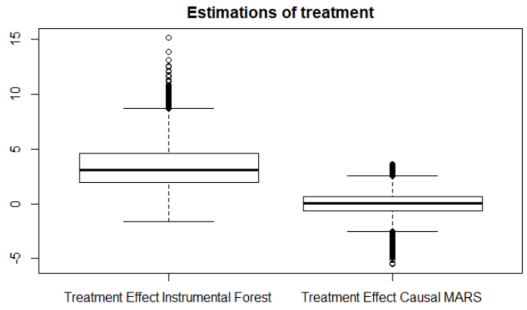

Figura 1·: elaboración propia

Como se aprecia, mientras *causal MARS* no parece estimar un efecto superior a 0, *instrumental forest* indica que la falta de seguro incrementa la medida de problemas mentales en 3 puntos en promedio. Resulta razonable pensar que causal MARS estima sesgadamente por la falta de control de variables de confusión que no hayan sido incluidas en el estudio.

4. Conclusiones

Las ventajas de la introducción de machine learning en la inferencia causal son relevantes, especialmente cuando se trata de identificar la heterogeneidad en el efecto de los tratamientos. Si en el pasado era necesario preparar el experimento y los subgrupos a analizar a priori, con antelación a la puesta en marcha de los experimentos si se deseaba estudiar posible heterogeneidad en el tratamiento, los nuevos métodos presentados en el trabajo permiten detectar esa posible heterogeneidad a posteriori de diseñar el experimento, y sin necesidad de correr múltiples contrastes de hipótesis.

Sin embargo, cabe destacar la importancia de los supuestos que se asumen para la aplicación de los modelos: *causal MARS*, *causal Boosting* y *PTO Forest* asumen la ignorabilidad en el tratamiento y podrían dar lugar a estimaciones sesgadas del efecto del tratamiento. En caso de que no esté claro que se dé el supuesto de ignorabilidad del que parten los citados algoritmos, Generalized Random Forest con variables instrumentales de Athey, Tibshirani y Wager [2] puede resultar una opción mucho más recomendable, ya que ayuda a prevenir ese sesgo. Por tanto, si bien la introducción de las técnicas de *machine learning* amplía las herramientas del inestigador y puede ayudar a encontrar variables que causan heterogeneidad en el efecto del tratamiento de forma más sencilla, sigue siendo vital entender los supuestos de los modelos y el diseño y la asignación del tratamiento: en los casos en que la ignorabilidad sea un supuesto razonable en el experimento, y sea porque el investigador

[7] ysec mide los años que los individuos continúan estudiando tras superar la primaria. Comprende los años de educación secundaria, pero también puede comprender los de educación universitaria y superior.

entiende que el tratamiento no depende de ninguna variable o porque se dispone y se puede controlar toda variable que lo afecta, *causal MARS* y *causal Boosting* pueden resultar muy adecuadas para afrontar el problema. Si la ignorabilidad es un supuesto discutible, y se intuye que hay variables de confusión no observadas, es probable que estos métodos estimen sesgadamente y un enfoque basado en variables instrumentales sea mucho más apropiado.

Una última observación que debe hacerse hace referencia a la valoración de los modelos. Su introducción ha sido muy reciente y la literatura aún está en desarrollo, por lo que no existe un consenso claro acerca de cómo valorar y comparar el ajuste de los modelos descritos durante el trabajo. Schuler et al. [11] proponen varios criterios de validación para medir el ajuste de los modelos: *μ-risk*, basado en modelos que midan por separado el *outcome* de los tratados y no tratados, y ajusten lo mejor posible una función de riesgo basada en cada *outcome*, y el *t-risk*, basado en emparejar individuos similares a raíz de la distancia de *mahalanobis*, computar una estimación del efecto del tratamiento y optimizar una función de riesgo basada en dicha estimación. No obstante, sigue sin haberse establecido un procedimiento consolidado para valorar el ajuste de los modelos de inferencia causal, cosa que el investigador debería tener en cuenta.

Con todo, la introducción del *machine learning* en el campo de la inferencia causal están en pleno desarrollo, su aportación resulta tremendamente útil para la investigación, pero, lejos de ser algoritmos que puedan aplicarse con ligereza, resulta clave tener un conocimiento profundo de los supuestos de los modelos y de las características del experimento y el tratamiento.

Referencias

[1]. Jason Abrevaya. Estimating the effect of smoking on birth outcomes using a matched panel data approach, enero 2004.

[2]. Susan Athey, Julie Tibshirani y Stefan Wager. Generalized Random Forest, arXiv:1610.01271v3 [stat.ME], Current version, July 2017. Annals of Statistics, (forthcoming).
https://arxiv.org/abs/1610.01271
https://github.com/swager/grf7.

[3]. Susan Athey y Guido W. Imbens. Yanyag Kong y Viras Ramachandra. An introduction to recursive partitioning for heterogeneous treatment effects estimation using causalTree package, septiembre 2016.

[4]. Susan Athey y Guido W. Imbens. Recursive partioning for heterogeneous causal effects, diciembre 2015.

[5]. Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker. The Oregon health insurance experiment: evidence from the first year*.

[6]. KJ Jager, C. Zoccali, A. Macleod y FW Dekker. Counfounding: what it is and how to deal with it, *abc of epidemiology, International Society of Nephrology*, 2008.

[7]. Jiuyong Li, Saisai Ma, Thuc Duy Le, Lin Liu y Jixue Liu. Causal Decision Trees, *School of Information Technology and Mathematical Sciences, University of South Australia, Australia Mawson Lakes, SA 5095*, 2015.

[8]. Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H. Shah, Trevor Hastie y Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high-dimensions, julio 2017.

[9]. Paul R.Rosenbaum y Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association, Vol. 79, No. 387*, septiembre 1984.

[10]. Donald B.Rubin. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology, Vol. 66, No. 5, 688-701*, 1974.

[11]. Alejandro Schuler, Michael Baiocchi, Robert Tibshirani y Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects, 2018.

[12]. Stefan Wager y Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests, julio 2017

Apéndice

Neyman

En su trabajo de 1923, Neyman plantea un estudio para valorar terrenos de cultivo. A la hora de calcular el rendimiento estimado para dicho cultivo, recurre a una expresión de valor esperado como la siguiente:

(1) ai = $\sum_{k=1}^{m} Uik$ / m

donde ai sería el rendimiento esperado para una variedad de cultivo i, siendo i el índice para posible tipo de cultivo, de i= 1 hasta v, y siendo k cada una de las parcelas disponibles para plantar, siendo m el total, en el terreno de cultivo de que se dispone.

Rubin

Suponiendo el ejemplo que introduce Rubin (1974), en un experimento con dos sujetos de estudio:

El efecto causal promedio que se quiere estimar es:

(4) $$\frac{1}{M} * \sum_{j=1}^{M} [yj(E) - yj(C)]$$

Siendo M la cantidad de sujetos en el estudio, y E la exposición a un tratamiento dado y C la exposición a su alternativa (en el caso de un tratamiento binario). Dicha ecuación implica promediar las diferencias entre los outcome potenciales que se darían de asignarse E o C a cada individuo j de la muestra de estudio. El problema radica en que dichos outcomes potenciales no son observables, y se deben inferir.

De haberse asignado aleatoriamente E y C a cada sujeto del experimento, veríamos, con equiprobabilidad al primer individuo recibiendo el tratamiento E y al segundo C que, al contrario, dando como valor esperado:

(5) $\frac{1}{2} * [y1(E) - y2(C)_{+} \frac{1}{2} * [y2(E) - y1(C)]$

Así pues, bajo aleatoriedad en la asignación del tratamiento, el efecto causal esperado estimado en los datos observados se corresponde con el efecto causal promedio planteado en la ecuación (4).

Lalonde dataset

El dataset contiene datos acerca de 614 personas, recopilados para el estudio. 185 reciben la formación, y 429 permanecen como muestra de control, y se recoge información sobre 10 variables distintas para cada individuo. Ocho de las variables son candidatas que se podrían considerar variables de confusión. Se recogen porque el investigador intuye que pueden estar correlacionadas con el *outcome*. La variable *treat* recoge de forma binaria si el individuo ha recibido la formación (1) o si es parte de la muestra de control (0). Las variables *educ* y *nodegree* recogen información acerca de los estudios de los individuos: *educ* los años de formación y *nodegree* la tenencia (0) o no (1) de estudios universitarios. Age mide la edad, *black* si el individuo es (1) o no (0) de etnicidad negra, hispan lo propio para la hispana, y *married* el estado civil (1 casado, 0 no). *Re74* y *re75* recogen rentas percibidas años antes de la formación, y *re78* recoge la renta tras la formación laboral y es la variable dependiente.

Validación cruzada para causal Boosting

Para la selección del mejor modelo por validación cruzada, se propone la siguiente ecuación:

(7) $\sum X\ e\ V$({Gk(x,1) − Gk(x,0)} −{HK(x,1) − HK(x,0)}) ^2

donde Gk(x,1) − Gk(x,0) supone el efecto del tratamiento estimado por el modelo de boosting con k árboles, donde k puede ir de 1 hasta K, y HK(x,1) − HK(x,0) representaría el efecto del tratamiento estimado con los datos de la muestra de validación en el modelo boosting con los K máximos árboles, con la misma estructura que la construida en la muestras de entrenamiento pero reajustando las predicciones de los nodos terminales del árbol.