PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Taller 3: Material Complementario

Datos para la evaluación de impacto en políticas públicas
Diplomado en Ciencia de Datos para Políticas Públicas

29 de Septiembre 2022

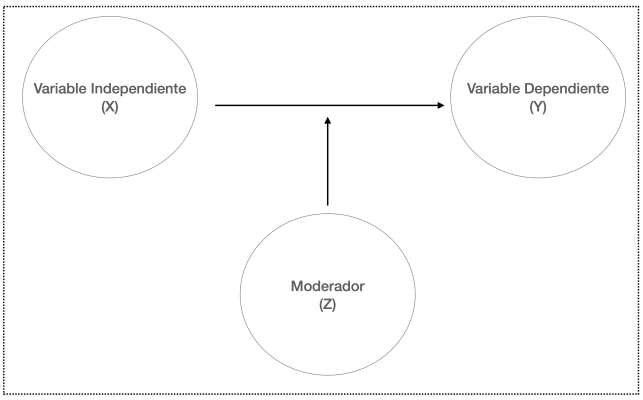
I Términos no lineales

Resumen

En línea con los supuestos de la regresión OLS para que sea el **B**est **L**inear **U**nbiased **E**stimate (mejor estimación lineal insesgada) se pueden incorporar las transformaciones no lineales. Esto nos ayuda a capturar de manera más apropiada el patrón de los datos reduciendo los problemas de **sesgo** y **eficiencia** en nuestras estimaciones. Este tipo de transformaciones son muy utiles a nivel sustantivo y técnico, ya que nos permiten variar nuestras hipótesis y mejorar la calidad del modelo.

Las más utilizadas en la disciplina son:

1. **Término Interacción**: Se incluye un término interacción en la regresión cuando pensamos que el "efecto" de una variable independiente "X" sobre una variable dependiente "Y" podría ser moderada dependiendo el valor de una variable independiente "Z". En ese sentido, los moderadores pueden ser categóricos (género, grupo etario, etc.) o continuos (edad, nivel de estres, etc.) y pueden amortiguar o exacerbar efectos. Esto es lo que se conoce como moderación:



Esto queda representado bajo la forma:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon_1$$

¿Cómo se interpretan las interacciones?

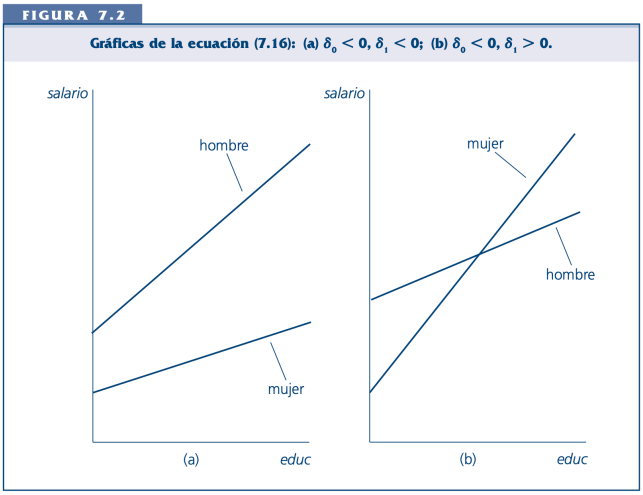
a. Referirse al efecto principal.

b. Indicar si la interacción presenta un resultado estadísticamente significativo. En otras palabras, indicar si las pendientes son distintas entre los grupos.

Para determinar las pendientes podemos usar (ejemplo sencillo):

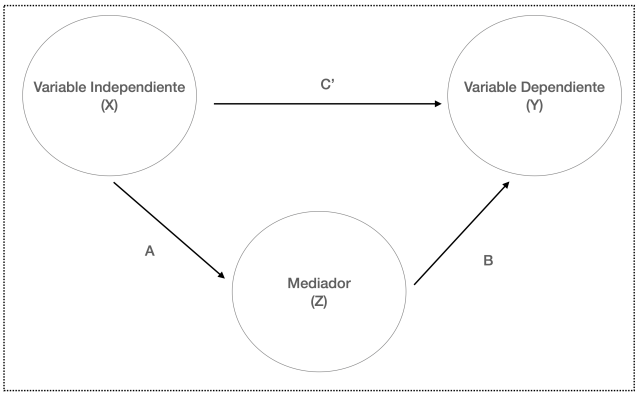
	Z=0	Z=1
Y =	$(\beta_1 + \beta_3 * 0)X = \beta_1 X$	$(\beta_1 + \beta_3 * 1)X = (\beta_1 + \beta_3)X$

c. Presentar visualmente la interacción e indicar los patrones observados para cada grupo.



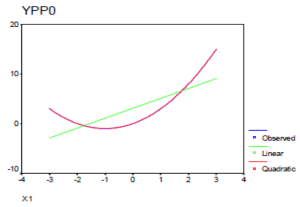
Moderación ≠ Mediación

Ojo que esto es distinto a una mediación. Este tipo de análisis responde a la pregunta de cuánto de la relación entre "X" e "Y" es explicada por "Z". En otras palabras, una variable mediadora se encuentra en el camino entre una variable predictora (independiente) y una variable de resultado (dependiente), siendo así un mecanismo de acción a través del cual opera una variable sobre otra. **Idealmente deberían estar ordenadas en una secuencia temporal que sea concordante**.

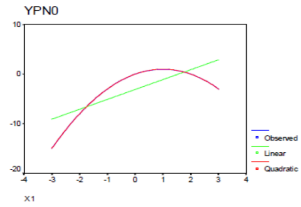


2. **Términos Cuadráticos (polinomios)**: Incluir estos términos nos permite dar cuenta de mejor manera de una relación en la que la pendiente aumenta y luego decrece (convava) o caso contrario (convexa). En otras palabras hay un punto de cambio en la dirección de la pendiente (asocación de X con Y).

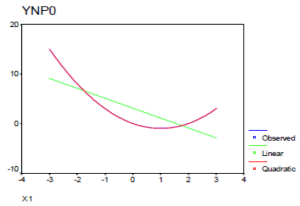
b1 positive, b2 positive; $Y = 2X + X^2$



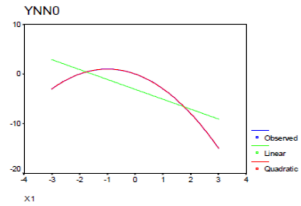
b1 positive, b2 negative; $Y = 2X - X^2$



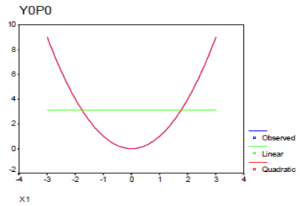
b1 negative, b2 positive; $Y = -2X + X^2$



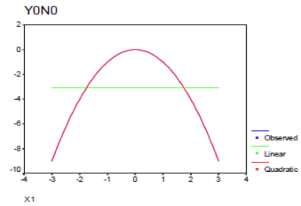
b1 negative, b2 negative; $Y = -2X - X^2$



b1 zero, b2 positive; $Y = X^2$

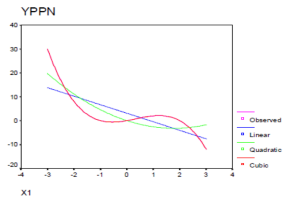


b1 zero, b2 negative; $Y = -X^2$

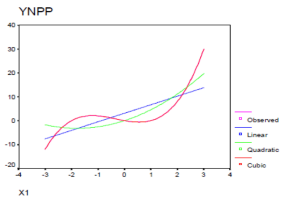


También podemos incorporar términos cúbicos:

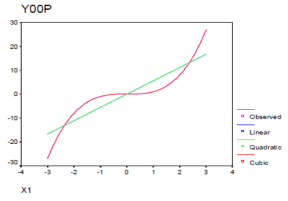
b1 positive, b2 positive, b3 negative; $Y = 2X + X^2 - X^3$



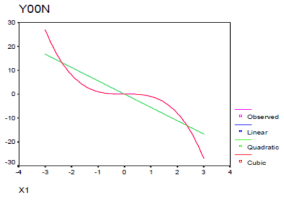
b1 negative, b2 positive, b3 positive; $Y = -2X + X^2 + X^3$



b1 zero, b2 zero, b3 positive; $Y = X^3$



b1 zero, b2 zero, b3 negative; $Y = -X^3$



En definitiva, este tipo de transformaciones nos ayudan a **flexibilizar la forma funcional de nuestra asociación**. En consecuencia vamos a capturar de mejor manera el patrón de los datos. Un coeficiente cuadrático o cúbico estadísticamente significativo nos sugiere que la relación no es lineal. A su vez, ambos términos deben ser interpretados en función del β principal que representa la asociación de X e Y.

3. **Términos logarítmicos**: Se incluye este tipo de términos cuando hay una relación que crece de manera exponencial en vez de aritméticamente. Esto con el fin de normalizar la distribución de la variable (recuerde que nuestro objetivo es hacer más creíble los supuestos de OLS). En general estos términos se interpretan como el cambio de la variable dependiente (Y) ante una variación porcentual de la variable independiente (X). Esto de la forma:

$$Y_i = \beta_1 * log(X_i)$$

La siguiente tabla puede ayudar al calculo y estimación de variables transformadas a logarítmos. Recuerda que cada interpretación se sitúa en un contexto de investigación particular.

Modelo	Regresión	Variable Dep. (Y)	Variable Indep. (X)	Interpretación del regresor (β_1)
Nivel - Nivel	$Y_i = \beta_0 + \beta_1 X_i + u_i$	Y	X	$\Delta Y = \beta_1 \Delta X$
Nivel – Log	$Y_i = \beta_0 + \beta_1 \log(X_i) + u_i$	Y	log(X)	$\Delta Y = \left(\frac{\beta_1}{100}\right) \%\Delta X$
Log - Nivel	$\log(Y_i) = \beta_0 + \beta_1 X_i + u_i$	log(Y)	X	$\%\Delta Y = (100\beta_1)\Delta X$
Log-Log	$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + u_i$	log(Y)	Log(X)	$\%\Delta Y = \beta_1 \%\Delta X$

1. El modelo Nivel-Nivel representa las variables en su escala original, por lo que simplemente se interpreta como cualquier modelo de regresión OLS. Es decir, un cambio de una unidad en X, afecta en β_1 unidades a Y.

2. En el modelo Nivel-Log la variable independiente es un logaritmo, por lo que se interpreta como un incremento del 1% de cambio en X es asociado a un cambio en Y de $(0,01 * \beta_1)\%$.

3. En el modelo Log-Nivel la variable dependiente es un logaritmo, entonces se interpreta como un incremento de 1 unidad en X es asociado a un cambio en Y de $(100 * \beta_1)\%$.

4. En el modelo Log-Log ambas variables están en escala logarítmica por lo que se interpreta como un incremento del 1% en X es asociado a un cambio en Y de $\beta_1\%$.

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

MVDB

Contexto del Estudio

Para el diseño de una política pública orientada a reducir el sedentarismo en Chile, nos interesa comprender la relevancia de ciertas variables sobre la probabilidad de que una persona sea sedentaria. Para realizar esto utilizaremos la base de datos `ISSP_2011_Salud.dta` que corresponde al módulo de Salud de la International Social Survey Programme para el año 2011. La ISSP es un programa de colaboración internacional que realiza encuestas anuales sobre diversos temas relevantes para las ciencias sociales (desigualdad social, salud, medio ambiente, religión, etc).

Dada la disponibilidad de información, diseñaremos nuestro propio indicador binario de sedentarismo utilizando la variable `V57` de la base de datos:

1. Nunca; 2. Una vez al mes o menos frecuentemente; 3. Varias veces al mes; 4. Varias veces a la semana; 5. Todos los días; 8. No sabe; 9. No contesta

Para esto debemos realizar las siguientes actividades:

a. A partir de la variable `V57` de su base de datos genere un indicador binario que señale 1 si la persona es sedentaria y 0 si es que no lo es. Justifique su análisis, genere la nueva variable y valide su recodificación de las categorias a partir de la variable original.

b. ¿Cuál es la proporción de personas sedentarias para el caso chileno?

Según la literatura hay una serie de factores que podrían estar asociados a la probabilidad de sedentarismmo en Chile. Entre ellos se mencionan como los más relevantes el índice de masa corporal, el género, la edad, la educación, si es fumador y si esta casado/a. Con esta información responda:

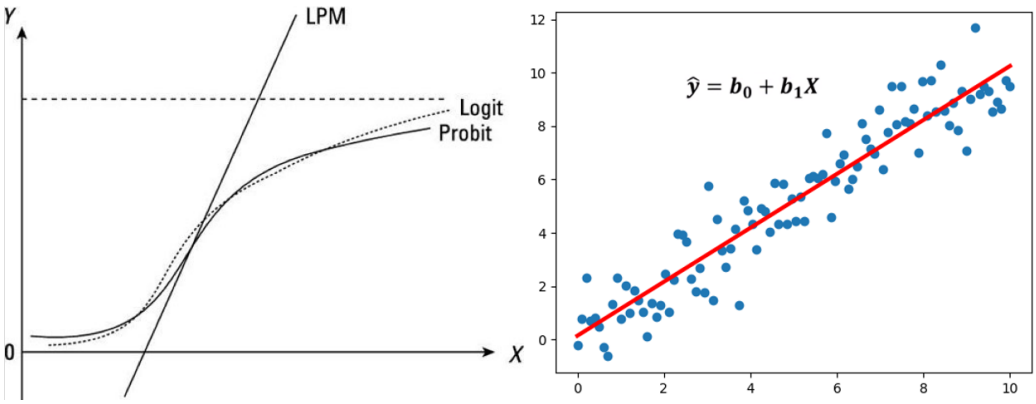
c. Explore su base de datos, genere las recodificaciones necesarias para las variables que capturen las dimensiones indicadas en el enunciado y valide sus resultados.

d. Seleccione solo las variables generadas y limpie su base de datos de casos perdidos. ¿Cuántas observaciones se perdieron de la base original? ¿Cuáles son las dimensiones finales de su base de datos?

Modelos de regresión para variables dependientes binarias

Diversos fenómenos sociales son difíciles de dimensionar a través de una escala cuantitativa. Medir el índice de masa corporal, el peso o la altura son cuestiones sencillas de ubicar dentro de una escala, sin embargo, actitudes, percepciones, eventos o fenómenos de interés son medidos con indicadores cualitativos.

Una especificación binaria para la variable dependiente nos permite estudiar la probabilidad de que un fenómeno pueda estar presente (éxito) o ausente (fracaso) en una realidad específica. Así por ejemplo, podemos estudiar la probabilidad de que una persona trabaje o no trabaje, sea pobre o no, tenga cancer o no, etc. **El modelamiento con datos binarios busca estimmar la probabilidad de éxito (ocurrencia) o fracaso (no ocurrencia) de un evento como una función de un conjunto de variables independientes**.



Nota: Para los valores promedio los tres modelos llegan a resultados parecidos, pero cuando cambian esos valores es necesario realizar ajustes.

a. Modelo de probabilidad lineal (MLP)

Antes de masificarse los modelos logit y probit, se empleba con extensión el modelo de probabilidad lineal. Este modelo es representado como:

$$p_i = \beta_0 + \sum_{k=1}^{K-1} \beta_k X_i k + \epsilon_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

Donde: $\epsilon_i = p_i - \hat{p}_i$, $E(\epsilon) = 0$, $var(\epsilon) = \sigma_e^2$ y la estimación de los parámetros del modelo se realiza vía mínimos cuadrados (con el supuesto de normalidad). El β_k corresponde al cambio en la probabilidad de éxito ($Y_i = 1$) cuando X_k aumenta en una unidad dejando el resto de las variables constantes.

a. Estime, reporte e interprete el siguiente modelo de probabilidad lineal:

$$p(Sedentarismo)_i = \beta_0 + \beta_1 IMC$$

b. Estime, reporte e interprete el siguiente modelo de probabilidad lineal:

$$p(Sedentarismo)_i = \beta_0 + \beta_1 IMC + \beta_2 Mujer + \beta_3 Etario + \beta_4 Educ + \beta_5 Fumador + \beta_6 Casado$$

Beneficios del MLP:

- No necesitamos aprender otro método de estimación.
- Interpretación sencilla de los coeficientes.
- Fácil de estimar y poco demandante en términos computacionales.
- Primera aproximación a modelos más sofisticados.

Problemas del MLP:

- Probabilidades predichas pueden caer fuera del intervalo [0,1], particularmente cuando $\hat{p}$ adopta valores extremos.
- El modelo puede presentar problemas de eficiencia en sus estimadores por heterocedasticidad.
- Los errores no presentaran una distribución normal.

b. Modelos probit binario (MPB)

En la práctica, no todos los modelos cumplen los supuestos necesarios como la normalidad e independencia de los errores (test de Durbin-Watson). En los casos en que los datos no sigan una distribución normal, es posible ajustar un modelo lineal generalizado. En términos simples se busca una función para el valor esperado de Y_i, tal que se pueda retornar a un modelo lineal clásico. Dicha función es conocida como la **función de enlace**. En particular el modelo que tiene una variable dependiente con distribución Bernoulli y que utiliza una función de enlace probit se denomina modelo probit binario (no lineal):

$$p_i = \Phi(\beta_0 + \sum_{k=1}^{K-1} \beta_k X_i k + \epsilon_i) = \Phi(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i)$$

Donde $\Phi(...)$ corresponde a la función de distribución normal acumulativa. Los $\beta' s$ (efectos marginales) dependen de los valores de las otras variables explicativas del modelo. En definitiva se observará un efecto marginal para cada individuo de la muestra.

a. Estime, reporte e interprete el siguiente modelo probit:

$$p(Sedentarismo)_i = \Phi(\beta_0 + \beta_1 IMC + \beta_2 Mujer + \beta_3 Etario + \beta_4 Educ + \beta_5 Fumador + \beta_6 Casado)$$

Problemas del MPB

- A *priori* no tenemos un valor directo del efecto para cada β del modelo. En otras palabras, la magnitud del β no nos dice nada sustantivo respecto al valor de nuestra variable dependiente. La dirección del efecto solo nos dice si dada cierta covariable dismminuye o aumenta la probabilidad de que ocurra el fenómeno observado.

Solución: Estimar las probabilidades predichas de $p_i = 1$ dado ciertos valores de las covariables (generalmente el promedio o valores de interés). Estimar el valor z y buscar en la tabla de distribución normal.

- El modelo probit involucra evaluar una integral por lo que es matemáticamente más complejo de estimar y demanda mayor potencia computacional.

c. Modelos logísticos binarios (MLB)

Un modelo que tiene una variable dependiente con distribución Bernoulli y que utiliza una función de enlace `logit` se denomina modelo logístico (no lineal). El modelo logit se utiliza para estimar la probabilidad de ocurrencia de p_i y se define como:

$$log(\frac{p_i}{1-p_i}) = \frac{1}{1-exp(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_k X_{ki}+\epsilon_i)}$$

Los β's del modelo logístico no tienen una interpretación, ya que corresponden al logaritmo de la razón de odds y por tanto, el $exp(\beta)$ puede interpretarse como la razón de odds. **¿Pero qué es una odds?**

Se define una odds o chance como la probabilidad de ocurrencia versus la probabilidad de no ocurrencia de un evento. Formalmente:

$$odds_i = \frac{p_i}{1-p_i}$$

Por ejemplo, imagine que la probabilidad de ser sedentario con más de 60 años está dada por la razón 4/5. Entonces:

$$odds_i = \frac{\frac{4}{5}}{1-\frac{4}{5}} = \frac{0.8}{0.2} = 4$$

Si las $odds > 1$, entonces obtenemos el "número de veces más probable de que ocurra el evento". En cambio, si las $odds < 1$ entonces se toma $1-odds$ y se interpreta como "x menos probable de que ocurra el evento". En este casos las odds o chances de ser sedentario versus no serlo es 4 veces más para una persona mayor de 60 años.

Ahora, nos interesa comparar las odds de dos grupos y para eso obtenemos la **razón de odds** (*odds ratio*) que equivale a:

$$\theta = \frac{odds_1}{odds_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

En el contexto de regresión logística, la razón de odds cuantifica cuánto más probable es la ocurrencia del evento ante el cambio en una unidad en la variable X_i. En definitiva, será la medida de asociación y tamaño del efecto entre una variable independiente y una variable dependiente binaria. Entonces:

- Si $\beta_i > 0$, entonces $\theta_i > 1$ y por tanto X_i será un factor que aumente la probabilidad de ocurrencia del evento.
- Si $\beta_i = 0$, entonces $\theta_i = 1$ y por tanto X_i no será un factor determinante sobre la probabilidad de ocurrencia.
- Si $\beta_i < 0$, entonces $\theta_i < 1$ y por tanto X_i será un factor que disminuya la probabilidad de ocurrencia del evento.

A continuación realice las siguientes tareas:

a. Estime, reporte e interprete el siguiente modelo probit:

$$log(\frac{Sedentario_i}{1-sedentario_i}) = \frac{1}{1-exp(\beta_0+\beta_1 IMC+\beta_2 Mujer+\beta_3 Etario+\beta_4 Educ+\beta_5 Fumador+\beta_6 Casado)}$$

b. Interprete el β_4. ¿Qué debe hacer antes de interpretar el β del modelo?

c. Evalúe le capacidad predictiva del modelo.

d. Compare las probabilidades predichas para los tres modelos a partir de las siguientes características:

- IMC=25
- Mujer
- Menor de 35 años, entre 35 y 60 años y mayor de 60 años
- Educación promedio
- No fumador
- No casado

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

III Experimentos Aleatorios

a. Cuantificando el efecto de un programa

¿Cuál es el efecto de un programa de estudio asistido en outputs académicos?

Esta pregunta se basan en los efectos del **cambio en una variable *X*** (*treatment*: seleccionado para obtener el beneficio), **sobre una variable *Y*** (*outcome*: resultado en una evaluación estandarizada). Para dar respuestas precisas, debemos cuantificar estas variaciones.

La siguiente ecuación define una línea recta que representa la relación entre un programa de formación asistido por tecnología en resultados educativos. En otras palabras, nos indica la variación de estas *variables de resultado* (o outcome) a partir de un tratamiento específico.

Como ya hemos visto en clases, la recta se puede escribir como:

$$Outcome = E(Y|X) = Y = \beta_0 + \beta_1 MindSpark + \mu$$

1. Identifiquemos cada uno de los términos de la ecuación

$$Y,\ \beta_0,\ \beta_1,\ \mu$$

2. ¿Qué podemos determinar si conocemos el valor de β_0 y β_1?

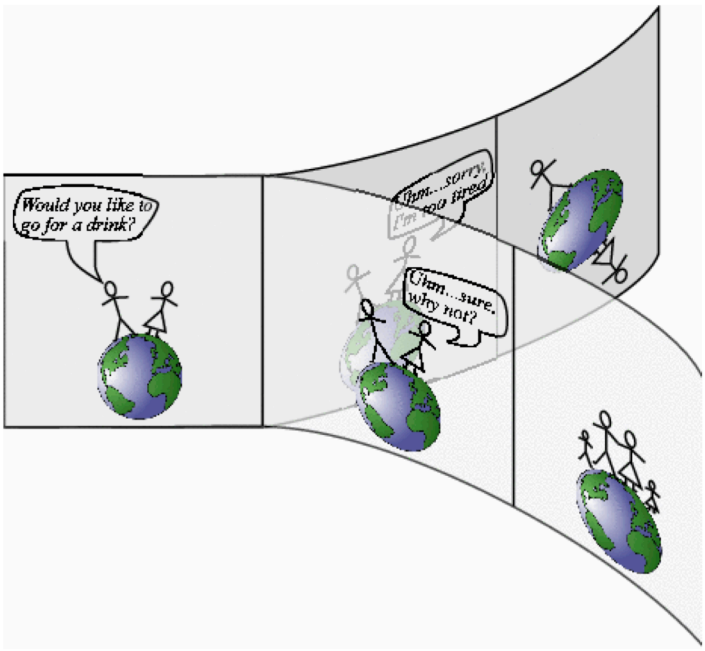
Podríamos determinar la variación asociada en resultados académicos a partir de la participación en el programa. En otras palabras, identificar la manera en que los resultados académicos cambian a medida de que se dispone de programa de apoyo educacional asistido por tecnología o no.

3. ¿Podríamos predecir algún valor?

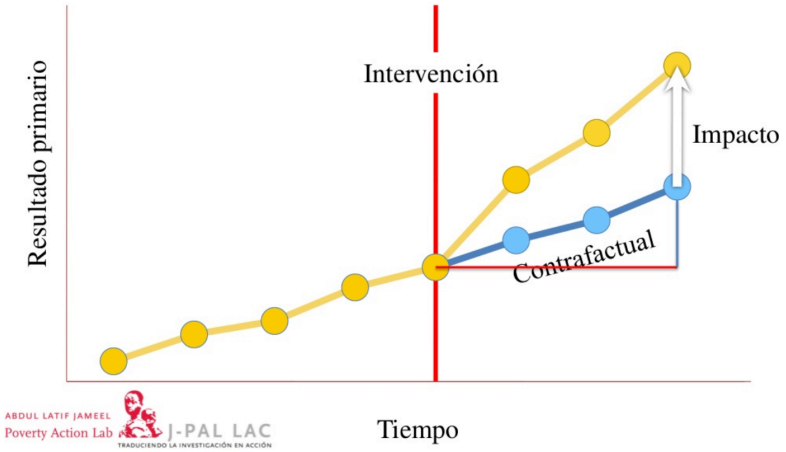
Podríamos predecir los resultados académicos para quienes reciben y no reciben el tratamiento.

4. ¿Podría ser esto un efecto causal?

Problema del contrafactual: La imposibilidad de observar una variable de resultado (Y_i) en la misma unidad y al mismo tiempo para dos condiciones diferentes.



Podríamos verlo graficamente[1]:



No podemos estimar el efecto neto de un tratamiento: $Y_i(t) - Y_i(c) = \tau_i$ al mismo tiempo en la misma unidad. Por lo tanto, es imposible observar el efecto del tratamiento.

¿Qué podemos hacer?

[1]Figura disponible en la siguiente presentación: https://www.povertyactionlab.org/sites/default/files/documents/3.El%20dise%C3%B1o%20experimental_%20Martin%20Valdivia.pdf

Estimar un **efecto causal promedio**[2]:

$$E(Y|t) = E(Y_i(t=1) - Y_i(t=0)) = \tau_{ate}$$

Sabemos que $Yi(t=1)$ es el valor en la variable resultado si el sujeto i fue tratado. $Yi(t=0)$ es el valor si el sujeto i no fue tratado.

Para obtener el efecto del tratamiento en un diseño experimental aleatorizado debemos considerar los siguientes supuestos:

- **Independencia**: la asignación al tratamiento es independiente de la variable resultado para tratados $(Y_i(t))$ y controles $(Y_i(c))$, y de cualquier otra variable en la población (en este caso el análisis de balance en la próxima sección nos ayuda a dar cuenta de este supuesto), formalmente queda expresado de la siguiente manera:

$$Y_i(t), Y_i(c)) \perp D_i$$

En consecuencia la asignación al grupo de tratamiento o controlno incide sobre el valor esperado. Por lo tanto, la asignación aleatoria es condición de posibilidad de la independencia y se deduce que, en **los promedios**, la única diferencia entre $(D_t$ y $(D_c$ corresponde a que un grupo fue tratado y el otro no. En definitiva, se puede estimar el ATE como el efecto promedio del tratamiento para el grupo tratamiento y en el grupo de control:

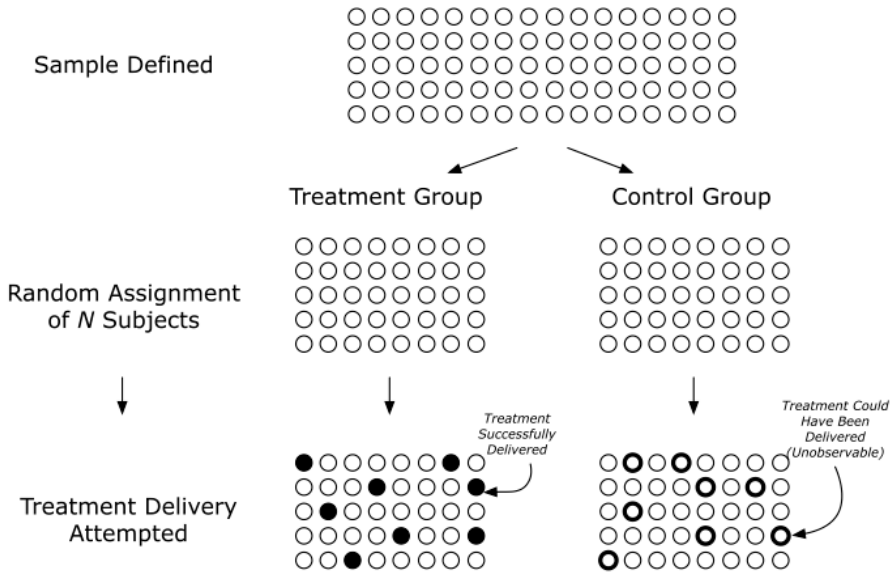
$$\frac{1}{N}\sum_{i=1}^{N} Y_i(t) - \frac{1}{N}\sum_{i=1}^{N} Y_i(c)$$

- **Restricción de exclusión**: los resultados del experimento dependen únicamente de la exposición al tratamiento $(D_i = 1)$ y no de la asignación entre grupo de tratamiento y control (quien recibe o no el beneficio). En términos formales nos permite identificar el efecto causal. En definitiva se espera que todos los individuos asignados al tratamiento sean expuestos al estímulo, a su vez, los asignados al grupo de control no son expuestos al estímulo:

$$Y_i(1, d) = Y_i(0, d)$$

- **SUTVA**: no hay interferencia (*spillovers*) en la respuesta al tratamiento de un individuo en comparación a otro. En otras palabras, si la observación i se expone al tratamiento t, el valor de la variable resultado Y se mantendrá igual sin importar el tipo de asignación ni el tipo de tratamiento que reciban las otras observaciones i (sería problemático que la asignación del seguro a un individuo afectara el resultado de salud en otro individuo). Esto nos garantiza de que **solo existan dos respuestas** ante la condición de tratamiento (asignación) y de tratado, independientes de las otras unidades. Por lo tanto, podemos estimar la causalidad como al diferencia en la variable resultado para el grupo de tratamiento y el grupo de control.

[2]Sobre el diseño tradicional en Broockman, David E. and Kalla, Joshua and Sekhon, Jasjeet S., The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs (September 20, 2017). Available at SSRN: https://ssrn.com/abstract=2742869 or http://dx.doi.org/10.2139/ssrn.2742869

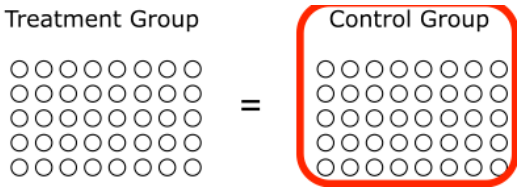


Podemos obtener este resultado a partir de una regresión lineal con una variable independiente binaria:

$$E(Y|t) = \beta_0 + \beta_1 MindSpark + \mu$$

b. Validez del diseño experimental

La **clave** de este diseño experimental es la **asignación aleatoria** de los participantes al tratamiento (de manera natural). En este caso la asignación al tratamiento (T_i) es independiente de la variable de resultado para tratados ($Y_i(t)$), controles ($Y_i(c)$) y de cualquier otra variable. Por lo tanto, la asignación aleatoria garantiza que en los **promedios**, la única diferencia entre tratados y controles es que un grupo fue tratado y el otro no, por lo que el error (μ) será aleatorio.



c. Pruebas de balance

Las pruebas de **balance** o verificaciones de aleatorización las utilizamos para evaluar que el efecto identificado no se deba a otras variables más que al tratamiento. De esta manera, el balance permite asegurar que la asignación aleatoria al tratamiento está equilibrada con respecto a estas variables y que, por lo tanto, la diferencia observada entre los tratados y controles no se deba a estos posibles factores de confusión (o variables omitidas). **Si las diferencias entre tratados y controles son significativas, habría que recurrir a otros análisis.**

¿Cómo podríamos evaluar si, en este set de covariables, los grupos están balanceados?