

Tarea 1 y 2

Fecha de entrega: Miércoles 11 de mayo (11:59 pm).

Nota: Deberán subir a *Canvas* dos archivos. Un **primer archivo** de texto con sus respuestas. Este archivo debe ser auto-suficiente. Esto quiere decir que con revisar este archivo debe ser posible calificar su tarea en su totalidad. Para este archivo pueden utilizar el formato de su preferencia (e.g. pdf, L^AT_EX, Word u hojas escritas a mano y escaneadas). El **segundo archivo** es un archivo de soporte que genere todos sus resultados. Típicamente este será un archivo de programación, mismo que al ejecutarse replicaría los resultados que estás reportando en tu primer archivo. Dado que ustedes son libres de elegir el software que utilizarán, este puede ser un R-script, Do-File o similar.

The Economic Consequences of Increasing Sleep Among the Urban Poor

Conscientes de que las áreas urbanas pobres en los países en vías de desarrollo presentan ambientes complicados que pueden interferir con un buen periodo de sueño, Pedro Bessone, Gautam Rao y Frank Schilbach (2021) diseñaron un experimento aleatorizado en la ciudad de Chennai, India, con el fin de investigar los efectos que tiene un periodo más prolongado de descanso sobre la productividad, el ingreso, los aspectos cognitivos, la toma de decisiones, el bienestar y la oferta de trabajo.

Para su experimento, los autores reunieron a 452 adultos de ingresos bajos de Chennai y estudiaron su desempeño laboral y de descanso por 28 días. El desempeño laboral se cuantificaba a partir de la realización de una actividad por parte de los participantes durante el periodo de estudio. El estado anímico se medía con una serie de encuestas y el desempeño de descanso se medía con actígrafos que los participantes utilizaban.

Durante los primeros ocho días del estudio todos los participantes estuvieron en una condición de control con el fin de reunir información de referencia. Con este grupo de individuos se llevaron a cabo dos experimentos de manera independiente, como pueden consultar en el paper. El primer experimento **no será de nuestro interés. Es decir, para propósitos de la tarea podrían simplemente ignorarlo.**¹ El segundo experimento buscaba medir el impacto de las siestas. Este tratamiento empezó el noveno día del estudio y consistió en que a los individuos de tratamiento se les asignaba un espacio individual

¹Dicho experimento consistió en separar a los participantes en tres grupos. Al primer grupo se le dio información de la importancia del sueño e insumos para mejorar su calidad de sueño (e.g. ventiladores, colchones y almohadas). Al segundo grupo se lo mismo que al primer grupo más un incentivo económico: por cada minuto de incremento en sueño registrado por el actígrafo con respecto a sus datos de referencia, se le daba una rupia india. El tercer grupo era el de control

en las oficinas del estudio para tomar una siesta de 30 minutos. Aquellos que no desearan dormir tenían que permanecer en silencio en el espacio designado para las siestas. Este espacio tenía camas, sábanas, almohadas, tapones de oídos y cubreojos disponibles. Los individuos de control cada día eran aleatorizados a que durante los 30 minutos de siesta del grupo de tratamientos trabajaran o tuvieran tiempo de descanso donde eran libres de decidir qué hacer (excepto dormir o trabajar).

Es importante aclarar que, para el tratamiento de siesta, los autores decidieron estratificar a los participantes de acuerdo con sus horas de sueño y ganancias en los primeros siete días.

Al final de este documento encontrarán un diccionario de las variables disponibles en las bases de datos (Tabla 1). En *Canvas* podrás descargar las tres bases de datos que necesitarás para llevar a cabo esta tarea:

- *Baseline*: contiene el promedio para cada variable de los 452 participantes durante los primeros siete días; deberás considerar esta base como los datos de referencia.
- *Endline*: contiene el promedio para cada variable de los 452 participantes durante los últimos siete días; deberás considerar esta base para analizar los posibles impactos que puede haber tenido el tratamiento.
- *Completa*: es la base de datos panel que contiene la información de los 452 individuos durante los 28 días que duró el estudio. La variable *day_in_study* está disponible solo en esta base e indica a qué día corresponde cada observación (entre 1 y 28).

A lo largo de la tarea y para las preguntas que lo requieran, será importante que reportes tus procedimientos, las pruebas de hipótesis que realices, tus resultados y tus gráficas; estas últimas deberán tener título y nombre en los ejes.

1. RCTs

En la primera parte de la tarea analizaremos la implementación del experimento de siestas, indicado con la variable T_{nap} . Para esta parte solo utilizaremos las bases de *Baseline* y *Endline*.

1. **Balance.** Empezaremos analizando el balance de los datos. Pese a que el diseño del experimento utilizó estratificación por hora de sueño y ganancias, para esta pregunta supondremos que tal división no existió. Utiliza la base de *Baseline* y elige, utilizando tu criterio, al menos 10 variables para evaluar el balance en la asignación del tratamiento, T_{nap} . Crea una tabla de balance donde evalúes el balance de las variables de forma individual y conjunta. Reporta claramente tus resultados en una tabla. Haz una muy breve descripción de lo que obtienes y da una conclusión.
2. **Efectos de tratamiento.** En esta parte utilizarás solo la base *Endline*. Asumiendo que la distribución de tratamiento (T_{nap}) sí fue aleatoria, realiza los siguientes incisos: [Nota: para cada inciso reporta el p-value, lo necesitarás más adelante]
 - a) Estima los efectos del tratamiento (*Average Treatment Effect*) utilizando el estimador de Neyman.
 - b) Estima los efectos del tratamiento (*Average Treatment Effect*) utilizando una estimación OLS con errores heterocedásticos sin controles.
 - c) Estima los efectos del tratamiento (*Average Treatment Effect*) utilizando una estimación OLS agregando controles. Tú deberás decidir qué controles utilizar y justificar cómo y por qué los elegiste.
 - d) Reporta los resultados de los tres incisos anteriores en una tabla con los formatos de regresión que empleamos usualmente. Escribe una breve conclusión de lo que obtuviste.
 - e) De los tres incisos, elige tu estimador favorito y sigue dicho método para estimar los ATE para las variables nap_time_mins , $sleep_report$, $happy$, $cognitive$ ² y $typing_time_hr$. Reporta tus resultados con formato de tabla e interpreta los resultados obtenidos. Indica claramente si los efectos encontrados son pequeños o grandes.
3. **Fischer Exact Test.** Definamos la hipótesis nula como: el tratamiento de las siestas no tiene efecto alguno sobre la productividad para las personas.
 - a) Utiliza el *Fischer Exact Test* para evaluar H_0 , y reportar el p-value. Emplea 1,000 simulaciones.

²Tú deberás generar una variable que agrupe las variables cognitivas y explicar cómo lo hiciste.

- b) Contrasta el p-value que obtuviste con los p-values de las tres estimaciones que realizaste en la segunda pregunta. ¿Qué conclusiones puedes obtener?
4. **Estratificación.** En el diseño del RCT los autores estratificaron utilizando dos variables: ganancias y horas de sueño. Para cada variable, los autores calcularon la mediana y clasificaron a los individuos como aquellos por arriba o debajo de la mediana. Con esta clasificación formaron 4 grupos y aleatorizaron la asignación de T_{nap} al interior de cada grupo.
- a) Crea los 4 grupos y reporta el número de observaciones en tratamiento y control para cada grupo. ¿Tiene esta tabla los valores que hubieras esperado ex-ante?
 - b) Utilizando un estimador de Neyman, estima el *Average Treatment Effect* (ATE) para cada estrato y el ATE agregado. Utiliza como variables dependientes las que señalamos en el último inciso de la pregunta 2.
 - c) Repite el inciso anterior, pero estimando el efecto heterogéneo para los 4 grupos utilizando una regresión con interacciones. Reporta todos tus resultados (del inciso anterior y este) en una misma tabla. Compara tus resultados en términos de valores estimados y errores estándar. Escribe un breve párrafo al respecto.
5. **Atrición.** Una situación a la que se pueden enfrentar diversos autores de experimentos aleatorizados es la pérdida de observaciones. Para los siguientes incisos supondremos tal situación en el estudio de Bessone, Rao y Schilbach. Con este propósito, generamos una variable en las bases de datos, llamada *drop_indicator*.³ Los participantes que consideraremos no terminaron el estudio tendrán un valor de *drop_indicator* igual a 1.
- a) Utilizando la variable *drop_indicator* reporta cuántas observaciones de control y tratamiento de T_{nap} se perdieron. Reporta dicha pérdida en términos porcentuales también.
 - b) Utilizando las mismas variables que empleaste para la tabla de balance de la primera pregunta, haz el test adecuado para evaluar si la atrición afecta: (i) la validez interna y (ii) la validez externa.
 - c) ¿Cuál es la conclusión que puedes dar a partir de las tablas que realizaste en el inciso anterior?
6. **Lee Bounds.**
- a) Reporta qué porcentaje de la muestra son *always respondents*, *never respondents* y *selective respondents*. Describe en este contexto qué quiere decir el supuesto de monotonicidad y si hace sentido de forma intuitiva.

³Esta variable no existe en la base de datos original y sólo fue producida con los propósitos de esta tarea.

- b)* Estima los *Lee Bounds* para el efecto de tratamiento de productividad asumiendo que el supuesto de monotonicidad es válido.
- c)* ¿Cómo se compara este resultado con respecto al que obtuviste en los primeros tres incisos de la pregunta (2)? ¿Debería o no estar el intervalo de los Lee Bounds centrado en los valores estimados en la pregunta 2?

2. Matching

En esta sección seguiremos trabajando sólomente con las bases de datos de *Baseline* y *Endline*. En la sección anterior descubrimos que, como resultado de la atrición, se pierde el balance. En esta sección asumiremos que las observaciones de atrición (i.e. los individuos con *drop_indicator* igual a 1 se pierden). Utilizando las observaciones restantes, analizaremos si *matching* nos permite regresar a nuestro estimador del RCT.

1. **Coarsened Exact Match.** Utilizando las variables *female* y *education* lleva a cabo una estimación del ATE. Reporta si, como resultado del procedimiento CEM, existieron pérdidas de observaciones y cuántas fueron (en caso de que hayan).
2. **Nearest Neighbor.** Utilizando este método estima el TOT utilizando las variables *female*, *age* y *sleep-report* (de *Baseline*). Emplea la distancia de *Mahalanobis* y compara el resultado de utilizar uno, cinco y diez nearest neighbors.
3. **PSM.** Por último, en esta pregunta utilizaremos el PSM para estimar el TOT.
 - a) Para empezar, estima el propensity score utilizando cinco de las variables que empleamos para la tabla de balance de la primera pregunta (las que tu quieras). Contrasta el resultado de obtener el PSM con los métodos: (i) MPL, (ii) probit y (iii) logit. Reporta el resultado de las tres estimaciones en una tabla de regresiones con el formato típico.
 - b) Utilizando los tres métodos de la pregunta anterior, reporta en una tabla por separado el valor promedio y la desviación estándar del PSM y el reponderador para el grupo de tratamiento y el grupo de control de *T_nap*.
 - c) Utilizando el estimador de *probit* para el PSM, calcula el reponderador y con este ponderador vuelve a producir la tabla de balance con las diez variables que utilizaste en la pregunta 1 de RCTs; es decir, incluirás las cinco variables que usaste para el PSM y las que dejaste fuera. ¿Qué concluyes de este ejercicio?
 - d) Independientemente del resultado de la pregunta anterior, utilizando el estimador de *probit* para el PSM estima el TOT para productividad y las cinco variables dependientes del último inciso de la pregunta 2. Reporta tus resultados en tablas y contrasta los resultados con lo obtenido en los incisos anteriores.

3. Diferencias en Diferencias

En esta última sección de la tarea utilizaremos las bases de *Baseline*, *Endline* y *Completa*. En particular, la base de datos *Completa* es un panel donde tenemos observado para los individuos los días del estudio por separado.

1. **2x2.** Utilizando todas las observaciones, es decir sin considerar el problema de atrición, estima el TOT con los datos de *Baseline* y *Endline* para formar una tabla de 2x2 con la variable productividad. Produce dicha tabla e indica con qué estimación de OLS podrías producir esta tabla.
2. **2x2 parteB.** Repite el ejercicio anterior, pero ahora forma *Baseline* y *Endline* partiendo de la base *Completa*. De hecho, no es necesario que generes dos nuevas bases de datos, sino simplemente que uses los datos de *Completa* para producir tu tabla de 2x2 donde distingas tratamiento y control con *T_nap*, y antes vs después con la variable *day_in_study*. Genera el resultado con OLS también y, por último, agrega controles para hacer tu estimación. ¿Qué tipo de errores consideras que deberías estar empleando? y ¿por qué?
3. **Comparación.** ¿Cómo se compara este resultado con el de los primeros tres incisos de la pregunta (2)? Haz un breve párrafo descriptivo donde compares los resultados.
4. **Panel.** Reproduce los resultados de las Figura A.V y A.VI. En ambos casos solo deberás hacer las Figuras para el experimento de Nap.
5. **Tendencia paralela.** Ahora imagina que quisieras utilizar DiD para corregir el problema de atrición. Evalúa si hay evidencia a favor o en contra de que se cumpla el supuesto de tendencia paralela una vez que restringimos nuestras observaciones a aquellas después de la atrición (i.e. eliminando los individuos con *drop_indicator=1*). Indica claramente la especificación que estás empleando y la prueba de hipótesis a evaluar.
6. **Matching y DiD.** Ahora combinaremos los métodos anteriores.
 - a) Independientemente del resultado del inciso anterior, propón cómo podrías emplear *matching* combinado con DiD para resolver el problema de la tendencia paralela.
 - b) Usando la solución que propusiste en el inciso anterior, estima el TOT a lo largo del tiempo y el TOT de forma general utilizando el estimador de *DiD* y *matching*. ¿Cómo se compara tu resultado con los tres incisos de la pregunta 2?
 - c) Reproduce el resultado anterior con las cinco variables del último inciso de la pregunta 2. Discute el resultado obtenido.

Cuadro 1: Descripción de variables.

Variable	Descripción
<i>pid</i>	Participant ID
<i>time_in_office</i>	Tiempo en la oficina (horas)
<i>absences_baseline</i>	Ausencias durante los primeros siete días
<i>age</i>	Edad del participante
<i>female</i>	1 = female
<i>education_</i>	Años de educación
<i>treatment_group</i>	0 = Control, 1 = Sleepaid Treatment Group o 2 = Sleepaid & Incentive Treatment Group
<i>sleep_night</i>	Tiempo total de horas dormidas (horas)
<i>nap_time_mins</i>	Minutos dormidos durante la siesta
<i>no_of_children_</i>	Número de hijos del participante
<i>T_nap</i>	0 = Nap Control Group, 1 = Nap Treatment Group
<i>act_inbed</i>	Total de tiempo que el participante pasó en la cama (horas)
<i>an_12_number_of_awakenings</i>	Número de veces que la participante se despertó durante la noche
<i>an_13_average_awakening_length</i>	Promedio de tiempo interrumpido de sueño (minutos)
<i>typing_time_hr</i>	Horas de trabajo
<i>earnings</i>	Ganancias del día por el trabajo de recolección de datos
<i>productivity</i>	Productividad registrada por hora durante el trabajo de recopilación de datos
<i>tot_earnings</i>	Ganancias totales por día
<i>deposit_today_amount</i>	Monto de dinero depositado en lockers
<i>withdraw_today_amount</i>	Monto de dinero extraído de lockers
<i>unemployed</i>	Status actual de empleo: 0 = Desempleado, 1 = Trabajo temporal, 2 = Empleado de tiempo parcial, 3 = Empleado de tiempo completo, 4 = Ama de casa y empleada a tiempo parcial o completo o 5 = Solo ama de casa
<i>c27</i>	Número de siestas que la participante toma durante el día
<i>c28_h</i>	Horas promedio que el participante dedica a siestas durante el día
<i>c28_m</i>	Minutos promedio que el participante dedica a siestas durante el día

Variable	Descripción
<i>d1</i>	1 = el participante consume tabaco o Buyo
<i>prior_savings</i>	Ahorros de la participante
<i>go_to_bed</i>	1 = El participante se levantó en medio de la noche y permaneció despierto o realizó otra actividad
<i>awake_self_report</i>	Minutos que la participante reportó haber permanecido despierta en la noche
<i>out_of_bed</i>	Minutos que el participante estuvo fuera de la cama durante la noche
<i>happy</i>	Dummy =1 si el participante reportó ese día estar feliz o muy feliz. Las alternativas eran estar nada feliz, un poco infeliz o algo feliz
<i>ds_g1_satisfaction</i>	1 = El participante no está satisfecho con su vida, 10 = El participante está satisfecho con su vida
<i>energy</i>	Dummy =1 si el participante reportó ese día tener energía todo el tiempo, la mayor parte del tiempo o buena parte del tiempo. La alternativa era tener energía solo durante momentos, pocos momentos o en ningún momento.
<i>stress</i>	Dummy =1 si el participante reportó ese día estar estresado algunas veces, con frecuencia moderada o con mucha frecuencia. La alternativa era estar estresado nunca o casi nunca.
<i>sleep_eff</i>	Eficiencia de sueño
<i>sleep_report</i>	Horas registradas de sueño por el actígrafo
<i>daily_savings</i>	Monto que se dedicó a ahorros por día
<i>pay_hf</i>	Pago derivado del desempeño cognitivo en <i>Hearts and Flowers</i>
<i>pay_corsi</i>	Pago derivado del desempeño cognitivo en <i>Corsi Block Span</i>
<i>pay_pvt</i>	Pago derivado del desempeño cognitivo en <i>Psychomotor Vigilance Task</i>
<i>corsi_measure</i>	Desempeño numérico en la actividad cognitiva <i>Corsi Block Span</i>
<i>hf_measure</i>	Desempeño numérico en la actividad cognitiva <i>Hearts and Flowers</i>
<i>pvt_measure</i>	Desempeño numérico en la actividad cognitiva <i>Psychomotor Vigilance Task</i>
<i>treat_s</i>	1 = la participante pertenece al primer grupo de tratamiento (devices & encouragement)

Variable	Descripción
<i>treat_s_i</i>	1 = el participante pertenece al segundo grupo de tratamiento (devices & incentives)
<i>health_bsl</i>	índice de salud de los primeros siete días
<i>drop_indicator</i>	Dummy que indica si en el ejercicio con atrición asumiremos que estos datos se perdieron por atrición