Problem Set 2:
Multiple hypothesis testing

Economía Aplicada

Universidad de
SanAndrés

Alumnos:
Juan Diego Barnes
Fransisco Legaspe
Rodrigo Martin
Diego Fasan

Profesor: Martín Rossi
Tutores: Gastón García Zavaleta
Tomás Pacheco

1. Multiple hypothesis testing

1. Comenzamos replicando la tabla del paper de Attanasio et al. (2020), los resultados de las regresiones se presentan en la tabla 1.[1]

Tabla 1: Treatment Effects on Various Outcomes

	Treatment Effect		
	Point Estimate	**SE**	**Sample Size**
Panel A. Child's Cognitive Skills at Follow-up			
Bayley: Cognitive	0.250	(0.063)	1263
Bayley: Receptive Language	0.174	(0.063)	1,262
Bayley: Expressive Language	0.029	(0.062)	1,261
Bayley: Fine Motor	0.073	(0.059)	1,261
MacArthur: Words	0.086	(0.064)	1,321
MacArthur: Complex Phrases	0.057	(0.056)	1,321
Panel B. Child's Socio-Emotional Skills at Follow-up			
ICQ: Difficult (-)	0.073	(0.045)	1,325
ICQ: Unsociable (-)	0.037	(0.055)	1,325
ICQ: Unstoppable (-)	0.031	(0.054)	1,325
ECBQ: Inhibitory Control	0.000	(0.058)	1,322
ECBQ: Attentional Focusing	0.072	(0.048)	1,322
Panel C. Material Investments at Follow-up			
FCI: Play Materials	0.217	(0.064)	1,325
FCI: Coloring and Drawing Books	-0.125	(0.056)	1,325
FCI: Toys for Movement	-0.049	(0.065)	1,325
FCI: Toys for Shapes	0.426	(0.088)	1,325
FCI: Shop-Bought Toys	0.019	(0.061)	1,325
Panel D. Time Investments at Follow-up			
FCI: Play Activities	0.280	(0.051)	1,325
FCI: Told Story	0.153	(0.064)	1,325
FCI: Read to Child	0.403	(0.068)	1,325
FCI: Played with Toys	0.183	(0.060)	1,325
FCI: Named Things	0.144	(0.048)	1,325

Notas: Las medidas seguidas por (-) han sido invertidas de manera que una puntuación más alta se refiere a un comportamiento mejor. Los efectos relacionados con los factores latentes están en puntos logarítmicos. Los errores estándar (entre paréntesis) fueron corregidos por clisteres a nivel municipal.

En la presente tabla podemos ver el impacto de recibir los estímulos en distintas áreas. La tabla tiene cuatro paneles, en donde se detallan los resultados de cada tipo de actividad o habilidad y sus metodologías por categorías como, habilidades cognitivas, socioemocionales, inversiones materiales y de tiempo. Resulta relevante ver que para la inversión en material no está claro como responde el total de actividades monitoreadas. En su mayoría son positivos pero algunos responden negativamente. El resto, en general, hay una respuesta en promedio mayor para el grupo tratado. En el primer panel, el desarrollo cognitivo aumenta un 0.25 para

[1] En este punto nos alejamos de la consigna, debido a que al poner los outcomes en columnas y ajustar la tabla al tamaño de la pagina esta queda ilegible, de modo que optamos por presentar los datos de la misma manera que lo realiza el paper en su tabla 2.

el grupo en tratamiento y el lenguaje receptivo un 0.17. En el panel B, no hay efectos muy relevantes. Los efectos más relevantes del panel D, son el impacto del 0.28 en actividades de juego y un 0.40 en leerle a los chicos. Todo los efectos son adicionales para el grupo en tratamiento.c

2. Cuando estamos testeando múltiples hipótesis, sea por tener múltiples *outcomes*, tratamientos o submuestras, los procedimientos estándar de inferencia no son confiables. Podríamos encontrar que un efecto es significativo, por puro azar. En conjunto las probabilidad de un error de tipo I (falso positivos) aumenta con la cantidad de realizaciones. Mas formalmente la probabilidad de falso positivo aumenta con la cantidad de hipótesis, una medida de esto es el *Family-Wise Error Rate (FWER)*, que es la probabilidad de cometer al menos un falso positivo.

$$FWER = P(h_1 = FP \vee ... \vee h_n = FP) = 1 - P(h_1 = FP \wedge ... \wedge h_n = FP)$$

$$= 1 - [P(h_1 \neq FP) \times ... \times P(h_n \neq FP)]$$

Implícitamente en la formula se asume independencia entre las hipótesis, al expresar la probabilidad conjunta como el producto de las probabilidades individuales, de modo que claramente como las probabilidades se encuentran entre 0 y 1 la probabilidad de falso positivo aumenta con el numero de hipótesis.

3. Ahora que conocemos el problema que se presenta al testear hipótesis múltiples, realizaremos 3 diferentes estrategias de corrección:

   - Bonferroni.
   - Holm.
   - Benjamini, Krieger y Yekutieli.

   En el caso del paper de Attanasio et al. (2020), se intenta estimar efecto de un único tratamiento sobre diferentes outcomes. El paper presenta estos outcomes en 4 grupos o paneles según que tan relacionados los considera. Para las correcciones, adoptamos una postura conservadora, considerando los 21 outcomes como 21 hipótesis que se quieren testear a la vez (no por grupos), basándonos en que el tratamiento es el mismo para todos los outcomes. La Tabla 2, presentan los resultados de realizar estas correcciones sobre los p-valores y/o el nivel de significatividad.

Tabla 2: Treatment Effects on Various Outcomes

	Treatment Effect			Multiple Hypothesis Correction			Sample Size
	Point Estimate	**SE**	**P-Values**	**Bonferroni** (1)	**Holm** (2)	**Benjamini et al.** (3)	
Panel A. Child's Cognitive Skills at Follow-up							
Bayley: Cognitive	0.250	(0.063)	0.00	0.00	Yes	0.00	1263
Bayley: Receptive Language	0.174	(0.063)	0.01	0.15	No	0.01	1,262
Bayley: Expressive Language	0.029	(0.062)	0.64	1.00	No	0.41	1,261
Bayley: Fine Motor	0.073	(0.059)	0.22	1.00	No	0.22	1,261
MacArthur: Words	0.086	(0.064)	0.18	1.00	No	0.18	1,321
MacArthur: Complex Phrases	0.057	(0.056)	0.32	1.00	No	0.27	1,321
Panel B. Child's Socio-Emotional Skills at Follow-up							
ICQ: Difficult (-)	0.073	(0.045)	0.11	1.00	No	0.12	1,325
ICQ: Unsociable (-)	0.037	(0.055)	0.51	1.00	No	0.37	1,325
ICQ: Unstoppable (-)	0.031	(0.054)	0.57	1.00	No	0.40	1,325
ECBQ: Inhibitory Control	0.000	(0.058)	1.00	1.00	No	0.51	1,322
ECBQ: Attentional Focusing	0.072	(0.048)	0.14	1.00	No	0.14	1,322
Panel C. Material Investments at Follow-up							
FCI: Play Materials	0.217	(0.064)	0.00	0.02	Yes	0.00	1,325
FCI: Coloring and Drawing Books	-0.125	(0.056)	0.03	0.59	No	0.04	1,325
FCI: Toys for Movement	-0.049	(0.065)	0.45	1.00	No	0.34	1,325
FCI: Toys for Shapes	0.426	(0.088)	0.00	0.00	Yes	0.00	1,325
FCI: Shop-Bought Toys	0.019	(0.061)	0.75	1.00	No	0.43	1,325
Panel D. Time Investments at Follow-up							
FCI: Play Activities	0.280	(0.051)	0.00	0.00	Yes	0.00	1,325
FCI: Told Story	0.153	(0.064)	0.02	0.40	No	0.03	1,325
FCI: Read to Child	0.403	(0.068)	0.00	0.00	Yes	0.00	1,325
FCI: Played with Toys	0.183	(0.060)	0.00	0.06	Yes	0.00	1,325
FCI: Named Things	0.144	(0.048)	0.00	0.07	No	0.01	1,325

Notas: Las medidas seguidas por (-) han sido invertidas de manera que una puntuación más alta se refiere a un comportamiento mejor. Los efectos relacionados con los factores latentes están en puntos logarítmicos. Los errores estándar (entre paréntesis) fueron corregidos por clisteres a nivel municipal. En las correcciones (1) y (3) se prestan los p-valores corregidos, en la corrección (2) se indica si la hipótesis fue rechazada al 0.05 de nivel de significatividad.

En la Tabla 2 podemos ver los valores observados en la Tabla 1 en sus dos primeras columnas y además su p-valor correspondiente. Además se estimó mediante las tres estrategias alternativas, mencionadas anteriormente, el p-valor para cada outcome.

4. Bonferroni nos ofrece una corrección simple, pero muy restrictiva. Busca corregir el *FWER*, plantenado que al estar evaluando m hipótesis, tenemos que trabajar al nivel de significatividad α/m. La estrategia desarrollada por Holm (1979) hace que la corrección sea menos restrictiva que la de Bonferroni, pero no corrige el problema de asumir independencia entre las hipótesis. La estrategia de Benjamini, Krieger and Yakutieli (2006), la mas aplicada en la actualidad, tiene como ventajas, que considera la dependencia entre los outcomes (si los outcomes son perfectamente dependientes, los pvalues corregidos son iguales a los originales), esto implica un mayor poder estadístico. Para ver en detalle como aplicarlo se pude consultar Anderson (2008). La estrategia de Bonferroni y Benjamini, Krieger and Yakutieli corrigen los p-valores, mientras que la corrección de Holm corrige el nivel de significatividad contra el que se contrastan los resultados.

5. Como podemos ver en la primera fila de resultados, el p-valor no cambia con las nuevas estrategias. En el caso de Holm, nos dice que es significativo al menos al 5 % por lo que, si bien es más laxo, no contradice a los otros p-valores. Esto es distinto en el outcomes lenguaje receptivo, donde la estimación original, devuelve un p-valor con significatividad al 1 % mientras que Bonferroni y Holm lo estiman con una confianza menor al 95 % Holm (Bonferroni aun menor). La estrategia Bonferroni nos dificulta encontrar resultados significativos en todo el panel de habilidades socioemocionles, mientras que nuestro resultado original lo podía ser en al 15 % (en dificultad y foco de atención). En general, los p-valores mayores al 0.15 los penaliza más. Con

respecto a los valores arrojados por la estrategia de Benjamini, Krieger y Yekutieli, son muy similares a los p-valores originales, incluso mejora la confianza de aceptación. Esto se da en lenguaje expresivo, frases complejas, juguetes de movimiento y otros outcomes más. Luego si comparamos Holmes, vemos que este acepta la inferencia si el modelo original tiene p-valores bajos o si Bonferroni esta por debajo de 0.07. El resto de las filas, las rechaza.

Referencias

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C., and Rubio-Codina, M. (2020). Estimating the production function for human capital: results from a randomized controlled trial in colombia. *American Economic Review*, 110(1):48–85.