



Semana 1

Regresión Lineal

DCDPP - Datos para la evaluación de Políticas Públicas | PUC | 19 de agosto, 2022

 **Pablo A. Celhay** |  pacelhay@uc.cl

Outline

1. Motivación
2. Logística del curso
3. Introducción
 - Causalidad
 - Experimentos aleatorios
4. Validez, RCTs y estudios observacionales
 - Experimentos
 - Datos Observacionales

1. Motivación

Motivación

The screenshot shows the header of the diarioUchile website. It includes the logo 'diarioUchile' with the text 'Año X, 7 de agosto de 2018', a 'MENÚ' button, a weather icon showing clouds, and the text 'Santiago, Ahora: 7°C | Min: 8°C - Max: -17°C'. To the right is the 'radio Uchile' logo. Below the header, a news article titled 'Plan Auge: una década de críticas' is displayed under the 'SALUD' category. The article discusses the 10-year anniversary of the Plan Auge and its critics. At the bottom of the article, there are links for 'NOTICIAS' and 'BLOGS Y OPINIÓN'.

The next section is from 'EL MOSTRADOR'. It features a large 'm' logo and the title 'Cómo financiar el Plan Auge: ¿a quién le pasa la cuenta el gobierno?'. Below the title, it says 'por EL MOSTRADOR | 8 mayo, 2003'. The text discusses the critique of the IVA's impact on low-income families. To the right is a small image of a person.

The final section is from 'emol. Nacional'. It shows the 'emol' logo with 'Nacional' underneath, the date 'Santiago: Martes 07 de agosto del 2018 | Actualizado 09:23', and a navigation bar with links like 'Noticias', 'Economía', 'Deportes', etc. Below the navigation bar is a search bar with the text 'Buscar' and a link 'amarillas.com'.

Por primera vez en sus diez años, plan AUGE se queda sin recursos para mejoras

Nuevo decreto, previsto para 2016, debía incluir nuevas coberturas en las 80 enfermedades ya garantizadas. Pero las correcciones previstas para esta norma, evaluadas en \$6.047 millones, no fueron incluidas en proyecto de ley de presupuesto.

Motivación

The screenshot shows the header of the diarioUchile website. It includes the logo 'diarioUchile' with the text 'Año X, 7 de agosto de 2018', a 'MENÚ' button, weather information 'Santiago, Ahora: 7°C | Min: 8°C - Max: -17°C', and a 'radio Uchile' logo. Below the header, a news article titled 'Plan Auge: una década de críticas' is displayed under the 'SALUD' category. The article discusses the 10-year anniversary of the Plan Auge and its critics. At the bottom of the page, there are links for 'NOTICIAS' and 'BLOGS Y OPINIÓN'.

La discusión del plan
AUGE/GES ha
estado centrada en
sus costos

The screenshot shows an article from 'EL MOSTRADOR' dated May 8, 2003. The title is 'Cómo financiar el Plan Auge: ¿a quién le pasa la cuenta el gobierno?'. The text discusses the impact of the IVA (Value Added Tax) increase on low-income families. The 'emol' logo is visible at the bottom right.

¿Qué falta en la discusión
sobre el impacto de esta
política pública?

Por primera vez en sus diez años, plan AUGE se queda sin
recursos para mejoras

Nuevo decreto, previsto para 2016, debía incluir nuevas coberturas en las 80 enfermedades ya garantizadas. Pero las correcciones previstas para esta norma, evaluadas en \$6.047 millones, no fueron incluidas en proyecto de ley de presupuesto.

The screenshot shows the header of the emol.com Nacional website. It includes the logo 'emol. Nacional', the date 'Martes 07 de agosto del 2018 | Actualizado 09:23', and a navigation bar with links for Noticias, Economía, Deportes, Espectáculos, Tendencias, Autos, Servicios, Chile, Mundo, Tecnología, Educación, Documentos, Multimedia, Buscar, and amarillas.com.

Motivación

- Este es un curso de **métodos para el análisis de impacto de las políticas públicas**

Los objetivos son:

- O1. Aplicar análisis de regresión lineal con datos reales
- O2. Comunicar resultados para su uso en políticas públicas
- O3. Identificar los principales problemas de regresión estadística
- O4. Aplicar el concepto de causalidad en análisis de impacto de políticas públicas
- O5. Formular preguntas de análisis de impacto de políticas públicas
- O6. Distinguir el método más apropiado para responder estas preguntas
- O7. Criticar estudios/papers/reportes de políticas públicas donde se utilicen datos y métodos estadísticos

2. Logística del curso

Metodología

- Formato Cátedra y Taller
 - Viernes: Clase con [Pablo Celhay](mailto:pacelhay@uc.cl) ( pacelhay@uc.cl)
 - Sábados: Talleres de aplicación práctica en [RStudio](#) y dudas de los estudiantes.
 - [Jose Conejeros](mailto:jdconejeros@uc.cl) ( jdconejeros@uc.cl)
- Se requiere una participación activa del(la) alumno(a)

Metodología

- Lecturas semanales asignadas que complementan la clase
 - Papers asignados
 - S&W: Stock, J. and Watson, M., Introduction to Econometrics, 3rd edition. (Disponible en Biblioteca SJ-UC  [aquí](#))
 - A&P: Angrist, J. D., & Pischke, J. S. (2014). Mastering 'metrics: The path from cause to effect. Princeton University Press. (Disponible en Biblioteca SJ-UC  [aquí](#))
 - GMPRV: Gertler, P. et al. (2017). "La evaluación de Impacto en la Práctica", Segunda edición. Grupo Banco Mundial y Banco Inter Americano del Desarrollo. (Descargable  [aquí](#))
 - Otras lecturas durante el curso

Logística y Contenido del programa

Ver el calendario de actividades en clase



Trabajos y evaluación

- 1 Trabajo grupal aplicado a políticas públicas y evaluación de impacto: **30% (Policy Brief)**
 - Proponer y aplicar una evaluación de impacto a una política pública
 - Grupos de 3 a 4 estudiantes como máximo
 - Con evaluación de pares
- Presentación Final del trabajo grupal **30% (Policy Brief)**
 - Pregunta de investigación, metodología de evaluación, análisis, conclusiones y limitaciones.
 - 15 minutos por grupo
- 2 Tareas de análisis de datos: **20% c/u**
 - Se trabaja en R (default)
 - Se entregan scripts ORDENADOS además de respuestas escritas sobre las preguntas de la tarea
 - Es fundamental asistir a clases y talleres
 - Trabajo en parejas

¿Preguntas?



3. Introducción

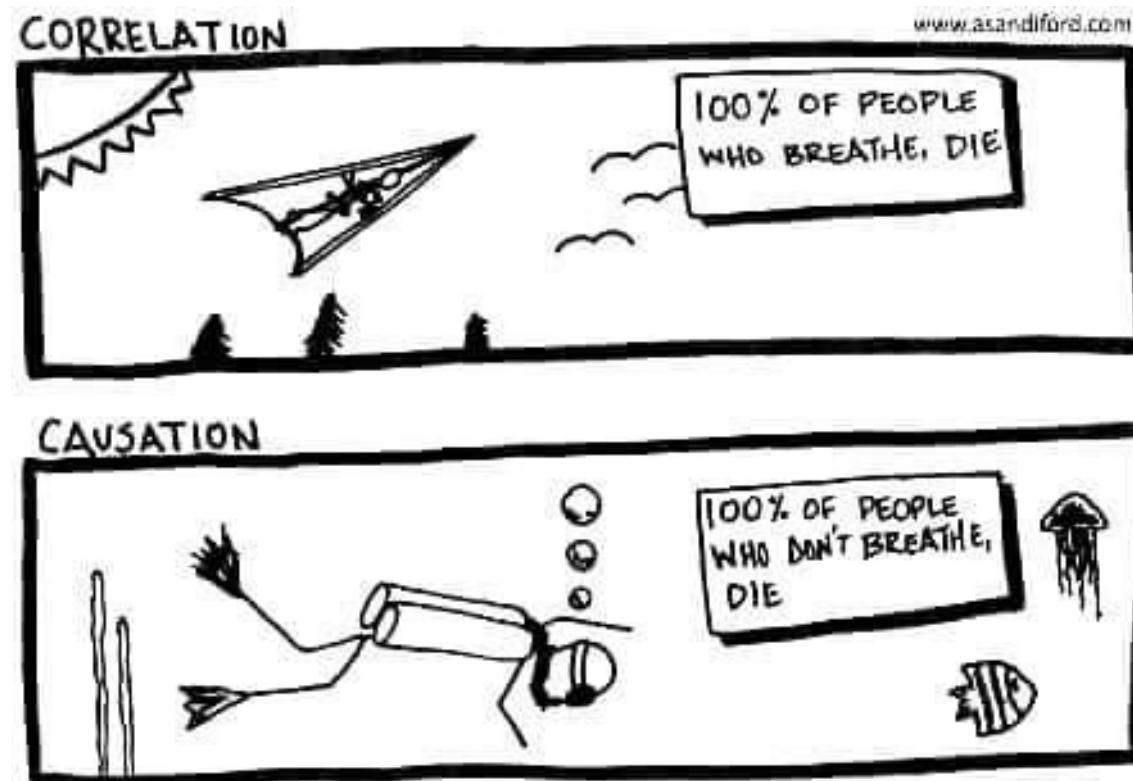
Introducción

¿Qué es la econometría?

- Un conjunto de técnicas de estadística que permiten examinar **relaciones empíricas** entre distintas variables
 - Empírica: Utilizamos datos
 - Relación:
 - Causal: ¿Cuál es la consecuencia de un cambio en una variable X sobre una variable Y ?
 - Asociación/Correlación: ¿De qué manera X e Y se "mueven conjuntamente"? ¿Cómo **predicen** movimientos en X los movimientos en Y ?
- Lo más común es inferir sobre el comportamiento empírico de la población mediante una muestra aleatoria
- Se repetirán muchos conceptos del curso de análisis de datos I

Introducción

- Una diferencia importante es que en este curso siempre nos estaremos preguntando si la relación que vemos entre X e Y es causal



Introducción

- Una diferencia importante es que en este curso siempre nos estaremos preguntando si la relación que vemos entre X e Y es causal

Illusions of causality: how they bias our everyday thinking and how they could be reduced

Helena Matute^{1}, Fernando Blanco¹, Ion Yarritu¹, Marcos Diaz-Lago¹, Miguel A. Vadillo² and Itxaso Barberia^{3,4}*

¹Departamento de Fundamentos y Métodos de la Psicología, Universidad de Deusto, Bilbao, Spain; ²Primary Care and Public Health Sciences, King's College London, London, UK; ³Departamento de Psicología Básica, Universitat de Barcelona, Barcelona, Spain; ⁴EventLab, Departamento de Personalidad, Evaluación y Tratamiento Psicológico, Universitat de Barcelona, Barcelona, Spain

Illusions of causality occur when people develop the belief that there is a causal connection between two events that are actually unrelated. Such illusions have been proposed to underlie pseudoscience and superstitious thinking, sometimes leading to disastrous consequences in relation to critical life areas, such as health, finances, and wellbeing. Like optical illusions, they can occur for anyone under well-known conditions. Scientific thinking is the best possible safeguard against them, but it does not come intuitively and needs to be taught. Teaching how to think scientifically should benefit from better understanding of the illusion of causality. In this article, we review experiments that our group has conducted on the illusion of causality during the last 20 years. We discuss how research on the illusion of causality can contribute to the teaching of scientific thinking and how scientific thinking can reduce illusion.



Introducción

- **Discusión en clase:** ¿Por qué es importante para la política pública?
- Un estudio publicado en *Nature* (Rauscher, Shaw and Ky, 1993) sugiere que escuchar a Mozart por 10-15 minutos puede aumentar temporalmente el CI en 8 o 9 puntos. Poco tiempo después el estado de Georgia en EE.UU comenzó a repartir CDs de música clásica a los padres de niñas(os) menores de edad. Esta iniciativa fue seguida por Colorado, Florida, etc.
- **Discusión en clase:** ¿Es esta una buena política pública para aumentar el CI de niñas(os)?

Introducción

- PROGRESA: Programa de transferencias condicionales en México 1997

Figure 1 CCTs in the World, 1997 and 2008



Introducción

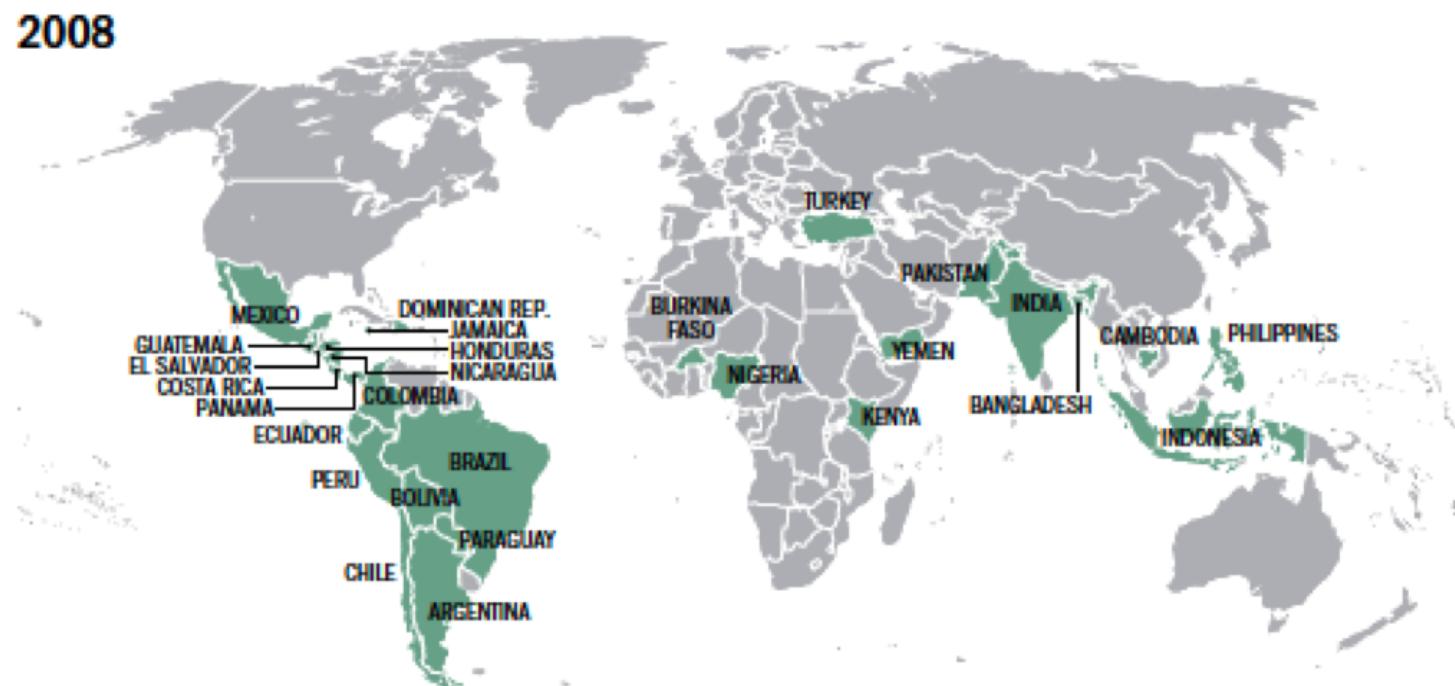
- Genera una serie de estudios sobre impacto

- Angelucci, Manuela, and Orazio Attanasio. 2008. "Oportunidades: Program Effects on Consumption, Low Participation, and Methodological Issues." *Economic Development and Cultural Change*.
- Angelucci, Manuela, and Orazio Attanasio. 2005. "Conditional Cash Transfers in Mexico: Using a Survey to Evaluate PROGRESA." In *Conditional Cash Transfer Studies*, London.
- Barham, Tania. 2005a. "Providing Conditional Cash Transfers to the Poor: A Report to Congress." Unpublished manuscript, Department of Agriculture, University of California at Berkeley, CA.
- Das, Jishnu, Quy-Toan Do, and Bruno Ferman. 2006. "Conditional Cash Transfer Programs." *World Bank Research Observer* 20 (1): 57–80.
- de Janvry, Alvaro, and Alan Gertler. 2000. "Final Report: The Impact of PROGRESA on Health." Unpublished manuscript, International Food Policy Research Institute, Washington, DC.
- Chilavert, Daniel, and Alan Gertler. 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment." *American Economic Review* 94 (2): 336–41.
- Gertler, Paul, Sebastián Martínez, and Marta Rubio-Codina. 2006. "Investing in Cash Transfers to Raise Long-Term Living Standards." Policy Research Working Paper 3994, World Bank, Washington, DC.
- Behrman, Jere R., and John Hoddinott. 2000. "An Evaluation of the Impact of PROGRESA on Pre-school Child Height." Unpublished manuscript, International Food Policy Research Institute, Washington, DC.
- _____. 2005. "Programme Evaluation with Unobserved Heterogeneity and Selective Implementation: The Mexican PROGRESA Impact on Child Nutrition." *Oxford Bulletin of Economics and Statistics* 67 (4): 547–69.



Introducción

- Rápida expansión de estos programas en el mundo



Source: World Bank.

Introducción

- Hay dos tipos de relaciones causales que forman preguntas comunes en el mundo de las PpPb:

1. El efecto de una variable sobre otra:

- ¿En cuánto afecta la nutrición materna en el peso al nacer del(a) niño(a)?
- ¿Cuánto afectan las campañas electorales los resultados electorales?
- ¿Cuánto afecta la educación los salarios?
- **¿Otras?**

2. El efecto de un programa social:

- ¿Cuánto cae la el % VIH con políticas de información sexual/sanitaria?
- ¿Cómo afectan la política de jardines infantiles el desarrollo infantil?
- ¿El CAE aumenta la matrícula universitaria? **¿Para quiénes?**
- **¿Otras?**

3.1 Causalidad

Causalidad

- Causalidad: Una acción específica tiene una consecuencia específica (medible)
- $Y = f(Z, X)$
 - Y es una función de muchos factores Z y de X .
 - Y: peso al nacer, Z: edad de la madre , X: # de controles prenatales
 - La meta es establecer o conocer la relación causal entre X e $Y \rightarrow$ **medir** cambios en Y que **sean atribuibles** a X .
 - Si vemos datos hoy de X e Y probablemente encontramos que más controles prenatales aumentan el peso al nacer.
 - **¿corr(X,Y)?**
- Pero, ¿es esta una relación causal? Evidencia mixta

Causalidad

- Más ejemplos:
 - ¿El cinturón de seguridad (X) reduce mortalidad por accidentes (Y)?
 - ¿La contaminación industrial (X) aumenta la temperatura del planeta (Y)?
 - ¿Venir a clases (X) mejora mi nota final (Y)?
- El concepto clave para responder estas preguntas sobre causalidad es el de **contrafactual** o contrafáctico
 - ¿Qué hubiese pasado en ausencia de la política pública?
 - El escenario ideal es poder comparar el resultado Y cuando X ocurre y cuando X no ocurre
- **Discusión en clase:** ¿Cuál es el contrafactual de los tres ejemplos arriba?
- **Discusión en clase:** ¿Cómo podríamos usar el concepto del contrafactual para estudiar una relación causal entre el número de controles prenatales sobre el peso al nacer?

Causalidad

- El problema fundamental es que no observamos el contrafactual:
 - No observamos que pasaría a los que ocupan el cinturón de seguridad, si estos mismos no lo hubiesen ocupado.
 - No observamos la temperatura del planeta actual bajo distintos tipos de emisiones industriales en el mismo tiempo.
 - No observamos a los mismos estudiantes en clases con distinto número de estudiantes en un mismo período.
- La meta (y el arte) del análisis empírico apunta a generar este contrafactual a partir de datos para poder estimar una relación causal entre (X) e (Y).
 - Situación ideal: Experimento aleatorio controlado
 - Situación más común: Datos observacionales analizados con técnicas econométricas

3.2 Experimentos aleatorios

Experimentos aleatorios controlados: Introducción

- Conceptualmente(!), este es el método ideal para estimar el efecto causal de un tratamiento (programa, intervención,etc.)
 - También se conocen cómo estudios aleatorios, experimentos sociales, experimento aleatorio, etc.
- Los RCT comparan el resultado Y para dos grupos que son similares salvo en que uno recibió el tratamiento y el otro no:
 - Los participantes **voluntariamente** deciden ser parte del estudio.
 - Esta muestra de voluntarios se divide aleatoriamente entre grupos de tratamiento y control
 - "Cara" o "sello"... con un software
 - Se ofrece el tratamiento al grupo de tratamiento
 - Grupo de control no se le ofrece el tratamiento, pero podría recibir un placebo
 - **Discusión en clase:** ¿Cuál es el contrafactual? ¿Cuál es el efecto del programa?

Experimentos aleatorios controlados: Introducción

- En las ciencias sociales, RCT son cada vez más frecuentes. Aún muy por debajo de ciencias médicas.
 - Pobreza, Trabajo, Salud, Educación, Crimen, etc.
- Algunos ejemplos:
 - Efectos de desparasitación en niñas(os) - Kenya
 - Efectos de seguros de salud en la salud de los asegurados - USA (RAND)
 - Efectos de campañas para votar sobre participación electoral - USA (RAND)
 - Efectos de transferencias condicionadas sobre salud y educación - México (Progresa)
- **Discusión en clase:** ¿Qué significa **aleatorio**?
- **Discusión en clase:** ¿Asignación aleatoria al tratamiento es lo mismo que una muestra aleatoria?

¿Por qué necesitamos econometría?

- Meta principal: desarrollar la habilidad de crítica analítica a estudios empíricos
- RCT se consideran como un benchmark pero:
 - No siempre los podemos implementar
 - No siempre responden las preguntas más relevantes
 - Pero si proveen un buen estimador del efecto causal cuando son bien ejecutados
- Si no podemos hacer un RCT, ¿entonces?
 - Usar datos observacionales (CASEN, datos administrativos, etc.)
 - Problema: La variable de interés (X) no fue asignada aleatoriamente
- Aprenderemos técnicas de análisis de regresión que permitan estimar efectos causales (bajo ciertos supuestos)
- Ver calendario de actividades en clase

Lecturas próxima semana

- S&W 1; 9.1, 9.4
- S&W 4.1-4.6, 5.1-5.2
- <https://www.vox.com/science-and-health/2018/8/29/17790118/alcohol-lancet-health-study>

Introducción: Repaso

- La primera clase hablamos de dos ideas claves en el análisis empírico:
 - Causalidad:
 - "X tiene un efecto causal en Y" → un cambio en Y es atribuible a X
 - \neq correlación (causalidad ⇒ correlación)
- **Discusión en clase:** ¿Qué es el contrafactual? ¿Qué hubiese pasado si...? **No Observable ... Replicable** ... con métodos/supuestos
- Experimento aleatorio:
 - Ideal desde un punto de vista metodológico → replica contrafactual en promedio (al menos)
 - Esto no significa que todo lo que es importante desde una Pol Pub se pueda responder con un RCT
- Hoy, revisaremos una pregunta causal de política pública con métodos experimentales y no experimentales (obervacionales)
- Pero primero discutamos qué hace que un diseño empírico (experimentales y no experimentales) sea válido o creíble

Validez

- La validez se refiere a la credibilidad de los resultados de un estudio:
 - Cualquier estudio debe tener la meta de alcanzar un alto estándar de validez
 - Hay grados de validez (no es bipolar). Dependerá de supuestos en el método
- Validez Interna:
 - La evidencia muestra un link entre X e Y pero, ¿hasta dónde es este una relación causal?
 - Clave: ¿hay otro factor que no sea X que pueda ser responsable de la asociación observada entre X e Y ?
 - Generalmente los RCT tienen validez interna por construcción

Discusión en clase: ¿Por qué?

En estudios no experimentales dependerá de los supuestos (estadísticos) en el análisis

Validez

- La validez se refiere a la credibilidad de los resultados de un estudio:
 - Cualquier estudio debe tener la meta de alcanzar un alto estándar de validez
 - Hay grados de validez (no es bipolar). Dependerá de supuestos en el método
- Validez Externa:
 - ¿Son los resultados extrapolables/generalizables el resto de la población u otras sub poblaciones?
 - Ejemplo: El efecto de la contaminación (X) sobre los casos de intoxicación (Y) en Quintero, ¿es válido para la región de Aysén?
 - ¿Para Puchuncaví?
 - ¿Hay otros estudios similares que llegan a la misma conclusión?
 - Para mejorar la validez externa:
 - Mejorar la selección de la muestra: ej, que sea mas representativa.
 - Replicar resultados en otros períodos, poblaciones o geografía
- Los conceptos de validez son claves para dar juicio de la calidad de la evidencia que tenemos al frente.

Validez

- Un estudio publicado en *Nature* (Rauscher, Shaw and Ky, 1993) sugiere que escuchar a Mozart por 10-15 minutos puede aumentar temporalmente el CI en 8 o 9 puntos. Poco tiempo después el estado de Georgia en EE.UU comenzó a repartir CDs de música clásica a los padres de niñas(os) menores de edad. Esta iniciativa fue seguida por Colorado, Florida, etc.
- **Discusión en clase:** ¿Es esta una buena política pública para aumentar el CI de niñas(os)?
- ¿Cómo evaluamos la calidad de este estudio en base a lo discutido hasta aquí?
 - Validez interna:
 - Validez externa:

Experimentos vs Datos observacionales

- La pregunta de PpPp hoy es:
 ¿Qué efecto tiene el tamaño de la clase sobre el rendimiento de las(os) niñas(os)?
- Hay dos formas (al menos) de diseñar estudios para responder esta pregunta

Discusión en clase:

1:

2:



4. Validez, RCTs y estudios observacionales

Experimentos

- Datos experimentales: El experimento Tennessee STAR (e.g., Hanushek 1999)
 - Proyecto STAR (Student-Teacher Achievement Ratio)
 - 1980s
 - Cuatro años de duración
 - Estudia el efecto del tamaño de la clase en $K - 3^{\text{ero}}$ básico
 - Estudiantes que entran en un determinado año son asignados **aleatoriamente** "randomizados" a uno de tres grupos
 - Clase pequeña (13-17 estudiantes)
 - Clase regular (22-25 estudiantes)
 - Clase regular + asistente
 - Se comparan los resultados de niñas(os) en clases pequeñas con aquellos de clases regulares
 - Veamos niñas(os) en Kindergarten
- **Discusión en clase:** ¿Cuáles son los grupos de tratamiento y control en este caso?

Experimentos

- La gran ventaja de un RCT (experimento aleatorio controlado) es la validez interna que ofrece
- **Discusión en clase:** ¿Por qué es válido internamente el estudio en este caso?
- Validez interna: Podemos confiar que las **diferencias en resultados** entre grupo de tratamiento y control representan una **asociación causal** del tratamiento
- Grupo de tratamiento y control son similares salvo en su **condición de tratamiento**

Experimentos

- ¿Funcionó la asignación aleatoria?
 - Podemos revisar el protocolo del experimento
 - Revisar documentos en los que muestren que se siguió el protocolo
 - Explorar si los grupos son similares
 - Esto es importante pues necesitamos que el grupo de control sea un contrafactual válido para el grupo de tratamiento
 - Esto se cumple sólo si **antes del tratamiento** el grupo de control y tratamiento son similares

Experimentos

Para replicar los ejercicios puede revisar el archivo: [Ejemplo_semana1.R](#)

```
library(AER) # Contiene los datos  
data(STAR) # Cargamos los datos  
colnames(STAR)
```

```
[1] "gender"      "ethnicity"    "birth"        "stark"        "star1"  
[6] "star2"       "star3"        "readk"        "read1"        "read2"  
[11] "read3"       "mathk"        "math1"        "math2"        "math3"  
[16] "lunchk"      "lunch1"       "lunch2"       "lunch3"       "schoolk"  
[21] "school1"     "school2"      "school3"      "degreek"     "degree1"  
[26] "degree2"     "degree3"      "ladderk"      "ladder1"     "ladder2"  
[31] "ladder3"     "experiencek" "experience1"  "experience2" "experience3"  
[36] "tethnicityk" "tethnicity1" "tethnicity2"  "tethnicity3" "systemk"  
[41] "system1"      "system2"      "system3"      "schoolidk"   "schoolid1"  
[46] "schoolid2"   "schoolid3"
```

Experimentos

```
# Generamos los datos con las variables de interés
df ← STAR %>%
  tidyverse::drop_na(stark, readk, mathk) %>%
  mutate(str = if_else(stark == "small", 1, 0),
        free_lunch = if_else(lunchk == "free", 1, 0),
        male = if_else(gender == "male", 1, 0),
        black = if_else(ethnicity == "afam", 1, 0)) %>%
  rowwise() %>%
  mutate(tscorek = (readk + mathk))
# Vemos los datos
df %>% dplyr::select(str, readk, mathk, tscorek, free_lunch, male, black) %>% head(n=5)
```

```
# A tibble: 5 × 7
# Rowwise:
  str readk mathk tscorek free_lunch male black
  <dbl> <int> <int>    <int>      <dbl> <dbl> <dbl>
1     1    447    473     920        0     0     0
2     1    450    536     986        0     0     1
3     0    439    463     902        1     1     1
4     0    448    559    1007        0     1     0
5     1    447    489     936        1     1     0
```

Experimentos

Características antes del tratamiento

	Tratados	Controles	Diferencia	p-value
Free Lunch (%)	0.4732	0.4874	-0.014	0.3225
Male (%)	0.5150	0.5128	0.002	0.8828
Black (%)	0.3113	0.3252	-0.014	0.2970

Discusión en clases: ¿En base a esta evidencia, "funcionó" la asignación aleatoria?

Experimentos

- Volviendo a la pregunta de política pública: ¿Cuál es el efecto del tamaño de la clase sobre los resultados escolares?
 - ¿Cómo respondemos esta pregunta usando los datos?
1. Estimación: Comparar los resultados promedio de niñas(os) en clases pequeñas con niñas(os) en clases regulares
 2. Test de hipótesis: Probar que la hipótesis nula de que **no hay efectos**
 3. Intervalos de confianza: Estimar un intervalo de confianza para la diferencia de promedios entre grupo de tratamiento y grupo de control

Discusión en clases: ¿Cómo serían los datos?

Experimentos

Resultados según el tamaño de la clase

Samll class in K	Media	SD	N
0	918.20	72.21	4048
1	931.94	76.36	1738
	922.33	73.75	5786

Discusión en clases: ¿Cuál es la diferencia en resultados promedio entre tener una clase pequeña y una clase regular?

Experimentos

Discusión en clases: ¿Cómo armamos el test de hipótesis?

Pizarra



Experimentos

- Conclusiones
 - Estar en una clase pequeña versus una clase regular:
 - Aumenta el puntaje del examen en 13.13
 - La diferencia es significativa estadísticamente (distinta de cero)

- **¿Significativa desde el punto de vista de política pública?**

1. ¿Qué significa 13.13 puntos adicionales? ¿Es grande?

2. Concepto adicional: Tamaño del efecto(Effect size) = $\frac{\text{Efecto en Y}}{\text{Desviación Estándar de Y}}$

3. Tamaño del efecto = $\frac{13.13}{73.74}$

4. Tamaño del efecto = 0.18 desviaciones estándar del examen

5. Util para revisar costo efectividad

4. Validez, RCTs y estudios observacionales

Estudio observacional

- ¿Qué pasa si no tenemos o podemos hacer un RCT?

Más común/real

- ¿Cómo estimamos el efecto de estar en una clase pequeña versus regular?
- Necesitamos un estudio **observacional**
- Tomar datos existentes de niñas(os) en clases pequeñas y comparar sus resultados con niñas(os) en clases regulares
- Preocupación principal de validez interna: ¿Por qué hay niñas(os) en clases pequeñas y otras(os) en clases regulares o más grandes?
- **Discusión en clase:** ¿Por qué no nos preocupamos de esto en el RCT (caso anterior)?
- **Discusión en clase:** ¿Por qué es importante preguntarnos esto?

Estudio observacional

- Datos observacionales: Base pública de datos sobre niñas(os) en clases de distintos tamaños
- Muestra de niñas(os) en clases desde $K = 6$ ($n=420$) en California
- Tamaño de clase medida como la razón entre estudiantes y profesoras(es)
- Esta medida del tamaño de clase no nos dice nada sobre que otras cosas cambian en clases de distinto tamaño
- ¿Cómo se ve la relación empírica entre tamaño de la clase y desempeño?

Estudio observacional

```
data(CASchools)
glimpse(CASchools)
```

Rows: 420

Columns: 14

```
$ district      <chr> "75119", "61499", "61549", "61457", "61523", "62042", "685...
$ school       <chr> "Sunol Glen Unified", "Manzanita Elementary", "Thermalito ...
$ county        <fct> Alameda, Butte, Butte, Butte, Fresno, San Joaquin, ...
$ grades        <fct> KK-08, KK-08, KK-08, KK-08, KK-08, KK-08, KK-08, KK...
$ students      <dbl> 195, 240, 1550, 243, 1335, 137, 195, 888, 379, 2247, 446, ...
$ teachers      <dbl> 10.90, 11.15, 82.90, 14.00, 71.50, 6.40, 10.00, 42.50, 19...
$ calworks     <dbl> 0.5102, 15.4167, 55.0323, 36.4754, 33.1086, 12.3188, 12.90...
$ lunch         <dbl> 2.0408, 47.9167, 76.3226, 77.0492, 78.4270, 86.9565, 94.62...
$ computer      <dbl> 67, 101, 169, 85, 171, 25, 28, 66, 35, 0, 86, 56, 25, 0, 3...
$ expenditure   <dbl> 6384.911, 5099.381, 5501.955, 7101.831, 5235.988, 5580.147...
$ income        <dbl> 22.690001, 9.824000, 8.978000, 8.978000, 9.080333, 10.4150...
$ english       <dbl> 0.000000, 4.583333, 30.000002, 0.000000, 13.857677, 12.408...
$ read          <dbl> 691.6, 660.5, 636.3, 651.9, 641.8, 605.7, 604.5, 605.5, 60...
$ math          <dbl> 690.0, 661.9, 650.9, 643.5, 639.9, 605.4, 609.0, 612.5, 61...
```

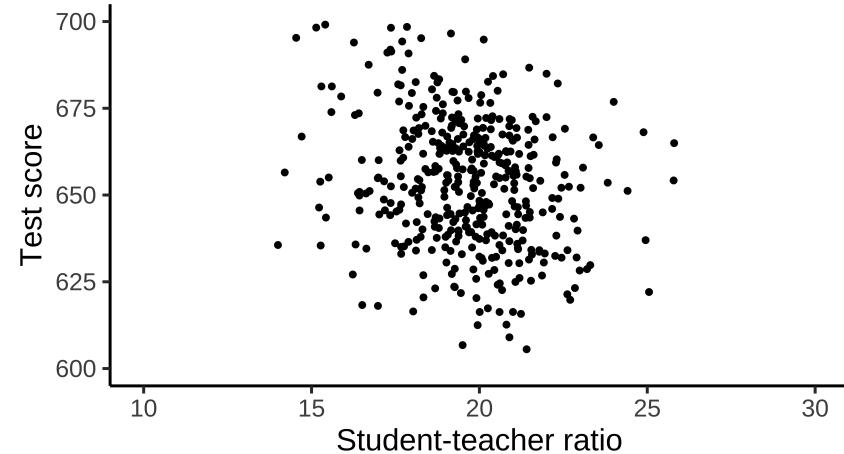
Estudio observacional

```
# Generamos una data de respaldo solo con las modificaciones a la data
df2 ← CASchools %>%
  mutate(ratio=students/teachers,
        small;if_else(ratio<20, 1, 0)) %>%
  rowwise() %>%
  mutate(tscorek=mean(c(read, math)))

# Vemos los datos
df2 %>% dplyr::select(tscorek, students, teachers, ratio, small) %>% head(n=5)
```

```
# A tibble: 5 × 5
# Rowwise:
  tscorek students teachers ratio small
    <dbl>     <dbl>     <dbl> <dbl> <dbl>
1   691.      195     10.9  17.9     1
2   661.      240     11.1  21.5     0
3   644.     1550     82.9  18.7     1
4   648.      243      14    17.4     1
5   641.     1335     71.5  18.7     1
```

Estudio observacional



Discusión en clase: ¿Qué muestra esta figura?

¿Cómo contestamos la pregunta de política pública?

- Tomemos los datos y generemos una variable: $= 1$ si el $\frac{\text{número de estudiantes}}{\text{profes}} < 20$ y $= 0$ si $\frac{\text{número de estudiantes}}{\text{profes}} \geq 20$
- Y hagamos lo mismo que antes: test de hipótesis

Estudio observacional

Resultados según el tamaño de la clase

Samll class	Media	SD	N
0	650.077	17.854	181
1	657.246	19.385	239
	654.157	19.053	420

Discusión en clases: ¿Cuál es la diferencia en resultados promedio entre tener una clase pequeña y una clase regular?

Estudio observacional

Discusión en clase: ¿Respondimos la pregunta sobre el efecto del tamaño de clase en desempeño escolar?

- ¿Qué pregunta nos tenemos que hacer? (pista: contrafactual)
- ¿Cómo podemos revisar la validez interna de este resultado? (pista: lo hicimos para el RCT)
- Comparar otras características que no debiesen verse afectadas por el tratamiento

Características antes del tratamiento

	Tratados	Controles	Diferencia	p-value
Free Lunch (%)	41.8127	48.5247	-6.712	0.0115
English as a second language (%)	12.6212	19.9236	-7.302	0.0001
Average Income (in 000's of \$)	16.3045	14.0120	2.293	0.0005

Discusión en clases: ¿Es el grupo de control un buen **estimador del contrafactual**?

Conclusiones

- En el contexto de un experimento ideal, la diferencia en resultados promedio entre grupo de control y tratamiento nos puede entregar un **buen estimador** del efecto causal de una política pública o intervención.
- En el contexto de un estudio observacional, la diferencia observada entre grupos **comúnmente** no nos entrega un estimador causal. Depende de los supuestos.
- Necesitamos otros métodos pues los RCT no son frecuentes
 - O demoran mucho y las decisiones políticas son urgentes
- La regresión lineal es una herramienta útil y poderosa para generar estimadores causales.
 - Su forma más simple (bivariada) no es mejor en términos de validez interna que lo que acabamos de ver (de hecho lo anterior es una regresión bivariada)
 - Pero podemos utilizarla con otros métodos para obtener validez interna en la comparación de grupo de tratamiento y control cuando no tenemos un RCT
 - Para esto, paciencia y estudio en este curso **sobre todo en cuanto a lenguaje e interpretación**

Regresión lineal Bivariada

- La última vez concordamos que para poder establecer un link causal entre X e Y debíamos descartar que otras variables sean responsables de esta asociación
- El análisis de regresión permite hacer esto pues podemos **dejar constante** o **controlar estadísticamente** estas otras variables
- Hoy: introducción a el modelo de regresión bivariada
 - La regresión bivariada no nos permite controlar por otros factores pero es un buen punto de partida para introducir ideas claves
 - Bivariada: Dos variables (X e Y)
- Partamos por el concepto de la función de regresión poblacional (FRP)
 - Por ahora, asumamos que la relación entre X e Y es lineal y que la FRP puede escribirse como:

$$Y = \beta_0 + \beta_1 X + \mu$$

Regresión lineal Bivariada

- Partamos por el concepto de la función de regresión poblacional (FRP)
 - Por ahora, asumamos que la relación entre X e Y es lineal y que la FRP puede escribirse como:

$$Y = \beta_0 + \beta_1 X + \mu$$

- Y : **variable dependiente** o **variable de resultado**
- X : **variable independiente** o **variable explicatoria**
- μ : **error de la regresión**; contiene todas las otras variables distintas a X que también afectan a Y
- β_0 y β_1 son parámetros poblacionales de interés
 - β_0 es la constante de la línea
 - β_1 es la pendiente de la línea

Regresión lineal Bivariada

- Ejemplo

$$\text{Puntaje Prueba} = \beta_0 + \beta_1 \text{Tamano Clase} + \mu$$

Discusión en clase: ¿Qué cosas contiene μ en este caso?

- A veces la FRP se escribe de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

- Aquí, i representa el dato (Y, X) para la observación i
- Por ejemplo en la regresión: $\text{Puntaje Prueba} = \beta_0 + \beta_1 \text{Tamano Clase} + \mu$, i puede ser una niña o un colegio

Regresión lineal Bivariada

- Cómo bien saben para estimar parámetros poblacionales (ej., un promedio) y inferencia estadística hay que recolectar datos y aplicar un estimador a esa muestra de datos
- Por ejemplo el promedio
- **Discusión en clase:** ¿Cómo se ven los datos en nuestro ejemplo?

Base de datos

Observación(i)	Y	X
1	605	20
2	612	18
...
...
n	607	25

Regresión lineal Bivariada

- Esto nos lleva a la versión muestral de la regresión: Función de Regresión Muestral (FRM)
- El “ $\hat{\cdot}$ ” indica que es una estimación

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\mu}$$

ó

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Regresión lineal Bivariada

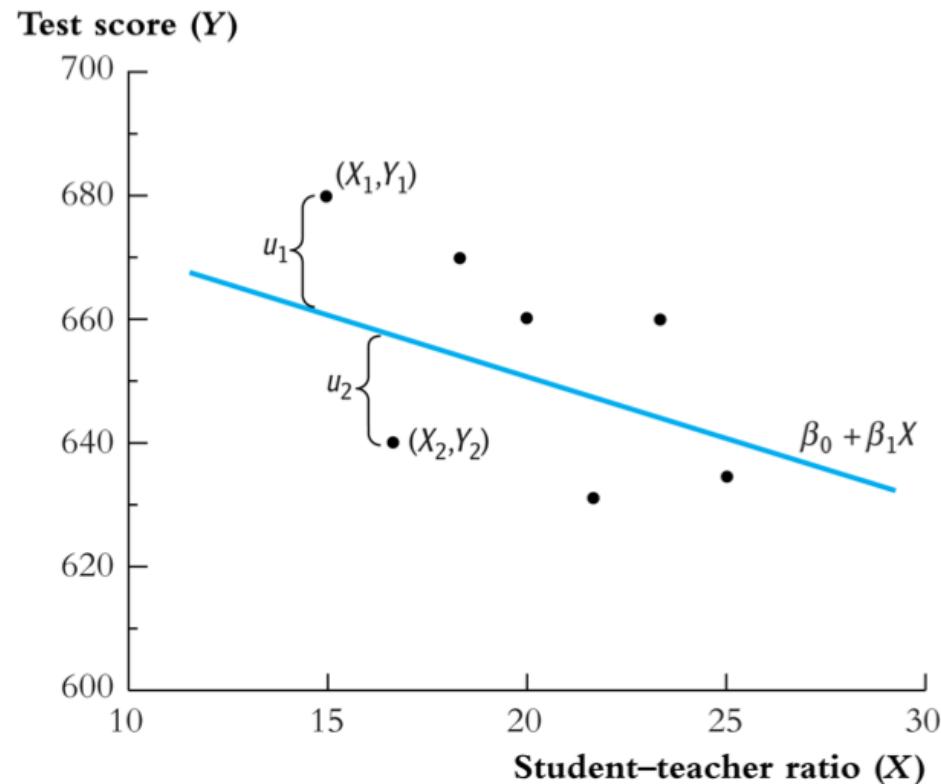
$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\mu}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- $\hat{\beta}_0$ es un estimador de β_0 y representa el valor predicho de Y cuando $X = 0$
- $\hat{\beta}_1$ es un estimador de β_1
 - $\hat{\beta}_1 = \frac{\Delta \hat{Y}}{\Delta X}$
 - Representa el cambio predicho en Y **asociado** a un aumento en X en una unidad
- \hat{Y} Es lo que el modelo predice que será Y para un cambio dado en X
- $\hat{\mu}_i$: Residuo. Diferencia entre el valor predicho de Y y el valor de la variable dependiente para la observación i .

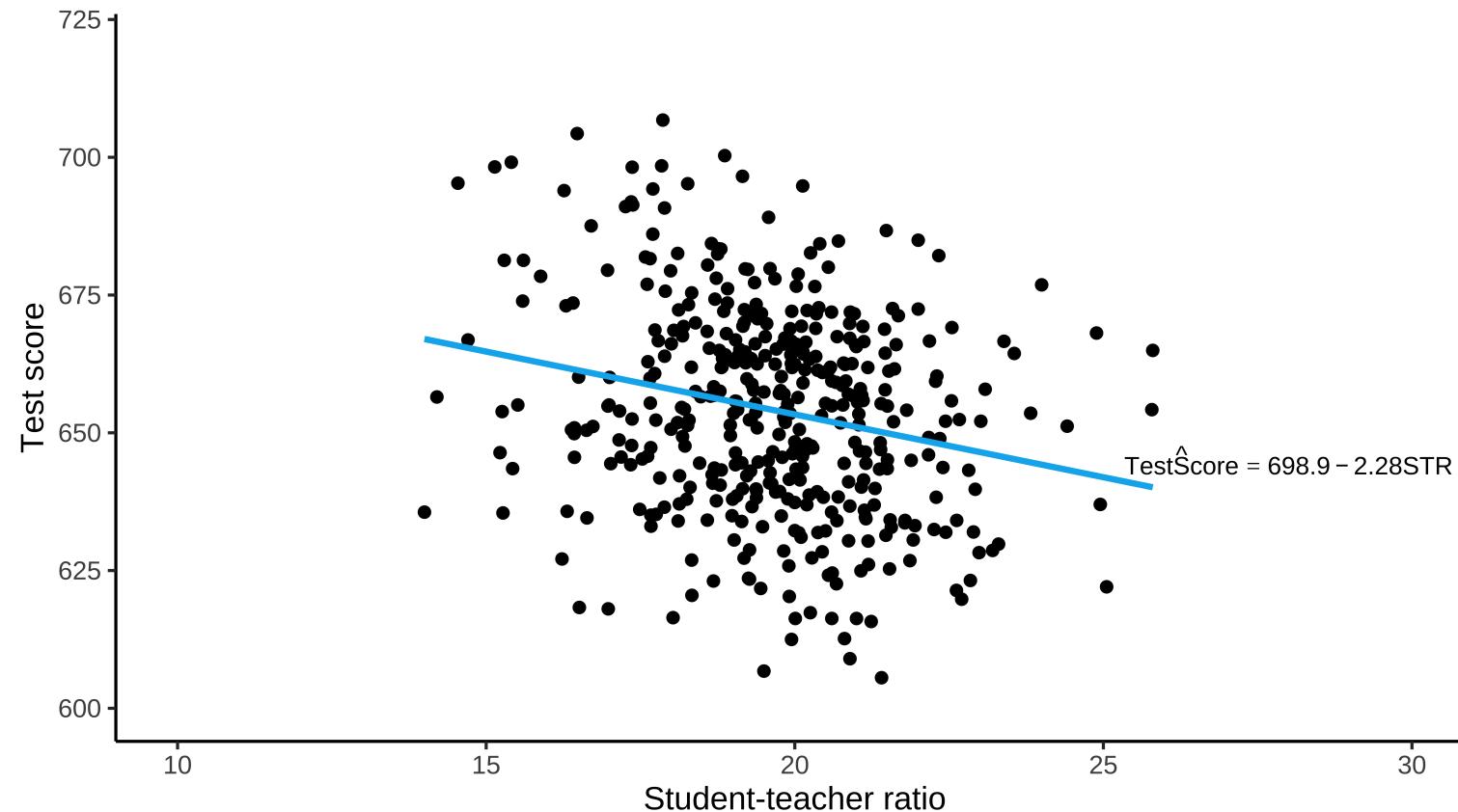
Regresión lineal Bivariada

- Visualización

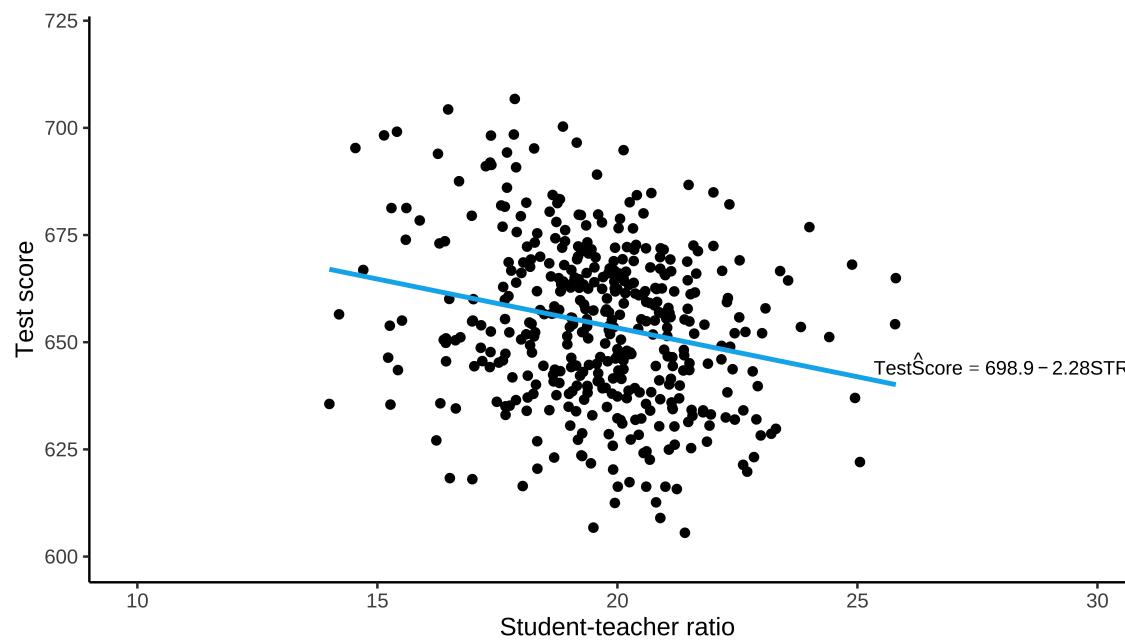


Regresión lineal Bivariada

- Visualización



Regresión lineal Bivariada

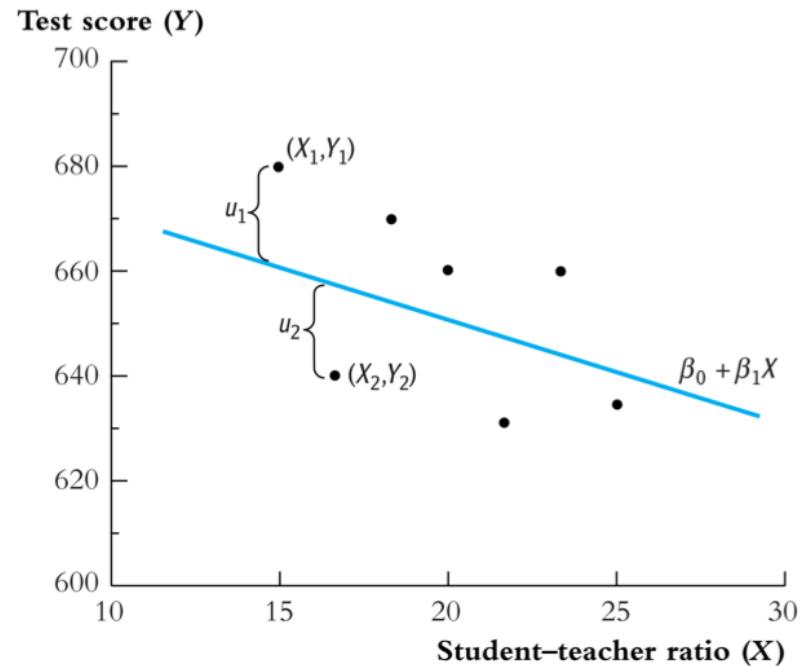


Discusión en clases: Basándonos en lo estimado, ¿cuál es el puntaje predicho para un valor de STR de 20?

Discusión en clases: Basándonos en lo estimado, ¿cuál es el cambio predicho en puntaje asociado a un aumento en la razón estudiante/profe de dos unidades?

Regresión lineal Bivariada

- Uno podría proponer distintos estimadores de β_0 y β_1
- El más aplicado: Mínimos Cuadrados Ordinarios (MCO/OLS)
- Bajo ciertas condiciones MCO tiene propiedades muy deseables
- Idea: Minimizar la suma del cuadrado de los residuos (verticalmente)



$$\text{Minimizar } \sum_{i=1}^n \hat{\mu}_i^2 = \min \sum_{i=1}^n Y_i - \hat{Y}_i^2 = \min \sum_{i=1}^n Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X)^2$$

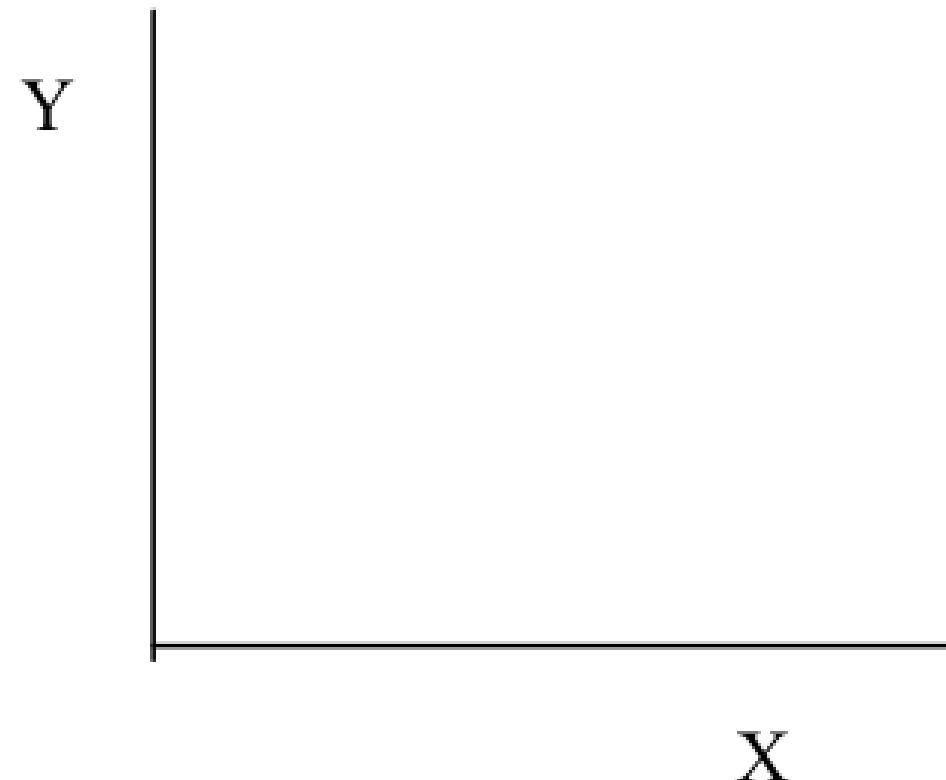
Regresión lineal Bivariada

- Al minimizar $\sum_{i=1}^n \hat{\mu}_i^2$, podemos derivar fórmulas para los estimadores de $\hat{\beta}_0$ y $\hat{\beta}_1$
- Las fórmulas en el contexto de la regresión bivariada son:
 - $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
 - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- Examinemos la fórmula de $\hat{\beta}_1$:
 - Denominador positivo \Rightarrow el signo del numerador determina el signo de $\hat{\beta}_1$
 - Numerador: Si Y está por sobre su media cuando X está por sobre su media (en promedio) el numerador es positivo. Cuando Y está por debajo de su media cuando X está por sobre su media el numerador es negativo.
 - Numerador es el mismo que el de la **correlación parcial**:

$$r_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Regresión lineal Bivariada

Discusión en clase: ¿Qué podemos decir del signo de $\hat{\beta}_1$?

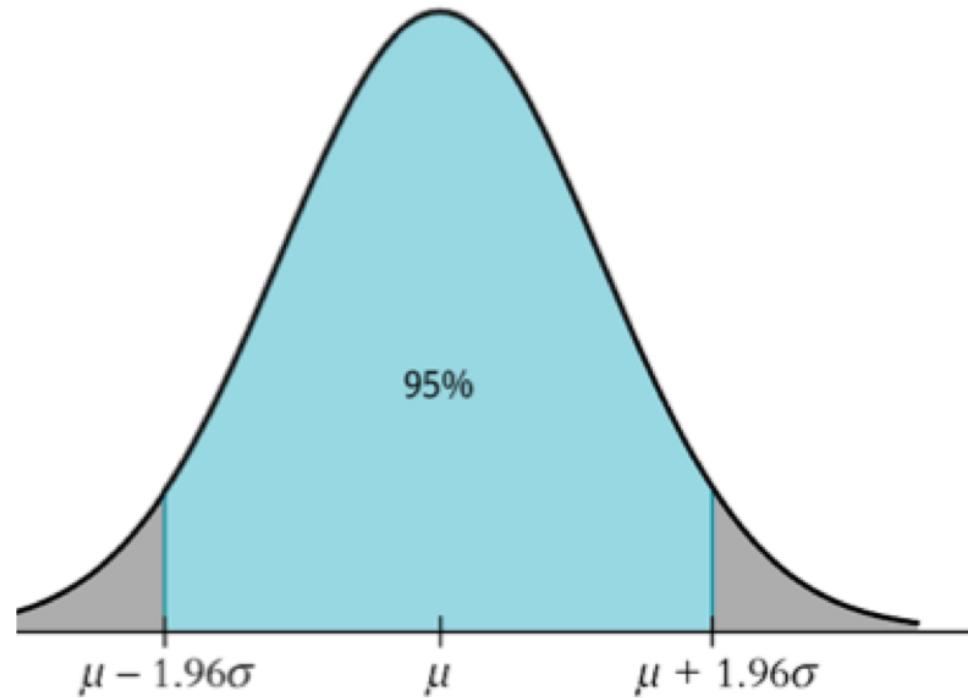


Distribución muestral de los estimadores MCO

- En MCO:
 - $\hat{\beta}_0$ es un estimador de β_0
 - $\hat{\beta}_1$ es un estimador de β_1
 - $\hat{\beta}_0$ y $\hat{\beta}_1$ siguen una distribución muestral
- Nos enfocamos en la distribución muestral de $\hat{\beta}_1$ ya que β_1 usualmente este es el parámetro de interés
- Necesitamos una distribución muestral para poder testear la hipótesis de que $H_o : \beta_1 = 0$ y construir un intervalo de confianza β_1
- La distribución de $\hat{\beta}_1$ es similar (análoga) a la distribución del promedio muestral que vieron en el curso pasado.

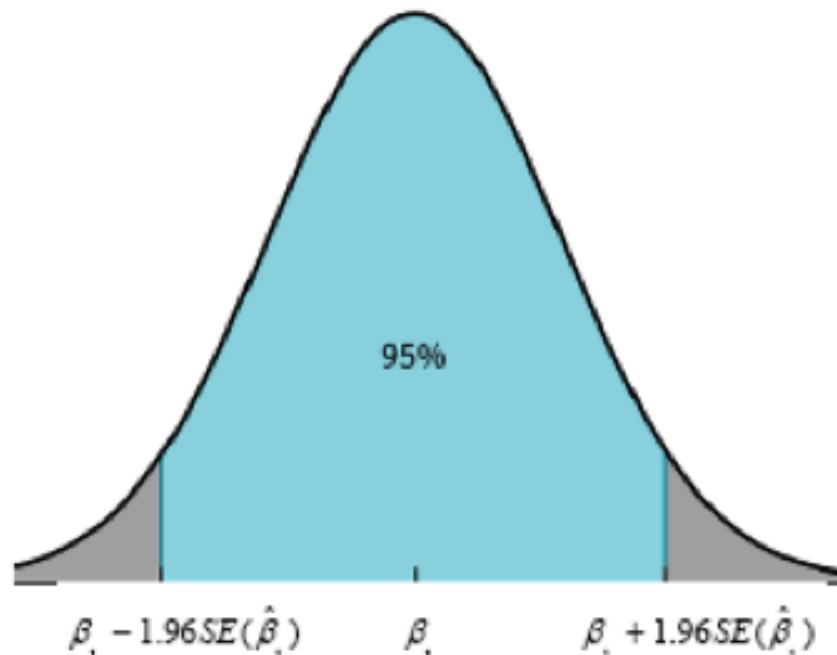
Distribución muestral de los estimadores MCO

La distribución del promedio es normal con un promedio igual a la media poblacional (μ) [
 $x \rightarrow N(\mu, \sigma^2)$]



Distribución muestral de los estimadores MCO

Para muestras grandes, la distribución de $\hat{\beta}_1$ es normal con media igual al valor poblacional (verdadero) de la pendiente de la FRP, β_1 . Esto se mantiene bajo ciertos supuestos que hablaremos en breve.



Distribución muestral de los estimadores MCO

- Tal como se calcula la varianza y error estándar de un promedio para armar los IC, ahora calculamos el error estándar de $\hat{\beta}_1$:

$$SE(\hat{\beta}_1) = s_{\hat{\beta}_1} = \sqrt{\frac{n}{n-2} \frac{\sum_{i=1}^n (X_i - \bar{X})^2 (\hat{\mu}_i)^2}{[\sum_{i=1}^n (X_i - \bar{X})^2]^2}}$$

- Intervalo de confianza para el 95% de significancia estadística: $\hat{\beta}_1 \pm 1.96 \cdot s_{\hat{\beta}_1}$
- Testeamos la hipótesis nula de que: $H_0 : \beta_1 = a$

$$t = \frac{\hat{\beta}_1 - a}{s_{\hat{\beta}_1}}$$

- al 95% rechazamos la hipótesis nula si $t > 1.96$ para un test de dos colas
- p-value: Si la hipótesis nula es verdad, es la probabilidad de obtener un test estadístico tan extremo o más extremo de lo que observamos

Ejemplo

Call:

```
lm_robust(formula = tscorek ~ ratio, data = df2, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	698.93	10.3644	67.436	9.487e-227	678.560	719.306	418
ratio	-2.28	0.5195	-4.389	1.447e-05	-3.301	-1.259	418

Multiple R-squared: 0.05124 , Adjusted R-squared: 0.04897

F-statistic: 19.26 on 1 and 418 DF, p-value: 1.447e-05

Discusión en clases:

- ¿ $\hat{\beta}_0, \hat{\beta}_1, SE(\hat{\beta}_1)$?
- Interpretación: ¿ $\hat{\beta}_0, \hat{\beta}_1$?
- Testear $H_0 : \hat{\beta}_1 = 0$ al 5% de confianza. ¿Por qué $H_0 : \hat{\beta}_1 = 0$ y no $H_0 : \hat{\beta}_1 = 10$? ¿Cómo interpretamos el resultado de este?

Ejemplo

```
# A tibble: 420 x 4
# Rowwise:
#> # ... with 410 more rows
```

	district	tscorek	ratio	small
	<chr>	<dbl>	<dbl>	<dbl>
1	75119	691.	17.9	1
2	61499	661.	21.5	0
3	61549	644.	18.7	1
4	61457	648.	17.4	1
5	61523	641.	18.7	1
6	62042	606.	21.4	0
7	68536	607.	19.5	1
8	63834	609	20.9	0
9	62331	612.	19.9	1
10	67306	613.	20.8	0

Ejemplo

Call:

```
lm_robust(formula = tscorek ~ ratio, data = df2, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	698.93	10.3644	67.436	9.487e-227	678.560	719.306	418
ratio	-2.28	0.5195	-4.389	1.447e-05	-3.301	-1.259	418

Multiple R-squared: 0.05124 , Adjusted R-squared: 0.04897

F-statistic: 19.26 on 1 and 418 DF, p-value: 1.447e-05

Discusión en clases:

- ¿ $\hat{\beta}_0, \hat{\beta}_1, SE(\hat{\beta}_1)$?
- Interpretación: ¿ $\hat{\beta}_0, \hat{\beta}_1$?
- Testear $H_0 : \hat{\beta}_1 = 0$ al 5% de confianza. ¿Por qué $H_0 : \hat{\beta}_1 = 0$ y no $H_0 : \hat{\beta}_1 = 10$? ¿Cómo interpretamos el resultado de este test?

Regresión bivariada con variable cualitativa

- La información cualitativa puede ser capturada por una variable binaria: variable que toma dos valores (ej., uno y zero), conocidas como variables "dummy"
- Ejemplos:
 - Mujer: = 1 si la persona i es mujer
 - Mujer: = 0 si la persona i es hombre
- Los valores 0, 1 son arbitrarios pero llevan a una fácil interpretación de la regresión lineal

Discusión en clase: ¿Cómo se ven los datos?

Regresión bivariada con variable cualitativa

Discusión en clase: ¿Cómo se ven los datos?

Base de datos

ID	ytrabaj	esc	sexo	ecivil
1	344	12	1	0
2	1318	16	0	1
...
...
n	2020	10	1	1

¿Qué pasa con la regresión lineal cuando X es binaria/dummy?

Ejemplo: Brecha salarial por género

Pregunta de interés: ¿Las mujeres ganan menos que los hombres? ¿Cuánto menos?

```
data(wage1)  
wage1
```

	wage	educ	exper	tenure	nonwhite	female	married	numdep	smsa	northcen
1	3.10	11	2	0	White	Female	Notmarried	2	1	0
2	3.24	12	22	2	White	Female	Married	3	1	0
3	3.00	11	2	0	White	Male	Notmarried	2	0	0
4	6.00	8	44	28	White	Male	Married	0	1	0
5	5.30	12	7	2	White	Male	Married	1	0	0
6	8.75	16	9	8	White	Male	Married	0	1	0
7	11.25	18	15	7	White	Male	Notmarried	0	1	0
8	5.00	12	5	3	White	Female	Notmarried	0	1	0
9	3.60	12	26	4	White	Female	Notmarried	2	1	0
10	18.18	17	22	21	White	Male	Married	0	1	0
11	6.25	16	8	2	White	Female	Notmarried	0	1	0
12	8.13	13	3	0	White	Female	Notmarried	0	1	0
13	8.77	12	15	0	White	Male	Married	2	1	0
14	5.50	12	18	3	White	Male	Notmarried	0	1	0
15	22.20	12	31	15	White	Male	Married	1	1	0

Ejemplo: Brecha salarial por género

Pregunta de interés: ¿Las mujeres ganan menos que los hombres? ¿Cuánto menos?

Ingresos según género

female	n	media	sd
Female	252	4.588	2.529
Male	274	7.099	4.161

Discusión en clase: En base a esta información, ¿cómo armamos un test para la H_0 : "No hay diferencias?"

Ejemplo: Brecha salarial por género

Discusión en clase: En base a esta información, ¿cómo armamos un test para la H_0 : "No hay diferencias?"

```
t.test(wage ~ female, alternative = c("two.sided"),
       data=wage1,
       paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

Welch Two Sample t-test

```
data: wage by female
t = -8.44, df = 456.33, p-value = 4.243e-16
alternative hypothesis: true difference in means between group Female and group Male is not equal to zero
95 percent confidence interval:
-3.096690 -1.926971
sample estimates:
mean in group Female   mean in group Male
        4.587659           7.099489
```

Ejemplo: Brecha salarial por género

- Podemos también utilizar regresión lineal para calcular esto de manera más directa:

$$y_{trabaj} = \beta_0 + \beta_1 mujer_i + \mu_i$$

- **Discusión en clase:** En base a esta información,

- ¿cuál es el valor predicho de los salarios?
 - ¿valor predicho para hombres?
 - ¿para mujeres?
- ¿Cómo interpretamos β_0 ? ¿Cómo interpretamos β_1 ?

Ejemplo: Brecha salarial por género

- Podemos también utilizar regresión lineal para calcular esto de manera más directa:

$$y_{trabaj} = \beta_0 + \beta_1 mujer_i + \mu_i$$

Call:

```
lm_robust(formula = wage ~ relevel(female, ref = "Male"), data = wage1,  
          se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.099	0.2514	28.24	2.453e-107
relevel(female, ref = "Male")Female	-2.512	0.2976	-8.44	3.125e-16
	CI Lower	CI Upper	DF	
(Intercept)	6.606	7.593	524	
relevel(female, ref = "Male")Female	-3.097	-1.927	524	

Multiple R-squared: 0.1157 , Adjusted R-squared: 0.114

F-statistic: 71.23 on 1 and 524 DF, p-value: 3.125e-16

Discusión en clase: ¿Cómo armamos un test para la H_0 : "No hay diferencias?

Ejemplo: Brecha salarial por género

- En este ejemplo utilizamos a el grupo "hombres" como **grupo base** o **grupo de comparación**, i.e. el grupo sobre el cuál comparamos. También se conoce cómo grupo de referencia o **categoría omitida**.
- ¿Cómo cambia la interpretación de los coeficientes si ahora aplicamos la siguiente regresión?

$$y_{trabaj} = \alpha_0 + \alpha_1 hombre_i + \mu_i$$

- ¿cuál es el grupo base ahora?
- ¿qué representa $\hat{\alpha}_1$?
- ¿cómo se relaciona $\hat{\alpha}_1$ con $\hat{\beta}_1$?

Ejemplo: Brecha salarial por género

$$y_{trabaj} = \alpha_0 + \alpha_1 hombre_i + \mu_i$$

Call:

```
lm_robust(formula = wage ~ female, data = wage1, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	4.588	0.1593	28.79	5.013e-110	4.275	4.901	524
femaleMale	2.512	0.2976	8.44	3.125e-16	1.927	3.097	524

Multiple R-squared: 0.1157 , Adjusted R-squared: 0.114

F-statistic: 71.23 on 1 and 524 DF, p-value: 3.125e-16

Contexto de evaluación de impacto

- Para estimar el efecto causal de un programa usamos regresión lineal
 - Supongamos que tenemos un RCT
 - Tenemos datos de nuestra variable de interés Y (ej., puntaje académico)
 - Definimos $Tratado = 1$ si la observación i esta en el grupo de tratamiento
 - Definimos $Tratado = 0$ si la observación i esta en el grupo de control
- **Discusión en clase:** ¿Qué regresión estimamos?

$$puntaje_i = \beta_0 + \alpha_1 tratado_i + \mu_i$$

- **Discusión en clase:** ¿Qué representa β_0, β_1 ?

Ejemplo: Proyecto STAR

```
# A tibble: 15 × 7
# Rowwise:
  str readk mathk tscorek free_lunch male black
  <dbl> <int> <int>    <int>      <dbl> <dbl> <dbl>
1     1   447    473     920        0     0     0
2     1   450    536     986        0     0     1
3     0   439    463     902        1     1     1
4     0   448    559    1007        0     1     0
5     1   447    489     936        1     1     0
6     0   431    454     885        1     1     0
7     0   395    423     818        1     0     1
8     0   451    500     951        0     0     0
9     1   478    439     917        0     0     1
10    1   455    528     983        0     0     0
11    0   430    473     903        0     1     0
12    0   437    468     905        0     1     0
13    1   474    559    1033        0     1     0
14    1   424    494     918        0     1     0
15    0   490    528    1018        0     1     0
```

Ejemplo: Proyecto STAR

- Usando diferencias de medias

Puntaje según tratamiento

str	n	media	sd
0	4048	918.201	72.214
1	1738	931.942	76.359

Ejemplo: Proyecto STAR

- Usando diferencias de medias

```
t.test(tscorek ~ str, alternative = c("two.sided"),
       data=df,
       paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

Welch Two Sample t-test

```
data: tscorek by str
t = -6.3768, df = 3129.1, p-value = 2.075e-10
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-17.965473 -9.515634
sample estimates:
mean in group 0 mean in group 1
918.2013         931.9419
```

Ejemplo: Proyecto STAR

- Usando regresión lineal

Call:

```
lm_robust(formula = tscorek ~ str, data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	918.20	1.135	808.936	0.000e+00	915.976	920.43	5784
str	13.74	2.155	6.377	1.943e-10	9.517	17.96	5784

Multiple R-squared: 0.007297 , Adjusted R-squared: 0.007125

F-statistic: 40.67 on 1 and 5784 DF, p-value: 1.943e-10

Discusión en clases: ¿Cuál es la diferencia de puntaje entre grupos? ¿Es significativa al 5%? ¿Cómo lo sabemos?

Supuestos detrás de los estimadores MCO

- Estamos interesados en la siguiente FRP: $Y_i = \beta_0 + \beta_1 X_i + \mu_i$
- Pueden haber otros métodos (distintos a MCO). ¿Por qué ocupar MCO?
 - Bajo ciertos supuestos, MCO tiene propiedades **deseables**.

Supuestos detrás de los estimadores MCO

- Cuatro supuestos claves:
 - **S1** Esperanza condicional del error es igual a cero
 - $E(\mu_i/X_i) = 0 \rightarrow E(Y_i/X_i) = \beta_0 + \beta_1 X_i$
 - X_i y μ_i son independientes (!Crucial!)
 - **S2** (X_i, Y_i) para $i = 1, \dots, n$ son iid
 - Basta tener una muestra aleatoria
 - **S3** (X_i, u_i) tienen un cuarto momento finito
 - Colas de la dist. (kurtosis) chicas \rightarrow sin outliers graves: verificar
 - **S4** $Var(u_i|X_i) = \sigma^2$: Homocedasticidad
 - Varianza del error igual para todo valor de X_i . No se cumple: usar `lm_robust`

Supuestos detrás de los estimadores MCO

- Con (S1 - S3), los estimadores de MCO tienen las siguientes propiedades:
 - **P1 Sesgo:**
 - $\hat{\beta}_1$ es un estimador insesgado de β_1 si $E[\hat{\beta}_1] = \beta_1$
 - Si tomamos muchas muestras (aleatorias) y aplicamos MCo a cada una, y promediamos esos estimadores, el promedio sería igual a β_1
 - En promedio, le "achuntamos" a β_1 .
 - ¿Por qué es importante? Si $\hat{\beta}_1$ es insesgado, entonces en valor esperado represente el efecto causal de X en Y
 - **P2 Consistencia:**
 - A medida que el tamaño de la muestra aumenta ("n tiende a infinito") MCO produce estimadores que se acercan al valor poblacional
 - **P3 Distribución normal:**
 - Para una muestra suficientemente grande los estimadores MCO se distribuyen normal (i.e. la → Podemos utilizar la distribución normal para hacer test de hipótesis)

Supuesto 1: Media Condicional del Error=0

- Para efectos de este curso (causalidad, ev de impacto) probablemente el supuesto más crucial es

$$E[\mu_i | X_i] = 0$$

- Si μ_i esta correlacionado con X_i entonces este supuesto NO se cumple.
- Siempre nos preguntaremos si es que este supuesto se cumple en el curso.
 - ¿Qué otras cosas en μ_i podrían estar correlacionadas con X_i ?
 - ¿Qué otros factores, distintos a X_i , afectan a Y_i ?
 - ¿Están estos correlacionados con X_i ?

Supuesto 1: Ejemplo "Cigarros y Salud"

- Queremos estimar la siguiente relación:

$$salud_i = \beta_0 + \beta_1 cigs_i + \mu_i$$

- $salud_i$ es una medida de la calidad de la salud de una persona i (ej., autor reporte "buena" o "mala") y $cigs_i$ representa el número de cigarros a la semana que esta misma persona fuma
- **Discusión en clase:** ¿Qué hay en μ_i ?
- La pregunta clave es si $\hat{\beta}_1$ es un estimador insesgado de β_1
 - El promedio de μ_i , ¿es el mismo para todos los niveles de $cigs_i$?
- Para simplificar: μ_i es el número de horas destinada al ejercicio (ejercicio_hrs)

Supuesto 1: Ejemplo "Cigarros y Salud"

- (S1) implica que el promedio de horas a la semana de ejercicio en la población es el mismo sin importar el número de cigarros que una persona fume. O, es el mismo para personas con distintos niveles de consumo de cigarros.
- Entonces el supuesto es que:
 - $E[ejercicio_hrs|cigs = 0] = E[ejercicio_hrs|cigs = 4] = E[ejercicio_hrs|cigs = 50]$
- **Discusión en clase:** ¿Es razonable?
- Pregunta tipo, ¿es razonable asumir que el número de horas que una persona ejercita no está correlacionado con el número de cigarros que fuma a la semana?

Ejemplo 2: Tamaño de la clase y desempeño académico

- Proyecto STAR con datos experimentales y regresión bivariada

Call:

```
lm_robust(formula = tscorek ~ str, data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	918.20	1.135	808.936	0.000e+00	915.976	920.43	5784
str	13.74	2.155	6.377	1.943e-10	9.517	17.96	5784

Multiple R-squared: 0.007297 , Adjusted R-squared: 0.007125

F-statistic: 40.67 on 1 and 5784 DF, p-value: 1.943e-10

- **Discusión en clase:** ¿Se cumple (S1)?

Ejemplo 2: Tamaño de la clase y desempeño académico

- **Discusión en clase:** ¿Se cumple (S1)?
- Un buen punto de comparación cuando vemos regresiones que tratan de estimar una relación causal es pensar en el experimento ideal.
 - X_i es asignado aleatoriamente
 - Entonces, cualquier variable en μ debería ser independiente de X_i
 - En un RCT ideal $E(\mu|X_i) = 0$
 - Con datos observacionales, esto es más difícil de creer y hay que pensar más en esto

Ejemplo 3: Tamaño de la clase y desempeño académico

- Proyecto STAR con datos observacionales y regresión bivariada

Call:

```
lm_robust(formula = tscorek ~ ratio, data = df2, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	698.93	10.3644	67.436	9.487e-227	678.560	719.306	418
ratio	-2.28	0.5195	-4.389	1.447e-05	-3.301	-1.259	418

Multiple R-squared: 0.05124 , Adjusted R-squared: 0.04897

F-statistic: 19.26 on 1 and 418 DF, p-value: 1.447e-05

- **Discusión en clase:** ¿Se cumple (S1)?
- Si $\hat{\beta}_1$ está sesgado, ¿qué hacemos?

Ejemplo 4: Discusión en clase

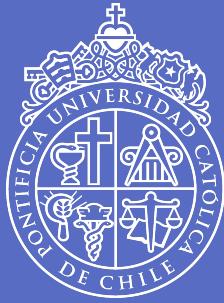
- Tengo una muestra de N estudiantes y veo dos variables (sexo y tamaño de clase). Si la intención es estimar el efecto del tamaño de la clase sobre desempeño académico con una regresión bivariada, ¿qué muestra ocupó? (pista: calcular proporción de hombres y mujeres en cada tipo de clase)

Muestra 1: Promedio del número de estudiantes por celda

	Hombres	Mujeres	Total
Clase grande	15	16	31
Clase chica	8	9	17
Total	23	25	48

Muestra 1: Promedio del número de estudiantes por celda

	Hombres	Mujeres	Total
Clase grande	9	22	31
Clase chica	12	5	17
Total	21	27	48



Semana 1

Regresión Lineal

19 de agosto, 2022

 **Pablo A. Celhay** |  pacelhay@uc.cl

Diseño y formato de la presentación:  José Daniel Conejeros |  jdconejeros@uc.cl |  JDConjeros