Pontificia Universidad Católica de Chile
Departamento de Ingeniería Industrial y de Sistemas
ICS2563 - Econometría Aplicada
Sección 1

Tarea 2

29 de marzo de 2022

2022-1 - Profesor Patricio Domínguez

Aspectos generales

- La tarea puede ser desarrollada en forma indivual o en parejas (2 estudiantes).
- La fecha de entrega de la tarea es el **7 de Abril a las 19:59 hrs** en el portal del curso en Canvas.
- La entrega debe incluir el script de `R` (o el dofile de `STATA`) y un informe de análisis en formato pdf.
- Sobre el **informe**:
  - El informe debe incorporar las respuestas a todas las preguntas, incorporando las figuras y/o tablas que estime conveniente.
- Sobre el **código**:
  - Cada entrega debe incluir el script de `R` (o el dofile de `STATA`) desde donde se corre todo el análisis.
  - Cada análisis debe ser desarrollado por ustedes mismos utilizando vectores y matrices; el uso de paquetes estadísticos será penalizado. Se indicará en las preguntas cuando deben realizar cálculos "a mano", es decir, utilizando fórmulas vistas en clases, vectores y matrices.
  - Es importante que el código esté adecuadamente comentado para facilitar su corrección. Por ejemplo, indicar qué parte del informe/preguntas se desarrollan en cada sección del código.
  - **Reproducibilidad:** El código debiera estar escrito de manera tal que cualquier persona lo pueda correr, y reproducir los resultados desde su computador. En el caso de crear datos aleatorios recomendamos fijar una semilla que permita reproducir los resultados tal cual queden especificados en el informe de reporte.

- Parte de la tarea es que se vean enfrentados a tomar decisiones. Si deben aplicar criterio, háganlo y justifiquen sus elecciones.
- Ante dudas o preguntas, se recomienda fuertemente la utilización del foro de preguntas del curso. Las respuestas pueden servir a otros compañeros, que se enfrentan a las mismas dudas. No se permite publicar respuestas.
- La tarea tiene un total de **100 puntos**, en donde la presentación del informe posee 5 puntos (El informe debe incorporar las respuestas a todas las preguntas, y las figuras y/o tablas que estime necesarias), y la presentación del script 5 puntos (todos los cálculos realizados).
- **Bonificacion:** Como un incentivo al uso del procesador de texto LATEX, se entregarán 5 puntos adicionales a quienes lo usen para escribir su informe.

Descripción de la tarea

El objetivo de esta tarea es trabajar con métodos de estimación de coeficientes de una regresión lineal. Además, aplicaremos el Teorema de Frisch-Waugh-Lovell y conoceremos algunos detalles del cálculo del coeficiente de determinación y la estimación de errores standard. Para ello, utilizaremos **las mismas bases de datos de la Tarea 1.**

A continuación se presentan las indicaciones presentadas anteriormente en la Tarea 1 a modo de recordatorio. Deben juntar (*merge*) distintas bases de datos (BDD) a nivel de colegios. En cada una de las BDD hay un identificador de establecimientos escolares (RBD) que le permitirán unir la información proveniente de diferentes fuentes. En concreto, debe conectar las siguientes BDD:

1) SIMCE 8 básico 2019: Puntaje promedio SIMCE de todos los establecimientos que rindieron la prueba ese año. Descargar desde este Link.

2) Matrícula por establecimiento: Número de estudiantes por establecimiento reconocidos por el MINEDUC. Descargar desde este Link. y poner atención a variable mat_total.

3) Datos de estudiantes prioritarios por establecimiento: Número de estudiantes prioritarios para recibir la Subvención Escolar Preferencial (SEP) por establecimiento. Descargar desde este Link y poner atención a variable n_prio.

Preliminar:

Lea los diccionarios de las bases de datos para identificar las variables presentes en cada una de ellas. Luego, crea una nueva variable "proporción estudiantes prioritarios" dividiendo la

cantidad de estudiantes prioritarios por el total de estudiantes matriculados en cada colegio. Utilizaremos esta variable como una variable proxy del índice de *vulnerabilidad social* del establecimiento educacional. Finalmente, inspeccione la BDD y familiarícese con las variables y sus respectivas observaciones.

- Se recomienda fuertemente el uso de funciones que puedan usar múltiples veces.
- En la mayoría de las preguntas debarán trabajar con la variable del puntaje SIMCE matemáticas **estandarizada**, excepto en casos donde ello implique que la variable no quede bien definida.

Preguntas

1. Métodos de Estimación (10 puntos)

Para esta pregunta considere el siguiente modelo de regresión lineal simple.

$$pje\_simce\_mat_i = \alpha + \beta * prop\_prio_i + \epsilon_i$$

Donde pje_simce_mat corresponde al puntaje SIMCE en Matemáticas **estandarizado**.

a) Estime los coeficientes (α, β) utilizando el método de máxima verosimilitud. Especifique en el código (*script*) la función de verosimilitud que será maximizada. Recomendamos que utilice las función `optim` de `R` u otra equivalente.

b) Estime los coeficientes (α, β) utilizando el método de mínimos cuadrados ordinarios. Para esta pregunta puede utilizar la solución conocida de mínimos cuadrados ordinarios.

2. Coeficiente de determinación (15 puntos)

Para la siguiente pregunta, considere que pje_simce_mat corresponde al puntaje SIMCE en Matemáticas **estandarizado**.

a) Determine el coeficiente de determinación (R^2) del siguiente modelo de regresión lineal simple:

$$pje\_simce\_mat_i = \alpha + \beta * prop\_prio_i + \epsilon_i$$

b) Determine el coeficiente de determinación (R^2) del siguiente modelo de regresión lineal múltiple:

$$pje\_simce\_mat_i = \beta_0 + \beta_1 * prop\_prio_i + \beta_2 * mat\_total_i + \beta_3 * PP_i + \beta_4 * PS_i + \epsilon_i$$

donde PP y PS son indicadores (variables binarias o dummies) iguales a 1 si el establecimento es particular pagado o no, o si es particular subvencionado o no, respectivamente. Debe construir estas variables.

c) Discuta los valores obtenidos en a) y b).

3. Error estándar (10 puntos)

Estime (en forma manual, esto es explicitando cada uno de los cálculos) el valor de los coeficientes (β) y su error estándar (σ_β) de la regresión multivariada estimada en 2.b). Recuerde utilizar pje_simce_mat como el puntaje SIMCE en Matemáticas **estandarizado**.

4. Binscatter FWL (15 puntos)

Utilice el teorema de FWL para construir un binscatter que muestre la relación entre la proporción de estudiantes prioritarios y el puntaje simce en matemáticas (**estandarizado**), controlando por las siguientes variables: mat_total, PP, PS. Construya un binscatter con 20 observaciones.

Luego, construya un segundo binscatter de 20 observaciones no residualizadas que obtendría en un modelo de regresión lineal simple.

Muestre ambos binscatter incluyendo la recta de ajuste MCO para ambos casos y compare.

5. Análisis de una tabla de regresión (40 puntos)

En esta parte debe **construir una tabla** que compare el coeficiente de determinación, el estimador puntual (β) y la desviación estándar (σ_β) del coeficiente de proporción de estudiantes prioritarios (prop_prio) de distintas regresiones.

Los valores de la Tabla 1 adjunta son ficticios, usted **debe completar luego de estimar cada una de las regresiones** con las condiciones identificadas en las filas de regresores y la variable dependiente correspondiente. En otras palabras, considere que la tabla contiene: El valor del coeficiente de proporción de estudiantes prioritarios (β) y su desviación estándar (entre paréntesis), y el coeficiente de determinación.

Consta de dos paneles según la especificación de la variable dependiente:

- **A**: log(pje_simce_mat), logaritmo del puntaje simce de matemáticas según la escala SIMCE **original**.
- **B**: pje_simce_mat, valor **estandarizado** del puntaje SIMCE matemáticas.

Además, las columnas de la tabla corresponden a:

- (1) Regresión bivariada simple entre proporción de estudiantes prioritarios y la variable dependiente, utilizando la base de datos a nivel individual (todas las observaciones).
- (2) Regresión multivariada de proporción de estudiantes prioritarios sobre la variable dependiente, incluyendo como covariables los regresores adicionales I: mat_total, y las siguientes variables binarias: PP, PS. Se utiliza la base de datos a nivel individual (todas las observaciones).

- (3) Regresión multivariada de proporción de estudiantes prioritarios sobre la variable dependiente, incluyendo como covariables los regresores adicionales I: mat_total, y las siguientes variables binarias: PP, PS, y además una variable adicional que distribuye de la siguiente manera: $X_1 \sim N(\mu = 0, \sigma^2 = 0{,}015) +$ prop_prio. Utilizando la base de datos a nivel individual (todas las observaciones).
- (4) Regresión bivariada simple entre proporción de estudiantes prioritarios y la variable dependiente, utilizando la base de datos del binscatter (20 observaciones).
- (5) Regresión multivariada de proporción de estudiantes prioritarios sobre la variable dependiente, incluyendo como covariables los regresores adicionales I: mat_total, y las siguientes variables binarias: PP, PS. Utilizando la base de datos del binscatter (20 observaciones).
- (6) Regresión multivariada de proporción de estudiantes prioritarios sobre la variable dependiente, incluyendo como covariables los regresores adicionales I: mat_total, y las siguientes variables binarias: PP, PS, y además una variable adicional que distribuye de la siguiente manera: $X_1 \sim N(\mu = 0, \sigma^2 = 0{,}015) +$ prop_prio. Utilizando la base de datos del binscatter (20 observaciones).

Hint: usted debe construir el regresor adicional II tal como se indica. Debe simular datos por lo que se le recomienda fijar una semilla. Además, se utiliza la variable prop_prio que se ha utilizado en toda la tarea.

Una posible organización de los resultados se indica en la tabla a continuación:

Cuadro 1: Comparación de modelos

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: log(pje_simce_mat)						
Prop. de estudiantes prioritarios	0.05	0.05	0.05	0.05	0.05	0.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Regresores adicionales I	No	Sí	Sí	No	Sí	Sí
X_1	No	No	Sí	No	No	Sí
Número de observaciones	n	n	n	p	p	p
R^2	0.2	0.2	0.2	0.2	0.2	0.2
Panel B: pje_simce_mat						
Prop. de estudiantes prioritarios	0.05	0.05	0.05	0.05	0.05	0.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Regresores adicionales I	No	Sí	Sí	No	Sí	Sí
X_1	No	No	Sí	No	No	Sí
Número de observaciones	n	n	n	p	p	p
R^2	0.2	0.2	0.2	0.2	0.2	0.2

Notas: la tabla muestra resultados de 12 regresiones organizadas en dos paneles.

En el panel A se observan resultados del coeficiente de proporción de estudiantes prioritarios para diferentes regresiones que utilizan el logaritmo del puntaje simce matemáticas (original) como variable dependiente.

En el panel B se observan resultados del coeficiente de proporción de estudiantes prioritarios para diferentes regresiones que utilizan el puntaje simce estandarizado de matemáticas como variable dependiente.

Todos los resultados incluyen el valor coeficiente de proporción de estudiantes prioritarios β y su desviación estándar (σ_β, entre paréntesis), además del coeficiente de determinación (R^2). Reporte, además, el total de observaciones de cada regresión.

A la izquierda de cada fila se indican las covariables (regresores) incluidos en cada regresión. Finalmente, columnas (1)-(3) muestran resultados para regresiones utilizando todas las observaciones de la base de datos, mientras que las columnas (4)-(6) muestran resultados utilizando solo 20 observaciones, similares a las utilizadas para construir un Binscatter.

Analice los resultados que obtiene al completar la tabla y discuta:

a) Compare el coeficiente de determinación entre las regresiones que utilizan todas las observaciones y aquellas que utiliza para el binscatter de 20 observaciones.

b) Analice el rol de los "regresores adicionales I" tanto en la estimación puntual como la desviación estándar del coeficiente de proporción de estudiantes prioritarios (β, σ_β), y el coeficiente de determinación. Discuta.

c) Analice el rol de X_1 tanto en la estimación puntual como la desviación estándar del coeficiente de proporción de estudiantes prioritarios (β, σ_β), y el coeficiente de determinación. Discuta.

d) Compare los resultados en los paneles A y B y comente respecto de las diferencias observadas en el coeficiente de determinación, el valor del estimador puntual (β) y la desviación estándar (σ_β) del coeficiente de proporción de estudiantes prioritarios de las distintas regresiones.