NOMBRE: Juan Ignacio García Forteza
Francisco Javier Hortal Correa

SECCIÓN: 1

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ingeniería Industrial y de sisitemas

ICS2563 — Econometría Aplicada — 1' 2022

Tarea 3 – Respuesta Pregunta 1

1)

El sesgo tiene un valor de -0.2518647. (Basado en ayudantía 5, 2022-1) (Desarrollo en R)

2)

Consideramos la regresión larga; pje_simce_mat = $\beta_0 + \beta_1$PROP_PRIO + β_2MAT_TOTAL y su regresión corta asociada, pje_simce_mat = $\lambda_0 + \lambda_1$PROP_PRIO. Calculamos λ_1 mediante MCO en regresión simple:

$$\lambda_1 = \frac{Cov(PROP\_PRIO, pje\_simce\_mat)}{Var(PROP\_PRIO)}$$
$$= \frac{Cov(\beta_0 + \beta_1 PROP\_PRIO + \beta_2 MAT\_TOTAL)}{Var(PROP\_PRIO)}$$
$$= \beta_1 + \beta_2 \frac{Cov(PROP\_PRIO, MAT\_TOTAL)}{Var(PROP\_PRIO)}$$

Teniendo de esta manera el sesgo $\delta = \beta_2 \frac{Cov(PROP\_PRIO, MAT\_TOTAL)}{Var(PROP\_PRIO)}$ con tres componentes. Luego la regresión corta subestima el modelo.

$\beta_2 = 0.0002942866$. Notamos que es positivo, luego esperaríamos de acuerdo al cálculo anterior que la covarianza entre la proporción de alumnos prioritarios y el total de alumnos matriculados sea negativa, ya que la varianza siempre es positiva.

$Cov(PROP\_PRIO, MAT\_TOTAL) = -52.83594$. El valor efectivamente es negativo, lo que significa que su correlación es negativa. Luego, en promedio, un aumento de la proporción de alumnos prioritarios refleja una disminución del total de alumnos. Ahora estamos en condiciones de calcular el sesgo teórico:

$$Var(PROP\_PRIO) = 0.06173516$$
$$\delta = \beta_2 \frac{Cov(PROP\_PRIO, MAT\_TOTAL)}{Var(PROP\_PRIO)} = -0.2518647$$

Esto que nos da exactamente el mismo valor calculado anteriormente de manera empírica.

3)

i)

Beta regresión corta: -2.515116

Beta regresión larga: -2.263104

Sesgo: -0.2520118

Vemos que los resultados son parecidos a lo obtenido en los ejercicios anteriores. Los dos betas son negativos y aumenta en la regresión larga, lo que significa que la variable de matriculas totales influyo en que la proporción de profesores no disminuya tanto el puntaje de matemáticas.

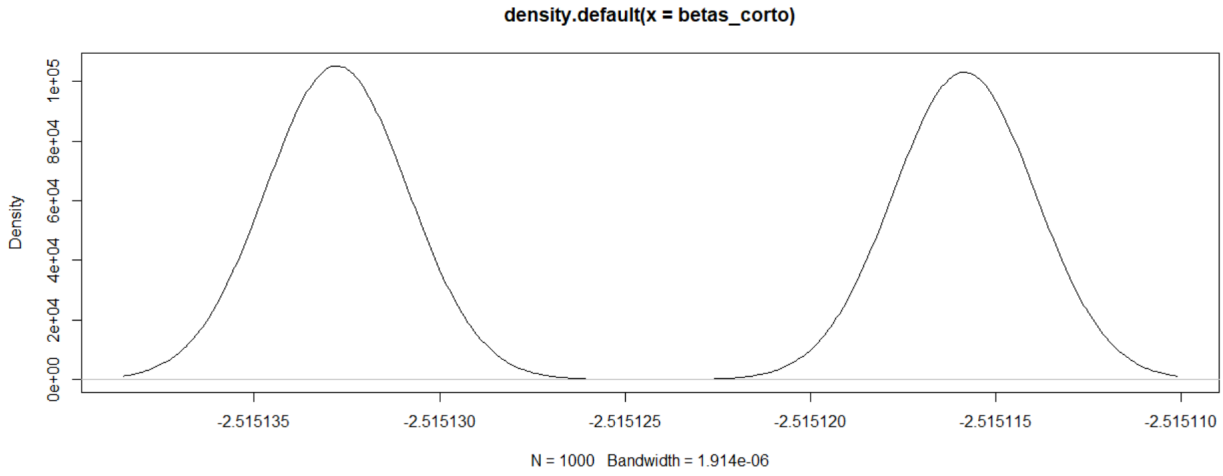
ii)

Se calculan 1000 regresiones como las de i) en R. se guardan los betas y sesgos.

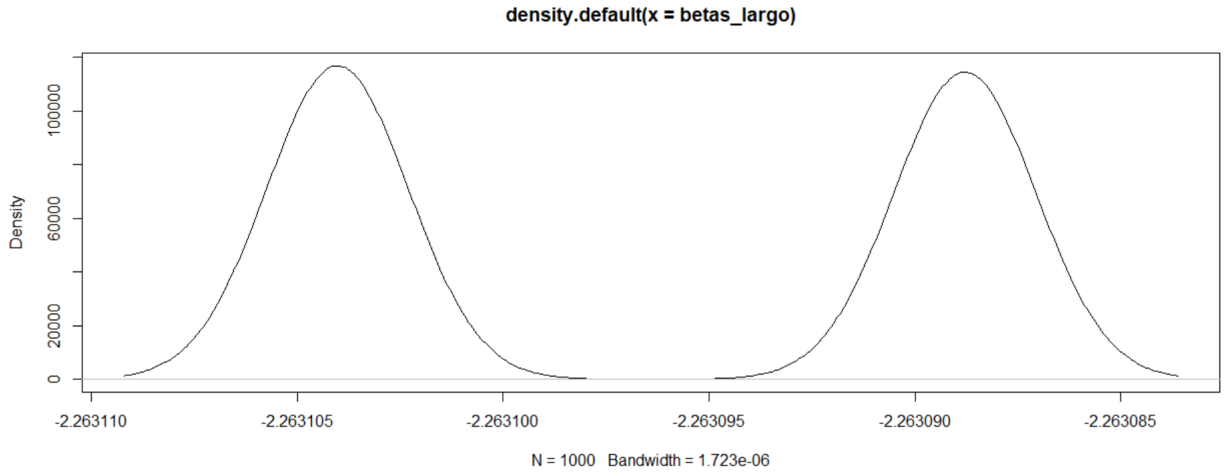
iii)

Se procede a mostrar los histogramas de los betas y sesgos.

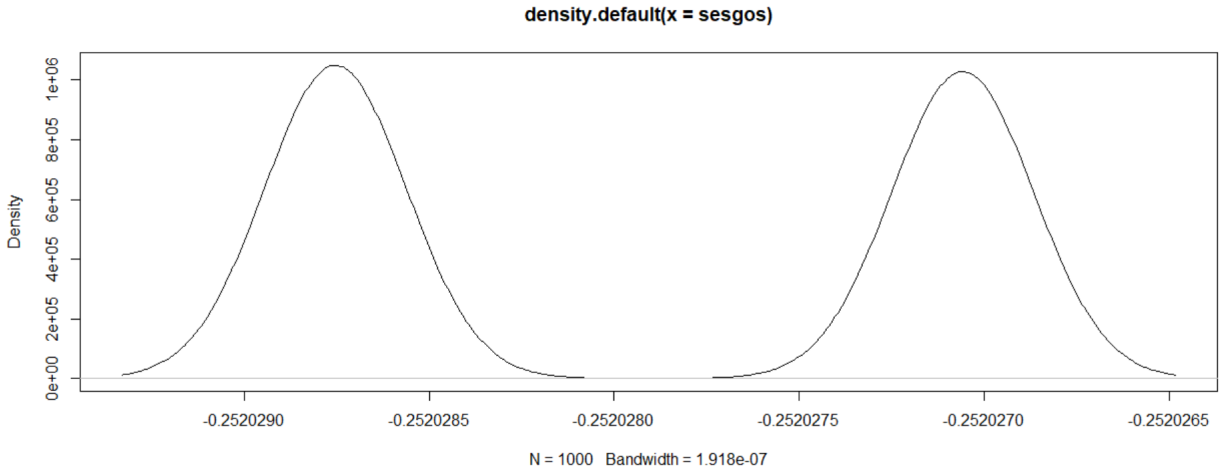
Beta regresión corta:



Beta regresión larga:



Sesgo:



Podemos ver en el histograma de el sesgo que hay dos campanas. Esto se puede deber a que haya alguna variable omitida que diferencie los sesgos de un grupo de establecimientos con otros. También se puede ver simetría en las campanas, lo que es buena señal de que son datos aleatorios y representan bien al universo completo.

NOMBRE: Juan Ignacio García Forteza
Francisco Javier Hortal Correa

SECCIÓN: 1

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ingeniería Industrial y de sisitemas

ICS2563 — Econometría Aplicada — 1' 2022

Tarea 3 – Respuesta Pregunta 2

1) Para esta pregunta se agregó a la base la base de datos la tasa de profesores por estudiante y se utilizaron otras variables ya existentes para realizar 4 regresiones distintas con variables que podrían afectar el puntaje simce de un establecimiento.

El primer modelo es una regresión entre el puntaje estandarizado simce matemáticas, tasa de profesores por estudiante y porcentaje de estudiantes prioritarios del establecimiento.

El segundo modelo es una regresión entre el puntaje estandarizado simce matemáticas, tasa de profesores por estudiante y el porcentaje de estudiantes mujeres respecto al total de alumnos del establecimiento.

El tercer modelo es una regresión entre el puntaje estandarizado simce matemáticas, tasa de profesores por estudiante y el puntaje simce de lenguaje estandarizado.

Finalmente el cuarto modelo es una regresión entre el puntaje estandarizado simce matemáticas, tasa de profesores por estudiante y si el establecimiento es rural o no.

A continuación se adjunta una tabla resumen con los principales resultados obtenidos.

Modelo	β	R2 ajustado
1	-8.246576	0.3913
2	-46.837265	0.07226
3	-30.723710	0.6694
4	-35.13113	0.07041

Como se puede ver en los resultados obtenidos, en el moedelo 2 y 4 se presenta una baja relación entre las variables, esto se puede afirmar al ver R2 ajustado ya que este es muy cercano a 0. Sin embargo en los modelos 1 y 3 se puede notar una mayor relación entre las variables de la regresión, sobretodo en el modelo número 4.

2) El modelo escogido es el número 3, el cual es una regresión entre el puntaje estandarizado simce matemáticas, tasa de profesores por estudiante y el puntaje simce de lenguaje estandarizado. Este modelo viene siendo el escogido por nosotros porque como se puede observar en la tabla del ítem anterior, es el que tiene un R2 ajustado más cercano a 1, lo que quiere decir que es el modelo que mejor ajusta la relación de las variables. El valor del coeficiente de la tasa de profesores por estudiante de esta regresión se puede interpretar que si tiene una relación con el puntaje simce matemáticas ya que como se puede ver con el R2 ajustado este está relativamente cercano a 1.

3) A pesar que el moedelo elegido tiene un R2 ajustado alto, el coeficiente de la tasa de profesores por

estudiante podría estar sesgado respecto a su efecto causal sobre el puntaje simce. Esto se debe a que son muchas las variables que pueden influir en esto y al no ser consideradas pasan a ser variables omitidas, lo que produce un sesgo en el coeficiente. Una posible fuente de sesgo es que se sabe la cantidad de matriculados totales de cada establecimiento pero no se sabe por ejemplo la cantidad de alumnos que son repitentes o el promedio de los alumnos del establecimiento, lo cual podría influir de manera directa en el coeficiente de la regresión ya que es de esperarse que alumnos con mayor promedio obtengan mejores resultados. Acá aparece sin embargo otro problema que vendría siendo la posibilidad de que la dificultad entre los establecimientos no sea la misma, lo que podría llevar a que puedan haber promedios altos con bajo nivel de conocimientos y promedios no tan altos con buen nivel de conocimiento.
Otra posible fuente de sesgo puede ser la disponibilidad de profesores, ya que es de esperarse que en regiones mas pobladas como lo es la región metropolitana por ejemplo exista una mayor disponibilidad de docentes mientras que regiones mas alejadas de la capital como Arica o Magallanes quizás no cuenten con mucha disponibilidad de docentes.
Esto podría tener amenazas sobre la regresión estimada ya que se podría sobreestimar o menospreciar la relación entre cantidad de profesores por alumno y el puntaje obtenido en el simce, lo que podría llevar a conclusiones erroneas y esto a que se realicen planificaciones a futuro no adecuadas. Por ejemplo si se encontrara una baja relación entre la proporción de profesores y el puntaje simce se podría decidir reducir el número de docentes por establecimiento, lo que podría tener un efecto perjudicial para la educación.

NOMBRE: Juan Ignacio García Forteza
Francisco Javier Hortal Correa

SECCIÓN: 1

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ingeniería Industrial y de sisitemas

ICS2563 — Econometría Aplicada — 1' 2022

Tarea 3 – Respuesta Pregunta 3

1)

Variable	Treatment	Control	p value for difference
Promedio Pje Simce Matemáticas Estandarizado	0.0169356	-0.004214702	0.5177415
Promedio Pje Simce Len Estandarizado Lenguaje	0.01718112	-0.004430001	0.5059704
Particular (=1)	0.0738255	0.07289639	0.9128392
Subvencionado (=1)	0.4354027	0.431538	0.8104186
Municipal (=1)	0.4748322	0.4726368	0.8923924
Hombres (=1)	0.519782	0.5176446	0.5059932
Mujeres (=1)	0.480218	0.4823548	0.5061189
Convenio SEP Vigente (=1)	0.8330537	0.8617781	0.01623583
Estudiantes Prioritarios (=1)	0.5625222	0.5629648	0.9570702
Establecimientos Rurales (=1)	0.2474832	0.235345	0.3852026
Región Metropolitana (=1)	0.2919463	0.2825005	0.5218593
Grupo Social Alto (=1)	0.07298658	0.0709496	0.8090977
Grupo Social Bajo (=1)	0.2005034	0.1992213	0.9214755
Matriculas Totales	573582	2305764	
Establecimientos Totales	1192	4623	

Vemos que el grupo de control es de 4.623 establecimientos y el grupo de tratamiento es de tan solo 1.192. Esta es una gran diferencia de más de 4 veces, sin embargo es una gran cantidad de establecimientos para un experimento. Aún es más grande la muestra considerando el número de alumnos matriculados (573.582 en tratamiento y 2.305.764 en control.

Podemos ver que no hay grandes diferencias en los p values de la mayoría de variables, por lo que estas son estadísticamente significativas, sin embargo destaca la variable Convenio SEP Vigente. Esta tiene un pvalue de 0.01623583, el que fue el menor pvalue de todos. Debido a que pvalue en este caso es menor a 0.05, la variable no alcanza un intervalo de confianza del 95%. Se rechaza la hipótesis nula y se acoge la hipótesis alternativa, lo que nos dice que los datos son significativamente distintos. Se puede concluir que la población de control tiene un mayor porcentaje de establecimientos con convenio SEP vigente que los establecimientos utilizados en el experimento. Variables como tipo de colegio, sexo, grupo social, región y promedio simce tienen una diferencia pequeña de promedios entre control y tratamiento que alcanza en todos los casos el intervalo de confianza de 95%. Se cree que la tabla en su conjunto está balanceada, ya que solo una variable tuvo problemas y las otras no. Con solo una variable la tabla de balances aún es válida,

pero siempre considerando que la variable SEP no es confiable. La diferencia entre los datos SEP se puede deber a que muchos colegios tienen el convenio y pocos no lo tienen. Si la muestra fuera aún más grande es probable que los resultados fueran más cercanos entre tratamiento y control, ya que si ocurre que por mala suerte algunos de los pocos colegios caen en control, la diferencia puede llegar a ser mayor. Es conveniente que el grupo de control y el de tratamiento tengan un tamaño similar.

2)

Se crearon 4 regresiones lineales. Se decidio ir en aumento de número de variables para poder analizar la variación en el coeficiente β de treat, el cual nos dirá lo que afecto el tratamiento en el puntaje simce de matemáticas suponiendo que las variables omitidas no sean estadísticamente significativas.

Regresión 1:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.76031    0.06160 -12.342   <2e-16 ***
treat        0.02450    0.03205   0.764    0.445
p_mujeres    1.56750    0.12412  12.629   <2e-16 ***
```

Regresión 2:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.97243    0.05643  17.232   <2e-16 ***
treat        0.02189    0.02521   0.869    0.385
p_mujeres    0.86007    0.09833   8.747   <2e-16 ***
PROP_PRIO   -2.47173    0.04127 -59.885   <2e-16 ***
```

Regresión 3:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.01531    0.05637  18.011  < 2e-16 ***
treat        0.01920    0.02507   0.766    0.444
p_mujeres    0.86949    0.09779   8.891  < 2e-16 ***
PROP_PRIO   -2.64690    0.04643 -57.003  < 2e-16 ***
RURAL_RBD    0.21750    0.02695   8.070 8.45e-16 ***
```

Regresión 4:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.792736   0.039424  20.108  < 2e-16 ***
treat          0.004035   0.017271   0.234  0.81527
p_mujeres     -0.217676   0.068842  -3.162  0.00158 **
PROP_PRIO     -1.067530   0.047360 -22.541  < 2e-16 ***
RURAL_RBD     -0.056285   0.019101  -2.947  0.00323 **
CONVENIO_SEP  -0.086775   0.028463  -3.049  0.00231 **
pje_simce_leng  0.664849   0.008334  79.778  < 2e-16 ***
```

Podemos ver que El coeficiente β de la variable treat parte en 0.02450, luego 0.02189, 0.01920 y finaliza en 0.004035. Esta disminución se debe a que las variables que se van agregando a los modelos están opacando

a el efecto de treat en el puntaje simce matemáticas.Esto no quiere decir que la variable treat no sea un factor de cambio, sino que otras variables están incidiendo aún más en el pje simce de matemáticas. También podemos ver unos p_value pequeños en todas las variables de los modelos excepto en treat, donde el p_value es más grande. Esto nos dice que las otras variables tienen un nivel de significancia mayor a treat.

3)

Se puede ver gracias a los datos expuestos que en promedio los establecimientos que pertenecieron al grupo de tratamiento obtuvieron mejor puntaje simce tanto en lenguaje como en matemáticas. La diferencia de puntajes estandarizados fue de 0.0211503 en matemáticas y de 0.02161112 en lenguaje. Podemos ver que la diferencia es bastante pequeña. El p-value en ambas pruebas simce fue superior al 95% de confianza, lo que entrega certeza. También se concluyo en 3.2 que la variable treat tenía una incidencia opacada por otras variables. Podemos atribuir que haya sido pequeña la variación a que el incremento en profesores también fue pequeño con un 0.04 porciento de profesores más por alumno, pero es un buen indicador que en ambas pruebas allá subido el puntaje promedio. Se puede concluir gracias a lo expuesto anteriormente que aumentar el número de profesores por estudiante aumenta el rendimiento en la prueba simce, pero que un incremento pequeño en profesores también otorgará un incremento pequeño en el resulado simce.