Tercer_TP_tagged

May 15, 2019

1 Respuesta 1

Como iniciación con los datos, comenzamos con ver que datos tenemos, mirando al menos los valores de estos enl las priemras filas, y tambien viendo el tipo de datos que tenemos, mas los estadísticos más basicos de estos. Recordamos aquí que la variable respuesta es el logaritmo del peso del bebé al nacer (`lbwght`).

[0]:
```
summary(iv_data)
```

```
     cigtax          cigprice        fatheduc        motheduc          parity
 Min.   : 2.0   Min.   :103.8   Min.   : 1.00   Min.   : 2.00   Min.   :1.000
 1st Qu.:15.0   1st Qu.:122.8   1st Qu.:12.00   1st Qu.:12.00   1st Qu.:1.000
 Median :20.0   Median :130.8   Median :12.00   Median :12.00   Median :1.000
 Mean   :19.6   Mean   :130.7   Mean   :13.19   Mean   :13.13   Mean   :1.614
 3rd Qu.:26.0   3rd Qu.:137.0   3rd Qu.:16.00   3rd Qu.:15.00   3rd Qu.:2.000
 Max.   :38.0   Max.   :152.5   Max.   :18.00   Max.   :18.00   Max.   :6.000
      male            white            lbwght          packs
 Min.   :0.0000   Min.   :0.0000   Min.   :3.135   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:4.682   1st Qu.:0.00000
 Median :1.0000   Median :1.0000   Median :4.787   Median :0.00000
 Mean   :0.5189   Mean   :0.8438   Mean   :4.768   Mean   :0.08846
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.883   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.0000   Max.   :5.602   Max.   :2.00000
    lfaminc
 Min.   :-0.6931
 1st Qu.: 2.9178
 Median : 3.3142
 Mean   : 3.2768
 3rd Qu.: 3.7495
 Max.   : 4.1744
```

A primera vista ya nos encontramos con algunas situaciones, como mínimo, destacables en los datos.

Vemos que por las columnas de educacion (`fatheduc` y `motheduc`), tenemos un rango de 2 a 18 donde el promedio y la mediana estan alrededor de 12 y 13. No sabemos exactamente a que corresponden estos valores, si a años equivalentes de estudios, o tal vez a alguna noción categorica de estudio, como así también podrían ser datos estrictamente ordinales de nivel alcanzado.

Algo similar ocurre con la variable dependiente `lbwght`. En principio contamos con que esta variable es el logaritmo del peso al nacer, posiblmente en unidades de libras. Pero si consideramos que el rango de esta variable está contenido dentro de `[3.1, 5.61]`, entonces tenemos, que, en libras, los recién nacidos de esta base están entre $ exp(3.1) = 22.19 $ y $ exp(5.61) = 273 $ libras, para toda la muestra. Sin saber las unidades exactas de estos valores, podría pensarse que esta varialbe tiene un rango poco común.

Explorando un poco más, los datos dejan algunas dudas sobre la mesa, como sobre qué relación existe entre los paquetes (`packs`) consumidos durante el embarazo y el precio de los cigarros. Vemos que existen varios casos para los cuales no se consumieron paquetes, pero existe un precio para los cigarros (consumidos?). Ver recuadro de abajo para notar esto.

Algo similar ocurre con los ingresos familiares, donde podemos tener ingresos menores a `1`, pero como no se aclara la unidad de estos, no queda claro si es más un error en la data.

Miramos la relacion entre las variables explicativas y la var dependiente.

En primer lugar analizamos el logaritmo del peso al nacer, pero separado por cada factor para las variables de `male` y `white`. Vemos que no parecen haber grandes diferencias en los niveles de los factores i.e. , mirando cada gráfico debajo por separado, notamos que la variabilidad de la variable de respuesta se ve modificada cuando comparamos sus valores dentro de cada nivel del factor, sin embargo la mediana comparativa no parece afectarse demasiado, así como tampoco sus cuartiles altos.

Es importante notar que a partir de estos graficos no se puede concluir que estas variables no pueden aportar información al modelo, si se utilizan como variables de control.

1.1 Matriz de correlaciones

Como era de esperar, vemos que existe una correlación negativa entre las variables de `packs` y la variable explicativa, mientras que el resto de las variables están correlacionadas positivamente.

Obviamente era de esperar que haya tanta correlación positiva entre el precio de los cigarrillos y los impuestos, y que tal vez no debamos utilizar las dos variables en el modelo, para no romper supuestos de multi-colinealidad.

[0]:
```
round(cor(iv_data[, std_cols] ), 2)
```

A matrix: 8 Œ 8 of type dbl

	cigtax_std	cigprice_std	fatheduc_std	motheduc_std	pari
cigtax_std	1.00	0.88	0.08	0.05	0.01
cigprice_std	0.88	1.00	0.09	0.06	-0.01
fatheduc_std	0.08	0.09	1.00	0.64	-0.05
motheduc_std	0.05	0.06	0.64	1.00	-0.10
parity_std	0.01	-0.01	-0.05	-0.10	1.00
lbwght_std	0.04	0.04	0.07	0.04	0.07
packs_std	0.02	0.01	-0.18	-0.22	0.04
lfaminc_std	0.03	0.11	0.41	0.40	-0.06

2 Respuesta 2

Primero ajustamos un modelo lineal por cuadrados mínimos del peso al nacer en función de la cantidad de paquetes por día consumidos por la madre durante el embarazo.

[0]:
```
modelo1 = lm(iv_data$lbwght ~ iv_data$packs)
```

Estudiamos los coeficientes obtenidos, su significación estadística junto con una medida general del modelo a partir de los resultados de la tabla siguiente:

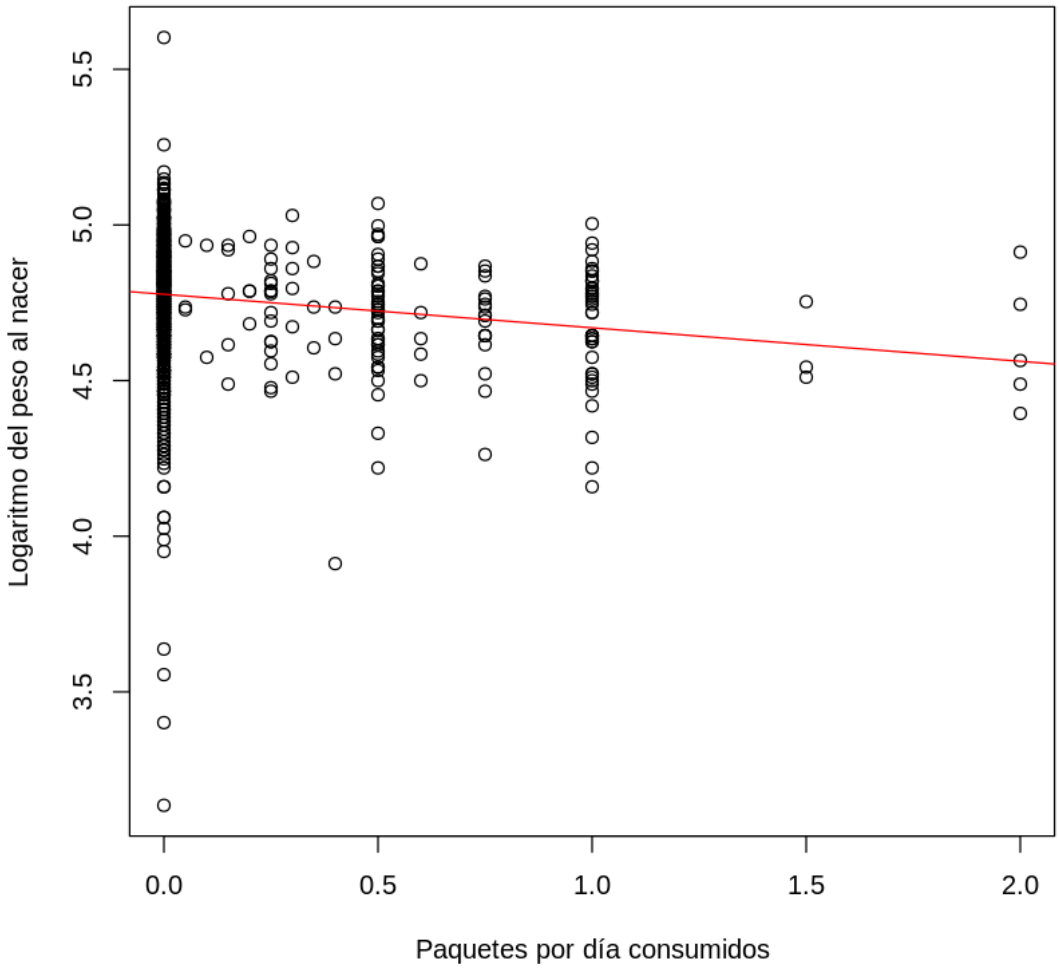
[0]:
```
summary(modelo1)
```

```
Call:
lm(formula = iv_data$lbwght ~ iv_data$packs)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64156 -0.08571  0.02168  0.11655  0.82506

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.777055   0.005674 841.946  < 2e-16 ***
iv_data$packs -0.107613   0.020167  -5.336 1.14e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1859 on 1189 degrees of freedom
Multiple R-squared:  0.02339,Adjusted R-squared:  0.02257
F-statistic: 28.47 on 1 and 1189 DF,  p-value: 1.136e-07
```

Vemos que la variable explicativa propuesta es relevante para determinar el peso del bebé al nacer ya que el coeficiente estimado es estadísticamente significativo. Sin embargo, analizando el coeficiente R^2 vemos que su valor es bajo.

Si graficamos ambas variables en un gráfico de dispersión junto con la recta obtenida mediante el ajuste lineal :

[0]:
```
plot(iv_data$packs, iv_data$lbwght,
     xlab = "Paquetes por día consumidos", ylab = "Logaritmo del peso al nacer")
abline(modelo1$coefficients, col="red")
```



Vemos que los residuos en este modelo dependen claramente de la variable explicativa, lo que nos da que la variable explicativa estaría rompiendo con el principio de exogeneidad. Luego no estamos cumpliendo con las hipotesis minimas del modelo lineal en este caso, y posiblemente así estemos omitiendo variables explicativas para este modelo.

Como segundo modelo, intentamos explicar la misma variable de interés, agregando el resto de las variables sugeridas

Obtenemos nuevamente que las variables explicativas son relevantes y el coeficiente R^2 aumenta en este caso. Sin embargo, sigue siendo bajo como para explicar la variabilidad de los datos.

3 Respuesta 3

La variable `packs` podría estar correlacionada con otros factores de la salud de la madre (y por consiguiente del bebé en gestación) no observados por el modelo . Esta correlación afectaría las estimaciones al no estar garantizada la exogeneidad, rompiendo con la consistencia del modelo tradicional.

Ejemplos de estas variables no observadas podrían ser situaciones como el estrés sufrido durante el embarazo, la edad de la madre al momento de la gestación (como factor de salud), o también por otras cuestiones demográficas de la madre.

4 Respuesta 4

Para validar si `cigprice` es un instrumento válida para `packs` esta variable debe ser exógeno (no predeterminado), para obtener que no está correlacionado con los demás no-observables.

A su vez la variable debe tener matriz de correlación de rango completa, con la variable de control (en este caso `packs`) .

Como punto a favor para `cigprice`, el precio de cada paquete está correlacionado con la cantidad de paquetes consumidos. Un factor en contra podría ser la posible correlación del precio de los cigarrillos con proteccionismo del estado debido a la importancia de los impuestos en el precio de este producto en particular entendidos como política pública.

5 Respuesta 5

La variable `cigprice` tiene sentido como instrumento y no como control de este estudio, debido a la pregunta que motiva el estudio. Cuando el investigador está interesado en estudiar el efecto de cierta variable endógena A en la variable respuesta B se considerará a A como una variable control y se buscarán variables instrumentales para garantizar la consistencia del modelo. Por el contrario, en caso de no tener interés en una variable A que sea exógena, esta se podrá considerar como variable instrumental, para ayudar con la consistencia del modelo.

[0]:
```
ivfit <- ivreg(lbwght ~ packs + male + lfaminc + parity| cigprice + male +␣
 ↪lfaminc + parity, data = iv_data)

summary(ivfit, df = Inf , diagnostics = TRUE)
```

```
Call:
ivreg(formula = lbwght ~ packs + male + lfaminc + parity | cigprice +
    male + lfaminc + parity, data = iv_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74202 -0.06786  0.06545  0.16988  0.90060

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.459462   0.344501  12.945   <2e-16 ***
packs       0.718020   1.298157   0.553    0.580
```

```
malemale     0.051280   0.032137   1.596    0.111
lfaminc      0.062245   0.072211   0.862    0.389
parity       0.008669   0.016078   0.539    0.590

Diagnostic tests:
                  df1  df2 statistic p-value
Weak instruments    1 1186     0.688   0.407
Wu-Hausman          1 1185     0.948   0.331
Sargan              0   NA        NA      NA

Residual standard error: 0.2845 on Inf degrees of freedom
Multiple R-Squared: -1.282,Adjusted R-squared: -1.29
Wald test: 10.75 on 4 DF,  p-value: 0.0295
```

6 Respuesta 6

Al comparar los resultados obtenidos entre el modelo tradicional de minimos cuadrados y el modelo de 2 pasos, incorporando la proyección de la variable endógena hacia el espacio de variables instrumentales (con `cigprice`), encontramos una gran diferencia el efecto de cantidad de packs sobre el peso del recién nacido. El cambio es tal que hay un cambio de signo en el coeficiente estimado y este deja de ser estadísticamente significativo.

[0]:
```
base_msg <-"Coeficientes para"
print(sprintf("%s MC ordinarios son: %s", base_msg,␣
 ↪paste(modelo2$coefficients,collapse=" ")))
print(sprintf("%s MC 2 pasos son: %s", base_msg,␣
 ↪paste(mc_2etapas2$coefficients, collapse=" ")))
print(sprintf("%s MC2E directo son: %s", base_msg, paste(ivfit$coefficients,␣
 ↪collapse=" ")))
```

```
[1] "Coeficientes para MC ordinarios son: 4.67508802892718 -0.10118362895115
0.0338913417855693 0.0172654709782682 0.0168782447746321"
[1] "Coeficientes para MC 2 pasos son: 4.45946217779144 0.718020461255999
0.0512798738939455 0.0622454118678936 0.00866886567873625"
[1] "Coeficientes para MC2E directo son: 4.45946217779145 0.718020461255977
0.0512798738939451 0.0622454118678924 0.00866886567873645"
```

7 Respuesta 7

En el primer paso de este modelo, incluimos todos los regresores salvo los endogenos i.e. la variable instrumental para el regresor endógeno mas el resto de los regresores, que son instrumentos en si mismos. La idea es lograr la proyección de la variable endógena a este nuevo espacio. Si no lo hacemos, entonces no estaríamos utilizando la información que nos proporcionan el resto de las variables exógenas, para explicar `packs`.

Como se ve en los resultados de la celda anterior, los coeficientes de los modelos tipo *MC2E* son equivalentes salvo alguna diferencia menor debido a redondeos numéricos. En cambio los errores estandar parecen disminuir.

8 Respuesta 8

El estadístico F sirve como base para comparar la hipotesis nula de que el modelo es mejor que el modelo "trivial" (solo intercepto) y no da información acerca de si hay correlaciones entre los regresores y los residuos.

En este caso vemos que el `F-statistic` en el primer paso de *MC2E* da un valor por debajo de 10 y también vemos que el coeficiente para `cigprice` no parece ser estadísticamente significativo, ayudando a cultivar la idea de que no es un buen instrumento, dado que buscamos ver que el instrumento tiene efecto sobre la variable endógena.

Si consideramos que el primer paso ya está sesgado, entonces tendremos valores sesgados para el modelo del segundo paso.

La afirmación sobre el valor del `F-statistic` menor a 10 no necesariamente es correcta pero su recíproca es mas importante para nostros, especialmente dado el caso de tener un solo instrumento. Como regla, si su valor es menor a 10 entonces estamos ante un indicio de un instrumento débil, ya que los valores de corte para los instrumentos están en general por encima del valor de corte del F-test.

9 Respuesta 9

Via un modelo sobre-identificado y vía *GMM* , `cigtax` también puede ser considerado como una variable instrumental si esgrimimos las mismas razones que antes para `cigprice`. Aquí creeemos que este instrumento es ortogonal con las variables no observadas y que no afectan a la variable explicada, si no a través de la cantidad de `packs` que consumen las embarazadas. En este caso podría justificarse mejor la exogeneidad debido a que se tiene en cuenta el proteccionismo del estado con respecto a la salud.

Como primer resultado notorio en la estimación vía *GMM*, los 3 modelos son muy similares en sus resultados, tanto como en el valor de los coeficientes así como también en la significancia de los coeficientes, que resultaron ser pobres según los t-estadísticos. Es posible creer que puedan haber relaciones entre los regresores y el peso del bebé al nacer que sean por razones azarosas y no por efectos reales.

[0]:
```
# como las VI deberian tener covarianza = 0 con los residulos del modelo lineal␣
 ↪original,
# en teoria el j-test de sobre-identificacion tiene como hip. nula de que los␣
 ↪coeficientes de las
# VI z1, z2 son nulos para un modelo de estimacion de residuos "originales",␣
 ↪dado
# una VI. Ergo esto es lo mismo que mirar los t-statistics y el coeficiente␣
 ↪para
# el siguiente modelo de regresar los residuos del modelo simple con los␣
 ↪instrumentos (incluye exogenas)
sobre_ident = lm(modelo2$residuals ~ iv_data$male + iv_data$lfaminc +␣
 ↪iv_data$parity + iv_data$cigprice + iv_data$cigtax +1 )
```

```
summary(sobre_ident)
```

```
Call:
lm(formula = modelo2$residuals ~ iv_data$male + iv_data$lfaminc +
    iv_data$parity + iv_data$cigprice + iv_data$cigtax + 1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63198 -0.08730  0.02007  0.11683  0.84532

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.0337084  0.1194681   0.282    0.778
iv_data$malemale -0.0004578  0.0107296  -0.043    0.966
iv_data$lfaminc   0.0002597  0.0076214   0.034    0.973
iv_data$parity   -0.0001760  0.0061269  -0.029    0.977
iv_data$cigprice -0.0004798  0.0011063  -0.434    0.665
iv_data$cigtax    0.0014631  0.0014489   1.010    0.313

Residual standard error: 0.1844 on 1185 degrees of freedom
Multiple R-squared:  0.001658,Adjusted R-squared:  -0.002555
F-statistic: 0.3935 on 5 and 1185 DF,  p-value: 0.8535
```

10 Respuesta 10

Para el modelo estimado via *GMM* robusto, y con los dos instrumentos antes mencionados más la variable endogena de packs, vemos que estamos en el caso de sobre-identificación.

La hipótesis nula en los J-tests de los modelos *GMM* no pueden ser rechazados, dando a entender que los modelos se encuentran bien calculados y no se encuentran mal-estimados ya que no parecería que hay efecto de las variables instrumentales sobre los residuos del modelo original.

[0]:
```
# nrow(iv_data) * summary(sobre_ident)$r.squared
sobre_ident_test = linearHypothesis(sobre_ident,
                                    c("iv_data$cigprice = 0", "iv_data$cigtax = 0"),
                                    test = "Chisq")
# sobre_ident_test
# computar el p-valor correcto del J-statistic
pchisq(sobre_ident_test[2, 5], df = 1, lower.tail = FALSE)
```

0.16070263946391

En adición, lo mismo puede decirse si calculamos a mano el estadístico y su significancia, vía el test de Hausman. Tenemos un 16% de probabilidad de ver un evento así con estos datos. Con lo cual parecen ser instrumentos válidos.

11 11 - Instrumentos débiles

Estudie la posibilidad de instrumentos débiles en el modelo sobre identificado

[0]:
```
# computamos el estadístico F robusto

linearHypothesis(mc_2etapas1,
                 "iv_data$cigprice = 0",
                 vcov = vcovHC, type = "HC1")
```

A anova: 2 Œ 4

Res.Df	Df	F	Pr(>F)
<dbl>	<dbl>	<dbl>	<dbl>
1187	NA	NA	NA
1186	1	0.6999517	0.4029685

12 Respuesta 11

Existen razones para sospechar de la fortaleza de estos instrumentos, específicamente al mirar los resultados de la primera etapa de *MC2E* donde estamos estimando la variable endógena vía los instrumentos.

Se ve que los los coeficientes de las VI propuestas no parecen ser muy diferentes de 0, y el estadístico t nos dice que no podemos rechazar la hipótesis nula de que sus coficientes sean valores nulos.

Del mismo modo, al mirar el estadístico F vemos que tenemos un valor de $7 < 10$ lo que da el indicio de que estos instrumentos son débiles para explicar la varianza de `packs`. Luego posiblmente estemos en una situación donde los coeficientes de `packs` para el modelo *MC2E* estén sesgados, dado que no tenemos instrumentos los suficientemente importantes.

[0]:
```
```