College Proximity: Instrumental Variables

Mariana Lugo Lizzy Gamboa

2022-11-20

Modelación en Ciencia de Datos

Variables Instrumentales

Para la elaboración de la práctica se utilizan las siguientes librerías:

```
install.packages("AER")
library(AER)
library(table1)
library(dplyr)
library(kableExtra)
library(ggplot2)
library(stargazer)
library(stats)
library(glmnet)
library(ivreg)
library(GGally)
```

1. Adjunte el paquete AER y cargue los datos de CollegeDistance.

Para el ejercicio se utilizan los datos de la encuesta *High School and Beyond* realizada por el Departamento de Educación en 1980, con seguimiento en 1986. La encuesta incluye a estudiantes de aproximadamente 1,100 preparatorias.

Se cargan los datos los cuales contienen 4,739 observaciones y 14 variables. Las variables se definen de la siguiente manera:

1. gender: variable categórica que indica el género.
2. ethnicity: variable categórica que indica la etnia (African-American, Hispanic or other).
3. score: variable numérica de la puntuación del año base de la prueba compuesta . Estas son pruebas de rendimiento que se dan a los estudiantes de último año de secundaria de la muestra.
4. fcollege: variable categórica que indica si el padre es graduado de college.
5. mcollege: variable categórica que indica si la madre es graduada de college.
6. home: variable categórica que indica si la familia es dueña de la casa.
7. urban: variable categórica que indica si la escuela está en una zona urbana.
8. unemp: variable numérica de la tasa de desempleo del condado en 1980.
9. wage: variable numérica del salario estatal por hora en la industria manufacturera en 1980.
10. distance: distancia a la 4-year colleg (en 10 millas).
11. tuition: colegiatura promedio de 4-year college estatal (en 1000 USD).
12. education:número de años de educación.

13. income: variable categórica que indica si el ingreso familiar es mayor a 25,000 USD al año.
14. region: variable categórica que indica la región (West u otra)

Se imprime una muestra de los datos:

gender	ethnicity	score	fcollege	mcollege	home	urban	unemp	wage	distance	tuition	education	income	region
male	other	39.15	yes	no	yes	yes	6.2	8.09	0.2	0.88915	12	high	other
female	other	48.87	no	no	yes	yes	6.2	8.09	0.2	0.88915	12	low	other
male	other	48.74	no	no	yes	yes	6.2	8.09	0.2	0.88915	12	low	other
male	afam	40.40	no	no	yes	yes	6.2	8.09	0.2	0.88915	12	low	other
female	other	40.48	no	no	no	yes	5.6	8.09	0.4	0.88915	13	low	other
male	other	54.71	no	no	yes	yes	5.6	8.09	0.4	0.88915	12	low	other

```
data(CollegeDistance)
kable(head(CollegeDistance),booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

2. Obtenga una descripción general del conjunto de datos. Específicamente, tome en cuenta que la variable distancia (la distancia a la escuela más cercana en 10 millas) servirá como instrumento en las estimaciones. Utilice un histograma para visualizar la distribución de la distancia.

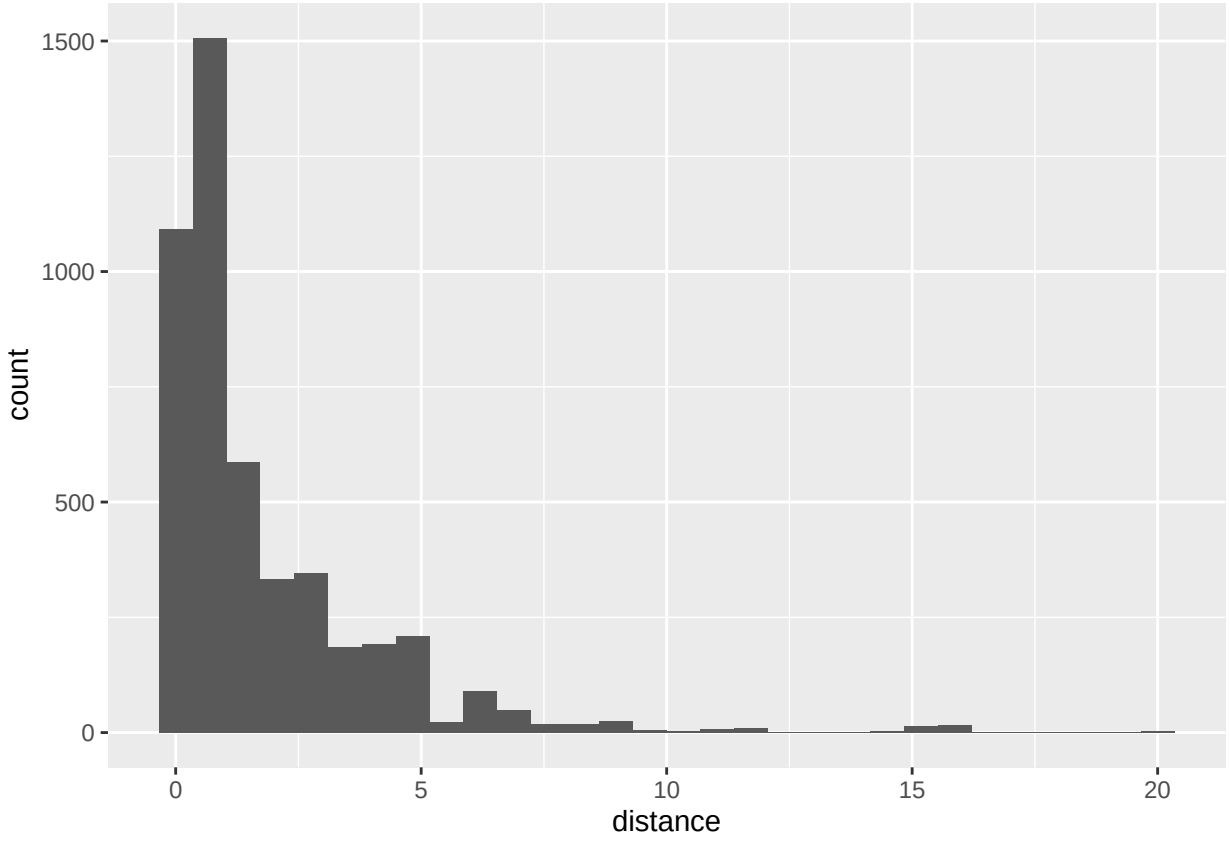
Se muestra la descripción general de las variables. Específicamente, se puede notar el alrededor del 77% de las escuelas no están en zonas urbanas y que la media de la distancia son 18 millas. Las familias son de ingresos bajos donde la mayoría de los alumnos tiene padres y madres que no tiene un título de college.

```
table1::table1(urban~ ., data = CollegeDistance)
```

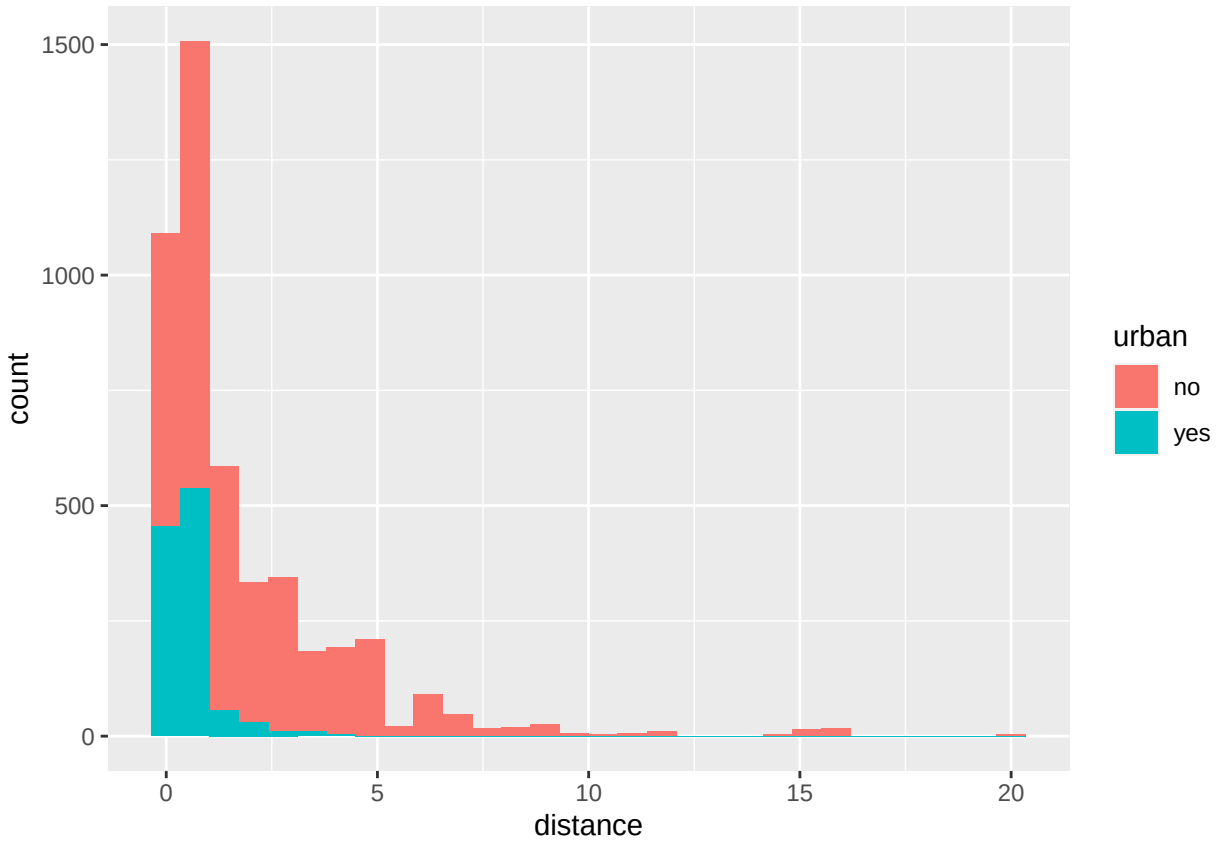
	Overall
	(N=4739)
urban	
no	3635 (76.7%)
yes	1104 (23.3%)
gender	
male	2139 (45.1%)
female	2600 (54.9%)
ethnicity	
other	3050 (64.4%)
afam	786 (16.6%)
hispanic	903 (19.1%)
score	
Mean (SD)	50.9 (8.70)
Median [Min, Max]	51.2 [29.0, 72.8]
fcollege	
no	3753 (79.2%)
yes	986 (20.8%)
mcollege	
no	4088 (86.3%)
yes	651 (13.7%)
home	
no	852 (18.0%)
yes	3887 (82.0%)
unemp	
Mean (SD)	7.60 (2.76)
Median [Min, Max]	7.10 [1.40, 24.9]
wage	
Mean (SD)	9.50 (1.34)
Median [Min, Max]	9.68 [6.59, 13.0]
distance	
Mean (SD)	1.80 (2.30)
Median [Min, Max]	1.00 [0, 20.0]
tuition	
Mean (SD)	0.815 (0.340)
Median [Min, Max]	0.824 [0.258, 1.40]
education	
Mean (SD)	13.8 (1.79)
Median [Min, Max]	13.0 [12.0, 18.0]
income	
low	3374 (71.2%)
high	1365 (28.8%)
region	
other	3796 (80.1%)
west	943 (19.9%)

Se muestra la distribución de la distancia general y separando por zona urbana con el histograma y una gŕafica de caja y brazos. La distribución general de la distancia está sesgada a la derecha. Además con la gráfica de caja y brazos podemos observar que aquellas escuelas en zona urbana tienen un menor promedio de distancia que las que nos están en un área urbana.

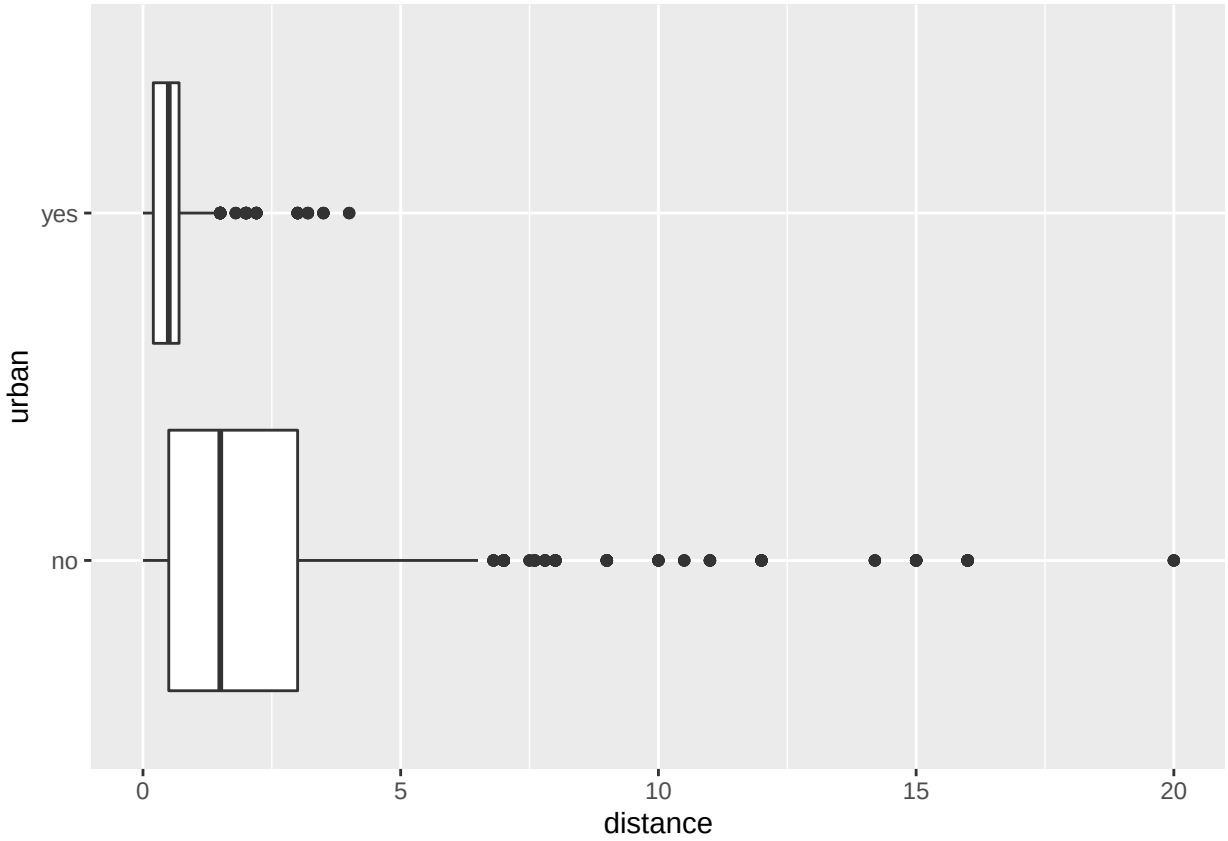
```
ggplot(CollegeDistance, aes(x = distance)) +
  geom_histogram()
```




```
ggplot(CollegeDistance, aes(x = distance, fill = urban)) +
  geom_histogram()
```




```
ggplot(CollegeDistance, aes(x = distance, y = urban)) +
geom_boxplot()
```



3. Estime dos regresiones cuyos resultados no son confiables debido al problema de selección. Guarde estos resultados para compararlos posteriormente con los obtenidos mediante el enfoque de variables instrumentales aplicado por Card (1993). Estime por ejemplo el modelo:

El modelo 1 se define como:

$$log|wage| = \beta_0 + \beta_1 log|education| + u$$

El modelo 2 se define como:

$$log|wage| = \beta_0 + \beta_1 education + \beta_2 ethnicity + \beta_3 unemp + \beta_4 gender + \beta_5 region + \beta_6 urban + u$$

```
m1<-lm(log(wage)~log(education), CollegeDistance)
stargazer(m1, type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                log(wage)
## ------------------------------------------------
## log(education)                 0.029*
##                                (0.016)
##
## Constant                       2.166***
##                                (0.043)
##
## ------------------------------------------------
## Observations                    4,739
## R2                              0.001
## Adjusted R2                     0.0004
## Residual Std. Error       0.144 (df = 4737)
## F Statistic             3.071* (df = 1; 4737)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
m2<-lm(log(wage)~education+ ethnicity+ unemp+ gender+region+urban, CollegeDistance)
stargazer(m2, type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                log(wage)
## ------------------------------------------------
## education                       0.001
##                                 (0.001)
##
## ethnicityafam                  -0.064***
##                                 (0.006)
##
## ethnicityhispanic              -0.049***
##                                 (0.005)
##
## unemp                          0.013***
##                                 (0.001)
##
## genderfemale                   -0.009**
##                                 (0.004)
##
## regionwest                     -0.021***
##                                 (0.005)
##
## urbanyes                        0.008*
```

```
##                                (0.005)
##
## Constant                      2.159***
##                                (0.017)
##
## ------------------------------------------------
## Observations                    4,739
## R2                              0.106
## Adjusted R2                     0.104
## Residual Std. Error       0.136 (df = 4731)
## F Statistic           79.892*** (df = 7; 4731)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

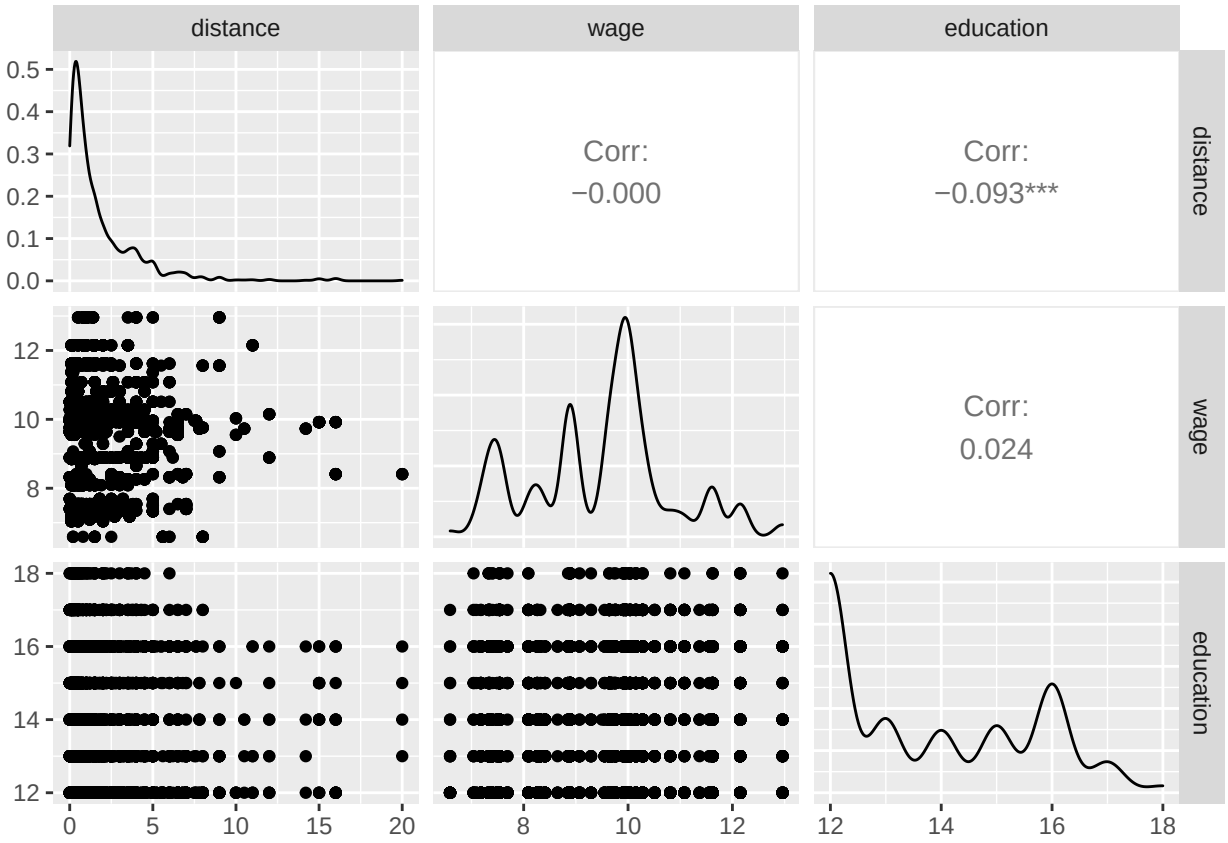
4. ¿Por qué se puede utilizar la distancia a la escuela como instrumento? Justifique el motivo por el cual esta variable podría ser utilizada como un instrumento.

Se utiliza la variable de distancia como instrumento para la educación ya que la distancia a una universidad con programas de 4 años estará correlacionada con la decisión de tener y/o terminar con el título o grado de college, pero puede ser que esta variable, per se, no sea una buena predictora del salario (fuera de que éste aumenta la educación de manera exógena). Por lo tanto, la distancia es un instrumento válido para la eduación. En su artículo, Card establece que aquellos alumnos que crecieron sin una universidad cerca enfrentan mayores costos en su educación, ya que la opción de vivir en casa se imposibilita. Mayores costos implican una menor inversión en educación de mayor calidad, sobretodo para alumnos de familias de ingresos menores.

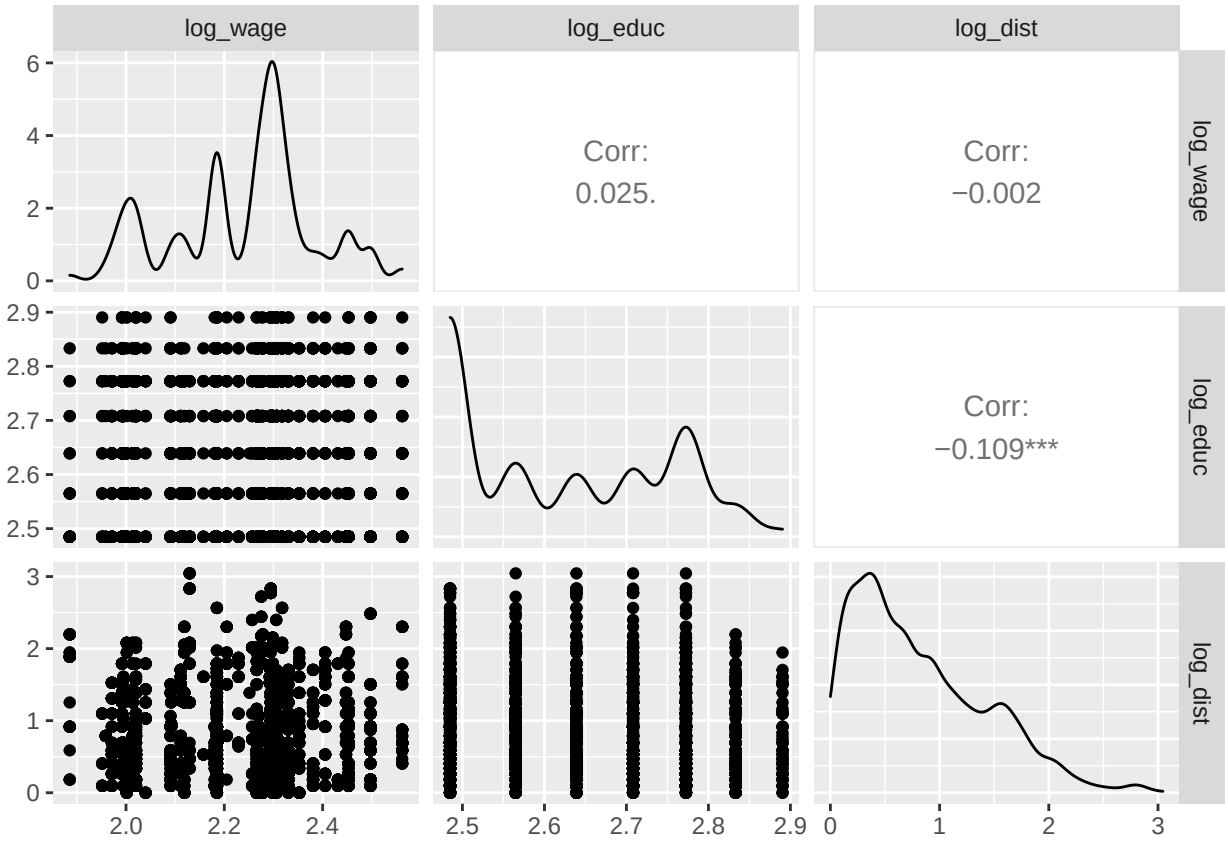
5. Calcule las correlaciones de la distancia del instrumento con la educación regresora endógena y la variable dependiente salario. ¿Qué parte de la variación en la educación se explica por la regresión de la primera etapa que utiliza la distancia como regresor?

Se muestra las correlaciones de las variables mencionadas. Se puede observar que existen una correlación entre la distancia y la variable de educación. La correlación se mantien al aplicar logarítmos a las variables. Además, la variación explicada por la distancia en el en la educación es la R^2, es decir 0.008683088.

```
cor_vars<- CollegeDistance %>% select(distance, wage, education)
ggpairs(cor_vars)
```




```
log_cor_vars<- CollegeDistance %>% mutate(log_wage=log(wage),
                                          log_educ=log(education),
                                          log_dist=log(distance+1)) %>%
                                          select(c(log_wage, log_educ, log_dist))
ggpairs(log_cor_vars)
```




```
R2 <-summary(lm(education ~ distance, data = CollegeDistance))$r.squared
print(R2)
```

```
## [1] 0.008683088
```

6. Repita las estimaciones anteriores utilizando IV, es decir, utilice la distancia como un instrumento para la educación en ambas regresiones mediante la función ivreg(). Guarde sus resultados y obtenga los errores estándar robustos para ambos modelos.

Se corren los dos modelos con la función *ivreg()* :

```
m3 <- ivreg(log(wage) ~ education | distance, data = CollegeDistance)


m4<-ivreg(log(wage) ~ unemp + ethnicity + gender + urban + education + region | . - education + distanc
stargazer(m3, m4,type="text")
```

```
##
## ===================================================
##                             Dependent variable:
##                         -----------------------------
##                                   log(wage)
##                             (1)               (2)
## ---------------------------------------------------
```

```
## unemp                                      0.014***
##                                             (0.001)
##
## ethnicityafam                              -0.030***
##                                             (0.011)
##
## ethnicityhispanic                          -0.031***
##                                             (0.008)
##
## genderfemale                                -0.007
##                                             (0.005)
##
## urbanyes                                     0.006
##                                             (0.006)
##
## education                    0.001          0.066***
##                             (0.013)         (0.014)
##
## regionwest                                  -0.013*
##                                             (0.007)
##
## Constant                   2.221***         1.242***
##                             (0.173)         (0.203)
##
## ------------------------------------------------------
## Observations                 4,739            4,739
## R2                           0.001           -0.546
## Adjusted R2                 0.0004           -0.548
## Residual Std. Error 0.144 (df = 4737) 0.179 (df = 4731)
## ======================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```

Se obtienen los errores estándar robustos para ambos modelos:

```
coeftest(m3, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2212812  0.1603013 13.8569   <2e-16 ***
## education   0.0014406  0.0116057  0.1241   0.9012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(m4, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##                     Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)        1.24217641  0.20046744  6.1964 6.267e-10 ***
```

```
## unemp              0.01408441  0.00096071 14.6604 < 2.2e-16 ***
## ethnicityafam     -0.02986152  0.01015586 -2.9403 0.0032946 **
## ethnicityhispanic -0.03145030  0.00811982 -3.8733 0.0001088 ***
## genderfemale      -0.00741953  0.00521274 -1.4233 0.1547021
## urbanyes           0.00587934  0.00627821  0.9365 0.3490807
## education          0.06577026  0.01427816  4.6064 4.205e-06 ***
## regionwest        -0.01256492  0.00651871 -1.9275 0.0539752 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Verifique que sus resultados coinciden con los obtenidos cuando utiliza el procedimiento de estimación mediante las dos regresiones de MC2E para ambos modelos.

Se crea la función para la regresión del estimador de mínimos cuadrados en dos etapas (MC2E). Se muestra que los resultados coinciden con las estimaciones del inciso anterior.

```
MC2E <- function(Y, X, W = NULL, Z, data) {
 # regresión de la primera etapa y valores estimados
 fs_model <- lm(as.formula(paste(X, "~", paste(c(Z, W), collapse = "+"))), data = data)
 X_fitted <- fs_model$fitted.values

 # regresión de la segunda etapa
 ss_model <- lm(as.formula(paste(Y, "~", paste(W, collapse = "+"), "+ X_fitted")),  data = data)

 # coeficientes de la segunda etapa
 return(
   coefficients(ss_model)
 )}


kable(MC2E(Y = "log(wage)", X = "education", Z = "distance", data = CollegeDistance))
```

	x
(Intercept)	2.2212812
X_fitted	0.0014406

```
kable(MC2E(Y = "log(wage)",
     X = "education",
     W = c("unemp", "ethnicity", "gender", "urban", "region"),
     Z = "distance",
     data = CollegeDistance))
```

	x
(Intercept)	1.2421764
unemp	0.0140844
ethnicityafam	-0.0298615
ethnicityhispanic	-0.0314503
genderfemale	-0.0074195
urbanyes	0.0058793
regionwest	-0.0125649
X_fitted	0.0657703

8. Interprete sus resultados en el contexto del problema, esto es, cuáles son los rendimientos de la educación estimados y el efecto del sesgo de selección en el modelo.

La regresión del MC2E considera otras variables demográficas como control que son significativas para el salario (definido como los rendimientos de la educación) además de los años de educación. Éstas son:

- La tasa de desempleo en el condado tiene una relación positiva con el salario: entre mayor tasa, mayor es es el salario. La interpretación es que condados con mayores tasas de desempleo pueden ser indicativos de una mayor escasez de oferta laboral, por lo tanto se refleja en mayores salarios.
- La etnia tiene una relación negativa con el salario cuando son afroamericanos o hispanos.
- La región oeste tiene una relación negativa con el salario.
- La educación resultó significativa, con una relación positiva después de los controles y quitando el sesgo por la distancia. Sin embargo, aún quitando el sesgo de la distancia y considerando las otras variables de control, la educación tiene el efecto de que por un año adicional de educación se espera que aumente el ingreso por aproximadamente 6.7%.

9. Realice las pruebas de verificación de variables instrumentales (endogeneidad de la regresora, relevancia del instrumento y exogeneidad del instrumento).

Para poder hacer las pruebas de *endogeneidad** y *exogeneidad** necesitamos las pruebas de Hausman y Sargan respectivamente, esto podemos observarlo en los siguientes resultados:

Donde Hausman nos dice que existe endogeneidad:

```
summary(m4, diagnostics = T)
```

```
##
## Call:
## ivreg(formula = log(wage) ~ unemp + ethnicity + gender + urban +
##     education + region | . - education + distance, data = CollegeDistance)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.585984 -0.117674  0.001466  0.140398  0.464518
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.2421764  0.2030180   6.119 1.02e-09 ***
## unemp              0.0140844  0.0009601  14.669  < 2e-16 ***
## ethnicityafam     -0.0298615  0.0105535  -2.830  0.00468 **
## ethnicityhispanic -0.0314503  0.0080279  -3.918 9.07e-05 ***
## genderfemale      -0.0074195  0.0052321  -1.418  0.15623
## urbanyes           0.0058793  0.0063263   0.929  0.35275
## education          0.0657703  0.0144298   4.558 5.30e-06 ***
## regionwest        -0.0125649  0.0069397  -1.811  0.07027 .
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    1 4731     48.96 2.98e-12 ***
## Wu-Hausman          1 4730     35.94 2.18e-09 ***
## Sargan              0   NA        NA       NA
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1786 on 4731 degrees of freedom
## Multiple R-Squared: -0.546,  Adjusted R-squared: -0.5483
## Wald test: 49.16 on 7 and 4731 DF,  p-value: < 2.2e-16
```

```
m7 <- lm(residuals(m4) ~ education + ethnicity + unemp + gender +
           region + urban + distance, data = CollegeDistance)


Chi_test <- linearHypothesis(m7,
                             "distance = 0",
                             test = "Chisq")
Chi_test
```

```
## Linear hypothesis test
##
## Hypothesis:
## distance = 0
##
## Model 1: restricted model
## Model 2: residuals(m4) ~ education + ethnicity + unemp + gender + region +
##     urban + distance
##
##   Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1   4731 87.338
## 2   4730 86.679  1   0.65869 35.944  2.031e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos ver los grados de libertad es igual a 1, sin embargo deberiamos de tener $k-1$ en este caso

```
pchisq(Chi_test[2,5], df = 0, lower.tail = FALSE)
```

```
## [1] 0
```

Rechazamos la hipótesis de que ambos instrumentos son exógenos.

Para verificar la **relevancia** de un instrumento se debe usar la prueba F, si se tiene más de un instrumento. Para el caso de usar solo la distancia como instrumento, es suficiente con la regresión, es decir, el parametro que mide la relación del coeficiente, de igual manera hacermos el ejercicio:

```
m6 <- lm(formula = log(wage) ~ education + distance, data = CollegeDistance)

linearHypothesis(m6,
                 "distance = 0",
                 vcov = vcovHC, type = "HC1")
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```
## distance = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ education + distance
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1   4737
## 2   4736  1 0.0026 0.9596
```

```
m5 <- lm(formula = log(wage) ~ education + ethnicity + unemp + gender +
           region + urban + distance, data = CollegeDistance)

linearHypothesis(m5,
                 "distance = 0",
                 vcov = vcovHC, type = "HC1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## distance = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ education + ethnicity + unemp + gender + region +
##     urban + distance
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F    Pr(>F)
## 1   4731
## 2   4730  1 36.6 1.562e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como se ha visto en la significancia de los modelos programados anteriormente, la variable distancia con la educación es un instrumento relevante cuando se consideran otras variables dentro del modelo.

10. Pruebe si su instrumento es débil. Tome en cuenta que las familias con un fuerte interés por la educación pueden mudarse a vivir cerca de las universidades. Además, los vecindarios cercanos a las universidades pueden tener mercados laborales más fuertes reflejados en ingresos más altos. Tales características pueden invalidar el instrumento, ya que introducen variables no observadas que influyen en los ingresos, pero que no pueden ser capturadas por años de escolaridad, la medida de educación.

En este caso con las pruebas anteriores, podemos decir que nuestro instrumento es débil y los estimadores estan sesgados, por lo que nuestras estimaciones pueden ser poco confiables.

Bibliografía

- Card, David, (1993), Using Geographic Variation in College Proximity to Estimate the Return to Schooling, No 4483, NBER Working Papers, National Bureau of Economic Research, Inc.

- Introduction to Econometrics with R: https://www.econometrics-with-r.org/12-ivr.html