**6.3.****X Implementación de Variables Instrumentales en Stata**

Para el ejemplo de la implementación de variables instrumentales en Stata se utilizará el ejemplo del programa *canasta.* En este caso, nuevamente se quiere estimar el impacto de este programa sobre la talla para la edad de los participantes:

(1)

donde es el indicador de participación en el programa es la variable de resultado y *X* es un vector de características observables del individuo que determinan *Y*. El problema al que nos enfrentamos es que se viola el supuesto de independencia condicional, es decir, hay una relación entre la probabilidad de participar en el programa y algunas características no observables que también pueden afectar la variable de interés:

(2)

Esto puede suceder porque las madres con mayor capital social son las que más conocimiento tienen acerca de los alimentos que deben ser proporcionados a los niños y, así mismo, son estas madres las que más probabilidad tienen de enterarse de la existencia de un programa como *canasta*. Haciendo la estimación por mínimos cuadrados ordinarios (MCO) concluiríamos que el programa incrementa la talla para la edad en 0.2489 desviaciones estándar.

|  |
| --- |
| Resultado 1:  Source | SS df MS Number of obs = 4000  -------------+------------------------------ F( 6, 3993) = 39.49  Model | 68.9063178 6 11.4843863 Prob > F = 0.0000  Residual | 1161.23317 3993 .290817222 R-squared = 0.0560  -------------+------------------------------ Adj R-squared = 0.0546  Total | 1230.13949 3999 .307611775 Root MSE = .53927  ------------------------------------------------------------------------------  ha\_nchs | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  D | **.2489481** .0171935 14.48 0.000 .2152392 .282657  personas | -.0093891 .004544 -2.07 0.039 -.018298 -.0004802  orden\_n | -.0632222 .0263134 -2.40 0.016 -.1148111 -.0116333  ocupado\_jefe | -.0364222 .0223229 -1.63 0.103 -.0801875 .0073431  educa\_jefe | .005974 .002325 2.57 0.010 .0014157 .0105322  ingresos\_h~e | .0001628 .000101 1.61 0.107 -.0000353 .0003609  \_cons | -.0840199 .0452452 -1.86 0.063 -.1727257 .004686  ------------------------------------------------------------------------------ |

Sin embargo, este resultado puede también deberse al hecho de que las madres que participan en el programa tienen mayor capital social y, por las razones anteriormente explicadas, tienen niños con una talla para la edad superior. Dado que el estimador de MCO es:

(3)

y hay una relación positiva entre los no observables y la probabilidad de ser tratado (, entonces el estimador de MCO estaría sobreestimando el impacto del programa (.

Para solucionar este problema, se utilizan dos instrumentos. El primero de ellos es la distancia entre el hogar y el centro operador de este programa. Es razonable suponer que hay una relación entre la probabilidad de participar en el programa y la distancia dado que entre menor sea ésta, menor el costo en transporte y tiempo de participar y por lo tanto mayor la probabilidad de hacer parte de *canasta.* El otro instrumento utilizado es el número de oficinas operadoras en el municipio. Se utiliza porque es razonable suponer que entre mayor oficinas operadoras del programa en el municipio, mayor conocimiento tienen los habitantes sobre el programa. Además, estos instrumentos se pueden considerar exógenos dado que no hay razón aparente para suponer que hogares con menor distancia al centro operador, o residentes en municipios con gran cantidad de oficinas operadoras, tienen una mayor talla para la edad. Para formalizar estas hipótesis es necesario hacer las pruebas de relevancia y exogeneidad de los instrumentos.

**6.3.X.1 Prueba de relevancia**

Lo que hacemos para evaluar la relevancia de los instrumentos es estimar una regresión donde la variable dependiente es el indicador de tratamiento y los regresores son los instrumentos y las otras variables explicativas:

|  |
| --- |
| Resultado 2:  . reg D of\_op distancia  Source | SS df MS Number of obs = 4000  -------------+------------------------------ F( 2, 3997) = 64.00  Model | 31.0125359 2 15.506268 Prob > F = 0.0000  Residual | 968.411464 3997 .242284579 R-squared = 0.0310  -------------+------------------------------ Adj R-squared = 0.0305  Total | 999.424 3999 .24991848 Root MSE = .49222  ------------------------------------------------------------------------------  D | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  of\_op | **.033219** .0037878 8.77 0.000 .0257928 .0406453  distancia | **-.0000486** 6.97e-06 -6.98 0.000 -.0000623 -.000035  \_cons | .4983069 .0191481 26.02 0.000 .4607659 .5358478  ------------------------------------------------------------------------------  .  . reg D $X of\_op distancia  Source | SS df MS Number of obs = 4000  -------------+------------------------------ F( 7, 3992) = 27.95  Model | 46.689176 7 6.66988229 Prob > F = 0.0000  Residual | 952.734824 3992 .238661028 R-squared = 0.0467  -------------+------------------------------ Adj R-squared = 0.0450  Total | 999.424 3999 .24991848 Root MSE = .48853  ------------------------------------------------------------------------------  D | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  personas | -.0213117 .0041031 -5.19 0.000 -.029356 -.0132673  orden\_n | .0972524 .0238059 4.09 0.000 .0505796 .1439253  ocupado\_jefe | .0753753 .0201885 3.73 0.000 .0357946 .1149561  educa\_jefe | -.0057376 .0021042 -2.73 0.006 -.0098631 -.0016121  ingresos\_h~e | .0003076 .0000914 3.37 0.001 .0001284 .0004868  of\_op | **.0332178** .0037605 8.83 0.000 .0258451 .0405905  distancia | **-.0000487** 6.92e-06 -7.05 0.000 -.0000623 -.0000352  \_cons | .4527444 .0438179 10.33 0.000 .3668368 .538652  . test distancia=of\_op=0  ( 1) - of\_op + distancia = 0  ( 2) distancia = 0  F( 2, 3992) = 65.00  Prob > F = 0.0000 |

Dado que encontramos que los coeficientes asociados a los instrumentos son estadísticamente significativos, hay evidencia de que existe una relación estadísticamente significativa entre el número de oficinas operadoras en el municipio y la distancia del hogar al centro operador, y la probabilidad de participar en el tratamiento.

**6.3.X.2 Prueba de exogeneidad**

El segundo requisito que debe cumplir un instrumento es el de exogeneidad. Esto es, que los instrumentos no estén relacionados con las variables no observables. Si esto no se cumple, los estimadores de variables instrumentales no serían consistentes ni insesgados. La prueba de exogeneidad que se presenta en este caso consiste en estimar los residuales de la ecuación (1):

(4)

y, posteriormente, evaluar si existe alguna relación estadísticamente significativa entre estos residuales y los instrumentos a utilizar. Esto se hace simplemente estimando una regresión por MCO donde la variable dependiente está compuesta por los residuales estimados y los regresores son los instrumentos:

. (5)

Si los coeficientes de la matriz son estadísticamente diferentes de cero, existe evidencia de que hay una relación entre los instrumentos y las variables no observables. Por lo tanto, lo ideal para garantizar que los instrumentos son exógenos es no rechazar la hipótesis nula de que los coeficientes de la matriz son iguales a cero. En el ejemplo del programa *canasta*:

|  |
| --- |
| Resultado 3:  . reg ha\_nchs D $X  Source | SS df MS Number of obs = 4000  -------------+------------------------------ F( 6, 3993) = 39.49  Model | 68.9063178 6 11.4843863 Prob > F = 0.0000  Residual | 1161.23317 3993 .290817222 R-squared = 0.0560  -------------+------------------------------ Adj R-squared = 0.0546  Total | 1230.13949 3999 .307611775 Root MSE = .53927  ------------------------------------------------------------------------------  ha\_nchs | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  D | .2489481 .0171935 14.48 0.000 .2152392 .282657  personas | -.0093891 .004544 -2.07 0.039 -.018298 -.0004802  orden\_n | -.0632222 .0263134 -2.40 0.016 -.1148111 -.0116333  ocupado\_jefe | -.0364222 .0223229 -1.63 0.103 -.0801875 .0073431  educa\_jefe | .005974 .002325 2.57 0.010 .0014157 .0105322  ingresos\_h~e | .0001628 .000101 1.61 0.107 -.0000353 .0003609  \_cons | -.0840199 .0452452 -1.86 0.063 -.1727257 .004686  ------------------------------------------------------------------------------  .  . predict uhat, residuals  . reg uhat of\_op distancia  Source | SS df MS Number of obs = 4000  -------------+------------------------------ F( 2, 3997) = 0.26  Model | .151132554 2 .075566277 **Prob > F = 0.7710**  Residual | 1161.08204 3997 .290488376 R-squared = 0.0001  -------------+------------------------------ Adj R-squared = -0.0004  Total | 1161.23317 3999 .290380888 Root MSE = .53897  ------------------------------------------------------------------------------  uhat | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  of\_op | **.0002942** .0041475 0.07 0.943 -.0078373 .0084257  distancia | **5.48e-06** 7.63e-06 0.72 0.472 -9.47e-06 .0000204  \_cons | -.0118993 .0209666 -0.57 0.570 -.0530055 .0292069  ------------------------------------------------------------------------------ |

Vemos que los instrumentos no tienen relación estadísticamente significativa con los residuales. Ninguna variable es significativa y el modelo en conjunto tampoco lo es, lo que nos indica que es probable que los instrumentos sean exógenos.

**6.3.X.3 Mínimos cuadrados ordinarios en dos etapas**

En este caso hacemos variables instrumentales pero mostramos paso por paso. Lo primero que se debe hacer es hacer una estimación de las probabilidades de participar con base en los regresores y los instrumentos. Luego, las probabilidades predichas se utilizan como instrumento de la participación en el programa:

|  |
| --- |
| Resultado 4:  . probit D $X distancia of\_op  Iteration 0: log likelihood = -2771.4366  Iteration 1: log likelihood = -2675.7646  Iteration 2: log likelihood = -2675.6035  Iteration 3: log likelihood = -2675.6035  Probit regression Number of obs = 4000  LR chi2(7) = 191.67  Prob > chi2 = 0.0000  Log likelihood = -2675.6035 Pseudo R2 = 0.0346  ------------------------------------------------------------------------------  D | Coef. Std. Err. z P>|z| [95% Conf. Interval]  -------------+----------------------------------------------------------------  personas | -.056137 .0108297 -5.18 0.000 -.0773628 -.0349112  orden\_n | .2556066 .0626281 4.08 0.000 .1328577 .3783555  ocupado\_jefe | .1947698 .0528018 3.69 0.000 .0912802 .2982595  educa\_jefe | -.015059 .0054907 -2.74 0.006 -.0258205 -.0042975  ingresos\_h~e | .0008236 .0002426 3.40 0.001 .0003482 .0012991  distancia | -.0001267 .0000181 -7.00 0.000 -.0001622 -.0000912  of\_op | .0867556 .0099521 8.72 0.000 .0672497 .1062614  \_cons | -.1219782 .1145348 -1.06 0.287 -.3464623 .1025059  ------------------------------------------------------------------------------  .  . predict phat  . ivreg2 ha\_nchs $X (D=phat), first  First-stage regressions  -----------------------  First-stage regression of D:  OLS estimation  --------------  Estimates efficient for homoskedasticity only  Statistics consistent for homoskedasticity only  Number of obs = 4000  F( 6, 3993) = 32.69  Prob > F = 0.0000  Total (centered) SS = 999.424 Centered R2 = 0.0468  Total (uncentered) SS = 1952 Uncentered R2 = 0.5120  Residual SS = 952.6320316 Root MSE = .4884  ------------------------------------------------------------------------------  D | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  personas | .0000419 .0045089 0.01 0.993 -.0087981 .0088819  orden\_n | -.000604 .0250734 -0.02 0.981 -.0497619 .0485539  ocupado\_jefe | .0006032 .0213597 0.03 0.977 -.0412737 .0424801  educa\_jefe | -2.57e-06 .002166 -0.00 0.999 -.0042491 .004244  ingresos\_h~e | -3.32e-07 .0000951 -0.00 0.997 -.0001868 .0001861  trat\_hat | 1.000707 .087604 11.42 0.000 .8289546 1.17246  \_cons | -.0004202 .0563133 -0.01 0.994 -.1108256 .1099852  ------------------------------------------------------------------------------  Included instruments: personas orden\_n ocupado\_jefe educa\_jefe  ingresos\_hogar\_jefe trat\_hat  ------------------------------------------------------------------------------  Partial R-squared of excluded instruments: 0.0316  Test of excluded instruments:  F( 1, 3993) = 130.49  Prob > F = 0.0000  Summary results for first-stage regressions  -------------------------------------------  Variable | Shea Partial R2 | Partial R2 | F( 1, 3993) P-value  D | 0.0316 | 0.0316 | 130.49 0.0000  Underidentification tests  Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)  Ha: matrix has rank=K1 (identified)  Anderson canon. corr. N\*CCEV LM statistic Chi-sq(1)=126.58 P-val=0.0000  Cragg-Donald N\*CDEV Wald statistic Chi-sq(1)=130.72 P-val=0.0000  Weak identification test  Ho: equation is weakly identified  Cragg-Donald Wald F-statistic 130.49  See main output for Cragg-Donald weak id test critical values  Weak-instrument-robust inference  Tests of joint significance of endogenous regressors B1 in main equation  Ho: B1=0 and overidentifying restrictions are valid  Anderson-Rubin Wald test F(1,3993)=4.92 P-val=0.0266  Anderson-Rubin Wald test Chi-sq(1)=4.93 P-val=0.0264  Stock-Wright LM S statistic Chi-sq(1)=4.92 P-val=0.0265  Number of observations N = 4000  Number of regressors K = 7  Number of instruments L = 7  Number of excluded instruments L1 = 1  IV (2SLS) estimation  --------------------  Estimates efficient for homoskedasticity only  Statistics consistent for homoskedasticity only  Number of obs = 4000  F( 6, 3993) = 5.41  Prob > F = 0.0000  Total (centered) SS = 1230.139487 Centered R2 = 0.0553  Total (uncentered) SS = 1243.678874 Uncentered R2 = 0.0656  Residual SS = 1162.070617 Root MSE = .539  ------------------------------------------------------------------------------  ha\_nchs | Coef. Std. Err. z P>|z| [95% Conf. Interval]  -------------+----------------------------------------------------------------  D | .**2197716** .0966029 2.27 0.023 .0304333 .4091098  personas | -.0100115 .0049739 -2.01 0.044 -.0197602 -.0002628  orden\_n | -.0606025 .0276502 -2.19 0.028 -.1147959 -.0064091  ocupado\_jefe | -.0340712 .0235896 -1.44 0.149 -.0803061 .0121636  educa\_jefe | .0058022 .0023902 2.43 0.015 .0011175 .010487  ingresos\_h~e | .0001716 .0001049 1.64 0.102 -.0000341 .0003772  \_cons | -.0709511 .0621135 -1.14 0.253 -.1926914 .0507892  ------------------------------------------------------------------------------  Underidentification test (Anderson canon. corr. LM statistic): 126.579  Chi-sq(1) P-val = 0.0000  ------------------------------------------------------------------------------  Weak identification test (Cragg-Donald Wald F statistic): 130.487  Stock-Yogo weak ID test critical values: 10% maximal IV size 16.38  15% maximal IV size 8.96  20% maximal IV size 6.66  25% maximal IV size 5.53  Source: Stock-Yogo (2005). Reproduced by permission.  ------------------------------------------------------------------------------  Sargan statistic (overidentification test of all instruments): 0.000  (equation exactly identified)  ------------------------------------------------------------------------------  Instrumented: D  Included instruments: personas orden\_n ocupado\_jefe educa\_jefe  ingresos\_hogar\_jefe  Excluded instruments: phat  ------------------------------------------------------------------------------ |

**6.3.X.4 Variables instrumentales en Stata**

El comando en Stata para estimar un modelo mediante variables instrumentales es *ivreg2.* A continuación se muestra cómo se utiliza este comando en el contexto del programa *canasta:*

|  |
| --- |
| Resultado 5:  . ivreg2 ha\_nchs $X (D=distancia of\_op), first  First-stage regressions  -----------------------  First-stage regression of D:  OLS estimation  --------------  Estimates efficient for homoskedasticity only  Statistics consistent for homoskedasticity only  Number of obs = 4000  F( 7, 3992) = 27.95  Prob > F = 0.0000  Total (centered) SS = 999.424 Centered R2 = 0.0467  Total (uncentered) SS = 1952 Uncentered R2 = 0.5119  Residual SS = 952.734824 Root MSE = .4885  ------------------------------------------------------------------------------  D | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  personas | -.0213117 .0041031 -5.19 0.000 -.029356 -.0132673  orden\_n | .0972524 .0238059 4.09 0.000 .0505796 .1439253  ocupado\_jefe | .0753753 .0201885 3.73 0.000 .0357946 .1149561  educa\_jefe | -.0057376 .0021042 -2.73 0.006 -.0098631 -.0016121  ingresos\_h~e | .0003076 .0000914 3.37 0.001 .0001284 .0004868  distancia | -.0000487 6.92e-06 -7.05 0.000 -.0000623 -.0000352  of\_op | .0332178 .0037605 8.83 0.000 .0258451 .0405905  \_cons | .4527444 .0438179 10.33 0.000 .3668368 .538652  ------------------------------------------------------------------------------  Included instruments: personas orden\_n ocupado\_jefe educa\_jefe  ingresos\_hogar\_jefe distancia of\_op  ------------------------------------------------------------------------------  Partial R-squared of excluded instruments: 0.0315  Test of excluded instruments:  F( 2, 3992) = 65.00  Prob > F = 0.0000  Summary results for first-stage regressions  -------------------------------------------  Variable | Shea Partial R2 | Partial R2 | F( 2, 3992) P-value  D | 0.0315 | 0.0315 | 65.00 0.0000  Underidentification tests  Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)  Ha: matrix has rank=K1 (identified)  Anderson canon. corr. N\*CCEV LM statistic Chi-sq(2)=126.16 P-val=0.0000  Cragg-Donald N\*CDEV Wald statistic Chi-sq(2)=130.27 P-val=0.0000  Weak identification test  Ho: equation is weakly identified  Cragg-Donald Wald F-statistic 65.00  See main output for Cragg-Donald weak id test critical values  Weak-instrument-robust inference  Tests of joint significance of endogenous regressors B1 in main equation  Ho: B1=0 and overidentifying restrictions are valid  Anderson-Rubin Wald test F(2,3992)=2.42 P-val=0.0895  Anderson-Rubin Wald test Chi-sq(2)=4.84 P-val=0.0889  Stock-Wright LM S statistic Chi-sq(2)=4.83 P-val=0.0892  Number of observations N = 4000  Number of regressors K = 7  Number of instruments L = 8  Number of excluded instruments L1 = 2  IV (2SLS) estimation  --------------------  Estimates efficient for homoskedasticity only  Statistics consistent for homoskedasticity only  Number of obs = 4000  F( 6, 3993) = 5.33  Prob > F = 0.0000  Total (centered) SS = 1230.139487 Centered R2 = 0.0548  Total (uncentered) SS = 1243.678874 Uncentered R2 = 0.0651  Residual SS = 1162.702759 Root MSE = .5391  ------------------------------------------------------------------------------  ha\_nchs | Coef. Std. Err. z P>|z| [95% Conf. Interval]  -------------+----------------------------------------------------------------  D | **.2102978** .0967891 2.17 0.030 .0205946 .400001  personas | -.0102136 .0049767 -2.05 0.040 -.0199678 -.0004595  orden\_n | -.0597519 .0276622 -2.16 0.031 -.1139689 -.0055349  ocupado\_jefe | -.0333079 .0236003 -1.41 0.158 -.0795636 .0129479  educa\_jefe | .0057464 .0023911 2.40 0.016 .00106 .0104329  ingresos\_h~e | .0001744 .000105 1.66 0.097 -.0000313 .0003801  \_cons | -.0667076 .0621803 -1.07 0.283 -.1885789 .0551636  ------------------------------------------------------------------------------  Underidentification test (Anderson canon. corr. LM statistic): 126.161  Chi-sq(2) P-val = 0.0000  ------------------------------------------------------------------------------  Weak identification test (Cragg-Donald Wald F statistic): 65.005  Stock-Yogo weak ID test critical values: 10% maximal IV size 19.93  15% maximal IV size 11.59  20% maximal IV size 8.75  25% maximal IV size 7.25  Source: Stock-Yogo (2005). Reproduced by permission.  ------------------------------------------------------------------------------  Sargan statistic (overidentification test of all instruments): 0.361  Chi-sq(1) P-val = 0.5480  ------------------------------------------------------------------------------  Instrumented: D  Included instruments: personas orden\_n ocupado\_jefe educa\_jefe  ingresos\_hogar\_jefe  Excluded instruments: distancia of\_op  ------------------------------------------------------------------------------ |

Comparando con las estimaciones en el resultado 1 vemos que efectivamente los estimadores de mínimos cuadrados ordinarios están sesgados. Mientras que por MCO el impacto del programa se estima alrededor de 0.2489, por variables instrumentales éste toma el valor de 0.2109, lo cual representa un error de aproximadamente el 18% del impacto del programa.