Tarea calificada

Módulo 1

Sonia Asto Mercado - 20166142

Claudia Vivas Alejandro - 20141150

1.1 Show that this data corresponds to a non-experimental design by testing the mean equivalence for baseline variables between the treatment and control groups.

Table 1: Test de medias

Variable	p-value
Edad	0.0188
Edadˆ2	0.0255
Secundaria completa	0.5378
Sexo	0.9344
Agua potable	0.0000
Casado	0.0000
Inodoro con cisterna	0.0086
Suelo de tierra	0.0000
Paredes de esteras	0.1078
Techo de esteras	0.0000
Tiene hijos	0.0000
Numero de hijos	0.0000
Ingreso antes del programa	0.0000

En un diseño experimental, se cumple el supuesto de que el esperado de las variables no observables son iguales en el grupo de control y tratamiento. Este supuesto se puede verificar realizando un test de medias por cada variable observable que se tiene en la

muestra. No obstante, en el presente caso nos encontramos ante un diseño no experimental, el cual no fue construido de forma aleatorio, de esa forma se esperaría que se rechace la hipótesis nula del test de medias

La hipótesis nula de este test sostiene que las medias de los grupos que se comparan son iguales a un nivel de significancia. En la siguiente tabla 1, se observan los p-value de las hipótesis nulas del test de medias. Dicho lo anterior, si los valores de la hipótesis nula son menores que 0.10, 0.05 o 0.01 se podría afirmar que se rechaza la hipótesis nula y, por tanto, los grupos no tienen medias iguales. Las variables **agua potable, casado, suelo de tierra, techo de esteras, tiene hijos, numero de hijos, ingreso antes del programa** presentan valores mayores a los valores presentados anteriormente, por lo tanto, no se cumplen que los grupos de control y tratamiento tienen medias iguales.. Entonces es plausible afirmar que nos encontramos ante un diseño no experimental.

1.2 Find the average treatment effect on the treated for the earnings variable by estimating a cross-sectional linear regression. Interpret your results. Use the following specification:

$$\text{earnings1} = \alpha_0 + \beta_1 treatment + \alpha_1 age + \alpha_2 age^2 + \alpha_3 educ5 + \alpha_4 sex + \alpha_5 married + \alpha_6 water + \alpha_7 toilet + \alpha_8 floor + \alpha_9 walls + \alpha_10 ceiling + \alpha_11 children + \alpha_12 nchildren + \alpha_13 training. + \alpha_14 duration0 + \varepsilon_i$$

Se pide hallar el indicador ATT usando el estimador OLS, el cual se verá reflejado en el valor estimado que tome β_1. Siguiendo la especificación indicada, en la primera columna de la tabla 2, se observa que las personas que fueron parte del programa PROJOVEN aumentaron sus ingresos en s/ 68.87, en comparación al grupo de personas que no recibió el programa. Este resultado es estadísticamente significativo al 1% , asimismo se observa que el error estándar es de 20.63 y el intervalo de confianza es muy ancho, lo que lleva a decir que el estimador OLS genera mucha variabilidad en el coeficiente

estimado.

1.3 Re do 1.2 after dropping educ5, sex, age and age^2 variables from the model. What happens with the estimated treatment effects? Interpret your results.

Los resultados de esta regresión están en la tabla 2. Se observa que las personas que participaron en el programa PROJOVEN aumentaron sus ingresos en s/ 59.38, en comparación al grupo de personas que no fueron parte de PROJOVEN. Este resultado es estadísticamente significativo al 1% , asimismo el error estándar es de 21.14. En comparación con el modelo anterior, se observa que el efecto se redujo en aproximadamente s/ 9 y el error estándar aumenta. Lo cual se interpreta como una mayor variabilidad en el coeficiente estimado,lo cual se debería a la omisión de variables de control importantes como la edad, educación y sexo, variables que son importantes en la literatura y respaldado por la teoría de la ecuación de Mincer en la estimación de salarios. Asimismo, se observa que en la primera regresión el R^2 es de 0.1104, mientras que en el segundo modelo es de 0.0455, asi el primer modelo captura mejor la variabilidad de todas las variables, asimismo, el $R^2_{ajustado}$ nos permite comparar que modelo ajusta mejor las variabilidades, en el primer modelo es de 0.0873, mientras que en el segundo modelo es de 0.0275. Con ese indicador se vuelve a mostrar que el primer modelo presenta un mejor ajuste y mejores resultados. Por último, el estadistico F indica que todas las variables son significativas para ambos modelos.

1.4 Find the average treatment effect on the treated for the earnings variable by estimating a difference-in-differences linear regression. Interpret your results. Include the same control variables as in 1.2

En la tabla 3, el estimador de diferencias en diferencias (representado por la variable ”did”) indica que el efecto del programa PROJOVEN tuvo un efecto positivo sobre el ingreso de los participantes, pues aumentó su ingreso en s/ 137.69, este resultado es

Table 2: OLS regressions

	(1.2)	(1.3)
tratamiento	68.866***	59.381***
	(20.634)	(21.142)
edad	0.158	
	(63.125)	
edad al cuadrado	0.107	
	(1.549)	
secundaria completa	-9.565	
	(25.516)	
sexo	118.837***	
	(19.006)	
casado	-5.089	-13.574
	(33.649)	(34.531)
agua potable	6.983	9.677
	(24.756)	(25.506)
inodoro con cisterna	6.343	7.926
	(24.934)	(25.709)
suelo de tierra	3.571	1.768
	(20.513)	(21.109)
paredes de esteras	6.529	5.199
	(22.193)	(22.839)
techo de esteras	-26.654	-17.343
	(24.041)	(24.734)
tiene hijos	-87.211*	-107.755**
	(46.665)	(47.525)
numero de hijos	47.129*	46.322*
	(27.621)	(28.108)
antes del entrenamiento	7.021	9.763
	(28.279)	(28.873)
duracion antes del tratamiento (dias)	0.045	0.065
	(0.080)	(0.083)
Constant	76.488	172.709***
	(637.510)	(22.116)
Observations	594	594

Standard errors in parentheses

Dependent variable: earnings. No experimental case.

Data source: PROJOVEN.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

estadísticamente significativo al 1% y el error estándar es de 21.24 (muy cerca al resultado del segundo caso del estimador OLS). El error estándar y el nivel de signifcancia no difieren mucho del estimador OLS, sin embargo la magnitud del efecto del programa PROJOVEN sobre los jovenes que fueron parte de este es casi el doble, esta diferencia proviene del control de los efectos contemporaneos, los cuales son cancelados gracias al estimador differencias en differencias. Cabe acotar que los datos que disponemos son de corte transversal, aunque idealmente se debería contar con un panel de datos, de esta forma se podria controlar mejor los efectos contemporaneos.

1.5. Repeat your analysis in 1.4 separately for men and women. Interpret your results.

A continuación se realizan dos estimaciones, en la cuales se condiciona el outcome de interés a la variable observable sexo, los resultados se observan en la tabla 3. En primer lugar, la regresión de diferencias en diferencias condicionada a los varones indica que el efecto de participar en el programa PROJOVEN fue positivo sobre el ingreso de los participantes , ya que su ingreso aumentó en s/ 136.82 en comparación a los varones que no partiparon en el programa, este resultado es estadísticamente significativo al 1% y el error estándar es de 34.57. En segundo lugar, la regresión de diferencias en diferencias condicionada a las mujeres indica que el efecto de participar en el programa PROJOVEN fue positivo sobre el ingreso de las participantes , ya que su ingreso aumentó en s/ 138 en comparación a las mujeres que no partiparon en el programa, este resultado es estadísticamente significativo al 1% y el error estándar es de 25.82. Se puede concluir que, existió un mayor efecto positivo sobre las mujeres en comparación a los varones, en ambos casos el coeficiente estimado es estadísticamente significativo al 1%, sin embargo el error estándar fue mayor para la estimación del grupo de varones en comparación con el de las mujeres.

1.6 Use the psmatch2.ado program in STATA to estimate the nearest-neighbor

Table 3: Diff - Diff regressions

	D-D	D-D varones	D-D mujeres
tratamiento	-68.259***	-91.019***	-56.580***
	(16.324)	(25.992)	(20.396)
year	28.368*	43.502*	17.084
	(15.073)	(24.636)	(18.258)
did	137.690***	136.819***	138.001***
	(21.245)	(34.571)	(25.820)
edad	-1.312	44.869	9.821
	(37.919)	(66.713)	(44.629)
edad al cuadrado	0.226	-0.911	-0.072
	(0.931)	(1.639)	(1.095)
secundaria completa	-34.611**	-82.913***	-28.790*
	(15.328)	(31.387)	(16.721)
sexo	111.084***	0.000	0.000
	(11.417)	(.)	(.)
casado	-7.398	94.459*	-30.281
	(20.213)	(49.909)	(21.186)
agua potable	1.296	-3.261	2.821
	(14.871)	(24.785)	(18.034)
inodoro con cisterna	10.792	-13.234	20.654
	(14.978)	(25.497)	(17.954)
suelo de tierra	6.972	-2.474	19.787
	(12.322)	(19.792)	(15.333)
paredes de esteras	-1.741	9.250	3.287
	(13.331)	(23.018)	(15.830)
techo de esteras	-20.514	-22.196	-31.795*
	(14.441)	(22.652)	(18.446)
tiene hijos	-68.599**	-112.626	-84.028***
	(28.031)	(91.553)	(28.379)
numero de hijos	38.624**	121.921*	41.006**
	(16.592)	(73.054)	(16.000)
antes del entrenamiento	-1.541	-24.181	16.134
	(16.987)	(26.602)	(22.033)
duracion antes del tratamiento (dias)	0.062	0.117*	0.029
	(0.048)	(0.066)	(0.074)
Constant	55.959	-242.641	-54.582
	(383.023)	(672.840)	(450.762)
Observations	1188	512	676

Standard errors in parentheses

Dependent variable: earnings. No experimental case.

Data source: PROJOVEN.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

propensity score matching estimator. Use a logit model to predict the probability of participation (propensity score) using the following specification:

$$\rho_i = \alpha_0 + \alpha_1 age + \alpha_2 age^2 + \alpha_3 educ5 + \alpha_4 sex + \alpha_5 married + \alpha_6 water + \alpha_7 toilet + \alpha_8 floor + \alpha_9 walls + \alpha_1 0 ceiling + \alpha_1 1 children + \alpha_1 2 nchildren + \alpha_1 3 training. + \alpha_1 4 duration0 + \alpha_1 5 earnings1 \varepsilon_i$$

Find the average treatment effect on the treated. Interpret your results.

Mediante la estimación del propensity score, se determina qué variables afectan la probabilidad de ser seleccionado para el programa, al 5% de significancia (ver tabla 4). Respecto a las características de la persona, se encuentra que a mayor edad, las personas tienen menos probabilidad de pertenecer al programa, aunque existe una edad mínima a partir de la cual afecta positivamente esta probabilidad. Estar casado es un factor que influye negativamente en la probabilidad de ser partícipes del programa o no. Además de ello, a mayor salario previo del individuo, antes del tratamiento, menos oportunidades de participar del programa PROJOVEN tiene la persona. Asimismo, se observa que las características de la vivienda de la persona también son importantes, dado que tener agua potable, suelo de tierra y techos de esteras aumentan la probabilidad de participar, mientras que tener paredes de esteras la disminuye. Por otro lado, si el individuo tuvo entrenamiento previo disminuye la probabilidad de ser beneficiario del programa. De acuerdo con la estimación realizada, el soporte común se encuentra en el siguiente rango [0.00835595, 0.99902051]. Dentro de ella, se encuentra 295 personas que son ”controles” y 299 que son ”tratados” que están divididas en 6 bloques.

En la tabla 5 mediante la estimación ”one-to-one matching”, el efecto promedio del tratamiento sobre los tratados (ATT) es 107.40 soles. En otras palabras, los que participan del programa PROJOVEN tienen un ingreso mayor en 107.40 soles frente a aquellos que no participaron del programa, pero que tuvieron similar probabilidad de

Table 4: Logit Regression

	(1)
edad	-1.586**
	(0.706)
edad al cuadrado	0.041**
	(0.017)
secundaria completa	-0.440
	(0.288)
casado	-1.574***
	(0.393)
agua potable	0.657**
	(0.272)
inodoro con cisterna	-0.104
	(0.277)
suelo de tierra	1.534***
	(0.219)
paredes de esteras	-1.066***
	(0.248)
techo de esteras	1.366***
	(0.273)
tiene hijos	-0.567
	(0.534)
numero de hijos	0.148
	(0.320)
estudio	-1.155***
	(0.382)
duracion antes del tratamiento (dias)	0.004***
	(0.002)
ingresos mensuales antes del programa	-0.004***
	(0.001)
Constant	15.768**
	(7.102)
Observations	594

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: treatment. No experimental case.

Data source: PROJOVEN.

Table 5: PSM with one nearest neighbor

	Difference
Unmatched	71.063***
	(18.294)
ATT	107.408***
	(37.704)
Observations	594

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Dependent variable: earnings. No experimental case.
Data source: PROJOVEN.

participación. El ATT tiene un t-estadístico de 2.85, por lo que el p-value, para una data de 598 observaciones, es 0.00452. Entonces, podemos decir, que el efecto promedio del tratamiento sobre los tratados es significativo al 1%. Además de ello, el error estándar de este modelo es 37.704, por encima de los modelos que se mostrarán más adelante. Esto se debe a que al usar individuos muy comparables, el nivel de información que se obtiene es poca, de manera que la variancia es más alta.

1.7 Re do 1.6 but this time use 5 neighbors rather than 1 in the matching estimation. What do you observe? Interpret your results.

En la tabla 6, la estimación del ATT se realiza utilizando 5 vecinos para el matching de cada individuo tratado, de manera que el ATT es de 90.16 soles (17 soles menos aproximandamente que el ATT con solo un vecino cercano). El t-estadístico en este caso es 3 y, por ende, el p-value es 0.00281. Ello significa que el efecto promedio del tratamiento sobre los tratados es significativo al 1%. Por otro lado, vemos que el error estándar (30.03) es menor cuando se estima el efecto con más vecinos cercanos, esto debido a que el nivel de información que se tiene es más rico, ya que hay más individuos ”menos comparables”

Table 6: PSM with five nearest neighbors

	Difference
Unmatched	71.063***
	(18.294)
ATT	90.160***
	(30.033)
Observations	594

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Dependent variable: earnings. No experimental case.
Data source: PROJOVEN.

1.8 Re do 1.6 but this time use kernel matching. What do you observe? Interpret your results.

Table 7: PSM with Kernel distribution

	Difference
Unmatched	71.063***
	(18.294)
Constant	98.891***
	(32.535)
Observations	594

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Dependent variable: earnings. No experimental case.
Data source: PROJOVEN.

Para este caso, se utiliza la distribución de kernel en el cual comparamos a cada individuo del grupo tratamiento con todos los individuos del grupo de control que están en el soporte común. Con ello, se obtiene un ATT de 98.89 soles. Es decir, la diferencia promedio entre los ingresos de los individuos beneficiarios y aquellos que no participaron del programa es de 98.89 soles. Este efecto es mayor que el encontrado cuando se estima el ATT con 5 vecinos cercanos (1.7), pero menor cuando se estima solo con

un vecino (1.6). De igual manera, se puede observar que el error estándar es mayor a lo obtenido en el caso (1.7), pero menor a lo obtenido en el (1.6). Por último, el p-value que se obtiene en este caso es 0.00247. Al igual que los 2 casos anteriores, el ATT es significativo al 1%.

1.9 Assess the plausibility of the conditional independence assumption (CIA) by implementing the balancing covariate test (use the pstest command in STATA)

Table 8: Balancing Covariate Test

	Treated	Control	p value
Edad	19.755	19.443	0.114
Edadˆ2	396.320	383.520	0.111
Secundaria completa	0.852	0.710	0.000
Sexo	0.433	0.497	0.119
Casado	0.084	0.082	0.949
Agua potable	0.691	0.679	0.745
Inodoro con cisterna	0.668	0.654	0.729
Suelo de tierra	0.577	0.597	0.630
Paredes de esteras	0.638	0.0.582	0.166
Techo de esteras	0.376	0.297	0.043
Tiene hijos	0.151	0.186	0.258
Numero de hijos	0.208	0.279	0.166
Training previo	0.198	0.146	0.096
Duración del training	56.480	26.599	0.007
Ingreso antes del programa	74.012	79.125	0.631
Observations		596	

Data source: PROJOVEN

El supuesto de independencia condicional (CIA, por sus siglas en inglés) nos dice que la participación no está determinada por variables no observadas, sino que se atribuye exclusivamente a características observables. Para ello se ha aplicado una prueba de balance en las variables observables de la data. En la tabla 8, se revela que la mayoría de las variables están balanceadas, es decir, su p-value es mayor a 0.05, por lo que no se puede rechazar la hipótesis nula de que la diferencia entre el grupo de tratamiento y de

control es cero.

Teniendo en cuenta ello, podemos decir que ambos grupos están balanceados y, por ende, son estadísticamente similares al 5% de significancia en las siguientes características: edad, género, estado civil (casado), acceso a agua potable, inodoro con cisterna, suelo de tierra, paredes de esteras, si tienen hijos, número de hijos, si tuvieron entrenamiento previo y salario. Además, podemos decir que tanto el grupo de tratamiento y de control son estadísticamente distintos respecto a si tienen secundaria completa, si tienen techo de estera y la duración del training previo.

1.10. Compare the estimates and the assumptions from the diff-in-diff and matching estimators. Was this program effective? What is the most credible estimator in the context of this program?

Para utilizar el Propensity Score Matching, se necesita cumplir con el supuesto de independencia condicional (CIA) y el supuesto de soporte común. El primer supuesto se cumple, ya que la probabilidad de participar en el programa PROJOVEN solo depende de variables observables. Esto se puede observar en la tabla 8, en la cual se muestra que las variables se encuentran balanceadas en su mayoría. Sin embargo, es importante mencionar que pueden existir más variables, que afectan la probabilidad, que podrían tomarse en cuenta para hacer más precisa la estimación. Estas variables pueden ser las características sobre la escolaridad de los padres, participación previa en otros programas sociales, así como información laboral (empleado o no, tipo de empleo, entre otros).

En relación al segundo supuesto (soporte común), este supuesto se cumple, debido a que existe una región entre las distribuciones de la probabilidad de ser elegido para el programa del grupo control y tratamiento. Como se mencionó anteriormente, el rango de soporte común es [0.00835595, 0.99902051], lo que quiere decir que las distribuciones

de probabilidad de ambos grupos son muy similares, lo que genera que exista esta región entre ellos y nos permita realizar la estimación del efecto del programa mediante esta metodología.

Por otro lado, el supuesto de identificación del estimador de diferencias en diferencias sostiene que el outcome para el grupo de control y tratamiento, en ausencia del tratamiento, generaría que el outcome siga tendencias paralelas, gráficamente se observaría que el outcome para el grupo de control y tratado en ausencia del tratamiento siguen tendencias paralelas. Asimismo, este supuesto ayuda a controlar el sesgo de selección propio de un diseño no experimental, lo cual implica que el término de perturbación o los factores no observables sean iguales en ambos grupos antes y después del tratamiento. Este supuesto se puede verificar graficando las tendencias de los outcomes antes del tratamiento o realizando un test placebo, donde se escoge un periodo falso de tratamiento y un grupo de tratamiento falso y se estima el modelo por diff-diff. Si los resultados demuestran un impacto nulo entonces el supuesto de tendencias paralelas se cumple. No obstante, no contamos con datos panel para verificar el supuesto de tendencias paralelas, de forma que no podemos afirmar que se cumple el supuesto de identifación del estimador de diferencias en diferencias.

- Efectividad:

  En las cuatro estimaciones realizadas (una con Diff-Diff y tres con PSM) se encuentra que el programa PROJOVEN si aumenta los ingresos de los participantes. Sin embargo, es importante resaltar que el efecto obtenido de la estimación Diff-Diff (s/ 137.69) es mayor que el efecto obtenido con la metodología PSM con un vencino (s/ 107.41), con cinco vecinos (s/ 90.16) y con la distribución de Kernel (s/ 98.90). Respecto a los errores estándar, podemos decir que el estimador de diferencias en diferencias (21.245) es menor a los registrados por el estimador de propensity score matching con un vecino (37.704), con cinco vecinos (30.033) y

con la distribución de kernel (32.535). Además, los estimadores reportan un nivel de confianza al 99% para todos los casos.

- Credibilidad:

  Debido a que el supuesto de identifación del estimador de diferencias en diferencias no es verificado y, por tanto, no se puede afirmar que se controla el sesgo de selección, descartamos al estimador de diferencias en diferencias como el más confiable. Por otro lado, el estimador de propensity score matching cumple con los supuestos de indetificación CIA y soporte común, asi que este estimador queda totalmente identificado. En conclusión, consideramos que el estimador de propensity score matching es el que performa mejor para estimar los efectos causales del programa PROJOVEN.