Manejo de datos Jadenkä 2018

Por: Gabriella Wong, Innovations for poverty action

15 February 2019

Este documento tiene por propósito describir el manejo de datos para la evaluación de impacto del programa JADENKÄ. Aquí veremos: (1) Criterios de limpieza de estudiantes, (2) Descripción de limpieza de datos, (3) Análisis por desgaste de la muestra, (4) características generales de las bases de datos finales.

Para la creación de este documento se utilizó Markdown.

(1) Criterios de limpieza

Tenemos una serie de estudiantes con consentimiento (encuesta realizada) que no deberían tomarse en consideración para el análisis de impacto del programa, pues en diversas ocasiones, son estudiantes que no pertenecen a la muestra (estudiantes falsos), estudiantes que fueron entrevistados dos veces (duplicados), o problemas de recolección de datos que no permiten la identificación de los estudiantes.

A continuación, se detalla los potenciales criterios para eliminar estudiantes de las bases de datos del programa Jadenkä:

Criterios a considerar para la eliminación de estudiantes en baseline y endline:

a. Duplicados: Encuestas que se hicieron dos veces.
b. Desconocidos: Estudiantes cuya identidad es desconocida debido a problemas en la recolección de datos. Por ejemplo, con nombre "EEA".
c. Falsificados: Estudiantes que según reporte de verificaciones telefónicas de LB no son reconocidos por los maestros (reporte de falsificaciones línea base)[1].
d. Capacitación: Encuestas que fueron subidas al servidor como parte de capacitaciones a encuestadores
e. Audio checks: Encuestas que fueron reconocidas como falsas a través de audio checks[2]. *(Solo aplica para endline)*

(2) Limpieza de línea de base

En esta sección, describimos cuáles fueron los puntos a limpiar para la base de datos de estudiantes de linea de base. Para ello, se utiliza el dofile *BL_estudiantes_cleaning*. En primer lugar, el objetivo principal de la limpieza de datos de línea base es poder eliminar a las observaciones que no cumplen con los criterios específicos para formar parte de dicha muestra. En tal sentido, se ha depurado a los estudiantes duplicados, los que no pueden ser reconocidos, y los falsos por recolección de datos. En segundo lugar, el dofile cumple la funcion de recrear los codigos únicos a estudiantes. Por último, el tercer punto de la limpieza de datos de la línea de base implica corregir errores que sucedieron durante la recolección, como por ejemplo, un error en fechas, de códigos, o de otra información relevante para el estudio.

El archivo "BL_dropped_students" contiene todas las observaciones que han sido eliminadas de la data cruda de línea base.

En total, cuenta con 43 observaciones. Éstas se dividen de la siguiente manera:

Tabla de incidencias - Linea de base

Tipo de correcciones	Numero de observaciones	Porcentaje del total (%)
Sesiones de capacitación	2	4.65
Estudiantes Duplicados	12	27.91
Estudiantes falsificados	27	62.79
Estudiantes no identificados	2	4.65
Total	43	100%

Como podemos observar, el 62.79% del total de incidencias corresponde a falsificaciones detectadas en la línea de base. Éstas se identificaron mediante llamadas telefónicas realizadas a los maestros, en donde se les nombraba el nombre del estudiante que fue entrevistado dentro de su sección. De tal manera, los maestros identificaban si el estudiante nombrado pertenecía a su sección o no. Y por lo tanto, estos 27 estudiantes son aquellos cuyos maestros no reconocieron pertenecer a su escuela.

De esta manera, la muestra total para la línea de base ha sufrido los siguientes cambios:

Tabla de cambios en muestra - Linea de base

Tipo de correcciones	Numero de observaciones
Base de datos SurveyCTO	3653
Observaciones por depurar	43
Observaciones sin consentimiento	21
Total observaciones para analisis	**3589**

---

[1] Durante la recolección de línea final (octubre), realizamos llamadas telefónicas a los maestros para reconfirmar si los estudiantes enlistados son efectivamente estudiantes suyos. De esta manera, pudimos reconocer a los estudiantes que no deben pertenecer a la muestra. Estos serán eliminados ex - post la creación de los códigos asignados por IPA.

[2] Por ejemplo, encuestas que aparecen con consentiemiento "Sí", pero al escuchar el audio, encuestadora se encuentra en una carretera, y no en una escuela teniendo conversación con estudiante.

A continuación la siguiente tabla muestra el total de incidencias encontradas durante la recolección de la línea final.

(3) Limpieza de linea final

Repetimos el mismo proceso para la muesta de la linea final.

Tabla de incidencias - Linea final

Tipo de correcciones	Numero de observaciones	Porcentaje del total (%)
Sesiones de capacitación	42	50.60
Estudiantes Duplicados	25	30.12
Estudiantes falsificados	16	19.28
Estudiantes no identificados	0	0.00
Total	83	100%

La tabla anterior muestra cómo está distribuido las observaciones que fueron eliminadas de la base de datos para la línea final. En total, son 43 observaciones que fueron dropeadas por los motivos descritos; siendo las sesiones de capacitación la principal causa con 50.60%.

Tabla de cambios en muestra - Linea final

Tipo de correcciones	Numero de observaciones
Base de datos SurveyCTO	4392
Estudiantes no encontrados	1071
Observaciones por depurar	83
Observaciones sin consentimiento	70
Total observaciones para analisis	**2518**

Del total de observaciones recogidas, y luego de eliminar las observaciones por status de estudiantes, así como las observaciones por depurar y las que no cuentan con consentimiento, la muestra final para la linea final es de 2518 observaciones.

Summary de la sección

Hasta ahora hemos revisado cómo se ha trabajado las bases de datos desde su recolección de datos, hasta la limpieza de su información, depurando información que no debía formar parte del análisis por los motivos descritos en la primera sección. El siguiente paso para continuar con la evaluación de impacto, es describir el desgaste de la muestra. Este paso busca entender si la probabilidad de pertenecer al grupo de los no entrevistados está relacionado con la asignación al tratamiento. De esta manera, en caso encontremos efectos significativos, puede que estemos subestimado (menos control y más tratamiento), o sobreestimando (viceversa) el efecto del programa. Caso contrario, la evaluación de impacto brindará efectos robustos de la evaluación.

(4) Análisis de desgaste

El marco de la muestra efectiva es el total de estudiantes que se encuentran en la línea final, y ésta es de 2518 observaciones. Esto implica una pérdida de 1071 respecto a la línea de base, o a un nivel de desgaste de 0.30.

En esta sección, vamos a presentar dos análisis: 1. Tablas de ortogonalidad: sobre los puntajes y data administrativa de línea final con información de línea base. 2. Análisis de desgaste: sobre la probabilidad de ser una observación perdida sobre la asignación al tratamiento, el género, y la zona de donde se encuentra el estudiante.

1. Tablas de ortogonalidad

Tabla: Estudiantes LF con información sobre puntaje LB

La tabla de ortogonalidad a describir a continuación fue realizada manteniendo solo a los estudiantes que se encuentran en la línea final, pero con información de la línea base. El objetivo es entender que no hemos perdido balance al tener desgaste en la muestra, y por tanto, que los estudiantes siguen siendo comparables en LF.

El primero modelo ha sido corregido por errores estándar a nivel de escuela, mientras que el segundo modelo toma encuenta los efectos fijos por variables de estratificación: área geográfica y tipo de grado. Asimismo, todo puntaje ha sido doble estandarizado.

```
(note: file D:\Box Sync\Panama Math\10_Analysis&Results\01 Data analysis\data\dta\BL\temp\cluster_y_strata.dta not
found)
file D:\Box Sync\Panama Math\10_Analysis&Results\01 Data analysis\data\dta\BL\temp\cluster_y_strata.dta saved

                                        Control:  Tratamien~e:  Tratamien~l:  p-value f~y:
                                               _           _           _           _
  ¿Cuál es el sexo del estudian~?:mean      0.502       0.511       0.504       0.753
                    time_elapsed:mean     17.872      17.976      18.241       0.754
                                nr:mean      7.795       7.560       7.493       0.624
        Puntaje - Línea de base:mean      0.108       0.049       0.169       0.151
         Baseline Ansiedad (sd):mean     -0.029       0.019      -0.165       0.039
                       bl_egra_z:mean      0.080       0.121       0.228       0.043
                EGRA en ngabere:mean      0.001       0.022       0.002       0.745
  Baseline ethnomath score (sd):mean     -0.017       0.051       0.034       0.570
    Baseline cultura score (sd):mean     -0.053       0.052       0.027       0.148
                Estudiante ngäbe:mean      0.934       0.948       0.938       0.329
                                  N:_    811.000     881.000     826.000           .
```

Resumen: Observamos que EGRA en español, y ansiedad matemática son estadísticamente significativas entre asignaciones. Respecto a EGRA, la diferencia es estadísticamente significativa entre control-intercultural, y bilingüe-intercultural (ver p-values por brazo a partir de la cuarta a la sexta columna). En el índice de ansiedad matemática se repite la misma incidencia.

Tabla: Estudiantes LF con información sobre escuelas de LF

Ahora queremos describir los cambios que han podido surgir con la pérdida de estudiantes durante la línea final. De este modo, vamos a realizar la tabla de ortogonalidad con información de data adminsitrativa proporcionada por el MEDUCA, con las escuelas que pertenecen a la línea final.

```
file D:\Box Sync\Panama Math\10_Analysis&Results\01 Data analysis\data\dta\BL\temp\admin_balance.dta saved
```

	Control:	Bilingue~1):	Intercul~2):	Overall:	p-value f~y:
Area:mean	3.368	3.371	3.363	3.367	0.999
se	0.157	0.156	0.156	0.090	.
DISTRITO:mean	14.048	14.113	14.048	14.070	0.996
se	0.584	0.629	0.652	0.358	.
CORREGIMIENTO:mean	59.232	61.782	60.048	60.351	0.848
se	3.051	3.445	3.126	1.850	.
r_area_mgmt:mean	3.048	2.976	3.113	3.046	0.774
se	0.133	0.135	0.138	0.078	.
# total de estudiantes latino~y:mean	35.416	34.847	38.556	36.271	0.574
se	2.235	2.468	3.227	1.543	.
Tipo de preescolar:mean	2.184	2.185	2.153	2.174	0.911
se	0.059	0.059	0.061	0.034	.
# Total de estudiantes en esc~l:mean	35.031	35.158	38.938	36.380	0.663
se	2.910	3.177	4.163	1.994	.
# Total de estudiantes en esc~l:mean	15.200	15.092	14.900	15.063	0.980
se	1.127	1.052	1.045	0.619	.
# Total de estudiantes Ngabe ~ :mean	31.000	31.540	33.573	32.035	0.684
se	1.931	2.192	2.450	1.268	.
# Total de estudiantes Ngabe ~ :mean	33.512	35.131	37.141	35.269	0.677
se	2.526	2.872	3.280	1.677	.
# Total de estudiantes Ngabe ~ :mean	17.966	16.842	17.508	17.446	0.688
se	0.967	0.926	0.864	0.529	.
Servicio de luz eléctrica o p~e:mean	1.290	1.283	1.379	1.318	0.567
se	0.072	0.071	0.069	0.041	.
Escuela con Luz eléctrica:mean	0.354	0.398	0.344	0.366	0.709
se	0.049	0.050	0.049	0.028	.
La escuela tiene panel solar:mean	0.670	0.690	0.742	0.701	0.548
se	0.049	0.051	0.046	0.028	.
# salones de madera:mean	2.214	2.292	3.037	2.519	0.386
se	0.434	0.440	0.516	0.270	.
# salones de concreto:mean	4.143	3.500	5.037	4.253	0.579
se	0.570	0.699	1.507	0.589	.
Material de la escuela:mean	1.435	1.362	1.493	1.430	0.580
se	0.091	0.077	0.096	0.051	.
Base de la cual proviene la i~o:mean	1.336	1.371	1.339	1.349	0.814
se	0.042	0.044	0.043	0.025	.
Indica si tiene cocina:mean	0.410	0.403	0.338	0.384	0.610
se	0.056	0.058	0.055	0.033	.
Cuenta con alguna fuente de a~a:agua potab~s	0.908	0.888	0.862	0.886	0.535
se	0.026	0.029	0.032	0.017	.
Indica si tiene huerta:mean	0.370	0.377	0.475	0.410	0.454
se	0.066	0.067	0.066	0.038	.
# letrinas que tiene la escuela:mean	0.846	0.972	1.095	0.969	0.106
se	0.080	0.081	0.089	0.048	.
# sanitarios que tiene la esc~l:mean	0.462	0.458	0.486	0.469	0.990
se	0.127	0.114	0.209	0.089	.
N:_	125.000	124.000	124.000	373.000	.

Tabla: Estudiantes LF con información sobre escuelas de LF

Finalmente, vamos a revisar las características de las escuelas que quedan en LF con información de la base de datos de directores, con el fin de re-evaluar el balance por características de las escuelas, y hacer contraste con la data administrativa.

```
(note: file D:\Box Sync\Panama Math\10_Analysis&Results\01 Data analysis\data\dta\BL\temp\school_data_EL.dta not found)
file D:\Box Sync\Panama Math\10_Analysis&Results\01 Data analysis\data\dta\BL\temp\school_data_EL.dta saved
```

	Control:	Tratamien~e:	Tratamien~l:	p-value f~y:
¿La escuela es unigrado o mul~g:mean	0.615	0.550	0.638	0.419
Cuántos turnos tiene la escue~?:mean	1.355	1.265	1.304	0.349
¿Su escuela cuenta con maestr~a:mean	0.455	0.451	0.393	0.659
Lengua predominante:Español	0.774	0.844	0.778	0.313
Lengua predominante:Ngäbere	0.461	0.402	0.453	0.645
Lengua predominante:Buglere	0.000	0.025	0.000	0.222
¿Su escuela tiene un preescol~ :mean	0.686	0.723	0.709	0.810
¿Su escuela tiene un CEFACEI ~ :mean	0.471	0.446	0.485	0.792
¿Su escuela tiene un prejardín?:mean	0.729	0.767	0.658	0.592
¿Su escuela tiene un jardín?:mean	0.971	0.945	0.959	0.718
Originalmente, ¿la ecuela fue~o:mean	1.017	1.025	1.026	0.917
¿De qué material es la escuela?:mean	1.267	1.238	1.265	0.956

¿De qué material son las aula~d:mean	1.193	1.295	1.186	0.288	
¿De qué material son las aula~d:mean	1.904	1.797	1.793	0.812	
¿El CEFACEI (u otro programa ~ :mean	1.160	1.206	1.367	0.381	
¿De dónde viene el agua que u~ :mean	1.526	1.516	1.496	0.992	
¿Cómo es la situación de la e~u:mean	2.052	2.098	2.034	0.734	
¿Con qué tipo de servicio bás~o:mean	1.912	1.908	1.705	0.502	
¿Tienen conexión a internet?:mean	0.397	0.355	0.388	0.767	
¿Qué tipo de servicio sanitar~ :mean	1.693	1.694	1.658	0.902	
¿El lugar donde se encuentra ~ :mean	0.026	0.049	0.060	0.537	
N:_	117.000	122.000	117.000	.	

2. Análisis de desgaste

Los siguientes modelos muestran la probabilidad de no estar presente en la línea final y su relación con la asignación al programa.

Al ver el Modelo 1, no se encuentra una relación significativa entre la pérdida de estudiantes y su asignación al tratamiento. Tampoco se encuentra una relación estadística por género del estudiante, ni por tipo de grado estudiantil.

	(1) NLF	(2) NLF	(3) NLF	(4) NLF	(5) NLF	(6) NLF	(7) NLF
Bilingue (T1)	-0.041		-0.016		-0.044		0.008
	(-1.40)		(-0.45)		(-0.63)		(0.19)
Intercultural~2)	-0.005		0.015		0.035		0.045
	(-0.18)		(0.43)		(0.51)		(1.07)
¿Cuál es el se~?		-0.002	-0.011				
		(-0.12)	(-0.41)				
Formal				-0.018	-0.017		
				(-0.57)	(-0.30)		
No formal				0.012	0.057		
				(0.35)	(0.91)		
Área Comarcal						0.078**	0.053
						(3.31)	(1.28)
Constant	0.314***	0.299***	0.300***	0.305***	0.307***	0.247***	0.229***
	(15.35)	(21.44)	(11.77)	(10.72)	(6.21)	(13.34)	(9.21)
Observations	3589	3589	3589	3589	3589	3589	3589

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

No obstante, la probabilidad de pertenecer al grupo de desgaste está relacionada estadísticamente con el pertenecer al área comarcal (zona toma el valor de 1 cuando es área comarcal). Es decir, los estudiantes que viven en zonas comarcalaes tienen más opciones de estar en el grupo de desgaste, pues existe mayor ausentismo escolar, así como dificultad en el acceso a la escuela; causas que pueden permitir no encontrar al estudiante durante la visita. Al revisar el modelo 7 (interacción de tratamiento con zona), ésta no es significativa. Esto indica que la cantidad de estudiantes perdidos de las áreas comarcales no es significativamnete distinta entre los grupos de tratamiento y control.

6. ANEXOS

Para revisar de manera detallada los outputs de este documento, correr los siguientes do-files:

- BL_estudiantes_clean.do
- BL_estudiantes_prep.do
- EL_estudiantes_clean.do
- EL_estudiantes_prep.do
- Attrition_test.do

Todos los do-files trabajados se encuentran en la carpeta *"dofiles"* dentro de *"Analisis&Report"*

** THE END **