Data for
Development Impact

Construcción de Datos

Manejo y limpieza

Rony Rodriguez-Ramírez

July 11, 2020

LAMBDA

Construcción de Datos

Desglose de tareas de trabajo de datos

- Dividimos el proceso de trabajo de datos en cuatro etapas:
  1. De-identificación
  2. Limpieza de datos
  3. Construcción de variables
  4. Análisis de datos
- Cada una de estas etapas tiene entradas y salidas bien definidas.
- Para cada etapa, debe haber una carpeta de códigos y un conjunto de datos correspondiente
- Los nombres de códigos, conjuntos de datos y salidas para cada etapa deben ser consistentes
- El código, los datos y los resultados de cada una de estas etapas deben pasar por al menos una ronda de revisión de código.

Recopilación e importación de datos.

- Para esta presentación, asumiremos que ya ha recopilado e importado sus datos.
- En la práctica, sin embargo, la limpieza de datos comienza antes de que termine la recopilación de datos
- Las siguientes tareas de recopilación de datos que, cuando se realizan correctamente, hacen que la limpieza de datos sea mucho más fácil:
  1. Programación de encuestas
  2. Monitoreo de calidad de datos
  3. Importación de datos

De-identificación

Input: Los datos sin procesar (raw data)

- Contiene solo información recibida directamente del campo.
- Los archivos de datos sin procesar deben almacenarse en la carpeta de datos sin procesar exactamente como se recibieron.
- Tenga en cuenta cómo y dónde se almacenan: no se pueden volver a crear y casi siempre contienen datos confidenciales
- Los archivos de datos sin procesar nunca deben editarse directamente
- Si los datos sin procesar contienen información confidencial, deben estar encriptados
- Asegúrese de tener una copia de seguridad de los datos sin procesar (con suerte, nunca necesitará usarlos)

Output: Datos de-idenficados

- Versión de trabajo del conjunto de datos que se puede compartir dentro del equipo de investigación sin riesgo
- Contiene solo información recibida directamente del campo
- No contiene identificadores directos
- No es necesariamente anónimo
- Por lo general, la desidentificación no debe afectar la usabilidad de los datos.

Identificar identificadores directos

- Lo primero que necesita para de-identificar los datos es una lista de todos los identificadores directos en el conjunto de datos
- Idealmente, todas las variables potencialmente identificables se marcan durante el diseño del cuestionario.
- Herramientas útiles:
  - `Escaneo PII` de JPAL
  - `SdcMicro` del Banco Mundial
  - `Iecodebook` de DIME Analytics

Eliminar identificadores directos

Una vez que tenga una lista de todos los identificadores directos, para cada uno de ellos, pregúntese (y el plan de preanálisis):

1. ¿Será necesaria esta variable para el análisis?
   - Si la respuesta es no, simplemente suéltela
   - No tenga miedo de descartar demasiadas variables: siempre puede recuperarlas (pero no puede deshacer una fuga de datos)
2. ¿Puedo codificar o construir una variable que enmascara la PII?

Data cleaning

Data cleaning

Durante la limpieza de datos, observará cuidadosamente cada variable en su conjunto de datos. Los objetivos de este proceso son:

- Hacer que el conjunto de datos sea fácilmente utilizable y comprensible.
- Documentar puntos de datos individuales y patrones que pueden sesgar el análisis.

Introducción

- La limpieza es probablemente la tarea de trabajo de datos que consume más tiempo, y se sentirá tentado a omitirla.
- Sin embargo, este es el momento cuando realmente conozca sus datos.
- Explore su conjunto de datos usando tabulaciones (tab), summ y diagramas (plots) descriptivos.
- Conocer bien su conjunto de datos permitirá hacer análisis
- Limpiar bien sus datos le ahorrará tiempo.

Output: Una base de datos limpia

¿Qué creen que nos referimos con una base de datos limpia?

Output: Una base de datos limpia

- Al final de este proceso, debe tener un conjunto de datos que sea esencialmente el mismo que descargó del servidor.
- La principal diferencia es que el conjunto de datos limpios debería ser más fácil de entender para cualquiera que lo abra por primera vez.

Output: Una base de datos limpia

- También debe remontarse fácilmente al instrumento de encuesta.
- Por lo general, se creará un conjunto de datos limpios para cada fuente de datos o instrumento de encuesta.
- El conjunto de datos limpios y no identificados, más la documentación que lo respalda, son los primeros datos de salida de su proyecto: un conjunto de datos publicable.

Output: Documentación

Algunas piezas de documentación deben acompañar al conjunto de datos limpios:

- Un diccionario de variables, o libro de códigos, que enumera detalles sobre las variables en el conjunto de datos.
- Los instrumentos utilizados para recopilar los datos.
- Un registro completo de las correcciones realizadas a los datos sin procesar, incluida una explicación cuidadosa sobre el proceso de toma de decisiones involucrado
- Un informe que documente cualquier irregularidad adicional y patrones de distribución encontrados en los datos.

Unique ID

- Lo primero que desea buscar cada vez que abre un nuevo conjunto de datos por primera vez es
  1. Unidad de observación.
  2. Identificación única y totalmente variable de identificación.
- Antes de separar los datos identificables de los no identificados, asegúrese de saber cómo cruzar ambos utilizando la ID única

Propiedades deseables de una variable ID

1. Identificación única
2. Totalmente identificable
3. Anónimo
4. Constante dentro de un proyecto

¿Cómo probaría si una variable se identifica de manera única y completa?

Unique ID

Comandos para probar que la variable en Stata:

- `isid`
- `codebook`

¿Cuál es la unidad de observación?

HHID	Village	District	HH number	HH head	HHH Age
022501	25	2	1	Andrew	52
022502	25	2	2	Patrick	48
023207	32	2	7	Charles	29
023205	32	2	5	Jeffrey	37
012501	25	1	1	Walter	48
011103	11	1	3	Anne	26
011205	12	1	5	Lawrence	61
024502	45	2	2	Dennis	45
024501	45	2	1	Nancy	41

¿Cuál es la unidad de observación?

Clinic ID	Clinic Number	District	Patient	Age
02452	542	2	Andrew	52
02543	543	2	Patrick	48
02156	156	2	Charles	29
01152	152	1	Jeffrey	37
01152	152	1	Walter	49
01238	238	1	Anne	26
01122	122	1	Lawrence	61
02122	122	2	Dennis	45
02122	122	2	Nancy	41

Usando `iefieldkit` para resolver entradas duplicadas

- El comando `ieduplicates` ayuda a identificar y resolver duplicados en datos de encuestas sin procesar.
- El comando genera un informe de todas las entradas duplicadas de una variable (en Excel) y elimina los duplicados del conjunto de datos hasta que se resuelven.
- El informe de Excel se utiliza para documentar los casos de entrevistas duplicadas y cómo se resolvieron.

Correcciones en el ingreso de datos

- Durante la recopilación de datos, particularmente durante la recopilación de datos primarios, es probable que los enumeradores y supervisores informen los problemas
- Durante el monitoreo de la calidad de los datos, es probable que también identifique los problemas que deben abordarse
- Ejemplos de eso son errores tipográficos, identificaciones incorrectas y nuevas encuestas
- Es importante registrar todos estos problemas y las comunicaciones sobre ellos.
- Este es el único caso, aparte de las ID duplicadas, cuando cambiará los puntos de datos durante la limpieza de datos
- **Haga todas las correcciones en un do file**, no manualmente, y recuerde documentar de dónde proviene la información

Crear un conjunto de datos anotados

Label variables (etiquetas)

Al limpiar un conjunto de datos, debe asegurarse de que todas las variables estén correctamente etiquetadas, de modo que sea fácil entender qué representa cada variable:

- Verifique que todas las variables tengan etiquetas de variables.
- Las etiquetas de las variables deben explicar qué es la variable y, si ese es el caso, en qué unidad está.
- Las etiquetas no pueden tener más de 80 caracteres.

Cifrando variables (encode)

- La base de datos limpia no debe contener variables de cadena (string variables, i.e., `texto`), excepto
  - Nombres propios que no son categorías
  - Dígitos con ceros a la izquierda o ID largos (más de 15 dígitos)
- Eso significa que las variables de cadena (string) deben transformarse en variables categóricas o factores etiquetados.
- Tenga en cuenta las preguntas abiertas: presentan un riesgo mucho mayor de divulgación estadística.
- Verifique que todas las variables categóricas tengan value labels (etiquetadas con valor).

Encoding variables en Stata

- En Stata, la mejor práctica es usar la codificación tanto con la etiqueta (`label`) como con las opciones de no extensión (`noextend`).
- Otros comandos útiles: `label define, label value, label dir, label list, labelbook`.
- Si usó SurveyCTO, usó la etiqueta de columna: stata y los datos se importaron correctamente, este paso puede no ser necesario.

Ejemplo:
```
encode dist name, generate(dist id) label(district)
                  noextend
```

Extended missing values

- Durante la recopilación de datos primarios, use códigos como -88, -9, -777 para representar diferentes razones de datos faltantes, como ”no sabe”, ”se negó a responder”, etc.
- Es necesario eliminar estos valores, ya que de lo contrario sesgarán los medios.
- Si los cambiamos a todos como perdidos, perderemos información.
- Use valores perdidos extendidos para mantener la información, pero aún así le dice a Stata que los trate como perdidos.

Extended missings values en Stata

- El missing value regular en Stata es: .
- Use missing values extendidos para representar la misma razón de la falta de datos en todo el proyecto:
  - .d = ”No sé / Don’t know”
  - .r = ”Se negó a responder / Refused to answer”
  - .s = ”Saltado / Skipped”
- Los valores perdidos también se pueden etiquetar.

Extended missings values en Stata

- Para Stata, números $< . < .a < .b < .c < \cdot < .z$
- Así que reemplace esto:

```
sum HH_ingreso if employment != .
```

Con esto:

```
sum HH_ingreso if employment < .
sum HH_ingreso if !missing(employment)
```

Renombrando variables

- No cambie los nombres de las variables que provienen de una encuesta, incluso si no le gustan las convenciones de nomenclatura utilizadas en el cuestionario.
- Cambiar el nombre de las variables hará que sea más difícil encontrar la correspondencia entre las variables y las preguntas de la encuesta.

Artículo sobre el nombre de las variables en Medium

Usando iefieldkit para anotar una base de datos

- El comando iecodebook le ayuda a realizar la mayoría de las tareas descritas anteriormente (con la excepción de la codificación)
- El comando genera (en Excel) una lista de todas las variables en el conjunto de datos y sus etiquetas, y les aplica los cambios para simplificar el proceso.
- El informe de Excel se utiliza para documentar las modificaciones realizadas en el conjunto de datos durante la limpieza.

Otras tareas en la limpieza de datos

Recategorizando valores listado como ”Otros”

- Las variables categóricas generalmente tienen una opción abierta ”otro, especificar” que se guarda como una cadena
- Las respuestas que aparecen con frecuencia en la pregunta abierta se pueden incluir como una nueva categoría en la variable categórica
- Eso generalmente se realiza durante el piloto o las comprobaciones de alta frecuencia, pero es posible que aún se omitan categorías relevantes

Eliminar variables de la encuesta

-
- Algunas variables se crean para ser utilizadas dentro de la encuesta y para las comprobaciones de la encuesta.
- Ese es el caso de la mayoría de los campos de cálculo, así como las notas y las variables de duración.
- Las variables que no forman parte del cuestionario en sí pueden eliminarse del conjunto de datos limpio.

Ordenar variables

- Se recomienda que las variables en el conjunto de datos final sigan un cierto orden como en el cuestionario
- Si creó nuevas variables durante la limpieza de datos, por ejemplo, para cambiar los códigos de la lista, probablemente estarán fuera de servicio
- Es posible que desee reordenar esas variables para que el conjunto de datos sea más fácil de leer y comparar con el cuestionario.

Convenciones para guardar archivos

Guardando archivos

- Durante el proceso de limpieza de datos, es posible que haya guardado varios archivos intermedios, por ejemplo, si limpió módulos largos por separado para que su código sea más legible.
- Después de limpiar sus datos y fusionarlos nuevamente, querrá guardar un conjunto de datos limpios final, que contenga todas las variables de su encuesta.

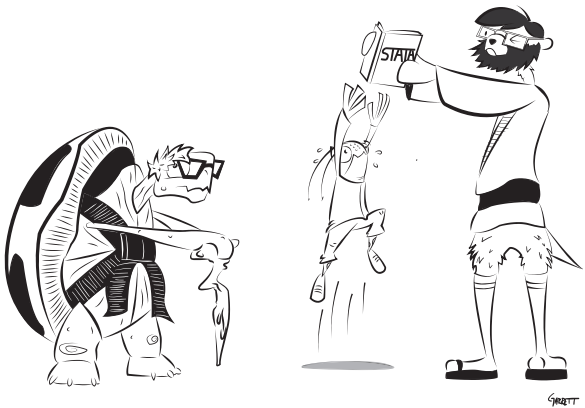
Guardando archivos

- Este nuevo conjunto de datos probablemente será bastante pesado. Use `compress` para guardar sus variables en el formato más económico (i.e., tamaño).
- A menudo es deseable guardar su conjunto de datos en una versión anterior de Stata, por lo que otros miembros de su equipo no tendrán conflictos de versión. Para hacer esto, use `saveold`.

Nombrando archivos

- Asegúrese de que todos los archivos de salida, conjuntos de datos y otros estén etiquetados de forma clara y única, es decir: ”desc_stats_tmt_only.xls” ”input_plan_adm_data.dta”
- A menudo es deseable tener los nombres de sus conjuntos de datos y archivos do vinculados, por lo que es fácil entender qué archivos do están creando qué conjunto de datos, como ”merge.do” y ”merged.dta” o ”cleaning.do” y ”clean.dta”.
- No utilice _v1, _v2, etc. para ningún archivo final. Esto conduce a errores en los archivos do que dependen de estos archivos cuando se agrega una nueva versión.
- Está bien usar _v1, _v2, etc. para versiones antiguas de archivos si realmente necesita mantener un archivo

STATA TIME



Nos vemos mañana.