Data for
Development Impact

Análisis de Datos

Construcción de Datos

Rony Rodriguez-Ramírez

January 30, 2022

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.

¿Qué es la construcción de datos?

- La construcción de variables significa procesar los puntos de datos tal como se proporcionan en los datos sin procesar para que sean adecuados para el análisis
- Es en esta etapa que los datos sin procesar se transforman en datos de análisis.
- Esto se hace mediante la creación de variables derivadas (p. Ej., Dummies, índices, interacciones).

¿Qué es la construcción de datos?

- Esto se hace mediante la creación de variables derivadas (p. Ej., Dummies, índices, interacciones).
- Es la única etapa en la que se realizarán cambios en los puntos de datos.
- La construcción está estrechamente relacionada con el diseño de investigación y el diseño de cuestionarios.
- Idealmente, la construcción del indicador debe realizarse justo después de la limpieza de datos, de acuerdo con el plan de pre-análisis.

¿Qué es la construcción de datos?

- La construcción está estrechamente relacionada con el diseño de investigación y el diseño de cuestionarios.
- Idealmente, la construcción del indicador debe realizarse justo después de la limpieza de datos, de acuerdo con el plan de pre-análisis.
- Aquí es cuando utilizará más el conocimiento de los datos que adquirió y la documentación que creó durante el paso de limpieza.
- A menudo es útil comenzar a buscar comparaciones y otra documentación fuera del editor de código

¿Qué es la construcción de datos?

- Inputs:
  - Una o más bases de datos limpias.
  - Master data.
- Outputs:
  - Una o más base de datos para análisis.
  - Un codebook para cada base de datos para análisis.
- Tareas:
  - Unidad de observación (en la encuesta) $\longrightarrow$ Unidad de análisis.
  - Pregunta de la encuesta $\longrightarrow$ Indicador de análisis.

¿Por qué la construcción es una tarea separada de la limpieza de datos?

1. Para diferenciar claramente los datos recopilados originalmente del resultado de las decisiones de procesamiento de datos
2. Para garantizar que la definición de variable sea coherente en todas las fuentes de datos.
   - Durante la limpieza de datos, creamos un output por cada input,
   - Durante la construcción de datos, podemos tener múltiples entradas y salidas.
   - Por ejemplo, podemos tener varias rondas de recopilación de datos que se limpiarán por separado, pero queremos que todas se construyan de la misma manera.

¿Qué planear con anticipación?

- ¿Cuáles son los indicadores finales necesarios para responder una pregunta de investigación?
- ¿Cómo se definen y calculan?
- ¿Cuáles son los pasos para llegar allí?
- ¿Cómo lidiar con diferentes rondas de recopilación de datos?

¿Por qué la construcción es una tarea separada de la limpieza de datos?

- En la práctica, la construcción de datos a menudo ocurre al mismo tiempo que el análisis de datos.
- A medida que analiza los datos, serán necesarias diferentes variables construidas, así como subconjuntos y otras alteraciones de los datos.
- Sin embargo, incluso si la construcción termina antes del análisis solo en el orden en que se ejecuta el código, es importante considerarlos como pasos diferentes

Construyendo variables analíticas

Creando nuevas variables

- Crear nuevas variables en lugar de sobrescribir la información original.
- Las variables construidas deben tener nombres intuitivos y funcionales.
- Ordene el conjunto de datos de modo que las variables relacionadas estén cercanas entre sí

Abordar los valores atípicos

Durante la construcción, abordará los problemas que observó en los datos durante la limpieza, incluidos valores atípicos y valores faltantes:

- Lo único que no desea hacer es dejar una observación completa debido a valores atípicos

Abordar los valores atípicos

Durante la construcción, abordará los problemas que observó en los datos durante la limpieza, incluidos valores atípicos y valores faltantes:

- Lo único que no desea hacer es dejar una observación completa debido a valores atípicos
- Las formas comunes de abordar los valores atípicos son los recortes y winsor.

Abordar los valores atípicos

Cómo tratar los valores atípicos e imputar valores perdidos son preguntas de investigación, pero hay algunas cosas a tener en cuenta.

- Asegúrese de documentar cuál fue el enfoque elegido por el equipo y por qué decidió usarlo en un caso particular

Abordar los valores atípicos

Cómo tratar los valores atípicos e imputar valores perdidos son preguntas de investigación, pero hay algunas cosas a tener en cuenta.

- Asegúrese de documentar cuál fue el enfoque elegido por el equipo y por qué decidió usarlo en un caso particular
- Estas decisiones pueden afectar la distribución de variables y los resultados observados, por lo tanto, mantenga la variable original en el conjunto de datos en lugar de reemplazarla.

Unidades estandarizadas

Asegúrese de que haya coherencia entre las variables construidas:

- Recomendamos codificar preguntas sí/no como 1 y 0 o TRUE y FALSE, para que puedan usarse numéricamente como frecuencias en medias y como variables dummies en regresiones.

Unidades estandarizadas

Asegúrese de que haya coherencia entre las variables construidas:

- Recomendamos codificar preguntas sí/no como 1 y 0 o TRUE y FALSE, para que puedan usarse numéricamente como frecuencias en medias y como variables dummies en regresiones.
- Para las variables categóricas no binarias, verifique que las etiquetas y los niveles tengan la misma correspondencia entre las variables que usan las mismas opciones.

Unidades estandarizadas

Asegúrese de que haya coherencia entre las variables construidas:

- Recomendamos codificar preguntas sí/no como 1 y 0 o TRUE y FALSE, para que puedan usarse numéricamente como frecuencias en medias y como variables dummies en regresiones.
- Para las variables categóricas no binarias, verifique que las etiquetas y los niveles tengan la misma correspondencia entre las variables que usan las mismas opciones.
- Las variables numéricas que se comparan o agregan deben convertirse a la misma escala o unidad de medida

Unidades estandarizadas

Asegúrese de que haya coherencia entre las variables construidas:

- Recomendamos codificar preguntas sí/no como 1 y 0 o TRUE y FALSE, para que puedan usarse numéricamente como frecuencias en medias y como variables dummies en regresiones.
- Para las variables categóricas no binarias, verifique que las etiquetas y los niveles tengan la misma correspondencia entre las variables que usan las mismas opciones.
- Las variables numéricas que se comparan o agregan deben convertirse a la misma escala o unidad de medida
- Cuando convierta unidades, establezca las tasas de conversión en el archivo maestro do usando globals.

Crear medidas agregadas

- El caso más simple de nuevas variables a crear son los indicadores agregados.

Crear medidas agregadas

- El caso más simple de nuevas variables a crear son los indicadores agregados.
- Saltar al paso donde realmente crea estas variables parece intuitivo, pero también puede causarle muchos problemas, ya que pasar por alto los detalles puede afectar sus resultados

Crear medidas agregadas

- El caso más simple de nuevas variables a crear son los indicadores agregados.
- Saltar al paso donde realmente crea estas variables parece intuitivo, pero también puede causarle muchos problemas, ya que pasar por alto los detalles puede afectar sus resultados
- Es importante verificar y volver a verificar las asignaciones de valores de las preguntas, así como sus escalas, antes de construir nuevas variables basadas en ellas.

Crear medidas agregadas

- El caso más simple de nuevas variables a crear son los indicadores agregados.
- Saltar al paso donde realmente crea estas variables parece intuitivo, pero también puede causarle muchos problemas, ya que pasar por alto los detalles puede afectar sus resultados
- Es importante verificar y volver a verificar las asignaciones de valores de las preguntas, así como sus escalas, antes de construir nuevas variables basadas en ellas.
- Observe las distribuciones de las variables originales y construidas.

Crear medidas agregadas

- Muchas veces es más fácil lidiar con conjuntos de datos largos al agregar
- Tenga en cuenta cómo se tratan los valores perdidos.

Crear medidas agregadas

En Stata, los diferentes comandos tratan las faltas de manera diferente:

- gen income_total = income_wage + income_rent + income_sales
- egen income_total = rowtotal(income_wage income_rent income_sales)
- egen income_total = rowtotal(income_wage income_rent income sales), m
- collapse (sum) income_wage_hh = income_wage
- collapse (mean) income_wage_hh_mean = income_wage
- collapse (median) income_wage_hh_median = income_wage

Merging base de datos

- La fusión puede cambiar tanto el número de observaciones como el valor de las variables.
- Tenga cuidado al combinar conjuntos de datos que no tienen los mismos identificadores.
- Stata y R tratan los valores en conflicto de manera diferente: R crea dos variables y Stata mantiene las entradas del conjunto de datos maestros.

Merging base de datos

options	Description
Options	
keepusing(*varlist*)	variables to keep from using data; default is all
generate(*newvar*)	name of new variable to mark merge results; default is **_merge**
nogenerate	do not create **_merge** variable
nolabel	do not copy value-label definitions from using
nonotes	do not copy notes from using
update	update missing values of same-named variables in master with values from using
replace	replace all values of same-named variables in master with nonmissing values from using (requires **update**)
noreport	do not display match result summary table
force	allow string/numeric variable type mismatch without error
Results	
assert(*results*)	specify required match results
keep(*results*)	specify which match results to keep
sorted	do not sort; datasets already sorted

sorted does not appear in the dialog box.

Reshaping observaciones

- La remodelación cambia la unidad de observación
- En Stata, la remodelación tiene una sintaxis muy única.
- Debe tener variables de identificación en el conjunto de datos para poder remodelarlo
- La remodelación creará observaciones en el conjunto de datos largo incluso cuando faltan variables

¿Qué son datos tidy?

> Todas las familias felices se parecen unas a otras, pero cada familia infeliz lo es a su manera
>
> León Tolstói (Ana Karenina)

"Al igual que las familias, las base de datos tidy son todos iguales, pero cada base es desordenada a su manera. Los datos tidy proporcionan una forma estandarizada de vincular la estructura de una base de datos (su diseño físico) con su semántica (su significado)" (Wickham, p. 2).

¿Qué son datos tidy?

- Los datos tidy son una forma estándar de mapear el significado de un conjunto de datos a su estructura.
- Los datos son desordenados u ordenados dependiendo de cómo las filas, columnas y tablas se combinan con observaciones, variables y tipos.
  1. Cada variable forma una columna;
  2. Cada observación forma una fila;
  3. Cada tipo de unidad de observación forma una tabla.

Previniendo errores

Previniendo errores

- Las tareas de construcción más complejas implican cambiar la estructura de los datos, como la muestra y la unidad de observación.
- Las fusiones, remodelaciones y colapsos pueden cambiar el número de observaciones y crear entradas faltantes.
- Asegúrese de leer acerca de cómo cada comando trata las observaciones faltantes
- Si está subconjustando sus datos, elimine las observaciones explícitamente, indicando por qué lo está haciendo y cómo cambió el conjunto de datos

Escribir pseudocódigo

- Describa los pasos para crear su indicador en español simple.
- Refinar los subpasos involucrados.
- Cuando entres en demasiados detalles, escribe el código.
- Piense en los posibles errores que pueden aparecer en cada subpaso.

Escribir `pesudocódigo`: Piense en los resultados esperados

- Piense en cómo el comando que está usando trata los valores perdidos.
- Intenta predecir el resultado que obtendrás.
  - ¿Se fusionarán todas las observaciones?
  - ¿Cambiará el número de observaciones?
  - ¿Se crearán los valores faltantes?

Documente los resultados observados

- Explore los resultados reales de la operación.
- Escriba en los comentarios lo que sucedió.
- Agregue comentarios al código explicando consecuencias inesperadas

Construya checks en su código

- Probar la unidad de observación y la variable ID
- Lanzar mensajes de error o romper el código si
  - Confirme los resultados esperados.
  - Compruebe si las salidas están cambiando cuando vuelva a ejecutar el código
- Use assert en Stata.

Outputs de construcción

Base de datos construídas

- Un conjunto de datos construido está construido para responder una pregunta de análisis;
- Diferentes piezas de análisis pueden requerir diferentes muestras o diferentes unidades de observación;
- Puede tener tantos conjuntos de datos construidos como sea necesario para el análisis;
- No se preocupe si no puede crear un único conjunto de datos de análisis ”canónico”;
- Si tiene varios conjuntos de datos construidos, asígneles un nombre cuidadoso para saber cuándo usarlos.

Documentar el output

- La documentación es una salida de construcción tan relevante como el código y los datos.

Documentar el output

- La documentación es una salida de construcción tan relevante como el código y los datos.
- Alguien que no esté familiarizado con el proyecto debe poder comprender el contenido de los conjuntos de datos de análisis y los pasos dados para crearlos.

Documentar el output

- La documentación es una salida de construcción tan relevante como el código y los datos.
- Alguien que no esté familiarizado con el proyecto debe poder comprender el contenido de los conjuntos de datos de análisis y los pasos dados para crearlos.
- La construcción de datos implica la traducción de puntos de datos concretos a mediciones más abstractas.

Documentar el output

- Documente exactamente cómo se deriva o calcula cada variable

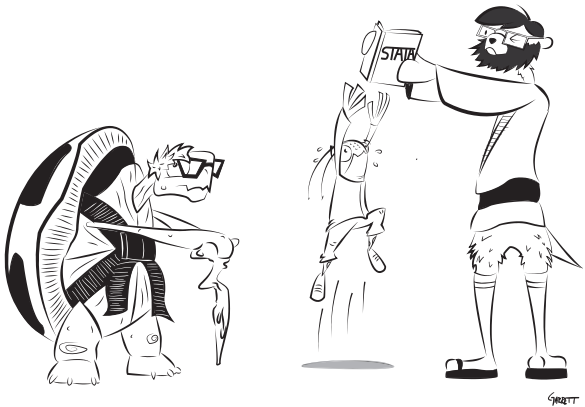
Documentar el output

- Documente exactamente cómo se deriva o calcula cada variable
- Registre cuidadosamente cómo se combinaron, recodificaron y escalaron variables específicas, y haga referencia a esos registros en el código

Documentar el output

- Documente exactamente cómo se deriva o calcula cada variable
- Registre cuidadosamente cómo se combinaron, recodificaron y escalaron variables específicas, y haga referencia a esos registros en el código
- Esto puede ser parte de una discusión más amplia con su equipo sobre la creación de protocolos para la definición variable, lo que garantizará que los indicadores se definan de manera consistente en todos los proyectos.

STATA TIME



Hasta la próxima semana.