-
-

- Home
- About
-

Type text to search here... ☐

Home > Data Management > Update to Import COVID-19 post

Update to Import COVID-19 post

24 March 2020 Chuck Huber, Associate Director of Statistical Outreach 5 Comments

Like 1 Tweet

In my last post, I mentioned that I did not want to distribute my **covid19.ado** file because “it could be rendered useless if or when Johns Hopkins changes its data”. I wrote that on March 19, 2020, and the data changed on March 23, 2020. This will likely happen again (and again, and again …). I may post updates in the future as the data change, but you may need to adapt sooner than I can post. So let’ s see how we can update our code to adapt to the changing data.

Let’ s begin by running the code from my last blog post.

```
local URL = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/"
forvalues month = 1/12 {
    forvalues day = 1/31 {
        local month = string(`month', "%02.0f")
        local day = string(`day', "%02.0f")
        local year = "2020"
        local today = "`month'-`day'-`year'"
        local FileName = "`URL'`today'.csv"
        clear
        capture import delimited "`FileName'"
        capture confirm variable ïprovincestate
        if _rc == 0 {
            rename ïprovincestate provincestate
            label variable provincestate "Province/State"
        }
        capture save "`today'", replace
    }
}
clear
forvalues month = 1/12 {
    forvalues day = 1/31 {
        local month = string(`month', "%02.0f")
        local day = string(`day', "%02.0f")
        local year = "2020"
        local today = "`month'-`day'-`year'"
        capture append using "`today'"
    }
}
```

Something looks wrong when we **describe** our data.

```
. describe

Contains data
  obs:        11,341
 vars:            17
------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------
provincestate   str43   %43s                  Province/State
countryregion   str32   %32s                  Country/Region
lastupdate      str19   %19s                  Last Update
confirmed       long    %8.0g                 Confirmed
```

```
deaths          int     %8.0g                 Deaths
recovered       long    %8.0g                 Recovered
latitude        float   %9.0g                 Latitude
longitude       float   %9.0g                 Longitude
fips            long    %12.0g                FIPS
admin2          str21   %21s                  Admin2
province_state  str28   %28s                  Province_State
country_region  str32   %32s                  Country_Region
last_update     str19   %19s                  Last_Update
lat             float   %9.0g                 Lat
long_           float   %9.0g                 Long_
active          long    %12.0g                Active
combined_key    str44   %44s                  Combined_Key
------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

We have variables with similar names, such as **provincestate** and **province_state**, **countryregion** and **country_region**, and so forth. The variable names have changed in the newer raw files. But we must have the same variable names when we **append** the data.

I looked through the most recent raw data files and identified the date on which the data changed. You can do this without opening the files. You can simply **describe** the data from your local disk or cloud account.

The raw data from March 22, 2020, use the old variable names.

```
. describe using 03-22-2020.dta

Contains data
  obs:           309                          24 Mar 2020 11:48
 vars:             8
------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------
provincestate   str28   %28s                  Province/State
countryregion   str32   %32s                  Country/Region
lastupdate      str19   %19s                  Last Update
confirmed       long    %12.0g                Confirmed
deaths          int     %8.0g                 Deaths
recovered       long    %12.0g                Recovered
latitude        float   %9.0g                 Latitude
longitude       float   %9.0g                 Longitude
------------------------------------------------------------------------
Sorted by:
```

The raw data from March 23, 2020, use the new variable names.

```
. describe using 03-23-2020.dta

Contains data
  obs:         3,415                          24 Mar 2020 11:48
 vars:            12
------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------
fips            long    %12.0g                FIPS
admin2          str21   %21s                  Admin2
province_state  str28   %28s                  Province_State
country_region  str32   %32s                  Country_Region
last_update     str19   %19s                  Last_Update
lat             float   %9.0g                 Lat
long_           float   %9.0g                 Long_
confirmed       long    %12.0g                Confirmed
deaths          int     %8.0g                 Deaths
recovered       long    %12.0g                Recovered
active          long    %12.0g                Active
combined_key    str44   %44s                  Combined_Key
------------------------------------------------------------------------
Sorted by:
```

We could write some clever code to distinguish between files created before and after March 23. But a simple alternative is to use **capture rename** to change the variable names where necessary in the raw data files.

Let’ s try this on the raw data file for March 23 before we incorporate it into the rest of our code.

```
. use 03-23-2020.dta

. capture rename province_state provincestate

. capture rename country_region countryregion

. capture rename last_update lastupdate

. capture rename lat latitude
. capture rename long longitude

. describe

Contains data from 03-23-2020.dta
  obs:         3,415
 vars:            12                          24 Mar 2020 11:48
------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------
fips            long    %12.0g                FIPS
admin2          str21   %21s                  Admin2
provincestate   str28   %28s                  Province_State
countryregion   str32   %32s                  Country_Region
lastupdate      str19   %19s                  Last_Update
latitude        float   %9.0g                 Lat
longitude       float   %9.0g                 Long_
confirmed       long    %12.0g                Confirmed
deaths          int     %8.0g                 Deaths
recovered       long    %12.0g                Recovered
active          long    %12.0g                Active
combined_key    str44   %44s                  Combined_Key
------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

The variable names in the new data now match the variable names in the old data. Some variables in the newer data did not appear in the old data. Those new variables will be appended to the final dataset but will not contain any data for dates prior to March 23.

The updated code below will import the raw data from the Johns Hopkins GitHub repository as of March 23, 2020. I have displayed the new commands in red.

```
local URL = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/"
forvalues month = 1/12 {
     forvalues day = 1/31 {
          local month = string(`month', "%02.0f")
          local day = string(`day', "%02.0f")
          local year = "2020"
          local today = "`month'-`day'-`year'"
          local FileName = "`URL'`today'.csv"
          clear
          capture import delimited "`FileName'"
          capture confirm variable ïprovincestate
          if _rc == 0 {
               rename ïprovincestate provincestate
               label variable provincestate "Province/State"
          }
          capture rename province_state provincestate
          capture rename country_region countryregion
          capture rename last_update lastupdate
          capture rename lat latitude
          capture rename long longitude
          capture save "`today'", replace
          }
}
clear
forvalues month = 1/12 {
     forvalues day = 1/31 {
          local month = string(`month', "%02.0f")
          local day = string(`day', "%02.0f")
          local year = "2020"
          local today = "`month'-`day'-`year'"
          capture append using "`today'"
     }
}
```

We can verify that this worked by describing the resulting data.

```
. describe

Contains data
  obs:        11,341
```

```
 vars:            12
------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------
provincestate   str43   %43s                  Province/State
countryregion   str32   %32s                  Country/Region
lastupdate      str19   %19s                  Last Update
confirmed       long    %8.0g                 Confirmed
deaths          int     %8.0g                 Deaths
recovered       long    %8.0g                 Recovered
latitude        float   %9.0g                 Latitude
longitude       float   %9.0g                 Longitude
fips            long    %12.0g                FIPS
admin2          str21   %21s                  Admin2
active          long    %12.0g                Active
combined_key    str44   %44s                  Combined_Key
------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

Let’ s save this dataset so we can use it later.

```
. save covid19_raw
file covid19_raw.dta saved
```

Please note that we have not checked and cleaned these data. The code above and the resulting data should be used for instructional purposes only.

I will show you how to convert the raw data to time-series data in my next post.

Categories: Data Management Tags: coronavirus, COVID-19, imort

5 Comments **The Stata Blog** **Disqus' Privacy Policy** **1 Login**

♡ **Recommend** Tweet Share **Sort by Newest**

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Lorena Guadalupe Barberia • 11 hours ago

This is very useful! Thank you! I found an error in one line for Brazil. I know you are only using the data repository, but I think it underscores that users need to check the data carefully for consistency.

Reply • Share ›

inwoner_van_de_stad_Gent • 14 hours ago • edited

It's actually easier if you use their (newly changed) time series data.
My ado file is here:
https://github.com/StataAfi...

Reply • Share ›

> **Lorena Guadalupe Barberia** inwoner_van_de_stad_Gent • 10 hours ago
>
> I tried to download and install your ado, but I am not getting the merged data set. It only saves separate files by date and stops at that point on my computer following your instructions.
>
> Reply • Share ›

Ben Shillitoe • 15 hours ago

Just one quick note for cleaning purposes, along the way UK has been recoded as United Kingdom in the John Hopkins Dataset. Pre-cleaning, there are two data sets for the UK data

Reply • Share ›

Jean-Claude Arbaut • a day ago • edited

As a matter of fact, this character "ï" in the first variable name comes from the byte order mark (here it's "EF BB FF" in hexadecimal). To deal with this, add the option "encoding(utf-8)" to -import delim-. See https://en.wikipedia.org/wi...

Reply • Share ›

Subscribe **Add Disqus to your siteAdd DisqusAdd** **Do Not Sell My Data**

Import COVID-19 data from Johns Hopkins University
RSSTwitterFacebook

关注 5,891

Subscribe to the Stata Blog

Receive email notifications of new blog posts

Name

Email Address*

Subscribe

Recent articles

- Update to Import COVID-19 post
- Import COVID-19 data from Johns Hopkins University
- Just released from Stata Press: *Introduction to Time Series Using Stata, Revised Edition*
- Bayesian inference using multiple Markov chains

- Adding recession shading to time-series graphs

Archives

- 2020
- 2019
- 2018
- 2017
- 2016
- 2015
- 2014
- 2013
- 2012
- 2011
- 2010

Categories

- Blogs
- Company
- Data Management
- Graphics
- Mathematics
    - Linear Algebra
    - Numerical Analysis
- Performance
    - Hardware
    - Memory
    - Multiprocessing
- Programming
    - Mata
- Resources
    - Documentation
    - Meetings
- Stata Products
    - New Books
    - New Products
- Statistics

Tags

#StataProgramming ado ado-command ado-file Bayes Bayesian bayesmh binary biostatistics conference do-file econometrics endogeneity estimation Excel format gmm graphics import marginal effects margins Mata meeting mlexp nonlinear model numerical analysis OLS power precision probit programming putexcel random numbers runiform() sample size SEM simulation Stata matrix command Stata matrix function statistics syntax time series treatment effects users group vector autoregression

Links

- Stata
- Stata Press
- The Stata Journal
- Stata FAQs
- Statalist
- Statalist archives
- Links to others

Top www.stata.com

Terms of use