Reshaping Data Using Black Spruce Data

Andy Grogan-Kaylor

2025-10-11

1 Background

This is a handout about the process of `reshape`-ing data from *wide* to *long* and *vice versa*.

Chihara and Hesterberg (2018) provide a data set concerning the growth of Black Spruce Trees. According to these authors:

> "Black spruce (Picea mariana) is a species of a slow-growing coniferous tree found across the northern part of North America. It is commonly found on wet organic soils. In a study conducted in the 1990s, a biologist interested in factors affecting the growth of the black spruce planted its seedlings on sites located in boreal peatlands in northern Manitoba, Canada (Camill et al. 2010). The data set Spruce contains a part of the data from the study (Table 1.8). Seventy-two black spruce seedlings were planted in four plots under varying conditions (fertilizer–no fertilizer, competition–no competition), and their heights and diameters were measured over the course of 5 years. The researcher wanted to see whether the addition of fertilizer or the removal of competition from other plants (by weeding) affected the growth of these seedlings."

2 Get The Data

```
clear all

use "https://github.com/agrogan1/multilevel/raw/master/reshaping-data/Spruce.dta", clear
```

```
label variable Tree "Tree number"

label variable Competition "C (competition), CR (competition removed)"

label variable Fertilizer "F (fertilized), NF (not fertilized)"

label variable Height0 "Height (cm) of seedling at planting"

label variable Height5 "Height (cm) of seedling at year 5"

label variable Diameter0 "Diameter (cm) of seedling at planting"

label variable Diameter5 "Diameter (cm) of seedling at year 5"

label variable Ht_change "Change (cm) in height"

label variable Di_change "Change (cm) in diameter"
```

3 Describe The Data

```
describe
```

```
Contains data from https://github.com/agrogan1/multilevel/raw/master/reshaping-data/S
> pruce.dta
 Observations:            72
    Variables:             9                  26 Apr 2020 12:18
-------------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------------
Tree            long    %12.0g                Tree number
Competition     long    %12.0g     Competition
                                              C (competition), CR (competition
                                                removed)
Fertilizer      long    %12.0g     Fertilizer
                                              F (fertilized), NF (not fertilized)
Height0         double  %10.0g                Height (cm) of seedling at planting
Height5         double  %10.0g                Height (cm) of seedling at year 5
```

```
Diameter0       double  %10.0g                Diameter (cm) of seedling at planting
Diameter5       double  %10.0g                Diameter (cm) of seedling at year 5
Ht_change       double  %10.0g                Change (cm) in height
Di_change       double  %10.0g                Change (cm) in diameter
-------------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

4 Keep Only Relevant Variables

It is often *very useful* when working with longitudinal data to `keep` only the relevant variables to have a *manageable data set* to work with.

```
keep Tree Competition Fertilizer Height0 Height5 Diameter0 Diameter5
```

5 List Out A Sample Of The Data

```
list in 1/10
```

```
     | Tree   Compet~n   Fertil~r   Height0   Height5   Diameter0   Diamet~5 |
     |-----------------------------------------------------------------------|
  1. |    1         NC          F        15        60    1.984375        7.4 |
  2. |    2         NC          F         9      45.2    1.190625        5.2 |
  3. |    3         NC          F        12        42   1.7859375        5.7 |
  4. |    4         NC          F      13.7      49.5      1.5875        6.4 |
  5. |    5         NC          F        12      47.3      1.5875        6.2 |
     |-----------------------------------------------------------------------|
  6. |    6         NC          F        12      56.4      1.5875        7.4 |
  7. |    7         NC         NF      16.8      43.5    1.984375        4.9 |
  8. |    8         NC         NF      14.6      49.2    1.984375        5.4 |
  9. |    9         NC         NF        16        54    1.984375        7.1 |
 10. |   10         NC         NF      15.4        45    1.984375        5.1 |
     +-----------------------------------------------------------------------+
```

6 Wide Compared To Long Data

The data are currently in *wide* format, where *every row is an individual*, and *every individual has a single row of data.* For a given measure, each time point is in a *different column of data.*

In *long* format, *every row is an individual-observation*, and *every individual has multiple rows of data.* For a given measure, each time point is in the *same column of data*, and the different time points are distinguished by a *time* variable.

7 Reshaping The Data

We are going to `reshape` the data from *wide* format to *long* format.

7.1 Steps In Reshaping Data

1. Only `keep` the relevant variables, as we did just above.
2. `rename` each independent or dependent variable from each time point so that it has the `stub-time` format.

> `stub-time` format
>
> Notice how the variables in this data set are already in the `stub-time` format. If the variables had a different format, e.g. `height_five_years`, `height_zero_years`, it would usually be easier to rename them e.g. `rename height_five_years height5`, and `rename height_zero_years height0`.[1]

3. Look at the data using `browse` or `list` to make sure that the `reshape` command worked properly.

7.2 Use `reshape`

In the reshape command below, notice that we only include the variables that we consider to be *time varying*. Variables that are not included are considered to be *time invariant*. `Tree` is an *id* variable that is already in the data. `year` is a time variable that we are creating. We do not include `Competition` or `Fertilizer` in our reshape command because those are variables that do not change over time.

[1] In recent versions of Stata, there are advanced ways of dealing with variables with names such as `x1suffix`, `x2suffix`, `x3suffix`, etc.. See `help reshape` for information on these new approaches. However, I still find it is often easier to `rename`’ variables before `reshape`-ing them.

```
reshape long Height Diameter, i(Tree) j(year)
```

```
(j = 0 5)

Data                               Wide   ->   Long
-----------------------------------------------------------------------------
Number of observations               72   ->   144
Number of variables                   7   ->   6
j variable (2 values)                     ->   year
xij variables:
                        Height0 Height5   ->   Height
                    Diameter0 Diameter5   ->   Diameter
-----------------------------------------------------------------------------
```

`id` variable

The `id` variable, whatever it is named, has to uniquely identify the observations. A useful command here is `isid`, e.g. `isid id`. If your `id` variable is not unique, it is often due to missing values. `drop if id == .` usually solves the problem. Because `Tree` is the id variable in this dataset, the appropriate command would be `drop if Tree == .`.

7.3 Use `list` To Look At A Sample Of The Data

```
list in 1/20
```

```
     | Tree   year   Compet~n   Fertil~r   Height    Diameter |
     |--------------------------------------------------------|
  1. |    1      0         NC          F       15    1.984375 |
  2. |    1      5         NC          F       60         7.4 |
  3. |    2      0         NC          F        9    1.190625 |
  4. |    2      5         NC          F     45.2         5.2 |
  5. |    3      0         NC          F       12   1.7859375 |
     |--------------------------------------------------------|
  6. |    3      5         NC          F       42         5.7 |
  7. |    4      0         NC          F     13.7      1.5875 |
  8. |    4      5         NC          F     49.5         6.4 |
  9. |    5      0         NC          F       12      1.5875 |
 10. |    5      5         NC          F     47.3         6.2 |
     |------------------------------------------------------|
 11. |    6      0        NC          F      12     1.5875 |
 12. |    6      5        NC          F    56.4        7.4 |
 13. |    7      0        NC         NF    16.8   1.984375 |
 14. |    7      5        NC         NF    43.5        4.9 |
 15. |    8      0        NC         NF    14.6   1.984375 |
     |------------------------------------------------------|
 16. |    8      5        NC         NF    49.2        5.4 |
 17. |    9      0        NC         NF      16   1.984375 |
 18. |    9      5        NC         NF      54        7.1 |
 19. |   10      0        NC         NF    15.4   1.984375 |
 20. |   10      5        NC         NF      45        5.1 |
     +------------------------------------------------------+
```

References

Camill, Philip, Laura Chihara, Brad Adams, Christian Andreassi, A. N. N. Barry, Sahir Kalim, Jacob Limmer, Mike Mandell, and Greg Rafert. 2010. "Early Life History Transitions and Recruitment of Picea Mariana in Thawed Boreal Permafrost Peatlands." *Ecology*. https://doi.org/10.1890/08-1839.1.

Chihara, Laura M., and Tim C. Hesterberg. 2018. *Mathematical Statistics with Resampling and r*. *Mathematical Statistics with Resampling and R*. https://doi.org/10.1002/9781119505969.