Homework 1: Data Fundamentals

A. Short answer

B. Empirical exercise

Submission

Start Over

B. Empirical exercise

In this exercise you will demonstrate basic knowledge about *data structures* and data documentation. Along the way, you will be introduced to a few R commands. You will do this using a data set constructed by David Card for his well-known study analyzing the effect of education on wages:

Card, D., "Using Geographic Variation in College Proximity to Estimate the Return to Schooling", in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, E. Christophides, et al., eds, Toronto: University of Toronto Press (1995).

These data will be featured prominently in Part II of the course when we will replicate some of Card's analysis. For now, we'll take the opportunity describe the key features of his data, as we would if we had conducted his analysis ourselves.

The version of Card's data we will use comes from the `wooldridge` package. You don't have to worry about installing packages in this environment because that has been taken care of for you, but if you were replicating this exercise on your machine you would need to and here's the command to do it:

```
install.packages("wooldridge")
```

Then you would load the package using the `library` function:

```
library()
```

Once loaded, all of the package's exported functions and objects become directly accessible in your R session. This means you can use those functions and objects as if they were part of the base R distribution without needing to reference the package name.

Finally, you may want to explicitly load the Card data into your R environment. The Card data set is named `card` in the `wooldridge` package. (That's `card`, all lower case. Case matters in R.) It's generally not necessary to explicitly load the data with the data function after the relevant package is attached, but it will helpful with the project.

Question 1:

First, use the `library()` and `data()` functions to load the `wooldridge` package and `card` data set.

There will be a few coding questions in the homework assignments that we need to grade so that you are on track to continue. This is one of them. So, before moving on make sure you click Submit Answer. If you have completed the code chunk correctly, you will get a "Correct" response in a green-shaded box below the chunk. Errors will be indicated in a red box.

R Code Start Over Run Code Submit Answer

```r
library(wooldridge)
data(card)
```

Amazing! Correct!

Provenance

Before we look at the structure of the data, let's do a little *provenance* work. (Just a little.) Go to the paper linked to above to answer the next six questions.

Question 2:

Card obtained the data from the _____.

NLSYM

Correct!

Question 3:

The source of Card's data is a survey that began in _____ with _____ young men age 14-24.

1966, 5525

Correct!

Question 4:

The same young men were surveyed again in selected years through _____ , effectively creating a _____ data set where the unit of observation is the person- _____ .

1981, longitudinal, education

Submit Answer

Question 5:

The survey was not a random sample of the US population because men from neighborhoods with a high concentration of _____ residents were over-sampled.

non-white

Correct!

Question 6:

Card's analysis is based on the 1976 survey when the youngest respondents are _____. By 1976, attrition had reduced the sample size to _____ observations. After filtering the sample on observations with valid education and wage data, Card is left with an analysis sample of _____ young men.

24, 3694, educated

Submit Answer

Continue

Data Documentation and Structure

Now let's turn to documentation and structure.

The `wooldridge vignette` provides descriptions of the variables contained in the data set. Use the vignette to answer the next few questions.

Question 7:

The `key` variable in the data set is _____.

id

Correct!

Question 8:

The `wage` variable is measured in _____. The `lwage` variable is the _____ transformation of `wage`.

cents, log

Correct!

Question 9:

The variable `exper` measures labor-market experience as ______.

age - educ - 6

Correct!

The `str()` function, which provides an overview of the data type, size, and content in a data set. Apply it to determine the structure of the `card` data set and answer the questions that follow.

R Code Start Over Run Code

```r
str(card)
```

```
 $ age     : int  29 27 34 27 34 26 33 29 28 29 ...
 $ fatheduc: int  NA 8 14 11 8 9 14 14 12 12 ...
 $ motheduc: int  NA 8 12 12 7 12 14 14 12 12 ...
 $ weight  : num  158413 380166 367470 380166 367470 ...
 $ momdad14: int  1 1 1 1 1 1 1 1 1 1 ...
 $ sinmom14: int  0 0 0 0 0 0 0 0 0 0 ...
 $ step14  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reg661  : int  1 1 1 0 0 0 0 0 0 0 ...
 $ reg662  : int  0 0 0 1 1 1 1 1 1 1 ...
 $ reg663  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reg664  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reg665  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reg666  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reg667  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reg668  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reg669  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ south66 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ black   : int  1 0 0 0 0 0 0 0 0 0 ...
 $ smsa    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ south   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ smsa66  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ wage    : int  548 481 721 250 729 500 565 608 425 515 ...
 $ enroll  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ KWW     : int  15 35 42 25 34 38 41 46 32 34 ...
 $ IQ      : int  NA 93 103 88 108 85 119 108 96 97 ...
 $ married : int  1 1 1 1 1 1 1 1 4 1 ...
 $ libcrd14: int  0 1 1 1 0 1 1 1 0 1 ...
 $ exper   : int  16 9 16 10 16 8 9 9 10 11 ...
 $ lwage   : num  6.31 6.18 6.58 5.52 6.59 ...
 $ expersq : int  256 81 256 100 256 64 81 81 100 121 ...
 - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

Question 10:

The `card` data set contains _____ observations and _____ variables.

3010, 34

Correct!

Question 11:

What data type is `lwage`? _____. How about `wage`? ______. (Use the full-name description of the data type in your answers.)

numeric, integer

Correct!

Question 12:

The third person in the data set is _____ years old, has _____ years of education, has _____ years of experience, and reported a wage of $ ______ .

34, 12, 16, 7.21

Correct!

The `skim()` function provided by the `skimr` package is another useful tool for data documentation Load `skimr` via a `library()` command and then "skim" the `card` data. Answer a few more questions based on the `skim()` output.

R Code Start Over Run Code

```r
library(skimr)
skim(card)
```

Data summary

Name	card
Number of rows	3010
Number of columns	34
_______________________	
Column type frequency:	
numeric	34
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50
id	0	1.00	2581.75	1500.54	2.00	1275.50	2541.00
nearc2	0	1.00	0.44	0.50	0.00	0.00	0.00
nearc4	0	1.00	0.68	0.47	0.00	0.00	1.00
educ	0	1.00	13.26	2.68	1.00	12.00	13.00
age	0	1.00	28.12	3.14	24.00	25.00	28.00
fatheduc	690	0.77	10.00	3.72	0.00	8.00	10.00
motheduc	353	0.88	10.35	3.18	0.00	8.00	12.00
weight	0	1.00	321185.26	170645.80	75607.00	122798.00	365200.00
momdad14	0	1.00	0.79	0.41	0.00	1.00	1.00
sinmom14	0	1.00	0.10	0.30	0.00	0.00	0.00
step14	0	1.00	0.04	0.19	0.00	0.00	0.00
reg661	0	1.00	0.05	0.21	0.00	0.00	0.00
reg662	0	1.00	0.16	0.37	0.00	0.00	0.00
reg663	0	1.00	0.20	0.40	0.00	0.00	0.00
reg664	0	1.00	0.06	0.25	0.00	0.00	0.00
reg665	0	1.00	0.21	0.41	0.00	0.00	0.00
reg666	0	1.00	0.10	0.29	0.00	0.00	0.00
reg667	0	1.00	0.11	0.31	0.00	0.00	0.00
reg668	0	1.00	0.03	0.17	0.00	0.00	0.00
reg669	0	1.00	0.09	0.29	0.00	0.00	0.00
south66	0	1.00	0.41	0.49	0.00	0.00	0.00
black	0	1.00	0.23	0.42	0.00	0.00	0.00
smsa	0	1.00	0.71	0.45	0.00	0.00	1.00
south	0	1.00	0.40	0.49	0.00	0.00	0.00
smsa66	0	1.00	0.65	0.48	0.00	0.00	1.00
wage	0	1.00	577.28	262.96	100.00	394.25	537.50
enroll	0	1.00	0.09	0.29	0.00	0.00	0.00
KWW	47	0.98	33.54	8.61	4.00	28.00	34.00
IQ	949	0.68	102.45	15.42	50.00	93.00	103.00
married	7	1.00	2.27	2.07	1.00	1.00	1.00
libcrd14	13	1.00	0.67	0.47	0.00	0.00	1.00
exper	0	1.00	8.86	4.14	0.00	6.00	8.00
lwage	0	1.00	6.26	0.44	4.61	5.98	6.29
expersq	0	1.00	95.58	84.62	0.00	36.00	64.00

Question 13:

How many variables have missing data? _____ .

6

Correct!

Question 14:

What percentage of young men in the sample are missing `IQ` test scores? _____ % . (Answer to 1 decimal place, for example: ``99.9'' percent)

31.5

Correct!

Question 15:

What percentage of the sample are Black? _____ %. Is that representative of the US population in 1976? (Yes/No) _____ .

23, no

Correct!

Finally, use the `object.size` function to estimate the amount of memory allocated to store the Card data.

R Code Start Over Run Code

```r
memory_size <- object.size(card)
print(memory_size)
```

```
438416 bytes
```

```
[1] 0.4181061
```

Question 16:

Based on `object.size` the Card data take up _____ MB in memory. (Round to 3 digits)

0.418

Correct!

Continue

Previous Topic Next Topic