Translating Data from MySQL to Stata

Michael Johnson and Phil Schumm

Research Computing Group
Department of Health Studies
University of Chicago

August 23, 2004

(Supported by National Institute on Aging grant P01 AG18911-01A1)

MySQL

What is MySQL?

- Open-Source database management system
- Cross platform
- High performance
- Popular back-end for web applications

MySQL and Research

- CAPI using MySQL backend
- Online data collection (e.g., multi-center clinical trial)

Data Extraction Goals

Our goals for a data extraction method:

- Reproducible/Trackable
  - Text-based representation
  - Preserve when data was generated, where it was extracted from
  - Preserve basic information regarding how tables were set-up
  - Include this information within dataset itself.
- Cross-platform
- Simple

MyStata

MyStata is:

- Python script
- Uses MySQLdbAPI
- Generates pair of .do and .dct files for each MySQL table

Basic Example

```
Table Creation:
CREATE TABLE foobar(
foo INT,
bar TEXT
);

INSERT INTO foobar('foo', 'bar') VALUES (1, 'Hello');
INSERT INTO foobar('foo', 'bar') VALUES (2, 'World');
```

Basic Example

Invocation:
```
python MyStata.py -d test -u mjohnson foobar
```

Generates:
```
foobar.do and foobar.dct
```

Basic Example

```
Stata Run:
. do foobar
                ...
. list

     +-------------+
     | foo     bar |
     |-------------|
  1. |   1   Hello |
  2. |   2   World |
     +-------------+

. notes li

_dta:
  1.  Extracted by MyStata from database test on 2004-08-19 16:07:17

foo:
  1.  Column foo, Type = int(11),Null = YES, Key =

bar:
  1.  Column bar, Type = text,Null = YES, Key =
```

ENUM fields

- In MySQL, these are strings columns where possible values are restricted to specific set.
  - smoker ENUM('Yes','No', 'Refused', 'Unknown')
- Encoded as numeric variables in Stata dataset, with appropriate value label attached.
- Want to preserve possible options even if all possible values are not used.

ENUM example

```
CREATE TABLE enum_example(
subject char(4),
smoker enum('Yes','No', 'Refused', 'Unknown'),
affected enum('Yes','No', 'Refused', 'Unknown')
);

insert into enum_example(subject, smoker, affected)
    values ('0001','Yes','No');
insert into enum_example(subject, smoker, affected)
            values ('0002','No','Unknown');
```

ENUM example

```
. do enum_example

. list

     +-----------------------------+
     | subject   smoker   affected |
     |-----------------------------|
  1. |    0001      Yes         No |
  2. |    0002       No    Unknown |
     +-----------------------------+
```

ENUM example

```
. describe

Contains data
  obs:             2
 vars:             3
 size:            32 (99.9% of memory free)   (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
subject         str4   %9s                    *
smoker          long   %8.0g       smoker     *
affected        long   %8.0g       affected   *
                                               * indicated variables have notes
-------------------------------------------------------------------------------
```

ENUM example

```
. label list
affected:
           1 No
           2 Refused
           3 Unknown
           4 Yes
smoker:
           1 No
           2 Refused
           3 Unknown
           4 Yes
```

SET columns

- These represents sets of strings
- MySQL returns a string containing a comma separated list (not useful for analysis)
- Variable is created for each possible member of the set

Other Issues

- Column names
- String lengths
- Fuzzy dates
- Time