



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Stata tip 89: Estimating means and percentiles following multiple imputation

Peter A. Lachenbruch
Oregon State University
Corvallis, OR
peter.lachenbruch@oregonstate.edu

1 Introduction

In a statistical analysis, I usually want some basic descriptive statistics such as the mean, standard deviation, extremes, and percentiles. See, for example, Pagano and Gauvreau (2000). Stata conveniently provides these descriptive statistics with the `summarize` command's `detail` option. Alternatively, I can obtain percentiles with the `centile` command. For example, with `auto.dta`, we have

```
. sysuse auto
(1978 Automobile Data)
. summarize price, detail
```

Price			
	Percentiles	Smallest	
1%	3291	3291	
5%	3748	3299	
10%	3895	3667	Obs 74
25%	4195	3748	Sum of Wgt. 74
50%	5006.5		Mean 6165.257
		Largest	Std. Dev. 2949.496
75%	6342	13466	
90%	11385	13594	Variance 8699526
95%	13466	14500	Skewness 1.653434
99%	15906	15906	Kurtosis 4.819188

However, if I have missing values, the `summarize` command is not supported by `mi estimate` or by the user-written `mim` command (Royston 2004, 2005a,b, 2007; Royston, Carlin, and White 2009).

2 Finding means and percentiles when missing values are present

For a general multiple-imputation reference, see *Stata 11 Multiple-Imputation Reference Manual* (2009). By recognizing that a regression with no independent variables estimates the mean, I can use `mi estimate: regress` to get multiply imputed means. If I wish to get multiply imputed quantiles, I can use `mi estimate: qreg` or `mi estimate: sqreg` for this purpose.

I now create a dataset with missing values of `price`:

```
. clonevar newprice = price
. set seed 19670221
. replace newprice = . if runiform() < .4
(32 real changes made, 32 to missing)
```

The following commands were generated from the multiple-imputation dialog box. I used 20 imputations. Before Stata 11, this could also be done with the user-written commands `ice` and `mim` (Royston 2004, 2005a,b, 2007; Royston, Carlin, and White 2009).

```
. mi set mlong
. mi register imputed newprice
(32 m=0 obs. now marked as incomplete)
. mi register regular mpg trunk weight length
. mi impute regress newprice, add(20) rseed(3252010)
```

```
Univariate imputation      Imputations =      20
Linear regression          added =      20
Imputed: m=1 through m=20  updated =      0
```

Variable	Observations per <i>m</i>			
	complete	incomplete	imputed	total
newprice	42	32	32	74

(complete + incomplete = total; imputed is the minimum across *m* of the number of filled in observations.)

```
. mi estimate: regress newprice
```

```
Multiple-imputation estimates      Imputations =      20
Linear regression                  Number of obs =      74
                                   Average RVI   =     1.3880
                                   Complete DF    =      73
                                   DF:      min    =     19.46
                                   avg          =     19.46
                                   max          =     19.46
DF adjustment: Small sample
                                   F( 0, .) = .
Within VCE type: OLS              Prob > F    = .
```

newprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_cons	5693.489	454.9877	12.51	0.000	4742.721 6644.258

From this output, we see that the estimated mean is 5,693 with a standard error of 455 (rounded up) compared with the complete data value of 6,165 with a standard error of 343 (also rounded up). However, we do not have estimates of quantiles. This could also have been done using `mi estimate: mean newprice` (the `mean` command is near the bottom of the estimation command list for `mi estimate`).

We can apply the same principle using `qreg`. For the 10th percentile, type

```
. mi estimate: qreg newprice, quantile(10)
Multiple-imputation estimates      Imputations      =      20
.1 Quantile regression           Number of obs   =      74
                                Average RVI      =     0.2901
                                Complete DF      =      73
                                DF:      min      =     48.05
                                avg       =     48.05
                                max       =     48.05
DF adjustment:      Small sample
                                F(      0,      .) =      .
                                Prob > F      =      .
```

newprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	3495.635	708.54	4.93	0.000	2071.058	4920.212

Compare the value of 3,496 with the value of 3,895 from the full data. We can use the simultaneous estimates command for the full set:

```
. mi estimate: sqreg newprice, quantiles(10 25 50 75 90) reps(20)
Multiple-imputation estimates      Imputations      =      20
Simultaneous quantile regression  Number of obs   =      74
                                Average RVI      =     0.6085
                                Complete DF      =      73
DF adjustment:      Small sample   DF:      min      =     23.19
                                avg       =     26.97
                                max       =     31.65
```

newprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
q10 _cons	3495.635	533.5129	6.55	0.000	2408.434	4582.836
q25 _cons	4130.037	237.1932	17.41	0.000	3642.459	4617.614
q50 _cons	5200.238	441.294	11.78	0.000	4292.719	6107.757
q75 _cons	6620.232	778.8488	8.50	0.000	5025.49	8214.974
q90 _cons	8901.985	1417.022	6.28	0.000	5971.962	11832.01

3 Comments and cautions

The `qreg` command does not give the same result as the `centile` command when you have complete data. This is because the `centile` command uses one observation, while the `qreg` command uses a weighted combination of the observations. It will have somewhat shorter confidence intervals, but with large datasets, the difference will be

small. A second caution is that comparing two medians can be tricky: the difference of two medians is not the median difference of the distributions. I have found it useful to use percentiles because there is a one-to-one relationship between percentiles if data are transformed. In our case, there is plentiful evidence that **price** is not normally distributed, so it would be good to look for a transformation and impute those values.

This method of using regression commands without an independent variable can provide estimates of quantities that otherwise would be difficult to obtain. For example, it is much faster than finding 20 imputed percentiles and then combining them with Rubin's rules, and it is much less onerous and prone to error.

4 Acknowledgment

This work was supported in part by a grant from the Cure JM Foundation.

References

- Pagano, M., and K. Gauvreau. 2000. *Principles of Biostatistics*. 2nd ed. Belmont, CA: Duxbury.
- Royston, P. 2004. Multiple imputation of missing values. *Stata Journal* 4: 227–241.
- . 2005a. Multiple imputation of missing values: Update. *Stata Journal* 5: 188–201.
- . 2005b. Multiple imputation of missing values: Update of ice. *Stata Journal* 5: 527–536.
- . 2007. Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *Stata Journal* 7: 445–464.
- Royston, P., J. B. Carlin, and I. R. White. 2009. Multiple imputation of missing values: New features for mim. *Stata Journal* 9: 252–264.
- StataCorp. 2009. *Stata 11 Multiple-Imputation Reference Manual*. College Station, TX: Stata Press.