Quantile Imputation of Missing Data

Matteo Bottai

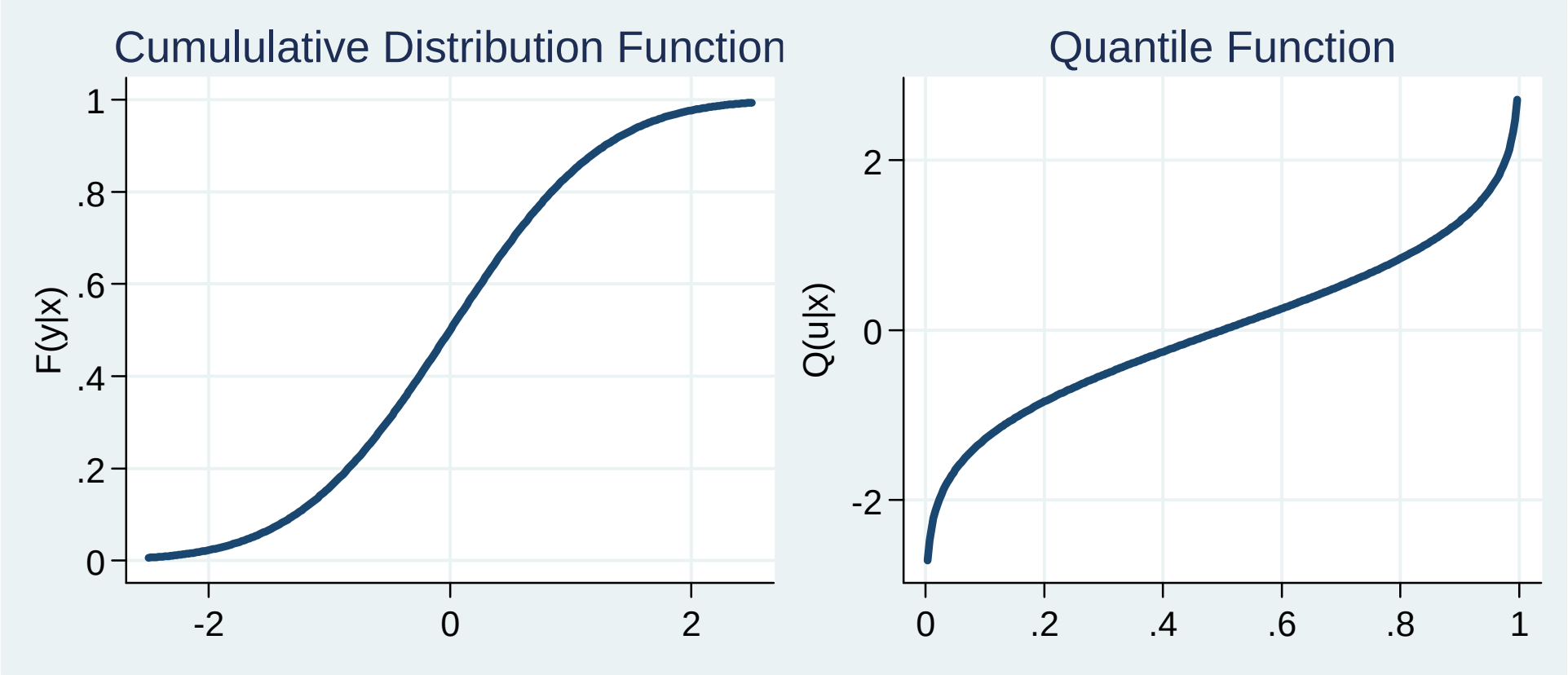
Thanks to

Nicola Orsini and Yulia Marchenko

Conditional Quantile Function

Let $F(y|x)$ be the cumulative distribution function of Y given x.
The quantile function of Y given x is

$$Q(u|x) = \inf\{y: F(y|x) \geq u\}$$



Quantile Imputation

Probability transformation theorem (Casella & Berger, 1990)

If $Y \sim F(y|x)$ and $U \sim U(0,1)$
Let $W = Q(U|x)$
Then $W \sim F(y|x)$

Suppose $Y_i \sim F(y|x_i)$ and $U \sim U(0,1)$ (Bottai & Zhen, 2011)

If $\hat{Q}(u|x_i) \xrightarrow{P} Q(u|x_i)$ for every $u \in (0,1)$
Let $\hat{Y}_i = \hat{Q}(U|x_i)$
Then $\hat{Y}_i \xrightarrow{D} F(y|x_i)$

Conditional Quantile Estimators

There are several possible estimators for $Q(u|x_i)$.

For this presentation we use quantile regression (Koenker, 1978)

Possible nonparametric alternatives include:

- Local logistic model and adjusted Nadaraya-Watson (Hall et al, 1999)
- Mixed data types (Li & Racine, 2008)

Example 1: `mi laplace` with missing data (1 of 3)

x1	x2	y	y_miss
0.31	1	7.70	7.70
0.69	0	7.75	7.75
0.56	1	9.75	.
0.37	1	7.52	7.52

Example 1: `mi laplace` with missing data (2 of 3)

```
. mi register imputed y_miss
(308 m=0 obs. now marked as incomplete)

. mi laplace y_miss x1 x2 , add(5)

Multiple imputation                             Imputations =        5
Laplace regression                                    added =        5
Imputed: m=1 through m=5                            updated =        0

------------------------------------------------------------------------
                     |                Observations per m
                     |-----------------------------------------------
            Variable |   Complete   Incomplete   Imputed |     Total
---------------------+-----------------------------------+-----------
             y_miss |        692          308       308 |      1000
------------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Example 1: `mi laplace` with missing data (3 of 3)

```
. mi estimate: regress y_miss
Multiple-imputation estimates                     Imputations     =          5
Linear regression                                 Number of obs   =       1000
                                                  Average RVI     =     0.1363
                                                  Largest FMI     =     0.1264
                                                  Complete DF     =        999
                                                  DF:     min     =     211.06
                                                          avg     =     211.06
DF adjustment:   Small sample                             max     =     211.06
                                                  F(   0,      .) =          .
Within VCE type:          OLS                     Prob > F        =          .
------------------------------------------------------------------------------
      y_miss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   7.362527   .0761562    96.68   0.000     7.212403    7.512652
------------------------------------------------------------------------------

. regress y
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   7.363576   .0729357   100.96   0.000     7.220451      7.5067
------------------------------------------------------------------------------
```

Example 2: `mi laplace` with missing and censored data (1 of 3)

x1	y	y_miss_cens	cens
0.31	7.70	7.70	.
0.69	7.75	.	6.66
0.56	9.75	.	6.90
0.37	7.52	.	.

Example 2: `mi laplace` with missing and censored data (2 of 3)

```
. mi register imputed y_miss_cens
(515 m=0 obs. now marked as incomplete)

. mi laplace y_miss_cens x1 x2 , add(5) c(cens)

Missing observations         214
Censored observations        301
Uncensored observations      485

Multiple imputation                          Imputations =        5
Laplace regression                                added =        5
Imputed: m=1 through m=5                        updated =        0

------------------------------------------------------------------
                    |               Observations per m
                    |-----------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
       y_miss_cens |        485          515       515 |      1000
------------------------------------------------------------------
```

Example 2: `mi laplace` with missing and censored data (3 of 3)

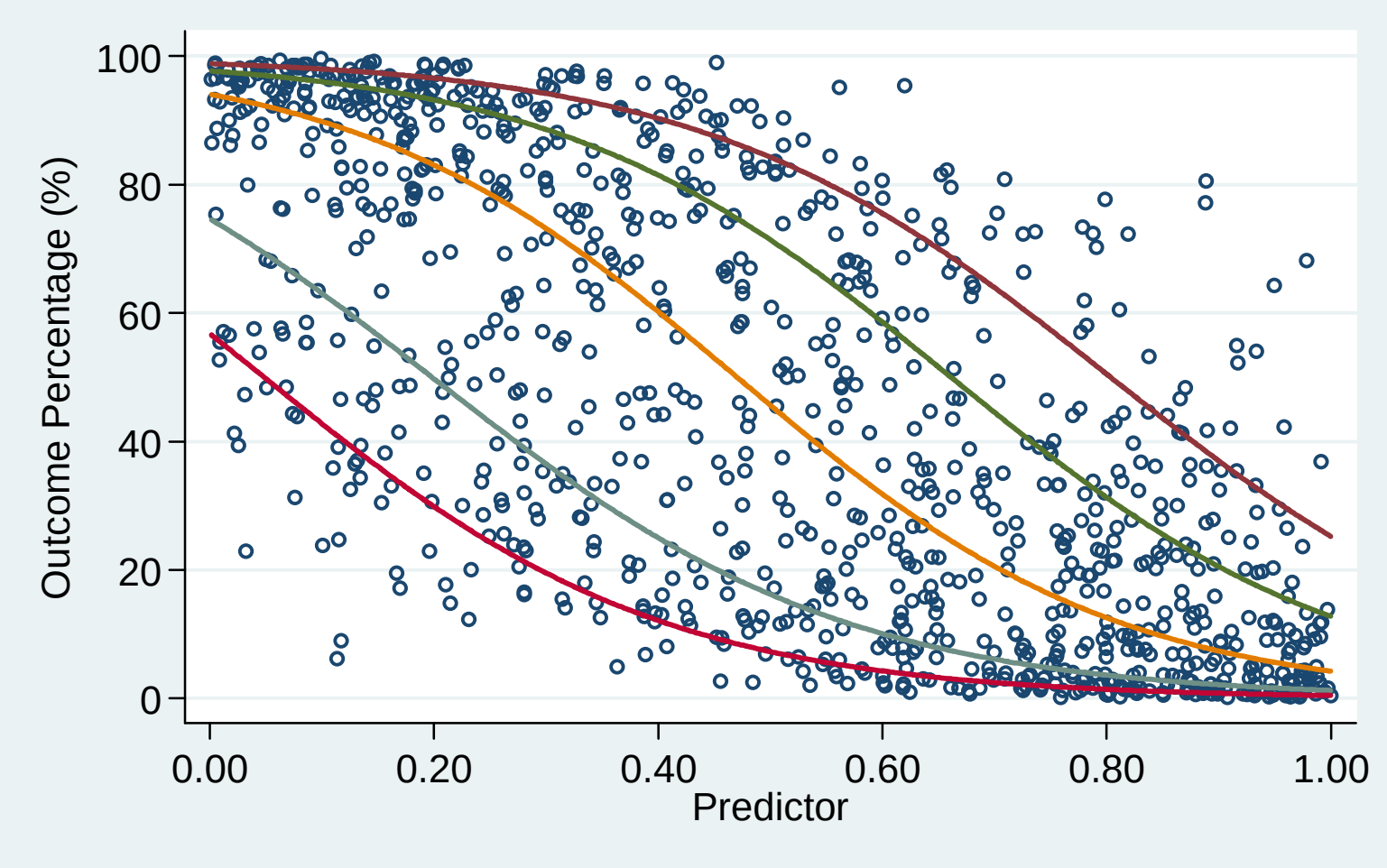
```
. mi estimate: laplace y_miss_cens x1 x2

------------------------------------------------------------------------------
 y_miss_cens |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   5.396186   .1605569    33.61   0.000     5.078263    5.714109
          x2 |   2.552206   .1042797    24.47   0.000     2.343483    2.760929
       _cons |   3.514986   .1013782    34.67   0.000     3.314881     3.71509
------------------------------------------------------------------------------

. laplace y x1 x2

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q50          |
          x1 |   5.429232   .1386361    39.16   0.000      5.15718    5.701284
          x2 |    2.59261   .0830899    31.20   0.000     2.429559    2.755661
       _cons |   3.561758   .0702685    50.69   0.000     3.423867    3.699649
------------------------------------------------------------------------------
```

Example 3: `mi laplace` with transformed data (1 of 2)



Example 3: `mi laplace` with transformed data (2 of 2)

Unlike the mean, quantiles are invariant to increasing transformations

$$Q_Y(u) = g^{-1}\{Q_{g(Y)}(u)\}$$

One can

1. Apply a convenient transformation to the outcome
2. Impute the missing values of the transformed variable
3. Transform the imputed values back to the original scale

For example, for bounded outcomes one can use a logit transform (Bottai et al, 2011). This ensures imputed values are within feasible region and may facilitate modeling.

Summary

Quantile Imputation

- ✓ Can impute ignorable missing data and non-informative censoring
- ✓ Works seamlessly with transformations (e.g. bounded outcomes)
- ✓ Can be applied to dependent data (e.g. repeated measures)

- ✘ Is computational slow
- ✘ Requires modeling quantile functions or selecting smoothing parameters