12. Multiple Imputation For Questionnaire Data

Scale Scores

Single questions items lack validity because they tap into only one aspect of a complex variable

A scale score is composite variable computed by summing or averaging the responses to questionnaire items measuring the same theme

Scale scores so a better job of characterizing psychological characteristics than do single items

Prorated Scale Scores (Averaging The Available Items)

Prorated scale scores are computed by averaging the available item responses

e.g., A respondent answers 4 of 10 items, the scale score is the mean of the 4 complete items

Equivalent to imputing with a person's mean

Proration Example

Proration

ID	Q_1	Q_2	Q_3	Scale
1	1	2	1	1.3
2	5	NA	4	4.5
3	3	2	4	3.0
4	NA	3	NA	3.0

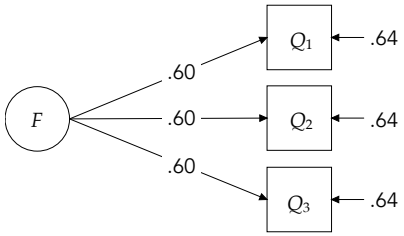
Person-Mean Imputation

ID	Q_1	Q_2	Q_3	Scale
1	1	2	1	1.3
2	5	4.5	4	4.5
3	3	2	4	3.0
4	3.0	3	3.0	3.0

Proration Requirements

Proration requires two very strict assumptions: identical item means and inter-item correlations (parallel factor structure) and an MCAR mechanism

	Q_1	Q_2	Q_3
Q_1	1.00		
Q_2	0.36	1.00	
Q_3	0.36	0.36	1.00
Means	3.00	3.00	3.00



Scale-Level And Item-Level Imputation

Scale-level imputation: (a) compute scale scores, treating the scale as missing when one or more items is missing, (b) impute the scale scores

Item-level imputation: (a) impute the items, (b) compute the scale score from the filled-in items

Item-level imputation usually provides a (VERY) dramatic gain in precision

Imputation Strategy 1: Impute Scales And Ignore Items

Original data

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Imputation variables

ID	$Scale_X$	$Scale_Y$
1	4	NA
2	NA	NA
3	9	7
4	NA	10
…		
200	13	7

Imputation Strategy 2: Impute Items Then Compute Scales

Original data

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Imputation variables

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Body Attitudes Data

Questionnaire data from a study of body attitudes in a sample of 500 middle school students

Variables include body mass index (BMI), five questionnaire items measuring negative body attitudes, and past history of being bullied (0 = never bullied, 1 = history of being bullied)

All questionnaire items measured on a 7-point scale

bodyatt.dat

Variable	Name	Missing %	Scaling
Identifier variable	ID	0	Integer index
History of being bullied	BULLIED	10.4	0 = not bullied, 1 = bullied
Body mass index	BMI	8.0	Continuous
Attitude item 1 (hips too broad)	BATT1	13.4	7-point ordinal scale
Attitude item 2 (desire to be thinner)	BATT2	12.2	7-point ordinal scale
Attitude item 3 (too thick)	BATT3	0	7-point ordinal scale
Attitude item 4 (body looks swollen)	BATT4	12.4	7-point ordinal scale
Attitude item 5 (belly looks pregnant)	BATT5	0	7-point ordinal scale

Negative Body Attitudes Questionnaire

	Strongly Disagree ⟷ Strongly Agree						
1. My hips seem to broad to me.	1	2	3	4	5	6	7
2. I have a strong desire to be thinner.	1	2	3	4	5	6	7
3. I think I'm too thick.	1	2	3	4	5	6	7
4. Some parts of my body look swollen.	1	2	3	4	5	6	7
5. My belly looks as if I'm pregnant.	1	2	3	4	5	6	7

Substantive Analysis

Negative body attitudes scale score regressed on body mass index and a bullied indicator

$$BODYATT_i = \beta_0 + \beta_1(BMI_i) + \beta_2(BULLIED_i) + \varepsilon_i$$

Multiply impute the missing item responses, then compute and analyze the composite

FCS Imputation Model Sequence

Current iteration — Previous iteration

Impute Y_1 conditional on current Y_2 and Y_3

$$Y_{1i}^{(t)} = \gamma_{10} + \gamma_{11}\left(Y_{2i}^{(t-1)}\right) + \gamma_{12}\left(Y_{3i}^{(t-1)}\right) + e_{1i}$$

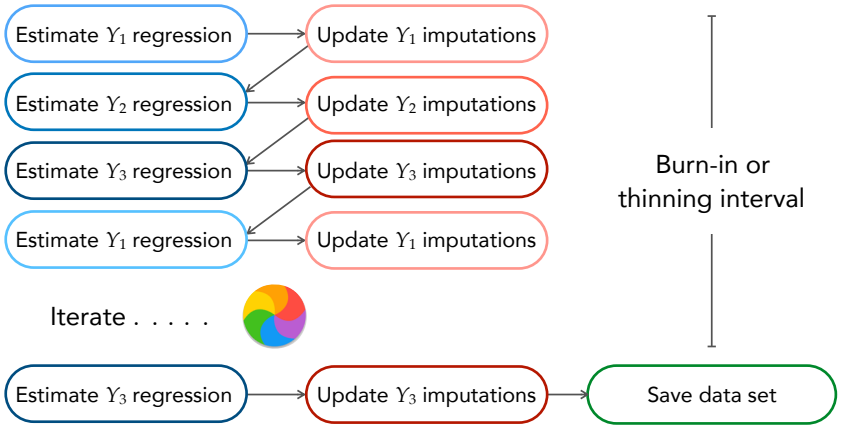
Impute Y_2 conditional on Y_3 and updated Y_1

$$Y_{2i}^{(t)} = \gamma_{20} + \gamma_{21}\left(Y_{1i}^{(t)}\right) + \gamma_{22}\left(Y_{3i}^{(t-1)}\right) + e_{2i}$$

Impute Y_3 conditional on updated Y_1 and Y_2

$$Y_{3i}^{(t)} = \gamma_{30} + \gamma_{31}\left(Y_{1i}^{(t)}\right) + \gamma_{32}\left(Y_{2i}^{(t)}\right) + e_{3i}$$

MCMC Algorithm



Imputation Regression Models

MCMC models the underlying latent scores for all categorical variables, threshold parameters convert discrete imputes for output data

$$BULLIED^{*(t)} \sim + BMI^{(t-1)} + BATT_1^{*(t-1)} + BATT_2^{*(t-1)} + BATT_3^{*(t-1)} + BATT_4^{*(t-1)} + BATT_5^{*(t-1)}$$

$$BMI^{(t)} \sim BULLIED^{*(t)} + BATT_1^{*(t-1)} + BATT_2^{*(t-1)} + BATT_3^{*(t-1)} + BATT_4^{*(t-1)} + BATT_5^{*(t-1)}$$

$$BATT_1^{*(t)} \sim BULLIED^{*(t)} + BMI^{(t)} + BATT_2^{*(t-1)} + BATT_3^{*(t-1)} + BATT_4^{*(t-1)} + BATT_5^{*(t-1)}$$

$$BATT_2^{*(t)} \sim BULLIED^{*(t)} + BMI^{(t)} + BATT_1^{*(t)} + BATT_3^{*(t-1)} + BATT_4^{*(t-1)} + BATT_5^{*(t-1)}$$

$$BATT_3^{*(t)} \sim BULLIED^{*(t)} + BMI^{(t)} + BATT_1^{*(t)} + BATT_2^{*(t)} + BATT_4^{*(t-1)} + BATT_5^{*(t-1)}$$

$$BATT_4^{*(t)} \sim BULLIED^{*(t)} + BMI^{(t)} + BATT_1^{*(t)} + BATT_2^{*(t)} + BATT_3^{*(t)} + BATT_5^{*(t-1)}$$

$$BATT_5^{*(t)} \sim BULLIED^{*(t)} + BMI^{(t)} + BATT_1^{*(t)} + BATT_2^{*(t)} + BATT_3^{*(t)} + BATT_4^{*(t)}$$

Blimp Imputation Script For Analysis In Mplus

```
DATA: bodyatt.dat;
VARIABLES: id bullied bmi batt1 batt2 batt3 batt4 batt5;
ORDINAL: bullied batt1-batt5;
MISSING: 999;
FCS: bullied bmi batt1-batt5;
SEED: 90291;
NIMPS: 20;
BURN: 20000;
THIN: 10000;
CHAINS: 4 processors 4;
OPTIONS: psr;
SAVE: separate = imps_*.dat;
```

Blimp Imputation Script For Analysis In R, SAS, SPSS, Stata

```
DATA: bodyatt.dat;
VARIABLES: id bullied bmi batt1 batt2 batt3 batt4 batt5;
ORDINAL: bullied batt1-batt5;
MISSING: 999;
FCS: bullied bmi batt1-batt5;
SEED: 90291;
NIMPS: 20;
BURN: 20000;
THIN: 10000;
CHAINS: 4 processors 4;
OPTIONS: psr;
SAVE: stacked0 = imps_stacked.dat;
```


Output Data Information

Stacked file format (R, SAS, SPSS, Stata)

```
VARIABLE ORDER IN SAVED DATA:

  imp# id bullied bmi batt1 batt2 batt3 batt4 batt5
```

Separate file format (Mplus)

```
VARIABLE ORDER IN SAVED DATA:

  id bullied bmi batt1 batt2 batt3 batt4 batt5
```

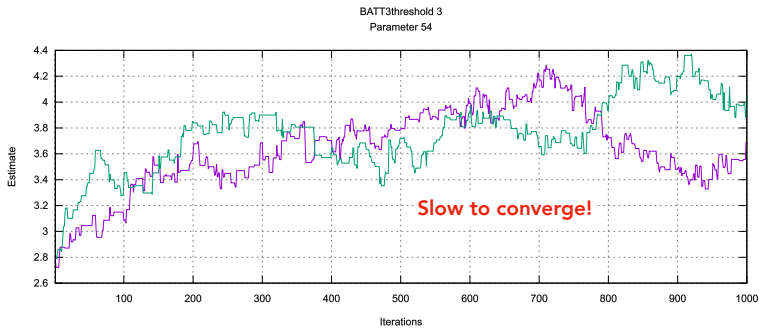

Blimp Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations across 4 chains   | Highest PSR | Parameter # |
                          51 to 100      |       2.024 |          54 |
                         101 to 200      |       2.263 |          29 |
                                    ...
                        8301 to 16600    |       1.075 |          54 |
                        8351 to 16700    |       1.072 |          55 |
                        8401 to 16800    |       1.066 |          55 |
                        8451 to 16900    |       1.058 |          55 |
                        8501 to 17000    |       1.053 |          55 |
                        8551 to 17100    |       1.048 |          55 |
                        8601 to 17200    |       1.045 |          55 |
                        8651 to 17300    |       1.042 |          55 |
                        8701 to 17400    |       1.042 |          55 |
                        8751 to 17500    |       1.041 |          55 |
```

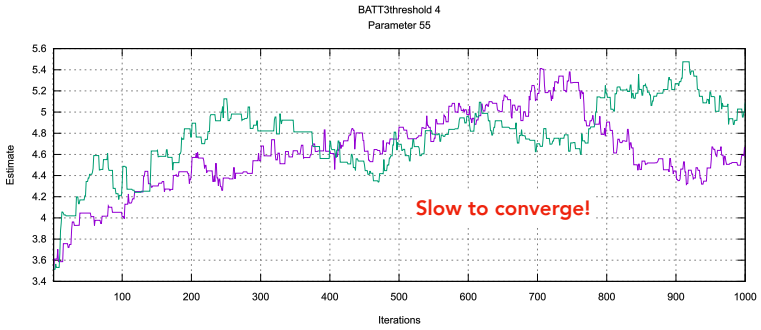

Blimp Trace Plots

Thresholds from 1000 iterations (parameter #54)



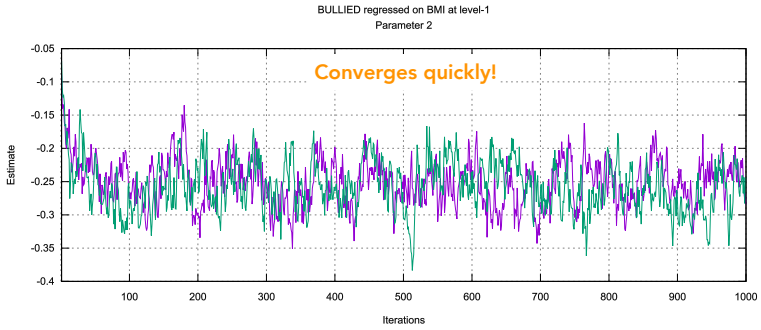
Blimp Trace Plots

Thresholds from 1000 iterations (parameter #55)



Blimp Trace Plots

Slope estimates from 1000 iterations



Summary Of Pooled Estimates

Analysis results from 20 imputed data sets

Parameter	Est.	*SE*	*z*	*p*
Intercept	20.00	0.18	110.61	< .001
BMI slope	0.44	0.06	6.90	< .001
BULLIED slope	4.57	0.49	9.37	< .001

Interpretations

The negative body attitude scale mean for the not bullied group (the intercept) is 20

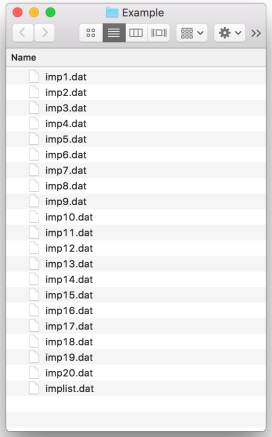
Controlling for BMI, the attitude mean for the bullied group was 4.57 points higher

Being consistently bullied is associated with worse body attitudes (higher negative attitudes)

Mplus Imputation Format

Mplus requires imputed data sets as separate files

Blimp creates a text file containing the names of the data sets, and this file serves as the input data for subsequent analyses



Mplus Imputation Analysis Script

```
DATA:
file = imps_list.dat;
type = imputation;
VARIABLE:
names = id bullied bmi batt1 batt2 batt3 batt4 batt5;
usevariables = bullied bmi bodyatt;
DEFINE:
bodyatt = batt1 + batt2 + batt3 + batt4 + batt5;
center bmi (grandmean);
MODEL:
bodyatt on bmi (b1)
  bullied (b2);
MODEL TEST:
b1 = 0; b1 = b2;
OUTPUT:
stdyx;
```


Mplus Output

```
MODEL FIT INFORMATION

Number of Free Parameters                        4

...

Wald Test of Parameter Constraints

          Value                            104.578
          Degrees of Freedom                 2
          P-Value                          0.0000
```


Mplus Output

```
MODEL RESULTS

                                                    Two-Tailed   Rate of
                  Estimate       S.E.  Est./S.E.    P-Value   Missing

 BODYATT   ON
    BMI              0.444      0.064      6.902      0.000     0.237
    BULLIED          4.566      0.487      9.370      0.000     0.116

 Intercepts
    BODYATT         20.004      0.181    110.609      0.000     0.066

 Residual Variances
    BODYATT         12.669      0.851     14.881      0.000     0.115
```

Mplus Output

```
STANDARDIZED MODEL RESULTS

...

R-SQUARE

    Observed                                        Two-Tailed   Rate of
    Variable        Estimate       S.E.  Est./S.E.    P-Value    Missing

    BODYATT            0.201      0.035      5.818      0.000      0.142
```


R Imputation Analysis Script

```
library(mitml)
library(plyr)

# read stacked imputation data
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
impdata <- read.table(paste0(getwd(), "/imps_stacked.dat"), header = F)
names(impdata) <- c("imputation", "id", "bullied", "bmi", "batt1", "batt2", "batt3", "batt4", "batt5")
impdata <- impdata[impdata$imputation > 0, ]

# compute scale score in each imputed data set and center covariate
impdata$bodyatt <- impdata$batt1+impdata$batt2+impdata$batt3+impdata$batt4+impdata$batt5
impdata <- ddply(impdata, c("imputation"), transform, bmic = scale(bmi, center = T, scale = F))

# analysis and pooling
implist <- as.mitml.list(split(impdata, impdata$imputation))
analysis <- with(implist, lm(bodyatt ~ bmi + bullied))
estimates <- testEstimates(analysis, var.comp = T)
estimates
```


R Imputation Analysis Script

```
# estimate empty model with no predictors
emptymodel <- with(implist, lm(bodyatt ~ 1))

# compare models with Wald test (e.g., MI version of omnibus F test)
testModels(analysis, emptymodel, method = "D1")

# compare models with likelihood ratio test (e.g., MI version of chi-square diff)
testModels(analysis, emptymodel, method = "D3")
```


R Output

```
Final parameter estimates and inferences obtained from 20 imputed data sets.

             Estimate Std.Error   t.value         df   P(>|t|)       RIV       FMI
(Intercept)    20.004     0.181   110.307   4472.479     0.000     0.070     0.066
bmic            0.444     0.064     6.886    354.943     0.000     0.301     0.236
bullied         4.566     0.489     9.345   1460.244     0.000     0.129     0.115

                   Estimate
Residual~~Residual   12.746

Unadjusted hypothesis test as appropriate in larger samples.
```

R Output

```
Model comparison calculated from 20 imputed data sets.
Combination method: D1

    F.value      df1      df2    P(>F)      RIV
     51.632        2 1065.081    0.000    0.207

Unadjusted hypothesis test as appropriate in larger samples.



Model comparison calculated from 20 imputed data sets.
Combination method: D3

   F.value      df1      df2    P(>F)      RIV
    44.327        2 740.157    0.000    0.259
```


SAS Imputation Analysis Script

```
/* read stacked imputation data */
data impdata (where = (_imputation_ gt 0));
infile '/folders/myfolders/imps_stacked.dat';
input _imputation_ id bullied bmi batt1 batt2 batt3 batt4 batt5;
bodyatt = batt1 + batt2 + batt3 + batt4 + batt5;
run;

/* center covariate at grand means in each data set */
proc means data = impdata noprint;
var bmi;
by _imputation_;
output out = grandmeans (drop =_type_ _freq_) mean = bmimean; run;

data impdata;
merge impdata grandmeans;
by _imputation_;
bmic = bmi - bmimean;  run;
```


SAS Imputation Analysis Script

```
/* analyze imputations */
proc reg data = impdata outest = estimates covout noprint;
model bodyatt = bmic bullied;
by _imputation_;
run;

/* pool estimates and standard errors */
proc mianalyze data = estimates;
modeleffects Intercept bmic bullied;
run;

/* omnibus test of model fit */
proc mianalyze data = estimates mult;
modeleffects bmic bullied;
run;
```


SAS Output

The MIANALYZE Procedure

Model Information	
Data Set	WORK.ESTIMATES
Number of Imputations	20

Variance Information (20 Imputations)							
	Variance				Relative	Fraction	
Parameter	Between	Within	Total	DF	Increase in Variance	Missing Information	Relative Efficiency
Intercept	0.002042	0.030745	0.032888	4472.5	0.069723	0.065596	0.996731
bmic	0.000917	0.003197	0.004160	354.94	0.301008	0.235660	0.988354
bullied	0.025930	0.211461	0.238687	1460.2	0.128755	0.115279	0.994269

Parameter Estimates (20 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	20.004302	0.181352	19.64876	20.35984	4472.5	19.921946	20.086056	0	110.31	<.0001
bmic	0.444112	0.064496	0.31727	0.57096	354.94	0.394290	0.492267	0	6.89	<.0001
bullied	4.565642	0.488556	3.60729	5.52399	1460.2	4.157527	4.783428	0	9.35	<.0001

SAS Output

The MIANALYZE Procedure

Model Information	
Data Set	WORK.ESTIMATES
Number of Imputations	20

Variance Information (20 Imputations)

Parameter	Variance: Between	Variance: Within	Variance: Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
bmic	0.000917	0.003197	0.004160	354.94	0.301008	0.235660	0.988354
bullied	0.025930	0.211461	0.238687	1460.2	0.128755	0.115279	0.994269

Parameter Estimates (20 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
bmic	0.444112	0.064496	0.317269	0.570955	354.94	0.394290	0.492267	0	6.89	<.0001
bullied	4.565642	0.488556	3.607294	5.523989	1460.2	4.157527	4.783428	0	9.35	<.0001

Multivariate Inference
Assuming Proportionality of Between/Within Covariance Matrices

Avg Relative Increase in Variance	Num DF	Den DF	F for H0: Parameter=Theta0	Pr > F
0.206559	2	1065.1	51.63	<.0001

SPSS Imputation Analysis Script

```
* set working directory.
CD "YOUR-FILE-PATH".

* read stacked imputation data.
DATA LIST free file = "imps_stacked.dat"
 /imputation_ id bullied bmi batt1 batt2 batt3 batt4 batt5.
MISSING VALUES all (999).

* compute scale score and center covariate.
COMPUTE bodyatt = batt1 + batt2 + batt3 + batt4 + batt5.
AGGREGATE
 /break = imputation_
 /bmimean = mean(bmi).
COMPUTE bmic = bmi - bmimean.
```


SPSS Imputation Analysis Script

```
* initiate pooling routines.
SORT CASES by imputation_.
SPLIT FILE layered by imputation_.

* analysis and pooling.
REGRESSION
 /descriptives mean stddev corr sig n
 /dependent bodyatt
 /method = enter bmic bullied.
```


SPSS Analysis Output

Coefficients[a]

imputation_	Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.
.00	1	(Constant)	19.621	.207		94.843	.000
		bmic	.418	.080	.293	5.208	.000
		bullied	4.656	.665	.394	7.003	.000
1.00	1	(Constant)	20.086	.177		113.688	.000
		bmic	.394	.057	.293	6.970	.000
		bullied	4.158	.465	.376	8.935	.000
20.00	1	(Constant)	20.076	.178		113.056	.000
		bmic	.448	.058	.324	7.748	.000
		bullied	4.734	.472	.419	10.025	.000
Pooled	1	(Constant)	20.004	.181		110.307	.000
		bmic	.444	.064		6.886	.000
		bullied	4.566	.489		9.345	.000

a. Dependent Variable: bodyatt

Stata Imputation Analysis Script

```
// set working directory
cd "YOUR-FILE-PATH"

// read stacked data
clear
infile imp id bullied bmi batt1 - batt5 using “imps_stacked.dat”

// recode missing data in original data (imp = 0)
recode bullied bmi batt1 - batt5 (999 = .)

// compute scale score and center covariate.
gen bodyatt = batt1 + batt2 + batt3 + batt4 + batt5
egen impmeans = mean(bmi), by(imp)
gen bmic = bmi - impmeans

// convert to mi data , analysis and pooling
mi import flong, m(imp) id(id) imputed(bullied - bmic) clear
mi estimate, cmdok: regress bodyatt bmic bullied
```


Stata Output

```
Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =        500
                                                Average RVI       =     0.1658
                                                Largest FMI       =     0.2368
                                                Complete DF       =        497
DF adjustment:   Small sample                   DF:     min       =     183.45
                                                        avg       =     314.94
                                                        max       =     424.97
Model F test:       Equal FMI                   F(   2,  335.0)   =      51.56
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
     bodyatt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmic |     .44392   .0645241     6.88   0.000     .3166152    .5712249
     bullied |   4.563943   .4886984     9.34   0.000     3.602653    5.525233
       _cons |   19.98675   .1813459   110.21   0.000      19.6303    20.34319
------------------------------------------------------------------------------
```