Simple Missing Data Treatments

Utrecht University Winter School: Missing Data in R

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Bad Methods

- Deletion-Based Methods
- Deterministic Imputation Methods

OK Methods

Comparisons

Bad Methods (These almost never work)

Listwise Deletion (Complete Case Analysis)

- Use only complete observations for the analysis
  - Very wasteful (can throw out lots of useful data)
  - Loss of statistical power

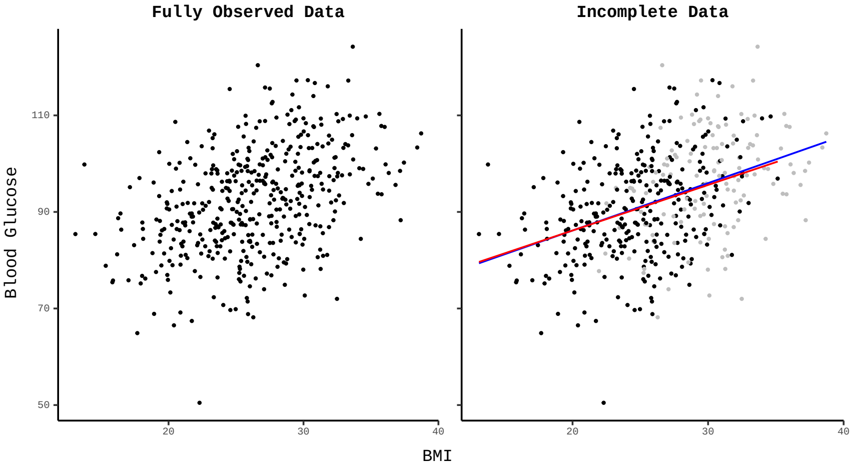
Pairwise Deletion (Available Case Analysis)

- Use only complete pairs of observations for analysis
  - Different samples sizes for different parameter estimates
  - Can cause computational issues

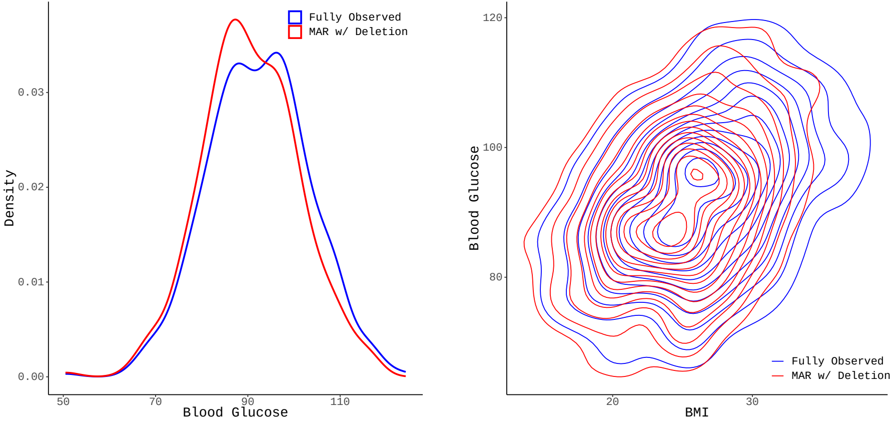
Example

```r
## Read in some example data:
dat0 <- dat1 <- readRDS(paste0(dataDir, "diabetes_norm.rds"))

## Simulated missingness based on 'bmi':
m <- simLinearMissingness(data  = dat1,
                          pm    = 0.30,
                          preds = "bmi",
                          auc   = 0.85)$r

## Impose missing on 'glu' according to the missingness above:
dat1[m, "glu"] <- NA
```

Example


Fully Observed Data
Incomplete Data
Blood Glucose
110
90
70
50
20
30
40
BMI

Example

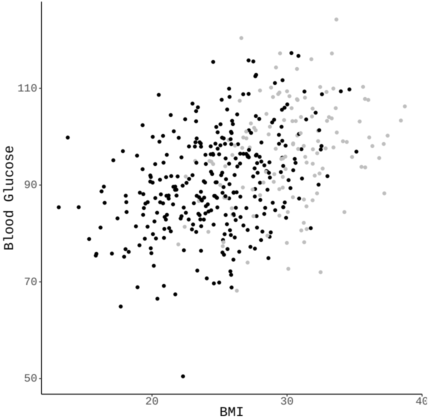


Fully Observed
MAR w/ Deletion
Density
0.00
0.01
0.02
0.03
50
70
90
110
Blood Glucose
Blood Glucose
80
100
120
20
30
BMI
Fully Observed
MAR w/ Deletion

Bad Methods (These almost never work)

(Unconditional) Mean Substitution

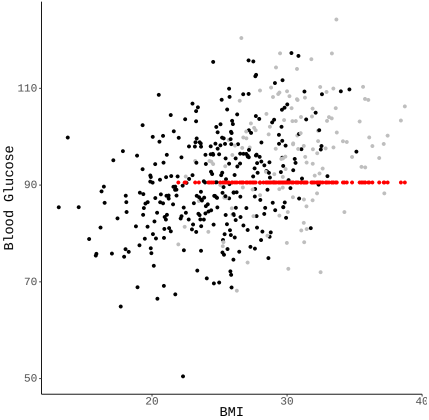
- Replace Y_{mis} with $\bar{Y}_{obs}$
  - Negatively biases regression slopes and correlations
  - Attenuates measures of linear association



Bad Methods (These almost never work)

(Unconditional) Mean Substitution

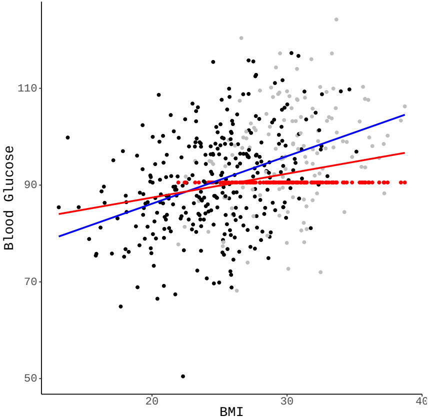
- Replace Y_{mis} with $\bar{Y}_{obs}$
  - Negatively biases regression slopes and correlations
  - Attenuates measures of linear association



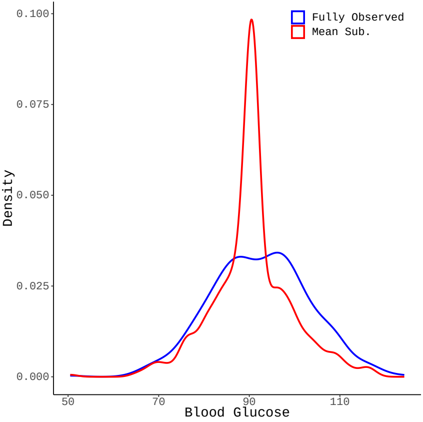
Bad Methods (These almost never work)

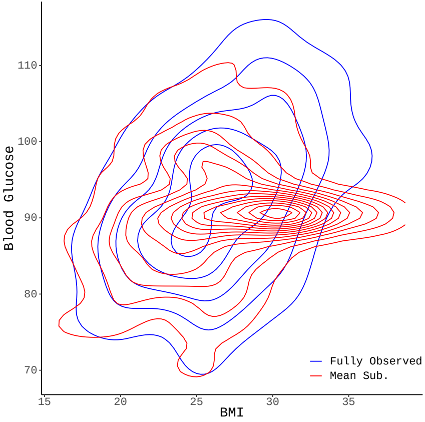
(Unconditional) Mean Substitution

- Replace Y_{mis} with $\bar{Y}_{obs}$
  - Negatively biases regression slopes and correlations
  - Attenuates measures of linear association



Example


Fully Observed
Mean Sub.
0.100
0.075
0.050
0.025
0.000
Density
50
70
90
110
Blood Glucose


110
100
90
80
70
Blood Glucose
15
20
25
30
35
BMI
Fully Observed
Mean Sub.

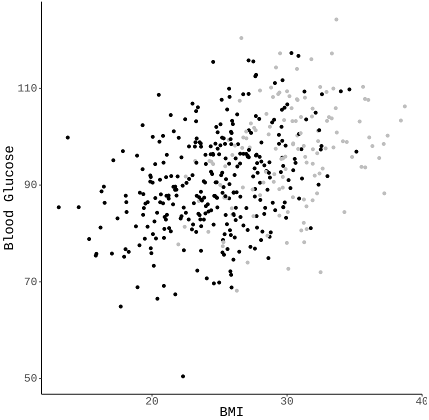
Implementation

```r
dat1[m, "glu"] <- mean(dat1$glu, na.rm = TRUE)

miceOut <- mice(data = dat1, m = 1, maxit = 1, method = "mean")
impData <- complete(miceOut, 1)
```

Bad Methods (These almost never work)

Deterministic Regression Imputation
(Conditional Mean Substitution)

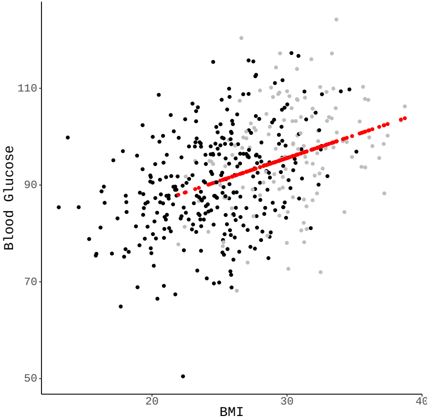
- Replace Y_{mis} with $\widehat{Y}_{mis}$ from some regression equation
  - Positively biases regression slopes and correlations
  - Inflates measures of linear association



Bad Methods (These almost never work)

Deterministic Regression Imputation
(Conditional Mean Substitution)

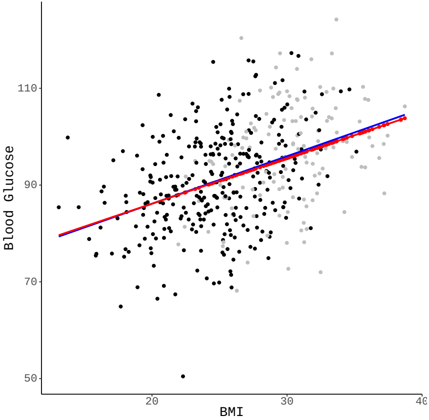
- Replace Y_{mis} with $\widehat{Y}_{mis}$ from some regression equation
  - Positively biases regression slopes and correlations
  - Inflates measures of linear association



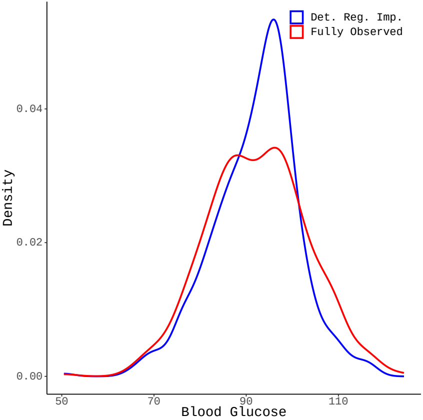
Bad Methods (These almost never work)

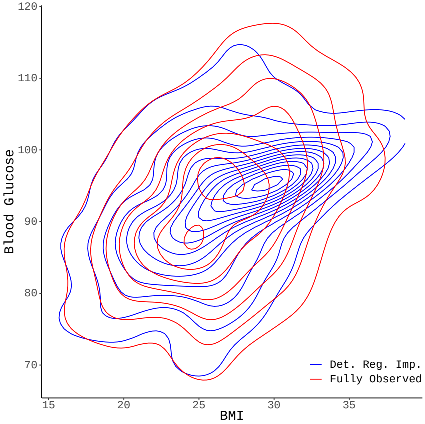
Deterministic Regression Imputation
(Conditional Mean Substitution)

- Replace Y_{mis} with $\widehat{Y}_{mis}$ from some regression equation
  - Positively biases regression slopes and correlations
  - Inflates measures of linear association



Example


Det. Reg. Imp.
Fully Observed
Density
0.04
0.02
0.00
50
70
90
110
Blood Glucose


120
110
100
90
80
70
Blood Glucose
15
20
25
30
35
BMI
Det. Reg. Imp.
Fully Observed

Implementation

```r
miceOut <- mice(data = dat1, m = 1, method = "norm.predict")
impData <- complete(miceOut, 1)
```

Bad Methods (These almost never work)

General Issues with Deletion-Based Methods

- Biased parameter estimates unless data are MCAR
- Generalizability issues

General Issues with Simple Single Imputation Methods

- Biased parameter estimates even when data are MCAR
- Attenuates variability in any treated variables

Bad Methods (These almost never work)

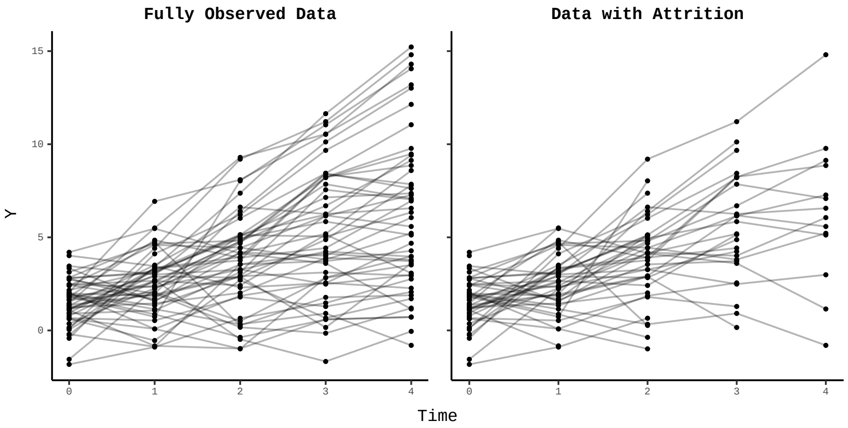
Averaging Available Items (Person-Mean Imputation)

- Compute aggregate scores using only available values
  - Missing data must be MCAR
  - Each item must contributes equally to the aggregate score

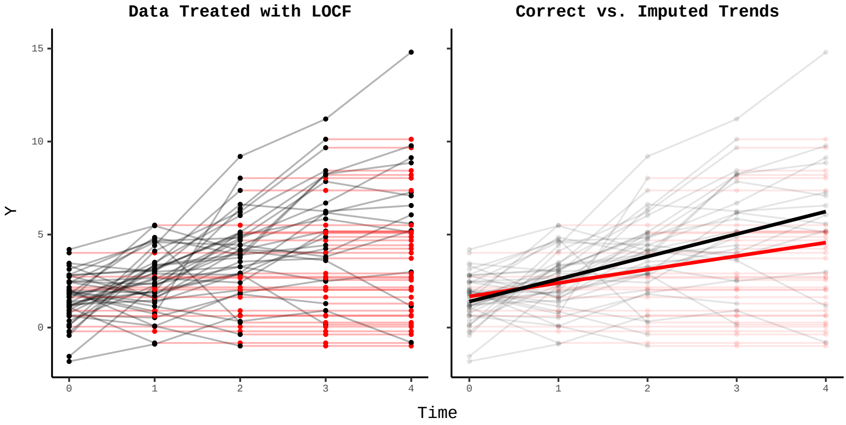
Last Observation Carried Forward (LOCF)

- Replace post-dropout values with the most recent observed value
  - Assume that dropouts would maintain their last known values
  - Attenuates estimates of growth/development

LOCF

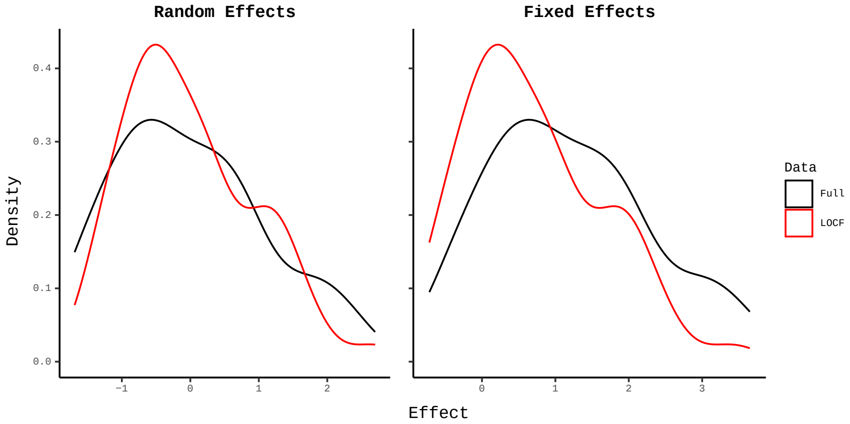

Fully Observed Data
Data with Attrition
Y
15
10
5
0
0
1
2
3
4
Time

LOCF


Data Treated with LOCF
Correct vs. Imputed Trends
Y
15
10
5
0
0
1
2
3
4
Time

Example

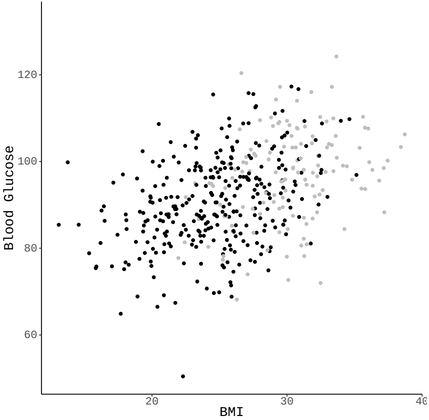
```
## Fit some multilevel regression models
fit1 <- lmer(y ~ t + (t | id), data = fullData)
fit2 <- lmer(y ~ t + (t | id), data = locfData)
```



OK Methods (These sometimes work)

Stochastic Regression Imputation

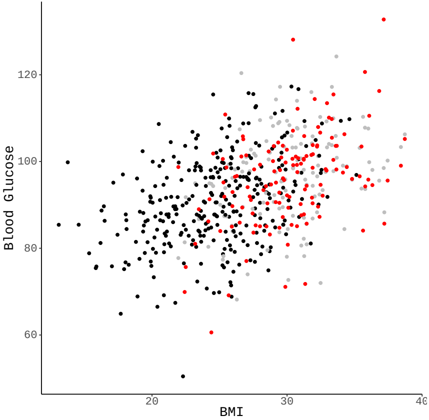
- Fill Y_{mis} with $\widehat{Y}_{mis}$ plus some random noise.
  - Produces unbiased parameter estimates and predictions
  - Computationally efficient
  - Attenuates standard errors
  - Makes CIs and prediction intervals too narrow



OK Methods (These sometimes work)

Stochastic Regression Imputation

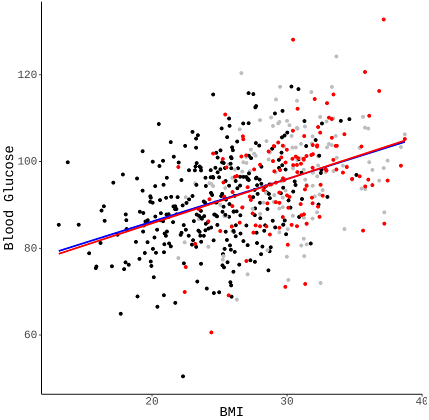
- Fill Y_{mis} with $\widehat{Y}_{mis}$ plus some random noise.
  - Produces unbiased parameter estimates and predictions
  - Computationally efficient
  - Attenuates standard errors
  - Makes CIs and prediction intervals too narrow



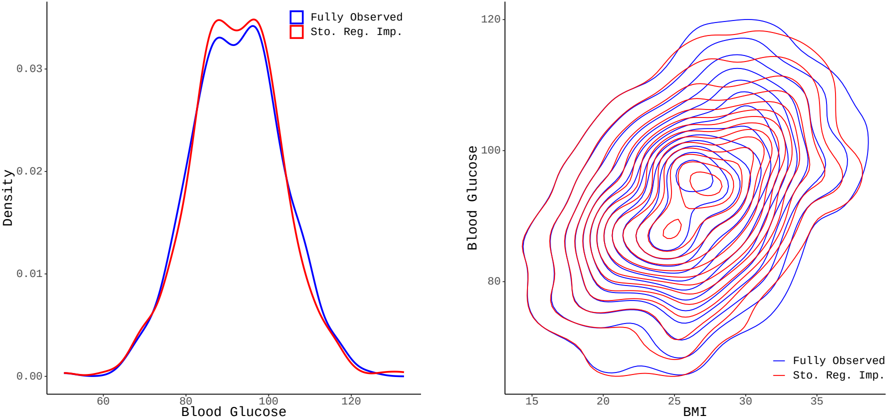
OK Methods (These sometimes work)

Stochastic Regression Imputation

- Fill Y_{mis} with $\widehat{Y}_{mis}$ plus some random noise.
  - Produces unbiased parameter estimates and predictions
  - Computationally efficient
  - Attenuates standard errors
  - Makes CIs and prediction intervals too narrow



Example



Density
0.00
0.01
0.02
0.03
60
80
100
120
Blood Glucose
🟦 Fully Observed
🟥 Sto. Reg. Imp.
Blood Glucose
80
100
120
15
20
25
30
35
BMI
Fully Observed
Sto. Reg. Imp.

Implementation

```r
miceOut <- mice(data = dat1, m = 1, seed = 42, method = "norm.nob")
impData <- complete(1)
```

OK Methods (These sometimes work)

Nonresponse Weighting

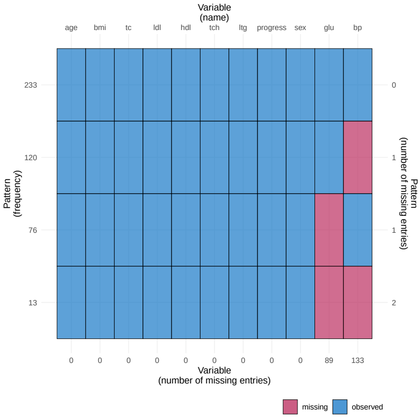
- Weight the observed cases to correct for nonresponse bias
  - Popular in survey research and official statistics
  - Only worth considering with *Unit Nonresponse*
  - Doesn’t make any sense with *Item Nonresponse*

Comparison

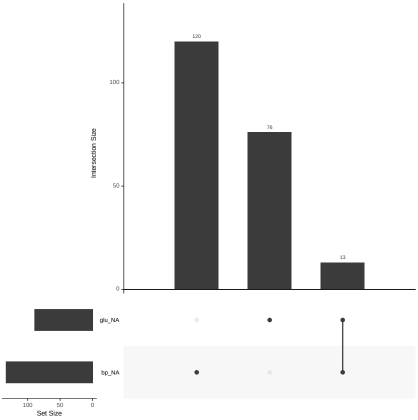
Run a Monte Carlo simulation to compare the treatments.

- Use the synthetic diabetes data as the population.
- Simulate MAR missingness.
  - Blood Glucose
    - $PM = 20\%, P(M) \sim \{bmi, age\}$
  - Blood Pressure
    - $PM = 30\%, P(M) \sim \{bmi, tc\}$
- Treat the missing data as above.
- Use the treated data to estimate several statistics.
- Repeat the process 250 times and pool the results.

Comparison



Variable
(name)
age
bmi
tc
ldl
hdl
tch
ltg
progress
sex
glu
bp
Pattern
(frequency)
233
120
76
13
Pattern
(number of missing entries)
0
1
1
2
0
0
0
0
0
0
0
0
0
89
133
Variable
(number of missing entries)
missing
observed



Intersection Size
100
50
0
120
76
13
glu_NA
bp_NA
100
50
0
Set Size

Comparison

	MS	DRI	SRI	CC	FO
glu	91.30	92.43	92.41	91.30	92.35
bp	97.29	95.50	95.55	97.29	95.50

Variable Means

	MS	DRI	SRI	CC	FO
glu	92.98	105.69	120.70	112.26	117.29
bp	117.22	139.89	175.35	158.40	177.85

Variable Variances

Comparison

$$Y_{BP} = \beta_0 + \beta_1 X_{BMI} + \beta_2 X_{Glucose} + \beta_3 X_{Age} + \varepsilon$$

	MS	DRI	SRI	CC	FO
β_0	65.06	31.85	39.41	40.65	39.74
β_{bmi}	0.39	0.51	0.62	0.61	0.66
β_{glu}	0.16	0.44	0.32	0.32	0.30
β_{age}	0.16	0.19	0.22	0.19	0.22
R^2	0.12	0.37	0.26	0.21	0.25

Linear Regression Estimates

Comparison

	age	bmi	bp	tc	ltg
MS	29.46	15.42	25.29	92.26	1.74
DRI	44.53	21.78	66.47	118.07	2.32
SRI	43.99	21.66	61.16	118.12	2.32
CC	29.76	14.28	50.35	67.44	1.85
FO	40.52	21.19	58.19	107.83	2.28

Covariances with Blood Glucose

Comparison

	age	bmi	tc	ltg	glu
MS	33.11	12.29	34.57	1.51	25.29
DRI	54.20	22.46	92.89	2.54	66.47
SRI	54.19	22.27	90.86	2.54	61.16
CC	36.42	14.87	50.73	2.05	50.35
FO	52.50	22.65	86.00	2.60	58.19

Covariances with Blood Pressure