Simple Missing Data Treatments

Utrecht University Winter School: Missing Data in R

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

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

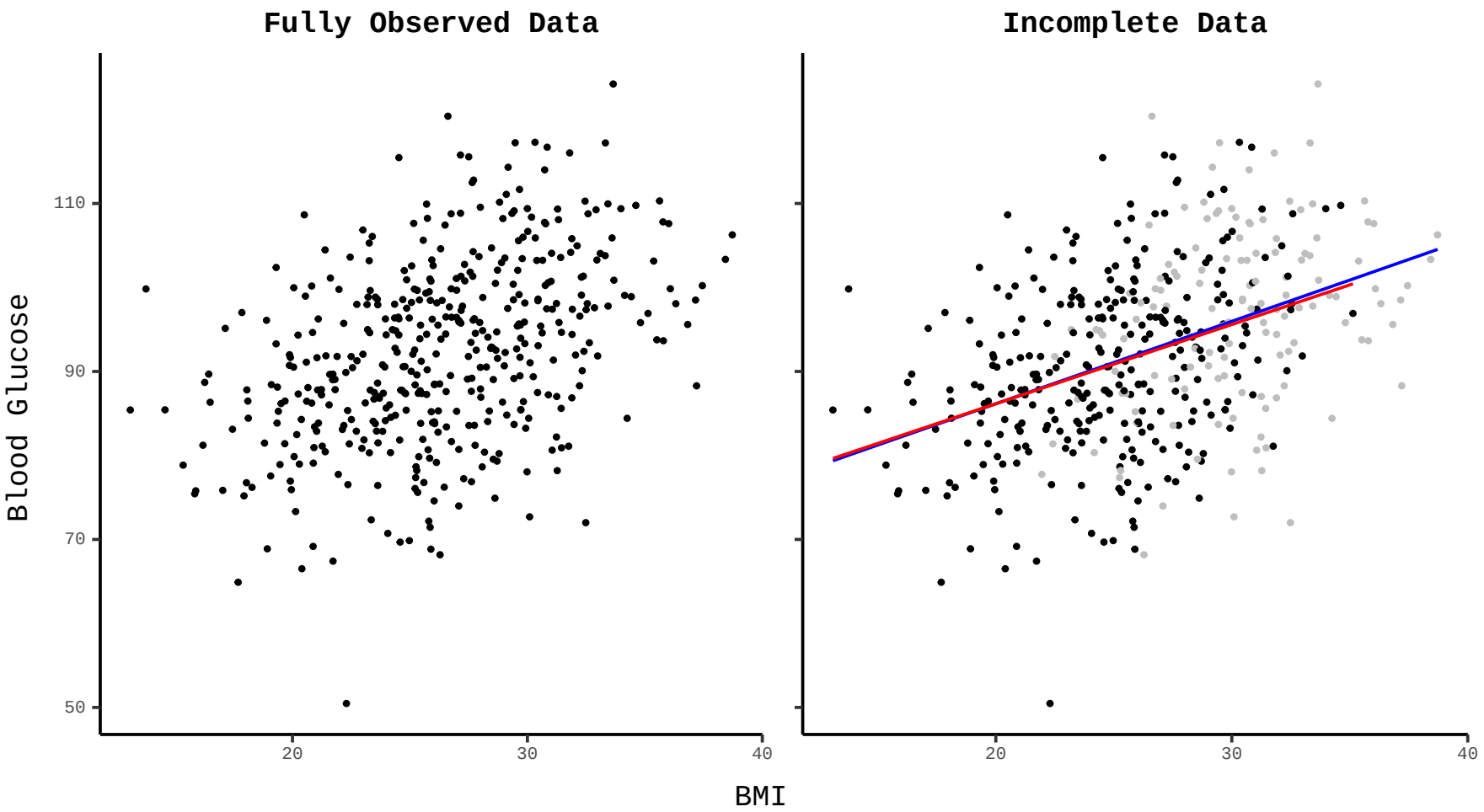
Pairwise Deletion (Available Case Analysis)

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

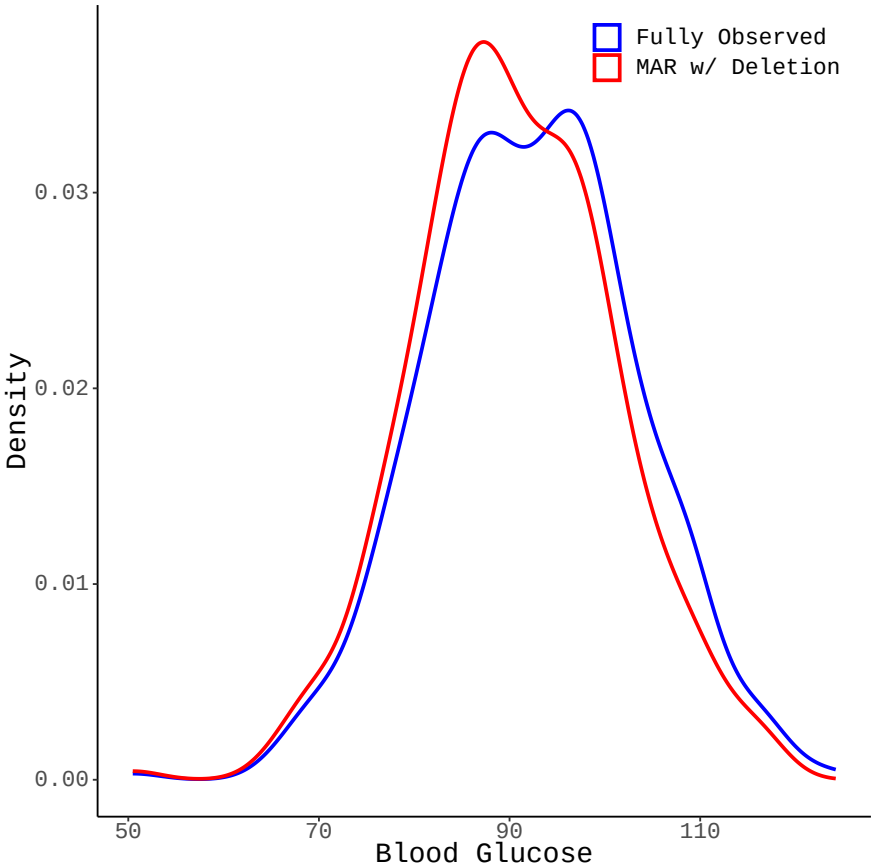
Example

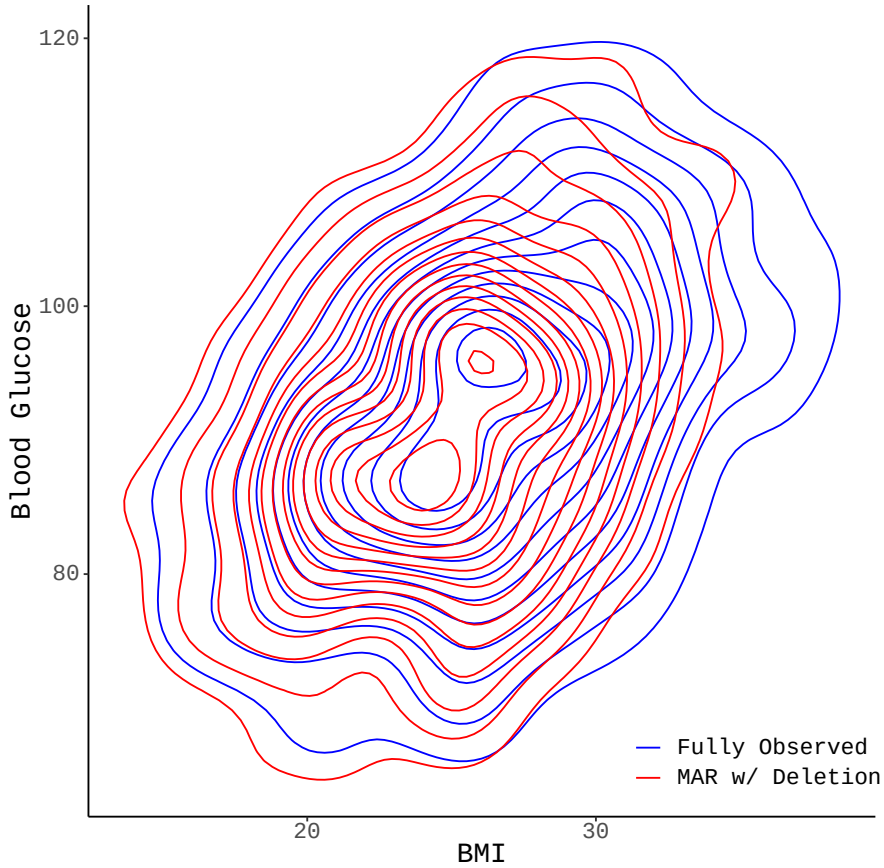
```r
## Read in some example data:
dat0 <- dat1 <- readRDS(paste0(dataDir, "diabetes_norm.rds"))

## 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 missingess above:
dat1[m, "glu"] <- NA
```

Example


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

Example

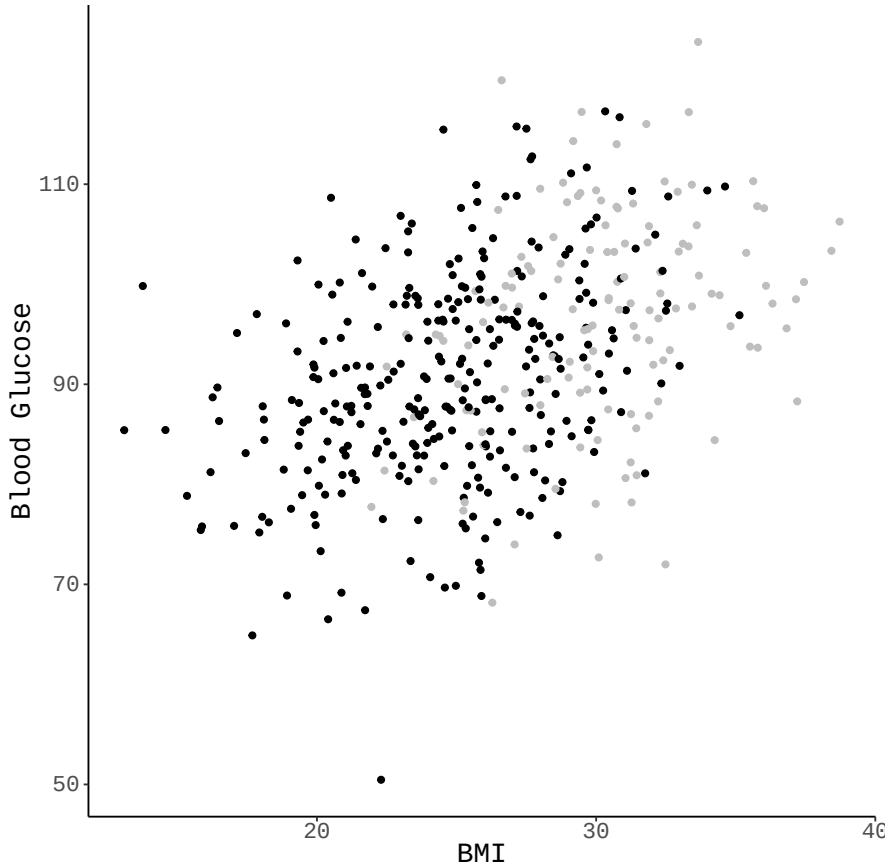

Fully Observed
MAR w/ Deletion
Density
0.03
0.02
0.01
0.00
50
70
90
110
Blood Glucose


120
100
80
Blood Glucose
20
30
BMI
Fully Observed
MAR w/ Deletion

Bad Methods (These almost never work)

(Unconditional) Mean Substitution

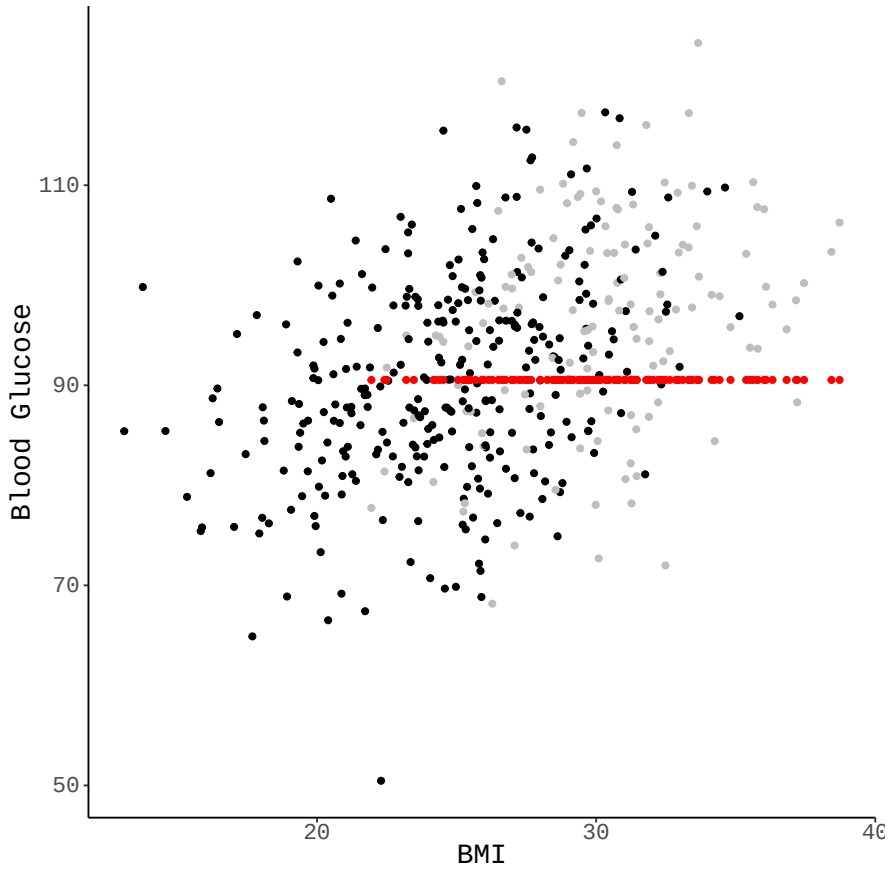
- Replace Y_{mis} with $\bar{Y}_{obs}$
  - Negatively biases regression slopes and correlations
  - Attenuates measures of linear association



Bad Methods (These almost never work)

(Unconditional) Mean Substitution

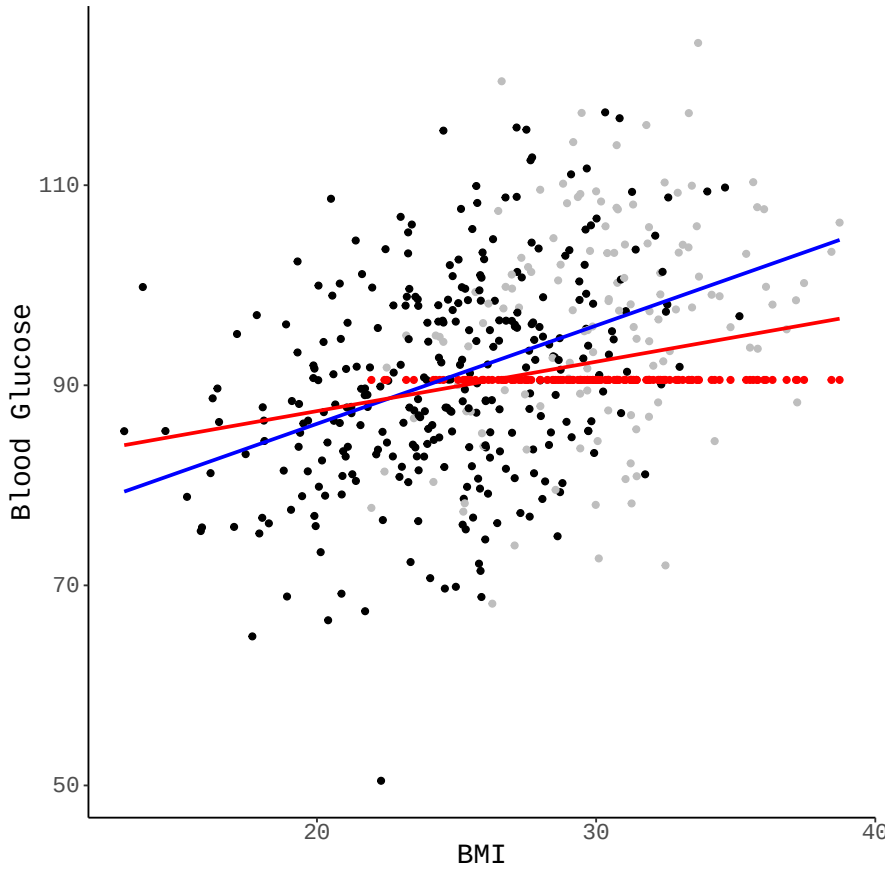
- Replace Y_{mis} with $\bar{Y}_{obs}$
  - Negatively biases regression slopes and correlations
  - Attenuates measures of linear association



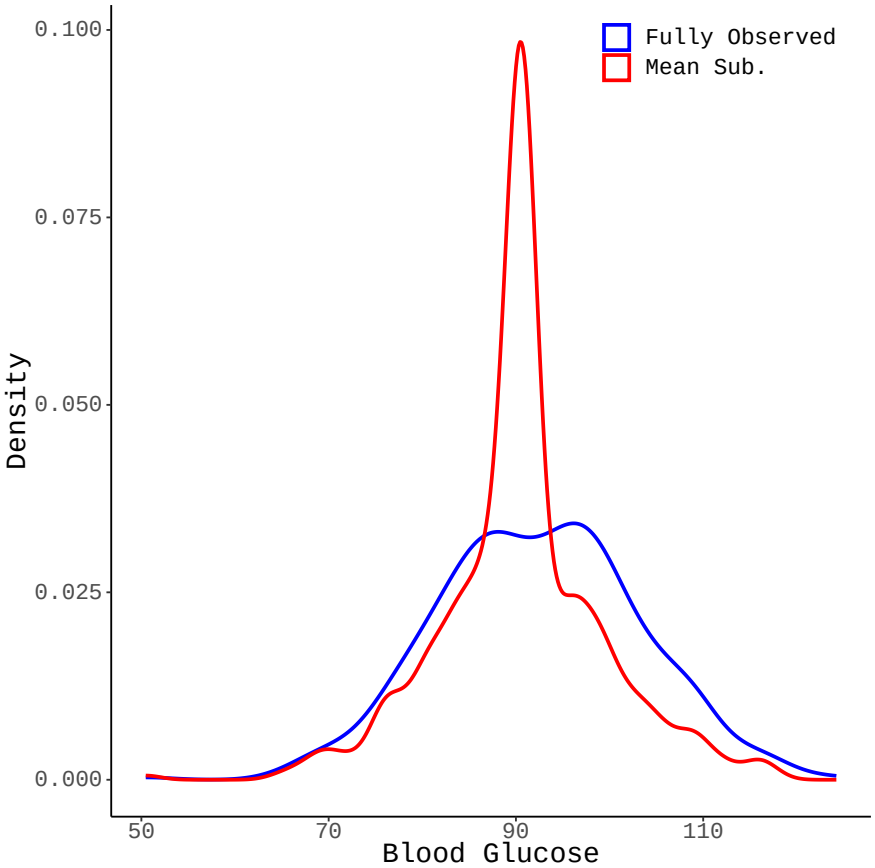
Bad Methods (These almost never work)

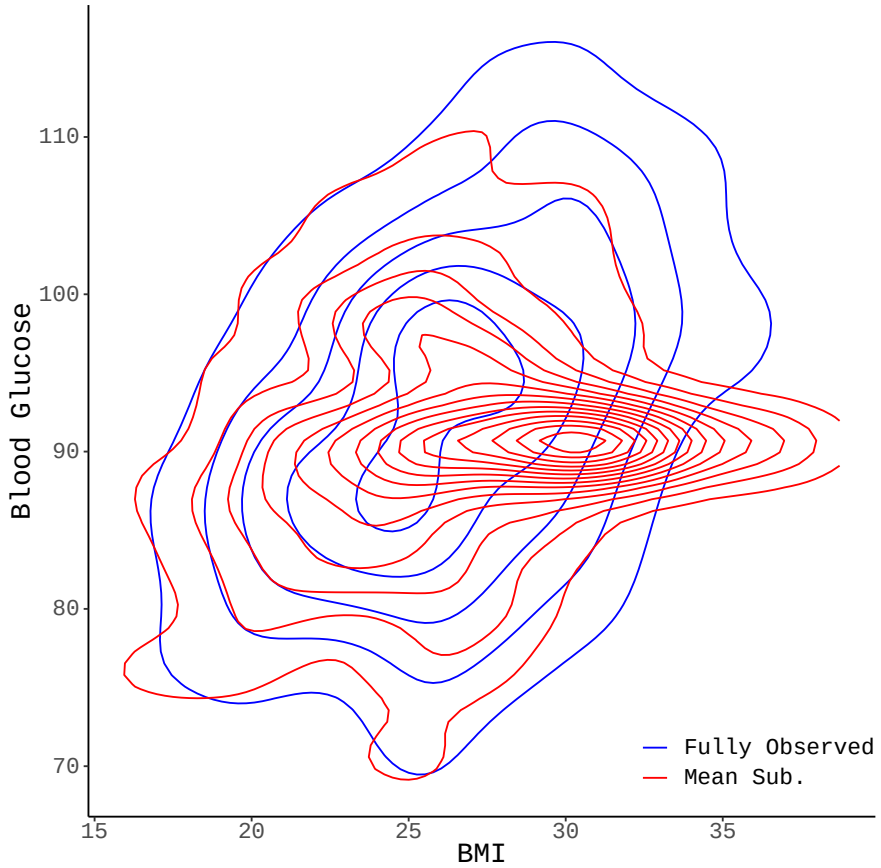
(Unconditional) Mean Substitution

- 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.

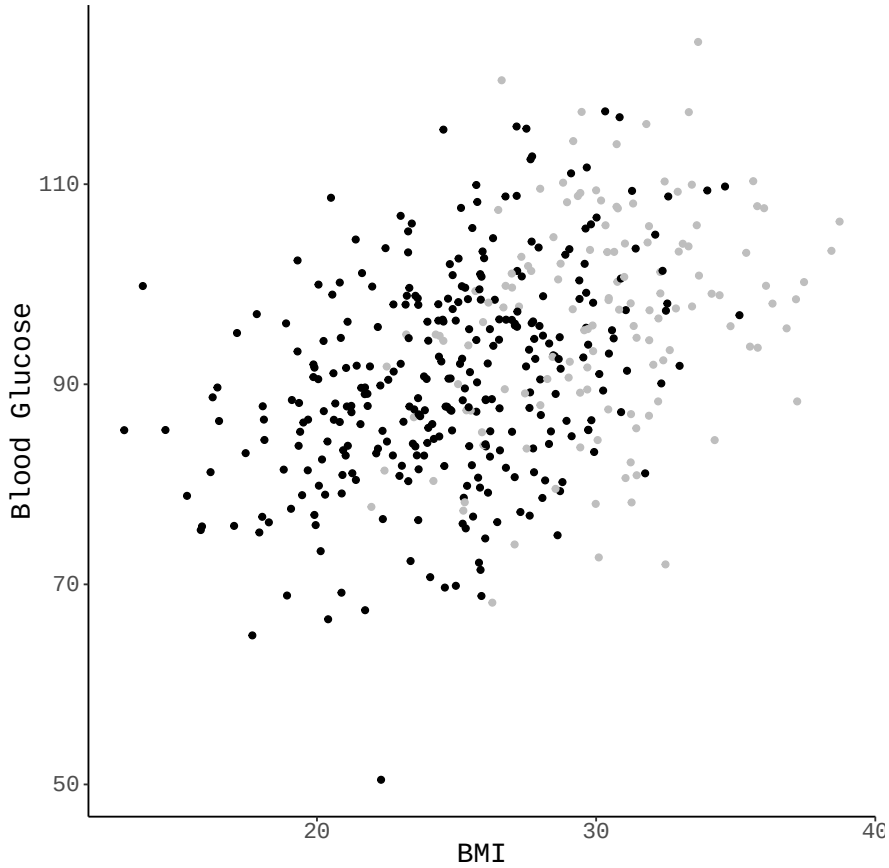
Implementation

```
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)

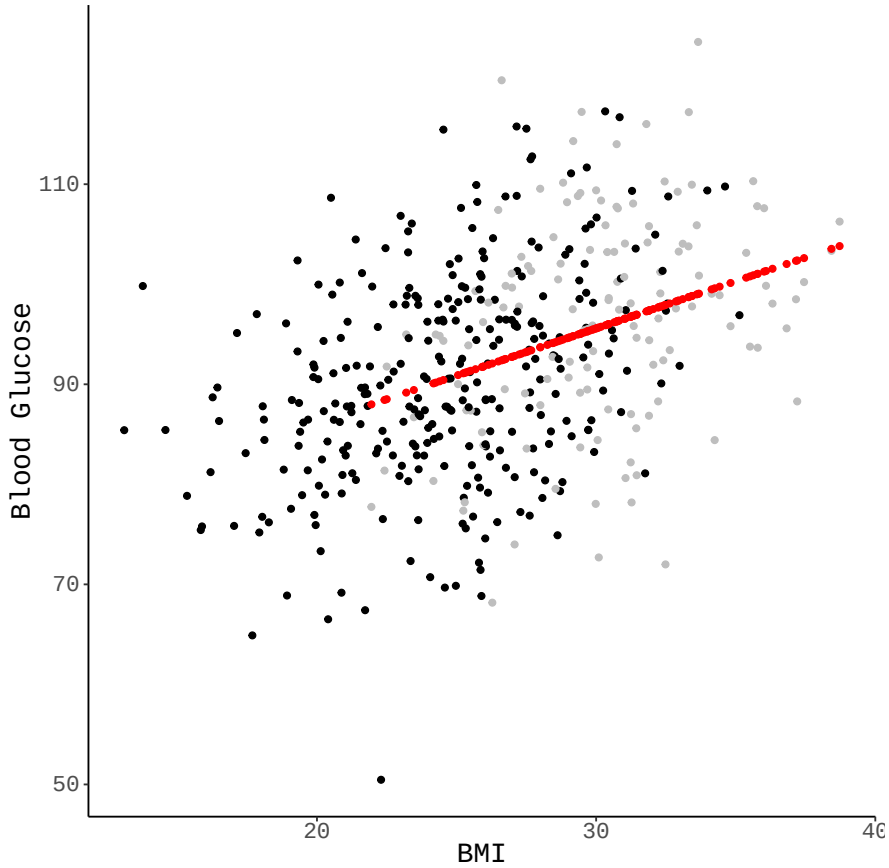
- Replace Y_{mis} with $\widehat{Y}_{mis}$ from some regression equation
  - Positively biases regression slopes and correlations
  - Inflates measures of linear association



Bad Methods (These almost never work)

Deterministic Regression Imputation
(Conditional Mean Substitution)

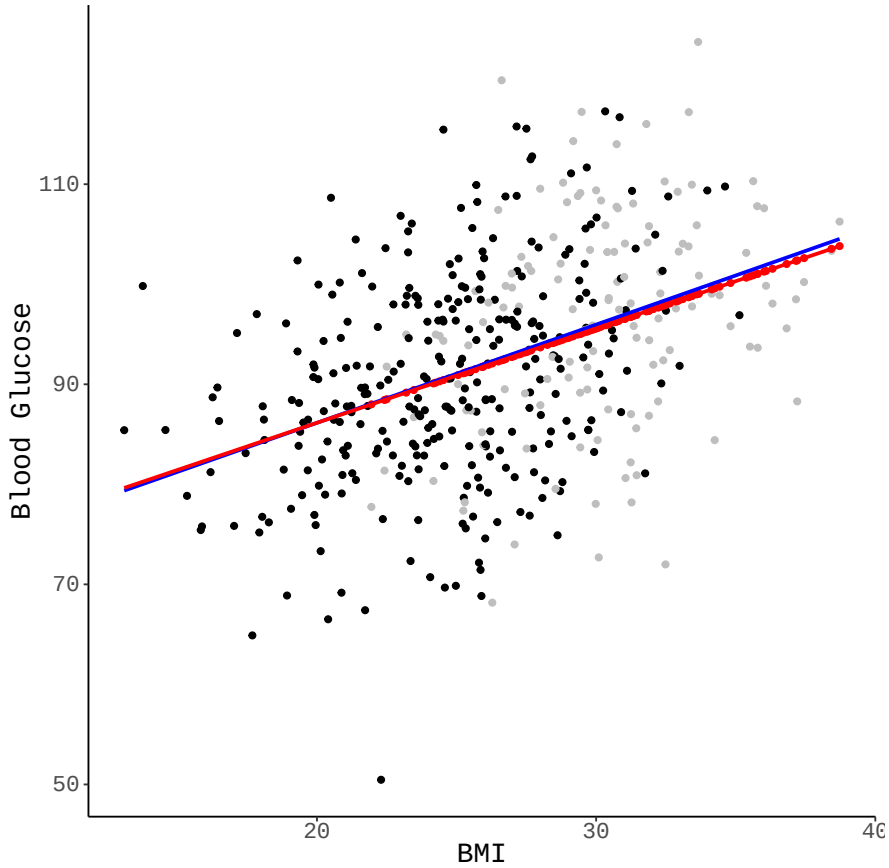
- Replace Y_{mis} with $\widehat{Y}_{mis}$ from some regression equation
  - Positively biases regression slopes and correlations
  - Inflates measures of linear association



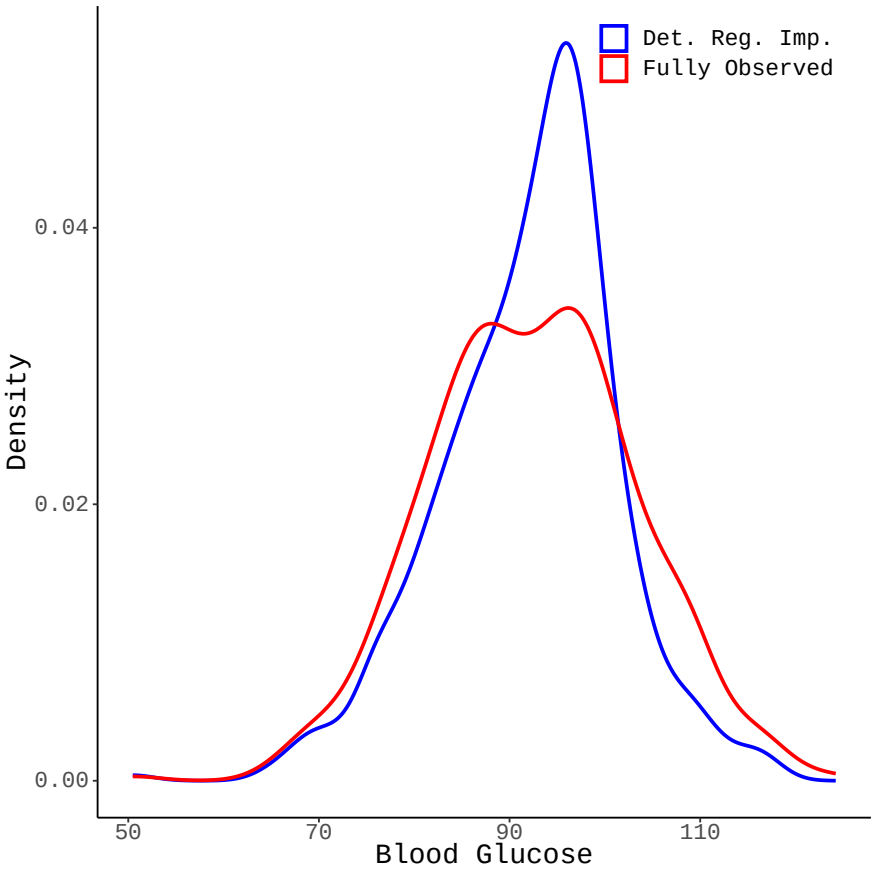
Bad Methods (These almost never work)

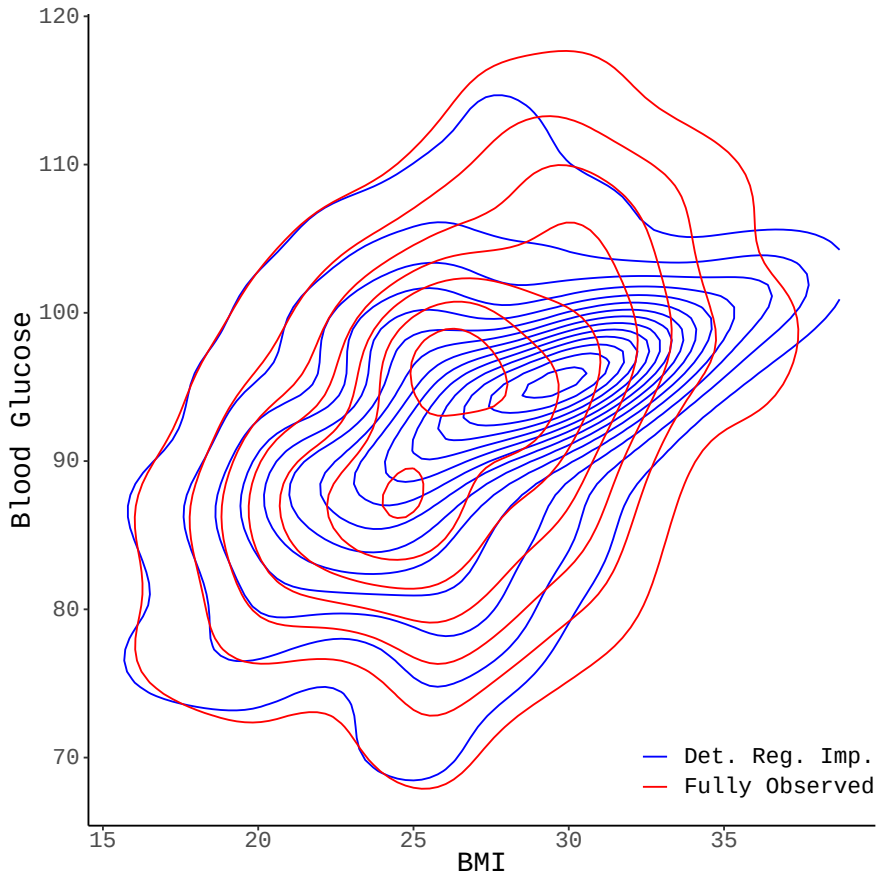
Deterministic Regression Imputation
(Conditional Mean Substitution)

- 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
0.04
0.02
0.00
Density
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

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

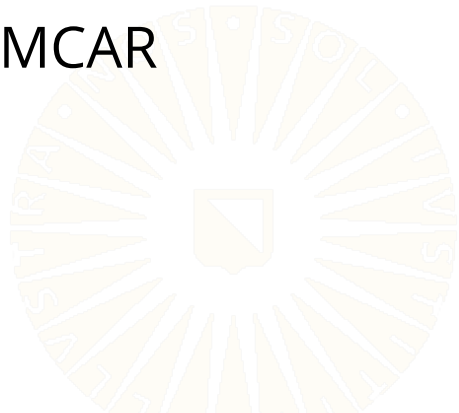
Bad Methods (These almost never work)

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



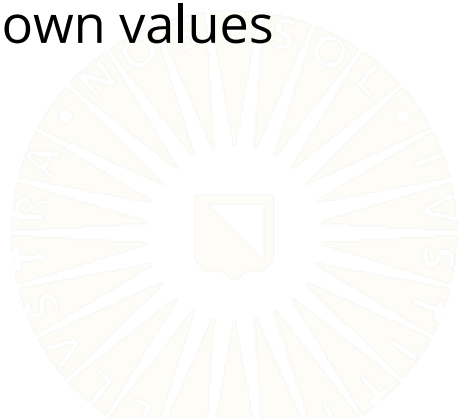
Bad Methods (These almost never work)

Averaging Available Items (Person-Mean Imputation)

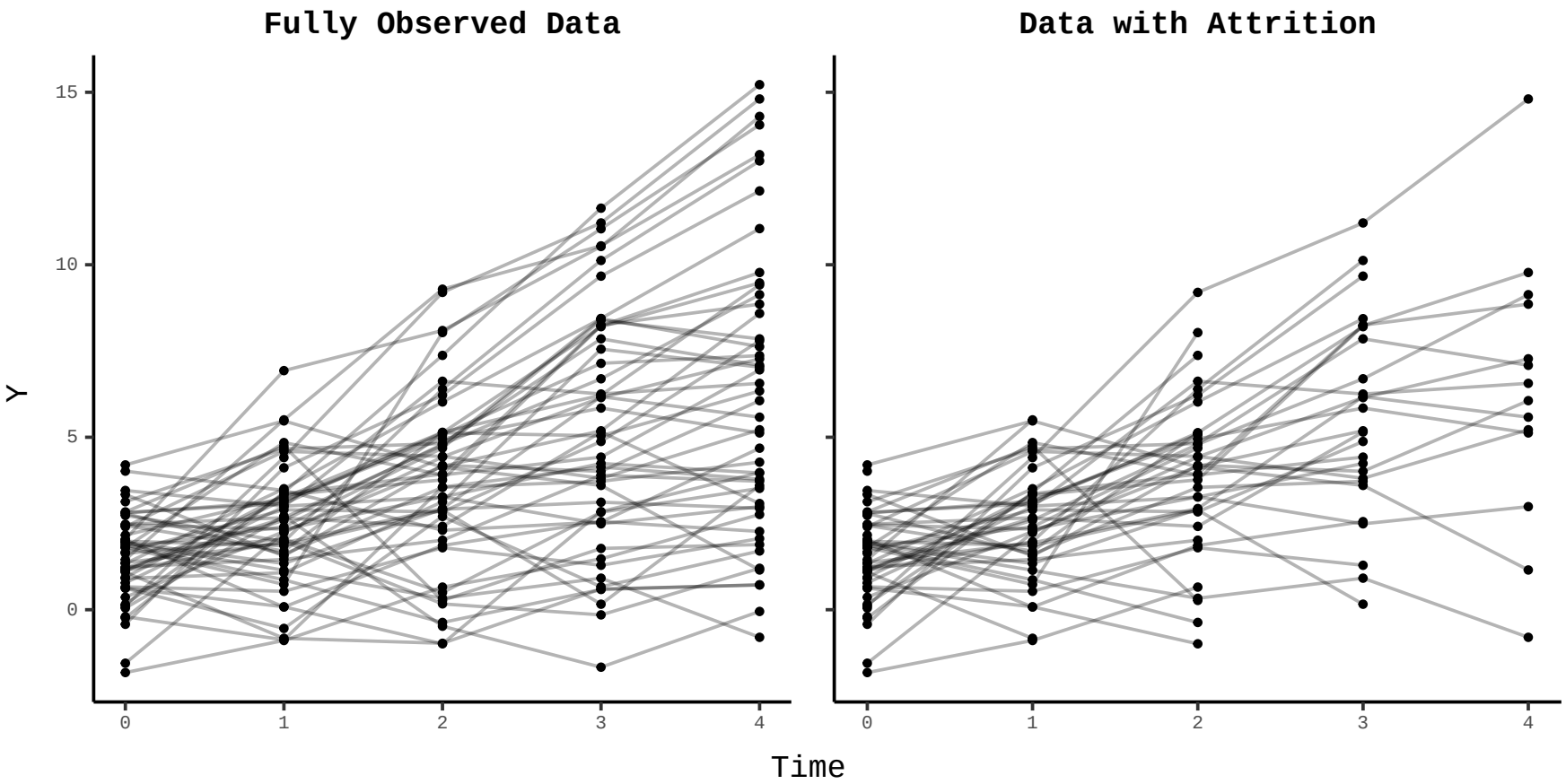
- Compute aggregate scores using only available values
  - Missing data must be MCAR
  - Each item must contributes equally to the aggregate score

Last Observation Carried Forward (LOCF)

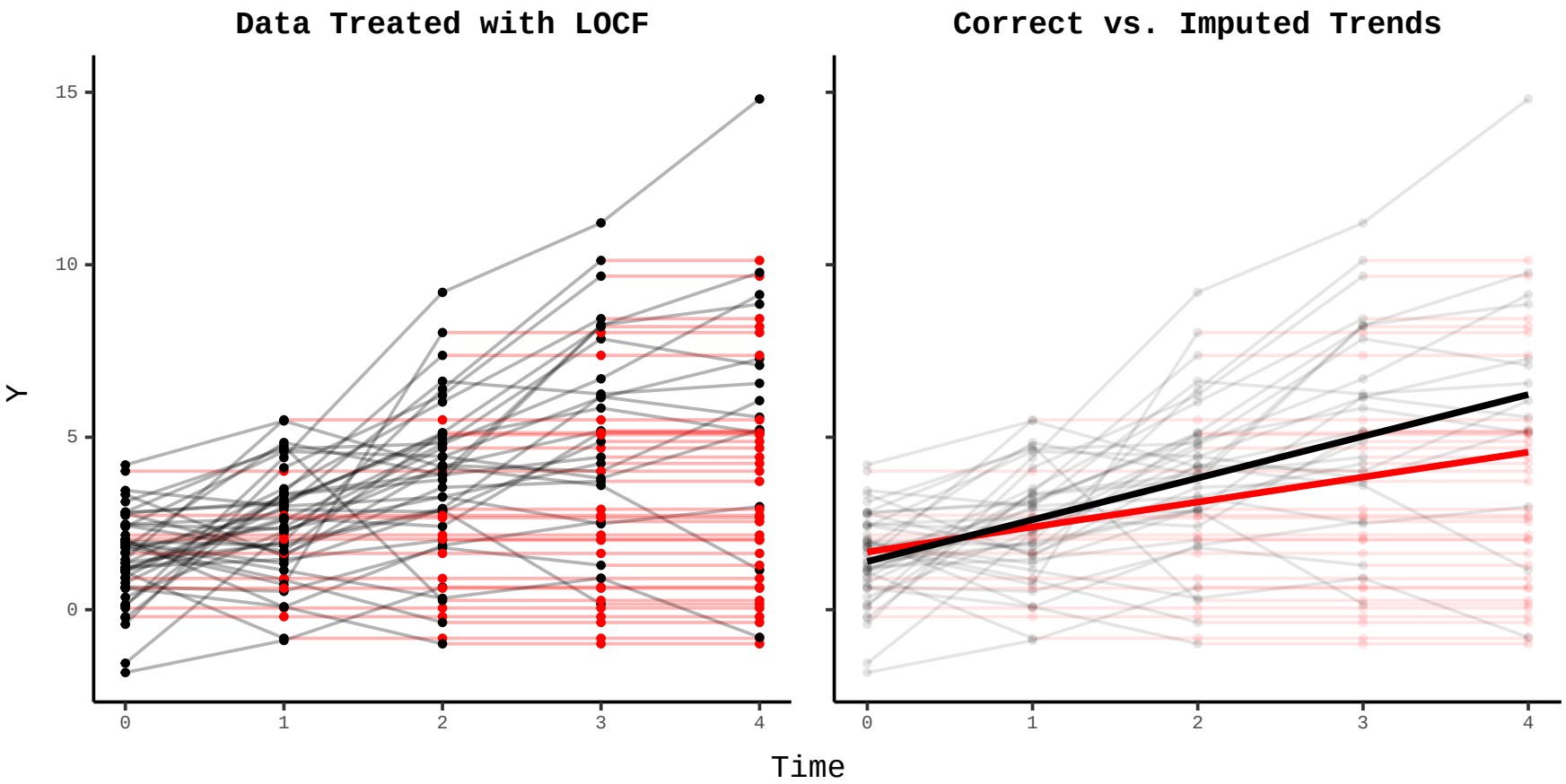
- Replace post-dropout values with the most recent observed value
  - Assume that dropouts would maintain their last known values
  - Attenuates estimates of growth/development



LOCF

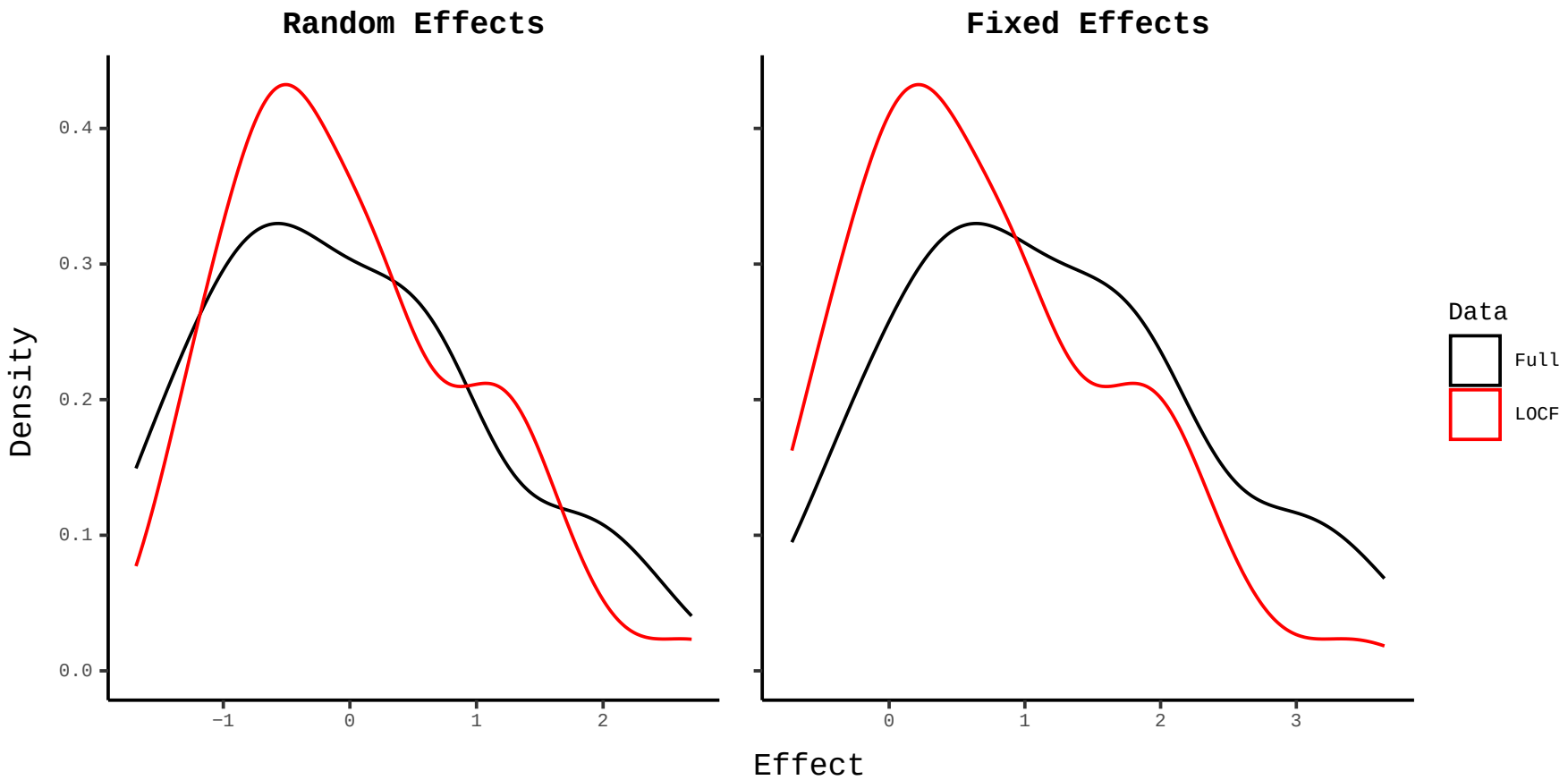

Fully Observed Data
Data with Attrition
Y
15
10
5
0
0
1
2
3
4
Time

LOCF


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

Example

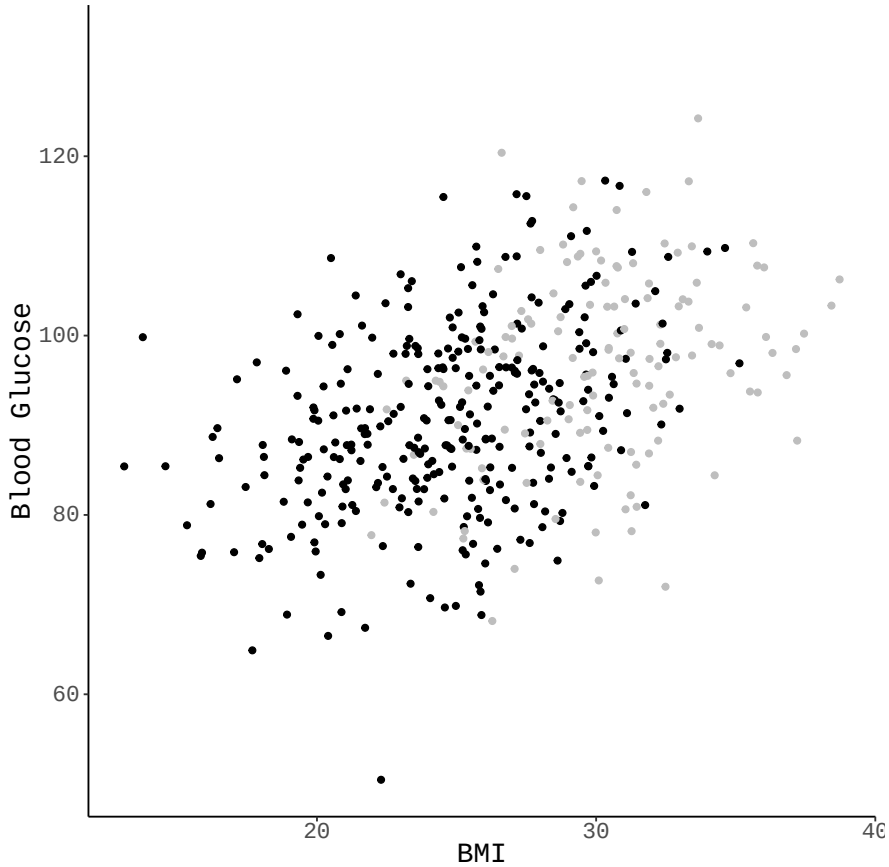
```
## Fit some multilevel regression models
fit1 <- lmer(y ~ t + (t | id), data = fullData)
fit2 <- lmer(y ~ t + (t | id), data = locfData)
```



OK Methods (These sometimes work)

Stochastic Regression Imputation

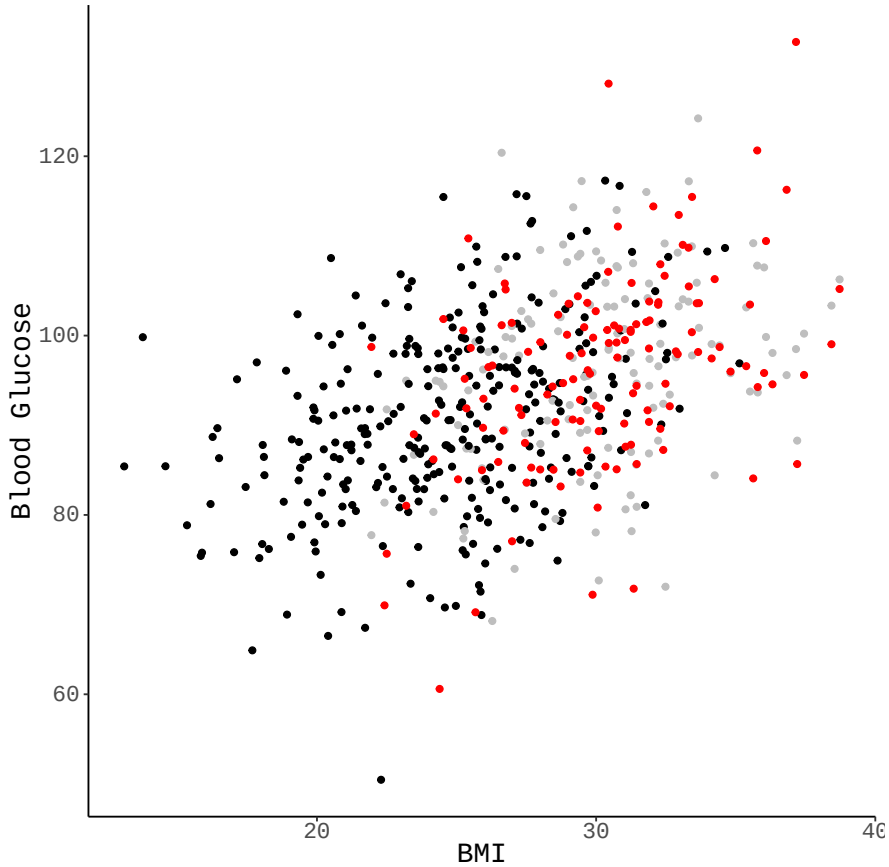
- Fill Y_{mis} with $\widehat{Y}_{mis}$ plus some random noise.
  - Produces unbiased parameter estimates and predictions
  - Computationally efficient
  - Attenuates standard errors
  - Makes CIs and prediction intervals too narrow



OK Methods (These sometimes work)

Stochastic Regression Imputation

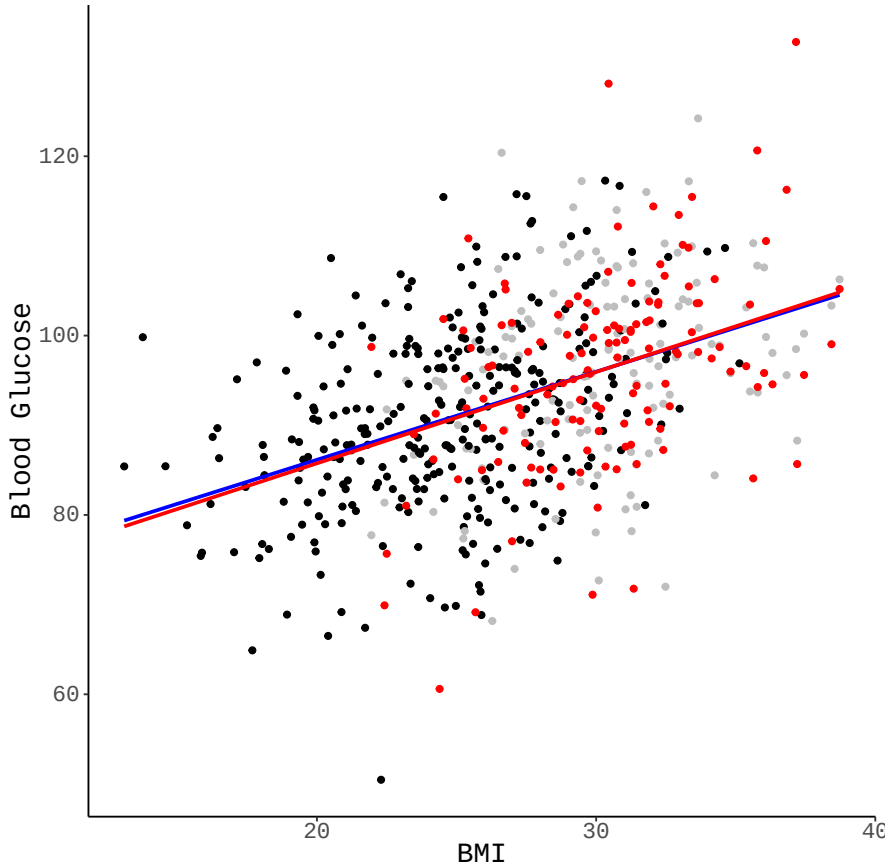
- Fill Y_{mis} with $\widehat{Y}_{mis}$ plus some random noise.
  - Produces unbiased parameter estimates and predictions
  - Computationally efficient
  - Attenuates standard errors
  - Makes CIs and prediction intervals too narrow



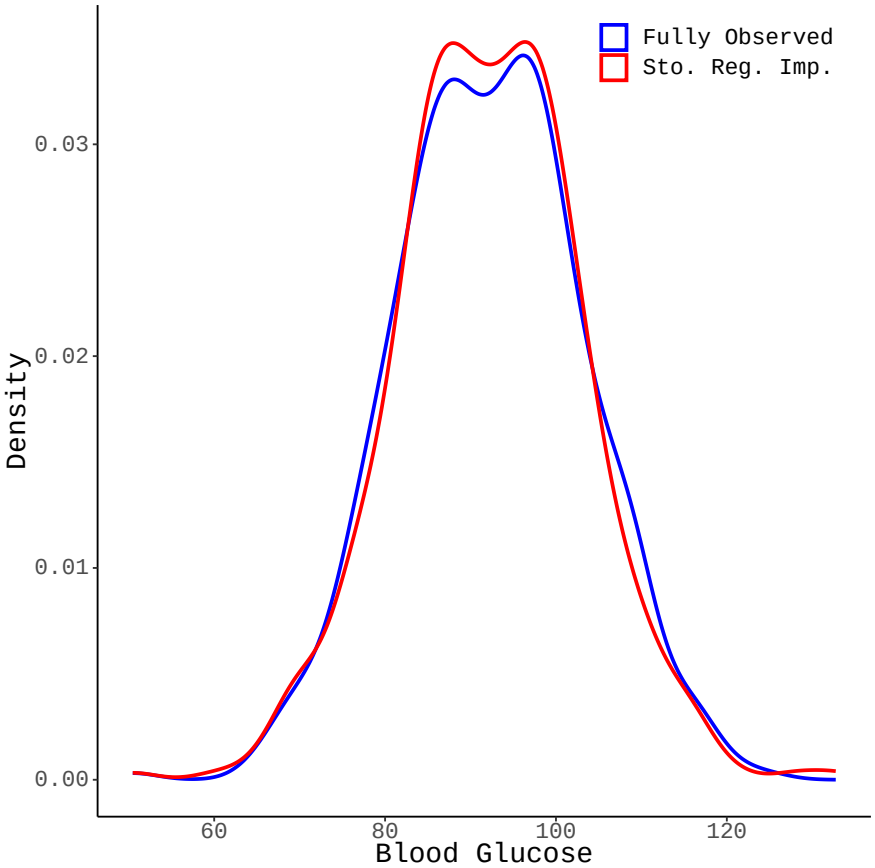
OK Methods (These sometimes work)

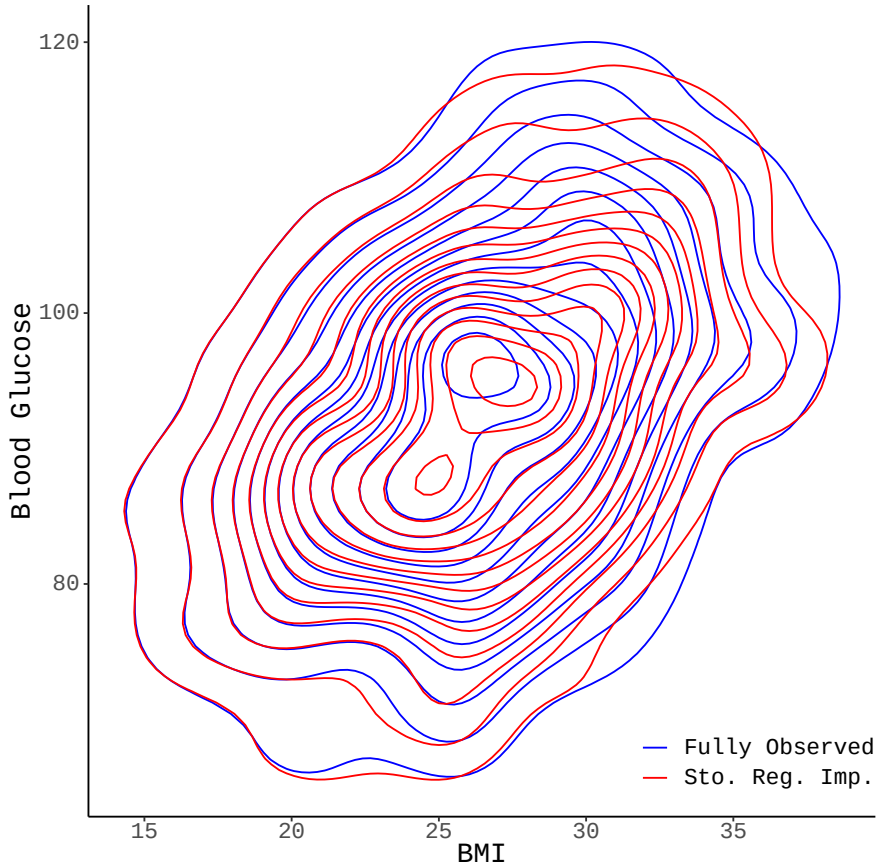
Stochastic Regression Imputation

- 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


Fully Observed
Sto. Reg. Imp.
Density
0.03
0.02
0.01
0.00
60
80
100
120
Blood Glucose


Blood Glucose
120
100
80
15
20
25
30
35
BMI
Fully Observed
Sto. Reg. Imp.

Implementation

```
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*