Missing Data 1

MSBBSS01: Survey data analysis, week 46, BOL - 1.023

Stef van Buuren, Gerko Vink

Nov 13, 2023

Course Overview

Nature and impact of missing data

Ad-hoc techniques

Multiple imputation

Generating imputations, univariate

Stef van Buuren

Gerko Vink

Course Overview

Why deal with missing data?

- Missing data are everywhere
- Missing data are the **heart of statistics**
- Ad-hoc fixes do not (always) work
- **Multiple imputation** is broadly applicable, yields correct statistical inferences
- Goal: get you comfortable with use of `mice` for imputing survey data

Course materials

- Osiris
- Content
- Exercises and practicals at <www.gerkovink.com/sda>

Reading materials

- Van Buuren, S. and Groothuis-Oudshoorn, C.G.M. (2011). `mice`: Multivariate Imputation by Chained Equations in `R`. Journal of Statistical Software, 45(3), 1–67. https://www.jstatsoft.org/article/view/v045i03
- Van Buuren, S. (2018). Flexible Imputation of Missing Data. Second Edition. Chapman & Hall/CRC, Boca Raton, FL. https://stefvanbuuren.name/fimd

Chapman & Hall/CRC
Interdisciplinary Statistics Series

Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

`mice` software

1. CRAN: `mice 3.16.0`

- `install.packages("mice")`

2. Github: `mice 3.16.8`

- `devtools::install_github("amices/mice")`

Schedule

Slot	Time	What	Topic
A	16.30-17.30	L	Missing data, ad-hoc methods
	17.30-17.45		COFFEE/TEA
B	17.45-18.15	L	Multiple imputation, univariate
C	18.15-19.00	P	Three vignettes

Nature and impact of missing data

Definition of missing values

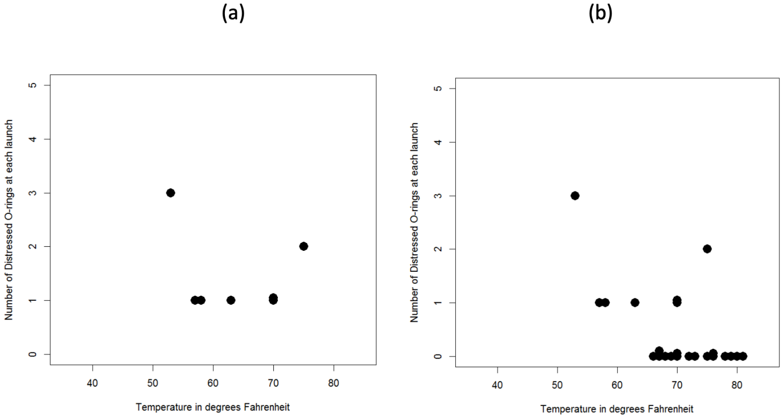
- Missing values are those values that are not observed
- Values do exist in theory, but we are unable to see them

Challenger space shuttle - 28 Jan 1986 - 7 deaths

Challenger space shuttle - 28 Jan 1986 - 7 deaths

- What made the Challenger crash?

Figure 1.1 (a) Data examined in the pre-launch teleconference; **(b)** Complete data.



DARK

WHY
WHAT YOU DON'T KNOW
MATTERS

DATA

DAVID J. HAND

What is dark data?

Dark data are concealed from us, and that very fact means we are at risk of misunderstanding, of drawing incorrect conclusions, and of making poor decisions.

Dark data types (1/2)

- DD-Type 1: Data We Know Are Missing
- DD-Type 2: Data We Don't Know are Missing
- DD-Type 3: Choosing Just Some Cases
- DD-Type 4: Self-Selection
- DD-Type 5: Missing What Matters
- DD-Type 6: Data Which Might Have Been
- DD-Type 7: Changes with Time
- DD-Type 8: Definitions of Data
- DD-Type 9: Summaries of Data
- DD-Type 10: Measurement Error and Uncertainty

Dark data types (2/2)

- DD-Type 11: Feedback and Gaming
- DD-Type 12: Information Asymmetry
- DD-Type 13: Intentionally Darkened Data
- DD-Type 14: Fabricated and Synthetic Data
- DD-Type 15: Extrapolating beyond Your Data

Definition of missing values

- Missing values are those values that are not observed
- Values do exist in theory, but we are unable to see them
- One possible reasons is non-response

Types of non-response

Two types of non-response

- unit non-response: no observed response at all for a case
- item non-response: some, but not all, responses are missing for a case

You can classify missing values in three groups:

- Missing values that should have been observed (unintentional)
- Missing values that should not have been observed (intentional)
- Missing values whose true value can be deduced from the observed data (deductive missings)

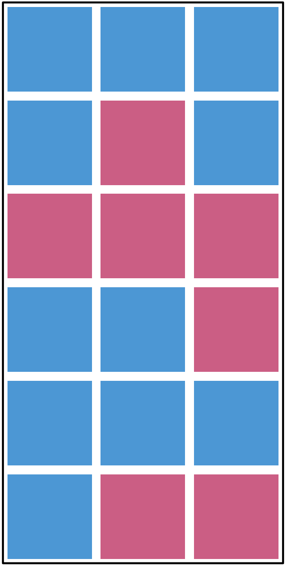
Intentionality vs Response

	Intentional	Unintentional
Unit nonresponse	Sampling	Refusal Self-selection
Item nonresponse	Branching Matrix Sampling	Skip question Coding error

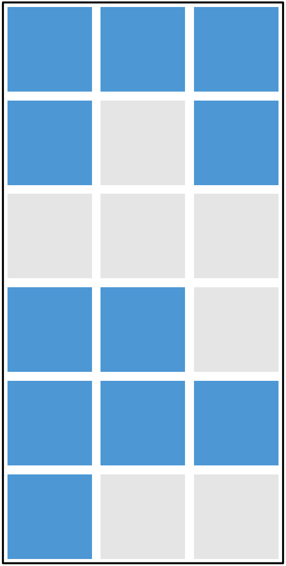
Some confusing terminology

- Complete data = Observed data + Unobserved data
- Incomplete data = Observed data
- Missing data = Unobserved data
- Complete cases = subset of rows in the observed data without missing values
- Complete variables = subset of columns in the observed data without missing values

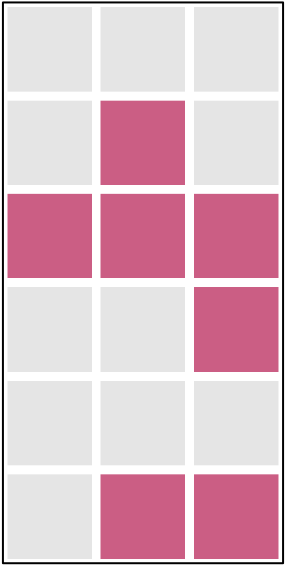
Complete data



Incomplete data = observed data



Missing data = unobserved data



Why values can be missing

Missingness can occur for a lot of reasons. For example

- death, dropout, refusal
- routing, experimental design
- join, merge, bind
- too far away, too small to observe
- power failure, budget exhausted, bad luck

Consequences of missing data

- Cannot calculate, not even the mean
- Less information than planned
- Enough statistical power?
- Different analyses, different n's
- Systematic biases in the analysis
- Appropriate confidence interval, P-values?

Missing data can severely complicate interpretation and analysis

Strategies to deal with missing data

- Prevention
- Ad-hoc methods, e.g., single imputation, complete cases
- Weighting methods
- Likelihood methods, EM-algorithm
- Multiple imputation

Ad-hoc techniques

Listwise deletion, complete-case analysis

- Analyze only the complete records
- Advantages
  - Simple (default in most software)
  - Unbiased under MCAR
  - Conservative standard errors, significance levels
  - Two special properties in regression

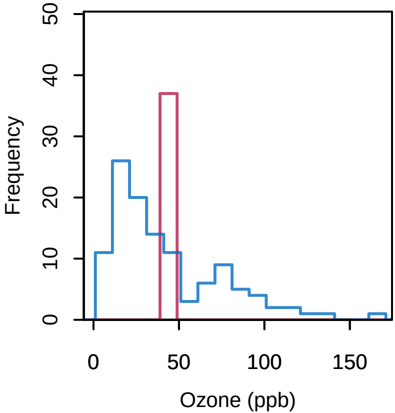
Listwise deletion, complete-case analysis

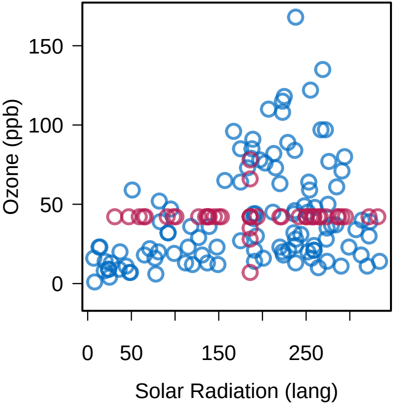
- Disadvantages
  - Wasteful
  - May not be possible
  - Larger standard errors
  - Biased under MAR, even for simple statistics like the mean
  - Inconsistencies in reporting

Mean imputation

- Replace the missing values by the mean of the observed data
- Advantages
  - Simple
  - Unbiased for the mean, under MCAR

Mean imputation





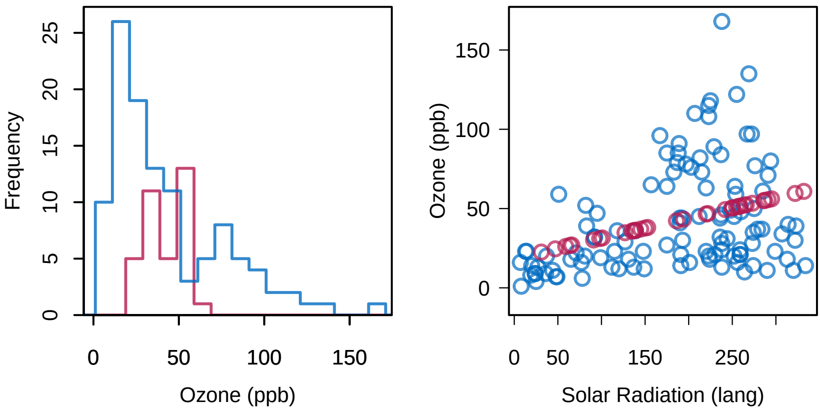
Mean imputation

- Disadvantages
  - Disturbs the distribution
  - Underestimates the variance
  - Biases correlations to zero
  - Biased under MAR
- AVOID (unless you know what you are doing)

Regression imputation

- Also known as **prediction**
  - Fit model for Y^{obs} under listwise deletion
  - Predict Y^{mis} for records with missing Y's
  - Replace missing values by prediction
- Advantages
  - Under MAR, unbiased estimates of regression coefficients
  - Good approximation to the (unknown) true data if explained variance is high
- Favourite among data scientists and machine learners

Regression imputation



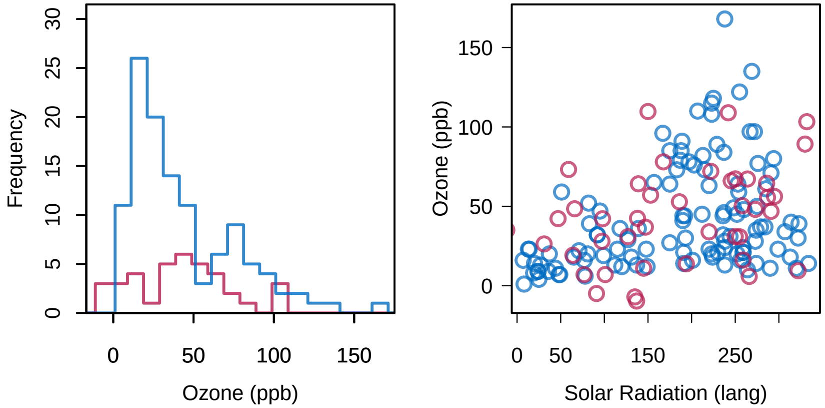
Regression imputation

- Disadvantages
  - Artificially increases correlations
  - Systematically underestimates the variance
  - Too optimistic P-values and too short confidence intervals
- AVOID. Harmful to statistical inference

Stochastic regression imputation

- Like regression imputation, but adds appropriate noise to the predictions to reflect uncertainty
- Advantages
  - Preserves the distribution of Y^{obs}
  - Preserves the correlation between Y and X in the imputed data

Stochastic regression imputation



Stochastic regression imputation

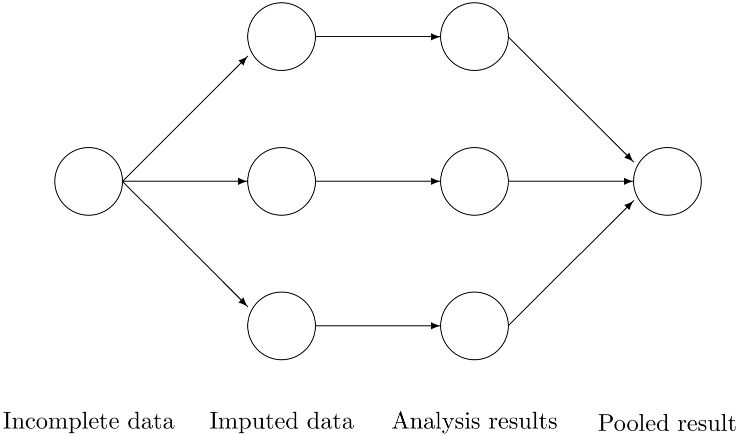
- Disadvantages
  - Symmetric and constant error restrictive
  - Single imputation: does not take uncertainty imputed data into account, and incorrectly treats them as real
  - Not so simple anymore

Overview of assumptions needed

	Mean	Unbiased Reg Weight	Correlation	Standard Error
Listwise	MCAR	MCAR	MCAR	Too large
Pairwise	MCAR	MCAR	MCAR	Complicated
Mean	MCAR	–	–	Too small
Regression	MAR	MAR	–	Too small
Stochastic	MAR	MAR	MAR	Too small
LOCF	–	–	–	Too small
Indicator	–	–	–	Too small

Multiple imputation

Multiple imputation



Acceptance of multiple imputation

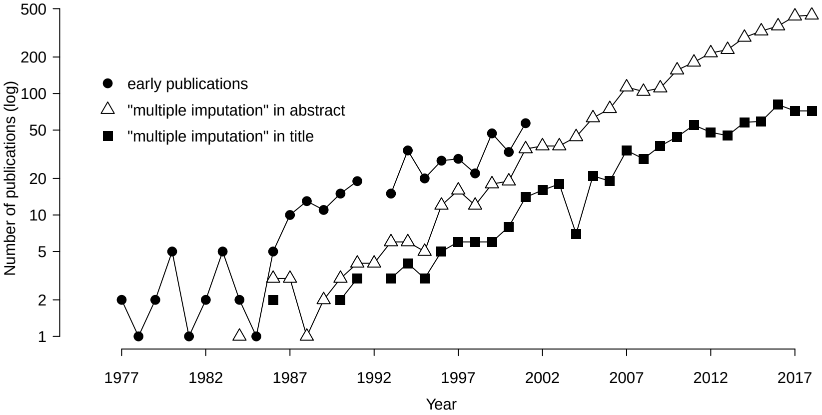


Figure 1: Source: Scopus (April 3, 2019)

Estimand

- Q is a quantity of scientific interest in the population.
- Q can be a vector of population means, population regression weights, population variances, and so on.
- Q may not depend on the particular sample, thus Q cannot be a standard error, sample mean, p-value, and so on.

Goal of multiple imputation

- Estimate Q by $\hat{Q}$ or $\bar{Q}$ accompanied by a valid estimate of its uncertainty.
- What is the difference between $\hat{Q}$ or $\bar{Q}$?
  - $\hat{Q}$ and $\bar{Q}$ both estimate Q
  - $\hat{Q}$ accounts for the sampling uncertainty
  - $\bar{Q}$ accounts for the sampling **and** missing data uncertainty

Pooled estimate $\bar{Q}$

$\hat{Q}_\ell$ is the estimate of the ℓ-th repeated imputation

$\hat{Q}_\ell$ contains k parameters, represented as a $k \times 1$ column vector

Pooled estimate $\bar{Q}$ is simply the average

$$\bar{Q} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{Q}_\ell$$

Within-imputation variance

Average of the complete-data variances as

$$\bar{U} = \frac{1}{m}\sum_{\ell=1}^{m} \bar{U}_\ell,$$

where $\bar{U}_\ell$ is the variance-covariance matrix of $\hat{Q}_\ell$ obtained for the ℓ-th imputation

$\bar{U}_\ell$ is the variance is the estimate, *not* the variance in the data

Within-imputation variance is large if the sample is small

Between-imputation variance

Variance between the m complete-data estimates is given by

$$B = \frac{1}{m-1}\sum_{\ell=1}^{m}(\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})',$$

where $\bar{Q}$ is the pooled estimate.

The between-imputation variance is large there many missing data

Total variance

The total variance is *not* simply $T = \bar{U} + B$

The correct formula is

$$\begin{aligned} T &= \bar{U} + B + B/m \\ &= \bar{U} + \left(1 + \frac{1}{m}\right) B \qquad (1) \end{aligned}$$

for the total variance of $\bar{Q}_m$, and hence of $(Q - \bar{Q})$ if $\bar{Q}$ is unbiased

The term B/m is the simulation error

Three sources of variation

In summary, the total variance T stems from three sources:

1. $\bar{U}$, the variance caused by the fact that we are taking a sample rather than the entire population. This is the conventional statistical measure of variability;
2. B, the extra variance caused by the fact that there are missing values in the sample;
3. B/m, the extra simulation variance caused by the fact that $\bar{Q}_m$ itself is based on finite m.

Variance ratio's (1)

Proportion of the variation attributable to the missing data

$$\lambda = \frac{B + B/m}{T}$$

Relative increase in variance due to nonresponse

$$r = \frac{B + B/m}{\bar{U}}$$

These are related by $r = \lambda/(1 - \lambda)$.

Variance ratio's (2)

Fraction of information about Q missing due to nonresponse

$$\gamma = \frac{r + 2/(\nu + 3)}{1 + r}$$

This measure needs an estimate of the degrees of freedom ν (c.f. section 2.3.6)

Relation between γ and λ

$$\gamma = \frac{\nu + 1}{\nu + 3}\lambda + \frac{2}{\nu + 3}.$$

The literature often confuses γ and λ.

Statistical inference for $\bar{Q}$ (1)

The $100(1-\alpha)\%$ confidence interval of a $\bar{Q}$ is calculated as

$$\bar{Q} \pm t_{(\nu,1-\alpha/2)}\sqrt{T},$$

where $t_{(\nu,1-\alpha/2)}$ is the quantile corresponding to probability $1-\alpha/2$ of t_ν.

For example, use $t(10, 0.975) = 2.23$ for the 95% confidence interval for $\nu = 10$.

Statistical inference for $\bar{Q}$ (2)

Suppose we test the null hypothesis $Q = Q_0$ for some specified value Q_0. We can find the P-value of the test as the probability

$$P_s = \Pr\left[F_{1,\nu} > \frac{(Q_0 - \bar{Q})^2}{T}\right]$$

where $F_{1,\nu}$ is an F distribution with 1 and ν degrees of freedom.

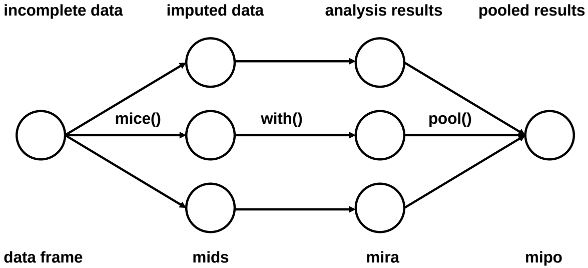
How large should m be?

Classic advice: $m = 3, 5, 10$. More recently: set m higher: 20–100.

Some advice:

- Use $m = 5$ or $m = 10$ if the fraction of missing information is low, $\gamma < 0.2$.
- Develop your model with $m = 5$. Do final run with m equal to percentage of incomplete cases.

Multiple imputation in `mice`



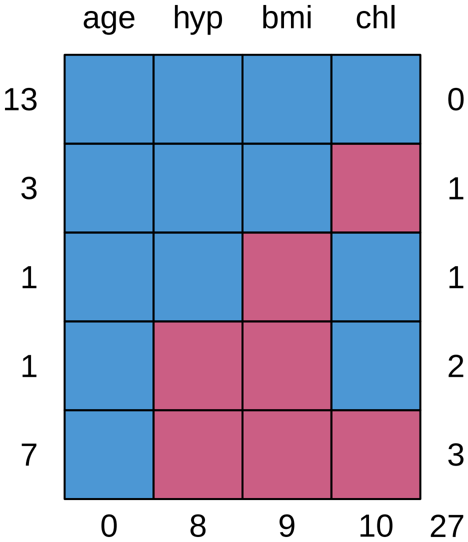
Inspect the data

```r
library("mice")
head(nhanes)
```

```
##   age  bmi hyp chl
## 1   1   NA  NA  NA
## 2   2 22.7   1 187
## 3   1   NA   1 187
## 4   3   NA  NA  NA
## 5   1 20.4   1 113
## 6   3   NA  NA 184
```

Inspect missing data pattern

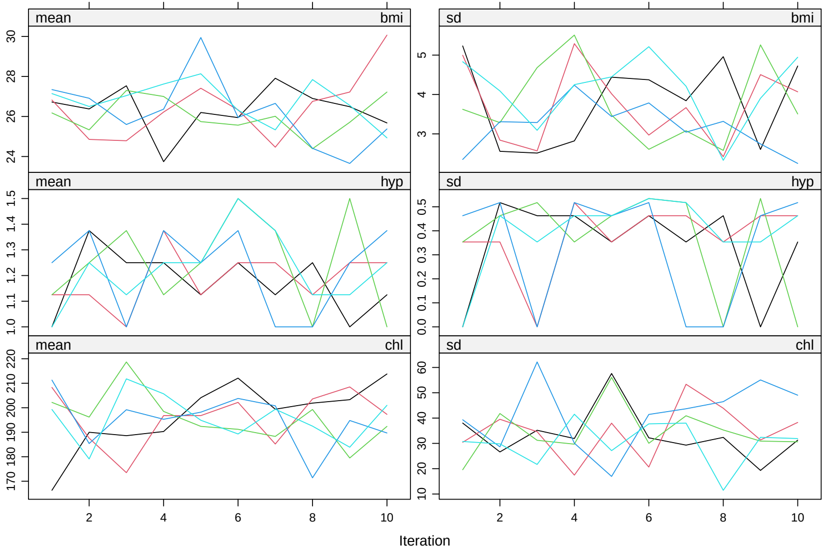
```r
md.pattern(nhanes)
```



Multiply impute the data

```
imp <- mice(nhanes, print = FALSE, maxit=10, seed = 24415)
```

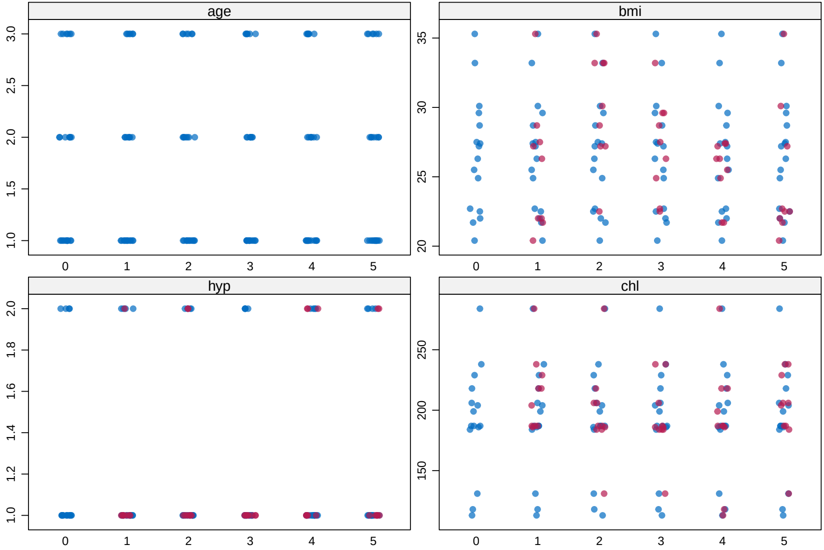
Inspect the trace lines for convergence



Stripplot of observed and imputed data

```r
stripplot(imp, pch = 20, cex = 1.2)
```

Stripplot of observed and imputed data



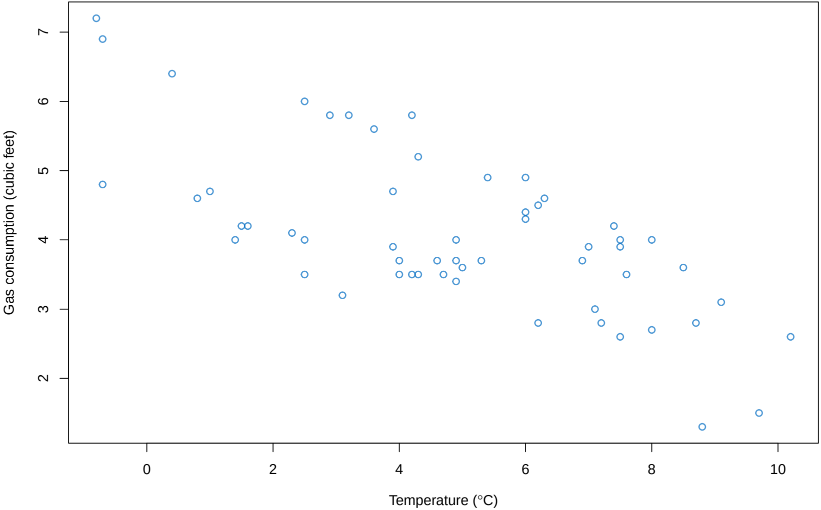
Fit the complete-data model

```r
fit <- with(imp, lm(bmi ~ age))
est <- pool(fit)
summary(est)
```

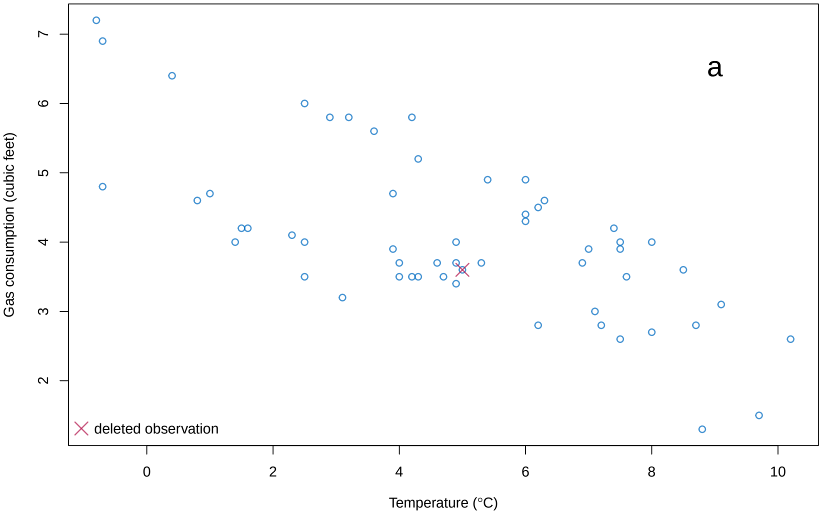
```
##          term estimate std.error statistic    df  p.valu
## 1 (Intercept)    30.01      2.44     12.32  8.01 1.73e-0
## 2         age    -1.94      1.12     -1.73 11.92 1.10e-0
```

Generating imputations, univariate

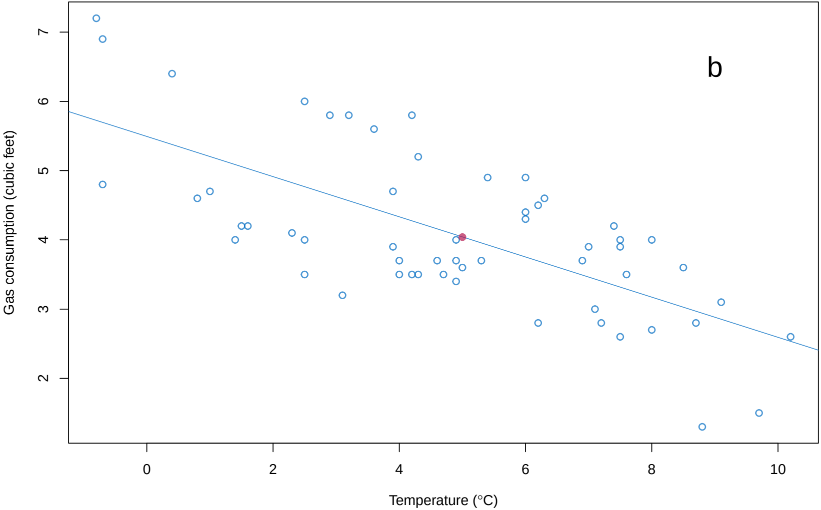
Relation between temperature and gas consumption



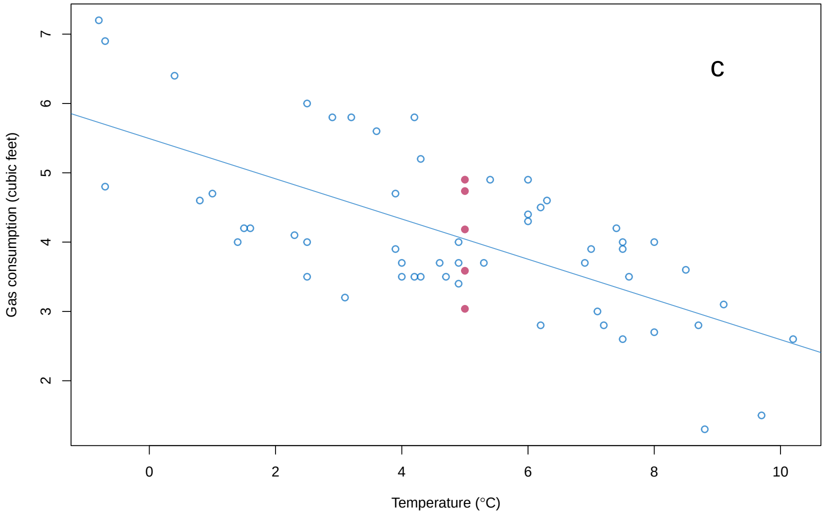
We delete gas consumption of observation 47



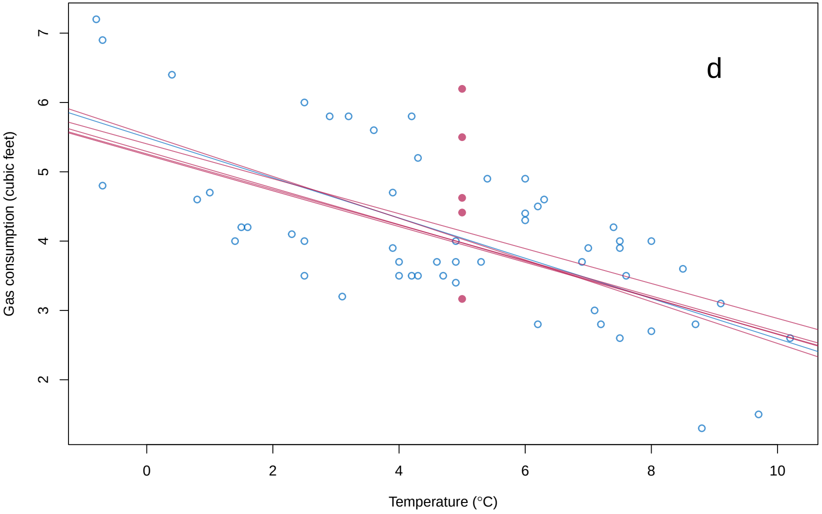
Predict imputed value from regression line



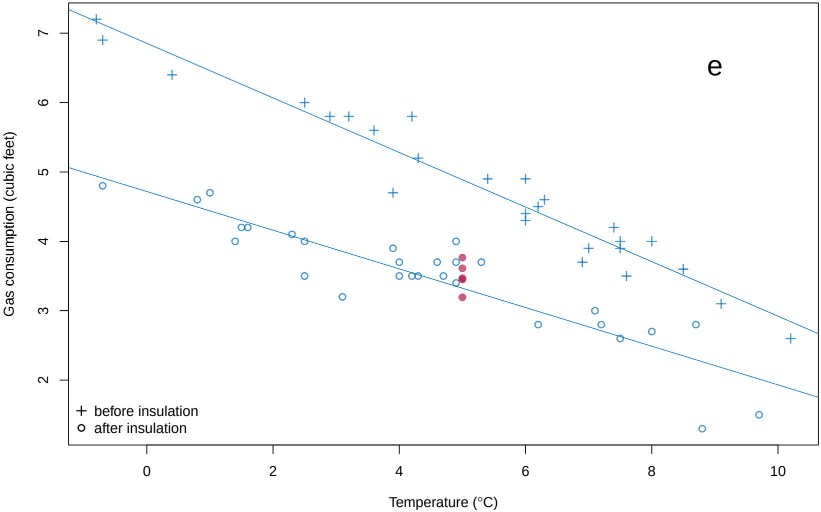
Predicted value + noise



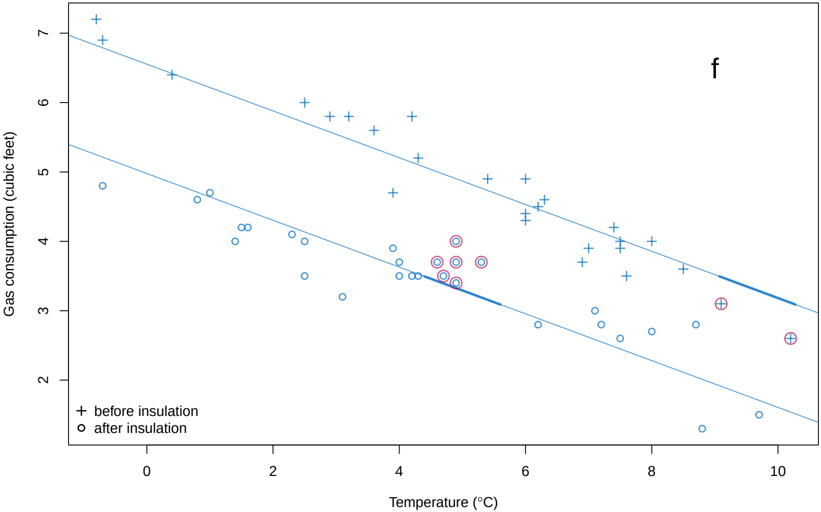
Predicted value + noise + parameter uncertainty



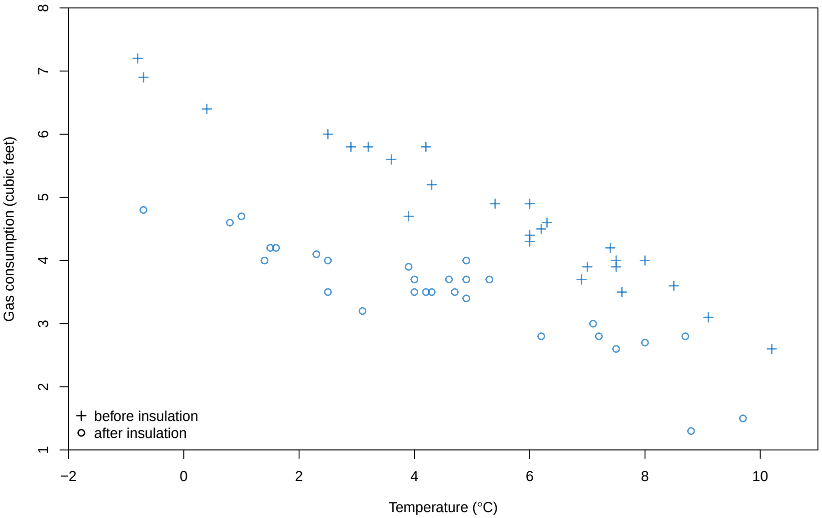
Imputation based on two predictors



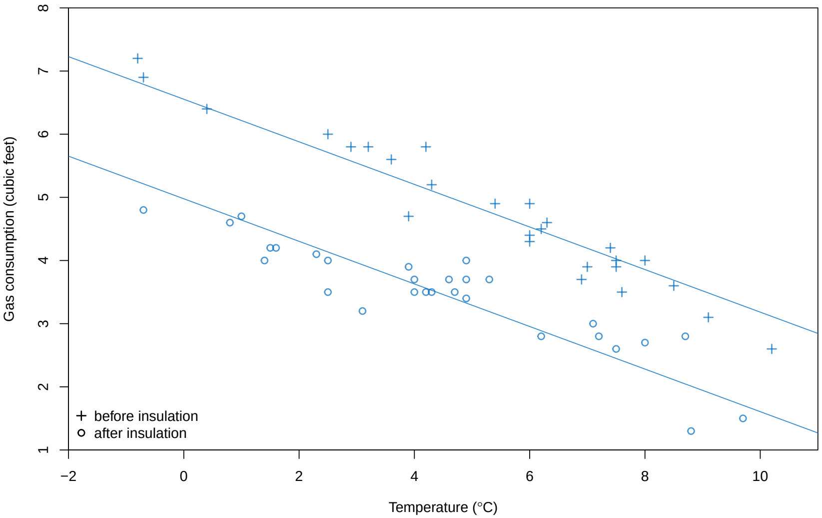
Drawing from the observed data



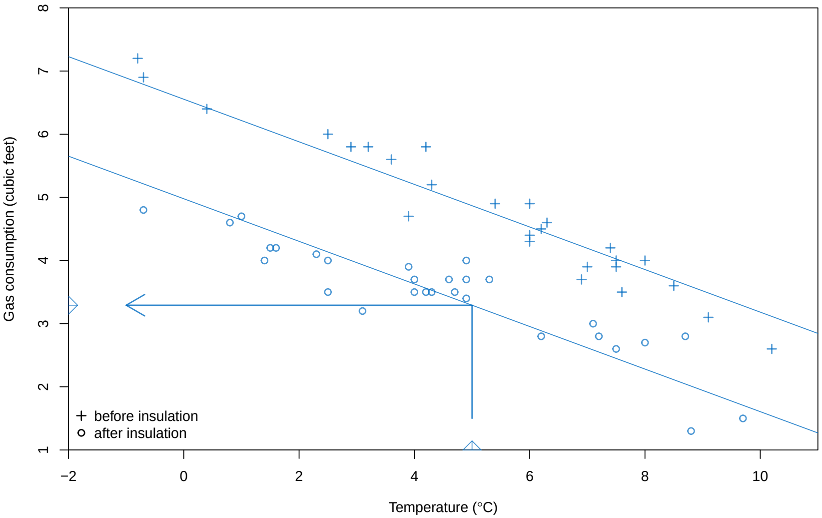
Predictive mean matching



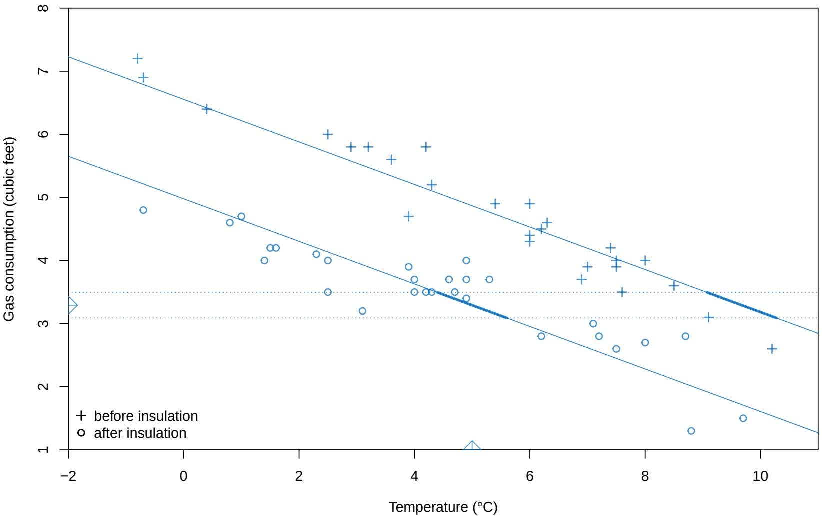
PMM: Add two regression lines



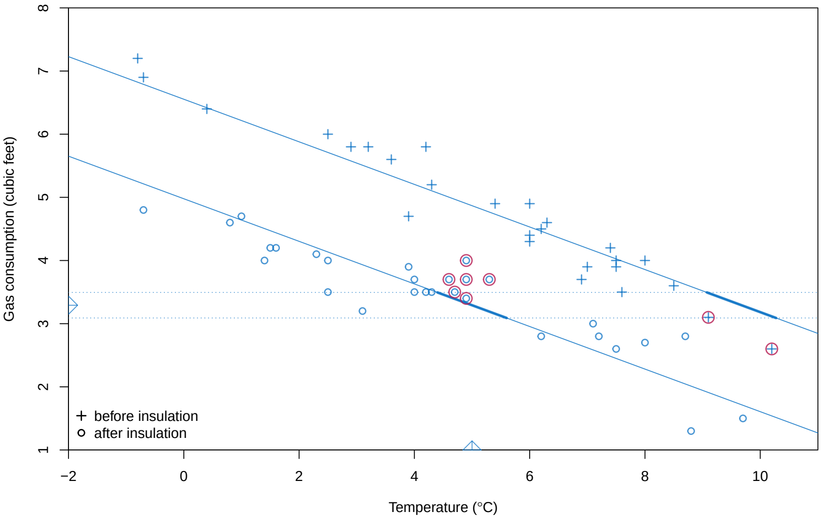
PMM: Predicted given 5°,C, ‘after insulation’



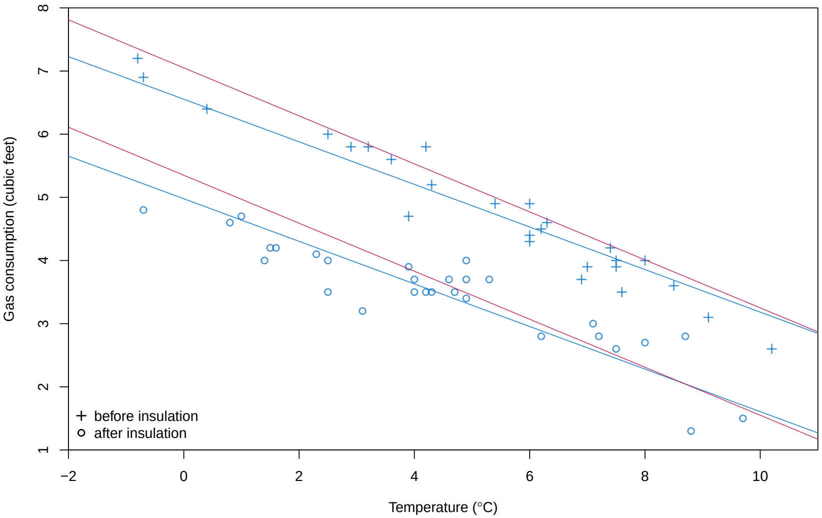
PMM: Define a matching range $\hat{y} \pm \delta$



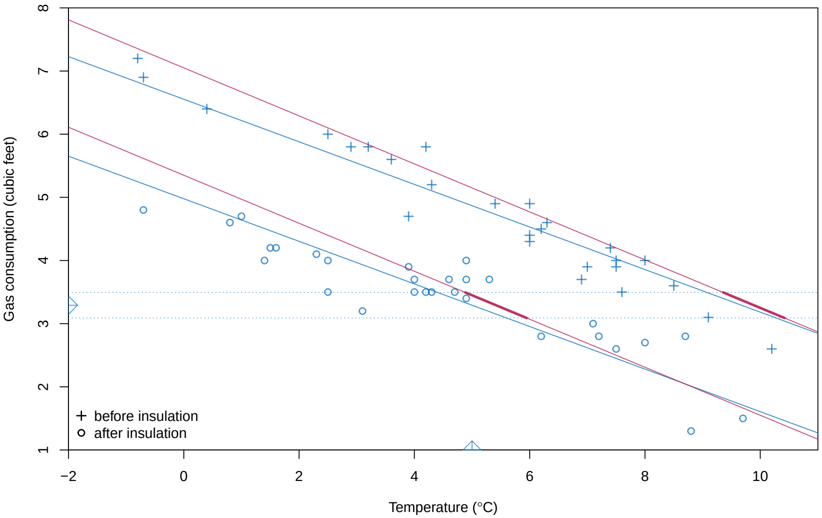
PMM: Select potential donors



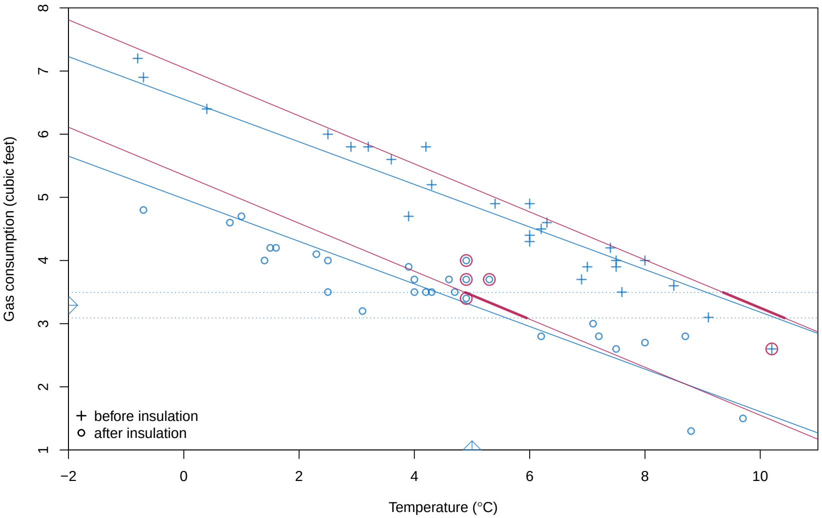
PMM: Bayesian PMM: Draw a line



PMM: Define a matching range $\hat{y} \pm \delta$



PMM: Select potential donors

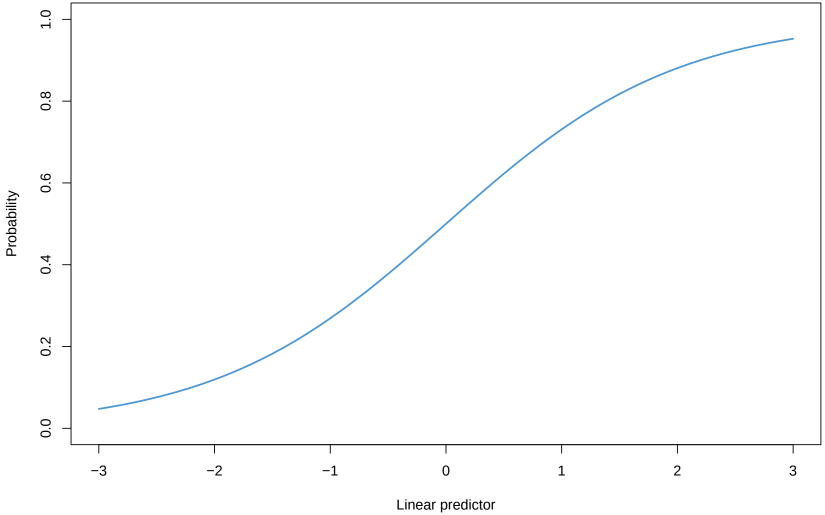


Imputation of a binary variable

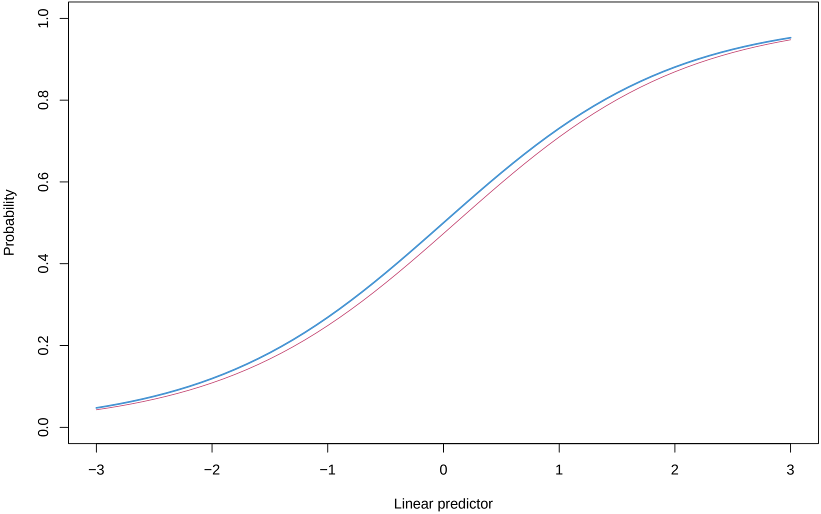
- Logistic regression

$$\Pr(y_i = 1 | X_i, \beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

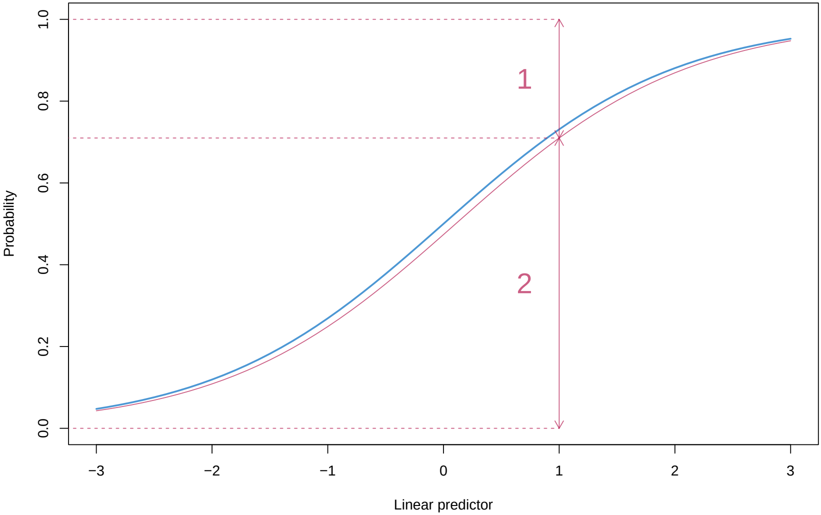
Fit logistic model



Draw parameter estimate



Read off the probability



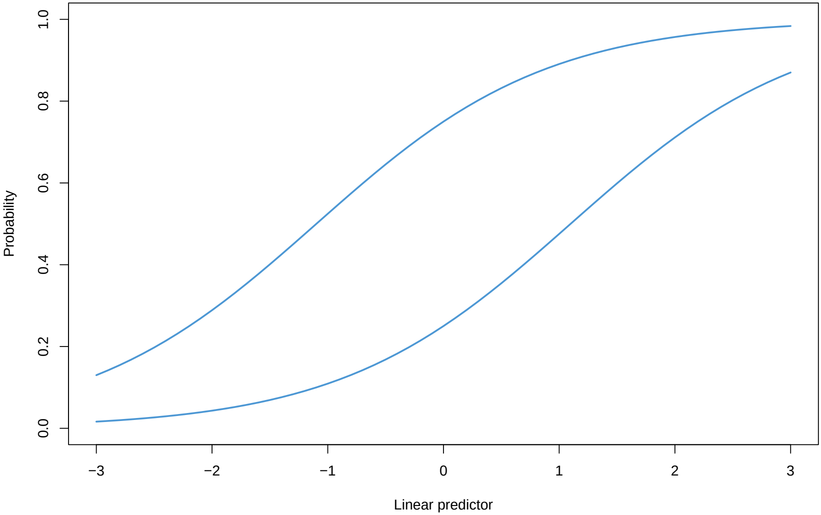
Impute ordered categorical variable

- K ordered categories $k = 1, \ldots, K$
- *ordered logit model*, or
- *proportional odds model*

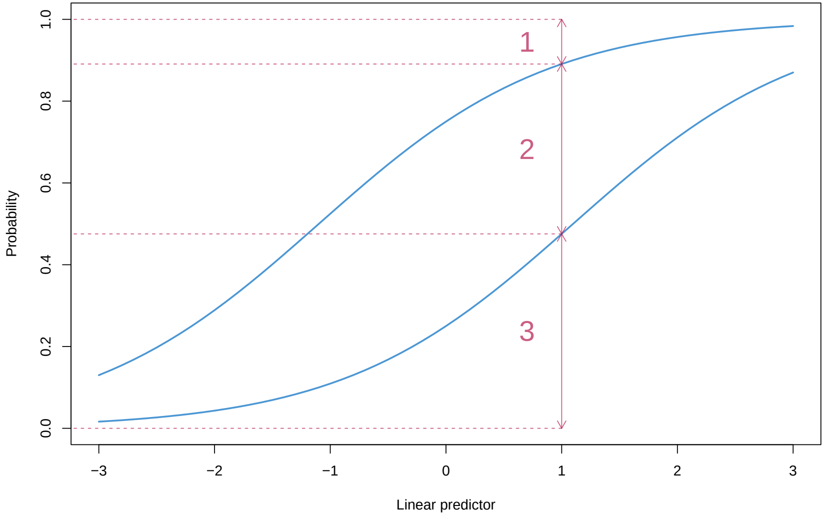
$$\Pr(y_i = k | X_i, \beta) = \frac{\exp(\tau_k + X_i \beta)}{\sum_{k=1}^{K} \exp(\tau_k + X_i \beta)}$$

-

Fit ordered logit model



Read off the probability



Built-in imputation functions

https://amices.org/mice/reference/index.html

Next week

- Aproaches to multivariate missing data
- MICE algorithm
- Pooling
- Workflows
- Specification of imputation model
- Multilevel data