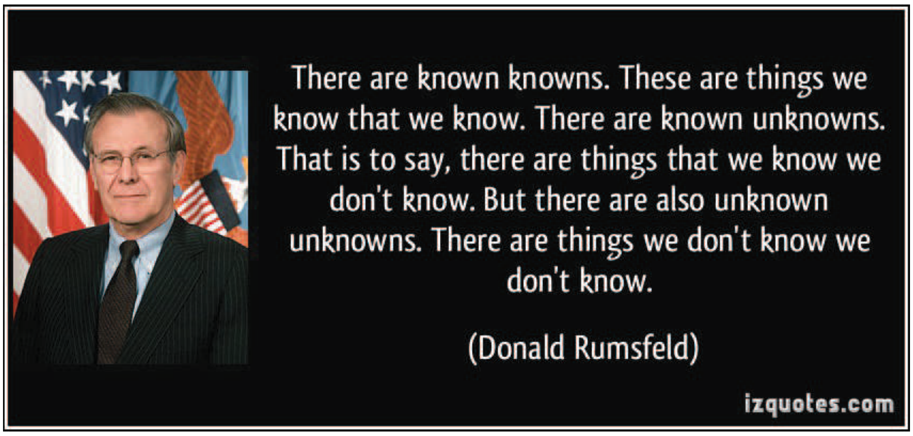
Missing Data: Problems, risks and solutions

Ariel Alonso Abad

Catholic University of Leuven

Missing data: The unknown unknowns



United States Secretary of Defense, Donald Rumsfeld talking about the missing WMD

Missing data problem

- Missing data: Ubiquitous presence in science
- Highly technical/mathematical field
- Only basic definitions, principles and methods
  - Less mathematically involved
  - Easy to implement in standard packages
- Advanced topics
  - Pattern mixture models
  - Selection models
  - Shared parameter models
  - Bodyguard theorem

Titanic

On the 10 April 1912 the largest passenger steamship in the world left Southampton England, to New York City. At 23:40 on 14 April, it struck an iceberg and sank at 2:20 the following morning, resulting in the deaths of 1,517 people in one of the deadliest peacetime maritime disasters in history.

Titanic

On the 10 April 1912 the largest passenger steamship in the world left Southampton England, to New York City. At 23:40 on 14 April, it struck an iceberg and sank at 2:20 the following morning, resulting in the deaths of 1,517 people in one of the deadliest peacetime maritime disasters in history.

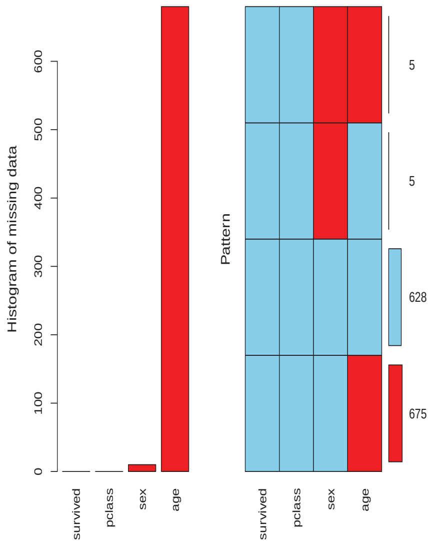
Titanic: Missing data

- Information on 1313 passengers
- Variables:
  - Survival: Y values 1/0.
  - age in years.
  - class: 1st, 2nd, 3rd.
  - sex: 1 male, 0 female.
- Adjusting by age, had class and gender an effect on survival?

Reading the data in R

```
> ## Needed libraries
>
> library(mice)
> library(lattice)
> library(VIM)
> library(aod)
> library(BaM)
>
> ## Reading the data
>
> titanic.missing <- read.table("titanicmissing.txt", header=T, sep=",")
> head(titanic.missing,10)
>
   survived pclass sex     age
1         1    1st   0 29.0000
2         0    1st   0  2.0000
3         0    1st   1 30.0000
4         0    1st   0 25.0000
5         1    1st   1  0.9167
6         1    1st   1 47.0000
7         1    1st   0 63.0000
8         0    1st   1 39.0000
9         1    1st   0 58.0000
10        0    1st   1 71.0000
11        0    1st   1 47.0000
12        1    1st   0 19.0000
13        1    1st   0      NA
14        1    1st   1      NA
15        0    1st   1      NA
>
```


Titanic: Missing data

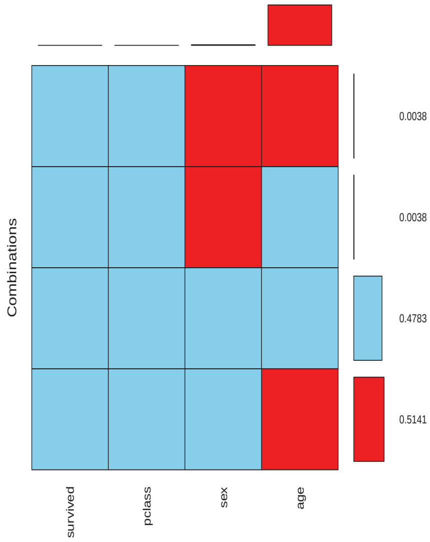



```
> ## Exploring the missingness
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)

> barMiss(titanic.missing[,c("survived","age")])
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```

Titanic: Missing data

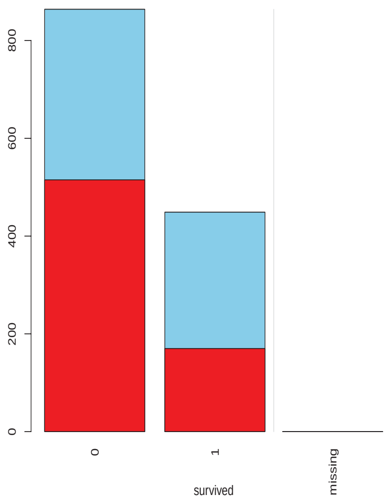



```
> ## Exploring the missingness
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)

> barMiss(titanic.missing[,c("survived","age")])
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```


Titanic: Missing data

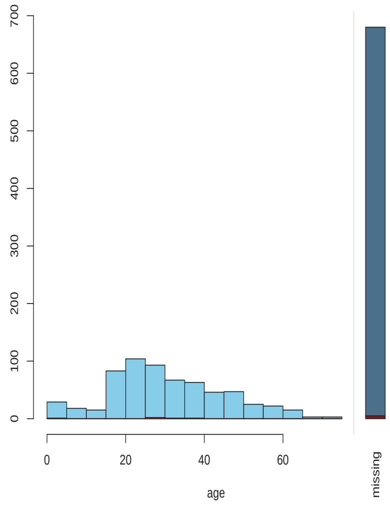



```
> ## Exploring the missingness
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)

> barMiss(titanic.missing[,c("survived","age")])
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```

Titanic: Missing data

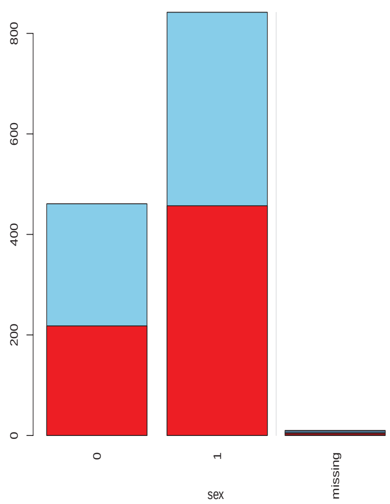



```
> ## Exploring the missingness
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)

> barMiss(titanic.missing[,c("survived","age")])
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```


Titanic: Missing data

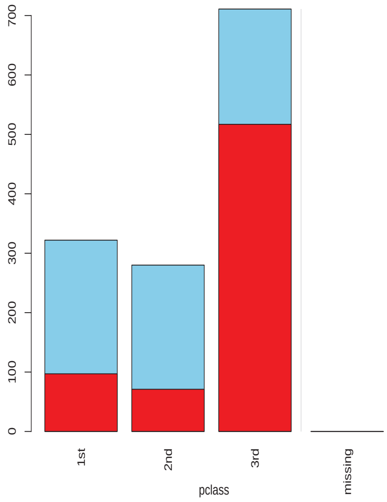



```
> ## Exploring the missingness
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)

> barMiss(titanic.missing[,c("survived","age")])
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```

Titanic: Missing data

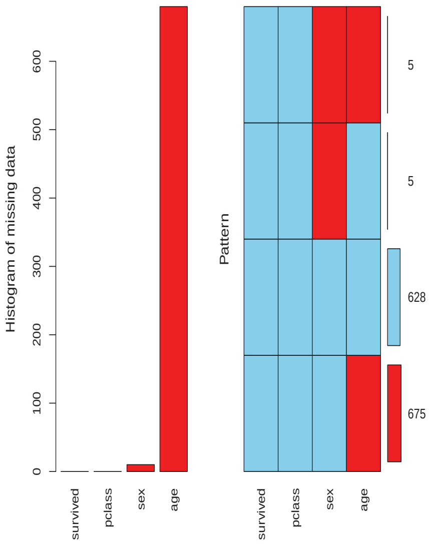



```
> ## Exploring the missingness
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)

> barMiss(titanic.missing[,c("survived","age")])
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```


Titanic: Missing data



- Only 628 completers.
- Age: more than 50% missing.
- Complete case analysis.

Titanic: Model

Analysis Model

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot class_2 + \beta_3 \cdot class_3 + \beta_4 \cdot age$$

Equivalently

$$P(Y = 1|class, sex, age) = \frac{e^{\beta_0+\beta_1 sex+\beta_2 class_2+\beta_3 class_3+\beta_4 age}}{1 + e^{\beta_0+\beta_1 sex+\beta_2 class_2+\beta_3 class_3+\beta_4 age}}$$

Analyzing the data in R

```
> ## Fitting a logistic regression model for the complete cases
>
> titanic.logistic.omit<-glm(survived ~ pclass + sex + age, family=binomial, data = titanic.missing)
> summary(titanic.logistic.omit)
>
Call:
glm(formula = survived ~ pclass + sex + age, family = binomial,
    data = titanic.missing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9519  -0.6500  -0.3172   0.5857   2.6875

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.439356   0.470042   9.445  < 2e-16 ***
pclass2nd   -1.466980   0.282904  -5.185 2.16e-07 ***
pclass3rd   -2.793512   0.338744  -8.247  < 2e-16 ***
sex         -3.085718   0.240780 -12.816  < 2e-16 ***
age         -0.047645   0.008747  -5.447 5.12e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 862.77  on 627  degrees of freedom
Residual deviance: 536.08  on 623  degrees of freedom
  (685 observations deleted due to missingness)
AIC: 546.08

Number of Fisher Scoring iterations: 5
>
```

Analyzing the data in R

```
> ## Global effect of class
>
> wald.test(b=coef(titanic.logistic.omit), Sigma=vcov(titanic.logistic.omit),
+ Terms=2:3)
>

Wald test:
----------

Chi-squared test:
X2 = 68.3, df = 2, P(> X2) = 1.4e-15

>
```


Titanic: Results

Coefficient	Explanation	Estimate	Std. Error	z value	p-value
β_0	Intercept	4.43	0.470	9.45	0.00
β_1	sex	-3.09	0.241	-12.82	0.00
β_2	2nd	-1.47	0.282	-5.19	0.00
β_3	3rd	-2.79	0.339	-8.25	0.00
β_4	age	-0.05	0.009	-5.45	0.00

χ^2 test for the effect of class

χ^2	df	p-value
68.3	2	0.00

Analyzing the data in R

```
> ## Odds ratios
> exp(cbind(OR =titanic.logistic.omit$coefficients,
+ confint(titanic.logistic.omit)))
>
Waiting for profiling to be done...
                     OR        2.5 %       97.5 %
(Intercept) 84.72034359 34.83140641 220.47300701
pclass2nd    0.23062102  0.13097332   0.39776333
pclass3rd    0.06120586  0.03083499   0.11664061
sex          0.04569723  0.02805441   0.07223138
age          0.95347203  0.93687359   0.96961325
>
```

- Odds of survival 77% smaller in 2nd class than in 1st class
- Odds of survival 93% smaller in 3rd class than in 1st class
- Odds of survival 95% smaller for men than for women
- An increase of one year in age is associated with a decrease of 5% in the odds of survival

Titanic: Results

Probability of surviving by gender and class after fixing age to its mean value $age = 31.27$

	Probability of Surviving		
Sex	1st	2nd	3rd
Male	0.47	0.18	0.05
Female	0.95	0.82	0.54

- Huge effect of class
  - Males: 1st class nine times more chance than 3rd class.
  - Females: 1st class two times more chance than 3rd class.
- Huge effect of gender: in 1st, 2nd and 3rd class women had 2, 5 and 11 times more chance to survive than men.

Missing data

- Common in many scientific investigations
  - A questionnaire got lost
  - Some subjects did not report their income
  - A machine got broken
- Determining the appropriate analytic approach is a major question
  - Throw them away?
  - Make a guess about their values?
  - Use the information available?
- Development of statistical methods has been an active area of research.

Missing data

Missing data:

Observations that are intended to be made but are not made.

Two possible, but distinct, goals

- Make inferences that would apply to the population targeted by the complete sample.
- Make inferences that would apply to those subject remaining in the study, or with complete data (on relevant variables).

We will focus on the first of these.

Missing data: Common strategies

Complete Cases: In complete cases (CC), sometimes also called listwise deletion (LD), all cases with missing values are deleted. Following deletion, conventional methods are used to derive estimates from the remaining, complete cases. **Default in many software.**

Available Cases: In available cases (AC), also called pairwise deletion (PD), each moment is estimated separately using cases with values for the pertinent variables.

Mean substitution: Special case of imputation. Substitution of missing values with the simple (grand) mean (MS).

Last Observation Carried Forward: Special case of imputation. Whenever a value is missing, the last observed value is substituted (LOCF).

Does it really matter?

Missing data: Simulation I

Generated 50 000 observations from the bivariate normal (X, Y) with correlation $\rho = 0.5$. So at the population level

$$\mu_X = E(X) = 0 \qquad \text{Var}(X) = 1$$
$$\mu_Y = E(Y) = 0 \qquad \text{Var}(Y) = 1$$
$$\rho = \text{Corr}(X, Y) = 0.5$$

Three settings where Y is always observed and X sometimes missing

- X is missing with probability 0.5.
- X is missing if $Y < 0$.
- X is missing if $X < 0$.

X is missing with probability 0.5

Obs	X	Y
1	-1.62	-0.05
2	0.49	0.13
3	-0.19	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

X is missing with probability 0.5

- CC means deleting cases where X is missing.

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

Obs	X	Y
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮
⋮	⋮	⋮
⋮	⋮	⋮

X is missing with probability 0.5

- CC means deleting cases where X is missing.
- Following deletion, conventional methods are used to derive estimates from the remaining, complete cases.

	Population values			Complete Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.007	0.99	0.499
Y	0	1		0.001	1.00	

X is missing with probability 0.5

- AC: each moment is estimated separately using cases with values for the pertinent variables.

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

- To calculate $\overline{Y}$
- To calculate $\overline{X}$
- To calculate Corr(X, Y)

X is missing with probability 0.5

- AC: each moment is estimated separately using cases with values for the pertinent variables.

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

- To calculate $\overline{Y}$
- To calculate $\overline{X}$
- To calculate Corr(X, Y)

X is missing with probability 0.5

- AC: each moment is estimated separately using cases with values for the pertinent variables.

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

- To calculate $\overline{Y}$
- To calculate $\overline{X}$
- To calculate Corr(X, Y)

X is missing with probability 0.5

- AC: each moment is estimated separately using cases with values for the pertinent variables.
- E(Y) and Var(Y) would be estimated using all the cases.
- E(X), Var(X), and cov(X, Y) would be estimated using only the cases with values for X.

	Population values			Available Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.007	0.99	0.499
Y	0	1		0.003	1.00	

X is missing with probability 0.5

- Mean imputation: All missing values in X are imputed using $\bar{X}_n$.
- Following imputation, conventional methods are used to derive estimates.

	Population values			Mean imputation		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.007	0.496	0.249
Y	0	1		0.003	1.00	

X is missing if $Y < 0$

Obs	X	Y
1	-1.62	-0.05
2	0.49	0.13
3	-0.19	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

	X	Y
1	NA	-0.05
2	0.49	0.13
3	NA	-0.59
4	-0.29	0.79
5	NA	-2.25
6	0.94	0.07
7	NA	-0.82
8	NA	-0.12
9	NA	-0.38
10	NA	-0.89
⋮	⋮	⋮

X is missing if $Y < 0$

- CC means deleting cases where X is missing since $Y < 0$.

	Population values			Complete Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.397	0.841	0.185
Y	0	1		0.798	0.363	

- AC: Moments estimated using cases with values for the pertinent variables.

	Population values			Available Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.397	0.841	0.185
Y	0	1		0.005	0.996	

X is missing if $Y < 0$

- Mean imputation.

	Population values			Mean imputation		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.399	0.421	0.009
Y	0	1		0.003	1.00	

Simulation I: Conclusion

Case I: X is missing with probability 0.5

- No major problems.
- CC and AC worked fine.
- MS failed.

Case II: X is missing if $Y < 0$. All methods seem to fail.

Case III: X is missing if $X < 0$. Not shown here but again all methods seem to fail.

What is going on?

Missing data mechanism

Everything depends on how the missing values got missing. Let missingness be the probability that a value is missing (P (Missing))

- Missing not at random (MNAR): P (Missing) depends on both observed and missing values.
- Missing at random (MAR): P (Missing) depends only on observed values.
- Missing completely at random (MCAR): P (Missing) depends neither on observed nor on unobserved values.

Missing not at random: MNAR

The probability that an observation is missing depends on subject information that is not observed, like the value of the missing observation itself

- Studying mental health: People who have been diagnosed as depressed may report their mental status less often than others.
- Asking for income level: Missing data may be more likely to occur when the income level is relatively high/low.

MNAR: Highly problematic.

The only way to obtain unbiased estimates of the parameters is to model missingness. In other words, we would need to write a model that accounts for the missing data mechanism and **hope** this model is approximately correct.

Missing at random: MAR

The probability that an observation is missing depends on subject information that is present, i.e., missingness can be described using observed subject variables

- Depressed people may be less inclined to report their income and, hence, missingness in income will be related to depression. If mental status is always observed then missingness in income is MAR.
- Depressed people may also have a lower income. A high rate of missing data among depressed individuals $\Rightarrow$ observed mean income might be much larger than it would be without missing data.
- The probability of drop out may depend on the treatment received.

MAR: No simple methods.

Generally, under MAR, simple techniques like complete and available case analysis and overall mean imputation, give biased results.

Missing completely at random: MCAR

The probability that an observation is missing is not related to any other subject characteristics

- Equipment malfunctioned.
- The weather was terrible.
- Data were not entered correctly.

MCAR: Most methods work.

Although very inefficient, some simple techniques like complete and available case analysis will give unbiased results under MCAR. However, MS and LOCF **do not** work in this setting neither.

Missing Data: Formal definitions

Formal Definitions

- Let $\boldsymbol{z} = (\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis})$ be the vector variables whose joint distribution depends on the parameter vector $\boldsymbol{\theta}$.
- The locations of the missing values are summarized by the vector $\boldsymbol{r}$ with $r_i = 0$ if z_i is missing and 1 otherwise. If $p(\boldsymbol{r}|\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}, \phi)$ is the distribution of $\boldsymbol{r}$ then
  - MCAR: $p(\boldsymbol{r}|\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}, \phi) = p(\boldsymbol{r}|\phi)$.
  - MAR: $p(\boldsymbol{r}|\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}, \phi) = p(\boldsymbol{r}|\boldsymbol{z}_{obs}, \phi)$.
  - MNAR: $p(\boldsymbol{r}|\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}, \phi)$.

Simulation I: Remarks

Three settings where Y is always observed and X sometimes missing

- MCAR: X is missing with probability 0.5.
- MAR: X is missing if $Y < 0$.
- MNAR: X is missing if $X < 0$.

Titanic: Simulations II

Simulations mimicking Titanic data set

- Age simulated mimicking the original data.
- Gender: $sex \sim \text{Bernoulli}(0.5)$. For men $sex = 1$.
- Only two classes considered $class = 1$ indicating first class.
- Survival (Y) like in case study and

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = 2.18 + 1.93 \cdot class - 3.04 \cdot sex - 0.04 \cdot age$$

Titanic: The incomplete data

Generating the missing data

- 2500 datasets were generated each with 1000 passengers.
- Missing data created for age.
- The probability of age being missing depending on:
  - Class: First class less chance of missing age
  - Survival: Survivors less chance of missing age
  - Missing generating mechanism MAR.

$$\text{logit}\left[P(r = 0|class, Y)\right] = 2.11 - 1.5 \cdot class - 2.85 \cdot Y$$

$r = 0$ implies that age is missing.

Titanic: Simulations II

Analysis

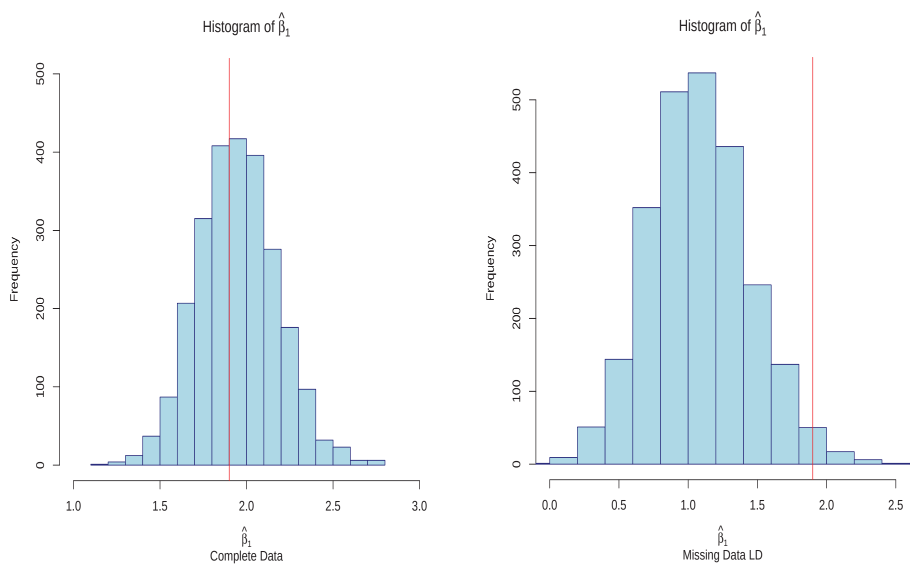
Model:

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = \beta_0 + \beta_1 \cdot class + \beta_2 \cdot sex + \beta_3 \cdot age$$

Two types of analysis

- Data sets without missing values (Complete data).
- Data sets with missing values analyzed using complete cases (also called listwise deletion LD). Around 50% of the observations in each dataset had missing age.

Simulations II: Results



How to handle missing data?

Three methods to handle missing values:

- Complete Cases: Just analyse individuals with complete data.
- Multiple imputation (MI): Stochastically fill in missing values using observed data
  - Create multiple complete datasets
  - Apply complete-data estimators to each
  - Combine estimates (Rubin's Rules)
- Inverse probability weighting (IPW): Like Complete Cases but weight every individual by the inverse of P(no-missing)=(his probability of not having missing information).

Important

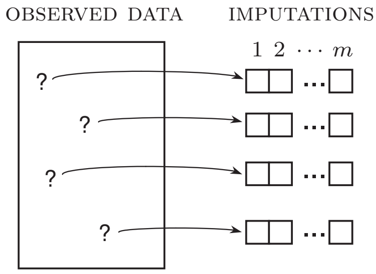
CC valid only under MCAR. IPW and MI valid under MAR!!!

Multiple imputations

Why multiple imputations?

- Single imputation techniques overestimate precision, since no correction is made for the uncertainty introduced from imputing the missing observations.
- This additional variability in the estimates is made explicit by generating multiple completed data sets.
- Each time replace missing values $\boldsymbol{z}_{mis}$ by draws from the conditional distribution $f(\boldsymbol{z}_{mis}|\boldsymbol{z}_{obs}, \widehat{\psi})$, rather than by the average of that distribution.

A simulation-based approach to missing data



- Generate $m > 1$ plausible versions of $\boldsymbol{z}_{mis}$.
- Analyze each of the m datasets by standard complete-data methods.
- Combine the results.

Rubin (1987) calls this the repeated-imputation inference method

How to impute the missing values

Imputation model: For instance for the Titanic simulations one can impute age for subject i using the model $f(\boldsymbol{z}_{mis}|\boldsymbol{z}_{obs}, \boldsymbol{\psi})$

$$age_i = \gamma_0 + \gamma_1 class_i + \gamma_2 sex_i + \gamma_3 Y_i + \epsilon_i$$

- First this model is fitted to the completers to estimate the parameters $\widehat{\boldsymbol{\psi}} = (\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3, \widehat{\sigma}^2_\epsilon)$.
- Note that there is also variability introduced from replacing $\boldsymbol{\psi}$ in $f(\boldsymbol{z}_{mis}|\boldsymbol{z}_{obs}, \boldsymbol{\psi})$ by an estimate.
- However, we usually have an estimate for the variation in $\widehat{\boldsymbol{\psi}}$: $\widehat{\boldsymbol{\psi}} \sim N(\boldsymbol{\psi}, \widehat{\Sigma}_\psi)$.
- Drawing $\boldsymbol{\psi}$ from $N(\widehat{\boldsymbol{\psi}}, \widehat{\Sigma}_\psi)$ accounts for this additional variation.

The imputation algorithm

- Draw $\psi^{(k)}$ from $N(\widehat{\psi}, \widehat{\Sigma}_\psi)$
- Draw $\boldsymbol{z}_{mis}^{(k)}$ from $f(\boldsymbol{z}_{mis} \,|\boldsymbol{z}_{obs}, \psi^{(k)})$
- Using the completed data $(\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}^{(k)})$, calculate an estimate $\widehat{\boldsymbol{\theta}}^{(k)}$ for the parameter $\boldsymbol{\theta}$ of interest, as well as its covariance matrix $\boldsymbol{U}^{(k)}$
- Repeat this m times
- Note that $\boldsymbol{U}^{(k)}$ reflects the sampling uncertainty, i.e., the uncertainty in the estimates of $\boldsymbol{\theta}$ due to the fact that only a finite sample is available.
- We can now obtain inferences for $\boldsymbol{\theta}$ from pooling the estimates

$$\widehat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{k=1}^{m} \widehat{\boldsymbol{\theta}}^{(k)}$$

The imputation algorithm

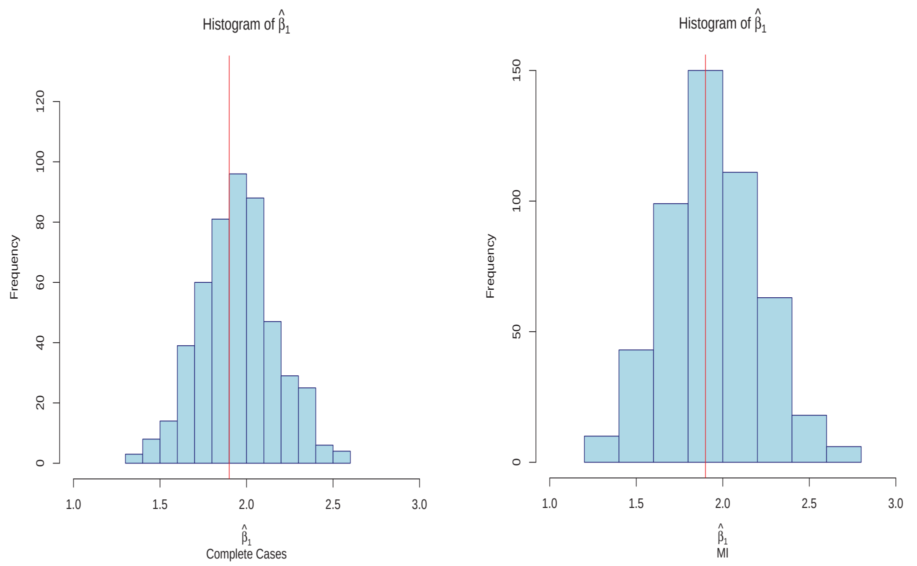
- The covariance matrix of $\widehat{\boldsymbol{\theta}}$ equals

$$\text{var}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{W}} + \left(\frac{m+1}{m}\right)\widehat{\boldsymbol{B}}$$

where $\widehat{\boldsymbol{W}} = \dfrac{\sum_{k=1}^{m} \boldsymbol{U}^{(k)}}{m}$ and $\widehat{\boldsymbol{B}} = \dfrac{\sum_{k=1}^{m}\left(\widehat{\boldsymbol{\theta}}^{(k)} - \widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}^{(k)} - \widehat{\boldsymbol{\theta}}\right)'}{m-1}$.

- $\widehat{\boldsymbol{W}}$ represents the within-imputation variance, representing sampling uncertainty
- $\widehat{\boldsymbol{B}}$ represents the between-imputation variance, representing the uncertainty in imputing the missing observations as well as the uncertainty in the estimation of ψ.
- Typically, m will be small: $m = 5, 10$ already yields a major improvement over single imputation.

Titanic simulation: MI results (500 data sets and $m = 5$)



Multiple imputation in R

- Several packages available: Amelia, VIM, mice...
- Different algorithms
  - Amelia: Bootstrapped EM algorithm
  - VIM: Iterative Robust Model-based Imputation (irmi)
  - mice: Chained equations algorithm (CEA)
- CEA has been found to work well in a variety of simulation studies (Schunk 2008; Drechsler and Rassler 2008; Giorgi et al. 2008)
- Area of active research

Observations and warnings

- Variables used to impute a missing outcome may themselves be incomplete
- Rows or columns in the data can be ordered, e.g., as with longitudinal studies
- Variables can be of different types (e.g., binary, unordered, ordered, continuous), thereby making the application of theoretically convenient models, such as the multivariate normal, inappropriate
- Imputation can create impossible combinations (e.g. pregnant fathers), or destroy deterministic relations in the data (e.g. sum scores)
- Imputations can be nonsensical (e.g. body temperature of the dead)

Titanic data: MI

- We impute each missing value 100 times and fitted the following model to every imputed data set

$$\text{logit}\left[P(Y=1|class, sex, age)\right] = \beta_0 + \beta_1 sex + \beta_2 class_2 + \beta_3 class_3 + \beta_4 age$$

- We obtained then 100 estimates for the parameters of interest with $k = 1, 2, \dots, 100$

$$\widehat{\boldsymbol{\theta}}^{(k)} = (\widehat{\beta}_0^{(k)}, \widehat{\beta}_1^{(k)}, \widehat{\beta}_2^{(k)}, \widehat{\beta}_3^{(k)}, \widehat{\beta}_4^{(k)})$$

- We combined all these estimates using the Rubin's rules previously described.

Multiple imputation in R

```
> ###### Titanic multiple imputation
>
> ## Studying the patterns of missiness
>
> pattern=md.pattern(titanic.missing)
> pattern
    survived pclass sex age
628        1      1   1   1   0
  5        1      1   0   1   1
675        1      1   1   0   1
  5        1      1   0   0   2
           0      0  10 680 690
>
> pairs=md.pairs(titanic.missing)
> pairs
>
$rr                                      $rm
         survived pclass  sex age                 survived pclass sex age
survived     1313   1313 1303 633        survived        0      0  10 680
pclass       1313   1313 1303 633        pclass          0      0  10 680
sex          1303   1303 1303 628        sex             0      0   0 675
age           633    633  628 633        age             0      0   5   0


$mr                                      $mm
         survived pclass sex age                  survived pclass sex age
survived        0      0   0   0         survived        0      0   0   0
pclass          0      0   0   0         pclass          0      0   0   0
sex            10     10   0   5         sex             0      0  10   5
age           680    680 675   0         age             0      0   5 680
>
```


Multiple imputation in R

```
> ## Imputing the missing values
>
> imp <- mice(titanic.missing, m=100)
> imp
>
Multiply imputed data set
Call:
mice(data = titanic.missing, m = 100)
Number of multiple imputations:  100
Missing cells per column:
survived   pclass      sex      age
       0        0       10      680
Imputation methods:
survived   pclass      sex      age
      ""       ""    "pmm"    "pmm"
VisitSequence:
sex age
  3   4
PredictorMatrix:
         survived pclass sex age
survived        0      0   0   0
pclass          0      0   0   0
sex             1      1   0   1
age             1      1   1   0
Random generator seed value:  NA
>
## Imputations are generated according to the default method, which is, for numerical data, predictive
## mean matching (pmm) (Little 1988).
```

Diagnostic checking

- An important step in multiple imputation is to assess whether imputations are plausible
- Imputations should be values that could have been obtained had they not been missing
- Imputations should be close to the data
- Data values that are clearly impossible (e.g. negative counts, pregnant fathers) should not occur in the imputed data
- Imputations should respect relations between variables, and respect the appropriate amount of uncertainty about their *true* values
- Diagnostic checks on the imputed data provide a way to check the plausibility of the imputations

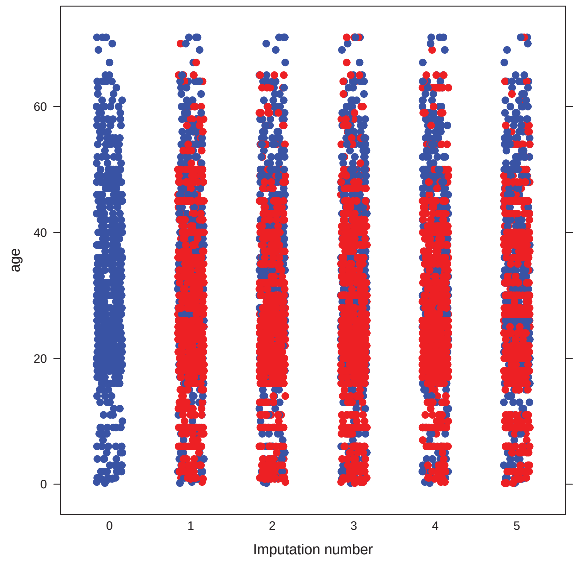
Diagnostic checking in R

```
> ## Imputed values for age. Each row corresponds to a missing entry in age.
> ## The columns contain the multiple imputations.
> imp$imp$age[1:10,1:5]
>
    1       2       3  4  5
13 60 27.0000 19.0000 55 22
14 57  0.9167 17.0000 48 26
15 47 28.0000 50.0000 47 31
30 28 28.0000 56.0000 55 40
33 22 37.0000 39.0000 24 30
36 50 50.0000 64.0000 61 27
41 30 34.0000  0.9167 34 38
46 62 46.0000 54.0000 36 39
47 61 58.0000 46.0000 61 24
53 45 37.0000 54.0000 21 23
>
> ## The complete data combine observed and imputed data.
> ## The first completed data set can be obtained as  (only first 10 passenger shown)
>
> complete(imp,1)[1:10,]
   survived pclass sex     age
1         1    1st   0 29.0000
2         0    1st   0  2.0000
3         0    1st   1 30.0000
4         0    1st   0 25.0000
5         1    1st   1  0.9167
6         1    1st   1 47.0000
7         1    1st   0 63.0000
8         0    1st   1 39.0000
9         1    1st   0 58.0000
10        0    1st   1 71.0000
>
```

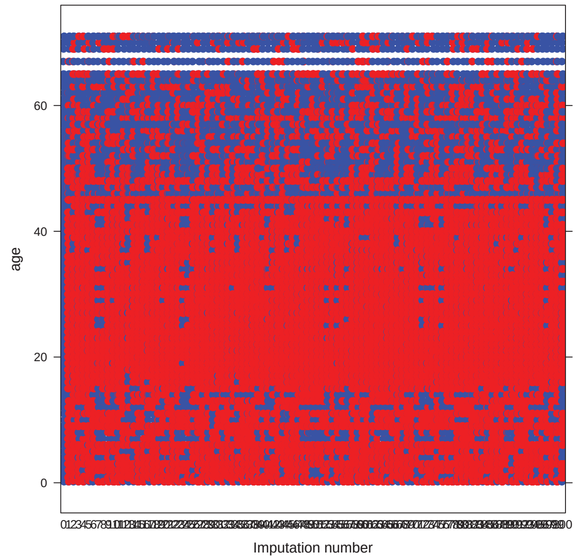
Diagnostic checking in R

```
> ## It is often useful to inspect the distributions of original and the imputed
> ## data. The complete() function extracts the original and the imputed data
> ## sets from the imp object as a long (row-stacked) matrix. The col vector
> ## separates the observed (blue) and imputed (red) data for age
>
> com <- complete(imp, "long", inc=T)
> col <- rep(c("blue","red")[1+as.numeric(is.na(imp$data$age))],101)
> stripplot(age~.imp, data=com, jit=TRUE, fac=0.8, col=col, pch=20, cex=1.4,
+           xlab="Imputation number")
>
```


Distributions: Original versus imputed data



Distributions: Original versus imputed data



Imputation methods

Method	Description	Scale type	Default
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression, non-Bayesian	numeric	
mean	Unconditional mean imputation	numeric	
2l.norm	Two-level linear model	numeric	
logreg	Logistic regression	factor, 2 levels	Y
polyreg	Polytomous (unordered) regression	factor, >2 levels	Y
lda	Linear discriminant analysis	factor	
sample	Random sample from the observed data	any	

- The method argument of mice() specifies the imputation method per column and overrides the default
- Columns that need not be imputed have method "",i.e.,

```
imp <- mice(titanic.missing, meth = c("", "", "logreg", "pmm"), m=100)
```

Analysis of imputed data in R

```
> ## Analyzing the imputed data sets
>
> fit <- with(data=imp, exp=glm(survived ~ pclass + sex + age, family=binomial))
>
> ## Creating a data set with the results of all the analysis
>
> MI.matrix<-matrix(0,100,5)
> for(k in 1:100){ MI.matrix[k,]<-coefficients(fit$analyses[[k]])}
> MI.results=data.frame(Intercept=MI.matrix[,1], pclass2=MI.matrix[,2],
+          pclass3=MI.matrix[,3], sex=MI.matrix[,4], age=MI.matrix[,5])
> MI.results[1:10,]
>
   Intercept   pclass2   pclass3       sex         age
1   3.321512 -1.201354 -2.606778 -2.437407 -0.03492116
2   4.042564 -1.412543 -2.858506 -2.657812 -0.04863579
3   4.217690 -1.531627 -3.031196 -2.593078 -0.05211627
4   3.504774 -1.316043 -2.749440 -2.387495 -0.03891783
5   4.399160 -1.584609 -3.001377 -2.631284 -0.05634107
6   3.668436 -1.331814 -2.821121 -2.402105 -0.04331810
7   3.686304 -1.385195 -2.826104 -2.452432 -0.04270390
8   3.597697 -1.306929 -2.874065 -2.417954 -0.04120242
9   3.751935 -1.395021 -2.781783 -2.437738 -0.04433574
10  3.598338 -1.283901 -2.764597 -2.450878 -0.04034921
>
```


Analysis of imputed data in R

```
> ## Combining the results using Rubin's rule
>
> est <- pool(fit)
> summary(est)
                    est          se          t       df     Pr(>|t|)       lo 95       hi 95
(Intercept)  3.62632240 0.464628361   7.804781 212.6374 2.664535e-13  2.71045483  4.54218997
pclass2     -1.30813713 0.248053338  -5.273612 639.4154 1.832732e-07 -1.79523474 -0.82103951
pclass3     -2.76475931 0.262026202 -10.551461 474.9145 0.000000e+00 -3.27963337 -2.24988524
sex         -2.48033948 0.168420817 -14.727036 919.6779 0.000000e+00 -2.81087322 -2.14980575
age         -0.04111961 0.009669708  -4.252415 171.3019 3.470195e-05 -0.06020674 -0.02203248

            nmis       fmi    lambda
(Intercept)   NA 0.5505477 0.5463401
pclass2       NA 0.2382350 0.2358561
pclass3       NA 0.3160640 0.3131898
sex           10 0.1416097 0.1397450
age          680 0.6210089 0.6166097
>
> ## The column fmi contains the fraction of missing information, i.e. the proportion of the
> ## variability that is attributable to the uncertainty causedby the missing data.
>
```

Titanic results: CC+MI

		CC		MI	
Coefficient	Explanation	Estimate	Std. Error	Estimate	Std. Error
β_0	Intercept	4.43	0.470	3.63	0.465
β_1	sex	-3.09	0.241	-2.48	0.168
β_2	2nd	-1.47	0.282	-1.31	0.248
β_3	3rd	-2.79	0.339	-2.76	0.262
β_4	age	-0.05	0.009	-0.04	0.009

- Some differences in the estimates of β_1 (sex) and β_2 (2nd class indicator)
- Although p-values differ we get the same qualitative conclusions
- It is not always like this

Inverse probability weighting (IPW)

Suppose we have the following data

Group	A			B			C		
Response	1	1	1	2	2	2	3	3	3

then the average response is 2. However if we observed

Group	A			B			C		
Response	1	?	?	2	2	2	?	3	3

then the average response is $13/6 = 2.17$ which is biased.

Inverse probability weighting (IPW)

Suppose we have the following data

Group		A			B			C	
Response	1	?	?	2	2	2	?	3	3
$P(\text{Response})$		$\frac{1}{3}$			1			$\frac{2}{3}$	
$\dfrac{1}{P(\text{Response})}$		3			1			$\frac{3}{2}$	

Calculate weighted average

$$\frac{1\cdot 3+(2+2+2)\cdot 1+(3+3)\cdot \dfrac{3}{2}}{3+1+1+1+\dfrac{3}{2}+\dfrac{3}{2}}=2$$

Thus IPW has eliminated the biased. Notice that this example is MAR.

Titanic: Simulations II

Simulations mimicking Titanic data set

- Age simulated mimicking the original data.
- Gender: $sex \sim \text{Bernoulli}(0.5)$. For men $sex = 1$.
- Only two classes considered $class = 1$ indicating first class.
- Survival (Y) like in case study and

$$\text{logit}\left[P(Y=1|class, sex, age)\right] = 2.18 + 1.93\cdot class - 3.04\cdot sex - 0.04\cdot age$$

Titanic: The incomplete data

Generating the missing data

- 2500 datasets were generated each with 1000 passengers.
- Missing data created for age.
- The probability of age being missing depending on:
  - Class: First class less chance of missing age
  - Survival: Survivors less chance of missing age
  - Missing mechanism MAR

$$\text{logit}\,[P(r=0|class, Y)] = 2.11 - 1.5 \cdot class - 2.85 \cdot Y$$

$r = 0$ implies that age is missing.

Titanic: Simulations II

Analysis

Model:

$$\text{logit}\,[P(Y=1|class, sex, age)] = \beta_0 + \beta_1 \cdot class + \beta_2 \cdot sex + \beta_3 \cdot age$$

Type of analysis

- Inverse probability weighting (IPW)

Inverse probability weighting (IPW)

Create the new variable r with $r = 0$ when age is missing and $r = 1$ when age is observed

Passenger	survived	class	sex	age	r
1	0	0	0	NA	0
2	0	1	0	30.44	1
3	1	1	1	26.60	1
4	0	0	0	NA	0
5	1	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	
197	1	0	1	28.67	1
198	1	0	1	28.88	1
199	0	1	1	22.77	1
200	0	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	

Titanic simulation: IPW

Recall that $r = 1$ if age is observed. One can then fit the model

$$\text{logit}\left[P(r = 1|class, Y)\right] = \alpha_0 + \alpha_1 \cdot class + \alpha_2 \cdot Y$$

to get the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

Passenger	survived	class	sex	age	r
1	0	0	0	NA	0
2	0	1	0	30.44	1
3	1	1	1	26.60	1
4	0	0	0	NA	0
5	1	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	

Titanic simulation: IPW

The weight associated with subject i is

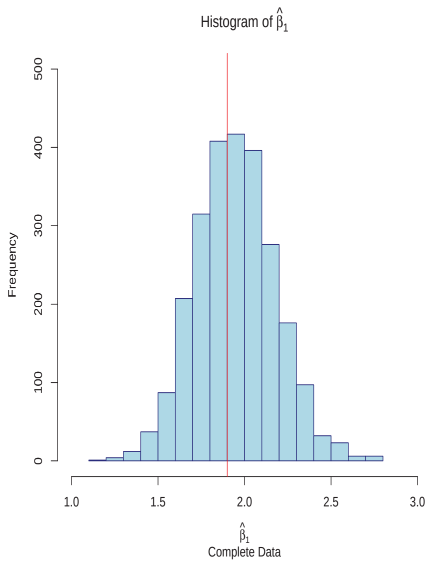
$$w_i = \frac{1}{P(r_i = 1|class_i, sex_i)} = 1 + \exp\left(1 + \hat{\alpha}_0 + \hat{\alpha}_1 \cdot class_i + \hat{\alpha}_2 \cdot Y_i\right)$$

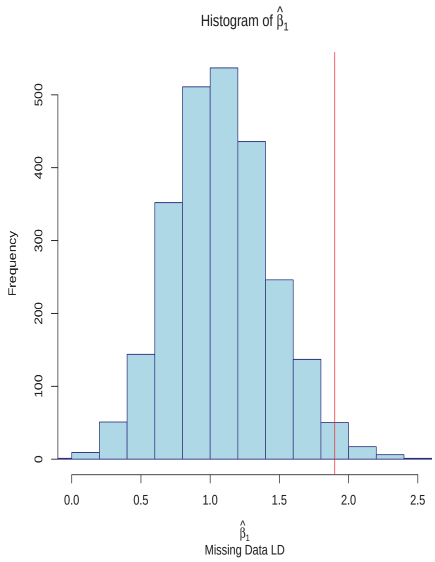
Data are again analyzed with model

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = \beta_0 + \beta_1 \cdot class + \beta_2 \cdot sex + \beta_3 \cdot age$$

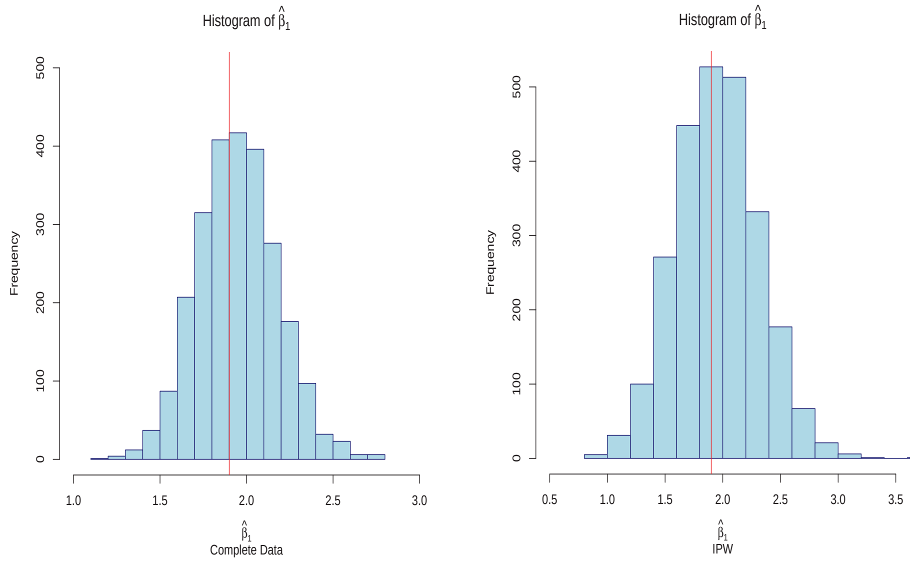
but this time using a weighted logistic regression, i.e. passing the previous w_i weights to the fitting procedure.

Titanic simulation: Complete Case Analysis





Titanic simulation: IPW results



Titanic data: IPW

Let now $r = 1$ if age and sex are observed. One can then fit the model

$$\text{logit}\left[P(r = 1|class, Y)\right] = \alpha_0 + \alpha_1 \cdot class_2 + \alpha_2 \cdot class_3 + \alpha_3 \cdot Y$$

to get the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$.

The weight associated with subject i is

$$\begin{aligned} w_i &= \frac{1}{P(r_i = 1|class_{2i}, class_{3i}, sex_i)} \\ &= 1 + \exp\left(1 + \hat{\alpha}_0 + \hat{\alpha}_1 \cdot class_{2i} + \hat{\alpha}_2 \cdot class_{3i} + \hat{\alpha}_3 \cdot Y_i\right) \end{aligned}$$

Titanic data: IPW

Let now $r = 1$ if age and sex are observed. One can then fit the model

$$\text{logit}\left[P(r = 1|class, Y)\right] = \alpha_0 + \alpha_1 \cdot class_2 + \alpha_2 \cdot class_3 + \alpha_3 \cdot Y$$

to get the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$.

Data are again analyzed with model

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot class_2 + \beta_3 \cdot class_3 + \beta_4 \cdot age$$

but this time using a weighted logistic regression, i.e. passing the previous w_i weights to the fitting procedure.

IPW in R

```
> ###### Titanic IPW
>
> ## Creating the missing data indicator variable r
>
> titanic.missing$r<-as.numeric(!is.na(titanic.missing$age))*as.numeric(!is.na(titanic.missing$sex))
> head(titanic.missing,15)
>
   survived pclass sex     age r        w
1         1    1st   0 29.0000 1 1.373464
2         0    1st   0  2.0000 1 1.526999
3         0    1st   1 30.0000 1 1.526999
4         0    1st   0 25.0000 1 1.526999
5         1    1st   1  0.9167 1 1.373464
6         1    1st   1 47.0000 1 1.373464
7         1    1st   0 63.0000 1 1.373464
8         0    1st   1 39.0000 1 1.526999
9         1    1st   0 58.0000 1 1.373464
10        0    1st   1 71.0000 1 1.526999
11        0    1st   1 47.0000 1 1.526999
12        1    1st   0 19.0000 1 1.373464
13        1    1st   0      NA 0 1.373464
14        1    1st   1      NA 0 1.373464
15        0    1st   1      NA 0 1.526999
>
```

IPW in R

```
> ## Fitting the logistic regression model to  calculate the probabilities of being complete
>
> titanic.ipw.glm<-glm(r ~ pclass + survived, data=titanic.missing,family=binomial)
> summary(titanic.ipw.glm)
>
Call:
glm(formula = r ~ pclass + survived, family = binomial, data = titanic.missing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7488  -0.7745  -0.7745   0.8119   1.6435

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6406     0.1447   4.426 9.58e-06 ***
pclass2nd     0.2999     0.1856   1.616   0.1062
pclass3rd    -1.6911     0.1559 -10.848  < 2e-16 ***
survived      0.3444     0.1377   2.502   0.0124 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1817.7  on 1312  degrees of freedom
Residual deviance: 1538.4  on 1309  degrees of freedom
AIC: 1546.4

Number of Fisher Scoring iterations: 4
>
```


IPW in R

```
> ## Calculating the weights: Inverse Probabilities
>
> titanic.missing$w<-1/fitted(titanic.ipw.glm)
> head(titanic.missing,15)
>
   survived pclass sex     age r        w
1         1    1st   0 29.0000 1 1.373464
2         0    1st   0  2.0000 1 1.526999
3         0    1st   1 30.0000 1 1.526999
4         0    1st   0 25.0000 1 1.526999
5         1    1st   1  0.9167 1 1.373464
6         1    1st   1 47.0000 1 1.373464
7         1    1st   0 63.0000 1 1.373464
8         0    1st   1 39.0000 1 1.526999
9         1    1st   0 58.0000 1 1.373464
10        0    1st   1 71.0000 1 1.526999
11        0    1st   1 47.0000 1 1.526999
12        1    1st   0 19.0000 1 1.373464
13        1    1st   0      NA 0 1.373464
14        1    1st   1      NA 0 1.373464
15        0    1st   1      NA 0 1.526999
>
```

IPW in R

```
> titanic.results.ipw<- glm(survived ~ pclass + sex + age, data=titanic.missing, weights=titanic.missing$w,
+ family=binomial)
> summary(titanic.results.ipw)
>
Call:
glm(formula = survived ~ pclass + sex + age, family = binomial,
    data = titanic.missing, weights = titanic.missing$w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3652  -0.8523  -0.5784   0.7815   4.6110

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.766252   0.322451  11.680  < 2e-16 ***
pclass2nd   -1.294939   0.220977  -5.860 4.63e-09 ***
pclass3rd   -2.720643   0.221377 -12.290  < 2e-16 ***
sex         -2.659578   0.159116 -16.715  < 2e-16 ***
age         -0.041524   0.006168  -6.732 1.67e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1687.8  on 627  degrees of freedom
Residual deviance: 1112.7  on 623  degrees of freedom
  (685 observations deleted due to missingness)
AIC: 1081.6

Number of Fisher Scoring iterations: 4
>
```


Titanic data: CC, MI and IPW

		CC		IPW	
Coefficient	Explanation	Estimate	Std. Error	Estimate	Std. Error
β_0	Intercept	4.43	0.470	3.77	0.322
β_1	sex	-3.09	0.241	-2.66	0.159
β_2	2nd	-1.47	0.282	-1.29	0.221
β_3	3rd	-2.79	0.339	-2.72	0.221
β_4	age	-0.05	0.009	-0.04	0.006
		MI			
β_0	Intercept	3.63	0.465		
β_1	sex	-2.48	0.168		
β_2	2nd	-1.31	0.248		
β_3	3rd	-2.76	0.262		
β_4	age	-0.04	0.009		