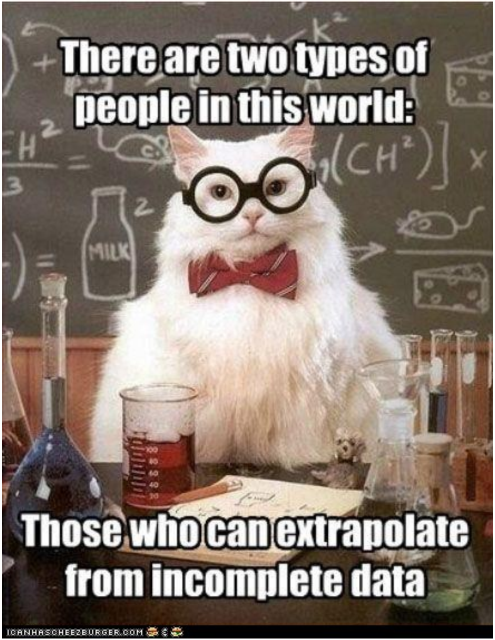
Missing Data

Author: Nicholas G Reich

This material is part of the ***statsTeachR*** *project*

Today's Lecture

- Types of missing data
- Ways to describe missing data
- Multiple imputation


There are two types of people in this world:
Those who can extrapolate from incomplete data
ICANHASCHEEZBURGER.COM

Best practices

Hard to argue with an approach that does the following:

- quantify the completeness of covariate data
- present and discuss patterns of or reasons for missing data
- provide details about your approach for handling missing data in the analysis

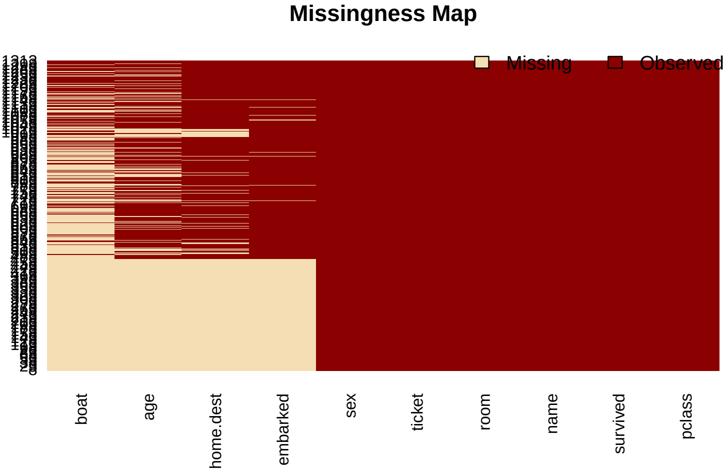
Proposed guidelines for reporting missing covariate data (Burton and Altman 2004)

Quantifying missing data

```
library(Hmisc)
getHdata(titanic)
colnames(titanic)

##  [1] "pclass"    "survived"  "name"      "age"       "embarked" 
##  [6] "home.dest" "room"      "ticket"    "boat"      "sex"      

na.pattern(titanic)

## pattern
## 0000000000 0000000010 0000010000 0000010010 0000100000 0000100010 
##        279        315          6         27          4          2 
## 0001000000 0001000010 0001010000 0001010010 0001100010 0001110010 
##         51         95          7         41          8        478 
```

Quantifying missing data

```r
library(Amelia)
missmap(titanic)
```



Quantifying missing data

What percentage of each variable's observations are missing?

```
nrow(titanic)

## [1] 1313

colSums(is.na(titanic))

##    pclass  survived      name       age  embarked home.dest      room
##         0         0         0       680       492       559         0
##    ticket      boat       sex
##         0       966         0
```

Formal Missing Data Classifications

Missing Completely at Random (MCAR)

- No data, observed or unobserved, are related to missingness.

Missing at Random (MAR)

- No unobserved data are related to missingness, but missingness may depend on observed data.

Missing Not at Random (MNAR) or unignorable missingness

- Missingness relationship cannot be simplified: it depends on unobserved data!

What kind of missingness did the titanic dataset have?

Missing Completely at Random (MCAR)

- No data, observed or unobserved, are related to missingness.

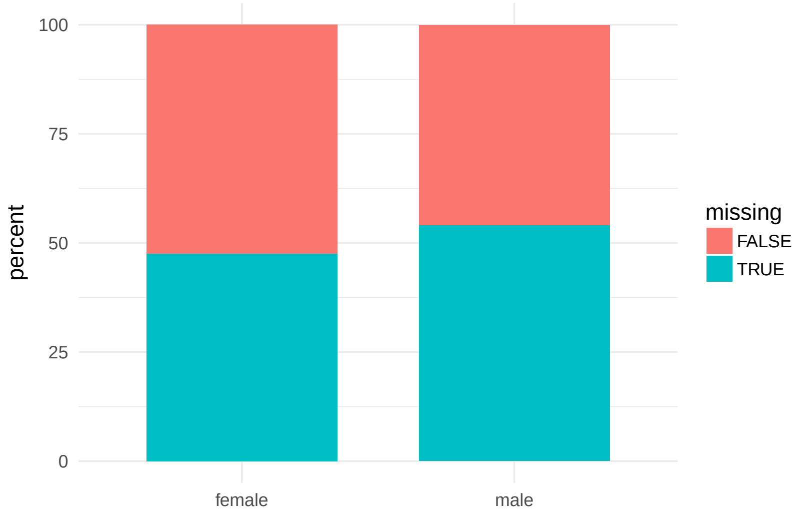
Missing at Random (MAR)

- No unobserved data are related to missingness, but missingness may depend on observed data.

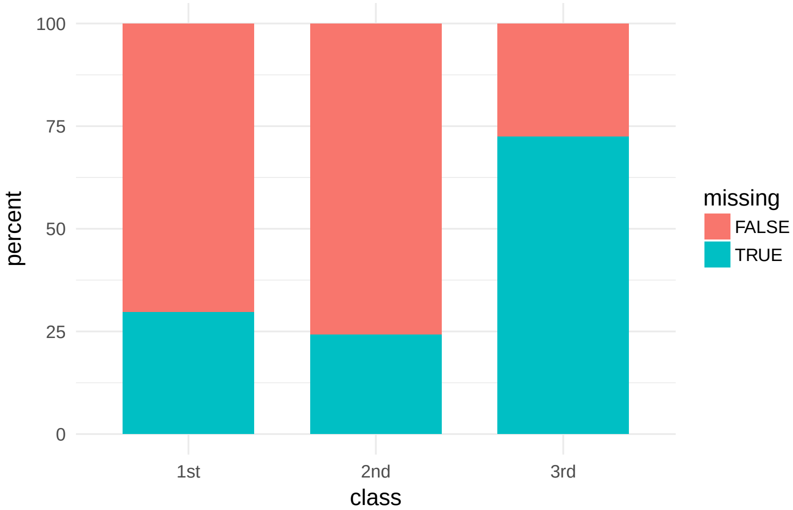
Missing Not at Random (MNAR) or unignorable missingness

- Missingness relationship cannot be simplified: it depends on unobserved data!

What kind of missingness did the titanic dataset have?



What kind of missingness did the titanic dataset have?



Example code used to create the last graphic

Harder than it should be, it felt like... Code adapted from this page.

```r
t3 <- titanic %>%
  group_by(pclass, age_mis) %>%
  summarise(count=n()) %>%
  mutate(perc=count/sum(count))

ggplot(t3, aes(x = pclass, y = perc*100, fill = age_mis)) +
  geom_bar(stat="identity", width = 0.7) +
  labs(x = "class", y = "percent", fill = "missing") +
  theme_minimal(base_size = 14)
```

Testing for the different types of data

Tests about the type of data you have

- MAR vs. MNAR: Not a definitive test here. Best option is to use your domain-specific knowledge about the data.
- MCAR vs. MAR: Little's test can weigh evidence for/against these two settings.

Little's H_0: The data is MCAR

Low p-values suggest that the data are MAR; high p-values suggest they are MCAR.

```r
test <- BaylorEdPsych::LittleMCAR(titanic[,c("pclass", "survived", "age", "sex"

## this could take a while

test$p.value

## [1] 0
```

Types of analyses for missing data

Analysis strategies (in rough order of desirability, low to high)

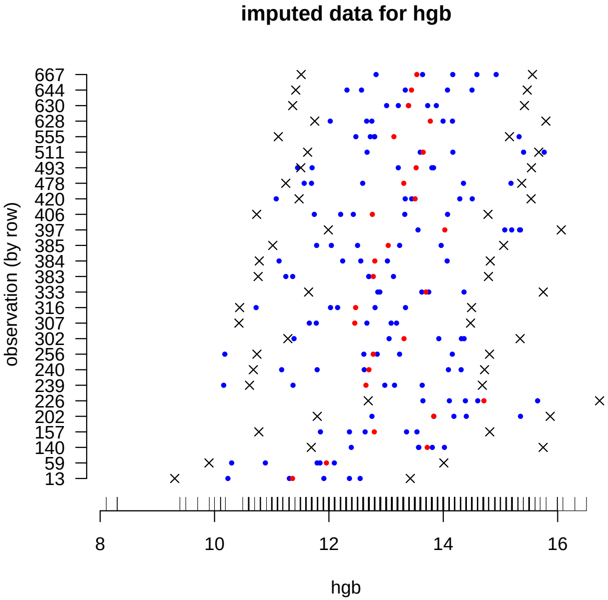
- MCAR only: Complete case a.k.a. "listwise deletion"
- Ad-hoc methods (e.g. mean imputation)
- Weighting methods
- MAR: Likelihood-based approaches (e.g. EM algorithm)
- MAR: Multiple Imputation (many flavors)
- MAR: Bayesian methods

Multiple imputation

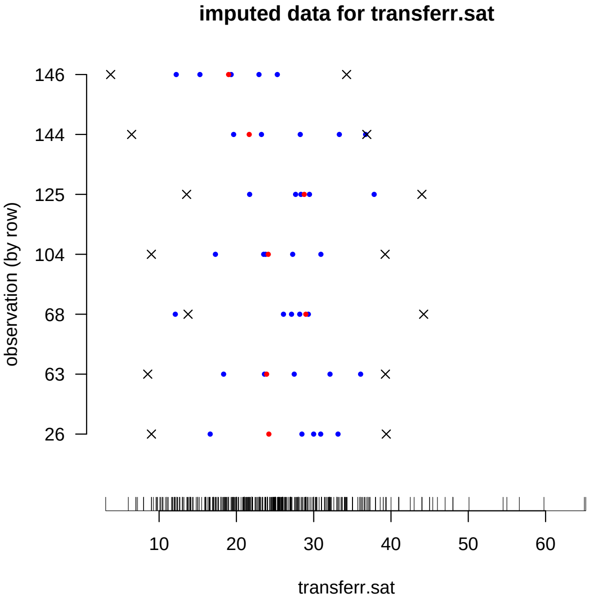
General approach

- For each missingness pattern, a model is built to use the available covariates to estimate the missing covariates.
- Random samples are taken from the predictive distribution to create multiple “complete” datasets.
- Typically, 10-15 datasets is seen as being sufficient.
- Coefficient and SE estimates are combined across datasets.

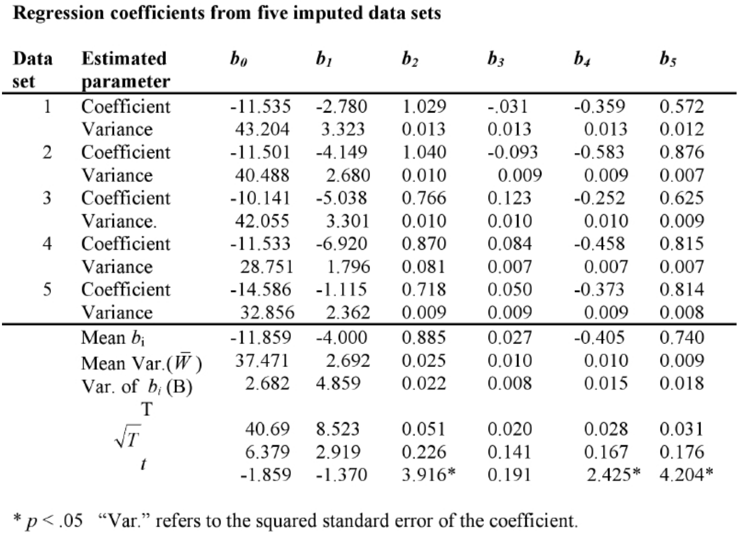
Multiple imputation: example



Multiple imputation: example



Multiple imputation results



Regression coefficients from five imputed data sets

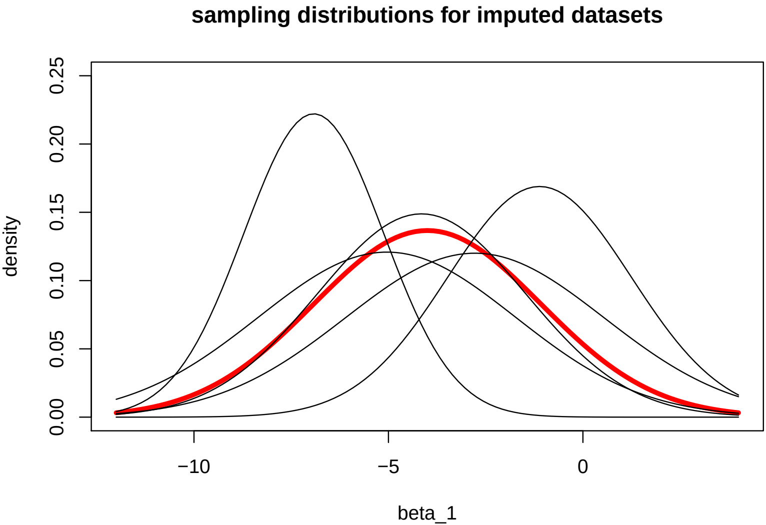
Data set	Estimated parameter	b_0	b_1	b_2	b_3	b_4	b_5
1	Coefficient	-11.535	-2.780	1.029	-.031	-0.359	0.572
	Variance	43.204	3.323	0.013	0.013	0.013	0.012
2	Coefficient	-11.501	-4.149	1.040	-0.093	-0.583	0.876
	Variance	40.488	2.680	0.010	0.009	0.009	0.007
3	Coefficient	-10.141	-5.038	0.766	0.123	-0.252	0.625
	Variance.	42.055	3.301	0.010	0.010	0.010	0.009
4	Coefficient	-11.533	-6.920	0.870	0.084	-0.458	0.815
	Variance	28.751	1.796	0.081	0.007	0.007	0.007
5	Coefficient	-14.586	-1.115	0.718	0.050	-0.373	0.814
	Variance	32.856	2.362	0.009	0.009	0.009	0.008
	Mean b_i	-11.859	-4.000	0.885	0.027	-0.405	0.740
	Mean Var.($\bar{W}$)	37.471	2.692	0.025	0.010	0.010	0.009
	Var. of b_i (B)	2.682	4.859	0.022	0.008	0.015	0.018
	T	40.69	8.523	0.051	0.020	0.028	0.031
	$\sqrt{T}$	6.379	2.919	0.226	0.141	0.167	0.176
	t	-1.859	-1.370	3.916*	0.191	2.425*	4.204*

* $p < .05$ "Var." refers to the squared standard error of the coefficient.

DC Howell, Treatment of Missing Data – Part II.

Multiple imputation results

The final estimated sampling distribution for each β is an average of the sampling distributions from each imputed dataset.



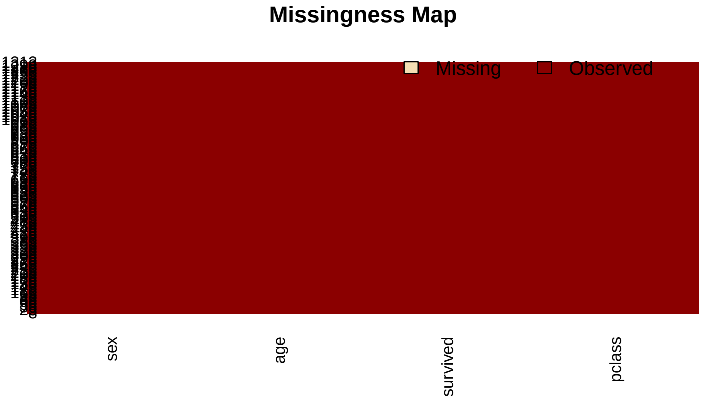
Multiple imputation software

There are two commonly used implementations of multiple imputation in R:

- MICE: http://www.stefvanbuuren.nl/mi/
- To be used together: Amelia (runs the MI) and Zelig (fits models to, among other things, MI datasets): http://gking.harvard.edu/amelia and http://zeligproject.org/

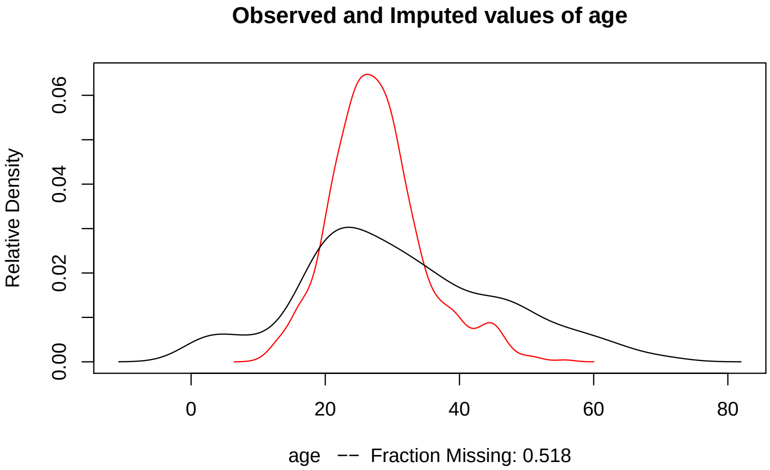
Multiple imputation for titanic data

```
t2 <- titanic[,c("pclass", "survived", "age", "sex")]
imp_titanic <- amelia(x = t2, m = 10, noms=c("sex", "pclass"))
missmap(imp_titanic$imputations$imp1)
```



Multiple imputation for titanic data

```
plot(imp_titanic, which.vars = "age")
```



Multiple imputation for titanic data

```
t2 <- t2[complete.cases(t2),] ## only include complete cases
m_full <- glm(survived~sex+age+pclass, data=t2, family=binomial)
summary(m_full)$coef
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  4.52216290 0.471007573   9.601041 7.914121e-22
## sexmale     -3.08670894 0.241062738 -12.804588 1.545447e-37
## age         -0.04930858 0.008732002  -5.646882 1.633840e-08
## pclass2nd   -1.49522913 0.281986441  -5.302486 1.142363e-07
## pclass3rd   -2.84127142 0.338897350  -8.383870 5.121522e-17
```

```
library(Zelig)
m_imp <- zelig(survived~sex+age+pclass, model="logit", data=imp_titanic)
```

```
summary(m_imp)
```

```
## Model: Combined Imputations
##             Estimate Std.Error z value  Pr(>|z|)
## (Intercept)  3.87172  0.407296   9.506 0.000e+00 ***
## sexmale     -2.50917  0.165980 -15.117 0.000e+00 ***
## age         -0.04625  0.008081  -5.723 1.048e-08 ***
## pclass2nd   -1.38023  0.234200  -5.893 3.784e-09 ***
## pclass3rd   -2.86393  0.239519 -11.957 0.000e+00 ***
## ---
```

Best practices

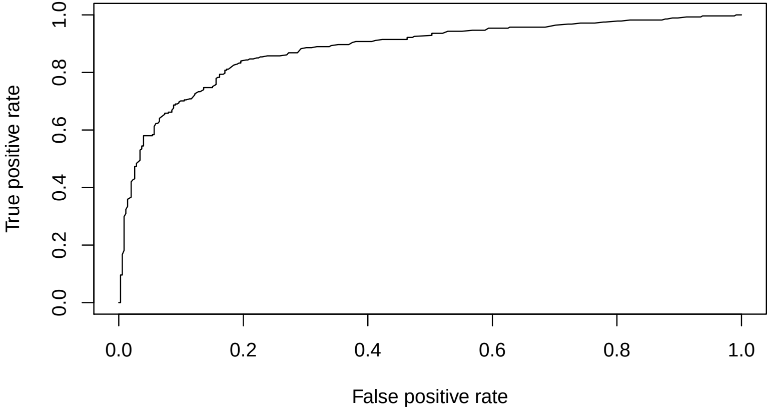
Hard to argue with an approach that does the following:

- quantify the completeness of covariate data
- present and discuss patterns of or reasons for missing data
- provide details about your approach for handling missing data

Proposed guidelines for reporting missing covariate data (Burton and Altman 2004)

Bonus: ROC for Titanic data

```
library(ROCR)
pred <- prediction(predict(m_full, type="response"), t2$survived)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf)
```


True positive rate
False positive rate