How to bootstrap the Netflix data

Art B. Owen

Department of Statistics

Stanford University

Statistics as usual

	Variable 1	$\cdots$	Variable C
Case 1			
$\vdots$			
Case R			

1) Variables are named entities:
   - E.g. pressure, volume, income $\cdots$
   - They persist
2) Cases are anonymous replicates
   - Sampled IID from some F
   - Of no inherent interest

Under statistic as usual $\cdots$

. . . we only care about cases because they show relationships among variables.

Variables by variables

Rating	Viewer 1	Viewer 2	Viewer 3	$\cdots$	Viewer C
Movie 1	4	4	1	$\cdots$	4
Movie 2	5	5	NA	$\cdots$	NA
Movie 3	3	3	NA	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Movie R	NA	5	3	$\cdots$	4

Sometimes specific rows and columns are both of persistent interest:

IPs	$\times$	books	$\longrightarrow$	purchases
terms	$\times$	documents	$\longrightarrow$	counts
candidate	$\times$	interviewer	$\longrightarrow$	rating
nodes	$\times$	more nodes	$\longrightarrow$	labeled edges

Triples

	Movie	Viewer	Rating
Case 1	1	1	4
Case 2	1	2	4
Case 3	2	1	5
⋮	⋮	⋮	⋮
Case N	R	C	4

- Now cases are anonymous
- We don't store the NAs
- 2 categorical variables
  with lots of levels
- Not independent:
  - Cases 1 & 2 share a movie
  - Cases 1 & 3 share a viewer

How should we bootstrap and cross-validate data like this?

Should we resample cases? leave out cases?

Sample reuse as usual

Cross validation

1) Pairs (X_i, Y_i) are IID from F
2) Leave out some pairs
3) Fit $\hat{Y} = f(X)$ from retained pairs
4) Predict held out Y's

Bootstrap

1) Data X_i are IID from F (unknown)
2) Estimate F by $\hat{F}$ (known)
3) Sample X_i^* from $\hat{F}$
4) X_i are to F as X_i^* are to $\hat{F}$

$\cdots$ in a nutshell

What we'd like

Bootstrap

1) For complicated methods (e.g. spectral bi-clustering) we would like to
   a) resample the data a few times,
   b) refit the model,
   c) see what is stable
2) For simple things (eg. do Harry Potter readers like Mozart?) we want to know if small effects are real

Cross-validation

1) We want to leave out some known data and predict them to simulate filling gaps in the matrix
2) We want to avoid leaving out one whole row

The answer

	Bootstrap	Cross-validation
By Elements	Wrong	OK (awkward)
By Rows and Cols	OK (approx)	Good

Random effects model

$$X_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \qquad i = 1, \ldots, R \qquad j = 1, \ldots, C$$

$a_i \sim N(0, \sigma_A^2)$ e.g. plants

$b_j \sim N(0, \sigma_B^2)$ e.g. environments

$\varepsilon_{ij} \sim N(0, \sigma_E^2)$

Used in agriculture

Studied for decades

$\hat{\mu}$ is $\bar{X}_{\bullet\bullet}$

No bootstrap exists for $V(\hat{\mu})$

None can exist $\cdots$

$\cdots$ McCullagh (2000)

We can't even bootstrap a balanced $\bar{X}$!

McCullagh (2000)

For $\hat{\mu} = \bar{X}_{\bullet\bullet} = \frac{1}{R}\frac{1}{C}\sum_{i=1}^{R}\sum_{j=1}^{C} X_{ij}$

Boot-I Resample from $N = RC$ values

Boot-II Resample R rows indep of C columns

$$V(\hat{\mu}) = \frac{\sigma_A^2}{R} + \frac{\sigma_B^2}{C} + \frac{\sigma_E^2}{RC}$$ true var

$$E(\hat{V}_I(\hat{\mu})) \doteq \left(\sigma_A^2 + \sigma_B^2 + \sigma_E^2\right)\frac{1}{RC}$$ way too small

$$E(\hat{V}_{II}(\hat{\mu})) \doteq \frac{\sigma_A^2}{R} + \frac{\sigma_B^2}{C} + \frac{3\sigma_E^2}{RC}$$ close

Boot-I is seriously flawed, Boot-II is close

Generalization (O 2007)

Recall

$$i = 1, \ldots R \quad \text{rows}$$

$$j = 1, \ldots C \quad \text{columns}$$

$$X_{ij} = \mu + a_i + b_j + \varepsilon_{ij}$$

Now allow missing values

$$Z_{ij} = \begin{cases} 1 & ij \text{ observed} \\ 0 & \text{else} \end{cases}$$

Nonnormal data

$$E(a_i) = E(b_j) = E(\varepsilon_{ij}) = 0$$

$$V(a_i) = \sigma_A^2, \quad V(b_j) = \sigma_B^2, \quad V(\varepsilon_{ij}) = \sigma_E^2$$

Sample size quantities

$$n_{i\bullet} = \sum_{j=1}^{C} Z_{ij} \qquad n_{\bullet j} = \sum_{i=1}^{R} Z_{ij}$$

$$\nu_A = \frac{1}{N}\sum_{i=1}^{R} n_{i\bullet}^2 \qquad \nu_B = \frac{1}{N}\sum_{j=1}^{C} n_{\bullet j}^2$$

$$V(\hat{\mu}_x) = \sigma_A^2 \frac{\nu_A}{N} + \sigma_B^2 \frac{\nu_B}{N} + \sigma_E^2 \frac{1}{N} \qquad \text{true var}$$

$$E(\hat{V}_I(\hat{\mu})) \doteq \left(\sigma_A^2 + \sigma_B^2 + \sigma_E^2\right)\frac{1}{N} \qquad \text{way too small}$$

$$E(\hat{V}_{II}(\hat{\mu})) \doteq \sigma_A^2 \frac{\nu_A}{N} + \sigma_B^2 \frac{\nu_B}{N} + \sigma_E^2 \frac{3}{N} \qquad \text{close enough}$$

Typically $1 \ll \nu \ll N$

Netflix: $\nu_{\text{Movies}} \doteq 56{,}200 \quad \nu_{\text{Cust}} \doteq 646 \quad N \doteq 100{,}000{,}000$

Generalizations

Non-constant variance: let

$$V(a_i) = \sigma^2_{A(i)} \quad V(b_j) = \sigma^2_{B(j)} \quad V(\varepsilon_{ij}) = \sigma^2_{E(i,j)}$$

uniformly bounded away from 0 and ∞

Then as $N \to \infty$

$$\frac{E(\hat{V}_{II}(\hat{\mu}_x)) - V(\hat{\mu}_x)}{V(\hat{\mu}_x)} = O(\epsilon_N)$$

Where

$$\epsilon_N = \max\left(\frac{1}{R}, \frac{1}{C}, \frac{1}{\nu_A}, \frac{1}{\nu_B}, \max_i \frac{n_{i\bullet}}{N}, \max_j \frac{n_{\bullet j}}{N}\right)$$

The 3 remains but gets swamped

Netflix data

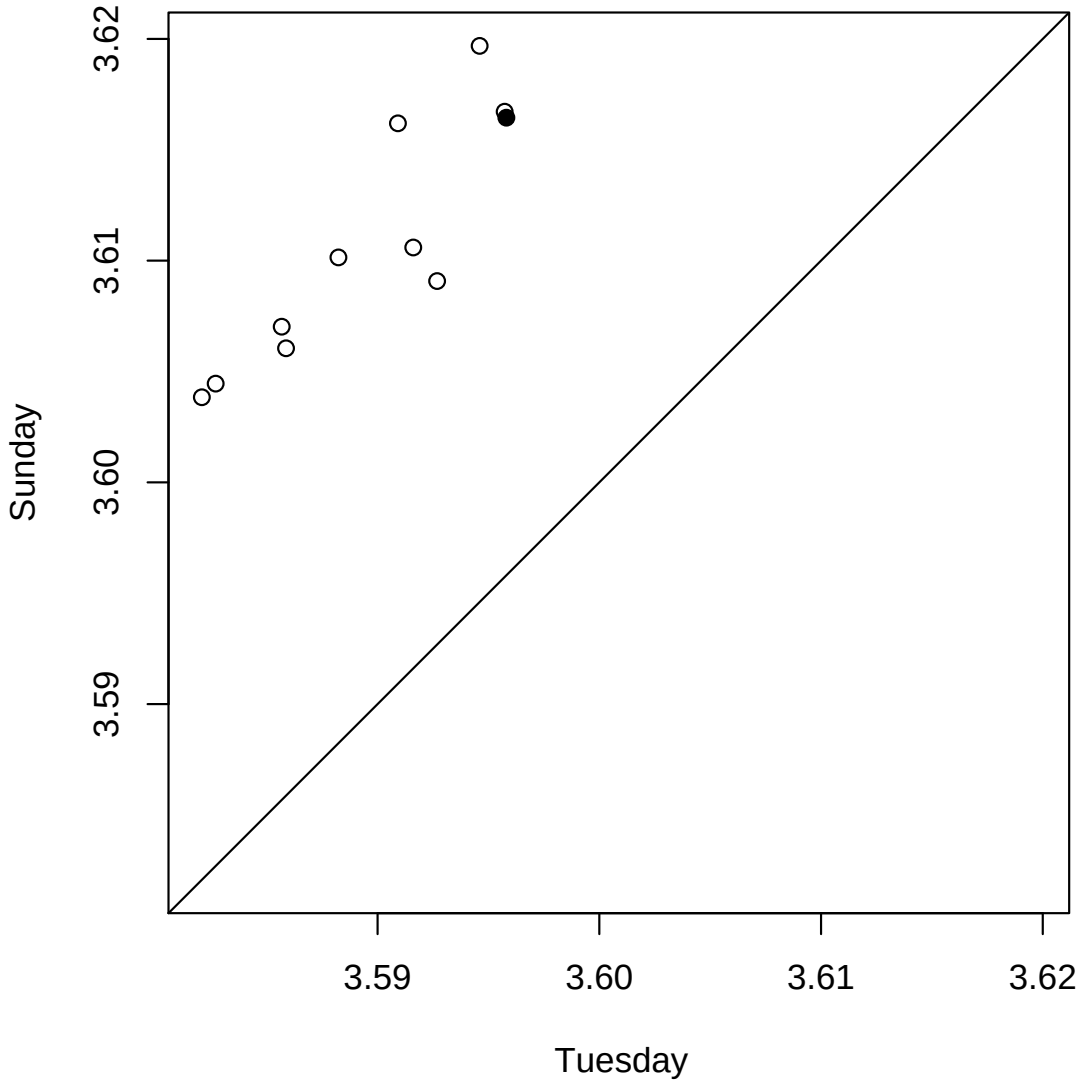
- $N = 100{,}480{,}507$ ratings $17{,}770$ movies $480{,}189$ customers
- It would be fun to look for small effects linked to customer demographics
- But those are not available for privacy reasons
- So I look at the day of the week effect
- Ratings made on Tuesdays are a tad low (average 3.596)

Advantage over Tuesday

Mon	Tue	Wed	Thu	Fri	Sat	Sun
0.002	0.000	0.009	0.008	0.010	0.019	0.021

A very small effect. Is it real? Maybe. The sample size is large.

10 bootstraps



- Sunday vs Tuesday
- 10 resamplings (open)
- Real data (solid)
- Small but real: $p < 2 \times 10^{-5}$.

Maybe:

a) Worse movies rated on Tuesday

b) or each movie does worse on Tuesday

c) or tougher customers

d) or each customer tougher on Tuesday

Place your bets!

Some details

$$Z_{ij} = 1 \iff \text{Cust i rates movie j}$$

$$Y_{ij} = \text{Rating} \in \{1, 2, 3, 4, 5\}$$

$$D_{ij}^{\text{Tue}} = 1 \iff \text{Rating on Tuesday}$$

$$\hat{\mu}_{\text{Tue}} = \frac{\sum_{ij} Z_{ij} D_{ij}^{\text{Tue}} Y_{ij}}{\sum_{ij} Z_{ij} D_{ij}^{\text{Tue}}}$$

$$\mu_{\text{Tue}} = \frac{\sum_{ij} \mathbb{E}(Z_{ij} D_{ij}^{\text{Tue}} Y_{ij})}{\sum_{ij} \mathbb{E}(Z_{ij} D_{ij}^{\text{Tue}})}$$

$$\hat{\theta} = \hat{\mu}^{\text{Tue}} - \hat{\mu}^{\text{Sun}}$$

$$\theta = \mu^{\text{Tue}} - \mu^{\text{Sun}}$$

Use Taylor expansions (delta method)

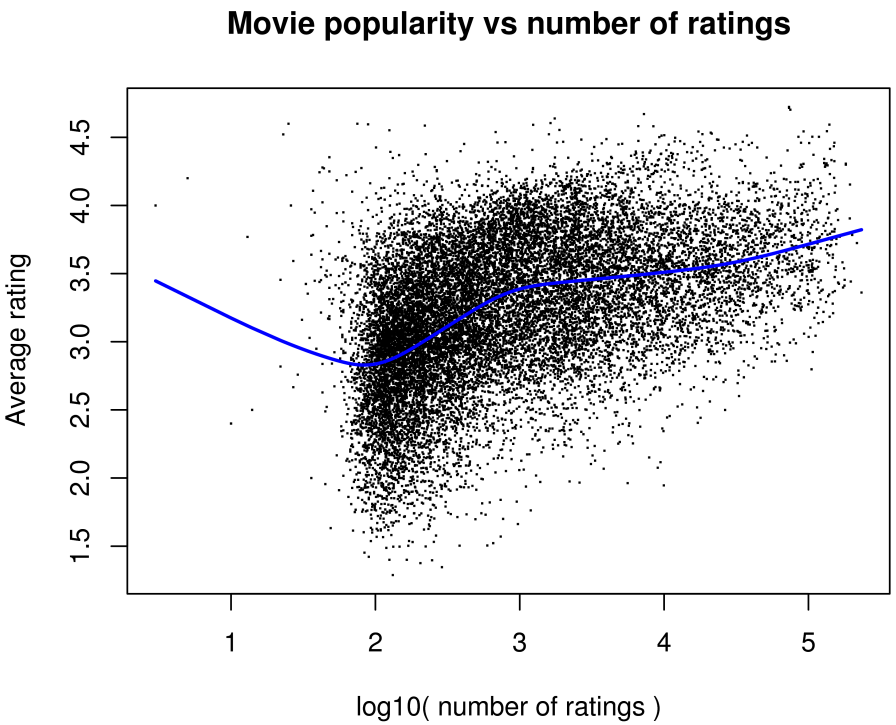
Modelling Z_{ij}

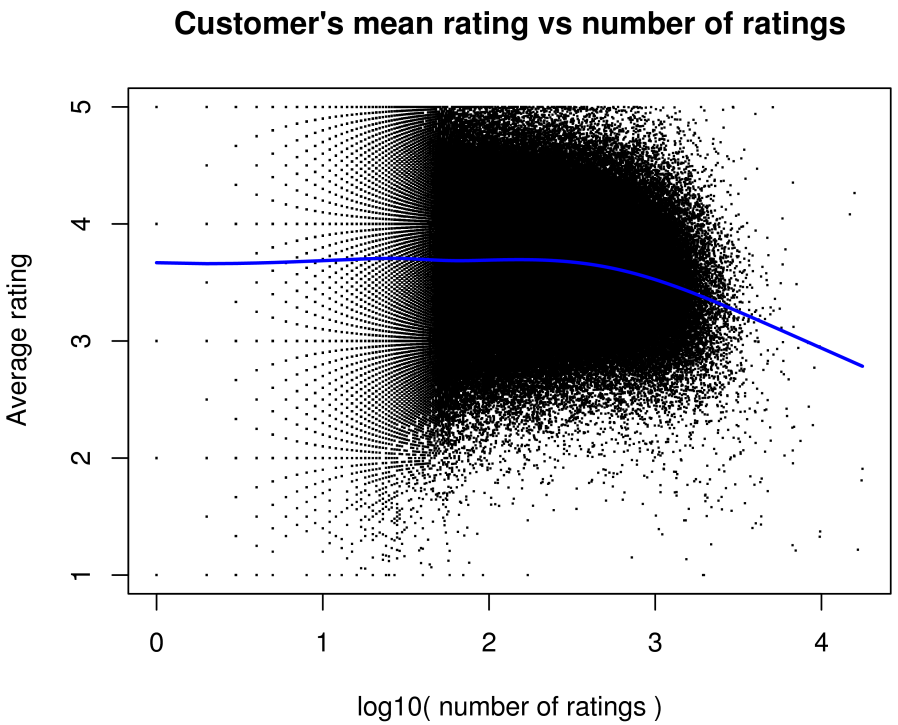
- We do not model the missingness
- Analysis is conditional on Z_{ij}
- No need to resample unobserved Y_{ij}'s

Can/should we do that?

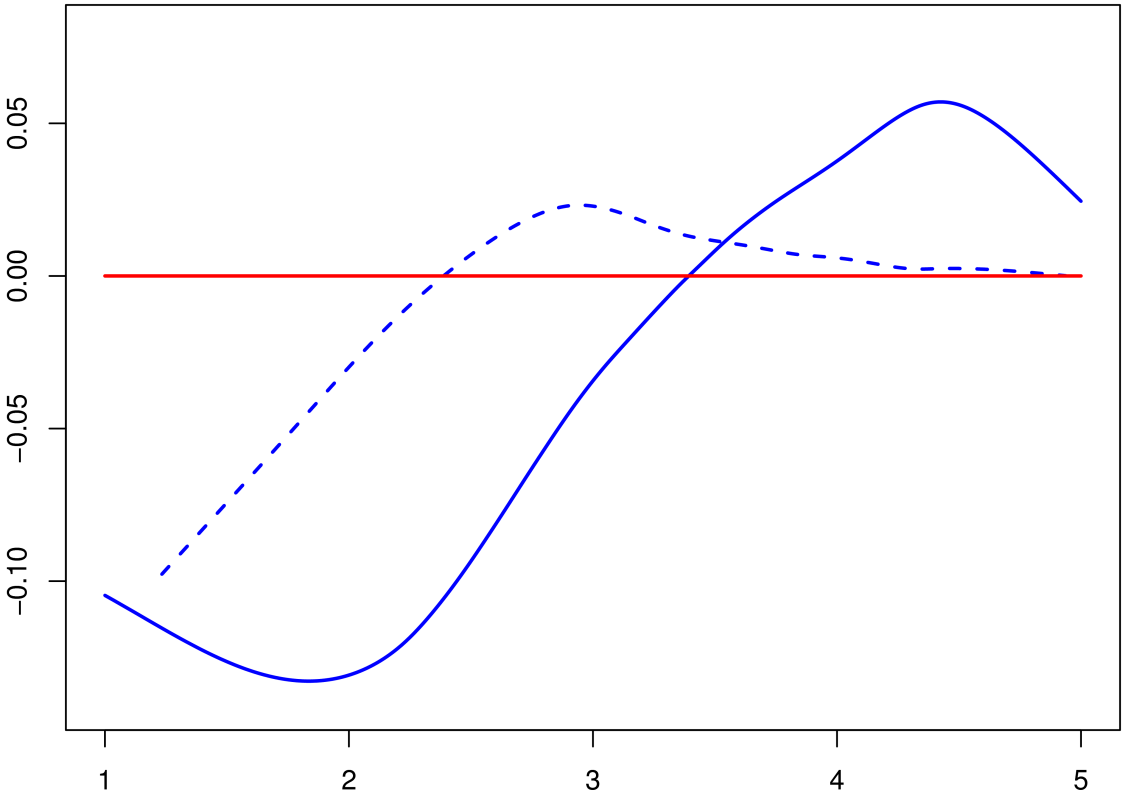
- Missingness is very important
- Less so if you're predicting ratings that were actually made
- Modelling Z_{ij} requires untestable assumptions (from outside the data)

The data


Movie popularity vs number of ratings
Average rating
1.5
2.0
2.5
3.0
3.5
4.0
4.5
1
2
3
4
5
log10(number of ratings)


Customer's mean rating vs number of ratings
Average rating
1
2
3
4
5
0
1
2
3
4
log10(number of ratings)

Sunday vs Tuesday again



- Solid curve:
  - Cust avg score on x axis
  - Cust Sun$-$Tue on vertical axis
  - Get $100{,}000$s of points
  - Smooth the points
- Dashed curve:
  - Same for movies

Bootstrap conclusion

1) Don’t resample matrix entries

2) Resample row and col entities independently