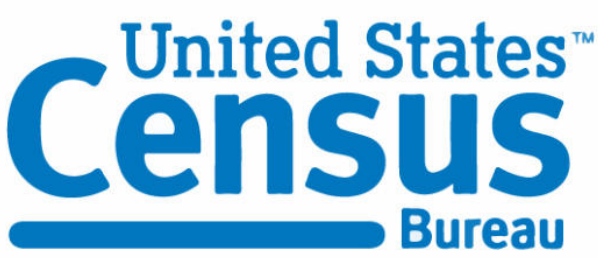
Linear Regression with Linked Data

Martin Slawski Emanuel Ben-David



The PDHP Workshop Series
University of Michigan

August 21, 2019

Disclaimer: The views in this presentation are those of the speakers, and do not necessarily represent those of the Census Bureau.

Motivation

- The US Census Bureau is a principal agency of the U.S. Federal Statistical System.
- It is responsible for producing data about the American people and economy.
- These data help researchers, businesses and government officials make informed decisions, policies and allocate funds.
- Combining and merging data form multiple sources reduce the cost of data collection, research, and the burden on responders.
- In particular, when some information about the target population is available in other existing data (surveys, administrative data).

Record linkage

- Combining two data files requires correct identification of entities on both data files (matching).
- Record linkage (RL), or entity resolution, is formally the task of pairing records from two heterogeneous data files and identifying which pairs belong to the same entity.
- Examples: linking ...
  - ... IRS data to Census data,
  - ... health registries to administrative claims,
  - ... social security or national insurance administrative databases of workers to longitudinal surveys of businesses,
  - ... the housing assistance program files to the national health interview survey.

Which records belong to the same individual?

f.name	m.name	l.name	m.o.b	lives in
Emanuel	Hyatt	Bendavid	Mar	New York, NY
Emmanuel	Ben	David	Dec	Washington, DC
Emanuel	NA	Ben-Dawid	Nov	Stanford, CA
Emanuel	NA	Ben-David	Mar	Ashland, OR
E.	NA	Ben-Davit	Nov	San Diego, CA

- Maybe the same individual.
- Variations can be due to misspelling, typographical error, faulty record system or legal name changes.

Fellegi – Sunter model for RL

- Fellegi - Sunter (FL) model is a probabilistic model for RL.
- File A has n_a records and file B has n_b records.
- Each pair of records $(a, b) \in A \times B$ need to be compared for RL.
- In general, $A \times B$ is partitioned into the sets of matches M and non-matches U.
- There are $n_a \times n_b$ record pairs whose match/non-match status have to be determined.
- Assume that each record $a \in A$ and $b \in B$ has K fields (attributes) in common, these are called the matching variables

$$Z^A = (z_1^A, \ldots, z_K^A) \quad \text{and} \quad Z^B = (z_1^B, \ldots, z_K^B).$$

File A

first_name	last_name	location	gender	income
elisha	clarke	newark	female	55118.31
lauren	block	newark	male	55246.25
bridget	alderson	chicago	female	55032.86
kyle	bullock	newark	male	55130.33
livia	broadby	newark	female	55047.96
thomas	ryan	new york	male	55220.51
bayley	clarke	phoenix	female	55367.62
ella	wilde	newark	female	55209.56
ella	reid	newark	male	55302.07
emiily	tinus	henderson	male	55091.73
nicholas	longo	newark	female	54879.95
nicholas	maynard	irvine	male	55370.54
nikki	boxer	louisville	male	54982.01
tristan	humphreys	newark	male	55444.76
lara	meaney	newark	male	55103.44
shana	sarantou	newark	female	55127.66
nicholas	coleman	newark	female	55068.10
tara	kusuma	newark	female	54882.36
alexander	reid	newark	male	55076.98
katharine	ritzau	newark	male	55225.48
nicholas	goode	newark	male	54887.72
william	mccarthy	newark	female	55262.94
blake	ho	virginia beach	male	55115.68

File B

fN	lN	city	gender	contact	age	stress_level	race	education	years_exp
noah	green	newark	female	(162) 735-9037	23	2	W	Some	28
finn	mcphail	newark	male	(817) 524-4381	31	4	M	Pro	9
damien	george	newark	male	(817) 056-5555	51	4	W	Some	15
logan	bellchambers	newark	male	(416) 062-3591	55	1	W	H	25
liam	meaney	newark	male	(228) 845-7654	53	2	As	Some	4
sarah	nurse	newark	female	(239) 653-9589	38	2	W	Pro	11
jackson	belperio	newark	male	(478) 804-6467	32	5	W	Asso	4
alexandra	dibben	new york	female	(649) 635-1853	22	1	W	None-deg	30
imogen	ryan	newark	male	(683) 306-7405	33	3	W	PhD	10
jacob	nguyen	newark	male	(943) 359-1467	29	3	W	No	9
bradley	campbell	newark	male	(253) 950-0909	36	3	B	Pro	19
judah	trevan	newark	female	(785) 804-3583	40	3	W	Pro	21
tahlia	pitt-lancaster	newark	male	(672) 105-3234	37	5	W	Nurs	7
alana	klemm	newark	male	(951) 299-2432	57	2	W	Some	30
stephanie	chandler	newark	female	(898) 200-6367	35	2	W	No	10
sarah	hobson	newark	male	(844) 367-6970	49	1	W	H	14
livia	campbell	chicago	female	(492) 677-5118	29	1	W	Nurs	27
lauren	quill	el paso	female	(379) 412-2720	36	4	W	Asso	16
liam	fey	newark	male	(412) 699-6391	26	4	W	Some	10
caitlin	binkowski	newark	female	(834) 903-3335	43	1	B	Some	5
sean	milburn	houston	female	(630) 222-6779	30	4	M	MA	8
aiden	ryan	mesa	male	(358) 826-9976	47	4	W	Pro	20

Two steps in RL

- Two main steps are
  - constructing a comparison vector $\gamma(a, b)$ for all pairs of records that are potentially a match (i.e., $(a, b) \in M$)
  - a decision rule that declares whether the pair (a, b) is a match, possible match or non-match
- When record pairs are compared on K matching variables the comparison vector $\gamma(a, b) = (\gamma_1(a, b), \ldots, \gamma_K(a, b))$.
- The k-th component $\gamma_k(a, b)$ is formed based on the level of similarity between the k-th matching variable. For example

$$\gamma_k(a, b) = \begin{cases} 0 & \text{different} \\ \vdots & \\ L_k & \text{identical} \end{cases}$$

shows how similar are z_k^A and z_k^B, for example agreement $= 1$ or disagreement $= 0$ on `Gender`.

Blocking to increase efficiency

- The blocking criteria is the minimum characteristics necessary for a pair of records to be considered a match.
- For example, the first six letters of the last name and the first letter of the first name.
- The pairs of records are then compared only if they pass the blocking criteria.
- The main goal of blocking is to reduce the number of comparisons.
- The choice of blocking criteria depends on **recall** and **precision**.
    - Recall is a measure of how many relevant records are included by the blocking scheme.
    - Precision is a measure of how many of the total records retrieved by the blocking scheme are relevant.

Linkage rules in the FL model of RL

- Given a comparison vector $\gamma = \gamma(a, b)$ we wish to designate the associated pair as a match (decision A_1), a possible match (decision A_2), or a non-match (decision A_3).
- A linkage rule is a mapping $\gamma \mapsto \{A_1, A_2, A_3\}$.
- The FL model considers the ratio $R = P(\gamma|M)/P(\gamma|U)$ and designates a pair as

  (a) a match, If $R > U$

  (b) a possible match, if $L \leq R \leq U$

  (c) a non-match, if $R < L$,

- The upper and lower cutoff thresholds U and L are determined by a priori error bounds on false matches and false non-matches.

Linkage error

- Errors in linkage occur when there is no common unique identifiers in both files (such as `SSN, EIN, HPID`).
- Such identifiers are often removed to protect privacy or confidentiality of the individuals.
- Unique identifiers can be different across multiple files. For example, education, health and tax records use different personal identifiers.
- Linkage between such data files then rely on some quasi-identifiers (matching variables), such as name, date of birth, addresses (or their combination) etc.

Some sources of linkage error

Field	Type	Examples
Names	Case	John Smith | JOHN SMITH
	Nicknames	Charles | Chuck
	Synonyms	William | Bill
	Prefixes	Dr. John Smith
	Suffixes	John Smith, III
	Punctuation	O'Malley | Smith-Taylor | Smith, Jr.
	Spaces	John Smith, Jr
	Digits	J2ohn Smith
	Initials	AM | A.M. | Anne Marie
	Transposition	John Smith | Smith John
Address	Abbreviations	RD | Road | DR | Drive
Dates	Format	01012013 | 01-01-2013 | 01JAN2013
	Invalid values	Month = 13 | Day = 32 | Birth year = 2025 | Date = 29FEB2013
Social Security Number	Format	999999999 | 999-99-9999 | 999 99 9999
Geographic locations	Abbreviations	NC | North Carolina
	FIPS codes	North Carolina = 37
	SSA codes	North Carolina = 34
	ZIP Codes	99999 | 99999-9999
	Concatenation of State and county codes	Mecklenburg County, NC | 37119
Sex	Format	Male / Female | M / F | 1 / 2

- Dusetzina SB, Tyree S, Meyer AM, et al. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. An Overview of Record Linkage Methods. Available from: https://www.ncbi.nlm.nih.gov/books/NBK253312/

Two types of linkage error

- There are two types of error associated with RL:
  - *false matches* (Type I errors): those non-matches erroneously designated as matches.
  - *false non-matches* (Type II errors): those matches erroneously designated as non-matches.
- **Does linkage error have any impact on the statistical analysis of the linked data ?**
  - What are the effect of false matches (mismatches)?
  - What are the effect of false non-matches (missed matches)?
- **How can we adjust the analysis to alleviate adverse effects of linkage error?**

  In this workshop we try to answer these questions by focusing on the case of linear regression.

Problem Statement

- Suppose we are interested a linear regression model where:
  - the response variable Y resides in file A
  - some of the predictors reside in file A and others in file B, i.e., $\mathbf{x} = (\mathbf{x}^A, \mathbf{x}^B)$.
- Linear regression:

$$y_i^A = (\mathbf{x}_i^A)^\top \beta^A + (\mathbf{x}_i^B)^\top \beta^B + \epsilon$$

- The superscript A or B indicates in which file the variable resides.

Remark: In the analysis of the linked data files we distinguish between two cases:

(a) the analyst has access to both files and carries out RL as well.

(b) the analyst has access only to the linked file.

File A: the response variable $Y = \texttt{income}$

first_name	last_name	location	gender	income
elisha	clarke	newark	female	55118.31
lauren	block	newark	male	55246.25
bridget	alderson	chicago	female	55032.86
kyle	bullock	newark	male	55130.33
livia	broadby	newark	female	55047.96
thomas	ryan	new york	male	55220.51
bayley	clarke	phoenix	female	55367.62
ella	wilde	newark	female	55209.56
ella	reid	newark	male	55302.07
emiily	tinus	henderson	male	55091.73
nicholas	longo	newark	female	54879.95
nicholas	maynard	irvine	male	55370.54
nikki	boxer	louisville	male	54982.01
tristan	humphreys	newark	male	55444.76
lara	meaney	newark	male	55103.44
shana	sarantou	newark	female	55127.66
nicholas	coleman	newark	female	55068.10
tara	kusuma	newark	female	54882.36
alexander	reid	newark	male	55076.98
katharine	ritzau	newark	male	55225.48
nicholas	goode	newark	male	54887.72
william	mccarthy	newark	female	55262.94
blake	ho	virginia beach	male	55115.68

File B contains the predictors

fN	lN	city	gender	contact	age	stress_level	race	education	years_exp
noah	green	newark	female	(162) 735-9037	23	2	W	Some	28
finn	mcphail	newark	male	(817) 524-4381	31	4	M	Pro	9
damien	george	newark	male	(817) 056-5555	51	4	W	Some	15
logan	bellchambers	newark	male	(416) 062-3591	55	1	W	H	25
liam	meaney	newark	male	(228) 845-7654	53	2	As	Some	4
sarah	nurse	newark	female	(239) 653-9589	38	2	W	Pro	11
jackson	belperio	newark	male	(478) 804-6467	32	5	W	Asso	4
alexandra	dibben	new york	female	(649) 635-1853	22	1	W	None-deg	30
imogen	ryan	newark	male	(683) 306-7405	33	3	W	PhD	10
jacob	nguyen	newark	male	(943) 359-1467	29	3	W	No	9
bradley	campbell	newark	male	(253) 950-0909	36	3	B	Pro	19
judah	trevan	newark	female	(785) 804-3583	40	3	W	Pro	21
tahlia	pitt-lancaster	newark	male	(672) 105-3234	37	5	W	Nurs	7
alana	klemm	newark	male	(951) 299-2432	57	2	W	Some	30
stephanie	chandler	newark	female	(898) 200-6367	35	2	W	No	10
sarah	hobson	newark	male	(844) 367-6970	49	1	W	H	14
livia	campbell	chicago	female	(492) 677-5118	29	1	W	Nurs	27
lauren	quill	el paso	female	(379) 412-2720	36	4	W	Asso	16
liam	fey	newark	male	(412) 699-6391	26	4	W	Some	10
caitlin	binkowski	newark	female	(834) 903-3335	43	1	B	Some	5
sean	milburn	houston	female	(630) 222-6779	30	4	M	MA	8
aiden	ryan	mesa	male	(358) 826-9976	47	4	W	Pro	20

Linkage based on common variables

File A

Matching Variables	Response	Predictors
z_1	y_1^A	x_1^A
z_2	y_2^A	x_2^A
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
	.	.
		.
		.
z_n	y_n^A	x_n^A

File B

Matching Variables	Predictors
z_1	x_1^B
z_2	
.	.
.	.
.	.
.	.
	.
	.
z_m	x_m^B

Linked data file

Response	Predictors	Predictors
y_1^A	x_1^A	x_1^B
y_2^A	x_2^A	x_2^B
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.		.
y_k^A	x_k^A	x_k^B

- Note that the size of the linked data file is $k \leq \min\{n, m\}$.

Lahiri & Larsen approach

- Lahiri & Larsen (JASA 2005) consider the linear regression of a response variable Y that resides in file A and a set of covariates $\mathbf{x}$ that reside in file B.
- The linear regression with no false matches is

$$y_i = \mathbf{x}_i^\top \beta^* + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \epsilon_i \perp\!\!\!\perp \mathbf{x}_i, \;\; i = 1, \ldots, n.$$

- Because of false matches the observed response for entity i is not necessarily y_i.
- Lahiri & Larsen model the observed responses in the linked data file as

$$\widetilde{y}_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij}, \text{ for } i \neq j. \end{cases}$$

Näive estimator vs. Lahiri & Larsen estimator

- The näive estimator of β is the least squares estimator using $\widetilde{y}_i$ instead of y_i, i.e.,

  $\widehat{\beta}_N = (X^\top X)^{-1} X^\top \widetilde{\mathbf{y}}$, where $X = (\mathbf{x}_1 \dots \mathbf{x}_n)^\top$ and $\widetilde{\mathbf{y}} = (\widetilde{y}_1 \dots \widetilde{y}_n)^\top$.

- Lahiri & Larsen define an unbiased estimator of β^* using the stochastic matrix $Q = (q_{ij})$

$$\widehat{\beta}_{LL} = \left(X^\top Q^\top Q X \right)^{-1} X^\top Q^\top \tilde{\mathbf{y}}.$$

- Because $E[\tilde{\mathbf{y}}|\mathbf{y}] = Q\mathbf{y}$, the näive estimator is biased:

$$\begin{aligned} E\left[\widehat{\beta}_N\right] &= E\left[E\left[\widehat{\beta}_N|\mathbf{y}\right]\right] \\ &= (X^\top X)^{-1} X^\top Q X \beta^* \neq \beta^*, \text{ unless } Q = I. \end{aligned}$$

Lahiri & Larsen estimator and Blocking

In a special case, the Lahiri-Larsen estimator can be shown to be equivalent to weighted "blockwise" averaging, where

- the blocks are defined by exact agreement on a set of matching variables $B_\ell = \{i : \mathbf{z}_i = v_\ell\}$, $\ell = 1, \ldots, L$:

  e.g., the matching variables $\mathbf{z}$ might be given by `ZIPCODE` and `Gender`, and the blocks correspond to observations having the same (`ZIPCODE`, `Gender`)-combination.

- we assume that within each block, matching is done uniformly at random, i.e., for observation i in block B_ℓ, i.e.,

$$\widetilde{y}_i = y_j \text{ for any } j \in B_\ell \text{ with probability } \frac{1}{|B_\ell|}.$$

  .

- Weighted block-wise averaging means linear regression based on $(\bar{\mathbf{x}}_\ell, \bar{y}_\ell)_{\ell=1}^L$ with weights $\{|B_\ell| = \#\text{obs. in block } B_\ell\}_{\ell=1}^L$.

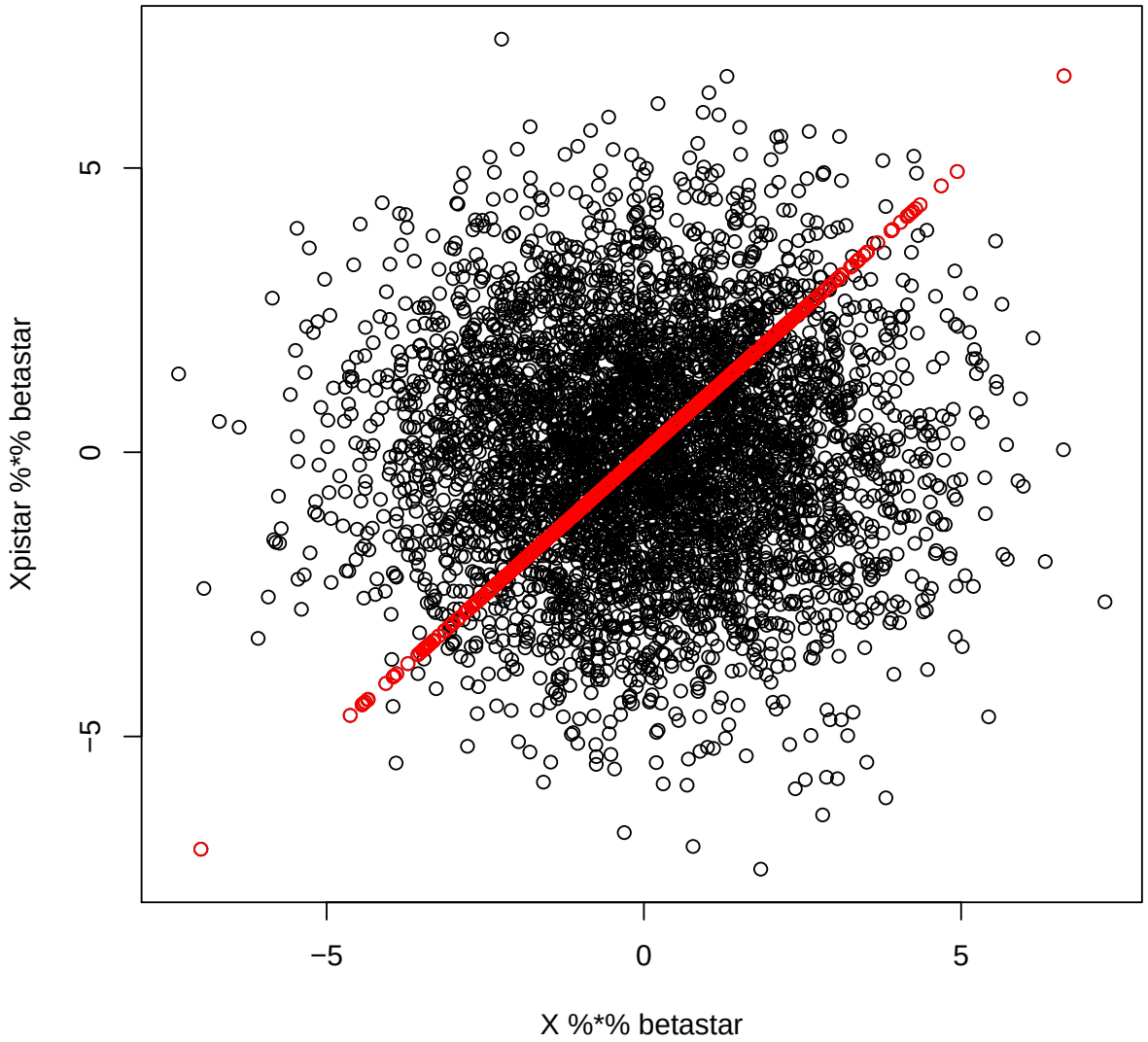
Toy example

See scripts/test_lahiri_larsen.R:

```
n <- 5000
blockindex <- sample(1:500, n, replace = TRUE)
d <- 5
X <- matrix(rnorm(n*d), nrow = n, ncol = d)
betastar <- c(1,1,-1,-1,0)
...
...
...
sigma <- 0.2
xi <- sigma * rnorm(n)
y <- Xpistar %*% betastar  + xi
```

Toy example (c'ted)

Response before and after linkage (shuffling):



Toy example (c'ted)

```
source("../code/lahiri_larsen.R")
#approach I (general Q, but slow if Q has block structure)
Q <- generate_Q_block(blockindex)
betaQ <- lahiri_larsen(X, y, Q)


#approach II (tailored to block structure, faster)
betaQcheck <- coef(lahiri_larsen_block(X, y, blockindex))

# naive least squares
coef(lm(y ~ X - 1))
     X1           X2           X3           X4           X5
 0.07205  0.11612075 -0.10753189 -0.07813942  0.01627887

# Lahiri-Larsen
betaQcheck
XbarX.1     XbarX.2     XbarX.3     XbarX.4     XbarX.5
1.0092      0.9905      -0.9975     -1.0014     -0.0076
```

Shortcomings

The Lahiri-Larsen approach requires knowledge or accurate estimates of Q. The latter requires a good understanding of the linkage process.

The approach ensures unbiasedness. However, the MSE (mean squared error) could still be substantial.

Modeling the effect of false matches as permutation

- In this approach we assume that
  - the analyst has access only to the linked file.
  - all the observed predictors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ reside in file B.
  - The linkage error results in a permutation of the observed predictors (or equivalently responses) in the linked data file (what does this assumption imply?)
  - Therefore, there is an unknown permutation π^* such that

$$y_i = \mathbf{x}_{\pi^*(i)}^\top \beta^* + \epsilon.$$

  - Objectives: estimate β^*, π^* or both.

Remark: non-linear setup

- In non-linear regression setup, the model can be expressed as

$$y_i = f^* \left(\mathbf{x}_{\pi^*(i)} \right) + \epsilon_i$$

- Objective I: learning f^* from data, but we do not know π^*.
- Objective II: leveraging the functional relationship to recover π^*.
- If f^* accurately predicts y, learning π^* should be easier.
- Note that the joint observations are $\left(y_i, \mathbf{x}_{\pi^*(i)} \right)$ and not $(y_i, \mathbf{x}_i)$.

A quick remark on the effect of false non-matches

- In the setup above we have ignored the effect of false non-matches.
- In general, the effect of false non-matches in RL is similar to that of non-response in survey sampling (selection bias and non-representativeness).
- To some degree the effect of false non-matches can be quantified, but adjusting the analysis is more difficult problem.
- In the context of non-ignorable sample selection see the recent paper

  "Measures of the Degree of Departure from Ignorable Sample Selection"

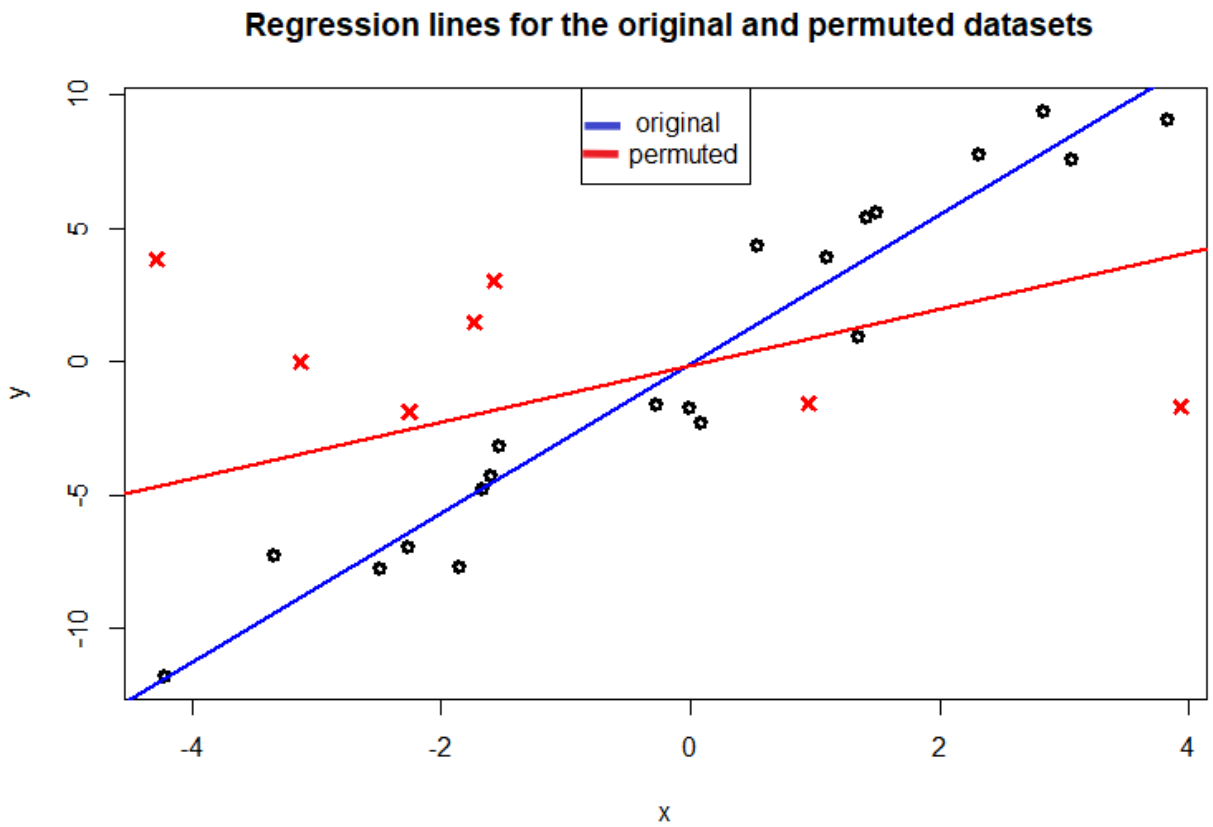
  by Phil Boonstra, Brady West, Roderick Little and Rebecca Andridge (2019).

Back to the linear regression setup

- In general, both β^* and π^* are regarded as unknown.
- This situation is described as *regression with permutation* (or *shuffling*).
- Variants of this problem in recent machine learning literature are studied under the term *unlabeled sensing*.
- Let Π^* denote the matrix representation of the permutation π^*. Then the linear model can be expressed as

$$\mathbf{y} = \Pi^* X \beta^* + \boldsymbol{\epsilon}.$$

A Toy example



- Even small number of false matches can have a significant effect on the least squares (LS) regression line.

Illustration with El Nino data set

	obs	year	month	day	date	latitude	longitude	zon.winds	mer.winds	humidity	air.temp	s.s.temp
1	4060	93	5	9	930509	-0.02	-109.96	-2.1	2.1	81.2	26.8	27.02
2	4061	93	5	10	930510	-0.02	-109.96	-3.4	1.4	84.2	26.95	26.91
3	4062	93	5	11	930511	-0.02	-109.96	-3.8	2.2	84.9	26.98	26.78
4	4063	93	5	12	930512	-0.02	-109.96	-3	1.5	86.9	26.93	26.74
5	4064	93	5	13	930513	-0.02	-109.96	-4.5	1.9	87.6	27.01	26.82
6	4065	93	5	14	930514	-0.02	-109.96	-5	1.3	85.6	26.96	26.68
7	4066	93	5	15	930515	-0.02	-109.96	-4.5	0.3	83.4	26.89	26.82
8	4067	93	5	16	930516	-0.02	-109.97	-1.9	0	82.4	26.82	27.08
9	4068	93	5	17	930517	-0.02	-109.97	-0.8	4.3	85.1	27.01	27.33
10	4069	93	5	18	930518	-0.02	-109.96	-2	5.8	85.7	27.19	27.13
11	4070	93	5	19	930519	-0.02	-109.96	-2.9	4.4	83	27.15	26.99
12	4071	93	5	20	930520	-0.02	-109.96	-3.3	3.5	83.5	27.09	26.88
13	4072	93	5	21	930521	-0.02	-109.97	-4.3	3	85.1	27.06	26.79
14	4073	93	5	22	930522	-0.02	-109.97	-4.4	2.6	86.5	27.03	26.71

- The data set contains oceanographic and surface meteorological readings ($94K$) taken from a series of buoys positioned throughout the equatorial Pacific.

See
scripts/analyze_elnino-large.R and
scripts/analyze_elnino-small.R

Illustration with El Nino data set

```
# correctly linked
elnino <- read.csv("../data/elnino-large/elnino.csv", header = TRUE)
# affected by linkage error
elnino_merged <- read.csv("../data/elnino-large/elnino-merged.csv", hea

head(elnino)

obs year month day   date latitude longitude zon.winds mer.winds humidi
1 4060    93     5   9 930509    -0.02   -109.96      -2.1       2.1
2 4061    93     5  10 930510    -0.02   -109.96      -3.4       1.4
3 4062    93     5  11 930511    -0.02   -109.96      -3.8       2.2
4 4063    93     5  12 930512    -0.02   -109.96      -3.0       1.5
5 4064    93     5  13 930513    -0.02   -109.96      -4.5       1.9
6 4065    93     5  14 930514    -0.02   -109.96      -5.0       1.3
  air.temp s.s.temp
1    26.80    27.02
2    26.95    26.91
3    26.98    26.78
4    26.93    26.74
```

Regression modeling

Attributes:

- buoy identifier
- location (latitude and longitude)
- five climate measurements (`zon`, `mer`, `humidity`, `air.temp`, `s.s.temp`)
- The regression model is

$$\begin{aligned} \texttt{air.temp} &= \beta_0 + \beta_{\texttt{zon}} \cdot \texttt{zon} + \beta_{\texttt{mer}} \cdot \texttt{mer} \\ &+ \beta_{\texttt{humidity}} \cdot \texttt{humidity} + \beta_{\texttt{s.s.temp}} \cdot \texttt{s.s.temp}. \end{aligned}$$

Regression modeling

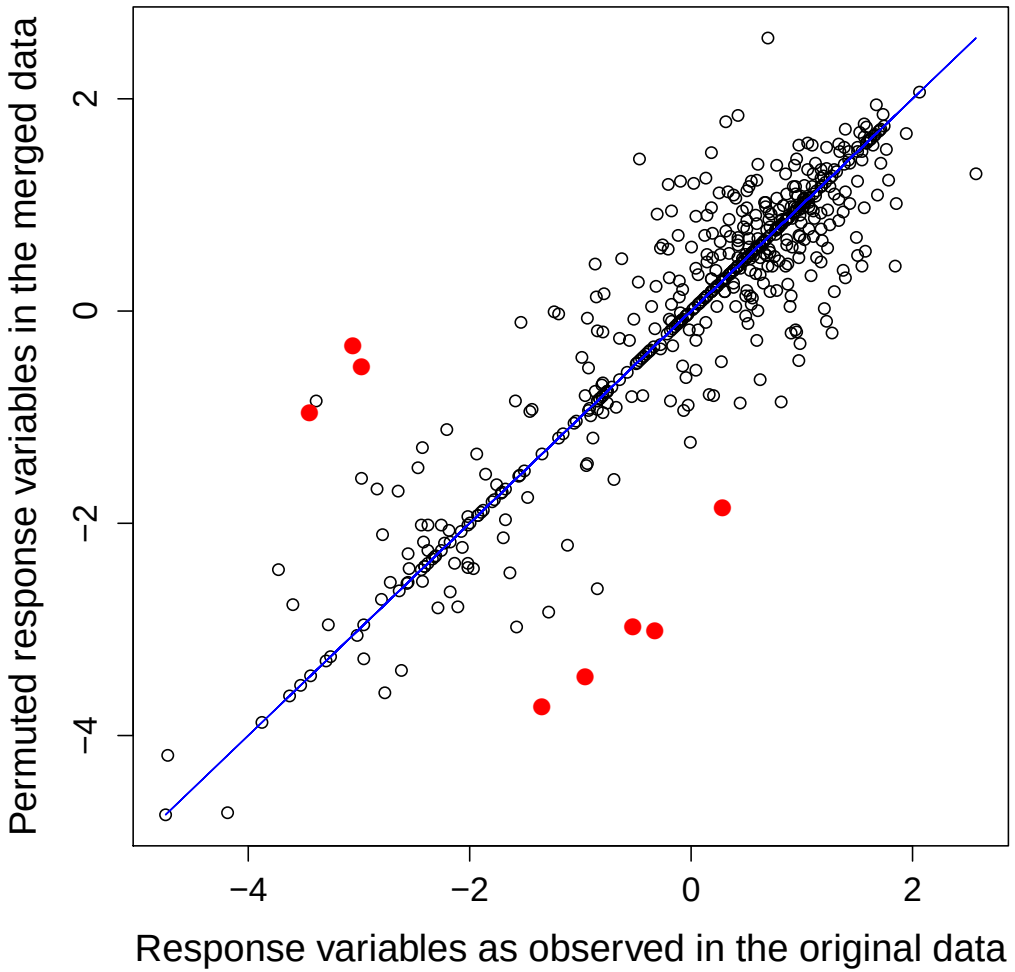
```
lm0 <- lm(air.temp ~ s.s.temp + zon.winds + mer.winds + humidity,
          data = elnino)

summary(lm0)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.1515937  0.0492926  104.51   <2e-16 ***
s.s.temp     0.8443225  0.0010995  767.90   <2e-16 ***
zon.winds   -0.0551962  0.0005494 -100.46   <2e-16 ***
mer.winds   -0.0309646  0.0005882  -52.64   <2e-16 ***
humidity    -0.0223441  0.0003415  -65.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5094 on 93930 degrees of freedom
Multiple R-squared:  0.9075,Adjusted R-squared:  0.9075
F-statistic: 2.303e+05 on 4 and 93930 DF,  p-value: < 2.2e-16
```

Experiment

- We divide the data set into two files A and B.
- A contains the response variable.
- B contains all predictor variables.
- We use (latitude, longitude) as quasi-identifiers for linking A and B (using the `fastLink` package in R).

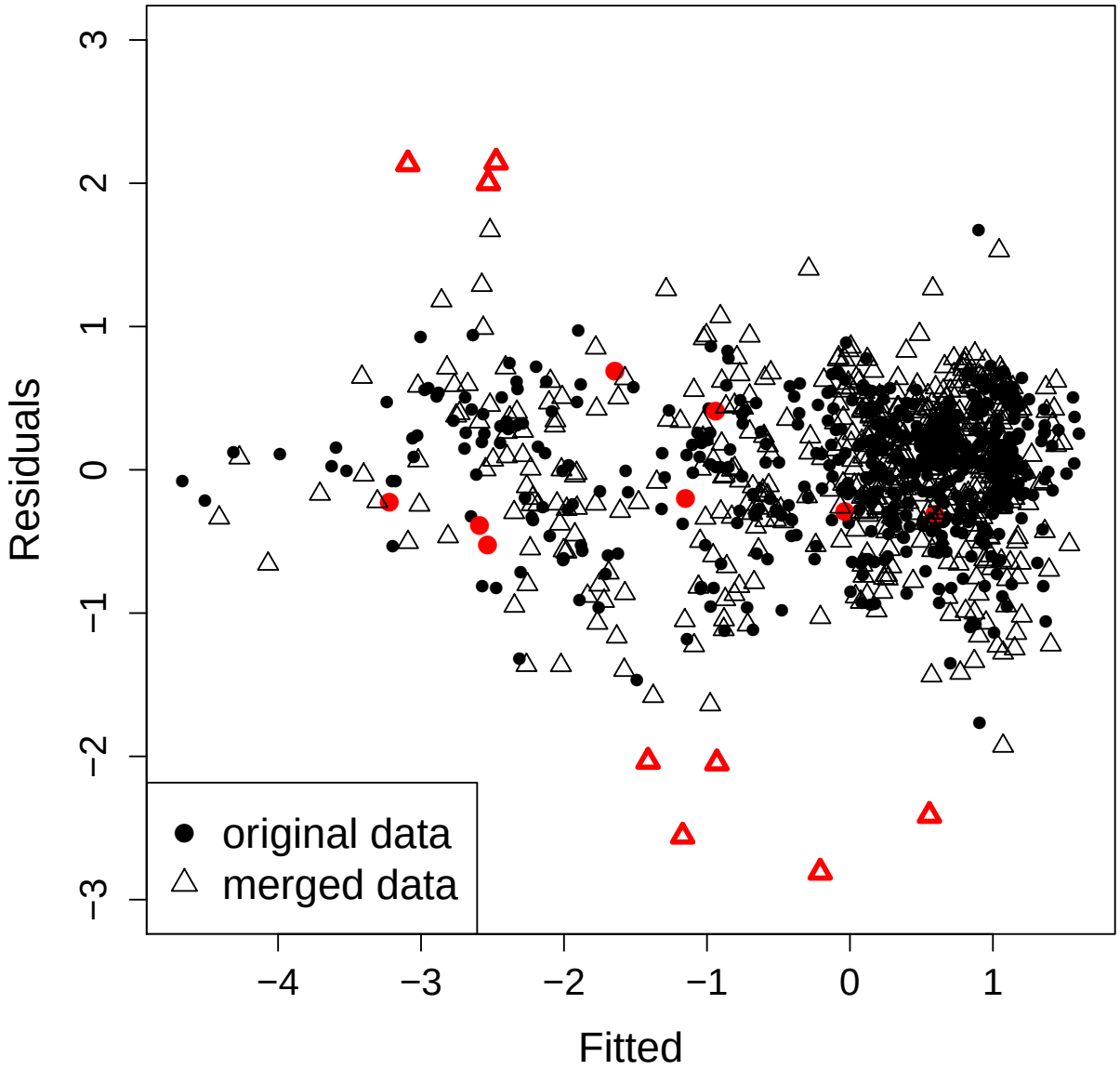
The effect of linkage error on the responses

from `scripts/analyze_elnino-small.R`:



$\mathbf{y}$ (original)	$\mathbf{y}$ (merged)
⋮	⋮
-0.96	-3.45
-3.45	-0.96
0.28	-1.86
-3.06	-0.33
-0.33	-3.02
-2.98	-0.53
-0.53	-2.98
-1.35	-3.73
⋮	⋮

Plot of regression residuals



Can we adjust the analysis?

- The plots confirm that it is not advisable to ignore linkage error.
- Goal: estimate the regression coefficients and/or permutation that are
  - computationally efficient,
  - statistically optimal (in some sense).
- First approach: Least Squares Estimation

$$\min_{\beta, \Pi \in \mathcal{P}_n} \|\mathbf{y} - \Pi X \beta\|_2^2 \qquad (\Pi\text{-LS}),$$

  where $\mathcal{P}_n$ is set of all permutation matrices of dimension n.
- Bad news: solving the problem (Π-LS) is NP-hard (a variant of QAP – see next slide).

Linear and quadratic assignment problems

- Let M and N be two n-by-n matrices.
- Two well-known combinatorial optimization problems are:

$$\text{(LAP)} \qquad \min_{\Pi \in \mathcal{P}_n} \operatorname{tr}(\Pi M)$$

$$\text{(QAP)} \qquad \min_{\Pi \in \mathcal{P}_n} \operatorname{tr}\left(\Pi M \Pi^\top N\right)$$

- LAP can be solved in polynomial time $O(n^3)$ via Hungarian or auction algorithms.
- When β^* is known (LS) problem becomes an (LAP) and thus computationally a feasible problem.

Linear assignment problem and sorting

In fact, it is easy to show that for known β^* the following optimization problems are equivalent:

$$\min_{\Pi \in \mathcal{P}_n} \|\mathbf{y} - \Pi X\beta^*\|_2^2 \tag{1}$$

$$\min_{\Pi \in \mathcal{P}_n} -\operatorname{tr}(\Pi(X\beta^*)\mathbf{y}^\top) \tag{2}$$

$$\max_{\Pi \in \mathcal{P}_n} \langle y, \Pi(X\beta^*)\rangle = \sum_{i=1}^{n} y_{(i)}(X\beta^*)_{(i)}, \tag{3}$$

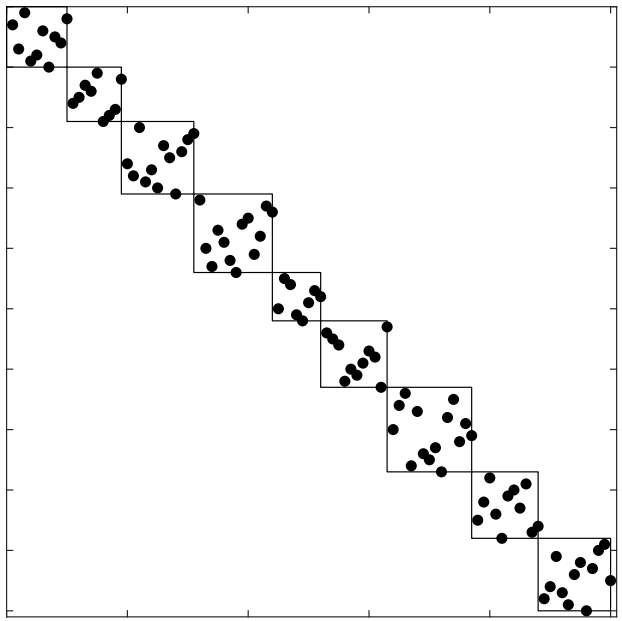
where the subscript (i) denotes the i-th order statistic, e.g.,

$$y_{(1)} < \ldots < y_{(n)}.$$

Bottom line:
the optimal Π can be found by sorting the entries of $\mathbf{y}$ and $X\beta^*$.

Block-structured permutations

Common scenario: after index re-ordering, permutation matrix Π^* has block structure, i.e., mis-matches occur only within blocks (typically defined by agreement on matching variables).



Given knowledge of the blocks, the optimal permutation can be computed block by block.

Statistical issues with the approach (Π-LS)

$$\min_{\beta, \Pi \in \mathcal{P}_n} \|\mathbf{y} - \Pi X \beta\|_2^2 \qquad (\Pi\text{-LS}),$$

Even if we could solve (Π-LS), the resulting estimators of β^* and Π^* are inconsistent: Abid et al. ('17), S. & Ben-David ('19), Hsu et al. ('17)

Consistency generally requires a lower bound on the **S**ignal-to-**N**oise **R**atio

$$\mathtt{SNR} = \frac{\|\beta^*\|_2^2}{\sigma^2}.$$

- Hsu et al. (2017) shows a minimax-type lower bound of $\mathtt{SNR} = \Omega(d/\log\log n)$ for consistent of estimation of β^*.
- Estimation of Π^* is even much harder, cf. Pananjady et al. (2018).

Toy example

See scripts/test_optimal_matching.R.

How small must σ be to recover π^*?

```
set.seed(1019)
n <- 1000
# random division into blocks
blockindex <- sample(1:200, n, replace = TRUE)
x <- rnorm(n)

# generate random permutations within each block
...
...
xpistar <- x[pistar]
sigma <- 0.001
xi <- sigma * rnorm(n)
y <- 2*xpistar + xi
###
source("../code/optimal_matching.R")
pihat <- optimal_matching(x, y, blockindex)
pihat <- optimal_matching(x, y, blockindex=rep(1,n))
sum(pihat != pistar)
```

Sparse permutations

- Without any constraints on the permutation, estimation is infeasible from both computational and statistical viewpoints.
- In the context of record linkage, it is reasonable to assume that Π^* only moves a small fraction of the labels.

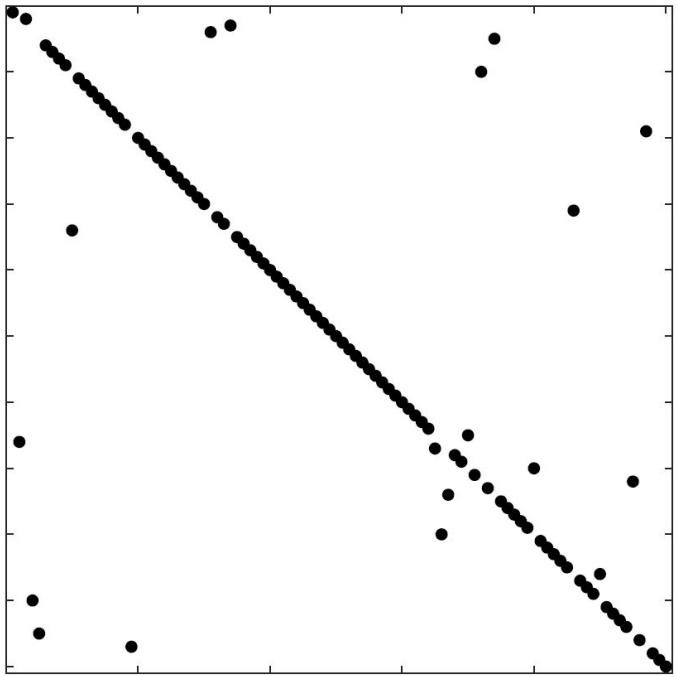


Figure: Matrix representation of a sparse permutation ($\bullet = 1$).

(LS) under sparsity constraint on π^*

- A permutation π is k-sparse if $|\{i : \pi(i) \neq i\}| \leq k$.
- Equivalently, a permutation is k-sparse if its Hamming distance from the identity

$$d_H(\pi, Id) := \sum_{i=1}^{n} \mathbf{I}(\pi(i) \neq i)$$

  is less than or equal to k.
- The set of k-sparse permutations is denoted by $\mathcal{P}_{n,k}$.
- Under the sparsity constraint, we solve

$$\left(\widehat{\beta}, \widehat{\Pi}\right) = \arg \min_{\beta, \Pi \in \mathcal{P}_{n,k}} \|\mathbf{y} - \Pi X \beta\|_2^2 \qquad \text{(C-}\Pi\text{-LS)}$$

Relaxation of constraint (LS)

- If we set $e^* = \Pi X\beta^* - X\beta^*$, then we observe that e^* is k-sparse (i.e., has at most k non-zero entries) whenever π^* is k-sparse.
- A relaxation of problem (C-Π-LS) is thus

$$\begin{aligned} & \min_{\beta,e} \|\mathbf{y} - X\beta - e\|_2^2 \\ \text{subject to} \quad & \|e\|_0 \leq k \end{aligned}$$

- A further (ℓ_1) relaxation leads to

$$\begin{aligned} & \min_{\beta,e} \|\mathbf{y} - X\beta - e\|_2^2 \\ \text{subject to} \quad & \|e\|_1 \leq c \end{aligned}$$

Two-stage estimation

1. Solve

$$\underset{\beta,e}{\operatorname{argmin}} \|\mathbf{y} - X\beta - e\|_2^2 + \lambda\|e\|_1$$

to obtain an estimate $(\widehat{\beta}, \widehat{e})$ of (β^*, e^*).

2. Plug-in estimate of Π^* given $\widehat{\beta}$,

$$\begin{aligned} \widehat{\Pi} &= \underset{\Pi \in \mathcal{P}_{n,k}}{\operatorname{argmin}} \|\mathbf{y} - \Pi\widehat{y}\|^2 \\ &= \underset{\Pi \in \mathcal{P}_{n,k}}{\operatorname{argmax}} y^\top \Pi \widehat{y}, \qquad \widehat{y} := X\widehat{\beta}. \end{aligned}$$

Remarks.

(1) Formulation 1. can be shown to be equivalent to using Huber's robust regression estimator, cf. She & Owen (JASA, 2012).

(2) The maximization in 2. requires linear integer programming. It is computationally easier to maximize over $\mathcal{P}_n$ ($\rightarrow$ sorting).

Demo

See scripts/test_robust_regression.R

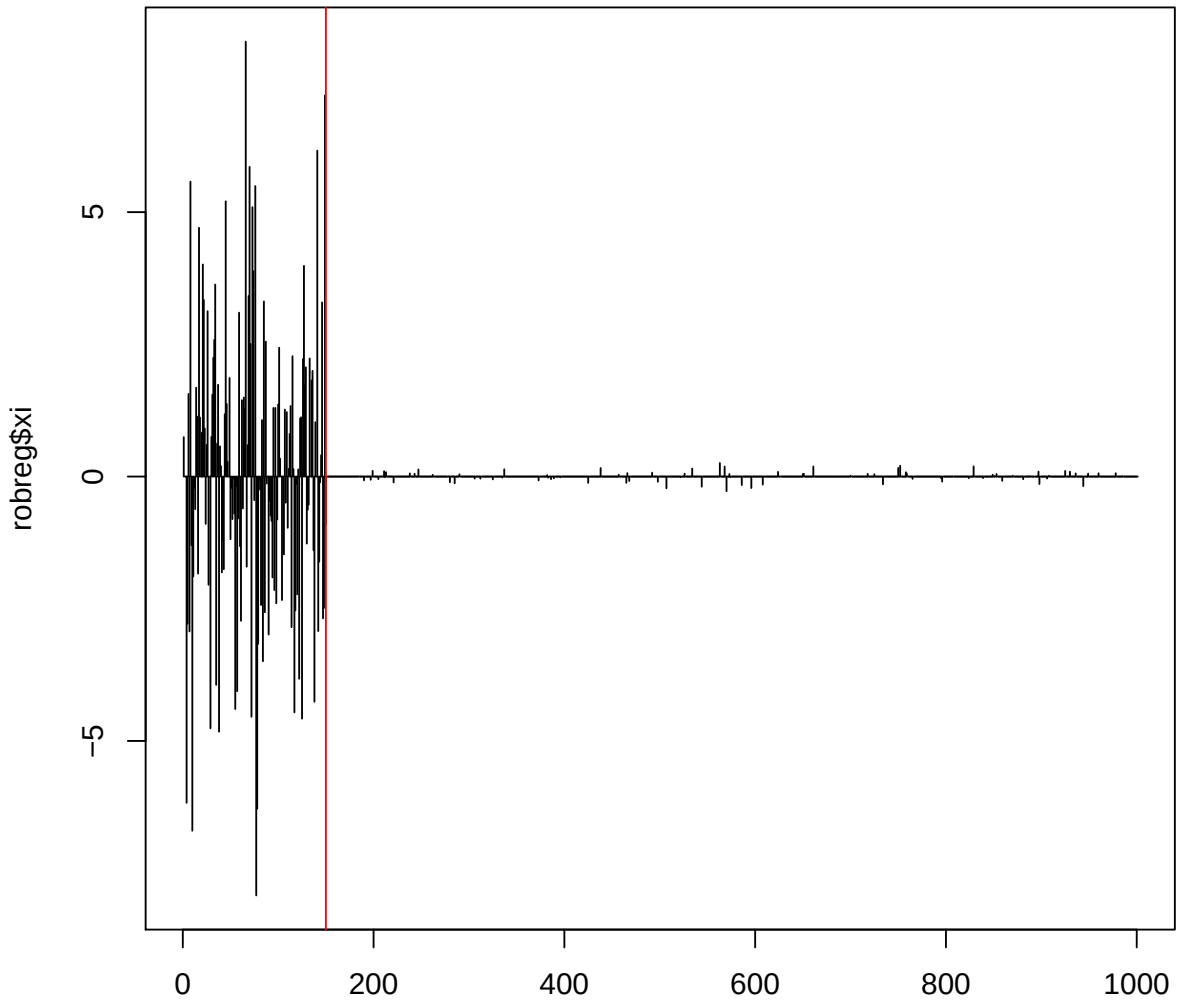
```
set.seed(1019)
n <- 1000
blockindex <- sample(1:200, n, replace = TRUE)
d <- 5
X <- matrix(rnorm(n*d), nrow = n, ncol = d)
betastar <- c(1,1,-1,-1,0)

# randomly shuffly first 15% of observations
alpha <- 0.15
pistar <- c(sample(n*0.15), (n*0.15 + 1):n)

Xpistar <- X[pistar,]
sigma <- 0.2
xi <- sigma * rnorm(n)
y <- Xpistar %*% betastar  + xi
```

Demo (c'ted)

```
source("../code/robust_regression.R")
library(MASS)

rlm0<- rlm(X, y)

robreg <- robust_regression(X, y, lambda = 1.345 * rlm0$s)

plot(robreg$xi, type = "h")
abline(v = n*alpha, col = "red")
# estimation error of robust estimator
sqrt(sum((robreg$beta - betastar)^2))
0.03976109
# estimation error of naive estimator
sqrt(sum((coef(lm(y ~ X - 1)) - betastar)^2))
0.3291098
```

Identification of mismatched data

Identification of mismatched observations via $\widehat{e}$ ($\widehat{e_i}$ vs. obs. index i):



Theoretical guarantees

- $\widehat{\beta}$ is consistent for a certain choices of λ as long as

$$k \lesssim \frac{n-p}{\log(n/k)}$$

- Given $\widehat{\beta}$, the permutation Π^* can be recovered with high probability in time $O(n \log n)$ as long as

$$\texttt{SNR} = \frac{\|\beta^*\|_2^2}{\sigma^2} \gtrsim n^5$$

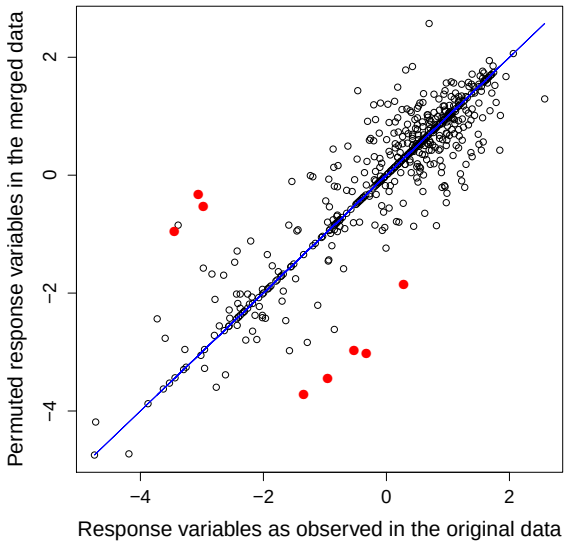
Performance of $\widehat{\beta}$ using the El Nino data set

	interc	zon	mer	humidity	s.s.temp	RMSE	ℓ_2-dist.	error
$\widehat{\beta}^{\text{oracle}}$	5.15	$-.056$	$-.031$	$-.022$	.844	.509	0	.260 $(4.6{\cdot}10^{-3})$
$\widehat{\beta}^{\text{ols}}$	6.72	$-.037$	$-.045$	$-.017$	.774	.771	1.57	.276 $(4.9{\cdot}10^{-3})$
$\widehat{\beta}$	5.74	$-.044$	$-.037$	$-.016$	.806	.773	.59	.267 $(4.8{\cdot}10^{-3})$

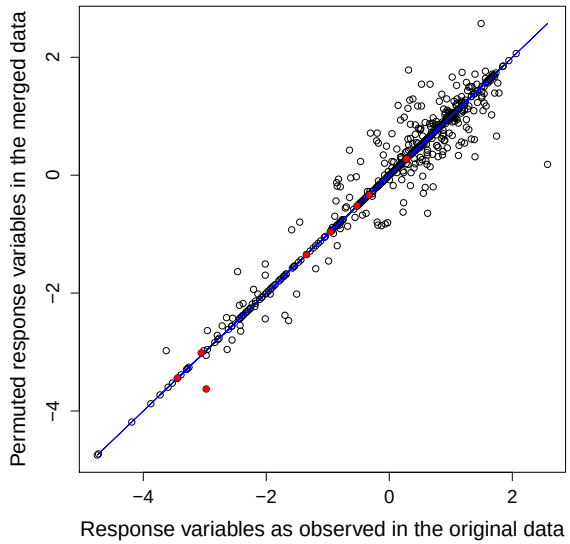
- The naive LS regression increases the residual sum of squares by 27%
- Compared to the naive estimator, our approach improves the estimate of β^* by a factor of one half in terms of ℓ_2-distance.

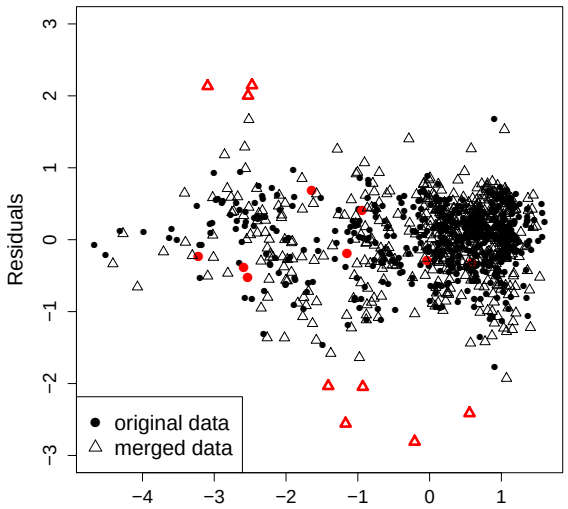
Recovering Π^*

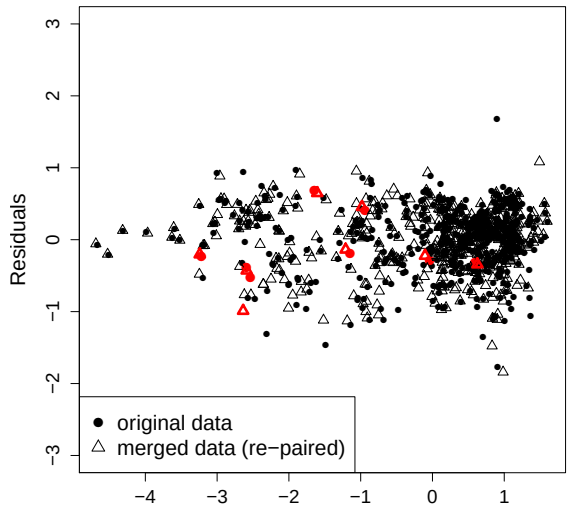
Before



After







Mixture modeling approach

- Let the binary random variable Λ_i be the match status indicator, i.e., $\Lambda_i = 1$ when the pair $(\mathbf{x}_i, y_i)$ refers to the same entity and $\Lambda_i = 0$ otherwise.
- We model

$$y_i|(\mathbf{x}_i, \Lambda_i = 1) \sim \mathcal{N}(\mathbf{x}_i^\top \beta, \sigma^2), \text{ and } y_i|(\mathbf{x}_i, \Lambda_i = 0) \sim F_y.$$

- Suppose a fraction $q = k/n$ of the linked data is mismatched.
- Therefore, the distribution of $y_i|\mathbf{x}_i$ is the mixture distribution

$$y_i|\mathbf{x}_i \sim (1-q)\mathcal{N}(\mathbf{x}_i^\top \beta, \sigma^2) + qF_y.$$

- Multiplying by the marginal density of $\mathbf{x}_i$, we can similarly model the joint density of $(\mathbf{x}_i, y_i)$ as a mixture density.

Composite likelihood method for estimation

- We can estimate the parameters in the mixture model via the method of maximum likelihood (ML).
- Note that ML estimation applies to **i.i.d** observations, but in the presence of false matches a fraction of jointly observed pairs $(\mathbf{x}_i, y_i)$ are not independent.
- Under correlated observations we resort to **composite likelihood method** (also called pseudo-likelihood method).
- Let $\theta = (\beta, \sigma^2, q)$ denote the mixture model parameters; F_y can be estimated from the empirical distribution of y.
- Assuming a classical linear regression model with Gaussian errors, denote

$$L_{i1}(\theta|\mathbf{x}, \Lambda_i = 1) = (1-q)\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i^\top \beta}{\sigma}\right)$$

Composite likelihood method

- Similarly, assuming that $F_y = N(0, \sigma_y^2)$

$$L_{i2}(\theta|\mathbf{x}_i, \Lambda_i = 0) = q\frac{1}{\sigma_y}\phi\left(\frac{y_i}{\sigma_y}\right).$$

- The Composite likelihood method estimates θ by maximizing

$$\ell(\theta) = \sum_{i=1}^{n} \log\left((1-q)\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i^\top \beta}{\sigma}\right) + q\frac{1}{\sigma_y}\phi\left(\frac{y_i}{\sigma_y}\right)\right).$$

- We use the sandwich estimator for estimating the covariance of $\hat{\theta}$, i.e.,

$$\widehat{\mathrm{Cov}}(\hat{\theta}) = \{-\ell(\theta)''\}^{-1}\left\{\sum_{i=1}^{n} \ell(\theta)_i' \ell(\theta)_i'^\top\right\}\{-\ell(\theta)''\}^{-1}|_{\theta=\hat{\theta}}$$

Some general remarks on the composite likelihood method

- We can employ the EM algorithm to maximize the log of the composite likelihood function $\ell(\theta)$.
- Under some regularity conditions the composite likelihood method guarantees that the estimators are consistent.
- The composite likelihood method may be generalized to GLM setting as well.

Demo

See scripts\test_mixture.R

```
X <- as.matrix(read.csv("../data/simulated/Xsim.csv", header = FALSE))
y <- as.matrix(read.csv("../data/simulated/ysim.csv", header = FALSE))
source("../code/mixture_model.R")

res0 <- fit_mixture0(X, y, control = list(init = "robust"))
tausq <- mean(y^2)
f0 <- function(z) dnorm(z, mean = 0, sd = sqrt(tausq))

res <- fit_mixture(X, y, f0, control = list(init = "robust"))
res
$betahat
          [,1]
V1  0.4408806
V2 -0.8655929
$sigmahat
[1] 0.5407414
$alphahat
[1] 0.8847033
```

Some references

- P. Lahiri and M. Larsen. *Regression Analysis with Linked Data* (2005).
- A. Pananjady, M. Wainwright, and T. Cortade. *Denoising Linear Models with Permuted Data* (2018).
- – *Linear Regression with Shuffled data: Statistical and Computational Limits of Permutation Recovery* (2018).
- M. Slawski and E. Ben-David. *Linear Regression with Sparsely Permuted Data*. Electronic Journal of Statistics (2019).
- – *A Two-Stage Approach to Multivariate Linear Regression with Sparsely Mismatched Data*. Pre-print (2019).
- M. Slawski, G. Diao, and E. Ben-David. *A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data*.

Acknowledgment

- We would like to acknowledge other collaborators on this ongoing research:
  - Guoqing Diao, George Mason University
  - Ping Li, Baidu Research

Thanks for your attention!