Bootstrapping divergence weighted independence graphs for design based survey analysis

July 2008

joe.whittaker@lancaster.ac.uk

Joint work with Chung-Feng Kao

- Introduction
- Outline methodology
- Examples
- Stability experiment

Aims

- Design based inference is the paradigm for analysis of survey data:
  - parameters are functions of population units,
  - variation arises from sampling lists.
- Context: exploratory analysis of survey data, need overview of joint distributions of subsets.
- We want 'graphical models' that make no appeal to super populations, probability modelling or likelihood.
- Use population measures of independence strength, $G = (V, E) \rightarrow (V, E, W)$.

Design based survey analysis

Cochran (1977), finite population.

- Theoretical paradigm is to estimate parameters

$$Q = \sum_{t \in \mathcal{P}} y^t$$

defined on the population $\mathcal{P} = \{t | t = 1, \ldots, \mathrm{N}\}$,

from the sample $\mathcal{S}$

$$\widehat{Q} = \sum_{t \in \mathcal{S}} y^t.$$

- Inference: repeated sampling of $\mathcal{S}$ from $\mathcal{P}$.

Population proportions

k survey variables $(y_1, y_2, \ldots, y_k)$.

• Population proportion in r-th category of i-th vble

$$\phi_i(r) = \sum_{t=1}^{N} I_{\{y_i^t = r\}} \, N^{-1}$$

where I is indicator function.

Suppose y_i discrete in ordered (wlog) set $r = 0, 1, \ldots, M_i - 1$,

discretise innately continuous variables,

assign integers to categorical variables.

Proportions as expectations

Easier to express measures as expectations of rvs under SRS with replacement from population:

Y_i takes values $\{y_i^1, \ldots, y_i^N\}$ with prob N^{-1}.

- Population proportions may be written

$$\phi_i(r) = \mathsf{E}_{\mathcal{P}} I_{\{Y_i = r\}}$$

Bivariate proportions $\phi_{ij}(r, s) = \mathsf{E}_{\mathcal{P}} I_{\{Y_i = r \cap Y_j = s\}}$

extends to conditional distributions,

higher dimensions.

Shannon entropy $-\mathsf{E}_{\mathcal{P}} \log \phi(Y)$

measures departure of ϕ from uniformity.

Divergence against independence

- $\mathsf{Inf}_{\mathcal{P}}(Y_i \perp\!\!\!\perp Y_j) = \mathsf{E}_{\mathcal{P}} \log \dfrac{\phi_{ij}(Y_i, Y_j)}{\phi_i(Y_i)\, \phi_j(Y_j)}.$

Double sum over the population.

Inf measures how nearly $\phi_{ij}(r_i, r_j)$ factorises into product when averaged over the population.

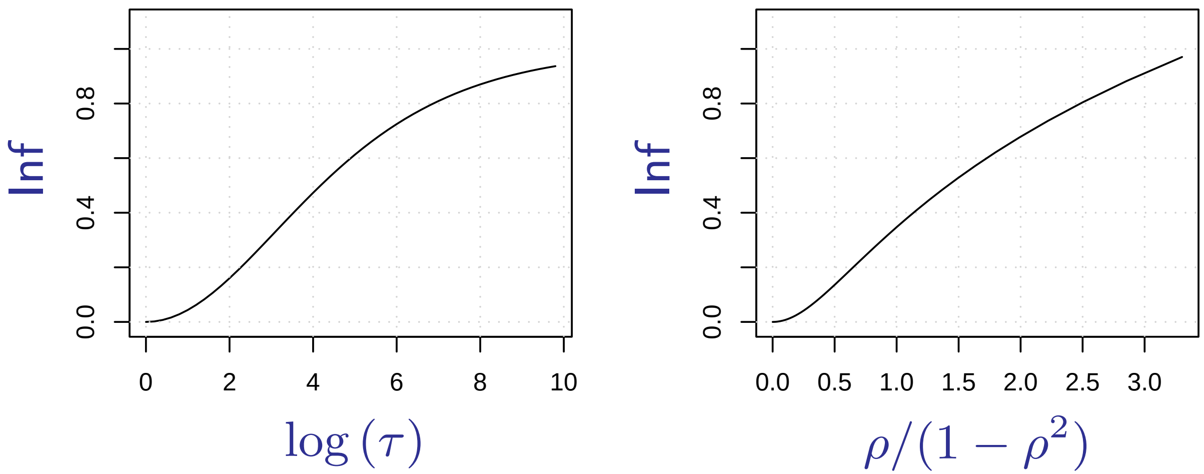
Well known properties.

Conditional independence: generalise to three or more dimensions.

- $\mathsf{Inf}_{\mathcal{P}}(Y_i \perp\!\!\!\perp Y_j | Y_k) = \mathsf{E}_{\mathcal{P}} \log \dfrac{\phi_{ij|k}(Y_i, Y_j | Y_k)}{\phi_{i|k}(Y_i | Y_k)\, \phi_{j|k}(Y_j | Y_k)}.$

Magnitude of Inf

- Compare to cpr and to correlation coefficient:
  - two binary rvs, equi-probable margins, cpr τ,
  - two standard Normal rvs, corr ρ,
    $\left(\text{Inf}(X \perp Y) = -\frac{1}{2}\log(1-\rho^2)\right)$.



Inf measured in bits using $\log_2$.

Divergence weighted independence graph

- The DWIG is the graph (V, E, W) with edge weights

$$w_{ij} = \text{Inf}_{\mathcal{P}}(Y_i \perp\!\!\!\perp Y_j | Y_{\setminus ij}),$$

where $_{\setminus ij}$ indicates the remaining vbles in set.

w_{ij} is the extra information for predicting Y_i provided by Y_j when the rest have been taken into account.

- Graph is complete, all edges appear,

natural display sets edge width and tone $\propto w_{ij}$.

Toy example: 4 binary variables

Proportions linearly increase over the 16 categories in standard order:

$0.7, 1.5, 2.2, 2.9, 3.7, 4.4, 5.1, 5.9, 6.6, 7.4, 8.1, 8.8, 9.6, 10.3, 11.0, 11.8$%

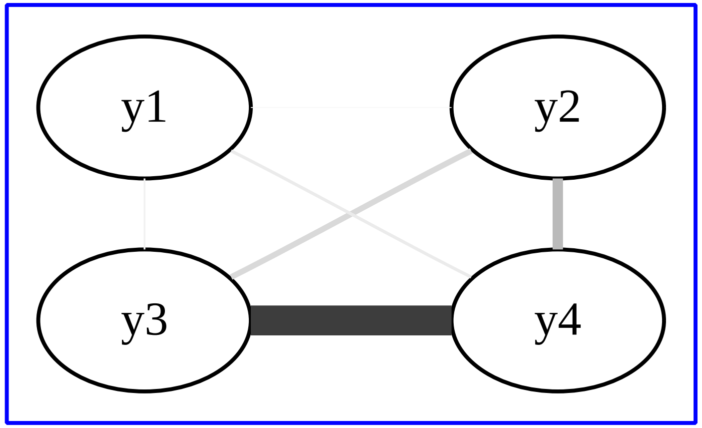
Divergences

	Y_1	Y_2	Y_3	Y_4
Y_1	0.00	0.453	0.936	1.584
Y_2	0.453	0.00	2.910	5.296
Y_3	0.936	2.910	0.00	15.161
Y_4	1.584	5.296	15.161	0.00

symmetric, positive, diagonal values zero.

Largest 15.161 mbits,

set max edge width/tone to 20 mbits.



set max 20 mbits
actual 15.1611

Relative strengths now apparent,

Y_4 interacts most strongly, most predictable,

Y_1 is the least predictable.

Sensitive to the setting of thickest width.

Remarks

A DWIG gives an overview of a joint distribution.

- Informative as to
  - conditional associations
    strengths, symmetries, structure;
  - prediction and approximate separation:
    which are the key predictors;
  - approximations:
    thresholding a DWIG gives an UG.
- Weights based on alternative CI statements to describe joint distribution possible:
  Bayes nets, chain graphs,

The entropy estimate of the divergence

In design based inference a parameter is a functional of the finite population, and estimated using the same recipe on the sample.

- Divergence against independence

$$
\begin{aligned}
\widehat{w}_{ij} &= \mathsf{E}_{\mathcal{S}} \log \frac{\widehat{\phi}_{ij}(Y_i, Y_j)}{\widehat{\phi}_i(Y_i)\,\widehat{\phi}_j(Y_j)} \qquad \text{from def of Inf} \\
&= \mathsf{E}_{\mathcal{S}} \log \widehat{\phi}_{ij}(Y_i, Y_j) - \mathsf{E}_{\mathcal{S}} \log \widehat{\phi}_i(Y_i) - \mathsf{E}_{\mathcal{S}} \log \widehat{\phi}_j(Y_j)
\end{aligned}
$$

linear comb of entropies of sample proportions.

Easily generalises to cond indep.

Young women smoking: GHS data

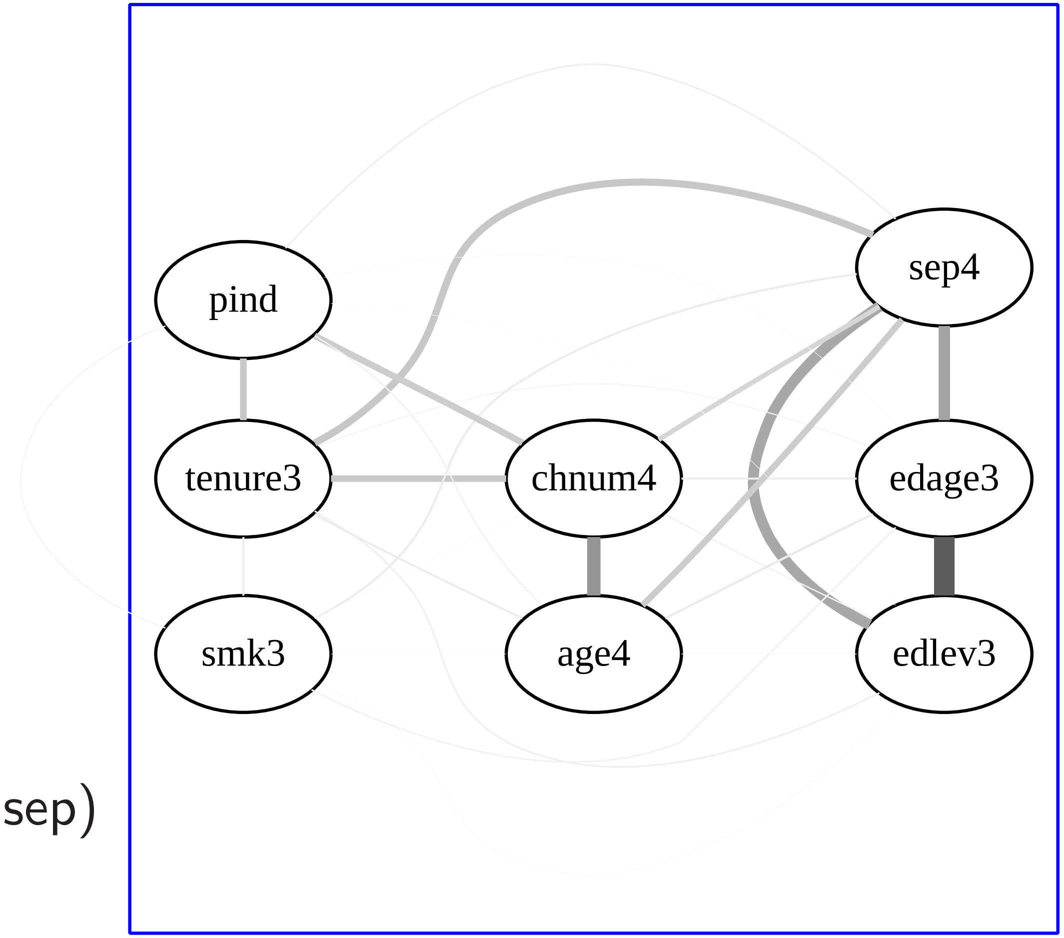
http://www.data-archive.ac.uk

Data from GH surveys: 01/02, 02/03, 03/04,

- 4651 young women aged 20-34, 8 variables.

DWIG for young women smoking: all years

smk3		
curr	quit	never
1598	659	2394



set max 200 mbits
actual 126.4376

nn(smk3) = (pind,tenure,edage,sep)

strongest edge: (edlev,edage)

Munich rent data

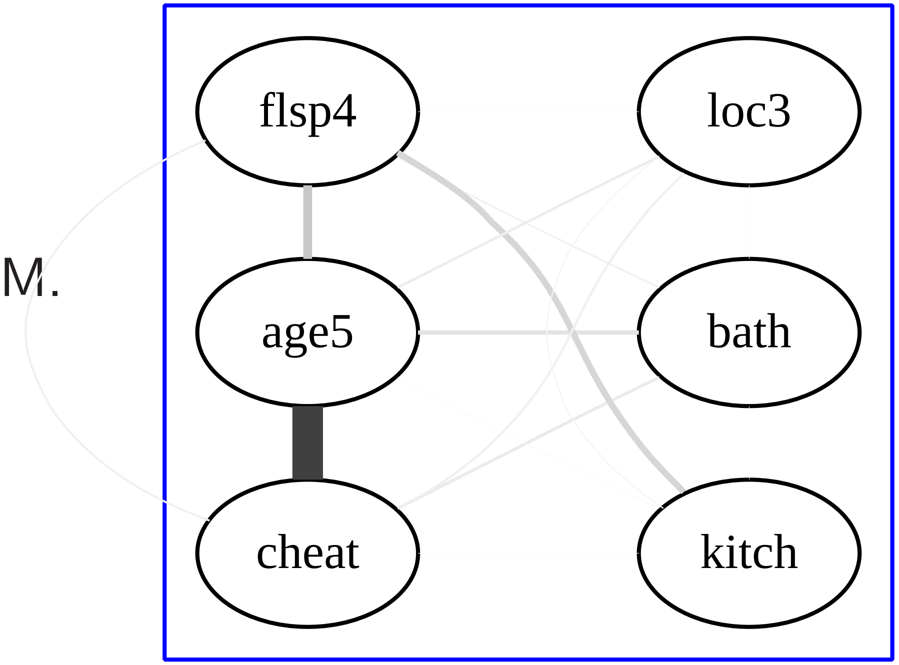
Stasinopoulos and Rigby (2000)

Rent survey, April 1993.

Response variable, monthly net rent, DM.

DWIG for explanatory variables

- 1969 obs, 6 variables.



set max 150 mbits
actual 112.2773

Conclude:

one strong edge, 112 mbits is small,

two other edges relatively strong,

five others visible, seven others invisible.

→ Divergence: how is it estimated? is it stable?

Employee satisfaction data 2001

Source: http://stars.ac.uk

Medical sales force, about 10k in UK.

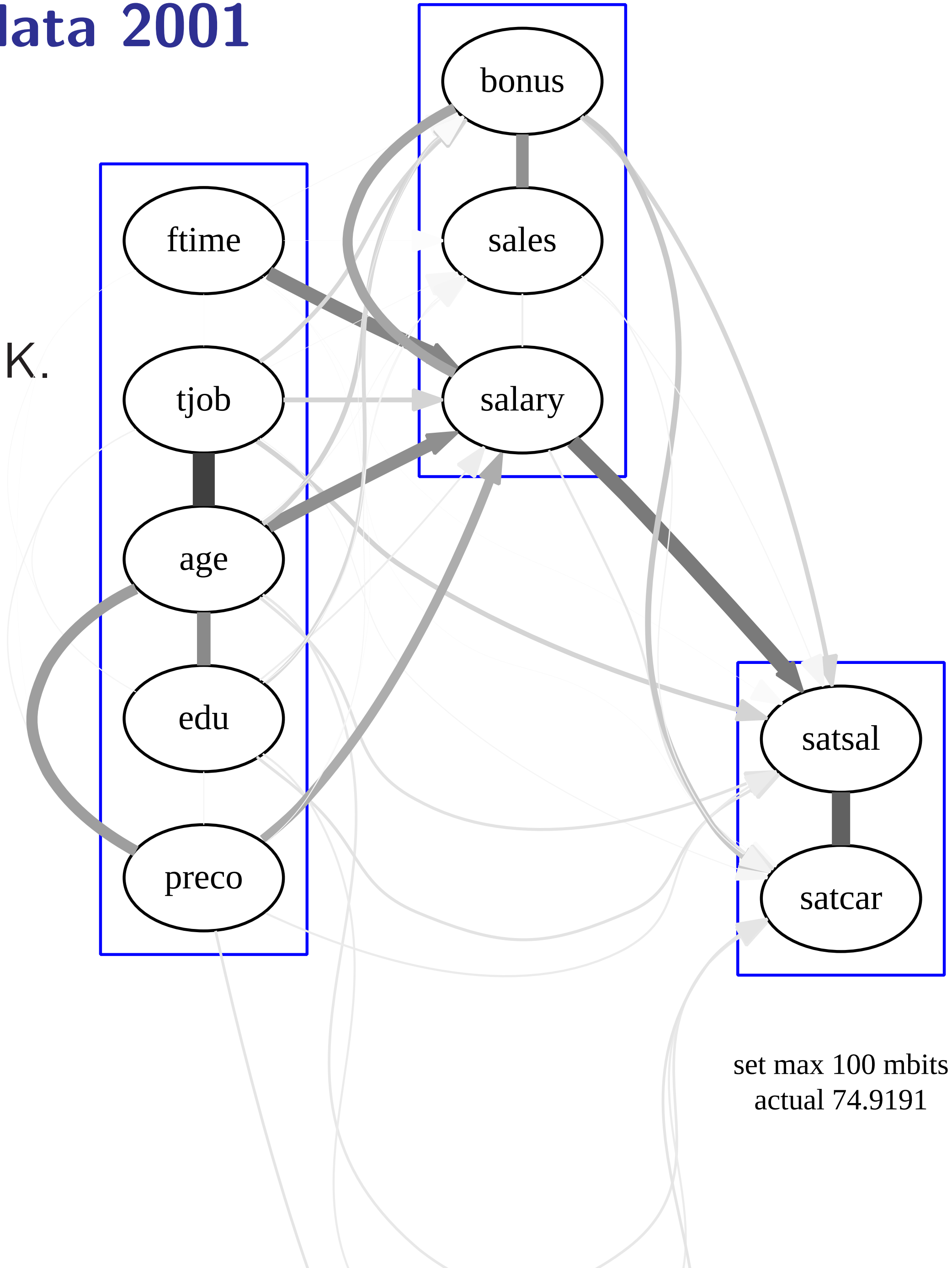
Survey: 9k questionnaires,

20% response →1758,

remove missing values →

- 1272 observations, 16 variables.

Chain graph on 3 blocks.



Deviance approximation to the divergence

Numerically can evaluate w_{ij} from logistic regession deviance.

- Rewrite

$$
\begin{aligned}
\widehat{w}_{ij} &= \mathsf{E}_{\mathcal{S}} \log \widehat{\phi}_{i|\backslash i}(Y_i|Y_j, Y_{\backslash ij}) - \mathsf{E}_{\mathcal{S}} \log \widehat{\phi}_{i|\backslash ij}(Y_i|Y_{\backslash ij}) \\
&= \max_{\theta} \mathsf{E}_{\mathcal{S}} \log \phi_{i|\backslash i}(Y_i|Y_{\backslash i}, \theta_{\backslash i}) - \max_{\theta} \mathsf{E}_{\mathcal{S}} \log \phi_{i|\backslash ij}(Y_i|Y_{\backslash ij}, \theta_{\backslash ij})
\end{aligned}
$$

Terms proportional to logistic reg log-likelihoods.

- Choose response vble Y_i,
  fit saturated model on other vbles, w/wo Y_j.

If d_{ij} is difference in two residual deviances,
then $\widehat{w}_{ij} = d_{ij}/(2N_{\mathcal{S}} \log(2))$.

- Use multinomial logistic regression for $3+$ levels of response.

- Optimising under the main effects model gives a deviance difference to approximate $\widehat{w}_{ij}$.

d_{ij} is not symmetric in i and j,

choose the max: $\tilde{w}_{ij} = \max(d_{ij}, d_{ji})$.

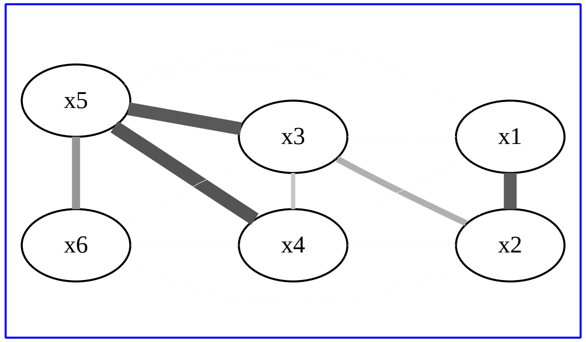
A bootstrapping experiment

Concern to show divergences are stable.

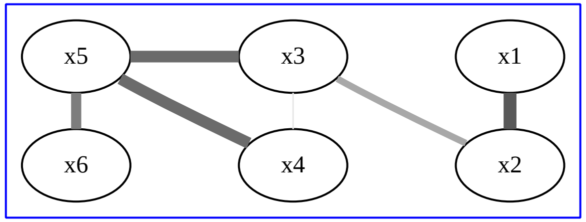
Slight perturbation population proportions
$\rightarrow$ slightly perturbed divergences.

Take pop from fixed k-dim prob:
$k = 6$ vbles, categ $2^5 \times 3$, $N_{\mathcal{P}} = 10000$.

Pop dwig from ent & dev approx:

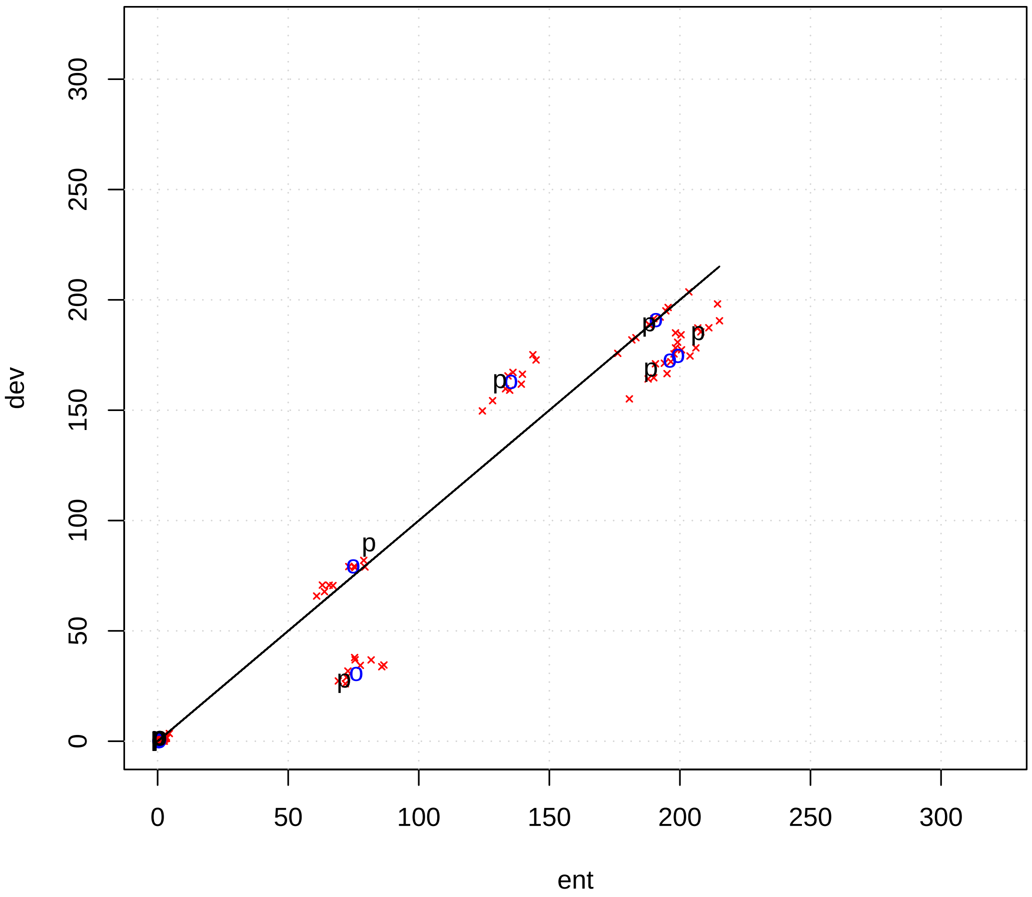


set max 300 mbits
actual 200.8153



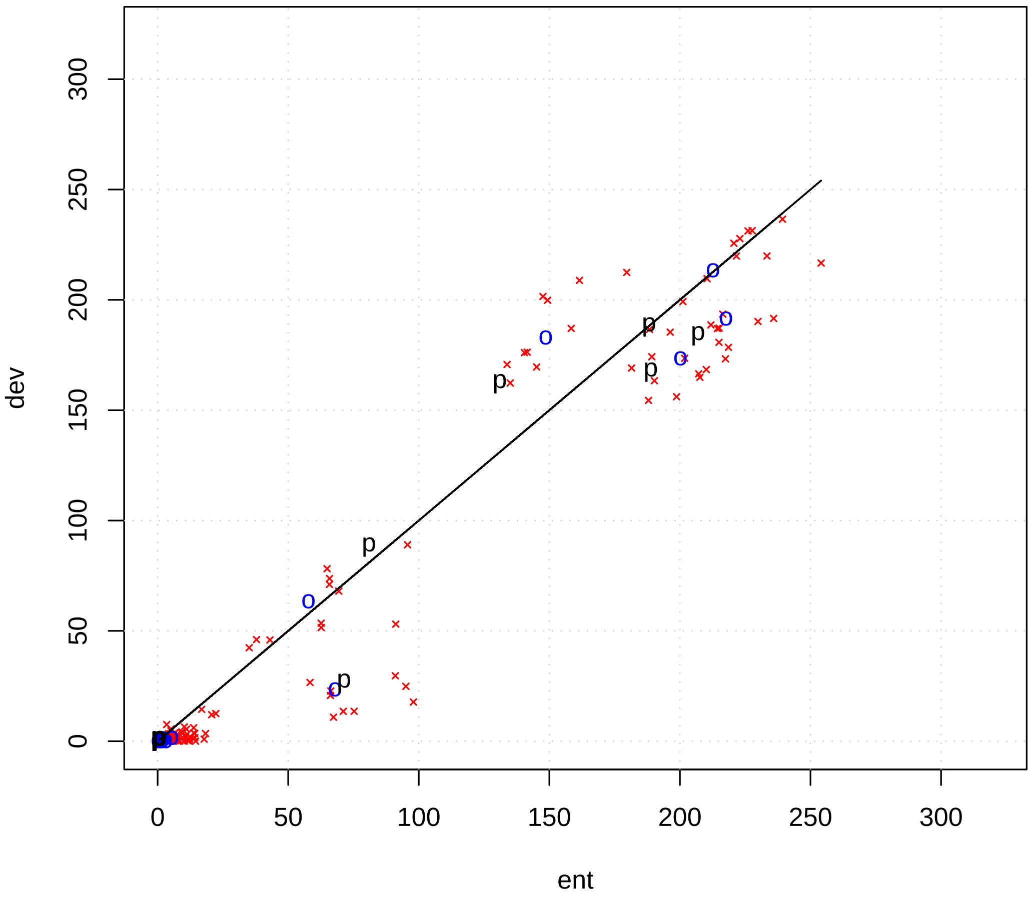
set max 300 mbits
actual 192.1975

Divergences from 10 boots: $N_s = 3200$



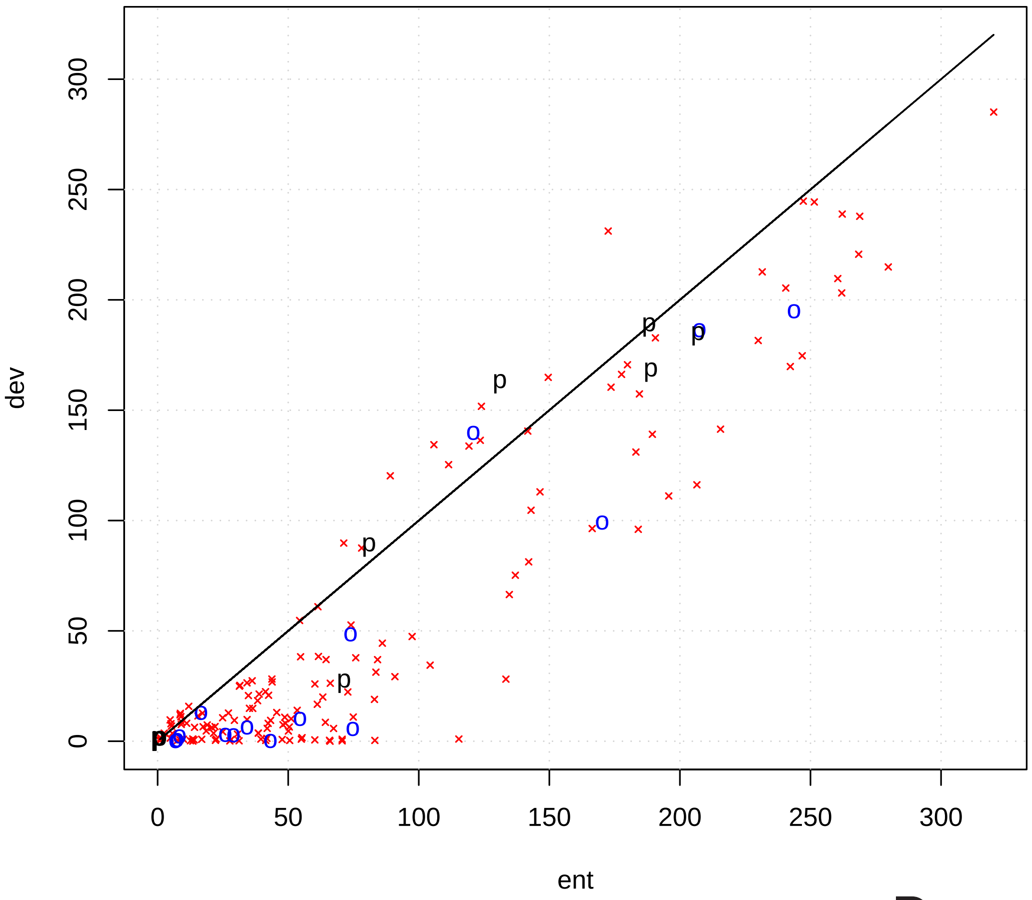
- ent vs dev,
  - 150 points = edges x repetitions.
- 6 non zero divergences, clear pattern.
- Get different dwig from dev or ent,
  - not on 45^o line,
  - measure different things, interactions,
  - but give similar edge ordering
  - ent $\geq$ dev?
- Variation in ent or dev much the same
  - good, as needed for inference.
  - Smaller when div zero.
  - High intra-boot correlation ent and dev.

Divergences from 10 boots: $N_s = 800$



Pattern still clear.

Divergences from 10 boots: $N_s = 200$



Pattern no longer evident.

Summary

- Divergence measures extend CI and graphs to design-based survey framework.

- Useful tool for EDA in varying dimensions, of marginal and conditional tables.

- The relative weights gives coherence to graph.

- Use graphViz for display, R-package dwig.

- Bootstrap inference works, however statistical criterion for stability based on bootstrap dist with established theory is needed,

eg to make statements $k \approx 10$ needs $N_s \approx 1000$.