OH SHIT, IT'S A SURVEY

Daniel Simpson
University of Toronto

with Lauren Kennedy, Alex Gao, and Andrew Gelman

CEASELESS FEASTS OF SCHADENFREUDE

LIKE IT OR NOT, MOST DATA IS UNREPRESENTATIVE

- Life is nothing but suffering and pain.
- One of the big statistical challenges is dealing with this unrepresentative data
- Unsurprisingly, this is a quite well studied problem, but it is currently having a bit of a resurgence.

WHAT WE WOULD LIKE TO HAPPEN

- In a dream scenario, we would know the way in which our data is non-representative. Preferably it would be sampled **probabilistically**

- For example, DHS surveys are carefully **designed** and this known design can be used to undo the non-representativeness of the data.

- Most of the time, we do this using an elaborate system of weights that are roughly related to the probability that a particular observation would be in the data set

DON'T DREAM IT'S OVER

- Let's say we want to estimate the population average. The straightforward estimate

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

is hopelessly biased.

- But if we know enough about the design we can attach a weight w_j to each observation to make a new estimator

$$\bar{y}_w = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j}$$

GENERALLY SPEAKING WEIGHTED SURVEY ESTIMATORS ARE UNBIASED

- If the weights are chosen correctly, $\bar{y}_w$ will be unbiased!
- And if you're really clever, you can often work out the variance of the estimator
- A similar scheme works for regression and GLMs (but not anything with a random effect!)
- All of this is to say that most survey work makes extensive use of weights.

WE SHOULD ALL PROBABLY KNOW THAT UNBIASEDNESS IS ONLY A USEFUL PROPERTY IF A LOT OF OTHER THINGS ARE GOING WELL

ALL MY HAPPINESS IS GONE

- Most unrepresentative data does not come from a simple probability sample.
- Non-response, non-compliance, or a difficult to define population will all tank the most traditional survey methods and introduce a lot of bias.
- There are weights-based methods that try to deal with this, but I think everyone agrees that this makes stuff very hard.

SURVEYS AS A PREDICTION PROBLEM

- Much like in the rest of statistics, it's worth asking ourselves, "What if we give up unbiasedness and focus instead on accuracy?"

- Little's 1997 paper is perhaps the best description of this.

- The idea is that if we can **predict** the response in the unobserved population, then we can use those predictions to estimate any population quantity of interest!

SURVEYS AS A PREDICTION PROBLEM

- So if I have an unobserved member of the population with covariates vector x_i, my job is now to **predict** the corresponding observation $\tilde{y}_i = \tilde{y}_i(x_i, y_{\text{obs}})$

- The corresponding estimate of the population mean would then be

$$\frac{1}{N}\left(\sum_{j=1}^{n} y_i + \sum_{j=n+1}^{N} \tilde{y}_j\right)$$

- This is unbiased if the predictions are

MEET YOUR NEW ASSUMPTIONS

- Weights-based methods make assumptions about the design and implementation of the data collection .

- Prediction-based methods replace these assumptions with the assumption that **the unobserved data can be well predicted by the observed sample.**

- So this concept is certainly not a panacea, but it can be quite useful

BUT WHAT ARE WE PREDICTING BASED ON?

- In a lot of situations, the auxiliary information contained in the covariate vector x_i are (or have been coerced into) categorical variables (eg state, race/ethnicity, gender, etc).

- But some variables are decidedly **not** categorical, like age, location

- But let’s deal with the first case

THE MAGNIFICENT MISTER P

HOW DO WE PREDICT WITH CATEGORICAL VARIABLES

- If we had enough observations for each combination of covariates, we could just use the cell average as a good prediction of the average outcome in that cell. (No pooling of information)

- Or we could take the overall average. (Complete pooling of information.

- Or we could use a multilevel (GLMM) model to borrow strength appropriately.

BUT WHAT DO WE DO ONCE WE'VE FIT A MODEL?

- We need to predict every unobserved person!

- We do this by constructing a **poststratification matrix** that tells us how many people have each combination of covariate levels (ie how many are in each cell).

- We then use the model to estimate the total response in the cell.

FOR EXAMPLE

- If we have a binary response (eg “Did this person die of liver cancer?”) then in the jth cell (corresponding to Female, 34-39, Oregon, Smoker, Heavy Drinker, Favourite Britney Album is Blackout) has N_j people in it.

- If we predict the probability of dying of liver cancer for people in cell j is $\tilde{p}_j$, then our estimated number of deaths is either $N_j\tilde{p}_j$ **or** $Binomial(N_j, \tilde{p}_j)$

WHAT ARE THE ADVANTAGES OF MRP

- The best thing about MRP is that we we have freedom to choose our regression model.
- This means we can use the regularization properties of multilevel regression to control the variance from small counts in some cells
- It is also very natural to think of small-area estimation with MRP

IS MRP A WEIGHTING METHOD?

- Kinda? If there are weights, they are going to be data-dependent. (And they might be negative)

- If the data is fit with a linear model containing **all interactions** and with **no regularization**, you can show that this is equivalent to using the classical poststratification weights

$$w_j = \frac{N_j}{n_j}$$

- But usually we don't have all of these interactions and the "weights" are difficult to interpret (sometimes negative etc)

WHAT IS THE OBVIOUS PROBLEM WITH MRP

- Where does the poststratification matrix come from????
- Decadal censuses are decadal and do not measure important things. The ACS is too small to fill in a lot of cells. Same with specialized surveys for smaller populations.
- Some limited work has been done on modelling and regularizing the poststratificaiton matrix and this is an **important practical problem we need to solve**

BUT WHAT IF OUR CATEGORIES ARE ORDINAL?

ORDER MUST BE MAINTAINED

- Using multilevel modelling is super useful when the stratifying variables are nominal categorical variables.

- But what about when there's an order?

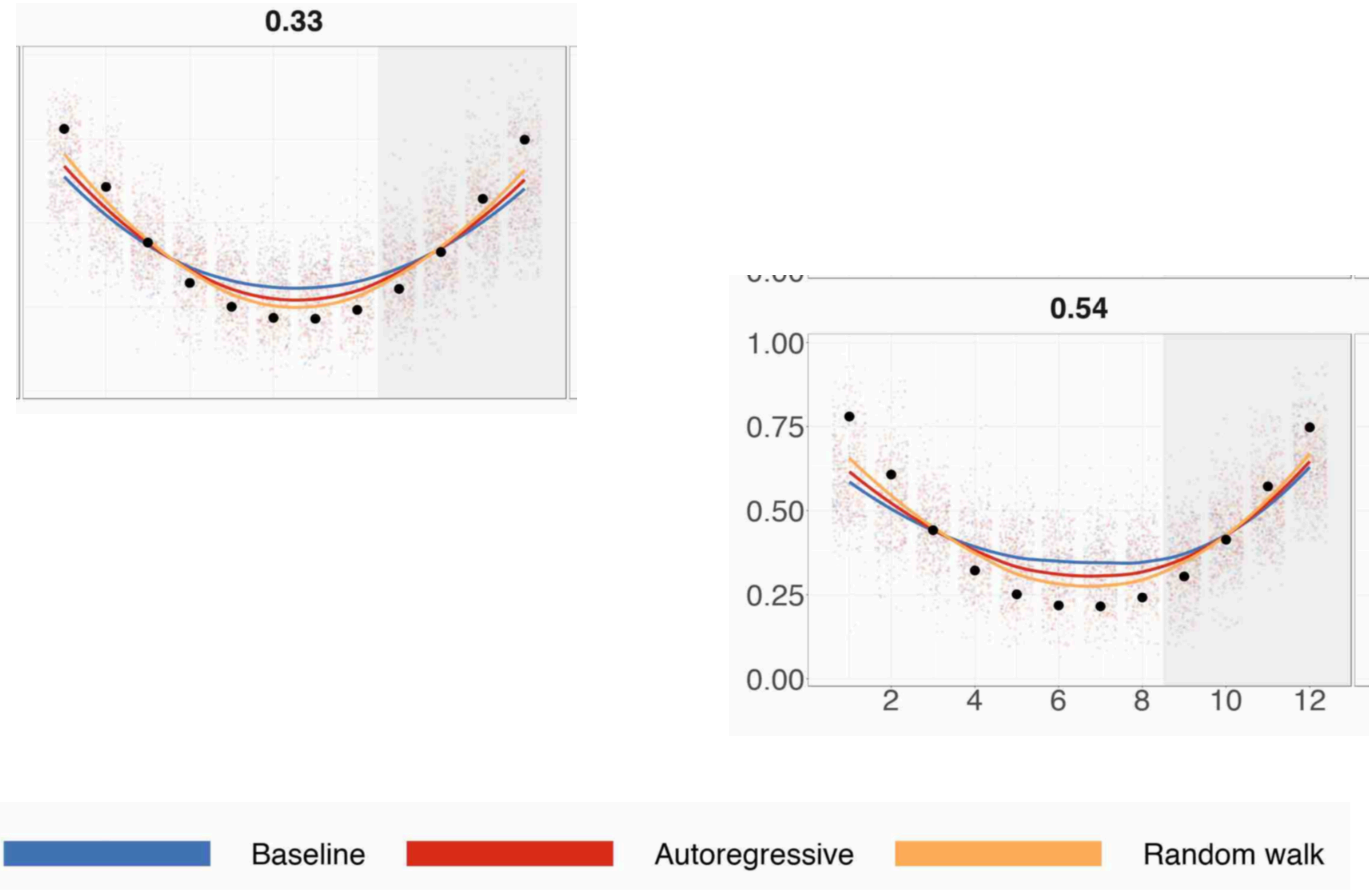
- When dealing with weights, you usually have to collapse categories into each other to balance the bias with the variance.

- But if we are being predictive we probably can just model it!

AN EXAMPLE: AGE

- Age is, by all accounts, a continuous variable that is ordered.
- So if I wanted to model something that depends on age, I would naturally use a time series or spline model for flexibility.
- If I just use an iid random effect, i’m probably going to miss any non-linear effect

A COMPARISON



BUT DOESN'T THAT MAKE IT HARD TO BUILD THE POSTSTRAT MATRIX?

- So it's definitely true that we can make better predictions that model the effect of ordinal variables.

- But that's only half of the job: we also need to generalize to the population.

- Alex struggled to fine 60 year old Asian men in Alabama earning more than $150k in the ACS.

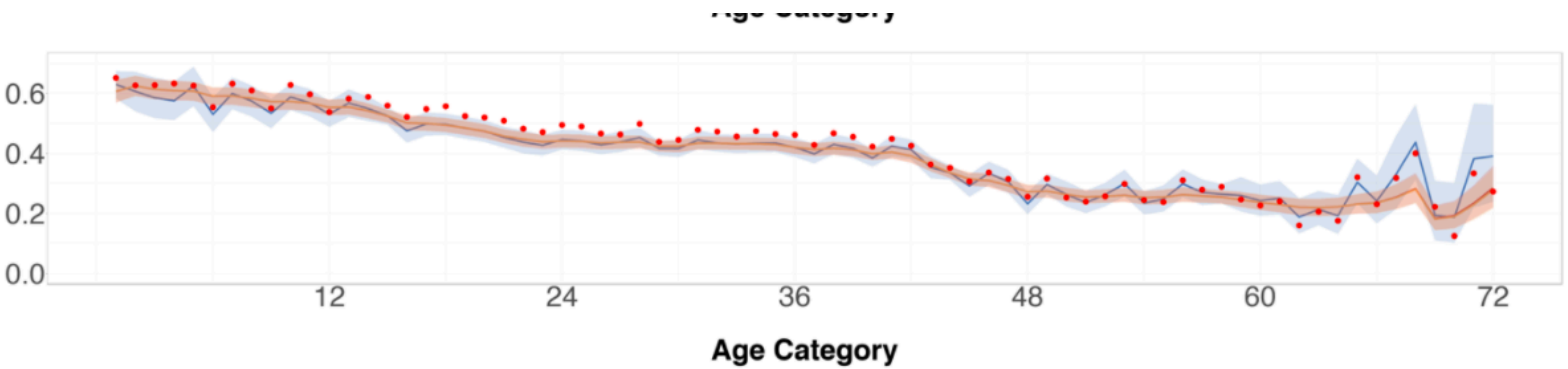
A KEY POINT THAT WE DIDN'T UNDERSTAND

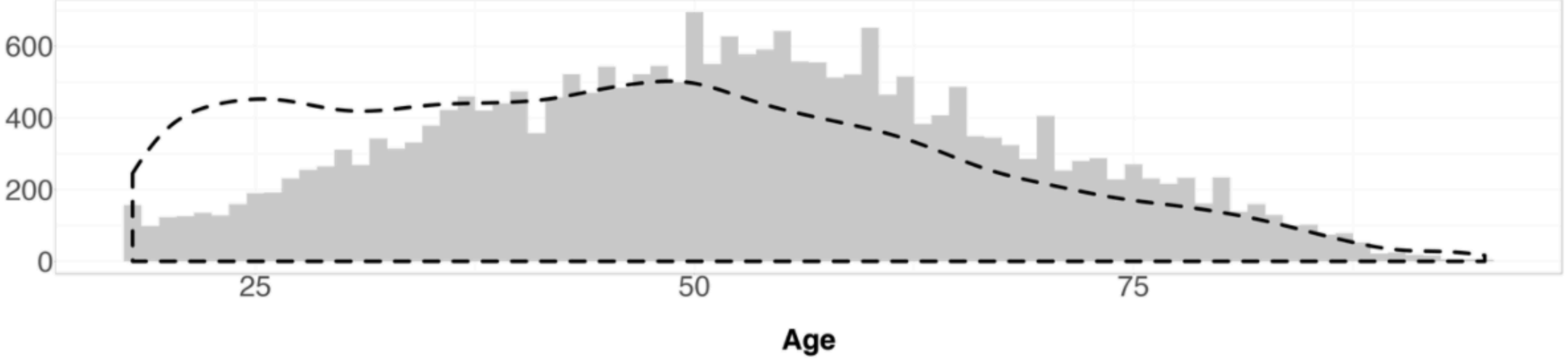
- We can predict the at a scale finer than the poststratification cells.

- So although we predict for every age, we can use these to get estimates in the post-strat row for 18-30!

- **Estimate at the natural scale for the data, predict at the natural scale for the population information!**

STABILITY


Age Category
0.6
0.4
0.2
0.0
12
24
36
48
60
72
Age Category


600
400
200
0
25
50
75
Age

IF OUR SURVEY IS AT A FINER LEVEL THAN OUR POPULATION, WE CAN USE THAT

- Age isn’t uniformly distributed in age bins.
- Income isn’t uniformly distributed in income bins.
- We can probably do something about this!

CHOICES

THIS BRINGS US TO ONE OF THE REALLY IMPORTANT QUESTIONS

- How do we **know** that Model A is better than Model B for survey data?
- How do we do variable selection?
- What does variable selection even mean?

NO INSTRUMENTAL BREAKS

- Generally speaking, we want three things:

  1. We need the prediction variables to be correlated with the response. (So if we use a non-linear model it should be a good one).
  2. We need our poststratification variables to be correlated with the result
  3. We want to avoid poststratifying on instrumental variables (ie variables that are correlated with selection but not the outcome)

HOW DO WE MONITOR THESE THINGS?

- Model selection for surveys is hard

- Lumley and Scott (2015) show that if you have survey weights, using those weights to compute leave-one-out cross validation scores is a good idea

TYPICAL USE OF CROSS VALIDATION

- Take your data y and pull out k observations (here either k is 1 or around 10% of your observations)
- Fit your model on the remaining observations.
- Compute the **log-score** of your remaining observation

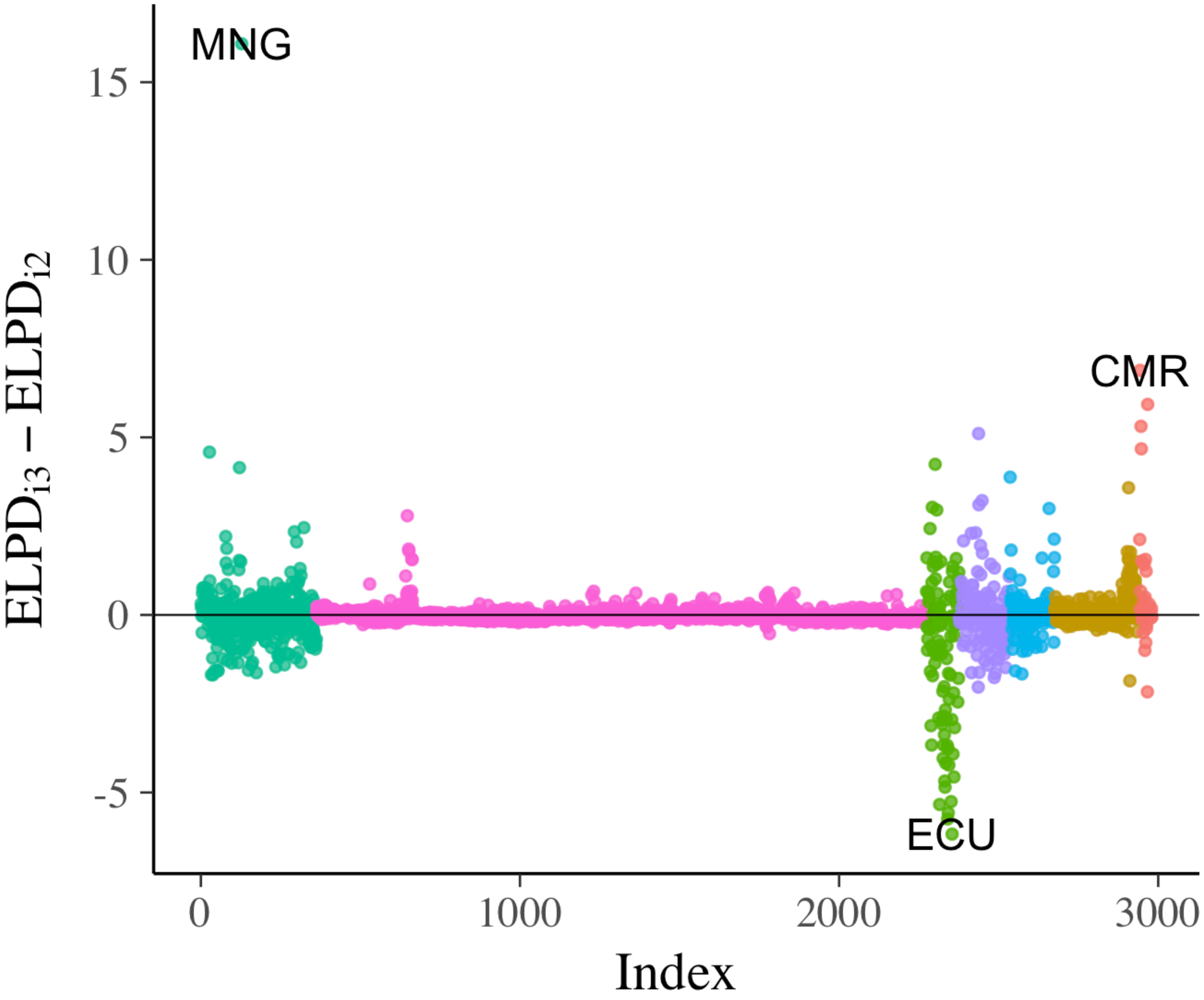
$$\text{elpd}_j = \frac{1}{k}\sum_{i=1}^{k} \log p(y_i \mid y_{-k})$$

- Repeat this as many times as possible and report the average elpd.

BIGGER IS ALWAYS BETTER

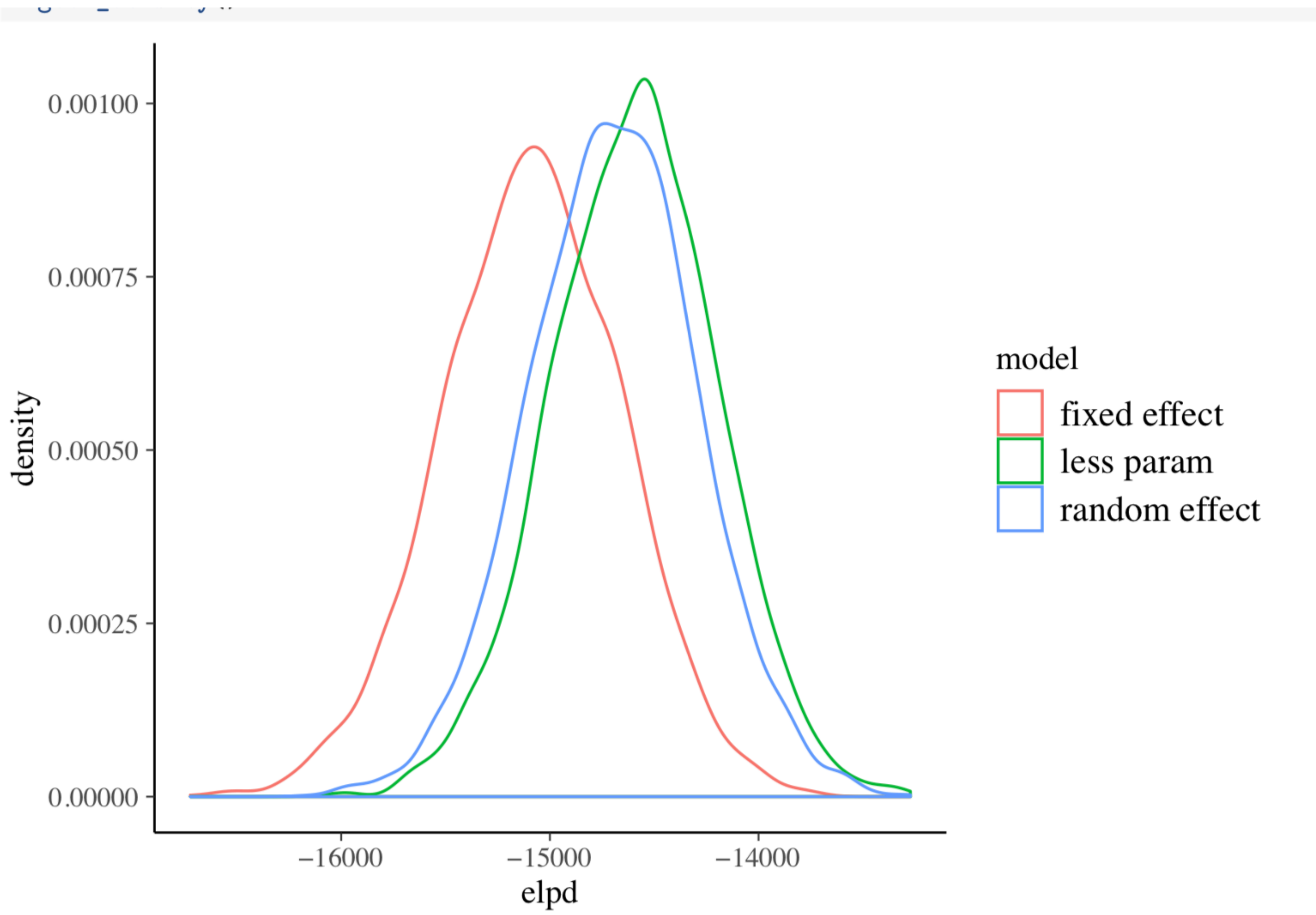
- This is then typically used to compare between two models
- The assumption is that bigger is better.
- Why? Because elpd converges (under independence assumptions) to a constant minus the Kullback-Leibler divergence between the prediction and the data generating distribution.
- But there’s more information than just the sum…

MORE INFORMATION THAN JUST COMPUTING A SUM



CAN WE USE MRP HERE?

- Maybe? The idea is that we can compute that leave-one-out log-predictive density for each point in our sample and then use MRP to estimate the **population leave-one-out expected log-predictive density.**

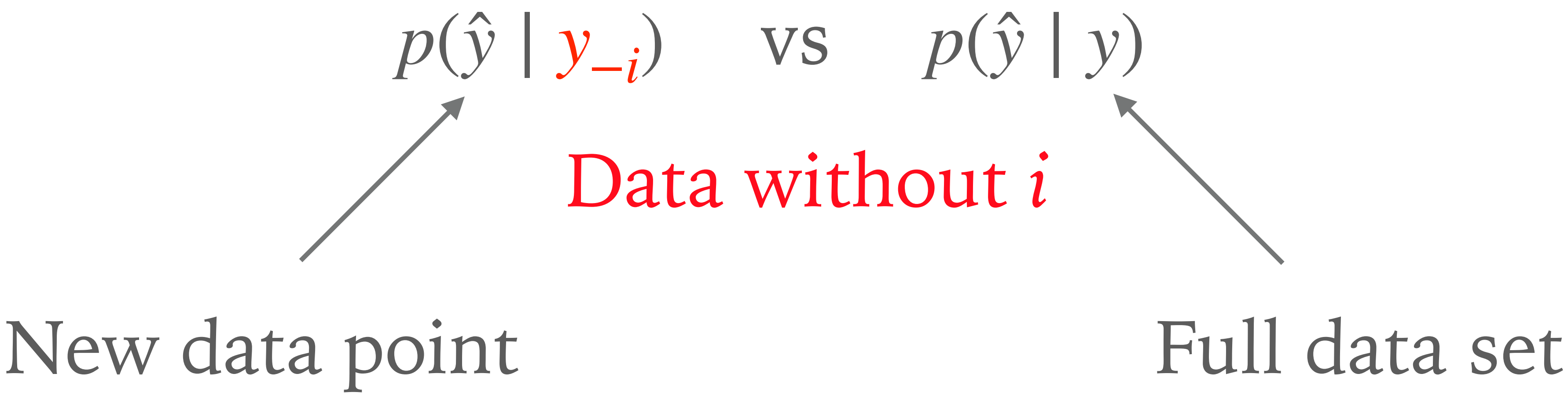


THE SECRET UTILITY OF LEAVE ONE OUT DISTRIBUTIONS

- But what if we didn't care about comparing with another distribution?
- There's still value here.
- Why? Because what does it mean if the leave-one-out predictive distribution is different from the whole-data predictive distribution?
- It is **important** to know about these influential points!
- They are very similar to high-leverage points in linear regression.

JUST A MOMENT

- For observation i, we want to compare



$$p(\hat{y} \mid y_{-i}) \quad \text{vs} \quad p(\hat{y} \mid y)$$

Data without i

New data point

Full data set

- How do we tell if these things are similar?
- **Idea:** The full data predictive distribution should be a good importance sampling proposal for the loo distribution, ie

$$\operatorname{Var}_{\theta \sim p(\theta \mid y_{-i})} \frac{p(\theta \mid y_{-i})}{p(\theta \mid y)} < \infty$$

OK THIS IS HARD TO CHECK IN GENERAL

- In general, we only have access to a sample of these importance weights.
- This makes things hard.
- But some classical statistics comes to the rescue:

The extreme tail of a distribution converges
to a Generalized Pareto Distribution

- So while we might not be able to check analytically if the importance weights has a finite variance, we can estimate the distribution of the extreme weights, **which gives us the same information.**

THE GENERALIZED PARETO DISTRIBUTION

- The generalized Pareto distribution has the

$$p(z) = \frac{1}{\sigma}\left(1 + kz\right)^{-1/k-1}$$

- The key parameter is *k*, which controls how many moments the tail distribution has.
- We can estimate *k* by k-hat, which tells us how many moments a specific sample appears to have.
- This is an extremely useful, and easy to compute, quantity. Because if k-hat is large, then the LOO predictive distribution is **very** different from the full data predictive distribution!

SHOULD WE BE LEAVING MORE THAN ONE THING OUT?

- Should we be leaving out strata? Or cells?

- Should we be leaving out strongly dependent things?

- Should we leave out the future? The past? A window?