Estimating Preferences Over Data to Inform Statistical Disclosure Control Decisions

Elan Segarra

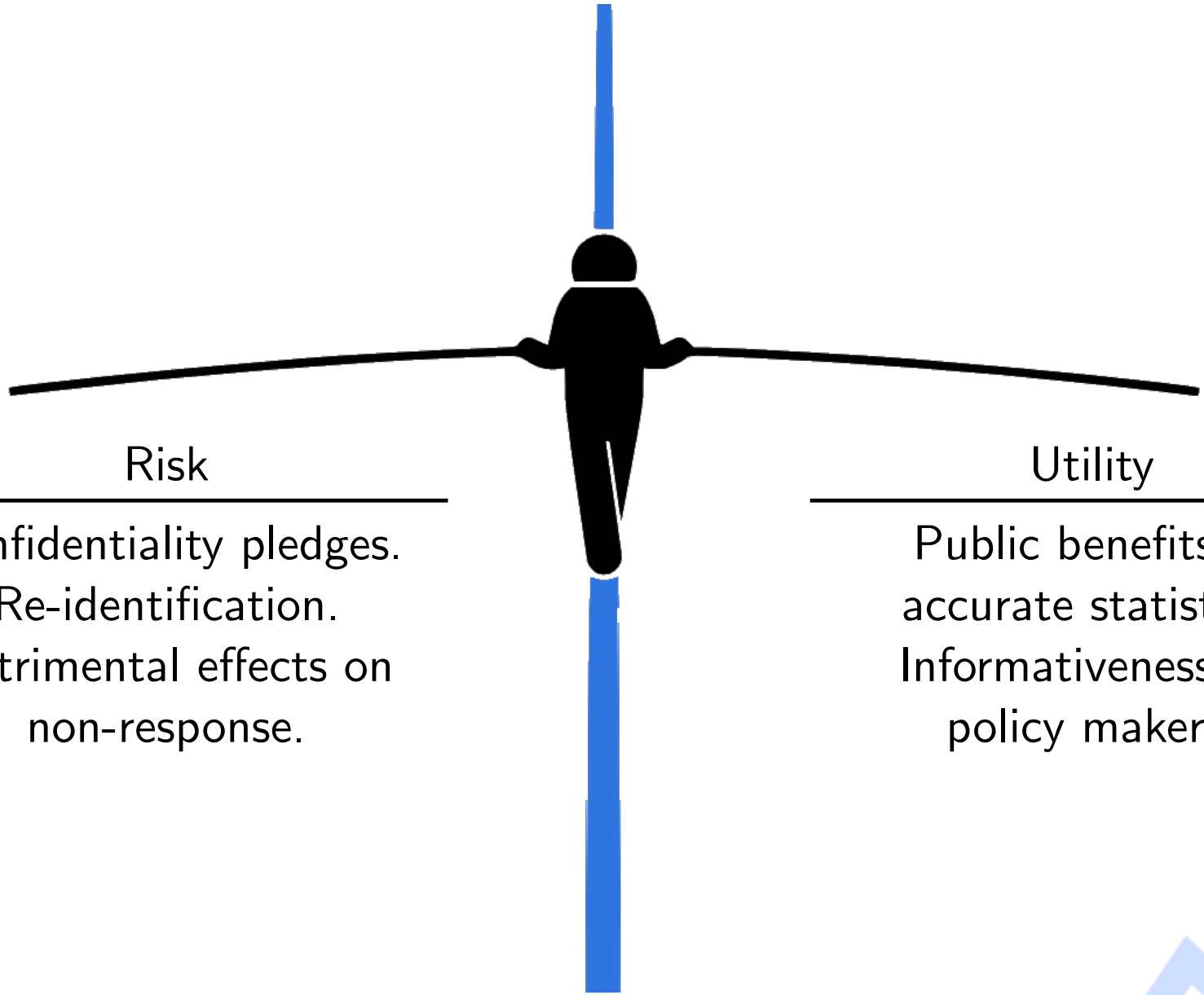
U.S. Bureau of Labor Statistics

Data Privacy Protection and the Conduct of Applied Research

May 17th, 2024

Disclaimer: The views expressed herein are those of the author(s) and do not necessarily reflect those of the Federal Government, Department of Labor, or the Bureau of Labor Statistics. All results have been reviewed to ensure that no confidential information is disclosed.

Statistical Disclosure Control: Implicit Balancing Act



Risk

Confidentiality pledges.
Re-identification.
Detrimental effects on non-response.

Utility

Public benefits of accurate statistics.
Informativeness for policy makers.

Two Margins of Choice within SDC Methods

Extensive Margin	**Intensive Margin**
How much total acceptable risk?	Where to distribute the accuracy?
Cell Suppression	
Which cells are sensitive? Acceptable identifiable bounds on suppressions?	Which complementary suppressions?
Differential Privacy	
What is the privacy budget (i.e. ε)?	How to allocate ε across potential publications?

The work presented here focuses on the intensive dimension

Project Overview

Goal: Quantify data users' preferences over statistics and incorporate them into the intensive margin of SDC decisions.

Approach:

1. Estimate a nested logit model of consumer preferences.
2. Generate valuations for potential statistics using preferences.
3. Optimize SDC intensive decisions using valuations.

Application:

- Census of Fatal Occupational Injuries: Data being consumed.
- Google Analytics: Pageview data used to estimate preferences.
- Significant heterogeneity in preferences over characteristics with highest value on breakdowns by employment status.
- Estimated cell/table valuations are used in 2 SDC approaches: cell suppression and differential privacy.

Application: Subject Data - CFOI

The Census of Fatal Occupational Injuries (CFOI):

- Annual census of fatal work injuries collected since 1992.
- Compiled by a Federal-State cooperative program which collects data info multiple sources (police reports, news, OSHA investigations, etc.).
- Compiled data includes narratives, injury codes (OIICS), geography, timing, and demographic information.

Application: Subject Data - CFOI

The Census of Fatal Occupational Injuries (CFOI):

- Annual census of fatal work injuries collected since 1992.
- Compiled by a Federal-State cooperative program which collects data info multiple sources (police reports, news, OSHA investigations, etc.).
- Compiled data includes narratives, injury codes (OIICS), geography, timing, and demographic information.

Disclosure control is particularly difficult for CFOI:

- It is a census, the counts are small, and some data is public.
- BLS publishes many tables/figures/statistics using CFOI (e.g. industry and occupational breakdowns).
- Currently uses cell suppression to protect confidentiality.

Model of Consumer Choice Over Statistics

Three important objects make up the model:

1. A **statistic** is an individual scalar.
   - Ex: The count of work fatalities in the construction sector
2. A **publication** is a collection of statistics.
   - Ex: A table/figure of work fatality counts by industry
3. A **market** consists of a set of publications at a specific time from which a data consumer can choose.
   - Ex: On BLS.gov a data consumer can choose to view fatality counts by employee status, industry, occupation, or age group

Key insight: Observing which publications are chosen (i.e. clicked on) reveals preferences over the underlying statistics

Model of Consumer Choice: Nested Logit

Consumer i has indirect utility from publication p in market t given by

$$U_{ipt} = \underbrace{\frac{1}{|\mathcal{S}_p|}\sum_{s\in\mathcal{S}_p} X_{st}\beta + Z_{pt}\alpha + \xi_{pt}}_{\equiv\delta_{pt}} + \varsigma_{ig} + (1-\sigma)\epsilon_{ipt}$$

$\mathcal{S}_p$: Set of statistics included in publication p
X_{st} : Observable characteristics of statistic s (eg ind. breakdown)
Z_{pt} : Observable characteristics of publication p (eg bar chart)
ξ_{pt} : Unobservable stat. characteristics (eg ugly presentation)
ς_{ig} : Unobservable correlated nest shock (eg broken site link)
ϵ_{ipt} : Unobservable characteristics (eg researcher vs layperson)

Objects of interest: β and α quantify preferences across characteristics

Nest Structure

Model of Consumer Choice: Utility Maximization

If users choose the publication (e.g. table) that maximizes their indirect utility, then the observed "market share" of publication p in time period t is given by

$$s_{pt} = \frac{\exp\left(\frac{\delta_{pt}}{1-\sigma}\right)}{D_g^{\sigma} \sum_h D_h^{1-\sigma}} \qquad \text{where} \qquad D_g = \sum_{k \in g} \exp\left(\frac{\delta_{kt}}{1-\sigma}\right)$$

Note: "Market share" $= s_{pt} = \frac{q_{pt}}{M_t} = \frac{\text{pageviews}}{\text{site visitors}}$

Estimation:

1. Inversion step (the "magic" of logit):

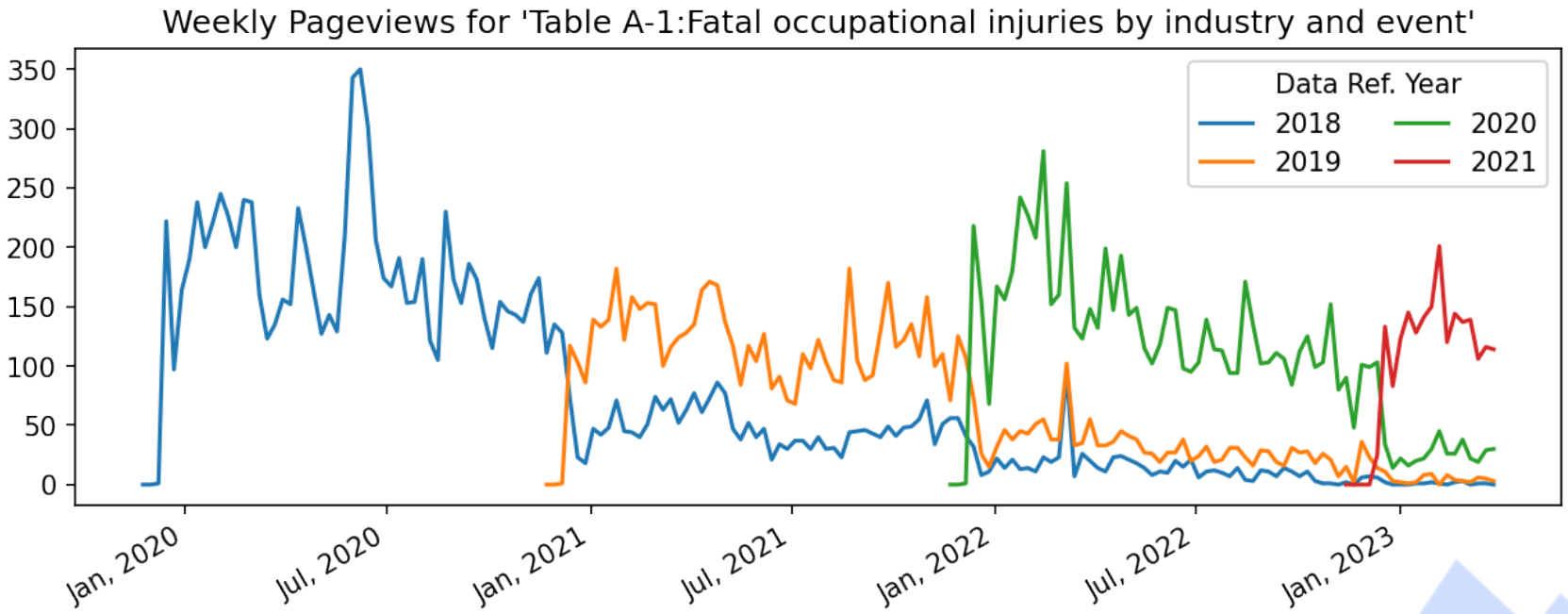
$$\ln s_{pt} - \ln s_{0t} = \tilde{X}_{pt}\beta + Z_{pt}\alpha + \sigma \ln s_{p|g} + \xi_{pt}$$

2. Estimate using instrumental variables.

Data: Google Analytics

Google Analytics

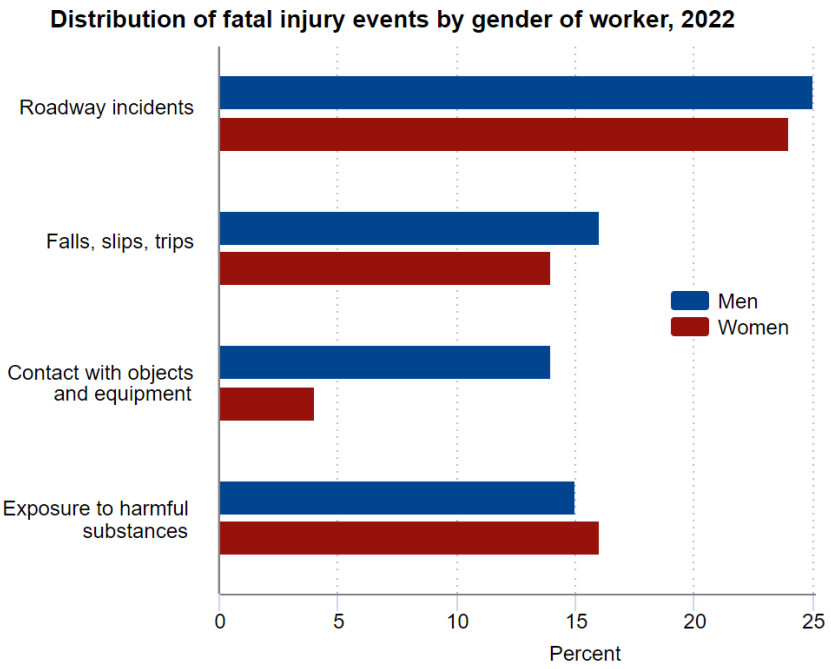
- Granular data on pageviews, duration, and even demographics
- There are 28 different tables/figures each published over multiple reference years
- Relative pageviews function as our measure of choice among consumers



Data: Extracting Characteristics

Characteristics are manually coded for each site, such as

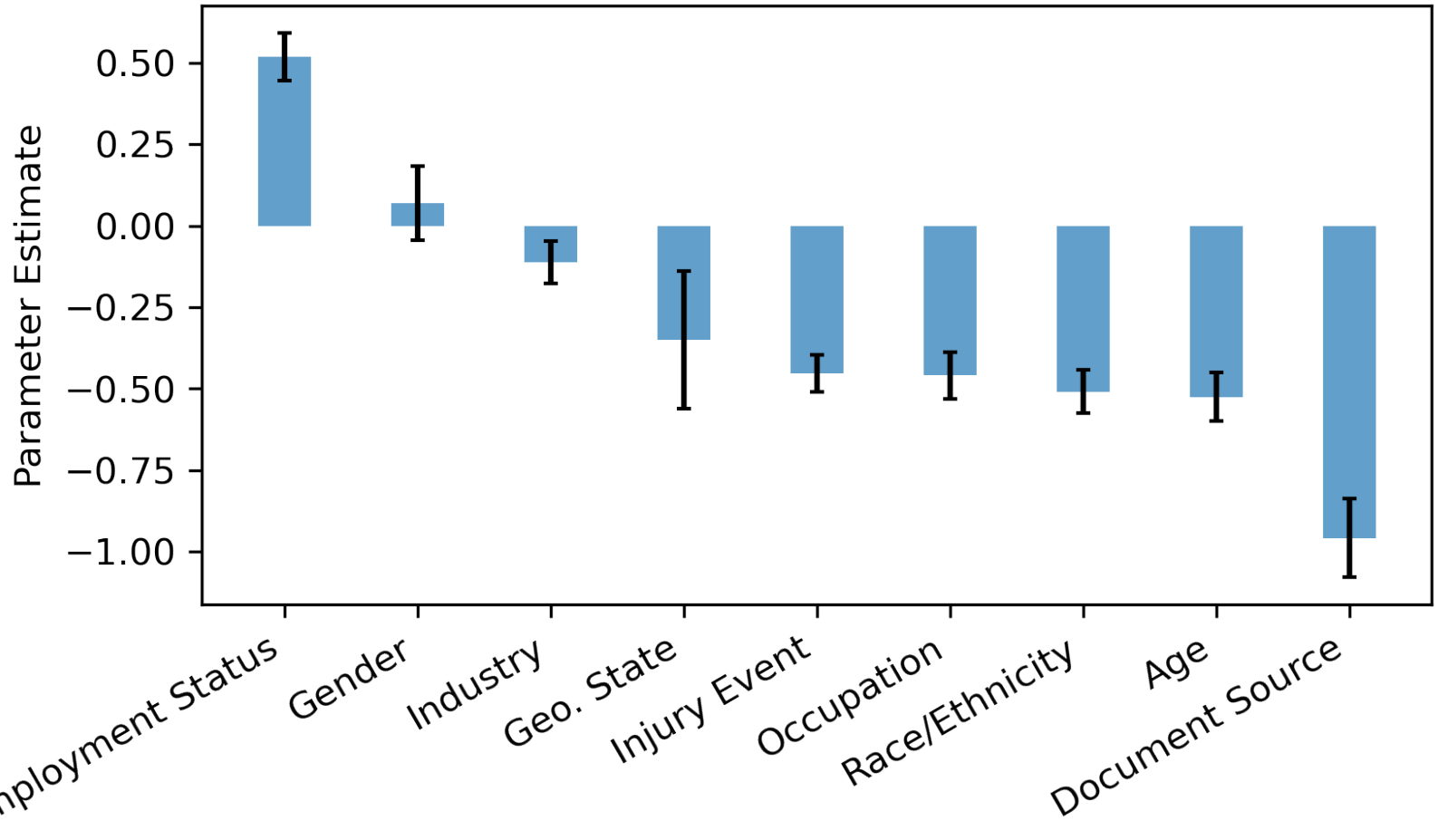
- *ind* = includes breakdown by industry sectors
- *format* = bar chart, table, time series etc.
- *multiyear* = includes more than one year of data
- *curr_year* = includes most recent RY (at view time)
- *exposition* = includes exposition along with data



$$\Rightarrow \quad \widetilde{x}_{pt} = \begin{bmatrix} curr\_year \\ ind \\ gender \\ event \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ \vdots \end{bmatrix}$$

$$z_{pt} = \begin{bmatrix} multiyear \\ format\_bar \\ exposition \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

Estimation Results: Statistic Characteristics



Though the exact values of $\widehat{\beta}$ are difficult to interpret, their relative ordering reflects which breakdowns are more valued by data consumers.

Statistical Disclosure Control (SDC)

Consider these 2 tables of synthetic CFOI statistics:

Table 1: Fatalities by Age and Year

	Year			
Age	2019	2020	2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404

Table 2: Fatalities by Age and Industry

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

We consider two SDC contexts:

1. Cell Suppression Problem
2. Differential Privacy

How can we use our estimated preferences to inform our SDC methods? $\Rightarrow$ Computing valuations over cells and/or tables.

Computing Valuations

Could construct mean utilities, $\mathbb{E}\left[\widehat{U}_{ipt}\right] \Rightarrow$ several issues.

Computing Valuations

Could construct mean utilities, $\mathbb{E}\left[\widehat{U}_{ipt}\right] \Rightarrow$ several issues. Instead we define the estimated valuation of publication p as

$$\widehat{v}_p = \mathbb{E}\left[P(\text{Choose pub. } p)\right] = \frac{\exp\left(\tilde{x}_p\widehat{\beta} + z_p\widehat{\alpha}\right)}{\sum_{k=1}^{P}\exp\left(\tilde{x}_k\widehat{\beta} + z_k\widehat{\alpha}\right)}.$$

This approach assesses value using a hypothetical market including only the potential cells/tables under consideration.

Example

Given those 2 table options and their characteristics:

- $\widehat{v}_1 = P(\text{Choose Table 1}) = 0.361$
- $\widehat{v}_2 = P(\text{Choose Table 2}) = 0.639$

The publication by age and industry is 77% more valuable (on average) than the publication by age and reference year.

Disclosure Application: Cell Suppression Problem

SDC Method: Tabular Cell Suppression

- Sensitive cells are suppressed to protect confidentiality.
- Additional cells (i.e. complementary/secondary) often need to be suppressed.

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Disclosure Application: Cell Suppression Problem

SDC Method: Tabular Cell Suppression

- Sensitive cells are suppressed to protect confidentiality.
- Additional cells (i.e. complementary/secondary) often need to be suppressed.

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Primary Suppression

Disclosure Application: Cell Suppression Problem

SDC Method: Tabular Cell Suppression

- Sensitive cells are suppressed to protect confidentiality.
- Additional cells (i.e. complementary/secondary) often need to be suppressed.

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Primary Suppression

One Set of Complementary Suppressions

One Set of Complementary Suppressions

There are often multiple options for complementary suppressions. Estimated valuations *over the cells* can help guide decisions.

Disclosure Application: Cell Suppression Problem

Table 1: Fatalities by Age and Year

Age	Year			
	2019	2020	2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404

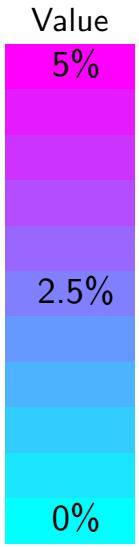


Table 2: Fatalities by Age and Industry

Age	Industry (for 2021)			
	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Individual cell valuations are used as an input into any CSP solver to find optimal complementary suppressions.

Disclosure Application: Cell Suppression Problem

Table 1: Fatalities by Age and Year

Age	Year 2019	2020	2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404

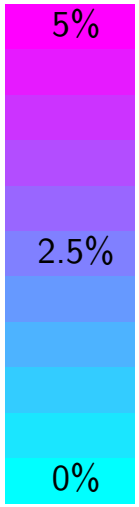


Table 2: Fatalities by Age and Industry

Age	Industry (for 2021) Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Individual cell valuations are used as an input into any CSP solver to find optimal complementary suppressions.

Note: The optimization will be limited by the granularity available in the preference data. For example:

- Without separate publications for Trade versus Mfg, estimated valuations are constant across industries.
- Can result in multiple optima.

Disclosure Application: Differential Privacy

Differential Privacy (DP):

- Algorithm property that provides a provable confidentiality guarantee and typically involve noise injection.
- Increasing adoption of DP across the NSAs.
- Typically involve a privacy budget, $\varepsilon > 0$.
  - Larger $\varepsilon \Rightarrow$ less noise and less security
  - Smaller $\varepsilon \Rightarrow$ more noise and more security

Disclosure Application: Differential Privacy

Differential Privacy (DP):

- Algorithm property that provides a provable confidentiality guarantee and typically involve noise injection.
- Increasing adoption of DP across the NSAs.
- Typically involve a privacy budget, $\varepsilon > 0$.
  - Larger $\varepsilon \Rightarrow$ less noise and less security
  - Smaller $\varepsilon \Rightarrow$ more noise and more security

Simple heuristic for leveraging estimated valuations:

- Allocate the privacy budget, ε, among publications, e.g.
  - For Table 1 and 2 use 0.361ε and 0.639ε.
  - More generally, for P potential publications
    1. Estimate valuations: $\widehat{v_1}, \ldots, \widehat{v_P}$
    2. For publication p use $\widehat{v_p}\varepsilon$ in the DP mechanism
- Maintains aggregate risk guarantee ε while ensuring more accuracy for higher valued statistics.

Conclusion

Summary:

- Simple framework for leveraging preferences over data:
  1. Estimate a nested logit model of consumer preferences.
  2. Generate valuations for potential statistics using preferences.
  3. Optimize over SDC intensive decisions using valuations.
- Proof of concept using CFOI:
  - Estimated preferences using GA pageviews.
  - Found significant heterogeneity in prefs over characteristics.
  - Illustrated use of cell/table valuations in two SDC methods: Cell Suppression and Differential Privacy

Future Work:

- Incorporate data from the BLS custom query tool.
- Explore the framework with other SDC methods and other BLS data products.
- Consider alternative information sources to identify preferences of other stakeholders.

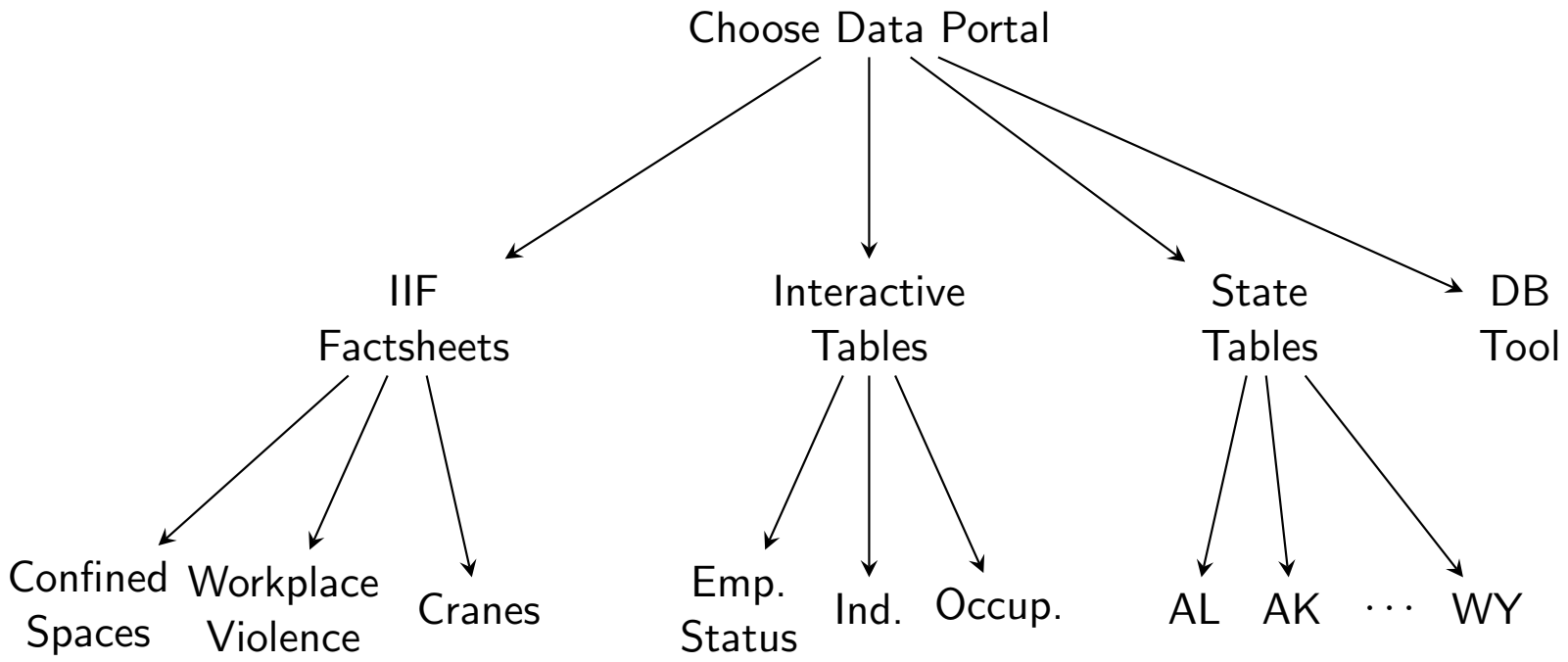
CONTACT INFORMATION

Thank You!

Elan Segarra
U.S. Bureau of Labor Statistics
Office of Compensation and Working Conditions
Segarra.Elan@BLS.gov

Nested Logit Model: Decision Tree

The nests (i.e. groups of related statistics/tables) line up with the likely route a user takes to access the statistic or table:



Can possibly generalize to other access mediums (e.g. twitter or API) or even to entire catalog of BLS data products.

Estimation

To put all of this in a more familiar form, if we define

$$y_{jt} = \ln q_{jt} - \ln q_{0t}$$

Then we have

$$y_{jt} = X_{jt}\beta + \sigma \ln s_{j|g} + \xi_{jt}$$

and since y_{jt}, X_{jt}, and $s_{j|g}$ are all observed we can use traditional regression methods to estimate β and σ.

Estimation

To put all of this in a more familiar form, if we define

$$y_{jt} = \ln q_{jt} - \ln q_{0t}$$

Then we have

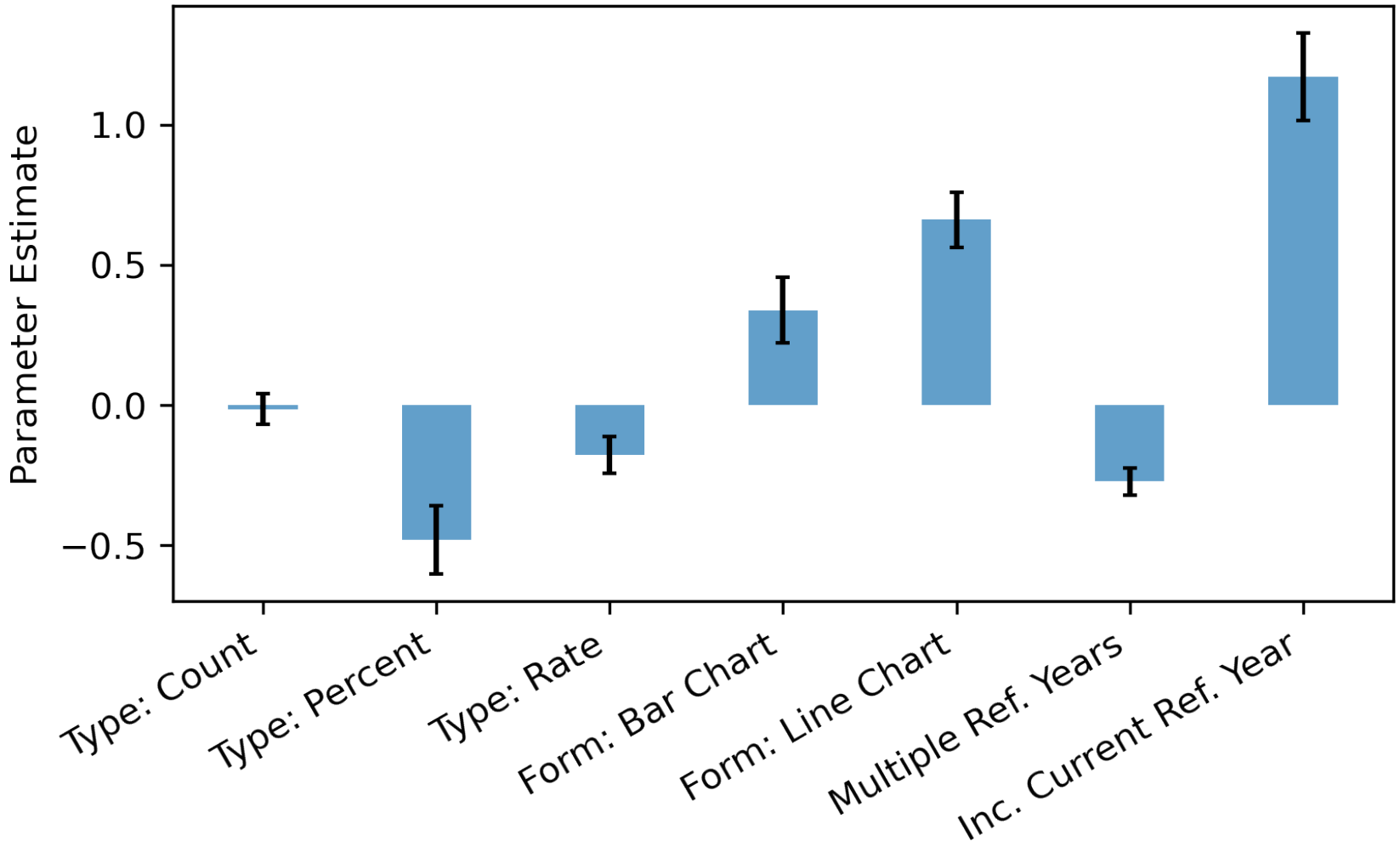
$$y_{jt} = X_{jt}\beta + \sigma \ln s_{j|g} + \xi_{jt}$$

and since y_{jt}, X_{jt}, and $s_{j|g}$ are all observed we can use traditional regression methods to estimate β and σ.

Open question: Can we use OLS or is there endogeneity that would require something like IV?

- Reminder: ξ_{jt} are unobs. table characteristics
- Seems like obs. table characteristics (X_{jt}) are exogenously determined by BLS

Estimation Results: Publication Characteristics



Similarly, the exact values of $\widehat{\theta}$ are difficult to interpret, but they suggest which publication chars. are valued by data consumers.

Regression Results (Full)

	(1)	(2)
Employment Status	0.518**	0.518**
	(0.037)	(0.038)
Industry	-0.112**	-0.112**
	(0.030)	(0.033)
Occupation	-0.459**	-0.459**
	(0.034)	(0.036)
Gender	0.069	0.069
	(0.072)	(0.058)
Injury Event	-0.453**	-0.453**
	(0.025)	(0.029)
Age	-0.525**	-0.525**
	(0.042)	(0.038)
Geo. State	-0.350**	-0.350**
	(0.088)	(0.107)
Race/Ethnicity	-0.509**	-0.509**
	(0.034)	(0.034)
Document Source	-0.958**	-0.958**
	(0.054)	(0.061)
Constant	-3.308**	-3.308**
	(0.182)	(0.118)

** indicates significance at the 0.01 level.

	(1)	(2)
Type: Count	-0.016	-0.016
	(0.030)	(0.028)
Type: Percent	-0.482**	-0.482**
	(0.075)	(0.062)
Type: Rate	-0.179**	-0.179**
	(0.036)	(0.034)
Form: Bar Chart	0.337**	0.337**
	(0.051)	(0.060)
Form: Line Chart	0.660**	0.660**
	(0.058)	(0.051)
Multiple Ref. Years	-0.274**	-0.274**
	(0.021)	(0.025)
Inc. Current Ref. Year	1.170**	1.170**
	(0.059)	(0.080)
Nest Shares	0.358**	0.358**
	(0.029)	(0.037)
Mkt FE	Yes	Yes
Robust SE	No	Yes
N	5870	5870

** indicates significance at the 0.01 level.