SOCIOLOGICAL METHODOLOGY

4

INFERRING LOGIT MODELS FROM EMPIRICAL MARGINS USING PROXY DATA

*Ju-Sung Lee**
*Kathleen M. Carley**

We examine several approaches for inferring logit models from empirical margins of predictor covariates and conditional margins containing the means of a binary response for each covariate margin. One method is to fit proxy data to the conditional response using the beta distribution, a process we call "margin analysis." Proxy data can obtained using three approaches: (1) implementing the iterative proportional fitting (IPF) procedure on the margin totals, (2) sampling from a larger relevant data source such as the census, and (3) enumerating, or sampling from, the combinatoric space of all possible tables constrained by the margins. The first procedure is a well-studied approach for estimating contingency tables from margins, but it does not necessarily maintain the associations between the covariates unless seeded with an initial

This work was supported in part by the Internal Revenue Service (IRS) for research on errors (TIRNO08E00040) and by the Office of Naval Research (ONR) for research on dynamic networks (N00014-08-1-1186) and metric robustness (N00014-06-1-0104) . Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Internal Revenue Service, the Office of Naval Research, or the U.S. government. Direct correspondence to Kathleen M. Carley, ISR, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA; e-mail: Kathleen.carley@cs.cmu.edu.

*Carnegie Mellon University

table containing those associations. In the second approach, which is appropriate for analyzing sociodemographic covariates, we can use a large census sample adjusting for sampling biases observed in the empirical margins. However, the appropriateness of using a census proxy depends substantially on how similar the sampling pools are. Our third approach entails exploring the combinatoric space of all contingency tables constrained by the margins while considering the associations among the covariates. We aggregate the logit models estimated from each table in that space into a single model. This approach is more robust than the first two as it considers multiple proxies. While the estimated logit models from each approach are generally similar to one another, for the low-dimensional tables we explore in this paper, the combinatoric approach incurs wider standard errors, which renders potentially significant coefficients insignificant. Finally, we suggest weighting the combinatoric models with evidence-relevant probabilities obtained using the multivariate Pólya distribution.

1. INTRODUCTION

Model estimation for contingency tables is driven by the extent of information available about those tables, from exact cell frequencies to odds-ratios to fixed margin totals. When all cell frequencies are available, the log-linear approach is appropriate, and it estimates coefficients for predicting frequencies expected by the marginal totals as well as interactions (Bishop, Fienberg, and Holland 2007). In the case of "partial information release," a contingency table (or distribution of tables) is implied and demands alternative estimation approaches.[1]

A common form of partial information is marginal totals (or simply margins), and these often summarize the data (including some dependent response) or describe a sample while maintaining confidentiality—for example, aggregated census data (Dobra, Karr, and Sanil 2003; Fienberg 2005). The iterative proportional fitting (IPF) procedure, another expectation-based estimation approach, provides contingency tables constrained to margin totals (Deming and Stephan 1940).[2] Additional information can include conditional margins, which

[1] We credit Aleksandra Slavković for the phrase "partial information release."

[2] In the basic implementation, the initial seeded table is uninformative and contains uniform values.

not only adds a dimension but also offers multidimensional margins, as opposed to the typical unidimensional margins. The benefits of inferring contingency tables and models from partial information are significant, especially to research that extends and synthesizes the reports and models of published material, such as meta-analysis.[3]

The data examined in this paper pertain to tax compliance behavior, particularly whether or not respondents have committed tax evasion, or intentional tax error,[4] on their recent tax returns.[5] Since the 1970s, a host of studies have modeled this behavior with various sociodemographic, behavioral, and attitudinal predictors. A number of these studies report marginal totals of their sample as well as conditional margins typically in the form of proportion of respondents in each margin category whose response is tax evasion.

We outline an approach for estimating a logit model from proxy data whose margins are consistent with empirical sociodemographic sample count margins and are fitted to the conditional response margins.[6] The foremost complication is accounting for the association between the covariates. Bishop and colleagues summarize some measures of association highlighting the work by Goodman and Kruskal (1954, 1959, 1963, 1972). Single measures of association, such as the correlation coefficient or chi-squared-based measures, are considered inappropriate for tables larger than 2 × 2, as they do not adequately capture the ways a table can deviate from independence. Since association is often a "multidimensional concept," research suggests that a multidimensional measure be employed for tables larger than 2 × 2. As our work pertains to fixed margin totals, the class of relevant measures is called "margin sensitive." In this paper, we propose using

[3] While we may obtain the original data from the authors of the published works, our attempts were unfruitful.

[4] The term "intentional error" stems from the IRS classification of incorrect portions of tax returns as being one of two types of errors: intentional and inadvertent. The IRS, however, cannot definitively categorize an error as either intentional or inadvertent; without substantial evidence of willful intention by the taxpayer, the error is usually categorized as inadvertent. While we use the term "noncompliance" also synonymously with "intentional error," in other writings it might be used to indicate either kind of error.

[5] The time horizon varies from study to study, spanning past year's returns to 5-years to lifetime.

[6] We use the phrase "logit model" synonymously with "logistic regression."

the complete table to capture associations using a measure that offers probabilistic comparisons to other tables.

In order to estimate a logit model, we require either covariate data or a contingency table, which can be enumerated into data. Of course, the margins of the proxy data will need to match the empirical margins. Our first table is derived from the IPF and represents our control as it largely assumes independence among the covariates. Next, we examine the use of a secondary data set, which is presumptively similar to the data underlying the margins. Since the margin covariates we examine are strictly sociodemographic, we turn to census data as a potential proxy.

Our third approach eschews fitting to conditional margins from just a single contingency table, and it explores the combinatoric space of all possible contingency tables constrained by the sample margins and aggregates the logit models produced by each hypothetical model. Focusing on a single combinatoric solution that best resembles the census proxy is tantamount to seeding the IPF with informative values, including census data, in order to maintain some level of association. Deming and Stephan (1940), Friedlander (1961), Causey (1984), Bartholdy (1991), and Little and Wu (1991) have investigated informative seeding of the IPF.

The aggregation of model coefficients can also be adjusted, if the data underlying the reported sample margins presumptively exhibit some association, by weighting each constituent model by its degree of relevance to the empirical association. For these weights, we employ the multivariate Pólya distribution, which offers a probabilistic measure of fit between some evidence table and a candidate table. Finally, we deal with additional complications due to inconsistent empirical margins (indicative of missing data) and multiple ways of exploiting conditional margins.

2. LINEAR PREDICTION OF CONDITIONAL RESPONSE MARGINS

Among the prominent sociodemographic predictors often found in tax compliance studies are gender, age, education, and self-employment. While gender and self-employment are binary indicators, the others, though discrete, span a range large enough to warrant condensing into a smaller set of categories (or bins), as seen in many studies that report margins of sociodemographic variables.

The cases we explore include only discrete bivariate and trivariate distributions, or contingency tables, of the predictors and a binary response.[7] As a diagnostic exercise, we first explore how well simple linear prediction can reconstitute the joint response distribution, expressed as a table of proportions of the covariate contingency table. For the diagnostics that follow, we restrict our options to five cases: 2 × 2, 2 × 3, 3 × 4, 2 × 3 × 2, and 2 × 3 × 4. These dimensions are associated with the following covariate combinations:[8]

x_1	x_2	x_3	n_1^{bins}	n_2^{bins}	n_3^{bins}	$n_{\text{Total}}^{\text{bins}}$
Gender	Self-employment	–	2	2	–	4
Gender	Age	–	2	3	–	6
Age	Education	–	3	4	–	12
Gender	Age	Self-employment	2	3	2	12
Gender	Age	Education	2	3	4	24

We fix the values for a covariate i of size n_i^{bins} to span the discrete range $\{0, \ldots, n_i^{\text{bins}} - 1\}$; for example, gender and self-employment take on values of 0 and 1, while the age covariate can be 0, 1, or 2.

For our diagnostics, we want to assess the degree of predictive error incurred by table dimensionality. To this end, we generate simulated data and predict the response by fitting a logit model to data margins. Our response, y, is a binary variable, so we draw uniform Bernoulli probabilities for $p(y = 1)$ for each cell of the contingency table:

$$p_{ij[k]} \sim \text{Unif}(0, 1), \tag{1}$$

where $i \in \{0, \ldots, n_1^{\text{bins}} - 1\}$, $j \in \{0, \ldots, n_2^{\text{bins}} - 1\}$, and, if the data are trivariate, $k \in \{0, \ldots, n_3^{\text{bins}} - 1\}$. For sampling uniform contingency tables, we draw a prior probability vector of length $n_{\text{Total}}^{\text{bins}}$ from a uniform Dirichlet and employ it in its conjugate multinomial draw for a sample

[7] The response technically constitutes an additional dimension/contingency.

[8] For age and education, we condense the covariates into three and four categories, respectively. This manner of condensing large-ranged covariates is not uncommon.

set of cells totaling to n:[9, 10]

$$\boldsymbol{q} \sim \text{Dirichlet}(\boldsymbol{\alpha} = (1, \ldots, 1))$$
$$\boldsymbol{x} \sim \text{Multinomial}(n, \boldsymbol{p} = \boldsymbol{q})$$

Note that $\boldsymbol{p}$ in the multinomial is a parameter label and distinct from the uniformly drawn $\boldsymbol{p}$ of equation (1). For these diagnostics, we set the sample size to $n = 200$. For each contingency table $\boldsymbol{x}$ and response distribution $\boldsymbol{p}$, we infer a logit model, which can be accomplished by either enumerating the $\boldsymbol{x}$ into a full data matrix or performing weighted logistic regression. For example, for the two-dimensional data, we have the following model:[11]

$$\text{logit}[p(y = 1)] = \beta_0 + \beta_1 x_1 + \beta x_2$$

In order to assess the accuracy of IPF and margin analysis (an alternative fitting method we will explain shortly), we calculate the proportional joint response distribution predicted from the logit on either the actual sample $\boldsymbol{x}$ or some proxy derived from the sample (e.g., IPF), and compare that to the original response $\boldsymbol{p}$ using mean square error (MSE).

We draw a set of random response probabilities $\boldsymbol{p}$ and a set of random sample distributions $\boldsymbol{x}$. We then pair each joint response distribution $\boldsymbol{p}$ with each joint data distribution $\boldsymbol{x}$, yielding a convoluted population of $\boldsymbol{x}$, $\boldsymbol{p}$ pairings. We keep only those pairings that are guaranteed to provide convergent models; specifically, we omit those that contain zero margins or any joint response cell frequency of zero as these seem to produce nonconverging results. Due to this manner of rejection sampling, our initial set of draws varies according to its dimensionality:

[9] We could accomplish this by multiplying the Dirichlet draw $\boldsymbol{q}$ by the sample size $n + 1$ and taking the floor; however, this approach requires that we reject draws that do not sum to n, which is typically about 4/5 of the draws, or adjust the sample, which can lead to biases; therefore, this method is less efficient than using the multinomial.

[10] Our convention is that a **boldfaced** variable—for example, $\boldsymbol{X}$ or $\boldsymbol{x}$—denotes a vector or matrix of quantities as does a set of values held within a pair of parentheses—for example, $(x_0, \ldots, x_n)$; nonboldfaced variables denote scalars.

[11] The logit$[x]$ is the log odds of x, $\log[\frac{x}{1-x}]$; conversely, logit$^{-1}[x]$ represents the inverse-logit, $\frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+e^{-x}}$.

Dimensions	***p***, ***x*** Draws	Potential Samples	Valid Samples
2 × 2	20	400	365
2 × 3	25	625	376
3 × 4	50	2500	317
2 × 3 × 2	50	2500	317
2 × 3 × 4	560	313600	392

We next summarize IPF and introduce margin analysis, a procedure to fit covariate data to conditional response margins.

2.1. *Iterative Proportional Fitting Procedure*

The iterative proportional fitting (IPF) procedure is an approach for estimating a contingency table from marginal totals and an initial seed table (Deming and Stephan 1940; Friedlander 1961; Fienberg 1970; Dobra and Fienberg 2001). A uniform seed table yields cells with expected values, similar to those produced by a log-linear model with no interaction terms. We have an unknown contingency table with cell values $\boldsymbol{x}$ from whose marginal totals we estimate $\hat{\boldsymbol{x}}$. For an uninformed table, we choose initial seed values of $\hat{x}_{ij}^{(0)} = 1$ and repeat the following:[12]

$$\hat{x}_{ij}^{(2\eta-1)} = \frac{\hat{x}_{ij}^{(2\eta-2)} x_{i\cdot}}{\sum_{k=0}^{J} \hat{x}_{ik}^{(2\eta-2)}} \quad \text{and} \quad \hat{x}_{ij}^{(2\eta)} = \frac{\hat{x}_{ij}^{(2\eta-1)} x_{\cdot j}}{\sum_{k=0}^{I} \hat{x}_{kj}^{(2\eta-1)}},$$

where $I =$ (the number of rows) and $J =$ (the number of columns) and the known marginal totals (i.e., row sums and column sums) are[13]

$$x_{i\cdot} = \sum_{k=0}^{J-1} x_{ij} \quad \text{and} \quad x_{\cdot j} = \sum_{k=0}^{I-1} x_{kj}$$

[12] Alternatively, the seeded table can contain nonuniform values, in which case the IPF will maintain some of the interactions.

[13] Boldfaced $\boldsymbol{x}_{i\cdot}$ denotes the vector of margin values while nonboldfaced $x_{i\cdot}$ refers to some scalar statistic, typically the sum over the unlabeled '·' margin, in this case the second dimension or column.

until we reach a predetermined convergence condition

$$\left(\sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \hat{x}_{ij}^{(t)} - \hat{x}_{ij}^{(t-1)} \right) < 1 \times 10^{-10}.$$

Not surprisingly, more information leads to greater accuracy in the IPF estimation. For trivariate data, three sets of two-dimensional margins provide a better fit than three sets of one-dimensional margins; however, papers typically report only one-dimensional margins. Beckman, Baggerly, and McKay (1996) elaborate on Deming and Stephan's approach for estimating multidimensional IPF tables from more than two unidimensional margins.

Furthermore, conditional margins imply an additional dimension and provide not only an accompanying set of margins but a subset of the multidimensional margins. For instance, in the 2×2 case (i.e., [g]ender by [s]elf-employment), an additional conditional response margin (i.e., y = noncompliance) implies $2 \times 2 \times 2$ data. However, the conditional response margins offer only two of the three two-dimensional margins (i.e., $g \times y$ and $s \times y$).[14]

Alternatively, we can perform IPF on the sample and response margins separately to obtain the nonresponse IPF table

$$f_{\text{IPF}}(\boldsymbol{x}(1-\boldsymbol{p})) = f_{\text{IPF}}(\boldsymbol{x}) - f_{\text{IPF}}(\boldsymbol{xp}),$$

where $f_{\text{IPF}}(\boldsymbol{z})$ is our function for the IPF algorithm and returns the estimated contingency table fitted to a known set of margins associated with table $\boldsymbol{z}$. We can also obtain the IPF joint response distribution $\boldsymbol{p}_{\text{IPF}}$:

$$\boldsymbol{p}_{\text{IPF}} = \frac{f_{\text{IPF}}(\boldsymbol{xp})}{f_{\text{IPF}}(\boldsymbol{x})}.$$

We can then either expand both the rounded response and nonresponse IPF tables into a full data set, including the binary response variable, and perform a straightforward logistic regression or, to be more precise, perform a weighted logistic regression.[15] As the IPF ratio approach

[14] While it might be possible to enhance IPF to use only some of the required multidimensional margins, we have not found this in the literature.

[15] IPF often yields noninteger cell values.

exploits more information than the IPF, the estimated tables are more similar (according to MSE or χ^2) to the combined data and response contingency table than tables estimated from three sets of unidimensional margins.[16] If the sample or conditional margins are inconsistent due to missing data, we would need to augment the lesser margins. If only the sample margins are consistent, or easily made consistent, we can instead use the IPF-estimated table solely from the sample margins and fit a logit model to the conditional margins using the "margin analysis" procedure we detail next.

2.2. *Margin Analysis*

As an alternative to the IPF ratio approach, we offer a way to fit the sample margins to the conditional response margins using actual or proxy covariate data. This approach is also suitable when the conditional margins yield inconsistent response population sizes due to missing data. Given a proxy contingency table with cells $\hat{\boldsymbol{x}}$ (either IPF-derived or from another source such as a census), we wish to infer a logit model from conditional response margins, $\boldsymbol{p}$, and margins associated with a table with unknown cells $\boldsymbol{x}$. For example, given a 2 × 2 contingency table, we have margins of length I and J and the following cell counts of the proxy:

		x_2	
		0	1
x_1	0	$\hat{x}_{00}$	$\hat{x}_{01}$
	1	$\hat{x}_{10}$	$\hat{x}_{11}$

For some set of parameter coefficients $\boldsymbol{\beta}$, we obtain a predicted probability for each covariate combination, i and j where $i, j \in \{0, 1\}$

$$\hat{p}_{ij} = \text{logit}^{-1}(\boldsymbol{\beta}[i\ j]^{\mathrm{T}}),$$

and we compute the predicted conditional response margins (i.e., weighted mean of the response for category for each covariate):

[16] While the analysis supporting this assertion does not appear in this paper, it is available from the authors upon request. Furthermore, we have yet to test for the extent to which a complete set of multidimensional margins is superior to IPF ratio.

$$\hat{p}_{i\cdot} = \frac{\sum_{j=0}^{J-1} \hat{p}_{ij}\hat{x}_{ij}}{\sum_{j=0}^{J-1} \hat{x}_{ij}} \quad \text{and} \quad \hat{p}_{\cdot j} = \frac{\sum_{i=0}^{I-1} \hat{p}_{ij}\hat{x}_{ij}}{\sum_{i=0}^{I-1} \hat{x}_{ij}}.$$

We employ the beta distribution to fit our predicted marginal probabilities to the empirical conditional margins $\boldsymbol{p_{i\cdot}}$ and $\boldsymbol{p_{\cdot j}}$.[17] The log-likelihood, for the two-dimensional table, is[18, 19]

$$\begin{aligned}\mathcal{L} = &\sum_{i=0}^{I-1} \log[\text{Beta}(\hat{p}_{i\cdot}|\alpha = (1 - p_{i\cdot})x_{i\cdot} + 1, \beta = p_{i\cdot}x_{i\cdot} + 1)] \\ &+ \sum_{j=0}^{J-1} \log[\text{Beta}(\hat{p}_{\cdot j}|\alpha = (1 - p_{\cdot j})x_{\cdot j} + 1, \beta = p_{\cdot j}x_{\cdot j} + 1)]\end{aligned}$$

We find the maximum likelihood $\boldsymbol{\beta}$ using the Newton-Raphson gradient descent optimization algorithm.[20] We obtain the covariance matrix of the estimated coefficients by taking the negative inverse of the Hessian or inverse of the Fisher information matrix, produced by the Newton-Raphson algorithm.[21] If the conditional response margins yield inconsistent total response counts (i.e., $\sum_i p_{i\cdot}x_{i\cdot} \neq \sum_j p_{\cdot j}x_{\cdot j}$), the MA approach will fit toward an overall mean response, weighted by each dimension's total: $\sum_i p_{i\cdot}x_{i\cdot}/\sum_i x_{i\cdot} + \sum_j p_{\cdot j}x_{\cdot j}/\sum_j x_{\cdot j}$. This situation

[17] The beta distribution is suitable because the solution converges to the same estimates and covariance as the logit, when solved with full data, including response, rather than sample and conditional margins.

[18] Each additional margin would add a beta log-likelihood term.

[19] $\mathcal{L}$ denotes a log-likelihood.

[20] While an analytical solution exists for non-logit models (i.e., probabilities are treated as the response to a straightforward linear regression), one does not for the logit model. An analytical solution is intractable due to the logit transformation. If we were fitting to simple [0,1] probability, we can obtain an analytical solution for the two covariate models; we omit the solution from the paper, but it is available upon request.

[21] For example, the variance V for a univarate model (i.e., intercept and one covariate) is

$$V = -[I^{-1}] = -\left[\frac{\delta\mathcal{L}}{\delta\beta_i\beta_j}\right].$$

can arise if there are missing data, leading to sample margins having different totals, and/or the sample or conditional margins are erroneous, producing different response proportions, even with normalized margins. For the data we examine in this paper, we are faced with both complications.

2.3. *Diagnostic Results*

For each pair of covariate data $\boldsymbol{x}$ and joint response distribution $\boldsymbol{p}$, we estimate a predicted distribution $\hat{\boldsymbol{p}}$ from model coefficients $\boldsymbol{\beta}$ predicted through

1. the logit model using $\boldsymbol{x}$ and $\boldsymbol{p}$ (i.e., the full data);
2. margin analysis (MA) to fit the actual data $\boldsymbol{x}$ to the conditional response margins derived from $\boldsymbol{p}$;
3. margin analysis using IPF-derived tables from $\boldsymbol{x}$ fitting also to the conditional response margins derived from $\boldsymbol{p}$; and
4. a weighted logit model on the IPF ratio (IPF$_\mathrm{R}$).

We compare each $\hat{\boldsymbol{p}}$ to its associated $\boldsymbol{p}$ using mean square error (MSE):

$$\mathrm{MSE} = \sqrt{\sum_{i=0}^{I-1} \sum_{j=0}^{J-1} (\hat{p}_{ij} - p_{ij})^2}.$$

We also obtain the following diagnostic MSEs:[22, 23]

	Minimum				Mean				Maximum			
Dim.	Logit	MA	IPF	IPF$_\mathrm{R}$	Logit	MA	IPF	IPF$_\mathrm{R}$	Logit	MA	IPF	IPF$_\mathrm{R}$
2 × 2	0.000	0.000	0.004	0.031	0.255	0.255	0.306	0.392	0.961	0.961	1.285	2.079
2 × 3	0.100	0.100	0.110	0.085	0.499	0.509	0.536	0.526	1.137	1.123	1.772	1.301
3 × 4	0.532	0.537	0.542	0.361	0.965	0.987	0.975	0.889	1.581	1.558	1.568	1.730
2 × 3 × 2	0.480	0.481	0.501	0.368	0.897	0.920	0.935	0.902	1.482	1.693	2.306	1.937
2 × 3 × 4	0.872	0.872	0.888	0.911	1.359	1.374	1.364	1.339	1.890	1.915	1.890	2.053

Higher dimensionality incurs an increase in the potential nonlinearity of the response distribution so we expect a concomitant increase in

[22] We obtain a similar pattern in the results when we employ a χ^2 test of fitness between the actual and predicted response count tables. These are available from the authors upon request.

[23] We ignore extreme, degenerative models (i.e., coefficients >5).

the MSEs. However, since the IPF_R can model nonlinearities in the response, we are not surprised to see its lower MSE relative to the others.

Because we want to better understand the effect of dimensionality of error, we consider several relevant predictors: cell count, degrees of freedom, and the length of the diagonal. If we have I row margins, J column margins, and K slices or layers, the number of unknowns or degrees of freedom is[24]

$$\begin{array}{ccc} \text{Unknown cells} - & \text{Known margins} & = \text{Unknown variables } (df) \\ I \times J \times K & -\, I + (J-1) + (K-1) = & IJK - I - J - K + 2 \end{array}$$

We require only one full set of margins, I, to provide us with a total sample count, which is why we require only $J - 1$ and $K - 1$ parts of the other margins. The diagonal $\sqrt{I^2 + J^2 + [K^2]}$ is a measure of dimensional size. For two-dimensional tables, we omit the K^2 term. The predictors are then

I	J	K	Number of Cells	Number known	Number unknown	Diagonal
2	2	–	4	3	1	2.83
2	3	–	6	4	2	3.61
3	4	–	12	6	6	5.00
2	3	2	12	5	7	4.12
2	3	4	24	7	17	5.39

In order to control for the full dimensional space, we normalize the MSE, dividing by the maximum possible, which for proportions is $\sqrt{IJK}$. The mean and standard deviations of the normalized MSEs are

			Logit		MA		IPF		IPF_R	
I	J	K	μ_MSE	σ_MSE	μ_MSE	σ_MSE	μ_MSE	σ_MSE	μ_MSE	σ_MSE
2	2	–	0.127	0.094	0.127	0.094	0.153	0.097	0.196	0.138
2	3	–	0.204	0.080	0.208	0.079	0.219	0.090	0.215	0.099
3	4	–	0.279	0.054	0.285	0.057	0.282	0.054	0.257	0.064
2	3	2	0.259	0.058	0.266	0.065	0.270	0.070	0.260	0.065
2	3	4	0.277	0.036	0.281	0.037	0.278	0.035	0.273	0.040

[24] Here we use I, J, K to indicate the full length of each dimension, not the highest index. We also, attribute the term "slice" to Deming (1940).

Even with the normalization, there remains the increase of error with dimensionality with a slight drop at 2 × 3 × 2 for the first three measures, suggesting that the diagonal might be the best predictor. Furthermore, the marginal increase in error drops and even reverses, suggesting a nonlinear trend. Also, the IPF and $\mathrm{IPF_R}$ generally exhibit the lowest minimum and the highest maximum MSEs (previous table), so we see larger standard deviations above. When we predict the normalized MSE by the candidate predictors separately, we find that the diagonal predictor confers the highest adjusted-R^2s. Due to the nonlinear nature of the MSE pattern, we also examine quadratic models.[25]

Predictor	Logit	MA	IPF	$\mathrm{IPF_R}$	Logit^2	MA^2	IPF^2	$\mathrm{IPF_R^2}$
Intercept	−0.013^	−0.015^	0.037***	0.115***	−0.491***	−0.539***	−0.412***	−0.002
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Diagonal	0.058***	0.059***	0.048***	0.030***	0.301***	0.326***	0.277***	0.090***
	(0.002)	(0.002)	(0.002)	(0.020)	(0.020)	(0.021)		(0.026)
$\mathrm{Diagonal}^2$					−0.029***	−0.032***	−0.028***	−0.007*
					(0.002)	(0.002)	(0.003)	(0.003)
Adj-R^2	0.373	0.372	0.268	0.090	0.423	0.429	0.314	0.092
n	1743 (for all models)							

The models are modestly improved with the quadratic term. However, the pattern is inconclusive as we examined only five dimensional forms; it remains possible that the error trend reaches a plateau of ∼0.300.

3. TAX COMPLIANCE DATA

Vogel (1974) collected data on tax compliance behavior and attitudes toward taxation from a sample of the Swedish population in 1974. For the subsequent analysis, we focus on three covariates that feature prominently in the literature of tax noncompliance: gender, self-employment, and age. Other tax compliance studies that offer margins include Houston and Tran (2001) and Mason and Calvin (1978).[26]

[25] The significance stars in our regression models follow the standard nomenclature:

$$p < \begin{cases} 0.001 & \text{if ‘***’} \\ 0.01 & \text{if ‘**’} \\ 0.05 & \text{if ‘*’} \\ 0.1 & \text{if ‘\^{}’} \end{cases}$$

[26] We select Vogel’s study as he reports a more varied set of margins and employs a larger sample size.

We highlight some research on the effects of these covariates on tax compliance.[27]

3.1. *Gender*

Being male is consistently associated with higher levels of noncompliance. Social psychological evidence points to men exhibiting a higher degree of anti-authoritarian and risk-seeking attitudes than women. Most tax evasion studies report gender effects that are consistent with this observation (Vogel 1974; Mason and Calvin 1978; Tittle 1980; Jackson and Milliron 1986; Baldry 1987; Porcano 1988; Collins, Millliron, and Toy 1992). While men have traditionally committed more crimes than women, these crimes are often borne of masculine physicality and circumstance, factors that do not necessarily contribute to tax evasion. Hence, some studies, such as those conducted by Houston and Tran (2001) and Friedland, Maital, and Rutenberg (1978), find a reverse trend; however, their effect and sample sizes are too small for the claim to be significant.

3.2. *Self-Employment*

Self-employment offers nonwithholding income and, consequently, additional opportunities to evade taxes, so it is no surprise that the evasion rate for self-employed individuals is consistently higher than that for those who are not. Since reports of this observation appear as margins (Vogel 1974; Houston and Tran 2001; Schuetze 2002) and both significant and insignificant model predictions (Aitken and Bonneville 1980; Groenland and van Veldhoven 1983; Porcano 1988; Andreoni et al. 1998; Wahlund 1992; Slemrod et al. 2001), the finding is inconclusive and further complicated by self-employment's moderate association with other sociodemographics, particularly gender and age. Furthermore, risk-seeking individuals, who are more likely to evade taxes,

[27] While typical tax compliance studies include additional sociodemographic and nonsociodemographic covariates, we focus on these three covariates as they are commonly reported (or included in models) and allow us to maintain our goal of providing logit models amenable to meta-analysis.

might also be more likely to be self-employed; this linkage has not been sufficiently explored in the literature.

3.3. *Age*

On the other hand, increasing age has a diminishing effect on noncompliance, allegedly due to increasing conservatism as well as risk aversion and increased perception of risk (Mason and Calvin 1978; Wahlund, 1992). Furthermore, a host of studies, across several decades, all report an overall diminishing effect for age on noncompliance (Vogel 1974; Mason and Calvin 1978; Houston and Tran 2001; Friedland et al. 1978; Baldry 1987; Jackson and Milliron 1986; Andreoni, Erard, and Feinstein 1998; Ritsema, Thomas, and Ferrier 2003). Jackson and Milliron (1986) offer that generational and life cycle differences may be responsible for some of the inconsistent findings, and Porcano (1988) and Collins, Milliron, and Toy (1992) report no significant effect due to age, when controlling for attitudes and personality traits. While taxpayers might gain further knowledge about the tax system and, potentially, ways to evade as they age, increased income also becomes a factor in compliance (Mason and Calvin 1978), although some evidence points to a nonlinear pattern (Jackson and Milliron 1986).

3.4. *Association Among Sociodemographic Variables*

In addition to IPF estimated contingency tables, we examine the predictive capabilities of a large proxy sample—namely, an $n = 10{,}000$ nationally representative subsample of the U.S. Year 2000 Census' Public Use Micro Sample (PUMS) data.[28, 29] There is broad agreement that these sociodemographic traits are mildly correlated.[30] Men, on average,

[28] We drew our subsamples from the 5% PUMS using person-level weights.

[29] We use the U.S. 2000 Census largely because of convenience as it is employed in a larger project of which this work is a part. While data from an earlier and/or Swedish census would be more appropriate, the use of the census is illustrative and does not detract from the importance of the alternative combinatoric approach, which relies only moderately on census data.

[30] For consistency with the margins we analyze later, the "Age" covariate here has been condensed into three categories, 20–29, 30–59, and 60–70; we exclude individuals whose ages fall outside those three groups.

are more likely to be self-employed than women (according to both U.S. and Swedish sources) or are more likely to have nonwithholding jobs (Mason and Calvin 1978), all of which impact noncompliance.[31] These interrelationships between sociodemographic traits require one to consider them concurrently in any inference exercise. Gender is often a prominent predictor even in those models that include attitudes and personality traits. An inspection of the Pearson correlation coefficients and χ^2 statistics for our census sample reveals this to be the case:[32]

	Pearson correlation ρ		χ^2 statistic	
	Gender	Age	Gender	Age
Age	−0.012***		1.68	
Self-employment	0.104*	−0.075***	107.15***	81.18***

While the associations are nominal, two exhibit significance at the $p < 0.001$ level for our large sample.[33] The χ^2 test mitigates the relationship between gender and age, which is not surprising as the age structure differences are subsumed by our coarse condensation of age.[34]

[31] Several reports point to over 2.5 times more Swedish men than women being self-employed throughout 2000–2010, which is greater than the 1.9 ratio we find in our PUMS subsample (Eklund and Vesju 2008; Brunk and Andersson 2009). Also, the definition of self-employment in Sweden appears comparable to that in the United States (Brunk and Andersson 2009).

[32] We employ this sample throughout the rest of the paper. We also argue that these correlations of the U.S. census population do not substantially differ with the Swedish population.

[33] Incidentally, the correlation coefficients stop exhibiting significance at a subsample level of approximately $n = 500$, which is not surprising for low correlations.

[34] Also, if we reduce our sample to $n = 1000$ (which is the size of the margin sample we will analyze), the χ^2 maintain significance, with the age and S.E. relationship reduced to $\chi^2 = 8.08, p < 0.05$:

x_1	x_2	χ^2
Gender	Age	0.21
Gender	S.E.	10.91**
Age	S.E.	8.08*

4. INFERRING LOGIT MODELS FROM EMPIRICAL MARGINS

We explore logit model inference for the 2 × 2 (gender by self-employment), 2 × 3 (gender by age), and 2 × 3 × 2 (gender by age by self-employment) cases.[35]

4.1. *2 × 2 Analysis*

For our first inference, we estimate a logit model for the following margins reported by Vogel (1974):

Covariate	Categories/Bins	% Noncompliant	n^{Vogel}	$\|\|n\|\|$
Gender	Female = 0	21.7	506	0.416
	Male = 1	32.3	709	0.584
Self-employment (S.E.)	No = 0	27.9	967	0.901
	Yes = 1	37.1	106	0.099

For comparison of the sample covariate margins, we present the normalized row and column margins of the initial PUMS as well as the PUMS if we condition on row (gender) and column (self-employment) separately:[36]

	Row/Gender		Column/S.E.		Raw	Norm
Data	0	1	0	1	MSE	MSE
Vogel	0.416	0.584	0.901	0.099	–	–
PUMS	0.504	0.496	0.904	0.096	0.1241	0.0621
Row conditioned	0.416	0.584	0.899	0.101	0.0038	0.0019
Column conditioned	0.509	0.491	0.901	0.099	0.1234	0.0622

The MSE columns allows us to measure the similarity of each proxy distribution to the Vogel margins. The raw MSE is directly calculated from a vector form by concatenating the normalized row and column margins:

[35] We introduce our approach incrementally for pedagogical reasons. The procedure for obtaining the combinatoric models becomes more complicated with the increasing size and number of dimensions, and we feel the reader would be better served by being exposed to the less complicated cases first.

[36] We restrict the age range to 20–70, as reflected in Vogel’s data.

$$\mathrm{MSE} = \sqrt{\sum [(v_{0\cdot}, v_{1\cdot}, v_{\cdot 0}, v_{\cdot 1}) - (x_{0\cdot}, x_{1\cdot}, x_{\cdot 0}, x_{\cdot 1})]^2}$$

where v refers to the normalized Vogel margins and x refers to the normalized margins of a proxy PUMS table. The normalized MSE divides the raw by the maximum possible MSE, which is $\sqrt{1+1+1+1} = 2$ for the 2 × 2 case. The PUMS as a source of comparison can be relevant if we maintain the belief that it reflects the true sampling distribution better than the IPF, and we have no other option for choice of proxy.

We observe clear differences between Vogel's sample and our unconditioned PUMS sub-sample. The head of a household who also files tax returns tends to be male, which would explain why Vogel's data exhibits an overrepresentation of males. This suggests row-conditioned (i.e., conditioned on gender) treatment of our PUMS sample might be appropriate. In fact, the low MSE suggests that the relationships between gender and self-employment are similar across the Vogel and PUMS samples despite the differences in their sampling pools. Given the high MSE for the column-conditioned PUMS, we can probably dismiss the notion that Vogel's data are biased by self-employment instead of gender.

As Vogel's margins sum to two different sample sizes of 1215 and 1073, due to missing data, and for the sake of simplicity, we will proceed with our analysis by assuming a sample size of $n = 1000$ applied over the normalized row and column sums. Also, for consistency, we impose this sample size on the PUMS proxies as well, which will continue to exhibit the normalized margins obtained from the 10K sample. The IPF and the IPF ratio methods require the sample and response margins to be consistent; hence, we must account for the missing data, which we detail in Appendix B.

In order to explore the combinatoric space of contingency table solutions that satisfy the margins, we can express the constraints on the cell values to the following integer programming problem:

$$\begin{aligned}
x_{00} + x_{01} \qquad\qquad\quad &= 416\\
x_{10} + x_{11} &= 584\\
x_{00} + \qquad x_{10} \qquad\quad &= 901\\
x_{01} + \qquad x_{11} &= \ \ 99,
\end{aligned}$$

where all values $\boldsymbol{x} \geq 0$. This reduces to the following equations with one unknown:

$$\begin{aligned} x_{10} &= -x_{11} + 584 \\ x_{01} &= -x_{11} + 99 \\ x_{00} &= x_{11} + 317, \end{aligned}$$

where the sole constraint is $0 \leq x_{11} \leq 99$.

We can now enumerate all 99+1 combinatoric tables that fulfill the marginal totals.[37] We compare each combinatoric table $\boldsymbol{x}$ to the IPF-derived contingency table, the row and column conditioned PUMS (R and C), and the initial PUMS (0) with MSE; the last three have been normalized to a sample size of 1000 (i.e., their normalized tables have been multiplied by 1000). We also compare, with MSE, the logit coefficients ($\boldsymbol{\beta}$) produced by marginal analysis on each of the proxies and the empirical conditional margins to those of coefficients borne from each combinatoric table ($\boldsymbol{\beta}_{\boldsymbol{x}}$). The beta-binomial likelihood function is parameterized as follows:[38]

$$\begin{aligned} p_{00} &= 0.217, & x_{00} &= 506 \\ p_{01} &= 0.323, & x_{01} &= 709 \\ p_{10} &= 0.279, & x_{10} &= 967 \\ p_{11} &= 0.371, & x_{11} &= 106 \end{aligned}$$

For illustrative purposes, we display the first few and last few combinatoric distributions along with the MSE results in Table 1.

The labels “R” and “C” refer to the row- and column-conditioned PUMS as the proxy distribution and “0” refers to the initial 10K PUMS sample. As an inquiry into internal consistency, we highlight those combinatoric solutions that minimize the MSE between them and each of the proposed proxy distributions; we also highlight

[37] For more detailed investigation into bounds on constrained tables, see Dobra and Fienberg (2001) and Dobra et al. (2003).

[38] While we normalize the size of the candidate proxy table to 1000 so that the constraint equations are more readable, we retain the empirical sample counts for accuracy in the fit. This disparity does not have any noticeable consequence on the final logit models.

TABLE 1
First and Last Three Combinatoric Results

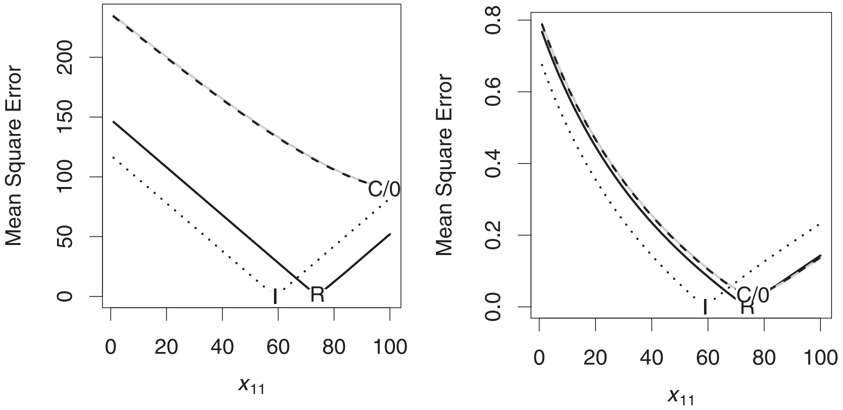
Combinatoric Table $\boldsymbol{x}$				Mean Square Error Between $\boldsymbol{x}$ or $\boldsymbol{\beta}_x$ and...							
x_{00}	x_{01}	x_{10}	x_{11}	IPF	β_{IPF}	R	β_{R}	C	β_{C}	0	β_0
317	99	584	0	116	0.679	146.0	0.764	146.0	0.764	234.2	0.788
318	98	583	1	114	0.657	144.0	0.743	144.0	0.743	232.3	0.766
319	97	582	2	112	0.636	142.0	0.722	142.0	0.722	230.5	0.745
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
414	2	487	97	78	0.222	48.1	0.136	48.1	0.136	91.0	0.128
415	1	486	98	80	0.228	50.1	0.141	50.1	0.141	90.5	0.133
416	0	485	99	82	0.233	52.1	0.146	52.1	0.146	90.1	0.138

those solutions that minimize the error of the model coefficients, as shown below:

Combinatoric Table $\boldsymbol{x}$				Mean Square Error Between $\boldsymbol{x}$ or $\boldsymbol{\beta}_x$ and							
x_{00}	x_{01}	x_{10}	x_{11}	IPF	β_{IPF}	R	β_{R}	C	β_{C}	0	β_0
375	41	526	58	**0**	**0.000**	30.1	0.091	134.7	0.112	134.0	0.111
390	26	511	73	30	0.092	**2.0**	**0.002**	113.7	0.037	113.1	0.036
416	0	485	99	82	0.233	52.0	0.143	**90.2**	0.136	**90.1**	0.138
392	24	509	75	34	0.104	4.5	0.013	111.2	**0.035**	110.6	**0.034**

A **boldfaced** MSE score points to the combinatoric table, which offers the minimum MSE for the associated column. Of course, we expect to find a perfect, or near-perfect, minimum for IPF since an IPF estimated table exists within the constraints of any set of margins. While the minimum row-conditioned PUMS (R) offers almost-zero minimal MSEs, the associated table differs significantly from the IPF solution (MSE = 30.1), suggesting they are not interchangeable and one's choice model must be selected with care.

The column conditioned PUMS and the initial PUMS samples are similar (normalized marginal MSE is 0.0041), so we observe identical combinatoric solutions. However, the minimal MSE solutions for table and model coefficients differ, indicating that the MSE function for tables and coefficients have divergent solution topologies especially when the proxy has no close combinatoric solutions.



(a) MSE of Data Distribution

(b) MSE of β Coefficients

FIGURE 1. MSE of IPF, $\text{PUMS}_{\text{gender}}$ and PUMS_0. For the range of possible values of x_{11}, we display in (a) the mean square error between the distribution $\boldsymbol{x}$ and the IPF (I, dotted), gender-conditioned PUMS (R, solid black), self-employment-conditioned PUMS (C, solid gray), and nonconditioned PUMS ("0", dashed). In (b), we display the MSE between the β"s of each combinatoric logit model and of the marginal analysis of the IPF, the conditioned and unconditioned PUMS. The raw PUMS (0) and self-employment (C) curves overlay one another.

Still, we need to consider the empirical association between "gender" and "self-employment". The correlation coefficient from an IPF-derived table will be almost zero, $\rho_{\text{IPF}} = 0.001$, while the row-conditioned PUMS and its minimal combinatoric solution exhibit higher correlations: 0.100 and 0.103 respectively. The χ^2 statistics are respectively 0.002 (IPF), 10.22 (R), and 10.64 (combinatoric), with the PUMS R and combinatoric maintaining nearly the same association level shown in the unconditioned PUMS subsample of equivalent size. While the exact association for the Swedish population was unavailable to us at the time of writing, some investigation points to the correlation lying closer to the PUMS than IPF given the claim that significantly more men than women are self-employed.

Figure 1 shows how each of the MSEs vary across the space of joint distributions. We see that the MSEs can become appreciably large relative to their respective minima. The MSEs of the coefficients in particular span a range large enough to warrant concern over selecting the correct proxy data and, consequently, being accurate about the association among the covariates.

In Table 2, we present the logit models that have been estimated using the IPF, the row and column conditioned PUMS, and the initial PUMS.[39] We also perform a weighted logit on IPF_R in which the conditional margin is incorporated into the IPF estimation, with the response and nonresponse tables obtained separately through IPF.

We introduce and then aggregate the "All" models in which we estimate model coefficients from each of the combinatoric tables through MA. In combining the combinatoric models, we compute the mean of each covariate coefficient and, for the standard errors, we combine both within- and between-solution variance.[40,41] The "All_M" model fits each solution to self-employment margins augmented for missing data.[42] "All_R" is similar to IPF_R but examines combinations of both the response and nonresponse tables; however, this approach also requires consistent margins (i.e., missing data augmentation).[43]

Finally, in "All_W," we enhance the combinatoric approach by weighting each model's estimates by how well each combinatoric table reflects some weighted empirical evidence—namely, the row-conditioned PUMS as its margin MSE with the Vogel margins its

[39] For comparison, we offer those models based on combinatoric tables closest to each treatment of the PUMS data in Appendix C.

[40] We employ the approach of Gelman et al. (2003) to combining the variance across of M sample sets:

$$T = \frac{(n-1)}{n} W + \frac{1}{n} B,$$

where M is number of estimations and the between-sample variation is

$$B = \frac{n}{M-1} \sum_{i=1}^{M} \left(\beta_i - \overline{\beta}\right)^2$$

and the within-sample variation is

$$W = \frac{1}{M} \sum_{i=1}^{M} \sigma_i^2.$$

These errors are nearly identical to those obtained by multiple imputation.

[41] It would be inappropriate to apply meta-analytic fixed- or random-effects estimation of the combined effects as each solution is not new evidence.

[42] See Appendix B for details on missing data augmentation.

[43] In this paper, we infer an aggregated model from the combinatoric response/nonresponse tables for only the 2 × 2 case.

TABLE 2
Logit Models for Gender × Self-Employment (2 × 2)

X	IPF	R	C	0	$\mathrm{IPF_R}$	All	$\mathrm{All_M}$	$\mathrm{All_R}$	$\mathrm{All_W}$
I	-1.328^{***} (0.112)	-1.284^{***} (0.108)	-1.252^{***} (0.101)	-1.251^{***} (0.101)	-1.328^{***} (0.122)	-1.349^{***} (0.149)	-1.370^{***} (0.149)	-1.361^{***} (0.136)	-1.322^{***} (0.132)
G	0.545^{***} (0.149)	0.522^{***} (0.150)	0.505^{***} (0.149)	0.505^{***} (0.149)	0.545^{***} (0.148)	0.596^{**} (0.188)	0.596^{**} (0.187)	0.586^{***} (0.163)	0.566^{***} (0.171)
S	0.415 (0.226)	0.345 (0.226)	0.338 (0.227)	0.339 (0.227)	0.417 (0.224)	0.517 (0.330)	0.505 (0.331)	0.482^{*} (0.277)	0.449 (0.293)
$\mathcal{L}$	11.3	11.3	11.0	11.0	−591.4	–	–	–	–
n	1000 (for all models)								

Note: The predictors are I = Intercept, G = Gender and S = Self-Employment. The model are estimated from the following proxy data:

1. IPF
2. Row (R)
3. Column-conditioned PUMS (C)
4. 10K PUMS (0)
5. IPF ratio ($\mathrm{IPF_R}$)
6. a. All of the combinatoric tables (All)
   b. Augmented for the missing self-employment data ($\mathrm{All_M}$)
   c. Ratio combinatoric in which the response and nonresponse tables are independently solved ($\mathrm{All_R}$)
   d. Combinatoric weighted by the Pólya probabilities (All_W)

lowest.[44] We employ the Pólya distribution to measure that fit:[45]

$$\mathcal{L} = \log\left[\text{Pólya}\left(\boldsymbol{x}_i \middle| \kappa \cdot \boldsymbol{x}_{\text{R}}^{\text{PUMS}} + 1\right)\right], \tag{2}$$

where i is an index to one of the combinatoric solutions (i.e., $\boldsymbol{x}_i$), $\boldsymbol{x}_{\text{R}}^{\text{PUMS}} = (389, 27, 510, 74)$, and $\kappa = 0.01$, our belief in the relevance of the row-conditioned PUMS to the Swedish sample. Despite the similarity in the margins of the Vogel and the row-conditioned PUMS, we acknowledge that the 2000 PUMS and the sample of the 1974 Swedish population are different enough to warrant a low weight on the PUMS as evidence.[46]

We easily convert the log-likelihoods into probabilistic weights:

$$w_i = \exp(\mathcal{L}_i - \mathcal{L}_{\max}).$$

In effect, we weight those combinatoric tables that resemble the row-conditioned PUMS slightly higher than those that do not. While other measures of association, such as the correlation coefficient, are applicable to the 2 × 2 case, the Pólya approach is more easily interpretable and applicable to higher dimensional cases.

[44] The Gelman approach for aggregated model coefficients, described in an earlier footnote, is adjusted so that $\overline{\beta}$ is now a weighted mean as is the mean within-sample variance W. Furthermore, the between-sample variation becomes weighted variance.

[45] The multivariate Pólya distribution is a Dirichlet prior on a multinomial and is the multivariate analogue of the beta-binomial. Essentially, we obtain the probability of some data $\boldsymbol{n}$ arising from the probability distributions specified by the Dirichlet parameterized with evidence, $\boldsymbol{x}$, in this case the κ weighted, row-conditioned PUMS. The density of the Pólya is

$$\Pr(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{n!}{\prod_k (n_k!)} \frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma\left(n + \sum_k \alpha_k\right)} \prod_k \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)},$$

where Γ is the gamma function, n_k is the number of items of $\boldsymbol{x}$ in group k, and $n = \sum_k n_k$.

[46] We can alternatively employ the Swedish self-employment gender ratio in constructing an evidence table, but we save that for future investigation.

In all of the models, “gender” remains a prominent and significant predictor whereas the significance and effect size of “self-employment” wavers, confirming those findings that self-employment’s involvement in tax compliance is not a foregone conclusion as we might intuitively expect. The models based on single proxy naturally display narrower coefficient standard errors than the aggregated models. However, we notice that the errors of the ratio combinatoric model, “All_{R}”, exhibit shrinkage from the other “All” models. We would expect further shrinkage if we were to combine both the combinatoric ratio and evidence-weighted approaches.

4.2. *2 × 3 Analysis*

To demonstrate how larger dimensionality complicates the combinatoric approach, we extend our analysis to the covariate pair “gender” × “age” whose dimensions are 2 × 3. Vogel’s age margin is expressed with five categories, which we condense into three.

Covariate	Categories	% Evasion	n	Combined % Categories		Combined % Evasion	n	$\|\|n\|\|$
Age	20–29 = 0	38.8	288	23.7 = 0			288	0.237
	30–39 = 1	31.5	230	18.9	30–59 = 1	30.4	488	0.401
	40–59 = 2	29.5	258	21.2				
	60–69 = 3	19.6	226	18.6	60+ = 2	18.1	440	0.362
	70+ = 4	16.5	214	17.6				

The normalized row and column margins and MSE scores are shown below.

	Row/Gender		Column/Age			Raw	Norm
Data	0	1	0	1	2	MSE	MSE
Vogel	0.416	0.584	0.237	0.401	0.362	–	–
PUMS	0.504	0.496	0.218	0.486	0.296	0.1661	0.0678
Row conditioned	0.416	0.584	0.219	0.486	0.295	0.1096	0.0447
Column conditioned	0.504	0.496	0.237	0.401	0.362	0.1247	0.0510

While the row-conditioned PUMS remains superior, its advantage is only slight largely because there is only a small association between gender and age. We suspect survey bias on both gender and age to be largely responsible for the observed margin differences rather than a significant difference in the population age structure.[47] We again assume a sample size of $n = 1000$, and the equations simplify to the following:[48]

$$\begin{aligned} x_{00} &= x_{11} + x_{12} - 347 \\ x_{01} &= -x_{11} + 401 \\ x_{02} &= - x_{12} + 362 \\ x_{10} &= -x_{11} - x_{12} + 584, \end{aligned}$$

with the free parameters being $\{x_{11}, x_{12}\}$ and, again, subject to the constraint that all values $\boldsymbol{x} \geq 0$. We find 69,438 total solutions and obtain the minimal MSE results shown in Table 3.

In this case, the rounded IPF produces margins that do not coincide with a combinatoric solution—hence, the nonzero MSE. The correlation coefficients for IPF, R, and C are respectively −0.014, −0.024, and −0.043, while the empirical PUMS correlation is −0.012; these imply that the IPF-derived table here is the best candidate for model inference. Upon inspecting the χ^2, we obtain 0.008 (IPF), 0.112 (R), and 0.177 (C), of which the last is closest to the PUMS statistic 0.21.

None of the PUMS proxies are close to any of the combinatoric tables, with the closest being column-conditioned PUMS (C). Despite that, we find parity in the model coefficients $\boldsymbol{\beta}$ across the board, with the minima β MSEs all less than 0.05. However, we suspect the lack of association between the covariates renders the inference process insensitive to the structure in the table. The similarity of models is confirmed in Table 4, in which the effect sizes are more consistent than those shown in Table 2. We observe that "age" remains a consistently significant predictor, while the effect of "gender" is mitigated only in the unweighted combinatoric models.

[47] The 1974 Swedish population data were unavailable at the time of writing. However, the margins for the 2000 Swedish population are gender then age concatenated (0.512, 0.488, 0.163, 0.544, 0.293), and they resemble the 2000 U.S. population (MSE = 0.0808) more so than the Vogel margins (MSE = 0.2216).

[48] Refer to Appendix A for details on the simplification.

TABLE 3
MSE Results for Gender × Age (2 × 3)

Combinatoric Table $\boldsymbol{x}$			Mean Square Error Between $\boldsymbol{x}$ or $\boldsymbol{\beta}_x$ and...							
x_{00}	x_{01}	x_{02}								
x_{10}	x_{11}	x_{12}	IPF	β_{IPF}	R	β_{R}	C	β_{C}	0	β_0
98	167	151	**1.4**	0.003	78.8	0.014	75.0	0.037	107.2	0.034
139	234	211								
98	169	149	4.2	**0.000**	79.5	0.017	75.0	0.039	107.2	0.036
139	232	213								
97	161	158	13.3	0.019	**77.7**	0.010	76.1	0.031	108.1	0.024
140	240	204								
106	139	171	49.7	0.014	86.4	**0.007**	93.9	0.034	121.9	0.029
131	262	191								
87	174	155	20.4	0.050	81.5	0.039	**72.7**	0.034	105.5	0.023
150	227	207								
83	217	116	89.2	0.036	127.6	0.036	110.6	**0.013**	133.5	0.012
154	184	246								
87	174	155	20.4	0.050	81.5	0.039	72.7	0.034	**105.5**	0.023
150	227	207								
84	204	128	65.2	0.036	109.2	0.032	93.2	0.016	119.8	**0.007**
153	197	234								

4.3. *2 × 3 × 2 Analysis*

Our final analysis combines the three covariates into a single model. The normalized margins are shown below.

	Row/Gender		Column/S.E.			Slice/Age		Raw	Norm
Data	0	1	0	1	2	0	1	MSE	MSE
Vogel	0.416	0.584	0.237	0.401	0.362	0.901	0.099		
PUMS	0.504	0.496	0.218	0.486	0.296	0.904	0.096	0.1661	0.0480
Conditioned on ...									
row	0.416	0.584	0.219	0.486	0.295	0.899	0.101	0.1096	0.0317
column	0.504	0.496	0.237	0.401	0.362	0.905	0.095	0.1248	0.0360
slice	0.504	0.496	0.217	0.487	0.296	0.901	0.099	0.1656	0.0478

The row-conditioned PUMS maintains only a slight advantage over column-conditioning.

Again, we employ Gaussian elimination to determine the unknown variables.[49] Since the solution space is now greatly expanded,

[49] See Appendix A for further details.

TABLE 4
Logit Models for Gender × Age (2 × 3)

Predictor	IPF	R	C	0	IPF_R	All	All_M	All_W
Intercept	-0.726^{***} (0.141)	-0.736^{***} (0.141)	-0.696^{***} (0.138)	-0.705^{***} (0.137)	-0.728^{***} (0.151)	-0.666^{***} (0.174)	-0.663^{***} (0.175)	-0.696^{***} (0.158)
Gender	0.563^{***} (0.154)	0.549^{***} (0.152)	0.540^{***} (0.152)	0.534^{***} (0.151)	0.569^{***} (0.151)	0.592 (0.444)	0.593 (0.447)	0.582 (0.315)
Age	-0.539^{***} (0.096)	-0.529^{***} (0.095)	-0.529^{***} (0.096)	-0.527^{***} (0.095)	-0.539^{***} (0.095)	-0.607^{**} (0.233)	-0.613^{**} (0.234)	-0.575^{***} (0.166)
$\mathcal{L}$	13.8	13.8	13.7	13.5	−568.4	–	–	–
n	1000 (for all models)							

especially for the sample count of $n = 1000$, we resort to a sampling of the combinatoric space instead of enumerating it. First, we sample a solution for the entire self-employment 2×2 slice by using a uniform Dirichlet prior on a multinomial:[50]

$$\boldsymbol{q}_{\cdot\cdot 1} \sim \text{Dirichlet}\,(\boldsymbol{\alpha} = (1, 1, 1, 1, 1, 1))$$
$$\boldsymbol{x}_{\cdot\cdot 1} \sim \text{Multinomial}\,(n = 99; \boldsymbol{p} = \boldsymbol{q}_{\cdot\cdot 1}),$$

where each draw $\boldsymbol{x}_{\cdot\cdot 1} = (x_{001}, x_{011}, x_{021}, x_{101}, x_{111}, x_{121})$. Given the solution for $\boldsymbol{x}_{\cdot\cdot 1}$, we then sample the remaining variables for a solution for the partial row $\boldsymbol{x}_{1\cdot 0}$ subject to

$$x_{100} + x_{110} + x_{120} = r_1 - (x_{101} + x_{111} + x_{121}),$$

where the row sum is $r_1 = \sum x_{1\,..}$ The sampling process is similar to the one shown above:

$$\boldsymbol{q}_{1\cdot 0} \sim \text{Dirichlet}(\boldsymbol{\alpha} = (\alpha_{100}, \alpha_{110}, \alpha_{120}) = (1, 1, 1))$$
$$\boldsymbol{x}_{1\cdot 0} \sim \text{Multinomial}(n = r_1 - (x_{101} + x_{111} + x_{121}); \boldsymbol{p} = \boldsymbol{q}_{1\cdot 0}).$$

With $\boldsymbol{x}_{\cdot\cdot 1}$ and $\boldsymbol{x}_{1\cdot 0}$, we obtain the rest of the sampled contingency table. However, due to the additional constraints of all values to be greater than or equal to 0 (i.e., $\boldsymbol{x} \geq 0$), we are not guaranteed a valid solution.[51] Instead of constraining our sampling strategy further, we simply reject all those solutions that do not satisfy the constraint. Out of 200,000 drawn samples of $\{\boldsymbol{x}_{1\cdot 0}, \boldsymbol{x}_{\cdot\cdot 1}\}$, we obtain 74,382 valid solutions.[52] The minimal MSE results apper in Table 5.[53]

[50] In fact, this is the equivalent of drawing random variates from an uninformed Pólya.

[51] Using the basis approach for finding unique linear algebraic solutions incurs the same negative value issue.

[52] Even if each of the free variables took on as few as 10 values, we would still be looking at 10 million tables. Furthermore, the enumeration of tables subject to constraints is not a trivial issue. Some early work on the enumeration of margin constrained tables was done by Gail and Mantel (1977). For improving the sampling of constrained tables, we would investigate work by Diaconis and Sturmfels (1998) and also the generalized shuttle algorithm, in Dobra et al. (2003), which can incorporate additional constraints.

[53] Due to space constraints, we omit the minimum MSE combinatoric tables.

TABLE 5
MSE Results for Gender × Age × Self-Employment (2 × 3 × 2)

IPF	β_{IPF}	R	β_{R}	C	β_{C}	S	β_{S}	0	β_0
0.0	**0.000**	72.5	0.062	82.5	0.122	106.1	0.106	106.7	0.105
19.1	0.050	63.8	0.017	76.6	0.074	97.2	0.059	97.7	0.059
196.6	**0.009**	204.4	0.066	227.8	0.123	237.7	0.107	237.5	0.107
41.0	0.092	**58.3**	0.085	83.3	0.115	93.6	0.113	94.1	0.116
147.3	0.055	160.0	**0.013**	176.8	0.070	192.8	0.056	192.9	0.056
58.3	0.113	79.0	0.099	**58.7**	0.108	87.0	0.089	87.9	0.086
183.1	0.118	198.9	0.060	194.4	**0.015**	197.9	0.015	198.6	0.018
52.9	0.264	73.0	0.277	61.9	0.292	**85.6**	0.274	85.9	0.271
287.2	0.110	304.6	0.055	295.7	0.023	299.6	**0.008**	300.2	0.008
52.9	0.264	73.0	0.277	61.9	0.292	85.6	0.274	**85.9**	0.271
311.0	0.103	327.2	0.049	318.9	0.029	322.2	0.008	322.7	**0.005**

Since we are unable to explore the entire space of solutions, the above minima are estimates, with the exception of the first line, in which we compare the actual IPF solution to the rest. Interestingly, the additional dimension produces MSEs for the nearest R and C, which are lower than those for 2 × 3, both normalized and absolute. Given that the normalized margin MSE for 2 × 3 is worse than the MSE for this and the 2 × 2 cases, we again suspect that the structural similarity we gain through conditioning on gender while including self-employment as a covariate is responsible.

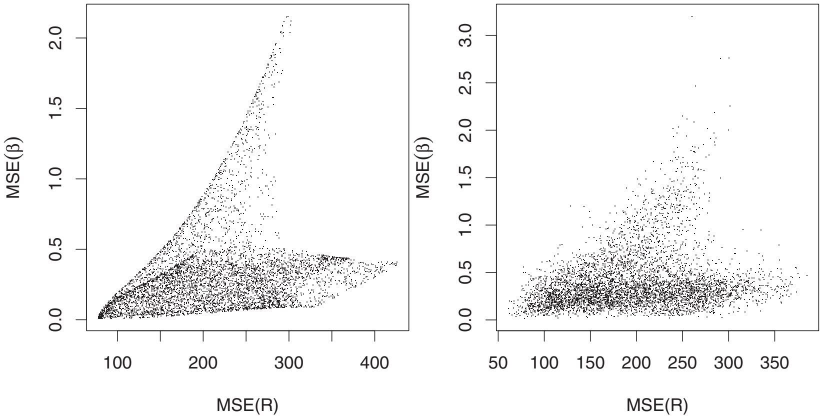
Still, the minima are far from ideal. The difference between the real IPF MSE and the sampled MSE suggests that closer solutions to R, C, and 0 also exist. Disparities between solutions associated with minima proxies and the minima $\boldsymbol{\beta}$ MSEs confirm our belief that the relationship between the $\boldsymbol{\beta}$ MSEs and proxy MSEs is complicated. We confirm this in Figure 2. One region exhibits some correlation while, in another denser region, the MSE of the coefficients is insensitive to the MSE of the proxy.

In Table 6, we present the logit models. As we would expect, the coefficients for weighted combinatoric model "All$_{\mathrm{W}}$" land in between those of the unweighted combinatoric and the proxy models. The conservative finding is that "self-employment" has a comparable yet insignificant effect on tax evasion. So far, in all the cases, the combinatoric "All" models tend to stretch out the covariate coefficients such

TABLE 6
Logit Models for Gender × Age × Self-Employment (2 × 3 × 2)

		PUMSconditionedon							
X	IPF	R	C	S	0	IPF_R	All	All_M	All_W
I	-0.772^{***} (0.141)	-0.739^{***} (0.137)	-0.681^{***} (0.132)	-0.694^{***} (0.131)	-0.694^{***} (0.131)	-0.771^{***} (0.153)	-0.716^{***} (0.183)	-0.729^{***} (0.183)	-0.735^{***} (0.167)
G	0.561^{***} (0.154)	0.522^{***} (0.155)	0.501^{**} (0.154)	0.501^{**} (0.154)	0.500^{**} (0.153)	0.570^{***} (0.151)	0.619 (0.406)	0.619 (0.405)	0.599^{*} (0.300)
A	-0.536^{***} (0.096)	-0.549^{***} (0.098)	-0.550^{***} (0.098)	-0.550^{***} (0.098)	-0.548^{***} (0.098)	-0.541^{***} (0.095)	-0.603^{**} (0.218)	-0.604^{**} (0.217)	-0.576^{***} (0.161)
S	0.435 (0.235)	0.467^{*} (0.236)	0.488^{*} (0.239)	0.471^{*} (0.237)	0.468^{*} (0.236)	0.407 (0.229)	0.406 (0.382)	0.394 (0.384)	0.420 (0.344)
$\mathcal{L}$	19.2	19.2	18.9	18.8	18.8	−566.8	–	–	–
n	1000 (for all models)								

Note: The predictors are I = Intercept, G = Gender, A = Age, and S = Self-Employment.



(a) 2 × 3

(b) 2 × 3 × 2

FIGURE 2. MSEs of PUMS R Proxy by $\boldsymbol{\beta}$. We obtain $m = 5,000$ samples of the combinatoric tables and display the MSEs for the PUMS R by the PUMS R $\boldsymbol{\beta}$ for each of the 2 × 3 and 2 × 3 × 2 cases.

that their absolute effect sizes are larger than the proxy models. In this 2 × 3 × 2 case, the effect is countered by one of the covariates (i.e., self-employment). Also, an inspection of the coefficient distributions (not shown) reveals those for gender and ages to be skewed left and right, respectively, with the tails stretching into more positive values for gender and more negative for age. These tails reflect that region of highly correlated MSE PUMS R proxy and MSE $\boldsymbol{\beta}$ in Figure 2. Hence, the "All" model coefficients are influenced by those combinatoric tables that stray drastically from the proxies, particularly the PUMS R.

5. Conclusion

We explored several methods for estimating logit models from empirical margins of sociodemographic covariates and conditional responses for tax noncompliance. These methods are particularly valuable in estimating predictive models when only margins are published. Models estimated with our approaches can be employed in expansive work such as meta-analysis. Often in meta-analysis the reported models and/or margins need to be reworked and made consistent with one another in order to be combined.

The contingency table estimated from a uniformly seeded iterative proportional fitting of margins is suitable when there exists little or no association between the margin covariates. The use of a larger proxy data source, in our case a subsample of the U.S. Census PUMS can serve as an alternative proxy provided the margin proportions and covariate associations are sufficiently similar to either those of the census sample or a conditioned treatment of it.

However, both proxy approaches require a fitting procedure to the conditional response margin. For this, we fit our predicted response to the beta distribution (i.e., marginal analysis [MA]), which yields models similar to the logistic regression, of which the standard errors account for the fact we are fitting to conditional mean, and not actual, responses. On the other hand, the IPF ratio, which incorporates the conditional margins into IPF estimation, requires no such ancillary fitting as it directly produces a joint response distribution. According to the diagnostics, we can generally obtain more accurate models using an IPF ratio rather than MA, particularly if the association between the response and covariates is nonlinear.

Both the IPF+MA and the IPF ratio approaches require consistent margins, which we lack due to missing data. For IPF+MA, it is sufficient to account for just the margin sample sizes, for which the most straightforward approach is to proportionally pad those counts (i.e., normalize and scale by the desired sample size). As for inconsistent response margins, the MA approach will fit to the equivalent of a weighted mean, which is potentially problematic if the resulting response count exceeds a known limit. In order to employ the IPF ratio method (i.e., IPF over both the response and sample margins), it is necessary to impose additional assumptions while augmenting the sample and conditional response margins. We realize that, with MA, linear prediction becomes more tenuous with increasing dimensionality, the best measure of which is, in the cases we examined, the diagonal of the covariate contingency table.

Since empirical margins for discrete data explicitly constrain the space of contingency table solutions, we exploit this constraint and offer linear logit models that aggregate the coefficients from each combinatoric solution satisfying the marginal (and other) constraints. This method entails enumerating these combinatoric solutions, or sampling from them if the space is too large to enumerate. The estimated coefficients are aggregated, accounting for each model's standard errors and the between-model variance.

This approach is improved when we weight each logit model in a manner that reflects the relevance of each combinatoric table to some known empirical distribution. The multivariate Pólya provides us with an interpretable probability density for how well the proposed candidate contingency table matches some similar empirical table. Despite its being a scalar measure, the probability weight is sufficient for our purposes since all we need to know is how similar the combinatoric proxy is to our evidence. However, the weight for the evidence itself, κ, was subjectively yet conservatively determined, and a more precise manner of deciding κ is warranted.

The weighted models naturally exhibit standard errors that are constrained relative to those produced by the unweighted aggregated models, but they are still wider than those from single-proxy predictions. While our method of obtaining samples of valid combinatoric tables is generalizable, at least up to the $2 \times 3 \times 2$ case, we recognize that the rejection sampling approach will need to be optimized (1) for larger and more numerous dimensions and (2) in order to obtain those contingency table solutions that at least surround the mode of the Pólya probability (i.e., the evidence table itself). The unweighted model is the natural and conservative choice when we lack supplementary information regarding the association among the covariates.

Admittedly, we have not fully explored the ways in which the combinatoric approach can be of use to our problem. We recognize that it can be extended to the response and nonresponse tables as well as to the possible solutions for augmenting the missing data, thereby obviating the need for MA. However, this approach can be computationally intensive with increased span of covariates, even with rejection sampling from the space of possible solutions, and begs for a more efficient approach. Furthermore, our investigation reveals that an aggregated model based on separate combinatoric response and nonresponse tables incur narrower errors than the combinatoric base table + MA approach. For simplicity in our models, we have omitted interactions and nonlinear terms and anticipate examining them in future research.

Additionally, the weighted or unweighted combinatoric approach can be extended to model estimation beyond the logit. For example, for nonbinary dependent variables, we can aggregate estimates from combinatoric log-linear models in the same manner that we created the logit models.

APPENDIX A: MATRIX FORMULATION OF MARGIN EQUATIONS

For the 2 × 3 case, we have the following linear system:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{00} \\ x_{01} \\ x_{02} \\ x_{10} \\ x_{11} \\ x_{12} \end{bmatrix} = \begin{bmatrix} r_0 \\ r_1 \\ c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 416 \\ 584 \\ 237 \\ 401 \\ 362 \end{bmatrix}.$$

Employing Gaussian elimination, we obtain

$$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{00} \\ x_{01} \\ x_{02} \\ x_{10} \\ x_{11} \\ x_{12} \end{bmatrix} = \begin{bmatrix} c_0 - r_1 \\ c_1 \\ c_2 \\ r_1 \end{bmatrix} = \begin{bmatrix} -347 \\ 401 \\ 362 \\ 584 \end{bmatrix},$$

for which the unknown variables are $\{x_{11}, x_{12}\}$.

For the 2 × 3 × 2 case, the initial linear system is

$$\begin{bmatrix} 1&1&1&0&0&0&1&1&1&0&0&0\\ 0&0&0&1&1&1&0&0&0&1&1&1\\ 1&0&0&1&0&0&1&0&0&1&0&0\\ 0&1&0&0&1&0&0&1&0&0&1&0\\ 0&0&1&0&0&1&0&0&1&0&0&1\\ 1&1&1&1&1&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&1&1&1&1&1&1 \end{bmatrix} \begin{bmatrix} x_{000}\\ x_{010}\\ x_{020}\\ x_{100}\\ x_{110}\\ x_{120}\\ x_{001}\\ x_{011}\\ x_{021}\\ x_{101}\\ x_{111}\\ x_{121} \end{bmatrix} = \begin{bmatrix} r_0\\ r_1\\ c_0\\ c_1\\ c_2\\ s_0\\ s_1 \end{bmatrix} = \begin{bmatrix} 416\\ 584\\ 237\\ 401\\ 362\\ 901\\ 99 \end{bmatrix}.$$

Through elimination, we obtain

$$\begin{bmatrix} 1&0&0&0&-1&-1&0&-1&-1&-1&-2&-2\\ 0&1&0&0&1&0&0&1&0&0&1&0\\ 0&0&1&0&0&1&0&0&1&0&0&1\\ 0&0&0&1&1&1&0&0&0&1&1&1\\ 0&0&0&0&0&0&1&1&1&1&1&1 \end{bmatrix} \begin{bmatrix} x_{000}\\ x_{010}\\ x_{020}\\ x_{100}\\ x_{110}\\ x_{120}\\ x_{001}\\ x_{011}\\ x_{021}\\ x_{101}\\ x_{111}\\ x_{121} \end{bmatrix}$$

$$= \begin{bmatrix} -c_1 - c_2 - s_1 \\ c_1 \\ c_2 \\ r_1 \\ s_1 \end{bmatrix} = \begin{bmatrix} -862 \\ 401 \\ 362 \\ 584 \\ 99 \end{bmatrix}.$$

In doing so, we obtain the 7 unknown variables, each of which appears in more than one equation: $\{x_{110}, x_{120}, x_{011}, x_{021}, x_{101}, x_{111}, x_{121}\}$. The complementary set contains the solvable variables and are associated with exactly one coefficient in the Gaussian form.

APPENDIX B: MISSING VOGEL DATA

There is clearly some missing information in the data from Vogel (1974), as evidenced by the inconsistent margin totals:

$$\begin{aligned} 709 + 506 &= 1215 \text{ (male + female count)} \\ 967 + 106 &= 1073 \text{ ($\sim$self-employed + self-employed count)} \\ \Delta &= 142 \text{ (difference),} \end{aligned}$$

where "~" denotes logical negation, or "not." The missing data are also responsible for the disparate counts of evaders:

$$\begin{aligned} &(709)(0.323) + (506)(0.217) \\ &\quad = 338.809 \text{ (number of evaders according to gender)} \\ &(967)(0.279) + (106)(0.371) \\ &\quad = 309.119 \text{ (number of evaders according to self-employment)} \end{aligned}$$

Furthermore, the proportions of evaders in the sample are also incongruous:

$$p_{\text{gender}} = \frac{338.809}{1215} = 0.2788551$$

$$p_{\text{s.e.}} = \frac{309.119}{1073} = 0.2880885.$$

We thus seek to augment the missing self-employment data by first inferring the base self-employment by response contingency table:

$$\begin{array}{lll}
(967)(0.279) = 269.793 & \Rightarrow \sim \text{self-employed} & \text{evaders} \\
(106)(0.371) = 39.326 & \Rightarrow \quad \text{self-employed} & \text{evaders} \\
967 - 279 \quad = 697 & \Rightarrow \sim \text{self-employed} & \sim \text{evaders} \\
106 - 39 \quad\;\; = 67 & \Rightarrow \quad \text{self-employed} & \sim \text{evaders.}
\end{array}$$

We augment each of the self-employment categories equally to the sample count exhibited by gender:

$$\begin{aligned} 967x + 106x &= 1215 \\ x &= 1.132339. \end{aligned}$$

We then obtain the following augmented sample counts for self-employment:

$$\begin{aligned} 967x &= 1094.972 \\ 106x &= 120.028. \end{aligned}$$

This gives us the following contingency table, with the potential augmentation:

$$\begin{array}{cc|c}
 & \multicolumn{2}{c}{\text{Self-employment}} \\
\text{gender} & \hat{n}_{00} = 697 + a & \hat{n}_{01} = 67 + b \\ \hline
 & \hat{n}_{10} = 270 + c & \hat{n}_{11} = 39 + 142 - (a + b + c)
\end{array},$$

where $\{a,\ b,\ c\}$ denote the missing data, subject to the following constraints,

$$\sum \hat{n}_{\cdot 1} = 339 \text{ number of evaders according to gender}$$

$$\sum \hat{n}_{0\cdot} = 1095 \text{ augmented} \sim\!\text{self-employed count}$$

$$\sum \hat{n}_{1\cdot} = 120 \text{ augmented self-employed count,}$$

and also the presumptive constraint that the difference in evasion rates across self-employment statuses remains the same:

$$\frac{\hat{n}_{11}}{\hat{n}_{1\cdot}} - \frac{\hat{n}_{01}}{\hat{n}_{0\cdot}} = 0.371 - 0.279 = 0.092.$$

We obtain the solution

$$\begin{aligned} a &= 102.431 \\ b &= 9.56889 \\ c &= 25.5689, \end{aligned}$$

which gives us the augmented contingency table

$$\begin{array}{c} \text{S.E.} \\ \text{Gender}\ \begin{array}{c|c} 799.43100 & 76.56889 \\ \hline 295.56890 & 43.43121 \end{array} \xrightarrow{\text{Rounded}} \begin{array}{c|c} 799 & 77 \\ \hline 296 & 43 \end{array}. \end{array}$$

The margins and response rates for self-employment are shown below.

	Marginal Counts		Evasion Rates	
Sample	~Self-employed	Self-employed	~Self-employed	Self-employed
Original	957	106	0.279	0.371
Augmented	1095	120	0.2703196	0.3583333

Note that an alternative, minimal *ad hoc* adjustment, targeting just the proportion of evaders by adding 3 and 32, yields the roughly same results.

APPENDIX C: COMBINATORIC TABLE MODELS BASED ON PUMS

We present models affiliated with those combinatoric tables most similar to the conditioned and unconditioned PUMS distributions. Interestingly, conditioning on age or self-employment (C or S) yields distributions whose combinatoric solution is shared by the initial PUMS (0):

	2×2		2×3		2×3×2		
Predictor	R	C/0	R	C/0	R	C	S/0
Intercept	−1.283***	−1.260***	−0.726***	−0.718***	−0.752***	−0.792***	−0.676***
	(0.108)	(0.108)	(0.143)	(0.147)	(0.106)	(0.139)	(0.116)
Gender	0.522***	0.508**	0.546***	0.517***	1.249***	0.650***	0.513***
	(0.151)	(0.165)	(0.153)	(0.153)	(0.167)	(0.099)	(0.101)
Age			−0.526***	−0.517***	−0.933***	−0.560***	−0.576***
			(0.096)	(0.095)	(0.167)	(0.099)	(0.101)
S.E.	0.340	0.201			0.478	0.498*	0.348
	(0.226)	(0.247)			(0.247)	(0.244)	(0.231)

The “R” model of the 2 × 3 × 2 case has unusually large effect sizes. We suspect that either the combinatoric sample needs to be expanded or those solutions that are similar to the row-conditioned PUMS need to be explicitly sought.

REFERENCES

Aitken, Sherie S., and Laura Bonneville. 1980. *A General Taxpayer Opinion Survey*. Prepared for Office of Planning and Research, Internal Revenue Service. Washington, DC: CSR, Inc.

Andreoni, James, Brian Erard, and Jonathan Feinstein. 1998. “Tax Compliance.” *Journal of Economic Literature* 36:818–60.

Baldry, Jonathan C. 1987. “Income Tax Evasion and the Tax Schedule: Some Experimental Results.” *Public Finance* 42:357–83.

Bartholdy, Kasper. 1991. “A Generalization of the Friedlander Algorithm for Balancing of National Accounts Matrices.” *Computer Science in Economics and Management* 4:165–74.

Beckman, Richard J., Keith A. Baggerly, and Michael D. McKay. 1996. “Creating Synthetic Baseline Populations.” *Transportation Research Part A: Policy and Practice* 30:415–29.

Bishop, Yvonne M., Stephen E. Fienberg, and Paul W. Holland. 2007. *Discrete Multivariate Analysis: Theory and Practice*. New York: Springer.

Brunk, Thomas, and Paul Andersson. 2009. "Sweden: Self Employed Workers." Technical Report SE0801019Q, Oxford Research.

Causey, Beverley D. 1984. "Estimation Under Generalized Sampling of Cell Proportions for Contingency Tables Subject to Marginal Constraints." *Communications in Statistics—Theory and Methods* 13:2487–94.

Collins, Julie H., Valerie C. Milliron, and Daniel R. Toy. 1992. "Determinants of Tax Compliance: A Contingency Approach." *Journal of the American Taxation Association* 14:1–29.

Deming, W. Edwards, and Frederick F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." *Annals of Mathematical Statistics* 11:427–44.

Diaconis, Persi, and Bernd Sturmfels. 1998. "Algebraic Algorithms for Sampling from Conditional Distributions." *Annals of Statistics* 26:363–97.

Dobra, Adrian, and Stephen E. Fienberg. 2001. "Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals with Applications to Disclosure Limitation." *Statistical Journal of the United Nations ECE* 18:363–71.

Dobra, Adrian, Alan F. Karr, and Ashish P. Sanil. 2003. "Preserving Confidentiality of High-Dimensional Tabulated Data: Statistical and Computational Issues." *Statistics and Computing* 13:363–70.

Eklund, Stina, and Altin Vesju. 2008. "Incentives to Self-Employment Decision in Sweden: A Gender Perspective." Technical Report SE-103 33, Ministry of Enterprise, Energy and Communications, Stockholm, Sweden.

Fienberg, Stephen E. 1970. "An Iterative Procedure for Estimation in Contingency Tables." *Annals of Mathematical Statistics* 41:907–17.

———. 2005. "Confidentiality and Disclosure Limitation." Pp. 463–69 in *Encyclopedia of Social Measurement,* vol.1, edited by Kimberly Kempf-Leonard. Boston, MA: Elsevier Academic Press.

Friedland, Nehemiah, Shlomo Maital, and Aryeh Rutenberg. 1978. "A Simulation Study of Income Tax Evasion." *Journal of Public Economics* 10:107–16.

Friedlander, D. 1961. "A Technique for Estimating a Contingency Table, Given the Marginal Totals and Some Supplementary Data." *Journal of the Royal Statistical Society,* Series A (General), 124:412–20.

Gail, Mitchell, and Nathan Mantel. 1977. "Counting the Number of $\underline{r} \times \underline{c}$ Contingency Tables with Fixed Margins." *Journal of the American Statistical Association* 72:859–62.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall.

Goodman, Leo A., and William H. Kruskal. 1954. "Measures of Association for Cross Classifications." *Journal of the American Statistical Association* 49: 732–64.

———.1959. "Measures of Association for Cross Classifications II: Further Discussions and References." *Journal of the American Statistical Association* 54: 123–63.

———.1963. “Measures of Association for Cross Classifications III: Approximate Sampling Theory.” *Journal of the American Statistical Association* 58:310–64.

———.1972. “Measures of Association for Cross Classifications IV: Simplification of Asymptotic Variances.” *Journal of the American Statistical Association* 67:415–21.

Groenland, Edward A. G., and Gery M. van Veldhoven. 1983. “Tax Evasion Behavior: A Psychological Framework.” *Journal of Economic Psychology* 3:129–44.

Houston, Jodie, and Alfred Tran. 2001. “A Survey of Tax Evasion Using the Randomized Response Technique.” *Advances in Taxation* 13:69–94.

Jackson, Betty R., and Valerie C. Milliron. 1986. “Tax Compliance Research: Findings, Problems, and Prospects.” *Journal of Accounting Literature* 5:125–65.

Little, Roderick J. A., and Mei-Miau Wu. 1991. “Models for Contingency Tables with Known Margins When Target and Sampled Populations.” *Journal of the American Statistical Association* 86:87–95.

Mason, Robert, and Lyle D. Calvin. 1978. “A Study of Admitted Income Tax Evasion.” *Law and Society Review* 13:73–89.

Porcano, Thomas M. 1988. “Correlates of Tax Evasion.” *Journal of Economic Psychology* 9:47–67.

Ritsema, Christina M., Deborah W. Thomas, and Gary D. Ferrier. 2003. “Economic and Behavioral Determinants of Tax Compliance: Evidence from the 1997 Arkansas Tax Penalty Amnesty Program.” Presented at the 2003 IRS Research Conference.

Schuetze, Herb J. 2002. “Profiles of Tax Noncompliance among the Self-Employed in Canada: 1969 to 1992.” *Canadian Public Policy / Analyse de Politiques* 28: 219–38.

Slemrod, Joel, Marsha Blumenthal, and Charles Christian. 2001. “Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota.” *Journal of Public Economics* 79:455–83.

Tittle, Charles R. 1980. *Sanctions and Social Deviance: The Question of Deterrence*. New York: Praeger.

Vogel, Joachim. 1974. “Taxation and Public Opinion in Sweden: An Interpretation of Recent Survey Data.” *National Tax Journal* 27:499–513.

Wahlund, Richard. 1992. “Tax Changes and Economic Behavior: The Case for Tax Evasion.” *Journal of Economic Psychologys* 13:657–77.