A note on estimation of parametric binary outcome models with pure choice-based data

Nail Kashaev *
nkashaev@uwo.ca

December, 2020

Abstract I propose a generalized method of moments type procedure to estimate parametric binary choice models when the researcher only observes pure choices-based or presence-only data and has some information about the distribution of the covariates. This auxiliary information comes in the form of moments. I present an application based on the data on all COVID-positive tests from British Columbia. Publicly available demographic information on the population in British Columbia allows me to estimate the conditional probability of a person being COVID-positively tested conditional on observed demographics.

JEL classification: C2, C81, I19.

Keywords: Pure choice-based data, presence-only data, data combination, missing data, epidemiology, novel coronavirus.

*Department of Economics, University of Western Ontario.

1. Introduction

This note considers the problem of estimation of parametric binary outcome models with a pure choice-based or presence-only data samples. In pure choice-based datasets, the researcher only observes information about a particular group. For example, datasets on car accidents contain information only about drivers and cars involved in accidents (e.g., age, gender, sobriety level, weather conditions, and car model); store loyalty program datasets contain information on only those customers that made a purchase in that store (e.g., demographic information and information about products that were purchased); and medical test centers often collect information (e.g., age, gender, and health history) only about those who got sick or infected.

It is well known that the pure choice-based sample in itself does not identify the parameters of the conditional choice probabilities (Lancaster & Imbens, 1996). To identify and estimate the model, one needs some independent source of information. In this note I propose a simple, based on generalized method of moments (GMM), procedure that uses several moments (e.g., average age) from the whole population as this independent source. Lancaster & Imbens (1996) propose a semiparametric estimation procedure when an additional random *sample* of all covariates from the whole population is available. In practice, obtaining a sample from the whole population might be too costly or even impossible since some of the covariates may be observable only in the the pure choice-based sample.[1] My procedure, in contrast to Lancaster & Imbens (1996), only requires knowledge of a finite set of moments of covariates and can be applied in settings where the researcher can only get access to some aggregate marginal moments.

In many cases, for example in rare events studies or in marketing research, there is more complete information for a given choice outcome (Graham et al., 2004, Pearce & Boyce, 2005). In the opposite case, full data on covariates is only available for the whole population, while information for particular choices is private and unobservable (e.g., different types of elections). My procedure allows that either the population or the choice-based distribution of covariates to be replaced by only finite information (such as several moments or quantiles). Most importantly, for some cases, information on some regressors can be completely missing from, for example, population data.

The proposed procedure is based in inverse probability weighting. Informally, I minimize the difference between the observed moments from the population and the

[1] For instance, in datasets on car accidents the sobriety level of drivers is usually only observed after the accident has happened.

moments computed from the pure choice-based data weighted by the inverse of the probability of being in the pure choice-based sample. Since the latter probability is a function of the parameter of interest, I can estimate it. A similar idea is used in the literature on the multinomial sampling schemes (e.g., Manski & Lerman, 1977, Cosslett, 1981a,b, Imbens, 1992, and Tripathi, 2011) and in the literature on data sets with non-random attrition or models of missing data (e.g., Hellerstein & Imbens, 1999 and Nevo, 2003). In the former the weights are defined by the nature of the sampling scheme. In the latter, sometimes, to avoid the effects of attrition, panel data sets are augmented with new units randomly drawn from the original population, so-called refreshment samples. These refreshment samples are used to conduct the inverse probability weighting.[2]

As an empirical application, I estimate the conditional probability of a person being COVID-positively tested conditional on gender and age based on the data for all tested individuals in British Columbia. The results suggest that, conditional on being tested, men are more likely than women to get COVID-positive test results uniformly for all age groups. Moreover, for both males and females, the probability of getting COVID-positive test results as a function of age has two local maxima: the first one for those who are 20-29 years old, the second one if for those who are older that 90 years old.

This paper is organized as follows. Section 2 formally defines the underlying data generating process, the data structures, and the estimator. I show the empirical application in Section 3. Section 4 concludes.

[2]See Ridder (1992) and Hirano et al. (1998) for detailed discussions.

2. Model and Estimation

2.1. Main Model and Data Structure

Let $\mathbf{y} \in \{0, 1\}$ be binary outcome variable and $\mathbf{x}$ be a random vector of attributes supported on $X \subseteq \mathbb{R}^{d_x}$.[3] Assume that for all $x \in X$

$$\Pr(\mathbf{y} = 1 | \mathbf{x} = x) = G(x; \theta_0),$$

where $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is a vector of unknown parameters to be estimated, and G is a known function. Let q denote the unconditional probability of $\mathbf{y} = 1$. That is, $q = \Pr(\mathbf{y} = 1) = \mathbb{E}[G(\mathbf{x}; \theta_0)]$. In this section, I assume that q is known. Later on, I discuss how this assumption can be relaxed.

Assumption 1. *The analyst observes*

(i) a sample of independent and identically distributed (i.i.d.) observations $\{\tilde{\mathbf{x}}_i\}_{i=1}^n$ *from* $F_{\mathbf{x}|\mathbf{y}}(\cdot|1)$*;*

(ii) a vector $\bar{h}_x$ *such that* $\bar{h}_x = \mathbb{E}[h_x(\mathbf{x})]$ *for some known function* $h_x : X \to \mathbb{R}^{d_h}$*;*

Assumption 1(i) states that the sample is pure choice-based – the dependent variable $\mathbf{y}_i$ equals to 1 for everyone in the sample. Without extra information we can not learn anything about θ_0 (or $\Pr(\mathbf{y} = 1|\mathbf{x} = \cdot)$), since there is no variation in the outcome variable. This extra information comes in the form of Assumption 1(ii). In particular, I assume that there are known moments of $\mathbf{x}$ (not $\tilde{\mathbf{x}}$) captured by $\bar{h}_x$. Function h_x can take different forms. In applications, one usually has information about expected values or different quantiles of covariates. For instance, $h_x(x) = x$ implies that the econometrician knows expected value of $\mathbf{x}$. If $h_x(x) = (\mathbb{1}(x_i \leq x_i^*))_{i=1,\ldots,d_x}$ and $\bar{h}_x = (1/2)_{i=1,\ldots,d_x}$, then the econometrician knows the median of every component of $\mathbf{x}$. In my empirical application $h_x(x) = x$. Often, the vector $\bar{h}_x$ can only be consistently estimated. I extend my analysis to this case in Section 2.3.

The estimation strategy is easy to motivate. First, assume for a moment that $\mathbf{x}$ is

[3] Throughout the paper, deterministic vectors and functions are denoted by lower-case regular font Latin letters (e.g., x) and random objects by bold letters (e.g., $\mathbf{x}$). Capital letters are used to denote supports of random variables (e.g., $\mathbf{x} \in X$). I denote the support of a conditional distribution of $\mathbf{x}$ conditional on $\mathbf{z} = z$ by X_z. $F_{\mathbf{x}}(\cdot)$ ($f_{\mathbf{x}}(\cdot)$) and $F_{\mathbf{x}|\mathbf{z}}(\cdot|z)$ ($f_{\mathbf{x}|\mathbf{z}}(\cdot|z)$) denote the c.d.f. (p.d.f.) of $\mathbf{x}$ and $\mathbf{x}$ conditional on $\mathbf{z} = z$, respectively.

a continuously distributed random variable with p.d.f. $f_{\mathbf{x}}$. From Bayes' Rule

$$f_{\mathbf{x}|\mathbf{y}}(x|1) = \frac{\Pr(\mathbf{y}=1|\mathbf{x}=x)}{\Pr(\mathbf{y}=1)} f_{\mathbf{x}}(x) = \frac{G(x;\theta_0)}{q} f_{\mathbf{x}}(x)$$

for all $x \in X$. Hence,

$$\bar{h}_x = \mathbb{E}\left[h_x(\mathbf{x})\right] = \int_X h_x(x) f_{\mathbf{x}}(x) dx = \int_X \frac{q h_x(x)}{G(x;\theta_0)} f_{\mathbf{x}|\mathbf{y}}(x|1) dx = \mathbb{E}\left[\frac{q h_x(\mathbf{x})}{G(\mathbf{x};\theta_0)} \Big| \mathbf{y}=1\right]$$

The above intuition generalizes to settings where $\mathbf{x}$ might not admit a p.d.f as the following lemma demonstrates.

Lemma 1. *For a given function function $h_x : X \to \mathbb{R}^{d_h}$, if (i) $G(\mathbf{x};\theta_0) > 0$ with probability 1, and (ii) $\mathbb{E}\left[\left|\frac{h_x(\mathbf{x})}{G(\mathbf{x};\theta_0)}\right|\right] < \infty$, then*

$$\mathbb{E}\left[h_x(\mathbf{x})\right] = \mathbb{E}\left[\frac{q h_x(\mathbf{x})}{G(\mathbf{x};\theta_0)} \Big| \mathbf{y}=1\right].$$

Proof. The statement of the lemma follows from the following

$$\mathbb{E}\left[\frac{q h_x(\mathbf{x})}{G(\mathbf{x};\theta_0)} \Big| \mathbf{y}=1\right] = \frac{\mathbb{E}\left[\frac{q h_x(\mathbf{x}) \mathbb{1}(\mathbf{y}=1)}{G(\mathbf{x};\theta_0)}\right]}{\Pr(\mathbf{y}=1)} = \mathbb{E}\left[\frac{h_x(\mathbf{x}) \mathbb{1}(\mathbf{y}=1)}{G(\mathbf{x};\theta_0)}\right] =$$
$$= \mathbb{E}\left[\frac{h_x(\mathbf{x}) G(\mathbf{x};\theta_0)}{G(\mathbf{x};\theta_0)}\right] = \mathbb{E}\left[h_x(\mathbf{x})\right],$$

where the first equality follows from the definition of the conditional expectation, the second equality follows from the definition of q, and the third one follows from the law of iterated expectations. ■

Lemma 1 allows me to construct a system of moments that can be used to estimate the model. Define

$$m(x;\theta) \equiv q \frac{h_x(x)}{G(x;\theta)} - \bar{h}_x.$$

I have the following system of unconditional moment conditions

$$\mathbb{E}\left[m(\mathbf{x};\theta_0)\right] = 0. \tag{1}$$

The identification and estimation problem then is simply a question of the uniqueness of the solution to the system of equations (1).

I use the standard two-step GMM to estimate the parameters of interest. Denote

$$\hat{m}(\theta) = 1/n \sum_{i=1}^{n} m(\tilde{\mathbf{x}}_i; \theta), \qquad \tilde{V} = 1/n \sum_{i=1}^{n} m(\tilde{\mathbf{x}}_i, \tilde{\theta}) m'(\tilde{\mathbf{x}}_i, \tilde{\theta}) - \hat{m}(\tilde{\theta}) \hat{m}'(\tilde{\theta})$$
$$\tilde{\theta} = \arg\min_{\theta \in \Theta} \hat{m}'(\theta) \hat{m}(\theta), \qquad \hat{\theta} = \arg\min_{\theta \in \Theta} \hat{m}'(\theta) \tilde{V}^{-1} \hat{m}(\theta).$$

Then under the standard regularity conditions listed below, my GMM estimator is consistent and asymptotically normal.

Theorem 2.1. *If (i) Θ and X are compact; (ii) $G(x;\theta)$ and its derivative with respect to θ are continuous and bounded away from zero for all $\theta \in \Theta$ and $x \in X$; (iii) θ_0 is the unique solution to $\mathbb{E}[m(\tilde{\mathbf{x}};\theta)] = 0$; (iv) The matrix $V = \mathbb{E}[q^2 h_x(\tilde{\mathbf{x}}) h'(\tilde{\mathbf{x}}) / G^2(\tilde{\mathbf{x}};\theta_0)] - \bar{h}_x \bar{h}'_x$ is nonsingular; and (iv) The matrix $A = \mathbb{E}[q h_x(\tilde{\mathbf{x}}) \partial_{\theta'} G(\tilde{\mathbf{x}};\theta_0) / G^2(\tilde{\mathbf{x}};\theta_0)]$ is of full column rank; then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, (A' V^{-1} A)^{-1})$$

Proof. Consistency and asymptotic normality of the estimator can be proved as in Newey & McFadden (1994). ■

2.2. Alternative Data Structure

Suppose that instead of observing a pure choice-based sample and some moments from the distribution of $\mathbf{x}$, the analyst observes a sample from the population $\{\mathbf{x}_i\}_{i=1}^{m}$ and some moments from the distribution of the "pure choice-based" population. That is, the analyst knows $\bar{h}_{\tilde{x}} = \mathbb{E}[h_{\tilde{x}}(\mathbf{x}) | \mathbf{y} = 1]$ for some known function $h_{\tilde{x}} : X \to \mathbb{R}^{d_h}$. In this situation we still can apply Lemma 1 and obtain the following moment condition

$$\mathbb{E}[m(\mathbf{x};\theta)] = \mathbb{E}\left[\frac{G(\mathbf{x};\theta_0) h_{\tilde{x}}(\mathbf{x})}{q} - \bar{h}_{\tilde{x}}\right].$$

Given the above moment condition, one just needs to apply Theorem 2.1 to it and to get consistent and asymptotically normal estimator of θ_0.

2.3. Accounting for Sampling Error in the Moment Restrictions

In the above analysis I assume that $\bar{h}_x$ is known exactly. In this section I generalizes the result to the case when $\bar{h}_x$ is estimated from an auxiliary sample. Assume that we observe an independent sample $\{\mathbf{x}_i\}_{i=1}^m$ along with the primary sample $\{\tilde{\mathbf{x}}_i\}_{i=1}^n$. In this situation we can either use the auxiliary sample to estimate $\bar{h}_x$ by $\hat{h}_x = 1/m \sum_i h_x(\mathbf{x}_i)$, or use the primary sample to estimate $\bar{h}_{\tilde{x}}$ by $\hat{h}_{\tilde{x}} = 1/m \sum_i h_{\tilde{x}}(\tilde{\mathbf{x}}_i)$. If the sample size of the primary sample, n, is relatively bigger than the size of the auxiliary sample, m, (i.e., m/n converges to zero), then one can ignore the sampling error in estimation of $\bar{h}_{\tilde{x}}$ and can use the estimation procedure described in Section 2.2. In the opposite case, when n/m converges to zero, it is better to use the estimation procedure described in Section 2.1.[4]

If m/n converges to an integer $k \neq 0$ then, following Hellerstein & Imbens (1999), I can describe the asymptotic behavior as follows. Suppose we have n observations of $\mathbf{t}$, where $\mathbf{t}_i$ consists of $(\tilde{\mathbf{x}}_i, \mathbf{h}_{i1}, \mathbf{h}_{i2}, \ldots, \mathbf{h}_{ik})$. The observations $\mathbf{h}_{ij}$ are used to estimate $\bar{h}_x$ by $\sum_{i=1}^n \sum_{j=1}^k \mathbf{h}_{ij}/(nk)$. Denote

$$\hat{m}(\theta) = 1/n \sum_{i=1}^n \left[\frac{q h_x(\tilde{\mathbf{x}}_i)}{G(\tilde{\mathbf{x}}_i;\theta)} - 1/k \sum_{j=1}^k \mathbf{h}_{ij} \right], \quad \tilde{V} = 1/n \sum_{i=1}^n m(\tilde{\mathbf{x}}_i;\tilde{\theta}) m'(\tilde{\mathbf{x}}_i;\tilde{\theta}) - \hat{m}(\tilde{\theta})\hat{m}'(\tilde{\theta})$$
$$\tilde{\theta} = \arg\min_{\theta \in \Theta} \hat{m}'(\theta)\hat{m}(\theta), \qquad \hat{\theta} = \arg\min_{\theta \in \Theta} \hat{m}'(\theta)\tilde{V}^{-1}\hat{m}(\theta)$$

Similarly to Theorem 1 the following theorem describes the large sample properties of the estimator.

Theorem 2.2. *If (i) Θ and X are compact; (ii) $G(x;\theta)$ and its derivative with respect to θ are continuous and bounded away from zero for all $\theta \in \Theta$ and $x \in X$; (iii) θ_0 is the unique solution to $\mathbb{E}[m(\tilde{\mathbf{x}};\theta)] = 0$; (iv) The matrix $V = \mathbb{E}[q^2 h_x(\tilde{\mathbf{x}}) h_x'(\tilde{\mathbf{x}})/G^2(\tilde{\mathbf{x}};\theta_0)] - 1/k^2 \sum_{j=1}^k \mathbb{E}[\mathbf{h}_{1j}] \sum_{j=1}^k \mathbb{E}[\mathbf{h}_{1j}']$ is nonsingular; (iv) The matrix*

$$A = \mathbb{E}\left[q h_x(\tilde{\mathbf{x}}) \partial_{\theta'} G(\tilde{\mathbf{x}};\theta_0)/G^2(\tilde{\mathbf{x}};\theta_0) \right]$$

is of full column rank; then

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, (A'V^{-1}A)^{-1})$$

Proof. The proof follows the same steps as the proof of Theorem 2.1. ■

[4] This case is not considered in Hellerstein & Imbens (1999). Nevo (2003) mentions that in this case his procedure does not work.

2.4. Local Identification

In the estimation part I assume that the parameters of interest are point identified. In this subsection I provide sufficient conditions for local identification of θ_0. I say that θ_0 is locally identified if $\mathbb{E}\left[\partial_{\theta'} m(\tilde{\mathbf{x}};\theta)\right]$ has full column rank for all θ in some neighborhood of θ_0. Generally, the above rank condition is hard to verify. However, for the special case that is used in my empirical application I can derive a simple sufficient conditions.

Lemma 2. *Let $h_x(x) = x$ and $G(x;\theta) = F(x'\theta)$, where $F(\cdot)$ is strictly increasing or decreasing. Assume that (i) $G(\tilde{\mathbf{x}};\theta) > 0$ with probability 1 for all $\theta \in \Theta$; (ii) $\mathbb{E}\left[\partial_{\theta'} m(\tilde{\mathbf{x}};\theta)\right]$ exists for all $\theta \in \Theta$; and (iii) $\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'\right]$ is a positive definite matrix. Then θ_0 is locally identified.*

Proof. Assume without loss of generality that $F(\cdot)$ is a strictly increasing function. Fix some $\theta \in \Theta$. Denote

$$\alpha(x;\theta) = \frac{qF'(x'\theta)}{G(x;\theta)^2},$$
$$A(\theta) = \mathbb{E}\left[\alpha(\tilde{\mathbf{x}};\theta)\tilde{\mathbf{x}}\tilde{\mathbf{x}}'\right].$$

Then

$$\mathbb{E}\left[\partial_{\theta'} m(\tilde{\mathbf{x}};\theta)\right] = -A(\theta).$$

$A(\theta)$ is positive definite since $\alpha(\tilde{x};\theta) > 0$ with probability 1 for all θ. Hence, θ_0 is locally identified. ■

Lemma 2 has a direct application to the Probit and the Logit models since $F(\cdot)$ can be the logistic or the standard normal c.d.f. Lemma 2 does require knowing moments for each of the covariates. However, in some cases, if the analyst cannot obtain a moment for one of them, she can use different moments of the rest of the covariates to identify the parameter of interest. For instance, consider the following example.

Example 1. Let

$$G(\tilde{x};\theta_0) = \frac{1}{1+\exp\{\theta_{00} + \theta_{01}\tilde{x}_1 + \theta_{02}\tilde{x}_2\}}$$

$\tilde{x} = (\tilde{x}_1, \tilde{x}_2) \in \{-1, 0, 1\} \times \{0, 1\}$ is distributed according to

$$\begin{cases} \Pr(\tilde{\mathbf{x}}_1 = 1, \tilde{\mathbf{x}}_2 = 1) = 2/3 \\ \Pr(\tilde{\mathbf{x}}_1 = 0, \tilde{\mathbf{x}}_2 = 0) = \Pr(\tilde{\mathbf{x}}_1 = -1, \tilde{\mathbf{x}}_2 = 0) = 1/6 \\ \Pr(\tilde{\mathbf{x}}_1 = 0, \tilde{\mathbf{x}}_2 = 1) = \Pr(\tilde{\mathbf{x}}_1 = -1, \tilde{\mathbf{x}}_2 = 1) = \Pr(\tilde{\mathbf{x}}_1 = 1, \tilde{\mathbf{x}}_2 = 0) = 0, \end{cases}$$

and

$$h_x(\tilde{x}) = \begin{bmatrix} 1 \\ \tilde{x}_1 \\ \tilde{x}_1^2 \end{bmatrix}$$

That is, instead of using a moment of $\mathbf{x}_2$, I use the second moment of $\mathbf{x}_1$ to identify θ_{02}. Then

$$-\mathbb{E}\left[\partial_{\theta'} m(\tilde{\mathbf{x}}; \theta)\right] / q = e^{\theta_0}/6 \begin{bmatrix} 4e^{\theta_1+\theta_2} + 1 + e^{-\theta_1} & 4e^{\theta_1+\theta_2} - e^{-\theta_1} & 4e^{\theta_1+\theta_2} \\ 4e^{\theta_1+\theta_2} - e^{-\theta_1} & 4e^{\theta_1+\theta_2} + e^{-\theta_1} & 4e^{\theta_1+\theta_2} \\ 4e^{\theta_1+\theta_2} + e^{-\theta_1} & 4e^{\theta_1+\theta_2} - e^{-\theta_1} & 4e^{\theta_1+\theta_2} \end{bmatrix}$$

and the right hand side matrix has the same rank as

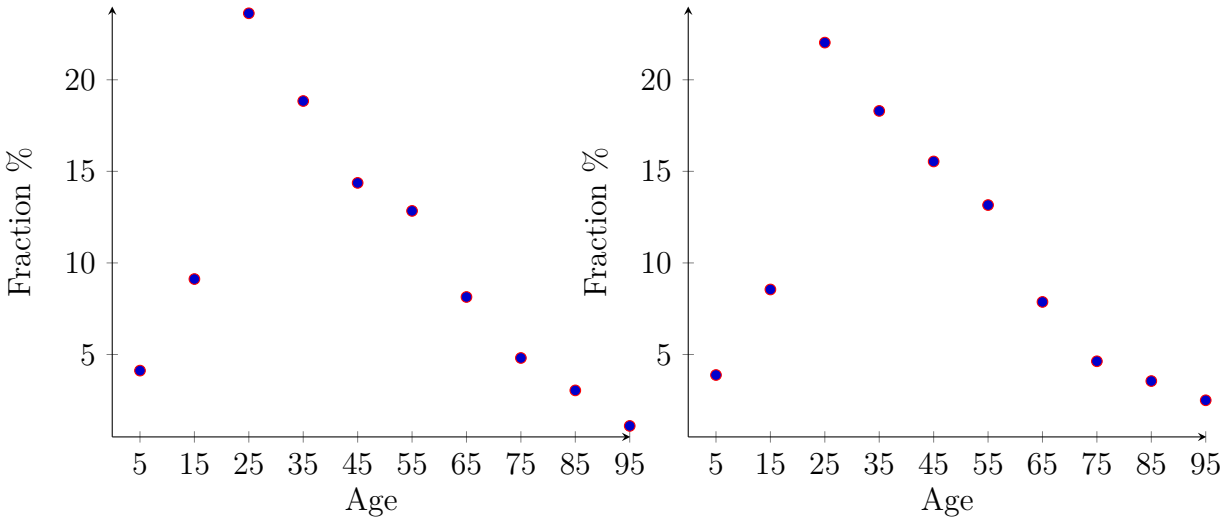
$$\begin{bmatrix} 4e^{\theta_1+\theta_2} + 1 + e^{-\theta_1} & 4e^{\theta_1+\theta_2} - e^{-\theta_1} & 4e^{\theta_1+\theta_2} \\ -2e^{-\theta_1} - 1 & 2e^{-\theta_1} & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

The latter has full rank for all θ. Hence, θ_0 is locally identified.

Note that $h_x(\tilde{\mathbf{x}})$ can be treated as an instrument for $\tilde{\mathbf{x}}$. In the above example, since $\tilde{\mathbf{x}}_1$ is correlated with $\tilde{\mathbf{x}}_2$, $\tilde{\mathbf{x}}_1^2$ contains enough information to identify the coefficient in front of $\tilde{\mathbf{x}}_2$.

3. An Empirical Application

In this section, I use the propose procedure to estimate the conditional probability of being COVID-positively tested (the rate of positive results among those tested or the test yield) conditional on gender and age using the data on all COVID-tested individuals in British Columbia. Although this probability does not exactly measure



(a) Males **(b)** Females

Figure 1 – Age distribution in the sample for different genders. Age axis correspond to the mid points of age brackets (e.g, 15 corresponds to the 10-19 bracket).Sample size $= 38,036$.

the probability of being infected by the virus, it is informative about the infection rate and can be used to construct bounds on it (Manski & Molinari, 2020, Stoye, 2020). My procedure, under the assumptions I make, allows to derive these bounds conditional on different demographic characteristics.

In the data I only observe age and gender of those who were tested positively. As a result, I have a pure choice-based sample. In order to estimate the model, I augment the data set by information on the age and gender distribution in British Columbia.

3.1. Primary Sample

The data used in the analysis is publicly available and is provided by the BC Centre for Disease Control [5] Every observation contains information about the age, gender, and residency of tested individual.[6] Age is characterized by 10 binary variable: $I_{<10,i}$, $I_{10-19,i}$, $I_{20-29,i}$,...,$I_{>90,i}$. So $I_{j,i} = 1$ if individual i belongs to age group j. The gender variable, Male_i, is also binary and is equal to 1 if individual i is male and 0 otherwise. I exclude 116 observations that had missing values. As a result, I end up having $38,036$ observation.

[5] http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data

[6] The data was obtained on December 8, 2020.

Almost half of individuals in the sample are females (49.1 percent). Figure 1 presents age distribution for males and females in the sample. The distribution is unimodal with the pick at 20-29 years old age group.

3.2. Auxiliary Information

In order to construct the system of moment conditions, I need to choose function h_x and estimate $\bar{h}_x$ and q. The total number of tested individuals is reported at the website of the BC Centre for Disease Control is $1,235,006$.[7] Thus, the q is estimated as the ratio of the number of COVID-positive tests to the number of tested individuals: $38036/1,235,006 = 0.0308$. Since the age and gender distribution of individuals is publicly available, I picked $h_x(x) = x$. The estimates of $\bar{h}_x$ were obtained from the official website of the Government of British Columbia[8]. The moments used in estimation are presented in Table 1.

Table 1 – Moments used in estimation

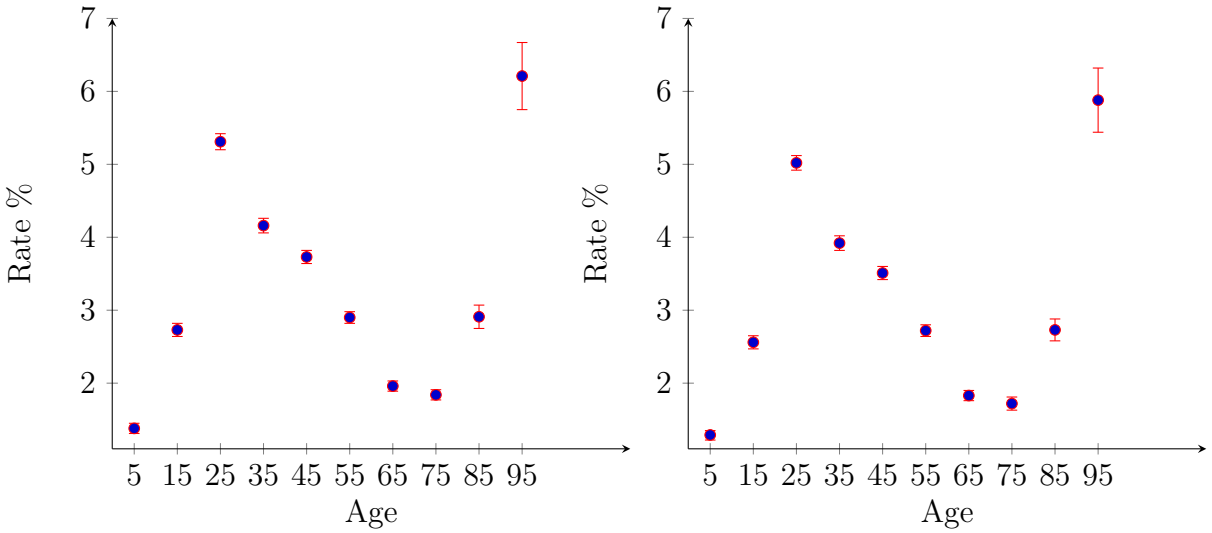
h_x	Male	$I_{<10}$	I_{10-19}	I_{20-29}	I_{30-39}	I_{40-49}	I_{50-59}	I_{60-69}	I_{70-79}	I_{80-89}	$I_{>90}$
$\bar{h}_x$	0.495	0.093	0.103	0.136	0.142	0.128	0.143	0.13	0.082	0.036	0.009

3.3. Estimation

I model the probability of being positively tested conditional on age and gender as the Probit model. That is, $G(x, \theta_0) = \Phi(x'\theta_0)$, where x includes a constant, nine age group dummy variables, and the gender dummy variable; $\Phi(\cdot)$ is the standard normal c.d.f. The estimated coefficients together with corresponding standard error are presented in Table 2. All coefficients a significant at the 5 percent significance level.

[7] The data was obtained on December 7, 2020.

[8] https://www2.gov.bc.ca/gov/content/data/statistics/people-population-community/population/population-estimates



(a) Males **(b)** Females

Figure 2 – Probability of being COVID-positively tested for different age groups and gender (Panels (a) and (b)). Dots indicate point estimates. Bars indicate 95 percent confidence intervals.

Table 2 – Estimation results

Variable	$\hat{\theta}$	std error
Constant	-1.5651	0.0192
<10	-0.6655	0.0218
10-19	-0.3848	0.0207
20-29	-0.078	0.02
30-39	-0.1946	0.0201
40-49	-0.2458	0.0203
50-59	-0.3587	0.0203
60-69	-0.5245	0.0207
70-79	-0.5493	0.0216
80-89	-0.3567	0.0229
Male	0.0274	0.0049

In Figure 2 I present estimates (together with 95 percent confidence intervals) of the rate of positive results for different age groups for males and females. There are two local maxima at "20-29" and ">90" age groups. I also estimated the relative

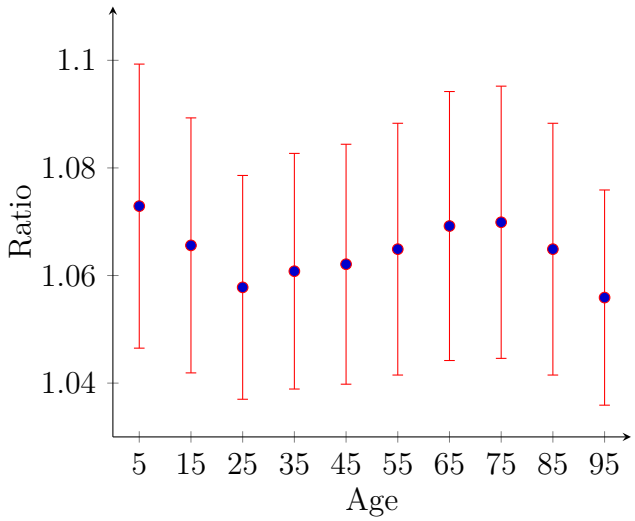


Figure 3 – The ratio of probabilities of being COVID-positively tested between males and females. Dots indicate point estimates. Bars indicate 95 percent confidence intervals.

risk of being positively tested as the rate for males divided by the rate for females for different age groups. The results, presented in Figure 3, suggest that uniformly over age groups males are about 1.06 times more likely to be COVID-positively tested. Moreover, for the age groups "20-29" and ">90", this risk for males is the lowest.

3.4. Discussion

Apart from the parametric specification for G, in order to use the age and gender distribution I have to assume that probability of being tested does not depend on conditioning variables. This may not be true if, for instance, younger individuals get more tested than the older ones. However, the results for the relative risk are robust to this type of dependencies as long as probability of being does not depend on gender.

Since all covariates are binary, one can conclude without any estimation procedure that since there are 50.49 percent of females in British Columbia and 50.9 percent of infected are males, then the rate of positive results should be higher for males (assuming that the probability of being tested does not depend on gender). My procedure allows to unpack one more level of observed heterogeneity captured by age without knowing the joint distribution of age and gender. Moreover, if I had a continuously distributed covariate (e.g., income level), then my method would require knowing only the average income in order to estimate the model parameters.

4. Conclusion

This paper proposes a method to estimate binary models with a pure-choice based data or a data with unobserved responses in the presence of additional information that comes in the form of the finite set of moments. Importantly, the pure-choice based data problem is dual to the problem of data with unobserved response. Hence, the procedure can be used in estimation of inverse probability weights in data sets with non-random attrition even if the refreshment sample is much smaller than the primary sample. I applied the procedure to estimate the probability of being COVID-positively tested conditional on demographics using the data from British Columbia.

References

Cosslett, S. (1981a). Efficient estimation of discrete-choice models. *Structural analysis of discrete data with econometric applications*, (pp. 51–111).

Cosslett, S. (1981b). Maximum likelihood estimator for choice-based samples. *Econometrica*, (pp. 1289–1316).

Graham, C., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19(9), 497–503.

Hellerstein, J. & Imbens, G. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, 81(1), 1–14.

Hirano, K., Imbens, G., Ridder, G., & Rebin, D. (1998). Combining panel data sets with attrition and refreshment samples.

Imbens, G. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, (pp. 1187–1214).

Lancaster, T. & Imbens, G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics*, 71(1), 145–160.

Manski, C. & Lerman, S. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, (pp. 1977–1988).

Manski, C. F. & Molinari, F. (2020). Estimating the covid-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics*, 21(1), 43–52.

Newey, W. & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4, 2111–2245.

Pearce, J. & Boyce, M. (2005). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43(3), 405–412.

Ridder, G. (1992). An empirical evaluation of some models for non-random attrition in panel data. *Structural Change and Economic Dynamics*, 3(2), 337–355.

Stoye, J. (2020). Bounding disease prevalence by bounding selectivity and accuracy of tests: The case of covid-19. *arXiv preprint arXiv:2008.06178*.

Tripathi, G. (2011). Gmm based inference with stratified samples when the aggregate shares are known. *Journal of Econometrics*.