Estimation of parametric binary outcome models with degenerate pure choice-based data with application to COVID-19-positive tests from British Columbia

Nail Kashaev [*]

nkashaev@uwo.ca

May, 2021

Abstract I propose a generalized method of moments type procedure to estimate parametric binary choice models when the researcher only observes degenerate 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-19-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-19-positively tested conditional on 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 paper considers the problem of estimation of parametric binary outcome models with degenerate pure choice-based or presence-only data samples. In degenerate 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 degenerate 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 paper, I show that the parameters of the conditional choice probabilities can be point-identified if one uses several moments (e.g., average age) from the whole population as this independent source. Based on these moments, I propose a computationally simple, generalized method of moments (GMM) type procedure. 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 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.

As an empirical application, I estimate the conditional probability of a person being

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

COVID-19-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, for both males and females, the probability of getting COVID-19-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 than 90 years old. Moreover, the estimated ratio of the risk for males to the risk for females is greater than 1 for all ages suggesting that men are more likely than women to get COVID-19-positive tested uniformly for all age groups. Interestingly, this ratio does not change much with age. In this application, I only used publicly available data on those who were positively tested and only information on the fraction of females and the fractions of different age groups in the population (i.e., no information on the joint distribution of age and gender is required).

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 because of privacy concerns (e.g., different types of elections). My procedure allows 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 on inverse probability weighting. Informally, I minimize the difference between the observed moments from the population and the 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] In contrast to these literatures, my main result does not require the existence of the auxiliary sample to estimate the parameters of interest. Imbens & Lancaster (1994) use additional moment conditions to achieve efficiency gains. In contrast to my setting, Imbens & Lancaster (1994) can estimate their model without these moments.

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. In Section 4, I provide several extensions of the baseline model. Section 5 concludes.

2. 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, $q = \Pr(\mathbf{y} = 1) = \mathbb{E}\left[G(\mathbf{x}; \theta_0)\right]$. In this section, I assume that q is known. Later on, I discuss how this assumption can be relaxed.

Assumption 1. *The analyst observes*

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

[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.

(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* $\bar{h}_x = \mathbb{E}\left[h_x(\mathbf{x})\right]$ *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 $h_x(x) = (\mathbb{1}\left(x_i \leq x_i^* \right))_{i=1,\ldots,d_x}$ 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 $h_x(x) = x$. Note that h_x may only include marginal moments of covariates. No information on the joint distribution is needed.

Often, the vector $\bar{h}_x$ can only be consistently estimated. In this situations one will use auxiliary samples of covariates from the population. Importantly, in contrast to existing methods, these samples can come from completely different sources and contain no information about the correlation structure between covariates. For instance, if $\mathbf{x}$ includes age and income variables, one would only need one sample that contains age data, and possibly a completely independent sample that has information about income.[4] Even if the auxiliary samples are not available for different reasons (e.g., privacy concerns), statistical agencies often report summary statistics of these samples. These summary statistics can be used as proxies for $\bar{h}_x$.

The estimation strategy is easy to motivate. First, assume for a moment that $\mathbf{x}$ is a continuously distributed random variable with p.d.f. $f_{\mathbf{x}}$. From Bayes' Rule

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

[4] I extend my analysis to this case in Section 4.2.

for all $x \in X$. Hence,

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

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) $\mathbb{E}\left[\left|\left|\frac{h_x(\mathbf{x})}{G(\mathbf{x};\theta_0)}\right|\right|\right] < \infty$, then*

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

Proof. The statement of the lemma follows from the following

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

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

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

I have the following system of unconditional moment conditions (the expectation is taken with respect to $\tilde{\mathbf{x}}$, not $\mathbf{x}$)

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

2.1. Identification

The identification problem is simply a question of the uniqueness of the solution to the system of equations (1). Given that function G is usually highly nonlinear, it is hard to provide primitive conditions for local or global identification for a general model (for examples of sufficient conditions see, for instance, Rothenberg, 1971, Newey & McFadden, 1994, and Komunjer, 2012). However, for the special case that is used in my empirical application, I can derive simple sufficient conditions for global identification.

Lemma 2. *Let $h_x(x) = x$ and $G(x;\theta) = F(x^\intercal\theta)$, where $F(\cdot)$ is strictly increasing or decreasing continuously differentiable function. Assume that (i) Θ is convex set; (ii) $G(\tilde{\mathbf{x}};\theta) > 0$ with probability 1 for all $\theta \in \Theta$; and (iv) $\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal\right]$ is invertible. Then if $q \in (0,1)$ is known (can be consistently estimated), then θ_0 is globally identified*

Proof. Without loss of generality assume that F is strictly increasing. Note that

$$-\mathbb{E}\left[\partial_{\theta^\intercal} m(\tilde{\mathbf{x}};\theta)\right] = \mathbb{E}\left[\frac{qF'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal\right],$$

where F' is the derivative of F. Since for any $\lambda \neq 0$

$$\lambda^\intercal\mathbb{E}\left[\frac{qF'(\tilde{\mathbf{x}}^\intercal\theta)}{F'(\tilde{\mathbf{x}}^\intercal\theta)^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal\right]\lambda = \mathbb{E}\left[\frac{qF'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)^2}(\lambda^\intercal\tilde{\mathbf{x}})^2\right] > 0,$$

we can conclude that $\mathbb{E}\left[\partial_{\theta^\intercal} m(\tilde{\mathbf{x}};\theta)\right]$ is negative-definite matrix. Hence, by Theorem 6 in Gale & Nikaido (1965), $g(\cdot) = \mathbb{E}\left[m(\tilde{\mathbf{x}};\cdot)\right]$ is injective and θ_0 is globally 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.. Condition (iii) is the standard rank condition that requires absence of multicollinearity. 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 at least locally 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 $h_x(\tilde{x})=(1,\tilde{x}_1,\tilde{x}_1^2)^\intercal$. 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

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

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 $\tilde{\mathbf{x}}_2$.

Lemma 2 requires knowledge of the aggregate share $q = \Pr(\mathbf{y} = 1)$. Note, however, that if $d_h \geq d_\theta + 1$, then we can treat q as an unknown extra parameter and still identify the model. Appendix A formalizes this intuition by providing a sufficient condition for global identification when q is unknown.

I conclude this section by noting that identification of θ_0 does not imply that the population averages can be identified. For instance, with known θ_0, we can identify the marginal effect $\dfrac{\partial \Pr(\mathbf{y} = 1|\mathbf{x} = x)}{\partial x}$ for a given x, but we can only derive bounds for the average marginal effect $\mathbb{E}\left[\dfrac{\partial \Pr(\mathbf{y} = 1|\mathbf{x})}{\partial x}\right]$.

2.2. Estimation

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

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

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 $\mathbb{E}\left[m(\tilde{\mathbf{x}};\theta)\right] = 0$; (iv) The matrix $V = \mathbb{E}\left[q^2 h_x(\tilde{\mathbf{x}})h^\intercal(\tilde{\mathbf{x}})/G^2(\tilde{\mathbf{x}};\theta_0)\right] - \bar{h}_x\bar{h}_x^\intercal$ is nonsingular; and (iv) The matrix $A = \mathbb{E}\left[q h_x(\tilde{\mathbf{x}})\partial_{\theta^\intercal} G(\tilde{\mathbf{x}};\theta_0)/G^2(\tilde{\mathbf{x}};\theta_0)\right]$ is of full column rank; then*

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

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

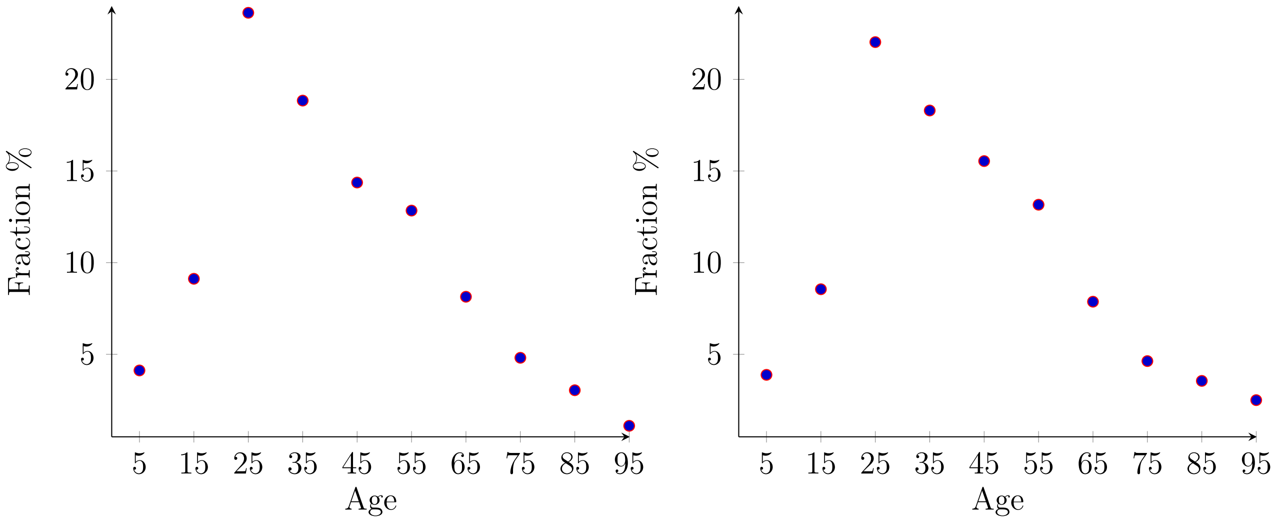
The above estimation procedure assumes that q is known. Note, however, that if $d_h \geq d_\theta + 1$, then we can treat q as an unknown extra parameter and still estimate the model provided the conditions for Theorem 2.1 are satisfied with $(\theta_0^\intercal, q)^\intercal$ instead of θ_0.

I conclude this section by noting that even if the parameter of interest is not point identified, we deal with a system of unconditional moment conditions. Hence, one can always construct confidence sets for the parameter of interest or the identified set using methods proposed in, for instance, Chernozhukov et al. (2007) or Andrews & Soares (2010).

3. An Empirical Application

In this section, I use the proposed procedure to estimate the conditional probability of being COVID-19-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-19-tested individuals in British Columbia. Although this probability does not exactly measure 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.



(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 $= 38,036$.

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: $\mathrm{I}_{<10,i}$, $\mathrm{I}_{10-19,i}$, $\mathrm{I}_{20-29,i}$,...,$\mathrm{I}_{>90,i}$. So $\mathrm{I}_{j,i} = 1$ 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.

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.

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

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

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 parameter q is estimated as the ratio of the number of COVID-19-positive tests to the number of tested individuals: $38036/1,235,006 = 0.0308$. Since the information about different age groups and gender 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 2.

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, $G(x, \theta_0) = \Phi(x^\intercal \theta_0)$, 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 errors are presented in Table 2. All coefficients are significant at the 5 percent significance level.

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 risk of being positively tested as the rate for males divided by the rate for females for different age groups. The results,

[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

Table 2 – Estimation results. The results are rounded to the third digit.

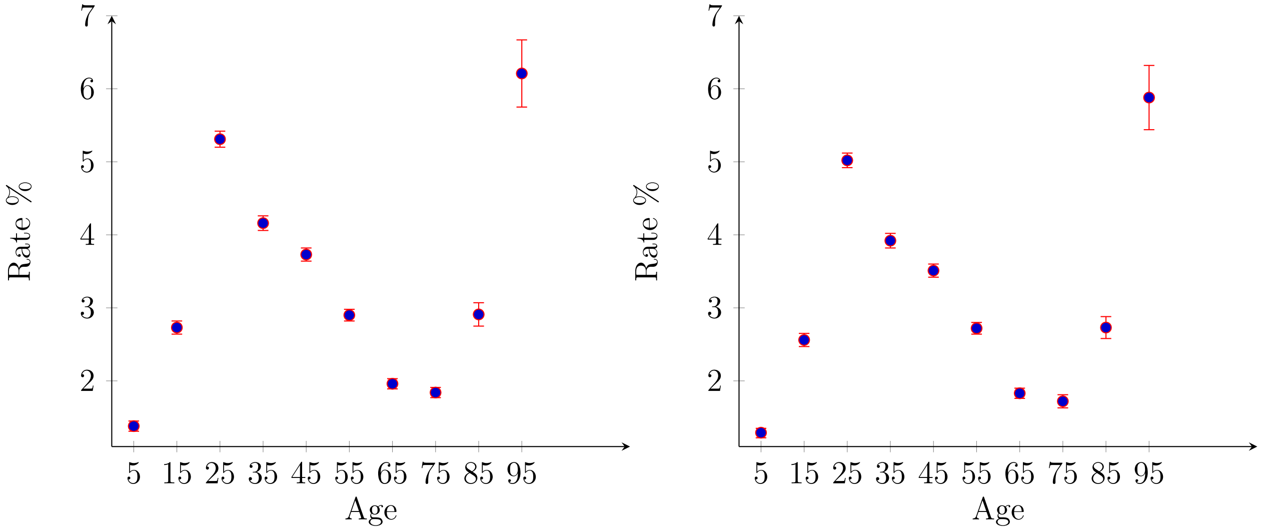
Variable	$\hat{\theta}$	std error
Constant	-1.565	0.019
<10	-0.666	0.022
10-19	-0.385	0.021
20-29	-0.078	0.02
30-39	-0.195	0.02
40-49	-0.246	0.02
50-59	-0.359	0.02
60-69	-0.525	0.021
70-79	-0.549	0.022
80-89	-0.357	0.023
Male	0.027	0.005

presented in Figure 3, suggest that uniformly over age groups males are about 1.06 times more likely to be COVID-19-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, to use the age and gender distribution, I have to assume that the 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 the probability of being tested 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



(a) Males

(b) Females

Figure 2 – Probability of being COVID-19-positively tested for different age groups and gender (Panels (a) and (b)). Dots indicate point estimates. Bars indicate 95 percent confidence intervals.

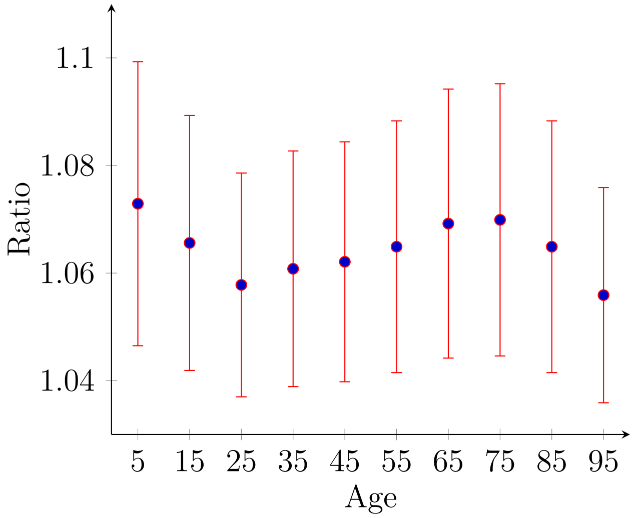


Figure 3 – The ratio of probabilities of being COVID-19-positively tested between males and females. Dots indicate point estimates. Bars indicate 95 percent confidence intervals.

model parameters.

4. Extensions

4.1. 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 $\bar{h}_{\tilde{x}} = \mathbb{E}\left[h_{\tilde{x}}(\mathbf{x})|\mathbf{y}=1\right]$ 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

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

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

4.2. 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 4.1. In the opposite case, when n/m converges to zero, it is better to use the estimation procedure described in Section 2.[9]

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 $(\tilde{\mathbf{x}}_i, \mathbf{h}_{i1}, \mathbf{h}_{i2}, \dots, \mathbf{h}_{ik})$. 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

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

$$\tilde{\theta} = \arg\min_{\theta\in\Theta} \hat{m}^\intercal(\theta)\hat{m}(\theta), \qquad \hat{\theta} = \arg\min_{\theta\in\Theta} \hat{m}^\intercal(\theta)\tilde{V}^{-1}\hat{m}(\theta)$$

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

Theorem 4.1. *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 $\mathbb{E}\left[m(\tilde{\mathbf{x}};\theta)\right] = 0$; (iv) The matrix $V = \mathbb{E}\left[q^2 h_x(\tilde{\mathbf{x}}) h_x^\intercal(\tilde{\mathbf{x}})/G^2(\tilde{\mathbf{x}};\theta_0)\right] - 1/k^2 \sum_{j=1}^{k} \mathbb{E}\left[\mathbf{h}_{1j}\right] \sum_{j=1}^{k} \mathbb{E}\left[\mathbf{h}_{1j}^\intercal\right]$ is nonsingular; (iv) The matrix*

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

is of full column rank; then

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

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

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

5. 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 a finite set of moments. The pure-choice based data problem is dual to the problem of data with unobserved response. Hence, the procedure can be used in the 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-19-positively tested conditional on demographics using the data from British Columbia. The results indicate that males of all ages are more likely to be positively tested than females. However, this relative risk is similar for all age groups.

References

Andrews, D. W. & Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1), 119–157.

Chernozhukov, V., Hong, H., & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5), 1243–1284.

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).

Gale, D. & Nikaido, H. (1965). The jacobian matrix and global univalence of mappings. *Mathematische Annalen*, 159(2), 81–93.

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).

Imbens, G. W. & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4), 655–680.

Komunjer, I. (2012). Global identification in nonlinear models with moment restrictions. *Econometric Theory*, (pp. 719–729).

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.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, (pp. 577–591).

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*.

A. Additional Identification Result

Lemma 3. *Let conditions of Lemma 2 are satisfied but $q \in (0,1)$ is unknown. Assume, moreover, that (i) $\bar{h}_c = \mathbb{E}\left[h_c(\mathbf{x})\right]$ is known for some known $h_c : X \to \mathbb{R}$; (ii) the determinant of*

$$\mathbb{E}\left[\begin{array}{cc} \dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal & \left(\dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)}h_c(\tilde{\mathbf{x}}) - 1\right)\dfrac{\tilde{\mathbf{x}}}{2F(\tilde{\mathbf{x}}^\intercal\theta)} \\ \left(\dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)}h_c(\tilde{\mathbf{x}}) - 1\right)\dfrac{\tilde{\mathbf{x}}^\intercal}{2F(\tilde{\mathbf{x}}^\intercal\theta)} & -\dfrac{h_c(\tilde{\mathbf{x}})}{F(\tilde{\mathbf{x}}^\intercal\theta)} \end{array}\right]$$

is positive for all $\theta \in \Theta$ and $q' \in (0,1)$. Then θ_0 and q are globally identified.

Proof. Note that if we add the extra moment condition captured by h_c, then the matrix of derivatives of the extended system with respect to $(\theta^\intercal, q)^\intercal$ evaluated at $(\theta^\intercal, q')^\intercal$ equals to

$$B = \mathbb{E}\left[\begin{array}{cc} -\dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal & \dfrac{\tilde{\mathbf{x}}}{F(\tilde{\mathbf{x}}^\intercal\theta)} \\ -\dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)^2}h_c(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^\intercal & \dfrac{h_c(\tilde{\mathbf{x}})}{F(\tilde{\mathbf{x}}^\intercal\theta)}. \end{array}\right]$$

Similar to the proof of Lemma 2, by Theorem 6 in Gale & Nikaido (1965), to establish global identification it is sufficient to show that $(B + B^\intercal)/2$ is a negative-definite matrix. Note that

$$-(B + B^\intercal)/2 = \mathbb{E}\left[\begin{array}{cc} \dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal & \left(\dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)}h_c(\tilde{\mathbf{x}}) - 1\right)\dfrac{\tilde{\mathbf{x}}}{2F(\tilde{\mathbf{x}}^\intercal\theta)} \\ \left(\dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)}h_c(\tilde{\mathbf{x}}) - 1\right)\dfrac{\tilde{\mathbf{x}}^\intercal}{2F(\tilde{\mathbf{x}}^\intercal\theta)} & -\dfrac{h_c(\tilde{\mathbf{x}})}{F(\tilde{\mathbf{x}}^\intercal\theta)} \end{array}\right]$$

Since conditions of Lemma 2 are satisfied, the upper block $\mathbb{E}\left[\dfrac{q'F'(\tilde{\mathbf{x}}^\intercal\theta)}{F(\tilde{\mathbf{x}}^\intercal\theta)^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal\right]$ is a positive-definite matrix, hence all its principal minors are positive. Condition (ii) implies that the determinant of $-(B + B^\intercal)/2$ is positive. Hence, by Sylvester's criterion $-(B + B^\intercal)/2$ is positive-definite. ∎