BRIDGING THE DIGITAL DIVIDE AND MITIGATING CYBER SECURITY RISKS IN CANADA

Peter MacKenzie

A DISSERTATION SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN ECONOMICS

YORK UNIVERSITY

TORONTO, ONTARIO

May 2025

Abstract

Canada's rapid digital transformation has created significant opportunities but also intensified existing inequalities and cyber security vulnerabilities. To better understand these challenges, an analysis is conducted at both individual and firm levels using recent Statistics Canada data and advanced econometric methods. At the individual level, data from the 2020 Canadian Internet Use Survey reveal how socioeconomic and demographic characteristics such as age, education, income, gender, and Indigenous identity influence digital engagement. A survey-weighted debiased Lasso logit model captures complex interactions among these factors, while cluster analysis assesses how provincial pandemic measures affected internet use and digital adoption during COVID-19.

Firm-level analysis incorporates data from the 2021 Canadian Survey of Digital Technology and Internet Use and the 2021 Canadian Survey of Cyber Security and Cybercrime. A Business Digital Usage Score quantifies firms' adoption of advanced technologies such as cloud computing, data analytics tools, and artificial intelligence. Stochastic frontier analysis evaluates how close firms are to their technological frontier. A survey-weighted debiased Lasso logit model identifies factors associated with both digital adoption and cyber security vulnerabilities across industries and firm sizes.

The Independence of Irrelevant Alternatives assumption in Multinomial Logit models is critically evaluated through simulation experiments examining the performance of the Hausman-McFadden (HM) specification test. The HM test is assessed under a number of data-generating scenarios used to mirror real-world applied research scenarios. To address

challenges posed by high-dimensional data, the study introduces a Hausman test based on a debiased Lasso estimator.

Acknowledgments

I would first like to express my sincere gratitude to my supervisor, Dr. Joann Jasiak, whose guidance, mentorship, and support have been invaluable throughout my doctoral journey. I am equally grateful to my co-supervisor, Dr. Purevdorj Tuvaandorj, for his collaboration, insightful feedback, and mentorship, which greatly enhanced the quality of my dissertation.

My thanks extend to my fellow PhD students and colleagues at York University, who have made my graduate experience both enjoyable and enriching.

Finally, I would like to thank my parents, John and Janet, as well as my brother, Michael, for their love, encouragement, and support, which have been instrumental in my achievements. Most importantly, I thank my fiancée, Rachel, whose patience, love, and unwavering belief in me have carried me through this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Digital technologies are fundamentally reshaping economies and societies around the world. Innovations such as cloud computing, digital payments, and the Internet of Things have improved productivity and connectivity. However, these changes have also introduced challenges including unequal access to technology and increased cyber security risks. Canada provides an important case study as significant investments have been made to expand digital infrastructure. Yet, inequalities in technology access persist across regions and demographic groups. These issues are critical for policy makers, businesses, and researchers aiming to promote inclusive economic growth and technological advancement.

The motivation for this thesis is to better understand the digital divide and cyber security risks in Canada. The research explores these issues at two levels: individuals and firms. At the individual level, the thesis investigates how socioeconomic and demographic characteristics shape digital engagement. At the firm level, the research focuses on patterns of technology adoption and the relationship between digital advancement and cyber security vulnerability. The thesis makes methodological contributions by addressing econometric challenges that arise in these choice set contexts.

Four primary research questions guide this work. First, how do socioeconomic and demographic factors influence individual digital engagement and literacy in Canada? Second, how

did the COVID-19 pandemic influence internet usage and digital adoption among Canadians? Third, what factors determine Canadian firms' adoption of advanced digital technologies and their resulting exposure to cyber security risks? Fourth, how can econometric modeling techniques used to study these questions be improved, particularly in high-dimensional settings?

This thesis consists of three related papers, each addressing aspects of these research questions and together providing a coherent framework for understanding digital inequality and cyber security in Canada.

The first chapter examines individual-level digital inequality using data from the 2020 Canadian Internet Use Survey (CIUS). This chapter identifies the main socioeconomic and demographic determinants of digital engagement including age, income, education, gender, and Indigenous status. It introduces a Digital Literacy Score that quantifies individuals' ability to effectively use digital technologies for activities such as online banking and digital payments. Methodologically, this chapter uses a survey-weighted Lasso logistic regression model designed to handle numerous variables and interactions. Results indicate significant disparities in digital engagement and literacy, highlighting populations that require targeted support. The COVID-19 pandemic's impact on digital behavior is also examined, revealing varied patterns of adoption across provinces and demographic groups.

The second paper analyzes digital technology adoption and cyber security practices among Canadian firms. It utilizes data from two 2021 Statistics Canada surveys: the Survey of Digital Technology and Internet Use and the Survey of Cyber Security and Cybercrime. The paper develops a Business Digital Usage Score that quantifies how extensively firms have adopted technologies such as cloud computing and enterprise management systems. Using stochastic frontier analysis, this research measures firms' efficiency in adopting these technologies. It further employs high-dimensional logistic regression models with debiased Lasso methods to explore determinants of both technological efficiency and cyber security vulnerabilities. The results highlight industry-specific and firm-size-specific differences in digital adoption and cyber security practices, providing insights for policy aimed at enhanc-

ing productivity and reducing risk.

The third paper contributes methodologically to the econometric literature on discrete choice modeling. It critically evaluates the Independence of Irrelevant Alternatives (IIA) assumption underlying Multinomial Logit (MNL) models. This chapter conducts simulation experiments to assess the performance of the Hausman-McFadden test statistic for evaluating the IIA assumption, especially in high-dimensional settings. A Hausman test based on a debiased Lasso estimator is developed. This approach attempts to provide more reliable results in empirical research involving a large number of explanatory variables. The methodological advances in this paper help to ensure robust empirical analyses for economists working with complex data structures.

Together, these three chapters make substantive and methodological contributions to the literature. They provide updated empirical evidence on digital inequality and cyber security risk in Canada. They identify specific vulnerable populations and industry sectors requiring targeted interventions. Methodologically, they advance econometric techniques applicable beyond this context, helping to improve inference in empirical economic studies.

The remainder of the thesis is structured as follows. Chapter 2 investigates the determinants of digital engagement and literacy among Canadians, focusing on individual-level inequalities and the pandemic's impact. Chapter 3 extends the analysis to the firm level, assessing digital technology adoption, productivity, and cyber security risks. Chapter 4 contributes by evaluating the IIA assumption in Multinomial Logit models and proposing a debiased Lasso-based Hausman test. Chapter 5 concludes by summarizing key findings, policy implications, and directions for future research.

Chapter 2

Digital Divide: An Empirical Study of CIUS 2020

Coauthorship Statement

This chapter is based on joint work titled “Digital Divide: Empirical Study of CIUS 2020" with Dr. Joann Jasiak and Dr. Purevdorj Tuvaandorj. All authors contributed equally to the conceptualization, data preparation, empirical modeling, analysis, and writing of the manuscript.

A version of this chapter is currently under second round revision review at `The Canadian Journal of Economics`.

2.1 Introduction

The digital divide represents a gap between those who can fully participate in the digital world and those who cannot. It is determined by the availability of digital infrastructure, such as high- speed internet, and the ability and willingness of individuals to engage with digital technologies, which depend on their socio-economic and demographic characteristics.

As Canada and other major economies explore the implementation of “digital money” or

Central Bank Digital Currencies (CBDC), it becomes crucial to understand the extent of the digital divide in Canada and reveal the characteristics of Canadians affected by it. Limited individual engagement with digital technologies is an obstacle in the advancement of internet-based services, including digital banking, education, and emerging financial technologies such as CBDC. Individual connectivity is necessary to ensure a balanced growth of the economy and fair participation for Canadians in an increasingly digital society.

The Canadian government has taken significant steps to develop internet availability by investing heavily in digital infrastructure. As a result of the High-Speed Access for All: Canada's Connectivity Strategy, along with a $1.7 billion investment from the 2019 federal budget, 94.15% of Canadians reported having internet access at home in 2020. However, access to the internet does not ensure the use of internet, and on-line services. Individuals must be able to afford internet services and devices, have access to them, and be willing to use them.

This paper uses the 2020 Canadian Internet Use Survey (CIUS) to investigate how socio-economic and demographic characteristics of Canadians influence their engagement with digital technologies. CIUS 2020 contains more information than the previous installments of CIUS conducted by Statistics Canada, especially on the use of internet for online banking and digital payments. It also includes more individual characteristics, such as visible minority status and Aboriginal identity. Our objective is to provide an updated and comprehensive study to inform Canadians and future policymakers given Canada's aging society and the increasing role of digital banking and cashless transactions.

Our first contribution is in applying survey-adapted Lasso inference methods to identify key socio-economic and demographic individual characteristics that determine the use of internet, e-mail, online banking and digital payments through virtual wallets and credit cards. The novel `svy LLasso` estimator for logit models (Jasiak and Tuvaandorj, 2023) allows us to analyze a large number of explanatory variables and their interactions, while incorporating survey weights to ensure the results are representative of the Canadian population. We also

use multiple correspondence analysis (MCA) of qualitative socio-demographic variables from CIUS 2020 to reveal combined effects of multiple individual characteristics.

Limited usage of the internet by the individuals who can access and afford it is often caused by a poor level of digital literacy, defined as "the ability to use information and communication technologies to find, evaluate, create, and communicate information, requiring both cognitive and technical skills" (American Library Association, n.d.). Our second contribution is in approximating digital literacy in Canada by designing a composite score based on digital technology usage. This score allows us to compare and rank digital literacy across various population segments.

The COVID-19 pandemic has highlighted the importance of both digital access and digital literacy. To study the effects of the pandemic, we use cluster analysis to identify two distinct groups of Canadians: one that increased their use of digital technologies during the pandemic and another that was less inclined to embrace these changes. We then observe that the COVID-19 measures introduced by different provinces possibly influenced digital adoption rates.

In our empirical study, we reveal several interesting new results based on CIUS 2020 reflecting recent trends in Canadian society. We find that women are more likely than men to use email and score higher in digital literacy. Recent immigrants and visible minorities score high in digital literacy too. Among the recent immigrants the English-speaking ones use more online banking. We also observe that visible minorities are more frequently using virtual wallets. In contrast, certain groups of individuals such as those age 65 and over, those with low income, the unemployed, single older individuals, and those with only a high school education, especially residents of Manitoba and Maritime provinces — are becoming more disconnected from an increasingly digital society. While the CIUS data excludes First Nations on reserve, we find that off-reserve First Nations use less internet and score lower on digital literacy than non-Aboriginals. This raises concerns given Canada's aging population and slow progress of truth and reconciliation initiatives, and emphasizes the need to address

the digital divide affecting the disadvantaged groups.

Consistent with earlier research, we find that age, income, and education influence digital technology usage. These individual characteristics were found relevant in the past study of internet use and online activity level in Canada by Haight et al. (2014) based on CIUS 2010, and remain statistically significant, in addition to the gender and recent immigrant status. Similar findings have also been reported in empirical research conducted abroad (Reddick et al., 2020; Robinson et al., 2015; Cullen, 2001; Friedline et al., 2020). However in the past, women and recent immigrants were found to be less likely to use online activities in Canada (Haight et al., 2014) and the U.S. (Zickuhr and Smith, 2012). We observe that internet connectivity among women has increased, and the access gap between immigrants and Canadian citizens identified by Haight et al. (2014) has diminished. The high digital literacy of women, especially those employed, as documented in our study, may be attributed to their growing participation in STEM and technology-intensive fields. In addition, Canada's high-skilled immigration policies and Federal Skilled Trades Program may be positively influencing the digital literacy and technology usage of immigrants.

The availability of the visible minority status in CIUS 2020 allows us to provide new results, which are mostly encouraging. However, visible minorities may not be accessing internet and digital payments equally in all provinces. Our study of interaction effects reveals that visible minority in Manitoba are less likely to use email, which points to the problem of regional disparities in Canada.

In Canada, the digital divide has traditionally been characterized by the rural-urban gap, with urban areas generally exhibiting higher levels of digital engagement compared to rural counterparts (Carson, 2013). Our results indicate that rural residents are not only less likely to use the internet but also emails and virtual wallets, with this issue concerning both Ontario and Quebec. However, we find no evidence of rural residents using less online banking and credit cards than the urban residents. Because of the limited scope of CIUS (2020), our study does not cover on-reserve Aboriginal communities, whose access to broadband internet

use in Canada is discussed by (Koch, 2022) in the context of the aforementioned federal Connectivity Strategy in rural communities and First Nations reserves.

Recent studies have also explored the intra-urban divide, focusing on disparities driven by factors like education, income, and other socio-economic variables (Reddick et al., 2020; Dewan and Riggins, 2005; Wavrock et al., 2022). These studies show that significant inequalities persist, particularly among vulnerable groups such as low-income households and seniors. We complement these findings by studying the interactions of variables, including low income and age. In particular, low income or single Canadians who are more than 65 years old, or single and French-speaking Canadians are found in our study to be disadvantaged.

Koch (2022), Reddick et al. (2020), and Van Deursen and Van Dijk (2019) have underscored the importance of digital literacy and usage in understanding the full scope of the divide. Reddick et al. (2020) and Van Deursen and Van Dijk (2019) argue that digital literacy is a major obstacle in access to broadband internet in the U.S. and Netherlands respectively. Koch (2022) addresses the importance of designing government funded initiatives to improve the digital literacy in Canada. Our digital literacy score is a new instrument of analysis indicating clearly which segments of Canadian population need to be given priority in this respect: old, low income, with low educational attainment, First Nations, and residents of Maritime provinces.

The COVID-19 pandemic caused significant changes in the digital lives of Canadians (Koch, 2022; Wavrock et al., 2022; Engert and Huynh, 2022). With physical distancing measures and stay-at-home orders, Canadians increasingly turned to digital platforms for work, shopping, education, and social interaction (Aston et al., 2020; Deng et al., 2020). Education and income levels played a crucial role in determining internet access and the ability to fully participate in this digital shift (Wavrock et al., 2022), which was not homogeneous across Canada according to our results. Our evidence shows that relatively less Canadians adopted new digital technologies during the pandemic in the Maritime provinces.

In the existing literature, the digital divide is conceptualized across three levels: access (material access to technology), usage (the ability to effectively engage with digital tools), and outcomes (the tangible benefits resulting from digital usage) (Van Deursen and Van Dijk, 2019; Ferreira et al., 2021). Early research predominantly concentrated on the first-level digital divide, which pertains to the supply of internet and physical access to technology (Cullen, 2001; Van Dijk and Hacker, 2003), whereas recent research has shifted focus toward the demand side—the factors influencing usage and outcomes.

Among the 5.85% of the Canadian population without internet access, both demand-side and supply-side barriers persist (Jordan, 2019). Demand-side barriers account for the majority of reasons and include lack of interest (50.83%), high service costs (26.08%), and the high cost of equipment (2.48%). Supply-side barriers, such as service unavailability, represent a smaller portion, accounting for 6.75% of non-adoption reasons. These barriers limit people's connectivity to the internet and digital technologies, with demand-side issues being the predominant obstacles.

On the demand side, personal characteristics and digital literacy, influence the preferences for digital technologies impacting their usage. For instance, Chen, Engert, Huynh, O'Habib, Wu and Zhu (2022) find that the use of debit and credit cards has generally increased since the pandemic; however, some subsets of Canadians continue to prefer cash for transactions (Henry et al., 2023; Engert and Huynh, 2022). This indicates that even when access is available, demand-side factors such as personal preferences and digital literacy affect the usage of digital technologies and the benefits derived from them.

As digital payment adoption rises, significant challenges persist for many First Nations communities, including limited internet access and difficulties maintaining access to cash (Chen, Engert, Huynh and O'Habib, 2022). The shift from traditional to digital banking has altered consumer behavior and accelerated the closure of physical bank branches as institutions prioritize digital optimization (Aversa et al., 2022). While this shift offers convenience, it raises concerns about financial exclusion (Kamdjoug et al., 2021), in particular given lower

digital literacy of (off-reserve) First Nations documented in our study. Cash usage in Canada has declined sharply, with only one in three transactions now involving physical cash (Huynh, 2017). A concern of the Bank of Canada is the increasing interest in cryptocurrency which are not regulated and highly volatile. Despite global interest, cryptocurrency adoption so far remains low in Canada (Huynh et al., 2020; Adrian and Mancini-Griffoli, 2019). The data on cryptocurrency use is available in CIUS (2020). However, the sample of respondents is too small for valid inference and virtual wallets are explored instead in this paper. Our study provides reliable data-based insights that can help identifying groups to be disadvantaged in a future cashless economy, or with a fully digitalized banking system.

The paper is organized as follows: Section 3.2 describes the CIUS 2020 dataset. Section 2.3 lays out the paper's estimation and inference approach. Section 2.4 presents the `svyLLasso` results, MCA diagrams, and digital literacy approximations. Section 2.5 provides additional analyses examining the impact of COVID-19 on the digital divide and a comparison with the results of Haight et al. (2014) on CIUS 2010. We conclude in Section 2.6. The online appendix provides a description of the sampling and weighting scheme used in CIUS 2020, technical details of the methods used in the paper, further information on the digital literacy score, and additional estimation results.

2.2 Data description

This section describes the CIUS 2020 survey and the variables used in our empirical analysis. CIUS 2020 is the most relevant data source on Canadian internet usage and comprises $17,409$ observations on households across Canada. The survey includes answers from Canadians 15 years of age and older living in one of Canada's ten provinces. The survey has a cross-sectional design, which uses both landline and cellular phone numbers from Statistics Canada's dwelling frame. Statistics Canada uses stratified sampling at the census metropolitan area and census agglomeration level. The overall response rate to the survey is 41.6%.

CIUS 2020 data are appropriately weighted using sample weights. Statistics Canada provides the weight variables, which are based on independent estimates for various age and sex groups in each province and account for survey non-response, among other factors (see Appendix A for the stratification scheme and survey weights). Properly weighting the data allows the sample of the Canadian population used in CIUS 2020 to accurately represent the entire population.

To assess the digital divide, we study the demographic and socio-economic characteristics of CIUS 2020 respondent, which appear as the explanatory variables in logit models of the use of the internet and selected internet-based services. Sections 2.2.1 and 2.2.2 describe these dependent and explanatory (independent) variables, respectively.

2.2.1 Dependent variables

We consider the logit models of internet use and of the use of internet-based services, which are internet use, email use, online-banking, virtual wallet, and credit card payments. The first variable (internet use) reveals the social connectivity of Canadians. The latter four dependent variables are chosen to examine the readiness of Canadians to transition towards digital financial technologies.

The following five questions from CIUS 2020 serve as the dependent variables for our logistic models:

- **Internet use (Model 1):** "During the past three months have you used the internet from any location?" This binary question, with responses *Yes* or *No*, helps determine the factors affecting whether a Canadian individual has access to the internet.

- **Online banking (Model 2):** "During the past three months have you conducted online banking?" This question gauges the demographic factors influencing a person's proficiency and trust in conducting online financial transactions.

- **Email use (Model 3):** "During the past three months have you sent and received

emails?" The use of emails is a basic marker of digital literacy and provides insights into the user's familiarity with standard online communication tools.

- **Virtual wallet usage (Model 4):** "During the past twelve months have you used a virtual wallet to pay for goods over the internet?" This question identifies factors affecting whether Canadians use virtual wallets for payments.

- **Online credit card use (Model 5):** "During the past twelve months did you use a credit card previously entered or entered at the time of purchase to pay for goods over the internet?" This question provides insights into the trust and usability of online financial transactions among Canadians.

In Models 2–5, we categorize the possible responses as 1) *Yes*, 2) *No*, and 3) *Not stated*. We test the Independence of Irrelevant Alternatives (IIA) hypothesis to determine whether to include the *Not stated* category in Section 2.4.1.

Sample sizes

Model 1, concerning internet use, encompasses the full sample with 17,409 respondents. For Models 2 and 3, representing online banking and email use respectively, we excluded the *Not stated* responses based on the results of the Hausman test for IIA. Model 2 is thus based on 17,135 respondents after excluding 274 *Not stated* responses, while Model 3 comprises 17,268 respondents following the exclusion of 141 *Not stated* responses.

Models 4 and 5 are each based on data from 12,124 respondents. The reduction of the sample size, when compared to the full sample, arises from the sequential structure of CIUS design. The survey filters respondents based on their internet usage and other specific activities, such as expenditure on digital goods and services. As a result, certain alternatives have a probability of zero. Additionally, 307 responses were marked as *Not stated* and excluded based on the results of a Hausman test for IIA.

2.2.2 Explanatory (Independent) variables

The selected explanatory variables in the logistic regression models 1 to 5 provide a comprehensive profile of the respondents, capturing their socioeconomic, and demographic characteristics. These encompass income, education, employment status, Aboriginal identity, visible minority status, immigration status, age, gender, location, type of household, language spoken at home, and province. All variables are multi-categorical and detailed in the regression tables.

For many explanatory variables, a *Not stated* category exists as well and is retained in the regressions. Exclusion of this category could introduce bias, given that respondents who selected *Not stated* for one question often provided answers to others.

Each model omits the categories associated with a representative individual as the comparison category for the logistic regression. That representative individual has the following characteristics – urban, age 45–54, male, non-Aboriginal, English and non-official language speaker, not employed, some post-secondary education, not a visible minority, family household with children under 18, income of $\$52,204$–$\$92,485$, landed immigrant (recent immigrant), and from the province Alberta.

2.3 Survey-weight adjusted logit Lasso inference

We consider 41 explanatory (independent) categorical variables. Some of these variables are expected to have direct effects on the dependent variables of the model, while others are included to account for potential interaction effects. For instance, household type and income variables may exhibit cross-effects on dependent variables like internet use and online banking. Accounting for second-order interactions results in 674 control variables, a number that is relatively large compared to the sample size. However, there is no a priori guidance on which variables should enter the model.

In situations where a model contains many regressors, Lasso variable selection techniques

are known to flexibly reduce the dimensionality of the data and select variables with higher predictive power for explaining the categorical dependent variables of interest. For these reasons, we adopt the logistic Lasso approach, well-suited for this problem. It possesses optimality properties under a sparsity assumption and leads to automatic variable selection (Belloni et al., 2014; Mullainathan and Spiess, 2017).

The survey weights play a crucial role in ensuring the generalizability of survey results to the entire Canadian population. However, existing Lasso-based estimation and inference methods, including the commonly used logit Lasso variable selection, require adjustment for survey weights. This paper employs a new logistic Lasso variable selection method for binary choice models in a survey environment, termed the `svy LLasso`, which is described below. The asymptotic properties of the `svy LLasso` estimator are given in Jasiak and Tuvaandorj (2023).

Let θ denote the parameter vector of the logistic regression including the slope parameters β and intercept α. The (non-negative) tuning parameter used in the Lasso is denoted by λ. A survey-weighted logistic Lasso is based on minimizing the weighted negative log-likelihood function $L(\theta)$ subject to ℓ_1 penalty on the parameter vector:

$$\min_{\theta=(\alpha,\beta')'\in\mathbb{R}^{p+1}} \left(-L(\theta) + \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{2.3.1}$$

where $L(\theta) = n^{-1} \sum_{i=1}^{n} w_i (y_i x_i'\theta - \log(1+\exp(x_i'\theta)))$, $x_i'\theta = \alpha + \tilde{x}_i'\beta$, and $(y_i, x_i')' \in \mathbb{R}^{p+1}, i = 1, \ldots, n$, are the pairs of dependent and independent observations with the corresponding strictly positive survey weights $w_i, i = 1, \ldots, n$. The sampling scheme used in CIUS 2020 is akin to simple stratified sampling (Cameron and Trivedi, 2009), so we treat w_i as given, and $\{(y_i, x_i')'\}_{i=1}^{n}$ as independent.

Note that, as is standard in the Lasso literature, only the "slope" parameters in $\beta = (\beta_1, \ldots, \beta_p)'$ are penalized in (2.3.1). We fit the model (2.3.1) using the `R` package `glmnet`. For the tuning parameter λ, we use the package's default value chosen by 10-fold cross

validation with the loss function "`auc`" (area under the ROC curve).

Prior to inference being made on the coefficients, the `svy LLasso` estimator needs to be transformed to ensure valid results. Due to its computational and conceptual simplicity, we use a survey-version of the debiased Lasso (DB) method proposed by Zhang and Zhang (2014), Javanmard and Montanari (2014) and Xia et al. (2020) as the main inferential tool for the logit coefficients and the average marginal effects (AMEs) after variable selection by `svy LLasso`. It is based on the following one-step estimator constructed from the initial `svy LLasso` estimator $\hat{\theta}$:

$$\tilde{\theta}^{DB} \equiv \hat{\theta} + H(\hat{\theta})^{-1} S(\hat{\theta}),$$

where $H(\cdot)$ and $S(\cdot)$ are the (sample) Hessian and the score functions for the full parameter vector in the logistic model. The one-step (or DB) estimator removes the bias of the initial `svy LLasso` estimator and has an asymptotic normal distribution, thus facilitating standard t-ratio-based inference.

An alternative transformation method considered is the survey-logit versions of the selective inference (SI) procedure proposed by Lee et al. (2016) and Taylor and Tibshirani (2018), and the $C(\alpha)$ (or Neyman orthogonalization) method after Lasso variable selection proposed by Belloni et al. (2016) to make inference on the model parameters and AMEs. The former method is based on a one-step estimator denoted as $\tilde{\theta}^{SI}$ and the test statistic in the latter is labelled as C_α. See Appendix A.1.1 for a brief description of these methods and Jasiak and Tuvaandorj (2023) for further theoretical analyses.

2.4 Empirical results

This section reports the empirical results. Section 2.4.1 shows the `svy LLasso` logit estimation results for Models 1-5. We analyze the logit models with interaction effects in Section 2.4.2. We report the outcomes of the multiple correspondence analysis in Section 2.4.3 and present the digital divide score in Section 2.4.4.

2.4.1 svy LLasso logit models

As stated in Section 2.2, the online banking, email use, virtual wallet, and credit card dependent variables have three categories: *Yes*, *No*, and *Not stated*. We use first the survey-weighted Hausman-McFadden test of the IIA hypothesis to see if we can remove the *Not stated* observations from the logit models. The online banking variable has a Hausman-McFadden statistic of 0.05 with a p-value of 1, which is strong evidence in favor of IIA. Hence, we use the restricted specification of Model 2, removing the *Not stated* observations from the model. The dependent variables email use, virtual wallet, and credit card use have Hausman-McFadden statistics $-0.95, -0.77$, and -1.29. Therefore, conventionally, the *Not stated* observations are removed from these models as well.

We report the empirical results, including the `svy LLasso` estimates and the test results based on the debiased Lasso estimates of the logit model coefficients and AMEs, $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ in Tables 2.1–2.5 below. Tables A.2–A.6 in the online Appendix A.3 present the results of the selective inference and $C(\alpha)$ procedures, which are consistent with the debiased Lasso results.

The estimation of the internet use Model 1 reveals which explanatory variables influence a person's internet connectivity. The estimation of Models 2-5 shows which explanatory variables are essential for the use of internet-based devices. We consider evidence of a digital divide in Models 1-5 under the following conditions: a) When some explanatory variables are selected by the `svyLLasso` and statistically significant while others are not, the divide is between individuals with and without the characteristics represented by the selected variables. b) When some categories within an explanatory variable are selected by the `svyLLasso` and statistically significant while others are not, this suggests a divide between individuals belonging to the selected categories and those who do not. c) When all categories of an explanatory variable are selected by the `svyLLasso` and statistically significant, but the coefficients either have different signs or take noticeably different values, the digital divide is indicated by these distinctions in coefficient signs or magnitudes.

Model 1: Internet use. Table 2.1 presents the results based on the internet use model and reveals the explanatory variables influencing an individual's connectivity and usage of the internet.

The model reveals a rural-urban divide in internet access. Specifically, rural Canadians have a 1.7% lower probability of having used the internet in the prior three months compared to their urban counterparts. This disparity corroborates the findings of *Canada's connectivity strategy*. Despite substantial federal investments to bolster rural internet access, this discrepancy persists, emphasizing the enduring challenges rural residents confront in bridging the digital divide.

All age group categories are selected by `svy LLasso` and are statistically significant. Younger age brackets, specifically those between 15 and 44, have positive coefficients and AME values. In contrast, those aged 55 and above have negative coefficients and AME values. Individuals in the *25-34* age category are 5.4 percentage points more likely to be internet users, while the eldest group (*65 and older*) is 8.2 percentage points less likely compared to the reference group of *45-54* years.

Several demographic factors are selected by `svy LLasso` and statistically significant. For instance, those who are employed, predominantly English speakers, university graduates, and high earners are more likely to use the internet. Conversely, individuals residing in the province of Quebec who are older, have a high school education or less, identify as a visible minority, are single, and have low incomes have a lower likelihood of internet usage.

The results for internet connectivity are generally close to the findings of past research on internet connectivity in Canada (Haight et al., 2014; Friedline et al., 2020; Jordan, 2019). However, there are differences concerning the gender or immigration variables, which are not selected by `svy LLasso` or found to be statistically significant for internet use in our analysis.

Model 2: Online banking. Table 2.2 presents the findings from the online banking model, exploring the factors that influence Canadians' adoption and use of digital financial technologies. The online banking model is of particular significance, as the current online banking systems may share functional parallels with potential digital financial technologies, like a CBDC system.

The results indicate that younger, employed, high-income, and university-educated Canadians are more inclined to utilize online banking. Factors such as lower educational attainment, lower income, identification as a visible minority, and being aged 55 or older reduce the likelihood of online banking usage. As indicated by the absolute value of AMEs, the most impactful variables include the age category of *65 and older*, employment status, and educational attainment of *High school or less*.

Individuals in the *65 and older* age category display a notable divergence in behavior, being 15.4 percentage points less likely to use online banking than those in the *45-54* age group. Employment status is another prominent determinant; specifically, those employed exhibit a 10.6 percentage point heightened likelihood of using online banking relative to their unemployed counterparts. Educational credentials further accentuate the divide. Those whose highest educational achievement is *High school or less* are 11.4 percentage points less likely to engage in online banking than individuals with at least *Some post-secondary* education. Notably, the `svy LLasso` did not select variables such as *Location, Gender, Aboriginal identity*, and *Province* as influential determinants in the model.

Model 3: Email use. Table 2.3 presents the results for email usage, a metric that gauges Canadians' digital social and professional connectivity. Email is one of the most commonly used internet service among those that are connected. While there are similarities in the variables influencing both email usage and online banking, as seen in the selections by `svy LLasso` and the statistically significant explanatory variables, there are also intriguing distinctions.

One notable difference is the influence of location. While the *Rural* category is associated with a reduced likelihood of email use, it does not significantly affect online banking. A plausible interpretation is geographical necessity: rural Canadians might lean towards online banking due to their distance from physical bank branches. Additionally, rural employment might not demand as extensive email communication as certain urban jobs.

Gender dynamics offer another dimension of differentiation. The *Female* category is influential in the email use model, though its impact, as evidenced by the debiased Lasso AME, is relatively modest. The difference in email use based on gender could reflect occupational patterns, with women potentially occupying more office roles that necessitate email, in contrast to blue-collar roles that might be more prevalent among men.

In Table 2.3, the variable with the largest estimated AME (in absolute value) is the language variable category *English, French, and Non-official language*. However, despite the large AME estimate, the variable is not selected by `svy LLasso`. The oldest age category, *65 and older* is selected by `svy LLasso` and is statistically significant. This category has the second largest AME; those *65 and older* are 10 percentage points less likely than those in the age group *45-54* to send and receive emails.

Educational background influences email usage patterns. Those with a *University degree* or higher exhibit a stronger propensity for email use than those with only *Some post-secondary education*. Individuals with an education level of *High school or less* show a diminished likelihood. The trend seen in educational attainment might arise from the nature of jobs accessible at different educational levels, often linking higher education qualifications to roles that require frequent email communication.

Model 4: Virtual wallet. Some internet users make online payments using virtual wallets. Table 2.4 details the explanatory variables influencing the adoption of virtual wallets in Canada, a crucial variable for research on digital currencies in the country. While previous models assessed Canadians' internet connectivity and use of other digital technologies, the

virtual wallet model will show what factors currently affect the uptake of digital payment methods.

Age emerges as an important determinant in virtual wallet use. All age group categories, excluding those aged *35-44* were selected by `svy LLasso` and statistically significant. Younger Canadians have the highest probability of using a virtual wallet. The age group, *15-24*, has an 11.2 percentage point increase in the likelihood of using a virtual wallet than the base age group of *45-54*. In contrast, older Canadians, especially those *65 and older*, demonstrate a decreased likelihood. The age group *65 and older* is the least likely to use a virtual wallet compared to the age group *45-54*. The debiased Lasso AME for the oldest age group shows that those *65 and older* are 8.3 percentage points less likely than the reference age group to use a virtual wallet.

The coefficient for *Visible minority* is chosen by `svy LLasso` and statistically significant. *Visible minority* has a positive AME on the use of a virtual wallet. This result is striking, considering the variable category *Visible minority* in previous results has either not been selected by `svy LLasso` or had a negative effect on the dependent variable. The AME shows that a person identifying as a visible minority is 5.2 percentage points more likely to use a virtual wallet than a person who is not a visible minority. The positive *Visible minority* coefficient might reflect the increased use of foreign cryptocurrencies like Alipay and WeChat pay by visible minorities in Canada.

Income and education stand out as influential variables. Specifically, Canadians earning $\$146,560$ or more are more likely to use a virtual wallet than those in the base income category of $\$52,204$–$\$92,485$. Additionally, individuals holding a *University degree* are 8 percentage points more likely to use a virtual wallet than those with *Some post-secondary* education.

Model 5: Credit card. Credit card payments are the most popular way to make purchases online. Table 2.5 presents findings on the explanatory variables influencing Cana-

dians' use of credit cards for online transactions—a crucial understanding considering the anticipated card component of a potential CBDC. Given the frequent use of credit and debit cards in the current Canadian financial landscape, these insights are pivotal for the successful integration of a CBDC.

The model reveals some interesting results. Again, age is a significant determinant: younger Canadians, specifically those in the *15-24* age bracket, are 8.8 percentage points less likely to utilize a credit card for online purchases when compared to the reference group of *45-54*. Education emerges as another prominent factor. Individuals with a *High school or less* education level show a reduced likelihood for online credit card transactions. Those with a *University degree* are more inclined towards such transactions.

Economic and regional factors have an impact on whether Canadians use credit cards for online shopping. People residing in Quebec and those in the lowest income bracket are less likely to use credit cards for online purchases. On the other hand, English-speaking Canadians, those with a university degree, people who are employed, residents of family households without children under 18, and Ontarians are more likely to use credit cards for online shopping.

Summary of results. Our analysis identifies age, education, and income as key factors in digital adoption: younger, more educated, and higher-income individuals are more digitally engaged, revealing a socio-economic digital divide. Contrary to previous research, immigration status and gender did not have an impact on the digital divide, with immigrants and women displaying similar levels of digital engagement as other groups. Additionally, visible minorities are increasingly adopting new technologies like virtual wallets, and younger demographics prefer alternative digital payment methods over traditional ones, indicating a shifting digital landscape.

Table 2.1: Lasso Logistic Regression Results for Internet Use Dependent Variable

Variables	Categories	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Intercept		3.428	3.246^{***}	0.000	–	–
Location	Rural	−0.225	-0.287^{***}	0.001	-0.017^{***}	0.001
Age	15–24	0.627	1.235^{***}	0.000	0.054^{***}	0.000
	25–34	0.161	0.683^{**}	0.007	0.033^{*}	0.014
	35–44	0.038	0.548^{*}	0.016	0.027^{*}	0.035
	55–64	−0.721	-0.527^{**}	0.003	-0.032^{*}	0.014
	65 and older	−1.570	-1.262^{***}	0.000	-0.082^{***}	0.000
Gender	Female	0.013	0.099	0.200	0.006	0.211
Aboriginal	Aboriginal	–	-0.497^{*}	0.021	-0.032^{*}	0.011
Language	English	0.354	0.598^{*}	0.037	0.035^{*}	0.044
	French	–	0.246	0.435	0.013	0.464
	Non-official	–	0.065	0.836	0.004	0.842
	English and French	–	0.793	0.124	0.036	0.231
	French and Non-official	–	−0.533	0.544	−0.035	0.495
	English, French and Non-official	–	−1.434	0.193	$-0.118^{\cdot}$	0.067
Employment	Employed	0.514	0.574^{***}	0.000	0.032^{***}	0.000
Education	High school or less	−0.911	-0.971^{***}	0.000	-0.058^{***}	0.000
	University degree	0.451	0.519^{***}	0.000	0.027^{***}	0.000
Minority	Visible minority	−0.048	-0.352^{*}	0.037	-0.021^{*}	0.034
Household type	Family w/o children under 18	–	−0.029	0.872	−0.002	0.875
	Single	−0.596	-0.665^{***}	0.000	-0.043^{***}	0.001
	Other household type	–	0.149	0.635	0.008	0.656
Income	$52,203 and lower	−0.536	-0.475^{***}	0.000	-0.028^{***}	0.000
	$92,486–$146,559	–	0.092	0.469	0.005	0.486
	$146,560 and higher	0.359	0.547^{***}	0.001	0.028^{***}	0.001
Immigration	Non-landed immigrant	–	−0.258	0.176	−0.014	0.211
Province	NL	–	−0.31	0.111	$-0.019^{\cdot}$	0.091
	PEI	–	−0.272	0.155	−0.017	0.135
	NS	–	−0.298	0.123	−0.018	0.104
	NB	–	−0.101	0.586	−0.006	0.585
	QC	−0.296	-0.448^{*}	0.026	-0.027^{*}	0.034
	ON	0.039	−0.018	0.911	−0.001	0.913
	MB	–	-0.501^{*}	0.013	-0.032^{**}	0.006
	SK	–	-0.413^{*}	0.037	-0.026^{*}	0.024
	BC	0.031	0.095	0.602	0.005	0.613

Note: $n = 17,409$. $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ denote the debiased Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{\cdot}$ $p < 0.1$.

Table 2.2: Lasso Logistic Regression Results for Online Banking Dependent Variable

Variables	Categories	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Intercept		1.120	0.625**	0.009	–	–
Location	Rural	–	−0.092	0.154	−0.015	0.167
Age	15–24	–	0.045	0.721	0.007	0.734
	25–34	0.414	0.637***	0.000	0.092***	0.000
	35–44	0.267	0.540***	0.000	0.079***	0.000
	55–64	−0.071	−0.324***	0.000	−0.052***	0.001
	65 and older	−0.718	−0.873***	0.000	−0.154***	0.000
Gender	Female	–	0.089	0.107	0.014	0.123
Aboriginal	Aboriginal	–	−0.248	0.105	−0.040	0.106
Language	English	0.000	0.509**	0.005	0.081**	0.007
	French	–	0.598**	0.005	0.086*	0.012
	Non-official	–	0.337•	0.090	0.050	0.122
	English and French	–	0.239	0.526	0.036	0.561
	French and Non-official	–	−0.280	0.648	−0.046	0.647
	English, French and Non-official	–	−0.127	0.858	−0.020	0.861
Employment	Employed	0.662	0.653***	0.000	0.106***	0.000
Education	High school or less	−0.637	−0.686***	0.000	−0.114***	0.000
	University degree	0.331	0.409***	0.000	0.062***	0.000
Minority	Visible minority	−0.135	−0.303**	0.003	−0.048**	0.005
Household type	Family w/o children under 18	0.078	0.305***	0.000	0.048***	0.001
	Single	−0.166	−0.137	0.137	−0.022	0.167
	Other household type	–	0.372*	0.042	0.054•	0.068
Income	$52,203 and lower	−0.265	−0.252***	0.001	−0.041**	0.002
	$92,486–$146,559	–	0.123	0.132	0.019	0.153
	$146,560 and higher	0.086	0.252**	0.004	0.039**	0.006
Immigration	Non-landed immigrant	–	−0.082	0.463	−0.013	0.486
Province	NL	–	−0.015	0.905	−0.002	0.909
	PEI	–	−0.068	0.604	−0.011	0.615
	NS	–	−0.079	0.540	−0.012	0.552
	NB	–	−0.068	0.603	−0.011	0.614
	QC	–	−0.101	0.460	−0.016	0.474
	ON	–	−0.032	0.748	−0.005	0.758
	MB	–	−0.383**	0.004	−0.063**	0.003
	SK	–	−0.114	0.377	−0.018	0.389
	BC	–	−0.010	0.931	−0.002	0.934

Note: $n = 17,135$. $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ denote the debiased Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, • $p < 0.1$.

Table 2.3: Lasso Logistic Regression Results for Email Use Dependent Variable

Variables	Categories	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Intercept		1.960	1.964***	0.000	–	–
Location	Rural	−0.158	−0.207**	0.005	−0.021**	0.007
	15–24	0.390	0.658***	0.000	0.058***	0.000
	25–34	0.444	0.742***	0.000	0.063***	0.000
	35–44	0.294	0.585***	0.000	0.051***	0.000
	55–64	−0.425	−0.343**	0.004	−0.035**	0.009
	65 and older	−1.036	−0.899***	0.000	−0.100***	0.000
Gender	Female	0.087	0.151*	0.021	0.015*	0.025
Aboriginal	Aboriginal	–	−0.473**	0.008	−0.051**	0.004
Language	English	0.402	0.301	0.179	0.030	0.207
	French	–	−0.118	0.640	−0.012	0.644
	Non-official	−0.047	−0.225	0.353	−0.023	0.357
	English and French	–	0.426	0.327	0.037	0.395
	French and Non-official	–	−0.302	0.669	−0.032	0.656
	English, French and Non-official	–	−1.908*	0.019	−0.272***	0.001
Employment	Employed	0.411	0.457***	0.000	0.045***	0.000
Education	High school or less	−0.790	−0.851***	0.000	−0.088***	0.000
	University degree	0.750	0.828***	0.000	0.072***	0.000
Minority	Visible minority	−0.192	−0.346**	0.008	−0.035*	0.011
Household type	Family w/o children under 18	–	−0.055	0.644	−0.005	0.655
	Single	−0.456	−0.571***	0.000	−0.062***	0.000
	Other household type	–	−0.052	0.824	−0.005	0.828
Income	$52,203 and lower	−0.383	−0.323***	0.000	−0.033***	0.000
	$92,486–$146,559	–	0.088	0.371	0.008	0.391
	$146,560 and higher	0.329	0.441***	0.000	0.040***	0.000
Immigration	Non-landed immigrant	0.016	0.147	0.304	0.015	0.311
Province	NL	–	−0.240	0.120	−0.025	0.111
	PEI	–	−0.174	0.265	−0.018	0.260
	NS	–	−0.387*	0.012	−0.041**	0.008
	NB	–	−0.251•	0.098	−0.026•	0.090
	QC	−0.154	−0.326*	0.044	−0.033•	0.050
	ON	0.164	0.069	0.577	0.007	0.582
	MB	–	−0.466**	0.004	−0.050**	0.002
	SK	–	−0.364*	0.021	−0.038*	0.015
	BC	0.236	0.260•	0.077	0.024•	0.073

Note: $n = 17,268$. $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ denote the debiased Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, • $p < 0.1$.

Table 2.4: Lasso Logistic Regression Results for Virtual Wallet Dependent Variable

Variables	Categories	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Intercept		−2.038	−2.650***	0.000	–	–
Location	Rural	−0.220	−0.609***	0.000	−0.057***	0.000
Age	15–24	0.300	0.867***	0.000	0.112***	0.000
	25–34	0.207	0.619***	0.000	0.075***	0.000
	35–44	–	0.334**	0.005	0.039**	0.003
	55–64	−0.308	−0.608***	0.000	−0.057***	0.000
	65 and older	−0.548	−1.009***	0.000	−0.083***	0.000
Gender	Female	–	−0.091	0.280	−0.010	0.277
Aboriginal	Aboriginal	–	0.040	0.872	0.004	0.869
Language	English	–	0.129	0.596	0.014	0.597
	French	–	0.132	0.653	0.015	0.640
	Non-official	–	−0.410	0.116	−0.040	0.153
	English and French	–	0.105	0.829	0.012	0.822
	French and Non-official	–	−0.618	0.448	−0.055	0.534
	English, French and Non-official	–	−0.907	0.359	−0.073	0.495
Employment	Employed	–	0.020	0.853	0.002	0.852
Education	High school or less	–	−0.066	0.568	−0.007	0.568
	University degree	0.027	0.254**	0.009	0.028**	0.008
Minority	Visible minority	0.162	0.453***	0.001	0.052***	0.001
Household type	Family w/o children under 18	–	0.064	0.547	0.007	0.543
	Single	–	0.033	0.797	0.004	0.793
	Other household type	–	0.121	0.621	0.014	0.606
Income	$52,203 and lower	–	0.080	0.551	0.009	0.541
	$92,486–$146,559	–	0.155	0.203	0.017	0.189
	$146560 and higher	0.233	0.563***	0.000	0.066***	0.000
Immigration	Non-landed immigrant	–	0.129	0.372	0.014	0.380
Province	NL	–	−0.270	0.178	−0.027	0.214
	PEI	–	−0.311	0.131	−0.030	0.169
	NS	–	−0.282	0.163	−0.028	0.198
	NB	–	−0.065	0.757	−0.007	0.760
	QC	–	−0.126	0.532	0.013	0.539
	ON	–	0.043	0.759	0.005	0.756
	MB	–	−0.420*	0.032	−0.040$^{\cdot}$	0.058
	SK	–	−0.215	0.270	−0.022	0.298
	BC	–	0.080	0.636	0.009	0.627

Note: $n = 12,124$. $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ denote the debiased Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{\cdot}$ $p < 0.1$.

Table 2.5: Lasso Logistic Regression Results for Credit Card Use Dependent Variable

Variables	Categories	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Intercept		1.334	1.100^{***}	0.000	–	–
Location	Rural	–	−0.125	0.134	−0.020	0.140
Age	15–24	−0.363	-0.522^{***}	0.000	-0.088^{***}	0.000
	25–34	–	0.055	0.630	0.008	0.644
	35–44	–	0.135	0.188	0.020	0.213
	55–64	–	−0.022	0.830	−0.003	0.835
	65 and older	–	−0.053	0.653	−0.008	0.662
Gender	Female	–	−0.004	0.958	−0.001	0.959
Aboriginal	Aboriginal	–	0.198	0.306	0.029	0.347
Language	English	0.216	0.019	0.928	0.003	0.933
	French	−0.192	-0.679^{**}	0.006	-0.116^{**}	0.005
	Non-official	–	−0.044	0.844	−0.007	0.849
	English and French	–	−0.185	0.646	−0.030	0.644
	French and Non-official	–	−0.777	0.263	−0.141	0.202
	English, French and Non-official	–	$-1.352^{\cdot}$	0.088	-0.266^{*}	0.035
Employment	Employed	0.002	$0.148^{\cdot}$	0.083	$0.023^{\cdot}$	0.091
Education	High school or less	−0.411	-0.453^{***}	0.000	-0.073^{***}	0.000
	University degree	0.357	0.490^{***}	0.000	0.073^{***}	0.000
Minority	Visible minority	–	-0.235^{*}	0.044	-0.037^{*}	0.046
Household type	Family w/o children under 18	0.035	0.335^{***}	0.000	0.051^{***}	0.000
	Single	–	0.317^{**}	0.002	0.046^{**}	0.005
	Other household type	–	0.161	0.430	0.024	0.463
Income	$52,203 and lower	−0.073	-0.286^{**}	0.004	-0.046^{**}	0.005
	$92,486–$146,559	–	0.097	0.306	0.015	0.328
	$146,560 and higher	–	0.084	0.393	0.013	0.415
Immigration	Non-landed immigrant	–	0.151	0.232	0.024	0.237
Province	NL	–	$-0.287^{\cdot}$	0.081	$-0.047^{\cdot}$	0.072
	PEI	–	0.078	0.637	0.012	0.655
	NS	–	−0.045	0.783	−0.007	0.788
	NB	–	−0.012	0.944	−0.002	0.946
	QC	−0.112	−0.042	0.798	−0.007	0.810
	ON	0.029	0.241^{*}	0.042	$0.037^{\cdot}$	0.051
	MB	–	0.035	0.829	0.005	0.837
	SK	–	0.022	0.891	0.003	0.895
	BC	–	0.211	0.131	0.031	0.161

Note: $n = 12,124$. $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ denote the debiased Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{\cdot}$ $p < 0.1$.

2.4.2 Interaction effects

To enhance the model's predictive ability, we include the relationships between the explanatory variables, incorporated as interaction terms in the logit Models 1-5 estimated by `svy LLasso`. First, we examine whether the second-order specification with interaction terms is more appropriate than the first-order specification in Models 1–5. Accordingly, we compare the mean-squared 10-fold cross-validation (CV) error of the adaptive Lasso estimator (see Bühlmann and van de Geer (2011) for a detailed treatment) for both specifications with and without the interactions, using `R` package `polywog`. Table 2.6 reports the result.

The results show that a linear specification for Models 1, 4, and 5 results in smaller mean-squared errors, while a second-order specification might be preferred for Models 2 and 3 in terms of the prediction error.

For Models 2 and 3, after fitting the second-order model with 674 variables by `svy LLasso`, we make inference on the coefficients using the debiased Lasso procedure.

Table 2.6: Order selection

	CV error				
Models	1	2	3	4	5
1st order	0.395	0.955	0.644	0.684	0.942
2nd order	0.396	0.944	0.643	0.692	0.945
sample size	17409	17135	17268	12124	12124

Note: The table reports the mean-squared 10-fold cross-validation error for first-order model with 41 covariates and the second-order model with 674 covariates based on adaptive Lasso estimator obtained using the `R` package `polywog`. Models 1-5 are internet use, online banking, email use, virtual wallet, and credit card, respectively.

Tables 2.7 and 2.8 present the interaction results for extended Models 2 and 3 of online banking and email usage, respectively. The tables include only those variables which are statistically significant at the 5% level, as indicated by their coefficient p-values. We have omitted significant interaction variables involving *Not stated* responses due to interpretability concerns.

In the extended online banking Model 2, alongside the notable negative influence of the variables *High school or less* and *Visible minority*, an interesting pattern of the interaction variables emerges. Specifically, the age group *15-24* interacting with *Family without children under 18* has a significant positive effect on online banking use. This result highlights the distinct digital behavior of younger individuals in specific family settings, emphasizing the role of age and household composition in digital engagement.

The interaction between the *65 and older* age group and *Single* households is particularly revealing. This combination is significant, corroborates the finding in Section 2.1 and shows the digital divide disproportionately affects older, single individuals, especially those with lower incomes. This finding is crucial as it explores demographic effects combined with socioeconomic factors, suggesting a more complex picture where subsets of the older population are particularly at a disadvantage digitally. It is a pattern that warrants attention, as it highlights a segment of the population that might be struggling to keep pace with the rapid digitization of financial services. The Canadian population is also aging, which makes these findings even more important for policymakers.

In the email usage model, the interaction of the age category *65 and older* with the lower income category (*$52,203 and lower*) further underscores this concern. The significant negative impact on digital engagement among older, lower-income individuals indicates the challenges this demographic faces in accessing and utilizing digital technologies. It paints a picture of a group being left behind in the digital landscape, emphasizing the need for targeted interventions to bridge this gap.

Overall, the second-order interaction terms illustrate the complex relationship between the use of digital technologies and the different demographic and soci-economic characteristics of the user. The results point toward the presence of the digital divide in Canada.

Table 2.7: Lasso Logistic Regression with Interactions for Online Banking Dependent Variable

Variables	Categories	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value
Intercept		1.013	3.151^{**}	0.006
Language	English	–	-2.663^{**}	0.007
	High school or less	−0.597	-1.554^{*}	0.012
	Visible minority	−0.114	-1.292^{*}	0.050
Location × *Immigration*	(Rural) × (Non-landed immigrant)	–	-0.991^{*}	0.029
Location × *Province*	(Rural) × (QC)	–	0.769^{*}	0.050
	(Rural) × (ON)	–	0.577^{*}	0.035
Age × *Language*	(15-24) × (English)	–	-1.465^{*}	0.049
	(15-24) × (English and French)	–	-5.084^{**}	0.006
Age × *Employment*	(15-24) × (Employed)	–	0.681^{*}	0.024
Age × *Education*	(15-24) × (University degree)	–	1.514^{*}	0.010
Age × *Household type*	(15-24) × (Family w/o children under 18)	0.291	1.177^{***}	0.000
	(15-24) × (Single)	–	1.087^{*}	0.025
	(15-24) × (Other household type)	–	2.096^{**}	0.006
	(65 and older) × (Single)	−0.065	-0.857^{*}	0.044
Gender × *Language*	(Female) × (English)	0.068	0.752^{*}	0.047
Gender × *Employment*	(Female) × (Employed)	0.153	0.342^{*}	0.017
	(Female) × (University degree)	–	-0.378^{*}	0.013
Language × *Education*	(English) × (High school or less)	–	1.643^{***}	0.001
Language × *Income*	(English) × ($146,560 and higher)	–	1.405^{*}	0.019
Language × *Immigration*	(English) × (Non-landed immigrant)	–	1.480^{***}	0.001
Language × *Education*	(French) × (High school or less)	–	1.331^{*}	0.014
Language × *Household type*	(French) × (Single)	–	-1.802^{*}	0.012
Language × *Immigration*	(French) × (Non-landed immigrant)	–	1.254^{*}	0.042
Language × *Education*	(Non-official) × (High school or less)	–	1.144^{*}	0.026
Language × *Immigration*	(Non-official) × (Non-landed immigrant)	–	0.963^{*}	0.044
Language × *Employment*	(French and Non-official) × (Employed)	–	3.790^{*}	0.041
Employment × *Income*	(Employed) × ($146,560 and higher)	–	-0.464^{*}	0.036
Household type × *Income*	(Family w/o children under 18) ×($52,203 and lower)	–	-0.650^{*}	0.016
	(Single) × ($52,203 and lower)	–	-0.659^{*}	0.013

Note: $n = 17,135$. The coefficients shown in this table are found to be significant at the 5% level based on their estimated p-values. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 2.8: Lasso Logistic Regression with Interactions for Email Use Dependent Variable

Variables	Categories	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value
Intercept		1.936	4.597^{**}	0.002
Age	55-64	−0.321	-2.188^{*}	0.035
Language	English	–	-2.799^{*}	0.028
	French	–	-5.688^{*}	0.033
	English, French and Non-official	–	-33.857^{***}	0.001
Location × *Age*	(Rural) × (35-44)	–	0.750^{*}	0.039
	(Rural) × (65 and older)	–	0.745^{**}	0.008
Location × *Language*	(Rural) × (English, French and Non-official)	–	32.644^{*}	0.041
Age × *Immigration*	(25-34) × (Non-landed immigrant)	–	-1.195^{*}	0.035
Age × *Province*	(25-34) × (MB)	–	1.766^{*}	0.023
Age × *Language*	(55-64) × (English)	–	1.945^{*}	0.022
	(55-64) × (French)	–	2.074^{*}	0.026
Age × *Province*	(55-64) × (MB)	–	1.439^{*}	0.022
Age × *Language*	(65 and older) × (English)	–	1.705^{*}	0.048
Age × *Income*	(65 and older) × ($52,203 and lower)	−0.223	-0.697^{*}	0.040
Language × *Income*	(English) × ($146,560 and higher)	–	1.857^{*}	0.022
Language × *Province*	(French) × (MB)	–	6.461^{*}	0.018
Minority × *Province*	(Visible minority) × (MB)	–	-1.825^{**}	0.005

Note: $n = 17,268$. The coefficients shown in this table are found to be significant at the 5% level based on their estimated p-values. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

2.4.3 Multiple correspondence analysis

The dependent and explanatory variables discussed in Sections 2.1 and 2.2 are categorical. Hence, the relationships between these variables can be analyzed using the multiple correspondence analysis (MCA) for associations between categorical variables. Its advantage is that the MCA displays graphically complex dependencies involving the interactions between groups of variables. We consider it as complementary to the logit estimation results, as the MCA results are less rigorous, although easy to interpret. In particular, the MCA allows us to explore further the interaction effects, including both the dependent and explanatory variables.

Figures 2.1 and 2.2 display the variable categories represented in two-dimensional space.

Internet use, email use and online banking. Figure 2.1 presents the associations between the explanatory and dependent variables appearing in Models 1-3 of internet use, email use, and online banking. The green-labelled variable categories are the dependent variables in our logit models and the supplemental variables in the MCA. The red-labelled categories are the explanatory variables in our logit models.

The groupings of variable categories illustrate graphically the underlying structure of the data. The most apparent grouping of variable categories is in the top left quadrant of the graph. This grouping includes individuals who did not use the internet, email or online banking in the last three months. Grouped with these dependent variable categories are the explanatory categories *65 years and older, Not employed, Single, High school or less*, and people who earn less than $\$52,204$ a year. Our logistic regressions identified these explanatory variables as statistically significant.

In the lower right quadrant of the plot, we see another grouping. The dependent variable categories of people who used the internet, email and online baking are in this quadrant grouped relatively close to the variables *University degree*, income of $\$92,485$–$\$146,559$, income greater than $\$146,559$, *Families with children under 18, Employed*, and age group

categories *45-54, 35-44, and 25-34*. In Tables 2.1, 2.2, and 2.3, these variables are all statistically significant and have positive coefficients.

Virtual wallet and credit card use. Figure 2.2 illustrates the associations between the dependent and explanatory variables appearing in Models 4-5 of the virtual wallet and credit card use. In the top right quadrant of the plot, the dependent variable category *Did not use credit card* is grouped with the explanatory variable categories $\$52,203$–$\$92,485$, *Single, High school or less, Not employed*, income less than $\$52,204$, *15-24*, and *65 and older*. In Table 2.5, we see that `svy LLasso` has selected the lowest age group category *15-24* and *High school or less*. The MCA grouping around *No credit card* usage is relatively consistent with the variables selected by `svy LLasso`.

The top left quadrant of the plot has the dependent variable category *Used virtual wallet*. The explanatory variables grouped around *Used virtual wallet* are *Urban, 25-34, ON and AB*. In Table 2.4, the explanatory variables selected by `svy LLasso` are all the age group categories, *Rural*, *Visible minority*, the highest income category, and *University degree*. The grouping around the *Used virtual wallet* is mostly consistent with the variable categories selected by `svy LLasso`.

`svy LLasso` selected the variable *Visible minority*. Although it is not in the close grouping of variables around virtual wallet, it is in the same quadrant of the graph. *Visible minority* is closely grouped with *Landed immigrant*, which is consistent with Figure 2.1.

2.4.4 Digital literacy score

The use the internet and internet-based services is determined not only by the demographic and socio-economic characteristics of an individual, but also by their digital ability, or digital literacy. The digital literacy is an outcome of various socio-economic characteristics. It is unobserved, i.e. latent as there is no CIUS question that provides direct information about the digital literacy of the respondents. We create a measure (score) of digital literacy and

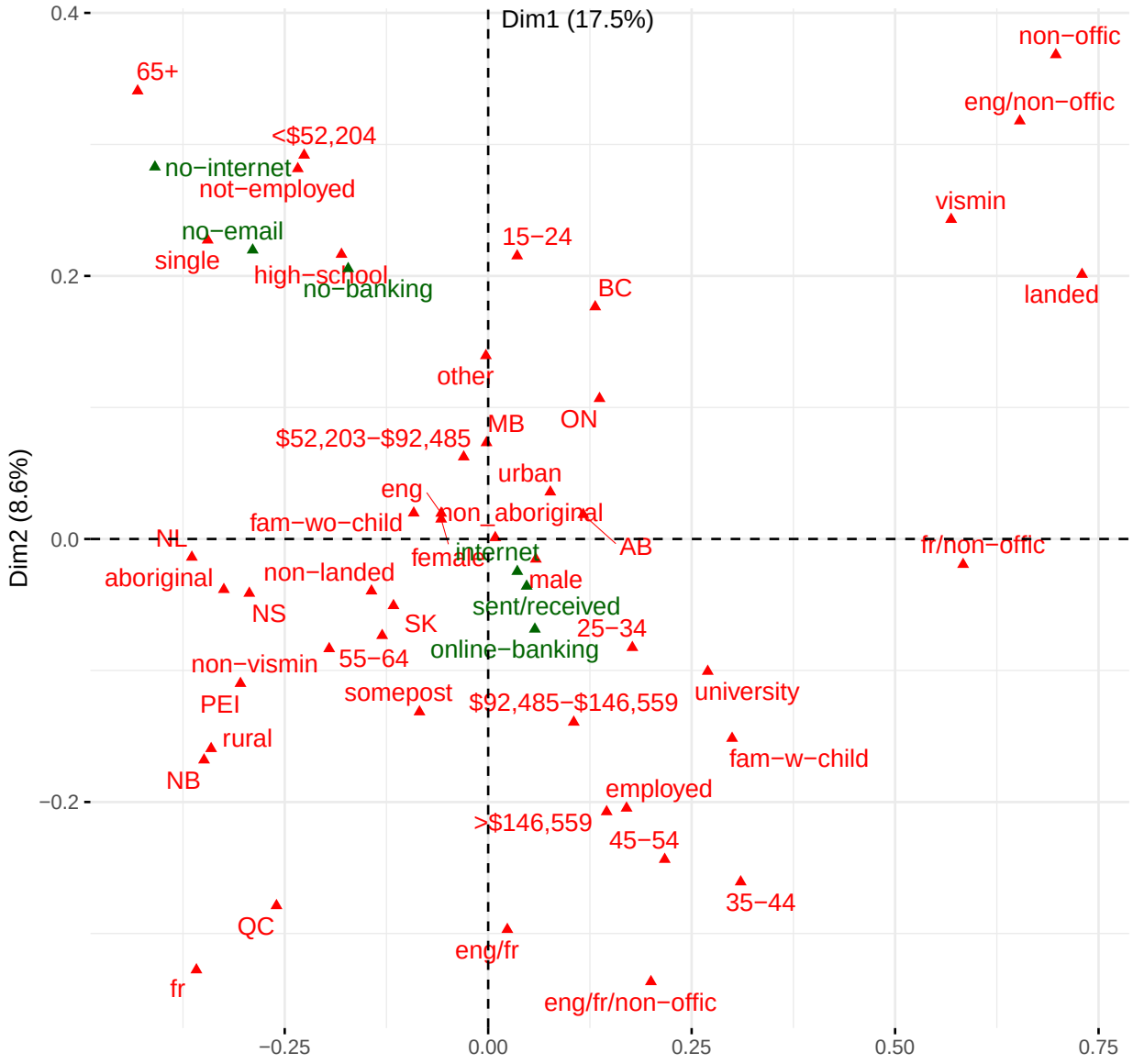


Figure 2.1: Coordinate plot for Internet Use, Email Use and Online Banking

Note: Figure 2.1 shows the multiple correspondence analysis coordinate plot for the dependent variables internet use, email use, and online banking (labelled in green). The explanatory variables in this analysis are labelled in red.

apply it to groups of individuals distinguished in the previous sections to assess the digital divide in a more rigorous way.

We analyze the distributional properties of the score of digital literacy in the entire sample. We also compare its values in the groups of individuals distinguished with respect

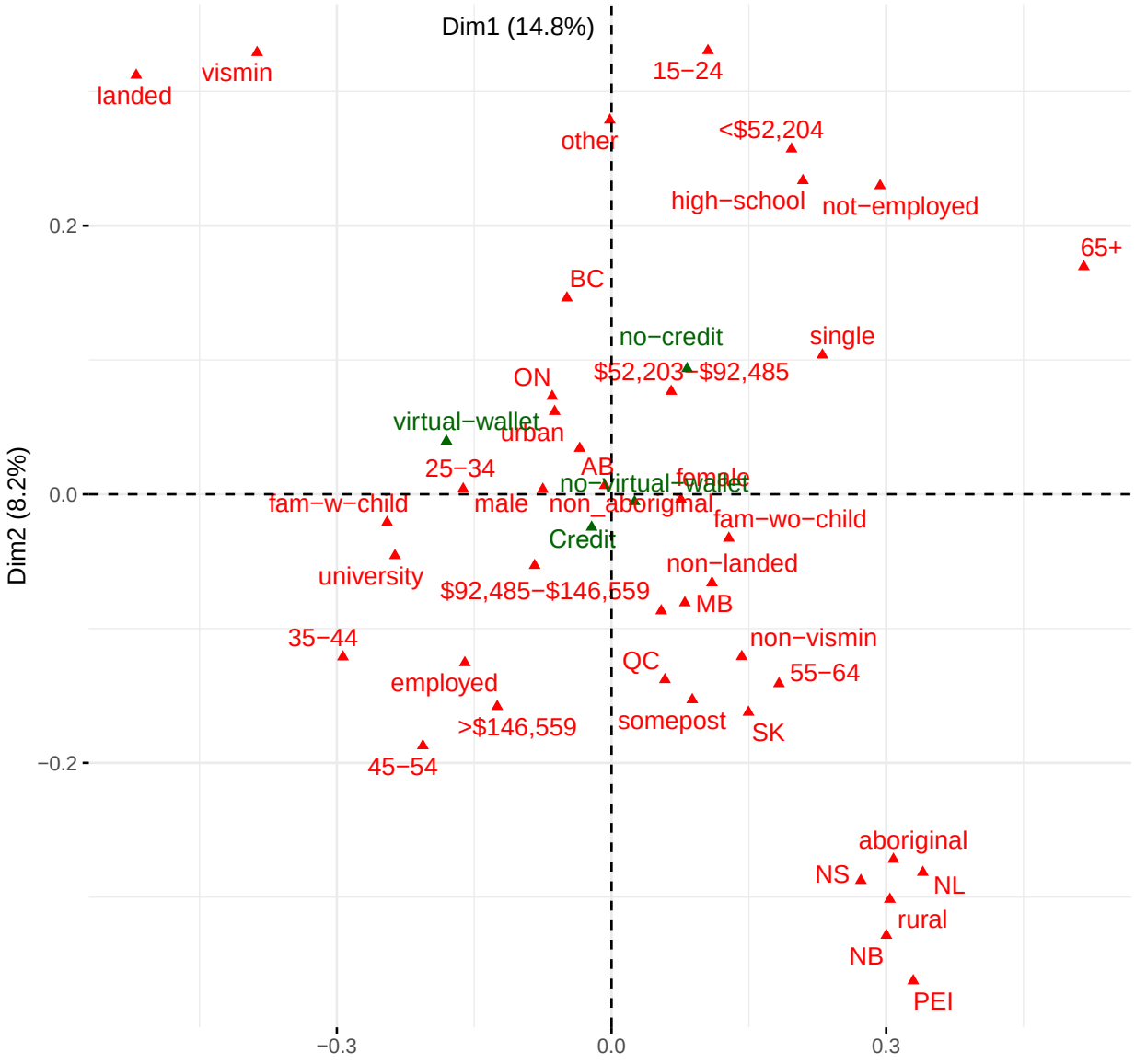


Figure 2.2: Coordinate plot for Virtual Wallet and Credit Card Use

Note: Figure 2.2 shows the multiple correspondence analysis coordinate plot for the dependent variables virtual wallet and credit card (labelled in green). The explanatory variables in this analysis are labelled in red.

to location, age, gender, education level and immigration status. We consider noticeable differences in the value of the digital literacy score as the evidence of digital divide between, or inside the groups.

The digital literacy score comes from survey respondents' answers to 10 questions in

CIUS 2020 (see Appendix A.2 for the list of 10 questions our score comprises). Respondents that answer *Yes* to these questions receive 1 point per *Yes* response [1].

The higher the score (out of 10), the higher the perceived digital literacy of the respondent. Next, we compute the average scores for the aforementioned groups of individuals and display these results in Table 2.9. The average score from all respondents in Table 2.9 is 7.11, and the standard deviation is equal to 2.15. Therefore, respondents answered an average of seven questions with *Yes*.

The first group of individuals we investigate is distinguished with respect to the location. The first two rows of the table show the average score out of 10 for survey respondents residing in urban and rural locations. Urban residents score higher in digital literacy than rural residents. This rural/urban divide is consistent with our `svy LLasso` and MCA results that show a divide between rural and urban residents regarding internet connectivity.

The age group variable shows one of the largest divides regarding digital literacy score. The oldest age group category *65 years and older* has the lowest digital literacy score in our study. The youngest age group also scores relatively lower than the three middle-age categories. Due to the type of questions that make up the digital literacy score, younger respondents may have been less likely to answer *Yes* to these questions. Many of the questions focus on making online purchases and using digital technology, potentially skewing toward people in the middle age groups.

There is no major difference in the scores of males and females. The lack of a digital divide among genders is consistent with our `svy LLasso` results. However, in contrast to gender, we observe a digital literacy gap between *Aboriginal* and *non-Aboriginal* respondents. *Aboriginal* individuals, on average, score lower on the digital literacy scale. This result may reflect broader socioeconomic and geographical challenges faced by *Aboriginal* communities.

Employment status, educational attainment and income all show large discrepancies in

[1] All relevant questions were asked to a subset of 12,431 CIUS 2020 respondents. After removing *Not stated* answers from this subset, we are left with 11,874 observations. We compared our digital literacy score with other samples where *Not stated* answers were replaced by multiple imputation methods following Van Buuren (2018). Results were consistent across models.

the digital literacy scores across their categories. *Employed* individuals scored an average of almost one point higher than *Non-employed*. Individuals with low educational attainment of a *High school or less* score the second lowest only to respondents *65 years and older* on our digital literacy score test. Educational attainment of a *University degree* shows an average of more than a point difference in their digital literacy score compared to those with a *High school or less*.

The lowest income category of individuals making $\$52,203$ *and lower* has the second lowest digital literacy score. The digital literacy score increases as income categories increase, with the highest income category having the highest digital literacy score. These results are very consistent with the Lasso inference results. `svy LLasso` selected employment status, income, and education variables. The debiased Lasso results also showed that employment, income, and education categories are relevant explanatory variables in almost every logit model specification.

Both groups distinguished with respect to immigration status and visible minority status show surprising results. The immigration status variable category *Landed immigrant* has a slightly higher digital literacy score than *Non-landed immigrant* (non-immigrant/non-recent immigrant). The variable *Visible minority* scores higher on our digital literacy score than the category *Non-visible minority*.

From our MCA results, we know that the variable categories *Landed immigrant* and *Visible minority* are grouped together, suggesting that many recent immigrants are also visible minorities. New immigrants to Canada often have to use the internet and online resources when applying to immigrate to Canada and become citizens. These requirements could explain why visible minorities and recent immigrants have slightly higher digital literacy scores than non-visible minorities and non-immigrants.

The digital literacy scores for provinces vary. The Maritime provinces—Newfoundland and Labrador (NL), Prince Edward Island (PEI), Nova Scotia (NS), and New Brunswick (NB)—along with Manitoba (MB) and Saskatchewan (SK), score the lowest. In contrast,

British Columbia (BC), Ontario (ON), and Alberta (AB) have the highest digital literacy scores, with these provinces showing almost identical results. The MCA results from each plot consistently group the Maritime provinces with the rural category, which explains their lower scores on the digital literacy scale.

Individuals who have used a virtual wallet score the highest on our digital literacy score, with an average score of 8.31. Canadians currently using virtual wallets have very high digital literacy, much higher than the average Canadian.

2.5 Additional analyses and robustness checks

This section presents additional analyses and robustness checks for the digital divide. In Section 2.5.1, we investigate the effects of COVID-19 on the digital divide and also consider the influence of provincial safety (stringency) measures on technology adoption during the period covered in CIUS. In Section 2.5.2, we compare internet use and its determinants from CIUS 2010 to CIUS 2020 to study the evolution of the digital divide over the past decade.

2.5.1 Impact of COVID-19 on the digital divide

The onset of the COVID-19 pandemic has reshaped financial behaviors, notably transitioning from traditional to digital transaction and communication methods. We observed a decline in cash usage at the pandemic's onset, followed by an increased adoption of digital payments (Chen, Engert, Huynh and O'Habib, 2021; Chen, Engert, Huynh, O'Habib and Zhu, 2021). This shift is evident in the rise of mobile payment usage from 11% in November 2020 to 17% by April 2021. The analysis is based on questions posed to CIUS respondents regarding changes in their usage of various digital technologies during the COVID-19 pandemic.

To explore behavioral changes in the use of digital technology, we employ k-means clustering on CIUS survey questions related to online activities during the pandemic. Using the elbow method and silhouette scores, we determine the optimal number of clusters. Profiling

Table 2.9: Digital Literacy Score

Variables	Categories	Digital Literacy Score
Location	Urban	7.18
	Rural	6.68
Age	15–24	7.12
	25–34	7.71
	35–44	7.61
	45–54	7.18
	55–64	6.66
	65 and older	6.06
Gender	Male	7.12
	Female	7.10
Aboriginal identity	Non-Aboriginal	7.12
	Aboriginal	6.86
Employment status	Employed	7.38
	Not employed	6.61
Education	High school or less	6.48
	Some post-secondary	6.99
	University degree	7.71
Visible minority status	Visible minority	7.33
	Not a visible minority	7.03
Household type	Family with children under 18	7.41
	Single	6.72
	Family w/o children under 18	6.98
	Other household type	7.15
Income	$52,203 and lower	6.46
	$52,204–$92,485	6.84
	$92,486–$146,559	7.23
	$146,560 and higher	7.82
Immigration status	Landed immigrant	7.28
	Non-landed immigrant	7.07
Province	NL	6.84
	PEI	6.88
	NS	6.81
	NB	6.85
	QC	7.03
	ON	7.15
	MB	6.99
	SK	6.94
	BC	7.20
	AB	7.21
Virtual wallet	Used virtual wallet	8.31
	No virtual wallet	6.94

Note: This table presents the average Digital Literacy Scores derived from 10 questions in CIUS 2020 data. Scores range from 0 to 10, with higher values indicating greater digital literacy. The table categorizes respondents based on various demographic and socioeconomic factors.

these clusters and comparing them with demographics, we derive centroids for each cluster to better understand their characteristics. The k-means cluster analysis identifies two distinct clusters, which we label as *Digital Adopters* and *Digital Resisters*.

In the digital literacy score distributions for the identified clusters, we observe the *Digital Adopters* cohort has a median digital literacy score of 8.0, with an interquartile range from 7.0 to 9.0, indicating a consistent, higher proficiency in digital literacy within this group. Conversely, the *Digital Resisters* cohort displays a more dispersed distribution, with a median score of 6.0 and an interquartile range spanning from 4.0 to 7.0. The heterogeneity in this cluster indicates a broader spectrum of digital engagement behaviors; see Figure 2.3 for a graphical representation of these distributions.

Figure 2.5 depicts the demographic composition of the *Digital Adopters* cluster. The demographics with the highest representation in the *Digital Adopters* cluster include those who are employed, have a university education, are part of a family with children, and have an income of $162,800 or more. In contrast, the demographics with the lowest representation in the cluster are the unemployed, individuals with high school education, singles, and those with an income of $44,119 or less.

We compare the percentages of individuals in each cluster engaging in online banking, email, online credit card usage, and virtual wallet usage. The *Digital Adopters* demonstrate a higher inclination to use these digital tools compared to the *Digital Resisters*. A grouped bar chart in Figure 2.4 visually represents this distinction, clearly emphasizing the differences in digital engagement between the clusters. *Digital Adopters*, on average, have a much higher digital literacy score.

Incorporation of the stringency dataset

To analyze the impact of governmental responses to COVID-19 on digital adoption, we use the stringency dataset developed by Cheung et al. (2021). This dataset, adapted from methodology developed by Oxford University's Blavatnik School of Government for the Ox-

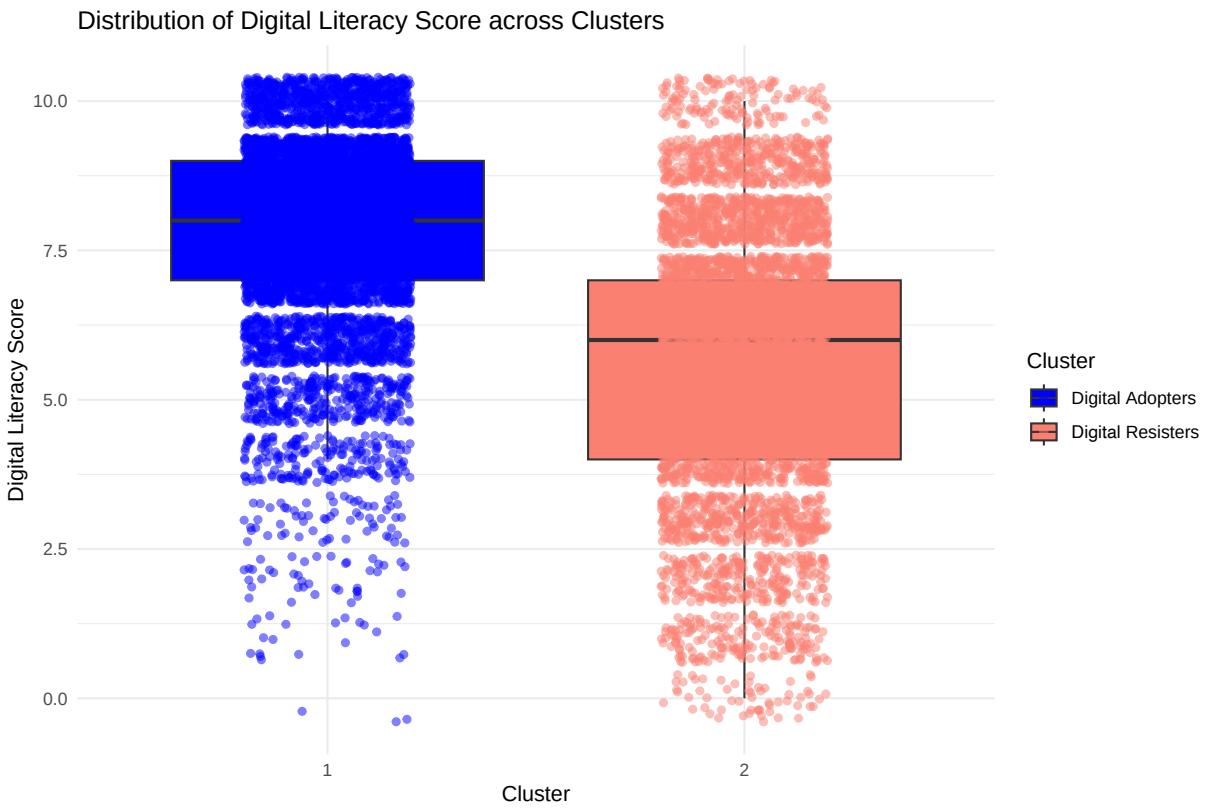


Figure 2.3: Digital Literacy Score By Cluster

Note: Figure 2.3 shows a boxplot of the *Digital Adopters* and *Digital Resisters* digital literacy scores. The black bar inside the boxplot shows the average score for each group.

ford COVID-19 Government Response Tracker, measures the stringency of containment restrictions across Canadian provinces. We use the timeline from January 1, 2020, until the end of CIUS data collection on March 3, 2021. Figure 2.7 shows the breakdown of the average stringency for each province over the specified time frame.

We compare the distribution of *Digital Adopters* across provinces with either above or below-average stringency restrictions, utilizing the violin plot illustrated in Figure 2.8. Provinces with above-average stringency measures demonstrate a concentrated prevalence of *Digital Adopters*, suggesting a potential correlation between increased stringency and higher digital adoption. Conversely, provinces with below-average stringency measures exhibit a broader and lower prevalence of Digital Adopters, indicating varied digital adoption rates under less stringent conditions.

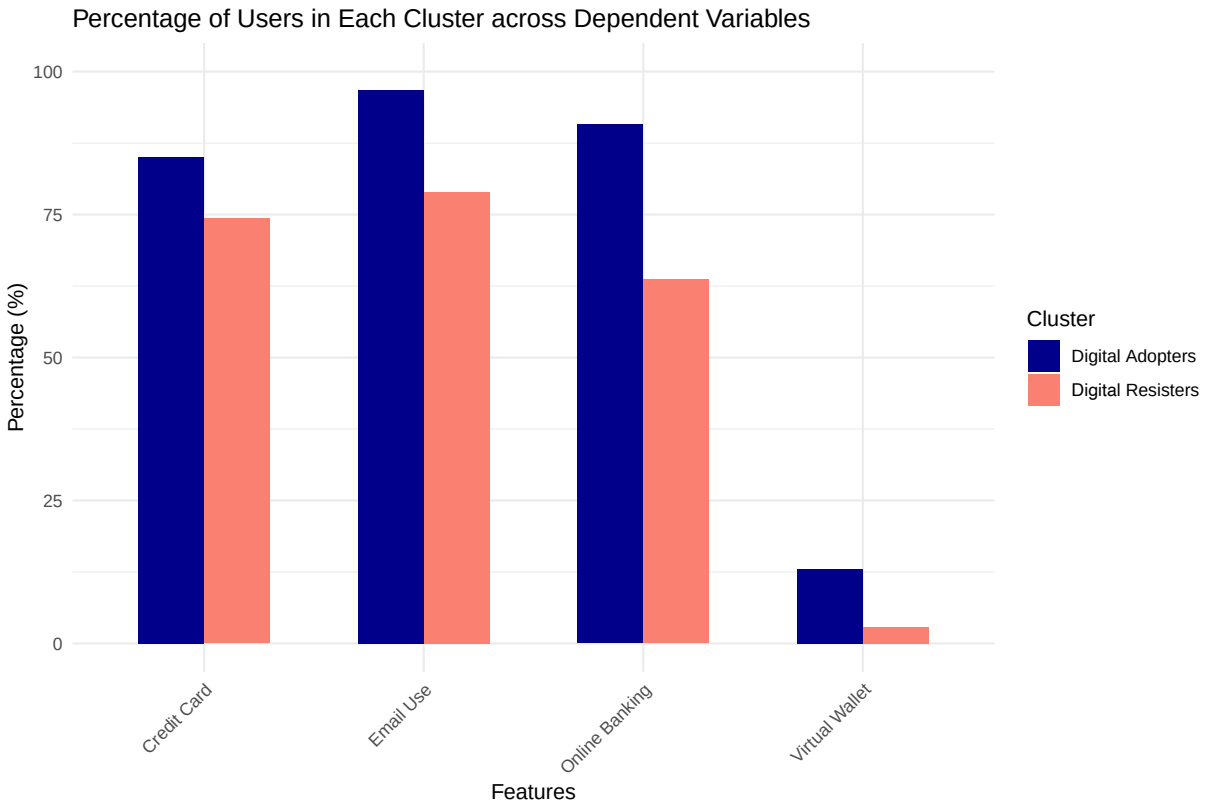


Figure 2.4: Percentage of Respondents in Each Cluster Using Credit Card Online, Email, Online Banking, and Virtual Wallets

To quantify this relationship, we calculate the (Pearson's) correlation between average stringency and the percentage of individuals in the *Digital Adopters* cluster. The test yields a correlation coefficient 0.327 with a p-value 0.357, suggesting that the correlation could be attributed to chance variation rather than a true underlying relationship.

2.5.2 Comparison with CIUS 2010

As the use of digital technologies in Canada increases, it is crucial to compare our findings with prior data and research. In this section, we trace the change in the digital divide by comparing internet use in CIUS 2010 and CIUS 2020 data. In particular, we shed light on whether the digital divide has grown or narrowed over time. This comparison allows us to identify persisting gaps as well as areas where progress has been made. Understanding these

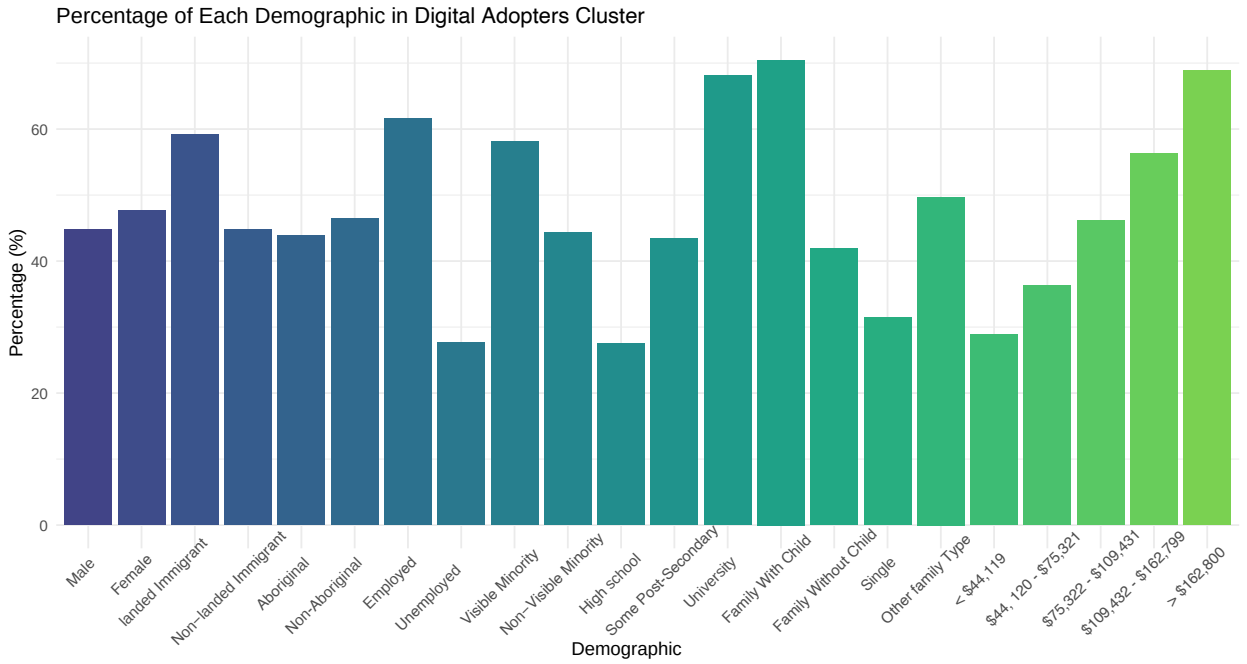


Figure 2.5: Demographics of Digital Adopters Cluster

Note: The bar graph in this figure shows the percentage of the *Digital Adopters* in each specified demographic group.

trends is vital for policymakers and stakeholders in crafting strategies and interventions to bridge the digital divide effectively.

Haight et al. (2014) conduct a study using the internet use variable from the CIUS 2010 data.[2] According to Haight et al. (2014), 80% of Canadians aged 16 and above were connected to the internet in 2010, which represents a notable increase compared to previous years, with figures of 73% in 2007, 68% in 2005, 64% in 2003, and a mere 51% in 2000.

After accounting for the survey weights, the proportion of respondents who were connected to and use the internet (in the past three months) is 92.2% in the online CIUS 2020 data, a notable 11.9% increase over the 10-year span.[3]

To examine this pattern further, we estimate a survey-weighted logit model for the in-

[2]We limit our comparison with Haight to internet use, as not all variables used by Haight are available in the online CIUS 2010 dataset.

[3]The online CIUS 2020 and CIUS 2010 data are available at:
https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/NUVBX2
https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/YUIPZ7

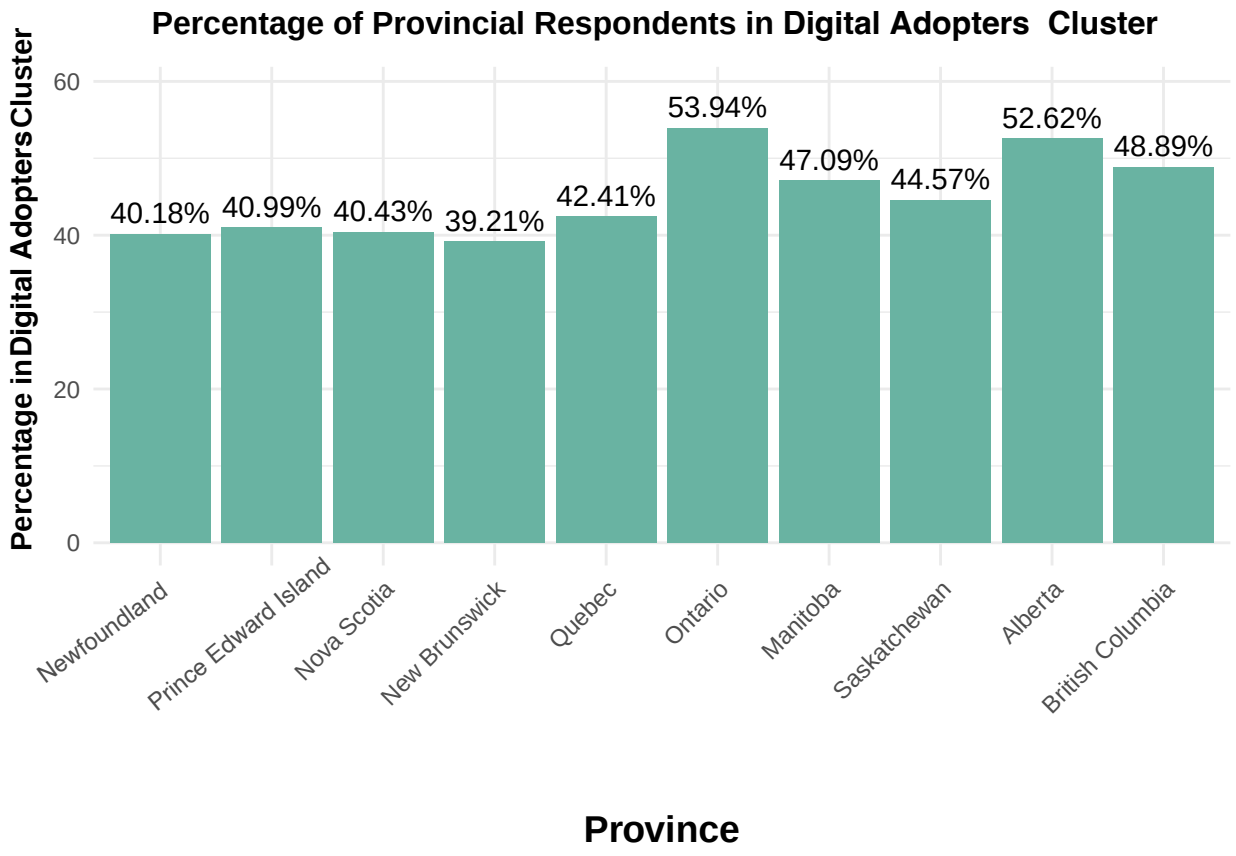


Figure 2.6: Percentage of Digital Adopters from each Province

Note: Figure 2.6 shows a bar graph of the percentage of each province in the *Digital Adopters* cluster.

ternet use dependent variable, closely mimicking the specification of Haight et al. (2014) using the online version of CIUS 2010. We then estimate the same model using comparable variables in CIUS 2020 data. Since the latter dataset does not include immigration status or high school graduation information used by Haight et al. (2014), we estimate the model without these variables. Similarly to Haight et al. (2014), we use the age variable without grouping it into different age cohorts. However, note that the quintiles of the income variable slightly differ between CIUS 2010 and CIUS 2020, and the internet use variable in CIUS 2010 is an indicator of whether respondents have used the internet in the past 12 months, as opposed to the 3 months in CIUS 2020.

Table 2.10 reports the survey-logit estimation results. The coefficient estimates for the internet use variable in CIUS 2010 are qualitatively similar to those reported in Haight

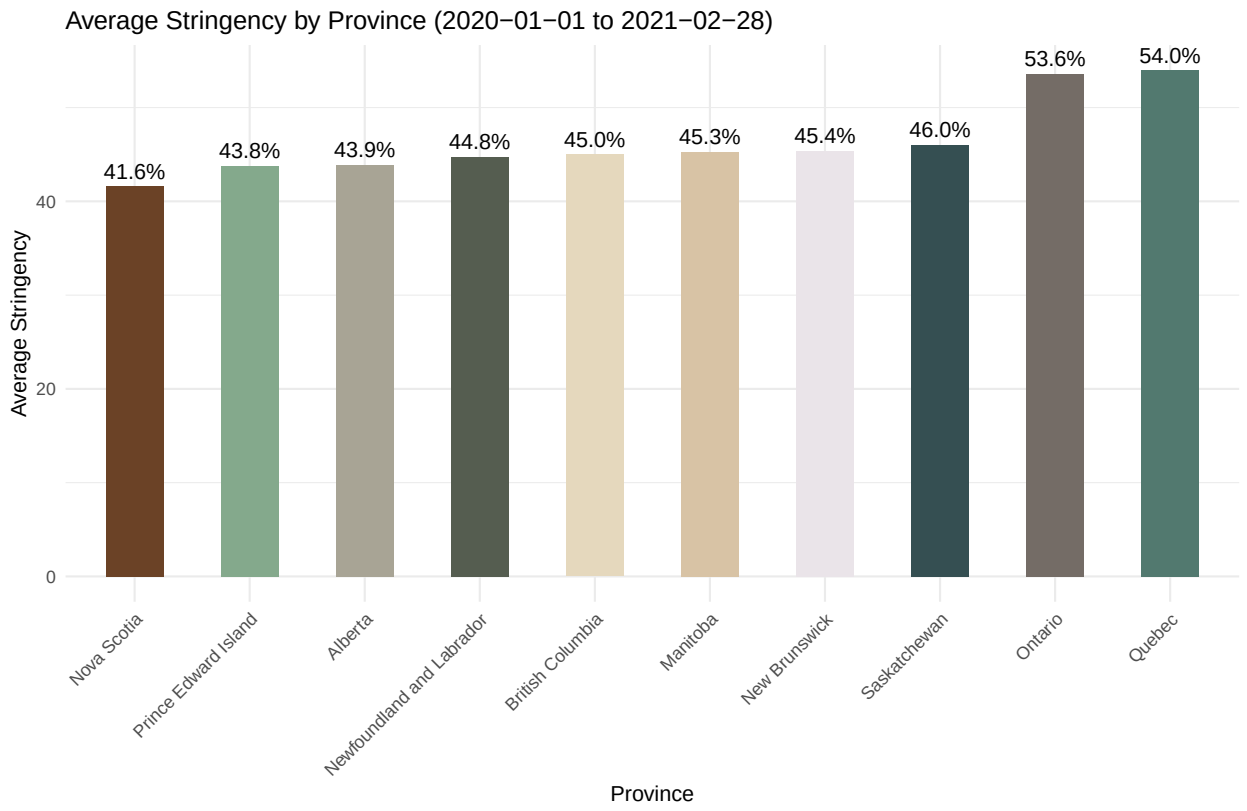


Figure 2.7: COVID-19 Stringency Index by Province

Note: The bar graph in this figure displays the average COVID-19 stringency restrictions for each province from January 1, 2020, to March 3, 2021. It is based on the stringency dataset developed by Cheung et al. (2021), which follows the methodology of Oxford University's Blavatnik School of Government, as used in the Oxford COVID-19 Government Response Tracker.

et al. (2014). The AME estimates suggest that a digital divide in Canada has narrowed across several crucial demographic dimensions. The effect of income across the quintiles 2–5 (Income 1–5 denote the dummy variables for the income brackets) relative to the income quintile 1 has decreased. The gap between individuals with a high school degree or less, or a university degree relative to those with some post-secondary education appears to have narrowed. Also, the effect of whether a respondent is currently a student or not (Student is the corresponding dummy) is absent in CIUS 2020, in contrast with the highly significant AME estimate of 0.127 in CIUS 2010 data.

Regarding rural vs. urban dummies, the coefficient estimates point toward a persistent gap. Interestingly, the negative effect of age on internet use seems to have decreased but

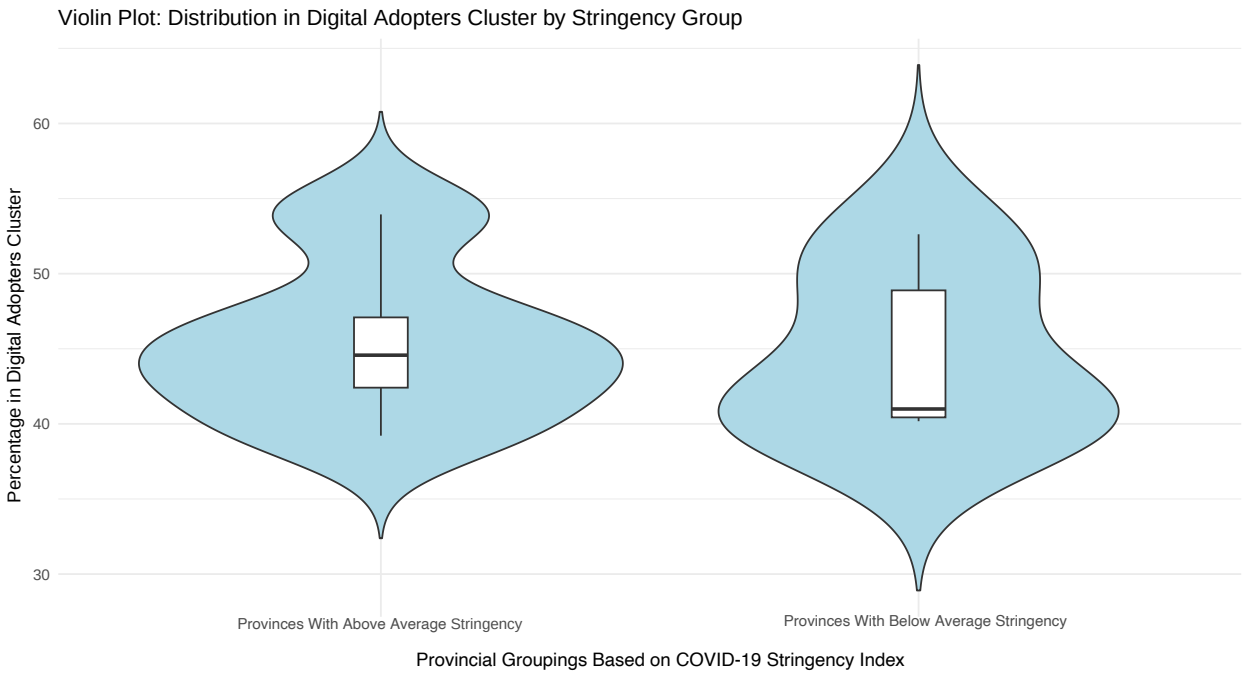


Figure 2.8: COVID-19 Stringency Index and Percentage of Observations in the Digital Adopters Cluster

Note: This figure shows a violin plot of two different distributions. The provinces have been divided into two groups: one with above-average provincial COVID-19 stringency restrictions and the other with below-average. The plot on the left shows the distribution of *Digital Adopters* in the above-average group, and the one on the right shows the distribution in the below-average group.

still persists, and gender has no effect on internet use in both datasets.

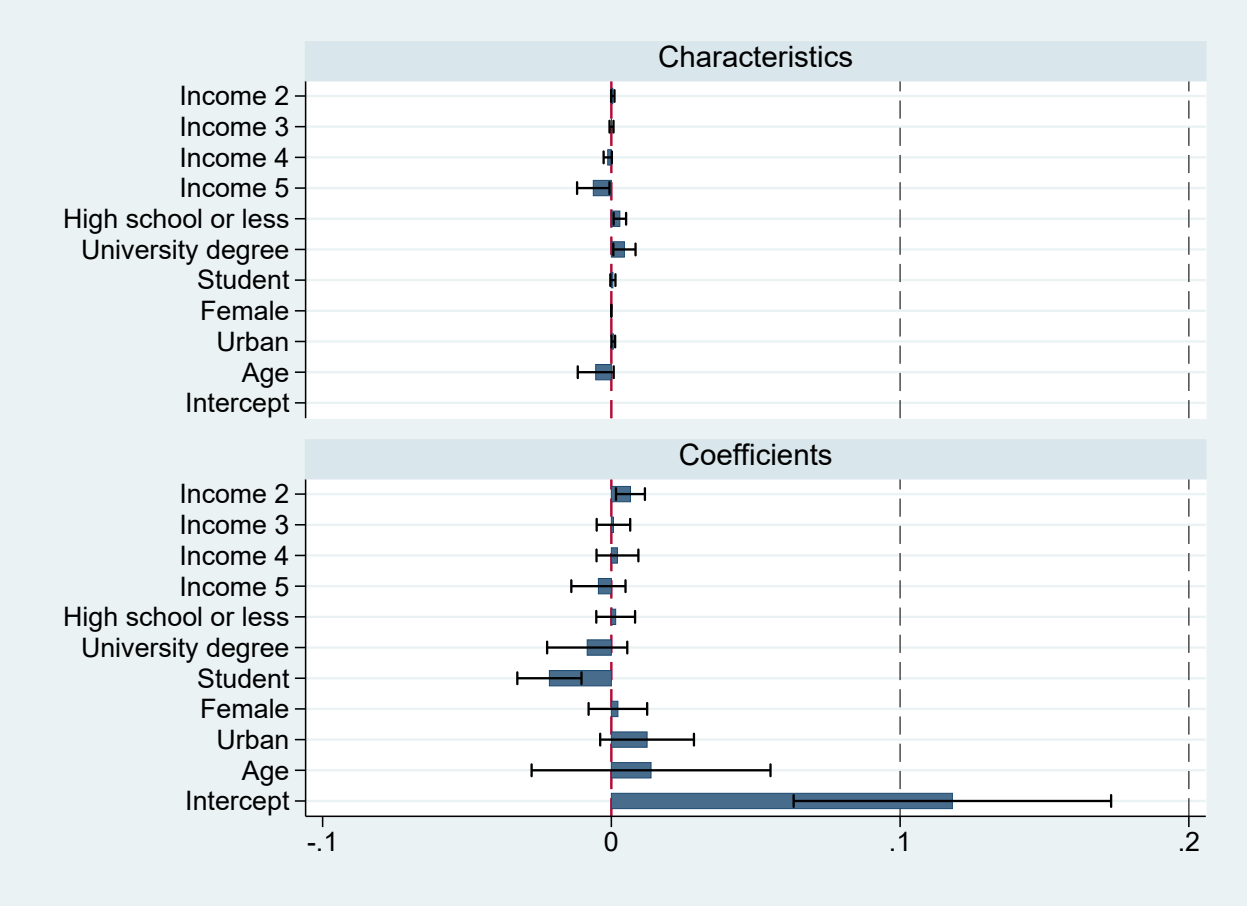
To explain the 11.9% internet use differential between CIUS 2010 and CIUS 2020 data, we perform the Oaxaca-Blinder decomposition, which decomposes the gap in the internet use rates in CIUS 2010 and CIUS 2020 data into portions due to differences in coefficients and characteristics (regressors).

Table 2.11 and Figure 2.9 display the twofold Oaxaca-Blinder decomposition, where CIUS 2020 model coefficients are used as the reference coefficients. The decomposition results suggest that we can attribute approximately -0.4% of the 11.9% difference to group differences in characteristics (i.e., age, education, gender), and the remaining 12.3% to differences in coefficients. This result is expected since the regressors in both models are comparable, and the coefficients exhibit some variations, as reported in Table 2.10.

The variable-by-variable twofold decomposition reveals that two key factors, namely the

percentage of individuals in the second income bracket and the survey respondents who are currently enrolled as students, mainly drive the variation in internet usage. Not surprisingly, the change in the intercept exceeds zero by a large margin and is highly significant, as many dummy variables, e.g., employment status, family type, province, and age cohort, are omitted in the survey-logit specifications.

Figure 2.9: Regressor-by-Regressor Variation in the Oaxaca-Blinder Decomposition



Note: This figure plots the regressor-by-regressor variations in the Oaxaca-Blinder decomposition of the difference in the estimated conditional probabilities of internet use in CIUS 2010 and CIUS 2020 data based on survey-weighted logit model. The CIUS 2020 model coefficients are used as the reference coefficients. Income 2–5 are the dummy variables for the income quintiles.

Table 2.10: Survey-Logit Estimates for Internet Use Dependent Variable

		Coefficient estimates	
Variables	Categories	CIUS 2010	CIUS 2020
Intercept		3.620***(0.156)	4.764***(0.277)
Location	Rural (omitted)	–	–
	Urban	0.188***(0.059)	0.330***(0.075)
Age	Age	−0.695***(0.030)	−0.659***(0.045)
Gender	Female	0.035 (0.066)	0.078 (0.074)
	Male (omitted)	–	–
Education	High school or less	−1.040***(0.065)	−0.996***(0.077)
	Some post-secondary (omitted)	–	–
	University degree	0.770***(0.144)	0.509***(0.156)
	Student	1.222***(0.261)	−0.362 (0.285)
Income	Income 1 (omitted)	–	–
	Income 2	0.397***(0.084)	0.717***(0.086)
	Income 3	0.904***(0.090)	0.940***(0.112)
	Income 4	1.207***(0.106)	1.309***(0.145)
	Income 5	1.888***(0.136)	1.671***(0.184)
	Observations	22,623	17,409
		AME estimates	
Variables	Categories	CIUS 2010	CIUS 2020
Intercept		–	–
Location	Rural (omitted)	–	–
	Urban	0.020** (0.006)	0.019***(0.004)
Age	Age	−0.072***(0.002)	−0.039***(0.002)
Gender	Female	0.004 (0.007)	0.005 (0.004)
	Male (omitted)	–	–
Education	High school or less	0.197***(0.014)	0.098***(0.011)
	Some post-secondary (omitted)	–	–
	Student	0.127***(0.028)	−0.021 (0.017)
	University degree	0.080***(0.014)	0.030***(0.009)
Income	Income 1 (omitted)	–	–
	Income 2	−0.108***(0.007)	−0.059***(0.005)
	Income 3	0.041***(0.009)	0.042***(0.005)
	Income 4	0.094***(0.009)	0.055***(0.007)
	Income 5	0.126***(0.010)	0.077***(0.009)
	Observations	22,623	17,409

Note: This table reports the survey-weighted logit estimates for comparable variables between CIUS 2010 and CIUS 2020. The CIUS 2020 model coefficients are used as the reference coefficients. The top panel reports the coefficient estimates while the bottom panel reports the corresponding AME estimates. The standard errors are in parantheses. Income 1–5 are the dummy variables for the income quintiles. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

2.6 Conclusion

This paper examines the CIUS 2020 data using the `svy LLasso` estimator of logit models with internet use, online banking, email, virtual wallet, and online credit card usage as the dependent variables. We find large disparities in digital literacy across different demographic groups within Canada. Younger Canadians, particularly those aged 15 to 24, are choosing to use emerging financial technologies like virtual wallets, moving away from traditional methods such as online banking and credit cards. Individuals with high incomes, high educational attainment, and stable employment are the most engaged with digital technologies. They consistently utilize a broader array of digital services and are more likely to adopt new innovations. Our findings on women, visible minorities, and immigrants were contrary to previous findings. Women, especially those in the workforce, are more likely to use email and exhibit higher digital literacy. Visible minorities and recent immigrants are also showing strong digital engagement, with new immigrants often matching or even exceeding Canadian-born citizens.

Regions such as Manitoba and the Maritime provinces face significant digital barriers. These areas are home to populations with lower digital literacy and engagement, particularly among seniors and lower-income individuals. Prioritizing these provinces in national strategies is crucial to ensuring the benefits of a digital economy are shared equitably across all Canadians.

Our comparison with CIUS 2010 data shows the evolution of the digital divide over the past decade, with increased overall digital connectivity but a persistent urban rural divide. Older Canadians, especially seniors, exhibit lower engagement with digital technologies, reflecting a digital divide along age lines. This divide is most acute among individuals who are older and single or older with low income, emphasizing the intersectionality of age and income in digital exclusion.

The COVID-19 pandemic highlighted the essential role of digital access and literacy. While we observed a positive correlation between stricter provincial public health measures

and digital adoption, it was not statistically significant suggesting these public health policies did not alter the already established patterns of the digital divide. Even though the use of digital technologies increased during the pandemic, pre-existing disparities continued to shape digital engagement, despite the varying strictness of restrictions across provinces.

Given the current state of the digital divide, the potential introduction of a CBDC could disproportionately disadvantage individuals from lower socioeconomic classes who may struggle to adapt to new digital monetary systems. There is a need for targeted investments in digital literacy and infrastructure—not only in rural areas but also in lower-income urban communities. Providing internet access alone is insufficient; comprehensive strategies that include education and support are essential to equip all Canadians to participate fully in an increasingly digital economy. As Canada moves toward a potential cashless society, these efforts are crucial to prevent the widening of the digital divide.

Limitations and areas for future research. The CIUS 2020 data used in our analysis covers only the ten Canadian provinces, excluding the territories and Aboriginal reserves. This omission may underrepresent the true extent of the digital divide, particularly in remote and rural areas such as on reserves where unique socioeconomic and geographical challenges affect digital connectivity. Additionally, although the proportion of unbanked individuals in Canada is exceedingly small—nearly 99% of respondents in the 2017 Methods-of-Payment Survey reported having a bank account—the unbanked population predominantly resides in the Northern territories and reserves, areas not captured in our data. While CIUS includes participants who self-identify as Aboriginal, it does not gather information directly from First Nations reserves. This could skew perceptions of digital inclusivity among Indigenous populations. Comparisons with the 2017 Aboriginal People's Survey show similar internet usage rates among Indigenous respondents, but it is essential to approach these findings with an understanding of the unique challenges faced by these communities.

Table 2.11: Twofold Oaxaca-Blinder Decomposition of Internet Use Difference in CIUS 2010 and CIUS 2020

Variables	Categories	Survey-logit	p-value
	CIUS 2020	0.922***	0.000
	CIUS 2010	0.803***	0.000
Decomposition	Difference	0.119***	0.000
	Characteristics	−0.004	0.270
	Coefficients	0.123***	0.000
	Characteristics		
Location	Rural (omitted)	–	–
	Urban	0.001*	0.040
Age	Age	−0.005	0.088
Gender	Male (omitted)	–	–
	Female	0.000	0.778
Education	High school or less	0.003**	0.008
	Student	0.001	0.235
	Some post-secondary (omitted)	–	–
	University degree	0.005*	0.021
Income	Income 1 (omitted)	–	–
	Income 2	0.001	0.089
	Income 3	0.000	0.883
	Income 4	−0.001	0.085
	Income 5	−0.006*	0.029
	Coefficients		
Location	Rural (omitted)	–	–
	Urban	0.012	0.135
Age	Age	0.014	0.515
Gender	Male (omitted)	–	–
	Female	0.002	0.661
Education	High school or less	0.002	0.662
	Student	−0.022***	0.000
	Some post-secondary (omitted)	–	–
	University degree	−0.008	0.238
Income	Income 1 (omitted)	–	–
	Income 2	0.007**	0.010
	Income 3	0.001	0.807
	Income 4	0.002	0.569
	Income 5	−0.004	0.349
	Intercept	0.118***	0.000
	Observations	40,032	

Note: This table reports the regressor-by-regressor variations in the Oaxaca-Blinder decomposition of the difference in the internet use rates in CIUS 2010 and CIUS 2020 data based on survey-weighted logit model. Income 2–5 are the dummy variables for the income quintiles. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Chapter 3

Digital Adoption and Cyber Security: An Analysis of Canadian Businesses

Coauthorship Statement

This chapter is based on joint work titled “Digital Adoption and Cyber Security: An Analysis of Canadian Businesses" with Dr. Joann Jasiak and Dr. Purevdorj Tuvaandorj. All authors contributed equally to the conceptualization, data preparation, empirical modeling, analysis, and writing of the manuscript.

This paper has not yet been submitted for publication.

3.1 Introduction

The digital transformation of business operations is reshaping the Canadian economy, offering opportunities for increased productivity and innovation. For businesses, adopting digital technologies is no longer optional; it is essential for maintaining competitiveness in a globalized market. However, with greater reliance on digital tools comes an increased exposure to cyber security risks. Understanding this *trade-off*—the efficiency gains from digital adoption versus the vulnerabilities it introduces is critical for Canadian businesses.

This study examines the extent of digital technology adoption and cyber security practices among Canadian firms, using a novel dataset that combines information from the 2021 Canadian Survey of Digital Technology and Internet Use (SDTIU) and the 2021 Canadian Survey of Cyber Security and Cybercrime (CSCSC). To our knowledge, this is the first study to analyze the most recent version of the CSCSC dataset, offering an opportunity to explore both digital adoption and cyber security. These surveys provide detailed data on firms' adoption of advanced technologies such as cloud computing, artificial intelligence, and enterprise management systems, as well as the frequency, severity, and impact of cyber security incidents on business operations.

The primary objective of this paper is to explore the trade-off between digital adoption and cyber security risk, by examining how technological advancement can enhance productivity while simultaneously increasing firms' vulnerability to cyber threats. Specifically, we analyze which types of firms are adopting digital technologies, how adoption levels vary across industries, and whether these technologies are increasing cyber security incidents. We also investigate the firm-specific factors such as size, industry, and employee characteristics that shape both technological efficiency and the ability to manage cyber security risks. Our methodological contributions include introducing the Business Digital Usage Score (BDUS) to quantify digital adoption, enabling a comparative analysis of firms' engagement with digital technologies.

Given the large number of firms and the high dimensionality of mostly qualitative explanatory variables in our dataset, we introduce high-dimensional logit models estimated via a Lasso-penalized maximum likelihood estimator with survey weights to ensure representativeness of Canadian firms. We use a debiasing method for inference on the selected model's coefficients and establish its asymptotic validity. To assess how closely firms operate to their technological usage frontier, we apply stochastic frontier analysis. Additionally, we employ k-means clustering to categorize firms by technological efficiency, facilitating the identification of distinct profiles of digital adoption and efficiency.

The literature on digital technology adoption emphasizes its role in improving firm productivity. Larger firms are often better positioned to implement these technologies, benefiting from economies of scale and greater access to resources (Leung et al., 2008). In contrast, smaller firms frequently encounter barriers, including high costs of implementation and limited technical expertise, which can hinder their ability to fully realize the potential benefits of digital adoption. Ferrari (2012) demonstrates that industry-specific differences in digital readiness significantly affect adoption rates, while Aghimien et al. (2021) highlight the role of regional policies in either enabling or constricting firms' technological advancement. Together, these studies emphasize the structural factors shaping disparities in digital technology adoption across firms.

In Canada, Bilodeau et al. (2019), using data from the 2017 CSCSC, show the widespread reliance of Canadian businesses on digital technologies. They report that about 92% of Canadian businesses used one or more digital technologies or services in 2017, with significant increases in the adoption of websites and social media integration since 2013. In addition, just over one-fifth of Canadian businesses reported being impacted by cyber security incidents that affected their operations, with 54% noting that these incidents prevented employees from carrying out day-to-day work and about 30% experiencing additional repair or recovery costs.

While digital adoption can boost productivity, it also introduces new risks, particularly related to cyber security. The Geneva Association, a leading international think tank of the insurance industry, defines cyber risks as breaches in confidentiality, availability, and data integrity, posing operational threats to firms that increasingly rely on interconnected digital systems (Eling et al., 2016). Cebula and Young (2010) expand on this definition, framing cyber risks as disruptions that extend beyond information technology (IT) systems to affect broader business stability. In Canada, the Toronto Public Library system experienced significant disruptions following a cyberattack, while the Nova Scotia Health Department faced operational delays and data breaches during a similar event (Bridge and Zoledziowski,

2024; Bousquet, 2023).

The adoption of advanced technologies such as Internet of Things (IoT) and enterprise management systems could increase cyber security risks. Blichfeldt and Faullant (2021) highlight that while these systems can improve productivity, their complexity often introduces integration challenges that, if poorly managed, can undermine business operations. Additionally, the widespread use of interconnected devices has expanded the potential attack surface for cybercriminals, requiring firms to invest more heavily in cyber security infrastructure.

Remote work adoption during the COVID-19 pandemic accelerated the reliance on digital tools but also introduced new risks and challenges. Hackney et al. (2022) find that firms effectively utilizing digital technologies during the pandemic demonstrated resilience in maintaining operations under lockdown conditions. However, Aczel et al. (2021) and Kitagawa et al. (2021) report that the rapid shift to remote work led to heightened employee burnout and increased risks of phishing attacks, underscoring the broader implications of accelerated digital adoption.

Cyber insurance is a tool for mitigating cyber security risks. According to the OECD (OECD, 2017), the cyber insurance market doubled in size between 2015 and 2020, fueled by firms' increasing awareness of cyber threats. However, Fitch Ratings (Fitch Ratings, 2021) highlights that high premiums and restrictive coverage policies hinder adoption, especially among smaller firms. Globally, cyber security spending is projected to surpass $170 billion by 2026 (Gartner, Inc., 2021).

Despite the growing demand for cyber insurance, firms may strategically underinvest in cyber security measures due to the unobservable nature of such investments. Ahnert et al. (2022) argue that firms may prioritize visible innovations over less transparent risk mitigation strategies, as clients are often unable to directly evaluate cyber security expenditures. This dynamic creates a paradox in which firms recognize the increasing risks of digital adoption but fail to allocate sufficient resources to mitigate them effectively.

The paper is organized as follows. Section 3.2 describes the datasets used in the study,

the SDTIU and the CSCSC, as well as the construction of the associated scores and variables. Section 3.3 outlines the paper's main contribution and methodological framework: the survey-weighted debiased logit Lasso method. Section 3.4 presents the empirical results, discussing the key determinants of technological efficiency and cyber security vulnerabilities. We conclude in Section 3.5 with a discussion of the implications of our findings for Canadian businesses and policymakers. The appendix is divided into two parts: Appendix B.1 provides additional technical details on the survey-weighted debiased logit Lasso model, and Appendix B.2 outlines the questions that comprise the BDUS and the Cyber Security Incidence indicator.

3.2 Data Description and Variable Construction

The empirical analysis is based on a merged dataset from two Statistics Canada surveys: the 2021 SDTIU and the 2021 CSCSC. These surveys were merged using firm size and industry classifications from the North American Industry Classification System (NAICS), enabling an integrated study of digital adoption and cyber security risks among Canadian businesses.[1]

The SDTIU focuses on digital technology adoption, including metrics such as internet usage, e-commerce participation, and ICT adoption, while the CSCSC examines cyber security practices and the impact of cyber incidents on businesses. The merged dataset includes both qualitative variables, such as the implementation of specific technologies and cyber security measures, and quantitative variables, such as cyber security expenditure and incident-related costs. Together, the surveys cover numerous variables relevant to digital adoption and cyber security. The firms vary in size, ranging from small enterprises with fewer than 10 employees to large corporations with over 500 employees. The industries covered include manufacturing, retail, professional services, and information technology.

[1]Methodological differences between the surveys include variations in enterprise size definitions (for small firms: SDTIU defines small firms as those with 5–49 employees, while CSCSC defines them as 10–49 employees) and potential differences in primary respondents' understanding of their enterprise's operations. Additionally, the surveys differ in target populations, including NAICS and enterprise size requirements.

Both surveys apply stratified sampling by industry and firm size, with survey weights correcting for selection probabilities, non-responses, and sampling biases. These weights ensure that the results are representative of the broader population of Canadian enterprises. The SDTIU achieved a response rate of 73%, while the CSCSC had a response rate of 65%. The pre-matched weighted samples cover 327,567 enterprises for the SDTIU and 185,644 for the CSCSC. After merging, the final dataset comprises a weighted sample of 179,657 enterprises, ensuring comprehensive coverage across various business demographics.

Below, we introduce aggregate measures of digital technology usage: BDUS in Section 3.2.1, a separate measure of Business Technological Efficiency (based on k-means clustering) in Section 3.2.2, which examines whether firms encounter significant challenges in implementing the technologies they adopt, and an indicator for Cyber Security Incidence in Section 3.2.3.

3.2.1 Business Digital Usage Score

The BDUS is a quantitative variable constructed to evaluate the digital engagement of Canadian firms by measuring the technologies they have adopted. It condenses responses from the 2021 SDTIU into a single, interpretable score that reflects the cumulative utilization of digital tools across ten distinct domains, such as cloud computing, digital payment systems, artificial intelligence (AI), smart devices, online sales, government digital connectivity, fiber-optic internet, and company websites (see Appendix B.2 for the exact questions).

For each of the ten domains, we create a binary variable equal to 1 if the firm reports using that technology. Summing these indicators yields a score between 0 and 10, with higher values indicating more extensive digital engagement. Although it is conceivable that advanced technologies (e.g., AI) have disproportionate impacts on firm productivity compared to simpler ones (e.g., a basic website), weighting them by perceived importance would introduce additional assumptions. Firms also have heterogeneous technological needs, some rely heavily on cloud computing for scalable data storage, while others benefit more from online

payment systems or data analytics. Therefore uniform summation provides a transparent, easily replicable index of a firm's overall digital footprint.

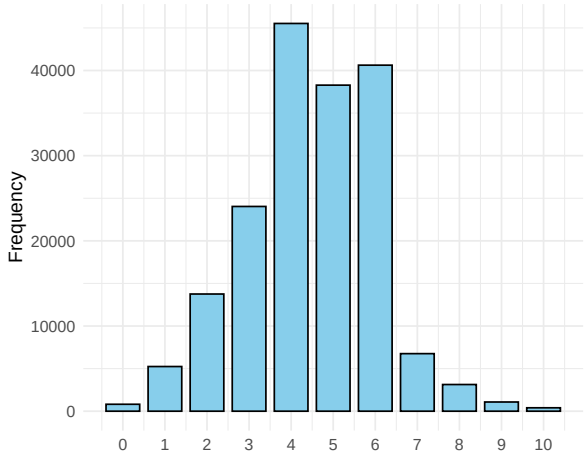


Figure 3.1: Histogram of Business Digital Usage Scores

Figure 3.1 shows the distribution of BDUS scores. The majority of firms fall within scores of 3 to 7, suggesting moderate overall adoption levels, with mild peaks at 4 and 6. At the extremes, a small proportion of firms report virtually no digital engagement (score = 0), whereas another small subset achieves near-comprehensive adoption (scores of 9 or 10). This dispersion highlights the heterogeneity of adoption patterns: some businesses implement only a handful of relevant technologies, while others adopt a wide array.

It is important to note that the BDUS does not measure the intensity of usage or the ease of integration. Rather, it provides a concise snapshot of whether certain well-known tools have been adopted in at least a basic form. We use the BDUS to investigate the characteristics of firms that adopt digital technologies through a stochastic frontier model in Section 3.4, measuring how close businesses are to their technological frontier. Additionally, we use the BDUS to assess whether having a more extensive digital profile correlates with cyber risk exposure (Table 3.1).

3.2.2 Business Technological Efficiency

While the BDUS captures which digital technologies firms adopt, it does not gauge how well these technologies are integrated. To address this, we construct a measure of Technological Efficiency by applying a k-means clustering algorithm to a set of SDTIU survey items about technological implementation challenges. Specifically, these survey items parallel the BDUS domains but ask whether the firm experiences difficulties using each technology, whereas the BDUS questions simply ask if firms use each technology. The exact questions used in this clustering exercise are listed in Appendix B.2.

Let $\{z_{i1}, z_{i2}, \ldots, z_{i10}\}$ be binary indicators for firm i, capturing whether it reports a challenge in each of the ten domains. A response of "Yes" indicates an operational problem in using the technology associated with that domain. We apply k-means clustering to these ten binary variables and determine that $k = 2$ is the optimal cluster count, resulting in two groups: *Digitally Efficient* and *Not Digitally Efficient*. Figure 3.2 displays the percentage of reported challenges in each domain for these two clusters. Firms in the latter cluster (52,505 businesses) report significantly more issues than those in the former (127,152 businesses). For example, 70.71% of *Not Digitally Efficient* firms cite difficulties with AI, compared to 23.58% in the *Digitally Efficient* group. Similarly, 57.55% of *Not Digitally Efficient* firms encounter challenges with cloud computing, versus only 3.21% among *Digitally Efficient* firms.

The high incidence of challenges among *Not Digitally Efficient* firms does not imply that they fail to adopt these tools. Some businesses report both a high BDUS score (indicating widespread adoption) and frequent operational issues, suggesting partial or suboptimal implementation. Even in the *Digitally Efficient* cluster, a non-trivial share of firms faces difficulties in at least one domain.

This k-means classification yields a binary variable, Technological Efficiency, which we use as a dependent variable in one of the logit models in Section 3.4. Whereas the BDUS measures the extent of adoption, the Technological Efficiency grouping reflects the firm's

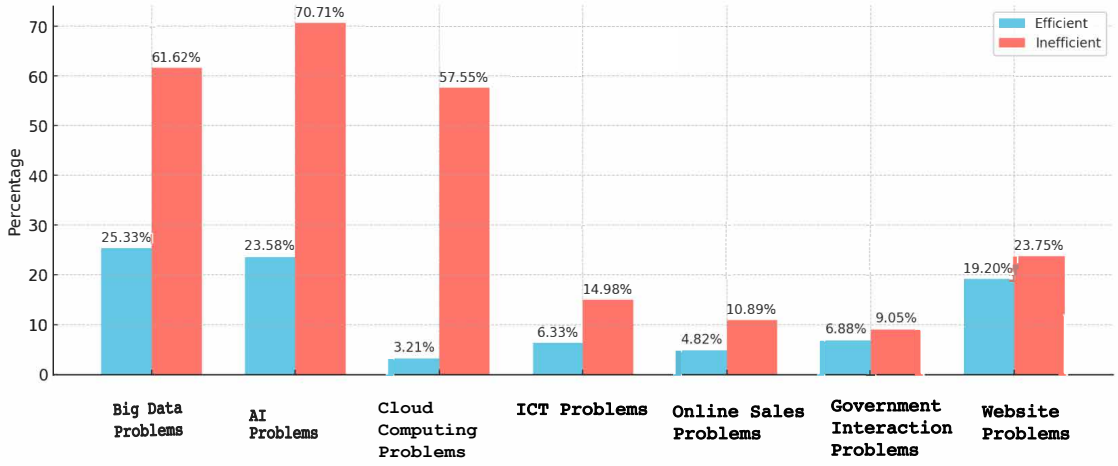


Figure 3.2: Percentage of Technological Problems Between Efficient and Inefficient Clusters

ability to use the technology effectively. In this sense, the two measures are complementary: the BDUS indicates how many digital tools a firm adopts, while the logit model using the Technological Efficiency variable reveals whether the adopted technologies are being used efficiently.

3.2.3 Cyber Security Incidence

To examine the cyber security challenges faced by Canadian businesses, we construct a set of variables based on survey responses related to the occurrence of cyber security incidents. The primary variable, *Cyber Security Incidence*, indicates whether a business was affected by any cyber security incident in 2021. This binary variable is coded as 1 if the business reported experiencing one or more types of cyber security incidents, and 0 otherwise. These incidents range from theft of assets, business data, or intellectual property to disruptions of business activities. Of the 179,656 surveyed firms, 32,371 (18.0%) reported at least one such incident.

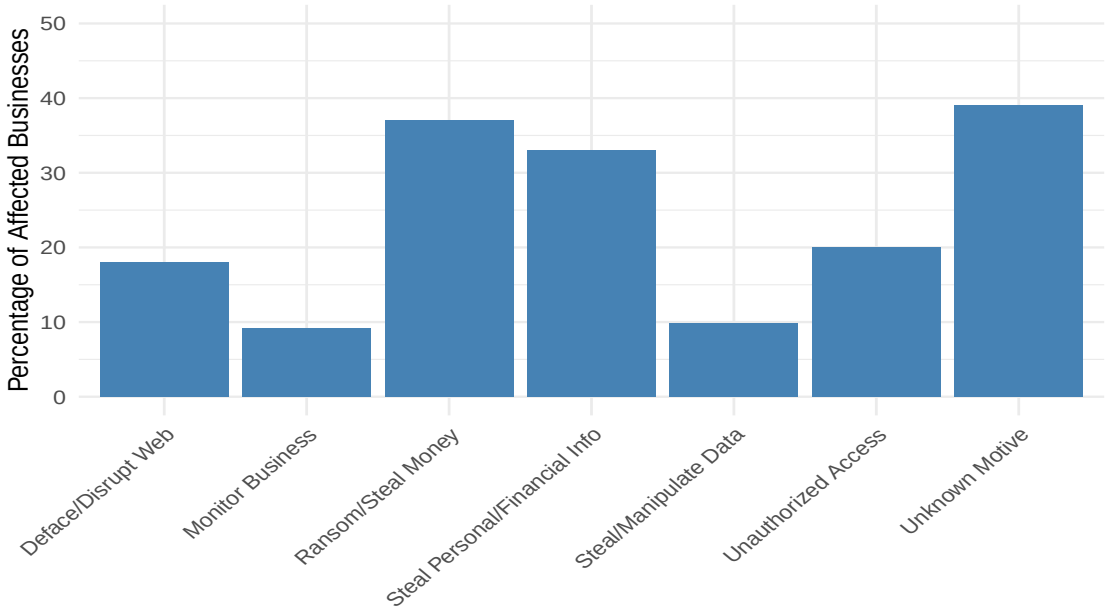


Figure 3.3: Types of Issues Reported by Businesses After Cyber Security Incident (Percentage of Affected Businesses)

Figure 3.3 shows the occurrence of these incident types among the 32,371 firms that experienced any cyber incident. The most frequently cited issues included incidents to steal money or demand ransom payment (37%), incidents with an unknown motive (39%), incidents to steal personal or financial information (33%), incidents to disrupt or deface the business or web presence (18%), and incidents to access unauthorised or privileged areas (20%). Fewer respondents reported incidents to steal or manipulate intellectual property or business data (9.8%) and incidents to monitor and track business activity (9.2%).

We use this binary *Cyber Security Incidence* variable in subsequent analyses for two primary reasons. It serves as a dependent variable in the cyber security incidence logit model (Section 4), where we identify which firm characteristics and digital adoption practices predict a higher likelihood of experiencing an incident. The *Cyber Security Incidence* variable is also included as a regressor in our stochastic frontier analysis to examine whether having experienced a breach correlates with digital technology usage. The rationale is that a prior cyber incident may influence firms' subsequent decisions or capabilities regarding

digital adoption and security investments. A full list of cyber incidence types appears in Appendix B.2.

3.3 Estimation Methodology: Survey Weighted Debiased Lasso

To estimate models predicting technological efficiency and cyber security incidents, we employ a survey-weighted logistic Lasso (hereafter `svy LLasso`) estimation technique and introduce a debiasing method for inference. This approach is well-suited to our analysis, as it effectively handles high-dimensional data and incorporates a diverse set of variables, including digital technology usage, cyber security practices, and firm characteristics. Our primary Lasso models include over 50 independent variables, while second-order interaction models expand to more than 200 variables.

To accurately represent the population of Canadian businesses, we adapt the standard logistic Lasso method to incorporate Statistics Canada survey weights. To this end, we define the following general ℓ_1-penalized maximum likelihood estimator for a Generalized Linear Model with survey weights, where the `svy LLasso` estimator is a special case:

$$\hat{\theta} = \underset{\theta=(\alpha,\beta')'\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \left(-L_n(\theta) + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

where:

- $\theta = (\alpha, \beta')'$ includes the intercept α and coefficients $\beta \in \mathbb{R}^p$,
- $L_n(\theta) = -n^{-1} \sum_{i=1}^{n} w_i g(y_i, x_i'\theta)$ is the survey-weighted log-likelihood, where $g(y, x'\theta)$ is the negative log-density function (see Appendix B.1.1 for a detailed description), x_i is the regressor vector for firm i, y_i is the outcome variable, and w_i is the survey weight,
- $\lambda \sum_{j=1}^{p} |\beta_j|$ is the penalty term, with tuning parameter λ, enforcing sparsity by shrink-

ing less relevant coefficients to zero.

For our logit models, this approach selects the variables most predictive of technological efficiency and cyber security incidents while accounting for survey weights. This framework was adopted by Jasiak and Tuvaandorj (2023) and Jasiak et al. (2024), with the former developing an inference method different from the one we consider below.

Inference via Debiasing. For statistical inference on model parameters and average marginal effects (AME) (Appendix B.1), we adapt the debiased Lasso method from Zhang and Zhang (2014) and Javanmard and Montanari (2014) to the survey context described above. Given the ℓ_1-penalized maximum likelihood estimator $\hat{\theta}$, the debiasing (DB) method applies a one-step correction:

$$\tilde{\theta} = \hat{\theta} + \hat{H}(\hat{\theta})^{-1} S(\hat{\theta}),$$

where $\hat{H}(\hat{\theta})$ and $S(\hat{\theta})$ are the negative Hessian and score function of the weighted log-likelihood function $L_n(\theta)$. The adjustment term $\hat{H}(\hat{\theta})^{-1} S(\hat{\theta})$ corrects the bias introduced by the ℓ_1-penalized variable selection, enabling reliable inference. This variant, which uses the standard Hessian, follows Xia et al. (2023), who consider standard GLMs, adapted here for survey weights. We estimate the asymptotic variance of $n^{1/2} S(\theta_0)$ using a sample information matrix $\hat{I}(\hat{\theta})$ (see Appendix B.1.1), where θ_0 represents the unknown true values of θ.

For a nonlinear parameter function $\rho(\theta)$ (an $r \times 1$ vector, possibly n-dependent), e.g., the AME, we define the debiased estimator:

$$\tilde{\rho} = \rho(\hat{\theta}) + \dot{\rho}(\hat{\theta})' \hat{H}(\hat{\theta})^{-1} S(\hat{\theta}), \quad \dot{\rho}(\theta) = \frac{\partial \rho(\theta)'}{\partial \theta}. \tag{3.3.1}$$

Asymptotic Validity. We establish the asymptotic validity of Wald-type inference in the following proposition. To keep the exposition simple, the underlying framework, definitions,

and assumptions are provided in Appendix B.

Proposition 3.3.1 (Asymptotic Validity of Survey Debiasing Estimator) *Let Assumption 1 hold, and assume that*

- $\lambda = C\sqrt{\frac{\log p}{n}}$ *with* $C = O(1)$, $p \geq 1$,
- $p^2/n \to 0$, *and* $m_0 \log p \sqrt{\frac{p}{n}} \to 0$ *as* $n \to \infty$, *where* m_0 *is the number of non-zero coefficients of* θ_0,
- $\rho(\theta)$ *(with fixed* $r < p+1$*) is differentiable near* θ_0 *with a locally Lipschitz Jacobian* $\dot{\rho}(\theta)$, *and* $\lambda_{\min}(\dot{\rho}(\theta_0)'\dot{\rho}(\theta_0)) > \lambda_l > 0$, *where* $\lambda_{\min}$ *denotes the minimum eigenvalue.*

Then:

$$\left(\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\dot{\rho}(\hat{\theta})\right)^{-1/2} n^{1/2}(\tilde{\rho} - \rho(\theta_0)) \xrightarrow{d} N(0, I_r).$$

The order of the tuning parameter λ is standard in the literature (Bühlmann and van de Geer, 2011; Negahban et al., 2012; van de Geer et al., 2014; Hastie et al., 2015). The assumptions on the number of covariates p and model sparsity m_0 align with those in Xia et al. (2023). In particular, the condition $m_0 \log p \sqrt{p/n} \to 0$ is stronger than the condition $m_0 \frac{\log p}{\sqrt{n}} \to 0$ assumed by van de Geer et al. (2014). However, unlike van de Geer et al. (2014), no direct assumption is imposed on the sparsity of the inverse Hessian (or information matrix).

The assumption of a locally Lipschitz Jacobian $\dot{\rho}(\theta)$ is slightly stronger than the usual continuous differentiability condition required for testing nonlinear hypotheses (see, e.g., Newey and McFadden (1994, Section 9) and Hansen (2022*a*,*b*)). Under this assumption, the estimation error arising from the estimation of θ_0 and $\rho(\theta_0)$ becomes negligible.

Using this proposition, we make inference on the parameters θ_0 and $\rho(\theta_0)$. In sum, the debiasing approach above offers a straightforward and robust way to analyze high-dimensional survey data, ensuring both variable selection and valid inference in the presence of survey weights.

3.4 Empirical Results

The empirical analysis examines the relationship between digital technology adoption and cyber security vulnerabilities among Canadian businesses. We begin by assessing whether broader digital adoption correlates with increased cyber risk (Section 3.4.1) using rank correlations between the BDUS and various security measures. Next, we estimate a Stochastic Frontier Analysis (SFA) model (Section 3.4.2), treating the BDUS as an "output" to identify which factors drive a firm toward its digital usage frontier. We then employ logit models with Lasso selection for two binary outcomes: Technological Efficiency (Section 3.4.3) and Cyber Security Incidence (Section 3.4.4), allowing us to determine whether the same firm characteristics or digital practices that foster greater adoption also enhance implementation efficiency or increase cyber vulnerability.

The sample sizes for the analyses are as follows: the `svy LLasso` results for business efficiency are based on a weighted sample of 175,428 businesses, excluding those with a BDUS score less than 2. The cyber security incidence model uses the full merged dataset, encompassing 179,657 businesses. The `svy LLasso` procedure is implemented using `R`, where we utilize the default value of the tuning parameter λ from the `R` package `glmnet` (see Appendix B.1).

3.4.1 Digital Technology Adoption and Cyber Security

We explore whether a broader digital presence for a firm correlates with heightened cyber risk. Table 3.1 reports polychoric and polyserial correlations between the BDUS (an ordinal score from 0 to 10) that measures the amount of technology a firm has adopted and various cyber security measures, including firms cyber security spending (continuous variable), a binary indicator for whether or not a firm experienced a cyber incident, and whether or not a firm paid ransom as a result of a cyber security breach. We use polychoric correlation for ordinal to binary comparisons and polyserial correlation for the ordinal to continuous

comparison.

The positive and significant correlation between BDUS and both cyber security spending ($\rho = 0.156^{***}$) and experiencing an incident ($\rho = 0.083^{***}$) suggests that while digitally engaged firms invest more in security, they also face greater exposure to attacks. Firms with a higher BDUS are also more likely to have to pay a ransom as a result of a cyber security breach ($\rho = 0.266^{***}$). The absence of any cyber security measures correlates negatively with BDUS ($\rho = -0.068^{***}$), indicating that firms with minimal digital footprints may perceive fewer threats but also forgo basic protective actions.

Table 3.1: Polychoric/Polyserial Correlations Between BDUS and Cyber Security Measures

Cyber Security Measure	**Correlation**	**p-value**
Cyber security spending (numeric)	0.156	$< 0.001^{***}$
Experienced a cyber security incident (binary)	0.083	$< 0.001^{***}$
Firm implemented cloud storage (binary)	0.108	$< 0.001^{***}$
Firm paid ransom (binary)	0.266	$< 0.001^{***}$
Firm implemented no cyber security measures (binary)	-0.068	$< 0.001^{***}$

Notes: The table reports polychoric correlations for ordinal–binary comparisons and polyserial correlations for the ordinal BDUS and the continuous measure of cyber security spending. All correlations are statistically significant at the 1% level. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Sample size: 179,657.

These correlations provide preliminary evidence of a trade-off: as firms adopt more digital tools (higher BDUS), they may both allocate more resources to cyber security and become more frequent targets of attacks. This finding motivates the proceeding empirical analyses.

3.4.2 Digital Technology Adoption by Canadian Businesses

We estimate an SFA model to examine which firm-level and industry-level characteristics bring businesses closer to their "digital usage frontier." Treating BDUS (a quantitative score 0–10) as an "output" — that is, the extent to which a firm adopts digital tools — allows us to separate random variation from systematic shortfalls in adoption. The frontier approach is particularly useful because some firms may lag in adopting new technologies for reasons beyond mere chance (e.g., internal constraints or strategic decisions). In this specification, a

positive coefficient indicates that the corresponding variable moves firms closer to the frontier of digital adoption.

Table 3.2 presents the maximum likelihood estimates for the logistic SFA model. We use a logistic specification to account for the bounded nature of BDUS, recognizing that moving from, for example, two to three adopted technologies represents a more substantial increase in digital capacity than moving from eight to nine. The explanatory variables are grouped into three categories: *Firm characteristics*, *Digital technologies*, and *Cyber security measures*. After controlling for industry fixed effects, the estimated mean efficiency score across all firms is 77%, suggesting that while most businesses integrate some digital tools, gaps remain.

Table 3.2 indicates that several firm characteristics and digital practices significantly influence how close a business is to its digital usage frontier. *Medium* (0.085^{***}) and *Large* (0.173^{***}) firms both exhibit higher BDUS levels than the *Small* reference category, while *Working from home* (0.147^{***}) also aligns positively with adoption. Among digital technologies, *Open source solutions* (0.090^{***}), *Client information management* (0.183^{***}), and *Online advertising* (*Paid*: 0.039^{**}; *Free*: 0.075^{***}) yield significantly higher BDUS.
Certain industry sectors register negative coefficients: *Mining/Utilities/Construction* (-0.095^{***}), *Wholesale/Retail/Transport* (-0.055^{**}), and *Education/Health* (-0.066^{***}). Whereas *Manufacturing* (0.043^{***}) and *Arts/Accommodation/Food* (0.106^{***}) both exceed the baseline category of *Professional services*.

For cyber security measures, the share of *Female employees in ICT roles* (0.002^{***}) is the only strongly significant predictor, indicating that a higher percentage of women in ICT correlates with greater digital adoption. The variance parameter estimates confirm that systematic inefficiency (σ_u) dominates random noise (σ_v), suggesting that much of the under-adoption of digital tools is rooted in structural constraints rather than chance.

These results imply that resource capacity (firm size), remote-work arrangements, and the availability of specific digital tools or managerial practices (e.g., open source, client infor-

Table 3.2: Logistic Stochastic Frontier Model for BDUS

Variable	**Coef.**	**Std. Error**	**z-value**	**p-value**
Intercept	1.721	0.021	82.800	$< 0.001^{***}$
Firm characteristics				
Medium-sized firm	0.085	0.012	6.840	$< 0.001^{***}$
Large-sized firm	0.173	0.016	10.920	$< 0.001^{***}$
Working from home	0.147	0.015	9.630	$< 0.001^{***}$
Mining/Utilities/Construction	−0.095	0.025	−3.800	$< 0.001^{***}$
Manufacturing	0.043	0.015	2.910	0.004^{***}
Wholesale/Retail/Transport	−0.055	0.022	−2.530	0.011^{**}
Education/Health	−0.066	0.017	−3.820	$< 0.001^{***}$
Arts/Accommodation/Food	0.106	0.022	4.840	$< 0.001^{***}$
Other services	0.001	0.027	0.030	0.980
Digital technologies				
Blockchain	0.058	0.030	1.920	0.055^{*}
Open source technologies	0.090	0.014	6.380	$< 0.001^{***}$
Client information management	0.183	0.012	15.540	$< 0.001^{***}$
Paid advertising	0.039	0.017	2.260	0.024^{**}
Free advertising	0.075	0.014	5.440	$< 0.001^{***}$
Firm provides ICT training	0.030	0.029	1.020	0.310
Sales-related problems	0.097	0.040	2.450	0.014^{**}
Cyber security measures				
Gender in ICT roles (% female)	0.002	0.000	5.670	$< 0.001^{***}$
Gender in cyber security (% female)	−0.001	0.000	−0.460	0.644
Cyber security certification	0.014	0.013	1.020	0.306
Cyber security practices	−0.007	0.015	−0.450	0.655
Cyber security training	0.029	0.018	1.620	0.106
Cyber security insurance	−0.019	0.014	−1.410	0.157
Cyber security incidents	0.010	0.017	0.600	0.551
Variance parameters:				
$\ln(\sigma_v^2)$	−4.265	0.130	−32.920	$< 0.001^{***}$
$\ln(\sigma_u^2)$	−2.001	0.088	−22.630	$< 0.001^{***}$
σ_v	0.119	0.008	15.440	$< 0.001^{***}$
σ_u	0.368	0.016	22.630	$< 0.001^{***}$
σ_u/σ_v	3.103	0.021	151.180	$< 0.001^{***}$

Notes: Reference categories: Small firm (firm size) and Professional services (industry). All other variables are binary. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. A positive coefficient indicates that the variable moves the firm closer to its digital usage frontier. Sample size: 179,657.

mation management) can move Canadian firms closer to their feasible frontier of technology usage. Industries such as Construction or Education/Health appear to face adoption barriers, while manufacturing and consumer-facing sectors integrate digital tools more readily.

Although extensive digital adoption may heighten cyber threats (Section 3.4.1), most of the cyber security variables analyzed here do not notably influence how far along a firm is on the adoption curve.

3.4.3 Technological Efficiency of Canadian Businesses

The SFA analysis, using BDUS as the dependent variable in Section 3.4.2, measures the extent of digital adoption but does not capture how effectively these tools are utilized. To address this distinction, we introduce a binary variable for *Technological Efficiency*, derived through k-means clustering. This clustering algorithm categorizes firms based on whether they report few or many operational difficulties in using adopted technologies. A firm is classified as "Technologically Efficient" if it experiences relatively few challenges across multiple domains (see Section 3.2 for details). We restrict the sample to businesses with BDUS ≥ 2, ensuring a baseline level of digital engagement prior to evaluating efficiency.

Table 3.3 presents the `svy LLasso` results for the probability of a firm being Technologically Efficient. The independent variables are grouped into three categories: *Firm characteristics*, *Digital technologies*, and *Cyber security measures*. A positive coefficient indicates that the variable increases the likelihood of efficient technology use, while a negative coefficient suggests a reduced likelihood. The final two columns report the AMEs and their corresponding p-values.

Table 3.3: Debiased Logit Lasso Estimation Results for Technological Efficiency

Variables	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Intercept	−0.060	0.105	0.606	0.122	0.606
Firm characteristics					
Medium firm	0.551	0.856	$< 0.001^{***}$	0.158	$< 0.001^{***}$
Large firm	0.366	1.237	$< 0.001^{***}$	0.102	$< 0.001^{***}$
Remote work	0.509	0.664	$< 0.001^{***}$	0.139	$< 0.001^{***}$

Variables	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Female in ICT roles (1–20%)	.	0.486	0.254	0.023	0.720
Female in ICT roles (21–40%)	.	0.150	0.688	0.139	0.153
Female in ICT roles (41–60%)	0.211	1.066	0.087*	0.123	0.379
Female in ICT roles (>60%)	.	0.915	0.273	−0.032	0.214
Foreign market	.	0.041	0.873	0.031	0.566
Mining/Utilities/Construction	−0.772	−1.108	< 0.001***	0.020	0.458
Manufacturing	.	0.130	0.411	−0.065	0.044**
Wholesale/Retail/Transport	−0.300	−0.416	0.024**	−0.030	0.392
Education/Health	.	−0.193	0.354	0.113	0.006***
Arts/Accommodation/Food	0.256	0.808	0.002***	0.146	0.001***
Other services	0.524	1.086	< 0.001***	0.146	0.001***
Digital technologies					
Blockchain usage	0.086	1.066	0.153	0.046	0.428
ICT training	0.045	0.314	0.374	0.070	0.327
Online orders	.	−0.211	0.166	0.006	0.884
AI	.	0.205	0.520	0.264	< 0.001***
IoT	1.765	1.971	< 0.001***	0.156	< 0.001***
Computer network	0.661	0.989	< 0.001***	−0.001	0.968
Customer relationship management	.	−0.008	0.965	−0.141	< 0.001***
Electronic data interchange	−0.680	−0.883	< 0.001***	−0.101	0.007**
Enterprise resource planning	.	−0.636	0.005***	−0.168	0.050*
Big data usage	0.365	1.350	0.019**	0.076	0.012**
Open source technologies	0.288	0.519	0.007***	−0.022	0.730
Advertising	−0.327	−0.440	0.010**	0.110	< 0.001***
Free advertising	0.431	0.740	< 0.001***	0.073	0.025**
Website	0.399	0.454	0.010**	−0.006	0.852
Company apps	.	−0.041	0.839	−0.087	< 0.001***

Variables	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Social media	−0.287	−0.587	$< 0.001^{***}$	−0.027	0.240
Fiber optic	.	−0.179	0.198	−0.046	0.113
Online sales	−0.064	−0.300	0.083^{*}	−0.013	0.607
Client information management	0.064	0.094	0.498	−0.068	0.019^{**}
Cyber security measures					
Female in cyber security roles (1–20%)	.	−0.141	0.708	−0.012	0.815
Female in cyber security roles (21–40%)	.	−0.077	0.799	0.030	0.443
Female in cyber security roles (41–60%)	.	0.202	0.391	0.049	0.166
Female in cyber security roles (>60%)	0.062	0.329	0.123	−0.019	0.487
Cyber security employees (1–2)	.	−0.127	0.445	−0.003	0.946
Cyber security employees (3+)	.	−0.020	0.941	−0.014	0.622
Cyber security insurance	.	−0.090	0.590	0.014	0.537
Employee monitoring	.	−0.087	0.574	−0.184	$< 0.001^{***}$

Notes: All numeric values are rounded to three decimals. $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ denote the debiased logit Lasso coefficient and AME estimates, respectively. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Reference categories: Small firm, 0% female in ICT roles, 0% female in cyber security roles, 0 cyber security employees, and Industry: Professional services. Sample size: 175,428.

Table 3.3 shows that several firm characteristics significantly increase the probability of using digital tools efficiently. *Medium* ($\widetilde{\text{AME}}^{DB} = 0.158^{***}$) and *Large* ($\widetilde{\text{AME}}^{DB} = 0.102^{***}$) firms exhibit higher AME relative to *Small* firms. *Remote work* ($\widetilde{\text{AME}}^{DB} = 0.139^{***}$) is also associated with increased efficiency. Among digital technologies, the *IoT* ($\widetilde{\text{AME}}^{DB} = 0.156^{***}$) and *Big data analytics* ($\widetilde{\text{AME}}^{DB} = 0.076^{**}$) both show strong positive effects, while advertising efforts, such as free advertising ($\widetilde{\text{AME}}^{DB} = 0.073^{**}$), also correlate with efficient usage. For industries, *Arts/Accommodation/Food* ($\widetilde{\text{AME}}^{DB} = 0.146^{***}$) and *Education/Health* show a positive marginal effect ($\widetilde{\text{AME}}^{DB} = 0.113^{***}$) compared to *Professional services*.

Firms endowed with more resources, due to scale (medium or large size) or operational flexibility (remote work) can allocate staff and capital to integrate digital systems more effectively. Real-time connectivity from IoT and big data appears to reinforce structured work flows, while advertising activities may align with better-organized digital platforms. Consumer-facing industries, such as Arts/Accommodation/Food may capitalize on digital marketing tools or consumer facing technologies such as reservation systems more readily than sectors facing heavier regulatory or operational barriers.

The variables that reduce the likelihood of efficient digital implementation include *Electronic data interchange* ($\widetilde{\text{AME}}^{DB} = -0.101^{***}$) and *Enterprise resource planning* ($\widetilde{\text{AME}}^{DB} = -0.168^{**}$). The AME values suggest that more complex systems can pose challenges with technological adoption. The *Manufacturing* industry ($\widetilde{\text{AME}}^{DB} = -0.065^{**}$) has a negative effect based on AMEs, while certain cyber security variables, such as *Employee monitoring* ($\widetilde{\text{AME}}^{DB} = -0.184^{***}$), also correlate negatively with Technological Efficiency.

These negative or insignificant AMEs point to organizational or regulatory factors that can undermine digital adoption benefits. Complex software solutions (EDI, ERP) often require robust training and IT resources; without sufficient support, firms may experience integration hurdles. Industries like Mining/Utilities/Construction may face specialized work flows or regulatory strictures that may obstruct rapid digital platform adoption. Certain cyber security practices (employee monitoring) can introduce procedural friction or negative employment sentiment that overshadows efficiency gains if not carefully managed.

3.4.4 Cyber Security Incidence

The determinants of *Cyber Security Incidence* variable are analyzed using a binary dependent variable introduced in Section 3.2. The *Cyber Security Incidence* variable equals 1 if a firm reported experiencing at least one cyber security incident during 2021, and 0 otherwise; among 179,656 surveyed firms, 32,371 (18.0%) reported an incident. Cyber incidents encompass a range of adverse events, including theft of business assets, data breaches, disruptions

to business activities, intellectual property losses, and other cyber-related issues. Although a large number of independent variables are included in the analysis, relatively few emerge as statistically significant predictors, indicating that cyber risk is shaped by a limited subset of factors.

Table 3.4: `svy LLasso` Results for Cyber Security Incidence

Variables	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Intercept	−1.904	−2.989	$< 0.001^{***}$	.	.
Firm characteristics					
Medium firm	.	0.047	0.716	0.007	0.723
Large firm	.	0.370	0.038^{**}	0.057	0.030^{**}
Remote work	.	0.104	0.476	0.015	0.488
Female in ICT roles (1–20%)	.	−0.325	0.236	−0.042	0.291
Female in ICT roles (21–40%)	.	0.389	0.178	0.060	0.155
Female in ICT roles (41–60%)	.	0.123	0.781	0.018	0.783
Female in ICT roles (>60%)	.	−0.305	0.660	−0.040	0.694
Mining/Utilities/Construction	.	0.728	0.002^{***}	0.119	0.001^{***}
Manufacturing	.	0.464	0.002^{***}	0.071	0.001^{***}
Wholesale/Retail/Transport	.	0.398	0.034^{**}	0.059	0.032^{**}
Education/Health	.	0.058	0.771	0.008	0.777
Arts/Accommodation/Food	.	0.235	0.394	0.034	0.394
Other services	.	0.484	0.083^{*}	0.075	0.066^{*}
Digital technologies					
Blockchain usage	.	0.202	0.667	0.030	0.663
ICT training	.	0.361	0.168	0.055	0.151
Online orders	.	0.032	0.835	0.005	0.841
AI	.	−0.289	0.207	−0.038	0.255
IoT	.	0.148	0.336	0.021	0.345

Variables	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Computer network	.	0.016	0.909	0.002	0.912
Customer relationship management		0.129	0.419	0.019	0.426
Electronic data interchange	.	−0.176	0.293	−0.024	0.323
Enterprise resource planning	.	0.137	0.477	0.020	0.480
Big data usage	.	−0.414	0.307	−0.053	0.374
Open source technologies	.	−0.102	0.528	−0.014	0.550
Confidential cloud	0.161	0.466	$< 0.001^{***}$	0.067	0.002^{***}
Personal device	.	0.222	0.120	0.031	0.141
VPN	.	0.003	0.986	0.000	0.986
Payment services	.	−0.123	0.663	−0.017	0.682
Client information management	.	0.082	0.569	0.012	0.582
Website	.	−0.098	0.616	−0.014	0.624
Company apps	.	0.112	0.556	0.016	0.563
Social media	.	−0.173	0.255	−0.025	0.265
Online sales	.	0.072	0.663	0.010	0.673
Cyber security measures					
Anti malware	.	−0.236	0.323	−0.034	0.327
Web security	.	−0.115	0.502	−0.016	0.517
Email security	.	0.272	0.271	0.037	0.304
Network security	.	0.087	0.691	0.012	0.704
Data security	.	0.164	0.362	0.023	0.374
POS security	.	−0.115	0.502	−0.016	0.522
Software security	.	−0.251	0.171	−0.034	0.198
Hardware security	.	0.203	0.259	0.029	0.266
Password security	0.159	0.267	0.143	0.038	0.157
Access security	.	0.008	0.963	0.001	0.964
Female in cyber security roles (1–20%)	.	−0.042	0.904	−0.006	0.908

Variables	`svy LLasso`	$\tilde{\theta}^{DB}$	p-value	$\widetilde{\text{AME}}^{DB}$	p-value
Female in cyber security roles (21–40%)	.	0.062	0.819	0.009	0.823
Female in cyber security roles (41–60%)	.	0.188	0.434	0.028	0.434
Female in cyber security roles (>60%)	.	−0.424	0.054*	−0.056	0.082*
Cyber security employees (1–2)	.	0.527	0.061*	0.074	0.07*
Cyber security employees (3+)	.	0.059	0.737	0.008	0.743
Cyber security insurance	.	−0.321	0.050**	−0.043	0.073*
Cyber consultant	.	−0.018	0.941	−0.003	0.943
Cyber information	0.295	0.214	0.450	0.030	0.459
Cyber training	.	0.195	0.254	0.028	0.259
Cyber policy	.	−0.108	0.507	−0.015	0.527
Cyber practice	0.051	0.254	0.331	0.036	0.346
Employee monitoring	.	0.242	0.122	0.035	0.123
Risk assessment	.	0.199	0.219	0.029	0.226
Cyber team: white employees only	.	−0.049	0.785	−0.007	0.794
Cyber team: minority employees only	.	0.424	0.152	0.065	0.131
Cyber certification	.	0.096	0.570	0.014	0.580
No cyber security measures	.	−0.179	0.634	−0.024	0.657

Notes: All numeric values are rounded to three decimals. $\tilde{\theta}^{DB}$ and $\widetilde{\text{AME}}^{DB}$ denote the debiased logit Lasso coefficient and AME estimates, respectively. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Reference categories: Small firm, 0% female in ICT roles, 0% female in cyber security roles, 0 cyber security employees, cyber team: diverse employees. Sample size: 179,657.

Table 3.4 displays the `svy LLasso` estimates for the probability of experiencing a cyber security incident. Among the variables with positive and statistically significant associations are being a large firm ($\widetilde{\text{AME}}^{DB} = 0.057^{**}$) and adopting confidential cloud solutions ($\widetilde{\text{AME}}^{DB} = 0.067^{***}$). In addition, certain industry categories Mining/Utilities/Construction, Manufacturing, Wholesale/Retail/Transport, and Other services exhibit statistically signif-

icant positive coefficients.

Two factors show negative and statistically significant effects on the likelihood of a cyber incident. Having over 60% female employees in cyber security roles ($\widetilde{\text{AME}}^{DB} = -0.056^{*}$) and holding cyber security insurance ($\widetilde{\text{AME}}^{DB} = -0.043^{*}$) are both associated with lower probabilities of experiencing an incident.

These results suggest that larger enterprises may face heightened cyber risks, potentially reflecting the extensive data infrastructures and more valuable assets characteristic of bigger firms. Reliance on confidential cloud solutions might increase exposure to attacks, as remote and cloud-based systems use electronic methods to store data. Lower incidence rates among firms with higher proportions of female cyber security personnel could indicate that diverse cyber teams are more adept at preventing or responding to threats. Likewise, the negative association of cyber security insurance with incidence likelihood hints that insured firms may adopt more robust preventive strategies to manage their risk profiles.

3.4.5 Interaction Effects

Using an adaptive Lasso approach with polynomial expansions, we investigate whether second-order (interaction) terms enhance the predictive accuracy of our *Technological Efficiency* and *Cyber Security Incidence* specifications. We estimate both a first-order (linear) model and a second-order model that includes all pairwise interactions among the explanatory variables. See Bühlmann and van de Geer (2011) for the theoretical background.

Table 3.5 reports mean-squared cross-validation (CV) errors under each specification, along with the penalty parameter λ_{cv}. For the *Technological Efficiency* model, allowing second-order terms substantially lowers the CV error, suggesting that interactions play an important role in explaining efficiency gains from digital technologies. In the *Cyber Security Incidence* model, the simpler first-order specification yields a slightly lower CV error, indicating that higher-order interactions do not improve predictions of cyber security incidence likelihood.

A second-order polynomial specification is justified for the *Technological Efficiency* model, therefore we re-estimate the `svy LLasso` model with interaction terms involving firm size, remote work, key technologies, and industry classifications. Table 3.6 shows the significant interaction terms selected by the `svy LLasso` estimator, along with their debiased parameter estimates $\tilde{\theta}^{DB}$ and p-values. Only the interaction coefficients that were selected by the Lasso and statistically significant at the 5% level are included in Table 3.6.

Table 3.5: Cross-Validation Results for Models With and Without Interaction Terms

Model	**Degree**	λ_{cv}	**CV Error**	**Preferred Specification**
Technological Efficiency	1	0.00261	0.92442	Second-order
	2	0.00029	0.34142	
Cyber Security Incidence	1	0.00685	0.87247	First-order
	2	0.03132	0.86486	

Notes: The table shows the mean-squared CV error from an adaptive Lasso specification with polynomial expansions of different degrees (1 = linear, 2 = second-order interactions). For each model, λ_{cv} denotes the penalty parameter that minimizes the CV error. Based on these metrics, the second-order polynomial is preferred for the Technological Efficiency model, while a first-order specification is preferred for the Cyber Security Incidence model. Sample size: 179,657.

Table 3.6: `svy LLasso` with Interactions: Technological Efficiency

Variables	**Lasso**	$\tilde{\theta}^{DB}$	**p-value**
Intercept	1.095	3.953	0.012**
Remote work	2.641	4.295	0.002***
Online orders	−2.008	−4.555	< 0.001***
IoT	2.249	4.873	0.009***
Computer network	−0.369	−3.173	0.006***
Fiber optic	−1.197	−3.019	0.009***
Medium firm × Remote work	−0.884	−1.428	0.040**
Medium firm × IoT	1.901	5.554	< 0.001***
Medium firm × Cyber security insurance	−0.470	−1.257	0.009***
Medium firm × Website	1.878	2.934	0.004***

Variables	**Lasso**	$\tilde{\theta}^{DB}$	**p-value**
Medium firm × Company apps	−2.170	−4.019	0.002***
Medium firm × Manufacturing	−2.471	−5.658	< 0.001***
Medium firm × Other services	−2.677	−8.226	< 0.001***
Remote work × Female in ICT roles (1–20%)	0.024	−16.870	0.008***
Remote work × CRM	0.340	2.496	0.018**
Remote work × Open source technologies	−1.689	−2.598	0.005***
Remote work × Website	−1.864	−3.230	0.003***
Remote work × Social media	2.361	3.824	< 0.001***
ICT training × Foreign market	0.298	7.453	0.002***
ICT training × CIM	−0.240	−4.743	0.002***
Female in ICT roles (1–20%) × Open source	0.136	9.104	0.005***
Large firm × CRM	−1.116	−4.636	0.004***
Online orders × Computer network	1.278	2.872	0.004***
Online orders × ERP	−0.047	−3.533	0.002***
Online orders × Social media	0.939	3.102	0.002***
Online orders × Manufacturing	0.453	3.207	0.001***
Foreign market × EDI	1.704	5.321	0.003***
Foreign market × CIM	−0.622	−2.730	0.035**
IoT × Computer network	−0.890	−3.443	< 0.001***
IoT × CIM	0.663	2.285	0.032**
IoT × Manufacturing	−2.497	−4.684	0.002***
Computer network × CRM	−2.041	−2.983	0.006***
Computer network × Advertising	−0.191	−3.003	0.018**
Computer network × Fiber optic	1.525	3.576	< 0.001***
Computer network × Other services	−0.815	−5.782	0.039**
CRM × CIM	−0.105	−2.083	0.015**
CRM × Advertising	−1.041	−2.367	0.035**

Variables	**Lasso**	$\tilde{\theta}^{DB}$	**p-value**
EDI × Advertising	−4.335	−5.309	$< 0.001^{***}$
EDI × Online sales	−2.325	−3.303	0.006^{***}
ERP × Manufacturing	0.953	2.583	0.038^{**}
ERP × Wholesale/Retail/Transport	2.113	3.094	0.034^{**}
CIM × Mining/Utilities/Construction	1.256	6.701	$< 0.001^{***}$
CIM × Education/Health	−4.227	−6.218	$< 0.001^{***}$
Advertising × Free advertising	2.001	2.886	0.021^{**}
Advertising × Website	1.479	4.818	0.029^{**}
Advertising × Education/Health	−5.083	−9.202	$< 0.001^{***}$
Free advertising × Website	−0.161	−6.123	0.013^{**}
Free advertising × Social media	−0.298	−3.753	0.010^{**}
Free advertising × Wholesale/Retail/Transport	−3.066	−3.263	0.038^{**}
Free advertising × Education/Health	3.234	6.648	$< 0.001^{***}$
Website × Online sales	1.572	6.424	0.011^{**}
Website × Mining/Utilities/Construction	−1.338	−4.640	$< 0.001^{***}$
Company apps × Social media	5.323	10.627	$< 0.001^{***}$
Company apps × Manufacturing	−0.974	−3.853	0.030^{**}
Company apps × Arts/Accommodation/Food	−0.883	−9.426	0.033^{**}
Social media × Wholesale/Retail/Transport	−2.163	−4.058	$< 0.001^{***}$
Fiber optic × Arts/Accommodation/Food	0.388	4.972	0.049^{**}
Fiber optic × Other services	2.051	9.419	$< 0.001^{***}$
Online sales × Mining/Utilities/Construction	2.554	8.200	$< 0.001^{***}$
Online sales × Manufacturing	−0.384	−2.825	0.007^{***}
Online sales × Other services	−1.933	−9.947	$< 0.001^{***}$

Notes: Numeric values are rounded to three decimal places. $\tilde{\theta}^{DB}$ denotes the debiased logit Lasso coefficient estimate. Coefficients statistically significant at the 5% level based on their p-values are reported. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Variable names are abbreviated for brevity: CRM (Customer

Relationship Management), CIM (Client Information Management), EDI (Electronic Data Interchange), ERP (Enterprise Resource Planning). Sample size: 175,428.

Medium-sized firms exhibit positive and statistically significant interactions with the adoption of the IoT (*Medium firm* $\times$ *IoT*: $\tilde{\theta}^{DB} = 5.554^{***}$) and firm website use (*Medium firm* $\times$ *Website*: $\tilde{\theta}^{DB} = 2.934^{***}$). Remote work arrangements positively interact with social media (*Remote work* $\times$ *Social media*: $\tilde{\theta}^{DB} = 3.824^{***}$) and customer relationship management software (*Remote work* $\times$ *CRM*: $\tilde{\theta}^{DB} = 2.496^{**}$).

The largest statistically significant positive interaction occurs between company-specific applications and social media use (*Company apps* $\times$ *Social media*: $\tilde{\theta}^{DB} = 10.627^{***}$). The presence of female employees in ICT roles (1–20%) interacts positively and significantly with open-source technology adoption (*Female in ICT roles (1–20%)* $\times$ *Open source*: $\tilde{\theta}^{DB} = 9.104^{***}$). Online sales and the Mining, Utilities, and Construction industry have a positive statistically significant interaction (*Online sales* $\times$ *Mining/Utilities/Construction*: $\tilde{\theta}^{DB} = 8.200^{***}$). Additionally, free advertising positively interacts with firms in the Education and Health industry (*Free advertising* $\times$ *Education/Health*: $\tilde{\theta}^{DB} = 6.648^{***}$).

The interaction between EDI and advertising is negative and statistically significant (*Electronic data interchange* $\times$ *Advertising*: $\tilde{\theta}^{DB} = -5.309^{***}$). Similarly, medium-sized firms show a negative and statistically significant interaction with company-specific applications (*Medium firm* $\times$ *Company apps*: $\tilde{\theta}^{DB} = -4.019^{***}$). Remote work arrangements negatively interact with open-source technologies (*Remote work* $\times$ *Open source*: $\tilde{\theta}^{DB} = -2.598^{***}$). Online sales exhibit a negative interaction with firms in the Wholesale, Retail, and Transport industries (*Online sales* $\times$ *Wholesale/Retail/Transport*: $\tilde{\theta}^{DB} = -4.058^{***}$).

Medium-sized firms benefit from adopting IoT solutions and online platforms, likely due to greater resource availability compared to smaller firms. Remote work enhances the efficiency of communication-oriented technologies, such as CRM and social media. The remote work variable itself also exhibits a strong, statistically significant positive effect on firm technological efficiency. Workforce diversity in ICT roles is positively correlated with digital

efficiency, especially when adopting open-source systems. Industry-specific interactions produce varied effects depending on the technology and sector: online sales positively interact with mining, utilities, and construction, while advertising has a positive effect in the education and health sectors. Conversely, interactions like online sales with wholesale, retail, and transport are negatively associated with firm digital efficiency.

3.5 Conclusion

This paper contributes new evidence on how Canadian businesses navigate the trade-off between digital technology adoption and heightened cyber security risk. Using data from Statistics Canada's SDTIU and CSCSC surveys, we construct a BDUS to gauge overall adoption levels and then evaluate how effectively businesses use these tools by modeling their technological efficiency. In addition, we employ a survey-weight-adjusted Lasso estimator and introduce a debiasing method for high-dimensional logit models to identify the drivers of technological efficiency and cyber security risk.

The stochastic frontier analysis suggests that larger firms, remote work arrangements, and specific advanced technologies (e.g., open-source solutions, client information management systems) can push a firm closer to its digital "frontier." At the same time, a portion of businesses lag behind feasible adoption levels, as indicated by the high ratio of inefficiency to noise in the frontier estimations.

Firms must balance efficiency gains against growing cyber vulnerabilities. Firms that adopt more sophisticated digital technologies or store sensitive data in the cloud often face elevated risk exposure. Our `svy LLasso` model on cyber incidence confirm that large firms and those using cloud-based services are more likely to report cyber security incidents. However, the predictive power of cyber risk does not improve with second-order interaction effects, suggesting that firm size and core technological choices are the primary drivers of cyber exposure. The two main variables that decreased the likelihood of cyber incidents were

firms having cyber security insurance and firms that had a high representation of females in cyber security roles.

When it comes to technological efficiency simple linear relationships fail to capture the complexity of how organizational choices, workforce composition, and industry shape digital outcomes. By allowing second-order (interaction) terms, the `svy LLasso` approach shows that certain combinations of variables such as *medium firms* adopting IoT, or *female ICT representation* interacting with advanced tools like AI can be particularly conducive to efficiency improvements. On the other hand, friction in implementing complex software like EDI or ERP can negate some of these benefits.

The analysis demonstrates the importance of firm size and industry. While small firms are sometimes more "locally efficient," the resource advantages of larger organizations may facilitate deeper or more comprehensive integration of technologies. Industries also differ substantially. In resource- and asset-intensive sectors such as Mining or Construction, strong positive interactions emerge between targeted digital solutions and improved operational processes, whereas compliance-heavy fields like Education and Health exhibit more negative or complex relationships.

Gender composition in ICT roles has meaningful consequences for digital adoption outcomes. Although our results do not prove a causal mechanism, the recurring positive coefficients on interactions involving a share of female ICT staff and advanced technologies suggest that even partial gender diversity in technical teams can amplify the returns to adopting new tools. This pattern is also seen in broader research suggesting that heterogeneity in skill sets and perspectives can catalyze creative problem-solving.

The results highlight the balance firms must strike between achieving efficiency gains from digital technologies and managing increased cyber risks. Larger, digitally advanced firms approach their efficiency frontier yet face elevated cyber vulnerability, especially when security practices fall short. Remote work arrangements enhance both digital adoption and technological efficiency without increasing cyber risk exposure. Female representation in ICT

and cyber security roles consistently improves outcomes across adoption, efficiency, and cyber security. Certain cyber security practices, particularly obtaining cyber security insurance and ensuring gender diversity within cyber security roles significantly reduce incident likelihood. Reliance on cloud-based services, notably confidential cloud storage, emerges as a risk factor increasing cyber vulnerability. These results suggest policymakers should implement targeted digital strategies tailored by industry and firm size to boost technological efficiency, while simultaneously establishing baseline cyber security practices such as insurance coverage and workforce diversity to mitigate cyber threats effectively.

Chapter 4

Independence of Irrelevant Alternatives in the Multinomial Logit Model: A Simulation Study

4.1 Introduction

The multinomial logit (MNL) model is widely used in empirical economics and related disciplines due to its analytical convenience and ease of interpretation. Central to the validity of this model is the assumption of Independence of Irrelevant Alternatives (IIA), which states that the odds ratio between any two alternatives remains constant irrespective of other available choices (Hausman and McFadden, 1984).

Given the importance of IIA, Hausman and McFadden (1984) proposed a specification test, the Hausman–McFadden (HM) test, to detect violations of this assumption. The HM test compares parameter estimates from the full choice set against those derived from a restricted subset of alternatives. Under the null hypothesis that IIA holds these estimates should not differ systematically and the resulting statistic follows an asymptotic chi-square distribution. However, subsequent research highlighted significant practical challenges with

the HM test including negative test statistics that contradict theoretical expectations, finite-sample biases, and inaccurate variance estimation (Cheng and Long, 2007; Vijverberg, 2007). Because real-world choice environments often violate IIA, researchers have explored more flexible discrete choice models like nested logit and mixed logit which relax or circumvent the IIA property by allowing correlated errors across alternatives (McFadden and Train, 2000; Hensher and Greene, 2003).

Previous simulation studies have demonstrated possible size and power issues with the HM test. Fry and Harris (1996, 1998) found that the HM test exhibited substantial size distortions, particularly in smaller samples. Cheng and Long (2007) documented that the HM test significantly overrejects the null hypothesis at modest sample sizes ($N = 50$–500), only approaching nominal rejection rates at extremely large samples ($N = 10,000$). Vijverberg (2007) further emphasized the frequent occurrence of negative test statistics and variance estimation issues, demonstrating that corrections for these problems could reduce, but not eliminate size distortions. Fok and Paap (2019) reaffirmed these findings, highlighting both size distortion and limited power, particularly at moderate sample sizes and in cases of subtle IIA violations.

Econometric applications increasingly involve high-dimensional multinomial logit models with many covariates. Traditional maximum likelihood estimation performs poorly in these contexts due to identifiability problems and overfitting. the HM test can be problematic in small samples or higher-dimensional settings and when a model has a large number of covariates (Fry and Harris, 1996, 1998; Cheng and Long, 2007). Regularization methods such as the Lasso (Tibshirani, 1996), have been adapted to discrete choice models providing improved prediction and effective variable selection. However, standard Lasso estimators are biased due to shrinkage, complicating inference. Recently, debiased Lasso techniques have emerged to correct this shrinkage bias, enabling valid asymptotic inference even with high-dimensional data (van de Geer et al., 2014).

This paper evaluates the finite-sample performance of the classical HM test under vari-

ous simulation scenarios. The size and power of the HM test is examined when responding to changes in parameter dimensionality, coefficient magnitudes, covariate types (continuous, categorical, and mixed), and sample sizes. To assess power a nested logit data-generating process is used that deliberately induces violations of IIA through correlated alternatives. A debiased Lasso version of the HM test is developed, which leverages penalized likelihood estimation and a subsequent bias-correction step designed to address challenges arising in high-dimensional settings. Although we include the debiased Lasso approach for completeness, our analysis ultimately finds limited improvements from this method compared to classical approaches.

The remainder of the paper proceeds as follows. Section 4.2 describes the MNL framework and introduces the HM test, including both its classical form and a novel debiased Lasso adaptation designed for high-dimensional contexts. Section 4.3 outlines the Monte Carlo simulation setup, detailing scenarios used to assess size and power under various conditions—including variations in sample size, predictor dimensionality, and types of covariates (continuous, categorical, and mixed). Simulation results are presented in Section 4.4, emphasizing how both the classical and debiased approaches perform in detecting IIA violations. Finally, Section 4.5 summarizes key findings and offers practical recommendations for applying the HM test across diverse data environments.

4.2 Methods

First the classical MNL formulation is outlined, including maximum-likelihood estimation and the key role of IIA in determining consistency. Then the HM test is introduced, comparing the full MNL fit (all choice categories) to a restricted model that omits or collapses one of the alternatives.

The nested logit design is described, which can induce correlated errors, thereby violating IIA and gauging the HM test's empirical power. In high-dimensional settings, we also

use a debiased Lasso approach to handle cases where the number of predictors p is large. This extension preserves the theoretical underpinnings of the HM test while accommodating variable selection and shrinkage. The simulation schemes, including both IIA-compliant and nested logit scenarios, are detailed in Section 4.3, where we vary sample sizes, predictor dimensionalities, and types of covariates (continuous, categorical, or mixed).

4.2.1 Multinomial Logit (MNL) Model

Consider an outcome variable $Y \in \{1, 2, \ldots, J\}$ representing a discrete choice among J possible alternatives. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ denote the covariates for observation $i = 1, \ldots, n$. The multinomial logit (MNL) model assumes that

$$\Pr(Y = j \mid \boldsymbol{x}_i) \;=\; \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{1 + \sum_{\ell=2}^{J} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_\ell)}, \quad j \;=\; 2, \ldots, J, \tag{4.2.1}$$

with the first category serving as the baseline, i.e. $\boldsymbol{\beta}_1 = \boldsymbol{0}$. Here, each $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^\top \in \mathbb{R}^p$ is the coefficient vector for category j, and $\boldsymbol{x}_i$ is a p-dimensional covariate vector for observation i.

Equivalently, we define

$$\eta_{ij} \;=\; \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \quad \text{and} \quad \Pr(Y = 1 \mid \boldsymbol{x}_i) \;=\; \frac{1}{1 + \sum_{\ell=2}^{J} \exp(\eta_{i\ell})},$$

and for $j = 2, \ldots, J$,

$$\Pr(Y = j \mid \boldsymbol{x}_i) \;=\; \frac{\exp(\eta_{ij})}{1 + \sum_{\ell=2}^{J} \exp(\eta_{i\ell})}.$$

Likelihood Inference

Define the indicator $d_{ij} = \mathbf{1}(Y_i = j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. The log-likelihood function for the MNL model is

$$\ell(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J) \;=\; \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log \Pr(Y = j \mid \boldsymbol{x}_i), \tag{4.2.2}$$

subject to the normalization $\boldsymbol{\beta}_1 = \mathbf{0}$. Maximum likelihood estimates $\widehat{\boldsymbol{\beta}}_j$ can be obtained by numerically maximizing (4.2.2). Standard software packages implement this via iterative algorithms such as Newton Raphson.

Independence of Irrelevant Alternatives (IIA)

A key assumption of the MNL specification is the IIA. IIA requires that the ratio of choice probabilities for any two categories is independent of the other alternatives. Specifically, the odds

$$\frac{\Pr(Y = j \mid \boldsymbol{x}_i)}{\Pr(Y = m \mid \boldsymbol{x}_i)}$$

must be invariant to the presence or absence of alternative categories other than j and m. The HM test, described below, is a classical diagnostic for assessing potential violations of IIA.

4.2.2 Hausman–McFadden (HM) Test

Full vs. Restricted Estimates

Given J categories, let

$$\widehat{\boldsymbol{\theta}}_{\text{full}} \;=\; \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{2,\text{full}} \\ \vdots \\ \widehat{\boldsymbol{\beta}}_{J,\text{full}} \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\theta}}_{\text{rest}} \;=\; \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{2,\text{rest}} \\ \vdots \\ \widehat{\boldsymbol{\beta}}_{J,\text{rest}} \end{pmatrix}$$

denote the MLEs of the coefficient vectors in (i) the *full* model using all J categories, and (ii) a *restricted* model that omits one or more alternatives (e.g., re-estimating the logit only on $\{1, 2\}$, treating category 1 as baseline). Under IIA, the estimates should be consistent across the full and restricted problems up to sampling variation.

Test Statistic

The Hausman–McFadden test statistic compares these two sets of estimates. Let

$$\boldsymbol{\Delta} \;=\; \widehat{\boldsymbol{\theta}}_{\text{rest}} \;-\; \widehat{\boldsymbol{\theta}}_{\text{full}},$$

be the difference in coefficient estimates, and define the covariance matrices

$$\widehat{\boldsymbol{V}}_{\text{rest}} \;=\; \text{Var}\Big(\widehat{\boldsymbol{\theta}}_{\text{rest}}\Big), \quad \widehat{\boldsymbol{V}}_{\text{full}} \;=\; \text{Var}\Big(\widehat{\boldsymbol{\theta}}_{\text{full}}\Big).$$

Then the Hausman–McFadden statistic is given by

$$H \;=\; \boldsymbol{\Delta}^{\top}\Big(\,\widehat{\boldsymbol{V}}_{\text{rest}} - \widehat{\boldsymbol{V}}_{\text{full}}\Big)^{-1}\boldsymbol{\Delta}. \tag{4.2.3}$$

Under the null hypothesis of IIA (meaning the restricted and full estimators are both consistent for the same true parameters), H is asymptotically distributed as a chi-square random variable with degrees of freedom equal to the dimension of $\boldsymbol{\Delta}$. Because the restricted model omits one category, we compare only the overlapping parameters (those corresponding to categories 2 and 3) in both models, ensuring $\widehat{\boldsymbol{\theta}}_{\text{rest}}$ and $\widehat{\boldsymbol{\theta}}_{\text{full}}$ share the same dimension. The parameters associated with the dropped category in the full model are excluded from the Hausman comparison.

Nested Logit Simulation Design for IIA Violations

To explore how the Hausman–McFadden test behaves under controlled departures from IIA, we generate data according to a simplified *nested logit*-style correlation structure in the unobserved utilities. For each observation $i = 1, \ldots, n$, let the utility for choice j (with $j \geq 2$) be modeled as

$$U_{ij} \;=\; \boldsymbol{x}_i^\top \boldsymbol{\beta}_j \;+\; \nu_i \;+\; \varepsilon_{ij}, \tag{4.2.4}$$

where:

- $\boldsymbol{x}_i \in \mathbb{R}^p$ follows, for instance, an i.i.d. Gaussian distribution $N(\boldsymbol{0}, \boldsymbol{I}_p)$;
- $\boldsymbol{\beta}_j \in \mathbb{R}^p$ is a deterministic coefficient vector for category j;
- ν_i is a shared random term, capturing correlation among the utilities of alternatives $j = 2, \ldots, J$ within the same "nest";
- ε_{ij} is an i.i.d. Type-I extreme value error that underlies the usual MNL form (or can be subsumed into the MNL framework if we exponentiate and normalize as in (4.2.1)).

Error Term Assumptions. We assume $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$ is independent across i, and is also independent of every ε_{ij}. Each ε_{ij} follows an i.i.d. Type-I extreme value distribution with mean 0 and scale 1, independent across both i and j. Consequently, $\mathrm{Cov}(\nu_i, \varepsilon_{ij}) = 0$ for all i, j. This correlation structure induces a "nested" relationship among the alternatives that share the same ν_i, thereby violating the strict IIA property assumed by the classical MNL model.

In practice, once U_{ij} is realized for each j, we draw Y_i by selecting the category that yields the highest utility. Equivalently, the probability of selecting category j is

$$\Pr(Y_i = j \mid \boldsymbol{x}_i, \nu_i) \;=\; \frac{\exp\Big(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \nu_i\Big)}{1 + \sum_{\ell=2}^{J} \exp\Big(\boldsymbol{x}_i^\top \boldsymbol{\beta}_\ell + \nu_i\Big)},$$

which reduces to the standard MNL form only *conditional* on ν_i. Marginalizing over ν_i induces correlation among the unconditional probabilities of choices $2, \ldots, J$, leading to potential rejections of IIA under the HM test.

Implementation in Simulation Code

Algorithmically, the design proceeds as follows:

1. Generate $\{\boldsymbol{x}_i\}_{i=1}^n$, each i.i.d. $N(\boldsymbol{0}, \boldsymbol{I}_p)$.

2. Draw $\nu_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$.

3. Compute $\eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \nu_i$ for $j \in \{2, \ldots, J\}$ and $\eta_{i1} = 0$ (baseline).

4. Calculate choice probabilities p_{ij} via the softmax form

$$p_{ij} \;=\; \frac{\exp(\eta_{ij})}{\sum_{m=1}^{J} \exp(\eta_{im})}, \quad j = 1, \ldots, J,$$

   and then draw Y_i from the categorical distribution defined by $\Big(p_{i1}, \ldots, p_{iJ}\Big)$.

Debiased Lasso MNL Hausman Test

In high-dimensional settings where the number of covariates p can be large relative to the sample size n, the classical maximum likelihood estimates of the MNL model become unstable or even infeasible. Regularization approaches such as the Lasso (Tibshirani, 1996) can be used to perform variable selection and shrinkage, but the resulting estimates are biased due to the penalization. *Debiased* Lasso methods (van de Geer et al., 2014) correct for this shrinkage in a post-processing step, yielding estimates amenable to asymptotically valid inference.

Full-Model Debiased Lasso MNL

Let $\{\boldsymbol{x}_i, Y_i\}_{i=1}^n$ be the data, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $Y_i \in \{1, \ldots, J\}$. We first fit a penalized MNL model via

$$\min_{\{\boldsymbol{\beta}_j\}_{j=2}^J} \Big\{ -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^J d_{ij} \, \log \Pr(Y_i = j \mid \boldsymbol{x}_i; \{\boldsymbol{\beta}_j\}) \; + \; \lambda \sum_{j=2}^J \|\boldsymbol{\beta}_j\|_1 \Big\},$$

where $d_{ij} = \boldsymbol{1}(Y_i = j)$ and λ is a tuning parameter selected via cross-validation. Denote the resulting *penalized* estimates by $\widetilde{\boldsymbol{\beta}}_j$. A *one-step* correction is then applied to obtain the debiased (DB) estimates $\widehat{\boldsymbol{\beta}}_{j,\mathrm{DB}}$, leveraging approximate Hessian and information matrices computed from the fitted model. For $J = 3$ (to fix ideas), we stack category-2 and category-3 coefficients into

$$\widetilde{\boldsymbol{\theta}} \; = \; \big(\widetilde{\boldsymbol{\beta}}_2, \, \widetilde{\boldsymbol{\beta}}_3\big),$$

and perform a one step update:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{DB}} \; = \; \widetilde{\boldsymbol{\theta}} \; + \; \widehat{\boldsymbol{H}}^{-1} \widehat{\boldsymbol{S}},$$

where $\widehat{\boldsymbol{H}}$ approximates the Hessian of the penalized log-likelihood and $\widehat{\boldsymbol{S}}$ is the score vector. The resulting DB estimates asymptotic normality under suitable regularity conditions (van de Geer et al., 2014), allowing the construction of valid standard errors.

Restricted-Model Debiased Lasso Logit

To form the restricted-model estimates (where we temporarily drop categories $\{3, \ldots, J\}$ and retain only $Y \in \{1, 2, 3\}$), we similarly fit a penalized MNL model on the subset of observations with $Y_i \in \{1, 2, 3\}$. Denoting $\widetilde{\boldsymbol{\beta}}_{2,\mathrm{rest}}$ as the Lasso-fitted coefficients for the restricted MNL model, another one-step correction to obtain the debiased restricted estimate $\widehat{\boldsymbol{\beta}}_{2,\mathrm{rest,DB}}$.

Hausman–McFadden Statistic with Debiased Estimates

With $\widehat{\boldsymbol{\theta}}_{\mathrm{DB}}$ denoting the debiased MNL parameters for the full model and $\widehat{\boldsymbol{\theta}}_{\mathrm{rest,DB}}$ the debiased coefficients from the restricted (binary) logit, the Hausman–McFadden (HM) test statistic is then computed as

$$H_{\mathrm{DB}} \;=\; \Big(\widehat{\boldsymbol{\theta}}_{\mathrm{rest,DB}} \;-\; \widehat{\boldsymbol{\theta}}_{\mathrm{DB}}\Big)^{\top}\Big(\widehat{\boldsymbol{V}}_{\mathrm{rest,DB}} \;-\; \widehat{\boldsymbol{V}}_{\mathrm{DB}}\Big)^{-1}\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{rest,DB}} \;-\; \widehat{\boldsymbol{\theta}}_{\mathrm{DB}}\Big), \tag{4.2.5}$$

where $\widehat{\boldsymbol{V}}_{\mathrm{rest,DB}}$ and $\widehat{\boldsymbol{V}}_{\mathrm{DB}}$ are the estimated covariance matrices of the debiased restricted and full parameter vectors, respectively. Under the null hypothesis that IIA holds (and thus both debiased estimators are consistent for the same parameters), H_{DB} is asymptotically χ^2-distributed with degrees of freedom equal to the dimension of $\widehat{\boldsymbol{\theta}}_{\mathrm{rest,DB}} - \widehat{\boldsymbol{\theta}}_{\mathrm{DB}}$.

4.3 Simulation Design

The simulation study proceeds in two major parts. First, we examine the empirical *size* of the HM test by generating data under a correctly specified MNL model, where the IIA assumption holds and assessing whether the test rejects at or near its nominal significance level of 5%. Second, we investigate the *power* of the test by creating data from a nested logit framework that deliberately violates IIA, observing the test's frequency of detecting this misspecification.

Each simulation involves an outcome variable Y with four discrete alternatives, $Y \in \{1, 2, 3, 4\}$, where category 1 is designated as the baseline ($\boldsymbol{\beta}_1 = \boldsymbol{0}$). Across simulations, we vary the sample size n, considering values 500, 1000, and 5000, and also vary the dimensionality of the predictor vector $\boldsymbol{x}_i \in \mathbb{R}^p$, choosing among dimensions $p = 3, 5$, and 10.

The predictor variables are generated according to three distinct specifications. The first involves continuous predictors drawn from a mixture of normal and uniform distributions.

The second uses categorical predictors, generated as binary variables drawn from Bernoulli distributions. The third specification combines continuous and binary predictors, thus capturing a mixed data structure often encountered in applied research. Additionally, the true coefficient vectors $(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$ are assigned values reflecting small to moderate effect sizes, typically ranging from -1.5 to 1.5, capturing both positive and negative relationships with the predictor variables. This diversity in simulation scenarios enables a comprehensive examination of the HM test under varying conditions likely to be encountered in empirical practice.

Scenarios for Assessing Size. When we enforce IIA by design, each observation's utility for choice j is generated from the standard MNL form:

$$U_{ij} \;=\; \boldsymbol{\beta}_j^{\top} \boldsymbol{x}_i \;+\; \varepsilon_{ij}, \quad \varepsilon_{ij} \;\sim\; \text{i.i.d. Type-I extreme value.}$$

We draw Y_i according to the softmax probabilities implied by these utilities. Because this matches the classical MNL model, the HM test should not reject often if it is well calibrated. High rejection rates (substantially exceeding 5%) are treated as size distortions. To evaluate the empirical size of the HM test, we vary both the type of covariates (continuous, binary, and mixed) and the sample size ($n = 500, 1000, 5000$). Each simulation setting is replicated 1000 times. For each replication, we estimate the full MNL model (with four categories), then fit a restricted binary logit model omitting category $Y = 4$. The HM test statistic is computed and the fraction of replications rejecting the null hypothesis at the nominal 5% significance level is recorded.

Scenarios for Assessing Power. We then shift to *nested logit* data-generating processes, where correlation among certain alternatives is induced via nest-specific scale parameters. Concretely, we group $\{1, 2\}$ as one nest and $\{3, 4\}$ as another, specify nest-specific parameters (λ_A, λ_B), and generate the utilities for each choice following the typical nested logit form.

Observed predictors $\boldsymbol{x}_i$ (again, of dimension p) are drawn in the same ways described above (continuous, categorical, or mixed). Because the true data-generating process now violates IIA, the classical MNL-based HM test is expected to reject more frequently as n increases. We repeat each scenario (with, $n \in \{500, 1000, 5000\}$ and $p \in \{3, 5, 10\}$) for 1000 replications, again tracking the test's empirical rejection rate.

4.3.1 Estimation Approaches.

In all scenarios, the *classical* Hausman–McFadden test is computed by:

1. Fitting a full MNL model on $\{1, 2, 3, 4\}$ (via `multinom`),
2. Restricting to $\{1, 2, 3\}$ by dropping choice 4 (and re-fitting `multinom` on that subset),
3. Forming the test statistic from the difference in parameter estimates and their covariance matrices as in (4.2.3).

A *Debiased Lasso* counterpart is included when p becomes large relative to n. Specifically, we replace the two MNL fits with ℓ_1-penalized multinomial regressions (`glmnet`), followed by a one-step correction to remove shrinkage bias. This yields *debiased* estimates for the full and restricted models, which are then plugged into the same Hausman–McFadden testing framework (van de Geer et al., 2014).

Outcome Measures. For each setting, we compute the empirical rejection rate over 1000 replications at a nominal 5% level. Under the null (i.e., MNL with true IIA), this rate ideally remains near 5% (no size distortion). Under violations (nested logit), higher rejection rates indicate greater power.

Overall, the final design includes permutations of:

$$n \in \{100, 200, 500, 1000, 5000\},$$
$$p \in \{3, 5, 10\},$$
$$\text{covariate types: } \{\text{continuous, categorical, mixed}\},$$
$$\text{coefficient patterns: } \{\text{sparse, dense, nested-logit}\}.$$

We replicate each combination $n_{\text{sim}} = 1000$ times (or as specified) to obtain stable estimates of the Hausman–McFadden test's empirical rejection rates.

4.4 Results

We begin by examining the empirical size of the classical HM test when the MNL model is correctly specified and the IIA assumption holds. Under these conditions, the HM test should reject the null hypothesis at approximately the nominal significance level of 5%.

After, we assess the test's power to detect violations of IIA by generating data from a nested logit structure, intentionally violating the MNL model's IIA assumption. This power analysis allows us to gauge the sensitivity of the HM test to realistic departures from IIA. Finally, we extend our evaluation to the proposed debiased Lasso version of the HM test, investigating whether this approach improves finite-sample control of size and power in high-dimensional settings.

We organize our results by first discussing the size properties of the classical HM test, followed by an analysis of its power properties. Lastly, we present findings for the DB-Lasso HM framework, highlighting key improvements or limitations relative to the standard method.

4.4.1 Size Performance of the Classical MNL HM Test

We begin by examining the empirical size of the HM test under scenarios in which the MNL model is correctly specified (IIA holds). Table 4.1 (and the corresponding plots) summarize the rejection rates for three types of predictor variables (*continuous*, *mixed*, and *categorical*), varying the number of predictors ($p = 3, 5, 10$) and the sample size ($n = 500, 1000, 5000$). Ideally, under a nominal 5% test, these rejection rates should cluster around 5%.

Across all three data-generation schemes (continuous, mixed, and categorical), the rejection rates for *3*- and *5*-predictor models tend to hover closer to the nominal 5% level, with only mild fluctuations (e.g., between approximately 2% and 10% in most cases). When the model includes *10* predictors, the test systematically displays higher rejection frequencies, often exceeding 10% or more, signaling a tendency to reject IIA too often. This pattern emerges consistently, suggesting that in moderate- to high-dimensional settings, the classical HM test may reject too frequently even when IIA is true.

Although the overall size patterns are broadly similar, minor distinctions appear depending on how the predictors are generated. In the *continuous* and *mixed* designs, the rejection rates for 3- and 5-predictor models generally remain reasonably near the nominal level, often in the 5–10% range. For the 10-predictor models, however, continuous designs sometimes lead to even higher rejections (e.g., around 12–14%), whereas the mixed designs exhibit similarly elevated rates (10–12%). The *categorical* designs, by contrast, sometimes produce lower-than-nominal rejection frequencies (e.g., below 5%) for 3- and 5-predictor models, but again trend upward when the model size grows to 10 predictors.

An interesting finding concerns the evolution of rejection rates as n increases from 500 to 1000 and 5000. In many scenarios, especially with fewer predictors, the HM test moves slightly closer to the nominal 5% level or remains stable. However, in the 10 parameter settings, the rejection rates often do not shrink to near 5% even as the sample size grows. For example, certain continuous or mixed 10-predictor scenarios record rejection rates of 10–15% at $n = 500$ and remain elevated (above 10%) even by $n = 5000$. These results suggest

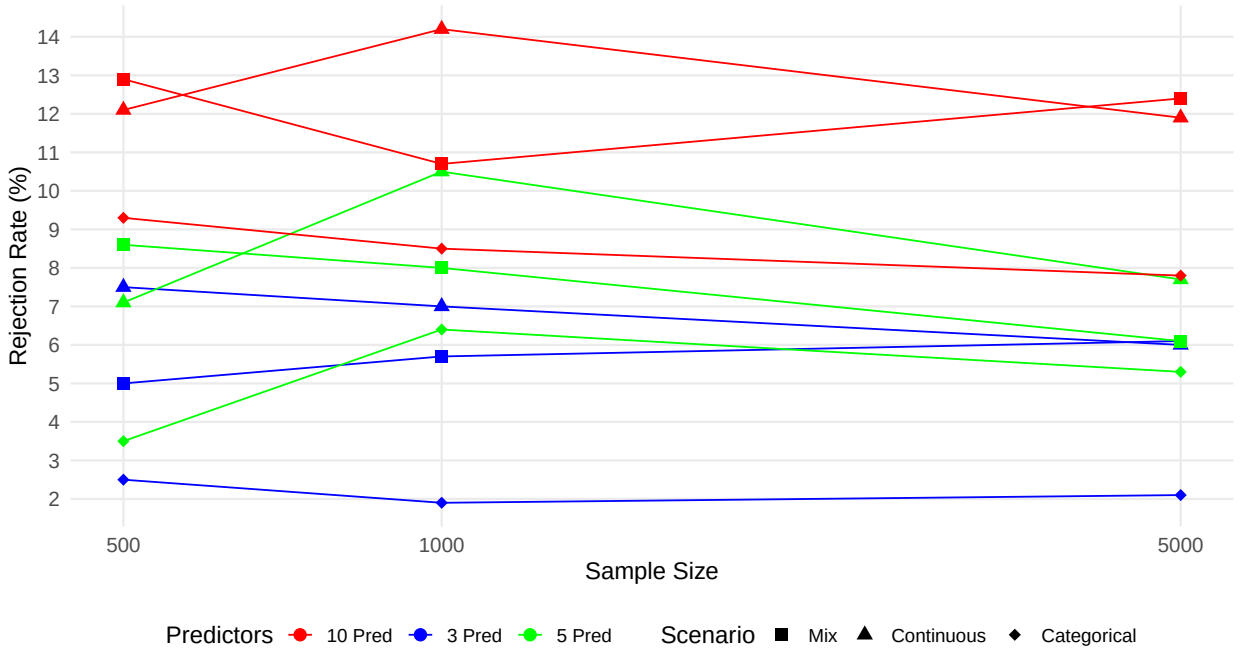


Figure 4.1: Rejection rates of the Hausman–McFadden test under varying parameter types and sample sizes

that simply increasing n is not always sufficient to control Type I error in higher-dimensional regressions.

The classical MNL HM test exhibits acceptable size control in lower-dimensional models (3–5 predictors) across the different data-generation schemes, with moderate deviations from the nominal 5% level. As the number of predictors rises to 10, systematic over-rejection emerges consistently, reflecting possible instability of maximum-likelihood estimators or inflated variability in high-dimensional MNL settings. In practical applications with numerous predictors, the classical HM test may overly reject the null hypothesis of IIA.

4.4.2 Power Performance of the HM Test under Nested Logit Violations

The empirical power of the HM test in detecting violations of the IIA assumption is tested by simulating data from a nested logit model. In this setting, the MNL model is misspecified:

Table 4.1: IIA Rejection Rates (%) Across Sample Sizes and Predictor Counts

Scenario	Predictor Count	n = 500	n = 1000	n = 5000
Mix	3 Pred	5.0	5.7	6.1
	5 Pred	8.6	8.0	6.1
	10 Pred	12.9	10.7	12.4
Continuous	3 Pred	7.5	7.0	6.0
	5 Pred	7.1	10.5	7.7
	10 Pred	12.1	14.2	11.9
Categorical	3 Pred	2.5	1.9	2.1
	5 Pred	3.5	6.4	5.3
	10 Pred	9.3	8.5	7.8

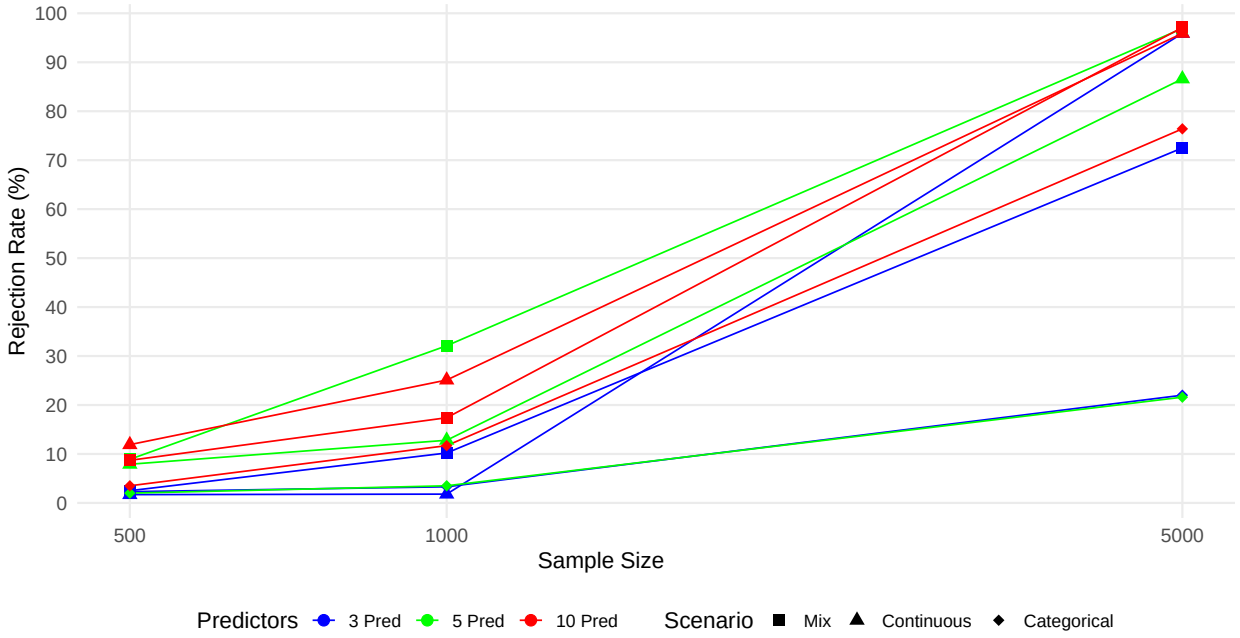


Figure 4.2: Rejection rates of the Hausman–McFadden test using a nested logit specification, across varying predictor scenarios and sample sizes.

IIA does not hold. Table 4.2 (and the corresponding plot 4.2) summarize rejection rates for three types of predictor variables (*continuous*, *mixed*, and *categorical*), varying the number of predictors ($p = 3, 5, 10$) and sample sizes ($n = 500, 1000, 5000$). For a well-powered test the rejection rates should increase substantially beyond the nominal 5% level as n grows and/or as the IIA violation becomes more pronounced.

Under the *mixed* variable design, the HM test displays low power when $n = 500$ and $p = 3$ (2.5% rejections), but increases to over 70% once n reaches 5000 for the same dimension. The increase from $n = 500$ to $n = 1000$ is smaller, from 2.5% to 10.2% with 3 predictors, and from 8.9% to 32.1% with 5 predictors. For *5* or *10* predictors, the test can exceed 90% rejection by $n = 5000$, but also demonstrates relatively low power in the other sample size scenarios.

In the *continuous* scenario, rejection rates also climb at higher n, yet show minimal improvement at the lower to mid sample range in certain cases. With 3 predictors, the rate only moves from 1.7% to 1.8% from $n = 500$ to $n = 1000$, before leaping to 95.9% at $n = 5000$. The 10-predictor models jump from 25.1% at $n = 1000$ to nearly 96% for $n = 5000$.

In the *categorical* design, the HM test shows small power at $n = 500$ and $n = 1000$, often below 12% across 3 or 5 predictors.Even at $n = 1000$, improvements are sometimes mild (e.g., from 2.3% to 3.3% for 3 predictors). Once $n = 5000$, the test achieves higher rejection rates, 76.4% for 10 predictors, though still less dramatic than in continuous or mixed scenarios when p is small. Purely categorical setups may require larger sample sizes to reveal IIA failures more decisively.

As the sample size grows, so does the frequency with which the HM test rejects the (incorrectly assumed) MNL model. The rate and extent of this increase depend on the dimensionality (p) and the nature of the predictors (continuous vs. categorical), but overall, the classical HM test demonstrates good power properties in large samples. For smaller n, detection is limited, particularly in categorical setups. These results affirm that with enough observations or model complexity, the HM test can effectively flag IIA violations, although the exact levels of power can vary considerably by data type and the number of predictors involved.

Table 4.2: Nested Logit MNL Hausman Test Rejection Rates in %

Predictor Count	n = 500	n = 1000	n = 5000
Mixed Scenario			
3 Pred	2.5%	10.2%	72.5%
5 Pred	8.9%	32.1%	96.9%
10 Pred	8.7%	17.4%	97.1%
Continuous Scenario			
3 Pred	1.7%	1.8%	95.9%
5 Pred	7.9%	12.8%	86.6%
10 Pred	11.9%	25.1%	95.9%
Categorical Scenario			
3 Pred	2.3%	3.3%	22.0%
5 Pred	2.0%	3.5%	21.6%
10 Pred	3.5%	11.7%	76.4%

4.4.3 Debiased Lasso HM Test

The classical maximum likelihood estimation-based HM test demonstrated acceptable size properties for moderate-dimensional scenarios (3–5 predictors). However, as documented in Section 4.4.1, significant size distortions occurred when the number of predictors increased to ten, even with large sample sizes ($n = 5000$). Given these limitations, a natural alternative is to consider regularization approaches that can handle high-dimensional predictor spaces effectively. The *Debiased Lasso* (DB-Lasso) approach is used to improve upon the classical HM test by mitigating issues related to variance inflation and instability in high-dimensional MNL settings.

The DB-Lasso approach leverages the strengths of Lasso regularization, which is appealing when the number of predictors p is large, as it can shrink coefficients and perform variable selection. Standard Lasso estimators are inherently biased due to the ℓ_1 penalty complicating inference on the coefficients the model produces. To correct for this bias a one-step "debiasing" procedure is implemented (van de Geer et al., 2014), adjusting the penalized estimates using an update derived from the model's score function and the inverse Hessian evaluated at the penalized solution. Once these debiased estimates are obtained for the *full*

model (all J alternatives) and the *restricted* model (with one alternative removed) the HM test statistic can be constructed using the difference between these two debiased estimators and their respective covariance matrices.

To investigate the finite-sample behavior of the HM test using the DB-Lasso method, we conduct the same *size* and *power* experiments described earlier, but replace the classical MNL fits with the DB-Lasso method. Under *size* scenarios (correctly specified), we find that the DB-Lasso HM test never rejects the null hypothesis, always falling below the nominal 5% level. Under *power* scenarios (nested logit data), we similarly observe no rejections when testing on the same data generating process as used in Section 4.4.2.

To further investigate the lack of power in the DB-Lasso HM test, additional scenarios beyond those initially considered in Section 4.4.2 are tested. A more pronounced violation of IIA is introduced by creating more strongly correlated nests and significantly increasing the magnitude of the regression coefficients. However, even under these deliberately severe conditions, the DB-Lasso approach still produces very low rejection rates. Rejections of the null remained infrequent and inconsistent, likely due to conservative inference resulting from the one-step debiasing correction around heavily penalized initial estimates. The HM test statistics tend to remain near zero, implying insufficient sensitivity to detect genuine IIA violations. Unlike the classical HM test which gains power as the sample size or magnitude of violation increases, the DB-Lasso variant fails to reject under practically relevant conditions rendering it unsuitable for empirical applications in testing the IIA assumption.

4.5 Conclusions

This paper examined the performance of the HM specification test for the IIA assumption in MNL models. Through a series of simulation experiments, we evaluated both the *size* of the test (when MNL is correctly specified and IIA holds) and its *power* (when data are generated from a nested logit framework that deliberately violates IIA). The classical HM

test exhibits acceptable size control in modest-dimensional settings (three to five predictors) but can display inflated Type I error as the number of covariates grows larger. In power simulations, we find that the HM test can detect significant violations of IIA when the sample size is sufficiently large, although it often lacks power in small samples or in subtle nesting scenarios.

We also implemented and evaluated a *Debiased Lasso* HM test. This adaptation applies penalized multinomial logit estimation followed by a one-step adjustment to correct for shrinkage bias in the Lasso estimates. While such debiasing techniques have proven effective in certain linear or binary logistic regressions, we consistently found that the debiased Lasso procedure severely under-rejects the null of IIA in our settings. For both the correctly specified (size) and the nested logit (power) simulations, the DB-Lasso HM test produced a near-zero rejection rate except in extreme nesting structures. It is suspected that the one-step corrections may dramatically inflate the variance estimates around the heavily penalized solution, leading to a very conservative test statistic.

Empirical researchers interested in testing the IIA property have several practical takeaways from these results. First, when the dimensionality of the predictor set is moderate (three to five variables) and sample sizes exceed a few thousand, the standard HM test appears reasonably stable and can detect meaningful violations of IIA. Second, in higher-dimensional contexts where there may be ten or more predictor variables, the classical HM test's performance degrades (both in size and power). Third, the proposed debiased Lasso version of the HM test does not offer a viable solution to the limitations identified in the classical MNL HM test. More sophisticated multi-step corrections or alternative forms of regularized estimation might be required to preserve power without inflating variance estimates.

These simulations highlight both the utility and the practical limitations of the HM test in MNL contexts. The test can be effective at large sample sizes (n) and moderate predictor dimensionality (p), its stability is not guaranteed in higher-dimensional or smaller-

sample applications. The debiased Lasso alternative, aimed at handling such scenarios, failed to exhibit sufficient rejection rates when IIA was violated. Future work might explore alternative multi-step corrections or inference approaches better suited to the penalized structure of multinomial logit models, facilitating valid IIA tests even in high-dimensional settings.

Chapter 5

Conclusion

Digital transformation in Canada is uneven. Some individuals and firms push the boundaries of adoption, while others remain persistently excluded or under-equipped. The first two studies collected here examine different aspects of this: disparities in digital access among individuals and the challenges firms face in balancing digital advancement with cyber security risk. The third study examines statistical tools available for analyzing complex, high-dimensional behavior in these contexts.

Across Canadian households, patterns of internet use and engagement with digital financial tools reveal persistent divides shaped by income, age, region, and education. Yet some findings run against prevailing narratives: women, particularly those in the workforce exhibit strong digital literacy, and recent immigrants often match or exceed the digital engagement of Canadian-born individuals. Residents of the Maritime provinces and Manitoba, along with lower-income seniors, remain disproportionately underconnected. These regional and demographic gaps suggest that infrastructure investments alone are insufficient; the ability to navigate and benefit from digital systems depends on a deeper intersection of resources, familiarity, and institutional trust.

Within the business sector, adoption of digital tools is widespread, but efficiency varies considerably. Larger firms and those using remote work arrangements tend to operate closer

to the digital frontier, yet face greater exposure to cyber threats, especially when relying on cloud-based infrastructure. The analysis shows that cyber security risks are not randomly distributed: firms with cyber insurance and more female representation in ICT roles are less likely to report cyber incidents. Gender diversity in technical teams is associated with gains not just in security outcomes but also in overall digital efficiency.

Analyzing behavior in these large survey settings increasingly requires high-dimensional estimation tools. Penalized models such as the Lasso offer flexibility, but standard inferential procedures struggle under regularization. Simulation results show that the widely used Hausman-McFadden test for IIA breaks down in higher-dimensional multinomial contexts, and naive debiasing methods fail to recover reliable inference. These findings reinforce the need for continued development of inference frameworks that match the complexity of modern data, particularly when policy conclusions hinge on subtle substitution patterns or interaction effects.

Targeted investments in digital literacy and infrastructure remain critical for addressing persistent divides in digital adoption across Canadian households and firms. At the enterprise level, improving cyber security practices and promoting workforce diversity, particularly gender representation in technical roles, offers potential benefits in both security outcomes and overall digital efficiency. Researchers and policymakers must advance statistical methods suited to high-dimensional datasets, as traditional methods often fall short under increasing complexity. Progress in statistical methods will enable clearer interpretation of large scale survey data, supporting more precise policy design and implementation.

Bibliography

Aczel, B., Kovacs, M., Van Der Lippe, T. and Szaszi, B. (2021), 'Researchers Working from Home: Benefits and Challenges', *PloS one* **16**(3), e0249127.

Adrian, T. and Mancini-Griffoli, T. (2019), 'The Rise of Digital Money, International Monetary Fund', *Annual Review of Financial Economics* **13**, 57–77.

Aghimien, D., Aigbavboa, C., Meno, T. and Ikuabe, M. (2021), 'Unravelling the Risks of Construction Digitalisation in Developing Countries', *Construction Innovation* **21**(3), 456–475.

Ahnert, T., Brolley, M., Cimon, D. A. and Riordan, R. (2022), 'Cyber Security and Ransomware in Financial Farkets', *Available at SSRN 4057505* .

American Library Association (n.d.), 'Digital Literacy'.
URL: *https://literacy.ala.org/digital-literacy/*

Aston, J., Vipond, O., Virgin, K. and Youssouf, O. (2020), Retail E-Commerce and COVID-19: How Online Shopping Opened Doors While Many were Closing, Technical report, Statistics Canada.

Aversa, J., Shaker, R., Cleave, E. and Salooja, N. (2022), 'Determinants of Bank Closures: Exploring the Relationship between Neighborhood Characteristics and Bank Branch Locations', *Papers in Applied Geography* **8**(4), 393–413.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014), 'High-Dimensional Methods and Inference on Structural and Treatment Effects', *Journal of Economic Perspectives* **28**(2), 29–50.

Belloni, A., Chernozhukov, V. and Wei, Y. (2016), 'Post-Selection Inference for Generalized Linear Models with Many Controls', *Journal of Business & Economic Statistics* **34**(4), 606–619.

Bilodeau, H., Lari, M. and Uhrb, M. (2019), 'Cyber Security and Cybercrime Challenges of Canadian Businesses, 2017', *Juristat: Canadian Centre for Justice Statistics* pp. 1–18.

Blichfeldt, H. and Faullant, R. (2021), 'Performance Effects of Digital Technology Adoption and Product & Service Innovation–A Process-Industry Perspective', *Technovation* **105**, 102275.

Bousquet, T. (2023), '100,000 Current and Past Nova Scotia Health, IWK, and Public Service Employees Had Their Payroll Information Stolen in MOVEit Breach', https://www.halifaxexaminer.ca/government/province-house/100000-current-past-nova-scotia-health-employees-had-their-payroll-information-stolen-in-moveit-breach/.

Bridge, S. and Zoledziowski, A. (2024), '1 Million Books and 4 Months Later, Toronto's Library Recovers from a Cyberattack', `https://www.cbc.ca/news/canada/toronto/toronto-library-ransomware-recovery-1.7126412`.

Bühlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media.

Cameron, A. C. and Trivedi, P. K. (2009), *Microeconometrics: Methods and Evaluations*, Cambridge University Press.

Carson, S. A. (2013), 'The Remote Rural Broadband Deficit in Canada', *Journal of Rural and Community Development* **8**(2).

Cebula, J. J. and Young, L. R. (2010), A Taxonomy of Operational Cyber Security Risks, Technical report, Software Engineering Institute, Carnegie Mellon University.

Chen, H., Engert, W., Huynh, K. P. and O'Habib, D. (2021), An Exploration of First Nations Reserves and Access to Cash, Technical report, Bank of Canada Staff Discussion Paper.

Chen, H., Engert, W., Huynh, K. P. and O'Habib, D. (2022), Identifying Financially Remote First Nations Reserves, Technical report, Bank of Canada Staff Discussion Paper.

Chen, H., Engert, W., Huynh, K. P., O'Habib, D., Wu, J. and Zhu, J. (2022), Cash and COVID-19: What Happened in 2021, Technical report, Bank of Canada.

Chen, H., Engert, W., Huynh, K. P., O'Habib, D. and Zhu, J. (2021), Cash and COVID-19: The Impact of the Second Wave in Canada, Technical report, Bank of Canada Staff Discussion Paper.

Cheng, S. and Long, J. S. (2007), 'Testing for iia in the multinomial logit model', *Sociological Methods & Research* **35**(4), 583–600.

Cheung, C., Lyons, J., Madsen, B., Miller, S. and Sheikh, S. (2021), The Bank of Canada COVID-19 Stringency Index: Measuring Policy Response Across Provinces, Technical report, Bank of Canada Staff Analytical Notes.

Cullen, R. (2001), 'Addressing the Digital Divide', *Online Information Review* **25**(5), 311–320.

Deng, Z., Morissette, R. and Messacar, D. (2020), Running the Economy Remotely: Potential for Working From Home During and After COVID-19, Technical report, Statistics Canada.

Dewan, S. and Riggins, F. J. (2005), 'The Digital Divide: Current and Future Research Directions', *Journal of the Association for Information Systems* **6**(12), 298–337.

Dufour, J.-M., Trognon, A. and Tuvaandorj, P. (2016), Generalized $C(\alpha)$ Tests in Estimating Functions with Serial Dependence, *in* W. K. Li, D. Stanford and H. Yu, eds, 'Advances

in Time Series Methods and Applications: the McLeod Festschrift', Springer, Berlin and New York, pp. 151–178.

Eling, M., Schnell, W. and Sommerrock, F. (2016), *Ten Key Questions on Cyber Risk and Cyber Risk Insurance*, The Geneva Association, Zurich. Available at `https://www.genevaassociation.org`.

Engert, W. and Huynh, K. P. (2022), Cash, COVID-19 and the Prospect for a Canadian Digital Dollar, Technical report, Bank of Canada Staff Discussion Paper.

Ferrari, A. (2012), *Digital Competence in Practice: An Analysis of Frameworks*, Vol. 10, Luxembourg: Publications Office of the European Union.

Ferreira, D., Vale, M., Carmo, R. M., Encalada-Abarca, L. and Marcolin, C. (2021), 'The Three Levels of the Urban Digital Divide: Bridging Issues of Coverage, Usage and Its Outcomes in VGI Platforms', *Geoforum* **124**, 195–206.

Fitch Ratings (2021), 'Sharply Rising Cyber Insurance Claims Signal Further Risk Challenges'.

URL: *https://www.fitchratings.com/research/insurance/sharply-rising-cyber-insurance-claims-signal-further-risk-challenges-15-04-2021*

Fok, D. and Paap, R. (2019), New misspecification tests for multinomial logit models, Econometric Institute Report EI2019-24, Econometric Institute, Erasmus University Rotterdam.

Friedline, T., Naraharisetti, S. and Weaver, A. (2020), 'Digital Redlining: Poor Rural Communities' Access to Fintech and Implications for Financial Inclusion', *Journal of Poverty* **24**(5-6), 517–541.

Fry, T. R. L. and Harris, M. N. (1996), 'A monte carlo study of tests for the independence of irrelevant alternatives property', *Transportation Research Part B: Methodological* **30**(1), 19–30.

Fry, T. R. L. and Harris, M. N. (1998), 'Testing for independence of irrelevant alternatives: Some empirical results', *Sociological Methods & Research* **26**(3), 401–423.

Gartner, Inc. (2021), 'Gartner Forecasts Worldwide Security and Risk Management Spending to Exceed $150 Billion in 2021'.
URL: *https://www.gartner.com/en/newsroom/press-releases/2021-05-17-gartner-forecasts-worldwide-security-and-risk-managem*

Hackney, A., Yung, M., Somasundram, K. G., Nowrouzi-Kia, B., Oakman, J. and Yazdani, A. (2022), 'Working in the Digital Economy: A Systematic Review of the Impact of Work-from-Home Arrangements on Personal and Organizational Performance and Productivity', *Plos one* **17**(10), e0274728.

Haight, M., Quan-Haase, A. and Corbett, B. A. (2014), 'Revisiting the Digital Divide in Canada: The Impact of Demographic Factors on Access to the Internet, Level of Online Activity, and Social Networking Site Usage', *Information, Communication & Society* **17**(4), 503–519.

Hansen, B. (2022*a*), *Econometrics*, Princeton University Press.

Hansen, B. (2022*b*), *Probability and Statistics for Economists*, Princeton University Press.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC press.

Hausman, J. and McFadden, D. (1984), 'Specification Tests for the Multinomial Logit Model', *Econometrica: Journal of the Econometric Society* pp. 1219–1240.

Henry, C. S., Engert, W., Sutton-Lalani, A., Hernandez, S., McVanel, D. and Huynh, K. P. (2023), 'Unmet Payment Needs and a Central Bank Digital Currency', *Journal of Digital Banking* **8**(3), 242–255.

Hensher, D. A. and Greene, W. H. (2003), 'The mixed logit model: The state of practice', *Transportation* **30**(2), 133–176.

Huynh, K. P. (2017), How Canadians Pay for Things, Technical report, Bank Of Canada.

Huynh, K. P., Nicholls, G. and Nicholson, M. W. (2020), 2019 Cash Alternative Survey Results, Technical report, Bank Of Canada.

Jasiak, J., MacKenzie, P. and Tuvaandorj, P. (2024), 'Digital Divide: Empirical Study of CIUS 2020', *ArXiv preprint arXiv:2301.07855* .

Jasiak, J. and Tuvaandorj, P. (2023), 'Penalized Likelihood Inference with Survey Data', *ArXiv preprint https://arxiv.org/abs/2304.07855* .

Javanmard, A. and Montanari, A. (2014), 'Confidence Intervals and Hypothesis Testing for High-Dimensional Regression', *The Journal of Machine Learning Research* **15**(1), 2869–2909.

Jordan, B. (2019), High-Speed Access for All: Canada's Connectivity Strategy, Technical report, Government of Canada.

Kamdjoug, J. R. K., Wamba-Taguimdje, S.-L., Wamba, S. F. and Kake, I. B. (2021), 'Determining Factors and Impacts of the Intention to Adopt Mobile Banking App in Cameroon: Case of SARA by Afriland First Bank', *Journal of Retailing and Consumer Services* **61**, 102509.

Kitagawa, R., Kuroda, S., Okudaira, H. and Owan, H. (2021), 'Working from home and productivity under the covid-19 pandemic: Using survey data of four manufacturing firms', *PLoS One* **16**(12), e0261761.

Koch, K. (2022), 'The Territorial and Socio-Economic Characteristics of the Digital Divide in Canada', *Canadian Journal of Regional Science* **45**(2), 89–98.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016), 'Exact Post-Selection Inference, with Application to the Lasso', *The Annals of Statistics* **44**(3), 907–927.

Leung, D., Meh, C. and Terajima, Y. (2008), 'Productivity in Canada: Does Firm Size Matter?', *Bank of Canada Review* **2008**(Autumn), 7–16.

McFadden, D. and Train, K. (2000), 'Mixed mnl models for discrete response', *Journal of Applied Econometrics* **15**(5), 447–470.

Mullainathan, S. and Spiess, J. (2017), 'Machine Learning: An Applied Econometric Approach', *Journal of Economic Perspectives* **31**(2), 87–106.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), 'A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers', *Statistical Science* **27**(4), 538 – 557.
URL: *https://doi.org/10.1214/12-STS400*

Newey, W. K. and McFadden, D. (1994), Large Sample Estimation and Hypothesis Testing, *in* R. F. Engle and D. L. McFadden, eds, 'Handbook of Econometrics, Volume 4', Amsterdam, chapter 36, pp. 2111–2245.

Neyman, J. (1959), Optimal Asymptotic Tests of Composite Statistical Hypotheses, *in* U. Grenander, ed., 'Probability and Statistics, the Harald Cramér Volume', Almqvist and Wiksell, Uppsala, Sweden, pp. 213–234.

OECD (2017), *Enhancing the Role of Insurance in Cyber Risk Management*, OECD Publishing, Paris. Available at OECD iLibrary.
URL: *https://doi.org/10.1787/9789264282148-en*

Reddick, C. G., Enriquez, R., Harris, R. J. and Sharma, B. (2020), 'Determinants of Broadband Access and Affordability: An Analysis of a Community Survey on the Digital Divide', *Cities* **106**, 102904.

Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M. and Stern, M. J. (2015), 'Digital Inequalities and Why They Matter', *Information, Communication & Society* **18**(5), 569–582.

Smith, R. J. (1987), 'Alternative Asymptotically Optimal Tests and Their Application to Dynamic Specification', *The Review of Economic Studies* **LIV**, 665–680.

Taylor, J. and Tibshirani, R. (2018), 'Post-selection Inference for l1-Penalized Likelihood Models', *Canadian Journal of Statistics* **46**(1), 41–61.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Van Buuren, S. (2018), *Flexible Imputation of Missing Data, Second Edition*, Chapman and Hall.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014), 'On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models', *The Annals of Statistics* **42**(3), 1166 – 1202.
URL: *https://doi.org/10.1214/14-AOS1221*

Van Deursen, A. J. A. M. and Van Dijk, J. A. G. M. (2019), 'The First-Level Digital Divide Shifts from Inequalities in Physical Access to Inequalities in Material Access', *New Media & Society* **21**(2), 354–375.

Van Dijk, J. and Hacker, K. (2003), 'The Digital Divide as a Complex and Dynamic Phenomenon', *The Information Society* **19**(4), 315–326.

Vijverberg, W. (2007), 'Testing for IIA with the Hausman-McFadden Test', *IZA Discussion Papers* .

Wavrock, D., Schellenberg, G. and Schimmele, C. (2022), Canadians' Use of the Internet

and Digital Technologies Before and During the COVID-19 Pandemic, Technical report, Statistics Canada.

Wooldridge, J. M. (2001), 'Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples', *Econometric Theory* **17**(2), 451–470.

Xia, L., Nan, B. and Li, Y. (2020), 'A Revisit to De-biased Lasso for Generalized Linear Models', *arXiv preprint arXiv:2006.12778* .

Xia, L., Nan, B. and Li, Y. (2023), 'Debiased Lasso for Generalized Linear Models with a Diverging Number of Covariates', *Biometrics* **79**(1), 344–357.

Zhang, C.-H. and Zhang, S. S. (2014), 'Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 217–242.

Zickuhr, K. and Smith, A. (2012), 'Digital Differences', Pew Research Center.
URL: *https://www.pewresearch.org/internet/2012/04/13/digital-differences/*

Appendix A

Technical Appendix for Paper 1

A.1 Sampling and weighting methodology in CIUS 2020

Sampling. The collection of CIUS 2020 is based on a stratified design employing probability sampling. The stratification is done at the province/census metropolitan area (CMA) and census agglomeration (CA) level where each of the ten provinces were divided into strata/geographic areas.

CIUS 2020 uses a frame that combines landline and cellular telephone numbers from the Census and various administrative sources with Statistics Canada's dwelling frame. Records on the frame are groups of one or several telephone numbers associated with the same address.

Each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records (the groups of telephone numbers) was next selected in each stratum. CIUS 2020 only selects one respondent randomly from each eligible household to complete an electronic questionnaire or to respond to a telephone interview.

The number of respondents for the 2020 CIUS was 17,409, which is 41.6% of the sample size 41,817.

Weighting. Each record within a stratum has an equal probability of selection given by

$$\frac{\text{the number of records sampled in the stratum}}{\text{the number of records in the stratum from the survey frame}}.$$

A short description of the survey weights calculation is as follows.[1]

1. The initial weight is the inverse of an adjusted version of the probability of selection given above.

2. The person weight is equal to *Initial Household weight* $\times$ *Factor 1* $\times$ *Number of Eligible Household Members (capped at 5)*, where *Factor 1* involves an adjustment for non-response among others.

3. The final person weight w_i is an adjusted version of the person weight above.

A.1.1 Inference with survey logistic Lasso

Since CIUS 2020 data were collected using a stratified sampling scheme which is close to simple stratified sampling where the units within each stratum are sampled independently with equal probability, we treat w_i as constant and given (Wooldridge (2001), Section 3), and $\{(y_i, x_i')'\}_{i=1}^n$ as independent.

From the weighted log-likelihood function $L(\theta) = n^{-1}\sum_{i=1}^n w_i(y_i x_i'\theta - \log(1+\exp(x_i'\theta)))$, the score function, the information and negative Hessian matrices can be obtained respectively as

$$S(\theta) = \frac{\partial L(\theta)}{\partial \theta} = n^{-1}\sum_{i=1}^n w_i x_i(y_i - \Lambda(x_i'\theta)), \tag{A.1.1}$$

$$I(\theta) = n^{-1}\sum_{i=1}^n w_i^2 x_i x_i' \Lambda(x_i'\theta)(1-\Lambda(x_i'\theta)), \tag{A.1.2}$$

$$H(\theta) = -\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'} = n^{-1}\sum_{i=1}^n w_i x_i x_i' \Lambda(x_i'\theta)(1-\Lambda(x_i'\theta)), \tag{A.1.3}$$

[1]Further details of the weighting procedure can be found in Section 10 of Microdata User Guide, CIUS 2020 at https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4432#a2

where $\Lambda(z) = \exp(z)/(1+\exp(z))$ is the logistic CDF. We report the marginal effects for each variable along with the coefficient estimates in the regression tables which is defined as follows. For a dummy regressor $\tilde{x}_{ij}, j = 1, \ldots, p; i = 1, \ldots, n$, the marginal effect (ME) is $\text{ME}_{ij}(\theta) \equiv \Lambda(x_i'\theta)|_{\tilde{x}_{ij}=1} - \Lambda(x_i'\theta)|_{\tilde{x}_{ij}=0}$. The average marginal effect (AME) of the j-th regressor is defined as

$$\text{AME}_j = \text{AME}_j(\theta_0) \equiv \text{E}\left[\frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i \text{ME}_{ij}(\theta_0)\right],$$

where θ_0 denotes the true value of θ and the expectation is taken with respect to the distribution of the regressors. An estimator of AME_j is

$$\widehat{\text{AME}}_j(\hat{\theta}) \equiv \frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i \left(\Lambda(x_i'\hat{\theta})|_{\tilde{x}_{ij}=1} - \Lambda(x_i'\hat{\theta})|_{\tilde{x}_{ij}=0}\right),$$

where $\hat{\theta} = (\hat{\alpha}, \hat{\beta}')'$ is an estimator of θ_0 e.g. `svy LLasso` estimator. Note also that

$$\frac{\partial\widehat{\text{AME}}_j(\hat{\theta})}{\partial\theta} \equiv \frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i \left\{\left[x_i\Lambda(x_i'\hat{\theta})(1-\Lambda(x_i'\hat{\theta}))\right]|_{\tilde{x}_{ij}=1} - \left[x_i\Lambda(x_i'\hat{\theta})(1-\Lambda(x_i'\hat{\theta}))\right]|_{\tilde{x}_{ij}=0}\right\}.$$

Debiased Lasso

The debiased Lasso method of Zhang and Zhang (2014), Javanmard and Montanari (2014) and Xia et al. (2020) is based the one-step estimator constructed from the initial Lasso estimator $\hat{\theta}$:

$$\tilde{\theta} \equiv \hat{\theta} + H(\hat{\theta})^{-1}S(\hat{\theta}).$$

The standard errors for the parameters are calculated using the distributional approximation: as $n \to \infty$

$$(\tau' H(\hat{\theta})^{-1} I(\hat{\theta}) H(\hat{\theta})^{-1}\tau)^{-1/2} n^{1/2}\tau'(\tilde{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\tau \in \mathbb{R}^{p+1}$ is a fixed vector with $\tau'\tau = 1$, and θ_0 denotes the true parameter vector. To obtain a confidence interval for $\text{AME}_j, j = 2, \ldots, p+1$, define a one-step estimator

$$\widetilde{\text{AME}}_j \equiv \widehat{\text{AME}}_j(\hat{\theta}) + \frac{\partial \widehat{\text{AME}}_j(\hat{\theta})}{\partial \theta'} H(\hat{\theta})^{-1} S(\hat{\theta}).$$

Then, under some regularity conditions as $n \to \infty$

$$\left(\frac{\partial \widehat{\text{AME}}_j(\hat{\theta})}{\partial \theta'} H(\hat{\theta})^{-1} I(\hat{\theta}) H(\hat{\theta})^{-1} \frac{\partial \widehat{\text{AME}}_j(\hat{\theta})}{\partial \theta} \right)^{1/2} n^{1/2}(\widetilde{\text{AME}}_j - \text{AME}_j) \xrightarrow{d} \mathcal{N}(0,1).$$

$C(\alpha)$/Orthogonalization method

We follow Belloni et al. (2016) who develop a $C(\alpha)$-type subvector inference procedure in a sparse high-dimensional generalized linear model by constructing an estimating equation orthogonalized against the direction of the nuisance parameter estimation (see Neyman (1959) for the $C(\alpha)$ test), and consider a survey version of their statistic.

Consider testing a scalar component θ_1 (e.g. the i-th element β_i of β) of θ. Partition the parameters as $\theta = (\theta_1, \theta_2')'$, $\theta_1 \in \mathbb{R}$, $\theta_2 \in \mathbb{R}^p$. Also partition the quantities in (A.1.1) and (B.1.4) as

$$S(\theta) = [S_1(\theta)', S_2(\theta)']', \; S_1(\theta) \in \mathbb{R}, \; S_1(\theta) \in \mathbb{R}^p,$$
$$H(\theta) = \begin{bmatrix} H_{11}(\theta) & H_{12}(\theta) \\ H_{21}(\theta) & H_{22}(\theta) \end{bmatrix}, \; H_{11}(\theta) \in \mathbb{R}, \; H_{21}(\theta) = H_{12}(\theta)' \in \mathbb{R}^p, \; H_{22}(\theta) \in \mathbb{R}^{p \times p}.$$

Consider the restriction $H_0 : \theta_1 = \theta_{01}$ and let $\tilde{\theta}^* = (\theta_{01}', \tilde{\theta}_2^{*\prime})'$, where $\tilde{\theta}_2^*$ is the logistic Lasso or Post-logistic Lasso (Belloni et al., 2016) estimator of θ_2. The survey $C(\alpha)$ statistic is then defined as

$$C_\alpha(\theta_{01}) \equiv n \, S(\tilde{\theta}^*)' D(\tilde{\theta}^*) \left(D(\tilde{\theta}^*)' I(\tilde{\theta}^*) D(\tilde{\theta}^*) \right)^{-1} D(\tilde{\theta}^*)' S(\tilde{\theta}^*), \tag{A.1.4}$$

where $D(\theta) \equiv [I_{k_1}, -H_{22}(\theta)^{-1}H_{21}(\theta)]'$. Here $D(\tilde{\theta}^*)'S(\tilde{\theta}^*) = S_1(\tilde{\theta}^*) - H_{12}(\tilde{\theta}^*)H_{22}(\tilde{\theta}^{**})^{-1}S_2(\tilde{\theta}^*)$ is the effective score function obtained by orthogonalizing the score function of the parameters of interest against the score function of the nuisance parameters. Under $H_0 : \theta_1 = \theta_{01}$ and appropriate regularity conditions $C_\alpha(\theta_{01}) \xrightarrow{d} \chi^2_1$ as $n \to \infty$.

For testing the restriction $H_0 : \psi(\theta_0) = 0$ on a scalar nonlinear parameter $\psi(\theta)$, following Dufour et al. (2016) and Smith (1987) consider the $C(\alpha)$ statistic:

$$C_\alpha(\psi_0) \equiv n \left(\frac{\partial \psi(\tilde{\theta}^*)}{\partial \theta'} H(\tilde{\theta}^*)^{-1} I(\tilde{\theta}^*) H(\tilde{\theta}^*)^{-1} \frac{\partial \psi(\tilde{\theta}^*)}{\partial \theta} \right)^{-1} \left(\frac{\partial \psi(\tilde{\theta}^*)}{\partial \theta'} H(\tilde{\theta}^*)^{-1} S(\tilde{\theta}^*) \right)^2, \quad \text{(A.1.5)}$$

where the auxiliary estimate $\tilde{\theta}^*$ satisfies $\psi(\tilde{\theta}^*) = 0$. Let $\text{AME}_{(1)}$ be the AME with respect to a dummy regressor with a coefficient θ_1. Then, $\text{AME}_{(1)} = 0$ if $\theta_1 = 0$, and it follows that $\text{AME}_{(1)}(\tilde{\theta}^*) = 0$ for $\tilde{\theta}^* = (0, \hat{\theta}_2')'$, where $\hat{\theta}_2$ is the `svy LLasso` estimate. It is easy to see that the statistic in (A.1.5) for the hypothesis $H_0 : \text{AME}_{(1)} = 0$ is numerically identical to the statistic in (A.1.4) for the hypothesis $H_0 : \theta_1 = 0$.

Selective inference

We also consider the survey-logit version of the "selective inference" method proposed by Lee et al. (2016) and Taylor and Tibshirani (2018). This method makes inference on the coefficients selected by the Lasso i.e. the target parameters determined from the data which are random. This feature makes the selective inference method conceptually different from the debiased Lasso and $C(\alpha)$ methods where the target parameters are the population parameters.

The key ingredient in this method is the one-step estimator which updates the estimates of the (non-zero) coefficients selected by the survey logistic Lasso, denoted as $\hat{\theta}_M$:

$$\tilde{\theta}_M \equiv \hat{\theta}_M + H_M(\hat{\theta}_M)^{-1} S_M(\hat{\theta}_M), \quad \text{(A.1.6)}$$

where $H_M(\cdot)$ and $S_M(\cdot)$ are the Hessian and the score functions of the logistic model cor-

responding to `svy LLasso` selected coefficients. Then, a test statistic constructed from the conditional distribution of the one-step estimator (A.1.6) given Lasso selection events is used to test a hypothesis on the `svy LLasso` selected coefficients. We refer to Jasiak and Tuvaandorj (2023) for further details of the method in a survey setting.

A.1.2 Multiple correspondence analysis

The Multiple Correspondence Analysis (MCA) is an analog of principal component analysis (PCA) for multiple categorical variables. MCA may provide a useful summary and visualization of survey data (with categorical variables) by revealing the variables that contribute the most to the variation in the data, identifying a set of observations with similar characteristics in their survey response and quantifying the degree of associations between different categories.

The MCA process works by taking J categorical variables, each having K_j categorical levels with the sum of these levels being equal to K. Given I observations we denote the indicator matrix as X. This indicator matrix is used to perform correspondence analysis. The correspondence analysis gives two different sets of factor scores, one set for the rows of the matrix and the other for the columns.

We denote the total of this table as N and set $Z \equiv X/N$. The vector r contains the sums of the rows of the matrix Z and the vector c the sums of the columns of Z. To compute the MCA, we need to define diagonal matrices $D_r \equiv \text{diag}(r)$ and $D_c \equiv \text{diag}(c)$. To find the factor scores we use the singular value decomposition

$$D_r^{-1/2}[Z - rc']D_c^{-1/2} = P\Delta Q'.$$

The matrix Δ is the diagonal matrix of singular values and the matrix of eigenvalues is

$\Lambda = \Delta^2$. We find the row and column factor scores as

$$F = D_r^{-1/2} P\Delta \quad \text{and} \quad G = D_c^{-1/2} Q\Delta.$$

In this paper, the MCA is done using what is called the Burt matrix defined as $B = X'X$. Using the Burt matrix gives the exact same factors as is the case with the indicator matrix X but also gives a better approximation of the captured inertia.

Various plots accompanying the MCA can be used to visualize a global pattern within the data. The coordinate plots which represent the variable categories in two dimensional space are provided in Figures 2.1 and 2.2.

A.2 Details on the digital literacy score

The calculation of the digital literacy score is based on the responses to the following 10 questions all of which have the following answers: *Yes, No, Valid skip, Don't know, Refusal, Not stated*. The questions 1-7 are "During the past three months, which of the following activities, related to communication, have you done over the Internet?" followed by:

1. "Have you used social networking websites or apps?"
2. "Have you made online voice calls or video calls?"
3. "Have you researched for information about community events?"
4. "Have you accessed the news?"
5. "Have you found locations and directions?"
6. "Have you researched for information on health?"
7. "Have you researched for information about goods or services?"

The remaining questions 8-10 are:

8. "During the past 12 months, how did you pay for the goods and services ordered over the Internet? Did you use an online payment service?"

9. "During the past 12 months, which of the following software related activities have you carried out using any device? Have you copied or moved files or folders?"

10. "Have you carried out any of the following to manage access to your personal data over the Internet during the past 12 months? Have you checked that the website where you provided personal data was secure e.g., https sites, safety logo or certificate?"

11,874 out of 12,431 possible respondents answered all the relevant questions and the remaining respondents had at least one question unanswered. Figure A.1 plots the weighted histogram and Table A.1 below reports the weighted descriptive statistics of the scores of 11,874 respondents who answered all the relevant questions. The results for the distribution of the digital literacy score from other samples, where *Not stated* answers were replaced by multiple imputation methods following Van Buuren (2018), were consistent with Table A.1 and Figure A.1 (see also footnote 1).

Table A.1: Descriptive Statistics for Digital Literacy Scores

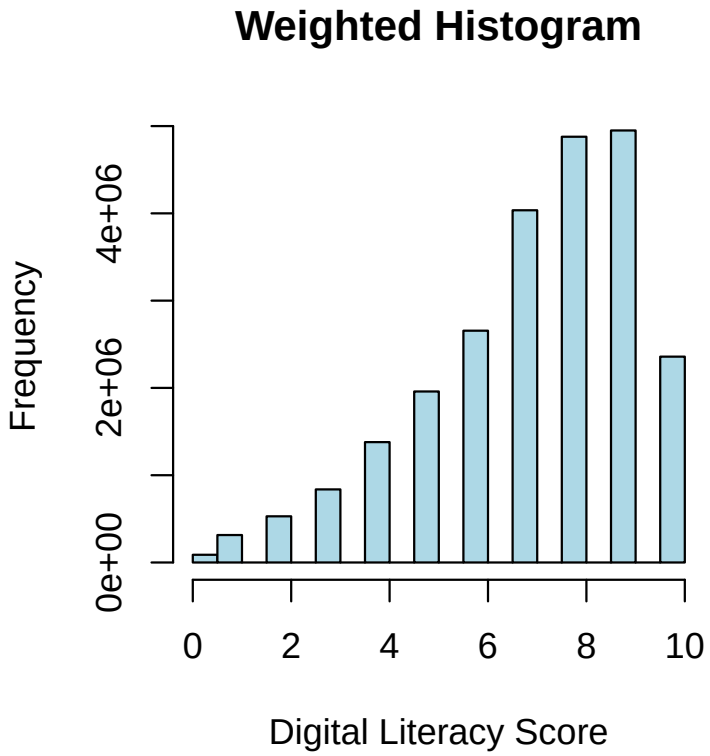
Weighted estimates						
Mean	Stdev	Skewness	Kurtosis	1st Quartile	Median	3rd Quartile
7.11	2.15	-0.85	0.26	6.00	8.00	9.00

Note: The table reports the weighted descriptive statistics of the scores of 11,874 respondents who answered all the relevant questions.

A.3 $C(\alpha)$ and selective inference results

This section presents the outcomes of the $C(\alpha)$ and SI for the logit coefficients and their AMEs for models 1 to 5. The results are very similar to the results reported in Section 2.4.

Figure A.1: Weighted Histogram of Digital Literacy Scores



Note: This figure plots the weighted histogram of the digital scores of 11874 respondents who answered all the relevant questions.

Therefore, these additional results further corroborate the findings from the `svy LLasso` and debiased Lasso results.

Since the SI is made only on the coefficients chosen by `svy LLasso`, it tends to have less coefficient and AME estimates that are significant based on the p-values than in the case of the debiased Lasso results. Besides the small differences in significance, the selective inference and debiased Lasso results were relatively consistent. The C_α test statistics and corresponding p-values are also consistent with the `svy LLasso` and debiased Lasso results.

Table A.2: Lasso Logistic Regression Results for Internet Use Dependent Variable

Variables	Categories	C_α	p-value	$\tilde{\theta}^{SI}$	p-value	$\widetilde{\text{AME}}^{SI}$	p-value
Intercept		80.184***	0.000	2.930***	0.000	–	–
Location	Rural	12.198***	0.000	−0.307	0.998	−0.018	0.996
Age	15–24	13.647***	0.000	1.207	0.140	0.056	0.123
	25–34	6.954**	0.008	0.67	0.326	0.036	0.323
	35–44	5.687*	0.017	0.555	0.628	0.032	0.627
	55–64	19.400***	0.000	−0.543	0.999	−0.032	0.996
	65 and older	176.970***	0.000	−1.296	1.000	−0.081	1.000
Gender	Female	1.620	0.203	0.093	0.756	0.005	0.757
Aboriginal	Aboriginal	5.304*	0.021	–	–	–	–
Language	English	3.628•	0.057	0.427**	0.002	0.025**	0.003
	French	0.609	0.435	–	–	–	–
	Non-official	0.043	0.836	–	–	–	–
	English and French	2.361	0.124	–	–	–	–
	French and Non-official	0.367	0.544	–	–	–	–
	English, French and Non-official	1.692	0.193	–	–	–	–
Employment	Employed	21.370***	0.000	0.609***	0.000	0.034***	0.000
Education	High school or less	189.022***	0.000	−0.888	1.000	−0.053	1.000
	University degree	12.502***	0.000	0.624***	0.000	0.032***	0.000
Minority	Visible minority	4.379*	0.036	−0.275	0.520	−0.016	0.515
Household type	Family w/o children under 18	0.026	0.872	–	–	–	–
	Single	18.073***	0.000	−0.665	1.000	−0.043	1.000
	Other household type	0.226	0.635	–	–	–	–
Income	$52,203 and lower	30.504***	0.000	−0.526	1.000	−0.031	1.000
	$92,486–$146,559	0.524	0.469	–	–	–	–
	$146,560 and higher	9.420**	0.002	0.491**	0.004	0.025**	0.003
Immigration	Non-landed immigrant	1.829	0.176	–	–	–	–
Province	NL	2.540	0.111	–	–	–	–
	PEI	2.023	0.155	–	–	–	–
	NS	2.383	0.123	–	–	–	–
	NB	0.297	0.586	–	–	–	–
	QC	5.130*	0.0240	−0.201	0.859	−0.011	0.832
	ON	0.017	0.895	0.183	0.469	0.011	0.470
	MB	6.162*	0.013	–	–	–	–
	SK	4.347*	0.037	–	–	–	–
	BC	0.250	0.617	0.264	0.604	0.015	0.604

Note: $n = 17,409$. C_α denotes the $C(\alpha)$ statistic which tests simultaneously the statistical significance of the coefficient and of its AME. $\tilde{\theta}^{SI}$ and $\widetilde{\text{AME}}^{SI}$ denote the selective inference one-step Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, • $p < 0.1$.

Table A.3: Lasso Logistic Regression Results for Online Banking Dependent Variable

Variables	Categories	C_α	p-value	$\tilde{\theta}^{SI}$	p-value	$\widetilde{\text{AME}}^{SI}$	p-value
Intercept		10.569**	0.001	1.049***	0.000	–	–
Location	Rural	2.036	0.154	–	–	–	–
Age	15–24	0.127	0.721	–	–	–	–
	25–34	23.173***	0.000	0.639***	0.000	0.094***	0.000
	35–44	22.858***	0.000	0.513***	0.000	0.078***	0.000
	55–64	12.714***	0.000	−0.331	0.978	−0.054	0.977
	65 and older	97.324***	0.000	−0.897	1.000	−0.161	1.000
Gender	Female	2.592	0.107	–	–	–	–
Aboriginal	Aboriginal	2.631	0.105	–	–	–	–
Language	English	7.723**	0.005	0.083	0.794	0.013	0.795
	French	8.042**	0.005	–	–	–	–
	Non-official	2.883•	0.090	–	–	–	–
	English and French	0.402	0.526	–	–	–	–
	French and Non-official	0.208	0.648	–	–	–	–
	English, French and Non-official	0.032	0.858	–	–	–	–
Employment	Employed	70.108***	0.000	0.659***	0.000	0.108***	0.000
Education	High school or less	129.578***	0.000	−0.668	1.000	−0.112	1.000
	University degree	24.964***	0.000	0.449***	0.000	0.069***	0.000
Minority	Visible minority	8.596**	0.003	−0.352	0.999	−0.058	0.999
Household type	Family w/o children under 18	11.558***	0.001	0.259*	0.033	0.042*	0.035
	Single	2.252	0.133	−0.191	0.972	−0.031	0.964
	Other household type	4.138*	0.042	–	–	–	–
Income	$52,203 and lower	11.843***	0.001	−0.280	1.000	−0.046	1.000
	$92,486–$146,559	2.270	0.132	–	–	–	–
	$146560 and higher	7.792**	0.005	0.184•	0.073	0.029•	0.075
Immigration	Non-landed immigrant	0.538	0.463	–	–	–	–
Province	NL	0.014	0.905	–	–	–	–
	PEI	0.270	0.604	–	–	–	–
	NS	0.376	0.540	–	–	–	–
	NB	0.270	0.603	–	–	–	–
	QC	0.547	0.460	–	–	–	–
	ON	0.103	0.748	–	–	–	–
	MB	8.498**	0.004	–	–	–	–
	SK	0.780	0.377	–	–	–	–
	BC	0.008	0.931	–	–	–	–
	AB (omitted)	–	–	–	–	–	–

Note: $n = 17,135$. C_α denotes the $C(\alpha)$ statistic which tests simultaneously the statistical significance of the coefficient and of its AME. $\tilde{\theta}^{SI}$ and $\widetilde{\text{AME}}^{SI}$ denote the selective inference one-step Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, • $p < 0.1$.

Table A.4: Lasso Logistic Regression Results for Email Use Dependent Variable

Variables	Categories	C_α	p-value	$\tilde{\theta}^{SI}$	p-value	$\widetilde{\text{AME}}^{SI}$	p-value
Intercept		50.422***	0.000	1.722***	0.000	–	–
Location	Rural	8.218**	0.004	−0.246	0.992	−0.025	0.990
Age	15–24	12.364***	0.000	0.664**	0.001	0.060**	0.001
	25–34	17.75***	0.000	0.737***	0.000	0.065***	0.000
	35–44	14.221***	0.000	0.589***	0.001	0.054***	0.001
	55–64	11.781***	0.001	−0.367	0.999	−0.037	0.997
	65 and older	94.367***	0.000	−0.929	1.000	−0.104	0.999
Gender	Female	5.024*	0.025	0.152•	0.061	0.015•	0.064
Aboriginal	Aboriginal	6.942**	0.008	–	–	–	–
Language	English	1.278	0.258	0.337**	0.002	0.033**	0.003
	French	0.219	0.640	–	–	–	–
	Non-official	0.881	0.348	−0.188	0.449	−0.019	0.447
	English and French	0.959	0.327	–	–	–	–
	French and Non-official	0.183	0.669	–	–	–	–
	English, French and Non-official	5.507*	0.019	–	–	–	–
Employment	Employed	23.347***	0.000	0.465**	0.002	0.046**	0.002
Education	High school or less	184.115***	0.000	−0.816	1.000	−0.085	1.000
	University degree	49.181***	0.000	0.873***	0.000	0.076***	0.000
Minority	Visible minority	7.085**	0.008	−0.322	0.977	−0.033	0.974
Household type	Family w/o children under 18	0.213	0.644	–	–	–	–
	Single	24.825***	0.000	−0.533	1.000	−0.058	1.000
	Other household type	0.049	0.824	–	–	–	–
Income	$52,203 and lower	17.168***	0.000	−0.381	1.000	−0.039	1.000
	$92,486–$146,559	0.800	0.371	–	–	–	–
	$146,560 and higher	11.925***	0.001	0.402***	0.001	0.037***	0.000
Immigration	Non-landed immigrant	1.044	0.307	0.122	0.842	0.012	0.844
Province	NL	2.420	0.120	–	–	–	–
	PEI	1.244	0.265	–	–	–	–
	NS	6.260*	0.012	–	–	–	–
	NB	2.730•	0.098	–	–	–	–
	QC	4.091*	0.043	−0.192	0.793	−0.019	0.779
	ON	0.191	0.662	0.278**	0.006	0.027**	0.006
	MB	8.440**	0.004	–	–	–	–
	SK	5.343*	0.021	–	–	–	–
	BC	2.417	0.120	0.439**	0.002	0.041**	0.002

Note: $n = 17,268$. C_α denotes the $C(\alpha)$ statistic which tests simultaneously the statistical significance of the coefficient and of its AME. $\tilde{\theta}^{SI}$ and $\widetilde{\text{AME}}^{SI}$ denote the selective inference one-step Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, • $p < 0.1$.

Table A.5: Lasso Logistic Regression Results for Virtual Wallet Dependent Variable

Variables	Categories	C_α	p-value	$\tilde{\theta}^{SI}$	p-value	$\widetilde{\text{AME}}^{SI}$	p-value
Intercept		70.764***	0.000	−2.222***	0.000	–	–
Location	Rural	23.924***	0.000	−0.627	1.000	−0.061	1.000
Age	15–24	29.036***	0.000	0.702**	0.002	0.083**	0.003
	25–34	25.121***	0.000	0.504***	0.001	0.057**	0.001
	35–44	7.735**	0.005	–	–	–	–
	55-64	14.361***	0.000	−0.747	1.000	−0.070	1.000
	65 and older	21.482***	0.000	−1.144	1.000	−0.096	1.000
Gender	Female	1.165	0.280	–	–	–	–
Aboriginal	Aboriginal	0.026	0.872	–	–	–	–
Language	English	0.281	0.596	–	–	–	–
	French	0.203	0.653	–	–	–	–
	Non-official	2.469	0.116	–	–	–	–
	English and French	0.047	0.829	–	–	–	–
	French and Non-official	0.574	0.448	–	–	–	–
	English, French, and Non-official	0.842	0.359	–	–	–	–
Employment	Employed	0.034	0.853	–	–	–	–
Education	High school or less	0.326	0.568	–	–	–	–
	University degree	6.847**	0.009	0.254	0.391	0.027	0.393
Minority	Visible minority	13.365***	0.000	0.242*	0.044	0.026•	0.053
Household type	Family w/o children under 18	0.362	0.547	–	–	–	–
	Single	0.066	0.797	–	–	–	–
	Other household type	0.244	0.621	–	–	–	–
Income	$52,203 and lower	0.356	0.551	–	–	–	–
	$92,486–$146,559	1.620	0.203	–	–	–	–
	$146,560 and higher	24.149***	0.000	0.506***	0.000	0.057***	0.000
Immigration	Non-landed immigrant	0.796	0.372	–	–	–	–
Province	NL	1.811	0.178	–	–	–	–
	PEI	2.279	0.131	–	–	–	–
	NS	1.947	0.163	–	–	–	–
	NB	0.095	0.757	–	–	–	–
	QC	0.39	0.532	–	–	–	–
	ON	0.094	0.759	–	–	–	–
	MB	4.605	0.032	–	–	–	–
	SK	1.219	0.270	–	–	–	–
	BC	0.224	0.636	–	–	–	–

Note: $n = 12,124$. C_α denotes the $C(\alpha)$ statistic which tests simultaneously the statistical significance of the coefficient and of its AME. $\tilde{\theta}^{SI}$ and $\widetilde{\text{AME}}^{SI}$ denote the selective inference one-step Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, • $p < 0.1$.

Table A.6: Lasso Logistic Regression Results for Credit Card Use Dependent Variable

Variables	Categories	C_α	p-value	$\tilde{\theta}^{SI}$	p-value	$\widetilde{\text{AME}}^{SI}$	p-value
Intercept		21.706***	0.000	1.193***	0.000	–	–
Location	Rural	2.250	0.134	–	–	–	–
Age	15–24	16.767***	0.000	−0.572	1.000	−0.098	1.000
	25–34	0.233	0.630	–	–	–	–
	35–44	1.735	0.188	–	–	–	–
	55–64	0.046	0.830	–	–	–	–
	65 and older	0.202	0.653	–	–	–	–
Gender	Female	0.003	0.958	–	–	–	–
Aboriginal	Aboriginal	1.046	0.306	–	–	–	–
Language	English	0.001	0.980	0.278*	0.011	0.044*	0.014
	French	8.287**	0.004	−0.408	0.972	−0.067	0.967
	Non-official	0.039	0.844	–	–	–	–
	English and French	0.211	0.646	–	–	–	–
	French and Non-official	1.252	0.263		–	–	–
	English, French, and Non-official	2.916•	0.088	–	–	–	–
Employment	Employed	3.006•	0.083	0.151	0.934	0.024	0.934
Education	High school or less	33.289***	0.000	−0.470	1.000	−0.077	1.000
	University degree	29.228***	0.000	0.480***	0.000	0.072***	0.000
Minority	Visible minority	4.063*	0.044	–	–	–	–
Household type	Family w/o children under 18	13.824***	0.000	0.219	0.542	0.035	0.543
	Single	9.465**	0.002	–	–	–	–
	Other household type	0.623	0.430		–	–	–
Income	$52,203 and lower	8.466**	0.004	−0.291	0.976	−0.047	0.975
	$92,486–$146,559	1.049	0.306	–	–	–	–
	$146,560 and higher	0.728	0.393	–	–	–	–
Immigration	Non-landed immigrant	1.430	0.232		–	–	–
Province	NL	3.045•	0.081	–	–	–	–
	PEI	0.223	0.637	–	–	–	–
	NS	0.076	0.783	–	–	–	–
	NB	0.005	0.944	–	–	–	–
	QC	0.076	0.783	−0.119	0.633	−0.019	0.620
	ON	4.051*	0.044	0.155	0.456	0.024	0.457
	MB	0.046	0.829	–	–	–	–
	SK	0.019	0.891	–	–	–	–
	BC	2.277	0.131	–	–	–	–

Note: $n = 12,124$. C_α denotes the $C(\alpha)$ statistic which tests simultaneously the statistical significance of the coefficient and of its AME. $\tilde{\theta}^{SI}$ and $\widetilde{\text{AME}}^{SI}$ denote the selective inference one-step Lasso estimates of the logit parameter and AME respectively. " – " denotes the variables not selected by `svy LLasso`. Comparison categories and variables with *Not stated* answers are not displayed for clarity and interpretability. Significance levels are marked as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, • $p < 0.1$.

Appendix B

Technical Appendix for Paper 2

B.1 Additional details on the implementation.

Tuning Parameter Selection. In the empirical analysis and the small Monte Carlo simulations below, the logit Lasso model `svyLLasso` is fitted using the `R` package `glmnet`. For the tuning parameter λ, we use the package's default value, determined by 10-fold cross-validation with the loss function `auc` (area under the ROC curve).

Average marginal effect in the logit model. We report the marginal effects (ME) for each variable along with the coefficient estimates for logit models in the tables. Let $\Lambda(z) = \exp(z)/(1+\exp(z))$ be the logistic distribution function. For a dummy regressor $\tilde{x}_{ij}$, where $j = 1, \ldots, p$ and $i = 1, \ldots, n$, the ME is defined as:

$$\mathrm{ME}_{ij}(\theta) \equiv \Lambda(x_i'\theta)|_{\tilde{x}_{ij}=1} - \Lambda(x_i'\theta)|_{\tilde{x}_{ij}=0}.$$

Given the survey weights $\{w_i\}_{i=1}^n$ corresponding to the observations $\{(y_i, x_i')'\}_{i=1}^n$, the AME of the j-th regressor is given by:

$$\mathrm{AME}_j = \mathrm{AME}_j(\theta_0) \equiv \mathrm{E}\left[\frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i \mathrm{ME}_{ij}(\theta_0)\right],$$

where θ_0 is the true parameter value, and the expectation is with respect to the regressors' distribution. An estimator for AME_j is:

$$\widehat{\text{AME}}_j(\hat{\theta}) \equiv \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \left(\Lambda(x_i'\hat{\theta})|_{\tilde{x}_{ij}=1} - \Lambda(x_i'\hat{\theta})|_{\tilde{x}_{ij}=0} \right),$$

where $\hat{\theta} = (\hat{\alpha}, \hat{\beta}')'$ is the estimated parameter vector, e.g., the `svy LLasso` estimator. The debiased logit Lasso estimator of AME_j is then constructed using the one-step iteration provided in the equation (3.3.1).

B.1.1 Post-Selection Inference for Survey-GLM

Consider the density of a scalar outcome variable y_i given a $(p+1) \times 1$ vector of covariates x_i (which includes a constant) specified as

$$f(y_i|x_i, \theta_0) = \exp(y_i x_i'\theta_0 - a(x_i'\theta_0))c(y_i), \quad i = 1, \ldots, n,$$

where θ_0 is the true value of the parameter vector $\theta \in \mathbb{R}^{p+1}$, and $a(\cdot)$ and $c(\cdot)$ are known functions. The combined SDTIU and CSCSC data were collected using a stratified sampling scheme, wherein units within each stratum are sampled independently with equal probability. In line with this, we treat w_i as fixed (see Wooldridge, 2001), and assume $\{(y_i, x_i')'\}_{i=1}^n$ to be independent.

Let $g(y, x'\theta) \equiv -\log f(y, x'\theta)$ and define the weighted log-likelihood function as:

$$L_n(\theta) \equiv -n^{-1} \sum_{i=1}^n w_i g(y_i, x_i'\theta). \tag{B.1.1}$$

The score function, the sample information and negative Hessian matrices corresponding to

(B.1.1) are defined as

$$S(\theta) \equiv \frac{\partial L_n(\theta)}{\partial \theta} = -n^{-1}\sum_{i=1}^{n} w_i x_i \dot{g}(y_i, x_i'\theta), \quad \dot{g}(y,t) \equiv \frac{\partial g(y,t)}{\partial t}, \tag{B.1.2}$$

$$\hat{I}(\theta) \equiv n^{-1}\sum_{i=1}^{n} w_i^2 x_i x_i' \dot{g}(y_i, x_i'\theta)^2, \tag{B.1.3}$$

$$\hat{H}(\theta) \equiv -\frac{\partial^2 L_n(\theta)}{\partial \theta \partial \theta'} = n^{-1}\sum_{i=1}^{n} w_i x_i x_i' \ddot{g}(y_i, x_i'\theta), \quad \ddot{g}(y,t) \equiv \frac{\partial^2 g(y,t)}{\partial t^2}. \tag{B.1.4}$$

Moreover, we define $H(\theta_0) \equiv \mathrm{E}[\hat{H}(\theta_0)]$ and $I(\theta_0) \equiv \mathrm{E}[\hat{I}(\theta_0)]$.

We will using the following notations in the assumptions and the proof of Proposition 3.3.1 below. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and the largest eigenvalue of a symmetric matrix A, respectively. For a real matrix $A = (a_{ij})$, let $\|A\|_\infty \equiv \max_{i,j}|a_{ij}|$, and $\|A\| = \sqrt{\mathrm{tr}(A'A)}$ and $\|A\|_2 = \sqrt{\lambda_{\max}(A'A)}$ denote its Frobenius and spectral norms, respectively. The sub-Gaussian norm of a random variable X is defined as $\|X\|_{\psi_2} \equiv \sup_{m\geq 1} m^{-1/2}(\mathrm{E}[|X|^m])^{1/m}$. The sub-Gaussian norm for the random vector is defined as $\|X\|_{\psi_2} \equiv \sup_{\|b\|=1}\|X'b\|_{\psi_2}$.

We establish the asymptotic validity of the debiasing method under the following assumptions imposed directly on the negative log-density function $g(y,t)$ which are similar to the assumptions employed in Xia et al. (2023). Let $X = [x_1, \ldots, x_n]'$.

Assumption 1 (Asymptotic validity)

(a) $\{(y_i, x_i')'\}_{i=1}^n$ *are independent with* $\max_{1\leq i\leq n} a_i < C_u < \infty$ *a.s. where*

$$a_i \in \{\|x_i\|_{\psi_2}, \|x_i\|_\infty, \|X\theta_0\|_\infty\}.$$

Moreover, w_i *is non-random with* $0 < C_l < w_i < C_u$ *for all* n, i.

(b) For $A \in \{H(\theta_0), I(\theta_0), \mathrm{E}[n^{-1}X'X]\}$, *there exist positive constants* λ_l *and* λ_u *such that* $0 < \lambda_l \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \lambda_u < \infty$.

(c) The function $g(y,t) \equiv a(t) - yt - \log c(y)$ *is convex in* $t \in \mathbb{R}$ *for all* y*, and twice differentiable with respect to* t *for all* (y,t)*. There exist a positive definite matrix* H *and* $\eta > 0$ *such that* $\lambda_{\min}(H) > \lambda_l > 0$ *and*

$$n^{-1}\sum_{i=1}^{n} \mathrm{E}[w_i(g(y_i, x_i'\theta) - g(y_i, x_i'\theta_0))] \geq \|H^{1/2}(\theta - \theta_0)\|^2 \tag{B.1.5}$$

for all $\|X(\theta - \theta_0)\|_\infty < \eta$*. Furthermore,* $\ddot{g}(y,t)$ *is Lipschitz with some constant* $L_0 > 0$*:*

$$\max_{t_0 \in \{x_i'\theta_0\}} \sup_{\max(|t-t_0|, |\tilde{t}-t_0|) \leq \eta} \sup_{y \in \mathcal{Y}} \frac{|\ddot{g}(y,t) - \ddot{g}(y,\tilde{t})|}{|t - \tilde{t}|} \leq L_0, \tag{B.1.6}$$

and

$$\max_{t_0 \in \{x_i'\theta_0\}} \sup_{y \in \mathcal{Y}} |\dot{g}(y, t_0)| \leq C_u, \tag{B.1.7}$$

$$\max_{t_0 \in \{x_i'\theta_0\}} \sup_{|t-t_0| \leq \eta} \sup_{y \in \mathcal{Y}} |\ddot{g}(y,t)| \leq C_u. \tag{B.1.8}$$

For discussions of these assumptions, we refer to Jasiak and Tuvaandorj (2023).

Simulations. We conduct a simulation experiment to verify the robustness of the debiased logit Lasso inference. We first generate $N = 10,000$ draws from a standard logit model as follows:

$$y_i \sim \text{Bernoulli}(\pi_i), \tag{B.1.9}$$

where $\theta_0 = (1, 1, 1, 0_{1\times(p-2)})'$, $\tilde{x}_{ij} \sim$ i.i.d. Bernoulli(0.5) for $j = 1, \ldots, p$ and $i = 1, \ldots, N$, $x_i = (1, \tilde{x}_i')'$, and $\pi_i = x_i'\theta_0$.

The population is then stratified into four strata of sizes 1,000, 2,000, 3,000, and 4,000. From each stratum, we draw 50 and 100 observations with replacement, yielding stratified samples of size $n = 200$ and $n = 400$, respectively. Observation weights are $w_i = 0.1, 0.2, 0.3, 0.4$, corresponding to the four strata. To evaluate the impact of regressor di-

mensionality, we set p such that $\frac{p}{n} \in \{0.01, 0.025, 0.05, 0.1, 0.25, 0.5\}$ for each $n \in \{200, 400\}$. The true AME for $\theta_{(2)} = \beta_1$ is 0.11.

We assess the empirical size of the tests by separately testing two null hypotheses:

$$H_0 : \theta_{(2)} = 1, \quad H_0 : \text{AME}_2 = 0.11. \tag{B.1.10}$$

The empirical sizes of the DB test and the standard survey t-test (t_{svy}) at the 5% nominal level are presented in the table below. The standard survey logit t_{svy} test overrejects by a wide margin, while the DB test exhibits reasonably accurate null rejection rates for both hypotheses in most cases, confirming its robustness to regressor dimensionality.

Table B.1: Empirical rejection frequencies of the tests for $H_0 : \theta_{(2)} = 1$ and $H_0 : \text{AME}_2 = 0.11$ at 5% level. Standard stratified sampling.

Tests	$p = 2$	$p = 5$	$p = 10$	$p = 20$	$p = 50$	$p = 100$
$H_0 : \theta_{(2)} = 1$, $n = 200$						
DB	5.0	4.4	3.7	3.1	4.5	3.3
t_{svy}	6.2	6.4	8.0	8.7	36.0	94.9
$H_0 : \text{AME}_2 = 0.11$, $n = 200$						
DB	5.4	5.3	4.6	3.7	3.5	1.4
t_{svy}	5.7	7.7	7.4	8.2	50.9	93.3

Tests	$p = 4$	$p = 10$	$p = 20$	$p = 40$	$p = 100$	$p = 200$
$H_0 : \theta_{(2)} = 1$, $n = 400$						
DB	4.8	4.4	6.0	3.7	5.6	3.9
t_{svy}	5.0	5.1	6.3	15.9	40.4	98.3
$H_0 : \text{AME}_2 = 0.11$, $n = 400$						
DB	4.5	4.9	5.8	5.0	4.6	3.3
t_{svy}	5.3	6.9	9.1	10.8	46.8	93.7

Notes: $n = 200, 400$. DB and t_{svy} denote the debiased Lasso and standard survey-weighted t tests respectively. 1000 simulation replications.

B.2 Survey Questions Used for Variable Construction

This appendix provides the survey questions used to construct the Cyber Score, the k-means clustering variables, and the Business Digital Usage Score (BDUS). Each set of questions corresponds to specific aspects of digital technology adoption and cyber security challenges.

B.2.1 Cyber Security Incidence Variable Construction

The Cyber Security Incidence variable is equal to 1 if a firm answers "Yes" to one of the following questions and 0 otherwise.

> *To the best of your knowledge, which cyber security incidents impacted your business in 2021? Select all that apply.*

- Incidents to disrupt or deface the business or web presence.
- Incidents to steal personal or financial information.
- Incidents to steal money or demand ransom payment.
- Incidents to steal or manipulate intellectual property or business data.
- Incidents to access unauthorised or privileged areas.
- Incidents to monitor and track business activity.
- Incidents with an unknown motive.

B.2.2 Questions Used for k-means Clustering

The k-means clustering variables were derived from responses to the following survey questions, which identify challenges businesses face in utilizing various digital and financial technologies. Each affirmative response indicates an inefficiency or challenge:

- Does your business face challenges with online transaction processing?

- Does your business face challenges with digital marketing?
- Does your business face challenges with data analytics?
- Does your business face challenges with integrating digital technologies into business operations?
- Does your business face challenges with big data?
- Does your business face challenges with artificial intelligence?
- Does your business face challenges with cloud computing?
- Does your business face challenges with ICT infrastructure?
- Does your business face challenges with government connectivity?
- Does your business face challenges with website operations?

B.2.3 Questions Used for Business Digital Usage Score (BDUS)

The BDUS is based on responses to questions about the adoption of specific digital technologies. A "Yes" response to any of the following indicates that the business has utilized the respective technology:

- Does your business use online transaction processing systems?
- Does your business use digital marketing platforms?
- Does your business use data analytics tools?
- Does your business integrate digital technologies into business operations?
- Does your business utilize big data technologies?
- Does your business employ artificial intelligence tools?

- Does your business use cloud computing services?
- Does your business maintain ICT infrastructure?
- Does your business interact with government systems digitally?
- Does your business operate its own website?

The BDUS score is computed as the total number of "Yes" responses to these questions, with higher scores indicating greater digital engagement.