Perceived Reality:
Does credit card security concern stop you from shopping online?

Jiarui Cai Jing Yang Qinghui Yu*

University of Toronto

April 4, 2017

Abstract

We investigate the marginal effect of perceived risk of credit card transaction on one's tendency to place orders online. We regress the number of orders an individual places on the web for the year against concerns for credit card transaction and other control factors. Different regression techniques are compared against each other, we test and conclude the significance of our regressors. We found that the effect is negative and significant, and a linear model is not powerful enough to explain the shopping behavior of an individual.

keywords: Applied Econometrics, OLS, 2SLS, Instrumental Variables, Heteroskedasticity

*The authors' names alphabetically ordered to indicate equal contribution

1 Introduction

Thanks to a surge in the number of Internet-enabled devices and a maturing logistic network, there has been a rapid growth in the e-commerce sector in the past decade. Reported by Statistics Canada, Canadian companies sold more than $136 billion in goods and services online in 2013, 11% up from $122 billion of 2012 [StatisticsCanada, 2013]. The quick growth is not without its reasons: e-commerce represents an efficient way for retailers to sell their goods, websites like Amazon.ca provides a long reach for the retailers and reduces the need of a physical store. As indicated by alexa.com, a web service that tracks the traffics of every website, Amazon.ca is the seventh most popular website in Canada, only behind the likes of Google and Facebook.[1] Despite the upward trend, many individuals refuse to shop online, often the case, they point to the reason that online shopping is not secure. In data compiled by Statistics Canada, one-fifth (19%) of Canadian refuses to shop online due to security concerns[StatisticsCanada, 2012a].

However, most websites are relatively safe from attacks and the number of security vulnerabilities are decreasing rapidly[Grossman, 2012]. When rare data breaches occur, password and credit card informations are never leaked. This is because servers store sensitive information in a *hash*, which is a one-way algorithm that guarantees any information cannot be reverse computed in any reasonable amount of time. Most attacks on the websites belong to a category called Denial of Service, in which the hacker produces artificial traffic to exhaust the computing capacity of the server, thus crashing it. This type of attack produces only mild inconveniences, not leakage of information. In light of this fact, we wonder if the security risk of online shopping is the true drive behind the individuals who refuse to spend money online. Researchers have pointed out that the perceived risk by consumers bears no correlation with online retailers privacy and security reality, and that perceived risk could play a part in an individual's tendency to shop online [Miyazaki and Fernandez, 2000].

In Canada, 93% of shoppers pay for online purchases with credit cards [Frederick, 2016]. Thus of the perceived security risks, credit card security is arguably the most prominent risk faced by

[1] http://www.alexa.com/topsites/countries/CA

consumers. Concern over the security of credit card payments holds Canadians back from buying more online [Grau, 2008]. Depending on the magnitude of this effect, it might be beneficial for the corporations and governments to spend resources to educate customers so that this mist of danger for online transaction fades for a more efficient way of shopping. Previous research has proven that more privacy and security disclosure made by an online merchant increases trust and revenue [Culnan et al., 1999].

In this paper, we will examine the effect of perceived credit card risk on one's tendency to shop online. OLS regression and instrumental variables are used to establish relationships between the number of orders of goods and services an individual places online in a year and his or her views on using credit card. We discuss the significance and magnitude of such effect and outline drawbacks to our approach. We also provide insights into different approaches or tips for improvements if one were to better tackle this problem.

2 Literature Review

In their paper *Literature derived reference models for the adoption of online shopping*[Chang et al., 2005], the authors summarized all of the previous works on perception of Internet security. They identified 6 studies that explored the effect of perceived privacy or security risks on purchasing intention or actual online purchasing behavior. They found that the previous works often do not agree with each other - half of the studies found a significant negative effect, and the others found no effect. The authors generalized online purchasing risks into two categories: product risk and financial risk. We are primarily interested in the latter. In this section, we will examine three papers that are of most relevance to our study.

2.1 Miyazaki and Fernandez, 2000

In *Internet Privacy and Security: An Examination of Online Retailer Disclosures*[Miyazaki and Fernandez, 2000], the authors examined the effect of disclosures of privacy and security practices by online retailers on consumers perceptions of risk and purchase intentions. Online shopping sites for 381 commercial enterprises based in the US were sampled and grouped into 17 shopping cat-

egories. The percentage of privacy and security statements on the website was compared against the risk perceptions and purchase likelihoods from a consumer survey in 1999. The researchers concluded that the prevalence of online privacy and security statements has negligible effect on both privacy and security perceptions. However, the percentage of privacy and security statements is positively related to online purchase tendencies.

2.2 Miyazaki and Fernandez, 2001

In 2001, Miyazaki and Fernandez further explored risk perceptions among consumers with varying levels of Internet experience and how their perceptions relate to online shopping activity in their paper *Consumer Perceptions of Privacy and Security Risks for Online Shopping*[Miyazaki and Fernandez., 2001]. Multiple regression analysis was applied to data from a survey of 160 respondents from major metropolitan areas in the US. This time, Miyazaki and Fernandez concluded that perceived risk toward online shopping is negatively related to online purchasing rates, but this impact is relatively insignificant.

2.3 Forsythe and Shi, 2003

In their paper *Consumer patronage and risk perceptions in Internet shopping*[Forsythe and Shi, 2003], Forsythe and Shi discusses the types of risks perceived by Internet shoppers and browsers and the relationship between those risks and online patronage. Data is taken from the Graphic, Visualization, and Usability Centres WWW User Survey 1988 from Georgia Institute of Technology with 641 valid responses. Regression analysis suggests that the perceived risks of online shoppers do not influence Internet patronage behaviors in an extensive and systematic way.

The above studies agree that the effect of perceived Internet security risk on online purchasing rates is insignificant. However, it must be noted that these findings come from survey data with limited sample size or dated before 2000. Seeing the massive development around the Internet, we should not assume that the results from these studies reflect current trends.

3 Data and Model

3.1 Overview and Preprocessing

The data used in this study is taken from The Canadian Internet Use Survey (CIUS) conducted by Statistics Canada in 2012.[StatisticsCanada, 2012b] The survey totals 22,615 samples and 132 variables. The variables can be classified into 4 categories: demographics, Internet access information, online shopping frequency and habits, and personal security information. An important point to note is that the public use microdata is not raw survey data, meaning that it had been edited, imputed and weighted by Statistics Canada to eliminate any integrity loopholes. Confidential information such as income and age of an individual was suppressed to protect the privacy of the respondents and was replaced with derived variable such as income quintile and age cohorts.[2]

Not all data in the microdata file is of use for our study, therefore we further processed the data before any numerical computations. This allows us to exactly define the scope of our study. It is implied from our research question that we are only concerned with Internet users, so we dropped any samples that reported no Internet usage in the past year. We also removed any individuals who placed over 365 orders per year. Intuitively, it is very unlikely for an individual to place multiple orders every day. Going through the data, it is obvious that such observations are extreme outliers and likely fake data. The survey also contains responses that is of no use to us, for example, one could *refuse* any question or answer *I do not know*. In this case, we either dropped the response or replaced it with a *no*. If the response is *I do not know* for questions related to the scope of our study, for example **do you use the Internet**, we dropped it. If a nonresponse more likely indicates a *no*, such as an *I do not know* for **are you concerned about using credit card online**, we changed it to *no*. If someone is truly concerned, it is unlikely he answers *I do not know*. After preprocessing, we have 16,366 valid observations left.

3.2 Variables

Based on our research interest, we chose *number of orders made online over the past year* as the dependent variable. The independent variable of interest is *" are you concerned about using credit*

[2] A more detailed explanation could be found from the documentation of the survey.

card over the Internet", for which the answers are *not at all concerned (0)*, *concerned (1)* or *very concerned (2)*.

On top of that, we also accounted for other factors that influence the number of orders made online. This includes an individual's demographic information such as one's age, gender, urban/rural status, highest education achieved, employment, household size, and income quintile. We expect Canadians residing in urban locations to have slightly lower odds of making an online purchase than did individuals residing in rural areas and smaller towns, based on 2007 data from Statistics Canada. This may be the result of more purchasing options at a retail stores in urban areas. We also expect a person's education to be positively correlated with his or her purchases, as better-educated respondents is predicted to be less concerned with Internet security [Hui and Wan, 2007] and thus more likely to make purchases online. Some extra factors we control are one's *home Internet access*, since online shopping activities are most likely to occur during one's leisure time. We also included *years of Internet usage*. Previous research suggests that the likelihood of purchasing on the Internet increases with the increase of one's experience on the Internet [Pesante, 2017]. Internet experience reduces the risk perception of online shopping [Miyazaki and Fernandez., 2001]. In addition, one's *weekly average number of hours on the Internet for personal use* is considered, as more exposure to the web results in more chances for purchase.

A summary statistics of the variables are presented in Table 1.

3.3 Model

We constructed a linear model as follows

$$\begin{aligned} NumOrd = \beta_0 + \beta_1 sc_1 + + \beta_2 female + \beta_3 employed \\ + \beta_4 hhsize + \beta_5 homea + \beta_6 urban + \beta_7 coll + \beta_8 univ \\ + \beta_9 incomeD2 + \beta_{10} incomeD3 + \beta_{11} incomeD4 + \beta_{12} incomeD5 \\ + \beta_{13} ageD2 + \beta_{14} ageD3 + \beta_{15} ageD4 + \beta_{16} ageD6 \\ + \beta_{17} hourD2 + \beta_{18} hourD3 + \beta_{19} hourD4 + \beta_{20} hourD5 + \beta_{21} hourD6 \\ + \beta_{22} yearD2 + \beta_{23} yearD3 + \beta_{24} yearD4 + \beta_{25} yearD5 + \mu \end{aligned} \tag{1}$$

sc_1 is a categorical variable that measures whether or not an individual is concerned about using credit card online, as described in section 3.2. $yearD2$, $yearD3$, $yearD4$, $yearD5$ are dummy variables that stands for $1-2$, $2-5$, $5-10$, and more than 10 years of Internet usage respectively. Notice we omitted $yearD1$, this is to avoid perfect collinearity between these variables. The base group $yearD1$ had less than one year of Internet experience. Similarly, $hourD2$, $hourD3$, $hourD4$, $hourD5$, and $hourD6$ represents $5-10$, $10-20$, $20-30$, $30-40$, and more than 40 hours of Internet usage per week for personal reasons respectively. The base group $hourD1$ had under 5 hours of usage. Likewise, $incomeD2$, $incomeD3$, $incomeD4$, and $incomeD5$ translates to household income quintiles of \$25,000-\$45,000, \$45,000-\$70,000, \$70,000-\$100,000, and over \$100,000 respectively. The base group $incomeD1$ had an income less than \$25,000. Meanwhile, $ageD2$, $ageD3$, $ageD4$, $ageD5$, and $ageD6$ represents age groups of 25-34, 35-44, 45-54, 55-64, and over 65 in the order specified. The base group $ageD1$ represents those under 25. $female$, $employed$, $urban$, and $homea$ are self-descriptive dummy variables that indicate if the observed individual is a female, employed, lives in an urban area and have access to the Internet at home respectively. $coll$ (college or some post-secondary) and $univ$ (university degree) are dummy variables for highest education level achieved. The base group had high school or less education background. Statistics Canada also provided probability weight for each sample, which is $\frac{1}{p_i}$, where p_i is the probability of sample i being selected for the survey. We use this weight in our regressions.

4 Methodology

4.1 Significance Tests

Since we are primarily concerned about the marginal effect of sc_1, we must make sure our estimate is statistically significant. To this end, we will be conducting t-test on the coefficient β_1. We propose the standard hypotheses:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

We will be evaluating the t-test at 1%, 5% and 10% significance level. If the variable is both economically significant and statistically significant, we should be seeing a reasonably large coefficient with a low p-value.

4.2 Functional Form

Functional form misspecification arises when a model does not adequately capture key features of the actual relationship between the indendent variable(x) and the dependent variable(y). This often amounts to not considering non-linear effects, such as those on x^2 or $log(y)$. In our case, because our regressors are all dummy variables, where taking log and squares makes no sense. Therefore, we will only be investigating interaction effects between variables, more specifically, the interaction effect between sc_1 and the control variables $age, income, hour$, and $year$.

4.3 Evaluating Regressor Endogeneity

One potential issue we have identified with this model is reverse causality between $NumOrd$ and sc_1. An individual's concern about using credit card over the Internet may increase with the number of orders he or she made online. It is likely that as one shops more online and shops from more websites, one will become more susceptible to credit card theft. If this is the case, we would have a simultaneity problem, since sc_1 depends on $NumOrd$. The zero conditional mean assumption of the OLS model would be violated. As a result, the OLS estimators may be inconsistent and biased.

To solve this problem, we need to find instrumental variables that are correlated with the endogenous variable sc_1 and uncorrelated with the error term in the OLS model. We have selected two instruments for sc_1: *concerned about conducting banking transactions over the Internet* ($cbanking$) and *do you currently use any paid versions of Internet security software to protect your computer* ($bsoft$). We will use t-test and F-test to test the relevance of the instruments in the original OLS equation and stage one of 2SLS. A good instrument should have significant relationship with sc_1 and insignificant relationship with $NumOrd$ when conditioned on the independent variables. The results and justification will be presented in section 5 and 6.

4.4 Two-Stage Least Squares

Having found an instrument for sc_1, we will be performing Two-Stage Least Squares Regression (2SLS) to solve the simultaneity issue between $NumOrd$ and sc_1.

In the first stage, we will predict a sc_1 using our instruments and control variables. We regress sc_1 on the instrument variables $cbanking, bsoft$, and the control variables used in equation (1). We obtain the equation

$$\hat{sc_1} = \alpha + \pi_1 cbanking + \pi_2 bsoft + \sum_{i=3}^{26} \pi_i z_i + v \tag{2}$$

where z_i are control variables from equation (1).

The first stage of 2SLS removes simultaneity between $NumOrd$ and sc_1 by partialing out the effect of endogenous error terms while leaving only the exogenous part of the variation in sc_1. Second, we will use these estimates to predict sc_1, and denote predicted sc_1 as $\hat{sc_1}$.

After obtaining the $\hat{sc_1}$, we regress $NumOrd$ on $\hat{sc_1}$ and the rest of the independent variables on $NumOrd$ in our second stage, specified below. This should give us an unbiased estimate of β_1

$$NumOrd = \beta_0 + \beta_1 \hat{sc_1} + \sum_{j=2}^{25} \beta_j z_j + \mu \tag{3}$$

where z_i are control variables from equation (1).

4.5 Heteroskedasticity

In multiple regression, we generally assume the error term is homoskedastic - meaning its variance is independent of the regressors. However, in this case, this assumption may not hold. Then we lose the ability to use significance tests in any meaningful way. We would be using Breusch-Pagan test to determine if heteroskedasticity is an issue here. It aims to use regressors to predict $\hat{\mu}^2$ with the equation $E[\mu^2] = \lambda_0 + \lambda_1 \hat{sc}_1 + \sum_{j=2}^{25} \lambda_j z_j + \gamma$.

$$H_0 : \forall\ i \in \{1 \dots 25\}\ \lambda_i = 0$$

$$H_1 : \exists\ i \in \{1 \dots 25\}\ \lambda_i \neq 0$$

If the model is heteroskedastic, the F test on the above regression should produce a small p-value. In this case, we would be using robust standard errors to preserve the variance estimate.

5 Results

5.1 OLS Results

OLS results are presented in Table 2. sc_1 has a marginal effect of -1.359 on $NumOrd$ and this is statistically significant at the the 1% level. This implies that an increase in the level of concern about using credit card over the Internet reduces the number of orders made online by 1.359 orders. This effect may seem small but taking into account that 84% of the respondents reported a total number of orders under 10, it is relatively significant.

Contrary to our expectations, the OLS estimates for $female$ and $urban$ are statistically insignificant at the 10% level, which is consistent with the results found in *A Multivariate Analysis of Web Usage*[Korgaonkar. and Wolin., 2017]. Similarly, compared to the base group $yearD1$, an increase in Internet usage history to $yearD2$, $yearD3$ or $yearD4$ has no statistically significant effect on $NumOrd$ at 10% level, holding all else fixed. However, $yearD5$ has a coeficient of 3.003 which is

statiscally significant at 1%. This suggests that only 10 or more years of Internet usage makes a positive difference in the number of orders.

On the other hand, *employed* and *homea* each has a positive marginal effect of 0.892 and 1.072 on $NumOrd$ respectively while *hhsize* has a negative marginal effect of -0.420. All three variables are statistically significant at 5%. In addition, compared to the base group who had high school or less education background, those in *coll* and *univ* makes 1.041, 3.470 additional orders at 5% and 1% significance level respectively. Also as predicted, increasing income and hours spend on the Internet per week increases $NumOrd$ and the effects are significant at 1%. It is observed that hours have a larger numerical effect on $NumOrd$. For instance, individuals who spend over 10 hours on the Internet per week make 10.510 more orders online than the base group who spend less than 5 hours per week. Lastly, at the 1% significance level, individuals in $ageD2$, $ageD3$ and $ageD4$ makes 3.837, 3.317 and 1.841 more orders online than those in $ageD1$. Individuals in $ageD5$ makes 1.087 more orders than those in $ageD1$ at the 10% significance level. However, there is no statistically significant difference between individuals in $ageD6$ and $ageD1$. This observation is unsuprising since children and senior people are predicted to shop less online, while people between these two groups generally consume more[McKeown and Brocca, 2009].

5.2 Functional Form Test Results

Functional form test results are reported in table 3. Each *svar* is a interaction term derived from $sc_1 \times var$. We have omitted the coefficients on non-interaction terms for compactness. It is seen that there are quite a few significant terms. In particular $sincomeD3, sageD3, sageD5, syearD4$ are significant at 10%, $suniv, syearD5$ significant at 5% and $shourD6$ at 1%. All of the significant interaction terms have a negative slope as expected.

5.3 2SLS Results

The results of 2SLS are reported in Table 4 and 5. $\hat{sc}_1$ has a coefficient of -0.767, which implies that a one level increase in concern about using credit card over the Internet reduces number of orders by 0.767 units. This is statiscally significant at the 10% level. For the rest of the variables, their estimates are relatively similar to the estimates of OLS, except for *employed*. Although the

marginal effect of *employed* is almost the same as its effect in OLS, it is insignificant in this case.

5.4 Heteroskedasticity Tests

The Breush-Pagan test returned a p-value of 0, it shows that the model is indeed heteroskedastic, which is what we expected. In our model, all of the regressors are categorical, which means any variation in the dependent variable implies variation in the error term (see figure 1). Realistically, homoskedasticity is very unlikely to hold. So we obtained all of our 2SLS estimates with robust error.

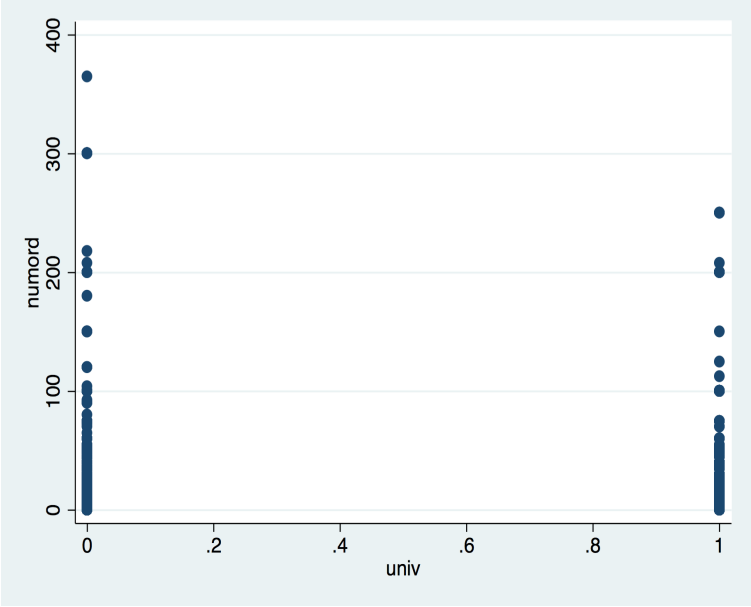


Figure 1: Heteroskedasticity in $NumOrd$

6 Discussions

6.1 The Validity of Instrumental Variables

In Section 4 we presented a set of two instrument variables: $cbanking, bsoft$. Now we present an argument regarding to their validity rooted in economic intuition and statistical significance.

People who purchased Internet security software particularly care about the safety of personal and financial information[Ranganathan and Ganapathy, 2002]; and thus they are more likely to be concerned about the usage of credit card online. This is supported by the p-value which is less

than 0.01 in first stage regression (see 4). The error term in the OLS model can be the online purchasing preferences of consumers, and quality of previous online shopping experience such as information satisfaction[Park and Kim, 2003], which are unobservable in this case. Clearly, these error terms are uncorrelated with the purchase of security software. Therefore, *do you currently use any paid versions of Internet security software to protect your computer* is a valid instrumental variable.

An online banking transaction is defined as a payment method that authorizes a financial transfer, such as bill payment. We argue that *cbanking* is correlated with sc_1, since online orders paid by credit card involve financial transactions, and such transaction is handled through the banking network, naturally consumers would be concerned about security of banking transactions as well. For a statistical argument, please see table 4, it is clear that *cbanking* satisfies the relevancy condition (significant at 1%). Following *An integrative model of organizational trust*[Mayer et al., 1995] and *Not So Different After All: A Cross-Discipline View of Trust* [Rousseau et al., 1998], consumer's trust on banking transaction is mainly correlated with the expectation that the bank will fulfill its obligations, while the trust in using credit card online relies on confidence in online merchants. Since the concerns are rooted in different sources, we argue that after partialling out the effect of concern about credit card and other independent variables on $NumOrd$, concern about banking transaction has no significant effect on $NumOrd$.

By controlling for variables such as education and Internet experience, the effect of a general perceived risk of the Internet on $NumOrd$ is partialled out. As outlined in section 3.2, an increase in education or Internet experience reduces perceived risk. Therefore, we argue that concern about banking transaction is relatively uncorrelated with the error term. Hence, *concerned about conducting banking transactions over the Internet* is a valid instrument. The use of both $cbanking, bsoft$ is justified by the results of the F-test (F statistic = 0.00) in the first stage.

6.2 2SLS

Contrary to OLS results, $\hat{sc}_1$ is only statistical significant under 90% confidence interval. This can be attributed to the larger standard error of $\hat{sc}_1$ estimate at 0.414 compared to 0.259 in OLS. In

addition, with the inclusion of *cbanking* and $bsoft$ in the first stage of 2SLS, part of the endogeneity of sc_1 is removed. Therefore, the $\hat{sc_1}$ has a smaller coefficient.

6.3 Functional Form

OLS models belong to a family of models called *parametric models*, where any variations in the data is captured by a finite number of parameters. Such families of models are great in theoretical frameworks, but generally is not powerful enough when applied to real-world data. These are favored by econometricians and statisticians for their interpretability. In their paper *Partially Adaptive Econometric Methods For Regression and Classification*[Hansen et al., 2010], the authors argue that the normal family distribution is not appropriate to model distributions that are skewed or have fat tails. The authors argue that we should be using more advanced semi-parametric and non-parametric models. In our case, the fact that we have found factors beyond simple control terms that affect $NumOrd$ shows our model does not capture variations beyond simple linear relationships. An example of a non-parametric model is provided in 7.

6.4 Limitations

We are limited by many factors in this study. First, the data is not what we would have liked. In an ideal setting, we would be collecting our own data. This would include a detailed profile of an individual, including his/her *place of birth, income, residency, age* in the form of a raw file. As the discussion about the error term in section 6.1, we would also be asking questions regarding to one's *previous online shopping experience* and whether or not he/she *prefer online or offline shopping.* The fact that we only have access to a pool of categorical variables means that we cannot model variations within the groups. We also cannot use Generalized Linear Squares regression without further imposing functional form assumptions. With the use of robust error, we could correct bias in variance estimate at the cost of efficiency of our estimators. Furthermore, we would also attempt different regression techniques such as ridge regression and categorical probit model on binarized $NumOrd$ and compare their strengths and goodness-of-fit. The choice of instrumental variable is hardly perfect in our study, given the data constraint we are working with. This would mean that the variance for our 2SLS estimate is large, contributing to the eroded statistical significance.

7 Conclusion

We developed and validated a model for investigating the effect of perceived credit card security risk on number of online orders by performing OSL and 2SLS regression. Furthermore, we evaluated the endogeneity of the perceived credit card security. We established that *concerns about conducting banking transactions over the Internet* and *do you currently use any paid version of Internet security software to protect your computer* influence concern about the using credit card over the Internet, but they are less correlated with the number of orders. Hence the pair can be used as intruments to improve our estimation of the effect of perceived credit card security risk on number of orders. All in all, our study concluded that concern about using credit card over the Internet is significant under 95% confidence interval in OLS regression and under 90% confidence interval in 2SLS regression.

The importance of Business-to-commerce (B2C) e-commerce to the future global economy is well known. To stimulate B2C e-commerce development, we need more well-designed online shopping sites that can provide a convenient, safe and trustworthy platform for online shoppers. These characterstics effectively attract and retain consumers. The results of our study suggest that reducing perceived risk of online credit card transactions is one way to achieve it.

References

M. K. Chang, W. Cheung, and V. S. Lai. Literature derived reference models for the adoption of online shopping. *Information & Management*, 42(4):543 – 559, 2005. ISSN 0378-7206. doi: http://dx.doi.org/10.1016/j.im.2004.02.006. URL `http://www.sciencedirect.com/science/article/pii/S0378720604000515`.

M. J. Culnan, S. J. Milberg, R. J. Bies, and M. B. Levy. Consumer privacy, 1999.

S. M. Forsythe and B. Shi. Consumer patronage and risk perceptions in internet shopping. *Journal of Business Research*, 56(11):867–875, March 2003. doi: 10.1016/S0148-2963(01)00273-9. URL `http://resolver.scholarsportal.info/resolve/01482963/v56i0011/867_cparpiis`.

J. Frederick. Canada ecommerce market profile, 2016. URL `http://www.pfsweb.com/pdf/Global-eCommerce-Book-Canada-2016.pdf`.

J. Grau. Canada b2c e-commerce, 2008.

J. Grossman. The state of website security. *IEEE Security & Privacy Magazine*, 10(4):91–93, 2012. doi: 10.1109/MSP.2012.111. URL `http://resolver.scholarsportal.info/resolve/15407993/v10i0004/91_tsows`.

J. Hansen, J. McDonald, P. Theodossiou, and B. Larsen. Partially adaptive econometric meth-

ods for regression and classification. *Computational Economics*, 36(2):153–169, 2010. doi: 10.1007/s10614-010-9226-y. URL http://resolver.scholarsportal.info/resolve/09277099/v36i0002/153_paemfrac.

T.-K. Hui and D. Wan. Factors affecting internet shopping behaviour in singapore: Gender and educational issues. *International Journal of Consumer Studies*, 31(3):310–316, 2007. doi: 10.1111/j.1470-6431.2006.00554.x. URL http://resolver.scholarsportal.info/resolve/14706423/v31i0003/310_faisbisgaei.

P. K. Korgaonkar. and L. D. Wolin. A multivariate analysis of web usage. *Journal of Advertising Research, March*, 1999:53, April 2017. URL http://go.galegroup.com.myaccess.library.utoronto.ca/ps/i.do?p=AONE&sw=w&u=utoronto_main&v=2.1&it=r&id=GALE.

J. Li, S. Ma, T. Le, L. Liu, and J. Liu. Causal decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):257–271, Feb 2017. ISSN 1041-4347. doi: 10.1109/TKDE.2016.2619350.

R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709, Jul 01 1995. URL http://myaccess.library.utoronto.ca/login?url=http://search.proquest.com.myaccess.library.utoronto.ca/docview/1305652827?accountid=14771. Last updated - 2013-02-24.

L. McKeown and J. Brocca. Internet shopping in canada: An examination of data, trends and patterns. *Business Special Surveys and Technology Statistics Division*, 2009. URL http://www.statcan.gc.ca/pub/88f0006x/88f0006x2009005-eng.pdf.

A. D. Miyazaki and A. Fernandez. Internet privacy and security: An examination of online retailer disclosures. *Journal of Public Policy & Marketing*, 19(1):54–61, 2000. ISSN 07439156. URL http://www.jstor.org/stable/30000487.

A. D. Miyazaki and A. Fernandez. Consumer perceptions of privacy and security risks for online shopping. *Journal of Consumer Affairs*, 35(1):27–44, March 2001. doi: 10.1111/j.1745-6606.2001.tb00101.x. URL http://go.galegroup.com.myaccess.library.utoronto.ca/ps/i.do?p=AONE&sw=w&u=utoronto_main&v=2.1&it=r&id=GALEPesante.

C.-H. Park and Y.-G. Kim. Identifying key factors affecting consumer purchase behavior in an online shopping context. *International Journal of Retail & Distribution Management*, 31(1):16–29, 2003. doi: 10.1108/09590550310457818. URL http://resolver.scholarsportal.info/resolve/09590552/v31i0001/16_ikfacpbiaosc.

L. Pesante. On risk, convenience, and internet shopping behavior. *Technical Communication, May*, 2001:225, April 2017. URL http://go.galegroup.com.myaccess.library.utoronto.ca/ps/i.do?p=AONE&sw=w&u=utoronto_main&v=2.1&it=r&id=GALE.

C. Ranganathan and S. Ganapathy. Key dimensions of business-to-consumer web sites. *Information & Management*, 39(6):457 – 465, 2002. ISSN 0378-7206. doi: http://dx.doi.org/10.1016/S0378-7206(01)00112-4. URL http://www.sciencedirect.com/science/article/pii/S0378720601001124.

D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer. Not so different after all: A cross-discipline view of trust. *The Academy of Management Review*, 23(3):393, Jul 01 1998. URL http://myaccess.library.utoronto.ca/login?url=http://search.proquest.com.myaccess.library.utoronto.ca/docview/1840076229?accountid=14771. Last updated - 2016-11-16.

StatisticsCanada. Canada year book, 2012a. URL http://www.statcan.gc.ca/pub/11-402-x/2012000/chap/retail-detail/retail-detail01-eng.htm.

StatisticsCanada. Persons file (public-use microdata file), statistics canada (producer). using chass (distributor). *Canadian/Household internet use survey*, 2012b.

StatisticsCanada. Digital technology and internet use, 2013. URL http://www.statcan.gc.ca/

daily-quotidien/140611/dq140611a-eng.htm.

Tables

Table 1: Summary Statistics

	Frequency	Percent	Cum. Percent
sc1			
Not concerned	4,262	26.04	26.04
Concerned	7,048	43.06	69.11
Very concerned	5,056	30.89	100.00
cbanking			
Not concerned	6,135	37.49	37.39
Concerned	6,248	38.18	75.66
Very concerned	3,983	24.34	100.00
homea			
No	736	4.50	4.50
Yes	15,630	95.50	100.00
hhsize			
1	3,751	22.92	22.92
2	6,082	37.16	60.08
3	2,666	16.29	76.37
4	3,867	23.63	100.00
gender			
Female	7,417	45.32	45.32
Male	8,949	54.68	100.00
income			
≤ 25000	2,290	13.99	13.99
25000 - 45000	3,078	18.81	32.80
45000 - 70000	3,542	21.64	54.44
70000 - 100000	3,714	22.69	77.13
≥ 100000	3,742	22.86	100.00
education			
high	4,889	29.87	29.87
coll	7,402	45.23	75.1
univ	4,075	24.90	100.00
employed			
No	5,071	30.98	30.98
Yes	11,295	69.02	100.00

	Frequency	Percent	Cum. Percent
age			
16 - 24	1,684	10.29	10.29
25 - 34	2,949	18.02	28.31
35 - 44	3,106	18.98	47.29
45 - 54	3,227	19.72	67.01
55 - 64	3,154	19.27	86.28
65+	2,246	13.72	100.00
hour			
≤ 5	7283	44.77	46.83
5 - 10	4,428	27.06	27.06
10 - 20	2,842	17.37	17.37
20 - 30	1,067	6.52	6.52
30 - 40	364	2.22	2.22
≥ 40	382	2.33	100
year			
≤ 1	322	1.96	5.22
1 - 2	533	3.26	3.26
2 - 5	1,926	11.77	11.77
5 - 10	4,635	28.32	28.32
≥ 10	8,950	54.69	54.69

Table 2: OLS Regression Results

Variable	**Coefficient**	**(Std. Err.)**
sc1	-1.359**	(0.259)
female	-0.228	(0.359)
employed	0.892*	(0.446)
hhsize	-0.420*	(0.210)
homea	1.072*	(0.485)
urban	-0.264	(0.340)
coll	1.041*	(0.418)
univ	3.470**	(0.558)
incomeD2	1.123*	(0.490)
incomeD3	2.115**	(0.567)
incomeD4	2.850**	(0.603)
incomeD5	5.160**	(0.583)
ageD2	3.837**	(0.684)
ageD3	3.317**	(0.713)
ageD4	1.841**	(0.614)
ageD5	1.087†	(0.633)
ageD6	-0.325	(0.549)
hourD2	2.536**	(0.388)
hourD3	4.660**	(0.536)
hourD4	6.222**	(0.935)
hourD5	6.089**	(1.304)
hourD6	10.510**	(2.500)
yearD2	-0.483	(0.633)
yearD3	-0.558	(0.603)
yearD4	0.816	(0.654)
yearD5	3.003**	(0.641)
Intercept	-2.107*	(1.068)

N	16366
R^2	0.103
$F_{(26,16339)}$	40.53

Significance levels : † : 10% * : 5% ** : 1%

Table 3: Functional Form test

Variable	**Coefficient**	**(Std. Err.)**
sc1	2.459*	(1.001)
... omitted results for non-interaction terms...		
scoll	-0.488	(0.577)
suniv	-1.716*	(0.705)
sincomeD2	-0.770	(0.555)
sincomeD3	-1.141†	(0.694)
sincomeD4	-0.068	(0.662)
sincomeD5	-1.045	(0.669)
sageD2	-0.072	(0.984)
sageD3	-1.281†	(0.758)
sageD4	-0.441	(0.780)
sageD5	-1.539†	(0.851)
sageD6	-0.010	(0.639)
shourD2	-0.593	(0.434)
shourD3	-0.144	(0.761)
shourD4	0.497	(1.387)
shourD5	-0.086	(1.955)
shourD6	-8.413**	(3.246)
syearD2	-0.673	(0.754)
syearD3	-0.857	(0.698)
syearD4	-1.333†	(0.732)
syearD5	-1.861*	(0.768)
Intercept	-6.374**	(1.486)
N	16366	
R^2	0.111	
$F_{(46,16319)}$	27.452	

Significance levels : † : 10% * : 5% ** : 1%

Table 4: 2SLS: First Stage

Variable	**Coefficient**	**(Std. Err.)**
cbanking	0.574**	(0.010)
bsoft	0.037*	(0.015)
female	0.046**	(0.015)
employed	0.038†	(0.020)
hhsize	0.002	(0.008)
homea	0.096*	(0.040)
urban	-0.002	(0.015)
coll	0.012	(0.020)
univ	-0.033	(0.022)
incomeD2	0.044	(0.029)
incomeD3	0.078**	(0.028)
incomeD4	0.058*	(0.028)
incomeD5	0.006	(0.029)
ageD2	0.019	(0.030)
ageD3	0.054†	(0.030)
ageD4	0.148**	(0.030)
ageD5	0.183**	(0.031)
ageD6	0.216**	(0.035)
hourD2	-0.057**	(0.018)
hourD3	-0.066**	(0.022)
hourD4	-0.118**	(0.035)
hourD5	-0.171**	(0.051)
hourD6	-0.078	(0.051)
yearD2	-0.053	(0.075)
yearD3	-0.070	(0.069)
yearD4	-0.073	(0.066)
yearD5	-0.112†	(0.067)
Intercept	0.376**	(0.083)

N	16366
R^2	0.404
$F_{(27,16338)}$	193.894

Significance levels : † : 10% * : 5% ** : 1%

Table 5: 2SLS: Second Stage

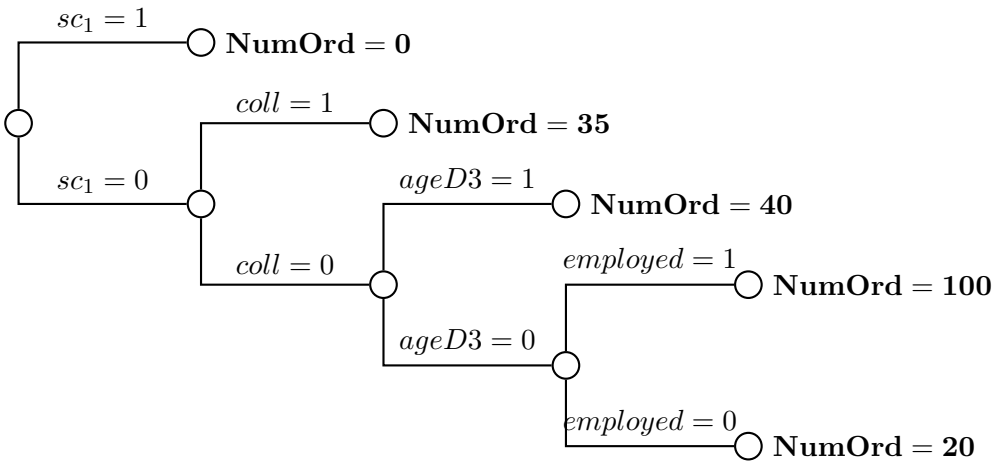
Variable	**Coefficient**	**(Std. Err.)**
sc1hat	-0.767†	(0.414)
female	-0.266	(0.362)
employed	0.873†	(0.450)
hhsize	-0.426*	(0.210)
homea	1.034*	(0.484)
urban	-0.266	(0.340)
coll	1.015*	(0.419)
univ	3.473**	(0.557)
incomeD2	1.121*	(0.490)
incomeD3	2.100**	(0.568)
incomeD4	2.858**	(0.604)
incomeD5	5.228**	(0.582)
ageD2	3.821**	(0.686)
ageD3	3.232**	(0.707)
ageD4	1.661**	(0.615)
ageD5	0.851	(0.645)
ageD6	-0.569	(0.579)
hourD2	2.584**	(0.397)
hourD3	4.714**	(0.535)
hourD4	6.325**	(0.933)
hourD5	6.224**	(1.308)
hourD6	10.540**	(2.514)
yearD2	-0.395	(0.627)
yearD3	-0.445	(0.599)
yearD4	0.973	(0.668)
yearD5	3.201**	(0.647)
Intercept	-2.721*	(1.117)

N	16366
R^2	0.1
$F_{(26,16339)}$	40.31

Significance levels : $\dagger$: 10% $*$: 5% $**$: 1%

Appendix A. Decision Trees

A decision tree is a non-parametric model that models non-linearity. It is especially effective when the regressors are categorical, which is so in our case. A toy tree looks like the following.



When choosing which path to split on, the decision tree will choose a variable on which when we condition the value, we gain the most information. This allows for an efficient and parsimonious way of representing non-linearities.

Obvious from above we know that a decision tree's power is directly proportional to its height, an infinitely high decision tree could represent all of the variation in the world. The most important factor when characterizing the decision tree's power is its height. In our experiment, we established a decision tree with height $h = 10$. A height of 10 is not powerful, but when using it on our data, it reports a R^2 score of 0.614, when reading through the prediction it makes, it is clear that they are accurate.

However, such model has very serious drawbacks that prevents their wide use in econometrics context. They are not interpretable - even though they could be made to.[Li et al., 2017] Most software packages implementing decision trees do not supply visualizing the decision structure made by the tree. They also do not provide partial effect estimation or statistical significance tests. In future econometrics works, these models could provide a supporting evidence on the sign and magnitude of partial effects of interest and not intended as a main tool for study.