The Equilibrium Effects of Asymmetric Information: Evidence from Consumer Credit Markets

Andres Liberman   Christopher Neilson   Luis Opazo   Seth Zimmerman

March 2018

Abstract

This paper exploits a large-scale natural experiment to study the equilibrium effects of information restrictions in credit markets. In 2012, Chilean credit bureaus were forced to stop reporting defaults for 2.8 million individuals (21% of adult population). We show that the theoretical effects of information deletion on aggregate borrowing and total surplus are ambiguous and depend on the pre-deletion demand and cost curves for defaulters and non-defaulters. Using panel data on the universe of bank borrowers in Chile combined with the deleted registry information, we implement machine learning techniques to measure changes in lenders' cost predictions following deletion. Deletion reduces (raises) predicted costs the most for poorer defaulters (non-defaulters) with limited borrowing histories. Using a difference-in-differences design, we find that individuals exposed to increases in predicted costs reduce borrowing by 6.4%, while those exposed to decreases raise borrowing by 11.8% following the deletion, for a 3.5% aggregate drop in borrowing. Using the difference-in-difference estimates as inputs into the theoretical framework, we find that deletion reduced aggregate welfare.

Keywords: Information asymmetry, consumer credit
JEL codes: G20, D14, D82

*Andres Liberman is at New York University, email: aliberma@stern.nyu.edu. Christopher Neilson is at Princeton University, email: cneilson@princeton.edu. Luis Opazo is at ABIF, email: lopazo@abif.cl. Seth Zimmerman is at University of Chicago Booth School of Business, email: seth.zimmerman@chicagobooth.edu. We thank Holger Mueller, Neale Mahoney, Philipp Schnabl, Johannes Stroebel, and seminar participants at Booth Micro Lunch, Boston College, CKGSB, CUHK, EIEF, Erasmus, ITAM, PBC Tsinghua, and Tilburg for comments and suggestions. Sean Hyland and Jordan Rosenthal-Kay provided excellent research assistance. This research was funded in part by the Fama-Miller Center for Research in Finance at the University of Chicago Booth School of Business. We thank Sinacofi for providing the data. All errors and omissions are ours only. First version: October 2017

1 Introduction

Many countries have institutions that limit the information available to consumer lenders. These include 'sunset' provisions that erase defaults after set periods of time, restrictions on the types of past borrowing outcomes and demographic variables that can be used to inform future lending decisions, and one-time purges of default records. The stated motivation for these policies is often that allowing lenders access to certain kinds of information unfairly reduces borrowing opportunities for individuals with past defaults that are no longer predictive of future outcomes (Miller 2003, Elul and Gottardi 2010, Steinberg 2014). These individuals may be disproportionately drawn from historically disadvantaged groups or have suffered from exposure to some negative past shock such as a natural disaster or economic downturn. The goal of limiting lender access to credit information is to improve credit access for these borrowers.

Several recent empirical studies confirm that deleting default records increases borrowing for beneficiaries (Bos and Nakamura 2014, González-Uribe and Osorio 2014, Herkenhoff, Phillips and Cohen-Cole 2016, Liberman 2016, Dobbie, Goldsmith-Pinkham, Mahoney and Song 2016). However, theory emphasizes that the implications these institutions have for aggregate lending and the distribution of acces to credit depend not just on how they affect the beneficiaries of deletion, but on the information asymmetries they induce in consumer credit markets and the equilibrium responses by lenders (Akerlof 1970, Jaffee and Russell 1976, Stiglitz and Weiss 1981). Individuals whose credit information is deleted benefit if lenders perceive them as more willing or able to repay their loans. But this gain may come at a cost to the non-defaulters with whom defaulters are pooled. The effects of information-limiting institutions depend on the tradeoff between these two groups.

This paper develops a simple framework for analyzing the equilibrium effects of information deletion, and uses it to study a 2012 policy change in the Chilean consumer credit market that forced all credit bureaus operating in the country to cease reporting individual-level information on defaults for the majority of defaulters. We have four key findings. First, we show that the effects of information deletion on aggregate borrowing and welfare outcomes are ambiguous, and can be computed using estimates of a) the slopes of the demand and average cost curves in the pre-deletion equilibrium and b) the changes in lenders' cost beliefs that result from deletion. Second, we use machine learning techniques to show that deletion shifts expected costs away from defaulters and towards non-defaulters, with the biggest increases in costs borne by lower-income non-homeowners with good credit records who most resemble high-cost individuals, i.e., defaulters. Third, we use a difference-in-differences design to show that increases in borrowing for the beneficiaries of deletion are more than offset by losses for non-defaulters, so that on net deletion caused borrowing to fall by 3.5% in both markets combined. Fourth, and finally, we incorporate difference-in-difference estimates into our theoretical framework to show that the welfare effects of deletion are also negative under a variety of assumptions about lenders' pricing strategies.

Underlying our analysis is a large-scale change in the information available to lenders in Chilean consumer credit markets. In February 2012, the Chilean Congress passed Law 20,575 (henceforth, the "policy change"), which contained a provision that forced all credit bureaus operating in the country to stop reporting individual-level information on certain defaults.[1] The policy change affected registry information for all individuals whose defaults as of December 2011 added up to less than 2.5 million Chilean pesos (CLP; roughly USD $5,000). This group made up 21% of all Chilean adults and 67% of all bank borrowers at the time of implementation. After the deletion, registry information no longer distinguished individuals with deleted records from those with no defaults. The policy change was a one-time deletion and did not affect how subsequent defaults were recorded. By three years after the one-time deletion, the count of individuals in the default registry had nearly returned to it's pre-deletion level and was still rising.

We study this change using an empirical framework that takes an unraveling model in the style of Akerlof (1970) and Einav, Finkelstein and Cullen (2010) as a baseline. We consider the effects of pooling multiple submarkets following the deletion of differentiating information. Focusing on a simple case with average cost pricing and linear demand and cost curves in high- and low-cost markets, we show that the welfare and borrowing effects of pooling the two markets are ambiguous. The key tradeoff is that, when the deleted information predicts costs, deletion reduces welfare losses from adverse selection for the high-cost group but increases these losses for the low-cost group. The magnitude of these effects depends on the levels and slopes of borrowers' demand and cost curves. We can compute the changes in lending and welfare from information deletion with credible estimates of the slopes of these curves in each submarket.

Our framework indicates that the effects of deletion depend on how it changes lenders' expectations about default outcomes for different groups of borrowers, and on the shape of borrowers' demand and cost curves. We estimate these objects using panel data on the universe of bank borrowers in Chile that includes the information deleted following the policy change and unavailable to lenders. Our empirical analysis

[1] See http://www.leychile.cl/Navegar?idNorma=1037366 for the complete text.

proceeds in two main steps. First, we use machine learning techniques to construct cost predictions with and without the deleted credit information. This helps us understand how lenders' cost expectations shift following deletion. Second, we use variation in cost expectations generated by information deletion to estimate the slope of borrowers' demand and cost curves. The second step employs a difference-in-differences identification strategy that looks at at how borrowing changes following deletion for individuals for whom cost expectations rise and fall.

Our administrative data track defaults and borrowing outcomes between 2009 and 2015, the period surrounding the policy change. Before the policy change, all banks had access to two types of information on lending outcomes for individuals. The first was a database of borrowing and default outcomes for bank borrowing only.[2] This dataset was shared by all banks through the Chilean banking regulator. This dataset allowed banks to observe bank borrowing and default even for loans originated at other banks. The second was the credit registry, which contained default amounts for both bank and non-bank loans. In addition to bank defaults, registry records included defaults on student loans, car loans, and retail credit cards. After the policy change, banks lost access to registry data but kept access to bank data. The effect of the policy was thus to cut off banks' information on non-bank defaults for borrowers dropped from the registry. We have access to both bank borrowing and registry data, including registry records from before the deletion to which banks were subsequently denied access.

We combine these data with a machine learning approach to show how deletion of credit registry data affects the predictions banks can make about borrowers' costs (Mullainathan and Spiess 2017). We use a random forest algorithm to construct two sets of predictions about borrower costs. The first uses both bank borrowing data and credit registry records, while the second uses only the bank borrowing data and not deleted registry records. Eliminating registry data reduces the R^2 in predictions of future defaults by 15%, and leads to systematic overestimates of default probabilities for borrowers without registry defaults and underestimates for borrowers with registry defaults.

We define exposure to the policy as percent increase in predicted costs following deletion. Because non-defaulters outnumber defaulters, exposure is positive (i.e., predicted costs rise) for 61% of the population. The individuals with the largest exposure borrow small amounts and do not have bank or non-bank defaults. They are on average poorer and less likely to own homes. These individuals resemble the borrowers for whom costs fall most dramatically, except that they do not show up on the default

[2] See (Liberman 2016) and (Cowan and De Gregorio 2003) for details on Chilean credit bureaus.

registry. In contrast, predicted costs for individuals who borrow large amounts with higher rates of bank default do not change after deletion.

We use a difference-in-differences analysis to unpack the causal effect of the deletion of information across the exposure distribution. This analysis compares changes in borrowing (and costs) for individuals exposed to increases and decreases in lenders' cost predictions to changes in borrowing for individuals for whom the deleted default records are uniformative about costs. To do this, we use snapshots of borrower and credit registry data at six month intervals leading up to and including the December 2011 snapshot to identify groups of borrowers who would have been exposed to positive, negative, and zero changes in cost predictions had deletion taken place at that time. We use interactions between the predicted exposure variables and a dummy equal to one for cohorts exposed to the actual deletion policy–the December 2011 snapshot– to estimate the effects of deletion in the positive- and negative-exposure group.

There are two assumptions underlying this analysis. The first is that trends in the positive- and negative-exposure groups would have evolved in parallel to the zero-exposure group in the absence of the deletion policy. We evaluate this assumption using a standard analysis of pretrends. The second is that deletion does not affect outcomes for individuals whose cost predictions do not change. We evaluate this assumption using a supplementary difference-in-differences analysis that compares changes in borrowing for the zero-exposure group following the February 2012 policy change to changes for cohorts of zero exposure borrowers the year before. We find no evidence that deletion affected borrowing for this group.

We find that quantities borrowed by the negative- and positive-exposure groups move in parallel to the zero exposure group during the pre-deletion period. Following deletion, borrowing jumps up by 11.7% for the group exposed to cost decreases (on a baseline mean of $141,000 CLP) and falls by 6.4% for the group exposed to cost increases (on a baseline mean of $215,000 CLP). Lenders' cost predictions fall by 29% in the former group and rise by 22% in the latter, corresponding to elasticities of lending to costs of -0.40 and -0.29 in the positive and negative exposure groups, respectively.

Because more borrowers are exposed to increases in predicted costs than to decreases, these estimates indicate that aggregate effect of deletion across the two groups was to reduce borrowing by 3.5%. The total value of the reduction in borrowing is about $20 billion CLP over a six-month period, approximately $40 million. Aggregate declines are largest as a share of borrowing for lower-income borrowers: borrowing drops by 4.2% overall for lower-income individuals and by 3.7% for individuals without mortgages.

Though deletion reduces borrowing, it could still raise total surplus if the individuals for whom borrowing rises value that borrowing more relative to costs than those for whom it falls. The total increase in borrowing for the negative exposure group is 31% the size of the decline for the negative exposure group, so each unit of borrowing in the positive exposure group would need to generate 2.9 times more surplus for this to be the case. This could happen if the high-cost market suffers more from adverse selection at baseline than the low cost market, driving a wedge between consumers valuation and marginal costs. However, our empirical analysis suggests that this is unlikely. Repeating our difference-in-difference analysis with actual costs as the dependent variable suggests that, in both the high- and low-cost markets, realized costs change only slightly as quantity varies. The signs of our estimated cost effects are consistent with adverse selection at the margin in both markets, but we cannot reject cost effects of zero.

The finding of limited effects on cost is consistent with the claim that declines in overall borrowing reflect welfare losses. We formalize this analysis by using the estimated quantity, predicted cost, and realized cost effects from our difference-in-difference analysis as inputs into our equilibrium model of pooling with adverse selection. Together with baseline mean values, these quasi-experimental effect estimates are sufficient to identify welfare effects under the assumption of average cost pricing. We find that pooling increases welfare losses from adverse selection by 72% of the losses relative to the efficient provision prior to pooling. The high-cost market switches from underprovision in the unpooled equlibrium to over-provision in the pooled market, with smaller welfare losses in the latter case. Underprovision increases in the high-cost market.

One limitation of the welfare analysis is that it imposes average cost pricing. In practice lenders likely mark up prices over costs.[3] Our analysis is hampered by the fact that we do not observe interest rates for specific loans. However, we show that our qualitative findings of a welfare loss hold over a wide range of possible markup values in the low- and high-cost markets. Another limitation is that our measure of default is only a proxy for lenders' total cost of making a particular loan. However, our findings do not change when we use alternate plausible cost measures. Finally, it is possible that deletion has additional welfare effects outside of the credit market. For example, there is mixed evidence that credit information may induce externalities in labor markets (Bos, Breza and Liberman 2016, Herkenhoff et al. 2016, Dobbie et al. 2016). We leave this question for future work.

Our paper contributes to several strands of literature on the role of information in

[3] E.g., Ausubel (1991) shows evidence of lack of competition in the US credit card market.

credit markets. The first explores the effects of the removal of default information on borrowing outcomes for beneficiaries of specific policy changes (Musto 2004, Brown and Zehnder 2007, Bos and Nakamura 2014, González-Uribe and Osorio 2014, Herkenhoff et al. 2016, Liberman 2016, Dobbie et al. 2016). These papers show that policies that eliminate default flags raise borrowing for beneficiaries relative to non-beneficiaries. Similar points are raised by the literature that studies the effects of "ban the box" policies that prevent employees from checking job applicants' criminal histories (e.g., (**?**)). We replicate the finding that deletion raises borrowing for beneficiaries, and extend it by showing that when these policies are implemented scale, losses for non-beneficiaries outweigh these gains. In doing so we link the empirical literature on policy experiments in consumer credit to an extensive theoretical literature emphasizing how supply-side responses to asymmetric information can reduce credit availability in equilibrium (Akerlof 1970, Jaffee and Russell 1976, Stiglitz and Weiss 1981).

From a methodological perspective, we innovate by linking a machine-learning approach to estimation of treatment effect heterogeneity with a simple 'sufficient statistics' (Chetty 2009) approach to the underlying economic model. This contributes to a second strand of research that estimates structural models of consumer credit markets and uses them to simulate the effects of policy changes (Adams, Einav and Levin 2009, Einav, Jenkins and Levin 2013, Einav, Jenkins and Levin 2012). Unlike these papers, we focus on a simple demand model and identify key elasticities governing lending outcomes using a natural experiment. This parallels recent empirical work in studies of insurance markets (Einav et al. 2010). In the final section of the paper we illustrate how this procedure can be used to study the effects of two hypothetical policies that limit information available to lenders, the deletion of bank default records in addition to credit registry records and the deletion of information on gender. The deletion of additional default information increases the spread of changes in predicted costs, with bigger gains for winners and losses for lowers than in the policy change as implemented. Deletion of information on gender, which parallels restrictions on demographic information available to lenders in credit markets (Munnell, Tootell, Browne and McEneaney 1996, Blanchflower, Levine and Zimmerman 2003, Pope and Sydnor 2011) and labor markets (**?**), reduces overall borrowing by % and borrowing for women by %.[4] A consistent theme that emerges from our findings is that the costs of information deletion policies fall mostly on individuals observably similar to the intended beneficiaries.

We also contribute to a growing literature within economics that uses machine learn-

[4] **?** study how the removal of credit information may induce equilibrium effects in labor markets when employers use this information to screen applicants.

ing to explore treatment effect heterogeneity given access to many possible mediating variables (Athey and Imbens 2016, Athey and Wagner 2017), and to generate counterfactuals estimates that allow for causal inference where no credible experiment exists (e.g. Burlig, Knittel, Rapson, Reguant and Wolfram (2017)).[5] See Varian (2016) or Mullainathan and Spiess (2017) for a review. In contrast, motivated by the economic theory of markets with asymmetric information, we focus on measures of predicted average costs as the key determinant of heterogeneous treatment effects. This means that, unlike in many machine learning applications, the mechanism underyling effect heterogeneity is transparent and has strong ties to economic theory. It also reduces the set of causal parameters required to apply this approach in other settings from a potentially large number of heterogeneous effects defined across interactions of mediator variables to a single (set of) elasticity(ies). Broadly, our approach is easy to implement and complements the 'big data' that is increasingly prevalent in credit markets and in economics more generally (Petersen and Rajan 2002, Einav and Levin 2014).

2 Economic Framework

2.1 Model

This section presents a framework to help interpret policies that limit credit information and to motivate our empirical approach. Our focus is on understanding how deletion affects welfare and borrowing outcomes through adverse selection, not moral hazard. This is consistent with the empirical application we study here, a one-time deletion based on characteristics that were predetermined at the time of policy announcement. Other applications for which it might be relevant are restrictions on the use of personal characteristics such as race or age in lending and insurance markets. With the goal of transparent empirical implementation, we focus on a simple model of market unraveling in the style of Akerlof (1970).

Consider a consumer credit market where lenders set prices on the basis of observable borrower characteristics but borrowers may have private information on their own costs. Assume for simplicity that the lending market is competitive, so that in equilibrium prices are equal to average costs. This market structure parallels the Einav et al. (2010) model of insurance under adverse selection in that lenders set prices and quan-

[5]Several papers employ machine learning techniques to study credit markets. These include Huang, Chen and Wang (2007), Khandani, Kim and Lo (2010) and **?**. These papers focus on using machine learning techniques to improve cost prediction. In contrast, we use ML techniques to study the effects of actual and counterfactual policy changes on borrowing.

tities are endogenously determined.

Individual borrowers are denoted by i. Lenders partition markets using two types of borrower observable characteristics. The first type, X_i, are always observable to lenders. The second, $Z_i \in \{0, 1\}$, are variables that will be deleted from the lender's information set, e.g., by the policy change. In what follows we supress X_i notation. One can think of this analysis as taking place within subgroups of borrowers defined by $X_i = x$. We model $Z_i = 1$ as being a default flag that predicts higher costs. There are a unit measure of borrowers in the market, of whom a fraction α have $Z_i = 0$ and a fraction $1 - \alpha$ have $Z_i = 1$. Demand and cost functions may vary across values of Z_i. Let $q_z(R)$, $MC_z(R)$, and $AC_z(R)$ denote the demand for credit, marginal cost, and average cost functions for type $Z_i = z$ as a function of the lender's (gross) offer rate R. $q_z(R)$ denotes the *average* quantity of credit purchased for individuals in the market, so that total market quantity is given by $\alpha q_0(R)$ for $Z_i = 0$ and $(1 - \alpha)q_1(R)$ for $Z_i = 1$. To guarantee unique equilibria, we assume that the (inverse) demand curve crosses the marginal cost curve from above exactly once in each market. For analytic tractability, we further assume that the demand and cost curves are linear.

2.1.1 Pre-deletion equilibria

When lenders observe Z_i, equlbria are defined by the intersection of inverse demand and average cost curves in each market. Letting $R_z(q)$ represent the inverse demand curve in each market, equilibrium quantities q_z^e are determined by $R_z(q_z^e) = AC_z(R_z(q_z^e))$. Let $AC_z^e = AC_z(R_z(q_z^e))$ denote the equlibrium average cost in each market.

Following Einav et al. (2010), the slopes of the cost and demand curves determine the losses in total surplus from asymmetric information. We focus on the empirically relevant case where there is adverse selection in both markets; i.e., where marginal cost curves are downward sloping. We illustrate this in Figure 1. The surplus-maximizing quantity q_z^* is determined by $R_z(q_z^*) = MC_z(q_z^*)$. We denote the surplus-maximizing rate as $R_z^* = R_z(q_z^*)$. Deadweight loss due to asymmetric information in market z is the area of the shaded triangle (denoted by "A" in the high cost market and "B" in the low-cost market in Figure 1, respectively), with total welfare loss in each market given by the formula:

$$DWL_z = \frac{1}{2}\left(q_z^* - q_z(AC_z^e)\right) \times \left(AC_z^e - MC_z(AC_z^e)\right). \tag{1}$$

2.1.2 Deletion policy

In the pooling equilibrium lenders no longer observe Z_i. Demand in the pooled market at price R is given by $q(R) = q_0(R) + q_1(R)$, and the pooled market average cost is $AC(R) = s(R)AC_0(R) + (1 - s(R))AC_1(R)$, where the low-cost share $s(R)$ is defined as $s(R) = \frac{q_0(R)}{\alpha q_0(R)+(1-\alpha)q_1(R)}$,. The equilibrium price/average cost AC^e and quantity q^e are determined by $AC^e = AC(R(q^e))$. The changes in average borrowing from pooling in each market are then given by:

$$\Delta q_z = q_z(AC^e) - q_z(AC_z^e),$$

and the average welfare loss by:

$$DWL_z = \frac{1}{2}\left(q_z^* - q_z(AC^e)\right) \times \left(AC^e - MC_z(AC^e)\right).$$

Changes in total surplus from pooling are determined by the relationship between the group-specific demand and cost curves and the pooled average costs. For individuals with $Z_i = 0$ at baseline, rising rates due to pooling increase surplus losses due to underprovision of credit. These additional losses are denoted by "D" in the left panel of Figure 1, the low-cost market. For individuals with $Z_i = 1$, the effects of pooling on total surplus are ambiguous. If $AC^e > R_1^*$, then the effects of the policy for this group are unambiguously positive, as pooling reduces the underprovision of credit due to adverse selection. If $AC^e < R_1^*$, then the effects are unclear. Losses from overprovision in the pooled market may outweigh losses from underprovision in the segregated market. Figure 1 illustrates the latter case, with surplus losses from overprovision equal to the area of triangle C in the left panel.

2.1.3 Measuring the effects of pooling

The equilibrium borrowing and welfare effects of pooling are determined by the slopes of the demand and cost curves in the high- and low-cost markets. Given observations of unpooled quantities q_z^e, costs AC_z^e, and slopes $\frac{dq_z}{dR}$ and $\frac{dAC_z}{dR}$, pooled equilibrium average costs and quantities are given by the solution to the system of equations

$$\begin{aligned} AC^p &= \frac{\alpha q_0^p}{\alpha q_0^p + (1-\alpha) q_1^p} AC_0^p + \frac{(1-\alpha) q_1^p}{\alpha q_0^p + (1-\alpha) q_1^p} AC_1^p \\ q_z^p &= q_z^e + \frac{dq_z}{dR}(AC^p - AC_z^e) \text{ for } z \in \{0,1\} \\ AC_z^p &= AC_z^e + \frac{dAC_z}{dR}(AC^p - AC_z^e) \text{ for } z \in \{0,1\} \end{aligned} \tag{2}$$

There are five equations and five unknowns, yielding an analytic solution for each value. Multiple equilibria are possible but, as we discuss below, not empirically relevant in the setting we consider here.

Computing welfare effects requires knowledge of the levels and slopes of marginal cost curves in addition to the demand and average cost curves. Here we exploit the observation that the equilibrium value of marginal cost $MC_z^e = \frac{dAC_z}{dq} q_z^e + AC_z^e$, and that with linear average cost curves $\frac{dMC_z}{dq} = 2\frac{dAC_z}{dq}$.

2.1.4 Model implications

The effects of pooling on aggregate borrowing and total welfare are ambiguous even in this simple model. We illustrate this point through an exercise in which we simulate equilibrium outcomes from pooling a low-cost and high-cost market. We take as a baseline low-cost and high-cost markets each with a unit measure of potential participants. In the high-cost market, baseline average quantity purchased q_1^e is 50 and baseline average cost AC_1^e is 0.25. In the low-cost market, q_0^e is 100 and AC_0^e is 0.15.

Figure 2 shows how borrowing and welfare effects of pooling vary depending on the slopes of the demand and cost curves in each market. Panel A presents a benchmark case in which there is adverse selection in the high-cost market but none in the low-cost market. Pooling reduces welfare losses from adverse selection in the high-cost market to almost zero, while increasing welfare losses from a baseline value of zero in the low-cost market. Quantity borrowed in the pooled market falls slightly, but welfare losses relative to efficient quantities also fall. Welfare gains in the high-cost group outweight losses in the low-cost group.

Panel B of figure 2 keeps market conditions in the high-cost market the same but adds a downward-sloping average cost curve to the low-cost market. Pooling again slightly reduces quantities borrowed overall. However, because the welfare costs of underprovision in the low-cost market are greater, welfare losses in the low-cost market now outweigh gains in the high-cost market, leading to declines in overall welfare.

Panel C of figure 2 also takes Panel A as a baseline, but makes the high-cost consumers more price-responsive. Here, welfare losses decline from pooling, and quantity borrowed increases. Gains for the high-cost group more than offset losses for the low-cost group by both measures.

2.2 Mapping the model to data

The key empirical challenge when taking this model of pooling to the data is estimating the slopes of demand and cost curves in the high- and low-cost markets. In principle, one could estimate the slopes of these curves using any exogenous shock to price in each market. We use shocks to lenders' beliefs about borrowers' average costs due to information deletion. Under an average cost pricing policy, these shocks translate directly into prices. Let $AC_z^e(x,z)$ denote the value of AC_z^e for individuals with characteristics $X_i = x$ and $Z_i = z$, and $AC^e(x)$ be the value of AC_z^e for individuals with $X_i = x$. We define exposure to the deletion policy E_i for individuals i in markets defined by $X_i = x$ and $Z_i = z$ as

$$E(x,z) = \log AC^e(x) - \log AC_z^e(x,z).$$

Although our model's predictions depend on differences in the level of costs, this stems from the assumption of linearity. In our empirical implementation we focus instead on the difference in log costs because it better reflects lenders' beliefs about changes in risk, and thus, exposure to the policy. For example, under our proposed measure of exposure, a 1% increase in the predicted default rate for borrowers whose baseline predicted default is 1% is comparable to a 10% increase in default for borrowers whose baseline is 10%.

The distribution of pooling effects across borrowers depend on the the distribution of $E(x,z)$. Groups for whom deleting information on Z_i has little affect on cost predictions will be less affected by pooling, either positively or negatively. Our empirical strategy is to look across markets with different values of $E(x,z)$ and observe how borrowing and cost outcomes change following deletion. We use a machine learning approach to choose the conditioning sets of covariates X_i and Z_i. We describe the empirical setting and our approach to analyzing it in the next two sections.

3 Empirical setting

3.1 Formal consumer credit and credit information in Chile

In Chile, formal consumer credit is supplied by banks and by other non-bank financial intermediaries, most notably department stores. As of December 2011 there were 23 banks operating in Chile, including one state owned and 11 foreign owned institutions, which had issued approximately $23 billion in non-housing consumer credit (i.e., credit cards, overdraft credit lines, and unsecured term loans).[6] As of the same month, the 9 largest non-banking lenders (all department stores) had a total consumer credit portfolio of approximately $5 billion. Although banks issue more credit, the number of department store borrowers is larger (14.7 million active non-bank credit cards, of which 5.4 million recorded a transaction during that month, versus 3.8 million consumer credit bank borrowers).[7]

Credit information in Chile is shared by a public credit registry that concerns banks and a private credit bureau that concerns non-banking lenders (Cowan and De Gregorio 2003, Liberman 2016). Banks are required by law to disclose their borrowers' outstanding balance and delinquencies. This information is aggregated by the banking regulator (SBIF) and is only made available to banks. Non-bank lenders are not regulated and disclose only delinquencies, not their borrower's outstanding balances. Financial information aggregators such as our data provider consolidate the data at the individual-level with the use of the unique tax ID. Any person may access a credit report with the use of this ID.[8] Banks (and non-bank lenders) access the credit bureau to obtain information for their lending decisions. As a result, banks may learn a borrower's total bank debt and bank delinquencies, but may only obtain the reported delinquencies from non-banks (i.e., cannot access non-bank debt balances). In turn, non-banks can only learn an individuals' bank and non-bank delinquencies, but not the level of bank or non-bank consumer credit. The key point here is that both bank and non-bank lenders rely on the default registry to run credit checks of potential borrowers.

[6] All information in this paragraph is publicly available through the local banking regulator's website, www.sbif.cl.

[7] Chile's population is approximately 17 million.

[8] As noted above, one publicly stated purpose of the Law was to regulate the use of credit information.

3.2 The policy change

In early 2012, the Chilean Congress passed Law 20,575 to regulate credit information.[9] The bill included a one-time "clean slate" provision by which credit bureaus and information aggregators would stop sharing information on individuals' delinquencies that were reported as of December 2011. This provision affected only borrowers whose delinquencies, including bank and non-bank debts, added up to at most 2.5 million pesos. According to press reports, the provision was a way to alleviate alleged negative consequences of the February 2010 earthquake, which had caused large damage to property and had ostensibly forced a number of individuals into financial distress. The Chilean Congress had already enacted a similar law that forced credit bureaus to stop reporting information on past defaults in 2002. Nevertheless, this new "clean-slate" was marketed as a one-time change, and indeed, all delinquencies incurred after December 2011 were subsequently subject to the regular treatment and reported by credit bureaus.

Following the passage and implementation on February 2012 of Law 20,575, credit bureaus stopped sharing information on defaults for roughly 2.8 million individuals, approximately 21% of the 13 million Chileans older than 15 years old.[10] In effect, this means that individuals who were in default on any bank or non-bank credit as of December 2011 for a consolidated amount below 2.5 million pesos appeared as having no delinquencies after the passage of the law. This is shown in Figure 3, where we plot the time series of the number of individuals in our data with any positive default reported through credit records as of the last day of each semester (ending in June or December).[11] The figure shows a large reduction in the number of individuals with consolidated defaults as of June 2012, after the policy change, relative to December 2011.[12] Interestingly, the figure shows a sharp increase in the number of affected individuals in the following semesters until December 2015, the last semester in our data. This is consistent with the fact that the policy was a one-time change, as future defaults were recorded and reported by credit bureaus, as well as with the fact that many individuals whose defaults were no longer reported did default on new obligations.

The policy change modified the information that lenders, bank and non-bank, could obtain on defaults at other lenders. After the policy change, department stores could

[9] See http://www.leychile.cl/Navegar?idNorma=1037366.

[10] Figure taken from press reports of the "Primer Informe Trimestral de Deuda Personal", U. San Sebastian.

[11] Due to data constraints, our data is limited to individuals who were present in the regulatory banking dataset prior to the passage of Law 20,575.

[12] There is no evidence of an aggregate increase in defaults following the February 2010 earthquake.

no longer verify any type of delinquencies, while banks could not observe whether individuals had defaulted on non-bank debt. However, banks could still verify whether an individual had bank defaults because the banking regulator's data was not subject to the policy change. Thus, the policy change induced a sharp information asymmetry between the banking industry as a whole and its borrowers, rather than creating asymmetries in the information available to each bank with respect to its borrowers.

The median interest rate charged to small borrowers rose following deletion. Figure 4 plots median interest rates for small and large consumer loans before and after the deletion. We observe a 4 percentage point increase in rates in the small loan market, a 15% rise from a base of 26%. Rates continue to rise following the policy change, reaching almost 35% (30% above the base pre-policy rate) by the fourth quarter following implementation. We do not observe changes in rates for larger borrowing amounts, which suggests that the effects we see are not driven by coincident changes in other determinants of borrowing rates. We show below that most new borrowing is done by borrowers with no defaults. This means that the median new loan can be thought of as belonging to this market.

3.3 Data and summary statistics

We obtain from Sinacofi, a privately owned Chilean credit bureau, individual-level panel data at the monthly level on the debt holdings and repayment status for the universe of bank borrowers in Chile from April 2009 until 2014. Sinacofi merged the data to measures of consolidated defaults from the credit registry. We observe registry data at six month intervals, in June and December of each year. As is typical in empirical research on consumer credit, microdata do not include interest rates.

We use these data to build a panel dataset that links registry snapshots to borrowing outcomes. We use the six registry snapshots from December 2009 through December 2011. We link each snapshot to bank borrowing and default outcomes over the six month period beginning two months after the snapshot (i.e., the six month interval beginning in February for the December snapshots, and the six-month interval beginning in August for the June snapshot). This alignment corresponds to the timing of the deletion policy, which took place in February 2012 based on the December 2011 registry records.

Table I reports summary statistics for this data. The first column is the full sample, which includes all individuals who show in the borrowing data. There are 23 million person-time period observations from 5.6 million individuals in the dataset. 37% of borrowers in our dataset have a positive value of consolidated default, with an average value in default of $554,000 CLP. 31% of the population, or 84% of all defaulters, have a default amount strictly between 0 and $2.5 million CLP, and are eligible for deletion. Consistent with our eligibility calculation, we observe deletion for 29% of all borrowers in the December 2011 cohort. Conditional on eligiblity for deletion, the average consolidated amount in default is $172,000 CLP.

The average bank debt balance for consumers is $7.8 million CLP. Unsecured consumer lending accounts for 28% of all debt, for an average of $2.2 million CLP. Mortgage debt accounts for the majority of the remainder. The average bank default balance (defined as debt on which payments are at least 90 days overdue) across all borrowers is $338,000 CLP, or 12% of the overall debt balance. For borrowers eligible for registry deletion, this average is $147,000. Comparing bank default balances to registry default balances shows that deletion eliminates banks' access to 15% ($= 100 \times (1 - 147/172)$) of the default amount among individuals whose balances in default falls below the deletion threshold.

We do not directly observe new borrowing or repayment. Thus, we define new consumer borrowing as any increase in an individual's consumer debt balance of at least 10% month over month, and the amount of new consumer borrowing as an indicator for new borrowing times the amount of the increase. In the full sample, 30% of consumers take out at least one new consumer loan in the six month period following each credit snapshot. The average amount of new borrowing is $184,000 CLP. We define new bank defaults analogously using borrowers' bank default balances. 17% of customers have a new bank default, with an average default amount of $37,000 CLP. In our analysis of the effects of information deletion, we focus on new consumer borrowing as the outcome of interest as defaults are most costly to lenders for uncollateralized borrowing.

The average age in our sample is 44, and 44% of borrowers are female. Our data include data on socioeconomic status for 9% of individuals overall. These data divide individuals into five groups by socioeconomic background. We use these data to generate predictions of socioeconomic status for all individuals in the sample using a machine learning approach. We describe this process in Appendix B. In our empirical analysis we split our sample by this predicted SES categorization. One strong predictor of SES classification is whether or not an individual has a home mortgage. We split by this categorization as well.

The second column of Table I describes our main analysis sample. We focus on borrowers who have a positive debt balance six months prior to the credit snapshot and consolidated default of $2.5 million CLP or less, including zero values. This group accounts for 97% of individuals and 95% of observations. The restriction on debt balances allows us to define a consistent sample across time. Without it, the structure of our data generate spurious increases in mean borrowing over time. This occurs because individuals are included in our sample only if they borrow at some point between 2009 and 2014. An individual with a zero debt balance in 2009 must borrow in the future; otherwise, she would not be included in the data. Subsetting on individuals with positive debt balances at baseline addresses this issue.[13] The restriction to consolidated defaults of $2.5 million CLP or less lets us focus on the part of the credit market where available information changed. Lenders were able to observe consolidated defaults above $2.5 million CLP both before and after the cutoff. Demographics and borrowing in the panel sample are similar to the full dataset.

The third column of Table I describes the the sample of individuals with positive borrowing. As we discuss in the next section, this is the sample we use for constructing cost predictions. They tend to be richer, and have much lower current default balances relative to overall borrowing (0.01 vs 0.09 in the full panel). Their rates of future default are also somewhat lower (0.05 vs. 0.08 in the full panel.

4 Machine learning cost predictions and registry deletion

4.1 Constructing cost predictions

The effect of deletion policies is to change the predictions lenders can make about costs for different kinds of borrowers. We take a machine learning approach that describes changes in cost predictions using a random forest (Mullainathan and Spiess 2017). The intuition underlying this approach is that banks make lending decisions by dividing potential borrowers into groups based on observable characteristics (Agarwal, Chomsisengphet, Mahoney and Stroebel Forthcoming). We have access to borrowers' observable characteristics but do not observe banks' grouping choices. The random forest repeatedly chooses sets of possible predictor variables at random and constructs a regression tree using those predictors. Each tree iteratively splits by the explanatory variables, choosing splits to maximize in-sample predictive power. The random forest obtains predictions by averaging over predictions from each tree. One way to think about this process in our context is as averaging over different guesses about which variables banks might use to classify borrowers.

[13] An alternate approach would be to take the population of all Chileans, irrespective of borrowing, as the sample. We do not have access to data on non-borrowers.

We do not observe banks' true costs. Instead, we focus our discussion on a simple measure of costs: an indicator variable equal to one if a borrower adds to his default balance in the six month period following each registry snapshot. When predicting default outcomes we focus on the sample of individuals who have new borrowing over that same period. We make this restriction because the goal of the exercise is to recover cost predictions for market participants.

We build each tree in our random forest by choosing variables at random from a set of 15 possible predictors. These consist of two lags (relative to the time of policy implementation) of new quarterly consumer borrowing, new quarterly total borrowing, consumer borrowing balance, secured debt balance, average cost, and available credit line, as well as a gender indicator. For pre-policy predictions, the set of variables also includes the credit registry default data. We set the number of trees in a forest to 150. Predictive power is not sensitive to other choices in this range. We choose other model parameters (how many variables to select for inclusion in each tree and the minimum number of observations in a terminal node in the tree) using a cross-validation procedure. For comparison, we also construct predictions using two alternate methods: a logistic LASSO and a naive Bayes classifier. See Appendix B for details on these approaches.

For each method, we construct two sets of predictions. The first set uses training data from the same registry cross-section as the outcome data. These predictions correspond to the best guess a lender can make about default outcomes using data available to them at the time of the loan. We use the contemporaneous predictions as our measure of lenders' best guess about average costs. For this set of predictions, differences between predicted values with and without the default information depend on differences in the average costs in each submarket in the separate-market equilibrium, potentially time-varying shocks to credit demand that move individuals with different covariate values along their cost curves, and endogenous responses to pooling (in the post-pooling time period).

To estimate the slopes of the demand and cost curves, we need to isolate variation in costs due to supply-side price shocks. Our second set of predictions helps us do this. This set of predictions uses training data from the December 2009 registry cross section to generate predictions for all other cross sections. Conditional on covariates, these predictions do not vary across cohorts in the remaining data, and therefore do not reflect the effects of time-varying demand shocks. They use only data from before pooling took place, so they do not reflect endogenous reponses to information deletion. Our empirical analysis uses exposure to deletion policy measured in the second set of predictions to instrument for changes in predicted costs based on the first. We construct both types of predictions using a training sample consisting of 10% of the observations in the relevant snapshot. We exclude the December 2009 data from our difference-in-differences analysis in all specifications, and excluded training data from our outcome analysis.

Table II compares in- and out-of-sample R^2 measures for the random forest to those from other prediction methods. We compute R^2 values using a linear regression of observed default outcomes on default predictions using each method, and present separate estimates for the training and testing periods. The contemporaneous random forest predictions have in-sample (out-of-sample) R^2 values of 0.147 (0.078) when including registry information. Without registry information, these values fall by roughly 15% to 0.135 (0.068). The pre-period random forest predictions have slightly lower training-sample R^2 values and slightly higher testing sample R^2 values, with a similar percentage drop from dropping registry information. Random forest predictions outperform the naive Bayes and logistic LASSO predictions.

4.2 The distribution of exposure to changes in predicted costs

In addition to reducing explanatory power, deletion affects the distribution of cost predictions across defaulters and non-defaulters. The upper panel of Figure 6 shows that predictions with and without deleted default information both track observed costs across the distribution of realized cost outcomes, on average. The cost predictions slightly overpredict costs at the bottom and middle of the cost distribution, and underpredict at the top. The figure focuses on predictions trained in pre-period data, but results are very similar using the predictions based on contemporaneous data.

As shown in the lower-left panel of the graph, differences in observed outcomes between borrowers with and without defaults tend to be small conditional on the full-information prediction. There are almost no borrowers with defaults at the bottom of the full-information predicted cost distribution, and few borrowers without defaults at the very top. In the deleted information predictions, defaulters shift towards the bottom of the distribution and non-defaulters towards to the top. Conditional on the predicted costs, defaulters have higher costs going forward.

Figure 7 explores the distribution of changes in predicted values from deletion in more detail. For each individual, E_i is the percentage change in cost prediction caused by deletion. The upper panel of Figure 7 plots the density of exposure E_i by default status using predictions from the pre-period training set. For non-defaulters, predicted costs rise for 89% of borrowers, with an average increase of 29%. For defaulters, predicted costs fall for 95% of borrowers, with an average drop of 32%. The exposure distribution for defaulters is bimodal, with one mode at zero and the other centered near a decline of 75%. More borrowers are non-defaulters than defaulters, so predicted costs increase for a majority (63%) of borrowers in the market. The lower panel shows a similar distribution of exposure using the contemporaneous training set.

Table III describes how observable attributes of borrowers vary by exposure. We split borrowers into three groups: the 'low-cost market', defined as individuals for whom cost predictions rise by at least 15% following deletion, the 'high-cost market,' defined as individuals for whom cost predictions fall by at least 15%, and the 'zero group,' defined as individuals for whom cost predictions change by less than 15% in either direction. An important point from this table is that most borrowers are exposed to cost increases from deletion: 52% of observations fall into the low-cost category, compared to 32% in the zero-change group and 16% in the high-cost group. Almost all borrowers in the high-cost group have bank defaults, while almost no borrowers in the low-cost group do.

Though the individuals in the low-cost market are more likely to come from high-SES backgrounds and have mortgages, the borrowers whose cost predictions rise most following deletion are those who resemble high-cost borrowers along these dimensions. FIgure 8 plots binned means of indicators for holding some mortgage debt at baseline (left panel) and coming from a high-SES background (right panel). Both graphs have upside-down V shapes. About 20% of borrowers in both the top and bottom deciles of the exposure distribution hold mortgage debt, compared to a maximum of about 30% for borrowers with modest positive exposure. Similarly, about 25% of borrowers in the top and bottom deciles of the exposure distribution come from high-SES backgrounds, compared to a maximum of over 60% for individuals with exposed to slight increases in cost predictions. Intuitively, the borrowers who benefit most from the policy are those who are difficult to distinguish from non-defaulters without access to the deleted information. In contrast, borrowers who are relatively unaffected by the policy are those for whom more accurate information about costs is available outside of the deleted registry.

5 Equilibrium effects

5.1 Empirical strategy

We isolate the effects of changes in lenders' cost beliefs on borrowing outcomes using a difference-in-differences approach. Intuitively, we compare changes in borrowing outcomes before and after deletion for individuals exposed to increases (and decreases) in cost beliefs to those for individuals with near-zero exposure. We construct cohorts of borrowers at six month intervals leading up to the policy change, including the month of the policy change itself. We then compare the effects of exposure to changes in cost expectations in the treated cohort to the effects of exposure in pre-treatment placebo cohorts.

Consider a sample of individuals who are either not exposed to changes in lender beliefs to deletion, or who are exposed to increases (decreases) in predicted costs. Within this sample, we estimate specifications of the form:

$$Y_{ic} = \gamma_c + \tau_c D_{ic} + X_{ic}\Psi_c + e_{ic}. \tag{3}$$

Y_{ic} is borrowing for individual i in cohort c, γ_c are cohort fixed effects, and X_{ic} are a set of individual covariates that include age, gender, and lagged borrowing and default outcomes. D_{ic} is an indicator equal to one if an individual is in the group exposed to increased (decreased) costs.

The coefficients of interest are the τ_c, which capture cohort-specific estimates of the effects of exposure to increases in lender cost predictions on borrowing. We normalize τ_c to be zero in the cohort immediateley prior to deletion. If deletion reduces borrowing for exposed individuals, we expect τ_c to be flat in the cohorts leading up to treatment, and then to become negative in the deletion cohort. We measure exposure using random forest predictions trained in the December 2009 pre-period. We define the zero-exposure group to be the set of individuals for whom $|E_{ic}| < 0.15$. Our findings are not affected by alternate thresholds for inclusion in this group.

This type of specification can recover the total effect of deletion on borrowing under two assumptions. The first is the standard difference-in-differences assumption that borrowing in the non-zero exposure groups follows parallel trends to the zero exposure group. We can evaluate this assumption by looking at pre-trends in the τ_c. The second assumption is that registry deletion does not affect borrowing outcomes for individuals in the zero-exposure group. If registry deletion raised (lowered) borrowing in the zero-exposure group, our estimates will understate (overstate) the gains in borrowing attributable to deletion. We revisit this assumption below using a supplementary difference-in-differences approach.

We also use the difference-in-difference specifications to estimate the effects of deletion on prices and realized costs. Under the assumption that lenders price based on their best predictions of average costs, we define $P_{ic} = AC(X_{ic}, Z_{ic})$ for pre-deletion values of c, and $P_{ic} = AC(X_{ic})$ for post-deletion values. We define average costs as $AC_{ic} = AC(X_{ic}, Z_{ic})$ across all values of c. Consistent with the idea that costs and price estimates should capture the best information available to lenders at the time of purchase, we use contemporaneously-trained random forest predictions to generate these outcome variables. Because the goal of these estimates is to capture how prices and costs change for loans following deletion, we estimate these specifications in the sample of individuals with positive borrowing.

Statistical inference is not straightforward in this setting. We would like to allow for correlation in error terms within the categories that lenders use to set prices, but we do not observe what these categories are. We use an auxiliary machine learning step to identify interactions of covariates within which individuals have similar cost values (i.e., each of these interactions identifies smaller "markets" where borrowers look similar to lenders). We then cluster standard errors in our regressions within groups defined by these interactions. There are 303 such groups in the full sample. Inference is robust to changes in the coarseness of these groupings.

5.2 The effects of information deletion on borrowing and costs

We first report how borrowing, predicted cost, and observed cost outcomes change for individuals with deleted default records relative to individuals without deleted default records. Using the full sample of borrower data in each registry snapshot, we estimate difference-in-difference specifications that interact cohort relative to default with an indicator variable for presence on the default registry. Panel A of Figure 9 and column 1 of Table IV report estimates when the dependent variable is the log of predicted costs. This variable is equal to the (log) prediction using registry defaults records in the pre-deletion period and the prediction that excludes these records in the post-deletion period. The log difference in cost predictions for defaulters relative to non-defaulters is steady in the year leading up to deletion, then falls by 0.66 after deletion, corresponding to a 52% decline in in lenders' cost expectations for defaulters relative to non-defaulters. This is consistent with evidence from Figures 6 and 7 that deletion reduces predicted costs for defaulters and raises them for non-defaulters.

Panel B of figure 9 and column 2 of Table IV report estimates when the dependent variable is new consumer borrowing. Borrowing is steady in the year leading up to deletion. In the six months following deletion borrowing for defaulters rises by just over $41,000 CLP relative to borrowing for non-defaulters. This is 46% of the base-period borrowing of $88,000 CLP for defaulters. Registry deletion raised borrowing for the beneficiaries of deletion relative to non-beneficiaries. However, relative gains reflect a combination of gains for defaulters and losses for non-defaulters, so they do not imply that the policy raised borrowing overall

Figure 10 and the right two panels of IV report estimates of equation 3. These estimates recover effects for borrowers exposed to positive and negative shocks to cost predictions relative to the group where cost predictions do not change following deletion. Lenders' cost expectations for both groups are flat in the year leading up to deletion. At the time of deletion, log cost predictions rise by 0.22 in the positive exposure group and fall by 0.30 in the negative exposure group. Pre-trends in borrowing are also flat for both groups in the year leading up to deletion. Following deletion, borrowing falls by $14,000 CLP in the positive exposure group, equal to 6.4% of pre-period mean for that group. Borrowing rises by $17,000 CLP for the negative exposure group, equal to 11.8% of the pre-deletion mean. The implied elasticity of borrowing with respect to changes in cost predictions is -0.29 (-0.40) in the positive (negative) exposure group.

These estimates indicate that the net effect of deletion was to reduce borrowing. The low-cost market consists of 2.1 million individuals. At an average loss of $14,000 CLP per person, the total loss is just under $30 billion CLP, or $60 million USD at an exchange rate of 500 CLP per dollar. The high cost market consists of 608,000 individuals, with an average gain of $17,000 CLP per person and a total gain of $10 billion CLP or $20 million USD. The net effect of deletion across the two markets was thus to reduce borrowing by $20 billion CLP, or 4% of the total borrowing across the two groups. To the extent the goal of deletion policy was to increase access to credit, it appears to have been counterproductive.

The effects of deletion are largest for the low-SES borrowers who are most exposed to changes in predicted costs. Table V repeats the analysis from Table IV, subsetting by whether borrowers have a mortgage at baseline, and by our predicted measure of socioeconomic status. Individuals without mortgages and lower-SES individuals are more responsive to changes in lenders' cost expectations, and experience larger percentage changes in borrowing. For individuals without mortages, exposure to increased (decreased) costs lowers (raises) borrowing by 7.1% (12.3%) of baseline values. For individuals with mortgages, the percent decline (rise) in quantity borrowed is 2.8% (9.7%).

For low-SES individuals, the percent decrease (increase) in quantity borrowed is 9.2% (12.4%) compared to 6.1% (7.7%) for high-SES individuals.

5.3 Net effects of deletion on welfare

5.3.1 Effects of deletion on costs

The deletion policy reduced overall borrowing, with declines for non-defaulters offsetting gains for defaulters. However, the policy may still have raised welfare if it transferred borrowing from individuals who value credit less relative to costs to individuals who value it more. To explore the welfare effects of pooling, we first assess the potential for market failure in the high- and low-cost markets using the slopes of the average cost curves in each market. We then tie this analysis explicitly to welfare using the framework from section 2.

Table VI reports the effects of deletion on realized average costs in the high- and low-cost market, in the full sample and split by mortgage and SES categories. At baseline, the average cost for borrowers in the low-cost market is 0.017, and the average cost in the high-cost market is 0.119. Deletion has little effect on average costs in either market. It slightly raises average costs for borrowers in the low-cost group and lowers average costs in the high-cost group. Because quantities fall in the low-cost group and rise in the high-cost group, the signs of these point estimates are consistent with downward-sloping average cost curves in both markets. However, in neither case can we reject an effect of zero at conventional levels of significance. These findings suggest that adverse selection is limited conditional on the information available to borrowers before deletion takes effect, and that welfare losses due to asymmetric information may be limited prior to deletion. Findings across subgroups are similar.

To assess the welfare effects of information deletion, we consider the following thought experiment: for a market at the average value of pooled average costs $AC(x)$, what is the welfare effect of moving from an equilibrium where lenders can condition prices on default flag z to one where they cannot? The mean value of $AC(x)$ is 0.050. Conditional on log $AC(x)$, costs are 43% lower for the low-cost type and 36% higher for the high cost type, for level values of $AC(x, z)$ of 0.029 and 0.069 in the low- and high-cost markets respectively. We estimate the slope of the demand curve in each market using results from Table IV and the slope of the average cost curve in each market using results from Table VI.

Panels A and B of Figure 11 show the demand, average cost, and marginal cost curves in the low-cost and high-cost markets, respectively. Demand curves reflect *aver-*

age quantity borrowed by an individual in each market. The pre-deletion equilibrium in each market is determined by the intersection of the demand and average cost curves. Equilibrium (q, p) pairs are $(112, 0.069)$ and $(252, 0.029)$ in the high- and low-cost markets, respectively. The average quantity borrowed across both markets is 220 and the average price is 0.033. Average cost curves slope down in both markets, leading to underprovision relative to the efficient quantity.

Demand is less elastic in the high-cost market than the low-cost market. This means that for some common offer price p in both markets, the share of high cost types in market rises with p. In our linear parameterization, the share of high-cost types in the market is equal to one for $p > 0.14$. Eliminating default records leads to adverse selection because the higher-cost types enter the market first as prices drop.

Panel C of Figure 11 shows the pooled demand, average cost, and marginal cost curves. The demand curve is piecewise linear, with the slope becoming flatter when the low-cost types enter the market at lower prices. The pooled average cost curve is the quantity-weighted average of the average cost curves in the low- and high-cost markets. The marginal cost curve follows the high-cost curve at very high prices, then shifts rapidly downward as the low-cost types enter the market. Equilibrium price and average quantity in the pooled market are given by the intersection of the pooled demand curve and the pooled average cost curve, with $(q, p) = (215, 0.035)$. Quantity borrowed declines on average, and prices rise.

The welfare effect of pooling differs in the high- and low-cost markets. In the low-cost market, pooling exacerbates welfare losses from underprovision. In the the high-cost market, the pooled price is below the intersection of the demand and marginal cost curves, so welfare losses in the pooling equilibrium come from overprovision.

Table VII summarizes the quantitative implications of this analysis. In the low-cost market, the equilibrium price rises from 0.029 before deletion to 0.035 afterward, while average costs do not meaningfully change. Quantity borrowed declines by an average of \$14,000 CLP per person, or a total of \$28 billion CLP. The welfare loss relative to the efficient quantity rises by 114% of the baseline value. In contrast, prices in the high-cost market drop from 0.069 to 0.035, and borrowing rises by \$28,000 CLP per person, or \$17 billion CLP in aggregate. Welfare losses in this market decline by 75%. Aggregating across markets, borrowing falls by \$11 billion CLP, and welfare losses rise by an amount equal to 72% of the welfare loss relative to the optimum at baseline.

5.3.2 Markups over average cost and costs over the long run

Our measures of average cost over a short time horizon are only a proxy for actual costs to lenders. Beyond that, the prices that lenders charge to consumers will include markups over average costs if borrowing markets are not perfectly competitive. We do not observe default rates or prices for specific loans. We therefore consider the robustness of our conclusions to different assumptions about long-run costs and markups.

Appendix Figure

Appendix Table

5.4 Effects for the zero-exposure group and persistence of effects over time

We test the assumption of no effect on the zero-exposure group using a second difference-in-differences specification. This specification exploits variation within borrower cohort by time relative to deletion. Let t index six-month periods relative to the period beginning in February of calendar year c. Within the zero exposure groups, we estimate equations of the for:

$$Y_{ict} = \gamma_c + \theta_t + \tau_t \times 1[c = c_T] + e_{ict}, \tag{4}$$

where Y_{ict} is borrowing for individual i in cohort c at time relative to deletion t. The γ_c and θ_t are cohort and event-time fixed effects, respectively. $1[c = C_T]$ is an indicator equal to one if c is the treated cohort c_0. Here, the τ_t are the effects of interest, capturing how borrowing changes after registry deletion in 2011 relative to changes at the same time of year in previous years.

This specification will capture unbiased estimates of the effect of registry deletion on borrowing for the zero-exposure group if time-of-year effects are the same in the 2011 and earlier borrowing cohorts. It differs substantially from our main approach in the requirements for unbiased estimation. In particular, our main approach differences out time-varying shocks that affect all borrowers by measuring outcomes relative to the zero-exposure group. This supplementary specification requires the strong asssumption that seasonal effects be constant across years.

We present our findings in Figure 12. We follow borrowing outcomes for a year before and after deletion, divided into six month windows. Borrowing grows more rapidly in the pre-deletion period for the 2011 cohort than it did in earlier cohorts, suggesting that seasonal effects may differ from year-to-year. Following deletion, the trend reverses, and borrowing falls for the cohort treated with information deletion relative to the control cohort. That is, following deletion borrowing falls relative to the pre-deletion baseline and even more relative to the pre-deletion trend for the zero-exposure group. Though the presence of pre-deletion trends argues for caution in interpretation, these findings are hard to reconcile with a claim that information deletion *raised* borrowing in the zero exposure group. It follows that our main estimates of the effects of deletion *underestimate* the decline in borrowing from the deletion policy, if anything.

5.5 Additional evidence: borrowing from non-banks

The effects of deletion on aggregate borrowing could be reduced if individuals subject to higher prices for bank credit shift towards non-bank borrowing, or increased if non-bank borrowing also declines. The largest non-bank lenders in Chile are department stores that issue credit cards.[14] We explore how borrowing changed at these institutions using aggregate data on retail credit card lending. Appendix Figures IA2, IA3, and IA4 show no distinct breaks in the total stock of retail credit cards, the number of retail credit cards used, or the amount transacted at the time of deletion.

These findings are consistent with the hypothesis that deletion reduced aggregate borrowing. Deletion effects in the credit card market may be smaller than in the consumer bank lending market because low-risk individuals are very unlikely to borrow in that market both before and after deletion. Median interest rates for credit card lending are 75% higher than for consumer bank lending just before deletion (45% vs. 26% in November 2011) and remain higher following deletion (e.g. 45% vs. 31% in February 2012). That few individuals subsitute from consumer credit to credit card borrowing is consistent with the observation that prices remained lower in the consumer credit market following default.

6 Evaluation of counterfactual deletion policies

Our approach provides a framework through which policymakers can predict the distributional and aggregate effects of changes in credit information. In this section we apply it to two hypothetical changes in the credit information available to lenders. The first is a deletion of banks' internal default records in addition the credit registry. This is a more radical version of the original policy. The second is a deletion of information about gender. The idea of eliminating the use of demographic information has parallels

[14] As of December 2011 the nine largest non-bank lenders (all department stores) had consumer credit portfolios of $5 billiion US, compared to $23 billion for banks.

in US anti-discrimination laws as applied to credit markets (Munnell et al. 1996, Blanchflower et al. 2003, Pope and Sydnor 2011).

In each case, we simulate the effects of counterfactual policies using the following procedure. First, we compute each individual's (log) exposure to the policy by estimating predicted costs with and without the deleted information. We then take our estimates of exposure to cost changes, and scale them by our estimated elasticity of borrowing with respect to costs of $-$ from Table IV. Table VIII presents changes in predicted costs and borrowing by quintile of exposure to cost changes. It also describes the characteristics of borrowers in each quintile.

Unsurprisingly, the more radical deletion option leads to larger changes in predicted costs than the actual deletion policy. Mean exposure in bottom (top) quintile is -0.81 (1.61) under the hypothetical policy, compared to -0.59 (0.65) in the policy as implemented. This implies larger effects on borrowing outcomes. We predict an increase of 13% for bottom-quintile individuals, a decline of 23% for top-quintile individuals, and overall reduction in borrowing of 5.8%. As with the observed policy, the negative effects of the comprehensive default deletion fall most heavily on smaller borrowers and lower-SES individuals.

The overall effects of deletion of gender information are smaller than those of default deletion, but fall most heavily on women and low-SES borrowers. Predicted log costs fall by an average of 0.36 in the bottom quintile of the exposure distribution and rise by 0.31 in the top quintile. 71% of borrowers in the top quintile, where costs rise the most, are women, compared to 35% in the bottom quintile and 43% in the opulation as a whole. 28% percent of bottom-quintile borrowers come from the highest SES group, compared to 19% in the high-SES group. The implication is eliminating the gender flag from credit records reduces borrowing for women by 1.1% on average, compared to an 0.16% reduction in the population as a whole.

7 Conclusion

This paper explores the equilibrium effects of information asymmetries on credit markets in the context of a large-scale policy change that forced credit bureaus to stop reporting past defaults for the majority of defaulters in the Chilean consumer credit market. We develop a simple economic framework that illustrates how the demand and cost curves in separate high- and low-cost markets determine the effects of pooling those markets on borrowing and welfare. The effects of deletion on both of these outcomes are ambiguous in sign.

After using a machine learning approach to construct cost predictions and document the distribution of exposure to increases and decreases in predicted costs from deletion, we estimate the demand and cost curves using a difference-in-differences design. Our core empirical finding is that losses from information deletion are regressive and outweigh gains in this setting: consumer borrowing falls by 3.5% after the policy change, with the largest losses for lower-income individuals with smaller borrowing balances. There is no evidence that the winners from the policy value borrowing sufficiently more than the losers to offset these losses.

Our findings suggest that although policies that limit information availability in credit markets can raise welfare, they should be deployed cautiously. Even if deletion lead to increased borrowing for defaulters, it may reduce lending over all. A feature of deletion policies is that the biggest losers tend to resemble the biggest winners on all characteristics observable to the lender other than the deleted information, so policies implemented with the goal of helping disadvantaged populations also have greatest risk of negative effects for these populations.

Our findings motivate a simple procedure by which policymakers can predict the distributional consequences of a proposed change in credit information. The procedure is to construct cost predictions before and after the change, and identify the individuals with the biggest gains and losses in predicted costs. These estimates can be used alone to classify likely winners and losers, can be paired with estimates of demand elasticities to predict changes in quantity borrowed, or can be combined with estimates of demand and cost elasticities to predict changes in welfare. This approach can also be applied to understanding how existing information-restricting institutions such as sunset provisions affect lending. We leave this exercise for future research.

References

Adams, William, Liran Einav, and Jonathan Levin, "Liquidity Constraints and Imperfect Information in Subprime Lending," *American Economic Review*, 2009, *99* (1), 49–84.

Agarwal, Sumit, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel, "Do Banks Pass Through Credit Expansions to Consumers who Want to Borrow?," *Quarterly Journal of Economics*, Forthcoming.

Akerlof, George A, "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 1970, *84* (3), 488–500.

Athey, Susan and Guido Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7353–7360.

——— **and Stefan Wagner**, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Working Paper*, 2017.

Ausubel, Lawrence M., "The Failure of Competition in the Credit Card Market," *The American Economic Review*, 1991, *81* (1), 50–81.

Blanchflower, David G, Phillip B Levine, and David J Zimmerman, "Discrimination in the small-business credit market," *The Review of Economics and Statistics*, 2003, *85* (4), 930–943.

Bos, Marieke and Leonard I Nakamura, "Should defaults be forgotten? evidence from variation in removal of negative consumer credit information," 2014.

———**, Emily Breza, and Andres Liberman**, "The Labor Market Effects of Credit Market Information," *Review of Financial Studies*, 2016, *Forthcoming.*

Breiman, Leo, "Random Forests," *Machine Learning*, 2001, *45*, 5–32.

———**, Jerom Friedman, Charles J. Stone, and R.A. Olshen**, *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.

Brown, M. and C. Zehnder, "Credit reporting, relationship banking, and loan repayment," *Journal of Money, Credit and Banking*, 2007, *39* (8), 1883–1918.

Burlig, Fiona, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram, "Machine Learning From Schools About Energy Efficiency," *NBER Working Paper*, 2017, (w23908).

Chetty, Raj, "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 2009, *1*, 451–488.

Cowan, Kevin and Jose De Gregorio, "Credit Information and Market Performance: The Case of Chile," in Margaret J. Miller, ed., *Credit Reporting Systems and the International Economy*, Vol. 4, Cambridge, MA: MIT Press, 2003, pp. 163–201.

Dobbie, Will, Paul Goldsmith-Pinkham, Neale Mahoney, and Jae Song, "Bad credit, no problem? Credit and labor market consequences of bad credit reports," Technical Report, National Bureau of Economic Research 2016.

Einav, Liran, Amy Finkelstein, and Mark R Cullen, "Estimating welfare in insurance markets using variation in prices," *The quarterly journal of economics*, 2010, *125* (3), 877–921.

—— **and Jonathan Levin**, "Economics in the age of big data," *Science*, 2014, *346* (6210), 1243089.

——**, Mark Jenkins, and Jonathan Levin**, "Contract Pricing in Consumer Credit Markets," *Econometrica*, 2012, *80* (4), 1387–1432.

——**,** ——**, and** —— , "The impact of credit scoring on consumer lending," *The RAND Journal of Economics*, 2013, *44* (2), 249–274.

Elul, Ronel and Piero Gottardi, "Bankruptcy: Is it enough to Forgive or must we also Forget?," 2010.

González-Uribe, Juanita and Daniel Osorio, "Information sharing and credit outcomes: Evidence from a natural experiment," 2014.

Herkenhoff, Kyle, Gordon Phillips, and Ethan Cohen-Cole, "The impact of consumer credit access on employment, earnings and entrepreneurship," Technical Report, National Bureau of Economic Research 2016.

Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jan Wang, "Credit Scoring with a Data Mining Approach Based on Support Vector Machines," *Expert Systems with Applications*, 2007, *33* (4), 847–856.

Jaffee, Dwight M and Thomas Russell, "Imperfect information, uncertainty, and credit rationing," *The Quarterly Journal of Economics*, 1976, pp. 651–666.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo, "Consumer Credit-Risk Models via Machine-Learning Algorithms," *Journal of Banking & Finance*, 2010, *34* (4), 2767–2787.

Liberman, Andres, "The Value of a Good Credit Reputation: Evidence from Credit Card Renegotiations," *Journal of Financial Economics*, 2016, *120* (3), 644–660.

Miller, Margaret J, *Credit reporting systems and the international economy*, Mit Press, 2003.

Mullainathan, Sendhil and Jann Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Munnell, Alicia H, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney, "Mortgage lending in Boston: Interpreting HMDA data," *The American Economic Review*, 1996, pp. 25–53.

Musto, David K, "What happens when information leaves a market? Evidence from postbankruptcy consumers," *The Journal of Business*, 2004, *77* (4), 725–748.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011, *12*, 2825–2830.

Petersen, Mitchell A and Raghuram G Rajan, "Does distance still matter? The information revolution in small business lending," *The journal of Finance*, 2002, *57* (6), 2533–2570.

Pope, Devin G and Justin R Sydnor, "What's in a Picture? Evidence of Discrimination from Prosper. com," *Journal of Human Resources*, 2011, *46* (1), 53–92.

Steinberg, Joseph, "Your privacy is now at risk from search engines– even if the law says otherwise," *Forbes*, June 2014.

Stiglitz, J.E. and A. Weiss, "Credit rationing in markets with imperfect information," *The American economic review*, 1981, *71* (3), 393–410.

Varian, Hal, "Causal Inference in Economics and Marketing," *Proceedings of the Natural Academy of Sciences*, 2016, *113* (27), 7310–7315.

Figures and Tables

Figure 1: Equilibria for high- and low-cost markets and under pooling

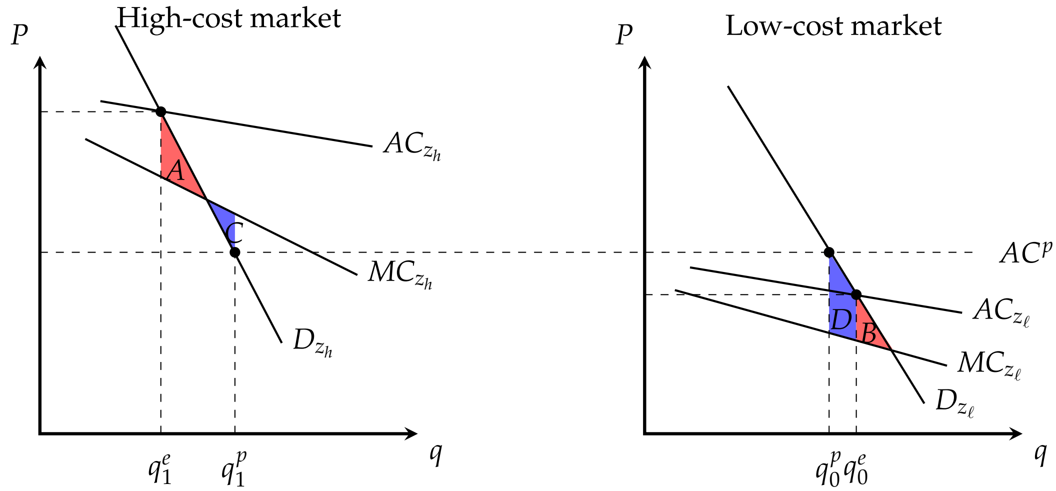
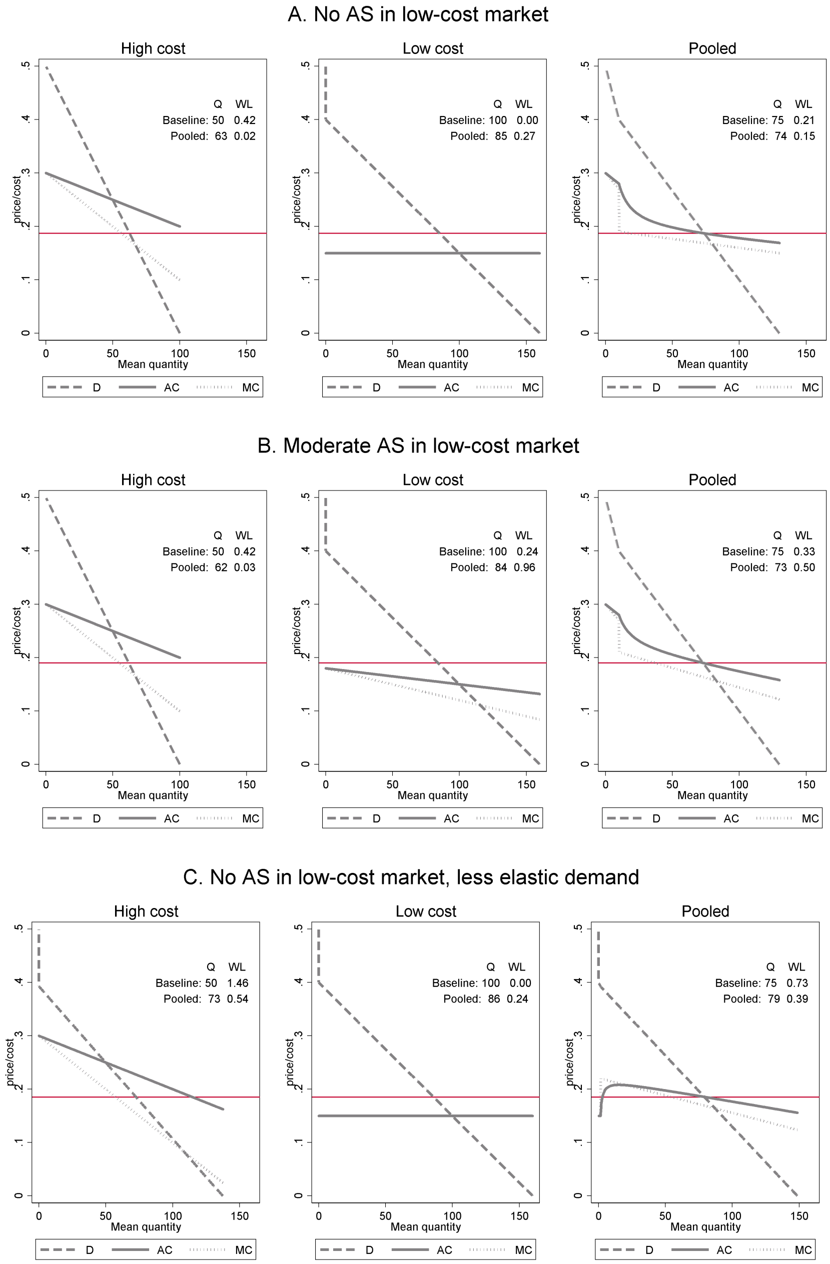


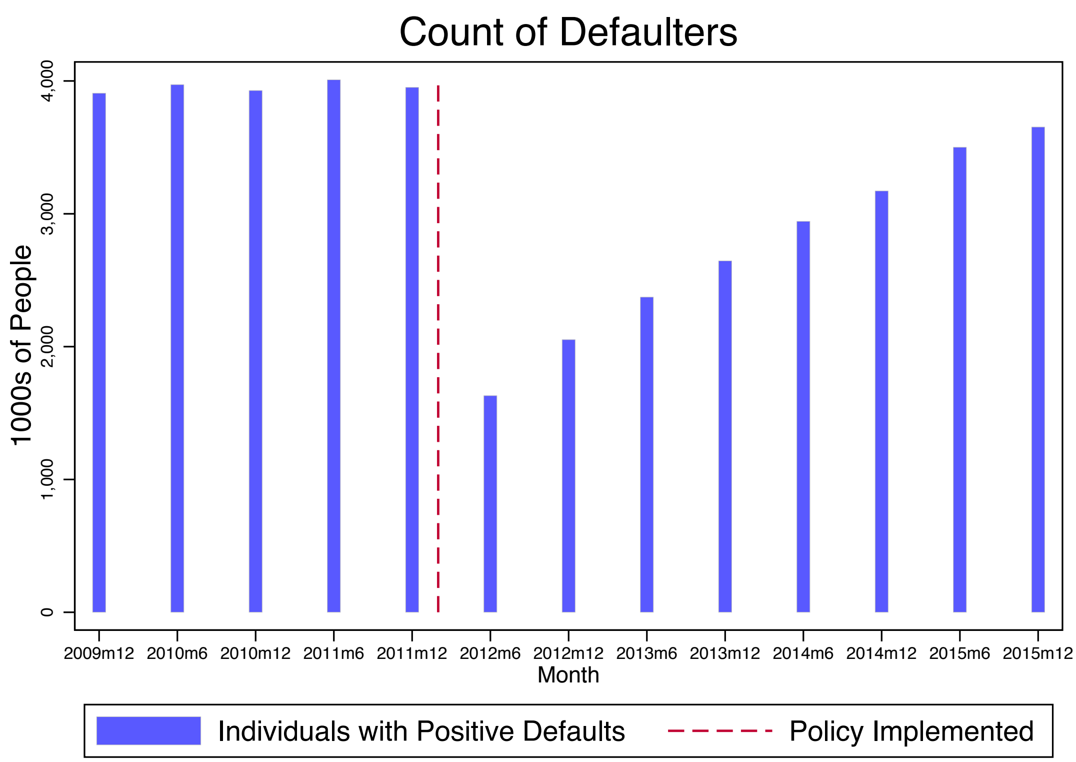
Diagram illustrating the economic framework. Left panel describes the high-cost market; right panel describes the low-cost market.

Figure 2: Simulated separating and pooling equilibria in high-cost and low-cost markets



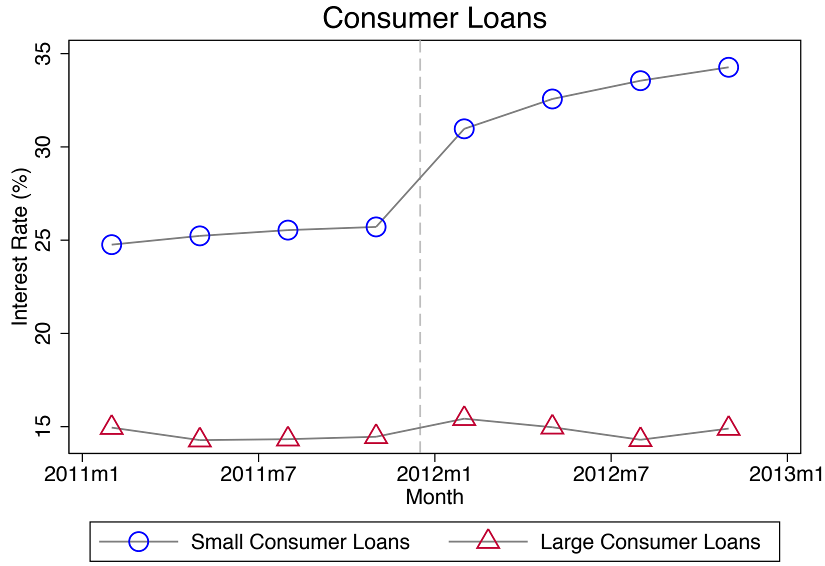
Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about slopes of average cost and demand curves in the high-cost and low-cost market. Common parameters across all panels: N individuals in high- and low-cost market normalized to one. High cost separate equilibrium (p,q)=(0.25, 50), low-cost separate equilibrium is (p,q)=(0.15,100). Parameters listed as $(\frac{dAC_0}{dq}, \frac{dAC_1}{dq}, \frac{dq_0}{dp}, \frac{dq_1}{dp})$ Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$.. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

Figure 3: Individuals with positive past defaults over time



Each bar represents the count of individuals in the credit registry with positive default values at six month intervals. The vertical line represents the implementation of the registry deletion policy.

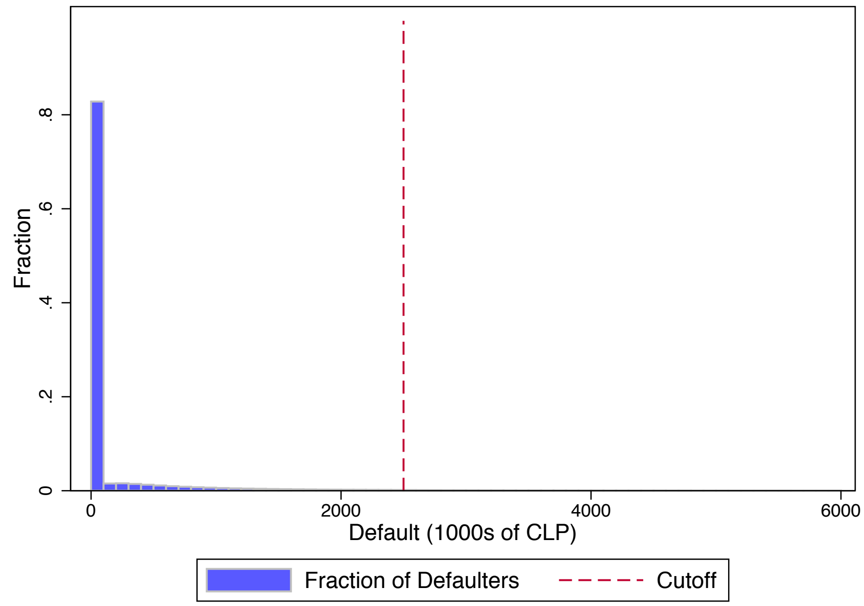
Figure 4: Interest rates



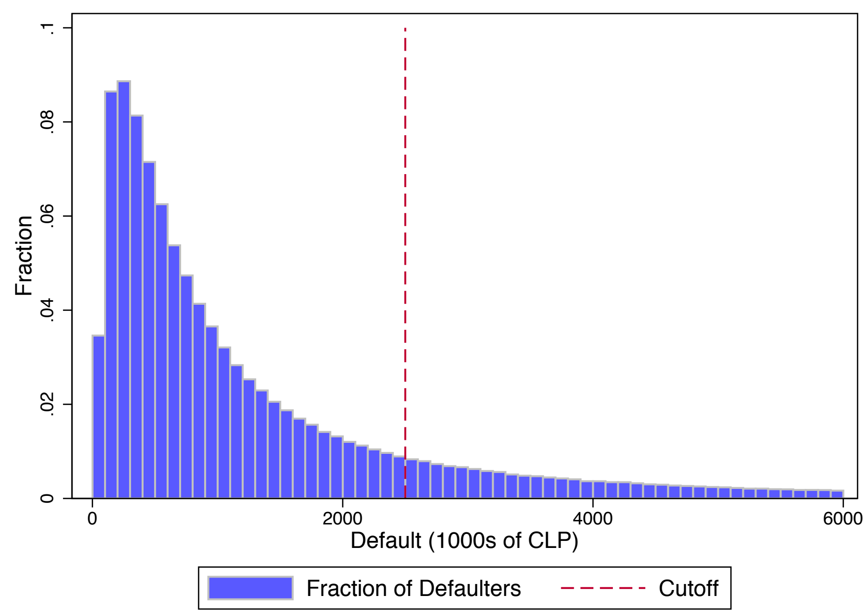
End of period median interest rates for small (top) and large (bottom) consumer loans issued by banks, by quarter relative to December 2011-February 2012. Information on rates obtained from website of Superintendencia de Bancos e Instituciones Financieras, www.sbif.cl.

Figure 5: Histogram of amount in default as of December 2011

Panel A: all individuals

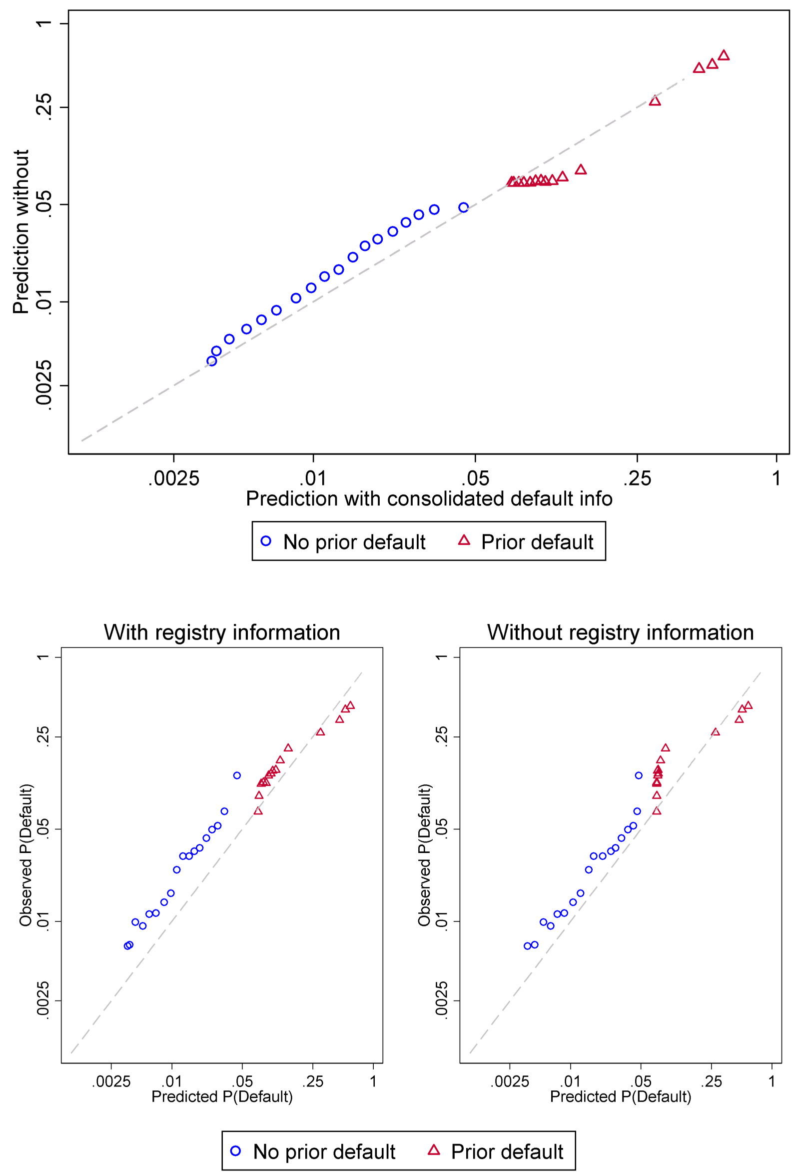


Panel B: conditional on positive default



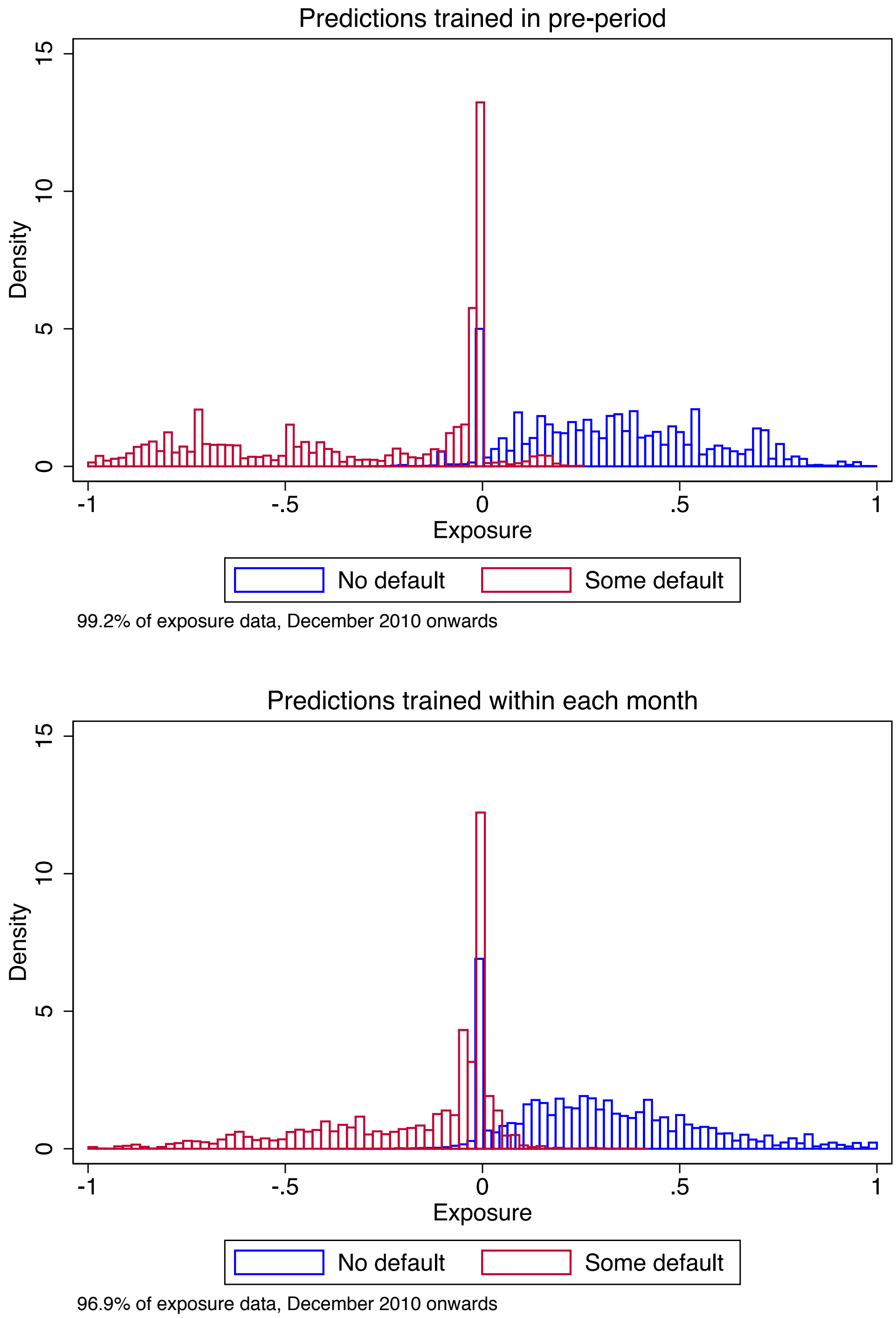
Panel A: Histogram of consolidated defaults as of December 2011, for amounts below \$6 million CLP (approximately \$3,000). Panel B: Histogram of consolidated defaults for individuals with positive defaults only.

Figure 6: Predictions with and without registry data



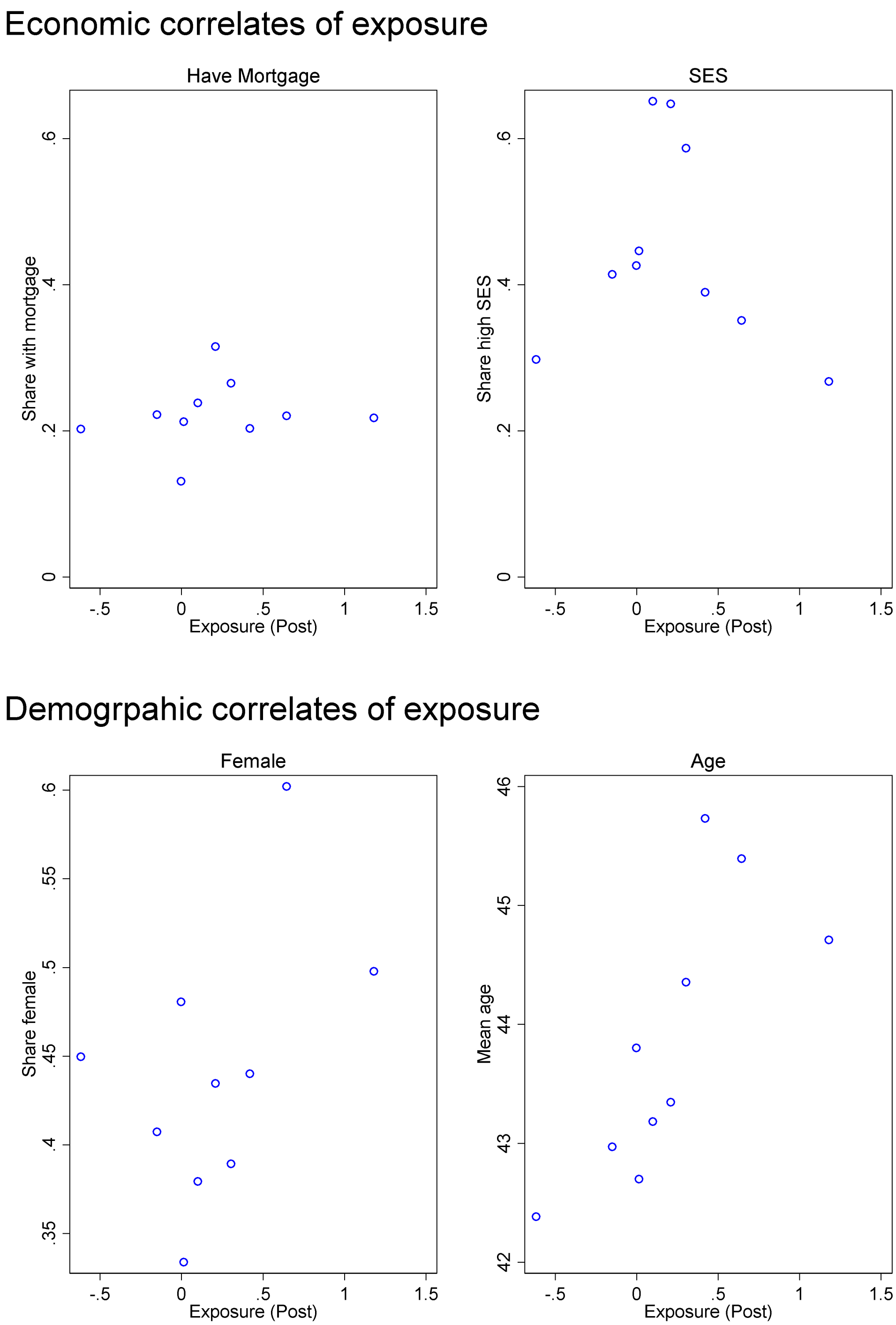
Binned means of random forest cost predictions (horizontal axis; log scale) vs. out-of-sample observed cost outcomes (vertical axis, log scale). Panel A pools over all individuals; panel B splits by default status in December 2011 credit registry. Our cost outcome measure is simply an indicator variable for at least one new default in the six month period beginning in February 2012, the date of registry deletion. Predictions are constructed using registry and borrowing data from December2009 and June 2010. See text for details. 45 degree lines are plotted for convenience.

Figure 7: Density of log exposure to information deletion



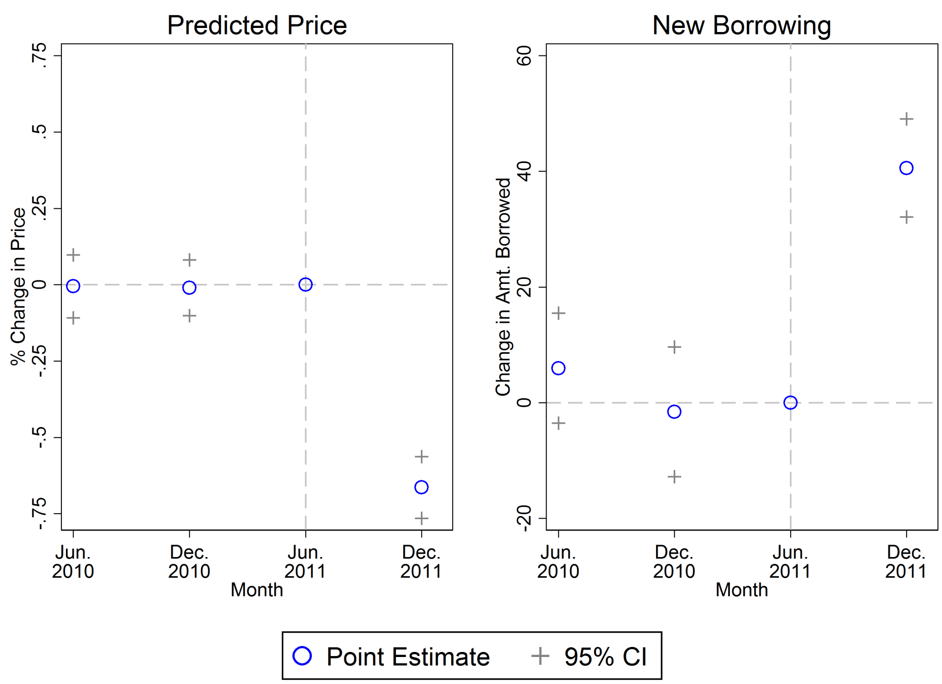
Histogram of exposure to predicted log cost changes by default status. Top: exposure generated from AC^{pre} predictions. Bottom: exposure generated from AC^{post} predictions. Red bars is exposure for defaulters, blue for non-defaulters. Defaulter mean pre-period (post) exposure is -0.32 (-0.17) and non-defaulter mean exposure is 0.33 (0.34). Graphs showexposure distribution between -1 and 1 for each group. Sample: borrower panel from December 2010 through December 2011.

Figure 8: Borrower SES and share of mortgage holders by exposure to information deletion



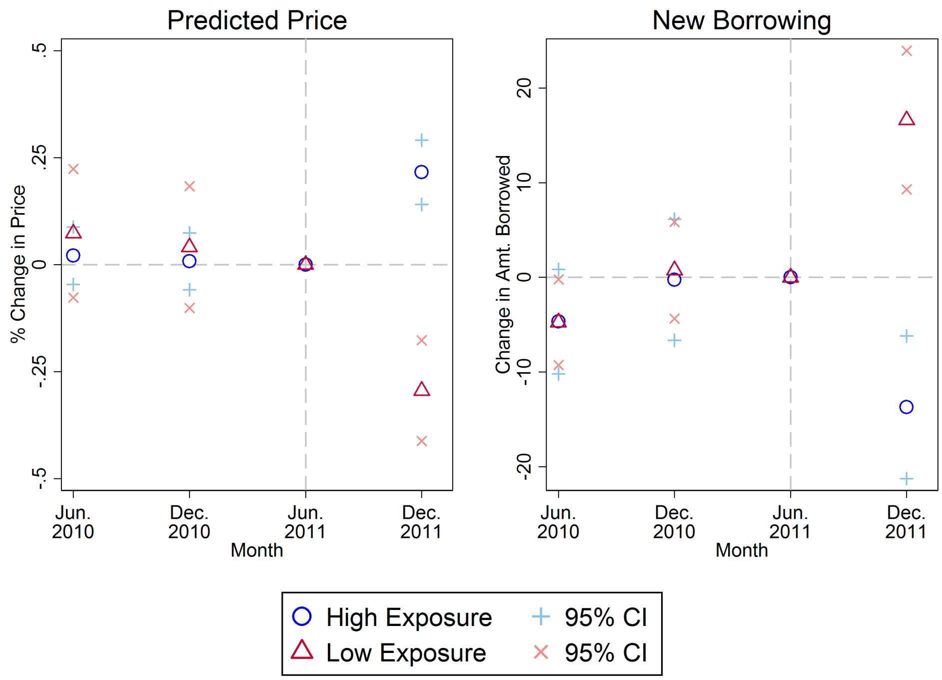
Binned means of indicators for having outstanding mortgage debt (left panel) and coming from a low-SES background (right panel) by decile of exposure distribution. Horizontal axis is log change in cost prediction from deletion. ML predictions come from pre-period training dataset.

Figure 9: Effects of registry deletion on defaulters relative to non-defaulters



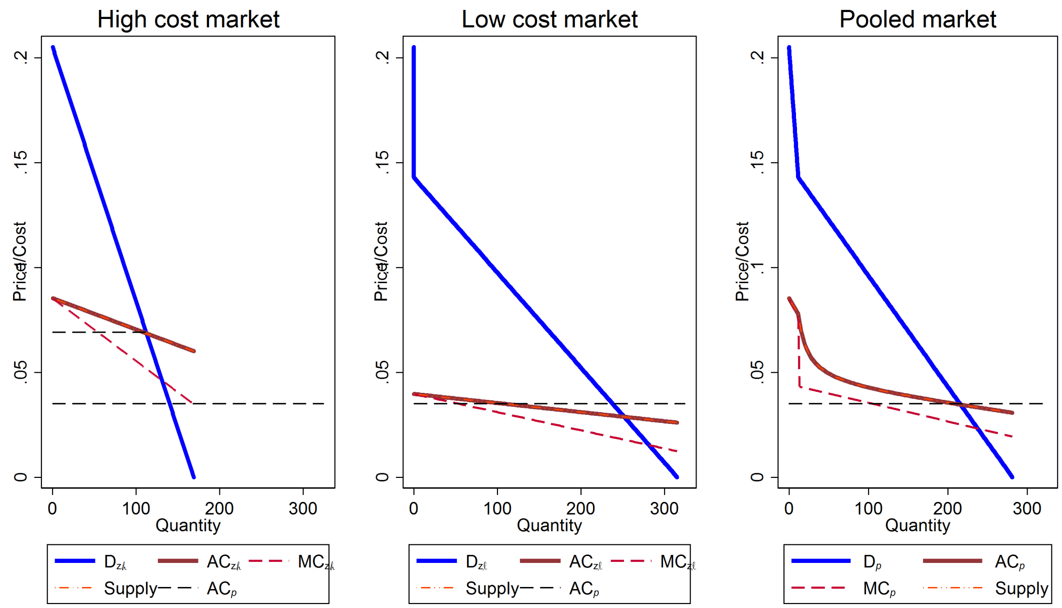
Difference-in-difference estimates and 95% CIs of the effects of prior default on predicted costs (left panel) and observed borrowing (right panel) using equation 3. Borrowing is measured over six month intervals with $t = 0$ in the six month period following deletion in February 2012. Consistent with the implementation of the deletion policy, default status is determined using registry snapshot three months prior to the start of each interval. Standard errors clustered at market level. See text for details.

Figure 10: Effects of registry deletion by exposure to changes in predicted costs



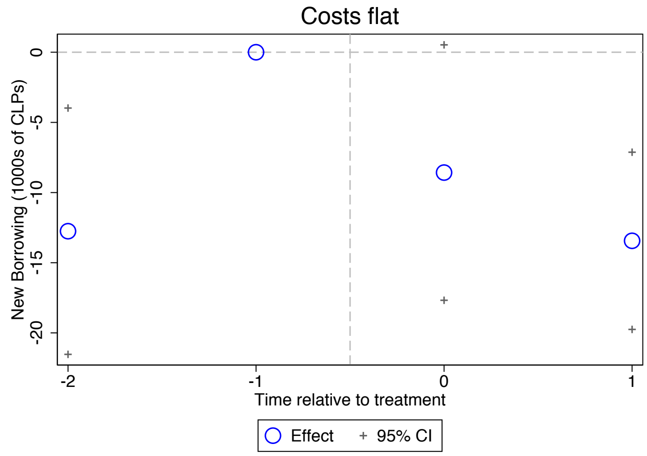
Difference-in-difference estimates and 95% CIs of the effects of exposure to changes in predicted costs on cost predictions (lef panel) and new borrowing (right panel) using equation 3. Each panel splits the sample into individual with positive and negative changes in predicted costs. Effects for each group are measured relative to the omitted category of no exposure to changes in predicted costs, defined as the bottom ten percent of the distribution of the absolute value of cost changes. Standard errors clustered at market level. See text for details.

Figure 11: Empirical estimates of different markets



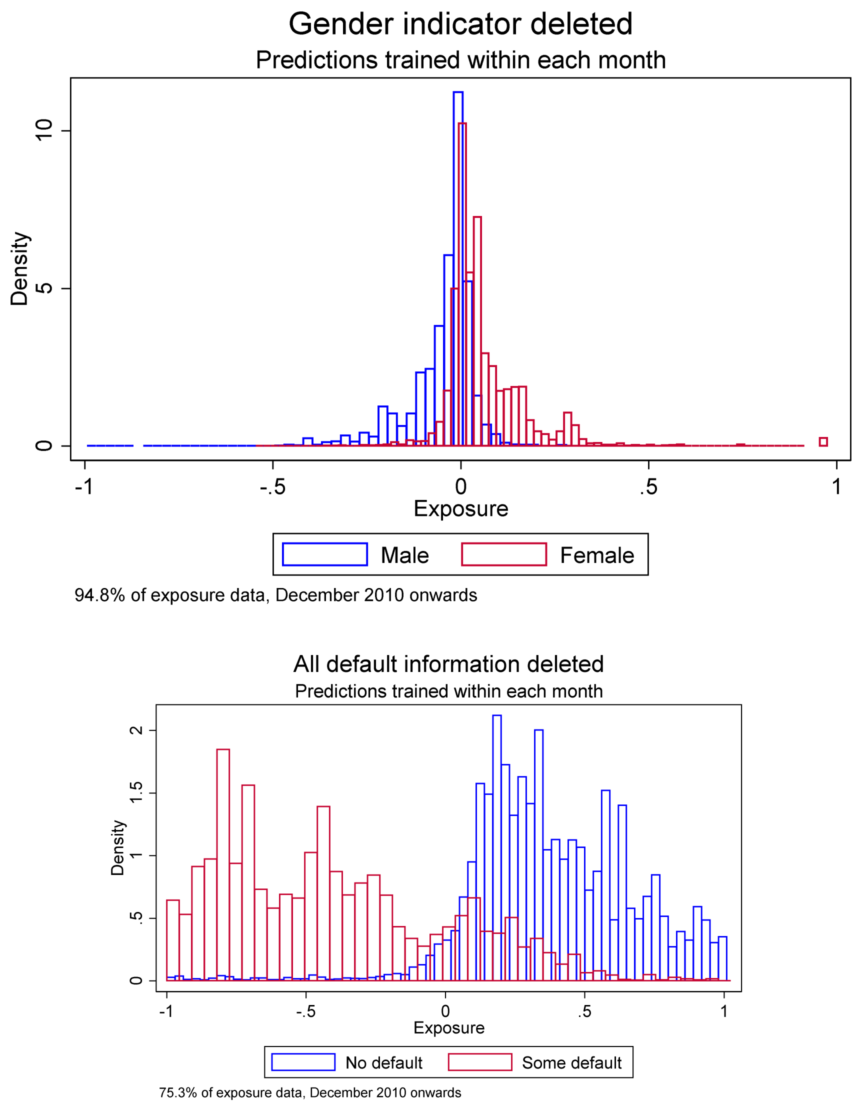
Empirical estimate of figure 1 using difference-in-difference estimates of slopes, assuming a markup of 0% in both markets.

Figure 12: Effects of registry deletion by exposure and time relative to deletion



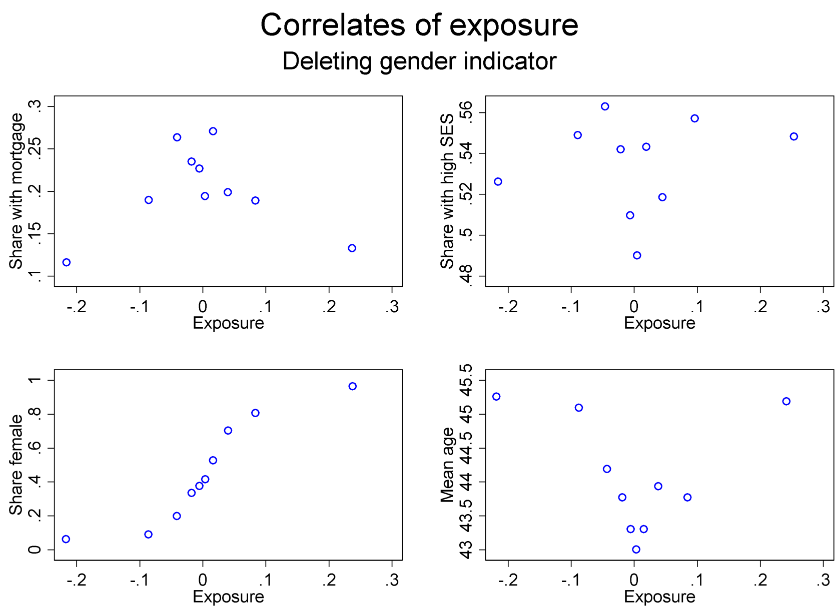
Difference-in-difference estimates and 95% confidence intervals for effects of exposure to changes in borrowing using equation 4 for the exposure-defined 'zero group' only. Horizontal axis in each graph is time in six month intervals relative to the February 2012 deletion policy. These estimates work by comparing changes in borrowing pre- and post-February 2012 to changes pre- and post-February 2011. The 'Costs flat' or zero group is the bottom 15% of the distribution of absolute values in cost changes. Exposure is measured using December 2011 registry data in the 'treatment' sample and in December 2010 in the 'control' sample. Standard errors clustered at market level. See text for details.

Figure 13: Distribution of exposure under counterfactual deletion policies



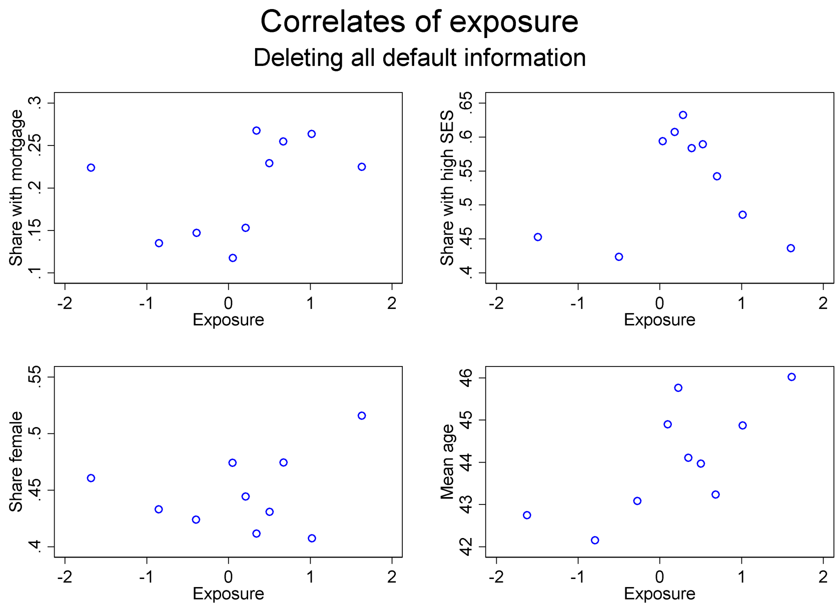
Histograms of exposure under counterfactual deletion policies. On top: log difference in prediced costs excluding and including a gender indicator variable, split by gender. Below: exposure defined when all default information is deleted from the credit registry, split by default amout. See text for details.

Figure 14: Correlates of exposure under counterfactual deletion policy – no gender



Binscatters of correlates of exposure under the counterfactual policy of deleting a gender indicator. See text for details.

Figure 15: Correlates of exposure under counterfactual deletion policy – all default information



Binscatters of correlates of exposure under the counterfactual policy of deleting all default information. See text for details.

Table I: Sample description

	All	In Panel	In Panel, Positive Borrowing
Any registry default	0.37	0.33	0.14
Deletion eligible	0.31	0.33	0.14
Observed deletion	0.29	0.30	0.17
Registry default Amt	554.50	182.00	54.45
Reg. default amt—reg. <2.5m	172.25	182.00	54.45
Debt Balance	7,768	7,675	13,075
Consumer Borrowing Balance	2,172	2,097	2,634
Have mortgage	0.19	0.19	0.24
Mortgage Balance	4,343	4,387	8,192
Any bank default	0.17	0.14	0.03
Bank default amt	338.09	155.81	31.06
Bank default amt— reg. <2.5m	147.46	155.81	31.06
Average Costs	0.12	0.09	0.01
New consumer Borrowing	0.31	0.32	1.00
New Consumer Borrowing Amt	184	190	650
New Bank Default	0.08	0.08	0.05
New bank default amt	36.57	27.28	14.55
Age	44.12	44.08	43.40
Female	0.44	0.45	0.45
Have SES	0.10	0.10	0.13
SES A	0.25	0.25	0.36
SES B	0.29	0.29	0.27
SES C	0.25	0.25	0.20
SES D & E	0.22	0.22	0.17
N (observations)	23,001,337	21,769,213	4,593,511
N of clusters	330	330	330
N (individuals)	5,577,605	5,433,403	2,314,786

Descriptive statistics on borrowing sample. Observations are at the person by half-year level. Data run from August 2009 through July 2012. Six-month snapshots run from February-July and August-January. Borrowing outcomes from each six month interval are linked to credit registry data from two months prior to the start of the interval (December and June, respectively). We refer to time periods by the registry month. Columns define samples. 'All' column is all Chilean consumer bank borrowers. 'In panel' is the set of borrowers with a positive balance six months prior to a given month. 'In panel, positive borrowing' is the subset of borrowers who additionally have new borrowing in the snapshot – a 10% random sample of this subset defines our machine learning training set, which we exclude from the main panel. See text for details. 'Positive default' and 'Default (amt)' are dummies for positive registry defaults and mean default amount conditional on some positive value, respectively. 'Borrowing' is mean consumer borrowing balance. 'New borrowing' is an indicator variable equal to one if quarterly consumer balance expands by 10%, and 'New borrowing, amt' is that indicator multiplied by the observed balance change. 'Debt,' 'New debt,' and 'New debt (amt)' are defined analogously but for all debt, including secured debt. SES categories are internal categorizations used by banks. 'Average costs' are the share of debt at least 90 days overdue divided by the total debt balance.

Table II: Log Likelihoods of Various Algorithms

	AC^{pre}		AC^{post}	
	Training	Testing	Training	Testing
Naive Bayes				
With registry info	−0.412	−0.682	−0.398	−0.633
Without registry info	−0.324	−0.516	−0.319	−0.478
Logistic LASSO				
With registry info	−0.183	−0.327	−0.183	−0.336
Without registry info	−0.187	−0.338	−0.188	−0.348
Random Forest				
With registry info	−0.176	−0.278	−0.173	−0.295
Without registry info	−0.180	−0.284	−0.177	−0.305

Mean binomial log likelihoods for each algorithm. Columns identify the sample in which the log likelihood value is calculated. The 'training' sample is a 10% random sample of borrowers with new borrowing in the July 2009 Snapshot (which trains AC^{pre}) and within each snapshot (AC^{post}). 'Testing' identifies the main sample used in our analysis, from which the training set is dropped. Rows identify prediction methods. Within each prediction method, the 'with registy info' row uses registry information in addition to the other, while the 'without registry info' row does not. See section 4 for the full list of predictors and Appendix B for details on the transformation of these predictors and the structure of each algorithm.

Table III: Demographics by exposure category

	Low cost market	Zero group	High cost market	Pooled
Positive Default	0.01	0.46	0.99	0.31
Amt. Default	52	696	456	566
New Borrowing	236	175	99	195
New Debt	468	356	156	384
Positive Bank Default	0.04	0.15	0.18	0.10
Low SES	0.50	0.56	0.71	0.55
Have Mortgage	0.25	0.18	0.18	0.22
Age	44.4	43.8	42.5	43.9
Female	0.47	0.41	0.46	0.45
Share of individuals	0.53	0.32	0.16	1
N	2,051,138	1,234,733	612,737	3,898,608

Baseline borrowing and demographic characteristics by exposure-generated market type in July 2011. Rows correspond to features of the sample and columns define market type. 'Positive default' is an indicator for whether individuals have positive default balances within the snapshot while 'Amt. Default' computes the mean default value conditional on having positive default. 'New borrowing' computes mean new borrowing across all individuals, as does new 'New debt.' 'Positive bank default' indicates positives bank default for individuals within the snapshot. 'Low SES' is an indicator flagging bank defined socioeconomic status.' Have mortgage' is an indicator flagging whether individuals have positive mortgage balances in the snapshot. 'Age' reports the mean age of individuals in the snapshot in years. 'Female' is flags gender reported to the bank. Share of individuals computes the share of total individuals in the snapshot contained in each market, while N reports the number of individuals (observations).

Table IV: Difference in differences by default and exposure

	Registry Information		*Low cost market*		*High cost market*	
	$\widehat{\text{Price}}$	New Borrowing	$\widehat{\text{Price}}$	New Borrowing	$\widehat{\text{Price}}$	New Borrowing
Jun. 2010	-0.01	5.97	0.02	-4.67^{+}	0.07	-4.74^{*}
	(0.05)	(4.84)	(0.03)	(2.81)	(0.08)	(2.30)
Dec. 2010	-0.01	-1.60	0.01	-0.25	0.04	0.75
	(0.05)	(5.70)	(0.03)	(3.25)	(0.07)	(2.59)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	-0.66^{***}	40.59^{***}	0.22^{***}	-13.72^{***}	-0.29^{***}	16.60^{***}
	(0.05)	(4.32)	(0.04)	(3.83)	(0.06)	(3.72)
Elasticity				-0.29		-0.40
Dep. Var. Base Period Mean			0.04	215.28	0.10	140.98
N Clusters	329	329	303	303	284	285
N Obs.	5,468,917	15,513,587	4,930,411	13,093,725	2,156,891	7,493,968
N Individuals	2,630,314	4,693,948	2,385,366	4,363,940	1,433,629	3,212,628
N Exposed Individuals			855,928	2,132,055	143,165	608,229

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 3. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. We take the log of '$\widehat{\text{Price}}$' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and price effect. '*N* exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Table V: Difference in differences by exposure, mortgage, and socioeconomic status

	Low cost market				*High cost market*			
	$\widehat{\text{Price}}$		New Borrowing		$\widehat{\text{Price}}$		New Borrowing	
By Mortgage Status								
	No Mortage	Mortgage	No Mortage	Mortgage	No Mortage	Mortgage	No Mortage	Mortgage
Jun. 2010	0.03	-0.05^{*}	-5.21	-3.84	0.11	-0.10	-6.08^{*}	-0.78
	(0.04)	(0.02)	(3.31)	(4.22)	(0.08)	(0.06)	(2.56)	(4.33)
Dec. 2010	0.02	-0.05^{+}	1.04	4.48	0.07	-0.09^{+}	0.46	5.59
	(0.04)	(0.03)	(3.29)	(5.66)	(0.08)	(0.05)	(2.81)	(4.16)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.20^{***}	0.22^{***}	-13.22^{***}	-8.85	-0.27^{***}	-0.42^{***}	15.73^{***}	19.78^{***}
	(0.04)	(0.04)	(3.72)	(6.91)	(0.07)	(0.05)	(4.06)	(5.11)
Elasticity			-0.35	-0.13			-0.46	-0.23
Dep. Var. Base Period Mean	0.05	0.03	185.39	318.06	0.10	0.09	127.19	204.06
N Clusters	303	292	303	293	281	268	281	272
N Obs.	3,734,294	1,196,117	10,148,532	2,945,193	1,742,719	414,172	6,135,611	1,358,357
N Individuals	1,894,374	558,811	3,566,538	923,617	1,167,470	284,204	2,649,628	606,131
N Exposed Individuals	636,066	219,862	1,609,450	522,605	118,441	24,724	497,783	110,446
By Socioeconomic Status								
	Low SES	High SES	Low SES	High SES	Low SES	High SES	Low SES	High SES
Jun. 2010	0.04	-0.00	-0.40	-2.78	0.12	-0.04	-1.32	-6.32
	(0.05)	(0.02)	(3.59)	(3.59)	(0.09)	(0.05)	(2.89)	(4.15)
Dec. 2010	0.01	-0.02	1.61	-1.59	0.08	-0.05	-1.03	6.47
	(0.04)	(0.02)	(3.09)	(4.22)	(0.08)	(0.04)	(2.55)	(4.53)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.22^{***}	0.20^{***}	-8.78^{***}	-21.31^{***}	-0.30^{***}	-0.32^{***}	9.27^{**}	18.78^{***}
	(0.05)	(0.03)	(2.58)	(4.82)	(0.07)	(0.05)	(3.05)	(5.47)
Elasticity			-0.41	-0.30			-0.41	-0.24
Dep. Var. Base Period Mean	0.07	0.02	95.12	347.84	0.16	0.05	75.44	243.48
N Clusters	303	302	303	302	278	283	279	285
N Obs.	1,943,879	2,986,532	6,999,869	6,093,856	941,603	1,215,288	4,617,114	2,876,854
N Individuals	1,201,766	1,347,265	2,768,287	2,021,242	721,292	755,356	2,021,269	1,378,643
N Exposed Individuals	366,368	489,560	1,109,738	1,022,317	94,855	48,310	421,652	186,577

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 3 over defined subsamples. Columns 1 through 4 are price and borrowing dif-in-dif effect estimates in the low cost market while columns 5 through 8 define the report estimates in the high cost market. Column headers report dependent variable at the top and subsample below. 'Elasticity' is borrowing effect scaled by base period outcome mean and price effect within each market-subsample. We take the log of '$\widehat{\text{Price}}$' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and price effect. '*N* exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Table VI: Difference in difference estimates on average costs

			Low cost market					*High cost market*			
	Registry Information	Pooled	No Mortgage	Have Mortgage	Low SES	High SES	Pooled	No Mortgage	Have Mortgage	Low SES	High SES
Jun. 2010	−0.00	0.02	0.03	-0.05^{*}	0.04	−0.00	0.07	0.11	−0.10	0.12	−0.04
	(0.05)	(0.03)	(0.04)	(0.02)	(0.05)	(0.02)	(0.08)	(0.08)	(0.06)	(0.09)	(0.05)
Dec. 2010	−0.01	0.00	0.01	-0.05^{+}	0.01	−0.02	0.04	0.06	-0.09^{+}	0.07	−0.05
	(0.05)	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(0.07)	(0.08)	(0.05)	(0.08)	(0.04)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	-0.09^{*}	0.02	0.01	0.03	0.01	0.03	−0.04	−0.03	−0.08	−0.06	−0.02
	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.06)	(0.07)	(0.05)	(0.07)	(0.05)
Elasticity		0.10	0.05	0.12	0.06	0.16	0.12	0.11	0.18	0.20	0.05
Dep. Var. Base Period Mean		0.04	0.05	0.03	0.07	0.02	0.10	0.10	0.09	0.16	0.05
N Clusters	329	303	303	292	303	302	284	281	268	278	283
N Obs.	5,468,917	4,930,411	3,734,294	1,196,117	1,943,879	2,986,532	2,156,891	1,742,719	414,172	941,603	1,215,288
N Individuals	2,630,314	2,385,366	1,894,374	558,811	1,201,766	1,347,265	1,433,629	1,167,470	284,204	721,292	755,356
N Exposed Individuals		855,928	636,066	219,862	366,368	489,560	143,165	118,441	24,724	94,855	48,310

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 3 where the dependent variable is average cost. 'Registry information' reports the estimated effect of deletion while other column headers only define subsamples. Columns 2-6 are estimated over the low cost market (as defined by exposure) while columns 7-11 are over the high cost market. We take the log of 'Average cost' when estimating the regressions but report the its base period mean in levels. 'Elasticity' is average cost effect scaled by the price effect. '*N* exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Table VII: Distribution of deletion effects

	Separate	Pooled	Difference
Low cost market			
Price	0.029	0.035	0.006
Average Cost	0.029	0.029	0.001
New Borrowing	251.561	237.796	-13.765
Welfare Loss	0.162	0.346	0.185
Aggregate New Borrowing	516	488	-28
Aggregate Welfare Loss	331,633	710,368	378,736
N Individuals	2,051,138	2,051,138	2,051,138
High cost market			
Price	0.069	0.035	-0.034
Average Cost	0.069	0.064	-0.004
New Borrowing	112.713	140.302	27.589
Welfare Loss	0.155	0.039	-0.117
Aggregate New Borrowing	69	86	17
Aggregate Welfare Loss	95,169	23,617	$-71,552$
N Individuals	612,737	612,737	612,737
Aggregate			
Price	0.033	0.035	0.002
Average Cost	0.038	0.037	-0.000
New Borrowing	585	574	-11
Welfare Loss	426,802	733,985	307,183
N Individuals	2,663,875	2,663,875	2,663,875

This table describes changes in key welfare metrics before and following deletion, assuming a markup of 0%.

Table VIII: Difference in differences by default and exposure

	Exposure	Positive Default	Amt. Default	NCV	NDV	New Bank Default	High SES	Age	Female
Panel A: All Default Deleted									
Q1	-0.98	0.65	450	82	118	0.16	0.16	44.9	0.51
Q2	-0.14	0.41	420	255	362	0.12	0.31	44.0	0.48
Q3	0.35	0.13	495	417	813	0.06	0.33	43.9	0.39
Q4	0.71	0.06	486	203	458	0.06	0.20	44.0	0.44
Q5	1.52	0.27	768	91	251	0.05	0.16	43.1	0.46
Pooled	0.29	0.30	504	209	401	0.09	0.24	44.0	0.45
Panel B: Gender Deleted									
Q1	-0.65	0.24	476	165	544	0.07	0.23	42.6	0.41
Q2	-0.10	0.28	604	317	681	0.09	0.33	43.7	0.39
Q3	0.02	0.34	549	240	371	0.13	0.23	43.8	0.31
Q4	0.12	0.40	437	220	264	0.11	0.20	43.8	0.49
Q5	0.82	0.24	459	106	143	0.06	0.18	45.9	0.66
Pooled	0.04	0.30	504	209	401	0.09	0.24	44.0	0.45

Baseline borrowing and demographic characteristics by quantile of the exposure distribution measured in December 2011. Rows correspond to quintiles of the exposure distribution. '% exposure' is log change in predicted cost times 100. 'MV¿0' is an indicator variable for a registry default. 'MV' is mean registry default amount. 'NCV' and 'NDV' are mean consumer and total borrowing amounts in the quarter before deletion, respectively. 'Consumo' and 'Deuda' are consumer and total debt balances. 'Mayor90' is overdue loan balance, 'AC' is average costs, and gender is an indicator for female. SES is an indicator equal to one for presence in the highest SES category

Internet Appendix

A Additional Results

Figure IA1: Percent increases in welfare loss by markup over average costs

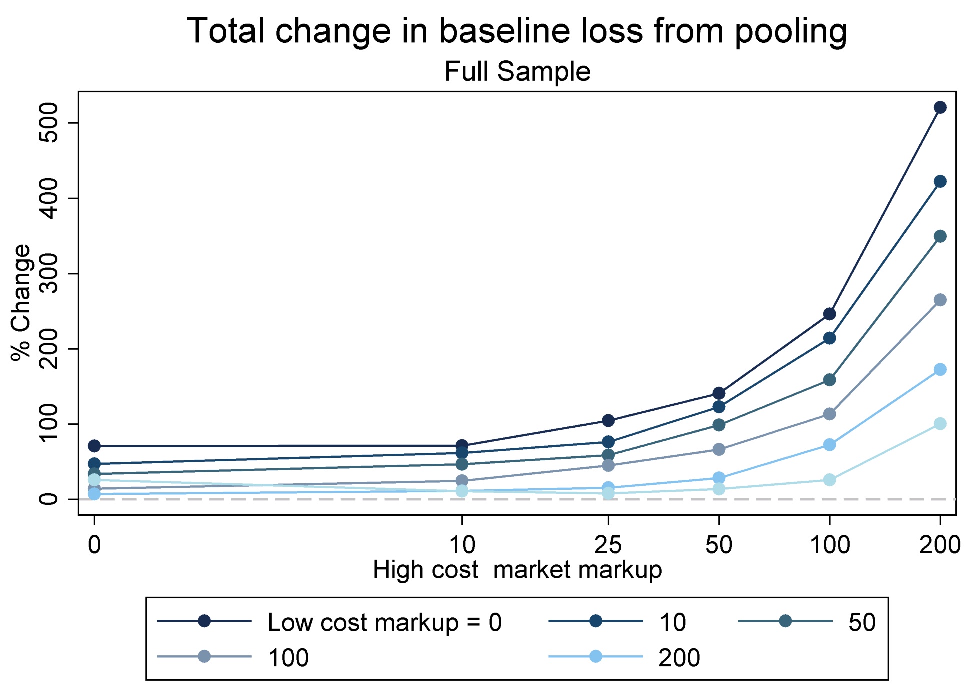
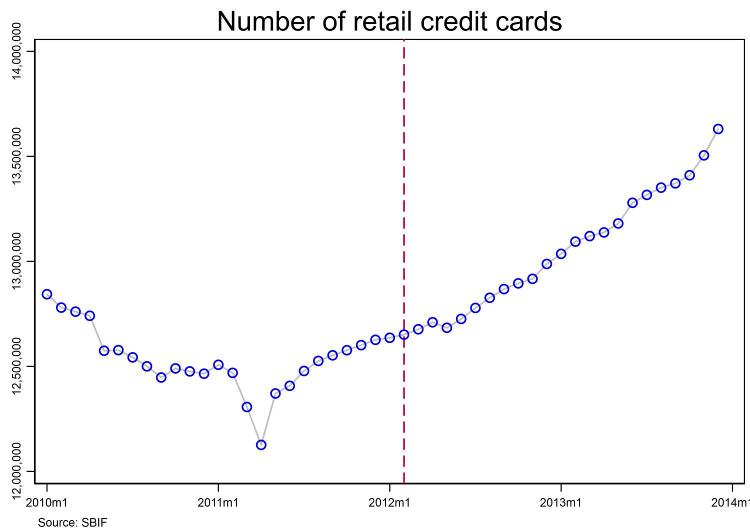


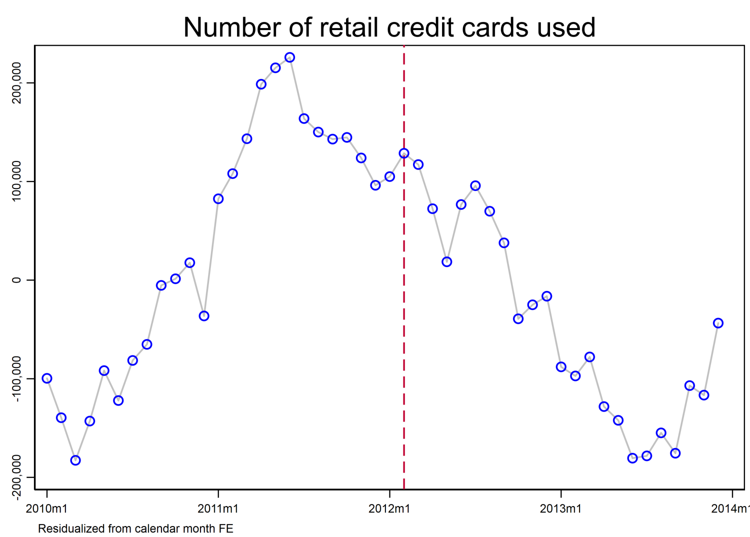
Figure displays percent changes in welfare relative to the efficient quantity from pooling the high- and low-cost markets at different markup levels of prices over average costs in each market. Percentage changes in welfare losses are measured relative to what is observed observed in the separate-markets equilibrium. Welfare losses in percentage terms are on the vertical axis, with higher values corresponding to larger losses. Each connected set of points corresponds to welfare losses at a given value of the markup in the low-cost market for a given value of the high-cost markup. Markups are computed relative to average costs in the high-cost and low-cost market. We set the markup in the pooled market as the quantity-weighted average of markups in the low- and high-cost markets, so average markup does not change with pooling.

Figure IA2: Stock of retail credit cards over time



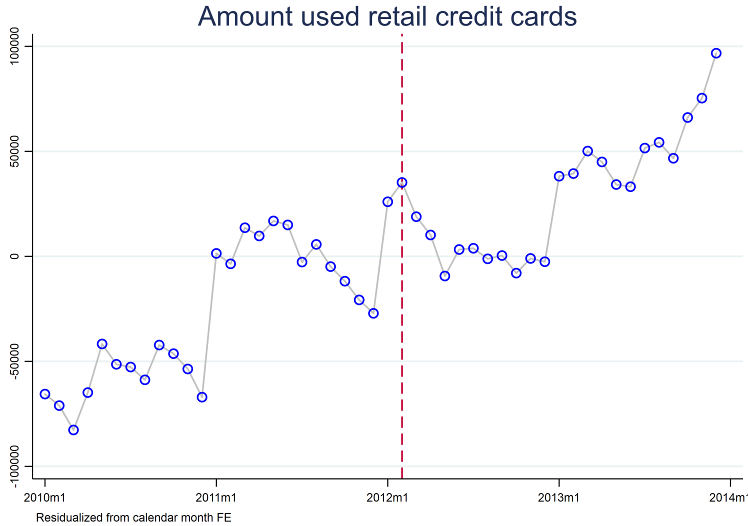
Stock of retail credit cards by month. Time of deletion policy noted with vertical line.

Figure IA3: Retail credit cards in use over time



Number of retail credit cards used by month. Time of deletion policy noted with vertical line. Source: SBIF.

Figure IA4: Number of retail credit card uses over time



Amount of retail credit purchases by month. Time deletion policy noted with vertical line. Source: SBIF.

Table IAI: Difference-in-difference predictions using long run cost measures

	Low cost market			*High cost market*		
	$\widehat{\text{Price}}$	Average Cost	New Borrowing	$\widehat{\text{Price}}$	Average Cost	New Borrowing
Jun. 2010	0.01	0.00	-7.09^{*}	0.03	0.03	-5.68^{+}
	(0.02)	(0.02)	(3.05)	(0.05)	(0.05)	(3.23)
Dec. 2010	0.01	0.01	-2.11	0.01	0.01	0.30
	(0.02)	(0.02)	(3.52)	(0.05)	(0.05)	(3.25)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.25^{***}	0.12^{***}	-13.28^{**}	-0.30^{***}	0.04	17.98^{***}
	(0.02)	(0.02)	(4.21)	(0.04)	(0.04)	(3.47)
Elasticity		0.48^{***}	-0.24^{***}		-0.12^{***}	-0.36^{***}
Dep. Var. Base Period Mean	0.08	0.08	214.70	0.14	0.14	165.09
N Clusters	307	307	307	299	299	300
N Obs.	4,961,674	4,961,674	13,163,613	2,519,339	2,519,339	8,117,207
N Individuals	2,394,399	2,394,399	4,373,700	1,571,258	1,571,258	3,422,263
N Exposed Individuals	765,941	765,941	1,967,865	134,306	134,306	589,628

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 3. Table is identical to Table IV but uses a one-year ahead measure of default to compute predicted costs. See section 5.4 for details. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. We take the log of '$\widehat{\text{Price}}$' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and price effect. '*N* exposed individuals' reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level.

Table IAII: Difference-in-difference predictions using logistic LASSO cost measures

	Low cost market			*High cost market*		
	$\widehat{\text{Price}}$	Average Cost	New Borrowing	$\widehat{\text{Price}}$	Average Cost	New Borrowing
Jun. 2010	0.61***	0.60***	−2.99	0.49***	0.49***	−2.56
	(0.11)	(0.11)	(5.04)	(0.07)	(0.07)	(2.57)
Dec. 2010	0.49***	0.48***	−12.59*	0.30***	0.30***	3.02
	(0.12)	(0.12)	(5.81)	(0.06)	(0.06)	(2.12)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.67***	0.32**	−26.48***	−0.35***	0.10+	15.56***
	(0.14)	(0.12)	(6.40)	(0.07)	(0.05)	(3.95)
Elasticity		0.48***	−0.18***		−0.28***	−0.49***
Dep. Var. Base Period Mean	0.04	0.04	220.85	0.18	0.18	90.60
N Clusters	298	298	306	123	123	123
N Obs.	4,732,430	4,732,430	12,397,248	1,073,283	1,073,283	6,004,974
N Individuals	2,314,089	2,314,089	4,219,921	845,732	845,732	2,451,377
N Exposed Individuals	1,137,949	1,137,949	2,511,392	196,199	196,199	778,351

Significance: + 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 3. Table is identical to Table IV but uses measures of default output from a logistic LASSO algorithm to compute predicted costs. See section 5.4 for details. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. We take the log of '$\widehat{\text{Price}}$' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and price effect. '*N* exposed individuals' reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level.

Table IAIII: Difference-in-difference predictions using naive Bayes cost measures

	Low cost market			*High cost market*		
	$\widehat{\text{Price}}$	Average Cost	New Borrowing	$\widehat{\text{Price}}$	Average Cost	New Borrowing
Jun. 2010	1.44***	1.44***	−9.37*	0.36	0.35	−7.84+
	(0.19)	(0.19)	(3.80)	(0.24)	(0.24)	(3.99)
Dec. 2010	1.41***	1.40***	−0.37	0.42+	0.41+	2.22
	(0.18)	(0.18)	(3.51)	(0.23)	(0.23)	(3.88)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	1.46***	0.66**	−10.43*	−2.95***	−0.17	17.40***
	(0.20)	(0.20)	(4.06)	(0.32)	(0.27)	(3.80)
Elasticity		0.45***	−0.03***		0.06***	−0.04***
Dep. Var. Base Period Mean	0.05	0.05	212.24	0.38	0.38	134.01
N Clusters	319	319	324	217	217	220
N Obs.	5,081,605	5,081,605	13,701,724	1,017,983	1,017,983	4,990,559
N Individuals	2,447,081	2,447,081	4,435,160	780,080	780,080	2,191,608
N Exposed Individuals	1,145,635	1,145,635	2,757,708	107,655	107,655	458,245

Significance: + 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 3. Table is identical to Table IV but uses measures of default output from a naive Bayes algorithm to compute predicted costs. See section 5.4 for details. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. We take the log of '$\widehat{\text{Price}}$' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and price effect. '*N* exposed individuals' reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level.

B Detail on the machine learning procedure

We generate cost predictions by regressing an innovation in default indicator against a large selection of features using a random forest algorithm. We create four sets of predictions trained on 10% of the data with new borrowing within each snapshot – approximately 8% of the overall data. Predictions are trained and predicted either within each 6-month post-December snapshot (AC^{post}), or only in the December 2009 snapshot (AC^{pre}). The random forests for each type are constructed with or without registry information. We use `python`'s `sklearn` package to perform our machine learning tasks (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot and Duchesnay 2011).

Our random forest regression design constructs regression trees using a feature vector of the following observable characteristics of each observation: a gender indicator, and one and two period lags of innovations in borrowing, innovations in total debt, total borrowing, total debt, average costs, and credit line information. We additionally include the default history deleted from the credit registry in some of the trees. In total, these trees have either thirteen or fourteen predictor variables.

We scale our features by binning their nonzero values into quartiles. This reduces noise in the feature vector and creates parsimonious regression trees. In our dataset, we find that this additionally decreases the time necessary to construct a random forest. Finally, we subset over only new borrowers in each period so that our cost estimates reflect costs conditional on borrowing.

To genearate our $\mathrm{AC}^{\mathrm{pre}}$ predictions, we train a model only using observations in the December 2009 snapshot. $\mathrm{AC}^{\mathrm{post}}$ predictions are generated using a training sample from each snapshot; these predictions are actually generated using a suite of models each tied to a particular snapshot.

We use three-fold cross validation combined with a grid search to pick parameters for each model. The parameters over which we search are the minimum number of observations in a terminal node (*minleaf*) and the number of features over which each tree can sample. We set the number of trees in a forest to 150. Predictive power is not sensitive to choices in this range. See figure IA5 and IA6 to see outcomes from this procedure.

Constructing random forests is (generally) a supervised learning task. Breiman (2001) defines a random forest a set of regression trees, $h_k = h(x, \Theta_k)$ where h is a tree and Θ_k is a random selection of observations and features from the training data, where each tree "votes" on the output given an observation. We pick splits in the data to reduce mean-squared error, as is common with regression tasks. We use this loss function and a regression task, despite our target variable existing only in $\{0, 1\}$, to ensure that our outputs are continuous on $[0, 1]$ and reflect probabilities. Our predictions are best thought of as a weighted average of default rate in pools of observations clustered together by similarity along a set of their covariates.

We additionally estimate a regression tree[15] to bin borrowers into smaller markets. We define a market as a set of observations M such that $h(x_i, \Theta)$ returns a prediction stemming from the same terminal node for all $i \in M$. We use this method to cluster borrowers into borrowers with similar features and default rates. These clusters therefore represent infered groups in the data at the level which we believe the treatment is applied and are analagous to the clusters defined in each tree in the forest.

Finally, we recreate the analysis above, exchanging the random forest algorithm for two other machine learning procedures that return classification probabilities. These are a naive Bayes classifier and a logistic LASSO. Our naive Bayes classifier first bins nonzero values along the feature vector into quartiles. Under the naive assumption of independence of features in the feature vector, the classifier constructs $P(\text{default}|X)$ using Bayes' formula under the assumption that $P(X|\text{default})$ is Gaussian, though this is functionally irrelevant due to binning.

For the logistic LASSO, we take the log of nonzero values of continuous features, generating a flag for zeros. We perform a logistic regression with a λ penalty term of the sum absolute value of the coefficients and use three-fold cross validation to pick λ for each model; see figure IA7.

Finally, we classify observations' socioeconomic status by training a random forest classifier on observations for whom the bank defined socioeconomic status group. Our three-fold cross validation procedure indicates that we are able to do this with approximately 35% accuracy using a random forest composed of 100 trees and built on a feature vector consisting of continuous measures of consumer debt, mortgage amount, debt balance, credit line, bank default, average cose, age, total default amount, and indicators for gender, new borrowing, and having positive borrowing cap.

[15]We estimate CART-style regression trees that split using variance reduction (Breiman, Friedman, Stone and Olshen 1984).

Figure IA5: Cross-validation output for AC^{pre} random forest predictions

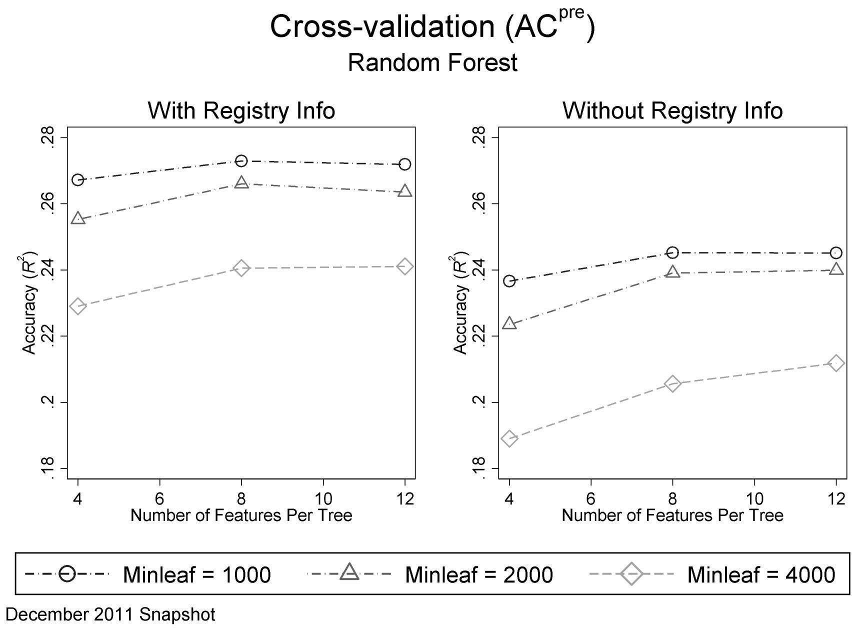
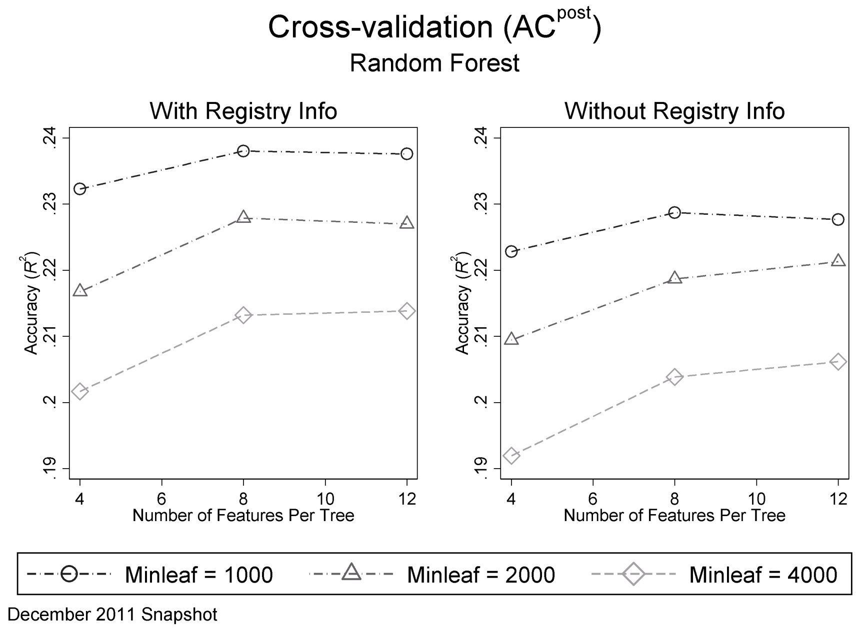


Figure IA6: Cross-validation output for AC^{post} random forest predictions



C Model details

For the high and low cost markets, indicated by their associated default flag $z \in \{0,1\}$, we can compute welfare loss by the following equations:

$$\text{(Inverse) Demand:} \quad q = \max\left(q_z^e + \frac{\mathrm{d}q_z}{\mathrm{d}p}(p - p_z^e),\ 0\right) \tag{5}$$

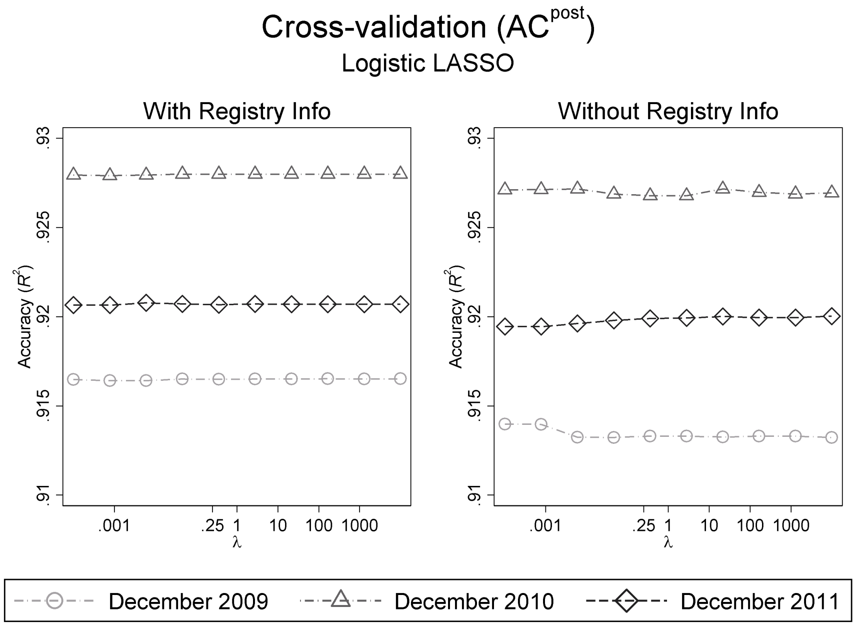
$$\text{Average Cost:} \quad p = \max\left(AC_z^e + \frac{\mathrm{d}AC_z}{\mathrm{d}q}(q - q_z^e),\ 0\right) \tag{6}$$

$$\text{Marginal Cost:} \quad p = \max\left(MC_z^e + 2\frac{\mathrm{d}AC_z}{\mathrm{d}q}(q - q_z^e),\ 0\right). \tag{7}$$

Absent average cost pricing, the efficient equilibrium equates marginal cost with demand. First we consider interior solutions given by the pair (p_z^*, q_z^*) that solves,

$$q = q_z^e + \frac{\mathrm{d}q_z}{\mathrm{d}p}(p - p_z^e)$$
$$p = MC_z^e + 2\frac{\mathrm{d}AC_z}{\mathrm{d}q}(q - q_z^e).$$

Figure IA7: Cross-validation output for AC^{post} logistic LASSO predictions



By plugging (7) into (5), we find the interior solution is,

$$q_z^{*\prime} = q_z^e + \frac{p_z^e - MC_e^z}{2\frac{\mathrm{d}AC_z}{\mathrm{d}q} - \left(\frac{\mathrm{d}q_z}{\mathrm{d}p}\right)^{-1}}$$

$$p_z^{*\prime} = \frac{MC_z^e - 2\frac{\mathrm{d}AC_z}{\mathrm{d}q}\frac{\mathrm{d}q_z}{\mathrm{d}p}p_z^e}{1 - 2\frac{\mathrm{d}AC_z}{\mathrm{d}q}\frac{\mathrm{d}q_z}{\mathrm{d}p}}.$$

Note, however, that these are unconstrained by the strict positivity. When $p_z^{*\prime} < 0$, $p_z^* = 0$. Therefore,

$$q_z^* = \begin{cases} q^{*\prime}_z & p^{*\prime} \geq 0 \\ q_z^e - \frac{\mathrm{d}q_z}{\mathrm{d}p}p_z^e & p^{*\prime}_z < 0 \end{cases}$$

Deadweight loss is computed as the difference in total surplus between market prices given by average cost and marginal cost pricing. Broadly, this is the area of the triangle formed below demand and above marginal cost between the points where average cost and marginal cost curves cross demand. However, if marginal cost is zero before it crosses demand, the traditional formulation,

$$DWL_z = \frac{1}{2}(q_z^* - q_z^e)(p_z^e - MC_z^e) \tag{8}$$

fails to account for the additional surplus lost, stemming from consumers who would have purchased quantities between q_z^e and q_z^* at prices below MC_z^e that exist when marginal cost goes to zero before it crosses demand. Assuming downwards sloping average and marginal cost curves, the proper formulation is then,

$$DWL_z = \frac{1}{2}(q_z^* - q_z^e)(p_z^e) - \frac{1}{2}(q_z^{0\prime} - q_z^e)(MC_z^e),$$

where $q_z^{0\prime}$ is the minimum of q_z^0, the smallest quantity at which marginal cost is zero, and the competitive equilibrium quantity q_z^*. If marginal cost is zero after it crosses demand, this gives the standard formula given in (8). Further, $q_z^0 \not< q_z^e$, and if it were, then $MC_z^e = 0$, assuming linearity. Note,

$$q_z^0 = q_z^e - \frac{MC_z^e}{2}\left(\frac{\mathrm{d}AC_z}{\mathrm{d}q}\right)^{-1}.$$

In the pooled case, we compute:

$$DWL_Z^{\text{pool}} = \frac{1}{2}(p^{\text{pool}})(q_z^* - q_z^{\text{pool}}) - \frac{1}{2}(MC_z^{\text{pool}})(q_z^0 - q_z^{\text{pool}}).$$

Adjustments are necessary to these formulas when cost curves slope upwards, in which case average cost price may give rise to an automatic surplus gain for buyers.

XXX

At equilibrium,

$$\begin{aligned} AC^* = {} & \frac{\sum_{z\in\{0,1\}}(Q_z^e - 2AC_z^e\frac{dQ_z}{dP})(1 - \frac{dAC_z}{dQ}\frac{dQ_z}{dP})}{\sum_{z\in\{0,1\}} -2(\frac{dQ_z}{dP}(1 - \frac{dAC_z}{dQ}\frac{dQ_z}{dP}))} \\ & + \frac{\sqrt{-4(\sum_{z\in\{0,1\}}(1 - \frac{dAC_z}{dQ}\frac{dQ_z}{dP})(AC_z^e\frac{dQ_z}{dP}(AC_z^e\frac{dQ_z}{dP} - Q_z^e)) + ((Q_z^e - 2AC_z^e\frac{dQ_z}{dP})(1 - \frac{dAC_z}{dQ}\frac{dQ_z}{dP}))^2)}}{\sum_{z\in\{0,1\}} -2(\frac{dQ_z}{dP}(1 - \frac{dAC_z}{dQ}\frac{dQ_z}{dP}))} \end{aligned}$$