Journal of Economic Theory and Econometrics, Vol. 34, No. 2, Jun. 2023, 51–72

The Effect of the Franchises on Local Bakeries with High-Dimensional Controls*

Soyul Lee† Chang Sik Kim‡

Abstract Since the 2013 designation of the bakery industry as a suitable business for small- and medium-sized enterprises in South Korea, there have been persistent questions about the validity of the designation and the suitability of the 500-meter store distance limits. This paper analyzes the effect of the entrance of a large franchise bakery on the closing rates of neighboring rival bakery using the Cox proportional hazard model. Our empirical results reveal that standard one-step variable selection like Lasso selection can lead to omitted variable bias in the Cox proportional hazard model, since it may exclude possible confounding variables from the estimation. We first show that the double-Lasso (Least Absolute Shrinkage and Selection Operator) selection method proposed by Belloni *et al.* (2014) can resolve the omitted variable biases through the Monte-Carlo simulations. The empirical results using the double-Lasso selection procedure suggest that the number of franchise bakeries within 200 meters of a bakery significantly increases the closure rate of that bakery, but that this effect becomes insignificant when we restrict our analysis to small- and medium-sized bakeries.

Keywords Suitable Business for SMEs, Entry Restriction, Many Covariates, Double-Lasso Selection, Cox Proportional Hazard Model

JEL Classification C55, L50

*We are grateful to the two anonymous reviewers for their thoughtful comments and suggestions.

†Graduate Student, Department of Economics, Sungkyunkwan University, E-mail: soyullee94@gmail

‡Corresponding Author, Department of Economics, Sungkyunkwan University, E-mail: skimcs@skku.edu

Submitted March 23, 2023, Accepted June 9, 2023

1. INTRODUCTION

Over the last several years, while small non-franchise bakeries have faced increasing operational difficulties due to financial constraints, the corporate franchise bakery industry has increased its market share based on its financial strength and marketing. The sales of franchise bakeries rose from 962.1 billion won in 2007 to 3.603 trillion won in 2016, and the market share of franchise bakeries in the bakery industry has risen from 40% to 60% over the same period. The number of franchise bakeries has also increased significantly. Between 2007 and 2016, Paris Baguette, which has the largest market share in the industry, expanded its number of franchisees from 1,350 to 3,420. Tous Les Jours, which has the second largest market share after Paris Baguette, also increased its number of franchisees from 886 to 1,323 over the same period.[1]

The Korea Bakers Association, which has owners of bakeries as its members, claims that the rapid increase in the number of large bakery franchises has uprooted existing confectionery technicians. According to the association's estimate, the number of small bakeries dropped sharply from 18,000 in the early 2000s to 5,000 by 2011. The association has claimed that the struggles faced by small bakeries were primarily due to excessive encroachment of franchise bakeries into small business districts, and it has therefore urged that the bakery market should be designated as a suitable business for small- and medium-sized enterprises (SMEs). Taking the association' claim into consideration, Korea Commission for Corporate Partnership (KCCP) selected the bakery industry as a suitable business for SMEs in 2013 to protect small bakeries. As a result, KCCP advised corporate franchise bakeries to refrain from expanding their businesses. Specifically, franchise and in-store bakeries larger than SMEs were allowed to open new franchises representing within 2% of the number of existing stores each year. They were also restricted from opening new franchisees within 500m of small bakeries. The bakery industry was again designated as a suitable business for SMEs in 2016, thus limiting large corporations from entering the market until 2019. Since the expiration of the designation, large franchises have signed an agreement with the Korea Bakers Association and operated under the same restrictions as when the bakery industry was designated as a suitable business

[1]Korea Agro-Fisheries & Food Trade Corporation (2011). A Survey on the Market Status of Processed Food Segmentation: Bakery Market (11-1541000-000745-01). https://www.at.or.kr/home/apko000000/index.action
Korea Agro-Fisheries & Food Trade Corporation (2018). A Survey on the Market Status of Processed Food Segmentation: Bakery Market (11-1543000-002285-01). https://www.at.or.kr/home/apko000000/index.action

for SMEs.

However, there are questions regarding the effectiveness of the designation of SMEs in terms of protecting non-franchised bakeries against franchised bakeries. The distance restriction guidelines are particularly controversial because even the regulators could not provide a clear basis for calculating the distance limit of 500m. If the distance limit were too large, then more bakeries could come in and compete, but their entries would have been blocked. Meanwhile, if the distance limit were too small, then the policy would fail to achieve its goal of protecting small bakeries.

This paper discusses the validity of designating the bakery industry as a suitable business for SMEs and the suitability of the 500m store distance limits. With its particular focus on distance restriction guidelines, this paper aims to determine whether large franchise bakeries have a significant effect on the closure of nearby bakeries. Specifically, this paper analyzes the effect of the entrance of a large franchise bakery on the closure of a neighboring rival bakery depending on the distance between two. This paper considers the characteristics of bakeries and commercial districts from two sources: food sanitation establishments in Seoul and Seoul's side street trade areas (SSTA) data. To control for the differing business environments for each bakery, we only use the bakeries that were on the SSTA between 2015 and 2019. The final data includes 1078 bakeries that were either closed between 2015 and 2019 or open until 2019. For business district level controls, 120 covariates including floating population, working population, and price of an apartment are used to control for environment effects on the survival of a bakery.

To accurately determine the closure rate of non-franchised bakeries, the selection of covariates is an important part of addressing the issues related to the large dimensional modeling. Including all covariates may result in overfitting of the data even if there are no large dimensionality problems. However, excluding relevant covariates that affect the survival of a bakery and the number of nearby competitors can generate an omitted variable bias. Hence, regularization of the model to avoid overfitting problems and the selection of relevant variables to prevent the omitted variable biases is crucial and represents a challenging task for correct model specification. In this paper, we employ the "double-Least Absolute Shrinkage and Selection Operator (Lasso)" estimators proposed by Belloni *et al.* (2014) for exogenous regressors to deal with the selection of relevant covariates.

Lin *et al.* (1998) demonstrates how the omission of variables in Cox proportional hazard (PH) regression causes bias in the resulting estimates. We pose the

problem in the framework of a Cox PH model as

$$\lambda_i(t|d_i,x_i) = \lambda_0(t)exp(d_i\alpha_0 + x_i'\theta_g),$$

where d_i is the variable of interest and x_i is a set of control variables. The main goal of our analysis is to conduct inference on the effect α_0 based on the correct specification of the model, including the proper selection of x_i covariates in the large dimensional model.

Belloni *et al.* (2014) show that a serious omitted variable bias can be detected in the estimates of α_0 with conventional post-single Lasso selection techniques. This bias can be particularly severe when there are some control variables that are not highly correlated with the dependent variable but are highly correlated with d_i. This is mainly because single-selection procedures exclude those control variables from the model since they are selected to be useless for predicting dependent variables, thus confounding the effect of d_i. The double-Lasso selection technique proposed by Belloni *et al.* (2014) resolves this problem by including control variables that are highly correlated with d_i in the model even if they are not highly correlated with the dependent variable. Therefore, to reduce the potential omitted variables bias, a control variable is only omitted in the double-Lasso selection if it correlates with neither the dependent variable nor the variable of interest.

In this paper, we apply the post-double-Lasso techniques in the estimation of the closure rate of non-franchise bakeries using Cox PH regression, and we find that the closure rate of a bakery is significantly affected by the number of franchise bakeries within 200m. However, when we limit our sample to small and medium-sized bakeries, the estimated effect of franchise bakeries using the same methodology becomes insignificant. These mixed results imply that the number of franchise bakeries nearby does not significantly affect the closure of a small or medium-sized bakery. We also address the effect of the entrance of franchise agents with high-dimensional data, and show that the double-Lasso selection procedure proposed by Belloni *et al.* (2014) can even eliminate the omitted variable bias in the Cox PH model.

The rest of this paper is as follows. Section 2 discusses the related previous literature and Section 3 describes the data set we use in this paper. In Section 4, we present our main model and detail some Monte Carlo simulations. Section 5 provides an empirical analysis using our model and Section 6 presents the conclusion.

2. LITERATURE REVIEW

Our study is related to research focusing on the effects of the geographical clustering of firms on the closure rates of bakeries. Many previous empirical studies have discussed how endogeneity issues are crucial in the interpretation of existing empirical results. As shown by Shaver and Flyer (2000)and Alcácer (2006), smaller, less capable firms are more likely to be clustered than larger, more capable firms, so clustering and performance may appear to have a simple negative relationship. More recent research has largely addressed this issue using panel data, which identifies changes in clustering levels around a focal firm. In this field of research, endogeneity is also a substantial problem for obtaining valid estimation results regarding the effect of the number of nearby franchise bakeries on the closure of a bakery. Since bakeries tend to enter more attractive commercial districts, bakeries that survive longer are more likely to be clustered than bakers that do not survive as long. To control unobserved heterogeneity, this paper uses high dimensional panel data for the characteristics of each business districts and the number of nearby competitors for each quarter to avoid any possible omitted variable biases.

Second, this paper is related to variable selection literature in that it constructs a model among many controls through variable selection. Out of many model selection techniques in the literature, the most widely used techniques are stepwise selection (Efroymson, 1966), Akaike Information Criterion (AIC) (Akaike, 1973, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978), Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani, 1996), and Adaptive Lasso (Zou, 2006). Lasso maintains a balance between larger and smaller models, but it does not possess the oracle property (Fan and Li, 2001; Fan and Peng, 2004). Meanwhile, Adaptive- Lasso obtains a convex objective, thus yielding oracle estimators by using a weighted L_1 penalty with weights determined by an initial estimator.

This paper aims to explore several model selection criteria including Lasso, Adaptive-Lasso, and double-Lasso for model selection among a lot of characteristics for each business district and bakery. Our study applies these model selection techniques to the Cox Proportional Hazard (PH) model to correctly select controls that affect the failure rate of a bakery.Belloni and Chernozhukov (2013) propose a post-double-Lasso selection procedure that first uses a Lasso regression to select covariates correlated with the outcome and then selects covariates correlated with the variable of interest. A final ordinary least squares regression includes the union of two sets of covariates, thus improving the properties of the estimators. We apply this procedure to properly capture the effect of

nearby competitors on the exit rate of non-franchise bakeries. In particular, this paper shows how the properties of the double Lasso selection estimator are improved compared to those of single selection estimators (Lasso, Adaptive-Lasso) in the Cox PH model by referring to Lin *et al.* (1998).

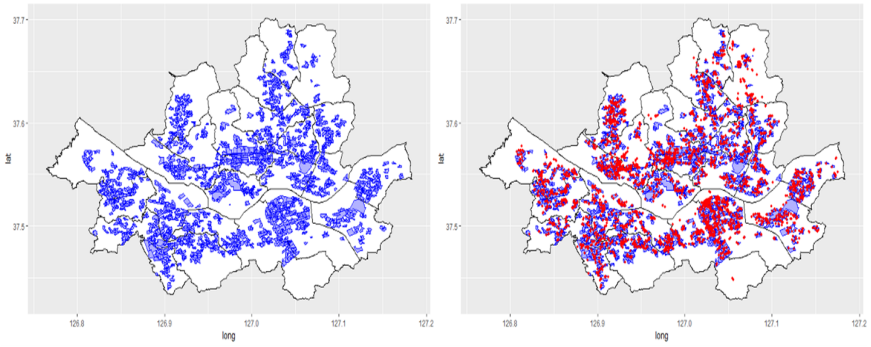
Our paper contributes most directly to the literature aiming to assess the effect of the entrance of franchise agents in a market on the survival of incumbent agents. The explosive expansion in market share of franchises has continually led to problems for other businesses and resulting controversy, but there have been very few studies examining this. Nam (2017) showed that the number of competitors and business-specific characteristics such as the age and size of a business significantly affect the shutdown rates of self-employed businesses. Choi *et al.* (2020) showed that FTC restrictions on the coffee franchise industry in South Korea significantly affect the closure rates of coffee shops. Yang (2016) addressed the effects of the franchise bakeries, like our analysis, but only with a small number of pre-selected control variables. Kang *et al.* (2018) also showed that franchise bakeries belonging to larger or older franchisors have a higher chance of survival, thus implying the higher exit rate of non-franchise bakeries.

3. DATA

This paper uses two data sources to obtain the characteristics of bakeries and commercial districts. First, we use the individual status of food sanitation establishments for the characteristics of each bakery. Local governments in Korea disclose the statuses of food sanitation establishments including each bakery within their districts. For this paper, we extract data on food sanitation establishments in the bakery industry in Seoul. The information available includes the business name, total address, licensing date, closing date, total size of the facility, etc. In addition to typical bakeries, the data also includes businesses dealing with items such as rice cake, cookies, waffle, and hamburgers that are somewhat considered to be alternatives to baked bread; we excluded these businesses from the analysis. We also excluded bakeries located within event stores and department stores from the analysis.

To control for the characteristics of business environments, we use data for Seoul's side street trade areas (SSTA) provided by the Seoul Metropolitan Government. SSTA data is used to support small business owners by providing information regarding the status of each area, such as the size of the floating population, nearby transportation facilities, and apartment prices around the area. Areas with dense housing and high concentrations of small businesses in Seoul

Figure 1: Seoul's Side Street Trade Areas (SSTA) and Bakeries in Seoul



are designated as SSTAs. Using the location information of bakeries, which we obtain from the status of food sanitation establishments, we pick out the bakeries that were on the SSTA between 2015 and 2019. Figure 1 shows the SSTA data and the coordinates of the bakeries, where blue and red dots indicate the SSTA and locations of bakeries, respectively. Most bakeries are located in the SSTA as expected.

Variable	Min.	Median	Mean	Max.	SD
T	1	20	29.4	216	28.41
Bakery200m	0	1	1.54	9	1.54
Bakery500m	0	5	5.32	23	3.25
Bakery800m	0	8	8.42	32	4.93
Fbakery200m	0	0	0.52	5	0.76
Fbakery500m	0	2	1.89	10	1.68
Fbakery800m	0	3	3.30	14	2.42

Table 1: Descriptive Statistics for Store-level Controls

Note: The unit duration is one quarter, and T represents the survival duration of bakeries. Bakery200m, Bakery500m, and Bakery 800m are the numbers of bakeries within 200m, 200m 500m, and 500m 800m of a bakery, respectively. "Fbakery" in variable names stands for franchise bakery.

The final data for this study includes 1078 bakeries that were in business for at least some period of time between 2015 and 2019. Among them, 334 bakeries went out of business during the same period. To control for the characteristics of individual bakeries, we prepare a dummy variable that shows whether a bakery

is a franchisee or in a retailer mart. Among 1078 bakeries, 302 are franchised bakeries and 40 are in-store bakeries in retailer marts. To measure the level of proximity to competitors, we count the number of in-radius bakeries. Specifically, we analyze the number of bakeries within 200m, between 200m and 500m, and between 500m and 800m. We also include year dummies in the explanatory variables to control for macroeconomic impacts on the closure of bakeries. Table 1 presents descriptive statistics for individual bakery level controls.

The original data for SSTA provides almost 1000 possible control variables for each district, but some variables do not sufficiently reflect the difference among areas enough or contain too specific level of information to use in the model estimation. Therefore, we delete and transform some controls as they can reveal the differences among areas clearly, thus decreasing the number of control variables to 120. Table 2 lists the district level control variates we utilize in this paper.

4. METHODOLOGY

4.1. OMITTED VARIABLE BIAS IN COX PROPORTIONAL HAZARD REGRESSION

In this section, we examine the bias of estimates caused by the omitted confounders in Cox PH regression by referring Lin *et al.* (1998) and show that this bias can be reduced by employing double-Lasso estimation, as we have discussed earlier. To analyze the impact of the number of nearby franchise bakeries on the failure rates of small and medium-sized bakeries, the simplest approach in Cox PH regression is to regress the failure rate on some measure of franchise bakery exposure and other covariates:

$$\lambda_i(t|d_i,x_i) = \lambda_0(t)exp(\alpha Fbakery500m_{it} + x'_{it}\beta), \tag{1}$$

In our case, $\lambda_i(t|d_i,x_i)$ is the failure rate of a bakery i, $Fbakery500m_{it}$ is the number of entire franchise bakeries within 500m of the bakery, and x_{it} is a vector of exogenous control variables. Because there are many control variables regarding a bakery and its neighborhood characteristics that should be considered in the model, the estimation of the Cox PH model in (1) involves some challenges related to dealing with high-dimensional data. Including all covariates would likely result in overfitting of the data, if not dimensionality problems. However, excluding covariates that correlate with both the survival of a bakery and the number of nearby competitors would introduce an omitted variable bias.

Variable Group	Number of Control Variables	Explanation
1. Controls for SSTA		
Floating population	32	Average floating population by gender, age group, and time zone in SSTA
Working population	6	Working population by gender and age group of the resident population in SSTA
Resident population	6	Resident population by gender and age group in SSTA
Apartment	6	Number of apartment households by area and market price in SSTA
Facility	8	Number of facilities that gather people in SSTA
2. Controls for Hinterland of SSTA		
Floating population	32	Average floating population by gender, age group, and time zone in hinterland of SSTA
Working population	6	Working population by gender and age group of the resident population in hinterland of SSTA
Resident population	6	Resident population by gender and age group in hinterland of SSTA
Apartment	6	Number of apartment households by area and market price in hinterland of SSTA
Facility	12	Number of facilities that gather people in hinterland of SSTA

Table 2: List of Seoul's Side Street Trade Areas (SSTA) Level Controls

Note: The Hinterland of SSTA is considered the 200m in-radius of SSTA.

Therefore, to deal with the model specification, we use the double-Lasso estimators proposed by Belloni *et al.* (2014) for exogenous regressors in our Cox-PH model. However, the double-Lasso in Belloni *et al.* (2014) was originally suggested in linear models, we need to consider the application of the double-Lasso approach in the Cox-PH model as follows.

Let T denote the survival time or failure time of interest and let D be the covariate of interest. Further, X represents confounding factors that affect D and T. Suppose that conditional on D, the confounder X is normally distributed with mean γD and unit variance.

$$X = \gamma D + \varepsilon, \quad \varepsilon \sim N(0,1) \tag{2}$$

The proportional hazards regression specifies that the hazard functions of T conditional on the sets of covariates (D, X) and (D) are, respectively,

$$\lambda(t|D,X) = \lambda_0(t)exp(\alpha D + \beta X) \tag{3}$$

and

$$\lambda(t|D) = \lambda_0^*(t)exp(\alpha^* D), \tag{4}$$

where $\lambda_0(\cdot)$ and $\lambda_0^*(\cdot)$ are arbitrary baseline hazard functions and (α,β)and (α^*) are unknown regression parameters. We want to ascertain the relationship between α and α^*, which represent the effect of the omitted variables in the Cox-PH model.

Let $F(x|D)$ be the distribution function of X given D. Also, let $f(t|\cdot)$ and $S(t|\cdot)$ denote the conditional density and survival functions of T, respectively. By elementary probability arguments,

$$\lambda(t|D) = \frac{f(t|D)}{S(t|D)} = \frac{\int_{-\infty}^{\infty} f(t|D,X)dF(x|D)}{\int_{-\infty}^{\infty} S(t|D,X)dF(x|D)}. \tag{5}$$

Under model (3),

$$\int_{-\infty}^{\infty} f(t|D,x)dF(x|D) = \int_{-\infty}^{\infty} \lambda_0(t)e^{\alpha D+\beta X} \times exp(-\Lambda_0(t)e^{\alpha D+\beta X})dF(x|D)$$

$$\int_{-\infty}^{\infty} S(t|D,x)dF(x|D) = \int_{-\infty}^{\infty} exp(-\Lambda_0(t)e^{\alpha D+\beta X})dF(x|D)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. Thus, Equation (5) becomes

$$\lambda(t|D) = \lambda_0(t)exp(\alpha D)h(t;D), \tag{6}$$

where

$$h(t;D) = \frac{\int_{-\infty}^{\infty} e^{\beta X} \times exp(-\Lambda_0(t)e^{\alpha D+\beta X})dF(x|D)}{\int_{-\infty}^{\infty} exp(-\Lambda_0(t)e^{\alpha D+\beta X})dF(x|D)}.$$

Under the Normality assumption of X, it can be shown that

$$h(t;D) = exp(\beta\gamma D + 0.5\beta^2)\frac{\int_{-\infty}^{\infty} exp(-\Lambda_0(t)e^{\beta^2+\alpha D+\beta X})exp(-\frac{(x-\gamma D)^2}{2})dx}{\int_{-\infty}^{\infty} exp(-\Lambda_0(t)e^{\alpha D+\beta X})exp(-\frac{(x-\gamma D)^2}{2})dx}. \tag{7}$$

The numerator of equation (7) can be transformed into $\int_{-\infty}^{\infty} exp\{-\Lambda_0(t)(e^{\beta^2}-1)e^{\alpha D+\beta X}\}exp\{-\Lambda_0(t)e^{\beta^2+\alpha D+\beta X}\}exp\{-\frac{(x-\gamma D)^2}{2}\}dx$. Since $exp\{-\Lambda_0(t)(e^{\beta^2}-1)e^{\alpha D+\beta X}\}$ goes faster to 1 than $exp(\beta\gamma D+0.5\beta^2)$ when $\beta \to 0$, $h(t;D)$ can be approximated by $exp(\beta\gamma D+0.5\beta^2)$ if $|\beta|$ is small. It then follows that

$$\begin{aligned}\lambda(t|D) \approx \lambda_0(t)exp(\alpha D)h(t;D) &= \lambda_0(t)exp(\alpha D+\beta\gamma D+0.5\beta^2) \\ &= \lambda_0(t)exp((\alpha+\beta\gamma)D+0.5\beta^2). \end{aligned} \tag{8}$$

Therefore, we have

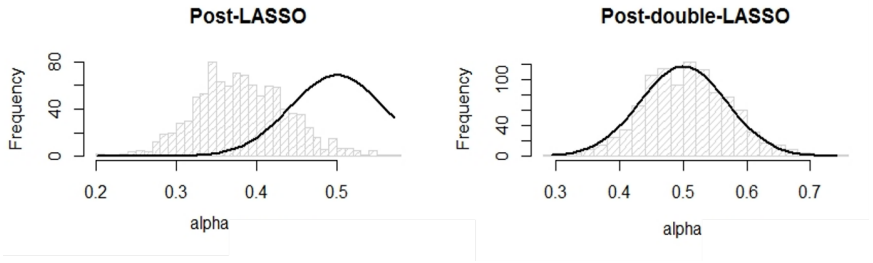
$$\alpha^* \approx \alpha+\beta\gamma, \tag{9}$$

which means that the post-single selection method can work poorly due to the omitted variable bias caused by dropping X when β is small enough even if γ is large.

In terms of the bias, $\beta\gamma$, the post-double selection technique should be able to resolve the problem. The basic concept of this technique is to use Lasso regression to select covariates that are most important for explaining the dependent variable as well as the explanatory variables of interest. In the most straightforward approach, the union sets of the selected variables could then be used as covariates in the main regression. In our context, the post-double selection method selects variables with two equations that contain the information from equations (2) and (3), estimating α with D and the union of the selected controls. In doing so, X is only omitted if its coefficient is small in both equations, which substantially limits the potential for omitted variables bias.

The Figure2 shows the finite sample distributions of α by the Monte-Carlo simulation that we extend the simulation scheme in Belloni *et al.* (2014) into a framework of Cox model, and the detailed descriptions of the simulation are given in Section 4.2. The right panel in Figure 2 shows the result of equation (3) where the model includes both the variable of interest and confounders. The figure shows that the finite-sample distribution of the post-double-Lasso estimator is close to a normal distribution. In contrast, the left panel in Figure 2 illustrates the problem with the traditional post-single selection estimator based on

Figure 2: Finite-sample Distributions of the Post-single Selection and the Post-double Selection



equation (4), as it can be seen that its distribution sharply deviates from the true α.

4.2. MONTE CARLO SIMULATIONS

In this section, we examine the finite-sample properties of the post-double-selection method and compare its performances to those of a standard post-single-selection method in the Cox model. The simulation schemes and the presentation format of the results in this section are the similar to those in Belloni *et al.* (2014) except that we consider the Cox PH model rather than linear models.

All of the simulation results are based on the model as

$$\lambda_i(t|d_i,x_i) = \lambda_0(t)exp(d_i\alpha_0 + x_i'\theta_g), \tag{10}$$

and

$$d_i = x_i'\theta_m + v_i, \tag{11}$$

where $v_i \sim N(0,1)$, the covariates $x_i \sim N(0,\Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$, $\alpha_0 = 0.5$, and the sample size n is set to 1000. The dimension of x_i is set to be $p = 200$. We set $\theta_{g,j} = c_y\beta_{0,j}$ and $\theta_{m,j} = c_d\beta_{0,j}$ with $\beta_{0,j} = (\frac{1}{j})^2$ for $j = 1,\cdots,200$. We can transform model (10) as

$$log(\frac{\lambda_i(t|d_i,x_i)}{\lambda_0(t)}) = d_i\alpha_0 + x_i'\theta_g \tag{12}$$

The constants c_y and c_d are chosen to generate the desired population values for the reduced form R^2's, i.e., the R^2's for equations (11) and (12). For each equation, we choose c_y and c_d to generate $R^2 = 0, 0.2, 0.4, 0.6$, and 0.8.

We report results for three different procedures: Two procedures are the standard post-single selection estimators–the post-Lasso and the post-adaptive-Lasso–which apply Lasso and adaptive-Lasso to equation (10) without penalizing α that is the coefficient on d_i, to select additional control variables from among x. Estimates of α are then obtained through Cox PH regression using d_i and the set of additional control variables selected in the Lasso step. For the post-double-Lasso, we run a Lasso of failure rate on x_i to select a set of predictors for failure rate $\lambda(t|d_i,x_i)$ and run a Lasso of d_i on x_i to select a set of predictors for d_i. α is then estimated by running a Cox PH regression of $\lambda(t|d_i,x_i)$ on d_i and the union of the sets of regressors selected in the two Lasso runs.

First Stage Structure	$R^2=0.2$ $R^2=0$		$R^2=0.2$ $R^2=0.8$		$R^2=0.8$ $R^2=0$		$R^2=0.8$ $R^2=0.8$	
Estimation procedure	RMSE	Size	RMSE	Size	RMSE	Size	RMSE	Size
Post-Lasso	0.131	0.539	0.092	0.128	0.410	0.987	0.094	0.158
Post-adaptive-Lasso	0.097	0.083	0.099	0.104	0.294	0.461	0.099	0.111
Post-double-Lasso	0.068	0.054	0.087	0.091	0.071	0.056	0.085	0.083

Table 3: Simulation Results for Selected R^2 Values

Note: The table reports root-mean-square-error (RMSE) and the size of the t-test for α under 5% significance level from a Monte Carlo simulation experiment. Results are based on 1000 simulation replications. Data are based on the model with coefficients that decay quadratically.

We start by summarizing the results in Table 3 for $(R_y^2,R_d^2)=(0,0.2)$,$(0,0.8)$, $(0.8,0.2)$, and $(0.8,0.8)$, where R_y^2 is the population R^2 in model (12) and R_d^2 in model (11). We report root-mean-square-error (RMSE) for estimating α and the size of 5% level tests. The results show that the post-double-Lasso procedure performs well without relying on ex-ante knowledge of the coefficients that go in the control functions, θ_g and θ_m. Meanwhile, the post-Lasso and post-adaptive-Lasso procedures generally do not perform as well as post-double-Lasso, and they are very sensitive to the value of R_d^2. While post-Lasso and post-adaptive-Lasso perform adequately when R_d^2 is small, their performances deteriorate quickly as R_d^2 increases. This lack of robustness is common in traditional variable selection methods such as Lasso that were designed for forecasting, not

causal inference.

We provide further details about the performance of the estimators in Figure 3 which plot the size of 5% level tests, bias for the post-Lasso, post-adaptive-Lasso, and post-double-Lasso estimators across the full set of R^2 values considered. The figures are plotted with the same scale for easier comparison, and the rejection frequencies for the post-Lasso were censored at 0.5 for improved readability. The most striking feature of the figures is the poor performances of the post-Lasso and post-adaptive-Lasso estimators. The two estimators have an order of magnitude more bias than the corresponding post-double-Lasso estimator, and they both perform poorly in terms of the size of tests across many different R^2 combinations. Specifically, the post-Lasso and post-adaptive-Lasso estimators do not reliably control size distortions or bias, as the controls except d_i are uncorrelated with the hazard rate (where second stage R^2 equals to 0) and the controls are highly correlated with d_i (where first stage R^2 is high). By contrast, the post-double-Lasso estimator performs relatively well across the full range of R^2 combinations considered.

The simulation results are favorable to the post-double-Lasso estimator. In the simulation, we can see that the post-double-Lasso procedure provides a valid estimate of the effect of d_i in the presence of many potential confounding variables. Overall, the simulation evidence supports why we need to use the post-double-Lasso estimator to obtain valid estimates in the presence of many potential confounding variables in a Cox PH model. Meanwhile, it also shows that the standard post-single selection procedure provides poor inference and is therefore not a reliable tool for empirical research when there possibly exist many potential confounding variables in various duration modeling.

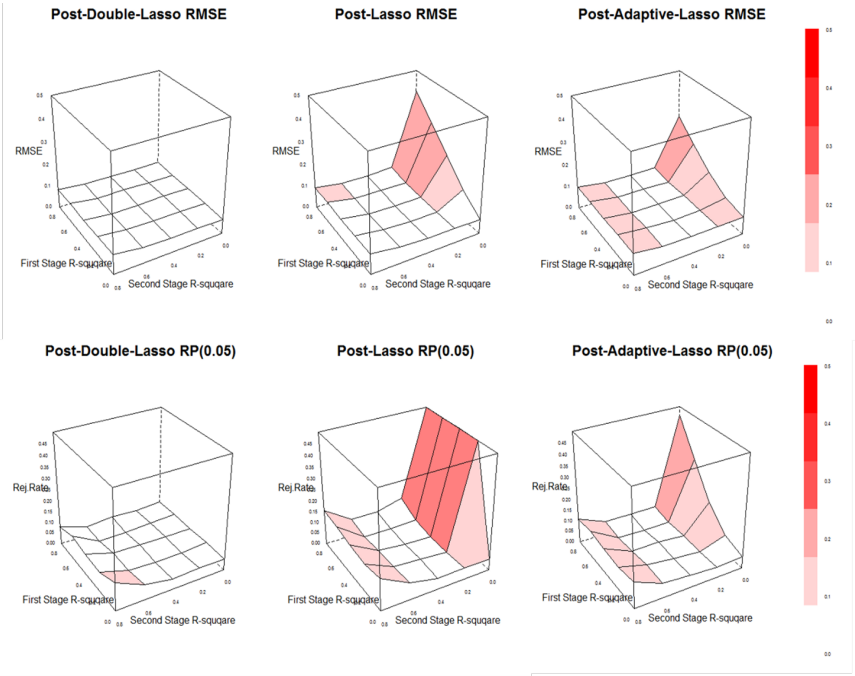
5. EMPIRICAL RESULTS

5.1. EFFECT ON THE CLOSURE OF ALL TYPES OF BAKERIES

The first set of results assesses the effect of the nearby entrance of franchise bakeries on the failure of all types of bakeries. To see how the impact of franchise bakeries changes with distance, we estimate the effect as changing the variable of interest from the number of franchise bakeries within 200m to the number of franchise bakeries between 500m and 800m. Using Lasso, adaptive-Lasso, and double-Lasso variable selection methods, we examine the differences in estimates.

As we have simulated in the preceding section, for the post-single selection procedures, variable selection is done on all controls except the variable of the

Figure 3: Rejection Frequencies for 5% Level Test and RMSE for Estimating the Effect of d_i



Note: This figure presents the root-mean-square-error (RMSE) and rejection frequency results for the 5% level test for estimating the effect of d_i from the simulation study. Results are reported for a post-double-Lasso selection procedure, a one-step post-Lasso estimator, and a one-step post-adaptive-Lasso estimator. The reduced form and first stage R^2 correspond to the population R^2 of model (12) and model (11), respectively. Note that the rejection frequencies are censored at 0.5.

interest.[2] Since this paper aims to estimate the effect of franchise bakeries, we do not report estimates on the other variables. Based on the variable selection methods, there are three major selected variables: a dummy variable for whether a bakery is in a mart, a dummy variable for whether a bakery is a franchisee, and the number of subway stations in a SSTA. All these variables are significant, and the signs of their coefficients are all reasonable. The closure rate of a bakery shop in a large supermarket (in-store bakery shop) is estimated as being higher than other bakery shops. This is potentially because in-store bakeries have often signed a contract for a given operating period regardless of its business profits. It is also valid that franchise bakeries have higher survival rates than independent bakeries considering that they receive management strategies for survival from corporate headquarters. The negative sign of the number of subway stations is sensible because it would increase accessibility.

As we explained in the preceding section, for the double-Lasso selection, we conduct an additional Lasso regression to select covariates correlated with the variable of interest. Since the number of bakeries in a market is closely related to the market conditions, over 70 covariates are typically selected in this stage. However, the inclusion of too many controls in a model can cause overfitting, thus leading to high model variance. Therefore, we limit the maximum number of variables in the model to 60 to guarantee model stability.[3]

Table 4 presents the coefficient estimates for the number of franchise bakeries in each radius. When using both post-Lasso selection and post-adaptive-Lasso selection, all estimated coefficients are insignificant, which means that the number of franchise bakeries does not affect the closure rate of a nearby bakery. However, with post-double-Lasso selection, it is confirmed that the number of franchise bakeries within 200m significantly increases the failure rate of a bakery. Further, the size of the coefficient is much larger than those in other selection methods. The results can be interpreted as showing that an additional franchise bakery within 200m of a bakery increases the quarterly rate of closure of the original bakery by about 19%.

In fact, 14 variables are selected in the Post-Lasso selection and 15 variables are selected in the Post–Adaptive Lasso Selection, and Double-Lasso selection add 51 more variables. [4] Given that the bias of the single-Lasso selection

[2]When attempting one-step variable selection even with the variable of interest, the number of franchise bakeries within 200m of a bakery continued to be selected by the Lasso and adaptive-Lasso variable selection techniques.

[3]Vittinghoff and McCulloch (2007) reveals that 5-9 events per predictor variable (EPV) in Cox regression do not show severe problems in estimation, compared to 10-16 EPV.

[4]We do not report the detailed set of variables in the paper, but selected variables with Post-

Variable Selection Procedure	Franchise bakeries in 200m	Franchise bakeries in 200m 500m	Franchise bakeries in 500m 800m
Post-Lasso-Selection	0.1236	-0.0117	0.0371
	(0.0968)	(0.0493)	(0.0252)
Post-Adaptive-Lasso-Selection	0.1089	0.0033	0.0225
	(0.0946)	(0.0347)	(0.0240)
Post- Double-Lasso-Selection	0.1733**	-0.0247	0.0611
	(0.1038)	(0.0524)	(0.0392)

Table 4: Estimated Effects of the Number of Franchise Bakeries on the Closure Rates of Nearby Bakeries

Note: Coefficient estimates for the number of franchise bakeries in radius. We regard the number of franchises within 200m of a bakery as the variable of interest in the first column. The number of franchise bakeries between 200m and 500m is in the 2nd column and the number of franchise bakeries between 500m and 800m is in the 3rd column. The coefficients in the 1st row are estimated using the post-Lasso selection technique while those in the 2nd row are estimated using post-adaptive-Lasso selection. The estimation is conducted using the Belloni *et al.* (2014)estimator for the last row. The values in parentheses are standard errors(Non-clustered). *,**,*** indicate 10, 5, and 1 percent significance levels. The number of observations is 12,760 with 1,078 bakeries.

or adaptive-Lasso selection comes from omitting controls that may have strong enough correlation with the key variables of interest, we try to find any correlation between omitted variables under the single-Lasso or adaptive-Lasso selection and the number of nearby franchise bakeries.[5] To see this, we first control the effect of selected 15 variables with Double-Lasso by regressing the number of nearby franchise bakery on the selected 15 variables. Next, we regress the residuals on 51 more variables with Double-Lasso selections and find very strong correlations. This additional result provide relevant empirical evidence how single Lasso selections can omit controls that can possibly have strong relationships with the key control variables in the main model.

Lasso and Post-Adaptive Lasso selections are similar, and the double-Lasso selection adds 51 additional variables.

[5]One referee kindly pointed out this issue, which can tighten the connection between our simulation and the empirical findings.

5.2. EFFECT ON THE CLOSURE OF SMALL AND MEDIUM-SIZED BAKERIES

The second set of results is focused to assess the effect of franchise bakeries on the failure only for nearby small- and medium-sized bakeries. Since we have estimated the closure rate of all the bakery shops including franchise bakeries in the Section 5.1, the competition among large franchise bakeries can cause overall higher closure rate in the previous section. Product differentiation or marketing differentiation of non-franchise bakeries can avoid the excessive competition among bakeries and cause different aspects of closure rates for the non-franchise bakeries. In this section, we conduct the same analysis as section 5.1. using a sample consisting solely of small- and medium-sized bakeries. There are 776 small and medium-sized bakeries in the data, and among them, 278 closed between 2015 and 2019.

One of interesting facts is that the number of selected control variables by Lasso and adaptive-Lasso decreased dramatically with the sample. For example, in the hinterland of SSTA, Lasso only selects the number of female residents as being in the 10s and 20s. This result implies that the closure of small and medium-sized bakeries might be much less dependent on market conditions than the closure of franchise bakeries. That is, small or mid-sized bakery shops are exposed to less competition compared to franchise bakeries possibly due to the product or marketing differentiation in the SSTA areas.

Table 5 lists the coefficient estimates for the number of franchise bakeries in each radius. Even with the post-double-Lasso selection, all estimated coefficients are insignificant, which means that the number of franchise bakeries does not significantly affect the closure rates of nearby small- and medium-sized bakeries.

On the other hand, when we focused on the effect of franchise bakeries on the failure only for nearby small- and medium-sized bakeries, one may want to consider the same set of explanatory variables used in the full sample in Table 4 instead of newly selected variables as in Table 5. In typical empirical analyses without variable selection, it is natural to use the same set of explanatory variables if they want to compare the effect of interest for the full sample vs. a subset.[6] Table 6 tabulates the estimation results with the same set of explanatory variables as in Table 4 without the different variable selection for the subset. With the same set of variables, we find very similar results compared to the Table 5. All of the estimated coefficients are insignificant as in Table 5, which implies

[6]We are grateful to referees for their valuable comments about this issue.

Variable Selection Procedure	Franchise bakeries in 200m	Franchise bakeries in 200m 500m	Franchise bakeries in 500m 800m
Post-Lasso-Selection	0.0386	-0.0298	0.0255
	(0.0734)	(0.0372)	(0.0262)
Post-Adaptive-Lasso-Selection	-0.0282	-0.0065	0.0243
	(0.1056)	(0.0386)	(0.0261)
Post- Double-Lasso-Selection	-0.0272	-0.0585	0.0254
	(0.1194)	(0.0586)	(0.0427)

Table 5: Estimated Effects of the Number of Franchise Bakeries on the Closure Rates of Nearby Small- and Medium-sized Bakeries

Note: Coefficient estimates for the number of franchise bakeries in radius with only small- and medium-sized bakeries. The values in parentheses are standard errors(Non-clustered). The number of observations is 8,095 with 776 bakeries.

that the selection procedure with only for nearby small- and medium-sized bakeries would not affect the result at all.

Variable Selection Procedure	Franchise bakeries in 200m	Franchise bakeries in 200m 500m	Franchise bakeries in 500m 800m
Post-Lasso-Selection	-0.0372	-0.0325	0.0272
	(0.111)	(0.0542)	(0.0282)
Post-Adaptive-Lasso-Selection	-0.0071	-0.0695	0.0190
	(0.111)	(0.0531)	(0.0267)
Post- Double-Lasso-Selection	-0.0143	-0.0590	0.0319
	(0.1196)	(0.0580)	(0.0433)

Table 6: Estimated Effects with the same set of explanatory variables as in Table 4

Note: Coefficient estimates for the number of franchise bakeries in radius with only small- and medium-sized bakeries. The values in parentheses are standard errors(Non-clustered). The number of observations is 8,095 with 776 bakeries.

Accordingly, we cautiously conclude that no statistical evidence for the number of nearby franchise bakeries affecting the closure rate of small- and medium-sized bakeries has been found at least in the hinterland of SSTA. Further, even if there is such an effect, the range of the influence area is not as wide as 500m. Our findings lead to somewhat different conclusions to some of previous literature, for example,Kang *et al.* (2018) found the exit rate of franchise bakeries in lower than other non-franchise bakeries with different sample periods and areas. In sum, we find little empirical evidence for the suitability of the designation of the

bakery industry for SMEs at least in terms of exit rates. One potential reason of no effect on small bakery may be due to the locational characteristics of Seoul, where enough consumers reside as a result of agglomeration economies. We may have different results in the cities outside of Seoul region, but the detailed data set outside of Seoul region is not currently available.

6. CONCLUSION

This paper discusses the validity of designating the bakery industry as a suitable business for SMEs and the suitability of the 500m store distance limits. In particular, the purpose of this paper is to determine whether the presence of large franchise bakeries significantly affects the closure of nearby bakeries while focusing on distance restriction guidelines. Our empirical results in this paper show that franchise bakeries have a significant impact on the closure of other bakeries within a 200m radius. However, when we focus on the closure rates of only small- and medium-sized bakery samples, there is no empirical evidence that franchise bakeries affect the closure of nearby small- and medium-sized non-franchised bakeries.

The main contribution of this paper to the literature can be summarized as follows. First, we show that the double-Lasso selection proposed by Belloni *et al.* (2014) can be applied to the Cox PH model, thus eliminating the possible bias caused by omitted variables. This paper also demonstrates the intuition behind why double-Lasso selection can resolve bias caused by omitted confounders following the approach in Lin *et al.* (1998) and show that the post-double-Lasso estimator outperforms post-single selection using Monte Carlo simulation. Second, we estimate the valid effect of the number of franchise bakeries on the closure rates of nearby bakeries, which cannot be found if we had used post-single selection. However, the application of post-double-Lasso makes it possible to validly estimate the effect of the number of franchise bakeries on the closure rates of nearby bakeries. This implies that post-Lasso and post-adaptive-Lasso procedures in the Cox PH model can generate biased estimates due to the omitted confounders.

REFERENCES

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in: B.N. Petrov and F. Csaki, eds., 2nd International Symposium on Information Theory (Akademia Kiado, Budapest).

Akaike, H. (1974). "A new look at the statistical model identification," IEEE Transactions on Automatic Control AC-19.

Alcácer, J. (2006). "Location choices across the value chain: How activity and capability influence collocation" *Management Science*, 52(10), 1457-1471.

Bates, T. (1995)."Analysis of Survival Rates Among Franchise and Independent Small Business Startups" *Journal of Small Business Management*,Vol. 33, No. 2, 26-36.

Belloni, A., and Chernozhukov, V. (2013). "Least squares after model selection in high-dimensional sparse models," in *Bernoulli*, 19(2), 521-547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81(2), 608-650.

Choi, Y.J., Kim,J.Y., and You, M.H. (2020). "Radius Restriction and Firms' Survival: Evidence from the Coffee Franchise Industry,"*Contemporary Economic Policy*, 38(3), 496-514.

Efroymson, M. (1966). "Stepwise regression - backward and forward look," Eastern Regional Meetings of the Institute of Mathematical Statistics. 1966.

Fan, J., and Li, R. (2010). "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, 96(456), 1348-1360.

Fan, J., and Peng, H. (2004). "Nonconcave penalized likelihood with a diverging number of parameters," *The annals of statistics*, 32(3), 928-961.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). "Assessing the sensitivity of regression results to unmeasured confounders in observational studies," *Biometrics*, 948-963.

Kang, J., Park, I., and Chun, H. (2018). "Survival Patterns between Franchise and Independent Stores: Evidence from Retail Bakeries," *Korean Journal of Industrial Organization* 26(4), 23-51.

Nam, Y. (2017). "Analysis on the Determinants of Exit of Self-Employed Businesses in Korea (in Korean)," Working Papers 2017-5, Economic Research Institute, Bank of Korea.

Schwarz, G. (1978). "Estimating the dimension of a model," *The annals of statistics*, 461-464.

Stanworth, J., D. Purdy, S. Price, and N. Zafiris (1998). "Franchise Versus Conventional Small Business Failure Rates in the US and UK: More Similarities than Differences," *International Small Business Journal*,16(3), 56-69.

Shaver, M. J., and Flyer, F. (2000). "Agglomeration economies, firm heterogeneity, and foreign direct investment in the United States," *Strategic management journals*, 21(12), 1175-1193.

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

Vittinghoff, E., and McCulloch, C. E. (2007). "Relaxing the rule of ten events per variable in logistic and Cox regression,"*American journal of epidemiology*, 165(6), 710-718.

Yang, J.S. (2007). "Study on the minimum distance restrictions: the effect of entering a large franchise on the closure of a nearby bakery,"*Korean Journal of Industrial Organization*,Proceeding, 148-184.

Zou, H. (2006). "The adaptive lasso and its oracle properties,"*Journal of the American statistical Vittinghoff association*,101(476), 1418-1429.

Korea Agro-Fisheries and Food Trade Corporation (2011). "A Survey on the Market Status of Processed Food Segmentation: Bakery Market (11-1541000-000745-01)."

Korea Agro-Fisheries and Food Trade Corporation (2019). "A Survey on the Market Status of Processed Food Segmentation: Bakery Market (11-1543000-002285-01)."