The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1117–1145
DOI: 10.1214/17-AOAS1039
© Institute of Mathematical Statistics, 2017

VARIABLE SELECTION FOR A CATEGORICAL VARYING-COEFFICIENT MODEL WITH IDENTIFICATIONS FOR DETERMINANTS OF BODY MASS INDEX[1]

BY JITI GAO*, BIN PENG† ZHAO REN‡ AND XIAOHUI ZHANG§

Monash University, University of Bath†, University of Pittsburgh‡ and University of Exeter§*

Obesity has become one of the major public health issues during the last three decades. A considerable number of determinants have been proposed for body mass index (BMI) by a large range of studies from multiple disciplines. In addition, it is well documented that impacts of these determinants are varying across demographic groups. However, little is known about the relative importance of these potential determinants and the varying impacts of all relatively important determinants. Using the shrinkage estimation technique, we propose a variable selection procedure for the categorical varying-coefficient model. We present a simulation study to exam performance of our method in different scenarios. We further apply the proposed method to examine the impacts of a large number of potential determinants on BMI using data from the 2013 National Health Interview Survey in the United States. By our method, the relevant determinants of BMI are identified through the variable selection procedure; and their varying impacts across demographic groups are quantified through the post-selection estimation.

1. Introduction. As a widely used measurement for body fat, body mass index (BMI) has been attracting significant attention from numerous researchers in multiple disciplines. The interest in measuring body fat came with increasing obesity in the last three decades, especially in developed countries. According to WHO estimates, the worldwide prevalence of obesity has more than doubled between 1980 and 2014. Obesity is a major risk factor for a large range of noncommunicable diseases [Fontaine et al. (2003), WHO (2015)]. It is thus crucial to identify and quantify the correlations between potential predictors and BMI. Empirical studies, which try to link particular lifestyle behaviors and other risk factors to BMI, may inform and guide policy makers to provide efficient incentives and interventions to reduce population BMI. Numerous studies have been seen in the last two decades and a large number of factors have been proposed as important drivers of increasing BMI [for references see Cawley (2011)]. Though there is an impressive amount of evidence on the individual importance of determinants, there is little

Received November 2016; revised February 2017.

[1]Supported in part by the Australian Research Council Discovery Grants Program under Grant numbers DP150101012 & DP170104421.

Key words and phrases. Body mass index, obesity, optimal variable selection, varying-coefficient regression.

guidance for policy makers about where cost-containment efforts [Stice, Shaw and Marti (2006)] should be focused. The inability of interventions to produce significant prevention effects may be due to incomplete understanding of the relative importance of predictors from various domains [Rehkopf et al. (2011)].

A lot of effort has been devoted to selecting the relatively important predictors for BMI in the last decade. Besides the conventional, but controversial, stepwise regression procedures [e.g., Von Kries et al. (2002)], some new statistical methods have been proposed or adopted recently to select determinants of BMI. For example, Huang et al. (2009) proposed a group bridge approach and applied it to determine risk factors on BMI of high school students. Rehkopf et al. (2011) adopted random forest, a tree-based analysis procedure, to rank the relative importance of risk factors for BMI among adolescent girls.

Despite the effort on selecting relatively important predictors for BMI, none of these studies simultaneously took into account the fact that impacts of determinants on BMI may vary across demographic groups. In fact, these varying impacts have been well documented in the literature. For example, Yu (2012) found that education attainment has different impacts on BMI in different gender, age and race groups. In particular, compared with college graduates, less educated whites and younger black women are more likely to be obese, and the differentials are larger for women than men, but weak or nonexistent among black men and older black women. Similar evidence has been found by a considerable number of studies, such as Colditz et al. (1991), Sobal, Rauschenbach and Frongillo (1992), Lipowicz, Gronkiewicz and Malina (2002), Zhang and Wang (2004) and so on. In order to capture such varying impacts, a common practice is to add interaction terms between selected BMI determinants and demographic variables into a regression model. The major shortcoming of this method is that it requires large degrees of freedom, which restrict the number of variables being allowed to have varying impacts on BMI. The choice of determinants having varying impacts, normally, serves to answer a specific research question, and therefore it is arbitrary and lacks statistical support. Furthermore, the method of adding interaction terms provides no statistical evidence to justify the importance of demographic variables, in terms of differencing the determinants' impacts on BMI.

In this paper, we provide a solution to the modeling issues existing in the literature of BMI studies using individual health survey data, that is, (1) how to allow for and quantify the varying impacts of determinants on BMI; (2) how to justify the relative importance of demographic variables in differencing potential determinants' impacts on BMI; and (3) how to identify the relatively important determinants of BMI. Data used in this study are from the 2013 National Health Interview Survey (NHIS) in the United States. There are 16,593 observations, 48 potential determinants and 32 demographic groups generated by 3 categorical variables (i.e., age group, gender and race).

To allow for and quantify the varying impacts of BMI determinants across demographic groups, we adopt the categorical varying-coefficient model proposed by

Li, Ouyang and Racine (2013), which specifies the impacts of BMI determinants as unknown functions of demographic variables. Different from the conventional practice of adding interaction terms to regression models, the categorical varying-coefficient model does not consume degrees of freedom that quickly when the number of demographic variables and/or BMI determinants increases.[2] Moreover, as documented in Li, Ouyang and Racine (2013), the selection of optimal bandwidths for categorical variables provides statistical justification on the relative importance of demographic variables in terms of differencing BMI determinants' impacts, and is able to serve as a filter to remove irrelevant demographic groups. For example, in our BMI study we are able to demonstrate that all demographic variables including age, gender and race are important in driving the BMI determinants' impacts to be different in different groups. We also find that gender and race are stronger in differencing the determinants' impacts on BMI than age. To identify the relatively important determinants of BMI, we adopt the group LASSO method proposed by Yuan and Lin (2006). In particular, we marry the categorical varying-coefficient model and the group LASSO method to simultaneously solve the aforementioned modeling issues in this BMI study.

The rest of the paper is organized as follows. We review the categorical varying-coefficient model of Li, Ouyang and Racine (2013), and introduce a variable selection procedure and its asymptotic results for the varying-coefficient model in Section 2. In Section 3, we conduct a Monte Carlo study to investigate the finite sample properties of the method. In Section 4, by using the 2013 NHIS data, we identify the important determinants of BMI and quantify their varying impacts on BMI across demographic groups. Section 5 concludes the paper with some discussions. The necessary assumptions required for the theoretical development are provided in the Appendix. Additional results and mathematical proofs are provided in the supplementary file of this paper [Gao et al. (2017)].

2. Methodology. In this study, a categorical varying-coefficient model is adopted to capture the varying impacts of a large range of factors on BMI across demographic groups. Varying-coefficient models have attracted considerable attention and gained popularity in the past two decades from both theoretical and practical aspects [e.g., Fan and Zhang (1999), Hastie and Tibshirani (1993), Li, Ouyang and Racine (2013), Li and Racine (2010), Wang and Xia (2009); and so forth]. As discussed in Wang and Xia (2009), including spurious regressors can degrade the estimation efficiency substantially. In order to address this issue, variable selection for varying-coefficient models has received increasing attention [Ma et al. (2015), Wang, Li and Huang (2008), Wang and Xia (2009)], but almost all of these existing variable selection methods for varying-coefficient models are specifically

[2]A detailed example is provided in Appendix S3 of the supplementary file [Gao et al. (2017)] to illustrate this difference.

for the setting that only continuous predictors or indexes enter the nonparametric specification of linear parameters. In fact, it is very common in empirical applications that categorical variables influence the regressors' impacts on the dependent variable, such as our BMI study in this paper.

To fill in the gap of literature and solve the modeling issues raised in BMI studies, we propose a variable selection procedure for the categorical varying-coefficient model below.

2.1. *Brief review*: *A categorical varying-coefficient model*. The model of Li, Ouyang and Racine (2013) is specified as follows:

$$Y_i = X_i' \beta_0(Z_i) + \varepsilon_i, \qquad i = 1, \ldots, N, \tag{2.1}$$

where $Z_i = (\bar{Z}_i', \tilde{Z}_i')'$ is an r-dimensional vector of discrete covariates with a support $\mathcal{D} = \bar{\mathcal{D}} \times \tilde{\mathcal{D}}$, $\bar{Z}_i = (Z_{i,1}, \ldots, Z_{i,\bar{r}})'$, $\tilde{Z}_i = (Z_{i,\bar{r}+1}, \ldots, Z_{i,r})'$ and $1 \le \bar{r} \le r$. Moreover, $\{\tilde{Z}_i, 1 \le i \le N\}$ is independent of all other variables and has no impact on $\beta_0(\cdot)$, which implies that $\tilde{Z}_i$ has no impact on Y_i at all. Therein, $\bar{Z}_i$ and $\tilde{Z}_i$ are referred to as relevant and irrelevant covariates, respectively. When $\bar{r} = r$, there is no irrelevant covariate existing in the system, that is, $\bar{Z}_i = Z_i$. To distinguish X_i from Z_i, they are referred to as regressors and covariates, respectively, hereafter. Based on the above description, the true model reduces to

$$Y_i = X_i' \beta_0(\bar{Z}_i) + \varepsilon_i, \qquad i = 1, \ldots, N, \tag{2.2}$$

where ε_i is a random error term; $X_i = (X_{i,1}, \ldots, X_{i,p})'$ is a p-dimensional vector of regressors; $\beta_0(z) = (\beta_{01}(z), \ldots, \beta_{0p}(z))'$ is a p-dimensional unknown coefficient function; and no information is known in advance to distinguish $\bar{Z}_i$ and $\tilde{Z}_i$. Moreover, both p and r are supposed to be fixed. This assumption is not that controversial. For example, in our BMI application, the sample size N is normally much larger than the number of potential predictors of X, that is, p, and the number of possible covariates Z is even smaller. In particular, N, p and r are 16,593, 48 and 3, respectively, in our BMI application. We refer to Section 4 for the details.

Applying model (2.2) to BMI data analysis allows us to capture the varying impacts of X, that is, potential predictors such as lifestyles and socio-economic factors, on BMI (indicated by Y) across demographic groups including gender, age group and race (denoted by Z). It is common practice to capture such kinds of varying impacts by adding interactions between the discrete Z variables and the X variables to a linear regression model, while it is straightforward to show that model (2.2) nests the latter model specification as a special case [cf. Appendix S3 of Gao et al. (2017)].

To carry on the regression, the kernel function of Aitchison and Aitken (1976) for an unordered covariate is adopted:

$$l(Z_{i,s}, z_s, \theta_s) = \begin{cases} 1, & \text{if } Z_{i,s} = z_s, \\ \theta_s, & \text{otherwise}, \end{cases} \tag{2.3}$$

where the range of θ_s is $[0, 1]$ for $s = 1, \ldots, r$. It can be seen that $\theta_s = 0$ leads to an indicator function and $\theta_s = 1$ gives a uniform weight function. Then (2.3) allows us to construct a product kernel function of the form

$$L(Z_i, z, \Theta) = \prod_{s=1}^{r} l(Z_{i,s}, z_s, \theta_s) = \prod_{s=1}^{r} \theta_s^{1(Z_{i,s} \neq z_s)}, \tag{2.4}$$

where $\Theta = (\theta_1, \ldots, \theta_r)'$. Therefore, for any $z \in \mathcal{D}$, the kernel-based OLS estimator is denoted as

$$\hat{\beta}(z) = \left[\sum_{j=1}^{N} X_j X_j' L(Z_j, z, \hat{\Theta}) \right]^{-1} \sum_{j=1}^{N} X_j Y_j L(Z_j, z, \hat{\Theta}),$$

where an optimal bandwidth $\hat{\Theta}$ is obtained by minimizing the following cross-validation criterion function:

$$CV(\Theta) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - X_i' \hat{\beta}_{-i})^2, \tag{2.5}$$

and the leave-one-out OLS estimator $\hat{\beta}_{-i}$ is defined as

$$\hat{\beta}_{-i} = \left[\sum_{j=1, j \neq i}^{N} X_j X_j' L(Z_j, Z_i, \Theta) \right]^{-1} \sum_{j=1, j \neq i}^{N} X_j Y_j L(Z_j, Z_i, \Theta).$$

It is convenient to introduce some notation here. For an r-dimensional vector $z = (z_1, \ldots, z_r)' \in \mathcal{D}$, we partition z as $z = (\bar{z}', \tilde{z}')'$ conformably with Z_i, where $\bar{z} = (z_1, \ldots, z_{\bar{r}})'$ and $\tilde{z} = (z_{\bar{r}+1}, \ldots, z_r)'$. Correspondingly, we partition Θ as $\Theta = (\bar{\Theta}', \tilde{\Theta}')'$, where $\bar{\Theta} = (\theta_1, \ldots, \theta_{\bar{r}})'$ and $\tilde{\Theta} = (\theta_{\bar{r}+1}, \ldots, \theta_r)'$. Due to space limitations, all assumptions needed for the lemmas and theorems in this paper are stated in the Appendix, and all mathematical proofs are provided in the supplementary file [Gao et al. (2017)]. Given that our study is based on Li, Ouyang and Racine (2013), we borrow two results from them and summarize them in the following lemma.

LEMMA 2.1. *Let* $\hat{\Theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)' = \operatorname{argmin}_{\Theta \in [0,1]^p} CV(\Theta)$.

1. *Under Assumptions 1 and 2.1,* $\hat{\theta}_s = O_P(\frac{1}{N})$ *for* $s = 1, \ldots, r$.
2. *Under Assumptions 1 and 2.2,* $\hat{\theta}_s = O_P(\frac{1}{\sqrt{N}})$ *for* $s = 1, \ldots, \bar{r}$, *and* $\lim_{N \to \infty} \Pr(\hat{\theta}_{\bar{r}+1} = 1, \ldots, \hat{\theta}_r = 1) \geq \alpha$ *for some* $\alpha \in (0, 1)$.

Lemma 2.1 summarizes Theorems 1 and 3 of Li, Ouyang and Racine (2013) and provides an asymptotic theory of smoothing parameters $\hat{\Theta}$. In particular, the rate of convergence of $\hat{\theta}_s$ depends on whether there is an irrelevant covariate or not, rather than the identification requirements stated in Assumptions 2.1 or 2.2. For

details, see Theorems 1 and 3 of Li, Ouyang and Racine (2013). It is worthwhile to mention that, for nonparametric/varying-coefficient models with at least one covariate as a continuous variable, the asymptotic theory of selected smoothing parameters through cross-validation has also been well developed [cf. Hall, Li and Racine (2007) and Li and Racine (2010)].

For a covariate z_s, if we obtain $\hat{\theta}_s = 1$, we can safely remove z_s from the model.[3] To some extent, this provides a variable selection procedure for the covariates. Hereafter, with a slight abuse of notation, we assume that we have removed all detected irrelevant covariates according to Lemma 2.1, that is, those z_s with $\hat{\theta}_s = 1$, and the remaining covariates of the ith observation are still represented by $Z_i = (\bar{Z}_i', \tilde{Z}_i')'$ as before. However, clearly there is a positive probability such that no $\tilde{Z}_i$ exists. The purpose of this variable selection on covariates is to reduce the total number of distinct realizations of z from our sample $\{Z_1, \ldots, Z_N\}$.

2.2. *Variable selection on X_i.* For model (2.2) with all detected irrelevant covariates removed, we propose a variable selection procedure to identify regressors of X_i with a nonzero coefficient when both p and r are fixed. Assume that there exists an unknown set $U^c \subseteq \{1, \ldots, p\}$ satisfying that $E|\beta_{0j}(\bar{Z}_i)|^2 = 0$ if and only if $j \in U^c$, where $\beta_{0j}(\bar{Z}_i)$ denotes the jth element of $\beta_0(\bar{Z}_i)$. To simplify notation, we assume that, in the true model, $U = \{1, \ldots, p^*\}$ and $U^c = \{p^* + 1, \ldots, p\}$, where the integer p^* satisfies $1 \le p^* \le p$. In other words, only the first p^* variables in X_i have nonzero coefficients and our goal is to identify U and U^c.

Let m denote the number of realizations of z by observing $\{Z_1, \ldots, Z_N\}$. Obviously m converges to the cardinality of $\mathcal{D}$ in probability with nondegenerate probability imposed on i.i.d. Z_i as N diverges to ∞. Since m is finite and observable, our parameters of interest can be characterized by the following $m \times p$ matrix B with the underlying true coefficient function B_0. For the sake of presentation, denote

$$
\begin{aligned}
\underset{m\times p}{B} &= (\beta_1, \ldots, \beta_m)' = (b_1, \ldots, b_p), \\
\underset{p\times 1}{\beta_j} &= (\beta_{j,1}, \ldots, \beta_{j,p})' \qquad \text{for } j = 1, \ldots, m, \\
\underset{m\times 1}{b_s} &= (\beta_{1,s}, \ldots, \beta_{m,s})' \qquad \text{for } s = 1, \ldots, p, \\
\underset{m\times p}{B_0} &= \big(\beta_0(z^1), \ldots, \beta_0(z^m)\big)' = (b_{01}, \ldots, b_{0p^*}, 0, \ldots, 0), \\
\underset{m\times 1}{b_{0s}} &= \big(\beta_{0s}(z^j), \ldots, \beta_{0s}(z^j)\big)' \qquad \text{for } s = 1, \ldots, p^*,
\end{aligned} \tag{2.6}
$$

where z^j, $j = 1, \ldots, m$, denotes the jth realization of $z \in \mathcal{D}$.

[3] Although one cannot always achieve $\hat{\theta}_s = 1$ for all irrelevant covariates simultaneously, as stated in Lemma 2.1, there is always a certain positive probability that we can recognize a covariate as irrelevant; that is, the probability of $\hat{\theta}_s = 1$ for the corresponding covariate is positive.

Notice that the last $p - p^*$ columns of B_0 are zero columns. By treating entries in each column of B_0 as a group, the selection on the regressor of X_i is, essentially, to identify those groups (i.e., columns) of the matrix B_0 with all entries as zero. Following the spirit of Yuan and Lin (2006), we consider the following regularized least squares estimator:

$$\hat{B} = (\hat{\beta}_{\gamma,1}, \ldots, \hat{\beta}_{\gamma,m})' = (\hat{b}_{\gamma,1}, \ldots, \hat{b}_{\gamma,p}) = \operatorname*{argmin}_{B \in \mathbb{R}^{m \times p}} Q_\gamma(B), \tag{2.7}$$

and

$$Q_\gamma(B) = \sum_{j=1}^{m} \sum_{i=1}^{N} (Y_i - X_i' \beta_j)^2 L(Z_i, z^j, \hat{\Theta}) + \sum_{s=1}^{p} \gamma_s \|b_s\|, \tag{2.8}$$

where $\hat{\Theta}$ is the smoothing parameter vector obtained from Lemma 2.1; b_s ($s = 1, \ldots, p$) is the sth column of B as denoted in (2.6); $\sum_{s=1}^{p} \gamma_s \|b_s\|$ is the group-wise regularizer and defined as the weighted sum of the ℓ_2 norms of all the column vectors in B; and $\gamma = (\gamma_1, \ldots, \gamma_p)'$ represents the weight that controls the group-wise regularizer.

REMARK 2.1. If we ignore the optimal bandwidth selection and use an indicator function to replace all kernel functions, we essentially have an adaptive version of a group LASSO model [cf. Yuan and Lin (2006)]. On the other hand, if we set all γ_s's to 0, we end up with the model proposed in Li, Ouyang and Racine (2013). Due to the features of BMI data, we combine both methods together and try to filter out any redundant information as much as possible.

Our first theorem is stated below.

THEOREM 2.1. *Suppose Assumptions 1–3 hold.*

1. *Let* $\gamma^* = (\gamma_1, \ldots, \gamma_{p^*})'$ *and* $\frac{\|\gamma *\|}{\sqrt{N}} \to \omega_1$, *where* ω_1 *is a constant satisfying* $0 \le \omega_1 < \infty$. *Then* $\|\hat{\beta}_{\gamma,j} - \beta_0(\bar{z}^j)\| = O_P(N^{-1/2})$ *for* $j = 1, \ldots, m$, *where* $\bar{z}^j = (z_1^j, \ldots, z_{\bar{r}}^j)'$.
2. *Let* $\frac{1}{\sqrt{N}} \min_{s \in \{p^*+1, \ldots, p\}} \gamma_s \ge \omega_2$, *where* ω_2 *is a sufficiently large constant. Then* $\Pr(\|\hat{b}_{\gamma,j}\| = 0) \to 1$ *for* $j = p^* + 1, \ldots, p$.

The first result of Theorem 2.1 states that if the regularizer weight is not too large, then the estimator (2.7) always has optimal $\sqrt{N}$ consistency. The second result implies that when the regularizer weight is at level $\sqrt{N}$, estimator (2.7) can successfully identify those regressors with a zero coefficient. To satisfy the assumptions in Theorem 2.1, all elements of γ can be simply set at level $\sqrt{N}$. However, with such γ, Theorem 2.1 does not imply any asymptotic normality property

of the estimator (2.7), while in Li, Ouyang and Racine (2013) the asymptotic normality property has been achieved for the oracle estimator.[4] Specifically, the oracle estimator is defined as

$$\hat{\beta}_{\mathrm{ora}}(\bar{z}^j) = \left(\sum_{i=1}^{N} X_{iU} X_{iU}' L(Z_i, z^j, \hat{\Theta})\right)^{-1} \sum_{i=1}^{N} X_{iU} Y_i L(Z_i, z^j, \hat{\Theta}), \tag{2.9}$$

where $j = 1, \ldots, m$ and $X_{iU} = (X_{i,1}, \ldots, X_{i,p^*})'$.

In fact, with a more careful data-driven choice of γ, we can further achieve the asymptotic normality whenever there is no irrelevant covariate with the help of following the oracle property for our estimator (2.7).

THEOREM 2.2. *Under conditions of Theorem 2.1,* $\|\hat{\beta}_{\gamma,jU} - \hat{\beta}_{\mathrm{ora}}(\bar{z}^j)\| = O_P(\frac{\|\gamma^*\|}{N})$ *for* $j = 1, \ldots, m$, *where* $\hat{\beta}_{\gamma,jU} = (\hat{\beta}_{\gamma,j1}, \ldots, \hat{\beta}_{\gamma,jp^*})'$; $\hat{\beta}_{\gamma,js}$ *denotes the* s*th element of* $\hat{\beta}_{\gamma,j}$ *for* $j = 1, \ldots, m$ *and* $s = 1, \ldots, p^*$; *and* γ^* *is denoted in Theorem 2.1.*

To achieve an asymptotic normality for the estimator (2.7), the convergence rate of $\hat{\beta}_{\gamma,jU}$ to $\hat{\beta}_{\mathrm{ora}}(\bar{z}^j)$ has to be much faster than $\frac{1}{\sqrt{N}}$. The oracle property in Theorem 2.2 implies such a result as long as $\|\gamma^*\|$ is much smaller than $\sqrt{N}$. Therefore, the simple choice of $\sqrt{N}$ level for γ is not sufficient.

To achieve a desired asymptotic normality property for the estimator (2.7), we propose a data-driven choice of γ, which can yield an even faster rate of convergence of an order of $o_P(\frac{1}{\sqrt{N}})$ to the oracle estimator. From now on, we assume that whenever the true coefficient is nonzero, that is, $b_{0s} \neq 0$ for $s = 1, \ldots, p^*$, its ℓ_2 norm is much larger than root N level, that is, $\|b_{0s}\| \gg \frac{1}{\sqrt{N}}$ for $s = 1, \ldots, p^*$. This assumption is not controversial in the current fixed dimension setting in which $\|b_{0s}\|$ is some positive constant as N increases.

Similarly to Wang and Leng (2007) and Wang and Xia (2009), our data-driven regularizer weight is as follows:

$$\gamma = \tilde{\gamma}\big(\|\tilde{b}_1\|^{-1}, \ldots, \|\tilde{b}_p\|^{-1}\big)', \tag{2.10}$$

where $\tilde{\gamma}$ is a scalar, $\tilde{b}_s$ is the sth column of the unregularized estimator $\tilde{B}$, and $\tilde{B}$ is obtained from (2.8) by simply choosing $\gamma_1 = \cdots = \gamma_p = 0$ as follows:

$$\tilde{B} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_m)' = (\tilde{b}_1, \ldots, \tilde{b}_p) = \operatorname*{argmin}_{B \in \mathbb{R}^{m \times p}} Q(B) \tag{2.11}$$

[4]Notice that the word "oracle" refers to those estimators provided in Li, Ouyang and Racine (2013) by assuming we know the true set U. Here we completely ignore the inefficiency brought in the model by the irrelevant covariates $\tilde{Z}_i$. The asymptotically efficient estimator is obtained when we know both the set U and the irrelevant covariates. However, this can only be done at a certain probability based on Lemma 2.1.

and

$$Q(B) = \sum_{j=1}^{m}\sum_{i=1}^{N}(Y_i - X_i'\beta_j)^2 L(Z_i, z^j, \hat{\Theta}). \tag{2.12}$$

Under Assumption 3.1, the first result of Theorem 2.1 and the assumption of $\|b_{0s}\| \gg \frac{1}{\sqrt{N}}$ for $s = 1, \ldots, p^*$, it is easy to verify that $\|\tilde{b}_s\|^{-1} = o_P(\sqrt{N})$ for $s = 1, \ldots, p^*$ and $\|\tilde{b}_s\| = O_P(1/\sqrt{N})$ for $s = p^* + 1, \ldots, p$. Then the intuition of choosing γ as (2.10) is straightforward. The unregularized estimator $\tilde{B}$ is an $\sqrt{N}$ consistent estimator. It provides information on how likely each column of B_0 is a zero column. In other words, smaller $\|\tilde{b}_j\|$ implies that the jth column is more likely to be zero, and hence suggests a larger regularizer on $\|b_j\|$. In particular, given that $\|\tilde{b}_s\|^{-1} = o_P(\sqrt{N})$ for $s = 1, \ldots, p^*$, Theorem 2.2 implies the desired rate of $o_P(\frac{1}{\sqrt{N}})$ for $\hat{\beta}_{\gamma, jU}$ to be the oracle estimator $\hat{\beta}_{\text{ora}}(\bar{z}^j)$. Given the form of γ in (2.10), the selection on the vector γ reduces to the selection on the scalar $\tilde{\gamma}$. Note that the properties of $\|\tilde{b}_j\|^{-1}$ for $j = 1, \ldots, p$ imply that a large enough constant $\tilde{\gamma}$ would satisfy all the conditions on γ. More specifically, we select the constant $\tilde{\gamma}$ by the following modified BIC-type (MBIC) criterion:

$$MBIC_{\tilde{\gamma}} = \ln RSS_{\tilde{\gamma}} + df_{\tilde{\gamma}} \cdot \frac{\ln N}{N},$$

where $df_{\tilde{\gamma}}$ is the number of nonzero coefficients identified by $\hat{B}_{\tilde{\gamma}}$, and $RSS_{\tilde{\gamma}}$ is defined as $RSS_{\tilde{\gamma}} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}(Y_i - X_i'\hat{\beta}_{\tilde{\gamma}, j})^2 L(Z_i, z^j, \hat{\Theta})$. The weight parameter is obtained by

$$\hat{\tilde{\gamma}} = \operatorname*{argmin}_{\tilde{\gamma}} MBIC_{\tilde{\gamma}}. \tag{2.13}$$

Recall the true set of nonzero coefficients is denoted by $U = \{1, \ldots, p^*\}$. Let $S_{\hat{\tilde{\gamma}}} = \{j : \|\hat{b}_{\hat{\tilde{\gamma}}, j}\| > 0, 1 \le j \le p\}$ indicate the set of relevant variables identified by the regularized estimator $\hat{B}_{\hat{\tilde{\gamma}}}$ with the weight parameter $\hat{\tilde{\gamma}}$ chosen by (2.13). Then we have the following theorem.

THEOREM 2.3. *Suppose that* $\|b_{0s}\| \gg \frac{1}{\sqrt{N}}$ *for* $s = 1, \ldots, p^*$. *Under conditions of Theorem* 2.1, *the weight parameter selected by the modified BIC-type criterion* (2.13) *can do the following*:

1. *Identify the true model consistently, that is*, $\Pr(S_{\hat{\tilde{\gamma}}} = U) \to 1$ *as* $N \to \infty$.
2. *Achieve asymptotic normality, that is*,

$$\sqrt{N}(\hat{\beta}_{\hat{\tilde{\gamma}}, jU} - \beta_{0U}(z^j)) \to_D N(0, \Sigma(z^j)) \tag{2.14}$$

for the relevant covariate case defined in Assumption 2, and for $j = 1, \ldots, m$, *where*

$$\begin{aligned}
\Sigma(z^j) &= A^{-1}(z^j)\Omega(z^j)A^{-1}(z^j),\\
A(z^j) &= E\big[X_{iU}X_{iU}'|z^j\big]\Pr(z^j),\\
\Omega(z^j) &= E\big[\varepsilon_i^2 X_{iU}X_{iU}'|z^j\big]\Pr(z^j),\\
\beta_{0U}(z^j) &= \big(\beta_{01}(z^j), \ldots, \beta_{0p^*}(z^j)\big)',
\end{aligned}$$

and X_{iU} *has been defined in* (2.9).

3. *For the irrelevant covariate case defined in Assumption 2,*

$$\hat{\beta}_{\hat{\gamma},jU} - \beta_{0U}(\bar{z}^j) = O_P\left(\frac{1}{\sqrt{N}}\right) \tag{2.15}$$

for $j = 1, \ldots, m$, *where* $\beta_{0U}(\bar{z}^j) = (\beta_{01}(\bar{z}^j), \ldots, \beta_{0p^*}(\bar{z}^j))'$.

When there is no irrelevant covariate (i.e., $r = \bar{r}$ and $Z_i = \bar{Z}_i$), the asymptotic normality result of (2.14) is based on the limiting distribution of $\sqrt{N}(\hat{\beta}_{\mathrm{ora}}(z^j) - \beta_{0U}(z^j))$, which is established by applying Theorem 2 of Li, Ouyang and Racine (2013) on the oracle model. In practice, one may want to establish a consistent estimate for $\Sigma(z^j)$ for $j = 1, \ldots, m$, which can be immediately obtained following the procedure provided in Theorem 2 of Li, Ouyang and Racine (2013), assuming $S_{\hat{\gamma}} = U$:

$$\hat{\Sigma}(z^j) = \hat{A}^{-1}(z^j)\hat{\Omega}(z^j)\hat{A}^{-1}(z^j),$$

where $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}_{\hat{\gamma},jU}$, $\hat{\Omega}^{-1}(z^j) = \frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_i^2 X_{iU}X_{iU}'L(Z_i, z^j, \hat{\Theta})$, and $\hat{A}^{-1}(z^j) = \frac{1}{N}\sum_{i=1}^N X_{iU}X_{iU}'L(Z_i, z^j, \hat{\Theta})$.

However, when there are irrelevant covariates (i.e., $r > \bar{r}$), the asymptotic distribution of $\sqrt{N}(\hat{\beta}_{\mathrm{ora}}(\bar{z}^j) - \beta_{0U}(\bar{z}^j))$ remains unknown even for the oracle estimator, and hence we only obtain $\sqrt{N}$ consistency in (2.15). In this case, the asymptotic distribution of $\sqrt{N}(\hat{\beta}_{\mathrm{ora}}(\bar{z}^j) - \beta_{0U}(\bar{z}^j))$ can be established by using a bootstrap method as documented in Li, Ouyang and Racine (2013).

In this section, we propose a regularized estimator for the categorical varying-coefficient model and obtain its superior statistical properties. In particular, the coefficients of the proposed categorical varying-coefficient model possess a natural group structure. To take advantage of the structure, we apply a group-wise regularizer to improve accuracy of variable selection and parameter estimation. Moreover, we apply a data-driven method, that is, a modified BIC-type criterion, to select the weight parameter, which further boosts the performance and helps to achieve an asymptotic normality property for the estimator, especially when no irrelevant covariate presents.

3. Monte Carlo evidence. In this section, we conduct a comprehensive Monte Carlo (MC) study to show the finite-sample performance of our method and a range of competing methods. To each generated data set $\{Y_i, X_i, Z_i\}$, first, we apply model (2.2) and estimate the optimal bandwidths. Following Lemma 2.1 and its discussion in Section 2.1, we remove irrelevant covariates to reduce the number of groups based on the realizations of Z_i.[5] Second, we identify the irrelevant regressors by estimating $\hat{B}$ through (2.7). Last, we estimate the model excluding irrelevant covariates and regressors by the unregularized estimator proposed in Li, Ouyang and Racine (2013). The purpose of the last step is to further reduce the possible bias.

To compare the finite-sample performance of our method with some competing ones and put all the methods on equal footing, we use their adaptive versions for all LASSO related methods. More specifically, for each data set, we conduct (a) an adaptive version group LASSO estimation method, (b) an adaptive version of the LASSO estimation method, and (c) a stepwise estimation method. In particular, the group LASSO method (denoted by GroupL) is essentially a special case of (2.2), that is, with all bandwidths equal to 0. Alternatively, without taking into account the varying impacts of X on Y according to Z, we apply methods (b) and (c) to the linear regression model (3.1) below (denoted by LASSO1 and SW1, respectively). Moreover, we apply methods (b) and (c) to the linear regression model (3.2) below (denoted by LASSO2 and SW2, respectively), where the varying impacts of X on Y are (particularly) captured by the interaction terms between X on Z. It is a very common practice in empirical studies [e.g., Yu (2012)]:

$$Y_i = \bigl(X_i', Z_{i,11}, \ldots, Z_{i,1c_1-1}, \ldots, Z_{i,r1}, \ldots, Z_{i,rc_r-1}\bigr)'\beta_0^* + \varepsilon_i, \tag{3.1}$$

$$\begin{aligned} Y_i = \bigl(&X_i', (Z_{i,11}X_i)', \ldots, (Z_{i,1c_1-1}X_i)', \\ &(Z_{i,r1}X_i)', \ldots, (Z_{i,rc_r-1}X_i)'\bigr)'\beta_0^* + \varepsilon_i, \end{aligned} \tag{3.2}$$

where $Z_{i,jk} = 1$ if the jth element of Z_i is k with $k = 1, \ldots, c_j - 1$; $Z_{i,jk} = 0$, otherwise.

Notice that when X_i does not exist in a model (3.1), that is, only categorical variables are included, special treatment [Gertheiss and Tutz (2010)] can be considered. We avoid using more complicated ways to introduce interactions in model (3.2) since it is almost impossible to exhaust all possibilities.

We consider three scenarios in terms of the data-generating process (DGP). In the first two scenarios, the DGPs are based on two categorical varying-coefficient models, that is, without and with the irrelevant covariate included in Z_i, respectively. And the DGP of the third scenario is a conventional linear regression model. Details of the DGPs are as follows:

[5]Refer to Li, Ouyang and Racine (2013) for extensive evidence on the performance of bandwidth selection in a finite sample.

Scenario 1: Let $p = 10$, $p^* = 5$, and $Y_i = (1, X_i')'\beta_0(Z_i) + \varepsilon_i$, where $X_i = H_i + V_i$ and $Z_i = (Z_{i,1}, \ldots, Z_{i,r})'$. For $\forall j = 1, \ldots, r$, $Z_{i,j}$ is i.i.d. over i and takes a value from $\{0, 1, 2\}$ with probability $\{0.25, 0.25, 0.5\}$, respectively. V_i is i.i.d. over i and follows $N(Z_{i,1}/2 \cdot i_{p-1}, \sqrt{Z_{i,1}+1} \cdot I_{p-1})$, in which I_{p-1} denotes the $(p-1)$-dimensional identity matrix and i_{p-1} represents the $(p-1)$-dimensional vector with all entries being one; H_i is i.i.d. over i and follows $N(i_{p-1}, I_{p-1})$; and ε_i is i.i.d. over i and follows $N(0, 1)$. Let $\beta_{0j}(Z_i)$ denote the jth element of the coefficient function $\beta_0(Z_i)$ for $j = 1, \ldots, p$.

Two sub-scenarios are designed as without and with the irrelevant covariate included in Z_i, respectively:

- *Scenario* 1.1: Relevant Covariate Case (i.e., $\bar{r} = r$). For $\forall j \le 5$,

$$\beta_{0j}(Z_i) = \begin{cases} 2 + 2j, & \text{if the remainder of } \sum_{k=1}^{r} Z_{i,k}/2 \text{ is 0,} \\ 1 + 2j, & \text{otherwise;} \end{cases}$$

  for $\forall j > 5$, $\beta_{0j} = 0$.
- *Scenario* 1.2: Irrelevant Covariate Case (i.e., $\bar{r} = 1$). For $\forall j \le 5$,

$$\beta_{0j}(Z_i) = \begin{cases} 2 + 2j, & \text{if the remainder of } Z_{i,1}/2 \text{ is 0,} \\ 1 + 2j, & \text{otherwise;} \end{cases}$$

  for $j > 5$, $\beta_{0j} = 0$.

Scenario 2: Let $Y_i = (1, X_i')'\beta_0 + \varepsilon_i$, where $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})'$, and $\beta_{0j} = 5$ with $j \le 5$ and $\beta_{0j} = 0$ with $j > 5$. All the other variables are generated in exactly the same way as for Scenario 1.

Under Scenario 1, model (2.2) is correctly specified, while models (3.1) and (3.2) are misspecified. Therefore, we expect our estimator performs better than the other methods. Under Scenario 2, all models [i.e., (2.2), (3.1) and (3.2)] are correctly specified, and so we expect reasonable performance from all the estimators.

To evaluate model performance, we examine three measures. They are (1) the percentage of missed true regressors (FNR); (2) the percentage of falsely selected noise regressors (FPR);[6] and (3) the mean squared prediction error (MSPE). We calculate MSPE, in the spirit of Chu, Li and Reimherr (2016), as follows:

$$MSPE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{-i} - y_i)^2, \tag{3.3}$$

[6]To be clear, all binary variables and interaction terms in (3.1) and (3.2) are considered as redundant information. For example, if we identify some interaction terms as relevant regressors by the LASSO method for model (3.2), these variables are counted as falsely selected.

where $\hat{y}_{-i}$ denotes the leave-one-out prediction for the ith individual (i.e., we implement estimation without the observation of the ith individual, and then use the estimated parameters to predict y_i for the ith individual). For each method under each scenario, we report averaged, over 1000 replications, FNR and FPR, and the root of averaged MSPE, denoted as RME. Note that the estimated RME should ideally converge to the standard deviation of ε_i (i.e., 1 in our MC design). Therefore, an estimated RME closing to 1 is an indicator for good model performance of the corresponding method.

In this MC study, we also consider a range of different settings for (N, r). In particular, we consider N of 2000, 4000 and 8000, which are reasonable, if not much smaller, sample sizes in empirical applications. With regard to the size of r, we set it as 2, 3 and 4. It is noteworthy that as $r = 4$, we already have 81 demographic groups based on our DGP, and so it is more than enough to demonstrate that the current setting covers our case study perfectly. For example, in our BMI study, 3 covariates (and 32 groups) are reasonably considered, which is supported by the BMI literature [cf. Yu (2012)].

We summarize the simulation results in Table 1. As expected, under Scenarios 1.1 and 1.2, our estimator (denoted as Varying-Coef) and group LASSO estimator (denoted as GroupL) outperform all other methods in general. As models (3.1) and (3.2) are misspecified, it is not surprising that LASSO1, LASSO2, SW1 and SW2 do not perform well. The RME's estimated by our estimator and group LASSO method, under different settings, are all close to 1, that is, the true standard deviation of ε_i. However, those estimated by LASSO and stepwise methods are far away from 1, which is an indication for less accurate estimates. Note that the true regressor can almost be identified by our estimator and group LASSO method, that is, FNR's are zero; in contrast, FNR's from SW1, SW2 and LASSO2 are considerably large. FPR's from Varying-Coef and GroupL are very low compared to those from all other methods. Not surprisingly, under Scenario 2, all methods perform relatively well except SW1 and SW2.

We now take a close look at these results from Varying-Coef and GroupL, as both of them can address two questions raised in the Introduction, that is, (1) allowing for and quantifying the varying impacts, and (2) identifying the relatively important determinants. However, only our method is able to address the question of "how to justify the relative importance of demographic variables" by looking at the estimates of the optimal bandwidths based on Lemma 2.1. Compared to the group LASSO method, the better performance of the varying-coefficient setting is due to the following two reasons: (1) The varying-coefficient setting uses optimal bandwidths throughout Scenarios 1.1, 1.2 and 2, and so the RMEs of Varying-Coef are closer to 1 as expected; and (2) For Scenario 1.2, the varying-coefficient setting can potentially throw away more possible irrelevant variables, and so that reduces the number of groups based on the realizations of Z_i. In other words, each group can potentially include more samples after we remove extra covariates from the system. For the sake of space, we report the histograms of the estimates on the

TABLE 1
Monte Carlo Simulation Results

			Varying-Coef			GroupL			LASSO1			SW1			LASSO2			SW2		
	r	***N***	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**
Scenario 1.1	2	2000	0.9871	0.0000	0.0357	0.9869	0.0000	0.0381	4.1390	0.0000	0.2497	4.2554	0.0833	0.1843	3.4135	0.0158	0.6561	3.4344	0.0143	0.1852
		4000	0.9942	0.0000	0.0076	0.9941	0.0000	0.0078	4.1424	0.0000	0.2168	4.2504	0.0833	0.1888	3.3459	0.0166	0.6567	3.4270	0.0143	0.2081
		8000	0.9970	0.0000	0.0031	0.9966	0.0000	0.0036	4.1458	0.0000	0.1878	4.2528	0.0833	0.1889	3.2587	0.0170	0.6567	3.3922	0.0143	0.2406
	3	2000	0.9354	0.0000	0.0404	0.9321	0.0000	0.0445	4.2912	0.0000	0.2682	4.4940	0.0769	0.1192	4.1787	0.0160	0.6589	4.4234	0.0121	0.1145
		4000	0.9801	0.0000	0.0118	0.9794	0.0000	0.0149	4.2993	0.0000	0.2160	4.4440	0.0769	0.1827	4.1822	0.0164	0.6574	4.3875	0.0121	0.1459
		8000	0.9909	0.0000	0.0068	0.9907	0.0000	0.0075	4.3031	0.0000	0.2401	4.4092	0.0769	0.2376	5.9979	0.0165	0.6577	4.3494	0.0121	0.1729
	4	2000	0.8038	0.0000	0.0921	0.7565	0.0000	0.0934	4.3319	0.0000	0.2264	4.6127	0.0714	0.0583	4.5710	0.0163	0.6583	4.6233	0.0105	0.0684
		4000	0.9585	0.0000	0.0758	0.8932	0.0000	0.0802	4.3393	0.0000	0.1693	4.5909	0.0714	0.0812	4.5658	0.0158	0.6565	4.6252	0.0105	0.0854
		8000	0.9986	0.0000	0.0660	0.9477	0.0000	0.0690	4.3433	0.0000	0.1602	4.5340	0.0714	0.1375	4.2906	0.0162	0.6561	4.6150	0.0105	0.1123
Scenario 1.2	2	2000	0.9929	0.0000	0.0379	0.9868	0.0000	0.1639	3.4909	0.0000	0.1383	3.6608	0.0833	0.1074	1.2706	0.0160	0.6348	2.0078	0.0143	0.0984
		4000	0.9970	0.0000	0.0130	0.9942	0.0000	0.1161	3.4944	0.0000	0.1068	3.6639	0.0833	0.1055	1.2326	0.0170	0.6298	2.0124	0.0143	0.0980
		8000	0.9985	0.0000	0.0043	0.9972	0.0000	0.0748	3.4940	0.0000	0.0918	3.6637	0.0833	0.1059	1.0383	0.0159	0.6176	2.0115	0.0143	0.0969
	3	2000	0.9898	0.0000	0.1423	0.9323	0.0000	0.2575	3.4904	0.0000	0.1321	3.6599	0.0769	0.1008	1.0752	0.0150	0.6311	2.0089	0.0121	0.0979
		4000	0.9954	0.0000	0.0759	0.9797	0.0000	0.2367	3.4897	0.0000	0.0998	3.6568	0.0769	0.1025	1.2851	0.0158	0.6266	2.0082	0.0121	0.0979
		8000	0.9977	0.0000	0.0196	0.9909	0.0000	0.0912	3.4932	0.0000	0.0828	3.6611	0.0769	0.0998	1.4331	0.0171	0.6202	2.0120	0.0121	0.0977
	4	2000	0.9881	0.0000	0.3168	0.7860	0.0000	0.3586	3.4892	0.0000	0.1111	3.6572	0.0714	0.0935	1.2057	0.0162	0.6304	2.0064	0.0105	0.0972
		4000	0.9942	0.0000	0.2656	0.8854	0.0000	0.3034	3.4904	0.0000	0.0884	3.6560	0.0714	0.0965	1.3948	0.0163	0.6264	2.0134	0.0105	0.0971
		8000	0.9972	0.0000	0.1584	0.9356	0.0000	0.2104	3.4941	0.0000	0.0784	3.6585	0.0714	0.0966	1.1489	0.0157	0.6171	2.0088	0.0105	0.0977

TABLE 1
(*Continued*)

			Varying-Coef			GroupL			LASSO1			SW1			LASSO2			SW2		
	r	***N***	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**	**RME**	**FNR**	**FPR**
Scenario 2	2	2000	0.9972	0.0000	0.0002	0.9883	0.0000	0.0000	0.9985	0.0000	0.0000	2.7182	0.0833	0.0291	1.0647	0.0156	0.0820	2.6802	0.0143	0.0290
		4000	0.9988	0.0000	0.0000	0.9945	0.0000	0.0000	0.9994	0.0000	0.0000	2.7172	0.0833	0.0286	0.9938	0.0156	0.0813	2.6815	0.0143	0.0289
		8000	0.9992	0.0000	0.0000	0.9971	0.0000	0.0000	0.9995	0.0000	0.0000	2.7210	0.0833	0.0277	0.9967	0.0165	0.0802	2.6863	0.0143	0.0285
	3	2000	0.9953	0.0000	0.0008	0.9320	0.0000	0.0000	0.9980	0.0000	0.0000	2.6872	0.0769	0.0303	1.0855	0.0154	0.0814	2.6761	0.0121	0.0297
		4000	0.9980	0.0000	0.0000	0.9811	0.0000	0.0000	0.9992	0.0000	0.0000	2.6869	0.0769	0.0308	0.9935	0.0156	0.0811	2.6818	0.0121	0.0285
		8000	0.9989	0.0000	0.0000	0.9913	0.0000	0.0000	0.9995	0.0000	0.0000	2.7014	0.0769	0.0277	0.9967	0.0167	0.0803	2.6826	0.0121	0.0291
	4	2000	0.9939	0.0000	0.0012	0.7858	0.0000	0.0000	0.9990	0.0000	0.0000	2.6670	0.0714	0.0325	0.9879	0.0156	0.0810	2.6789	0.0105	0.0289
		4000	0.9971	0.0000	0.0000	0.8932	0.0000	0.0000	0.9995	0.0000	0.0000	2.6764	0.0714	0.0301	0.9939	0.0158	0.0811	2.6823	0.0105	0.0286
		8000	0.9986	0.0000	0.0000	0.9478	0.0000	0.0000	0.9996	0.0000	0.0000	2.6661	0.0714	0.0327	0.9968	0.0167	0.0803	2.6777	0.0105	0.0295

1. Varying-Coef represents our variable selection method; GroupL represents the group LASSO method; LASSO1 represents applying the LASSO method to model (3.1); LASSO2 represents applying the LASSO method to model (3.2); SW1 represents applying the stepwise method to model (3.1); SW2 represents applying the stepwise method to model (3.2).

2. Note that the estimated RME should converge to the standard deviation of ε_i (i.e., 1 in our MC design). Therefore, an estimated RME closing to 1 is an indicator for good model performance of the corresponding method.

bandwidth of irrelevant covariates with corresponding discussions in the supplementary file of this paper [Gao et al. (2017)].

4. An application to BMI.

4.1. *Data*. Data used in this empirical study are from the 2013 National Health Interview Survey (NHIS) in the United States. The NHIS is conducted annually through face-to-face interviews. Our analysis focuses on adults aged 18 and over. BMI is calculated based on self-reported height and weight. We exclude underweight individuals (BMI less than 18.5) from our analysis, and focus on such individuals with normal weight and overweight. There are three reasons for us to do so. First, underweight is a much less prevalent health problem in developed countries like the U.S. In particular, in the NHIS data underweight accounts for a very small proportion, that is, 1.8 percent of the whole sample. Second, factors causing (or relating to) underweight are very much different from those for overweight or obesity. For example, eating disorders, such as anorexia nervosa and bulimia, lack of nutrition, and a hypermetabolism state, are considered as causes of underweight [Ali and Lindström (2006)], while unhealthy lifestyles and poor socio-economic factors are the major determinants of overweight and obesity (as discussed below in detail). However, information on these potential determinants of underweight is not available in the NHIS. Last but not least, for common factors causing both underweight and overweight, their impacts on BMI might have different signs. For example, mental health problems, such as depression, can cause both BMI increase from normal weight to overweight level (positive impact on BMI) [Faith et al. (2011)] and BMI decrease from normal weight to underweight level (negative impact) [Carey et al. (2014)]. This kind of "U" shape impact of determinants on BMI is hardly captured by our method.[7] In the end we use the natural logarithm transformed BMI in our analysis because BMI scores are skewed toward higher values in our sample [Zeng et al. (2013)].

Through a systematic review of the literature on overweight and obesity, we test impacts of 48 factors[8] [i.e., regressors X in the model (2.2)] on BMI, including lifestyle factors such as physical activity [Galani and Schneider (2007)], alcohol consumption [Colditz et al. (1991)], smoking habits [Cawley and Scholder (2013)] and so on; socio-economic factors [Cohen et al. (2013)] such as education, income, working arrangement, etc.; and some other factors such as marital status [Sobal, Rauschenbach and Frongillo (1992)], duration of US residence [Oza-Frank and

[7]We thank one referee for pointing out that quantile regression can serve as an alternative modeling method for BMI [Koenker (2005), Zhao, Zhang and Liu (2014)]; see Section 5 for a detailed discussion.

[8]The number of factors tested is restricted by information available in the data set. For example, energy intake and dietary habit are important factors for BMI and obesity [see, for example, Hill and Peters (1998)], but information about food consumption is not available in the NHIS.

Cunningham (2010)] and depression [Faith et al. (2011)]. As discussed, a range of previous studies shows that the impacts of regressors X on BMI are varying across demographic groups [Colditz et al. (1991), Sobal, Rauschenbach and Frongillo (1992), Zhang and Wang (2004)]. Therefore, we choose categorical variables of age, gender and ethnicity as covariates, that is, Z in our model. By excluding such individuals with underweight and those having missing values of any variable involved in the model, we end up with a data set having 16593 observations. Definitions and summary statistics for all variables are presented in Table 2. Furthermore, Table 3 lists all 32 (i.e., $m = 32$) possible realizations of the covariates.

4.2. *Summary of the main findings*.

4.2.1. *Variable selection*. First of all, we implement (2.5) to estimate the optimal bandwidth parameters. Results are reported in Table 4. It can be seen that all three covariates are relevant; however, their influences on the impacts of regressors on BMI are quite different. In particular, ethnicity and gender have relatively stronger influences than age group because the smoothing parameters associated with *ethnicity* and *sex* are much smaller than that of *age*.

Based on these smoothing parameters, we then apply our method to identify the relevant and irrelevant regressors to BMI. The optimal weight parameter selected by the modified BIC-type criterion through (2.13) is $\hat{\hat{\gamma}} = 3.2$. Table 5 presents the result of variable selection through equation (2.7). 24 regressors, out of 48 in total, are identified as relevant, and the others are irrelevant to BMI.

In particular, while our estimate suggests that exercise is correlated with BMI, the level of intensity and frequency does matter. For example, compared to never doing vigorous (or strength) activity, doing such a level of exercise less than once per week has almost no effect on BMI, while doing it more than once per week starts to change BMI. In terms of light/moderate activity, however, people have to do it more than three times per week to see some effect on BMI. Results from our study may provide guidance for policy makers to adopt more efficient incentives to avoid overweight or obesity, that is, encouraging people to do more intensive exercise or to do moderate exercise more frequently rather than simply promoting exercise at any intensive level with any frequency.

Both the status of drinking and smoking and their consumption level are relevant to BMI. No impact from computer use can be seen. For socio-economic factors, education, income, and the two highest levels of occupational social class (OSC) (*occup1* and *occup2* compared to lowest OSC, i.e., *occup5*), and health professional visit in the last 12 months are identified as relevant regressors for BMI, but the two lower levels of OSC (*occup3* and *occup4* compared to *occup5*), working arrangement, working hours, house ownership, health insurance coverage and medical care expenditure are irrelevant to BMI. Among all other factors, indicators on duration of living in the U.S. (i.e., born in the U.S. and living in the U.S. more

TABLE 2
Data description and summary statistics

Variable	Definition	Mean	St.D.
Y			
BMI	body mass index	27.96	6.01
Z			
sex	0 for female and 1 for male	0.49	0.50
age	0 for age < 25, 1 for $25 \leq$ age ≤ 44, 2 for $45 \leq$ age ≤ 64, and 3 for age ≥ 65	1.39	0.75
race	0 for white, 1 for black, 2 for asian, 3 for all the other races	0.33	0.67
X			
Lifestyle factors			
vig_l0	1 if never do vigorous activities, 0 otherwise (reference group)	0.45	0.50
vig_l1	1 if do vigorous activities less than once per week, 0 otherwise	0.04	0.19
vig_l2	1 if do vigorous activities more than one time and less than three times per week, 0 otherwise	0.28	0.45
vig_l3	1 if do vigorous activities more than three times per week, 0 otherwise	0.23	0.42
mod_l0	1 if never do light/moderate activities, 0 otherwise (reference group)	0.35	0.48
mod_l1	1 if do light/moderate activities less than once per week, 0 otherwise	0.02	0.15
mod_l2	1 if do light/moderate activities more than one time and less than three times per week, 0 otherwise	0.29	0.46
mod_l3	1 if do light/moderate activities more than three times per week, 0 otherwise	0.33	0.47
str_l0	1 if never do strength activities, 0 otherwise (reference group)	0.66	0.47
str_l1	1 if do strength activities less than once per week, 0 otherwise	0.02	0.14
str_l2	1 if do strength activities more than one time and less than three times per week, 0 otherwise	0.20	0.40
str_l3	1 if do strength activities more than three times per week, 0 otherwise	0.12	0.32
smk_ed	1 if current every day smoker, 0 otherwise	0.13	0.34
smk_sd	1 if current some day smoker, 0 otherwise	0.04	0.20
smk_f	1 if former smoker, 0 otherwise	0.20	0.40

TABLE 2
(*Continued*)

Variable	Definition	Mean	St.D.
smk_n	1 if never smoke, 0 otherwise (reference group)	0.62	0.48
cigsday	number of cigarettes per day	1.98	5.52
alc1yr	1 if Ever had 12+ drinks in any one year, 0 otherwise	0.72	0.45
alc_life	1 if Had 12+ drinks in entire life, 0 otherwise	0.13	0.33
alc_c0	1 if do not drink at all currently, 0 otherwise (reference group)	0.26	0.44
alc_c1	1 if current infrequent drinker, 0 otherwise	0.12	0.33
alc_c2	1 if current light drinker, 0 otherwise	0.36	0.48
alc_c3	1 if current moderate drinker, 0 otherwise	0.19	0.39
alc_c4	1 if current heavier drinker, 0 otherwise	0.06	0.25
cpuse_0	1 if never or almost never use computer, 0 otherwise (reference group)	0.15	0.35
cpuse_1	1 if use computer for some/most days, 0 otherwise	0.18	0.38
cpuse_2	1 if use computer on every day, 0 otherwise	0.67	0.47
Socio-economic factors			
educ1	number of years of school completed	15.54	3.08
occup1	1 if management, business, science, and arts occupations, 0 otherwise	0.38	0.49
occup2	1 if service occupations, 0 otherwise	0.18	0.38
occup3	1 if sales and office occupations, 0 otherwise	0.23	0.42
occup4	1 if natural resources, construction, and maintenance occupations, 0 otherwise	0.09	0.29
occup5	1 if production, transportation, and material moving occupations, 0 otherwise (reference group)	0.12	0.33
working	1 if working or with job last week, 0 otherwise	0.88	0.32
unemp	1 if looking for job last week, 0 otherwise	0.05	0.21
nowork	1 if not working at a job last week, 0 otherwise	0.05	0.22
retired	1 if retired, 0 otherwise (reference group)	0.02	0.15

TABLE 2
(*Continued*)

Variable	Definition	Mean	St.D.
wrkhrs	hours worked last week	35.46	17.28
lnincome	nature logrithm of total earnings last year	10.20	0.94
houseown	1 if own or being bought the house, 0 otherwise	0.56	0.50
notcov	1 if not have health insurance coverage, 0 otherwise	0.20	0.40
hp	1 if ever seen/talked to health professional in the last 12 months, 0 otherwise	0.79	0.40
hce_l1	1 if amount family spent for medical care is 0, 0 otherwise (reference group)	0.13	0.33
hce_l2	1 if amount family spent for medical care is less than $500 but more than 0, 0 otherwise	0.37	0.48
hce_l3	1 if amount family spent for medical care is less than $1999 but more than $500, 0 otherwise	0.30	0.46
hce_l4	1 if amount family spent for medical care is less than $2999 but more than $2000, 0 otherwise	0.09	0.29
hce_l5	1 if amount family spent for medical care is less than $4999 but more than $3000, 0 otherwise	0.06	0.24
hce_l6	1 if amount family spent for medical care is $5000 or more, 0 otherwise	0.06	0.23
Other factors			
married	1 if married or de facto, 0 otherwise	0.51	0.50
us_born	1 if born in the US, 0 otherwise	0.81	0.39
us_m15	1 if stay in the US for more than 15 years, 0 otherwise	0.12	0.32
us_m5l15	1 if stay in the US for more than 5 years but less than 15 years, 0 otherwise	0.06	0.24
us_l5	1 if stay in the US for less than 5 years, 0 otherwise (reference group)	0.02	0.12
citizenp	1 if U.S. citizen, 0 otherwise	0.90	0.30
mental	1 if have depression/anxiety/emotional problem, 0 otherwise	0.01	0.12
rg_ne	1 if live in north east, 0 otherwise	0.16	0.37
rg_mw	1 if live in midwest, 0 otherwise	0.21	0.41
rg_sth	1 if live in south, 0 otherwise	0.36	0.48
rg_west	1 if live in west, 0 otherwise (reference group)	0.27	0.44

TABLE 3

List of realizations of covariates in the data and the percentage of observations for each group

	Male									Female								
	Age				Ethnicity					Age				Ethnicity				
GI	**<25**	**[25, 45)**	**[45, 65)**	**≥65**	**W**	**B**	**A**	**O**	**Perc**	**GI**	**<25**	**[25, 45)**	**[45, 65)**	**≥65**	**W**	**B**	**A**	**O**	**Perc**
1	x				x				3.9%	17	x				x				4.1%
2	x					x			1.0%	18	x					x			0.7%
3	x						x		0.3%	19	x						x		0.3%
4	x							x	0.1%	20	x							x	0.1%
5		x			x				17.0%	21		x			x				17.9%
6		x				x			4.3%	22		x				x			2.9%
7		x					x		1.6%	23		x					x		1.9%
8		x						x	0.4%	24		x						x	0.4%
9			x		x				14.6%	25			x		x				14.4%
10			x			x			3.1%	26			x			x			2.3%
11			x				x		1.0%	27			x				x		1.1%
12			x					x	0.2%	28			x					x	0.3%
13				x	x				2.6%	29				x	x				2.5%
14				x		x			0.4%	30				x		x			0.2%
15				x			x		0.1%	31				x			x		0.1%
16				x				x	0.1%	32				x				x	0.1%

GI = Group Index
Perc = Percentage of the whole sample
M = Male, F = Female
W = White, B = Black, A = Asian, O = Other

TABLE 4
Estimated bandwidths for covariates

Sex	0.1158	*Age group*	0.1979	*Ethnicity*	0.0703

than 15 years compared to living in the U.S. less than 5 years), living in the south (compared to living in the west), marital status and mental health problems are robust factors for BMI; however, living in the US more than 5 years but less than 15 years (compared to less than 5 years), citizenship, living in either the northeast or the middle west (compared to living in the west) have no impact on BMI.

For comparison purposes, in this BMI study we also estimate the other five models applied in Section 3, that is, the group LASSO method, the LASSO method applied to models (3.1) and (3.2), respectively, and the stepwise method

TABLE 5
List of relevant and irrelevant variables to BMI

Relevant variable	**Irrelevant variable**
Lifestyle factors	*Lifestyle factors*
vig_l2	vig_l1
vig_l3	mod_l1
mod_l3	mod_l2
str_l2	str_l1
str_l3	smk_sd
smk_ed	cpuse_1
smk_f	cpuse_2
cigsday	*Socio-economic factors*
alc1yr	occup3
alc_life	occup4
alc_c1	working
alc_c2	unemp
alc_c3	nowork
alc_c4	wrkhrs
Socio-economic factors	houseown
educ1	notcov
occup1	hce_l2
occup2	hce_l3
lnincome	hce_l4
hp	hce_l5
Other factors	hce_l6
us_born	*Other factors*
us_m15	us_m5l15
rg_sth	citizenp
married	rg_ne
mental	rg_mw

TABLE 6
Model Comparison on RME

	Vary-Coeff	GroupL	LASSO1	SW1	LASSO2	SW2
RME	0.1562	0.1609	0.1657	0.2714	0.1646	0.2846

applied to models (3.1) and (3.2), respectively. X and Z in models (3.1) and (3.2) have the same specification as what has been discussed in Section 4.1. It is worthwhile to mention that such variables selected by our method are exactly the same as those selected by the group LASSO method. To compare model performance, we calculate root leave-one-out mean squared prediction errors (RME) $RME = (\sum_{i=1}^{N}(\hat{y}_{-i} - y_i)^2/N)^{1/2}$ for each model in Table 6,[9] where $\hat{y}_{-i}$ denotes the leave-one-out prediction for the ith individual. It can be seen that our method outperforms all the other five models with the lowest RME. It is also interesting to see that the group LASSO method performs as the second best, followed by LASSO methods applied to model (3.2) (the one taking account of varying impacts of X on BMI through interaction terms between X and Z). The LASSO method applied to model (3.1) (i.e., no varying impact is accounted for) performs worse than its counterpart. Performance of the stepwise method is the worst among all options. Besides the superior performance of our method, these results also demonstrate, to some extent, that the varying impacts of potential factors on BMI are widely presented.

4.2.2. *Varying impacts*. To quantify the effects of relevant regressors on BMI, we conduct a post-selection estimation using the unregularized estimation method for the varying-coefficient model only including the relevant regressors [i.e., equation (2.9)]. For the sake of space limitation, in the supplementary file [Gao et al. (2017)] we provide the full estimation results, including point and confidence interval estimates for the relevant determinants' impacts on BMI across demographic groups. Generally speaking, these estimated coefficients confirm that the selected variables are truly relevant to BMI. Because none of these regressors have their effects over all 32 groups to be constant zero, given zero is not consistently covered by the, at least 95%, CIs[10] of the 32 varying-effects of each regressor.

[9]We also calculate RME for each of the 32 demographic groups from each method. Because of space limitations, these results are provided in the supplementary file [Gao et al. (2017)].

[10]We cannot obtain CI's for the estimates provided in (2.7). After using the procedure of variable selection, following Wang and Xia (2009), we are able to calculate the 95% CIs through bootstrap for the post-selection estimates. See Theorem 2 and the discussions under Theorem 4 of Li, Ouyang and Racine (2013) for details. We point out that these CI's should be interpreted with caution. Indeed, these CI estimates might not be reliable without further justifying the variable selection bias issue. One sufficient condition for the validity of post-selection CIs is that all true relevant regressors

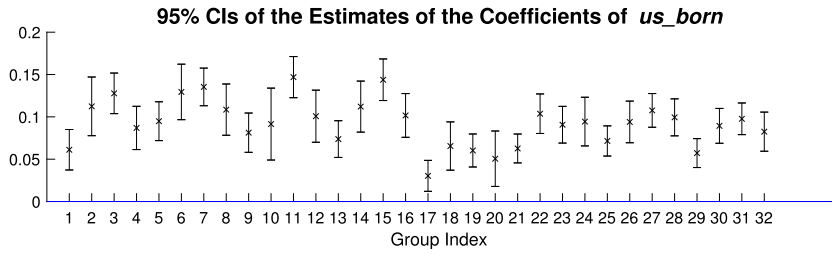

FIG. 1. *The post-selection estimates for a relevant regressor of us_born.*

Taking the regressor of *us_born* as an example, its varying effects on BMI across 32 demographic groups are shown in Figure 1. The demographic groups are indicated in the horizontal axis (for details, see Table 3). "×" represents the point estimate from the post-selection estimation, and the vertical line represents the 95% CI estimate. Two results emerge from this figure. First, the post-selection results show that the estimated effects of *us_born* on BMI are positive for all groups, which confirms that the regressor of *us_born* is truly relevant to BMI. Second, the effects of *us_born* on BMI are apparently varying across the 32 demographic groups. In particular, the effects are higher for males (groups 1–16) than females (groups 17–32) when age and race are the same, that is, group 1 vs group 17, 2 vs 18, and so forth. Furthermore, the differences are more significant for Asian groups. As shown in Figure 1, there is almost no overlap between the two corresponding CI estimates, that is, group 3 vs group 19, 7 vs 23, 11 vs 27, and 15 vs 31. Comparing across groups having the same gender and age range, *us_born* normally has higher impacts for Asian people. Taking the four youngest male groups as an example, being born in US increases BMI by 12.78% for Asians, which is higher than the increases of 6.11%, 11.24% and 8.69% for white, black and all other races, respectively.

5. Conclusions with discussions. In order to solve some challenging modeling and statistical issues existing in the literature of BMI studies, we propose a variable selection procedure for the categorical varying-coefficient model. We examine the impacts of a wide range of potential factors proposed in the huge literature on BMI and obesity by using data from the 2013 NHIS in the United States. Specifically, (1) we allow for and quantify the varying impacts of determinants on BMI by using a varying-coefficient setting; (2) we systematically justify the relative importance of demographic variables in differencing potential determinants' impacts on BMI by looking at the optimal bandwidths of demographic group variables; (3) we identify the relatively important determinants of BMI by using a group LASSO technique.

---

are successfully identified by (2.7). We refer readers to Dezeure et al. (2015) and Bühlmann and Mandozzi (2014) for other sufficient conditions with further theoretical justification.

Correspondingly, we also derive some asymptotic properties for the data-driven procedure documented in this paper. Our theoretical results show that the true model can be successfully detected with probability going to 1 under certain mild conditions. In addition, the proposed estimator also achieves asymptotic normality on the true (oracle) model whenever there is no irrelevant covariate.

In this study, we have not investigated any asymptotic behavior for the case where both p and r diverge to infinity. If we ignore the optimal bandwidth selection by using the indicator function to replace all kernel functions and let p and r diverge to infinity (let alone the fact that the number of demographic groups grows exponentially with r), the theoretical study reduces to that investigated by Lounici et al. (2011). However, to the best of our knowledge, how to achieve the optimal bandwidths for model (2.2) remains unknown for the high-dimensional case. We will pursue this in a future study.

In the end, as suggested by one referee, it is worthwhile to mention that the quantile regression model [Koenker (2005)] is an alternative approach if the interest is in some specific range (e.g., low or high) of BMI observations. In fact, a similar variable selection problem under the quantile categorical varying-coefficient model is considered by Zhao, Zhang and Liu (2014). Through using a penalized approach with both LASSO and fused LASSO [Tibshirani et al. (2005)] penalties, their method particularly advocates the fusion of categories of determinants for each regressor, hence less emphasizing varying impacts among different categories, which is the focus of our approach via a group LASSO penalty. The major difference between the proposed quantile regression procedure in Zhao, Zhang and Liu (2014) and our method is that the former cannot justify the relative importance of demographic variables while our method achieves this goal by adopting a kernel function to select optimal bandwidth in (2.5). For studies particularly interested in specific ranges of BMI, it would be more interesting to enable the corresponding quantile categorical varying-coefficient model to retrieve the information of demographic variables by properly marrying a bandwidth selection procedure and group LASSO-type penalty. We leave it as a future project.

APPENDIX: ASSUMPTIONS

ASSUMPTION 1. 1. $\{X_i, Z_i, Y_i\}_{i=1}^N$ are i.i.d. In addition, $\max_{\bar{z}\in\bar{\mathcal{D}}} \|\beta_0(\bar{z})\| < \infty$.

2. $E[Y_i^2|X_i = x, \bar{Z}_i = \bar{z}]$ is bounded on $(x, \bar{z}) \in \mathbb{R}^p \times \bar{\mathcal{D}}$.

3. Let $\sigma_\varepsilon^2(x, \bar{z}) = E[\varepsilon_i^2|X_i = x, \bar{Z}_i = \bar{z}]$ and $\sigma_\varepsilon^2(\bar{z}) = E[\sigma_\varepsilon^2(X_i, \bar{z})|\bar{Z}_i = \bar{z}]$. Then $E[\sigma_\varepsilon^2(X_i, \bar{z})X_i X_i'|\bar{Z}_i = \bar{z}]$ is positive definite for all $\bar{z} \in \bar{\mathcal{D}}$.

4. For $s = 1, \ldots, r$, the sth component of $z = (z_1, \ldots, z_r)'$ takes c_s different values in $\{0, 1, \ldots, c_s - 1\}$. Moreover, $2 \le \min_{1\le s\le r} c_s \le \max_{1\le s\le r} c_s < \infty$.

ASSUMPTION 2. 1. *Relevant Covariate Case: that is,* $\bar{r} = r$.
Define $L_{ij,\Theta} = L(Z_i, Z_j, \Theta)$, $m(Z_i) = E[X_i X_i' | Z_i]$ and

$$\eta_\beta(Z_j) = \big(E\big[X_i X_i' L_{ij,\Theta} | Z_j\big]\big)^{-1} E\big[X_i X_i' \beta(Z_i) L_{ij,\Theta} | Z_j\big].$$

Then $\Theta = 0_{r\times 1}$ are the only values of $\Theta = (\theta_1, \ldots, \theta_r)'$ that make

$$\sum_{z\in\mathcal{D}} \Pr(z)\big[\eta_\beta(z) - \beta_0(z)\big]' m(z)\big[\eta_\beta(z) - \beta_0(z)\big] = 0.$$

2. *Irrelevant Covariate Case: that is,* $\bar{r} < r$.
For $\tilde{Z}_i = (Z_{i,\bar{r}+1}, \ldots, Z_{i,r})'$, $\{\tilde{Z}_i, 1 \le i \le N\}$ is independent of all other variables and has no impact on $\beta_0(\cdot)$. Define $\bar{L}_{ij,\bar{\Theta}}$, $\bar{\eta}_\beta(\bar{Z}_j) = (E[X_i X_i' \bar{L}_{ij,\bar{\Theta}} | \bar{Z}_j])^{-1} \times E[X_i X_i' \beta(\bar{Z}_i) \bar{L}_{ij,\bar{\Theta}} | \bar{Z}_j]$ and $\bar{m}(\bar{Z}_i) = E[X_i X_i' | \bar{Z}_i]$. Then the only values of $\Theta = (\theta_1, \ldots, \theta_{\bar{r}})'$ that make

$$\sum_{\bar{z}\in\bar{\mathcal{D}}} \Pr(\bar{z})\big[\bar{\eta}_\beta(\bar{z}) - \beta_0(\bar{z})\big]' m(\bar{z})\big[\bar{\eta}_\beta(\bar{z}) - \beta_0(\bar{z})\big] = 0$$

are $\bar{\Theta} = 0_{\bar{r}\times 1}$. $\theta_s \in [0, 1]$ for $s = \bar{r} + 1, \ldots, r$.

ASSUMPTION 3. 1. For a random variable $\bar{Z}_i \in \bar{\mathcal{D}}$ and $\beta_0(\bar{Z}_i) = (\beta_{01}(\bar{Z}_i), \ldots, \beta_{0p}(\bar{Z}_i))'$, suppose there exists an integer $0 < p^* \le p$ such that $0 < E|\beta_{0j}(\bar{Z}_i)|^2 < \infty$ for $j = 1, \ldots, p^*$ and $E|\beta_{0j}(\bar{Z}_i)|^2 = 0$ for $j = p^* + 1, \ldots, p$.
2. For any $\bar{z} \in \bar{\mathcal{D}}$, $0 < \alpha_1 \le \rho_{\min} \le \rho_{\max} \le \alpha_2 < \infty$, where $\rho_{\min}$ and $\rho_{\max}$ denote the minimum and maximum eigenvalues of $E[X_i X_i' | \bar{z}]$, respectively, and α_1, α_2 are two universal positive constants.

Assumptions 1 and 2 are identical to those in Li, Ouyang and Racine (2013). Note that since the support $\mathcal{D}$ is finite, we automatically have $\Pr(z) = \Pr(Z_i = z) > \alpha_3 > 0$ with some universal constant α_3 for any $z \in \mathcal{D}$. Assumption 3.2 ensures all eigenvalues of $E[X_i X_i' | \bar{z}]$ are bounded uniformly.

SUPPLEMENTARY MATERIAL

Supplement to "Variable selection for a categorical varying-coefficient model with identifications for determinants of body mass index" (DOI: 10.1214/17-AOAS1039SUPP; .pdf). In this supplementary file, we provide a detailed presentation and discussion on (1) mathematical proofs of the main results, (2) estimation procedure of our method, (3) extra simulation results, and (4) other estimation results from the BMI study.

REFERENCES

AITCHISON, J. and AITKEN, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63** 413–420. MR0443222

ALI, S. M. and LINDSTRÖM, M. (2006). Socioeconomic, psychosocial, behavioural, and psychological determinants of BMI among young women: Differing patterns for underweight and overweight/obesity. *Eur. J. Public Health* **16** 324–330.

BÜHLMANN, P. and MANDOZZI, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comput. Statist.* **29** 407–430. MR3261821

CAREY, M., SMALL, H., YOONG, S. L., BOYES, A., BISQUERA, A. and SANSON-FISHER, R. (2014). Prevalence of comorbid depression and obesity in general practice: A cross-sectional survey. *Br. J. Gen. Pract.* **64** e122–e127.

CAWLEY, J. (2011). *The Oxford Handbook of the Social Science of Obesity*. Oxford Univ. Press, Oxford.

CAWLEY, J. and SCHOLDER, S. V. H. K. (2013). The demand for cigarettes as derived from the demand for weight control. Technical Report, National Bureau of Economic Research.

CHU, W., LI, R. and REIMHERR, M. (2016). Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. *Ann. Appl. Stat.* **10** 596–617. MR3528353

COHEN, A. K., RAI, M., REHKOPF, D. H. and ABRAMS, B. (2013). Educational attainment and obesity: A systematic review. *Obes. Rev.* **14** 989–1005.

COLDITZ, G. A., GIOVANNUCCI, E., RIMM, E. B., STAMPFER, M. J., ROSNER, B., SPEIZER, F. E., GORDIS, E. and WILLETT, W. C. (1991). Alcohol intake in relation to diet and obesity in women and men. *Am. J. Clin. Nutr.* **54** 49–55.

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840

FAITH, M. S., BUTRYN, M., WADDEN, T. A., FABRICATORE, A., NGUYEN, A. M. and HEYMSFIELD, S. B. (2011). Evidence for prospective associations among depression and obesity in population-based studies. *Obes. Rev.* **12** e438–e453.

FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

FONTAINE, K. R., REDDEN, D. T., WANG, C., WESTFALL, A. O. and ALLISON, D. B. (2003). Years of life lost due to obesity. *J. Amer. Medical Assoc.* **289** 187–193.

GALANI, C. and SCHNEIDER, H. (2007). Prevention and treatment of obesity with lifestyle interventions: Review and meta-analysis. *Int. J. Public Health* **52** 348–359.

GAO, J., PENG, B., REN, Z. and ZHANG, X. (2017). Supplement to "Variable selection for a categorical varying-coefficient model with identifications for determinants of body mass index." DOI:10.1214/17-AOAS1039SUPP.

GERTHEISS, J. and TUTZ, G. (2010). Sparse modeling of categorial explanatory variables. *Ann. Appl. Stat.* **4** 2150–2180. MR2829951

HALL, P., LI, Q. and RACINE, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Rev. Econ. Stat.* **89** 784–789.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HILL, J. O. and PETERS, J. C. (1998). Environmental contributions to the obesity epidemic. *Science* **280** 1371–1374.

HUANG, J., MA, S., XIE, H. and ZHANG, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

LI, Q., OUYANG, D. and RACINE, J. S. (2013). Categorical semiparametric varying-coefficient models. *J. Appl. Econometrics* **28** 551–579. MR3064528

LI, Q. and RACINE, J. S. (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory* **26** 1607–1637. MR2738011

LIPOWICZ, A., GRONKIEWICZ, S. and MALINA, R. M. (2002). Body mass index, overweight and obesity in married and never married men and women in Poland. *Am. J. Human Biol.* **14** 468–475.

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

MA, S., CARROLL, R. J., LIANG, H. and XU, S. (2015). Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates. *Ann. Statist.* **43** 2102–2131. MR3375878

OZA-FRANK, R. and CUNNINGHAM, S. A. (2010). The weight of US residence among immigrants: A systematic review. *Obesity Reviews* **11** 271–280.

REHKOPF, D. H., LARAIA, B. A., SEGAL, M., BRAITHWAITE, D. and EPEL, L. (2011). The relative importance of predictors of body mass index change, overweight and obesity in adolescent girls. *Int. J. Pediatr. Obes.* **6** 233–242.

SOBAL, J., RAUSCHENBACH, B. S. and FRONGILLO, E. A. (1992). Marital status, fatness and obesity. *Soc. Sci. Med.* **35** 915–923.

STICE, E., SHAW, H. and MARTI, C. N. (2006). A meta-analytic review of obesity prevention programs for children and adolescents: The skinny on interventions that work. *Psychol. Bull.* **132** 667–691.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

VON KRIES, R., TOSCHKE, A. M., KOLETZKO, B. and SLIKKER, W. (2002). Maternal smoking during pregnancy and childhood obesity. *Am. J. Epidemiol.* **156** 954–961.

WANG, H. and LENG, C. (2007). Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.* **102** 1039–1048. MR2411663

WANG, L., LI, H. and HUANG, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103** 1556–1569. MR2504204

WANG, H. and XIA, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104** 747–757. MR2541592

WHO (2015). Obesity and overweight Fact Sheet No. 311, Working paper. Available at http://www.who.int/mediacentre/factsheets/fs311/en/.

YU, Y. (2012). Educational differences in obesity in the United States: A closer look at the trends. *Obes.* **20** 904–908.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZENG, W., EISENBERG, D. T., JOVEL, K. R., UNDURRAGA, E. A., NYBERG, C., TANNER, S., REYES-GARCÍA, V., LEONARD, W. R., CASTANO, J., HUANCA, T. et al. (2013). Adult obesity: Panel study from native Amazonians. *Econ. Hum. Biol.* **11** 227–235.

ZHANG, Q. and WANG, Y. (2004). Socioeconomic inequality of obesity in the United States: Do gender, age, and ethnicity matter? *Soc. Sci. Med.* **58** 1171–1180.

ZHAO, W., ZHANG, R. and LIU, J. (2014). Regularization and model selection for quantile varying coefficient model with categorical effect modifiers. *Comput. Statist. Data Anal.* **79** 44–62. MR3227986

J. GAO
DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS
MONASH UNIVERSITY
VIC 3145
AUSTRALIA
E-MAIL: Jiti.Gao@monash.edu

B. PENG
DEPARTMENT OF ECONOMICS
UNIVERSITY OF BATH
BATH BA2 7JP
UNITED KINGDOM
E-MAIL: bp495@bath.ac.uk

Z. REN
DEPARTMENT OF STATISTICS
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA 15260
USA
E-MAIL: zren@pitt.edu

X. ZHANG
DEPARTMENT OF ECONOMICS
UNIVERSITY OF EXETER
EXETER EX4 4PU
UNITED KINGDOM
E-MAIL: x.zhang1@exeter.ac.uk