Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Estimable group effects for strongly correlated variables in linear models

Min Tsao [*]

Department of Mathematics, Southern University of Science and Technology, 1088 Xueyuan Avenue, Nanshan District, Shenzhen, Guangdong, 518055, China
Department of Mathematics & Statistics, University of Victoria, Victoria, British Columbia, Canada V8W 3R4

ARTICLE INFO

Article history:
Received 20 July 2017
Received in revised form 11 January 2018
Accepted 8 March 2018
Available online xxxx

Keywords:
Linear models
Multicollinearity
Ordinary least-squares regression
Estimable group effects
Strongly correlated predictor variables

ABSTRACT

In ordinary least-squares regression, strongly correlated predictor variables generate multicollinearity known to cause poor estimation of individual parameters of such variables and consequently difficulties in inference and prediction with the estimated model. We construct a theoretical model to study the impact of such multicollinearity on estimation of linear combinations of these parameters and uncover linear combinations that can be remarkably accurately estimated. Our results show that this type of multicollinearity represents a redistribution of information that allows some linear combinations to be extremely accurately estimated at the expense of other linear combinations becoming inestimable. Based on insights gained from studying this theoretical model, for all linear models with strongly correlated predictor variables, we develop a simple method for finding linear combinations of parameters of these variables that may be accurately estimated. Such linear combinations can be used to help resolve the aforementioned difficulties, and they provide another tool for handling multicollinearity in ordinary least-squares regression.

1. Introduction

Strongly correlated predictor variables appear often in data sets from observational studies such as social and medical studies. Such variables generate a multicollinearity problem for the (ordinary) least-squares regression. To describe the problem, consider linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_{q-1}]$ a known $n \times q$ design matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{q-1})^T$ an unknown $q \times 1$ vector of regression parameters, and $\boldsymbol{\varepsilon}$ an $n \times 1$ vector of random errors with mean zero and variance $\sigma^2\mathbf{I}$. Suppose the first p variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ are strongly correlated ($p < q < n$). Then, the resulting multicollinearity problem manifests through unusually large variances of least-squares estimators $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ for $\beta_1, \beta_2, \ldots, \beta_p$, rendering the estimators unreliable. This also makes it difficult to make inference and predications using the estimated model. Discussions about this type of multicollinearity can be found in many books on linear models, *e.g.*, Draper and Smith (1998), Belsley et al. (2004), and Montgomery et al. (2013). In practice, the problem is often handled by modified least-squares methods, such as Ridge regression (Hoerl and Kennard, 1970) and principal component regression. However, these methods are more complicated than the least-squares regression in terms of implementation, inference and interpretation. One of the

[*] Correspondence to: Department of Mathematics & Statistics, University of Victoria, Victoria, British Columbia, Canada V8W 3R4.
E-mail addresses: caom@sustc.edu.cn, mtsao@uvic.ca.

https://doi.org/10.1016/j.jspi.2018.03.003
0378-3758/© 2018 Elsevier B.V. All rights reserved.

objectives of this paper is to explore the use of the least-squares estimates for handling the consequences of multicollinearity, instead of avoiding it through more complicated alternatives, thereby making better use of the least-square regression in the presence of multicollinearity.

Although it is widely known that multicollinearity causes poor estimation of $\beta_1, \beta_2, \dots, \beta_p$ in the least-squares regression, we conducted a comprehensive search of the literature but found no discussion about its impact on the estimation of linear combinations of $\beta_1, \beta_2, \dots, \beta_p$. Silvey (1969) studied the estimation of linear combinations of all parameters $\beta_0, \beta_1, \dots, \beta_{q-1}$ but his results are not concerned with such an impact. To study this impact and to make the problem well defined, we consider

$$\Xi' = \{\xi(\mathbf{w}') \mid \xi(\mathbf{w}') = w_1'\beta_1 + w_2'\beta_2 + \cdots + w_p'\beta_p\}, \tag{2}$$

where $\mathbf{w}' = (w_1', w_2', \dots, w_p')^T$ is any $p \times 1$ vector satisfying $\sum_{i=1}^{p} |w_i'| = 1$. We call set Ξ' the class of *normalized group effects* of the p strongly correlated variables, each $\xi(\mathbf{w}')$ a *group effect* and each vector $\mathbf{w}'$ a *weight vector*. We choose constraint $\sum_{i=1}^{p} |w_i'| = 1$ instead of $\sum_{i=1}^{p} (w_i')^2 = 1$ because it allows Ξ' to include commonly used weighted averages such as $\frac{1}{p}\sum_{i=1}^{p} \beta_i$, even though it is technically more difficult to handle as it is non-smooth. We say a group effect is *estimable* if its minimum variance unbiased linear estimator has a variance that is smaller than or comparable to the error variance σ^2.

Individual parameters $\beta_1, \beta_2, \dots, \beta_p$ are special group effects in Ξ' but they are not estimable. We are interested in finding estimable effects in Ξ'; as none of the underlying parameters is estimable, such estimable effects are of theoretical interest. More importantly, the estimable effects are useful for inference and estimation concerning $\beta_1, \beta_2, \dots, \beta_p$, and for knowing when accurate predictions can be made with the estimated model. For example, if $\xi(\mathbf{w}')$ is estimable and its estimated value $\hat{\xi}(\mathbf{w}')$ is significantly different zero, then we may reject the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ and conclude that one or more parameters in this group are not zero. Also, the unknown parameters $\beta_1, \beta_2, \dots, \beta_p$ satisfy

$$\hat{\xi}(\mathbf{w}') \approx w_1'\beta_1 + w_2'\beta_2 + \cdots + w_p'\beta_p,$$

which is a constraint that reduces the parameter space for the p variables from $\mathbb{R}^p$ to essentially a line in $\mathbb{R}^p$. Such a dimension reduction can be used for estimating the unknown parameters. The set of weight vectors $\{\mathbf{w}'\}$ associated with the estimable effects also defines the region in the space of the p strongly correlated variables over which accurate predictions can be made using the least-squares estimated model.

To study estimable group effects in Ξ', we first construct a uniform model containing a single group of strongly correlated predictor variables with a uniform correlation structure. For this model, the level of multicollinearity can be quantified and its impact on estimation of group effects is mathematically tractable. Under this model, we find an optimal effect that benefits from multicollinearity in that the variance of its minimum-variance unbiased linear estimator actually decreases when the level of multicollinearity increases; that is, this effect can be more accurately estimated when the level of multicollinearity goes higher. It is rather surprising that such a linear combination of the parameters exists as these parameters themselves are not estimable at high levels of multicollinearity. Further, this optimal effect can be substantially more accurately estimated than a similar effect of a group of uncorrelated variables. The optimal effect has a simple interpretation as a *variability weighted average* of the underlying parameters. At any given level of multicollinearity, all estimable effects under the uniform model are located around it.

With the knowledge that multicollinearity can be an advantage in estimating group effects, we look for estimable effects of strongly correlated variables in a general linear model (1) which may contain more than one group of such variables with unspecified correlation structures, as well as not strongly correlated variables. For such a general model, multicollinearity becomes difficult to quantify, which makes it difficult to formulate the problem of finding estimable effects analytically. Although in principle it is possible to find such effects numerically for any group of strongly correlated variables, the computational effort required may be prohibitive. To avoid these difficulties, we make use of insights gained from the uniform model and generalize the variability weighted average effect to a group of strongly correlated variables in model (1). The (generalized) variability weighted effect is easy to compute and always accurately estimated in our numerical examples. It provides a simple means for finding other estimable effects in (1) as these are also located around it.

The focus of this paper is on finding estimable effects, but we will briefly discuss how such effects may be used to ensure prediction accuracy of the least-squares estimated model. More applications are given in Tsao (2017), and one of these is a constrained local regression method that uses the variability weighted average effect to estimate the underlying parameters of strongly correlated variables. This method complements the Ridge regression and other penalized least-squares methods such as Lasso (see, *e.g.*, Hastie et al., 2015) in that it is a local method for estimating parameters of only strongly correlated variables; the ordinary least-squares estimates of other parameters are unchanged.

The rest of this paper is organized as follows. In Section 2, we introduce the uniform model under which we reduce Ξ' to a subclass Ξ; this subclass is only "$(1/2^p)$th" the size of Ξ' but it contains all effects that can be most accurately estimated. We then find the optimal variability weighted average effect through Ξ and give a characterization of all estimable effects under the uniform model using this effect as a reference point. In Section 3, we go beyond the uniform model and generalize the variability weighted average effect to a group of strongly correlated variables in model (1). We give numerical examples demonstrating the remarkable accuracy at which it is estimated under severe multicollinearity. We also briefly discuss the application of estimable effects in the prediction accuracy problem of the least-squares estimated model. We conclude with a few remarks in Section 4. All proofs of theorems and corollaries are given in the Appendix.