

Chapter 1

Estimation Strategies

Abstract Inference under uncertain prior information has been common for several decades in statistical and related literature. In the context of a regression model, we introduce the basic notion of full model, submodel, pretest, and shrinkage estimation strategies. We briefly discuss some penalty estimators and compare it with nonpenalty estimators.

Keywords Regression model · Pretest and shrinkage estimation · Penalty estimation

1.1 Introduction

For the past few decades, the use of *uncertain prior information* (UPI) or nonsample information (NSI) has been common in the statistical inference of *conditionally specified models*. Uncertain prior information is generally incorporated in a model via a conjectured constraint on the model parameters, thus giving rise to candidate submodels. The constraint on the parameters is usually obtained through expert opinion or from past data. Alternatively, candidate submodels can be obtained by using existing variable selection techniques, such as AIC or BIC, among others, when models are assumed to be sparse. Generally speaking, a submodel is more feasible and practical in a host of applications. When the conjectured constraint holds, or the submodel based on variable selection approach is true, analysis of such submodels leads to efficient statistical inferences than would be obtained through the full model. No matter how the submodel is obtained, either by variable selection technique or by imposing the constraints based on expert opinion, the estimators based on a submodel may become considerably biased and inefficient if the submodel is not the true representative of the data at hand. Recently, simultaneous variable selection and estimation of submodel parameters has become quite popular. However, such procedures may also produce biased estimators of the candidate submodel parameters. An example of a procedures along this line of thought is *penalty estimation* and their variants.

A different, but related breed of estimators, existed in the literature for more than five decades, is the so-called pretest and shrinkage estimators which attempt to incorporate uncertain prior information into the estimation procedure. Bancroft (1944) suggested the *pretest estimation* (PE) strategy. The pretest estimation approach uses a test to decide the estimator based on either the full model or a submodel. Broadly speaking, Bancroft (1944) suggested two seemingly unrelated problems on pretesting strategy. One is a data pooling problem from various sources based on a pretest approach and other one is related to simultaneous model selection and pretest estimation problem in regression model. Since then a bulk of research has been done using the pretest estimation procedure in many applications. However, the pretest estimator can also be seen as arising indirectly from the study of the admissibility of the sample mean vector in a multivariate normal distribution (Stein 1956; James and Stein 1961). In fact, the James–Stein shrinkage estimation technique may be regarded as a smooth version of the Bancroft (1944) pretest estimation procedure.

Since its inception, both pretest and shrinkage estimation strategy has received considerable attention from researchers and practitioners. More importantly, it has been analytically demonstrated that shrinkage estimation strategy outshines the classical estimators in many scenarios. During the past two and half decades, Ahmed and his co-researchers, among others, have demonstrated that shrinkage estimators dominate the classical estimators in terms of *mean squared error* (MSE) for a host of statistical models. A detailed description of shrinkage estimation and large sample estimation techniques in a regression model can be found in Ahmed (1997). Pretest, shrinkage and likelihood-based methods continue to play vital roles in statistical inference. These strategies can be used for both variable selection and post estimation. Further, pretest and shrinkage methods provide extremely useful techniques for combining data from various sources. For instance, estimation methodologies in the arena of meta analysis are essentially of this type.

To fix the idea of pretest and shrinkage estimation, let us consider for the rest of this chapter the estimation problem in a multiple linear regression model and the remaining discussion follows. If $\mathbf{Y} = (y_1, y_2, \dots, y_n)'$ is a vector of responses, and the superscript $(\cdot)'$ denotes the transpose of a vector or matrix, and $\mathbf{X}$ is an $n \times p$ fixed design matrix, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is an unknown vector of parameters, then we have the following regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1.1)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ is the vector of unobservable random errors.

Suppose that we have a set of covariates to fit a regression model for the purpose of predicting a response variable. If a *a priori* is known or suspected that a subset of the covariates do not significantly contribute to the overall prediction of the average response, they may be left aside and a model without these covariates may be considered (sparse model). In other situations, a subset of the covariates may be considered as a nuisance, meaning that they are not of primary interest to the researcher but they cannot be completely ignored either (model with nuisance parameters). In such cases, the effect of the nuisance parameters must be taken into account in estimating

the coefficients of the remaining regression parameters. A candidate model for the data that involves only the important covariates (or active predictors) in predicting the response is called a restricted model or a submodel since we use a subset of the covariates after putting some restrictions on the nuisance covariates. On the other hand, a model that takes all the variables into consideration is termed as the unrestricted model/candidate full model or simply the full model.

Full Model Estimation

Based on the sample information only, the full model or unrestricted estimator (UE) of the regression coefficient is given by

$$\hat{\beta}^{\text{UE}} = (X'X)^{-1}X'Y.$$

Uncertain Prior Information

Suppose based on prior information it is possible to restrict the regression coefficients to a given subspace as follows:

$$H\beta = h,$$

where, H is a known matrix and h is a vector of known constants.

Auxiliary Information

A more interesting application of the above restriction is that β can be partitioned as $\beta = (\beta_1', \beta_2')'$. The subvectors β_1 and β_2 are assumed to have dimensions p_1 and p_2 respectively, with $p_1 + p_2 = p$. In high-dimensional data analysis, it is assumed that the model is sparse. In other words, it is plausible that β_2 may be set to a null vector. This auxiliary information regarding subvector β_2 can be obtained by applying existing variable selection procedures.

Submodel Estimation

Thus, under the subspace restriction the submodel estimator or restricted estimator (RE) is given by

$$\hat{\beta}^{\text{RE}} = \hat{\beta}^{\text{UE}} - (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\hat{\beta}^{\text{UE}} - h).$$

Similarly, a submodel estimator can be constructed using auxiliary information, that is $\beta_2 = \mathbf{0}$.

1.1.1 Pretest and Shrinkage Estimation

In a regression setup, pretest and shrinkage estimation arise when we have prior information about a subset of the available covariates. The prior information may or may not positively contribute to the estimation process. Nevertheless, it may be

advantageous to use the UPI when sample information is limited or the quality of the data at hand is poor, or even the estimate based on sample data may not be reliable. It is however, important to note that the consequences of incorporating UPI depend on the quality or usefulness of the information being added into the estimation process. Any uncertain prior information may be tested before incorporating it into the model. Based on the idea of Bancroft (1944), uncertain prior information may be validated through a pretest, and depending on such validation, the information may be incorporated into the model as a parametric restriction, thus choosing between the submodel and the full model estimation procedure.

Once we have a submodel estimator along with the full model estimator, we can test the validity of the subspace information, using a suitable test statistic ϕ_n . In pretest estimation framework, we consider testing the restriction in the form of the following null hypothesis:

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}.$$

The preliminary test estimator (PE) or simply pretest estimator for the regression parameter $\boldsymbol{\beta}$ is obtained as

$$\hat{\boldsymbol{\beta}}^{\text{PE}} = \hat{\boldsymbol{\beta}}^{\text{UE}} - (\hat{\boldsymbol{\beta}}^{\text{UE}} - \hat{\boldsymbol{\beta}}^{\text{RE}})I(\phi_n < c_{n, \alpha}), \quad (1.2)$$

where $I(\cdot)$ is an indicator function, and $c_{n, \alpha}$ is the $100(1 - \alpha)$ percentage point of the test statistic ϕ_n .

In the framework proposed by Stein (1956), the shrinkage estimator or Stein-type estimator takes a hybrid approach by shrinking the full model estimator to a plausible alternative estimator (submodel estimator). In this framework, the estimates are essentially being shrunken toward the submodel estimators.

A Stein-type shrinkage estimator (SE) $\hat{\boldsymbol{\beta}}^{\text{S}}$ of $\boldsymbol{\beta}$ can be defined as

$$\hat{\boldsymbol{\beta}}^{\text{S}} = \hat{\boldsymbol{\beta}}^{\text{RE}} + (\hat{\boldsymbol{\beta}}^{\text{UE}} - \hat{\boldsymbol{\beta}}^{\text{RE}}) \left\{ 1 - k\phi_n^{-1} \right\}, \quad k \geq 3.$$

The *positive-part shrinkage estimator* (PSE) has been suggested in the literature. A PSE has the form

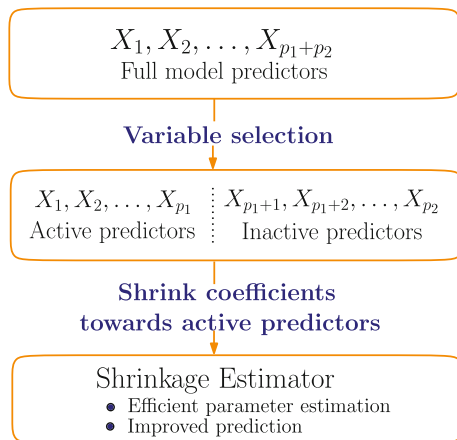
$$\hat{\boldsymbol{\beta}}_1^{\text{S}+} = \hat{\boldsymbol{\beta}}_1^{\text{RE}} + (\hat{\boldsymbol{\beta}}_1^{\text{UE}} - \hat{\boldsymbol{\beta}}_1^{\text{RE}}) \left\{ 1 - k\phi_n^{-1} \right\}^+,$$

where we define the notation $z^+ = \max(0, z)$. This adjustment controls for the over-shrinking problem in SE.

Shrinkage estimation method combines estimates from the candidate full model and a submodel in an optimal way dominating the full model estimator.

For a sparse model, a schematic flowchart of shrinkage estimation is shown in Fig. 1.1.

Fig. 1.1 Flowchart of shrinkage estimation strategy



1.1.2 Penalty Estimators

The penalty estimators are members of the penalized least squares family and they are obtained by optimizing a quadratic function subject to a penalty. Popular penalty estimators include the least absolute and shrinkage operator (LASSO), adaptive LASSO, group LASSO, the smoothly clipped absolute deviation (SCAD), and min-max concave penalty (MCP), among others. They are so called because the penalty term is used in the penalized least squares to obtain the estimate of the regression parameters. The main idea in such estimation rule is that the regression coefficients with weak or no effects are shrunk toward the zero vector when the model is sparse. The process often eliminates some of the covariates completely by shrinking their coefficients to exactly zero. Thus, the procedure performs variable selection and shrinkage together. In a sense, the procedure selects a submodel and estimates the regression parameters in the submodel. This technique is fruitful when the model is sparse, and applicable when the number of predictors (p) is greater than the number of observations (n), a so-called ultrahigh dimensional model.

In shrinkage estimation strategies, if prior information about a subset of the covariates is available, then the estimates are obtained by incorporating the available information. However, in the absence of prior information, one might go with the usual variable selection process to sort out the significant covariates. A typical variable selection process is likely to drop some of the covariates from the model. This is equivalent to having some prior knowledge about the dropped covariates that their coefficients are zero. Any suitable model selection criteria such as AIC, BIC, or penalty can be considered to achieve this objective.

Examining side-by-side the shrinkage and penalty estimation strategies, we see that the output of penalty estimation resembles shrinkage methods as it shrinks and selects the variables simultaneously. However, there is an important difference in how the shrinking works in penalty estimation when compared to the shrinkage estimation.

The penalty estimator shrinks the full model estimator toward zero and depending on the value of the tuning or penalty parameter, it sets some coefficients to zero exactly. Thus, penalty procedure does variable selection automatically by treating all the variables equally. It does not single out nuisance covariates, or for that matter, the UPI, for special scrutiny as to their usefulness in estimating the coefficients of the active predictors. However, SCAD, MCP, and adaptive LASSO, on the other hand, are able to pick the right set of variables while shrinking the coefficients of the regression model. Lasso-type regularization approaches have some advantages of generating a parsimony sparse model, but are not able to separate covariates with small contribution and covariates with no contributions. This could be a serious problem if there was a large number of covariates with small contributions and were forced to shrink toward zero. In the reviewed published studies on high-dimensional data analysis, it has been assumed that the signals and noises are well separated.

Apparently, there has been no study in the reviewed literature at one place which compares the relative performance of pretest, shrinkage, and penalty estimators. One of the objectives of the book is to present the relative properties of pretest, shrinkage, and penalty estimators in some useful models. Therefore, it is worth exploring the performance of the penalty and shrinkage estimators when it is suspected *a priori* that the parameters may be reduced to a subspace.

This presentation is fundamental, because pretest and Stein's methods may be appropriate for model selection problems and there is a growing interest in this area to fill the gap between two competitive strategies. Further, there is a dire need to investigate a more realistic case where there are small signals in the model and it cannot be effectively removed from the noise. The goal of this monograph is to discuss some of the issues involved in the estimation of the parameters in a host of models that may be over-parameterized by including too many variables in the model using shrinkage and pretest estimation strategies. For example, in genomics research, it is common practice to test a candidate subset of genetic markers for association with disease (e.g., Zeggini et al. 2007). Here the candidate subset is found in a certain population by doing genome wide association studies. The candidate subset is then tested for disease association in a new population. In this new population, it is possible that genetic markers not found in the first population are associated with disease. Shrinkage strategy is generally used to trade-off between bias and efficiency in a data adaptive way in order to provide meaningful analytical solutions to problems in genetic epidemiology and other applications.

1.2 Organization of the Book

This book is divided into six chapters.

In Chap. 2, we presented linear shrinkage and shrinkage pretest strategies for estimating the normal and Poisson means from a single sample, respectively. Asymptotic properties of the estimators have been proposed and compared with the classical estimators.

In Chap. 3, we extend the idea of pretest shrinkage strategies to a multiple sample case. Several estimators are proposed when multiple samples are available. For the multisample case, we demonstrate that the suggested shrinkage strategy is superior to the classical estimation strategy based on the sample data alone.

Shrinkage and pretest estimation in a multiple regression model are introduced in Chap. 4. Asymptotic bias and risk expressions of the estimators have been derived and the performance of shrinkage estimators is compared with the classical estimators using a real data example and through Monte Carlo simulation experiments. Several penalty estimators such as LASSO, adaptive LASSO, and SCAD estimators have been discussed. Monte Carlo simulation studies have been used to compare the performance of shrinkage and penalty estimators.

We discuss shrinkage, pretest, and penalty estimation, in partially linear model in Chap. 5. The risk properties of the estimators have been studied using asymptotic distributional risk and Monte Carlo simulation studies.

In Chap. 6 we considered shrinkage estimation in a Poisson regression model. This model assumes the response variable to have a Poisson distribution, and also that the logarithm of its expected value can be modeled by a linear combination of unknown parameters. We appraise the performance of shrinkage, penalty, and maximum likelihood estimators in this context with real data example and through Monte Carlo simulation experiments.

References

- Ahmed, S. E. (1997). Asymptotic shrinkage estimation: The regression case. *Applied Statistical Science*, *II*, 113–139.
- Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significances. *Annals of Mathematical Statistics*, *15*, 190–204.
- James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 361–379).
- Stein, C. (1956). The admissibility of hotelling's T^2 -test. *Mathematical Statistics*, *27*, 616–623.
- Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., et al. (2007). Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, *316*(5829), 1336–1341.

Penalty, Shrinkage and Pretest Strategies

Variable Selection and Estimation

Ahmed, S.E.

2014, IX, 115 p. 6 illus., Softcover

ISBN: 978-3-319-03148-4