CHAPTER 14

Big Data Analytics

Big data is a marquee topic. The term "big data" commonly refers to data sets that have one or more of these features: high dimension, high variety, high volume, and high velocity for the speed of collecting data. The availability of big data opens new areas of theoretical and empirical analysis such as incorporating artificial intelligence and machine learning techniques (e.g., Athey 2018; Athey et al. 2017; Chernozhukov et al. 2018; Chernozhukov, Hausman, and Newey 2018); one-dimensional asymptotics and multidimensional asymptotics (e.g., Bai and Saranadasa 1996; Hsiao and Zhou 2018; Phillips and Moon 1999); dimension reduction (e.g., Chen et al. 2019; Chudik, Kapetanios, and Pesaran 2018; Fan and Kim 2018); functional dynamics (e.g., Cai et al. 2018; Chang, Hu, and Park 2018; Chang, Kim, and Park 2018; Li et al. 2018; Phillips 1974; Robinson 1976); smooth versus abrupt structural change (e.g., Chen and Hong 2012; Sun, Hong, and Wang 2018, 2019); combining data of different sources and/or different time frequencies (e.g., Chen 2018; Chow and Lin 1976; Hsiao 1979c; Maddala 1971b); one stage modeling or multi-stage modeling (e.g., Lindley and Smith 1972). For general discussion, see Athey (2019), Hsiao (2020a), and Varian (2014). In this chapter, instead of providing a comprehensive discussion of these topics, I shall briefly introduce the basic machine learning algorithms and discuss the challenges to panel data analysis from the perspective of essential roles of econometric analysis: (i) empirically verifying economic theory, (ii) predicting the future, and (iii) measuring or simulating impacts of social policies.

Section 14.1 briefly reviews some machine learning methods that are useful to econometricians. Section 14.2 covers inference with high-dimensional data. Section 14.3 discusses inference methods for a low-dimensional set of variables of interest to economists in light of the presence of high-dimensional data. Section 14.4 examines the challenges to the prediction procedures.

14.1 MACHINE LEARNING ALGORITHMS

The sheer size of the data requires powerful data manipulation tools. It is often necessary to do some exploratory data analysis to summarize information in the data. Machine learning (ML) refers the application of artificial intelligence (AI) to process a massive amount of data through *pattern recognition* and improving from *experience*. In this way, ML provides computer systems with the ability to automatically perform specific tasks or predictions. ML can generally be classified into *supervised learning* and *unsupervised learning*. Supervised learning is to identify information with specific objectives (target variables). It consists of labeled data with features. Unsupervised learning has no specific

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

objective. The machine studies data to identify patterns. There is no answer key. The machine determines groups by parsing the data.

There are different types of ML algorithms. Each has a unique characteristic for a specific use. Among the popular learning models are:

14.1.1 Regression Analysis

This involves predicting the outcome using the regression methods in statistics to show the relationships between a dependent variable y and explanatory (or conditional) variables, $\boldsymbol{x}$, when the value of conditional variables changes.

14.1.2 Decision Trees and the Random Forest Algorithm

The decision tree approach is an *explorative* modeling approach. The approach classifies an observed sample (*training sample*) using a tree-like model to split the data into subsets based on the presence of certain attributes. A decision tree takes a flowchart structure. The *root* is the *target variable* (say, weather outlook), an internal *node* signifies the presence of an attribute, the *branch* is a connecting node that connects the data groups that possess certain attributes (nodes) (say, high, normal, or low humidity), and the *leaf* node (*class label*) is a terminal node that signifies observations or a percentage of observations possessing the relevant attributes (say, sunny, normal humidity, weak wind speed). However, the growing of a tree (sample splits) needs to be "trained." At each node, each candidate splitting a field must be sorted before its best split can be found. The best split is determined in a recursive manner called *recursive partitioning*. The recursion is stopped when the subset at a node all has the same value of the target variable or when further splitting does not improve the prediction of the target variable based on a certain criterion (say, prediction mean square error).

The advantages of the decision tree approach are that it is computationally simple, as it performs classification without having to estimate the parameter of the underlying model; and it provides a clear indication of which features (or variables) are most important for prediction. The disadvantages are that it is prone to classification errors and sampling errors, particularly if the (training) sample is relatively small in relation to classes (attributes). To overcome the disadvantages of making decision relying on a single decision tree, Breiman (1996, 2001) suggest random forest algorithms to form a decision from multiple trees working together (*ensemble*). The process of generating *multiple uncorrelated trees* that have the same probability distribution as the sample is through randomizing the observed features and randoming samples from the observed sample with replacement (the bootstrap method; Efron 1982). The steps are:

1. If there are n observations in the data, randomly draw n sample *with replacement* from the original data.
2. If there are M input variables (features), a number $m \ll M$ is specified such that at each node, m variables are selected at random out of the M, and the best split on these m variables is used to split the node.
3. The value of n and m are held constant during the growth of the forest.

14.1.3 Support Vector Machines and K-Means Clustering

These types of algorithms are data-driven methods to find similarities between data points and categorize them into a number of different groups. The splitting process is through

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

finding the optimal *hyperplane* (the distance from the hyperplane to the closest vector points) or through *agglomerative* (taking the two most similar clusters and merging them) or *divisive* (objects starting in the same cluster are divided into separate clusters). The process is stopped when a desired number of groups is achieved.

14.1.4 Neural Networks

Systems perform the task by emulating human behavior. The patterns in the data are detected through three layers: an (observed) input layer, an (observed) output layer, and one or many hidden layers, *neurons*. The weight of the neuron inside the hidden layer is determined through some statistical method (e.g., White 1989).

All these AI algorithms are aimed at getting good *out-of-sample predictions*. A *good-in-sample fit* could fail miserably out-of-sample. To avoid overemphasis of "model accuracy" with a given data set in those programs, often *penalty terms* based on statistical criteria (e.g., LASSO (Tibshirani 1996), Bayesian neural networks (Mullachery et al. 2018)) are added to the objective functions. *Cross-validation* is also used to avoid deceptive accuracy of a selected model by randomly dividing the data set into a "training" sample in which a model is selected, and a "validating" sample in which the predictability of the selected model is evaluated by some criteria.

The first two types of algorithms are often used in supervised learning. The third type of algorithm is often used for unsupervised learning. The fourth type of algorithm is mostly used for *reinforcement* learning, where the rules (of the game) and objectives are clearly defined. However, a combination of two or more types of algorithms have also been used for either supervised or unsupervised learning to improve the efficiency in determining correlations and relationships among variables

ML algorithms work well when there is a clearly defined task, performance metric, and learning experience. ML is a powerful tool for extracting and scaling up economically meaningful information from massive data warehouses. There is a fast-growing literature on ML. Athey and Imbens (2019), Bresson (2020), Fomby (2020), and Varian (2014) have provided reviews of ML methods that are useful to economists and econometricians. Abraham et al. (2021) provided examples showing that using the ML tool can provide significant enhancement in the areas of automation, economic growth, structural transformation, and distribution of income. Matillion (2019) and Nevale (2019), among others, have provided helpful user guides.

14.2 INFERENCE WITH HIGH-DIMENSIONAL DATA

The basic analytic tool for the big data approach is the same as traditional cross-sectional analysis that when $\boldsymbol{x}_i = \boldsymbol{x}_j = \boldsymbol{a}$,

$$E(y_i|\boldsymbol{x}_i = \boldsymbol{a}) = E(y_j|\boldsymbol{x}_j = \boldsymbol{a}). \tag{14.2.1}$$

The difference between y_i and y_j is attributed to the difference between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as well as the outcomes of a chance mechanism. In the traditional cross-section approach, whether $E(y_i|\boldsymbol{x}_i)$ is linear or nonlinear, the dimension of $\boldsymbol{x}_i$ is fixed, and usually small. The big data approach tries to capture the heterogeneity (or difference) between y_i and y_j through observed high-dimensional data. In other words, the dimension of $\boldsymbol{x}_i$ in the big data approach is typically high and could also increase with the accumulations of data.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

14.2.1 Regression Approach-Variable Selection in High Dimension

Big data typically contains information on a large number of variables, say m, where m could even exceed the size of the sample n. The curse of dimensionality poses severe challenges to extract meaningful models. One approach to address this problem is to select a subset of available variables based on the idea of *sparsity* in neural science. That is, it is to reduce the number of variables required to describe the statistical relationship, in order to improve model interpretation and prediction.

Let $\boldsymbol{w}_i = (w_{1i}, \ldots, w_{mi})$ be a vector of variables of a random sample of size n. A widely used approach in a linear regression framework is to solve the optimization problem,

$$\min \sum_{i=1}^{n} L(y_i - \sum_{j=1}^{p} \beta_j w_{ji}) + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta}), \tag{14.2.2}$$

where L(.) denotes the loss function, $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is a penalty function for the inclusion of variables w_{ji} measured in terms of β_j, and $\boldsymbol{\lambda}$ is a vector of tuning parameters set by the researcher. For instance, Tibshirani (1996) suggests $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ to be proportional to the l_1 norm, $|\boldsymbol{\beta}|$, of $\boldsymbol{\beta}$ that yields the well known *Least Absolute Shrinkage and Selection Operator* (LASSO). Su et al. (2016) use the l_2 norm, $||\boldsymbol{\beta}||$, and Zou and Hastie (2005) use a combination of l_1 and l_2 norm.

Alternatively, Chudik et al. (2018) put the selection of variables in a hypothesis testing framework and suggest a *One Covariate at a Time Multiple Testing* (OCMT) method. The implication of OCMT starts with a fundamental set of k-dimensional covariates $\boldsymbol{x}_i = (\boldsymbol{x}'_{1i}, \ldots, \boldsymbol{x}'_{ki})'$ that a researcher considers important among $\boldsymbol{w}_i = (\boldsymbol{x}'_i, \boldsymbol{z}'_i)$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $M_x = I_n - X(X'X)^{-1}X'$. The coefficients $\boldsymbol{\phi}$ for a linear regression model,

$$y_i = \boldsymbol{x}'_i\boldsymbol{\beta} + \boldsymbol{z}'_i\boldsymbol{\phi} + u_i, \quad i = 1, \ldots, n, \tag{14.2.3}$$

are $[EZ'M_xZ]^{-1}[EZ'M_x\boldsymbol{y}]$, where E denotes the expected value operator, and $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m-k})'$. The decision rule of whether to include z_{li} in the model is based on one-at-a-time regression of y_i on $\boldsymbol{x}_i$ and z_{li}. The variable z_l is included as an additional explanatory variable if the estimated t-value for ϕ_l exceeds the critical value functon,

$$C_{lp}(n, \delta) = \Phi^{-1}\left(1 - \frac{p}{2n^{\delta}}\right), \tag{14.2.4}$$

where p is the conventional significance level, $\Phi(\cdot)$ is the standard normal distribution function, and δ is determined by the researcher on the importance of the net contribution of z_l variable in relation to the overall net contribution of the complete set of $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m-k})'$. When $\delta = 0$, the critical value is just the conventional t-test for $\phi_l = 0$ with significance level p.

Both the penalty function approach and the OCMT work well in selecting a small set of regressors if the regressors are only weakly correlated. If the regressors are highly correlated, then the selected subset of variables may include too many noncausal variables. However, multicollinearity is frequently encounted in empirical studies.

Suppose the correlated $\boldsymbol{z}_l = (z_{l1}, \ldots, z_{ln})'$, $l = 1, \ldots, m-k$ can be decomposed in the factor form,

$$\boldsymbol{z}_l = F\boldsymbol{\gamma}_l + \boldsymbol{\varepsilon}_l \tag{14.2.5}$$

where $F = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r)$ are the $n \times r$ common factors, and $\boldsymbol{\varepsilon}_l$ are uncorrelated with F, then

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{\phi} + \boldsymbol{u} \tag{14.2.6}$$

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

can be rewritten in the form

$$\boldsymbol{y} = X\boldsymbol{\beta} + F\boldsymbol{\delta} + E\boldsymbol{\phi} + \boldsymbol{u}, \tag{14.2.7}$$

where the $r \times 1$ vector $\boldsymbol{\delta} = \Gamma\boldsymbol{\phi}$, $\Gamma = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{m-k})$, and $E = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{m-k})$. The common factors, F, can be estimated from Z as $\sqrt{n}$ times the eigenvectors corresponding to the largest r eigenvalues of the determinant equations,

$$\left|\frac{1}{n}Z'Z - \delta I_{m-k}\right| = 0, \tag{14.2.8}$$

where $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m-k})'$. Let $\tilde{X} = (X\ F)$ and $\tilde{M} = (I_n - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')$. Multiplying $\tilde{M}$ to (14.2.7) yields

$$\boldsymbol{y}^* = E^*\boldsymbol{\phi} + \boldsymbol{u}^* \tag{14.2.9}$$

where $\boldsymbol{y}^* = \tilde{M}\boldsymbol{y}, E^* = \tilde{M}E$, and $\boldsymbol{u}^* = \tilde{M}\boldsymbol{u}$. E^* is now uncorrelated. Hence, Sharifvaghefi (2020) suggests applying the OCMT to (14.2.9) to select a subset of Z for inclusion in model (14.2.3) if the t-value of the estimated $\hat{\boldsymbol{\phi}}_l$ exceeds the critical value given by (14.2.4). Sharifvaghefi (2020) further derived the mean square error of the estimated $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}}^*$ for the model (14.2.3) after implementing his generalized OCMT variable selection method, where $\boldsymbol{\phi}^*$ denotes the subset of coefficients $\boldsymbol{\phi}$ after selecting the subset of $\boldsymbol{z}$, $\boldsymbol{z}^*$, for inclusion in the model.

14.2.2 Nonparametric Method

Traditional nonparametric methods to capture the heterogeneity across individuals consider $E(y_i|\boldsymbol{x}_i)$ as a nonlinear function of $\boldsymbol{x}_i$ where the dimension of $\boldsymbol{x}_i$ is fixed, such as using the nearest-neighbor matching, kernel methods or the series method (e.g., Chen 2007; Li and Racine 2007). However, if the dimension of the conditional covariates, $\boldsymbol{x}_i$, increases with sample size, n, then there is a question about their applicability. Wager and Athey (2018), Wang (2020), and Wang and Huang (2019), etc. have explored the use of ML methodology to obtain nonparametric estimates of heterogeneous individuals. Their big data approach assumes that the outcomes of an individual can be captured by all those with observed features $\boldsymbol{x}_i$ in the *neighborhood*. In other words, the big data approach of predicting individual outcomes depends on the norm of selecting the closest neighborhood.

14.2.2.1 Random Forest Approach

The random forest algorithm, introduced by Breiman (2001), is used to obtain predictions by taking averages of the predictions from a number of decision trees constructed through random drawing of a subset of features together with random drawing from the dataset with replacement. It can be considered a data-driven nearest neighbor method except that the neighbor is a decision tree. Wager and Athey (2018) (WA) suggest building decision (causal) trees to resemble regression analogues as closely as possible.

Suppose there are n independent samples of $(y_i, \boldsymbol{x}_i)$ where $\boldsymbol{x}_i$ is of dimension k. WA suggest using the random forest algorithm to recursively split the feature space ($m \ll M$) until it is partitioned into a set of leaves L, each of which contains only a few training samples. Then, given a (test) point $\boldsymbol{X}$, $E(y|\boldsymbol{X})$ is identified by the leaf $L(\boldsymbol{X})$ containing $\boldsymbol{X}$,

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

$$\hat{\mu}(\boldsymbol{X}) = \frac{1}{\sum_{i=1}^{n} 1(\boldsymbol{x}_i \in L(\boldsymbol{X}))} \sum_{i=1}^{n} y_i \, 1(\boldsymbol{x}_i \in L(\boldsymbol{X})), \tag{14.2.10}$$

where $1(\cdot)$ denotes the indicator function and the splits are restricted so that each leaf (decision) of the tree must contain m or more sample observations. Since the number of observations in each leaf is different, (14.2.10) may be considered as a weighted average of the nearest neighborhood estimator where the closeness is defined in terms of leaves.

To improve the efficiency of (14.2.10), WA suggest splitting the sample randomly into G groups, $g = 1, \ldots, G$, with replacement,

$$\hat{\mu}_g(\boldsymbol{X}) = \frac{1}{\sum_{i \in g} 1(\boldsymbol{x}_i \in L(\boldsymbol{X}))} \sum_{i \in g} y_i \, 1(\boldsymbol{x}_i \in L(\boldsymbol{X})), \tag{14.2.11}$$

then obtain the ensemble estimate

$$\hat{\mu}(\boldsymbol{X}) = \frac{1}{G} \sum_{g=1}^{G} \hat{\mu}_g(\boldsymbol{X}). \tag{14.2.12}$$

WA give conditions for (14.2.12) to be asymptotically normally distributed. However, it is difficult to obtain the asymptotic variance by adapting the random forests algorithm because the estimation of variances of random forests estimators is very complicated.

14.2.2.2 Nearest Neighbor Approach

Fan, Lv, and Wang (2018) modify and enhance the traditional nonparametric nearest neighbors method with the machine learning algorithms to estimate the outcomes of heterogeneous individuals. The classical k-nearest neighbor (k-NN) for nonparametric regression is that given a fixed point X, the sample is relabeled in terms of the order of some distance measure, say the Euclidean norm,

$$||\boldsymbol{x}_{(1)} - X|| \leq ||\boldsymbol{x}_{(2)} - X|| \leq \cdots \leq ||\boldsymbol{x}_{(n)} - X||, \tag{14.2.13}$$

where $||\cdot||$ denotes the Euclidean distance. Let $y_{(i)}$ denote the relabeled y_i associated with $\boldsymbol{x}_{(i)}$ from the pair $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$. The 1-nearest neighbor estimator (1-NN) of $E(y|X)$ is

$$\hat{\mu}_{1-NN}(X) = y_{(1)}. \tag{14.2.14}$$

The k-NN estimator is the one that uses the average of the first k nearest neighbors,

$$\hat{\mu}_{k-NN}(X) = \frac{1}{k} \sum_{i=1}^{k} y_{(i)}. \tag{14.2.15}$$

Fan et al. (2018) propose a distributional nearest neighbors estimator (DNN). The steps are:

Step 1: From the sample, randomly draw a subsample of size $j \leq n$.

Step 2: Let $\hat{\mu}_{1,j}(X)$ denote the 1-nearest neighbor estimator based on a subsample of size j.

Step 3: Repeat steps 1 and 2 with different sample size such that $j_1 < j_2 < \cdots < j_s$ and $j_s \leq n$.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Step 4: Obtain the 1-nearest neighbor estimator as the average of s 1-nearest neighbor estimator

$$\hat{\mu}_{1,s}(X) = \frac{1}{s}\sum_{l=1}^{s}\hat{\mu}_{1,j_l}(X). \tag{14.2.16}$$

Because the estimator $\hat{\mu}_{1,s}(X)$ uses a subsample of size s from the sample of size n, there are $\binom{n}{s}$ subsample combinations. Out of these total subsample combinations, there are $\binom{n-1}{s-1}$ of $(y_{(1)}, \boldsymbol{x}_{(1)})$. Fan et al. (2018) name (14.2.16) the distributional nearest neighbors estimator (DNN) because the weight given to $y_{(1)}$ is different for different subsample size. Making use of the DNN estimator has an equivalent L-statistic representation (Serfling 1980),

$$D_n(s)(X) = \binom{n}{s}^{-1}\left\{\binom{n-1}{s-1}y_{(1)} + \binom{n-2}{s-2}y_{(2)} + \cdots + \binom{s-1}{s-1}y_{(n-s+1)}\right\}, \tag{14.2.17}$$

Fan et al. (2018) prove that the DNN estimator is asymptotically normally distributed. Wang and Huang (2019) demonstrate that the bootstrap method (Efron 1982) can be used to estimate the variance (Tu and Ping 1989). Furthermore, because the convergence rate depends on the sample size and the subsample size j_s, Wang and Huang (2019) suggest improving the efficiency of the simple average 1-nearest neighbor estimator (14.2.17) by taking the weighted average of subsample estimates with the weight depending on the size of j_s.

14.3 INFERENCE FOR LOW-DIMENSIONAL PARAMETERS IN THE PRESENCE OF HIGH-DIMENSIONAL DATA

Economists or econometricians are typically interested in the fundamental causal relations of a few variables of interest. However, factors affecting the outcomes of variables are numerous. A common method is to treat the aggregate of the impact of numerous omitted variables as random variables. For instance, suppose

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, n, \tag{14.3.1}$$

where y_i is the outcome variable of interest, $\boldsymbol{x}_i$ are the k-dimensional included explanatory variables, and v_i represents the impact of omitted variables. If $\boldsymbol{x}_i \perp v_i$, then regressing y_i on $\boldsymbol{x}_i$ can yield a consistent estimator of the unknown parameter $\boldsymbol{\beta}$ under very general conditions. However, if $\boldsymbol{x}_i$ is not orthogonal to v_i, regressing y_i on $\boldsymbol{x}_i$ will not yield a consistent estimator of $\boldsymbol{\beta}$.

The big data approach is to decompose

$$v_i = \eta_i^o + u_i, \tag{14.3.2}$$

where u_i is independent of $\boldsymbol{x}_i$, but η_i^o is correlated with $\boldsymbol{x}_i$. The infinite dimensional nuisance parameter η_i^o is assumed as

$$\eta_i^o = h(\boldsymbol{z}_i), \tag{14.3.3}$$

where $\boldsymbol{z}_i$ consists of infinite dimensional excluded variables that are independent of u_i. The big data often provide measurement of $\boldsymbol{z}_i$ whose dimension could exceed n. The basic approach is to use a machine learning method, such as LASSO (Tibshirani 1996)

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

or the Bayesian neural network (Mullachery et al. 2018) to select a subset of z_i, z_i^*, to approximate (14.3.3),

$$h(z_i) = \hat{h}(z_i^*) + \varepsilon_i, \tag{14.3.4}$$

where $h(z_i)$ or $\hat{h}(z_i^*)$ could be z_i or z_i^* itself or its transformation.

Substituting (14.3.4) into (14.3.1) yields

$$y_i - \hat{h}(z_i^*) = y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i + u_i. \tag{14.3.5}$$

Regressing y_i^* on $\boldsymbol{x}_i$, the scaled estimation error of $\hat{\boldsymbol{\beta}}$ is

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \to \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i u_i\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i\left(h(z_i) - \hat{h}(z_i^*)\right)\right]. \tag{14.3.6}$$

The first term of (14.3.6) is asymptotically normally distributed with mean $\mathbf{0}$ under fairly general conditions. However, the second term could be noncentered and even diverge in high-dimensional or highly complex settings because the rate of convergence of $\hat{h}(z_i^*)$ to $h(z_i)$ in the mean square error sense will typically be $(n)^{-b}$ with $b < \frac{1}{2}$. Hence, the second term could be of stochastic order $\sqrt{n}(n)^{-b} \to \infty$ as $\boldsymbol{x}_i$ are typically not centered at zero. Moreover, since $\boldsymbol{\beta}$ is unknown, it is not feasible to find $\hat{h}(z_i^*)$ with unknown v_i.

14.3.1 Sample Split Double/Debiased Estimator

To remove the bias of regressing y_i^* on $\boldsymbol{x}_i$, Belloni et al. (2014a, 2014b) and Chernozhukov, Chetverikov, et al. (2018) suggest a double/debiased estimator by first partialling out the impact of z_i on y_i and $\boldsymbol{x}_i$ through $y_i - E(y_i|z_i)$ and $\boldsymbol{x}_i - E(\boldsymbol{x}_i|z_i)$. Let $\tilde{y}_i = y_i - \hat{E}(y_i|z_i^*)$ and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \hat{E}(\boldsymbol{x}_i|z_i^*)$, where $\hat{E}(y_i|z_i^*)$ and $\hat{E}(\boldsymbol{x}_i|z_i^*)$ are the estimated $E(y_i|z_i)$ and $E(\boldsymbol{x}_i|z_i)$, respectively. Then regressing $\tilde{y}_i$ on $\tilde{\boldsymbol{x}}_i$ as in Robinson (1988a) and Newey (1994),

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'\right)^{-1}\left(\sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{y}_i\right). \tag{14.3.7}$$

The scaled estimation error of (14.3.7) is

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{N}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'\right)^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{N}\tilde{\boldsymbol{x}}_i u_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\hat{E}(\boldsymbol{x}_i|z_i^*) - E(\boldsymbol{x}_i|z_i)\right]\left[\hat{E}(y_i|z_i^*) - E(y_i|z_i)\right]\right\} + c. \tag{14.3.8}$$

The first term on the right-hand side of (14.3.8) is asymptotically normally distributed with mean $\mathbf{0}$ under fairly general conditions. The second term depends on the product of the estimation error $(\hat{E}(\boldsymbol{x}_i|z_i^*) - E(\boldsymbol{x}_i|z_i))$ and $(\hat{E}(y_i|z_i^*) - E(y_i|z_i))$. Suppose the estimation error for the former is $(n)^{-a}$ and the latter is $(n)^{-d}$; then the upper bound for the second term is $\sqrt{n}(n)^{-(a+d)}$, which can go to zero even when $\hat{E}(\boldsymbol{x}_i|z_i^*)$ and $\hat{E}(y_i|z_i^*)$ converge to $E(\boldsymbol{x}_i|z_i)$ and $E(y_i|z_i)$ at relatively slow rates. Therefore, as long as the term c in (14.3.8) is $o(1)$, the estimator (14.3.7) is well behaved.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

To find $\hat{E}(\boldsymbol{x}_i|z_i^*)$ and $\hat{E}(y_i|z_i^*)$, Belloni et al. (2014a, 2014b) suggest the following steps:

Step 1: Select a set of relatively small number of control variables that are considered useful to predict $\boldsymbol{x}_i$.

Step 2: Conditional on the selected control variables in step 1, select additional variables that help predict y_i.

Step 3: Combine the selected control variables in steps 1 and 2 as z_i^*. Then approximate $\hat{E}(\boldsymbol{x}_i|z_i^*)$ and $\hat{E}(y_i|z_i^*)$ by the usual nonparametric method, say the kernel method or series-based method.

The term c in (14.3.8) contains terms such as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\boldsymbol{x}_i - \hat{E}(\boldsymbol{x}_i|z_i^*)\big)\big(h(z_i) - \hat{h}(z_i^*)\big), \tag{14.3.9}$$

which involves $\frac{1}{\sqrt{n}}$ normalized sums of products of structural unobservables from model (14.3.1)–(14.3.2) with estimation errors in learning the nuisance functions $E(y_i|z_i)$ and $E(\boldsymbol{x}_i|z_i)$, which are generally related because the same data are used to form the estimators. If observations are independent, then a sample split can ensure $c = o(1)$. For instance, suppose the observations are randomly split into two groups, and $\hat{E}(\boldsymbol{x}_i|z_i^*)$ and $\hat{E}(y_i|z_i^*)$ are estimated using only observations in the first group. Then by substituting $\hat{E}(y_i|z_i^*)$ and $\hat{E}(\boldsymbol{x}_i|z_i^*)$ into the second group of observations, (14.3.9) will converge to zero; hence, the sample split double/debiased estimator (14.3.7) is asymptotically normally distributed with the mean equal to its true value.

While sample splitting could help to establish the term c to vanish in probability, its application can also lead to substantial loss of efficiency, as only a subset of available sample data is used. However, their efficiency can be regained by reversing the role of the first and second groups of data to obtain another sample split double/debiased estimator (14.3.7). Taking the average of the two subsample estimators can regain the efficiency loss due to sample splitting. As a matter of fact, Chernozhukov, Chetverikov, et al. (2018) suggest using an m-fold version of cross-fitting. Their empirical applications show that $m = 4$ or 5 works better than $m = 2$.

14.3.2 Orthogonal Projection Approach

Consider a model of the form (14.3.1)–(14.3.3),

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{v}, \tag{14.3.10}$$

$$\boldsymbol{v} = H(\boldsymbol{z}) + \boldsymbol{u}, \tag{14.3.11}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, $Z = (z_1, \ldots, z_n)'$, $\boldsymbol{v} = (v_1, \ldots, v_n)'$, $H(\boldsymbol{z}) = (h(z_1), \ldots, h(z_n))'$, $\boldsymbol{u} = (u_1, \ldots, u_n)'$ and $E(\boldsymbol{u}|X, Z) = \boldsymbol{0}$. If

$$h(z_i) = z_i'\boldsymbol{\gamma}, \tag{14.3.12}$$

and rank $(Z) \leq n - k$, one can eliminate the impact of $H(\boldsymbol{z})$ on $\boldsymbol{y}$ through the projection matrix $M_z = (I_n - Z(Z'Z)^{-}Z')$, where A^{-} denotes the generalized inverse of matrix A.

Multiplying M_z to (14.3.10) yields

$$\boldsymbol{y}^* = X^*\boldsymbol{\beta} + \boldsymbol{u}^*, \tag{14.3.13}$$

where $\boldsymbol{y}^* = M_z\boldsymbol{y}$, $X^* = M_zX$, and $\boldsymbol{u}^* = M_z\boldsymbol{u}$.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Under the assumption that $E(\boldsymbol{u}|X,Z)=0$, the estimator

$$\hat{\boldsymbol{\beta}}^{*}=(X^{*'}X^{*})^{-1}(X^{*'}\boldsymbol{y}^{*}) \tag{14.3.14}$$

is consistent under fairly general conditions on $X, Z,$ and $\boldsymbol{u}$ (Hsiao and Zhou 2021).

Many big data sets are considered random draws from a common population. If u_i is independently identically distributed with constant variance σ^2, the asymptotic covariance matrix of (14.3.14) can be approximated by

$$\hat{\sigma}^2\left(\frac{1}{n}X^{*'}X^{*}\right)^{-1}, \tag{14.3.15}$$

where $\hat{\sigma}^2=\frac{1}{n-k-p}\sum_{i=1}^{n}(y_i^{*}-\boldsymbol{x}_i^{*'}\hat{\boldsymbol{\beta}}^{*})^2$, k denotes the dimension of $\boldsymbol{x}_i$ and p denotes the dimension of linearly independent $\boldsymbol{z}_i$. When n is large, $\hat{\sigma}^2$ can be simply approximated by $\frac{1}{n}\sum_{i=1}^{n}(y_i^{*}-\boldsymbol{x}_i^{*'}\hat{\boldsymbol{\beta}}^{*})^2$. If u_i is heteroscedastic, $\mathrm{Var}(u_i)=\sigma_i^2\neq\mathrm{Var}(u_j)=\sigma_j^2$, following White (1980), the asymptotic covariance matrix can be approximated by

$$\left(\frac{1}{n}X^{*'}X^{*}\right)^{-1}\left(\frac{1}{n}X^{*'}\hat{V}^{*}X^{*}\right)\left(\frac{1}{n}X^{*'}X^{*}\right)^{-1}, \tag{14.3.16}$$

where $\hat{V}^{*}$ is an $n\times n$ diagonal matrix with diagonal elements equal to $(y_i^{*}-\boldsymbol{x}_i^{*'}\hat{\boldsymbol{\beta}}^{*})^2$.

When $h(z)$ is a nonlinear function, then $H^{*}(z)=M_zH(z)\neq\mathbf{0}$ and

$$\mathrm{plim}\frac{1}{n}X^{*'}H^{*}(z)=\mathrm{plim}\,\frac{1}{n}X'H(z)=\boldsymbol{a}, \tag{14.3.17}$$

while $\boldsymbol{a}$ could be different from zero. Then (14.3.14) is inconsistent. If $h(z_i)$ can be approximated by a sieve function such that (e.g. Chen 2007),

$$\left|h(z_i)-\sum_{j=1}^{k_n}\pi_j p_j(z_i):\pi_1,\ldots,\pi_{kn}\,\varepsilon\,R\right|\leq c^{k_n}, \tag{14.3.18}$$

where $|c|<1$, $\{p_j(z_i),\ j=1,2,\ldots\}$ is a sequence of known basis function such as power series, Fourier series, splines or waves, and k_n increases with n such that (14.3.18) approaches to zero when $n\to\infty$. Then orthogonal projection method can still be implemented:

Let $Z^{*}=(Z,H^{*}(z))$ where $H^{*}(z)=(h^{*}(z_1),\ldots,h^{*}(z_n))$ and $h^{*}(z_i)=\sum_{j=1}^{k_n}\pi_j p_j(z_i)$. Multiplying

$$M_z^{*}=(I_n-Z^{*}(Z^{*'}Z^{*})^{-}Z^{*'}). \tag{14.3.19}$$

to the model (14.3.10) yields

$$\tilde{\boldsymbol{y}}=\tilde{X}\boldsymbol{\beta}+\tilde{\boldsymbol{v}}, \tag{14.3.20}$$

where $\tilde{\boldsymbol{y}}=M_{z^{*}}^{*}\boldsymbol{y}$, $\tilde{X}=M_{z^{*}}^{*}X$, and $\tilde{\boldsymbol{v}}=M_{z^{*}}^{*}\boldsymbol{v}$. As long as $\left(\frac{1}{n}\tilde{X}'\tilde{X}\right)$ converges to a nonsingular matrix as $n\to\infty$, regressing $\tilde{\boldsymbol{y}}$ on $\tilde{X}$ yields a consistent and asymptotically normally distributed estimator of $\boldsymbol{\beta}$. (Hsiao and Zhou 2021).

14.3.3 Monte Carlo

Hsiao and Zhou (2021) conducted Monte Carlo simulation to examine the finite sample performance of the least squares estimator, the orthogonal projection estimator and the double/debiased estimator (Belloni et al. 2014a, 2014b; and Chernozhukov, Chetverikov, et al. 2018), which is implemented in the following steps as suggested by Belloni et al. (2014a, 2014b):

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Step 1: Select a subset of $\mathbf{z}$, $\mathbf{z}_1^*$, based on regressing $\mathbf{x}_i$ on $\mathbf{z}_i$ using the Lasso method.

Step 2: Select a subset of $\mathbf{z}$, $\mathbf{z}_2^*$, based on regressing y_i on $\mathbf{z}_i$ using the Lasso method.

Step 3: Regressing y_i on $\mathbf{x}_i$ and the union of $\mathbf{z}_{1i}^*$ and $\mathbf{z}_{2i}^*$ with an intercept.

Hsiao and Zhou (2021) let the dimension of $\mathbf{z}$, k_z, increases with sample size n at $k_z = \left[n^{1/2}\right]$ with $n = 200, 500, 1000, 2000$, where $[\cdot]$ denotes the integer part of $\cdot$. They consider the data generating processes of the following forms.

14.3.3.1 Model with Linearly Omitted Variables

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \sum_{j=1}^{p} \gamma_j z_{ji} + u_i, \quad i = 1, 2, \ldots, n, \tag{14.3.21}$$

where they assume $a = 0.5$, $b_1 = 1$, $b_2 = 2$, and $\gamma_j = cd^{j-1}$ for $j = 1, \ldots, p$ with $c = 0.3$ and $d = 0.7$, and generate of $\mathbf{w}_i = \left(x_{i1}, x_{i2}, z_{i1}, \ldots, z_{ip}\right)'$ as weakly cross-sectionally dependent

$$w_{pi} = \left(1 + b^2\right)\xi_{pi} + b\xi_{p+1,i} + b\xi_{p-1,i}, \tag{14.3.22}$$

where ξ_{pi} is a random draw from $0.5\left(\chi^2(1) + 1\right)$, and they let $b = 1$.

14.3.3.2 Model with Nonlinearly Omitted Variables

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + 0.1h(\mathbf{z}_i) + u_i, \quad i = 1, 2, \ldots, n, \tag{14.3.23}$$

with $h(z) = (z'\gamma)^2$ and $\gamma_j = 0.3\,(1/j)$ for $j = 1, \ldots, p$, and $\mathbf{w}_i = \left(x_{i1}, x_{i2}, z_{i1}, \ldots, z_{ip}\right)'$ are generated from a three-factor model.

For these two DGPs, they let the error term $u_i \sim IIDN\left(0, \sigma_i^2\right)$ and $\sigma_i^2 \sim IID$ $0.5\left(\chi^2(1) + 1\right)$ for $i = 1, \ldots, n$. They set $k_z = \left[n^{1/2}\right]$ with $n = 200, 500, 1000$, and 2000, and the number of replication is set at 1000 times.

For each estimator, Hsiao and Zhou (2021) compute the mean of bias, RMSE, and empirical size, for the nominal size of 5% which is calculated as empirical rejection frequency of t-statistics based on the estimated covariance matrix (14.3.16). Tables 14.1 and 14.2 reproduce their Tables 4 and 6. They find the following: (1) As expected, the OLS estimation ignoring the impact of z_j's will lead to biased estimation when the omitted variables and variables of interest are correlated. (2) The machine learning method using LASSO to select important covariates among z_j's in general works well. (3) The orthogonal projection method using a projection matrix consisting of all z_j's also works well whether $h(\mathbf{z})$ is linear or nonlinear in $\mathbf{z}$. For all simulations, the bias of the orthogonal projection method is quite small, and the empirical size is quite close to nominal value. In order words, the statistical inference can be conducted based on the orthogonal projection method.

14.3.4 Empirical Examples

14.3.4.1 Income Tax Deduction on Net Financial Assets

Hsiao and Zhou (2021) revisited the effect of 401(k) eligibility on net financial assets studied in Chernozhukov et al. (2017). They follow Chernozhukov et al. (2017) to use net

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Table 14.1. *Estimation results of b_1 and b_2 for (14.3.21)*

		b_1			b_2		
		OLS	Orth. Proj.	Lasso	OLS	Orth. Proj.	Lasso
$N = 200$	bias	−0.0012	−0.0019	−0.0047	−0.0041	0.0013	−0.0079
	RMSE	0.0671	0.0673	0.0670	0.0360	0.0378	0.0388
	size	4%	5.5%	3%	1.3%	6.3%	5.2%
$N = 500$	bias	0.0214	0.0002	−0.0025	0.0322	0.0009	−0.0023
	RMSE	0.0331	0.0262	0.0262	0.0533	0.0436	0.0430
	size	4.2%	5.7%	5%	7.7%	6.1%	5.6%
$N = 1000$	bias	−0.0240	−0.002	−0.0032	0.0202	0.0016	−0.0042
	RMSE	0.0322	0.0219	0.0226	0.0386	0.0330	0.0332
	size	7.5%	5.2%	5.6%	8.4%	5.9%	5%
$N = 2000$	bias	−0.0061	−0.0014	−0.0036	0.0047	−0.0003	−0.0034
	RMSE	0.0171	0.0162	0.0164	0.0161	0.0157	0.0160
	size	1.3%	4.4%	5%	0.9%	5.3%	5.4%

Notes: 1. Size is calculated as empirical frequency using 1.96 as critical value.
2. "Orth. Proj." and "Lasso" stand for the orthogonal projection estimation and the double/debiased estimation, respectively.
Source: Hsiao and Zhou (2021).

Table 14.2. *Estimation results of b_1 and b_2 for (14.3.23)*

		b_1			b_2		
N		OLS	OP_Series	Lasso	OLS	OP_Series	Lasso
200	bias	0.4211	0.0512	−0.0058	1.6764	−0.0949	−0.1545
	RMSE	0.4290	0.1036	0.0822	1.6791	0.1520	0.1860
	size	0%	11%	5%	100%	15%	26%
500	bias	2.4284	−0.0010	−0.0486	1.9615	−0.0551	−0.0842
	RMSE	2.4293	0.0748	0.0840	1.9624	0.0863	0.1055
	size	100%	0.8%	3.3%	100%	10%	15%
1000	bias	1.2671	−0.0208	−0.0606	1.6092	−0.0092	−0.0849
	RMSE	1.2679	0.0535	0.0759	1.6097	0.0481	0.0953
	size	100%	4.2%	15%	100%	4.2%	41%
2000	bias	1.0946	−0.0396	−0.0611	2.3238	0.0205	0.0033
	RMSE	1.0950	0.0504	0.0681	2.3240	0.0378	0.0308
	size	100%	15%	35%	100%	3.5%	0.5%
10000	bias	2.4204	−0.0101	−0.0307	2.2930	0.0114	0.0337
	RMSE	2.4203	0.0173	0.0337	2.2930	0.0183	0.0365
	size	100%	5.1%	38%	100%	6.5%	50%

Notes: "OP-Series" refers to orthogonal projection using $(\mathbf{e}_n, \mathbf{Z}, \sin(Z\pi), \ldots, \sin(5Z\pi))$
Source: Hsiao and Zhou (2022, Table 9)

financial assets as the outcome variable, y, which is defined as the sum of IRA balances, 401(k) balances, checking accounts, U.S. saving bonds, other interest-earning accounts in banks and other financial institutions, other interest-earning assets (such as bonds held personally), stocks, and mutual funds less non-mortgage debt. The variable of interest is the treatment variable, x, an indicator for being eligible to enroll in a 401(k) plan. The vector of raw covariates, Z, consists of age, income, family size, years of education, a married indicator, a two-earner status indicator, a defined benefit pension status indicator, an IRA participation indicator, and a home ownership indicator.

To analyze the effect of 401(k) eligibility on net financial assets, they consider OLS and orthogonal projection method. They also replicate the estimation results of LASSO, regression tree, and random forest using the two-folds procedure suggested by

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Table 14.3. *Estimation of β_x for the effects of 401(k) eligibility on net financial assets*

	β_x					
	OLS	Orth. Proj. Linear	Orth. Proj. Noninear	Lasso	Reg. Tree	Random Forest
Estimate	19559	5896	5896	7718	8745	9180
s.e.	(1245)	(1523)	(1523)	(1796)	(1488)	(1526)

Note: Standard errors are in parentheses.
Source: Hsiao and Zhou (2021).

Table 14.4. *Estimation of β_x for the effects of unemployment insurance bonus on unemployment duration*

	β_x					
	OLS	Orth. Proj. Linear	Orth. Proj. Noninear	Lasso	Reg. Tree	Random Forest
Estimate	−0.104	−0.090	−0.090	−0.081	−0.083	−0.076
s.e.	(0.044)	(0.043)	(0.044)	(0.036)	(0.037)	(0.037)

Note: Standard errors are in parentheses.
Source: Hsiao and Zhou (2021).

Chernozhukov et al. (2017). The estimation results are summarized in Table 14.3. Similar to the findings of Chernozhukov et al. (2017), the results are quite sensitive to the methods used. From Table 14.3, they note that OLS provides the largest estimate of the treatment effects, while the orthogonal projection method produces the lowest estimate for the effects of 401(k) eligibility on net financial assets. The sensitivity of the estimates could be an indication that the included covariates Z cannot fully take account of the correlation between the treatment variable and the errors in the model.

14.3.4.2 Unemployment Insurance and Unemployment Duration

Hsiao and Zhou (2021) also revisit the effect of unemployment insurance (UI) bonus on unemployment duration. Following Chernozhukov et al. (2017), they focus only on the most generous compensation scheme, treatment 4, and drop all individuals who received other treatments. In this treatment, the bonus amount is high and the qualification period is long compared to other treatments, and claimants are eligible to enroll in a workshop. The treatment variable, x, is an indicator variable for treatment 4, and the outcome variable, y, is the log of duration of unemployment for the UI claimants. The vector of covariates, Z, consists of age group dummies, gender, race, number of dependents, quarter of the experiment, location within the state, existence of recall expectations, and type of occupation. They also replicate the estimation results of LASSO, regression tree, and random forest using two folds from Chernozhukov et al. (2017). The estimation results are summarized in Table 14.4. Contrary to the study of the impact of 401(k), all machine learning/big data approaches as well as the orthogonal projection approach yield estimates of similar magnitude.

14.3.5 Asymptotic Covariance Matrix of Regression Coefficients When Data Are Not Necessarily Randomly Drawn

The assumption of big data being randomly drawn is not necessarily appropriate as argued by Abadie et al. (2020). For instance, suppose the proportion of sample not randomly drawn

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

is less than 1 as $n \to \infty$. For those units that are randomly drawn, $\boldsymbol{\beta}$ can be estimated by the least squares method. For those units that are nonrandomly drawn, say subject to truncation or censoring, $\boldsymbol{\beta}$ is estimated by the methods discussed in Chapter 7. Then the asymptotic covariance matrix of regression coefficients can be approximated by

$$(X'X)^{-1}X'VX(X'X)^{-1} \tag{14.3.24}$$

where V is an $n \times n$ diagonal matrix with diagonal elements equal to $(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2$, where $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.

If information on whether a sample is drawn randomly or nonrandomly is unknown, one could just estimate $\boldsymbol{\beta}$ assuming all observations are from a nonrandom sample since those procedures presumably would remain consistent, if samples are randomly drawn, even though they are not efficient. However, in the big data context, presumably consistency consideration should dominate efficiency consideration.

14.4 PREDICTION

14.4.1 Data-Based or Causal Predictive Models

Being able to predict the unknown outcome accurately is very important to decision makers. Often a decision could involve billions of dollars. For instance, Netflix has budgeted $1 billion for the then yet-to-be-produced *Avatar 2* animation series. A widespread belief exists that big data are soon able to predict our every move. Netflix has already applied big data forecasts for decision making prior to commencing production of its own TV shows. Meteorologists have used big data weather forecasts to obtain accurate predictions.

If the outcome of a variable h-periods ahead, $y_{i,t+h}$, is deterministic, then it is possible to obtain good prediction. If $y_{i,t+h}$ is stochastic, there is no way one can obtain perfect prediction. Prediction of $y_{i,t+h}$ has to rely on the information available at t, I^t; then

$$\begin{aligned} y_{i,t+h} &= E(y_{i,t+h}|I^t) + \varepsilon_{i,t+h} \\ &= g_{i,t+h}(\cdot) + \varepsilon_{i,t+h},\ h = 0, 1, \ldots \end{aligned} \tag{14.4.1}$$

where $i \geq 1$, and $\varepsilon_{i,t+h}$ represents the impact of unknown factors between t and $t + h$. If $y_{i,t+h}$ is deterministic, $\varepsilon_{i,t+h}$ can be pushed to zero. If $y_{i,t+h}$ is stochastic, $\varepsilon_{i,t+h}$ is in general not equal to zero. For instance, weather forecasts are still inaccurate beyond a week. Neither does the existence of vast amounts of data on earthquakes short of a reliable causal model help the precision of earthquake prediction.

The goal for both the data-based or causal approach is to construct a predictive model that is as close to $g_{i,t+h}(\cdot)$ as possible. The data-based approach is to use the AI algorithms to mine the high-dimensional and high-volume data to generate a predictive model. Let $f_{i,t+h}(\boldsymbol{x}_{i,t+h})$ be the selected predictive model for $g(\cdot)$. A data-based predictive model could be generated through the following steps:

Step 1: Variable selection.

AI algorithms to mine the data to find relevant predictors, say $\boldsymbol{x}_{it}$.

Step 2: Construct a simple predictive model (*Sparsity*).

To avoid "overfitting" or a "noise distorting signal" within a sample, construct a predictive model $f_{i,t+h}(\boldsymbol{x}_{it}^*; \hat{\boldsymbol{\theta}}^*)$ by minimizing a penalized objective function of the form,

$$\sum_{t=1}^{T} L(y_{i,t+h} - f_{i,t+h}(\boldsymbol{x}_{it}; \boldsymbol{\theta})) + \lambda h(\boldsymbol{\theta}), \tag{14.4.2}$$

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

where $L(\cdot)$ denotes the loss function to predict the outcome $y_{i,t+h}$, say $L(y_{i,t+h} - f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta})) = (y_{i,t+h} - f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta}))^2$, $h(\boldsymbol{\theta})$ is the penalty function, say the Euclidian norm of $\boldsymbol{\theta}$, $||\boldsymbol{\theta}||$, or the sum of absolute value of θ_k, $\sum_{k=1}^{K} |\theta_k|$, and λ is the tuning parameter set by an investigator (Tibshirani 1996). Denote the resulting model by $f_{i,t+h}(\boldsymbol{x}^*_{it};\hat{\boldsymbol{\theta}}^*)$, where $\boldsymbol{x}^*_{it}$ and $\hat{\boldsymbol{\theta}}^*$ are the selected subset of $\boldsymbol{x}_{it}$ and $\boldsymbol{\theta}$.

Step 3: Post-sample validation.

Divide the data into two (or k) subsets, using one subset of data to estimate the predictive model, $f_{i,t+h}(\boldsymbol{x}^*_{it};\hat{\boldsymbol{\theta}}^*)$. Substitute $f_{i,t+h}(\boldsymbol{x}^*_{it};\hat{\boldsymbol{\theta}}^*)$ into the second set of data to evaluate the adequacy of $f_{i,t+h}(\boldsymbol{x}^*_{it};\hat{\boldsymbol{\theta}}^*)$. If the model is considered unsatisfactory, repeat steps 1-3 until a satisfactory model is found.

Step 4: Model Averaging.

With big data, the possibility exists to expand the list of conditional covariates. Moreover, there could have several competing economic models, say $f^l_{i,t+h}(\boldsymbol{x}^l_{it};\hat{\boldsymbol{\theta}}^l)$, $l = 1, \ldots, m$. With the model uncertainty, model average predictions are often suggested (e.g., Bates and Granger 1969; Hsiao and Wan 2014; Hsiao and Zhou 2019; Elliott and Timmermann 2017). Kotchoni et al. (2019) suggest a *regularized data-rich averaging approach* that consists of predicting $y_{i,t+h}$ through the following steps:

Step 1: Randomly divide the data set into m roughly equal subsets.

Step 2: Use the AI algorithm to search for the relevant predictors, $\boldsymbol{x}^l_{it}$, from the lth data set. Construct $f^l_{i,t+h}(\boldsymbol{x}^{*l}_{it};\hat{\boldsymbol{\theta}}^l)$ that minimizes

$$\sum_{t=1}^{T} [y_{i,t+h} - f^l_{i,t+h}(\boldsymbol{x}^l_{it};\boldsymbol{\theta}^l)]^2 + \lambda h(\boldsymbol{\theta}^l), \tag{14.4.3}$$

Step 3: Predict $y_{i,t+h}$ by

$$\hat{y}_{i,t+h} = \frac{1}{m}\sum_{l=1}^{m} \hat{y}^l_{i,t+h}, \tag{14.4.4}$$

where

$$\hat{y}^l_{i,t+h} = f^l_{i,t+h}(\boldsymbol{x}^{*l}_{it};\hat{\boldsymbol{\theta}}^l).$$

The traditional econometric approach of constructing a predictive model can follow essentially the same steps. The only difference is in the identification of the subset of relevant predictors and whether to impose prior constraints derived from economic theories. For instance, economists or econometricians would select the subset of variables based on what they considered important causal models, such as the Klein–Goldberger (1955) macro-econometric models for the U.S., the dynamic stochastic general equilibrium model (e.g., Sbordone et al. 2010) or some stability conditions, etc.

The pros and cons of data based or causal predictions may be considered by decomposing the prediction errors of predicting $y_{i,t+h}$ from a predictive model $f_{i,t+h}(\boldsymbol{x}_{it};\hat{\boldsymbol{\theta}})$ into three components,

$$\begin{aligned} y_{i,t+h} - f_{i,t+h}(\boldsymbol{x}_{it};\hat{\boldsymbol{\theta}}) = {} & [g_{i,t+h}(\cdot) - f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta})] \\ & + [f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta}) - f_{i,t+h}(\boldsymbol{x}_{it};\hat{\boldsymbol{\theta}})] + \varepsilon_{i,t+h}. \end{aligned} \tag{14.4.5}$$

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

The first component of (14.4.5) is the specification error, the second component is due to the estimation error between the pseudo parameter $\boldsymbol{\theta}$ and its estimator $\hat{\boldsymbol{\theta}}$, and the third component is the innovation error. Thus, the mean square prediction error, $E[y_{i,t+h} - f_{i,t+h}(\boldsymbol{x}_{it};\hat{\boldsymbol{\theta}})]^2$, consists of the mean square error due to misspecification, $E[g_{i,t+h}(\cdot) - f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta})]^2$, the mean square error due to the sample estimation error $E[f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta}) - f_{i,t+h}(\boldsymbol{x}_{it};\hat{\boldsymbol{\theta}})]^2$, and the mean square error of the innovation $\varepsilon_{i,t+h}$, $E(\varepsilon_{i,t+h})^2$. The last component is beyond the control of any investigator. The second component depends on the complexity of the predictive model $f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta})$ and sample variability, as well as the degrees of freedom. There is a trade-off between the specification error and the estimation error. The data scientists and econometricians or statisticians have come up with various suggestions to balance the two, such as Akaike's (1973) information criterion (AIC), the Bayesian information criterion (BIC) (Schwarz 1978), and Bayesian model averaging (Steel 2011).

Prediction is different from identifying causal factors of an outcome. The mean square specification error $E[g_{i,t+h}(\cdot) - f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta})]^2$ depends on the closeness of $g_{i,t+h}(\cdot)$ and $f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta})$. Any variable that reduces the error $[g_{i,t+h}(\cdot) - f_{i,t+h}(\boldsymbol{x}_{it};\boldsymbol{\theta})]$ can be considered as a useful predictor, regardless of whether it is a causal factor of $y_{i,t+h}$. However, whether a variable is a useful predictor depends on the predictive horizon h. Take the crime rate in a region and police concentration ratio in the region as an example. If police assignments depend on the crime rate in the region, then police concentration ratio is a good predictor for the crime rate in the region in the short run, say $h = 1, 2, 3$, and AI algorithms are most likely to pick up the police concentration ratio in a region as a predictor for a region's crime rate. However, the police concentration ratio is not a causal factor for the crime rate in a region. To obtain good prediction for a multi-period-ahead forecast, one needs to identify the fundamental causal factors for a region's crime rate, say per-capita GDP, education level, etc. In other words, for a short-run prediction of a region's crime rate, the police concentration ratio could be a good predictor, but for the long-run prediction, including the police concentration ratio could be adding irrelevant noise to the predictive model.

Besides the difference in selecting the subset of variables, the data-based approach usually does not put any prior restrictions on the selected subset of variables, while the economic or econometric approach may impose prior restrictions based on economic theories. It is conceivable that combining the data-based approach with the causal approach can potentially yield a more accurate predictive model. As a matter of fact, Chen, Hsieh, and Lin (2020), CHL, have suggested a *matrix factorization and equilibrium collaborative filtering algorithm* that combines the traditional machine learning algorithm with the economic model of matching (Shapley and Shubik 1972) to predict the matching of men and women on an online dating platform and is able to demonstrate that it can yield better prediction than just relying on the data based approach or causal approach.

The CHL algorithm is based on postulating the net utility function for a male client i matching a female j, as

$$d^{1*}_{ij,t} = \boldsymbol{\beta}_1' \boldsymbol{z}_{jt} + \boldsymbol{\beta}_2' \boldsymbol{s}^1(\boldsymbol{x}_{it}, \boldsymbol{z}_{jt}) + \alpha_{1i} - \tau_{ij,t} + \varepsilon^1_{ij,t}, \tag{14.4.6}$$

and a female client j matching a male i, as

$$d^{2*}_{ji,t} = \boldsymbol{\gamma}_1' \boldsymbol{x}_{it} + \boldsymbol{\gamma}_2' \boldsymbol{s}^2(\boldsymbol{z}_{jt}, \boldsymbol{x}_{it}) + \alpha_{2j} - \tau_{ji,t} + \varepsilon^2_{ji,t}, \tag{14.4.7}$$

where $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{jt}$ denote the attributes of male i and female j, respectively, $\boldsymbol{s}^1(\cdot)$ and $\boldsymbol{s}^2(\cdot)$ are some distance measures between i and j, $\tau_{ij,t}$ and $\tau_{ji,t}$ are the matching costs, α_{1i} and α_{2j} are the individual-specific effects, and $\varepsilon^1_{ij,t}$ and $\varepsilon^2_{ji,t}$ are the random-error terms,

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

independent of $\boldsymbol{x}_{it}$, $\boldsymbol{z}_{jt}$,α_{1i}, α_{2j}, $\tau_{ij,t}$ and $\tau_{ji,t}$. Let the dummy variable $d^1_{ij,t} = 1$ if the ith male likes the jth female and 0 otherwise, and $d^2_{ji,t} = 1$ if the jth female likes the ith male and 0 otherwise. The observed $(d^1_{ij,t}, d^2_{ji,t})$ take the value,

$$\begin{aligned} d^1_{ij,t} &= 1, \quad \text{if} \quad d^{1*}_{ij,t} > 0, \\ &= 0, \quad \text{otherwise}, \end{aligned} \tag{14.4.8}$$

and

$$\begin{aligned} d^2_{ji,t} &= 1, \quad \text{if} \quad d^{2*}_{ji,t} > 0, \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{14.4.9}$$

Then

$$P(d^1_{ij,t} = 1) = \int_{-(\boldsymbol{\beta}_1' \boldsymbol{z}_{jt} + \boldsymbol{\beta}_2' \boldsymbol{s}^1(\boldsymbol{x}_{it}, \boldsymbol{z}_{jt}) + \alpha_{1i} - \tau_{ij,t})}^{\infty} f(\varepsilon^1_{ij,t})\, d\varepsilon^1_{ij,t} \tag{14.4.10}$$

$$P(d^2_{ji,t} = 1) = \int_{-(\boldsymbol{\gamma}_1' \boldsymbol{x}_{it} + \boldsymbol{\gamma}_2' \boldsymbol{s}^2(\boldsymbol{z}_{jt}, \boldsymbol{x}_{it}) + \alpha_{2j} - \tau_{ji,t})}^{\infty} f(\varepsilon^2_{ji,t})\, d\varepsilon^2_{ji,t}. \tag{14.4.11}$$

The unknown parameters of models (14.4.10) and (14.4.11) can be estimated by the methods discussed in Chapter 6.

Conditional on $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \alpha_{1i}, \alpha_{2j}, \tau_{ij,t}, \tau_{ji,t})$, CHL derive the predictive model of recommended list of candidates for a male client with features $\boldsymbol{x}_{it}$ to match female candidates $\boldsymbol{z}_{jt}$, or a female client with features $\boldsymbol{z}_{jt}$ to match a male client with features $\boldsymbol{x}_{it}$ by solving the equilibrium conditions of the transferable utility matching model proposed by Choo and Siow (2006),

$$\begin{aligned} &n_{\boldsymbol{x}_{it}} \text{Prob}(d^1_{ij,t} = 1 \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{jt}, \tau_{ij,t}, \alpha_{1i}) \\ &\quad = n_{\boldsymbol{z}_{jt}} \text{Prob}(d^2_{ji,t} = 1 \mid \boldsymbol{z}_{jt}, \boldsymbol{x}_{it}, \tau_{ji,t}, \alpha_{2j}), \end{aligned} \tag{14.4.12}$$

where $n_{x_{it}}$ and $n_{z_{jt}}$ are the total number of type $\boldsymbol{x}_{it}$ male and type $\boldsymbol{z}_{jt}$ female, respectively. Choo and Siow (2006) show that under the assumption $\tau_{ij,t} = \tau_{ji,t}$, the equilibrium matching satisfies the following system of equations:

$$\begin{aligned} &\mu_{xzt} = \mu^{0.5}_{xot}\, \mu^{0.5}_{ozt} \exp\left(\frac{U_{xzt} + V_{xzt}}{2}\right) \\ &\mu_{xot} + \sum_j \mu_{xzt} = n_{x_{it}} \quad \forall \boldsymbol{x}_{it} \\ &\mu_{ozt} + \sum_j \mu_{xzt} = n_{z_{jt}} \quad \forall \boldsymbol{z}_{jt}, \end{aligned} \tag{14.4.13}$$

where U_{xzt} and V_{xzt} denote the expected utility of a man with type x_{it} married to a woman with type $\boldsymbol{z}_{jt}$, and the expected utility of a woman with type $\boldsymbol{z}_{jt}$ married to a man with type $\boldsymbol{x}_{it}$, respectively, μ_{xzt} is the number of type $\boldsymbol{x}_{it}$ men who marry type $\boldsymbol{z}_{jt}$ women, and μ_{xot}, μ_{ozt} are the number of type $\boldsymbol{x}_{it}$ men and type $\boldsymbol{z}_{jt}$ women who remain single.

Based on (14.4.12) and (14.4.13), CHL suggested a *matrix factorization and equilibrium collaborative filtering algorithm* that can consist of the following steps:

Step 1: Use matrix factorization algorithm to identify user item attributes (*content filtering*) through past user actions (collaborating filtering), say, minimizing the prediction error of matching in terms of the product of male attributes $\boldsymbol{x}_{it}$ and female attributes $\boldsymbol{z}_{jt}$ by

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

$$\min \sum_{t=1}^{T} \sum_{i=1} \sum_{j=1} (y_{ij,t} - \boldsymbol{x}_{it}' \boldsymbol{z}_{jt})^2 + \frac{\lambda_1}{2} \|\boldsymbol{x}_{it}\|^2 + \frac{\lambda_2}{2} \|\boldsymbol{z}_{jt}\|^2, \tag{14.4.14}$$

where $y_{ij,t} = 1$ if ith men liked jth woman and $y_{ij,t} = -1$ otherwise[1], $\|\cdot\|$ denotes the Euclidian norm, λ_1 and λ_2 are the tuning parameters.

Step 2: Conditional on $\boldsymbol{x}_{it}, \boldsymbol{z}_{jt}$, and the price $\tau_{ij,t}$, $\tau_{ji,t}$, estimate the binary model parameters $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \alpha_{1i}, \alpha_{2j})$ using data $(d_{ij,t}, d_{ji,t})$. (CHL use a linear probability model).

Step 3: Solve the matching equilibrium as defined in (14.4.12) and (14.4.13) to recommend a list of candidates with features $\boldsymbol{z}_{jt}$ to a client with features $\boldsymbol{x}_{it}$ or a recommended list of candidates with features $\boldsymbol{x}_{it}$ to a client with feature $\boldsymbol{z}_{jt}$.

Step 3 is equivalent to imposing prior restrictions derived from economic theory to the prediction model. CHL compare the predictive performance of their proposed algorithm with the matrix factorization algorithm popular in data science (e.g., Netflix uses it), using the Taiwan online matching platform data which contain more than 490,000 men and women. They pick up the top 10 women in terms of marriage probability to recommend to the male client with feature $\boldsymbol{x}_{it}$ by taking a short-cut of their algorithm by:

1. In step 1 of the matrix factorization algorithm, instead of using observed $\boldsymbol{x}_{it}, \boldsymbol{z}_{jt}$, they treat them as "latent variables" to be estimated. This is actually a pure two-way factor model (interactive model discussed in Chapter 10). Hence, there is no step 2 in their short-cut algorithm.
2. Solving (14.4.13) to derive the recommended list of candidate, the short-cut algorithm assumes $U_{xzt} + V_{zxt} = \boldsymbol{x}_{it}' \boldsymbol{z}_{jt}$, where $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{jt}$ are the estimated latent (or factors and factor loading) in their simplified step 1.

They show that their recommended lists predict the like clicks substantially better than recommended lists relying on the popular matrix factorization algorithm alone.

14.4.2 Aggregate or Disaggregate Predictive Models

Most big data approaches are concerned with finding the predictive models for the micro units. On many occasions, policy makers are interested not in the micro outcomes, but in the average (or aggregate) outcomes. This raises a number of issues such as: (a) how to summarize the average information in the micro units; (b) when micro units are heterogeneous, how to classify the micro units into relatively homogeneous groups (in the sense that the differences between units in a group can be attributed to the chance mechanism); (c) if micro units are considered heterogeneous, whether one shall take a fixed-coefficients or random-coefficients approach to link the micro predictive models with the aggregate predictive model.[1]

14.4.2.1 Aggregation Methods

Consider aggregating y_{it} over N cross-sectional units by the method,

$$y_t = \sum_{i=1}^{N} w_i y_{it}, \tag{14.4.15}$$

[1] $y_{ij,t} = 0$ if the ith man did not click "like" on the jth woman at t.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

subject to

$$w_i \geq 0, \text{ and } \sum_{i=1}^{N} w_i = 1. \tag{14.4.16}$$

The conventional simple average aggregation method is to let $w_i = \frac{1}{N}$, then

$$\bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} y_{it}. \tag{14.4.17}$$

However, many aggregation methods satisfy (14.4.15) and (14.4.16). For instance, Hsiao, Shen and Zhou (2021) suggest to choose $w_i = w_i^{*2}$ or $w_i = w_i^{**2}$ where w_i^* or w_i^{**} are the corresponding element of the eigenvectors, $\boldsymbol{w}^* = (w_1^*, \ldots, w_N^*)'$ and $\boldsymbol{w}^{**} = (w_1^{**}, \ldots, w_N^{**})'$ that corresponds to the smallest and largest eigenvalue of the $N \times N$ matrix

$$\frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{y}_t - \bar{\boldsymbol{y}})(\boldsymbol{y}_t - \bar{\boldsymbol{y}})'. \tag{14.4.18}$$

Then

$$y_t^* = \sum_{i=1}^{N} w_i^{*2} y_{it}, \tag{14.4.19}$$

$$y_t^{**} = \sum_{i=1}^{N} w_i^{**2} y_{it}. \tag{14.4.20}$$

Suppose there are the N micro units, y_{it} is the sum of a long-run component μ_{it} and a transitory component v_{it} where $Ev_{it} = 0$; then

$$E\bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} \mu_{it}, \tag{14.4.21}$$

$$Ey_t^* = \sum_{i=1}^{N} w_i^{*2} \mu_{it}, \tag{14.4.22}$$

and

$$Ey_t^{**} = \sum_{i=1}^{N} w_i^{**2} \mu_{it}, \tag{14.4.23}$$

and

$$Var(y_t^*) \leq Var(\tilde{y}_t) \leq Var(y_t^{**}). \tag{14.4.24}$$

In other words, using w_i^{*2} as the weight gives more weight to those units that have a smaller variation of v_{it}. Using w_i^{**2} as the weight gives more weight to these units that have a larger variation of v_{it}. Using the simple average aggregation method gives equal weight to all cross-sectional units v_{it}. We would expect that using $w_i = w_i^{*2}$ tends to give a smooth long-term trend and using $w_i = w_i^{**2}$ tends to give a more volatile trend, while the simple

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

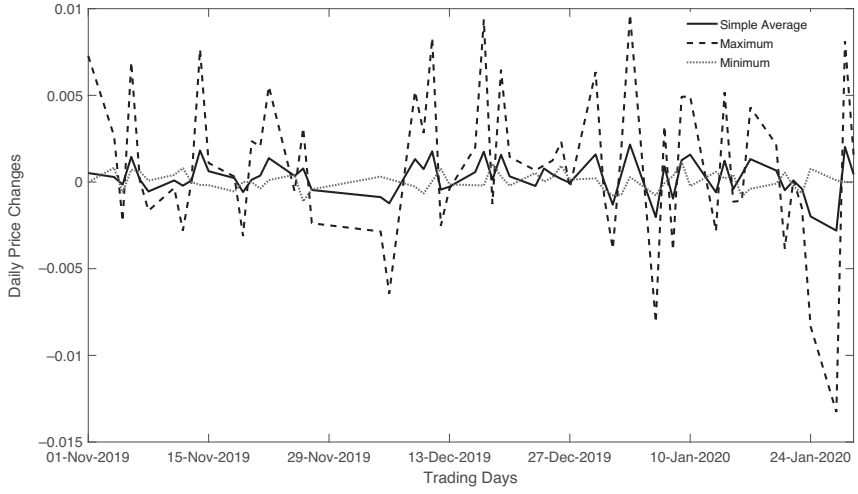

Figure 14.1. Comparison of aggregation of the stock price changes of top 15 US companies.
Source: Hsiao, Shen and Zhou (2021).

average aggregation method gives a long-term trend that is in between. Figure 14.1 plots the aggregate of the top 15 U.S. companies, stock price changes over time in terms of $\bar{y}_t$, y_t^*, and y_t^{**}.

If all micro units have identical $\mu_{it} = \mu_{jt} = \mu_t$, then all three aggregation methods have the same long-term trend. If μ_{it} are different across i, then $E\bar{y}_t \neq Ey_t^* \neq Ey_t^{**}$. Three different aggregation methods yield three different aggregate long-run relations. If the focus is on finding the long-run relationship between the aggregate variables, aggregation should be the one that uses the eigenvector corresponding to the smallest eigenvalue of (14.4.18). If the interest is on the volatility, then aggregation should be using the one that corresponds to the largest eigenvalue of (14.4.18).

To summarize, aggregation is a convenient way to summarize the information in the micro units. If micro units are homogeneous (in the sense of y_{it} conditional on some covariates), aggregation using the eigenvector corresponds to the smallest eigenvalue of (14.4.18), yielding more information about the underlying trend. If interest is in the volatility, then aggregation using the eigenvector corresponding to the largest eigenvalue of (14.4.18) should be used. If micro units are heterogeneous, then decision makers should specify the focus of their deliberation before one can consider what aggregation method to adopt.

14.4.2.2 Coherence and Reconciliation Approach to Partition Micro Units into Homogeneous Groups

When micro units are heterogeneous, one may partition micro units into G subgroups where members in each group are considered homogeneous in the sense that differences among them are due to the working of the chance mechanism. If units within the group are homegenous, there is no aggregation bias. The prediction based on the aggregate of micro units, say $\hat{y}_{g,t+h}$, has the expected value $E(\hat{y}_{g,t+h}|I_t)$, which is identical to $E(\hat{y}_{i,t+h}|I_t)$ for $i \in g$th group. Then we can consider partitioning micro units into homogeneous groups based on the idea of coherence and reconciliation in the forecasting literature (e.g., Hyndman et al. 2016; Wickramasuriya et al. 2019).

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

If all the predictive models for the aggregate or disaggregate data are unbiased predictors, coherence is a constraint that the forecasts of aggregates should be equal to the sum of the corresponding disaggregated forecasts. Therefore, suppose one wishes to partition all N micro units into G homogeneous groups. Let $\hat{\boldsymbol{y}}_{g,t+h}$ denote the h period-ahead forecast based on aggregating all the members in the gth group, and let $\hat{y}_{gi,t+h}$ denote the forecast of $y_{gi,t+h}$ of the ith member in the gth group based on the information of I_t. Let $\hat{\boldsymbol{y}}_{G,t+h} = (\hat{\boldsymbol{y}}_{1,t+h}, \ldots, \hat{\boldsymbol{y}}_{G,t+h})'$ denote $G \times 1$ aggregate predictions for the G groups and $\hat{\boldsymbol{y}}_{t+h} = (\hat{y}_{1,t+h}, \ldots, \hat{y}_{N,t+h})'$ denote the $N \times 1$ disaggregate predictions. Then coherence means

$$E(\hat{\boldsymbol{y}}_{G,t+h}|I_t) = S_G E(\hat{\boldsymbol{y}}_{t+h}|I_t), \tag{14.4.25}$$

where S_G is a $G \times N$ summing matrix $S = (s_{gi})$ that partitions N micro units into G mutually exclusive groups so that the sum of each column in S_G is equal to 1.[2] For instance, suppose $G = 2$; then S could be of the form

$$S_G = \begin{bmatrix} 1\,1\,0\,1\,0\cdots \\ 0\,0\,1\,0\,1\cdots \end{bmatrix}.$$

Reconciliation is the process of adjusting forecasts to make them coherent. The reconciliation can be considered to find S_G such that

$$\min \sum_{t=1}^{T} (\boldsymbol{y}_{G,t+h} - S_G \hat{\boldsymbol{y}}_{t+h})' (\boldsymbol{y}_{G,t+h} - S_G \hat{\boldsymbol{y}}_{t+h}), \tag{14.4.26}$$

subject to

$$s_{gi} \text{ either 1 or 0 and } \sum_{g=1}^{G} s_{gi} = 1 \text{ for } i = 1, \ldots, N, \tag{14.4.27}$$

where $\boldsymbol{y}_{G,t+h}$ is a $G \times 1$ vector of $(\boldsymbol{y}_{1,t+h}, \ldots, \boldsymbol{y}_{G,t+h})$, where $\boldsymbol{y}_{g,t+h}$ denotes the aggregate of $y_{i,t+h}$, for $i \in g$th group.

However, S_G is derived conditional on G. A priori, one does not know the number of relatively homogeneous groups in N cross-sectional units. One way to select the number of groups, G, could be to use Bayesian information criterion (Schwarz 1978) by choosing G to minimize

$$\frac{1}{T} \sum_{t=1}^{T} \left(\boldsymbol{y}_{G,t+h} - S_G \hat{\boldsymbol{y}}_{t+h}\right)' \left(\boldsymbol{y}_{G,t+h} - S_G \hat{\boldsymbol{y}}_{t+h}\right) + \frac{G \log T}{T}. \tag{14.4.28}$$

14.4.2.3 Fixed Coefficients versus Random Coefficients for Modeling Heterogeneous Forecasting Models

When micro units are heterogeneous, the fundamental relations underlying the micro series and macro series are different (e.g., Amemiya and Wu 1972; Lewbel 1992, 1994; Pesaran 2003; Stoker 1993; Theil 1954; Trivedi 1985). If the heterogeneous parameters of the micro units stay constant over time and T is large, in principle, one can estimate a micro behavioral relationship using individual time series data, then consider which aggregation methods yield the most useful summary information to decision makers. When

[2] Note that the definition of the "summing" matrix S_G is different from that of Wickramasuriya et al. (2019). So is the derivation of S_G.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

T is finite, it is not feasible to estimate a large number of micro-predictive models. Using the random-coefficients model estimation method discussed in Chapter 13 to link the aggregate predictive model and the micro predictive models could be a viable alternative.

For instance, consider a predictive model for the aggregate data that takes the form

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t, \tag{14.4.29}$$

where $y_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$, $\boldsymbol{x}_t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_{it}$. Let the micro model be denoted as

$$y_{it} = \boldsymbol{x}_t'\boldsymbol{\beta}_i + u_{it}, \quad i = 1, \ldots, N. \tag{14.4.30}$$

If the micro predictive model (14.4.30) with $\boldsymbol{\beta}_i$ fixed is treated as the basic model, then the aggregate predictive model is not (14.4.29) because of the aggregation bias, as discussed by Stoker (1993), among others. On the other hand, if the coefficients $\boldsymbol{\beta}_i$ for (14.4.30) are randomly distributed with mean $\boldsymbol{\beta}$ and constant variance–covariance matrix, aggregating (14.4.30) yields (14.4.29) (Zellner 1966), and then both $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}$ can be obtained from the approach discussed in Chapter 13.

When micro units are heterogeneous and time series dimension T is small, the feasibility of using the random-coefficients framework to take account of the heterogeneity may be considered from the perspective of following two examples:

(a) Predicting Monthly Demand for Kilowatt-Hour

In a study of Ontario, Canada, regional electricity demand, Hsiao et al. (1989) estimate a model of the form

$$y_{it} = \gamma_i y_{i,t-1} + \boldsymbol{\delta}_i'\boldsymbol{d}_{it} + \boldsymbol{\beta}_i'\boldsymbol{x}_{it} + u_{it}, \tag{14.4.31}$$

where y_{it} denotes the logarithm of monthly kilowatt-hour or kilowatt demand for region i at time t, $\boldsymbol{d}_{it}$ denotes 12 monthly dummies, and $\boldsymbol{x}_{it}$ denotes climatic factor and the logarithm of income, own price, and price of its close substitutes, all measured in real terms. Four different specifications are considered:

1. The coefficients $\boldsymbol{\theta}_i' = (\gamma_i, \boldsymbol{\delta}_i', \boldsymbol{\beta}_i')$ are fixed and different for different region.
2. The coefficients $\boldsymbol{\theta}_i' = \boldsymbol{\theta}' = (\gamma, \boldsymbol{\delta}', \boldsymbol{\beta}')$ for all i.
3. The coefficients vectors $\boldsymbol{\theta}_i$ are randomly distributed with common mean $\boldsymbol{\theta}$ and covariance matrix Δ.
4. The coefficients $\boldsymbol{\beta}_i$ are randomly distributed with common mean $\bar{\boldsymbol{\beta}}$ and covariance matrix Δ_{11}, and the coefficients γ_i and $\boldsymbol{\delta}_i$ are fixed and different for different i.

Monthly data for Hamilton, Kitchener–Waterloo, London, Ottawa, St. Catherines, Sudbury, Thunder Bay, Toronto, and Windsor from January 1967 to December 1982 were used to estimate these four different specifications. Comparisons of the one-period-ahead root mean square prediction error

$$\sqrt{\sum_{t=T+1}^{T+f} (y_{it} - \hat{y}_{it})^2 / f}$$

from January 1983 to December 1986 are summarized in Tables 14.5 and 14.6. As one can see from these tables, treating heterogeneity as fixed (model 1) dominates the simple pooling (model 2) and random-coefficients (model 3) formulations predictions for regional demand and average demand. The mixed fixed- and random-coefficients model (model 4) performs the best, in terms of predicting both regional and average demand. It appears that combining information across regions together with a proper account of regional-specific

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Table 14.5. *Root-mean-square prediction error of log kilowatt-hours (one-period-ahead forecast)*

	Root mean square error			
Municipality	Region-specific	Pooled	Random coefficients	Mixed
Hamilton	0.0865	0.0535	0.0825	0.0830
Kitchener–Waterloo	0.0406	0.0382	0.0409	0.0395
London	0.0466	0.0494	0.0467	0.0464
Ottawa	0.0697	0.0523	0.0669	0.0680
St. Catharines	0.0796	0.0724	0.0680	0.0802
Sudbury	0.0454	0.0857	0.0454	0.0460
Thunder Bay	0.0468	0.0615	0.0477	0.0473
Toronto	0.0362	0.0497	0.0631	0.0359
Windsor	0.0506	0.0650	0.0501	0.0438
Unweighted average	0.0558	0.0586	0.0568	0.0545
Weighted average[a]	0.0499	0.0525	0.0628	0.0487

[a] The weight is kilowatt-hours of demand in the municipality in June 1985.
Source: Hsiao et al. (1989, p. 584).

Table 14.6. *Root-mean-square prediction error of log kilowatts (one-period-ahead forecast)*

	Root mean square error			
Municipality	Region-specific	Pooled	Random coefficients	Mixed
Hamilton	0.0783	0.0474	0.0893	0.0768
Kitchener–Waterloo	0.0873	0.0440	0.0843	0.0803
London	0.0588	0.0747	0.0639	0.0586
Ottawa	0.0824	0.0648	0.0846	0.0768
St. Catharines	0.0531	0.0547	0.0511	0.0534
Sudbury	0.0607	0.0943	0.0608	0.0614
Thunder Bay	0.0524	0.0597	0.0521	0.0530
Toronto	0.0429	0.0628	0.0609	0.0421
Windsor	0.0550	0.0868	0.0595	0.0543
Unweighted average	0.0634	0.0655	0.0674	0.0619
Weighted average[a]	0.0558	0.0623	0.0673	0.0540

[a] The weight is kilowatt-hours of demand in the municipality in June 1985.
Source: Hsiao et al. (1989, p. 584).

factors is capable of yielding better predictions for regional and average demand than the approach of simply using regional-specific data (model 1) or pooling (model 2).

(b) Demand for Money

Hsiao, Shen, and Fujiki (2005) considered the estimation of Japan's prefecture nominal money demand equation of the form

$$y_{it} = \gamma_i y_{i,t-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta}_i + \alpha_i + u_{it}, \quad |\gamma_i| < 1, \quad i = 1, \ldots, N, \tag{14.4.32}$$

where the error u_{it} is covariance stationary over time. Then assuming the aggregate relation takes the form

$$y_t = \gamma y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + v_t \tag{14.4.33}$$

where $y_t = \frac{1}{N}\sum_{t=1}^{T} y_{it}$, $\boldsymbol{x}_t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_{it}$ is a misspecified relation as shown by Amemiya and Wu (1973), Thiel (1954), etc. However, if $(\gamma, \boldsymbol{\beta}')'$ is considered as the average

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Table 14.7. *Least squares estimation of aggregate money demand function*

Dependent variable	Sample period	Variable	Parameter estimate	Standard error
M2	1980.IV–2000.IV	Intercept	1.30462	0.28975
		Real GDP	−0.15425	0.04538
		RM2(−1)	1.07022	0.02790
		Bond rate	−0.00186	0.00069
	1992.IV–2000.IV	Intercept	−0.16272	0.85081
		Real GDP	0.00847	0.06772
		RM2(−1)	1.00295	0.02248
		Bond rate	−0.00250	0.00140
M1	1980.IV–2000.IV	Intercept	0.46907	0.21852
		Real GDP	−0.01857	0.01700
		RM2(−1)	0.98964	0.01249
		Bond rate	−0.00566	0.00135
	1992.IV–2000.IV	Intercept	−0.68783	2.10228
		Real GDP	0.08414	0.14898
		RM2(−1)	0.96038	0.01999
		Bond rate	−0.01005	0.00283

Source: Hsiao, Shen, and Fujiki (2005, Table 5).

Table 14.8. *Random-coefficient estimates of Japan prefecture money demand equation*

	M1		M2	
	Coefficient	Standard error	Coefficient	Standard error
Lagged money	0.656	0.034	0.0533	0.069
Income	0.0881	0.0114	0.0473	0.064
Bond rate	−0.0476	0.006	−0.009	0.003
Constant	−2.125	0.038	0.043	0.0239
	Variance–covariance matrix of $M1(\gamma_i, \beta_i')$			
	0.015			
	−0.001	0.0177		
	0.001	−0.059	0.0005	
	−0.024	−0.0588	−0.023	2.017
	Variance–covariance matrix of $M2(\gamma_i, \beta_i')$			
	0.068			
	−0.0831	0.062		
	0.002	0.0003	0.0014	
	−0.13	−0.107	−0.009	0.8385

Source: Hsiao, Shen, and Fujiki (2005, Table 1).

relationship between y_{it} and $(y_{i,t-1}, \boldsymbol{x}_{it})$, where γ_i and $\boldsymbol{\beta}_i$ are randomly distributed around $(\gamma, \boldsymbol{\beta}')$, then $(\gamma, \boldsymbol{\beta}')$ can be meaningfully estimated.

Table 14.7 provides Hsiao, Shen, and Fujiki's (2005) estimates of the aggregate relations between (real) money demand, (real) GDP and (five-year) bond rate. They are unstable and sensitive to the time period covered. Depending on the sample period covered, the estimated relations are either of wrong sign or statistically insignificant. The estimated long-run income elasticities are 75.23 for M1 and 11.04 for M2, respectively, an "incredible" magnitude.

Table 14.8 provides their random-coefficients model estimates of the mean relation between (real) money demand and (real) GDP and (five-year) bond rate for the 40 Japanese prefectures. The estimated short-run income elasticity for M1 and M2 is 0.88 and 0.47,

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

Table 14.9. *Error sum of squares (ESS) and predicted error sum of squares (PES) for disaggregate and aggregate data*

	M1		M2	
	Aggregate data	Disaggregate data	Aggregate data	Disaggregate data
EES	3.78×10^9	1.35×10^6	3.59×10^{43}	7.45×10^{42}
PES	2.51×10^{10}	5.75×10^7	9.55×10^{45}	2.04×10^{43}

Source: Hsiao, Shen, and Fujiki (2005, Table VIII).

respectively. The long-run income elasticity is 2.56 for M1 and 1.01 for M2. These results appear to be consistent with economic theory and the broadly observed facts about Japan. The average growth rate for M2 in the 1980s is about 9.34%. The inflation rate is 1.98%. The real M2 growth rate is 7.36%. The real growth rate of GDP during this period is 4.13%. Taking into account the impact of the five-year bond rate falling from 9.332% at 1980.I to 5.767 at 1989.IV, the results are indeed very close to the estimated long-run income elasticities based on disaggregate data analysis.[3]

If "heterogeneity" is indeed present in micro units, then shall we predict the aggregate outcome based on the summation of estimated micro relations, or shall we predict the aggregate outcomes based on the estimated aggregate relations? Unfortunately, there is not much work on this specific issue. In choosing between whether to predict aggregate variables using aggregate (H_a) or disaggregate equations (H_d), Grunfeld and Griliches (1960) suggest using the criterion of:

$$\text{Choose } H_d \text{ if } \boldsymbol{e}_d'\boldsymbol{e}_d < \boldsymbol{e}_a'\boldsymbol{e}_a, \text{ otherwise choose } H_a$$

where $\boldsymbol{e}_d$ and $\boldsymbol{e}_a$ are the estimates of the errors in predicting aggregate outcomes under H_d and H_a, respectively. Hsiao, Shen, and Fujiki (2005) provide a simulation comparison based on artificially generated time series data for each prefecture based on the observed stylized facts. Table 14.9 presents the within-sample fit comparisons in the first row and the post-sample prediction comparison in the second row. Based on random-coefficient estimation of $(\gamma_i, \boldsymbol{\beta}')$ and $(\gamma, \boldsymbol{\beta}')$, both criteria unambiguously favor predicting aggregate outcomes by summing the outcomes from the disaggregate equations in this particular example.

The above two examples appear to indicate that random-coefficients or mixed fixed- and random-coefficients models could be viable alternatives to take account heterogeneity across cross-sectional units when T is small.

14.4.3 Combining Data of Different Sources and/or Different Time Frequencies

Big data can take a variety of forms and can be nonstructured. For instance, it is argued that "sentiment" can help predict the outcome (e.g., Baker et al. 2016; Ranco et al. 2015). However, there are no formal rules to convert qualitative information to numerical numbers convenient for statistical analysis. Neither is there much discussion on the compatability of data from different sources with regard to whether to combine or not to combine data from different sources. If data from different sources all contain some information about some parameters $\boldsymbol{\delta}$, then the compatability of combining different sources could be considered

[3] The reason that the random-coefficients approach works here could be because although the coefficients are different for different prefactures, they also satisfy the de Finetti (1964) exchangeability assumption.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

from the traditional likelihood principle. For instance, suppose there are two data sets, I and II; then a likelihood ratio statistic of the form can be constructed,

$$L = L^{I+II}(\hat{\boldsymbol{\delta}}) - L^{I}(\hat{\boldsymbol{\delta}}^{I}) - L^{II}(\hat{\boldsymbol{\delta}}^{II}), \tag{14.4.34}$$

where $L^{I}(\hat{\boldsymbol{\delta}}^{I})$, $L^{II}(\hat{\boldsymbol{\delta}}^{II})$, and $L^{I+II}(\hat{\boldsymbol{\delta}})$, denote the maximum log-likelihood value evaluated at the MLE of $\boldsymbol{\delta}$ from data set I, data set II and the combined data set. Under the null that both sets of data contain information about the common parameter $\boldsymbol{\delta}$, if sample size is large, $-2L$ is asymptotically chi-square distributed with the degree of freedom equal to the dimension of $\boldsymbol{\delta}$. The question is, what is the appropriate level of significance? If the costs of mistakenly accepting the pooling hypothesis and rejecting the pooling hypothesis are the same, Maddala (1971b) suggested using something like a 25%–30% level of significance, rather than the conventional 5%, in our preliminary test of significance.

The specifications of the maximum-likelihood estimates and their variance–covariances merely summarize the likelihood function in terms of the location of its maximum and its curvature around the maximum. It is possible that the information contained in the likelihood function is not fully expressed by these. When the compatibility of different data sources is investigated, it is useful to plot the likelihood function extensively. For this purpose, Maddala (1971b) suggested that one should also tabulate and plot the relative maximum likelihoods of each data set,

$$R_M(\boldsymbol{\delta}) = \frac{\underset{\boldsymbol{\theta}}{\text{Max}}\ L(\boldsymbol{\delta},\boldsymbol{\theta})}{\underset{\boldsymbol{\delta},\boldsymbol{\theta}}{\text{Max}}\ L(\boldsymbol{\delta},\boldsymbol{\theta})} \tag{14.4.35}$$

for different values of $\boldsymbol{\delta}$, where $\boldsymbol{\theta}$ represents the set of nuisance parameters, $\max_{\theta} L(\boldsymbol{\delta},\boldsymbol{\theta})$ denotes the maximum of L with respect to $\boldsymbol{\theta}$, given $\boldsymbol{\delta}$ and $\max_{\delta,\theta} L(\boldsymbol{\delta},\boldsymbol{\theta})$ denotes the maximum of L with respect to both $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$. The plot of (14.4.35) summarizes almost all the information contained in the data set on $\boldsymbol{\delta}$. Hence, the shapes and locations of the relative maximum likelihoods will reveal more information about the compatibility of the different bodies of data than a single test statistic can. Maddala (1971b) used a simple econometric model relating to the demand for food in the United States, considered by Tobin (1950) to illustrate the basic idea.

The demand equation from the cross-sectional data is specified as

$$y_{1i} = \delta_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + u_i, \qquad i = 1, \ldots, N, \tag{14.4.36}$$

where y_{1i} is the logarithm of the average food consumption of the group of families at a point in time, and z_{1i} and z_{2i} are the logarithms of the average income of the ith family and the ith family size, respectively. The time series demand function is

$$y_{2t} = \beta_0 + \beta_1(x_{1t} - \beta_2 x_{2t}) + \beta_3(x_{2t} - x_{2,t-1}) + v_t, \qquad t = 1, \ldots, T. \tag{14.4.37}$$

where y_{2t}, x_{1t}, and x_{2t} are the logarithms of the food price index, per-capita food supply for domestic consumption, and per-capita disposable income, respectively. The income elasticity of demand, δ, was assumed common to both regressions, namely, $\gamma_1 = \beta_2 = \delta$. The error term u_i and v_t were independent of each other and were assumed independently normally distributed, with zero means and constant variances σ_u^2 and σ_v^2, respectively.

The results of the cross-sectional estimates are

$$\hat{y}_{1i} = \underset{}{0.569} + \underset{(0.0297)}{0.5611 z_{1i}} + \underset{(0.0367)}{0.2540 z_{2i}}, \tag{14.4.38}$$

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

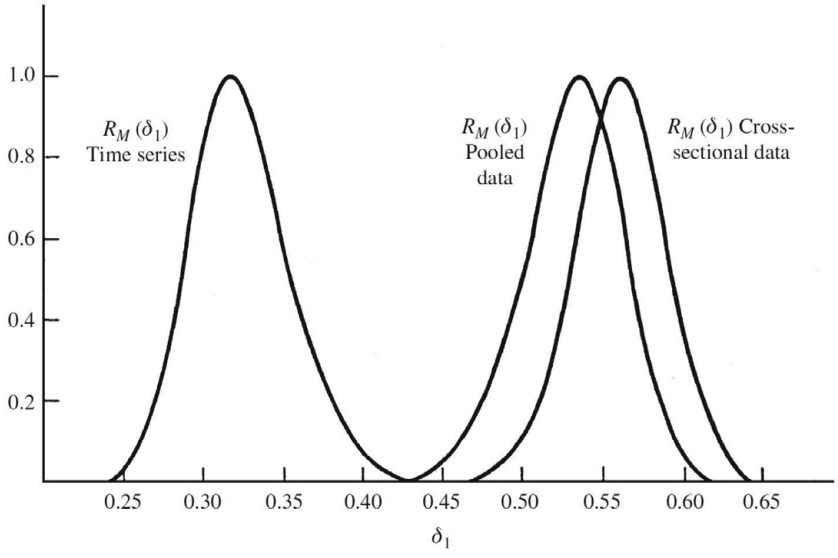

Figure 14.2. Relative maximum likelihood for the parameters δ_1.
Source: Maddala (1971b, Fig. 1).

where standard errors are in parentheses. The results of the time series regression are

$$\hat{y}_{2t} = 7.231 + \underset{(0.0612)}{1.144x_{2t}} - \underset{(0.0906)}{0.1519}(x_{2t} - x_{2,t-1}) - \underset{(0.4010)}{3.644x_{1t}}. \tag{14.4.39}$$

The implied income elasticity, β_2, is 0.314.

When the cross-sectional estimate of γ_1, 0.56, is introduced into the time-series regression, the estimated β_1 is reduced to -1.863, with a standard error of 0.1358. When γ_1 and β_2 are estimated simultaneously by the maximum-likelihood method, the estimated γ_1 and β_1 are 0.5355 and -1.964, with a covariance $\begin{bmatrix} 0.00206 & 0.00827 \\ & 0.04245 \end{bmatrix}$.

Although there is substantial improvement in the accuracy of the estimated coefficient using the combined data, the likelihood-ratio statistic turns out to be 17.2, which is significant at the 0.001 level with one degree of freedom. It strongly suggests that in this case we should not pool the time series and cross-sectional data.

Figure 14.2 reproduces Maddala's plot of the relative maximum likelihood $R_M(\delta_1)$ for the parameter δ_1 (the income elasticity of demand) in the Tobin model from cross-sectional data alone, from time series data alone, and from the pooled sample. The figure reveals that the information on δ_1 provided by the time-series data is almost as precise as that provided by the cross-sectional data (otherwise, the likelihood function would be relatively flat). Furthermore, there is very little overlap between the likelihood functions from time series and cross-sectional data. Again, this unambiguously suggests that the data should not be pooled.[4]

[4] The results are based on both u_i and v_t are independently normally distributed. In practice, careful diagnostic checks should be performed before exploring the pooling issue. As a matter of fact, Izan (1980) redid the analysis by assuming v_t to follow a first-order autoregressive process. The resulting likelihood ratio test statics accepted the null of pooling.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

In addition to the issue of poolability, there is also an issue of how to best combine data recorded at different time frequencies. Both literature on interpolation or extrapolation are considered in discrete interval data framework (e.g. Chow and Lin 1976; Hsiao 1979c). However, the speed of collecting data raises the complicated issues of continuous time modeling (e.g., Cai et al. 2018; Chang, Hu, and Park 2018; Li, Robinson, and Shang 2019; Phillips 1974; Robinson 1976) versus the modeling in terms of discrete time interval data considered in this book and how to combine discrete time interval data and continuous time data.

14.4.4 Structural Changes

The previous discussions are based on the assumption that parameters stay constant over time. However, there exist possibilities of structural changes due to changes in policies (e.g. Lucas 1976), or technologies or external conditions that can lead to changes in decision rules. For example, the *Financial Times* reported on September 10, 2020, that foreign exchange became the new playground for investors looking to profit from sharp moves in prices after central banks' aggressive response to COVID-19 robbed them of opportunities in bond markets. The foreign exchange volatility index was well above levels seen 12 months before. If the break point is known and there are a large number of post-break sample observations, presumably one can use post-break data to construct predictive models. If the break point is unknown or is near the end of sample period, Wang et al. (2013) suggested ignoring the volatilities of structural break and just using all of the observed sample to construct an appropriate time series model for forecasting volatilities. Pesaran and Pick (2011) demonstrated that (simple) averaging forecasts over different estimation windows in general yields a lower bias and root mean square prediction error than forecasts based on a single estimation window. On the other hand, Sun et al. (2020), SHLWZ, note that there could be multiple candidate predictive models, and the availability of big data allows an investigator to construct predictive models with parameters changing smoothly over time rather than changing abruptly at a given break point. Therefore, they suggest taking varying weights on a local time-varying average method to take account of possible structural breaks over time.

Suppose there are M candidate predictive models for predicting y_{t+h}, and all the predictive models are time-varying parameters models. For simplicity, we assume each predictive model takes the form,

$$y_t^{(j)} = \boldsymbol{x}_t^{(j)'}\boldsymbol{\beta}_t^{(j)} + \varepsilon_t^{(j)}, \quad t = 1, \ldots, T, \tag{14.4.40}$$

where $x_t^{(j)}$ denotes the subset of the countably infinite number of covariates in the information set I^t. SHLWZ assume $\boldsymbol{\beta}_t^{(j)}$ is a smooth function of the ratio t/T as in Cai (2007), Chen and Hong (2012), Robinson (1989), etc., such that

$$\boldsymbol{\beta}_s^{(j)} \approx \boldsymbol{\beta}_t^{(j)}, \quad s \in (t - Th, t + Th), \tag{14.4.41}$$

then $\boldsymbol{\beta}_t^{(j)}$ can be estimated by

$$\hat{\boldsymbol{\beta}}_t^{(j)} = (X^{(j)'} K_t X^{(j)})^{-1} (X^{(j)'} K_t \boldsymbol{y}) \tag{14.4.42}$$

where $\boldsymbol{y} = (y_1, \ldots, y_T)'$, $X^{(j)} = (\boldsymbol{x}_t^{(j)'})$, and $K_t = \text{diag}\{k_{1t}, \ldots, k_{Tt}\}$ is a smooth kernel with $k_{st} = k(\frac{s-t}{Th})$ being a prespecified symmetric probability density function,

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015

and h is a bandwith parameter such that $h \to 0$ and $Th \to \infty$ as $T \to \infty$. Then y_t is predicted by

$$\hat{\mu}_t^{(j)} = \boldsymbol{x}_t^{(j)'} \hat{\boldsymbol{\beta}}_t^{(j)}. \tag{14.4.43}$$

To further reduce the bias of $\hat{\boldsymbol{\beta}}_t^{(j)}$ and $\hat{\mu}_t^{(j)}$, SHLWZ suggest using a jackknife estimator in lieu of (14.4.42), which is defined as K_{-t}=diag$\{k_{1,t}, k_{2,t}, \ldots, k_{(t-1),t}, 0, k_{t+1,t} \ldots, k_{T,t}\}$,

$$\tilde{\boldsymbol{\beta}}_t^{(j)} = (X^{(j)'} K_{-t} X^{(j)})^{-1} (X^{(j)'} K_{-t} \boldsymbol{y}) \tag{14.4.44}$$

and

$$\tilde{\mu}_t^{(j)} = \boldsymbol{x}_t^{(j)'} \tilde{\boldsymbol{\beta}}_t^{(j)}. \tag{14.4.45}$$

To take into account that there are M candidate predictive models, SHLWZ further suggest a varying weighting scheme after obtaining the best local time averaging of the jth predictive model by minimizing

$$\sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{M} w_j^M \tilde{\mu}_{jt}^{(j)} \right)^2 \tag{14.4.46}$$

subject to

$$\sum_{j=1}^{M} (w_j^M)^2 = 1. \tag{14.4.47}$$

where $\tilde{\mu}_{jt}^{(j)}$ denote the jth predictive model based on local jackknife averaging. They show by allowing the weights to change smoothly over time, the average squared error of their time-varying average method is asymptotically equivalent to the local averaged squared error of the infeasible best possible average error of these M candidate models.[5]

[5] The potential covariates for each candidate predictive model could be large. To further improve the prediction accuracy of the time-varying model average method, Sun, Hong, Wang, and Zhang (2020) have suggested a parsimonious time-varying forward-validating model averaging that selects model averaging weights and a subset of regressors from each predictive model's chosen covariates, $x_t^{(j)}$, simultaneously.

Downloaded from https://www.cambridge.org/core. Pennsylvania State University, on 31 May 2025 at 18:40:34, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009057745.015