Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 694-716
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Revisiting Akaike's Final Prediction Error and the Generalized Cross Validation Criteria in Regression from the Same Perspective: From Least Squares to Ridge Regression and Smoothing Splines

Jean Raphael Ndzinga Mvondo*, Eugène-Patrice Ndong Nguéma

Department of Mathematics, Ecole Polytechnique, Yaoundé I, Cameroon
Email: *rndzinga@yahoo.fr

How to cite this paper: Ndzinga Mvondo, J.R. and Ndong Nguéma, E.-P. (2023) Revisiting Akaike's Final Prediction Error and the Generalized Cross Validation Criteria in Regression from the Same Perspective: From Least Squares to Ridge Regression and Smoothing Splines. *Open Journal of Statistics*, **13**, 694-716.
https://doi.org/10.4236/ojs.2023.135033

Received: August 2, 2023
Accepted: September 25, 2023
Published: September 28, 2023

Abstract

In regression, despite being both aimed at estimating the Mean Squared Prediction Error (MSPE), Akaike's Final Prediction Error (FPE) and the Generalized Cross Validation (GCV) selection criteria are usually derived from two quite different perspectives. Here, settling on the most commonly accepted definition of the MSPE as the expectation of the squared prediction error loss, we provide theoretical expressions for it, valid for any linear model (LM) fitter, be it under random or non random designs. Specializing these MSPE expressions for each of them, we are able to derive closed formulas of the MSPE for some of the most popular LM fitters: Ordinary Least Squares (OLS), with or without a full column rank design matrix; Ordinary and Generalized Ridge regression, the latter embedding smoothing splines fitting. For each of these LM fitters, we then deduce a computable estimate of the MSPE which turns out to coincide with Akaike's FPE. Using a slight variation, we similarly get a class of MSPE estimates coinciding with the classical GCV formula for those same LM fitters.

Keywords

Linear Model, Mean Squared Prediction Error, Final Prediction Error, Generalized Cross Validation, Least Squares, Ridge Regression

1. Introduction

In many branches of activity, the data analyst is confronted with the need to model

a continuous numeric variable Y (the *response*) in terms of one or more explanatory other variables (called *covariates* or *predictors* or *regressors*) $X_1,\cdots,X_p$ in a population Ω through a model

$$Y=\Phi(X)+\varepsilon, \tag{1}$$

where[1]: $X=(X_1,\cdots,X_p)^{\mathrm{T}}$; $\Phi(\cdot)$ is a function from $\mathbb{R}^p\to\mathbb{R}$, generally unknown, called the *regression function*; ε is an unobserved *error term*, also called *residual error* in the model (1).

In our developments in this paper, and contrary to popular tradition, we will not need that the variables X and ε necessarily be stochastically independent. However, the usual minimal assumption for (1) is that the variables X and ε satisfy:

$$(\mathscr{A}_0).\quad \mathbb{E}(\varepsilon\,|\,X)=0.$$

Though sometimes far more debatable, we will also admit that the postulated regression model (1) satisfies the *homoscedasticity* assumption for the residual error variance:

$$(\mathscr{A}_1).\quad \mathbb{V}\mathrm{ar}(\varepsilon\,|X)=\sigma^2>0.$$

Now, as is well known, Assumption $(\mathscr{A}_0)$ implies that

$$\Phi(X)=\mathbb{E}(Y\,|\,X),\quad \text{i.e.}\quad \forall \boldsymbol{x}\in\mathbb{R}^p,\quad \Phi(\boldsymbol{x})=\mathbb{E}(Y\,|X=\boldsymbol{x}). \tag{2}$$

However, that conditional expectation function can almost never be analytically computed in practical situations. The aim of the regression analysis of $Y\,|\,X$ (*i.e.* "Y given X") is rather to estimate the unknown regression function $\Phi(\cdot)$ in (1) based on some observed data

$$(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n)\in\mathbb{R}^p\times\mathbb{R},\quad \text{i.i.d.}\quad \overset{\mathscr{D}}{\sim}(X,Y), \tag{3}$$

collected on a sample of size n drawn from Ω. If achieved, this will result in a computable function $\hat{\Phi}:\mathbb{R}^p\to\mathbb{R}$ so that the final practical model used to express the response variable Y in terms of the vector of predictors X will be:

$$Y=\hat{\Phi}(X)+\tilde{\varepsilon}, \tag{4}$$

where $\tilde{\varepsilon}$ is the residual error in the modeling. But once we get such a fit for the regression model (1), an obvious question then arises: how can one measure the accuracy of that computed fit (its so called *goodness-of-fit*)?

For some specific regression models (generally parametric ones and, most notably, the LM fitted by OLS with a full column rank design matrix), various measures of accuracy of their fit to given data, with closed form formulas, have been developed. But such specific and easily computable measures of accuracy are not universally applicable to all regression models. So they cannot be used to compare two arbitrarily fitted such models. Opposite to that, in the late 1960s, Akaike decisively introduced an approach (but in a much broader context in-

[1]By default, in this paper, vectors and random vectors are column vectors, unless specified otherwise; and A^{T} denotes the transpose of the matrix (or vector) A.

cluding time series modeling) which has that desirable universal feature [1], and is thus now generally recommended to compare different regression fits for the same data, albeit, sometimes, at some computational cost. It is based on estimating the *prediction error* of the fit. But various definitions and measures of that error have been used. By far, the most popular is the *Mean Squared Prediction Error* (MSPE), but there appears to be a myriad ways of defining it and/or theoretical estimates of it. A detailed lexicon on the matter is given in [2] and the references therein.

These various definitions and theoretical estimates of the MSPE are, undoubtedly, insightful in their respective motivations and aims, but they ultimately make the subject of prediction error assessment utterly complicated and confusing for most non expert users. This is compounded by the fact that many of these definitions and discussions of prediction error liberally mix theoretical estimates (*i.e.* finite averages over sample items) with intrinsic numeric characteristics of the whole population (expressed through *expectations* of some random variable). In that respect, while being both aimed at estimating the MSPE, the respective classical derivations of Akaike's Final Prediction Error (FPE) and Craven and Wahba's Generalized Cross Validation (GCV) stem from two quite different perspectives [1] [3]. This makes it not easy, for the non expert user, to grasp that these two selection criteria might even be related in the first place.

The first purpose of this paper is to settle on the definition of the MSPE most commonly known by users to assess the prediction power of any fitted model, be it for regression or else. Then, in that framework, we shall provide a conceptually simpler derivation of the FPE as an estimate of that MSPE in a LM fitted by OLS when the design matrix is full column rank. Secondly, we build on that to derive generalizations of the FPE for the LM fitted by other well known methods under various scenarios, generalizations seldom accessible from the traditional derivation of the FPE. Finally, we show that, in that same unified framework, a minor variation in the derivation of the MSPE estimates yield the well known formula of the GCV for all these various LM fitters. For the latter selection criterion, previous attempts [4] [5] have been made to provide a derivation of it different from the classical one as an approximation of the leave-one-out cross validation (LOO-CV) estimate of the MSPE. We view our approach as a more straightforward and inclusive derivation of the GCV score.

To achieve that, we start, in Section 2, by reviewing the prediction error viewpoint in assessing how a fitted regression model performs and to settle on the definition of the MSPE most commonly known by users to measure that performance (while briefly recalling the alternative one most given in regression textbooks). Then, in Section 3, focusing specifically on the LM, we provide theoretical expressions of that MSPE measure valid for any arbitrary LM fitting method, be it under random or non random designs. In the next sections, these expressions are successively specialized to some of the best known LM fitters to deduce, for each, computable estimates of the MSPE, a class of which yielding

the FPE and a slight variation the GCV. Those LM fitters include: OLS, with or without a full column rank design matrix (Section 4), Ridge Regression, both Ordinary and Generalized (Section 5), the latter embedding smoothing splines fitting. Finally, Section 6 draws a conclusion and suggests some needed future work.

As customary, we summarize the data (3) through the design matrix and the response vector:

$$\boldsymbol{X} = \left(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\right)^{\mathrm{T}} \in \mathcal{M}_{n,p}\left(\mathbb{R}\right) \quad \text{and} \quad \boldsymbol{y} = \left(y_1, \cdots, y_n\right)^{\mathrm{T}} \in \mathbb{R}^n. \tag{5}$$

It is important to emphasize that technically, each observed response y_i is a realization of a random variable. Consequently, the same will be true of the vector $\boldsymbol{y}$. On the other hand, the matrix $\boldsymbol{X}$ is considered as fixed or random, depending on whether the covariates $X_1, \cdots, X_p$ are viewed as fixed or random variables. According to which one holds, one talks of *fixed design* or *random design*. Our presentation will encompass both scenarios. For convenience, we will use the same symbol to denote a random variable and its realization. Nonetheless, the distinction between the latter two will be apparent from context.

2. Prediction Error of a Fitted Regression Model

2.1. The Prediction Error Viewpoint for Assessing a Fitted Regression Model

Using the data (3), assume we got an estimate $\hat{\Phi}(\cdot)$ of the unknown regression function $\Phi(\cdot)$ in the regression model (1). Thus, we fitted the model (1) to express the response Y in terms of the vector of covariates $X = \left(X_1, \cdots, X_p\right)$ in the population under study Ω. From a predictive perspective, assessing the goodness-of-fit of that fitted model amounts to answering the question: *How well the fitted model is likely to predict the response variable Y on a future individual based on its covariates values* $X = \left(X_1, \cdots, X_p\right)$?

To try to formalize an answer, consider a new member of Ω, drawn independently from the sample of Ω which produced the data (3), and assume we have observed its covariates vector $X = X_0 \in \mathbb{R}^p$, but not its response value $Y = Y_0 \in \mathbb{R}$. Nonetheless,served value of the latter. This is the *prediction problem of* $Y \mid X$ ("Y given X") on that individual. With model (4) fitted to the response variable Y, it seems natural to predict the unknown value of $Y = Y_0$ on that individual, given from the former, we want to get an idea of the unob $X = X_0$, by:

$$\hat{Y}_0 = \hat{\Phi}\left(X_0\right). \tag{6}$$

But there is, obviously, an error attached to such a prediction of $Y = Y_0$ value by $\hat{Y}_0$, called the *prediction error* or *predictive risk* of $\hat{Y}_0$ *w.r.t.* Y_0.

2.2. The Mean Squared Prediction Error

The prediction error of $\hat{Y}_0$ *w.r.t.* Y_0 needs to be assessed beforehand to get an idea of the quality of the fit provided by the estimated regression model (4) for

$Y \mid X$ in the population under study. To that aim, the global measure mostly used is the Mean Squared Prediction Error (MSPE), but one meets a bewildering number of ways of defining it in the literature [2], creating a bit of confusion in the mindset of the daily user of Statistics. Yet, most users would accept as definition of the MSPE:

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) = \mathbb{E}\left(\hat{Y}_0 - Y_0\right)^2, \tag{7}$$

an intrinsic characteristic of the population which we need to estimate from the available data (3).

In this paper, an unconditional expectation like the *r.h.s.* of (7) is meant to integrate over all the random variables involved in the integrand. A conditional expectation, on the other hand, acts the same, except for the conditioning random variable(s). As such, *en route* to getting $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$, we will pass successively through:

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}, \boldsymbol{y}, X_0\right) = \mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}, \boldsymbol{y}, X_0\right], \tag{8a}$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}, \boldsymbol{y}\right) = \mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}, \boldsymbol{y}\right], \tag{8b}$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) = \mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}\right], \tag{8c}$$

representing, each, the MSPE conditioned on some information, which might be relevant in its own right. In particular, $\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ is the relevant definition of the MSPE in the case of a fixed design. But, as a consequence of moving successively through (8a)-(8c) to get (7), handling the fixed design case will not need a special treatment because computing (8c) will be a necessary intermediate step.

Obviously, for most fitted regression models, trying to compute $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ or $\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ is a hopeless task. However, it is known that *K*-fold cross validation (CV) can be used to estimate these quantities in a nearly universal manner, imposing no distributional assumption and using a quasi automatic algorithm [6]. Nonetheless, cross validation has its own defects such as a high extra computational cost[2], the impact of the choice of *K* and, generally, upwards bias. As for the latter defect, Borra and Di Ciaccio [2] showed in an extensive simulation study that Repeated Corrected CV, a little popularized correction to *K*-fold cross validation developed by Burman [7], performed pretty well and outperformed, on some well known nonlinear regression fitters and under a variety of scenarios, all the other theoretically more technically involved suggested measures of the MSPE alluded to above.

Fortunately, after fitting a Linear Model to given data by a chosen method,

[2]However, that defect is more and more mitigated these days, thanks to the availability of increasingly user-friendly parallel programming environments in popular statistical software systems, provided one has a computer allowing to exploit such possibilities, *e.g.* a laptop with several *core processors*.

there is, at least for the most common methods, no need to shoulder the computational cost attached to CV to estimate the MSPE of the fitted model. It is the purpose of this article to show that for the most well known LM fitters, one can transform (7)-(8c) in a form allowing to deduce, in a straightforward manner, two estimates of the MSPE computable from the data, which turn out to coincide, respectively, with the FPE and the GCV. This, therefore, will yield a new way to get these two selection criteria in linear regression completely different from how they are usually respectively motivated and derived.

2.3. Prediction Error and Sources of Randomness in the Regression Process

The rationale behind settling on (7) or (8c) as definition of the MSPE lies in the fact that, from our standpoint, any global measure of the prediction error in the population computed from a sample collected in it must account both for the randomness in drawing that sample and in that of drawing a future individual. Consequently, each of the expectations in (7)-(8c) should be understood as integrating over all the possible sources of randomness entering in the process of computing the prediction $\hat{Y}_0$ of Y_0, safe the conditioning variables, if any:

1) the realized, but unobserved random n-vector $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^{\mathrm{T}}$ of sample residual errors in the model (1) for the data (3). This vector results from the fact that for the observed sample of n individuals, the model (1) implies that

$$y_i = \Phi(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \cdots, n, \quad \text{with } \varepsilon_1, \cdots, \varepsilon_n \text{ i.i.d.} \overset{\mathscr{D}}{\sim} \varepsilon; \tag{9}$$

2) $X_0 \in \mathbb{R}^p$, the vector of covariates for the potential newly independently drawn individual for which the unknown value $Y_0 \in \mathbb{R}$ of the response Y is to be predicted;

3) ε_0, the error of the model (1) for that new individual, $Y_0 = \Phi(X_0) + \varepsilon_0$, $\varepsilon_0 \overset{\mathscr{D}}{\sim} \varepsilon$.

4) and, in case of a random design, the entries in the design matrix $\boldsymbol{X}$.

The key assumption to assess the prediction error of a regression model is then:

($\mathscr{A}_2$). *The random couple* $(\varepsilon_0, X_0) \overset{\mathscr{D}}{\sim} (\varepsilon, X)$ *and is independent from* $(\varepsilon, \boldsymbol{X})$.

2.4. Measuring the Prediction Error in Regression: *Sample Based Definition*

While the GCV score is classically derived, indeed, as an estimate of the MSPE as given by (7), through a two-stage process where the intermediate step is the well known LOO-CV estimate of that MSPE, the traditional derivation of the FPE criterion stems from a completely different angle. Actually, the latter angle is the one presented in most textbooks on regression [8] [9]. In it, with the regression function Φ in (1) estimated through $\hat{\Phi}$, a function computed from the data (3), the prediction error of that fit is rather measured by how well the vector of predicted responses on sample items, $\hat{\boldsymbol{y}} = (\hat{y}_1, \cdots, \hat{y}_n)^{\mathrm{T}}$, with $\hat{y}_i = \hat{\Phi}(\boldsymbol{x}_i)$, $i = 1, \cdots, n$,

estimates the vector of exact responses on those items, $\blacksquare = (\Phi_1, \cdots, \Phi_n)^{\mathrm{T}}$, with $\Phi_i = \Phi(\boldsymbol{x}_i)$, $i = 1, \cdots, n$. In that respect, one considers the *mean average squared error*, also called *risk*,

$$\mathrm{MASE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[(\hat{y}_i - \Phi_i)^2 \mid \boldsymbol{X}\right], \tag{10}$$

with bias-variance decomposition

$$\mathrm{MASE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(\hat{y}_i \mid \boldsymbol{X}) + \frac{1}{n}\sum_{i=1}^{n} b_i^2. \tag{11}$$

where $b_i = \mathbb{E}(\hat{y}_i \mid \boldsymbol{X}) - \Phi_i$ is the conditional bias, given $\boldsymbol{X}$, in estimating Φ_i by $\hat{y}_i$.

But the relation to the more obvious definition (8c) of the conditional MSPE is better seen when one considers, instead, the *average predicted squared error* ([8] Chapter 3, page 42] or *prediction risk* ([9] Chapter 2, page 29):

$$\mathrm{PSE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(y_i^* - \hat{y}_i\right)^2 \mid \boldsymbol{X}\right], \tag{12}$$

where $y_1^*, \cdots, y_n^*$ are putative responses assumed generated at the respective predictor values $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ through model (1), but with respective errors $\varepsilon_1^*, \cdots, \varepsilon_n^*$ independent from the initial ones $\varepsilon_1, \cdots, \varepsilon_n$. Nonetheless, there is a simple relation between (10) and (12):

$$\mathrm{PSE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X}) = \sigma^2 + \mathrm{MASE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X}). \tag{13}$$

hence minimizing $\mathrm{PSE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X})$ *w.r.t.* $\hat{\Phi}$ is the same as doing so for $\mathrm{MASE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X})$.

In its classical derivation for linear regression (see, *e.g.*, [10], pages 19-20), the FPE selection criterion is an estimate of $\mathrm{PSE}(\hat{\boldsymbol{y}} \mid \boldsymbol{X})$. Now, with the terminology elaborated in [5], measuring the prediction error by the latter amounts to using a *Fixed-X* viewpoint as opposed to the *Random-X* one when measuring it instead through $\mathrm{MSPE}(\hat{Y}_0, Y_0)$. But an even more common terminology to distinguish these two approaches to estimating the predictive performance of a regression method qualifies the first as *in-sample* prediction and the second one as *out-of-sample* prediction. It should be said that while in the past, the prediction error was mostly evaluated using the in-sample paradigm, facing the complexity of data met in modern statistics, noticeably high dimensional data, many researchers in regression have advocated or used the out-of-sample viewpoint, though this might be through either (7), (8b), or (8c), depending on the author(s). In that respect, in addition to the aforementioned paper of Trosset and Tibshirani, we may cite, *e.g.*, Breiman and Spector [11], Leeb [12], Dicker [13], Dobriban and Wager [14].

Note, however, that the prediction error viewpoint in assessing the quality of a regression fit is not without its own demerits. Indeed, in [15] and [16], it is highlighted that in the specific case of smoothing splines regression, one can find a fit which is optimal from the prediction error viewpoint, but which clearly un-

dersmoothes the data, resulting in a wiggly curve. But the argument appears to be more a matter of visual esthetics because the analysis in those papers targets the regression function Φ, which is, indeed, probably the main objective of many users of univariate nonparametric regression. Nonetheless, when measuring the prediction error through $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$, the target is rather the response Y, *i.e.* Φ + *error*, in which the wiggliness is inherently embedded. Do not forget that when formulating his Final Prediction Error, Akaike was working on real world engineering problems [17]. Hence his interest in targeting the actual output in his predictions.

3. MSPE of a Linear Model Fitter

From now on, we focus attention on the generic LM, with $\beta = \left(\beta_1, \cdots, \beta_p\right)^{\mathrm{T}}$:

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon = X^{\mathrm{T}}\beta + \varepsilon, \tag{14}$$

to be fitted to the data (3). Because of its ease of tractability and manipulation, the LM is the most popular approach to estimating the regression function $\Phi(\cdot)$ in (1). It is mostly implemented by estimating β through the Ordinary Least Squares criterion. However, several other approaches have been designed to estimate β based on various grounds, such as Weighted Least Squares, Total Least Squares, Least Absolute Deviations, LASSO [18] and Ridge Regression [19]. Furthermore, some generally more adaptive nonparametric regression methods proceed by first nonlinearly transforming the data to a scale where they can be fitted by a LM. Due to its more than ever quite central role in statistical modeling and practice, several books have been and continue to be fully devoted to the presentation of the LM and its many facets such as: [20] [21] [22] and [23].

3.1. Some Preliminaries

For the sample of n individuals with recorded data (3), the general regression Equation (9) becomes, in the case of the LM (14):

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\beta + \varepsilon_i, \quad i = 1, \cdots, n, \tag{15}$$

or, better globally summarized in matrix form,

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon, \quad \text{with} \quad \varepsilon = \left(\varepsilon_1, \cdots, \varepsilon_n\right)^{\mathrm{T}}. \tag{16}$$

Then Assumptions ($\mathscr{A}_0$) and ($\mathscr{A}_1$) respectively imply here:

$$\mathbb{E}\left(\varepsilon \mid \boldsymbol{X}\right) = \boldsymbol{0} \quad \text{and} \quad \mathbb{M}\mathrm{cov}\left(\varepsilon \mid \boldsymbol{X}\right) = \sigma^2 \boldsymbol{I}_n, \tag{17}$$

with $\boldsymbol{I}_n$ the n-by-n identity matrix. Consequently,

$$\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{X}\beta \quad \text{and} \quad \mathbb{M}\mathrm{cov}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \sigma^2 \boldsymbol{I}_n. \tag{18}$$

Fitting the LM (14) boils down to estimating the p-vector of parameters β. We call *Linear Model fitter*, or *LM fitter*, any method allowing to achieve that. First, we consider an arbitrarily chosen such method. It uses the data (3) to

compute $\hat{\beta} \in \mathbb{R}^p$, an estimate of β. The precision of that estimate[3] can be assessed through its *Mean Squared Error matrix*:

$$\mathbb{M}\text{se}\left(\hat{\beta}\right) = \mathbb{E}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}}\right]. \tag{19a}$$

But to reach $\mathbb{M}\text{se}\left(\hat{\beta}\right)$ generally requires passing through the conditional Mean Squared Error matrix of $\hat{\beta}$ given the design matrix $\boldsymbol{X}$:

$$\text{Mse}\left(\hat{\beta} \mid \boldsymbol{X}\right) = \mathbb{E}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \mid \boldsymbol{X}\right]. \tag{19b}$$

The relationship between the two is:

$$\text{Mse}\left(\hat{\beta}\right) = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{M}\text{se}\left(\hat{\beta} \mid \boldsymbol{X}\right)\right]. \tag{19c}$$

Those two matrices will play a key role in our MSPE derivations to come.

The precision of an estimate of a vector parameter like β is easier to assess when its Mean Squared Error matrix coincides with its covariance matrix. Hence, our interest in:

Definition 1 *In the LM* (14), $\hat{\beta}$ *is an unbiased estimate of* β, *conditionally on* $\boldsymbol{X}$, *if* $\mathbb{E}\left(\hat{\beta} \mid \boldsymbol{X}\right) = \beta$.

Then one has: $\mathbb{M}\text{se}\left(\hat{\beta} \mid \boldsymbol{X}\right) = \mathbb{M}\text{cov}\left(\hat{\beta} \mid \boldsymbol{X}\right)$, the conditional covariance matrix of $\hat{\beta}$ given $\boldsymbol{X}$.

Since $\mathbb{E}\left(\hat{\beta}\right) = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}\left(\hat{\beta} \mid \boldsymbol{X}\right)\right]$, it is immediate that if $\hat{\beta}$ is an unbiased estimate of β, conditionally on $\boldsymbol{X}$, then $\mathbb{E}\left(\hat{\beta}\right) = \beta$, *i.e.* $\hat{\beta}$ is an unbiased estimate of β (unconditionally). So the former property is stronger than the latter, but is more useful in this context. Note also that it implies: $\mathbb{M}\text{se}\left(\hat{\beta}\right) = \mathbb{M}\text{cov}\left(\hat{\beta}\right)$, the *covariance* matrix of $\hat{\beta}$. On the other hand, when $\hat{\beta}$ is a biased estimate of β, the bias-variance decomposition of $\mathbb{M}\text{se}\left(\hat{\beta}\right)$ might be of interest:

$$\mathbb{M}\text{se}\left(\hat{\beta}\right) = \mathbb{B}\text{ias}\left(\hat{\beta}\right)\mathbb{B}\text{ias}\left(\hat{\beta}\right)^{\mathrm{T}} + \mathbb{M}\text{cov}\left(\hat{\beta}\right), \tag{20}$$

where $\mathbb{B}\text{ias}\left(\hat{\beta}\right) = \mathbb{E}\left(\hat{\beta}\right) - \beta$.

3.2. MSPE of a Linear Model Fitter: *Base Results*

In the prediction setting of Section 2.1 applied to the LM (14), one has, for the new individual:

$$Y_0 = X_0^{\mathrm{T}}\beta + \varepsilon_0, \tag{21}$$

with $\varepsilon_0 \overset{\mathscr{D}}{\sim} \varepsilon$, but unknown. It would then be natural to predict the response value Y_0 by

$$\tilde{Y}_0 = X_0^{\mathrm{T}}\beta = \Phi\left(X_0\right), \tag{22}$$

were the exact value of the parameter vector β available. Since that is not typ-

[3]Keeping in line with our stated convention of denoting a random variable and its realization by the same symbol, whereas, technically, an *estimate* of a parameter is a realization of a random variable called *estimator* of that parameter, we use the term *estimate* here for both. So when an estimate appears inside an expectation or a covariance notation, it is definitely the estimator which is meant.

ically the case, one rather predicts Y_0 by the computable quantity

$$\hat{Y}_0 = X_0^{\mathrm{T}}\hat{\beta} = \hat{\Phi}(X_0). \tag{23}$$

The goal here is to find expressions, in this context, for the MSPE of $\hat{Y}_0$ *w.r.t.* Y_0 as given by (7)-(8c), manageable enough to allow the derivation of computable estimates of that MSPE for the most common LM fitters. The starting point to get such MSPE expressions is the base result:

Theorem 1 *For any* $\hat{\beta}$ *estimating* β *in the LM* (14), *one has*, *under* ($\mathscr{A}_0$), ($\mathscr{A}_1$), ($\mathscr{A}_2$):

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}, \boldsymbol{y}, X_0\right) = \sigma^2 + \mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot X_0 X_0^{\mathrm{T}}\right], \tag{24a}$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}, \boldsymbol{y}\right) = \sigma^2 + \mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right], \tag{24b}$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta} \mid \boldsymbol{X}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right], \tag{24c}$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]. \tag{24d}$$

Proof. From $\hat{Y}_0 - Y_0 = X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right) - \varepsilon_0$, we first get:

$$\left(\hat{Y}_0 - Y_0\right)^2 = \varepsilon_0^2 - 2\varepsilon_0 X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right) + \left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2. \tag{25}$$

Now, from (8a) and using Assumption ($\mathscr{A}_2$),

$$\begin{aligned}
&\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}, \boldsymbol{y}, X_0\right) \\
&= \mathbb{E}\left(\varepsilon_0^2 \mid X_0\right) - 2\mathbb{E}\left(\varepsilon_0 \mid X_0\right) X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right) + \left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2 \\
&= \mathbb{V}\mathrm{ar}\left(\varepsilon_0 \mid X_0\right) + \left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2 = \sigma^2 + \left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2,
\end{aligned} \tag{26}$$

the latter because $\mathbb{E}\left(\varepsilon_0 \mid X_0\right) = \mathbb{E}\left(\varepsilon \mid X\right) = 0$, and so $\mathbb{E}\left(\varepsilon_0^2 \mid X_0\right) = \mathbb{V}\mathrm{ar}\left(\varepsilon_0 \mid X_0\right) = \mathbb{V}\mathrm{ar}\left(\varepsilon \mid X\right) = \sigma^2$. On the other hand, $\left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2 = \mathrm{tr}\left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2$ because $\left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2$ is a scalar. Thus

$$\left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\right]^2 = \mathrm{tr}\left[X_0^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} X_0\right] = \mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} X_0 X_0^{\mathrm{T}}\right],$$

which, inserted in (26), yields (24a).

Thanks to (24a) and identity (8b), we have

$$\begin{aligned}
\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}, \boldsymbol{y}\right) &= \mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}, \boldsymbol{y}\right] \\
&= \mathbb{E}_{X_0}\left\{\mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}, \boldsymbol{y}, X_0\right]\right\}, \quad \text{since } X_0 \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{y}) \\
&= \mathbb{E}_{X_0}\left\{\sigma^2 + \mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot X_0 X_0^{\mathrm{T}}\right]\right\} \\
&= \mathbb{E}_{X_0}\left(\sigma^2\right) + \mathbb{E}_{X_0}\left\{\mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot X_0 X_0^{\mathrm{T}}\right]\right\}
\end{aligned}$$

$$= \sigma^2 + \mathrm{tr}\left\{\mathbb{E}_{X_0}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot X_0 X_0^{\mathrm{T}}\right]\right\}$$

$$= \sigma^2 + \mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot \mathbb{E}_{X_0}\left(X_0 X_0^{\mathrm{T}}\right)\right]$$

$$= \sigma^2 + \mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right].$$

Likewise, (24b) and (8c) give:

$$\begin{aligned}
\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) &= \mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}\right] = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}, \boldsymbol{y}\right] \\
&= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\sigma^2 + \mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]\right\} \\
&= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left(\sigma^2\right) + \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]\right\} \\
&= \sigma^2 + \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\mathrm{tr}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]\right\} \\
&= \sigma^2 + \mathrm{tr}\left\{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]\right\} \\
&= \sigma^2 + \mathrm{tr}\left\{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\mathrm{T}}\right] \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right\} \\
&= \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta} \mid \boldsymbol{X}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right].
\end{aligned}$$

From relations (24c) and (7), one gets:

$$\begin{aligned}
\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) &= \mathbb{E}\left(\hat{Y}_0 - Y_0\right)^2 = \mathbb{E}_{\boldsymbol{X}}\mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left\{\sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta} \mid \boldsymbol{X}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]\right\} \\
&= \mathbb{E}_{\boldsymbol{X}}\left(\sigma^2\right) + \mathbb{E}_{\boldsymbol{X}}\left\{\mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta} \mid \boldsymbol{X}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]\right\} \\
&= \sigma^2 + \mathrm{tr}\left\{\mathbb{E}_{\boldsymbol{X}}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta} \mid \boldsymbol{X}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]\right\} \\
&= \sigma^2 + \mathrm{tr}\left\{\mathbb{E}_{\boldsymbol{X}}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta} \mid \boldsymbol{X}\right)\right] \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right\} \\
&= \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right].
\end{aligned}$$

□

The above result is interesting in that it imposes no assumption on $\hat{\beta}$, hence it is valid for any LM fitter. But an immediate important subcase is provided in:

Corollary 2 *If, conditional on* $\boldsymbol{X}$, $\hat{\beta}$ *estimates* β *unbiasedly in the LM* (14), *then, under Assumptions* ($\mathscr{A}_0$), ($\mathscr{A}_1$) *and* ($\mathscr{A}_2$):

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{cov}\left(\hat{\beta} \mid \boldsymbol{X}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right], \tag{27a}$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{cov}\left(\hat{\beta}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]. \tag{27b}$$

4. MSPE When Fitting the LM by Ordinary Least Squares

By far, the most popular approach to estimating the parameter vector β in the LM (14) is through minimizing the Ordinary Least Squares (OLS) criterion us-

ing the observed data (3):

$$\hat{\beta} = \hat{\beta}_{\mathrm{OLS}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\mathrm{T}} \beta \right)^2. \tag{28}$$

The properties of $\hat{\beta}_{\mathrm{OLS}}$ as an estimate of β depend on whether the design matrix $\boldsymbol{X}$ has full column rank or not. This remains true when studying the corresponding MSPE as well.

4.1. MSPE in the LM Fitted by OLS with *X* of Full Column Rank

4.1.1. LM Fitted by OLS with *X* Full Column Rank

Here, we consider the LM (14) fitted under:

($\mathscr{A}_3$). $\operatorname{rank}(\boldsymbol{X}) = p$, *i.e.* $\boldsymbol{X}$ *is a full column rank matrix.*

That assumption is known to be equivalent to saying that the square matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is nonsingular, thus implying that

$$\hat{\beta}_{\mathrm{OLS}} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y}. \tag{29}$$

Furthermore, given Assumptions ($\mathscr{A}_0$)-($\mathscr{A}_1$), (18) holds, so

$$\mathbb{E}\left(\hat{\beta}_{\mathrm{OLS}} \mid \boldsymbol{X} \right) = \beta \quad \text{and} \quad \mathbb{M}\mathrm{cov}\left(\hat{\beta}_{\mathrm{OLS}} \mid \boldsymbol{X} \right) = \sigma^2 \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1}. \tag{30}$$

We also recall that under these assumptions, with $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\mathrm{OLS}}$ the residual response vector and $\|\cdot\|$ the Euclidean norm, a computable unbiased estimate of the residual variance σ^2 is:

$$\hat{\sigma}^2 = \frac{\mathrm{SSR}}{n-p}, \quad \text{with} \quad \mathrm{SSR} = \|\hat{\boldsymbol{e}}\|^2 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\mathrm{T}} \hat{\beta}_{\mathrm{OLS}} \right)^2, \tag{31}$$

the latter being the *sum of squared residuals* in the OLS fit to the data.

Now, from the first identity in (30), we deduce that when $\boldsymbol{X}$ is full column rank, $\hat{\beta}_{\mathrm{OLS}}$ is an unbiased estimate of β, conditionally on $\boldsymbol{X}$. Hence, combining the second identity in (30) with Corollary 2 yields:

Theorem 3 *In the LM* (14) *fitted by OLS with Assumptions* ($\mathscr{A}_0$), ($\mathscr{A}_1$), ($\mathscr{A}_2$) *and* ($\mathscr{A}_3$),

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X} \right) = \sigma^2 \left(1 + \mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}} \right) \right] \right), \tag{32a}$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \right) = \sigma^2 \left(1 + \mathrm{tr}\left\{ \mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \right] \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}} \right) \right\} \right). \tag{32b}$$

Proof. From (24c),

$$\begin{aligned}
\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X} \right) &= \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_{\mathrm{OLS}} \mid \boldsymbol{X} \right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}} \right) \right] \\
&= \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{cov}\left(\hat{\beta}_{\mathrm{OLS}} \mid \boldsymbol{X} \right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}} \right) \right] \\
&= \sigma^2 + \mathrm{tr}\left[\sigma^2 \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}} \right) \right] \\
&= \sigma^2 + \sigma^2 \mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}} \right) \right] \\
&= \sigma^2 \left(1 + \mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}} \right) \right] \right).
\end{aligned}$$

Now, using (32a), one gets:

$$\begin{aligned}\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) &= \mathbb{E}_{\boldsymbol{X}}\mathbb{E}\left[\left(\hat{Y}_0 - Y_0\right)^2 \mid \boldsymbol{X}\right] \\ &= \mathbb{E}_{\boldsymbol{X}}\left\{\sigma^2\left(1+\mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\cdot\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)\right]\right)\right\} \\ &= \sigma^2\left\{\mathbb{E}_{\boldsymbol{X}}\left(1+\mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\cdot\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)\right]\right)\right\} \\ &= \sigma^2\left\{1+\mathbb{E}_{\boldsymbol{X}}\left(\mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\cdot\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)\right]\right)\right\} \\ &= \sigma^2\left\{1+\mathrm{tr}\left(\mathbb{E}_{\boldsymbol{X}}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\cdot\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)\right]\right)\right\} \\ &= \sigma^2\left(1+\mathrm{tr}\left\{\mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right]\cdot\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)\right\}\right),\end{aligned}$$

from which the result is got. □

4.1.2. The FPE and the GCV in the LM Fitted by OLS with *X* of Full Column Rank

From (32b) in Theorem 3, we deduce a closed form computable estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$, using data (3), by estimating, respectively:

- the residual variance σ^2 by $\hat{\sigma}^2$ given by (31);
- the $p\times p$ expectation matrix $\mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right]$ by the observed $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$;
- the $p\times p$ expectation matrix $\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)$ by (given that $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ are *i.i.d.* $\overset{\mathscr{D}}{\sim} X_0$):

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}. \tag{33}$$

Therefore, one estimates $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ by

$$\begin{aligned}\widehat{\mathrm{MSPE}}\left(\hat{Y}_0, Y_0\right) &= \hat{\sigma}^2\left\{1+\frac{1}{n}\mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\cdot\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\right]\right\} = \hat{\sigma}^2\left[1+\frac{1}{n}\mathrm{tr}\left(\boldsymbol{I}_p\right)\right] \\ &= \hat{\sigma}^2\left(1+\frac{p}{n}\right) = \frac{n+p}{n-p}\cdot\frac{\mathrm{SSR}}{n} = \hat{S}^2\cdot\frac{n+p}{n-p} = \mathrm{FPE},\end{aligned} \tag{34}$$

with

$$\hat{S}^2 = \mathrm{SSR}/n = \|\hat{\boldsymbol{e}}\|^2/n = \left\|\boldsymbol{y}-\boldsymbol{X}\hat{\beta}_{\mathrm{OLS}}\right\|^2/n, \tag{35}$$

the usual Maximum Likelihood Estimator of the residual variance σ^2 in the LM (14) when the residual error ε is assumed to follow a $\mathscr{N}\left(0,\sigma^2\right)$ Gaussian distribution.

We see that the final estimate $\widehat{\mathrm{MSPE}}\left(\hat{Y}_0, Y_0\right)$ obtained for $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ coincides with Akaike's Final Prediction Error (FPE) goodness-of-fit criterion for the LM (14) fitted by OLS [1]. The main difference between the derivation above and the traditonal one is that the latter uses the sample viewpoint of the prediction error reviewed in Section 2.4. The latter excludes the possibility that covariates values on a future individual might be completely unrelated to the observed $\boldsymbol{x}_i$'s in the sample (3). In particular, it does not account for any po-

tential random origin for the design matrix $\boldsymbol{X}$, a situation often encountered in certain areas of application of the LM such as in econometrics.

On the other hand, estimating, instead, $\mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)$ by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/(n-p)$ yields as estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$:

$$\hat{\sigma}^2\left(1+\frac{p}{n-p}\right)=\frac{n\cdot\mathrm{SSR}}{(n-p)^2}=\frac{\hat{S}^2}{(1-p/n)^2}=\mathrm{GCV}, \tag{36}$$

the traditional GCV estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ in the LM fitted by OLS with $\boldsymbol{X}$ full column rank.

Remark 1 *Note that the very way the two estimates* (34) *and* (36) *of* $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ *were derived above implies that they can also validly serve, each, as an estimate of the conditional* $\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ *given by* (32*a*). *This will remain true for all the estimates derived for the MSPE under the other scenarios examined in this paper.*

4.2. MSPE in the LM Fitted by OLS with *X* Not of Full Column Rank

4.2.1. LM Fitted by OLS with *X* Column Rank Deficient

Although Assumption ($\mathscr{A}_3$) is routinely admitted by most people when handling the LM, one actually meets many concrete instances of data sets where it does not hold. Fortunately, with the formalization by Moore [24], Penrose [25] and, especially, Rao [26] of the notion of *generalized inverse* (short: *g*-inverse) of an arbitrary matrix, it became possible to handle least squares estimation in the LM without having to assume the design matrix $\boldsymbol{X}$ necessarily of full column rank.

To begin with, it is shown in most textbooks on the LM that whatever the rank of the design matrix $\boldsymbol{X}$, a vector $\hat{\beta}\in\mathbb{R}^p$ is a solution to the OLS minimization problem (28) *if, and only if,* $\hat{\beta}$ is a solution to the so called *normal equations*:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\beta}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}. \tag{37}$$

When Assumption ($\mathscr{A}_3$) holds, the unique solution to the normal equations is clearly $\hat{\beta}=\hat{\beta}_{\mathrm{OLS}}$ given by (29). When that's not the case, the square matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is singular, hence does not have a regular inverse $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$. Nevertheless, even then it can be shown that the normal Equation (37) are always consistent. But the apparent negative thing is that they then have infinitely many solution vectors $\hat{\beta}$, actually all vectors $\hat{\beta}\in\mathbb{R}^p$ of the form:

$$\hat{\beta}=\hat{\beta}_{\mathrm{OLS}}^{-}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \tag{38}$$

where $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}$ is any *g*-inverse of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ in the sense of Rao. Given that multitude of possible OLS estimates of β in this case, one may worry that this may hinder any attempt to get a meaningful estimate of the MSPE in the fitted LM. But we are going to show that such a worry is not warranted.

When Assumption ($\mathscr{A}_3$) does not hold, in spite of there being as many solutions $\hat{\beta}_{\mathrm{OLS}}^{-}$ to the normal Equation (37) as there are *g*-inverse s $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}$ of

$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, *i.e.* infinitely many, it is a remarkable well known fact that the fitted response vector

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}_{\mathrm{OLS}}^{-} = H_{\boldsymbol{X}} \cdot \boldsymbol{y}, \quad \text{with} \quad H_{\boldsymbol{X}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}, \tag{39}$$

is the same, whichever *g*-inverse $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}$ of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is used to compute $\hat{\beta}_{\mathrm{OLS}}^{-}$ through (38). This stems from the hat matrix $H_{\boldsymbol{X}}$ being equal to the matrix of the orthogonal projection of $\mathbb{R}^n$ into the range space of $\boldsymbol{X}$ ([22] Appendix A). Therefore, the residual response vector

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\mathrm{OLS}}^{-} = \left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\boldsymbol{y}, \tag{40}$$

does not vary with the *g*-inverse $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}$ either. Then we will need the result:

Lemma 4 *With* $r_{\boldsymbol{X}} = \operatorname{rank}(\boldsymbol{X})$, *one has*: $\operatorname{tr}\left(H_{\boldsymbol{X}}\right) = r_{\boldsymbol{X}}$ *and* $\mathbb{E}\left(\left\|\hat{\boldsymbol{e}}\right\|^2 \mid \boldsymbol{X}\right) = \left(n - r_{\boldsymbol{X}}\right)\cdot\sigma^2$.

Proof. On the one hand, one has:

$$\operatorname{tr}\left(H_{\boldsymbol{X}}\right) = \operatorname{tr}\left[\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\right] = \operatorname{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]. \tag{41}$$

Now, $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}$ being a g-inverse of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, it comes that $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is an idempotent matrix ([22] Appendix A, page 509]. Hence

$$\operatorname{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = \operatorname{rank}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = \operatorname{rank}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) = \operatorname{rank}(\boldsymbol{X}) = r_{\boldsymbol{X}}. \tag{42}$$

Relations (41) and (42) give $\operatorname{tr}\left(H_{\boldsymbol{X}}\right) = r_{\boldsymbol{X}}$. On the other hand,

$$\hat{\boldsymbol{e}} = \left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\boldsymbol{y} = \left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\left(\boldsymbol{X}\beta + \varepsilon\right) = \left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\boldsymbol{X}\beta + \left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\varepsilon. \tag{43}$$

Now, $H_{\boldsymbol{X}}$ being the matrix of the orthogonal projection of $\mathbb{R}^n$ into the range space of $\boldsymbol{X}$, then $H_{\boldsymbol{X}}\boldsymbol{X} = \boldsymbol{X}$. Hence,

$$\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\boldsymbol{X} = \boldsymbol{X} - H_{\boldsymbol{X}}\boldsymbol{X} = 0. \tag{44}$$

From (43) and (44), we get $\hat{\boldsymbol{e}} = \left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\varepsilon$. Therefore:

$$\left\|\hat{\boldsymbol{e}}\right\|^2 = \varepsilon^{\mathrm{T}}\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)^{\mathrm{T}}\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\varepsilon = \varepsilon^{\mathrm{T}}\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)^2\varepsilon = \varepsilon^{\mathrm{T}}\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\varepsilon.$$

Under Assumptions ($\mathscr{A}_0$),($\mathscr{A}_1$), and thanks to the known identity which gives the expectation of a quadratic form ([27] Appendix B, page 170], one has:

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{e}}\right\|^2 \mid \boldsymbol{X}\right] &= \left[\mathbb{E}(\varepsilon \mid \boldsymbol{X})\right]^{\mathrm{T}}\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\mathbb{E}(\varepsilon \mid \boldsymbol{X}) + \operatorname{tr}\left[\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\mathbb{M}\mathrm{cov}(\varepsilon \mid \boldsymbol{X})\right] \\ &= \operatorname{tr}\left[\left(\boldsymbol{I}_n - H_{\boldsymbol{X}}\right)\sigma^2\boldsymbol{I}_n\right] = \sigma^2\left[\operatorname{tr}\left(\boldsymbol{I}_n\right) - \operatorname{tr}\left(H_{\boldsymbol{X}}\right)\right] = \sigma^2\left(n - r_{\boldsymbol{X}}\right)\end{aligned}$$

□

An unbiased estimate of the LM residual variance σ^2 in this case is thus known to be:

$$\hat{\sigma}_{\boldsymbol{X}}^2 = \frac{\mathrm{SSR}}{n - r_{\boldsymbol{X}}}, \quad \text{with} \quad \mathrm{SSR} = \left\|\hat{\boldsymbol{e}}\right\|^2 = \left\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\mathrm{OLS}}^{-}\right\|^2 = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\beta}_{\mathrm{OLS}}^{-}\right)^2. \tag{45}$$

We will also need the mean vector and covariance matrix of $\hat{\beta}_{\mathrm{OLS}}^{-}$. First,

$$\mathbb{E}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\beta, \tag{46a}$$

which shows that $\hat{\beta}_{\mathrm{OLS}}^{-}$ is a biased estimator of β when Assumption ($\mathscr{A}_3$)

does not hold. But in spite of that, note that from (39),

$$\mathbb{E}\left(\hat{\boldsymbol{y}} \mid \boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{X}\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right) = \boldsymbol{X}\mathbb{E}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right) = H_{\boldsymbol{X}}\boldsymbol{X}\beta = \boldsymbol{X}\beta = \mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right). \quad (46b)$$

On the other hand,

$$\mathbb{M}\mathrm{cov}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right) = \sigma^2\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)_S^{-}, \text{ with } \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)_S^{-} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}\right]^{\mathrm{T}}, \quad (47a)$$

the symmetric and positive semi-definite matrix $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)_S^{-}$ also being a g-inverse of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Then

$$\begin{aligned}\mathbb{M}\mathrm{cov}\left(\hat{\boldsymbol{y}} \mid \boldsymbol{X}\right) &= \mathbb{M}\mathrm{cov}\left(\boldsymbol{X}\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right) = \boldsymbol{X}\mathbb{M}\mathrm{cov}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right)\boldsymbol{X}^{\mathrm{T}} \\ &= \sigma^2\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)_S^{-}\boldsymbol{X}^{\mathrm{T}} = \sigma^2 H_{\boldsymbol{X}},\end{aligned} \quad (47b)$$

again independent of the g-inverse $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}$ of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ used to compute $\hat{\beta}_{\mathrm{OLS}}^{-}$.

4.2.2. Preliminary for the MSPE in the LM Fitted by OLS without Assumption ($\mathscr{A}_3$)

Our first aim here is to examine the MSPE in the LM when fitted by OLS under the assumption that the design matrix $\boldsymbol{X}$ might not have full column rank. So β has been estimated through $\hat{\beta}_{\mathrm{OLS}}^{-}$ given by (38) and we are interested in $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) = \mathbb{E}\left(\hat{Y}_0 - Y_0\right)^2$, where $\hat{Y}_0 = X_0^{\mathrm{T}}\hat{\beta}_{\mathrm{OLS}}^{-}$ is taken as prediction of Y on an independently sampled new individual for whom $X = X_0 \in \mathbb{R}^p$ would have been observed, but not $Y = Y_0 \in \mathbb{R}$. We are going to use the results of Section 3.2. First note that since, from the above, $\hat{\beta}_{\mathrm{OLS}}^{-}$ is a biased estimator of β, Corollary 2 does not apply here. Nonetheless, from Theorem 1, we get:

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right)\cdot\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)\right], \quad (48a)$$

$$\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_{\mathrm{OLS}}^{-}\right)\cdot\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)\right]. \quad (48b)$$

Our estimation of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ in this case will be based on those two identities and:

Lemma 5 *In the LM* (14) *with Assumptions* ($\mathscr{A}_0$)-($\mathscr{A}_1$),

$$\mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right)\cdot\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = r_{\boldsymbol{X}}\cdot\sigma^2, \quad \text{with } r_{\boldsymbol{X}} = \mathrm{rank}\left(\boldsymbol{X}\right). \quad (49)$$

Proof. For

$$A(\boldsymbol{X}) = \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right)\cdot\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = \mathrm{tr}\left\{\mathbb{E}\left[\left(\hat{\beta}_{\mathrm{OLS}}^{-} - \beta\right)\left(\hat{\beta}_{\mathrm{OLS}}^{-} - \beta\right)^{\mathrm{T}} \mid \boldsymbol{X}\right]\cdot\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right\},$$

$$\begin{aligned}A(\boldsymbol{X}) &= \mathrm{tr}\left\{\boldsymbol{X}\cdot\mathbb{E}\left[\left(\hat{\beta}_{\mathrm{OLS}}^{-} - \beta\right)\left(\hat{\beta}_{\mathrm{OLS}}^{-} - \beta\right)^{\mathrm{T}} \mid \boldsymbol{X}\right]\cdot\boldsymbol{X}^{\mathrm{T}}\right\} \\ &= \mathrm{tr}\left\{\mathbb{E}\left[\boldsymbol{X}\left(\hat{\beta}_{\mathrm{OLS}}^{-} - \beta\right)\left(\hat{\beta}_{\mathrm{OLS}}^{-} - \beta\right)^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}} \mid \boldsymbol{X}\right]\right\} \\ &= \mathrm{tr}\left\{\mathbb{E}\left[\left(\hat{\boldsymbol{y}} - \boldsymbol{X}\beta\right)\left(\hat{\boldsymbol{y}} - \boldsymbol{X}\beta\right)^{\mathrm{T}} \mid \boldsymbol{X}\right]\right\} \\ &= \mathrm{tr}\left[\mathbb{M}\mathrm{cov}\left(\hat{\boldsymbol{y}} \mid \boldsymbol{X}\right)\right], \text{ thanks to (46b)} \\ &= \sigma^2\mathrm{tr}\left(H_{\boldsymbol{X}}\right), \text{ by (47b).}\end{aligned}$$

Hence (49), thanks to Lemma 4. □

4.2.3. The FPE and the GCV in the LM Fitted by OLS with X Column Rank Deficient

Given (48a), we first estimate $\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)$ by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ as in (33), which entails the following preliminary estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ in the present case, using the last lemma:

$$\widetilde{\mathrm{MSPE}}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) = \sigma^2 + \frac{1}{n}\mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_{\mathrm{OLS}}^{-} \mid \boldsymbol{X}\right)\cdot \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = \sigma^2\left(1+\frac{r_X}{n}\right). \tag{50}$$

But since $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right) = \mathbb{E}_{\boldsymbol{X}}\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$, then (51) also gives a preliminary estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$. Then, estimating σ^2 by $\hat{\sigma}_X^2$ given by (45), our final estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ in this case, computable from data, is:

$$\mathrm{FPE} = \hat{\sigma}_X^2\left(1+\frac{r_X}{n}\right) = \hat{S}^2\cdot\frac{n+r_X}{n-r_X}, \tag{51}$$

which is also an estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$. It is denoted FPE because it generalizes (34) in assessing the goodness-of-fit of OLS in the LM when the design matrix $\boldsymbol{X}$ is column rank deficient. The remarkable feature is that this estimate is the same, whichever g-inverse $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-}$ of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ was used to get the estimate $\hat{\beta}_{\mathrm{OLS}}^{-}$ of β in (38).

Estimating, instead, $\mathbb{E}\left(X_0X_0^{\mathrm{T}}\right)$ by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/\left(n-r_X\right)$ yields as estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$:

$$\hat{\sigma}_X^2\left(1+\frac{r_X}{n-r_X}\right) = \frac{n\cdot\mathrm{SSR}}{\left(n-r_X\right)^2} = \frac{\hat{S}^2}{\left(1-r_X/n\right)^2} = \mathrm{GCV}, \tag{52}$$

the GCV estimate of $\mathrm{MSPE}\left(\hat{Y}_0, Y_0\right)$ in the LM fitted by OLS when $\boldsymbol{X}$ is not full column rank.

5. MSPE when Fitting the LM by Ridge Regression

The design matrix $\boldsymbol{X}$ being column rank deficient means that its p columns are linearly dependent, or almost so. This happens when there is multicollinearity among the p regressors $X_1, \cdots, X_p$, and thus some of them are redundant with some others. However, when this occurs, computing an OLS estimate $\hat{\beta}_{\mathrm{OLS}}^{-}$ of β, given by (38), in a numerically stable manner is not easy and requires using carefully designed Numerical Linear Algebra programming routines. The difficulty stems from the fact that this requires, at least implicitly, to find, along the way, the exact rank of $\boldsymbol{X}$, which is difficult to achieve, precisely because of the multicollinearity among its columns. It can then be of much interest to have a method which can fit the LM without having to bother about the exact rank of $\boldsymbol{X}$. This is precisely what Ridge regression (RR) tries to achieve.

Hoerl and Kennard presented two variants of Ridge Regression [19]. In the initial one (the default), which some have termed Ordinary Ridge Regression (ORR), to fit the LM (14), one estimates β through regularizing the OLS criterion (28) by a ridge constraint, yielding:

$$\hat{\beta} = \hat{\beta}_\lambda = \underset{\beta\in\mathbb{R}^p}{\mathrm{argmin}}\left[\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\beta\right)^2 + \lambda\|\beta\|^2\right], \tag{53}$$

for some $\lambda > 0$, a penalty parameter to choose appropriately. The unique solution to (54) is known to be (whatever the rank of $\boldsymbol{X}$):

$$\hat{\beta}_\lambda = G_\lambda^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \quad \text{with} \quad G_\lambda = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p. \tag{54}$$

Hoerl and Kennard presented ORR in [19] assuming that the design matrix $\boldsymbol{X}$ was full column rank (*i.e.* our Assumption ($\mathscr{A}_3$)), which also requires that $n \geq p$. But because the symmetric matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is always at least semi-positive definite, imposing $\lambda > 0$ entails that the $p \times p$ matrix G_λ is symmetric and positive definite (SPD), whatever the rank of $\boldsymbol{X}$ and the ranking between the integers *n* and *p*. That is why, in what follows, we do not impose any rank constraint on $\boldsymbol{X}$ (apart being trivially ≥1 because $\boldsymbol{X}$ is nonzero). No specific ranking either is assumed between *n* and *p*.

However, hereafter, since it does not require extra work, we directly consider the extended setting of Generalized Ridge Regression (GRR) which, to fit the LM (14), estimates β through solving the minimization problem:

$$\hat{\beta} = \hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\beta \right)^2 + \lambda \cdot \|\beta\|_{\boldsymbol{D}}^2 \right], \tag{55}$$

where λ is as in ORR and $\boldsymbol{D}$ is a $p \times p$ symmetric and semi-positive definite (SSPD) matrix, both given, while $\|\beta\|_{\boldsymbol{D}}^2 = \beta^{\mathrm{T}}\boldsymbol{D}\beta$. The solution of (56) is still (55), but now with

$$G_\lambda = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{D}, \tag{56}$$

under the assumption that the latter is SPD. Since smoothing splines fitting can be cast in the GRR form (56), what follows applies to that hugely popular non-parametric regression method as well.

5.1. The MSPE Issue for Ridge Regression

More than for Least Squares, depending on the unspecified parameter λ, it is critical to assess how Ridge Regression fits the LM for given data in order to be able to select the best λ value, *i.e.* the one ensuring the best fit. From the prediction error point of view stated in this article, this amounts to choosing the $\lambda > 0$ for which the RR fit has the smallest MSPE. It, thus, requires to estimate $\mathrm{MSPE}_\lambda\left(\hat{Y}_0, Y_0\right) = \mathbb{E}\left(\hat{Y}_0 - Y_0\right)^2$ for any given λ value, where X_0 and Y_0 are as before, while $\hat{Y}_0 = X_0^{\mathrm{T}}\hat{\beta}_\lambda$. Traditionally, estimating $\mathrm{MSPE}_\lambda\left(\hat{Y}_0, Y_0\right)$ in this context is mostly done using the Generalized Cross Validation (GCV) criterion initially developed by Craven and Wahba for selecting the best value of the smoothing parameter in a smoothing spline [3]. That GCV is obtained as a variation of the LOO-CV. Here, to estimate $\mathrm{MSPE}_\lambda\left(\hat{Y}_0, Y_0\right)$, we take a different route. We first note that the fitted sample response vector is

$$\hat{\boldsymbol{y}}_\lambda = \boldsymbol{X}\hat{\beta}_\lambda = H_\lambda \boldsymbol{y}, \quad \text{with} \quad H_\lambda = \boldsymbol{X}G_\lambda^{-1}\boldsymbol{X}^{\mathrm{T}}. \tag{57}$$

On the other hand, from (55),

$$\mathbb{E}\left(\hat{\beta}_\lambda \mid \boldsymbol{X}\right) = G_\lambda^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = T_\lambda \beta, \quad \text{with} \quad T_\lambda = G_\lambda^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \neq \boldsymbol{I}_p. \tag{58}$$

So, again, Corollary 2 does not apply. But, using Theorem 1,

$$\mathrm{MSPE}_\lambda\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_\lambda \mid \boldsymbol{X}\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right], \tag{59a}$$

$$\mathrm{MSPE}_\lambda\left(\hat{Y}_0, Y_0\right) = \sigma^2 + \mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_\lambda\right) \cdot \mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)\right]. \tag{59b}$$

For estimating those quantities, we will need some preliminary results.

5.2. Preliminary Results for Estimating the MSPE in Ridge Regression

First, two simple, but remarkable identities about the matrices H_λ and T_λ given in (58) and (59).

Lemma 6 $T_\lambda = \boldsymbol{I}_p - \lambda G_\lambda^{-1}\boldsymbol{D}$ and $\lambda \boldsymbol{X} G_\lambda^{-1}\boldsymbol{D} = \left(\boldsymbol{I}_n - H_\lambda\right)\boldsymbol{X}$.

Proof. The first identity is easily got from (59) and (57). Indeed, one has:

$$\boldsymbol{I}_p = G_\lambda^{-1} G_\lambda = G_\lambda^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + G_\lambda^{-1}\lambda\boldsymbol{D} = T_\lambda + \lambda G_\lambda^{-1}\boldsymbol{D},$$

As for the second one,

$$\lambda \boldsymbol{X} G_\lambda^{-1}\boldsymbol{D} = \boldsymbol{X} G_\lambda^{-1}\left(G_\lambda - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) = \boldsymbol{X} - \boldsymbol{X} G_\lambda^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \left(\boldsymbol{I}_n - H_\lambda\right)\boldsymbol{X}. \quad \square$$

$\square$

Next, a key preliminary about the Mean Squared Error matrix of $\hat{\beta}_\lambda$ as an estimate of β:

Lemma 7 *Under Assumptions* ($\mathscr{A}_0$)-($\mathscr{A}_1$), *one has*:

$$\mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_\lambda \mid \boldsymbol{X}\right) \cdot \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = \sigma^2 \mathrm{tr}\left(H_\lambda^2\right) + \left\|\left(\boldsymbol{I}_n - H_\lambda\right)\boldsymbol{X}\beta\right\|^2. \tag{60}$$

Proof. Inserting (20) in $\mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_\lambda \mid \boldsymbol{X}\right) \cdot \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = \mathrm{tr}\left[\boldsymbol{X}\mathbb{M}\mathrm{se}\left(\hat{\beta}_\lambda \mid \boldsymbol{X}\right)\boldsymbol{X}^{\mathrm{T}}\right]$ gives:

$$\mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_\lambda \mid \boldsymbol{X}\right) \cdot \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] = \mathrm{tr}(A) + \mathrm{tr}(B),$$

with $A = \boldsymbol{X}\mathbb{M}\mathrm{cov}\left(\hat{\beta}_\lambda \mid \boldsymbol{X}\right)\boldsymbol{X}^{\mathrm{T}}$ and $B = \boldsymbol{X}\mathbb{B}\mathrm{ias}\left(\hat{\beta}_\lambda\right) \cdot \left[\boldsymbol{X}\mathbb{B}\mathrm{ias}\left(\hat{\beta}_\lambda\right)\right]^{\mathrm{T}}$. Now,

$$A = \mathbb{M}\mathrm{cov}\left(\hat{\boldsymbol{y}}_\lambda \mid \boldsymbol{X}\right) = H_\lambda \mathbb{M}\mathrm{cov}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) H_\lambda = \sigma^2 H_\lambda^2$$

$$\mathrm{tr}(B) = \mathrm{tr}\left\{\left[\boldsymbol{X}\mathbb{B}\mathrm{ias}\left(\hat{\beta}_\lambda\right)\right]^{\mathrm{T}} \boldsymbol{X}\mathbb{B}\mathrm{ias}\left(\hat{\beta}_\lambda\right)\right\} = \left\|\boldsymbol{X}\mathbb{B}\mathrm{ias}\left(\hat{\beta}_\lambda\right)\right\|^2$$

$$= \left\|\boldsymbol{X}\left(T_\lambda - \boldsymbol{I}_p\right)\beta\right\|^2 = \left\|\lambda \boldsymbol{X} G_\lambda^{-1}\boldsymbol{D}\beta\right\|^2 = \left\|\left(\boldsymbol{I}_n - H_\lambda\right)\boldsymbol{X}\beta\right\|^2,$$

the last three identities using (59) and Lemma 6. $\square$

With the above lemma, we are now in a position to be able to estimate $\mathrm{MSPE}_\lambda\left(\hat{Y}_0, Y_0\right)$ in Ridge Regression. We examine, hereafter, two paths for achieving that: one leads to the FPE, the other one to the GCV.

5.3. Estimating the MSPE in Ridge Regression by the FPE

Here, we first estimate $\mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)$ by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ in (60a). Then, given (61), this suggests the preliminary estimate of $\mathrm{MSPE}_\lambda\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ in RR:

$$\begin{aligned}\widetilde{\mathrm{MSPE}}_{\lambda}^{0}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right) &= \sigma^2 + \frac{1}{n}\mathrm{tr}\left[\mathbb{M}\mathrm{se}\left(\hat{\beta}_{\lambda} \mid \boldsymbol{X}\right)\cdot \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right] \\ &= \sigma^2\left[1+\frac{\mathrm{tr}\left(H_{\lambda}^2\right)}{n}\right] + \frac{\left\|\left(\boldsymbol{I}_n - H_{\lambda}\right)\boldsymbol{X}\beta\right\|^2}{n},\end{aligned} \tag{61}$$

which is also, therefore, a preliminary estimate of $\mathrm{MSPE}_{\lambda}\left(\hat{Y}_0, Y_0\right)$. It is only a preliminary estimate because even given λ, it still depends on the two unknowns σ^2 and β. Interestingly, it can be shown ([8] Chapter 3, page 46) that (62) coincides with $\mathrm{PSE}\left(\hat{\boldsymbol{y}}_{\lambda} \mid \boldsymbol{X}\right)$ given by (62) in the present setting.

It is useful to note that (62) depends on β only through the squared bias term $\left\|\left(\boldsymbol{I}_n - H_{\lambda}\right)\boldsymbol{X}\beta\right\|^2$. To estimate the latter, let $\hat{\boldsymbol{e}}_{\lambda} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{\lambda} = \left(\boldsymbol{I}_n - H_{\lambda}\right)\boldsymbol{y}$, the vector of sample residuals, and $\mathrm{SSR}_{\lambda} = \left\|\hat{\boldsymbol{e}}_{\lambda}\right\|^2$, the sum of squared residuals in the Ridge Regression fit. Then, thanks to a well known identity ([9] Chapter 2, page 38),

$$\mathbb{E}\left(\mathrm{SSR}_{\lambda} \mid \boldsymbol{X}\right) = \left\|\left(\boldsymbol{I}_n - H_{\lambda}\right)\boldsymbol{X}\beta\right\|^2 + \sigma^2\mathrm{tr}\left[\left(\boldsymbol{I}_n - H_{\lambda}\right)^2\right],$$

implying that $\mathrm{SSR}_{\lambda} - \sigma^2\mathrm{tr}\left[\left(\boldsymbol{I}_n - H_{\lambda}\right)^2\right]$ is an unbiased estimate of $\left\|\left(\boldsymbol{I}_n - H_{\lambda}\right)\boldsymbol{X}\beta\right\|^2$. Hence a general formula for computing an estimate of the MSPE in Ridge Regression:

$$\mathrm{FPE}\left(\hat{\sigma}^2\right) = \frac{2\hat{\sigma}^2\mathrm{tr}\left(H_{\lambda}\right) + \mathrm{SSR}_{\lambda}}{n}, \tag{62}$$

where $\hat{\sigma}^2$ is a chosen estimate of σ^2, possibly computed from the RR fit, thus dependent on λ.

We denoted $\mathrm{FPE}\left(\hat{\sigma}^2\right)$ the estimate of the MSPE given by (63) for the reason to follow. Indeed, probably the most popular estimate of the residual variance σ^2 from an RR fit is the one proposed by Wahba in the context of smoothing splines [28]:

$$\hat{\sigma}_1^2 = \frac{\mathrm{SSR}_{\lambda}}{n - \mathrm{tr}\left(H_{\lambda}\right)}. \tag{63}$$

Now, if one uses $\hat{\sigma}^2 = \hat{\sigma}_1^2$ in (63), an algebraic manipulation easily leads to:

$$\mathrm{FPE}\left(\hat{\sigma}_1^2\right) = \frac{\mathrm{SSR}_{\lambda}}{n}\cdot\frac{n + \mathrm{tr}\left(H_{\lambda}\right)}{n - \mathrm{tr}\left(H_{\lambda}\right)} = \mathrm{FPE}, \tag{64}$$

recovering the classical formula of the FPE for this setting, but this time as an estimate of the MSPE rather than the PSE.

5.4. Estimating the MSPE in Ridge Regression by the GCV

Here, we estimate $\mathbb{E}\left(X_0 X_0^{\mathrm{T}}\right)$ in (60a) rather by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/\left[n - \mathrm{tr}\left(H_{\lambda}\right)\right]$. With (61), this suggests a preliminary estimate $\widetilde{\mathrm{MSPE}}_{\lambda}^{1}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ of $\mathrm{MSPE}_{\lambda}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ in RR, correspondig to $\widetilde{\mathrm{MSPE}}_{\lambda}^{0}\left(\hat{Y}_0, Y_0 \mid \boldsymbol{X}\right)$ where the denominators n have been replaced by $n - \mathrm{tr}\left(H_{\lambda}\right)$. Using the same unbiased estimate of $\left\|\left(\boldsymbol{I}_n - H_{\lambda}\right)\boldsymbol{X}\beta\right\|^2$ as in the previous section and an estimate $\hat{\sigma}^2$ of σ^2, we get another general formula for computing an estimate of the MSPE in RR:

$$\mathrm{GCV}\left(\hat{\sigma}^2\right)=\frac{\hat{\sigma}^2\mathrm{tr}\left(H_\lambda\right)+\mathrm{SSR}_\lambda}{n-\mathrm{tr}\left(H_\lambda\right)}. \tag{65}$$

We denoted it $\mathrm{GCV}\left(\hat{\sigma}^2\right)$ because again taking $\hat{\sigma}^2=\hat{\sigma}_1^2$, one gets:

$$\mathrm{GCV}\left(\hat{\sigma}_1^2\right)=\frac{\mathrm{SSR}_\lambda}{n\left[1-n^{-1}\mathrm{tr}\left(H_\lambda\right)\right]^2}=\mathrm{GCV}, \tag{66}$$

the well known formula of Craven and Wahba's GCV [3] for this setting, but this time derived without any reference to Cross Validation.

6. Conclusion and Perspectives

In this work, the goal was not to derive new and better selection criteria to assess the goodness-of-fit in regression, but rather to show how one can derive the well known Akaike's FPE and, Craven and Wahba's GCV [3] as direct estimates of the measure of prediction error most commonly known to users, which is not how they are traditionally derived. We achieved this for some of the best known linear model fitters, with the two derivations differing only slightly for each of them. But, nowadays, in regression, use of the FPE criterion is generally not recommended because much better performing criteria are known [29], while GCV has its own shortcomings in certain settings (*e.g.* small sample size), though hugely popular and almost the best in some difficult high dimensional situations [12]. It is then our hope that, in the future, one can, through the same unified framework used in this paper, derive new and better selection criteria, different from other already available such proposals for the same setting, among which we can cite the AICc [30], the modified GCV [31] [32], the modified RGCV and $\mathrm{R_I GCV}$ [33] [34], RC_p, RC_p^+ and $\widehat{\mathrm{RC}_p}$ [5].

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Akaike, H. (1970) Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics*, **22**, 203-217. https://doi.org/10.1007/BF02506337

[2] Borra, S. and Di Ciaccio, A. (2010) Measuring the Prediction Error. A Comparison of Cross-Validation, Bootstrap and Covariance Penalty Methods. *Computational Statistics & Data Analysis*, **54**, 2976-2989. https://doi.org/10.1016/j.csda.2010.03.004

[3] Craven, P. and Wahba, G. (1979) Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, **31**, 377-403. https://doi.org/10.1007/BF01404567

[4] Li, K.-C. (1985) From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation. *Journal of the Japan Statistical Society*, **38**, 119-130.

[5] Rosset, S. and Tibshirani, R. (2020) From Fixed-X to Random-X Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estima-

tion. *JASA*, **115**, 138-151. https://doi.org/10.1080/01621459.2018.1424632

[6] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer-Verlag, New York.

[7] Burman, P. (1989) A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and Repeated Learning-Testing Methods. *Biometrika*, **76**, 503-514. https://doi.org/10.1093/biomet/76.3.503

[8] Hastie, T. and Tibshirani, R. (1990) Generalized Additive Models. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

[9] Eubank, R.L. (1999) Nonparametric Regression and Spline Smoothing. 2nd Edition, Marcel Dekker, New York. https://doi.org/10.1201/9781482273144

[10] McQuarrie, A.D.R. and Tsai, C.-L. (1998) Regression and Time Series Model Selection. World Scientific Publishing Co. Re. Ltd, Singapore. https://doi.org/10.1142/3573

[11] Breiman, L. and Spector, P. (1992) Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, **60**, 291-319. https://doi.org/10.2307/1403680

[12] Leeb, H. (2008) Evaluation and Selection of Models for Out-of-Sample Prediction When the Sample Size Is Small Relative to the Complexity of the Data-Generating Process. *Bernoulli*, **14**, 661-690. https://doi.org/10.3150/08-BEJ127

[13] Dicker, L.H. (2013) Optimal Equivariant Prediction for High-Dimensional Linear Models with Arbitrary Predictor Covariance. *The Electronic Journal of Statistics*, **7**, 1806-1834. https://doi.org/10.1214/13-EJS826

[14] Dobriban, E. and Wager, S. (2018) High-Dimensional Asymptotics of Prediction. *Annals of Statistics*, **46**, 247-279. https://doi.org/10.1214/17-AOS1549

[15] Lukas, M.A. (2014) Performance Criteria and Discrimination of Extreme Undersmoothing in Nonparametric Regression. *Journal of Statistical Planning and Inference*, **153**, 56-74. https://doi.org/10.1016/j.jspi.2014.05.006

[16] Lukas, M.A., de Hoog, F.R. and Anderssen, R.S. (2015) Practical Use of Robust GCV and Modified GCV for Spline Smoothing. *Computational Statistics*, **31**, 269-289. https://doi.org/10.1007/s00180-015-0577-7

[17] Kitagawa, G. (2008) Contributions of Professor Hirotogu Akaike in Statistical Science. *Journal of the Japan Statistical Society*, **38**, 119-130. https://doi.org/10.14490/jjss.38.119

[18] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *JRSS*, *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[19] Hoerl, A.E. and Kennard, R. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[20] Searle, S.R. (1971) Linear Models. John Wiley & Sons, New York.

[21] Rao, C.R. (1973) Linear Statistical Inference and Its Applications. 2nd Edition, Wiley, New York. https://doi.org/10.1002/9780470316436

[22] Rao, C.R., Toutenburg, H., *et al.* (2008) Linear Models and Generalizations: Least Squares and Alternatives. 3rd Edition, Springer-Verlag, Berlin.

[23] Olive, D.J. (2017) Linear Regression. Springer International Publishing, Berlin. https://doi.org/10.1007/978-3-319-55252-1

[24] Moore, E.H. (1920) On the Reciprocal of the General Algebraic Matrix (Abstract).

Bulletin of the AMS, **26**, 394-395.

[25] Penrose, R. (1955) A Generalized Inverse for Matrices. *Proceedings—Cambridge Philosophical Society*, **51**, 406-413. https://doi.org/10.1017/S0305004100030401

[26] Rao, C.R. (1962) A Note on a Generalized Inverse of a Matrix with Applications to Problems in Mathematical Statistics. *JRSS*, *Series B*, **24**, 152-158. https://doi.org/10.1111/j.2517-6161.1962.tb00447.x

[27] Kendrick, D.A. (2002)) Stochastic Control for Economic Models. 2nd Edition, VTEX Ltd., Vilnius.

[28] Wahba, G. (1978) Improper Priors, Spline Smoothing and the Problem of Guarding against Model Errors in Regression. *JRSS*, *Series B*, **40**, 364-372. https://doi.org/10.1111/j.2517-6161.1978.tb01050.x

[29] Härdle, W., Hall, P. and Marron, J.S. (1988) How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum-(with Discussion). *JASA*, **83**, 86-89. https://doi.org/10.2307/2288922

[30] Hurvich, C.M., Simonoff, J.S. and Tsai, C.-L. (1998) Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *JRSS*, *Series B*, **60**, 271-293. https://doi.org/10.1111/1467-9868.00125

[31] Cummins, D.J., Filloon, T.G. and Nychka, D. (2001) Confidence Intervals for Nonparametric Curve Estimates: Toward More Uniform Pointwise Coverage. *JASA*, **96**, 233-246. https://doi.org/10.1198/016214501750332811

[32] Kim, Y.-J. and Gu, C. (2004) Smoothing Spline Gaussian Regression: More Scalable Computation via Efficient Approximation. *JRSS*, *Series B*, **66**, 337-356. https://doi.org/10.1046/j.1369-7412.2003.05316.x

[33] Lukas, M.A. (2006) Robust Generalized Cross-Validation for Choosing the Regularization Parameter. *Inverse Probability*, **22**, 1883-1902. https://doi.org/10.1088/0266-5611/22/5/021

[34] Lukas, M.A. (2008) Strong Robust Generalized Cross-Validation for Choosing the Regularization Parameter. *Inverse Probability*, **24**, Article ID: 034006. https://doi.org/10.1088/0266-5611/24/3/034006