ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Čížek, Pavel

Working Paper

(Non) Linear Regression Modeling

Papers, No. 2004,11

Provided in Cooperation with:
CASE - Center for Applied Statistics and Economics, Humboldt University Berlin

Suggested Citation: Čížek, Pavel (2004) : (Non) Linear Regression Modeling, Papers, No. 2004,11, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin

This Version is available at:
https://hdl.handle.net/10419/22185

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

(Non) Linear Regression Modeling

Pavel Čížek[1]

Tilburg University, Department of Econometrics & Operations Research
Room B 516, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
P.Cizek@uvt.nl

We will study causal relationships of a known form between random variables. Given a model, we distinguish one or more dependent (endogenous) variables $\boldsymbol{Y} = (Y_1, \ldots, Y_l), l \in N$, which are explained by a model, and independent (exogenous, explanatory) variables $\boldsymbol{X} = (X_1, \ldots, X_p), p \in N$, which explain or predict the dependent variables by means of the model. Such relationships and models are commonly referred to as regression models.

A regression model describes the relationship between the dependent and independent variables. In this paper, we restrict our attention to models with a form known up to a finite number of unspecified parameters. The model can be either linear in parameters,

$$\boldsymbol{Y} = \boldsymbol{X}^\top \boldsymbol{\beta_0} + \varepsilon,$$

or nonlinear,

$$\boldsymbol{Y} = h(\boldsymbol{X}, \boldsymbol{\beta_0}) + \varepsilon,$$

where $\boldsymbol{\beta}$ represents a vector or a matrix of unknown parameters, ε is the error term (fluctuations caused by unobservable quantities), and h is a known regression function. The unknown parameters $\boldsymbol{\beta}$ are to be estimated from observed realizations $\{y_{1i}, \ldots, y_{li}\}_{i=1}^n$ and $\{x_{1i}, \ldots, x_{pi}\}_{i=1}^n$ of random variables $\boldsymbol{Y}$ and $\boldsymbol{X}$.

Here we discuss both kinds of models, primarily from the least-squares estimation point of view, in Sects. 1 and 2, respectively. Both sections present the main facts concerning the fitting of these models and relevant inference. The main focus is however on the estimation of regression models with near and exact multicollinearity, whereby we are more concerned about statistical rather than numerical side of this phenomenon.

1 Linear Regression Modeling

Let us first study the linear regression model $Y = \boldsymbol{X}^\top \boldsymbol{\beta_0} + \varepsilon$ assuming $E(\varepsilon|\boldsymbol{X}) = 0$. Unless said otherwise, we consider here only one dependent

variable Y. The unknown vector $\boldsymbol{\beta^0} = (\beta_1^0, \ldots, \beta_p^0)$ is to be estimated given observations $\mathbf{y} = (y_1, \ldots, y_n) \in R^n$ and $\{\mathbf{x_i}\}_{i=1}^n = \{(x_{1i}, \ldots, x_{pi})\}_{i=1}^n$ of random variables Y and X; let us denote $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})^\top \in R^{n \times p}$ and let $\mathbf{x_{.k}}$ be the kth column of $\mathbf{X}$. Thus, the linear regression model can be written in terms of observations as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\varepsilon} = \sum_{k=1}^{p} \mathbf{x_{.k}} \beta_k^0 + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in R^n$.

Sect. 1.1 summarizes how to estimate the model (1) by the method of least squares. Later, we specify what ill-conditioning and multicollinearity are in Sect. 1.2 and discuss methods dealing with it in Sects. 1.3–1.9.

1.1 Fitting of Linear Regression

Let us first review the least squares estimation and its main properties to facilitate easier understanding of the fitting procedures discussed further. For a detailed overview of linear regression modeling see [10].

The *least squares* (LS) approach to the estimation of (1) searches an estimate $\mathbf{b}$ of unknown parameters $\boldsymbol{\beta_0}$ by minimizing the sum of squared differences between the observed values y_i and the predicted ones $\hat{y}_i(\mathbf{b}) = \mathbf{x_i}^\top \mathbf{b}$.

Definition 1. *The least squares estimate of linear regression model (1) is defined by*

$$\mathbf{b^{LS}} = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} \{y_i - \hat{y}_i(\boldsymbol{\beta})\}^2 = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x_i}^\top \boldsymbol{\beta})^2. \tag{2}$$

This differentiable problem can be expressed as minimization of

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}$$

with respect to $\boldsymbol{\beta}$ and the corresponding first-order conditions are

$$-\mathbf{X}^\top \mathbf{y} + \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = 0 \quad \Longrightarrow \quad \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}. \tag{3}$$

They are commonly referred to as normal equations and identify the global minimum of (2) as long as the second order conditions $\mathbf{X}^\top \mathbf{X} > 0$ hold; that is, the matrix $\mathbf{X}^\top \mathbf{X}$ is supposed to be positive definite, or equivalently, nonsingular.[1] Provided that $\mathbf{X}^\top \mathbf{X} > 0$ and $E(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$, the LS estimator is unbiased and can be found as a solution of (3)

$$\mathbf{b^{LS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{4}$$

Additionally, it is the best unbiased linear estimator of (1), see [1, Thm. 1.2.1].

[1] This assumption is often specified in terms of the underlying random variable X: $E(\boldsymbol{X}\boldsymbol{X}^\top) > 0$ is positive definite.

Theorem 1. *(Gauss-Markov) Assume that* $E(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$, $E(\boldsymbol{\varepsilon}^2|\mathbf{X}) = \sigma^2\mathbf{I_n}$, *and* $\mathbf{X}^\top\mathbf{X}$ *is non-singular. Let* $\mathbf{b} = \mathbf{C}^\top\mathbf{y}$, *where* $\mathbf{C}$ *is a* $t \times p$ *matrix orthogonal to* $\mathbf{X}$, $\mathbf{C}^\top\mathbf{X} = \mathbf{I}$. *Then* $Var(\mathbf{b}) - Var(\mathbf{b^{LS}}) > 0$ *is a positive definite matrix for any* $\mathbf{b} \neq \mathbf{b^{LS}}$.

Finally, the LS estimate actually coincides with the maximum likelihood estimates provided that random errors ε are normally distributed (in addition to the assumptions of Thm. 1) and shares then the asymptotic properties of the maximum likelihood estimation (see [1, Chap. 1]).

Computing LS estimates

The LS estimate $\mathbf{b^{LS}}$ can be and often is found by directly solving the system of linear equations (3) or evaluating formula (4), which involves a matrix inversion. Both direct and iterative methods for solving systems of linear equations are presented in Chap. 4, Part II, of this Handbook. Although this straightforward computation may work well for many regression problems, it often leads to an unnecessary loss of precision, see [9, Chap. 2], and additionally, it is not very suitable if the matrix $\mathbf{X}^\top\mathbf{X}$ is ill-conditioned[2] or nearly singular (multicollinearity) because it is not numerically stable. Being concerned mainly about statistical consequences of multicollinearity, the numerical issues regarding the identification and treatment of ill-conditioned regression models are beyond the scope of this paper; let us refer an interested reader to [15, Chap. 9], [12, Chap. 3], [9, Chap. 2], and recent monograph [3].

Let us now briefly review a class of numerically more stable algorithms for the LS minimization. They are based on orthogonal transformations. Assuming a matrix $\mathbf{Q} \in R^n$ is an orthonormal matrix, $\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I_n}$,

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Qy} - \mathbf{QX}\boldsymbol{\beta})^\top(\mathbf{Qy} - \mathbf{QX}\boldsymbol{\beta}).$$

Thus, multiplying a regression model by an orthonormal matrix does not change it from the LS point of view. Since every matrix $\mathbf{X}$ can be decomposed into the product $\mathbf{Q}_x\mathbf{R}_x$ (the QR decomposition), where $\mathbf{Q}_x$ is an orthonormal matrix and $\mathbf{R}_x$ is an upper triangular matrix, pre-multiplying (1) by $\mathbf{Q}_x^\top$ produces

$$\mathbf{Q}_x^\top\mathbf{y} = \mathbf{R_x}\boldsymbol{\beta} + \mathbf{Q}_x^\top\boldsymbol{\varepsilon}, \tag{5}$$

where $\mathbf{R_x} = (\mathbf{R_1}, \mathbf{R_2})^\top$ and $\mathbf{R_1} \in R^{p\times p}$ is an upper triangular matrix and $\mathbf{R_2} \in R^{(n-p)\times p}$ is a zero matrix. Hence, the sum of squares to minimize can be written as

$$(\mathbf{Q}_x^\top\mathbf{y} - \mathbf{R_x}\boldsymbol{\beta})^\top(\mathbf{Q}_x^\top\mathbf{y} - \mathbf{R_x}\boldsymbol{\beta}) = (\mathbf{y_1} - \mathbf{R_1}\boldsymbol{\beta})^\top(\mathbf{y_1} - \mathbf{R_1}\boldsymbol{\beta}) + \mathbf{y_2}^\top\mathbf{y_2},$$

where $\mathbf{y_1} \in R^p$ and $\mathbf{y_2} \in R^{n-p}$ form $\mathbf{Q}_x^\top\mathbf{y} = (\mathbf{y_1}^\top, \mathbf{y_2}^\top)^\top$. The LS estimate is then obtained from the upper triangular system $\mathbf{R_1}\boldsymbol{\beta} = \mathbf{y_1}$, which is trivial to

[2] A regression problem is called ill-conditioned if a small change in data causes large changes in estimates.

solve by backward substitution. There are many algorithms for constructing a suitable QR decomposition for finding LS estimates, such as the Householder or Givens transformations; see Chap. 4, Part II, of this Handbook, or [8, Chaps. 7–8], [12, Chap. 3], [15, Chap. 9], and [3, Chaps. 1 and 2].

LS inference

Linear regression modeling does not naturally consist only of obtaining a point estimate $\mathbf{b^{LS}}$. One needs to measure the variance of the estimates in order to construct confidence intervals or test hypotheses. Additionally, one should assess the quality of the regression fit. Most such measures are based on the regression residuals $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$. We briefly review the most important regression statistics, and next, indicate how it is possible to compute them if the LS regression is estimated by means of the orthogonalization procedure described in the previous paragraph.

The most important measures used in statistics to assess model fit and inference are the total sum of squares

$$TSS = (\mathbf{y} - \bar{y})^\top (\mathbf{y} - \bar{\mathbf{y}}) = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$, the residual sum of squares

$$RSS = \mathbf{e}^\top \mathbf{e} = \sum_{i=1}^{n} e_i^2,$$

and the complementary regression sum of squares

$$RegSS = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = TSS - RSS.$$

Using these quantities, the regression fit can be evaluated, for example, the coefficient of determination $R^2 = 1 - RSS/TSS$ as well as many information criteria (modified $\bar{R}^2$, Mallows and Akaike criteria, etc.). Additionally, it can be used to compute the variance of the estimates in simple cases. The variance of the estimates can be estimated by

$$Var(\mathbf{b^{LS}}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{S}^{-1} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}, \tag{6}$$

where $\mathbf{S}$ represents an estimate of the covariance matrix $Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. Provided that the model is homoscedastic, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I_n}$, the residual variance σ^2 can be estimated as an average of squared residuals $s^2 = \mathbf{e}^\top \mathbf{e}/n$. Apart from the residual variance, one needs also an inverse of $(\mathbf{X}^\top \mathbf{X})^{-1}$, which will often be a by-product of solving normal equations.

Let us now describe how one computes these quantities if a numerically stable procedure based on the orthonormalization of normal equations is used.

Let us assume we already constructed a QR decomposition of $\mathbf{X} = \mathbf{Q_x R_x}$. Thus, $\mathbf{Q_x Q_x^\top} = \mathbf{I}$ and $\mathbf{Q_x^\top X} = \mathbf{R_x}$. RSS can be computed as

$$\begin{aligned} RSS = \mathbf{e}^\top\mathbf{e} &= (\mathbf{y} - \mathbf{Xb})^\top(\mathbf{y} - \mathbf{Xb}) = (\mathbf{y} - \mathbf{Xb})^\top\mathbf{Q_x Q_x^\top}(\mathbf{y} - \mathbf{Xb}) \\ &= (\mathbf{Q_x^\top y} - \mathbf{R_x Xb})^\top(\mathbf{Q_x^\top y} - \mathbf{R_x Xb}). \end{aligned}$$

Consequently, RSS is invariant with respect to orthonormal transformations (5) of the regression model (1). The same conclusion applies also to TSS and $RegSS$, and consequently, to the variance estimation. Thus, it is possible to use the data in (5), transformed to achieve better numerical stability, for computing regression statistics of the original model (1).

1.2 Multicollinearity

Let us assume that the design matrix $\mathbf{X}$ fixed. We talk about *multicollinearity* when there is a linear dependence among the variables in regression, that is, the columns of $\mathbf{X}$.

Definition 2. *In model (1), the exact multicollinearity exists if there are real constants* $a_1, \ldots, a_p$ *such that* $\sum_{k=1}^{p} |a_k| > 0$ *and* $\sum_{k=1}^{p} a_k \mathbf{x_{.k}} = 0$.

The exact multicollinearity (also referred to as reduced-rank data) is relatively rare in linear regression models unless the number of explanatory variables is very large or even larger than the number of observations, $p \geq n$.[3] When the number p of variables is small compared to the sample size n, near multicollinearity is more likely to occur: there are some real constants $a_1, \ldots, a_p$ such that $\sum_{k=1}^{p} |a_k| > 0$ and $\sum_{k=1}^{p} a_k \mathbf{x_{.k}} \approx 0$, where $\approx$ denotes approximate equality. The multicollinearity in data does not have to arise only as a result of highly correlated variables (e.g., more measurements of the same characteristic by different sensors or methods), which by definition occurs in all applications where there are more variables than observations, but it could also result from the lack of information and variability in data.

Whereas the exact multicollinearity implies that $\mathbf{X}^\top\mathbf{X}$ is singular and the LS estimator is not identified, the near multicollinearity permits non-singular matrix $\mathbf{X}^\top\mathbf{X}$. The eigenvalues $\lambda_1 \leq \ldots \leq \lambda_p$ of matrix $\mathbf{X}^\top\mathbf{X}$ can give some indication concerning multicollinearity: if the smallest eigenvalue λ_1 equals zero, the matrix is singular and data are exactly multicollinear; if

[3] This happens often in agriculture, chemometrics, sociology, etc. For example, [9] uses data on the absorbances of infra-red rays of many different wavelength by chopped meat, whereby the aim is to determine the moisture, fat, and protein content of the meat as a function of these absorbances. The study employs measurements at 100 wavelengths from 850 nm to 1050 nm, which gives rise to many possibly correlated variables.

λ_1 is close to zero, near multicollinearity is present in data.[4] Since measures based on eigenvalues depend on the parametrization of the model, they are not necessarily optimal and it is often easier to detect multicollinearity by looking at LS estimates and their behavior as discussed in next paragraph. See [3] and [18] for more details on detection and treatment of ill-conditioned problems.

The multicollinearity has important implications for LS. In the case of exact multicollinearity, matrix $\mathbf{X}^\top\mathbf{X}$ does not have a full rank, hence the solution of the normal equations is not unique and the LS estimate $\mathbf{b^{LS}}$ is not identified. One has to introduce additional restrictions to identify the LS estimate. On the other hand, even though near multicollinearity does not prevent the identification of LS, it negatively influences estimation results. Since both the estimate $\mathbf{b^{LS}}$ and its variance are proportional to the inverse of $\mathbf{X}^\top\mathbf{X}$, which is nearly singular under multicollinearity, near multicollinearity inflates $\mathbf{b^{LS}}$, which may become unrealistically large, and variance $Var(\mathbf{b^{LS}})$. Consequently, the corresponding t-statistics are typically very low. Moreover, due to the large values of $(\mathbf{X}^\top\mathbf{X})^{-1}$, the least squares estimate $\mathbf{b^{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ reacts very sensitively to small changes in data. See [13] for a more detailed treatment and real-data examples of the effects of multicollinearity.

There are several strategies to limit adverse consequences of multicollinearity provided that one cannot improve the design of a model or experiment or get better data. First, one can impose an additional structure on the model. This strategy cannot be discussed in details since it is model specific, and in principle, it requires only to test a hypothesis concerning additional restrictions. Second, it is possible to reduce the dimension of the space spanned by $\mathbf{X}$, for example, by excluding some variables from the regression (Sects. 1.3 and 1.4). Third, one can also leave the class of unbiased estimators and try to find a biased estimator with smaller variance and mean squared error. Assuming we want to judge the performance of an estimator $\mathbf{b}$ by its mean squared error (MSE), the motivation follows from

$$\begin{aligned} MSE(\mathbf{b}) &= E[(\mathbf{b}-\boldsymbol{\beta^0})(\mathbf{b}-\boldsymbol{\beta^0})^\top] \\ &= E[\{\mathbf{b}-E(\mathbf{b})\}\{\mathbf{b}-E(\mathbf{b})\}^\top] + [E\{E(\mathbf{b})-\boldsymbol{\beta^0}\}][E\{E(\mathbf{b})-\boldsymbol{\beta^0}\}]^\top \\ &= Var(\mathbf{b}) + Bias(\mathbf{b})Bias(\mathbf{b})^\top. \end{aligned}$$

Thus, it is possible that introducing a bias into estimation in such a way that the variance of estimates is significantly reduced can improve the estimator's MSE. There are many biased alternatives to the LS estimation as discussed in Sects. 1.5–1.9 and some of them even combine biased estimation with variable selection. In all cases, we present methods usable both in the case of near and exact multicollinearity.

[4] Nearly singular matrices are dealt with also in numerical mathematics. To measure near singularity, numerical mathematics uses the conditioning numbers $d_k = \sqrt{\lambda_k/\lambda_1}$, which converge to infinity for singular matrices (as λ_1 approaches zero). Matrices with very large conditioning numbers are called ill-conditioned.

1.3 Variable Selection

The presence of multicollinearity may indicate that some explanatory variables are linear combinations of the other ones,[5] and consequently, do not improve explanatory power of a model. Hence, they could be dropped from the model provided there is some justification for dropping them also on the model level rather than just dropping them to fix data problems. As a result of removing some variables, the matrix $\mathbf{X}^\top\mathbf{X}$ would not be (nearly) singular anymore.

Eliminating variables from a model is a special case of model selection procedures, which are discussed in details in Chap. 1, Part III, of this Handbook. An overview and comparison of many classical variable selection is given, for example, in [9] and [69]. For discussion of computational issues related to model selection, see [9] and [8, Chap. 8]. Here we briefly discuss methods specific for variable selection within a single regression model and some more general model selection methods that are useful both in the context of variable selection and of biased estimation discussed in Sects. 1.5–1.9.

Backward elimination

A simple and often used method to eliminate non-significant variables from regression is *backward elimination*, a special case of stepwise regression. Backward elimination starts from the full model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ and identifies variable $\mathbf{x_{.k}}$ such that

1. its omission results in smallest increase of RSS; or
2. it has the smallest t-statistics $t_k = b_k^{LS}/\sqrt{s_k^2/(n-p)}$, where s_k^2 is an estimate of b_k^{LS} variance, or any other test statistics of $H_0 : \beta_{0k} = 0$; or
3. its removal causes smallest change of prediction or information criteria characterizing fit or prediction power of the model. A well-known examples of information criteria are modified coefficient of determination $\bar{R}^2 = 1 - (n+p)\mathbf{e}^\top\mathbf{e}/n(n-p)$, Akaike information criterion [21], $AIC = \log(\mathbf{e}^\top\mathbf{e}/n) + 2p/n$, and Schwarz information criterion [77], $SIC = \log(\mathbf{e}^\top\mathbf{e}/n) + p\ln n/n$, where n and p represents sample size and the number of regressors, respectively.

Next, the variable $\mathbf{x_{.k}}$ is excluded from regression by setting $b_k = 0$ if (i) one did not reach a pre-specified number of variables yet or (ii) the test statistics or change of the information criterion lies below some selected significance level.

Although backward elimination, which can be also viewed as a pre-test estimator (see [17]), is often used in practice, it involves largely arbitrary choice of the significance level. In addition, it has rather poor statistical properties caused primarily by discontinuity of the selection decision, see [64]. Moreover,

[5] This is more often a "feature" of data rather than of the model.

even if a stepwise procedure is employed, one should take care of reporting correct variances and confidence intervals valid for the whole decision sequence. Inference for the finally selected model as if it were the only model considered leads to significant biases, see [97], [91], and [35]. Backward elimination also does not perform well in the presence of multicollinearity and it cannot be used if $p > n$. Finally, let us note that a nearly optimal and admissible alternative is proposed in [65].

Forward selection

Backward elimination cannot be applied if there are more variables than observations, and additionally, it may be very computationally expensive if there are many variables. A classical alternative is *forward selection*, where one starts from an intercept-only model and adds one after another variables that provide the largest decrease of RSS. Adding stops when the F-statistics

$$R = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}}(n - p - 2)$$

lies below a pre-specified critical 'F-to-enter' value. The forward selection can be combined with the backward selection (e.g., after adding a variable, one performs one step of backward elimination), which is known as a stepwise regression [16]. Its computational complexity is discussed in [9].

Note that most disadvantages of backward elimination apply to forward selection as well. In particular, correct variances and confidence intervals should be reported, see [9, Sect. 3.3] on their approximations. Moreover, forward selection can be overly aggressive in selection in the respect that if a variable $\mathbf{x}$ is already included in a model, forward selection primarily adds variables orthogonal to $\mathbf{x}$, thus ignoring possibly useful variables that are correlated with $\mathbf{x}$. To improve upon this, [40] proposed least angle regression, considering correlations of to-be-added variables jointly with respect to all variables already included in the model.

All-subsets regression

Neither forward selection, nor backward elimination guarantee the optimality of the selected submodel, even when both methods lead to the same results.[6] An alternative approach – *all-subsets regression* – is based on forming a model for each subset of explanatory variables. Each model is estimated and a selected prediction or information criterion, which quantifies the unexplained variation of the dependent variable and the parsimony of the model, is evaluated. Finally, the model attaining the best value of a criterion is selected and variables missing in this model are omitted.

[6] This happens especially when a pair of variables has jointly high predictive power; for example, if the dependent variable $\mathbf{y}$ depends on the difference of two variables $\mathbf{x_1} - \mathbf{x_2}$.

This approach deserves several comments. First, one can use many other criteria instead of AIC or SIC. These could be based on the test statistics of a joint hypothesis that a group of variables has zero coefficients, extensions or modifications of AIC or SIC, general Bayesian predictive criteria, criteria using non-sample information, model selection based on estimated parameter values at each subsample and so on. See the next subsection, [25], [79], [52], [54], [98], [55], for instance, and Chap. 1, Part III, of this Handbook for more detailed overview.

Second, the evaluation and estimation of all submodels of a given regression model can be very computationally intensive, especially if the number of variables is large. This motivated tree-like algorithms searching through all submodels, but once they reject a submodel, they automatically reject all models containing only a subset of variables of the rejected submodel, see [39]. These so-called branch-and-bound techniques are discussed in [9], for instance.

An alternative computational approach, which is increasingly used in applications where the number of explanatory variables is very large, is based on the *genetic programming* (genetic algorithm, GA) approach, see [89]. GAs searches through the space of all submodels which are represented by "chromosomes" – a $p \times 1$ vectors $\mathbf{m_j} \in \{0,1\}^p$ of indicators marking whether a variable is included in a submodel. To choose the best submodel, one has a population $\mathcal{P} = \{\mathbf{m_j}\}_{j=1}^{J}$ of submodels that compete with each other by means of information or prediction criteria. The submodels in the population can combine their characteristics (chromosomes), sometimes additionally affected by a random mutation, to create their offsprings $\mathbf{m_j^*}$. Whenever an offspring $\mathbf{m_j^*}$ performs better than the original model $\mathbf{m_j}$ (parent), $\mathbf{m_j^*}$ replaces $\mathbf{m_j}$ in population $\mathcal{P}$. Repeating this process searches among all submodels and provides a rather effective way of obtaining the best submodel, especially when the number of explanatory variables is very high. See Chap. 6, Part II, of this handbook and [4] for introduction to genetic programming.

Cross validation

Cross validation (CV) is a general model-selection principle, proposed already in [81], which chooses a specific model in a similar way as the prediction criteria. CV compares models, which can include all variables or exclude some, based on their out-of-sample performance, which is measured typically by MSE. To achieve this, a sample is split to two disjunct parts: one part is used for estimation and the other part for checking the fit of the estimated model on "new" data, which were not used for estimation, by comparing the observed and predicted values.

Probably the most popular variant is the leave-one-out cross-validation (LOU CV), which can be used not only for model selection, but also for choosing nuisance parameters (e.g., in nonparametric regression, see [6]). Assume we have a set of models $\mathbf{y} = h_k(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}$ defined by regression functions

$h_k, k = 1, \ldots, M$, that determine variables included or excluded from regression. For model given by h_k, LOU CV evaluates

$$CV_k = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i,-i}\right)^2, \tag{7}$$

where $\hat{y}_{i,-i}$ is the prediction at $\mathbf{x_i}$ based on the model $\mathbf{y_{-i}} = h_k(\mathbf{X_{-i}}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon_{-i}}$ and $\mathbf{y_{-i}}, \mathbf{X_{-i}}, \boldsymbol{\varepsilon_{-i}}$ are the vectors and matrices $\mathbf{y}, \mathbf{X}, \boldsymbol{\varepsilon}$ without their ith elements and rows, respectively. Thus, all but the ith observation are used for estimation and the ith observation is used to check the out-of-sample prediction. Having evaluated CV_k for each model, $k = 1, \ldots, M$, we select the model commanding the minimum $\min_{k=1,\ldots,M} CV_k$.

Unfortunately, LOU CV is not consistent as far as the linear model selection is concerned. To make CV a consistent model selection method, it is necessary to omit n_v observations from the sample used for estimation, where $\lim_{n\to\infty} n_v/n = 1$. This fundamental result derived in [78] places a heavy computational burden on the CV model selection. Since our main use of CV in this paper concerns nuisance parameter selection, we do not discuss this type of CV any further. See [9, Chap. 5] and Chap. 1, Part III, of this Handbook for further details.

Example 1. We compare several mentioned variable selection methods using a classical data set on air pollution used originally by [68], who modeled mortality depending on 15 explanatory variables ranging from climate and air pollution to socioeconomic characteristics and who additionally demonstrated instabilities of LS estimates using this data set. We refer to the explanatory variables of data Pollution simply by numbers 1 to 15.

Table 1. Variables selected from Pollution data by different selection procedures. RSS is in brackets.

Number of variables	Forward selection	Backward elimination	All-subset selection
1	9 (133695)	9 (133695)	9 (133695)
2	6, 9 (99841)	6, 9 (99841)	6, 9 (99841)
3	2, 6, 9 (82389)	2, 6, 9 (82389)	2, 6, 9 (82389)
4	2, 6, 9, 14 (72250)	2, 5, 6, 9 (74666)	1, 2, 9, 14 (69154)
5	1, 2, 6, 9, 14 (64634)	2, 6, 9, 12, 13 (69135)	1, 2, 6, 9, 14 (64634)

We applied the forward, backward, and all-subset selection procedures to this data set. The results reported in Table 1 demonstrate that although all

three methods could lead to the same subset of variables (e.g, if we search a model consisting of two or three variables), this is not the case in general. For example, searching for a subset of four variables, the variables selected by backward and forward selection differ, and in both cases, the selected model is suboptimal (compared to all-subsets regression) in the sense of the unexplained variance measured by RSS.

1.4 Principle Components Regression

In some situations, it is not feasible to use variable selection to reduce the number of explanatory variables or it is not desirable to do so.[7] Since such data typically exhibit (exact) multicollinearity and we do not want to exclude some or even majority of variables, we have to reduce the dimension of the data in another way.

A general method that can be used both under near and exact multicollinearity is based on the *principle components analysis* (PCA), see Chap. 6, Part III, of this Handbook. Its aim is to reduce the dimension of explanatory variables by finding a small number of linear combinations of explanatory variables $\mathbf{X}$ that capture most of the variation in $\mathbf{X}$ and to use these linear combinations as new explanatory variables instead the original one. Suppose that $\mathbf{G}$ is an orthonormal matrix that diagonalizes matrix $\mathbf{X}^\top\mathbf{X}$: $\mathbf{G}^\top\mathbf{G} = \mathbf{I}$, $\mathbf{X}^\top\mathbf{X} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}^\top$, and $\mathbf{G}^\top\mathbf{X}^\top\mathbf{X}\mathbf{G} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ is a diagonal matrix of eigenvalues of $\mathbf{X}^\top\mathbf{X}$.

Definition 3. *Assume without loss of generality that* $\lambda_1 \geq \ldots \geq \lambda_p$ *and* $\mathbf{g_1}, \ldots, \mathbf{g_p}$ *are the corresponding eigenvectors (columns of matrix* $\mathbf{G}$*). Vector* $\mathbf{z_i} = \mathbf{X}\mathbf{g_i}$ *for* $i = 1, \ldots, p$ *such that* $\lambda_i > 0$ *is called the* i*th principle component (PC) of* $\mathbf{X}$ *and* $\mathbf{g_i}$ *represents the corresponding loadings.*

PCA tries to approximate the original matrix $\mathbf{X}$ by projecting it into the lower-dimensional space spanned by the first k eigenvectors $\mathbf{g_1}, \ldots, \mathbf{g_k}$. It can be shown that these projections capture most of the variability in $\mathbf{X}$ among all linear combinations of columns of $\mathbf{X}$, see [7, Thms. 9.1–9.3].

Theorem 2. *There is no standardized linear combination* $\mathbf{X}\mathbf{a}$*, where* $\|a\| = 1$*, that has strictly larger variance than* $\mathbf{z_1} = \mathbf{X}\mathbf{g_1}$*:* $Var(\mathbf{X}\mathbf{a}) \leq Var(\mathbf{z_1}) = \lambda_1$*. Additionally, the variance of the linear combination* $\mathbf{z} = \mathbf{X}\mathbf{a}$*,* $\|a\| = 1$*, that is uncorrelated with the first* k *principle components* $\mathbf{z_1}, \ldots, \mathbf{z_k}$ *is maximized by the* $(k+1)$*-st principle component* $\mathbf{z} = \mathbf{z_{k+1}}$ *and* $\mathbf{a} = \mathbf{g_{k+1}}$*,* $k = 1, \ldots, p-1$*.*

[7] The first case can occur if the number of explanatory variables is large compared to the number of observations. The latter case is typical in situations when we observe many characteristics of the same type, for example, temperature or electroimpulse measurements from different sensors on a human body. They could be possibly correlated with each other and there is no a priori reason why measurements at some points of a skull, for instance, should be significant while other ones would not be important at all.

Consequently, one chooses a number k of PCs that capture a sufficient amount of data variability. This can be done by looking at the ratio $L_k = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$, which quantifies the fraction of the variance captured by the first k PCs compared to the total variance of $\mathbf{X}$.

In the regression context, the chosen PCs are used as new explanatory variables, and consequently, PCs with small eigenvalues can be important too (see [56]). Therefore, one can alternatively choose the PCs that exhibit highest correlation with the dependent variable $\mathbf{y}$ because the aim is to use the selected PCs for regressing the dependent variable $\mathbf{y}$ on them, see [56]. Moreover, for selecting "explanatory" PCs, it is also possible to use any variable selection method discussed in Sect.1.3. Recently, [53] proposed a new data-driven PC selection for PCR obtained by minimizing MSE.

Next, let us assume we selected a small number k of PCs $\mathbf{Z_k} = (\mathbf{z_1}, \ldots, \mathbf{z_k})^\top$ by some rule such that matrix $\mathbf{Z_k^\top Z_k}$ has a full rank, $k \leq p$. Then the *principle components regression* (PCR) is performed by regressing the dependent variable $\mathbf{y}$ on the selected principle components $\mathbf{Z_k}$, which have a (much) smaller dimension than original data $\mathbf{X}$, and consequently, multicollinearity is diminished or eliminated, see [5, Chap. 8]. We estimate this new model by LS,

$$\mathbf{y} = \mathbf{Z_k}\boldsymbol{\gamma} + \eta = \mathbf{X}\mathbf{G_k}\boldsymbol{\gamma} + \eta,$$

where $\mathbf{G_k} = (\mathbf{g_1}, \ldots, \mathbf{g_k})^\top$. Comparing it with the original model (1) shows that $\boldsymbol{\beta} = \mathbf{G_k}\boldsymbol{\gamma}$. It is important to realize that in PCR we first fix $\mathbf{G_k}$ by means of PCA and then estimate $\boldsymbol{\gamma}$.

Finally, concerning different PC selection criteria, [24] demonstrates the superiority of the correlation-based PCR (CPCR) and convergence of many model-selection procedures toward the CPCR results. See also [37] for a similar comparison of CPRC and PCR based on GA variable selection.

Example 2. Let us use data Pollution to demonstrate several important issues concerning PCR. First, we identify PCs of the data. The fraction of variance explained by the first k PCs as a function of k is depicted on Fig. 1 (dashed line). On the one side, almost all of the $\mathbf{X}$ variance is captured by the first PC. On the other side, the percentage of the $\mathbf{y}$ variance explained by the first k PCs (solid line) grows and reaches its maximum relatively slowly. Thus, the inclusion of about 7 PCs seems to be necessary when using this strategy.

On the other hand, using some variable selection method or checking the correlation of PCs with the dependent variable $\mathbf{y}$ reveals that PCs 1, 3, 4, 5, 7 exhibit highest correlations with $\mathbf{y}$ (higher than 0.25), and naturally, a model using these 5 PCs has more explanatory power ($\bar{R}^2 = 0.70$) than for example the first 6 PCs together ($\bar{R}^2 = 0.65$). Thus, considering not only PCs that capture most of the $\mathbf{X}$ variability, but also those having large correlations with the dependent variable enables building more parsimonious models.

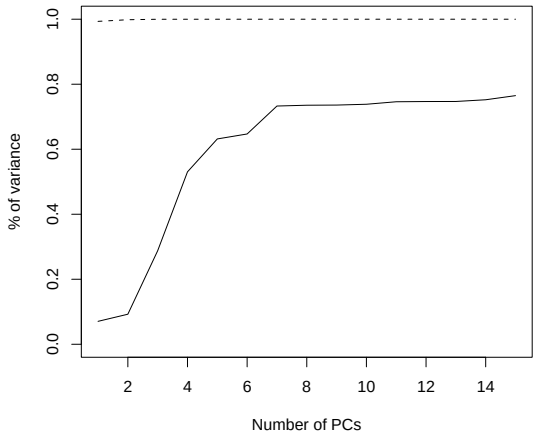

Fig. 1. Fraction of the explained variance of $\mathbf{X}$ (dashed line) and $\mathbf{y}$ (solid line) by the first k PCs.

1.5 Shrinkage Estimators

We argued in Sect. 1.2 that an alternative way of dealing with unpleasant consequences of multicollinearity lies in biased estimation: we can sacrifice small bias for a significant reduction in variance of an estimator so that its MSE decreases. Since it holds for an estimator b and a real constant $c \in R$ that $Var(c\mathbf{b}) = c^2 Var(\mathbf{b})$, a bias of the estimator $\mathbf{b}$ towards zero, $|c| < 1$, naturally leads to a reduction in variance. This observation motivates a whole class of biased estimators – *shrinkage estimators* – that are biased towards zero in all or just some of their components. In other words, they "shrink" the Euclidean norm of estimates compared to that of the corresponding unbiased estimate. This is perhaps easiest to observe on the example of the Stein-rule estimator, which can be expressed in linear regression model (1) as

$$\mathbf{b^{SR}} = \left(1 - \frac{k\mathbf{e}^\top \mathbf{e}}{n\mathbf{b^{LS}}^\top \mathbf{X}^\top \mathbf{X}\mathbf{b^{LS}}}\right)\mathbf{b^{LS}}, \tag{8}$$

where $k > 0$ is an arbitrary scalar constant and $\mathbf{e}^\top\mathbf{e}/n$ represents an estimate of the residual variance ([13, Chap. 6]). Apparently, the Stein-rule estimator just multiplies the LS estimator by a constant smaller than one. See [17] for an overview of this and many other biased estimators.

In the following subsections, we discuss various shrinkage estimators that perform well under multicollinearity and that can possibly act as variable selection tools as well: the ridge regression estimator and its modifications (Sect. 1.6), continuum regression (Sect. 1.7), the Lasso estimator and its variants (Sect. 1.8), and partial least squares (Sect. 1.9). Let us note that there

are also other shrinkage estimators, which either do not perform well under various forms of multicollinearity (e.g., Stein-rule estimator) or are discussed in other parts of this chapter (e.g., pre-test and PCR estimators in Sects. 1.3 and 1.4, respectively).

1.6 Ridge Regression

Probably the best known shrinkage estimator is the *ridge estimator* proposed and studied by [50]. Having a non-orthogonal or even nearly singular matrix $\mathbf{X}^\top\mathbf{X}$, one can add a positive constant $k > 0$ to its diagonal to improve conditioning.

Definition 4. *Ridge regression (RR) estimator is defined for model (1) by*

$$\mathbf{b^{RR}} = (\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \tag{9}$$

for some ridge parameter $k > 0$.

"Increasing" the diagonal of $\mathbf{X}^\top\mathbf{X}$ before inversion shrinks $\mathbf{b^{RR}}$ compared to $\mathbf{b^{LS}}$ and introduces a bias. Additionally, [50, Thm. 4.3] also showed that the derivative of $MSE(\mathbf{b^{RR}})$ with respect to k is negative at $k = 0$. This indicates that the bias

$$Bias(\mathbf{b^{RR}}) = -k(\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}\boldsymbol{\beta}$$

can be smaller than the decrease in variance (here for a homoscedastic linear model with error variance σ^2)

$$Var(\mathbf{b^{RR}}) - Var(\mathbf{b^{LS}}) = \sigma^2(\mathbf{X}^\top\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+k\mathbf{I})^{-1} - \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

caused by shrinking at least for some values of k. The intervals for k where RR dominates LS are derived, for example, in [33], [13, Chap. 7], and [10, Sect. 3.10.2]. Moreover, the improvement in $MSE(\mathbf{b^{RR}})$ with respect to $MSE(\mathbf{b^{LS}})$ is significant under multicollinearity while being negligible for nearly orthogonal systems. A classical result for model (1) under $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I_n})$ states that $MSE(\mathbf{b^{RR}}) - MSE(\mathbf{b^{LS}}) < 0$ is negative definite if $k < k_{max} = 2\sigma^2/\boldsymbol{\beta}^\top\boldsymbol{\beta}$, see [13, Sect. 7.2], where an operational estimate of k_{max} is discussed too. Notice however that the conditions for the dominance of the RR and other some other shrinkage estimators over LS can look quite differently in the case of non-normal errors [86].

In applications, an important question remains: how to choose the ridge parameter k? In the original paper [50], the use of the ridge trace, a plot the components of the estimated $\mathbf{b^{RR}}$ against k, was advocated. If data exhibit multicollinearity, one usually observes a region of instability for k close to zero and then stable estimates for large values of ridge parameter k. One should choose the smallest k lying in the region of stable estimates. Alternatively, one could search for k minimizing $MSE(\mathbf{b^{RR}})$; see the subsection on generalized RR for more details. Furthermore, many other methods for model selection

could be employed too; for example, LOU CV (Sect. 1.3) performed on a grid of k values is often used in this context.

Statistics important for inference based on RR estimates are discussed in [50] and [13] both for the case of a fixed k as well as in the case of some data-driven choices. Moreover, [13] describes algorithms for a fast and efficient RR computation.

To conclude, let us note that the RR estimator $\mathbf{b^{RR}}$ in model (1) can be also defined as a solution of a restricted minimization problem

$$\mathbf{b^{RR}} = \underset{\mathbf{b}:\|\mathbf{b}\|_2^2 \leq r^2}{\operatorname{argmin}} \ (\mathbf{y} - \mathbf{X}\mathbf{b})^\top (\mathbf{y} - \mathbf{X}\mathbf{b}), \tag{10}$$

or equivalently as

$$\mathbf{b^{RR}} = \underset{\mathbf{b}}{\operatorname{argmin}} \ (\mathbf{y} - \mathbf{X}\mathbf{b})^\top (\mathbf{y} - \mathbf{X}\mathbf{b}) + k\|\mathbf{b}\|_2^2, \tag{11}$$

where r represents a tuning parameter corresponding to k (see [83]). This approach was used by [70], for instance. Moreover, formulation (10) reveals one controversial issue in RR: re-scaling the original data to make $\mathbf{X}^\top\mathbf{X}$ a correlation matrix. Although there are no requirements of this kind necessary for theoretical results, standardization is often recommended to make influence of the constraint $\|\mathbf{b}\|_2^2 \leq r^2$ same for all variables. There are also studies showing adverse effects of this standardization on estimation, see [13] for a discussion. A possible solution is generalized RR, which assigns to each variable its own ridge parameter (see the next paragraph).

Generalized ridge regression

The RR estimator can be generalized in the sense that each diagonal element of $\mathbf{X}^\top\mathbf{X}$ is modified separately. To achieve that let us recall that this matrix can be diagonalized: $\mathbf{X}^\top\mathbf{X} = \mathbf{G}^\top\mathbf{\Lambda}\mathbf{G}$, where $\mathbf{G}$ is an orthonormal matrix and $\mathbf{\Lambda}$ is a diagonal matrix containing eigenvalues $\lambda_1, \ldots, \lambda_p$.

Definition 5. *Generalized ridge regression (GRR) estimator is defined for model (1) by*

$$\mathbf{b^{GRR}} = (\mathbf{X}^\top\mathbf{X} + \mathbf{G}\mathbf{K}\mathbf{G}^\top)^{-1}\mathbf{X}^\top\mathbf{y} \tag{12}$$

for a diagonal matrix $\mathbf{K} = \operatorname{diag}(k_1, \ldots, k_p)$ *of ridge parameters.*

The main advantage of this generalization being ridge coefficients specific to each variable, it is important to know how to choose the matrix $\mathbf{K}$. In [50] and [13], the following result is derived.

Theorem 3. *Assume that* $\mathbf{X}$ *in model (1) has a full rank,* $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I_n})$*, and* $n > p$*. Further, let* $\mathbf{X} = \mathbf{H}\mathbf{\Lambda}^{1/2}\mathbf{G}^\top$ *be the singular value decomposition of* $\mathbf{X}$ *and* $\boldsymbol{\gamma} = \mathbf{G}^\top\boldsymbol{\beta_0}$*. The MSE-minimizing choice of* $\mathbf{K}$ *in (12) is* $\mathbf{K} = \sigma^2 \operatorname{diag}(\gamma_1^{-2}, \ldots, \gamma_p^{-2})$.

An operational version (feasible GRR) is based on an unbiased estimate $\hat{\gamma}_i = \mathbf{G}^\top \mathbf{b}^{\mathbf{LS}}$ and $s^2 = (\mathbf{y} - \mathbf{H}\hat{\gamma})^\top(\mathbf{y} - \mathbf{H}\hat{\gamma})$. See [50] and [13], where you also find the bias and MSE of this operational GRR estimator, and [88] for further extensions of this approach. Let us note that the feasible GRR (FGRR) estimator does not have to possess the MSE-optimality property of GRR because the optimal choice of $\mathbf{K}$ is replaced by an estimate. Nevertheless, the optimality property of FGRR is preserved if $\lambda_i \gamma_i^2 \leq 2\sigma^2$, where λ_i is the (i,i)th-element of $\mathbf{\Lambda}$, see [42].

Additionally, given an estimate of MSE-minimizing $\hat{\mathbf{K}} = \mathrm{diag}(\hat{k}_1, \ldots, \hat{k}_p)$, many authors proposed to choose the ridge parameter k in ordinary RR as a harmonic mean of $\hat{k}_i, i = 1, \ldots, p$; see [51], for instance.

Almost unbiased ridge regression

Motivated by results of [13], [59] proposed to correct GRR for its bias using the first-order bias approximation. This yields almost unbiased GRR (AUGRR) estimator

$$\mathbf{b}^{\mathbf{AUGRR}} = (\mathbf{X}^\top\mathbf{X} + \mathbf{GKG}^\top)^{-1}(\mathbf{X}^\top\mathbf{y} + \mathbf{KG}^\top\boldsymbol{\beta_0}).$$

The true parameter value $\boldsymbol{\beta_0}$ being unknown, [71] defined a feasible AUFGRR estimator by replacing the unknown $\boldsymbol{\beta_0}$ by $\mathbf{b}^{\mathbf{FGRR}}$ and $\mathbf{K}$ by the employed ridge matrix. Additionally, a comparison of the FGRR and feasible AUGRR estimators with respect to MSE proved that FGRR has a smaller MSE than AUGRR in a wide range of parameter space. Similar observation was also done under a more general loss function in [87]. Furthermore, [22] derived exact formulas for the moments of the feasible AUGRR estimator.

Further extensions

RR can be applied also under exact multicollinearity, which arises for example in data with more variables than observations. Although the theory and application of RR is the same as in the case of full-rank data, [13], the computational burden $O(np^2 + p^3)$ becomes too high for $p > n$. A faster algorithm with computational complexity only $O(np^2)$ was found by [47].

Example 3. Using data Pollution once again, we estimated RR for ridge parameter $k \in (0, 10)$ and plotted the estimated coefficients $\mathbf{b}^{\mathbf{RR}}$ as functions of k (ridge trace plot), see Fig. 2. For the sake of simplicity, we restricted ourselves only to variables that were selected by some variable selection procedure in Table 1 (1, 2, 6, 9, 12, 13, 14). The plot shows the effect of ridge parameter k on slope estimates ($k = 0$ corresponds to LS). Apparently, slopes of some variables are affected very little (e.g., variable 1), some significantly (e.g., the magnitude of variable 14 increases more than twice), and some variables shrink extremely (e.g., variables 12 and 13). In all cases, the biggest change occurs between $k = 0$ and $k = 2$, after which then estimates stabilize. The vertical dashed line in Fig. 2 represents the CV estimate of k ($k_{CV} = 6.87$).

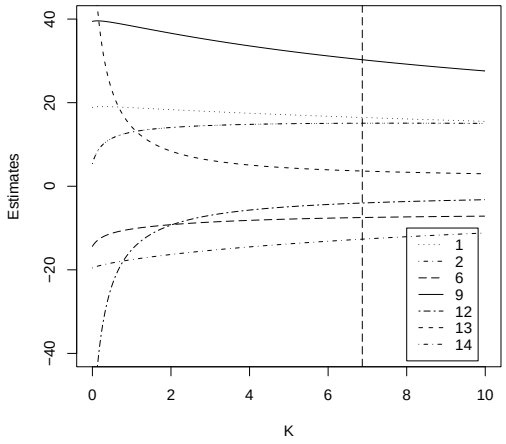

Fig. 2. Ridge trace plot for variables 1, 2, 6, 9, 12, 13, 14 of data Pollution. The vertical line represents the CV-choice of k.

1.7 Continuum Regression

RR discussed in Sect. 1.6 is very closely connected with the *continuum regression* proposed by [31] as a unifying approach to the LS, PCR, and partial least squares (see Sect. 1.9) estimation.

Definition 6. *A continuum regression (CR) estimator* $\mathbf{b}^{\mathbf{CR}}(\alpha)$ *of model (1) is a coefficient vector maximizing function*

$$T_\alpha(\mathbf{c}) = (\mathbf{c}^\top \mathbf{s})^2 (\mathbf{c}^\top \mathbf{S}\mathbf{c})^{\alpha-1} = (\mathbf{c}^\top \mathbf{X}^\top \mathbf{y})^2 (\mathbf{c}^\top \mathbf{X}^\top \mathbf{X}\mathbf{c})^{\alpha-1}, \tag{13}$$

for a given value of parameter $\alpha \geq 0$ *and a given length* $\|c\|$*, where* $\mathbf{S} = \mathbf{X}^\top \mathbf{X}$ *and* $\mathbf{s} = \mathbf{X}^\top \mathbf{y}$*.*

This definition yields estimates proportional to LS for $\alpha = 0$, PCR for $\alpha \to \infty$, and yet-to-be-discussed partial least squares for $\alpha = 1$. Apart from this, the advantage of CR is that one can adaptively select among the methods by searching an optimal α. To determine α, [31] used CV.

The relationship between RR and CR was indicated already in [82], but the most important result came after uncovering possible discontinuities of CR estimates as a function of data and α by [29]. In an attempt to remedy the discontinuity of the original CR, [30] not only proposed to maximize

$$T_\delta(\mathbf{c}) = (\mathbf{c}^\top \mathbf{s})^2 (\mathbf{c}^\top \mathbf{S}\mathbf{c})^{-1} |\mathbf{c}^\top \mathbf{S}\mathbf{c} + \delta|^{-1},$$

for $\delta \geq 0$ instead of $T_\alpha(\mathbf{c})$ from Def. 6 (δ can be chosen by CV, see [30]), but also proved the following proposition.

Theorem 4. *If a regressor* $\mathbf{b_f}$ *is defined according to*

$$\mathbf{b_f} = \underset{\|\mathbf{c}\|=1}{\operatorname{argmax}} f\{K^2(\mathbf{c}), V(\mathbf{c})\},$$

where $K(\mathbf{c}) = \mathbf{y}^\top \mathbf{X}\mathbf{c}$, $V(\mathbf{c}) = \|\mathbf{X}\mathbf{c}\|^2$, $f(K^2, V)$ *is increasing in* K^2 *for constant* V, *and increasing in* V *for constant* K^2, *and finally, if* $\mathbf{X}^\top\mathbf{y}$ *is not orthogonal to all eigenvectors corresponding to the largest eigenvalue* λ_{max} *of* $\mathbf{X}^\top\mathbf{X}$, *then there exists a number* $k \in (-\infty, \lambda_{max}) \cup [0, +\infty]$ *such that* $\mathbf{b_f}$ *is proportional to* $(\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$, *including the limiting cases* $k \to 0, k \to \pm\infty$, *and* $k \to -\lambda_{max}$.

Thus, the RR estimator fundamentally underlies many methods dealing with multicollinear and reduced rank data such as mentioned PCR and partial least squares. Notice however that negative values of the ridge coefficient k have to be admitted here.

Finally, let us note that CR can be extended to multiple-response-variables models [32].

1.8 Lasso

The ridge regression discussed in Sect. 1.6 motivates another shrinkage method: *Lasso* (least absolute shrinkage and selection operator) by [84]. Formulation (10) states that RR can be viewed as a minimization with respect to an upper bound on the L_2 norm of estimate $\|\mathbf{b}\|_2$. A natural extension is to consider constraints on the L_q norm $\|\mathbf{b}\|_q$, $q > 0$. Specifically, [84] studied case of $q = 1$, that is L_1 norm.

Definition 7. *The Lasso estimator for the regression model (16) is defined by*

$$\mathbf{b^L} = \underset{\|\boldsymbol{\beta}\|_1 \le r}{\operatorname{argmin}} \ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{14}$$

where $r \ge 0$ *is a tuning parameter.*

Lasso is a shrinkage estimator that has one specific feature compared to ordinary RR. Because of the geometry of L_1-norm restriction, Lasso shrinks the effect of some variables and eliminates influence of the others, that is, sets their coefficients to zero. Thus, it combines regression shrinkage with variable selection, and as [84] demonstrated also by means of simulation, it compares favorably to all-subsets regression. Considering variable selection, Lasso could be formulated as a special case of least angle regression [40]. To achieve the same kind of shrinking and variable-selection effects for all variables, they should be standardized before used in Lasso; see [9, Sect. 3.11] for details.

As far as the inference for the Lasso estimator is concerned, [61] recently studied the asymptotic distribution of Lasso-type estimators using L_q-norm condition $\|\boldsymbol{\beta}\|_q \le r$ with $q \le 1$, including behavior under nearly-singular designs.

Further, it remains to find out how Lasso estimates can be computed. Equation (14) indicates that one has to solve a restricted quadratic optimization problem. Setting $\beta_j^+ = \max\{\beta_j, 0\}$ and $\beta_j^- = -\min\{\beta_j, 0\}$, the restriction $\|\boldsymbol{\beta}\| \leq r$ can be written as $2p+1$ constraints: $\beta_j^+ \geq 0, \beta_j^- \geq 0$, and $\sum_{j=1}^{p}(\beta_j^+ - \beta_j^-) \leq r$. Thus, convergence is assured in $2p+1$ steps. Additionally, the unknown tuning parameter r is to be selected by means of CV. Finally, although solving (14) is straightforward in usual regression problems, it can become very demanding for reduced-rank data, $p > n$. Reference [72] treated lasso as a convex programming problem, and by formulating its dual problem, developed an efficient algorithm usable even for $p > n$.

Example 4. Let us use data Pollution once more to exemplify the use of Lasso. To summarize the Lasso results, we use the same plot as [84] and [40] used, see Fig. 3. It contains standardized slope estimates as a function of the constraint $\|b\| \leq r$, which is represented by an index $r/\max\|\mathbf{b}\| = \|\mathbf{b^L}\|/\|\mathbf{b^{LS}}\|$.[8] Moreover, to keep the graph simple, we plotted again only variables that were selected by variable selection procedures in Table 1 (1, 2, 6, 9, 12, 13, 14).

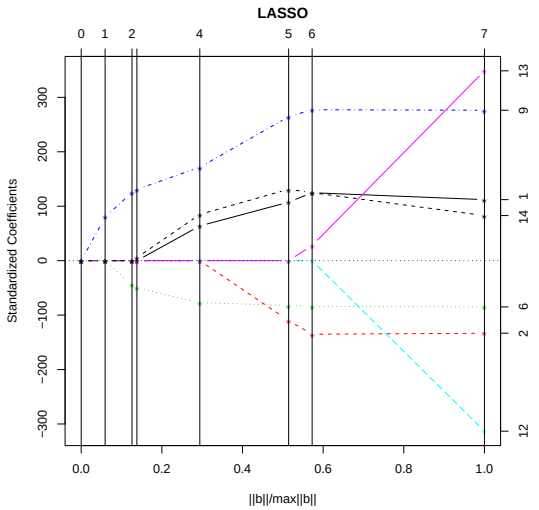

Fig. 3. Slope coefficients for variables 1, 2, 6, 9, 12, 13, 14 of data Pollution estimated by Lasso at different constraint levels, $r/\max\|\mathbf{b}\|$. The right axis assigns to each line the number of variable it represents and the top axis indicates the number of variables included in the regression.

In Fig. 3, we can observe which variables are included in the regression (have a nonzero coefficient) as tuning parameter r increases. Clearly, the order

[8] The LS estimate $\mathbf{b^{LS}}$ corresponds to the $\mathbf{b^L}$ under $r = \infty$, and thus, renders the maximum of $\|\mathbf{b^L}\|$.

in which the first of these variables become significant – 9, 6, 14, 1, 2 – closely resembles the results of variable selection procedures in Table 1. Thus, lasso combines shrinkage estimation and variable selection: at a given constraint level r, it shrinks coefficients of some variables and removes the others by setting their coefficients equal to zero.

1.9 Partial Least Squares

A general modeling approach to most of the methods covered so far was CR in Sect. 1.7, whereby it has two "extremes": LS for $\alpha = 0$ and PCR for $\alpha \to \infty$. The *partial least squares* (PLS) regression lies in between – it is a special case of (13) for $\alpha = 1$, see [31]. Originally proposed by [19], it was presented as an algorithm that searches for linear combinations of explanatory variables best explaining the dependent variable. Similarly to PCR, PLS also aims especially at situations when the number of explanatory variables is large compared to the number of observations (e.g., in chemometrics, sociology, etc.). Here we present the PLS idea and algorithm themselves as well as the latest results on variable selection and inference in PLS.

Having many explanatory variables $\mathbf{X}$, the aim of the PLS method is to find a small number of linear combinations $\mathbf{T_1} = \mathbf{X c_1}, \ldots, \mathbf{T_q} = \mathbf{X c_q}$ of these variables, thought about as latent variables, explaining observed responses

$$\hat{\mathbf{y}} = b_0 + \sum_{j=1}^{q} \mathbf{T_j} b_j \tag{15}$$

(see [45] and [48]). Thus, similarly to PCR, PLS reduces the dimension of data, but the criterion for searching linear combinations is different. Most importantly, it does not depend only on $\mathbf{X}$ values, but on $\mathbf{y}$ too.

Let us now present the PLS algorithm itself (see [10, Sect. 3.10] for more details and [45] for an alternative formulation), which defines yet another shrinkage estimator, see [58] and [46]. The indices $\mathbf{T_1}, \ldots, \mathbf{T_q}$ are constructed one after another. Estimating the intercept by $b_0 = \bar{\mathbf{y}}$, let us start with centered variables $\mathbf{z_0} = \mathbf{y} - \bar{\mathbf{y}}$ and $\mathbf{U_0} = \mathbf{X} - \bar{\mathbf{X}}$ and set $k = 1$.

1. Define the index $\mathbf{T_k} = \mathbf{U_{k-1}}(\mathbf{U_{k-1}^\top z_{k-1}})$. This linear combination is given by the covariance of the unexplained part of the response variable $\mathbf{z_{k-1}}$ and the unused part of explanatory variables $\mathbf{U_{k-1}}$.
2. Regress the current explanatory matrix $\mathbf{U_{k-1}}$ on index $\mathbf{T_k}$

   $$\mathbf{w_k} = (\mathbf{T_k^\top T_k})^{-1} \mathbf{T_k^\top U_{k-1}}$$

   and the yet-unexplained part of response $\mathbf{z_{k-1}}$ on index $\mathbf{T_k}$

   $$b_k = (\mathbf{T_k^\top T_k})^{-1} \mathbf{T_k^\top z_{k-1}},$$

   thus obtaining the kth regression coefficient.

3. Compute residuals, that is the remaining parts of explanatory and response variables: $\mathbf{U_k} = \mathbf{U_{k-1}} - \mathbf{T_k}\mathbf{w_k}$ and $\mathbf{z_k} = \mathbf{z_{k-1}} - \mathbf{T_k}b_k$. This implies that the indices $\mathbf{T_k}$ and $\mathbf{T_l}$ are not correlated for $k < l$.
4. Iterate by setting $k = k + 1$ or stop if $k = q$ is large enough.

This algorithm provides us with indices $\mathbf{T_k}$, which define the analogs of principle components in PCR, and the corresponding regression coefficients b_k in (15). The main open question is how to choose the number of components q. The original method proposed by [93] is based on cross validation. Provided that CV_k from (7) represents the CV index of PLS estimate with k factors, an additional index $\mathbf{T_{k+1}}$ is added if Wold's R criterion $R = CV_{k+1}/CV_k$ is smaller than 1. This selects the first local minimum of the CV index, which is superior to finding the global minimum of CV_k as shown in [73]. Alternatively, one can stop already when Wold's R exceeds 0.90 or 0.95 bound (modified Wold's R criteria) or to use other variable selection criteria such as AIC. Recent simulation study [63] showed that modified Wold's R is preferable to Wold's R and AIC. Furthermore, similarly to PCR, there are attempts to use GA for the component selection, see [62], for instance.

Next, the first results on the asymptotic behavior of PLS appeared first during last decade. The asymptotic behavior of prediction errors was examined by [49]. The covariance matrix, confidence and prediction intervals based on PLS estimates were first studied by [36], but a more compact expression was presented in [74]. It is omitted here due to many technicalities required for its presentation. There are also attempts to find a sample-specific prediction error of PLS, which were compared by [41].

Finally, note that there are many extensions of the presented algorithm, which is usually denoted PLS1. First of all, there are extensions (PLS2, SIMPLS, etc.) of PLS1 to models with multiple dependent variables, see [57] and [44] for instance, which choose linear combinations (latent variables) not only within explanatory variables, but does the same also in the space spanned by dependent variables. A recent survey of these and other so-called two-block methods is given in [90]. PLS was also adapted for on-line process modeling, see [76] for a recursive PLS algorithm. Additionally, in an attempt to simplify the interpretation of PLS results, [85] proposed orthogonalized PLS. See [96] for further details on recent developments.

Example 5. Let us use again data Pollution, although it is not a typical application of PLS. As explained in Sects. 1.7 and 1.9, PLS and PCR are both based on the same principle (searching for linear combinations of original variables), but use different objective functions. To demonstrate, we estimated PLS for 1 to 15 latent variables and plotted the fraction of the $\mathbf{X}$ and $\mathbf{y}$ variance explained by the PLS latent variables in the same way as in Fig. 1. Both curves are in Fig. 4. Almost all of the variability in $\mathbf{X}$ is captured by the first latent variable, although this percentage is smaller than in the case of PCR. On the other hand, the percentage of the variance of $\mathbf{y}$ explained by the first k latent

variables increases faster than in the case of PCR, see Fig. 4 (solid vs. dotted line).

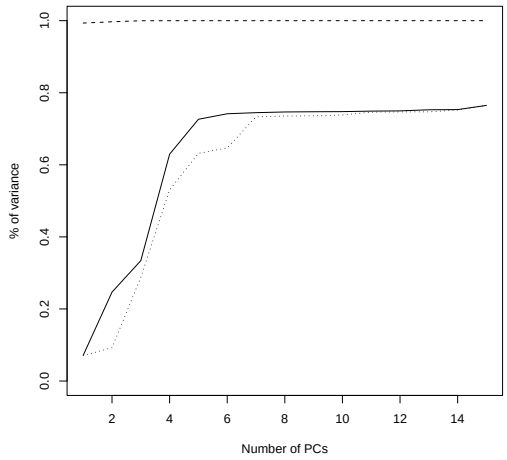

Fig. 4. Fraction of the explained variance of $\mathbf{X}$ (dashed line) and $\mathbf{y}$ (solid line) by the first k latent variables in PLS regression and by first k PCs (dotted lines).

1.10 Comparison of the Methods

Methods discussed in Sects. 1.3–1.9 are aiming at the estimation of (nearly) singular problems and they are often very closely related, see Sect. 1.7. Here we provide several references to studies comparing the discussed methods.

First, an extensive simulation study comparing variable selection, PCR, RR, and PLS regression methods is presented in [44]. Although the results are conditional on the simulation design used in the study, [44] found that PCR, RR, and PLS are, in the case of ill-conditioned problems, highly preferable to variable selection. The differences between the best methods, RR and PLS, are rather small and the same holds for comparison of PLS and PCR, which seems to be slightly worse than RR. An empirical comparison of PCR and PLS was also done by [92] with the same result. Next, the fact that neither PCR, nor PLS asymptotically dominates the other method was proved in [49] and further discussed in [48]. A similar asymptotic result was also given by [80]. Finally, the fact that RR should not perform worse than PCR and PLS is supported by Thm. 4 in Sect. 1.7.

2 Nonlinear Regression Modeling

In this section, we study the nonlinear regression model

$$y_i = h(\mathbf{x_i}, \boldsymbol{\beta_0}) + \varepsilon_i, \tag{16}$$

$i = 1, \ldots, n$, where $h : R^p \times R^k \to R$ is a known regression function and $\boldsymbol{\beta_0}$ is a vector of k unknown parameters. Let us note that the methods discussed in this section are primarily meant for truly nonlinear models rather than intrinsically linear models. A regression model is called intrinsically linear if it can be unambiguously transformed to a model linear in parameters. For example, the regression model $y = \beta_1 x/(\beta_2 + x)$ can be expressed as $1/y = 1/\beta_1 + \beta_2/\beta_1 x$, which is linear in parameters $\theta_1 = 1/\beta_1$ and $\theta_2 = \beta_2/\beta_1$. Transforming a model to its linear form can often provide better inference, such as confidence regions, although one has to be aware of the effects of the transformation on the error-term distribution.

We first discuss the fitting and inference in the nonlinear regression (Sects. 2.1 and 2.2), whereby we again concentrate on the least square estimation. For an extensive discussion of theory and practice of nonlinear least squares regression see monographs [2], [11], and [14]. Second, similarly to the linear modeling section, methods for ill-conditioned nonlinear systems are briefly reviewed in Sect. 2.3.

2.1 Fitting of Nonlinear Regression

In this section, we concentrate on estimating the vector $\boldsymbol{\beta_0}$ of unknown parameters in (16) by *nonlinear least squares*.

Definition 8. *The nonlinear least squares (NLS) estimator for the regression model (16) is defined by*

$$\mathbf{b^{NLS}} = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} \{y_i - \hat{y}_i(\boldsymbol{\beta})\}^2 = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} \{y_i - h(\mathbf{x_i}, \boldsymbol{\beta})\}^2. \tag{17}$$

Contrary to the linear model fitting, we cannot express analytically the solution of this optimization problem for a general function h. On the other hand, we can try to approximate the nonlinear objective function using the Taylor expansion because the existence of the first two derivatives of h is an often used condition for the asymptotic normality of NLS, and thus, could be readily assumed. Denoting $\mathbf{h}(\mathbf{b}) = \{h(\mathbf{x_i}, \mathbf{b})\}_{i=1}^n$ and $S_n(\mathbf{b}) = \sum_{i=1}^n [y_i - h(\mathbf{x_i}, \mathbf{b})]^2$, we can state the following theorem from [1, Chap. 4].

Theorem 5. *Let ε_i in (16) are independent and identically distributed with $E(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$ and $Var(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2 \mathbf{I_n}$ and let B be an open neighborhood of $\boldsymbol{\beta_0}$. Further, assume that $h(\mathbf{x}, \mathbf{b})$ is continuous on B uniformly with respect to $\mathbf{x}$ and twice continuously differentiable in B and that*

1. $\lim_{n\to\infty} S_n(\mathbf{b}) \neq 0$ *for* $\mathbf{b} \neq \boldsymbol{\beta_0}$*;*
2. $[\partial \mathbf{h}(\mathbf{b})/\partial \mathbf{b}^\top]^\top [\partial \mathbf{h}(\mathbf{b})/\partial \mathbf{b}^\top]/n$ *converges uniformly in* B *to a finite matrix* $\mathbf{A}(\mathbf{b})$*, such that* $\mathbf{A}(\boldsymbol{\beta_0})$ *is nonsingular;*
3. $\mathbf{h}(\mathbf{b^1})^\top [\partial^2 \mathbf{h}(\mathbf{b^2})/\partial b_j \partial b_k]/n$ *converges uniformly for* $\mathbf{b^1}, \mathbf{b^2} \in B$ *to a finite matrix for all* $j, k = 1, \ldots, k$*.*

Then the NLS estimator $\mathbf{b^{NLS}}$ *is consistent and asymptotically normal*

$$\sqrt{n}\left(\mathbf{b^{NLS}} - \boldsymbol{\beta_0}\right) \to N(0, \sigma^2 \mathbf{A}(\boldsymbol{\beta_0})^{-1}). \tag{18}$$

Hence, although there is no general explicit solution to (17), we can assume without loss of much generality that the objective function $S_n(\mathbf{b})$ is twice differentiable in order to devise a numerical optimization algorithm. The second-order Taylor expansion provides then a quadratic approximation of the minimized function, which can be used for obtaining an approximate minimum of the function, see [14]. As a result, one should search in the direction of the steepest descent of a function, which is given by its gradient, to get a better approximation of the minimum. We discuss here the incarnations of these methods specifically for the case of quadratic loss function in (17).

Newton's method

The classical method based on the gradient approach is Newton's method, see [8] and [14] for detailed discussion. Starting from an initial point $\mathbf{b^1}$, a better approximation is found by taking

$$\begin{aligned} \mathbf{b^{k+1}} &= \mathbf{b^k} - \mathbf{H}^{-1}(\mathbf{r}^2, \mathbf{b^k})\mathbf{J}(\mathbf{r}, \mathbf{b^k}) = \\ &= \mathbf{b^k} - \left[\mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{J}(\mathbf{h}, \mathbf{b^k}) + \sum_{l=1}^{n} r_i(\mathbf{b}) \mathbf{H}(h_i, \mathbf{b^k})\right]^{-1} \mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{r}(\mathbf{b^k}), \end{aligned} \tag{19}$$

where $\mathbf{r}(\mathbf{b}) = \{[y_i - h(x_i, \mathbf{b})]\}_{i=1}^n$ represents residuals, $\mathbf{J}(\mathbf{f}, \mathbf{b}) = \partial \mathbf{f}(\mathbf{b})/\partial \mathbf{b}^\top$ is the Jacobian matrix of a vector function $\mathbf{f}(\mathbf{b})$, and $\mathbf{H}(\mathbf{f}, \mathbf{b}) = \partial^2\{\sum_{i=1}^n \mathbf{f}(\mathbf{b})\}$ $/\partial \mathbf{b} \partial \mathbf{b}^\top$ is the Hessian matrix of $\mathbf{f}(\mathbf{b})$.

To find $\mathbf{b^{NLS}}$, equation (19) is iterated until convergence is achieved. This is often verified by checking whether the relative change from $\mathbf{b^k}$ to $\mathbf{b^{k+1}}$ is sufficiently small. Unfortunately, this criterion can indicate a lack of progress rather than convergence. Instead, [2, Sect. 2.2] proposed to check convergence by looking at some measure of orthogonality of residuals $\mathbf{r}(\mathbf{b^k})$ towards the regression surface given by $\mathbf{h}(\mathbf{b^k})$, since the identification assumption of model (16) is $E(\mathbf{r}(\boldsymbol{\beta_0})|\mathbf{X}) = 0$. See [8], [2], [3], and [12] for more details and further modifications.

To evaluate iteration (19), it is necessary to invert the Hessian matrix $\mathbf{H}(\mathbf{r}^2, \mathbf{b})$. From the computational point of view, all issues discussed in Sect. 1 apply here too and one should use a numerically stable procedure, such as QR decomposition, to perform the inversion. Moreover, to guarantee that (19) leads to a better approximation of the minimum, that is

$\mathbf{r}(\mathbf{b^{k+1}})^\top \mathbf{r}(\mathbf{b^{k+1}}) \leq \mathbf{r}(\mathbf{b^k})^\top \mathbf{r}(\mathbf{b^k})$, the Hessian matrix $\mathbf{H}(\mathbf{r}^2, \mathbf{b^k})$ needs to be positive definite, which in general holds only in a neighborhood of $\boldsymbol{\beta_0}$ (see the Levenberg-Marquard method for a solution). Even if it is so, the step in the gradient direction should not be too long, otherwise we "overshoot." Modified Newton's method addresses this by using some fraction α_{k+1} of iteration step $\mathbf{b^{k+1}} = \mathbf{b^k} - \alpha_{k+1}\mathbf{H}^{-1}(\mathbf{r}^2, \mathbf{b^k})\mathbf{J}(\mathbf{r}, \mathbf{b^k})$. See [43], [8], and [28] for some choices of α_{k+1}.

Gauss-Newton method

The Gauss-Newton method is designed specifically for NLS by replacing the regression function $h(\mathbf{x_i}, \mathbf{b})$ in (17) by its first-order Taylor expansion, see [14] and [8]. The resulting iteration step is

$$\mathbf{b^{k+1}} = \mathbf{b^k} - \left\{\mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{J}(\mathbf{h}, \mathbf{b^k})\right\}^{-1} \mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{r}(\mathbf{b^k}). \tag{20}$$

Being rather similar to Newton's method, it does not require the Hessian matrix $\mathbf{H}(\mathbf{r}^2, \mathbf{b})$, which is "approximated" by $\mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{J}(\mathbf{h}, \mathbf{b^k})$ (both matrices are equal in probability for $n \to \infty$ under assumptions of Thm. 5, see [1, Chap. 4]). Because it only approximates the true Hessian matrix, this method belongs to the class of quasi-Newton methods. The issues discussed in the case of Newton's method apply also to the Gauss-Newton method.

Levenberg-Marquard method

Depending on data and the current approximation $\mathbf{b^k}$ of $\mathbf{b^{NLS}}$, the Hessian matrix $\mathbf{H}(\mathbf{b^k})$ or its approximations such as $\mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{J}(\mathbf{h}, \mathbf{b^k})$ can be badly conditioned or not positive definite, which could even result in divergence of Newton's method (or a very slow convergence in the case of modified Newton's method). The Levenberg-Marquard method addresses the ill-conditioning by choosing the search direction $\mathbf{d_k} = \mathbf{b^{k+1}} - \mathbf{b^k}$ as a solution of

$$\left\{\mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{J}(\mathbf{h}, \mathbf{b^k}) + \tau \mathbf{I_p}\right\} \mathbf{d_k} = -\mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{r}(\mathbf{b^k}) \tag{21}$$

(see [67]). This approach is an analogy of RR discussed in Sect. 1.6, and hence, it limits the length of the innovation vector $\mathbf{d_k}$ compared to the (Gauss-)Newton method. See [8] and [3] for a detailed discussion of this algorithm. There are also algorithms combining both Newton's and the Levenberg-Marquard approaches by using at each step the method that generates a larger reduction in objective function.

Although Newton's method and its modifications are most frequently used in applications, the fact that they find local minima gives rise to various improvements and alternative methods. They range from simple starting the minimization algorithm from several (randomly chosen) initial points to general global-search optimization methods such as genetic algorithms mentioned in Sect. 1.3 and discussed in more details in Chapts. 5 and 6 of this Handbook.

2.2 Statistical Inference

Similarly to linear modeling, the inference in nonlinear regression models is mainly based, besides the estimate $\mathbf{b^{NLS}}$ itself, on two quantities: the residual sum of squares $RSS = \mathbf{r}(\mathbf{b^{NLS}})^\top \mathbf{r}(\mathbf{b^{NLS}})$ and the (asymptotic) variance of the estimate $Var(\mathbf{b^{NLS}}) = \sigma^2 \mathbf{A}(\boldsymbol{\beta_0})^{-1}$, see (18). Here we discuss how to compute these quantities for $\mathbf{b^{NLS}}$ and its functions.

RSS will be typically a by-product of a numerical computation procedure, since it constitutes the minimized function. RSS also provides an estimate of σ^2: $s^2 = RSS/(n-k)$. The same also holds for the matrix $\mathbf{A}(\boldsymbol{\beta_0})$, which can be consistently estimated by $\mathbf{A}(\mathbf{b^{NLS}}) = \mathbf{J}(\mathbf{h}, \mathbf{b^k})^\top \mathbf{J}(\mathbf{h}, \mathbf{b^k})$, that is, by the asymptotic representation of the Hessian matrix $\mathbf{H}(\mathbf{r}^2, \mathbf{b})$. This matrix or its approximations are computed at every step of (quasi-)Newton methods for NLS, and thus, it will be readily available after the estimation.

Furthermore, the inference in nonlinear regression models may often involve a nonlinear (vector) function of the estimate $f(\mathbf{b^{NLS}})$; for example, when we test a hypothesis (see [14] for a discussion of NLS hypothesis testing). Contrary to linear functions of estimates, where $Var(\mathbf{A}\mathbf{b^{NLS}} + \mathbf{a}) = \mathbf{A}^\top Var(\mathbf{b^{NLS}})\mathbf{A}$, there is no exact expression for $Var[f(\mathbf{b^{NLS}})]$ in a general case. Thus, we usually assume the first-order differentiability of $f(\cdot)$ and use the Taylor expansion to approximate this variance. Since

$$f(\mathbf{b}) = f(\boldsymbol{\beta_0}) + \frac{\partial f(\boldsymbol{\beta_0})}{\partial \mathbf{b}^\top}(\mathbf{b} - \boldsymbol{\beta_0}) + o(\|\mathbf{b} - \boldsymbol{\beta_0}\|),$$

it follows that the variance can be approximated by

$$Var[f(\mathbf{b^{NLS}})] \doteq \frac{\partial f(\mathbf{b^{NLS}})}{\partial \mathbf{b}^\top} Var(\mathbf{b^{NLS}}) \frac{\partial f(\mathbf{b^{NLS}})}{\partial \mathbf{b}}.$$

Hence, having an estimate of $Var(\mathbf{b^{NLS}})$, the Jacobian matrix of function f evaluated at $\mathbf{b^{NLS}}$ provides a first-order approximation of the variance of $f(\mathbf{b^{NLS}})$.

2.3 Ill-conditioned Nonlinear System

Similarly to linear modeling, the nonlinear models can also be ill-conditioned when the Hessian matrix $H(\mathbf{r}^2, \mathbf{b})$ is nearly singular or does not even have a full rank, see Sect. 1.2. This can be caused either by the nonlinear regression function h itself or by too many explanatory variables relative to sample size n. Here we discuss extensions of methods dealing with ill-conditioned problems in the case of linear models, Sects. 1.5–1.9, to nonlinear modeling: ridge regression, Stein-rule estimator, Lasso, and partial least squares.

First, one of early nonlinear RR was proposed by [34], who simply added a diagonal matrix to $\mathbf{H}(\mathbf{r}^2, \mathbf{b})$ in equation (19). Since the nonlinear modeling is done by minimizing of an objective function, a more straightforward way is to use the alternative formulation (11) of RR and to minimize

$$\sum_{i=1}^{n}\{y_i - h(\mathbf{x_i}^\top, \boldsymbol{\beta})\}^2 + k\sum_{j=1}^{p}\beta_j^2 = \mathbf{r}(\boldsymbol{\beta})^\top \mathbf{r}(\boldsymbol{\beta}) + k\|\boldsymbol{\beta}\|_2^2, \tag{22}$$

where k represents the ridge coefficient. See [70] for an application of this approach.

Next, equally straightforward is an application of Stein-rule estimator (8) in nonlinear regression, see [60] for a recent study of positive-part Stein-rule estimator within the Box-Cox model. The same could possibly apply to Lasso-type estimators discussed in Sect. 1.8 as well: the Euclidian norm $\|\boldsymbol{\beta}\|_2^2$ in (22) would just have to be replaced by another L_q norm. Nevertheless, the behavior of Lasso within linear regression has only recently been studied in more details, and to my best knowledge, there are no results on Lasso in nonlinear models yet.

Finally, there is a range of modifications of PLS designed for nonlinear regression modeling, which either try to make the relationship between dependent and explanatory variables linear in unknown parameters or deploy an intrinsically nonlinear model. First, the methods using linearization are typically based on approximating a nonlinear relationship by higher-order polynomials (see quadratic PLS by and INLR approach by [26]) or a piecewise constant approximation (GIFI approach, see [27]). For their recent overview see [96]. Second, several recent works introduced intrinsic nonlinearity into PLS modeling. Among most important contributions, there are [75] and [66] modeling the nonlinear relationship using a forward-feed neural network, [95] and [38] transforming predictors by spline functions, and [23] using fuzzy-clustering regression approach, for instance.

References

1. Amemiya T (1985) Advanced Econometrics. Harvard University Press, Cambridge, USA
2. Bates D M, Watts D G (1988) Nonlinear Regression Analysis and Its Applications. John Wiley & Sons, USA
3. Björk A (1996) Numerical Methods for Least Squares Problems. SIAM Press, Philadelphia, PA, USA
4. Chambers L (1998) Practical Handbook of Genetic Algorithms: Complex Coding Systems, Volume III. CRC Press, USA
5. Gunst R F and Mason R L (1980) Regression Analysis and Its Application: a Data-Oriented Approach. Marcel Dekker, Inc., New York, USA
6. Härdle W (1992) Applied Nonparametric Regression. Cambridge University Press, Cambridge, UK
7. Härdle W and Simar L (2003) Applied Multivariate Statistical Analysis. Springer, Germany
8. Kennedy W J, Gentle J E (1980) Statistical Computing. Marcel Dekker, Inc., New York
9. Miller A (2002) Subset Selection in Regression, Chapman & Hall/CRC, USA

10. Rao C R and Toutenberg H (1999) Linear Models, Springer, New York
11. Seber G A F and Wild C J (1989) Nonlinear Regression, Wiley, New York
12. Thisted R A (1991) Elements of Statistical Computing. Chapman and Hall, London New York
13. Vinod H D and Ullah A (1981) Recent Advances in Regression Methods. Marcel Dekker Inc., New York Basel
14. Amemiya T (1983) Non-linear regression models. In Griliches Z and Intriligator M D (eds) Handbook of Econometrics, Volume 1. North-Holland Publishing Company, Amsterdam
15. Barlow J L (1993) Numerical aspects of solving linear least squares problems. In Rao C R (ed) Handbook of Statistics, Volume 9. Elsevier, Amsterdam London New York Tokyo
16. Efroymson (1960) Multiple regression analysis. In Ralston A and Wilf H S (eds) Mathematical Methods for Digital Computers, Vol. 1, Wiley, New York
17. Judge G G and Bock M E (1983) Biased estimation. In Griliches Z and Intriligator M D (eds) Handbook of Econometrics, Volume 1. North-Holland Publishing Company, Amsterdam
18. Leamer E E (1983) Model choice and specification analysis. In Griliches Z and Intriligator M D (eds) Handbook of Econometrics, Volume 1. North-Holland Publishing Company, Amsterdam
19. Wold H (1966) Estimation of principle components and related models by iterative least squares. In Krishnaiaah (ed) Multivariate analysis. Academic Press, New York
20. Brandt J, Hein W (2001) Polymer materials in joint surgery. In: Grellmann W, Seidler S (eds) Deformation and fracture behavior of polymers. Engineering materials. Springer, Berlin Heidelberg New York
21. Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723
22. Akdeniz F, Yüksel G, and Wan A T K (2003) The moments of the operational almost unbiased ridge regression estimator. Applied Mathematics and Computation, in press.
23. Bang Y H, Yoo C K, Lee I-B (2003) Nonlinear PLS modeling with fuzzy inference system. Chemometrics and Intellingent Laboratory Systems 64: 137–155
24. Barros A S and Rutledge D N (1998) Genetic algorithm applied to the selection of principal components. Chemometrics and Intelligent Laboratory Systems 40: 65–81
25. Bedrick E J and Tsai C-L (1994) Model Selection for Multivariate Regression in Small Samples. Biometrics 50: 226–231.
26. Berglund A and Wold S (1997) INLR, implicit nonlinear latent variable regression. Journal of Chemometrics 11: 141–156
27. Berglund A, Kettaneh, Wold S, Bendwell N, and Cameron D R (2001) The GIFI approach to non-linear PLS modelling. Journal of Chemometrics 15: 321–336
28. Berndt E R, Hall B H, Hall R E, and Hausman J A (1974) Estimation and Inference in Nonlinear Structural Models. Annals of Econometric and Social Measurement 3: 653–666
29. Björkström A and Sundberg R (1996) Continuum regression is not always continuous. Journal of Royal Statistical Society B 58: 703–710
30. Björkström A and Sundberg R (1999) A generalized view on continuum regression. Scandinavian Journal of Statistics 26: 17–30

31. Brooks R and Stone M (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. Journal of Royal Statistical Society B 52: 237–269
32. Brooks R and Stone M (1994) Joint continuum regression for multiple predicants. Journal of American Statistical Association 89: 1374–1377
33. Chawla J S (1990) A note on ridge regression. Statistics & Probability Letters 9: 343–345
34. Dagenais M G (1983) Extension of the ridge regression technique to non-linear models with additive errors. Economic Letters 12: 169–174
35. Danilov D and Magnus J R (2003) On the harm that ignoring pretesting can cause. Journal of Econometrics, in press
36. Denham M C (1997) Prediction intervals in partial least squares. Journal of Chemometrics 11: 39–52
37. Depczynski U, Frost V J, and Molt K (2000) Genetic algorithms applied to the selection of factors in principal component regression. Analytica Chimica Acta 420: 217–227
38. Durand J-F, Sabatier R (1997) Additive spline for partial least squares regression. Journal of American Statistical Association 92: 1546–1554
39. Edwards D and Havranek T (1987) A fast model selection procedure for large families of models. Journal of American Statistical Association 82: 205–213
40. Efron B, Hastie T, Johnstone I, and Tibshirani R (2003) Least Angle Regression. Annals of Statistics, in press
41. Faber N M, Song X-H, and Hopke P K (2003) Sample specific standard error of prediction for partial least squares regression. Trends in Analytical Chemistry 22: 330–334
42. Farebrother R W (1976) Further results on the mean square error of ridge estimation. Journal of Royal Statistical Society B 38: 248–250.
43. Fletcher R and Powell M J D (1963) A rapidly convergent descent method for minimization. Computer Journal 6: 163–168
44. Frank I E, Friedman J H, Wold S, Hastie T, and Mallows C (1993) A statistical view of some chemometrics regression tools. Technometrics 35 / 2: 109–148
45. Garthwaite P H (1994) An interpretation of partial least squares. The Journal of American Statistical Association 89: 122–127
46. Coutis C (1996) Partial least squares algorithm yields shrinkage estimators. The Annals of Statistics 24: 816–824
47. Hawkins D M and Yin X (2002) A faster algorithm for ridge regression of reduced rank data. Computational Statistics & Data analysis 40: 253–262
48. Helland I S (2001) Some theoretical aspects of partial least squares regression. Chemometrics and Intellingent Laboratory Systems 58: 97–107
49. Helland I S and Almoy T (1994) Comparison of Prediction Methods When Only a Few Components are Relevant. Journal of the American Statistical Association 89: 583–591
50. Hoerl A E and Kennard R W (1970) Ridge regression: biased estimation of nonorthogonal problems. Technometrics 12: 55–67
51. Hoerl A E, Kennard R W, and Baldwin K F (1975) Ridge regression: some simulations. Communications in Statistics 4: 105–123
52. Hughes A W and Maxwell L K (2003) Model selection using AIC in the presence of one-sided information. Journal of Statistical Planning and Inference 115: 379–411

53. Hwang J T G and Nettleton D (2003) Principal components regression with data-chosen components and related methods. Technometrics 45: 70–79
54. Ibrahim J G and Ming-Hui C (1997) Predictive Variable Selection for the Multivariate Linear Model. Biometrics 53: 465–478
55. Jian W and Liu X (2003) Consistent model selection based on parameter estimates. Journal of Statistical Planning and Inference, in press.
56. Jolliffe I T (1982) A note on the use of the principle components in regression. Applied Statistics 31 / 3: 300–303
57. Jong S (1993) SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18: 251–263
58. Jong S (1995) PLS shrinks. Journal of Chemometrics 9: 323–326
59. Kadiyala K (1984) A class of almost unbiased and efficient estimators of regression coefficients. Ecnomic Letters 16: 293–296.
60. Kim M, Hill R C (1995) Shrinkage estimation in nonlinear regression: the Box-Cox transformation. Journal of Econometrics 66: l–33
61. Knight K and Fu W (2000) Asymptotics for Lasso-type estimators. The Annals of Statistics 28: 1356–1389
62. Leardi R and Gonzáles A L (1998) Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemometrics and Intellingent Laboratory Systems 41: 195–207
63. Li B, Morris J, and Martin E B (2002) Model section for partial least squares regression. Chemometrics and Intellingent Laboratory Systems 64: 79–89
64. Magnus J R (1999) The traditional pretest estimator. Theory of Probability and Its Applications 44 / 2: 293–308
65. Magnus J R (2002) Estimation of the mean of a univariate normal distribution with known variance. The Econometrics Journal 5, 225–236
66. Malthouse E C, Tamhane A C, and Mah R S H (1997) Nonlinear partial least squares. Computers chem. Engng 21 / 8: 875–890
67. Marquardt D W (1963) An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics 11: 431–441
68. McDonald G C and Schwing R C (1973) Instabilities of regression estimates relating air pollution to mortality. Technometrics 15: 463–482
69. Miller A J (1984) Selection of Subsets of Regression Variables. Journal of the Royal Statistical Society, Series A, 147 / 3: 389–425
70. Ngo SH, Kemény S, and Deák A (2003) Performance of the ridge regression methods as applied to complex linear and nonlinear models. Chemometrics and Intellingent Laboratory Systems 67: 69-78
71. Ohtani K (1986) On small sample properties of the almost unbiased generalized ridge estimator. Communications in Statistics, Theory and Methods 22: 2733–2746
72. Osborne M R, Presnell B, and Turlach B A (1999) On the Lasso and its dual. Journal of Computational and Graphical Statistics 9: 319–337
73. Osten D W (1988) Selection of optimal regression models via cross-validation. Journal of Chemometrics 2: 39–48
74. Phatak A, Reilly P M, Pendilis A (2002) The asymptotic variance of the univariate PLS estimator. Linear Algera and its Applications 354: 245–253
75. Qin S and McAvoy T (1992) Nonlinear PLS modeling using neural networks. Computers in Chemical Engineering 16: 379–391

76. Qin S J (1997) Recursive PLS algorithms for adaptive data modeling. Computers in Chemical Engineering 22 / 4: 503–514
77. Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6: 461–464
78. Shao J (1993) Linear model selection by cross-validation. Journal of American Statistical Association 88: 486–494
79. Shi P and Tsai C-L (1998) A note on the unification of the Akaike information criterion. Journal of the Royal Statistical Society, Series B, 60: 551–558
80. Stoica P and Söderström T (1998) Partial least squares: a first-order analysis. Scandinavian Journal of Statistics 25: 17–26
81. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. Journal of Royal Statistical Society B 36: 111–147
82. Sundberg R (1993) Continuum regression and ridge regression. Journal of Royal Statistical Society B 55: 653–659
83. Swamy P A V B, Mehta J S, and Rappoport P N (1978) Two methods of evaluating Hoerl and Kennard's ridge regression. Communications in Statistics A 12: 1133–1155
84. Tibshirani R (1996) Regression shrinkage and selection via Lasso. Journal of Royal Statistical Society B 58: 267–288
85. Trygg J and Wold S (2002) Orthogonal projections to latent structures, O-PLS. Journal of Chemometrics 16 / 3: 119–128
86. Ullah A, Sristava V K, and Chandra R (1983) Properties of shrinkage estimators in linear regression when disturbances are not normal. Journal of Econometrics 21: 289–402
87. Wan A T K (2002) On generalized ridge regression estimators under collinearity and balanced loss. Applies Mathematics and Computation 129: 455–467
88. Wang S G and Chow S C (1990) A note on adaptive generalized ridge regression estimator. Statistics & Probability Letters 10: 17–21
89. Wasserman G S and Sudjianto A (1994) All subsets regression using a generic search algorithm. Computers and Industrial Engineering 27: 489–492
90. Wegelin J A (2000) A survey of partial least squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle, 2000.
91. Weiss R E (1995) The influence of variable selection: a bayesian diagnostic perspective. Journal of the American Statistical Association 90: 619–625
92. Wentzell P D, Montoto L V (2003) Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. Chemometrics and Intelligent Laboratory Systems 65: 257–279
93. Wold S (1978) Cross-validation estimation of the number of components in factor and principal components analysis. Technometrics 24: 397–405
94. Wold S, Kettaneh-Wold N, and Skagerberg B (1989) Nonlinear PLS modelling. Chemometrics and Intelligent Laboratory Systems 7: 53–65
95. Wold S (1992) nonlinear partial least squares modelling II. Spline inner relation. Chemometrics and Intellingent Laboratory Systems 14: 71–84
96. Wold S, Trygg J, Berglund A, Atti H (2001) Some recent developments in PLS modeling. Chemometrics and Intelligent Laboratory Systems 58: 131–150
97. Zhang P (1992) Inference after variable selection in linear regression models. Biometrika 79 / 4: 741–746
98. Zheng X and Loh W-Y (1995) Consistent variable selection in linear models. Journal of the American Statistical Association 90: 151–156

99. Ross D W (1977) Lysosomes and storage diseases. MA Thesis, Columbia University, New York

Index