Chapter 5

Least Squares Regression: Biased Regression Methods

5.1 Introduction

This chapter deals with *Gaussian* regression models. The techniques considered fall under the umbrella of *biased* regression methods, which are useful in situations where there are many explanatory variables, but there is ill-conditioning of the (X^tX) matrix since the columns $X_{j.} : j = 1, \ldots, r$ have strong linear dependence. The matrix X^tX may be singular. In this chapter, principal component regression and partial least squares regression are considered; in the next, *penalised* techniques, such as ridge regression, least angle regression and LASSO are considered.

Regularised regression (in particular Lasso), enables us to select a subset of the explanatory variables as does least-angle regression, which is an automatic variable selection method, which improves the forward stepwise technique.

5.2 The Generalised Inverse

Recall that the basic equation for multiple linear regression is

$$Y = X\beta + \epsilon$$

where β is a p-vector of unknown parameters, X is an $n \times p$ design matrix and $\epsilon \sim N(0, \sigma^2 I)$. The assumption of Gaussian errors may be relaxed, but if we assume that $\epsilon_1, \ldots, \epsilon_n$ are independent, mean 0 and each with variance σ^2, then we estimate the parameters by *ordinary least squares* (OLS) to obtain the OLS estimator; $\widehat{\beta}_{OLS}$ minimises

$$(Y - X\beta)^t(Y - X\beta).$$

Any β which satisfies the so called *normal equation*

$$(X^tX)\beta = X^tY$$

gives a minimum; this is the OLS (ordinary least squares) estimate. If X^tX is invertible, the OLS estimator is unique and:

$$\widehat{\beta}_{OLS} = (X^tX)^{-1}X^tY.$$

In many situations, for example chemometrics (e.g. food research, environmental pollution studies) it is often the case that the number of variables exceeds the number of observations, so that (X^tX) is not invertible.

In *experimental design*, it is often very useful to consider an *over parametrised* model, where the parameters have an intuitive interpretation. For example, consider the model:

$$Y_{ij} = \alpha + \beta_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0,\sigma^2)$ are independent of each other, α is an overall average and β_i is the effect of treatment i. There are m different treatments and n_i experimental units subject to treatment i. For example, consider $n_1 = 2$, $n_2 = 2$, $n_1 = 1$, so that $n = 5$ (the total number of experimental units). As a linear model, this may be written as

$$Y = X\beta + \epsilon$$

where this is short-hand for:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{pmatrix}$$

where the observations and errors have been put into vectors and relabelled.

Here X is a 5×4 matrix of rank 3. Clearly X^tX is not invertible.

Suppose that X is $n \times r$, but is of rank t where $t < r$. Then X^tX does not have an inverse. A generalised inverse G always exists.

Definition 5.1 (Generalised Inverse)**.** *A generalised inverse G of a matrix A is a matrix that satisfies: $AGA = A$.*

Lemma 5.2. *Let A be a symmetric $n \times n$ matrix of rank $r \leq n$. There exists a generalised inverse.*

Proof By rearranging the rows and columns of A, it may be written as $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ where A_{11} is an $r \times r$ matrix of rank r. Let

$$G = \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

then an elementary computation yields: $AGA = A$. □

In cases where X^tX is not invertible, any generalised inverse G of X^tX gives a solution to the normal equations: $\widehat{\beta} = GX^tY$, but the solution is not unique. The estimation, however, is unique for *estimable* functions:

Lemma 5.3. *Let v^t be a linear combination of the rows of X. For such a v, the function $v^t\beta$ is said to be estimable. For estimable functions, the OLS estimator $v^t\widehat{\beta}$ is uniquely defined.*

Proof Let G_1 and G_2 be any two generalised inverses of X^tX. Let $\widehat{\beta}_1 = G_1X^tY$ and $\widehat{\beta}_2 = G_2X^tY$. Then, since $X^tX\widehat{\beta}_i = X^tY$ for $i = 1, 2$,

$$0 = (X^tX)(\widehat{\beta}_1 - \widehat{\beta}_2)$$

and hence for any linear combination λ^t of the rows of X^tX, $\lambda^t\widehat{\beta}_1 = \lambda^t\widehat{\beta}_2$ and hence for any vector $v = X\lambda$ for some λ and linear combinations of such. The result follows. □

Going back to the example from experimental design, we can see that α (the 'overall' average) is not estimable; it cannot be obtained in this way. But $\alpha + \beta_i$ is estimable for $i = 1, 2, 3$. These correspond to the average value for the outcome from treatments $1, 2, 3$ respectively. Also, $\beta_1 - \beta_2$ (the difference in average outcome from using treatments 1 and 2) is estimable; $v^t = (1, 1, 0, 0) - (1, 0, 1, 0) = (0, 1, -1, 0)$ gives $v^t\beta = \beta_1 - \beta_2$, where $\beta = (\alpha, \beta_1, \beta_2, \beta_3)^t$.

5.2.1 Moore-Penrose Generalised Inverse

One way (an obvious way) to obtain a generalised inverse is as follows: let V be the orthonormal matrix and Λ the diagonal matrix, with entries $\lambda_1 \geq \lambda_2 \geq \ldots$ such that

$$X^tX = V\Lambda V^t.$$

and let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$. Then $X^tX = V\Lambda V^t$ and, if $\lambda_r > 0$, the inverse is well defined and satisfies: $(X^tX)^{-1} = V\Lambda^{-1}V^t$.

Now consider the situation where $\lambda_1 \geq \ldots \geq \lambda_t > 0$ and $\lambda_{t+1} = \ldots = \lambda_r = 0$. Let

$$V = (v_1 | \ldots | v_r)$$

and let $\widetilde{V} = (v_1|\ldots|v_t)$. While there is a whole family of generalised inverses, the *Moore-Penrose* generalised inverse is uniquely defined. If X^tX is of rank t, its Moore-Penrose generalised inverse is defined as:

$$G = \widetilde{V}\widetilde{\Lambda}^{-1}\widetilde{V}^t.$$

where $\widetilde{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_t)$. It is clear that this is a generalised inverse since

$$(X^tX) = (\widetilde{V}|\hat{V})\begin{pmatrix} \widetilde{\Lambda} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \widetilde{V}^t \\ \widehat{V} \end{pmatrix} = \widetilde{V}\Lambda\widetilde{V}.$$

The *generalised-inverse regression* estimator is:

$$\widehat{\beta}_{gir} = GX^tY.$$

In the absence of further information, the Moore-Penrose generalised inverse is the default option in the software. The fitted values are $\widehat{Y}_{gir} = X\widehat{\beta}_{gir}$.

The fitted values are estimators of μ, the mean vector and these are always estimable (since each mean is obtained from an individual row of X). As indicated above, each choice of generalised inverse gives the same answer when computing the OLS estimates of estimable functions. The estimator $\widehat{\beta}_{gir}$ minimises the error sum of squares within the t dimensional subspace spanned by $\widetilde{V} := (v_{.1}|\ldots|v_{.t})$. The estimator is *conditionally* unbiased. That is, if X^tX is of rank t, then it is unbiased. If X^tX is of rank greater than t, then the estimator is biased.

5.3 Principal Component Regression

In many situations, X^tX may be invertible, but some eigenvalues may be small. This can lead to instability since, for the estimator $\widehat{\beta} = (X^tX)^{-1}X^tY$, the covariance is $\text{Cov}(\widehat{\beta}) = \sigma^2(X^tX)^{-1}$, so low eigenvalues of X^tX can lead to a large variance for some $v^t\widehat{\beta}$ where v is a unit vector.

One attempt to deal with this is the so-called *Principal Component* regression, *PC regression.*

Firstly, ‘principal component’ needs to be defined. For random variables, the covariance matrix Σ of a p-variate random vector $X = (X_1, \ldots, X_p)^t$ can be decomposed as: $\Sigma = P\Lambda P^t$, where P is an orthonormal $p \times p$ matrix and Λ is a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ where $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Let $Y = P^tX$, then $\text{Cov}(Y) = \Lambda$. The random variables $(Y_1, \ldots, Y_p)$ are the *principal components* of X. Note that

$$Y_j = \sum_{k=1}^{p} X_k P_{kj}.$$

Now consider an $n \times p$ data matrix X, with n p-variate observations. The empirical covariance matrix is the matrix S with entries:

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \overline{X}_{.i})(X_{kj} - \overline{X}_{.j}).$$

A decomposition into principal components can be carried out on a covariance matrix:

$$S = VDV^t$$

where V is orthonormal, $D = \text{diag}(d_1, \ldots, d_p)$ arranged in decending order: $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$. Let

$$Z = XV$$

then Z is an $n \times p$ data matrix with empirical covariance D.

To carry out a PC regression, we first *centre* Y and X by subtracting the column means from each entry. These are stored; it implies that, in the analysis, for a model

$$Y = \beta_0 + \sum_{j=1}^{r}\beta_j x_j + \epsilon,$$

$\beta_0 = 0$. Let S_{XX} denote the covariance matrix for X. After centring, $X^tX = (n-1)S_{XX}$.

The technique of PCR is described and it is shown to be equivalent to using $\widehat{\beta} = \widetilde{V}\widetilde{\Lambda}^{-1}\widetilde{V}^tX^tY$ when the number of eigenvalues used corresponds to the number of principal components.

Firstly, having centred both Y and X, make the singular value decomposition $X^tX = V\Lambda V^t$ where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ and $\lambda_1 \geq \ldots \geq \lambda_r \geq 0$. Take the t largest eigenvalues, where t is chosen according to a criterion for example $\frac{\sum_{j=1}^{t}\lambda_j}{\sum_{j=1}^{r}\lambda_j} \geq 0.9$ (the principal components account for over 90% of the variation) and let $\widetilde{V} = (v_{.1}|\ldots|v_{.t})$, the first t columns of V.

Let $Z = X\widetilde{V}$. The entries of Z are the scores of the first t principal components of X.

A PC regression is carried out by regressing Y on Z produced in this way (rather than X). The estimated regression coefficients for the t principal components are:

$$\widehat{\gamma}_{\text{pcr}} = (Z^tZ)^{-1}Z^tY.$$

Using $\widetilde{V}^t\widetilde{V} = I$ and $\widetilde{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_t)$ (only the restricting to non-zero eigenvalues) gives:

$$\widehat{\gamma}_{\text{pcr}} = (\widetilde{V}^tX^tX\widetilde{V})^{-1}\widetilde{V}^tX^tY = (\widetilde{V}^tV\Lambda V^t\widetilde{V})^{-1}\widetilde{V}^tXY = \widetilde{\Lambda}^{-1}\widetilde{V}^tX^tY.$$

Therefore, in the situation where $\lambda_1 \geq \ldots \geq \lambda_t > 0$ and $\lambda_{t+1} = \ldots = \lambda_r = 0$:

$$\widehat{\beta}_{\text{gir}} = \widetilde{V}\widehat{\gamma}_{\text{pcr}}.$$

The fitted values turn out to be exactly the same for PCR and for GIR when the Moore-Penrose generalised inverse of rank t is used:

$$\widehat{Y}_{\text{pcr}} = Z\widehat{\gamma}_{\text{pcr}} = X\widetilde{V}(\Lambda^{-1}\widetilde{V}^t X^t Y) = X\widehat{\beta}_{\text{gir}} = \widehat{Y}_{\text{gir}}.$$

Having computed $\widehat{\gamma}_{\text{pcr}}$, the coefficients when regression is performed on the transformed variables, it is usual to transform them into coefficients of the original input variables. Given the t vector $\widehat{\gamma}^{(t)}_{\text{pcr}}$, set

$$\widehat{\beta}_{\text{pcr}} = \widetilde{V}\widehat{\gamma}_{\text{pcr}}.$$

In practise, the rank of $X^t X$ and hence the number of components is an unknown *metaparameter* to be determined from the data. The number of components may be determined (for example) by Kaiser's criterion.

Warning PCR attempts to related Y to X when there is severe collinearity. It may fail dramatically. The extraction of the principal components of X is made without any reference to Y. There is therefore no reason to suppose that Y is highly correlated with any of the principal components selected; indeed, Y may have its highest correlation with the components that have been dropped from the analysis.

5.3.1 Shrinkage Methods

In this notation, the biased estimators considered here are all of the form

$$\widehat{\beta} = V(F\Lambda^{-1})V^t X^t Y,$$

where $F = \text{diag}(f_1, \ldots, f_r)$ f_j is the jth shrinkage factor, with the convention that if $f_j = 0$ then $\lambda_j^{-1} f_j = 0$. For example, for a t component PCA regression, $f_j = 1$ for $j = 1, \ldots, t$ and $f_j \lambda_j^{-1} = 0$ for $j \geq t + 1$.

5.4 Partial Least Squares Regression

Partial Least Squares Regression (PLSR) is an attempt to deal with the deficiency of PCR. Factors are constructed out of the regressor variables, which are useful for predicting Y, the response. Therefore, while PCR only uses the X variables to construct factors, PLSR uses both X and Y to determine the factors.

PLSR describes a family of techniques; PLSR is usually obtained by an algorithm rather than an optimisation procedure. The result is a sequence of prediction models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ which give increasingly accurate predictions of the output variable Y. The 'best' PLSR model is the one that minimises a cross-validation estimate of the prediction error.

The question of whether or not the cross-validation criterion selects the best model and in what sense is still an open problem. The PLSR algorithm of Wold, Martens, and Wold (2002) is now given. In the following, variances and covariances refer to the statistical variances and covariances, which are easily calculated when the columns are centred.

The following algorithm reduces the regression problem to a series of *simple* regression problems, choosing the 'best' one-dimensional regressor (a linear combination of the variables availble) at each stage.

Recall that, for a simple linear regression problem $y_j = \alpha + \beta x_j + \epsilon$,

$$X^t X = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & n\overline{x^2} \end{pmatrix}$$

so that

$$(X^t X)^{-1} = \frac{1}{n\text{Var}(x)} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

Recall, for OLS regression,

$$\widehat{\beta} = (X^t X)^{-1} X^t Y$$

and for simple linear regression, this reduces to:

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \frac{1}{\text{Var}(x)} \begin{pmatrix} \overline{x^2}\overline{y} - \overline{x}\overline{xy} \\ \text{Cov}(x,y) \end{pmatrix}.$$

Here $\text{Var}(x) = \frac{1}{n}\sum_{j=1}^n (x_j - \overline{x})^2$ and $\text{Cov}(x,y) = \frac{1}{n}\sum_{j=1}^n (x_j - \overline{x})(y_j - \overline{y})$.
If the variables have been centred, then $\alpha = 0$ and the estimate of β may be expressed as:

$$\widehat{\beta} = \frac{\text{Cov}(x,y)}{\text{Var}(x)}.$$

At each stage of PLSR, we choose a linear combination of the columns of X which is orthogonal to previous choices, which has maximum correlation with the response variable Y. The algorithm proceeds as follows:

1. Standardise the columns of the $n \times r$ design matrix X so that each column has mean 0 and standard deviation 1. Let $\overline{y} = \frac{1}{n}\sum_{j=1}^n Y_j$ (average before centring). Centre the n-vector Y (the response vector) by subtracting $\overline{y}$ from each entry, so that it has mean 0. Set $X^{(0)} = X$ (after centring and standardising) and $Y^{(0)} = Y$ (after centring).

   Note Since the columns have been standardised, $\text{Var}(X_{.i}) = 1$ for each column $i = 1, \ldots, r$.

2. For $k = 1, \ldots, t$

- For $j = 1, \ldots, r$ *regress* $Y^{(k-1)}$ on $X_{.j}^{(k-1)}$ (simple linear regression) to get the OLS regression coefficient:

  $$\widehat{\beta}_{k-1,j} = \frac{\text{Cov}(X_{.j}^{(k-1)}, Y^{(k-1)})}{\text{Var}(X_{.j}^{(k-1)})}$$

  Since we have centred the columns, for vectors x and y, $\text{Cov}(x, y) = \frac{1}{n}\sum x_j y_j = \frac{1}{n}x^t y$ and $\text{Var}(x) = \frac{1}{n}x^t x = \frac{1}{n}\sum x_j^2$.

- Having carried out the *simple* linear regression for each of the covariates, we take a suitable convex combination of these. We compute the weighted average

  $$Z_k = \sum_{j=1}^{r} w_{k-1,j}\widehat{\beta}_{k-1,j}X_{.j}^{(k-1)}$$

  as a *predictor* of $Y^{(k-1)}$ where we choose weights $w_{k-1,j} \propto \text{Var}(X_{.j}^{(k-1)})$. That is, we take the relative value, or weight, of predictor variable j proportional to the variance of $X_{.j}^{(k-1)}$ Thus:

  $$Z_k \propto \sum_{j=1}^{r} \left(\text{Cov}(X_{.j}^{(k-1)}, Y^{(k-1)})\right) X_{.j}^{(k-1)}.$$

  Since the columns of X have been centred, the covariance is simply the inner product: for a centred n-vector x and an n-vector y, $\text{Cov}(x, y) := \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.

- Regress $Y^{(k-1)}$ on Z_k to get the OLS regression coefficient

  $$\widehat{\theta}_k = \frac{\text{Cov}(Z_k, Y^{(k-1)})}{\text{Var}(Z_k)}$$

  and the residual vector

  $$Y^{(k)} := Y^{(k-1)} - \widehat{\theta}_k Z_k.$$

- For $j = 1, \ldots, r$ regress $X_{.j}^{(k-1)}$ on Z_k to get the OLS regression coefficient:

  $$\widehat{\phi}_{kj} = \frac{\text{Cov}(Z_k, X_{.j}^{(k-1)})}{\text{Var}(Z_k)}$$

  and residual vector

  $$X_{.j}^{(k)} := X_{.j}^{(k-1)} - \widehat{\phi}_{kj} Z_k.$$

  Stop when $\text{Cov}(X_{.j}^{(k)}, Y^{(k)}) = 0$ for each $j = 1, \ldots, r$.

3. The PLSR function fitted with t components is, therefore, given by:

$$\widehat{Y}_{plsr}^{(t)} = \overline{y}\mathbf{1}_n + \sum_{k=1}^{t} \widehat{\theta}_k Z_k.$$

Empirical studies strongly suggest that PLSR gives better results than PCR; for similar prediction accuracy, fewer components are needed and that both PCR and PLSR are substantially better than OLS when there is ill conditioning in the X matrix.