Contents

1
Linear Models

In this chapter we assume that the reader is familiar with the basic concepts of linear algebra as given, for example, in Strang (2003). Acquaintance with simple linear regression and concepts such as correlations and covariance matrices, and the multivariate normal distribution is also assumed. Such background material can be found in many introductory statistics books, such as Ross (2004).

1.1 The model

Consider the random vector $(Y, \mathbf{X})$ where $Y \in \mathbb{R}$ and $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$. We model Y as a function of $\mathbf{X}$, perturbed by noise. The goal is to understand the relation between Y and $\mathbf{X}$, and in particular be able to predict Y from $\mathbf{X}$. The distribution of $\mathbf{X}$ is often not of interest, so we begin by considering $\mathbf{X}$ fixed and later condition on it. When $\mathbf{X}$ is fixed we denote its value by $\mathbf{x}$.

Given $\mathbf{x}$, the general additive model is

$$Y = r(\mathbf{x}) + \varepsilon, \tag{1.1}$$

where $r : \mathbb{R}^p \to \mathbb{R}$ and the *error* ε satisfies $\mathrm{E}(\varepsilon) = 0$. The function $r(\mathbf{x})$ is called the *regression function*, Y is called the *dependent variable*, and the components of $\mathbf{x} = (x_1, \ldots, x_p)$ are called *covariates*, predictors, *independent variables* or features in the computer science literature. A typical example considered in the econometrics literature takes Y to be income, and the components of $\mathbf{x}$ to be the number of years of education, years

of experience, age, and other such factors deemed relevant predictors of income. In biostatistics, Y could be a health related measurement such as systolic blood pressure, or the indicator of a disease, and the covariates could be, for example, various blood test measurements.

The best mean squared error predictor of Y under (1.1) is $r(\mathbf{x})$, that is,

$$r(\mathbf{x}) = \arg\min_{g} \mathrm{E}(Y - g(\mathbf{x}))^2, \tag{1.2}$$

where the minimization is over all functions $g(\mathbf{x})$, assuming $\mathrm{E}Y^2$ exists. This fact is equivalent to the identity

$$\arg\min_{a\in\mathbb{R}} \mathrm{E}(Y - a)^2 = \mathrm{E}Y, \tag{1.3}$$

see Exercise 1.11.1.

A statistical problem is formulated by assuming the function $r(\mathbf{x})$ in (1.1) belongs to some known class of functions $\mathcal{R}$. If this class is finite-dimensional, then we may write $\mathcal{R} = \{r(\mathbf{x}, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^p$ is the parameter space. If in addition the distribution of ε is in some finite dimensional parametric class of distributions such as the normal $\mathcal{N}(0, \sigma^2)$, then the model is said to be parametric. If $\mathcal{R}$ is infinite-dimensional and ε is not restricted to a finite-dimensional parametric family, the model is considered non-parametric. Otherwise, the model is semiparametric. See Section ref.

In this chapter we consider the *linear model*

$$r(\mathbf{x}, \boldsymbol{\beta}) = \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}\boldsymbol{\beta}, \tag{1.4}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathsf{T}} \in \mathbb{R}^p$, a column vector of real parameters. Linearity refers to the coefficients β_i, and $r(\mathbf{x})$ is not restricted to be linear in $\mathbf{x}$; in (1.4) we can have, for example, a quadratic model like $r(x) = \beta_1 + \beta_2 x + \beta_3 x^2$, by taking $\mathbf{x} = (x_1, x_2, x_3)$ with $x_1 = 1$, $x_2 = x$, and $x_3 = x^2$. When the linear model (1.4) contains more than one covariate the term *multiple regression* is often used.

We observe n pairs $(Y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ satisfying (1.1) and (1.4), where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. Specifically, we have

$$Y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \ \text{ with } \ \mathrm{E}(\varepsilon_i) = 0, \quad i = 1, \ldots, n. \tag{1.5}$$

The main focus is inference on $\boldsymbol{\beta}$, and prediction of new Y values given their corresponding covariate vectors $\mathbf{x}$. When the covariates are properly scaled, a large component β_ℓ of $\boldsymbol{\beta}$ suggests that the corresponding covariate x_ℓ is strongly related to Y.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathsf{T}}$ be the column vector consisting of the dependent variables, let $\mathbb{X} = [x_{ij}]$ be the $n \times p$ matrix having entries x_{ij}, and let $\boldsymbol{\varepsilon}$ be the column vector of errors. The matrix $\mathbb{X}$ is known as the *design matrix*. We can rewrite the model (1.5) for the whole sample in matrix notation as

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{with} \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \tag{1.6}$$

or equivalently $\mathrm{E}(\mathbf{Y}) = \mathbb{X}\boldsymbol{\beta}$. Various additional assumptions on the ε_i's will appear; they may be uncorrelated, have a common variance, be identically distributed and normally distributed.

Simple linear regression is the special case where $\mathbb{X}$ is $n \times 2$, with first column given by $(1, \ldots, 1)^{\mathsf{T}}$, and second column by $(x_1, \ldots, x_n)^{\mathsf{T}}$, so that (1.6) becomes

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \ldots, n; \tag{1.7}$$

thus $r(x_i) = \beta_1 + \beta_2 x_i$ is a function of the single variable x_i.

A famous example of regression was the study of the relation between the heights of 1,078 fathers and their sons by the 19th-century scientist Sir Francis Galton. He noticed that fathers who were taller or shorter than the average father tended to have sons who likewise were taller or shorter than the average son, but to a lesser degree than their fathers. Galton called this phenomenon *regression to the mean*.

To provide a formal explanation of this phenomenon, let (Y_i, x_i) denote the heights of the son and father in the i^{th} pair. Writing the intercept in (1.7) as $\beta_1 - \beta_2\overline{x}$ we obtain the alternate parametrization of the model

$$Y_i = \beta_1 + \beta_2(x_i - \overline{x}) + \varepsilon_i, \;\; \varepsilon_i \quad i = 1, \ldots, n, \tag{1.8}$$

where $\overline{x} = n^{-1}\sum_{i=1}^{n} x_i$. In this way we see that β_2 expresses the contribution of a father's deviation $x_i - \overline{x}$ from the mean, to his son's height, Y_i. Taking averages in (1.8) we obtain

$$Y_i - \overline{Y} = \beta_2(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}).$$

Hence, the deviation of the son's height from the average is, apart from a random, zero mean error, proportional to that of the father. If $0 < \beta_2 < 1$ then the son's deviation from his generation's mean is expected to be smaller than his father's. This phenomenon holds under certain conditions, see Exercise 1.11.2

For another example of a linear model, let Y_i measure a univariate response from two treatment groups, one of size m, and another of size $n-m$. In this case we may write

$$Y_i = \beta_1 + \varepsilon_i \quad i = 1, \ldots, m, \quad Y_i = \beta_2 + \varepsilon_i \quad i = m+1, \ldots, n, \tag{1.9}$$

so that β_j denotes the mean treatment response in group $j = 1, 2$. This model may be written in the form (1.6) by letting the matrix $\mathbb{X}$ have rows $\mathbf{x}_i = (1, 0)$ for $1 \le i \le m$ and $\mathbf{x}_i = (0, 1)$ for $m+1 \le i \le n$. Variables taking only the values 0 and 1 as above are called *dummy* or *indicator* variables. *Analysis of variance* models are constructed as linear models with dummy variables, while *analysis of covariance* concerns models with both dummy and continuous covariates.

Analysis of the linear model is made simple by casting it in the language of linear algebra. In particular, let

$$\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\} \tag{1.10}$$

denote the *range space* of the matrix $\mathbb{X}$ consisting of all vectors that can be achieved by multiplying $\mathbb{X}$ on the right by $\boldsymbol{\beta}$. It is easy to see that $\mathcal{C}(\mathbb{X})$ is the linear subspace spanned by forming all linear combinations of the columns of $\mathbb{X}$, hence $\mathcal{C}(\mathbb{X})$ may also be referred to as the *column space* of $\mathbb{X}$. Now the model (1.6) can be simply expressed as $\mathrm{E}(\mathbf{Y}) \in \mathcal{C}(\mathbb{X})$.

If the columns of the design matrix $\mathbb{X}$ are not linearly independent, that is, if some columns of $\mathbb{X}$ can be expressed as linear combinations of others, the model is said to exhibit *multicollinearity*. Letting the *null space* of $\mathbb{X}$ be given by

$$\mathcal{N}(\mathbb{X}) = \{\mathbf{b} : \mathbb{X}\mathbf{b} = 0\},$$

infinitely many vectors yield the same value of $\mathbb{X}\boldsymbol{\beta}$ when $\mathrm{rank}(\mathbb{X}) < p$, since for any $\mathbf{b} \in \mathcal{N}(\mathbb{X})$ we have $\mathbb{X}\boldsymbol{\beta} = \mathbb{X}(\boldsymbol{\beta}+\mathbf{b})$. In this case different values of $\boldsymbol{\beta}$ yield the same value of $\mathrm{E}(\mathbf{Y}) = \mathbb{X}\boldsymbol{\beta}$, and therefore generate the same distribution on the data. Hence $\boldsymbol{\beta}$ cannot be estimated, and so is not *identifiable*, see Chapter ???. Linear independence of the columns and identifiability of a parameter vector $\boldsymbol{\beta}$ can be achieved without changing $\mathcal{C}(\mathbb{X})$ by a suitable deletion of columns from $\mathbb{X}$. Henceforth, in this chapter unless explicitly stated otherwise, we assume that the columns of $\mathbb{X}$ are linearly independent, necessitating the condition that $n \geq p$, since $\mathrm{rank}(\mathbb{X}) \leq \min\{n, p\}$. In this case, the columns of $\mathbb{X}$ form a basis for $\mathcal{C}(\mathbb{X})$.

In many situations in science one has many more variables than observations. For example, DNA sequencing is often done on a small sample of n individuals, while the number p of DNA loci that form the covariates is very large, making p much greater than n. We deal with such situations in Chapter ref.

1.2 Linear spaces and projections

1.2.1 Least squares

For column vectors $\mathbf{u}$ and $\mathbf{v}$ define the inner product $\mathbf{u}^\mathsf{T}\mathbf{v} = \sum u_i v_i$, and the norm $||\mathbf{u}|| = (\mathbf{u}^\mathsf{T}\mathbf{u})^{1/2}$. Given an $n \times p$ matrix $\mathbb{X}$ and a column vector $\boldsymbol{\beta} \in \mathbb{R}^p$, the squared length $||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2$ of the vector $\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}$ in $\mathbb{R}^n$ is a measure of how well the parameter $\boldsymbol{\beta}$ predicts the observed $\mathbf{Y}$ though the given design matrix $\mathbb{X}$. The *least squares* approach to estimating $\boldsymbol{\beta}$ in the model (1.6) consists of finding a value of the parameter $\widehat{\boldsymbol{\beta}}$ that minimizes this squared distance, that is,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i\boldsymbol{\beta})^2. \tag{1.11}$$

If we were to predict the value Y_i from its covariates $\mathbf{x}_i$, the least squares approach now provides the estimator $\mathbf{x}_i\widehat{\boldsymbol{\beta}}$. Thus, the least squares approach

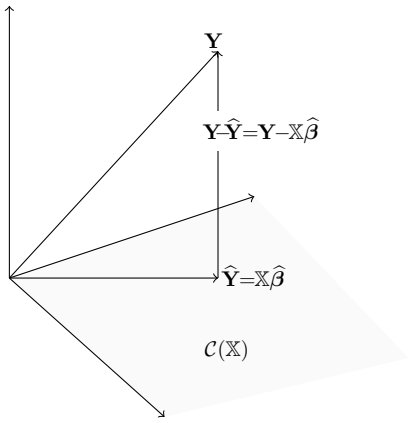

Figure 1.1. The normal equation (1.13) is depicted in this figure, where the gray xy plane is $\mathcal{C}(\mathbb{X})$, $\widehat{\mathbf{Y}}$ is the projection of $\mathbf{Y}$ on $\mathcal{C}(\mathbb{X})$, and $\mathbf{Y} - \widehat{\mathbf{Y}}$ is orthogonal to any vector in $\mathcal{C}(\mathbb{X})$.

minimizes the sum of the squared deviations of the predictors for the given data. More generally, when the unknown function r in the model (1.1) is of the parametric form $r(\mathbf{x}_i, \boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \Theta$, the least squares estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\Theta} \sum_{i=1}^{n} (Y_i - r(\mathbf{x}_i, \boldsymbol{\beta}))^2. \tag{1.12}$$

We now compute $\widehat{\boldsymbol{\beta}}$ of (1.11), starting with a geometric explanation, which we formalize later. As seen in Figure 1.1, the vector $\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}$ should be orthogonal to $\mathcal{C}(\mathbb{X})$, that is, to every vector in $\mathcal{C}(\mathbb{X})$. Writing $\mathbf{u} \perp \mathbf{v}$ to denote that the column vectors $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, that is, $\mathbf{u}^\mathsf{T}\mathbf{v} = 0$, and using (1.10), we thus require

$$\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}} \perp \mathbb{X}\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p;$$

equivalently,

$$\boldsymbol{\beta}^\mathsf{T}\mathbb{X}^\mathsf{T}\left(\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\right) = 0 \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p.$$

If for some vector $\mathbf{u}$ we have $\boldsymbol{\beta}^\mathsf{T}\mathbf{u} = 0$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$ then $\mathbf{u} = \mathbf{0}$, and hence we finally obtain the so called *normal equation:*

$$\mathbb{X}^\mathsf{T}\left(\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{0}, \quad \text{or equivalently} \quad \mathbb{X}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}^\mathsf{T}\mathbf{Y}. \tag{1.13}$$

Here "normal" is in the sense of orthogonal, rather than bearing any relation to the normal distribution. By Exercise 1.11.3, when $\mathbb{X}$ has linearly independent columns the matrix $\mathbb{X}^\mathsf{T}\mathbb{X}$ is invertible, so by the second identity

in (1.13) we obtain the explicit solution

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\mathbf{Y}. \tag{1.14}$$

Exercise 1.11.4 concerns the computation of $\widehat{\boldsymbol{\beta}}$ in the analysis of variance model (1.9). Exercise 1.11.6 shows that when the variations about the averages in the father and son's generations are equal, then the estimator $|\widehat{\beta_2}| \leq 1$. Exercise 1.11.7 shows that the normal equation (1.13) can be obtained by differentiation.

The following proposition verifies that the estimator $\widehat{\boldsymbol{\beta}}$ given by (1.14) has the least squares property (1.11).

Proposition 1.2.1 *The estimator $\widehat{\boldsymbol{\beta}}$ given by (1.14) is the unique vector in $\mathbb{R}^p$ that minimizes $||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2$.*

Proof: For an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^p$, by adding and subtracting we obtain

$$\begin{aligned}
||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2 &= ||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}} + \mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 \\
&= ||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2 + ||\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 + 2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbb{X}^T(\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) \\
&= ||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2 + ||\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 \\
&\geq ||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2.
\end{aligned}$$

The cross term in the second line vanishes by (1.13), showing that $\widehat{\boldsymbol{\beta}}$ minimizes $||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2$. That $\widehat{\boldsymbol{\beta}}$ is the unique minimizer follows from the fact that the inequality above is equality if and only if $\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = 0$, which implies $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ since $\mathbb{X}$ is full rank. ■

Using $\widehat{\boldsymbol{\beta}}$ given by (1.14), the vector $\widehat{\mathbf{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$ is the *projection* of $\mathbf{Y}$ onto $\mathcal{C}(\mathbb{X})$, which can be expressed as $\mathbb{P}\mathbf{Y}$ where

$$\mathbb{P} = \mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}. \tag{1.15}$$

It is easy to verify that the $n \times n$ matrix $\mathbb{P}$ satisfies

$$\mathbb{P}^{\mathsf{T}} = \mathbb{P} \quad \text{and} \quad \mathbb{P}^2 = \mathbb{P}. \tag{1.16}$$

Any matrix $\mathbb{P}$ that satisfies (1.16) is called an orthogonal projection matrix, or simply, a *projection matrix*. The matrix $\mathbb{P}$ in (1.15) is the projection matrix onto $\mathcal{C}(\mathbb{X})$, that is, the matrix satisfying (1.16) having $\mathcal{C}(\mathbb{P}) = \mathcal{C}(\mathbb{X})$.

The following lemma records an important property of $\mathbb{P}$, and its complementary projection $\mathbb{N}$, defined below.

Lemma 1.2.2 *Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ and $\mathbb{P} = \mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}$. Then $\mathbb{N} := \mathbb{I} - \mathbb{P}$ is a projection matrix, and*

$$\operatorname{tr}(\mathbb{P}) = p \quad \textit{and} \quad \operatorname{tr}(\mathbb{N}) = n - p,$$

that is, the trace of a projection matrix equals the dimension of the subspace upon which it projects.

Proof: It is easily verified that $\mathbb{N}$ is symmetric and satisfies $\mathbb{N}^2 = \mathbb{N}$, and hence is a projection. Now using the cyclic invariance $\mathrm{tr}(\mathbb{AB}) = \mathrm{tr}(\mathbb{BA})$ to reorder the matrices inside the trace, $\mathrm{tr}(\mathbb{P}) = \mathrm{tr}(\mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}) = \mathrm{tr}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{X}) = \mathrm{tr}(\mathbb{I}) = p$. It follows that $\mathrm{tr}(\mathbb{N}) = n - p$. ∎

For uniqueness of $\mathbb{P}$ and further properties of projections, see Exercise 1.11.8.

Proposition 1.2.3 below shows that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ and provides its variance. The proof is straightforward, see Exercise 1.11.9.

Proposition 1.2.3 *Let $\widehat{\boldsymbol{\beta}}$ be given by (1.14). Then,*

1. If $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ *then* $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

2. If $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{I}$ *then* $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$.

The condition $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{I}$ means that the errors have equal variances and are uncorrelated. Equality of variance is known as *homoscedasticity*.

If the columns of $\mathbb{X}$ are linearly dependent, then $\mathbb{X}^\mathsf{T}\mathbb{X}$ is singular, and hence its determinant will be zero. If the columns are nearly linearly dependent, then the matrix $\mathbb{X}^\mathsf{T}\mathbb{X}$ will have a small determinant, and hence $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$ will have large components, indicating that the estimation of $\boldsymbol{\beta}$ will be imprecise. On the other hand, the projection $\widehat{\mathbf{Y}}$, considered as a prediction of $\mathbf{Y}$, is the same for all design matrices whose range is $\mathcal{C}(\mathbb{X})$.

1.2.2 Generalized least squares

The sum of squares that is minimized in (1.11) assigns equal weights to the squared deviations $(y_i - \mathbf{x}_i\boldsymbol{\beta})^2$. This case is called *ordinary least squares* (OLS). If the errors ε_i in (1.6) are uncorrelated but have unequal variances, then intuitively one should assign higher weight to the more 'reliable' observations, that is, those with smaller variance. In fact, will see in (1.20) that when observations are uncorrelated their weights should be inversely proportional to their variances.

To take into account correlated errors, consider the model

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{with} \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{V}, \tag{1.17}$$

where $\mathbb{V}$ is a known $n \times n$ positive definite matrix. Note that when the observations are generated by taking linear combinations or averages of uncorrelated data values with unknown, constant variance, then their correlation structure, $\mathbb{V}$, will be known, though σ^2 may not be.

As $\mathbb{V}$ is positive definite there exists a non-singular symmetric $n \times n$ matrix $\mathbb{U}$ such that $\mathbb{V} = \mathbb{U}\mathbb{U}^\mathsf{T}$. Multiplying on the left by $\mathbb{U}^{-1}$ in (1.17), and letting $\mathbf{Z} = \mathbb{U}^{-1}\mathbf{Y}$, we obtain

$$\mathbf{Z} = \mathbb{W}\boldsymbol{\beta} + \boldsymbol{\eta} \quad \text{where} \quad \mathbb{W} = \mathbb{U}^{-1}\mathbb{X} \quad \text{and} \quad \boldsymbol{\eta} = \mathbb{U}^{-1}\boldsymbol{\varepsilon}, \tag{1.18}$$

We now have $\mathrm{Var}(\boldsymbol{\eta}) = \mathrm{Var}(\mathbb{U}^{-1}\boldsymbol{\varepsilon}) = \sigma^2\mathbb{U}^{-1}\mathbb{V}\mathbb{U}^{-\mathsf{T}} = \sigma^2\mathbb{I}$, where $\mathbb{U}^{-\mathsf{T}} = (\mathbb{U}^{-1})^{\mathsf{T}}$.

Applying the least squares principle to the model (1.18) we compute $\boldsymbol{\beta}$ that minimizes the sum of squares

$$||\mathbf{Z} - \mathbb{W}\boldsymbol{\beta}||^2 = (\mathbf{Y} - \mathbb{X}\boldsymbol{\beta})^{\mathsf{T}}\mathbb{V}^{-1}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}), \tag{1.19}$$

which is similar to $||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2$ considered in (1.11), but weighted by the factor $\mathbb{V}^{-1}$. This approach is known as *generalized least squares*(GLS). In the special case that $\mathbb{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, the diagonal matrix with the variances of the observations on its diagonal, (1.19) reduces to

$$||\mathbf{Z} - \mathbb{W}\boldsymbol{\beta}||^2 = \sum_{i=1}^{n}(Y_i - \mathbf{x}_i\boldsymbol{\beta})^2/\sigma_i^2, \tag{1.20}$$

a sum of squares with weights inversely proportional to the variances.

For the model (1.17), calculating the least squares estimator of $\boldsymbol{\beta}$, that is, the vector $\widehat{\boldsymbol{\beta}}$ that minimizes (1.19), by (1.14), results in

$$\widehat{\boldsymbol{\beta}} = (\mathbb{W}^{\mathsf{T}}\mathbb{W})^{-1}\mathbb{W}^{\mathsf{T}}\mathbf{Z} = (\mathbb{X}^{\mathsf{T}}\mathbb{V}^{-1}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\mathbb{V}^{-1}\mathbf{Y}. \tag{1.21}$$

Estimators of the form (1.21) are known as *generalized least squares estimators*. It is easy to see that this estimator is unbiased for $\boldsymbol{\beta}$ in the model (1.17).

1.3 The Gauss-Markov Theorem

Linear estimators based on the data $\mathbf{Y}$ are of the form $\mathbb{C}\mathbf{Y}$ for some matrix $\mathbb{C}$ that may depend on $\mathbb{X}$. Estimators (1.14) and (1.21) are linear and are also unbiased for $\boldsymbol{\beta}$ whenever the error vector $\boldsymbol{\varepsilon}$ has mean zero. The Gauss-Markov Theorem below provides the minimal variance estimator among all linear estimators of linear combinations of $\boldsymbol{\beta}$ in the model

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{with} \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{I} \quad \text{and} \quad \mathbb{X} \in \mathbb{R}^{n\times p}. \tag{1.22}$$

Theorem 1.3.1 (Gauss-Markov) *Let model* (1.22) *hold, and suppose that for given row vectors* $\mathbf{a} \in \mathbb{R}^p$ *and* $\mathbf{c} \in \mathbb{R}^n$ *the linear estimator* $\mathbf{cY}$ *satisfies* $\mathrm{E}(\mathbf{cY}) = \mathbf{a}\boldsymbol{\beta}$. *Then* $\mathrm{Var}(\mathbf{cY}) \geq \mathrm{Var}(\mathbf{a}\widehat{\boldsymbol{\beta}})$.

In words, $\mathbf{a}\widehat{\boldsymbol{\beta}}$ has the smallest variance among linear unbiased estimators of $\mathbf{a}\boldsymbol{\beta}$.

Proof: We have $\mathbf{a}\boldsymbol{\beta} = \mathrm{E}(\mathbf{cY}) = \mathbf{c}\mathbb{X}\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, implying $\mathbf{a} = \mathbf{c}\mathbb{X}$. Therefore

$$\begin{aligned}\frac{1}{\sigma^2}\mathrm{Var}(\mathbf{a}\widehat{\boldsymbol{\beta}}) = \mathbf{a}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbf{a}^{\mathsf{T}} &= \mathbf{c}\mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\mathbf{c}^{\mathsf{T}} \\ &= ||\mathbb{P}\mathbf{c}^{\mathsf{T}}||^2 \leq ||\mathbf{c}^{\mathsf{T}}||^2 = \frac{1}{\sigma^2}\mathrm{Var}(\mathbf{cY}),\end{aligned}$$

where the third equality uses the fact that $\mathbb{P} = \mathbb{P}^{\mathsf{T}}\mathbb{P}$, and the inequality follows from Exercise 1.11.8 part *6*, which says that projection decreases length. ∎

This result says that if we use $\mathbf{x}\widehat{\boldsymbol{\beta}}$ to predict a new observation Y having covariate vector $\mathbf{x}$, then our prediction has minimal variance over all predictors of the form $\mathbf{cY}$ that have expectation $E(Y)$ that is, $\mathbf{x}\boldsymbol{\beta}$.

Corollary 1.3.2 *For the model of* (1.22), $\mathrm{Var}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}})$ *is nonnegative definite for any linear unbiased estimator* $\widetilde{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$.

The proof of this corollary is left as Exercise 1.11.10.

A more general result holds. In the model (1.17), the estimator (1.21) has the minimal variance matrix among linear unbiased estimators. The proof is left to the reader in Exercise 1.11.11.

1.4 Design matrices with dependent columns

We briefly discuss the case where the matrix $\mathbb{X}$ in the model (1.22) is not of full rank, that is, its columns are not linearly independent. For a situation where a non full rank design matrix appears naturally, see Exercise 1.11.12. When $\mathbb{X}$ is not of full rank $\boldsymbol{\beta}$ is not unique and hence not identifiable, and therefore cannot be estimated. However, the expectation $\mathrm{E}(\mathbf{Y})$ or linear combinations thereof can be estimated. Regardless of whether $\mathbb{X}$ is full rank or not, the model (1.22) can be expressed in the form $\mathrm{E}(\mathbf{Y}) \in \mathcal{C}(\mathbb{X})$ and $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbb{I}$. By deleting suitable columns from $\mathbb{X}$, we can obtain a full rank matrix $\mathbb{W}$ such that $\mathcal{C}(\mathbb{W}) = \mathcal{C}(\mathbb{X})$. The matrix $\mathbb{P} = \mathbb{W}(\mathbb{W}^{\mathsf{T}}\mathbb{W})^{-1}\mathbb{W}^{\mathsf{T}}$ is the projection matrix onto $\mathcal{C}(\mathbb{X})$. The projection of any vector $\mathbf{v}$ onto $\mathcal{C}(\mathbb{X})$ is unique since if there were two distinct projections, that is, two vectors in $\mathcal{C}(\mathbb{X})$ that minimize the distance from $\mathbf{v}$, then by a straightforward convexity argument, their linear combination would be closer to $\mathbf{v}$, contradicting the assumption that the two vectors are projections. It follows that the projection matrix to $\mathcal{C}(\mathbb{X})$ is unique since its operation on all vectors is unique.

The following result considers estimation of linear combinations of $\mathrm{E}(\mathbf{Y})$, and applies even in the case where $\mathbb{X}$ is not full rank.

Theorem 1.4.1 *Consider model* (1.22) *with* $\mathbb{X}$ *not necessarily of full rank, and let* $\mathbf{a}$ *be a row vector in* $\mathbb{R}^n$, *and* $\widehat{\mathbf{Y}} = \mathbb{P}\mathbf{Y}$. *Then* $\mathbf{a}\widehat{\mathbf{Y}}$ *is the unbiased linear estimator of* $\mathbf{a}\mathrm{E}(\mathbf{Y})$ *with the smallest variance.*

The proof is left to Exercise 1.11.13.

1.5 Variance estimation

The next result concerns estimation of the variance σ^2 in the model (1.22).

Theorem 1.5.1 *Under the model* (1.22),

$$S^2 := \frac{1}{n-p}||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2 \tag{1.23}$$

is an unbiased estimator of σ^2.

Proof: First note that $\mathrm{E}(\mathbf{Y}) = \mathbb{X}\boldsymbol{\beta} = \mathrm{E}(\mathbb{X}\widehat{\boldsymbol{\beta}})$. Now, since $\mathbb{P}\mathbf{Y} = \widehat{\mathbf{Y}} = X\widehat{\boldsymbol{\beta}}$, we have

$$\mathrm{E}(||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2) = \mathrm{E}\{[(\mathbb{I}-\mathbb{P})\mathbf{Y}]^\mathsf{T}[(\mathbb{I}-\mathbb{P})\mathbf{Y}]\} = \mathrm{E}[\mathbf{Y}^\mathsf{T}(\mathbb{I}-\mathbb{P})^\mathsf{T}(\mathbb{I}-\mathbb{P})\mathbf{Y}].$$

The latter expression is a real number and hence equal to its trace. Applying $\mathrm{tr}(\mathbb{A}\mathbb{B}) = \mathrm{tr}(\mathbb{B}\mathbb{A})$ with $\mathbb{A} = \mathbf{Y}^\mathsf{T}(\mathbb{I}-\mathbb{P})^\mathsf{T}$ and $\mathbb{B} = (\mathbb{I}-\mathbb{P})\mathbf{Y}$, and the properties (1.16) of $\mathbb{P}$, we obtain

$$\mathrm{E}(||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2) = \mathrm{E}\{\mathrm{tr}[(\mathbb{I}-\mathbb{P})\mathbf{Y}\mathbf{Y}^\mathsf{T}(\mathbb{I}-\mathbb{P})^\mathsf{T}]\} = \sigma^2\mathrm{tr}(\mathbb{I}-\mathbb{P}) = \sigma^2(n-p),$$

with the final equality holding by Lemma 1.2.2. ■

1.6 Least squares estimation under linear constraints

Linear models often are considered with constraints and are of the form

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{with} \quad \mathbb{D}\boldsymbol{\beta} = \mathbf{c} \tag{1.24}$$

for a given matrix $\mathbb{D} \in \mathbb{R}^{q\times p}$ with q independent rows, and a vector $\mathbf{c} \in \mathbb{R}^q$. Estimation of $\boldsymbol{\beta}$ under the constraint (1.24) amounts to estimating $r(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ subject to knowing the precise value of the function at some given points, that is $r(\mathbf{d}_i) = c_i$, for $i = 1, \ldots, q$, where $\mathbf{d}_1, \ldots, \mathbf{d}_q$ are the rows of $\mathbb{D}$, and $\mathbf{c}^\mathsf{T} = (c_1, \ldots, c_q)$. Estimating $\boldsymbol{\beta}$ in the model (1.24) is also needed for testing whether the linear constraint holds. Exercise 1.11.14 discusses that case that the rows of $\mathbb{D}$ are not independent.

Theorem 1.6.1 provides the least squares estimator for the constrained linear model (1.24). See Exercise 1.11.15 for an illustration of its use with $\mathbf{c} \neq \mathbf{0}$. Examples with $\mathbf{c} = \mathbf{0}$ may arise when considering the regression function $\mathrm{E}(Y) = r(\mathbf{x})\boldsymbol{\beta} = \sum x_j\beta_j$, where we wish to estimate $\boldsymbol{\beta}$ under constraints like $\beta_i = \beta_j$ for some of the coefficients, that is equality of the contribution of the corresponding covariates to the response Y, or perhaps $\beta_i = 2\beta_j$, reflecting known constraints on their relative contribution.

Linear independence of the rows of $\mathbb{D}$ implies $q \leq p$. When estimating $\boldsymbol{\beta}$ under the constraint $\mathbb{D}\boldsymbol{\beta} = \mathbf{c}$, the case $q = p$ is trivial since then (1.24) is equivalent to $\boldsymbol{\beta} = \mathbb{D}^{-1}\mathbf{c}$, determining $\boldsymbol{\beta}$ uniquely.

Assuming $\mathrm{E}\mathbf{Y} \in \mathcal{C}(\mathbb{X})$, the least squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ was defined in (1.14). The constrained model (1.24) can be expressed as $\mathrm{E}(\mathbf{Y}) \in M_{\mathbf{c}}$ where $M_{\mathbf{c}} = \{\mathbb{X}\boldsymbol{\beta} : \mathbb{D}\boldsymbol{\beta} = \mathbf{c}\}$. We denote the least squares estimator of $\boldsymbol{\beta}$ in this model by $\widehat{\boldsymbol{\beta}}_{M_{\mathbf{c}}}$. The nearest point in $M_{\mathbf{c}}$ to a given $\mathbf{Y} \in \mathbb{R}^n$ exists and is unique, as $M_{\mathbf{c}}$ is closed and convex. By definition of $\widehat{\boldsymbol{\beta}}_{M_{\mathbf{c}}}$ this point is $\mathbb{X}\widehat{\boldsymbol{\beta}}_{M_{\mathbf{c}}}$, and $\widehat{\boldsymbol{\beta}}_{M_{\mathbf{c}}}$ is unique since $\mathbb{X}$ has full rank. We define $\mathbb{P}_{M_{\mathbf{c}}}$ by the relation $\mathbb{P}_{M_{\mathbf{c}}}\mathbf{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}_{M_{\mathbf{c}}}$, that is, $\mathbb{P}_{M_{\mathbf{c}}}$ represents the projection onto $M_{\mathbf{c}}$.

The projection matrix $\mathbb{P}_{M_{\mathbf{0}}}$ can be obtained explicitly from the proof of Theorem 1.6.1. The matrix $\mathbb{P}$ will denote the projection onto $\mathcal{C}(\mathbb{X})$ as usual. In the case $\mathbf{c} \neq \mathbf{0}$, $\mathbb{P}_{M_{\mathbf{c}}}$ is not a projection matrix, and, in fact, is affine, but not linear, that is, $P_{M_{\mathbf{c}}}\mathbf{Y} = \mathbb{A}\mathbf{Y} + \mathbf{d}$ for some matrix $\mathbb{A}$ and vector $\mathbf{d} \neq \mathbf{0}$. More precisely, from (1.25) below we have $\mathbb{P}_{M_{\mathbf{c}}}Y = \mathbb{P}_{M_{\mathbf{0}}}\mathbf{Y} + \mathbb{X}\mathbf{b} - \mathbb{P}_{M_{\mathbf{0}}}\mathbb{X}\mathbf{b}$ which is of the form $\mathbb{A}\mathbf{Y} + \mathbf{d}$ with $\mathbf{d} = \mathbb{X}\mathbf{b} - P_{M_{\mathbf{0}}}\mathbb{X}\mathbf{b}$, which vanishes if and only if $\mathbb{X}\mathbf{b} \in M_{\mathbf{0}}$, that is, if and only if $\mathbb{D}\mathbf{b} = \mathbf{0}$, and hence if and only if $\mathbf{c} = \mathbf{0}$.

The following theorem formally proves the relation (1.25) which is made geometrically clear in Figure ref. From (1.25) we then derive the least squares estimator of the constrained linear model.

Theorem 1.6.1 *Let $Y = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbb{X} \in \mathbb{R}^{n\times p}$ with $p < n$, having p independent columns, and let $E(\boldsymbol{\varepsilon}) = 0$. Let $\mathbb{D} \in \mathbb{R}^{q\times p}$ have q independent rows, and $\mathbf{b}$ be any solution to $\mathbb{D}\mathbf{b} = \mathbf{c}$. Then*

$$\mathbb{P}(\mathbf{Y} - \mathbb{X}\mathbf{b}) - \mathbb{P}_{\mathbf{M_0}}(\mathbf{Y} - \mathbb{X}\mathbf{b}) = \mathbb{P}\mathbf{Y} - \mathbb{P}_{\mathbf{M_c}}\mathbf{Y}. \tag{1.25}$$

and the constrained least squares estimator of $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}}_{M_{\mathbf{c}}} := \arg \min_{\boldsymbol{\beta}\,:\,\mathbb{D}\boldsymbol{\beta}=\mathbf{c}} ||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2,$$

is given by

$$\widehat{\boldsymbol{\beta}}_{M_{\mathbf{c}}} = \widehat{\boldsymbol{\beta}} + (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{D}^{\mathsf{T}}\left[\mathbb{D}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{D}^{\mathsf{T}}\right]^{-1}(\mathbf{c} - \mathbb{D}\widehat{\boldsymbol{\beta}}). \tag{1.26}$$

Proof: First consider the case $\mathbf{c} = \mathbf{0}$. Recall that $M_{\mathbf{0}} = \{\mathbb{X}\boldsymbol{\beta} : \mathbb{D}\boldsymbol{\beta} = \mathbf{0}\}$ and set $W = \mathcal{C}(\mathbb{X}) \cap M_{\mathbf{0}}^{\perp}$, where $M_{\mathbf{0}}^{\perp}$ denotes the orthogonal complement of $M_{\mathbf{0}}$. Forming $M_{\mathbf{0}}$ by imposing q linear constraints on $\mathcal{C}(\mathbb{X})$, a subspace of dimension p, leaves $M_{\mathbf{0}}$ with dimension $p - q$. The subspace W has dimension q, being the orthogonal complement of $M_{\mathbf{0}}$ in $\mathcal{C}(\mathbb{X})$.

We claim that $W = \mathcal{C}(\mathbb{G})$, where $\mathbb{G} = \mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{D}^{\mathsf{T}}$. First, note that a vector $\mathbf{u} \in \mathcal{C}(\mathbb{G})$ satisfies $\mathbf{u} = \mathbb{G}\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^q$. Such a vector is clearly in $\mathcal{C}(\mathbb{X})$ since $\mathbb{G}$ has the matrix $\mathbb{X}$ on its left side. In order to conclude that $\mathcal{C}(\mathbb{G}) \subseteq W$ it remains to show that $\mathbf{u} \in M_{\mathbf{0}}^{\perp}$, which is equivalent to showing that $\mathbf{u}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta} = 0$ provided $\mathbb{D}\boldsymbol{\beta} = \mathbf{0}$. We have

$$\mathbf{u}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\alpha}^{\mathsf{T}}\mathbb{G}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\alpha}^{\mathsf{T}}[\mathbb{D}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}]\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\alpha}^{\mathsf{T}}\mathbb{D}\boldsymbol{\beta} = 0.$$

To show that $\mathcal{C}(\mathbb{G}) = W$ we note that

$$q \geq r(G) \geq r(X'G) = r(D) = q,$$

where the first equality follows from $\mathcal{C}(G) \subseteq W$. Since $\mathcal{C}(G)$ has dimension q, the same dimension as W, the two subspaces must be equal.

As in (1.15), it follows that the projection matrix onto W is $\mathbb{P}_W = \mathbb{G}(\mathbb{G}^\mathsf{T}\mathbb{G})^{-1}\mathbb{G}^\mathsf{T}$, that is,

$$\mathbb{P}_W := \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\left[\mathbb{D}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\right]^{-1}\mathbb{D}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}. \tag{1.27}$$

As W and $M_\mathbf{0}$ are orthogonal complements in $\mathcal{C}(\mathbb{X})$, we have $\mathbb{P} = \mathbb{P}_W + \mathbb{P}_{M_\mathbf{0}}$, where $\mathbb{P} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$, the projection onto $\mathcal{C}(\mathbb{X})$ given by (1.15), and $\mathbb{P}_{M_\mathbf{0}}$ the projection to $M_\mathbf{0}$.

$$\begin{aligned}\mathbb{X}\widehat{\boldsymbol{\beta}}_{M_\mathbf{0}} = \mathbb{P}_{M_\mathbf{0}}\mathbf{Y} = (\mathbb{P} - \mathbb{P}_W)\mathbf{Y} &= \mathbb{P}\mathbf{Y} - \mathbb{P}_W\mathbf{Y} \\ &= \mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\left[\mathbb{D}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\right]^{-1}\mathbb{D}\widehat{\boldsymbol{\beta}}. \end{aligned}\tag{1.28}$$

As $\mathbb{X}$ has full rank, $\mathbb{X}\mathbf{u} = \mathbb{X}\mathbf{v}$ implies $\mathbf{u} = \mathbf{v}$, yielding (1.26) for the case $\mathbf{c} = \mathbf{0}$. Alternatively, one may multiply the equality above on the left by $(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$.

For the case where $\mathbf{c} \neq 0$, let $\mathbf{b}$ be any solution to the equation $\mathbb{D}\mathbf{b} = \mathbf{c}$. We have

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{M_\mathbf{c}} &= \arg\min_{\boldsymbol{\beta}\,:\,\mathbb{D}\boldsymbol{\beta}=\mathbf{c}} ||\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}||^2 = \arg\min_{\boldsymbol{\beta}\,:\,\mathbb{D}\boldsymbol{\beta}=\mathbf{c}} ||\mathbf{Y} - \mathbb{X}\mathbf{b} - \mathbb{X}(\boldsymbol{\beta} - \mathbf{b})||^\mathbf{2} \\ &= \arg\min_{\boldsymbol{\gamma}\,:\,\mathbb{D}\boldsymbol{\gamma}=\mathbf{0}} ||\mathbf{Y} - \mathbb{X}\mathbf{b} - \mathbb{X}\boldsymbol{\gamma}||^2 + \mathbf{b} = \widehat{\boldsymbol{\gamma}} + \mathbf{b}, \end{aligned}\tag{1.29}$$

where $\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\,:\,\mathbb{D}\boldsymbol{\gamma}=\mathbf{0}} ||\mathbf{Y} - \mathbb{X}\mathbf{b} - \mathbb{X}\boldsymbol{\gamma}||$, and where we have made the change of variable $\boldsymbol{\gamma} = \boldsymbol{\beta} - \mathbf{b}$ in the third equality, which shifts the argmin by $\mathbf{b}$. By definition $\mathbb{X}\widehat{\boldsymbol{\gamma}} = \mathbb{P}_{M_\mathbf{0}}(\mathbf{Y} - \mathbb{X}\mathbf{b})$, and now by multiplying (1.29) on the left by $\mathbb{X}$ we obtain

$$\mathbb{P}_{M_\mathbf{c}}\mathbf{Y} = P_{M_\mathbf{0}}(\mathbf{Y} - \mathbb{X}\mathbf{b}) + \mathbb{X}\mathbf{b}. \tag{1.30}$$

We obtain identity (1.25) by simply adding $\mathbb{P}\mathbf{Y}$ on both sides of the equation and using $\mathbb{P}\mathbb{X}\mathbf{b} = \mathbb{X}\mathbf{b}$. Using $\mathbb{P}_{M_\mathbf{0}} = \mathbb{P} - \mathbb{P}_W$ we obtain from (1.30)

$$\mathbb{X}\widehat{\boldsymbol{\beta}}_{M_\mathbf{c}} = (\mathbb{P} - \mathbb{P}_W)(\mathbf{Y} - \mathbb{X}\mathbf{b}) + \mathbb{X}\mathbf{b}. \tag{1.31}$$

Again using $\mathbb{P}\mathbb{X}\mathbf{b} = \mathbb{X}\mathbf{b}$, we obtain

$$\mathbb{X}\widehat{\boldsymbol{\beta}}_{M_\mathbf{c}} = \mathbb{P}\mathbf{Y} - \mathbb{P}_W\mathbf{Y} + \mathbb{P}_W\mathbb{X}\mathbf{b}. \tag{1.32}$$

Now $\mathbb{D}\mathbf{b} = \mathbf{c}$ implies $\mathbb{P}_W\mathbb{X}\mathbf{b} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\left[\mathbb{D}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\right]^{-1}\mathbf{c}$, and together with the last equality in (1.28), (1.32) becomes (1.26) multiplied by $\mathbb{X}$ on the left. Since $\mathbb{X}$ has full rank, (1.26) follows by arguing as after (1.28). ∎

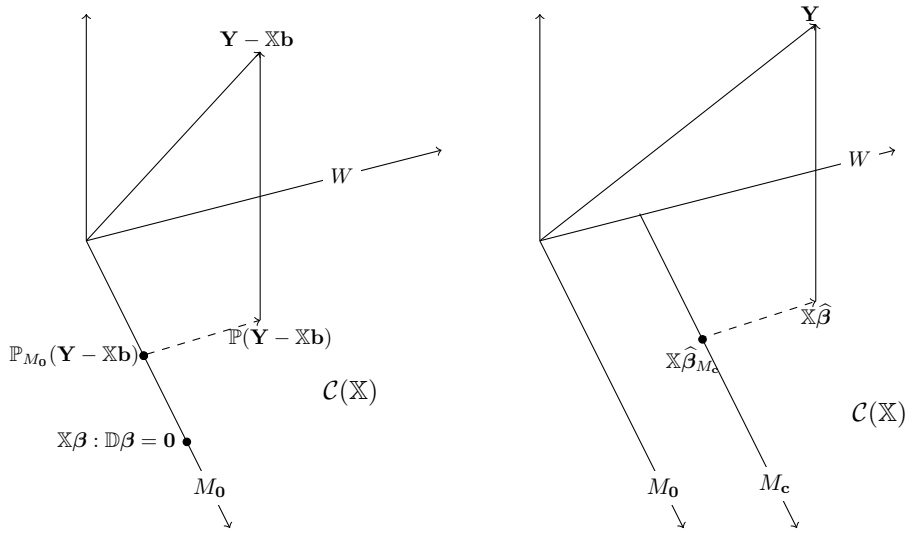

Figure 1.2. The quantities in the discussion of constrained estimation are depicted. Relations such as (1.25), (1.30), and (1.40) below can easily be seen from the diagram.

1.7 Normal errors

In previous sections only first and second moment assumptions are required, and in particular, no distributional form for the error vector $\boldsymbol{\varepsilon}$ was assumed. In order to provide distributional results for testing and estimation we now add assumptions on the distribution of the observations. More specifically, we will assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I})$, or equivalently that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and are independent, $i = 1, \ldots, n$.

1.7.1 Why normal?

To quantify the precision of estimates, or to test hypotheses with a given level of significance, one must make assumptions on the distribution of the data or resort to asymptotic results. Typically normality is assumed for a variety of reasons. One justification is that errors that arise, approximately, from an accumulation of many independent small errors, will be approximately normal by the Central Limit Theorem. Secondly, the normal assumption is mathematically convenient. Indeed, when choosing a model for data analysis, it is sometime better to choose a useful model, rather than a model that is perhaps more realistic, but too complicated.

Suppose the assumption of normal errors is invoked and used when it does not in fact hold. Subject to various conditions, such as those given in Section 1.9.2 below that the row vectors of the design matrix are i.i.d., it will be shown that the asymptotic distribution of estimators and test

statistics for linear models with non-normal error distributions have the same asymptotic distribution as that obtained under the assumption of normality. We will briefly discuss ways to test normality in Section 1.10.

1.7.2 Maximum likelihood estimators and their distribution

In order to provide distributional results for testing and estimation, we consider the following model:

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I}). \tag{1.33}$$

The likelihood function at $\mathbf{y} \in \mathbb{R}^n$ is given by the density

$$f(\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-||\mathbf{y}-\mathbb{X}\boldsymbol{\beta}||^2/2\sigma^2}. \tag{1.34}$$

It is easy to that $(\widehat{\boldsymbol{\beta}}, \|\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|^2/n)$ is the maximum likelihood estimator (MLE) of $(\boldsymbol{\beta}, \sigma^2)$, where $\widehat{\boldsymbol{\beta}}$ is the least squares estimator given by (1.14), see Exercise 1.11.16. The fact that the least squares estimator of $\boldsymbol{\beta}$ coincides with the maximum likelihood estimator (MLE), adds justification to the least squares approach. Hence, in addition to the optimality properties of least squares estimators given by the Gauss-Markov Theorem in Section 1.3, $\widehat{\boldsymbol{\beta}}$ also has further properties derived from the theory of maximum likelihood estimation. Replacing $\mathbb{X}\boldsymbol{\beta}$ in (1.34) by a general regression function $r(\mathbf{x}, \boldsymbol{\beta})$, in view of (1.12), we find maximum likelihood and least squares estimators coincide for any parametric regression function $r(\mathbf{x}, \boldsymbol{\beta})$, including non linear functions of $\mathbf{x}$ and $\boldsymbol{\beta}$ such as $\beta_1 x/(\beta_2 + x)$.

Proposition 1.7.1 *Let model* (1.33) *hold and let* $\widehat{\boldsymbol{\beta}}$ *and* S^2 *be given by* (1.14) *and* (1.23)*, respectively. Then*

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}) \quad \textit{and} \quad \frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p},$$

and the estimators $\widehat{\boldsymbol{\beta}}$ *and* S^2 *are independent.*

Proof: To compute the distribution of S^2, we first note

$$||\mathbf{Y} - \mathbb{X}\widehat{\beta}||^2 = ||(I - \mathbb{P})\mathbf{Y}||^2 = ||\mathbb{N}\mathbf{Y}||^2 = ||\mathbb{N}\boldsymbol{\varepsilon}||^2,$$

where in the second equality we used the notation $\mathbb{N} = \mathbb{I} - \mathbb{P}$, and in final equality applied (1.15) to obtain $\mathbb{P}\mathbb{X} = \mathbb{X}$, and therefore $\mathbb{N}\mathbb{X} = (\mathbb{I}-\mathbb{P})\mathbb{X} = \mathbb{O}$, where $\mathbb{O}$ is a matrix of zeros, and

$$\mathbb{N}\mathbf{Y} = \mathbb{N}\mathbb{X}\boldsymbol{\beta} + \mathbb{N}\boldsymbol{\varepsilon} = \mathbb{N}\boldsymbol{\varepsilon}. \tag{1.35}$$

As $\mathbb{N}$ is symmetric we may diagonalize it as $\mathbb{N} = \mathbb{T}^\mathsf{T}\mathbb{\Lambda}\,\mathbb{T}$, where $\mathbb{T}$ is orthogonal, that is, $\mathbb{T}^\mathsf{T}\mathbb{T} = \mathbb{T}\mathbb{T}^\mathsf{T} = \mathbb{I}$ and $\mathbb{\Lambda}$ is a diagonal matrix having the eigenvalues λ_i of $\mathbb{N}$ along its diagonal. By Exercise 1.11.18, since $\mathbb{N}^2 = \mathbb{N}$, all the eigenvalues $\mathbb{\Lambda}$ of $\mathbb{N}$ satisfy $\lambda^2 = \lambda$, and hence take on only the values

zero and one. As $\mathrm{tr}(\mathbb{N})$ is the sum of the eigevalues of $\mathbb{N}$, in this case it is also the number of eigenvalues equal to 1. By Lemma 1.2.2, $\mathrm{tr}(\mathbb{N}) = n - p$, so $\mathbb{\Lambda}$ has $n - p$ ones and p zeros along its diagonal. Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbb{I})$ implies $\mathbb{T}\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbb{T}\mathbb{T}^\mathsf{T}) = \mathcal{N}(0, \sigma^2\mathbb{I})$, we obtain $\mathbb{T}\boldsymbol{\varepsilon} \stackrel{d}{=} \boldsymbol{\varepsilon}$, yielding

$$||\mathbb{N}\boldsymbol{\varepsilon}||^2 = \boldsymbol{\varepsilon}^\mathsf{T}\mathbb{N}^\mathsf{T}\mathbb{N}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\mathsf{T}\mathbb{N}^2\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\mathsf{T}\mathbb{N}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\mathsf{T}\mathbb{T}^\mathsf{T}\mathbb{\Lambda}\,\mathbb{T}\boldsymbol{\varepsilon} \stackrel{d}{=} \boldsymbol{\varepsilon}^\mathsf{T}\mathbb{\Lambda}\,\boldsymbol{\varepsilon}.$$

In particular, as $\mathbb{\Lambda}$ is diagonal with $n - p$ ones and p zeros on its diagonal,

$$\frac{(n-p)S^2}{\sigma^2} \stackrel{d}{=} \frac{\boldsymbol{\varepsilon}^\mathsf{T}\mathbb{\Lambda}\,\boldsymbol{\varepsilon}}{\sigma^2} = \sum_{i=1}^{n} \lambda_i \left(\frac{\varepsilon_i}{\sigma}\right)^2 \stackrel{d}{=} \sum_{i=1}^{n-p} \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi^2_{n-p},$$

as claimed.

To show that $\widehat{\boldsymbol{\beta}}$ and S^2 are independent we note that

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \mathbb{N}\mathbf{Y}) = \mathrm{Cov}((\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbf{Y}, \mathbb{N}\mathbf{Y}) = \sigma^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{N}^\mathsf{T} = \mathbb{O}^\mathsf{T},$$

since $\mathbb{X}^\mathsf{T}\mathbb{N}^\mathsf{T} = (\mathbb{N}\mathbb{X})^\mathsf{T} = \mathbb{O}^\mathsf{T}$, as above (1.35). Since $(\widehat{\boldsymbol{\beta}}, \mathbb{N}\mathbf{Y})$ is a linear transformation of $\mathbf{Y}$ it has a multivariate normal distribution, and since the two components have covariance zero, they are independent. In particular $\widehat{\boldsymbol{\beta}}$ is independent of any function of $\mathbb{N}\mathbf{Y}$, such as S^2.

Using again that a linear transformation of a multivariate normal is again multivariate normal, to complete the distirbutional claim on $\widehat{\boldsymbol{\beta}}$ we only need to compute its mean and variance, for which we refer to Proposition 1.2.3. ∎

Here is a geometric explanation of the above result. First assume that $\mathcal{C}(\mathbb{X}) = H$, where $H = \mathbb{R}^p \times \{\mathbf{0}\}^{n-p}$, that is, the subspace of all vectors whose last $n - p$ coordinates are zero. In particular, as the final $n - p$ coordinates of $\mathbb{X}\boldsymbol{\beta}$ are zero, $Y_i = \varepsilon_i$ for $i = p + 1, \ldots, n$, and furthermore the projection $\mathbb{P}$ of $\mathbf{Y}$ onto $\mathcal{C}(\mathbb{X})$ simply sets the final $n - p$ coordinates of $\mathbf{Y}$ to zero, that is, $\mathbb{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}} = (Y_1, \ldots, Y_p, 0, \ldots, 0)^\mathsf{T}$. Hence $||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2 = ||\mathbf{Y} - \widehat{\mathbf{Y}}|| = \sum_{i=p+1}^{n} Y_i^2 = \sum_{i=p+1}^{n} \varepsilon_i^2$, implying $\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$ as well as independence of S^2 and $\widehat{\boldsymbol{\beta}}$ or equivalently $\widehat{\mathbf{Y}}$.

When $\mathcal{C}(\mathbb{X})$ is an arbitrary p-dimensional space, it can always be rotated to coincide with H. Formally, the required rotation can be accomplished by an orthogonal matrix $\mathbb{T}$ that transforms the columns of $\mathbb{X}$ to a basis for H, that is, by a matrix $\mathbb{T}$ such that $\mathcal{C}(\mathbb{T}\mathbb{X}) = H$.

In order to construct such a matrix recall that the row space of $\mathbb{X}$ has dimension p. Hence the null space $\{\mathbf{t} \in \mathbb{R}^n : \mathbf{t}\mathbb{X} = 0\}$ has dimension $n - p$, and for any $(n - p) \times n$ matrix $\mathbb{V}$ whose rows form an orthonormal basis to this space we must have $\mathbb{V}\mathbb{X} = \mathbb{O}$. Now let $\mathbb{T}$ be the $n \times n$ matrix whose last $n - p$ rows are those of $\mathbb{V}$, and whose first p rows are vectors $\mathbf{t}_i \in \mathbb{R}^n,\ i = 1, \ldots, p$ which complete the rows of $\mathbb{V}$ to an orthonormal basis of $\mathbb{R}^n$. As $\mathbb{T}$ is full rank, the space $\mathcal{C}(\mathbb{T}\mathbb{X})$ has dimension p. But now the relation $\mathbb{V}\mathbb{X} = \mathbb{O}$, saying that the last $n-p$ coordinates of vectors in $\mathcal{C}(\mathbb{T}\mathbb{X})$ are zero, implies $\mathcal{C}(\mathbb{T}\mathbb{X}) = H$.

Rotating the whole n-dimensional space by such an orthogonal transformation $\mathbb{T}$ yields a new linear model with $\mathbf{Y}, \mathbb{X}$ and $\boldsymbol{\varepsilon}$ replaced by $\mathbb{T}\mathbf{Y}, \mathbb{T}\mathbb{X}$ and $\mathbb{T}\boldsymbol{\varepsilon}$. Since $\boldsymbol{\varepsilon} \stackrel{d}{=} \mathbb{T}\boldsymbol{\varepsilon}$, this case now reduces to the previous.

1.7.3 Consistency

If we are considering our model as one in a sequence of models of the form $\mathbf{Y}_{(n)} = \mathbb{X}_{(n)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{(n)}$, with $\mathbf{Y}_{(n)}$, and $\boldsymbol{\varepsilon}_{(n)} \in \mathbb{R}^n$ and where $\mathbb{X}_{(n)}$ is $n \times p$, all with a common parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we may ask also about the asymptotic properties of the estimator $\widehat{\boldsymbol{\beta}}_{(n)}$ in (1.14) with $\mathbb{X}$ and $\mathbf{Y}$ replaced by $\mathbb{X}_{(n)}$ and $\mathbf{Y}_{(n)}$ respectively, as $n \to \infty$.

Since $\widehat{\boldsymbol{\beta}}_{(n)}$ is an unbiased estimator of $\boldsymbol{\beta}$ by Part *1* of Proposition 1.2.3, Chebyshev's inequality yields

$$P(|\widehat{\beta}_{(n)\,i} - \beta_i| > \epsilon) \leq \frac{\mathrm{Var}(\widehat{\beta}_{(n)\,i})}{\epsilon^2} = \frac{\sigma^2(\mathbb{X}_{(n)}^{\mathsf{T}}\mathbb{X}_{(n)})_{ii}^{-1}}{\epsilon^2},$$

by Part *2* of Proposition 1.2.3, where $\mathbb{A}_{ii}$ is the i^{th} diagonal element of the matrix $\mathbb{A}$. Hence, if $(\mathbb{X}_{(n)}^{\mathsf{T}}\mathbb{X}_{(n)})_{ii}^{-1}$ converges to zero as $n \to \infty$, then $\widehat{\beta}_{(n)\,i} \xrightarrow{p} \beta_i$. If such convergence holds for all $i = 1, \ldots, p$, then $\widehat{\boldsymbol{\beta}}_{(n)} \xrightarrow{p} \boldsymbol{\beta}$, that is, $\widehat{\boldsymbol{\beta}}_{(n)}$ is a consistent sequence of estimators of $\boldsymbol{\beta}$. The consistency of the sequence $\widehat{\boldsymbol{\beta}}_{(n)}$ can hold for many choices of sequences of design matrices $X_{(n)}$.

1.7.4 Wald statistics and confidence sets

We derive some statistics related to estimation of functions of $\boldsymbol{\beta}$. These statistics can be used for testing hypotheses and forming confidence intervals and sets as in Chapter ref.

We consider the model (1.6): $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$. The hypothesis $H_0 : \beta_i = \beta_{i\,0}$ for some fixed $\beta_{i\,0}$ and $1 \leq i \leq p$ can be tested by means of the statistic

$$T = \frac{(\widehat{\beta}_i - \beta_{i\,0})}{\sqrt{S^2((\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1})_{ii}}}.$$

By Proposition 1.7.1, $\mathrm{Var}(\widehat{\beta}_i) = \sigma^2((\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1})_{ii}$, and it now follows that under H_0 the statistic T has the t_{n-p} distribution. Confidence intervals and tests for β_i may now be constructed as shown in Chapter ref.

We may construct confidence sets for the entire vector $\boldsymbol{\beta}$ under the same assumptions. Proposition 1.7.1 implies that under $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}}\sigma^{-2}(\mathbb{X}^{\mathsf{T}}\mathbb{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \sim \chi_p^2,$$

and replacing σ^2 by the consistent estimator S^2, we can base our inference on the statistic

$$W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}} S^{-2} (\mathbb{X}^{\mathsf{T}}\mathbb{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{1.36}$$

Proposition 1.7.1 implies that $W/p \sim F_{p,n-p}$, hence the ellipsoid centered at $\widehat{\boldsymbol{\beta}}$, defined by

$$\{\boldsymbol{\beta} : W \leq pF_{p,n-p,1-\alpha}\}$$

is a $1-\alpha$ confidence set for $\boldsymbol{\beta}$, where $F_{p,n-p,1-\alpha}$ is the $1-\alpha$ quantile of the $F_{p,n-p}$ distribution. It is easy to verify that $W \xrightarrow{d} \chi^2_p$ as $n \to \infty$. We will see in Section 1.9.2 that this convergence also holds without the assumption of normality under certain condtions.

1.8 Likelihood ratio tests

In this section our goal is to test the null hypothesis $H_{\mathbf{c}}$ that $\boldsymbol{\beta}$ satisfies the constraint $\mathbb{D}\boldsymbol{\beta} = \mathbf{c}$ of (1.24) against the alternative that $\boldsymbol{\beta}$ is any other vector in $\mathbb{R}^p$,; in other words, the null is equivalent to $\mathrm{E}(\mathbf{Y}) \in M_{\mathbf{c}}$ where $M_{\mathbf{c}} = \{\mathbb{X}\boldsymbol{\beta} : \mathbb{D}\boldsymbol{\beta} = \mathbf{c}\}$. The latter space is an affine subspace of $\mathbb{R}^p$, which is a linear subspace when $\mathbf{c} = \mathbf{0}$. Hypotheses about equality between some of the components of $\boldsymbol{\beta}$, that is, that $\beta_i = \beta_j$ for some i and j, or equality between linear combinations of $\boldsymbol{\beta}$, can be expressed in (1.24) with $\mathbf{c} = \mathbf{0}$. We continue to assume model (1.33) as in Section 1.7, that is, that $\mathbf{Y} \sim \mathcal{N}(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I})$, or equivalently $\mathrm{E}(\mathbf{Y}) \in \mathcal{C}(\mathbb{X})$, with errors satisfying $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I})$.

Given any subset $C \subset \mathbb{R}^n$ set $\widehat{\mathbf{Y}}_C = \arg\min_{\boldsymbol{\mu}\in C} ||\mathbf{Y} - \boldsymbol{\mu}||^2$, and $\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{X})}$. A direct calculation, akin to computing maximum likelihood estimators for normal errors, as in Exercise 1.11.16, shows that

$$\max_{\boldsymbol{\mu}\in C, \sigma^2\in\mathbb{R}^+} \left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-||\mathbf{Y}-\boldsymbol{\mu}||^2/2\sigma^2} \right\} = \frac{1}{(2\pi||\mathbf{Y} - \widehat{\mathbf{Y}}_C||^2/n)^{n/2}} e^{-n/2}. \tag{1.37}$$

The generalized likelihood ratio test statistic for testing

$$H_{\mathbf{c}} : \mathrm{E}(\mathbf{Y}) \in M_{\mathbf{c}} \quad \text{versus} \quad H^a_{\mathbf{c}} : \mathrm{E}(\mathbf{Y}) \in M^c_{\mathbf{c}} \cap \mathcal{C}(\mathbb{X})$$

is therefore

$$\Lambda := \frac{\max_{\boldsymbol{\mu}\in\mathcal{C}(\mathbb{X}), \sigma^2\in\mathbb{R}^+} \left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-||\mathbf{Y}-\boldsymbol{\mu}||^2/2\sigma^2} \right\}}{\max_{\boldsymbol{\mu}\in M_{\mathbf{c}}, \sigma^2\in\mathbb{R}^+} \left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-||\mathbf{Y}-\boldsymbol{\mu}||^2/2\sigma^2} \right\}} = \frac{||\mathbf{Y} - \widehat{\mathbf{Y}}_{M_{\mathbf{c}}}||^n}{||\mathbf{Y} - \widehat{\mathbf{Y}}||^n}. \tag{1.38}$$

The quantities in the denominator and numerator of (1.38) measure the lack of fit of the unconstrained model $\mathrm{E}(\mathbf{Y}) \in \mathcal{C}(\mathbb{X})$, and the null model $\mathrm{E}(\mathbf{Y}) \in M_{\mathbf{c}}$, respectively. The null hypothesis that $\mathrm{E}(\mathbf{Y}) \in M_{\mathbf{c}}$ is rejected

when Λ is large, indicating that the lack of fit for the model restricted to have its mean vector in the null parameter space is large relative to that of the model which requires only that $\mathrm{E}(\mathbf{Y}) \in \mathcal{C}(\mathbb{X})$.

In order to construct a test, we need to determine the distribution of Λ. By Proposition 1.2.3, $\frac{1}{\sigma^2}||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 \sim \chi^2_{n-p}$. Furthermore,

$$\mathbf{Y} - \widehat{\mathbf{Y}}_{M_\mathbf{c}} = \mathbf{Y} - \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}} = (\mathbb{I} - \mathbb{P})\mathbf{Y} + (\mathbb{P} - \mathbb{P}_{M_\mathbf{c}})\mathbf{Y}, \qquad (1.39)$$

where $P_{M_\mathbf{c}}$ is an affine operator projection onto $M_\mathbf{c}$ given by (1.30). We readily see that the two terms on the right-hand side of (1.39) are orthogonal since $\mathbb{I} - \mathbb{P}$ projects to $\mathcal{C}(\mathbb{X})^\perp$, and $\mathbb{P}$ and $\mathbb{P}_{M_\mathbf{c}}$ both map to $\mathcal{C}(\mathbb{X})$. In particular, we obtain the Pythagorean identity

$$||\mathbf{Y} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}||^2 = ||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 + ||\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}||^2. \qquad (1.40)$$

This equation and other identities and orthogonality relations appearing in this discussion can be seen and understood from figure 1.2. The orthogonality of the two summands on the right-hand side of (1.39), which are jointly normally distributed, being an affine transformation of a multivariate normal vector, implies their independence. Since the decision whether to reject $H_\mathbf{c}$ or not depends on whether the likelihood ratio statistic Λ is larger than some critical value or not, in view of (1.40), it suffices to study the distribution of the monotone function of Λ given by

$$F := \frac{||\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}||^2/q}{||\mathbf{Y} - \widehat{\mathbf{Y}}||^2/(n-p)}. \qquad (1.41)$$

For the case $\mathbf{c} = \mathbf{0}$, arguments identical to those of Proposition 1.7.1 show that $\frac{1}{\sigma^2}||\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{0}}||^2 \sim \chi^2_q$ under $H_\mathbf{0}$. Also by 1.7.1, $||\mathbf{Y} - \widehat{\mathbf{Y}}||^2/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$, and the independence of the two terms on the right hand side of (1.40), it follows by definition of the F distribution that

$$F \sim F_{q,n-p}. \qquad (1.42)$$

Hence, a level-α likelihood ratio test rejects $H_\mathbf{0}$ when $F > k$, where k is the $1-\alpha$ percentile of the $F_{q,n-p}$ distribution.

In order to consider the case $\mathbf{c} \neq \mathbf{0}$ we use (1.25) which yields, for any $\mathbf{b}$ satisfying $\mathbb{D}\mathbf{b} = \mathbf{c}$,

$$||\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}||^2 = ||\mathbb{P}(\mathbf{Y} - \mathbb{X}\mathbf{b}) - \mathbb{P}_{\mathbf{M_0}}(\mathbf{Y} - \mathbb{X}\mathbf{b})||^\mathbf{2}. \qquad (1.43)$$

The hypothesis $H_\mathbf{c} : \mathrm{E}(\mathbf{Y}) \in M_\mathbf{c}$ is equivalent to $H_\mathbf{0} : \mathrm{E}(\mathbf{Y}) - \mathbb{X}\mathbf{b} \in \mathbf{M_0}$. It follows that under $H_\mathbf{c}$ the right-hand side of (1.43) has a χ^2_q distribution and that the statistic F defined in (1.41) is again distributed as $F_{q,n-p}$.

Finally in order to provide a useful expression for the likelihood ratio test we compute $\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}$ explicitly. First, write this expression as $\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widehat{\boldsymbol{\beta}}_{M_\mathbf{c}}$. Note that from (1.26) we have $\mathbb{X}\widehat{\boldsymbol{\beta}}_{M_\mathbf{c}} - \mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\left[\mathbb{D}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\right]^{-1}(\mathbf{c} - \mathbb{D}\widehat{\boldsymbol{\beta}})$. It then follows by a straightfor-

ward calculation that

$$||\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}||^2 = (\mathbb{D}\widehat{\boldsymbol{\beta}} - \mathbf{c})^\mathsf{T} \left[\mathbb{D}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\right]^{-1} (\mathbb{D}\widehat{\boldsymbol{\beta}} - \mathbf{c}).$$

We summarize these results the following

Proposition 1.8.1 *1. If* $\mathbf{Y} \sim \mathcal{N}(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I})$ *and in particular* $\mathrm{E}(\mathbf{Y}) \in \mathcal{C}(\mathbb{X})$*, then* $||\mathbf{Y} - \widehat{\mathbf{Y}}||^2$ *and* $||\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}||^2$ *are statistically independent.*

2. If in addition we have $\mathrm{E}(\mathbf{Y}) \in M_\mathbf{c}$*, then* $||\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{M_\mathbf{c}}||^2/\sigma^2 \sim \chi^2_q$*.*

3. The likelihood ratio test statistic F *of* (1.41) *can be written as*

$$F = \frac{(\mathbb{D}\widehat{\boldsymbol{\beta}} - \mathbf{c})^\mathsf{T} \left[\mathbb{D}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{D}^\mathsf{T}\right]^{-1} (\mathbb{D}\widehat{\boldsymbol{\beta}} - \mathbf{c})}{qS^2}, \tag{1.44}$$

with S^2 *defined in* (1.23)*. Under the hypothesis* $H_\mathbf{c} : \mathrm{E}(\mathbf{Y}) \in M_\mathbf{c}$ *we have* $F \sim F_{q,n-p}$*.*

1.9 Random design matrices

1.9.1 Conditioning on a random $\mathbb{X}$

In some applications the design matrix, that is, the covariates, are determined by the designer of the experiment, e.g., a chemist who chooses the inputs in an experiment. However, in most cases $\mathbb{X}$ is random, often with some assumptions. If we are interested only in the relation between Y and its covariate vector $\mathbf{x}$ and in predicting Y from $\mathbf{x}$, and not in the distribution of the covariates, we may condition on $\mathbb{X}$, which in this case becomes *ancillary*, see Section ref.

We now assume

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbb{X}) = \mathbf{0} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbb{X}) = \sigma^2\mathbb{I}. \tag{1.45}$$

Under (1.45), all previous results up to and including Section 1.6 that hold assuming only first and second moment conditions on the errors for a fixed design matrix now hold conditionally given a full rank random $\mathbb{X}$. The results assuming normal errors also hold conditionally under the assumption that the conditional distribution of $\boldsymbol{\varepsilon}$ given $\mathbb{X}$ is normal.

For example, from Proposition 1.2.3 $\mathrm{E}(\widehat{\boldsymbol{\beta}}|\mathbb{X}) = \boldsymbol{\beta}$, $\mathrm{Var}(\widehat{\boldsymbol{\beta}}|\mathbb{X}) = \sigma^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$, and as in Theorem 1.5.1 $\mathrm{E}(S^2|\mathbb{X}) = \sigma^2$. Furthermore, the Gauss-Markov Theorem holds with variances replaced by conditional variances.

1.9.2 Independent covariate vectors

In Section 1.7.3 we considered a sequence of non-random design matrices $\mathbb{X}_{(n)}$ and, under certain conditions, proved consistency of the sequence $\widehat{\boldsymbol{\beta}}_{(n)}$ of least squares estimators. We now consider consistency and asymptotic

normality when the design matrices are allowed to be random. Sequences of random design matrices with independent and identically distributed rows arise naturally when assuming the covariate vectors are drawn independently having the distribution of some random vector $\mathbf{X}$. To show consistency we also assume $Q = \mathrm{E}[\mathbf{X}\mathbf{X}^\mathsf{T}]$ exists and is invertible. Also taking the error distribution into consideration, we moreover assume that $(\varepsilon_i, \mathbf{X}_i)$, or equivalently $(Y_i, \mathbf{X}_i)$, are i.i.d. for $i = 1, \ldots, n$. As our argument for consistency depends on an application of the Law of Large Numbers, these assumptions can be relaxed somewhat. Further, we assume that (1.45) is satisfies with $\mathbb{X}_{(n)}$ replacing $\mathbb{X}$.

Under the stated conditions $n^{-1}\mathbb{X}_{(n)}^\mathsf{T}\mathbb{X}_{(n)} = n^{-1}\sum_{i=1}^n \mathbf{X}_i^\mathsf{T}\mathbf{X}_i$ converges in probability, by the Law of Large Numbers, to Q, and $n(\mathbb{X}_{(n)}^\mathsf{T}\mathbb{X}_{(n)})^{-1}$ likewise converges to Q^{-1}. To show the consistency of $\widehat{\boldsymbol{\beta}}_{(n)}$, write

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(n)} &= (\mathbb{X}_{(n)}^\mathsf{T}\mathbb{X}_{(n)})^{-1}\mathbb{X}_{(n)}^\mathsf{T}\mathbf{Y} = (\mathbb{X}_{(n)}^\mathsf{T}\mathbb{X}_{(n)})^{-1}\mathbb{X}_{(n)}^\mathsf{T}(\mathbb{X}_{(n)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + n(\mathbb{X}_{(n)}^\mathsf{T}\mathbb{X}_{(n)})^{-1}\mathbb{X}_{(n)}^\mathsf{T}\boldsymbol{\varepsilon}/n. \qquad (1.46)\end{aligned}$$

The j^{th} component of $\mathbb{X}_{(n)}^\mathsf{T}\boldsymbol{\varepsilon}/n$ is equal to $\sum_{i=1}^n X_{ij}\varepsilon_j/n$, where X_{ij}, $j = 1, \ldots, p$ are i.i.d. having the distribution of the j^{th} component of $\mathbf{X}$. Since (1.45) gives that the errors have mean zero and finite variance, we have $\mathrm{E}(X_{ij}\varepsilon_j) = 0$, see Exercise 1.11.20. Now the Law of Large Numbers yields $\sum_{i=1}^n X_{ij}\varepsilon_j/n \xrightarrow{p} 0$. Together with the fact that $n(\mathbb{X}_{(n)}^\mathsf{T}\mathbb{X}_{(n)})^{-1}$ converges to Q^{-1}, consistency follows from (1.46).

We next show that under the same conditions, the sequence of estimators $\widehat{\boldsymbol{\beta}}_{(n)}$ is asymptotically normal. Rewriting (1.14) using some straightforward manipulations, see Exercise 1.11.21, shows that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{(n)} - \boldsymbol{\beta}) = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^\mathsf{T}\mathbf{X}_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_i^\mathsf{T}\varepsilon_i\right). \qquad (1.47)$$

As above we have that $n^{-1}\sum_{i=1}^n \mathbf{X}_i^\mathsf{T}\mathbf{X}_i \xrightarrow{p} Q$, and setting $\Omega = \mathrm{Var}(\mathbf{X}_i^\mathsf{T}\varepsilon_i)$ the Central Limit Theorem and Slutsky's Lemma, see Section ref, yield

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{(n)} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, Q^{-1}\Omega Q^{-1}).$$

If ε_i and $\mathbf{X}_i$ are independent, using the conditional variance formula in (1.45) to see that $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I$, we obtain $\Omega = \sigma^2 Q$, and further that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{(n)} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2 Q^{-1}), \qquad (1.48)$$

which should be compared to the claim on the distribution of $\widehat{\boldsymbol{\beta}}$ for a fixed design matrix, in Proposition 1.7.1. As in the case for fixed design matrices, these results may be used for testing hypotheses and forming confidence regions for $\boldsymbol{\beta}$.

Under the assumptions leading to (1.48) we can construct a confidence interval for a function $g(\boldsymbol{\beta}) \in \mathbb{R}$ of the regression parameters using Theorem

?? to see that the statistic

$$(g(\hat{\boldsymbol{\beta}}_{(n)})) - g(\boldsymbol{\beta}))^{\mathsf{T}} S^{-2} \dot{g}(\boldsymbol{\beta})(\mathbb{X}_{(n)}^{\mathsf{T}} \mathbb{X}_{(n)}) \dot{g}(\boldsymbol{\beta})^{\mathsf{T}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{1.49}$$

which is a more general case of the Wald statistic, has asymptotically the χ_1^2 distribution. Here $\dot{g}(\boldsymbol{\beta}) = (\frac{\partial}{\partial \beta_1} g(\boldsymbol{\beta}), \ldots, \frac{\partial}{\partial \beta_p} g(\boldsymbol{\beta}))$. Replacing the above function g by $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^p$ in (1.49), the resulting Wald statistic has the χ_p^2 asymptotic distribution.

Similar results hold without assuming normality of $\boldsymbol{\varepsilon}$. Under the conditions leading to (1.48), the latter implies

$$n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}} S^{-2} Q (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

is asymptotically χ_p^2 under H_0. Replacing Q by its consistent estimator $\frac{1}{n}\mathbb{X}^{\mathsf{T}}\mathbb{X}$ does not affect its asymptotic χ^2 distribution, and we obtain the statistic (1.36), which is a special case of the Wald statistic , see Chapter ref.

1.9.3 Multiple correlations

In this section we assume that the first column of the design matrix $\mathbb{X}$ consists of ones, corresponding to linear regression with an intercept. In this case, the normal equation (1.13) implies that $\mathbf{1}^{\mathsf{T}}(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{0}$, where $\mathbf{1}^{\mathsf{T}}$ is a row consisting of n ones. It follows that the average of the components of $\mathbf{Y}$ and the average of the components of $\widehat{\mathbf{Y}}$ coincide, both being equal to $\overline{Y} = \sum_i Y_i/n$. Let $\overline{\mathbf{Y}} \in \mathbb{R}^n$ be the vector having all coordinates equal to $\overline{Y}$. The quantity R^2 defined by

$$R^2 = \frac{||\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}||^2}{||\mathbf{Y} - \overline{\mathbf{Y}}||^2} = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}, \tag{1.50}$$

is known as the *multiple correlation coefficient* , and is also sometimes called the coefficient of determination. The notation $R^2_{Y\mathbb{X}}$ is often used in order to express the dependence of R^2 on the set of covariates. This quantity measures the part of the total variance $\sum_{i=1}^n (Y_i - \overline{Y})^2$ of the Y's which is explained by the regression model. In Figure 1.3 we see that that $R^2 = \cos(\theta)$, where θ is the angle between the vectors $\mathbf{Y} - \overline{\mathbf{Y}}$ and $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$. It is not hard to see that for a single covariate X one has $R^2_{YX} = r(Y, X)^2$, the sample correlation for (X_i, Y_i), see Exercise 1.11.22. In general, R^2 is the square of the sample correlation for the pairs $(Y_i, \widehat{Y}_i)$.

To see this fact note that

$$r(Y, \widehat{Y})^2 = \left(\frac{(\mathbf{Y} - \overline{\mathbf{Y}})^{\mathsf{T}} (\widehat{\mathbf{Y}} - \overline{\mathbf{Y}})}{||\mathbf{Y} - \overline{\mathbf{Y}}||\ ||\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}||} \right)^2,$$

and that this latter expression is equal to R^2 of (1.50) by virtue of the identity

$$(\mathbf{Y} - \overline{\mathbf{Y}})^{\mathsf{T}} (\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}) = (\mathbf{Y} - \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}} - \overline{\mathbf{Y}})^{\mathsf{T}} (\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}) = ||\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}||^{\mathbf{2}}$$

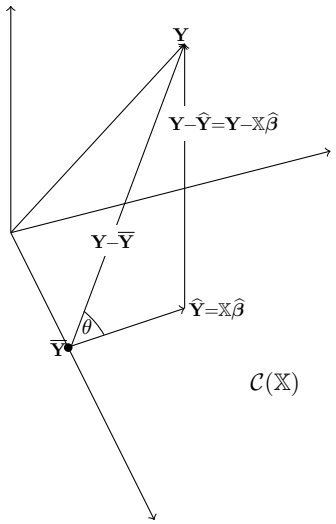

Figure 1.3. The figure demonstrates the Pythagorean identity $||\mathbf{Y} - \overline{\mathbf{Y}}||^2 = ||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 + ||\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}||^2$, and $R^2 = \cos(\theta)$.

by the orthogonality of $(\mathbf{Y} - \widehat{\mathbf{Y}})$ and $(\widehat{\mathbf{Y}} - \overline{\mathbf{Y}})$.

Suppose we partition the design matrix $\mathbb{X}$ as $\mathbb{X} = (\mathbb{U}, \mathbb{V})$ where $\mathbb{U}$ is $n \times q$ and $\mathbb{V}$ is $n \times (p-q)$, and in Figure 1.3 we replace the space spanned by the vector $\overline{\mathbf{Y}}$ by a space spanned by the columns of $\mathbb{V}$, replacing also the vector $\overline{\mathbf{Y}}$ by $\widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{V})}$, the projection of $\mathbf{Y}$ into $\mathcal{C}\mathbb{V}$. Now set θ to be the angle between the vectors $\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{V})}$ and $\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{V})}$, where as usual, $\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{X})}$. Then $\cos(\theta)$ is denoted by $R^2_{Y\mathbb{U}|\mathbb{V}}$, the *multiple partial correlation* that measures the relation between the dependent variable Y and the covariates in $\mathbb{U}$ given those in $\mathbb{V}$. Formally, we have

$$R^2_{Y\mathbb{U}|\mathbb{V}} = \frac{||\widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{X})} - \widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{V})}||^2}{||\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathcal{C}(\mathbb{V})}||^2}.$$

The multiple correlation R^2 is sometimes used as a measure of the quality of a model, a high R^2 suggesting a better predictive power. It is easy to see that if the covariates in one model include all those of another model together with some additional ones, then the model with more covariates will have a larger value of R^2. This fact follows from the observation that $||\mathbf{Y} - \widehat{\mathbf{Y}}||^2$ decreases if the space $\mathcal{C}(\mathbb{X})$ is replaced by a larger space, and therefore $||\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}||^2$ must increase, by the Pythagorean identity $||\mathbf{Y} - \overline{\mathbf{Y}}||^2 = ||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 + ||\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}||^2$. For this reason one should be careful in using R^2 for model selection, as doing do may lead to large models that *overfit* the data. This phenomenon is discussed in more detail in Chapter ref.

1.10 Residual analysis

In previous sections various assumptions on the errors ε_i in the model (1.6) appear. For instance, the errors may be assumed to have equal variances, that is, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$, a property known as *homoscedacticity*, or be independent and normally distributed. As the errors ε_i are not observed, in order to determine if such assumptions are reasonable we instead consider the vector of *residuals* given by $\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}$. One may use formal or descriptive tests on $\widehat{\varepsilon}_i$, such as plotting histograms to assess normality, see Section ref.

Note that by (1.13) $\mathbb{X}^{\mathsf{T}}\widehat{\boldsymbol{\varepsilon}} = \mathbf{0}$. In particular, in the common case that the first column of $\mathbb{X}$ consists of 1's, we always have $\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i = 0$. Therefore the latter identity cannot be considered evidence that $\mathrm{E}(\varepsilon_i) = 0$, which is part of our assumptions.

For testing normality of the residuals, and the assumption of equal variances, a qqplots elsewhere, but we should have $F^{-1}(U)$, order stats - which we have etc. ...used, which graphs the points $(\Phi^{-1}((i-1/2)/n), \widehat{\varepsilon}_{(i)})$, where Φ is the standard normal distribution function, and $\hat{\varepsilon}_{(i)}$ are the residuals in increasing order.

This plot is motivated as follows: given a sample $U_1, \ldots, U_n$, where $U_i \sim \mathcal{U}[0,1]$, the order statistics of the sample satisfy $U_{(i)} \sim Beta(i, n+1-i)$ and therefore $\mathrm{E}(U_{(i)}) = i/(n+1)$. Hence, we expect that a plot of $(i/(n+1), U_{(i)})$, the order statistics of the uniform sample against their expectations, would be nearly linear and close to the main diagonal. Applying the same monotone transformation to both coordinates of these points we again expect a nearly diagonal linear plot. In particular, applying the transformation Φ^{-1}, the points $(\Phi^{-1}(i/(n+1)), \Phi^{-1}(U_{(i)}))$ should lie on the diagonal, and as $\Phi^{-1}(U) \sim \mathcal{N}(0,1)$, we can expect this same pattern in the plot $(\Phi^{-1}(i/(n+1)), \widehat{\varepsilon}_{(i)})$ if the residuals come from a standard normal distribution. If the residuals are not standard, but all have the same normal distribution, the plot would still be nearly linear, with a slope depending on the standard deviation of $\widehat{\varepsilon}_i$, see Exercise 1.11.23. A continuity-type correction leads to the definition of the normal probability plot above, with the factor 1/2.

In homoscedastic models with normal errors, the residuals should look like i.i.d. normal variables. However, in some regression data sets, it may be the case that a certain natural ordering of these errors will reveal a deviation from the i.i.d. property. A run test on $\widehat{\varepsilon}_i$ in the suspected order can be used to detect such a deviation.

Residual plots are scatter plots of the residuals in some given order. For example, suppose the data arise from experiments performed sequentially in time. A plot of the residuals ordered by time, that is, a scatter plot of the points $(t_i, \widehat{\varepsilon}_i)$, or the points $(i, \widehat{\varepsilon}_{t(i)})$ where t_i is the time of the i^{th} experiment and $t_{(i)}$ are their order statistics, may reveal a pattern of deviation from independence, or may show a departure from homoscedasticity

by having, for example, residuals whose variability increases in time. As an example, consider a chemical experiment, where the input covariates consist of reagents that deteriorate in time, resulting in lower response values Y. A plot of the residuals against time will tend to have initially positive residuals, and then negative ones, indicating a violation of the i.i.d. assumption on the errors. If reagents' deterioration also leads to an increase in the variance of the response values, then again the same plot should reveal this deviation from homoscedasticity by showing an increase in the dispersion of the residuals as a function of time.

Often, the ordering of the residuals is determined with respect to a covariate, say x, which need not be in the model. In this case, various anomalies may show up in a scatter plot of the points $(x_i, \widehat{\varepsilon}_i)$. For example, if the covariate x appears linearly in the model, but its true relation to the response Y is, say, quadratic, then we may see a quadratic pattern in the residual plot.

1.11 Exercises

Exercise 1.11.1 *Verify identities* (1.2) *and* (1.3).

Exercise 1.11.2 *Consider the model*

$$Y = \beta_1 + \beta_2 X + \epsilon,$$

where X and ϵ are independent random variables with finite variance, and $\mathrm{Var}(Y) = \mathrm{Var}(X)$. *In terms of Galton's regression of the heights of fathers and sons, this condition means that the variability doesn't change between generations. The condition can also be achieved by standardizing the variables. Show that $|\beta_2| \leq 1$.*

Exercise 1.11.3 *With $p \leq n$, show that a matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ has independent columns if and only if $\mathbb{X}^{\mathsf{T}}\mathbb{X}$ is invertible.*

Exercise 1.11.4 *The model in this exercise is the same as that of* (1.9)*, with different notation. Consider $Y_i = \beta_1 x_i + \beta_2 u_i + \varepsilon_i$ for $i = 1, \ldots, n$, where for $i = 1, \ldots, m < n$ we have $x_i = 1$ and $u_i = 0$, and for for $i = m+1, \ldots, n$ we have $x_i = 0$, and $u_i = 1$. Write these relation in the form $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$; write $\mathbb{X}$ explicitly, and compute $\widehat{\boldsymbol{\beta}}$.*

In this exercise you can think of two groups, the first group consists of the first m, and the remaining $n - m$ subjects, all members of each group having the same mean. In this case x and u are called dummy variables.

Exercise 1.11.5 *For each of the model* (1.7) *and* (1.8) *of simple linear regression, write down the matrix $\mathbb{X}$ and the estimator $\widehat{\boldsymbol{\beta}}$ explicitly using* (1.14). *Note that the calculation is simpler for* (1.8). *Compare the results. Obtain these estimators also by differentiation. Show that the line $Y =$*

$\widehat{\beta}_1 + \widehat{\beta}_2 x$ passes through the point $(\bar{x}, \bar{Y})$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

Exercise 1.11.6 *Continuing Exercise 1.11.2 consider the model* (1.7)*. Compute $\widehat{\boldsymbol{\beta}}_2$, and show if $\sum_i (x_i - \bar{x})^2 = \sum_i (Y_i - \bar{Y})^2$ then $|\widehat{\boldsymbol{\beta}}_2| \leq 1$.*

Exercise 1.11.7 *Show that the normal equation* (1.13) *is equivalent to $\frac{\partial}{\partial \boldsymbol{\beta}} ||Y - \mathbb{X}\boldsymbol{\beta}||^2 = \mathbf{0}$, so the least squares estimator* (1.14) *can be obtained also by straightforward differentiation.*

Exercise 1.11.8

1. *Show that two $n \times p$ matrices $\mathbb{X}$ and $\mathbb{U}$ satisfy $\mathcal{C}(\mathbb{X}) = \mathcal{C}(\mathbb{U})$ if and only if there exists a nonsingular $p \times p$ matrix $\mathbb{A}$, such that $\mathbb{U} = \mathbb{X}\mathbb{A}$.*
2. *Show that the projection matrix defined in* (1.15) *based on $\mathbb{X}$ is equal to that based on $\mathbb{U}$. In other words, the projection matrix to a given subspace is unique.*
3. *Prove that if $\mathbb{P}$ satisfies* (1.16) *so does $\mathbb{N} = \mathbb{I} - \mathbb{P}$.*
4. *Prove that if $\mathbb{P}$ satisfies* (1.16) *then for every column vector $\mathbf{v} \in \mathbb{R}^n$ we have $(\mathbb{I} - \mathbb{P})\mathbf{v} \perp \mathbb{P}\mathbf{v}$. Therefore the equation $\mathbf{v} = \mathbb{P}\mathbf{v} + (\mathbb{I} - \mathbb{P})\mathbf{v}$ decomposes $\mathbf{v}$ into a sum of two orthogonal projections.*
5. *Prove that $\mathbb{P}\mathbf{v} = \mathbf{v}$ for any vector $\mathbf{v} \in \mathcal{C}(\mathbb{X})$, and that $\mathbb{P}\mathbb{X} = \mathbb{X}$.*
6. *Prove that any $\mathbf{v} \in \mathbb{R}^n$ satisfies $||\mathbf{v}|| \geq ||\mathbb{P}\mathbf{v}||$.*

Exercise 1.11.9 *Prove Proposition 1.2.3 and its analog for the weighted least squares estimator* (1.21) *associated with the model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 V$. In particular show that again $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.*

Exercise 1.11.10 *Prove Corollary 1.3.2 by using Theorem 1.3.1 and that $\mathrm{Var}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}})$ is nonnegative definite if and only if $\mathrm{Var}(\mathbf{a}\widetilde{\boldsymbol{\beta}}) \geq \mathrm{Var}(\mathbf{a}\widehat{\boldsymbol{\beta}})$ for all row vectors $\mathbf{a} \in \mathbb{R}^p$. In particular the latter relation implies $\mathrm{Var}(\widetilde{\beta}_j) \geq \mathrm{Var}(\widehat{\beta}_j)$ for $j = 1, \ldots, p$.*

Exercise 1.11.11 *Prove the Gauss-Markov Theorem for a general variance matrix: if $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbb{V}$, then the estimator of* (1.21) *has the minimal variance matrix among linear unbiased estimators.*

Exercise 1.11.12 *The model* (1.9) *is often expressed in the form*

$$Y_i = \mu + \alpha_1 + \varepsilon_i \quad i = 1, \ldots, m, \qquad Y_i = \mu + \alpha_2 + \varepsilon_i \quad i = m+1, \ldots, n.$$

This model arises naturally, for example, when we study reactions Y_i to one of two treatments. Then μ is thought of as the mean reaction level without treatment, and α_j are the treatment effects, so that $\mu + \alpha_j$ are the mean levels under treatments $j = 1, 2$. Express this model in the form of (1.6)*, with $\boldsymbol{\beta}^{\mathsf{T}} = (\mu, \alpha_1, \alpha_2)$ and an $n \times 3$ design matrix $\mathbb{X}$. Show that $\mathbb{X}$ is not of full rank.*

Exercise 1.11.13 *Prove Theorem 1.4.1 by using Theorem 1.3.1 with* $\mathbb{W}$ *replacing* $\mathbb{X}$ *in the model* (1.22).

Exercise 1.11.14 *Consider the model* (1.24)*, and assume that for a given* $\mathbf{c}$ *there exists a solution* $\boldsymbol{\beta}$ *to the equation* $\mathbb{D}\boldsymbol{\beta} = \mathbf{c}$*, but suppose* $\mathbb{D}$ *does not have independent columns. Show that it is possible to delete a number of rows from* $\mathbb{D}$ *and the same number of elements from* $\mathbf{c}$ *so that* $\mathbb{D}$ *will independent rows, and obtain an equivalent model.*

Exercise 1.11.15 *Suppose for* $i = 1, 2, 3$ *that* n_i *independent, unbiased measurements are taken on the three angles of a triangle. Apply Theorem 1.6.1 to find the least squares estimators of the three angles.*

Exercise 1.11.16 *Show that if* $\mathbf{Y} \sim N(\mathbb{X}\boldsymbol{\beta}, \sigma^2 I)$ *then the MLE of* $(\boldsymbol{\beta}, \sigma^2)$ *is* $(\widehat{\boldsymbol{\beta}}, \frac{1}{n}||\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}||^2)$.

Exercise 1.11.17 *In the model* (1.17) *show that* $\widehat{\boldsymbol{\beta}}$ *of* (1.21) *is the MLE of* $\boldsymbol{\beta}$*, and find the MLE of* σ^2.

Exercise 1.11.18 *Prove that if a square matrix* $\mathbb{A}$ *satisfies* $\mathbb{A}^2 = \mathbb{A}$ *then its eigenvalues are all zero or one.*

Exercise 1.11.19 *Prove* (1.37) *and* (1.38).

Exercise 1.11.20 *Prove that under the model* (1.45) *we have* $\mathrm{E}(X_{ij}\varepsilon_j) = 0$.

Exercise 1.11.21 *Prove* (1.47).

Exercise 1.11.22 *Show that for a single covariate* X *we have* $R^2_{Y \cdot X} = Corr(Y, X)^2$*, where* $Corr(Y, X)$ *is the sample correlation between* Y *and* X.

Exercise 1.11.23 *Given a sample* $X_1, \ldots, X_n$*, its normal probability plot is a scatter plot of the points* $(\Phi^{-1}((i - 0.5)/n, X_{(i)})$*, where* $X_{(i)})$ *are the order statistics of the sample. If* $X_i \sim \mathcal{N}(\mu, \sigma^2)$*, how do you expect* μ *and* σ *to appear in the plot?*

References

Ross, S.M. (2004), *Introduction to probability and statistics for engineers and scientists*. Academic Press.

Strang, G. (2003), *Introduction to linear algebra*. Wellesley Cambridge Pr.