Lecture Notes on Linear Regression

Mário A. T. Figueiredo,
Departamento de Engenharia Electrotécnica e de Computadores,

Instituto Superior Técnico, Lisboa, Portugal

Latest update: March 2010

1 Least Squares Linear Regression

We are given a set of input-output pairs, $\mathcal{T} = \{(\mathbf{x}_{(1)}, y_{(1)}), ..., (\mathbf{x}_{(n)}, y_{(n)})\}$, where each $\mathbf{x}_{(i)} \in \mathbb{R}^p$ and $y_{(i)} \in \mathbb{R}$. The goal is to estimate a linear "regression function"of the form

$$g(\mathbf{x}, \boldsymbol{\beta}, \beta_0) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_j,$$

where $\boldsymbol{\beta} = [\beta_1, ..., \beta_p]^T$ is the vector of "regression coefficients", β_0 is the "intercept coefficient", and x_j denotes the j-component of vector $\mathbf{x}$. The components of the $\mathbf{x}_{(i)}$ are usually called the "explanatory variables" or "independent variables". The classical estimation criterion is the minimization of the mean squared error on $\mathcal{T}$, that is,

$$\begin{aligned}
\left(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0\right) &= \arg\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^{n} \left(y_{(i)} - g(\mathbf{x}_{(i)}, \boldsymbol{\beta}, \beta_0)\right)^2 \\
&= \arg\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^{n} \left(y_{(i)} - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{i,j} \right)^2 \\
&= \arg\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^{n} \left(y_{(i)} - \beta_0 - \boldsymbol{\beta}^T \mathbf{x}_{(i)}\right)^2,
\end{aligned}$$

where $x_{i,j}$ is the j-th component of $\mathbf{x}_{(i)}$. We can start by getting rid of β_0; without loss of generality, we can assume that each regressor x_j was "centered"by having its mean removed from the training set, that is,

$$\sum_{i=1}^{n} x_{i,j} = 0, \tag{1}$$

for $j = 1, ..., p$, and that the "response" variables also have zero sample mean,

$$\sum_{i=1}^{n} y_{(i)} = 0.$$

Under these conditions, it's trivial to show (and is left as an exercise to the reader) that the minimization with respect to β_0 does not depend of $\boldsymbol{\beta}$ and leads to $\widehat{\beta}_0 = 0$. We are thus left with

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_{(i)} - \sum_{j=1}^{p} \beta_j \, x_{i,j} \right)^2 \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_{(i)} - \boldsymbol{\beta}^T \mathbf{x}_{(i)} \right)^2 ; \end{aligned} \tag{2}$$

this can be written in matrix notation as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 , \tag{3}$$

where $\|\cdot\|^2$ denotes the square of the standard Euclidean norm, $\mathbf{y} = [y_{(1)}, ..., y_{(n)}]^T \in \mathbb{R}^n$, and $\mathbf{X}$ is a $n \times p$ matrix with $\mathbf{x}_{(i)}^T$ in the i-th column, that is,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{(1)}^T \\ \vdots \\ \mathbf{x}_{(n)}^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}.$$

Solving (3) simply requires taking the gradient with respect to $\boldsymbol{\beta}$ and equating to zero. Elementary calculus leads to

$$\begin{aligned} \nabla \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 &= \nabla \left(\mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} \right) \\ &= 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\,\mathbf{X}^T\mathbf{y}, \end{aligned} \tag{4}$$

where ∇ here denotes the gradient with respect to (w.r.t.) $\boldsymbol{\beta}$. Setting to zero and solving for $\boldsymbol{\beta}$ leads to

$$\widehat{\boldsymbol{\beta}} = \text{solution w.r.t. } \boldsymbol{\beta} \text{ of } \left\{ \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \right\}; \tag{5}$$

the system in (5) is called the system of "normal equations". If matrix $\mathbf{X}^T\mathbf{X}$ is non-singular, that is, if $\mathbf{X}$ has p non-zero singular values (equivalently, $\mathbf{X}$ has p linearly independent columns), then $\mathbf{X}^T\mathbf{X}$ has inverse and

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}^T\mathbf{X} \right)^{-1} \mathbf{X}^T\mathbf{y}, \tag{6}$$

which is known as the *ordinary least squares* (OLS) estimate.

The regressed values at the training points $\{\mathbf{x}_{(1)}, ..., \mathbf{x}_{(n)}\}$ are given by

$$\widehat{y}_{(i)} = \widehat{\boldsymbol{\beta}}^T \mathbf{x}_{(i)};$$

collecting all these estimates in a vector $\widehat{\mathbf{y}} = [\widehat{y}_{(i)}, ..., \widehat{y}_{(n)}]^T$ allows writing

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}, \tag{7}$$

where matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is usually called the "hat matrix". Matrix $\mathbf{H}$ is the orthogonal projector onto the space generated by the columns of $\mathbf{X}$; it's action is to project the observed $\mathbf{y}$ onto that space. Accordingly, as any orthogonal projection matrix, $\mathbf{H}$ is idempotent, that is $\mathbf{HH} = \mathbf{H}$; in fact,

$$\begin{aligned} \mathbf{HH} &= \mathbf{X}\overbrace{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}}^{\text{identity}}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ &= \mathbf{H}. \end{aligned} \tag{8}$$

In other words, $\widehat{\mathbf{y}}$ is the closest (in Euclidean norm) vector to $\mathbf{y}$, in the subspace spanned by the columns of $\mathbf{X}$.

2 Some Properties of Least Squares Regression

To obtain some statistical properties of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, a few assumptions about the generation of the observations $\mathbf{y} = [y_{(1)}, ..., y_{(n)}]^T$ are needed. For now, let's simply assume that each $y_{(i)}$ is obtained by adding a zero-mean random perturbation $w_{(i)}$ to a "true", or "noiseless" value $\boldsymbol{\beta}\mathbf{x}_{(i)}$. It's also assumed that all these random perturbations are statistically independent. In vector notation, these assumptions can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \tag{9}$$

where $\mathbf{w} = [w_{(1)}, ..., w_{(n)}]^T$ is a sample of a random vector of mean $[0, ..., 0]^T$ and covariance matrix $\sigma^2\mathbf{I}$, where $\mathbf{I}$ denotes an identity matrix of appropriate dimensions. Another fundamental assumption is that $\mathbf{X}$ and $\boldsymbol{\beta}$ are fixed, deterministic quantities, and all statistical variability in $\mathbf{y}$ is due to the random perturbation/noise $\mathbf{w}$.

Under the assumptions described in the previous paragraph, it is easy to conclude that $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is an unbiased estimate. In fact,

$$\begin{aligned} \mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}\left[\mathbf{y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\,\boldsymbol{\beta} \\ &= \boldsymbol{\beta}, \end{aligned}$$

where (9) was invoked to write $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$, since $\mathbf{w}$ has zero mean and $\mathbf{X}\boldsymbol{\beta}$ is a deterministic constant vector.

It is also simple to obtain the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ which, since $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right] = \boldsymbol{\beta}$, is equal to

$$\begin{aligned}\mathbf{cov}\left[\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right] &= \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}\right)^T\right] \\ &= \mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{\text{OLS}}\,\widehat{\boldsymbol{\beta}}_{\text{OLS}}^T\right] - \boldsymbol{\beta}\boldsymbol{\beta}^T. \end{aligned} \tag{10}$$

To obtain the covariance $\mathbf{cov}\left[\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right]$, the following well known fact is used: if $\mathbf{V} \in \mathbb{R}^d$ is a d-dimensional random vector with $(d \times d)$ covariance matrix $\mathbf{C}$, then the covariance of $\mathbf{U} = \mathbf{AV}$ is $\mathbf{cov}[\mathbf{U}] = \mathbf{ACA}^T$, where $\mathbf{A}$ is any matrix with d columns. The other key fact, obvious from (9), is that

$$\mathbf{cov}[\mathbf{y}] = \mathbf{cov}[\mathbf{w}] = \sigma^2\mathbf{I},$$

since $\mathbf{X}\boldsymbol{\beta}$ is deterministic. Putting these two facts together,

$$\begin{aligned}\mathbf{cov}\left[\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right] &= \mathbf{cov}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{cov}[\mathbf{y}]\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{cov}[\mathbf{y}]\,\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}, \end{aligned} \tag{11}$$

where we have used the fact that $\mathbf{X}^T\mathbf{X}$ is symmetric, so $(\mathbf{X}^T\mathbf{X})^{-1}$ is also symmetric.

If it is further assumed that the perturbation vector $\mathbf{w}$ is Gaussian, then $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is also Gaussian, because it is a linear function of a Gaussian variable ($\mathbf{y}$ is the sum of a deterministic constant with a Gaussian variable, thus is a Gaussian variable); formally,

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right),$$

where $\mathcal{N}(\mathbf{t}, \mathbf{C})$ denotes a multivariate Gaussian of mean $\mathbf{t}$ and covariance matrix $\mathbf{C}$.

3 Gauss-Markov Theorem

One of the often invoked reasons to use least squares regression is the Gauss-Markov theorem. This theorem states that, among all linear unbiased estimates of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{\text{OLS}}$ has minimal variance: $\boldsymbol{\beta}_{\text{OLS}}$ is BLUE (*best linear unbiased estimate*). Of course this does not mean that there can't exist nonlinear or biased estimates of $\boldsymbol{\beta}$ with smaller variance. Next, the Gauss-Markov theorem is presented and proved. In this section, we use the

formally more correct convention of denoting random variables and vectors with capital letters.

Gauss-Markov Theorem: *Let* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}$, *where* $\boldsymbol{\beta} \in \mathbb{R}^p$ *is an (unknown) deterministic vector,* $\mathbf{X}$ *is a (known) deterministic* $n \times p$ *matrix with rank* p, *and* $\mathbf{W} \in \mathbb{R}^n$ *is a random vector of zero mean and covariance matrix* $\sigma^2\mathbf{I}$. *Let* $\widehat{\boldsymbol{\beta}} : \mathbb{R}^n \to \mathbb{R}^p$ *be a function defined by*

$$\widehat{\boldsymbol{\beta}}(\mathbf{z}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{z}.$$

Then, $\widehat{\boldsymbol{\beta}}(\mathbf{Y}) \in \mathbb{R}^p$ *is a random vector with the following properties:*

(i) $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}(\mathbf{Y})\right] = \boldsymbol{\beta}$, *that is* $\widehat{\boldsymbol{\beta}}$ *is an unbiased estimator of* $\boldsymbol{\beta}$.

(ii) $\mathbf{cov}\left[\widehat{\boldsymbol{\beta}}(\mathbf{Y})\right] = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$.

(iii) *Let* $\widetilde{\boldsymbol{\beta}} : \mathbb{R}^n \to \mathbb{R}^p$ *be some other linear function (i.e., which can be written as* $\widetilde{\boldsymbol{\beta}}(\mathbf{z}) = \mathbf{P}\mathbf{z}$*), which is also an unbiased estimator of* $\boldsymbol{\beta}$, *i.e., such that* $\mathbb{E}\left[\widetilde{\boldsymbol{\beta}}(\mathbf{Y})\right] = \boldsymbol{\beta}$; *then,*

$$\mathbf{cov}\left[\widetilde{\boldsymbol{\beta}}(\mathbf{Y})\right] \geq \mathbf{cov}\left[\widehat{\boldsymbol{\beta}}(\mathbf{Y})\right].$$

Before the presentation of the proof, let us briefly examine part (iii) of the theorem. First of all, recall that an inequality of the form $\mathbf{A} \geq \mathbf{B}$, involving two square matrices, denotes that $\mathbf{A} - \mathbf{B}$ is a positive semi-definite matrix, that is, $\mathbf{v}^T(\mathbf{A} - \mathbf{B})\,\mathbf{v} \geq 0$ (equivalently, $\mathbf{v}^T\mathbf{A}\,\mathbf{v} \geq \mathbf{v}^T\mathbf{B}\,\mathbf{v}$), for any vector $\mathbf{v}$. Part (iii) of the theorem thus implies that any linear combination $\mathbf{v}^T\widetilde{\boldsymbol{\beta}}(\mathbf{Y})$ of elements of $\widetilde{\boldsymbol{\beta}}(\mathbf{Y})$ can't have smaller variance than the same linear combination $\mathbf{v}^T\widehat{\boldsymbol{\beta}}(\mathbf{Y})$ of elements of $\widehat{\boldsymbol{\beta}}(\mathbf{Y})$. In particular, taking $\mathbf{v} = \mathbf{e}_i$ (a vector with a 1 in position i and zeros everywhere else) leads to

$$\begin{aligned}
\mathrm{var}\left[\widetilde{\beta}_i(\mathbf{Y})\right] &= \mathbf{cov}\left[\mathbf{v}^T\widetilde{\boldsymbol{\beta}}(\mathbf{Y})\right] \\
&= \mathbf{v}^T\mathbf{cov}\left[\widetilde{\boldsymbol{\beta}}(\mathbf{Y})\right]\mathbf{v} \\
&\geq \mathbf{v}^T\mathbf{cov}\left[\widehat{\boldsymbol{\beta}}(\mathbf{Y})\right]\mathbf{v} \\
&= \mathbf{cov}\left[\mathbf{v}^T\widehat{\boldsymbol{\beta}}(\mathbf{Y})\right] \\
&= \mathrm{var}\left[\widehat{\beta}_i(\mathbf{Y})\right], \qquad (12)
\end{aligned}$$

showing that the variance of each individual component of $\widetilde{\boldsymbol{\beta}}(\mathbf{Y})$ is no smaller than the corresponding component of $\widehat{\boldsymbol{\beta}}(\mathbf{Y})$.

Proof: Parts (i) and (ii) of the theorem were proved in the previous section. To prove part (iii), we begin by recalling that if $\mathbf{U}$ and $\mathbf{V}$ are two random vectors of the same dimension, then,

$$\mathbf{cov}[\mathbf{U} + \mathbf{V}] = \mathbf{cov}[\mathbf{U}] + \mathbf{cov}[\mathbf{V}] + \mathbf{cov}[\mathbf{U}, \mathbf{V}] + \mathbf{cov}[\mathbf{V}, \mathbf{U}] \qquad (13)$$

where

$$\mathbf{cov}[\mathbf{U},\mathbf{V}] = \mathbb{E}\left[(\mathbf{U}-\mathbb{E}[\mathbf{U}])\,(\mathbf{V}-\mathbb{E}[\mathbf{V}])^T\right] = \mathbb{E}[\mathbf{U}\mathbf{V}^T] - \mathbb{E}[\mathbf{U}]E[\mathbf{V}]^T$$

is the so-called cross-covariance[1]. Application of equality (13) to $\widetilde{\boldsymbol{\beta}}(\mathbf{Y}) - \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ yields

$$\mathbf{cov}[\widetilde{\boldsymbol{\beta}}(\mathbf{Y}) - \widehat{\boldsymbol{\beta}}(\mathbf{Y})] = \mathbf{cov}[\widetilde{\boldsymbol{\beta}}(\mathbf{Y})] + \mathbf{cov}[\widehat{\boldsymbol{\beta}}(\mathbf{Y})] - \mathbf{cov}[\widetilde{\boldsymbol{\beta}}(\mathbf{Y}), \widehat{\boldsymbol{\beta}}(\mathbf{Y})] - \mathbf{cov}[\widehat{\boldsymbol{\beta}}(\mathbf{Y}), \widetilde{\boldsymbol{\beta}}(\mathbf{Y})]. \tag{14}$$

Recalling that $\widetilde{\boldsymbol{\beta}}$ is a linear estimator, $\widetilde{\boldsymbol{\beta}}(\mathbf{Y}) = \mathbf{P}\mathbf{Y}$, the condition of unbiasedness can be stated as follows: for any $\boldsymbol{\beta} \in \mathbb{R}^p$, if $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}$, where $\mathbf{W}$ is as defined above,

$$\mathbb{E}\left[\mathbf{P}\mathbf{Y}\right] = \mathbf{P}\mathbb{E}[\mathbf{Y}] = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

because $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}$ and $\mathbf{W}$ has zero expected value. For $\boldsymbol{\beta} = \mathbf{P}\mathbf{X}\boldsymbol{\beta}$ to be true for any $\boldsymbol{\beta}$ it is necessary that $\mathbf{P}\mathbf{X} = \mathbf{I}$.

We need one last fact concerning covariances: let $\mathbf{V}$ be a random vector, and $\mathbf{A}$ and $\mathbf{B}$ two matrices of the same dimension, then

$$\begin{aligned}
\mathbf{cov}[\mathbf{A}\mathbf{U}, \mathbf{B}\mathbf{U}] &= E[\mathbf{A}\mathbf{U}\mathbf{U}^T\mathbf{B}^T] - E[\mathbf{A}\mathbf{U}]E[\mathbf{B}\mathbf{U}]^T \\
&= \mathbf{A}\left(E[\mathbf{U}\mathbf{U}^T] - E[\mathbf{U}]E[\mathbf{U}]^T\right)\mathbf{B}^T \\
&= \mathbf{A}\,\mathbf{cov}[\mathbf{U}]\,\mathbf{B}^T.
\end{aligned}$$

Applying this fact to compute $\mathbf{cov}[\widetilde{\boldsymbol{\beta}}(\mathbf{Y}), \widehat{\boldsymbol{\beta}}(\mathbf{Y})]$, using that fact that $(\mathbf{X}^T\mathbf{X})^{-1}$ is symmetric, leads to

$$\begin{aligned}
\mathbf{cov}[\mathbf{P}\mathbf{Y}, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] &= \mathbf{P}\,\overbrace{\mathbf{cov}[\mathbf{Y}]}^{\sigma^2\mathbf{I}}\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2\mathbf{P}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \mathbf{cov}[\widehat{\boldsymbol{\beta}}(\mathbf{Y})].
\end{aligned} \tag{15}$$

Repeating for $\mathbf{cov}[\widehat{\boldsymbol{\beta}}(\mathbf{Y}), \widetilde{\boldsymbol{\beta}}(\mathbf{Y})]$ yields

$$\begin{aligned}
\mathbf{cov}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \mathbf{P}\mathbf{Y},] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{cov}[\mathbf{Y}]\,\mathbf{P}^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \mathbf{cov}[\widehat{\boldsymbol{\beta}}(\mathbf{Y})],
\end{aligned} \tag{16}$$

---

[1]The proof of this equality is elementary:

$$\begin{aligned}
\mathbf{cov}[\mathbf{U}+\mathbf{V}] &= E[(\mathbf{U}+\mathbf{V})(\mathbf{U}+\mathbf{V})^T] - E[\mathbf{U}+\mathbf{V}]E[\mathbf{U}+\mathbf{V}]^T \\
&= E[\mathbf{U}\mathbf{U}^T] + E[\mathbf{V}\mathbf{V}^T] + E[\mathbf{U}\mathbf{V}^T] + E[\mathbf{V}\mathbf{U}^T] - (E[\mathbf{U}] + E[\mathbf{V}])(E[\mathbf{U}] + E[\mathbf{V}])^T;
\end{aligned}$$

regrouping and interpreting the terms leads to (13).

because $\mathbf{PX} = \mathbf{I}$ implies that $\mathbf{X}^T\mathbf{P}^T = \mathbf{I}$.

Finally, inserting (15) and (16) into (14) leads to

$$\mathbf{cov}[\widetilde{\boldsymbol{\beta}}(\mathbf{Y}) - \widehat{\boldsymbol{\beta}}(\mathbf{Y})] = \mathbf{cov}[\widetilde{\boldsymbol{\beta}}(\mathbf{Y})] - \mathbf{cov}[\widehat{\boldsymbol{\beta}}(\mathbf{Y})]; \tag{17}$$

finally, since any covariance matrix is positive semi-definite, $\mathbf{cov}[\widetilde{\boldsymbol{\beta}}(\mathbf{Y}) - \widehat{\boldsymbol{\beta}}(\mathbf{Y})] \geq 0$, we obtain (12) concluding the proof.

4 Ridge Regression

When matrix $\mathbf{X}^T\mathbf{X}$ is singular, the ordinary least squares estimate, as given by (6), does not exist. One of the standard alternative criteria is the so-called *ridge regression*, which is defined as

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \right\}, \tag{18}$$

where $\lambda \geq 0$ is a parameter. The unconstrained minimization problem in (18) can also be seen as the Lagragian of the constrained problem

$$\begin{array}{ll} \min_{\boldsymbol{\beta}} & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ \text{subject to} & \|\boldsymbol{\beta}\|^2 \leq \tau, \end{array}$$

where λ is the Lagrange multiplier.

To solve (18), we begin by taking the gradient with respect to $\boldsymbol{\beta}$, which leads to

$$\nabla \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \right) = 2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} - 2\,\mathbf{X}^T\mathbf{y}. \tag{19}$$

Equating to zero and solving for $\boldsymbol{\beta}$ leads to

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{20}$$

Notice that since $\mathbf{X}^T\mathbf{X}$ is a symmetric matrix, it is positive semi-definite, *i.e.*, all its eigenvalues are non-negative. As a consequence, the condition $\lambda > 0$ is sufficient to guarantee that $\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)$ is positive definite, thus non-singular, and the ridge estimate exists regardless of matrix $\mathbf{X}$.

5 The Spectral View of OLS and Ridge Regression

To gain further insight into the OLS and ridge estimates, let's consider the singular value decomposition (SVD) of matrix $\mathbf{X}$, given by

$$\mathbf{X} = \mathbf{UDV}^T,$$

where $\mathbf{D}$ is a $p \times p$ diagonal matrix whose entries are the singular values of $\mathbf{X}$, $\mathbf{U}$ is an $n \times p$ matrix, whose columns are an ortho-normal basis for the p-dimensional subspace of $\mathbb{R}^n$ spanned by the p columns of $\mathbf{X}$ (thus $\mathbf{U}^T\mathbf{U} = \mathbf{I}$), and $\mathbf{V}$ is a $p \times p$ matrix, whose columns are a ortho-normal basis for the space spanned by the rows of $\mathbf{X}$ (thus $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$, *i.e.*, $\mathbf{V}^{-1} = \mathbf{V}^T$).

Consider first the OLS projection given by (7); using the SVD of $\mathbf{X}$, we can re-write (7) as

$$\begin{aligned}
\widehat{\mathbf{y}}_{\text{OLS}} &= \mathbf{U}\mathbf{D}\mathbf{V}^T \left(\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T\right)^{-1} \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T \left(\mathbf{V}\mathbf{D}\mathbf{D}\mathbf{V}^T\right)^{-1} \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{D}^{-2}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{U}^T\mathbf{y}, \qquad (21)
\end{aligned}$$

where the following fact was invoked: given two non-singular matrices $\mathbf{A}$ and $\mathbf{B}$, of compatible dimensions, $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. This expression for $\widehat{\mathbf{y}}_{\text{OLS}}$ shows that it is indeed an orthogonal projection of $\mathbf{y}$ onto the space spanned by the columns of $\mathbf{X}$; according to (21), this projection may be obtained by computing the inner product of $\mathbf{y}$ with every column of $\mathbf{U}$ and then combining these columns with weights equal to the corresponding projection. This is even more clearly seen by writing (21) more explicitly as

$$\widehat{\mathbf{y}}_{\text{OLS}} = \sum_{j=1}^{p} \mathbf{u}_j \left(\mathbf{y}^T\mathbf{u}_j\right), \qquad (22)$$

where $\mathbf{u}_j \in \mathbb{R}^n$ denotes the j-th column of $\mathbf{U}$.

Let us now consider the projection corresponding to the ridge estimate, which is given by

$$\widehat{\mathbf{y}}_{\text{ridge}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}. \qquad (23)$$

Inserting the SVD of $\mathbf{X}$, we have

$$\begin{aligned}
\widehat{\mathbf{y}}_{\text{ridge}} &= \mathbf{U}\mathbf{D}\mathbf{V}^T \left(\mathbf{V}\mathbf{D}\mathbf{D}\mathbf{V}^T + \lambda\mathbf{I}\right)^{-1} \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T \left(\mathbf{V}\mathbf{D}\mathbf{D}\mathbf{V}^T + \lambda\mathbf{V}\mathbf{V}^T\right)^{-1} \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T \left[\mathbf{V}\left(\mathbf{D}^2 + \lambda\mathbf{I}\right)\mathbf{V}^T\right]^{-1} \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\left(\mathbf{D} + \lambda\mathbf{I}\right)^{-2}\mathbf{D}\mathbf{U}^T\mathbf{y}. \qquad (24)
\end{aligned}$$

Noticing that matrices $\mathbf{D}$ and $(\mathbf{D} + \lambda\mathbf{I})$ are both diagonal, matrix $\mathbf{D}\left(\mathbf{D} + \lambda\mathbf{I}\right)^{-2}\mathbf{D}$ is also diagonal, given by

$$\mathbf{D}\left(\mathbf{D} + \lambda\mathbf{I}\right)^{-2}\mathbf{D} = \text{diag}\left\{\frac{d_1^2}{d_1^2 + \lambda}, \cdots, \frac{d_p^2}{d_p^2 + \lambda}\right\},$$

where $d_1, \ldots, d_p$ are the diagonal elements of $\mathbf{D}$, *i.e.*, the singular values of $\mathbf{X}$. Finally, we can explicitly write (24) as

$$\widehat{\mathbf{y}}_{\text{ridge}} = \sum_{j=1}^{p} \mathbf{u}_j \left(\frac{d_j^2}{d_j^2 + \lambda} \right) \left(\mathbf{y}^T \mathbf{u}_j \right). \tag{25}$$

This expression shows that (unlike in the OLS projection) each inner product $\mathbf{y}^T \mathbf{u}_j$ is linearly shrunk by a factor $d_j^2/(d_j^2 + \lambda) < 1$, before being used as a weight in the combination of the ortho-normal vectors $\mathbf{u}_1, \ldots, \mathbf{u}_p$. As a consequence, the inner products with basis vectors that corresponds to large singular values are almost left unaffected, while inner products with basis vectors associated with small singular values are more severely shrunk.

To understand what the ridge projection is doing, we need to understand the meaning of the singular values of $\mathbf{X}$, *i.e.*, the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Assuming that we are working with centered data (*i.e.*, satisfying (1)), the sample covariance of the set of observation points $\{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\}$ is $\mathbf{S} = \mathbf{X}^T\mathbf{X}/n$. Thus, the eigendecomposition of $\mathbf{S}$ can be written as

$$\mathbf{S} = \frac{1}{n} \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T.$$

In statistical terms, this is called the *principal component analysis* (PCA). Let us consider, without loss of generality, that the diagonal elements of $\mathbf{D}^2$ are sorted in non-increasing order, $d_1^2 \geq d_2^2 \geq \cdots \geq d_p^2$. The so-called *first normalized principal component*, $\mathbf{u}_1$ is the answer to the following question: in what direction of $\mathbb{R}^p$ does the set of points $\{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\}$ exhibit the largest variance? Moreover, d_1^2/n is the value of the variance of the data on this first principal direction. The remaining principal directions are the answer to similar questions, under the restriction of orthogonality with respect to the previously found directions. We can conclude that d_p^2 is a measure of the variance of the set of points $\{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\}$ along the p-th principal direction. In summary, what the ridge projection does is to apply a shrinkage to the coefficients of the projection in the directions where the variance is small, thus potentially yielding high variance estimates.

6 Dual Variables in Regression

6.1 Ordinary Least Squares

Recall that the OLS estimate of $\boldsymbol{\beta}$ is given by (6). Multiplying (on the left) the right hand side of (6) by the identity matrix $\mathbf{I} = \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$, we have

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{X}^T\mathbf{X} \left(\mathbf{X}^T\mathbf{X} \right)^{-1} \left(\mathbf{X}^T\mathbf{X} \right)^{-1} \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X} \left(\mathbf{X}^T\mathbf{X} \right)^{-2} \mathbf{X}^T\mathbf{y}, \tag{26}$$

which can be written as

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{X}^T\boldsymbol{\alpha}, \tag{27}$$

where

$$\boldsymbol{\alpha} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{y}. \tag{28}$$

The elements of $\boldsymbol{\alpha}$ are called the dual variables. Expression (27) reveals that $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ can be written as a linear combination of the columns of $\mathbf{X}^T$, that is, the rows of $\mathbf{X}$, that is, the set of points $\{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\}$. Formally,

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \sum_{i=1}^{n} \alpha_i \, \mathbf{x}_{(i)}. \tag{29}$$

For some new point $\mathbf{x}_{(*)}$, the predicted output is given by

$$\widehat{y}_{(*)} = \mathbf{x}_{(*)}^T \widehat{\boldsymbol{\beta}}_{\text{OLS}} = \sum_{i=1}^{n} \alpha_i \, \mathbf{x}_{(*)}^T \mathbf{x}_{(i)}, \tag{30}$$

which is a linear combination of the inner products of the new point $\mathbf{x}_{(*)}$ with all the points in the set $\{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\}$. This observation is of crucial importance to the development of the so-called kernel regression methods.

6.2 Ridge Regression

Let's consider now the case of ridge regression, as given by (20). This expression is the solution with respect to $\boldsymbol{\beta}$ of the linear system of equations

$$\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}, \tag{31}$$

which is equivalent to

$$\boldsymbol{\beta} = \frac{1}{\lambda}\,\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) = \mathbf{X}^T\boldsymbol{\alpha}, \tag{32}$$

where now

$$\boldsymbol{\alpha} = \frac{1}{\lambda}\,\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right). \tag{33}$$

At this point, the conclusion is similar to the one obtained in the OLS case: the coefficient vector estimate $\widehat{\boldsymbol{\beta}}$ can be written as a linear combination of points (as in (29)) and the predicted output at some new point $\mathbf{x}_{(*)}$ is also given by (30). What differs now from the OLS case is the computation of $\boldsymbol{\alpha}$, which is no longer given by (28). Inserting $\boldsymbol{\beta} = \mathbf{X}^T\boldsymbol{\alpha}$ into (33), we have

$$\boldsymbol{\alpha} = \frac{1}{\lambda}\,\left(\mathbf{y} - \mathbf{X}\mathbf{X}^T\boldsymbol{\alpha}\right) \tag{34}$$

which is equivalent to

$$\boldsymbol{\alpha} = \left(\lambda\mathbf{I} + \mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}. \tag{35}$$

Of course, $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$ can be computed from this $\boldsymbol{\alpha}$, *i.e.*,

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \mathbf{X}^T\left(\lambda\mathbf{I} + \mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}, \tag{36}$$

which thus is equivalent to (20). The key difference between (20) and (36) is that the latter involves inverting an $n \times n$ matrix, while the former requires the inversion of a $p \times p$ matrix. Which of the two is more convenient depends, of course, on the relative magnitude of p and n. For regression problems in high dimensional spaces, with few points, the dual form may be more efficient. Finally, it is worth pointing out again that this formulation will open the door to the introduction of kernel regression methods.

Bibliography

- T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.
- J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.