# Fitting Linear Statistical Models to Data by Least Squares III: Multivariate

Radu V. Balan, Brian R. Hunt and C. David Levermore
University of Maryland, College Park

Math 420: *Mathematical Modeling*
January 24, 2024 version

# Outline

1) Introduction to Linear Statistical Models
2) Linear Euclidean Least Squares Fitting
3) Auto-Regressive Processes
4) Linear Weighted Least Squares Fitting
5) Least Squares Fitting for Univariate Polynomial Models
6) Least Squares Fitting with Orthogonalization
7) Multivariate Linear Least Squares Fitting
8) General Multivariate Linear Least Squares Fitting

## 6. Multivariate Linear Least Squares Fitting

The least square method extends to settings with a multivariate dependent variable $\mathbf{y}$. Suppose we are given data $\{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^n$ where the $\mathbf{x}_j$ lie within a domain $\mathbb{X} \subset \mathbb{R}^p$ and the $\mathbf{y}_j$ lie in $\mathbb{R}^q$. The problem we will examine is now the following.

*How can you use this data set to make a reasonable guess about the value of $\mathbf{y}$ when $\mathbf{x}$ takes a value $\mathbb{X}$ that is not represented in the data set?*

In this setting $\mathbf{x}$ is called the *independent variable* while $\mathbf{y}$ is called the *dependent variable*. We will use weighted least squares to fit the data to a linear statistical model with $m$ parameter $q$-vectors $\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m\}$ in the form

$$\mathbf{f}(\mathbf{x}; \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m) = \sum_{i=1}^{m} \boldsymbol{\beta}_i f_i(\mathbf{x}) ,$$

where each basis function $f_i(\mathbf{x})$ is defined over $\mathbb{X}$ and takes values in $\mathbb{R}$.

We now define the $j^{\text{th}}$ residual by the vector-valued formula

$$\mathbf{r}_j(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m) = \mathbf{y}_j - \sum_{i=1}^{m} \boldsymbol{\beta}_i f_i(x_j) \,.$$

Introduce the $m{\times}q$-matrix $\boldsymbol{\mathcal{B}}$, the $n{\times}q$-vectors $\mathbf{Y}$ and $\mathbf{R}$, and the $n{\times}m$-matrix $\mathbf{F}$ by

$$\boldsymbol{\mathcal{B}} = \begin{pmatrix} \boldsymbol{\beta}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{\beta}_m^{\mathsf{T}} \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^{\mathsf{T}} \\ \vdots \\ \mathbf{y}_n^{\mathsf{T}} \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \mathbf{r}_1^{\mathsf{T}} \\ \vdots \\ \mathbf{r}_n^{\mathsf{T}} \end{pmatrix},$$

$$\mathbf{F} = \begin{pmatrix} f_1(\mathbf{x}_1) & \cdots & f_m(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ f_1(\mathbf{x}_n) & \cdots & f_m(\mathbf{x}_n) \end{pmatrix}.$$

We will assume the matrix $\mathbf{F}$ has rank $m$. The fitting problem then can be recast as finding $\boldsymbol{\mathcal{B}}$ so as to minimize the size of the vector

$$\mathbf{R}(\boldsymbol{\mathcal{B}}) = \mathbf{Y} - \mathbf{F}\boldsymbol{\mathcal{B}} \,.$$

As we did for univariate weighted least square fitting, we will minimize

$$q(\boldsymbol{\mathcal{B}}) = \tfrac{1}{2} \sum_{j=1}^{n} w_j \, \mathbf{r}_j(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m)^{\mathsf{T}} \mathbf{r}_j(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m) \,,$$

where the $w_j$ are positive weights. If we again let $\mathbf{W}$ be the $n \times n$ diagonal matrix whose $j^{\text{th}}$ diagonal entry is $w_j$ then this can be expressed as

$$\begin{aligned} q(\boldsymbol{\mathcal{B}}) &= \tfrac{1}{2}\mathrm{tr}\Big(\mathbf{R}(\boldsymbol{\mathcal{B}})^{\mathsf{T}}\mathbf{W}\mathbf{R}(\boldsymbol{\mathcal{B}})\Big) = \tfrac{1}{2}\mathrm{tr}\Big((\mathbf{Y} - \mathbf{F}\boldsymbol{\mathcal{B}})^{\mathsf{T}}\mathbf{W}(\mathbf{Y} - \mathbf{F}\boldsymbol{\mathcal{B}})\Big) \\ &= \tfrac{1}{2}\mathrm{tr}\Big(\mathbf{Y}^{\mathsf{T}}\mathbf{W}\mathbf{Y}\Big) - \mathrm{tr}\Big(\boldsymbol{\mathcal{B}}^{\mathsf{T}}\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{Y}\Big) + \tfrac{1}{2}\mathrm{tr}\Big(\boldsymbol{\mathcal{B}}^{\mathsf{T}}\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{F}\boldsymbol{\mathcal{B}}\Big) \,. \end{aligned}$$

Because $\mathbf{F}$ has rank $m$ the $m \times m$-matrix $\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{F}$ is positive definite. The function $q(\boldsymbol{\mathcal{B}})$ thereby has a strictly convex structure similar to that it had in the univariate case. It thereby has a unique global minimizer $\boldsymbol{\mathcal{B}} = \widehat{\boldsymbol{\mathcal{B}}}$ given by

$$\widehat{\boldsymbol{\mathcal{B}}} = (\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{Y} \,.$$

The fact that $\widehat{\boldsymbol{\mathcal{B}}}$ in a global minimizer again can be seen from the fact $\mathbf{F}^\top \mathbf{W} \mathbf{F}$ is positive definite and the identity

$$\begin{aligned} q(\boldsymbol{\mathcal{B}}) &= \mathrm{tr}\Big(\mathbf{Y}^\top \mathbf{W} \mathbf{Y}\Big) - \mathrm{tr}\Big(\widehat{\boldsymbol{\mathcal{B}}}^\top \mathbf{F}^\top \mathbf{W} \mathbf{F} \widehat{\boldsymbol{\mathcal{B}}}\Big) \\ &\quad + \mathrm{tr}\Big((\boldsymbol{\mathcal{B}} - \widehat{\boldsymbol{\mathcal{B}}})^\top \mathbf{F}^\top \mathbf{W} \mathbf{F} (\boldsymbol{\mathcal{B}} - \widehat{\boldsymbol{\mathcal{B}}})\Big) \\ &= q(\widehat{\boldsymbol{\mathcal{B}}}) + \mathrm{tr}\Big((\boldsymbol{\mathcal{B}} - \widehat{\boldsymbol{\mathcal{B}}})^\top \mathbf{F}^\top \mathbf{W} \mathbf{F} (\boldsymbol{\mathcal{B}} - \widehat{\boldsymbol{\mathcal{B}}})\Big) \,. \end{aligned}$$

In particular, this shows that $q(\boldsymbol{\mathcal{B}}) \geq q(\widehat{\boldsymbol{\mathcal{B}}})$ for every $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{m \times q}$ and that $q(\boldsymbol{\mathcal{B}}) = q(\widehat{\boldsymbol{\mathcal{B}}})$ if and only if $\boldsymbol{\mathcal{B}} = \widehat{\boldsymbol{\mathcal{B}}}$.

If we let $\widehat{\boldsymbol{\beta}}_i^\top$ be the $i^{\mathrm{th}}$ row of $\widehat{\boldsymbol{\mathcal{B}}}$ then the fit is given by

$$\widehat{\mathbf{f}}(x) = \sum_{i=1}^{m} \widehat{\boldsymbol{\beta}}_i f_i(x) \,.$$

The geometric interpretation of this fit is similar to that for the univariate weighted least squares fit.

**Example.** Use least squares to fit the affine model $f(\mathbf{x}; \mathbf{a}, \mathbf{B}) = \mathbf{a} + \mathbf{B}\mathbf{x}$ with $\mathbf{a} \in \mathbb{R}^q$ and $\mathbf{B} \in \mathbb{R}^{q\times p}$ to the data $\{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^n$. Begin by setting

$$\boldsymbol{\mathcal{B}} = \begin{pmatrix} \mathbf{a}^\mathsf{T} \\ \mathbf{B}^\mathsf{T} \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^\mathsf{T} \\ \vdots \\ \mathbf{y}_n^\mathsf{T} \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} 1 & \mathbf{x}_1^\mathsf{T} \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\mathsf{T} \end{pmatrix}.$$

Because

$$\mathbf{F}^\mathsf{T}\mathbf{W}\mathbf{Y} = \begin{pmatrix} \langle \mathbf{y}^\mathsf{T} \rangle \\ \langle \mathbf{x}\,\mathbf{y}^\mathsf{T} \rangle \end{pmatrix}, \qquad \mathbf{F}^\mathsf{T}\mathbf{W}\mathbf{F} = \begin{pmatrix} \langle 1 \rangle & \langle \mathbf{x}^\mathsf{T} \rangle \\ \langle \mathbf{x} \rangle & \langle \mathbf{x}\,\mathbf{x}^\mathsf{T} \rangle \end{pmatrix},$$

we find that

$$\begin{aligned} \widehat{\boldsymbol{\mathcal{B}}} &= (\mathbf{F}^\mathsf{T}\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^\mathsf{T}\mathbf{W}\mathbf{Y} = \begin{pmatrix} 1 & \langle \mathbf{x}^\mathsf{T} \rangle \\ \langle \mathbf{x} \rangle & \langle \mathbf{x}\,\mathbf{x}^\mathsf{T} \rangle \end{pmatrix}^{-1} \begin{pmatrix} \langle \mathbf{y}^\mathsf{T} \rangle \\ \langle \mathbf{x}\,\mathbf{y}^\mathsf{T} \rangle \end{pmatrix} \\ &= \begin{pmatrix} \langle \mathbf{y}^\mathsf{T} \rangle - \langle \mathbf{x} \rangle^\mathsf{T} \Big( \langle \mathbf{x}\,\mathbf{x}^\mathsf{T} \rangle - \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^\mathsf{T} \Big)^{-1} \Big( \langle \mathbf{x}\,\mathbf{y}^\mathsf{T} \rangle - \langle \mathbf{x} \rangle \langle \mathbf{y} \rangle^\mathsf{T} \Big) \\ \Big( \langle \mathbf{x}\,\mathbf{x}^\mathsf{T} \rangle - \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^\mathsf{T} \Big)^{-1} \Big( \langle \mathbf{x}\,\mathbf{y}^\mathsf{T} \rangle - \langle \mathbf{x} \rangle \langle \mathbf{y} \rangle^\mathsf{T} \Big) \end{pmatrix}. \end{aligned}$$

Because $\widehat{\boldsymbol{\mathcal{B}}}^{\mathsf{T}} = \begin{pmatrix} \hat{\mathbf{a}} & \widehat{\mathbf{B}} \end{pmatrix}$, by setting $\langle \mathbf{x} \rangle = \bar{\mathbf{x}}$ and $\langle \mathbf{y} \rangle = \bar{\mathbf{y}}$ we can express these formulas for $\hat{\mathbf{a}}$ and $\widehat{\mathbf{B}}$ simply as

$$\widehat{\mathbf{B}} = \Big\langle \mathbf{y}\,(\mathbf{x} - \bar{\mathbf{x}}\,)^{\mathsf{T}} \Big\rangle \Big\langle (\mathbf{x} - \bar{\mathbf{x}}\,)\,(\mathbf{x} - \bar{\mathbf{x}}\,)^{\mathsf{T}} \Big\rangle^{-1}\,, \qquad \hat{\mathbf{a}} = \bar{\mathbf{y}} - \widehat{\mathbf{B}}\bar{\mathbf{x}}\,.$$

The affine fit is therefore

$$\hat{\mathbf{f}}(\mathbf{x}) = \bar{\mathbf{y}} + \widehat{\mathbf{B}}(\mathbf{x} - \bar{\mathbf{x}}\,)\,.$$

**Remark.** The linear multivariate models considered above have the form

$$\mathbf{f}(\mathbf{x}; \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m) = \sum_{i=1}^{m} \boldsymbol{\beta}_i f_i(\mathbf{x})\,,$$

where each parameter vector $\boldsymbol{\beta}_i$ lies in $\mathbb{R}^q$ while each basis function $f_i(\mathbf{x})$ is defined over the bounded domain $\mathbb{X} \subset \mathbb{R}^p$ and takes values in $\mathbb{R}$. This is assumes that each entry of $\mathbf{f}$ is being fit to the same family — namely, the family spanned by the basis $\{f_i(\mathbf{x})\}_{i=1}^{m}$. Such families often are too large to be practical. We will therefore consider more general linear models.

## 7. General Multivariate Linear Least Squares Fitting

We now extend the least square method to the general multivariate setting. Suppose we are given data $\{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{n}$ where the $\mathbf{x}_j$ lie within a bounded domain $\mathbb{X} \subset \mathbb{R}^p$ while the $\mathbf{y}_j$ lie in $\mathbb{R}^q$. We will use weighted least squares to fit the data to a linear statistical model with $m$ real parameters in the form

$$\mathbf{f}(\mathbf{x}; \beta_1, \cdots, \beta_m) = \sum_{i=1}^{m} \beta_i \mathbf{f}_i(\mathbf{x}) \,,$$

where each basis function $\mathbf{f}_i(\mathbf{x})$ is defined over $\mathbb{X}$ and takes values in $\mathbb{R}^q$. We will minimize the $j^{\text{th}}$ residual, which is defined by the vector-valued formula

$$\mathbf{r}_j(\beta_1, \cdots, \beta_m) = \mathbf{y}_j - \sum_{i=1}^{m} \beta_i \mathbf{f}_i(x_j) \,.$$

Following what was done earlier, introduce the $m$-vector $\boldsymbol{\beta}$, the $nq$-vectors $\mathbf{Y}$ and $\mathbf{R}$, and the $nq{\times}m$ matrix $\mathbf{F}$ by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_n \end{pmatrix},$$

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}_1(\mathbf{x}_1) & \cdots & \mathbf{f}_m(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ \mathbf{f}_1(\mathbf{x}_n) & \cdots & \mathbf{f}_m(\mathbf{x}_n) \end{pmatrix}.$$

We will assume the matrix $\mathbf{F}$ has rank $m$. The fitting problem then can be recast as finding $\boldsymbol{\beta}$ so as to minimize the size of the vector

$$\mathbf{R}(\boldsymbol{\beta}) = \mathbf{Y} - \mathbf{F}\boldsymbol{\beta}\,.$$

We assume that $\mathbb{R}^q$ is endowed with an inner product. Without loss of generality we can assume that this inner product has the form $\mathbf{y}^\top \mathbf{G}\mathbf{z}$ where $\mathbf{G}$ is a symmetric, positive definite $q \times q$ matrix. We will minimize

$$q(\boldsymbol{\beta}) = \frac{1}{2} \sum_{j=1}^{n} w_j \, \mathbf{r}_j(\beta_1, \cdots, \beta_m)^\top \mathbf{G} \mathbf{r}_j(\beta_1, \cdots, \beta_m) \,,$$

where the $w_j$ are positive weights. If we let $\mathbf{W}$ be the symmetric, positive definite $nq \times nq$ block-diagonal matrix

$$\mathbf{W} = \begin{pmatrix} w_1\mathbf{G} & 0 & \cdots & 0 \\ 0 & w_2\mathbf{G} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_n\mathbf{G} \end{pmatrix} ,$$

then $q(\boldsymbol{\beta})$ can be expressed in terms of the weight matrix $\mathbf{W}$ as

$$\begin{aligned} q(\boldsymbol{\beta}) &= \tfrac{1}{2}\mathbf{R}(\boldsymbol{\beta})^\top \mathbf{W}\mathbf{R}(\boldsymbol{\beta}) = \tfrac{1}{2}(\mathbf{Y} - \mathbf{F}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{Y} - \mathbf{F}\boldsymbol{\beta}) \\ &= \tfrac{1}{2}\mathbf{Y}^\top \mathbf{W}\mathbf{Y} - \boldsymbol{\beta}^\top \mathbf{F}^\top \mathbf{W}\mathbf{Y} + \tfrac{1}{2}\boldsymbol{\beta}^\top \mathbf{F}^\top \mathbf{W}\mathbf{F}\boldsymbol{\beta} \,. \end{aligned}$$

Because $\mathbf{F}$ has rank $m$ the $m\times m$-matrix $\mathbf{F}^\top\mathbf{W}\mathbf{F}$ is positive definite. The function $q(\boldsymbol{\beta})$ thereby has the same strictly convex structure as it had in the univariate case. It therefore has a unique minimizer $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ where

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}^\top\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{W}\mathbf{Y}\,.$$

The fact that $\widehat{\boldsymbol{\beta}}$ in a minimizer again follows from the fact $\mathbf{F}^\top\mathbf{W}\mathbf{F}$ is positive definite and the identity

$$\begin{aligned} q(\boldsymbol{\beta}) &= \tfrac{1}{2}\mathbf{Y}^\top\mathbf{W}\mathbf{Y} - \tfrac{1}{2}\widehat{\boldsymbol{\beta}}^\top\mathbf{F}^\top\mathbf{W}\mathbf{F}\widehat{\boldsymbol{\beta}} + \tfrac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top\mathbf{F}^\top\mathbf{W}\mathbf{F}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ &= q(\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top\mathbf{F}^\top\mathbf{W}\mathbf{F}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\,. \end{aligned}$$

In particular, this shows that $q(\boldsymbol{\beta}) \geq q(\widehat{\boldsymbol{\beta}})$ for every $\boldsymbol{\beta} \in \mathbb{R}^m$ and that $q(\boldsymbol{\beta}) = q(\widehat{\boldsymbol{\beta}})$ if and only if $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.

**Remark.** The geometric interpretation of this fit is that same as that for the weighted least squares fit, except here the $\mathbf{W}$-inner product on $\mathbb{R}^{nq}$ is

$$(\mathbf{P}\,|\,\mathbf{Q})_{\mathbf{W}} = \mathbf{P}^\top\mathbf{W}\mathbf{Q}\,.$$

## Further Questions

We have seen how to use least squares to fit linear satistical models with $m$ parameters to data sets containing $n$ pairs when $m << n$. Among the questions that arise are the following.

- How does one pick a basis that is well suited to the given data?
- How can one avoid overfitting?
- Do these methods extended to nonlinear statistical models?
- Can one use other notions of smallness of the residual?