

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

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

Math 420: *Mathematical Modeling*

February 5, 2014 version

© 2014 B.R. Hunt and C.D. Levermore

## Outline

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

## 6. Multivariate Linear Least Squares Fitting

The least square method extends to settings with a multivariate dependent variable  $y$ . 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  $y$  when  $x$  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 in the form

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

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

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

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

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_1^\top \\ \vdots \\ \boldsymbol{\beta}_m^\top \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_n^\top \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \mathbf{r}_1^\top \\ \vdots \\ \mathbf{r}_n^\top \end{pmatrix},$$
$$\mathbf{F} = \begin{pmatrix} f_1(\mathbf{x}_1) & \cdots & f_m(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ f_1(\mathbf{x}_n) & \cdots & f_m(\mathbf{x}_n) \end{pmatrix}.$$

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

$$\mathbf{R}(\mathbf{B}) = \mathbf{Y} - \mathbf{FB}.$$

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

$$q(\mathbf{B}) = \frac{1}{2} \sum_{j=1}^n w_j \mathbf{r}_j(\beta_1, \dots, \beta_m)^\top \mathbf{r}_j(\beta_1, \dots, \beta_m),$$

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

$$\begin{aligned} q(\mathbf{B}) &= \frac{1}{2} \text{tr}(\mathbf{R}(\mathbf{B})^\top \mathbf{W} \mathbf{R}(\mathbf{B})) = \frac{1}{2} \text{tr}((\mathbf{Y} - \mathbf{F}\mathbf{B})^\top \mathbf{W} (\mathbf{Y} - \mathbf{F}\mathbf{B})) \\ &= \frac{1}{2} \text{tr}(\mathbf{Y}^\top \mathbf{W} \mathbf{Y}) - \text{tr}(\mathbf{B}^\top \mathbf{F}^\top \mathbf{W} \mathbf{Y}) + \frac{1}{2} \text{tr}(\mathbf{B}^\top \mathbf{F}^\top \mathbf{W} \mathbf{F} \mathbf{B}). \end{aligned}$$

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

$$\hat{\mathbf{B}} = (\mathbf{F}^\top \mathbf{W} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{W} \mathbf{Y}.$$

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

$$\begin{aligned} q(\mathcal{B}) &= \text{tr}(\mathbf{Y}^\top \mathbf{W} \mathbf{Y}) - \text{tr}(\hat{\mathcal{B}}^\top \mathbf{F}^\top \mathbf{W} \mathbf{F} \hat{\mathcal{B}}) \\ &\quad + \text{tr}((\mathcal{B} - \hat{\mathcal{B}})^\top \mathbf{F}^\top \mathbf{W} \mathbf{F} (\mathcal{B} - \hat{\mathcal{B}})) \\ &= q(\hat{\mathcal{B}}) + \text{tr}((\mathcal{B} - \hat{\mathcal{B}})^\top \mathbf{F}^\top \mathbf{W} \mathbf{F} (\mathcal{B} - \hat{\mathcal{B}})). \end{aligned}$$

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

If we let  $\hat{\beta}_i$  be the  $i^{\text{th}}$  row of  $\hat{\mathcal{B}}$  then the fit is given by

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

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  $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$ . Begin by setting

$$\mathbf{B} = \begin{pmatrix} \mathbf{a}^\top \\ \mathbf{B}^\top \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1^\top \\ \vdots \\ y_n^\top \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\top \end{pmatrix}.$$

Because

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

we find that

$$\begin{aligned} \hat{\mathbf{B}} &= (\mathbf{F}^\top \mathbf{W} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{W} \mathbf{Y} = \begin{pmatrix} 1 & \langle \mathbf{x}^\top \rangle \\ \langle \mathbf{x} \rangle & \langle \mathbf{x} \mathbf{x}^\top \rangle \end{pmatrix}^{-1} \begin{pmatrix} \langle \mathbf{y}^\top \rangle \\ \langle \mathbf{x} \mathbf{y}^\top \rangle \end{pmatrix} \\ &= \begin{pmatrix} \langle \mathbf{y}^\top \rangle - \langle \mathbf{x} \rangle^\top (\langle \mathbf{x} \mathbf{x}^\top \rangle - \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^\top)^{-1} (\langle \mathbf{x} \mathbf{y}^\top \rangle - \langle \mathbf{x} \rangle \langle \mathbf{y} \rangle^\top) \\ (\langle \mathbf{x} \mathbf{x}^\top \rangle - \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^\top)^{-1} (\langle \mathbf{x} \mathbf{y}^\top \rangle - \langle \mathbf{x} \rangle \langle \mathbf{y} \rangle^\top) \end{pmatrix}. \end{aligned}$$

Because  $\hat{\boldsymbol{\beta}}^\top = (\hat{\mathbf{a}} \quad \hat{\mathbf{B}})$ , by setting  $\langle \mathbf{x} \rangle = \bar{\mathbf{x}}$  and  $\langle \mathbf{y} \rangle = \bar{y}$  we can express these formulas for  $\hat{\mathbf{a}}$  and  $\hat{\mathbf{B}}$  simply as

$$\hat{\mathbf{B}} = \langle \mathbf{y} (\mathbf{x} - \bar{\mathbf{x}})^\top \rangle \langle (\mathbf{x} - \bar{\mathbf{x}}) (\mathbf{x} - \bar{\mathbf{x}})^\top \rangle^{-1}, \quad \hat{\mathbf{a}} = \bar{y} - \hat{\mathbf{B}}\bar{\mathbf{x}}.$$

The affine fit is therefore

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

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

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

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 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

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

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

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

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

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix},$$
$$\mathbf{F} = \begin{pmatrix} f_1(\mathbf{x}_1) & \cdots & f_m(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ f_1(\mathbf{x}_n) & \cdots & f_m(\mathbf{x}_n) \end{pmatrix}.$$

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

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

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

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

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

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

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  $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$  where

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

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

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

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

**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

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

## Further Questions

We have seen how to use least squares to fit linear statistical models with  $m$  parameters to data sets containing  $n$  pairs when  $m \ll n$ . 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 extend to nonlinear statistical models?
- Can one use other notions of smallness of the residual?