Fitting Linear Statistical Models to Data by Least Squares I: Introduction

Radu Balan, Brian R. Hunt and C. David Levermore

University of Maryland, College Park, MD

Math 420: *Mathematical Modeling*
February 2, 2021 version

Lectures on

Fitting Linear Statistical Models to Data by Least Squares

I. Euclidean Least Squares Fitting

II. Weighted Least Squares Fitting

III. Multivariate Least Squares Fitting

I. Euclidean Least Squares Fitting for Linear Models

Linear Statistical Models: Introduction

In modeling we are often faced with the problem of fitting data with some analytic expression. Suppose that we are studying a phenomenon that evolves over time and are given n distinct times $\{t_j\}_{j=1}^n$ and a measurement y_j of the phenomenon at each time t_j. We represent this data as the set of ordered pairs

$$\{(t_j, y_j)\}_{j=1}^n .$$

Each y_j might be a single number, which is the *univariate* case, or a vector of numbers, which is the *multivariate* case. We will treat the simpler univariate case first.

The basic problem we will examine is the following.

How can we use this data set to make a reasonable guess about what a measurment of this phenomenon might yield at other times?

Linear Statistical Models: Overfitting

Of course, you can always find functions $f(t)$ such that $y_j = f(t_j)$ for every $j = 1, \cdots, n$. For example, you can use Lagrange interpolation to construct a unique polynomial of degree at most $n - 1$ that does this. However, such a polynomial often exhibits wild oscillations that make it a useless fit. This problem is called *overfitting*.

Reasons that the problem of overfitting might arise include:

- the assumed form of $f(t)$ is ill-suited to matching the behavior of the phenomenon over the time interval being considered;
- the times t_j and measurements y_j are subject to error, so finding a function that fits the data exactly is not a good strategy even when the assumed form of $f(t)$ is well-suited to matching the behavior of the phenomenon over the time interval being considered.

Linear Statistical Models: Residuals

A strategy to help avoid these difficulties is to draw $f(t)$ from a family of suitable functions. Such a family is called a *model* in statistics. If we denote this model by $f(t; \beta_1, \cdots, \beta_m)$ where $m \ll n$ then the idea is to find values of $\beta_1, \cdots, \beta_m$ such that the graph of $f(t; \beta_1, \cdots, \beta_m)$ best fits the data. More precisely, we define the *residuals* $r_j(\beta_1, \cdots, \beta_m)$ by

$$y_j = f(t_j; \beta_1, \cdots, \beta_m) + r_j(\beta_1, \cdots, \beta_m), \quad \text{for every } j = 1, \cdots, n,$$

and try to minimize the $r_j(\beta_1, \cdots, \beta_m)$ in some sense.

The problem is simplified by restricting ourselves to models in which the parameters appear linearly — so-called *linear models*. Such a model is specified by the choice of a *basis* $\{f_i(t)\}_{i=1}^m$ and takes the form

$$f(t; \beta_1, \cdots, \beta_m) = \sum_{i=1}^{m} \beta_i f_i(t).$$

Linear Statistical Models: Examples

Example. One classic linear model is the family of all *polynomials* of degree at most ℓ. This family is often expressed as

$$f(t; \beta_0, \cdots, \beta_\ell) = \sum_{k=0}^{\ell} \beta_k \, t^k .$$

Here the index k runs from 0 to ℓ so that it matches the degree of each term in the sum. Therefore $m = \ell + 1$.

Example. If the underlying phenomena is *periodic* with period T then a classic linear model is the family of all *trigonometric polynomials* of degree at most ℓ. This family can be expressed as

$$f(t; \alpha_0, \cdots, \alpha_\ell, \beta_1, \cdots, \beta_\ell) = \alpha_0 + \sum_{k=1}^{\ell} \big(\alpha_k \cos(k\omega t) + \beta_k \sin(k\omega t)\big) ,$$

where $\omega = 2\pi / T$ its *fundamental frequency*. Notice that $m = 2\ell + 1$.

Linear Statistical Models: Translation Invariance

Remark. Linear models are linear in the parameters, but are typically nonlinear in the independent variable t. This is illustrated by the foregoing examples: the family of all polynomials of degree at most ℓ is nonlinear in t for $\ell > 1$; the family of all trigonometric polynomials of degree at most ℓ is nonlinear in t for $\ell > 0$.

Remark. When there is no preferred instant of time it is best to pick a model $f(t; \beta_1, \cdots, \beta_m)$ that is *translation invariant*. This means for every choice of parameter values $(\beta_1, \cdots, \beta_m)$ and time shift s there exist parameter values $(\beta_1', \cdots, \beta_m')$ such that

$$f(t + s; \beta_1, \cdots, \beta_m) = f(t; \beta_1', \cdots, \beta_m') \quad \text{for every } t\,.$$

Both models given on the previous slide are translation invariant. Can you show this? Can you find models that are not translation invariant?

General Univariate Linear Models: Introduction

It is just as easy to work in the general univariate setting in which we are given data

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

where the $\mathbf{x}_j$ are distinct points within a bounded domain $\mathbb{X} \subset \mathbb{R}^d$ and the y_j lie in $\mathbb{R}$. Here $\mathbf{x}$ is called the *independent variable* and y is called the *dependent variable*.

The problem we will examine now becomes the following.

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

General Univariate Linear Models: Example

We will consider linear statistical models with m parameters in the form

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

where each basis function $f_i(\mathbf{x})$ is defined over $\mathbb{X}$ and takes values in $\mathbb{R}$.
Example. One classic linear model is the family of all affine functions. If x_k denotes the k^{th} entry of $\mathbf{x}$ then this family can be written as

$$f(\mathbf{x}; a, b_1, \cdots, b_d) = a + \sum_{k=1}^{d} b_k\, x_k\,.$$

Alternatively, it can be expressed in vector notation as

$$f(\mathbf{x}; a, \mathbf{b}) = a + \mathbf{b} \cdot \mathbf{x}\,,$$

where $a \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^d$. Notice that here $m = d + 1$.

General Univariate Linear Models: Examples

Example. Similarly the family of all quadratic functions can be expressed in vector notation as

$$f(\mathbf{x}; a, \mathbf{b}, \mathbf{C}) = a + \mathbf{b} \cdot \mathbf{x} + \mathbf{x} \cdot \mathbf{C}\mathbf{x} \,,$$

where $a \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^d$ and $\mathbf{C} \in \mathbb{R}^{d \vee d}$. Here $\mathbb{R}^{d \vee d}$ denotes the set of all symmetric $d \times d$ real matrices. In this case $m = \frac{1}{2}(d+1)(d+2)$.

Remark. The dimension m for the family of polynomials in d variables of degree at most ℓ is

$$m = \frac{(d+\ell)!}{d!\,\ell!} = \frac{(d+1)(d+2)\cdots(d+\ell)}{\ell!} \,.$$

This grows like d^ℓ as d grows. This means that these models can become impractical when the dimension d is large. In such cases we can use custom built models rather than general ones.

General Univariate Linear Models: Residuals

Recall that given the data $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$ and any model $f(\mathbf{x}; \beta_1, \cdots, \beta_m)$, the *residual* associated with each $(\mathbf{x}_j, y_j)$ is defined by the relation

$$y_j = f(\mathbf{x}_j; \beta_1, \cdots, \beta_m) + r_j(\beta_1, \cdots, \beta_m) .$$

The linear model given by the *basis* $\{f_i(\mathbf{x})\}_{i=1}^m$ is

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

for which the residual $r_j(\beta_1, \cdots, \beta_m)$ is given by

$$r_j(\beta_1, \cdots, \beta_m) = y_j - \sum_{i=1}^m \beta_i f_i(\mathbf{x}_j) .$$

The idea is to determine the parameters $\beta_1, \cdots, \beta_m$ in the statistical model by minimizing the residuals $r_j(\beta_1, \cdots, \beta_m)$.

General Univariate Linear Models: Fitting Problem

This so-called *fitting problem* can be recast in terms of vectors. Define the m-vector $\boldsymbol{\beta}$, the n-vectors $\mathbf{y}$ and $\mathbf{r}$, and the $n \times m$-matrix $\mathbf{F}$ by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix},$$

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

We will assume the matrix $\mathbf{F}$ has rank m. The fitting problem is then the problem of finding a value of $\boldsymbol{\beta}$ that minimizes the size of

$$\mathbf{r}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{F}\boldsymbol{\beta}.$$

But what does "size" mean?

Euclidean Least Squares Fitting: Introduction

A popular notion of the size of a vector is the *Euclidean norm*, which is

$$\|\mathbf{r}(\boldsymbol{\beta})\| = \sqrt{\mathbf{r}(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{r}(\boldsymbol{\beta})} = \sqrt{\sum_{j=1}^{n} r_j(\beta_1, \cdots, \beta_m)^2} \,.$$

Minimizing $\|\mathbf{r}(\boldsymbol{\beta})\|$ is equivalent to minimizing $\|\mathbf{r}(\boldsymbol{\beta})\|^2$, which is the sum of the "squares" of the residuals.

For linear models $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{F}\boldsymbol{\beta}$, so we minimize

$$\begin{aligned} q(\boldsymbol{\beta}) = \tfrac{1}{2}\|\mathbf{r}(\boldsymbol{\beta})\|^2 = \tfrac{1}{2}\mathbf{r}(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{r}(\boldsymbol{\beta}) &= \tfrac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}) \\ &= \tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{y} + \tfrac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\beta} \,. \end{aligned}$$

Because this quadratic function of $\boldsymbol{\beta}$ is easy to minimize, this method is popular. We will use multivariable calculus to minimize it.

Euclidean Least Squares Fitting: Gradient

Recall that the *gradient* (if it exists) of a real-valued function $q(\boldsymbol{\beta})$ with respect to the m-vector $\boldsymbol{\beta}$ is the m-vector $\partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta})$ such that

$$\frac{\mathrm{d}}{\mathrm{d}s} q(\boldsymbol{\beta} + s\boldsymbol{\gamma})\Big|_{s=0} = \boldsymbol{\gamma}^{\mathrm{T}} \partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta}) \quad \text{for every } \boldsymbol{\gamma} \in \mathbb{R}^m .$$

In particular, for the quadratic $q(\boldsymbol{\beta})$ arising from our least squares problem we can easily check that

$$q(\boldsymbol{\beta} + s\boldsymbol{\gamma}) = q(\boldsymbol{\beta}) + s\boldsymbol{\gamma}^{\mathrm{T}}\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\beta} - \mathbf{F}^{\mathrm{T}}\mathbf{y}\right) + \tfrac{1}{2}s^2\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\gamma} .$$

By differentiating this with respect to s and setting $s = 0$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}s} q(\boldsymbol{\beta} + s\boldsymbol{\gamma})\Big|_{s=0} = \boldsymbol{\gamma}^{\mathrm{T}}\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\beta} - \mathbf{F}^{\mathrm{T}}\mathbf{y}\right) ,$$

from which we read off that the gradient is

$$\partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta}) = \mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\beta} - \mathbf{F}^{\mathrm{T}}\mathbf{y} .$$

Euclidean Least Squares Fitting: Hessian

Similarly, the derivative (if it exists) of the vector-valued function $\partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta})$ with respect to the m-vector $\boldsymbol{\beta}$ is the $m \times m$-matrix $\partial_{\boldsymbol{\beta\beta}} q(\boldsymbol{\beta})$ such that

$$\frac{\mathrm{d}}{\mathrm{d}s} \partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta} + s\boldsymbol{\gamma})\Big|_{s=0} = \partial_{\boldsymbol{\beta\beta}} q(\boldsymbol{\beta})\boldsymbol{\gamma} \quad \text{for every } \boldsymbol{\gamma} \in \mathbb{R}^m .$$

The symmetric matrix-valued function $\partial_{\boldsymbol{\beta\beta}} q(\boldsymbol{\beta})$ is the *Hessian* of $q(\boldsymbol{\beta})$.
For the quadratic $q(\boldsymbol{\beta})$ arising from our least squares problem we have

$$\partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta} + s\boldsymbol{\gamma}) = \mathbf{F}^{\mathrm{T}}\mathbf{F}(\boldsymbol{\beta} + s\boldsymbol{\gamma}) - \mathbf{F}^{\mathrm{T}}\mathbf{y} = \partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta}) + s\,\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\gamma} .$$

By differentiating this with respect to s and setting $s = 0$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}s} \partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta} + s\boldsymbol{\gamma})\Big|_{s=0} = \frac{\mathrm{d}}{\mathrm{d}s}\big(\partial_{\boldsymbol{\beta}} q(\boldsymbol{\beta}) + s\,\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\gamma}\big)\Big|_{s=0} = \mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\gamma} ,$$

from which we read off that the Hessian is

$$\partial_{\boldsymbol{\beta\beta}} q(\boldsymbol{\beta}) = \mathbf{F}^{\mathrm{T}}\mathbf{F} .$$

Euclidean Least Squares Fitting: Positive Definiteness

We now show that the $m \times m$ matrix $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is *positive definite*. We have

$$\gamma^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\gamma = (\mathbf{F}\gamma)^{\mathrm{T}}\mathbf{F}\gamma = \|\mathbf{F}\gamma\|^2 \geq 0 \qquad \text{for every } \gamma \in \mathbb{R}^m ,$$

which implies that $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is nonnegative definite. It will be positive definite if we can show that

$$\gamma^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\gamma \quad \Longrightarrow \quad \gamma = \mathbf{0} .$$

However, because $\gamma^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\gamma = \|\mathbf{F}\gamma\|^2$ we see that

$$\gamma^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\gamma \quad \Longrightarrow \quad \|\mathbf{F}\gamma\| = 0 \quad \Longrightarrow \quad \mathbf{F}\gamma = \mathbf{0} .$$

Because $\mathbf{F}$ has rank m, its columns are linearly independent, whereby

$$\mathbf{F}\gamma = \mathbf{0} \quad \Longrightarrow \quad \gamma = \mathbf{0} .$$

Therefore $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is positive definite.

Euclidean Least Squares Fitting: Minimizer

Because $\partial_{\boldsymbol{\beta}\boldsymbol{\beta}} q(\boldsymbol{\beta})$ is positive definite, the function $q(\boldsymbol{\beta})$ is *strictly convex*, whereby it has at most one (global) *minimizer*. We find this minimizer by setting the gradient of $q(\boldsymbol{\beta})$ equal to zero, yielding

$$\partial_{\boldsymbol{\beta}}\, q(\boldsymbol{\beta}) = \mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\beta} - \mathbf{F}^{\mathrm{T}}\mathbf{y} = \mathbf{0}\,.$$

Because the matrix $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is positive definite, it is invertible, whereby the above equation can be solved. The minimizer is found to be $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ where

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{y}\,.$$

Remark. In practice you should not compute $(\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}$ when $m > 2$. Rather, you should think of the right-hand side above as notation for the solution of the linear algebraic system $\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\beta} = \mathbf{F}^{\mathrm{T}}\mathbf{y}$. All that you need to compute is the solution $\widehat{\boldsymbol{\beta}}$ of this system.

Euclidean Least Squares Fitting: Uniqueness

The fact that $\widehat{\beta}$ is the *unique* global minimizer is seen by using the fact that $\mathbf{F}^{\mathrm{T}}\mathbf{y} = \mathbf{F}^{\mathrm{T}}\mathbf{F}\widehat{\beta}$ to obtain the identity

$$\begin{aligned} q(\beta) &= \tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} - \beta^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{y} + \tfrac{1}{2}\beta^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\beta \\ &= \tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} - \beta^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\widehat{\beta} + \tfrac{1}{2}\beta^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\beta \\ &= \tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} - \tfrac{1}{2}\widehat{\beta}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\widehat{\beta} + \tfrac{1}{2}(\beta - \widehat{\beta})^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}(\beta - \widehat{\beta}) \\ &= q(\widehat{\beta}) + \tfrac{1}{2}(\beta - \widehat{\beta})^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}(\beta - \widehat{\beta}) \,. \end{aligned}$$

Then the fact $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is positive definite implies that

$$\begin{aligned} q(\beta) \geq q(\widehat{\beta}) \quad &\text{for every } \beta \in \mathbb{R}^m \,, \\ q(\beta) = q(\widehat{\beta}) \quad &\Longleftrightarrow \quad \beta = \widehat{\beta} \,. \end{aligned}$$

Euclidean Least Squares Fitting: Affine Example

Example. For the affine model $f(t; \alpha, \beta) = \alpha + \beta t$ and data $\{(t_j, y_j)\}_{j=1}^n$ the matrix $\mathbf{F}$ has the form

$$\mathbf{F} = \begin{pmatrix} \mathbf{1} & \mathbf{t} \end{pmatrix}, \qquad \text{where} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}.$$

Then

$$\mathbf{F}^{\mathrm{T}}\mathbf{F} = \begin{pmatrix} \mathbf{1}^{\mathrm{T}}\mathbf{1} & \mathbf{1}^{\mathrm{T}}\mathbf{t} \\ \mathbf{t}^{\mathrm{T}}\mathbf{1} & \mathbf{t}^{\mathrm{T}}\mathbf{t} \end{pmatrix} = n \begin{pmatrix} 1 & \bar{t} \\ \bar{t} & \overline{t^2} \end{pmatrix},$$

and $\det\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right) = n^2\left(\overline{t^2} - \bar{t}^2\right) = n^2\sigma^2 > 0$, where

$$\bar{t} = \frac{1}{n}\sum_{j=1}^{n} t_j, \qquad \overline{t^2} = \frac{1}{n}\sum_{j=1}^{n} t_j^2, \qquad \sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(t_j - \bar{t})^2.$$

Here $\bar{t}$ and σ^2 are the sample mean and variance of t respectively.

Euclidean Least Squares Fitting: Affine Example

Then the $\widehat{\alpha}$ and $\widehat{\beta}$ that give the least squares fit are given by

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \widehat{\boldsymbol{\beta}} = (\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{y} = \frac{1}{n}\frac{1}{\sigma^2}\begin{pmatrix} \overline{t^2} & -\bar{t} \\ -\bar{t} & 1 \end{pmatrix}\begin{pmatrix} \mathbf{1}^{\mathrm{T}} \\ \mathbf{t}^{\mathrm{T}} \end{pmatrix}\mathbf{y}$$

$$= \frac{1}{\sigma^2}\begin{pmatrix} \overline{t^2} & -\bar{t} \\ -\bar{t} & 1 \end{pmatrix}\begin{pmatrix} \bar{y} \\ \overline{ty} \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} \overline{t^2}\,\bar{y} - \bar{t}\,\overline{ty} \\ \overline{ty} - \bar{t}\,\bar{y} \end{pmatrix},$$

where

$$\bar{y} = \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{y} = \frac{1}{n}\sum_{j=1}^{n} y_j\,, \qquad \overline{yt} = \frac{1}{n}\mathbf{t}^{\mathrm{T}}\mathbf{y} = \frac{1}{n}\sum_{j=1}^{n} y_j t_j\,.$$

These formulas can be expressed simply as

$$\widehat{\beta} = \frac{\overline{yt} - \bar{y}\,\bar{t}}{\sigma^2}\,, \qquad \widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{t}\,,$$

so $\widehat{\beta}$ is the sample covariance of y and t over the sample variance of t.

Euclidean Least Squares Fitting: Numerical Methods

Therefore the best fit is

$$\widehat{f}(t) = \widehat{\alpha} + \widehat{\beta}t = \bar{y} + \widehat{\beta}(t - \bar{t}\,) = \bar{y} + \frac{\overline{yt} - \bar{y}\,\bar{t}}{\sigma^2}(t - \bar{t}\,)\,.$$

Remark. In this example we inverted the matrix $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ to obtain $\widehat{\boldsymbol{\beta}}$. This was easy because our model had only two parameters in it, so $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ was only 2×2. The number of paramenters m does not have to be too large before this approach becomes slow or unfeasible. However for such m you can find $\widehat{\boldsymbol{\beta}}$ by using Gaussian elimination or some other *direct numerical method* to efficiently solve the linear system

$$\mathbf{F}^{\mathrm{T}}\mathbf{F}\boldsymbol{\beta} = \mathbf{F}^{\mathrm{T}}\mathbf{y}\,.$$

Such direct methods work because the matrix $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is positive definite. As we will see later, this step can be simplified by constructing the basis $\{f_i(t)\}_{i=1}^{m}$ so that $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is diagonal.

Euclidean Least Squares Fitting: Geometric View

The Euclidean least squares fit has a beautiful *geometric interpretation* in $\mathbb{R}^n$ equipped with the Euclidean scalar product

$$(\mathbf{p} \,|\, \mathbf{q}) = \mathbf{p}^{\mathrm{T}}\mathbf{q}\,.$$

The range of the $n \times m$ matrix $\mathbf{F}$ is given by

$$\mathrm{Range}(\mathbf{F}) = \left\{ \mathbf{F}\boldsymbol{\gamma} \,:\, \boldsymbol{\gamma} \in \mathbb{R}^m \right\}.$$

It is the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{F}$. Because $\mathbf{F}$ has rank m, its columns are linearly independent and $\mathrm{Range}(\mathbf{F})$ has dimension m.

Define $\widehat{\mathbf{r}} = \mathbf{r}(\widehat{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}$. For every $\boldsymbol{\gamma} \in \mathbb{R}^m$ we have

$$\begin{aligned}(\mathbf{F}\boldsymbol{\gamma} \,|\, \widehat{\mathbf{r}}) = (\mathbf{F}\boldsymbol{\gamma})^{\mathrm{T}}\widehat{\mathbf{r}} = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\widehat{\mathbf{r}} &= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\big(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}\big) \\ &= \boldsymbol{\gamma}^{\mathrm{T}}\big(\mathbf{F}^{\mathrm{T}}\mathbf{y} - \mathbf{F}^{\mathrm{T}}\mathbf{F}\widehat{\boldsymbol{\beta}}\big) = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{0} = 0\,.\end{aligned}$$

Therefore $\widehat{\mathbf{r}}$ is orthogonal to $\mathrm{Range}(\mathbf{F})$.

Euclidean Least Squares Fitting: Geometric View

We will express the fact that $\widehat{\mathbf{r}}$ is orthogonal to $\mathrm{Range}(\mathbf{F})$ as either

$$\widehat{\mathbf{r}} \perp \mathrm{Range}(\mathbf{F}) \qquad \text{or} \qquad \widehat{\mathbf{r}} \in \mathrm{Range}(\mathbf{F})^{\perp} .$$

Because $\widehat{\mathbf{r}} \perp \mathrm{Range}(\mathbf{F})$, we see that $\mathbf{y} = \mathbf{F}\widehat{\beta} + \widehat{\mathbf{r}}$ is the *orthogonal decomposition* of $\mathbf{y} \in \mathbb{R}^n$ into $\mathbf{F}\widehat{\beta} \in \mathrm{Range}(\mathbf{F})$ plus $\widehat{\mathbf{r}} \in \mathrm{Range}(\mathbf{F})^{\perp}$. Because $\mathbf{F}\widehat{\beta}$ and $\widehat{\mathbf{r}}$ are orthogonal we have the *Pythagorean relation*

$$\begin{aligned} \|\mathbf{y}\|^2 = (\mathbf{y} \,|\, \mathbf{y}) &= \big(\mathbf{F}\widehat{\beta} + \widehat{\mathbf{r}} \,\big|\, \mathbf{F}\widehat{\beta} + \widehat{\mathbf{r}}\big) \\ &= \big(\mathbf{F}\widehat{\beta} \,\big|\, \mathbf{F}\widehat{\beta}\big) + \big(\widehat{\mathbf{r}} \,\big|\, \mathbf{F}\widehat{\mathbf{r}}\big) = \big\|\mathbf{F}\widehat{\beta}\big\|^2 + \big\|\widehat{\mathbf{r}}\big\|^2 . \end{aligned}$$

Remark. Because the residual $\widehat{\mathbf{r}}$ is orthogonal to $\mathrm{Range}(\mathbf{F})$. it will have mean zero if $\mathbf{1} \in \mathrm{Range}(\mathbf{F})$, which is the case whenever the constant function 1 is in the linear span of the basis $\{f_i(\mathbf{x})\}_{i=1}^m$ for the model.

Euclidean Least Squares Fitting: Geometric View

Remark. Because $\widehat{\beta} = (\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{y}$, the components of the orthogonal decomposition $\mathbf{y} = \mathbf{F}\widehat{\beta} + \widehat{\mathbf{r}}$ can be expressed as

$$\mathbf{F}\widehat{\beta} = \mathbf{P}\mathbf{y}\,, \qquad \widehat{\mathbf{r}} = \mathbf{y} - \mathbf{F}\widehat{\beta} = (\mathbf{I} - \mathbf{P})\mathbf{y}\,,$$

where $\mathbf{P} = \mathbf{F}(\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}$ and $\mathbf{I}$ is the $n\times n$ identity matrix. It is easy to check that the $n\times n$ matrix $\mathbf{P}$ satisfies

$$\mathbf{P}^2 = \mathbf{P}\,, \qquad \mathbf{P}^{\mathrm{T}} = \mathbf{P}\,, \qquad \mathbf{P}\mathbf{F} = \mathbf{F}\,.$$

These properties can be used to show that:

- $\mathbf{P}$ is the orthogonal projection onto $\mathrm{Range}(\mathbf{F})$;
- $\mathbf{I} - \mathbf{P}$ is the orthogonal projection onto $\mathrm{Range}(\mathbf{F})^{\perp}$.

Further Questions

We have seen how to use Euclidean 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 do we pick a basis that is well suited to the given data? (This is explored in homework.)
- How can we avoid overfitting? (By keeping $m \ll n$ and being careful.)
- Do these methods extended to nonlinear statistical models? (Minimization can become extremely difficult.)
- Can we use other notions of smallness of the residual? (We see some in the next chapter.)