Outline of Three Lectures

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
1. Introduction to Linear Statistical Models

In modeling one is often faced with the problem of fitting data with some analytic expression. Let us suppose that we are studying a phenomenon that evolves over time. Given a set of $n$ times $\{t_j\}_{j=1}^n$ such that at each time $t_j$ we take a measurement $y_j$ of the phenomenon. We can represent this data as the set of ordered pairs

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

Each $y_j$ might be a single number or a vector of numbers. For simplicity, we will first treat the univariate case when it is a single number. The more complicated multivariate case when it is a vector will be treated later.

The basic problem we will examine is the following.

*How can you use this data set to make a reasonable guess about what a measurement of this phenomenon might yield at any other time?*
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 phenomena is called overfitting. There are two reasons why such difficulties arise.

- 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.

- The assumed form of $f(t)$ might be ill suited for matching the behavior of the phenomenon over the time interval being considered.
One strategy to help avoid these difficulties is to draw $f(t)$ from a family of suitable functions, which is called a \textit{model} in statistics. If we denote this model by $f(t; \beta_1, \cdots, \beta_m)$ where $m << n$ 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 will define the \textit{residuals} $r_j(\beta_1, \cdots, \beta_m)$ by the relation

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

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

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

$$f(t; \beta_1, \cdots, \beta_m) = \sum_{i=1}^m \beta_i f_i(t).$$
Example. The most classic linear model is the family of all polynomials of degree less than \( m \). This family is often expressed as

\[
f(t; \beta_0, \cdots, \beta_{m-1}) = \sum_{i=0}^{m-1} \beta_i t^i.
\]

Notice that here the index \( i \) runs from 0 to \( m - 1 \) rather than from 1 to \( m \). This indexing convention is used for polynomial models because it matches the degree of each term in the sum.

Example. Another classic linear model is the family of all trigonometric polynomials of degree at most \( l \) for some preiod \( T \). This family can be expressed as

\[
f(t; \alpha_0, \cdots, \alpha_l, \beta_1, \cdots, \beta_l) = \alpha_0 + \sum_{k=1}^{l} \left( \alpha_k \cos(k\omega t) + \beta_k \sin(k\omega t) \right),
\]

where \( \omega = 2\pi/T \) its fundamental frequency. Notice that here \( m = 2l + 1 \).
It is as easy to work in the more general setting in which we are given data

\[ \{(x_j, y_j)\}_{j=1}^n, \]

where the \( x_j \) lie within a bounded domain \( X \subset \mathbb{R}^p \) and the \( y_j \) lie in \( \mathbb{R} \). The problem we will examine now becomes the following.

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

We call \( x \) the *independent variable* and \( y \) the *dependent variable*. We will consider a linear statistical model with \( m \) real parameters in the form

\[
 f(x; \beta_1, \ldots, \beta_m) = \sum_{i=1}^{m} \beta_i f_i(x),
\]

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

$$f(x; \beta_0, \cdots, \beta_p) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i.$$ 

Alternatively, it can be expressed in vector notation as

$$f(x; a, b) = a + b^T x,$$

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

Remark. When the independent variable $x$ has dimension $p > 1$, the dimension $m$ of an associated general linear model grows rapidly as the model complexity increases. For example, the family of polynomials of degree at most $d$ over $\mathbb{R}^p$ has dimension

$$m = \frac{(p + d)!}{p! d!} = \frac{(p + 1)(p + 2) \cdots (p + d)}{d!}.$$
Given a linear model $f(x_j; \beta_1, \cdots, \beta_m)$, we fit it to the data by minimizing the residuals $r_j(\beta_1, \cdots, \beta_m)$ that are defined by the relation

$$y_j = f(x_j; \beta_1, \cdots, \beta_m) + r_j(\beta_1, \cdots, \beta_m).$$

This can be recast as a linear algebra problem. We introduce the $m$-vector $\beta$, the $n$-vectors $y$ and $r$, and the $n \times m$-matrix $F$ by

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

$$F = \begin{pmatrix} f_1(x_1) & \cdots & f_m(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_n) & \cdots & f_m(x_n) \end{pmatrix}.$$

We will assume the matrix $F$ has rank $m$. The fitting problem then becomes the problem of finding a value of $\beta$ that minimizes the size of the $n$-vector $r(\beta) = y - F\beta$. But what does “size” mean?
2. Linear Euclidean Least Squares Fitting

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

\[ |r(\beta)| = \sqrt{r(\beta)^\top r(\beta)} = \sqrt{\sum_{j=1}^{n} r_j(\beta_1, \ldots, \beta_m)^2}. \]

Minimizing \(|r(\beta)|\) is equivalent to minimizing \(|r(\beta)|^2\), which is the sum of the “squares” of the residuals. For linear models \(|r(\beta)|^2\) is a quadratic function of \(\beta\) that is easy to minimize, which is why the method is popular. Specifically, because \(r(\beta) = y - F\beta\), we minimize

\[ q(\beta) = \frac{1}{2}|r(\beta)|^2 = \frac{1}{2}r(\beta)^\top r(\beta) = \frac{1}{2}(y - F\beta)^\top (y - F\beta) = \frac{1}{2}y^\top y - \beta^\top F^\top y + \frac{1}{2}\beta^\top F^\top F\beta. \]
Let us recall some multivariable calculus. The gradient (if it exists) of a real-valued function $q(\beta)$ with respect to the $m$-vector $\beta$ is the $m$-vector $\partial_{\beta} q(\beta)$ such that

$$\left. \frac{d}{ds} q(\beta + s\gamma) \right|_{s=0} = \gamma^T \partial_{\beta} q(\beta)$$

for every $\gamma \in \mathbb{R}^m$.

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

$$q(\beta + s\gamma) = q(\beta) + s\gamma^T (F^T F \beta - F^T y) + \frac{1}{2} s^2 \gamma^T F^T F \gamma.$$  

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

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

from which we read off that

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

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

The symmetric matrix-valued function $\partial_{\beta \beta} q(\beta)$ is sometimes called the Hessian of $q(\beta)$.

For the quadratic $q(\beta)$ arising from our least squares problem you can easily check that

$$\frac{d}{ds} \partial_\beta q(\beta + s\gamma) \bigg|_{s=0} = \frac{d}{ds} (\partial_\beta q(\beta) + s F^T F \gamma) \bigg|_{s=0} = F^T F \gamma,$$

from which we read off that

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

Because $F$ has rank $m$, the $m \times m$-matrix $F^T F$ is positive definite.
Because ∂ββq(β) is positive definite, the function q(β) is strictly convex, whereby it has at most one global minimizer. We find this minimizer by setting the gradient of q(β) equal to zero, yielding

∂β q(β) = F^T F β − F^T y = 0.

Because the matrix F^T F is positive definite, it is invertible. The solution of the above equation is therefore β = ˆβ where

ˆβ = (F^T F)^{-1} F^T y.

The fact that ˆβ is a global minimizer can be seen from the fact F^T F is positive definite and the identity

q(β) = \frac{1}{2} y^T y − \frac{1}{2} ˆβ^T F^T F ˆβ + \frac{1}{2} (β − ˆβ)^T F^T F (β − ˆβ)

= q(ˆβ) + \frac{1}{2} (β − ˆβ)^T F^T F (β − ˆβ).

In particular, this shows that q(β) ≥ q(ˆβ) for every β ∈ R^m and that q(β) = q(ˆβ) if and only if β = ˆβ.
Remark. The least squares fit has a beautiful geometric interpretation with respect to the associated Euclidean inner product

\[ (p \mid q) = p^T q. \]

Define \( \hat{r} = y - F\hat{\beta} \). Observe that

\[ y = F\hat{\beta} + \hat{r} = F(F^TF)^{-1}F^Ty + \hat{r}. \]

The matrix \( P = F(F^TF)^{-1}F^T \) has the properties

\[ P^2 = P, \quad P^T = P. \]

This means that \( Py \) is the orthogonal projection of \( y \) associated with the Euclidean inner product onto the subspace of \( \mathbb{R}^n \) spanned by the columns of \( F \). Then the residual \( \hat{r} \) will be orthogonal to \( Py \) with respect to this inner product. This means that the residual will have mean zero for any data if and only if the constant function 1 is in the span of the basis functions for the model.
Example. If we want to find the Euclidean least squares fit of the affine model \( f(t; \alpha, \beta) = \alpha + \beta t \) to the data \( \{(t_j, y_j)\}_{j=1}^{n} \) then the matrix \( F \) has the form

\[
F = \begin{pmatrix} e & t \end{pmatrix}, \quad \text{where} \quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad t = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}.
\]

If we define

\[
\bar{t} = \frac{1}{n} \sum_{j=1}^{n} t_j, \quad \bar{t}^2 = \frac{1}{n} \sum_{j=1}^{n} t_j^2, \quad \sigma_t^2 = \frac{1}{n} \sum_{j=1}^{n} (t_j - \bar{t})^2,
\]

then we see that

\[
F^T F = n \begin{pmatrix} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{pmatrix},
\]

\[
\det(F^T F) = n^2 (\bar{t}^2 - \bar{t}^2) = n^2 \sigma_t^2 > 0.
\]

Notice that \( \bar{t} \) and \( \sigma_t^2 \) are the sample mean and variance of \( t \) respectively.
Then the $\hat{\alpha}$ and $\hat{\beta}$ that give the least squares fit are given by

$$(\hat{\alpha} \hat{\beta}) = \hat{\beta} = (F^T F)^{-1} F^T y = \frac{1}{n} \left( \frac{t^2 - \bar{t}}{\sigma_t^2} \right) \left( \bar{y} \frac{t}{t^2 - \bar{t}} \right)$$

$$= \frac{1}{\sigma_t^2} \left( \frac{t^2 - \bar{t}}{t^2 - \bar{t}} \right) \left( \frac{\bar{y}}{\bar{y} t} \right) = \frac{1}{\sigma_t^2} \left( \frac{t^2 \bar{y} - \bar{t} t y}{t y - \bar{t} \bar{y}} \right),$$

where

$$\bar{y} = \frac{1}{n} e^T y = \frac{1}{n} \sum_{j=1}^{n} y_j, \quad \bar{t} = \frac{1}{n} t^T y = \frac{1}{n} \sum_{j=1}^{n} y_j t_j.$$

These formulas for $\hat{\alpha}$ and $\hat{\beta}$ can be expressed simply as

$$\hat{\beta} = \frac{\bar{y} t - \bar{y} \bar{t}}{\sigma_t^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{t}.$$

Notice that $\hat{\beta}$ is the ratio of the covariance of $y$ and $t$ to the variance of $t$. 
The best fit is therefore
\[
\hat{f}(t) = \hat{\alpha} + \hat{\beta} t = \bar{y} + \hat{\beta} (t - \bar{t}) = \bar{y} + \frac{y t - \bar{y} \bar{t}}{\sigma_t^2} (t - \bar{t}).
\]

**Remark.** In the above example we inverted the matrix \( F^T F \) to obtain \( \hat{\beta} \). This was easy because our model had only two parameters in it, so \( F^T F \) was only \( 2 \times 2 \). The number of parameters \( m \) does not have to be too large before this approach becomes slow or unfeasible. However for fairly large \( m \) you can obtain \( \hat{\beta} \) by using Gaussian elimination or some other direct method to efficiently solve the linear system
\[
F^T F \beta = F^T y.
\]
Such methods work because the matrix \( F^T F \) is positive definite. As we will soon see, this step can be simplified by constructing the basis \( \{ f_i(t) \}_{i=1}^m \) so that \( F^T F \) is diagonal.
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 $m << n$. 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?