Fitting Linear Statistical Models to Data by Least Squares II: Weighted

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

Math 420: *Mathematical Modeling*
January 25, 2012 version

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

3. Linear Weighted Least Squares Fitting

The Euclidean norm treats every entry of $\mathbf{r}$ the same way. There are many times when that is a natural thing to do. But there are also times when it is natural to do other things. For example, if the times t_j are not uniformly distributed over the time interval under consideration then you might want to give each t_j a positive weight w_j proportional to the length of a subinterval it represents. In other words you can choose to minimize

$$q(\boldsymbol{\beta}) = \tfrac{1}{2} \sum_{j=1}^{n} w_j \, r_j(\beta_1, \cdots, \beta_m)^2 \, .$$

If we let $\mathbf{W}$ be the diagonal matrix whose j^{th} diagonal entry is w_j then this can be expressed as

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

Because $\mathbf{F}$ has rank m the $m \times m$-matrix $\mathbf{F}^\mathsf{T}\mathbf{W}\mathbf{F}$ is positive definite. The function $q(\boldsymbol{\beta})$ is thereby strictly convex, whereby it has a unique minimizer. We find this minimizer by setting the gradient of $q(\boldsymbol{\beta})$ equal to zero, yielding

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

Because $\mathbf{F}^\mathsf{T}\mathbf{W}\mathbf{F}$ is positive definite, it is invertible. The solution of the above equation is $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ where

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

The fact that $\widehat{\boldsymbol{\beta}}$ in a global minimizer can be seen from the fact $\mathbf{F}^\mathsf{T}\mathbf{W}\mathbf{F}$ is positive definite and the identity

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

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 $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.

The weighted least squares fit has a *geometric interpretation* with respect to the inner product associated with the weight matrix $\mathbf{W}$

$$(\mathbf{p} \,|\, \mathbf{q})_{\mathbf{W}} = \mathbf{p}^{\mathsf{T}} \mathbf{W} \mathbf{q} \,.$$

Define $\widehat{\mathbf{r}} = \mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}$. Observe that

$$\mathbf{y} = \mathbf{F}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{r}} = \mathbf{F}(\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{y} + \widehat{\mathbf{r}} \,.$$

The matrix $\mathbf{P} = \mathbf{F}(\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^{\mathsf{T}}\mathbf{W}$ has the properties

$$\mathbf{P}^2 = \mathbf{P} \,, \qquad \mathbf{P}^{\mathsf{T}}\mathbf{W} = \mathbf{W}\mathbf{P} \,.$$

This means that $\widehat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ associated with the $\mathbf{W}$-inner product $(\cdot \,|\, \cdot)_{\mathbf{W}}$ onto the subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{F}$. It follows that $\mathbf{y} = \widehat{\mathbf{y}} + \widehat{\mathbf{r}}$ is an orthogonal decomposition with respect to the $\mathbf{W}$-inner product, which yields the orthogonality relations

$$\widehat{\mathbf{y}}^{\mathsf{T}}\mathbf{W}\widehat{\mathbf{r}} = 0 \,, \qquad \mathbf{y}^{\mathsf{T}}\mathbf{W}\mathbf{y} = \widehat{\mathbf{y}}^{\mathsf{T}}\mathbf{W}\widehat{\mathbf{y}} + \widehat{\mathbf{r}}^{\mathsf{T}}\mathbf{W}\widehat{\mathbf{r}} \,.$$

The weighted least squares fit also has a *statistical interpretation* that is related to these orthogonality relations. If we normalize the weights so that

$$\sum_{j=1}^{n} w_j = 1\,,$$

then the weighted average of any sample $\{z_j\}_{j=1}^n$ is defined by

$$\langle z \rangle = \sum_{j=1}^{n} z_j w_j\,.$$

This weighted average is related to the $\mathbf{W}$-inner product by

$$\langle y\, z \rangle = \sum_{j=1}^{n} y_j z_j w_j = \mathbf{y}^{\mathsf{T}} \mathbf{W} \mathbf{z} = (\mathbf{y} \,|\, \mathbf{z})_{\mathbf{W}}\,.$$

The orthogonality relations can therefore be recast as

$$\langle \widehat{y}\, \widehat{r} \rangle = 0\,, \qquad \left\langle y^2 \right\rangle = \left\langle \widehat{y}^2 \right\rangle + \left\langle \widehat{r}^2 \right\rangle\,.$$

If the constant function 1 is in the span of the basis functions for the model then $\widehat{\mathbf{r}}$ will be orthogonal to the vector that has every entry equal to 1. It follows that

$$\langle \widehat{r} \rangle = 0\,, \qquad \langle \widehat{y} \rangle = \langle y \rangle = \bar{y}\,.$$

These formulas have the statistical interpretations that $\widehat{r}$ has mean zero while $\widehat{y}$ and y have the same mean. In that case the orthogonality relations are equivalent to

$$\langle (\widehat{y} - \bar{y}\,)\,\widehat{r}\,\rangle = 0\,, \qquad \left\langle (y - \bar{y}\,)^2 \right\rangle = \left\langle (\widehat{y} - \bar{y}\,)^2 \right\rangle + \left\langle \widehat{r}^{\,2} \right\rangle\,.$$

These formulas have the statistical interpretations that

$$\mathsf{Cov_s}(\widehat{y}, \widehat{r}) = 0\,, \qquad \mathsf{Var_s}(y) = \mathsf{Var_s}(\widehat{y}) + \mathsf{Var_s}(\widehat{r})\,,$$

where $\mathsf{Cov_s}$ and $\mathsf{Var_s}$ denote sample covariance and sample variance respectively. In particular, $\widehat{y}$ and $\widehat{r}$ are uncorrelated.

This statistical interpretation of the weighted least squares fit leads to a measure for the quality of the fit that is among the most commonly used. Specifically, the *coefficient of determination* R^2 is defined by

$$R^2 = \frac{\mathsf{Var}_{\mathsf{S}}(\widehat{y})}{\mathsf{Var}_{\mathsf{S}}(y)} = \frac{\left\langle (\widehat{y} - \bar{y}\,)^2 \right\rangle}{\left\langle (y - \bar{y}\,)^2 \right\rangle} = 1 - \frac{\left\langle \widehat{r}^{\,2} \right\rangle}{\left\langle (y - \bar{y}\,)^2 \right\rangle} = 1 - \frac{\mathsf{Var}_{\mathsf{S}}(\widehat{r})}{\mathsf{Var}_{\mathsf{S}}(y)}\,.$$

Because $\mathsf{Var}_{\mathsf{S}}(y) = \mathsf{Var}_{\mathsf{S}}(\widehat{y}) + \mathsf{Var}_{\mathsf{S}}(\widehat{r})$, we see that R^2 is simply the fraction of $\mathsf{Var}_{\mathsf{S}}(y)$ that is captured by the fit. In particular, we see that

$$0 \leq R^2 \leq 1\,.$$

Fits are considered to be better by this measure when R^2 is closer to 1. While R^2 can be a reasonable measure of the quality of a fit when being used to compare how well the same model fits different data, it is not good when being used to compare how well different models fit the same data. It is commonly used simply because it is easy to use.

4. Least Square Fitting for Univariate Polynomial Models

The family of all polynomials with degree less than m can be 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 is used for polynomial models because it matches the degree of each term. We will fit this linear model to the given data $\{(t_j, y_j)\}_{j=1}^n$ using weighted least squares with the weights so that

$$\sum_{j=1}^{n} w_j = 1 \, .$$

Recall that the weighted average of any quantities $\{z_j\}_{j=1}^n$ is then

$$\langle z \rangle = \sum_{j=1}^{n} z_j w_j \, .$$

Rather than use the monomials $\{t^i\}_{i=0}^{m-1}$ as the basis for this model, we use the following algorithm to construct a new basis $\{p_i(t)\}_{i=0}^{m-1}$ such that each $p_i(t)$ is a monic polynomial of degree i. We initialize

$$p_0(t) = 1\,, \qquad \sigma_0^2 = \left\langle p_0(t)^2 \right\rangle = \langle 1 \rangle = 1\,,$$
$$p_1(t) = t - \bar{t}\,, \qquad \sigma_1^2 = \left\langle p_1(t)^2 \right\rangle = \left\langle (t-\bar{t})^2 \right\rangle = \sigma_t^2\,.$$

Then given $p_{i-2}(t)$, $p_{i-1}(t)$, σ_{i-2}^2, and σ_{i-1}^2 for some $i \geq 2$ we compute

$$p_i(t) = \left(t - \frac{\left\langle t\, p_{i-1}(t)^2 \right\rangle}{\sigma_{i-1}^2} \right) p_{i-1}(t) - \frac{\sigma_{i-1}^2}{\sigma_{i-2}^2}\, p_{i-2}(t)\,, \quad \sigma_i^2 = \left\langle p_i(t)^2 \right\rangle.$$

We stop when $i = m-1$ and set

$$\widehat{f}(t) = \sum_{i=0}^{m-1} \widehat{\beta}_i p_i(t)\,, \qquad \text{where} \quad \widehat{\beta}_i = \frac{1}{\sigma_i^2} \langle p_i(t) y \rangle\,.$$

The polynomials $p_i(t)$ satisfy the orthogonality relations

$$\langle p_i(t)\, p_{i'}(t)\rangle = \delta_{ii'}\, {\sigma_i}^2 \qquad \text{for every } i,\, i' = 0,\, \cdots,\, m-1\,,$$

where $\delta_{ii'}$ is the Kronecker delta. Then the $m \times m$ matrix $\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{F}$ is diagonal with diagonal entries ${\sigma_i}^2$ while the m-vector $\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{y}$ has entries $\langle p_i(t) y\rangle$. The equation $\mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{F}\boldsymbol{\beta} = \mathbf{F}^{\mathsf{T}}\mathbf{W}\mathbf{y}$ thereby becomes simply

$$ {\sigma_i}^2 \beta_i = \langle p_i(t)\, y\rangle\,,$$

which yields the expression for $\widehat{\beta}_i$ given on the previous slide.

If we set $\widehat{y}_j = \widehat{f}(t_j)$ for every $j = 1,\, \cdots,\, n$ then another consequence of these polynomial orthogonality relations is the fact that

$$\left\langle (y - \bar{y}\,)^2 \right\rangle = \left\langle (\widehat{y} - \bar{y}\,)^2 \right\rangle + \left\langle \widehat{r}^{\,2} \right\rangle = \sum_{i=1}^{m-1} \frac{\langle p_i(t)\, y\rangle^2}{{\sigma_i}^2} + \left\langle \widehat{r}^{\,2} \right\rangle\,.$$

This shows exactly how much $\left\langle \widehat{r}^{\,2} \right\rangle$ will be reduced as m is increased.

Example. If we want to find the least squares fit of the data to a polynomial of degree less than 3 then our algorithm yields

$$p_0(t) = 1\,, \qquad p_1(t) = t - \bar{t}\,,$$
$$p_2(t) = (t-\bar{t}\,)^2 - \tau(t-\bar{t}\,) - \sigma^2\,,$$

where

$$\sigma^2 = \Big\langle (t-\bar{t}\,)^2 \Big\rangle\,, \qquad \tau = \frac{1}{\sigma^2}\Big\langle (t-\bar{t}\,)^3 \Big\rangle\,.$$

Moreover, $\sigma_0^2 = 1$, $\sigma_1^2 = \sigma^2$, and

$$\begin{aligned}\sigma_2^2 = \Big\langle p_2(t)^2 \Big\rangle &= \Big\langle (t-\bar{t}\,)^2 p_2(t) \Big\rangle \\ &= \Big\langle (t-\bar{t}\,)^4 \Big\rangle - \tau\Big\langle (t-\bar{t}\,)^3 \Big\rangle - \sigma^2\Big\langle (t-\bar{t}\,)^2 \Big\rangle \\ &= \Big\langle (t-\bar{t}\,)^4 \Big\rangle - \sigma^2\tau^2 - \sigma^4\,,\end{aligned}$$

while $\langle p_0(t)\, y\rangle = \langle y\rangle = \bar{y}$, $\langle p_1(t)\, y\rangle = \langle (t-\bar{t}\,)\, y\rangle$, and

$$\langle p_2(t)\, y\rangle = \Big\langle (t-\bar{t}\,)^2 y \Big\rangle - \tau\langle (t-\bar{t}\,)\, y\rangle - \sigma^2\bar{y}\,.$$

Therefore the fit is

$$\begin{aligned}\widehat{f}(t) &= \frac{\langle p_0(t)\, y\rangle}{\sigma_0^2}\, p_0(t) + \frac{\langle p_1(t)\, y\rangle}{\sigma_1^2}\, p_1(t) + \frac{\langle p_2(t)\, y\rangle}{\sigma_2^2}\, p_2(t) \\ &= \bar{y} + \frac{\langle (t-\bar{t})\, y\rangle}{\sigma^2}\,(t-\bar{t}) \\ &\quad + \frac{\left\langle (t-\bar{t})^2 y\right\rangle - \tau\langle (t-\bar{t})\, y\rangle - \sigma^2\bar{y}}{\sigma_2^2}\left((t-\bar{t})^2 - \tau(t-\bar{t}) - \sigma^2\right).\end{aligned}$$

Remark. Notice that we never had to explicitly solve a linear algebraic system in our solution of the above example. This should be contrast with our solution (given in an earlier example) of the simpler problem of fitting to the model $f(t;\alpha,\beta) = \alpha + \beta t$. In fact, you should notice that a solution of that earlier problem is contained within the solution of the above problem. This contrast shows there is some value in constructing an orthogonal basis for your model. We extend this idea in the next section.

5. Least Squares Fitting with Orthogonalization

We can generalize what we did for polynomial models to any linear model. Let $\{f_i(\mathbf{x})\}_{i=1}^m$ be a basis for some linear model. We can then use a variant of the Gramm-Schmidt algorithm to construct a new basis $\{g_i(\mathbf{x})\}_{i=1}^m$ that is orthogonal with respect to the inner product

$$(g \,|\, h) = \langle g(\mathbf{x})\, h(\mathbf{x}) \rangle \,.$$

The fact that $\mathbf{F}$ has rank m implies that $(\cdot \,|\, \cdot)$ is an inner product over the range of the model. We set $g_1(\mathbf{x}) = f_1(\mathbf{x})$ and for $i \geq 2$ compute

$$g_i(\mathbf{x}) = f_i(\mathbf{x}) - \sum_{i'=1}^{i-1} \frac{\langle f_i(\mathbf{x})\, g_{i'}(\mathbf{x}) \rangle}{\langle g_{i'}(\mathbf{x})^2 \rangle}\, g_{i'}(\mathbf{x}) \,.$$

We stop when $i = m$ and set

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{m} \widehat{\beta}_i g_i(\mathbf{x}) \,, \qquad \text{where} \quad \widehat{\beta}_i = \frac{\langle g_i(\mathbf{x})\, y \rangle}{\langle g_i(\mathbf{x})^2 \rangle} \,.$$

Remark. This algorithm for generating the basis $\{g_i(\mathbf{x})\}_{i=1}^m$ seems more complicated than the algorithm we used to generate the basis $\{p_i(t)\}_{i=0}^{m-1}$ for univariate polynomial models. This is because the structure of those polynomial models simplifies the more general algorithm.

If we set $\widehat{y}_j = \widehat{f}(\mathbf{x}_j)$ for every $j = 1, \cdots, n$ then the orthogonality relations satisfied by $\{g_i(\mathbf{x})\}_{i=1}^m$ imply

$$\left\langle (y - \bar{y}\,)^2 \right\rangle = \left\langle (\widehat{y} - \bar{y}\,)^2 \right\rangle + \left\langle \widehat{r}^2 \right\rangle = \sum_{i=1}^{m} \frac{\langle g_i(\mathbf{x})\,(y - \bar{y}\,)\rangle^2}{\langle g_i(\mathbf{x})^2 \rangle} + \left\langle \widehat{r}^2 \right\rangle .$$

This shows exactly how much $\left\langle \widehat{r}^2 \right\rangle$ will be reduced as m is increased.

Remark. Reducing $\left\langle \widehat{r}^2 \right\rangle$ does not always make the fit better. Indeed, sometimes the fit can get worse. This is the phenomenon of *overfitting*.

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?