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

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

Weighted Least Squares Fitting for Linear Models

Weighted Least Squares Fitting: Introduction

We will work in the *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 $\mathbb{R}$.
We will consider linear statistical models generated by a basis $\{f_i(\mathbf{x})\}_{i=1}^{m}$ where each $f_i(\mathbf{x})$ is defined over $\mathbb{X}$ and takes values in $\mathbb{R}$.
These models have the form

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

where $\beta_1, \cdots, \beta_m$ are real parameters.

Weighted Least Squares Fitting: Fitting Problem

Recall the *fitting problem* for such models cast in terms of vectors. Define the m-vector $\boldsymbol{\beta}$, the n-vector $\mathbf{y}$, and the $n \times m$-matrix $\mathbf{F}$ by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} f_1(\mathbf{x}_1) & \cdots & f_m(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ f_1(\mathbf{x}_n) & \cdots & f_m(\mathbf{x}_n) \end{pmatrix}.$$

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

$$\mathbf{r}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{F}\boldsymbol{\beta} = \begin{pmatrix} y_1 - f(\mathbf{x}_1; \beta_1, \cdots, \beta_m) \\ \vdots \\ y_n - f(\mathbf{x}_1; \beta_1, \cdots, \beta_m) \end{pmatrix} = \begin{pmatrix} r_1(\boldsymbol{\beta}) \\ \vdots \\ r_1(\boldsymbol{\beta}) \end{pmatrix}.$$

But what does "size" mean?

Weighted Least Squares Fitting: Weights

Euclidean least squares fitting measured the size of $\mathbf{r}$ by its Euclidean norm. The Euclidean norm treats every entry of $\mathbf{r}$ the same way. This is often a natural thing to do. But there are also times when it is natural to do other things. For example, if the points $\mathbf{x}_j$ are not uniformly distributed over the domain $\mathbb{X}$ then we might want to give each $\mathbf{x}_j$ a positive *weight* w_j proportional to the volume of the subset of $\mathbb{X}$ that it represents. In that case we can choose to minimize

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

If $\mathbf{W}$ is the diagonal matrix whose j^{th} diagonal entry is weight w_j then

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

Weighted Least Squares Fitting: Minimizer

Because $q(\boldsymbol{\beta})$ is a quadratic function of $\boldsymbol{\beta}$, we can use multivariable calculus to minimize it just as was done for the Euclidean case. The gradient and Hessian of $q(\boldsymbol{\beta})$ are found to be

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

Because $\mathbf{F}$ has rank m, the matrix $\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F}$ is positive definite. The function $q(\boldsymbol{\beta})$ is thereby strictly convex, whereby its minimizer is unique. It is found by setting the gradient of $q(\boldsymbol{\beta})$ equal to zero, yielding

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

Because $\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F}$ 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

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

Weighted Least Squares Fitting: Uniqueness

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

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

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

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

Weighted Least Squares Fitting: Geometric View

The weighted least squares fit has a *geometric interpretation* in $\mathbb{R}^n$ equipped with the scalar product associated with the weight matrix $\mathbf{W}$

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

Recall that the range of $\mathbf{F}$ is given by

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

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

$$\begin{aligned}\left(\mathbf{F}\boldsymbol{\gamma} \mid \widehat{\mathbf{r}}\right)_{\mathbf{W}} = (\mathbf{F}\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{W}\widehat{\mathbf{r}} = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{W}\widehat{\mathbf{r}} &= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{W}\left(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}\right)\\ &= \boldsymbol{\gamma}^{\mathrm{T}}\left(\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{y} - \mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F}\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{0} = 0\,.\end{aligned}$$

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

Weighted Least Squares Fitting: Geometric View

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

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

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

$$\begin{aligned}\|\mathbf{y}\|_{\mathbf{W}}^2 = (\mathbf{y}\,|\,\mathbf{y})_{\mathbf{W}} &= (\widehat{\mathbf{y}} + \widehat{\mathbf{r}}\,|\,\widehat{\mathbf{y}} + \widehat{\mathbf{r}})_{\mathbf{W}} \\ &= (\widehat{\mathbf{y}}\,|\,\widehat{\mathbf{y}})_{\mathbf{W}} + (\widehat{\mathbf{r}}\,|\,\widehat{\mathbf{r}})_{\mathbf{W}} = \|\widehat{\mathbf{y}}\|_{\mathbf{W}}^2 + \|\widehat{\mathbf{r}}\|_{\mathbf{W}}^2\,,\end{aligned}$$

where $\|\cdot\|_{\mathbf{W}}$ is the norm associated with the scalar product $(\cdot\,|\,\cdot)_{\mathbf{W}}$.

Weighted Least Squares Fitting: Geometric View

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

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

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

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

These properties can be used to show that:

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

Weighted Least Squares Fitting: Statistical View

The weighted least squares fit also has a *statistical interpretation* that is related to these geometric 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 **W**-scalar product by

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

The orthogonality and Pythagoean relations can then be recast as

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

Weighted Least Squares Fitting: Statistical View

If the constant function 1 is in the span of the basis 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 and Pythagorean relations are equivalent to

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

These formulas have the statistical interpretations that

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

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

Weighted Least Squares Fitting: Statistical View

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

$$R^2 = \frac{\mathrm{Var}_s(\hat{y})}{\mathrm{Var}_s(y)} = \frac{\langle(\hat{y} - \bar{y})^2\rangle}{\langle(y - \bar{y})^2\rangle} = 1 - \frac{\langle\hat{r}^2\rangle}{\langle(y - \bar{y})^2\rangle} = 1 - \frac{\mathrm{Var}_s(\hat{r})}{\mathrm{Var}_s(y)} .$$

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

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

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

Fitting for Univariate Polynomial Models: Introduction

The family of all polynomials with degree at most ℓ can be written as

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

The index k runs from 0 to ℓ so that it matches the degree of each term. We will fit this linear model to data $\{(t_j, y_j)\}_{j=1}^n$ using weighted least squares with weights $\{w_j\}_{j=1}^n$ normalized so that

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

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

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

Fitting for Univariate Polynomial Models: New Basis

Rather than use the monomials $\{t^k\}_{k=0}^{\ell}$ as the basis for this model, we use the following algorithm to construct a new basis $\{p_k(t)\}_{i=0}^{\ell}$ that is orthogonal with respect to the scalar product

$$(p \mid q) = \langle p(t)\, q(t) \rangle = \sum_{j=1}^{n} p(t_j)\, q(t_j)\, w_j \,,$$

and such that each $p_k(t)$ is a monic polynomial of degree k.
We set $\bar{t} = \langle t \rangle$ and initialize

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

Fitting for Univariate Polynomial Models: New Basis

Given $p_{k-2}(t)$, $p_{k-1}(t)$, σ_{k-2}^2, and σ_{k-1}^2 for some $k \geq 2$ we then set

$$\eta_{k-1} = \left\langle (t - \bar{t}\,) p_{k-1}(t)^2 \right\rangle ,$$

$$p_k(t) = \left(t - \bar{t} - \frac{\eta_{k-1}}{\sigma_{k-1}^2} \right) p_{k-1}(t) - \frac{\sigma_{k-1}^2}{\sigma_{k-2}^2} p_{k-2}(t) ,$$

$$\sigma_k^2 = \left\langle p_k(t)^2 \right\rangle .$$

We stop when $k = \ell$ and set

$$\widehat{f}(t) = \sum_{k=0}^{\ell} \widehat{\beta}_k p_k(t) , \qquad \text{where} \quad \widehat{\beta}_k = \frac{1}{\sigma_k^2} \left\langle p_k(t) y \right\rangle .$$

Fitting for Univariate Polynomial Models: Orthogonality

Remark. The polynomials $p_k(t)$ satisfy the orthogonality relations

$$\langle p_k(t)\, p_{k'}(t)\rangle = \delta_{kk'}\, \sigma_k^2 \qquad \text{for every } k,\, k' = 0, \cdots, \ell\,,$$

where $\delta_{kk'}$ is the Kronecker delta. Then the matrix $\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F}$ is diagonal with diagonal entries σ_k^2 while the vector $\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{y}$ has entries $\langle p_k(t) y\rangle$. The equation $\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F}\boldsymbol{\beta} = \mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{y}$ thereby becomes simply

$$\sigma_k^2 \beta_k = \langle p_k(t)\, y\rangle\,, \qquad \text{for } k = 0, \cdots .\ell\,,$$

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

Fitting for Univariate Polynomial Models: Orthogonality

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

$$\left\langle (y - \bar{y})^2 \right\rangle = \left\langle (\widehat{y} - \bar{y})^2 \right\rangle + \left\langle \widehat{r}^2 \right\rangle = \sum_{k=1}^{\ell} \frac{\left\langle p_k(t)\,(y - \bar{y}) \right\rangle^2}{\sigma_k^2} + \left\langle \widehat{r}^2 \right\rangle .$$

This shows exactly how much $\left\langle \widehat{r}^2 \right\rangle$ will be reduced as ℓ is increased.
Remark. One criterion for when to stop enlarging the model is when the addition of another basis function does not significantly reduce the residual variance $\left\langle \widehat{r}^2 \right\rangle$. For example, we could stop at ℓ if

$$\frac{\left\langle p_{\ell+1}(t)\,(y - \bar{y}) \right\rangle^2}{\sigma_{\ell+1}^2} < \tfrac{1}{4} \left\langle \widehat{r}^2 \right\rangle .$$

Fitting for Univariate Polynomial Models: Example

Example. If we want to find the least squares fit of the data to a polynomial of degree at most 2 then our algorithm yields

$$p_0(t) = 1\,, \qquad \sigma_0^2 = 1\,,$$
$$p_1(t) = t - \bar{t}\,, \qquad \sigma_1^2 = \sigma^2\,,$$
$$\eta_1 = \langle (t-\bar{t})p_1(t)^2 \rangle = \langle (t-\bar{t})^3 \rangle = \tau^3\,,$$
$$p_2(t) = \left(t - \bar{t} - \frac{\eta_1}{\sigma_1^2} \right) p_1(t) - \frac{\sigma_1^2}{\sigma_0^2}\, p_0(t)$$
$$= \left(t - \bar{t} - \frac{\tau^3}{\sigma^2} \right) (t-\bar{t}) - \sigma^2 = (t-\bar{t})^2 - \frac{\tau^3}{\sigma^2}(t-\bar{t}) - \sigma^2\,,$$

where $\bar{t}$, σ, and τ are given by the weighted averages

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

Fitting for Univariate Polynomial Models: Example

Moreover, we have

$$\sigma_2^2 = \langle p_2(t)^2 \rangle = \langle (t - \bar{t})^2 p_2(t) \rangle$$
$$= \langle (t - \bar{t})^4 \rangle - \frac{\tau^3}{\sigma^2} \langle (t - \bar{t})^3 \rangle - \sigma^2 \langle (t - \bar{t})^2 \rangle$$
$$= \langle (t - \bar{t})^4 \rangle - \frac{\tau^6}{\sigma^2} - \sigma^4 .$$

and

$$\langle p_0(t)\, y \rangle = \langle y \rangle = \bar{y} ,$$
$$\langle p_1(t)\, y \rangle = \langle (t - \bar{t})\, y \rangle = \langle (t - \bar{t})\, (y - \bar{y}) \rangle ,$$
$$\langle p_2(t)\, y \rangle = \langle (t - \bar{t})^2 y \rangle - \frac{\tau^3}{\sigma^2} \langle (t - \bar{t})\, y \rangle - \sigma^2 \bar{y}$$
$$= \langle (t - \bar{t})^2 (y - \bar{y}) \rangle - \frac{\tau^3}{\sigma^2} \langle (t - \bar{t})\, (y - \bar{y}) \rangle .$$

Fitting for Univariate Polynomial Models: Example

Therefore the weighted least squares fit is

$$\begin{aligned}\widehat{f}(t) &= \widehat{\beta}_0\, p_0(t) + \widehat{\beta}_1\, p_1(t) + \widehat{\beta}_2\, p_2(t) \\ &= \bar{y} + \widehat{\beta}_1\,(t-\bar{t}\,) + \widehat{\beta}_2\left((t-\bar{t}\,)^2 - \frac{\tau^3}{\sigma^2}(t-\bar{t}\,) - \sigma^2\right),\end{aligned}$$

where

$$\widehat{\beta}_1 = \frac{\langle p_1(t)\, y\rangle}{\sigma_1^2} = \frac{\langle (t-\bar{t}\,)\,(y-\bar{y}\,)\rangle}{\sigma^2},$$

$$\widehat{\beta}_2 = \frac{\langle p_2(t)\, y\rangle}{\sigma_2^2} = \frac{\langle (t-\bar{t}\,)^2 (y-\bar{y}\,)\rangle - \dfrac{\tau^3}{\sigma^2}\,\langle (t-\bar{t}\,)\,(y-\bar{y}\,)\rangle}{\langle (t-\bar{t}\,)^4\rangle - \dfrac{\tau^6}{\sigma^2} - \sigma^4}.$$

Fitting for Univariate Polynomial Models: Example

Remark. As the above example suggests, the $p_k(t)$ grow in complexity as k increases. This construction is seldom carried out for $\ell > 3$ due to both this growing complexity and the fact that polynomials of higher degree are rarely good statistical models for data sets.

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

Fitting with Orthogonalization: Introduction

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 Gram-Schmidt algorithm to construct a new basis $\{g_i(\mathbf{x})\}_{i=1}^m$ that is orthogonal with respect to the scalar product

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

The fact that $\mathbf{F}$ has rank m implies that $(\cdot \mid \cdot)$ is an scalar 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

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

Fitting with Orthogonalization: Orthogonality

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.

Remark. 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

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

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, which is the phenomenon of *overfitting*.

Further Questions

We have seen how to use weighted 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? (Use ones that effectively reduce the residual variance.)
- 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? (Yes, but none are as easy to implement as least squares.)