Polynomial Regression and Model Expansion

University of California, San Diego
Instructor: Armin Schwartzman

Dealing with curvature

- Suppose that some diagnostics reveal some curvature.
- We could transform the variables and fit a simple linear model.
- Instead of transforming the data, we can augment (enrich) the model. In particular, we can start by adding a quadratic term in the variable that showed some curvature.

Polynomial regression

- Suppose we have data $\{(x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, n\}$.
- A polynomial model is of the form:

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p$$

  This is a special case of linear regression, with variables $x_j = x^j, j = 0, \ldots, p$.
- Fitting the model by least squares regression amounts to minimizing:

$$\mathrm{SSE}(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i - \cdots b_p x_i^p)^2$$

  Let $(\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$ denote the solution.
- Under the standard assumptions on the errors, this corresponds to maximum likelihood estimation.
- We estimate σ^2 by

$$\widehat{\sigma}^2 = \frac{\mathrm{SSE}(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)}{n - p - 1}.$$

Polynomial regression

- Choosing (or estimating) the degree is non-trivial.
- The higher the degree, the richer and larger the model is.
  Comparing models sequentially via ANOVA may guide us in choosing a degree.
  (In general, it is preferable to use a model selection procedure. We will cover this in detail later in the course.)

Issues with the canonical polynomial basis

- Numerically, fitting polynomials may be unstable. Indeed, the design matrix is a Vandermonde matrix, known to be ill-conditioned:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^p \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^p \end{pmatrix}$$

- The LS coefficients change with the degree of the polynomial being fitted. This is because the columns of $\mathbf{X}$ are, in general, *not* orthogonal.

Orthogonal polynomials

- The idea is to define another basis for polynomials of degree less than or equal to p, so that the model becomes:

$$y = \alpha_0 + \alpha_1 \rho_1(x) + \cdots + \alpha_p \rho_p(x),$$

  where ρ_k is a polynomial of degree exactly k, and such that

$$\langle \rho_k, \rho_\ell \rangle := \sum_{i=1}^{n} \rho_k(x_i)\rho_\ell(x_i) = \mathbb{I}\{k = \ell\}.$$

  Note that $\langle \cdot, \cdot \rangle$ is a true inner product on the set of polynomials of degree at most p when p is less than or equal to the number of distinct x_i's.

- Orthogonal polynomials may be obtained by applying a Gram-Schmidt orthogonalization to the design matrix $\mathbf{X}$, although other stable techniques have been suggested.

  (In R, the function poly computes orthogonal polynomials by default.)

Piecewise polynomials

- Piecewise polynomials provide an even richer class of models.

- A general piecewise polynomial function is of the form

$$\sum_{k=0}^{K} (\beta_{k,0} + \cdots + \beta_{k,q} x^q) \; \mathbb{I}\{\xi_k < x \leq \xi_{k+1}\}$$

  $\xi_1 < \cdots < \xi_K$ are the knots, with $\xi_0 = -\infty$ and $\xi_{K+1} = \infty$.

- There are some drawbacks:
  - Visually not pleasing (always discontinuous).
  - The model is often large with $(K+1)(q+1)$ parameters.

Splines

- A spline of order $q+1$ (or degree q) is a piecewise polynomial of degree q with continuous derivatives up to order $q-1$.

  This is so if and only if adjacent pieces have the same values at the knots up to the $(q-1)$th derivative. Each such condition constitutes a linear constraint on the coefficients.

- A spline of order $q+1$ with knots $\xi_1, \dots, \xi_K$ is determined by $(K+1)(q+1) - Kq = K + q + 1$ parameters (or degrees of freedom).

Fitting a spline model

- A spline model can be fitted by least squares but subject to these additional (linear) constraints. This leads to the minimization of a quadratic subject to linear constraints.

- The preferred option is to use a basis for the model. Indeed, a spline model is a linear model. The following is a basis for spline of degree q and knots $\xi_1, \dots, \xi_K$:

$$1, x, \dots, x^q, (x-\xi_1)_+^q, \dots, (x-\xi_K)_+^q$$

  where $a_+ = \max(a, 0)$.

  This allows to fit the model by regular least squares.

B-spline basis

- Another basis, called B-spline basis, is usually used for fitting a splines model.

- Each element of the basis (called a B-spline) is localized, which makes for faster, more stable computations — the design matrix is block-diagonal.

- Suppose the knots are $\xi_1, \dots, \xi_K$, and ξ_0 and ξ_{K+1} are the boundary points. Let q be the degree.

  Define

$$\tau_j = \begin{cases} \xi_0 & \text{for } j = 1, \dots, q+1 \\ \xi_{j-q-1} & \text{for } j = q+2, \dots, q+K+1 \\ \xi_{K+1} & \text{for } j = q+K+2, \dots, 2q+K+2 \end{cases}$$

  Then the following are the B-splines

$$b_{j,q+1}(x) = \sum_{k=j}^{j+q+1} \frac{(x-\tau_k)_+^q}{\prod_{\ell=j,\ell\neq k}^{j+q+1}(\tau_k - \tau_\ell)}$$

B-spline basis

- We can define the B-splines recursively — this is modulo an irrelevant multiplicative constant.
- B-splines of order degree $q = 0$ are defined as

$$b_{j,1}(x) = \begin{cases} 1 & \text{if } \tau_j \leq x < \tau_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

- For $q \geq 0$, the B-splines of degree q are then defined as

$$b_{j,q+1}(x) = \frac{x - \tau_j}{\tau_{j+q} - \tau_j}\, b_{j,q}(x) + \frac{\tau_{j+q+1} - x}{\tau_{j+q+1} - \tau_{j+1}}\, b_{j+1,q}(x).$$

Cubic and natural splines

- Cubic splines are splines of degree 3

  They are very popular — part of the reason is that the (human) eye is apparently not sensitive to higher degrees of smoothness.

- Natural splines are cubic splines constrained to be linear before the first knot and after the last knot.

  There are K degrees of freedom in this model.

Smoothing splines

- Consider a model of the form

$$\mathbb{E}(y|x) = f(x)$$

  where we only assume that f is twice differentiable. How to fit this model?

- The following penalized least squares criterion is natural

$$\inf_g \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(x)^2 \mathrm{d}x$$

  where the infimum is over functions g that are twice differentiable.

- The parameter λ drives the degrees of freedom in the fit.

  $\lambda = 0$ corresponds to n degrees of freedom if the x_i's are all distinct.

  $\lambda = \infty$ corresponds to 2 degrees of freedom (simple linear regression).

Smoothing splines

- **Fact.** There is a minimizer of the above functional among natural splines with knots the distinct x_i's.
- Moreover, after choosing a basis for natural splines, the model becomes a weighted variant of ridge regression.
- The tuning parameter λ is chosen according to some criterion; in R, the function smooth.spline uses generalized cross validation (GCV), which provides an estimate for the prediction error.

Polynomials in several variables

- A polynomial model of degree q in p variables is of the following form:

$$\mathbb{E}(y|\mathbf{x}) = \sum_{s=0}^{q} \sum_{k_1+\cdots+k_p=s} \beta_{k_1,\ldots,k_p}\, x_1^{k_1} \cdots x_p^{k_p}$$

  where $\mathbf{x} = (x_1, \ldots, x_p)$.

- This means that, if we observe $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, we assume the model

$$y_i = \sum_{s=0}^{q} \sum_{k_1+\cdots+k_p=s} \beta_{k_1,\ldots,k_p}\, x_{i,1}^{k_1} \cdots x_{i,p}^{k_p} + \varepsilon_i$$

  where the errors satisfy $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0$.

Splines in several variables

- The simplest way to fit splines in several variables is to use the tensor product basis of one-dimensional B-splines.
- Just as for polynomials, the number of parameters becomes quite large very quickly as the dimension increases.
- Multivariate Additive Regression Splines (MARS) is a greedy method that aims at fitting a multivariate spline model.

Other bases for model expansion

Examples:

- Trigonometric polynomials (i.e. Fourier basis).
- Radial basis functions (RBF).
- Wavelets (and related families).

Other bases for model expansion

- In general, suppose we want to fit the model

$$\mathbb{E}(y|\mathbf{x}) = \sum_{j=0}^{p} \beta_j f_j(\mathbf{x}),$$

  where $f_0, \ldots, f_p$ are known functions.

- Then the model coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ can be fitted by least-squares and this corresponds to the MLE when the errors satisfy the standard assumptions.

- In that case, the design matrix based on data $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ is

$$\mathbf{X} = \begin{pmatrix} f_0(\mathbf{x}_1) & f_1(\mathbf{x}_1) & \cdots & f_p(\mathbf{x}_1) \\ f_0(\mathbf{x}_2) & f_1(\mathbf{x}_2) & \cdots & f_p(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ f_0(\mathbf{x}_n) & f_1(\mathbf{x}_n) & \cdots & f_p(\mathbf{x}_n) \end{pmatrix}$$