Lecture 25 — The linear model

25.1 The linear model

Example 25.1. When a string instrument sustains a note at a particular pitch, the resulting sound wave is periodic with some fixed frequency f (say 440Hz). For a "pure" tone at this pitch, the sound wave is a perfect sinusoidal wave with frequency f, but the sound produced by any real string instrument is not a pure tone. Instead, it is a superposition of sinusoidal waves with frequencies f, $2f$, $3f$, $4f$, etc., corresponding to different vibrating modes of the string. These frequencies are called the resonance harmonics, and the relative volumes, or amplitudes, of the resonance harmonics determine the timbre or "color" of the sound.

A recording device measures the sound wave produced by an instrument (sustaining a single note) at n points in time $t_1, \ldots, t_n$, where the measurements are contaminated by white noise. We consider the problem of estimating the amplitudes of the resonance harmonics for this instrument. Let $Y_1, \ldots, Y_n \in \mathbb{R}$ be measurements of the sound wave at these time points. Suppose, for simplicity, we have scaled our units so that the sine and cosine curves corresponding to the fundamental frequency f are $\sin(t)$ and $\cos(t)$. Then, assuming the existence of resonance harmonics up to frequency kf, we may model each measurement Y_i as

$$Y_i = \beta_1 \sin(t_i) + \beta_2 \cos(t_i) + \beta_3 \sin(2t_i) + \beta_4 \sin(2t_i) + \ldots + \beta_{2k-1} \sin(kt_i) + \beta_{2k} \cos(kt_i) + \varepsilon_i, \tag{25.1}$$

for some coefficients $\beta_1, \ldots, \beta_{2k} \in \mathbb{R}$, where the errors $\varepsilon_i \overset{IID}{\sim} \mathcal{N}(0, \sigma_0^2)$ correspond to the white noise and the variance σ_0^2 signifies the noise level. If we construct the matrix

$$X = \begin{pmatrix} \sin(t_1) & \cos(t_1) & \cdots & \sin(kt_1) & \cos(kt_1) \\ \sin(t_2) & \cos(t_2) & \cdots & \sin(kt_2) & \cos(kt_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sin(t_n) & \cos(t_n) & \cdots & \sin(kt_n) & \cos(kt_n) \end{pmatrix}$$

and denote its entries as x_{ij}, then we may write the above as

$$Y_i = \sum_{j=1}^{2k} \beta_j x_{ij} + \varepsilon_i,$$

or more succinctly in matrix notation, for all $i = 1, \ldots, n$, as

$$Y = X\beta + \varepsilon.$$

Here, Y denotes the column vector $(Y_1, \ldots, Y_n)$, β denotes the column vector $(\beta_1, \ldots, \beta_{2k})$, and ε denotes the column vector $(\varepsilon_1, \ldots, \varepsilon_n)$. This is called a **linear model**.

More generally, given a vector of **responses** $Y_1, \ldots, Y_n$, the linear model models each Y_i as a certain linear combination $\beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ of corresponding **covariates** $x_{i1}, \ldots, x_{ip}$, plus IID Gaussian errors. (The coefficients $\beta_1, \ldots, \beta_p$ are the same for all n responses $Y_1, \ldots, Y_n$.) We will treat the covariates as fixed and known constants. The values of the Gaussian errors are not directly observed; for simplicity, however, we'll assume in this lecture that their variance σ_0^2 is known. The parameters of the model are the regression coefficients $\beta_1, \ldots, \beta_p$. (Much of our analysis in this lecture may be extended to the more realistic setting where σ_0^2 is unknown, in which case it would also be a parameter of the model.)

25.2 Statistical inference

In the model of equation (25.1), the amplitudes of the k resonance harmonics are defined as $A_1 = \sqrt{\beta_1^2 + \beta_2^2}$, $A_2 = \sqrt{\beta_3^2 + \beta_4^2}$, ..., $A_k = \sqrt{\beta_{2k-1}^2 + \beta_{2k}^2}$. We will discuss the following inferential tasks:

- Estimate the amplitudes $A_1, \ldots, A_k$
- Provide confidence intervals corresponding to these estimates

Let $p = 2k$. To write down the likelihood for the linear model, note that $Y_1, \ldots, Y_n$ are independent and distributed as $Y_i \sim \mathcal{N}(\sum_j \beta_j x_{ij}, \sigma_0^2)$. Then

$$\text{lik}(\beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2} \left(Y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right),$$

and the log-likelihood is

$$l(\beta_1, \ldots, \beta_p) = -\frac{n}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2. \tag{25.2}$$

For any $\sigma_0^2 > 0$, this log-likelihood is maximized when $\beta_1, \ldots, \beta_k$ are the **least-squares** estimators minimizing the total squared error

$$\text{err} = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.$$

So the MLEs $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are equal to the least-squares estimators. To compute these MLEs, we solve the system of p equations for $m = 1, \ldots, p$

$$0 = \frac{\partial l}{\partial \beta_m} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} x_{im} \left(Y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right).$$

Letting X_m denote the mth column of X, these equations may be written as

$$0 = \frac{1}{\sigma_0^2} X_m^T (Y - X\beta),$$

or even more succinctly for all $m = 1, \ldots, p$ as $0 = \frac{1}{\sigma_0^2} X^T (Y - X\beta)$, where both sides of this equation are vectors of length p. Solving for β yields the MLEs/least-squares estimates

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{25.3}$$

To estimate an amplitude of a resonance harmonic, say $A_1 = \sqrt{\beta_1^2 + \beta_2^2}$, we may use the plugin estimate $\hat{A}_1 = \sqrt{\hat{\beta}_1^2 + \hat{\beta}_2^2}$.

To obtain a confidence interval for A_1, we will derive an approximate standard error for $\hat{A}_1$, by first deriving the sampling distribution of $\hat{\beta}$ and then applying the delta method. We compute the Fisher information $I_{\mathbf{Y}}(\beta) = -\mathbb{E}_\beta[\nabla^2 l(\beta)]$, by computing the second-order derivatives of l:

$$\frac{\partial^2 l}{\partial \beta_m \partial \beta_l} = -\frac{1}{\sigma_0^2} \sum_{i=1}^n x_{im} x_{il} = -\frac{1}{\sigma_0^2} X_m^T X_l$$

Then the Hessian matrix is $\nabla^2 l(\beta) = -\frac{1}{\sigma_0^2} X^T X$, so $I_{\mathbf{Y}}(\beta) = \frac{1}{\sigma_0^2} X^T X$. The distribution of $\hat{\beta}$ is then approximately $\mathcal{N}(\beta, \sigma_0^2 (X^T X)^{-1})$ for large n.[1]

We now apply the delta method: Defining $g(x, y) = \sqrt{x^2 + y^2}$, a Taylor expansion yields

$$\begin{aligned}
\hat{A}_1 - A_1 = g(\hat{\beta}_1, \hat{\beta}_2) - g(\beta_1, \beta_2) &\approx \frac{\partial g}{\partial x}(\beta_1, \beta_2) \times (\hat{\beta}_1 - \beta_1) + \frac{\partial g}{\partial y}(\beta_1, \beta_2) \times (\hat{\beta}_2 - \beta_2) \\
&= \frac{\beta_1}{\sqrt{\beta_1^2 + \beta_2^2}} (\hat{\beta}_1 - \beta_1) + \frac{\beta_2}{\sqrt{\beta_1^2 + \beta_2^2}} (\hat{\beta}_2 - \beta_2).
\end{aligned}$$

Letting $c_1 = \beta_1 / \sqrt{\beta_1^2 + \beta_2^2}$, $c_2 = \beta_2 / \sqrt{\beta_1^2 + \beta_2^2}$, $Z_1 = \hat{\beta}_1 - \beta_1$, and $Z_2 = \hat{\beta}_2 - \beta_2$, the above sampling distribution for $\hat{\beta}$ implies that (Z_1, Z_2) is approximately bivariate normal with mean 0 and covariance given by the upper-left 2×2 block of $\sigma_0^2 (X^T X)^{-1}$. So $\hat{A}_1 - A_1 \approx c_1 Z_1 + c_2 Z_2$ is approximately normal with mean 0 and variance

$$\begin{aligned}
\operatorname{Var}[c_1 Z_1 + c_2 Z_2] &= \operatorname{Cov}[c_1 Z_1 + c_2 Z_2, c_1 Z_1 + c_2 Z_2] \\
&= c_1^2 \operatorname{Var}[Z_1] + c_2^2 \operatorname{Var}[Z_2] + 2 c_1 c_2 \operatorname{Cov}[Z_1, Z_2] \\
&= c_1^2 \sigma_0^2 ((X^T X)^{-1})_{11} + c_2^2 \sigma_0^2 ((X^T X)^{-1})_{22} + 2 c_1 c_2 \sigma_0^2 ((X^T X)^{-1})_{12}.
\end{aligned}$$

Letting $\hat{c}_1 = \hat{\beta}_1 / \sqrt{\hat{\beta}_1^2 + \hat{\beta}_2^2}$ and $\hat{c}_2 = \hat{\beta}_2 / \sqrt{\hat{\beta}_1^2 + \hat{\beta}_2^2}$, we may estimate the standard error of $\hat{A}_1$ by

$$\hat{\text{se}} = \sqrt{\hat{c}_1^2 \sigma_0^2 ((X^T X)^{-1})_{11} + \hat{c}_2^2 \sigma_0^2 ((X^T X)^{-1})_{22} + 2 \hat{c}_1 \hat{c}_2 \sigma_0^2 ((X^T X)^{-1})_{12}},$$

and construct a 95% confidence interval for A_1 as $\hat{A}_1 \pm z(0.025)\hat{\text{se}}$.

[1] In fact, the distribution of $\hat{\beta}$ is exactly this multivariate normal distribution even for small n, because $Y = X\beta + \varepsilon \sim \mathcal{N}(X\beta, \sigma_0^2 I)$ so that $\hat{\beta} = (X^T X)^{-1} X^T Y \sim \mathcal{N}((X^T X)^{-1} X^T X \beta, \sigma_0^2 (X^T X)^{-1} X^T X (X^T X)^{-1}) = \mathcal{N}(\beta, \sigma_0^2 (X^T X)^{-1})$ by the property derived in Homework 1, Problem 4(a).