STATS 200: Introduction to Statistical Inference **Autumn 2016**

Lecture 12 — Parametric models and method of moments

In the last unit, we discussed hypothesis testing, the problem of answering a binary question about the data distribution. We will now turn to the question of how to estimate the parameter(s) of this distribution.

A **parametric model** is a family of probability distributions that can be described by a finite number of parameters[1]. We've already seen many examples of parametric models:

- The family of normal distributions $\mathcal{N}(\mu, \sigma^2)$, with parameters μ and σ^2.
- The family of Bernoulli distributions Bernoulli(p), with a single parameter p.
- The family of Gamma distributions Gamma(α, β), with parameters α and β.

We will denote a general parametric model by $\{f(x|\theta) : \theta \in \Omega\}$, where $\theta \in \mathbb{R}^k$ represents k **parameters**, $\Omega \subseteq \mathbb{R}^k$ is the **parameter space** to which the parameters must belong, and $f(x|\theta)$ is the PDF or PMF for the distribution having parameters θ. For example, in the $\mathcal{N}(\mu, \sigma^2)$ model above, $\theta = (\mu, \sigma^2)$, $\Omega = \mathbb{R} \times \mathbb{R}_+$ where $\mathbb{R}_+$ is the set of positive real numbers, and

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Given data $X_1, \ldots, X_n$, the question of which parametric model we choose to fit to the data usually depends on what the data values represent (number of occurrences over a period of time? aggregation of many small effects?) as well as visual examination of the shape of the data histogram. This question is discussed in the context of several examples in Rice Sections 8.2–8.3.

Our main question of interest in this unit will be the following: After specifying an appropriate parametric model $\{f(x|\theta) : \theta \in \Omega\}$, and given observations

$$X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta),$$

how can we estimate the unknown parameter θ and quantify the uncertainty in our estimate?

12.1 Method of moments

If θ is a single number, then a simple idea to estimate θ is to find the value of θ for which the theoretical mean of $X \sim f(x|\theta)$ equals the observed sample mean $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$.

Example 12.1. The **Poisson distribution** with parameter $\lambda > 0$ is a discrete distribution over the non-negative integers $\{0, 1, 2, 3, \ldots\}$ having PMF

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

[1] The number of parameters is fixed and cannot grow with the sample size

If $X \sim \text{Poisson}(\lambda)$, then it has mean $\mathbb{E}[X] = \lambda$. Hence for data $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, a simple estimate of λ is the sample mean $\hat{\lambda} = \bar{X}$.

Example 12.2. The **exponential distribution** with parameter $\lambda > 0$ is a continuous distribution over $\mathbb{R}_+$ having PDF

$$f(x|\lambda) = \lambda e^{-\lambda x}.$$

If $X \sim \text{Exponential}(\lambda)$, then $\mathbb{E}[X] = \frac{1}{\lambda}$. Hence for data $X_1, \ldots, X_n \overset{IID}{\sim} \text{Exponential}(\lambda)$, we estimate λ by the value $\hat{\lambda}$ which satisfies $\frac{1}{\hat{\lambda}} = \bar{X}$, i.e. $\hat{\lambda} = \frac{1}{\bar{X}}$.

More generally, for $X \sim f(x|\theta)$ where θ contains k unknown parameters, we may consider the first k **moments** of the distribution of X, which are the values

$$\begin{aligned} \mu_1 &= \mathbb{E}[X] \\ \mu_2 &= \mathbb{E}[X^2] \\ &\vdots \\ \mu_k &= \mathbb{E}[X^k], \end{aligned}$$

and compute these moments in terms of θ. To estimate θ from data $X_1, \ldots, X_n$, we solve for the value of θ for which these moments equal the observed sample moments

$$\begin{aligned} \hat{\mu}_1 &= \tfrac{1}{n}(X_1 + \ldots + X_n) \\ &\vdots \\ \hat{\mu}_k &= \tfrac{1}{n}(X_1^k + \ldots + X_n^k). \end{aligned}$$

(This yields k equations in k unknown parameters.) The resulting estimate of θ is called the **method of moments estimator**.

Example 12.3. Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \mu^2 + \sigma^2$. So the method of moments estimators $\hat{\mu}$ and $\hat{\sigma}^2$ for μ and σ^2 solve the equations

$$\begin{aligned} \hat{\mu} &= \hat{\mu}_1, \\ \hat{\sigma}^2 + \hat{\mu}^2 &= \hat{\mu}_2. \end{aligned}$$

The first equation yields $\hat{\mu} = \hat{\mu}_1 = \bar{X}$, and the second yields

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n}\left(\sum_{i=1}^n X_i^2 - 2\sum_{i=1}^n X_i\bar{X} + n\bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Example 12.4. Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$. If $X \sim \text{Gamma}(\alpha, \beta)$, then $\mathbb{E}[X] = \frac{\alpha}{\beta}$ and $\mathbb{E}[X^2] = \frac{\alpha+\alpha^2}{\beta^2}$. So the method of moments estimators $\hat{\alpha}, \hat{\beta}$ solve the equations

$$\begin{aligned} \frac{\hat{\alpha}}{\hat{\beta}} &= \hat{\mu}_1, \\ \frac{\hat{\alpha} + \hat{\alpha}^2}{\hat{\beta}^2} &= \hat{\mu}_2. \end{aligned}$$

Substituting the first equation into the second,

$$\left(\frac{1}{\hat{\alpha}}+1\right)\hat{\mu}_1^2=\hat{\mu}_2,$$

so

$$\hat{\alpha}=\frac{1}{\frac{\hat{\mu}_2}{\hat{\mu}_1^2}-1}=\frac{\hat{\mu}_1^2}{\hat{\mu}_2-\hat{\mu}_1^2}=\frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2}.$$

The first equation then yields

$$\hat{\beta}=\frac{\hat{\alpha}}{\hat{\mu}_1}=\frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2}.$$

12.2 Bias, variance, and mean-squared-error

Consider the case of a single parameter $\theta \in \mathbb{R}$. Any estimator $\hat{\theta} := \hat{\theta}(X_1,\ldots,X_n)$ is a statistic—it has variability due to the randomness of the data $X_1,\ldots,X_n$ from which it is computed. Supposing that $X_1,\ldots,X_n \overset{IID}{\sim} f(x|\theta)$ (so the parametric model is correct and the true parameter is θ), we can think about whether $\hat{\theta}$ is a "good" estimate of the true parameter θ in a variety of different ways:

- The **bias** of $\hat{\theta}$ is $\mathbb{E}_\theta[\hat{\theta}]-\theta$. Here and below, $\mathbb{E}_\theta$ denotes the expectation with respect to $X_1,\ldots,X_n \overset{IID}{\sim} f(x|\theta)$.
- The **standard error** of $\hat{\theta}$ is its standard deviation $\sqrt{\mathrm{Var}_\theta[\hat{\theta}]}$. Here and below, Var_θ denotes the variance with respect to $X_1,\ldots,X_n \overset{IID}{\sim} f(x|\theta)$.
- The **mean-squared-error (MSE)** of $\hat{\theta}$ is $\mathbb{E}_\theta[(\hat{\theta}-\theta)^2]$.

The bias measures how close the average value of $\hat{\theta}$ is to the true parameter θ; the standard error measures how variable is $\hat{\theta}$ around this average value. An estimator with small bias need not be an accurate estimator, if it has large standard error, and conversely an estimator with small standard error need not be accurate if it has large bias. The mean-squared-error encompasses both bias and variance: For any random variable X and any constant $c \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[(X-c)^2]&=\mathbb{E}[(X-\mathbb{E}X+\mathbb{E}X-c)^2]\\&=\mathbb{E}[(X-\mathbb{E}X)^2]+\mathbb{E}[2(X-\mathbb{E}X)(\mathbb{E}X-c)]+\mathbb{E}[(\mathbb{E}X-c)^2]\\&=\mathrm{Var}[X]+2(\mathbb{E}X-c)\mathbb{E}[X-\mathbb{E}X]+(\mathbb{E}X-c)^2\\&=\mathrm{Var}[X]+(\mathbb{E}X-c)^2,\end{aligned}$$

where we used that $\mathbb{E}X-c$ is a constant and $\mathbb{E}[X-\mathbb{E}X]=0$. Applying this to $X=\hat{\theta}$ and $c=\theta$,

$$\mathbb{E}_\theta[(\hat{\theta}-\theta)^2]=\mathrm{Var}[\hat{\theta}]+(\mathbb{E}_\theta[\hat{\theta}]-\theta)^2.$$

We obtain the **bias-variance decomposition** of mean-squared-error:

$$\text{Mean-squared-error}=\text{Variance}+\text{Bias}^2.$$

An important remark is that the bias, standard error, and MSE may depend on the true parameter θ and take different values for different θ. We say that $\hat{\theta}$ is **unbiased** for θ if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for *all* $\theta \in \Omega$.

Example 12.5. In the model $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, the method-of-moments estimator of λ was $\hat{\lambda} = \bar{X}$. Then

$$\mathbb{E}_\lambda[\hat{\lambda}] = \mathbb{E}_\lambda[\bar{X}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\lambda[X_i] = \lambda,$$

where the last equality uses $\mathbb{E}[X] = \lambda$ if $X \sim \text{Poisson}(\lambda)$. So $\mathbb{E}_\lambda[\hat{\lambda}] - \lambda = 0$ for all $\lambda > 0$, and $\hat{\lambda}$ is an unbiased estimator of λ. Also,

$$\text{Var}_\lambda[\hat{\lambda}] = \text{Var}_\lambda[\bar{X}] = \frac{1}{n^2}\sum_{i=1}^n \text{Var}_\lambda[X_i] = \frac{\lambda}{n},$$

where we have used that $X_1, \ldots, X_n$ are independent and $\text{Var}[X] = \lambda$ if $X \sim \text{Poisson}(\lambda)$. Hence the standard error of $\hat{\lambda}$ is $\sqrt{\lambda/n}$, and the MSE is λ/n. Note that both of these depend on λ—they are larger when λ is larger.

As we do not know λ, in practice to determine the variability of $\hat{\lambda}$ we may estimate the standard error by $\sqrt{\hat{\lambda}/n} = \sqrt{\bar{X}/n}$. For large n, this is justified by the fact that $\hat{\lambda}$ is unbiased with standard error of the order $1/\sqrt{n}$, so we expect $\hat{\lambda} - \lambda$ to be of this order. Hence the estimated standard error $\sqrt{\hat{\lambda}/n}$ should be very close to the true standard error $\sqrt{\lambda/n}$. (We expect the difference between $\sqrt{\lambda/n}$ and $\sqrt{\hat{\lambda}/n}$ to be of the smaller order $1/n$.)

Example 12.6. In the model $X_1, \ldots, X_n \overset{IID}{\sim} \text{Exponential}(\lambda)$, the method-of-moments estimator of λ was $\hat{\lambda} = 1/\bar{X}$. This estimator is biased: Recall Jensen's inequality, which says for any strictly convex function $g : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}[g(X)] > g(\mathbb{E}[X])$. The function $x \mapsto 1/x$ is strictly convex, so

$$\mathbb{E}_\lambda[\hat{\lambda}] = \mathbb{E}_\lambda[1/\bar{X}] > 1/\mathbb{E}_\lambda[\bar{X}] = 1/(1/\lambda) = \lambda,$$

where we used $\mathbb{E}_\lambda[\bar{X}] = \mathbb{E}_\lambda[X_1] = 1/\lambda$ when $X_1, \ldots, X_n \overset{IID}{\sim} \text{Exponential}(\lambda)$. So $\mathbb{E}_\lambda[\hat{\lambda}] - \lambda > 0$ for all $\lambda > 0$, meaning $\hat{\lambda}$ always has positive bias.

To compute exactly the bias, variance, and MSE of $\hat{\lambda}$, note that Exponential(λ) is the same distribution as Gamma($1, \lambda$). Then $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) \sim \text{Gamma}(n, n\lambda)$. (This may be shown by calculating the MGF of $\bar{X}$, as in the examples of Lecture 3.) The distribution of $\hat{\lambda} = 1/\bar{X}$ is called the Inverse-Gamma($n, n\lambda$) distribution, which has mean $\frac{\lambda n}{n-1}$ and variance $\frac{\lambda^2 n^2}{(n-1)^2(n-2)}$ for $n \geq 3$. So the bias, variance, and MSE are given by

$$\text{Bias} = \mathbb{E}_\lambda[\hat{\lambda}] - \lambda = \frac{\lambda n}{n-1} - \lambda = \frac{\lambda}{n-1},$$

$$\text{Variance} = \text{Var}_\lambda[\hat{\lambda}] = \frac{\lambda^2 n^2}{(n-1)^2(n-2)},$$

$$\text{MSE} = \frac{\lambda^2 n^2}{(n-1)^2(n-2)} + \left(\frac{\lambda}{n-1}\right)^2 = \frac{\lambda^2(n+2)}{(n-1)(n-2)}.$$