S&DS 242/542: Theory of Statistics

Lecture 11: Parametric models and method of moments

Parametric models

This unit of our course will be about fitting parametric models to data. We will discuss how to:

- Estimate unknown parameters of a model
- Construct confidence intervals and quantify uncertainty
- Test hypotheses about unknown parameters

We will explore frequentist and Bayesian approaches to these questions, and also think about these questions in contexts of model misspecification.

Parametric models

A **parametric model** is a family of probability distributions that can be described by a small number of parameters.

We've seen many examples already, including:

- $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.
- Bernoulli(p) with parameter $p \in [0, 1]$.
- Poisson(λ) with parameter $\lambda > 0$.
- Gamma(α, β) with parameters $\alpha, \beta > 0$.

Parametric models

We will denote a general parametric model by its PDF or PMF $f(x \mid \theta)$, which depends on a vector of k **parameters** $\theta \in \mathbb{R}^k$.

The set of allowable parameter values for the model is the **parameter space** — this may be all, or only a subset, of $\mathbb{R}^k$.

For example, in the $\mathcal{N}(\mu, \sigma^2)$ model, the parameters may be $\theta = (\mu, \sigma^2)$ and

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

The parameter space may be $\{(\mu, \sigma^2) \in \mathbb{R}^2 : \sigma^2 > 0\}$.

Choosing the model

Our choice of model may depend on many factors, including:

- What the data values represent. (Are they discrete or continuous measurements? Can they be negative?)
- Our understanding of the generative process for the data.
- Exploratory analysis and visual examination of the data.
- Considerations of computational time and cost.
- Considerations of how many parameters we can accurately learn given the amount of data that we have.
- Considerations of predictive accuracy, if the model is to make predictions on new unseen examples.

In this and next lecture, we will study the simple question: Assuming

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

how can we estimate the unknown parameter θ?

Method of moments

Method of moments for a single parameter

If $\theta \in \mathbb{R}$ is a single number, the **method of moments** estimator $\hat{\theta}$ is the value of θ for which the theoretical mean of the distribution $f(x \mid \theta)$ matches the sample mean $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$.

Example: Suppose $X_1, \ldots, X_n \overset{IID}{\sim}$ Poisson(λ).

If $X \sim$ Poisson(λ), then $\mathbb{E}[X] = \lambda$. (Lecture 2)

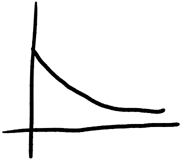
So the method-of-moments (MoM) estimator $\hat{\lambda}$ is just

$$\hat{\lambda} = \bar{X} = \frac{1}{n}(X_1 + \ldots + X_n).$$

Method of moments for a single parameter

Example: Suppose $X_1, \ldots, X_n \overset{IID}{\sim}$ Exponential(λ).

Exponential(λ) has PDF $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x > 0$



If $X \sim$ Exponential(λ) then $\mathbb{E}[X] = \frac{1}{\lambda}$.

$$\Big[\ \mathbb{E}[X] = \int_0^\infty \underbrace{x}_{u} \cdot \underbrace{\lambda e^{-\lambda x}\, dx}_{dv}$$

$$= -x \cdot e^{-\lambda x}\Big|_0^\infty - \int_0^\infty -e^{-\lambda x}\, dx$$

$$= \int_0^\infty e^{-\lambda x}\, dx = -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}.\ \Big]$$

The M-o-M estimator $\hat{\lambda}$ solves:

$$\frac{1}{\hat{\lambda}} = \bar{X} \Rightarrow \hat{\lambda} = \frac{1}{\bar{X}}.$$

Method of moments for multiple parameters

Equating the theoretical mean of $f(x \mid \theta)$ to the sample mean $\bar{X}$ gives one equation in the unknown parameters.

To estimate $\theta \in \mathbb{R}^k$ having k unknown parameters, in general we would need k equations. We may consider the first k **moments** of the distribution $X \sim f(x \mid \theta)$, which are the values

$$\mu_1 = \mathbb{E}[X], \quad \mu_2 = \mathbb{E}[X^2], \quad \ldots \quad \mu_k = \mathbb{E}[X^k].$$

The **method of moments estimator** $\hat{\theta}$ is the value of θ for which $\mu_1, \ldots, \mu_k$ match the observed sample moments

$$\begin{aligned}
\hat{\mu}_1 &= \tfrac{1}{n}(X_1 + \ldots + X_n) \\
\hat{\mu}_2 &= \tfrac{1}{n}(X_1^2 + \ldots + X_n^2) \\
&\vdots \\
\hat{\mu}_k &= \tfrac{1}{n}(X_1^k + \ldots + X_n^k)
\end{aligned}$$

Method of moments for multiple parameters

Example: Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. Let $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\mu_1 = \mathbb{E}[X] = \mu, \quad \mu_2 = \mathbb{E}[X^2] = \mathrm{Var}[X] + \mathbb{E}[X]^2 = \sigma^2 + \mu^2$$

Let $\hat{\mu}_1 = \bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$

$$\hat{\mu}_2 = \frac{1}{n}(X_1^2 + \ldots + X_n^2)$$

The M-o-M estimators $(\hat{\mu}, \hat{\sigma}^2)$ solve:

$$\hat{\mu} = \hat{\mu}_1, \quad \hat{\sigma}^2 + \hat{\mu}^2 = \hat{\mu}_2$$

$$\Rightarrow \hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \hat{\mu}_2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2$$

$$= \frac{1}{n}\sum_{i=1}^n X_i^2 - 2\bar{X}^2 + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$

$$= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

Method of moments for multiple parameters

Example: Let $X_1, \dots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$.

Gamma(α,β) has PDF $f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x>0$

If $X \sim \text{Gamma}(\alpha,\beta)$, then $\mathbb{E}[X] = \frac{\alpha}{\beta}$ and $\text{Var}[X] = \frac{\alpha}{\beta^2}$

(check this by calculus).

$\Rightarrow \mu_1 = \mathbb{E}[X] = \frac{\alpha}{\beta}, \quad \mu_2 = \text{Var}[X] + \mathbb{E}[X]^2 = \frac{\alpha + \alpha^2}{\beta^2}$

Let $\hat{\mu}_1 = \bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$

$\hat{\mu}_2 = \frac{1}{n}(X_1^2 + \dots + X_n^2)$.

Method of moments for multiple parameters

The M-o-M estimators $(\hat{\alpha}, \hat{\beta})$ solve:

$$\frac{\hat{\alpha}}{\hat{\beta}} = \hat{\mu}_1, \quad \frac{\hat{\alpha} + \hat{\alpha}^2}{\hat{\beta}^2} = \hat{\mu}_2$$

$$\Rightarrow \hat{\beta} = \frac{\hat{\alpha}}{\hat{\mu}_1} \text{ and } \frac{\hat{\alpha} + \hat{\alpha}^2}{(\hat{\alpha}/\hat{\mu}_1)^2} = \hat{\mu}_2$$

$$\Rightarrow \frac{\hat{\alpha} + \hat{\alpha}^2}{\hat{\alpha}^2} = 1 + \frac{1}{\hat{\alpha}} = \frac{\hat{\mu}_2}{\hat{\mu}_1^2}$$

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

$$= \frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$$

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

Generalized method of moments

Instead of choosing to match the means of $X, X^2, \ldots, X^k$, one may choose to match the means of other functions $T_1(X), T_2(X), \ldots, T_k(X)$.

For example, suppose $\theta \in \mathbb{R}$ is a single parameter, and let $T : \mathbb{R} \to \mathbb{R}$ be any function. A *generalized method of moments estimator* may choose θ so that the theoretical mean $\mathbb{E}_\theta[T(X)]$ matches the sample mean $\frac{1}{n}(T(X_1) + \ldots + T(X_n))$.

Here, we write $\mathbb{E}_\theta$ to indicate that the expectation is computed assuming that $X \sim f(x \mid \theta)$ with true parameter θ.

Generalized method of moments

Example: $X_1, \ldots, X_n \overset{IID}{\sim} \text{Pareto}(\alpha, 1)$.

Pareto$(\alpha,1)$ has PDF $f(x|\alpha) = \frac{\alpha}{x^{\alpha+1}}$ for $x>1$

$(=0$ for $x \leq 1)$.

If $X \sim \text{Pareto}(\alpha,1)$, then

$$\mathbb{E}[X] = \int_1^\infty x \cdot \frac{\alpha}{x^{\alpha+1}}\,dx = \int_1^\infty \frac{\alpha}{x^\alpha}\,dx$$

$$= \frac{\alpha}{-\alpha+1} x^{-\alpha+1}\Big|_1^\infty = \frac{\alpha}{\alpha-1} \text{ for } \alpha>1$$

(and $\mathbb{E}[X] = \infty$ if $\alpha \leq 1$)

The M-o-M estimator solves $\frac{\hat{\alpha}}{\hat{\alpha}-1} = \bar{X} \Rightarrow \hat{\alpha} = \frac{\bar{X}}{\bar{X}-1}$.

Generalized method of moments

Example: $X_1, \ldots, X_n \overset{IID}{\sim} \text{Pareto}(\alpha, 1)$. Consider instead $T(X) = \log X$.

$$\mathbb{E}[\log X] = \int_1^\infty (\log x) \cdot \frac{\alpha}{x^{\alpha+1}}\, dx \qquad [\text{let } u = \log x,\ x = e^u]$$

$$= \int_0^\infty u \cdot \frac{\alpha}{e^{u(\alpha+1)}}\, e^u\, du$$

$$= \int_0^\infty u \cdot \underbrace{\alpha e^{-\alpha u}}_{\text{PDF of Exponential}(\alpha)}\, du = \frac{1}{\alpha}$$

So a generalized M-o-M estimator $\hat{\alpha}$ based on $T(X) = \log X$

solves $\quad \dfrac{1}{\hat{\alpha}} = \dfrac{1}{n}(\log X_1 + \ldots + \log X_n) \Rightarrow \hat{\alpha} = \dfrac{n}{\log X_1 + \ldots + \log X_n}$.

Bias, variance, and mean-squared-error

Bias and variance

Consider a parameter $\theta \in \mathbb{R}$. Any estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is a statistic — i.e. a function of the observed data — and has variability due to the randomness of the data $X_1, \ldots, X_n$.

If $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ with true parameter θ, we can measure the accuracy of $\hat{\theta}$ via its bias and variance:

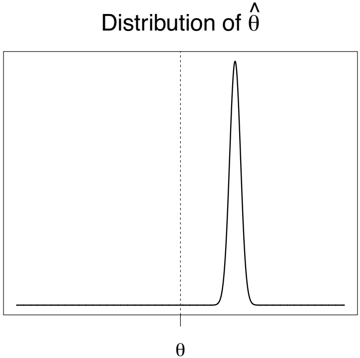
- The **bias** of $\hat{\theta}$ is $\mathbb{E}_\theta[\hat{\theta}] - \theta = \mathbb{E}_\theta[\hat{\theta}(X_1, \ldots, X_n)] - \theta$. Here $\mathbb{E}_\theta$ is the expectation computed assuming $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$.
- The **variance** of $\hat{\theta}$

$$\mathrm{Var}_\theta[\hat{\theta}] = \mathrm{Var}_\theta[\hat{\theta}(X_1, \ldots, X_n)]$$

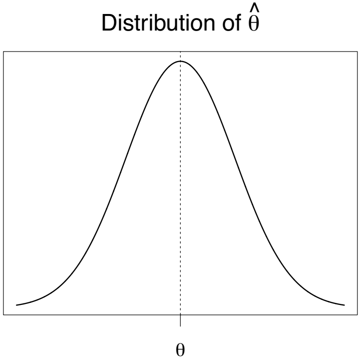
also computed assuming $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$.

The **standard error** of $\hat{\theta}$ is the standard deviation $\sqrt{\mathrm{Var}_\theta[\hat{\theta}]}$.

Bias and variance



Large bias



Large variance

Bias measures how close the average value of $\hat{\theta}$ is to the true parameter θ. Variance measures how variable is this estimate $\hat{\theta}$ around its average value.

Bias and variance

The **mean-squared-error (MSE)** of $\hat{\theta}$ is $\mathbb{E}_\theta[(\hat{\theta}-\theta)^2]$. It encompasses both bias and variance:

For any random variable Y, constant $c \in \mathbb{R}$,

$$\mathbb{E}[(Y-c)^2] = \mathbb{E}[(\underbrace{Y-\mathbb{E}Y}+\underbrace{\mathbb{E}Y-c})^2]$$

$$= \mathbb{E}[(Y-\mathbb{E}Y)^2] + 2\mathbb{E}[(Y-\mathbb{E}Y)(\mathbb{E}Y-c)] + \mathbb{E}[(\mathbb{E}Y-c)^2]$$

$$= \mathbb{E}[(Y-\mathbb{E}Y)^2] + 2(\mathbb{E}Y-c)\cdot\underbrace{\mathbb{E}[Y-\mathbb{E}Y]}_{=0} + (\mathbb{E}Y-c)^2$$

$$= \mathrm{Var}[Y] + (\mathbb{E}Y-c)^2$$

Apply w/ $Y=\hat{\theta}$, $c=\theta \Rightarrow \mathbb{E}_\theta[(\hat{\theta}-\theta)^2] = \mathrm{Var}_\theta[\hat{\theta}] + (\mathbb{E}_\theta\hat{\theta}-\theta)^2$

Bias and variance

This is the **bias-variance decomposition** of mean-squared-error:

$$\text{MSE} = \text{Variance} + \text{Bias}^2$$

Typically the bias, variance, and MSE all depend on the true parameter θ. That is, the accuracy of the estimator $\hat{\theta}$ may be different for different values of the true parameter θ.

We say that $\hat{\theta}$ is **unbiased** for θ if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for *all* possible parameter values θ belonging to the parameter space of the model.

Method of moments in the Poisson model

Recall, for $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, the method of moments estimator of λ was $\hat{\lambda} = \bar{X}$.

For $X_i \sim \text{Poisson}(\lambda)$, we have $\mathbb{E}_\lambda[X_i] = \text{Var}_\lambda[X_i] = \lambda$. Then

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

So $\mathbb{E}_\lambda[\hat{\lambda}] = \lambda$ for all $\lambda > 0$, meaning that $\hat{\lambda}$ is an unbiased estimator of λ. For the variance,

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

The standard error is $\sqrt{\frac{\lambda}{n}}$, and the MSE is variance + bias$^2 = \frac{\lambda}{n}$.

Estimating the standard error

We would often wish to report the standard error of $\hat{\lambda}$. Since the true standard error $\sqrt{\frac{\lambda}{n}}$ depends on λ, which is unknown, we typically report a *plug-in estimate* $\sqrt{\frac{\hat{\lambda}}{n}}$ for this standard error.

You may ask why we don't further account for the uncertainty of *this* estimate $\sqrt{\frac{\hat{\lambda}}{n}}$. We usually don't, because this additional error is much smaller than the standard error itself for large sample sizes n: If $\hat{\lambda} - \lambda \asymp \frac{1}{\sqrt{n}}$, then (by a Taylor expansion) $\sqrt{\frac{\lambda}{n}} - \sqrt{\frac{\hat{\lambda}}{n}} \asymp \frac{1}{n}$.

For example: If $n = 100$ and we estimate $\hat{\lambda} = 1$, we may report the standard error as $\sqrt{\frac{\hat{\lambda}}{n}} = 0.1$. The difference between this and the true standard error should be on the scale of $\frac{1}{n} = 0.01$, which is small compared to our reported standard error of 0.1.

Method of moments in the Exponential model

Recall, for $X_1, \ldots, X_n \overset{IID}{\sim}$ Exponential(λ), the method of moments estimator of λ was $\hat{\lambda} = 1/\bar{X}$. Note that

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

So $\bar{X}$ is an unbiased estimator of $1/\lambda$. However, this does not mean that $1/\bar{X}$ is an unbiased estimator of λ.

Recall Jensen's inequality: For any random variable Y taking values in (a, b) and any convex function $g : (a, b) \to \mathbb{R}$,

$$\mathbb{E}[g(Y)] \geq g(\mathbb{E}[Y]).$$

If Y is not a constant and g is strictly convex, then this inequality holds strictly. E.g. $\mathbb{E}[Y^2] > (\mathbb{E}[Y])^2$ as long as Y is not a constant.

Method of moments in the Exponential model

The function $g(x) = 1/x$ is strictly convex on the interval $(0, \infty)$ of possible values for $\bar{X}$, so

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

Then $\mathbb{E}_\lambda[\hat{\lambda}] - \lambda > 0$ for all $\lambda > 0$, meaning that $\hat{\lambda}$ has positive bias.

One may derive the exact bias and standard error in this example by using that $\hat{\lambda} = 1/\bar{X} \sim \text{Inverse-Gamma}(n, n\lambda)$. Then

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

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

For large n, we see that bias $\asymp \frac{1}{n}$, standard error $\asymp \frac{1}{\sqrt{n}}$, so MSE is dominated by variance rather than squared bias. This is a general phenomenon that we will observe again in later examples.