7. Models, Statistical Inference and Learning

7.2 Parametric and Nonparametric Models

A **statistical model** is a set of distributions $\mathfrak{F}$.

A **parametric model** is a set $\mathfrak{F}$ that may be parametrized by a finite number of parameters. For example, if we assume that data comes from a normal distribution then

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \;\; \mu \in \mathbb{R}, \; \sigma > 0 \right\}$$

In general, a parametric model takes the form

$$\mathfrak{F} = \{ f(x; \theta) : \; \theta \in \Theta \}$$

where θ is an unknown parameter that takes values in the **parameter space** Θ.

If θ is a vector and we are only interested in one component of θ, we call the remaining parameters **nuisance parameters**.

A **nonparametric model** is a set $\mathfrak{F}$ that cannot be parametrized by a finite number of parameters.

Some notation

If $\mathfrak{F} = \{ f(x; \theta) : \; \theta \in \Theta \}$ is a parametric model, we write

$$\mathbb{P}_\theta(X \in A) = \int_A f(x; \theta) dx$$

$$\mathbb{E}_\theta(X \in A) = \int_A x f(x; \theta) dx$$

The subscript θ indicates that the probability or expectation is defined with respect to $f(x; \theta)$; it does not mean we are averaging over θ.

7.3 Fundamental Concepts in Inference

7.3.1 Point estimation

Let $X_1, \dots, X_n$ be n iid data points from some distribution F. A point estimator $\hat{\theta_n}$ of a parameter θ is some function:

$$\hat{\theta_n} = g(X_1, \dots, X_n)$$

We define

$$\text{bias}(\hat{\theta_n}) = \mathbb{E}_\theta(\hat{\theta_n}) - \theta$$

to be the bias of $\hat{\theta_n}$. We say that $\hat{\theta_n}$ is **unbiased** if $\mathbb{E}_\theta(\hat{\theta_n}) = 0$.

A point estimator $\hat{\theta_n}$ of a parameter θ is **consistent** if $\hat{\theta_n} \xrightarrow{\text{P}} \theta$.

The distribution of $\hat{\theta_n}$ is called the **sampling distribution**.

The standard deviation of $\hat{\theta_n}$ is called the **standard error**, denoted by se:

$$\text{se} = \text{se}(\hat{\theta_n}) = \sqrt{\mathbb{V}(\hat{\theta_n})}$$

Often it is not possible to compute the standard error but usually we can estimate the standard error. The estimated standard error is denoted by $\hat{\text{se}}$.

Example. Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ and let $\hat{p_n} = n^{-1} \sum_i X_i$. Then $\mathbb{E}(\hat{p_n}) = n^{-1} \sum_i \mathbb{E}(X_i) = p$ so $\hat{p_n}$ is unbiased. The standard error is $\text{se} = \sqrt{\mathbb{V}(\hat{p_n})} = \sqrt{p(1-p)/n}$. The estimated standard error is $\hat{\text{se}} = \sqrt{\hat{p}(1-\hat{p})/n}$.

The quality of a point estimate is sometimes assessed by the **mean squared error**, or MSE, defined by:

$$\text{MSE} = \mathbb{E}_\theta \left(\hat{\theta_n} - \theta \right)^2$$

Theorem 7.8. The MSE can be rewritten as:

$$\text{MSE} = \text{bias}(\hat{\theta_n})^2 + \mathbb{V}_\theta(\hat{\theta_n})$$

Proof. Let $\bar{\theta_n} = \mathbb{E}_\theta(\hat{\theta_n})$. Then

$$
\begin{aligned}
\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 &= \mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2 && (1)\\
&= \mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n)^2 + 2(\bar{\theta}_n - \theta)\mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n)^2 + \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 && (2)\\
&= (\bar{\theta}_n - \theta)^2 + \mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n)^2 && (3)\\
&= \text{bias}^2 + \mathbb{V}_\theta(\hat{\theta}_n) && (4)
\end{aligned}
$$

Theorem 7.9. If $\text{bias} \to 0$ and $\text{se} \to 0$ as $n \to \infty$ then $\hat{\theta}_n$ is consistent, that is, $\hat{\theta}_n \xrightarrow{\text{P}} \theta$.

Proof. If $\text{bias} \to 0$ and $\text{se} \to 0$ then, by theorem 7.8, $\text{MSE} \to 0$. It follows that $\hat{\theta}_n \xrightarrow{\text{qm}} \theta$ -- and quadratic mean convergence implies probability convergence.

An estimator is **asymptotically Normal** if

$$\frac{\hat{\theta}_n - \theta}{\text{se}} \rightsquigarrow N(0, 1)$$

7.3.2 Confidence sets

A $1 - \alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \dots, X_n)$ and $b = b(X_1, \dots, _n)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta$$

In words, (a, b) traps θ with probability $1 - \alpha$. We call $1 - \alpha$ the **coverage** of the confidence interval.

Note: C_n **is random and** θ **is fixed!**

If θ is a vector then we use a confidence set (such as a sphere or ellipse) instead of an interval.

Point estimators often have a limiting Normal distribution, meaning $\hat{\theta}_n \approx N(\theta, \hat{\text{se}}^2)$. In this case we can construct (approximate) confidence intervals as follows:

Theorem 7.14 (Normal-based Confidence Interval). Suppose that $\hat{\theta}_n \approx N(\theta, \hat{\text{se}}^2)$. Let Φ be the CDF of a standard Normal and let $z_{\alpha/2} = \Phi^{-1}\left(1 - (\alpha/2)\right)$, that is, $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ and $\mathbb{P}(-z_{\alpha/2} < Z < z_{alpha/2}) = 1 - \alpha$ where $Z \sim N(0, 1)$. Let

$$C_n = \left(\hat{\theta}_n - z_{\alpha/2}\hat{\text{se}},\ \hat{\theta}_n + z_{\alpha/2}\hat{\text{se}}\right)$$

Then

$$\mathbb{P}_\theta(\theta \in C_n) \to 1 - \alpha$$

.

Proof.

Let $Z_n = (\hat{\theta}_n - \theta)/\hat{\text{se}}$. By assumption $Z_n \rightsquigarrow Z \sim N(0,1)$. Hence,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\hat{\theta}_n - z_{\alpha/2}\hat{\text{se}} < \theta < \hat{\theta}_n + z_{\alpha/2}\hat{\text{se}}\right) && (5)\\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\hat{\theta}_n - \theta}{\hat{\text{se}}} < z_{\alpha/2}\right) && (6)\\
&\to \mathbb{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) && (7)\\
&= 1 - \alpha && (8)
\end{aligned}$$

7.3.3 Hypothesis Testing

In **hypothesis testing**, we start with some default theory -- called a **null hypothesis** -- and we ask if the data provide sufficient evidence to reject the theory. If not we retain the null hypothesis.

7.5 Technical Appendix

- Our definition of confidence interval requires that $\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ for all $\theta \in \Theta$.
- A **pointwise asymptotic** confidence interval requires that $\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ for all $\theta \in \Theta$.
- An **uniform asymptotic** confidence interval requires that $\liminf_{n\to\infty} \inf \theta \in \Theta \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$.

The approximate Normal-based interval is a pointwise asymptotic confidence interval. In general, it might not be a uniform asymptotic confidence interval.