16 Classical Statistical Inference

In Chapter 15, we focused on estimating the mean and variance of a distribution given observed samples. In this chapter and the next, we look at the more general question of statistical inference, where this time we are estimating the parameter(s) of a distribution or some other quantity. We will continue to use the notation for estimators given in Definition 15.1.

16.1 Towards More General Estimators

We start the chapter with another example of point estimation.

Example 16.1 (Estimating the number of pink jelly beans)

Consider the jar of jelly beans shown in Figure 16.1. Suppose that we know that the jar has 1000 jelly beans. Our goal is to estimate the number of pink jelly beans. Let

$$\theta = \text{Number of pink jelly beans in the jar.}$$

To estimate θ, we randomly sample $n = 20$ jelly beans with replacement.

Figure 16.1 *This jar has 1000 jelly beans. How many of them are pink?*

Let X be the number of pink jelly beans that we observe in our sample of $n = 20$.

Observe that X is a random variable (r.v.) since the experiment is random. X can take on values from 0 to n. We use r.v. $\hat{\theta}(X)$ to denote our estimator of θ.

Question: What is a reasonable guess for what $\hat{\theta}(X)$ might look like?

Hint: It is easier to think about a specific instantiation of X. For example, suppose we observe $X = x$ pink jelly beans.

Answer: If we observe x jelly beans in our sample, then a reasonable estimate for the *fraction* of pink jelly beans is $\frac{x}{n}$. Hence we estimate the *number* of pink jelly beans is

$$\hat{\theta}(X = x) = \left(\frac{x}{n}\right) \cdot 1000, \qquad 0 \leq x \leq n. \tag{16.1}$$

Now, since (16.1) holds for every value of x, it follows that we can define

$$\hat{\theta}(X) = \left(\frac{X}{n}\right) \cdot 1000. \tag{16.2}$$

Question: Is $\hat{\theta}(X)$, as defined in (16.2), an unbiased estimator of θ?

Hint: It helps to start by considering the distribution of X.

Answer: Let us define

$$p = \frac{\theta}{1000}$$

to be the true fraction of pink jelly beans. Then,

$$X \sim \text{Binomial}(n, p),$$

and hence

$$\mathbf{E}[X] = np = n \cdot \frac{\theta}{1000}.$$

From this it follows that

$$\begin{aligned}\mathbf{E}\left[\hat{\theta}(X)\right] = \mathbf{E}\left[\frac{X}{n} \cdot 1000\right] &= \mathbf{E}[X] \cdot \frac{1000}{n} \\ &= n \cdot \frac{\theta}{1000} \cdot \frac{1000}{n} \\ &= \theta. \qquad \checkmark\end{aligned}$$

Thus, $\hat{\theta}(X)$ is an unbiased estimator of θ.

Question: Is $\hat{\theta}(x)$ a consistent estimator of θ?

Answer: Yes! To see this, we will show that $\mathbf{MSE}(\hat{\theta}) \to 0$, as $n \to \infty$. Note that n can be arbitrarily high because we're sampling with replacement.

We start by observing that $\mathbf{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta})$, by Lemma 15.5. Hence,

$$\begin{aligned}\mathbf{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta}) &= \left(\frac{1000}{n}\right)^2 \cdot np(1-p) \\ &= \left(\frac{1000}{n}\right)^2 \cdot n \cdot \frac{\theta}{1000}\left(1 - \frac{\theta}{1000}\right) \\ &= \frac{\theta(1000-\theta)}{n}.\end{aligned}$$

Clearly, $\mathbf{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$, so $\hat{\theta}$ is a consistent estimator, by Lemma 15.7.

16.2 Maximum Likelihood Estimation

In the previous section, we came up with what seemed like a reasonable estimator. However, there was no specific *method* for coming up with this estimator, nor the estimators in the prior chapter. In this section we describe a specific *methodology* for deriving an estimator. The methodology is called **maximum likelihood estimation (MLE)**. It is the *classical inference methodology* adopted by statisticians who consider themselves to be *frequentists*. In the next chapter we will investigate a different methodology for coming up with estimators which is preferred by the *Bayesian* statisticians.

In explaining the MLE method, to simplify notation we will assume that the sample data is just a single r.v., X, but in general it can be $X_1, X_2, \ldots, X_n$. For now we will assume that we have a single unknown, θ, that we are trying to estimate; we will later consider multiple unknowns. The goal is to derive $\hat{\theta}(X)$, which is a maximum likelihood estimator of θ based on the sample data X; we refer to this as an **ML estimator**. To create an ML estimator, we first consider an *arbitrary specific value* of the sample data, $X = x$, and ask,

> "What is the value of θ which maximizes the likelihood of seeing $X = x$?"

The expression that we derive will be a function of x. Since x is chosen *arbitrarily*, this allow us to define $\hat{\theta}$ as a function of the r.v. X.

Algorithm 16.2 (Creating an ML estimator) *Our goal is to estimate an unknown value, θ, given sample data represented by r.v. X.*

1. *Define*

$$\hat{\theta}_{ML}(X = x) = \underset{\theta}{argmax}\, \mathbf{P}\{X = x \mid \theta\}.$$

$\mathbf{P}\{X = x \mid \theta\}$ is called the **likelihood function** *and represents the probability that $X = x$, given a particular θ. The value of θ which maximizes the likelihood function is denoted by $\hat{\theta}_{ML}(X = x)$.*

2. *Convert $\hat{\theta}_{ML}(X = x)$, which is a function of x, for any arbitrary x, into r.v. $\hat{\theta}_{ML}(X)$, which is a function of a r.v., by replacing x with X.*

The MLE method is best illustrated via an example. Returning to Example 16.1, suppose that in our sample of $n = 20$ jelly beans we observe $X = 3$ jelly beans.

Question: What is $\mathbf{P}\{X = 3 \mid \theta\}$?

Answer: If we're given that there are θ pink jelly beans, then the fraction of pink jelly beans is $p = \frac{\theta}{1000}$. Hence, given $n = 20$, we have

$$\mathbf{P}\{X = 3 \mid \theta\} = \binom{20}{3}\left(\frac{\theta}{1000}\right)^3 \cdot \left(1 - \frac{\theta}{1000}\right)^{17}.$$

Figure 16.2 shows the probability that $X = 3$ under all possible values of θ from 0 to 1000, assuming $n = 20$.

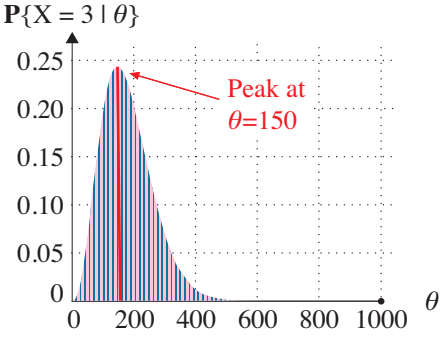

Figure 16.2 $\mathbf{P}\{X = 3 \mid \theta\}$ *as a function of θ, assuming $n = 20$.*

Question: Based on Figure 16.2, what value of θ maximizes $\mathbf{P}\{X = 3 \mid \theta\}$?

Answer: $\theta = 150$. So

$$\hat{\theta}_{\text{ML}}(X = 3) = \underset{\theta}{\text{argmax}}\, \mathbf{P}\{X = 3 \mid \theta\} = 150.$$

Question: What is the likelihood function, $\mathbf{P}\{X = x \mid \theta\}$?

Answer:

$$\mathbf{P}\{X = x \mid \theta\} = \binom{n}{x}\left(\frac{\theta}{1000}\right)^x \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}.$$

Question: What is $\hat{\theta}_{\text{ML}}(X = x) = \underset{\theta}{\text{argmax}}\, \mathbf{P}\{X = x \mid \theta\}$?

Answer: To answer this, we'll need to solve for the value of θ which maximizes the likelihood function:

$$\begin{aligned}
0 &= \frac{d}{d\theta}\mathbf{P}\{X = x \mid \theta\} \\
&= \frac{d}{d\theta}\binom{n}{x}\left(\frac{\theta}{1000}\right)^x \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x} \\
&= \binom{n}{x} \cdot \left(\frac{\theta}{1000}\right)^x \cdot (n-x) \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x-1} \cdot \frac{-1}{1000} \\
&\quad + \binom{n}{x} \cdot x \left(\frac{\theta}{1000}\right)^{x-1} \cdot \frac{1}{1000}\left(1 - \frac{\theta}{1000}\right)^{n-x}.
\end{aligned}$$

If we divide both sides by $\binom{n}{x} \cdot \left(\frac{\theta}{1000}\right)^{x-1} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-1-x}$, we are left with:

$$\begin{aligned}
0 &= -\frac{n-x}{1000} \cdot \frac{\theta}{1000} + \frac{x}{1000} \cdot \left(1 - \frac{\theta}{1000}\right) \\
0 &= -(n-x)\theta + x(1000 - \theta) \\
\theta &= \frac{1000x}{n}.
\end{aligned}$$

It is easily shown that the second derivative of the likelihood function is negative, and thus

$$\theta = \frac{1000x}{n}$$

is in fact the value of θ that maximizes the likelihood function. Hence,

$$\hat{\theta}_{\text{ML}}(X = x) = \frac{1000x}{n}. \tag{16.3}$$

Question: Given that

$$\hat{\theta}_{\text{ML}}(X = x) = \frac{1000x}{n}, \text{ for all } 0 \le x \le n,$$

what does this say about $\hat{\theta}_{\text{ML}}(X)$?

Answer:

$$\hat{\theta}_{\text{ML}}(X) = \frac{1000X}{n}.$$

Notice that this is the same estimator that we arrived at in (16.2); however, this time we followed a specific method (MLE) for coming up with the estimator.

16.3 More Examples of ML Estimators

Example 16.3 (Submissions to the Pittsburgh Supercomputing Center)

The number of jobs submitted daily to the Pittsburgh Supercomputing Center (PSC) follows a Poisson distribution with unknown parameter λ. Suppose that the numbers of job submissions on different days are independent. We observe the number of job submissions each day for a month, and denote these by $X_1, X_2, \ldots, X_{30}$. Our goal is to derive $\hat{\lambda}_{\text{ML}}(X_1, X_2, \ldots, X_{30})$, the ML estimator for λ.

Question: Before we do the computation, ask yourself: What do you expect the answer to be?

Hint: Recall that the parameter λ represents the mean of the Poisson distribution.

Answer: We are being asked to estimate the unknown parameter λ, which is the mean number of arrivals. It would make sense if this was simply the sample mean. That is:

$$\hat{\lambda}_{\text{ML}}(X_1, X_2, \ldots, X_{30}) = \frac{X_1 + X_2 + \cdots + X_{30}}{30}.$$

We now proceed to follow the MLE method, which will lead us to find that our intuition is in fact correct.

We write

$$\begin{aligned}
&\hat{\lambda}_{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) \\
&= \operatorname*{argmax}_{\lambda} \mathbf{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30} \mid \lambda\} \\
&= \operatorname*{argmax}_{\lambda} \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdots \frac{\lambda^{x_{30}} e^{-\lambda}}{x_{30}!} \\
&= \operatorname*{argmax}_{\lambda} \frac{\lambda^{x_1 + x_2 + \cdots + x_{30}} e^{-30\lambda}}{x_1! x_2! \cdots x_{30}!}.
\end{aligned}$$

To find the maximizing λ, we set the derivative of the likelihood function to 0:

$$\begin{aligned} 0 &= \frac{d}{d\lambda}\left(\frac{\lambda^{x_1+\cdots+x_{30}}e^{-30\lambda}}{x_1!x_2!\cdots x_{30}!}\right) \\ &= \frac{(x_1+\cdots+x_{30})\,\lambda^{x_1+\cdots+x_{30}-1}\cdot e^{-30\lambda} + \lambda^{x_1+\cdots+x_{30}}\cdot e^{-30\lambda}\cdot(-30)}{x_1!\cdots x_{30}!} \end{aligned}$$

Dividing both sides by the appropriate constants leaves us with

$$0 = (x_1+\cdots+x_{30}) + \lambda\cdot(-30). \tag{16.4}$$

Solving (16.4), and verifying that the second derivative is negative, yields

$$\lambda = \frac{x_1+\cdots+x_{30}}{30}$$

as the value of λ which maximizes the likelihood function.

Hence,

$$\hat{\lambda}_{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) = \frac{x_1+\cdots+x_{30}}{30}, \quad \forall x_1, \ldots, x_{30} \geq 0.$$

So

$$\hat{\lambda}_{\text{ML}}(X_1, X_2, \ldots, X_{30}) = \frac{X_1+X_2+\ldots X_{30}}{30},$$

as predicted.

16.4 Log Likelihood

Sometimes, rather than finding the value of θ that maximizes some probability, it is more convenient to maximize the log of that probability. Lemma 16.4 makes this clear.

Lemma 16.4 (Maximizing the log likelihood) *Given an unknown value, θ, that we are trying to estimate, suppose that we have sample data represented by r.v. X. Then,*

$$\hat{\theta}_{ML}(X = x) \equiv \underset{\theta}{argmax}\, \mathbf{P}\{X = x \mid \theta\} = \underset{\theta}{argmax}\, \log \mathbf{P}\{X = x \mid \theta\}.$$

Here, $\log \mathbf{P}\{X = x \mid \theta\}$ is referred to as the **log likelihood function**.

Proof: Maximizing the log likelihood is equivalent to maximizing the likelihood since log is a strictly increasing function. ■

Example 16.5 (Submissions to the PSC, revisited!)

Let's revisit Example 16.3, where the goal is to estimate λ. This time, however, we derive the estimator that maximizes the log likelihood:

$$\begin{aligned}
&\hat{\lambda}_{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) \\
&= \underset{\lambda}{\text{argmax}} \ \ln\left(\mathbf{P}\left\{X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30} \mid \lambda\right\}\right) \\
&= \underset{\lambda}{\text{argmax}} \ \ln\left(\mathbf{P}\left\{X_1 = x_1 \mid \lambda\right\} \cdot \mathbf{P}\left\{X_2 = x_2 \mid \lambda\right\} \cdots \mathbf{P}\left\{X_{30} = x_{30} \mid \lambda\right\}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\hat{\lambda}_{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) \\
&= \underset{\lambda}{\text{argmax}} \sum_{i=1}^{30} \ln \mathbf{P}\left\{X_i = x_i \mid \lambda\right\} \\
&= \underset{\lambda}{\text{argmax}} \sum_{i=1}^{30} \ln\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \\
&= \underset{\lambda}{\text{argmax}} \left(-30\lambda + \sum_{i=1}^{30} x_i \ln(\lambda) - \sum_{i=1}^{30} \ln(x_i!)\right) \\
&= \underset{\lambda}{\text{argmax}} \left(-30\lambda + \sum_{i=1}^{30} x_i \ln(\lambda)\right).
\end{aligned}$$

To find the maximizing λ, we set the derivative of the log likelihood function to 0:

$$0 = \frac{d}{d\lambda}\left(-30\lambda + \sum_{i=1}^{30} x_i \ln(\lambda)\right) = -30 + \left(\sum_{i=1}^{30} x_i\right) \cdot \frac{1}{\lambda}.$$

Hence,

$$\lambda = \frac{x_1 + x_2 + \cdots + x_{30}}{30}.$$

Thus again,

$$\hat{\lambda}_{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) = \frac{x_1 + x_2 + \cdots + x_{30}}{30}.$$

16.5 MLE with Data Modeled by Continuous Random Variables

When data is modeled by continuous random variables, we replace the probability mass function (p.m.f.) with the probability density function (p.d.f.) in expressing the likelihood. Definitions 16.6 and 16.7 provide a summary.

Definition 16.6 (MLE summary: single variable) *Given an unknown value, θ, that we wish to estimate:*
If the sample data is represented by **discrete** *r.v. X, then we define*

$$\hat{\theta}_{ML}(X = x) \equiv \underset{\theta}{argmax}\, \mathbf{P}\{X = x \mid \theta\}.$$

If the sample data is represented by **continuous** *r.v. X, we instead define*

$$\hat{\theta}_{ML}(X = x) \equiv \underset{\theta}{argmax}\, f_{X|\theta}(x).$$

Definition 16.7 (MLE summary: multiple variables) *Given an unknown value, θ, that we wish to estimate:*
If the sample data is represented by **discrete** *random variables $X_1, X_2, \ldots, X_n$, we define*

$$\hat{\theta}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \equiv \underset{\theta}{argmax}\, \mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n \mid \theta\}.$$

If the sample data is represented by **continuous** *random variables $X_1, X_2, \ldots, X_n$, we define*

$$\hat{\theta}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \equiv \underset{\theta}{argmax}\, f_{X_1, X_2, \ldots, X_n|\theta}(x_1, x_2, \ldots, x_n).$$

Example 16.8 (Time students spend on their probability homework)

Students often ask, "How long can I expect to spend on homework if I take the PnC probability class?" It turns out that the distribution of the time that students spend on homework is approximately distributed as Uniform$(0, b)$, where students can be viewed as independent in the time that they spend doing the homework. To get a feel for what b is, we survey three students. Let X_1, X_2, X_3 denote the times reported by the three students.

What is the ML estimator $\hat{b}_{\text{ML}}(X_1, X_2, X_3)$ for b?

$$\hat{b}_{\text{ML}}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \underset{b}{\text{argmax}}\ f_{X_1,X_2,X_3|b}(x_1, x_2, x_3).$$

$$f_{X_1,X_2,X_3|b}(x_1, x_2, x_3) = \begin{cases} \frac{1}{b^3} & \text{if } 0 < x_1, x_2, x_3 \leq b \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \frac{1}{b^3} & \text{if } b \geq \max\{x_1, x_2, x_3\} \\ 0 & \text{otherwise} \end{cases}.$$

Clearly $f_{X_1,X_2,X_3|b}(x_1, x_2, x_3)$ achieves its maximum when $b = \max\{x_1, x_2, x_3\}$. Therefore,

$$\hat{b}_{\text{ML}}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \max\{x_1, x_2, x_3\}$$

and

$$\hat{b}_{\text{ML}}(X_1, X_2, X_3) = \max\{X_1, X_2, X_3\}.$$

Question: Does $\hat{b}_{\text{ML}}$ feel like a good estimator of b? Is it what you would have expected?

Answer: Clearly, our estimate for b must be at least equal to the maximum of the samples. But it's not clear that our estimate shouldn't be *higher* than the maximum observed. In fact, if we've only made a few observations, one would expect b to be higher than the highest observation so far.

Question: Is $\hat{b}_{\text{ML}}$ an unbiased estimator?

Answer: This will be explored in Exercise 16.5, where you will show that $\hat{b}_{\text{ML}}$ is not an unbiased estimator, but can be made into one pretty easily.

We now turn to one more example involving continuous random variables.

Example 16.9 (Estimating the standard deviation of temperature)

The high temperature in Pittsburgh in June is (approximately) Normally distributed with a mean of $\mu = 79$ F. Suppose we would like to estimate the standard deviation, σ, of temperature. To do this, we observe the temperature on n randomly sampled independent June days, denoted by $X_1, X_2, \ldots, X_n$. Derive $\hat{\theta}_{\text{ML}}(X_1, X_2, \ldots, X_n)$, the ML estimator of σ.

We will use the log likelihood formulation:

$$\hat{\sigma}_{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \underset{\sigma}{\text{argmax}}\ \ln\left(f_{X_1,X_2,\ldots,X_n|\sigma}(x_1, x_2, \ldots, x_n)\right),$$

where

$$\begin{aligned}
&\ln\left(f_{X_1,\ldots,X_n|\sigma}(x_1,\ldots,x_n)\right)\\
&=\ln\left(\prod_{i=1}^{n} f_{X_i|\sigma}(x_i)\right)\\
&=\sum_{i=1}^{n}\ln\left(f_{X_i|\sigma}(x_i)\right)\\
&=\sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)\\
&=\sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}-\ln\sigma-\ln\sqrt{2\pi}\right)\\
&=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2-n\ln\sigma-n\ln\sqrt{2\pi}. \qquad (16.5)
\end{aligned}$$

To find the maximizing σ, we set the derivative of (16.5) to 0:

$$\begin{aligned}
0&=\frac{d}{d\sigma}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2-n\ln\sigma-n\ln\sqrt{2\pi}\right)\\
&=\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{n}{\sigma}\\
&=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2-n \qquad \text{(multiplying both sides by } \sigma).
\end{aligned}$$

This yields

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}.$$

Hence,

$$\hat{\sigma}_{\text{ML}}(X_1=x_1,X_2=x_2,\ldots,X_n=x_n)=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}},$$

and thus it follows that

$$\hat{\sigma}_{\text{ML}}(X_1,X_2,\ldots,X_n)=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}}. \qquad (16.6)$$

Question: How does $\hat{\sigma}_{\text{ML}}(X_1, X_2, \ldots, X_n)$ in (16.6) compare with $\sqrt{\overline{S^2}}$ from (15.2)?

Answer: These are the same.

16.6 When Estimating More than One Parameter

Sometimes we want to estimate more than one parameter of a distribution. This is done by defining an MLE that jointly optimizes over multiple parameters.

To see how this works, let's return to Example 16.9. Suppose this time we need to estimate both the mean, μ, and the standard deviation, σ, of the Normal distribution of temperature. Again we have n randomly sampled temperatures: $X_1, X_2, \ldots, X_n$. This time, we wish to derive a pair of ML estimators: $\hat{\mu}_{\text{ML}}(X_1, X_2, \ldots, X_n)$ and $\hat{\sigma}_{\text{ML}}(X_1, X_2, \ldots, X_n)$, where

$$\begin{pmatrix} \hat{\mu}\left(X_1 = x_1, \ldots, X_n = x_n\right) \\ \hat{\sigma}\left(X_1 = x_1, \ldots, X_n = x_n\right) \end{pmatrix} = \underset{\mu,\sigma}{\text{argmax}} \ \ln\left(f_{X_1,\ldots,X_n|\mu,\sigma}(x_1, \ldots, x_n)\right).$$

Our likelihood function, $g(\mu, \sigma)$, now depends on two parameters:

$$g(\mu, \sigma) = f_{X_1,X_2,\ldots,X_n|\mu,\sigma}(x_1, x_2, \ldots, x_n).$$

To find the pair (μ, σ) that maximizes $g(\mu, \sigma)$, we set both of the partial derivatives below to 0:

$$\frac{\partial \ln g(\mu, \sigma)}{\partial \mu} = 0 \qquad \text{and} \qquad \frac{\partial \ln g(\mu, \sigma)}{\partial \sigma} = 0.$$

From (16.5), we know that

$$\ln\left(g(\mu, \sigma)\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - n\ln\sigma - n\ln\sqrt{2\pi}.$$

Taking partial derivatives, we have that:

$$\frac{\partial \ln g(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \tag{16.7}$$

$$\frac{\partial \ln g(\mu, \sigma)}{\partial \sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{\sigma}. \tag{16.8}$$

Setting $\frac{\partial \ln g(\mu,\sigma)}{\partial \mu} = 0$ in (16.7) and $\frac{\partial \ln g(\mu,\sigma)}{\partial \sigma} = 0$ in (16.8) yields

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \text{and} \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}.$$

Substituting the expression $\mu = \frac{x_1+x_2+\cdots+x_n}{n}$ into the expression for σ, we get

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2}.$$

Hence we have that

$$\hat{\mu}(X_1 = x_1, \ldots, X_n = x_n) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

and

$$\hat{\sigma}(X_1 = x_1, \ldots, X_n = x_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2}.$$

Since these hold for all values of $x_1, \ldots, x_n$, we have that:

$$\hat{\mu}(X_1, \ldots, X_n) = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

and

$$\hat{\sigma}(X_1, \ldots, X_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{X_1 + X_2 + \cdots + X_n}{n}\right)^2}.$$

16.7 Linear Regression

We now turn to a different kind of estimation optimization problem, which is very common in data analysis. We are given n data points generated through some experiment. We can think of the ith data point as a pair of random variables, (X_i, Y_i) with value $(X_i = x_i, Y_i = y_i)$. We want to find the line that best fits the specific values: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, as shown in Figure 16.3. This is called **linear regression**.

As a concrete example, a company might be trying to understand how advertising is related to revenue. The company has data showing different periods where advertising was lower or higher, and the corresponding revenue during those

periods. The company would like to use this data to create a linear approximation of the relationship between advertising (x value) and revenue (y value).

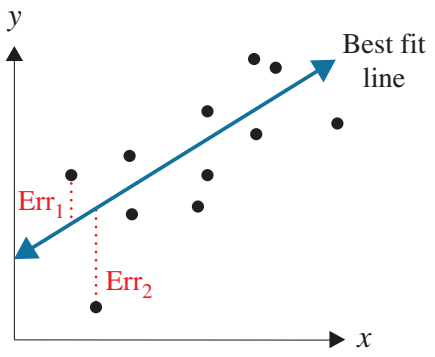

Figure 16.3 *An example of linear regression.*

Recall that a line in the x–y plane is determined by two parameters a and b, where

$$y = ax + b.$$

Our goal is to determine the values of a and b which define a line that best fits our data, where "best" is defined in Definition 16.10.

Definition 16.10 (Linear regression) *Let* $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ *be a set of data sample points. Suppose that* $\hat{a}$ *and* $\hat{b}$ *are estimators for the* a *and* b *parameters of a line fitting the sample points. For the purpose of estimation,* Y_i *is viewed as the* **dependent r.v.** *and* X_i *as the* **independent r.v.** *The* **estimated dependent r.v.** *is* $\hat{Y}_i$*, where*

$$\hat{Y}_i \equiv \hat{a}X_i + \hat{b}.$$

The **point-wise error** *is defined as the difference between the value of the estimated dependent r.v. and the true value for the ith point:*

$$\mathbf{Err}_i = Y_i - \hat{Y}_i.$$

The **sample average squared error (SASE)** *is then:*

$$\mathbf{SASE}\left(\hat{Y}_1, \ldots, \hat{Y}_n\right) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Err}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2. \qquad (16.9)$$

The goal of **linear regression** *is to find estimates* $\hat{a}$ *and* $\hat{b}$ *that minimize* $\mathbf{SASE}\left(\hat{Y}_1, \ldots, \hat{Y}_n\right)$.

Our plan is to derive estimators

$$\hat{a}\left((X_1, Y_1), \ldots (X_n, Y_n)\right) \qquad \text{and} \qquad \hat{b}\left((X_1, Y_1), \ldots (X_n, Y_n)\right),$$

which are functions of the data and which minimize $\mathbf{SASE}\left(\hat{Y}_1, \ldots, \hat{Y}_n\right)$ in (16.9).[1]

Question: What goes wrong if we try to set up $\hat{a}$ and $\hat{b}$ as ML estimators?

Answer: Observe that the likelihood function doesn't make sense here. There is no probability:

$$\mathbf{P}\left\{(X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n) \mid a, b\right\}$$

because once the X_i's are specified and a and b are specified, then the Y_i's are immediately specified.

The point is that we are not trying to maximize a likelihood function, but rather we're finding the $\hat{a}$ and $\hat{b}$ estimators that minimize the **SASE**. Other than that change in objective, however, the optimization setup is very similar to what we do under MLE, which is why we've included the topic in this chapter.

Question: How do we set up the optimization problem, replacing the likelihood function by the **SASE**?

Answer: For a given set of specific points, $(x_1, y_1), \ldots, (x_n, y_n)$, and a given choice of a and b, we define

$$g(a, b) = \mathbf{SASE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2.$$

Then,

$$\begin{pmatrix} \hat{a}\left((X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)\right) \\ \hat{b}\left((X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)\right) \end{pmatrix} = \operatorname*{argmin}_{a,b} g(a, b)$$

$$= \operatorname*{argmin}_{a,b} \left(\frac{1}{n} \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2 \right)$$

$$= \operatorname*{argmin}_{a,b} \left(\sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2 \right). \tag{16.10}$$

Question: How do we find the minimizing (a, b)?

[1] The **SASE** is reminiscent of the **MSE** that we define in Chapters 15 and 17, and in fact many books write **MSE** here. The main difference is that **SASE** is a *sample average* of squares, while **MSE** is an expectation of squares.

Answer: To find the pair (a, b) that minimizes $g(a, b)$, we set both of the partial derivatives below to 0:

$$\frac{\partial g(a,b)}{\partial a} = 0 \qquad \text{and} \qquad \frac{\partial g(a,b)}{\partial b} = 0.$$

We start with finding the minimizing b. By (16.10),

$$\begin{aligned}
0 &= -\sum_{i=1}^{n} \frac{\partial}{\partial b} \left(y_i - (ax_i + b)\right)^2 \\
&= 2\sum_{i=1}^{n} \left(y_i - (ax_i + b)\right) \\
&= \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right) \qquad \text{(divide both sides by 2)} \\
&= \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb.
\end{aligned}$$

Solving for b, we get:

$$\begin{aligned}
b &= \frac{\sum_{i=1}^{n} y_i}{n} - a\frac{\sum_{i=1}^{n} x_i}{n} \\
&= \overline{y} - a\overline{x},
\end{aligned} \tag{16.11}$$

where we define

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \text{and} \qquad \overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

We next find the minimizing a. By (16.10),

$$\begin{aligned}
0 &= -\sum_{i=1}^{n} \frac{\partial}{\partial a} \left(y_i - (ax_i + b)\right)^2 \\
&= 2\sum_{i=1}^{n} \left(y_i - (ax_i + b)\right) \cdot x_i \\
&= \sum_{i=1}^{n} y_i x_i - b\sum_{i=1}^{n} x_i - a\sum_{i=1}^{n} x_i^2. \qquad \text{(divide both sides by 2)}
\end{aligned}$$

To solve for a, it helps to first substitute in our optimizing b from (16.11):

$$\begin{aligned}
0 &= \sum_{i=1}^{n} y_i x_i - (\overline{y} - a\overline{x}) \sum_{i=1}^{n} x_i - a \sum_{i=1}^{n} x_i^2 \\
0 &= \sum_{i=1}^{n} x_i \left(y_i - \overline{y}\right) + \sum_{i=1}^{n} x_i a\overline{x} - a \sum_{i=1}^{n} x_i^2 \\
\sum_{i=1}^{n} x_i \left(y_i - \overline{y}\right) &= a \left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \overline{x} \right).
\end{aligned}$$

Hence,

$$\begin{aligned}
a &= \frac{\sum_{i=1}^{n} x_i \left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} x_i \left(x_i - \overline{x}\right)} \\
&= \frac{\sum_{i=1}^{n} x_i \left(y_i - \overline{y}\right) - \sum_{i=1}^{n} \overline{x} \left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} x_i \left(x_i - \overline{x}\right) - \sum_{i=1}^{n} \overline{x} \left(x_i - \overline{x}\right)} \quad \text{since } \sum_{i=1}^{n} \left(y_i - \overline{y}\right) = 0 = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \\
&= \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}. \qquad (16.12)
\end{aligned}$$

Hence, from (16.11) and (16.12), and substituting in $\hat{a}$ for a in (16.11), we have that

$$\begin{aligned}
\hat{b}\left((x_1, y_1), \ldots, (x_n, y_n)\right) &= \overline{y} - \hat{a}\overline{x} \\
\hat{a}\left((x_1, y_1), \ldots, (x_n, y_n)\right) &= \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}.
\end{aligned}$$

As these estimators are defined for all values of $(x_1, y_1), \ldots, (x_n, y_n)$, it follows that

$$\hat{b}\left((X_1, Y_1), \ldots, (X_n, Y_n)\right) = \overline{Y} - \hat{a}\overline{X} \qquad (16.13)$$

$$\hat{a}\left((X_1, Y_1), \ldots, (X_n, Y_n)\right) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right) \left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}. \qquad (16.14)$$

Using $\hat{a}$ and $\hat{b}$ from (16.13) and (16.14) guarantees our linear fit has minimal **SASE**.

Question: There's a natural interpretation for $\hat{b}$ in (16.13). What is it?

Answer: We can rearrange (16.13) to say

$$\overline{Y} = \hat{a}\overline{X} + \hat{b},$$

which makes perfect sense since we want $Y_i = aX_i + b$, and $\overline{Y}$ is the sample mean of the Y_i's and $\overline{X}$ is the sample mean of the X_i's.

Question: There's also a natural interpretation for $\hat{a}$ in (16.14) if we multiply the numerator and denominator by $\frac{1}{n-1}$. What is it?

Answer:

$$\hat{a}\left((X_1,Y_1),\ldots,(X_n,Y_n)\right) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} = \frac{\mathbf{Cov}(X,Y)}{\mathbf{Var}(X)}. \quad (16.15)$$

Specifically, the denominator of (16.15) is the (unbiased) sample variance of the X_i's, from Definition 15.8, and the numerator is the (unbiased) sample covariance between the X_i's and Y_i's.

Question: What can we say about the sign of $\hat{a}$ based on (16.15)?

Answer: When the covariance is positive, $\hat{a}$ will also be positive, meaning that the slope of the line is positive. This makes sense because it says that X and Y are positively correlated, meaning that when X goes up, Y goes up as well. Likewise, when the covariance is negative, the slope of the line is negative.

When doing regression, the goodness of fit of the line is denoted by a quantity called R^2, where higher R^2 is better.

Definition 16.11 ($\mathbf{R^2}$ goodness of fit) *Consider the set of data sample points* $\{(X_1 = x_1, Y_1 = y_1),\ldots,(X_n = x_n, Y_n = y_n)\}$ *with estimated linear fit:*

$$y = \hat{a}x + \hat{b}. \quad (16.16)$$

Define

$$\hat{y}_i \equiv \hat{a}x_i + \hat{b}$$

to be the estimated dependent value for the ith point. Let

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \text{and} \qquad \overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Then we define the **goodness of fit** *of the line (16.16) by*

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \qquad \text{where } 0 \le R^2 \le 1.$$

The R^2 metric is also called the **coefficient of determination**.

Question: How can we interpret R^2?

Answer: The subtracted term

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\text{sample average squared error}}{\text{sample variance}}$$

can be viewed as the sample average squared error in the estimators normalized by the sample variance of the data set. This term is thus sometimes referred to as "the fraction of unexplained variance." The hope is that this term is a small fraction, which means that R^2 is close to 1.

16.8 Exercises

16.1 **Estimating the bias of a coin**
A coin comes up heads with probability p and tails with probability $1 - p$. We do not know p. We flip the coin 100 times and observe X heads. Derive $\hat{p}_{\text{ML}}(X)$, the ML estimator for p.

16.2 **Battery lifetimes**
We have a bunch of batteries whose lifetimes are i.i.d. $\sim \text{Exp}(\lambda)$. Our goal is to determine λ. To do this, we sample the lifetimes of 10 batteries, whose lifetimes we represent by $X_1, X_2, \ldots, X_{10}$. Derive $\hat{\lambda}_{\text{ML}}(X_1, X_2, \ldots, X_{10})$, the ML estimator for λ.

16.3 **How many balls are blue?**
Suppose that you have a bin with four balls. Each ball is either yellow or blue (you don't know which). Your goal is to estimate the number of blue balls in the bin, which we'll refer to as θ.
To obtain your estimate, you sample three balls with replacement from the bin and note their colors. We let X_i denote the color of the ith ball, where we say that $X_i = 1$ if the ball is blue and $X_i = 0$ otherwise. Let $\hat{\theta}_{\text{ML}}(X_1, X_2, X_3)$ denote the ML estimator for θ.
Suppose we observed the specific sequence of colors: $1, 1, 0$. What is $\hat{\theta}_{\text{ML}}(X_1 = 1, X_2 = 1, X_3 = 0)$?

16.4 **Job CPU requirements follow a Pareto distribution**
After reading Chapter 10, you are well aware that job CPU requirements follow a Pareto(α) distribution. But for which value of α? To answer this question, we sample the CPU requirements of 10 jobs picked independently at random. Let $X_1, X_2, \ldots, X_{10}$ represent the CPU requirements of these jobs. Derive $\hat{\alpha}_{\text{ML}}(X_1, X_2, \ldots, X_{10})$, the ML estimator for α.

16.5 **Estimating the max of a distribution**
In Example 16.8, we saw that the time that students spend on their probability homework is distributed as $\sim \text{Uniform}(0, b)$. To estimate the maximum of this distribution, b, we surveyed three students independently at random, whose times we represented by X_1, X_2, X_3. We then derived the ML estimator $\hat{b}_{\text{ML}}(X_1, X_2, X_3)$ for b, showing that

$$\hat{b}_{\text{ML}}(X_1, X_2, X_3) = \max\{X_1, X_2, X_3\}.$$

(a) Is $\hat{b}_{\text{ML}}(X_1, X_2, X_3)$ an unbiased estimator of b?
(b) To make the estimator more accurate, we decide to generate more data samples. Suppose we sample n students. What is the ML estimator $\hat{b}_{\text{ML}}(X_1, X_2, \ldots, X_n)$? Is it biased when n is large?
(c) Can you think of an estimator $\hat{b}(X_1, \ldots, X_n)$ that is *not* the ML estimator, but is an unbiased estimator for all n? [Hint: You're going to want to scale up $\hat{b}_{\text{ML}}(X_1, \ldots, X_n)$.]

16.6 **Estimating the winning probability**
Team A has probability p of beating team B. We do not know p, but we can see that in the last 10 games played between A and B, team A won seven games and team B won three games. Assume that every game has a unique winner and that games are independent. Based on this information, formulate and compute the ML estimator for p.

16.7 **Disk failure probability estimation**
Suppose that every disk has probability p of failing each year. Assume that disks fail independently of each other. We sample n disks. Let X_i denote the number of years until the ith disk fails. Our goal is to estimate p. Derive $\hat{p}_{\text{ML}}(X_1, X_2, \ldots, X_n)$, the ML estimator for p.

16.8 **Practice with linear regression**
You are given five points: $(0, 5), (1, 3), (2, 1.5), (3.5, 0), (5, -3)$. Determine the best linear fit to these points and compute the R^2 goodness of fit for your estimate.

16.9 Acknowledgment

This chapter was written in collaboration with Weina Wang, who was a major contributor to the chapter contents and the exercises.