Discrete hazard functions and hitting streaks in baseball

Hazard functions

Suppose that a discrete random variable X has a probability mass function $p(x) = P(X = x)$. For simplicity, we will assume that the distribution of X is concentrated on the positive integers $1, 2, 3, \cdots$.

In many situations, we are interested in computing the conditional probability $P(X = x|X \leq x)$. For example, if X represents the year in which a person dies then $P(X = x|X \geq x)$[1] is the probability that the person dies in year x conditional on surviving up to year x. This conditional probability is called the (discrete) **hazard function** and we will denote it by $h(x)$. From the definition of conditional probability, it follows that

$$\begin{aligned} h(x) &= P(X = x|X \geq x) \\ &= \frac{P(X = x, X \geq x)}{P(X \geq x)} \\ &= \frac{P(X = x)}{P(X \geq x)} \\ &= \frac{p(x)}{\sum_{t \geq x} p(t)} \end{aligned}$$

If we are given the hazard function $h(x)$ then we can write the probability mass function as

$$p(x) = h(x) \prod_{t=1}^{x-1} (1 - h(t)) \quad \text{for } x = 1, 2, 3, \cdots$$

where $\prod$ denotes the product and $f(1) = h(1)$.

The shape of the hazard function is often of interest. A hazard function $h(x)$ that is increasing as x increases suggests that the random variable X improves with age while the opposite is true if $h(x)$ decreases as x increases. (Alternatively, a decreasing hazard function is often associated with "heavy tails" where more extreme data are observed whereas an increasing hazard function leads to lighter tails with few extreme observations.) When the hazard function is constant $h(x) = \theta$ for all x then the random variable X has a **geometric distribution** with probability mass function

$$p(x) = \theta(1 - \theta)^{x-1} \quad \text{for } x = 1, 2, 3, \cdots.$$

(The notion of a hazard function also applies to continuous distributions whose probability mass is concentrated on the positive real line. If X is a continuous random variable with cumulative distribution function F and density function f then we define its hazard function by

$$h(x) = \lim_{\delta \downarrow 0} \frac{1}{\delta} P(x \leq X \leq x + \delta | X \geq x) = \frac{f(x)}{1 - F(x)}.$$

[1] This conditional probability would be of interest to a life insurance company.

Given the hazard function $h(x)$, we can define the cumulative distribution function and density function by

$$\begin{aligned} F(x) &= 1 - \exp\left(-\int_0^x h(t)\,dt\right) \\ \text{and } f(x) &= h(x)\exp\left(-\int_0^x h(t)\,dt\right), \end{aligned}$$

respectively.)

Suppose that we have data $x_1, x_2, \cdots, x_n$ that we assume comes from a probability distribution with hazard function $h(x)$. Then a simple estimate of the hazard function is given by

$$\begin{aligned} \widehat{h}(x) &= \frac{\sum_{i=1}^{n} I(x_i = x)}{\sum_{i=1}^{n} I(x_i \geq x)} \\ &= \frac{\text{number of observations equal to } x}{\text{number of observations greater than or equal to } x} \end{aligned}$$

This is a good estimate of $h(x)$ provided that the denominator is reasonably large.

Application to hitting streaks

A number of people have looked at the probability distribution of hitting streaks in Major League Baseball (MLB). A simple model for this distribution is the geometric distribution (whose hazard function is constant), which can be justified if we assume that

(a) the probability that a player goes without a hit in a game is θ;

(b) this probability remains constant independently of the outcomes of previous games.

If (a) and (b) hold then the hazard function is $h(x) = \theta$. These assumptions are, of course, not terribly realistic. First, the probability of not getting a hit in a game varies from player to player. Second, the value of θ for a given player could vary considerably over the course of a season for a number of reasons.

Table 1 contains data from McCotter (2009)[2] on hitting streaks in MLB from 1957 to 2006. McCotter suggests that the number of long hitting streaks during this period exceeds what would occur simply by chance.

Figure 1 shows the estimated hazard function $\widehat{h}(x)$ for $x = 5, 6, \cdots, 34$ based on the simple estimate defined in the previous section along with a smooth estimate of the hazard function computed using a more sophisticated method. For the latter method, the estimates of the hazard function range from 0.35 for $x = 5$ to 0.20 for $x = 34$. (The estimate is much more reliable for smaller values of x than it is for larger values of x.) One explanation for the decrease is the fact that it is relatively easy for average hitters to have short hitting streaks while longer hitting streaks are generally attainable only by more accomplished hitters.

[2]McCotter, T. (2009) Hitting streaks don't obey your rules: Evidence that hitting streaks aren't just by-products of random variations. *The Baseball Research Journal.* **37**, 62-70.

Length	Frequency	Length	Frequency	Length	Frequency	Length	Frequency
5	22632	13	801	21	52	29	4
6	14470	14	552	22	38	30	9
7	9151	15	415	23	25	31	4
8	6081	16	270	24	22	32	0
9	4059	17	194	25	17	33	0
10	2645	18	129	26	8	34	1
11	1792	19	112	27	7	35+	5
12	1226	20	75	28	7		

Table 1: Frequency of hitting streaks from 1957-2006.

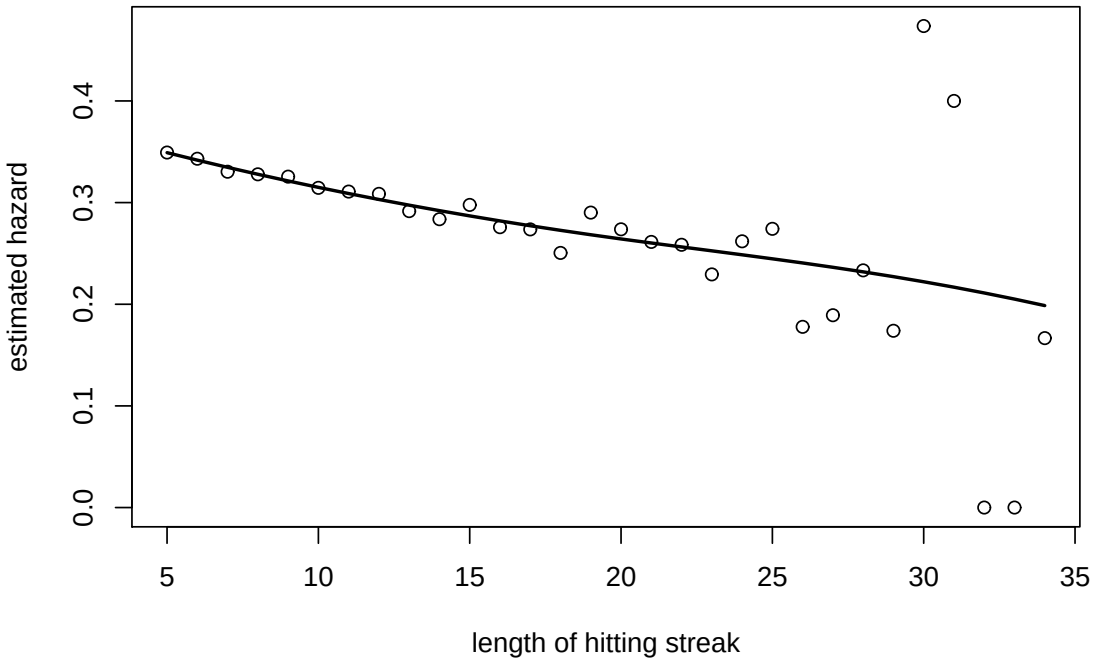

Figure 1: Estimates of the hazard function based on the data in Table 1.