Article can be accessed online at http://www.publishingindia.com

Analytically Yours

Analyzing Data Which are Curves - Functional Data Analysis

Arnab Kumar Laha*

In many areas of application, the data comes in the form of a curve or in other words a function. Consider a hotel having a fixed number of rooms. Suppose the hotel starts accepting booking for one year in advance. The information about the number of rooms booked for a particular date (say for 1st April, 2016) would be available for every day in the period 1st April, 2015 to 31st March, 2016. If the number of rooms booked 't' days prior to 1st April 2016 is denoted by b(t) then the curve $\{b(t) : 0 \leq t \leq 365\}$ is called the booking curve for 1st April 2016. Similar curves would be available for the number of rooms booked for 2nd April, 3rd April etc. Thus we would have the data in the form of curves $\{b_1, b_2, ..., b_n\}$. Statistical analysis of this kind of data is referred to as functional data analysis. One of the simplest questions to ask with booking curves data is about the shape of the average booking curve. The answer to this is simple: The mean booking curve is $\bar{b}(t) = \frac{1}{n}\sum_{i=1}^{n} b_i(t)$, $0 \leq t \leq 365$, which in other words is the point wise average of the booking curves. A careful reader by now would be wondering whether at all it is possible to observe $b_i(t)$ for every t. While theoretically it is possible, but in most practical situations $b_i(t)$ would only be observed at some values of t for example for t = 0, 1, 2, ..., 365. Then how does one get to the values of $b_i(t)$ for other values of t? The answer of course is through interpolation. One of the simplest approach is to do piecewise linear interpolation. In this method we simply connect the observed points by drawing straight lines between adjacent points. Technically, if one has observed $b_i(u)$ and $b_i(v)$ where u < v, then the value $b_i(t)$ for an intermediate point t between u and v is $b_i(t) = b_i(u) + \frac{b_i(v) - b_i(u)}{v - u}(t - u)$. Linear interpolation is easy to carry out and it gives a curve which is continuous. However, this curve may not be differentiable at the observed points. This leads to some analytic difficulties which prompts search for more complicated interpolation techniques. An elegant solution is obtained if one uses cubic splines for interpolating instead of the straight lines. In cubic spline interpolation a polynomial of degree three is used instead of the straight lines. This results in a curve which is twice differentiable and has many good properties. Thus cubic spline interpolation is often used for "reconstructing" the functions b_i from the observed values.

In Figure 1 below we give plots of 34 booking position curves for a given class of a long-distance train plying between two major cities of India. Each booking position curve gives the booking position for a train at a certain fixed time of the day starting from 60 days prior to the date of departure to one day prior to the day of departure. The negative booking position indicates that no reserved seat tickets are available and only Reservation Against Cancellation (RAC) or Waitlisted tickets are available.

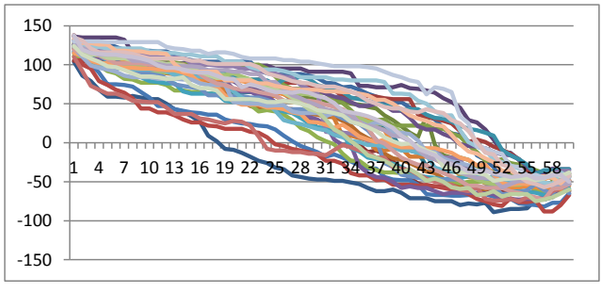


Fig. 1: The Booking Position Curves

The mean curve $\bar{b}(t)$ is calculated by taking the pointwise average of all the 34 curves. By pointwise average we mean that for every value of t we compute

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

$\bar{b}(t) = \frac{1}{34}\sum_{i=1}^{34} b_i(t)$. Now we can also compute SD(t), the standard deviation at each point t. In Figure 2 below we give the plot of the mean curve (denoted as AVG) along with curves AVG – 2 SD and AVG + 2 SD. If the booking position at time t is assumed to follow a normal distribution then approximately 95% of the booking positions on day t will lie within the interval (AVG(t) – 2 SD(t) , AVG + 2 SD(t)). An examination of Figure 2 indicates that if a booking is sought to be done up to five weeks before the day of departure of the train there is a high chance of obtaining an instant reservation whereas if the booking is done within the last week before the day of departure of the train there is very little chance of obtaining an instant reservation. Figure 3 below gives the probability of booking position being positive as a function of the number of days since the beginning of the booking.

In Figure 4 below we examine the standard deviation of the booking position as a function of the number of days since the beginning of booking. We find that the standard deviation is not constant over time. In fact the standard deviation of the booking position in the fifth and sixth weeks after the beginning of the booking is much higher as compared to that in the first or last week. Thus any modelling of this data needs to take care of this fact.

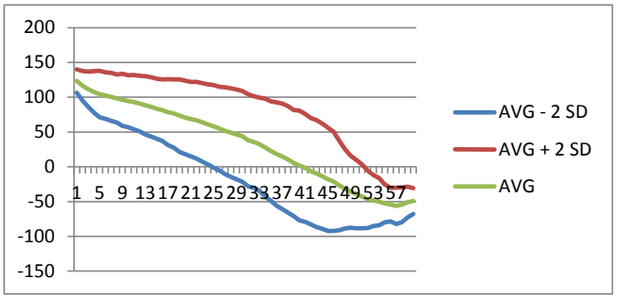


Fig. 2: The Average Booking Position Curve Along With the Curves Signifying the +/- 2 SD Deviations from the Same

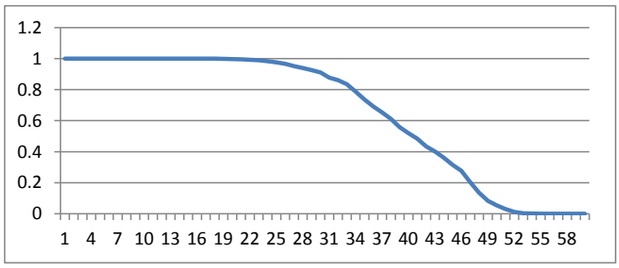


Fig. 3 : The Plot of Probability(Booking Position > 0) as a Function of Number of Days Since Opening of Booking

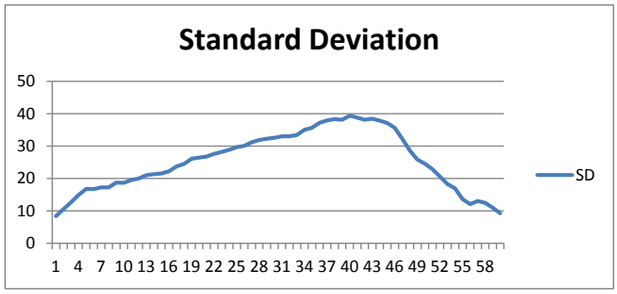


Fig. 4: Variation in the Standard Deviation of the Booking Positions over Days since Booking Opens

One of the key features of functional data is that observations at time points which are very close are often strongly correlated. This is expected because of the tangent line approximation which states that for small values of h, $f(t+h) \approx f(t) + hf'(t)$ where $f'(t)$ denotes the derivative of f at time t. Because of the presence of near perfect correlation between nearby time points the variance-covariance matrix, which is a key tool for describing variation in the multivariate context, often turns out to be ill-conditioned (near-singular). The Figure 5 below gives the correlation between the booking position on day t and the same on day t +1. It can be seen from the figure that for most values of t the correlation is very strong (greater than 0.95)

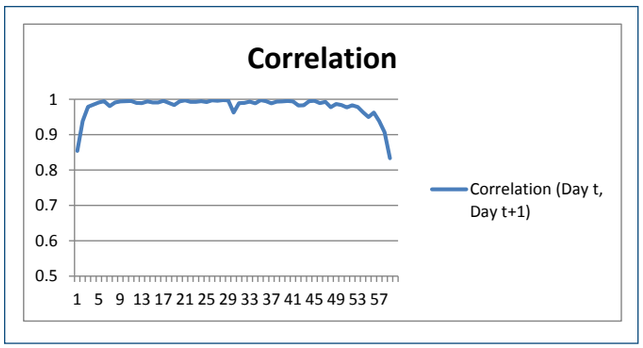
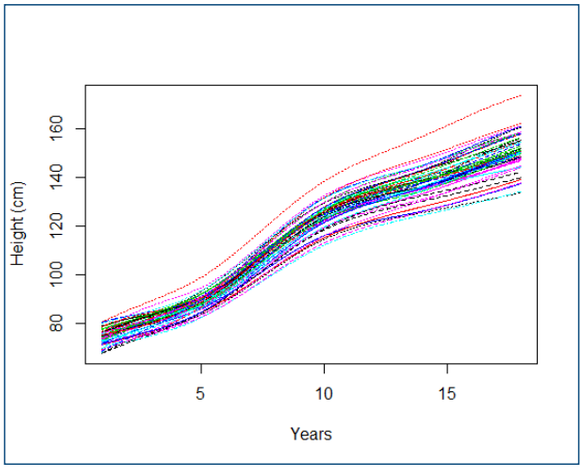
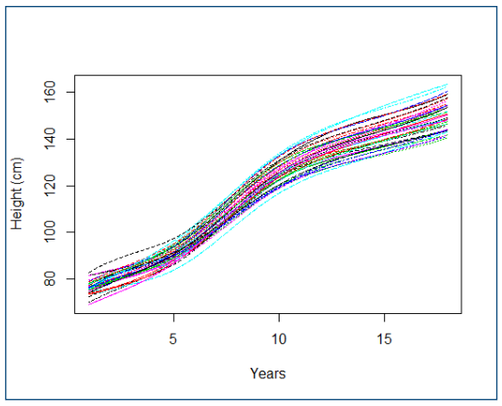


Fig. 5 : The variation in the correlation between the booking positions on day t and day (t+1) as a function of number of days since booking opening.

As a second example let us consider the data from the Berkeley Growth Study. In this study the heights of 54 girls and 39 boys were recorded at 31 different time points between the ages of 1 and 18 years (Tuddenham and Snyder, 1954). Figures 6(a) and 6(b) give the smoothed curves obtained using this data. One of the natural questions that comes to our mind is whether there is difference in the growth pattern in the two groups of male and female children.



(a)



(b)

Fig. 6: (a) Growth curves of the females (b) Growth curves of the males

Figure 7 below gives the plot of the average height of the male children (**average_m**) and female children (**average_f**). From the plot we see that the boys and girls have similar heights up to age 13 and from then onwards the boys become taller than the girls on average. Figure 9 gives the standard deviation of the heights of boys and girls at different ages. It is interesting to see that the variation in the heights of the girls exceeds those of the boys till about age 13 but after that the variation in the heights of boys becomes larger. After age 16 the variation in the heights of the boys and girls become almost equal.

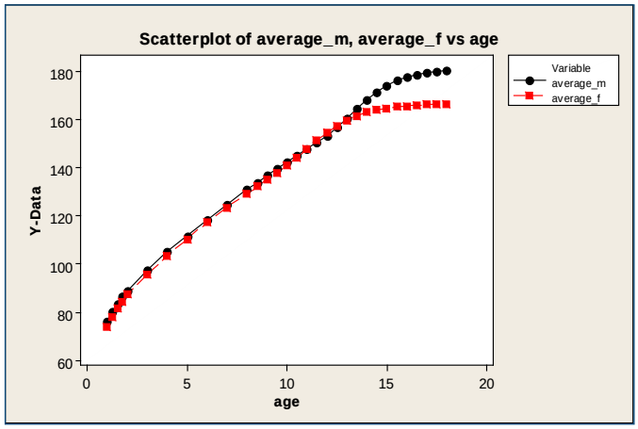


Fig. 8: Average Height of Boys and Girls at Different Ages in the Berkeley Growth Study

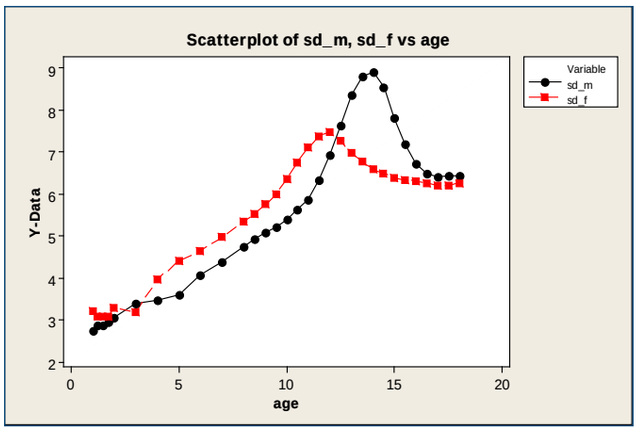


Fig. 9: The Standard Deviation in the Heights of the Boys and Girls at Different Ages

Readers desirous of learning more about Functional Data Analysis from an applied perspective may look at the book by Ramsey and Silverman (2002).

Acknowledgement: Ms. Poonam Rathi, Indian Institute of Management Ahmedabad collected the booking position data which has been used in this article.

References

Ramsey, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and Case studies.* Springer.

Tuddenham, R. D., & Snyder, M. M. (1954). Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development,* 1, pp.183–364.