

Student Projects and the London Olympics in 2012: Teaching and Learning through Modelling Sporting Performances

Introduction

In this note statistical approaches to modelling sporting performances are discussed. Final year projects based on such modelling have proved popular in the past. With the upsurge in interest in the next Olympic Games in London in 2012, such modelling could provide teachers of statistics with a rich source of meaningful student projects. Students may be more motivated to do such projects simply because London is hosting the Games.

Previous analyses of world record times in a number of sports, including male/female performance comparisons, provide excellent examples for students to get their hands dirty with real data and learn about pitfalls in applying routine modelling procedures. The Games comprise 28 sports, and within athletics alone there are 47 events for men and women. Teachers of statistics could find it very beneficial to think about the very wide range of projects possible by jumping on the Olympics bandwagon!

The literature on statistical aspects of modelling sporting performances is widespread, though much of it is not in the statistical literature. To keep discussion within bounds the present note focuses mainly on modelling athletic performances, using world records or data from the Olympic Games. This is often undertaken with a view to predicting the ultimate limits of human performance. A good review of the literature, to about 1996, is Schutz and Liu (1998); more up-to-date with a more comprehensive discussion of possible models is Kuper and Sterken (2008).

My interest in such modelling dates back about 10 years, when I started setting final year undergraduate projects on the topic. Typically the project starts off with a title like 'Modelling Athletic Performance' and is undertaken by students doing a Mathematics or joint Mathematics and Sport degree. The students are required to collect data, or augment existing data, on world record times, and the dates they were set, in athletics events from the 100 m to the marathon. World records or Olympic times can be used or compared. Field events and swimming data lend themselves to similar exercises to the ones described below.

It is possible to run these as 'straight' statistics projects that end up as exercises in non-linear model fitting that goes beyond what students meet in taught classes. They fit models of increasing complexity to the data, from linear to non-linear, with the aim of seeing if realistic predictions of the limits of performance can be obtained. For some events the answer is no, and a critical evaluation of the reasons is expected (see below). Often students become interested in non-statistical aspects of the problem; for example comparing male and female performances, or looking at the physiological reasons that govern performance, and the final product and title of the project will reflect this.

To stimulate new ideas for similar projects, I spent much of the recent summer break on a thorough (re-)reading of the relevant literature, and have re-analysed data used by those who have published on the topic. The following notes have been stimulated by this. Colleagues who teach statistics, in search of possible project topics, may find interest in the London Olympics as a useful stimulant to ideas.

On linearity and non-linearity

To support later discussion Figure 1 will be used. It shows the world records in the male and female marathons up to 1980, and thereafter year-on-year best times to 2005. (As I write this the men's world record has recently been broken, raising speculation about if and when the two-hour barrier will ever be broken.)

[Figure 1 about here]

A surprising amount of time has been devoted to fitting linear models of the form $y = \alpha + \beta x$ to this kind of data, and discussing the outcome. Here y is running time and x the date of performance. It is obvious from a theoretical point of view, and visually in Figure 1, that linear models are inappropriate. They lead to unbounded estimates of performances with a negative time. Exponential or power-law models, linearised as $\ln(y) = \alpha + \beta x$ or $\ln(y) = \alpha + \beta \ln(x)$, are better, but still inappropriate since they predict zero as the ultimate limit of performance. The simplest, and most widely used predictive model is exponential of the form $y = \alpha + \beta \exp(-\gamma x)$ that, with appropriate signs for the parameters, has a lower asymptotic limit of α . Some researchers (Deakin 1967, Neville and Whyte 2005, Kuper and Sterken 2008) have preferred four-parameter models such as logistic or Gompertz models which have an upper asymptote as well as a lower one.

There is no disputing the *theoretical* superiority of non-linear models with an asymptote, but the *practicalities* have been somewhat different. A notorious example is Chatterjee and Chatterjee's (1982) modelling of Olympic winning times from 1900 to 1976 for events from the 100 m to 800 m. This attracted considerable adverse comment in letters to the journal *Applied Statistics* in 1983 and 1984. Essentially others were unable to reproduce Chatterjee and Chatterjee's (1982) results. One problem was that the algorithm they used to fit the exponential model did not converge, and there was general agreement that the data available at the time did not warrant anything other than a linear model. McKeown and Sprevak (1992) present technical details of the issues involved.

To illustrate, Figure 2 shows the data for the 200 m used by Chatterjee and Chatterjee (1982) with the fitted lines from linear and quadratic regressions superimposed. They are virtually indistinguishable, and the parameter estimate for the quadratic term in the quadratic model is, unsurprisingly, not significant at the levels usually used. As McKeown and Sprevak (1992) and others have noted the quadratic model is an approximation to the exponential model, and if the quadratic term is not significant it suggests there is no point in trying to fit a non-linear model to the data.

[Figure 2 about here]

In non-technical terms, there is not much point in trying to predict asymptotes unless there is some evidence of 'bottoming-out' in performances. To some extent this has occurred for some events since 1976, but the problem has not gone away. Scully (2000), for example, models men's Olympic times from 1896 to 1996 for a variety of events. He tests the quadratic term in the quadratic model, finds that for most events it is significant, and concludes that fitting truly non-linear models is justified. He fails to comment on the obvious fact that the 1896 results are rather untypical and that if these are omitted there is then no evidence of non-linearity for the data for several events. The 1896 data are (silently) omitted from Chatterjee and Chatterjee's (1982) analysis.

From a modelling point of view, and issues concerning linearity, the marathon data in Figure 1 is interesting. The patterns for both males and females are obviously non-linear. However female participation in this event, as recognised by the recording of records, arrived late and from the early 1960s to about 1980 the progression in world records was rapid in comparison with what was happening with males, and approximately linear.

The patterns in data of these kind led to papers that suggested that female performances might eventually match or overtake those of males, of which Whipp and Wards' (1992) paper in *Nature*, provocatively entitled 'Will women soon outrun men?', is among the more notorious. They modelled average running velocity (rather than time) against date of performance and suggested, among other things that, on the basis of extrapolation from the linear patterns observed, 'the projected intersection for the marathon [for male and female performance] is 1998'. This and similar claims have attracted a surprising amount of comment (e.g., Neville and Whyte, 2005) since some researchers seem to feel obliged to pay attention to the topic.

My view is that it is, and always has been, nonsense to compare potential male and female performances on the basis of linear extrapolation, even if the data at any particular time exhibit linear trends. This is not to argue that females will never outperform males (although there are strong non-statistical arguments for making this case). It is simply that linear extrapolation is inappropriate, Notwithstanding the efforts of Paula Radcliffe, Figure1 suggests that world record performances are 'bottoming-out' for both sexes with a fairly consistent difference between males and females, the latter being slower.

The male data in Figure 1 exhibits other features of interest that I have not seen commented on often. Usually, modelling exercises assume a 'smooth' model, whether linear or non-linear. Basic data inspection suggest that a smooth model may often be inappropriate, Anybody doubting this should simply collect and plot world record data for events from the 100 m to the marathon. Most do not exhibit conveniently smooth behaviour. The male marathon data, which is actually 'better behaved' than

that for other distances, is a case in point. To my eye there are at least two ‘regimes’ evident – pre- and post- World War 2 – and possibly two regimes after World War 2.

This suggests that rather than fitting a single smooth model to the data some form of piecewise, change-point model might be of interest. Berthelot *et al.* (2008) is the only paper I know of that attempts something of the sort. Such models may have limited value for prediction since the relevant, later, part of the data may consist of very few observations. This suggests that weighted least squares analysis (least squares rather than maximum likelihood being most often used), giving more weight to later records, might be of interest, but I am unaware of any such attempts.

Other issues

The issue of data selection has been touched upon briefly above. The merit of using world record as opposed to Olympic data has been debated in the literature, as has the possibility of using year-on-year best times, or non-extreme results such as eighth best times (e.g., Grubb 1998). I have little comment to make here except to suggest that a comparative study of the results obtained from different treatments might be of interest.

In terms of events selected for study papers range from using all events from the 100 m to the marathon, to papers that use only a small subset of events, often without any explanation of why these were chosen. The constraints of journal publication probably preclude detailed and published inspection of all available data of the kind afforded in a student project; what can be seen in the latter is that many events do not lend themselves to ‘simple’ non-linear model fitting. I suspect that many papers that try to model only a subset of events select only those suited to their particular purpose. That is, the lack of generality of the models presented is silently passed over. Similarly, some papers silently pass over the fact that earlier times are not included in the analysis. This is often perfectly reasonable since earlier times are often obviously untypical, but explicit comment on the reasons for omission would be welcome.

Finally, how good are the models that have been put forward at predicting limits to human performance? Sometimes the question is posed as ‘are there limits to human performance?’. This doesn’t make sense as there obviously are such limits; it is more pertinent to ask if these can usefully be predicted given current data. For some events the answer has to be no, as there is not sufficient evidence of a ‘bottoming-out’ in performance to allow reliable predictions with non-linear models.

Where there is evidence of ‘bottoming-out’ I have sometimes suspected that you can get as good a prediction as any statistical model by estimating an asymptote ‘by eye’ and subtracting a small amount from the result. This is not as silly as it sounds. Many formal models have resulted in predictions that are worse than the current world record. *Ad-hoc* procedures, amounting to the arbitrary introduction of an extra ‘correction’ parameter, often based on residuals from the fitted model, have been used in several papers. Even so this does not necessarily result in a prediction better than

the current world record. For example, about half of Scully's (2000) analyses of 11 events, with such an adjustment, gave times worse than the current record.

Conclusion

Recent reading has led me to be quite sceptical about some of the effort that has gone into modelling limits of human performance from athletics records. I have, for the most part, refrained from identifying papers where some of the analysis is wrong, or where I think conclusions are misguided.

Nevertheless I confess to finding such analysis fun. At the very least the data are useful for learning about non-linear model fitting and the attendant pitfalls. It has been argued by others that what are essentially non-linear curve-fitting exercises, with a view to prediction, ignore the more fundamental physiological and other factors that govern performance. This is true, but it can also be argued that such factors generate the data that we try to model, and that curve-fitting attempts to capture such patterns as exist in the data, even if not explicitly modelling the causes.

More interestingly, and in the context of student projects, a lot of problems beyond the mechanics of fitting models to data, that any practising statistician needs to confront, are raised. Data and model selection and data inspection have been mentioned above. The issues concerning the usefulness or otherwise of linear and non-linear models are both interesting and subtle, and difficult to convey in a taught lecture course. We know, in theory, that a non-linear model is needed, but the data may either not justify anything that complicated or may require something much more complicated than the models we want to use. There may simply not be enough data for our purposes. Applied statisticians know this; a lot of statistics students need to appreciate the problems that can arise, and trying to model sporting performances can alert them to the complexity of 'real-life' data analysis.

Teaching and Learning Note

Colleagues within and outside the UK who read this article and are interested in pooling and sharing ideas for projects related to any of the Olympic sports may wish to contribute to a blog area of the Maths, Stats and OR web site, the precise address of which will be advertised later, once it is established.

References

Berthelot, G., Thibault, V., Tafflet, M., Escolano, S., El Helou, N., Jouven, X., Jermine, O., and Toussaint, J-F., 2008, The Citius end: world records progression announces the completion of a brief ultra-physiological quest, *PLoS ONE*, **3**, e1552.[doi:10.1371/journal.pone0001552](https://doi.org/10.1371/journal.pone0001552)

Chatterjee, S., and Chatterjee, S., 1982, New lamps for old: an exploratory analysis of running times in Olympic Games, *Applied Statistics*, **31**, 14-22.

Deakin, M.A.B., 1967, Estimating bounds on athletic performance, *Mathematical Gazette*, **51**, 100-103.

Grubb, H.J., 1998, Models for comparing athletic performances, *The Statistician*, **47**, 509-521.

Kuper, G.H., and Sterken, E., 2008, Modelling the development of world records in running, in *Statistical thinking in sports* (eds. J. Albert and R.H. Koning), 7-31, Chapman and Hall/CRC, Boca Raton.

McKeown, J.J., and Sprevak, D., 1992, Parameter estimation versus curve fitting: 'New lamps for old', *The Statistician*, **41**, 357-361.

Nevill, A.M., and Whyte, G., 2005, Are there limits to running world records?, *Medicine and Science in Sports and Exercise*, **37**, 1785-1788.

Schutz, R.W. and Liu, Y., 1998, Statistical modelling in track and field, in *Statistics in sport* (ed. J. Bennett), 173-196, Arnold, London.

Scully, G.W., 2000, Diminishing returns and the limit of athletic performance, *Scottish Journal of Political Economy*, **47**, 456-470.

Whipp, B.J., and Ward, S.A., 1992, Will women soon outrun men?, *Nature*, **355**, 25.

Figure Captions

Figure 1: Male and female performances in the marathon over time. Date is given as years since 1900; time is in seconds over two hours. Up to 1979 only world records are shown; thereafter the best annual times up to 2005 are shown. One early record for females is omitted.

Figure 2: Olympic winning times from 1900 to 1976 with the almost indistinguishable linear (solid line) and quadratic (dashed line) regression fits superimposed. Data are from Table 1 of Chatterjee and Chatterjee (1982).

running time (s)

20.0

20.5

21.0

21.5

22.0

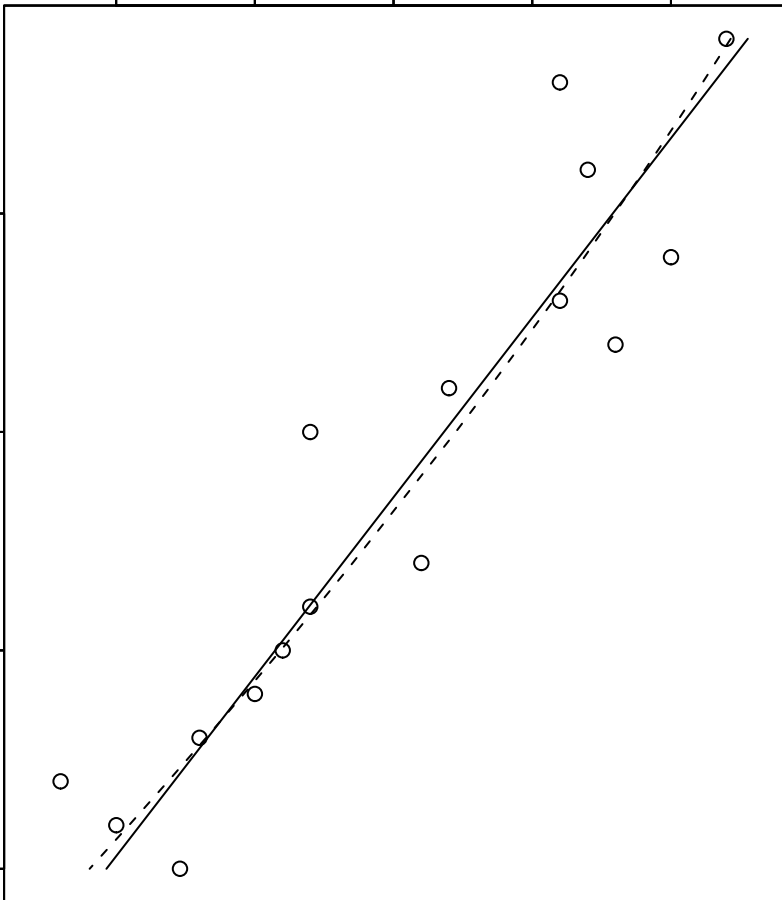
(year - 1896)/4

5

10

15

20



Female ○ Male +

