Appl. Statist. (1982),
31, *No.* 1, *pp.* 14–22

New Lamps for Old: An Exploratory Analysis of Running Times in Olympic Games

By SAMPRIT CHATTERJEE and SANGIT CHATTERJEE

New York University, New York *Northeastern University, Boston*

[Received June 1980. Final revision April 1981]

SUMMARY

In this paper, we analyse the winning times in the men's 100, 200, 400 and 800 metres run in the Olympic Games from 1900 to 1976. A non-linear model is fitted to each of the series, and an estimate of the best attainable time for each of these events is obtained. The entire data are then analysed as two-way classified data, to examine the effect on running times of distance of the race, and the year in which the Games were held. The "year effects" will estimate improvements in techniques and training methods over the years. The effect of the altitude of the meet on reported times is also studied. Two methods of analysis—mean and median fitting— are presented for comparative purposes. The times for these events in the 1980 Olympics are predicted using the fitted models and are compared with the observed times. A set of predictions for the 1984 Olympics is provided for future comparison.

Keywords: OLYMPIC RUNNING TIMES; MEDIAN FITTING; JACK-KNIFE; ASYMPTOTES; PREDICTION

0. INTRODUCTION

THIS paper analyses the winning times reported for men's track events in the Olympic Games from the period 1900 (Paris) to 1976 (Montreal). The analysis is divided into two main parts. In the first part, we look at individual events and attempt to find a model which best describes the observed times. From the data we try to estimate the best attainable time for a given event. Interesting conclusions are reached by examining the answers to these questions for various related but independent events (100, 200, 400 and 800 metres).

In the second part, we look at these four events simultaneously in the context of two-way classified data—years and distance. We examine the variation in year and distance effects and model them separately. Since the games are held at places of varying altitudes, we examine the effect of altitude as a covariate. Conventional linear models and a more robust method (Tukey, 1977) of analysis are performed and a comparative study is made. The fitted models are then used to predict the times for future Olympics. The performance of the model for 1980 is discussed, and predictions for 1984 are made. Several different predictions are made, one set based on individual series, and the other set based on all four series, leading to simultaneous predictions of the winning time for these four events.

1. ANALYSIS OF INDIVIDUAL EVENTS

1.1. *Selection of a Model*

The first step in the analysis was to construct a model for the observed times. An examination of the data shows that for these four events: (i) the times are improving (decreasing), (ii) the rate of improvement is decreasing, and from physiological consideration it is clear that (iii) there is a lower limit for the times of these events. If we let the time t_j be a function of the year j, then these considerations effectively rule out the linear model $t_j = \alpha + \beta j$ and the simple exponential models $t_j = \theta_1 \exp(j\theta_2)$. A simple model satisfying the above

0035–9254/82/31014 $2.00

physical considerations is

$$t_j = \theta_1 + \theta_2 \exp(j\theta_3) + \varepsilon_j, \quad \theta_2 > 0, \quad \theta_3 < 0, \tag{1}$$

where ε_j is a random disturbance. The best attainable time for an event is θ_1 ($= t_\infty$).

The above model is non-linear and was fitted in the study by a non-linear model fitting algorithm. The standard errors of the estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ were obtained by jack-knife (see Miller, 1974 and Tukey, 1977). The jack-knife estimates will be denoted by $\hat{\theta}_{1J}, \hat{\theta}_{2J}, \hat{\theta}_{3J}$, and the jack-knifed standard errors by S_{1J}, S_{2J}, S_{3J}, respectively. The jack-knife procedure not only provides us with estimates of standard errors, but also gives us an idea about the stability of estimates, and the sensitivity of the model to individual observations.

1.2. *Numerical Results and Analysis*

The data for the analysis are given in Table 1, and the results of the analysis are given in Table 2. The estimated values of $\theta_1, \theta_2, \theta_3$ for each of the four fitted models are given, as are the values of R^2 (the square of the multiple correlation coefficient).

TABLE 1
Winning times in men's running events in Olympic Games (1900–76) and altitude

Year	100 *metres*	200 *metres*	400 *metres*	800 *metres*	*Altitude (feet)*
1900	10·80	22·20	49·40	121·40	25
1904	11·00	21·60	49·20	116·00	455
1908	10·80	22·60	50·00	112·80	8
1912	10·80	21·70	48·20	111·90	46
1920	10·80	22·00	49·60	113·40	3
1924	10·60	21·60	47·60	112·40	25
1928	10·80	21·80	47·80	111·80	8
1932	10·30	21·20	46·20	109·80	340
1936	10·30	20·70	46·50	112·90	115
1948	10·30	21·10	46·20	109·20	8
1952	10·40	20·70	45·90	109·20	25
1956	10·50	20·60	46·70	107·70	3
1960	10·20	20·50	44·90	106·30	66
1964	10·00	20·30	45·10	105·10	45
1968	9·90	19·80	43·80	104·30	7349
1972	10·14	20·00	44·66	105·90	1699
1976	10·06	20·23	44·26	103·50	104

TABLE 2
Fitted model parameters, predicted values, jack-knifed values and standard errors

	100	200	400	800
$\hat{\theta}_1$	9·103	17·126	38·975	97·016
$\hat{\theta}_2$	1·935	5·405	11·430	21·871
$\hat{\theta}_3$	−0·0359	−0·0304	−0·0376	−0·0518
R^2	0·81	0·85	0·88	0·86
$\hat{\theta}_{1J}$	9·111	17·135	39·009	96·989
$\hat{\theta}_{2J}$	1·930	5·399	11·406	21·933
$\hat{\theta}_{3J}$	−0·0367	−0·0307	−0·0380	−0·0532
S_{1J}	0·039	0·073	0·154	0·593
S_{2J}	0·036	0·070	0·152	0·483
S_{3J}	0·0013	0·00070	0·00070	0·0022

The jack-knife values $\hat{\theta}_{1J}, \hat{\theta}_{2J}, \hat{\theta}_{3J}$ and their standard deviations S_{1J}, S_{2J}, S_{3J} are also given. Since each of the four events are analysed similarly, for brevity we present in detail only the analysis of the times of 200 metres run. The times for the event are plotted against the years in Fig. 1. Fig. 2 displays the residuals against the years. They all lie within ± 2 standard deviation, and appear to be randomly distributed. The fit is good, the value of R^2 is 0·85. The estimate of the best attainable time for the event by the jack-knife method is 17·135 with an estimated standard error of 0·073.

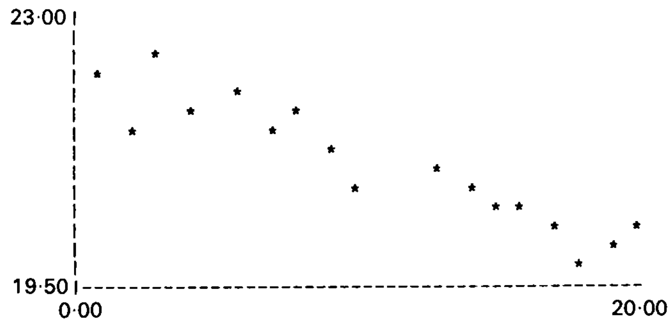

FIG. 1. Winning times for 200 metres plotted against year.

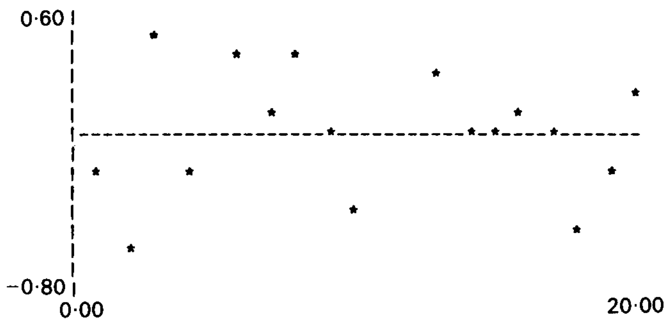

FIG. 2. Plot of residuals against year.

Before going on to discuss the results of the other events, we sidestep to look at a question which has often been raised with respect to athletic performances. Does altitude affect the times reported for these four events? While we cannot answer this question in general, we shall examine this question in detail for the four events under consideration. Of the 17 sites where the Olympic Games have been held, only Mexico City (7349 feet) and Munich (3493 feet) have elevations appreciably different from sea level. To see whether elevation affects the times, we plot the residuals from the fitted model against elevation. This plot of residuals for 200 metres is shown in Fig. 3. Except for points marked 1 and 2, the residuals are distributed randomly about altitude (another indication of the good fit of the model). The two points for which the altitude is appreciably different from sea level have negative residuals. This is in sharp contrast to the common folklore. We consider these residuals as just random, rather than concluding that running times get better with altitude. The same phenomenon is observed for each of our other three events. Thus, we conclude that for altitudes up to about 8000 feet and for these four running events, performance is not affected for top quality Olympic athletes, or the effects can be eliminated by training and conditioning. The effect of altitude on longer events is still an open question.

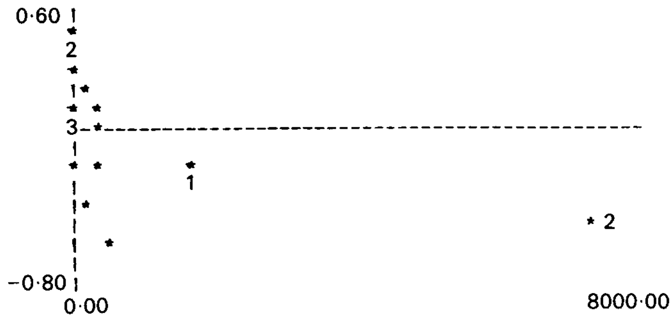

FIG. 3. Plot of residuals against altitude.

From examining Table 2, some general conclusions emerge. The estimate of the best attainable time for 200 metres is less than twice the best estimated time for 100 metres, indicating that there is a considerable start-up time for a sprint event, and once this speed is attained, it is maintained for at least 200 metres. A different picture emerges for 400 and 800 metres (the best time for 400 metres is greater than twice the best time for 200 metres, and the same holds for 800 and 400 metres).

This conclusion agrees with the results of Keller (1974) who from physiological and dynamical considerations arrived at the conclusion that races less than 290 metres are run as dashes. When we examine the estimated value of θ_3 across events, we find that the improvement in times has been more rapid in the longer events (the estimated value of θ_3 is monotonic with distance).

2. Two-way Classified Data

2.1. *Choice of Model*

In this section, we analyse all the four events together and try to isolate in the reported times the component which may be attributed to the year in which the meet was held, and the component which may be attributed to the distance for the event. The data fall in the classical format of two-way classified data with one observation in each cell. We start initially with the traditional additive linear model where

$$t_{ij} = \mu + Y_i + D_j + \varepsilon_{ij} \quad i = 1, 2, ..., 17, \quad j = 1, 2, 3, 4, \tag{8}$$

where t_{ij} is the observed time in the ith year for the jth event, μ represents a common effect, Y_i the effect of the ith year, D_j the effect of the jth event and ε_{ij} the random disturbance. We fitted the model in two ways: (i) the traditional analysis of variance approach, where the model parameters are estimated by mean effects (Box *et al.*, 1978) and (ii) by the more robust approach of median fitting (Tukey, 1977).

The residuals from both mean fitting and from median fitting display the same pattern, positive residuals occurring in the top right and bottom right, suggesting that the data reveal strong systematic departures from additivity. To ensure additivity, we look for a transformation, and consider the approach proposed by Tukey (1977) (see also Hoaglin, 1977 and McNeil, 1977). The technique consists of calculating for each fitted value a comparison value, CV, which is defined as

$$(CV)_{ij} = \frac{(\text{row effect } i) \times (\text{column effect } j)}{(\text{common value})},$$

where (row effect i) and (column effect j) are the estimated row and column effects for the ith row and jth column, respectively. The (common value) is the estimated value of μ in model (8). The choice of transformation is aided by plotting $(CV)_{ij}$ against fitted residuals. This plot is called a diagnostic plot. If the scatter plot of comparison values and residuals indicate a straight line, a power transformation is then available to remove non-linearity. If m is the slope of the diagnostic plot, then $(1-m)$ or some nearby numerically simple power is likely to be useful. The diagnostic plot from the median fit is shown in Fig. 4. The slopes of the diagnostic plot for mean fit and median fit are respectively both very near to unity, indicating a logarithmic transformation. Instead of working on the original scale, we now work on the transformed scale, logarithm of the times.

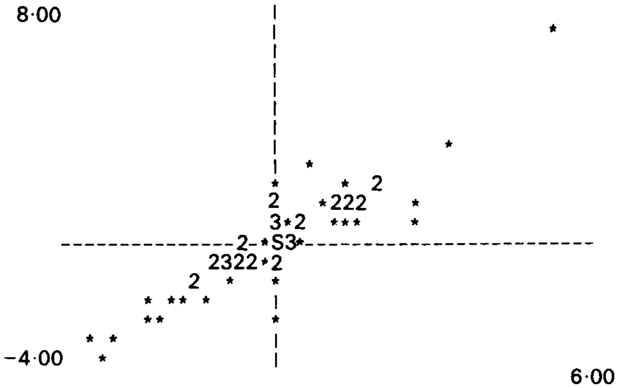

FIG. 4. Diagnostic Plot for median fit.

Tables 3 and 4 show the residuals from mean and median fitting of the transformed data. A diagnostic plot of the comparison values and transformed residuals has near zero slope (0·050) indicating that the transformation has worked effectively. We therefore take as our model:

$$\log t_{ij} = \mu' + Y'_1 + D'_j + \varepsilon_{ij}. \tag{9}$$

2.2. *Analysing the Year Effects*

A plot of median fitted row effects against year is given in Fig. 5; the plot for mean fitted row effects is very similar. Thus, a relationship similar to equation (1) seems appropriate in fitting the year effects. The fitted effects are

$$\text{Median fitted:} \quad Y'_i = -0{\cdot}072 + 0{\cdot}157\exp(-0{\cdot}73j), \tag{10}$$

$$\text{Mean fitted:} \quad Y'_i = -0{\cdot}080 + 0{\cdot}150\exp(-0{\cdot}07j). \tag{11}$$

A slightly more complex model incorporating the altitude at which the meet is held could have been fitted but is not done since altitude has a negligible effect.

2.3. *Analysing the Effect of Distance on Time*

A plot of the column effects obtained from the median fitting against distance d of the event is given in Fig. 6. A distinct non-linearity is observed. We found the column effects (D'_j as given in (9)) was best described in terms of d and $\sqrt{d}$. Although the description was arrived at from data analysis, the fitted model would have been suggested from considerations of the equations of dynamics of motion. Approximately the same descriptions emerge, whether the column effects are estimated by mean or median fitting. The fitted column effects are

$$\text{Mean fitted:} \quad D'_j = -3{\cdot}161 - 0{\cdot}00263 d_j + 0{\cdot}229\sqrt{d_j} \tag{12}$$

$$\text{Median fitted:} \quad D'_j = -3{\cdot}117 - 0{\cdot}00262 d_j + 0{\cdot}228\sqrt{d_j} \tag{13}$$

TABLE 3

Residuals from mean fitting of transformed data

	Display of residuals Column effects			
Row effects	−1·138	−0·437	0·360	1·216
0·061	−0·027	−0·007	−0·004	0·039
0·047	0·006	−0·020	0·006	0·008
0·050	−0·016	0·022	0·019	−0·024
0·029	0·005	0·002	0·003	−0·011
0·043	−0·009	0·002	0·018	−0·011
0·021	−0·006	0·005	−0·001	0·002
0·028	0·006	0·008	−0·004	−0·010
−0·004	−0·009	0·012	−0·006	0·003
−0·001	−0·012	−0·015	−0·002	0·029
−0·006	−0·007	0·010	−0·003	0·000
−0·010	0·007	−0·005	−0·006	0·004
−0·008	0·014	−0·012	0·009	−0·011
−0·030	0·007	0·004	−0·009	−0·003
−0·039	−0·004	0·004	0·005	−0·005
−0·057	0·004	−0·003	−0·006	0·005
−0·040	0·011	−0·010	−0·004	0·003
−0·047	0·010	0·008	−0·006	−0·013
−0·036	0·018	−0·004	−0·009	−0·005

TABLE 4

Residuals from median fitting of transformed data

	Display of residuals Column effects			
Row effects	−1·139	−0·437	0·360	1·216
0·059	−0·026	−0·007	−0·005	0·038
0·045	0·007	−0·020	0·006	0·007
0·048	−0·015	0·021	0·018	−0·024
0·027	0·006	0·002	0·003	−0·011
0·041	−0·008	0·002	0·018	−0·012
0·019	−0·005	0·005	−0·002	0·001
0·026	0·007	0·008	−0·004	−0·011
−0·006	−0·008	0·012	−0·007	0·003
−0·003	−0·011	−0·015	−0·003	0·028
−0·009	−0·006	0·009	−0·004	0·000
−0·013	0·008	−0·006	−0·007	0·004
−0·011	0·016	−0·013	0·009	−0·012
−0·032	0·008	0·004	−0·009	−0·003
−0·041	−0·003	0·003	0·005	−0·005
−0·059	0·005	−0·003	−0·007	0·005
−0·042	0·012	−0·011	−0·004	0·003
−0·049	0·011	0·008	−0·006	−0·013

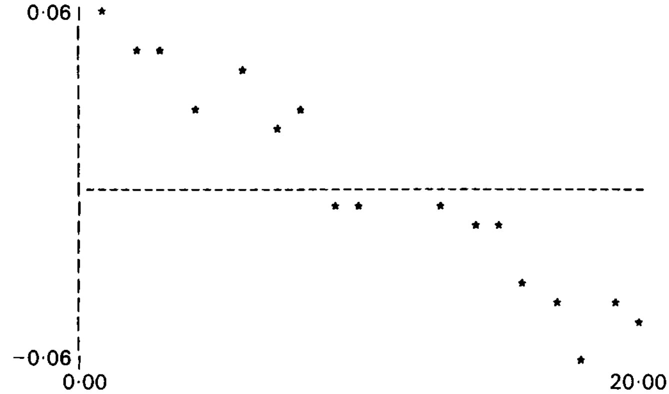

FIG. 5. Median fitted row effects against years.

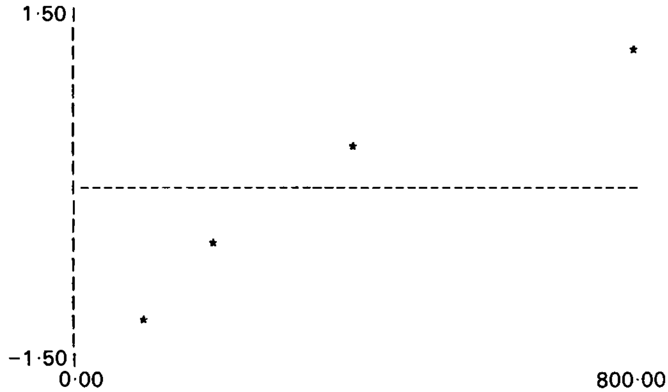

FIG. 6. Median fitted column effect against distance.

2.4. *Fitting the Model: Putting the Pieces Back*

After having obtained satisfactory descriptions of the row and column effects, we are now ready to put them together in (9) for a full specification of the model. To indicate the robustness of our analysis, we will present four models. We present two models fitted by the traditional analysis of variance approach (mean fitted) which includes and excludes altitude of the meet. Two similar models fitted by medians are also given. The fitted models are

$$\text{Mean fitted:} \quad \log t_{ij} = 0{\cdot}249 + 0{\cdot}150\exp(-0{\cdot}069j) - 0{\cdot}00263 d_j + 0{\cdot}229\sqrt{d_j}, \tag{14}$$

$$\text{Median fitted:} \quad \log t_{ij} = 0{\cdot}243 + 0{\cdot}157\exp(-0{\cdot}071j) - 0{\cdot}00262 d_j + 0{\cdot}228\sqrt{d_j}. \tag{15}$$

The residuals from these fitted models are acceptable. The residuals from the median fitted model shown in Table 5 reveal no discernible pattern. A similar picture emerges from the mean fitted model. Thus (14) and (15) may be taken as adequate models for the data. It is interesting to note that the mean fitted and median fitted models are nearly identical providing us with some assurance about the fitted model and the fitting procedures.

2.5. *Prediction from Fitted Models*

One of the aims of model fitting is to use the models for prediction. In our investigation on the Olympic times we have derived models which give adequate description of the data as evidenced by summary measures of goodness of fit as well as from a more microscopic examination of various residual plots. We expect therefore the models to predict well *but only if the external environment under which the model is built is unchanged.* The prediction for 1980 by the four methods is given in Table 6. The actual reported times are also given. From the table it is clear that the models have (except in one case) under-predicted in each of the four cases. The reason is not hard to find. In 1980 the external environment for the Olympic Games changed in that all the nations did not participate and there was a partial boycott. Some of the nations who did not participate have usually produced the athletes who have won these events. The reason for the failure of the models in predicting 1980 time is therefore understandable. This provides us with a lesson about using fitted models for prediction!

TABLE 5
Residuals of transformed data after removing median fitted row and column effects

	Display of residuals Column effects			
Row effects	−1·101	−0·395	0·395	1·254
0·061	−0·019	−0·005	0·005	0·045
0·060	0·000	−0·031	0·002	0·000
0·058	−0·016	0·016	0·020	−0·026
0·037	0·005	−0·004	0·004	−0·013
0·045	−0·003	0·003	0·025	−0·007
0·029	−0·006	0·000	0·000	0·000
0·036	0·006	0·003	−0·003	−0·012
0·002	−0·008	0·008	−0·003	0·003
0·000	−0·005	−0·014	0·005	0·033
−0·001	−0·005	0·006	−0·000	0·000
−0·004	0·008	−0·010	−0·004	0·004
−0·002	0·016	−0·017	0·012	−0·012
−0·025	0·010	0·002	−0·004	−0·002
−0·034	−0·001	0·001	0·009	−0·004
−0·050	0·005	−0·008	−0·004	0·004
−0·033	0·012	−0·015	−0·002	0·002
−0·040	0·011	0·004	−0·004	−0·013

TABLE 6
Predicted and observed times in 1980 for four running events (men)

Distance (metres)	Univariate model: Least squares	Univariate model: Jack-knife	Bivariate model: Median fit	Bivariate model: Mean fit	Observed
100	10·05	10·04	9·93	10·01	10·25
200	20·06	20·06	19·69	19·86	20·19
400	44·36	44·34	44·53	44·89	44·60
800	104·78	104·56	104·06	104·90	105·40

In Table 7, we give the predicted times for these events in 1984. They are given in the hopes that world political situations will permit the Olympic Games to return to its former status. Only 1984 will tell!

TABLE 7
Predicted 1984 times for four running events (men)

Distance (metres)	*Univariate model*		*Bivariate model*	
	Least square	*Jack-knife*	*Median fit*	*Mean fit*
100	10·01	10·00	9·92	9·98
200	19·98	19·97	19·66	19·80
400	44·16	44·14	44·42	44·77
800	104·38	104·16	103·82	104·62

ACKNOWLEDGEMENTS

Some of this work was done while the first author was visiting the Department of Mathematics, Auckland University, N.Z.

REFERENCES

BOX, G. E. P., HUNTER, W. G. and HUNTER, J. S. (1978). *Statistics for Experimenters*. New York: Wiley.
HOAGLIN, D. A. (1977). Direct approximation for chi-squared percentage points. *J. Amer. Statist. Ass.*, **72**, 508–516.
KAIREAN, J. and DALEY, A. (1977). *The Story of the Olympic Games*. Philadelphia: J. B. Lippincott & Co.
KELLER, J. B. (1974). Optimal velocity in a race. *Amer. Math. Monthly*, **81**, 474–480.
MCNEIL, D. R. (1977). *Interactive Data Analysis*. New York: Wiley.
MILLER, R. G. (1974). The jack-knife—a review. *Biometrika*, **61**, 1–15.
TUKEY, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.