Chapter 4

Regression Analysis

In Chapter 3 we used some basic statistical methods - data transformations and one-way and two-way analysis of variance - for analysing the Olympic performances in swimming, running, jumping and throwing events over the period from 1928 to 2000. These methods identified trends and showed how the performances in these sports improved over the years.

In this chapter we develop statistical models based on linear regression analysis. We also summarise the results obtained from these models and thus develop an overall view that will facilitate comparison of the performances in the four sports. This approach also facilitates the detection of outstandingly good and bad performances.

In fact, the two-way analysis of variance method for comparison is the special case of multiple linear regression in which there are two sets of binary predictor variables corresponding to the determinant and the covariate. So it is instructive to repeat the analyses given in Chapter 3 using multiple regression. We do this for each sport in turn, and then combine the results.

4.1 Swimming Performances

Figure 4.1 shows the results from fitting the multiple linear regression model to the men's swimming speeds in each of the 13 events for the 17 Olympics from 1928 to 2000. This table gives the regression coefficients for years and events, with 1928 as the baseline for year and the 50 meters freestyle as the baseline event. The table also gives the corresponding standard errors, from which confidence intervals may be computed, and p-values for testing the null hypotheses that the population coefficients are zero.

The bottom line of the table gives the r-squared and the adjusted r-squared (adjusted for the number of predictors used in the regression model), together with the residual sum of squares (rss), the residual standard deviation (s), and the p-value for testing the overall statistical significance of the model.

linear regression analysis: response = speed

predictor	coeff	St.Error	p-value
constant	1.8849	0.013819	0
year			0
1928	(0)		
1932	0.026254	0.012579	0.038932
1936	0.049603	0.012579	0.00013359
1948	0.056506	0.012579	1.5949e-005
1952	0.076478	0.013389	7.8156e-008
1956	0.124	0.012089	0
1960	0.15185	0.012089	0
1964	0.20992	0.01182	0
1968	0.2284	0.010807	0
1972	0.29366	0.010807	0
1976	0.3392	0.010925	0
1980	0.32986	0.010925	0
1984	0.35794	0.010807	0
1988	0.37103	0.010718	0
1992	0.3833	0.010718	0
1996	0.37995	0.010718	0
2000	0.40085	0.010718	0
event			0
50f	(0)		
100f	-0.21206	0.011424	0
100ba	-0.42584	0.011486	0
100br	-0.63388	0.012146	0
100bu	-0.36369	0.012146	0
200f	-0.39404	0.012146	0
200ba	-0.56907	0.01199	0
200br	-0.73635	0.011472	0
200bu	-0.54321	0.011757	0
200md	-0.60069	0.012607	0
400f	-0.50629	0.011424	0
400md	-0.69895	0.01199	0
1500f	-0.60134	0.011424	0

r-sq: 0.99302(0.99144) rss: 0.049054 df: 124 sd: 0.01989 p-value: 0

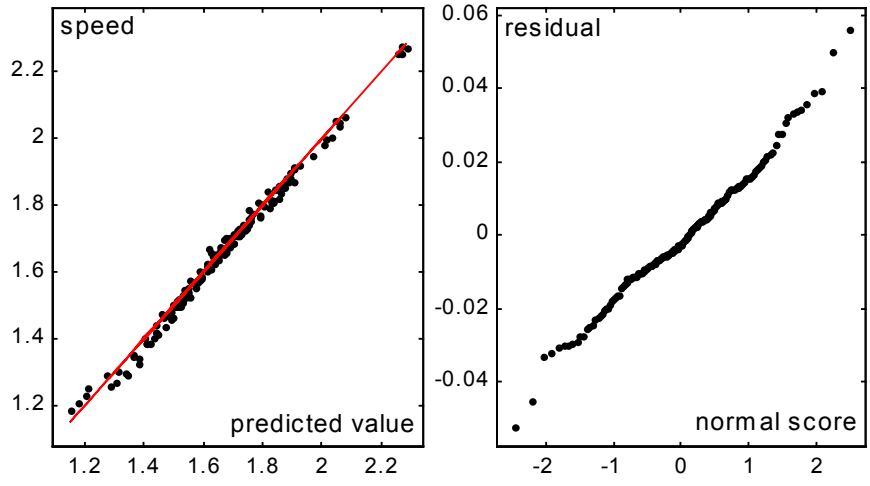

Figure 4.1: The results from fitting the multiple linear regression model to men swimming

The graphs below this table show (a) the relation between the speeds and their predicted values based on the model, and (b) the normal scores plot of residuals. These plots show (a) how well the model fits the data and (b) the plausibility of the normality assumption.

The coefficients for year increase steadily from 0 (the baseline value) in year 1928 to 0.40 in 2000. These coefficients show how the performances of the men swimmers improved over the whole period.

Since our objective is to compare the performances in the various sports (swimming, running, jumping and throwing), we need to make sure that these improvements are comparable.

How can we compare the performances in the different sports?

To answer this question, note that the performances in the swimming and running events are measured in terms of speeds, whereas the logarithms of the distances are used to compare the jumping and throwing performances.

A more valid way of comparing performances is based on percentage improvements. Taking logarithms is equivalent to using percentages, because an improvement of x percent from y to $y(1+x/100)$ corresponds to an improvement from $\log(y)$ to $\log(y)+\log(1+x/100)$ in the logarithms, that is, $\log(1+x/100)$. Note that if natural logarithms are used, this is close to $x/100$, or x percent. (This is because the Taylor expansion of $\ln(1+x/100)$ is $(x/100)+\frac{1}{2}(x/100)^2+...$). If the logarithms are taken to some other base, it doesn't change the result, because the same constant multiplies everything.

In transforming the distances we used logarithms to base 10. So to provide a valid basis for comparing the four sports, we should repeat the analysis for the swimming and running events after transforming the speeds by taking base 10 logarithms. Doing this enables us to use percentage improvements to compare all performances.

Figure 4.2 shows the corresponding result for the men's swimming performances, after transforming the speeds using base 10 logarithms.

linear regression analysis: response = log10(speed)

predictor	coeff	St.Error	p-value
constant	0.24809	0.004555	0
year			0
1928	(0)		
1932	0.0086616	0.0041464	0.038758
1936	0.015671	0.0041464	0.00024308
1948	0.01791	0.0041464	3.1743e-005
1952	0.025774	0.0044132	4.2907e-008
1956	0.037659	0.0039849	2.2204e-016
1960	0.045642	0.0039849	0
1964	0.061396	0.003896	0
1968	0.066472	0.0035621	0
1972	0.084367	0.0035621	0
1976	0.096479	0.0036011	0
1980	0.094157	0.0036011	0
1984	0.10128	0.0035621	0
1988	0.10448	0.003533	0
1992	0.10761	0.003533	0
1996	0.10675	0.003533	0
2000	0.11185	0.003533	0
event			0
50f	(0)		
100f	-0.035398	0.0037657	4.4409e-016
100ba	-0.087944	0.0037859	0
100br	-0.14299	0.0040036	0
100bu	-0.074835	0.0040036	0
200f	-0.082017	0.0040036	0
200ba	-0.12563	0.0039521	0
200br	-0.17623	0.0037815	0
200bu	-0.11905	0.0038754	0
200md	-0.13393	0.0041557	0
400f	-0.10967	0.0037657	0
400md	-0.16153	0.0039521	0
1500f	-0.13672	0.0037657	0

r-sq: 0.98894(0.98645) rss: 0.0053297 df: 124 sd: 0.006556 p-value: 0

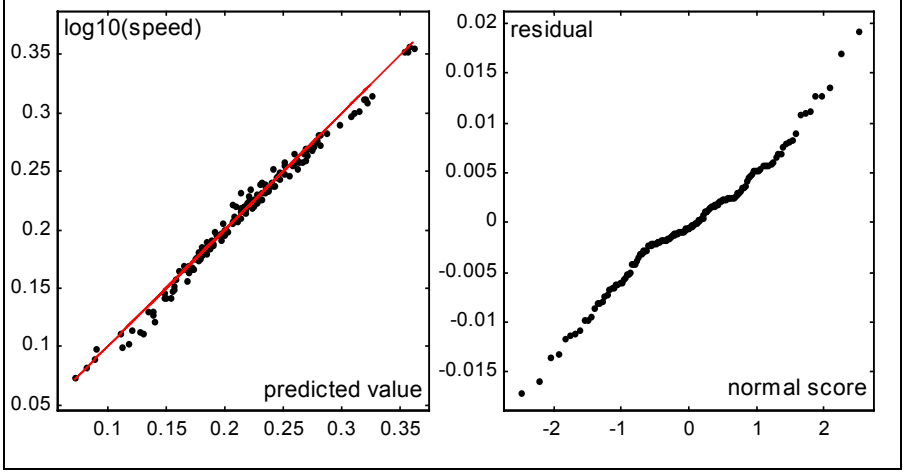

Figure 4.2: Men swimming using logarithm of speed to measure performance

The results show that, although the adjusted r-squared has decreased to 98.64%, the model still fits the data extremely well. The normal scores plot shows some slight curvature, but the normality assumption is still acceptable.

Figure 4.3 shows the corresponding result for the women's swimming performances.

linear regression analysis: response = log10(speed)

predictor	coeff	St.Error	p-value
constant	0.17514	0.0050049	0
year			0
1928	(0)		
1932	0.018356	0.0048257	0.00022808
1936	0.022796	0.0048257	6.4887e-006
1948	0.035266	0.0048257	3.6136e-011
1952	0.039985	0.0048257	2.2204e-013
1956	0.053553	0.0045992	0
1960	0.063823	0.0045992	0
1964	0.076603	0.0044447	0
1968	0.081536	0.0040394	0
1972	0.096085	0.0040394	0
1976	0.11605	0.0040766	0
1980	0.12144	0.0040766	0
1984	0.1214	0.0040394	0
1988	0.126	0.0040116	0
1992	0.12774	0.0040116	0
1996	0.12647	0.0040116	0
2000	0.13283	0.0040116	0
event			0
50f	(0)		
100f	-0.032624	0.003932	2.0917e-013
100ba	-0.086428	0.003932	0
100br	-0.13746	0.0041676	0
100bu	-0.072044	0.0040394	0
200f	-0.070815	0.0041676	0
200ba	-0.11308	0.0041676	0
200br	-0.16632	0.003932	0
200bu	-0.10831	0.0041676	0
200md	-0.12208	0.0043258	0
400f	-0.096865	0.003932	0
400md	-0.14382	0.0041155	0
800f	-0.10082	0.0041676	0

r-sq: 0.98775(0.98482) rss: 0.0054492 df: 117 sd: 0.0068245 p-value: 0

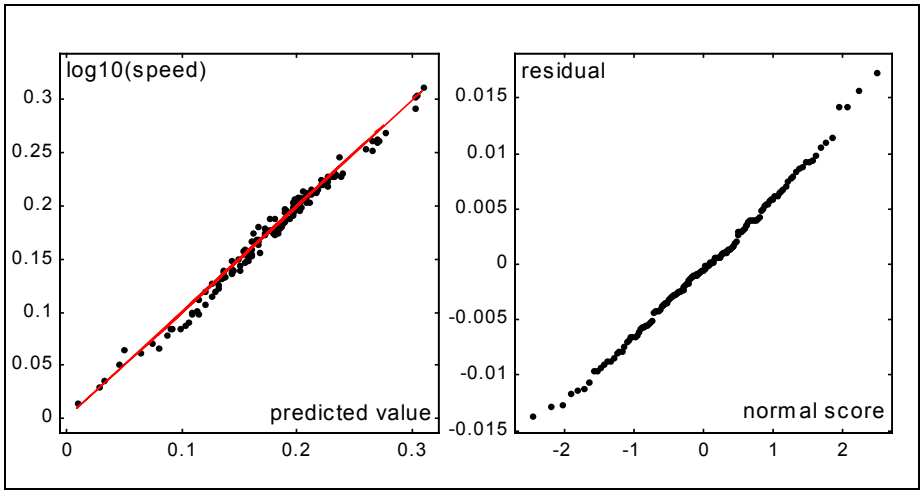

Figure 4.3: Women swimming using logarithm of speed to measure performance

In this case we see that the adjusted r-squared is 98.48%, again indicating that the model fits the data well. And again the plot of the speeds against their predicted values shows a close linear relation. The normal scores plot indicates that the normality assumption is reasonable.

4.2 Running Performances

Figure 4.4 shows the corresponding result for the men's running performances. Again the speeds are transformed using base 10 logs before fitting the model.

linear regression analysis: response = log10(speed)

predictor	coeff	St.Error	p-value
constant	0.96406	0.0023584	0
year			0
1928	(0)		
1932	0.01129	0.0028675	0.00015605
1936	0.012451	0.0028675	3.5001e-005
1948	0.016073	0.0028675	1.9898e-007
1952	0.019584	0.0028675	7.7326e-010
1956	0.021962	0.0028675	1.4857e-011
1960	0.028539	0.0028675	2.2204e-016
1964	0.030619	0.0028675	0
1968	0.039284	0.0028675	0
1972	0.035264	0.0028675	0
1976	0.036162	0.0028675	0
1980	0.031955	0.0028675	0
1984	0.040448	0.0028675	0
1988	0.042115	0.0028675	0
1992	0.040146	0.0028675	0
1996	0.044143	0.0028675	0
2000	0.041068	0.0028675	0
event			0
100	(0)		
110h	-0.083605	0.00184	0
200	-0.00047783	0.00184	0.79566
400	-0.044275	0.00184	0
400h	-0.083058	0.00184	0
800	-0.11664	0.00184	0
1500	-0.15995	0.00184	0

r-sq: 0.99272(0.99105) rss: 0.0027628 df: 96 sd: 0.0053646 p-value: 0

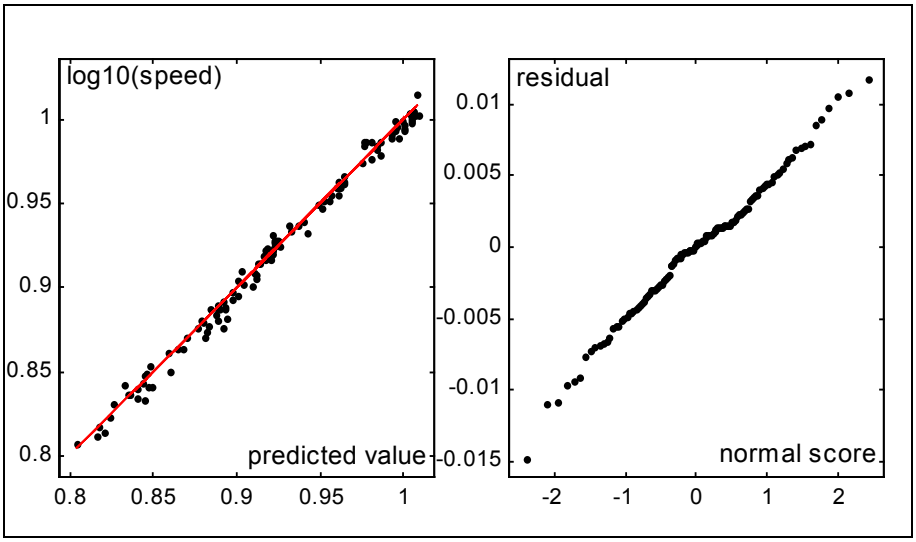

Figure 4.4: Men running using logarithm of speed to measure performance

In this case we see that the adjusted r-squared is 99.11%, indicating that the model fits the data very well. The plot of the speeds against their predicted values shows a close linear relation, and the normal scores plot indicates that the normality assumption is reasonable.

Figure 4.5 shows the corresponding result for the women's running performances.

linear regression analysis: response = log10(speed)

predictor	coeff	St.Error	p-value
constant	0.9062	0.0054228	0
year			0
1928	(0)		
1932	0.0090875	0.0075394	0.23398
1936	0.016512	0.0075394	0.033409
1948	0.018216	0.0069108	0.01127
1952	0.03131	0.0069108	3.9055e-005
1956	0.035835	0.0069108	4.2762e-006
1960	0.036488	0.0064807	9.1557e-007
1964	0.042276	0.006288	1.9536e-008
1968	0.049102	0.006288	4.2879e-010
1972	0.054101	0.006288	2.7357e-011
1976	0.058727	0.006288	2.2591e-012
1980	0.063475	0.006288	1.8607e-013
1984	0.060148	0.006288	1.0623e-012
1988	0.070118	0.006288	6.2172e-015
1992	0.064242	0.006288	1.2501e-013
1996	0.061879	0.006288	4.2721e-013
2000	0.063664	0.006288	1.6853e-013
event			0
100	(0)		
80h-100h	-0.071222	0.002595	0
200	-0.010067	0.0027298	0.00057559
400	-0.059043	0.0030773	0
800	-0.13176	0.0028814	0

r-sq: 0.9841(0.97747) rss: 0.0026238 df: 48 sd: 0.0073934 p-value: 0

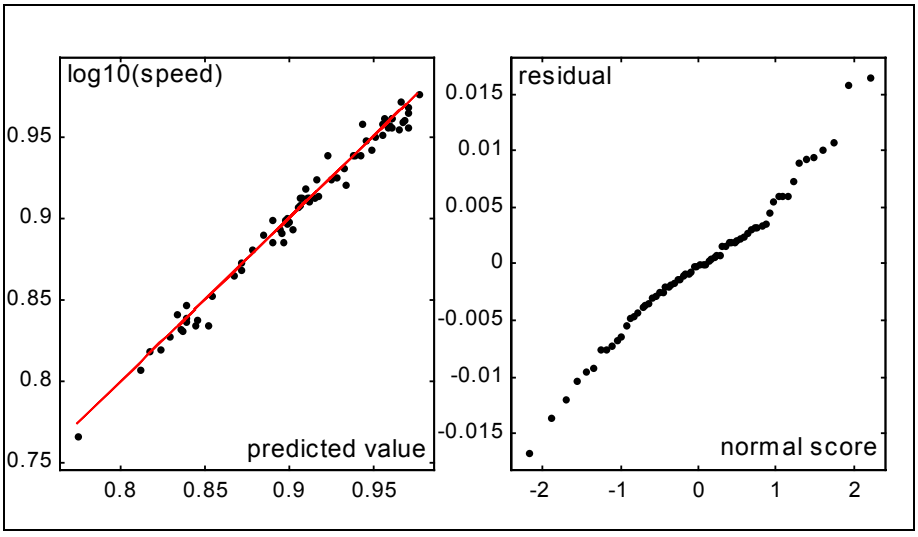

Figure 4.5: Women running using logarithm of speed to measure performance

The adjusted r-squared is now 97.75%, which is substantially less than the goodness-of-fit for the men's running events. This is not due to any particularly poorly fitting results, but rather to an overall larger variance in the speeds for the women's events. However, the plot of the log-transformed speeds against their predicted values does not show any curvature, and the normal scores plot indicates that the normality assumption is very reasonable.

4.3 Jumping and Throwing Performances

Figure 4.6 shows the results after fitting the multiple regression model for the men's and women's jumping performances.

The adjusted r-squared is now 99.91%, showing that this model provides an almost perfect fit to the jumping performances. It should be noted, however, that the r-squared statistic is high because the events form three distinct clusters, as can be seen from the plot of the logs of the distances against their predicted values. In fact there is some evidence that one of the performances in the middle cluster (corresponding to the long jump) is unusually low. This low outlier shows up more clearly in the normal scores plot.

Figure 4.7 shows the results for the men's and women's throwing performances.

The adjusted r-squared is 99.28%, which again indicates that the model fits the data very well. But as for the jumping performances, the plot of the logs of the distances against their predicted values is clustered, this time into two groups. The lower group contains the men's and women's shot put events. The normal scores plot highlights four low outliers, one of which is extremely low.

The extremely low outlier corresponds to the performance by Tapio Korjus from Finland, who won the men's javelin event at the Seoul Olympics in 1988 with a throw of 84.28 meters. To see how poor this throw was, note that this distance had been exceeded 32 years earlier at the 1956 Melbourne Olympics by Egil Danielsen from Norway with a throw of 85.71 meters. It is interesting to note that the overall best performance for its time among all Olympic throwing events, according to our model,

occurred way back in 1932. This outstanding performance was achieved by Matti Jarvinen, also from Finland, who threw the javelin a distance of 72.71 meters.

linear regression analysis: response = log distance

predictor	coeff	St.Error	p-value
constant	0.29335	0.0061069	0
year			0
1928	(0)		
1932	0.0079971	0.0079624	0.31911
1936	0.015641	0.0079624	0.053976
1948	0.005144	0.0075837	0.50011
1952	0.016917	0.0075837	0.02933
1956	0.029962	0.0075837	0.00020203
1960	0.041759	0.0075837	7.4595e-007
1964	0.049707	0.0075837	1.2615e-008
1968	0.06034	0.0075837	4.7075e-011
1972	0.057194	0.0075837	2.4721e-010
1976	0.058499	0.0075837	1.242e-010
1980	0.070969	0.0075837	1.8163e-013
1984	0.071087	0.0075837	1.7097e-013
1988	0.081497	0.0075837	8.8818e-016
1992	0.078709	0.0075837	3.5527e-015
1996	0.079278	0.007443	1.3323e-015
2000	0.073567	0.007443	2.3315e-014
event			0
M.High jump	(0)		
M.Long jump	0.57439	0.0038623	0
M.Triple jump	0.88619	0.0038623	0
W.High jump	-0.074486	0.0038623	0
W.Long jump	0.47923	0.0041128	0
W.Triple jump	0.81392	0.0090623	0

r-sq: 0.99932(0.99909) rss: 0.0078615 df: 62 sd: 0.01126 p-value: 0

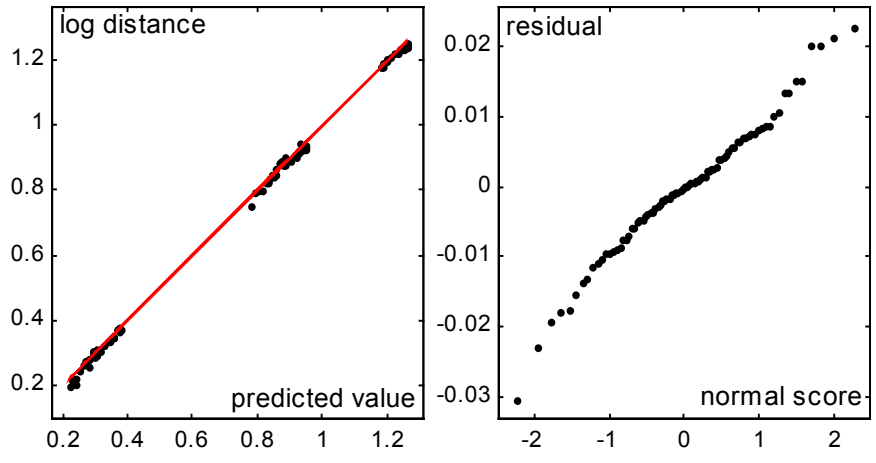

Figure 4.6: Jumping using logarithm of distance to measure performance

linear regression analysis: response = log distance

predictor	coeff	St.Error	p-value
constant	1.1737	0.010535	0
year			0
1928	(0)		
1932	0.013726	0.012871	0.28901
1936	0.032597	0.012871	0.013016
1948	0.023709	0.012479	0.060581
1952	0.060917	0.012479	4.4109e-006
1956	0.090379	0.012479	1.3118e-010
1960	0.10663	0.012479	2.6223e-013
1964	0.12168	0.012479	8.8818e-016
1968	0.1405	0.012479	0
1972	0.1599	0.012479	0
1976	0.17135	0.012479	0
1980	0.17891	0.012479	0
1984	0.16415	0.012479	0
1988	0.18873	0.012479	0
1992	0.17446	0.012479	0
1996	0.17394	0.012479	0
2000	0.1731	0.012479	0
event			0
M.shot put	(0)		
M.discus	0.49102	0.00727	0
M.javelin	0.62751	0.00727	0
M.hammer	0.55094	0.00727	0
W.shot put	-0.030087	0.0077146	0.00018274
W.discus	0.47001	0.00727	0
W.javelin	0.47665	0.007405	0

r-sq: 0.99418(0.99279) rss: 0.041331 df: 92 sd: 0.021195 p-value: 0

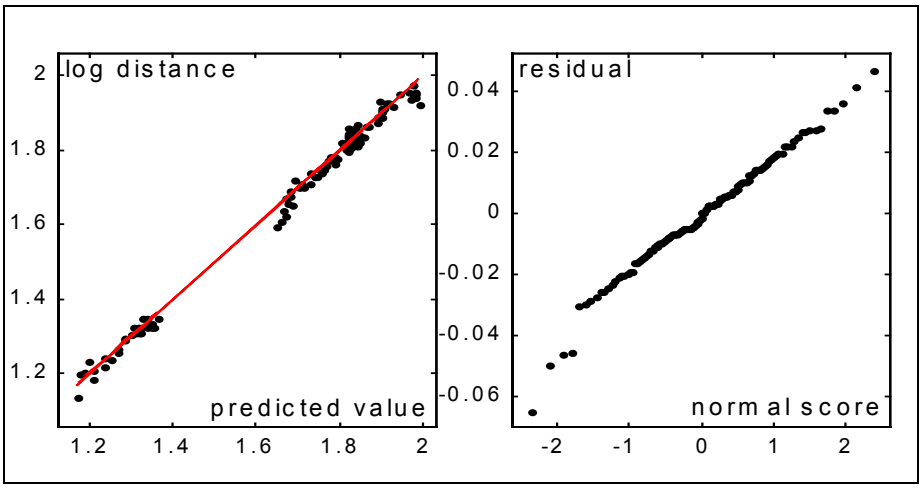

Figure 4.7: Throwing using logarithm of distance to measure performance

The other low outliers correspond to the winners of the women's discus events in 1928, 1932, and 1948. Given that these all occurred in earlier years, they provide some evidence that women's discus performances improved at a faster rate than other throwing events.

4.4 Linear Regression Model

Following are models of each events.

- The model for men's swimming is

$\log_{10}$ (speed) = 0.2481 + *year effect* + *event effect*

year effect:

[0('28) 0.0087('32) 0.0158('36) 0.0179('48) 0.0258('52) 0.0377('56) 0.0456('60) 0.0614('64) 0.0665('68) 0.0844('72) 0.0965('76) 0.0942('80) 0.0103('84) 0.1045('88) 0.1076('92) 0.1068('96) 0.1119(2000)]

event effect:

[0(50 free) –0.0354(100 free) –0.0879(100 back) –0.1430(100 breast) –0.0748(100 butterfly) –0.0820(200 free) –0.1256(200 back) –0.1762(200 breast) –0.1191(200 butterfly) –0.1339(200 medley) –0.1097(400 free) –0.1615(400 medley) –0.1367(1500 free)]

- The model for women's swimming is

$\log_{10}$ (speed) = 0.1751 + *year effect* + *event effect*

year effect:

[0('28) 0.0184('32) 0.0228('36) 0.0353('48) 0.0400('52) 0.0536('56) 0.0638('60) 0.0766('64) 0.0815('68) 0.0961('72) 0.1161('76) 0.1214('80) 0.1214('84) 0.1260('88) 0.1277('92) 0.1265('96) 0.1328(2000)]

event effect:

[0 (50 free) –0.0326(100 free) –0.0864(100 back) –0.1375(100 breast) –0.0720(100 butterfly) –0.0708(200 free) –0.11310(200 back) –0.1663(200 breast) –0.1083(200 butterfly) –0.1221(200 medley) –0.0969(400 free) –0.1438(400 medley) –0.1008(800 free)]

- The model for men's running is

$\log_{10}$ (speed) = 0.9641 + *year effect* + *event effect*

year effect:

[0('28) 0.0113('32) 0.0125('36) 0.0161('48) 0.0196('52) 0.0220('56) 0.0285('60) 0.0306('64) 0.0393('68) 0.0353('72) 0.0362('76) 0.0320('80) 0.0404('84) 0.0421('88) 0.0401('92) 0.0441('96) 0.0405(2000)]

event effect:

[0(100 sprint) –0.0836(100 hurdles) –0.0007(200 sprint) –0.0443(400 sprint) –0.0831(400 hurdles) –0.1166(800 sprint) –0.156(1500 sprint)]

- The model for women's running is

$\log_{10}$ (speed) = 0.9062 + *year effect* + *event effect*

year effect:

[0(‘28) 0.0091(‘32) 0.0165(‘36) 0.0182(‘48) 0.0313(‘52) 0.0358(‘56) 0.0365(‘60) 0.0423(‘64) 0.0491(‘68) 0.0541(‘72) 0.0587(‘76) 0.0635(‘80) 0.0602(‘84) 0.0701(‘88) 0.0642(‘92) 0.0619(‘96) 0.0637(2000)]

event effect:

[0(100 sprint) –0.0712(80 hurdles,100 hurdles) –0.0101(200 sprint) –0.0590(400 sprint) –0.1318(800sprint)]

- The model for jumping is

$\log_{10}$ (distance) = 0.2934 + *year effect* + *event effect*

year effect:

[0(‘28) 0.0080(‘32) 0.0156(‘36) 0.0051(‘48) 0.0169(‘52) 0.0300(‘56) 0.0418(‘60) 0.0497(‘64) 0.0603(‘68) 0.0572(‘72) 0.0585(‘76) 0.0710(‘80) 0.0711(‘84) 0.0815(‘88) 0.0787(‘92) 0.0791(‘96) 0.0736(2000)]

event effect:

[0(M.high jump) 0.5744(M.long jump) 0.8862(M.triple jump) –0.0745(W.high jump) 0.4792(W.long jump) 0.8139(W.triple jump)]

- The model for throwing is

$\log_{10}$ (distance) = 1.1737 + *year effect* + *event effect*

year effect:

[0(‘28) 0.0137(‘32) 0.0326(‘36) 0.0237(‘48) 0.0609(‘52) 0.0904(‘56) 0.1066(‘60) 0.1217(‘64) 0.1405(‘68) 0.1599(‘72) 0.1714(‘76) 0.1789(‘80) 0.1642(‘84) 0.1887(‘88) 0.1745(‘92) 0.1739(‘96) 0.1731(2000)]

event effect:

[0(M.shot put) 0.4910(M.discus) 0.6275(M.javelin) 0.5509(M.hammer) –0.0301(W.shot put) 0.4700(W.discus) 0.4767(W.javelin)]

In the next section we combine the results from these models.

4.5 Overall trends

Figure 4.8 graphs the coefficients for year from the models for the six sports (separating men's and women's swimming and running events but combining the men's and women's jumping and throwing events) shown in Figures 4.2 to 4.7.

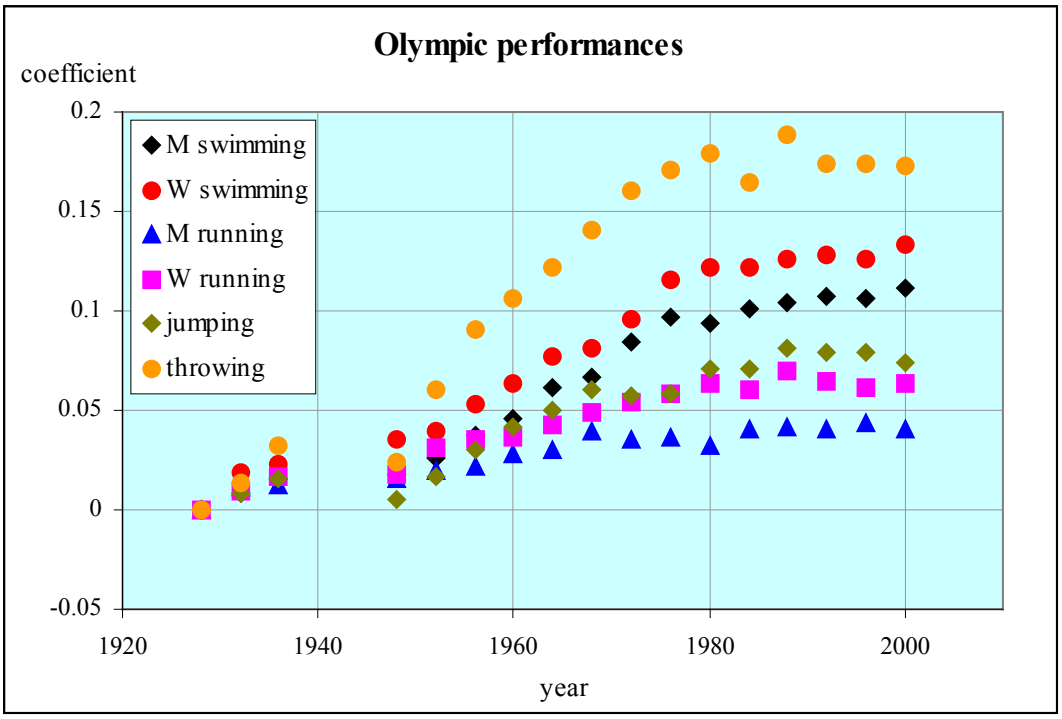

Figure 4.8: Comparison of performances in different sports

You can see from this graph that the performances increase with time, but more so in some events than others. The largest increase is in the throwing events, followed by the women's and the men's swimming events. The improvements in jumping and in the women's running over the years are similar. However the least improvements occur in the men's running events.

For each sport there is a steady improvement from 1928 to 1972, followed by a period in which there is little or no improvement. The main exception is in swimming, where both men's and women's performances improved noticeably at the Sydney 2000 Olympics.