Statistical description and inference study based on decathlon events data

Student name: Zhuo Xin

Student number: 1886015

Completion date: 8th January 2020

Abstract

this report is based on some data about decathlon events during 22 years, the aims are discovering some main features of data and using related statistical inferential methods to examine two assumptions that the one is about linear correlation between some events, and the other one is about linear regression model of decathlon scoring system. Correlation analysis, hypothesis test and linear regression are some fundamental methods in this report, fortunately, the hypothesis about correlation has been verified, although the linear regression of scoring system may be not reliable.

TABLE OF CONTENTS

Introduction

It has been reported by Tomlinson (2010) that the word 'decathlon' is from the Greek meaning 'ten prizes', an event in athletics comprising ten activities, four track and six field. Day one includes the 100 metres, long jump, shot-put, high jump, and 400 metres; day two includes the 110-metre hurdles, discus, pole vault, javelin, and 1,500 metres. However, decathlon was not always with ten events at beginning, the original version called 'pentathlon' included five items: long jump, discus, javelin, sprint, and a wrestling match. It is probably a great match to analyse participant's several types of sports ability.

Firstly, some main features of data would be shown through descriptive analysis section, following that, there are two sections conducting some statistical inferences in terms of linear correlation between some sports and linear regression respectively to verify if original hypothesis is correct. The purpose of the report is testing some assumptions via extracting useful data and executing statistical inferences methods.

Background

This report is based on a document recording some related data about decathlon events from 1985 to 2006, the data includes names and nationalities of athletes, years of events, results and points of each athlete. The main method to process data and data visualization is Python programming including Pandas, NumPy, Matplotlib and Seaborn modules.

Analysis on countries and points

1. Champions and excellent countries

We could assume that the country with highest total points of each year is a champion, the result could be shown after extracting via Python, from Figure 1, it seems that USA got ten times champion and following that is CZE whose frequency is seven. Besides, there were four countries with champions, it seems that American and Czech athletes dominated decathlon games from 1985 to 2006. From Appendix 1, the top ten countries with highest average total point are listed in descending order, JAM, BAR and UZB are the top three countries, especially for JAM, the mean of total point is larger than other countries apparently, but during this period, there were not champions from JAM, which may show that the overall performance of Jamaican athletes is excellent.

2. Simple analysis of Points of Each event

From Appendix 2, figure is mean point of each event, it would demonstrate athletes could probably get higher points in 110-metre hurdles, 100 metres, long jump and 400 metres, and there are three items of them are sprint events. In contrast,

Discus, javelin, shotput and 1500 metres are possibly not easy for athletes to get high points due to their lower mean points comparatively, specifically, it appears that three items of them are about throwing events.

3. Normality test of each event points

For verifying the hypothesis that points of each event are subject to normal distribution, first step is depicting histogram if the distribution of each event scoring, as shown by Figure 2, it appears that except for ‘high jump’ and ‘pole-vault’, other scoring of events may be subject to normal distribution, for more precise result, using Shapiro-Wilk test by Python may be a reliable way to make sure if previous assumption is correct, The null-hypothesis of this test is that the population is normally distributed, let significance level α=0.05, as shown by Figure 3, the second number in the round brackets is p-value, so it would indicate that only points of ‘100 metres’ does not reject null hypothesis and be subject to normal distribution.

Figure 1: Distribution of champions

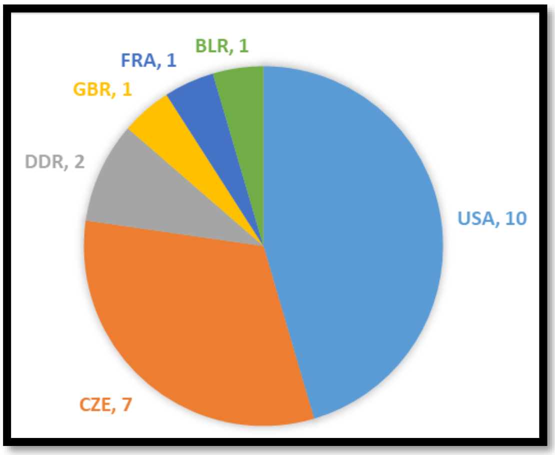


Figure 2: Distribution histograms of each event scoring

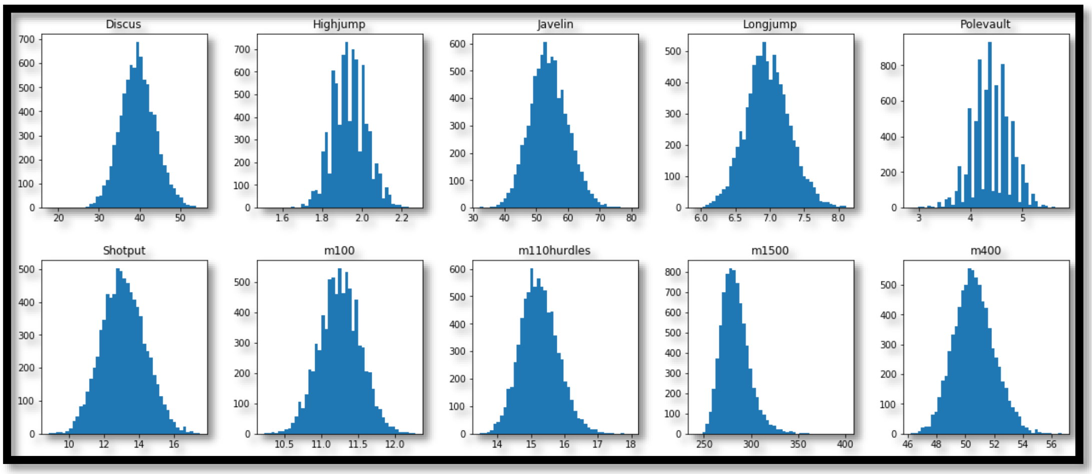
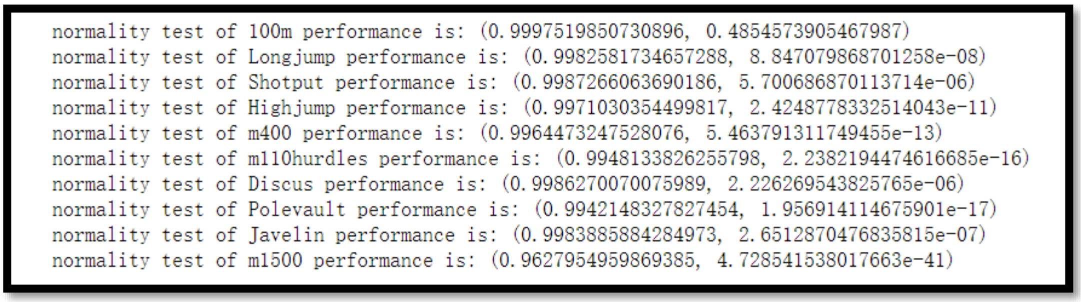


Figure 3: Results of Shapiro-Wilk normality test



```
normality test of 100m performance is: (0.9997519850730896, 0.4854573905467987)
normality test of Longjump performance is: (0.9982581734657288, 8.847079868701258e-08)
normality test of Shotput performance is: (0.9987266063690186, 5.700686870113714e-06)
normality test of Highjump performance is: (0.997103035449817, 2.4248778332514043e-11)
normality test of m400 performance is: (0.996447324752807б, 5.463791311749455e-13)
normality test of m110hurdles performance is: (0.9948133826255798, 2.2382194474616685e-16)
normality test of Discus performance is: (0.9986270070075989, 2.226269543825765e-06)
normality test of Polevault performance is: (0.9942148327827454, 1.956914114675901e-17)
normality test of Javelin performance is: (0.9983885884284973, 2.6512870476835815e-07)
normality test of m1500 performance is: (0.962795495986938б, 4.728541538017663e-41)
```

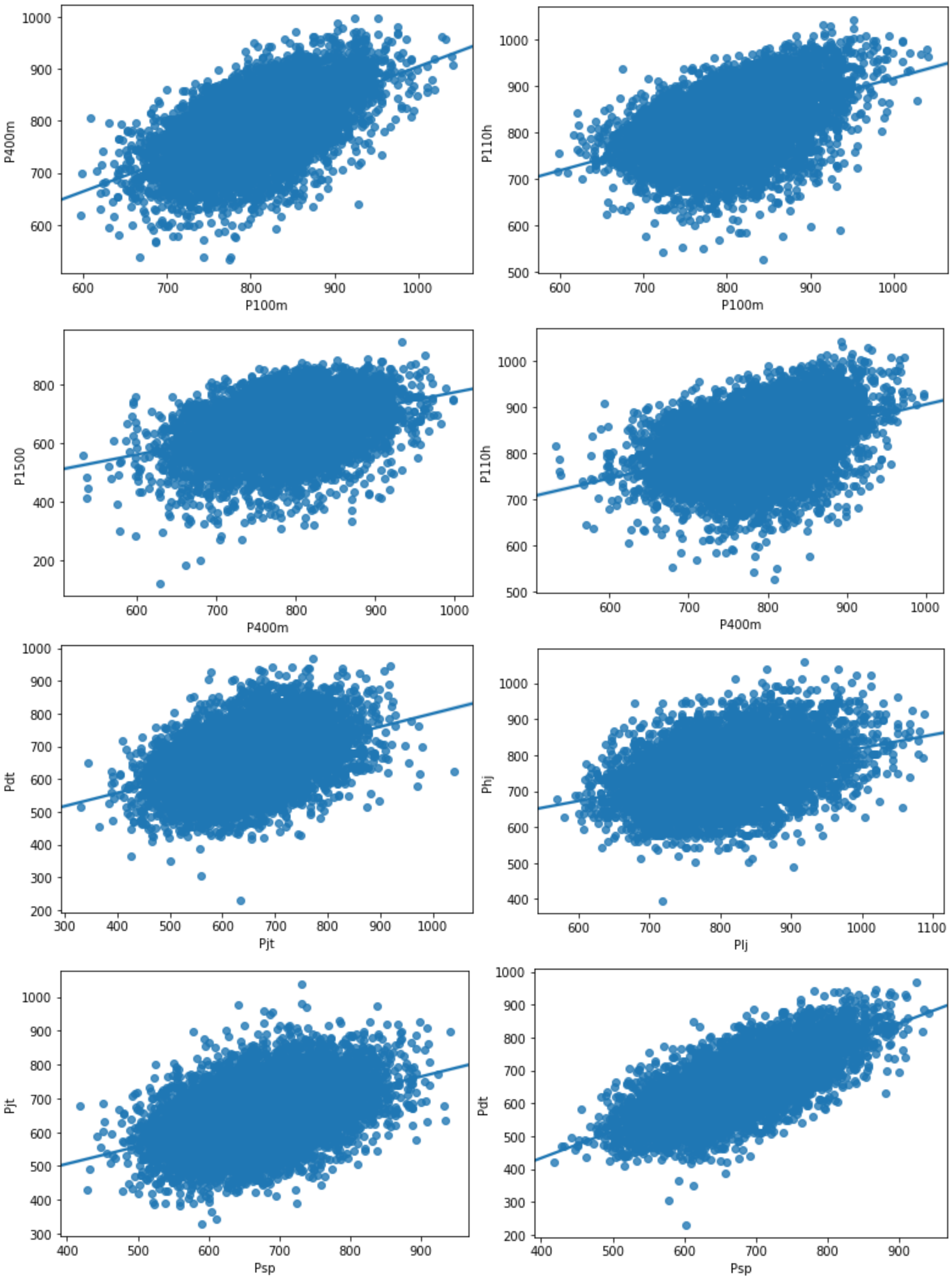
Linear correlation analysis

it could be argued that ten events in decathlon may be divided into three categories: running, throwing and jumping. In this case, 100 metres, 400 metres, 110-metre hurdles and 1,500 metres could be classified as running category, furthermore, long jump, high jump and pole vault could be in jumping category, eventually, shot-put, discus and javelin could be members in throwing type.
It could be assumed that there are some linear correlations between some events' score in every category.

1. scatters of each pair of variables

Using scatters of each pair of variables to check correlation in general. Firstly, as Figure 4 shown below, it appears that there are some pairs of variables with visible linear correlation, specifically, 1500 metres and pole vault are variables that do not probably have strong correlation with some other events in their categories, Costill et all (1976), cited in Maruo et all (2018) indicated that 'In sprinters large numbers of fast-twitch muscle fibres are required to accelerate in a transient period, whereas for long-distance runners a greater number of slow-twitch muscle fibres are required to maintain their own pace during a relatively long-lasting race', which means athletes would possibly use different muscle for different events, perhaps every part of muscle is not same strong as others, thus, they would have different sports ability in every event and get different score in sprint events and long-distance running, which may cause no obvious correlation between score of 1500 metres and some other running events.

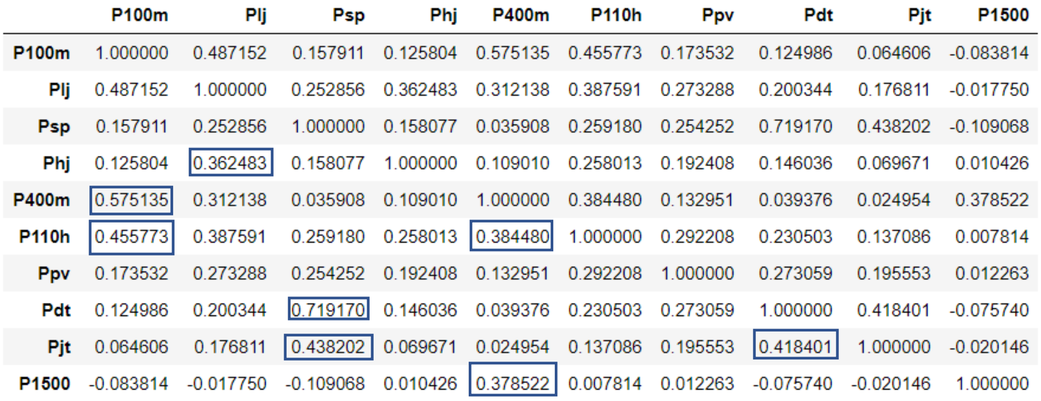
Figure 4: Scatters of each pair of variables



2. linear correlation coefficient

Calculating linear correlation coefficient to check correlation more precisely. According to pairs of events in Figure 4, using their sample data to calculate linear correlation coefficient would be a quantitative method to verify correlation, if correlation coefficient is less than 0.3, it may indicate a weak linear relationship, as Table 1 shown, all correlation coefficient is greater than 0.3. it would indicate that there is linear correlation between these chosen variables from Figure 4.

Table 1: Linear correlation coefficient

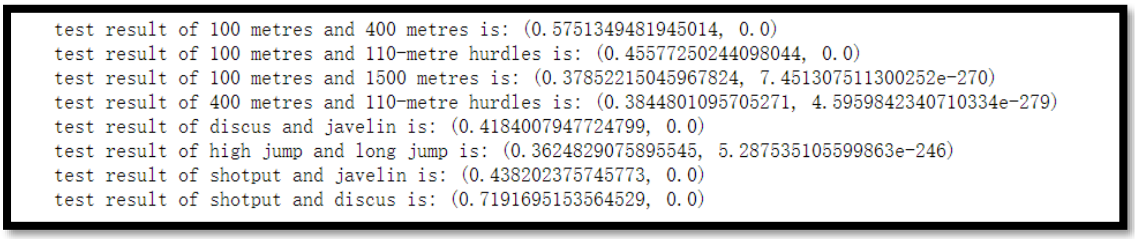


	P100m	Plj	Psp	Phj	P400m	P110h	Ppv	Pdt	Pjt	P1500
P100m	1.000000	0.487152	0.157911	0.125804	0.575135	0.455773	0.173532	0.124986	0.064606	-0.083814
Plj	0.487152	1.000000	0.252856	0.362483	0.312138	0.387591	0.273288	0.200344	0.176811	-0.017750
Psp	0.157911	0.252856	1.000000	0.158077	0.035908	0.259180	0.254252	0.719170	0.438202	-0.109068
Phj	0.125804	0.362483	0.158077	1.000000	0.109010	0.258013	0.192408	0.146036	0.069671	0.010426
P400m	0.575135	0.312138	0.035908	0.109010	1.000000	0.384480	0.132951	0.039376	0.024954	0.378522
P110h	0.455773	0.387591	0.259180	0.258013	0.384480	1.000000	0.292208	0.230503	0.137086	0.007814
Ppv	0.173532	0.273288	0.254252	0.192408	0.132951	0.292208	1.000000	0.273059	0.195553	0.012263
Pdt	0.124986	0.200344	0.719170	0.146036	0.039376	0.230503	0.273059	1.000000	0.418401	-0.075740
Pjt	0.064606	0.176811	0.438202	0.069671	0.024954	0.137086	0.195553	0.418401	1.000000	-0.020146
P1500	-0.083814	-0.017750	-0.109068	0.010426	0.378522	0.007814	0.012263	-0.075740	-0.020146	1.000000

3. the significance test of the correlation coefficient

Conducting t-test to examine reliability of the sample correlation coefficient. It could be assumed that the variables have bivariate normal distribution; their variances are equal; the observations are independent. The null-hypothesis of this test is that population correlation coefficient ρ is equal to zero, let significance level α=0.05. The result is shown in Figure 5, the second number in the round brackets is p-value, all calculated p-values are less than α, it seems that all eight pairs of variables rejected the null hypothesis, so it could be concluded that there are significant correlations between each pair verified.

Figure 5: results of the significance test of the correlation coefficient



```
test result of 100 metres and 400 metres is: (0.5751349481945014, 0.0)
test result of 100 metres and 110-metre hurdles is: (0.45577250244098044, 0.0)
test result of 100 metres and 1500 metres is: (0.37852215045967824, 7.451307511300252e-270)
test result of 400 metres and 110-metre hurdles is: (0.3844801095705271, 4.595984234071O334e-279)
test result of discus and javelin is: (0.4184007947724799, 0.0)
test result of high jump and long jump is: (0.3624829075895545, 5.287535105599863e-246)
test result of shotput and javelin is: (0.438202375745773, 0.0)
test result of shotput and discus is: (0.7191695153564529, 0.0)
```

Linear regression

It is reported by Barrow (2012) that there are ten events in decathlon, “in order to combine the results of these very different events - some give times and some give distances - a points system has been developed. Each performance is awarded a predetermined number of points according to a set of performance tables. These are added, event by event, and the winner is the athlete with the highest points total after ten events’’.

Therefore, it may be assumed that the scoring system of decathlon satisfies a linear regression model between the result of an event and the score of that. The linear regression model is:

$$y = \mathrm{a} + \beta x + \varepsilon$$

y represents points of events; x represents results of events; ε is an error term.

The estimated regression equation is:

$$\hat{y} = \hat{\mathrm{a}} + \hat{\beta} x$$

In this report, extracting results of 100 metres and points of 100 metres as an example, let results be independent variable x and points be dependent variable y, find the estimated regression equation of them and evaluate regression models.

1. The estimated regression equation

The estimated regression equation could be calculated by the method of least squares. After coping with data by Python, the result is shown by Table 2, the estimated regression equation is:

$$\hat{y} = 3238.5945 - 216.1561\hat{x}$$

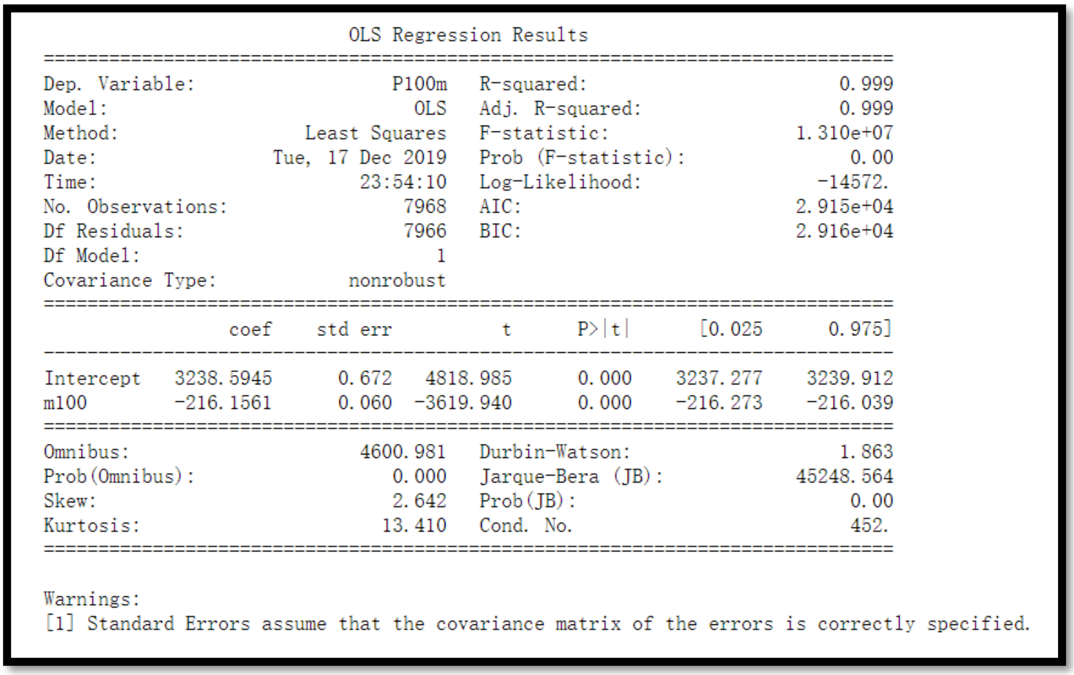
2. The goodness of fit

The goodness of fit could be examined by coefficient of determination R^2, when R^2 is closer to 1, it shows that the goodness of fit is probably better. As Table 2 shown, coefficient of determination R^2 is 0.999 which means goodness of fit may be great.

3. Hypothesis tests for the model parameters

To be specific, the model parameters could be examined by t-test. The null-hypothesis of these two tests are that α is equal to zero and β is equal to zero, let significance level α=0.05, as results shown in Table 2, two p-values of t-test are less than α, which possibly means that the estimated regression equation could reflect the linear relationship between variable x and y.

Table 2: information about linear regression



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   P100m   R-squared:                       0.999
Model:                             OLS   Adj. R-squared:                  0.999
Method:                  Least Squares   F-statistic:                 1.310e+07
Date:                 Tue, 17 Dec 2019   Prob (F-statistic):               0.00
Time:                         23:54:10   Log-Likelihood:                -14572.
No. Observations:                 7968   AIC:                         2.915e+04
Df Residuals:                     7966   BIC:                         2.916e+04
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3238.5945      0.672   4818.985      0.000    3237.277    3239.912
m100        -216.1561      0.060  -3619.940      0.000    -216.273    -216.039
==============================================================================
Omnibus:                     4600.981   Durbin-Watson:                   1.863
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            45248.564
Skew:                           2.642   Prob(JB):                         0.00
Kurtosis:                      13.410   Cond. No.                         452.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

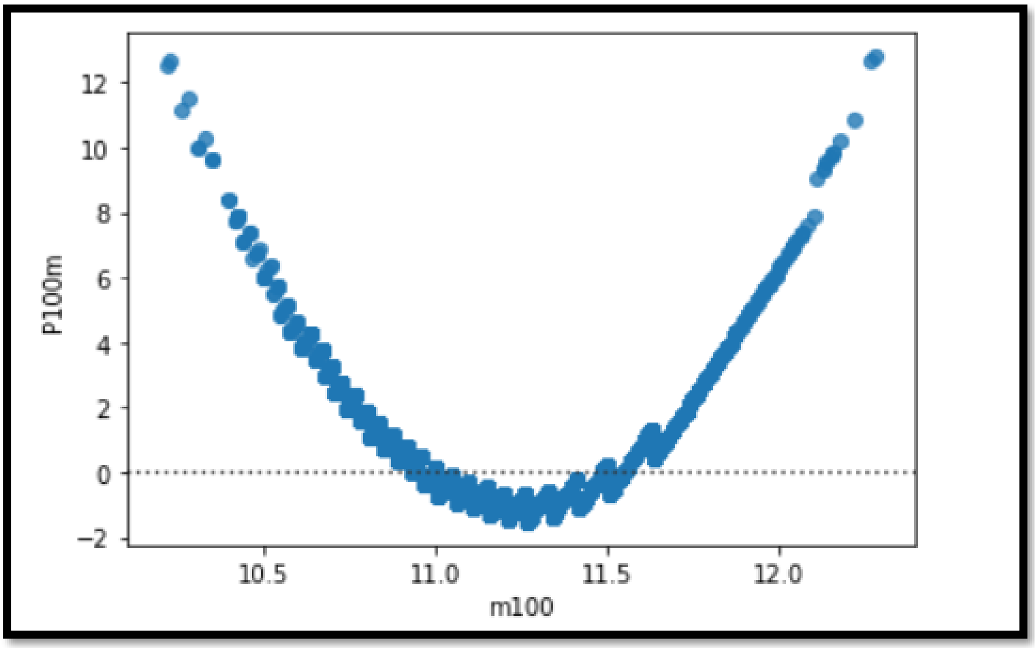
4. ANOVA for regression

ANOVA is a reliable method to inspect whether the linear relationship between dependent variables and independent variables is significant, which could be conducted by F-test. The null-hypothesis is that β is equal to zero, as results shown in Table 2, the p-value of the test is approximately equal to zero. It could be argued that there is a significant linear relationship between variable x and y.

5. Residual analysis

Residual analysis is a vital part, it is significant to examine whether the error term ε satisfies the assumption that is ε∽*N* (0, σ^2), therefore, residual plot is a reliable method, if this regression model of two variables x and y is correct, points in a residual plot would be appeared in a horizontal band area, however, in Figure 6, it seems that the regression model of two variables is probably improper.

Figure 6: Residual plot



Conclusion

In this research, some main features of data were illustrated, and there are two assumptions that the one proved to be correct is that there are some linear correlations between some events' score in every category. However, the other one proved to be unreliable is that scoring system of decathlon satisfies a linear regression model between the result of an event and the score of that. In this case, curve regression or other regression models may be considered for the scoring system in the future studies.

References

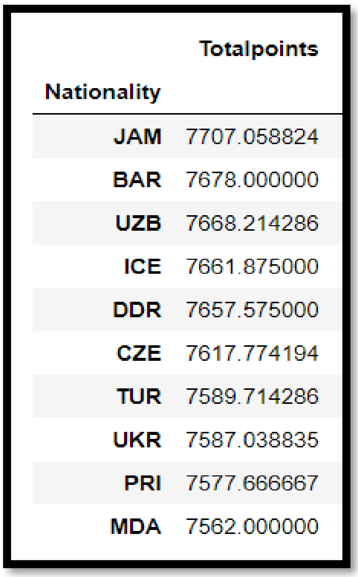
Barrow, J. 2012. Decathlon: the Art of Scoring Points. Available at: https://nrich.maths.org/8346 [Accessed:22 December 2019]

Maruo, Y. et all. 2018. *Long-Distance Runners and Sprinters Show Different Performance Monitoring – An Event-Related Potential Study* 9, 635. doi: 10.3389/fpsyg.2018.00653

Tomlinson, A. 2012. *A Dictionary of Sports Studies*. Oxford: Oxford University Press.

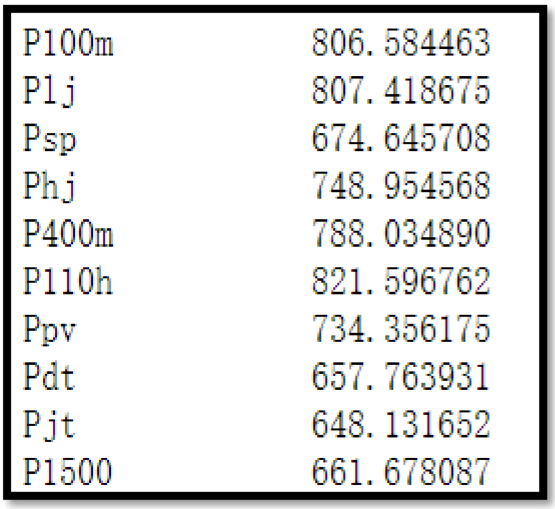
Appendix

Appendix 1: Top 10 highest mean of total points



	Totalpoints
Nationality	
JAM	7707.058824
BAR	7678.000000
UZB	7668.214286
ICE	7661.875000
DDR	7657.575000
CZE	7617.774194
TUR	7589.714286
UKR	7587.038835
PRI	7577.666667
MDA	7562.000000

Appendix 2: Mean points of each event



```
P100m          806.584463
Plj            807.418675
Psp            674.645708
Phj            748.954568
P400m          788.034890
P110h          821.596762
Ppv            734.356175
Pdt            657.763931
Pjt            648.131652
P1500          661.678087
```