Modelling the progression of world records in athletics - Events over middle distances

Jinda Zhang

1 Chapter 1 - Introduction to the problem

1.1 Discussion of the context

Data are available on the progression of world record times for the following events - 400 metres, 800 metres and 1500 metres for both men and women. The data are stored in Men400m.csv, Men800m.csv,Men1500m.csv, Women400m.csv, Women800m.csv and Women1500m.csv. Each file contains the following eight pieces of information for every ratified occasion on which the previous world record was beaten:

1.2 Aims of the proposed research

- For each event separately, fit and assess a model to the world record times.
- Use the best-fitting model type to compare the patterns of progress in the three events for men and women separately.
- Use the best-fitting model type to compare the patterns of progress for men and women in each event separately.

1.3 Questions of interest

- For each event separately, how fast is the progress of world record change over time?
- For men and women separately, how fast is the progress of world record change over time?
- For each event separately, how fast is the progress of world record change over time?
- When will the next world records breaking is going to happen?

1.4 Description of the study and variables involved

- Index - A serial number from 1 to n.
- Time - The new world record time (in seconds).
- Competitor - The name of the new world record holder.
- DOB - The new world record holder's date of birth (dd/mm/yyyy).
- Country - The country that the new world record holder represented (a 3-letter code).
- Venue - Where the new world record was set.
- Date - The date when the new world record was set (dd/mm/yyyy).
- Altitude - The altitude of city that events happened.
- Age - which is calculated by (Date-DOB)/365.
- Speed - Speed of athletics in match (m/s).

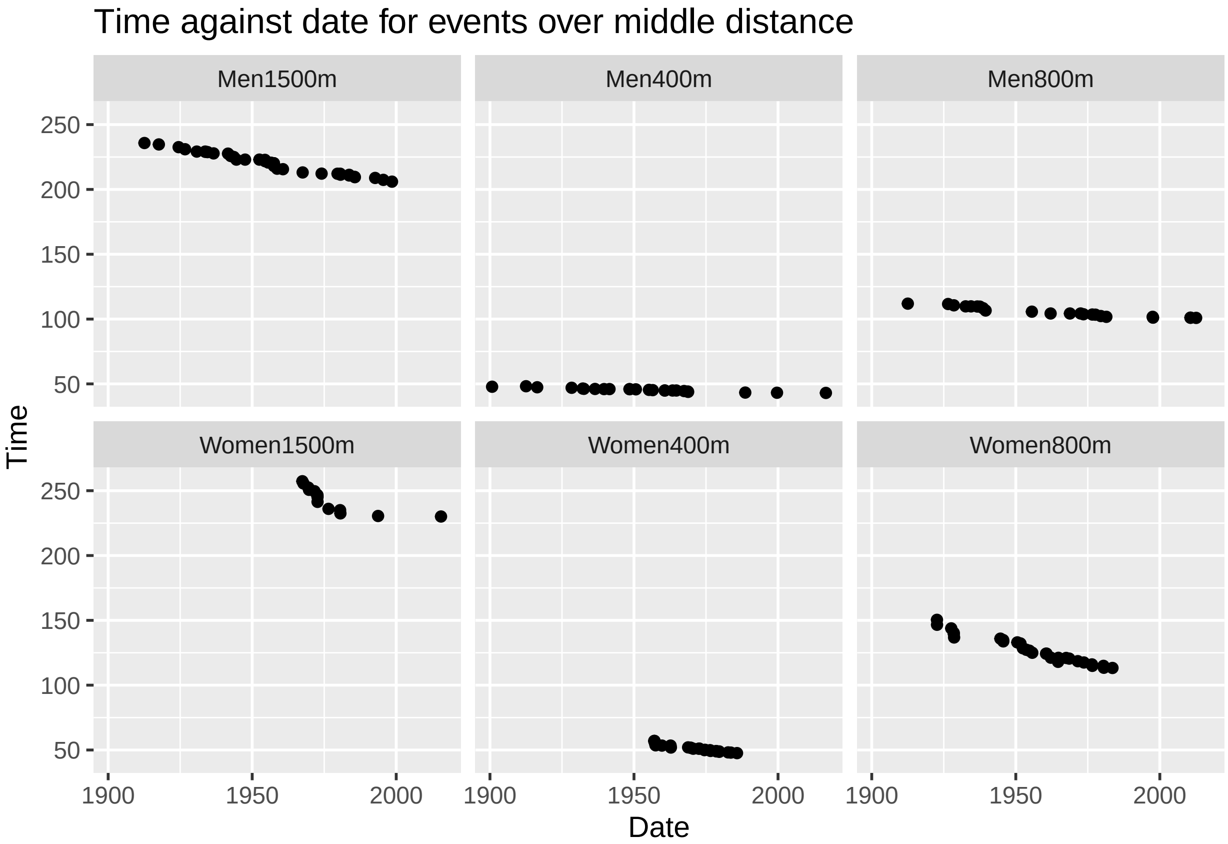
2 Chapter 2 - Description of the methods

2.1 Description of the statistical methods used

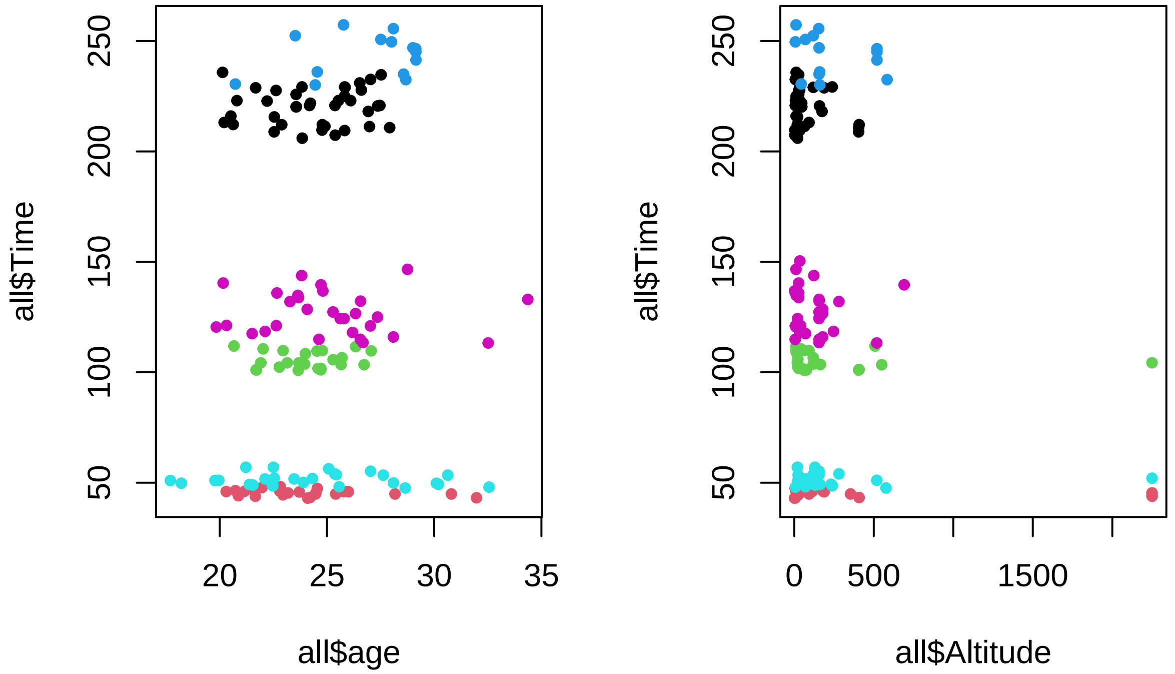
In this study, linear regression with its transformations, polynomial regression, Generalized additive models are being used to investigate the relationship between world records and date.

3 Chapter 3 - Analysis of the data

3.1 Exploratory data analysis



From the plot, we see that variable Time decreases as Date increases, as athletes takes less time to break world records. However, the plot for women seems non-linear, which indicates that linear models might not be appropriate here.



From the plot, we find that most athletics are among 20-35 years old, and 'Age' seems does not has a large effect on 'Time'. Most events happens in city that has altitude less 1000m. As a result, we consider 'Altitude' that are greater than 2000m as outliers in our study.

Table 1: Summary statistics on record Time by sex of 1500m events.

sex	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
F	14	243.5	9.3	230.1	235.25	245.8	250.425	257.3
M	37	220.4	8.2	206.0	212.20	220.8	227.600	235.8

Table 2: Summary statistics on record Time by sex of 400m events.

sex	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
F	27	51.3	2.8	47.60	49.11	51.0	53.4	57.0
M	21	45.6	1.4	43.03	44.90	45.9	46.2	48.2

Table 3: Summary statistics on record Time by sex of 800m events.

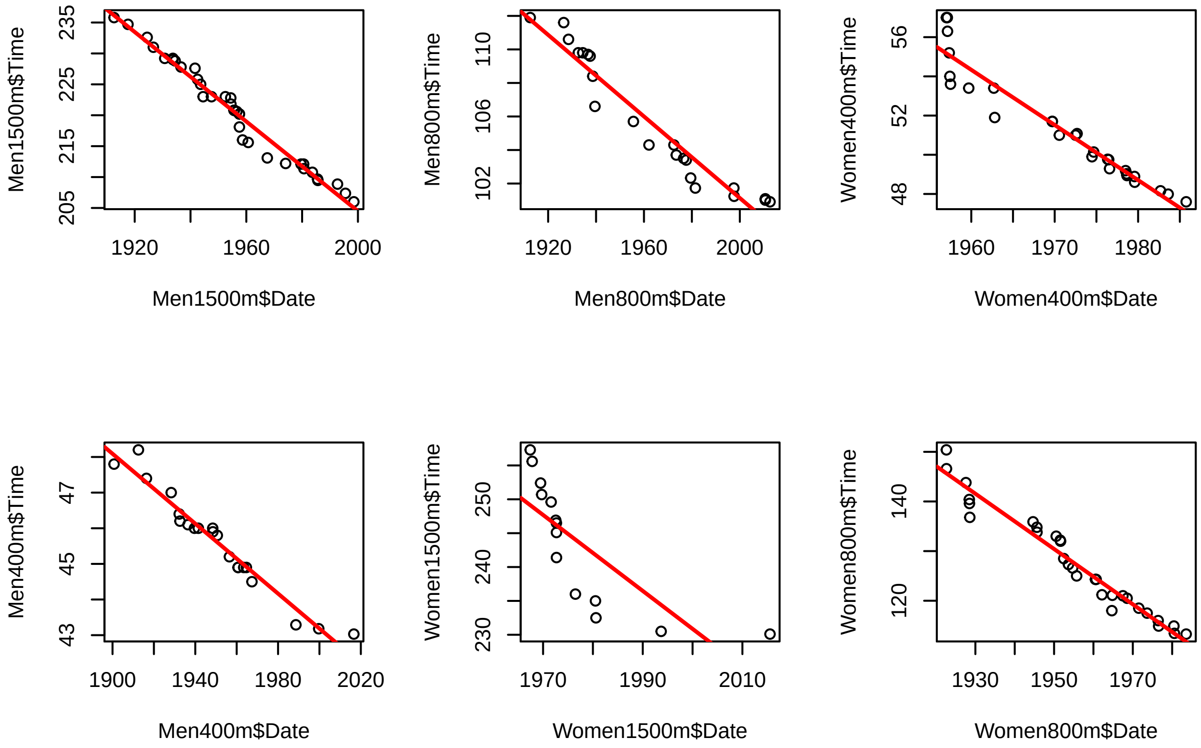
sex	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
F	30	127.5	10.4	113.28	119.00	125.8	134.550	150.4
M	22	105.6	3.9	100.91	101.88	104.3	109.675	111.9

For each event separately, we could see that man are faster than women in events of 400m, 800m, 1500m.

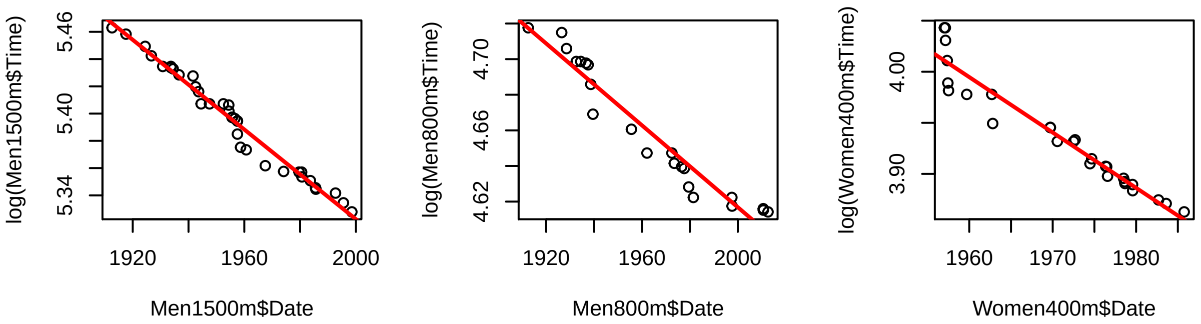
3.2 Analyses and model checks

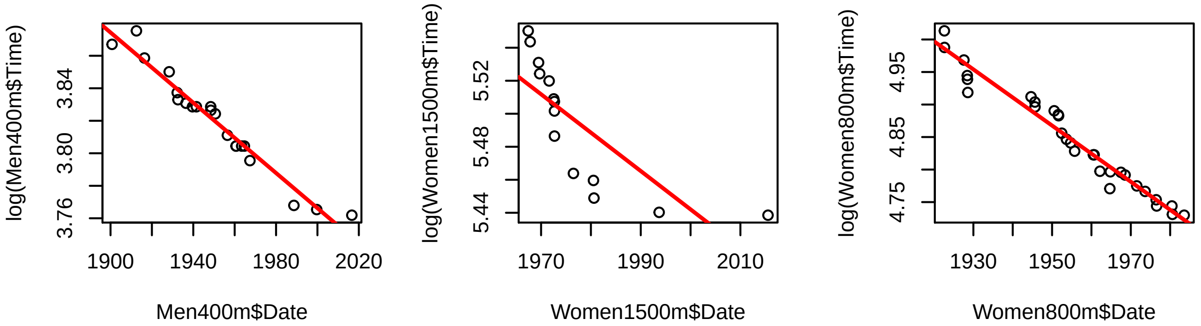
3.2.1 linear models

Firstly, we start by fitting linear models to investigate the relationship between variable Time and Date.



From the plot, we can see that linear model does not fit the data well, as there seems to exist curvature, which can not be treated well by linear model. What about log-linear models?

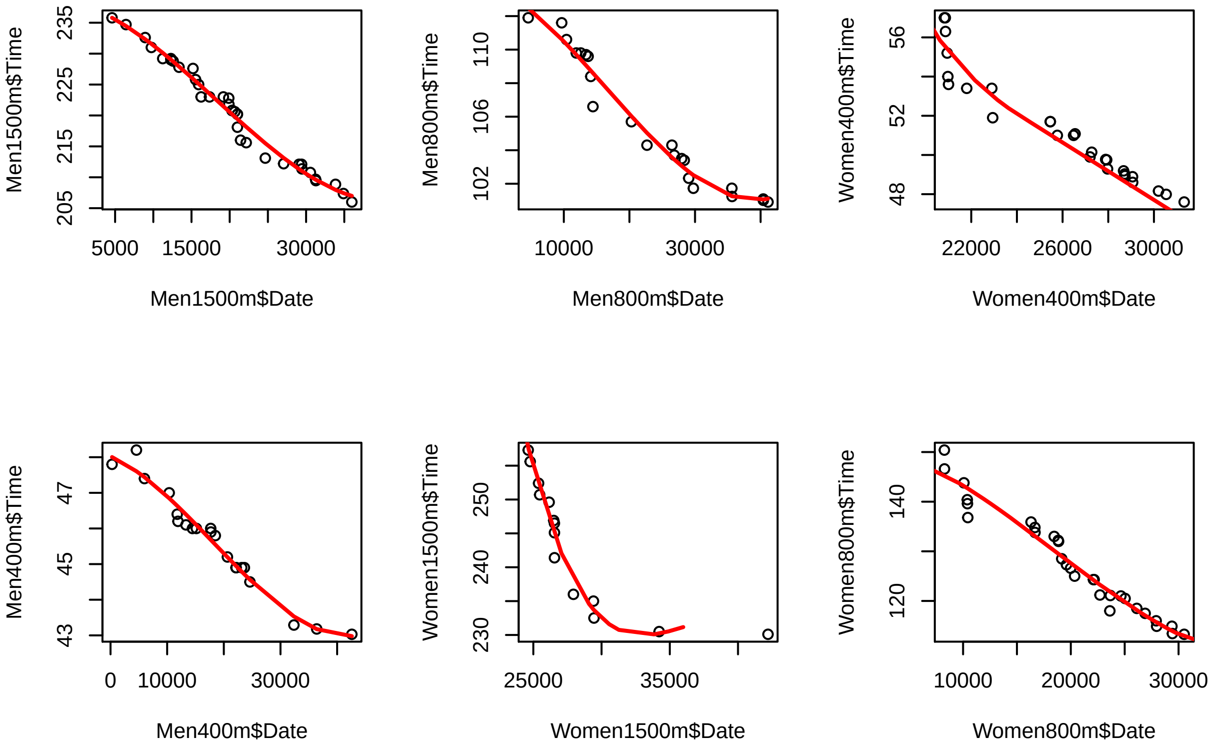




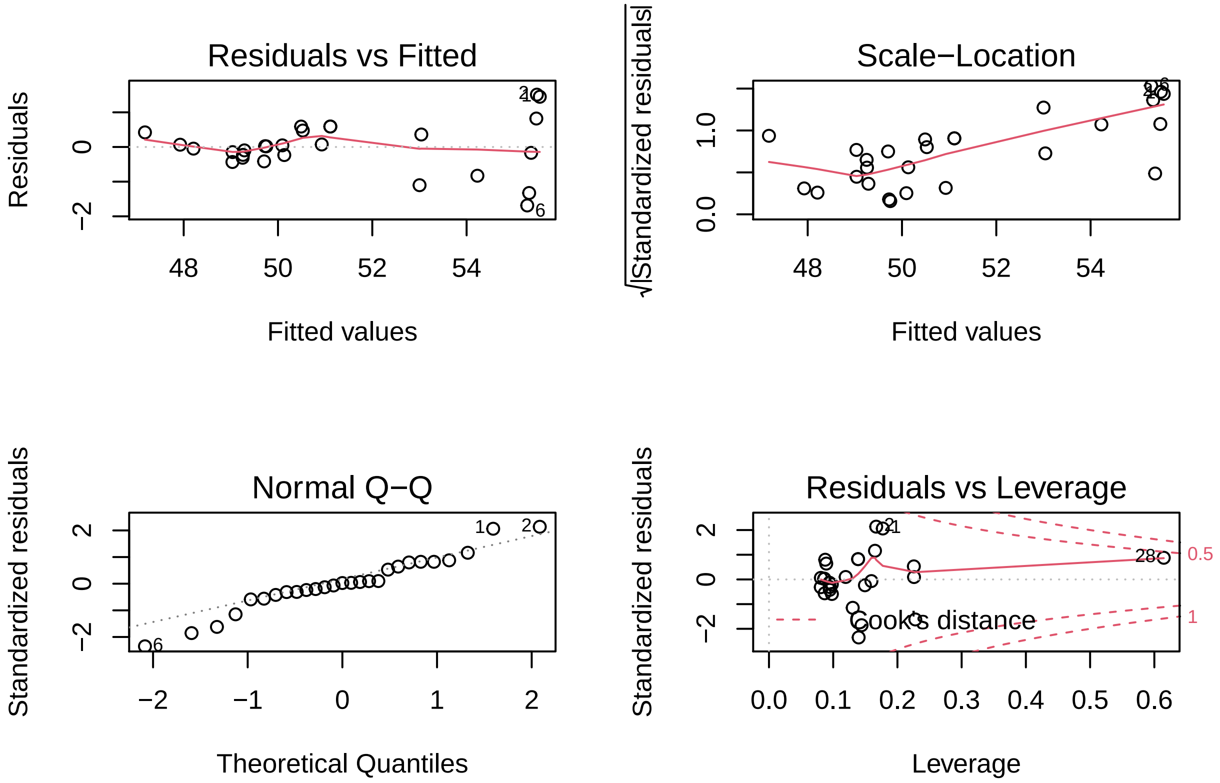
It seems that log transformation works well on Men1500m, Men400m, but does not fit women1500m, Women400m well, because there seems exist some curvature in above dataset.

3.2.2 Polynomial models

To fit polynomial models, we will first need to transform time into elapsed time from a starting point (say 1900-1-1).



Polynomial models seems to fit the data well. The next step is to check for model assumption. For example for Women1500m,

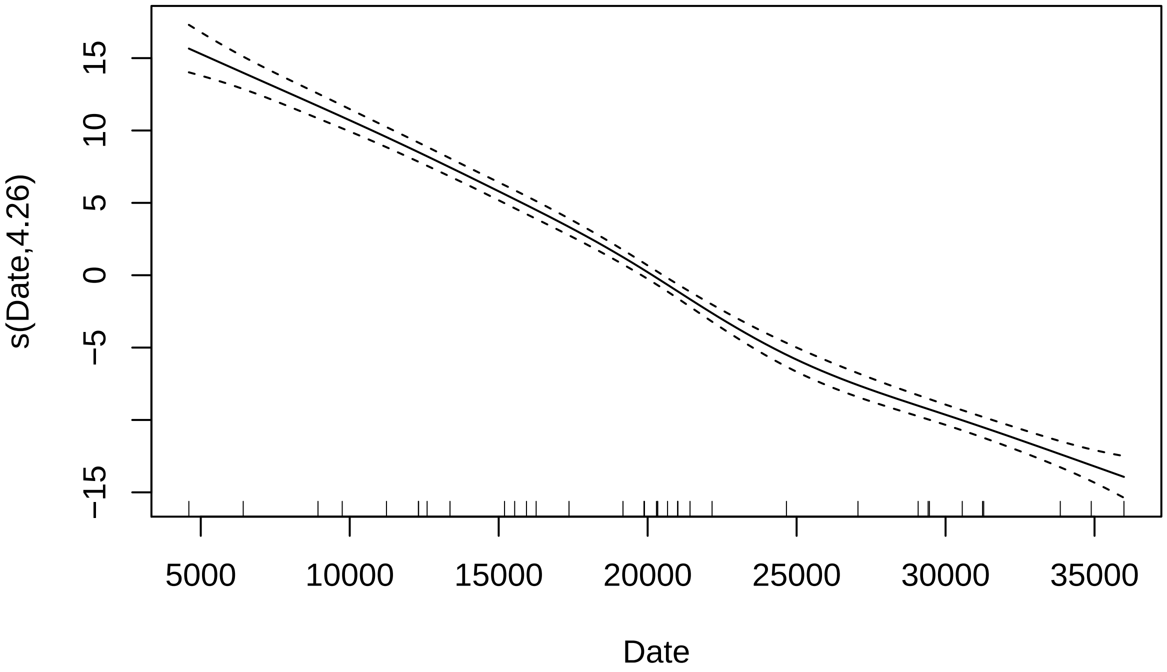


, we find that residuals are not independent around the horizontal axis, which might indicates polynomial models are also not appropriate for the data.

3.2.3 Generlized additive models

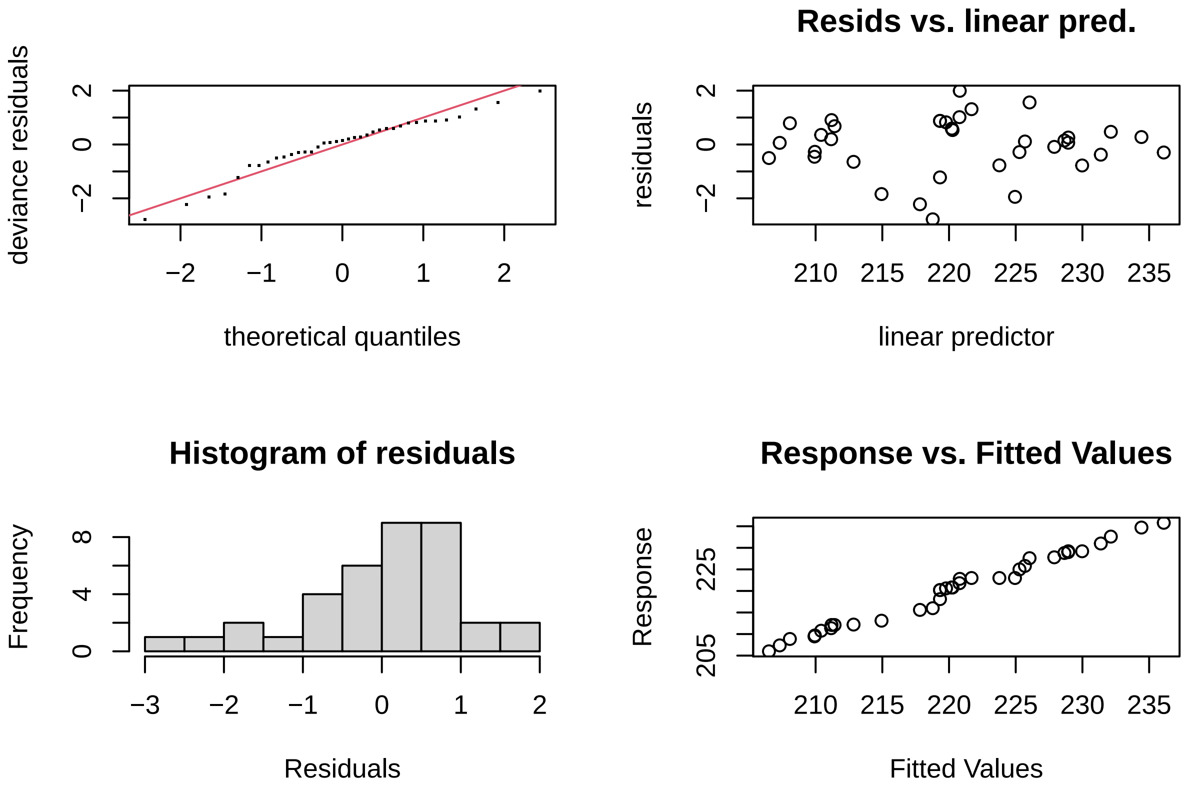
To choose the best-fitting GAM model for separate events, we use REML method to choose the smooth parameter for the GAM model. To compare models with different explanatory variables, we use AIC as selecting criteria.

The best fitting GAM model we choose for Men1500m is the following, as GAM model Time~s(Date) contains the lowest AIC, and largest adjusted R square, comparing with the Time~s(Date)+s(age), Time~s(Date)+s(age)+s(Altitude).



As output of the best-fitting model shows, there is an overall negative trend between Time and Date. Then, we start to check model assuption for the fitted Men1500m data.

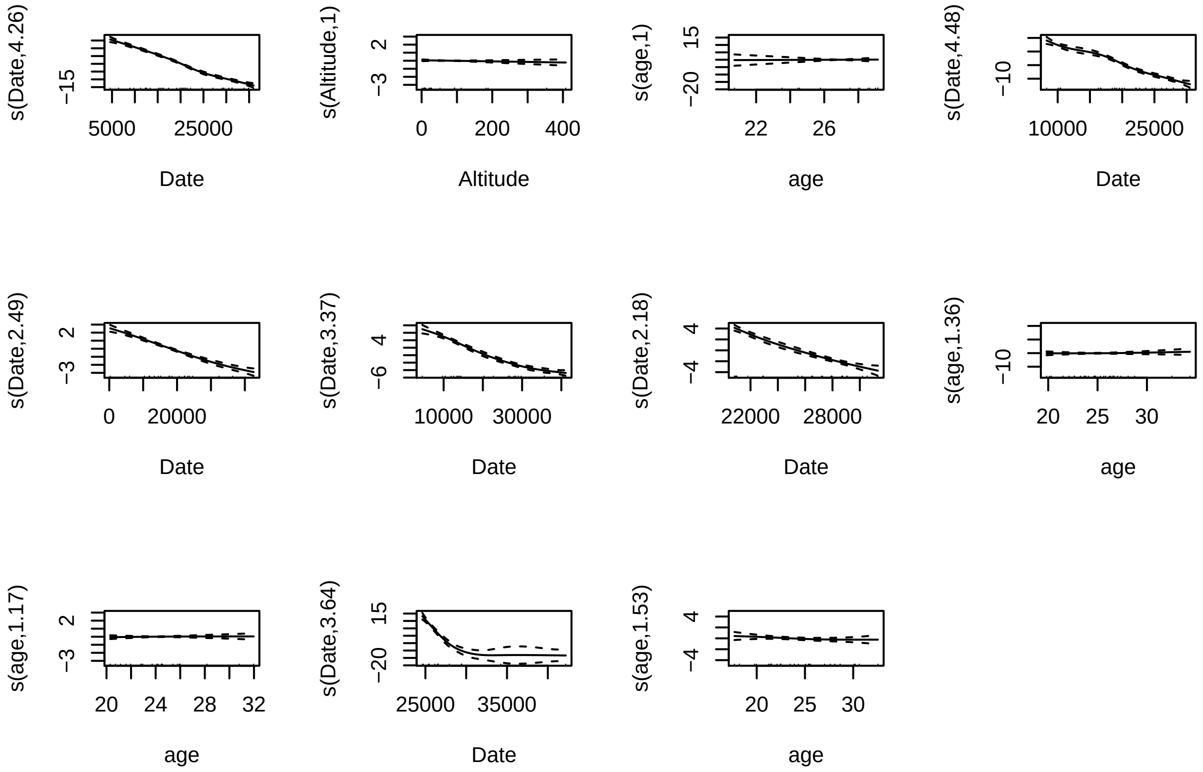
```
Family: gaussian
Link function: identity

Formula:
Time ~ s(Date)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 220.4376     0.1814    1216   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(Date) 4.256  5.191 374.9  <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

R-sq.(adj) =  0.982   Deviance explained = 98.4%
-REML = 60.593  Scale est. = 1.2169    n = 37

[1] 119.107
```



```
Method: REML   Optimizer: outer newton
full convergence after 4 iterations.
Gradient range [-2.705098e-06,2.081667e-07]
(score 60.59254 & scale 1.21686).
Hessian positive definite, eigenvalue range [0.1236008,17.65256].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

          k'  edf k-index p-value
s(Date) 9.00 4.26    0.42  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Normal Q-Q plot Shows that the residuals are almost normally distributed. Resids vs linear pred. plot shows that are residues are almost evenly distributed along the horizontal axis. Histogram of residuals shows that residuals are almost normally distributed. However, it is slightly left-skewed which might due to small sample size (38).

- Similarly, Time ~ s(Date) + s(age) + s(Altitude, bs = "cr", k = 3) is fitted for Men400m data; Time ~ s(Date) is fitted for Men800m data after checking AIC, R square, residual plots.
- Time ~ s(Date) + s(age, bs = "cr", k = 3) is fitted for Women1500m data, Time ~ s(Date) + s(age) is fitted for Women400m data, Time ~ s(Date) + s(age) is fitted for Women800m data



- As we can see, the derivative(dropping speed) varies as Date(days) increases. In particular for Men1500m data, the dropping speed is higher at Dates in between 15000~25000 days starting from 1900-1-1, which indicates world records for Men1500m develop rapidly around the year 1941 to year 1968.

3.3 Prediction for future records

In order to predict when future world records will happen, Date ~ s(Time) is fitted to evaluate the possible future world record date. For example, if we define next world records improves by 1 second. For Men1500m, we are able to explore when the next world records will happen (205,204,203,202,201,200) using the following piece of codes.

```
        1         2         3         4         5         6 
42225.41 41284.22 40343.03 39401.83 38460.64 37519.45 
```

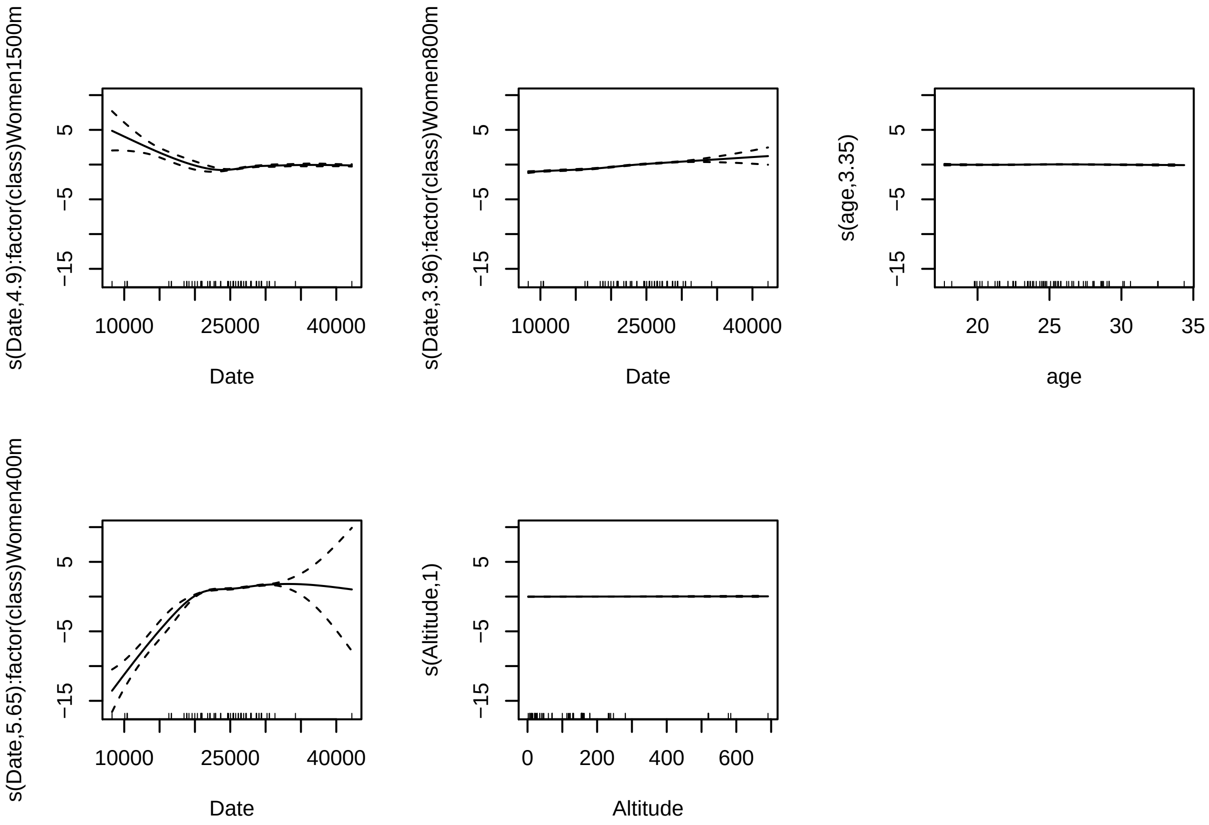
According to the result, for Men1500m world records, for 205s the record would probably appears at 2002. However, we noticed that the current world records for Men1500m is still 206s which is produced in 1974, which might indicates that 206s might be a threshold for male records at 1500m. Other explanatory variables, such as anthropometric parameters are needed for a more accurate prediction.

3.4 Pattern of progress for each event

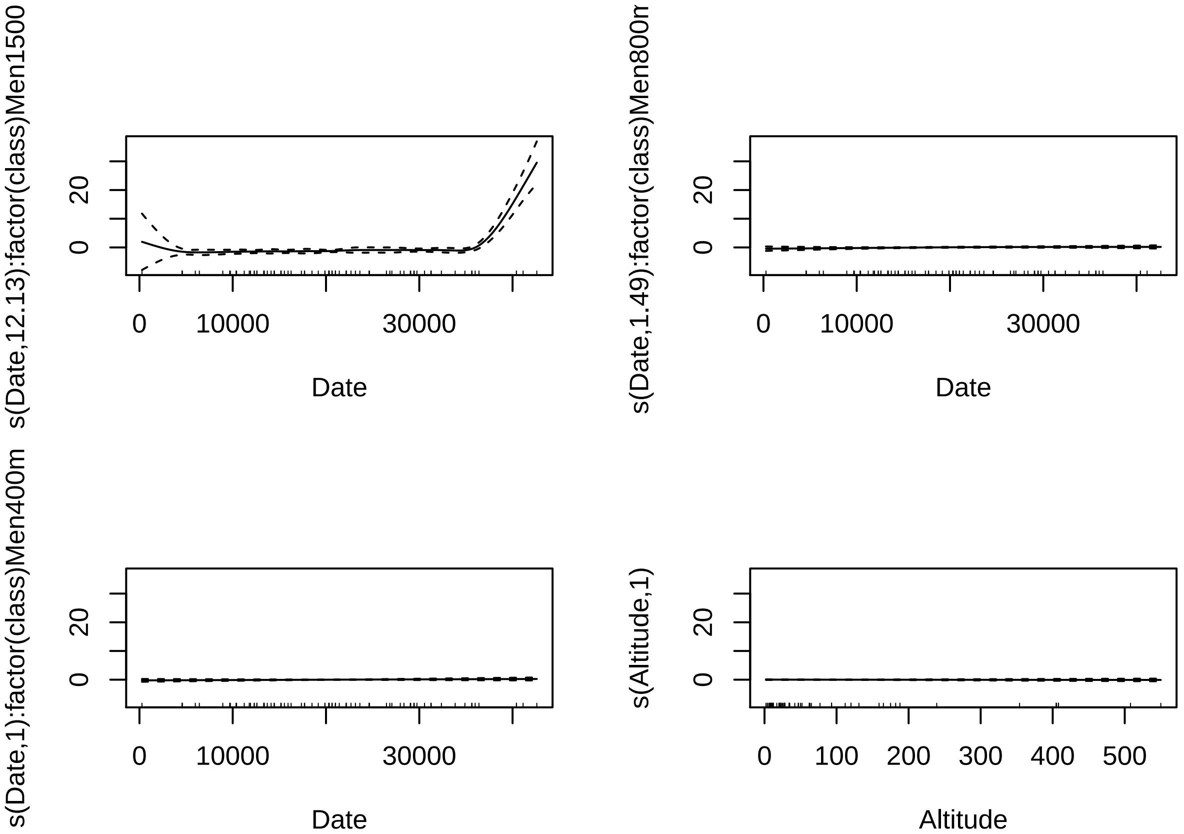
In order to compare the patterns of progress for each event, we start by combining data for Men and Women in three events.

There are three potential models that we can consider: Speed ~ s(Date, by = factor(class)) Speed ~ s(Date, by = factor(class)) + s(Altitude) Speed ~ s(Date, by = factor(class)) + s(Altitude) + s(age)

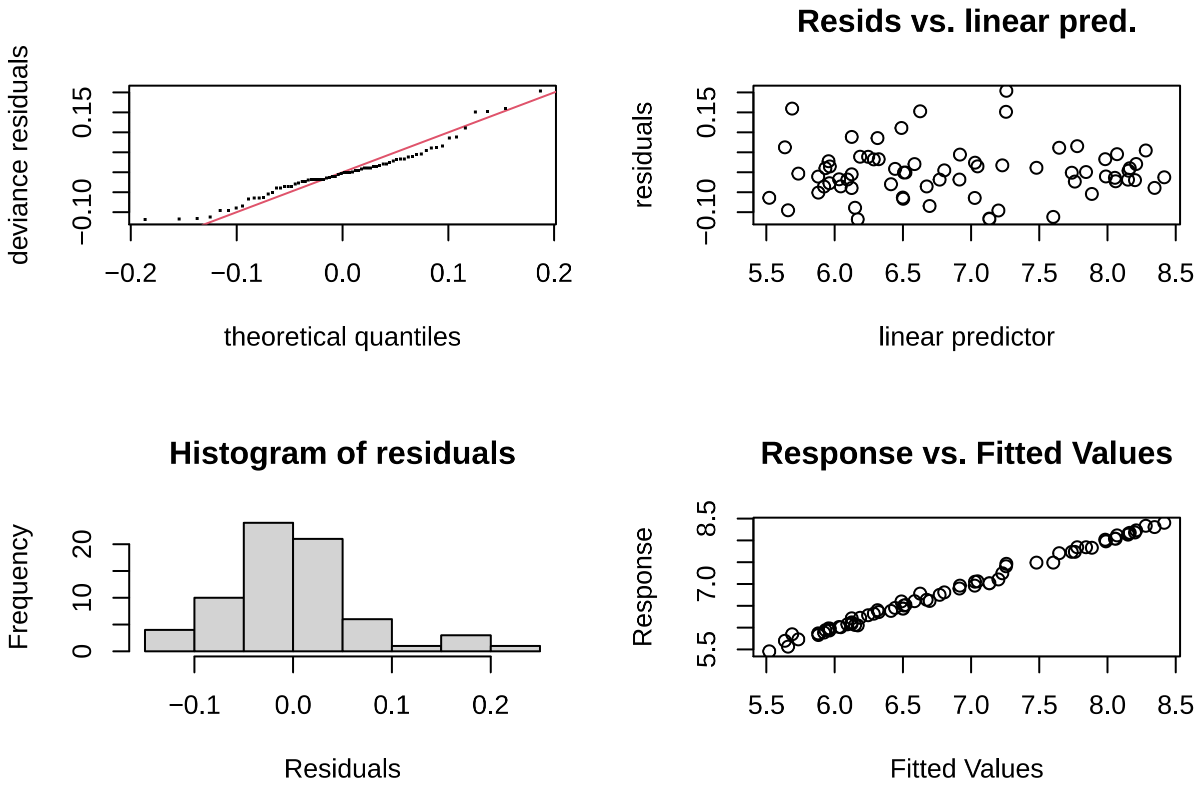
- If we choose the model by lowest AIC, model Speed ~ s(Date, by = factor(class)) + s(Altitude) + s(age) has lowest AIC -143.5269357, and largest R square adjusted.
- Similarly, we choose model Speed ~ s(Date, by = factor(class), k = 20) + s(Altitude) for its lowest AIC 144.6763988, and largest R square adjusted.



According to the plot, we could explore that relationship between Speed of world records and other explanatory variables. For women, we can see that the increasing speed of "Speed" variable even for both 400m,800m,1500m events. In addition to this, Altitude and age seems affect a little for these events in the plot. It might because most athletes are in early ages.



For men athletes , we find similar patterns that increasing speed of "Speed" variables are even in 400m, 800m, 1500m events. If we check the model assumption models, we could find model assumptions are well-fitted for the combined data models.



```
Method: REML   Optimizer: outer newton
full convergence after 11 iterations.
Gradient range [-1.694177e-05,2.853802e-06]
(score -39.12782 & scale 0.005796306).
Hessian positive definite, eigenvalue range [1.694125e-05,32.41615].
Model rank =  46 / 46

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                                  k'  edf k-index p-value
s(Date):factor(class)Women1500m 9.00 4.90    0.83   0.065 .
s(Date):factor(class)Women400m  9.00 5.65    0.83   0.060 .
s(Date):factor(class)Women800m  9.00 3.96    0.83   0.085 .
s(Altitude)                     9.00 1.00    0.91   0.205
s(age)                          9.00 3.35    0.95   0.275
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the first normal Q-Q plot, we can see that residuals are almost normally distributed. The histogram shows that residuals are almost bell-curved and has mean zero. Thus we conclude that model assumptions are well-fitted.

From the output of female athletes models, we can see that smoothing term of Date for each events is almost significant (0.055,0.8,0.055), edf of s(Altitude) is 1, which indicates that variable Altitude is a linear fit in the model.

3.5 Pattern of progress for each event

In order to compare the patterns of progress for each event, we start by combining data for each events.

3.6 Conclusions for specific questions

4 Chapter 4 - Conclusions and discussion

4.1 Summary of conclusions to all questions of interest

4.2 Discussion of any limitations of data and/or analysis

- Prediction of world records has limitations, such as human body limits. Other explanatory variables, such as anthropometric parameters are needed for a more accurate prediction.
- Windspeed is rarely recorded for distance above 200m, which affects athlete's running speed
- Size of the the dataset is too small for GAM's cross-validation in order to avoid overfitting.

4.3 Further analysis you did not have time to carry out