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

Jinda(Cecil) Zhang

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.
- 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 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 used to investigate the relationship between world records and date.

A **linear model** (LM), describes a quantitative response in terms of a linear combination of predictors (Faraway, 2009). Suppose we want to model the response Y in terms of three predictors, X_1,X_2, and X_3, we usually have to assume that it has some more restricted form, for example:

$$\widehat{\mathrm{y}} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \mathrm{X}_1 + \widehat{\beta}_2 \cdot \mathrm{X}_2 + \widehat{\beta}_3 \cdot \mathrm{X}_3$$

where

- β_{i}, i = 0,1,2,3 are unknown parameters.
- β_0 is called the intercept term.

Polynomials are widely used in situations where the response is curvilinear, as even complex nonlinear relationships can be adequately modeled by polynomials over reasonably small ranges of the x's. For example, the second-order polynomial regression in one variable can be represented as:

$y = \beta_0 + \beta_1 \cdot \mathrm{x} + \beta_2 \cdot \mathrm{x}^2 + \epsilon$ (Montgomery, Peck and Vining, 2012).

A **generalized additive model** (GAM) is a model with a linear predictor involving a sum of smooth functions of covariates. In general the model has a structure such as: $f(x_1, x_2, \ldots, x_n) = x_1^2 + x_2^2 + \cdots + x_n^2$

$$g(\mu_{\mathrm{i}}) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots$$

where

- $\mu_{\mathrm{i}} = E(Y_i)$ and Y_i follows some exponential family distribution.
- Y_i is a response variable, X_i is a row of the model matrix for any strictly parametric model components, θ is the corresponding parameter vector, and the f_j are smooth functions of the covariates, X_k. (Wood, 2006)

The **motivation** of using **Restricted Maximum likelihood** (REML) method is to fit a GAM model in order to avoid over-fitting for data with small sample size, which is only 38 in the athletic data set. To choose the best-fitting GAM model for separate events, we use REML method to choose the smooth parameter for the GAM model.

The idea underlying REML estimation was put forward by M. S. Bartlett in 1937. The first description of the approach applied to estimating components of variance in unbalanced data was by Desmond Patterson and Robin Thompson of the University of Edinburgh in 1971. REML approach is a particular form of maximum likelihood estimation that does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function calculated from a transformed set of data (Dodge, Yadolah, 2006).

3 Analysis of the Data

3.1 Exploratory data analysis

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	235.250	245.8	250.425	257.3
M	38	220.2	8.2	206.0	212.125	220.8	227.150	235.8

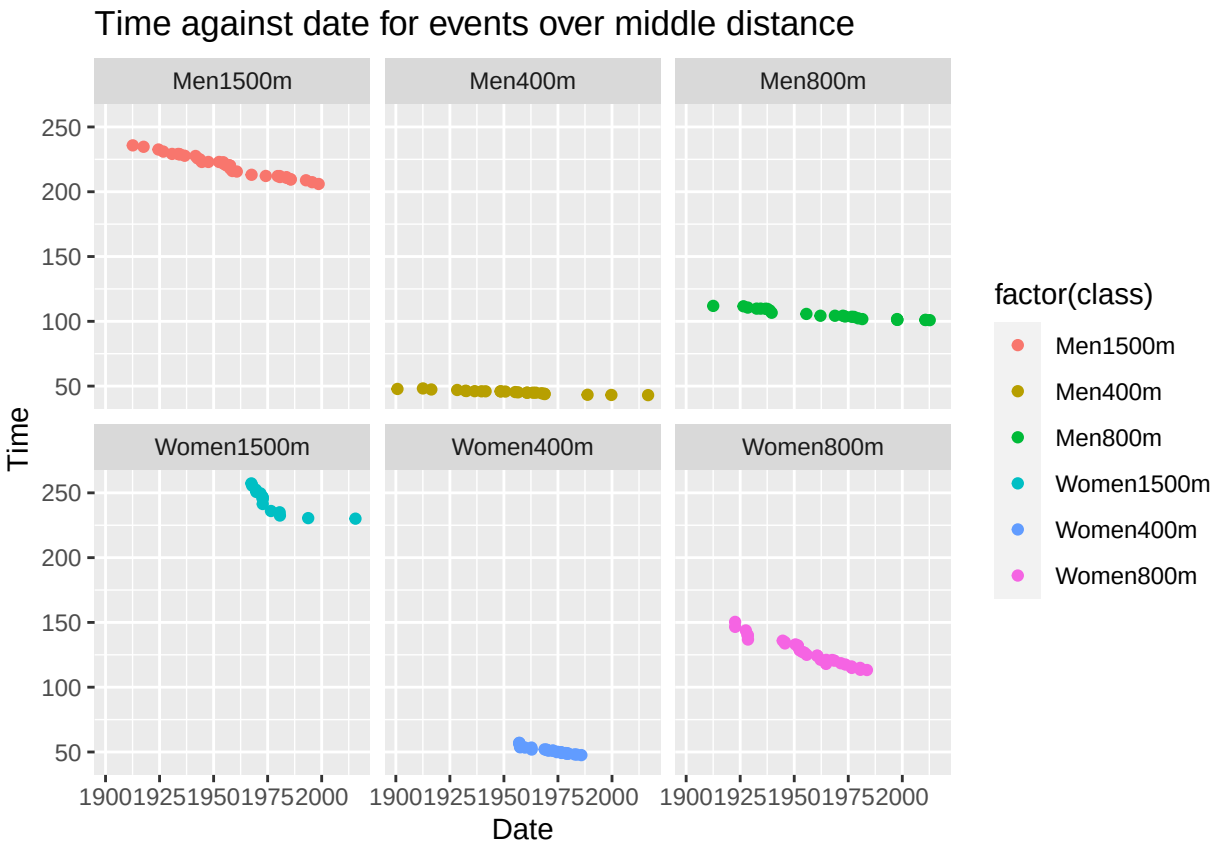
Table 2: Summary statistics on record Time by sex of 400m events.

sex	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
F	28	51.3	2.7	47.60	49.15	51.0	53.400	57.0
M	24	45.5	1.4	43.03	44.80	45.6	46.125	48.2

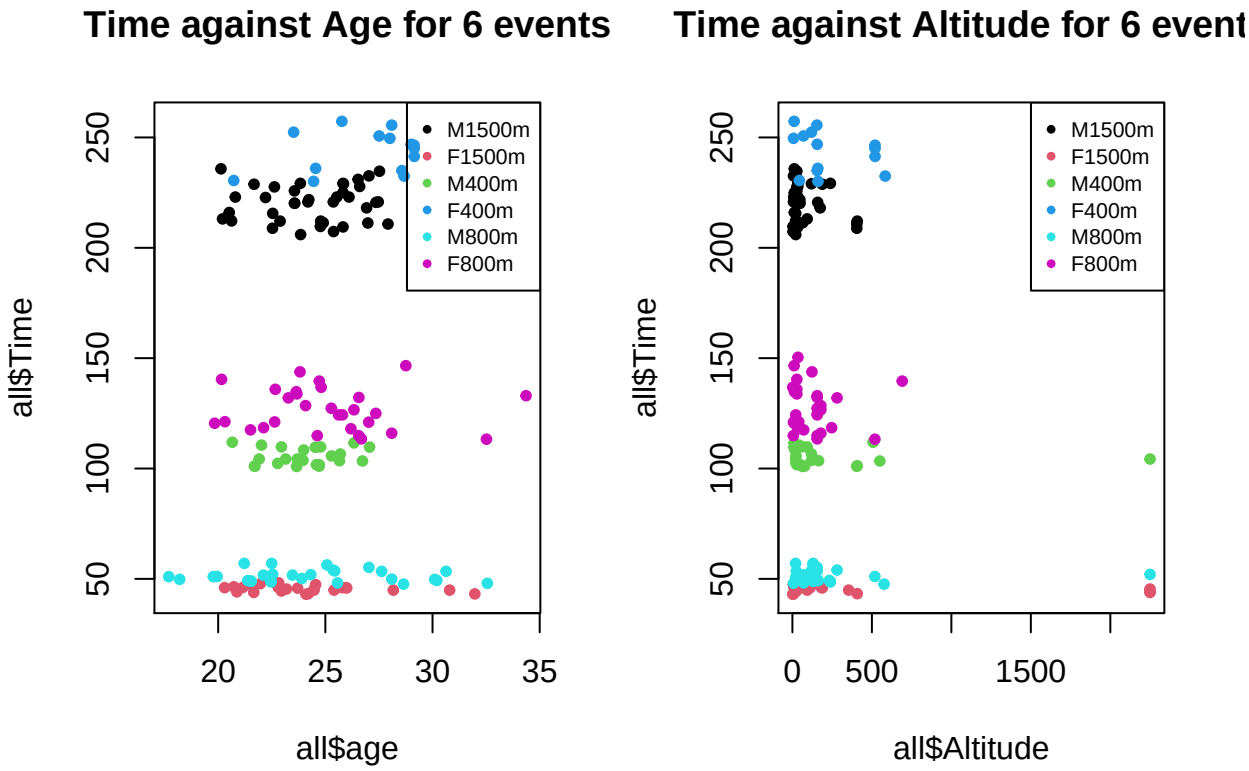
Table 3: Summary statistics on record Time by sex of 800m events.

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	24	105.3	3.8	100.91	101.73	104.3	109.625	111.9

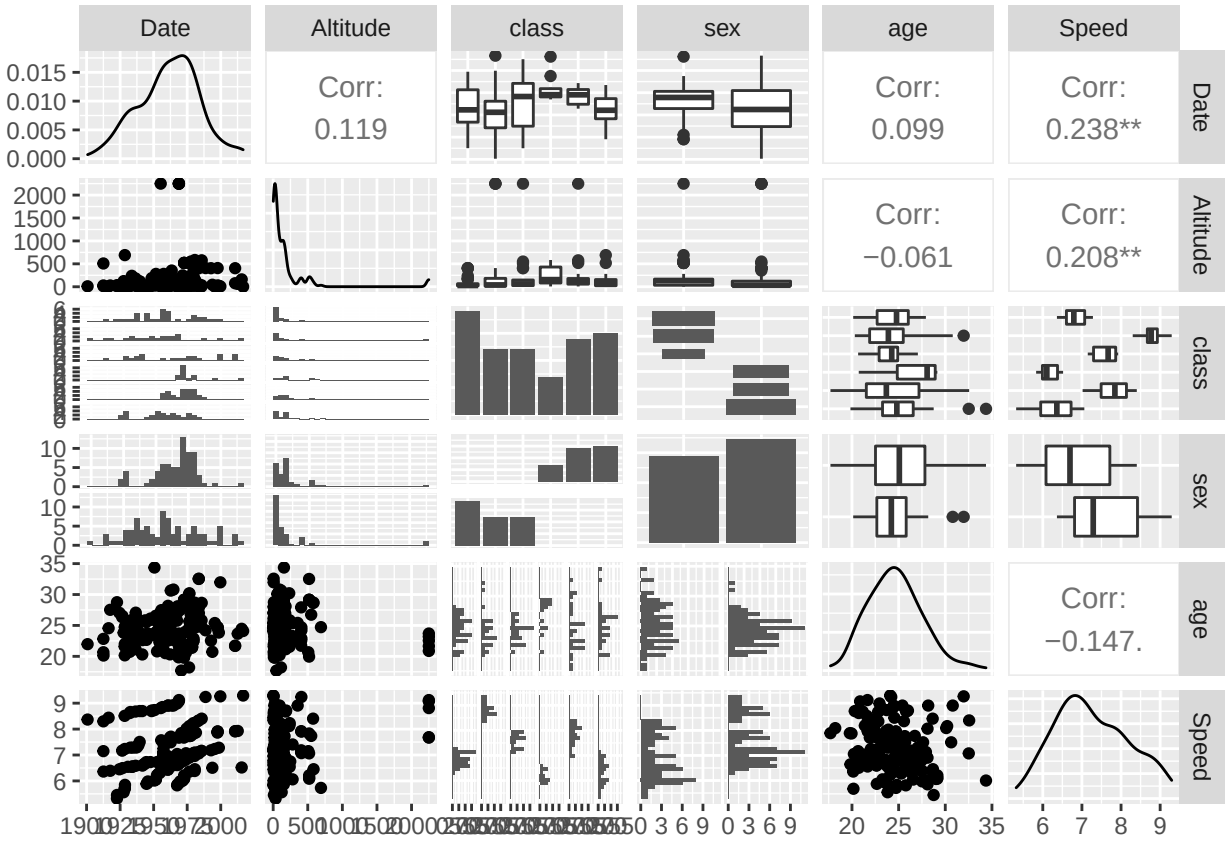
The above tables show that for each event separately, it can be seen that male are in average spend less time than female in all three 400m, 800m, 1500m events. The maximum and minimum of time for male events is both smaller then female events. In addition to this, it can seen that the standard deviation for male is smaller than female, which might indicates that female athletes has a faster progression rate than male athletes.



From the plots, we could see that the variable `Time` decreases as `Date` increases. In addition to the overall decreasing trend, it seems that there is a non-linear relationship for between `Time` and `Date` for female events, which indicates that linear models might not be appropriate to fit the data. More complicated model might be required to fit the data with curvatures.



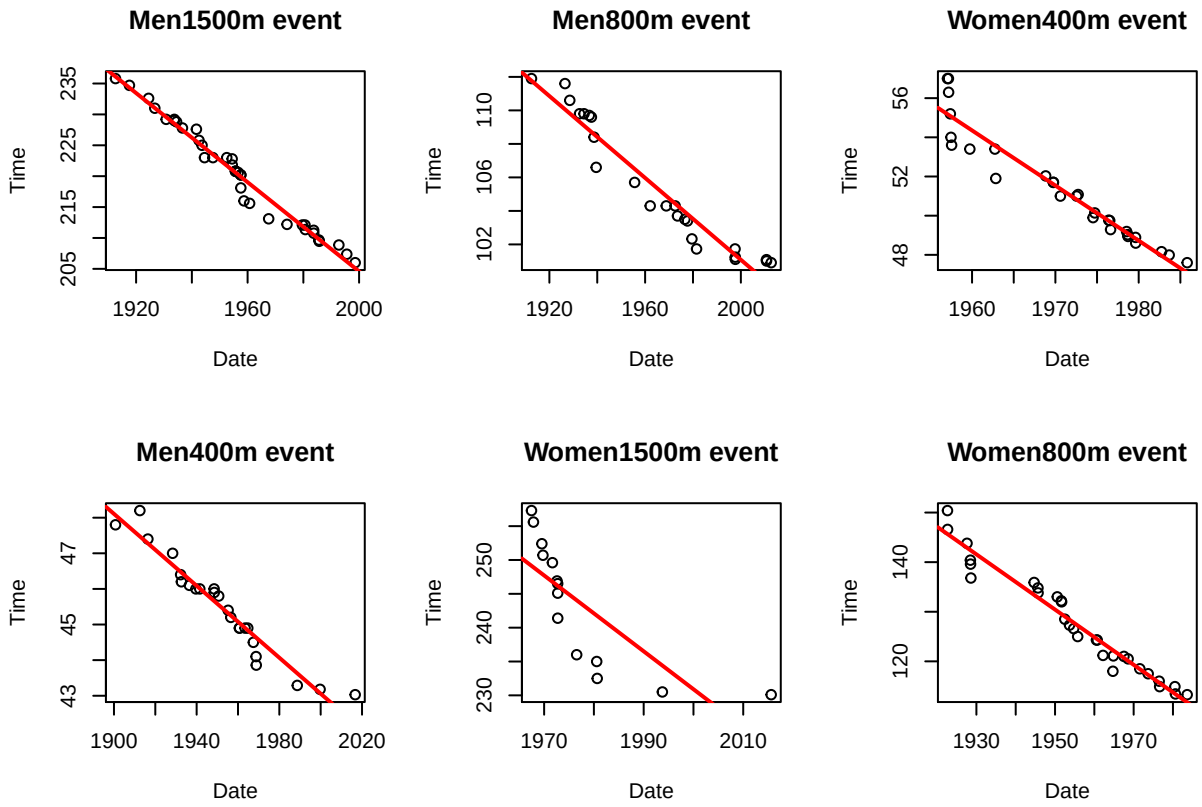
From the plot, we could see that for all six events, most competitions happens in city that has altitude less 500m. Data points that has altitude greater than 1500m might be considered as outliers and needs futher consideration while fitting the model. For all six events, we find that most athletics are among 20-35 years old, which are in their young age. In addition to this, variable 'Age' seems does not has a large effect on 'Time'.



The pairs plot shows that `Speed` and `Altitude` both have weak negative correlation with `age`, which are -0.147 and -0.061, respectively. `Speed` and `age` both have weak positive correlation with `Date`, which are 0.238, 0.099 separately. In addition to this, it is also worth noting that there is no strongly correlated pairs of variables.

3.2 Analyses and model checks

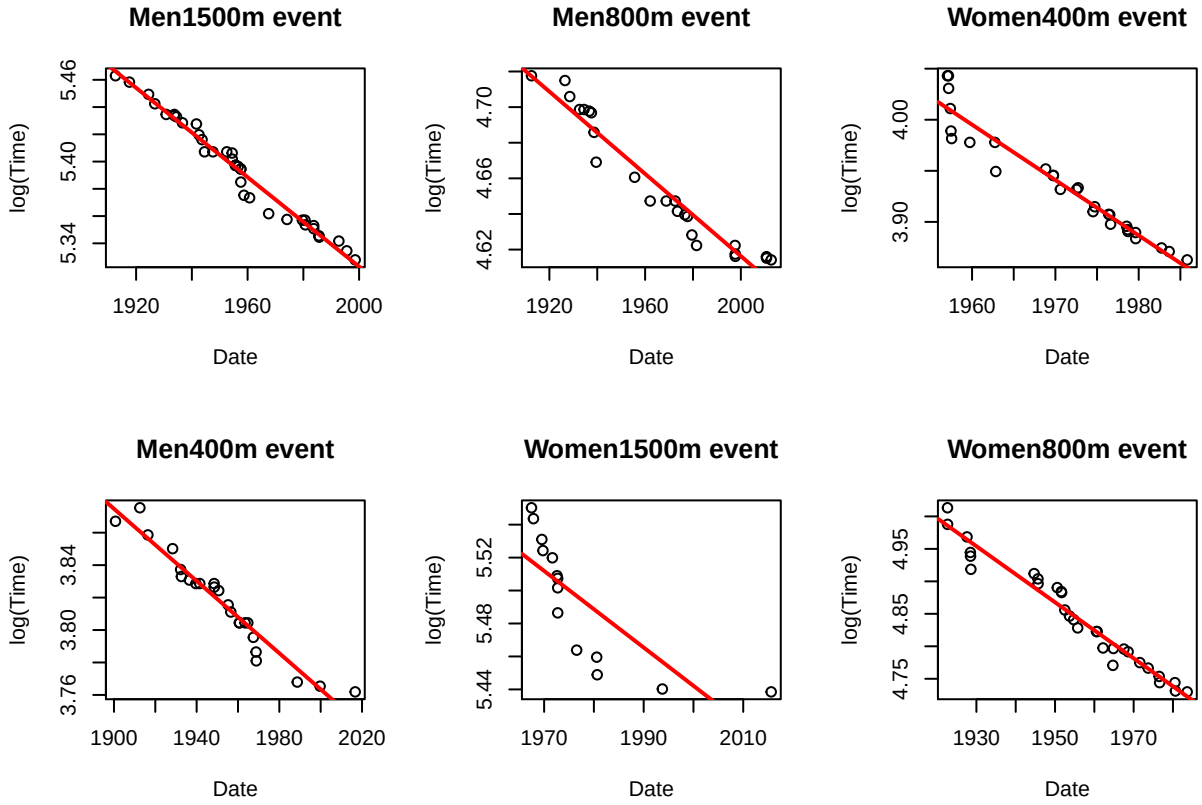
3.2.1 linear models



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

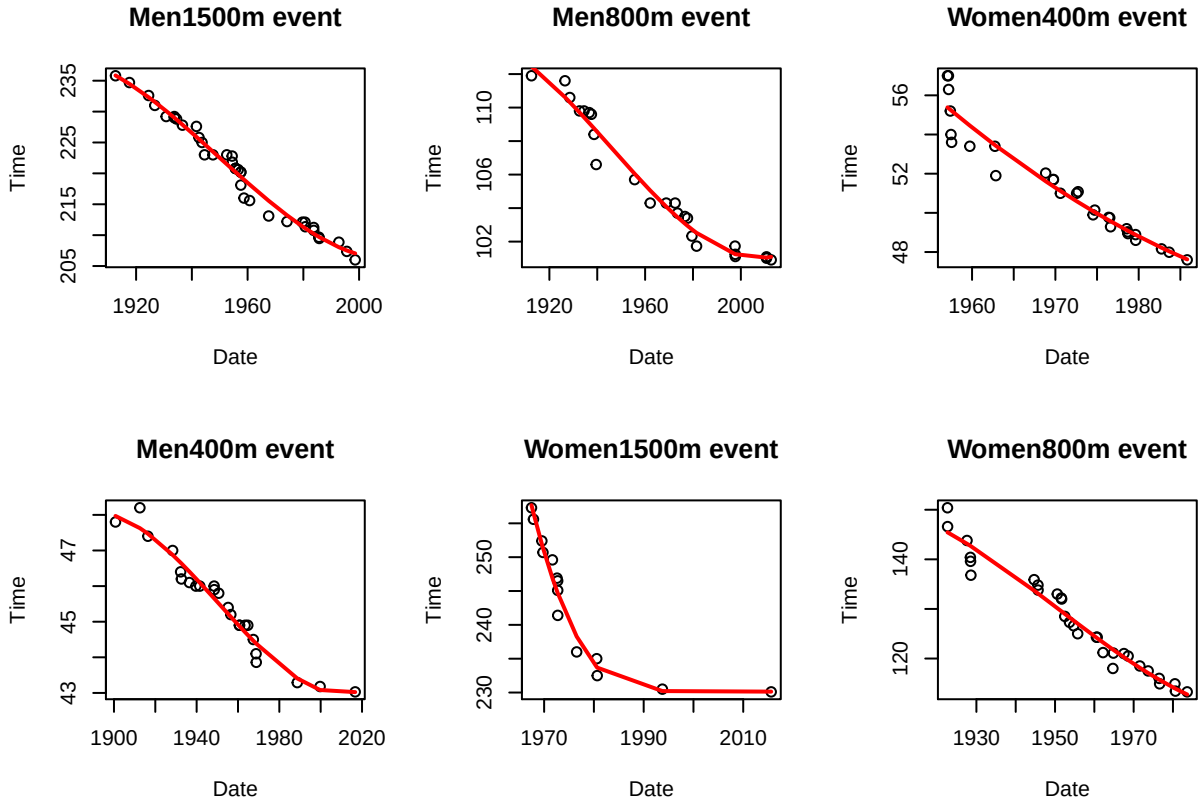
From the plot, we can see that linear model fit the data well for Men1500m data, and Women800m data. However, it does not fit the data well for Women400m and Women1500m data. Since linear models cannot fit the data with curvature well, what about log-linear models?

3.2.2 log-linear models

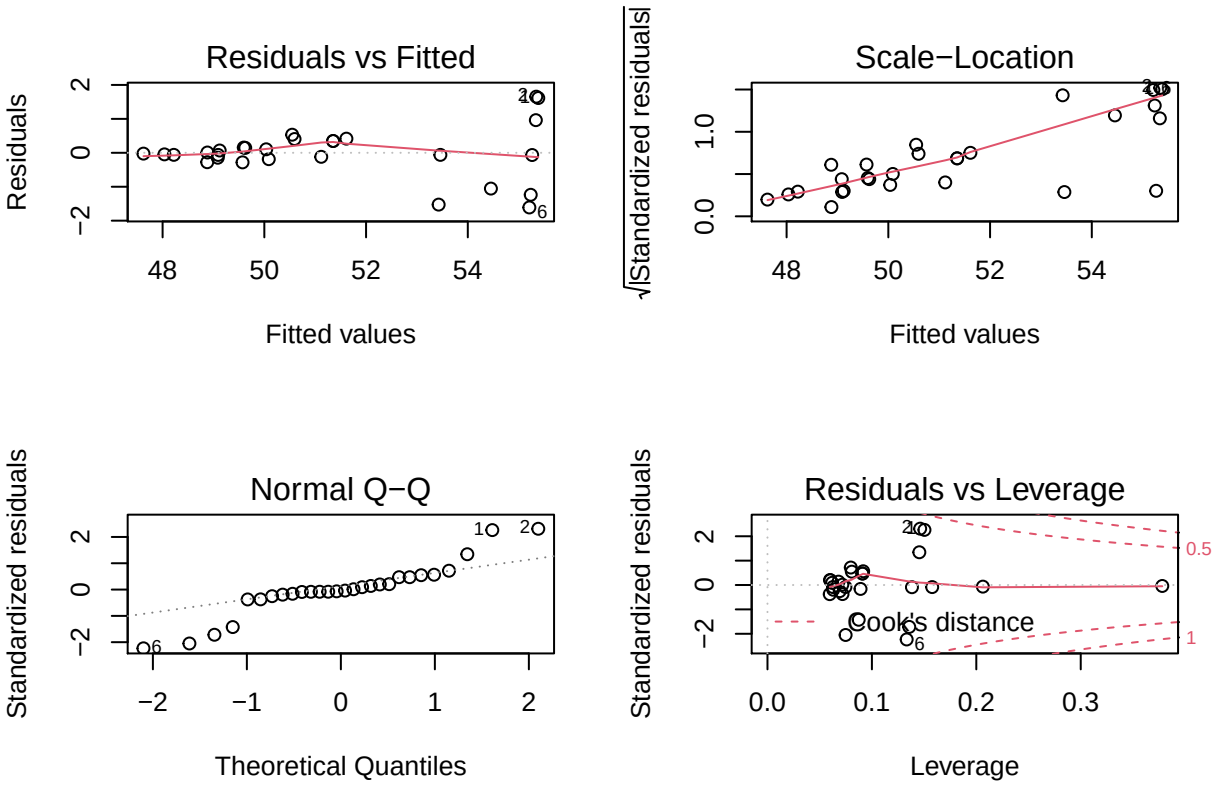


From the plot we could see that the log transformation does not help a lot in terms for model fitting. It seems that log transformation works well on Men1500m, Men400m, but does not fit women1500m, Women400m well, because there still exist some curvature in the above data.

3.2.3 Polynomial models



From the plot, we could see that a polynomial models seems to perform better than linear models. However, the models seem not to fit Men1500m event, Women400m event, Women1500m event well. The next step is to check for model assumption.



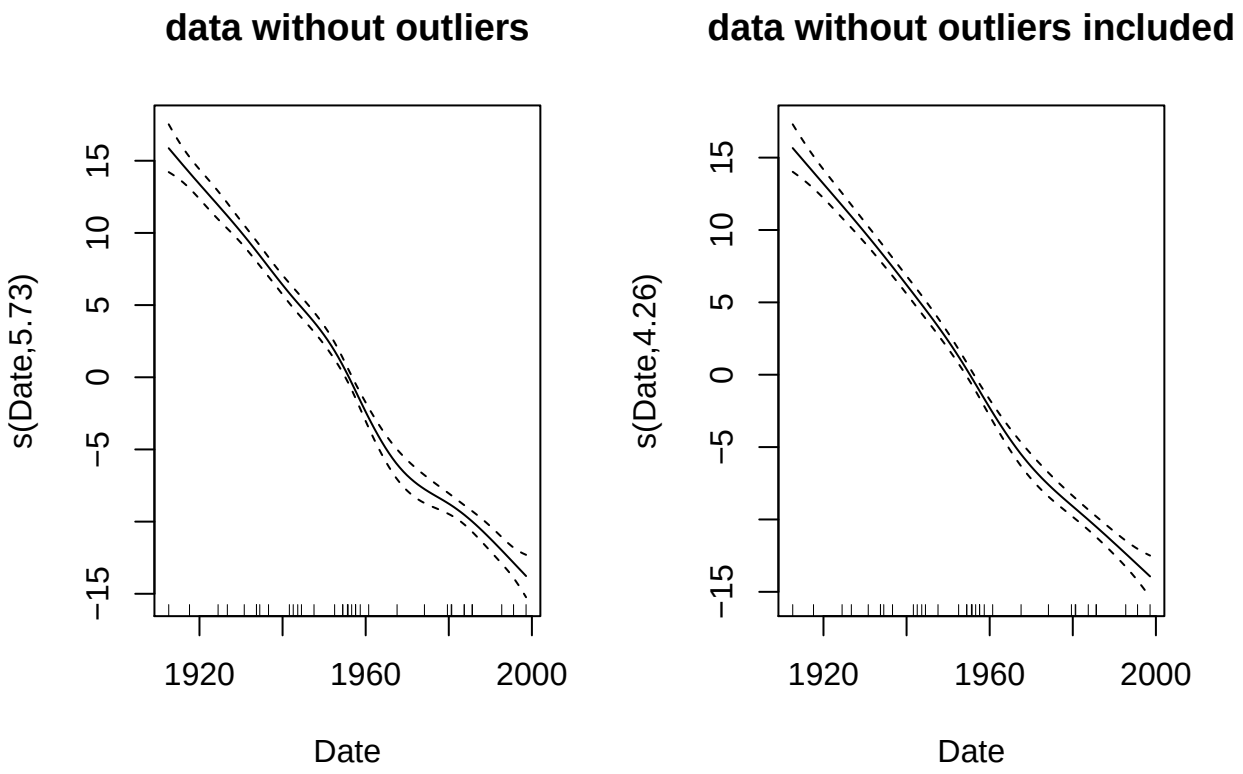
Take Men 1500m dataset as an example, we find that residuals are not independent around the horizontal axis, which might indicates polynomial models are also not appropriate for the data. It is worth noting that other dataset shows similar problems on polynomial models.

3.2.4 Generlized additive models

The **motivation** of using Restricted Maximum likelihood (REML) method is to fit a GAM model in order to avoid over-fitting for data with small sample size, which is only 38 in the athletic data set. 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). To investigate whether the data point that has *Altitude* greater than 1000m are potential **outliers**, we compare the plot that contains potential outliers and without potential outliers.



From the plot, we can see that the outlier does not affect the model greatly, and the conclusion remains the same. As a result, we decide not to remove outlier for Men1500m data, which is also the similar situations for other events.

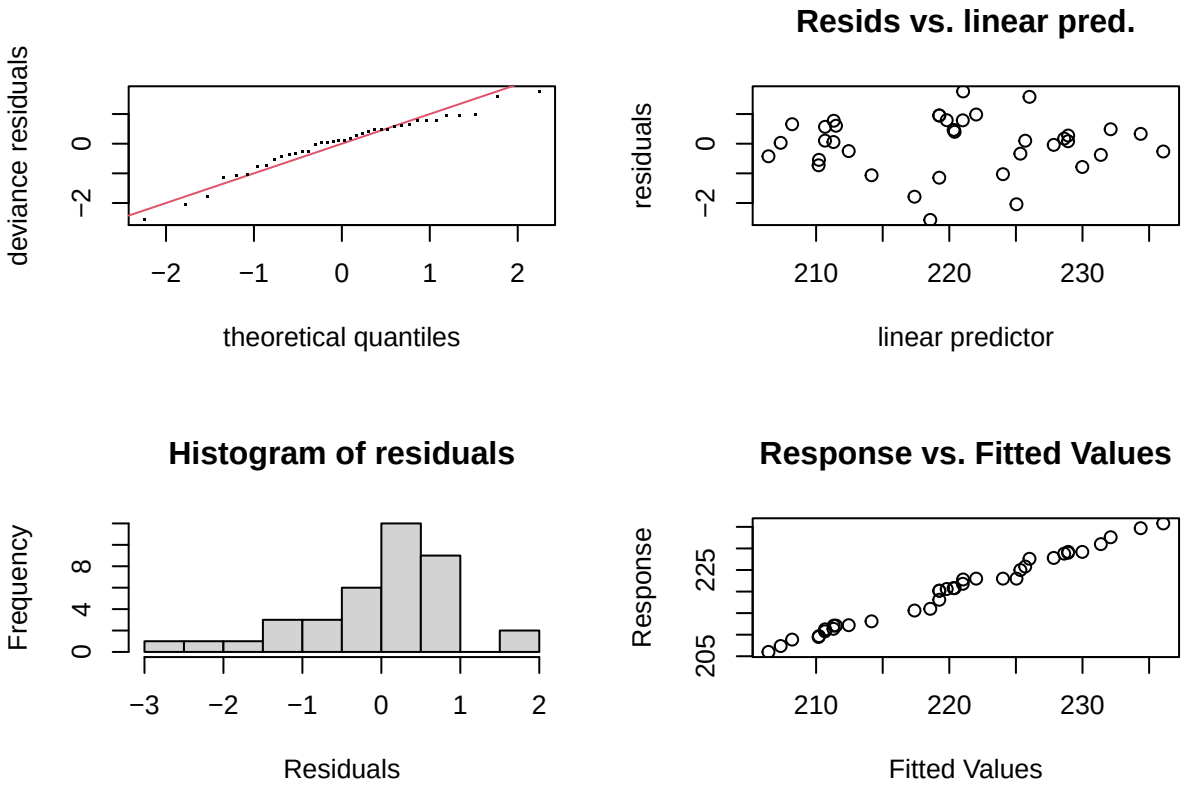
As output of the best-fitting model shows, there is an overall negative trend between Time and Date. As we can see, the derivative(dropping speed) varies as `Date` increases. In particular for Men1500m data, the dropping speed is higher at Dates in between year 1960 to year 1980, which suggest that the progression speed of Men1500m speed is faster during 1960 to 1980 than the rest of time.

```
Family: gaussian
Link function: identity

Formula:
Time ~ s(Date)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 220.1955     0.1639    1343   <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) 5.733  6.835 352  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R-sq.(adj) =  0.985   Deviance explained = 98.7%
-REML = 61.841  Scale est. = 1.0213    n = 38
```

```
[1] 116.6968
```

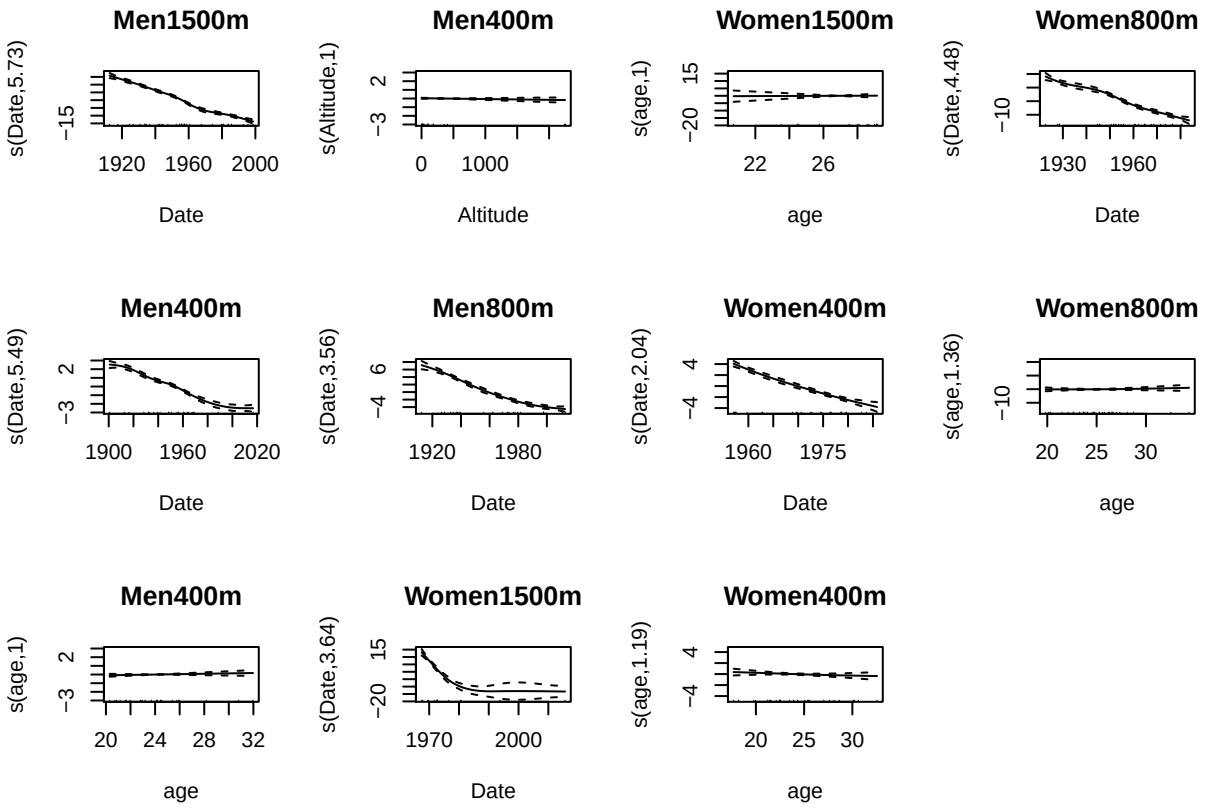
The output of summary statistics shows that smoothing term for variable `Date` is significant, and the model has high adjusted R-square with 0.985, and high deviance explained, which is 98.7%. Then, we check model assumption for the fitted Men1500m data.




```
Method: REML   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-2.053753e-05,1.378265e-05]
(score 61.84145 & scale 1.021337).
Hessian positive definite, eigenvalue range [0.1951777,18.31452].
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 5.73    0.49  <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).

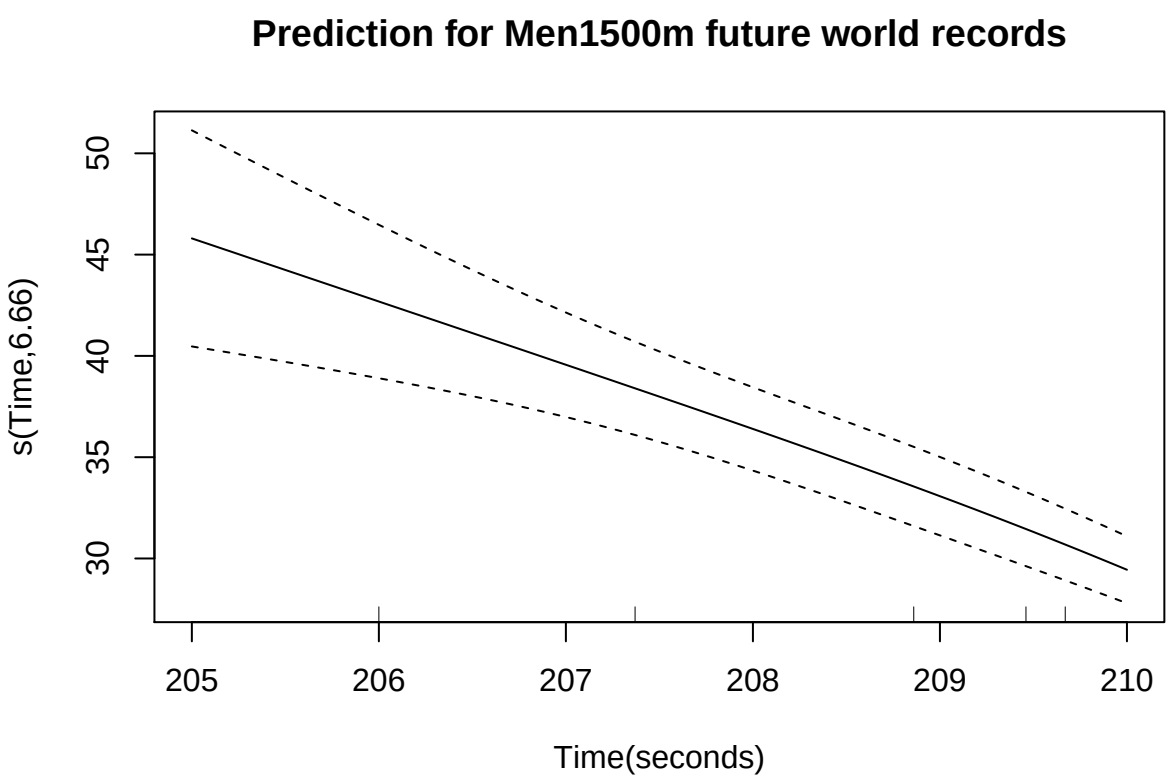
- Similarly, Time ~ s(Date) + s(age) + s(Altitude, bs = "cr", k = 3) is fitted for Men400m data; Time ~ s(Date) is fitted for Men800m data after checking AIC, R square, residual plots.
- Time ~ s(Date) + s(age, bs = "cr", k = 3) is fitted for Women1500m data, Time ~ s(Date) + s(age) is fitted for Women400m data, Time ~ s(Date) + s(age) is fitted for Women800m data. Outliers are included for Men1500m, Men400m, Men800m, Women1500m, Women400m, Women800m data after examined impact of including / excluding the outlier and that it made no difference.



- As we could see, `Time` decreases as `Date` increases for all other 6 events. In addition to this, the derivative(dropping speed) varies as Date increases. Altitude and age seems to have little impact on the Time.

3.3 Prediction for Future Records

In order to predict the date of future world records, Date ~ s(Time) is fitted to evaluate the possible future world record date for Men1500m event. If we define next world records improves by 1 second, which is observed by pattern of the previous data set. For example, for Men1500m event, now that the world record is 206 second, if we predict the $Time$ of next world records to be 205s, 204s, 203s, 202s, 201s, 200s, and we want to find $Date$ corresponding the above $Time$.




```
$`Possible future records(second)`
[1] 205 204 203 202 201

$`Date of possible future records`
[1] 2002.545 2005.660 2008.776 2011.891 2015.007
```

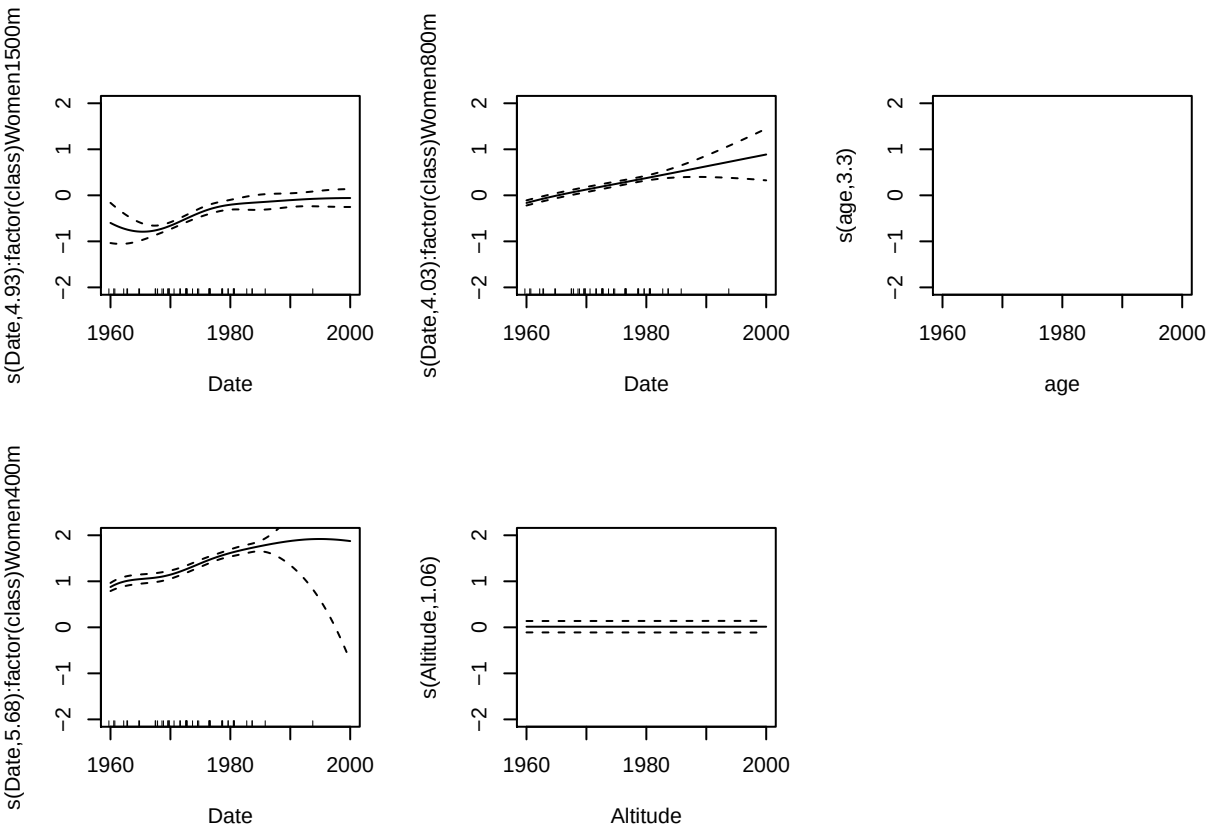
As the result above shown, for Men1500m world records, the next world record, which is 205s would probably appears at year 2002. However, we noticed that the current world records for Men1500m is still 206s which is produced in 1974, which might indicates that 206s might be a threshold for male world records at 1500m. In addition to this, other explanatory variables, such as anthropometric parameters might be needed for a more accurate prediction.

3.4 Pattern of Progress for Male/Female

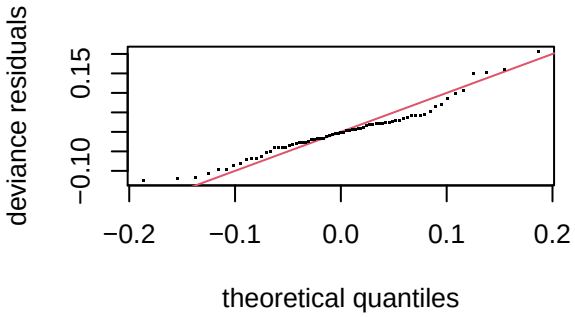
In order to compare the patterns of progress for Male/Female across events, we start by combining data for all events separately for male/female in order to compare the pattern under the same model formula. Note that to compare across different events, we use *Speed* as response variable, rather than *Time* to model the data since *Speed* is more impartial while comparing event with different distances.

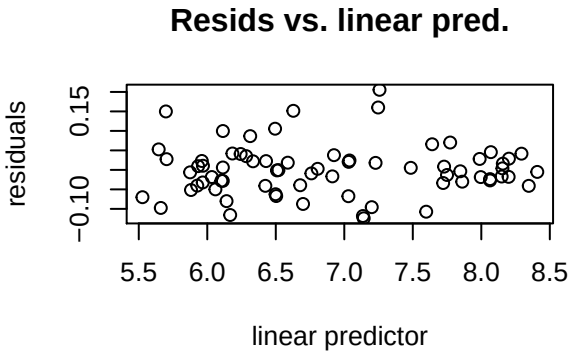
For female data, after fitting the model, there are three potential models that we could consider: Speed ~ s(Date, by = factor(class)) Speed ~ s(Date, by = factor(class)) + s(Altitude) Speed ~ s(Date, by = factor(class)) + s(Altitude) + s(age)

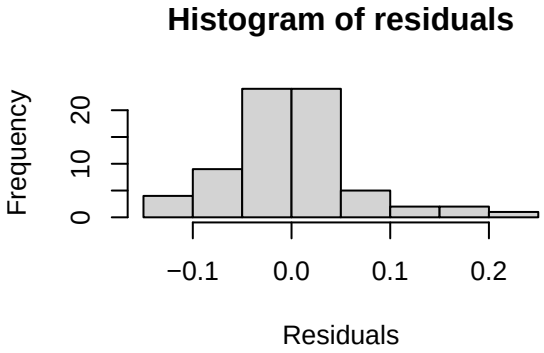
- If we choose the model by lowest AIC, model Speed ~ s(Date, by = factor(class)) + s(Altitude) + s(age) has lowest AIC -145.9555439, and largest R square adjusted.

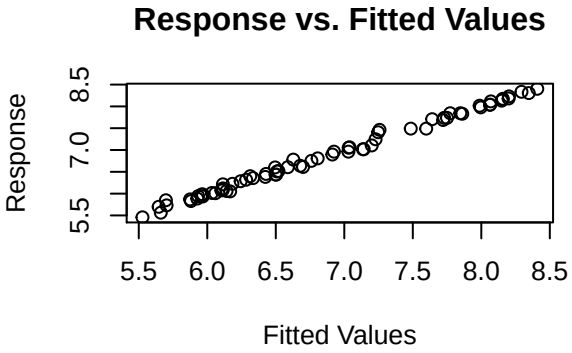


According to the plot, we could explore that relationship between Speed of world records and other explanatory variables. For female, we can see that the increasing in `Speed` for 400m and 800m events is faster than increasing speed of `Speed` for 1500m events, during 1960 to 1980, when most world records occurs. In addition to this, Altitude and Age seems to have a little effect on `Speed` for all three events for this model. It is worth noting that outliers(data point with altitude greater than 2000m) remain in model since they do not change the interpretation, as previously mentioned.





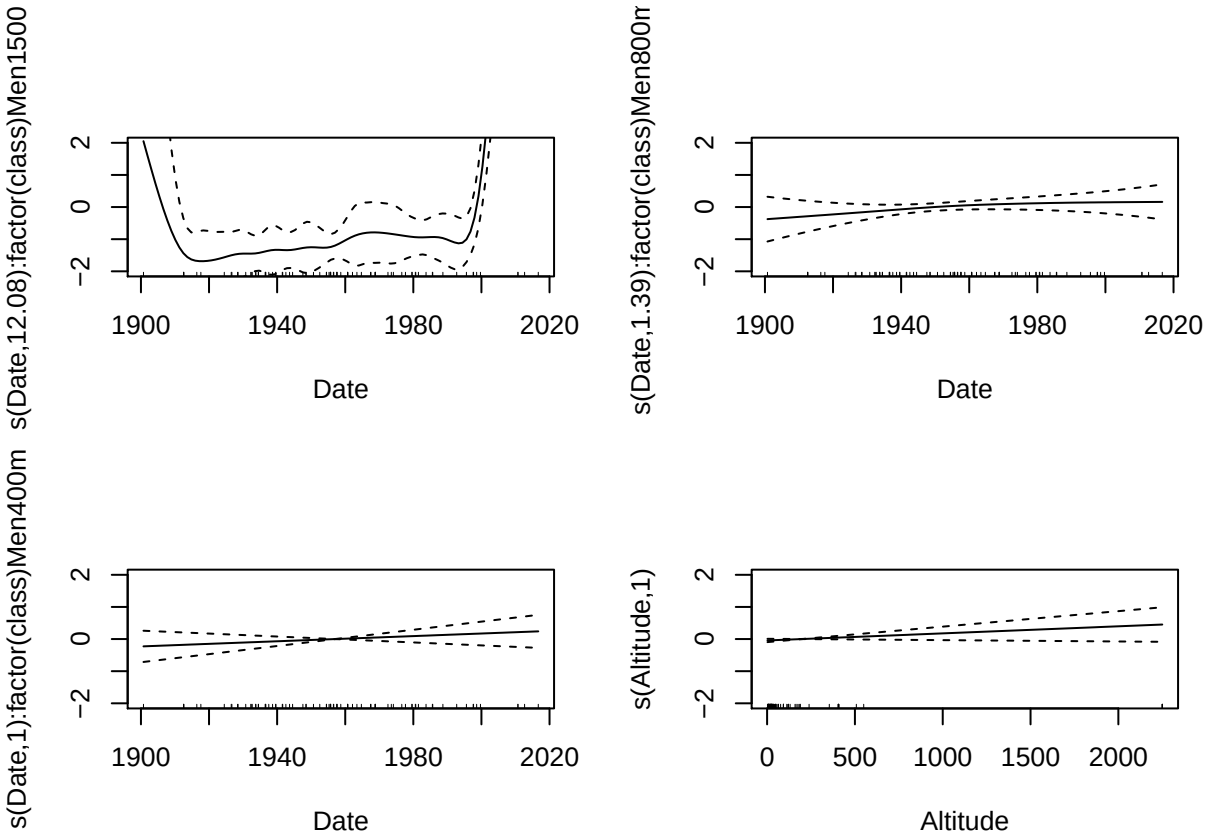





```
Method: REML   Optimizer: outer newton
full convergence after 10 iterations.
Gradient range [-2.043213e-06,1.021746e-05]
(score -39.8981 & scale 0.005774248).
Hessian positive definite, eigenvalue range [0.001988123,32.91575].
Model rank =  46 / 46

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 4.93    0.79    0.03 *
s(Date):factor(class)Women400m  9.00 5.68    0.79    0.02 *
s(Date):factor(class)Women800m  9.00 4.03    0.79    0.02 *
s(Altitude)                     9.00 1.06    0.91    0.22
s(age)                          9.00 3.30    0.91    0.23
---
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.

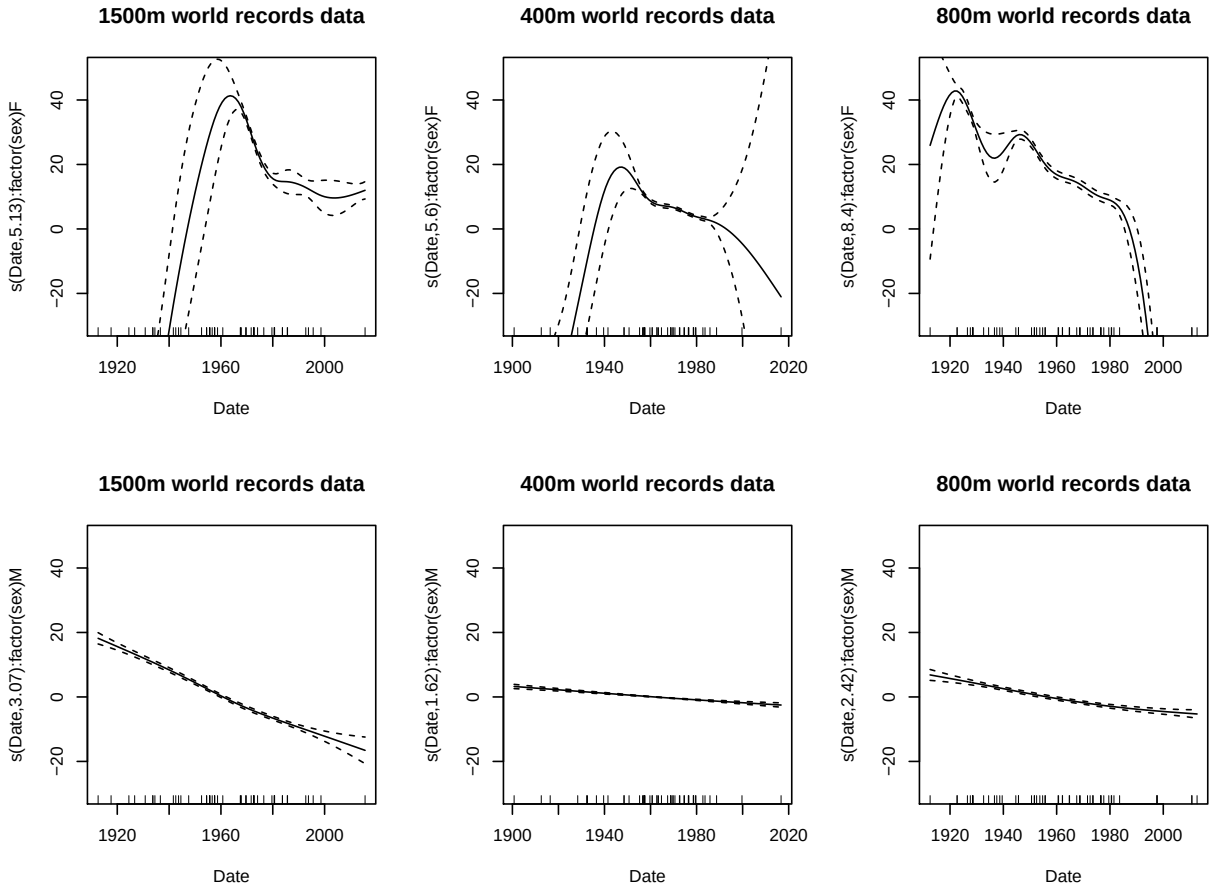


Similarly, we choose model Speed ~ s(Date, by = factor(class), k = 20) + s(Altitude) for its lowest AIC 154.4787124, and largest R square adjusted for male data.

For male athletes, if we choose 1920 to 1980 as our observation period, which corresponds to 20 to 80 in our *Date* axis. We could see that the curve of Men1500m data is for wiggy and steeper than Men400m and Men800m data, which indicates that increasing in `Speed` are faster for Men1500m than Men400m and Men800m, and the variance for Men1500m is also larger than Men400m and Men800m. As a result, we could conclude that Male has faster progression in events that has relatively longer distances(1500m).

3.5 Pattern of progress for Each Event

In order to compare gender performance for each event, we start by combining data for both male and female athletes to compare the pattern of progress for each event using the same model.



From the plot, we could see that for all three events, dropping speed for Time is faster for female than malefor data points from 1950 to 1980, when most world records was created. It could indicate that female hadfaster rate of progress than male in world records over middle distances.

4 Conclusions and discussion

4.1 Summary of conclusions to all questions of interest

- Male has faster progression in *Speed* in events that has relatively longer distances(1500m) than events that has relatively short distances(400m,800m) .
- Female has faster progression in *Speed* in events that has has relatively short distances(400m,800m).
- Female have faster rate of progress than male in world records over all middle distances events.

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.
- Wind speed 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, and for more accurate analysis.
- Other fitted method such as cross-validation method could used for comparison.

5 Reference

Dodge, Yadolah (2006). *The Oxford Dictionary of Statistical Terms*. Oxford [Oxfordshire]: Oxford University Press. ISBN 0-19-920613-9.

Faraway, J., 2009. *Linear models with R*. Boca Raton, Fla: Chapman & Hall/CRC.

Montgomery, D., Peck, E. and Vining, G., 2012. *Introduction to linear regression analysis*.

Patterson, H. D.; Thompson, R. (1971). *"Recovery of inter-block information when block sizes are unequal"*. Biometrika. 58 (3): 545. doi:10.1093/biomet/58.3.545

Wood, S., 2006. *Generalized additive models*. Boca Raton: Chapman & Hall/CRC.