Estimating Improvement Curves for Young Skiers

One-year Master thesis in Statistics, 2008
Department of Economics and Society, Dalarna University

（source: http://toolbox.svenskidrott.se/Downloads/27195/pics/bilder_138_N3_1.jpg）

Note: The photo above shows skiers of the age 14-17 at Svenska Skidspelen in 2006. The style of skiing is skating.

Authors：Rui Chen and Yuli Liang

Supervisor：Kenneth Carling

Submitted on: 22 June 2008

Abstract

This paper aims to estimate the improvement curves for young skiers. We selected 203 individual skiing participants from two races and obtained 967 repeated measures for these skiers in the recent 6 years (2003-2008). To reduce the confounding factors and make sure skiers' performance can be compared at the same level, reference race was used to standardize the velocities. After the standardization, non-linear mixed model was employed to fit the data. We find that the girl's improvement curve is more extended and the boy's curve is steep, which show the different effects of puberty. But the variations among different skiers for both boys and girls are similar.

Keywords: velocity; age; confounding factors; standardization; FPL model; PN model; DF model

Contents

1 Introduction

1.1 Background

In designing rules for competition and principles for systematic training, it is useful to know a young athlete's potential improvement curve. In terms of improvement, we usually use velocity as measurement. And set age as independent variable to measure how the velocity increases with age. From an ideal point, we should control, apart from age, all other factors which affect velocity. Then we can state how velocity depends on a relationship with age, not confounded by other variables. However, in competition it might be difficult to achieve such conditions for repeated measures. In some sports, we can set competitors under almost identical conditions. For example, in running, it is doable as an athlete can compete in an indoor arena on a flat track.

Such analysis indicates that males and females have different improvement curves where the curve for females has its greatest derivative around the age of 12. For males it is around 14 years of age and steeper than for the females. This can be simply explained by the puberty[1] effects (Abbassi, 1998). Conventionally, these curves are assumed to approximate other sports where the estimation of the curve of improvement is more challenging. For instance, out-door bike-racing where factors such as track, wind and pavement will confound the racing speed. That is, repeated measures which were collected under identical conditions are unlikely to be available.

1.2 Problem raised

For cross-country skiing (XC-skiing), the number of confounding factors is high, such as weather condition, race track and ski waxing. For example (see Table 2.1), the skier whose ID is H18-1 has two obviously different velocities (6.37m/s and 7.35m/s) in two races even

[1] Puberty refers to the process of physical changes by which a child's body becomes an adult body capable of reproduction.(http://en.wikipedia.org/wiki/Puberty)

though they took place in the same year of 2008. And by comparing different years (from 2004 to 2008) for the same races (Mora Pinglan, MP for short), big variation, which may be caused by weather conditions and race distance etc., was found in his velocities. Thus, to make an improvement curve of age one needs to eliminate these effects, which otherwise would confound the result of the analysis, make velocity standardized and comparable.

1.3 Aim

The aim of this thesis is to estimate an improvement curve (in terms of velocity) for the age interval 10 to 18 years. Due to an expected difference between boys and girls, we will estimate their curves separately. Reference intervals are needed because individual deviation from typical improvement must be taken into consideration. Thus, the estimate result should be made curves which are regarded as a typical curve complemented with reference intervals which expresses the individual deviation from the typical improvement.

1.4 Approach

We selected all XC-skiers of the age 15-18 in 2008 from Falun race (Lilla SS) and Mora race[2] on Saturday race[3] as a sample for analysis. These two races are chosen for three reasons: the location of them is geographically close to each other, which means that skiers are more likely to participate in both of the two races; they are popular for all young skiers at the age of 7-20, which means repeated measures can be obtained; they are individual races in the sense that we analyze skiers' individual performance. Then set those skiers as range, we search the archives for the results of historical races where this cohort has participated back to 2003 when the cohort was 10-13 years old. Nonlinear four-parameters logistic growth model is used (Karlsson, 2006). In the model, we set four parameters as both fixed and random effects to approximate the confounding factors.

[2] In MP, race took place on December in the year of 2002 and 2005-2007 while Lilla SS always runs in February. Thus, to make variable of year comparable, we consider December's race as the next year's. e.g. "Dec. 2007"="2008"

[3] There are two days races: Saturday and Sunday. On Saturday, they compete individually. On Sunday they race by teams which is not our analysis target.

2 Data

2.1 Data structure

After the treatment of the raw data[4], a new dataset was constructed which consists of 11 variables. ID was used to define skier instead of name and club in the raw data. Male and MP define gender and race as "1"for male and MP while "0" for female and Lilla SS. Age* is the age calculated by days (details can be found in the section 2.4). St_ Factor and St_Vel were computed by the formula in section 2.3. Table 2.1 gives an example of the dataset for the skier assigned ID "H18-1". The first row of the table origins from the raw data of a male skier named Kalle Eriksson from Domnarvets GOIF club whose race result is 26.15 minutes in the 10 km race of Lilla SS 2008 when he was 18 years old.

Table 2.1 An example of repeated measures from ID "H18-1"

ID	Male	Age	Age*	Year	MP	Time	Distance	Velocity	St_Vel	St_Factor
H18-1	1	18	17.73644	2008	0	26.15	10	6.373486	5.801744	1.098547
H18-1	1	18	17.61041	2008	1	22.683	10	7.347647	5.577086	1.317471
H18-1	1	17	16.61041	2007	1	28.75	10	5.797101	5.584109	1.038143
H18-1	1	16	15.61041	2006	1	13.617	5	6.119801	5.495549	1.113592
H18-1	1	15	14.61041	2005	1	11.51	5	7.240081	5.645188	1.282523
H18-1	1	14	13.61041	2004	1	6.217	2.5	6.702054	3.8461	1.742558

As is shown in the above table, each skier may have 12 observations (repeated measures) at most and 1 observation at least. In this case, skier H18-1 participated 6 times during our observed period; one is Lilla SS 2008 and the others at MP from 2004 until 2008.

2.2 Data description

We selected all skiers in the age interval 15-18 years who took part in Lilla SS or MP 2008. Thereafter we have searched for the participants in the two races in 2003-2008. All together

[4] The raw data of races Lilla SS and MP were mostly collected from website: www. faluik .se, www. Ifkmora.se and www.racetimer.se. Except for the results of 2005 and previous years of MP, which were obtained from Sandra Jansson' s office directly. The results of boys with age below 14 in 2004 MP are unavailable.

there are 203 skiers. For the 203 skiers, we have a total of 967 observations, of which 107 are boys and 96 are girls, 495 and 472 observations for each respectively. We tried to trace their results further back in 2000, but it was very difficult to get race data prior to 2003.

Table 2.2 The number of skiers in the data set divided by age and gender

age	boys	girls	total
18	23	15	38
17	25	24	49
16	32	25	57
15	27	32	59
total	107	96	203

Each skier should have 12 repeated measures if he or she participated in all the competitions. Then there would be 12×203=2436 observations in total. Now only 967 observations were used to estimate the curve. Is it credible for only use the observations available? Skiers missed races mostly due to four reasons:

1) Not training enough so they do not want to get bad results;

2) Sickness;

3) Participate in other ski races, especially some races which are particular for their age groups;

4) Participate in other kinds of sport events.

If 1) happened, the curve based on such data might overestimate the true improvement. Otherwise, the number of observations for particular skier can be regardless as the purpose of this paper is to analyze the relationship between velocities and age. However, it is difficult to make sure reason 1) happened because psycho problems exist. (Skiers probably say they will be absent in the race due to illness but actually they are not well training.) Thus, we assume that the missing values will not affect the shape of the improvement curves.

To make things clear and for further study, we provided the following table for the frequency of the number of observations. Most of the skiers have less than 10 repeated measures and on average 5 observations.

Table 2.3 The frequency of the numbers of repeated measures per skier in the data set

repeated measures	frequency	%	cumulate frequency	cumulate %
12	2	0.99	2	0.99
11	5	2.46	7	3.45
10	7	3.45	14	6.90
9	12	5.91	26	12.81
8	12	5.91	38	18.72
7	11	5.42	49	24.14
6	23	11.33	72	35.47
5	18	8.87	90	44.33
4	32	15.76	122	60.10
3	33	16.26	155	76.35
2	34	16.75	189	93.10
1	14	6.90	203	100

Then we provide the whole understanding of raw data through the box plot, see Figure 2.1. For boys, the velocity improves steadily as age increases but there exist big variations for age groups from 13 to 16. For girls, the relationship between age and velocity seems non-linear and slightly fluctuate after age 13. Large variations, which refer to un-standardized of data, also exist at the age interval between 12 and 18.

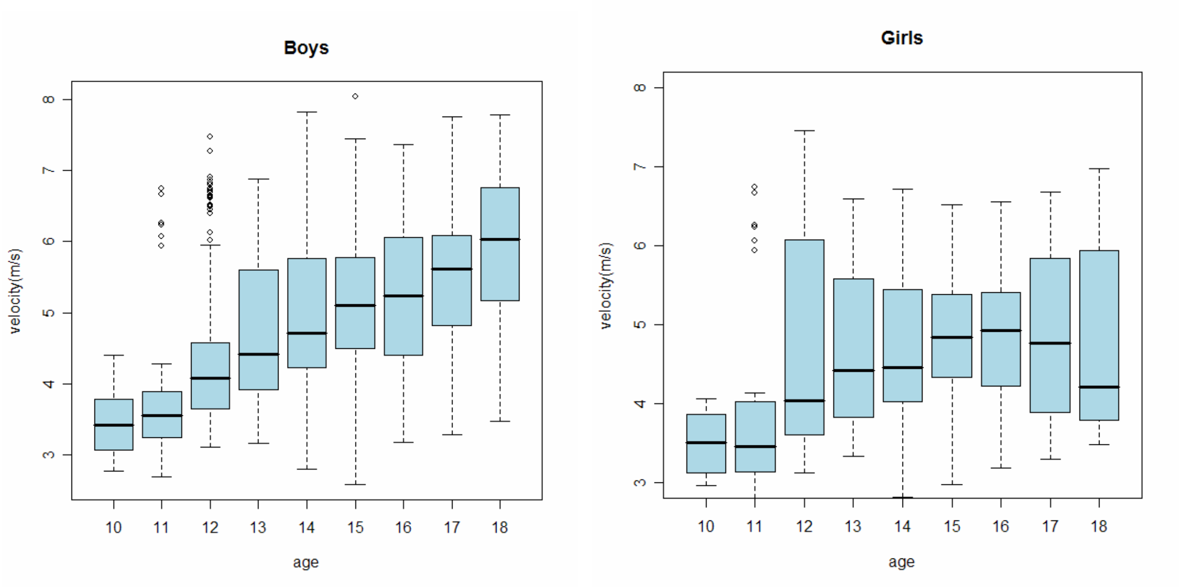


Figure 2.1 Box plots of skiers' velocity shown by age classes and gender

2.3 Standardization

It is unreasonable to compare velocities which are from different ages, different races and different years. Take H18-1 as an example again (see Table 2.1), in 2008 we can not tell exactly which speed is faster before standardizing the velocities, as race and weather conditions are different. And in the years between 2004 and 2008 in MP, whether or not there is an improvement of speed, as the velocities seems fluctuating. Because the race conditions such as track, weather (snowing or windy) and other confounding factors will affect the velocity. Thus a reference race is needed to standardize the velocity, to reduce the confounding factors and make sure skiers' performance can be compared at the same level.

Through comparison of the group data in two different races and 6 different years, we use the race of Lilla SS in 2004 as the reference race because the race distance is quite accurate and the other conditions are stable[5], and moreover the number of participant in this race is large in all age classes.

[5] Conclusions are made according to result of the race empirically.

Table 2.4 Medians of velocity in Lilla SS 2004 (m/s)

Boys	age	10	11	12	13	14	15	16	17	18
	V_{md04}^{age}	2.8694	3.2822	3.5087	3.6809	3.8461	5.2465	5.3304	5.4024	5.4024
Girls	age	10	11	12	13	14	15	16	17	18
	V_{md04}^{age}	2.7855	3.2085	3.2751	3.5294	3.6764	4.4404	4.6772	4.7984	4.7984

Note 1: 17 and 18 belong to the same class and hence it is not possible to compute the median for 17 and 18 years old separately.

Note 2: The median is calculated based on all participates, not only on those selected for the analysis of the thesis.

Standardization is carried out according to the formula below.

$$V_{std}^{i} = V^{i}_{obs} \times \frac{V_{md04}^{age}}{V_{md}^{age}}$$

Where V_{md04}^{age} stands for the median of velocity of particular age group in Lilla SS 2004, and V_{md}^{age} denotes the median of velocity of corresponding year, race, age and gender where original value of observation V_{obs}^{i} occurred.

To check whether or not the standardized method works, we need to make sure standardization not only works for middle values but also for other values, especially the high velocity values. Thus, we pick out boys at the age of 18 in MP 2008 to compare the velocities after standardized with reference race's velocities. If they are similar then the method does work.

Table 2.5 Deciles of boys in Lilla SS 2004 and MP 2008

race	10%	20%	30%	40%	50%	60%	70%	80%	90%
Lilla SS 2004	5.0700	5.1414	5.2124	5.3533	5.4012	5.4615	5.6103	5.7078	5.8127
MP 2008	5.1152	5.1798	5.2601	5.3164	5.4024	5.5339	5.5787	5.6252	5.7578

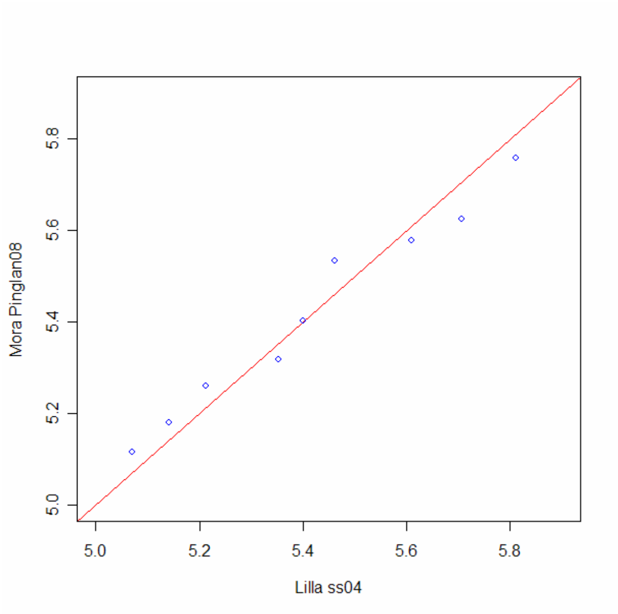


Figure 2.2 Deciles of boys of age 18 in Lilla SS 2004 and MP 2008

From the Figure 2.2, there is no large deviation between the plots and the diagonal which means the method seems to work well for all velocities. We also check for the boys' velocities of age 17, 16 and 15 and so on, and for girls, and similar results were found. Hence, it is sufficient to state that after standardization the data are comparable between different races and different years. This method of standardization works well with our data.

After standardization of data, more clear trends can be seen in the box plots (see Figure 2.3). It is interesting to note that both of boys' and girls' velocities have a jump around the age of 14. According to the physical theory, the youth become much stronger and is powerful during the puberty period, which should be especially true for boys. Also, skiers experienced a long training during this period. Thus, it is often to see a great improvement in that age group.

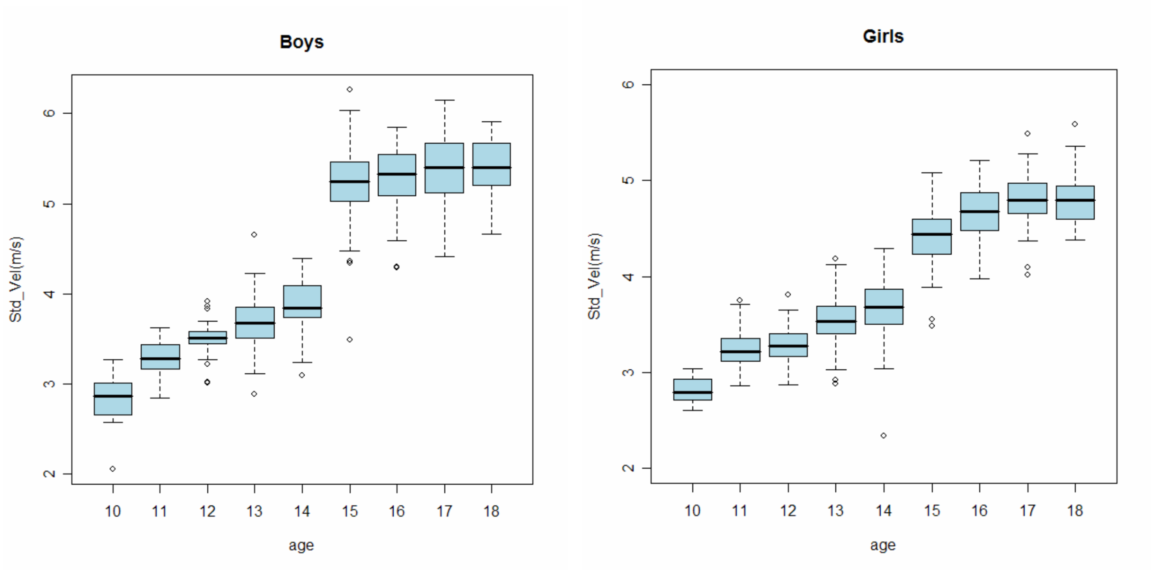


Figure 2.3 Box plots of skiers' standardized velocity shown by age classes and gender

2.4 New definition of age

One more thing should be considered, if we define the age variable more precisely, say, by day not by year, we may find more exact conclusions. As we know, skier who is born in January is likely to ski faster than a skier who is born in December. For this purpose, we redefine the independent variable as the actual age (use the date[6] when race took place minus the birthday of skier). The formula of calculating a more precise age is

$$\begin{aligned} age* = {} & [365 \times (raceyear - birthyear) + 30.4 \times (racemonth - birthmonth) \\ & + (raceday - birthday)] / 365 \end{aligned}$$

Where $age*$ stands for the actual age, and the coefficient 30.4 is chosen to regard all the possible numbers of days (28, 29, 30 and 31) in each month. And we assume there are 365 days in each year. An example of the new age can be found in the Table 2.1.

3 Method

Figure 2.2 suggests that the relationship between velocity and age is not linear, and a second

[6] The race month and race day of Lilla SS and MP are considered as February 15th and December 30th respectively as they usually take place on the middle of February and the end of December each year.

order polynomial is also unlikely to fit the data well. The shape indicates that the four parameters logistic model could be chosen.

3.1 Model definition

The four parameters logistic model (FPL) is often used in the biostatistics field, analyzing the dose-response relationship, which describes the change in effect on an organism caused by differing levels of exposure (or doses) to a stressor (usually a chemical). The non-linear logistic growth function used in four parameters is shown below:

$$f(x) = \beta_1 + \frac{\beta_2 - \beta_1}{1 + \exp((\beta_3 - x)/\beta_4)}$$

Where β_1 and β_2 are two asymptotes for the response variable, with β_1 giving its value at $-\infty$ and β_2 its value at ∞; β_3 gives the EC50 value[7], β_4 is a slope parameter that governs the steepness of the growth curve (Karlsson, 2006).

To illustrate how these parameters influence the shape of the curve, we change the values of β_3 and β_4 simultaneously and draw the curves. As β_1 and β_2 are called location parameters, which simply shifts the distribution of $f(x)$ so that the shape of the graph is unchanged (Casella and Berger, 2002). The following table and figure give some combinations of values for parameters β_1, β_2, β_3 and β_4 with corresponding curves.

Table 3.1 Combination of different values of parameters for FPL

Curve	β_1	β_2	β_3	β_4
1	3	6	14	4
2	3	6	15	1
3	3	6	17	0.5
4	4	4	14	0.5
5	3	6	17	0.2
6	3	6	15	0

[7] The value of x for which the response variable y has reached the midpoint $\beta_1 + |\beta_2 - \beta_1|/2$.

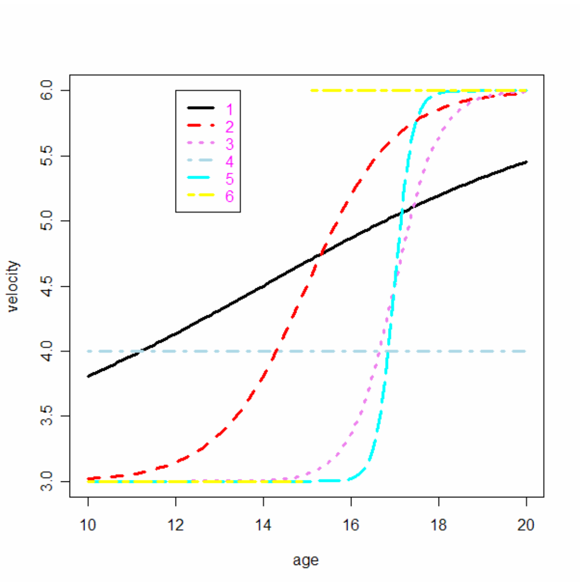


Figure 3.1 Six examples of FPL curves with different parameters values according to Table 3.1

In this thesis, skiers' performances are often affected significantly at different age groups. So use the velocity as the response variable, the age as the independent variable, the FPL in this case should be

$$f(age_t) = A + \frac{B-A}{1+\exp((xmid - age_t)/scale)} + \varepsilon_t \quad \text{(t=10, 11,...,18)}$$

A: minimum velocity of skiers in the whole life implied by the model

B: maximum velocity of skiers in the whole life implied by the model

xmid: value of age for which velocity has reached the midpoint A+(B-A)/2

scale: slope value of the improvement curve (can be interpreted as degree of improvement affect by puberty)

3.2 Model construction

The four parameters non-linear mixed logistic model belongs to non-linear mixed model. Thus, the "nlme" function in R[8] was used to estimate the model, with the "SSfpl" function for four parameters logistic model. Since we don't know if skiers have the same asymptotes, same xmid and same scale or not before the analysis. Thus in "nlme" function, parameters: A, B, xmid and scale are set as both fixed effects and random effects. Then we check for Akaike

[8] R: is a language and environment for statistical computing and graphics. (www.r-project.org)

information criterion (AIC) values. Smaller AIC values explained why all of parameters needed to be both fixed and random effects, instead parts of parameters be fixed effects and parts of them be random effects. Details of R codes are attached in the appendix I.

4 Result

4.1 Estimation

One of the best things one can do to ensure a successful nonlinear analysis is to obtain good starting values for the parameters-values from which convergence is quickly obtained (Douglas and Donald, 1988). For boys, starting values[9] of parameters were set as A=3.5, B=5.5, xmid=14.5, scal=0.3. For girls, to avoid maximum number of iterations reached without convergence, data of skiers whose number of observations is equal to and larger than 6 were used. (We also checked for boys with 6 and more observations, the results are similar as the results from all observations.) And start values were chose as A=3.2, B=4.8, xmid=14.5, scal=0.7. Estimates of parameters are showed in the Table 4.1 and 4.2.

Table 4.1 Fixed effects and random effects estimates of FPL model for Boys

	Fixed Effects		
	Value	Std.Error	t-value
A	3.5177	0.0316	110.1801
B	5.3273	0.0342	155.6467
xmid	14.0445	0.0411	341.8721
scale	0.2189	0.0174	12.5442
	Random Effects		
	Std.Dev		
A	0.1986		
B	0.2771		
xmid	0.2944		
scale	0.0001		
ε	0.2261		
Number of Observations: 495			
Number of Groups: 107			

[9] Start values were chose empirically in view of problem of convergence.

Table 4.2 Fixed effects and random effects estimates of FPL model for Girls where repeated measures are equal to and larger than 6

Fixed Effects			
	Value	Std.Error	t-value
A	3.1790	0.0441	72.0570
B	4.8714	0.0634	76.7791
xmid	14.0985	0.0866	162.7902
scale	0.7230	0.0565	12.7880
Random Effects			
	Std.Dev		
A	0.1863		
B	0.2480		
xmid	0.3434		
scale	<0.0001		
ε	0.1804		
Number of Observations: 302			
Number of Groups: 38			

Note: Retain 4 decimals.

For fixed effects, the estimates of the two asymptotes parameters for both boys and girls are reasonable as they reflect the velocities that most skiers will reach. And the xmid estimates for boys and girls are 14.0445 and 14.0985, which means both boys and girl will arrive at the midpoint of their velocities when they are 14 years old according to FPL model, which are 3.5177+(5.3273-3.5177)/2=4.4225m/s and3.1790+(4.8714-3.1790)/2=4.0252m/s respectively. However, the scale values for boys and girls are significantly different, 0.2189 and 0.7230 respectively. This is in accordance with the fact that boys did a greater improvement than girls who had a more slow and steady progress in their ski speed during the puberty period.

For random effects, the standard deviation of the scale for both boys and girls are extremely small (equal to or less than 0.0001) which indicates that the assumption that we believed scale has random effects for both boys and girls is not true. The standard deviation for xmid is large in both boys' and girls' models (0.2944 and 0.3434). That means the age when skiers reach the midpoint of their velocities differ substantially from individual to individual.

According to the results of the parameter estimation, the improvement curves can be drawn finally.

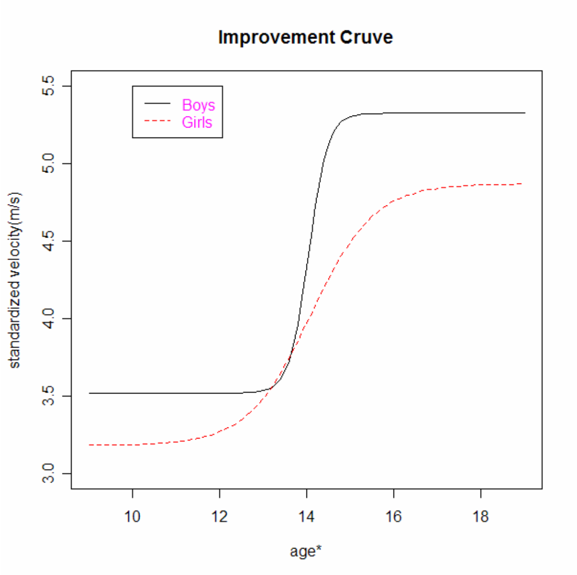


Figure 4.1 Improvement curves of velocity against age

For boys, the velocities increase slowly at the beginning then follow with a big jump at the age of 14 and after that stay still. It is a little different from the girls' curve which begins with increase and steady improvement. As we know, girls often develop their bodies earlier than boys (at the age of 12) but the improvement of performance will experienced a long period until the age of 18. For boys, their bodies and strength are better than girl even though they develop their bodies later than girls (at the age of 14) and after that they made a big improvement in a short period 14-15. Thus, the FPL model fits the data well.

4.2 Residual analysis

Residual analyses are required in the diagnosis of model and evaluate fitness of model. As box plots and QQ plots show residuals of the FPL models are approximately normal distributed with zero mean, except for younger male skiers, which indicates that the model fitted the data well. The trend exists in the younger male skiers indicates that it might be a problem with the logistic curve in estimating the velocity at the youngest boys. For this reason, we will compare this with two other models, see section 5.

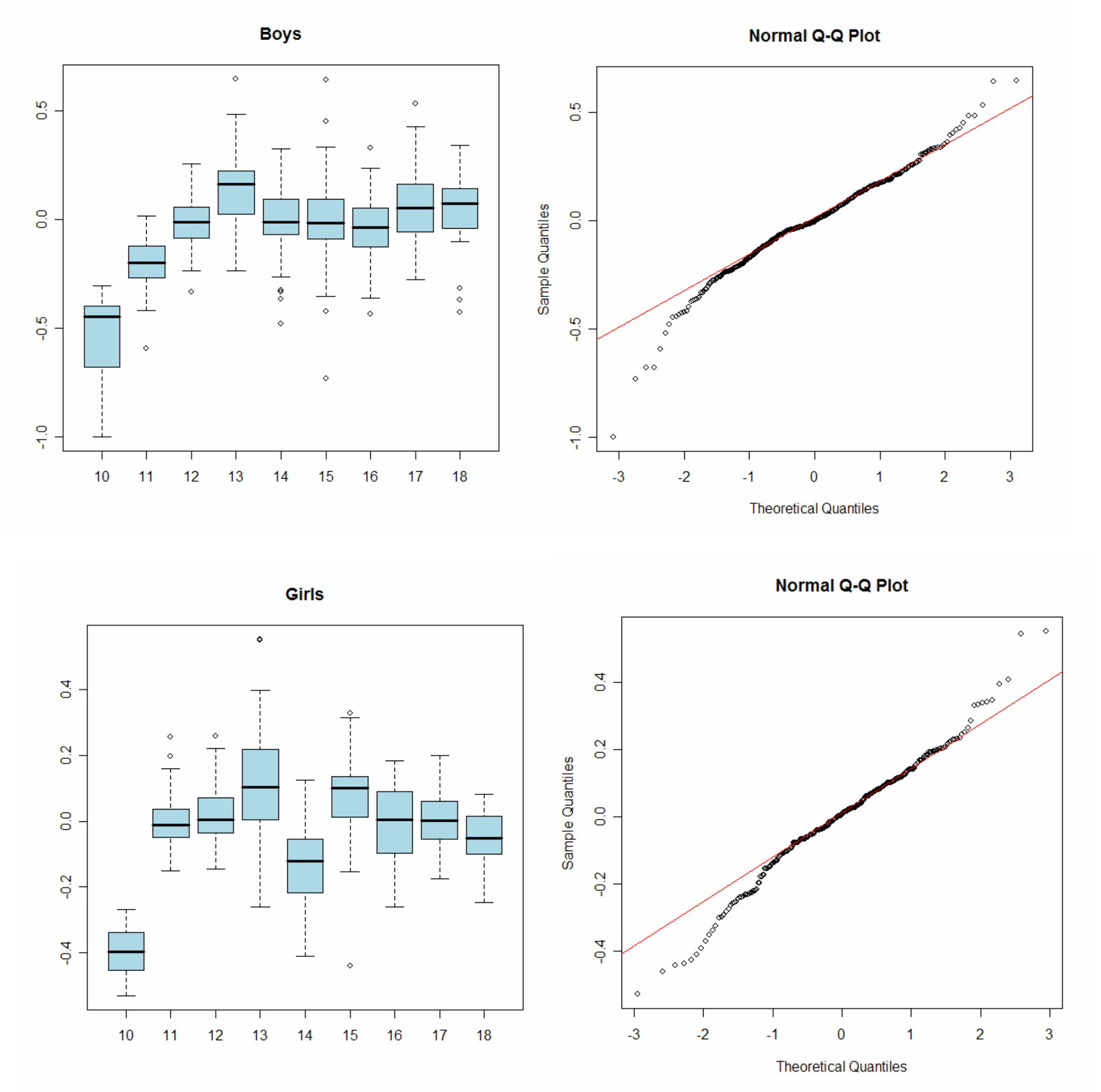


Figure 4.2 Residual analyses for FPL model divided by gender (box plots are drawn by age classes)

4.3 Some examples

In order to check whether the model fits the real data or not, some skiers with the most repeated measures were picked out to illustrate the model’s fit. There are two cases for boys and girls each. Their IDs are H15-8, H18-2, D15-7 and D18-9.

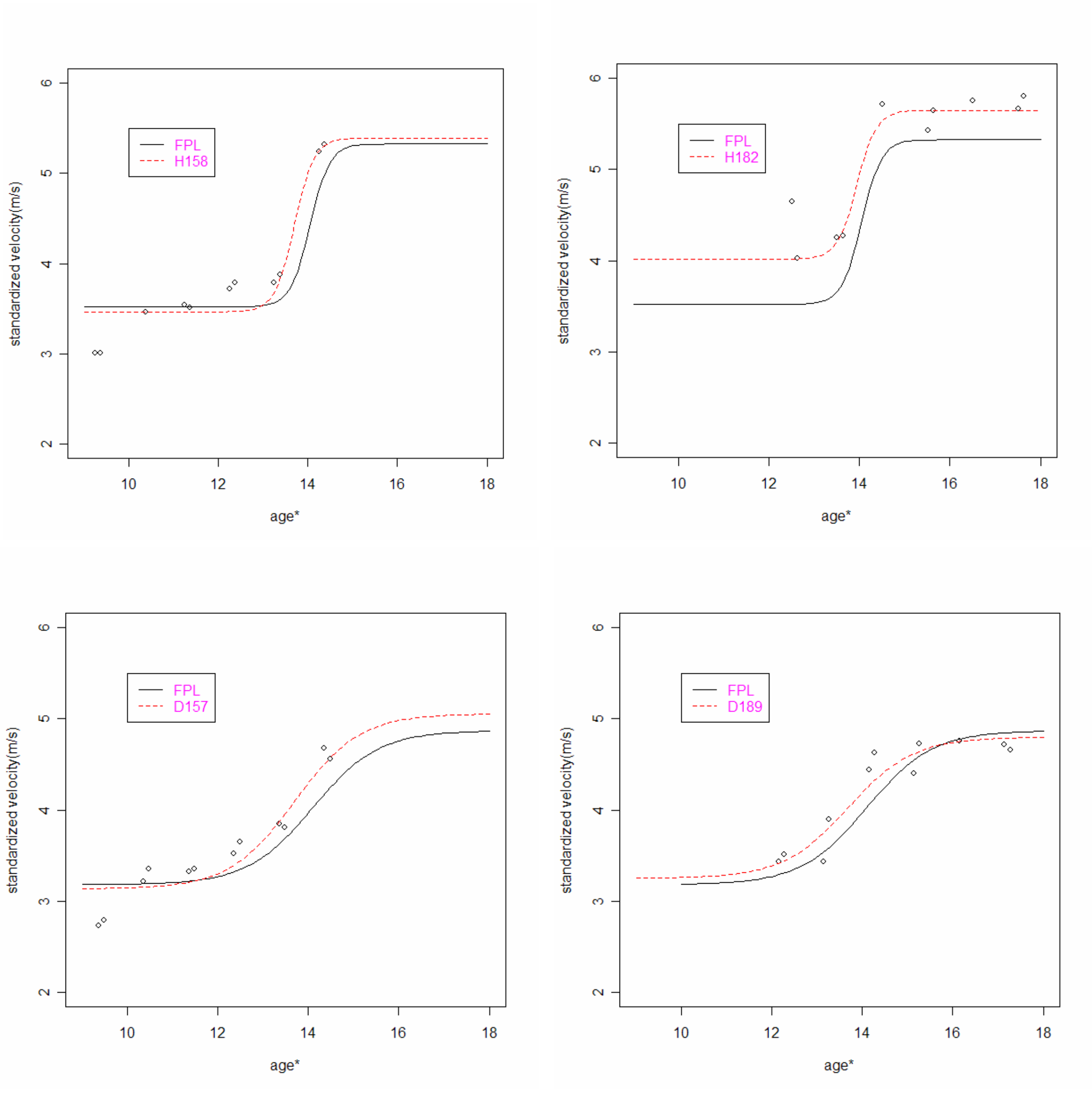


Figure 4.3 Four examples for fit of the FPL model

As Figure 4.3 shows, individual curves fit data more accurately than the FPL curve without random effects, which means random effects of their parameters do exist. Similar shapes and trends can be seen between the FPL curve and individual curves which explains the fixed effects. Thus, the FPL model fits our data well.

5 Comparison with other models

5.1 Between FPL and polynomial model

Nonlinear mixed-effects models extend linear mixed-effects models by allowing the regression function to depend nonlinearly on fixed and random effects. Because of its greater flexibility, an NLME model is generally more interpretable and parsimonious than a competitor empirical LME model based, say, on a polynomial function (Jose and Douglas, 2000). So we compare FPL with fifth-order polynomial model first. The fifth-order polynomial model (PN)

$$f(x) = \alpha + \sum_{i=1}^{5} \beta_i x^i + \varepsilon \quad (x \in (9,19))$$

PN uses the age classes as the independent variable and the standardized velocity as the response variable in the same data as previous FPL analysis.

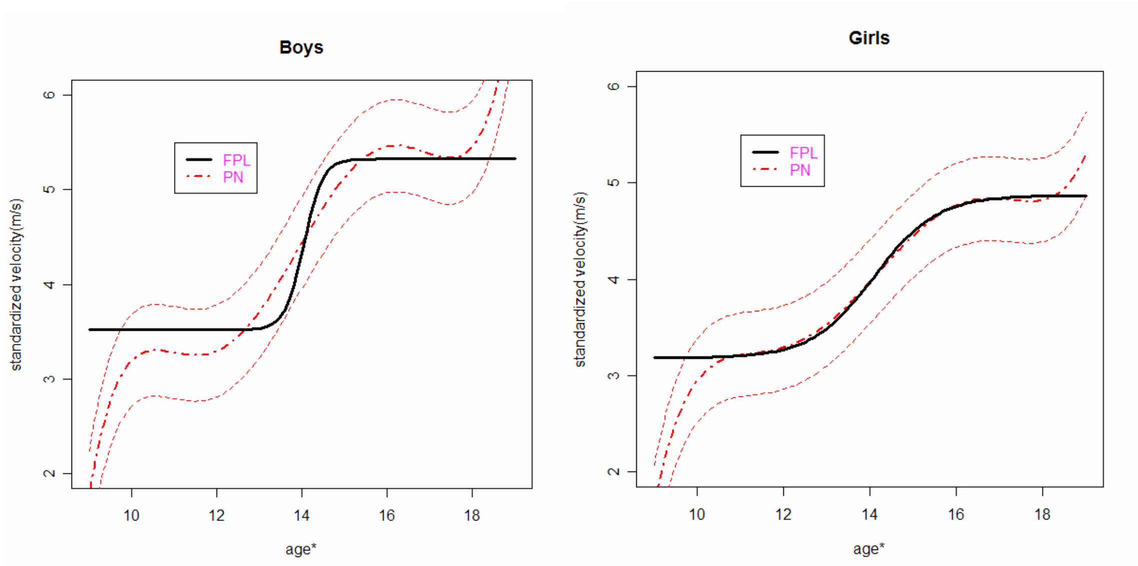


Figure 5.1 Comparison between FPL model and polynomial model divided by gender

In this polynomial model, we put ID as the random effect and consider it as normal distributed with null mean and constant variance. At 0.05 confidential level, the confidence interval of estimates with random effect is obtained (see the dashed lines in Figure 5.1). It means most of individual curves are in the area between two dashed lines.

From the comparison, we can see that for boys, there is a decreasing trend after age 11 in the polynomial curve which is not reasonable. Because physical theory and training outcome tell that there should be steady trends both before and after the puberty period. Thus the polynomial model is not as feasible as the FPL model, while for the girls the two curves are similar in the age interval (10,18) (see Appendix Ⅱ for more comparisons). In the FPL model, the AIC value of boys is 284.9648 and -12.6958 for girls; while in the PN model it is 474.4825 for boys, much bigger than FPL, and -3.5304 for girls, which also shows the FPL model is better than the PN model. There is another reason to convince us to choose the FPL model: FPL is more representative in reflecting the skiers' performances than the PN model which is more complex and is difficult to interpret the meaning of its parameters.

5.2 Between FPL and distribution-free model

Considered the age* is not integer values, it is necessary for distribution-free model to divide them into some levels (intervals). Each interval contains one year but due to many observations from age 13 to 15, we handle them by half a year. In total, there are 12 age levels between age 9 and 19. The form of distribution-free model (DF):

$$\text{Velocity} = \sum_{t=1}^{12} D_t \gamma_t + \varepsilon_t \qquad \text{and} \qquad D_t = \begin{cases} 1 \text{ if age =t} \\ 0 \text{ otherwise} \end{cases}$$

As the formula shows, in distribution-free model we defined "age" as a factor, which means each age group has different parameter. And "t" stands for the twelve age levels (denotes by 1, 2, 3… 12). Variable ID is set as random effects in this model. Improvement curves from these two models are plotted as Figure 5.2. Two skiers, whose ID are H18-2 and D15-7, are picked out again to draw their individual curves (see Figure 4.3) according to the DF model.

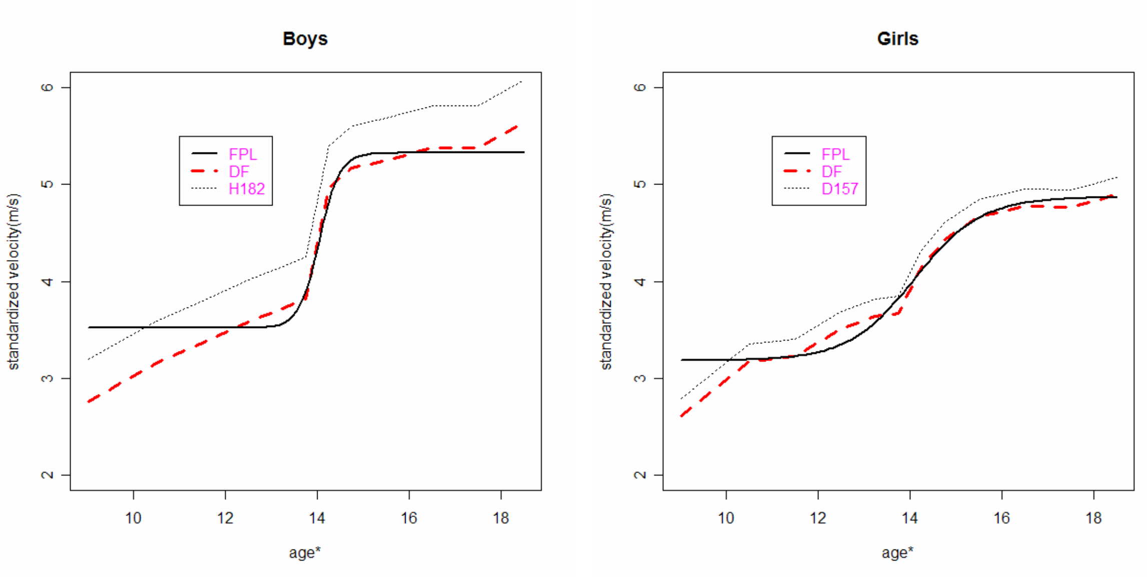


Figure 5.2 Comparison between FPL model and distribution-free model divided by gender

It seems the DF model fits better than the PFL, especially in youngest skiers where there is a reasonable and steady increase. However, we do not use the DF model based on following reasons: distribution free model belongs to non-parametric analysis which results are difficult to explain; we are mostly interested in the age interval around the puberty not in the beginning years of skiers skied; if we use the DF model to predict the velocities of skiers before 10 years old we will get negative values which is unbelievable. Thus, the FPL model is to be preferred over the DF model.

6 Discussion

The way in which we select the data may cause more observations in the age interval (13, 15) than any other age classes in our thesis. The result is more exact based on more observations than others. So for our estimation, there may be some bias at other ages especially at the age 10 and 18. It can be found from the comparisons between the FPL model and other alternative models (Figure 5.1 and 5.2). However, in our case, it is very difficult to get the data before 2003. Another important reason is that we are mostly interested in the puberty period (around the age 14) and not concern about the extreme ages.

The population of observed skiers was defined as those who are 15-18 years old in 2008. Then searching back, repeated measures were found during previous years until 2003. However, if we started to collect observations from 2003 based on those who were 10-13 years old, the different population will be obtained. And if our analysis based on this population the puberty effect might be not apparent because of the lower velocities of those skiers who quit the race after felt frustrated from bad results during the puberty period.

Another interesting thing is that young skiers who were born in the same year but not in the same academic year often have different performance. Generally, in a coach’s mind, the older skiers are thought to own better body conditions than the younger ones and the former will be paid more attention when they are doing training. With the good feedback, the former will be more confident in the performance and this preference will be continuous. At this time, the skier from the former academic year probably skis faster than the skier from the latter one. Time after time, the younger one may feel frustrated from bad training results and won’t attract much attention from the trainer. This may lead to a lower velocity in the races. It can be a potential topic for the further study if we divide the skiers by academic year.

Reference

Abbassi, V., (1998) Growth and Normal Puberty, *American Academy of Pediatrics*， volume 102, pp507-511

Casella, G., Berger, R. L., (2002) *Statistical Inference 2nd Edition,* Thomson Learning, pp116-118

Douglas, M. B., Donald, G. W., (1988) *Nonlinear Regression Analysis and Its Applications,* Wiley, pp72

Jose, C. P., Douglas, M. B., (2000) *Mixed Effects Models in S and S-PLUS,* Springer, pp274-276

Karlsson, A., (2006) *Estimation and Inference for Quartile Regression of Longitudinal Data with Application in Biostatistics*，Uppsala University

Appendix I：R Codes

Codes A: FPL model

```
## read the data ##
library(RODBC)
ch1<-odbcConnectExcel(“E:/d-essay/model/Velocity_st.xls”)
data<-sqlFetch(ch1, "Blad1")
odbcClose(ch1)
##estimating the FPL##
library(nlme)
male<-subset(data,data$Male==1)                                    #for boys
maleGr<-groupedData(St_Vel~Age_d|ID,data=male)                     #use group data
gc1<-nlme(St_Vel~SSfpl(Age_d,A,B,xmid,scale),fixed=A+B+xmid+scale~1,
random=pdDiag(A+B+xmid+scale~1),start=c(A=3.5,B=5.5,xmid=14.5,scale=.3),
data=maleGr)
summary(gc1)                                                       # FPL model for boys
boxplot(resid(gc1)~maleGr$Age_d,col="lightblue",main="Boys")       #box plot of residual
qqnorm(resid(gc1))                                      #chenk the residual for boys is normal distributed
qqline(resid(gc1),col=2)
ranef(gc1)                                              #for the purpose of get individual curves

male<-subset(data,data$Male==0)                                    #for girls
t<-table(male$ID)[(table(male$ID))>=6]
male16<-subset(male,male$ID%in%names(t))
male6Gr<-groupedData(St_Vel~Age_d|ID,data=male16)                  #make the model convergent, use
                                                                   the skiers whose observations are
                                                                   equal to and larger than 6
gc2<-nlme(St_Vel~SSfpl(Age_d,A,B,xmid,scale),fixed=A+B+xmid+scale~1,
random=pdDiag(A+B+xmid+scale~1),start=c(A=3.2,B=4.8,xmid=14.5,scale=.7),
data=male6Gr)
summary(gc2)                                                       # FPL model for girls
ranef(gc2)
boxplot(resid(gc2)~maleGr$Age_d,col="lightblue",main="Girls")      #box plot of residual
qqnorm(resid(gc2))                                      #chenk the residual for girls is normal distributed
qqline(resid(gc2),col=2)
##draw the improvement curve##
x<-male$Age_d
V<-function(x) 3.5177+(5.3273-3.5177)/(1+exp((14.0445-x)/0.2189))       #boys’ function of FPL
curve(V(x),xlim=c(9,19),ylim=c(3,5.5),main="Improvement Cruve",xlab="age*",
ylab="standardized velocity(m/s)")                                       #boys’ improvement curve
x<-male6$Age_d
```

```
V<-function(x) 3.1790+(4.8714-3.1790)/(1+exp((14.0985-x)/0.7230))          #girls’ function of FPL
curve(V(x),add=T,col=2,lty=2)                                                #girls’ improvement curve
legend(10,5.5,c("Boys","Girls"),col=c(1,2),text.col=6,lty=c(1,2))
```

Codes B: Pick out some examples

```
##the skier whose ID=H15-8##
H158<-subset(data,data$ID=="H15-8")
plot(H158$St_Vel~H158$Age_d,xlim=c(9,18),ylim=c(2,6),xlab="age*",
ylab="standardized velocity(m/s)")                                        #plot the velocities
V<-function(x) 3.5177+(5.3273-3.5177)/(1+exp((14.0445-x)/0.2189))       #the estimated equation
curve(V(x),add=T)
V<-function(x) (3.5177-5.443970e-02)+((5.3273+0.067101705)-(3.5177-5.443970e-02))/
(1+exp(((14.0445-3.410841e-01)-x)/(0.2189+7.507110e-08)))               #the individual equation
curve(V(x),add=T,col=2,lty=2)
legend(10,5.5,c("FPL","H158"),col=c(1,2),text.col=6,lty=c(1,2))
##the skier whose ID=H18-2##
H182<-subset(data,data$ID=="H18-2")
plot(H182$St_Vel~H182$Age_d,xlim=c(9,18),ylim=c(2,6),xlab="age*",
ylab="standardized velocity(m/s)")
V<-function(x) 3.5177+(5.3273-3.5177)/(1+exp((14.0445-x)/0.2189))       #the estimated equation
curve(V(x),add=T)
V<-function(x) (3.5177+4.973768e-01)+((5.3273+0.324473375)-(3.5177+4.973768e-01))/
(1+exp(((14.0445-1.006167e-01)-x)/(0.2189-4.227331e-08)))               #the individual equation
curve(V(x),add=T,col=2,lty=2)
legend(10,5.5,c("FPL","H182"),col=c(1,2),text.col=6,lty=c(1,2))
##the skier whose ID=D15-7##
D157<-subset(data,data$ID=="D15-7")
plot(D157$St_Vel~D157$Age_d,xlim=c(9,18),ylim=c(2,6),xlab="age*",
ylab="standardized velocity(m/s)")
V<-function(x) 3.1790+(4.8714-3.1790)/(1+exp((14.0985-x)/0.7230))      #the estimated equation
curve(V(x), add=T)
V<-function(x) (3.1790-0.04611786)+((4.8714+0.18557940)-(3.1790-0.04611786))/
(1+exp(((14.0985-0.402253135)-x)/(0.7230+6.304037e-09)))               #the individual equation
curve(V(x),add=T,col=2,lty=2)
legend(10,5.5,c("FPL","D157"),col=c(1,2),text.col=6,lty=c(1,2))
##the skier whose ID=D18-9##
D189<-subset(data,data$ID=="D18-9")
plot(D189$St_Vel~D189$Age_d,xlim=c(9,18),ylim=c(2,6),xlab="age*",
ylab="standardized velocity(m/s)")
V<-function(x) 3.1790+(4.8714-3.1790)/(1+exp((14.0985-x)/0.7230))       #the estimated equation
curve(V(x), add=T)
V<-function(x) (3.1790+0.06976502)+((4.8714-0.07096195)-(3.1790+0.06976502))/
(1+exp(((14.0985-0.410798308)-x)/(0.7230-5.230926e-09)))               #the individual equation
curve(V(x),add=T,col=2,lty=2)
```

```
legend(10,5.5,c("FPL","D189"),col=c(1,2),text.col=6,lty=c(1,2))
```

Codes C: Comparison between FPL and polynomial model

```
male<-subset(data,data$Male==1)                                    #boys
library(nlme)
gc3<-lme(St_Vel~Age_d+I(Age_d^2)+I(Age_d^3)+I(Age_d^4)+I(Age_d^5),random=~1|ID,data=male)
summary(gc3)                                                  #polynomial model for boys
cf<-fixef(gc3)                                                #extract the fixed effects estimates
rf<-ranef(gc3)                                                #extract the random effects estimates
x<-male$Age_d
V<-function(x) (cf[1]+cf[2]*x+cf[3]*x^2+cf[4]*x^3+cf[5]*x^4+cf[6]*x^5)
curve(V(x),xlim=c(9,19),ylim=c(2,6),xlab="age*",ylab="standardized velocity(m/s)",main="Boys",
lty=4,lwd=2,col=2)
sigma2 <- var(rf[,1])                             #calculate the variance of random effects estimates
vU <- qnorm(.975)*sqrt(sigma2)                    #the upper bound of confidence interval
vL <- -vU                                         #the lower bound of confidence interval
Vi<-function(x,i) (i+cf[1]+cf[2]*x+cf[3]*x^2+cf[4]*x^3+cf[5]*x^4+cf[6]*x^5)    #PN plus random effect
curve(Vi(x,vL),xlim=c(9,19),add=T,lty=2)
curve(Vi(x,vU),xlim=c(9,19),add=T,lty=2)
curve(3.5177+(5.3273-3.5177)/(1+exp((14.0445-x)/0.2189)),add=T,lwd=3)
legend(11,5.5,c("FPL","PN"),col=c(1,2),text.col=6,lty=c(1,4),lwd=c(3,2))
male<-subset(data,data$Male==0)                                       #girls
t<-table(male$ID)[(table(male$ID))>=6]
male6<-subset(male,male$ID%in%names(t))library(nlme)
gc4<-lme(St_Vel~Age_d+I(Age_d^2)+I(Age_d^3)+I(Age_d^4)+I(Age_d^5),random=~1|ID,data=male6)
summary(gc4)                                                  #polynomial model for girls
cf<-fixef(gc4)
rf<-ranef(gc4)
V<-function(x) (cf[1]+cf[2]*x+cf[3]*x^2+cf[4]*x^3+cf[5]*x^4+cf[6]*x^5)
curve(V(x),xlim=c(9,19),ylim=c(2,6),xlab="age*",ylab="standardized velocity(m/s)",main="Girls",
lty=4,lwd=2,col=2)
sigma2 <- var(rf[,1])                             #calculate the variance of random effects estimates
vU <- qnorm(.975)*sqrt(sigma2)                    #the upper bound of confidence interval
vL <- -vU                                         #the lower bound of confidence interval
Vi<-function(x,i) (i+cf[1]+cf[2]*x+cf[3]*x^2+cf[4]*x^3+cf[5]*x^4+cf[6]*x^5)    #PN plus random effect
curve(Vi(x,vL),xlim=c(9,19),add=T,lty=2)
curve(Vi(x,vU),xlim=c(9,19),add=T,lty=2)
curve(3.1790+(4.8714-3.1790)/(1+exp((14.0985-x)/0.7230)),add=T,lwd=3)
legend(11,5.5,c("FPL","PN"),col=c(1,2),text.col=6,lty=c(1,4),lwd=c(3,2))
```

Codes D: Comparison between FPL and distribution-free model

## read the data ##

```
library(RODBC)
ch2<-odbcConnectExcel("E:/d-essay/model/sample.xls")
```

```
data<-sqlFetch(ch2,"Sheet1")
odbcClose(ch2)

male<-subset(data,data$Male==1)
gc5<-lm(St_Vel~factor(Fage),random=~1|ID,data=male)              #estimating the distribution-free model
summary(gc5)
ranef(gc5)
plot(data$V11~data$Fage,col=2,type="l",ylim=c(2,6),xlab="age*",ylab="standardized velocity(m/s)",
main="Boys",lty=2,lwd=3)                                                   # DF for boys
lines(data$VH182~data$Fage,lty=3)
V<-function(x) 3.5177+(5.3273-3.5177)/(1+exp((14.0445-x)/0.2189))           #FPL for boys
curve(V(x),add=T,col=1)
legend(11,5.5,c("FPL","DF","H182"),col=c(1,2,1),text.col=6,lty=c(1,2,3),lwd=c(2,3,1))
gc6<-lme(St_Vel~factor(Fage),random=~1|ID,data=male6)
summary(gc6)
ranef(gc6)
plot(data$V00~data$Fage,col=2,type="l",ylim=c(2,6),xlab="age*",ylab="standardized velocity(m/s)",
main="Girls",lty=2,lwd=3)                                                  # DF for girls
lines(data$VD157~data$Fage,lty=3)
curve(3.1790+(4.8714-3.1790)/(1+exp((14.0985-x)/0.7230)),add=T,lwd=2)    #FPL for boys
legend(11,5.5,c("FPL","DF","D157"),col=c(1,2,1),text.col=6,lty=c(1,2,3),lwd=c(2,3,1))
```

Codes E: Some figures

##Figure 2.2##

```
male<-subset(data,data$Male==1)
boxplot(male$St_Vel~male$Age,col="lightblue",main="Boys",xlab="age",
ylab="Std_Vel(m/s)",ylim=c(2,6))                                              #box plot of boys
male<-subset(data,data$Male==0)
boxplot(male$St_Vel~male$Age,col="lightblue",main="Girls",xlab="age",        #box plot of girls
ylab="Std_Vel(m/s)",ylim=c(2,6))
```

##FPL of different combinations##

```
x<-data$Age                                                   #take the ages of all participants for example
V<-function(x) 3+(6-3)/(1+exp((14-x)/4))
curve(V(x),xlim=c(10,20),ylim=c(3,6),xlab="age",ylab="velocity",lwd=3)
V<-function(x) 3+(6-3)/(1+exp((15-x)/1))
curve(V(x),add=T,col=2,lty=2,lwd=3)
V<-function(x) 3+(6-3)/(1+exp((17-x)/0.5))
curve(V(x),add=T,col="violet",lty=3,lwd=3)
V<-function(x) 4+(4-4)/(1+exp((14-x)/0.5))
curve(V(x),add=T,col="lightblue",lty=4,lwd=3)
V<-function(x) 3+(6-3)/(1+exp((17-x)/0.2))
curve(V(x),add=T,col=5,lty=5,lwd=3)
V<-function(x) 3+(6-3)/(1+exp((15-x)/0))
```

```
curve(V(x),add=T,col=7,lty=6,lwd=3)
legend(12,6,c("1","2","3","4","5","6"),col=c(1,2,"violet","lightblue",
"5","7"),text.col=6,lty=c(1,2,3,4,5,6),lwd=c(3,3,3,3,3,3))                    # 6 curves in total
```

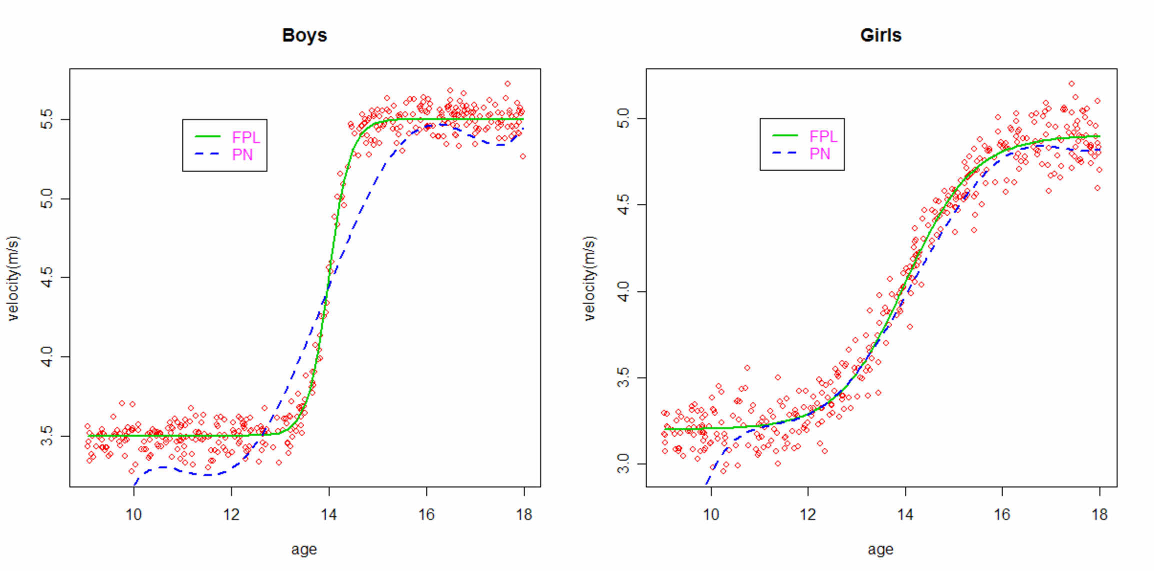
Codes F: Simulation

```
n=400
Age<-runif(n,9,18)                                                             # 400 skiers’ ages
x<-Age
plot(x,SSfpl(x,3.5,5.5,14,0.21)+rnorm(400,0,0.29^2),xlab="age",                # 400 boys’ values
ylab="velocity(m/s)",main="Boys",col=2)
curve(SSfpl(x,3.5,5.5,14,0.22),add=T,col=3,lwd=2)                             # FPL curve for boys
curve(V(x),add=T,col=4, lty=2,lwd=2)                                          # PN curve
legend(11,5.5,c("FPL","PN"),col=c(3,4),text.col=6,lty=c(1,2),lwd=c(2,2))
plot(x,SSfpl(x,3.2,4.9,14,0.7)+rnorm(400,0,0.34^2),xlab="age",                # 400 girls’ values
ylab="velocity(m/s)",main="Girls",col=2)
curve(SSfpl(x,3.2,4.9,14,0.7),add=T,col=3, lwd=2)                             # FPL curve for girls
curve(V(x),add=T,col=4, lty=2,lwd=2)                                              # PN curve
legend(11,5,c("FPL","PN"),col=c(3,4),text.col=6,lty=c(1,2),lwd=c(2,2))
```

Appendix II：Simulations in the purpose of comparing fitness between FPL and PN

Seeing from Figure 5.1, we find the results of comparison are completely different. For girls, it is nearly the same between the FPL curve and the PN curve; but for boys, it is not this situation. Thus it is necessary to do the simulation to see what happen.

400 skiers from age 9 to age 18 are existed at random, which is the uniform distribution. And the estimation results (Table 4.1 and Table 4.2) and the results of polynomial model will be used.



From the two figures above, for boys, the variation between the FPL curve and PN curve is due to the big jump in the FPL curve. Since the scale parameter in the FPL model is 0.22, the curve is very steep. For girls, the scale parameter is 0.7 and the curve is smooth. So the two curves have the similar shapes.