STATISTICS

Dalarna University

Master's Thesis 2007

Does randomly created relay teams perform worse than self-created?

------- An Empirical research on Falu Ski Winter Game

Author: Huan Liu and Yan Wu

Registered Nr: 840405 P206

820910 P404

Supervisor: Kenneth Carling

Date: 2007-08-22

TABLE OF CONTENTS

TABLE OF CONTENTS.....**1**
ABSTRACT.....**2**

1. Introduction……3

1.1 ski race in Dalarna……3
1.2 Background of Falu Winter Game……...4
1.3 Aim……...6

2. Data and imputation procedure...……...7

2.1 Data Description……...**7**
2.2 Handling of the missing observations……10
2.2.1 Imputation of age.....*11*
2.2.2 Imputation of Saturday's velocity.....*11*

3. Model and model specification……13

3.1 Model selection 13
3.2 Model fitting.....14
3.2.1 Model construction.....*14*
3.2.2 Estimate result.....*16*
3.2.3 ANOVA analysis.....*18*

4. Summary and conclusion.....19

Reference……..20
Appendix a……21
Appendix b…….....22

Abstract

In Sweden, cross-country skiing is a very popular sport in winter. This thesis studies on data collected in the Falu Winter Game which is held in the city of Falun in Sweden every winter. The Falu Winter Game is composed by two parts: Saturday's race and Sunday's race. The Sunday's race is a relay race in which two skiers make up a team. Most teams are self-created in the sense that two skiers agree to make up a team. However, some skiers for some reasons such as unexpected illness of companion would not be able to participate in the race. These skiers who have no team member may however report themselves to the organizers and the organizers made use of a lottery (randomize) these skiers into teams.

The aim of this thesis is to test whether there is an effect of randomizing skiers into teams. And by changing other influencing factors, how the effect will change. According to the data at hand, we use Linear mixed model to estimate the handicap. In this thesis, we use data from Sunday's race to see if randomly created relay teams perform worse than self-created.

After constructing three linear mixed models, we find that the output shows some interesting conclusion which reverses our consistent thinking. However, it is in accordance with the main purpose of holding Falu Winter Game: for increasing friendship and more fun from the race.

Key words: Falu Winter Game, randomized, self created, handicap, linear mixed model

1. Introduction

1.1 Ski race in Dalarna

The region of Dalarna, Sweden is best known as an excellent sporting resort in winter. When the area is covered in snow, this part is seen in a new light. The dense forest, mountains and lakes are frozen for a few months and become incredibly peaceful.

Figure 1a: A map of Sweden, with Dalarna in its centre. [1]

With this in mind it is not surprising that for Swedish this area is the most popular place for winter sports in the country. It is not just the tradition of downhill skiing which is taken care of here, but also cross-country skiing. Vasaloppet, which runs from Sälen to Mora in Sweden's Dalarna province, has

been an annual tradition since 1922, with historical roots from the 16th century. It is a classic Swedish ski race [2].

Figure 1b: A map of Dalarna

1.2 Background of Falu Winter Game

The cross-country race named “Lilla SS ”started in the early 1970s for children and youth. The ski competition in Falun (Falu Winter Games) was revamped three years ago by race leader Björn Helgåsen[3]. In 1990 there had been 1500 young cross countries skiers in Falun. However, the number is keep falling down every year. By facing this situation Helgåsen created Falu Winter Games to change the trend.

Surveys had shown that, the results and the prize of the competition are not very important for the competitors. They attend this race for fun and the race also offers good chances for every young amateur to communicate or make friends with each other and know more about skiing. They can learn how to cooperate with others. And it is also good for ski-masters to find out some young skiers who have ski talent. By now, the number of the young skiers is increasing and Falu Winter Games have

carried its point.

The Falu Winter Games consists of two races which are held on Saturday and Sunday. The competition on Saturday is a traditional one in classical style. All competitors are grouped according to their age. Skiers of the same age race the same distance and they start individually and their race time is clocked. The winner is the skier with the fastest time. The competitions on Sunday are relay races which are divided into different groups by age-interval and gender. In every age group, there are several teams (10-30). In each team there are two team members who race several laps on the race track. Each member needs to ski 3 laps, which may be 400 or 500 meters for each lap (Different age groups has different race distances). 400 meters distance for the skiers age 9-10, skiers whose age>10, the race distance is 500 meters.

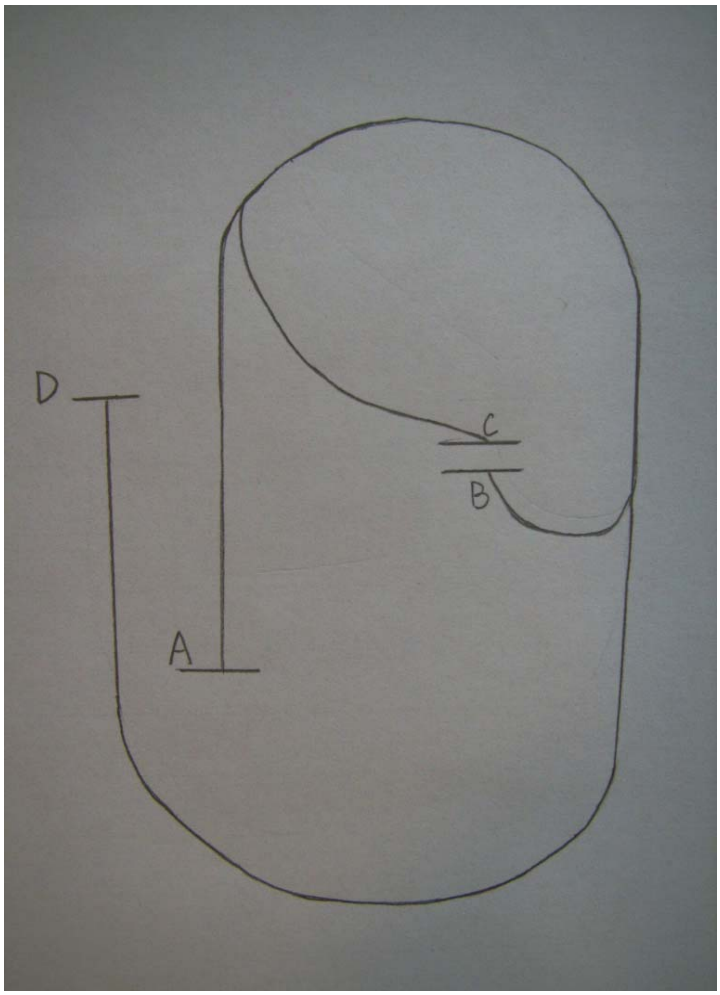

Figure 2: the map of Sunday's race track

From Figure 2, we know the racing track is roughly like this: the first skier starts from point A, ends at point B. This is the first lap. Then the second skier starts the relay from point C, ends at point B and the first one starts relay from point C again. Thus the lap 2 to 5 is repeated from point C to point B. However, at the sixth lap, the terminal is not at point A but at point D which is 50 meters away from A. Therefore we can clearly see from the map that the first and the sixth lap are longer than others and the sixth lap is the longest one.

Before the formal relay race, there is a prerun race which means the members in the team ski one lap each. The starting position in the following relay race is arranged according to the result rank of each team.

In the relay race the teams at the same age and gender start at the same time. Generally speaking, each competitor may attend the relay race with another team member who he already knows well. But for some reasons, there are a few competitors that cannot find his team member. Thus the organizers of the race may set these uneven skiers into a pool and teams are created by a lottery of the uneven skiers. The above content is derived from the interview of the holder.

These skiers combined by this way have not had the opportunity to practice together and the younger one might be put into an older age-interval. So these might imply a handicap, which we refer to as randomization effect.

1.3 Aim

The purpose of this thesis is to estimate if the handicap exists between self created teams and randomized teams and if the randomized teams perform worse than the self created teams. In this thesis the randomized is an influencing factor. We use data from all the skiers in every team in Sunday's races and construct three linear mixed models whose independent variables become more and more gradually. With the number of independent variables increasing in the models, the information about skier's abilities which included in the variables also becomes more. However, we set *RC* (random creating teams) as a fixed independent variable in three models. By comparing the change of the estimate of both *RC* variable and the random effect in

each model, it shows if the team created at random, it may bring out a positive or negative effect on the racing result. Furthermore, we analysis the direction of the effect and the reason of this phenomenon happens.

2. Data and imputation procedures

2.1 Data description

The data of this paper was collected from the website of Sweden Falu Idrottsklubb (Falu sports club) [4]. It provided all the information about the skiers' performance on the Saturday's individual race and Sunday's sprint in Falu Winter Game 2007. In this thesis, we collected 900 observations from 300 skiers' data and in 158 teams in the Sunday's race. We arranged the data into a new dataset which includes 20 variables. The proportion of gender: 45.3% is female, 54.7% is male. The proportion of participation: The proportion of team created: 98% of teams are self created, 2% of teams are created at random.

Table 1: name and description of the variables as well as some summary statistics

Variable	Describe	Short Name	Minimum	Maximum	Mean	Std.
ID number	the id number of each skier	id				
team	team number of each team in Sunday's race	team				
club	which club the skier comes from	club				
name		name				
gender	gender of each skier	gender				
age	age of each skier	age	8.000	21.000	12.320	2.472
race class (m)	the type of Sunday's relay race: 2*3(400) or 2*3(500)	class				
velocity of Saturday	each skier's velocity in Saturday, some are actual value some need to be imputed	vsat	66.180	366.750	216.310	58.487
distance of Saturday (km)	distance of Saturday's race	distance				
randomized created teams	the team is created by randomized or non-randomized 0 denotes by nonrandomized, 1 denotes by randomized	RC				
participation	participate in which day's race: Saturday, Sunday or both 0 denotes Saturday, 1	part				

	denotes Sunday, 2 denotes attend both					
velocity of prerun	velocity of each skier in Sunday's prerun	prerun	110.620	357.140	239.863	43.985
vs1,vs2,vs3	It denotes each team member's velocity in his/her first, second or third lap	vs1,vs2,vs3	78.910	379.750	254.536	51.883
team rank	the rank of each team in each member's first, second or third lap	rank				
member	It denotes team member's starting order, the first or the second	member				
av Sunday	It denotes the average of velocity of each skier in Sunday's race	av	139.530	362.480	254.343	36.584
t(min)	the racing time of one lap of each skier in Sunday's race	t	1.220	6.340	1.980	0.664

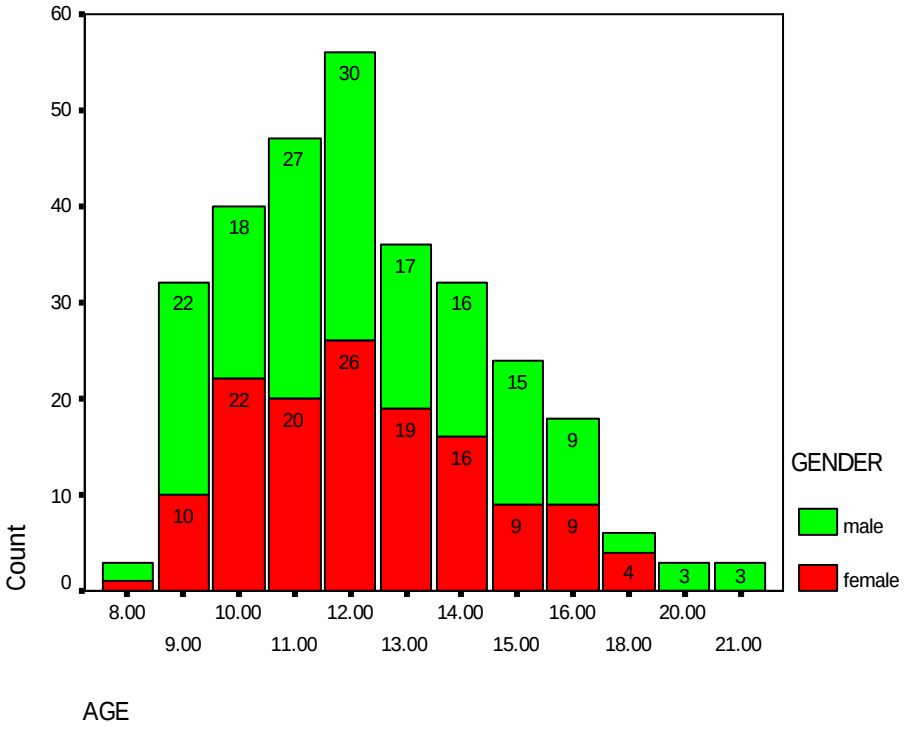

Figure 3: the ratio of gender and age of Sunday's race

From figure 3 above, we can see that the number of boy competitors is larger than girl competitors. And the skiers' age is most centralized on 10-14 years old. Therefore we selected the observations only from skiers whose age are under 15 years old.

2.2 Handling of the missing observations

Methods for handling incomplete data generally belong to one or more of the following categories [5]:

1. Case Deletion

Among older methods for missing data, the most popular is to discard units whose information is incomplete.

2. Imputation

The missing values are filled in and the resultant completed data are analyzed by standard methods. For valid inferences to result, modifications to the standard analyses are required to allow for the differing status of the real and the imputed values.

3. Reweighing

Reduce biases from case deletion by the judicious application of weights. After incomplete cases are removed, the remaining complete cases are weighted so that their distribution more closely resembles that of the full sample or population with respect to auxiliary variables.

In this thesis, according to the sample size is not large, we want to remain all the data and make the result precisely. Thus we choose the Imputation-Based Procedures to supplement the missing values in the data of variables.

2.2.1 Imputation of age:

In Saturday's game, the age of most competitors are known and accurate. However, in Sunday's race the skiers' ages are giving in age-intervals. For example, D17-18 shows the girl skier who is 17 or 18 years' old and H13-14 means the boy skier who is 13 or 14 years' old. For the purpose of acquiring the accurate age of every skier, we made use of imputation. So that no missing value in the age variable remains.

The specific method is that we found some other Falu ski races on the website. In the dataset we add "impute" as a new information variable. If the skier has accurate age we denote it as 0.

If his/her team member's age is accurate, we consider his/her age is same as the team member's and denote his/her "impute" variable as 1.

Setting 2 denotes we can find the skier in other races' list meanwhile his/her age on the list is accurate. We consider that age for the skier.

And setting 3 denotes that we pick up the maximum age of the age interval. For example the girl in D11-12, we impute that 12 is her age.

In the imputation of age procedure, we imputed 108 observations' data.

2.2.2 Imputation of Saturday's velocity

In the dataset the competitors who only attend Sunday's game do not have a velocity of Saturday. In these cases their Saturday's velocities are missing values. As a first step we have searched on the internet for other ski races in the vicinity of falun. Consider as an example in table 2. In this example two skiers at Falu Winter Game

also participated in the Race I.

Table 2: example of impute of Saturday's velocity

	Ski Race I	Falu Winter Game
Skier M	A	X
Skier N	B	C

For instance in the table M whose velocity of Saturday(X) is unknown. If we are succeeded in finding M in Ski Race I, we can calculate his/her velocity A in Race I. Then we need to find another skier N who attends both Race I and Falu Winter Game so that we can get the velocity B and C.

How can we choose N? By following 4 measure standardizations, we find the N in Race I who has relationship with M like this. The first measure standardization, N has the same gender, same age group and same club as M. If we can't find the N above-mentioned, we follow the second standardization. N has the same gender and same age group as M but maybe not in the same club as M. The third measure standardization is same age and club but maybe different gender. And the forth measure standardization is the same age but different gender and club. Then we can choose the N then calculate $x = \frac{AC}{B}$. We eliminate most of the missing values of velocity in Saturday.

For the remaining missing values we proceed to the following step. Saturday's velocity which can not be imputed after using the method above, we construct a linear model to do the imputation. The basic principle is using the data which *part* variable equal to 2 (attend both races in Saturday and Sunday) to estimate the Saturday's velocity of skiers who only attend the Sunday's game. (see the output in Appendix a).

We take a sample from ε in calculate the missing value of *vsat* by spacing random sampling method. After constructing the model and calculating the value, we can get the complete data of *vsat*.

Finally we imputed 105 observations of the data. Therefore after imputing age and velocity of Saturday's race there is no missing observation.

3. Model and model specification

3.1 Method Selection

Our aim is to analysis the handicap. Therefore we use data from all the skiers in Sunday's race and add *RC* as a variable to see if it significantly influences the racing time of every lap for each skier. If it is significant we conclude that the self created teams have a favor over other teams.

In general, the significant difference of randomized or self created teams may arise as an opportunity to practice together. So the teams which were created at random will cause the members to have less privity than the skiers from nonrandomized team. That is why the handicap may appear. We define *t* as dependent variable which measures the skier's racing time to complete one lap. We choose *vsat*, *lap*, *age*, *gender*, *prerun*, and *RC* as the independent variables which are ideal influencing factors to the racing result. We take each team as a unit (see the form of dataset in Appendix b) because by doing this it can eliminate the relativity of members in one team.

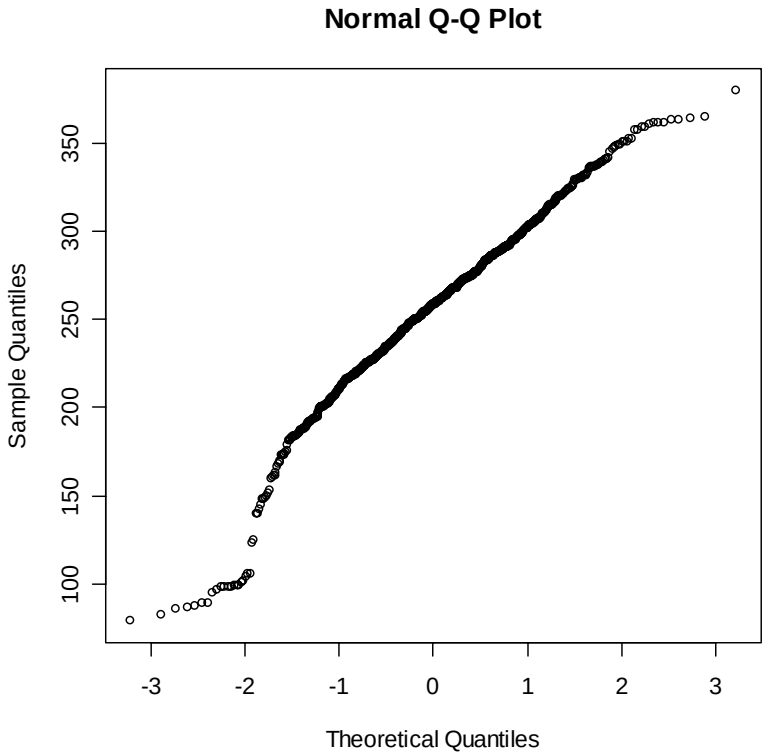

Figure 4: Normal QQ plot of t

From the shape of the Normal Q-Q Plot, we can consider the data of dependent variable obey the Normal distribution and the data be analyzed in this thesis is repeated measure data. There are several statistical methods used for analyzing repeated measures data. These include 1) separate analyses at each time point, 2) univariate analysis of variance, 3) univariate and multivariate analyses of time contrast variables, and 4) mixed model methodology [6]. In this paper, we will use mixed model methodology to analyze the data.

Therefore we choose the linear mixed model as the model form in this thesis. Mixed model are models where some of the independent variables are assumed to be fixed, while others are seen as randomly sampled from some population or distribution [7]. Thus mixed models are constructed by two parts: fixed part and random part. And mixed procedure is based on the general linear mixed model.

3.2 Model Fitting

3.2.1 Model construction

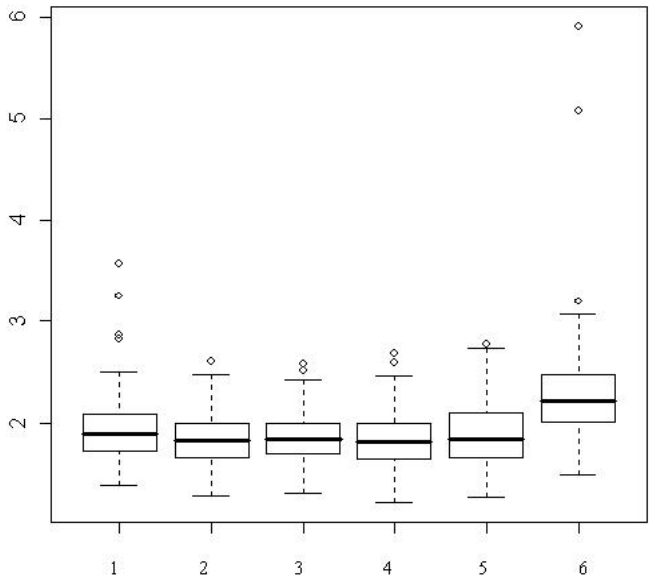

Figure 5: the box plot of racing time in six racing laps

The code in R is:

boxplot(t[age<15]~lap[age<15]+member [age<15])

From Figure 5, we can see the box plot of racing time in six laps. The first lap and sixth lap are longer than other laps while the sixth lap is the longest one. The reason we choose *t* as the dependent variable is because the difference of distance for every lap. (see Figure 2) Therefore we abandon choosing the velocity of Sunday's race as the dependent variable. Then we choose the racing time of every lap for each skier as the dependent variable for precise result.

And we set these factors as the independent variables.

1: *gender*,

2: *age*,

3: *vsat*,

4: *prerun*,

6: *rank*,

7: *lap*,

8: *RC*.

The data we choose is from all skiers in Sunday's race which can achieve our aim to predict the randomized groups. And the dataset we running in R of this model is arranged by every lap [Appendix b].

In this thesis, we construct three models in which we add independent variables gradually. However, the *RC* as a fixed independent variable exists in every model. By comparing the estimate of *RC*, random effect and fixed effect in three models, we can see the difference of racing result between self created teams and randomized teams. Furthermore, we can speculate the influence effect of randomize in team created.

The form of model 3 which has most independent variables in this thesis is:

$$t_{ilm} = \underbrace{RC_{ilm} + gender_{ilm} + vsat_{ilm} + age_{ilm} + lapo_{ilm} + prerun_{ilm}}_{fixed-effect} + \underbrace{(\underbrace{team}_{first-random-effect} (\underbrace{member}_{nested-random-effect}))}_{random-effect} + \varepsilon_{ilm}$$

Where,
i=1,2,…n denotes *team*, l=1,2,3 denotes *lap*, m=1,2 denotes *member*
RC denotes the randomly created teams, *team* and *member* are both random effect, *team* is the first random effect and *member* is the nested random effect.

Here we set both "*team*" and "*member*" as the random part, team is the first

random effect and member is the nested random effect. Because our train of thought is the racing time of the first lap of each skier may have interaction with his second and third lap. Also because of the co-operation between each team member in the relay race, the interaction exists between each member in every team. Thus, we try to decrease this interaction and make the racing time of each lap independently. Therefore we choose both "*team*" and "*member*" as the random part.

In this thesis we run the program of the linear mixed model in R with package "nlme".

3.2.2 Estimate result

For examine if team created at random has significant influence to the racing result, we compare the three models by adding variables gradually. By observing the running result, with variables becomes more and more, the estimate of *RC* is not change distinctly, the p value are all about 0.5 and not very significant. However part value of the random effect shows decreasing trend which is nice showing the variable control exists. Thus the construction of these models is proper.

Table 3: Estimate of the effect of randomly creating teams

	Model 1	Model 2	Model 3
RC	0.015	0.018	0.022
	(0.036)	(0.036)	(0.035)
Random effect			
Team	0.252	0.210	1.470e-05
Member	0.042	0.056	0.119
Fix Effect			
Gender	-0.117	-0.132	-0.007
	(0.048)	(0.042)	(0.025)
Prerun			-0.008
			(0.001)
Vsat			-0.001
			(0.0004)
Lapo			
Lapo 2	-0.112	-0.107	-0.085
	(0.029)	(0.030)	(0.032)
Lapo 3	-0.111	-0.111	-0.112
	(0.029)	(0.029)	(0.029)
Lapo 4	-0.069	-0.065	-0.042
	(0.029)	(0.030)	(0.033)
Lapo 5	-0.094	-0.094	-0.095
	(0.029)	(0.029)	(0.029)
Lapo 6	0.346	0.351	0.373
	(0.029)	(0.030)	(0.033)
Age			
Age 9		-0.075	-0.086
		(0.107)	(0.108)
Age 10		-0.208	-0.084
		(0.119)	(0.109)
Age 11		0.127	0.567
		(0.119)	(0.113)
Age 12		0.062	0.572
		(0.118)	(0.116)
Age 13		-0.051	0.609
		(0.121)	(0.124)
Age 14		-0.122	0.663
		(0.122)	(0.127)
Number of Obs: 900		Groups: member: 158	

* The value in brackets are standard deviation.

3.2.3 ANOVA analysis

		numDF	denDF	F-value	p-value
Modle 1	(Intercept)	1	487	6450.398	<.0001
	Gender	1	123	5.738	0.0181
	Lapo	5	487	76.111	<.0001
	RC	1	487	0.163	0.6865
Modle 2	(Intercept)	1	481	8661.732	<.0001
	Gender	1	123	7.765	0.0062
	Lapo	5	481	75.674	<.0001
	Age	6	481	6.613	<.0001
	RC	1	481	0.244	0.6213
Modle 3	(Intercept)	1	480	29668.917	<.0001
	Gender	1	122	28.489	<.0001
	Lapo	5	480	73.384	<.0001
	Age	6	480	26.836	<.0001
	RC	1	480	0.330	0.5659
	Prerun	1	122	199.208	<.0001
	Vsat	1	480	3.538	0.0606

The code in R is

```
b1<-lme(t~factor(gender)+factor(lapo)+factor(RC),random=~1|team/member,subset=age<15,na.action=na.exclude)
b2<-lme(t~factor(gender)+factor(lapo)+factor(age)+factor(RC),random=~1|team/member,subset=age<15,na.action=na.exclude)
b3<-lme(t~factor(gender)+factor(lapo)+factor(age)+factor(RC)+prerun+vsat,random=~1|team/member,subset=age<15,na.action=na.exclude)
lapo<-10*lap+member
```

4. Summary and conclusion

Regarding the relay race, we generally think that the results of non-randomized teams are better that the teams at random. Because the team members in the randomized teams are lack of cooperation privity with each other which will directly influence the racing results. But towards the investigation in the racing results of Falu Winter Game, the conclusion and our foreseer are opposite. When the team effect is small, the skiers' ability more measured through fixed effect. At the same time, the estimate of member effect quite large shows that the self created teams are not created by putting the "best" skiers together. The children who attend the race are not focus on if the team members are skillful enough and if they will take the crown. They may most hope that they can attend the race with their best friends and the main aim is for fun and for friendship. That is also the main purpose to hold Falu Winter Game. Through the race, it can make children promote their friendship and increase their favor on skiing.

Reference

[1] http://www.dalarna.se/template/turismPage____8381.aspx: Dalarna.se
[2] http://www.skiholiday.novasol.com/nov/999.nsf/0/ski_holiday_sweden_dalarna: ski holiday in Dalarna
[3] An interview of Björn Helgåsen, the race leader of Falu Winter Game, Lugnet sport center, Falun, 2007.3.17
[4] http://svenskidrott.se/Organisation.asp?OrgElementID=27195&CatId=422766 Sweden Falu Idrottsklubb (Falu sports club)
[5]. Joseph L. Schafer and John W. Graham. (2002). Missing Data: Our View of the State of the Art. P155-P159
[6] Littell, R.C., Henry P. R. and Ammerman, C. B. (1998). Statistical Analysis of Repeated Measures Data Using SAS Producedures. American Society of Animal Science. Florida Agric. Exp. Sta. Journal series no. R-05716. P. 1216.
[7] U Olsson (2002), Generalized Linear Models - an applied approach, Lund, Sweden

Appendix a

Table 1: estimate of the model in the imputation of Saturday's velocity

Dependent variable	**Coefficient**
vsat	
Independent variable	
prerun	0.528
av	0.152
age 10	33.522
age 11	87.784
age 12	35.061
age 13	–6.369
age 14	57.210
age 15	-20.555
age 16	-74.101
age 18	-58.105
age 20	104.367
age10:prerun	-0.152
age11:prerun	0.515
age12:prerun	-0.022
age13:prerun	0.151
age14:prerun	-0.056
age15:prerun	0.361
age16:prerun	0.527
age18:prerun	0.494
ε	

*Where, ε : residual random errors of the model. In R, we can get a series of the randomized number of the model.

Appendix b

Table 2: structure of data set in the linear mixed model

id	teamnumber	club	name	lap
1	1	Röros IL	HERMO Kristine W.	1
2	1	Röros IL	SKJEVDAL Kristina	1
3	2	Töcksfors IF	NILSSON Sofie	1
4	3	SK Bore	ERIKSEN Mikaela	1
5	4	Röros IL	GLöTHEIM Silje	1
6	4	Röros IL	HOLDEN Karianne	1
7	5	SK Leksand	OSKARSSON Sofi	1
8	5	SK Leksand	SARENMARK Alma	1
1	1	Röros IL	HERMO Kristine W.	2
2	1	Röros IL	SKJEVDAL Kristina	2
3	2	Töcksfors IF	NILSSON Sofie	2
4	3	SK Bore	ERIKSEN Mikaela	2
5	4	Röros IL	GLöTHEIM Silje	2
6	4	Röros IL	HOLDEN Karianne	2
7	5	SK Leksand	OSKARSSON Sofi	2
8	5	SK Leksand	SARENMARK Alma	2
1	1	Röros IL	HERMO Kristine W.	3
2	1	Röros IL	SKJEVDAL Kristina	3
3	2	Töcksfors IF	NILSSON Sofie	3
4	3	SK Bore	ERIKSEN Mikaela	3
5	4	Röros IL	GLöTHEIM Silje	3
6	4	Röros IL	HOLDEN Karianne	3
7	5	SK Leksand	OSKARSSON Sofi	3
8	5	SK Leksand	SARENMARK Alma	3

*The table only shows how we arrange the dataset in the model but it doesn't show the variable-items of our data.