RYERSON UNIVERSITY

Course Number	**Ckme136**
Course Title	**Data Analytics: Capstone Course**
Semester/Year	**Winter 2018**
Instructor	**Dr. Uzair**

Capstone Project	
Submission Date	**April 6, 2018**
Due Date	**April 6, 2018**

Name	Student ID	Signature*
Seyedpouria Modaresi	500442704	

Contents

Abstract:

https://github.com/pouriamo66/seyedpouria-modaresi

During the sport competitions, there are various factors, which effect on the performance of athletics. These findings will provide useful information to coaches and athletes to resolve weaknesses and get better results in future matches. This project is going to look at the results of the 2015 FINA World Swimming Championships and will explore the factors and impacts on the swimmers results and performances.

Introduction:

The project explores data from the 2015 FINA World Aquatics Championships, where we will quantify the relative speeds and variability among swimmers. We will perform a statistical analysis to assess the "current controversy" of the 2015 Worlds in which swimmers claimed that a slight current in the pool was affecting result.

Before we start working on our data, we need to have some background information about swimming computation.

1. **POOL**

look at the pool shows that ,there are tens lane in the pool ,the pool has 50 meters long for the event longer than 50 meters the swimmers turn around at end of the pool.

2. **STORKES**

There are four different strokes or ways of swimming, and each have own computation .they are

- Freestyle
- Breaststroke
- Butterfly
- Backstrok

Its important to know because of different in mechanism of strokes , swimmer swim in different speed

Other thing is Event at ther world Champion as being define by gender, distance, stroke

3. **ROUNDS**

The main question in this project is do swimmer swim faster in finals rounds. It is necessary to know that swimmer swam in which round and definition of rounds :

- **Heat** :First round
- **Semifinals**: Penultimate round in some events
- **Finals**: the final round

The main questions for this project are:

- Did the swimmers swim faster in final competition?
- Did lanes effect on swimmers performance?
- How does the performance of swimmers decline over long events?
- Does any differences between men and women in performance?
- Is there any significant differences between different countries swimmer performance?
- What is the relation between reaction time and performance?
- Which stroke was the most exciting one?

To answer to those questions this research explores relationships in the data, parameter estimation and computation of confidence intervals, and to establish further hypothesis for testing. Also After exploring the data, regression analysis will be use to find out relations between factors.

Literature Review :

From the beginning of the semester, I've been reading literature and watching webinars posted on Data analysis Sources and also posted on sport websites , to have better understanding of type of data , and what needs to be done to have better analyze.

Literatures about:

- ## swimming pools& swimming and Human factors

The first step in analyzing is data recognition. Therefore, it is important to know the rules and factors that are influence on swimmers performance.

1. **https://qz.com/761280/researchers-believe-certain-lanes-in-the-olympic-pool-may-have-given-some-swimmers-an-advantage/**

This website has useful information about how certain lanes in the Olympic pool may have given some swimmers an advantage.

2. https://swimswam.com/rio-olympic-test-event-showed-same-pool-bias-2-0/

This website mostly is about The "Current Controversy" of the 2015World Championships and answers this question: were Swimmers advantaged or disadvantaged depending on the direction and lane in which they swam. Moreover, gives information about all Strokes at the World Championships that are:

1. Freestyle
2. Breaststroke
3. Butterfly

- Backstroke

- ## R-studio

R is one of the most important language that Big Data students in Data science field should know or learn it. Ris a language that defines many of the complex concepts of statistics in a package and function and makes working with data much easier and faster.

1) **McKinney, Wes. R for Data Analysis. 1st ed. O'Reilly Media, 2012. Print.**

This book is mostly is about panda package in python and a globally recognized expert in python, so it has been more Valuable for concepts than an actual coding resource.
Learn R, Python & Data Science Online. Computer software.

2) **Datacamp. Datacamp.com. Web**
   Datacamp is a great online-based learning software that combines

- # Data Analysis

1. Wes McKinney . The Python for Data AnalysisMath. Statist. 33

(2013), no. 1, 1--. doi:10.1214/aoms/1349356084.
http://www.oreilly.com/catalog/errata.csp?isbn=9781449319793
this book is concerned with data manipulating, data processing, data cleaning, and data crunching in Python. It is a introduction to scientific computing in Python, and also it explains about the parts of the Python language and libraries that we need to effectively solve a broad set of data analysis problems.
2. CIND 110 & **CMTH 642** courses material

Topics of these courses include database architectures, formation of queries, queries themselves, data warehousing, relational database systems, NoSQL, and responsibilities of data management professionals.

- # Regression

1. CIND 110 & **CMTH 642** courses material

Topics of these courses include database architectures, formation of queries, queries themselves, data warehousing, relational database systems, NoSQL, and responsibilities of data management professionals.

2. "An Introduction to Machine Learning with Scikit-learn¶." Scikit-learn 0.18.1 Documentation. Scikit-learn 0.18.1. Web. http://scikitlearn. org/stable/tutorial/basic/tutorial.html.
Documentation and guides on one of the primary machine
Learning packages for python. Includes the primer on the code for
The package, as well as concrete examples of the package being
Used on the iris and digits data sets.

- # MYSQL

1. Seyed M.M. Tahaghoghi, Learning MySQL 65 ("O'Reilly Media, Inc.", 2006)

MySQL, the most popular open-source database, offers the power of a relational database in a package that's easy to set up and administer, and *Learning MySQL* provides all the tools you need to get started. This densely packed tutorial includes detailed instructions to help you set up and design an effective database, create powerful queries using SQL, configure MySQL for improved security, and squeeze information out of your data.

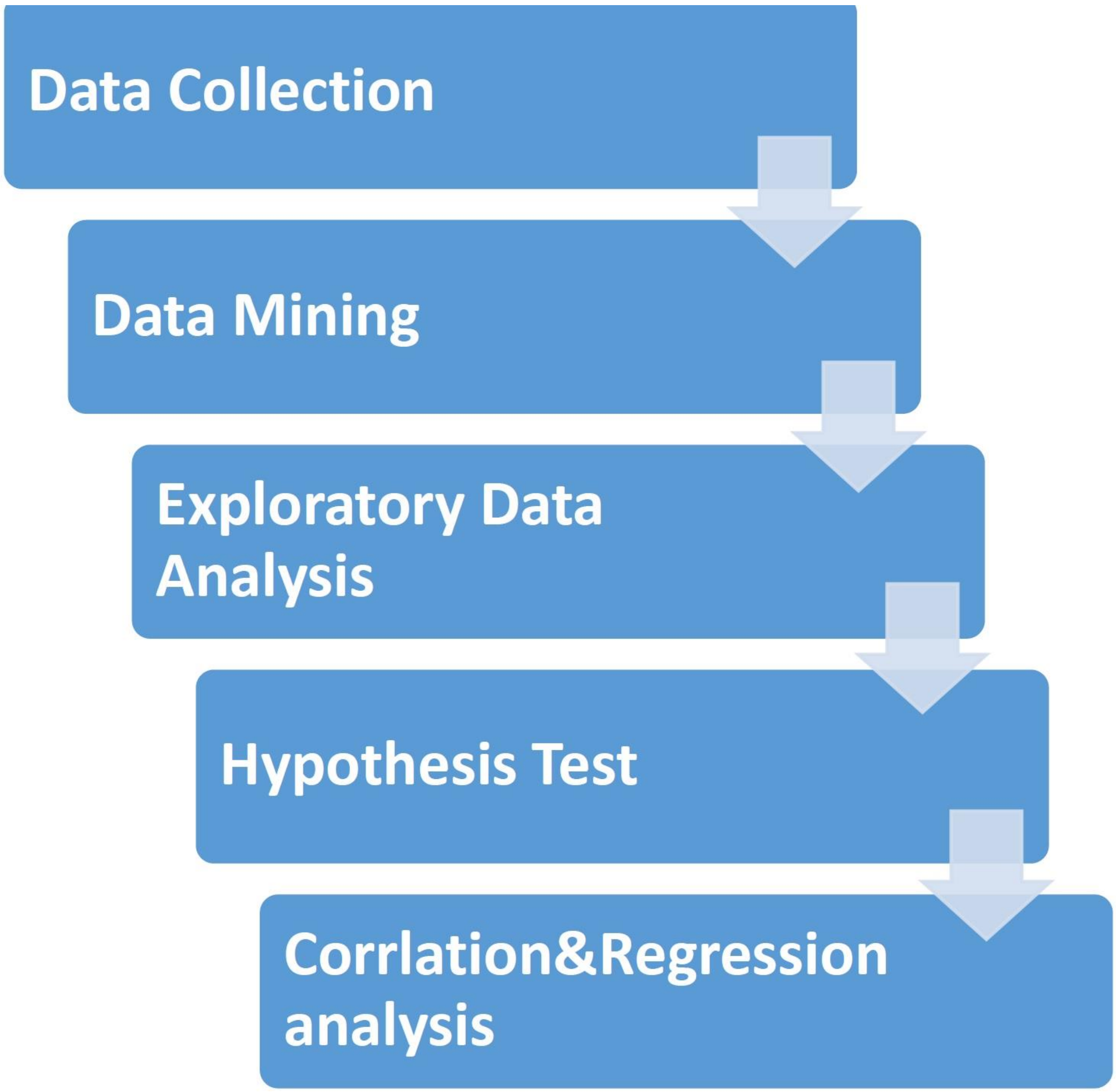
2. CIND 110 & **CMTH 119courses material**

Topics of these courses include database architectures, formation of queries, queries themselves, data warehousing, relational database systems, NoSQL, and responsibilities of data management professionals.

DATASET:

The dataset is about swimming results from 2015 World Aquatics Championships.Data is from OMEGA, official timekeepers of the World Championships. It Converted from XML to CSV format for having tidy dataset .
Accessed
http://www.omegatiming.com/File/Download?id=00010F0200FFFFFFFFFFFFFFFFFFFF08

Approach


Data Collection
Data Mining
Exploratory Data Analysis
Hypothesis Test
Corrlation&Regression analysis

Data Collection:

- Download the database in ‘csv’ format or from mysql
- Import the data to python

Data Mining:

- Identifying the source information.
- Picking the data points that need to be analyzed.
- Extracting the relevant information from the data.
- Drop missing values
- Interpreting and reporting the results.

Exploratory Data Analysis

- Do swimmers go faster in the finals?
- Do women swim the same way in semis and finals?
- Parameter estimates of difference between finals and semifinals
- How does the performance of swimmers decline over long events?
- How does the current effect depend on lane position?

Hypothesis Test

- With 95% confidence interval hypothesis test: Does lane assignment, affect performance?
- Parameter estimates of difference between finals and semifinals
- mean differences between odd and even splits
- Compute the 95% confidence intervals for the mean and median
- Find out p-value

Correlation Analysis

- Establish factors, such as average of split time,median , and any

Factors that prove relevant during the exploratory data analysis, and check the correlation between the factors

Regression Analysis

- Linear regression of average split time
- Test the fit of the model
- Conduct regression analysis using the independent factors

Result: (https://github.com/pouriamo66/seyedpouria-modaresi)

Do swimmers go faster in the final?

It seems swimmers swim faster in the finals .for example all of the personal best time of great American swimmers Michael Phelps happened in the final round of computations.

This arise the question, do swimmers swim faster in the finals than in other rounds? At face, this is reasonable question but we have to be more precise. For answering this question, first we need to solve some issues to make our result more clear.

- Which swimmers are we talking about? Individual swimmers or the whole field?
- Faster than heats? Faster than semifinals?
- For what strokes? For wat distances?

For being more specific in our question:

- Do individual swimmers swim faster in the finals compared to the other round? For Answering to this question we need to include all event in finals (50,100,200 meter freestyle, breaststroke, butterfly and backstroke)

We have almost ready to analysis our question but there is another issue that we need to solve it,

As it mentioned each strokes is swam at different way as we can see in figure (1) by ECDF by swim time of the rounds for the female 200 meter events.

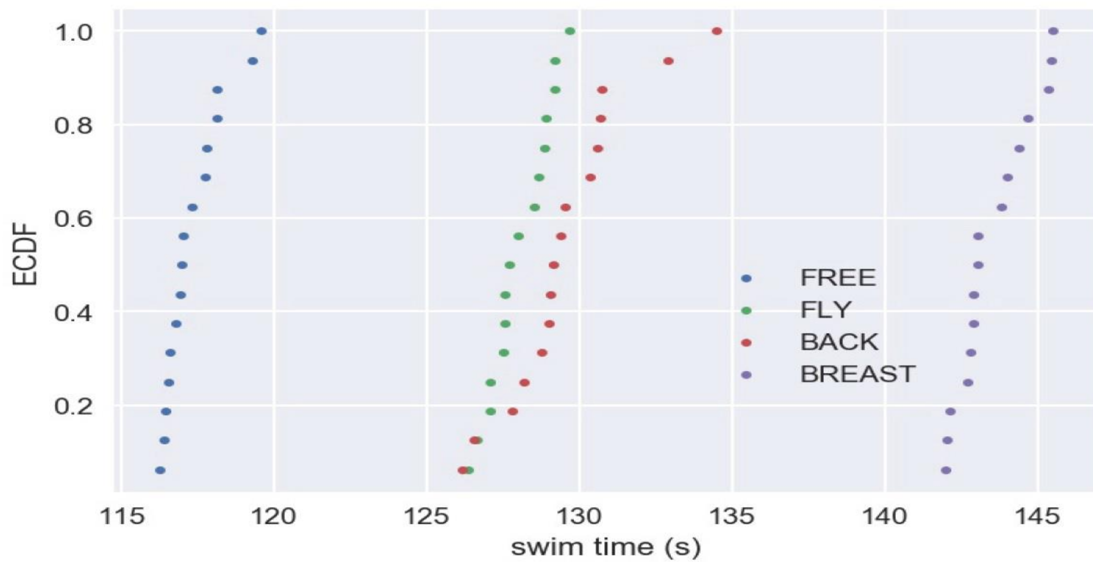


Figure (1)

Based on our EDA and parameter estimates, it is tough to discern improvement from the heats round to finals. We test the hypothesis that there is no difference in performance between the semifinals and finals. A permutation test is fitting for this. We will use p-value as the test statistic. Take an array of rounds times and an array of final times for each swimmer for each stroke/distance pair.

- Go through each array, and for each index, swap the entry in the respective final and semifinals array with a 50% probability.
- Use the resulting final and other rounds arrays to compute p-value.

STEPS:

1) To check this, we should first take a fast look at round saved in data to check the number of levels. To do so, first round data should be saved as a factor and then levels should be counted:

```
> swim1$round=as.factor(swim1$round)

> nlevels(swim1$round)

[1] 5
```

2) As number of levels are greater than two, we are obliged to test via ANOVA or Kruskal-Wallis to check weather "Split Swimming Time" are different through different "Rounds" or the differences are ignorable. To decide which test to be run, we should take care of Normality. As data size is big, by C.L.T it is capable of normality. As a result One way ANOVA is the solution and to do so, we run from regression parts:

```
> (l1=lm(formula = splitswimtime~0+round,data = swim1))

Call:

lm(formula = splitswimtime ~ 0 + round, data = swim1)

Coefficients:

roundFIN  roundPRE  roundSEM  roundSOP  roundSOS

   30.26     31.30     29.68     29.10     31.62

> summary(l1)

Call:

lm(formula = splitswimtime ~ 0 + round, data = swim1)

Residuals:

   Min     1Q Median     3Q    Max

-9.721 -1.928 -0.238  1.694 69.722

Coefficients:

        Estimate Std. Error t value Pr(>|t|)
```

roundFIN 30.26205   0.09378 322.70   <2e-16 ***

roundPRE 31.29770   0.03828 817.59   <2e-16 ***

roundSEM 29.68430   0.11262 263.57   <2e-16 ***

roundSOP 29.09750   1.80021   16.16   <2e-16 ***

roundSOS 31.62111   0.84863   37.26   <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.6 on 11359 degrees of freedom

(59 observations deleted due to missingness)

Multiple R-squared: 0.9867,   Adjusted R-squared: 0.9867

F-statistic: 1.687e+05 on 5 and 11359 DF, p-value: < 2.2e-16

3) The other result show that just the difference between "Final" round and "PRE" or "SEM" were significant:

Call:

lm(formula = splitswimtime ~ round, data = swim1)

Coefficients:

(Intercept)   roundPRE   roundSEM   roundSOP   roundSOS

30.2620   1.0356   -0.5778   -1.1645   1.3591

> summary(l1)

Call:

Min   1Q Median   3Q   Max

-9.721 -1.928 -0.238 1.694 69.722

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.26205    0.09378 322.696  < 2e-16 ***
roundPRE     1.03565    0.10129  10.224  < 2e-16 ***
roundSEM    -0.57775    0.14656  -3.942 8.12e-05 ***
roundSOP    -1.16455    1.80265  -0.646    0.518
roundSOS     1.35906    0.85379   1.592    0.111
```

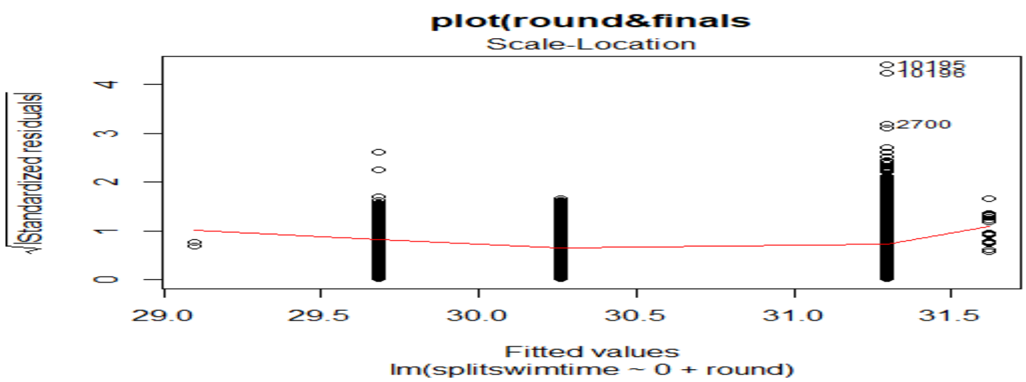


Figure (2)

As it is clear, p-value is less than 0.05 and this shows that the difference in different round is significant. By the estimate part it is obvious that "Final" round was faster than "SEM" and "SOP" but slower than "PRE" and "SOS" round

- **Did lanes effect on swimmers performance?**

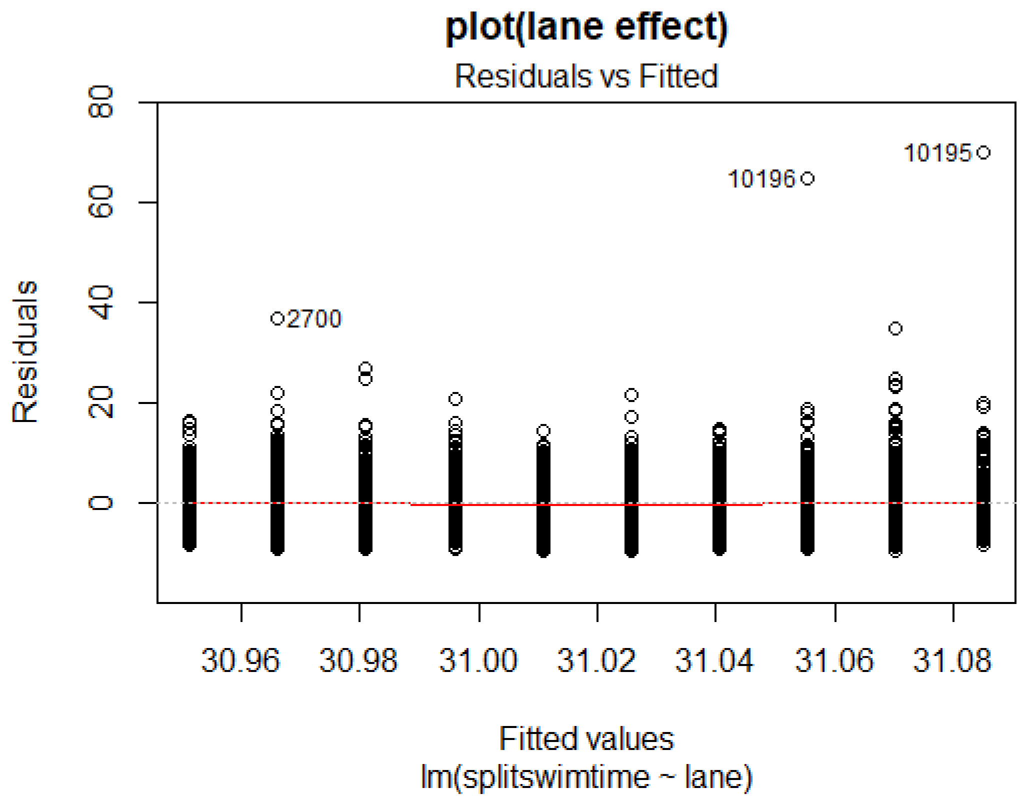
Apart from the physical differences, swimmers may gain psychological advantages, and the lane in which this happens depends on the swimmer. Some professional swimmers like to swim in lanes that are near the side, stating reasons which include being able to monitor all 7 other swimmers during 1 breath, as opposed to 2 breaths if the swimmer was near the middle lane. However, that swimmer specialised in freestyle, and the opposite may be true for breaststrokers and butterflyers, who may see other swimmers more clearly in the middle lanes.

STEPS:

1) To check this we use Regression models to understand weather swimmers in different lanes have recorded significant split swimming time or not.

```
> #Did lanes effect on swimmers performance?----
> swim1$lane=as.factor(swim1$lane)
> nlevels(swim1$lane)
[1] 10
> (l2=lm(formula = splitswimtime~lane,data = swim1))

Call:
lm(formula = splitswimtime ~ lane, data = swim1)

Coefficients:
(Intercept)       lane1       lane2       lane3       lane4       lane5
    31.4493     -0.3647     -0.3995     -0.5765     -0.8501     -0.7639
      lane6       lane7       lane8       lane9
    -0.3885     -0.5425     -0.1409      0.1057

> summary(l2)

Call:
lm(formula = splitswimtime ~ lane, data = swim1)

Residuals:
   Min     1Q Median     3Q    Max
-9.748 -1.905 -0.199  1.738 69.465

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.4493     0.1294 243.097  < 2e-16 ***
lane1        -0.3647     0.1670  -2.185 0.028943 *
lane2        -0.3995     0.1653  -2.416 0.015691 *
lane3        -0.5765     0.1665  -3.463 0.000536 ***
lane4        -0.8501     0.1640  -5.183 2.22e-07 ***
lane5        -0.7639     0.1653  -4.620 3.87e-06 ***
lane6        -0.3885     0.1648  -2.357 0.018433 *
lane7        -0.5425     0.1666  -3.257 0.001129 **
```

lane8	-0.1409	0.1671	-0.844	0.398938



Figure(3)

As it is clear the p-value is less than 0.05 and as a result it shows that Lane effect Swimmer Performance.

- **How does the performance of swimmers decline over long events?**

Recall the swimming pool is 50 meter long so for example during event of 800-meter swim, a swimmer swim 16 split. (The time is takes swimmer to swim one length of the pool called split).

Here is plot of Mikel Phelps split for his heat for 800-meterfree style Maybe most interesting feature of this plot is very fast split.

this is because the first split is different from last .the swimmers start race from starting blocks and this provide extra boost for their first split. The last split is also fast because the swimmer knows he can push hard because after the last split the race is over. Finally, as we can see during the race when the split gets longer the swimmer is slowing down.

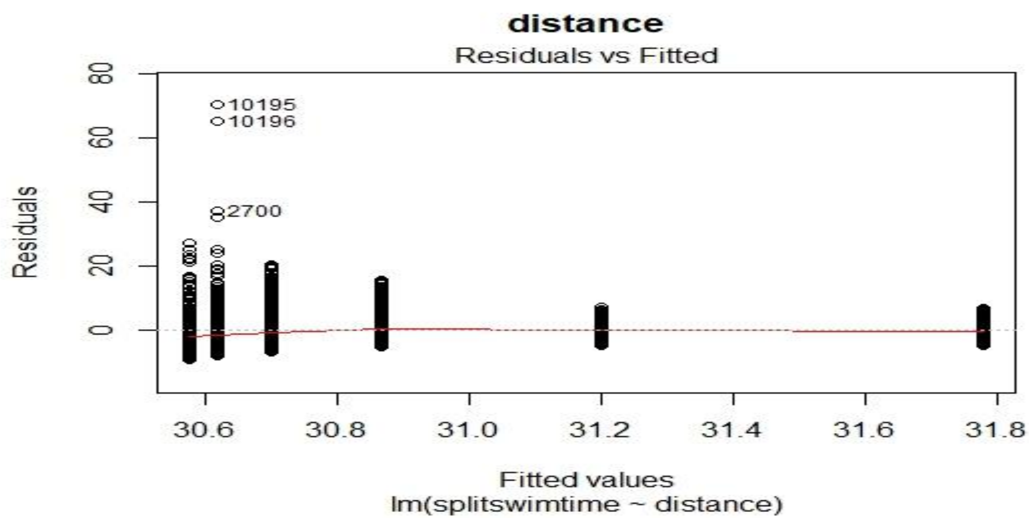
STEPS:

1) To check this we should run a regression model, use the distance as a continuous, quantitative variable, and see the effect of this on split swimming time. In this regards we run the model in two form: 1. with intercept an 2. Without intercept and then check if the model with intercept is good or not:

```
#How does the performance of swimmers decline over long events?----

> l3=lm(formula = splitswimtime~distance,data = swim1)
> summary(l3)

Call:
lm(formula = splitswimtime ~ distance, data = swim1)

Residuals:
   Min     1Q Median     3Q    Max 
-9.385 -1.917 -0.361  1.591 70.403 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.053e+01  4.978e-02  613.42   <2e-16 ***
distance    8.312e-04  6.260e-05   13.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.613 on 11362 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.01528,	Adjusted R-squared:  0.0152 
F-statistic: 176.3 on 1 and 11362 DF,  p-value: < 2.2e-16
```

2) l3.0=lm(formula = splitswimtime~0+distance,data = swim1)
> summary(l3.0)

```
Call:
lm(formula = splitswimtime ~ 0 + distance, data = swim1)

Residuals:
    Min      1Q  Median      3Q     Max
-16.561   6.357  20.972  25.559  98.125

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
distance 0.028954   0.000249   116.3   <2e-16 ***
```



Figure(4)

Both models are acceptable due to their p-value is less than 0.05 but model without intercept is much better because its adjusted R-square is much better. So we report that as the event gets longer, the split swimming time increase about 0.03 times, in other words, performance decline by the rate on 3% of its length.

- **Does any differences between men and women in performance?**

STEPS:

1) #Does any differences between men and women in performance?----

   ```
   > attach(swim1)
   > swim1$gender=as.factor(swim1$gender)
   ```

2)
   ```
   l4=lm(formula = splitswimtime~gender,data = swim1)
   > summary(l4)

   Call:
   lm(formula = splitswimtime ~ gender, data = swim1)

   Residuals:
      Min     1Q Median     3Q    Max 
   -8.496 -1.706 -0.016  1.324 68.434 

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
   (Intercept) 32.58568    0.04625  704.49   <2e-16 ***
   genderM     -2.89959    0.06291  -46.09   <2e-16 ***
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   ```

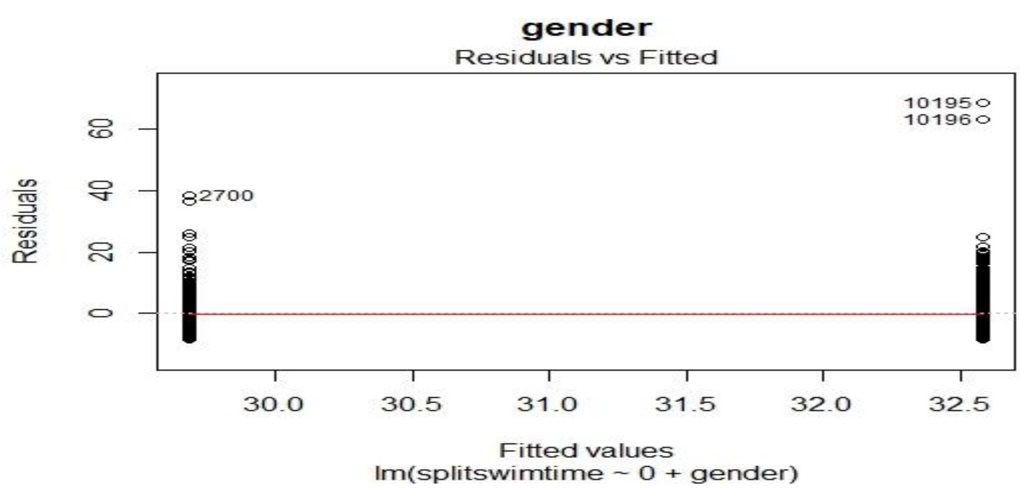


Figure (5)

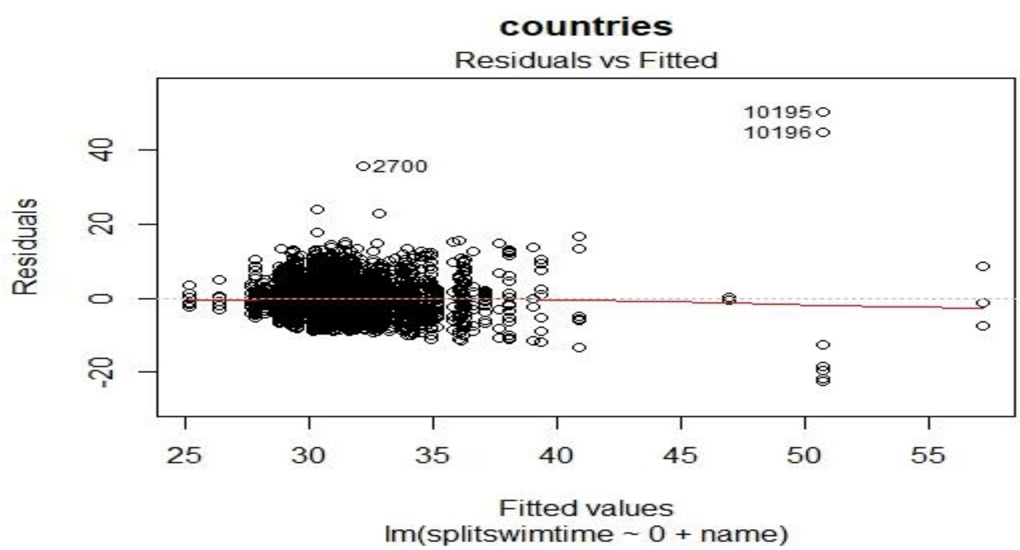
As is clear, significant differences detected between men and women and men perform better in this event.

- **Is there any significant differences between different countries swimmer performance?**

STEPS:

1) #Is there any significant differences between different countries swimmer performance?----

```
> swim1$name=as.factor(swim1$name)
> l5=lm(formula = splitswimtime~0+name,data = swim1)
> summary(l5)

Call:
lm(formula = splitswimtime ~ 0 + name, data = swim1)

Residuals:
    Min      1Q  Median      3Q     Max
-22.676  -1.712   0.015   1.600  50.304

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
nameAlbania          31.1326     0.6422  48.481   <2e-16 ***
```



Figure(6)

As is clear, significant differences detected between countries in this event

- **What is the relation between reaction time and performance?**

  To figure this out we should run the model twice and check the necessity of intercept:

STEPS:

1) #What is the relation between reaction time and performance?----

```
> l6=lm(formula = splitswimtime~reactiontime,data = swim1)
> summary(l6)

Call:
lm(formula = splitswimtime ~ reactiontime, data = swim1)

Residuals:
   Min     1Q Median     3Q    Max
-9.252 -1.968 -0.271  1.544 69.698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.7400     0.3855   51.21   <2e-16 ***
reactiontime  15.6507     0.5331   29.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.507 on 11360 degrees of freedom
  (61 observations deleted due to missingness)
Multiple R-squared:  0.07052,	Adjusted R-squared:  0.07044
F-statistic: 861.9 on 1 and 11360 DF,  p-value: < 2.2e-16
```

2) l6.0=lm(formula = splitswimtime~0+reactiontime,data = swim1)

```
> summary(l6.0)

Call:
lm(formula = splitswimtime ~ 0 + reactiontime, data = swim1)

Residuals:
    Min      1Q  Median      3Q     Max
-11.457  -2.449  -0.255   2.218  69.311

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
reactiontime 42.84964    0.05048   848.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Intercept is not needed and the relation is significant. It is shown that as the reaction time increases, performance declines at the rate of 42.85 times of reaction time.

- **Which stroke was the most exciting one?**

STEPS:

1) #Which stroke was the most exciting one?----

```
> swim1$stroke=as.factor(swim1$stroke)
> l7=lm(formula = splitswimtime~0+stroke,data = swim1)
> summary(l7)

Call:
lm(formula = splitswimtime ~ 0 + stroke, data = swim1)

Residuals:
   Min     1Q Median     3Q    Max 
-9.251 -1.871  0.000  1.820 66.868 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
strokeBACK   30.36854    0.10406   291.8   <2e-16 ***
strokeBREAST 34.15151    0.09731   350.9   <2e-16 ***
strokeFLY    29.46873    0.10222   288.3   <2e-16 ***
strokeFREE   30.44069    0.04094   743.6   <2e-16 ***
strokeMEDLEY 32.98095    0.09407   350.6   <2e-16 ***
---
```

It is shown that fly stroke was the most exciting one in this event. Now we should check weather others are significantly different or not. Of course at least one is different from others. And it is shown that Fly stroke is significantly more exciting than others:

2) swim1$stroke=relevel(x = swim1$stroke,"FLY")

```
> l7.0=lm(formula = splitswimtime~stroke,data = swim1)
> summary(l7.0)

Call:
lm(formula = splitswimtime ~ stroke, data = swim1)

Residuals:
   Min     1Q Median     3Q    Max 
-9.251 -1.871  0.000  1.820 66.868 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   29.4687     0.1022 288.298  < 2e-16 ***
strokeBACK     0.8998     0.1459   6.169 7.11e-10 ***
strokeBREAST   4.6828     0.1411  33.180  < 2e-16 ***
strokeFREE     0.9720     0.1101   8.827  < 2e-16 ***
strokeMEDLEY   3.5122     0.1389  25.284  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

This approach did provide information about factors, which effect swimmers performances. This shows that the difference in different round is significant but it does not mean swimmer swam faster at final. In addition, the results show us the performances are different between men and female, countries and lane. Our result also shows that fly stroke was the most exciting one in this event. At the end, the results also shows us some Swimmers were advantaged or disadvantaged depending on the direction, lane and even the time of event in which they swam.. As a result, swimmers have to be prepared for each match and analyze their previous performances with right and accurate analysis they can have a better record in the future tournament and also fix their weaknesses.

Appendix:

i. Dr.Uzair questions &student respond

1. If independent variable(e.g. gender) is categorical then you have to use logistic regression (glm function).

Due to what is written in our pamphlet if dependent variable is categorical, then we should use logistic regression:
(attached pic)

And in all types of regression given set of independent variables may be real valued, binary valued, or categorical valued. So the following line is correct:
> l4=lm(formula = splitswimtime~gender,data = swim1)
We also had some examples in our Module 4.1 that had sex as independent variable.

2. > l3.0=lm(formula = splitswimtime~0+distance,data = swim1): Why do you need 0+distance as independent variable. Why not only distance?

As in solution we take 2 steps:

1. Detecting the best model: the model is good that has the highest . The strategy is back ward. First all parameters (i.e. intercept and distance) is entered and then intercept removed from model. The firs model leads into and the second model leads to . this means the second model is better and we should go farther by it. The independent parts of lm function is written 0+distance to remove intercept from model.
2. The result is gotten.

3. Performance difference based on countries: How do you interpret the coefficient?

As different swimmers swam for different countries, their performance can be consider as sample of their countries performance. On the other hand name variable that contains countries name is categorical and if intercept is omitted the coefficient would be the mean value of performance. As a result the coefficient in result shows the mean value performance that their swimmer gained that tournament.

4. In the beginning you mentioned one-way anova will be applied. However, I couldn't see anova being used anywhere later.

As mentioned in previous question, when a categorical variable uses in linear regression, especially in simple linear regression, the coefficient would be mean values of dependent variable in each level of independent variable. Concentrate on simple linear models as all finding were based on!!! In model with intercept linear regression consider one of independent variable level as a based level and the mean value of dependent variable at that level would be consider as intercept estimation. Then the coefficients in other levels shows the difference of mean value of dependent variable at each level with the based level calculated as intercept. In model without intercept , the coefficients shows mean values of dependent variable in each level of independent variable.

In both cases, the F- statistics that are calculated based on the model is as F-statistics in anova. They are not equivalent and they are completely the same. Because the SSR needed for linear model F-statistic is the differences of observations from its predictions. When categorical variable used in simple linear regression, the predictions would be the mean value of dependent variable in the level. So SSR changes into SST gained in treatments. So using lm function and running F-statistics from is one-way anova.

5. Can you please explain how you interpret the outputs of lm model?

The lm model shows the linear relation of two sets of variable: dependent and independent variables. In fact it shows the effect of independent variables on dependent variables. JUST ONE WAY DIRECTED EFFECT!!!! For bidirectional effect we should use correlation measures.

The outputs from an lm function has a lot to say. The direct outputs show just the measure of effect. Summary of the lm function show the significance of model and show us weather it is useful to consider the linear relation or one can deny the relations. It also support us with and that shows the amount of variation that is modifiable by using the relation gained with model.

Plot of the lm model shows different ideas, such as outlier, the effective observations, normality of errors, Correlated independent variable or correlated error terms. So it can be used to adjust the model.

The coefficient gained from the lm model show the effect that an independent variable has on the dependent one when other independent variable is constant. The effect is defined as the amount of variation to the dependent variable when an independent variable changes just one unit while others are fixed.

ii. MYSQL

- (https://github.com/pouriamo66/seyedpouria-modaresi)
- Also attached

References:

i. https://qz.com/761280/researchers-believe-certain-lanes-in-the-olympic-pool-may-have-given-some-swimmers-an-advantage/

ii. .https://swimswam.com/rio-olympic-test-event-showed-same-pool-bias-2-0/

iii. McKinney, Wes. Python for Data Analysis. 1st ed. O'Reilly Media, 2012. Print.

iv. Ian Goodfellow, Yoshua Benjino, and Aaron Courville, *Deep Learning*, Unpublished: Book in preparation for MIT Press, http://www.deeplearningbook.org, 2016

v. .Wes McKinney . The Python for Data AnalysisMath. Statist. 33 (2013), no. 1, 1--. doi:10.1214/aoms/1349356084.

vi. http://www.oreilly.com/catalog/errata.csp?isbn=9781449319793

vii. CIND 110 & CMTH 642 courses material

viii. . "An Introduction to Machine Learning with Scikit-learn¶." Scikit-learn 0.18.1 Documentation. Scikit-learn 0.18.1. Web. http://scikitlearn. org/stable/tutorial/basic/tutorial.html.

ix. Seyed M.M. Tahaghoghi, Learning MySQL 65 ("O'Reilly Media, Inc.", 2006)

x. CIND 110 & CMTH 119courses material

xi. Julin L, Dapena J. Sprinters at the 1996 Atlanta Olympic Games did not hear the starter’s gun through the loudspeakers in the starting blocks. New Stud Athl. 2003;18(1):23–7.

xii. Omega Timing Web site [Internet]. Bienne: Omega SA; [cited 2015 Aug 5]. Available from: http://www.omegatiming.com. http://www.acsm-msse.org654 Official Journal of the American College of Sports Medicine APPLIED SCIENCES

xiii. Fe´de´ration Internationale de Natation. FINA Handbook 2015-17. Lausanne (Switzerland): FINA; 2013. p. 392.