INTRODUCTION

The models described in this report were created to predict the total number of medals that will be won by each country in the 2016 Olympics held in Rio de Janeiro. The models are based on information from 2000 to 2016 Olympic Games. The report will present linear mixed and generalised linear mixed models. The observations is figure 1 are correlated and a model that takes this into account could be more effective in presenting the data.

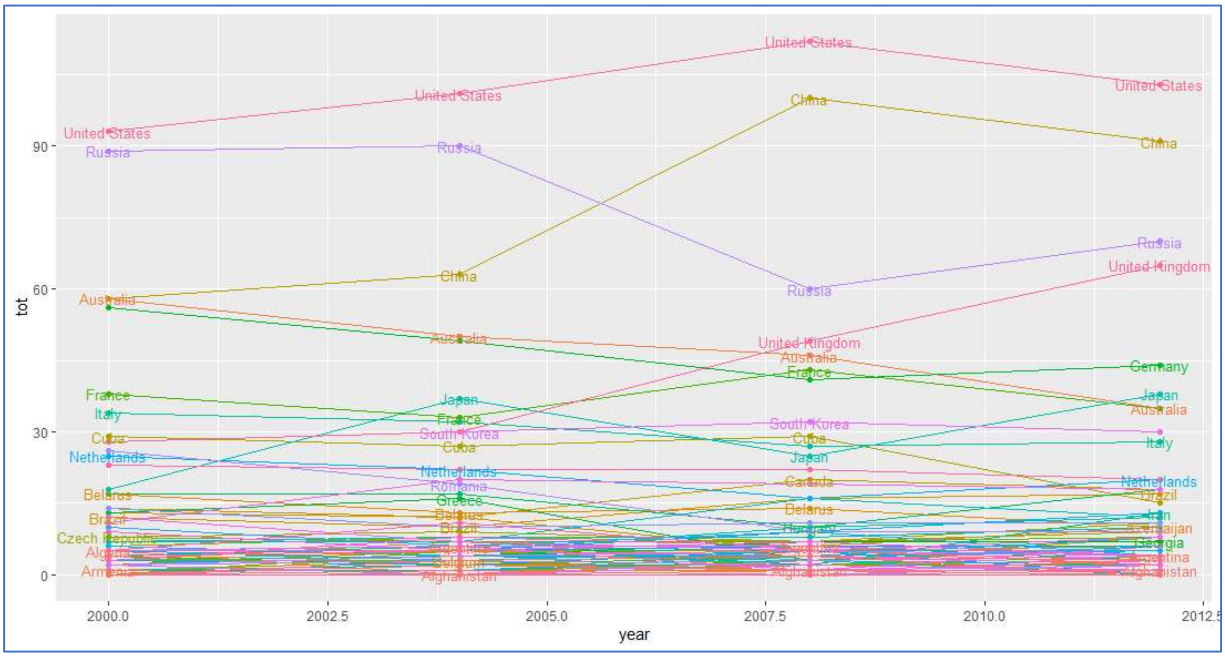


Figure 1

The second part of this report is an analysis of different models, followed by a conclusion that details the preferred model to predict the number of medals won by a country in the 2016 Olympics.

For a description of variables in the data set and exploratory analysis please refer to DanieleMolinari/Generalised-Linear-Models-Poisson-and-Negative-Binomial-distributions: Models to predict the total number of medals that will be won by each country in the 2016 Olympics held in Rio de Janeiro. (github.com)

ANALYSIS

The data set has been transformed into a longer format. The observations are not separated by year but instead they are all together in the same column. Furthermore, some variables have been log transformed.

In this phase, models will be created using the train data from 2000 to 2012. The first model, called LMM1, is a general model created using *glmer* function of the package *lme4*. The response is *tot* and all the explanatory variables have been used. This model considers a random effect for the intercept and the summary of this model is the following:

```
Linear mixed model fit by REML ['lmerMod']

Formula:
tot ~ gdp + pop + athletes + host + soviet + comm + muslim +
    oneparty + (1 | country)
   Data: train.data

REML criterion at convergence: 2749.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7113 -0.2463 -0.0047  0.2470  5.4773

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 110.75   10.524
 Residual              16.15    4.018
Number of obs: 431, groups:  country, 108

Fixed effects:
             Estimate Std. Error t value
(Intercept) -44.73064   10.16725  -4.399
gdp          -0.07524    0.37867  -0.199
pop           2.14069    0.73451   2.914
athletes      1.33486    0.41558   3.212
host         23.30255    3.16786   7.356
soviet        0.56615    4.32723   0.131
comm          6.62497    3.44274   1.924
muslim       -0.66525    2.86880  -0.232
oneparty     16.46349    6.64708   2.477

Correlation of Fixed Effects:
         (Intr) gdp    pop    athlts host   soviet comm   muslim
gdp      -0.224
pop      -0.348 -0.416
athletes  0.022 -0.206 -0.104
host      0.128 -0.168 -0.218 -0.156
soviet   -0.345  0.118  0.001 -0.025 -0.086
comm      0.046 -0.026  0.010 -0.091  0.186 -0.688
muslim   -0.301  0.031 -0.142  0.105  0.222 -0.139  0.173
oneparty -0.630  0.108 -0.131  0.012 -0.077  0.238 -0.281  0.026
```

From the output above it is possible to observe that not all the variables are statistically significative. The t-value for example of *gdp*, *soviet* and *muslim* have an absolute value lower than 1.96. This is also confirmed by the following confidence intervals and all the three variables contain zero. The same trend has been shown by confidence intervals calculated with the bootstrap method.

Confidence Interval Model 1		
	2.5%	97.5%
.sig01	8.87632422	11.7358422
.sigma	3.71664643	4.3380377
(Intercept)	-64.22706906	-25.3205601
gdp	-0.80100042	3.5398661
pop	0.72280027	0.56033933
athletes	0.53056614	2.1542889
host	17.22482468	29.3462704
soviet	-7.69245192	8.8422443
comm	0.04286601	13.1965100
muslim	-6.14037513	4.8211251
oneparty	3.78050475	29.1782317

The three subsequent models have been created by dropping not statistically significant variables one after another until reaching the fourth model that presents all variables with high t-value and none of the confidence intervals contain zero. At this point, the fifth and sixth models have been created with the same variables of model 4 but taking into account random effects also for the slope. LMM5 has an uncorrelated random effect between slop and intercept, whereas LMM6 has correlated random effects between the intercept and slop. LMM5 and LMM6 both showed problems of singularity and for this reason they will be not taken into account.

Figure 2 shows QQplot and residuals plot for model LMM4 and it is possible to notice that linearity and normality might be questionable, and no outlier are noticeable.

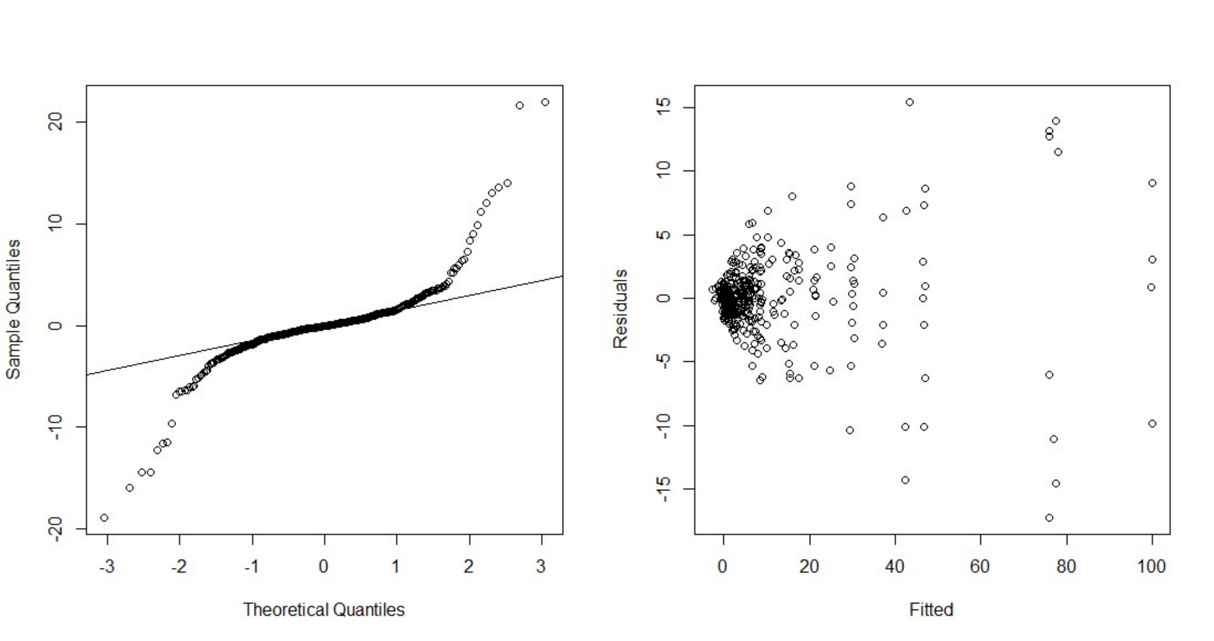


Figure 2

The following table compares all the four models in terms of AIC, BIC and p-value taken from the ANOVA function. It is possible to notice that the model with the lowest p-value is LMM4 which also has the lowest values of BIC. Therefore, it will be used to make predictions using the test data set.

ANOVA			
	AIC	BIC	Pr(>Chisq)
LMM4	2790.1	2822.7	
LMM3	2792.1	2828.7	0.8338
LMM2	2794.1	2834.7	0.8721
LMM1	2769.0	2840.8	0.8734

The second part of the analysis has been performed using function *glmer* from package *lme4*. As in the previous part, a first general model, called GLMM1, has been created considering all the variables in the data set and a random effect for the intercept. Afterwards, predictors have been dropped one by one based, in this case, on p-value. The fifth model (GLMM5) has all statistically significant variables and, at this point, GLMM6 model has been created considering random effects for the slope uncorrelated with the random effect for the intercept. However, its output shows problems of singularity, and it will be not taken into account. The seventh model (GLMM7) instead considers the random effect of the intercept and slope being correlated and all the variables included in this model seem to be statistically significative. To compare GLMM5 and GLMM7 the ANOVA function has been used and the results is as follow:

```
Data: train.data
Models:
GLMM5: tot ~ pop + athletes + host + comm + (1 | country)
GLMM7: tot ~ pop + athletes + host + comm + (1 + athletes | country)
      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
GLMM5    6 1768.6 1793.0 -878.28   1756.6
GLMM7    8 1768.3 1800.8 -876.13   1752.3 4.2967  2     0.1167
```

We cannot reject the null hypothesis and the random effect for the slope does not seem to be significant, therefore, GLMM5 will be used to make predictions.

Figure 3 shows QQplot and residuals plot for model GLMM5 and it is possible to notice that this model better describe the data because the points are consistently aligned with the black line, confirming the non-normality shown in the previous figure.

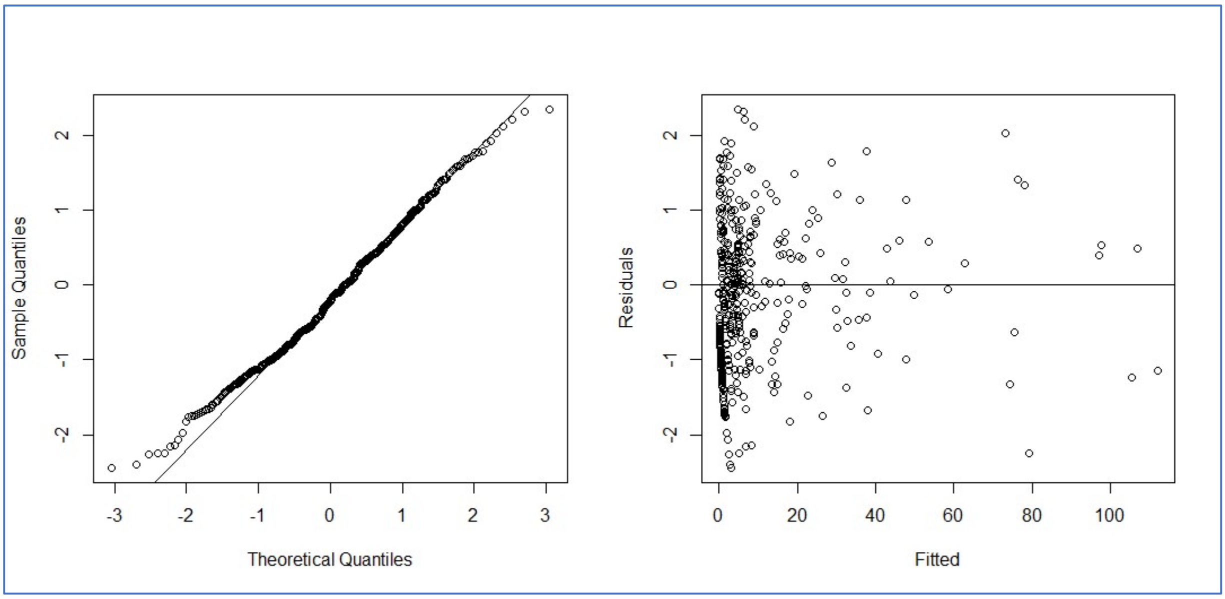
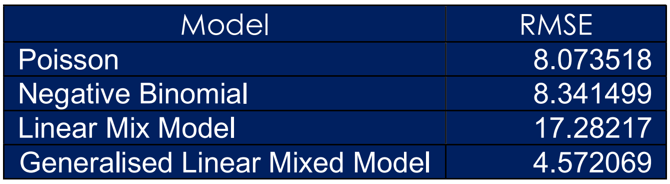


Figure 3

CONCLUSIONS

From report 1, that you can find in: DanieleMolinari/Generalised-Linear-Models-Poisson-and-Negative-Binomial-distributions: Models to predict the total number of medals that will be won by each country in the 2016 Olympics held in Rio de Janeiro. (github.com) two models have been taken into consideration: the one with Poisson distribution using *glm* function and the one with negative binomial distribution using *glm.nb* function. The root mean squared error have been calculated for these two models along with the LMM and GLMM models created in this report and they are reported in the following table. Generalised Linear Mixed Model has the lowest root mean squared error.



Model	RMSE
Poisson	8.073518
Negative Binomial	8.341499
Linear Mix Model	17.28217
Generalised Linear Mixed Model	4.572069

Figure 4 compares the model chosen described in report 1 and the model chosen in this report (GLMM5). They both have Poisson distribution but it is clear that the model created as a Generalised Linear Mix Model better follow the observations.

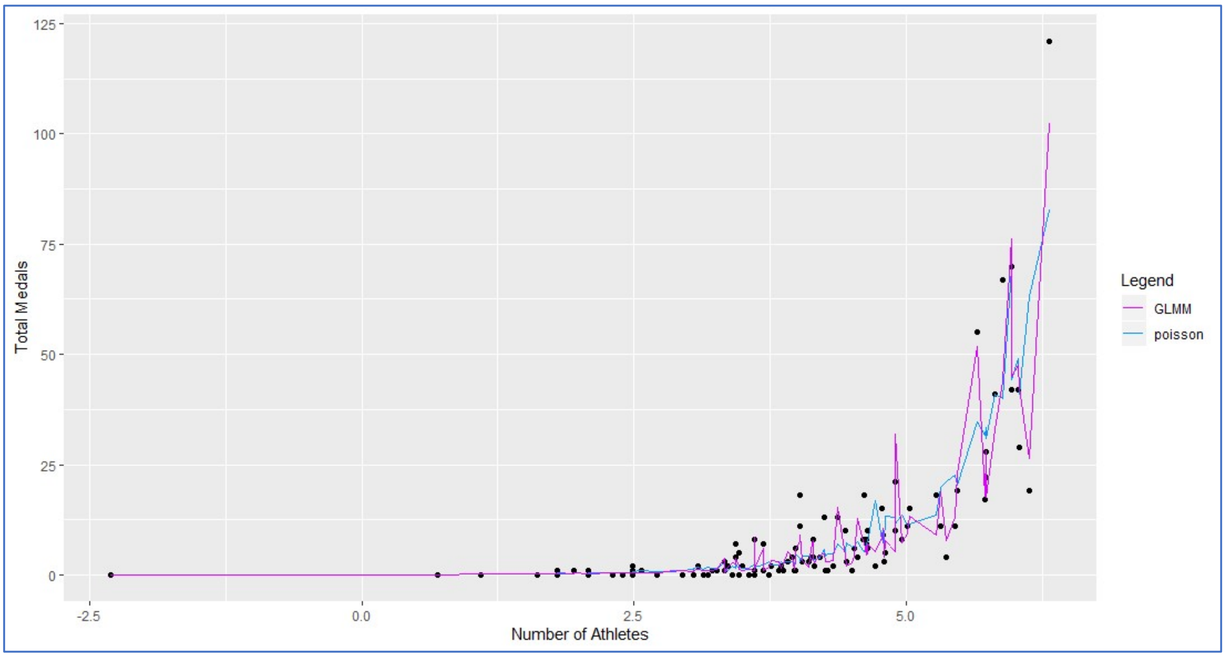


Figure 4

Figure 5 compares all four models, the two described in report 1 and the two described in this report. The plot shows the fitted values versus the observed values. A model that perfectly describes the data will have fitted values equal to the observed values, hence, the graph would show all the points lying on the black line. In this case the models created have some error in predicting the number of medals as seen in the previous table. However, the four models perform differently for each other. The LMM model presents some issue and should be probably more investigated. The other three have a more consistent behaviour and GLMM model in general has its points closer to the black line than the other two. That means that GLMM model better predicts the number of medals won by a country in the Olympic games of 2016.

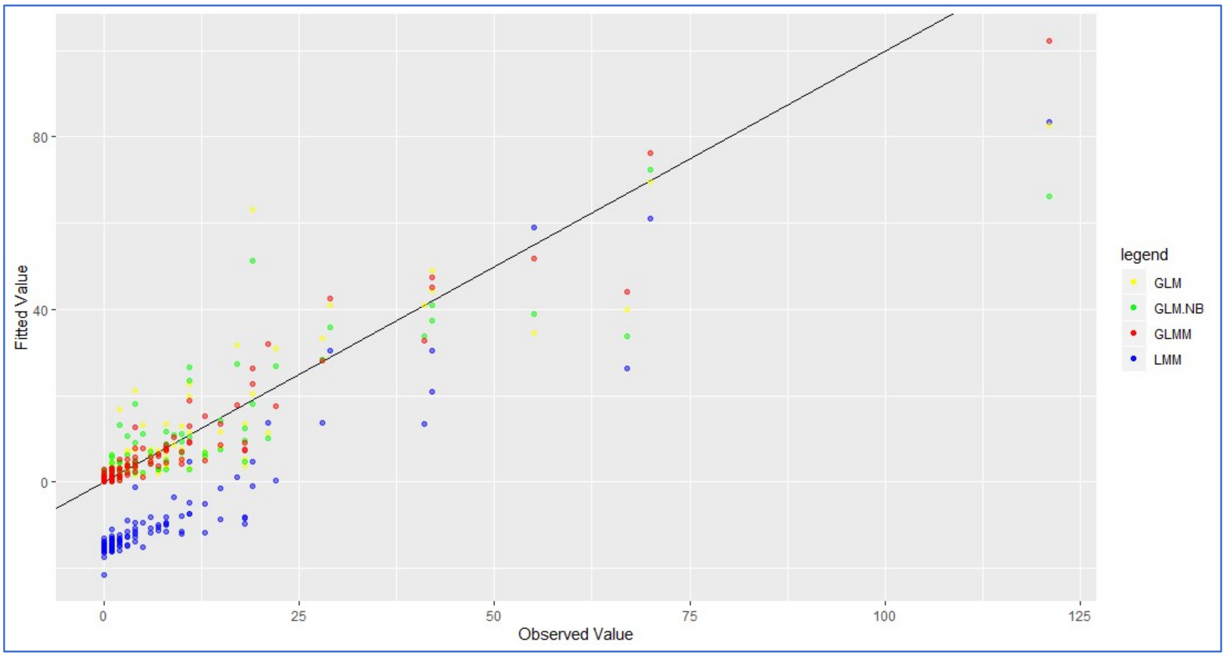


Figure 5: GLM = Generalise Linear Model using Poisson distribution; GLM.NB = Generalise Linear Model using Negative Binomial;
GLMM = Generalise Linear Mix Model; LMM = Linear Mix model

That also confirms what has been noticed in the figure 1. The observations are correlated within each other and there is a random effect for the intercept. Furthermore, as suggested from figure 2, a non-normal distribution better describes the data set.

FUTURE IMPROVEMENT

This analysis can be improved by adding more data to the data set about other Games, not only Olympics but also regional or local games, in order to have more information about athletes or any other factors which can affect their performance during Olympic games.

More detailed models can be created exploring for example the role of the other factors like *comm, muslim, oneparty* and *soviet*. Furthermore, Linear Mix Model probably deserve a more detailed investigation. Other distributions could be taken into account to create different models and try one of the methods suggested by the creator of *lme4* to deal with singularity. Better variable selection can be used for example starting with more simple models and gradually add predictors rather than dropping them from more complex models and checking thoroughly any interactions with factors. Finally, it would be equally interesting an examination with non-linear models.