GLMM Project - Prediction of number of medals won in 2016 Olympics using generalized linear mixed models

Preparer: Alexey Pankratov

Introduction

The aim of the GLMM Project is to build general linear mixed model that predicts the total number of medals won by countries. The model will be built on the dataset that includes data on the number of medals (total and gold) won by each country for 108 countries participating in the Rio 2016 Olympics, along with information on previous Olympic performance (from the 2000, 2004, 2008 and 2012 Games) and other variables. Other variables include:

- ***country***: the country's name,
- ***country.code***: the country's three-letter code,
- ***gdp***: the country's GPD in billions of US dollars during year YY. *Please note that original data has been transformed to from millions US dollars to trillions US dollars (i.e. divided by 1,000,000).*
- ***pop***: the country's population in millions in year YY. *Please note that original data has been transformed to from thousands to billions (i.e. divided by 1,000,000).*
- ***soviet***: 1 if the country was part of the former Soviet Union, 0 otherwise,
- ***comm***: 1 if the country is a former/current communist state, 0 otherwise,
- ***muslim***: 1 if the country is a Muslim majority country, 0 otherwise,
- ***oneparty***: 1 if the country is a one-party state, 0 otherwise,
- ***gold***: number of gold medals won in the Olympics,
- ***tot***: total number of medals won in the Olympics,
- ***totgold***: overall total number of gold medals awarded in the Olympics,
- ***totmedals***: overall total number of all medals awarded in the Olympics,
- ***altitude***: altitude of the country's capital city,
- ***athletes***: number of athletes representing the country in the Olympics,
- ***host***: 1 if the country has hosted/is hosting/will be hosting the Olympics, 0 otherwise.
- ***year***: year of the Olympics (2000, 2004, 2008, 2012 or 2016) – this variable is derived from the original table by transforming it.

In GLM Project, I focused on 3 generalized linear models: (1) Poisson regression model, (2) Zero inflated model (based on Poisson distribution) and (3) Basic Linear Model. However, the data is represented by multiple correlated observations per country. Therefore, independence assumption does not hold for abovementioned GLMs and result in biased standard errors and poorer prediction quality.

To correct this, I will build generalized linear mixed model that will include random effect for each country and allow to explore effect of each country separately. Accordingly, I will build the same models that I built initially but with country as a grouping factor, in particular – Poisson GLMM, Zero inflated GLMM and Linear Mixed Model (i.e. GLMM with Normal distribution and identity link function).

Exploratory analysis

Each row of original dataset contained observations for each country of all 5 Olympic games. In order to make analysis easier, I will make the data tidy by putting every observation in a separate row, i.e. each of the Olympic game is now in the separate row with additional variable "year" that specifies year of the Olympic game.

Now let's split the data into training data based on which I will build the model (games until 2012 including) and test data (2016 Olympics), on which I will check accuracy of prediction.

After transforming the dataset, let's explore relationship between variables starting with numerical variables in the training dataset.

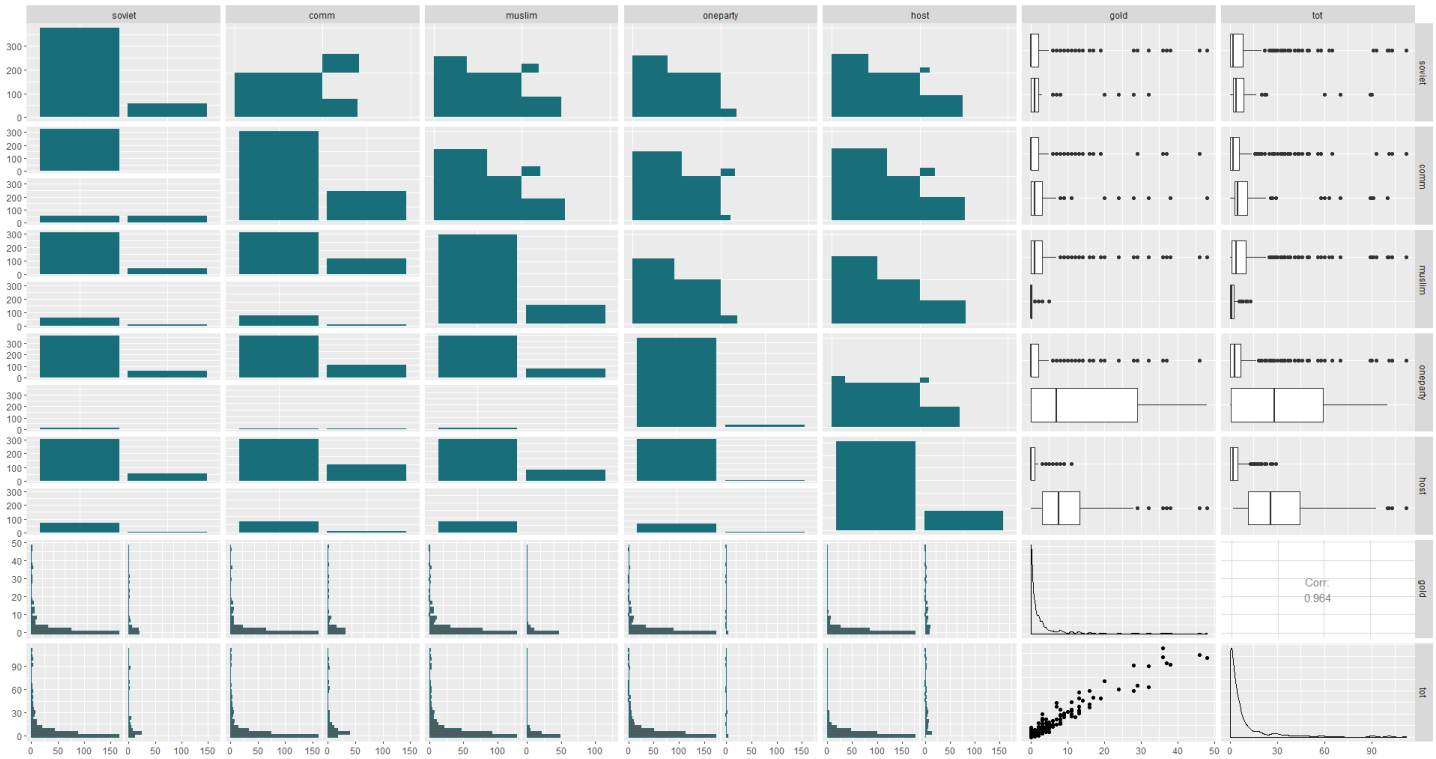
As it can be seen from the plot, there is a very strong positive correlation between total number of medals won and the number of gold medals won (0.964). Also you can notice that other variable correlate with total and gold medal won in a similar way, therefore I will focus only on prediction of total number of medals won as prediction of total number of all medals won in the Olympics will be also a good prediction of the number of gold medals won.

Also, there is a strong positive correlation between number of medal won (both total and gold) and the number of athletes representing the country. (correlation is 0.887). And GDP is a strong indicator of number of medals the country wins in the Olympics (correlation coefficient is 0.76)

At the same time, population is not as strongly correlated with medals won (corr. coefficient 0.416) as GPD and number of athletes.

Finally, it looks like altitude has almost no correlation with any other numerical variable (corr. coefficient is approx. -0.1 with all variables)

Now let's explore relationship between categorical variables



From the plot it looks like that ex-soviet and communist countries have slightly more medal on average than non-communist and Muslim countries have lower number of medals won than non-muslim, but at this stage it is not clear if this variable will be statistically significant or not. On the other hand, countries that hosted Olympics tend to have larger number of medals won then those that did not host. Also, countries with one-party system tend to win more medals on average. But this probably due to China as there are only three countries with one-party system, where the other two (Cuba and Eritrea) are dwarfed by China with regards to most of the variables.

Selection and assessment of generalized linear mixed models

First, I will try to build preliminary Poisson GLMM with random effect (random intercept) by country. I will build the model that includes all variables first. As I are building model to predict number of medals, i.e. counts, not rates, I will not use offset in the model. However, I also could have predicted number of medals won per athlete or per 1 person and in this case, I would have used expected number of medals won allowed for differences in population or number of athletes presenting the country as offset in the model. But instead, here I will have population and athletes directly as predictors in the model rather than offset.

```
> summary(poisson.glmm.preliminary)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: tot ~ athletes + gdp + pop + host + oneparty + muslim + soviet +      comm + altitude + (1 | country)
   Data: oldat_training
Control: glmerControl(check.nobs.vs.nRE = "ignore")

     AIC      BIC   logLik deviance df.resid
  1905.0   1949.7   -941.5   1883.0      420

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2916 -0.6962 -0.1661  0.4925  2.8826

Random effects:
 Groups  Name        Variance Std.Dev.
 country (Intercept) 0.9642   0.9819
Number of obs: 431, groups:  country, 108

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.0677474  0.1793988   0.378   0.7057
athletes     0.0023321  0.0003066   7.607 2.80e-14 ***
gdp          0.0288417  0.0150505   1.916   0.0553 .
pop          0.8654466  0.6438808   1.344   0.1789
host1        2.0082483  0.2874183   6.987 2.80e-12 ***
oneparty1   -0.4547422  0.7128917  -0.638   0.5235
muslim1     -0.4188141  0.2942735  -1.423   0.1547
soviet1     -0.3153519  0.4123433  -0.765   0.4444
comm1        1.4218997  0.3403783   4.177 2.95e-05 ***
altitude     0.0001069  0.0001710   0.625   0.5318
```

Looking at the summary of the Poisson regression model (Wald test), only “athletes”, “host” and “comm” variables in the above model has a p-value above 0.05. It means that only these variables are statistically significant and therefore all other variables can be removed from the model.

Now let's make a new Poisson GLMM that includes only statistically significant variables:

```
> summary(poisson.glmm.final)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: tot ~ athletes + host + comm + (1 | country)
   Data: oldat_training

     AIC      BIC   logLik deviance df.resid
  1904.5   1924.8   -947.2   1894.5      426

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2059 -0.6778 -0.1633  0.4931  2.8822

Random effects:
 Groups  Name        Variance Std.Dev.
 country (Intercept) 1.015    1.008
Number of obs: 431, groups:  country, 108

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0373681  0.1432760   0.261    0.794
athletes    0.0024024  0.0003022   7.949 1.88e-15 ***
host1       2.2068986  0.2788467   7.914 2.48e-15 ***
comm1       1.2787029  0.2412155   5.301 1.15e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> confint(poisson.glmm.final)
Computing profile confidence intervals ...
                    2.5 %      97.5 %
.sig01         0.849872181 1.207174294
(Intercept) -0.263377715 0.312917364
athletes       0.001805221 0.002993914
host1          1.666673881 2.778326727
comm1          0.807001606 1.771102114
```

All variables are considered statistically significant, which also supported by the fact that confidence interval of all coefficients does not contain 0. What is more, all coefficients are positive that means that number of medals increases with the number of athletes presented by the country and when the country is hosting the Olympics and if it is a current or former communist state.

In GLM Project I noticed that there is an excess of zeros in the data as lot of countries has not won a single medal (27.5% of total training data and similar share in test data).

Excess of zeroes is caused by the fact that the number of medals is finite, and some countries end up with 0 medals. For this case, zero-inflated models might be more appropriate. I will build the zero inflated model but with random intercept for each country.

As for Poisson GLMM, I first start with the preliminary zero-inflated GLMM with random-effects component for country that includes all predictors. To build the model I used R package “glmmTMB”. The notation for the model is the same as in “lme4” package. The basic glmmTMB fit that I used here is a zero-inflated Poisson model with a single zero-inflation parameter applying to all observations.
For more information on glmmTMB package, refer to the vignette below
https://cran.r-project.org/web/packages/glmmTMB/vignettes/glmmTMB.pdf)

```
> summary(zeroinfl.glmm.preliminary)
 Family: poisson  ( log )
Formula:          tot ~ athletes + gdp + pop + host + oneparty + muslim + soviet +      comm + altitude + (1 | country)
Data: oldat_test

     AIC      BIC   logLik deviance df.resid
   560.9    590.2   -269.4    538.9       95

Random effects:

Conditional model:
 Groups  Name        Variance Std.Dev.
 country (Intercept) 0.5411   0.7356
Number of obs: 106, groups:  country, 106

Conditional model:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.103e-02  2.071e-01   0.150  0.88092
athletes     9.073e-03  1.385e-03   6.549  5.8e-11 ***
gdp         -2.292e-02  5.328e-02  -0.430  0.66708
pop          1.491e-01  6.696e-01   0.223  0.82380
host1        3.781e-01  3.546e-01   1.066  0.28624
oneparty1   -8.285e-01  9.940e-01  -0.834  0.40456
muslim1      1.027e-02  2.767e-01   0.037  0.97040
soviet1     -3.002e-02  3.569e-01  -0.084  0.93297
comm1        8.656e-01  3.046e-01   2.842  0.00449 **
altitude     4.524e-05  1.575e-04   0.287  0.77387
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The summary shows that only “athletes” and “comm” are statistically significant predictors as p-values are below 0.05 for these variables. Accordingly, I will build final zero-inflated Poisson model including only these variables.

```
> summary(zeroinfl.glmm.final)
 Family: poisson  ( log )
Formula:          tot ~ athletes + comm + (1 | country)
Data: oldat_test

     AIC      BIC   logLik deviance df.resid
   549.4    560.1   -270.7    541.4      102

Random effects:

Conditional model:
 Groups  Name        Variance Std.Dev.
 country (Intercept) 0.5574   0.7466
Number of obs: 106, groups:  country, 106

Conditional model:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0739404  0.1646909   0.449 0.653457
athletes    0.0096227  0.0007519  12.797  < 2e-16 ***
comm1       0.7459697  0.2059574   3.622 0.000292 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> confint(zeroinfl.glmm.final)
                                            2.5 %      97.5 %
cond.(Intercept)                     -0.248847707 0.39672859
cond.athletes                         0.008148952 0.01109648
cond.comm1                            0.342300601 1.14963875
country.cond.Std.Dev.(Intercept)      0.591127432 0.94295094
```

Finally, for the sake of comparison I will also build basic type of linear mixed model based. I will start with the model that includes all predictors.

```
> summary(lmm.preliminary)
Linear mixed model fit by REML ['lmerMod']
Formula: tot ~ athletes + gdp + pop + host + oneparty + muslim + soviet +      comm + altitude + (1 | country)
   Data: oldat_training
Control: lmerControl(check.nobs.vs.nRE = "ignore")

REML criterion at convergence: 2504.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.2651 -0.2979 -0.0045  0.2911  5.3348

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 29.98    5.476
 Residual             10.66    3.265
Number of obs: 431, groups:  country, 108

Fixed effects:
              Estimate Std. Error t value
(Intercept) -2.3331260  0.9443771  -2.471
athletes     0.0828166  0.0051098  16.208
gdp          2.9044583  0.3115725   9.322
pop          7.9535989  3.6901507   2.155
host1        3.2553532  1.9424060   1.676
oneparty1   11.3131077  3.8082595   2.971
muslim1      0.1441562  1.5109799   0.095
soviet1      2.7573033  2.3076804   1.195
comm1        1.9163577  1.8937585   1.012
altitude    -0.0004366  0.0009075  -0.481
```

The summary output above does not contain p-values, as these can be based on different approximations depending on the design of the study. I can get some indication of the significance of the model parameters by obtaining approximate confidence intervals as shown below.

```
> confint(lmm.preliminary)
Computing profile confidence intervals ...
                    2.5 %       97.5 %
.sig01       4.5044782965  6.097012908
.sigma       3.0228660485  3.532492163
(Intercept) -4.1351980460 -0.554495409
athletes     0.0732056814  0.093212540
gdp          0.0023204303  0.003544181
pop          0.0008650038  0.014883743
host1       -0.5798951051  6.891870068
oneparty1    4.1130468541 18.530783872
muslim1     -2.7136092726  3.006383898
soviet1     -1.6053438628  7.130587626
comm1       -1.6830105263  5.490416182
altitude    -0.0021497595  0.001286426
```

The output shows that “athletes”, “gdp”, “pop” and “oneparty” are statistically significant predictors in the model as all other intervals contain 0. Also, it should be mentioned that all significant predictors have coefficient above 0 that means that the correlation is positive with number of medals.

Now in the final model all variables are considered statistically significant looking at the confidence intervals output below.

```
> summary(lmm.final)
Linear mixed model fit by REML ['lmerMod']
Formula: tot ~ athletes + gdp + pop + oneparty + (1 | country)
   Data: oldat_training

REML criterion at convergence: 2514

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.1494 -0.2947 -0.0103  0.2889  5.4315

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 30.42    5.516
 Residual             10.78    3.283
Number of obs: 431, groups:  country, 108

Fixed effects:
            Estimate Std. Error t value
(Intercept) -1.611048   0.661699  -2.435
athletes     0.088535   0.004267  20.749
gdp          2.916925   0.308295   9.461
pop          8.111151   3.703182   2.190
oneparty1   11.423188   3.655846   3.125
```

```
Computing profile confidence intervals
                    2.5 %       97.5 %
.sig01         4.6659454805  6.308936350
.sigma         3.0389600747  3.551607536
(Intercept) -2.9047814043 -0.320371596
athletes       0.0802097941  0.097092385
gdp            0.0023182441  0.003531013
pop            0.0008957827  0.015323835
oneparty1      4.3301189849 18.515988995
```

Comparison of the models. Conclusions and discussion

I have built the GLMMs and now I will compare its performance on test data and also compare results with GLMs that I selected in GLM Project.

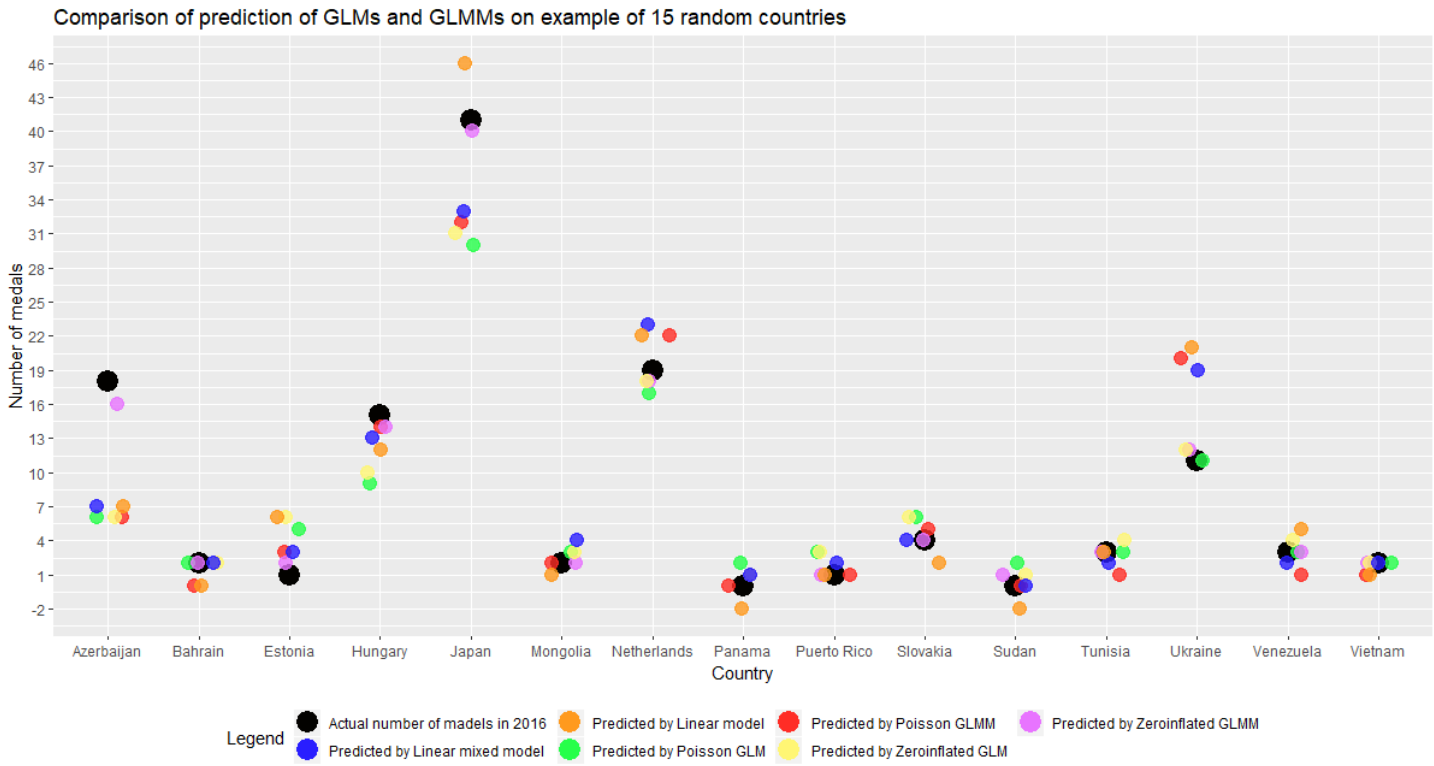
I will assess goodness of fit based on the 2 criteria:

1. **Akaike information criterion (AIC)**. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.
2. **Root-mean-square error (RMSE)** is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSE represents the square root of the mean square error, which in turn is the average squared difference between the fitted values and the actual value. The lower RMSE the better the model is.

	General Linear Mixed Models selected for final comparison			General Linear Models from GLM Project*		
Criteria of good fit	**Linear mixed model**	**Poisson GLMM**	**Zero inflated GLMM**	**Linear model**	**Poisson GLM**	**Zero inflated GLM**
AIC	**2,528**	**1,904**	**549**	**2,793**	**2,615**	**2,249**
RSME	**5.392**	**4.773**	**1.095**	**6.901**	**6.339**	**6.280**

* RMSE is slightly different from the ones calculated in the GLM Project due to the fact that in this assignment predicted number of medals was rounder to integers for consistent comparison with GLMMs. There is no impact on the conclusions made in the GLM Project.

Also, it worth looking at the predicted against actual plot to visualize predictive quality of each model. Below is the plot of the actual number of medals won in 2016 Olympic Games and predicted number by each of the abovementioned models for 15 randomly selected countries. I did not plot all countries as the plot would not be interpretable in this case.



Looking at the above criteria and the plot I can make the following conclusion:

1. As it was expected, GLMMs are better at predicting number of medals won than GLMs. RMSE is for GLMMs is higher for all GLMMs compared to GLMs, AIC is much lower for Poisson GLMM and Zero-inflated GLMM compared to GLM models. While AIC of LMM is only marginally higher than AIC of the best GLM (zero-inflated), its RMSE is lower. Better performance of GLMMs is achieved by taking country as a grouping factor. By doing this, these models take into account that multiple observations for each country are correlated and thus not normally distributed, while GLMs assume that the data is normally distributed and fail to account of the fact that it is not the case.
2. Among GLMMs, the best performance is achieved by Zero-inflated GLMM, its RMSE and AIC are much smaller than those of all other models. I can also see on the plot that although sometimes other models predict better for some countries, zero inflated GLMM is much closer to actual results on general (i.e. if looking across all countries). It is especially evident for the countries that with large number of medals. This supports our assumption that as there is only limited number of medals combined with large number of participating countries, a lot of countries do not win any medals. So, the data contains excessive umber of zeroes that distort the prediction results if not accounted for. Zero inflated GLM also performed best among GLMs in the GLM Project.
3. Linear mixed model is the worst predictor of number of medals among GLLMs. It is only marginally better than Poisson GLM by AIC and RMSE and even worse Zero inflated GLM if to compare their AIC. Linear model was also the worst among GLMs in the Assignment. This is expected, as generally linear (mixed) models do not predict counts as good as, for example, GLM based on Poisson distribution. Instead they are used for prediction of continuous data.

Considering the results above, the best model appears to be Zero inflated GLMM and therefore I will select this model as the final.

And as a final word I would like to discuss what can be done in terms of enhancing of goodness of fit of the model.

1. There is also a potential for search for the models with even better performance by introduction of the random slope (from the variables selected in final model) correlated or non-correlated with random intercept (country). However, this will likely to make the model less interpretable and also may result in the model overfitting.

Therefore, I included only random intercept by country, which is evident considering correlated observations in the data.

2. Regarding the data, it is also possible to transform the dataset to include gdp per capita instead of gdp and pop as variables.
3. Also, it may be worthy to include share of gdp of sport sector as predictor of medals won instead of total gdp and check how it impacts the model accuracy.

Thank you for your attention!