GLM Project – Prediction of number of medals won in 2016 Olympics

Preparer: Alexey Pankratov

Introduction

The aim of the project is to build general linear model that predicts the number of medals (gold and/tot all) 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 millions of US dollars during year YY,
- ***pop***: the country's population in thousands in year YY,
- ***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.

There is no restriction of the general linear models to use. I will focus on (1) Poisson regression model, (2) Zero inflated model and (3) Basic Linear Model, which is also a special case of GLM 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.

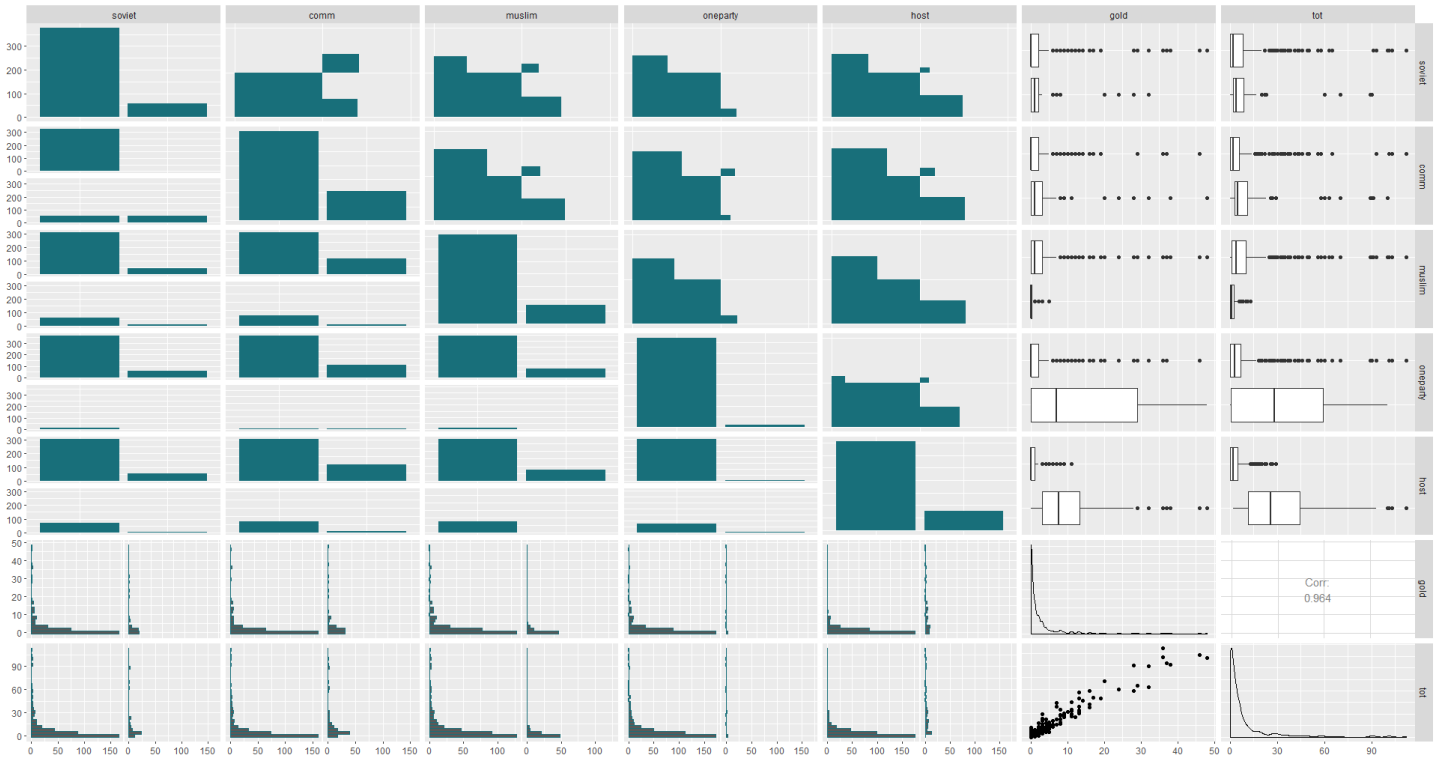
As it can be seen from the plot, there is a very strong positive correlation between total number of medals won and the number of gold medals won (0.964). Also you can notice that other variable correlate with total and gold medal won in a similar way, therefore I will focus only on prediction of total 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 if the difference significant is not clear at this stage. 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 general linear models

First, I will try to build preliminary Poisson regression model that includes all variables. 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.

```
Call:
glm(formula = tot ~ athletes + gdp + pop + host + oneparty +
    muslim + soviet + comm + altitude, family = poisson, data = oldat_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1744  -1.9032  -0.7772   0.7869   5.9514

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.647e-01  4.515e-02  16.936  < 2e-16 ***
athletes     4.364e-03  1.628e-04  26.802  < 2e-16 ***
gdp          4.694e-08  5.720e-09   8.206 2.28e-16 ***
pop         -7.697e-07  1.025e-07  -7.512 5.83e-14 ***
host1        1.059e+00  6.535e-02  16.199  < 2e-16 ***
oneparty1    6.267e-01  1.118e-01   5.604 2.09e-08 ***
muslim1     -4.010e-01  9.161e-02  -4.377 1.20e-05 ***
soviet1     -1.272e-01  7.197e-02  -1.768   0.0771 .
comm1        8.246e-01  6.943e-02  11.875  < 2e-16 ***
altitude    -7.322e-05  4.175e-05  -1.754   0.0795 .
---
```

```
> drop1(poisson.glm.preliminary, test = "F")
Single term deletions

Model:
tot ~ athletes + gdp + pop + host + oneparty + muslim + soviet +
    comm + altitude
          Df Deviance    AIC  F value    Pr(>F)
<none>         1470.0 2612.0
athletes  1   2165.7 3305.7 199.2633 < 2.2e-16 ***
gdp       1   1534.6 2674.6  18.5010 2.113e-05 ***
pop       1   1521.9 2661.9  14.8596 0.0001339 ***
host      1   1729.2 2869.3  74.2528 < 2.2e-16 ***
oneparty  1   1497.8 2637.9   7.9814 0.0049507 **
muslim    1   1490.9 2630.9   5.9849 0.0148371 *
soviet    1   1473.1 2613.1   0.8895 0.3461504
comm      1   1600.6 2740.6  37.4133 2.185e-09 ***
altitude  1   1473.1 2613.1   0.9041 0.3422226
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

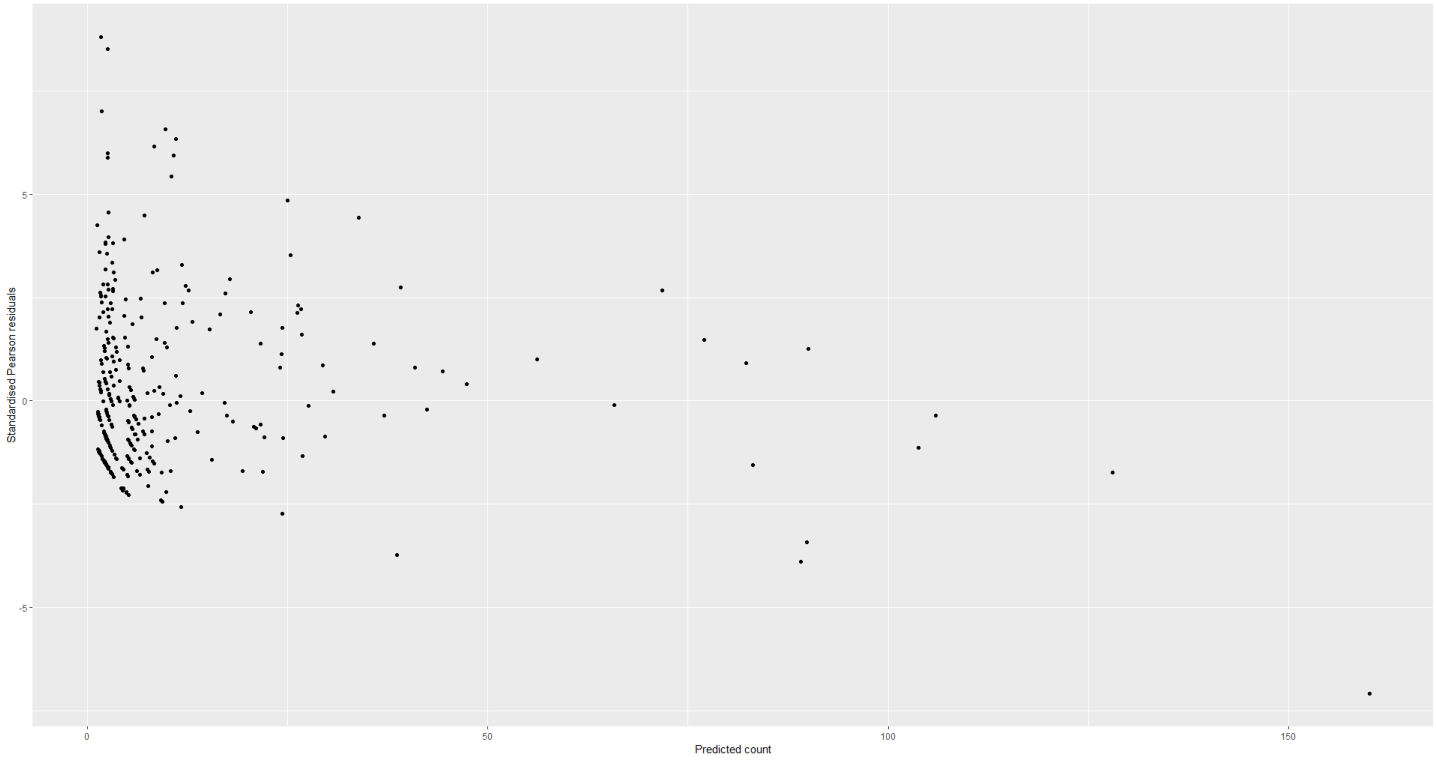
Looking at the summary of the Poisson regression model (Wald test), "soviet" and "altitude" coefficients in the above model has a p-value above 0.05. F-test provides the same results. It means that these variables are not statistically significant and therefore can be removed from the model.

Now let's make a new Poisson model without these variables:

```
Call:
glm(formula = tot ~ athletes + gdp + pop + host + oneparty +
    muslim + comm, family = poisson(), data = oldat_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1071  -1.8744  -0.7650   0.7641   5.6792

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.488e-01  3.873e-02  19.336  < 2e-16 ***
athletes     4.365e-03  1.627e-04  26.832  < 2e-16 ***
gdp          4.924e-08  5.600e-09   8.794  < 2e-16 ***
pop         -7.556e-07  1.013e-07  -7.458 8.75e-14 ***
host1        1.042e+00  6.332e-02  16.456  < 2e-16 ***
oneparty1    6.932e-01  1.036e-01   6.690 2.24e-11 ***
muslim1     -4.334e-01  8.985e-02  -4.824 1.41e-06 ***
comm1        7.520e-01  4.495e-02  16.727  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8036.3  on 430  degrees of freedom
Residual deviance: 1477.2  on 423  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 2615.2
```

Although all variables in this model are considered significant (t-value is below 0.05), looking at the residual deviance (*1477*), it is significantly higher that amount of chi squared on 423 degrees of freedom and 95% significance level (*472*) indicating a poor fit if the Poisson is the correct model for the response.

It is likely due to overdispersion. Below is the plot of dispersion of Poisson distribution. I can see in the plots that there are residuals for counts close to 0 that are above +8 of standard Pearson distribution.



Overdispersion is probably observed due to excess 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 using the same variables for as I used for Poisson.

Zero inflated model consists of two parts – Poisson and binomial. In zero-inflated models, it is possible to choose different predictors for the counts and for the zero-inflation. You might expect different variables to be driving win/lose vs. total number of medals. I will keep it simple and use the same covariate in both parts.

```
Call:
zeroinfl(formula = tot ~ gdp + athletes + pop + host + oneparty + muslim + comm, data = oldat_training)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-4.3549 -0.7488 -0.4062  0.4839  6.2918

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.077e+00         NA      NA       NA
gdp          4.884e-08         NA      NA       NA
athletes     4.085e-03         NA      NA       NA
pop         -6.565e-07         NA      NA       NA
host1        8.370e-01         NA      NA       NA
oneparty1    7.371e-01         NA      NA       NA
muslim1     -2.458e-01         NA      NA       NA
comm1        5.841e-01         NA      NA       NA

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.866e+00         NA      NA       NA
gdp         -3.939e-07         NA      NA       NA
athletes    -7.867e-02         NA      NA       NA
pop         -1.070e-05         NA      NA       NA
host1       -1.371e+01         NA      NA       NA
oneparty1   -1.329e-01         NA      NA       NA
muslim1      5.864e-02         NA      NA       NA
comm1       -7.475e-01         NA      NA       NA
```

*Note that NA is due to the fact that coefficient for parameter estimates pop and gdp are approximately zero because both population and gdp are significantly larger amounts then number of medals (hence you need to multiply population and gdp by a very small coefficient to get the number of medals). This can be solved by dividing variables pop and gdp by 1,000,000 (i.e. presenting population and GDP in million units). In this case coefficients will change with no impact on the characteristics of the model. Please refer Appendix 1 in the R code for comparison of both Zero inflated models.

I can check if zero inflated model better fits the data then Poisson model by performing a Vuong test of the two models. The Vuong non-nested test is based on a comparison of the predicted probabilities of two models that do not nest. Examples include comparisons of zero-inflated count models with their non-zero-inflated analogues (e.g., zero-inflated Poisson versus ordinary Poisson, or zero-inflated negative-binomial versus ordinary negative-binomial). A large, positive test statistic provides evidence of the superiority of model 1 over model 2, while a large, negative test statistic is evidence of the superiority of model 2 over model 1. Under the null that the models are indistinguishable, the test statistic is asymptotically distributed standard normal.

```
> vuong(zeroinfl.model, poisson.glm.final)
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                   7.113412 model1 > model2 5.6599e-13
AIC-corrected         6.815677 model1 > model2 4.6910e-12
BIC-corrected         6.210366 model1 > model2 2.6431e-10
```

The Vuong test compares the zero-inflated model with an ordinary Poisson regression model. Here I can see that our test statistic is significant, indicating that the zero-inflated model (model 1) is superior to the standard Poisson model (model 2).

Finally, for the sake of comparison I will also build basic type of linear model based on normal distribution with identity log-link (i.e. the same as lm() function in R or glm() functions using gaussian distribution with identity link function). I will start with the model that includes all predictors.

```
Call:
glm(formula = tot ~ athletes + gdp + pop + host + oneparty +
    muslim + soviet + comm + altitude, family = gaussian, data = oldat_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-21.617   -2.652    0.264    2.248   45.627

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.754e+00  5.256e-01  -5.240 2.54e-07 ***
athletes     9.133e-02  4.625e-03  19.745  < 2e-16 ***
gdp          3.496e-06  2.682e-07  13.036  < 2e-16 ***
pop          5.733e-06  2.031e-06   2.823  0.00499 **
host1        2.695e-01  1.250e+00   0.216  0.82946
oneparty1    1.145e+01  2.048e+00   5.590 4.10e-08 ***
muslim1      1.923e-01  8.153e-01   0.236  0.81369
soviet1      2.979e+00  1.239e+00   2.404  0.01665 *
comm1        1.437e+00  1.039e+00   1.384  0.16722
altitude    -2.722e-04  4.906e-04  -0.555  0.57934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 37.57018)

    Null deviance: 122161  on 430  degrees of freedom
Residual deviance:  15817  on 421  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 2797.9

Number of Fisher Scoring iterations: 2
```

As you can see from the summary output, t-values for coefficients of muslim, host, comm and altitude variables are significantly higher than 0.05 therefore they are not statistically significant so I will remove it from the final version of the basic linear model.

```
Call:
glm(formula = tot ~ athletes + pop + gdp + oneparty + soviet,
    family = gaussian, data = oldat_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-21.690   -2.497    0.447    2.228   45.402

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.708e+00  3.969e-01  -6.822 3.10e-11 ***
athletes     9.273e-02  3.364e-03  27.566  < 2e-16 ***
pop          5.875e-06  2.026e-06   2.900  0.00392 **
gdp          3.449e-06  2.652e-07  13.009  < 2e-16 ***
oneparty1    1.197e+01  1.955e+00   6.122 2.10e-09 ***
soviet1      4.211e+00  8.875e-01   4.745 2.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 37.45141)

    Null deviance: 122161  on 430  degrees of freedom
Residual deviance:  15917  on 425  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 2792.6

Number of Fisher Scoring iterations: 2
```

Now all variables are considered statistically significant.

Comparison of the models. Conclusions and discussion

Although I have already compared Poisson and Zero Inflated model, let's make a final comparison of all 3 models that were analysed above (ignoring preliminary versions that included all variables).

I will assess goodness of fit based on the 3 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. **Deviance** s a goodness-of-fit statistic for a statistical model. It represents twice the difference in maximized loglikelihoods evaluated at the saturated and current models. It is a generalization of the idea of using the sum of squares of residuals in ordinary least squares to cases where model-fitting is achieved by maximum likelihood. The lower residual deviance the better the model. It should be noted that there is no deviance calculation for Zero Inflated model as it is not a standard GLM but a mixed model that consists of two parts.
3. **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 RMSD 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.

Criteria of good fit	Linear model (Gaussian GLM with identity link function)	Poisson model	Zero inflated model
AIC	2,793	2,615	2,249
Deviance	15,917	1,477	NULL
Degrees of freedom	425	423	415
χ^2	474	472	463
RSME	6.886	6.349	6.239

*Degrees of freedom and chi squared are presented to show that they are approximately the same for each entity and that Deviance is comparable

Looking at the above criteria, the model that fits the data best is Zero inflation model as it shows the best AIC coefficient, and the lowest RSME.

Also as already mentioned above, residual deviance of normal linear model and Poisson model is higher than the 95th percentile of the chi squared on residual degrees of freedom, which indicates lack of fit. However, it is much lower than residual deviance of linear model that indicates a better fit compared to linear model.

However, looking at the RSME, it is not significantly lower in zero-inflated model than in Poisson model. Considering that zero-inflated (Poisson) model is a mixture of two distributions – Poisson and binomial - it is much less interpretable compared to general Poisson model. Having this in mind it may be reasonable to opt for Poisson model that although has slightly worse goodness of fit characteristics, but it is much simpler and more interpretable model.

It should be noted, that the number of athletes representing the country variable dominates each model, i.e. it serves as the main predictor of the number of medals won. Removing this variable significantly decreases predictive quality of every model, while removing other variables and keeping only 'athletes' also decreases prediction accuracy comparing with the final models above but not as much as if removing 'athletes'. The table below shows more than twice increase in RMSE in Poisson and Zero inflated model with the number of athletes predictor excluded making linear model the best model in this case (although still much worse than final models selected above). Same case is for AIC and deviance that increase significantly. At the same time, for the models that include only athletes as a predictor, all criteria of good fit are also significantly worse that in the selected final models above but not as bad as when athletes are removed.

	All variables are the same as in the final models above except for 'athletes' that are excluded			Only 'athletes' variable is included in the models		
Criteria of good fit	**Linear model**	**Poisson model**	**Zero inflated model**	**Linear model**	**Poisson model**	**Zero inflated model**
AIC	**3,233**	**3,314**	**2,946**	**3,004**	**3,144**	**2,595**
Deviance	**44,376**	**2,178**	**NULL**	**26,104**	**2,018**	**NULL**
Degrees of freedom	**426**	**424**	**417**	**430**	**430**	**428**
χ^2	**475**	**473**	**466**	**479**	**479**	**477**
RSME	**10.331**	**15.468**	**14.769**	**9.331**	**8.524**	**8.213**

Nevertheless, all models selected for final comparison (first table in this section) show better criteria of good fit therefore it is reasonable to keep both athletes and other variables that were kept in these models.

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. The zero-inflated Poisson model mixes two zero generating processes. The first process generates zeros. The second process is governed by a Poisson distribution that generates counts, some of which may be zero. Accordingly, in zero-inflated model it is possible to choose different predictors for the counts and for the zero-inflation. The model use in this assignment assumes the same variables both for the counts and for the zero-inflation. However, the model can be fine-tuned in the way that it the predicted number of medals won (distributed with Poisson distribution) could depend on one set of variables and prediction that 0 medal are won that follow binomial distribution is based on another set of variables.
2. For Poisson model it is also possible to predict not only the count of medals, but also the rate of medals won per athletes or per population as it has already been mentioned in the section “Selection and assessment of general linear models” above.
3. Regarding the data, it is also possible to transform the dataset to include gdp per capita instead of gdp and pop as variables.
4. 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!