Olympic Medal Prediction Rio 2016

Introduction

Olympic games are various sports competitions held every four years,in which huge number of athletes are participate from all over the world. When Olympic Games approach, there are endless number of predictions about how many medals each country will win. We would like to know how much accurate models can be built to predict number of medals won by each country.

Data Description

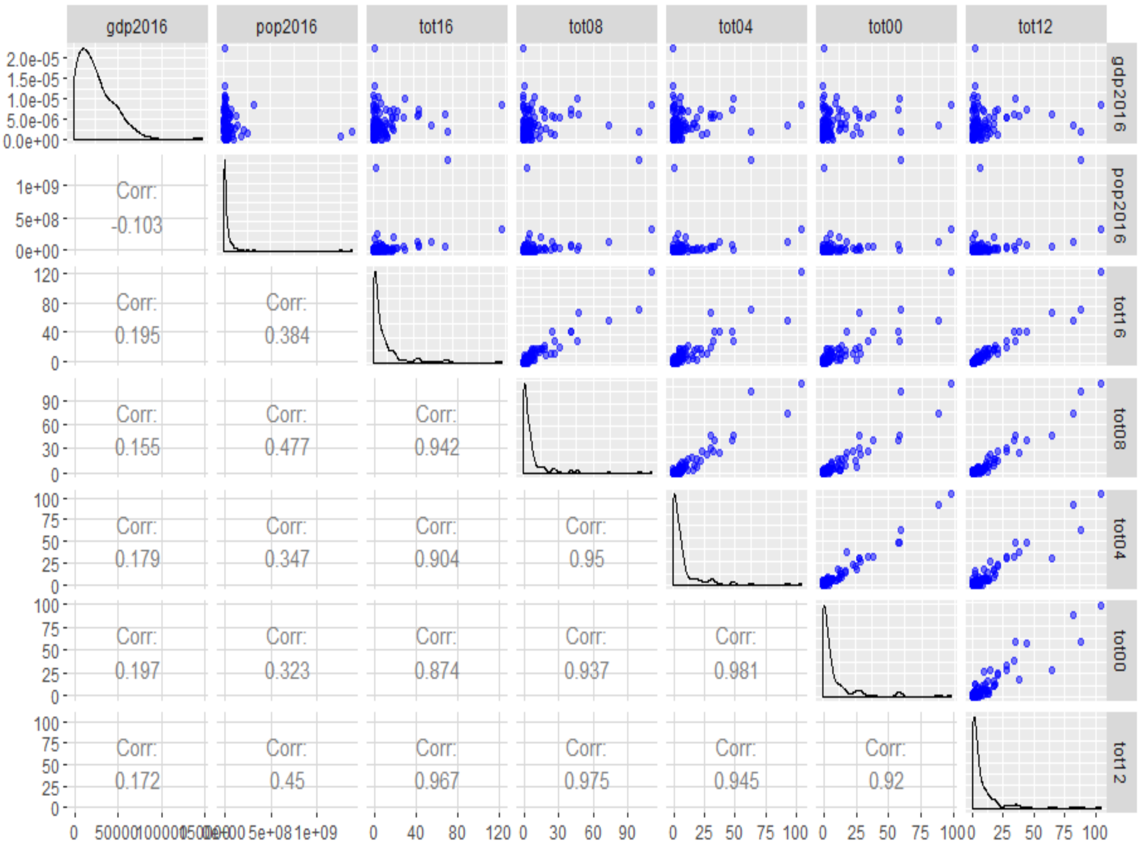
The data of the project is about 108 countries performance in Olympic Games since 2000.it comprises of two parts: total/gold medal won by each country and many political and economic factors that may affect our model .These factors include: Gross domestic product ,population, whether country was part of soviet union or former/current communist state,whether country is one-party state ,if Muslim country are majority of the country and whether the country is the host or was the host for the two past Olympic Games or will host the next one. Also the data includes the total number of medals in each Olympic games year and the number of athletes sent by each country.

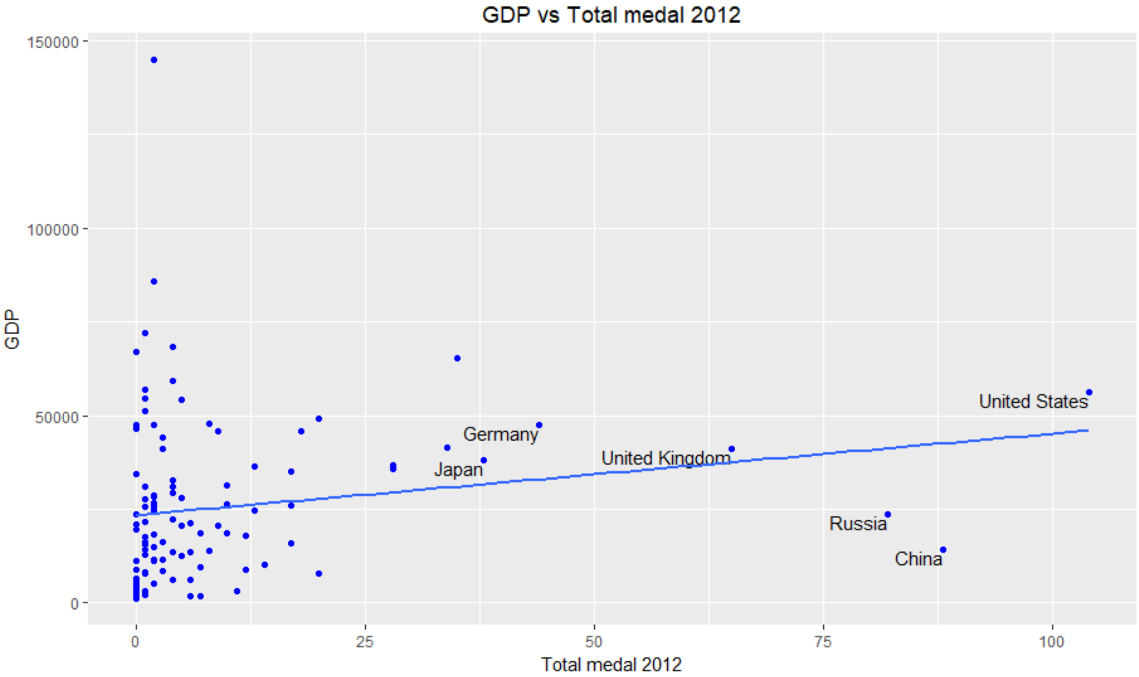
Pair plot is constructed. There is a positive relationship (high correlation) between the response variable total medal won in 2012 and all other variables. Also, all variables including the response are highly right skewed and that is an indicator of the existence of some outliers as shown in the figures below. There is a positive linear relationship between country's GDP per capita and population ,respectively and the total number of medals won in 2012.Countries with large population and GDP per capita tends to win more medals. But as shown in the figure some countries like Unites states,Russia and China won more medals than other countries compared to their populations and GDPs.

Methodology

Our goal is to predict total number of medals won by each country for 2016 Olympic Games in Brazil. The response variable of interest is the total medal won by each country. In our model, olympis games 2012 data is used as a training data and 2016 as a testing data. So for this purpose, total number of medal won in 2012 is used to fit the prediction model which is used to predict total medal won 2016. For the training data,the host city for Olympic Games 2012 was United Kingdom .The previous hosts were china and Greece and the future host was Brazil. For the testing data,the host city for Olympic Games 2016 was Brazil. The previous hosts were United Kingdom and China and the future host will be Japan. From the summary of the data,one observation (kosvo country data)with missing values is detected and deleted. Also, as the range of population and GDP per capita is large and there are outliers as mentioned before, a log transform is used. Five models with different

distributions for the response variable are built and compared to each other. These models are:


gdp2016
pop2016
tot16
tot08
tot04
tot00
tot12
2.0e-05
1.5e-05
1.0e-05
5.0e-06
0.0e+00
1e+09
5e+08
0e+00
120
80
40
0
90
60
30
0
100
75
50
25
0
Corr:
-0.103
Corr:
0.195
Corr:
0.384
Corr:
0.155
Corr:
0.477
Corr:
0.942
Corr:
0.179
Corr:
0.347
Corr:
0.904
Corr:
0.95
Corr:
0.197
Corr:
0.323
Corr:
0.874
Corr:
0.937
Corr:
0.981
Corr:
0.172
Corr:
0.45
Corr:
0.967
Corr:
0.975
Corr:
0.945
Corr:
0.92
0 50000100000150000
0e+00 5e+08 1e+09
0 40 80 120
0 30 60 90
0 25 50 75 100
0 25 50 75 100
0 25 50 75 100


GDP vs Total medal 2012
GDP
150000
100000
50000
0
United States
Germany
United Kingdom
Japan
Russia
China
0
25
50
75
100
Total medal 2012

- linear regression model :response variable is normally distributed.

- Generalized linear model :response variable is Poisson distribution

- Generalized linear model : response variable is Negative binomial distribution

-Zero inflation model : binomial distribution and Poisson or binomial distribution and negative binomial distribution.

To assess our models,the mean absolute error (MAE) ,root mean square error (RMSE) and mean forecast error are calculated. The mean absolute error (MAE)is the average of the absolute difference between the predicted value and the actual value across all observations. So it measures the average inaccuracy of the model estimations. The root mean square error(RMSE) is the square root of differences between predicted values and observed values.

Linear regression model

The linear regression model is the first model considered which the response variable is normally distributed.The model describes the number of medals for each country as a function of:

-total medal won at the previous Olympic Games 2008.

-total medal won at the Olympic Games 2004.

-total medal won at the olympic games 2000.

-GDP per capita -population

-Host,previous host and future host are indicator variables (1 or 0)

-Part of the former soviet union or not is an indicator variable (1 or 0)

-Former/current communist state is an indicator variable (1 or 0)

-One party state is an indicator variable(1 or 0)

-Mulsim majority country or not is an indicator variable(1 or 0)

A linear regression model is built with all the above variables as predictors.Due to the high correlation between all past olympic Games data since 2000, the variance inflation factor is used to detect the multicollinearity.Some of the variables with high variance inflation factor are execluded. Model best subset using many criterias like AIC and adjusted R^2 ,suggests that the total number of medal won at the previous Olympic Games 2008 is the best predictor for the total medal won by each country at 2012 Olympic games.From the model summary, 95% of the variability of the model is explained by the total medal won at previous Olympic games.

```
##Call:
## lm(formula = tot ~ totprev1, data = train)
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39939    0.42337   0.943    0.348
## totprev1     0.96061    0.02136  44.970   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.907 on 105 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9502
## F-statistic:  2022 on 1 and 105 DF,  p-value: < 2.2e-16
```

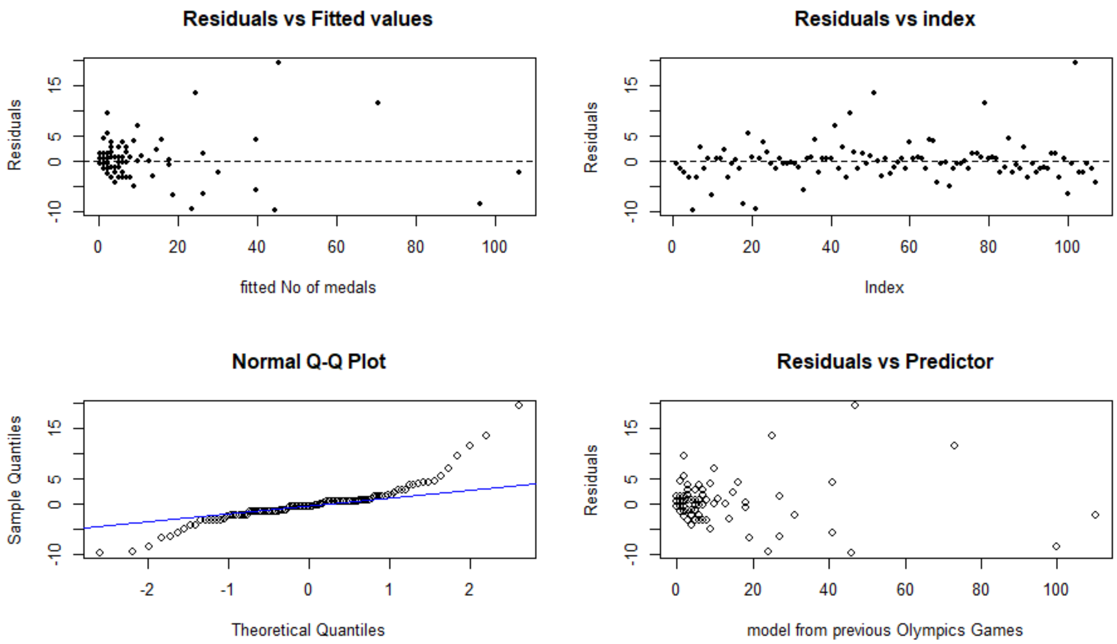
By checking the model assumptions From the below Figures:

1-Residuals vs fitted values : The mean of residuals is around zero and the variation of residuals is doubtfully constant as the existence of excessively outlying points.

2-Residual are indepen: Scatter plot of residuals with time sequence suggests that data observations collected over time and are not independent.

3-Normal Residuals: Normal probability plot of residuals shows the existence of outliers.

4-Residuals vs predictors :The mean of residuals remains close to zero. Also,the constant variability of residuals is doubtful cause of the excessively presence of outliers.



Generalized Linear Model (GLM)- Poisson Distribution

The response variable ,total number of medal won by each country,follows a Poisson distribution. when we have discrete response variable and predictors ,log transformation of linear model is equivalent to Poisson model. One assumption of Poisson Models is that the mean is equal to the variance. A Poisson model is fitted with the same predictors for the linear model. From the summary output,by looking at the standard error and the residual deviance is much greater than the degrees of freedom e.g. deviance/df,so there is a problem of over-dispersion. This means the response variable has variance not equal to the mean. Instead,the variance of the response variable is four times the mean. By testing the goodness-of-fit of the model with chi-square test based on the residual deviance and degrees of freedom,Poisson model doesn't fit the data (P< 0.05)

```
## Call:
## glm(formula = tot ~ log(GDPpercapita) + log(pop) + comm + muslim +
##     totprev1 + host, family = poisson, data = train, offset =
log(totmedals))
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -25.392331   1.849449 -13.730  < 2e-16 ***
## log(GDPpercapita)   0.615232   0.062147   9.900  < 2e-16 ***
## log(pop)            0.946223   0.084987  11.134  < 2e-16 ***
## comm1               0.564553   0.084439   6.686 2.29e-11 ***
## muslim1            -0.831361   0.162047  -5.130 2.89e-07 ***
## totprev1            0.014650   0.001635   8.960  < 2e-16 ***
## host1               1.022002   0.131921   7.747 9.40e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2004.2  on 106  degrees of freedom
## Residual deviance:  395.4  on 100  degrees of freedom
## AIC: 707.23
##
## Number of Fisher Scoring iterations: 5
```

```
1-pchisq(deviance(poiss.m.f),df.residual(poiss.m.f))
```

```
## [1] 0
```

Generalized Linear Model (GLM)- Quasi-Poisson

One way to deal with over-dispersion is to use quasi-Poisson model. Same model is fitted with same predictors but this time we use quasi-Poisson family. The outcome of the model account for overdispersion but the residual deviance has not changed. The dispersion parameter ,which was forced to 1 in the poisson model,is estimated at 3.8. The estimated coefficient of the quasi-Poisson fit and Poisson fit are the same. The only change is the predictors significance. The “F” test is used to determine the significance of the regression coefficient. Our final model is shown below with the most important predictors.

```
##
## Call:
## glm(formula = tot ~ log(GDPpercapita) + log(pop) + comm + muslim +
##     totprev1 + host, family = quasipoisson, data = train, offset =
log(totmedals))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7845  -1.5966  -0.8296   0.7481   4.8725
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -25.392331   3.703459  -6.856 5.92e-10 ***
## log(GDPpercapita)   0.615232   0.124448   4.944 3.09e-06 ***
## log(pop)            0.946223   0.170183   5.560 2.25e-07 ***
## comm1               0.564553   0.169085   3.339 0.001183 **
## muslim1            -0.831361   0.324493  -2.562 0.011898 *
## totprev1            0.014650   0.003274   4.474 2.03e-05 ***
## host1               1.022002   0.264167   3.869 0.000195 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.009871)
##
##     Null deviance: 2004.2  on 106  degrees of freedom
## Residual deviance:  395.4  on 100  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Generalized Linear Model (GLM)- Negative Binomial Distribution

Negative binomial regression is a generalization of Poisson regression without the assumption of equal mean and variance. Same model fitted with the same predictors except for the family it is negative binomial. The liklihood ratio test is used to determine the significance of the regression coefficient. Our final model is shown below with the most important predictors. By checking the result model,the coefficients have changed from the Quasi-Poisson model as they have different interpretations.

```
# Call:
## glm.nb(formula = tot ~ log(GDPpercapita) + log(pop) + comm +
##     muslim + totprev1, data = train, init.theta = 2.086303797,
##     link = log)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -14.640728   3.040217  -4.816 1.47e-06 ***
## log(GDPpercapita)   0.443068   0.101602   4.361 1.30e-05 ***
## log(pop)            0.764281   0.144111   5.303 1.14e-07 ***
## comm1               0.645963   0.204962   3.152  0.00162 **
## muslim1            -0.703418   0.267984  -2.625  0.00867 **
## totprev1            0.035998   0.005446   6.610 3.84e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.0863) family taken to
be 1)
##
##     Null deviance: 388.39  on 106  degrees of freedom
## Residual deviance: 116.47  on 101  degrees of freedom
## AIC: 547.53
##
## Number of Fisher Scoring iterations: 1
##              Theta:  2.086
##          Std. Err.:  0.452
##
##  2 x log-likelihood:  -533.530
```

```
1-pchisq(deviance(nb.m.f),df.residual(nb.m.f))
```

```
## [1] 0.1392955
```

By testing the goodness-of-fit of the model with chi-square test based on the residual deviance and degrees of freedom, the negative binomial model fits the data. (P> 0.05).

Zero inflated models-poisson distribution

The strict assumption of equal mean and variance of the Poisson model is often violated .High response values and many 0s lead to high variance. High variance can be handled using the negative binomial distribution. But when there are too many 0s than Poisson would predict,the zero-inflated model is the solution. Zero-inflated model fits two separate models at the same time. One is logistic (probit) model for counting zero's and the second one poisson or negative binomial for count data. As the variance is higher than the mean,zero inflated negative binomial model is used. Same predictors are used to fit the model. Wald test and likelihood ratio are used to compare different models with different predictors and below is the output of the final model. As the estimated theta parameter is significant that means the zero inflated Poisson model is not appropriate (overdispersion).

```
## Call:
## zeroinfl(formula = tot ~ log(GDPpercapita) + log(pop) + comm +
muslim +
##     totprev1, data = train, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.4166 -0.6621 -0.2753  0.4526  3.9381
##
## Count model coefficients (negbin with log link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -10.68878    3.19414  -3.346 0.000819 ***
## log(GDPpercapita)   0.30729    0.10837   2.835 0.004576 **
## log(pop)            0.59568    0.15003   3.971 7.17e-05 ***
## comm1               0.43224    0.19490   2.218 0.026574 *
## muslim1            -0.75048    0.25969  -2.890 0.003854 **
## totprev1            0.03388    0.00669   5.064 4.11e-07 ***
## Log(theta)          1.02317    0.24836   4.120 3.79e-05 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)          26.4128    15.9426   1.657   0.0976 .
## log(GDPpercapita)  -1.1349     0.6272  -1.809   0.0704 .
## log(pop)           -1.1850     0.7474  -1.585   0.1129
## comm1              -8.5004    49.6707  -0.171   0.8641
## muslim1            -1.3693     1.9976  -0.685   0.4930
## totprev1           -0.4381     0.3378  -1.297   0.1946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 2.782
## Number of iterations in BFGS optimization: 55
## Log-likelihood: -261.6 on 13 Df
```

Results and Discussion

The coefficients of the Poisson and Quasi-Poisson are the same and different from other models as they have different interpretations. Moreover ,the estimated standard error are also very similar except for the Poisson model(overdispersion) and linear model (different predictors). Comparing the logliklihood ,zero inflated model is the best one with slight difference from the negative binomial model. Poisson model is clearly the lowest one (overdispersion).Quasi-Poisson is not a likelihood based.The linear model is lower than zero inflated model. In terms of akaike information criterion (AIC),Poisson model has the highest value. Zero inflated model is the best one with slight difference from the negative binomial and the linear model is slightly more.AIC is calculated with likelihood function so it is not applied for Quasi-Poisson

```
##        Pois Qpois   Neg    lm  Zin
## logLik -347    NA -267 -297 -262
## Df        7     7    7    7   13

##       Pois Qpois    Neg     lm    Zin
## AIC 707.23    NA 547.53 599.27 549.16
```

Comparing number of zeros between the observed data and predicted data for different models,zero inflated model lead to the best result. Poisson model still the worst. The linear model are much better in modeling the zero counts.

```
## Train  Pois    NB  ZINB    lm
##    21     9    18    22    21

## Test Pois   NB ZINB   lm
##   25    9   18   23   21
```

The mean absolute error (MAE) for all models shown below. In terms of MAE and RMSE,the linear model is the best one. On average,the linear model prediction is different from the actual number of medal won by each country by 2.4. Quasi-Poisson model is the second best model in terms of MSE and RMSE. As the negative binomial variance is a quadratic function of the mean,that affect the model prediction especially with the large observed values. That explains the highest value of MAE and RMSE for negative binomial and zero-inflated model.

```
##      poiss.MAE Qpoiss.MAE   nb.MAE  zin.MAE   lm.MAE
## [1,]  4.338427   4.338427 16.60177 12.30066 2.414152

##      poiss.RMSE Qpoiss.RMSE  nb.RMSE zin.RMSE  lm.RMSE
## [1,]   7.116904    7.116904 93.97979 64.92177 3.870401
```

Below is the comparison of different models prediction for the top 10 countries at Olympic Games 2012. The actual medal won by each country ,the predicted medals and the difference between the actual and predicted medals. It shows that the linear model predicts Olympic Games reasonably good.

```
##            country Actual.medal2012 lm.Predicted.medal lm.diff
## 104  United States              104                106       2
## 18           China               88                 96       8
## 80          Russia               82                 71      11
## 103 United Kingdom               65                 46      19
## 36         Germany               44                 40       4
## 51           Japan               38                 24      14
## 5        Australia               35                 45      10
## 33          France               34                 40       6
## 48           Italy               28                 26       2
## 87     South Korea               28                 30       2
##     QPoiss.Predicted.medal Qpoiss.diff
## 104                    122          18
## 18                     128          40
## 80                      56          26
## 103                     65           0
## 36                      25          19
## 51                      20          18
## 5                       22          13
## 33                      22          12
## 48                      16          12
## 87                      16          12
```

From 963,the linear model predicts around 254 medals wrong. The table below shows the difference between the actual and predicted model in a decreasing order. United kingdom won 19 more medals than predicted by the linear model that exclude the host country impact. As the Quasi-Poisson model takes into account the host country impact,it predicts United Kingdom medals precisley. Other prediction error contribution come from he outliers Japan,Russia and Australia.

```
##            country Actual.medal2012 lm.Predicted.medal lm.diff
## 103 United Kingdom               65                 46      19
## 51           Japan               38                 24      14
## 80          Russia               82                 71      11
## 5        Australia               35                 45      10
## 45            Iran               12                  2      10
## 21            Cuba               14                 23       9
## 18           China               88                 96       8
## 41         Hungary               17                 10       7
## 10         Belarus               12                 19       7
## 33          France               34                 40       6
##     QPoiss.Predicted.medal Qpoiss.diff
## 103                     65           0
## 51                      20          18
## 80                      56          26
## 5                       22          13
## 45                       3           9
## 21                       7           7
## 18                     128          40
## 41                      10           7
## 10                       9           3
## 33                      22          12
```

Below we have prediction from different models for the total medals won by each country at Olympics Games 2016. As shown that the linear model has the best performance,we use it to predict the total medal won by each country 2016.The model predicts that United States ,China,Russia and United Kingdom will keep their positions in the top 5 countries. Brazil is predicted to win 17 medals using the linear model. But if we include the host impact and use the Quasi-Poisson model,it is predicted to win 34 medal. Although United States is predicted to win 4 less medals than previous Olympics Games, but it will still rank number one and win more medals than other countries. On average,its expected that the difference between linear model prediction and the true number of medals won by each country is 2.4.

```
##            country lm.Predicted.medal_2016 QPoiss.Pred.medal
## 104  United States                     100               113
## 18           China                      85               109
## 80          Russia                      79                64
## 103 United Kingdom                      63                31
## 36         Germany                      43                27
## 51           Japan                      37                25
## 5        Australia                      34                19
## 33          France                      33                20
## 48           Italy                      27                16
## 87     South Korea                      27                15
##     Zin.pred.medal
## 104            441
## 18             394
## 80             196
## 103             67
## 36              37
## 51              32
## 5               21
## 33              24
## 48              18
## 87              17
```

If we compare our model prediction with other prediction from [1] (call it model2). Same 5 countries to be in the top with same order. Also United States will win less medals compared tp 2012 but will still retain position number one in at the rank. As Japan will be the 2020 Olympic games host, model2 predicts that it will win more 8 medals. In contast to our model,which predicts that it will win one less medal. Both models predict United kingdom to win less medal by one or two. Taking the impact of the host country,model2 predicts Brazil to win 16 more medals than 2012.Same number of medals are predicted for Australia,France and Italy.

In order to improve the predictive model, we can include the total number of medals increasing over time. This can be done by adding another term to the linear model indicates the year of the Olympic Games(Time series)

References:

[1] Predicted 2016 medal table [ignoring Russian ban.
https://www.bbc.co.uk/news/magazine-36955132