Advanced Predictive Modeling

Assignment 1

2020-2021

Georgios Pediaditis

Introduction

The purpose of this report is the development of models for the prediction of the number of medals (gold/total or both) won by each country at Rio Olympics in 2016 based on results from previous Olympic Games. We will use the dataset rioolympics.csv.

Our first step is to explore the variables that are associated with the number of metals in 2012 Olympic Games.

Next we will develop our models based on the data up until 2012 and later on we will use these models in order to predict Olympic performance in the 2016 Olympic Games.

Data Loading

When we load the data from rioolympics.csv we notice that some missing values are read as Factors. So we read our data with na.strings="#N/A" and we convert the numeric values for soviet, comm, muslim, oneparty, and host to factors.

Data Handling

Next after we load the dataset we notice that it contains missing GDP values for Afghanistan, Cuba, and Syria Arab Republic. We noticed that Afghanistan and Syria Arab Republic are war torn countries and their missing data coincide with wars. Based on this observation we conclude that missing values aren't randomly missing so we decide the following.

For Afghanistan since we don't have any GDP data at 2000 but we do have data for 2004 up until 2016 we decide to apply Next Observation Carried Backward (NOCB) for 2000 missing value.

For Syria Arab Republic we have missing GDP data at 2016. We cant make any assumptions or extrapolation about these data so we decided to remove the particular observation from our dataset.

Finally, in the case of Cuba since we don't have any information about events that could cause the missing data we assume the missing value is randomly missing. Since missing data are at 2016 we decide to linearly extrapolate based on previous GDP data.

Exploratory analysis

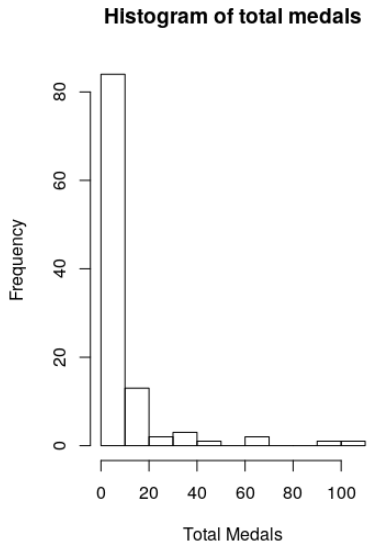
Since we want to perform exploratory analysis in 2012 data we start by choosing the columns for country, country.code, gdp12, pop12, soviet, comm, muslim, oneparty, gold12, tot12, totgold12, totmedals12, altitude, athletes12, and host.

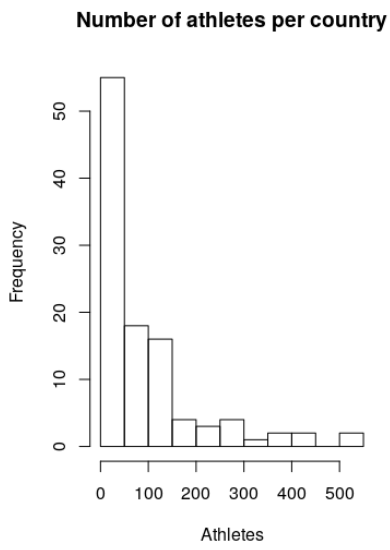
We summarize the columns for gdp12, pop12, gold12, tot12, altitude, and athletes12.

```
> summary(newoldat12[,c("gdp12","pop12","gold12","tot12","altitude","athletes12")])
     gdp12              pop12             gold12           tot12            altitude         athletes12
 Min.   :     800   Min.   :    105   Min.   : 0.000   Min.   :  0.000   Min.   : -28.00   Min.   :  0.0
 1st Qu.:   27580   1st Qu.:   4484   1st Qu.: 0.000   1st Qu.:  1.000   1st Qu.:  11.25   1st Qu.: 20.5
 Median :  174070   Median :  11045   Median : 1.000   Median :  3.000   Median :  76.00   Median : 47.0
 Mean   :  690776   Mean   :  56276   Mean   : 2.813   Mean   :  8.935   Mean   : 392.95   Mean   : 90.9
 3rd Qu.:  496902   3rd Qu.:  40080   3rd Qu.: 2.000   3rd Qu.:  9.000   3rd Qu.: 561.00   3rd Qu.:111.5
 Max.   :16155255   Max.   :1350695   Max.   :46.000   Max.   :103.000   Max.   :2850.00   Max.   :530.0
```

From the above we notice that there are potential outliers in almost all the summary columns. We can also see that our data have a lot of small values and view big ones.

As we can see from the following histograms we have a lot of countries with a small or none number of medals. We can also see that most countries participate in the Olympic games with a small number of athletes.





Before we proceed with variable comparison we notice that the number of gold with the total number of medals is correlated (0.964).

Fitting the model gold12~tot12 has a high R-squared 0.9311.

```
> summary(lm(gold12~tot12,newoldat12))

Call:
lm(formula = gold12 ~ tot12, data = newoldat12)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9487 -0.5719  0.2268  0.6100  7.1482

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.60996    0.19648  -3.104  0.00245 **
tot12        0.38312    0.01017  37.669  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.802 on 105 degrees of freedom
Multiple R-squared:  0.9311,    Adjusted R-squared:  0.9304
F-statistic:  1419 on 1 and 105 DF,  p-value: < 2.2e-16
```

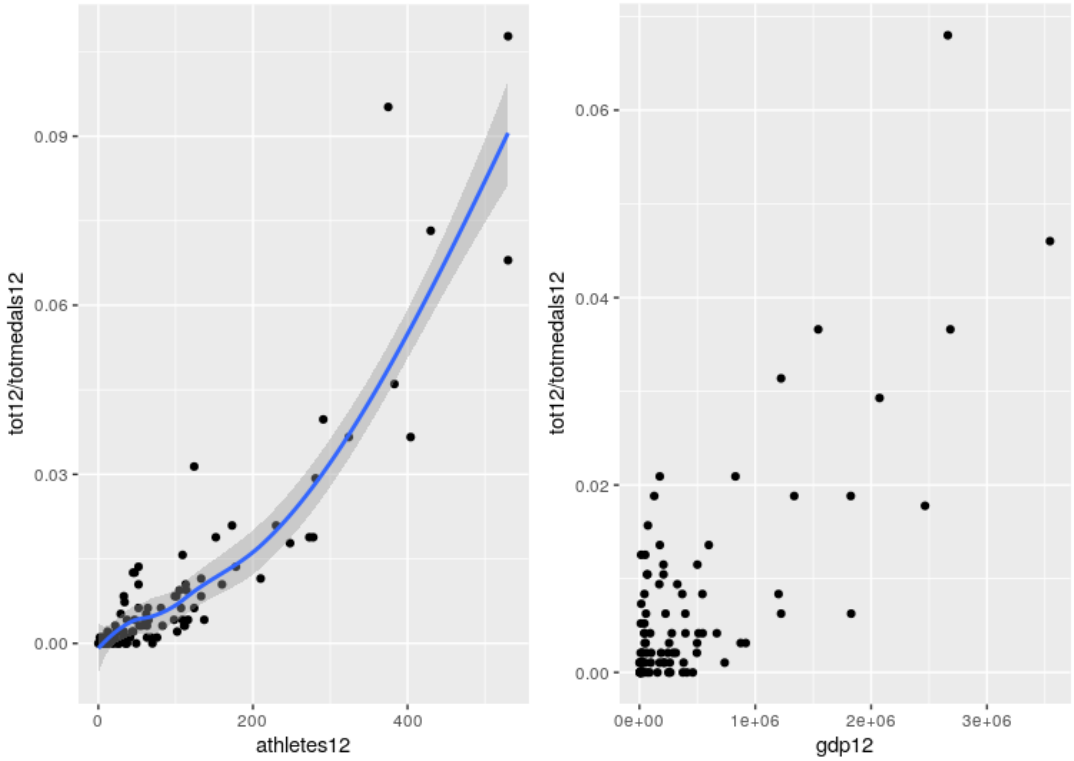
So we will decide to use the total number of medals for comparisons. If we can predict tot12 then we can use linear models to predict the gold12 with good accuracy.

Using cook distance we can see a lot of outliers. (104, 18, 80, 103, 51)

We decided not to remove them since although they pool the line upwards for big numbers for small numbers the don’t change our prediction that much.

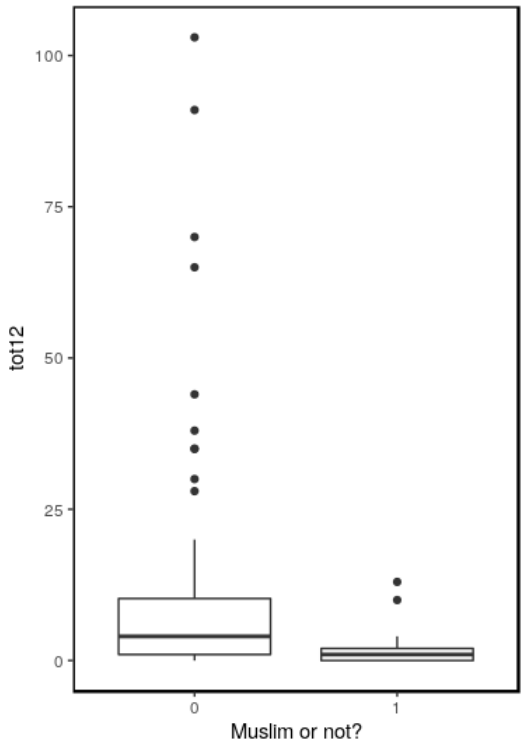
Particularly for big numbers this give us better predictions while for small numbers they don’t change our prediction enough to make a difference. 3.22 vs 3.1 for small numbers will be translated to 3 medals in both cases.

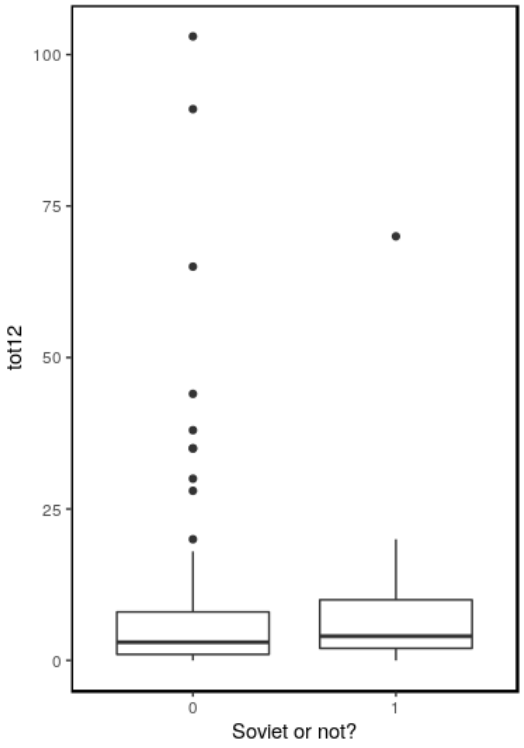
Plotting the graph rate of medals to the number of athletes we notice that the as the number of athletes increases the rate of medals increase quadratically. While plotting gdp with the rate of medals we see that the are positive corelated.

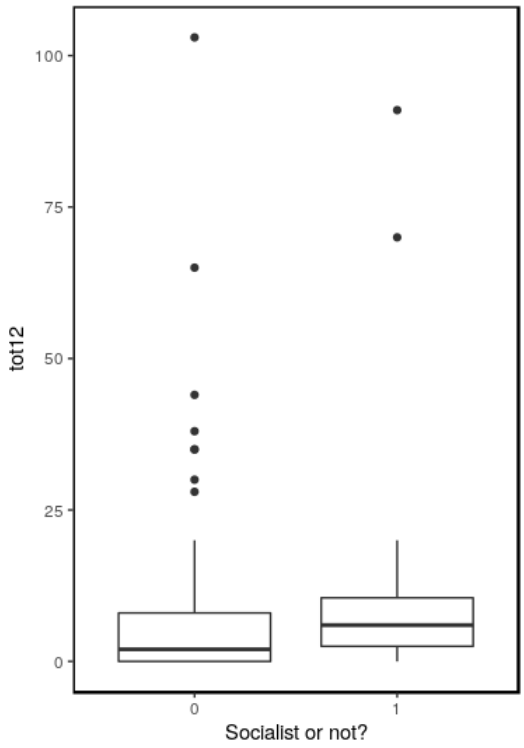


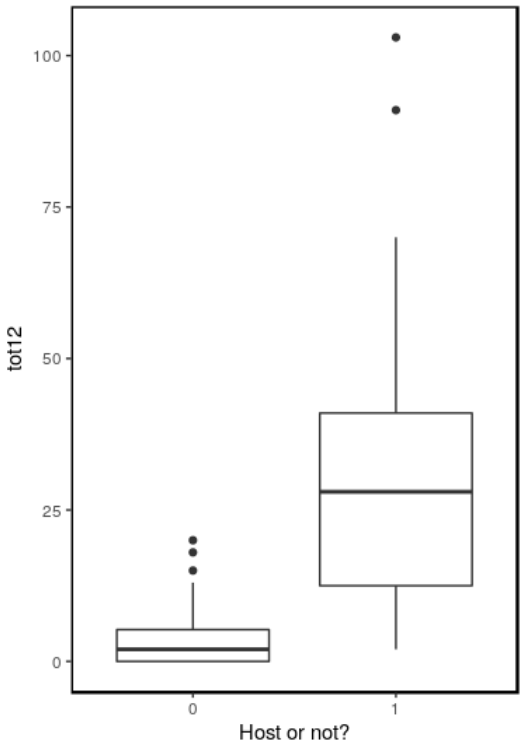
Finally from the following boxplots we can get some conclusions. Muslim countries get less medals than non Muslim countries. Non soviet countries although the get slightly less medals than ex soviet countries they have more upwards distributed medals numbers. The same applies for communist countries.

Finally, countries that host or hosted the Olympics in the past get more medals.


tot12
100
75
50
25
0
0
1
Muslim or not?


tot12
100
75
50
25
0
0
1
Soviet or not?


tot12
100
75
50
25
0
0
1
Socialist or not?


tot12
100
75
50
25
0
0
1
Host or not?

Models

In this part we will train our models with the data from 2000 up to 2012.

Binomial Distribution

Our first attempt is to try to fit a binomial distribution. Using all variables and testing with 3 different link functions (logit, probit, cloglog) show us that probit link function give us a better prediction.

	logit	probit	cloglog
Residual Deviance	1457.3	1360.7	1468.8

Trying to build a model with probit and only using significant variables (gdp, pop, comm, muslim, oneparty, athletes, and host) give us residual deviance 1231.5.

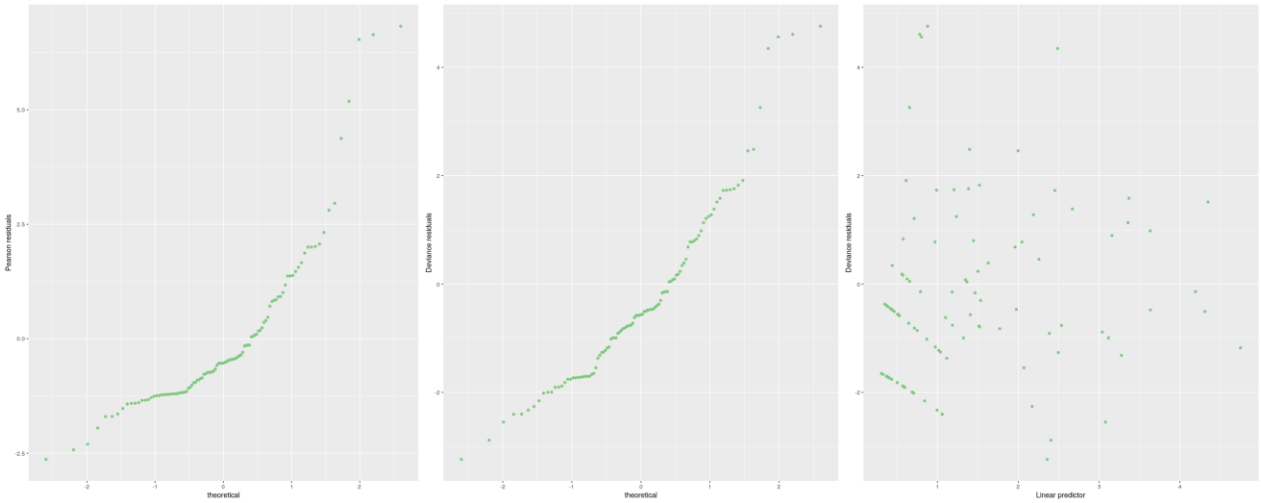
Testing our final model (tot,totmedals-tot)~gdp+pop+comm+muslim+oneparty+athletes+host give us residual deviance1361.4 while qchisq(df=428-7,p=0.95) give us 469.83. Our model isn't a good fit so our next step would be to try a different response distribution.

Poison regression

We try to apply poison regression. Fitting all the variables to our model will give us residual deviance 1459.4 which is not better than binomial distribution. Removing non significant columns (soviet, and altitude) give us 1466.0 residual deviance. The value is far away from qchisq(df=428-7,p=0.95) which is equal to 469.83.

We check our data for outliers and overdispersion.

Outliers



Based on the figures above we don't see any outliers or nonlinearity.

Overdispersion

Poisson regression assumes that E[Y]=Var[Y]. We can check whether there is overdispersion with dispersiontest.

```
> dispersiontest(mod_poison2,trafo=1)

        Overdispersion test

data:  mod_poison2
z = 7.6954, p-value = 7.054e-15
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
2.590096
```

The result above shows us that we have overdispersion.

We can either try the quasi-Poisson model or negative binomial model. Since variance is larger than the E[Y] (alpha>0) we try negative binomial model.

Negative binomial model

Applying negative binomial model to our data give us residual deviance 510.13 with a lot of non significant covariates. Important covariates are if the country is a communist country or Muslim, the altitude, number of athletes and if the country hosted in the past or present Olympic games.

Keeping these covariates give us residual deviance 509.34. Checking our model with qchisq(df=107-2,p=0.95) give us 471.95. That means that its an accepted model.

Prediction

Now we will use our models trained using the data from 2000 up until 2012 to predict the results for 2016.

Binomial model

For the binomial model using the model (gdp, pop, comm, muslim, openparty, athletes, and host) we created with the data from 2000 up until 2012 and in combination with the results from 2016 we get RMSE 20.27317

Poisson model

For the Poisson model using the model (gdp, pop, comm, muslim, oneparty, athletes, and host) we created with the data from 2000 up until 2012 and in combination with the results from 2016 we get RMSE 17.83276

Negative binomial model

For the negative binomial model using the model (comm, muslim, altitude, athletes, and host) we created with the data from 2000 up until 2012 and in combination with the results from 2016 we get RMSE 17.7556.

Conclusions

From the above we come to the conclusion that Negative binomial model give us better prediction (17.7556) compared to Poisson model and compared to Binomial model.

Our model using negative binomial model with the variables comm, muslim, altitude, athletes, and host give us a good enough fit and can predict the total number of medals given on our test set 2016. Furthermore from the number of total medals we can accurately predict using linear regression the number of golds (R-squared 0.9311)

Using Pearson residuals and comparing the expected values vs the actual values for 2016 we find out that our model mostly fails at Australia. 26 Medals in 2016 vs 343 expected medals. It also fails to South Korea, Kenya and Jamaica but to a lesser degree compared to Australia.

Improvements we could make at our model include checking why we got so big failures in countries like Australia etc. We should also checking for interactions and maybe try different a different model like quasi-Poisson model to see if it can give us a lower RMSE value.

Finally during the initial exploratory analysis we noticed in histograms that our data probably have excess zeros. So it would be a good idea to try zero inflated and hurdle models and see if they can provide us with a lower RMSE.