Math 651 Final Project

Mary Peng, Hillary Dunn, and Max Kearns

December 17, 2018

1 Abstract

2 Introduction—needs editing

Every four summers, the Olympic Games become the center of the world's attention, as elite athletes seek honor for both themselves and for their countries. Many countries associate tremendous national pride with their medal counts, since a nation's athletic competence also projects its soft power. For this last reason, some governments invest generously in their sports programs, in hopes of an elevated medal count in the next Olympics.

Given the fierce competition and high profile nature of medal counts, one may wonder what factors influence the number of medals that a country wins at the summer Olympics. Certainly countries with the largest economies and populations, such as the United States and China, commonly dominate the top of the billboard. However, Azerbaijan, which ranks 91st in population and 72nd in total GDP, also ranked in the top 20 countries by total medal count in the 2016 Summer Olympics.

Our team will develop a multiple linear regression model using various predictors, such as GDP, population size, size of athletic investment, and geographic location, to predict medal count. We will build the model on countries' total medal count for the 2004, 2008, and 2012 Olympics, and then project the model onto the 2016 Olympics to understand the accuracy of the model's predictions.

3 Methods and Materials

3.1 Data

Frame this as if our only data is population, per capita GDP, host, and soviet/communist

3.1.1 EDA–choose what we want to keep, and then format as figures in Appendix

Summary of Medal Count:

Distribution of Medal Count:

Distribution of Medal Count in 2008:

GDP Summary:

Population Summary:

Total Count of Communist / comm_soviet Countries:

Count of host, hosted within 8 years prior, or will be hosting within 8 years:

Histograms of Population:

Boxplot of log(GDP):

Boxplot of log(Population):

Scatterplot Matrix of Variables:

Correlation Matrix of Predictor Variables and Response Variable:

3.2 Model Building and Selection

3.2.1 Linear Model

3.2.2 Generalized Linear Model

The first choice for a glm is typically a Poisson model, which may provide a reasonable model for these data. This generalized model was created with the same variables as the linear model; population, GDP per capita, the host dummy variable, and the communist/soviet dummy variable. Like the linear model, the population and GDP per capita were log-transformed, but in this model it is unnecessary to transform the medal count. The full poisson model showed highly significant parameters, but the dispersion of this model was much greater than 1 (disp = 6.621), so a negative binomial model would likely provide a better fit for these data.

A negative binomial regression model assumes that dispersion is greater than 1, which is consistent with these data. Therefore, it allows for more accurate tests of the parameters. The best model again proved to be the full model, after a comparison of AIC between all subsets of variables (Table 2).

4 Results

5 Discussion and Conclusions

6 Bibliography

7 Appendix A

Table 1: Model Selection Diagnostics for a Negative Binomial Model

Pop	GDP/C	Host	Soviet	Parameters	Cp.nb	AIC.nb
1	0	0	0	2	198.46	2497.82
0	0	1	0	2	215.98	2633.68
0	1	0	0	2	336.41	2658.69
0	0	0	1	2	370.52	2695.53
1	0	1	0	3	108.70	2474.29
1	1	0	0	3	119.18	2428.33
1	0	0	1	3	180.88	2489.15
0	1	1	0	3	199.48	2601.80
0	0	1	1	3	206.80	2626.84
0	1	0	1	3	318.21	2614.08
1	1	0	1	4	58.39	2338.79
1	1	1	0	4	60.95	2417.52
1	0	1	1	4	87.77	2457.19
0	1	1	1	4	178.38	2556.83
1	1	1	1	5	5.00	2319.23

Table 2: Negative Binomial Model Coefficients

	Estimate	Std. Error	z value	p-value
Intercept	-10.1657683	0.6468624	-15.715504	0.00e+00
log(population)	0.4984760	0.0288138	17.299910	0.00e+00
log(gdp/capita)	0.4023884	0.0317426	12.676614	0.00e+00
Host	0.6926724	0.1593195	4.347693	1.38e-05
Comm/Soviet	1.0337603	0.0984874	10.496376	0.00e+00

8 Appendix B

```
knitr::opts_chunk$set(echo = FALSE, comment = NA)
base<-read.csv('data/base_data.csv', stringsAsFactors = F)
P__disp <- function(x) {
  pr <- sum(residuals(x, type="pearson")^2)
  dispersion <- pr/x$df.residual
  c(pr, dispersion)
}
library(dplyr)
library(qpcR)
library(MuMIn)

base.total = base[which(base$year!=2016),]
Olympic = base.total[,c(2,3,4,5,6,7,8)]

#Clean the data (Mary to add)

#Make new dataframe with GDP / capita
Olympic_v2 <- data.frame(year = Olympic$year, country = Olympic$country, count = Olympic$count, log_pop


library(knitr)
summary(base$count)
hist(base$count)
hist(base[which(base$year == c("2008")),c("count")])
summary(base$gdp)
summary(base$pop)
sum(base$comm_soviet)
nrow(base)-sum(base$comm_soviet)
sum(base$host)
par(mfrow=c(1,2))
#Histogram Population
hist(base$pop,main = "Untransformed",xlab = "Population")
#Histogram log(Population)
hist(log(base$pop),main = "Transformed",xlab = "log(Population)")
#Log(GDP)
boxplot(log(base$gdp),ylab = "log(GDP)")
#log(Population)
boxplot(log(base$pop),ylab = "log(Population)")
pairs(Olympic[,c(2,4,5,6,7)])
kable(cor(Olympic[,c(2,4,5,6,7)]))
Olympic.pois<-glm(count~log_pop + log_gdp_pcap + host + comm_soviet, data = Olympic_v2, family = poisso
P__disp(Olympic.pois)
library(MASS)
olympic.nb <- glm.nb(count~log_pop + log_gdp_pcap + host + comm_soviet, data = Olympic_v2)
library(leaps)
olympic.nb_leap <- leaps(y=Olympic_v2$count, x=Olympic_v2[,4:7])
Cp.nb<-round(olympic.nb_leap$Cp, 2)

which<-olympic.nb_leap$which
rownames(which) <-NULL
colnames(which)<-c('Pop', 'GDP/C', 'Host', 'Soviet')
```

```r
xList <- names(Olympic_v2)[4:7]
vec <- olympic.nb_leap$which

#Name the columns in the grid
names(vec) <- paste("X", 1:4, sep="")

#Build matrix of formula for every row
allModelsList <- apply(vec, 1, function(x) as.formula(
  paste(c("count ~ 1", xList[x]), collapse = "+")))

#Calculate the coefficients for all 16 models
allModelsResults <- lapply(allModelsList,
                           function(x) glm.nb(x, data=Olympic_v2))


AIC.nb<-matrix(unlist(lapply(allModelsResults, function(x) round(extractAIC(x),2))), ncol = 2, byrow = T
library(knitr)
kable(cbind(which, Parameters = olympic.nb_leap$size, Cp.nb, AIC.nb), caption = 'Model Selection Diagnos
tab<-summary(olympic.nb)$coefficients
colnames(tab)[4]<-'p-value'
rownames(tab)<-c('Intercept', 'log(population)', 'log(gdp/capita)', 'Host', 'Comm/Soviet')
kable(tab, format.args = list(justify = 'centre'), caption = 'Negative Binomial Model Coefficients')
```