Dan Forbush PA541 Homework 1

Dan Forbush

2/13/2019

## R-Code Setup

#set WD:  
setwd("~/Documents/UIC/PA 541 Data Analysis")  
  
#load libraries:  
library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 2.0.1 ✔ dplyr 0.7.8  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.3.1 ✔ forcats 0.3.0

## Warning: package 'tibble' was built under R version 3.5.2

## ── Conflicts ───────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
  
#load data:  
Olympics <- read.csv("~/Documents/UIC/PA 541 Data Analysis/Olympics.csv")

# Part 1

## Question 1

**The following will require you to use the tools and verbs we learned during week 1 to wrangle data. The results of these tasks will produce a tibble. You only need to copy and paste the tibble itself (what R reports) and not all of the variables or observations (i.e., don’t print out the whole dataset).**

**a. Calculate a new variable, called ‘total.medals’ which is the sum of gold, silver and bronze and add it to the Olympic dataset. (2pts)**

Olympics$total.medals <- Olympics$gold + Olympics$silver + Olympics$bronze

**b. For each country, how many gold medals has it won? (2pts)**

Olympics %>% group\_by(country) %>% summarise(gold.total = sum(gold, na.rm = TRUE))

## # A tibble: 118 x 2  
## country gold.total  
## <fct> <int>  
## 1 "Albania " 0  
## 2 "Algeria " 0  
## 3 American Samoa 0  
## 4 "Andorra " 0  
## 5 "Argentina " 0  
## 6 "Armenia " 0  
## 7 "Australia " 5  
## 8 "Austria " 37  
## 9 "Azerbaijan " 0  
## 10 "Belarus " 6  
## # … with 108 more rows

**c. For each year, how many total medals were given out? (2pts)**

Olympics %>% group\_by(year) %>% summarise(medal.total = sum(total.medals, na.rm = TRUE))

## # A tibble: 10 x 2  
## year medal.total  
## <int> <int>  
## 1 1980 115  
## 2 1984 117  
## 3 1988 138  
## 4 1992 168  
## 5 1994 183  
## 6 1998 205  
## 7 2002 234  
## 8 2006 252  
## 9 2010 258  
## 10 2014 295

**d. Which countries had the largest delegation of athletes in 1992? Create a tibble that contains only the variables country and athletes. (2pts)**

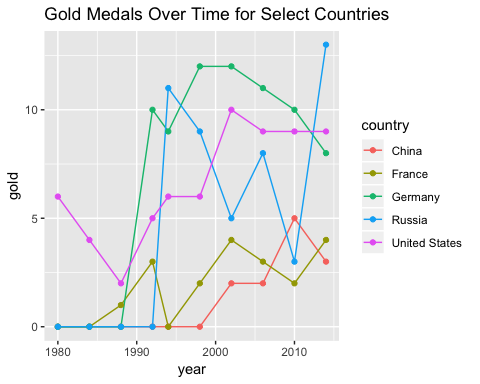
The United States had the largest delegation of athletes. The code is printed below:

Olympics %>%   
 filter(year == 1992) %>%   
 select(country, athletes) %>%   
 group\_by(country)%>%  
 summarise(total.athletes.1992 = sum(athletes)) %>%  
 arrange(desc(total.athletes.1992))

## # A tibble: 113 x 2  
## country total.athletes.1992  
## <fct> <int>  
## 1 "United States " 147  
## 2 Unified Team (Former Soviet) 129  
## 3 "Germany " 111  
## 4 "Canada " 109  
## 5 "France " 109  
## 6 "Italy " 107  
## 7 "Norway " 80  
## 8 Czechoslovakia 74  
## 9 "Switzerland " 74  
## 10 "Sweden " 73  
## # … with 103 more rows

**e. For the following five countries, plot the number of gold medals earned over time: United States, France, Germany, Russia, and China. (hint: see this site for help with creating a line plot** [**http://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization**](http://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization)**) (4pts)**

Olympics.gold <- Olympics %>%  
 filter(country %in% c("United States ","France ","Germany " ,"Russia " , "China ")) %>%  
 select(country, year, gold)  
  
ggplot(Olympics.gold, aes(x=year, y=gold, group=country)) +  
 geom\_line(aes(color=country)) +  
 geom\_point(aes(color=country)) +  
 labs(title = "Gold Medals Over Time for Select Countries")



## Question 2

**Create a new variable, ‘precip.cat’ that breaks the continuous variable precipitation down into three categories: (i) 0 - 25, (ii) >25 - 60, (iii) >60. Please label the categories “low”, “mid”, “high”. Set this new variable to be a factor.**

**What is the mean of total.medals in each of the precipitation categories?**

Olympics = Olympics %>%   
 mutate(precip.cat = case\_when(precipitation > 60 ~ 'High',  
 precipitation > 25 & precipitation <= 60 ~ 'Mid',  
 precipitation <= 25 ~ 'Low'))  
Olympics$precip.cat <- factor(Olympics$precip.cat, levels =c('Low', 'Mid', 'High'))  
class(Olympics$precip.cat)

## [1] "factor"

Olympics %>%   
 group\_by(precip.cat) %>%  
 summarise(mean.total.medals = mean(total.medals, na.rm=TRUE))

## # A tibble: 4 x 2  
## precip.cat mean.total.medals  
## <fct> <dbl>  
## 1 Low 0.569  
## 2 Mid 2.39   
## 3 High 1.53   
## 4 <NA> 8.55

## Question 3

**Run an ANOVA model to test the predictive capability of the percip.cat variable on total.medals.**

anova = aov(total.medals~precip.cat, data = Olympics)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## precip.cat 2 554 277.11 11.83 8.25e-06 \*\*\*  
## Residuals 1096 25670 23.42   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 23 observations deleted due to missingness

TukeyHSD(anova)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = total.medals ~ precip.cat, data = Olympics)  
##   
## $precip.cat  
## diff lwr upr p adj  
## Mid-Low 1.8168737 0.93679290 2.69695451 0.0000043  
## High-Low 0.9646582 0.07724692 1.85206944 0.0292489  
## High-Mid -0.8522155 -1.64420583 -0.06022523 0.0313736

**(a) What do you conclude from the results?**

The ANOVA test shows that there is a significant correlation between precipitation category and total medals. It gives a p value of 0.00000778, indicating that it is highly unlikely that the difference in the means of the total medals among precipitation categories was due to chance alone.

The Tukey results show that there is a strong statistical significance between middle and low precipitation countries, and there is a weaker statistical significance between low and high precipitation countries and medium and high precipitation countries.

**(b) What might cause us to find an association between these variables? Do you think precipitation is having a causal effect on total.medals? What else might be going on?**

I think what is going on here has to do with the Winter Olympics. The Winter Olympics mostly take place in snowy countries because a huge number of events require snow, such as skiing, snowboarding, luging, and others. Low precipitation countries are unlikely to have significant, if any, snow fall which they could use to practice in these events. Likewise, high precipitation countires probably are in tropical regions along the equator or other hot areas of the planet. They would likely be getting high amounts of rain, but not necessarily snow (though some probably do).

Therefore, these high and low precipitation countries may not participate and certainly wouldn’t succeed during the Winter Olympics. This would cause their average number of medals received to fall significantly. During half of the years in the dataset, those countries would receive 0 medals. Therefore, I don’t think that precipitation is having a causal effect on total medals received, but it does have an effect on which countries participate during which years of the Olympics.

# Part 2

**For part 2, let’s limit the Olympic data to only the year 2014 and only countries that participated in the Olympics that year (i.e., participate = 1). So, to begin part 2, create this new dataset.**

Olympics2014 <- Olympics %>%  
 filter(year == 2014 & participate == 1)

## Question 4

**Test the correlation between athletes and total.medals. What are your findings (i.e., what is the size of the correlation and is it significant)?**

cor.test(Olympics2014$athletes, Olympics2014$total.medals)

##   
## Pearson's product-moment correlation  
##   
## data: Olympics2014$athletes and Olympics2014$total.medals  
## t = 19.11, df = 86, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8504756 0.9332556  
## sample estimates:  
## cor   
## 0.8996646

This code indicates that the correlation between athletes and total medals is .899, which is very high, indicating that these two variables are very highly positively correlated. Additionally, the p-value is very small, indicating with a high degree of confidence that this correlation is statistically significant and is not due to chance alone.

## Question 5

**Run a simple linear regression predicting total.medals from athletes.**

model1 <- lm(total.medals ~ athletes, data = Olympics2014)  
summary(model1)

##   
## Call:  
## lm(formula = total.medals ~ athletes, data = Olympics2014)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.6140 -0.4339 0.1964 0.4365 19.6340   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.556596 0.394816 -1.41 0.162   
## athletes 0.120063 0.006283 19.11 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.168 on 86 degrees of freedom  
## Multiple R-squared: 0.8094, Adjusted R-squared: 0.8072   
## F-statistic: 365.2 on 1 and 86 DF, p-value: < 2.2e-16

**(a) Interpret the estimates of the slope and intercept coefficients in the context of the problem.**

The intercept of this model does not tell you much – only that the fit line intersects the y-axis when the number of athletes equals -.556596. This does not mean much in reality, other than the obvious: a country must send more than 0 athletes to receive any medals.

The slope coefficient tells you that with every additional athlete added, the number of total medals received increases on average by .120063.

**(b) What is the percentage of variation in total.medals explained by athletes?**

The r-squared value is .8094, and this means that approximately 80.94% of the total variation of total medals (away from the global mean) is explained by adding the athletes variable.

**(c) Predict the number of medals in a country if the number of athletes sent was 20.**

predict(model1, data.frame(athletes = 20))

## 1   
## 1.844664

The model predicts that a country that sends 20 athletes to the 2014 Olympics would receive 1.844664 total medals on average.

**(d) Explain any suspicion you might have that other factors might explain the observed relationship between the number of athletes and medals.**

I would first be concerned about whether the athletes variables fits with the assumptions necessary for the linear model. I would wonder if the athletes variable would fit with Simple Linear Regression assumption #5 – that athletes variable has homoskedasticity. I wonder if countries that send few athletes would have smaller variances in the total number of medals they receive, and if countries that send many athletes were have high variances in the total number of medals they receive.

Additionally, I think this relationship can be explained in part because of the selection process for how athletes reach the olympics. There are many qualifying rounds that weed out athletes who may perform below a certain competative standard. This acts as a causal mechanism to explain the correlation between athletes sent to the Olympics and total medals - the more athletes from a country that qualify for the Olympics, the more medals that country is likely to receive.

## Question 6

**Which country has the largest negative residual in our model from question 5?**

resid1 <- resid(model1)  
resid1[resid1==min(resid1)] #country 76 has the most negative residual

## 76   
## -8.613985

Olympics2014[76, ] #Switzerland has the most negative residual

## ID year host country temp precipitation elevation gold silver  
## 76 101 2014 0 Switzerland 37.8 65 4634 6 3  
## bronze population GDP participate athletes time total.medals  
## 76 2 7.996861 5.414 1 168 10 11  
## precip.cat  
## 76 High

**Which country has the largest positive residual? (hint see the script from week 4)**

resid1[resid1==max(resid1)] #country 58 has the largest positive residual

## 58   
## 19.63401

Olympics2014[58, ] #the Netherlands has the largest positive residual

## ID year host country temp precipitation elevation gold silver  
## 58 75 2014 0 Netherlands 42.4 66.6 862 8 7  
## bronze population GDP participate athletes time total.medals  
## 58 9 16.75496 4.1459 1 41 10 24  
## precip.cat  
## 58 High

**Tell me what these large positive and large negative residuals mean within the context of our data and model.**

Switzerland had the largest negative residual. In the context of the data, this means that Switzerland acquired a much lower number of total medals compared to what our model would predict given that Switzerland sent 168 athletes to the 2014 Olympics.

The Netherlands had the highest positive residual. In the context of the data, this means that the Netherlands acquired a much higher number of total medals compared to what our model would predict given that the Netherlands only sent 41 athletes to the 2014 Olympics.

## Question 7

**Fit a regression model with total.medals as the dependent variable and athletes, precipitation, and elevation as the predictors.**

model2 <- lm(total.medals~athletes + precipitation + elevation, data=Olympics2014)  
summary(model2)

##   
## Call:  
## lm(formula = total.medals ~ athletes + precipitation + elevation,   
## data = Olympics2014)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5832 -0.4817 0.2105 0.4212 19.4874   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.773e-01 8.689e-01 -0.434 0.665   
## athletes 1.206e-01 6.460e-03 18.663 <2e-16 \*\*\*  
## precipitation 2.794e-05 9.189e-03 0.003 0.998   
## elevation -6.386e-05 1.575e-04 -0.405 0.686   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.202 on 84 degrees of freedom  
## Multiple R-squared: 0.8098, Adjusted R-squared: 0.803   
## F-statistic: 119.2 on 3 and 84 DF, p-value: < 2.2e-16

**(a) What percentage of the variation in the dependent variable is explained by the predictors?**

The r-squared value is .8098, which means that this model accounts for about 80.98% of the variation in the total.medals variable.

**(b) Ignoring whether the predictor is significant or not, interpret the coefficient estimate on precipitation. Be specific when discussing the relationship.**

The estimated coefficient on the precipitation variable is 2.794 x 10-5, or .00002794. This means that with each additional millimeter of precipitation (i.e. with each additional unit as defined by the precipitation variable), the total number of medals in this model is predicted to increase by .00002794.

**(c) How do we interpret the p-value for precipitation?**

The p-value for the precipitation is .998, indicating that we definitely cannot reject the null-hypothesis that the true value of the coefficient on precipitation variable is 0. In other words, there is a very high likelihood that the precipitation coefficient calculated by our model is due to chance alone, and if we were to re-sample the dataset, we would see a coefficient of .00002794 or greater in 99.8% of all the samples.

## Question 8

**Compute a correlation test between the residuals of the model you ran in Question 7 with the fitted or predicted values from that model.**

resid2 <- resid(model2)  
predict2 <- predict(model2)  
cor.test(resid2, predict2)

##   
## Pearson's product-moment correlation  
##   
## data: resid2 and predict2  
## t = -5.5281e-17, df = 86, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.2094423 0.2094423  
## sample estimates:  
## cor   
## -5.961078e-18

**Compute a correlation test between the residuals and athletes.**

cor.test(resid2, Olympics2014$athletes)

##   
## Pearson's product-moment correlation  
##   
## data: resid2 and Olympics2014$athletes  
## t = -2.4077e-16, df = 86, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.2094423 0.2094423  
## sample estimates:  
## cor   
## -2.596297e-17

**What do you make of the results from these two correlation tests?**

Both the first correlation test and the second correlation test return a correlation value that is effectively zero, something that is confirmed by each test having a p-value of 1. In essence, this test confirms that the values we got for each correlation are due entirely to chance. These correlation results mean that the linear model calculation is working as intended.

The reason the first correlation test is 0 has to do with the Ordinary Least Square calculation that is done as part of a linear regression calculation. By minimizing the squares between the fit line and the data, the model is removing any correlation between the residuals (i.e. the variation that the model cannot account for) and the predicted values (i.e. the variation that the model can account for). When the linear model calculation is correct and the sum of the squares of the residuals is minimized, it makes perfect sense for this correlation to be 0.

Likewise with the second correlation test between athletes and the model residuals, the linear model calculates the predicted values using the athletes variable. The residuals are what is not accounted for in that calculation. If there were any correlation between the residuals and the athletes variable, that would mean the sum of the squares was not minimized and there would be a better fit line available that could be calculated.