CSC424: Assignment 1

Jasmine Dumas

January 20, 2016

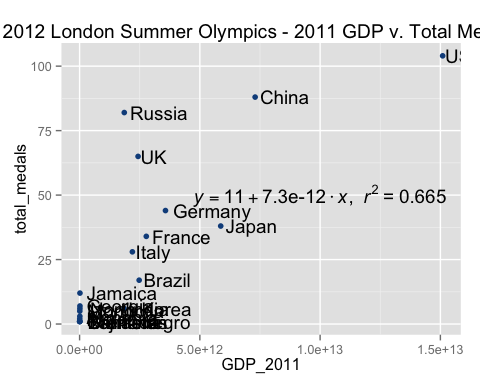
### Problem 1 (5 points – data exploration, visualization, and interpretation)

The Olympics are a great event to watch both in the Summer & Winter and it brings about a lot of national pride. While the games are being played many different facts and statistics emerge about athletes and their rigorous Olympic trial competitiveness, preparation, training, equipment, and sponsorship provided generously by their host country and athletic companies. With the 2012 Summer Olympic data I'm hoping to provide insight by trying to answer the following questions:  
1. *Do wealthier countries win more medals?*: **I would hypothesis that money plays a very significant component in winning medals and there should be a positive trend between those two variables.**2. *Does having a large talent pool of athelets to choose from ensure that a country will win alot of medals?*: **Statistically speaking I would hypothesis that a country with a large population would in then have a large pool athletes and be able to more selectively choose the best of the best and in that would equate to more medals.**

# read in the csv file  
setwd("/Users/jasminedumas/Desktop/depaul/CSC424")  
library(readr)  
olympics <- read\_csv("olympics.csv")  
colnames(olympics) <- c("ISOcountrycode", "Countryname", "GDP\_2011", "population\_2010", "Female\_count", "Male\_count", "Gold\_medals", "Silver\_medals", "Bronze\_medals") # column name should not start with a numeric value

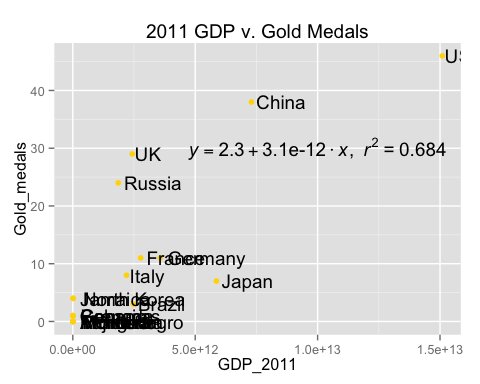
# Question 1  
# plot the relationship between 2010 GDP and total medals  
library(dplyr)

olympics <- mutate(olympics, total\_medals = Gold\_medals + Silver\_medals + Bronze\_medals)  
library(ggplot2)  
source('~/Desktop/R-directory/lm\_eqn.R') # adds the lm equation and R^2 value to plot  
  
gdp\_total <- ggplot(olympics, aes(x=GDP\_2011, y=total\_medals, label = Countryname)) +  
 geom\_point(color = "dodgerblue4") +  
 ggtitle("2012 London Summer Olympics - 2011 GDP v. Total Medals") +  
 geom\_text(hjust = -0.1, nudge\_x = 0.01,   
 check\_overlap = TRUE) +  
 annotate("text", x = 1e+13, y = 50, label = lm\_eqn(df=olympics, y=olympics$total\_medals, x=olympics$GDP\_2011), parse =TRUE)  
gdp\_total

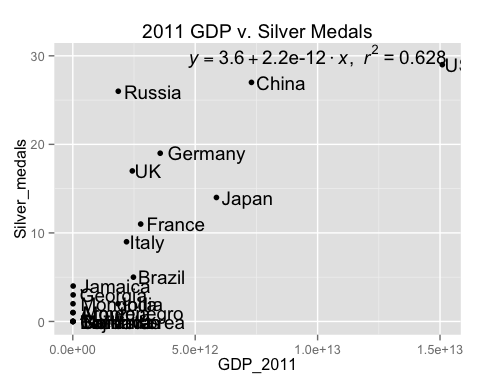


This scatter plot demonstrates the positive linear relationship of GDP to total amount of medals won by a country in the 2012 Olympics. Its quite clear that the more wealthier a country is the more medals they have won. There are some particular examples that deviate from this trend and that would for example Russia who has a much lower GDP (Gross Domestic Product) but significantly more overall medals comparative to Japan.

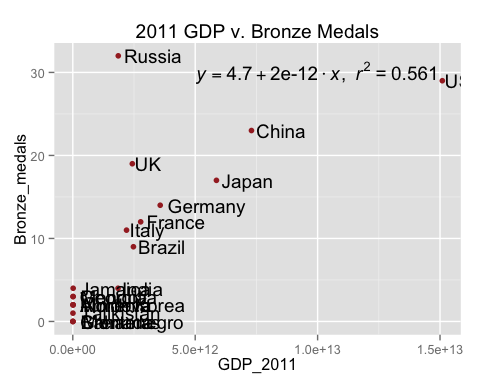
# Question 1 in detail  
# plot the relationship between 2010 GDP and each of the individauls medals  
gold <- ggplot(olympics, aes(x=GDP\_2011, y=Gold\_medals, label = Countryname)) +  
 geom\_point(color = "gold") +  
 ggtitle("2011 GDP v. Gold Medals") +  
 geom\_text(hjust = -0.1, nudge\_x = 0.01, check\_overlap = TRUE) +  
 annotate("text", x = 1e+13, y = 30, label = lm\_eqn(df=olympics, y=olympics$Gold\_medals, x=olympics$GDP\_2011), parse =TRUE)  
gold



silver <- ggplot(olympics, aes(x=GDP\_2011, y=Silver\_medals, label = Countryname)) +  
 geom\_point(color = "black") +  
 ggtitle("2011 GDP v. Silver Medals") +  
 geom\_text(hjust = -0.1, nudge\_x = 0.01, check\_overlap = TRUE) +  
 annotate("text", x = 1e+13, y = 30, label = lm\_eqn(df=olympics, y=olympics$Silver\_medals, x=olympics$GDP\_2011), parse =TRUE)  
silver

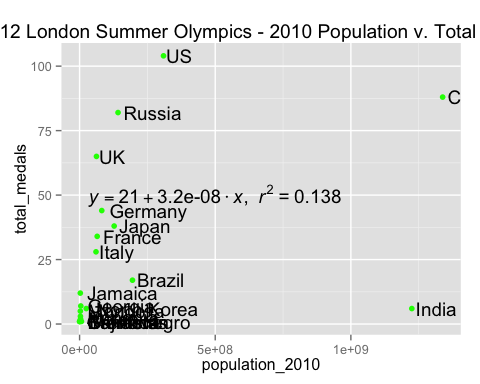


bronze <- ggplot(olympics, aes(x=GDP\_2011, y=Bronze\_medals, label = Countryname)) +  
 geom\_point(color = "brown") +  
 ggtitle("2011 GDP v. Bronze Medals") +  
 geom\_text(hjust = -0.1, nudge\_x = 0.01, check\_overlap = TRUE) +  
 annotate("text", x = 1e+13, y = 30, label = lm\_eqn(df=olympics, y=olympics$Bronze\_medals, x=olympics$GDP\_2011), parse =TRUE)  
bronze



These additional scatter plot goes to underscore the effect of GDP on the individual medal counts. For example Russia with a much lower GDP than the US earned more bronze medals in the games. Not all medals are obviously equal but with milliseconds and minimal points between first, second, and third place winning a medal in general is still seen as success for many countries.

# Question 2  
# plot the relationship between 2010 population and the total medals  
population <- ggplot(olympics, aes(x=population\_2010, y=total\_medals, label = Countryname)) +  
 geom\_point(color = "green") +  
 ggtitle("2012 London Summer Olympics - 2010 Population v. Total Medals") +  
 geom\_text(hjust = -0.1, nudge\_x = 0.01,   
 check\_overlap = TRUE) +  
 annotate("text", x = 5e+08, y = 50, label = lm\_eqn(df=olympics, y=olympics$total\_medals, x=olympics$population\_2010), parse =TRUE)  
  
population



This scatter plot shows a story different from my hypothesis and in fact having a populous country does not ensure that you will have many athletes in your population to choose from. For example the US has significantly lower population than India and China but has the most medals out of all of the countries.

### Problem 2: (10 points – regression analysis):

1. (2 points) Find the regression of LeafIndex on Latitude. Is latitude a useful predictor of leaf index? ***Latitude is a highly significant predictor with a p-value of less tha 0.05***

# read in data  
maple <- read\_table("maple.txt")  
fit\_lat <- lm(LeafIndex ~ Latitude, data = maple)  
summary(fit\_lat)

##   
## Call:  
## lm(formula = LeafIndex ~ Latitude, data = maple)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2348 -0.8488 0.0773 1.0074 3.3305   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.66716 3.05202 -0.546 0.589   
## Latitude 0.45369 0.07427 6.108 1.03e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.673 on 30 degrees of freedom  
## Multiple R-squared: 0.5543, Adjusted R-squared: 0.5394   
## F-statistic: 37.31 on 1 and 30 DF, p-value: 1.031e-06

1. (2 points) Repeat part (a) for the regression of LeafIndex on JulyTemp. **JulyTemp is a highly significant predictor with a p-value of less tha 0.05**

fit\_july <- lm(LeafIndex ~ JulyTemp, data = maple)  
summary(fit\_july)

##   
## Call:  
## lm(formula = LeafIndex ~ JulyTemp, data = maple)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.254 -1.288 0.096 1.245 3.212   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 40.74297 4.45498 9.145 3.51e-10 \*\*\*  
## JulyTemp -0.33318 0.06206 -5.368 8.23e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.789 on 30 degrees of freedom  
## Multiple R-squared: 0.49, Adjusted R-squared: 0.473   
## F-statistic: 28.82 on 1 and 30 DF, p-value: 8.233e-06

1. (6 points) Find the regression of LeafIndex on Latitude and JulyTemp. Compare the results of this analysis with your results from (a) and (b). How different are the slope coefficients in each case? What best explains the differences in their values? ***This regression fit did not indicate that the predictors were very significant (high p-values) but the overall p-value was significant which is suspicious. The coefficents in each case are all different except the signs for JulyTemp are all negative which is appropriate needeing moderate temperatures for tree growth. The correlation tests indicated a 1) very high negative correlation between the predictors 2) high correlation with latitude and the response 3) high negative corerlation with JulyTemp and the response. This multicolinarity bewteen the predictors explains the different in thier values amongst the regression fits***

fit <- lm(LeafIndex ~ JulyTemp + Latitude, data = maple)  
summary(fit)

##   
## Call:  
## lm(formula = LeafIndex ~ JulyTemp + Latitude, data = maple)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2082 -1.2363 0.1613 1.0551 3.2445   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.73184 11.42026 1.202 0.2389   
## JulyTemp -0.13524 0.09676 -1.398 0.1728   
## Latitude 0.31393 0.12388 2.534 0.0169 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.647 on 29 degrees of freedom  
## Multiple R-squared: 0.5824, Adjusted R-squared: 0.5536   
## F-statistic: 20.22 on 2 and 29 DF, p-value: 3.167e-06

# correlation tests  
cor(maple$Latitude, maple$JulyTemp)

## [1] -0.8072089

cor(maple$Latitude, maple$LeafIndex)

## [1] 0.7445147

cor(maple$JulyTemp, maple$LeafIndex)

## [1] -0.6999725

### Problem 3: (30 points – regression analysis)

1. (5 points) Before running any regressions make a prediction as to what the sign of the coefficient of each predictor should be expected to be. **My predictions for the signs of the coefficients are negative: *pctmin, fires, thefts, & pctold* and positive: *income*. My predictions about the signs are based on the notion that the negative coefficients are housing attributes and the positive coefficients are occupant attributes.**

(5 points) Obtain the correlation matrix for the variables PCT-MINOR FIRES THEFTS PCTOLD INCOME NEWPOL. Do the simple correlations support your predictions about the signs? ***By examining the newpol column all the correpsonding rows have a negative correlation sign for the housing attributes and the only positive correlation is with income. Yes, this correlation matrix supports my previous prediction about coefficient beta signs.***

chi\_home\_insurance\_original <- read\_table("chicinsur.txt")  
# make a copy of the data to remove irelevant columns  
chi\_home\_insurance <- chi\_home\_insurance\_original  
chi\_home\_insurance$zipcode <- NULL  
chi\_home\_insurance$fairpol <- NULL  
# correlation matrix  
cor(chi\_home\_insurance)

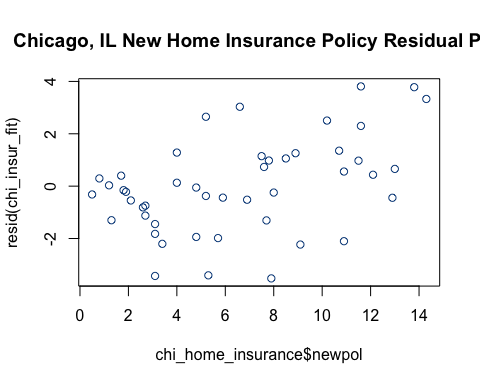
## pctmin fires thefts pctold newpol income  
## pctmin 1.0000000 0.5927956 0.2550647 0.2505118 -0.7594196 -0.7037328  
## fires 0.5927956 1.0000000 0.5562105 0.4122225 -0.6864766 -0.6104481  
## thefts 0.2550647 0.5562105 1.0000000 0.3176308 -0.3116183 -0.1729226  
## pctold 0.2505118 0.4122225 0.3176308 1.0000000 -0.6057428 -0.5286695  
## newpol -0.7594196 -0.6864766 -0.3116183 -0.6057428 1.0000000 0.7509780  
## income -0.7037328 -0.6104481 -0.1729226 -0.5286695 0.7509780 1.0000000

1. Run a multiple regression of NEWPOL on the variables listed above.
2. (5 points) Comment on the overall significance of the regression fit. ***The overall p-value for the regression fit is significant (p-value < 0.05). The adjusted R-squared value is moderatly high meaning the model explains 77% of the error.***
3. (5 points) Which predictors have coefficients that are significantly different from zero at the .05 level? ***pctmin, fires and pctold are significant predictors.***
4. (5 points) Do any of the predictors have signs that are different than suggested by their simple correlations? If so, explain what may be happening. If not, explain how such a thing can happen. ***The theft predictor has a different (positive) sign then previously predicted as negative. This sign change could be in the correlation matrix thefts did not exhibit a strong corerlation with the response variable. The sign change could also be due to more new home insurance policies to protect thier property from thefts.***
5. (5 points) Examine a plot of residuals versus predicted values. Do you see any problems? ***There are no observable patterns on the residual plot which points to the appropriateness of the model*.**

# multiple linear regression model to determine number of new home insurance policies  
chi\_insur\_fit <- lm(newpol ~ pctmin + fires + thefts + pctold + income, data = chi\_home\_insurance)  
summary(chi\_insur\_fit)

##   
## Call:  
## lm(formula = newpol ~ pctmin + fires + thefts + pctold + income,   
## data = chi\_home\_insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5235 -1.2134 -0.1544 1.0181 3.8096   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.0610686 2.8187818 4.279 0.000110 \*\*\*  
## pctmin -0.0594738 0.0131806 -4.512 5.3e-05 \*\*\*  
## fires -0.1018544 0.0480110 -2.121 0.039972 \*   
## thefts 0.0135616 0.0162371 0.835 0.408436   
## pctold -0.0643711 0.0158312 -4.066 0.000211 \*\*\*  
## income 0.0001164 0.0001804 0.645 0.522525   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.907 on 41 degrees of freedom  
## Multiple R-squared: 0.7939, Adjusted R-squared: 0.7688   
## F-statistic: 31.59 on 5 and 41 DF, p-value: 4.773e-13

# residual plot  
plot(chi\_home\_insurance$newpol, resid(chi\_insur\_fit), main= "Chicago, IL New Home Insurance Policy Residual Plot", col="dodgerblue4")



### Problem 4 (5 points –regression application):

Briefly describe an application for the multiple regression in a field of interest to you. Identify possible independent variables and the dependent variable for your application. If you read about the application from a research paper or news article, please provide a reference to it.

***I'm really interested in craft beer (like alot of people) but I'm particularly interested in all of the data that surrounds craft beer and the rating systems for ranking beer. This is a great application of multiple regression can can consist of numerous independent variables including: beer style, abv %, composition, & price to predict the dependent response for rating score. Here are a few links of my own multiple linear regression on determining rating score:*** [**https://github.com/jasdumas/regression/blob/master/beer/beer.Rmd**](https://github.com/jasdumas/regression/blob/master/beer/beer.Rmd) ***and a data vizualization about beer rating in craft breweries in Connecticut*** [**http://rpubs.com/jasdumas/Breweries-scrape**](http://rpubs.com/jasdumas/Breweries-scrape)