HW1-Matthew Fein-545

Matthew Fein

January 26, 2019

## Chocolate and Nobel Prizes

1. In this study Meserli operationally defines “number of Nobel laureates per capita” as a measure of a countrry’s overall cognitive function. I do not think that this is a reasonable approach. In general Nobel laureates do not seem to be representative of the population of a country as a whole but rather a small subset of the population which is of above average intelligence and cognitive function.
2. No, countries without any nobel prize winners are not included in the study (Brazil has 1 winner). This introduces the issue of selection bias which suggests that the sample used is not representative of the populations they are meant to repreent. Information bias is also an issue, especially if one considers number of nobel prizes as a poor measure of national cognitive functioning.

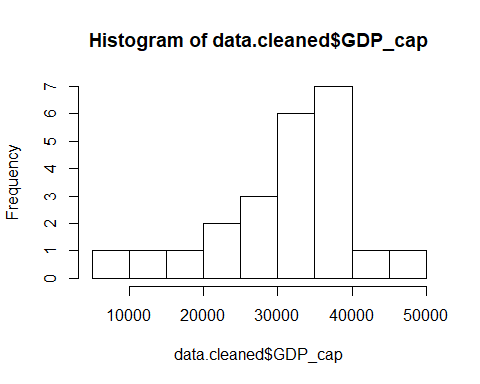
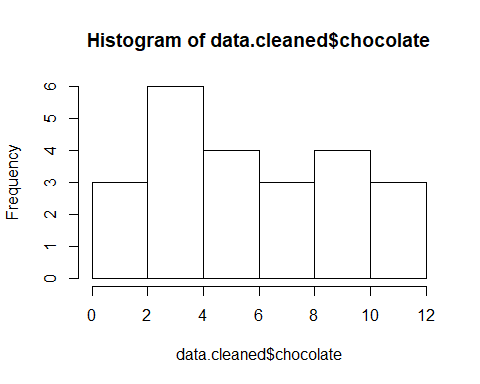
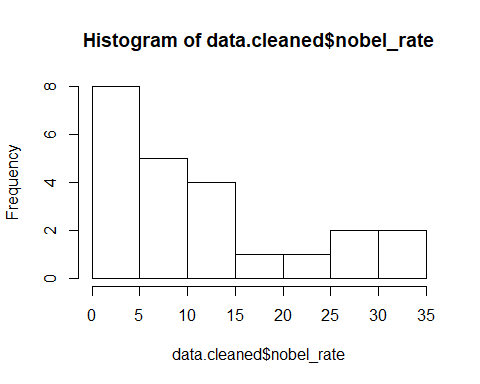
3)No, amount of nobel laureates and chocolate consumption are not measured on the same temporal scale. This is problematic because the data is presented in such a way that this difference in data collection time is not clear. Figure 1 is displaying correlational data, however, the x and y axes represent information collected over different amounts of time. There may be an effect of consumption during specific years and Nobel winneres but this can not be examined.

1. Summary statistics for select variables

data <- read.csv("nobel\_chocolate.txt", header = TRUE) # import data   
data.cleaned <- data[ ,1:3]  
data.cleaned[ ,4]<- data[ , 6] #create new df with only relevant variables  
colnames(data.cleaned)[4] <-"GDP\_cap"   
descriptives <-summary(data.cleaned[,c("nobel\_rate", "chocolate", "GDP\_cap")]) #create summary stats table  
SD<- round(apply(data.cleaned[,c("nobel\_rate", "chocolate", "GDP\_cap")], 2, sd, na.rm=TRUE), digits=2)  
descriptives<- rbind(descriptives, SD) #add row of SDs to table  
print(descriptives)

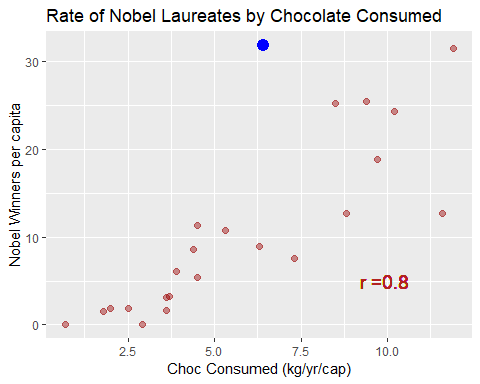
## nobel\_rate chocolate GDP\_cap   
## "Min. : 0.050 " "Min. : 0.700 " "Min. : 7418 "  
## "1st Qu.: 2.490 " "1st Qu.: 3.600 " "1st Qu.:27003 "  
## "Median : 8.622 " "Median : 4.500 " "Median :32881 "  
## "Mean :11.089 " "Mean : 5.804 " "Mean :30592 "  
## "3rd Qu.:15.790 " "3rd Qu.: 8.650 " "3rd Qu.:35929 "  
## "Max. :31.855 " "Max. :11.900 " "Max. :46733 "  
## SD "10.22" "3.28" "9467.66"

## NULL



The distribution for nobel rate is positively skewed, chocolate consumed is roughly uniform (or positively skewed) and GDP per capite is negatively skewed.

data.cleaned[ ,5] <- "Not Sweden" #create label for Sweden  
data.cleaned[data.cleaned$country == "Sweden", 5] <- "Sweden"  
r<- cor(data.cleaned$chocolate, data.cleaned$nobel\_rate) # calculate r  
NLxChoc <- data.cleaned %>% ggplot(aes(x=chocolate, y=nobel\_rate, color = V5)) +   
 geom\_point(color="#99000070", size=2) +  
 geom\_point(data=data.cleaned[data.cleaned$country == "Sweden",], color = "blue", size = 4) +  
 ggtitle("Rate of Nobel Laureates by Chocolate Consumed") +  
 labs(x="Choc Consumed (kg/yr/cap)", y="Nobel Winners per capita") +  
 geom\_text(label = "r =0.8 ", x=10, y=5, size=5, color="firebrick") # I couldn't get Sweden to have a label but it is blue...  
NLxChoc

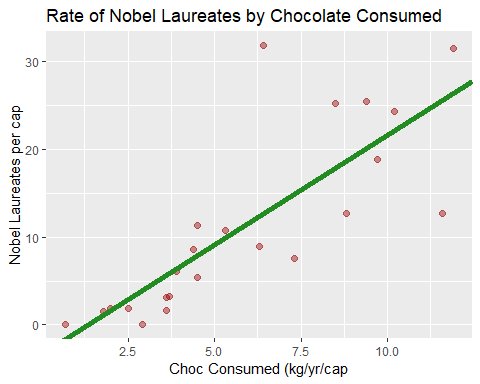
 There is a strong positive correlation between these two variables. I think this is an inappropriate measure because two very different time ranges are covered in each variable.

1. Meserli’s correlation value is 0.791 whereas mine is 0.8019. I would attribute this to rounding error since the values are so close together.
2. Meserli considers Sweden an outlier because it’s actual amount of Nobel laureates is more than twice as much as would be expected based on the model. He attributes this to either the Nobel Committee having a patriotic bias or Swedes being incredibly sensitive to the cognitive powers of chocolate.

lm.choc <- lm(nobel\_rate ~ chocolate, data = data.cleaned) #regress nobels against chocolate consumed  
summary(lm.choc)#model results with hypothesis tests and p values

##   
## Call:  
## lm(formula = nobel\_rate ~ chocolate, data = data.cleaned)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.888 -2.953 -0.213 1.992 19.279   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.400 2.699 -1.260 0.222   
## chocolate 2.496 0.407 6.133 4.37e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.26 on 21 degrees of freedom  
## Multiple R-squared: 0.6418, Adjusted R-squared: 0.6247   
## F-statistic: 37.62 on 1 and 21 DF, p-value: 4.374e-06

#make scatterplot with line  
s.1 <- data.cleaned %>% ggplot(aes(x=chocolate, y=nobel\_rate)) + geom\_point(color="#99000070", size=2) +  
 ggtitle("Rate of Nobel Laureates by Chocolate Consumed") +  
 labs(x="Choc Consumed (kg/yr/cap", y="Nobel Laureates per cap") +  
 geom\_abline(intercept= coef(lm.choc)[1] , slope=coef(lm.choc)[2], color="forestgreen", size=2)   
s.1



#plot residuals for residual analysis   
r.1 <- lm.choc %>% ggplot(aes(x=chocolate, y=.resid)) +   
 geom\_point(color="#99000070", size=2) +  
 ggtitle("Residual Plot") +  
 labs(x="Chocolate consumed (kg/yr/cap)", y="Residuals") +  
 geom\_hline(yintercept=0, color="gray50", lty=2)

a: y = -3.4 + 2.5x +e b: For every kg increase of chocolate consumed per year per capita a nation should win an additional 2.5 nobel prizes. c: I think that the residual plot does not illustrate any violations of the assumptions for conducting a hypothesis test on the data at hand. d: The slope is significantly different than 0, t(21)=6.13, p<.05.

#extract fitted value for Sweden  
lm.choc$fitted.values[data.cleaned$country == "Sweden"]

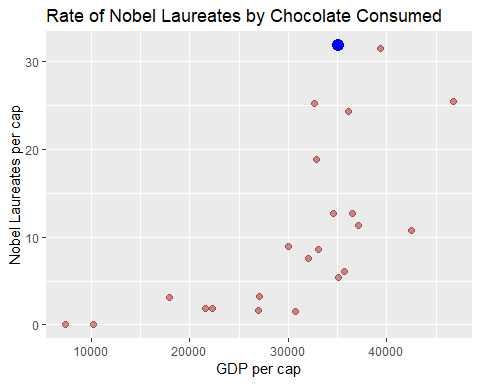
## 20   
## 12.57568

#extract residual error for Sweden  
lm.choc$residuals[data.cleaned$country == "Sweden"]

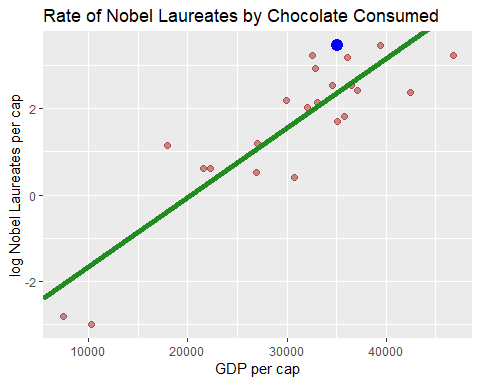
## 20   
## 19.27932

Both values are measured in Nobel Laureates per 10M in population (expected 12.57 and residual 19.23).

data.cleaned[ ,6] <- log(data.cleaned$nobel\_rate)  
colnames(data.cleaned)[6] <- "logNobel\_rate"  
lm.gdp <- lm(logNobel\_rate ~ GDP\_cap, data= data.cleaned)  
  
s.2 <- data.cleaned %>% ggplot(aes(x=GDP\_cap, y=nobel\_rate)) + geom\_point(color="#99000070", size=2) +  
 ggtitle("Rate of Nobel Laureates by Chocolate Consumed") +  
 geom\_point(data=data.cleaned[data.cleaned$country == "Sweden",], color = "blue", size = 4) +  
 labs(x="GDP per cap", y="Nobel Laureates per cap")   
s.3 <-data.cleaned %>% ggplot(aes(x=GDP\_cap, y=logNobel\_rate)) + geom\_point(color="#99000070", size=2) +  
 ggtitle("Rate of Nobel Laureates by Chocolate Consumed") +  
 geom\_point(data=data.cleaned[data.cleaned$country == "Sweden",], color = "blue", size = 4) +  
 labs(x="GDP per cap", y="log Nobel Laureates per cap") +  
 geom\_abline(intercept= coef(lm.gdp)[1] , slope=coef(lm.gdp)[2], color="forestgreen", size=2)  
s.2



s.3



lm.gdp <- lm(logNobel\_rate ~ GDP\_cap, data= data.cleaned)  
summary(lm.gdp)

##   
## Call:  
## lm(formula = logNobel\_rate ~ GDP\_cap, data = data.cleaned)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.3790 -0.6689 0.1134 0.5292 1.5190   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.267e+00 6.084e-01 -5.369 2.52e-05 \*\*\*  
## GDP\_cap 1.607e-04 1.903e-05 8.445 3.42e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8452 on 21 degrees of freedom  
## Multiple R-squared: 0.7725, Adjusted R-squared: 0.7617   
## F-statistic: 71.31 on 1 and 21 DF, p-value: 3.422e-08

A: The model with log Nobel rate is more linear. B: Sweden is no longer an outlier when log nobel rate is used. China and Brazil are the lowermost to the left. D: For every increase of one unit of per capita GDP there is a 0.0167 increase in log Nobel rate.

1. No, we can never infer causation when using correlational data.