SDGB 7840 HW 1: Chocolate & Nobel Prizes

Darshan Patel

1/18/2019

Read the *New England Journal of Medicine* article, “Chocolate Consumption, Cognitive Function, and Nobel laureates” (Messerli, F.H., Vol. 367(16), 1562-1564; 2012). A reconstruction of Messerli’s data is contained in the file “nobel\_chocolate.txt.” The information gathered in the dataset is from several different sources. The number of Nobel prize winnders is from Wikipedia and includes winners through November 2012, population information (used to compute the “nobel\_rate” variable) is from the World Bank and chocolate market size is from the Euromonitor International’s Passport Database.

Replicate Messerli’s analysis.

### Question 1: According to Messerli, what is the variable “number of Nobel laureates per capita” supposed to measure? Do you think it is a reasonable measure? Justify your answer.

Answer: According to Messerli, the variable “number of Nobel laureates per capita” represents the proportion of people with “superior cognitive function.” He believes that this would give a sense of the overall cognitive function of the country. In my opinion, this is not a reasonable measure because it is letting just a handful of people who have gotten an award to represent intelligence of a country. To better measure the cognitive function of a country, Messerli should use something that takes a greater proportion of the country’s population and measures their cognitive intelligence rather than scaling up the variable by 10 million.

### Question 2: Are countries without Nobel prize recipients included in Messerli’s study? If not, what type of bias(es) would that introduce?

Answer: Countries without Nobel prize recipients are not included in Messerli’s study. This introduces selection bias because data is selected subjectively, as in, if there was a Nobel laureate from that country.

### Question 3: Are the number of Nobel laureates per capita and chocolate consumption per capita measured on the same temporal scale? If not, how could this affect the analysis?

Answer: The number of Nobel laureates per capita and chocolate consumption per capita are not measured on the same temporal scale. As stated in the article, data for countries came from different times. Data was available for country from 2011, countries from 2010, from 2004 and from 2002. There is variation in the time collection of the Nobel laureates. In addition, the data for chocolate consumption may have been taken from one year alone, that could be far from when the Nobel prize was given out in some countries. There is no consistency here.

### Question 4: Create a table of summary statistics for the following variables: Nobel laureates per capita, GDP per capita, and chocolate consumption. Include the statistics: minimum, maximum, median, mean and standard deviation. Remember to include the units of measurement in your table.

Answer:

# Import tidyverse   
library(tidyverse)

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

## ✔ ggplot2 3.1.0.9000 ✔ purrr 0.2.5   
## ✔ tibble 2.0.1 ✔ dplyr 0.7.8   
## ✔ tidyr 0.8.1 ✔ stringr 1.3.0   
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## Warning: package 'tibble' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

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

# Read in the data and select only the columns of data needed  
df <- read\_delim("nobel\_chocolate.txt", delim = ',')

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## country = col\_character(),  
## totalpop = col\_integer(),  
## patentR = col\_integer(),  
## patentNR = col\_integer(),  
## fh\_cl = col\_integer(),  
## fh\_fotpsc5 = col\_integer(),  
## fh\_pr = col\_integer()  
## )

## See spec(...) for full column specifications.

df <- df %>% select("country", "nobel\_rate", "GDP\_cap", "chocolate")  
  
# Create a table of minimum, maximum, median, mean and standard  
# deviation of the 3 variables   
sum\_stats <- data.frame("min" = apply(df[,c("nobel\_rate", "GDP\_cap",   
 "chocolate")],2, min),   
 "max" = apply(df[,c("nobel\_rate", "GDP\_cap",  
 "chocolate")], 2, max),   
 "med" = apply(df[,c("nobel\_rate", "GDP\_cap",  
 "chocolate")], 2, median),   
 "mean" = apply(df[,c("nobel\_rate", "GDP\_cap",  
 "chocolate")], 2, mean),   
 "sd" = apply(df[,c("nobel\_rate", "GDP\_cap",  
 "chocolate")], 2, sd))  
  
# Rename the rows   
rownames(sum\_stats) <- c("Nobel laureates per capita, in number of Nobel laureates per 10 million population",   
 "GDP per capita, in units of currency per person",   
 "chocolate consumption, in kg/year/capita")  
  
# Print the summary statistics table in a viewable manner  
t(sum\_stats)

## Nobel laureates per capita, in number of Nobel laureates per 10 million population  
## min 0.05000  
## max 31.85500  
## med 8.62200  
## mean 11.08878  
## sd 10.21818  
## GDP per capita, in units of currency per person  
## min 7417.888  
## max 46733.359  
## med 32880.582  
## mean 30592.143  
## sd 9467.658  
## chocolate consumption, in kg/year/capita  
## min 0.700000  
## max 11.900000  
## med 4.500000  
## mean 5.804348  
## sd 3.279201

### Question 5: Create histograms for the following variables: Nobel laureates per capita, GDP per capita and chocolate consumption. Describe the shape of the distributions.

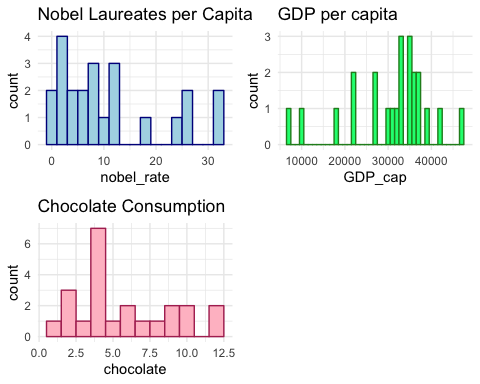
Answer:

# Import gridExtra  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

# Create histograms for each of the 3 variables   
g1 <- ggplot(df, aes(x = nobel\_rate)) + geom\_histogram(binwidth = 2,   
 color = "darkblue",  
 fill = "lightblue") +   
 ggtitle("Nobel Laureates per Capita") + theme\_minimal()  
g2 <- ggplot(df, aes(x = GDP\_cap)) + geom\_histogram(binwidth = 1000,   
 color = "forestgreen",   
 fill = "springgreen") +   
 ggtitle("GDP per capita") + theme\_minimal()  
g3 <- ggplot(df, aes(x = chocolate)) + geom\_histogram(binwidth = 1,   
 color = "maroon",   
 fill = "pink") +   
 ggtitle("Chocolate Consumption") + theme\_minimal()  
  
# Print the histograms in an aesthetic view  
grid.arrange(g1, g2, g3, ncol = 2)

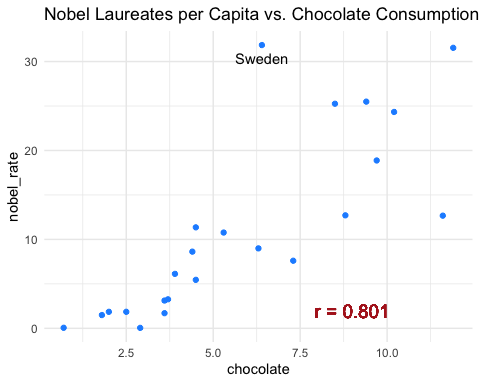


The distribution of noble laureates per capita is skewed right, as well as the distribution of chocolate consumption. The distribution of GDP per capita is slightly skewed left.

### Question 6: Construct a scatterplot of Nobel laureates per capita vs. chocolate consumption. Label Sweden on the plot. Compute the correlation between these two variables and add it to the scatterplot. How would you describe this relationship? Is correlation an appropriate measure? Why or why not?

Answer:

# Create scatterplot of NL per capita vs chocolate consumptions  
# Add the label of Sweden and the correlation value  
ggplot(df, aes(x = chocolate, y = nobel\_rate)) +   
 geom\_point(color = "dodgerblue") +   
 geom\_text(data = df[df$country == "Sweden",],   
 label = "Sweden", nudge\_y = -1.5) +   
 ggtitle("Nobel Laureates per Capita vs. Chocolate Consumption") +   
 geom\_text(label = paste("r = ",   
 round(cor(df$nobel\_rate, df$chocolate), 3),   
 sep = ""),   
 x=9, y=2, size=5, color="firebrick") +   
 theme\_minimal()



The relationship between nobel laureates per capita and chocolate consumption is fairly strong and linear. The correlation between nobel laureates per capita and chocolate consumption is . Correlation would be an appropriate measure here because the relationship between the two variables is evident and can be explained using a regression line.

### Question 7: What is Messerli’s correlation value? (Use the correlation value that includes Sweden.) Why is your correlation different?

Answer: Messerli’s correlation value is (with Sweden). The correlation calculated here is different because we used an updated data set. Namely, the data set here included winners through November 2012 whereas the one Messerli used only included winners through October 20, 2011.

### Question 8: Why does Messerli consider Sweden an outlier? How does he explain it?

Answer: Messerli considers Sweden as an outlier because its predicted number of Nobel laureates is when it is actually , more than twice the value expected. Messerli proposes that there is some patriotic bias by the Nobel Committee in Stockholm (the capital of Sweden) therefore they may be inclined to give out more Noble prizes to people in their country. Messerli also proposes that the Swedes are sensitive to chocolate, that even a small amount of chocolate greatly impacts their cognitive skills.

### Question 9: Regress Nobel laureates per capita against chocolate consumption (include Sweden.):

Answer:

# Regress NL against chocolate consumption  
model <- lm(nobel\_rate~chocolate, df)  
  
# Print the output of the model  
summary(model)

##   
## Call:  
## lm(formula = nobel\_rate ~ chocolate, data = df)  
##   
## 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

1. What is the regression equation? (Include units of measurement.)

# Print coefficients  
summary(model)$coefficients[,1]

## (Intercept) chocolate   
## -3.400366 2.496258

The regression equation is

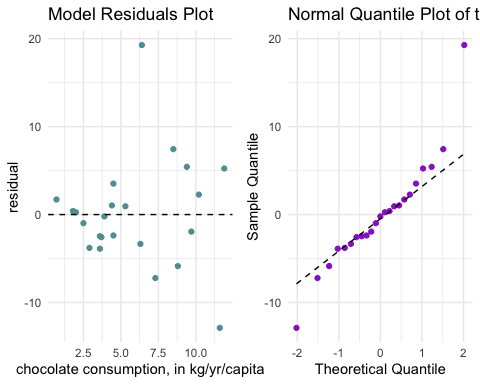
where is the number of Nobel laurates per million population and is the chocolate consumption in kg/yr/capita.

1. Interpret the slope.

A one unit increase in the chocolate consumption in kg/yr/capita is associated with a increase in the number of Nobel laureates per 10 million population.

1. Conduct a residual analysis to check the regression assumptions. Make all plots within one figure. Can we conduct hypothesis tests for this regression model? Justify your answer.

# Import broom  
library(broom)  
  
# Find the residuals  
residuals <- augment(model)[,c("chocolate", ".resid")]  
  
# Create a plot showing the residuals vs x values  
r1 <- ggplot(residuals, aes(x = chocolate, .resid)) +   
 geom\_point(color = "cadetblue") +   
 geom\_hline(yintercept = 0, linetype = "dashed") +   
 ggtitle("Model Residuals Plot") +   
 labs(x = "chocolate consumption, in kg/yr/capita", y = "residual") +   
 theme\_minimal()  
  
# Create a plot showing the normal quantile plot of the residuals  
r2 <- ggplot(residuals, aes(sample = .resid)) +   
 stat\_qq(color = "darkorchid") +   
 stat\_qq\_line(linetype = "dashed") +   
 ggtitle("Normal Quantile Plot of the Residuals") +   
 labs(x = "Theoretical Quantile", y = "Sample Quantile") +   
 theme\_minimal()  
  
# Print the plots in one figure  
grid.arrange(r1, r2, ncol =2)



The model residuals plot shows that the residuals are not fairly bounded as chocolate consumption increase. Rather, the residuals become more spread out as chocolate consumptions increase. The normal quantile plot of the residuals shows that the points form a relatively straight line. This means that the data comes from a normal distribution.

Therefore the constant variance assumption is violated while the normality assumption is satisfied. Hypothesis testing can be conducted for this regression model with caution.

1. Is the slope significant (conduct a hypothesis test and include the regression output)? Test at the level and specify the hypotheses being tested.

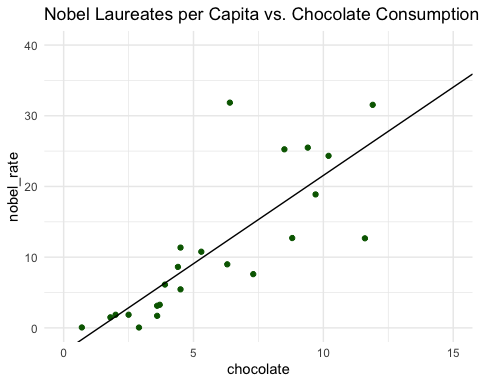
# Print the output of the model  
summary(model)

##   
## Call:  
## lm(formula = nobel\_rate ~ chocolate, data = df)  
##   
## 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

Let the null hypothesis be: and the alternative hypothesis be: . Let . Then the -statistic is calculated to be with an associated -value of roughly . This means that the slope, , is statistically significant since the -value is strictly less than .

1. Add the regression line to the scatterplot.

# Plot the regression line on the scatterplot above  
ggplot(df, aes(x = chocolate, y = nobel\_rate)) +   
 geom\_point(color = "darkgreen") +   
 geom\_abline(intercept = summary(model)$coefficients[1,1],  
 slope = summary(model)$coefficients[2,1]) +   
 lims(x = c(0, 15), y = c(0, 40)) +   
 ggtitle("Nobel Laureates per Capita vs. Chocolate Consumption") +   
 theme\_minimal()



### Question 10: Using the model, what is the number of Nobel laureates expected to be for Sweden? What is the residual? (Include units of measurement.)

Answer: According to the model, the number of Nobel laureates expected for Sweden is

nl\_sweden <- predict(model, newdata = df[df$country == "Sweden","chocolate"])  
nl\_sweden

## 1   
## 12.57568

Nobel laureates per 10 million population. The residuals, or difference between the predicted value and the actual number of Nobel laureates per 10 million population is

error <- df[df$country == "Sweden", "nobel\_rate"] - nl\_sweden  
error

## nobel\_rate  
## 1 19.27932

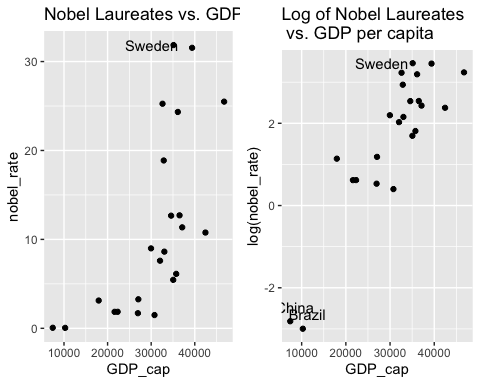
### Question 11: Now we will see if the variable GDP per capital (i.e., “GDP\_cap”) is a better way to predict Nobel laureates.

Answer: (a) In one figure, construct a scatter plot of (i) Nobel laureates vs. GDP per capita and (ii) log(Nobel laureates) vs. GDP per capita. Which plot is more linear? Label Sweden on both plots. On the second plot, label the two countries which appear on the bottom left corner.

# Find the countries that were outliers  
df[log(df$GDP\_cap) < 9.5,]

## # A tibble: 2 x 4  
## country nobel\_rate GDP\_cap chocolate  
## <chr> <dbl> <dbl> <dbl>  
## 1 Brazil 0.05 10264. 2.9  
## 2 China 0.06 7418. 0.7

# Plot GDP\_cap vs nobel\_rate and   
# label Sweden  
g3 <- ggplot(df, aes(x = GDP\_cap, y = nobel\_rate)) + geom\_point() + ggtitle("Nobel Laureates vs. GDP per capita") +   
 geom\_text(data = df[df$country == "Sweden",], label = "Sweden", nudge\_x = -5000)  
  
# Plot GDP\_cap vs nobel\_rate and  
# label Sweden and the other 2 outliers  
g4 <- ggplot(df, aes(x = GDP\_cap, y = log(nobel\_rate))) + geom\_point() + ggtitle("Log of Nobel Laureates \n vs. GDP per capita") + geom\_text(data = df[df$country == "Sweden",], label = "Sweden", nudge\_x = -7000) + geom\_text(data = df[log(df$GDP\_cap) < 9.5,], label = c("Brazil", "China"), nudge\_x = 1000, nudge\_y = 0.35)  
  
# Print the plot in 1 figure  
grid.arrange(g3, g4, ncol = 2)



The plot of log of Nobel Laureates per capita vs. GDP per capita is more linear than the plot of Nobel Laureates vs. GDP per capita.

1. Is Sweden still an outlier? Justify your answer.

Sweden does not appear to be an outlier in this scenario because its both its Nobel Laureates per capita value and log of Nobel Laureates per capita value does not deviate much from the other countries when plotted against the GDP per capita.

1. Regress log of Nobel laureates against GDP per capita. Provide the output and add the regression line to the scatterplot.

# Create a linear model to preduct log of Nobel   
# Laureates using GDP per capita  
model2 <- lm(log(nobel\_rate)~GDP\_cap, df)  
  
# Print the output of the model  
summary(model2)

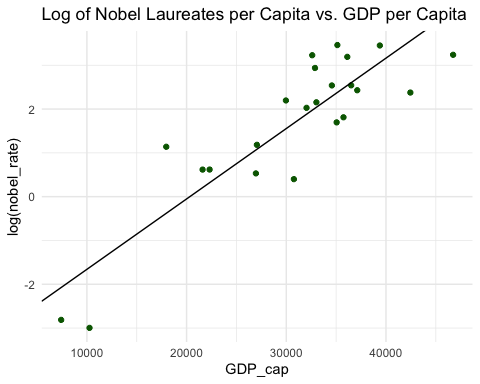
##   
## Call:  
## lm(formula = log(nobel\_rate) ~ GDP\_cap, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.3790 -0.6689 0.1134 0.5292 1.5190   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.267e+00 6.084e-01 -5.369 2.52e-05 \*\*\*  
## GDP\_cap 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

The regression line has the equation

where is the GDP per capita, in currency per person, and is the log of Nobel laureates per capita, in number of Nobel laurates per 10 million population. According to this model, using log(Nobel laureates per capita) and GDP per capita to predict Nobel laureates per capita was better than using chocolate consumption, and Nobel laureates, to do the same. Here the correlation value between both variables is while before it was . Furthermore, the -value associated with GDP\_cap is less than and so the null hypothesis, namely that , can be rejected and so the coefficient is statistically significant.

This is the regression line plotted on the scatterplot.

# Plot the regression line on the scatterplot  
ggplot(df, aes(x = GDP\_cap, y = log(nobel\_rate))) +   
 geom\_point(color = "darkgreen") +   
 geom\_abline(intercept = summary(model2)$coefficients[1,1],  
 slope = summary(model2)$coefficients[2,1]) +   
 ggtitle("Log of Nobel Laureates per Capita vs. GDP per Capita") +   
 theme\_minimal()



1. The log- is a multiplicative model: is For such a model, the slope is interpreted as follows: a unit increase in changes by approximately . For the above regression model, interpret the slope as so. Include units of measurement.

# Calculate the slope interpretation  
(exp(summary(model2)$coefficients[2,1]) - 1) \* 100

## [1] 0.01607463

A one unit increase in the GDP per capita, in currency per person is associated with a increase in the number of Nobel laureates per 10 million population.

### Question 12: Does increasing chocolate consumption cause an increase in the number of Nobel Laureates? Justify your answer.

Answer: Increasing chocolate consumption does not necessarily *cause* an increase in the number of Nobel Laureates. In the investigation here, a strong positive correlation between the two was found, with a correlation value of . However, just because there is a correlation does not imply there is causation. To make this statement, a controlled experiment/study would need to be done.