SDGB 7844 HW 1: Chocolate & Nobel Prizes

Jiayin Hu

2019/2/7

Before the questions and answers, we list the R packages we need in this homework report.

```
#setting the packages needed
require(tidyverse)
require(knitr)
require(gridExtra)
require(broom)
```

Q1.

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.

Messerli used the variable "number of Nobel laureates per capita" to measure overall national cognitive function. I do not think it is reasonable, because this variable could only reflect the population with superior cognitive function but not show the average cognitive function in each country.

Q2.

Are countries without Nobel prize recipients included in Messerli's study? If not, what types of bias(es) would that introduce?

Coutries without Nobel prize recipients are not included in Messerli's study. This introduces selection bias that some counties which consume certain amount of chocolate without Nobel prize recipients are not included. Such example does not shows positive correlation between chocolate consumption and cognitive function.

Q3.

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?

These two variables are not on the same temporal scale. The number of Nobel laureates per capita considers the Nobel prizes awarded through 2011. The periods of chocolate consuming data are different from countries, but all the data is from the recent 20 years. We cannot make sure whether people who lived before 20 century ate chocolate or not. Maybe they didn't eat chocolate but also got many Nobel prizes.

Q4.

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.

```
# Show the data
data <- read_delim("nobel_chocolate.txt", col_names = TRUE, delim = ",")
data
```

```
## # A tibble: 23 x 24
##    country nobel_rate chocolate chocmrkt     GDP GDP_cap totalpop literate
##    <chr>        <dbl>     <dbl>    <dbl>   <dbl>   <dbl>    <int>    <dbl>
##  1 Austra…       5.45       4.5    2514. 7.83e11  35053.   2.23e7     99.9
##  2 Austria      24.3       10.2     962. 3.04e11  36119.   8.42e6     99.9
##  3 Belgium       8.62       4.4    1006  3.65e11  33020.   1.10e7     99.9
##  4 Brazil        0.05       2.9    6214. 2.02e12  10264.   1.97e8     90.5
##  5 Canada        6.12       3.9    2647. 1.23e12  35739.   3.45e7     99.8
##  6 China         0.06       0.7    1864. 9.97e12   7418.   1.34e9     95.4
##  7 Denmark      25.3        8.5     592. 1.82e11  32602.   5.57e6     99.9
##  8 Finland       7.6        7.3     601. 1.73e11  32031.   5.39e6     99.9
##  9 France        8.99       6.3    5624. 1.96e12  29963.   6.54e7    100
## 10 Germany      12.7       11.6    8068. 2.83e12  34573.   8.18e7     99.9
## # ... with 13 more rows, and 16 more variables: lifeexp <dbl>,
## #   patentR <int>, patentNR <int>, articles <dbl>, internet <dbl>,
## #   mobile <dbl>, phone <dbl>, fh_cl <int>, fh_fotpsc5 <int>, fh_pr <int>,
## #   ti_cpi <dbl>, undp_gii <dbl>, undp_hdi <dbl>, wbgi_rle <dbl>,
## #   wbgi_vae <dbl>, wef_gend <dbl>
```

```
# Create summaary table
summary.data <- data[ , c(1, 2, 3, 6)]
my.summary <- do.call(data.frame,
                list(mean = apply(summary.data[ , 2:4], 2, mean),
                     sd = apply(summary.data[ , 2:4], 2, sd),
                     median = apply(summary.data[ , 2:4], 2, median),
                     min = apply(summary.data[ , 2:4], 2, min),
                     max = apply(summary.data[ , 2:4], 2, max)))
summary.table <- data.frame(t(my.summary))
names(summary.table) <- c("Nobel Rate (#/10M Population)", "Chocolate
Consumption (kg/yr/capita)", "GDP ($/capita)")

kable(summary.table, caption = "Summary Statistics of Each Variable", align =
"c")
```

Summary Statistics of Each Variable

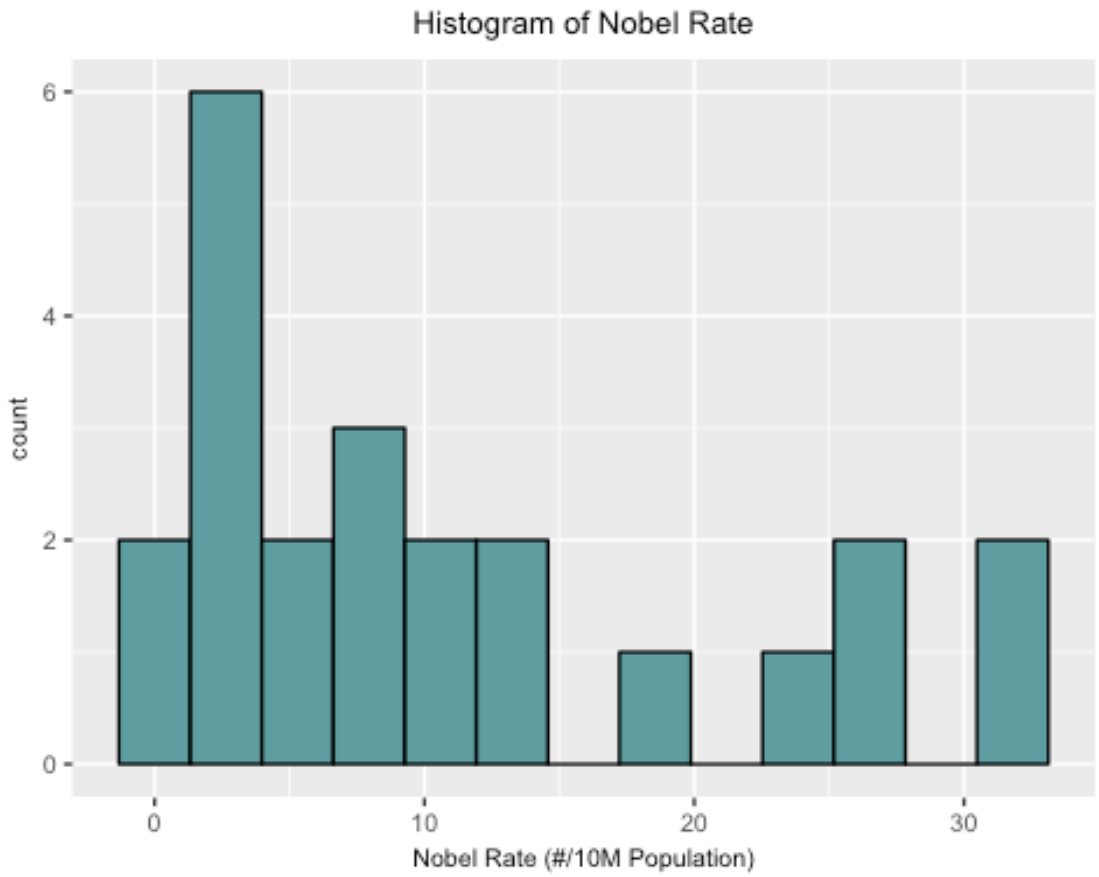
	Nobel Rate (#/10M Population)	Chocolate Consumption (kg/yr/capita)	GDP ($/capita)
mean	11.08878	5.804348	30592.143

sd	10.21818	3.279201	9467.658
median	8.62200	4.500000	32880.582
min	0.05000	0.700000	7417.888
max	31.85500	11.900000	46733.359

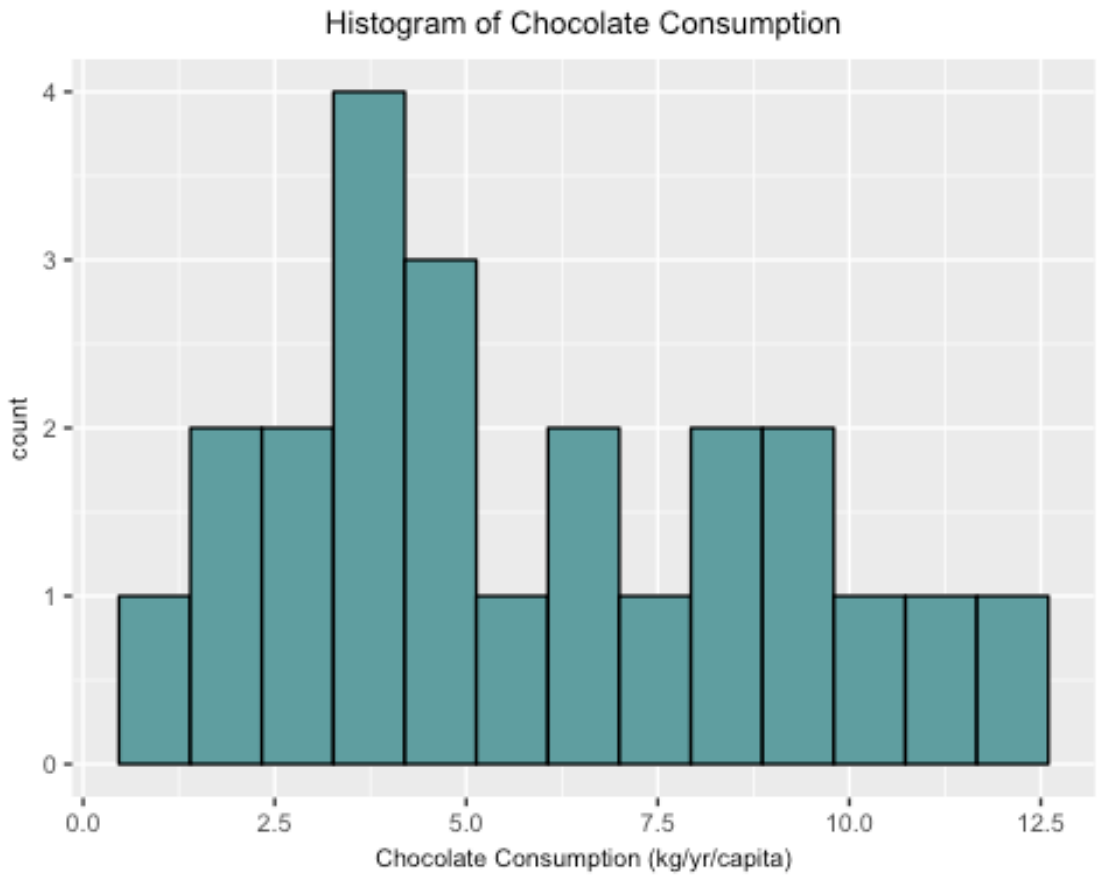
Q5.

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

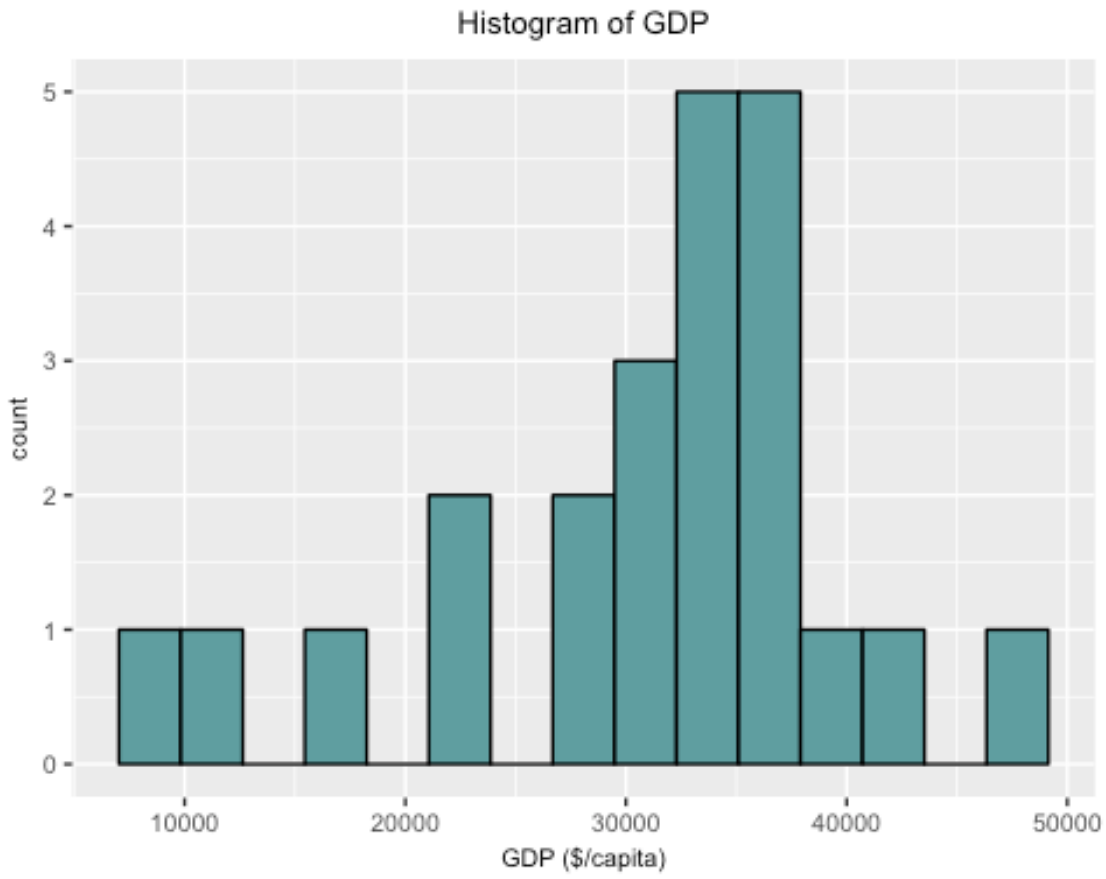
```
theme.info <- theme(plot.title = element_text(size = 10, hjust = 0.5),
                    axis.title = element_text(size = 8),
                    axis.text = element_text(size = 8))

nobel_hist <- summary.data %>% ggplot(aes(nobel_rate)) +
  geom_histogram(bins = 13, col = "black", fill = "cadetblue") +
  ggtitle("Histogram of Nobel Rate") +
  labs(x = "Nobel Rate (#/10M Population)") +
  theme.info

choco_hist <- summary.data %>% ggplot(aes(chocolate)) +
  geom_histogram(bins = 13, col = "black", fill = "cadetblue") +
  ggtitle("Histogram of Chocolate Consumption") +
  labs(x = "Chocolate Consumption (kg/yr/capita)") +
  theme.info

gdp_hist <- summary.data %>% ggplot(aes(GDP_cap)) +
  geom_histogram(bins = 15, col = "black", fill = "cadetblue") +
  ggtitle("Histogram of GDP") +
  labs(x = "GDP ($/capita)") +
  theme.info

nobel_hist
```


Histogram of Nobel Rate
count
6
4
2
0
0
10
20
30
Nobel Rate (#/10M Population)

```
choco_hist
```


Histogram of Chocolate Consumption
count
0
1
2
3
4
0.0
2.5
5.0
7.5
10.0
12.5
Chocolate Consumption (kg/yr/capita)

```
gdp_hist
```



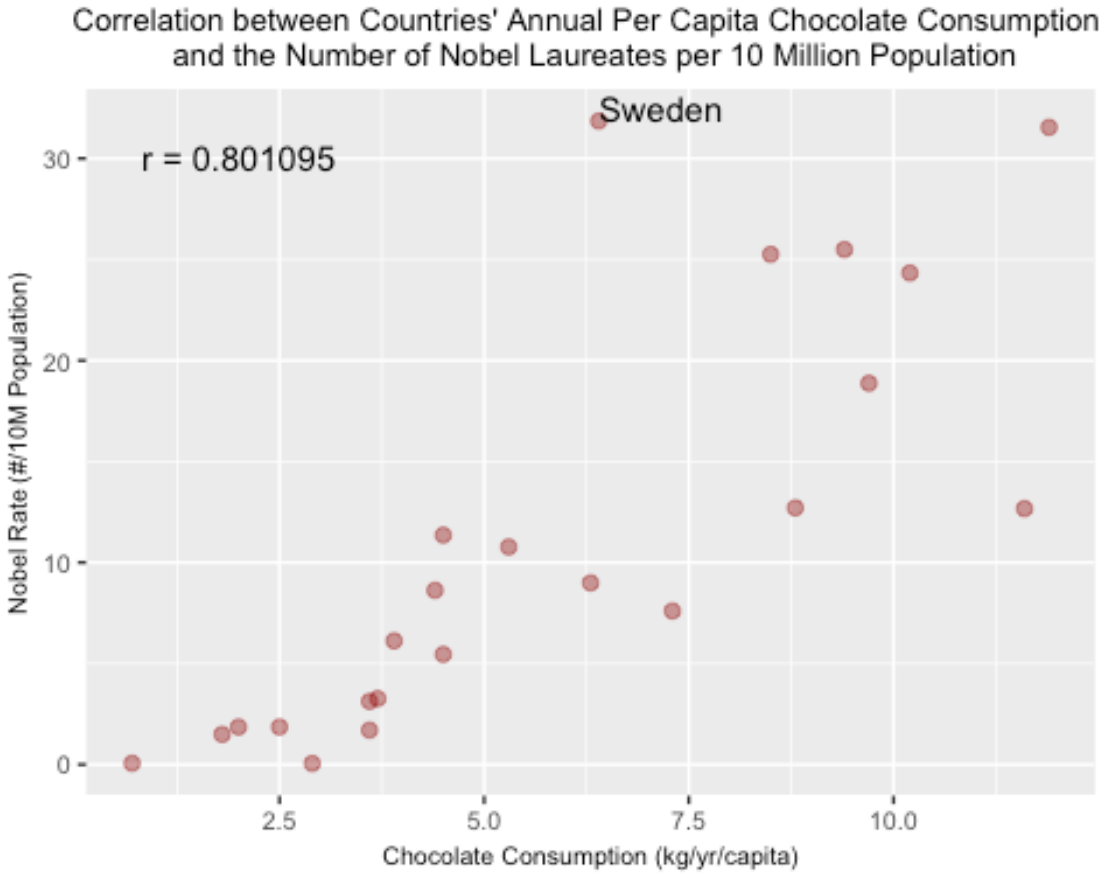
The histograms of Nobel rate and chocolate consumption are right skewed. The histogram of GDP is left skewed.

Q6.

Construct a scatterplot of Nobel laureates per capita vs. chocolate consumption. Label Sweden on your plot (on the computer, not by hand). 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?

```
s.1 <- summary.data %>% ggplot(aes(x = chocolate, y = nobel_rate,
label=country)) + geom_point(color = "#99000070", size = 2) +
  ggtitle("Correlation between Countries' Annual Per Capita Chocolate
Consumption \n and the Number of Nobel Laureates per 10 Million Population")
+
  labs(x = "Chocolate Consumption (kg/yr/capita)", y = "Nobel Rate (#/10M
Population)") +
  geom_text(aes(label = ifelse(country == "Sweden",
as.character(country),'')), hjust = 0, vjust=0) +
  theme.info

corr <- cor(summary.data$chocolate, summary.data$nobel_rate)
```

```
s.2 <- s.1 + annotate("text", x = 2, y = 30, label = paste("r = ",
as.character(eval(round(corr, 6))), sep = ""))

s.2
```



The correlation 0.801095 shows a strong positive linear relationship between chocolate consumption and Nobel prizes rate. Here, we consider correlation is appropriate because from the scatterplot we can observe that these two variables seem like to be described by a line.

Q7.

What is Messerli's correlation value? (Use the correlation value that includes Sweden.) Why is your correlation different?

Messerli's correlation value is 0.791. We use different dataset. The Nobel prize information we use includes 2012, but Messerli doesn't include it. And the sources of chocolate consumption are also different. These may cause the diversity.

Q8.

Why does Messerli consider Sweden an outlier? How does he explain it?

Because given per caita chocolate consumption of 6.4kg per year, we would predict Sweden should have prodeced a total of 14 Nobel laureates, yet 32 Nobel prizes are observed. Messerli explains this as the Nobel Commitee has some patriotic bias when assessing the candidates, or Swedes are very sensitive to chocolate so that minuscule amounts greatly enhance their cognition.

Q9.

Regress Nobel laureates per capita against chocolate consumption (include Sweden):

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

(b) Interpret the slope.

(c) 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.

(d) Is the slope significant (conduct a hypothesis test and include your regression output in your answer)? Test at the $\alpha = 0.05$ level and remember to specify the hypotheses you are testing.

(e) Add the regression line to your scatterplot.

```
lm.y <- lm(nobel_rate ~ chocolate, data = summary.data)
summary(lm.y)
```

```
##
## Call:
## lm(formula = nobel_rate ~ chocolate, data = summary.data)
##
## 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
```

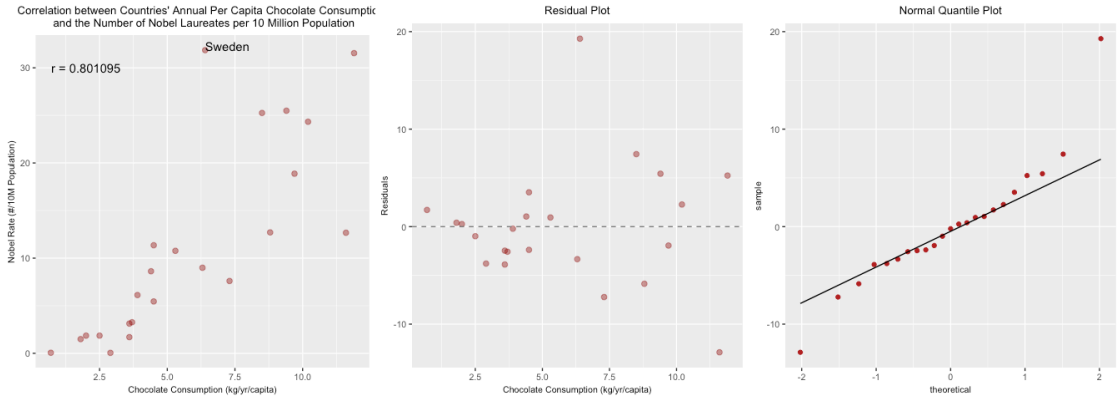
(a)The regression equation is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where $\hat{\beta}_0 = -3.400$(#/10M Population), $\hat{\beta}_1 = 2.496$ ((#/10M Population)/(kg/yr/capita)).

(b)The slope $\hat{\beta}_1$ means that 0.4 kg increase in chocolate consumption per capita per year is associated with 1 increasement in the number of Nobel prize per 10 million population in a given country.

```
library(egg)
y.augment <- augment(lm.y)
r.1 <- y.augment %>% ggplot(aes(x = chocolate, y=.resid)) +
  geom_point(color = "#99000070", size = 2) +
  ggtitle("Residual Plot") +
  labs(x = "Chocolate Consumption (kg/yr/capita)", y = "Residuals") +
  geom_hline(yintercept = 0, color = "gray50",  lty = 2) +
  theme.info


qq.1 <- y.augment %>%  ggplot(aes(sample=.resid)) +
  stat_qq(col = "firebrick") +
  stat_qq_line() +
  ggtitle("Normal Quantile Plot") +
  theme.info

grid.arrange(s.2, r.1, qq.1, ncol=3)
```



(c)For this model, data are all collected from valid sources so that it is convinced that x measurement is accurate and x and y are independent. But x variable is not fixed in this case. The residual plot shows a little bit heteroscedasticity but no obvious curved shape so we can consider x and y are linear relationship. From the normal quantile plot, the residuals shows a little bit fat-tail distribution rather than normal distribution. According to the analysis above, several assumptions are violated, so we could not conduct hypothesis tests for this model.

```
t <- lm.y$coef[2]/summary(lm.y)$coefficients[2,2]
p_value <- 2 * pt(t, df=nrow(data)-2, lower.tail=FALSE)
t

## chocolate
##  6.133414

p_value

##    chocolate
## 4.374029e-06
```

Set confidence level $\alpha = 0.05$. We have the two-sided hypothesis:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

The t-statitic is

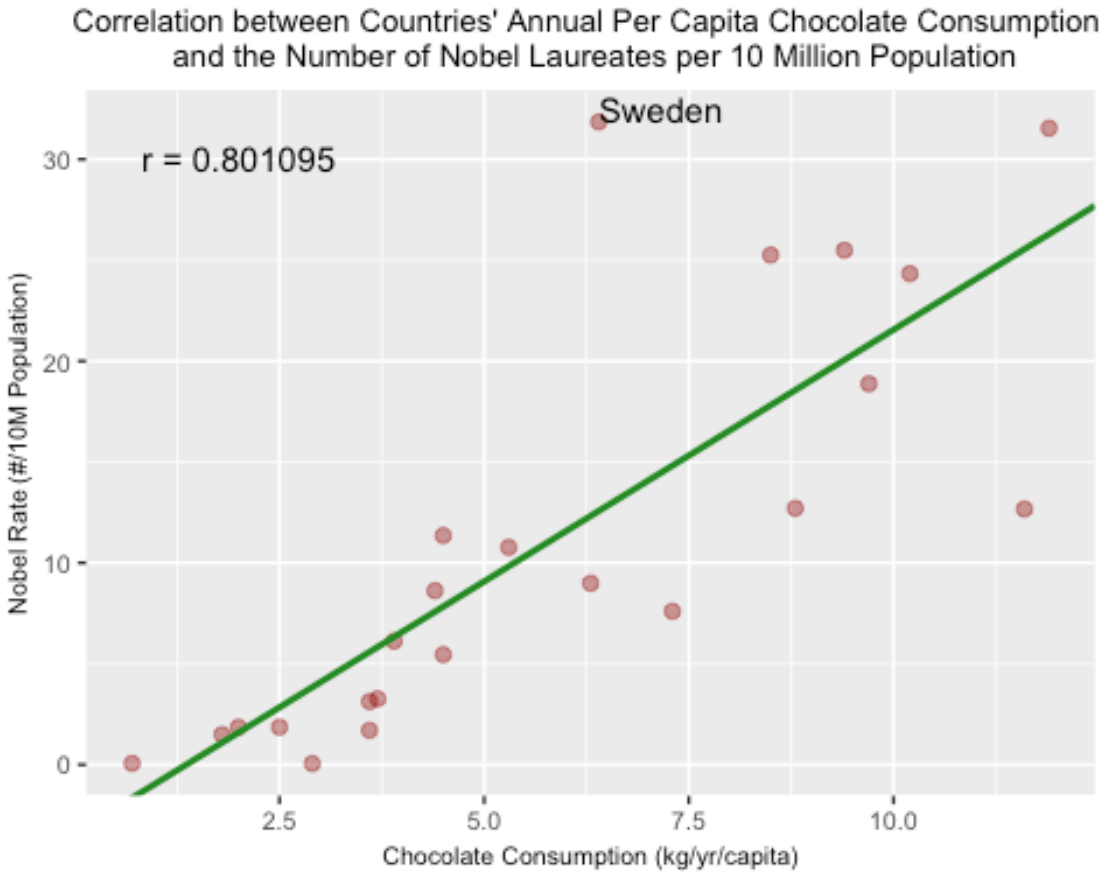
$$t = \frac{\hat{\beta}_1 - \beta_1^0}{\mathrm{SE}(\hat{\beta}_1)} = \frac{2.496 - 0}{0.407} = 6.133414$$

Then, calculate the p-value under t-distribution with $df = 23 - 2 = 21$.

$$p - \text{value} = 2\mathbb{P}(t > 6.133414) = 4.374029e - 06 < 0.05$$

Hence, we reject null hypothesis, which means the slope is significantly different from 0.

```
s.3 <- s.2 +
  geom_abline(intercept = coef(lm.y)[1] , slope=coef(lm.y)[2], color =
"forestgreen", size=1) +
  theme.info

s.3
```



Q10.

Using your model, what is the number of Nobel laureates expected to be for Sweden? What is the residual? (Remember to include units of measurement.)

```
lm.y$fitted.values[20]

##       20
## 12.57568

y.augment$.resid[20]

## [1] 19.27932
```

By using the model, the number of Nobel laureates expected to be for Sweden is 12.57568 per 10 million population. The residual is 19.27932 per 10 million population.

Q11.

Now we will see if the variable GDP per capita (i.e., "GDP cap") is a better way to predict Nobel laureates.

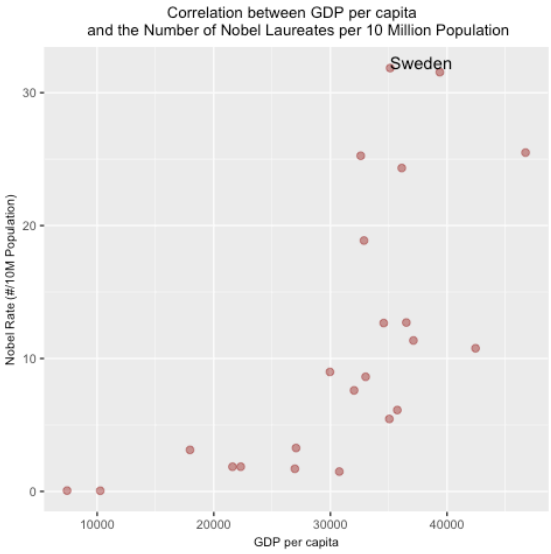
(a) In one figure construct a scatter plot of (i) Nobel laureates vs. GDP per capita and (ii) log(Nobel laureates) vs. GDP per capita. Which plot is more linear? Label Sweden on both plots. On the second plot, label the two countries which appear on the bottom left corner.

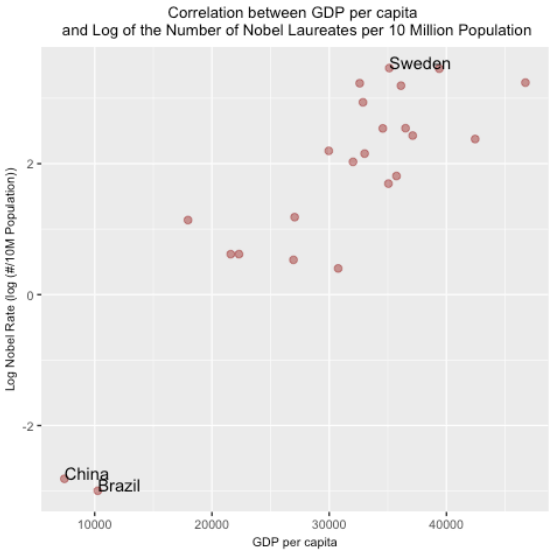
(b) Is Sweden still an outlier? Justify your answer.

(c) Regress Nobel laureates against log(GDP per capita). Provide the output and add the regression line to your scatterplot. (In practice, we would do a residual analysis here, but we will skip it to reduce the length of this assignment.)

(d) The log-y model is a multiplicative model: $log(y) = \beta_0 + \beta_1$ is $\hat{y} = e^{\beta_0 + \beta_1 x}$. For such a model, the slope is interpreted as follows: a unit increase in x changes y by approximately $(e^{\beta_1} - 1) \times 100\%$. For your regression, model interpret the slope (remember to include units of measurement).

```
s.4 <- summary.data %>% ggplot(aes(x = GDP_cap, y = nobel_rate,
label=country)) + geom_point(color = "#99000070", size = 2) +
  ggtitle("Correlation between GDP per capita  \n and the Number of Nobel
Laureates per 10 Million Population") +
  labs(x = "GDP per capita", y = "Nobel Rate (#/10M Population)") +
  geom_text(aes(label = ifelse(country == "Sweden",
as.character(country),'')), hjust = 0, vjust=0) +
  theme.info
s.5 <- summary.data %>% ggplot(aes(x = GDP_cap, y = log(nobel_rate),
label=country)) + geom_point(color = "#99000070", size = 2) +
  ggtitle("Correlation between GDP per capita \n and Log of the Number of
Nobel Laureates per 10 Million Population") +
  labs(x = "GDP per capita", y = "Log Nobel Rate (log (#/10M Population))") +
  geom_text(aes(label = ifelse(country == "Sweden",
as.character(country),'')), hjust = 0, vjust=0) +
  theme.info +
  geom_text(aes(label = ifelse(log(nobel_rate) < -2,
as.character(country),'')),hjust=0,vjust=0)
plot = grid.arrange(s.4, s.5, ncol = 2, nrow = 1)
```

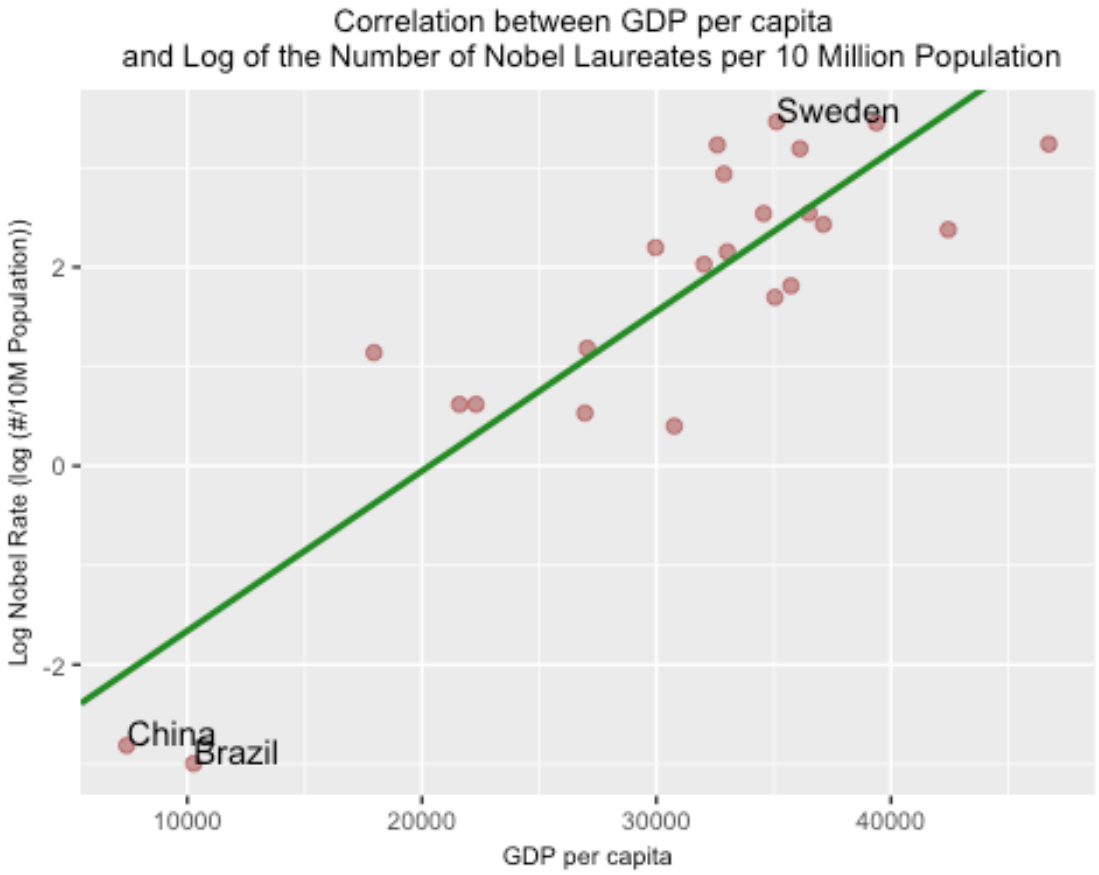




The second plot shows more linear.

(b) Sweden is not an outlier, because the point of Sweden is not so far from the points with near x-value.

```
lm.y1 <- lm(log(nobel_rate) ~ GDP_cap, data = summary.data)
summary(lm.y1)

##
## Call:
## lm(formula = log(nobel_rate) ~ GDP_cap, data = summary.data)
##
## 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

s.6 <- s.5 +
  geom_abline(intercept = coef(lm.y1)[1] , slope=coef(lm.y1)[2], color =
"forestgreen", size=1) +
  theme.info

s.6
```



(d)The regression model is

$$\hat{y}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}$$

where $\hat{\beta}_0 = -3.267e + 00$ (log(#/10M Population)), $\hat{\beta}_1 = 1.607e - 04$ (log(#/10M Population)/($/capita)). The interpretation for the slope is 10,000 dollar increase in GDP per capita is associated with the increase of the number of Nobel prize per 10 million population in a given country by approximately 60.7%.

Q12.

Does increasing chocolate consumption cause an increase in the number of Nobel Laureates? Justify your answer. Chocolate consumption does not cause increase in the number of Nobel Laureates. First, in this analysis, we have already pointed out that the data we use is not reasonable enough to show the relationship between these two variables. What's more, the regression model only shows the linear relationship but does not explain the causal relationship.