Homework 1

Yuwen(Suyi)Wu 5:45-8:00

02/02/2019

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.

From the article posted, the variable "number of Nobel laureates per capita" supposed to represent surrogate end point reflecting the proportion with superior cognitive function and thereby give us some measure of the overall cognitive function of a given country.

I don't think it is a reasonalbe measure because the variable "number of Nobel laureates per capita" could be related to the education level of whole country and its only represent to the top outstanding people of that country not for the whole population.

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

```
# import data
N_C<- read_delim("nobel_chocolate.txt", col_names=TRUE,delim=',')
nobel_prize <- select(N_C,country,nobel_rate)
# filter data without Nobel Prize
filter(nobel_prize,round(nobel_prize$nobel_rate,digits = 0)==0)
```

```
## # A tibble: 2 x 2
##   country nobel_rate
##   <chr>        <dbl>
## 1 Brazil        0.05
## 2 China         0.06
```

```
nrow(filter(nobel_prize,round(nobel_prize$nobel_rate, digits = 0)==0))
```

```
## [1] 2
```

the China and the Brazil have no Nobel prize recipients in study. In this case, the result would be with sample selection bias.

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?

The number of Nobel laureates per capita and chocolate consumption per capita not measured on the same scale.

From the article, all Nobel Prizes that were awarded through October 10, 2011, were included. Data on per capita yearly chocolate consumption in 22 countries was obtained from the year of 2011 for 1 country (Switzerland), from 2010 for 15 countries, from 2004 for 5 countries, and from 2002 for 1 country (China).

The temporal scale difference between number of Nobel laureates and chocolate consumption would affect the accuracy of regression and the model parameter during the model fitting.

The temporal scale for Nobel laureates is measured by a period of time(yearly), therefore the temperal scale for chololate consumption should also be measured by the same period of time(yearly). And the chocolate consumption of the countries that do not in the year same to the temporal scale for Nobel laureates should be changed to the same temporal scale of 2011.

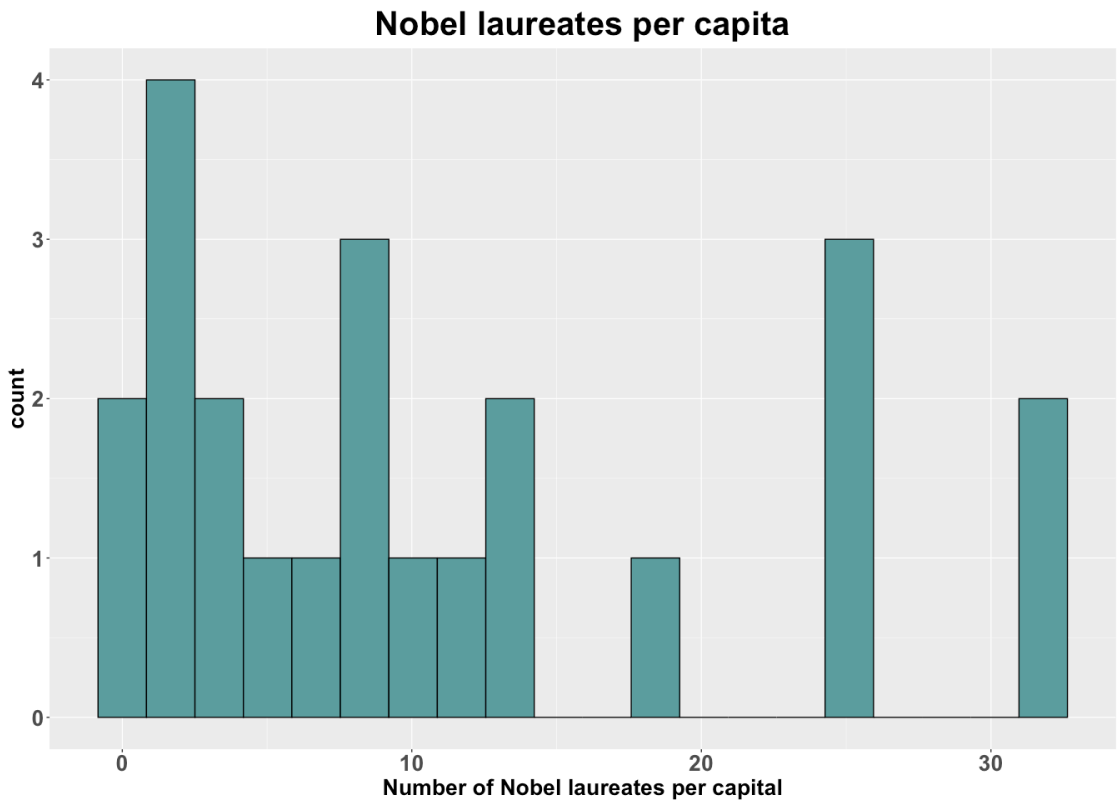
4. Create a table of summary statistics for the following variables: Nobel laureates per capita, GDP per capita, and chocolate consumption. Include the statistics: minimum, maximum, median, mean, and standard deviation. Remember to include the units of measurement in your table.

```
sum_stat <- N_C %>%
select(nobel_rate,GDP_cap,chocolate)%>%summarise_all(c("min","max","median","
mean","sd"))
sum_stat_t <- matrix(sum_stat,nrow = 3,ncol = 5)

colnames(sum_stat_t) <- c("min","max","median","mean","sd")
rownames(sum_stat_t) <- c("Nobel laureates per capita(people/10m)","GDP per
capita(dollar/capita)", "chocolate consumption(kg/yr/capita)")
sum_stat_t
```

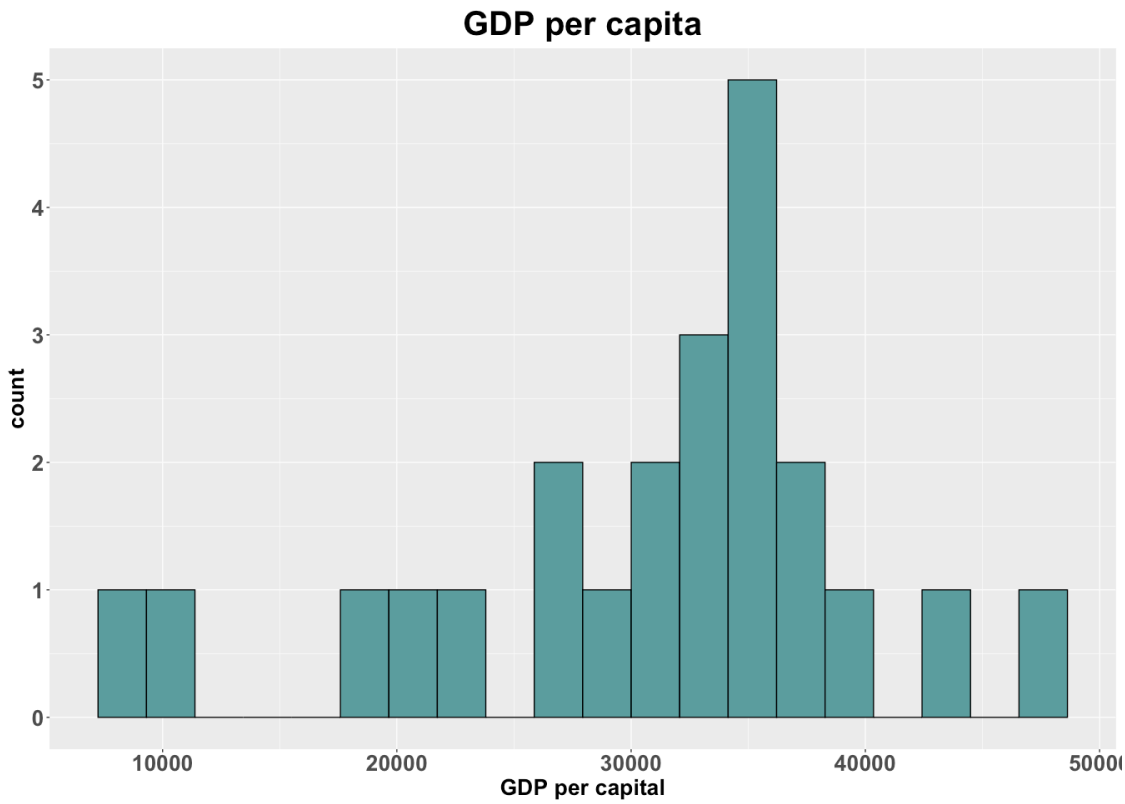
```
##                                        min      max      median   mean
## Nobel laureates per capita(people/10m) 0.05     31.855   8.622    11.08878
## GDP per capita(dollar/capita)          7417.888 46733.36 32880.58 30592.14
## chocolate consumption(kg/yr/capita)    0.7      11.9     4.5      5.804348
##                                        sd
## Nobel laureates per capita(people/10m) 10.21818
## GDP per capita(dollar/capita)          9467.658
## chocolate consumption(kg/yr/capita)    3.279201
```

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

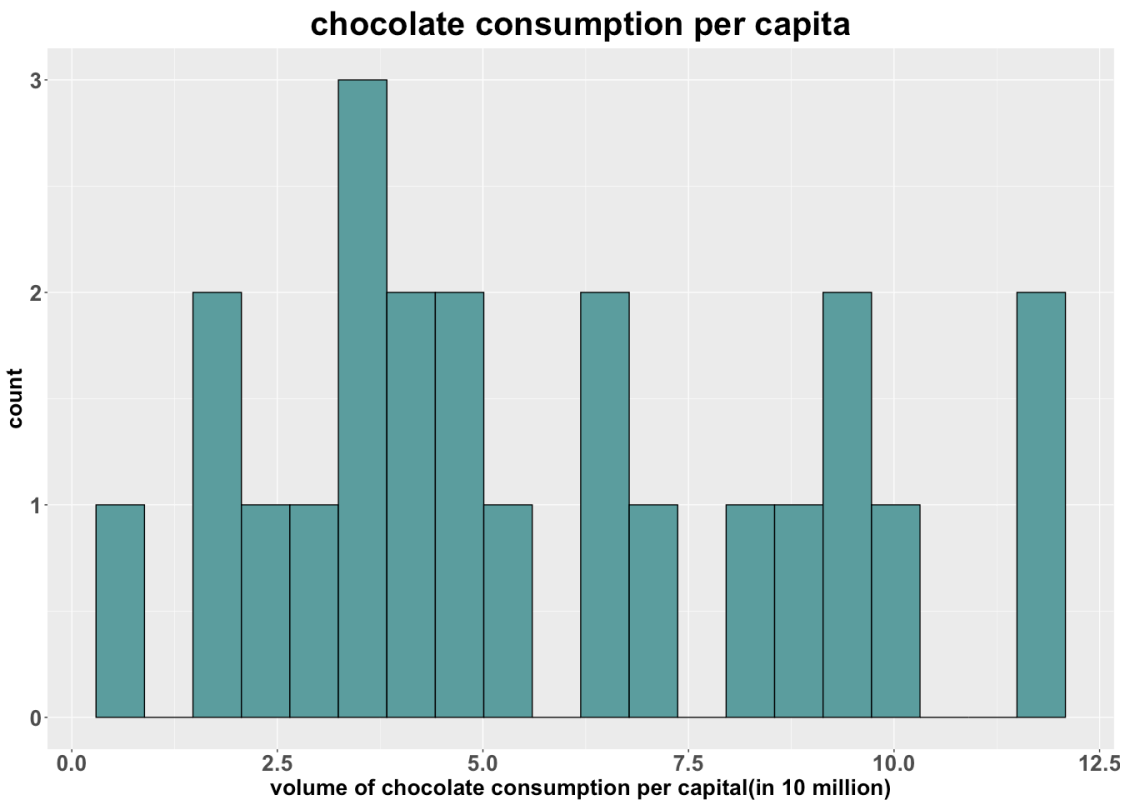
```
theme.info <- theme(plot.title = element_text(size=30, face = "bold",
hjust=0.5),
                    axis.title = element_text(size=20,face = "bold"),
                    axis.text=element_text(size=20, face = "bold"))

h.nobel <- N_C %>% ggplot(aes(nobel_rate)) + geom_histogram(bins=20,
col="black", fill="cadetblue") +
  ggtitle("Nobel laureates per capita") +
  labs(x="Number of Nobel laureates per capital") +
  theme.info
h.nobel
```




```
# The shape of Nobel laureates per capita is the right skewd distribution

h.GDP <- N_C %>% ggplot(aes(GDP_cap)) + geom_histogram(bins=20, col="black",
fill="cadetblue") +
  ggtitle("GDP per capita") +
  labs(x="GDP per capital") +
  theme.info
h.GDP
```



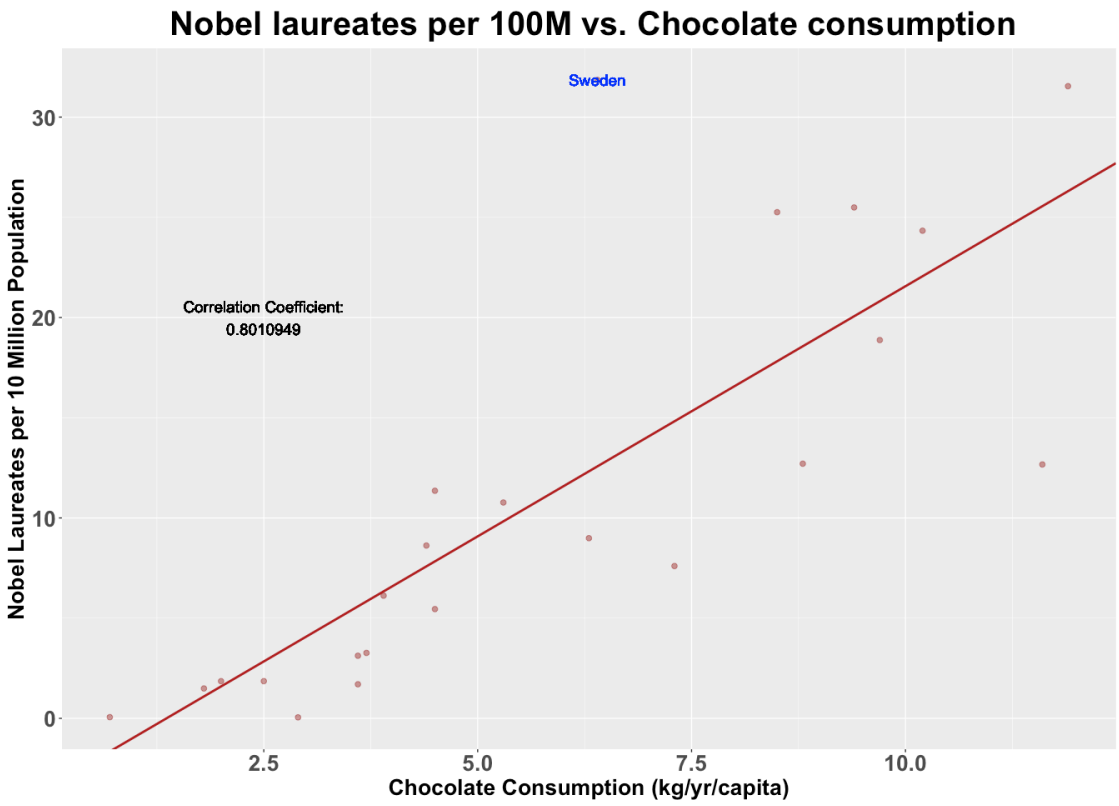

```
# The shape of GDP per capita is the left skewd distribution

h.chocolate <- N_C %>% ggplot(aes(chocolate)) + geom_histogram(bins=20,
col="black", fill="cadetblue") +
  ggtitle("chocolate consumption per capita") +
  labs(x="volume of chocolate consumption per capital(in 10 million)") +
  theme.info
h.chocolate
```




```
# The shape of chocolate consumption per capita is the right skewd
distribution
```

6. 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?

```
# complete regression model
lm.result <- lm(nobel_rate ~ chocolate, data = N_C)
# calculate correlation coefficient
cor_cn<-cor(N_C$nobel_rate,N_C$chocolate)
# plot the scatter point picture
s.NC <- N_C %>% ggplot(aes(x=chocolate, y=nobel_rate)) +
geom_point(color="#99000070", size=2) +
  ggtitle(" Nobel laureates per 100M vs. Chocolate consumption") +
  labs(x="Chocolate Consumption (kg/yr/capita)", y="Nobel Laureates per 10
Million Population") +
  geom_text(aes(label =
"Sweden",x=N_C$chocolate[N_C$country=="Sweden"],y=N_C$nobel_rate[N_C$country=
="Sweden"]),cex=5,col="blue")+
  geom_abline(intercept= lm.result$coefficients[1] ,
slope=lm.result$coefficients[2], color='firebrick', size=1)+
  geom_text(cex = 5, aes(label = "Correlation
Coefficient:\n0.8010949",x=2.5,y=20))+
  theme.info
s.NC
```




```
cor_cn

## [1] 0.8010949

lm.result

##
## Call:
## lm(formula = nobel_rate ~ chocolate, data = N_C)
##
## Coefficients:
## (Intercept)    chocolate
##      -3.400        2.496
```

The correlation between these Nobel laureates and chocolate consumption is 0.8010949, which is a strong positive correlation. This is a positive trend. The correlation is an appropriate measure, but the correlation doesn't imply causation. However, the correlation coefficient of 0.8 shows strong relation between this two variables.

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

Messerli's correlation value:0.791

My correlation value is 0.8010949

In Messerli's analysis, the number of Noble laureates is rounded to the integers. But in our model, we used the exact number with decimal to fit the model, which would provide more precise correlation coefficient. In this case, our corrlation coefficient would sllightly different from Messierli's. For instance, the Noble laureates of Sweden provided in article is 32 but in the dataset it is 31.855. Also, Our data is through Nov 27, 2012 whereas Messerli's analysis includes prize winners only through Oct. 1, 2011

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

Given Sweden's per capita chocolate consumption of 6.4kg per year, Messerli predict that Sweden should have produced a total of about 14 Nobel laureates, But Messerli obverved 32 Nobel laureates, which exceeds the expected number by a factor of more than 2. Therefore, he thought the Sweden should be an outlier. And he explained that this should be because the Nobel Committee in Stockholm has some inherent patriotic bias when assessing the candidates for this awards or Swedes are particularly sensitive to chocolate and even minuscule amounts greatly enhance their cognition.

9. 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.

ANSWER:

(a) Nobel Laureates(per 10 Million Population) = -3.400 (per 10 Million Population) + 2.496(per 10 Million Population)*Chocolate Consumption/(kg/yr/capita)

(b) Every additional kilogram increase per year per capita in chocolate consumption is associated with a 2.496 increase per 10 Million population in Nobel laureates.

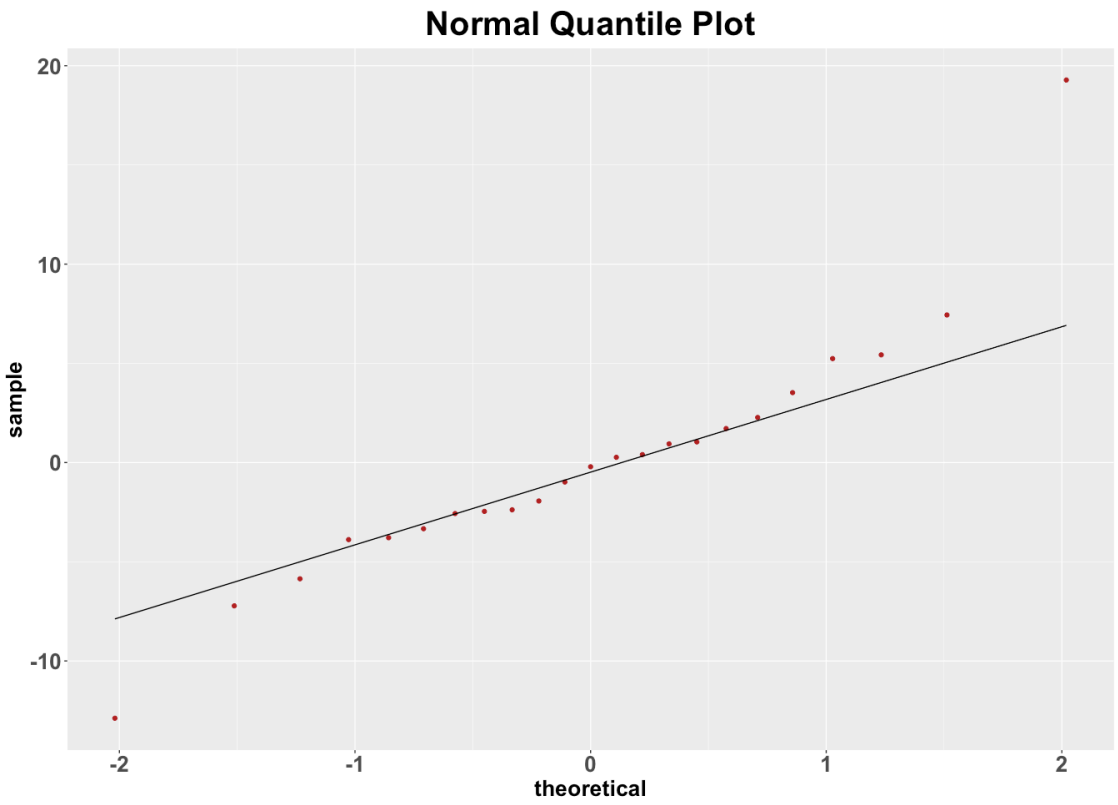
```
lm.result$residuals
```

```
##          1          2          3          4          5          6
## -2.3817942  2.2705364  1.0388316 -3.7887817 -0.2130395  1.7129855
##          7          8          9         10         11         12
##  7.4371746 -7.2223160 -3.3360582 -12.8882245 -0.9832786 -5.8607027
##         13         14         15         16         17         18
## -2.5707879  0.3991019  3.5232058  5.4275426 -2.4621622  0.2628503
##         19         20         21         22         23
## -3.8851622 19.2793160  5.2388981 -1.9383347  0.9401996
```

```
y.augment <- augment(lm.result)
y.augment
```

```
## # A tibble: 23 x 9
##    nobel_rate chocolate .fitted .se.fit  .resid   .hat .sigma .cooksd
##         <dbl>     <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>
##  1       5.45       4.5    7.83    1.41  -2.38  0.0507   6.39 4.07e-3
##  2      24.3       10.2   22.1     2.21   2.27  0.125    6.39 1.08e-2
##  3       8.62       4.4    7.58    1.42   1.04  0.0518   6.41 7.94e-4
##  4       0.05       2.9    3.84    1.76  -3.79  0.0791   6.35 1.71e-2
##  5       6.12       3.9    6.34    1.52  -0.213 0.0588   6.41 3.84e-5
##  6       0.06       0.7   -1.65    2.45   1.71  0.154    6.40 8.03e-3
##  7      25.3        8.5   17.8     1.71   7.44  0.0742   6.18 6.11e-2
##  8       7.6        7.3   14.8     1.44  -7.22  0.0529   6.20 3.93e-2
##  9       8.99       6.3   12.3     1.32  -3.34  0.0445   6.37 6.92e-3
## 10      12.7       11.6   25.6     2.70 -12.9   0.185    5.56 5.92e-1
## # … with 13 more rows, and 1 more variable: .std.resid <dbl>
```

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

p.residualqq
```

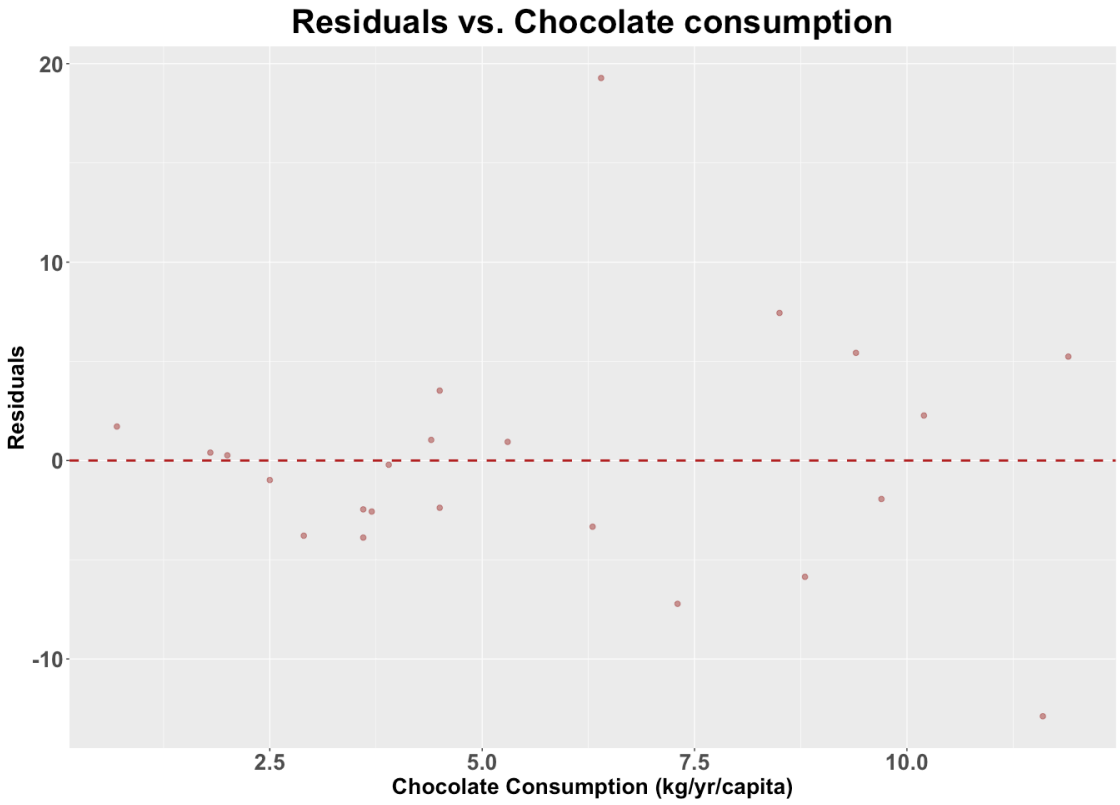


(c) If we want to use any hypothesis tests, etc. we must assume that:

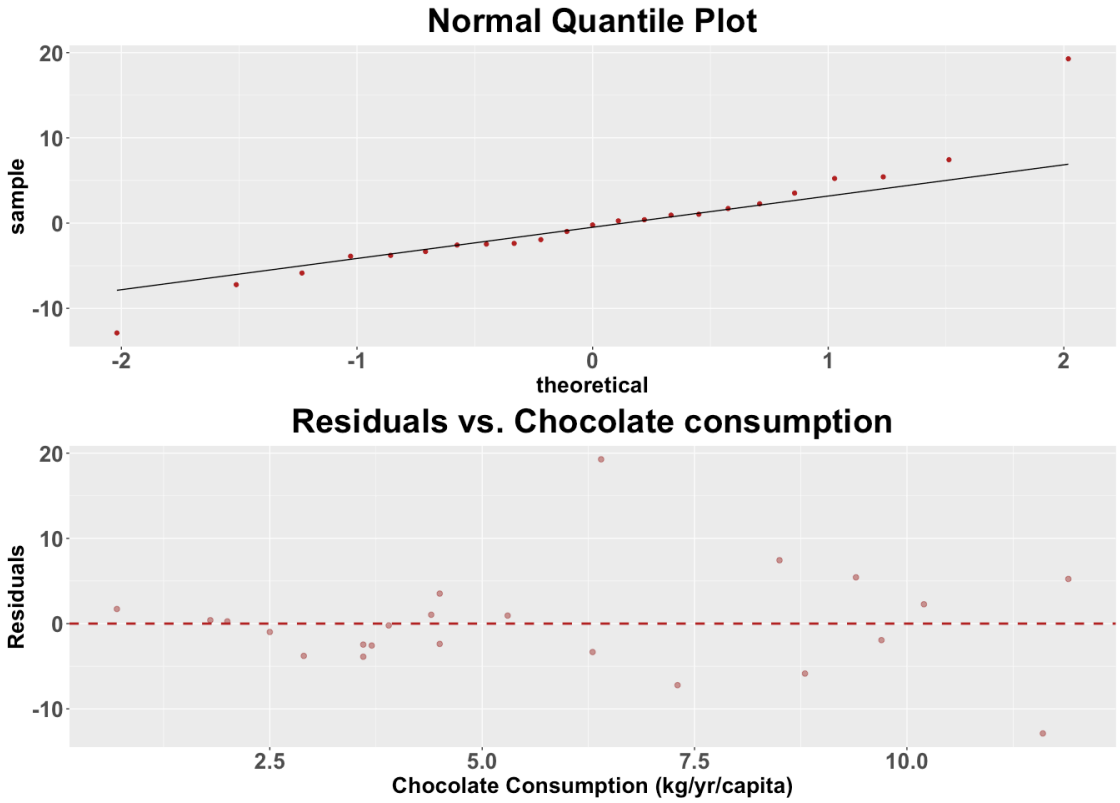
1. x values are fixed and are measured without error This assumption is not satisfied because the data are not in same temporal scale.
2. x and y are linearly related This assumption could be recognized satisfied based on above model and calculation.
3. errors are independent of each other This dataset is not the time series dataset, so we can assume the errors are independent of each other.
4. the points are evenly distributed above/below the regression line This could be check based on above scatter plot, so this assumption is satisfied.
5. errors have mean 0 The ordinary least squares method (OLS) ensures this
6. the errors are normally distributed

From the above qq plot, it could be seen that the residual with fat tail distribution. Therefore, this assumption does not satisfied.

```
## check for constant variance and normality ##
p.residual <- N_C %>% ggplot(aes(x = chocolate, y = lm.result$residuals),type
= "p",pch = 19) +
  geom_point(color="#99000070", size=2)+
  ggtitle("Residuals vs. Chocolate consumption") +
  labs(x="Chocolate Consumption (kg/yr/capita)", y="Residuals") +
  geom_hline(yintercept = 0, col="firebrick", lty=2, lwd=1)+
  theme.info
p.residual
```



```
grid.arrange(p.residualqq, p.residual)
```



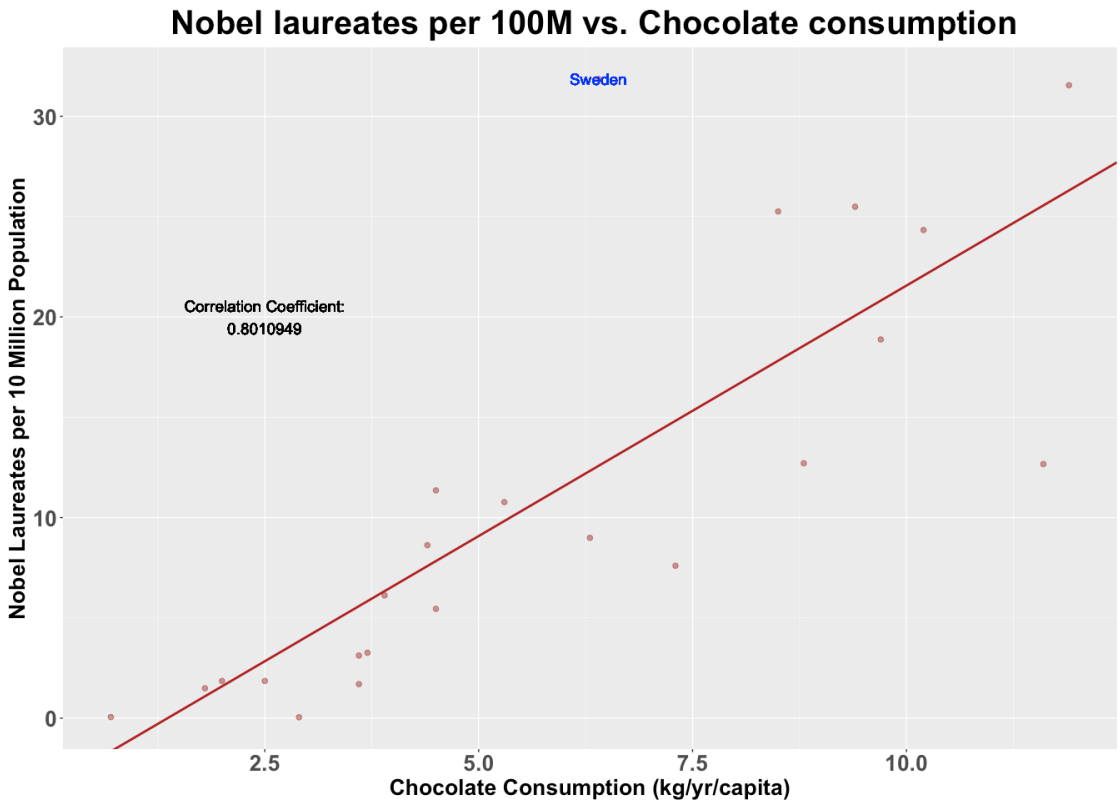
(d) The slope is significant. Hypothesis: H0 : $\beta 1 = 0$ H1 : $\beta 1 \neq 0$ ($\beta 1$ is the slope of chocolate consumption) $\alpha = 0.05$

increasing spread of points as x increases −→ constant variance assumption violated Therefore, we can not conduct the hypothesis tests for this regression model.

```
summary(lm.result)
```

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

# t = 6.133
# In this t test, the degrees of freedom is 21. And the P-value here is
4.37e-06, which is smaller than α = 0.05. Therefore, the β1 is significantly
different from zero and we can reject the H0 hypothesis.

s.NC
```



10. 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.)

```
## Nobel Laureates(per 10 Million Population) = -3.400 (per 10 Million
Population) + 2.496(per 10 Million Population)*Chocolate
Consumption/(kg/yr/capita)

Expected_Nobel_Sweden <- -3.400+2.496*N_C$chocolate[N_C$country=="Sweden"]
Expected_Nobel_Sweden

## [1] 12.5744

## The number of Nobel laureates expected to be for Sweden is 12.5744 per 10M
population.

Residual_Sweden <- N_C$nobel_rate[N_C$country =="Sweden"]-
```

```
Expected_Nobel_Sweden
Residual_Sweden

## [1] 19.2806

## The residual is 19.2806 per 10M population.
```

11. 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

(b) Nobel laureates vs. GDP per capita and

(ii) log(Nobel laureates) vs. GDP per capita.

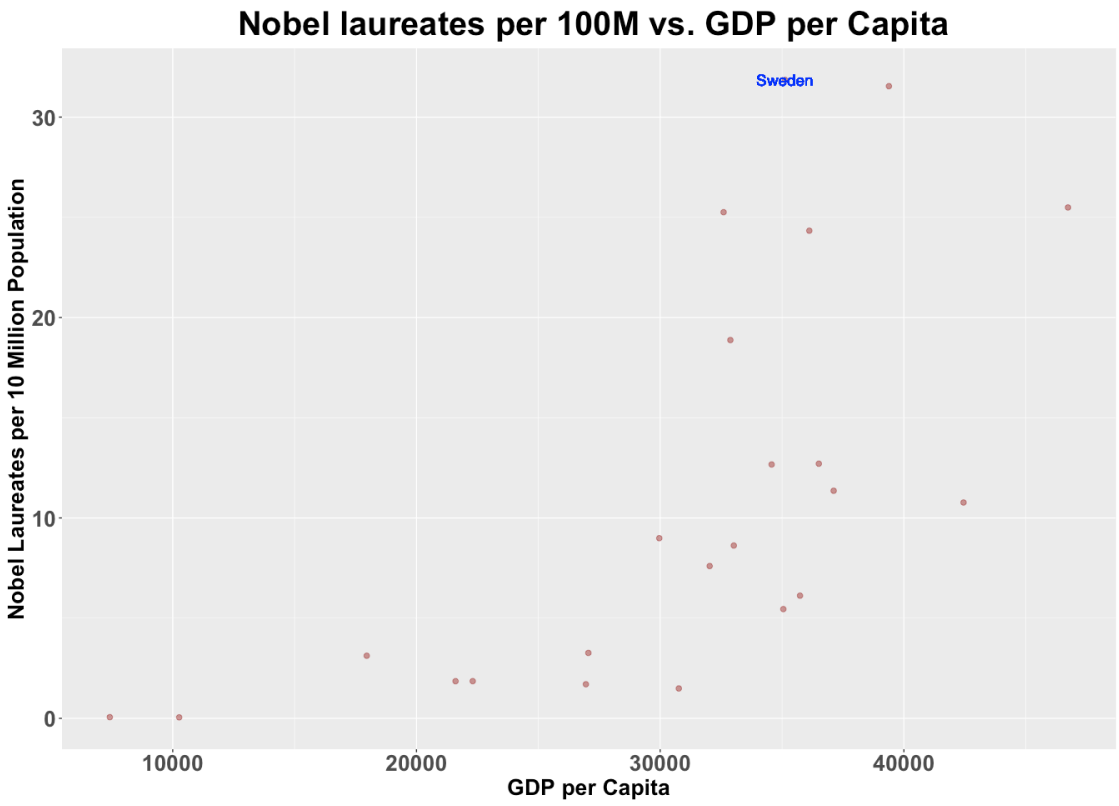
Which plot is more linear? Label Sweden on both plots. On the second plot, label the two countries which appear on the bottom left corner.

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

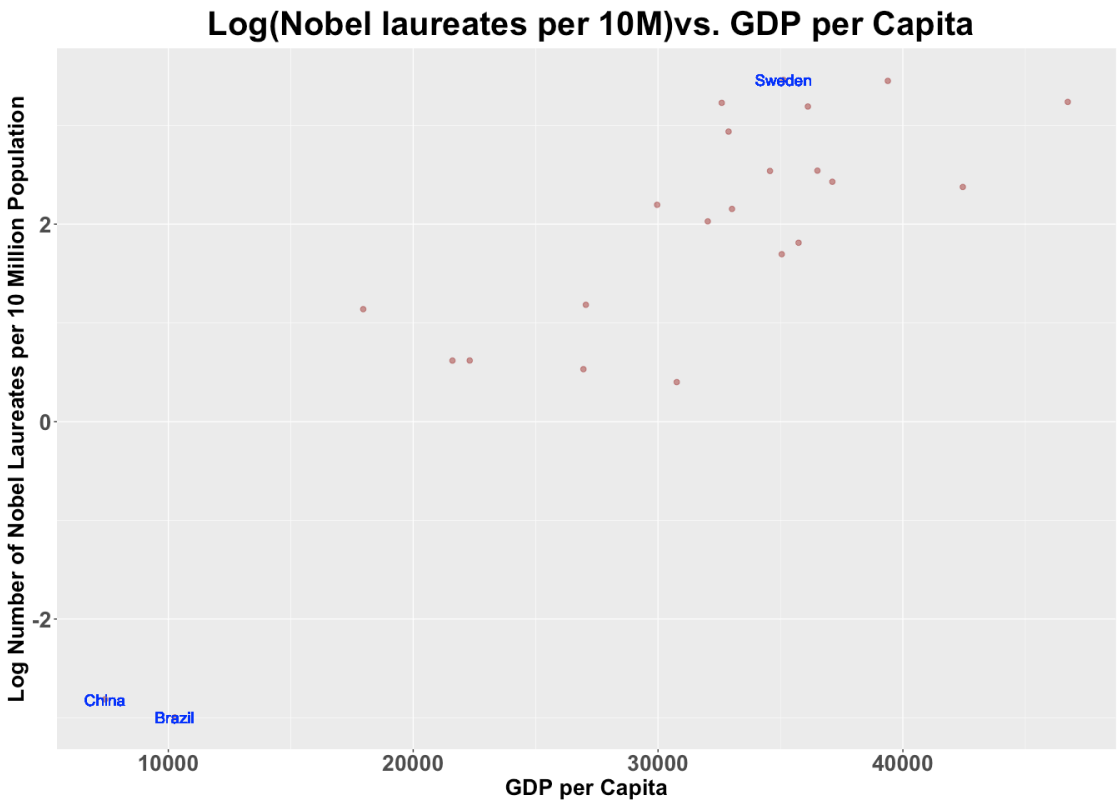
(c) Regress log(Nobel laureates) against GDP per capita. Provide the output and add the regression line to your scatterplot. (In practice, we would do a residual analysis here, but we will skip it to reduce the length of this assignment.)

(d) The log-y model is a multiplicative model: log(y) = β 0 + β 1 is y = e β 0 +β 1 x . For such a model, the slope is interpreted as follows: a unit increase in x changes y by approximately (e β 1 − 1) × 100. For your regression, model interpret the slope (remember to include units of measurement).

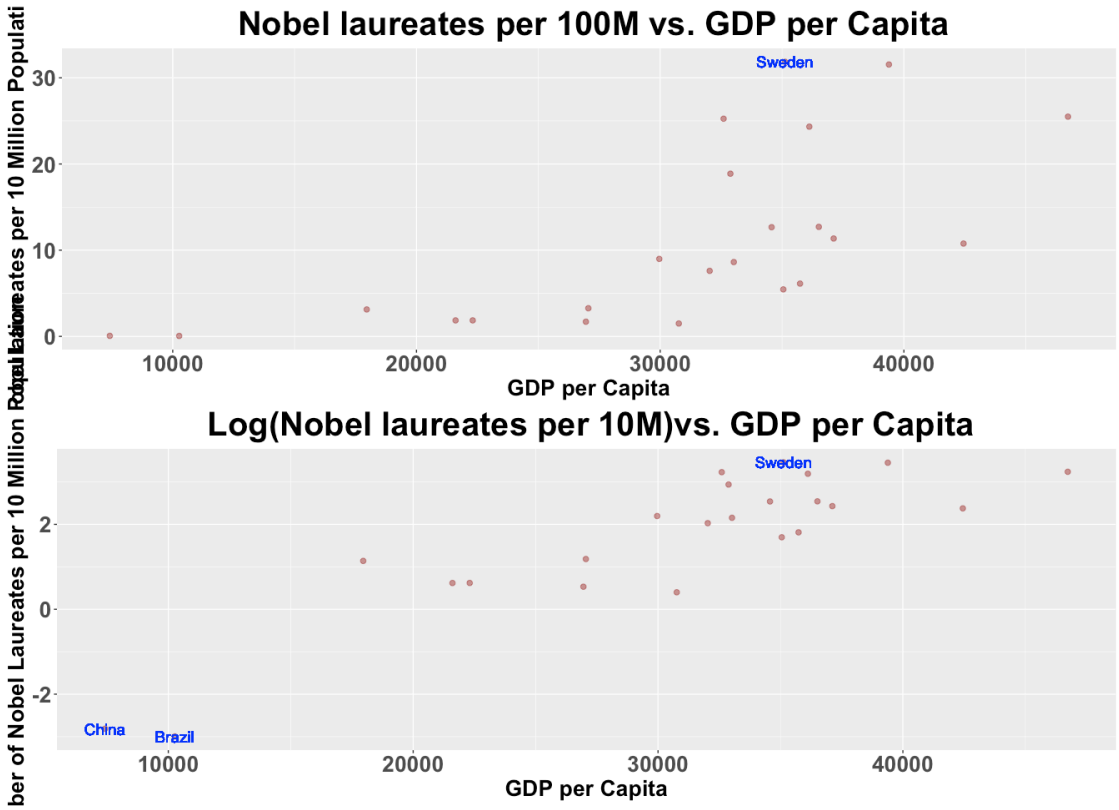
```
## (a) (i)
s.NG <- N_C %>% ggplot(aes(x=GDP_cap, y=nobel_rate)) +
geom_point(color="#99000070", size=2) +
  ggtitle(" Nobel laureates per 100M vs. GDP per Capita") +
  labs(x="GDP per Capita", y="Nobel Laureates per 10 Million Population") +
  geom_text(aes(label =
"Sweden",x=N_C$GDP_cap[N_C$country=="Sweden"],y=N_C$nobel_rate[N_C$country=="
Sweden"]),cex=5,col="blue")+
  theme.info
s.NG
```




```
## (a)(ii)
s.LOGNG <- N_C %>% ggplot(aes(x=GDP_cap, y=log(nobel_rate))) +
geom_point(color="#99000070", size=2) +
  ggtitle(" Log(Nobel laureates per 10M)vs. GDP per Capita") +
  labs(x="GDP per Capita", y="Log Number of Nobel Laureates per 10 Million
Population") +
  geom_text(aes(label =
"China",x=N_C$GDP_cap[N_C$country=="China"],y=log(N_C$nobel_rate[N_C$country=
="China"])),cex=5,col="blue")+
  geom_text(aes(label =
"Brazil",x=N_C$GDP_cap[N_C$country=="Brazil"],y=log(N_C$nobel_rate[N_C$countr
y=="Brazil"])),cex=5,col="blue")+
  geom_text(aes(label =
"Sweden",x=N_C$GDP_cap[N_C$country=="Sweden"],y=log(N_C$nobel_rate[N_C$countr
y=="Sweden"])),cex=5,col="blue")+
  theme.info
s.LOGNG
```



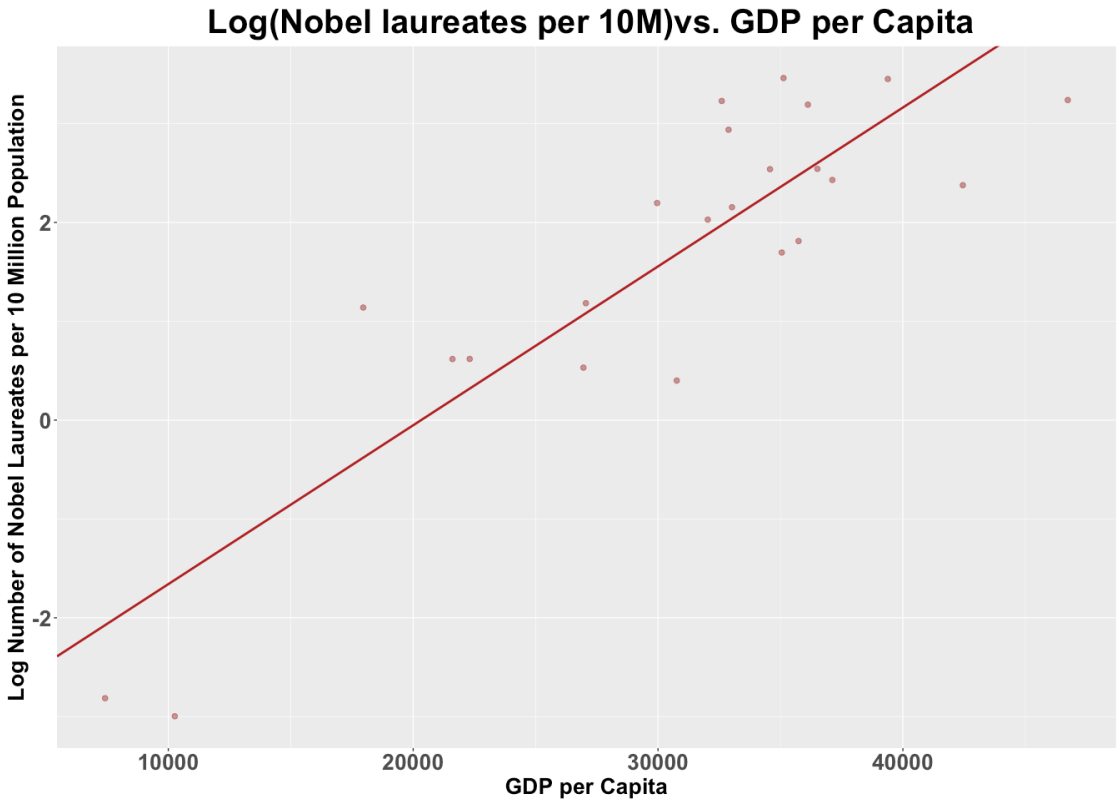

```
grid.arrange(s.NG,s.LOGNG)
```



```
# The log(Nobel Laureates) vs. GDP per capita is more linear.
```

From the above two scatter, the Sweden does not look like a outlier. The point clouded with other points togehter.

```
lm.logresult <- lm(log(nobel_rate) ~ GDP_cap, data = N_C)
lm.logresult

##
## Call:
## lm(formula = log(nobel_rate) ~ GDP_cap, data = N_C)
##
## Coefficients:
## (Intercept)      GDP_cap
##  -3.2665287    0.0001607

s.LOGNGrg <- N_C %>% ggplot(aes(x=GDP_cap, y=log(nobel_rate))) +
geom_point(color="#99000070", size=2) +
  ggtitle(" Log(Nobel laureates per 10M)vs. GDP per Capita") +
  labs(x="GDP per Capita", y="Log Number of Nobel Laureates per 10 Million
Population") +
  geom_abline(intercept= lm.logresult$coefficients[1] ,
slope=lm.logresult$coefficients[2], color='firebrick', size=1)+
  theme.info
s.LOGNGrg
```



```r
beta1 <- exp(lm.logresult$coefficients[2])
beta1
```

```
##  GDP_cap
## 1.000161
```

a unit increase in GDP per capital(person/perdollar/peryear) changes Number of Nobel Laureates by approximately 0.0161%

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

```r
summary(lm.result)
```

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

```r
pt(-6.133,21,lower.tail = FALSE)
```

```
## [1] 0.9999978
```

H0: correlation beta1 >= 0 H1: correlation beta1 < 0 alpha = 0.05

From the Linear Regression Result, the probability that correlation beta >= 0 is 99%

In this case, we can believe that the chocolate consumption has correlation with number of Nobel Laureates, however, correlation does not represent the causation which means we cannot say that increasing chocolate consumption cause an increase in the number of Nobel Laureates.