Small Assignment 1

(due September 10th 2020, 11:59 PM PDT)

This assignment will help you get familiar with using **R** to manipulate data. We have included comments in the code cells to help outline the required steps. We will not provide such comments in the future, but we encourage you to develop a habit of annotating your code in a similar way, so that another reader (perhaps yourself in the future!) can easily follow what you are doing.

If you have questions about using particular functions in **R**, remember that entering `?function` or `help(function)` will pull up a window outlining the syntax for the function and how the arguments should be entered.

Please submit this assignment via Gradescope. To do so you will need to generate a single pdf file of your answers. For small assignments, the pdf can be composed of photo images and/or screenshots if you are not using Jupyter. Please see the submission instructions for help uploading to Gradescope.

Part 1.

We are going to explore the nuts and bolts of regressions by doing one by hand. *Please do not use canned functions to calculate sample variances and covariances and to estimate regression coefficients.*

Start by reading in the `energy.csv` file.

In [1]:

```
# Read in the data and save it as an object
df <-read.csv('energy.csv', header=T)
```

View the structure of the dataset using the `head` function.

In [2]:

```
head(df)
```

A data.frame: 6 × 5

	country_name	country_code	energy_use	gdp	pop
	<chr>	<chr>	<dbl>	<dbl>	<int>
1	Algeria	DZA	37444.586	1.710000e+11	35725377
2	Armenia	ARM	2996.571	1.166202e+10	2977488
3	Bangladesh	BGD	27794.069	7.955435e+10	147969967
4	Botswana	BWA	2151.618	1.111302e+10	1933719
5	Brazil	BRA	248343.704	1.650000e+12	191765567
6	Brunei Darussalam	BRN	3629.024	1.439310e+10	388017

(a) Create two new variables, GDP ($) per capita and energy use (KT oil equivalent) per capita.

Hint: Use the `mutate` function (you will have to load `tidyverse` first).

In [3]:

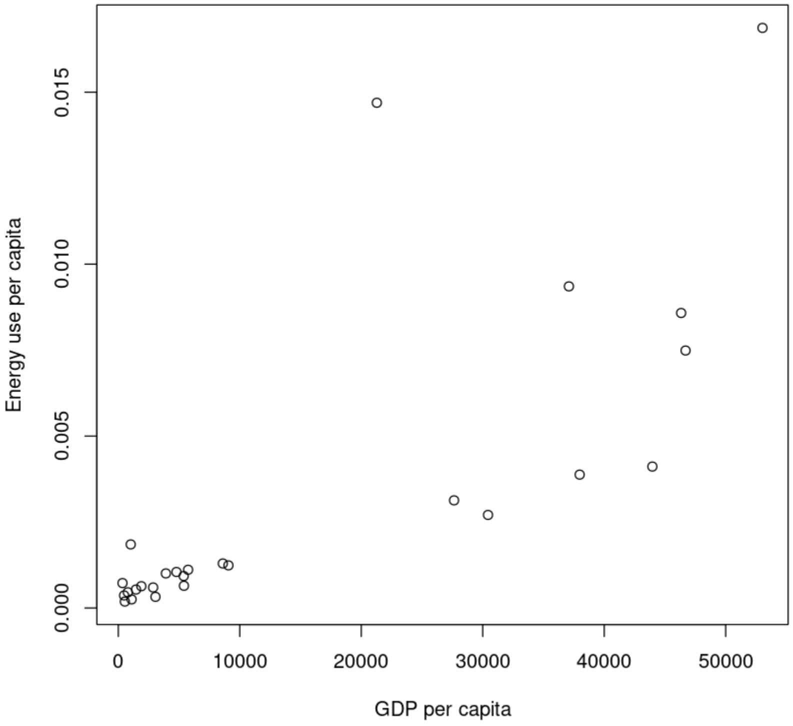
```
# View first few lines of data
library(tidyverse)
#Generate per capita variables
df <- mutate(df, gdp_pc= gdp/pop, energy_pc=energy_use/pop)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

✔ ggplot2 3.3.5     ✔ purrr   0.3.4
✔ tibble  3.1.4     ✔ dplyr   1.0.7
✔ tidyr   1.1.3     ✔ stringr 1.4.0
✔ readr   2.0.1     ✔ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

(b) Generate a scatter plot comparing energy use per capita (on the y-axis) to GDP per capita (on the x-axis) for the countries in your sample. *Hint: Use the* `plot` *function.*

In [4]:

```r
# Simple scatterplot
plot(df$gdp_pc, df$energy_pc, xlab='GDP per capita', ylab='Energy use per capita')
```



(c) Calculate the sample mean of GDP per capita, sample mean of energy per capita, covariance of GDP per capita and energy use per capita, and variance of GDP per capita.

*Hint: The functions `sum()` and `nrow()` will be useful. Create a vector by writing `c(object1, object2, object3,...)` .

In [5]:

```r
# Save means as new objects
xbar <-mean(df$gdp_pc)
ybar <-mean(df$energy_pc)

# Create new columns with residuals (x_i-\bar{x}) and (y_i-\bar{y}) as part of your data
df <- mutate(df, xres= gdp_pc-xbar, yres= energy_pc-ybar)

# Create new columns with the elements you need to calculate the covariance and variance,
df <- mutate(df, xy= xres*yres, xx=xres^2)

# Save the covariance and variance as new objects by summing over the relevant variables a
```

```
covxy<- sum(df$xy)/(nrow(df)-1)
varx<- sum(df$xx)/(nrow(df)-1)

# Display results, using a vector with four elements
c(xbar, ybar, covxy, varx) #note the use of c to create a vector!
```

```
15419.2698158678 · 0.0032316582767966 · 62.272034896757 · 326578025.706693
```

(d) Suppose y_i is energy use per capita in country i and x_i is GDP per capita in country i. Suppose your statistical model is

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ using your results in part **(c)**.

In [6]:

```
# Solve for \hat\beta_1 and \hat\beta_0 and save these objects
b1hat<- covxy/varx
b0hat<- ybar-xbar*b1hat

# Display results as a vector
c(b0hat,b1hat)
```

```
0.000291505442023208 · 1.90680419363809e-07
```

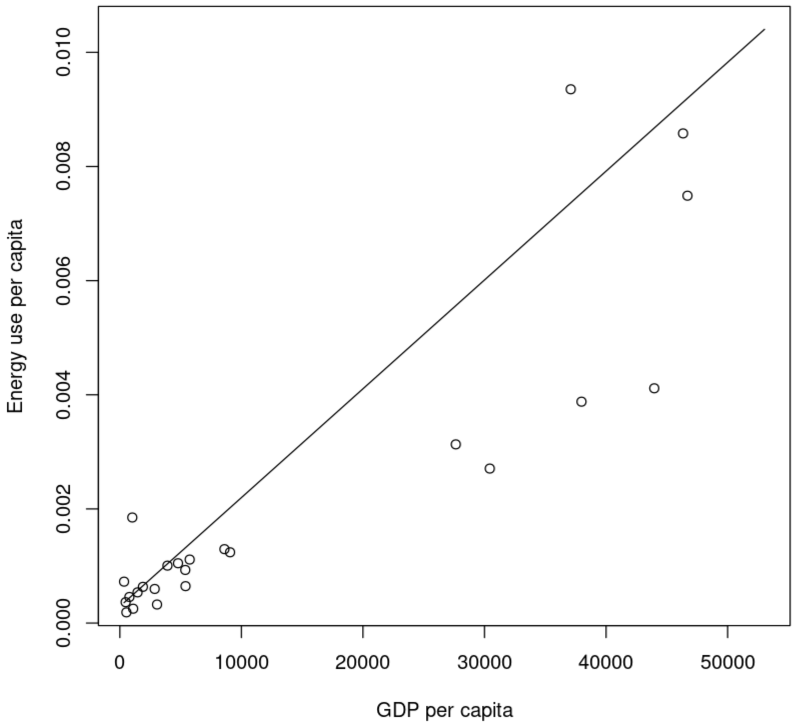
(e) Plot a line which indicates the predictions from your estimated model,

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

Hint: use the `curve` and `points` *functions.*

In [7]:

```
# Plot the line
curve(b0hat+b1hat*x, min(df$gdp_pc), max(df$gdp_pc), xlab='GDP per capita', ylab = 'Energy

# Overlay with data points
points(df$gdp_pc,df$energy_pc)
```



(f) Interpret your estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

$\hat{\beta}_0$ =0.00003 means that the model predicts that a country with 0 GDP would have energy use equivalent to .0003 KT of oil. $\hat{\beta}_1$ =1.9e-07 means that model predicts that each additional dollar of GDP leads to 1.9e-07 addition KT of oil equivalent energy use.

Part 2.

Suppose you estimate a weekly wage equation which suggests

$$\log(wage_i) = 6.45 + .08 * Ed_i + u_i$$

where Ed_i is the number of years of education completed by person i.

(a) Predict the wage for a person with 12 years of education (that is a person who completes high school).

In [8]:

```
exp(6.45+.08*12)
```

```
1652.42634686448
```

(b) Interpret your estimated β_1 =0.08.

An additional year of education is associated with 8% higher weekly wages.

In [ ]: