

Econometrics HW3 Part 3

Na SeungChan

HW2 Q3 : Previous Data Analysis

해당 파트는 HW2의 제출 내용을 그대로 복사해 온 것으로, 데이터 전처리의 편의를 위해서만 사용되고 문제풀이의 내용이 아님.

```
dataset <- read.csv('co2-emissions-vs-gdp.csv') %>%
  filter(Year == 2018) %>%
  filter(Code != '', Code != 'OWID_WRL')
```

```
dataset1 <- read.csv('energy.csv') %>%
  filter(Year == 2018) %>%
  filter(Code != '', Code != 'OWID_WRL')
```

```
full_data <- left_join(dataset, dataset1)
```

```
## Joining, by = c("Entity", "Code", "Year")
```

```
glimpse(full_data)
```

```
## Rows: 238
## Columns: 9
## $ Entity      <chr> "Afghanistan", "Alb~
## $ Code        <chr> "AFG", "ALB", "DZA"~
## $ Year        <int> 2018, 2018, 2018, 2~
## $ Annual.CO..emissions..per.capita. <dbl> 0.2948759, 1.732364~
## $ GDP.per.capita <dbl> 1934.555, 11104.166~
## $ X417485.annotations <chr> "", "", "", "", "",~
## $ Population..historical.estimates. <dbl> 36686788, 2877019, ~
## $ Continent   <chr> "", "", "", "", "",~
## $ Primary.energy.consumption.per.capita.kWh.person. <dbl> 1144.532, 14483.855~
```

Data from our world in data. Focused on countries in 2018 only. First dataset is from <https://ourworldindata.org/grapher/co2-emissions-vs-gdp> , Second dataset is from <https://ourworldindata.org/energy>

full_data are the merged data of above two.

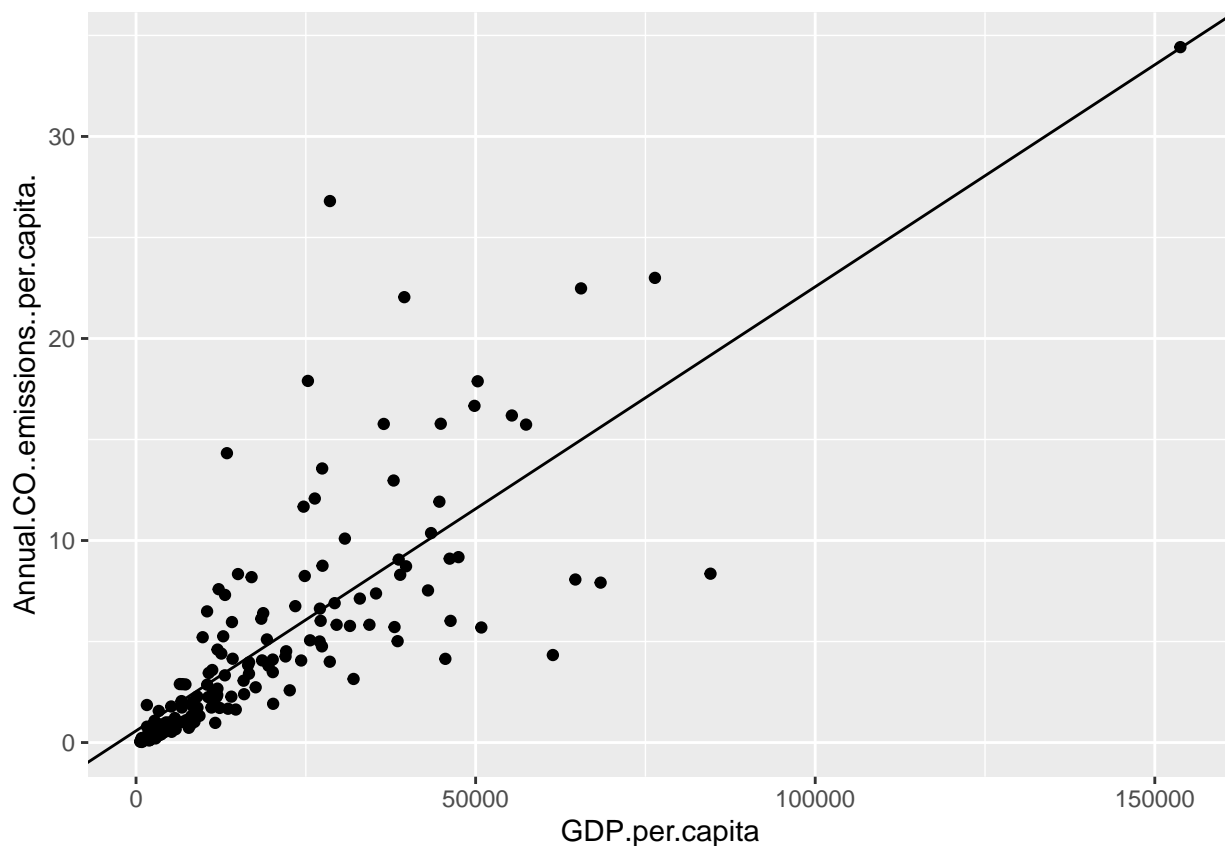
```
lm_a <- lm_robust(Annual.CO..emissions..per.capita. ~ GDP.per.capita, data = full_data)
summary(lm_a)
```

```
##
## Call:
## lm_robust(formula = Annual.CO..emissions..per.capita. ~ GDP.per.capita,
## data = full_data)
##
## Standard error type: HC2
##
## Coefficients:
```

```
##           Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)  0.5783892  2.564e-01  2.256 2.544e-02 0.072006 1.0847728 162
## GDP.per.capita 0.0002198  1.863e-05  11.798 1.400e-23 0.000183 0.0002565 162
##
## Multiple R-squared:  0.6284 ,   Adjusted R-squared:  0.6262
## F-statistic: 139.2 on 1 and 162 DF,  p-value: < 2.2e-16
```

```
ggplot(data = full_data, aes(x = GDP.per.capita, y = Annual.CO..emissions..per.capita.)) +
  geom_point() +
  geom_abline(intercept = 0.5783892, slope = 0.0002198)
```

```
## Warning: Removed 74 rows containing missing values (geom_point).
```



Research Question is about the causal correlation between the GDP per capita and annual CO2 emission per capita.

```
se(intercept) : 2.564e-01, se(slope) : 1.863e-05
```

p-value for slope $\Pr(>|t|) = <2.2e-16$ ***. This means the significance of the regression model.

Add additional variables X_2, \dots, X_k that may affect Y and estimate the multiple regression model. Compare the coefficient estimate of X_1 with (i). Note that the main research question is unchanged. The additional variables are supposed to eliminate possible omitted variable bias in the simple regression.

```
lm_b <- lm_robust(Annual.CO..emissions..per.capita. ~ GDP.per.capita + Primary.energy.consumption.per.capita..k
summary(lm_b)
```

```
##
## Call:
```

```
## lm_robust(formula = Annual.CO..emissions..per.capita. ~ GDP.per.capita +
## Primary.energy.consumption.per.capita..kWh.person., data = full_data)
##
## Standard error type: HC2
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept)      6.754e-01 2.466e-01  2.739
## GDP.per.capita      6.932e-05 5.621e-05  1.233
## Primary.energy.consumption.per.capita..kWh.person. 1.017e-04 4.283e-05  2.375
##              Pr(>|t|) CI Lower
## (Intercept)      0.00686 1.884e-01
## GDP.per.capita      0.21928 -4.168e-05
## Primary.energy.consumption.per.capita..kWh.person. 0.01873 1.714e-05
##              CI Upper DF
## (Intercept)      1.1624382 161
## GDP.per.capita      0.0001803 161
## Primary.energy.consumption.per.capita..kWh.person. 0.0001863 161
##
## Multiple R-squared: 0.7447 , Adjusted R-squared: 0.7416
## F-statistic: 75.44 on 2 and 161 DF, p-value: < 2.2e-16
```

Now I added X2 as 'energy use per capita.' This can be helpful because some of developed countries may use more non-fossil fuel energy.

p-value for 'energy use per capita' is smaller, and adjusted R^2 is bigger for the second model. But, 'the usage of eco-friendly energy' may be more helpful for this purpose, but making data like that was seriously hard data wrangling(keys are not consistent). Making those data can improve the quality of this research.

HW3 - Q3

You may continue on your previously collected dataset, or select another cross- sectional data. Let the variables of interest be Y and X1. You may also consider X2, ..., Xk that may directly affect Y . The main research question is "Does X1 affect Y ? If so, how much?".

- (i) Conduct a nonlinear regression of Y on X1 controlling for X2, ..., Xk and draw the fitted curve on the scatter plot and compare it with the fitted line of the linear regression.
- (ii) Provide the standard errors for the point estimates. Interpret the result.

HW3 - Q3.(i)

```
temp_med <- median(na.omit(full_data$Primary.energy.consumption.per.capita..kWh.person.))

full_data_bined <- full_data %>%
  mutate(HiEnergy = ifelse(Primary.energy.consumption.per.capita..kWh.person. >= temp_med, 1, 0))

remove(temp_med)

head(full_data_bined)

##      Entity Code Year Annual.CO..emissions..per.capita GDP.per.capita
## 1 Afghanistan AFG 2018      0.2948759    1934.555
## 2  Albania ALB 2018      1.7323643    11104.166
## 3   Algeria DZA 2018      4.1479607    14228.025
## 4 American Samoa ASM 2018      NA        NA
```

```
## 5 Andorra AND 2018 6.5922117 NA
## 6 Angola AGO 2018 0.7283953 7771.442
## X417485.annotations Population..historical.estimates. Continent
## 1 36686788
## 2 2877019
## 3 41927008
## 4 48445
## 5 75034
## 6 31273538
## Primary.energy.consumption.per.capita..kWh.person. HiEnergy
## 1 1144.532 0
## 2 14483.855 0
## 3 16012.625 0
## 4 29712.482 1
## 5 NA NA
## 6 3165.137 0
```

Hienergy 변수를 생성하여 통제를 진행하였다. 기준은 na.omit를 사용하여 na values인 국가들을 제외한 뒤의 중앙값이다. 단, 이와 같은 'NA 무작정 제거'가 통계적으로 문제가 될 수 있는데, 해당 데이터에서는 국가의 행정력이 잘 미치지 못하는 국가일수록 na가 발생할 가능성이 높아 NA에 체계적인 경향이 발생할 수 있기 때문이다. 차후 이와 같은 문제를 검토하기 위한 통계적 기법에 따른 연구를 수행하여야 할 것이다.

```
lm_c <- lm_robust(log(Annual.CO..emissions..per.capita.) ~ GDP.per.capita + HiEnergy + GDP.per.capita:HiEnergy ,
summary(lm_c)
```

```
##
## Call:
## lm_robust(formula = log(Annual.CO..emissions..per.capita.) ~
## GDP.per.capita + HiEnergy + GDP.per.capita:HiEnergy, data = full_data_bined)
##
## Standard error type: HC2
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|) CI Lower
## (Intercept) -1.5929412 1.566e-01 -10.174 4.617e-19 -1.9021573
## GDP.per.capita 0.0002086 1.957e-05 10.658 2.223e-20 0.0001699
## HiEnergy 2.9736159 1.910e-01 15.570 7.276e-34 2.5964458
## GDP.per.capita:HiEnergy -0.0001918 1.978e-05 -9.694 9.086e-18 -0.0002308
## CI Upper DF
## (Intercept) -1.2837251 160
## GDP.per.capita 0.0002472 160
## HiEnergy 3.3507860 160
## GDP.per.capita:HiEnergy -0.0001527 160
##
## Multiple R-squared: 0.8265 , Adjusted R-squared: 0.8233
## F-statistic: 204.5 on 3 and 160 DF, p-value: < 2.2e-16
```

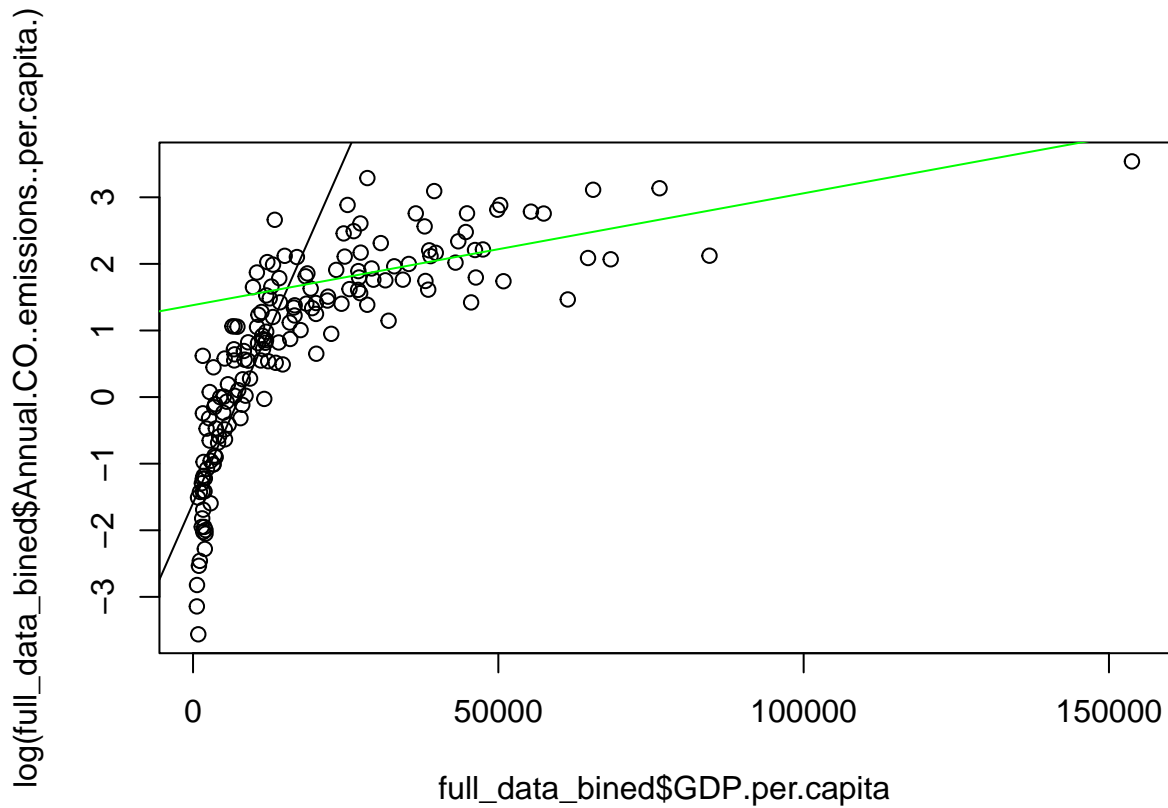
```
lm_c$coefficients
```

```
## (Intercept) GDP.per.capita HiEnergy
## -1.5929411999 0.0002085654 2.9736158920
## GDP.per.capita:HiEnergy
## -0.0001917723
```

```
plot(full_data_bined$GDP.per.capita, log(full_data_bined$Annual.CO..emissions..per.capita.))
```

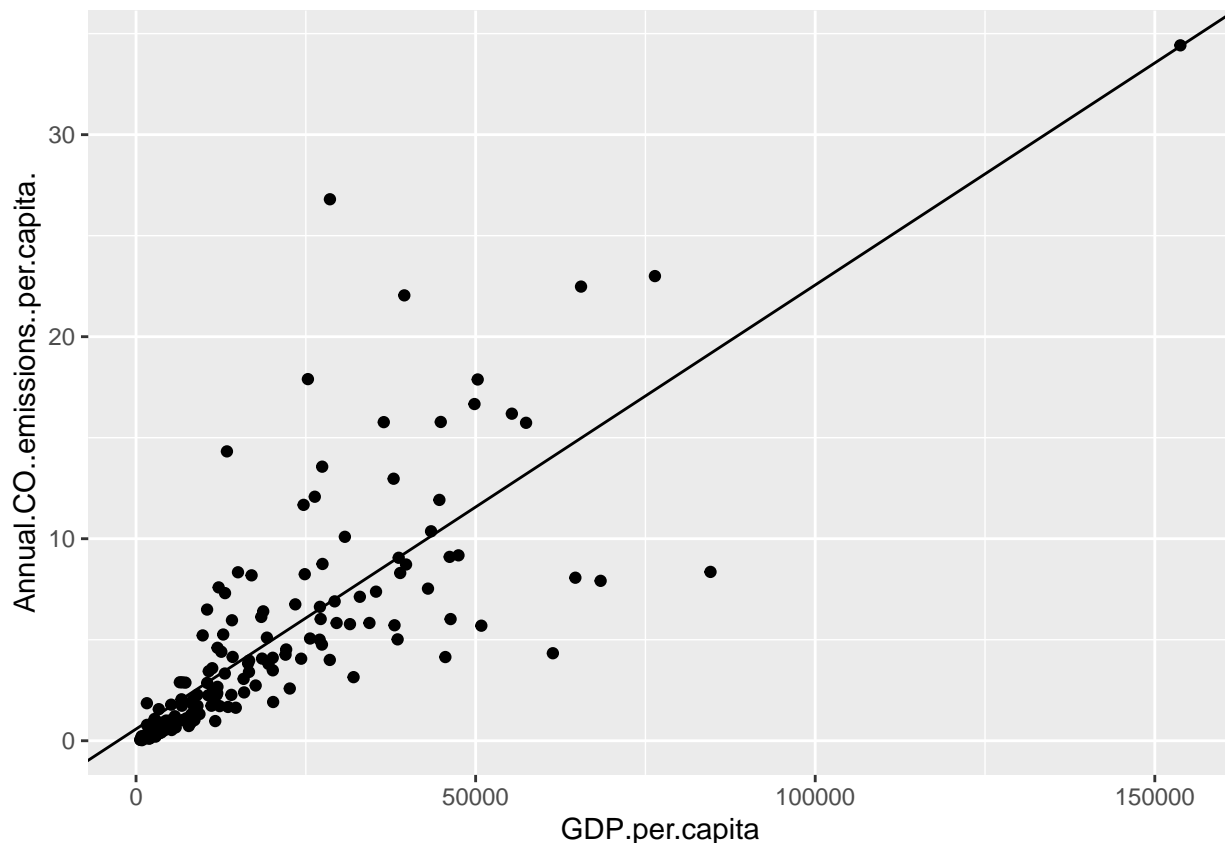
```
abline(coef = c(lm_c$coefficients[1], lm_c$coefficients[2]))
```

```
abline(coef = c(lm_c$coefficients[1] + lm_c$coefficients[3], lm_c$coefficients[2] + lm_c$coefficients[4]), col = 'green')
```



```
ggplot(data = full_data, aes(x = GDP.per.capita, y = Annual.CO2.emissions..per.capita.)) +  
  geom_point() +  
  geom_abline(intercept = 0.5783892, slope = 0.0002198)
```

```
## Warning: Removed 74 rows containing missing values (geom_point).
```



아래쪽 그림의 linear model과 비교했을 때 교호작용 term을 포함한 model은 가파른 기울기의 한 직선($H_i = 0$)과 상대적으로 완만한 직선($H_i = 1$)으로 구분되며, 교호작용 term의 유의성이 있을 것이라고 추측할 수 있다.

HW3 - Q3.(ii)

```
summary(lm_c)
```

```
##
## Call:
## lm_robust(formula = log(Annual.CO..emissions..per.capita.) ~
##   GDP.per.capita + HiEnergy + GDP.per.capita:HiEnergy, data = full_data_bined)
##
## Standard error type: HC2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|) CI Lower
## (Intercept)   -1.5929412  1.566e-01 -10.174 4.617e-19 -1.9021573
## GDP.per.capita    0.0002086  1.957e-05  10.658 2.223e-20  0.0001699
## HiEnergy        2.9736159  1.910e-01  15.570 7.276e-34  2.5964458
## GDP.per.capita:HiEnergy -0.0001918  1.978e-05  -9.694 9.086e-18 -0.0002308
##              CI Upper DF
## (Intercept)   -1.2837251 160
## GDP.per.capita    0.0002472 160
## HiEnergy        3.3507860 160
## GDP.per.capita:HiEnergy -0.0001527 160
```

##

Multiple R-squared: 0.8265 , Adjusted R-squared: 0.8233

F-statistic: 204.5 on 3 and 160 DF, p-value: < 2.2e-16

HiEnergy = 0인 경우의 회귀선은 $\log(y) = -1.5929 + 0.0002086x$

HiEnergy = 1인 경우의 회귀선은 $\log(y) = \text{lm_c}\$coefficients[1] + \text{lm_c}\$coefficients[3] + \text{lm_c}\$coefficients[2] + \text{lm_c}\$coefficients[4]x$

Interpret the result.

즉 x의 변화량으로 $E(Y|D,X)$ 의 변화량을 나눈 값은 $\text{lm_c}\$coefficients[2] + \text{lm_c}\$coefficients[3]D0$ 이다.