research_question_1

```
library(tidyverse)
library(readxl)
library(dplyr)
library(ggplot2)
```

```
#Read the dataset, select the columns to be used and drop the missing data
energy_rq1 <- read_excel("owid-energy-data.xlsx") %>%
  select(country, year, gdp, population, carbon_intensity_elec,fossil_share_elec,electricity_
  mutate(gdp_per_capita = gdp / population) %>%
  filter(!is.na(carbon_intensity_elec),
         !is.na(gdp_per_capita))

glimpse(energy_rq1)
```

```
Rows: 3,800
Columns: 8
$ country                <chr> "Afghanistan", "Afghanistan", "Afghanistan", "A~
$ year                   <dbl> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007,~
$ gdp                    <dbl> 11283793920, 11021273088, 18804871168, 21074343~
$ population             <dbl> 20130279, 20284252, 21378081, 22733007, 2356059~
$ carbon_intensity_elec  <dbl> 250.000, 217.391, 169.014, 241.758, 227.848, 21~
$ fossil_share_elec      <dbl> 35.417, 27.536, 21.127, 30.769, 29.114, 28.049,~
$ electricity_generation <dbl> 0.48, 0.69, 0.71, 0.91, 0.79, 0.82, 0.90, 1.01,~
$ gdp_per_capita         <dbl> 560.5384, 543.3414, 879.6333, 927.0372, 947.877~
```

Research question 1:

Does a country's GDP per capita significantly affect the carbon intensity of its electricity generation, and does this relationship differ between high-income and low-income countries?

- Outcome variable (include the name/description and type of variable):

- `carbon_intensity_elec` **(continuous)** measures the amount of carbon dioxide emitted per unit of electricity generated (grams of CO per kilowatt-hour). It captures how carbon-efficient a country's electricity production is — lower values indicate cleaner, more renewable-based power systems.

```
head(energy_rq1, 5)
```

```
# A tibble: 5 x 8
  country      year       gdp population carbon_intensity_elec fossil_share_elec
  <chr>       <dbl>     <dbl>      <dbl>                 <dbl>              <dbl>
1 Afghanistan  2000   1.13e10   20130279                  250                35.4
2 Afghanistan  2001   1.10e10   20284252                  217.               27.5
3 Afghanistan  2002   1.88e10   21378081                  169.               21.1
4 Afghanistan  2003   2.11e10   22733007                  242.               30.8
5 Afghanistan  2004   2.23e10   23560598                  228.               29.1
# i 2 more variables: electricity_generation <dbl>, gdp_per_capita <dbl>
```

Because the OWID Energy dataset does not include income classifications, we merged it with a separate World Bank income-group dataset (from same source) that provides each country's income level by year. This merge, done using country name and year, allows us to include income group as a key predictor and test whether the GDP–carbon-intensity relationship differs across income categories.

```
income <- read.csv(
  "https://raw.githubusercontent.com/owid/owid-datasets/master/datasets/Country%20Income%20Cl
)
head(income, 5)
```

```
       Entity Year Income.classification..World.Bank.2017.
1 Afghanistan 1987                                       1
2 Afghanistan 1988                                       1
3 Afghanistan 1989                                       1
4 Afghanistan 1990                                       1
5 Afghanistan 1991                                       1
```

```r
income <- income %>%
  rename(
    country = Entity,
    income_code = Income.classification..World.Bank.2017.,
    year = Year
  )

income <- income %>%
  mutate(income_group = case_when(
    income_code == 4 ~ "High income",
    income_code == 3 ~ "Upper middle income",
    income_code == 2 ~ "Lower middle income",
    income_code == 1 ~ "Low income"
  ))
```

```r
energy_income <- energy_rq1 %>%
  left_join(income, by = c("country", "year"))

head(energy_income)
```

```
# A tibble: 6 x 10
  country      year       gdp population carbon_intensity_elec fossil_share_elec
  <chr>       <dbl>     <dbl>      <dbl>                 <dbl>              <dbl>
1 Afghanistan  2000   1.13e10   20130279                  250                35.4
2 Afghanistan  2001   1.10e10   20284252                  217.               27.5
3 Afghanistan  2002   1.88e10   21378081                  169.               21.1
4 Afghanistan  2003   2.11e10   22733007                  242.               30.8
5 Afghanistan  2004   2.23e10   23560598                  228.               29.1
6 Afghanistan  2005   2.54e10   24404520                  220.               28.0
# i 4 more variables: electricity_generation <dbl>, gdp_per_capita <dbl>,
#   income_code <int>, income_group <chr>
```

Clean Data

```r
library(dplyr)

cleaned <- energy_income %>%
  filter(
    !is.na(carbon_intensity_elec),
```

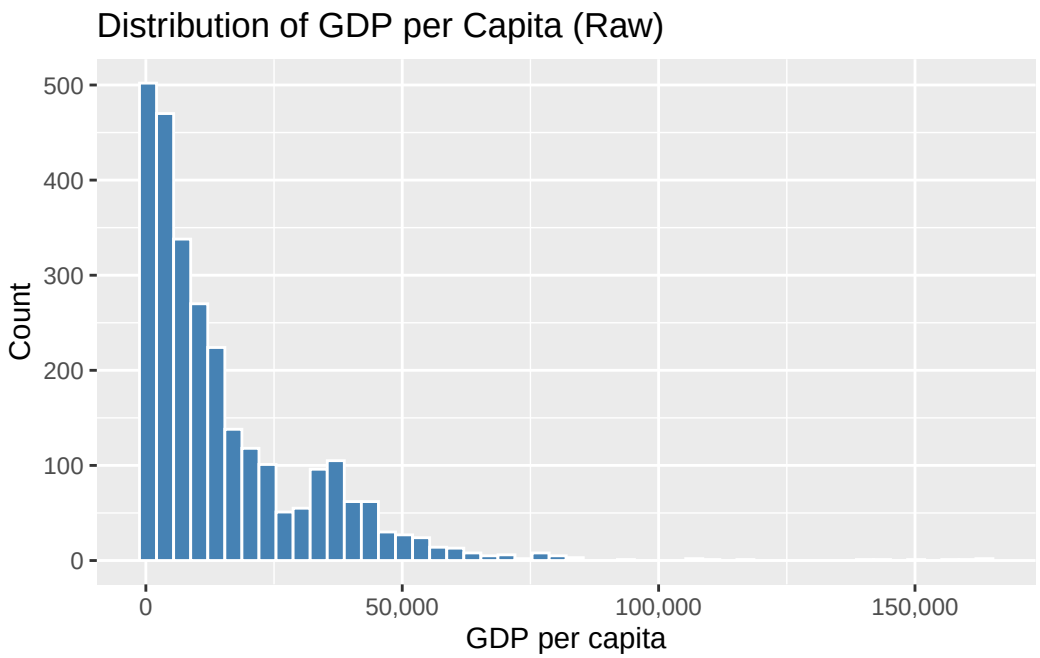
```
    !is.na(gdp_per_capita),
    !is.na(income_group)
  ) %>%
  mutate(
    log_gdp = log(gdp_per_capita),
    income_group = factor(income_group)   # ensure categorical
  )

cleaned$income_group <- factor(cleaned$income_group,
                               levels = c("Low income",
                                          "Lower middle income",
                                          "Upper middle income",
                                          "High income"))


head(cleaned)
```

```
# A tibble: 6 x 11
  country      year       gdp population carbon_intensity_elec fossil_share_elec
  <chr>       <dbl>     <dbl>      <dbl>                 <dbl>              <dbl>
1 Afghanistan  2000   1.13e10   20130279                  250                35.4
2 Afghanistan  2001   1.10e10   20284252                  217.               27.5
3 Afghanistan  2002   1.88e10   21378081                  169.               21.1
4 Afghanistan  2003   2.11e10   22733007                  242.               30.8
5 Afghanistan  2004   2.23e10   23560598                  228.               29.1
6 Afghanistan  2005   2.54e10   24404520                  220.               28.0
# i 5 more variables: electricity_generation <dbl>, gdp_per_capita <dbl>,
#   income_code <int>, income_group <fct>, log_gdp <dbl>
```

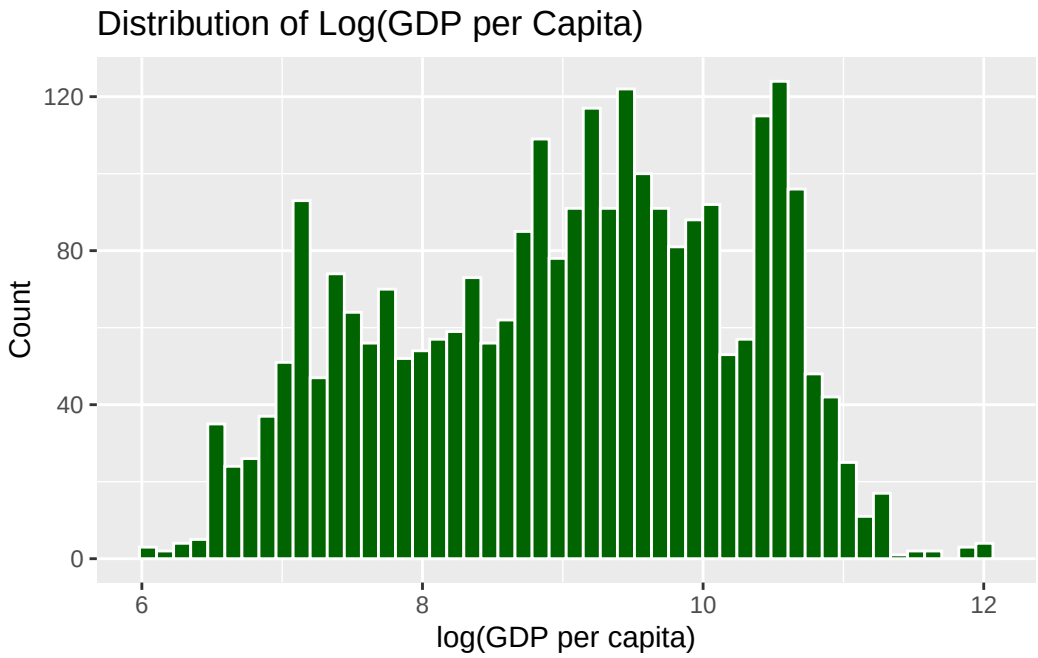
EDA

```
library(ggplot2)

ggplot(cleaned, aes(x = gdp_per_capita)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white") +
  scale_x_continuous(labels = scales::comma) +
  labs(
```

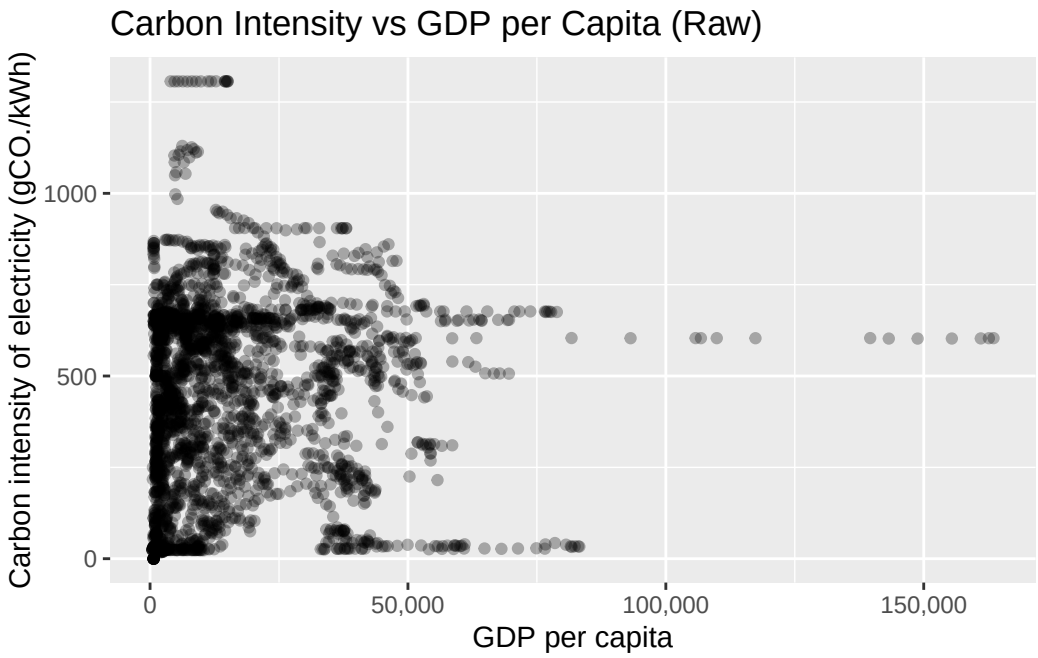
```
    title = "Distribution of GDP per Capita (Raw)",
    x = "GDP per capita",
    y = "Count"
  )
```



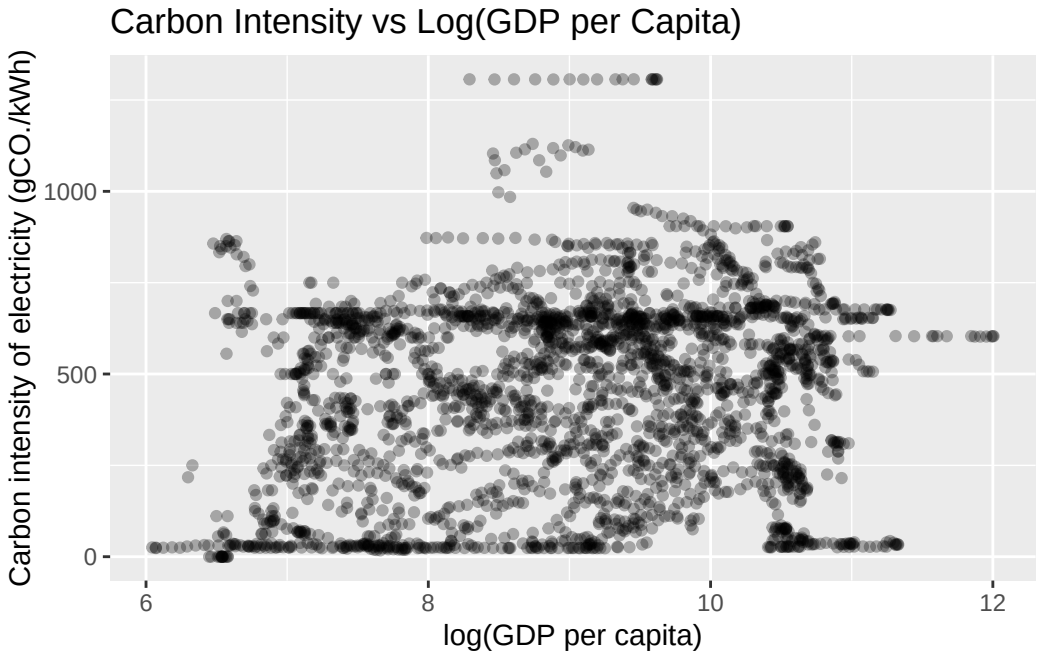

```
ggplot(cleaned, aes(x = log_gdp)) +
  geom_histogram(bins = 50, fill = "darkgreen", color = "white") +
  labs(
    title = "Distribution of Log(GDP per Capita)",
    x = "log(GDP per capita)",
    y = "Count"
  )
```




```r
ggplot(cleaned, aes(x = gdp_per_capita, y = carbon_intensity_elec)) +
  geom_point(alpha = 0.3) +
  scale_x_continuous(labels = scales::comma) +
  labs(
    title = "Carbon Intensity vs GDP per Capita (Raw)",
    x = "GDP per capita",
    y = "Carbon intensity of electricity (gCO /kWh)"
  )
```




```
ggplot(cleaned, aes(x = log_gdp, y = carbon_intensity_elec)) +
  geom_point(alpha = 0.3) +
  labs(
    title = "Carbon Intensity vs Log(GDP per Capita)",
    x = "log(GDP per capita)",
    y = "Carbon intensity of electricity (gCO /kWh)"
  )
```



In the histogram, GDP per capita is extremely right-skewed, with most countries clustered at low values and a long tail of very high-income countries. Log-transforming GDP produces a more symmetric distribution appropriate for linear modeling. In the scatter plot, the relationship between GDP and carbon intensity is curved and compresses most observations into a narrow region at low GDP levels. Using log(GDP) spreads the data more evenly and produces a clearer, more linear relationship suitable for regression.

Does GDP per capita affect carbon intensity, and does this relationship differ by income group?

This model requires an interaction term.

$$CI_i = \beta_0 + +\beta_1\, GDPpc_i + \beta_2\, IncomeGroup_i + \beta_3\, GDPpc_i \times IncomeGroup_i + \varepsilon_i.$$

```
# basic raw model
raw_model <- lm(
  carbon_intensity_elec ~ gdp_per_capita * income_group,
  data = cleaned
)

summary(raw_model)
```

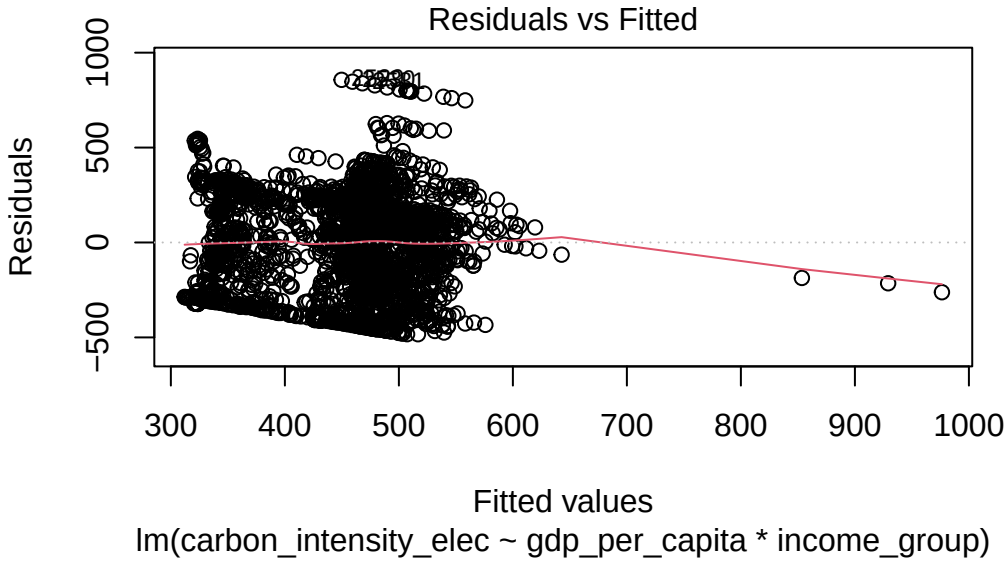
```
Call:
lm(formula = carbon_intensity_elec ~ gdp_per_capita * income_group,
    data = cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-483.35 -201.90   30.64  182.98  856.83

Coefficients:
                                                 Estimate Std. Error t value
(Intercept)                                    295.597927  15.802488  18.706
gdp_per_capita                                   0.039055   0.006596   5.921
income_groupLower middle income                104.885694  28.758743   3.647
income_groupUpper middle income                119.363430  35.163165   3.395
income_groupHigh income                        211.929026  26.496353   7.998
gdp_per_capita:income_groupLower middle income  -0.026713   0.007427  -3.597
gdp_per_capita:income_groupUpper middle income  -0.032730   0.006923  -4.728
gdp_per_capita:income_groupHigh income          -0.040055   0.006615  -6.055
                                                Pr(>|t|)
(Intercept)                                     < 2e-16 ***
gdp_per_capita                                  3.60e-09 ***
income_groupLower middle income                 0.000270 ***
income_groupUpper middle income                 0.000697 ***
income_groupHigh income                         1.84e-15 ***
gdp_per_capita:income_groupLower middle income  0.000328 ***
gdp_per_capita:income_groupUpper middle income  2.39e-06 ***
gdp_per_capita:income_groupHigh income          1.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 247.6 on 2741 degrees of freedom
Multiple R-squared:  0.06057,   Adjusted R-squared:  0.05817
F-statistic: 25.25 on 7 and 2741 DF,  p-value: < 2.2e-16
```

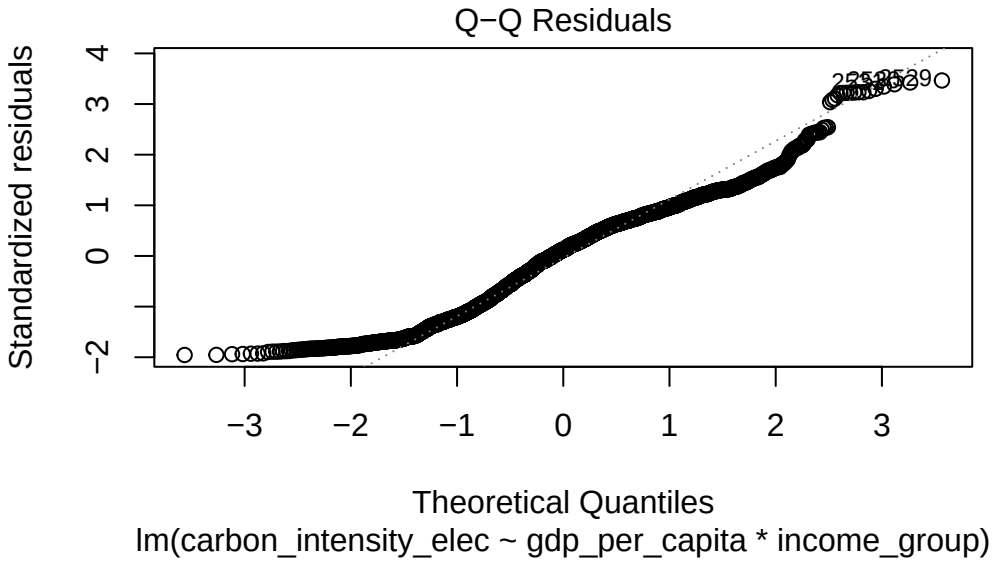
```
confint(raw_model)
```

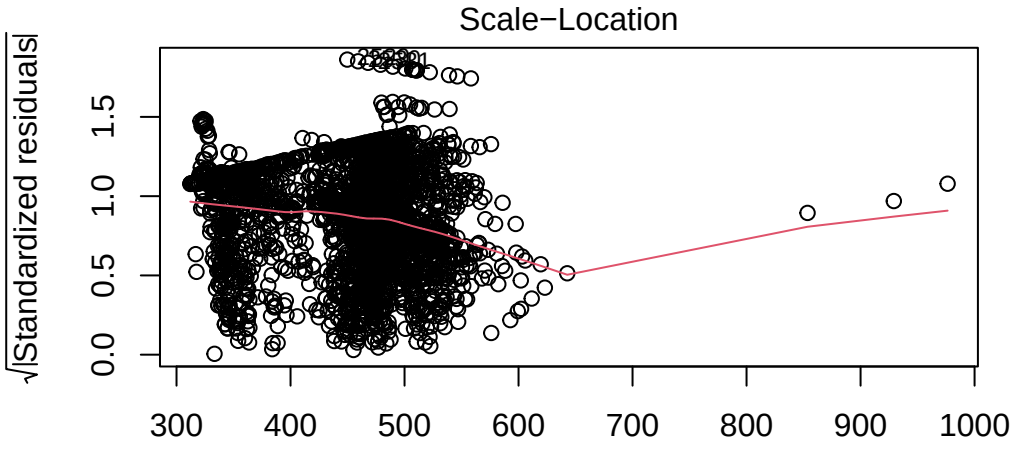
```
                                                    2.5 %       97.5 %
(Intercept)                                   264.61193755 326.58391696
gdp_per_capita                                  0.02612142   0.05198831
income_groupLower middle income                48.49469204 161.27669535
income_groupUpper middle income                50.41444568 188.31241377
```

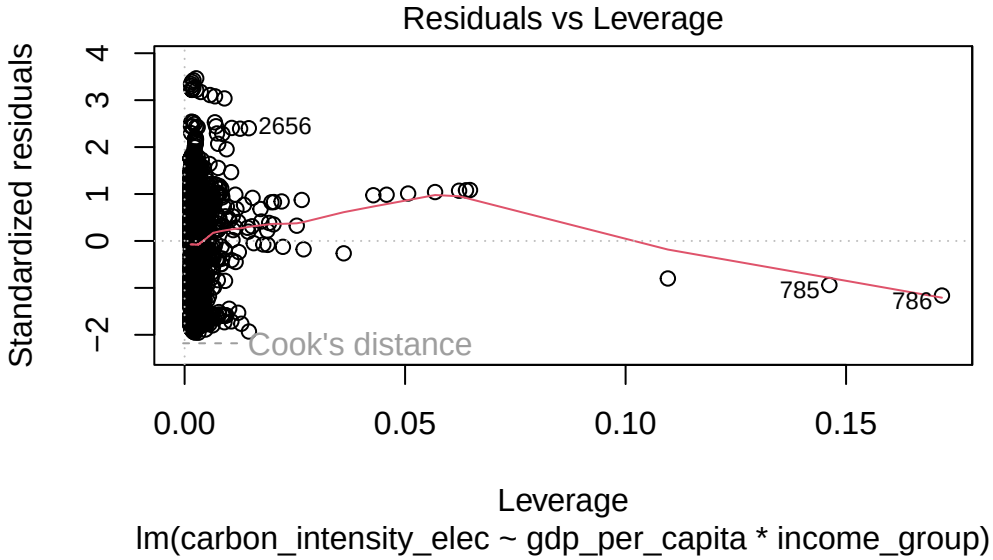
```
income_groupHigh income                           159.97418515 263.88386596
gdp_per_capita:income_groupLower middle income  -0.04127644  -0.01215023
gdp_per_capita:income_groupUpper middle income  -0.04630469  -0.01915448
gdp_per_capita:income_groupHigh income           -0.05302551  -0.02708484
```

```
plot(raw_model)
```









This raw model treats GDP per capita as a linear, untransformed, and evenly spaced continuous variable. As hinted in the EDA, this may be problematic due to the following observations:

- GDP per capita is very right skewed. Most countries have a very low GDP and a few have extremely high GDP.
- The scatterplot with raw GDP from earlier shows a curved, nonlinear pattern, while the model tries to force that curvature into a straight line.

How this affects the results?

The raw model forces linearity such that:

- All income groups have a positive main effect of GDP increasing carbon intensity.
- GDP is skewed so that small numeric changes at the low end represent tiny economic differences while numeric changes at the high end represent huge jumps in wealth. Thus, the slopes could be distorted and not represent true relationships.

Diagnostics

- Heteroskedasticity observed by fan shape in Residuals vs Fitted plot
- Curvature in Residuals vs Fitted suggesting linearity violated.

```
model <- lm(
  carbon_intensity_elec ~ log_gdp * income_group,
  data = cleaned
)

summary(model)
```

```
Call:
lm(formula = carbon_intensity_elec ~ log_gdp * income_group,
    data = cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-484.22 -203.46   28.01  181.72  853.69

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -404.65     126.40  -3.201  0.00138
log_gdp                                  104.75      17.02   6.156 8.55e-10
income_groupLower middle income          262.54     217.00   1.210  0.22644
income_groupUpper middle income          178.96     310.01   0.577  0.56381
income_groupHigh income                 1722.88     273.46   6.300 3.45e-10
log_gdp:income_groupLower middle income  -33.00      26.49  -1.245  0.21307
log_gdp:income_groupUpper middle income  -27.91      34.29  -0.814  0.41571
log_gdp:income_groupHigh income         -185.81      28.72  -6.471 1.15e-10

(Intercept)                             **
log_gdp                                 ***
income_groupLower middle income
income_groupUpper middle income
income_groupHigh income                 ***
```

```
log_gdp:income_groupLower middle income
log_gdp:income_groupUpper middle income
log_gdp:income_groupHigh income         ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 247.2 on 2741 degrees of freedom
Multiple R-squared:  0.06329,   Adjusted R-squared:  0.06089
F-statistic: 26.46 on 7 and 2741 DF,  p-value: < 2.2e-16
```

```
confint(model)
```

```
                                        2.5 %     97.5 %
(Intercept)                          -652.49450 -156.80330
log_gdp                                71.38483  138.11560
income_groupLower middle income      -162.96320  688.03799
income_groupUpper middle income      -428.92212  786.83574
income_groupHigh income              1186.65892 2259.09261
log_gdp:income_groupLower middle income  -84.94172   18.95155
log_gdp:income_groupUpper middle income  -95.14456   39.32257
log_gdp:income_groupHigh income      -242.12178 -129.50780
```

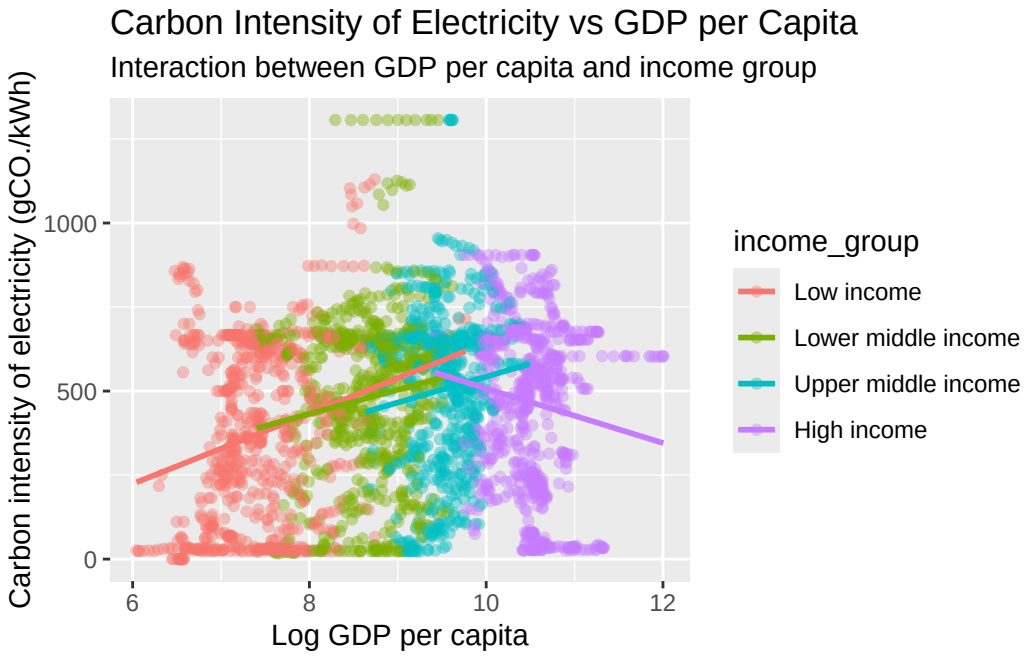
```
coefs <- coef(model)

#  Low income  (reference)
slope_low <- coefs["log_gdp"]

# Lower middle income
slope_lower_middle <- coefs["log_gdp"] +
  coefs["log_gdp:income_groupLower middle income"]

# Upper middle income
slope_upper_middle <- coefs["log_gdp"] +
  coefs["log_gdp:income_groupUpper middle income"]

# High income
slope_high <- coefs["log_gdp"] +
  coefs["log_gdp:income_groupHigh income"]

c(slope_low, slope_lower_middle, slope_upper_middle, slope_high)
```

```
  log_gdp   log_gdp   log_gdp   log_gdp
```

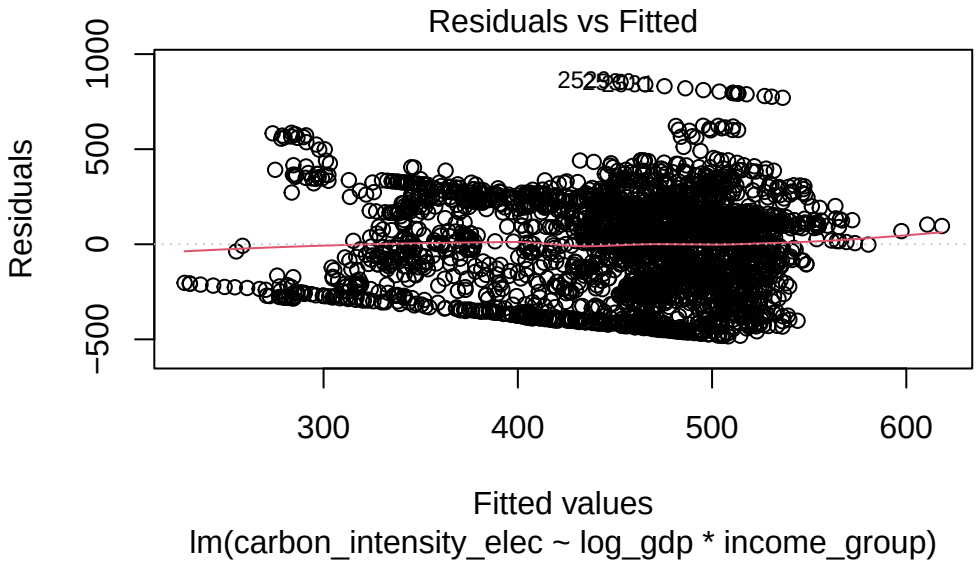
```
104.75022  71.75513  76.83922 -81.06457
```

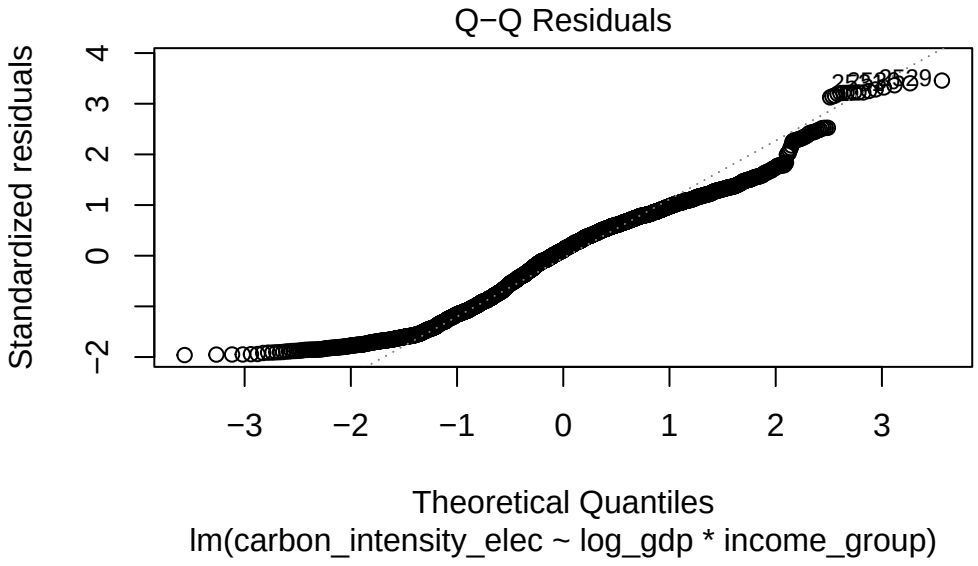
```r
library(ggplot2)

ggplot(cleaned, aes(x = log_gdp, y = carbon_intensity_elec, color = income_group)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Carbon Intensity of Electricity vs GDP per Capita",
    subtitle = "Interaction between GDP per capita and income group",
    x = "Log GDP per capita",
    y = "Carbon intensity of electricity (gCO /kWh)"
  )
```

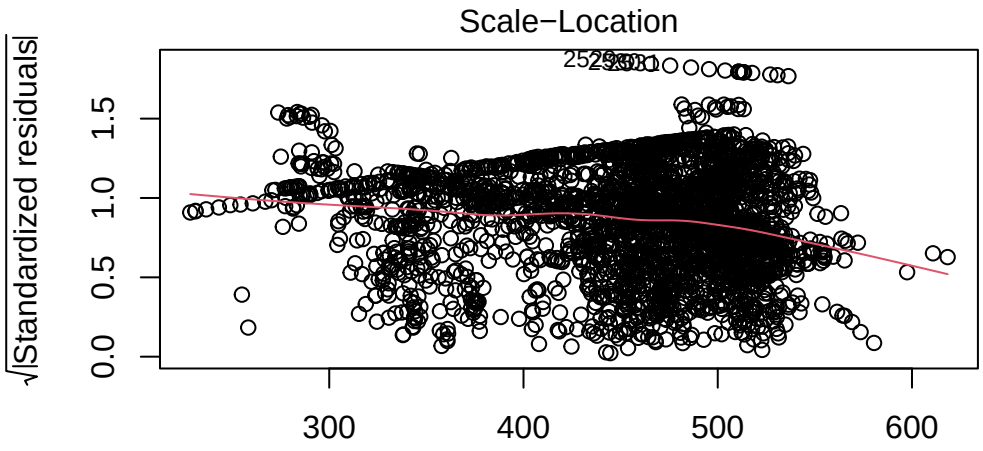
```
`geom_smooth()` using formula = 'y ~ x'
```

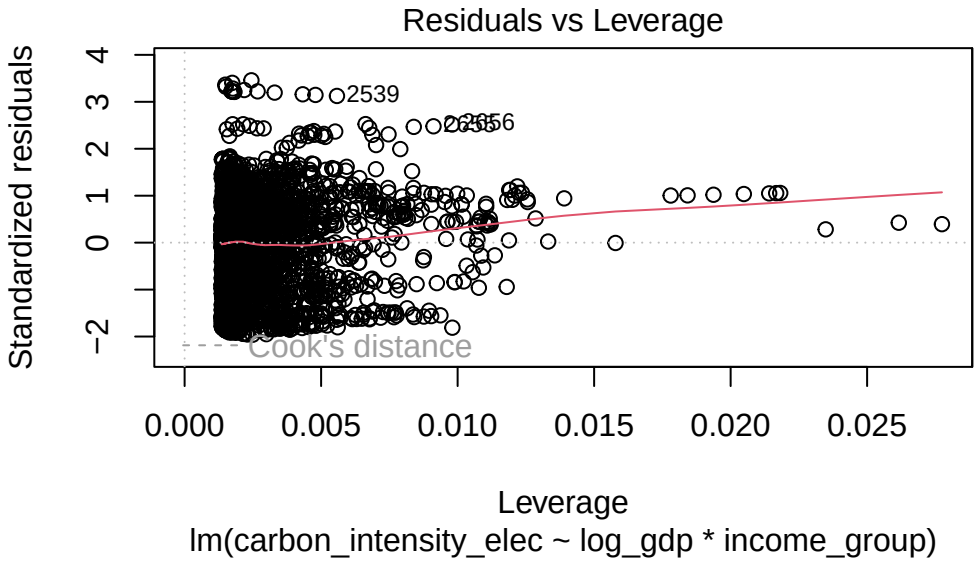



```r
plot(model)
```


Residuals vs Fitted
Residuals
1000
500
0
−500
300
400
500
600
Fitted values
lm(carbon_intensity_elec ~ log_gdp * income_group)


Q−Q Residuals
Standardized residuals
4
3
2
1
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(carbon_intensity_elec ~ log_gdp * income_group)






```
confint(model)
```

```
                                                2.5 %     97.5 %
(Intercept)                                -652.49450 -156.80330
log_gdp                                      71.38483  138.11560
income_groupLower middle income            -162.96320  688.03799
income_groupUpper middle income            -428.92212  786.83574
income_groupHigh income                    1186.65892 2259.09261
log_gdp:income_groupLower middle income     -84.94172   18.95155
log_gdp:income_groupUpper middle income     -95.14456   39.32257
```

`log_gdp:income_groupHigh income` `-242.12178 -129.50780`

The model we fit is:

$$CI_i = \beta_0 + +\beta_1 \log(GDPpc_i) + \beta_2 \, IncomeGroup_i + \beta_3 \left[\log(GDPpc_i) \times IncomeGroup_i\right] + \varepsilon_i.$$

The use of an **interaction term** allows the slope of GDP per capita to differ across income groups, which is essential for answering whether economic development operates differently in richer vs. poorer countries.

Model Interpretation

When holding all other variables constant:

- For **low-income countries**, a **one-unit increase in log(GDP per capita)** is associated with an **increase of 104.75 gCO /kWh** in carbon intensity, holding income group constant. This means that in low-income countries, economic growth corresponds to **more carbon-intensive electricity**.
- Holding log(GDP per capita) constant, **lower-middle-income countries** have carbon intensity values that are **262.54 gCO /kWh higher** on average than **low-income countries**, but the difference is **not statistically significant**.
- Holding log(GDP per capita) constant, **upper-middle-income countries** have carbon intensity values that are **178.96 gCO /kWh higher** on average than **low-income countries**, but this difference is **not statistically significant**.
- Holding log(GDP per capita) constant, **high-income countries** have carbon intensity levels that are **1,722.88 gCO /kWh higher** than **low-income countries**, and this difference **is statistically significant**. This reflects major structural differences in electricity systems across income levels.
- Compared to low-income countries, the slope relating log(GDP) to carbon intensity in lower-middle-income countries is **33.00 gCO /kWh smaller** for each unit increase in log(GDP), but the difference is **not statistically significant**.
- Compared to low-income countries, the slope is **27.91 gCO /kWh smaller** for every one-unit increase in log(GDP). This difference is **not statistically significant**.
- Compared to low-income countries, the slope is **185.81 gCO /kWh more negative** for each unit increase in log(GDP). This is **highly statistically significant**, and it completely reverses the sign of the GDP effect in high-income countries.
  - In low-income countries, GDP growth **increases** carbon intensity.
  - In high-income countries, GDP growth **decreases** carbon intensity.

This suggests that wealthy countries decarbonize as they grow, whereas poorer countries carbonize as they grow.

The reversal of the GDP effect between low and high-income countries indicates fundamentally different development pathways. Low-income countries appear to rely more heavily on coal, diesel, and other fossil fuels when expanding electricity access, causing carbon intensity to rise as GDP increases. High-income countries, however, may already have built or transitioned toward cleaner electricity systems, so additional economic growth is associated with investments in renewable energy, efficiency, and environmental regulations. As a result, high-income countries show declining carbon intensity as GDP increases, which suggests decarbonization, while low-income countries show increasing carbon intensity, reflecting carbonization during early development.

Model Assessment/Diagnostics

- Linearity: The Residuals vs Fitted plot does not show any major curvature so the linearity assumption is reasonable.
- Normality: The Q-Q plot shows slight deviation at the tail ends but is overall reasonable, so the normality of residuals assumption is not violated.
- Homoskedasticity: The Residuals vs Fitted plot shows moderate heteroskedasticity indicated via a fan shape in the residuals.
- Influential Observations: The residuals leverage plot does not show any observations with concerningly high Cook's Distance.