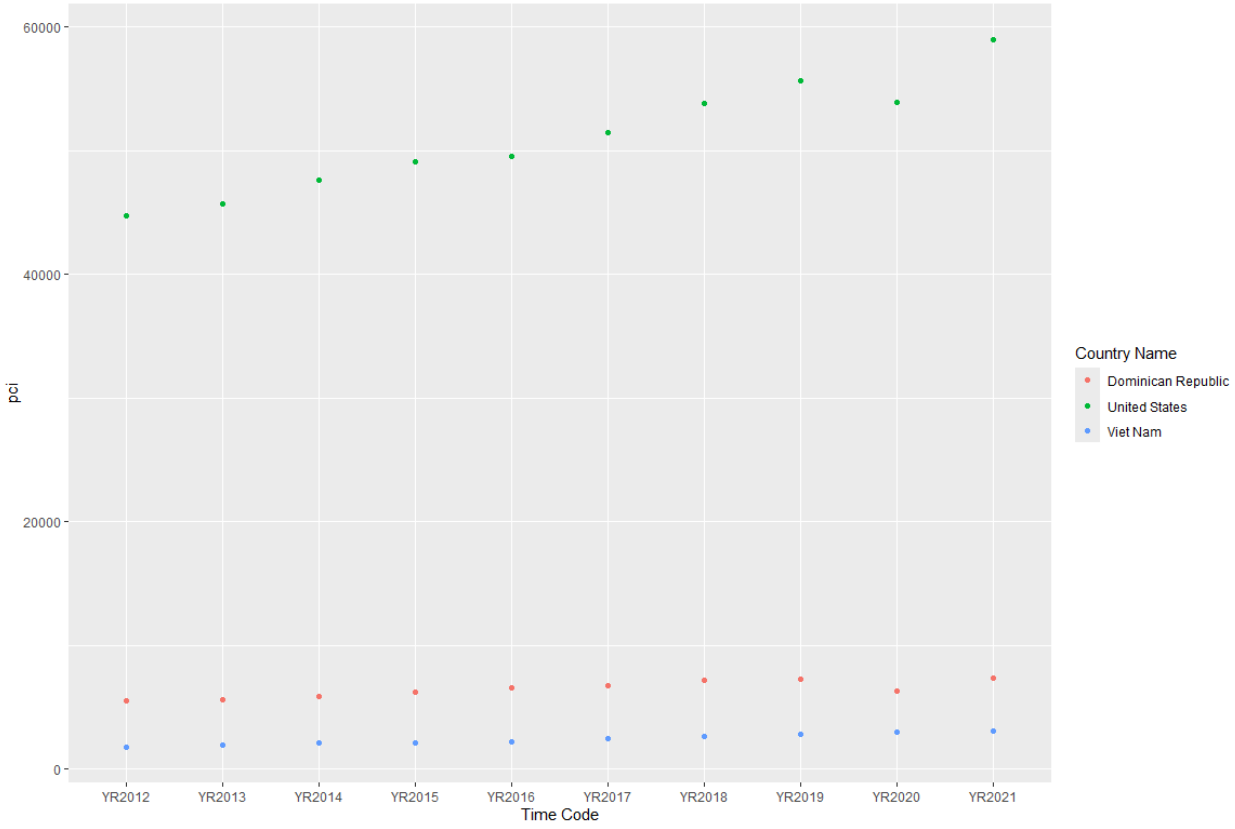
Per capita income, commonly referred to as PCI, represents the measure of the average amount of money earned by each person in a specific geographic area or country. It is used to evaluate the standard of living and quality of life of the population. The limitations of this measure however are that it does not account for wealth inequality, for example, a few wealthy individuals can skew the PCI to be higher than what most people in the region make, and it does not account for savings.

The question being asked is: Is there a difference in the trend in a World Bank development indicator between the three countries? For this report, the per capita incomes of the United States, the Dominican Republic, and Vietnam were investigated. The data was taken from the World Bank website and specifically the World Development Indicator database. It specifically represents the adjusted net national income per capita in current US$. Ten years were taken from the database of the website. The years 2022 and 2023 were available for selection on the website, but they did not have full data, so 2021 was used as the last year and 2012 was used as the first year.

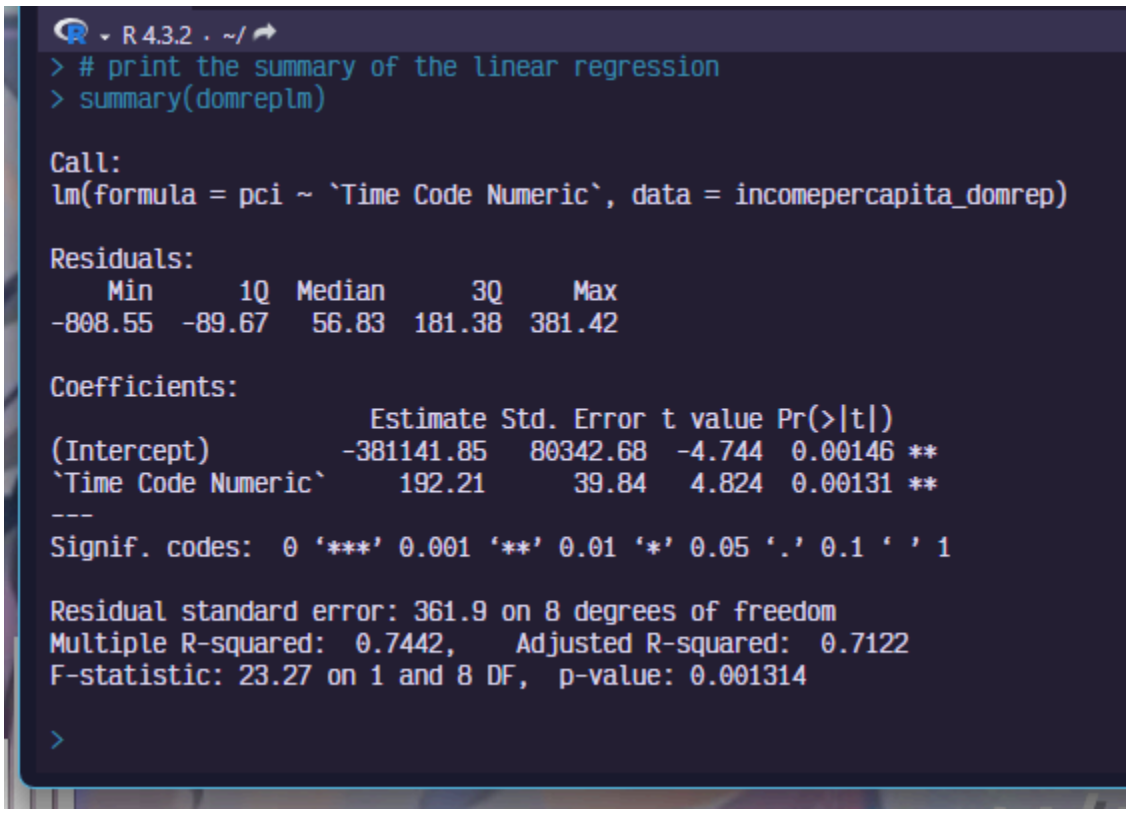
A scatterplot was created to plot the income per capita (y) vs time (x) with the country being a different color. The time data was not numeric, which sufficed for the scatterplot, but it was converted to be numeric inside a new column for the simple linear regression analysis.



From the scatterplot, it is clear that the United States generally has a much higher PCI than both the Dominican Republic and Vietnam, which are similar in PCI. In all years, the United States is much higher than the other two countries, and the Dominican Republic is slightly higher than

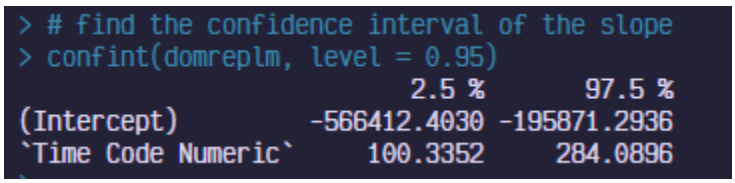
Vietnam. In addition, the Dominican Republic and Vietnam follow the same pattern of increasing and decreasing, they are always generally the same distance away from each other each year, while the United States sees sharper increases in its PCI between the years. It is noteworthy to mention that each country had a decrease in PCI in 2020, likely due to the COVID-19 pandemic slowing down business opportunities as less people left home and thus less people gave money to businesses.

Next, a simple linear regression analysis was ran in R for the Dominican Republic. The output is shown below. The simple linear regression equation to predict PCI based on time is $Y = 192.21x - 381141.85$. This indicates that each year, the PCI increases by $192.21. The y-intercept means that at year 1 CE, the PCI would be $-381,141.85, but this is not practical as the country did not exist then.

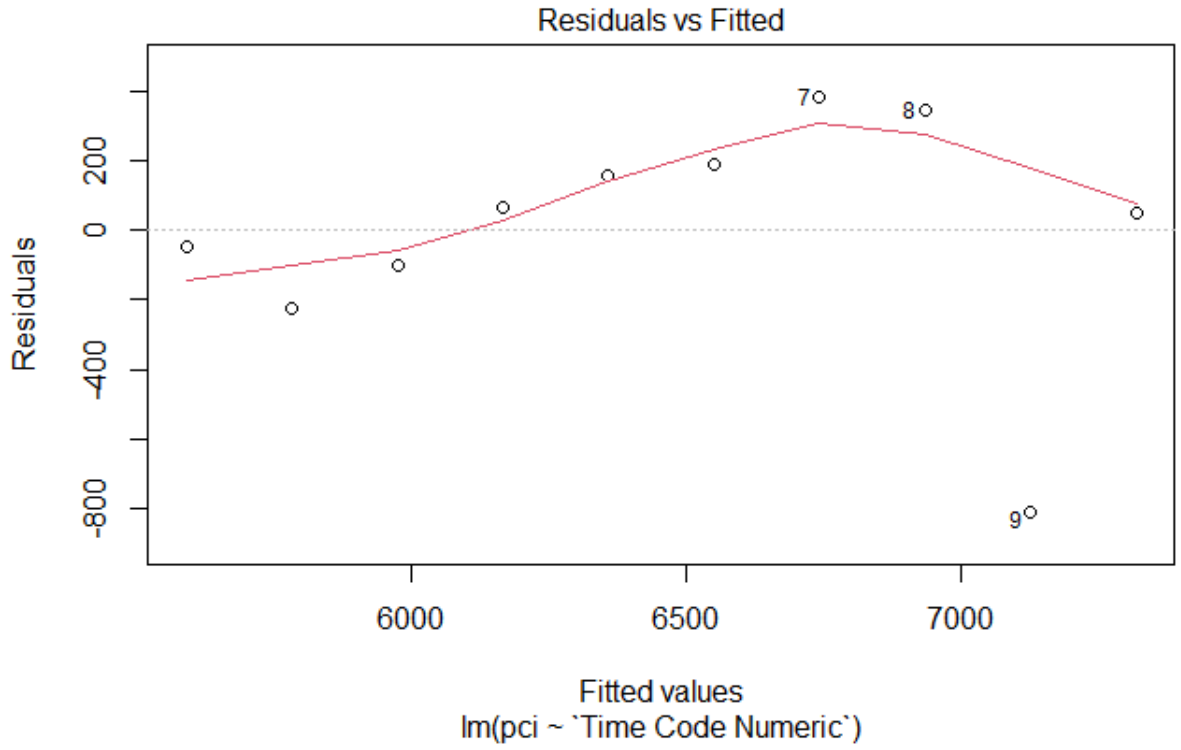


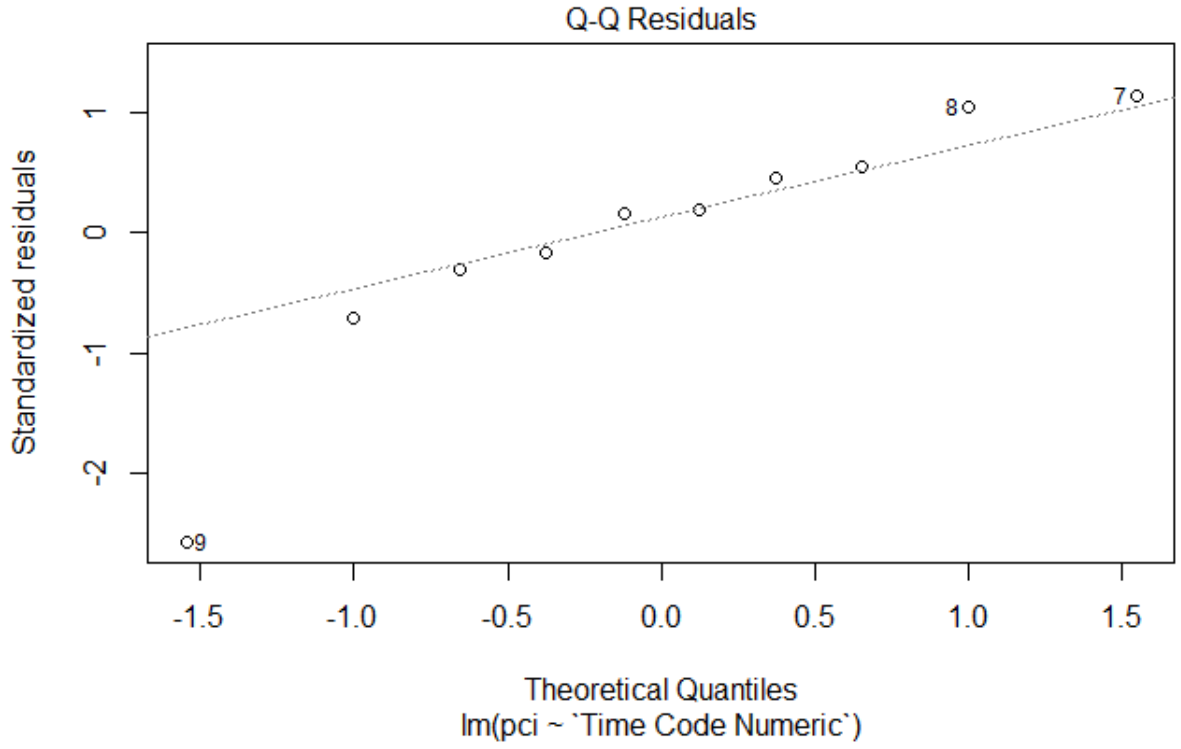
For the t-test for the slope, the test statistic is t = 4.824 and p = 0.00131 which is less than alpha = 0.05, which indicates that the slope of time code numeric is a statistically significant value.

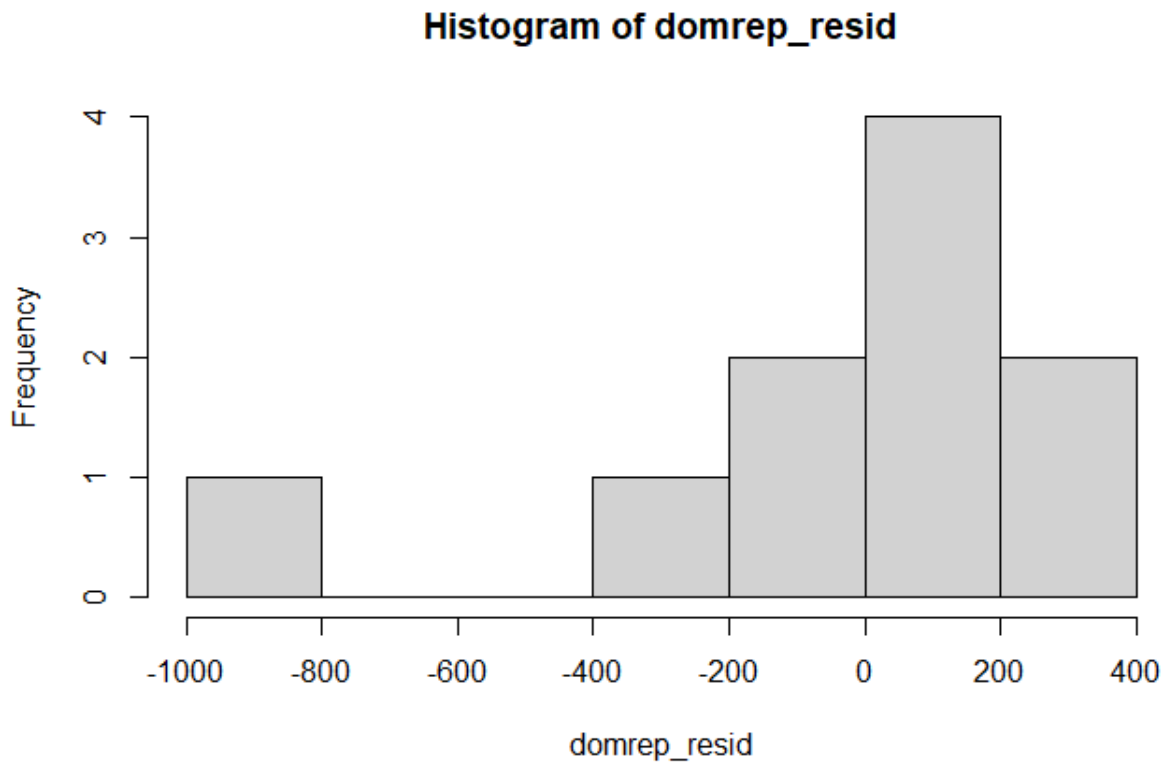
For the confidence interval for the slope, we are 95% confident that the yearly change in per capita income is between $100.34 and $284.09.



For the assumptions, the assumption of linearity may be violated because it looks like there is a positive linear trend in the residuals vs. fitted plot. The zero mean assumption is met because the mean is essentially zero. Independence and randomness are met because the PCI is taken as an average of people in a country. The normality assumption is met as the points follow the line in the Q-Q plot. Yet it seems there are outliers in the histogram of the residuals, making the data appear right skewed. The plots and output are shown below.




Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(pci ~ `Time Code Numeric`)
9
8
7
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
-2
-1
0
1


Histogram of domrep_resid
Frequency
domrep_resid
-1000
-800
-600
-400
-200
0
200
400
0
1
2
3
4

```
> mean(domrep_resid)
[1] 1.492244e-14
>
```

```
> predict.lm(pcidomrep, newx, interval = "confidence")
       fit      lwr      upr
1 7511.691 6941.608 8081.774
```

```
> predict.lm(pcidomrep, newx, interval = "prediction")
       fit      lwr     upr
1 7511.691 6501.042 8522.34
```

R Code:

```
# invoke the tidyverse library
library(tidyverse)

# plot the adjusted net national income per capita (y) vs time (x) with the country being a different color
ggplot(data = incomepercapita) +
  geom_point(mapping = aes(x=`Time Code`, y = `pci`, color = `Country Name`))

# simple linear regression analysis for Dominican Republic

# filter the dataset to only Dominican Republic
incomepercapita_domrep <- incomepercapita %>% filter(`Country Name` == "Dominican Republic")

# create a new column since the time code column has yr next to the numbers so it is seen as a row of characters
incomepercapita_domrep$`Time Code Numeric` <- as.numeric(gsub("YR", "", incomepercapita_domrep$`Time Code`))

# Create a simple linear regression equation to predict the indicator based on time.
domreplm <- lm(`pci` ~ `Time Code Numeric`, data = incomepercapita_domrep)

# print the summary of the linear regression
summary(domreplm)

# find the confidence interval of the slope
confint(domreplm, level = 0.95)

# Find the confidence interval and the prediction interval for the next year after the data was collected.
y <-incomepercapita_domrep$pci
x <- incomepercapita_domrep$`Time Code Numeric`
```

```
pcidomrep <- lm(y ~ x)
newx = data.frame(x = 2022)
predict.lm(pcidomrep, newx, interval = "confidence")
predict.lm(pcidomrep, newx, interval = "prediction")

# check the assumptions using the residuals
plot(domreplm)
domrep_resid <-resid(domreplm)
hist(domrep_resid)
mean(domrep_resid)
```