Executive Summary

The decline in the population growth rate has become a critical issue in many countries worldwide. This study aims to investigate the correlated factors in the decline of the population growth rate. In this report, we performed two studies:

1. Linear regression analysis: investigation relationship between the internet users rate in a range of countries and the population growth rate.

2. Hypothesis analysis: investigation whether there is a significant difference in the population growth rate between difference groups of countries by the score of Human Development Index (HDI).

Three publicly available datasets from Gapminder [1] were used in this study, comprising of the internet users rate, the population growth rate and the Human Development Index (HDI). Sample studies were conducted consisting of countries from around the world following two methodologies. Firstly, a simple regression analysis was conducted between the internet users and population growth rate. Then the Welch's Analysis of Variance (ANOVA) test was applied to investigate the mean differences between the HDI and the population growth rate.

Results of the first study showed a negative correlation between the internet user rate and the population growth rate, indicating that technology development has some relationship to demographic development. Results of the Welch's ANOVA test demonstrates a significant difference in mean population growth rate between countries with HDI scores lower than 0.700. However, no difference is found in mean population growth rate when HDI scores are higher than 0.700.

Our findings indicate that technology may have detrimental effects on society. Therefore, society urgently needs to devise and implement appropriate strategies to mitigate these negative impacts. Additionally, developing nations should enhance three factors of the HDI measure to address social issues associated with overpopulation. Furthermore, policymakers in developed countries might need to look for other factors that affect population growth. In conclusion, these two statistical analyses offer valuable insights for policymakers concerned with demographic development.

1. Introduction

Demographic development has been a crucial topic for many years with some countries facing a decline in population rate while others struggle to address overpopulation. The United Nation (UN) report [2] pointed out that the decline is mainly driven by decreasing fertility levels which is the most demographic policies concentrate on. This phenomenon has raised concerns among policymakers and researchers, as it may have significant implications for these countries' social and economic development [3]. Firstly, there has been a long-standing debate about whether technology, especially the internet, affects fertility levels. Several clinical studies [4-6] have suggested that wireless technology has a negative impact on male sperm quality. Likewise, a study [7] conducted in China has shown that women's fertility intentions are significantly affected by their internet usage. Secondly, worries are across different countries. On the one hand, developing countries are considering that overpopulation growth could impede the growth of the country and produce many societal issues. On the other hand, developed countries worry that the slowdown of population

growth result in serious issues such as population aging, labour and skill shortage [2]. The objectives of our investigation are:

- Provide important insights into society on whether the growth of the internet population is a potential factor of the declining population growth rate.

- Finding whether there is a significant difference in the population growth rate between developing and developed countries.

2. Data

To investigate our objectives, we chose three indicators from Gapminder [1]:

1. Population growth (annual %):

We assume that the population growth of a country is primarily affected by fertility level. One may be argued that the birth rate might be a better indicator for investigating fertility levels. However, due to insufficient data, birth rate data set is unable to march our goal.

2. Individuals using the Internet (% of population)

The recent report [8] shows that there is a total 5.16 billions of internet users worldwide, which around 64% of the world population has a mobile phone, 92.3% of users use mobile phone to access the internet. Therefore, in this study, we assume that most people access the internet via mobile devices.

3. Human Development Index (HDI)

The HDI score is judged by the health, education, and material well-being of a country. This score allows us to understand a country's developmental level. We assume that all data correctly reflect each country's situation.

2.1 Data Pre-processing

We are going to perform two studies across different countries, so a common year was chosen as the indicator to ensure the fairness of the data. Given the steady and consistent growth of technology, choosing a year closer to the present allows for better quality and accuracy of data on the internet users rate. Additionally, we excluded the years after 2020 due to some extreme events that created huge uncertainty, such as Covid-19 and the war in Ukraine. Therefore, we will select one year from 2015-2019. After removing all missing data points in each year, our analysis *Table 1* indicates that 2017 has the most sufficient points among the years selected and was used in our data analysis.

Table 1: available sample points of each data set

Year	HDI	Internet Users Rate	Population Growth Rate
2015	188	200	216
2016	188	203	216
2017	189	205	216
2018	189	163	216
2019	189	157	216

Table 2 given overview of three data sets. The population growth rate is our only dependent variable which will be presented in both studies. As a result, we built two data frames for two difference studies. *Table 3* shows the summary statistics for the linear regression analysis.

Table 2: Types of Variables in the analysis

Data	Type	Description
Population Growth Rate (%)	Numeric and Continuous	Dependent variable
Internet Users Rate (%)	Numeric and Continuous	Independent variable
Human Development Index	Categorial and ordinal	independent variable

Table 3: Summary of Date frame for regression analysis

Internet User Rate 2017 (%)			Population Growth Rate 2017 (%)			Sample (N)
Mean	**Median**	**Range**	**Mean**	**Median**	**Range**	
54.47	69.7	1.31 ~ 99.5	1.2251	1.17	-2.42 ~ 4.68	204

We categorized our HDI dataset by adopting the UN's category methods [9]: low, medium, high, very high. What's more, we divided high and very high pre-define group into three subgroups: somewhat high, high, very high. It based on a gap of 0.080. This classification method not only allows us for a more in-depth investigation of the relationship but also gets the similar sample size for each group. *Table 4* shows the summary statistics for our hypothesis study.

Table 4: Summary of Data frame for hypothesis study

Human Development Index 2017			Population Growth Rate 2017 (%)	
HDI Score	**Group Type**	**Sample (N)**	**Mean**	**Median**
>0.860	5 - Very High	36	2.530	2.660
0.789~0,859	4 - High	34	1.700	1.660
0.700~0.779	3 - Somewhat High	45	0.986	1.100
0.550~0.699	2 - Medium	40	0.805	0.492
<0.550	1 - Low	33	0.700	0.612

2.3 Limitation of the Data

Even though the data we used comes from a reliable and trustworthy source, it is important to highlight that there are still limitations that need to be acknowledged. Firstly, we only used data from a certain timeframe, which may limit our analysis to specific developmental environments. Hence, up-to-data study from the new data needs to be conducted. Secondly, the data may subject to bias or limitations, such as the possibility of bias in multicultural countries. Thirdly, the variables measured in the data may also be limited, which can affect the accuracy and applicability of our findings. For example, a country has a large of the population will only contribute to one data point in the data set. Finally, by recognising the limitations of the data, we ensure that our conclusions are based on a more general perspective and consider any potential sources of bias or limitations.

3. Linear regression between two variables

3.1 Finding Linear Relationship

Our first study is to investigate the relationship between the internet user rate and the population growth rate. *Figure 1* given overview of our two variables. The scatter plot shows a potential negative linear relationship between the population growth rate and internet usage rate. The histogram shows the distribution of independent variable x is not a normal distribution while the boxplot shows some outliers in the dependent variable y. Because we are conducting a linear regression analysis, we assume that the predictor variable x has no random process, but we do need to remove outliers.

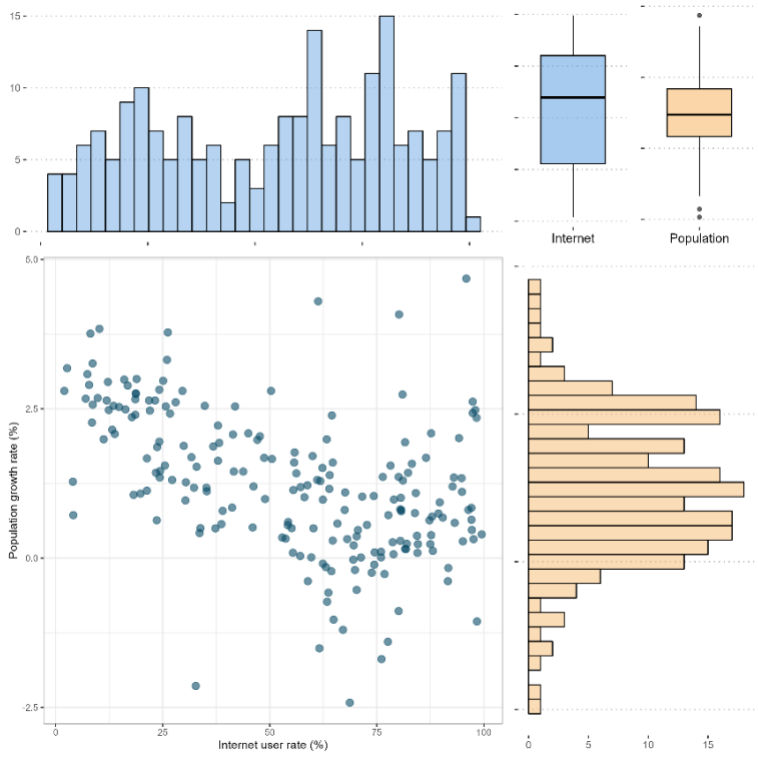

Figure 1: Combination plot of Internet users rate and Population growth rate

We removed three outliers from the population data set. *Figure 2* shows the Pearson correlation coefficient (r) between two data sets. As r is less than -0.50, hence, we assume that the population growth rate and the internet users rate has a negative linear relationship.

```
# Pearson correlation coefficient
## [1] -0.5472055
```

Figure 2: output of Pearson correlation coefficient in R Studio

3.2 Building model

The model summary in *Figure 3* shows that p_r < 2e-16, which indicates that there is a relationship between two variables. However, R^2 = 0.2994, which suggests the model may not fit data very well. It also shows two important variables of our model: β_0= 2.4915, β_1= -0.0227.

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.491546   0.150903  16.511   <2e-16 ***
## users_rate  -0.022717   0.002457  -9.246   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9796 on 200 degrees of freedom
## Multiple R-squared:  0.2994, Adjusted R-squared:  0.2959
## F-statistic: 85.48 on 1 and 200 DF,  p-value: < 2.2e-16
```

Figure 3: Simple Linear Model Overview in R Studio

Finally, we can build up a simple linear model *Equation (1)*:

$$y_i = 2.4915 - 0.0227x_i + e_i \quad (1)$$

y_i is response variable, x_i is predictor variable, e_i is the residuals error.

3.3 Assumptions

To make sure the model is reliable, three assumptions of residuals error must meet. They are Independence, normality, and homoscedasticity.

3.3.1. Independence

The Durbin-Watson (D-W) test can tell us that if there is any autocorrelation of residuals in our regression model [10]. The result on *Figure 4* shows p-value is greater than 0.05. This indicates that we should accept null hypothesis of Durbin-Watson test – no first order autocorrelation, which suggests that the autocorrelation of the residuals is not likely to affect our findings. In other words, we assume that the residuals are independent of each other.

```
library(car)
durbinWatsonTest(model)

##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1150268      2.228172   0.106
##  Alternative hypothesis: rho != 0
```

Figure 4: Output of Durbin Watson Test in R Studio

3.3.2. Normality

To check normality, the easier way is to use the quantile-quantile (Q-Q) plot and histogram. As *Figure 5* shows, we assume that the residuals are approximately normal.

3.3.3. Homoscedasticity

Figure 6 shows that residuals are distributed evenly between the red line, and the residuals also equally spread in across fitted values. Therefore, we assume that the residual is homoscedastic.

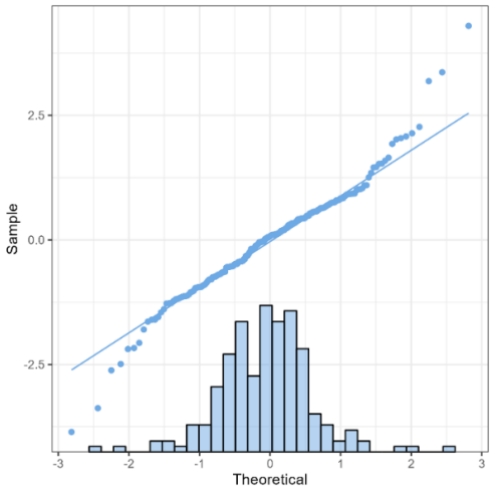

Figure 5: Q-Q Plot and histogram of residuals

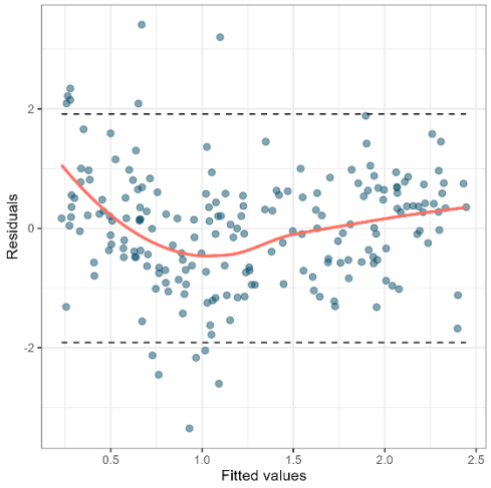

Figure 6: Residuals vs Fitted Values

3.3 Prediction interval and Confidence interval

Because our assumptions are not violated, we can continue to find the prediction interval. According to our model, we found prediction and confidence interval in 95% significant level. In the output *Figure 7*, we can see that the difference of prediction and confidence interval is very large, that indicates the prediction might not precise.

```
# explore two intervals
summary(pi)
##       fit              lwr               upr
##  Min.   :0.3073   Min.   :-1.7949   Min.   :2.410
##  1st Qu.:0.7161   1st Qu.:-1.3774   1st Qu.:2.810
##  Median :1.1427   Median :-0.9468   Median :3.232
##  Mean   :1.2506   Mean   :-0.8438   Mean   :3.345
##  3rd Qu.:1.8180   3rd Qu.:-0.2760   3rd Qu.:3.912
##  Max.   :2.3616   Max.   : 0.2544   Max.   :4.469
summary(ci)
##       fit              lwr               upr
##  Min.   :0.3073   Min.   :0.03287   Min.   :0.5818
##  1st Qu.:0.7161   1st Qu.:0.51949   1st Qu.:0.9126
##  Median :1.1427   Median :0.99440   Median :1.2910
##  Mean   :1.2506   Mean   :1.04893   Mean   :1.4522
##  3rd Qu.:1.8180   3rd Qu.:1.61590   3rd Qu.:2.0201
##  Max.   :2.3616   Max.   :2.05136   Max.   :2.6718
```

Figure 7: Summary of Prediction interval and Confidence interval

Finally, *Figure 8* shows our result on the linear regression analysis. It illustrates the 95% significant level of prediction and confidence interval. There are only few data points out of the prediction interval. Therefore, the simple linear model is fitted to our data. However, the wide prediction interval indicates that our linear model may not predict the future value precisely.

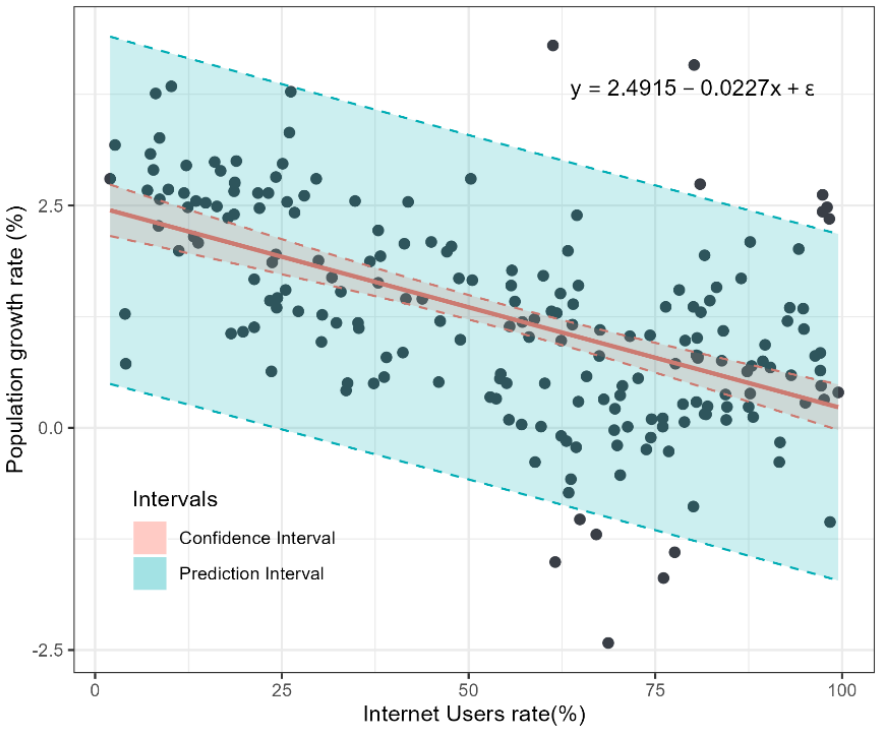

Figure 8: Scatter plot of Internet users rate vs population growth rate in 2017 (by countries)

.

4. Statistical Investigation of a hypothesis

4.1 Hypothesis Statement

In this investigation, we want to know if there is any difference in population growth rate between developed and developing countries. Assume μ_i (i is the group number) is the true mean of each group, our null hypothesis and alternative hypothesis are *Equation (2)*:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5,\ H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \qquad (2)$$

4.2 Explore Data

Figure 9 reveals some outliers that may impact our findings. Although these countries could be outliers for various reasons, external factors beyond the data are not included in this investigation. While retained few outliers that close to the tails to ensure a sufficiently large sample size, we removed eleven outliers. *Table 5* shows the size of each group after the removal.

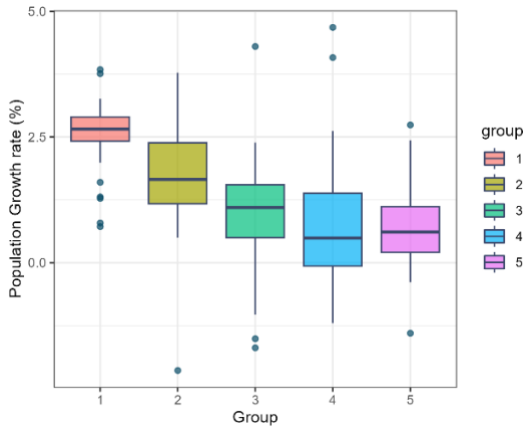

Figure 9: Box plot of each group

Table 5: Size of each group after removal of outliers

Group	Size
1	31
2	39
3	41
4	32
5	34

4.3 Assumptions

To analyse the differences in means of the five country groups, certain assumptions must be met prior to testing:

- All data points are independent

- the dependent variable in each group is normally distributed

- the dependent variable between groups has equal variance

4.3.1 Independence

While we are unable to verify the independence of each country's data, considering the sources are obtained from reputable organizations, thus we assume all data points are independence.

4.3.2 Normality

Figure 10 presents the quantile-quantile (Q-Q) plot of each group, all of them are approximately on the Q-Q line, so we assume that all our five groups of samples are normally distributed.

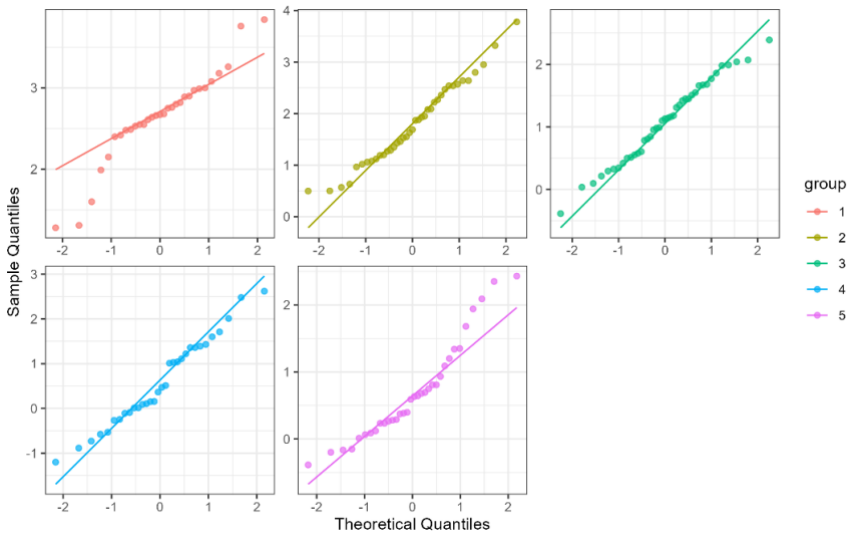

Figure 10: Normal Q-Q plot of each group

4.3.3 Homogeneity

To verify the homogeneity of variance among the groups, the Bartlett test is a useful tool for variance analysis involving more than two groups [11]. The test result in *Figure 11* indicates that p-value less that 0.05, giving the rejection of null hypothesis in the Bartlett test. Therefore, running a parametric test may produce an unreliable outcome. In such cases, a non-parametric test is the preferable option.

```
# variance check
bartlett.test(growth_rate ~ group, hdi_pop)
##  Bartlett test of homogeneity of variances
##
## data:  growth_rate by group
## Bartlett's K-squared = 8.9518, df = 4, p-value = 0.04832
```

Figure 11: Output of Bartlett Test In R Studio

4.4 Welch's Analysis of Variance (ANOVA) Test

Since the normality assumption holds, but there is an unequal variance among groups. We selected Welch's ANOVA as our method, because Welch's ANOVA is the preferred option for analysing heterogeneous groups and comparing means [12].

The output of Welch's ANOVA *Figure 12* given p < 2.2e-16. This suggests that we should reject the null hypothesis and accept the alternative hypothesis. Therefore, there is a significant difference in mean population growth rate among the different groups of HDI data.

```
# Welch's ANOVA
oneway.test(growth_rate ~ group, data = hdi_pop, var.equal = FALSE)
##  One-way analysis of means (not assuming equal variances)
##
## data:  growth_rate and group
## F = 50.335, num df = 4.000, denom df = 85.011, p-value < 2.2e-16
```

Figure 12: Output of Welch's ANOVA in R Studio

4.5 Post-hoc Test

The box plot in *Figure 9* given that the mean difference between group 5 and group 1 may be responsible for the rejection of null hypothesis. This might cause Type I error. To determine exactly where the differences from, a post-hoc test is necessary. We selected Games-Howell as our post-hoc test because it is a conservative and robust method for minimizing the risk of Type I error in groups with heterogeneous variances [13]. *Figure 13* shows the output of Games-Howell at a confidence level of 95%. The last column indicates significant difference of

```
games_howell_test(hdi_pop, growth_rate ~ group,
                        conf.level = 0.95, detailed = FALSE)
## # A tibble: 10 × 8
##    .y.         group1 group2 estimate conf.low conf.high    p.adj p.adj.signif
##  * <chr>       <chr>  <chr>     <dbl>    <dbl>     <dbl>    <dbl> <chr>
##  1 growth_rate 1      2        -0.852   -1.31     -0.395 1.80e- 5 ****
##  2 growth_rate 1      3        -1.61    -2.04     -1.19  0        ****
##  3 growth_rate 1      4        -2.06    -2.63     -1.50  0        ****
##  4 growth_rate 1      5        -1.89    -2.36     -1.41  2.02e-11 ****
##  5 growth_rate 2      3        -0.763   -1.24     -0.290 2.22e- 4 ***
##  6 growth_rate 2      4        -1.21    -1.81     -0.609 4.57e- 6 ****
##  7 growth_rate 2      5        -1.03    -1.55     -0.513 4.35e- 6 ****
##  8 growth_rate 3      4        -0.449   -1.03      0.130 2   e- 1 ns
##  9 growth_rate 3      5        -0.271   -0.762     0.221 5.39e- 1 ns
## 10 growth_rate 4      5         0.179   -0.438     0.795 9.25e- 1 ns
```

*"****" and "***" indicate significant, "ns" indicate not significant.*

Figure 13: Output of Games Howell Test in R-Studio.

each pair group, the last three row shows not significant. *Figure 14* visualizes the result from Games-Howell test. It shows the mean difference is getting smaller in higher HDI score pair groups.

The study result demonstrates a significant difference in mean population growth rate between countries with low and medium HDI scores. However, not enough evidence shows that there is a significant difference in mean population growth rate among countries with HDI scores higher than 0.700.

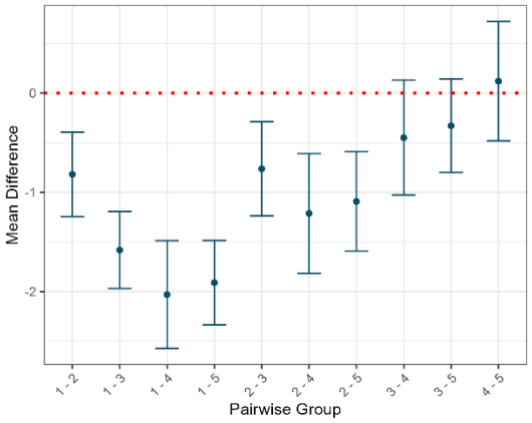

Figure 14: Pairwise Test Mean Difference in HDI Group

5. Concluding Remarks

As the result of linear regression analysis, we conclude that there is a negative correlation between the internet user rate and the population growth rate. It suggests that the internet may physically or mentally influence on a country's fertility level. However, this is not to imply the causation, many factors can affect them in the same way, such as economic pressure, better contraception, and so on. Furthermore, the analysis also given that the model is not applicable to some countries. Hence further studies need to be conducted on other factors and the outliers.

The findings of the hypothesis study provide valuable insights for governments and societies regarding the development of countries. Specifically, some developing countries (HDI < 0.700) that worry over population should focus on improving the HDI measures of health, education, and material well-being. Once these factors are improved, most the issue regarding overpopulation will resolve accordingly. On the contrary, when a country's HDI score greater than 0.700, there is not enough evidence to suggest that the further increase on HDI score will lead to a reduction in population growth. Therefore, government sectors may need to consider other factors while continuing to improve their systems.

These two studies only investigate the data from the year of 2017, and exclude other potential influential factors such as culture, politics and geography. Further studies can concentrate on specific regions or other timeframes.

Reference

[1] Gapminder. https://www.gapminder.org/data/ (accessed 20 April, 2023).

[2] D. UN, "World population prospects 2019: highlights," *United Nations Department for Economic and Social Affairs,* 2019.

[3] M. Herrmann. "The Global Population Will Soon Reach 8 Billion—Then What?" https://www.un.org/en/un-chronicle/global-population-will-soon-reach-8-billion-then-what (accessed 15 APR, 2023).

[4] J. A. Adams, T. S. Galloway, D. Mondal, S. C. Esteves, and F. Mathews, "Effect of mobile telephones on sperm quality: a systematic review and meta-analysis," (in eng), *Environ Int,* vol. 70, pp. 106-12, Sep 2014, doi: 10.1016/j.envint.2014.04.015.

[5] J. J. Oh, S. S. Byun, S. E. Lee, G. Choe, and S. K. Hong, "Effect of Electromagnetic Waves from Mobile Phones on Spermatogenesis in the Era of 4G-LTE," (in eng), *Biomed Res Int,* vol. 2018, p. 1801798, 2018, doi: 10.1155/2018/1801798.

[6] K. Balawender and S. Orkisz, "The impact of selected modifiable lifestyle factors on male fertility in the modern world," (in eng), *Cent European J Urol,* vol. 73, no. 4, pp. 563-568, 2020, doi: 10.5173/ceju.2020.1975.

[7] P. Liu, J. Cao, W. Nie, X. Wang, Y. Tian, and C. Ma, "The influence of internet usage frequency on women's fertility intentions—the mediating effects of gender role attitudes," *International Journal of Environmental Research and Public Health,* vol. 18, no. 9, p. 4784, 2021.

[8] "Digital Around The World." https://datareportal.com/global-digital-overview (accessed 20 April, 2023).

[9] UN, "HUMAN DEVELOPMENT REPORT 2021/2022," 08 SEP 2022. [Online]. Available: https://hdr.undp.org/system/files/documents/global-report-document/hdr2021-22pdf_1.pdf

[10] J. A. Tillman, "The power of the Durbin-Watson test," *Econometrica: Journal of the Econometric Society,* pp. 959-974, 1975.

[11] A. Hilton and R. A. Armstrong, "Statnote 5: Is one set of data more variable than another?," *Microbiologist,* vol. 2006, pp. 34-36, 2006.

[12] M. Delacre, C. Leys, Y. L. Mora, and D. Lakens, "Taking parametric assumptions seriously: Arguments for the use of Welch's F-test instead of the classical F-test in one-way ANOVA," *International Review of Social Psychology,* vol. 32, no. 1, 2019.

[13] A. Hilton and R. A. Armstrong, "Statnote 6: post-hoc ANOVA tests," *Microbiologist,* vol. 2006, pp. 34-36, 2006.

Appendix: R Code

Prerequire Packages

```
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(ggpubr)
```

2.Data

2.1 Data Preprocessing

```
# 1) Import Data ------------------------------------------
hdi <- read_csv("hdi_human_development_index.csv")
int_users_rate <- read_csv("internet_users.csv")
pop_growth_rate <-
  read_csv("population_growth_annual_percent.csv")

#Getting data points from last 2015-2019
hdi_count <- sapply(hdi[, (ncol(hdi) - 4):ncol(hdi)],
                    function(x) sum(!is.na(x)))
int_count <- sapply(int_users_rate[,(ncol(int_users_rate) - 5):
                                     (ncol(int_users_rate) - 1)],
                    function(x) sum(!is.na(x)))
pop_count <- sapply(pop_growth_rate[,
                                     (ncol(pop_growth_rate) - 5):
                                     (ncol(pop_growth_rate) - 1)],
                    function(x) sum(!is.na(x)))
cbind(hdi_count, int_count, pop_count)
rm(hdi_count, int_count, pop_count)

# 2) Data Cleaning and Filtering ----------------------------

int_ur_2017 <- data.frame(
  "country" = int_users_rate$country,
  "users_rate" = as.numeric(gsub("[^0-9.-]", "-",
                                 int_users_rate$"2017")))
# get population growth rate in 2017
pop_gr_2017 <- data.frame(
  "country" = pop_growth_rate$country,
  "growth_rate" = as.numeric(gsub("[^0-9.-]", "-",
                                  pop_growth_rate$"2017")))
# get HDI data in 2017
hdi_2017 <- data.frame(
  "country" = hdi$country,
  "HDI" = hdi$"2017")
# First dataset for regression analysis
int_pop <- na.omit(merge(int_ur_2017,
                         pop_gr_2017,
                         by = "country"))
# Second dataset for hypothesis investigation
hdi_pop <- na.omit(merge(hdi_2017,
                         pop_gr_2017,
                         by = "country"))
# remove unused dataframe
rm(int_ur_2017, hdi_2017, pop_gr_2017)
rm(hdi, int_users_rate, pop_growth_rate)

# Categorate HDI
hdi_pop[hdi_pop$HDI >= 0.860,
        "group"] <- "5"
hdi_pop[hdi_pop$HDI <= 0.859 & hdi_pop$HDI >= 0.780,
        "group"] <- "4"
hdi_pop[hdi_pop$HDI <= 0.779 & hdi_pop$HDI >= 0.700,
        "group"] <- "3"
hdi_pop[hdi_pop$HDI <= 0.699 & hdi_pop$HDI >= 0.550,
        "group"] <- "2"
hdi_pop[hdi_pop$HDI < 0.550,
        "group"] <- "1"
hdi_pop %>% count(group)
# remove HDI column
hdi_pop$HDI <- NULL
```

3.Regression Analysis

3.1 Finding Linear Relationship

```
# 1) Plot overview data scatter, box, histogram ---------------
p1 <- gghistogram(int_pop$users_rate, fill = "#6DA9E4",
                  xlab = "Internet Users Rate (%)") +
      theme_pubclean() +
      theme(axis.title.x = element_blank(),
            axis.title.y = element_blank(),
            axis.text.x = element_blank())
p2 <- gghistogram(int_pop$growth_rate, fill = "#F6BA6F",
                  xlab = "Population Growth Rate (%)") +
      theme_pubclean() +
      rotate() +
      theme(axis.title.x = element_blank(),
            axis.title.y = element_blank(),
            axis.text.y = element_blank())
p3 <- ggscatter(int_pop, "users_rate","growth_rate",
          size = 3, alpha = 0.6,
          color = "#0A4D68") +
      labs(x = "Internet user rate (%)",
           y = "Population growth rate (%)") +
      border()
b1 <- ggboxplot(int_pop$users_rate,
                fill = "#6DA9E4",
                xlab = "Internet") +
      theme_pubclean() +
      theme(axis.text.x = element_blank(),
            axis.title.y = element_blank(),
            axis.text.y = element_blank())
b2 <- ggboxplot(int_pop$growth_rate,
                fill = "#F6BA6F",
                xlab = "Population") +
      theme_pubclean() +
      theme(axis.text.x = element_blank(),
            axis.title.y = element_blank(),
            axis.text.y = element_blank())
bb <- ggarrange(b1, b2,
                ncol = 2, nrow = 1,
                widths = c(1,1), heights = c(1, 1))

# Figure 1
ggarrange(p1, bb, p3, p2,
          ncol = 2, nrow = 2,
          widths = c(3,1.5), heights = c(1.5, 3))

ggsave("Fig_1.png", width = 10, height = 10)
rm(p1, p2, p3, bb, b1, b2)

# 2) Pearson correlation coefficient and remove outliers---------

# Pearson correlation coefficient before remove outliers
cor(int_pop$users_rate, int_pop$growth_rate)
# with spearman before remove outlier
cor(int_pop$users_rate, int_pop$growth_rate,
    method = "spearman")
#  remove outliers ---------
outlier_max <- sort(int_pop$growth_rate, decreasing = T)[1]
int_pop <- int_pop[-which(int_pop$growth_rate == outlier_max), ]
outlier_min <- sort(int_pop$growth_rate)[1:2]
int_pop <- int_pop[-which(int_pop$growth_rate == outlier_min), ]
rm(outlier_max, outlier_min)
# Pearson correlation coefficient after remove outliers
cor(int_pop$users_rate, int_pop$growth_rate)
#with spearman after remove outlier
cor(int_pop$users_rate, int_pop$growth_rate,
    method = "spearman")
```

3.2 Simple linear model

```
# 1) Build up model-----------------------------------------
model <- lm(growth_rate ~ users_rate, data = int_pop)
# Explore the model
library(broom)
tidy(model)
augment(model)
summary(model)
anova(model)

# 2) Independent check--------------------------------------
library(car)
durbinWatsonTest(model)
```

```
# 3) Residuals Normality Check-----------------------------
# Figure 2
plot_qq <- ggqqplot(model$residuals,
                    add = c("qqline"),
                    conf.int = F,
                    color = "#6DA9E4",
                    #title = "Normal QQ plot & Histogram"
                    )
plot_hist <- ggplotGrob(gghistogram(model$residuals,
                                    fill = "#6DA9E4",
                                    xlab = "Resisduals") +
                                    theme_transparent())
plot_qq + annotation_custom(grob = plot_hist,
                            xmin = -3, xmax = 3,
                            ymin = -4.65, ymax = -1) + theme_bw()

ggsave("Fig_2.png", width = 5, height = 5)
rm(plot_qq, plot_hist)

# 4)Homoscedasticity check---------------------------------
df <- data.frame(.fitted = model$fitted.values,
                 .resid = model$residuals)
sd_resid <- sd(df$.resid)
df$upper_bound <- 1.96 * sd_resid
df$lower_bound <- -1.96 * sd_resid
# Figure 3
ggplot(df, aes(x = .fitted, y = .resid)) +
      geom_point(color = "#0A4D68", alpha = 0.5, size = 2) +
      geom_smooth(method = "loess", span = 0.75,
                  color = "#FF6D60", se = FALSE) +
      geom_line(aes(y = upper_bound), linetype = "dashed") +
      geom_line(aes(y = lower_bound), linetype = "dashed") +
      theme_bw() +
      labs(x = "Fitted values", y = "Residuals")

ggsave("Fig_3.png", width = 5, height = 4)
rm(df, sd_resid)
```

3.3 Prediction interval and Confidence interval

```
# 1) Find Interval-----------------------------------------
confint(model)
pi <- predict(model, int = "pred")
ci <- predict(model, int = "con")
# explore two intervals
summary(pi)
summary(ci)

# 2) Plot Linear Regression scatter plot with Interval----------
ggplot(int_pop, aes(x = users_rate, y = growth_rate)) +
      geom_point(col = "#393E46",size = 2) +
      geom_line(aes(y = pi[,1]), col = "#FF6D60") +
      geom_line(aes(y = pi[,2]), col = "#00ADB5",
                linetype = "dashed") +
      geom_line(aes(y = pi[,3]), col = "#00ADB5",
                linetype = "dashed") +
      geom_line(aes(y = ci[,2]), col = "#FF6D60",
                linetype = "dashed") +
      geom_line(aes(y = ci[,3]), col = "#FF6D60",
                linetype = "dashed") +
      geom_smooth(method = "lm", se = FALSE,
                  color = "#FF6D60",
                  formula = y ~ x) +
     annotate("text", x = 80, y =3.8,
           label = expression(y == 2.4915 + 0.0227*x + epsilon),
           parse = TRUE, color = "black") +
      scale_fill_manual(name = "Intervals",
                        values = c("#FF6D60", "#00ADB5")) +
      geom_ribbon(aes(ymin =ci[,2], ymax = ci[,3],
                      fill = "Confidence Interval"),
                  alpha = 0.2 , linetype = "dashed") +
      geom_ribbon(aes(ymin =pi[,2], ymax = pi[,3],
                      fill = "Prediction Interval"),
                  alpha = 0.2, linetype = "dashed") +
      theme_bw() +
      theme(legend.position = c(0.2, 0.2),
            legend.background =
              element_rect(fill ="transparent")) +
      labs(x = "Internet Users rate(%)",
           y = "Population growth rate (%)")
ggsave("Fig_4.png", width = 6, height = 5)
```

4 Hypothesis Test

4.1 Assumptions

```
# 1) Boxplot-----------------------------------------------
# Figure 5
ggplot(hdi_pop,
       aes(x = group, y = growth_rate, fill = group,)) +
       geom_boxplot(outlier.colour = "#0A4D68",
              alpha = 0.7, color= "#3C486B") +
       labs(x = "Group", y = "Population Growth rate (%)") +
       theme_bw()

ggsave("Fig_5.png", width = 5, height = 4)

# 2) Remove totally 11 outliers----------------------------
for (i in c("1", "2", "3", "4", "5")) {
  assign(paste0("group_", i),
         subset(hdi_pop$growth_rate, hdi_pop$group == i))
}
md <- sort(group_2)[1]
hdi_pop <- hdi_pop[-which(hdi_pop$growth_rate %in% md), ]
md <- max(group_3)
hdi_pop <- hdi_pop[-which(hdi_pop$growth_rate %in% md), ]
md <- sort(group_3)[1:3]
hdi_pop <- hdi_pop[-which(hdi_pop$growth_rate %in% md), ]
md <- sort(group_4, decreasing = T)[1:2]
hdi_pop <- hdi_pop[-which(hdi_pop$growth_rate %in% md), ]
md <- sort(group_5)[1]
hdi_pop <- hdi_pop[-which(hdi_pop$growth_rate %in% md), ]
md <- sort(group_5, decreasing = T)[1]
hdi_pop <- hdi_pop[-which(hdi_pop$growth_rate %in% md), ]
md <- sort(group_1)[1:2]
hdi_pop <- hdi_pop[-which(hdi_pop$growth_rate %in% md), ]

hdi_pop %>% count(group)

rm(md,group_1,group_2,group_3,group_4,group_5, i)

# 3) Normality with Q-Q plot-----------------------------------
# Figure 6
ggplot(hdi_pop,
       aes(sample = growth_rate,
           colour = group)
           ) +
       geom_qq() +
       theme_bw() +
       facet_wrap(~group, scales = "free") +
       geom_qq_line() +
       theme(strip.background = element_blank(),
             strip.text.x = element_blank()) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles")

ggsave("Fig_6.png", width = 8, height = 6, dpi = 300)

# 4) homogeneity of variances test-----------------------------
bartlett.test(growth_rate ~ group, hdi_pop)
```

4.2 Hypothesis test

```
# 1) Welch's ANOVA-----------------------------------
oneway.test(growth_rate ~ group, data = hdi_pop,
            var.equal = FALSE)

# 2) Pos-hod test-----------------------------
library(rstatix)
gh <- games_howell_test(hdi_pop, growth_rate ~ group,
                        conf.level = 0.95, detailed = FALSE)
gh
gh$group <- paste(gh$group1,"-", gh$group2)
# Figure 7
ggplot(gh,
       aes(x = group, y = estimate,
           ymin = conf.low, ymax = conf.high)) +
       geom_errorbar(width = 0.5, color = "#0A4D68") +
       geom_point(size = 1.5, color = "#0A4D68") +
       geom_hline(yintercept = 0, size =1,
             linetype = "dotted", col = "red") +
       theme_bw() +
       theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
       labs(x = "Pairwise Group", y = "Mean Difference")
ggsave("Fig_7.png", width = 5, height = 4, dpi = 300,)
```