Hwk5_bs2996

Bangda Sun

October 25, 2016

Part 1: Data for the US

i. Write function: percentile_ratio_discrepancies.

```
percentile_ratio_discrepancies <- function(P99, P99.5, P99.9, a){
  # input P99, P99.5, P99.9
  # return the sum square
  sse <- ((P99 / P99.9)^(-a+1) - 10)^2 +
         ((P99.5 / P99.9)^(-a+1) - 5)^2 +
         ((P99 / P99.5)^(-a+1) - 2)^2
  return(sse)
}
# Check the return
percentile_ratio_discrepancies(P99 = 1e6, P99.5 = 2e6, P99.9 = 1e7, a = 2)
```

```
## [1] 0
```

ii: Write function: exponent.multi_ratios_est

```
exponent.multi_ratios_est <- function(P99, P99.5, P99.9){
  # input P99, P99.5, P99.9
  # minimize the percentile_ratio_discrepancies using nlm()
  # initial value a determined by formula (4)
  a   <- (1 - (log(10) / log(P99 / P99.9)))
  est <- nlm(percentile_ratio_discrepancies, a, P99 = P99, P99.5 = P99.5, P99.9 = P99.9)
  return(est$estimate)
}
exponent.multi_ratios_est(P99 = 1e6, P99.5 = 2e6, P99.9 = 1e7)
```

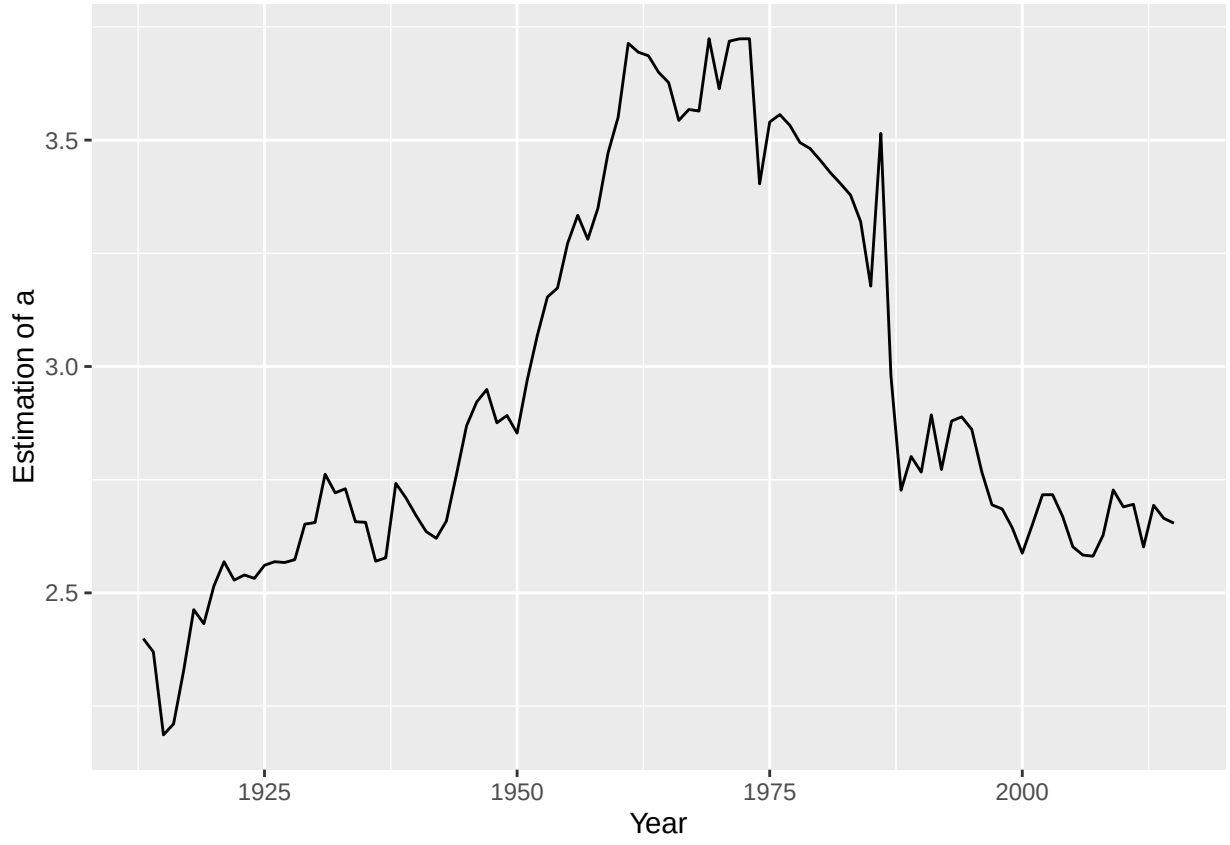
```
## [1] 2
```

iii: Write function and plotting the exponential estimates.

```
# set work directory and import data
setwd("C://Users//Bangda//Desktop//GR5206 Materials//Hwk5")
wtid <- read.csv("wtid-report.csv", header = TRUE)
str(wtid)
```

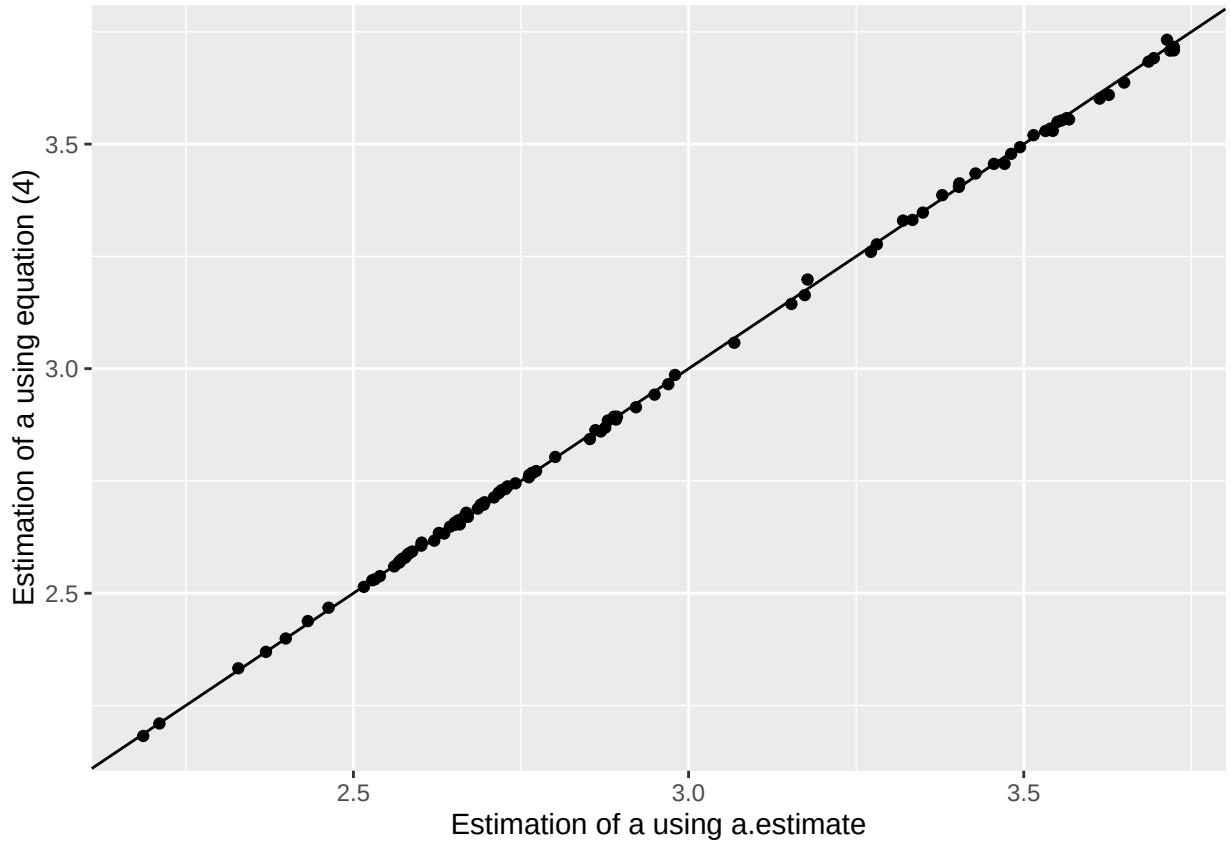
```
## 'data.frame':    103 obs. of  5 variables:
##  $ Country               : Factor w/ 1 level "United States": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Year                  : int  1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 ...
##  $ P99.income.threshold  : num  82677 76406 64409 77290 95327 ...
##  $ P99.5.income.threshold: num  135583 126911 122556 138102 154538 ...
##  $ P99.9.income.threshold: num  428630 410529 451668 518327 536356 ...
```

```r
names(wtid) <- c("Country", "Year", "P99", "P99.5", "P99.9")
# function a.estimate
a.estimate <- function(P99, P99.5, P99.9){
  # input P99, P99.5, P99.9
  # using exponent.multi_ratios_est to estimate each years' a
  a1 <- rep(NA, length(P99))
  for (i in 1:length(a1)){
    # if one of P99, P99.5, P99.9 is missing, return NA
    if (is.na(P99[i]) | is.na(P99.5[i]) | is.na(P99.9[i])){
      a1[i] <- NA
    } else {
      # if P99, P99.5, P99.9 all exists, estimate a
      a1[i] <- exponent.multi_ratios_est(P99 = P99[i], P99.5 = P99.5[i], P99.9 = P99.9[i])
    }
  }
  return(a1)
}
a1 <- a.estimate(P99 = wtid$P99, P99.5 = wtid$P99.5, P99.9 = wtid$P99.9)
df1 <- data.frame(Year = wtid$Year, a1 = a1)
library(ggplot2)
ggplot(df1, mapping = aes(x = Year, y = a1)) + geom_line() +
  labs(x = "Year", y = "Estimation of a")
```



iv. Comparison of two estimation

```
# using equation (4) to estimate
a2 <- (1 - (log(10)) / (log(wtid$P99 / wtid$P99.9)))
df2 <- data.frame(a1 = a1, a2 = a2)
ggplot(df2, mapping = aes(x = a1, y = a2)) + geom_point() +
  geom_abline(slope = 1, intercept = 0) +
  labs(x = "Estimation of a using a.estimate", y = "Estimation of a using equation (4)")
```




```
identical(a1, a2)
```

```
## [1] FALSE
```

These two estimates are highly correlated but not totally identical.

Part 2: Data for Other Countries

v. Estimate a for each country over time.

```
library(xlsx)
```

```
## Loading required package: rJava
```

```
## Loading required package: xlsxjars
```

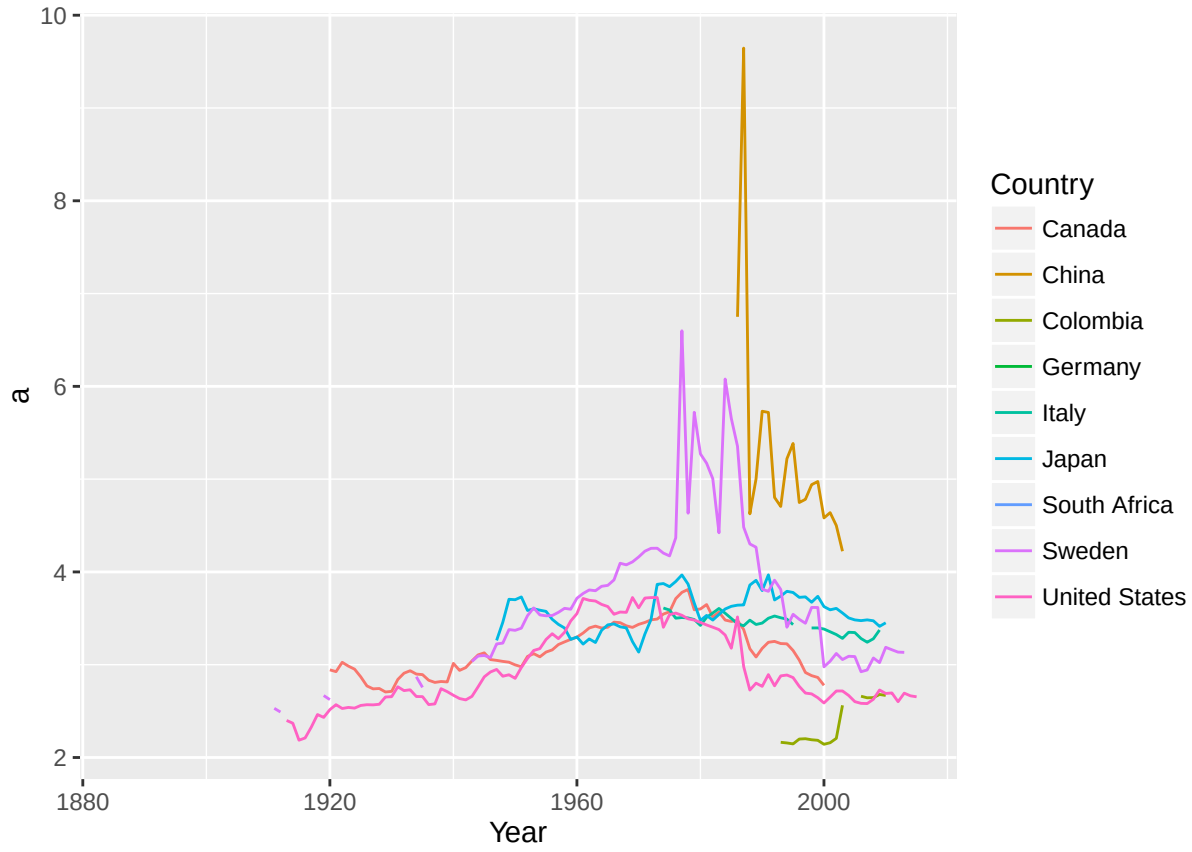
```r
# import the .xlsx file and save them into data.list
# data.list should has 9 data frame to store 9 countries' data
data.files <- list.files(pattern = "*.xlsx")
data.list <- vector("list", length(data.files))
for (i in 1:length(data.files)){
  data.list[[i]] <- read.xlsx(data.files[i], sheetName = "Series-layout A",
                              startRow = 2, header = TRUE)
}

# convert the all countries' data in a single dataframe
wtid3 <- data.frame(Country = character(0), Year = numeric(0),
                    AveIncomePerTU = numeric(0), P99 = numeric(0),
                    P99.5 = numeric(0), P99.9 = numeric(0))
for (j in 1:length(data.files)){
  # check if the variables exist in the dataframe
  if("P99.income.threshold" %in% colnames(data.list[[j]]) &
     "P99.5.income.threshold" %in% colnames(data.list[[j]]) &
     "P99.9.income.threshold" %in% colnames(data.list[[j]]) &
     "Average.income.per.tax.unit" %in% colnames(data.list[[j]])){
    # insert the value into the new dataframe
    wtid3 <- rbind(wtid3, data.frame(
      Country = data.list[[j]]$Country,
      Year = data.list[[j]]$Year,
      AveIncomePerTU = data.list[[j]]$Average.income.per.tax.unit,
      P99 = data.list[[j]]$P99.income.threshold,
      P99.5 = data.list[[j]]$P99.5.income.threshold,
      P99.9 = data.list[[j]]$P99.9.income.threshold
  ))
  } else {
    # if the variables are not in the dataframe, create them
    data.list[[j]]$Average.income.per.tax.unit <- rep(NA, length(data.list[[j]]$Year))
    data.list[[j]]$P99.income.threshold <- rep(NA, length(data.list[[j]]$Year))
    data.list[[j]]$P99.5.income.threshold <- rep(NA, length(data.list[[j]]$Year))
    data.list[[j]]$P99.9.income.threshold <- rep(NA, length(data.list[[j]]$Year))
    wtid3 <- rbind(wtid3, data.frame(
      # insert the value NA into the new dataframe
      Country = data.list[[j]]$Country,
      Year = data.list[[j]]$Year,
      AveIncomePerTU = data.list[[j]]$Average.income.per.tax.unit,
      P99 = data.list[[j]]$P99.income.threshold,
      P99.5 = data.list[[j]]$P99.5.income.threshold,
      P99.9 = data.list[[j]]$P99.9.income.threshold
    ))
  }
}
# estimate a for each country and each year
wtid3$a <- a.estimate(P99 = wtid3$P99, P99.5 = wtid3$P99.5, P99.9 = wtid3$P99.9)
```

vi. Plotting estimate of a over time

```
ggplot(data = wtid3) +
  geom_line(mapping = aes(x = Year, y = a, col = Country))
```

```
## Warning: Removed 71 rows containing missing values (geom_path).
```



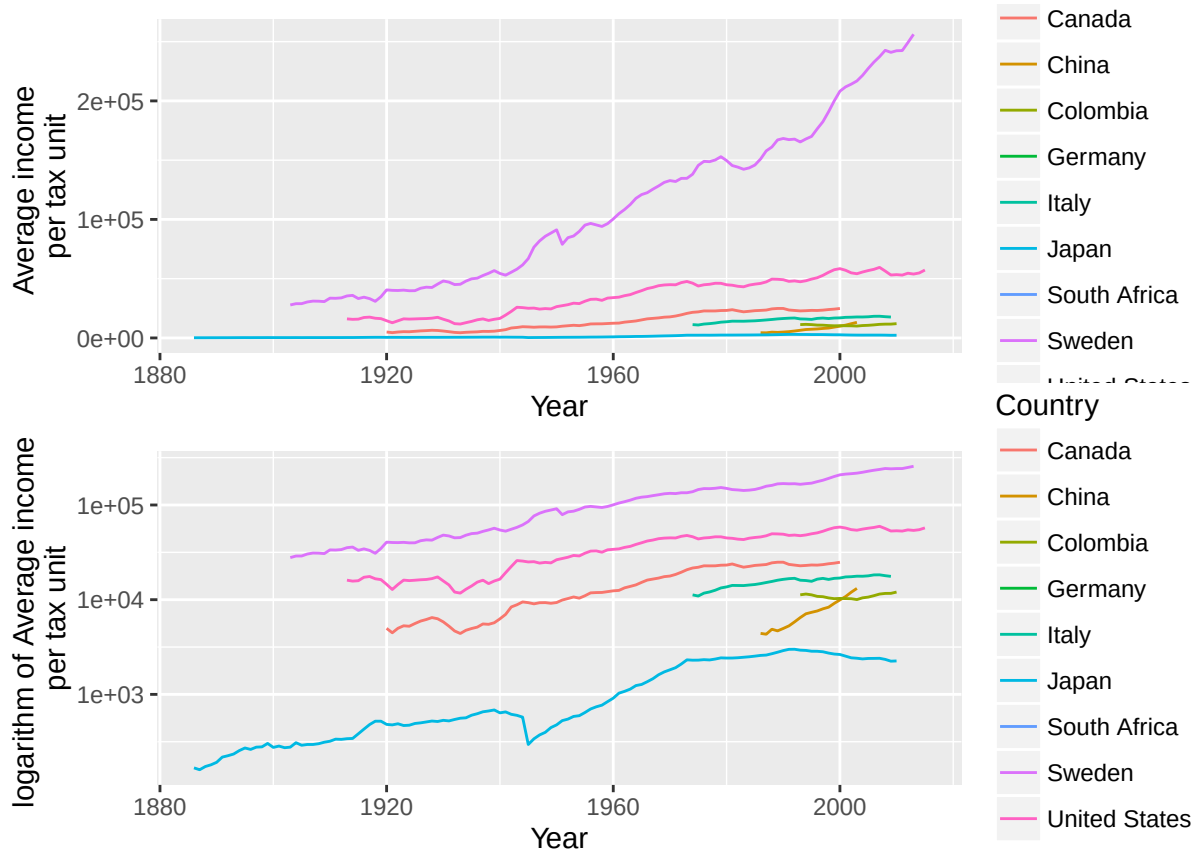
vii. Plotting average of income per tax unit

We can see that the average income per tax unit varies a lot among different countries, therefore we take logarithm to visualize them more clearly.

```
library(gridExtra)
g1 <- ggplot(data = wtid3) +
  geom_line(mapping = aes(x = Year, y = AveIncomePerTU, col = Country)) +
  labs(x = "Year", y = "Average income \nper tax unit")
g2 <- ggplot(data = wtid3) +
  geom_line(mapping = aes(x = Year, y = AveIncomePerTU, col = Country)) +
  scale_y_log10() +
  labs(x = "Year", y = "logarithm of Average income \nper tax unit")
grid.arrange(g1, g2, ncol=1)
```

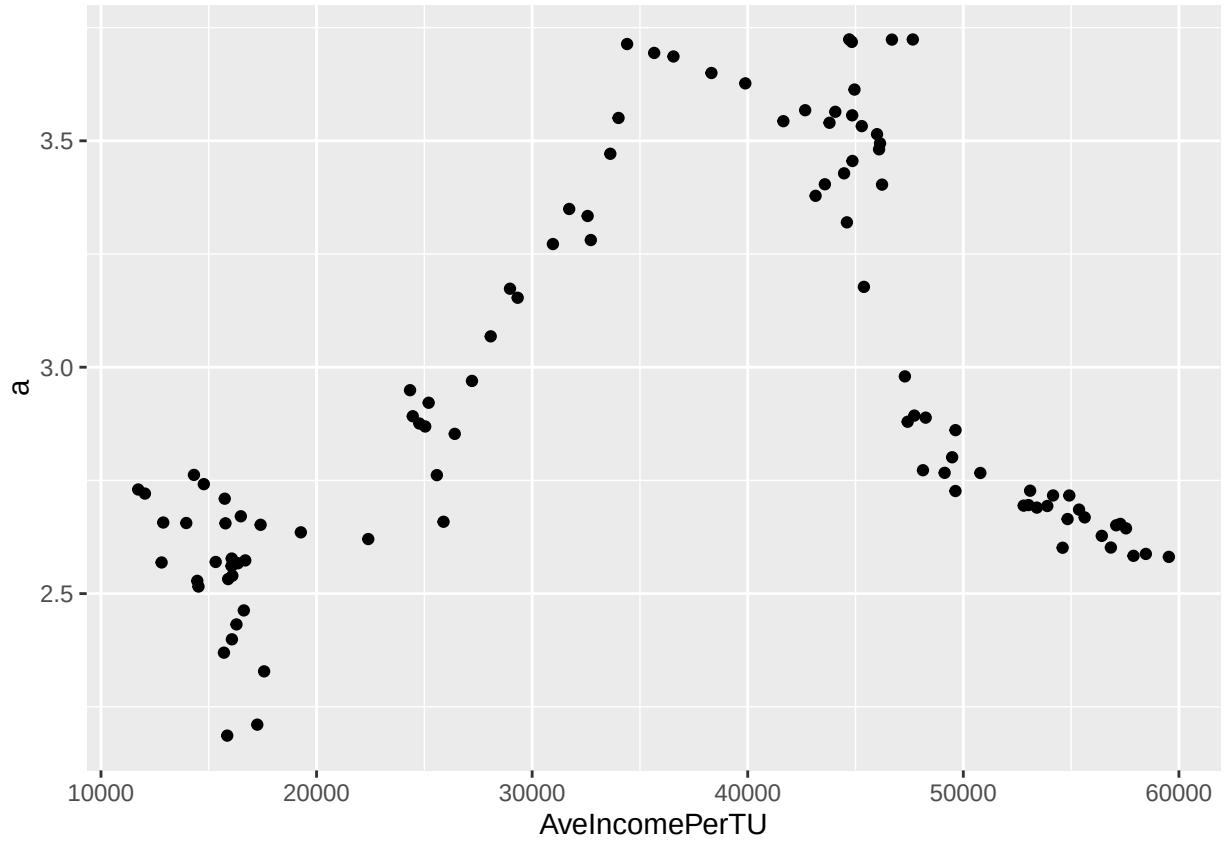
```
## Warning: Removed 10 rows containing missing values (geom_path).
```

```
## Warning: Removed 10 rows containing missing values (geom_path).
```



viii. Plotting a against average income for the US

```
# subset US data
wtid.us <- wtid3[wtid3$Country == "United States", ]
# ploting
ggplot(data = wtid.us) +
  geom_point(mapping = aes(x = AveIncomePerTU, y = a))
```



The graph shows that a is small at the very beginning, then it rises as average income rises (inequality decline), and then it goes down (inequality growth). Therefore Kuznets curve is not reasonable for the situation of US.

ix. Building quadratic regression model for US

```
lm1 <- lm(a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid.us)
summary(lm1)
```

```
##
## Call:
## lm(formula = a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50724 -0.18364 -0.02531  0.18689  0.54918
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.230e-01  1.515e-01   5.432 3.93e-07 ***
## AveIncomePerTU       1.394e-04  1.015e-05  13.740  < 2e-16 ***
## I(AveIncomePerTU^2) -1.891e-09  1.451e-10 -13.027  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2466 on 100 degrees of freedom
```

```
## Multiple R-squared:  0.6679, Adjusted R-squared:  0.6612
## F-statistic: 100.5 on 2 and 100 DF,  p-value: < 2.2e-16
```

As we can see, the coefficient of quadratic term is less than zero, which means the fitted function is a concave function, therefore the coefficient is not consistent with the hypothesis.

x. Building quadratic regression

```
reg <- function(country){
  # check if any necessary data exists
  if (all(is.na(wtid3$a[wtid3$Country == country]))){
    return(cat("Data of", country,"is missing."))
  } else {
    lm <- lm(a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid3[wtid3$Country == country,])
    return(summary(lm))
  }
}
reg("Canada")
```

```
##
## Call:
## lm(formula = a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid3[wtid3$Country ==
##     country, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49608 -0.09811 -0.00176  0.11890  0.46240
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.266e+00  1.057e-01  21.446  < 2e-16 ***
## AveIncomePerTU       1.241e-04  1.764e-05   7.037 6.70e-10 ***
## I(AveIncomePerTU^2) -3.361e-09  5.911e-10  -5.686 2.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1804 on 78 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:    0.6,  Adjusted R-squared:  0.5897
## F-statistic: 58.49 on 2 and 78 DF,  p-value: 3.037e-16
```

```
reg("China")
```

```
##
## Call:
## lm(formula = a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid3[wtid3$Country ==
##     country, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52834 -0.43065  0.04444  0.27953  3.13246
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.040e+01  2.328e+00   4.466 0.000453 ***
## AveIncomePerTU      -1.127e-03  6.043e-04  -1.865 0.081906 .
## I(AveIncomePerTU^2)  5.258e-08  3.594e-08   1.463 0.164182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 15 degrees of freedom
## Multiple R-squared:  0.3892, Adjusted R-squared:  0.3077
## F-statistic: 4.778 on 2 and 15 DF,  p-value: 0.0248
```

```
reg("Colombia")
```

```
##
## Call:
## lm(formula = a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid3[wtid3$Country ==
##     country, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26469 -0.10766 -0.04688  0.13801  0.39824
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.461e+01  1.870e+01   1.851   0.0870 .
## AveIncomePerTU      -6.095e-03  3.417e-03  -1.784   0.0978 .
## I(AveIncomePerTU^2)  2.867e-07  1.558e-07   1.841   0.0886 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1933 on 13 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4189, Adjusted R-squared:  0.3295
## F-statistic: 4.686 on 2 and 13 DF,  p-value: 0.02935
```

```
reg("Germany")
```

```
## Data of Germany is missing.
```

```
reg("Italy")
```

```
##
## Call:
## lm(formula = a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid3[wtid3$Country ==
##     country, ])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.107224 -0.040044 -0.007691  0.035728  0.125061
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.582e+00  4.598e-01   5.617 3.66e-06 ***
## AveIncomePerTU        1.594e-04  6.268e-05   2.544  0.01618 *
## I(AveIncomePerTU^2) -6.591e-09  2.103e-09  -3.134  0.00376 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05295 on 31 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7216, Adjusted R-squared:  0.7037
## F-statistic: 40.18 on 2 and 31 DF,  p-value: 2.465e-09
```

```
reg("Japan")
```

```
##
## Call:
## lm(formula = a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid3[wtid3$Country ==
##     country, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30181 -0.10102 -0.02892  0.07070  0.41642
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.729e+00  9.076e-02  41.086  < 2e-16 ***
## AveIncomePerTU       -5.136e-04  1.298e-04  -3.957 0.000201 ***
## I(AveIncomePerTU^2)   1.889e-07  3.823e-08   4.942 6.34e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1503 on 61 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  0.4618, Adjusted R-squared:  0.4441
## F-statistic: 26.17 on 2 and 61 DF,  p-value: 6.229e-09
```

```
reg("South Africa")
```

```
## Data of South Africa is missing.
```

```
reg("Sweden")
```

```
##
## Call:
## lm(formula = a ~ AveIncomePerTU + I(AveIncomePerTU^2), data = wtid3[wtid3$Country ==
##     country, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8008 -0.3044 -0.1088  0.1520  2.3314
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.012e+00  2.737e-01   3.699 0.000398 ***
## AveIncomePerTU        4.414e-05  4.259e-06  10.366  < 2e-16 ***
## I(AveIncomePerTU^2) -1.497e-10  1.492e-11 -10.034 9.47e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.554 on 79 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.5764, Adjusted R-squared:  0.5657
## F-statistic: 53.74 on 2 and 79 DF,  p-value: 1.842e-15
```

As we can see, the coefficients of quadratic term is positive for these countries: China, Colombia and Japan. And the coefficients are only reliable for Japan. Therefore Japan is compatible with the hypothesis.