Human Capital Index Hypotheses Testing

MINH CHAU

```
# load required packages
library(dplyr)
library(tidyr)
library(car)
library(ggplot2)
```

- Dataset required: `WorldBankData.csv`

(Note: This dataset comes from a publically available dataset from The World Bank. https://databank.worldbank.org/source/world-development-indicators (https://databank.worldbank.org/source/world-development-indicators).)

There are 8 variables in this real-world dataset, from 258 countries in 2016/2017:

- `Country.Name.Name` : name of country
- `Country.Code` : code given to country
- `Human.Capital.Index` : unitless number that goes from 0 to 1.
- `GDP.per.capita.PPP` in US Dollar. This is GDP per capita, but taking into account the purchasing power of the local currency, by comparing how much it costs to buy a basket of goods (e.g. food) compared to the reference currency (USD). (PPP stands for Purchasing Power Parity)
- `Health.Expenditure.per.capita` in US Dollar.
- `Tertiary.Education.Expenditure.per.student` in US Dollar.
- `Population` .- `Life.Expectancy.at.birth` in years.
- `Diabetes.Prevalence` in units of % of population ages 20 to 79.
- `Years.of.Compulsory.Education` in years.

Being a data set in real world, there are lots of missing data. Be wary of this!

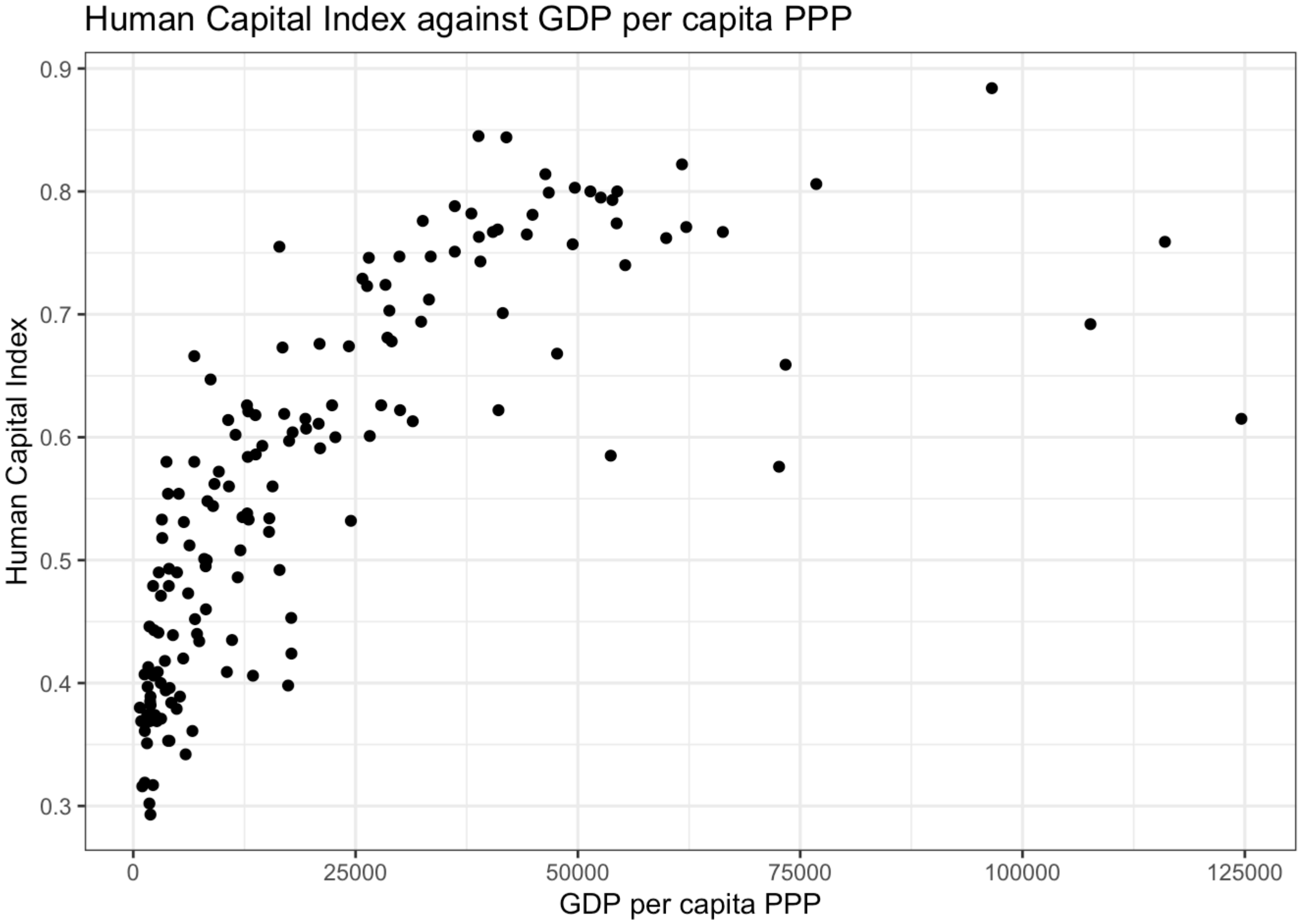
```
dta_wb <- read.csv("/Users/minhchau/Downloads/WorldBankData.csv")
```

First, let’s investigate `Human.Capital.Index` . As noted by Prime Minister Lee in his 2019 National Day Rally, Singapore topped the world on this Human Capital Index in 2018. Let’s try to see what are some of the possible variables that correlate with this.

(1a) Start off by plotting `Human.Capital.Index` (on the y-axis) versus `GDP.per.capita.PPP` on the x-axis.

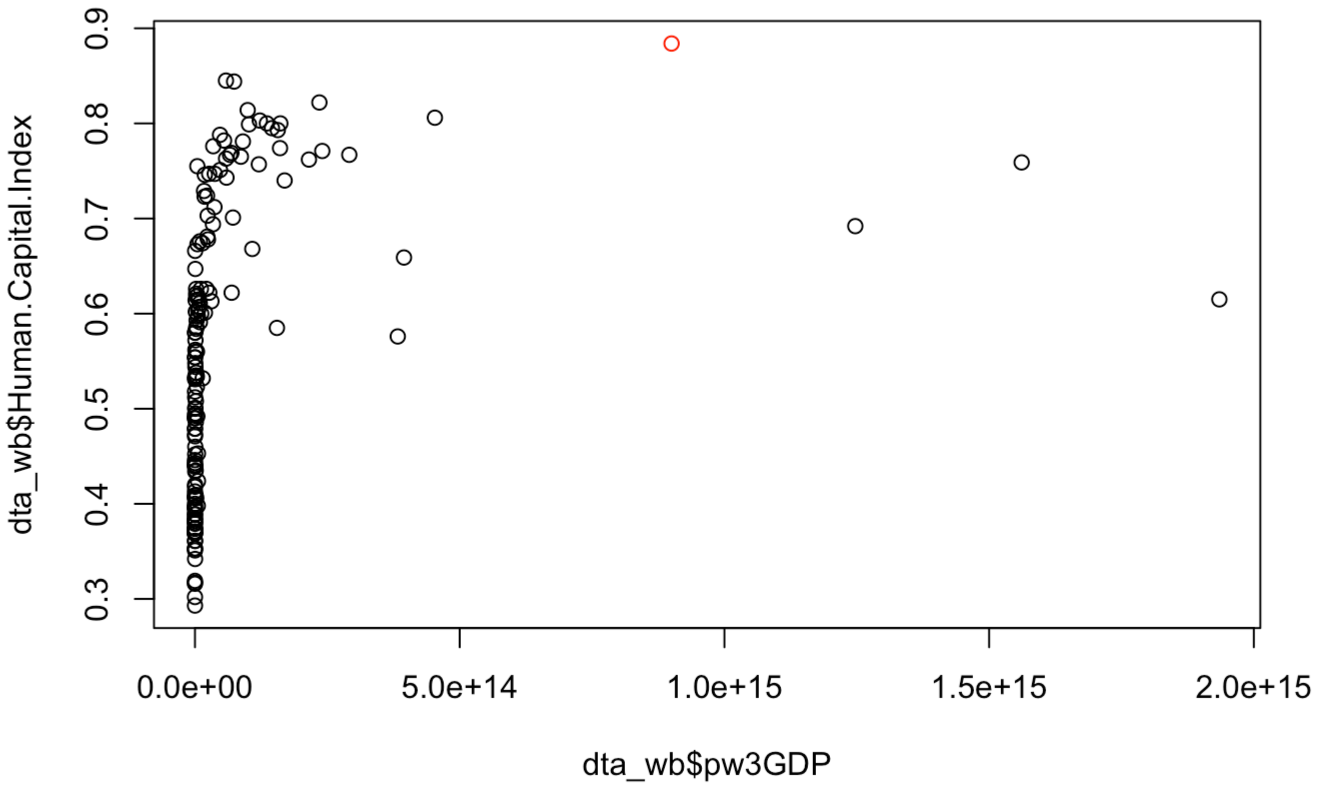
```
ggplot(dta_wb, aes(x=GDP.per.capita.PPP, y=Human.Capital.Index)) + ggtitle("Human Cap
ital Index against GDP per capita PPP") + xlab("GDP per capita PPP") +ylab("Human Cap
ital Index") +geom_point()+theme_bw()
```

```
## Warning: Removed 101 rows containing missing values or values outside the scale ra
nge
## (`geom_point()`).
```

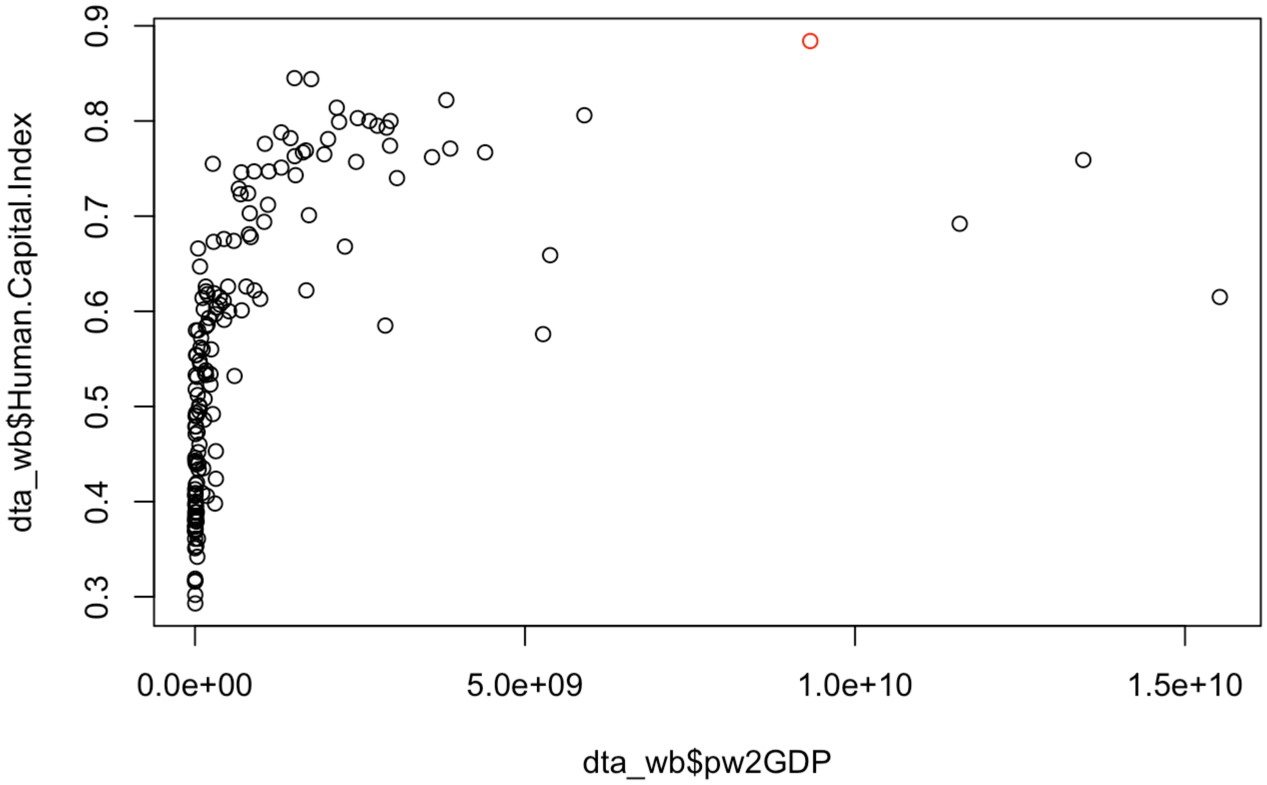


(1b) What type of transformation could you apply? Try a few functions that were shown in class: `x^2, x^3, ...`, `exp(x)`, `log(x)`. Make a plot that shows a linear relationship, and describe what you did. Highlight the dot that represents Singapore.

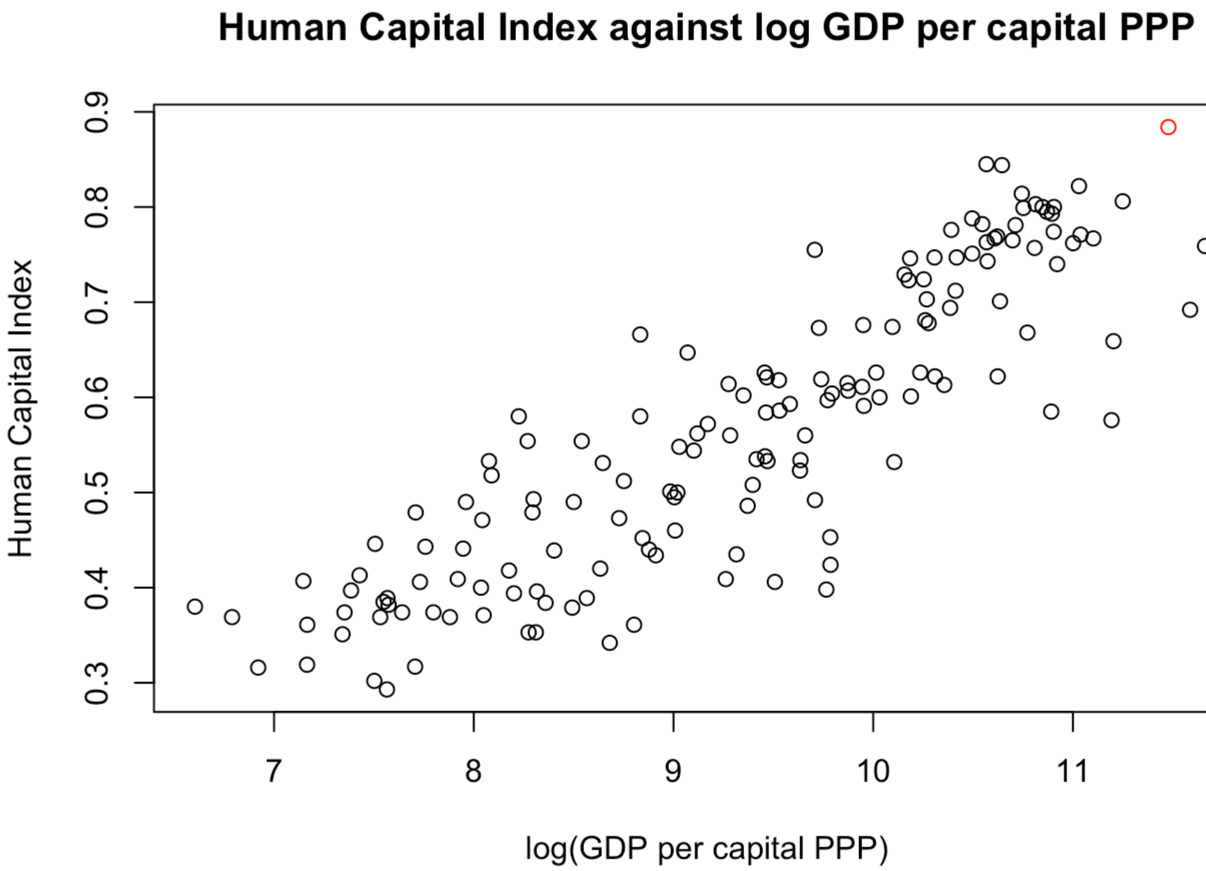
```
dta_wb$pw3GDP = (dta_wb$GDP.per.capita.PPP)^3
plot(dta_wb$pw3GDP, dta_wb$Human.Capital.Index, col=ifelse(dta_wb$Country.Name == "Si
ngapore", 'red', 'black'))
```



```
dta_wb$pw2GDP = (dta_wb$GDP.per.capita.PPP)^2
plot(dta_wb$pw2GDP, dta_wb$Human.Capital.Index, col=ifelse(dta_wb$Country.Name == "Si
ngapore", 'red', 'black'))
```



```
dta_wb$logGDP = log(dta_wb$GDP.per.capita.PPP)
plot(dta_wb$logGDP, dta_wb$Human.Capital.Index, col = ifelse(dta_wb$Country.Name ==
"Singapore", 'red', 'black'), main="Human Capital Index against log GDP per capital P
PP", xlab = "log(GDP per capital PPP)", ylab = "Human Capital Index")
```



(1c) Now that you have a plot of a linear relationship, run a linear regression using `lm()`, predicting `Human Capital Index`. Run `summary()` on the `lm` object to produce an output table. Interpret the output of the `lm()`. What do the `b's` regression coefficients estimates mean? (Interpret them and try to make sense of the numbers. How many countries made it into this regression? (What happened to the rest?) Comment on the `goodness-of-fit statistics`.

(1d) Do you think that log(GDP) a significant predictor in our linear regression model?

```
summary(lm(Human.Capital.Index ~ logGDP, dta_wb))
```

```
##
## Call:
## lm(formula = Human.Capital.Index ~ logGDP, data = dta_wb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21270 -0.04959  0.01103  0.06164  0.15487
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.432638   0.047260  -9.155 3.03e-16 ***
## logGDP       0.106843   0.005008  21.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07666 on 155 degrees of freedom
##   (101 observations deleted due to missingness)
## Multiple R-squared:  0.746,  Adjusted R-squared:  0.7443
## F-statistic: 455.2 on 1 and 155 DF,  p-value: < 2.2e-16
```

We state our hypotheses H0: ββlog(GDP)= 0 H1: ββlog(GDP) ≠ 0 where ββlog(GDP) represents the population slope parameter of the natural log- transformed GDP variable Given that p-value = 2e-16 < 0.05 at 5% level of significance We can reject the null hypothesis H0:ββlog(GDP)= 0, meaning that the population slope parameter of log(GDP) is statistically significantly different from zero. Thus we conclude that the log-transformed GDP variable (log(GDP)) is a significant predictor in our linear regression model.