BT1101

Lab 6: Linear Regression

Installing and loading packages

```r
# load required packages
library(dplyr)
library(tidyr)
library(ggplot2) # optional. we expect you to know base graphics, but allow ggplot if you find it easier
```

- We expect you to know base graphics

Part 1

Part One: Lab-Session Completion and Discussion

- 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.)

First, load in the dataset for this question. There are 8 variables in this real-world dataset, from 258 countries in 2016/2017:

- `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!

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

Loading datasets into R

```
dta_wb = read.csv('WorldBankData.csv')
```

Country.Name	Country.Code	Years.of.Compulsory.Education	Health.Expenditure.per.capita	Diabetes.Prevalence	GDP.per.capita.PPP	Tertiary.Education.Expenditure.per.student	Human.Capital.Index	Life.Expectancy.at.birth	Population
Afghanistan	AFG	9.0	162.78116	9.590000	1934.6368	NA	0.389	64.04700	36296400
Albania	ALB	9.0	759.66698	10.080000	12930.0677	14.80285	0.621	78.49500	2873457
Algeria	DZA	10.0	998.15375	6.730000	15266.4852	NA	0.523	76.29300	41389198
Andorra	AND	10.0	4978.70660	7.970000	NA	23.65730	NA	NA	77001
Angola	AGO	6.0	185.82040	3.940000	6650.5849	NA	0.361	61.80900	29816748
Antigua and Barbuda	ATG	11.0	976.38866	13.170000	25145.5412	NA	NA	76.51900	95426
Arab World	ARB	9.0	1014.55922	12.099230	17102.1269	NA	NA	71.43312	411898965
Argentina	ARG	14.0	1531.03836	5.500000	20843.1551	16.25644	0.611	76.73800	44044811
Armenia	ARM	12.0	876.85686	7.110000	9620.8185	9.72792	0.572	74.78200	2944809
Aruba	ABW	13.0	NA	11.620000	39454.6298	97.28195	NA	76.01000	105366
Australia	AUS	10.0	4529.88708	5.070000	49653.7159	NA	0.803	82.49756	24601860
Austria	AUT	13.0	5295.18177	6.350000	53879.2979	NA	0.793	81.64146	8797566
Azerbaijan	AZE	10.0	1193.05883	7.110000	17525.2796	23.31199	0.597	72.12300	9854033
Bahamas, The	BHS	12.0	1435.56751	13.170000	31581.1044	NA	NA	75.82300	381761
Bahrain	BHR	9.0	1866.29732	16.520000	47660.4799	NA	0.668	77.03800	1494074
Bangladesh	BGD	5.0	90.59840	8.380000	3998.4194	30.84674	0.479	72.80800	159670593
Barbados	BRB	11.0	1322.98551	13.570000	18526.0086	NA	NA	76.05700	286233
Belarus	BLR	9.0	1151.40885	5.180000	18915.9399	18.03424	NA	74.12927	9498264
Belgium	BEL	12.0	4667.88229	4.290000	49411.8691	NA	0.757	81.43902	11375158
Belize	BLZ	8.0	541.43433	17.110000	8500.4448	30.16545	NA	70.58800	375769
Benin	BEN	6.0	83.47637	0.990000	2276.5957	NA	0.406	61.17100	11175204
Bermuda	BMU	13.0	NA	13.000000	NA	NA	NA	81.44195	63874

Coding

Question 1

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. What do you notice? What type of relationship exists between the two variables? Is it linear?

- Plot Human.Capital.Index (on the y-axis) versus GDP.per.capita.PPP (on the x-axis)
- What type of relationship exists between the two variables?

(1b) What type of transformation could you apply? Try a few functions that were shown in class: `x^2, x^3, ...`, `exp(x)`, `log10(x)`. Make a plot that shows a linear relationship, and describe what you did.

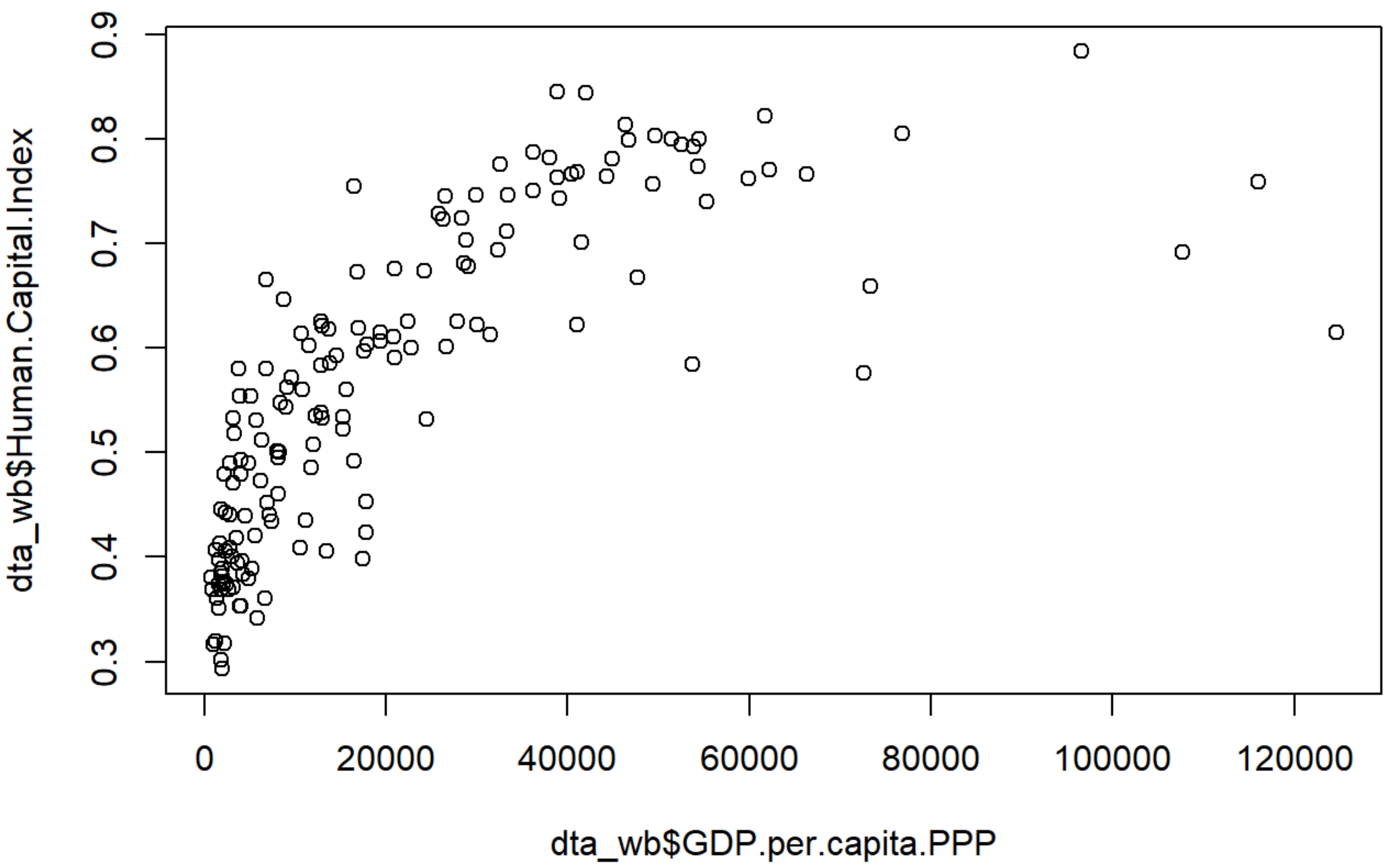
For fun: add code into your plot to highlight the dot that represents Singapore.

(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` coefficients mean? (Interpret them "in English" and try to make sense of the numbers, even if they might seem weird at first. How many countries made it into this regression? (What happened to the rest?) Comment on the goodness-of-fit statistics.

Q1(a)

(1a) Start off by plotting Human.Capital.Index (on the y-axis) versus GDP.per.capita.PPP on the x-axis. What do you notice? What type of relationship exists between the two variables? Is it linear?

```
plot(dta_wb$GDP.per.capita.PPP, dta_wb$Human.Capital.Index)
```



GDP per capita is correlated with Human Capital Index, such that countries with higher GDP per capita also tend to have higher Human Capital Index.

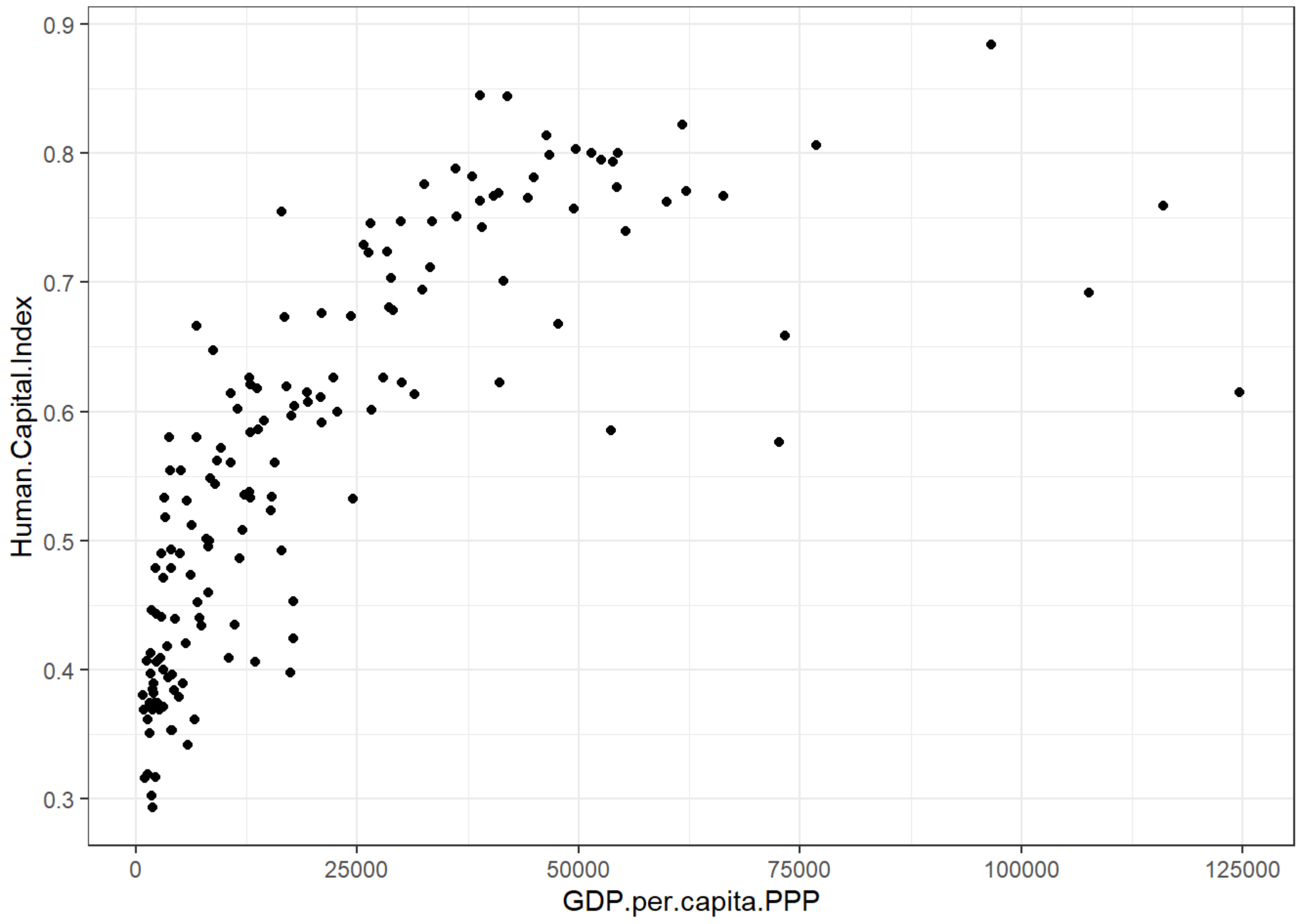
However the relationship does **not** seem to be linear.

Q1(a)

(1a) Start off by plotting Human.Capital.Index (on the y-axis) versus GDP.per.capita.PPP on the x-axis. What do you notice? What type of relationship exists between the two variables? Is it linear?

```
ggplot(dta_wb, aes(x=GDP.per.capita.PPP, y=Human.Capital.Index)) + geom_point() + theme_bw()
```

```
## Warning: Removed 101 rows containing missing values (geom_point).
```



Q1(b)

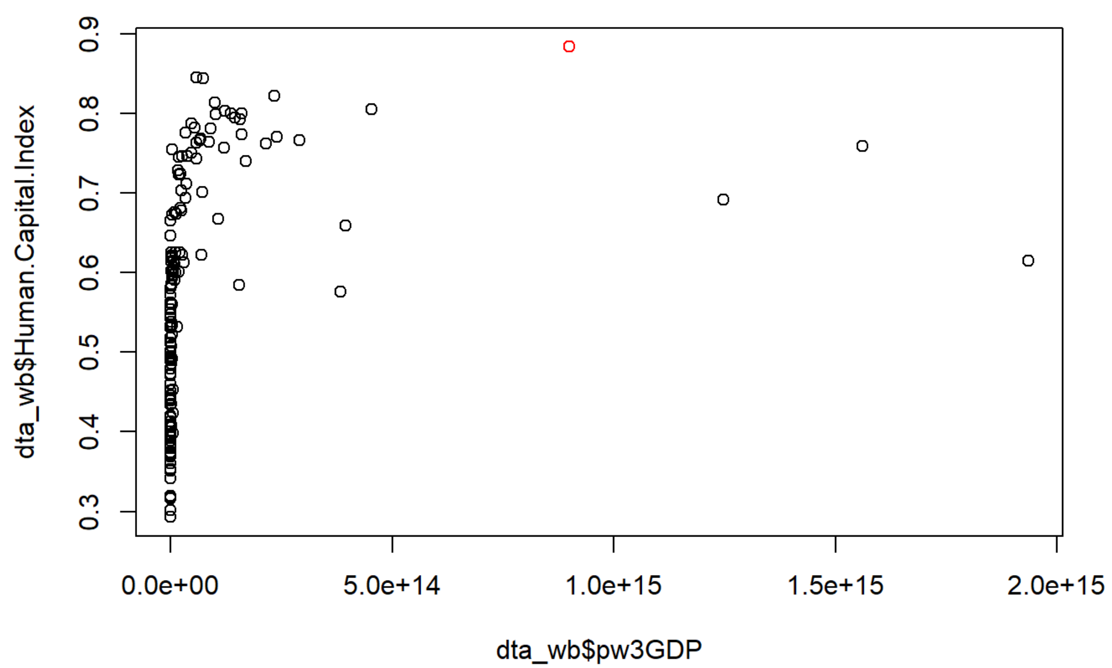
(1b) What type of transformation could you apply? Try a few functions that were shown in class: x^2, x^3, ..., exp(x), log10(x). Make a plot that shows a linear relationship, and describe what you did.

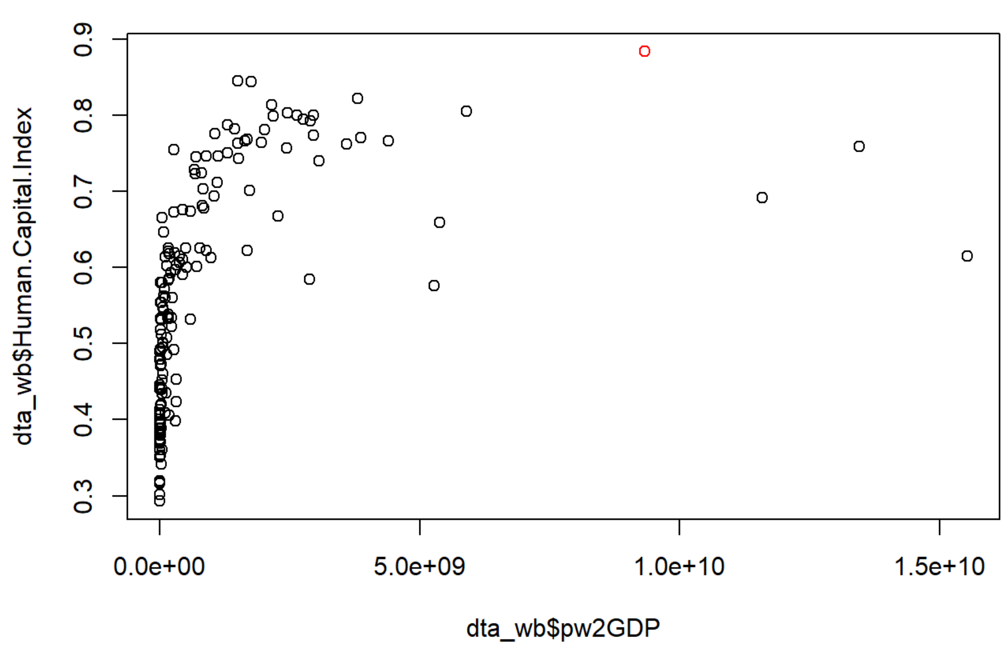
For fun: add code into your plot to highlight the dot that represents Singapore.

```
#powers

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=="Singapore", '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=="Singapore", 'red', 'black'))
```



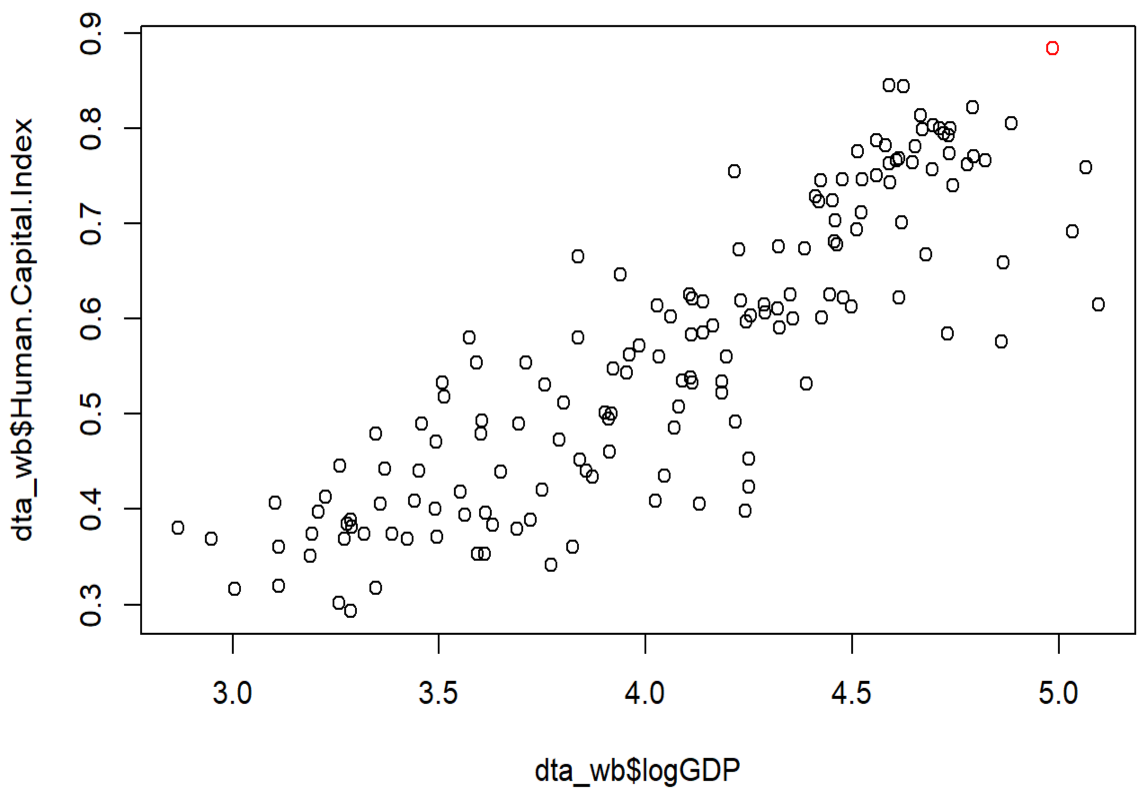


Q1(b)

(1b) What type of transformation could you apply? Try a few functions that were shown in class: x^2, x^3, ..., exp(x), log10(x). Make a plot that shows a linear relationship, and describe what you did.

For fun: add code into your plot to highlight the dot that represents Singapore.

```
# NOTE: log() is the natural log; log10() is logarithm of base 10.
dta_wb$logGDP = log10(dta_wb$GDP.per.capita.PPP)
plot(dta_wb$logGDP, dta_wb$Human.Capital.Index, col = ifelse(dta_wb$Country.Name=="Singapore", 'red', 'black'))
```



We see an exponential trend, similar to an example in the lecture slides. There may be several possible transformations to get a linear trend.

If we apply the base10 logarithm to GDP per capita, we find that there now seems to be a linear relationship between Human Capital Index and log-GDP-per-capita.

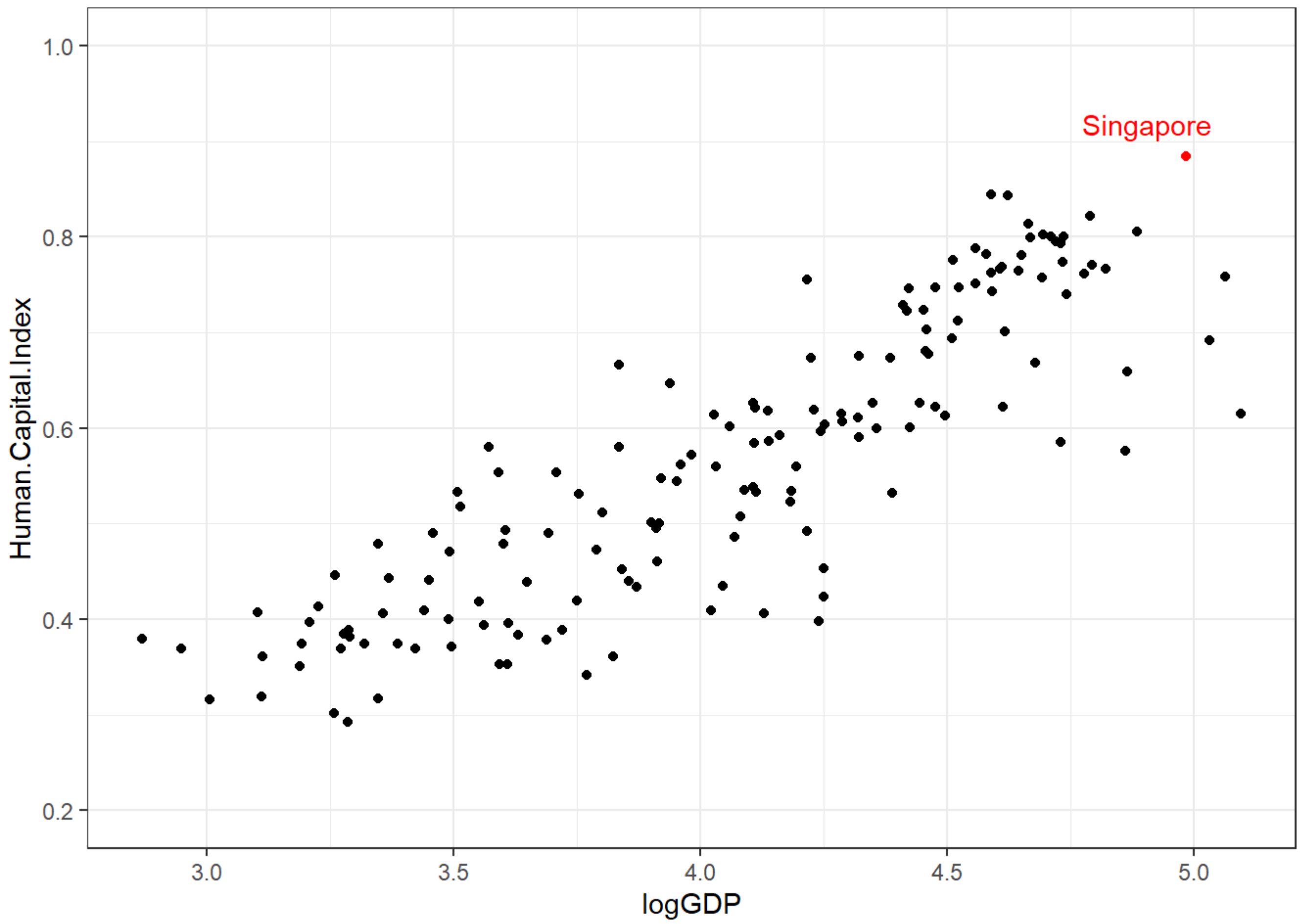
Note, natural log also acceptable, but less interpretable. If log10 GDP is 4.0, then we know that GDP is 10^4 or $10,000.

Q1(b)

(1b) What type of transformation could you apply? Try a few functions that were shown in class: x^2, x^3, ..., exp(x), log10(x). Make a plot that shows a linear relationship, and describe what you did.

For fun: add code into your plot to highlight the dot that represents Singapore.

```
# NOTE: log() is the natural log; log10() is logarithm of base 10.
dta_wb$logGDP = log10(dta_wb$GDP.per.capita.PPP)
ggplot(dta_wb, aes(x=logGDP, y=Human.Capital.Index)) + geom_point() +
  geom_point(data=subset(dta_wb, dta_wb$Country.Name=="Singapore"), color="red") +
  geom_text(data=subset(dta_wb, dta_wb$Country.Name=="Singapore"),
            aes(label=Country.Name), vjust=-1.0, hjust=0.8, color="red") +
  ylim(0.2, 1.0) + theme_bw()
```



Q1(c)

(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 coefficients mean? (Interpret them “in English” and try to make sense of the numbers, even if they might seem weird at first. How many countries made it into this regression? (What happened to the rest?) Comment on the goodness-of-fit statistics.

```
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) -0.43264    0.04726  -9.155 3.03e-16 ***
## logGDP       0.24602    0.01153  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
```

Discuss the interpretation of the units

Intercept (b_0) is -0.43
suggests average Human Capital Index for a country with logGDP =0 is -0.43 but HCI is between 0 and 1

slope (b_1) is 0.246
when log10-GDP of a country increases by 1 unit (i.e., GDP of a country increases by 10 times), we expect to see an average increase of Human Capital Index by 0.246

Q1(c)

(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 coefficients mean? (Interpret them "in English" and try to make sense of the numbers, even if they might seem weird at first. How many countries made it into this regression? (What happened to the rest?) Comment on the goodness-of-fit statistics.

Data
Data contains 258 countries. Use (nrow(dta_wb))
101 observations were deleted due to missingness of an imbalance data set.
Hence, 157 countries were used in the regression analysis.

Degrees of freedom
n-2 = 157-2 = 155 degrees of freedom, there is one independent variable logGDP plus an intercept

R-square is 0.746
Model explains almost 75% of the total variation of Human Capital Index.

F-test's p-value is significant implies that "not all beta1,… are not zero"
In this case with only one predictor, beta1 is statistically not zero.

Coding

Question 2

- Dataset required: `WorldBankData.csv`

Let's look at another set of variables in the same dataset. This time, let's consider `Health.Expenditure.per.capita`, `Diabetes.Prevalence`, and `Life.Expectancy.at.birth`.

(2a) If you had to design a predictive hypothesis with these three variables, what would it be? Which would be your dependent variable, and which would be your independent variables? Justify your answer. (Note, there is no necessarily "right" or "wrong" answer for this question, as is the case in real life, but there are more justifiable answers that you would feel more comfortable putting up to your boss!)

(2b) Plot the bivariate relationships between these three variables. (In other words, plot x-y scatterplots. There are 3 variables, so you'll need 3 scatterplots.) Please also apply the same transformation in (1b) to `Health.Expenditure.per.capita`. Comment on the relationship between the variables.

(2c) Run a multiple regression predicting `Life.Expectancy.at.birth` using the other two variables. Interpret the coefficients, spelling out what the numbers mean. Comment on your answers.

Q2(a)

(2a) If you had to design a predictive hypothesis with these three variables, what would it be? Which would be your dependent variable, and which would be your independent variables? Justify your answer. (Note, there is no necessarily “right” or “wrong” answer for this question, as is the case in real life, but there are more justifiable answers that you would feel more comfortable putting up to your boss!)

This is meant to be a question for discussion.

Q2(a)

(2a) If you had to design a predictive hypothesis with these three variables, what would it be? Which would be your dependent variable, and which would be your independent variables? Justify your answer. (Note, there is no necessarily “right” or “wrong” answer for this question, as is the case in real life, but there are more justifiable answers that you would feel more comfortable putting up to your boss!)

Life.Expectancy.at.birth

- Yes, it can be used as an outcome variable. We might seek to increase as a goal, especially if that is something of importance to society.

Health.Expenditure.per.capita

- Health.Expenditure.per.capita is potentially a policy variable that can be controlled by government. That should be a predictor / independent variable for interesting policy .
- Question: Could changing health expenditure improve the outcome variable, life expectancy?

Diabetes.Prevalence

- Yes, Diabetes.Prevalence could also be an outcome that one would want to optimize. If it’s just between Diabetes.Prevalence and Health.Expenditure.per.capita, I can definitely see Diabetes.Prevalence being the outcome variable.
- However, if we also have Life.Expectancy.at.birth, it makes more sense if Diabetes.Prevalence predicts Life.Expectancy.at.birth, than the other way around.

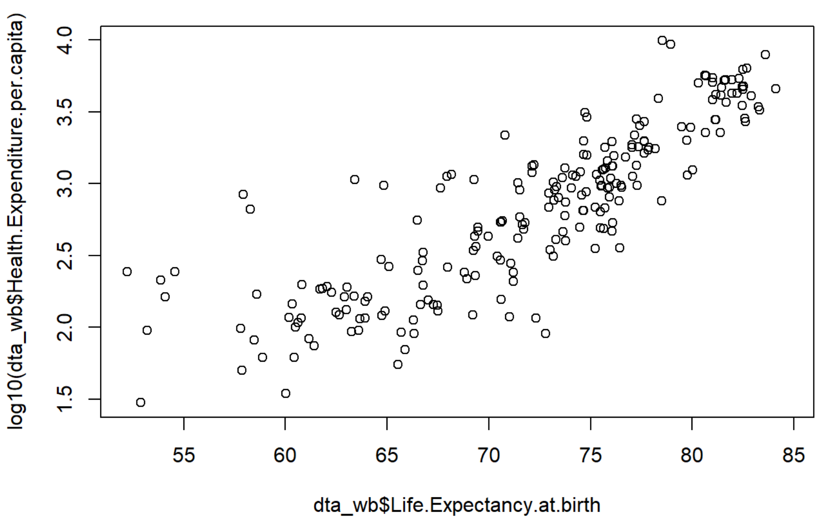
Q2(b)

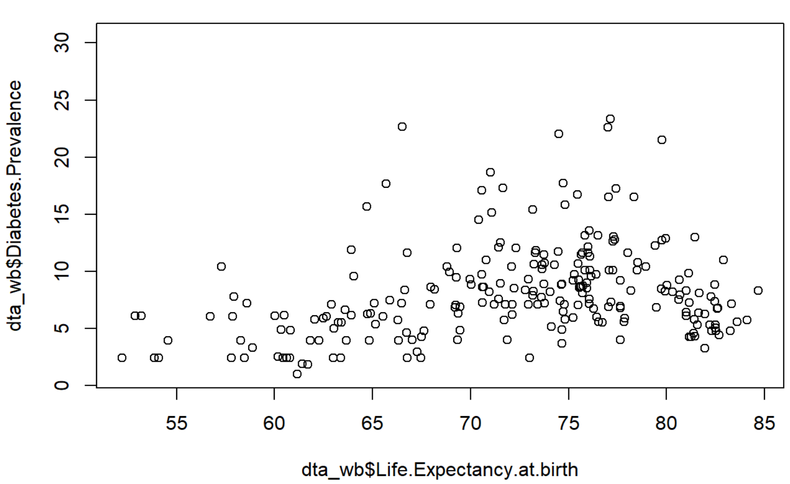
(2b) Plot the bivariate relationships between these three variables. (In other words, plot x-y scatterplots. There are 3 variables, so you'll need 3 scatterplots.) Please also apply the same transformation in (1b) to Health.Expenditure.per.capita. Comment on the relationship between the variables.

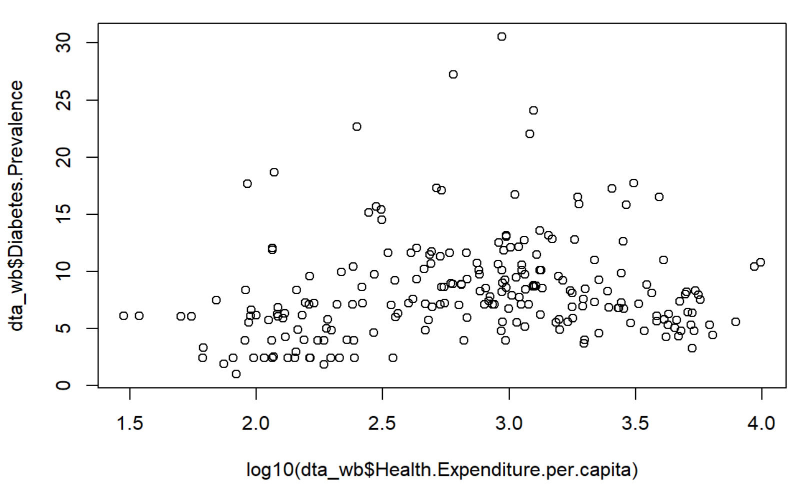
```
plot(dta_wb$Life.Expectancy.at.birth, log10(dta_wb$Health.Expenditure.per.capita))
```

```
plot(dta_wb$Life.Expectancy.at.birth, dta_wb$Diabetes.Prevalence)
```

```
plot(log10(dta_wb$Health.Expenditure.per.capita), dta_wb$Diabetes.Prevalence)
```







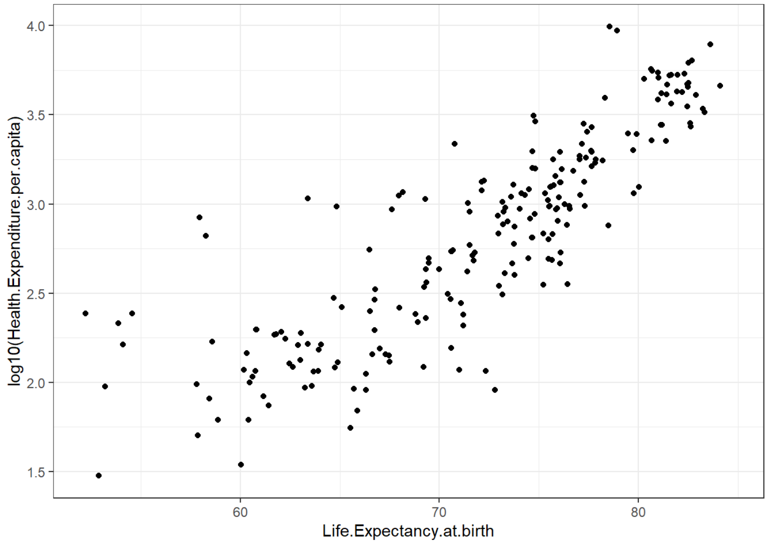
Q2(b)

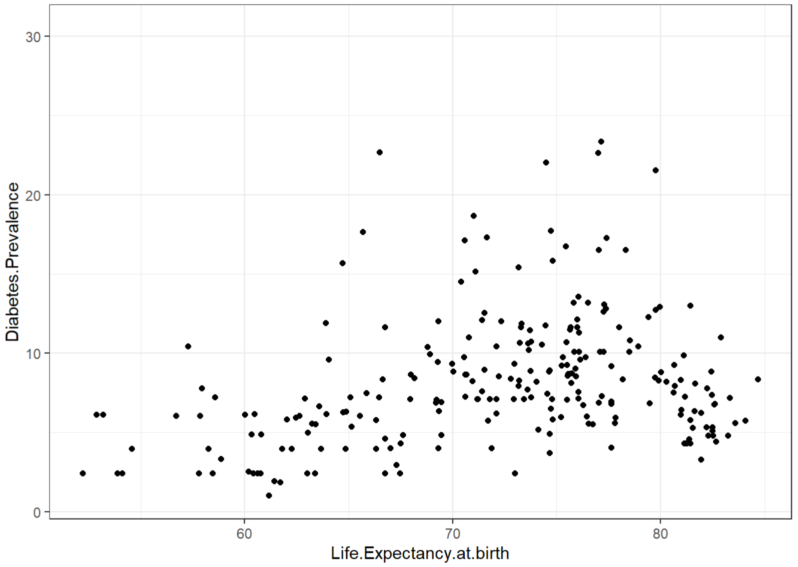
(2b) Plot the bivariate relationships between these three variables. (In other words, plot x-y scatterplots. There are 3 variables, so you'll need 3 scatterplots.) Please also apply the same transformation in (1b) to Health.Expenditure.per.capita. Comment on the relationship between the variables.

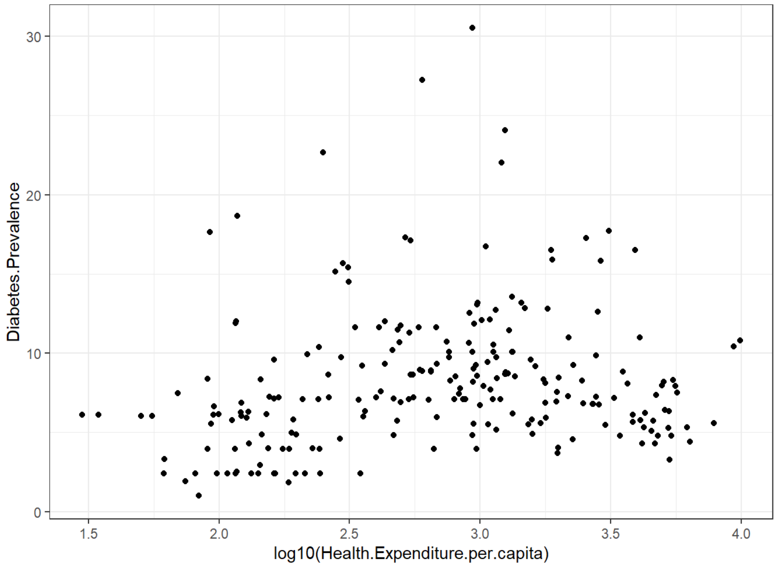
```
ggplot(dta_wb, aes(x=Life.Expectancy.at.birth, y=log10(Health.Expenditure.per.capita))) + geom_point() + theme_bw()
```

```
ggplot(dta_wb, aes(x=Life.Expectancy.at.birth, y=Diabetes.Prevalence)) + geom_point() + theme_bw()
```

```
ggplot(dta_wb, aes(x=log10(Health.Expenditure.per.capita), y=Diabetes.Prevalence)) + geom_point() + theme_bw()
```







Q2(c)

(2c) Run a multiple regression predicting `Life.Expectancy.at.birth` using the other two variables. Interpret the coefficients, spelling out what the numbers mean. Comment on your answers.

```
summary(lm(Life.Expectancy.at.birth ~ log10(Health.Expenditure.per.capita) + Diabetes.Prevalence,  dta_wb))
```

```
##
## Call:
## lm(formula = Life.Expectancy.at.birth ~ log10(Health.Expenditure.per.capita) +
##     Diabetes.Prevalence, data = dta_wb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0787  -1.4875   0.6018   2.0976  10.0565
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          39.61736    1.33051   29.78  < 2e-16 ***
## log10(Health.Expenditure.per.capita) 10.77368    0.45941   23.45  < 2e-16 ***
## Diabetes.Prevalence                   0.24448    0.06847    3.57 0.000438 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.796 on 218 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7364
## F-statistic: 308.4 on 2 and 218 DF,  p-value: < 2.2e-16
```

Note: Tutors should guide students on how to think through such models.

Significant predictors

- Both log-Health Expenditure and Diabetes prevalence are statistically significant predictors of Life Expectancy.

Interpretation of the intercept

- In a country with 0 log-Health expenditure and with 0 diabetes, life expectancy is 39.6 years on average

log-Health Expenditure per capita

- Every unit increase of log-Health Expenditure per capita (i.e., increasing Health Expenditure by 10x) is associated with an expected increase in life expectancy by 10.7 years!
- Note t-value is very large, and the p-value is very small,(< 0.05), thus we can reject the null hypothesis that this slope parameter of log Health Expenditure per capita = 0, i.e., the coefficient is significantly different from zero.

Diabetes prevalence

- How come every % increase in Diabetes prevalence is associated with an increase in life expectancy by 0.24 years?! Isn't this opposite, in that if prevalence of Diabetes is very high, then life expectancy should go down, right, since more people will die of diabetes?!
- This could be due to external factors that are not in our model. For example, perhaps countries that are more affluent have access to more sugary foods, and thus may experience higher rates of diabetes. The affluent countries also have access to better healthcare and hence may enjoy longer life expectancies.

confounding variables