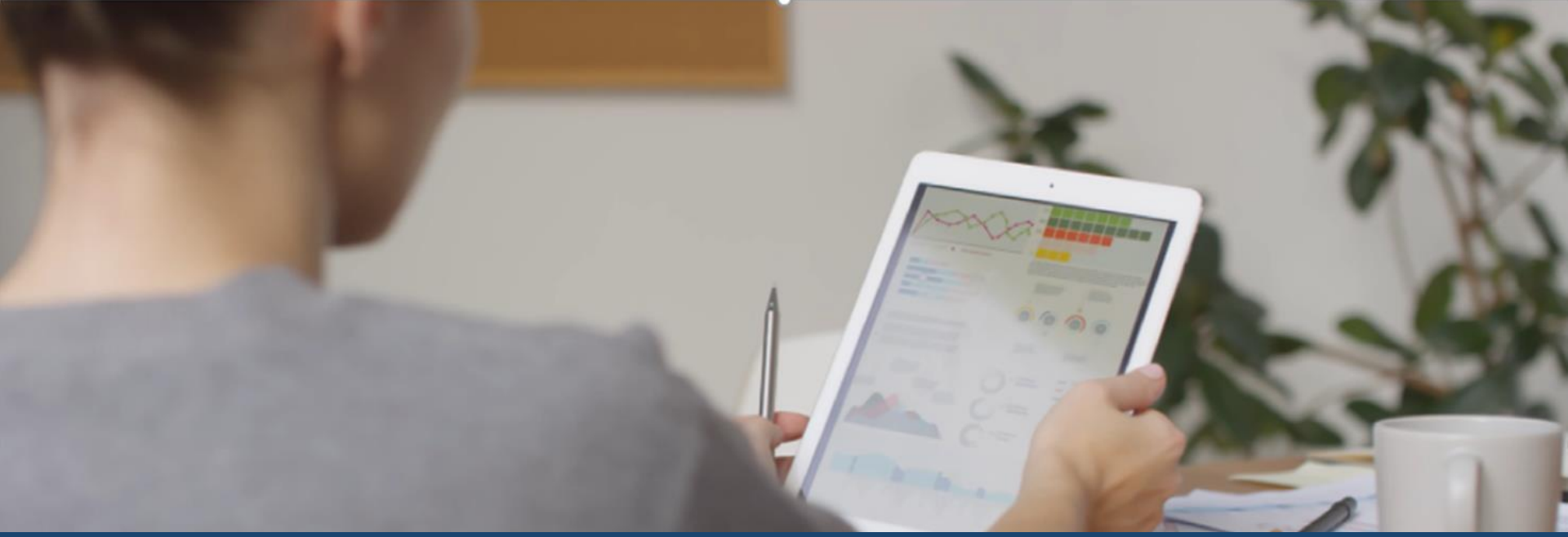
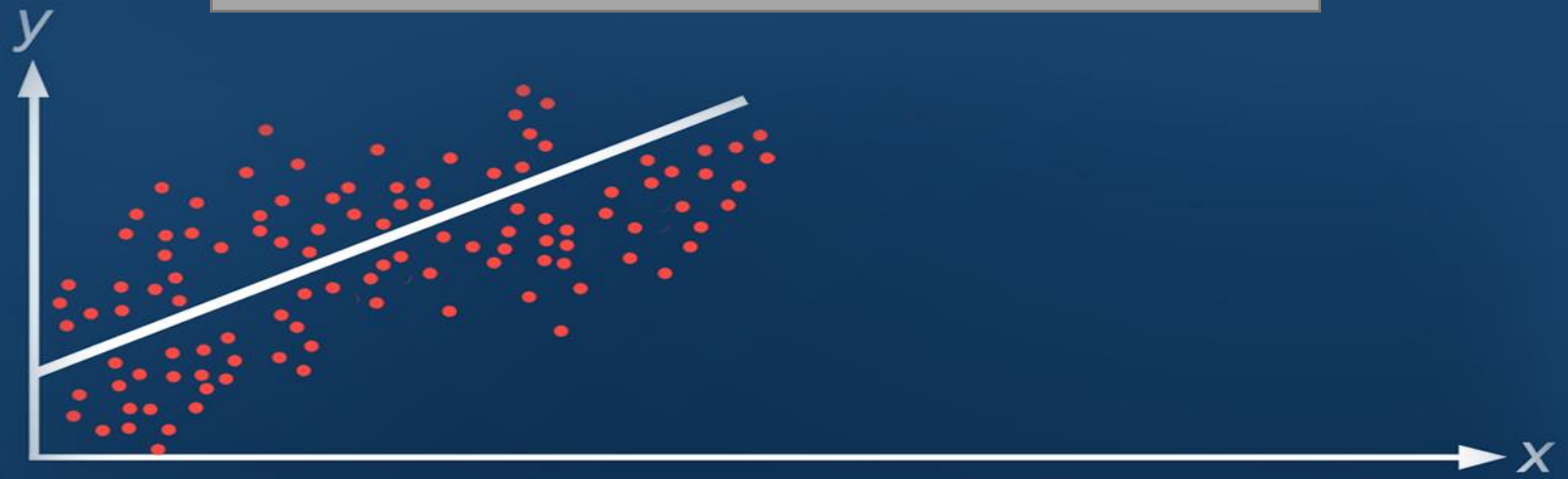


TBA2102 2020/2021 Semester 2
Tutorial 7



STRUCTURE OF TUTORIALS

Duration:

45 mins

Content:

- Tutorial 6 (Qn 2 and 3)

Tutorial 7

DATASET REQUIRED

Tutorial6_WorldBankData.csv

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

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

- **Human.Capital.Index :** Unitless number that goes from 0 to 1.
- **GDP.per.capita.PPP:** In $. 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 $.
- **Tertiary.Education.Expenditure.per.student**. In $.
- **Population**. In people.
- **Life.Expectancy.at.birth**. In years.
- **Diabetes.Prevalence**. In units of % of population ages 20 to 79.
- **Years.of.Compulsory.Education**. In years.

This being a real dataset, there is lots of missing data. Be wary of this!

QUESTION 2A

Now let's consider another set of variables in the same dataset:

- Health.Expenditure.per.capita
- Diabetes.Prevalence, and
- Life.Expectancy.at.birth.

- Design a predictive hypothesis with these three variables.

- Which would be your dependent variable, and which would be your independent variables? Justify your answer.

$$Life.Expectancy.at.birth. = b_0 + b_1 Health.Expenditure.per.capita + b_2 Diabetes.Prevalence$$

- Value of life expectancy
- Manipulability of Health.Expenditure.per.capita
- Why not Diabetes.Prevalence?

QUESTION 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.
- For the Health.Expenditure.per.capita variable, please also apply the same transformation in (1b) for the scatterplot.
- Comment on the relationship between the variables.

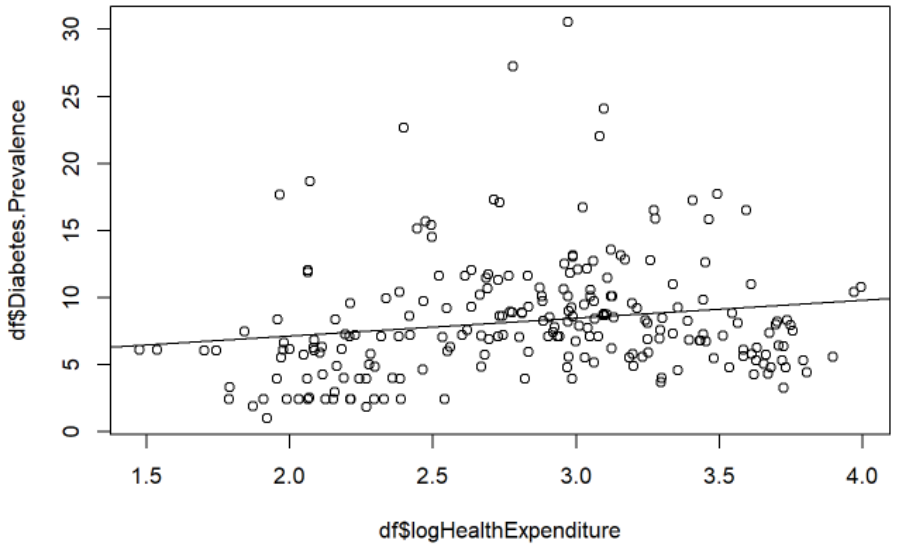
QUESTION 2B: DF$LOGHEALTHEXPENDITURE, DF$DIABETES.PREVALENCE

```
# log transform Health.Expenditure.per.capita
df$logHealthExpenditure<-log10(df$Health.Expenditure.per.capita)

plot(df$logHealthExpenditure, df$Diabetes.Prevalence)
```

- What is the nature of this relationship?

- We can draw best fit line with the [redacted] function.




```
abline(lm(Diabetes.Prevalence ~ logHealthExpenditure, data=df))
```

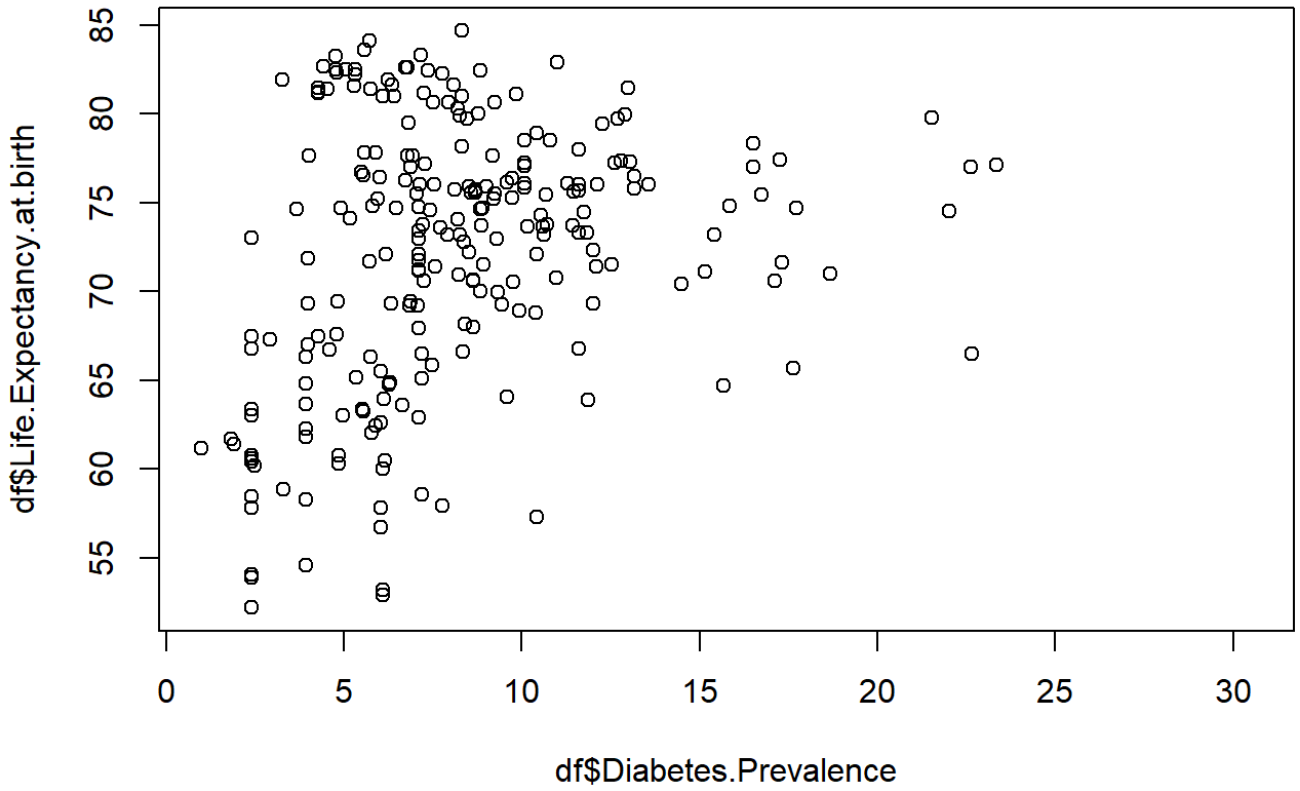
QUESTION 2B: LOGHEALTHEXPENDITURE, DF$LIFE.EXPECTANCY.AT.BIRTH

```
plot(df$logHealthExpenditure, df$Life.Expectancy.at.birth)
```

- What is the nature of this relationship?

QUESTION 2B: DF$DIABETES.PREVALENCE, DF$LIFE.EXPECTANCY.AT.BIRTH

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



- What is the nature of this relationship?

QUESTION 2C

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

QUESTION 2C

```
fit2<-lm(Life.Expectancy.at.birth ~ log10(Health.Expenditure.per.capita) + Diabetes.Prevalence, data=df)
summary(fit2)
```

```
##
## Call:
## lm(formula = Life.Expectancy.at.birth ~ log10(Health.Expenditure.per.capita) +
##     Diabetes.Prevalence, data = df)
##
## 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
```

t-distribution or z-distribution?

Based on results above? What is the result of the F-statistic?

- Interpret the intercept.
- Interpret the coefficient before log10(Health.Expenditure.per.capita).
- What is the nature of the relationship between Diabetes.Prevalence & Life.Expectancy.at.birth?

QUESTION 3

- Let's again return to **Human.Capital.Index** as our outcome of interest.

- According to the World Bank (see footnote), this index measures the amount of capital that a child born today can expect to attain by age 18, and is influenced by education and healthcare.

$$Human.Capital.Index = ? + \beta_1 Tertiary.Education.Expenditure.per.student + \beta_2 Healthcare$$

https://datacatalog.worldbank.org/dataset/human-capital-index

QUESTION 3A

A fellow student comes up with a hypothesis that quality of education should affect Human.Capital.Index.

- But something tells you that this is not so straightforward. What is the danger of putting Tertiary.Education.Expenditure.per.student into a linear model as a regressor?
- (Hint: Check its distribution by plotting and/or using the summary() function. Is there anything worth noting about this variable?) [2 marks]

$$Human.Capital.Index = \beta_0 + \beta_1 Tertiary.Education.Expenditure.per.student$$

```
summary(df$Tertiary.Education.Expenditure.per.student)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   3.121  16.412  27.276  43.677  38.411 334.000     208
```

```
sum(!is.na(df$Tertiary.Education.Expenditure.per.student))
```

```
## [1] 50
```

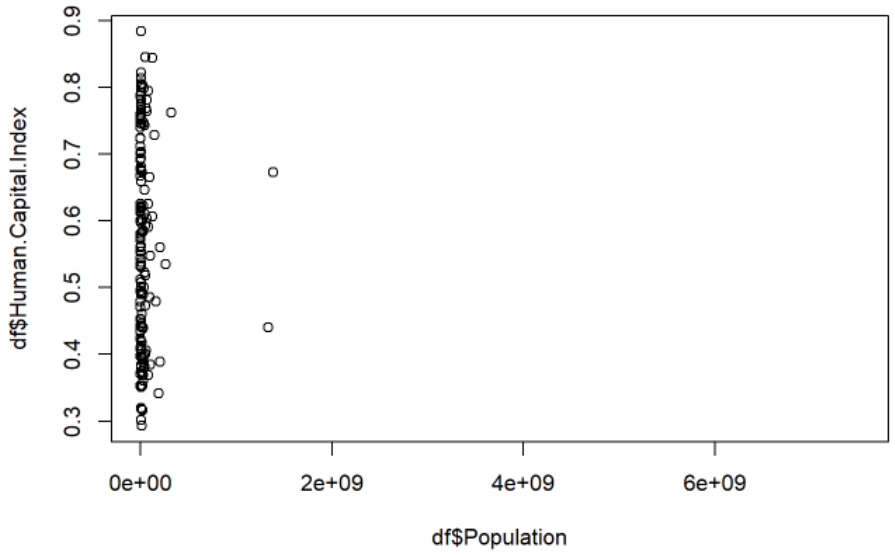
- Would you include this variable [Tertiary.Education.Expenditure.per.student] in the model?

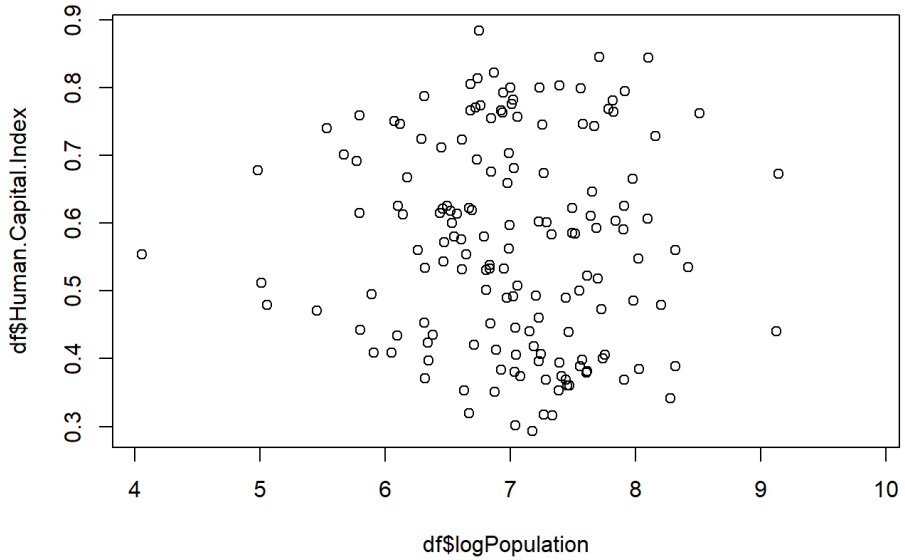
QUESTION 3B

- Is there a relationship between country population and its Human Capital Index?
- Is there a need to transform variables?
- Show this visually and using a linear model.

```
df$logPopulation=log10(df$Population)
plot(x=df$Population, y=df$Human.Capital.Index)
```

```
plot(x=df$logPopulation, y=df$Human.Capital.Index)
```





QUESTION 3B

- Is there a relationship between country population and its Human Capital Index? Is there a need to transform variables? Show this visually and using a linear model.

```
fit3<-lm(Human.Capital.Index ~ logPopulation, df)
summary(fit3)
```

```
##
## Call:
## lm(formula = Human.Capital.Index ~ logPopulation, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.272186 -0.144511 -0.001899  0.123382  0.313550
##
## Coefficients:
##                Estimate Std. Error    ?      Pr(>|t|)
## (Intercept)     0.65356    0.11074   5.902 2.19e-08 ***
## logPopulation -0.01231    0.01569  -0.785    0.434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1518 on 155 degrees of freedom
##   (101 observations deleted due to missingness)
## Multiple R-squared:  0.003959,   Adjusted R-squared:  -0.002468
## F-statistic: 0.616 on 1 and 155 DF,  p-value: 0.4337
```

t-distribution or z-distribution?

- Interpret the intercept.
- Interpret the coefficient before logPopulation

QUESTION 3C

Your manager is interested in factors that predict whether a country is above average on Human.Capital.Index or below average.

- First, do a median-split on Human.Capital.Index. Specifically, create a variable that is 1 (or TRUE) if the country's Human.Capital.Index is greater than or equal to the MEDIAN of all countries, and 0 (or FALSE) otherwise.

- Next, run a generalized linear model to predict whether a country will be above the median on Human.Capital.Index, using the following variables: GDP.per.capita.PPP, Health.Expenditure.per.capita, Life.Expectancy.at.birth and Diabetes.Prevalence.

- Apply transformations using your best judgment. [2 marks] Interpret the output of the model, and discuss the meaning of each of the coefficients. [2 marks]

QUESTION 3C

```
df$HCI.aboveMedian <- (df$Human.Capital.Index >= median(df$Human.Capital.Index, na.rm=T))

summary(glm(HCI.aboveMedian ~ log10(GDP.per.capita.PPP) + log10(Health.Expenditure.per.capita) + Life.Expectancy.
at.birth + Diabetes.Prevalence, df, family="binomial"))
```

```
##
## Call:
## glm(formula = HCI.aboveMedian ~ log10(GDP.per.capita.PPP) + log10(Health.Expenditure.per.capita) +
##     Life.Expectancy.at.birth + Diabetes.Prevalence, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4368  -0.1834   0.0396   0.2748   2.5107
##
## Coefficients:
##                                    Estimate Std. Error      ?  Pr(>|z|)
## (Intercept)                       -42.12004    9.68770  -4.348 1.38e-05 ***
## log10(GDP.per.capita.PPP)           6.13016    2.61398   2.345  0.01902 *
## log10(Health.Expenditure.per.capita) -0.95825    2.24115  -0.428  0.66896
## Life.Expectancy.at.birth            0.28392    0.10454   2.716  0.00661 **
## Diabetes.Prevalence                -0.11529    0.09372  -1.230  0.21862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 206.390  on 148  degrees of freedom
## Residual deviance:  67.554  on 144  degrees of freedom
##   (109 observations deleted due to missingness)
## AIC: 77.554
##
## Number of Fisher Scoring iterations: 7
```

t-distribution or z-distribution?

- Interpret the intercept.
- Interpret the coefficient before all the predictors

THANK YOU.

SEE YOU NEXT WEEK.