Assignment_One

Author - Eric Opoku -(Collaborators- Jessica Morgan and Chantal Valiquette)

October 7, 2022

```
name<- Sys.info()
name[7]
   user
"erico"
```

```
### Load the packages we will need for this file ####

library(AICcmodavg)  #AIC akaike information criterion
library(kableExtra)
library(pwr)
library(countrycode)
library(stargazer)
library(pwr)
library(readxl)
library(tidyverse) # load the installed package for each new session of R
library(ggdag)      # For plotting DAGs
library(dagitty)    # For working with DAG logic
library(broom)
library(faux) #  simulating data
library(modelsummary) #  regression tables
library(causaldata) #  data sets
library(here) #  directories and projects
library(plotly) #  directories and projects

set.seed(03262020) # random number generators; same numbers across machines
```

Question One

Directed acyclic graphs (DAGs) are employed to provide guidelines as to which covariates to include in the model, what sensitivity and type of analysis to conduct and the underlying assumptions to consider about the relationships between the variables.

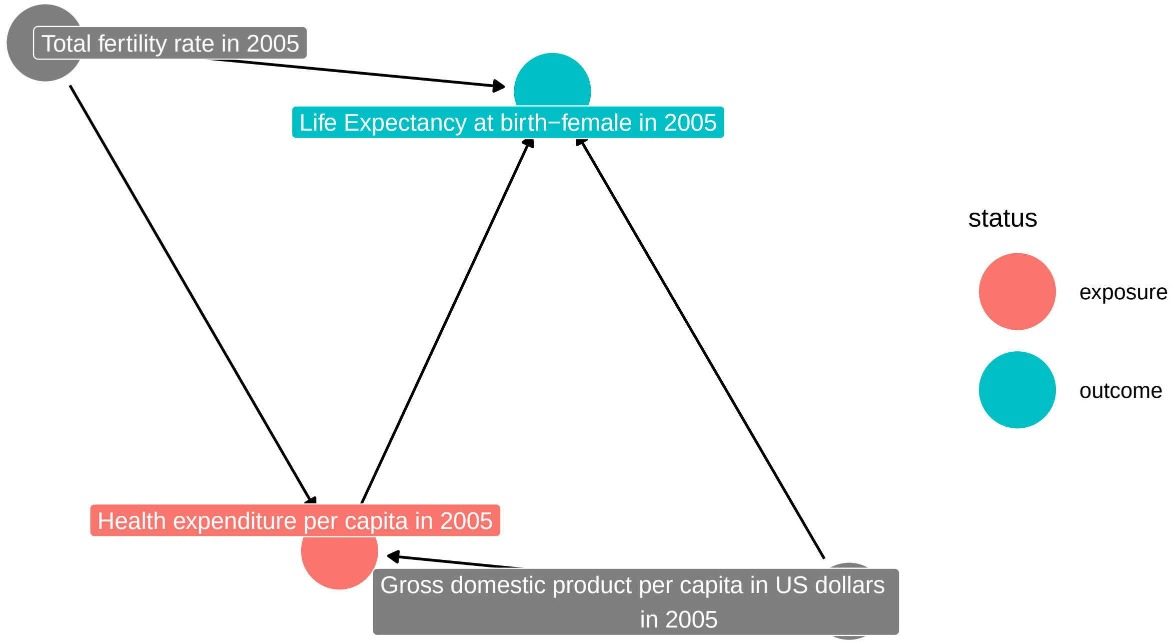
In this instance, the DAG shows that *Life Expectancy at birth for females* is the outcome or dependent variable. The *Health expenditure per capita* is the exposure (or main independent variable). *Gross domestic product per capita*, and *Total fertility rate* are the covariates which are 'controlled for' in the model.

The health literature shows that *Gross domestic product per capita*, and *Total fertility rate* have both direct effect on *Life Expectancy at birth for females* and indirect effect through *Health expenditure per capita*. With the direct effect, countries with *Gross domestic product per capita* may likely have higher standard and quality of life leading to higher *Life Expectancy at birth for females*. Also, higher *Total fertility rate* in a country may deprive females of higher standard and quality of life leading to lower *Life Expectancy at birth* for them.

With the indirect effect, higher *Gross domestic product per capita* and *Total fertility rate* may increase the financial capability and the concern to invest in health capital or increase health expenditure (for females), respectively which is also associated with *Life Expectancy at birth for females*. *Health expenditure per capita* may be associated with *Life Expectancy at birth for females* as *Health expenditure* may translated into good health (though not necessarily). It is worth stating that there is potential endogeneity in this association as some unobservable variables may affect both variables. An Instrumental variable or perhaps a 2 stage least squared approach may be appropriate for estimation.

The above arguments are all hypothesis.

```r
the_dag <- dagify(  # Create super basic DAG
  LEBF20052 ~ GDPPCUS2005 + HXPC2005 + TotFertRate2005,
  HXPC2005 ~ GDPPCUS2005+TotFertRate2005,
  exposure = "HXPC2005",
  outcome = "LEBF20052",
  labels = c(LEBF20052 = "Life Expectancy at birth-female in 2005",
             GDPPCUS2005 = "Gross domestic product per capita in US dollars
             in 2005",
             HXPC2005 = "Health expenditure per capita in 2005",
             TotFertRate2005="Total fertility rate in 2005")
)


set.seed(2545)  #prevents DAG shape from changing
ggdag_status(the_dag, use_labels= "label",
             text = FALSE) + theme_dag()
```


Total fertility rate in 2005
Life Expectancy at birth−female in 2005
Health expenditure per capita in 2005
Gross domestic product per capita in US dollars in 2005
status
exposure
outcome

```
# Adding a theme_dag() layer makes it have a white background; no axis labels
```

Question Two

In the regression output, we cannot talk about causation. Based on the DAG, we can only say that when we control for *Gross domestic product per capita* and *Total fertility rate*, there is an association (or not) between *Health expenditure per capita* and *Life Expectancy at birth for females.* This is because the model is not robust as it may require an instrumental variable for *Health expenditure per capita.* There is also potential endogeneity, and simultaneity bias in the model where some unobserved variables may affect both *Health expenditure per capita* and *Life Expectancy at birth for females.*

The DAG only shows that there could other complex relationships that could be elicited in a model hence one should be careful to interpret a model based on causality. I did not add other variables because they failed to prove economic significance.

```
mydata$HXPC2005<-as.numeric(mydata$HXPC2005)  #convert from character to numeric
Warning: NAs introduced by coercion
head(mydata$HXPC2005)
[1]  32.52102 171.86631 111.22996  37.99327 492.51460  87.83534
class(mydata$HXPC2005)
[1] "numeric"
lm_example<-lm(LEBF20052 ~ mydata$HXPC2005+ mydata$GDPPCUS2005 +
                 mydata$TotFertRate2005, data = mydata) #Linear regression model
summary(lm_example)

Call:
lm(formula = LEBF20052 ~ mydata$HXPC2005 + mydata$GDPPCUS2005 +
    mydata$TotFertRate2005, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-25.913  -4.466   0.603   4.736  18.938

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            84.7377139  1.4585234  58.098   <2e-16 ***
mydata$HXPC2005        -0.0009254  0.0011530  -0.803   0.4233
mydata$GDPPCUS2005      0.0002221  0.0001076   2.063   0.0406 *
mydata$TotFertRate2005 -5.3902988  0.3747854 -14.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.914 on 170 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6626,    Adjusted R-squared:  0.6566
F-statistic: 111.3 on 3 and 170 DF,  p-value: < 2.2e-16
```

```
#msummary(list("Simple"=lm_example),
 #         stars=c('*' = .1, '**' = .05, '***' = .01))  #regression table
```

Question Three

YES. The mean and standard deviation for *Life Expectancy at birth for females*, *Gross domestic product per capita* and *Health expenditure per capita* are huge. Perhaps, they should be transformed using a logarithm.

```
class(mydata$LEBF20052)     #summary(mydata$LEBF20052)   #View(mean_LEBF20052)
[1] "numeric"
mean_LEBF20052<-mean(mydata$LEBF20052, na.rm = TRUE)
print(mean_LEBF20052)   #69.85832
[1] 69.85832
sd_LEBF20052<-sd(mydata$LEBF20052, na.rm = TRUE)    #View(sd_LEBF20052)
print(sd_LEBF20052)   #11.88807
[1] 11.88807
count(mydata, LEBF20052, na.rm = TRUE)  #175 #sample size of life expectancy
# A tibble: 175 x 3
   LEBF20052 na.rm     n
       <dbl> <lgl> <int>
 1      41.1 TRUE      1
 2      42.7 TRUE      1
 3      43.4 TRUE      1
 4      43.5 TRUE      1
 5      45.5 TRUE      1
 6      46.0 TRUE      1
 7      46.3 TRUE      1
 8      46.4 TRUE      1
 9      46.9 TRUE      1
10      47.8 TRUE      1
# ... with 165 more rows



class(mydata$GDPPCUS2005)
[1] "numeric"
mean_GDPPCUS2005<-mean(mydata$GDPPCUS2005, na.rm = TRUE)
print(mean_GDPPCUS2005)   #9862.103
[1] 9862.103
sd_GDPPCUS2005<-sd(mydata$GDPPCUS2005, na.rm = TRUE)   #View(sd_GDPPCUS2005)
print(sd_GDPPCUS2005)   #16195
[1] 16195

#sample size of Gross domestic product per capita
count(mydata, GDPPCUS2005, na.rm = TRUE) #175  #sample size of life expectancy
# A tibble: 175 x 3
```

```
   GDPPCUS2005 na.rm     n
         <dbl> <lgl> <int>
 1        108. TRUE      1
 2        120. TRUE      1
 3        159. TRUE      1
 4        165. TRUE      1
 5        205. TRUE      1
 6        209. TRUE      1
 7        243. TRUE      1
 8        254. TRUE      1
 9        254. TRUE      1
10        262. TRUE      1
# ... with 165 more rows



class(mydata$HXPC2005)   #sample size of Health expenditure per capita
[1] "numeric"
mydata$HXPC2005<-as.numeric(mydata$HXPC2005)  #convert from character to numeric

mean_HXPC2005<-mean(mydata$HXPC2005, na.rm = TRUE)  #summary(mydata$HXPC2005)
print(mean_HXPC2005)  #713.3919
[1] 713.3919

sd_HXPC2005<-sd(mydata$HXPC2005, na.rm = TRUE)  #View(sd_HXPC2005)
print(sd_HXPC2005) #1329.515
[1] 1329.515
#sample size of Health expenditure per capita
count(mydata, HXPC2005, na.rm = TRUE) #174  #sample size of Health expenditure per capi
# A tibble: 175 x 3
   HXPC2005 na.rm     n
      <dbl> <lgl> <int>
 1     6.76 TRUE      1
 2     7.46 TRUE      1
 3     7.91 TRUE      1
 4    10.6  TRUE      1
 5    11.8  TRUE      1
 6    11.8  TRUE      1
 7    13.3  TRUE      1
 8    13.5  TRUE      1
 9    14.0  TRUE      1
10    14.1  TRUE      1
# ... with 165 more rows
```

```
summary(mydata$TotFertRate2005)  #Total fertility rate    #View(mydata)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.080   1.796   2.500   3.037   4.140   7.267 
class(mydata$TotFertRate2005)  #View(mean_TotFertRate2005)
[1] "numeric"
mean_TotFertRate2005<-mean_TotFertRate2005<-mean(mydata$TotFertRate2005,
                                                 na.rm = TRUE)
print(mean_TotFertRate2005)
[1] 3.036565
sd_TotFertRate2005<-sd(mydata$TotFertRate2005, na.rm = TRUE)
print(sd_TotFertRate2005)
[1] 1.571992
#View(sd_TotFertRate2005)
count(mydata, TotFertRate2005, na.rm = TRUE) #165
# A tibble: 165 x 3
   TotFertRate2005 na.rm     n
             <dbl> <lgl> <int>
 1            1.08 TRUE      1
 2            1.2  TRUE      1
 3            1.21 TRUE      1
 4            1.22 TRUE      1
 5            1.24 TRUE      1
 6            1.25 TRUE      1
 7            1.26 TRUE      3
 8            1.27 TRUE      1
 9            1.28 TRUE      1
10            1.29 TRUE      1
# ... with 155 more rows
```

```
desc_table = data.frame(
  Measure = c("Life_Expectancy_females", "GDP_per_capita",
              "Health_expenditure_per_capita", "Total_fertility_rate"),
  M_1   = c(mean(mydata$LEBF20052), mean(mydata$GDPPCUS2005), 713.39,
           mean(mydata$TotFertRate2005)),
  SD_1  = c(sd(mydata$LEBF20052), sd(mydata$GDPPCUS2005), 1329.52,
           sd(mydata$TotFertRate2005)), #there were NA's for health exp
  SS   = c(175, 175, 174,165))

#desc_table


kable(
  desc_table,
```

Table 1: Means, Standard Deviations and Sample size of Four Variables

Measure	*M*	*SD*	*Sample Size*
Life_Expectancy_females	69.86	11.89	175
GDP_per_capita	9862.10	16195.00	175
Health_expenditure_per_capita	713.39	1329.52	174
Total_fertility_rate	3.04	1.57	165

```
  col.names = c("Measure", "*M*", "*SD*", "*Sample Size*"),
  digits = 2,
  caption = "Means, Standard Deviations and Sample size of Four Variables"
  )
```

Question Four

Statistical significance is when one uses statistical tools such as p-value to conclude as to whether a relation between some variables makes sense or is significant or acceptable. Economic significance is when one uses economic theory to deem whether same relationship makes sense based on the story it tells.

Based on the p-value of the model, statistically, *Health expenditure per capita* has positive and significant association with *Life Expectancy at birth for females* (i.e., statistical significance). Thus, a 1000 Dollar increase in *Health expenditure per capita* is significantly associated with 4 years increase in *Life Expectancy at birth for females*.

From economics point of view, it is meaningful because, all other things being equal, higher health expenditure per capita may lead to good health and consequently, higher *Life Expectancy at birth for females* (i.e., economic significance).

It is however worth noting here that the model is not robust enough as *Health expenditure per capita* has a very low standard error. A low standard error (i.e., 0.001) means high t-statistics which means low p-value (i.e., o.000) and higher 'probability of significance' for the *Health expenditure per capita* variable meanwhile the R-squared is low (i.e., 0.2).

```
lm_simp<-lm(LEBF20052 ~ HXPC2005, data = mydata)


summary(lm_simp)

Call:
lm(formula = LEBF20052 ~ HXPC2005, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.910  -6.758   1.611   8.310  19.118 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.688e+01  9.097e-01  73.517  < 2e-16 ***
HXPC2005    3.995e-03  6.043e-04   6.612 4.59e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.57 on 172 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2027,    Adjusted R-squared:  0.198 
F-statistic: 43.72 on 1 and 172 DF,  p-value: 4.592e-10

msummary(list("Simple"=lm_simp),
         caption = "Coefficient-Level Estimates for a Model Fitted to Estimate Life
    Expectancy at birth for females.",
    statistic = c("conf.int",
            "standard error = {std.error}",
             "p-value = {p.value}"),
         stars=c('*' = .1, '**' = .05, '***' = .01))  #regression table
```

```
#lm(LEBF20052 ~ HXPC2005, data = mydata) %>%
 # tidy() %>%
  #kable(
   # caption = "Coefficient-Level Estimates for a Model Fitted to Estimate Life
    #Expectancy at birth for females.",
    #col.names = c("Predictor", "B", "SE", "t", "p"),
    #digits = c(0, 3, 3, 2, 4))
```

Table 2: Coefficient-Level Estimates for a Model Fitted to Estimate Life Expectancy at birth for females.

	Simple
(Intercept)	66.879*** [65.084, 68.675] standard error = 0.910 p-value = 0.000
HXPC2005	0.004*** [0.003, 0.005] standard error = 0.001 p-value = 0.000
Num.Obs.	174
R2	0.203
R2 Adj.	0.198
AIC	1318.3
BIC	1327.8
Log.Lik.	−656.137
F	43.718
RMSE	10.51

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Question Five

After controlling for *Gross domestic product per capita*, it is now realized that the association between *Health expenditure per capita* and *Life Expectancy at birth for females* was no more significant and the coefficient also reduced from 0.004 (Model A) to -0.001 (Model B). This could mean that *Health expenditure per capita* was only mediating the (significant) association between *Gross domestic product per capita* and *Life Expectancy at birth for females* in *model A*.

Now, *Health expenditure per capita* is associated with a higher standard errors comparative to *model A* (i.e., 0.002 in *model B*). Also, *model B* has a relatively higher R-squared compared to *model A*.

In the *model B*, a 1000 Dollar increase in *Gross domestic product per capita* is (significantly) associated with one year increase in *Life Expectancy at birth for females.* In the *model A*, a 1000 Dollar increase in *Health expenditure per capita* was (significantly) associated with 4 years increase in *Life Expectancy at birth for females.*

```r
lm_simp<-lm(LEBF20052 ~ HXPC2005, data = mydata)

lm_comp<-lm(LEBF20052 ~ HXPC2005 + GDPPCUS2005, data = mydata) #summary(lm_comp)

stargazer(
  lm_simp, lm_comp,
  type = "latex",
  title = "Two Regression Models Predicting Life Expectancy at birth for
  females",
  column.labels = c("Model A", "Model B"),
  colnames = FALSE,
  model.numbers = FALSE,
  dep.var.caption = " ",
  dep.var.labels = "Life Expectancy at birth for females",
  covariate.labels = c("Health expenditure per cap","GDP per capita"),
  keep.stat = c("rsq", "f"),
  notes.align = "l",
  out = "latex"
  )
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Oct 07, 2022 - 12:52:58 PM

Table 3: Two Regression Models Predicting Life Expectancy at birth for females

	Life Expectancy at birth for females	
	Model A	Model B
Health expenditure per cap	0.004^{***}	-0.001
	(0.001)	(0.002)
GDP per capita		0.001^{***}
		(0.0002)
Constant	66.879^{***}	65.828^{***}
	(0.910)	(0.937)
R^2	0.203	0.252
F Statistic	43.718^{***} (df = 1; 172)	28.813^{***} (df = 2; 171)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Question Six

To deal with outliers and skewness, I would recommend a logarithmic transformation of both *Gross domestic product per capita* and *Health expenditure per capita.* Thus, diminishing marginal utility of income suggests that a change in levels of *Gross domestic product (i.e., income) per capita* and *Health expenditure per capita* may impact a rich country lesser relative to a poor country. Hence, the best way to compare their impact on both countries is to use the percentage changes in both variables (i.e., logarithm).

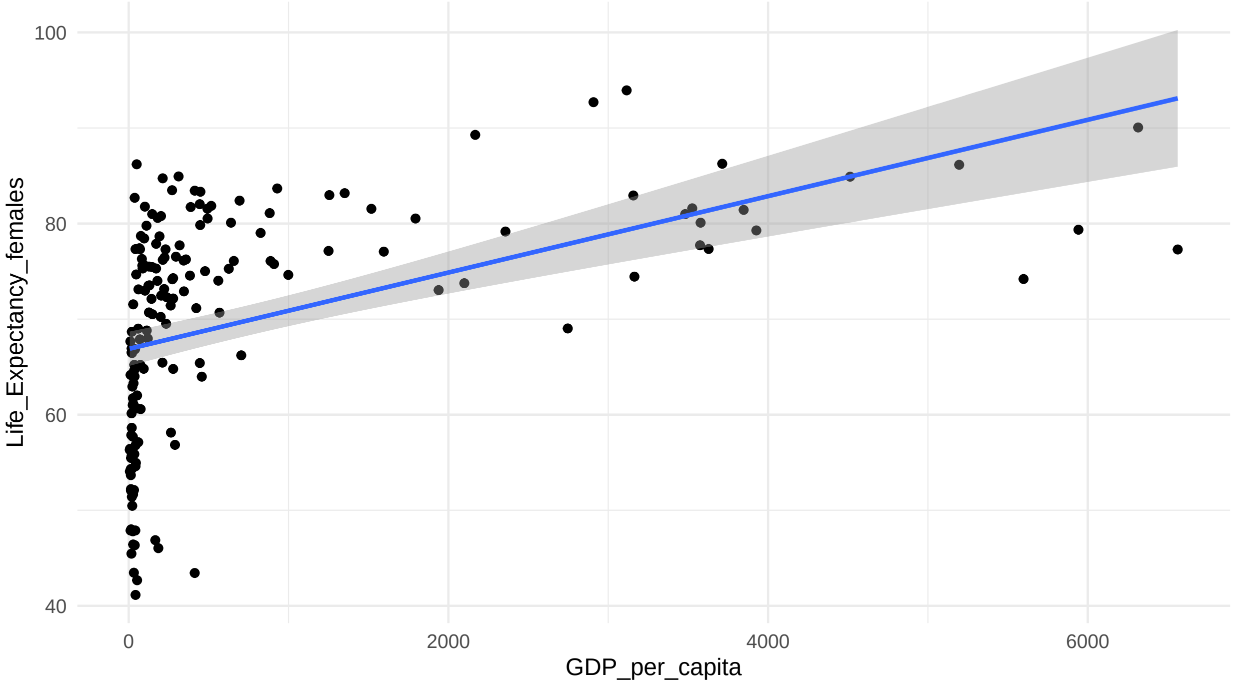
The R-squared for the previous *models A and B* (i.e., the levels models) more than doubled after taking the natural log of the variables from 0.203 and 0.252 to 0.475 and 0.493 in *models C and D*, respectively. This depicts that taking the natural log of the variables improved the fitness of the models.

The standard errors for *models C and D* increased (or improved) relative to *models A and B.*

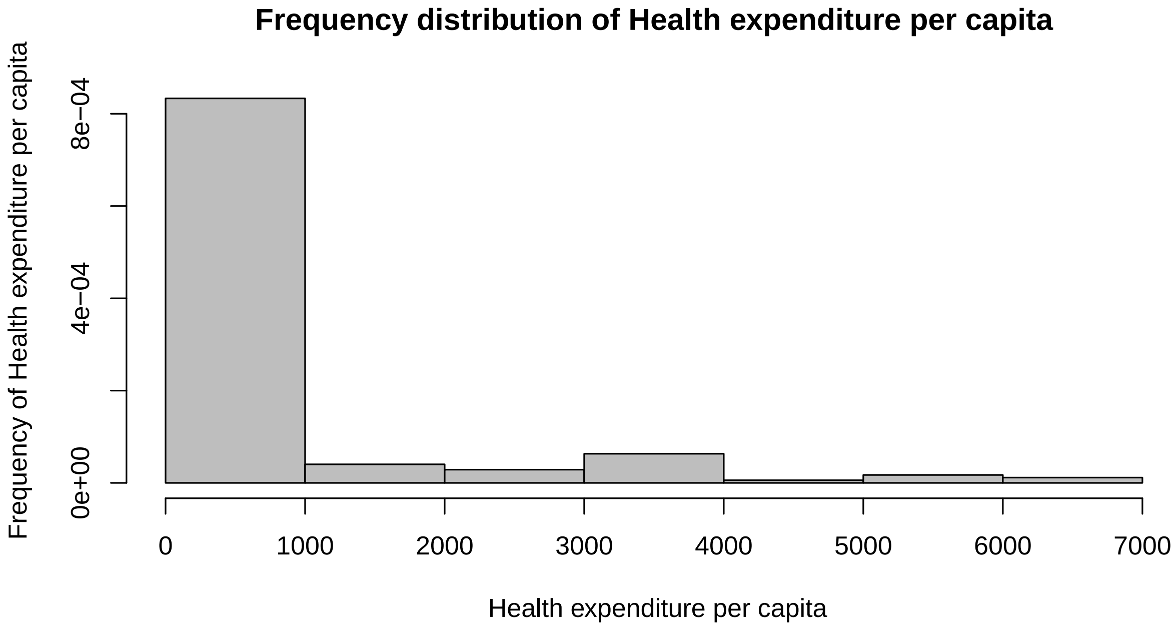
In the *model C*, the association between *log of Health expenditure per capita* and *Life Expectancy at birth for females* is (significantly) higher (i.e., 4.607) as compared to its counterpart in *model A* (i.e., 0.004).

Also, in *model D*, the association between *log of GDP per capita* and *Life Expectancy at birth for females* is (significantly) higher (i.e., 4.114) relative to its counterpart in *model B* (i.e., 0.001). It is worth noting here that *Health expenditure per capita* was not significant when both *models B and D* controlled for *GDP per capita.*

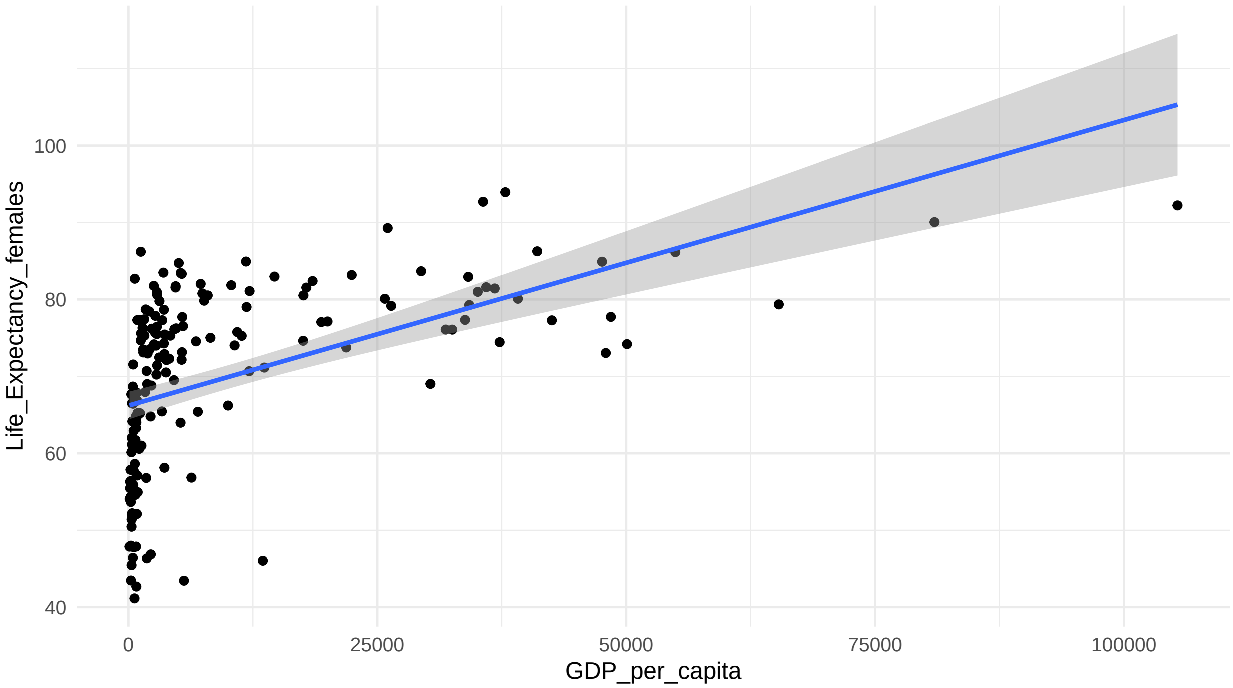
```
# health expenditure #outliers
ggplot(data=mydata,aes(x=HXPC2005,y=LEBF20052)) + geom_point() +
theme_minimal() + labs(x="GDP_per_capita",y="Life_Expectancy_females") +
geom_smooth(method='lm', formula= y~x) #potential nonlinearity
Warning: Removed 1 rows containing non-finite values (stat_smooth).
Warning: Removed 1 rows containing missing values (geom_point).
```



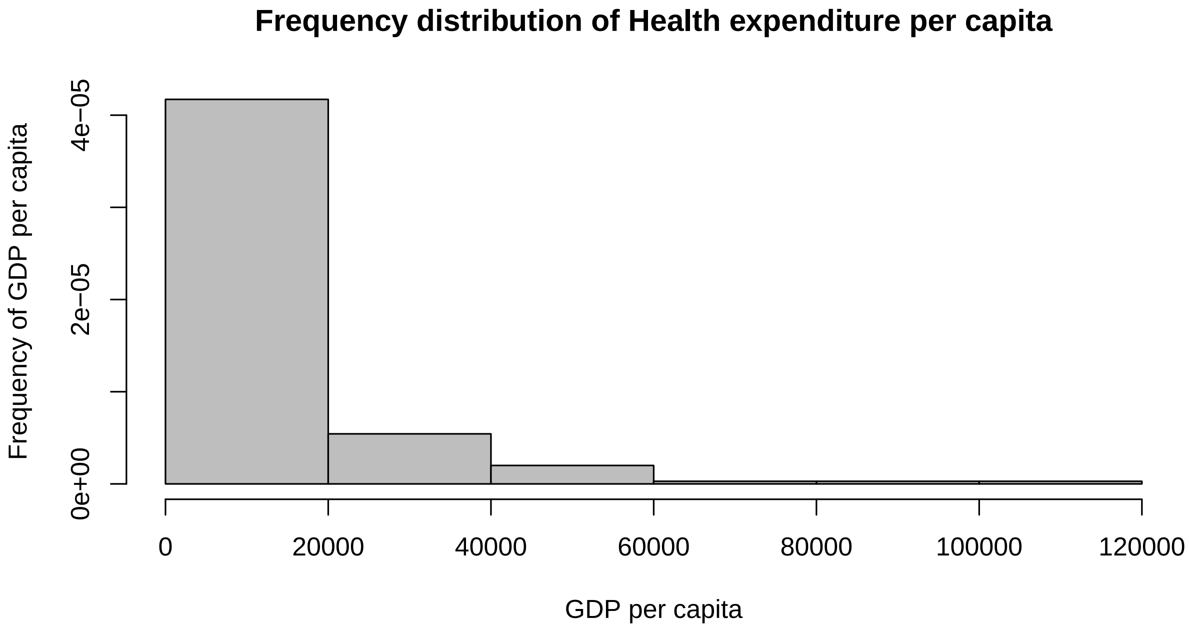
```


# health expenditure   #very skewed

h_HXPC2005<-hist(mydata$HXPC2005, breaks=5, prob=T,
                 main="Frequency distribution of Health expenditure per capita",
         ylab="Frequency of Health expenditure per capita",
         xlab=" Health expenditure per capita",
                    col="grey")
```



```
#log it



#GDP outliers
ggplot(data=mydata,aes(x=GDPPCUS2005,y=LEBF20052)) + geom_point() +
theme_minimal() + labs(x="GDP_per_capita",
y="Life_Expectancy_females") +
geom_smooth(method='lm', formula= y~x)
```



```
#GDP SKEWNESS
h_GDPPCUS2005<-hist(mydata$GDPPCUS2005, breaks=5, prob=T,
                main="Frequency distribution of Health expenditure per capita",
         ylab="Frequency of GDP per capita",
         xlab="GDP per capita",
                   col="grey")
```



```
#logarithm
#repeating previous regression with log of both GDP and HXPC
lm_simp_log<-lm(LEBF20052 ~ log(HXPC2005), data = mydata)

lm_comp_log<-lm(LEBF20052 ~ log(HXPC2005) + log(GDPPCUS2005), data = mydata) #summary(l


stargazer(
  lm_simp_log, lm_comp_log,
  type = "latex",
  title = "Two Regression Models Predicting Life Expectancy at birth for
  females (transformed)",
  column.labels = c("Model C", "Model D"),
  colnames = FALSE,
  model.numbers = FALSE,
  dep.var.caption = " ",
  dep.var.labels = "Life Expectancy at birth for females (transformed)",
  covariate.labels = c("log Health expenditure per cap", "log GDP per cap"),
  keep.stat = c("rsq", "f"),
  notes.align = "l",
  out = "latex"
  )
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Oct 07, 2022 - 12:52:59 PM

Table 4: Two Regression Models Predicting Life Expectancy at birth for females (transformed)

	Life Expectancy at birth for females (transformed)	
	Model C	Model D
log Health expenditure per cap	4.607***	0.889
	(0.369)	(1.541)
log GDP per cap		4.114**
		(1.657)
Constant	46.151***	32.448***
	(1.999)	(5.860)
R^2	0.475	0.493
F Statistic	155.523*** (df = 1; 172)	83.179*** (df = 2; 171)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Question Seven

After controlling for the geography dummy, the R-squared increased from 0.475 and 0.493 in the previous *models C and D* to 0.63 and 0.641 in the current *models E and F*, respectively. Thus, the model further improved.

The results depict that Africa is significantly associated with lesser *Life Expectancy at birth for females* relative to its Asian counterpart (i.e., reference variable). I am careful to explain the coefficients here since they are not marginal effects. However, for the sake of the assignment, I would say that being the African continent is significantly associated with 12 and 11.6 years lesser *Life Expectancy at birth for females* relative to their Asian counterparts for *models E and F*, respectively. The other variables were not significant and hence, i will not report them.

In addition, the association between *log of GDP per capita* and *Life Expectancy at birth for females* was positive and significant (i.e., Not marginal effects). For the sake of the assignment, I would say that a one percentage change in *log of GDP per capita* leads to a significantly 0.03 (i.e., 3.320/100) increase in *Life Expectancy at birth for females*.

Moreover, the significant impact or coefficients of *log of Health capital expenditure per capita* and the *log of GDP per capita* from the previous *models C and D* decreased significantly in the current *models E and F*. For instance, with regards to *log of Health capital expenditure per capita*, its association with *Life Expectancy at birth for females* decreased from 4.607 in *model C* to 3.195 in *model E* and that of *log of GDP per capita* decreased from 4.114 in *model D* to 3.32 in *model F*. This depicts that the coefficients in the previous *models C and*

D (without controlling for geography dummy) were overstated.

```
mydata$continent <- countrycode(sourcevar = mydata$Country,
                                origin = "country.name",
                                destination = "continent")

mydata$africa<-ifelse(mydata$continent=="Africa",yes= "1",no= "0")
count(mydata, africa, na.rm = TRUE) #52
```

A tibble: 2 x 3

africa na.rm n 1 0 TRUE 123 2 1 TRUE 52

```
mydata$asia<-ifelse(mydata$continent=="Asia",yes= "1",no= "0")
count(mydata, asia, na.rm = TRUE) #44
```

A tibble: 2 x 3

asia na.rm n 1 0 TRUE 131 2 1 TRUE 44

```
#i choose asia as dummy because they have second largest obs.

mydata$europe<-ifelse(mydata$continent=="Europe",yes= "1",no= "0")
count(mydata, europe, na.rm = TRUE) #40
```

A tibble: 2 x 3

europe na.rm n 1 0 TRUE 135 2 1 TRUE 40

```
mydata$americas<-ifelse(mydata$continent=="Americas",yes= "1",no= "0")
count(mydata, americas, na.rm = TRUE) #30
```

A tibble: 2 x 3

americas na.rm n 1 0 TRUE 145 2 1 TRUE 30

```r
mydata$oceania<-ifelse(mydata$continent=="Oceania",yes= "1",no= "0")
count(mydata, oceania, na.rm = TRUE) #9
```

A tibble: 2 x 3

oceania na.rm n 1 0 TRUE 166 2 1 TRUE 9

```r

#repeating Ques six regression
#dummy continent
#repeating previous regression with dummy continent
lm_simp_log_cont<-lm(LEBF20052 ~ log(HXPC2005)+ africa + oceania + europe
                     + americas, data = mydata)

lm_comp_log_cont<-lm(LEBF20052 ~ log(HXPC2005)+ africa + oceania + europe
                     + americas
                + log(GDPPCUS2005), data = mydata) #summary(lm_comp)



stargazer(
 lm_simp_log_cont, lm_comp_log_cont,
  type = "latex",
  title = "Two Regression Models Predicting Life Expectancy at birth for
  females (geography)",
  column.labels = c("Model E", "Model F"),
  colnames = FALSE,
  model.numbers = FALSE,
  dep.var.caption = " ",
  dep.var.labels = "Life Expectancy at birth for females (geography)",
  covariate.labels = c("log Health expenditure per cap", "Africa",
                       "Oceania", "Europe", "Americas",
                       "log GDP per cap"),
  keep.stat = c("rsq", "f"),
  notes.align = "l",
  out = "latex"
  )
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Oct 07, 2022 - 12:53:03 PM

Table 5: Two Regression Models Predicting Life Expectancy at birth for females (geography)

	Life Expectancy at birth for females (geography)	
	Model E	Model F
log Health expenditure per cap	3.195***	0.086
	(0.426)	(1.449)
Africa	−12.122***	−11.572***
	(1.567)	(1.568)
Oceania	−1.635	−1.240
	(2.671)	(2.645)
Europe	−1.765	−0.544
	(1.843)	(1.901)
Americas	−0.202	0.595
	(1.753)	(1.769)
log GDP per cap		3.320**
		(1.481)
Constant	57.515***	46.419***
	(2.332)	(5.461)
R^2	0.630	0.641
F Statistic	57.178*** (df = 5; 168)	49.627*** (df = 6; 167)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Question Eight

With this interaction, I am measuring the unit change in *Life Expectancy at birth for females* when the *log of Health expenditure per capita* changes by let's say 10 percent for Africa relative to their Asian counterparts.

It is meaningful to measure the above because 'returns' to *Health expenditure per capita* might be different for different continents. For instance, given a (relatively) lower investment in *Health expenditure per capita* in Africa, any change in *Health expenditure per capita* may more likely lead to higher returns relative to similar investment abroad or in Asia.

This analysis is based on 'the law of diminishing returns'. Thus, all other things being equal, as investment in health or health expenditure increases, returns to the investment increases to a point and cannot continue to increase anymore or stagnates or increases at a decreasing rate. In other words, continents such as Americas or Asia might have invested in health capital or have higher health expenditure relative to *Africa* especially due to drugs and technological costs and innovations. Hence, their 'returns to health capital investment/expenditure may be minimal relative to Africa.

Evidence from *models G and H* confirms the above hypothesis, though it is insignificant. Thus, a one percentage increase in the *log of Health expenditure per capita* increases *Life Expectancy at birth for females* by 0.0058 and 0.0032 in *models G and H*, respectively for Africa relative to their Asian counterparts.

-

```
#repeating Ques seven regression with interac
#repeating previous regression with interaction term

lm_simp_log_cont_int<-lm(LEBF20052 ~ log(HXPC2005) + africa + oceania + europe
                    + americas + log(HXPC2005):africa, data = mydata)

print(lm_simp_log_cont_int)
```

```
Call: lm(formula = LEBF20052 ~ log(HXPC2005) + africa + oceania + europe + americas
+ log(HXPC2005):africa, data = mydata)

Coefficients: (Intercept) log(HXPC2005) africa1
58.1570 3.0618 -14.4321
oceania1 europe1 americas1
-1.5809 -1.4812 -0.1032
log(HXPC2005):africa1
0.5836
```

```
lm_comp_log_cont_int<-lm(LEBF20052 ~ log(HXPC2005) + africa + oceania + europe
                    + americas
               + log(GDPPCUS2005)+ log(HXPC2005):africa, data = mydata) #summary(lm_con

print(lm_comp_log_cont_int)
```

Call: lm(formula = LEBF20052 ~ log(HXPC2005) + africa + oceania + europe + americas + log(GDPPCUS2005) + log(HXPC2005):africa, data = mydata)

Coefficients: (Intercept) log(HXPC2005) africa1
46.95740 0.06629 -12.84380
oceania1 europe1 americas1
-1.21749 -0.40990 0.63521
log(GDPPCUS2005) log(HXPC2005):africa1
3.26349 0.31885

```
stargazer(
  lm_simp_log_cont_int, lm_comp_log_cont_int,
  type = "latex",
  title = "Three Regression Models Predicting Life Expectancy at birth for
  females (Africa dummy)",
  column.labels = c("Model G", "Model H"),
  colnames = FALSE,
  model.numbers = FALSE,
  dep.var.caption = " ",
  dep.var.labels = "Life Expectancy at birth for females (Africa dummy)",
  covariate.labels = c("log of Health expenditure per cap", "Africa",
                       "Oceania", "Europe", "Americas",
                       "log GDP per cap sq", "log of Health Expenditure-Africa"),
  keep.stat = c("rsq", "f"),
  notes.align = "l",
  out = "latex"
  )
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Oct 07, 2022 - 12:53:04 PM

Table 6: Three Regression Models Predicting Life Expectancy at birth for females (Africa dummy)

	Life Expectancy at birth for females (Africa dummy)	
	Model G	Model H
log of Health expenditure per cap	3.062^{***}	0.066
	(0.486)	(1.455)
Africa	-14.432^{***}	-12.844^{***}
	(4.326)	(4.340)
Oceania	-1.581	-1.217
	(2.678)	(2.653)
Europe	-1.481	-0.410
	(1.912)	(1.953)
Americas	-0.103	0.635
	(1.765)	(1.778)
log GDP per cap sq		3.263^{**}
		(1.496)
log of Health Expenditure-Africa	0.584	0.319
	(1.018)	(1.014)
Constant	58.157^{***}	46.957^{***}
	(2.592)	(5.737)
R^2	0.631	0.641
F Statistic	47.512^{***} (df = 6; 167)	42.322^{***} (df = 7; 166)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Question Nine

Establishing causal relationship is difficult in the model because there are other covariates that may not have been controlled for (e.g., omitted variable bias). For instance, the Akaike Information criterion (AIC) showed that a model gets better the extent we make it better. Thus, when it comes to modelling, we can always do better. The model listed first always has the lowest AIC value and is thus the best fitting model as far these particular models are concerned. The best model (Model F) was when we controlled for relevant variables such as the continents and the *log of GDP per capita*.

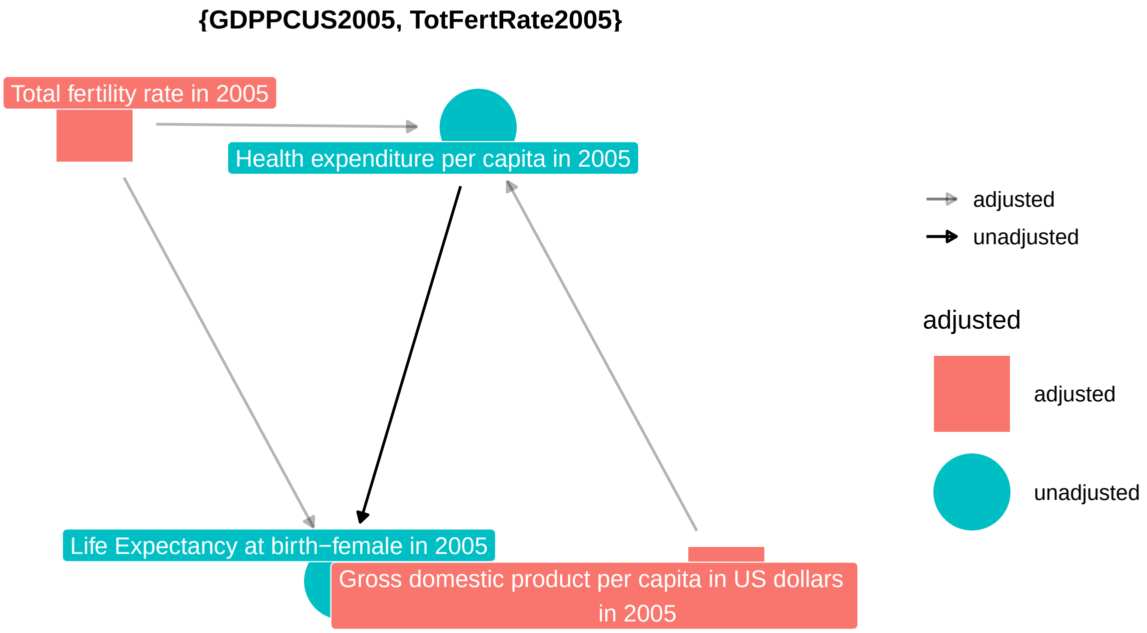
In addition, we cannot be certain that our outcome and exposure variables do not simultaneously impact each other (i.e., simultaneity bias). One can argue that same way countries that have higher *log of Health expenditure per capita* may have longer *life expectancy for females*, countries who have longer *life expectancy for females* are more likely to have higher *log of Health expenditure per capita*

Also, it could be that our model is impacted by other variables which are ignored from the model (i.e., potential endogeneity). For instance, there may be other unobserved factors that affect both the *life expectancy for females* and *log of Health expenditure per capita* that are not 'controlled for' in the model. Thus, a country like Americas may inherently cherish longevity relative to other continents and even though this cannot be measured due to lack of data availability, it may affect both the *life expectancy for females* and *log of Health expenditure per capita*. In addition, the figure depicts that *Health expenditure per capita* is unadjusted. Thus, the model may require an instrumental variable. This is the basic form of showing that some variables are not added in the model.

Moreover, there independent variables of a model could be associated with each other (e.g., multicollinearity).

By and large, the more we make the model robust, the more it becomes representative of the data. So we cannot conclude causality same way we can never know whether our model is the best it can ever be. This is the reason why economists do a lot of post-estimation testing.

```r
#stargazer(lm_comp, type = 'text')

#lm_full <- lm(LEBF20052 ~ HXPC2005 + GDPPCUS2005 + TotFertRate2005, data=mydata)

set.seed(33)  #prevents DAG shape from changing
ggdag_adjustment_set(the_dag, shadow = TRUE, text = FALSE,
                     use_labels="label") + theme_dag()
```


{GDPPCUS2005, TotFertRate2005}
Total fertility rate in 2005
Health expenditure per capita in 2005
Life Expectancy at birth−female in 2005
Gross domestic product per capita in US dollars in 2005
adjusted
unadjusted
adjusted
adjusted
unadjusted

```
#Let's do some little proof with the Akaike Information criterion (AIC)


models<-list(lm_simp, lm_simp_log, lm_simp_log_cont, lm_simp_log_cont_int, lm_comp, lm_c

mod.names<- c('lm_simp', 'lm_simp_log', 'lm_simp_log_cont', 'lm_simp_log_cont_int', 'lm_

aictab(cand.set = models, modnames = mod.names)

Model selection based on AICc:

                     K    AICc Delta_AICc AICcWt Cum.Wt      LL
lm_comp_log_cont     8 1190.46       0.00   0.60   0.60 -586.79
lm_comp_log_cont_int 9 1192.58       2.12   0.21   0.81 -586.74
lm_simp_log_cont     7 1193.42       2.96   0.14   0.95 -589.37
lm_simp_log_cont_int 8 1195.28       4.82   0.05   1.00 -589.20
lm_comp_log          4 1241.69      51.23   0.00   1.00 -616.73
lm_simp_log          3 1245.76      55.30   0.00   1.00 -619.81
lm_comp              4 1309.38     118.93   0.00   1.00 -650.57
lm_simp              3 1318.42     127.96   0.00   1.00 -656.14


#we can always do better when it comes to modelling
```

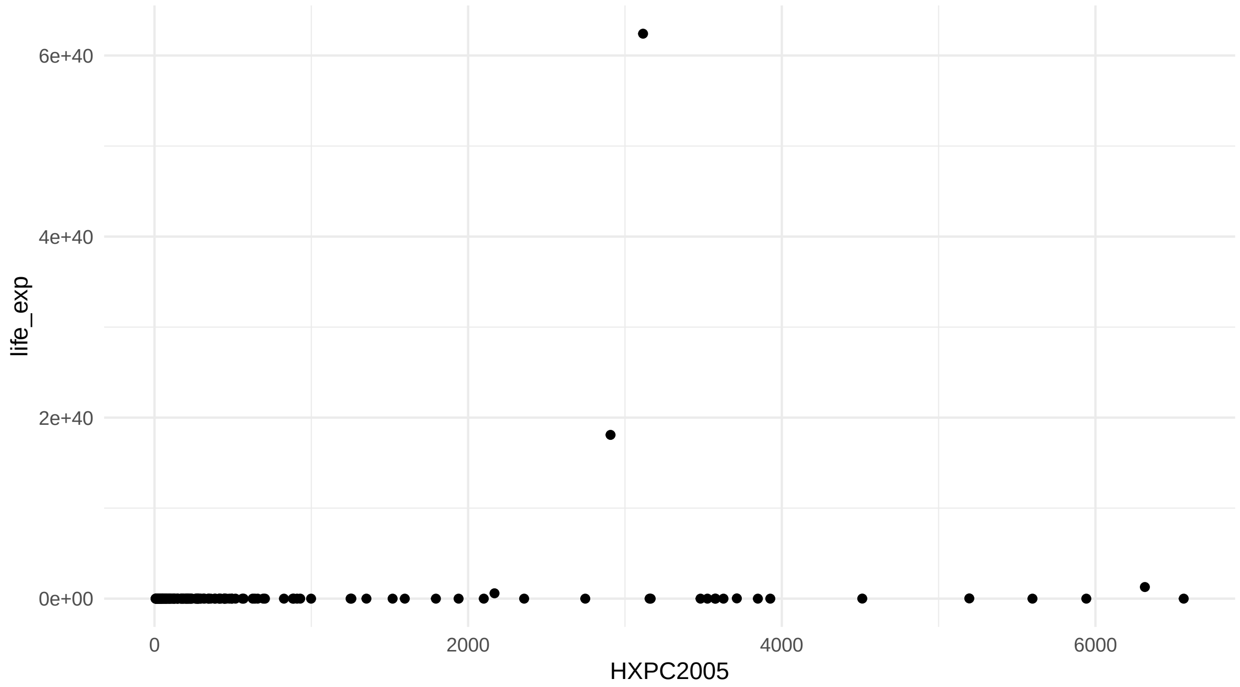
```
#lm_full_sq_conti is the best model
```

Question Ten

Based on the point made in question 9, it will always be advisable to run models with robust standard errors or take care of clustering. Hence, I will re-run the model with robust standard errors.

It is worth noting that if standard errors are not robust, then variables which were not supposed to be significant will end up being significant. For instance, in our model, we realize that the *classical* model has a relatively lower standard errors which means t-statistics are high, p-values are low, and the 'probability of significance' will be high for the variables relative to those in the *robust* standard errors model. Thus, some variables may be unnecessarily significant in the *classical* model only because the standard errors are not robust (or small). The standard errors for *log of Health expenditure per capita* is 1.541 and 2.188 for the *classical* and *robust* model, respectively. The standard errors for *log of GDP per capita* is 1.657 and 2.392 for the *classical* and *robust* model, respectively.

```
#Lets check if our assumption of homoskedasticity is satisfied.

mydata <- mydata %>% mutate(life_exp=exp(LEBF20052))
ggplot(mydata,aes(x=HXPC2005,y=life_exp)) + geom_point() + theme_minimal()
Warning: Removed 1 rows containing missing values (geom_point).
```



```
#repeating Ques eight regression
lm_comp_log<-lm(LEBF20052 ~ log(HXPC2005) + log(GDPPCUS2005), data = mydata) #summary(lm
```

	naive
(Intercept)	32.448***
	(5.860)
log(HXPC2005)	0.889
	(1.541)
log(GDPPCUS2005)	4.114**
	(1.657)
Num.Obs.	174
R2	0.493
R2 Adj.	0.487
AIC	1241.5
BIC	1254.1
Log.Lik.	−616.726
F	83.179
RMSE	8.38

* p < 0.1, ** p < 0.05, *** p < 0.01

```
#We may not be able to use kable here because of the naive and robust

msummary(list("naive"=lm_comp_log),
         stars=c('*' = .1, '**' = .05, '***' = .01))
```

```
# Same model with robust standard errors
msummary(list("naive"=lm_comp_log,
              "robust"=lm_comp_log),
         vcov=c("classical","robust"),
         stars=c('*' = .1, '**' = .05, '***' = .01))
```

	naive	robust
(Intercept)	32.448***	32.448***
	(5.860)	(7.999)
log(HXPC2005)	0.889	0.889
	(1.541)	(2.188)
log(GDPPCUS2005)	4.114**	4.114*
	(1.657)	(2.392)
Num.Obs.	174	174
R2	0.493	0.493
R2 Adj.	0.487	0.487
AIC	1241.5	1241.5
BIC	1254.1	1254.1
Log.Lik.	−616.726	−616.726
F	83.179	116.900
RMSE	8.38	8.38
Std.Errors	Classical	Robust

* p < 0.1, ** p < 0.05, *** p < 0.01

```r
packages <- knitr::write_bib(file = 'packages.bib')
packages
```

$AICcmodavg @Manual{R-AICcmodavg, title = {AICcmodavg: Model Selection and Multimodel Inference Based on (Q)AIC(c)}, author = {Marc J. Mazerolle}, year = {2020}, note = {R package version 2.3-1}, url = {https://CRAN.R-project.org/package=AICcmodavg}, }

$base @Manual{R-base, title = {R: A Language and Environment for Statistical Computing}, author = {{R Core Team}}, organization = {R Foundation for Statistical Computing}, address = {Vienna, Austria}, year = {2022}, url = {https://www.R-project.org/}, }

$broom @Manual{R-broom, title = {broom: Convert Statistical Objects into Tidy Tibbles}, author = {David Robinson and Alex Hayes and Simon Couch}, year = {2022}, note = {R package version 1.0.1}, url = {https://CRAN.R-project.org/package=broom}, }

$causaldata @Manual{R-causaldata, title = {causaldata: Example Data Sets for Causal Inference Textbooks}, author = {Nick Huntington-Klein and Malcolm Barrett}, year = {2021}, note = {R package version 0.1.3}, url = {https://github.com/NickCH-K/causaldata}, }

$countrycode @Manual{R-countrycode, title = {countrycode: Convert Country Names and Country Codes}, author = {Vincent Arel-Bundock}, year = {2022}, note = {R package version 1.4.0}, url = {https://vincentarelbundock.github.io/countrycode/}, }

$dagitty @Manual{R-dagitty, title = {dagitty: Graphical Analysis of Structural Causal Models}, author = {Johannes Textor and Benito {van der Zander} and Ankur Ankan}, year = {2021}, note = {R package version 0.3-1}, url = {https://CRAN.R-project.org/package=dagitty}, }

$dplyr @Manual{R-dplyr, title = {dplyr: A Grammar of Data Manipulation}, author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller}, year = {2022}, note = {R package version 1.0.10}, url = {https://CRAN.R-project.org/package=dplyr}, }

$faux @Manual{R-faux, title = {faux: Simulation for Factorial Designs}, author = {Lisa DeBruine}, year = {2021}, note = {R package version 1.1.0}, url = {https://github.com/debruine/faux}, }

$forcats @Manual{R-forcats, title = {forcats: Tools for Working with Categorical Variables (Factors)}, author = {Hadley Wickham}, year = {2022}, note = {R package version 0.5.2}, url = {https://CRAN.R-project.org/package=forcats}, }

$ggdag @Manual{R-ggdag, title = {ggdag: Analyze and Create Elegant Directed Acyclic Graphs}, author = {Malcolm Barrett}, year = {2022}, note = {R package version 0.2.6}, url = {https://github.com/malcolmbarrett/ggdag}, }

$ggplot2 @Manual{R-ggplot2, title = {ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics}, author = {Hadley Wickham and Winston Chang and Lionel Henry and Thomas Lin Pedersen and Kohske Takahashi and Claus Wilke and Kara Woo and Hiroaki Yutani and Dewey Dunnington}, year = {2022}, note = {R package version 3.3.6}, url = {https://CRAN.R-project.org/package=ggplot2}, }

$here @Manual{R-here, title = {here: A Simpler Way to Find Your Files}, author = {Kirill Müller}, year = {2020}, note = {R package version 1.0.1}, url = {https://CRAN.R-project.org/package=here}, }

$kableExtra @Manual{R-kableExtra, title = {kableExtra: Construct Complex Table with kable and Pipe Syntax}, author = {Hao Zhu}, year = {2021}, note = {R package version 1.3.4}, url = {https://CRAN.R-project.org/package=kableExtra}, }

$modelsummary @Manual{R-modelsummary, title = {modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready}, author = {Vincent Arel-Bundock}, year = {2022}, note = {R package version 1.0.2}, url = {https://vincentarelbundock.github.io/modelsummary/}, }

$plotly @Manual{R-plotly, title = {plotly: Create Interactive Web Graphics via plotly.js}, author = {Carson Sievert and Chris Parmer and Toby Hocking and Scott Chamberlain and Karthik Ram and Marianne Corvellec and Pedro Despouy}, year = {2021}, note = {R package version 4.10.0}, url = {https://CRAN.R-project.org/package=plotly}, }

$purrr @Manual{R-purrr, title = {purrr: Functional Programming Tools}, author = {Lionel Henry and Hadley Wickham}, year = {2020}, note = {R package version 0.3.4}, url = {https://CRAN.R-project.org/package=purrr}, }

$pwr @Manual{R-pwr, title = {pwr: Basic Functions for Power Analysis}, author = {Stephane Champely}, year = {2020}, note = {R package version 1.3-0}, url = {https://github.com/heliosdrm/pwr}, }

$readr @Manual{R-readr, title = {readr: Read Rectangular Text Data}, author = {Hadley Wickham and Jim Hester and Jennifer Bryan}, year = {2022}, note = {R package version 2.1.2}, url = {https://CRAN.R-project.org/package=readr}, }

$readxl @Manual{R-readxl, title = {readxl: Read Excel Files}, author = {Hadley Wickham and Jennifer Bryan}, year = {2022}, note = {R package version 1.4.1}, url = {https://CRAN.R-project.org/package=readxl}, }

$stargazer @Manual{R-stargazer, title = {stargazer: Well-Formatted Regression and Summary Statistics Tables}, author = {Marek Hlavac}, year = {2022}, note = {R package version 5.2.3}, url = {https://CRAN.R-project.org/package=stargazer}, }

$stringr @Manual{R-stringr, title = {stringr: Simple, Consistent Wrappers for Common String Operations}, author = {Hadley Wickham}, year = {2022}, note = {R package version 1.4.1}, url = {https://CRAN.R-project.org/package=stringr}, }

$tibble @Manual{R-tibble, title = {tibble: Simple Data Frames}, author = {Kirill Müller and Hadley Wickham}, year = {2022}, note = {R package version 3.1.8}, url = {https://CRAN.R-project.org/package=tibble}, }

$tidyr @Manual{R-tidyr, title = {tidyr: Tidy Messy Data}, author = {Hadley Wickham and Maximilian Girlich}, year = {2022}, note = {R package version 1.2.1}, url = {https://CRAN.R-project.org/package=tidyr}, }

$tidyverse @Manual{R-tidyverse, title = {tidyverse: Easily Install and Load the Tidyverse}, author = {Hadley Wickham}, year = {2022}, note = {R package version 1.3.2}, url = {https://CRAN.R-project.org/package=tidyverse}, }

[[25]] @Article{countrycode2018, title = {countrycode: An R package to convert country names and country codes}, author = {Vincent Arel-Bundock and Nils Enevoldsen and CJ Yetman}, journal = {Journal of Open Source Software}, year = {2018}, volume = {3}, number = {28}, pages = {848}, url = {https://doi.org/10.21105/joss.00848}, }

[[26]] @Article{dagitty2016, title = {Robust causal inference using directed acyclic graphs: the R package 'dagitty'}, journal = {International Journal of Epidemiology}, author = {Johannes Textor and Benito {van der Zander} and Mark S Gilthorpe and Maciej Liśkiewicz and George TH Ellison}, volume = {45}, number = {6}, pages = {1887–1894}, year = {2016}, doi = {10.1093/ije/dyw341}, }

[[27]] @Book{ggplot22016, author = {Hadley Wickham}, title = {ggplot2: Elegant Graphics for Data Analysis}, publisher = {Springer-Verlag New York}, year = {2016}, isbn = {978-3-319-24277-4}, url = {https://ggplot2.tidyverse.org}, }

[[28]] @Article{modelsummary2022, title = {{modelsummary}: Data and Model Summaries in {R}}, author = {Vincent Arel-Bundock}, journal = {Journal of Statistical Software}, year = {2022}, volume = {103}, number = {1}, pages = {1–23}, doi = {10.18637/jss.v103.i01}, }

[[29]] @Book{plotly2020, author = {Carson Sievert}, title = {Interactive Web-Based Data Visualization with R, plotly, and shiny}, publisher = {Chapman and Hall/CRC}, year = {2020}, isbn = {9781138331457}, url = {https://plotly-r.com}, }

[[30]] @Article{tidyverse2019, title = {Welcome to the {tidyverse}}, author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and

Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani}, year = {2019}, journal = {Journal of Open Source Software}, volume = {4}, number = {43}, pages = {1686}, doi = {10.21105/joss.01686}, }