Assignment1_Solution.R

alexh

2022-06-22

```
########## Assignment 1: Solution
# Creator: Alex Hoagland, alcobe@bu.edu
# Created: 6/18/2022
# Last modified: 6/18/2022
#
# PURPOSE
#   Solutions for Assignment 1
#
# NOTES:
#   - uses the Tidyverse package and Dplyr
################################################################################


##### Packages #####
```

Display name info

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

```
##    user
## "alexh"
```

```
# install.packages('tidyverse') # if needed, install the package
library(tidyverse) # call the relevant library
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(faux) # Useful package for simulating data
```

```
##
## ************
## Welcome to faux. For support and examples visit:
## https://debruine.github.io/faux/
## - Get and set global package options with: faux_options()
## ************
```

```
##
## Attaching package: 'faux'
```

```
## The following object is masked from 'package:purrr':
##
##     %||%
```

```
library(modelsummary)
library(causaldata)
library(here)
```

```
## here() starts at C:/Users/alexh/Dropbox/Teaching/HAD5744/2022_Fall/Assignments for 2021/Assig
nment1
```

```
# Load the data
library(readxl)
here::i_am("Assignments for 2021/Assignment1/Assignment1_Solution.R")
```

```
## here() starts at C:/Users/alexh/Dropbox/Teaching/HAD5744/2022_Fall
```

```
Dataset_1 <- read_excel(here("Assignments for 2021/Assignment1/Dataset 1.xlsx"))
##########


##### 1.-2. DAG -- multiple answers acceptable, not shown here ####
print("There are multiple acceptable answers for (1) and (2). I will skip these here.")
```

```
## [1] "There are multiple acceptable answers for (1) and (2). I will skip these here."
```

```
#####


##### 3. Summary Statistics ####
Dataset_1$HXPC2005 <- as.numeric(Dataset_1$HXPC2005) # There is a problem with HXPC not reading
 as numeric. Need to fix.
```

```
## Warning: NAs introduced by coercion
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:modelsummary':
##
##     SD
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
sumtable <- data.frame(describe(Dataset_1[,c('LEBF20052','GDPPCUS2005','HXPC2005','TotFertRate20
05',
                                               'PctUrb2005','PopGr2005')],
                          fast=TRUE, na.rm=TRUE))
sumtable <- sumtable[,-c(1,7)] # Drop the vars and range columns
htmlTable::htmlTable(format(sumtable, digits = 3),
                     header=c("N","Mean","Standard Deviation", "Minimum", "Maximum", "Standard E
rror"),
                     rnames=c("LEBF", "GDPPC", "HXPC", "Total Fertility Rate", "% Urban", "Popul
ation Growth"),
                     caption="Summary statistics: Based on 2005 Data.")
```

Summary statistics: Based on 2005 Data.

	N	Mean	Standard Deviation	Minimum	Maximum	Standard Error
LEBF	175	69.86	11.89	41.14	9.39e+01	8.99e-01
GDPPC	175	9862.10	16195.00	107.87	1.05e+05	1.22e+03
HXPC	174	713.39	1329.51	6.76	6.56e+03	1.01e+02
Total Fertility Rate	175	3.04	1.57	1.08	7.27e+00	1.19e-01
% Urban	175	53.62	23.14	9.50	1.00e+02	1.75e+00
Population Growth	175	1.46	1.32	-1.59	1.05e+01	9.95e-02

```
print("Note that the same size is not great here. Otherwise, there appears to be good variation
 on all variables, units look good, etc.")
```

```
## [1] "Note that the same size is not great here. Otherwise, there appears to be good variation
on all variables, units look good, etc."
```

```
##################################


##### 4. Univariate Regression ####
m1 <- lm(LEBF20052 ~ HXPC2005, data=Dataset_1)
msummary(list(m1),
         stars=c('*' = .1, '**' = .05, '***' = .01),
         fmt=2,
         statistic = c("s.e. = {std.error} (p = {p.value})","conf.int"),
         conf_level=.95,
         coef_rename=c("(Intercept)" = "Intercept", "GDPPCUS2005" = "GDPPC"),
         gof_omit = 'AIC|BIC|RMSE')
```

	Model 1
Intercept	66.88***
	s.e. = 0.91 (p = 0.00)
	[65.08, 68.68]
HXPC2005	0.00***
	s.e. = 0.00 (p = 0.00)
	[0.00, 0.01]
Num.Obs.	174
R2	0.203
R2 Adj.	0.198
F	43.718

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

```
print("Regression notes: The effect of an increase in HXPC on LEBF is a precise 0---there is no
 estimated impact of GDPPC on LEBF.")
```

```
## [1] "Regression notes: The effect of an increase in HXPC on LEBF is a precise 0---there is no
estimated impact of GDPPC on LEBF."
```

```
##################################


##### 5. Multivariate Regression ####
m2 <- lm(LEBF20052 ~ GDPPCUS2005 + HXPC2005, data=Dataset_1)
msummary(list(m1,m2),
         stars=c('*' = .1, '**' = .05, '***' = .01),
         fmt=2,
         statistic = c("s.e. = {std.error} (p = {p.value})","conf.int"),
         conf_level=.95,
         coef_rename=c("(Intercept)" = "Intercept", "GDPPCUS2005" = "GDPPC", "HXPC2005"= "HXPC",
"TotFertRate2005" = "Total Fertility Rate"),
         gof_omit = 'AIC|BIC|RMSE')
```

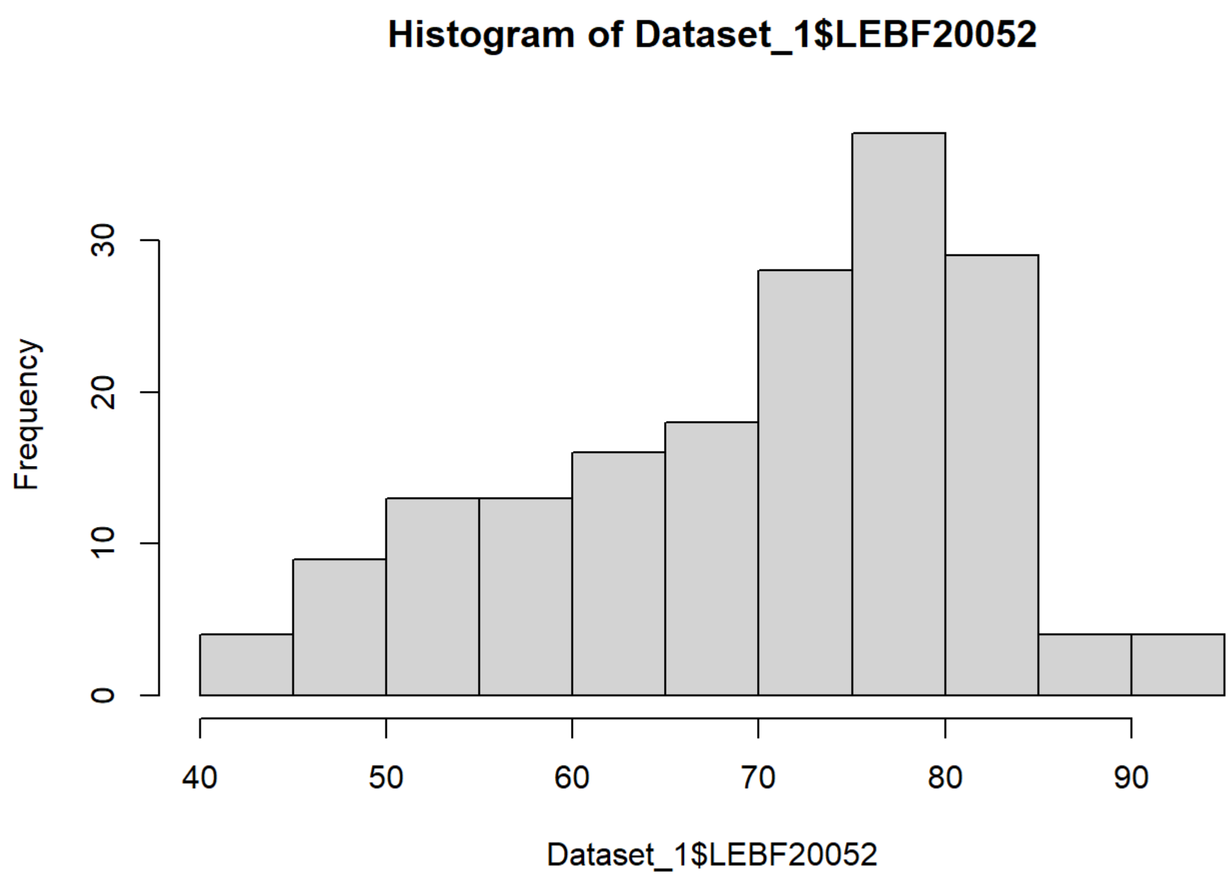
	Model 1	Model 2
Intercept	66.88***	65.83***
	s.e. = 0.91 (p = 0.00)	s.e. = 0.94 (p = 0.00)
	[65.08, 68.68]	[63.98, 67.68]
HXPC	0.00***	0.00
	s.e. = 0.00 (p = 0.00)	s.e. = 0.00 (p = 0.41)
	[0.00, 0.01]	[0.00, 0.00]
GDPPC		0.00***
		s.e. = 0.00 (p = 0.00)
		[0.00, 0.00]
Num.Obs.	174	174
R2	0.203	0.252
R2 Adj.	0.198	0.243
F	43.718	28.813

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

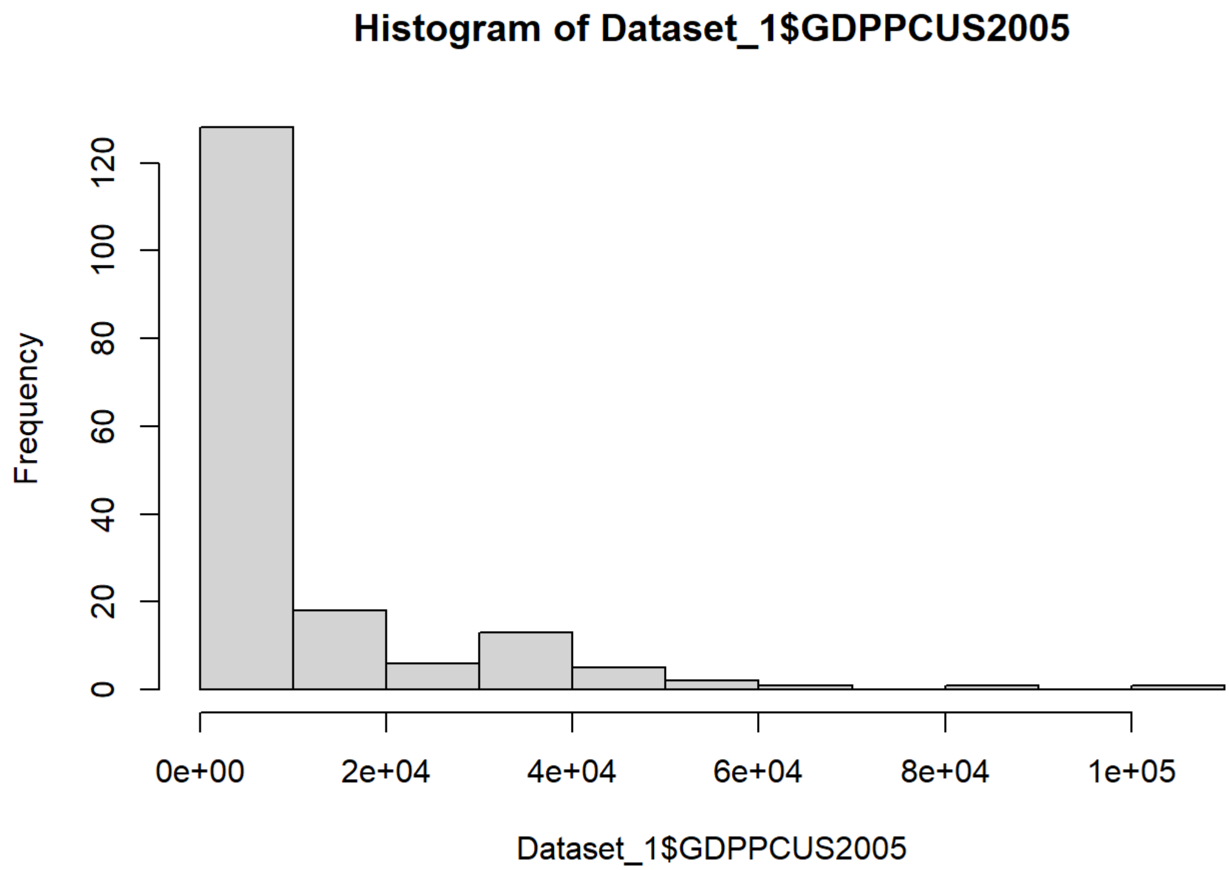
```
print("Controlling for GDP per capita eliminates the relationship between HXPC and LEBF, but the
re is still a measurement error here.")
```

```
## [1] "Controlling for GDP per capita eliminates the relationship between HXPC and LEBF, but th
ere is still a measurement error here."
```

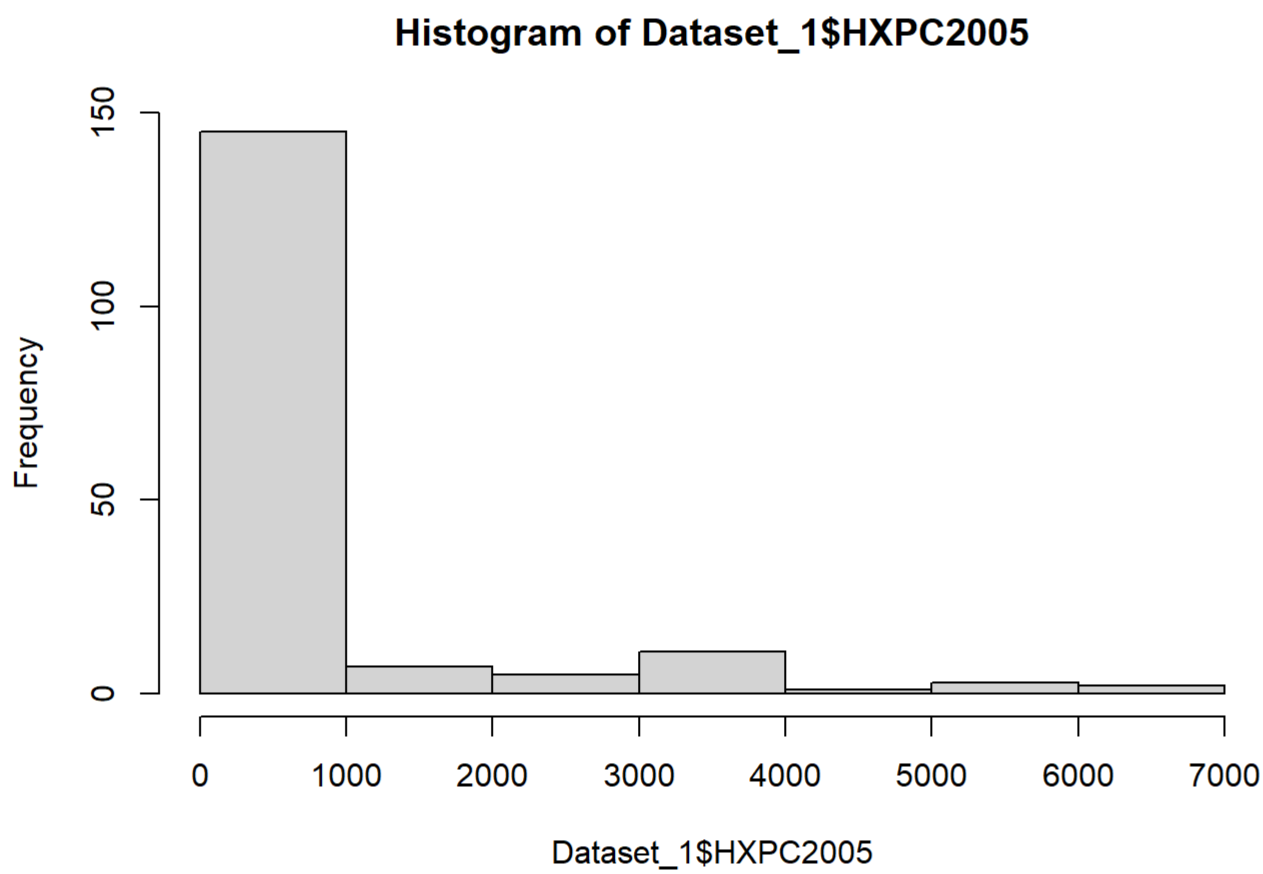
```
################################


##### 6. Transformations #####
hist(Dataset_1$LEBF20052) # Note that there is no skewness in LEBF, so no need for a transformat
ion there
```




```
hist(Dataset_1$GDPPCUS2005) # Lots of skewness here, recommend a log transform
```




```
hist(Dataset_1$HXPC2005) # Lots of skewness here, recommend a log transform
```




```
# Transform the data
Dataset_1 <- Dataset_1 %>% mutate(ln_gdppc = log(GDPPCUS2005),
                                  ln_hxpc = log(HXPC2005))


# New regression
m3 <- lm(LEBF20052 ~ ln_gdppc + ln_hxpc, data=Dataset_1)
msummary(list(m1,m2,m3),
         stars=c('*' = .1, '**' = .05, '***' = .01),
         fmt=2,
         statistic = c("s.e. = {std.error} (p = {p.value})","conf.int"),
         conf_level=.95,
         coef_rename=c("(Intercept)" = "Intercept", "ln_gdppc" = "ln(GDPPC)", "ln_hxpc"= "ln(HXP
C)",
                       "TotFertRate2005" = "Total Fertility Rate"),
         gof_omit = 'AIC|BIC|RMSE')
```

	Model 1	Model 2	Model 3
Intercept	66.88***	65.83***	32.45***
	s.e. = 0.91 (p = 0.00)	s.e. = 0.94 (p = 0.00)	s.e. = 5.86 (p = 0.00)

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

	Model 1	**Model 2**	**Model 3**
	[65.08, 68.68]	[63.98, 67.68]	[20.88, 44.01]
HXPC2005	0.00***	0.00	
	s.e. = 0.00 (p = 0.00)	s.e. = 0.00 (p = 0.41)	
	[0.00, 0.01]	[0.00, 0.00]	
GDPPCUS2005		0.00***	
		s.e. = 0.00 (p = 0.00)	
		[0.00, 0.00]	
ln(GDPPC)			4.11**
			s.e. = 1.66 (p = 0.01)
			[0.84, 7.39]
ln(HXPC)			0.89
			s.e. = 1.54 (p = 0.56)
			[−2.15, 3.93]
Num.Obs.	174	174	174
R2	0.203	0.252	0.493
R2 Adj.	0.198	0.243	0.487
F	43.718	28.813	83.179

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

```
print("After transforming the data, the results start to become more interpretable. Now, increas
ing GDP per capita by 10% is associated with almost a 1/2 year increase in life expectancy (0.4
1). However, there is no clear association between health expenditures and LEBF once we control
 for GDPPC.")
```

```
## [1] "After transforming the data, the results start to become more interpretable. Now, increa
sing GDP per capita by 10% is associated with almost a 1/2 year increase in life expectancy (0.4
1). However, there is no clear association between health expenditures and LEBF once we control
for GDPPC."
```

```
################################


##### 7. Geographic Dummies #####
# Read in crosswalk
crosswalk <- read_excel(here("Assignments for 2021/Assignment1/Country-Continent_Crosswalk.xlsx"
))

# Merge in info on continents and create dummies
Dataset_1 <- Dataset_1 %>% left_join(crosswalk, by=c("Country"))
Dataset_1 <- Dataset_1 %>%
  mutate(con_Africa = (Continent == "Africa"),
         con_Asia = (Continent == "Asia"),
         con_Europe = (Continent == "Europe"),
         con_Oceania = (Continent == "Oceania"),
         con_SA = (Continent == "South America"))

# New regression
m4 <- lm(LEBF20052 ~ ln_gdppc + ln_hxpc + con_Africa + con_Asia + con_Europe + con_Oceania + con
_SA, data=Dataset_1)
msummary(list(m1,m2,m3,m4),
         stars=c('*' = .1, '**' = .05, '***' = .01),
         fmt=2,
         statistic = c("s.e. = {std.error} (p = {p.value})","conf.int"),
         conf_level=.95,
         coef_rename=c("(Intercept)" = "Intercept", "ln_gdppc" = "ln(GDPPC)", "ln_hxpc"= "ln(HXP
C)",
                       "TotFertRate2005" = "Total Fertility Rate",
                       "con_AfricaTRUE" = "Africa", "con_AsiaTRUE" = "Asia", "con_EuropeTRUE" =
"Europe",
                       "con_OceaniaTRUE" = "Oceania", "con_SATRUE" = "South America"),
         gof_omit = 'AIC|BIC|RMSE')
```

	Model 1	Model 2	Model 3	Model 4
Intercept	66.88***	65.83***	32.45***	46.17***
	s.e. = 0.91 (p = 0.00)	s.e. = 0.94 (p = 0.00)	s.e. = 5.86 (p = 0.00)	s.e. = 5.52 (p = 0.00)
	[65.08, 68.68]	[63.98, 67.68]	[20.88, 44.01]	[35.27, 57.08]
HXPC2005	0.00***	0.00		
	s.e. = 0.00 (p = 0.00)	s.e. = 0.00 (p = 0.41)		
	[0.00, 0.01]	[0.00, 0.00]		
GDPPCUS2005		0.00***		
		s.e. = 0.00 (p = 0.00)		

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

	Model 1	Model 2	Model 3	Model 4
		[0.00, 0.00]		
ln(GDPPC)			4.11**	3.65**
			s.e. = 1.66 (p = 0.01)	s.e. = 1.49 (p = 0.02)
			[0.84, 7.39]	[0.71, 6.59]
ln(HXPC)			0.89	−0.34
			s.e. = 1.54 (p = 0.56)	s.e. = 1.45 (p = 0.81)
			[−2.15, 3.93]	[−3.21, 2.52]
Africa				−11.93***
				s.e. = 2.15 (p = 0.00)
				[−16.17, −7.68]
Asia				−0.74
				s.e. = 2.10 (p = 0.73)
				[−4.89, 3.42]
Europe				0.03
				s.e. = 2.09 (p = 0.99)
				[−4.10, 4.16]
Oceania				−1.42
				s.e. = 2.96 (p = 0.63)
				[−7.26, 4.41]
South America				1.23
				s.e. = 2.70 (p = 0.65)
				[−4.10, 6.55]
Num.Obs.	174	174	174	174
R2	0.203	0.252	0.493	0.641

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

	Model 1	Model 2	Model 3	Model 4
R2 Adj.	0.198	0.243	0.487	0.626
F	43.718	28.813	83.179	42.402

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

```
print("LEBF is significantly lower in African countries than in the rest of the world; no other
 significant differences are visible from this regression.")
```

```
## [1] "LEBF is significantly lower in African countries than in the rest of the world; no other
significant differences are visible from this regression."
```

```
###############################


##### 8. Interaction Terms ######
Dataset_1 <- Dataset_1 %>% mutate(interaction = ln_hxpc * con_Africa)

# New regression
m5 <- lm(LEBF20052 ~ ln_gdppc + ln_hxpc + con_Africa + con_Asia + con_Europe + con_Oceania + con
_SA + interaction, data=Dataset_1)
msummary(list(m1,m2,m3,m4,m5),
         stars=c('*' = .1, '**' = .05, '***' = .01),
         fmt=2,
         statistic = c("s.e. = {std.error} (p = {p.value})","conf.int"),
         conf_level=.95,
         coef_rename=c("(Intercept)" = "Intercept", "ln_gdppc" = "ln(GDPPC)", "ln_hxpc"= "ln(HXP
C)",
                       "TotFertRate2005" = "Total Fertility Rate",
                       "con_AfricaTRUE" = "Africa", "con_AsiaTRUE" = "Asia", "con_EuropeTRUE" =
"Europe",
                       "con_OceaniaTRUE" = "Oceania", "con_SATRUE" = "South America",
                       "PctUrban2005" = "% Urban", "inter_hxpc_urban" = "HXPC * % Urban"),
         gof_omit = 'AIC|BIC|RMSE')
```

	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	66.88***	65.83***	32.45***	46.17***	47.00***
	s.e. = 0.91 (p = 0.00)	s.e. = 0.94 (p = 0.00)	s.e. = 5.86 (p = 0.00)	s.e. = 5.52 (p = 0.00)	s.e. = 5.85 (p = 0.00)
	[65.08, 68.68]	[63.98, 67.68]	[20.88, 44.01]	[35.27, 57.08]	[35.46, 58.55]
HXPC2005	0.00***	0.00			

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

	Model 1	Model 2	Model 3	Model 4	Model 5
	s.e. = 0.00 (p = 0.00)	s.e. = 0.00 (p = 0.41)			
	[0.00, 0.01]	[0.00, 0.00]			
GDPPCUS2005		0.00***			
		s.e. = 0.00 (p = 0.00)			
		[0.00, 0.00]			
ln(GDPPC)			4.11**	3.65**	3.57**
			s.e. = 1.66 (p = 0.01)	s.e. = 1.49 (p = 0.02)	s.e. = 1.50 (p = 0.02)
			[0.84, 7.39]	[0.71, 6.59]	[0.60, 6.54]
ln(HXPC)			0.89	−0.34	−0.36
			s.e. = 1.54 (p = 0.56)	s.e. = 1.45 (p = 0.81)	s.e. = 1.45 (p = 0.80)
			[−2.15, 3.93]	[−3.21, 2.52]	[−3.24, 2.51]
Africa				−11.93***	−13.78***
				s.e. = 2.15 (p = 0.00)	s.e. = 4.72 (p = 0.00)
				[−16.17, −7.68]	[−23.10, −4.46]
Asia				−0.74	−0.81
				s.e. = 2.10 (p = 0.73)	s.e. = 2.12 (p = 0.70)
				[−4.89, 3.42]	[−4.99, 3.37]
Europe				0.03	0.12
				s.e. = 2.09 (p = 0.99)	s.e. = 2.11 (p = 0.95)
				[−4.10, 4.16]	[−4.04, 4.28]
Oceania				−1.42	−1.47

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

	Model 1	Model 2	Model 3	Model 4	Model 5
				s.e. = 2.96 (p = 0.63)	s.e. = 2.97 (p = 0.62)
				[−7.26, 4.41]	[−7.32, 4.39]
South America				1.23	1.18
				s.e. = 2.70 (p = 0.65)	s.e. = 2.71 (p = 0.66)
				[−4.10, 6.55]	[−4.17, 6.52]
interaction					0.45
					s.e. = 1.01 (p = 0.66)
					[−1.55, 2.45]
Num.Obs.	174	174	174	174	174
R2	0.203	0.252	0.493	0.641	0.642
R2 Adj.	0.198	0.243	0.487	0.626	0.624
F	43.718	28.813	83.179	42.402	36.946

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

```
print("Given the results in 7, it may be the case that health expenditures are high-return in areas with the lowest LEBF; hence, our interaction term looks at if increasing HXPC in African countries might improve LEBF. However, our regression still suggests no evidence that increasing health expenditures per capita is associated with lowering LEBF.")
```

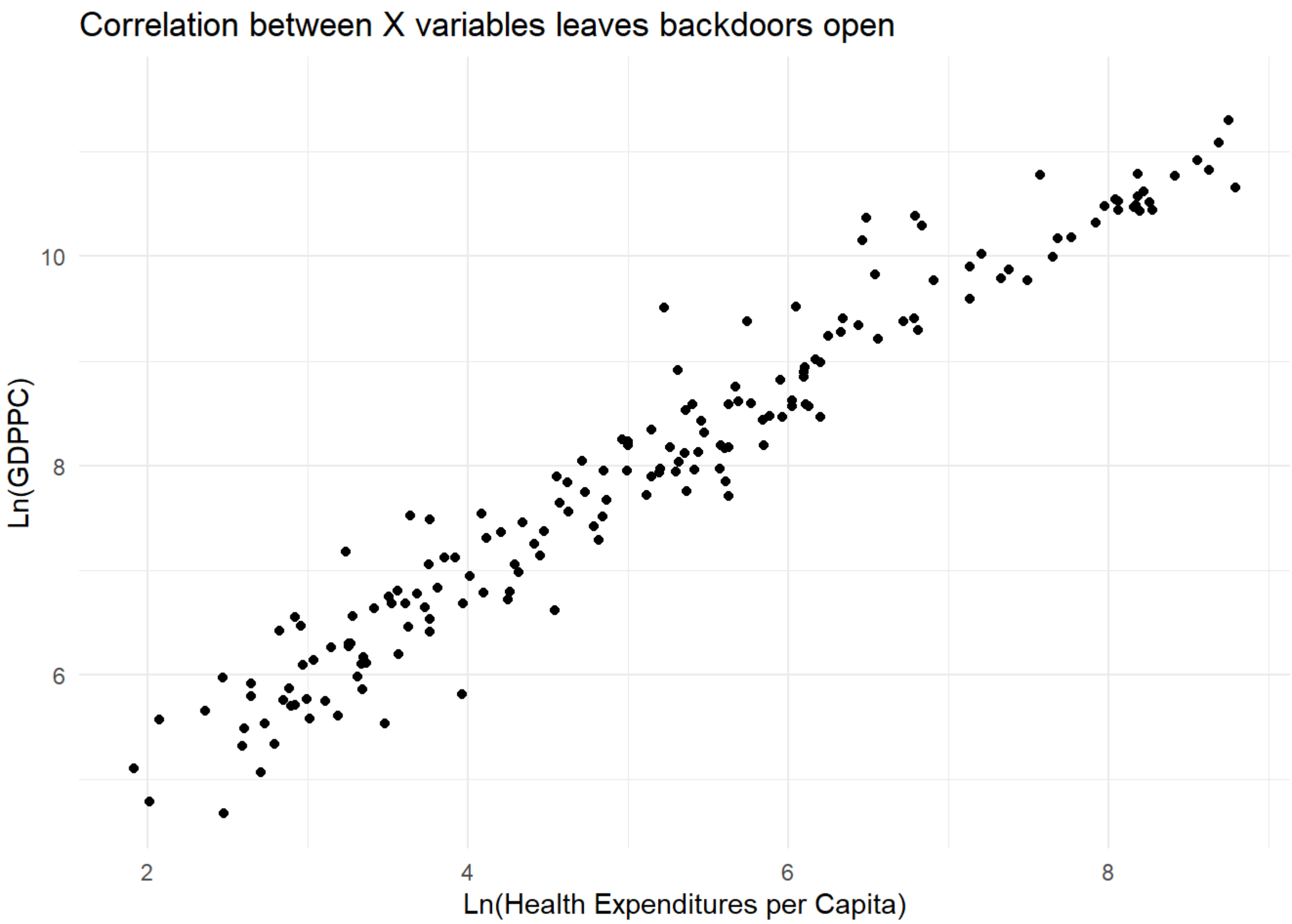
```
## [1] "Given the results in 7, it may be the case that health expenditures are high-return in areas with the lowest LEBF; hence, our interaction term looks at if increasing HXPC in African countries might improve LEBF. However, our regression still suggests no evidence that increasing health expenditures per capita is associated with lowering LEBF."
```

```
###############################


###### 9. Identification Problems ######
print("The main identification problem in this instance is that GDPPC and HXPC are so tightly correlated, there is not enough varaition in one without the other to correctly identify causal relationships.")
```

```
## [1] "The main identification problem in this instance is that GDPPC and HXPC are so tightly c
orrelated, there is not enough varaition in one without the other to correctly identify causal r
elationships."
```

```
ggplot(data=Dataset_1,aes(x=ln_hxpc)) + geom_point(aes(y=ln_gdppc)) +
  theme_minimal() + labs(x="Ln(Health Expenditures per Capita)",y="Ln(GDPPC)",title="Correlation
between X variables leaves backdoors open")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
###############################


##### 10. Standard Errors #####
library(miceadds)
```

```
## Loading required package: mice
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
## * miceadds 3.13-12 (2022-05-30 15:14:07)
```

```
# Compare robust standard errors and standard errors clustered by continent
msummary(list(m5,m5,m5),
         vcov=c("classical","robust",~Continent),
         stars=c('*' = .1, '**' = .05, '***' = .01),
         fmt=2,
         statistic = c("s.e. = {std.error} (p = {p.value})","conf.int"),
         conf_level=.95,
         coef_rename=c("(Intercept)" = "Intercept", "ln_gdppc" = "ln(GDPPC)", "ln_hxpc"= "ln(HXP
C)",
                       "TotFertRate2005" = "Total Fertility Rate",
                       "con_AfricaTRUE" = "Africa", "con_AsiaTRUE" = "Asia", "con_EuropeTRUE" =
"Europe",
                       "con_OceaniaTRUE" = "Oceania", "con_SATRUE" = "South America",
                       "PctUrban2005" = "% Urban", "inter_hxpc_urban" = "HXPC * % Urban"),
         gof_omit = 'AIC|BIC|RMSE')
```

	Model 1	Model 2	Model 3
Intercept	47.00***	47.00***	47.00***
	s.e. = 5.85 (p = 0.00)	s.e. = 8.18 (p = 0.00)	s.e. = 3.00 (p = 0.00)
	[35.46, 58.55]	[30.86, 63.14]	[41.09, 52.92]
ln(GDPPC)	3.57**	3.57	3.57***
	s.e. = 1.50 (p = 0.02)	s.e. = 2.30 (p = 0.12)	s.e. = 0.64 (p = 0.00)
	[0.60, 6.54]	[−0.98, 8.11]	[2.31, 4.82]
ln(HXPC)	−0.36	−0.36	−0.36
	s.e. = 1.45 (p = 0.80)	s.e. = 2.09 (p = 0.86)	s.e. = 0.70 (p = 0.61)
	[−3.24, 2.51]	[−4.49, 3.76]	[−1.75, 1.02]
Africa	−13.78***	−13.78**	−13.78***

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

	Model 1	Model 2	Model 3
	s.e. = 4.72 (p = 0.00)	s.e. = 5.37 (p = 0.01)	s.e. = 2.21 (p = 0.00)
	[−23.10, −4.46]	[−24.39, −3.17]	[−18.15, −9.41]
Asia	−0.81	−0.81	−0.81**
	s.e. = 2.12 (p = 0.70)	s.e. = 1.57 (p = 0.61)	s.e. = 0.39 (p = 0.04)
	[−4.99, 3.37]	[−3.92, 2.30]	[−1.57, −0.05]
Europe	0.12	0.12	0.12
	s.e. = 2.11 (p = 0.95)	s.e. = 1.29 (p = 0.92)	s.e. = 0.42 (p = 0.77)
	[−4.04, 4.28]	[−2.43, 2.68]	[−0.71, 0.96]
Oceania	−1.47	−1.47	−1.47***
	s.e. = 2.97 (p = 0.62)	s.e. = 1.92 (p = 0.45)	s.e. = 0.21 (p = 0.00)
	[−7.32, 4.39]	[−5.26, 2.33]	[−1.87, −1.06]
South America	1.18	1.18	1.18***
	s.e. = 2.71 (p = 0.66)	s.e. = 2.25 (p = 0.60)	s.e. = 0.18 (p = 0.00)
	[−4.17, 6.52]	[−3.27, 5.62]	[0.82, 1.53]
interaction	0.45	0.45	0.45
	s.e. = 1.01 (p = 0.66)	s.e. = 1.55 (p = 0.77)	s.e. = 0.39 (p = 0.25)
	[−1.55, 2.45]	[−2.61, 3.51]	[−0.32, 1.21]
Num.Obs.	174	174	174
R2	0.642	0.642	0.642
R2 Adj.	0.624	0.624	0.624
F	36.946	50.564	
Std.Errors	Classical	Robust	by: Continent

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

```
print("With robust standard errors, the impact of GDP on LEBF is no longer significant at the 9
0% confidence level. When clustering at the continent level, this result becomes more significan
t, and new continent relationships in LEBF emerge.")
```

```
## [1] "With robust standard errors, the impact of GDP on LEBF is no longer significant at the 9
0% confidence level. When clustering at the continent level, this result becomes more significan
t, and new continent relationships in LEBF emerge."
```

```
###############################
```