##-------------------------------------------------------------------------

# Preparation -------------------------------------------------------------

rm(list = ls()) # delete everything in the memory

library(ggplot2) # for graphics

library(Hmisc) # for rcorr

library(stargazer) # Regression table

library(effects) # for effect

# Read-in the world data

world.data <- read.csv("world.csv")

# Task 1 ------------------------------------------------------------------

# Examine the data

#---------------------------------------

# Describe Y

#---------------------------------------

# Numerical summary

summary(world.data $ women09)

# Graphical summary

g <- ggplot( world.data, aes( x = women09 ) ) + geom\_histogram()

g <- g + ylab("Number of countries") + xlab("Percentage women in parliament")

g

#---------------------------------------

# Describe X: per capita GDP

#---------------------------------------

# Numerical summary

summary(world.data $ gdp\_10\_thou)

# Graphical summary

g <- ggplot( world.data, aes( x = gdp\_10\_thou ) ) + geom\_histogram()

g <- g + ylab("Number of countries") + xlab("Per capita GDP (in 10,000 US dollars)")

g

#---------------------------------------

# Describe X-Y

#---------------------------------------

# To describe an X-Y relationship graphically, we draw what's called a

# scatterplot that shows the values of the X variable on the X-axis and

# the values of the Y variable on the Y-axis for all observations.

# To create a scatterplot, we use the geom\_point function (because

# each observation is denoted by a point.)

g <- ggplot(world.data, aes( x = gdp\_10\_thou, y = women09) ) + geom\_point()

g <- g + ylab("Percentage women in parliament") + xlab("Per capita GDP (in 10,000 US dollars)")

g

# Task 2 ------------------------------------------------------------------

# Omit observations with NAs

women.gdp <- world.data[is.na(world.data $ women09) == FALSE & is.na(world.data $ gdp\_10\_thou) == FALSE, ]

dim(women.gdp)

# Task 3 ------------------------------------------------------------------

# Calculate test-statistic & p-value

# The test statistic for a bivariate test of two numerical variables is

# (Pearson's) correlation coefficient.

# To calculate correlation coefficient (r), we first need to create a matrix object

# using the as.matrix function.

women.gdp.mat <- as.matrix(women.gdp[ c("gdp\_10\_thou","women09")])

# Then we use the rcorr function available in the Hmisc package.

rcorr(women.gdp.mat, type = "pearson")

# The correlation coefficient is 0.31 with the p-value < 0.001.

# The relationship is positive and highly statistically significant.

# Task 4 ------------------------------------------------------------------

# H0: the effect of gdp\_10\_thou on women09 is zero.

# Ha: the effect of gdp\_10\_thou on women09 is positive.

# Task 5 ------------------------------------------------------------------

# Regression

fit.wg <- lm(women09 ~ gdp\_10\_thou, data = women.gdp)

stargazer(fit.wg, type = "text")

# (a)

# Regression equation: women09 = 14.843 + 3.457 \* gdp\_10\_thou

# (b)

# Positive

# (c)

# An increase in per capita GDP (gdp\_10\_thou) by 10000 dollars will lead to an increase

# in female representation by 3.457 percentage points.

# An increase in per capita GDP (gdp\_10\_thou) by 1000 dollars will lead to an increase

# in female representation by 0.3457 percentage points.

# (d)

# The estimated coefficient for gdp\_10\_thou is highly statistically significant.

# We can reject the null hypothesis at 1% significance level.

# Task 6 ------------------------------------------------------------------

eff.wg <- effect(term = "gdp\_10\_thou", mod = fit.wg)

plot(eff.wg)

# Task 7 ------------------------------------------------------------------

stargazer(fit.wg, type = "text")

# (a)

# The R-squared is 0.093, which means that about 9% of the total variation in

# Female Representation is explained by per capita GDP.

# Many factors could potentially influence the variation in female representation

# across countries, and economic development appears to be one of them.

# R-squred of 9 % seems rather small, and it implies that either our model ignores

# some other important determinants, or the variation in female representation

# is quite random.

# (b)

# The Residual Std. Error is 10.379. That is, our predictions are off by about

# 10 percentage points, on average. I would say that this is rather big.

# Again, this implies that either our model ignores some other important

# determinants, or the variation in female representation is quite random.

# Task 8 ------------------------------------------------------------------

# Predicted values

women.gdp $ y.hat <- predict(fit.wg)

# Values for Rwanda

women.gdp[ women.gdp $ country == "Rwanda", c("women09", "y.hat") ]

# The predicted value of Female Representation for Rwanda is 14.91%, whereas

# the actual value is 56.3%. I would say that our prediction is very, very far off.

# Task 9 ------------------------------------------------------------------

# Do the analyses for PR and non-PR countries seperately to control for

# electoral system "manually".

# Re-label the PR variable

women.gdp $ pr <- factor(women.gdp $ pr\_sys, levels = c("No", "Yes"), labels = c("Non-PR", "PR"))

# Make sure it worked

table(women.gdp $ pr\_sys, women.gdp $ pr)

# Examine the X-Y relationship graphically

g <- ggplot(women.gdp, aes( x = gdp\_10\_thou, y = women09) ) + geom\_point()

g <- g + ylab("Percentage women in parliament") + xlab("Per capita GDP (in 10,000 US dollars)")

g <- g + facet\_grid( ~ pr)

g

# Create data sets for PR and non-PR

women.gdp.pr <- women.gdp[women.gdp $ pr == "PR", ]

women.gdp.mj <- women.gdp[women.gdp $ pr == "Non-PR", ]

# Separate regression models for PR and non-PR

fit.wg.pr <- lm(women09 ~ gdp\_10\_thou, data = women.gdp.pr)

fit.wg.mj <- lm(women09 ~ gdp\_10\_thou, data = women.gdp.mj)

stargazer(fit.wg.pr, type = "text")

stargazer(fit.wg.mj, type = "text")

# Alternatively, we can have one table that summarizes two (or more)

# regression models, as follows:

stargazer(fit.wg.pr, fit.wg.mj, type = "text")

# Task 10 -----------------------------------------------------------------

# The estimated effect of gdp\_10\_thou for PR countries.

# From the stargazer table above, we can see that the estimated effect of

# gdp\_10\_thou on women09 is positive (3.641) and statistically significant

# for PR countries (p < 0.01).

eff.wg.pr <- effect(term = "gdp\_10\_thou", mod = fit.wg.pr)

plot(eff.wg.pr)

# Task 11 -----------------------------------------------------------------

# The estimated effect of gdp\_10\_thou for non-PR countries.

# From the stargazer table above, we can see that the estimated effect of

# gdp\_10\_thou on women09 is positive (1.655) but not statistically significant.

# That is, we cannot reject the null hypothesis that the effect is zero.

eff.wg.mj <- effect(term = "gdp\_10\_thou", mod = fit.wg.mj)

plot(eff.wg.mj)

# Advanced: If you want to create a graph that shows the estimated effects

# of X on Y for two different scenarios (PR and non-PR), we can do the following.

# We will talk a lot more about this type of analyses in the Spring term.

fit.i <- lm(women09 ~ gdp\_10\_thou + pr + gdp\_10\_thou:pr, data = women.gdp)

eff.i <- effect(term = "gdp\_10\_thou:pr", mod = fit.i)

plot(eff.i)

# Task 12 -----------------------------------------------------------------

# Predicted values.

women.gdp.pr $ y.hat <- predict(fit.wg.pr)

# Values for Rwanda

women.gdp.pr[ women.gdp.pr $ country == "Rwanda", c("women09", "y.hat") ]

# The predicted value of Female Representation for Rwanda is now 19.45%

# (if we ignore electoral system, the prediction was 14.91%).

# This is still far off from the actual value (56.3%), but it is closer

# compared with the predicted value (14.91%) from the first model.

# end of file