QStep — Week 6 Lab: Homework Solutions

Load the necessary packages (if you haven't installed them in the lab, you'll need to install them as well). It's usually good practice to do this at the top of the worksheet.

```
library(tidyverse)
library(stargazer)
```

- create a vector `x` comprising six numbers: 4, 7, 9, 11, 11 and 20. Using the formula of the variance in this lab sheet, compute the variance of `x` 'by hand' (i.e. not using the `var()` function). The functions you need are `mean()`, `sum()`, `length()`, as well as some mathematical operators. Tip: you can make your life easier by breaking down the process into steps and storing intermediate results under some name. Verify your result with the `var()` function.

```
x <- c(4, 7, 9, 11, 11, 20)

# Solution 1 (nested functions)
var_x1 <- sum((x - mean(x))^2)/(length(x) - 1)
var_x1
```

```
## [1] 29.46667
```

```
# Solution 2 (breaking it down)
mean_x <- mean(x)
differences <- x - mean(x)
squared_differences <- differences^2
sum_of_squared_differences <- sum(squared_differences)
n_minus_1 <- length(x) - 1
var_x2 <- sum_of_squared_differences/n_minus_1
var_x2
```

```
## [1] 29.46667
```

```
# Verify it's the same
var(x)
```

```
## [1] 29.46667
```

```
var(x) == var_x1
```

```
## [1] TRUE
```

```
var(x) == var_x2
```

```
## [1] TRUE
```

- Go back to the `lijphart` dataset, regress `incarceration_2010` on `exec_parties_1981_2010` (i.e. use regression to explain the number of incarcerated people per 100,000 residents as a function of the index of consensus democracy).

```
lijphart <- read.csv("lijphart_data.csv")

lm1 <- lm(data = lijphart, incarceration_2010 ~ exec_parties_1981_2010)
summary(lm1)
```

```
##
## Call:
## lm(formula = incarceration_2010 ~ exec_parties_1981_2010, data = lijphart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -128.55  -62.76  -31.66   24.16  549.31
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              162.32      20.46   7.934 3.05e-09 ***
## exec_parties_1981_2010   -49.79      20.76  -2.398   0.0221 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.7 on 34 degrees of freedom
## Multiple R-squared:  0.1447, Adjusted R-squared:  0.1195
## F-statistic: 5.752 on 1 and 34 DF,  p-value: 0.0221
```

- How do you interpret the intercept and slope coefficients? What does this suggest for Lijphart's thesis about the effects of consensus democracy?

A country scoring 'zero' on the consensus democracy variable is predicted to have 162.3 incarcerated people per 100,000 citizen. An increase of one unit on the consensus democracy variable corresponds to a **decrease** of 49.8 incarcerated persons per 100,000. This is consistent with Lijphart's argument that consensus democracy makes for a 'kinder, gentler' democracy.

- Compute the β and α coefficients by hand using the formulas provided in section 4 of this lab sheet.

```
beta <- cov(lijphart$exec_parties_1981_2010, lijphart$incarceration_2010)/
  var(lijphart$exec_parties_1981_2010)

beta
```

```
## [1] -49.79416
```

```
alpha <- mean(lijphart$incarceration_2010) -
  beta*mean(lijphart$exec_parties_1981_2010)

alpha
```

```
## [1] 162.3196
```

- Then, plot the regression line on a scatter plot of the two variables.

```
ggplot(data = lijphart) +
  geom_point(mapping = aes(x = exec_parties_1981_2010,
                           y = incarceration_2010)) +
  geom_smooth(method = "lm",
              mapping = aes(x = exec_parties_1981_2010,
                            y = incarceration_2010), se = FALSE) +
  xlab("Executives-Party Index of Consensus Democracy") +
  ylab("Inmates per 100,000 citizens") +
  theme_minimal() #removes the ugly grey background
```
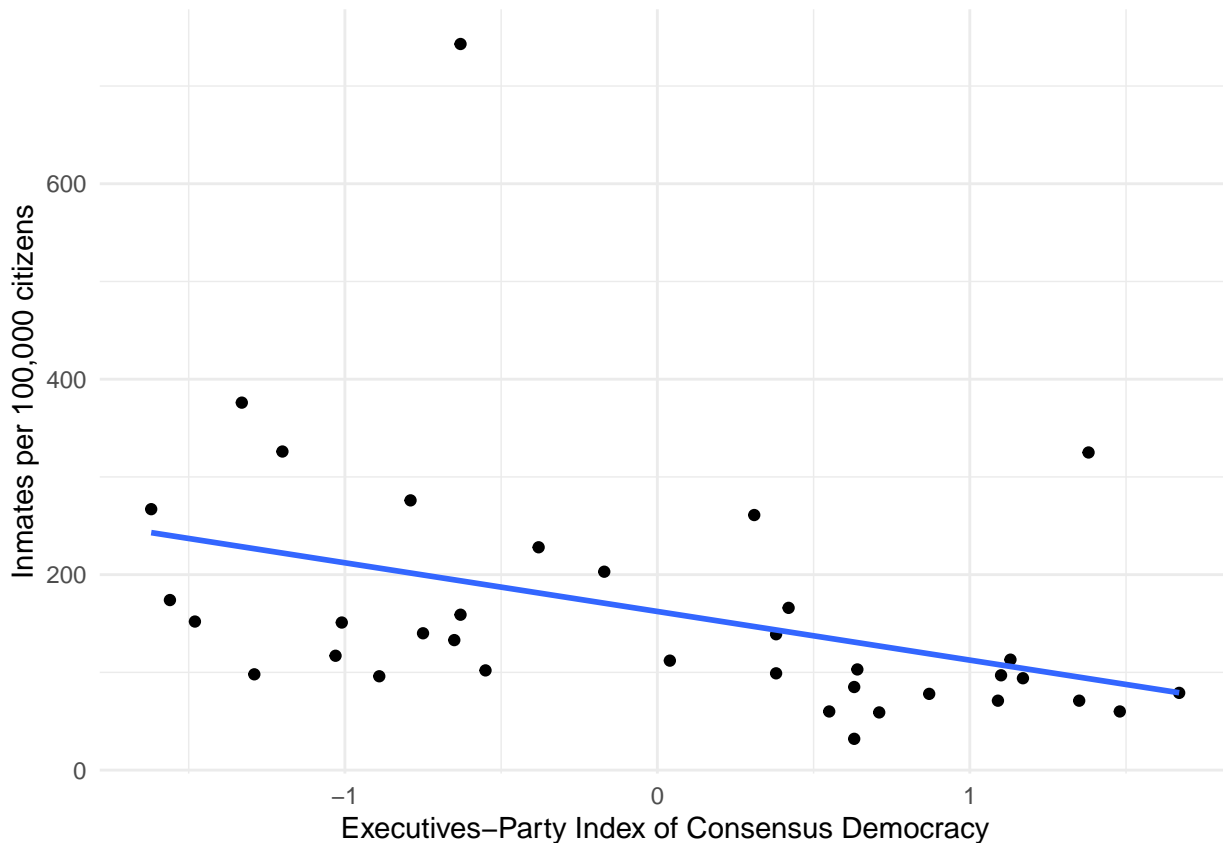

- Finally, store the regression output as a html file.

```
stargazer(lm1, out = "regression_output.html")
```

- Inspect the variable `europe` in the `lijphart` dataset, and consult the variable description in the lab sheet. What values does it take? What do they correspond to? Variables like these are known as 'dummy' or 'binary' variables.

```
table(lijphart$europe)
```

```
##
##  0  1
## 17 19
```

The variable takes the value of 1 for European countries and 0 for others.

- Now regress `womens_parl_representation_2010` on `europe`. How do we interpret substantively the slope coefficient? Tip: think about what it means to move from 0 to 1 on the independent variable, given what those values represent. What does the intercept mean? Tip: think about what it means for X to be 0, so that $Y = \alpha + \beta(0) = \alpha$.

```
lm2 <- lm(data = lijphart, womens_parl_representation_2010 ~ europe)
summary(lm2)
```

```
##
## Call:
## lm(formula = womens_parl_representation_2010 ~ europe, data = lijphart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.216  -7.429  -1.623   7.659  20.471
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.029      2.282   7.900 3.36e-09 ***
## europe        13.086      3.141   4.166 0.000201 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.41 on 34 degrees of freedom
## Multiple R-squared:  0.3379, Adjusted R-squared:  0.3184
## F-statistic: 17.35 on 1 and 34 DF,  p-value: 0.0002009
```

The slope can be interpreted as the predicted increase in percentage of women in parliament associated with moving from a non-European country (taking the value of 0) to a European country (taking the value of 1): in this case, the regression suggests that European countries have 13.1 percentage point higher female representation than non-European countries.

The intercept is the predicted value of the dependent variable for a non-European country. This is because the regression equation for a non-European country will be $Y = 18.029 + 0 \times (13.086) = 18.029$.

4

Extra: In fact, the predicted values are identical to the mean values of the subgroups: for a non-European country, the mean value of `womens_parl_representation_2010` is the intercept; for a European country, it's $Y = 18.029 + 1 \times (13.086) = 31.115$; that is, *the intercept plus the slope*.

```
mean(lijphart$womens_parl_representation_2010[lijphart$europe == 0])
```

```
## [1] 18.02941
```

```
# it's the same as the intercept of lm2
```

```
mean(lijphart$womens_parl_representation_2010[lijphart$europe == 1])
```

```
## [1] 31.11579
```

```
# it's the same as the intercept plus the slope
```

- Merge the `lijphart` dataset with the `ppi` dataset we used in week 4's lab (`ppi.csv` file), using `ccodealp` as the merging variable, and one of the functions of the `join()` family. Now compute the correlation coefficient of `parl_power_index` (the Parliamentary Power Index score computed by Fish and Kroeger, where higher = stronger parliament) and `exec_parties_1945_2010` (Lijphart's measure of 'consensus' democracy within institutions, where higher = more consensual). How do you interpret substantively the result? Does it make sense to you?

```
ppi <- read.csv("ppi.csv")

new_df <- full_join(lijphart, ppi, by = "ccodealp")

cor(new_df$parl_power_index,
    new_df$exec_parties_1945_2010, use = "complete.obs")
```

```
## [1] 0.5425377
```

A Pearson's r of 0.54 implies a positive, and fairly strong, correlation between parliamentary power and consensus democracy on the executives-party democracy. This makes sense, as a strong legislature relative to the executive is one of the features of consensus democracy.