CNQP3

2023-01-03

Load packages

```
library(tidyverse)
library(scales)
library(performance)
library(stargazer)
options(scipen = 999, digits = 2)
```

Read dataset

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

Question 1

1a.

```
sum(is.na(brazil$council.age))
```

```
[1] 99
```

- The author have no data on the age of the health council for 99 of the municipalities.

1b

```
ggplot(data = brazil, aes(x = 1, y = council.age)) +
  geom_boxplot(outlier.fill = "red", outlier.colour = "red") +
  labs(title = "Boxplot of the health council age",
       y = "Age",
       x = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        plot.title =element_text(size = 14) )
```

- The boxplot of the health council age above implies a median of 12 and 1st quartile and 3rd quartile values of 8 and 14 respectively.
- Two noticeable outliers were observed in the boxplot. The outliers are indicated with a red filled color.

1c.

```
summary(brazil$corruption)
```

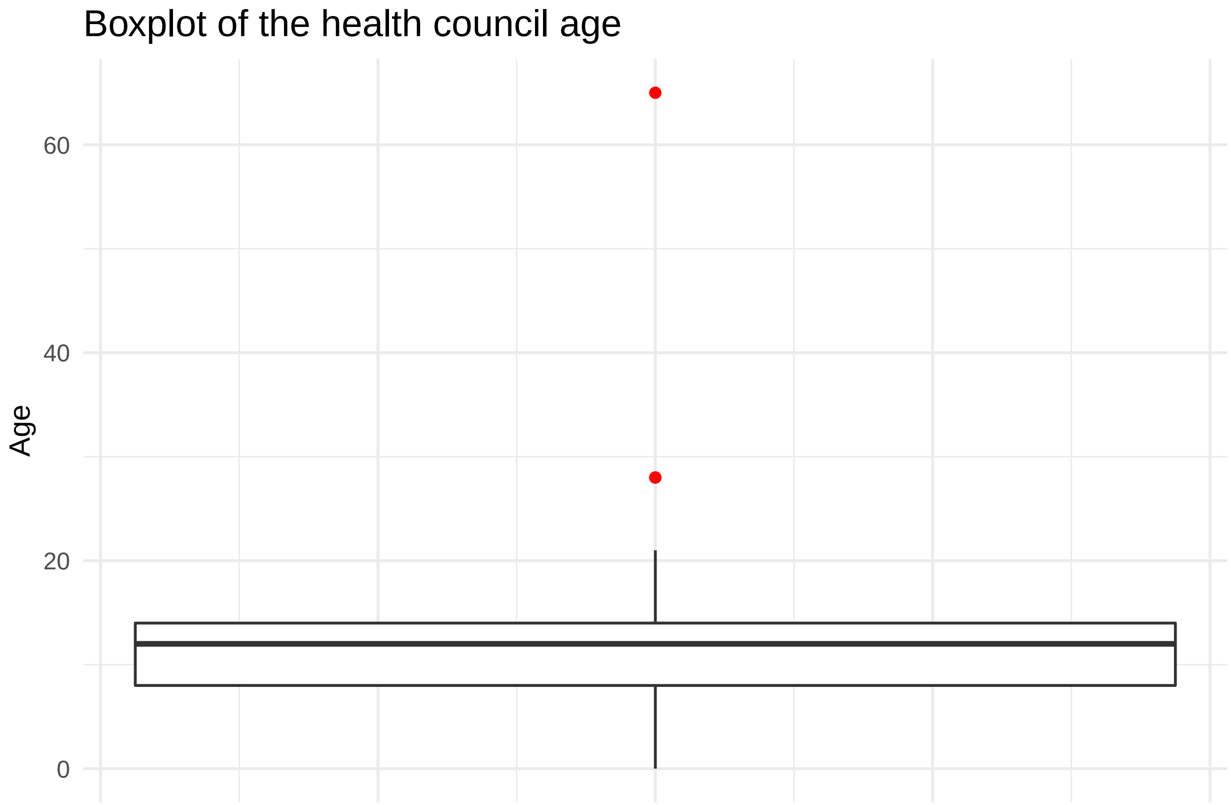


Figure 1: Boxplot of the health council age

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0       0      17      19      33     100 
```

- The mean and median of the `corruption` variable was observed to be 19 and 17 respectively.
- The mean is somewhat higher than the median as it would take into consideration all the values in the sample. As a result, the mean can be easily affected by an outlier. On the other hand, the median is a robust statistic as it's not easily affected by outliers.
- The estimate of the median implies that on the average, the corruption index of the municipalities that the author considered can be taken to be 17.

Question 2

2a.

```
model1 <- lm(corruption ~ council.age, data = brazil)
stargazer::stargazer(model1, type = "text")
```

```
===============================================
                        Dependent variable:    
                    ---------------------------
                            corruption         
-----------------------------------------------
council.age                  -0.340**          
                              (0.150)          
                                               
Constant                     23.000***         
                              (1.800)          
                                               
-----------------------------------------------
Observations                    881            
R2                             0.006           
Adjusted R2                    0.005           
Residual Std. Error      21.000 (df = 879)     
F Statistic            5.100** (df = 1; 879)   
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

2b.

- From the output of the simple linear regression above, it could be deduced that both the intercept and the `council.age` variates are significant at 5% level of significant.
- The estimate of the `council.age` implies that there would be on the average a 0.340 reduction in the corruption index as the age of the health council increases by a unit.
- On the other hand, the corruption index for a health council with age of zero is expected to be 23.0. This is the estimate of the intercept.
- The model is significant at 5% level of significant as the p-value of the F-statistic is less than 0.05.

2c.

We can interpret the regression coefficient as the average effect of council age on corruption under the following assumptions:

- Linear relationship: The relationship between `corruption` and `council.age` is expected to be linear.
- Independence: Observations are independent of each other
- Homoscedasticity: The variance of the residual is the same for any value of `council.age`.
- Normality: For any fixed value of corruption, council.age is normally distributed.

Question 3

3a.

```
model2 <- lm(corruption ~ council.age + margin + reelected + poverty, data = brazil)

stargazer(model1, model2, type = "text")
```

```
================================================================
                                Dependent variable:             
                    --------------------------------------------
                                     corruption                 
                             (1)                   (2)          
----------------------------------------------------------------
council.age               -0.340**              -0.290**        
                           (0.150)               (0.150)        
                                                                
margin                                            0.050         
                                                 (0.036)        
                                                                
reelected                                         -1.600        
                                                 (1.500)        
                                                                
poverty                                          0.150***       
                                                 (0.030)        
                                                                
Constant                  23.000***             15.000***       
                           (1.800)               (2.400)        
                                                                
----------------------------------------------------------------
Observations                 881                   877          
R2                          0.006                 0.037         
Adjusted R2                 0.005                 0.033         
Residual Std. Error   21.000 (df = 879)     20.000 (df = 872)   
F Statistic         5.100** (df = 1; 879) 8.500*** (df = 4; 872)
================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

3b.

- The estimated coefficient for margin in the model above is 0.050. This implies an increase in the corruption index for any elected major with a wide margin over the runner-up candidate in the previous election.
- Thus, the wider the margin between the elected major and the runner-up candidate, the higher the average corruption index.

3c.

- The model fit for the multiple linear regression seems to be a better fit than the simple linear regression as the adjusted R^2 for the multiple linear regression (0.033 0r 3.3%) is significantly higher than the simple linear regression (0.005 or 0.5%). Both models are significant at 5% level of significant.

3d.

```
test_df <- data.frame(council.age = c(10),
                      reelected = c(1),
                      margin = c(12),
                      poverty = c(50))

predict(model2, newdata = test_df)
```

```
 1
19
```

- The predicted corruption index score for a municipality health council that is 10 years old, that has a re-elected Major, where the Major won the last election by 12 percentage points, and where the poverty level is 50 is **19**.

Question 4

4a.

```
model3 <- lm(corruption ~ council.age + margin + reelected + poverty + council.age*reelected, data = bra

stargazer(model2, model3, type = "text")
```

```
===================================================================
                                   Dependent variable:
                      ---------------------------------------------
                                       corruption
                               (1)                    (2)
-------------------------------------------------------------------
council.age                 -0.290**                -0.120
                             (0.150)               (0.190)

margin                        0.050                  0.054
                             (0.036)               (0.036)

reelected                     -1.600                 3.200
                             (1.500)               (3.700)

poverty                     0.150***               0.150***
                             (0.030)               (0.030)

council.age:reelected                               -0.430
                                                   (0.300)

Constant                    15.000***              13.000***
                             (2.400)               (2.800)
```

```
-------------------------------------------------------------------
Observations                    877                     877
R2                             0.037                   0.040
Adjusted R2                    0.033                   0.034
Residual Std. Error      20.000 (df = 872)      20.000 (df = 871)
F Statistic            8.500*** (df = 4; 872) 7.200*** (df = 5; 871)
===================================================================
Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

4b.

- The estimated coefficient for margin in the model above is 0.054. This implies an increase in the corruption index for any elected major with a wide margin over the runner-up candidate in the previous election.
- Thus, the wider the margin between the elected major and the runner-up candidate, the higher the average corruption index.

4c.

```r
confint(model3, "poverty")
```

```
        2.5 % 97.5 %
poverty 0.095   0.21
```

- For the multiple model with interaction, the 95% confidence interval for the estimate of `poverty` is obtained to be (0.095, 0.21).

4d.

```r
cor.test(brazil$corruption, brazil$council.age)
```

```

    Pearson's product-moment correlation

data:  brazil$corruption and brazil$council.age
t = -2, df = 879, p-value = 0.02
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.14 -0.01
sample estimates:
   cor
-0.076
```

- From the Pearson product-moment correlation above, it could be deduced that there exist a low negative linear relationship between `corruption` and `council.age`.

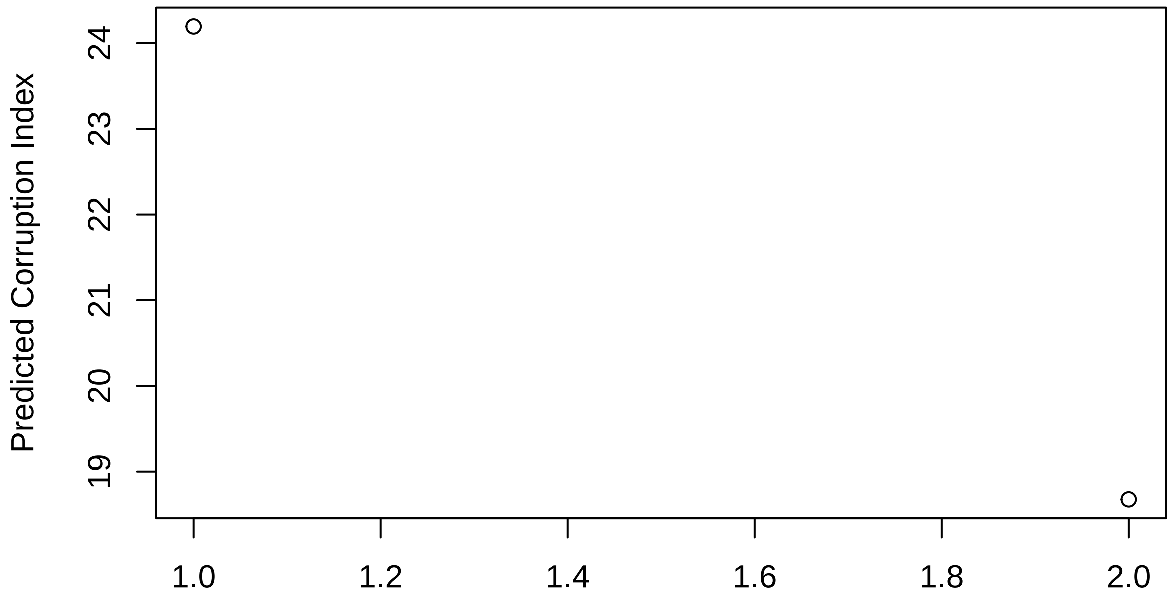
4e.

```r
test_df2 <- data.frame(council.age = c(0,20),
                       reelected = c(1,0),
                       margin = c(10, 10),
                       poverty = c(50, 50))
```

```
predict(model3, newdata = test_df2)
```

```
 1  2 
24 19 
```

```
plot(predict(model3, newdata = test_df2), ylab = "Predicted Corruption Index", xlab = "")
```



- The plot above also shows that as the `council.age` increases, the corruption index decreases. This is evident as the corruption index for `council.age == 0` is 24 and the corruption index for `council.age == 20` is 19.