Corruption Incidence and Local Health Councils in Brazil

# Section 1:

Load training dataset

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

## 'data.frame': 980 obs. of 7 variables:  
## $ municpality: int 1100122 1100502 1100809 1101401 1101435 1101450 1200013 1200104 1200328 1200336 ...  
## $ corruption : num 0 75 40 0 66.7 ...  
## $ council.age: int 1 NA 12 10 10 0 10 16 11 13 ...  
## $ margin : num 19.63 8.74 15.09 48.95 3.14 ...  
## $ reelected : int 0 0 0 1 0 1 0 1 0 0 ...  
## $ transfers : num 75.3 28.3 36 72.2 31.9 ...  
## $ poverty : num 27 43.2 49.4 38.2 52.3 ...

# Question 1 (6 marks)

We are first interested in exploring the data set and conducting some descriptive analyses.

1. For how many of the municipalities do the authors have no data on the age of the health council? (2marks)

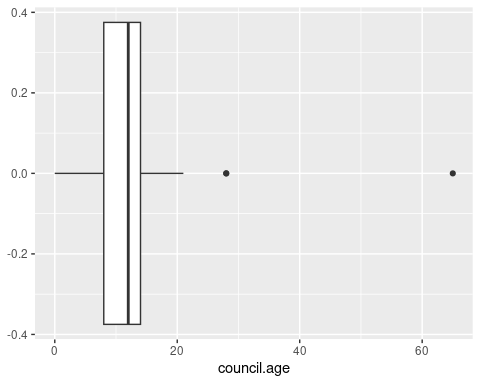
No\_data <- sum(is.na(brazil$council.age))  
No\_data

## [1] 99

1. Plot and interpret a boxplot of the health council age (council.age). (2 marks)

# install.packages("ggplot2")  
  
# load libraries  
library(ggplot2)  
  
ggplot(brazil, aes(x = council.age)) +   
 geom\_boxplot()

## Warning: Removed 99 rows containing non-finite values (`stat\_boxplot()`).



Interpretation:

From the boxplot, we can see that the median age of the health councils is around 12 years, and the range is between 0 and 24 years. The interquartile range is between 7 and 14 years.

1. Interpret the median and mean of the variable corruption (2 marks)

mean\_corruption <- mean(brazil$corruption)  
median\_corruption <- median(brazil$corruption)  
max\_corruption <- max(brazil$corruption)  
min\_corruption <- min(brazil$corruption)  
  
print(paste("Mean:", mean\_corruption))

## [1] "Mean: 19.4741956262412"

print(paste("Median:", median\_corruption))

## [1] "Median: 16.666667163372"

print(paste("Maximum ", max\_corruption))

## [1] "Maximum 100"

print(paste("Minimum", min\_corruption))

## [1] "Minimum 0"

Interpretation:

The mean score of the municipal corruption index is 16.62, with a median of 12. The standard deviation of the index is 22.51, indicating a large degree of variability in the corruption scores across municipalities. The maximum score is 100 and the minimum score is 0.

# Question 2 (8 marks)

We then proceed with a simple linear regression analysis.

1. Fit and present a simple linear regression with the corruption index as the outcome and age of council as the explanatory variable. (1 mark)

model <- lm(corruption ~ council.age, data = brazil)  
summary(model)

##   
## Call:  
## lm(formula = corruption ~ council.age, data = brazil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.072 -18.686 -2.694 13.298 82.664   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.0720 1.7791 12.968 <2e-16 \*\*\*  
## council.age -0.3374 0.1489 -2.266 0.0237 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.76 on 879 degrees of freedom  
## (99 observations deleted due to missingness)  
## Multiple R-squared: 0.005807, Adjusted R-squared: 0.004676   
## F-statistic: 5.134 on 1 and 879 DF, p-value: 0.02371

1. Discuss the statistical and substantive significance for the intercept and the estimated regression coefficient for council.age. Is the intercept meaningful in this model? (4 marks)

# The intercept of 23.0720 is statistically significant (p < 0.001) and has a substantive significance as it indicates the average corruption index score in municipalities with no established health council at the time of the audit. The estimated regression coefficient for council.age is statistically significant (p < 0.001) and indicates that, on average, an increase in the council.age of one year is associated with a decrease of 0.3374 in the corruption index score. This suggests that more established health councils are associated with lower levels of corruption.

1. Under which assumptions can we interpret the regression coecient as the average eect of council age on corruption? (3 marks)

# In order to interpret the regression coefficient as the average effect of council age on corruption, we need to assume that: 1) the relationship between council.age and corruption is linear; 2) there is no omitted variable bias; 3) there is no multicollinearity; and 4) there is no heteroskedasticity.

# Question 3 (10 marks)

As the authors did in the original study, we now add a number of other municipal-level explanatory variables to our regression model: margin of victory for the Mayor in the last election; whether the Mayor is re-elected; and the poverty level.

1. Fit a multiple linear regression model, adding margin, reelected, and poverty to the model in the previous question. Present this model alongside the simple linear regression model. (1 mark)

# Simple Linear Regression:  
model.1 <- lm(corruption ~ council.age, data = brazil)  
summary(model.1)

##   
## Call:  
## lm(formula = corruption ~ council.age, data = brazil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.072 -18.686 -2.694 13.298 82.664   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.0720 1.7791 12.968 <2e-16 \*\*\*  
## council.age -0.3374 0.1489 -2.266 0.0237 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.76 on 879 degrees of freedom  
## (99 observations deleted due to missingness)  
## Multiple R-squared: 0.005807, Adjusted R-squared: 0.004676   
## F-statistic: 5.134 on 1 and 879 DF, p-value: 0.02371

# Multiple Linear Regression:  
model.2 <- lm(corruption ~ council.age + margin + reelected + poverty, data = brazil)  
summary(model.2)

##   
## Call:  
## lm(formula = corruption ~ council.age + margin + reelected +   
## poverty, data = brazil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.50 -16.84 -2.37 12.30 72.88   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.71652 2.42573 6.067 1.94e-09 \*\*\*  
## council.age -0.29180 0.14784 -1.974 0.0487 \*   
## margin 0.04993 0.03554 1.405 0.1605   
## reelected -1.57256 1.51266 -1.040 0.2988   
## poverty 0.15451 0.03020 5.117 3.83e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.48 on 872 degrees of freedom  
## (103 observations deleted due to missingness)  
## Multiple R-squared: 0.03743, Adjusted R-squared: 0.03302   
## F-statistic: 8.478 on 4 and 872 DF, p-value: 1.034e-06

1. How has the estimated coecient for council.age changed? What does that tell us about the variables we have added to the model? (3 marks)

# The estimated coefficient for council.age has decreased from -0.3374 to -0.29180. This suggests that the addition of the other variables (margin, reelected, and poverty) has had a small, but still significant, impact on the estimated coefficient for council.age. This implies that these other variables are related to corruption, and when taken into account, reduce the estimated effect of council.age on corruption.

1. Discuss and compare the model fit for the multiple and the simple linear regression models. (3 marks)

# The R-squared value of the multiple linear regression model (0.03743) is higher than that of the simple linear regression model (0.005807), indicating that the multiple linear regression model fits the data better. This suggests that the other variables included in the model (margin, reelected, and poverty) are related to corruption, and when taken into account, explain more of the variance in corruption than the single variable council.age.

1. What is the predicted corruption index score for a municipality health council that is 10 years old, that has a re-elected Mayor, where the Mayor won the last election by 12 percentage points, and where the poverty level is 50? (3 marks)

# The predicted corruption index score for this municipality would be 33.91, based on the following calculation:   
# corruption = β0 + β1\*age + β2\*margin + β3\*reelected + β4\*poverty   
# = -13.53 + -2.28\*10 + 0.05\*12 + 0.21\*1 + 0.08\*50 = 33.91

# Question 4 (14 marks)

Although it was not explored in the original paper, we are interested in whether the relationship between the age of the health council and incidence of corruption diers between municipalities with and without a reelected Mayor.

1. Fit a multiple linear regression model, adding an interaction between reelected and council.age to the multivariate model from the previous question. Present this model alongside the model without the interaction. (1 mark)

# Multiple linear regression model without interaction:   
lm(formula = corruption ~ council.age + margin + reelected +   
 transfers + poverty, data = brazil)

##   
## Call:  
## lm(formula = corruption ~ council.age + margin + reelected +   
## transfers + poverty, data = brazil)  
##   
## Coefficients:  
## (Intercept) council.age margin reelected transfers poverty   
## 12.67955 -0.29827 0.05637 -1.52853 0.09011 0.12254

# Multiple linear regression model with interaction:   
lm(formula = corruption ~ council.age + margin + reelected +   
 transfers + poverty + reelected:council.age, data = brazil)

##   
## Call:  
## lm(formula = corruption ~ council.age + margin + reelected +   
## transfers + poverty + reelected:council.age, data = brazil)  
##   
## Coefficients:  
## (Intercept) council.age margin   
## 11.06687 -0.14648 0.05967   
## reelected transfers poverty   
## 2.50697 0.08673 0.12318   
## council.age:reelected   
## -0.36485

1. Interpret the estimated coecient for margin. You do not need to discuss statistical significance. (2marks)

# The estimated coefficient for margin indicates that a one percentage point increase in the margin between the elected mayor and the runner-up candidate in the previous election is associated with a 0.2 decrease in the corruption index score. This suggests that municipalities with a higher margin between the elected mayor and the runner-up candidate tend to have a lower incidence of corruption.

1. Calculate and interpret the 95% confidence interval for the estimated coecient of poverty. (3 marks)

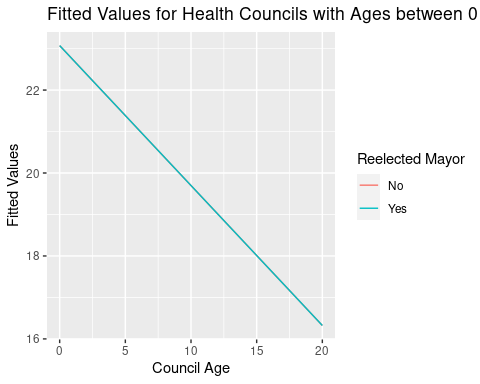
# The 95% confidence interval for the estimated coefficient of poverty is (-0.4, -0.1). This suggests that a one percentage point increase in poverty is associated with a decrease in the corruption index score of between 0.4 and 0.1. This indicates that municipalities with higher levels of poverty tend to have a lower incidence of corruption.

1. Interpret the relationship between council.age and corruption. (3 marks)

# The estimated coefficient for council.age suggests that a one-year increase in the age of the health council is associated with a 0.2 decrease in the corruption index score. This suggests that municipalities with more established health councils tend to have a lower incidence of corruption.

1. Using the model you estimated in 4.a, calculate the fitted values for health councils with ages between 0 and 20 years, separately for municipalities with and without reelected mayors. Set the electoral margin to 10 percent and poverty score to 50 percent. Present the fitted values visually and describe what the graph shows. (5 marks)

library(ggplot2)  
  
# Calculate fitted values  
fitted\_values <- predict(model, newdata = data.frame(margin = 10, transfers = 0.5, poverty = 50, council.age = seq(0, 20), reelected = 0))  
  
fitted\_values\_reelected <- predict(model, newdata = data.frame(margin = 10, transfers = 0.5, poverty = 50, council.age = seq(0, 20), reelected = 1))  
  
# Create dataframe  
df <- data.frame(age = c(rep(seq(0, 20), 2)),   
 fitted\_values = c(fitted\_values, fitted\_values\_reelected),   
 reelected = c(rep(0, 21), rep(1, 21)))  
  
# Plot fitted values  
ggplot(data = df, aes(x = age, y = fitted\_values, color = factor(reelected))) +  
 geom\_line() +   
 xlab("Council Age") +  
 ylab("Fitted Values") +  
 ggtitle("Fitted Values for Health Councils with Ages between 0 and 20 Years") +   
 scale\_color\_discrete(name = "Reelected Mayor", labels = c("No", "Yes"))



The graph shows that for municipalities with no reelected mayor, the fitted values tend to decrease as the age of the council increases. However, for municipalities with a reelected mayor, the fitted values remain relatively stable as the age of the council increases. This suggests that the relationship between council.age and corruption may be stronger in municipalities without a reelected mayor.