informative title

Clara le Draoulec, Maeva Kameni

March 22, 2022

Introduction:

quick background presentation + scientific question define all the variable:

CITATION: The first variable considered is the degree of unionization of the manufacturing labor force. Second, a dummy variable is used for the classifications right-to-work states and union- shop states. The third variable is a structural factor indicating the importance of relatively non-unionized industries in the economy of a given state. The fourth variable is the size of the agricultural labor force relative to the to- tal labor force.

Explanation of the variables in the text

```
## [1] "integer"
```

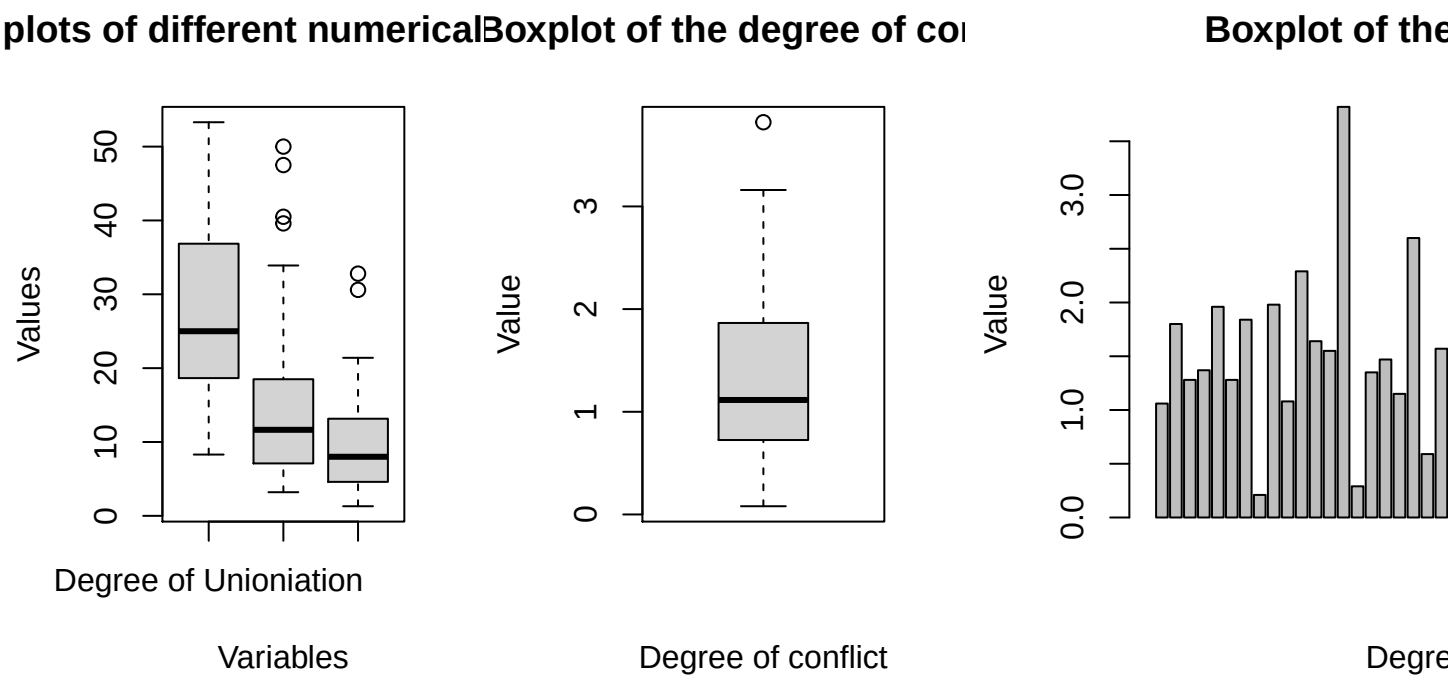
Exploratory Data Analysis:

data frame with 0 columns and 1 row

Table 1: Summary of the variable factors

State postal abbreviation	Degree of conflict	Degree of unionization	Union shop dummy	Sector	Ag force
Length:48	Min. :0.08	Min. : 8.3	Min. :0.00	Min. : 3.2	Min. : 1.3
Class :character	1st Qu.:0.74	1st Qu.:18.8	1st Qu.:0.00	1st Qu.: 7.1	1st Qu.: 4.6
Mode :character	Median :1.11	Median :25.0	Median :1.00	Median :11.7	Median : 8.0
NA	Mean :1.32	Mean :27.3	Mean :0.56	Mean :15.1	Mean : 9.9
NA	3rd Qu.:1.85	3rd Qu.:36.3	3rd Qu.:1.00	3rd Qu.:18.0	3rd Qu.:13.1

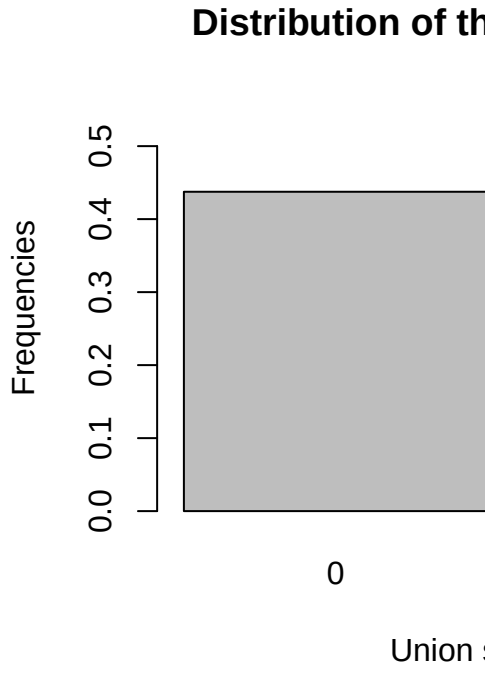
State postal abbreviation	Degree of conflict	Degree of unionization	Union shop dummy	Sector	Ag force
NA	Max. :3.82	Max. :53.3	Max. :1.00	Max. :50.0	Max. :32.8



Since there are some outliers, we will try to fit the model with or without them to access if these are influential points. (slide 68 Lecture 2)

Table 2: Summary of the variable factors

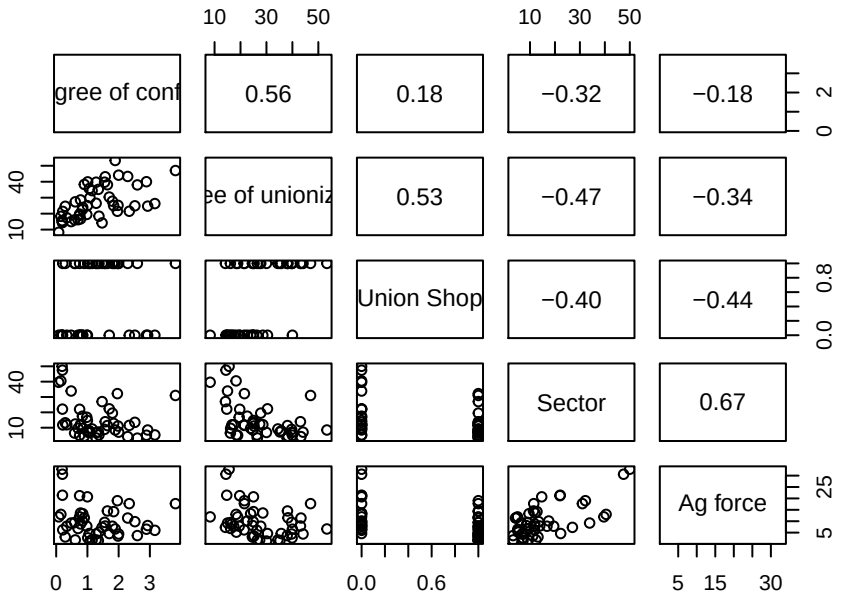
	State postal abbreviation	Degree of conflict	Degree of unionization	Union shop dummy	Sector
	Length:21	Min. :0.080	Min. : 8.30	Min. :0	Min. : 5
	Class :character	1st Qu.:0.340	1st Qu.:16.20	1st Qu.:0	1st Qu.:
	Mode :character	Median :0.780	Median :19.70	Median :0	Median
	NA	Mean :1.136	Mean :21.13	Mean :0	Mean :2
	NA	3rd Qu.:1.700	3rd Qu.:24.90	3rd Qu.:0	3rd Qu.:
	NA	Max. :3.160	Max. :40.00	Max. :0	Max. :5
	State postal abbreviation	Degree of conflict	Degree of unionization	Union shop dummy	Sector
	Length:27	Min. :0.210	Min. :14.2	Min. :1	Min. : 3
	Class :character	1st Qu.:1.035	1st Qu.:24.8	1st Qu.:1	1st Qu.:
	Mode :character	Median :1.370	Median :34.4	Median :1	Median
	NA	Mean :1.457	Mean :32.1	Mean :1	Mean :1
	NA	3rd Qu.:1.865	3rd Qu.:39.7	3rd Qu.:1	3rd Qu.:
	NA	Max. :3.820	Max. :53.3	Max. :1	Max. :3



For the dummy variable, a barplot is preferred to assess the distribution
Now we will analysis the distribution of the differents variable factors according to our groups namely Right-to-work states and Union-shop States (ca on maintien que c'est trop)

Let's see how the variables are linked to each other before beginning the model fitting. It could help us make sure that the model fits properly afterwards.

d with the numerical variables with the Pearson correla



Statisitics and model fitting:

First step:

- The first step in the analysis is a direct comparison of the observed levels of industrial conflict in the right-to-work stated and union-shop states for the period 1957-1962
- the measure of industrial conflict is the proportion of time lost due to work stoppages (man hours relative to total estimated working time)

(RESULT PAPER: AVERAGE PER YEAR: right-to-work : 0.19% // non-right to work:0.24% –> significative at 5%

AVERAGE TOTAL MAN HOURS LOST: right to work: 267,151 hours per year // union-shop : 712,603 hours per year –> significative –> causative factors: conceivably be favtor other than the existaence or non existence of a right to work law)

–> test the statements (better test t-test independant) :

- Agriculture > industry in right to work whereas industry > agriculture.

" Where the relatively non-unionized agricultural sector is of major importance in the economy of a state, the theoretical level of industrial conflict should be less than where this is not the case".

Second step:

- Try to explain the outcome variable using the variable factors (variance explained R^2 maybe adj-R^2) –> multiple linear regression (2 digits and ^ on estimates) : state how model fitted (ie, LS) CLEARLY describe how model selected define all terms

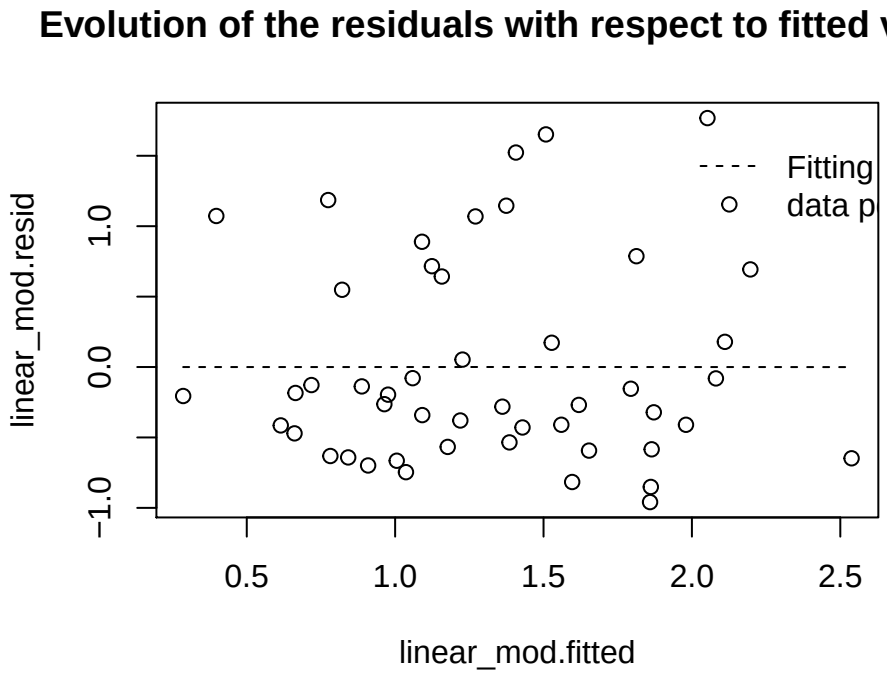
–> assumptions of the multiple linear regression: 1. A linear relationship between the dependent and independent variables 2. The independent variables are not highly correlated with each other 3. The variance of the residuals is constant 4. Independence of observation

–> state the assumptions: (ON THE ERROR) 1. errors have mean 0 2. errors are homoscedastic (same variance) 3. errors are uncorrelated 4. errors are normally distributed

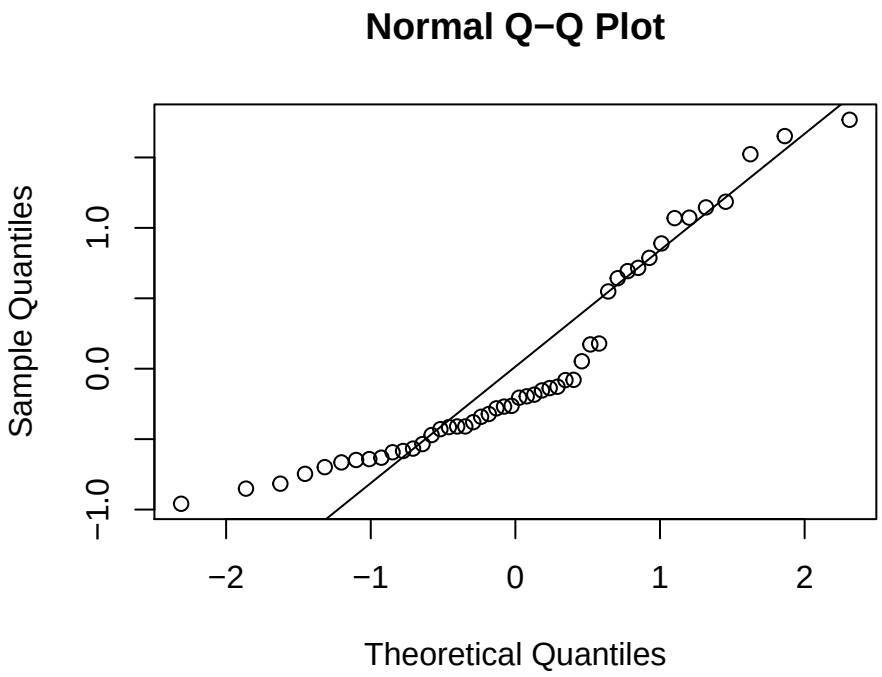
```
##
## Call:
## lm(formula = Degree_of_conflict ~ Degree_of_unionization + Union_shop_dummy +
##     Sector + Ag_force, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9590 -0.5432 -0.2347  0.5722  1.7676
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.216332   0.452069   0.479 0.634692
## Degree_of_unionization   0.049918   0.013104   3.809 0.000438 ***
## Union_shop_dummy        -0.287927   0.272070  -1.058 0.295837
## Sector                  -0.010426   0.013444  -0.776 0.442250
## Ag_force                 0.005709   0.021513   0.265 0.792003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7531 on 43 degrees of freedom
## Multiple R-squared:  0.3352, Adjusted R-squared:  0.2733
## F-statistic: 5.419 on 4 and 43 DF,  p-value: 0.001267
```

Model Assessment:

- State how the model fitted (Ordinary Least square) degree_of_conflict ~ remaining variable factors
- check the assumptions –> qqplot of the residuals and residuals vs. fitted



Commenter le plot



commenter le plot

Conclusion