36-402 Final Exam

Eu Jing Chua
eujingc

May 02, 2019

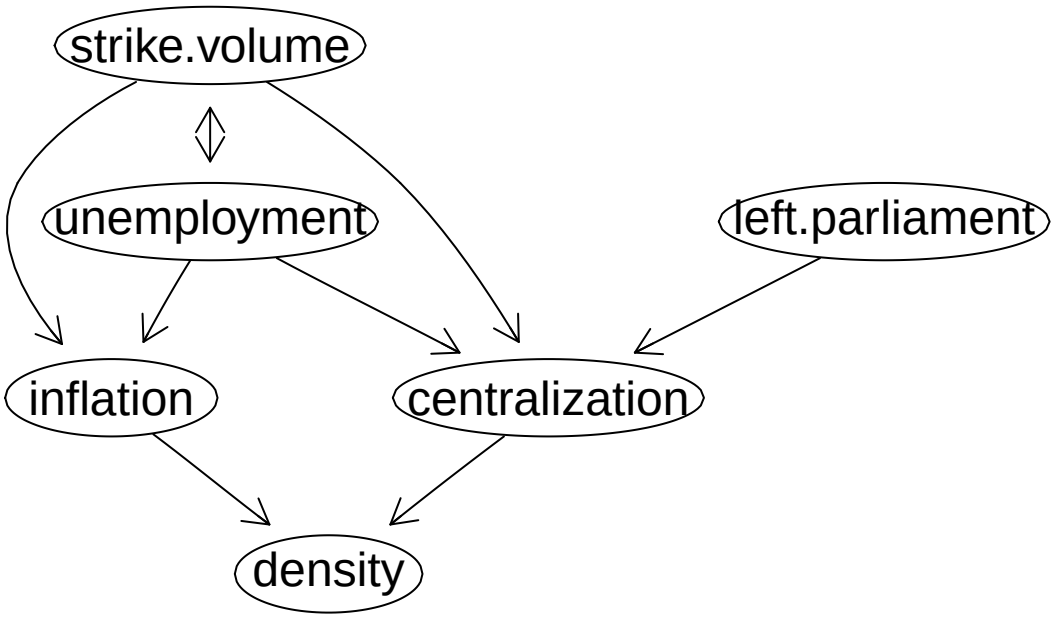
Question 1

The data set we are dealing with contains observations of factors deemed relevant to the frequency of strikes by organized workers. We are interested in knowing the causal effects of strikes, and thus use graphical modelling to find the various dependencies of each variable. We assume there are no hidden or latent variables in the data, and use the PC algorithm to infer the possible Markov equivalent Directed Acyclic Graphs (DAGs) that represent these dependencies. In the algorithm, we assume that all variables are normally distributed and linearly related to their parents, utilizing the `gaussCItest` at 0.95 significance level for independence testing.

Summary of Data

```
##  strike.volume     unemployment      inflation      left.parliament
##  Min.   :   0.0   Min.   : 0.000   Min.   :-2.900   Min.   : 8.16
##  1st Qu.:  19.0   1st Qu.: 1.200   1st Qu.: 2.700   1st Qu.:32.20
##  Median : 127.0   Median : 2.500   Median : 4.800   Median :42.50
##  Mean   : 288.7   Mean   : 3.555   Mean   : 5.957   Mean   :40.85
##  3rd Qu.: 360.0   3rd Qu.: 5.500   3rd Qu.: 8.200   3rd Qu.:49.70
##  Max.   :5918.0   Max.   :17.000   Max.   :27.500   Max.   :78.70
##
##  centralization     density
##  Min.   :0.000   Min.   :13.60
##  1st Qu.:0.250   1st Qu.:32.52
##  Median :0.375   Median :42.00
##  Mean   :0.456   Mean   :44.98
##  3rd Qu.:0.750   3rd Qu.:58.10
##  Max.   :1.000   Max.   :81.30
##                  NA's   :179
```

Looking at the above summary of the data, only for the variable `density` do we have missing data which we know is not missing at random but was due to the availability of the data only from 1960 onwards. As such, we chose not to omit all observations with missing `density`. However, since we are calculating the correlation matrix of our observations across multiple variables for `gaussCItest`, we can still utilize partial deletion to calculate as much of the correlation matrix as we can for pairwise complete observations.



Variables and their Relationships

Table 1: DAG 1 strike.volume → unemployment

	Parents	Children
strike.volume	None	inflation, centralization & unemployment
unemployment	strike.volume	inflation & centralization
inflation	strike.volume & unemployment	density
left.parliament	None	centralization
centralization	strike.volume, unemployment & left.parliament	density
density	inflation & centralization	None

Table 2: DAG 2 strike.volume ← unemployment

	Parents	Children
strike.volume	unemployment	inflation & centralization
unemployment	None	inflation, centralization & strike.volume
inflation	strike.volume & unemployment	density
left.parliament	None	centralization
centralization	strike.volume, unemployment & left.parliament	density
density	inflation & centralization	None

As we can see from the inferred DAG, there are two possible graphs as the edge between `strike.volume` and `unemployment` is undirected. Thus, we could possibly have a DAG where `strike.volume` → `unemployment`, or `strike.volume` ← `unemployment`. Every other relationship remains the same in both DAGs. Since both DAGs are possible given our observations, we take both DAGs into consideration in our analysis.

Question 2

Since we assumed linear models for all dependencies, we can estimate the coefficients of each linear model as well as the standard deviation of the regression noise under these models. Also, boostrapping with residual resampling is used to find 95% confidence intervals for each estimate of the coefficient and standard deviation:

Assuming the DAG where `strike.volume` ← `unemployment`,

Table 3: strike.volume against parents

	Coefficient	Lower	Upper
(Intercept)	169.50	112.10	228.90
unemployment	33.53	20.79	46.02

Assuming the DAG where `strike.volume` → `unemployment`,

Table 4: unemployment against parents

	Coefficient	Lower	Upper
(Intercept)	3.189000	2.9220000	3.447000
strike.volume	0.001269	0.0007702	0.001684

These other estimates hold for both graphs,

Table 5: inflation against parents

	Coefficient	Lower	Upper
(Intercept)	4.720000	4.1640000	5.24600
strike.volume	0.001692	0.0008679	0.00232
unemployment	0.210600	0.0915800	0.31710

Table 6: centralization against parents

	Coefficient	Lower	Upper
(Intercept)	0.6675000	0.5764000	0.7511000
strike.volume	-0.0001265	-0.0001693	-0.0000833
unemployment	-0.0187300	-0.0260800	-0.0115500
left.parliament	-0.0026550	-0.0044060	-0.0007794

Table 7: density against parents

	Coefficient	Lower	Upper
(Intercept)	21.320	19.4000	23.670
inflation	1.129	0.9187	1.329
centralization	35.140	32.1500	37.930

Table 8: Standard Deviations of regression noise for each endogeneous variable

	Noise SD	Lower	Upper
strike.volume	482.500	340.7000	618.0000
unemployment	2.968	2.7550	3.1900
inflation	4.479	4.1890	4.8110
centralization	0.296	0.2854	0.3081
density	10.510	9.8920	11.2700

Question 3

We are now interested in estimating the causal effect of `strike.volume` on density when we increase `strike.volume` by one standard deviation from its mean. Assuming the linear models hold, we want to eliminate all backdoor paths from `strike.volume` to `density` in order to remove confounding sources. Since there are two possible DAGs, we consider both cases:

Q3 a)

Assuming the DAG where `strike.volume` $\rightarrow$ `unemployment`, then there are no backdoor paths from `strike.volume` to`density`, hence we can use a linear model of `density` against `strike.volume` to estimate the effect:

In this case, we have a predicted expected change of -0.152 in the `density.`

Assuming the DAG where `strike.volume` $\leftarrow$ `unemployment`, then there are backdoor paths through unemployment from `strike.volume` to `density.` We can condition on `unemployment` to block all these backdoor paths, and hence construct a linear model of `density` against `strike.volume`, controlling for `unemployment.`

In this model, we have a predicted expected change of -0.237 in the `density`, controlling for `unemployment.`

Q3 b)

We can compare these estimated effects with the naive case where we linearly regress `density` against all other variables. In this case, we still estimate the effect on `density` when we increase `strike.volume` by one standard deviation from its mean, keeping all other variables at their mean.

The expected increase is 1.97 when `strike.volume` increases by one standard deviation and everything else remains at the mean. This is vastly different from the previous two estimates, which predicts a decrease in `density` as `strike.volume` increases, while this estimate predicts an increase in `density.`

Question 4

Given the possible causal inferences above, one important assumption made was that relationships were linear between endogeneous variables and their parents. One should verify the goodness-of-fit of our linear models before considering the implications of the causal inferences done above.

Q4 a) We could test each linearity assumption by doing a significance test for the difference in in-sample MSE between the linear regression and a non-parametric regression. In this case, we can use a kernel regression with bandwidths chosen with cross-validation. Such a non-parametric model will allow us to check for all kinds of mis-specifications in the linear model as it will converge to a significantly lower MSE. However, if the linear model was right, then it should also have a similar low MSE.

Thus we can test this by having H_0 : Linear model is right vs. H_a : Linear model is mis-specified, and then simulating data under the H_0. Specifically, we can add on the resampled residuals to the predictions of the linear model. We can then fit both linear models and kernel regressions to the data from the null hypothesis, then calculate the difference in MSE between the two models. This gives us a distribution for the difference in MSE under the null hypothesis.

Finally, we can take our observed difference in MSE from the two models fit to the actual data, and calculate a p-value for this observed difference under the above distribution.

Q4 b) Below are the plots of the simulated distributions of the MSE difference in models, assuming the null hypothesis, for each possible endogeneous variable against its parents. The red vertical line indicates our respective observed MSE differences. Table 9 summarizes these plots into their corresponding p-values for each test.

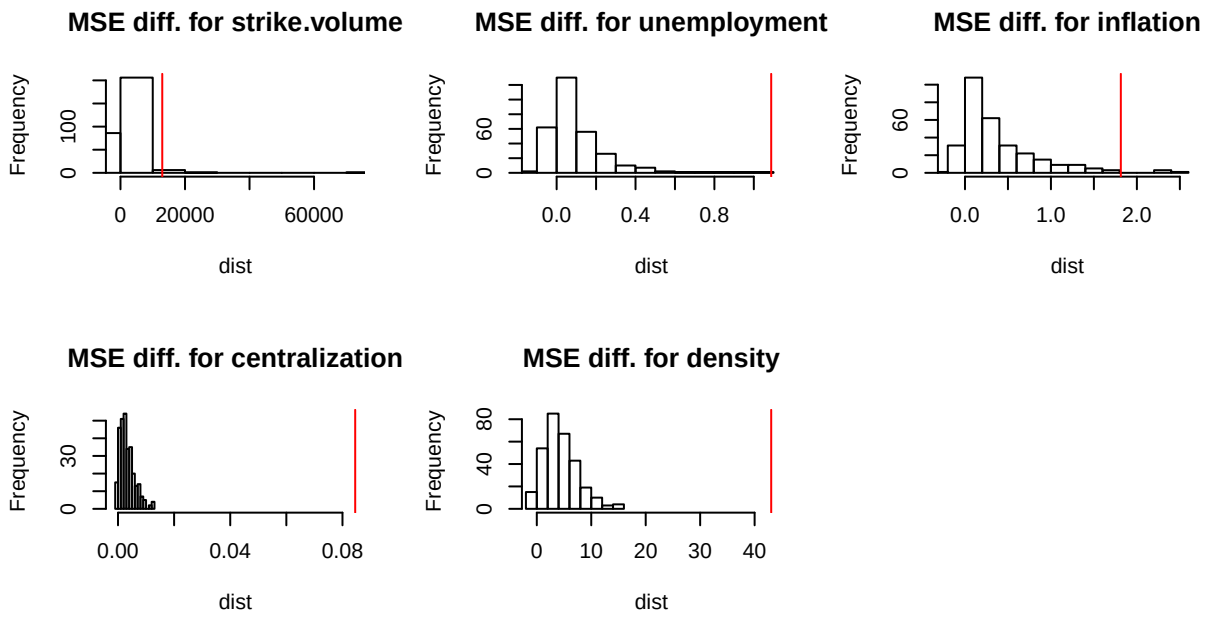


Table 9: Estimated p-values for goodness-of-fit of linear models

	x
strike.volume	0.0133
unemployment	0.0000
inflation	0.0133
centralization	0.0000
density	0.0000

Q4 c)

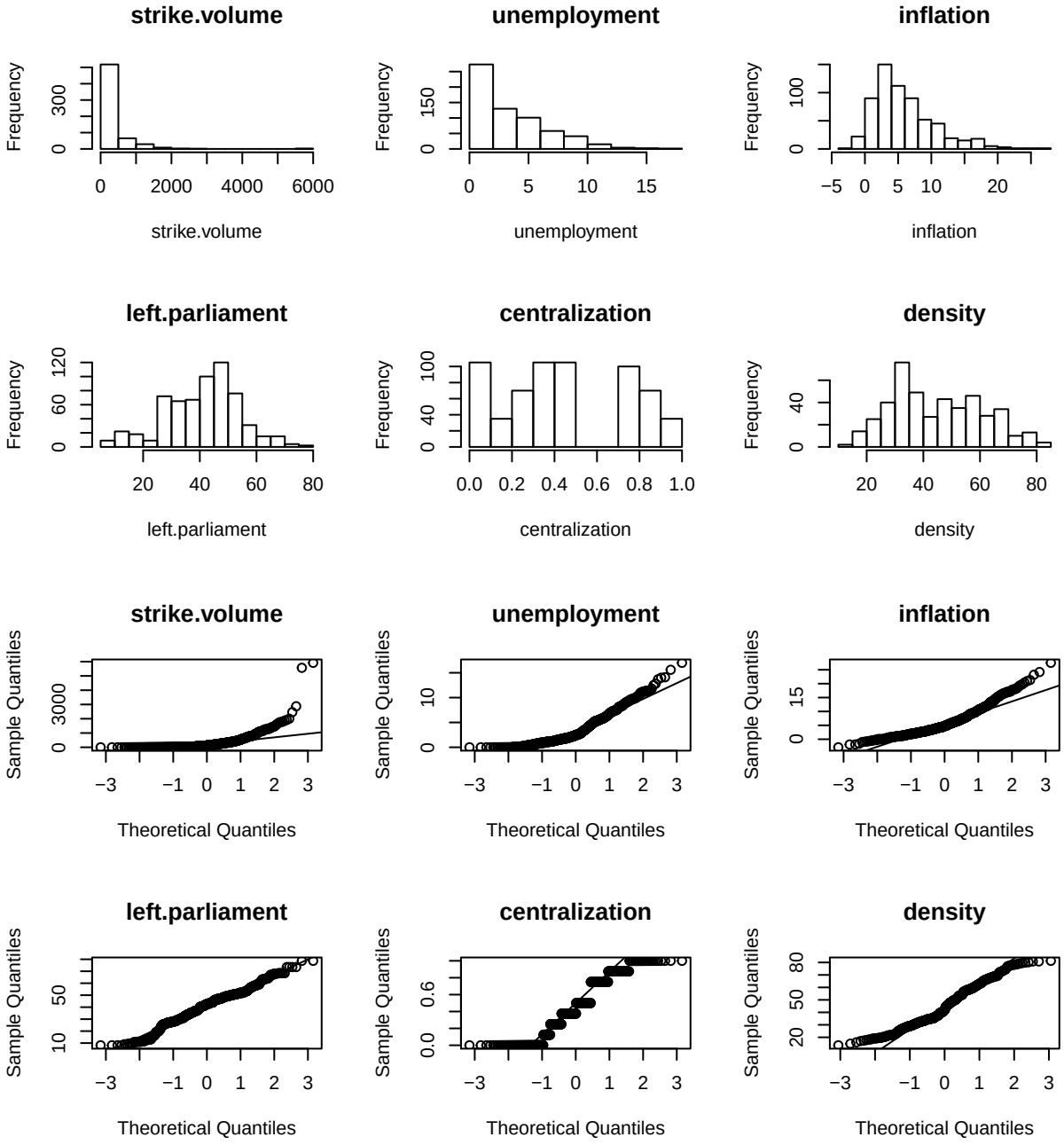
Assuming a 0.95 significance level test, we can see that for all possible parent-children relationships, the individual p-values are less than 0.05.

However if we wanted an overall conclusion for the linearity at 0.95 significance, then we can apply a conservative Bonferroni Correction for multiple testing, using $\alpha = \frac{0.05}{4} = 0.0125$ for each test since we have 4 tests for each DAG.

Even under this conservative correction, we see that the minimum p-value is approximately 0 (This p-value can either be from the relationship of `centralization` with its parents, or `density` and its parents. Regardless, both relationships are found in both possible DAGs, so the conclusion should hold for both DAGs). Since the minimum p-value for the test of non-linearity for all relationships is less than the corrected α in both DAGs, we conclude that there is sufficient evidence that we reject the overall null hypothesis at 0.95 significance, and that overall the linear model is mis-specified for endogeneous variables and their parents.

Question 5

We must evaluate some of the assumptions we made when inferring the DAGs based on the PC algorithm. Since the data was collected by a specialist in the field, lets assume that all relevant variables are observed such that there are no hidden or latent variables. Given so, we further assumed that all variables are Gaussian and are linearly related. First of all, the distribution of the observed data, followed by their normal Q-Q plots are as follows:



From the histograms and normal Q-Q plots, we can see that hardly any of the variables are roughly normally distributed with the exception of `left.parliament`. As for goodness-of-fit for our linear models, we can see from Q4 that in general this assumption is violated as there exist relationships, such as those of `density`

and `centralization` against their parents, that are significantly non-linear in either of the inferred DAGs. Although the Bonferroni Correction is conservative, we are conducting a relatively few number of multiple tests - four, so do we not lose too much power. Even with the adjusted α, we find that we still have sufficient evidence that not all of the relationships are linear.

Looking at the variables themselves, several of them have bounded ranges, such as `unemployment` which is a percentage that logically only lies within $[0, 100]$, or `centralization` which is a measure that as stated only lies within $[0, 1]$. Linear models of such variables against their parents will fail to respect the bounds of these variables. Secondly, it is not known as a fact that these variables are all linearly related to their parents; there is no reason to believe so other than to simplify the statistical modelling.

Hence, the base assumptions of our model do not hold: the variables are not all Gaussian, and are not all linearly related to their parents in the inferred DAG. The usage of `gaussCItest` as a conditional independence test is thus unsuitable for this dataset, so the inferred DAGs might not be accurate and on the whole is not reasonable.

Extra Credit

With the data collected, we make some heavy assumptions about the data collected, as well as the intrinsic relationships between the variables collected, in order to do some tractable statistical modelling. If these assumptions are to be believed, then we have two possible conclusions that the data agrees with:

- Strike volume is simply not affected by anything at all, and is just unpredictable.
- Strike volume is affected by unemployment, where as unemployment goes up, so does strike volume.

If the first conclusion is true, then there is nothing we can do to affect strike volume as it just happens on its own. This is entirely possible, as it could be that strikes just happen for such a wide variety of reasons that the observed variables just do not fully capture how one could effectively affect strike volume.

If the second conclusion is true, then lower unemployment rates are associated with lower strike volumes, and higher unemployment rates are associated with higher strike volumes. If this case, one can aim to indirectly lower strike volumes by lowering unemployment rate, which might make some sense as the population tends to be happier when they have jobs.

However these conclusions must be taken with a grain of salt as the assumptions made are very strong. In fact when checking the assumptions, there is sufficient evidence from the data to suggest that they in fact do not hold very well, and thus all the conclusions from analysis done with these assumptions must be treated with caution.

Moving forward, it is possible that not all relevant variables in the process of strikes have been collected. Thus, more studies can be done to investigate what other possible factors there are that affect the frequency of strikes. More observations of each variable would also help in developing more complex models that require more data to give accurate results.