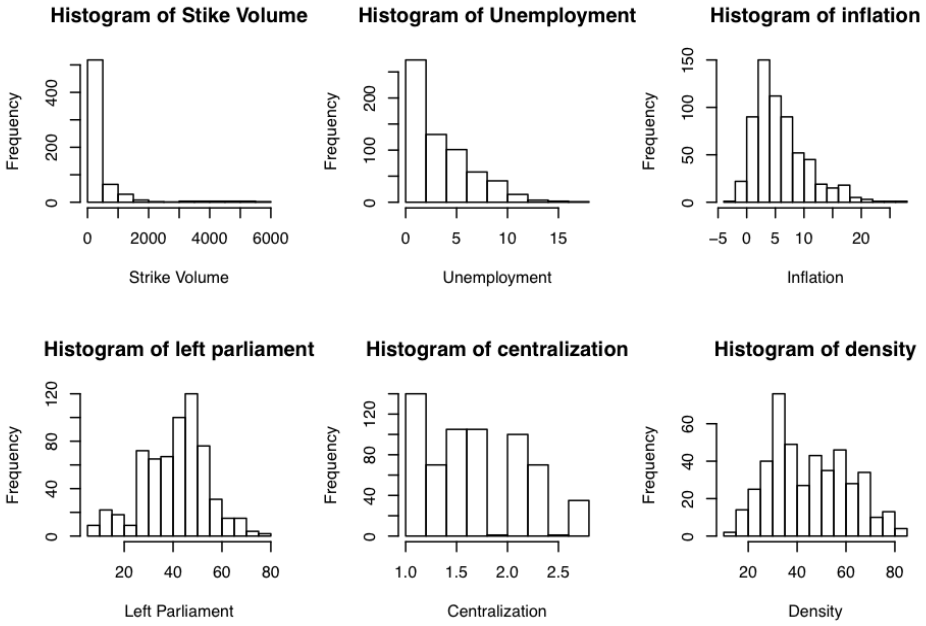
36402 Final Aaron Kruchten

Introduction

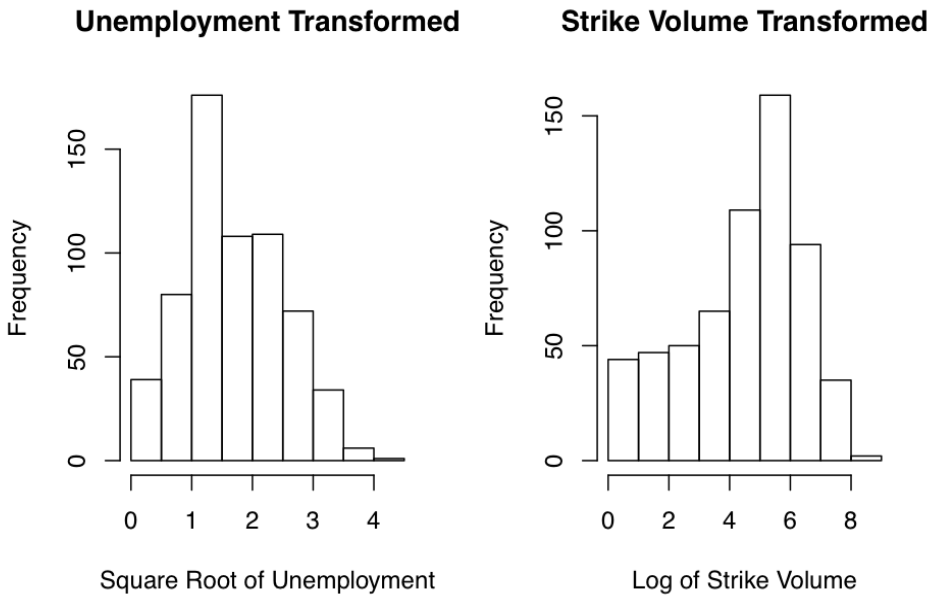
In this paper, we will attempt to find the factors which affect the frequency and severity of strikes by unionized and organized workers. We will attempt to do this by analyzing a data set which contains variables with information about strike volume, unemployment rate, inflation rate, parliamentary representation of social democratic and labor parties, centralization of leadership in a country's union, and union density. Our data set contains these variables for 18 different developed nations for the years between 1951-1985. Our first step in this analysis will be to form the directed acyclic graph for these variables to reach some insight about the causal effects between these variables.

Analysis

We will attempt to form the directed acyclic graph (DAG) by using the Gaussian independence conditional test and PC algorithm. One of the assumptions of the Gaussian independence conditional test is that all variables have a Gaussian distribution. We can check this assumption by inspecting the histograms of the variables we will put into our DAG. As one can see from the plots below, it seems that density, left parliament, and inflation are reasonably Gaussian with inflation having a small right skew. Centralization appears to be close to Gaussian however there appears to be no data points around 0.6, and many data points at a centralization of around 0.0. Applying few different transform to this data does not make it appear to be any more Gaussian so we will leave it as it is.



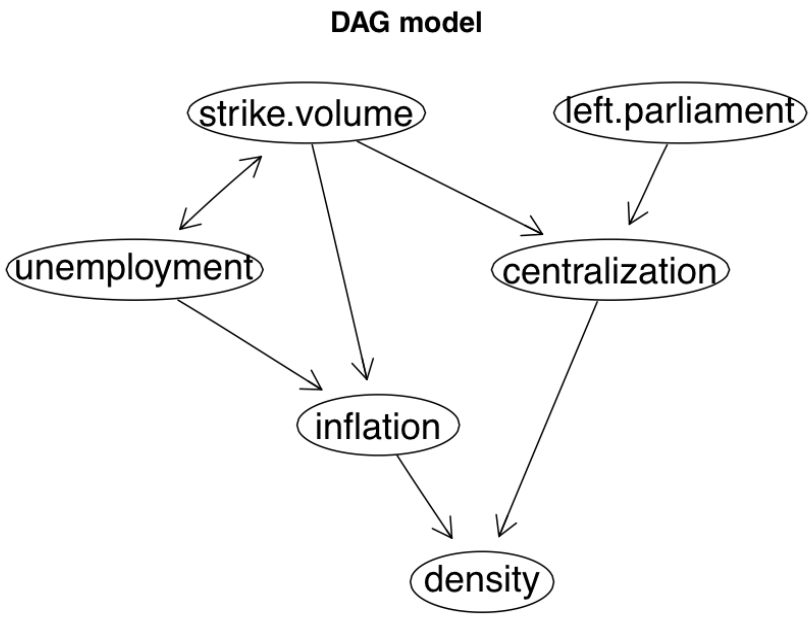
Below are the histograms of unemployment rate with a square root transform, and strike volume with a log transform. As one can see, these transforms appear to be closer to a Gaussian distribution than before. Therefore, we will use these transforms when forming the DAG.



We also observed that we have many NA's in our data set. Most notably, we have NA's for the years 1951-1959 for all countries in the density variable. There are a few ways we could consider dealing with this. One way is to impute that data using k-nearest neighbors to predict the values for this variable. We decided to not do this because of the data we have missing. That is we have data missing for all of the years between 1951-1959 and we feel that using k nearest neighbors may not give us an accurate prediction of what these density values are. For this reason, we decided to remove all rows that had an NA entry in them.

Below is the DAG for our data set. In the following table we report each variables children and parents. Additionally, we note that PC algorithm was unable to orient the edge between strike volume and unemployment.

Variables	Children	Parents
Strike Volume	Unemployment, Inflation, and Centralization	None
Left parliament	Centralization	None
Unemployment	Strike Volume and inflation	Strike Volume
Centralization	Density	Strike Volume and Left Parliament
Inflation	Density	Unemployment and Strike volume
Density	None	Inflation and Centralization



Now, using this DAG we will model each of the variables as a function of its parents. In the following tables, we list the coefficients for each parent in all of the models as well as an estimate of the standard deviation of the noise. For each of these estimated we included a bootstrapped 95% confidence interval. Lastly, because the PC algorithm failed to orient an edge between unemployment and strike.volume we consider both cases.

Table 2: Unemployment

		lower	upper
(Intercept)	0.9662	0.8198	1.0920
strike.volume	0.1912	0.1652	0.2207
SD of noise	0.6760	0.6296	0.7250

Table 3: Centralization

		lower	upper
(Intercept)	0.82400	0.72600	0.921000
strike.volume	-0.06210	-0.07420	-0.050000
left.parliament	-0.00256	-0.00444	-0.000469
SD of noise	0.28400	0.27400	0.298000

Table 4: Inflation

		lower	upper
(Intercept)	3.0140	2.0350	3.9920
unemployment	0.8145	0.2682	1.4040

		lower	upper
strike.volume	0.5295	0.3244	0.7264
SD of noise	4.0660	3.7220	4.3960

Table 5: Density

		lower	upper
(Intercept)	21.320	19.0600	23.830
inflation	1.129	0.8896	1.353
centralization	35.140	31.9900	38.210
SD of noise	10.520	9.8570	11.250

Table 6: Strike Volume

		lower	upper
(Intercept)	1.600	1.178	2.038
unemployment	1.477	1.249	1.681
SD of noise	1.879	1.763	1.995

We will estimate the affect that strike volume has on union density. As we observed before, the PC algorithm failed to orient the edge between strike volume and unemployment, and for that reason we must consider both models. The first DAG we will consider is where unemployment is a child of strike volume. In this model, there is no backdoor between strike volume and density and so we do not have to control for anything in our model. In the second DAG we must consider where strike volume is a child of unemployment we have a backdoor path from unemployment through inflation to density. To block this path we can control for inflation in our model. Lastly, we will also consider a third linear model where we predict the change in density with an increase of strike volume one standard deviation above the mean and holding all other variables at their mean.

We want to find the expected change in density when we increase the mean strike volume by one standard deviation. Our predictions can be found in the table below.

	Mean of Strike Volume	Estimate of Standard deviation	Change in Density
First Model	4.219	2.215	-1.056
Second Model	4.219	2.215	-4.065
Third model	4.219	2.215	2.946

An additional assumption made by the Gaussian independence conditional test that we used to form our DAG is that the relationship is linear between each of variables. We will now test to see if this assumption is plausible. We will do this by constructing a hypothesis test for the linear relationship between each of the variables. We will do this by first computing the MSE for a linear model and the MSE for a non-parametric model between two and computing the difference between these two values. Then, we will repeatedly re-sample our data set by bootstrapping the residuals of our linear model. We will then form our linear model and our non-parametric model on this new bootstrapped data set, and compute the difference in the MSE between these two models. We will then compute the p-value of our hypothesis test where the null is that the linear model is correct, and the alternative is that the linear model is not correct by counting the number of times the difference in MSE from the bootstrapped data is larger than the original MSE.

This works because of the fact that bootstrapping by re-sampling the residuals trusts that the model has

a correct shape for the regression, and we are assuming that our non-parametric regression curve fits the correct shape. If our linear model is not correct the bootstrapped data will change its distribution to be more linear and this will therefore lower its MSE. Because we assume the non-parametric curve fits the correct shape therefore the difference in MSE will be lowered.

Names	p_values
unemployment -> Strike Volume	0.000
Strike Volume -> Unemployment	0.019
Unemployment -> Inflation	0.054
Strike Volume -> Inflation	0.152
Strike Volume -> Centralization	0.004
Left Parliament -> Centralization	0.000
Centralization -> Density	1.000
Inflation -> Density	0.217

I think it is mostly unreasonable to model the relationship between variables as linear. In the above hypothesis tests, our null hypothesis is that the relationship is linear. Many of the p-values in the above table are very low, and I think most people would agree that it is reasonable to reject the null in the first, second, fifth, and sixth tests. This is half of the cases we considered. However, hypothesis tests are also sensitive to how much data we have. The null in this case is that the relationship between variables is actually linear. It could be the case that the relationship is not actually linear, but is close enough so that a linear model could still be useful. We could check this by visually inspecting the plots as well as visually checking the assumptions of linear regression such as no clear pattern in the residuals, the residuals are normally distributed and others.

Conclusion

We will now discuss the quality of our model. As we discussed before, the algorithm we used to form our DAG assumes that each variable is Gaussian distributed, and that the relationship between each variable is linear. The assumption that each variable is Gaussian distributed appeared to be incorrect at first, however we were able to correct it by transforming unemployment and strike volume. After this transformation unemployment and strike volume were now closer to normal than they were before, however unemployment now has a lot of entries that occur at one on the x axis and strike volume now has a very heavy left tail. Additionally, the histogram for centralization does not appear to be very close to normal. It appears to be closer to a uniform distribution than a normal distribution, and there is also a gap in the center of the histogram implying that there are no values of centralization between 1.8 and 2.0.

The algorithm we used to form our DAG also assumes a liner relationship between each of the variables. Testing this assumption with a formal hypothesis test we found that this likely did not hold for four of the 8 edges in our model. Therefore, the Gaussian assumption we made appears to be incorrect for at least one of variables, and the linear relationship between the variables assumption appears to be incorrect for at least four of the edges we tested. Additionally, our model failed orient an edge between strike volume and unemployment. Because of this, it is unclear from the DAG if strike volume causes unemployment or unemployment causes strike volume.

We will also discuss if the model is reasonable according to real world knowledge. We believe that for the most part the model is reasonable. Most of the edge orientations make sense and one could likely provide a reason for why it would be caused. For example, it makes sense that strike volume could affect inflation. If strike volume increases then companies that produce goods may have increase their costs in order to produce goods. They may have to hire new workers at a higher cost or rehire their old workers at a higher price. This would cause an increase in the costs of goods which could increase inflation. Additionally, we believe that the edge between strike volume and unemployment should likely be oriented with strike volume towards unemployment. We think it is more likely that strike volume will impact unemployment than unemployment

will affect strike volume. For example, if a union decides to strike a company may end up firing all of the employees which could lead to a large increase in unemployment.

Summary of Findings

We will now summarize our findings and provide advice to policy makers. In the above paper, we studied the factors that may lead to an increase in strike volume. To do this we attempted to form a graph that would show the causal effects between different variables. We attempted to form this by using an algorithm which makes some testable assumptions about the relationships between the variables and the variables themselves. Using formal statistical tests to check these assumptions, we found that these assumptions may not be correct, however we attempted to correct these assumptions as much as possible. Therefore, the recommendations made in this paper should be accepted with caution.

From our graph, we found that the only variable that may have a causal effect on strike volume is unemployment. We can estimate this effect and we found that when the square root of unemployment increases by one we expect the log of strike volume to increase by approximately 1.5. However, our model was inconclusive in whether or not unemployment has a causal effect on strike volume or strike volume has a causal effect on unemployment. We also note that because of the transformations we made on this data set this result may be difficult to interpret. In future studies, we may consider estimating the effect unemployment has on strike volume without any transformation.

Even though our study did not find much information about the factors that can affect strike volume, we were able to identify possible effects strike volume may have on other variables in our data set. We found that strike volume likely impacts centralization and inflation and we have inconclusive evidence that it may affect unemployment. Additionally, the impact strike volume has on inflation may be confounded by unemployment, or in other words the effect strike volume has on inflation may be due to the relationship between strike volume, unemployment, and inflation and not just strike volume and inflation. For this reason to try to estimate the effect strike volume has on inflation we must control for it, or take its effect into account.

Estimating these effects we found that when the log of strike volume increases by one we expect the amount of inflation in the economy to increase by about .5 percent. Additionally, we found that when the log of strike volume increases by one we expect centralization to decrease by .06. This implies that when people decide to go on strike the centralization of authority in unions appear to decrease. This may imply that strikes are usually not a unanimous decision, and when strikes occur it may lead to a separation of unions or different authorities in unions. Additionally, as we stated before our model was inconclusive of if unemployment affects strike volume or if strike volume affects unemployment. Therefore, we will also estimate the effect that strike volume has on unemployment. We found that when the log of strike volume increases by one we expect the square root of unemployment to increase by .19. This together with the previous result imply that strike volume and unemployment tend to increase together however it is not clear which causes the other.

Our model also shows that there is no causal effect from strike volume density, however there is a path from strike volume to density in our model so we there is still a relationship between strike volume and density. We again consider the two possibilities where unemployment is caused from strike volume and strike volume is caused from unemployment. When there is a causal effect from strike volume to unemployment we found that an increase in one of log strike volume leads to a decrease of about 1 in the union density. When there is a causal effect from unemployment to strike volume we found that this effect changes to a decrease of about 4 in the union density.