402 final strikes

Kaili Chen

5/9/2019

Contents

Load data, check where NA values are.

```
strikes = read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/19/exams/2/strikes.csv")
sum(is.na(strikes)) == sum(is.na(strikes$density))
```

```
## [1] TRUE
```

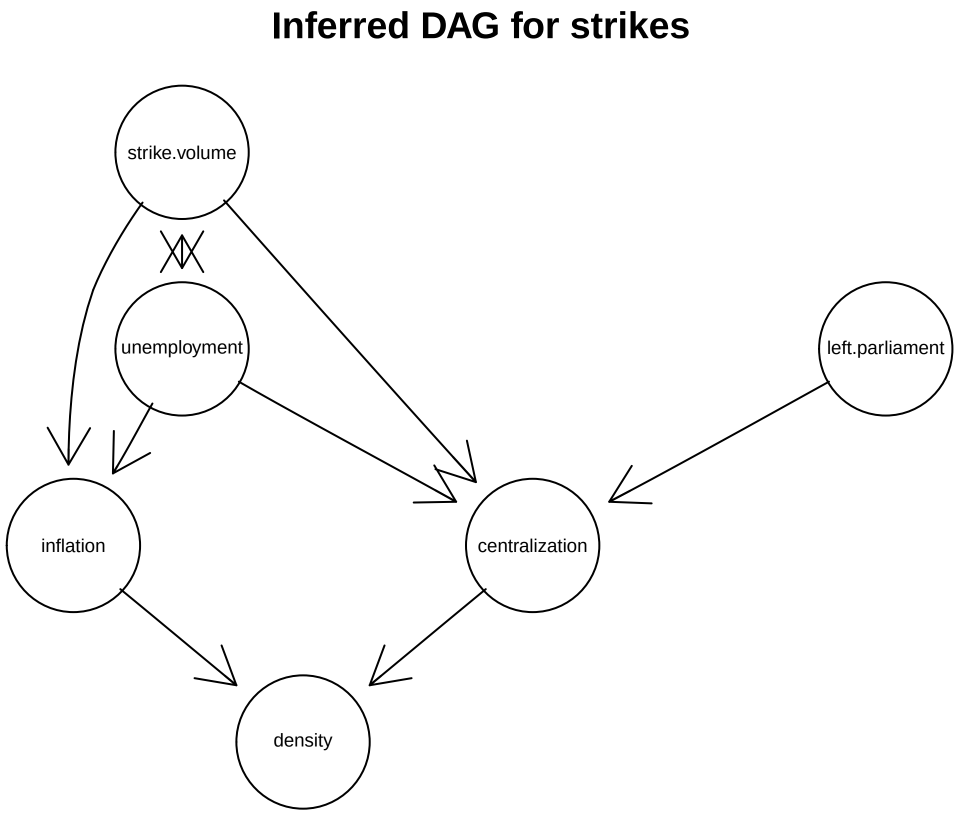
The only column with NA values is density

1.

(10) Use pc() from pcalg to obtain a graph, assuming all relations between variables are linear. Report the causal parents (if any) and children (if any) of every variable. If the algorithm is unable to orient one or more of the edges, report this, and in later parts of this problem, consider all the graphs which result from different possible orientations.

```
library(pcalg)
#Since density is considered in the graph model
#use the dataset after removing observations with NA values in the 'density' column
strikes.complete <- na.omit(strikes)
#I am leaving country and year out because country contains characters(not numeric values) and year is
strikes.dag <- pc(suffStat=list(C=cor(strikes.complete[,c(-1,-2)]), n=nrow(strikes.complete)),
               indepTest=gaussCItest, labels=colnames(strikes.complete)[c(-1,-2)],
               alpha=0.05)
plot(strikes.dag, main="Inferred DAG for strikes")
```

```
## Loading required namespace: Rgraphviz
```



Strike.volume is the parent of inflation, unemployment and centralization. Unemployment is the parent of strike.volume, inflation and centralization. Inflation is the parent of strike.volume and unemployment. Density has no descedents. Centralization is the parent of density. Left.parliament is the parent of centralization. The algorithm is unable to orient the edge between strike.volume and unemployment. Therefore, I will consider both directions in later tasks.

2.

(12) Linearly model each variable as a function of its parents. Report the coefficients (to reasonable precision), the standard deviation of the regression noise (ditto), and 95% confidence intervals for all of these, as determined by bootstrapping the residuals.

The estimates and 95% confidence interval for coefficients and standard deviation of the regression noise are presented in the tables below.

```r
#Resample residuals:
resample <- function(x) {
  sample(x,size=length(x),replace=TRUE)
}
```

1) strikevolume + unemployment->inflation

```r
mdl1 = lm(inflation~strike.volume+unemployment, data = strikes)
est1 = coefficients(mdl1)
sigma1 = sqrt(sum(mdl1$residuals^2)/(length(strikes)-2))
coefs.strikes.lm.isu <- function(df) {
  fit <- lm(inflation~strike.volume+unemployment,data=df)
  sigma = sqrt(sum(fit$residuals^2)/(length(df)-2))
  return(c(coefficients(fit),sigma))
}
```

```r
sim.strikes.resids.isu <- function() {
  model = lm(inflation~strike.volume+unemployment, data = strikes)
  new.data<-strikes
  noise <- resample(residuals(model))
  new.data[,"inflation"] <- fitted(model) + noise
  return(new.data)
}
strikes.lm.samp.dist.resids.isu <- replicate(1000,
   coefs.strikes.lm.isu(sim.strikes.resids.isu()))
ci1 = t(apply(strikes.lm.samp.dist.resids.isu,1,quantile,c(0.5,0.025,0.975)))
colnames(ci1) = c("estimate","2.5%","97.5%")
rownames(ci1) = c("intercept","strike.volume","unemployment","sd(noise)")
library(knitr)
kable(signif(ci1,3))
```

	estimate	2.5%	97.5%
intercept	4.72000	4.17e+00	5.26000
strike.volume	0.00169	9.77e-04	0.00246
unemployment	0.20900	9.88e-02	0.32900
sd(noise)	45.50000	4.22e+01	49.30000

2) strikevolume->unemployment

```r
mdl2 = lm(unemployment~strike.volume, data = strikes)
est2 = coefficients(mdl2)
coefs.strikes.lm.us <- function(df) {
  fit <- lm(unemployment~strike.volume,data=df)
  sigma = sqrt(sum(fit$residuals^2)/(length(df)-2))
  return(c(coefficients(fit),sigma))
}
sim.strikes.resids.us <- function() {
  model = lm(unemployment~strike.volume, data = strikes)
  new.data<-strikes
  noise <- resample(residuals(model))
  new.data[,"unemployment"] <- fitted(model) + noise
  return(new.data)
}
strikes.lm.samp.dist.resids.us <- replicate(1000,
   coefs.strikes.lm.us(sim.strikes.resids.us()))
ci2 = t(apply(strikes.lm.samp.dist.resids.us,1,quantile,c(0.5, 0.025,0.975)))
colnames(ci2) = c("estimate","2.5%","97.5%")
rownames(ci2) = c("intercept","strike.volume","sd(noise)")
kable(signif(ci2,3))
```

	estimate	2.5%	97.5%
intercept	3.20000	2.93e+00	3.48000
strike.volume	0.00126	8.22e-04	0.00177
sd(noise)	30.30000	2.80e+01	32.60000

3) strike.volume+unemployment+left.parliament->centralization

```
mdl3 = lm(centralization~strike.volume+unemployment+left.parliament, data = strikes)
est3 = coefficients(mdl3)
coefs.strikes.lm.cs <- function(df) {
  fit <- lm(centralization~strike.volume+unemployment+left.parliament,data=df)
  sigma = sqrt(sum(fit$residuals^2)/(length(df)-2))
  return(c(coefficients(fit),sigma))
}
sim.strikes.resids.cs <- function() {
  model = lm(centralization~strike.volume+unemployment+left.parliament, data = strikes)
  new.data<-strikes
  noise <- resample(residuals(model))
  new.data[,"centralization"] <- fitted(model) + noise
  return(new.data)
}
strikes.lm.samp.dist.resids.cs <- replicate(1000,
   coefs.strikes.lm.cs(sim.strikes.resids.cs()))
ci3 = t(apply(strikes.lm.samp.dist.resids.cs,1,quantile,c(0.5, 0.025,0.975)))
colnames(ci3) = c("estimate","2.5%","97.5%")
rownames(ci3) = c("intercept","strike.volume","unemployment","left.parliament","sd(noise)")
kable(signif(ci3,3))
```

	estimate	2.5%	97.5%
intercept	0.670000	0.594000	7.46e-01
strike.volume	-0.000127	-0.000174	-7.85e-05
unemployment	-0.018500	-0.026200	-1.04e-02
left.parliament	-0.002680	-0.004390	-1.00e-03
sd(noise)	3.010000	2.890000	3.13e+00

4) unemployment->strike.volume

```
mdl4 = lm(strike.volume~unemployment, data = strikes)
est4 = coefficients(mdl4)
coefs.strikes.lm.su <- function(df) {
  fit <- lm(strike.volume~unemployment,data=df)
  sigma = sqrt(sum(fit$residuals^2)/(length(df)-2))
  return(c(coefficients(fit),sigma))
}
sim.strikes.resids.su <- function() {
  model = lm(strike.volume~unemployment, data = strikes)
  new.data<-strikes
  noise <- resample(residuals(model))
  new.data[,"unemployment"] <- fitted(model) + noise
  return(new.data)
}
strikes.lm.samp.dist.resids.su <- replicate(1000,
   coefs.strikes.lm.us(sim.strikes.resids.su()))
ci4 = t(apply(strikes.lm.samp.dist.resids.su,1,quantile,c(0.5, 0.025,0.975)))
colnames(ci4) = c("estimate","2.5%","97.5%")
rownames(ci4) = c("intercept","unemployment","sd(noise)")
kable(signif(ci4,3))
```

	estimate	2.5%	97.5%
intercept	2.76e+02	234.0000	318.00

	estimate	2.5%	97.5%
unemployment	3.92e-02	-0.0184	0.13
sd(noise)	4.93e+03	3660.0000	6430.00

5) centralization + inflation->density

```
mdl5 = lm(density~inflation+centralization, data = strikes.complete)
est5 = coefficients(mdl5)
coefs.strikes.lm.dic <- function(df) {
  fit <- lm(density~inflation+centralization,data=df)
  sigma = sqrt(sum(fit$residuals^2)/(length(df)-2))
  return(c(coefficients(fit),sigma))
}
sim.strikes.resids.dic <- function() {
  model = lm(density~inflation+centralization, data = strikes.complete)
  new.data<-strikes.complete
  noise <- resample(residuals(model))
  new.data[,"density"] <- fitted(model) + noise
  return(new.data)
}
strikes.lm.samp.dist.resids.dic <- replicate(1000,
   coefs.strikes.lm.dic(sim.strikes.resids.dic()))
ci5 = t(apply(strikes.lm.samp.dist.resids.dic,1,quantile,c(0.5, 0.025,0.975)))
colnames(ci5) = c("estimate","2.5%","97.5%")
rownames(ci5) = c("intercept","inflation","centralization","sd(noise)")
kable(signif(ci5,3))
```

	estimate	2.5%	97.5%
intercept	21.30	18.900	23.60
inflation	1.13	0.913	1.37
centralization	35.10	32.100	38.20
sd(noise)	90.00	83.900	96.30

3.

You should find that there is no edge between strike volume and union density (neither is the parent of the other), but that there is at least one directed path linking them (either density is an ancestor of strike volume, or the other way around)

(a)

(8) Find the expected change in the descendant from a one-standard-deviation increase in the ancestor above its mean value.

This question is similar as asking if strike.volume is increased from 288.68 by the amount of 493.1012, by how much is density affected? According to the graph generated above, density directly depends on inflation and contralization. Inflation directly depends on strike.volume and unemployment. Centralization directly depends on unemployment, strike.volume and left.parliament. Since our ancestor in this relationship is strike.volume, we can see left.parliament as a fixed variable. Thus, the question reduces to how the change of strike.volume affects inflation and centralization and consequently, how the change of inflation and centralization directly affects their child – 'density'. The first step can be broken down to how strike.volume directly affects inflation and centralization and how strike.volume directly affect unemployment, which can directly affect inflation and centralization. Since in the graphical model, we assume any parent-child relationship is linear, we can read off the coefficients in the linear model from Q2 as the effect of a unit change in the parent on the child.

```r
#ancestor: strike.volume
#descendent: density
u = mean(strikes$strike.volume)
#one standard deviation of ancestor
change = sd(strikes$strike.volume)
# strike.volume increase 493.1012, by how much is density changed?
# 1) effect of increasing strike.volume by 493.1012 on inflation (inflation~strike.volume)
effect1 = ci1["strike.volume","estimate"]*change
# 2) effect of increasing strike.volume by 493.1012 on centralization(centralization~strike.volume)
effect2 = ci3["strike.volume","estimate"]*change
# 3) effect of increasing strike.volume by 493.1012 on unemployment(unemployment~strike.volume)
effect3 = ci2["strike.volume","estimate"]*change
# 4) effect of changing unemployment by 3) on inflation (inflation~unemployment)
effect4 = ci1["unemployment","estimate"]*effect3
# 5) effect of changing unemployment by 3) on centralization (centralization~unemployment)
effect5 = ci3["unemployment","estimate"]*effect3
# 6) effect of changing inflation in total by 1)+4)
effect6 = effect1+effect4
# 7) effect of changing centralization in total by 2)+5)
effect7 = effect2+effect5
# 8) effect of changing inflation on density (density~inflation)
effect8 = ci5["inflation","estimate"] * effect6
# 9) effect of changing centralization on density(density~centralization)
effect9 = ci5["centralization","estimate"] * effect7
# 10) change of density in total 8) + 9)
signif(effect8+effect9,3)
```

```
## [1] -1.51
```

According to the calculation shown in the code, the expected change in density from a one-standard-deviation increase in strike.volume above its mean value is -1.51.

(b)

(5) Linearly regress the descendant on all the other variables, including the ancestor. According to this regression, what is the expected change in the descendant, when the ancestor increases one SD above its mean value and all other variables are at their mean values?

```r
mdl_3b = lm(density~. -density-country-year, data = strikes.complete)
signif(coefficients(mdl_3b)["strike.volume"]*change,3)
```

```
## strike.volume
##          1.97
```

By linearly regressing density on strike.volume, unemployment, inflation, left.parliament and centralization, the coefficient of strike.volume in the model tells us that incraesing the ancestor(strike.volume) by one SD above its mean value, and controlling all the other variables, results in an increase of 1.97 in the descendent(density).

4.

Check the linearity assumption for each variable which has a parent.

(a)

(5) Describe your method, and why it should work.

I conducted a hypothesis test on the linearity between variables that has a parent-child relationship in the DAG generated above. The null hypothesis is that there is a linear relationship between them. I used the original data to simulate 100 data frames based on the linear model, and then I calcualte the differences in in-sample MSEs between the linear model and the non-parametric model on the simulated data frames. By calculating how many times the difference in MSEs based on the simulated data are greater than the estimated difference in MSE on the original data, I get the p-value. From the p-value, I can draw conclusions on the whether hypothesis seems plausible. This method would work because if the parametric model is right, it should predict as well as, or even better than,the non-parametric one, and we can check whether $MSE(\hat{\theta}) - MSE_{np}(\hat{\mu})$ is sufficiently small. [Chapter9, ADAfaEPoV]

(b)

(5) Report the p-value for each case, to reasonable precision.

```
#test linearity between strike.volume and unemployment
mdl2.lm = lm(unemployment~strike.volume, data = strikes)
mdl2.np = npreg(unemployment~strike.volume, data = strikes)
sim.lm <- function(linfit,test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(strike.volume = test.x)
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, unemployment = y.sim)
    return(sim.frame)
}
calc.D <- function(data) {
    MSE.p <- mean((lm(unemployment~strike.volume, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(unemployment~strike.volume, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D <- replicate(100, calc.D(sim.lm(mdl2.lm, strikes$strike.volume)))
d.hat = mean(mdl2.lm$residual^2) - mdl2.np$MSE
sum(null.samples.D > d.hat)/100
```

```
## [1] 0
```

Since the p-value is small, we can confidently reject the linearity between strike.volume and unemployment. Thus, the linearity assumptions on both directions (strike.volume->unemployment and unemployment->strike.volume) is violated.

```
#test linearity between strike.volume and inflation
inf.sv.lm = lm(inflation~strike.volume, data = strikes)
inf.sv.np  = npreg(inflation~strike.volume, data = strikes)
sim.lm2 <- function(linfit, test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(strike.volume = test.x)  #predictor vec = test.x
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, inflation = y.sim) #response vec = y.sim
    return(sim.frame)
}
calc.D2 <- function(data) {
    MSE.p <- mean((lm(inflation~strike.volume, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(inflation~strike.volume, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D2 <- replicate(100, calc.D2(sim.lm2(inf.sv.lm, strikes$strike.volume)))
d.hat2 = mean(inf.sv.lm$residual^2) - inf.sv.np$MSE
sum(null.samples.D2 > d.hat2)/100
```

```
## [1] 0.04
```

Since p value is small(0.03), we can confidently reject the linearity between strike.volume and inflation

```
#test linearity between unemployment and inflation
inf.u.lm = lm(inflation~unemployment, data = strikes)
inf.u.np  = npreg(inflation~unemployment, data = strikes)
sim.lm3 <- function(linfit, test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(unemployment = test.x)  #predictor vec = test.x
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, inflation = y.sim) #response vec = y.sim
    return(sim.frame)
}
calc.D3 <- function(data) {
    MSE.p <- mean((lm(inflation~unemployment, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(inflation~unemployment, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D3 <- replicate(100, calc.D3(sim.lm3(inf.u.lm, strikes$strike.volume)))
d.hat3 = mean(inf.u.lm$residual^2) - inf.u.np$MSE
sum(null.samples.D3 > d.hat3)/100
```

```
## [1] 0.71
```

Since p value very high, we don't have sufficient evidence to reject the linearity between strike.volume and inflation. We don't have enough data to draw conclusions.

```
#test linearity between centralization and stike.volume
ctr.sv.lm = lm(centralization~strike.volume, data = strikes)
ctr.sv.np  = npreg(centralization~strike.volume, data = strikes)
sim.lm4 <- function(linfit, test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(strike.volume = test.x)  #predictor vec = test.x
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, centralization = y.sim) #response vec = y.sim
    return(sim.frame)
}
calc.D4 <- function(data) {
    MSE.p <- mean((lm(centralization~strike.volume, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(centralization~strike.volume, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D4 <- replicate(100, calc.D4(sim.lm4(ctr.sv.lm, strikes$strike.volume)))
d.hat4 = mean(ctr.sv.lm$residual^2) - ctr.sv.np$MSE
sum(null.samples.D4 > d.hat4)/100
```

```
## [1] 0
```

Since p value is small, we can confidently reject the linearity between strike.volume and centralization

```
#test linearity between unemployment and centralization
ctr.u.lm = lm(centralization~unemployment, data = strikes)
ctr.u.np  = npreg(centralization~unemployment, data = strikes)
sim.lm5 <- function(linfit, test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(unemployment = test.x)  #predictor vec = test.x
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, centralization = y.sim) #response vec = y.sim
    return(sim.frame)
}
calc.D5 <- function(data) {
    MSE.p <- mean((lm(centralization~unemployment, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(centralization~unemployment, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D5 <- replicate(100, calc.D5(sim.lm5(ctr.u.lm, strikes$unemployment)))
d.hat5 = mean(ctr.u.lm$residual^2) - ctr.u.np$MSE
sum(null.samples.D5 > d.hat5)/100
```

```
## [1] 0
```

Since p value is small, we can confidently reject the linearity between unemployment and centralization

```
#test linearity between left.parliament and centralization
ctr.lp.lm = lm(centralization~left.parliament, data = strikes)
ctr.lp.np  = npreg(centralization~left.parliament, data = strikes)
sim.lm6 <- function(linfit, test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(left.parliament = test.x)  #predictor vec = test.x
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, centralization = y.sim) #response vec = y.sim
    return(sim.frame)
}
```

```
calc.D6 <- function(data) {
    MSE.p <- mean((lm(centralization~left.parliament, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(centralization~left.parliament, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D6 <- replicate(100, calc.D6(sim.lm6(ctr.lp.lm, strikes$left.parliament)))
d.hat6 = mean(ctr.lp.lm$residual^2) - ctr.lp.np$MSE
sum(null.samples.D6 > d.hat6)/100
```

```
## [1] 0
```

Since p value is small, we can confidently reject the linearity between left.parliament and centralization

```
#test linearity between density and centralization
d.ctr.lm = lm(density~centralization, data = strikes)
d.ctr.np  = npreg(density~centralization, data = strikes)
sim.lm7 <- function(linfit, test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(centralization = test.x)  #predictor vec = test.x
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, density = y.sim) #response vec = y.sim
    return(sim.frame)
}
calc.D7 <- function(data) {
    MSE.p <- mean((lm(density~centralization, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(density~centralization, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D7 <- replicate(100, calc.D7(sim.lm7(d.ctr.lm, strikes$centralization)))
d.hat7 = mean(d.ctr.lm$residual^2) - d.ctr.np$MSE
sum(null.samples.D7 > d.hat7)/100
```

```
## [1] 0
```

Since p value is small, we can confidently reject the linearity between density and centralization

```
#test linearity between density and inflation
d.inf.lm = lm(density~inflation, data = strikes)
d.inf.np  = npreg(density~inflation, data = strikes)
sim.lm8 <- function(linfit, test.x) {
    n <- length(test.x)
    sim.frame <- data.frame(inflation = test.x)  #predictor vec = test.x
    sigma <- summary(linfit)$sigma * (n - 2)/n
    y.sim <- predict(linfit, newdata = sim.frame)
    y.sim <- y.sim + rnorm(n, 0, sigma)
    sim.frame <- data.frame(sim.frame, density = y.sim) #response vec = y.sim
    return(sim.frame)
}
calc.D8 <- function(data) {
    MSE.p <- mean((lm(density~inflation, data = data)$residuals)^2)
    MSE.np.bw <- npregbw(density~inflation, data = data)
    MSE.np <- npreg(MSE.np.bw)$MSE
    return(MSE.p - MSE.np)
}
null.samples.D8 <- replicate(100, calc.D8(sim.lm8(d.inf.lm, strikes$inflation)))
d.hat8 = mean(d.inf.lm$residual^2) - d.inf.np$MSE
sum(null.samples.D8 > d.hat8)/100
```

```
## [1] 0.21
```

Since p value is greater than 0.05, we don't have enough evidence to reject the linearity between density and inflation.

(c)

(5) What is your overall judgment about whether it is reasonable to model each endogenous variable as linearly related to its parents? If you need more information than just p-values to reach a decision, describe it.

Based on 4(b), since 6 out of 8 linear models assumed are rejected by the specification test, it is not reasonable to model these 6 endogenous variables as linearly related to its parents. 2 out of 8 linear models have high p-values in the specification test, meaning that we don't have enough data to draw conclusions. In gerneral, it is not reasonable to model all the endogenous variables as linearly related to their parents. In fact, we are confident that 6 of the relationship is not linear. Besides, we need more data to decide whether it is reasonable to treat the 2 variables (density and inflation) as linearly related to their parents.

5.

(10) Discuss the overall adequacy of the model, on both statistical grounds (goodness-of-fit, appropriateness of modeling assumptions, etc.) and substantive, scientific ones (whether it makes sense, given what is known about the processes involved).

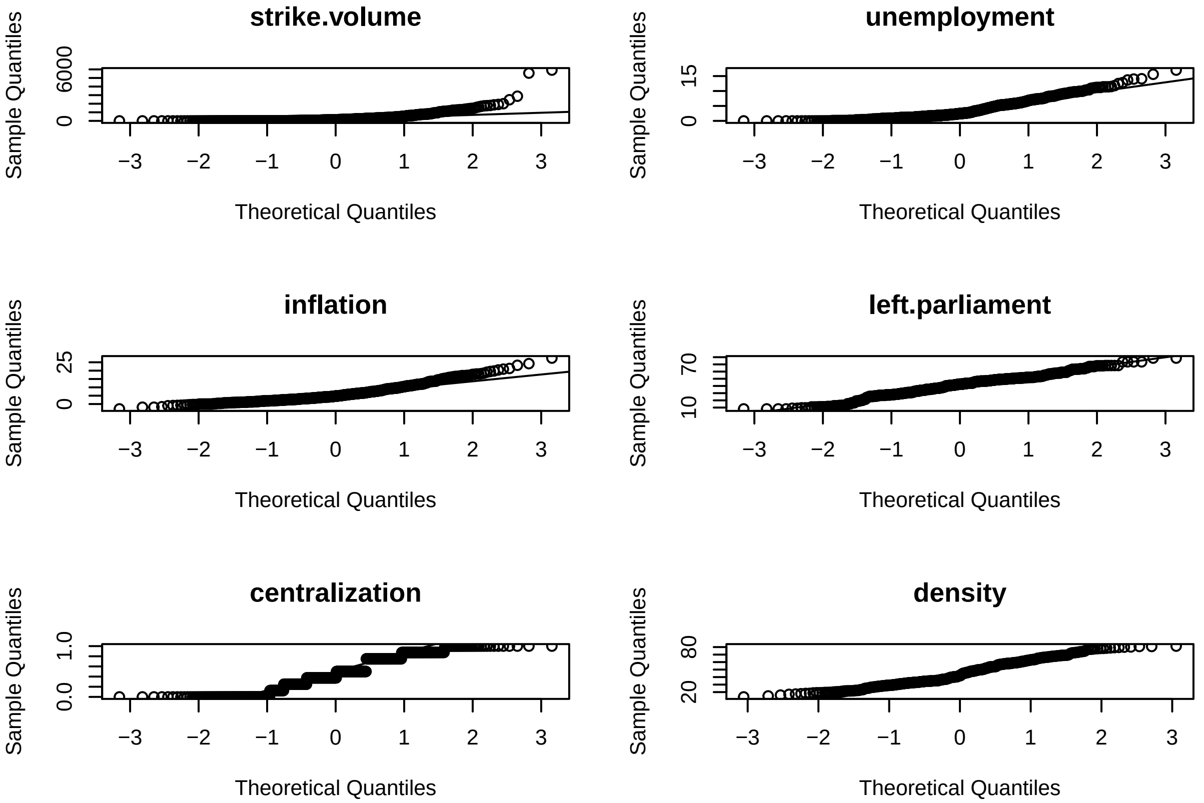
The Gaussian conditional independence test presumes that all of the variables involved have a joint distribution which is multivariate Gaussian. This entails that each variable has a Gaussian marginal distribution, and that each variable is a linear function of any of the others (plus independent Gaussian noise).[Solutions to HW12]

First, based on the specification tests conducted in 4c, we can say that it is very implausible to assume that each variable is a linear function of any of the others (plus independent Gaussian noise). In particular, the only two pairs of variables that may have linear relationship are "inflation and unemployment", as well as "density and inflation". According to the estimated coefficients of linear models calculated in in Q2, the increase in unemployment rate would lead to an increase in inflation. However, this does not make scientific sense because the relationship between inflation and unemployment has traditionally been an inverse correlation. Moreover, the model suggests an increase in inflation would cause an increase in union density. This relationship is plausible based on a scientific research.[Blaschke, S. (2000). Union Density and European Integration: Diverging Convergence. European Journal of Industrial Relations, 6(2), 217–236. https://doi.org/10.1177/095968010062006]

Second, we can check Q-Q plots for each variable to evaluate the assumption that each variable has a Gaussian marginal distribution.

```
#[Solutions to HW12]
# Create a mosaic of Gaussian Q-Q plots for each column in a data frame
  # Decorate each plot with the traditional ine through the quartiles
# Inputs: data frame or array (data)
# Outputs: none
# Side-effect: creates a mosaic of Gaussian Q-Q plots, with lines
# Presumes: all columns in data have names
```

```r
# Presumes: all columns in data are numeric
qqnorm.mosaic <- function(data) {
    p <- ncol(data)
    # Figure out how many rows and columns we need --- try to avoid
    # lots of wasted space
    rows <- ceiling(sqrt(p))
    columns <- ceiling(p/rows)
    par(mfrow=c(rows, columns))
    for (var in colnames(data)) {
        qqnorm(data[,var], main=var)
        qqline(data[,var])
    }
}
# Make Gaussian QQ plots for everything except year and country
qqnorm.mosaic(strikes[,c(-1,-2)])
```



Looking at the Q-Q plots suggests that only two variables (left.parliament and density) have approximately evenly Gaussian marginal distributions.

Third, The PC algorithm crucially assumes that all of the causally relevant variables are included in the data ("causal sufficiency"). More precisely, any variable not included has to be completely independent of the ones which are. For instance, if the communication barrier between empolyees and managers mattters, and have some sort of systematic relationship to the variables we have included, PC offers no guarantees at all. It is, logically, very hard to use observations on the variables you do have to conclude that there must be some other variable you don't have. [Solutions to HW12]

To sum up, although this model offers a systematic way to represent the interrelation between the variables in our observed data, the assumptions for the PC algorithm that is used to generate this model are challenged by the data itself, and we can't really rely heavily on this model since there might be some latent variables not included in our dataset but in fact play a role in this interrelation.

Summary report (extra credit)

(5) Write a one-page summary of your findings. As much as possible, use ordinary language (as opposed to mathematical formulas, computer code, or statistical jargon). You may find it useful to adopt the perspective of trying to give advice to a policy-maker who would like to know what actions to take to reduce (or increase) strikes.

Given a dataset which includes events of strikes in 18 developed (OECD) countries during 1951–1985, I use statistical methods to analyze the significance of six varibales related to a strike, and the interrelation between the six variables. These variables are Strike volume, defined as "days [of work] lost due to industrial disputes per 1000 wage salary earners"; Unemployment rate (percentage); Inflation rate (consumer prices, percentage); "parliamentary representation of social democratic and labor parties".(For the United States, this is the fraction of Senate seats held by the Democratic Party.); a measure of the centralization of the leadership in that country's union movement, on a scale of 0 to 1; Union density, the fraction of salary earners belonging to a union (only available from 1960).

The statistical model generated by a computer algorithm implies a couple of inferences on how the six factors are related. Unfortunately, since a lot of the assumptions I used to generate the model were being questioned by the observations, I was only able to generate a few well-supported inferences.

First, an increase in unemployment rate will lead to an increase in inflation. This inference is against the well-known Phillips curve (inverse relation between unemployent rate and inflation) but it is what the observations suggests. A possible explanations could be that the real relationship between inflation and unemployment is actually more complicated than the Phillips curve suggests, and in the data we are given, we only have limited number of observations and limited number of variables we are concerned.

Second, the data both support that an increase in inflation rate can lead to an increase in union density.

It is hard to use these two inferences to make recommendations on what actions to take to reduce strikes. However, policy makers are suggested to seek factors that may affect strikes outside of the six variables considered here, and that they are encouraged to collect more observations on existing factors in order to draw convincing conclusions on what factors affects the occurrence of strikes and how these factors affect the occurrence of strikes (induce or prevent).