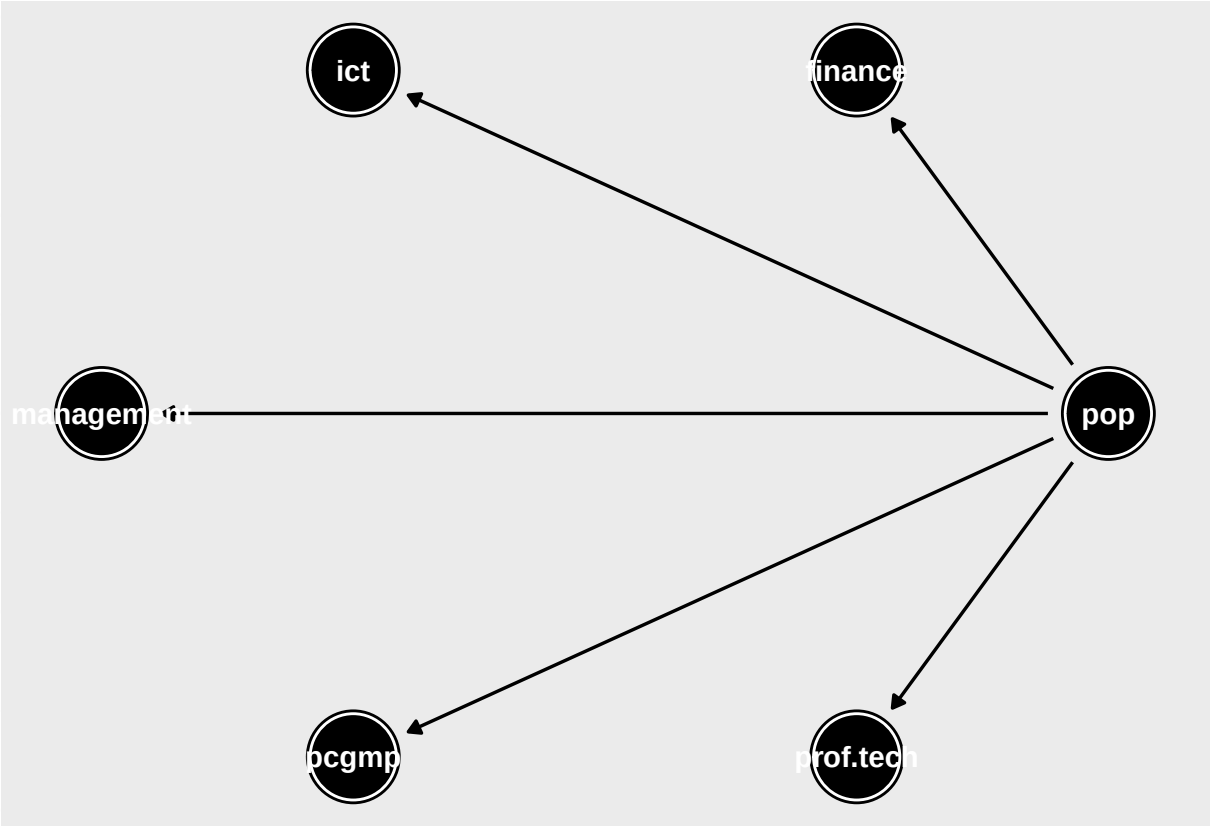
36-402 Homework 12

Eu Jing Chua
eujingc

May 19, 2019

Question 1

Q1 a)



However, multiple other graphs are possible as the theory says nothing about the relationships between per-capita output and the four industries, in which case any set of relationships would be compatible with the theory.

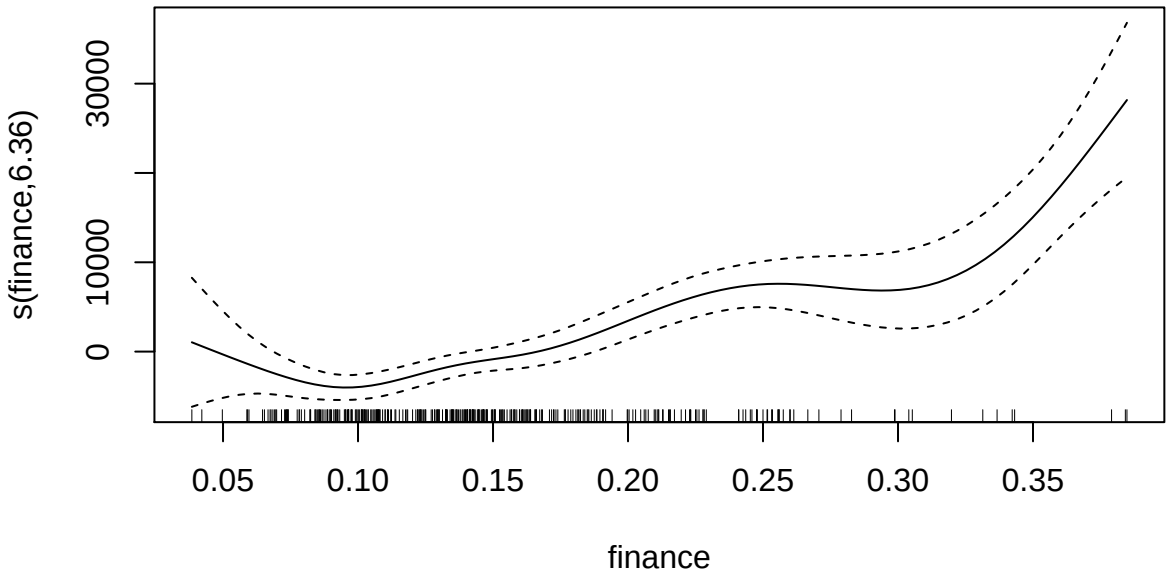
Q1 b)

`pcgmp` $\not\perp\!\!\!\perp$ `finance` as there is an open path through population.
`pcgmp` $\perp\!\!\!\perp$ `finance` | `pop` as all paths are closed.
`pcgmp` $\perp\!\!\!\perp$ `finance` | `pop`, `management` as all paths are closed.

Q1 c)

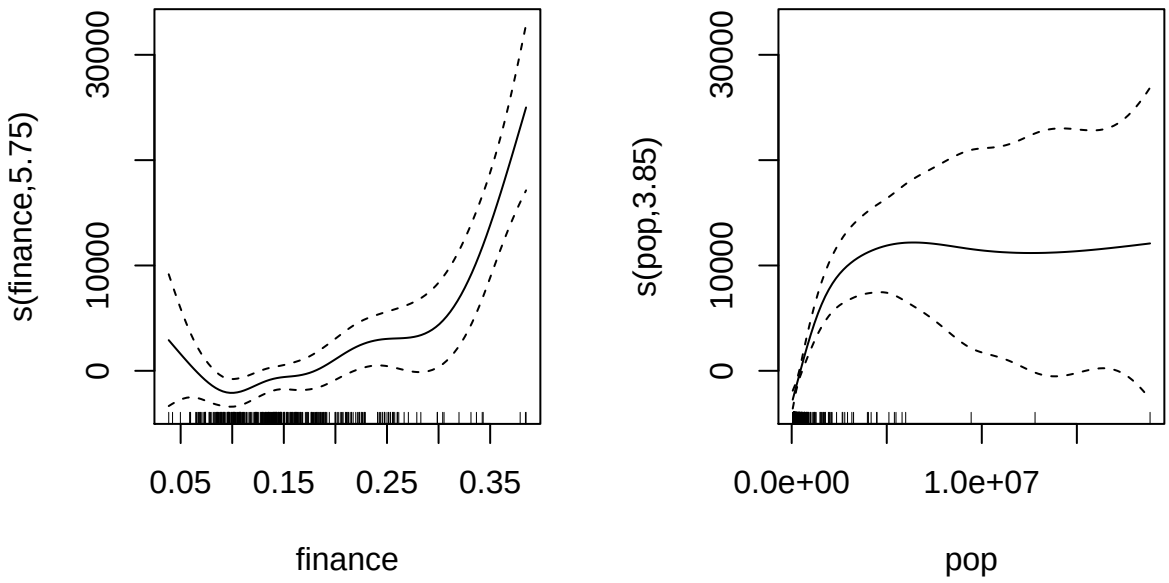
We can roughly test for the dependence between variables by assuming an additive mixture model using spline smoothing for each variable.

In the first model, we test for '*pcgmp*' $\not\perp\!\!\!\perp$ '*finance*', in which we only model `pcgmp` against `finance`. The partial response is as follows:



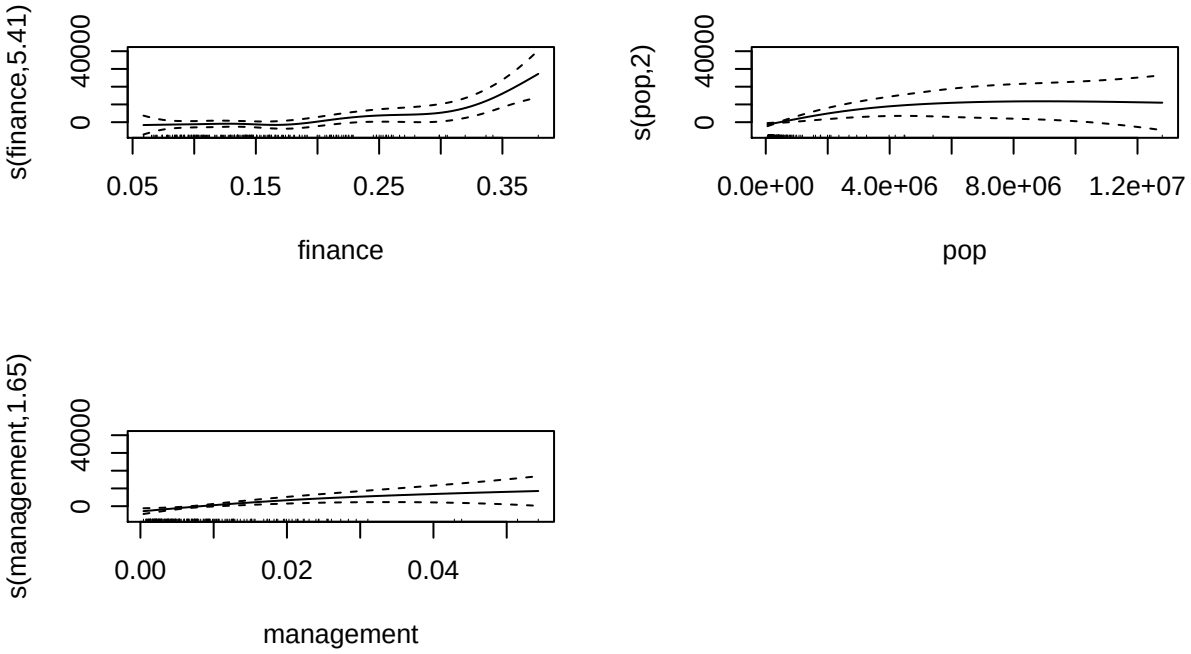
We can see a relationship exists between `pcgmp` and `finance`, where in general as `finance` increases, so does `pcgmp`. Thus, this seems to provide evidence that the dependence exists.

In the second model, we test for '$pcgmp$' ⫫ '$finance$' | 'pop', in which we model `pcgmp` against `finance`, controlling for `pop`. The partial responses are as follows:



However, the additive model shows that after controlling for `pop`, the partial response of `finance` still has a similar relationship as before, where higher values of `finance` are related to higher values of `pcgmp`. Thus, the data does not support this conditional independence.

In the third model, we test for '*pcgmp*' ⫫ '*finance*' | '*pop*', '*management*', in which we model `pcgmp` against `finance`, controlling for `pop` and `management`. The partial responses are as follows:

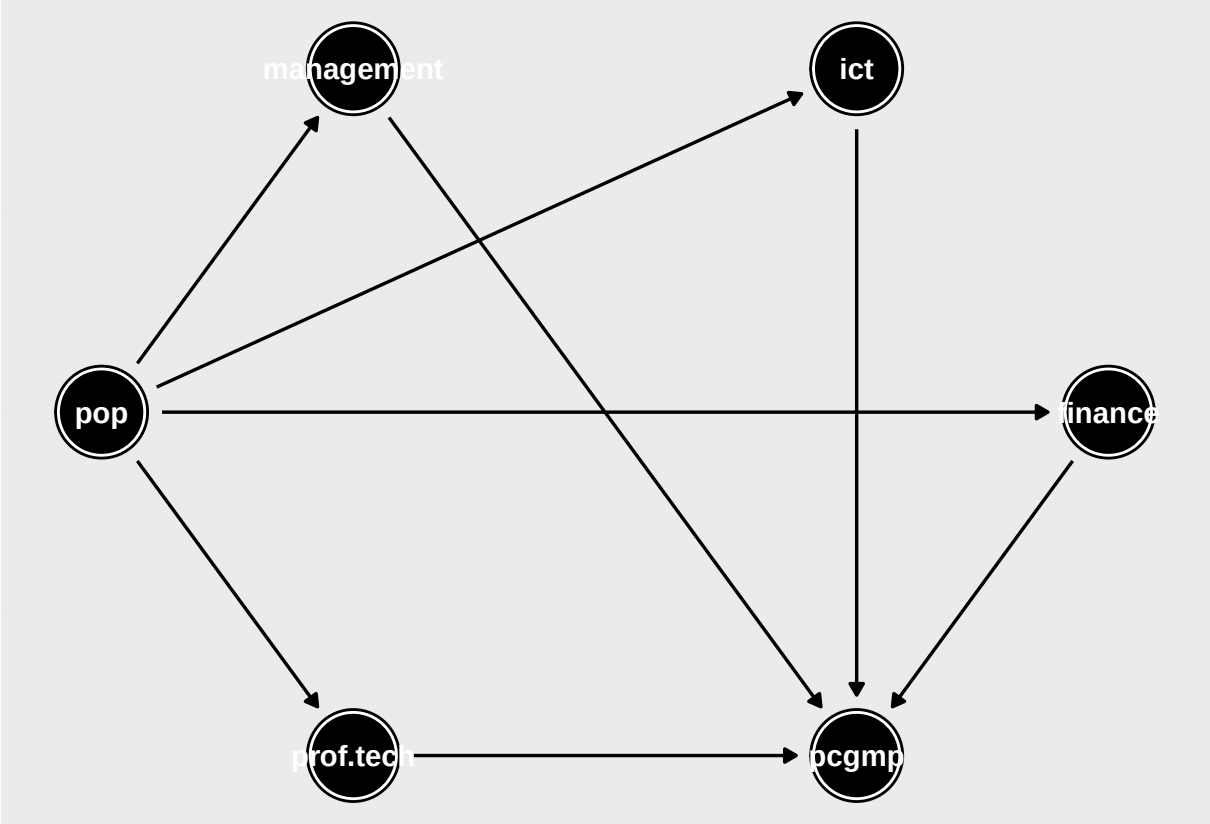


Similar to above, we can see that the partial response of `finance` again shows an increasing relationship, even after controlling for both variables. However, once again the data does not support this conditional independence.

Q1 d) Assuming the DAG above is right and that the relations can be modeled with an additive model with spline smoothing, we can first fit an additive model for `pcgmp` against `pop` with the whole dataset, using spline smoothing. With this model, we then predict the `pcgmp` where we fix `pop` to be the population size of Pittsburgh, and then another prediction where we fix `pop` to be double the previous size. The difference in these two predictions would then be the estimated average effect, assuming the DAG was right.

Question 2

Q2 a)



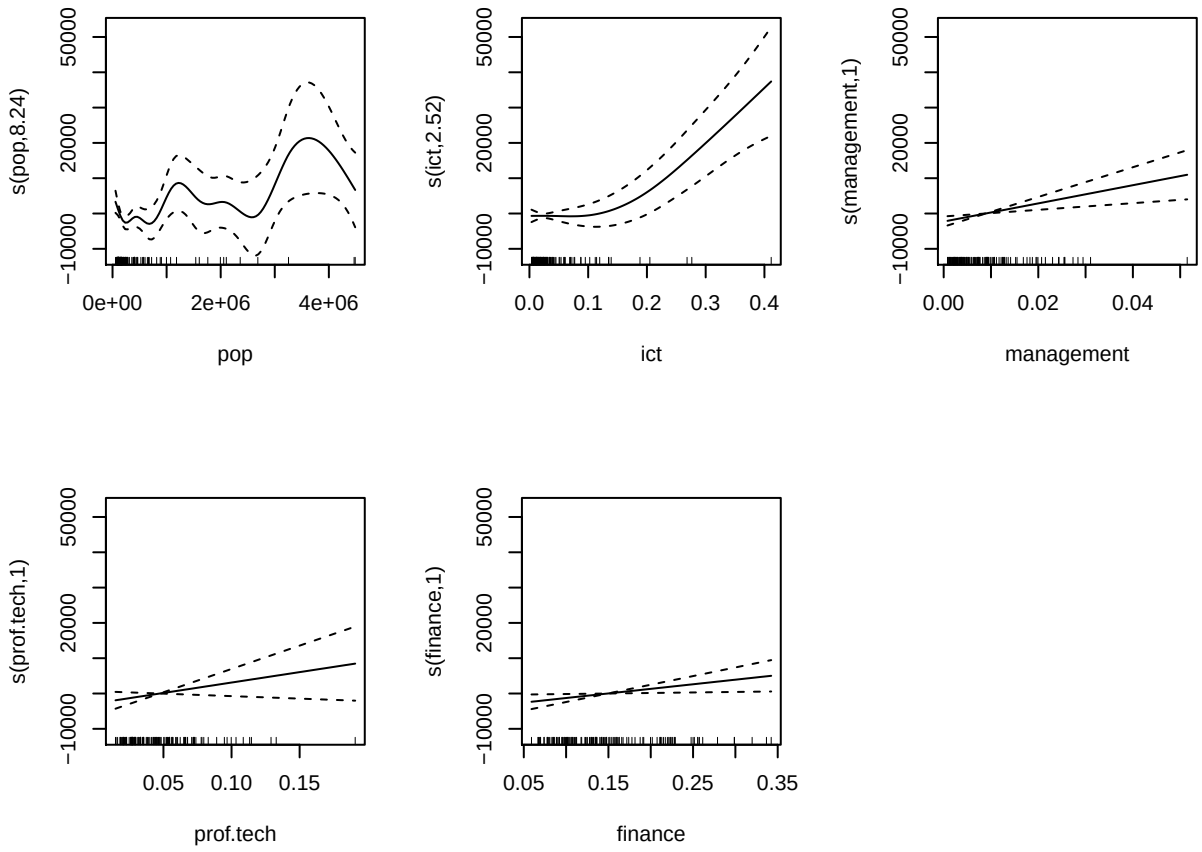
Q2 b)

Under this DAG, `pop` $\not\perp\!\!\!\perp$ `pcgmp`.
However, `pop` $\perp\!\!\!\perp$ `pcgmp` | `ict, management, prof.tech, finance`.

Q2 c) Under this DAG, `ict` $\perp\!\!\!\perp$ `finance` | `pop`. This also holds true for the previous DAG.

Q2 d) Under this DAG, `pop` $\perp\!\!\!\perp$ `pcgmp` | `ict, management, prof.tech, finance`. However, this is not true for the previous DAG.

Q2 e)



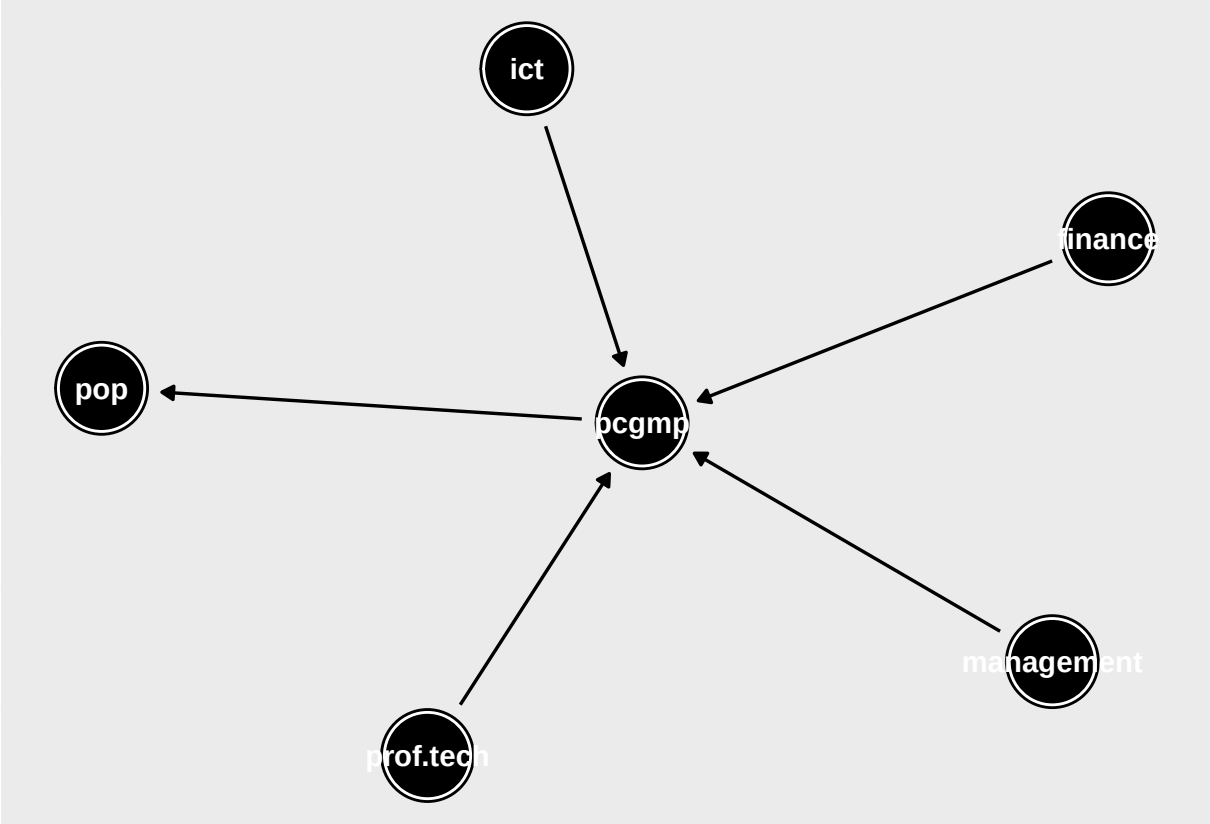
Assuming an additive model with spline smoothing once again, we can see that `pop` does not seem to be independent of `pcgmp` given the rest of 4 variables as its partial response is highly non-horizontal. However, according to this theory and the associated DAG, these variables should be conditionally independent.

Q2 f) This data set may not have all the variables to account for all sources of confounding when it comes to testing statistical independences. For example, the proportions from the four industries do not all add up to 1.0, so there are more industries that contribute to `pcgmp`, as well as many other possible geographical features.

Q2 g) Assuming the DAG above is right and that the relations can be modeled with an additive model with spline smoothing, we can first fit an additive model for `pcgmp` against `pop` and control for `ict, management, prof.tech, finance` with the whole dataset, using spline smoothing. With this model, we then predict the `pcgmp` where we fix `pop` to be the population size of Pittsburgh for the dataset, and then similarly another prediction where we fix `pop` to be double the previous size. The difference in these two predictions would then be the estimated average effect, assuming the DAG was right.

Question 3

Q3 a)

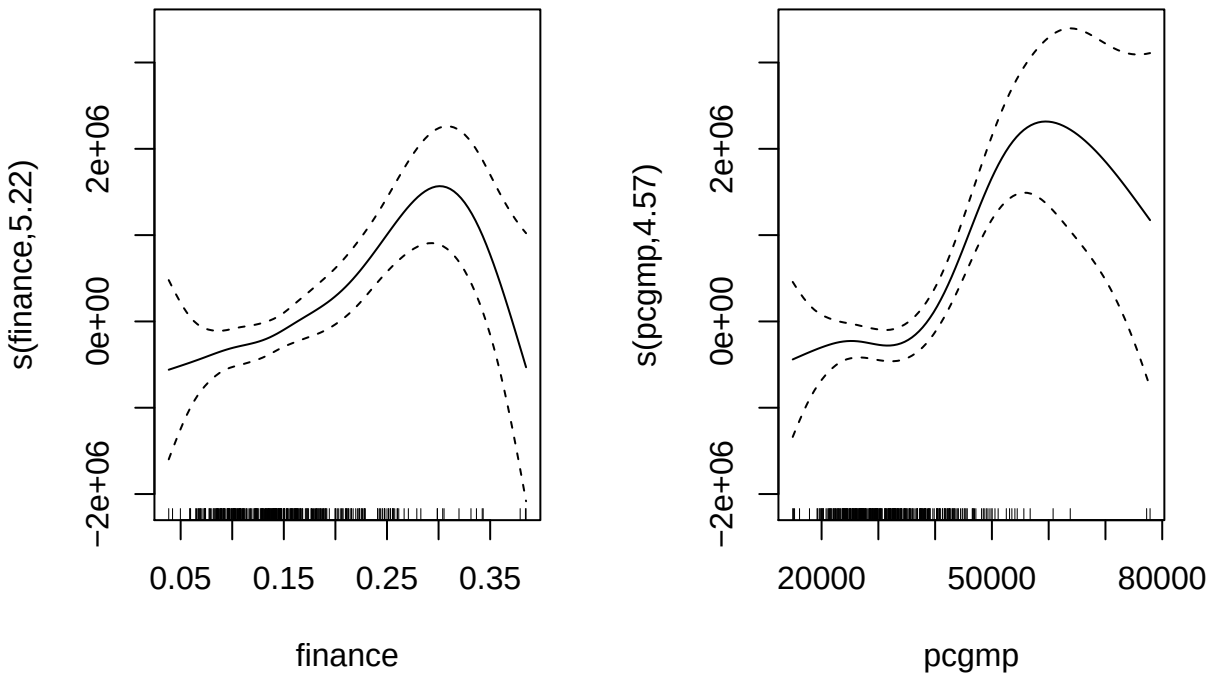


Q3 b) This is not possible as in DAG 1, `pop` is the parent of every variable and those are the only relations. In order to close any path, we would have to condition on `pop`, the middle of any fork. However in this DAG, all 4 industry proportions are the parents of `pcgmp` and so all paths between industries are colliders. Since `pop` is a descendent of `pcgmp`, conditioning on pop makes all the colliders open paths. Thus there is no set of variables that can close all paths in both DAGs.

Q3 c) `ict` $\perp\!\!\!\perp$ `finance` | `pop` in both DAGs.

Q3 d) `pop` $\perp\!\!\!\perp$ `finance` | `pcgmp` in this DAG, but not the previous two.

Q3 e) We fit an additive model with spline smoothing of `pop` against `finance`, controlling for `pcgmp`.

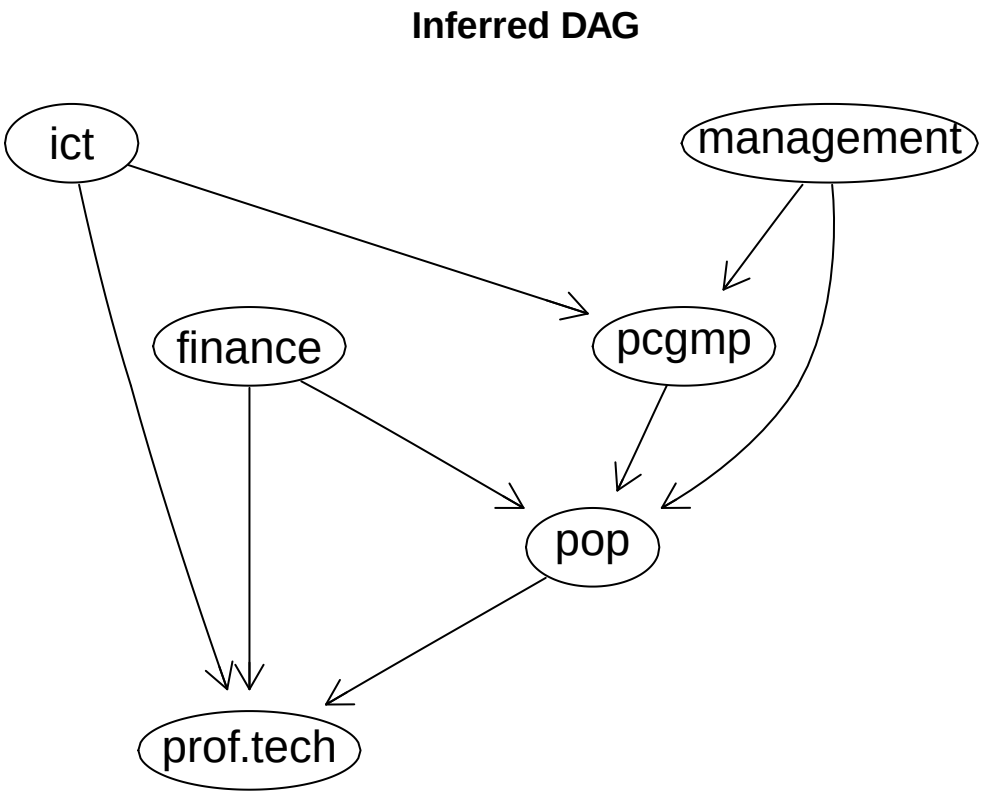


In the partial response of `finance`, we can see that there seems to be a relationship between `pop` and `finance`, where if they were independent controlling for `pcgmp` we would have seen roughly a horizontal response. Thus, the data does not seem to support the conditional independence.

Q3 f) In this model, increasing population would not affect `pcgmp` at all, as when we intervene and increase `pop` the new DAG we get is one with all incoming edges to `pop` removed. Thus, `pop` there would be no path from `pcgmp` and thus they would be independent.

Question 4

Q4 a)



Q4 b) In the inferred DAG, we can see that 3 of the industry proportions, namely `ict`, `finance` and `management` are exogeneous. However, we see that professional tech does not affect anything, but is in fact affected by `ict`, `finance` and `pop`.
We see that `pcgmp` is affected by `management` and `ict` only, and then goes on to affect `pop`.
Finally, `pop` is only affected by `finance`, `pcgmp` and `management`.

Q4 c)

Table 1: `pcgmp` against parents

	Coefficient	SE
(Intercept)	25865	923.35
ict	79954	22295.00
management	306470	53250.00

Table 2: `pop` against parents

	Coefficient	SE
(Intercept)	-1.1584e+06	3.5740e+05
finance	4.9154e+06	1.1993e+06
pcgmp	2.5831e+01	9.4385e+00

	Coefficient	SE
management	1.4046e+07	8.4403e+06

Table 3: `prof.tech` against parents

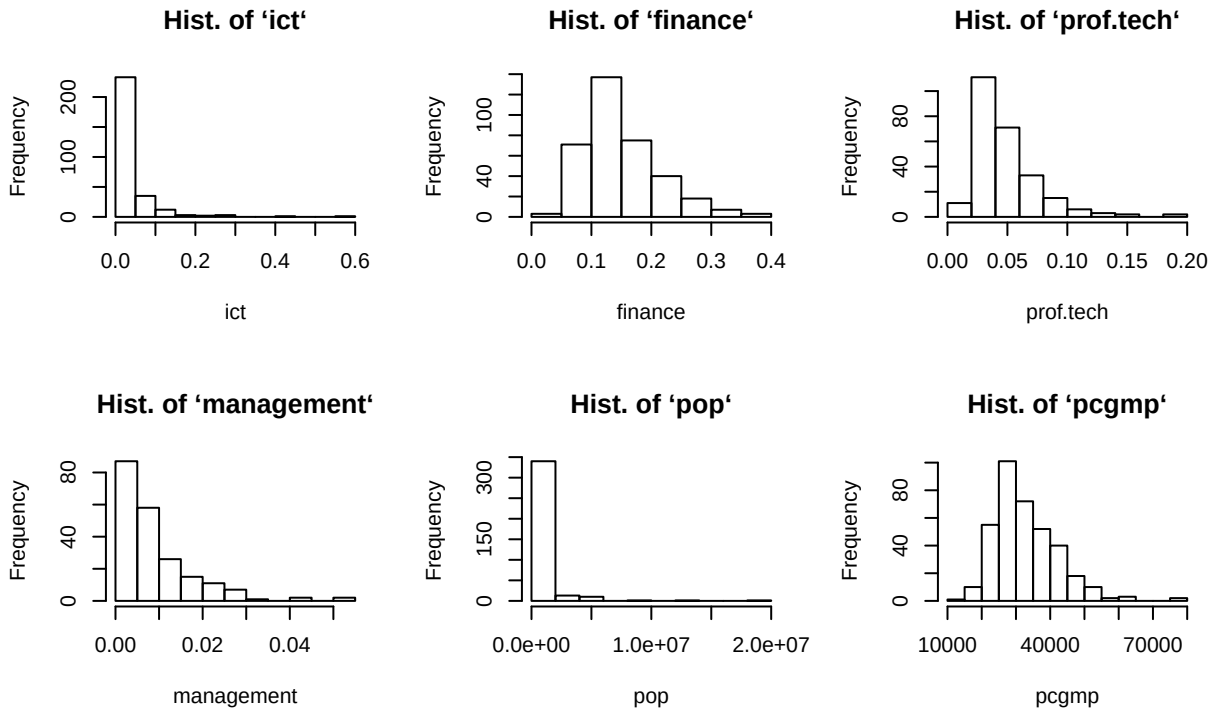
	Coefficient	SE
(Intercept)	0.0207	0.00619
ict	0.1970	0.06040
finance	0.1130	0.03180
pop	0.0000	0.00000

All of the coefficients of parent variables have positive signs, which indicate positive relationships under our linear model, where an increase in the parent variable corresponds to an increase in the variable itself too, controlling for all other parents.

Q4 d) According to the current DAG, changing the population size would have no effect on `pcgmp` as by doing so we remove all incoming edges to `pop`, leaving no open paths from `pop` to `pcgmp`.
However, by increasing the share of `ict` by 10%, we predict this causes `pcgmp` to increase by approximately 7995.4, when controlling for `management`.

Q4 e) The conditional independence test assumse that each variable's marginal distribution as well as pairwise distribution is Gaussian.

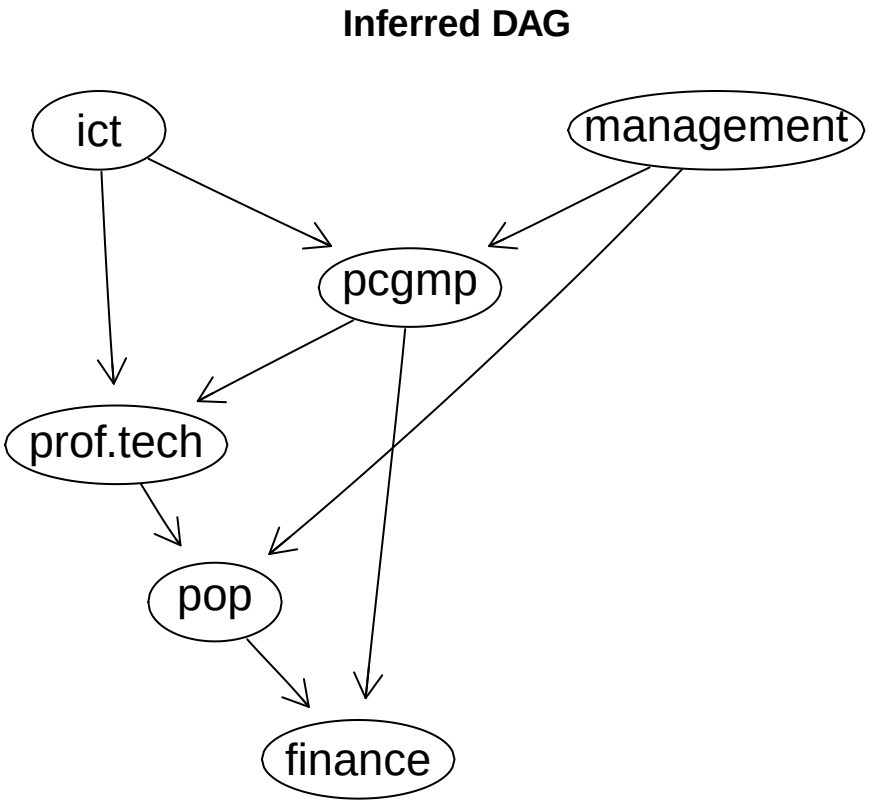
Q4 f)



This assumption is not plausible as from the rough distribution plots above, we can see that most of the variables have skewed distributions that are not approximately normal.

Question 5

Q5 a)

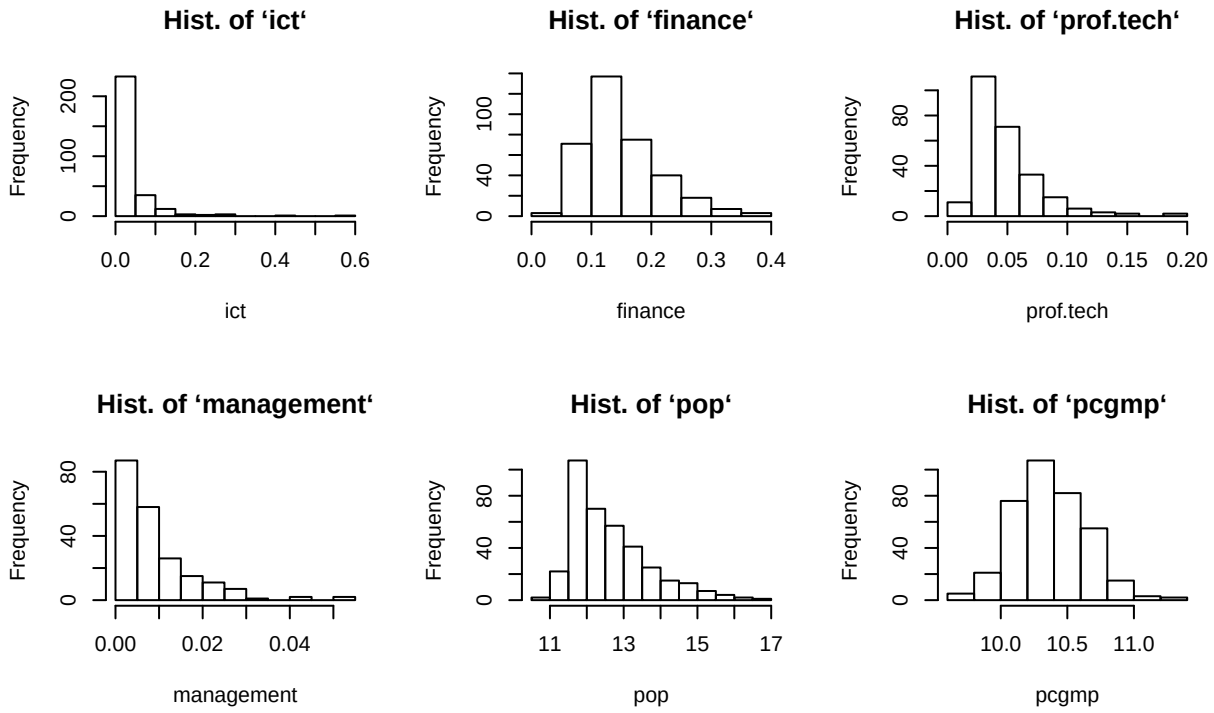


Q5 b) In this new graph, `finance` is no longer exogeneous but rather is affected by `pop` and `pcgmp`. We also see that now `prof.tech` affects `pop`, which is reverse of the previous DAG. Also, `pcgmp` no longer directly affects `pop`.

Q5 c) In the new DAG, when we change the population size, we still will observe no effect on `pcgmp` as this operation will result in no open paths again in the new DAG.

When we increase `ict` by 10%, we now predict that this causes `log(pcgmp)` to increase by around 0.19.

Q5 d)



Afte this transformation, the distributions of `log(pop)` and `log(pcgmp)` are slightly less skewed, but other variables such as `ict`, `prof.tech` and `management` are still very skewed and still do not satisfy the assumption well.

Question 6

We still face the same problem identified in Q2 f), which is that there are probably missing variables from this dataset that have not been observed. This is a crucial assumption about the PC algorithm, which is that there are no hidden or latent variables in our data for the algorithm to work. This cannot really be checked, as the underlying Markov property we use can only test for conditional independence between observed variables. If there are unobserved variables, there is no way to use tests of conditional independence to infer dependence on some hidden variable.