Statistical Coding Example (with project explanations)

Dorsa Massihpour

May 3, 2017

1 DAG Diagrams

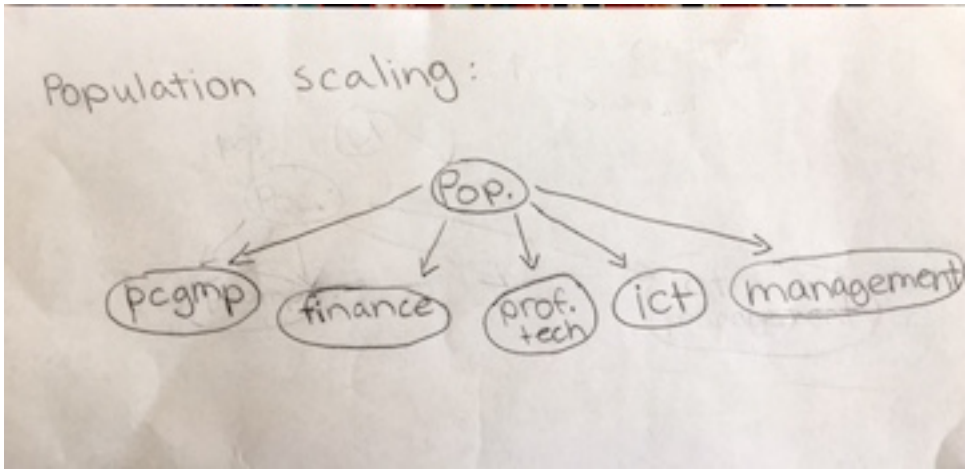
The goal of this assignment is to explore these three theories, all of which predict a positive association between city size and per-capita income, but differ in how they explain it, and in what a city should do to increase or deacrease its income.

1. Population-scaling A simple theory, supported by some of the original researchers on urban scaling, is that increasing population causes higher per-capita income, and also separately causes more of the city's economy to be in high-value industries, such as the four industries contained in thedata set. Population, on this theory, is the common cause of all the other variables in our data set.
2. Central-place theory A different theory is that high-value industries tend to be sited in larger cities to have access to more customers. According to this theory, then, population causes industry shares, and industry shares cause per-capita income, but there is no direct effect of population on income.
3. Exogenous industries Yet a third theory is that different cities acquire different industries more or less by chance (access to supplies or geographic advantages, successful early entrants to the market, good policy, dumb luck, etc.); that some industries pay much better than others; and that people move to places where the income level is high, and are pretty indifferent to everything else about the city.

The variables in this dataset are population of a metropolitan city ('pop'), per-capita gross metropolitan product or GDP ('pcgmp'), and proportion of the economy dedicated to each of the following four sectors: professional and technical ('prof.tech'), information and communications technologies ('ict'), financial ('finance'), and management services ('management').

```
library(knitr)
setwd("~/Desktop/36402")
gmpdata=read.csv("gmp-2006.csv", header=TRUE)
gmpdata=gmpdata[order(gmpdata$pop),] #366 rows
gmpdata.clean=na.omit(gmpdata) #133 rows
gmpdata.clean=gmpdata.clean[order(gmpdata.clean$pop),]
```

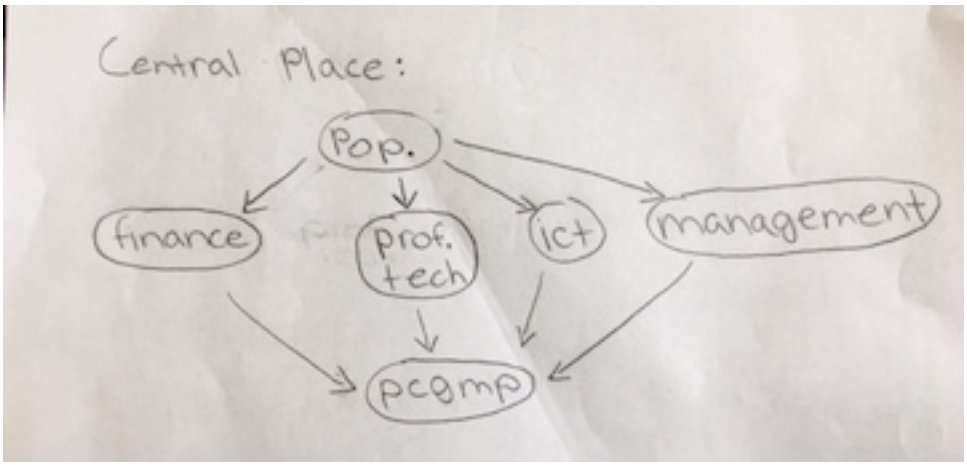
DAG for Population-scaling theory:



```
#Code for DAG above use DiagrammeR package
# library(DiagrammeR)
# mermaid("
#           graph LR
#           P(Pop.)-->G(PCGMP)
#           P-->F(Finance)
#           P-->T(Prof. tech)
#           P-->I(Ict)
#           P-->M(Management)
#           ")
```

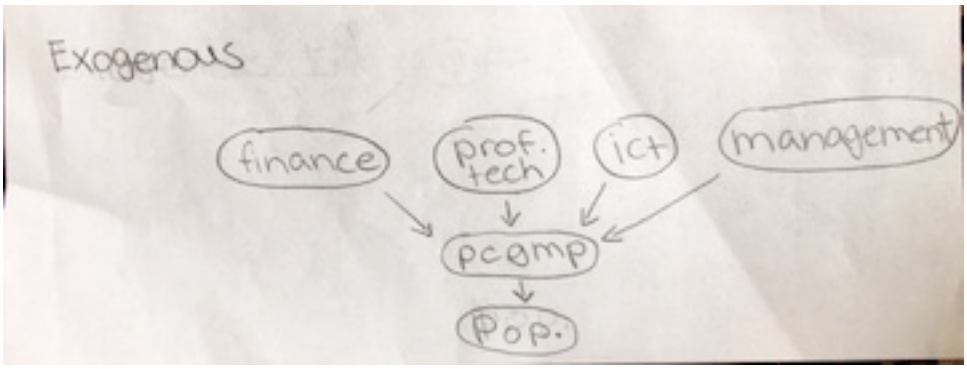
For this DAG, it is also possible that pcgmp, or per-capita income, also causes more of the city's economy to be in high-value industries. This means we could have arrows pointing from pcgmp to finance, prof. tech, ict, and management (separately).

DAG for Central-place theory:




```
#Code for DAG above use DiagrammeR package
# library(DiagrammeR)
# mermaid("
#           graph LR
#           P(Pop.)-->F(Finance)
#           P-->T(Prof. tech)
#           P-->I(Ict)
#           P-->M(Management)
#           F-->G(PCGMP)
#           T-->G
#           I-->G
#           M-->G
#           ")
```

DAG for Exogenous industries:




```
#Code for DAG above use DiagrammeR package
# library(DiagrammeR)
# mermaid("
#           graph LR
#           F(Finance)-->G(PCGMP)
#           T(Prof. tech)-->G
#           I(Ict)-->G
#           M(Management)-->G
#           G-->P(Pop)
#           ")
```

An alternative DAG is possible here if people are drawn to cities not only because there is higher income, but also because it contains high-value industries. Thus there could be arrows from each of finance, prof. tech, ict, and management pointing to population.

EDA

For our EDA, we note that the dataset has 233 rows with at least one column of missing data. Also, the pcgmp, population, ict, management and prof.tech variables are heavily right skewed. The finance variable is roughly normally distributed. In estimating causal effects, we use a cleaned version of the data set where all rows with at least one NA column are removed. This is because the predict function we use for finding causal effects may produce NA if we have any NA values in our dataset. We could have alternatively used the parameter na.rm=TRUE when taking the mean of the predict values.

2 Estimating Causal Effects

For each of the theories, we must look at all the back door paths from population to pcgmp to determine if need to control for any variables in order to estimate the average causal effect of a 10% increase in city population on pcgmp. Similarly, to estimate the average causal effect of increasing the share of a city's economy coming from professional and technical services by 10 percentage points, we must look at all the back door paths from prof.tech to pcgmp. I used the DAG diagrams that I used above to identify paths (instead of the alternatives that I described).

In the Population Scaling Theory, there are no back door paths from population to pcgmp. Thus we do not have to control for anything (or check if any variables satisfy the back door criterion at all) in our regression of population on pcgmp.

There is 1 back door path from prof.tech to pcgmp:
prof.tech<-pop->pcgmp
Population satisfies the back door criterion because it blocks the only back door path from prof.tech to pcgmp and is not a descendant of prof.tech. Thus, when regressing prof.tech on pcgmp to find the average causal effect of increasing the share of a city's economy coming from prof.tech by 10 percentage points, we must also add population as a covariate.

In the Central Place Theory, there are again no back door path from population to pcgmp.
Here are the back door paths from prof.tech to pcgmp:
prof.tech<-pop->finance->pcgmp
prof.tech<-pop->ict->pcgmp
prof.tech<-pop->management->pcgmp
Thus, when regressing prof.tech on pcgmp to find the chosen causal effect, we must include either (population), (ict, finance, and management), or (population, ict, finance, and management) as covariates since none of them are ancestors of prof.tech (and thus satisfy the back door criterion). We will choose the latter set.

In the Exogenous Industries Theory, there is 1 back door path from population to pcgmp:
pop<-pcgmp
We do not control for anything when finding the desired causal effect for the same reason described in the previous theory. Since there are no intermediary variables, we do not control for anything when modeling population on pcgmp to find the desired average causal effect (pop and pcgmp are dependent without controlling for anything).
There is also 1 back door path from prof.tech to pcgmp: prof.tech<-pcgmp
But again, we do not control for anything in our regression of prof.tech on pcgmp because these two variables are unconditionally dependent.

Model Selection

In order to select an accurate model, I fitted both a GAM and a linear regression model for each of the two desired causal effects, for each of the three theories. (Note that I could have picked any nonparametric method, but chose a GAM because it is computationally faster than most methods, such as a kernel regression) However, for the first coveted causal effect (population on pcgmp), we noted above that all three theories yield controlling for no other covariates. Thus we first fit a linear regression model with population as the predictor and pcgmp as the response variable. To check if our data satisfy the necessary conditions to use a linear regression, we first look at a plot of population vs. residuals:

Figure 1: Plots Violating Linear Regression Assumptions

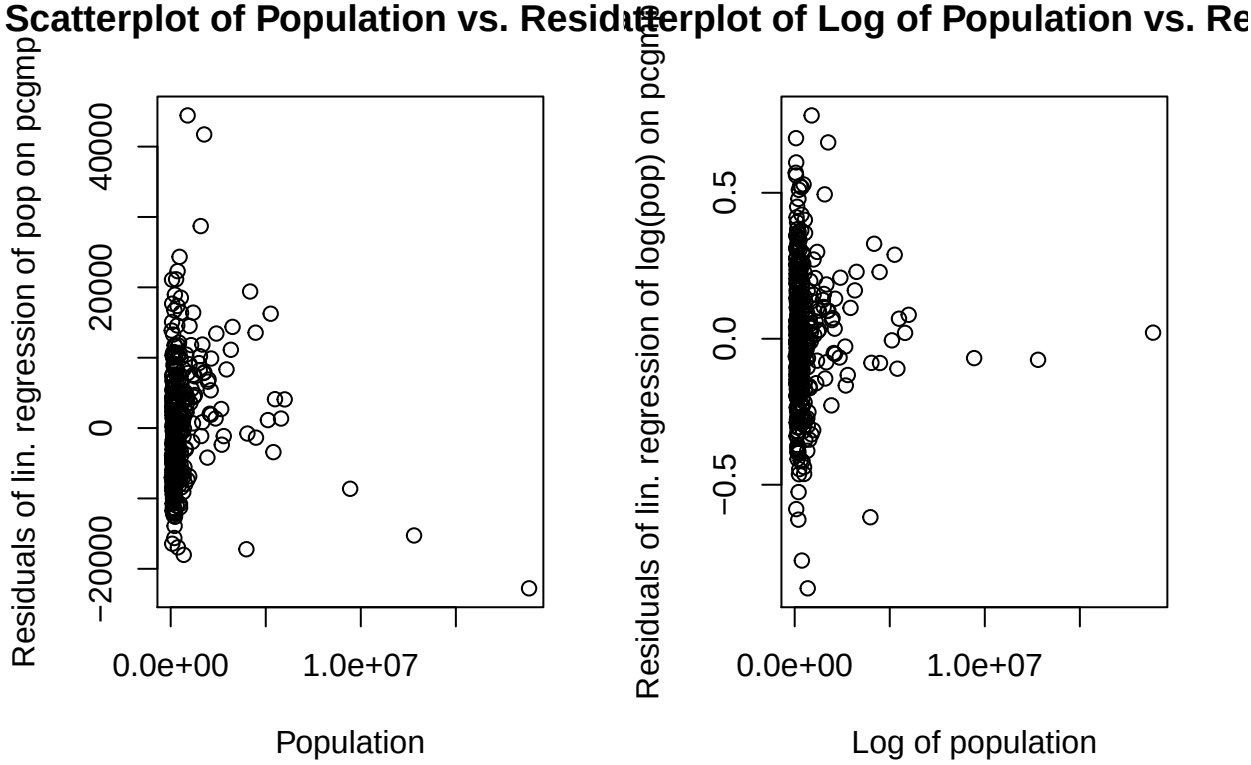
```
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-12. For overview type 'help("mgcv-package")'.
```

```
par(mfrow=c(1,2))
#Effect of pop on pcgmp for all three theories:
#Linear model and diagnostics
model1apop<-lm(pcgmp~pop,data=gmpdata)
plot(gmpdata$pop, model1apop$residuals, main="Scatterplot of Population vs. Residuals", xlab="Populatio
```

```
model1apop.2<-lm(log(pcgmp)~log(pop), data=gmpdata)
plot(gmpdata$pop, model1apop.2$residuals, main="Scatterplot of Log of Population vs. Residuals", xlab="
```



As we can see in Figure 1, there is a clear pattern in this plot, as there are more data points at smaller populations, and less with lower overall values at higher populations. Even when we transform the highly skewed pcgmp and population using the log function in the regression, this plot does not change much. In order for there to be a linear relationship between population and pcgmp, there should have been a random scatter of points in this plot. Since this first assumption is not satisfied, we do not need to check the other assumptions, mainly constant variances, normal distribution of errors, and irreducible errors with mean 0. Thus, we fit a GAM model with the same response except now the predictor will be smoothed. (From here on, all described GAM models will have all predictors smoothed). Although 4 of the variables are right skewed , GAM models have relaxed assumptions about the distribution of predictors, response variables and their errors. The mean squared prediction error (ubre score) for the first GAM model is lower (60,444,912 vs. 62,137,268), thus we proceed with our calculations using this model.

To find the causal effect of a increasing the share of a city's economy coming from prof.tech services by 10 percentage points, we could follow the same process described above for each of the three theories. However, when we run into a case where all the linear regression assumptions are satisfied, it will be difficult to decide between the linear regression and GAM solely on that basis. Thus, for the second causal effect I decided to use a GAM instead of a linear regression model for three main reasons: its lack of asumptions on data allow it to uncover the true patterns present in our dataset, contain less bias, and control smoothness to prevent overfitting. The only disadvantage is that it converges slightly more slowly compared to linear models.

For the first theory, we fit 2 GAM models with pcgmp as the response variable: one with population and prof.tech as predictors, and another with the log of these variables as predictors. The second model has a lower mean squared prediction error, so we proceed with this one (57283083 vs. 58884319). For the second theory, we again fit 2 GAM models with pcgmp as the response variable: one with population, ict, finance,

management, and prof.tech as predictors, and another with all variables except finance transformed with the log function. The second model again has a lower mean squared prediction error (46278565 vs. 48777908), so we proceed with this one. For the third theory, we once again fit 2 GAM models with pcgmp as the response variable : one with prof.tech as the predictor, and another with the log of prof.tech as the predictor. The first model has a lower mean squared prediction error, so we proceed with this one (59986328 vs. 65718665).

```
#Gam model and diagnostics
model1bpop<-gam(pcgmp~s(pop), data=gmpdata)
model1bpop$gcv.ubre
```

```
##   GCV.Cp
## 62137268
```

```
model1bpop.2<-gam(pcgmp~s(log(pop)), data=gmpdata)
model1bpop.2$gcv.ubre
```

```
##   GCV.Cp
## 60444912
```

```
df1a<-gmpdata
df1a$pop<-df1a$pop+.1

#Effect of prof.tech for Pop. Scaling theory:
model2aprof.tech<-gam(pcgmp~s((pop))+s((prof.tech)), data=gmpdata.clean)
model2aprof.tech$gcv.ubre
```

```
##   GCV.Cp
## 58884319
```

```
model2bprof.tech<-gam(pcgmp~s(log(pop))+s(log(prof.tech)), data=gmpdata.clean)
model2bprof.tech$gcv.ubre
```

```
##   GCV.Cp
## 57283083
```

```
df1b<-gmpdata.clean
df1b$prof.tech<-df1b$prof.tech+.1

#Effect of prof.tech for Central place theory:
model3aprof.tech<-gam(pcgmp~s(pop)+s(ict)+s(finance)+s(management)+s(prof.tech),data=gmpdata.clean)
model3aprof.tech$gcv.ubre
```

```
##   GCV.Cp
## 48777908
```

```
model3bprof.tech<-gam(pcgmp~s(log(pop))+s(log(ict))+s(finance)+s(log(management))+s(log(prof.tech)),data
model3bprof.tech$gcv.ubre
```

```
##   GCV.Cp
## 46278565
```

```r
#Effect of prof.tech Exogenous industries:
model4aprof.tech<-gam(pcgmp~s(prof.tech), data=gmpdata.clean)
model4aprof.tech$gcv.ubre
```

```
##   GCV.Cp
## 59986328
```

```r
model4bprof.tech<-gam(pcgmp~s(log(prof.tech)), data=gmpdata.clean)
model4bprof.tech$gcv.ubre
```

```
##   GCV.Cp
## 65718665
```

```r
#Bootstrapping: this code was taken from Professor Shalizi's textbook and modified
resample <- function(x) {
sample(x, size = length(x), replace = TRUE)
}
resample.data.frame <- function(data) {
sample.rows <- resample(1:nrow(data))
return(data[sample.rows, ])
}
resample.gmpdata <- function() { resample.data.frame(gmpdata.clean) }

est.lm1 <- function(data) {
  model1bpop.2<-gam(pcgmp~s(log(pop)), data=data)
  estimator<-mean(predict(model1bpop.2,  newdata=df1a))-mean(predict(model1bpop.2,newdata=data))
    return(estimator)
}

est.lm2 <- function(data) {
  model2bprof.tech<-gam(pcgmp~s(log(pop))+s(log(prof.tech)), data=data)
    estimator<-mean(predict(model2bprof.tech,  newdata=df1b))-mean(predict(model2bprof.tech,newdata=data
    return(estimator)
}

est.lm3 <- function(data) {
    model3bprof.tech<-gam(pcgmp~s(log(pop))+s(log(ict))+s(finance)+s(log(management))+s(log(prof.tech))
    estimator<-mean(predict(model3bprof.tech,  newdata=df1b))-mean(predict(model3bprof.tech,newdata=data
    return(estimator)
}

est.lm4 <- function(data) {
    model4aprof.tech<-gam(pcgmp~s(prof.tech), data=data)
    estimator<-mean(predict(model4aprof.tech,  newdata=df1b))-mean(predict(model4aprof.tech,newdata=data
    return(estimator)
}

rboot <- function(statistic, simulator, B) {
tboots <- replicate(B, statistic(simulator()))
if (is.null(dim(tboots))) {
tboots <- array(tboots, dim = c(1, B))
}
```

```
return(tboots)
}

bootstrap <- function(tboots, summarizer, ...) {
summaries <- apply(tboots, 1, summarizer, ...)
return(t(summaries))
}

bootstrap.se <- function(statistic, simulator, B) {
bootstrap(rboot(statistic, simulator, B), summarizer = sd)
}

bootstraperrors1<-bootstrap.se(est.lm1, resample.gmpdata, B=1e4)
bootstraperrors2<-bootstrap.se(est.lm2, resample.gmpdata, B=1e4)
bootstraperrors3<-bootstrap.se(est.lm3, resample.gmpdata, B=1e4)
bootstraperrors4<-bootstrap.se(est.lm4, resample.gmpdata, B=1e4)
```

Table 1: Estimated Average Causal Effects and Uncertainty Estimates

```
results1=c(round(mean(predict(model1bpop.2, newdata=df1a))-mean(predict(model1bpop.2,newdata=gmpdata)),
results2=c(bootstraperrors1, bootstraperrors2, bootstraperrors3, bootstraperrors4)
main=cbind(data.frame(results1), data.frame(results2))
colnames(main)=c("Estimate", "Bootstrap Std. Error")
rownames(main)=c("Theories 1-3, effect of 10% pop. increase on pcgmp", "Theory 1, effect of 10% increas
kable(main)
```

	Estimate	Bootstrap Std. Error
Theories 1-3, effect of 10% pop. increase on pcgmp	-0.001	687.7069
Theory 1, effect of 10% increase in share of cities prof/tech services on pcgmp	14369.139	11982.2672
Theory 2, effect of 10% increase in share of cities prof/tech services on pcgmp	3614.771	9113.1594
Theory 3, effect of 10% increase in share of cities prof/tech services on pcgmp	31143.606	25724.8351

Justification of Regressions:

For the Exogenous Industries Theory, the chosen GAM gives a valid estimates of the desired causal effect because prof.tech is exogenous (the fact that a city's economy tends to be in this high value industry occurs by chance). This randomization means that prof.tech has no parents, and thus no back door paths that link it to anything. Thus the back door criterion is satisfied if we condition on nothing.

For the first two thoeries our GAM models give us valid estimates of the causal effects: we are either using the back door criterion to get rid of any confounding effects on pcgmp by variables other than the effecting variables we chose, or their are no back door paths and thus no confounders, or the back door paths are directly from our predictors to pcgmp. For example, we mentioned that there are 4 back door paths from prof.tech to pcgmp in the 2nd theory. By controlling for either population or the 3 other industries (since neither are descendants of prof.tech), we can isolate the effect of prof.tech on pcgmp. Thus, if we change prof.tech and it has an effect on population, which affects finance and in turn, pcgmp, this (back door) effect on pcgmp will not be included in our measurement. Similarly, there is 1 back door path from prof.tech to pcgmp in the first theory, which is blocked by population. This means that when we change prof.tech, it may be correlated with a change in population that adds to the final effect on pcgmp. Conditioning on population (since it is not a descendant of prof.tech) will give us only the direct effect of prof.tech on pcgmp.

3. Finding and Testing Conditional Independence Relations

For each theory, we now look at conditional independences relations that hold in one theory but not the other two theories.
First let's consider all the paths from pcgmp to pop in each theory (rewritten from above).
Theory 1:
pcgmp<-pop

Theory 2:
pcgmp<-prof.tech<-pop
pcgmp<-finance<-pop
pcgmp<-ict<-pop
pcgmp<-management<-pop

Theory 3:
pcgmp->pop

In theory 2, pcgmp and pop are independent conditioning on finance, ict, management, and prof.tech. However this is not the case in theory 1 or 3, as no matter what variables we condition on, pcgmp and pop will be dependent.

Next let's consider all the paths from pop to finance in each theory.
Theory 1:
finance<-pop

Theory 2:
pop->finance
pop->prof.tech->pcgmp<-finance
pop->ict->pcgmp<-finance
pop->management->pcgmp<-finance

Theory 3:
pop<-pcgmp<-finance

In theory 3, pop and finance are independent given pcgmp, since pcgmp blocks the only path between pop and finance. However, in theory 2, there is a collider on pcgmp, so conditioning on this would unblock the path from finance to pop, making them conditionally dependent on pcgmp. In theory 1, no matter what we condition on, finance and pop will be dependent.

Finally let's consider all the paths from finance to pcgmp. Theory 1:
finance<-pop->pcgmp
Theory 2:
finance->pcgmp
finance<-pop->prof.tech->pcgmp
finance<-pop->ict->pcgmp
finance<-pop->management->pcgmp

Theory 3:
finance->pcgmp

In theory 1, finance and pcgmp are independent given population, since population blocks the only path between these two variables. In theory 2, there is a direct path from finance to pcgmp, so these two will always be dependent, no matter what is conditioned on. In theory 3, the same logic holds.

4. Testing Conditional Independence relations

To test whether the conditional independence relations from problem 3 hold, we will use the following strategy which is taken from the solutions to the extra credit problem in homework 10: given the variables X, Y, and Z, if we want to tests whether X ad Y are independent conditional on Z, we linearly regress X on Z and Y on Z, recording two sets of residuals Rx and Ry. The partial correlation between X and Y given Z is the correlation between Rx and Ry. To test the hypothesis that Rx and Ry are uncorrelated, we look at the p-value for the slope in a regession of Rx on Ry or Ry on Rx. If we find a partial correlation, it implies conditional dependence because it would not be possible to get a coefficient of 0 (or very close to 0) when regressing the residuals on each other if all values of X and Y affect each other. However, if we do not find any partial correlation, it does not imply conditional independence, except if we have a gaussian distribution of the values. This is because it could be possible for a different function of the X values, such as X^2, to have a correlation with Y (both given Z), but the current untransformed X's may produce an slope coefficient of 0 when regressing their residuals on those of Y. This is an example of where our method will go wrong in missing dependencies between X and Y. Additionally, no or very little partial correlation could mean that many X or Y values are missing, not that they are independent.

Using the method we described above, we test for theory 1 whether finance and pcgmp are independent given population.

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.2.5
```

```
##
## Attaching package: 'broom'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     bootstrap
```

```
lm.DAG1.pcgmp=lm(pcgmp~pop, data=gmpdata.clean)
lm.DAG1.finance=lm(finance~pop, data=gmpdata.clean)
lm.DAG1.test=lm(lm.DAG1.pcgmp$residuals ~ lm.DAG1.finance$residuals)
signif(summary(lm.DAG1.test)$coefficients, 2)
```

```
##                            Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                -1.5e-12        720 -2.1e-15    1.000
## lm.DAG1.finance$residuals  2.5e+04      13000  2.0e+00    0.052
```

```
est.lm5 <- function(data) {
    lm.DAG1.pcgmp=lm(pcgmp~pop, data=data)
    lm.DAG1.finance=lm(finance~pop, data=data)
    lm.DAG1.test=lm(lm.DAG1.pcgmp$residuals ~ lm.DAG1.finance$residuals)
    estimator<-glance(lm.DAG1.test)$p.value
    return(estimator)
}
```

Since the p-value is .052, we cannot reject the null hypothesis that finance and pcgmp are independent given population.
For theory 2, we test whether pcgmp and pop are independent conditioning on finance, ict, management, and prof.tech based on our data.

```
lm.DAG2.pcgmp=lm(pcgmp~finance, data=gmpdata.clean)
lm.DAG2.pop=lm(pop~finance, data=gmpdata.clean)
lm.DAG2.test=lm(lm.DAG2.pcgmp$residuals ~ lm.DAG2.pop$residuals)
signif(summary(lm.DAG2.test)$coefficients, 2)
```

```
##                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           -1.1e-12    7.2e+02 -1.5e-15    1e+00
## lm.DAG2.pop$residuals  5.0e-03    1.1e-03  4.6e+00    1e-05
```

```
est.lm6 <- function(data) {
    lm.DAG2.pcgmp=lm(pcgmp~finance, data=data)
    lm.DAG2.pop=lm(pop~finance, data=data)
    lm.DAG2.test=lm(lm.DAG2.pcgmp$residuals ~ lm.DAG2.pop$residuals)
    estimator<-glance(lm.DAG2.test)$p.value
    return(estimator)
}
```

The p-value of .00001 is less than an alpha value of .05, so we reject the null hypothesis of independence.
We test for theory 3, whether pop and finance are independent controlling for pcgmp, based on our data. The p-value of .00018 is less than an alpha value of .05, so we again rejec the null hypothesis of independence.

```
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.6e-11      54000 3.1e-16  1.00000
## lm.DAG3.finance$residuals  3.5e+06     920000 3.9e+00  0.00018
```

Table 2: P-values and Bootstrapped Std. Errors of Conditional Independence Testing

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.2.5
```

```
bootstraperrors5<-bootstrap.se(est.lm5, resample.gmpdata, B=1e4)
bootstraperrors6<-bootstrap.se(est.lm6, resample.gmpdata, B=1e4)
bootstraperrors7<-bootstrap.se(est.lm7, resample.gmpdata, B=1e4)

results1=c(glance(lm.DAG1.test)$p.value, glance(lm.DAG2.test)$p.value, glance(lm.DAG3.test)$p.value)
results2=c(bootstraperrors5, bootstraperrors6,bootstraperrors7)
```

```
main=cbind(data.frame(results1), data.frame(results2))
colnames(main)=c("P-value", "Bootstrapped P-values")
rownames(main)=c("Indep. of finance and pcgmp given pop", "Indep. pcgmp and pop given finance", "Indep.
kable(main)
```

	P-value	Bootstrapped P-values
Indep. of finance and pcgmp given pop	0.0516084	0.2355676
Indep. pcgmp and pop given finance	0.0000102	0.0143175
Indep. of pop and finance given pcgmp	0.0001783	0.0189130

5. How to Increase Per-capita Income

In order to determine whether it is better for a city who wants to increase its pcgmp to favor population growth or increaseing the share of its economy devoted to prof.tech, we refer back to table 1. A log of 10% population increase actually leads to a negative effect on pcgmp, although with a relatively high certainty compared to the other three estimates. Since for all three theories, a 10% increase in the economy's share of prof.tech leads to large positive increases in pcgmp, the city should choose to increase its share of prof.tech.