STAT 673 Project
A Statistic Model Comparison: OLS, IV estimation, and LiNGAM using 1966-76 US School Survey Data

Akihiko Mori, 230142775

April 10, 2022

1 Introduction

There are two main types of statistical causal inference: one is forward-looking research, and the other is backward-looking research. The former is a method of statistically estimating the causal effect of an outcome from interventions on a research subject, such as in a randomized controlled trial, based on the experimental and observational data of the subject (Chalmers et al., 1981). The former method has a long history and is actively used in the natural sciences, especially in agriculture and a related field (Mann, 1949), although the method of randomized controlled trials has been popular in Economics recently (Banerjee et al., 2016). One of the aims of the experiment is to estimate the outcome by introducing one or more independent variables, also called "factors," through prior operations by the researcher. The planning involves selecting and specifying, with appropriate objectives, the independent, dependent, and control variables that must be held constant to prevent external factors from influencing the results, planning the provision of the experiment under statistically optimal conditions given the constraints of available resources, and this includes assessing validity.

The latter, backward-looking research, is a method where the intervention has already been applied to the research subject, and the response and factors data are observed to make statistical causal inferences (Stock and Trebbi, 2003). In general, backward-looking research is the same as statistical inference, which is used to determine the difference between variations in the original data that are random variations or the effect of a well-specified causal mechanism (DiNardo, 2010). This data analysis is used in social science, epidemiology, and computer science, where a proper experimental design is difficult.

There are many tools used for statistical causal inference. Two main analytical streams for the causal inference can be applied to estimate intervention effects. Some models can be conditioned on covariates such that conditional exchangeability is satisfied and models that cannot be conditioned on covariates such that conditional exchangeability is satisfied. The former include covariate selection in the model, estimation of average treatment effects, and propensity score matching analysis (Lousdal, 2018). The latter typically includes operating variable methods, regression discontinuity designs, and difference-in-differences methods (Hernán and Robins, 2010). Others include covariance structure analysis and, more recently, LiNGAM, an application of machine learning.

2 Research Question

This paper compares simple and multiple linear regressions, instrumental variable (IV) estimation, and linear non-gaussian acyclic model (LiNGAM) methods in estimating the causal relationship between income and years of schooling for each person using a survey data US education and income, using R from each model. The fundamental hypothesis follows Card's assumption that income in 1976 is determined by academic years (Card, 1993). We emphasize that our interest is only the relationship of schooling to returns; therefore, general residual analysis and the hypothesis test of interest of parameter is conducted for the linear regression model but not focus on the F-test, goodness of fit for each linear model.

3 Method

Data Preparation

The data set used for this comparison is The National Longitudinal Survey of Older and Younger Men (NLSM), which is a survey data set used by David Card [1]. Card uses the young cohort who were in their late teens and early twenties in 1966. This data set is US 14-24 aged men cohort between 1966 and 1981, which contains over 5,000 observations with gender, duration of schooling and education, income, and proximity of the address to the school.

```
# Data set up
x <- read.csv("nls.csv",as.is=TRUE)

## Error in file(file, "rt"):  cannot open the connection

x$wage76 <- as.numeric(x$wage76)

## Error in eval(expr, envir, enclos):  object 'x' not found

x$lwage76 <- as.numeric(x$lwage76)

## Error in eval(expr, envir, enclos):  object 'x' not found

x1 <- x[is.na(x$lwage76)==0,]

## Error in eval(expr, envir, enclos):  object 'x' not found

x1$exp <- x1$age76 - x1$ed76 - 6 # working years after school

## Error in eval(expr, envir, enclos):  object 'x1' not found

x1$exp2 <- (x1$exp^2)/100 # experienced squared divided by 100

## Error in eval(expr, envir, enclos):  object 'x1' not found

x1$age2 <- x1$age76^2

## Error in eval(expr, envir, enclos):  object 'x1' not found
```

[1] The data set is online and can be downloaded from his website `http://davidcard.berkeley.edu/data_sets$`

OLS Model

Card (1993) posits that income in 1976 is determined by the individual's years of education:

$$\text{Income}_i = \alpha + \beta\text{Education}_i + \text{Unobserved}_i$$

This relationship is that income in 1976 for individual i is determined by their education level and other unobserved characteristics such as the unemployment rate in the place they live. Then we would like to estimate β.

Simple Plotting

```
# simple plot of linear regression
lm1 <- lm(lwage76 ~ ed76, data=x1)

                ## Error in is.data.frame(data):  object 'x1' not found

plot(x1$ed76,x1$lwage76, xlab="Years of Education",
     ylab="Log Wages (1976)")

## Error in h(simpleError(msg, call)):  error in evaluating the argument 'x' in selecting
                  a method for function 'plot':  object 'x1' not found

abline(a=lm1$coefficients[1],b=lm1$coefficients[2],lwd=3)

  ## Error in abline(a = lm1$coefficients[1], b = lm1$coefficients[2], lwd = 3):  object
                                      'lm1' not found
```

Figure 1: Plot of log wage and years of schooling in 1976 with OLS

Figure 1 is a simple plot of the relationship between log wages in 1976 and the years of education. We can see a positive relationship. At glance, people who do not graduate from high school (less than 12 years of education) earn less on average than those who attend university (more than 12 years of education). There is much overlap between the distribution.

```
# display the result of simple linear regression
stargazer(lm1, header=FALSE, type='latex', label="table1",
          single.row = TRUE,column.sep.width = "1pt",omit.stat=c("f", "ser"))

## Error in .stargazer.wrap(..., type = type, title = title, style = style, :  object 'lm1'
not found
```

The estimates of OLS can be used for the average effect of schooling on income. Table **??** shows the OLS estimate. The coefficient estimate of the relationship between years of schooling and log wages is 0.052. The coefficient can be interpreted as the percentage increases in wages associated with one additional year of increase in schooling. Since this value is logged, the predicted percentage change in wages, measured at the mean of wages, is 5.4%.

```
exp(log(mean(x1$wage76))+lm1$coefficients[2])/mean(x1$wage76)

## Error in mean(x1$wage76):  object 'x1' not found
```

This converted estimate suggests a high return to schooling in their life. However, the model makes a number of important assumptions about how the data is generated.

Check Assumptions

In general, there are four assumptions in linear regression model about the data.

(1) **Linearity of the data**

The relationship between the predictor (x) and the outcome (y) is assumed to be linear.

(2) **Normality of residuals**

The residual errors are assumed to be normally distributed.

(3) **homoscedasticity**

The residuals are assumed to have a constant variance (Homogeneity of residuals variance).

(4) **Independence of residuals error terms**

In these assumptions, we should check whether or not these assumptions hold true.

Potential problems include:

(1) *Non-linearity* of the outcome - predictor relationships

(2) *Heteroscedasticity*: Non-constant variance of error terms.

(3) *Presence of influential values* in the data that can be:

Outliers: extreme values in the outcome (y) variable

High-leverage points: extreme values in the predictors (x) variable

All these assumptions and potential problems can be checked by producing some diagnostic plots visualizing the residual errors.

```
# Check the assumption for the linear regression
par(mfrow = c(2, 2))
plot(lm1)

## Error in h(simpleError(msg, call)):  error in evaluating the argument 'x' in selecting a
method for function 'plot':  object 'lm1' not found
```

(1) The plot of the residuals and fitted values indicate no pattern in the residual plot. This suggests that we can assume a linear relationship between the predictors and the outcome variables.

(2.1) The QQ plot of residuals is used to check the normality assumption visually. Since the normal probability plot of residuals approximately follows a straight line, in this case, all the points fall approximately along this reference line, so we can assume normality.

(2.2) The plot of scale-location and standardized residuals shows whether residuals are spread equally along with the ranges of predictors. It's good if a horizontal line with equally spread points exists. In our case, this is the case: it can be seen that the variability (variances) of the residual points are constant with the value of the fitted outcome variable, suggesting constant variances in the residuals errors.

```
# Outliers check
x1 %>% group_by(ed76) %>% identify_outliers(lwage76)

## Error in group_by(., ed76):  object 'x1' not found
```

(3.1) An outlier is a point that has an extreme outcome variable value. The presence of outliers may affect the interpretation of the model because it increases the RSE. As the table shows, several outliers exceed three standard deviations. We should examine these outliers.

(3.2) If it has extreme predictor x values, a data point has high leverage. This can be detected by examining the leverage statistic or the hat-value. In our case, there is no high leverage point in the data. That is, all data points have a leverage statistic below $2(p + 1)/n = 4/3000 = 0.0013$.

Multiple Regression

We are interested in the effect of schooling on income; however, we would like to account for how other variables may also influence income. In general, we can guess that work experience is also an essential determinant of income. In addition to work experience, race and the region are also possible key factors to current income. Therefore,

$$\text{Income}_i = \alpha + \beta\text{Education}_i + \gamma\text{Work}_i + \cdots + \text{Unobserved}_i.$$

This equation represents that income in 1976 for individual i is determined by their education level, their experience, and other characteristics such as race, where the individual grew up and where the one is currently living. What we would like to estimate is β.

To build a regression model that includes all of the predictor variables that are statistically significantly related to the response variable, we would like to do a stepwise regression: forward, backward and both.

```
# data set for multilinear regression
mx1<-data.frame(ed76=x1$ed76,exp=x1$exp,exp2=x1$exp2,black=x1$black,
                reg76r=x1$reg76r,smsa76r=x1$smsa76r,smsa66r=x1$smsa66r,
                reg662=x1$reg662,reg663=x1$reg663,reg664=x1$reg664,
                reg665=x1$reg665,reg666=x1$reg666,reg667=x1$reg667,reg668=x1$reg668,
                reg669=x1$reg669,lwage76=x1$lwage76)

## Error in data.frame(ed76 = x1$ed76, exp = x1$exp, exp2 = x1$exp2, black = x1$black, :  object
'x1' not found

#define intercept-only model
lm.0<-lm(lwage76 ~ 1, data=mx1)

## Error in is.data.frame(data):  object 'mx1' not found

#define model with all predictors
lm.all<-lm(lwage76 ~ ., data=mx1)

## Error in is.data.frame(data):  object 'mx1' not found

#perform forward stepwise regression
lm.forw <- step(lm.0, direction='forward',
                scope=formula(lm.all), trace=1,steps = 100)

## Error in terms(object):  object 'lm.0' not found
```

```
#view results of backward stepwise regression
lm.forw$anova

## Error in eval(expr, envir, enclos):  object 'lm.forw' not found

plot(x = 1:10, y = lm.forw$anova$AIC,
     main = "AIC and variable selection",
```

```
     xlim = c(1,10), "l")
```

```
## Error in h(simpleError(msg, call)):  error in evaluating the argument 'y' in selecting a
method for function 'plot':  object 'lm.forw' not found
```

Forward stepwise selection is a variable selection method that: Begins with a model that contains no variables, only the intercept. Then starts adding the most significant variables one after the other. This step function takes AIC as a criterion for choosing the best model.

```
#perform backword stepwise regression
lm.back <- step(lm.all, direction='backward',
                scope=formula(lm.0), trace=1,steps = 100)
```

```
## Error in terms(object):  object 'lm.all' not found
```

```
#perform forward stepwise regression
lm.both <- step(lm.0, direction='both',
                scope=formula(lm.all), trace=1,steps = 10)
```

```
## Error in terms(object):  object 'lm.0' not found
```

```
#view results of backward stepwise regression
lm.back$anova
```

```
## Error in eval(expr, envir, enclos):  object 'lm.back' not found
```

```
plot(x = 1:3, y = lm.back$anova$AIC,
     main = "AIC and variable selection",
     xlim = c(1,3), "l")
```

```
## Error in h(simpleError(msg, call)):  error in evaluating the argument 'y' in selecting a
method for function 'plot':  object 'lm.back' not found
```

Backward elimination is a variable selection method which: Begins with a model that contains all variables. Then starts eliminating the least significant variables one after the other. This step function takes AIC as a criteria for choosing the best model.

```
# display comparison models
stargazer::stargazer(lm1,lm.forw,lm.back,lm.both,label = "tab2",
                     single.row = TRUE,column.sep.width = "1pt",omit.stat=c("f", "ser"))
```

```
## Error in .stargazer.wrap(..., type = type, title = title, style = style, :  object 'lm1'
not found
```

Table **??** shows OLS estimates of the regression to schooling on log income. Estimates of the coefficient of years of schooling on log wages vary from 0.052 to 0.075. This table presents the results in the traditional way. This way of presentation gives the reader a sense of the extent to which the estimates vary depending

on the exact specification of the model. The long regressions suggest that the effect of education on income is larger than in the short regressions, although the relationship does not change as the number of explanatory variables increases. The effect appears to stabilize at around 0.074. Therefore, under the standard assumptions of the OLS model, we can estimate that an additional year of schooling causes a person's income to increase by about 7.5%

Causal Pathways

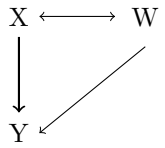

Figure 2: Dual Path Causal Graph

Longer regressions do not always reach a correct solution. Figure 2 illustrates the problem. The figure shows a confounded linear model in which two distinct causal pathways for X on Y exist. There is a direct causal effect of X on Y and an indirect causal effect of X on Y, which is mediated by W. Our ultimate goal is to estimate all three parameters. So of particular interest is determining whether or not $b = 0$. However, as shown in the figure, a backdoor relationship affecting Y through W would overestimate b.

In such a causal relationship, if all causal-related variables are omitted and the model is constructed with only the variable relationships between the target and response variables; then, the model suffers from omitted variable bias. Conversely, if more predictors are added, and yet each variable is correlated with each other, the problem of multicollinearity arises.

Figure 2 shows that x and w are causally related and therefore correlated. In other words, "causality implies correlation." This correlation between independent variables makes it difficult for the algorithm to separate the effect of x on y from the effect of w on y. When multiple causal relationships exist, it is difficult to express them in only one regression equation.

Check Assumptions

As shown in OLS model, there are four assumptions in linear regression model about the data.

(1) **Linearity of the data**

The relationship between the predictor (x) and the outcome (y) is assumed to be linear.

(2) **Normality of residuals**

The residual errors are assumed to be normally distributed.

(3) **homoscedasticity**

The residuals are assumed to have a constant variance (Homogeneity of residuals variance).

(4) **Independence of residuals error terms**

In these assumptions, we should check whether or not these assumptions hold true.

Potential problems include:

(1) *Non-linearity* of the outcome - predictor relationships

(2) *Heteroscedasticity*: Non-constant variance of error terms.

(3) *Presence of influential values* in the data that can be:

Outliers: extreme values in the outcome (y) variable

High-leverage points: extreme values in the predictors (x) variable

All these assumptions and potential problems can be checked by producing some diagnostic plots visualizing the residual errors.

```
# Check Multilinear Regression Assumptions
par(mfrow = c(2, 2))
plot(lm.both)
```

```
## Error in h(simpleError(msg, call)):  error in evaluating the argument 'x' in selecting a
method for function 'plot':  object 'lm.both' not found
```

Overall, there is nothing serious problem to violate the assumptions.

IV Model

The instrumental variable estimator (IV estimator) is the most common technique to solve the confounding model in econometrics or epidemiology (Klungel et al., 2015). IV model is used when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment. Figure 3 shows that the instrumental variable (Z) has a direct causal effect on X but is not determined by the unobserved characteristic. There is an arrow from Z to X and an arrow from U to X, but no arrow from U to Z.

Intuitively, IV is used when the explanatory variable of interest is correlated with the error term, and ordinary least squares or ANOVA would yield biased results. An effective instrument can reveal the causal effect of the explanatory variable on the dependent variable because it induces a change in the explanatory variable but has no independent effect on the dependent variable.

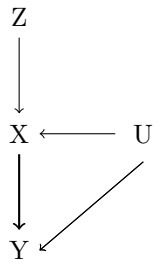

Figure 3: IV Causal Graph

An IV estimator is derived by a structural equation model (Bielby and Hauser, 1977), which is in general:

$$y = f_y(x, e_U)$$
$$x = f_x(z, e_U).$$

This equations represents that the value of variable y is determined by a function, f_y, the value of variable x and the error variable e_U, and the value of variable x is determined by a function, f_x, of the value of an instrumental variable Z and the error variable e_U on the right side.

However, most of the case, we assume the linear model in the parameter β; so rearranging the equation above into a matrix algebra:

$$\mathbf{y} = \mathbf{X}\beta + v$$
$$\mathbf{X} = \mathbf{Z}\Delta + \mathbf{E},$$

where $\mathbf{y}$ is a one-dimensional vector of the outcome of interest $\mathbf{y}$, $\mathbf{X}$ is a matrix of the observed explanatory variables $\{1, x_i\}$, β is a vector of the model parameters, and v is a one-dimensional vector of the error term v_i for the first equation. In addition, $\mathbf{Z}$ is a matrix of the instrumental variables $\{1, z_i\}$, Δ is a matrix of the relationship between the explanatory variables and the instrumental variables, and E is a matrix of unobserved characteristics determining the explanatory variables.

We can shuffle the matrix equations, and obtain:

$$\beta = \left(\mathbf{Z}^T\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{y} - \left(\mathbf{Z}^T\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{v}.$$

Since we assume that the expected value of error term is zero, we obtain the instrumental variable estimator:

$$\hat{\beta}_{IV} = \left(\mathbf{Z}^T\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{y}$$

Properties of IV Estimator

In order to set an ideal instrumental variable, it must satisfy three properties (Angrist and Pischke, 2014):

1. The variable directly affects the policy variable of interest (Z $\rightarrow$ X).
2. The variable is independent of the unobserved characteristics that affect the policy variable and the outcome of interest ($U \not\rightarrow Z$).
3. The variable affects the policy variable independently of the unobserved effect ($X = dZ + U$).

Given a variable that satisfies these three properties, IV estimator can be used to estimate the desired causality.

Now we are back to the discussion of how Card used OLS to estimate returns to schooling. Card finds that an extra year of schooling increases income by approximately 7.5%. The unobserved characteristics of the young men may determine both the amount of education that they get and the income they earn. In Figure 4, the problem is illustrated with a causal arrow running from the unobserved characteristic to both income and education.

According to Christofides et al. (1995), Card argues that young ones who grow up near a 4-year university will have lower costs of attending college and are thus more likely to get another year of education. Also, Card argues that growing up close to a 4-year college is unlikely to be determined by unobserved characteristics that also determine the amount of education that the young man gets and the income that the young man earns (Card, 2001). In the graph, the assumption is represented as an arrow from distance to college to education and no arrow from unobserved characteristics to distance to college.

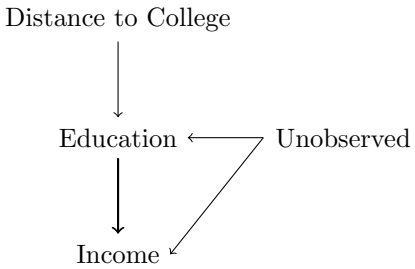

Figure 4: IV Causal Graph for Returning School

Multiple instruments are used by Card, where instruments for experience, as well as the distance to college (Christofides et al., 1995). Experience is measured as the difference between age and years of education, but education is confounded. The paper also utilizes age as the instrument for experience and squared age as the instrument for experience-squared. The instrumental variable procedure derived above is used to estimate the returns to schooling, accounting for all three instruments.

```
# Data set for income
y <- x1$lwage76
```

```
## Error in eval(expr, envir, enclos):  object 'x1' not found
```

```
# Data set for observed matrix
X <- cbind(1,x1$ed76, x1$exp, x1$exp2, x1$black, x1$reg76r,
           x1$smsa76r, x1$smsa66r, x1$reg662, x1$reg663,
           x1$reg664, x1$reg665, x1$reg666, x1$reg667,
           x1$reg668, x1$reg669)
```

```
## Error in cbind(1, x1$ed76, x1$exp, x1$exp2, x1$black, x1$reg76r, x1$smsa76r, :  object 'x1'
not found
```

```
# Data set for instrumental variables
Z1 <- cbind(1,x1$nearc4, x1$age76, x1$age2, x1$black,
            x1$reg76r,x1$smsa76r, x1$smsa66r, x1$reg662,
            x1$reg663,x1$reg664, x1$reg665, x1$reg666,
            x1$reg667, x1$reg668, x1$reg669)
```

```
## Error in cbind(1, x1$nearc4, x1$age76, x1$age2, x1$black, x1$reg76r, x1$smsa76r, :  object
'x1' not found
```

```
# Estimate the instrumental variable estimator
invZTX <- solve(t(Z1)%*%X)
```

```
## Error in t(Z1):  object 'Z1' not found
```

```
ZTy    <- t(Z1)%*%y
```

```
## Error in t(Z1):  object 'Z1' not found
```

```
beta_iv<- invZTX%*%ZTy
```

```
## Error in eval(expr, envir, enclos):  object 'invZTX' not found
```

```
## Calculate t statistics for significance test
# calculate sigma squared hat
N=length(y)
```

```
## Error in eval(expr, envir, enclos):  object 'y' not found
```

```
k=ncol(X)-1
```

```
## Error in ncol(X): object 'X' not found
```

```
sigma2 <- t(y - X %*% beta_iv) %*% (y - X %*% beta_iv) / (N-k-1)
```

```
## Error in t(y - X %*% beta_iv):  object 'y' not found
```

```
# allocate the lists of standard error and t stat for alpha
se.b <- t.b  <- c(rep(0,k+1))
```

```
## Error in eval(expr, envir, enclos):  object 'k' not found
```

```
# calculate standard error for estimates
for (i in 1:(k+1)){
  se.b[i] <- sqrt(sigma2 * invZTX[i,i])
}
```

```
## Error in eval(expr, envir, enclos):  object 'k' not found
```

```
# calculate t statistics for each estimates
for (i in 1:(k+1)){
  t.b[i] <- beta_iv[i]/se.b[i]
}
```

```
## Error in eval(expr, envir, enclos):  object 'k' not found
```

```
# critical value at significant level 0.05
cv025 <- qt(0.975, N-k-1, lower.tail=TRUE)
```

```
## Error in qt(0.975, N - k - 1, lower.tail = TRUE): object 'N' not found
```

```
# show beta_iv and cv
beta_iv_cv<-data.frame(coef=c(beta_iv),se=c(se.b),
                       cv05=c(beta_iv-cv025*se.b),cv95=c(beta_iv+cv025*se.b),
           row.names = c("intercept","ed76","exp","exp2","black","reg76r",
                         "smsa76r","smsa66r","reg662","reg663","reg664",
                         "reg665","reg666","reg667","reg668","reg669"))
```

```
## Error in data.frame(coef = c(beta_iv), se = c(se.b), cv05 = c(beta_iv - :  object 'beta_iv'
not found
```

```
print(xtable(beta_iv_cv))
```

```
## Error in xtable(beta_iv_cv):  object 'beta_iv_cv' not found
```

The instrumental variable estimates for returns to schooling are substantially higher than the OLS estimates. This is contrary to my expectation. In general, I expected the OLS estimates of the causal effect of schooling on income to be biased upward. The concern is that the OLS might be picking up on the fact that the family environment determines both college attendance and access to higher-earning jobs. This result suggests that the OLS estimates are biased downward. The reason for the upward bias of the instrumental variable estimates will be clarified in future studies, since this is a comparison of three models: OLS, the instrumental variable estimation, and LiNGAM.

Note that the equations of interest for the instrumental variable method are structural, not regression. So, the validation of the assumptions is different from the usual regression model.

LiNGAM

The instrumental variables (IV) method is one of the causal discoveries in statistics. It is used to estimate causal relationships, utilizing a structural equation modelling (SEM), when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment. Shimizu (2006) introduced a non-Gaussian version of the linear acyclic SEM with no latent confounders, known as a linear non-Gaussian acyclic model, abbreviated as LiNGAM:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i$$

where e_i are exogenously latent variables, and b_{ij} are the connection strengths from x_j to x_i. With the causal ordering of the variables x_i, denoted by $k(i)$, the causal relations of the variables x_i can be graphically represented by using a directed acyclic graph in Figure 5. The exogenous variables e_i follow non-Gaussian distributions but one of e_i at most Gaussian, with zero mean and non-zero variance and are mutually independent. The independence assumption between e_i implies that there are no latent confounding variables. In matrix form, the LiNGAM model is written as:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

where the matrix $\mathbf{B}$ collects the connection strengths b_{ij} , and the vectors $\mathbf{x}$ and $\mathbf{e}$ collect the observed variables x_i and the exogenous variables e_i, respectively.

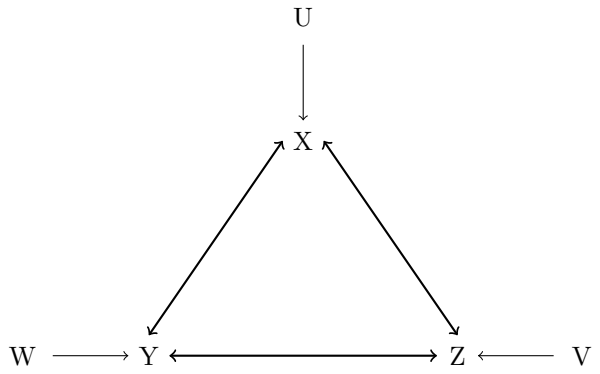

Figure 5: Directed Acyclic Graph (Causal Pathway)

According to Shimizu (2014), the fundamental concept of LiNGAM and its estimation is closely related to the independent component analysis (ICA). ICA is a non-Gaussian variant of factor analysis, and the ICA model for the observed variables x_i can be defined as follows:

$$x_i = \sum_{j=1}^{d} a_{ij} s_j$$

We do not go beyond the detail of ICA and how to solve using ICA and work the algorithm to its estimation.

Assumptions to LiNGAM

LiNGAM has the basic four assumptions (Shimizu, 2006):

1. Linearity: simplified model in this case but can be extended
2. Acyclicity: causal structure (Parent-child relationship of the data variables)
3. Non-Gaussian: semi-parametric or at most one is Gaussian error distribution
4. Markov Property: all nodes are independent of their non-descendants when conditioned on their parents

The major difference between LiNGAM and the estimation method of the instrumental variables is: LiNGAM is computationally capable of performing a causal search without human intervention, while IV is in that the researcher constructs the IV, and causal inference is made statistically from its structural equations.

So we would like to apply LiNGAM estimation to the Card's data of returns of schooling. The data used is the same data used in Multiple regression and IV estimation. The LiNGAM estimation method uses the R package of "pcalg: Methods for Graphical Models and Causal Inference" [2] to estimate the connection strength matrix **B**.

The R code and the resulting matrix **B** is shown below.

```
# Data setup
X <- cbind(x1$lwage76, x1$ed76, x1$exp, x1$exp2, x1$black, x1$reg76r,
           x1$smsa76r, x1$smsa66r, x1$reg662, x1$reg663,
           x1$reg664, x1$reg665, x1$reg666, x1$reg667,
           x1$reg668, x1$reg669)

## Error in cbind(x1$lwage76, x1$ed76, x1$exp, x1$exp2, x1$black, x1$reg76r, :  object 'x1'
not found

# LiNGAM function to estimate causal discoveries
lingam.res <- lingam(X, verbose = 2)

## Error in ncol(X): object 'X' not found
```

```
# as(lingam.res, "amat")
# display the connection strenght matrix
print(xtable(lingam.res$Bpruned, type = "latex"))

## Error in xtable(lingam.res$Bpruned, type = "latex"):  object 'lingam.res' not found
```

A causal graph from the estimated connection strength matrix yielded results that contradicted my expectations. In this graph, to simplify the graph, I grouped the following variables together and eliminated the non-connected variables. Two of $\{exp\}$ and $\{exp2\}$ is grouped to $\{work\}$, three of $\{reg76r\}$, $\{smsa76r\}$, and $\{smsa66r\}$ is a set to $\{regions\}$. The causality graph suggests that "income" affects "education,"

[2] https://cran.r-project.org/web/packages/pcalg/pcalg.pdf

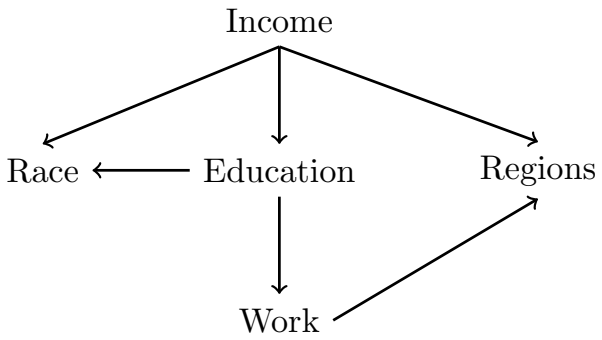

Figure 6: Directed Acyclic Graph of Returning to School

"race," and "years of employment"; "education" affects "race" and "years of employment"; and "years of employment" affects "area of residence." The results are the opposite of the commonly accepted causal relationship.

4 Discussion and Conclusion

This paper represents the comparison of the estimations of higher education on individual income levels, using microeconomic survey data on men aged between 14 and 24 in 1966, with four statistical models: (1) Simple Linear Regression, (2) Multiple Linear Regression, (3) the Instrumental Variable Estimation, and (4) Linear Non-Gaussian Acyclic Model.

Under the standard T testing approach, we cannot reject that the estimated parameter, the effect of schooling on income, every linear regression model is zero. Furthermore, there are no severe issues with each linear regression's assumptions.

```
## Error in data.frame(simple = c(lm1$coefficients[1], lm1$coefficients[2], :  object 'lm1'
not found
## Error in xtable(comp, type = "latex"):  object 'comp' not found
```

In all models, it is an implication that education in school has a relationship to income. In other words, on average, an association between schooling and individual income typically exists. The IV model reveals a possibility that education may have an impact on income using college proximity as an instrumental variable. However, the effect showed different results in each model. The single regression analysis estimated a smaller relationship between college education and income than the multiple regression analysis. Perhaps the simple regression suffers from an omitted variable bias, the relationship between education and income was more extensive when variables other than education, such as years of work, were added to the regression model.

An introduction of the instrumental variable method, however, called the relationship into question. The result of the instrumental variable estimate the effect of schooling to returns substantially higher than the OLS estimates. Because what the IV results show is that the influence relationship between each other by OLS is biased downward. In many studies, the introduction of the instrumental variable method has strengthened the causal relationship, but our IV model has weakened it for education and income. Whether this is due to a problem in the setting of the control variables or a problem in the structural equations awaits future detailed study.

A more interesting result is that when the recently developed LiNGAM is applied to the Card's data, it indicates the opposite of the common-sense causal relationship. That is, income may have an effect on education. This experimental result was obtained in this study because the data were directly applied to the LiNGAM model without considering time constraints among variables. In future studies, it is expected that introducing a temporal constraint condition will result in a more accurate statistical causal search. Despite being young history, causal discovery is a promising field that may help bridge the gap between machine and human knowledge.

References

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton university press.

Banerjee, A. V., Duflo, E., and Kremer, M. (2016). The influence of randomized controlled trials on development economics research and on development policy. *The state of Economics, the state of the world*, pages 482–488.

Bielby, W. T. and Hauser, R. M. (1977). Structural equation models. *Annual review of sociology*, 3(1):137–161.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160.

Chalmers, T. C., Smith Jr, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., and Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled clinical trials*, 2(1):31–49.

Christofides, L. N., Grant, E. K., Vanderkamp, J., and Swidinsky, R. (1995). *Aspects of labour market behaviour: Essays in honour of John Vanderkamp*. University of Toronto Press.

DiNardo, J. (2010). Natural experiments and quasi-natural experiments. In *Microeconometrics*, pages 139–153. Springer.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Klungel, O., Uddin, M. J., de Boer, A., Belitser, S., Groenwold, R., Roes, K., et al. (2015). Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods. *Pharm Anal Acta*, 6(353):2.

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging themes in epidemiology*, 15(1):1–7.

Mann, H. B. (1949). Analysis and design of experiments: Analysis of variance and analysis of variance designs. Technical report.

Shimizu, S. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Shimizu, S. (2014). Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98.

Stock, J. H. and Trebbi, F. (2003). Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194.