Problem Set 1 - Econometrics

Giacomo Lo Conte, Kun Wu, Francesca Eustacchi, Neeharika Kakunuri, Noor Ahmed Khoso

January 24, 2022

1 Theory I: DAGs and Potential Outcomes

a) Consider the following threshold model with a binary treatment D, an additional (binary) covariate x, the outcome y and an i.i.d error term with mean zero ϵ_i

$$y_i = \alpha + \beta D_i + \gamma x_i + \rho \epsilon_i \tag{1}$$

$$x_i = \mathbb{1}(\kappa + \delta D_i + \pi \epsilon_i > c) \tag{2}$$

The second part of equation 2 means that $x_i = 1$ if the sum on the right hand side is greater than the threshold c and $x_i = 0$ otherwise. Assuming that $\delta = 0$, show that the estimator of the average treatment effect (ATE), $\Delta = E(y|D = 1, x = X) - E(y|D = 0, x = X)$ equals $\beta + bias$, characterise the bias (i.e. derive the exact formula) and explain what the bias means.

Consider the expected value of y conditioned to $x = X$ and $D = 0$ or $D = 1$.

$$\begin{aligned} y_{D=1} &= E[y|D = 1, x = X] = E[\alpha + \beta + \gamma x + \rho\epsilon|D = 1, x = X] \\ y_{D=0} &= E[y|D = 0, x = X] = E[\alpha + \gamma x + \rho\epsilon|D = 0, x = X] \end{aligned} \tag{3}$$

$$\begin{aligned} y_{D=1} &= E[y|D = 1, x = X] = \alpha + \beta + \gamma E[x|D = 1, x = X] + \rho E[\epsilon|D = 1, x = X] \\ y_{D=0} &= E[y|D = 0, x = X] = \alpha + \gamma E[x|D = 0, x = X] + \rho E[\epsilon|D = 0, x = X] \end{aligned} \tag{4}$$

Since $\delta = 0$, $E[x|D] = E[\mathbb{1}(\kappa + \pi\epsilon > c)|D] = E\left[\mathbb{1}\left(\epsilon > \frac{c}{\pi}\right)|D\right]$ both when $D = 1$ and $D = 0$. The equation of Δ can be written as:

$$\Delta = \alpha + \beta + \gamma E[x|D] + \rho E[\epsilon|D = 1, x = X] - \alpha - \gamma E[x|D] - \rho E[\epsilon|D = 0, x = X] \tag{5}$$

Cancelling out the equal terms, we arrive at:

$$\Delta = \beta + \rho(E[\epsilon|D=0, x=X] - E[\epsilon|D=1, x=X]) = \beta + bias \tag{6}$$

If the error term is correlated with the treatment, then the term in the parentheses is different from 0 and it is a bias. This is a selection bias since the outcome across the two groups is influenced by unobservable variables correlated with the treatment captured by the error term. The sign of the bias is positive if these variables impact the outcome in the same direction the treatment affect them. Otherwise it will be negative.

b) Now assume that $\delta \neq 0$. Write down a DAG that represents the model. Should the researcher account for x to deconfound the treatment effect?

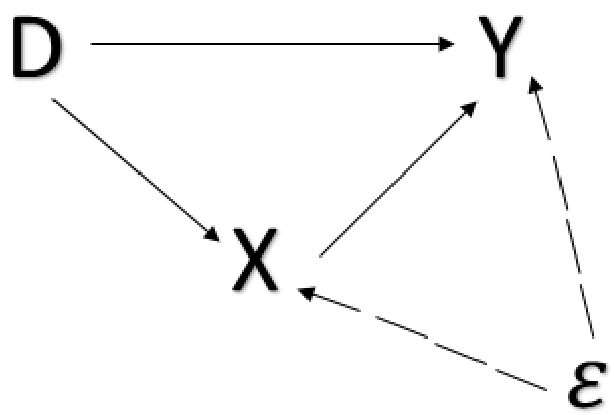

Figure 1: DAG of the theoretical model

The researcher should not account for x to deconfound the treatment effect because x is a mediator, and thus conditioning on it would lead to a selection bias. In this case, x is not as good as randomly assigned because it is a function of D.

c) Now assume that $\delta \neq 0$ and $\gamma = 0$. Using the potential outcomes framework, show that the bias $E(y|D=1, x=X) - E(y|D=0, x=X) - \beta$ can be re-written as a weighted average of $E(\epsilon|D=1, x=X) - E(\epsilon|D=0, x=X)$ across group switch $x=1$ and $x=0$. Provide a brief interpretation of the bias term.

Compute $\Delta = E(y|D=1, x=X) - E(y|D=0, x=X)$, when $\gamma = 0$.

$$\begin{aligned} \Delta &= E[\alpha + \beta + \rho\epsilon|D=1, x=X] - E[\alpha + \rho\epsilon|D=0, x=X] \\ \Delta &= \beta + \rho(E[\epsilon|D=1, x=X] - E[\epsilon|D=0, x=X]) \end{aligned} \tag{7}$$

Consider now the term in parentheses in equation (7). Notice that x can assume only two values 1 and 0, then we can rewrite $E[\epsilon|x = X]$ as $E[\epsilon|x = 1 \cup x = 0]$. Since this represents the union of the two possible events (x is either 0 or 1, but it can never be something different from this), we can split the condition based on the frequency they appear in the dataset with.

$$\begin{aligned} E[\epsilon|D = 1, x = X] - E[\epsilon|D = 0, x = X] = E[\epsilon|D = 1, x = 1]\frac{E[x = 1|X]}{E[x = X]} + E[\epsilon|D = 1, x = 0]\frac{E[x = 0|X]}{E[x = X]} - \\ + E[\epsilon|D = 0, x = 1]\frac{E[x = 1|X]}{E[x = X]} + E[\epsilon|D = 0, x = 0]\frac{E[x = 0|X]}{E[x = X]} \end{aligned} \tag{8}$$

Since $E[x] = E[x = 1|X] + E[x = 0|X]$, we can write $p = \frac{E[x=1|X]}{E[x]}$ and $1 - p = \frac{E[x=0|X]}{E[x]}$. The previous equation then can be rewritten as:

$$\begin{aligned} E[\epsilon|D = 1, x = X] - E[\epsilon|D = 0, x = X] = p\Bigg(E[\epsilon|D = 1, x = 1] - E[\epsilon|D = 0, x = 1] \Bigg) + \\ + (1 - p)\Bigg(E[\epsilon|D = 1, x = 0] - E[\epsilon|D = 0, x = 0] \Bigg) \end{aligned} \tag{9}$$

Notice that p is less or equal than 1 and greater or equal than 0. Moreover they sum to 1. Such a difference can be interpreted as a weighted mean of the difference $\Delta_E = E[\epsilon|D = 1, x = X] - E[\epsilon|D = 0, x = X]$ across $x = 0$ and $x = 1$ groups. Now consider Δ again. As we proved $\Delta = \beta + \rho\Delta_E \Rightarrow \rho\Delta_E = \Delta - \beta$. Considering that ρ is a constant, we can write $\rho E[\epsilon] = E[\rho\epsilon] = E[e]$. Then it becomes:

$$\begin{aligned} E(y|D = 1, x = X) - E(y|D = 0, x = X) - \beta = p\Bigg(E[e|D = 1, x = 1] - E[e|D = 0, x = 1] \Bigg) + \\ (1 - p)\Bigg(E[e|D = 1, x = 0] - E[e|D = 0, x = 0] \Bigg) \end{aligned} \tag{10}$$

d) Suppose $\beta > 0$. Derive conditions under which the inclusion of x would lead to the underestimation of β, i.e. $E(\hat{\beta}) < \beta$.

Consider the OLS estimator $\hat{\beta}$:

$$\begin{aligned} \hat{\beta} &= \frac{Cov(y, D)}{Var(D)} = \frac{Cov(\alpha + \beta D + \gamma x + \rho\epsilon, D)}{Var(D)} \\ \hat{\beta} &= \frac{Cov(\alpha, D)}{Var(D)} + \frac{Cov(\beta D, D)}{Var(D)} + \frac{Cov(\gamma x, D)}{Var(D)} + \frac{Cov(\rho\epsilon, D)}{Var(D)} \\ \hat{\beta} &= \beta + \gamma\frac{Cov(x, D)}{Var(D)} + \rho\frac{Cov(\epsilon, D)}{Var(D)} \end{aligned} \tag{11}$$

Consider now the expected value of $\hat{\beta}$:

$$E[\hat{\beta}] = E\left[\beta + \gamma\frac{Cov(x, D)}{Var(D)} + \rho\frac{Cov(\epsilon, D)}{Var(D)}\right] = \beta + \gamma E\left[\frac{Cov(x, D)}{Var(D)}\right] \tag{12}$$

For $E[\hat{\beta}] < \beta$, we must have $\gamma E\left[\frac{Cov(x, D)}{Var(D)}\right] < 0$. Consider then this last term:

$$E\left[\frac{Cov(x, D)}{Var(D)}\right] = E\left[\frac{Cov[\mathbb{1}(\kappa + \delta D + \pi\epsilon), D]}{Var(D)}\right] \tag{13}$$

The sign of this term is given by the sign of δ. So the conditions for it to be true is that either $\gamma > 0$ and $\delta < 0$ or $\gamma < 0$ and $\delta > 0$.

2 Theory and Simulation

2.1 The Gender Pay Gap

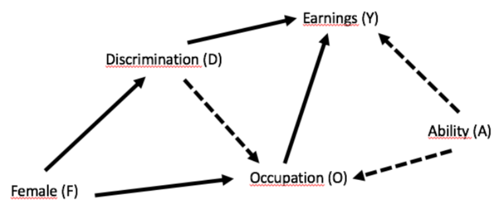

Figure 1 – Directed Acyclical Graph: Gender Discrimination and Earnings

Figure 2: DAG of the theoretical model

A researcher wants to estimate the effect of gender discrimination on earnings. To deconfound the causal effect, he/she develops the causal diagram shown in Figure 1. All variables except A are observed in the data set. F is a dummy variable that equals unity if a person is female. 0 is a set of dummy variables for broad occupational categories. D is a dummy variable indicating if a person is being discriminated or not. Assume that only women are discriminated against. The arrow from D to O is dashed because it is theoretically unclear whether we should expect an effect of discrimination on occupation.

a) Provide an intuitive explanation for each arrow in Figure 1. Provide an explanation for the absence of an arrow between F and Y . In your view, should there be additional arrows and/or variables in the diagram? If so, which ones?

(a.1) DAGs Arrow Explanation

1. Female on Occupation (F → O)

When O varies as F differs, it represents the changes in likelihood of having different choice range for occupation conditioning on gender. When a negative association is present, it implies that $F = 1$ leads to decline in the likelihood of observing $O = 1$, in which female is selected out in employment participation by gender disparities for given occupations. Instead of tangible economic reasons, preference over gender-specific occupations are introduced commonly by social norm or cultural belief.

Outflow and inflow of self-selection emerges when the gender-specific advantage is subjectively perceived for given occupations. Only a few extreme cases exist where candidates selection generates unequal opportunity by design, such as military service where being male has dominant conscription advantages against female.

2. Occupation on Earnings (O → Y)

When O is positively associated with Y, higher earnings are expected from occupations which are categorised as in $O = 1$. Wages vary across occupations given by non-pecuniary characteristics of different jobs, compensation for skill training cost, institutional factors which obstruct the market equalisation of wage, hierarchy introduced by restricted access to resource and opportunities to limited circles of eligibility, valuation of personal natural ability and cognitive skills, etc. The present occupation dummy implies a binary categorisation of aforementioned elements regarding the determination of wages.

3. Female on Discrimination (F $\rightarrow$ D)

When F is positively associated with D, it is more likely to encounter discrimination for female than male on the basis of gender. When only gender is considered regarding discrimination behaviour, it links to the belief about a particular conceptualised masculinity and femininity. Acceptance and expectation derive from social or cultural customs, and norms and contribute to the gender stereotypes, which leads to presence of gender superiority to another in this case "against female". The presence and severity of discrimination vary across societies, social groups, social roles, as well as vary at equality movement progress.

4. Discrimination on Earnings (D $\rightarrow$ Y)

When D is negatively correlated with Y, being discriminated against reduces the likelihood of having higher earnings. Based on the characteristics of the social groups, discrimination introduces the unjustified distinction on earnings, where the valuation treatment for identical labour output is unequal. All sources of discrimination, such as race, gender, age, religion, or sexual orientation, confront the labour market equilibrium, where employers forego the competitive advantage due to discriminatory employment decision. There are variations of the D$\rightarrow$Y direct effect based on societal factors, such as inter-group attitudes, equality movement, and anti-discrimination legislation process.

5. Discrimination on Occupation (D $\rightarrow$ O)

When D has a negative effect on O, being discriminated against is less likely to have broader range of occupation choice and more likely to be associated with $O = 0$. Compared to occupational integration, discrimination introduces both horizontal segregation (share) and vertical segregation (hierarchy). Preference and self-selection formed based on the cultural and social norms affect the socioeconomic status that limits the access to resources, which lead to occupational inequality and economic inefficiency of labour allocation. Similar to D$\rightarrow$E, the direct effect of D$\rightarrow$O varies across occupations and within occupations, and the effect also depends on the equality-relevant societal factors.

6. Ability on Occupation (A $\rightarrow$ O)

A positive correlation between A and O implies that the likelihood to have a broad choice range of occupation is higher for individuals who score higher in the measure of ability. More competent individuals are expected to perform better in cognitive skill formation and vocational skill acquisition, and therefore characterised with higher productivity, which receives higher valuation across occupations that leads to more job opportunities.

7. Ability on Earnings (A → Y)

A positive correlation between A and Y implies that competency is associated with higher earnings. The difference in ability will affect the scale of operation: the greater the scale of operation will potentially lead to greater output. Hence, the difference in ability will generate discrepancy in output, and the valuation of higher productivity will be compensated in monetary term.

(a.2) Absence of Arrow between F and Y

Holding all the other factors fixed, when it's only the gender making the earnings different, such relationship is captured by the degree, or presence, of discrimination. When there is no discrimination and no discerning on gender, the earnings are associated with the different payment for occupations and productivity. Thus, the characteristics of having different earnings are fully explained by either the presence of discrimination, or that of different occupation categories, or the ability, the binary variable of gender alone cannot have a direct effect on earnings.

(a.3) Potential Additional Arrows and Variables

The arrows of F → A and A → D carry no meaningful direct effects. For instance for F on A, with expected coefficients estimates with zero, since the gender and ability are independent with each other.

Education. Considering the wage gap in earnings and occupational segregation from O→Y and D→O, we are interested in the educational attainment which is involved with both. Education is denoted with E, we assume that between education and ability, ability dominates the influence on the outcome of education, say the degree obtained, instead of assuming education affects the ability. Therefore, we distinguish between the skills that obtain from education and the ability that individuals possess as invariant innate characteristics.

We assume that A has direct effect on E, D also has direct effect on E, and E will have direct effect on O. As in our assumption, ability will affect the education outcome as well as the degree obtain which presumably will be positively correlated with O. Note that we have considered educational disparities which implies a negative effect on E from discrimination. This is expressed in figure 3 .

Marital Status. In figure 4, Considering the occupation self-selection due to the marital status, we are interested in the preference over occupations as well as the discrimination against marital status, as with the extreme case of housewife. Denote marital status with M as dummy variable that equals to unity if married, we assume M has negative direct effect on O, and M will have positive effect on D. For the discrimination towards marital status, we concern about when M=0, single will face unequal treatment of labour requirement and unjustified employment.

Childbearing. Very similar to M, we are interested about the preference over occupation as well

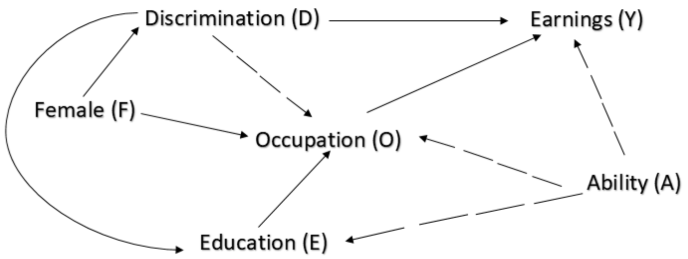

Figure 3: DAG with Education

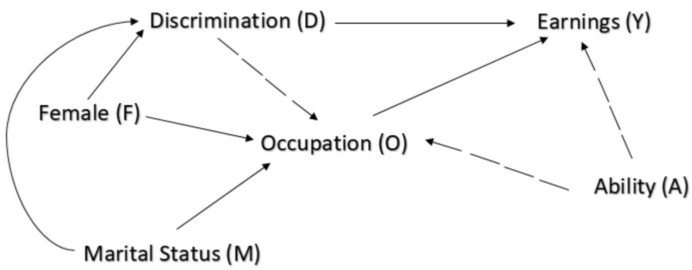

Figure 4: DAG with Marital Status

as the discrimination against childbearing individuals. Denote childbearing with C as dummy variable which equals to unity when there are children in the household. We consider the same paths and effects for the additional C as in the DAG for M. We consider that for the discrimination towards childbearing individuals are more severe. The concern may rise for the negative effect on occupation suitability when C=1, and the difficulties merge for mothers when the social norm has strong preference over particular expectations for motherhood. This is expressed in the DAG in figure 5.

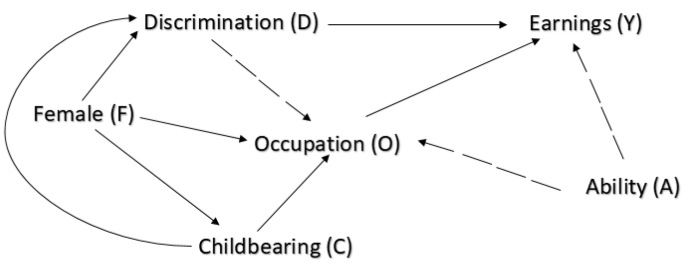

Figure 5: DAG with Childbearing

Equality Environment. Considering the potential variation on the prevalence and severity of discrimination for D→O, D→E, F→D, we are interested in the state of equality movement for the society that amend the discriminatory behaviours across the population. Denote the equality environment with G as scale ladder as shown in figure 6, the higher score represents more advanced equality progress, the less discriminatory activities. We consider G would only have positive effect on D.

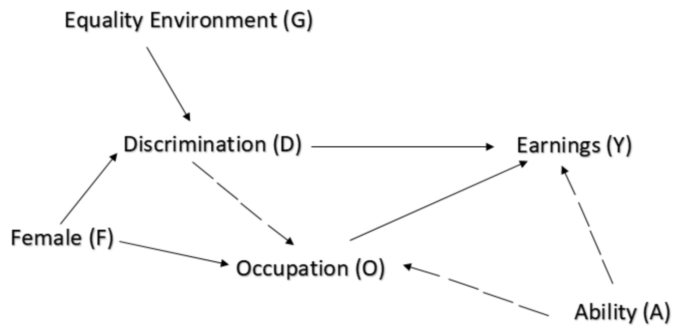

Figure 6: DAG with Equality Environment

b) Write out the paths from D to Y . Now assume that you observe ability and can control for it. Explain whether it makes sense (or not) to additionally control for the following: i) only F ; ii) only O; iii) both.

path 1, between D and Y: D→Y.

path 2, without F and A: D→O→Y.

path 3, without F: D→O←A→Y.

path 4, without A: D←F→O→Y.

path 5, with all: D←F→O←A→Y.

i) When A is observable and controlled, path 3 and path 5 with presence of A are reduced to path 2 and path 4, respectively. From the arrows indicated in the paths we can conclude that without F, the DAG is closed. It is unnecessary to control for F. Further, since F has indirect effect on Y via D and via O, to add F in the model specification for controlling F is not informative. In addition, to add F would potential open paths that are unobserved or not concerned.

ii) When A is controlled, from the DAG we may conclude that O is a mediator between D and Y. This bad control will further interfere with the bias on D's effect on Y.

iii) Path 2 represents the case when F and A are both controlled. In addition to F, further controlling for O means we control for a mediator, which leads to the selection bias where with the presence of discrimination, there is self-selection towards preference over occupations. The direction of the bias can be obtained by estimation.

c) Assume that A is unobservable. Explain why controlling for O can lead to collider bias.

When A is unobservable, thus the backdoor path from A cannot be closed. When controlling for O, collider bias may arise for the paths between F and A. We may establish spurious correlation between D and A, which are considered as two uncorrelated independent variables.

d) Illustrate the collider problem in a simulation based on the above causal diagram. To

do so, create a (simulated) dataset that represents all the arrows in Figure 1. It is sufficient to approximate O with one dummy variable. You will have to run several regressions; at the least, show the following regressions: i) Y on D; ii) Y on D controlling for O; iii) Y on D controlling for O and A. Run further regressions if needed and explain why the coefficients differ (or not). The task is here to show convincingly that controlling for variables on the causal path can lead to collider bias. This is what researchers would do in methodological papers, conference discussions or referee reports.

The data set simulated data set contains 1000 observations. We have used the tool R for simulating this data set. We have generated dummy variables for *"Female"*, *"Occupation"* and *"Discrimination"*. To simulate our model we picked parameters $a = 0.5$, $b = 2$, $c = 3$, $d = -1.2$, $e = -2.1$, $f = 0.25$, $g = 23.5$, $h = 5.8$, $i = -3.03$ and $k = 12.8$ to compare our estimated parameters. The simulated model is described by the following set of equations:

$$\begin{aligned} &Discrimination = \mathbb{1}(a + b^{*}Female + u_D > 2.8) \\ &Occupation = \mathbb{1}(c + d^{*}Discrimination + e^{*}Female + f^{*}Ability + u_O > 7) \\ &Earnings = g + h^{*}Ability + i^{*}Discrimination + k^{*}Occupation + u_E \end{aligned} \tag{14}$$

Table 1: Regression summary for Earnings on Discrimination

	Dependent variable:
	Earnings
Discrimination	-5.893^{***}
	(1.606)
Constant	142.849^{***}
	(1.073)
Observations	1,000
R^2	0.013
Adjusted R^2	0.012
Residual Std. Error	25.246 (df = 998)
F Statistic	13.465^{***} (df = 1; 998)
Note:	*p<0.1; **p<0.05; ***p<0.01

This table reports estimation results from an OLS Model of *Earnings* on *Discrimination*. Standard error reported in parentheses in column.

In table 1, we see the summary of the regression for earnings upon discrimination. We see that the estimate of discrimination is significant at 1 percent. The estimate is -5.893 indicating that in the presence of discrimination, earning fall by 5.893 units. However, when we compare the parameter i which is associated with Discrimination , we that the decrease in wages is lesser than -3.03. This implies that our estimate is biased when we control only for discrimination.

Table 2: Regression Summary for Earnings, Discrimination controlling for Occupation

	Dependent variable:
	Earnings
Discrimination	-3.231^{**}
	(1.426)
Occupation	26.543^{***}
	(1.574)
Constant	122.630^{***}
	(1.527)
Observations	1,000
R^2	0.232
Adjusted R^2	0.231
Residual Std. Error	22.279 (df = 997)
F Statistic	150.887^{***} (df = 2; 997)
Note:	*p<0.1; **p<0.05; ***p<0.01

This table reports estimation results from an OLS Model of *Earnings* on *Discrimination* and *Occupation*. Standard error reported in parentheses in column.

In table 2 we see the summary of regression of the effect of discrimination on earnings while controlling for occupation. The estimates are significant at 1 percent. And interestingly we see that earnings reduce by 3.231 units which is significant with the set parameter i which is 3.03 units. With an increase in the occupational categories, we witness that earnings increase by 25.543 euros. Additionally, the explanatory power has now increased from 0.013 from regression in table 1 to 0.231. The bias of discrimination is slightly better, though is still significant since we face the case of a "bad control" bias. This bias derives from the fact that occupation is a mediator of discrimination on earnings.

In table 4 we observe the summary of the regression earnings explained by discrimination while controlling for ability and occupation. The regression estimates are significant at 1 percent. Additionally, the models, explanatory power as explained by R^2 to 0.977. We also see that the coefficient of discrimination is -3.189 which is significant when compared to the parameter we have set. Moreover, we see that the coefficient is more significant when compared to the coefficient we have obtained in regression 2, in table 2. Therefore, using the control for ability and occupation has improved our model, because, ability is a confounder in our model. Including ability we have closed all the back-doors in our model we have ruled out the bias affecting the discrimination coefficient.

Table 3: Regression Summary for Earnings , Discrimination controlling for occupation and ability

	Dependent variable:
	Earnings
Discrimination	-3.189^{***}
	(0.246)
Occupation	12.771^{***}
	(0.282)
Ability	5.799^{***}
	(0.032)
Constant	39.634^{***}
	(0.531)
Observations	1,000
R^2	0.977
Adjusted R^2	0.977
Residual Std. Error	3.845 (df = 996)
F Statistic	$14{,}200.460^{***}$ (df = 3; 996)
Note:	*p<0.1; **p<0.05; ***p<0.01

This table reports estimation results from an OLS Model of *Earnings* on *Discrimination*, *Occupation* and *Ability*. Standard error reported in parentheses in column.

We tried to run two additional regressions, to observe the behaviour of the bias when we exclude some controls from the regressions. We begin by estimating Y as a function of discrimination and ability as represented in the results shown in 4. As we see notice, There is a decrease of 4.373 units in earnings due to discrimination when we include ability into the model. The explanatory power of the model is remains close to the explanatory power in model 3. Additionally, all the estimates are significant at 1 percent. Furthermore, this regression is of great importance as it highlights the bad control problem when we use occupation. The reason is that, since we cannot be sure about the link between discrimination and occupation, we want to see if ruling it out from the regression the estimates are unbiased. We keep ability since it is an important element for earnings. By comparing with the other regressions (and the true parameter), we notice the estimates are biased. The reason is that without including occupation there is a back-door in our model we do not close.

In table 5, we see the regression summary for the earnings explained by occupation and ability. Also these estimates are biased. Indeed, without controlling for discrimination we would have a bias due to the non-inclusion for a confounder (discrimination).

Table 4: Regression Summary for Earnings , Discrimination and Ability

	Dependent variable:
	Earnings
Discrimination	-4.373^{***}
	(0.428)
Ability	6.193^{***}
	(0.054)
Constant	43.002^{***}
	(0.918)
Observations	1,000
R^2	0.930
Adjusted R^2	0.930
Residual Std. Error	6.720 (df = 997)
F Statistic	$6,639.879^{***}$ (df = 2; 997)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

This table reports estimation results from an OLS Model of *Earnings* on *Discrimination* and *Ability*. Standard error reported in parentheses in column.

Table 5: Regression Summary for Earnings, Occupation while controlling for Ability

	Dependent variable:
	Earnings
Occupation	13.159***
	(0.303)
Ability	5.799***
	(0.035)
Constant	37.927***
	(0.555)
Observations	1,000
R^2	0.973
Adjusted R^2	0.973
Residual Std. Error	4.155 (df = 997)
F Statistic	18,175.270*** (df = 2; 997)
Note:	*p<0.1; **p<0.05; ***p<0.01

This table reports estimation results from an OLS Model of *Earnings* on *Occupation* and *Ability*. Standard error reported in parentheses in column.

2.2 Application: Miscarriages and the Outcomes of Subsequent Children

In a new paper, the authors study the effect of a mother having a miscarriage on the outcomes of sub-sequent children. Prior research has shown that miscarriages, while common, can have traumatic effects on women and, by extension, on families. It is thus plausible that a miscarriage affects the outcomes of children that were subsequently conceived and born, for example through changes in parenting styles.

The authors undertake several steps towards establishing causality. They refer to a large number of studies showing that the likelihood of having miscarriages appear to be unrelated to mother or family characteristics, and provide balancing tests in support of this assumption. A second challenge is to find a suitable control group. They restrict the sample to families with two children; the treatment group had a miscarriage in between the births of both children, the control group had no miscarriage.

a) While this identification strategy appears plausible at first, the choice of control group may induce a bad control problem (i.e. the choice is equivalent to conditioning on a mediator). Construct a DAG to explain where the bad control problem could lie and why this might bias the estimates (Hint: it has to do with the decision to have another child after a miscarriage). Discuss under what conditions the assumption that having a miscarriage is random is sufficient for establishing causality.

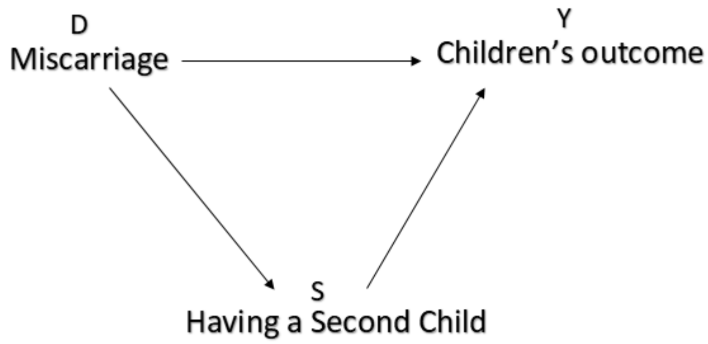

Figure 7: DAG of the theoretical model

In this case, (S), which indicates the choice of having a second child is not as good as randomly assigned. In fact, S is a function of miscarriage. If we condition on S we could have an upward or downward bias, and both are plausible. In case of downward bias, we expect a direct effect of miscarriage on the decision to have a second child as negative. In this case, women who have had a miscarriage are less likely to have a second child, because they could have a traumatic shock related to the miscarriage that discourage them to tempt another pregnancy. Lower S would lead lower Y, but the effect would be reduced by the bias.

In case of upward bias, we expect a direct effect of miscarriage on the decision to have a second child as positive. In this case, women who have had a miscarriage are more likely to have a second child, which could be explained by the fact that they are more mentally prepared because the consequence of the trauma of a miscarriage has already been addressed. Lower S would lead to lower Y, but the effect would be augmented by the bias.

The assumption of having miscarriage as random is sufficient for establishing causality if having a second child (S) is independent from having had miscarriage (D).

b) Using potential outcomes notation, derive the bias in the estimation of the ATE that results from the bad control problem found in a). Explain in what direction the bias could likely go.

In general, when we use the potential outcome notation, we define the casual effect of a treatment as the difference between the expected values of the outcome across the treatment and the control group:

$$\Delta = E[Y|D=1] - E[Y|D=0] \tag{15}$$

In the specific case, the author is controlling the treatment effects only for families with two children

(S=1) and for those families who suffered a miscarriage between the two pregnancies. Suppose that the outcome is defined according the following Data Generating Process:

$$Y = \alpha + \beta D + \gamma S + X'\delta + u \tag{16}$$

The underlying assumption of the authors is that there is no selection bias in the estimation of the Average Treatment Effect, i.e.:

$$\begin{aligned} ATE = E[Y|D=1,S=1] - E[Y|D=0,S=1] = \alpha + \beta + \gamma E[S|D=1] + E[X'|D=1]\delta \\ - \alpha - \gamma E[S|D=0] - E[X'|D=0]\delta = \beta \end{aligned} \tag{17}$$

Even assuming that $E[X'|D=1] = E[X'|D=0]$, it is hard to justify that the number of children a couple decide to have is not affected by the treatment itself. Specifically, couples that suffered a miscarriage might be scared of having a new pregnancy. Moreover, a miscarriage might be caused by problems that are likely to affect future pregnancies too. For these reasons, we think that the effect of a miscarriage on the number of the child is negative, i.e. $E[S|D=1] < E[S|D=0]$, and the casual effect is:

$$ATE = \beta + \gamma(E[S|D=1] - E[S|D=0]) \tag{18}$$

The sign of the bias then will be positive if $\gamma < 0$ and negative otherwise. It is difficult to evaluate what is the impact of the number of child on parenting style-associated outcomes. On one hand parents can be more "expert" and can transfer their experience accumulated with the first children on the subsequent ones. On the other, parents may have less time to take care of more children. We consider the latter to be less relevant, then we expect γ to be positive. For this reason, the bias will be negative and estimating this equation will underestimate the treatment effect.

c) Propose tests that could potentially show that the bias is quantitatively unimportant (after all, no identification strategy is perfect; so showing that a bias does not matter is often what is needed). Please be brief here.

The likelihood of having a second child is possible to vary with the treatment. We should then compare the distribution of family-size (number of children) in both groups. If the average frequency of two-children families across the two groups in our data set is similar, we can restrict the sample to two-children families and we can state that the bias is negligible.

Problem Set 1

```
rm(list=ls())
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
#Creating a dummy for female
set.seed(5)
Female<-ifelse(rnorm(1000)>0, 1, 0)
sum(Female)
```

```
## [1] 509
```

```
#Creating a dummy for dscirmination
a<-0.5
b<-2
set.seed(3)
Discrimination<-ifelse(a+b*Female+rnorm(length(Female), 1, 2)>2.8, 1, 0)
sum(Discrimination)
```

```
## [1] 446
```

```
#Creating a variable called Ability
set.seed(7)
Ability<-rnorm(length(Female), 16, 4)
summary(Ability)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.107  13.368  15.910  16.012  18.712  27.868
```

```
#Creating a dummy for occupation
c<-3
d<--1.2
e<-2.1
f<-0.25
set.seed(8)
Occupation<-ifelse(c+d*Discrimination+e*Female+f*Ability+rnorm(length(Female), 1, 2)>7, 1, 0)
sum(Occupation)
```

```
## [1] 717
```

```r
#setting parameters g , h , i and k to compare the estimates from the below regressions
g<-23.5
h<-5.8
i<--3.03
k<-12.8
set.seed(9)
Earnings<-g+h*Ability+i*Discrimination+k*Occupation+rnorm(length(Female), 16, 4)
summary(Earnings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   61.58  123.83  140.38  140.22  157.60  211.45
```

```r
#Regression for Y on D
stargazer(lm(Earnings~Discrimination), type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              Earnings
## -----------------------------------------------
## Discrimination              -5.893***
##                              (1.606)
##
## Constant                   142.849***
##                              (1.073)
##
## -----------------------------------------------
## Observations                  1,000
## R2                            0.013
## Adjusted R2                   0.012
## Residual Std. Error      25.246 (df = 998)
## F Statistic           13.465*** (df = 1; 998)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
#regression for Y on D + O as a control
stargazer(lm(Earnings~Discrimination+Occupation), type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              Earnings
## -----------------------------------------------
## Discrimination              -3.231**
##                              (1.426)
##
## Occupation                  26.543***
##                              (1.574)
##
## Constant                   122.630***
```

```
##                                  (1.527)
##
## -----------------------------------------------
## Observations                      1,000
## R2                                0.232
## Adjusted R2                       0.231
## Residual Std. Error       22.279 (df = 997)
## F Statistic           150.887*** (df = 2; 997)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
#regression for Y on D + O and A as a control
stargazer(lm(Earnings~Discrimination+Occupation+Ability), type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              Earnings
## -----------------------------------------------
## Discrimination               -3.189***
##                               (0.246)
##
## Occupation                   12.771***
##                               (0.282)
##
## Ability                       5.799***
##                               (0.032)
##
## Constant                     39.634***
##                               (0.531)
##
## -----------------------------------------------
## Observations                      1,000
## R2                                0.977
## Adjusted R2                       0.977
## Residual Std. Error        3.845 (df = 996)
## F Statistic         14,200.460*** (df = 3; 996)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
#Additional regressions
stargazer(lm(Earnings~Discrimination+Ability), type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              Earnings
## -----------------------------------------------
## Discrimination               -4.373***
##                               (0.428)
##
```

```
## Ability                      6.193***
##                               (0.054)
##
## Constant                     43.002***
##                               (0.918)
##
## -----------------------------------------------
## Observations                   1,000
## R2                             0.930
## Adjusted R2                    0.930
## Residual Std. Error      6.720 (df = 997)
## F Statistic         6,639.879*** (df = 2; 997)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
stargazer(lm(Earnings~Occupation+Ability), type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              Earnings
## -----------------------------------------------
## Occupation                   13.159***
##                               (0.303)
##
## Ability                      5.799***
##                               (0.035)
##
## Constant                     37.927***
##                               (0.555)
##
## -----------------------------------------------
## Observations                   1,000
## R2                             0.973
## Adjusted R2                    0.973
## Residual Std. Error      4.155 (df = 997)
## F Statistic         18,175.270*** (df = 2; 997)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```