Problem Set 1

This problem set is intended to shape your understanding of what a causal relation is and why '**correlation** does not imply **causation**'. We will the positive correlation between income and health (the *income health gradient*) discussed in lecture as an example. Suppose in the following we are interested in the causal effect of income on health and we know that the true relation is:

$$Health_{it} = \beta_0 + \beta_1 Income_{it} + \beta_2 Education_{it} + \beta_3 Age_{it} + \epsilon_{it} \tag{1}$$

Question 1 - Data Structures

A) **Question:** *What type of data would you need to estimate this equation? Why?*

Answer: The data is indexed by two subscripts: i refers to different data-points, labeled i, within the same time period and t refers to data from different time periods, labeled t. If the different i are individuals which stay the same over time (so we track the same individual over time) we speak of 'panel data' whereas if i refers to a sample from group i (where the composition of this sample might change over time) we speak of 'repeated cross-sectional data'. So we need one of the two.

B) **Question:** *If you had only uni-dimensional data, what estimating equation would you use? Either*

$$Health_t = \beta_0 + \beta_1 Income_{it} + \beta_2 Education_t + \beta_3 Age_{it} + \epsilon_t$$

or

$$Health_t = \beta_0 + \beta_1 Income_t + \beta_2 Education_t + \beta_3 Age_t + \epsilon_t$$

Answer: If we have only uni-dimensional data we can only estimate the second equation. In the first equation some of the data has two subscripts and hence would require two-dimensional data.

Question 2 - Data Quality

You get your data and while you can perfectly measure income, you are unable to measure education and age perfectly.

A) **Question:** *Algebraically, show what the inconsistency on the OLS estimate would be if you just run a regression of health on income, so you omit education and age. In what direction do you expect this inconsistency to go? (Assume for this question that the income measure is independent of the error term.)*

Answer: Define for convenience $\mathbf{y}$, $\mathbf{x}$, $\mathbf{e}$ and $\mathbf{a}$ as the vectors containing the data for health, income, education and age, respectively. Further, suppose we converted all data in mean-deviations form. We know that when using this data without including a constant term we get numerically the same coefficient estimates as when using the original data and including a constant. Since we are not interested in the constant here, we can use the data in mean-deviations form and omit the constant, which makes the algebra easier.

Then when calculating the mentioned OLS estimate we calculate

$$\hat{\beta}_1 = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \tag{2}$$

We plug in (as usual) for $\mathbf{y}$. You arrive at:

$$\begin{aligned}\hat{\beta}_1 - \beta_1 &= \beta_2 \cdot (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{e} + \beta_3 \cdot (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{a} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\epsilon \\ &= \beta_2 \cdot \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1}\frac{\mathbf{x}'\mathbf{e}}{N} + \beta_3 \cdot \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1}\frac{\mathbf{x}'\mathbf{a}}{N} + \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1}\frac{\mathbf{x}'\epsilon}{N}\end{aligned}$$

To calculate the inconsistency we take plims. Note that $\mathbf{x}'\mathbf{x} = \Sigma_i x_i^2$ and similarly for $\mathbf{x}'\mathbf{e}$ and $\mathbf{x}'\mathbf{a}$. Note that $plim\frac{\Sigma_i x_i^2}{N} = E(x_i^2) = Var(x_i)$ since $E(x_i) = 0$ because every is in mean-deviations form.[1] This same argument holds for writing $plim\frac{x'a}{N} = Cov(x,a)$, $plim\frac{x'e}{N} = Cov(x,e)$ and $plim\frac{x'\epsilon}{N} = Cov(x,\epsilon) = 0$. Hence we can rewrite

$$plim(\hat{\beta}_1 - \beta_1) = \beta_2 \cdot \frac{Cov(x,e)}{Var(x)} + \beta_3 \cdot \frac{Cov(x,a)}{Var(x)} \tag{3}$$

$Var(x)$ will be strictly positive (if it was 0 we could not calculate the OLS estimator) and hence the bias depends on both the covariance terms and the true coefficients. Presumably $\beta_2 > 0$, $\beta_3 < 0$, $Cov(x,e) > 0$ and $Cov(x,a) > 0$. So the first part is positive whereas the second part is negative. Hence, since we don't know the exact quantities involved, the sign of the bias is indetermined.

B) **Question:** *While you are not able to measure education and age perfectly, you are able to do so with some error. The error is random so that you know what you observe is:*

$$\tilde{a}_i = a_i + \nu_i \tag{4}$$

and

$$\tilde{e}_i = e_i + \upsilon_i \tag{5}$$

What is the bias in your OLS estimate from using these measures? (Consider in turn that you only measure age with error and that you only measure education with error.)

Answer: Let us consider the case where age is measured with error and eduction is not in the regression. Again convert all data into mean-deviations form. We can then rewrite the true relation as

$$\mathbf{y} = \beta_1\mathbf{x} + \beta_2\tilde{\mathbf{a}} + \epsilon - \beta_2\nu \tag{6}$$

Note that $\tilde{\mathbf{a}}$ and ν are surely correlated! So when calculating OLS

$$\hat{\beta}_{OLS} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\tilde{\mathbf{a}} \\ \tilde{\mathbf{a}}'\mathbf{x} & \tilde{\mathbf{a}}'\tilde{\mathbf{a}} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{x}'(\epsilon - \beta_2\nu) \\ \tilde{\mathbf{a}}'(\epsilon - \beta_2\nu) \end{pmatrix}$$

the plim of the 2nd element of the last vector is surely not 0.

This result shows how having measurement error in one variable can render **all** coefficient estimates biased! The intuition is the following: Suppose the mis-measurement is very severe such that the age variable you have is mostly noise. If there is some covariance between your variables of interest (here: income) and the mis-measured variable (here: age) then the coefficients on income will 'jump in' for the effect of age, or: the effect of higher age will be attributed to be an effect of higher income. Note that this only works if the true age and income are correlated.

Further you can understand why the measurement error bias on the coefficient of the mis-measured variable is downwards ('attenuation bias'): If it becomes pure noise, it will not explain anything and the coefficient on it will be 0 on average.

[1] We assumed here that all x_i have the same variance.

C) **Question:** *Compare your predicted bias in A and B. Discuss which approach you would take and why.*

Answer: Another way of asking the question would be to say: Suppose we have a mis-measured variable. Should we include it or should be drop it? Firstly: Both will lead to a bias, the measurement error bias or the omitted variable bias, respectively. So you have the choice between two poor options. Which one you choose might depend on

- whether one can sign the bias,
- the extent to which one prefers upward or downward biased estimates,
- the relative size of the measurement error in the variables and
- the strength of the correlation between the omitted variables and both income and health.

Question 3 - Causal Relationship

Question: *For each of the factors listed above (education, occupation, age, sex, marital status and ethnicity) discuss the way in which the correlation of these issues might affect your ability to estimate the causal effect of income on health. For each factor, assume it is the ONLY omitted variable in regression.*

Answer: Suppose the correct specification was (forgetting about sex and ethnicity)

$$y_i = \beta_1 \cdot x_i + \beta_2 \cdot e_i + \beta_3 \cdot o_i + \beta_4 \cdot a_i + \beta_5 \cdot m_i + \varepsilon_i \tag{7}$$

where y_i is some measure of the health of individual i and x_i is individual i's income. Now suppose we omit (because we just forget or we do not have data) each of education (e_i), occupation (o_i), age (a_i) and marital status (m_i) in turn. Then we have shown in class (for a simpler case, however) that the sign of the bias on $\hat{\beta}_1$ will depend on $\beta_j, j \in 1, 2, 3, 4$ and the covariance of the omitted variable with x_i. I would expect (though arguments different from mine might predict something else and be more convincing) that

- **education** influences health positively (better access to information) and is positively correlated with income (higher marginal productivity) - resulting in an upward bias, so we overestimate the coefficient β_1 on average,
- **occupation** might measure whether you work in a mine, which would influence health negatively and occupation would be negatively correlated with income - resulting again in an upward bias,
- **age** would probably influence health negatively and be correlated positively with income, - resulting in an downward bias, so we underestimate the coefficient β_1 on average,
- **marital status** would influence income positively (because of tax discount, e.g., or assortative matching) and be maybe positively related to health (because of some assortative matching in the 'marriage market') - resulting in an upward bias, so we overestimate the coefficient β_1 on average.

Note how these biases make intuitive sense. For example consider the downward bias induced by the omission of **age**: If indeed older people earn more but have a poorer health status, then when you regress health on income without having age in the regression, you might even find that "having more income decreases your health status". But this would only be concluded since you have forgotten a_i in the regression. If you had included it the coefficient on income would measure how higher income affects health *given the same age level*!

By the way: if you were to write a paper on this, this whole design would probably not get you very far. People would continue to tell you which variables you should have included (and some of these might not be measurable) or they might tell you that a good health status itself might cause a higher income. In any case even the weakest version of **A3** would be violated.

Question 4 - Data Interpretation

Some scholars have used cross-country comparisons between the US and UK to address some of these concerns. The argument they provide is that the main excluded factor is access to medical care, which varies dramatically by education and income in the US. In the UK, the NHS provides free insurance so this is not a problem.

A) **Question:** *What is the assumption such a strategy would make to "identify" the causal relationship of income on health? Does this make sense?*

Answer: Clearly one way income influences health is through access to health care. Now, if you are not interested in this part of the effect of income on health, but rather want to measure the sum of all other effects of income on health (whatever the exact channel, e.g. better nutrition, might be) then the UK data seems useful since the access to health care is theoretically the same for citizens of all income levels.

You might then be tempted to us the UK data and just run a regression of health status on income and some other controls. However, this would only work under the assumption that no other controls should be included and there is no reverse causation - and this assumption needs to be a reasonable one.

B) **Question:** *If you believe the identifying assumption of this paper, does this evidence support a causal effect of income on health? Why or why not? (Please look at the question sheet for the graphs.)*

Answer: If you believe that there are no omitted variable or reverse causality problems ('identifying assumption', cause it ensures that **A3** holds), then you would conclude that income has an effect on health independent of access to health care. (We did not run the regression, but from the graph it seems the coefficient would be positive and significant.)

By the way, if the only thing different between the US and the UK was the access to health care, the lower slope in the UK would actually indicate that access to health care might be a channel through which income effects health status - hence in the US the total effect of income on health status is bigger.

C) **Question:** *What else might explain the observed similarities and differences between the US and UK?*

Answer: Again, all the arguments made above would not make a good paper. The similarities, i.e. that income and health are positively correlated, might for example be driven by the omitted 'ability' of the individual (that is to say: cleverness). This might drive up both income and health and cause the correlation observed in the data. The problem is this 'ability' is inherently difficult to measure. And even if, somebody would probably come up with another omitted variable. Can you think of one?

And the differences between the US and UK might be driven by the fact that other features of the health system apart from free access are different in the UK and US, too. Or you might argue that the upper two or three deciles (which are on the x-axis in the graph) have actually a higher income in the US than in the UK - so we actually compare apples with pears.

D) **Question:** *Can you think of a test to distinguish between the theory in A and the theory in B? Please describe?*

Answer: If we wanted to check our 'identifying assumption' , we might just take one omitted variable, say education, and split the sample into those with high education and low education and see how low education guys have poor health and income and high education guys have good health and good income.

