Question 3 - Causal Relationship

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.

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

$$y_i = \beta_1 \cdot x_i + \beta_2 \cdot e_i + \beta_3 \cdot o_i + \beta_4 \cdot a_i + \beta_5 \cdot m_i + \varepsilon_i \tag{1}$$

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. What is the assumption such a strategy would make to "identify" the causal relationship of income on health? Does this make sense?

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.

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

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. What else might explain the observed similarities and differences between the US and UK?

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. Can you think of a test to distinguish between the theory in A and the theory in B? Please describe?

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.