**Question 1.** Consider the following causal claims:

- "He was hired for the job because he is a man. A woman would be less likely to get the job."
- "He was hired for the job because he is reliable."
- "He was hired for the job because he had a university degree."

What criteria can you use to determine if these are valid causal claims? If it is a valid claim, how could you estimate the claimed causal effect?

**Question 2.** Consider a world an outcome, $Y$, a binary self-selected treatment $T = 1$ for treated and $= 0$ for not treated, and a randomly assigned encouragement to treatment, $W = 1$ for encouraged and $= 0$ for not encouraged. Consider the claim: "Because the $W$ is randomly assigned, it is by construction a valid instrument for $T$. True or false?

**Question 3.** A doctor considers a patient's age ($A$), blood pressure ($B$), and blood sugar ($S$), and only these three, before assigning the patient to treatment ($T = 1$ as opposed to $= 0$, no treatment). If the outcome of interest is $Y$ then this implies that the regression

$$Y_i = a + bA_i + cB_i + dS_i + eT_i + \epsilon_i$$

will yield an unbiased estimate of the treatment effect. True or false?

**Question 4.** In the paper "The Costs of Remoteness: Evidence from German Division and Reunification", the researchers investigate the importance of market access for economic development. They use as a natural experiment the construction of the Iron Curtain which divided Germany into Eastern and Western parts, between which all trade was stopped.

The sample consists of all West German cities with a population of more than 20,000 in a base year. The researchers have observations on a measure of economic development (denoted by $y$) before and after the construction of the Iron Curtain. Their treatment group is cities within 75km of the border (which are presumed to lose access to markets with division of Germany) and the control group is all other cities.

1. Explain why a simple comparison of $y$ in treatment and control groups after division is unlikely to provide a good estimate of the effect of losing market access.
2. What equation would you estimate with this data and what parameters of this equation tell you about the causal effect of interest?
3. You now also have more observations on $y$ both before and after the construction of the Iron Curtain. Why might this information be useful and how would you use it?

**Question 5.** Define the propensity score, $p(T|X)$ which is the probability that an individual receives treatment $T$ conditional on all observable characteristics $X$. If there is selection into treatment on unobservables then the matching assumption that $\{Y_1, Y_0\}$ is independent of $p(T|X)$ will be satisfied. True, false or uncertain? (MSc Only)

**Question 6.** An investigator considers the linear model:

$$\begin{aligned}\log Earnings_i = \gamma_1 + \beta_1 Age_i + \beta_2 FullTimeEmployed_i + \beta_3 Tenure_i \\ + \gamma_2 Education_i + \gamma_3 White_i + \gamma_4 Female_i + \epsilon_i\end{aligned}$$

where $FullTimeEmployed_i$ is a dummy variable indicating whether individual $i$ was employed full time in period $t$, $White_i$ is a dummy indicating whether individual $i$ is of white race, and $Female_i$ is a dummy taking the value 1 if individual $i$ is female. The available sample of a cross-section of individuals is indexed by $i = 1, ..., N$. (Optional, but needed to do below questions. Will be discussed in following lectures)

1. Explain what would happen if instead of the variables $White_i$ and $Female_i$ were to use the complementary variables $Non-White_i$ and $Male_i$ defined in the obvious way.

2. Suppose we define the interaction variables $Z_i \equiv Age_i \times White_i$ and $W_i \equiv Age_i \times Female_i$. What would we achieve by introducing these two variables as additional regressors?

3. Discuss possible reasons why the regressors may violate exogeneity assumptions with respect to the error term, i.e., regressors and disturbances may be statistically related. Which regressors do you consdier the most suspect in this regard?

**Question 7.** Suppose we are interested in the effect of treatment $T$ on $Y$. We estimate a regression of the form

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 T_i + \hat{\epsilon}_i$$

$Y$ is a continuous outcome variable and $T$ is a discrete treatment indicator that is 1 if an individual received the treatment and 0 otherwise. Suppose that you estimate $\hat{\beta}_1$ using OLS.

1. Show that can be decomposed two components, the treatment on the treated (TOT), and a selection bias term (SB) (Hint: TOT $= E\left[Y_{1i} - Y_{0i} | T_i = 1\right]$ and SB $= E\left[Y_{0i} | T_i = 1\right] - Y_{0i}\left[T_i = 0\right]$ where $Y_{1i}$ is the outcome in the state of the world where an individual receives the treatment and $Y_{0i}$ is the outcome in the state of the world where an individual does not receive the treatment.)

2. Suppose treatment was conditionally randomly assigned based on some characteristic $X$. Must you include $X$ in the regression to ensure that $\hat{\beta}_1$ converges to true $\beta_1$? (MSc Only)

3. Now suppose that treatment was not randomly assigned but you find a control group which looks similar to your treatment group. You collect some more data so you can estimate a regression of the form:

$$Y_{it} = \hat{\beta}_0 + \hat{\beta}_1 T_i + \hat{\beta}_2 A_t + \hat{\beta}_3 (A_t \times T_i) + \hat{\epsilon}_{it}$$

   The variable $A$ is 0 at time $t = 1$ and 1 at time $t = 2$. The variable $T$ is 1 for the treatment group (in all periods) and 0 for the control group. The treatment occurs between time $t = 1$ and $t = 2$. Show how this regression estimates the effect of $T$ on $Y$. What assumptions does this estimation method imply?

4. Suppose you had 3 periods of data now, so that you could estimate a regression of the form

$$Y_{it} = \hat{\beta}_0 + \hat{\beta}_1 T_i + \hat{\beta}_2 A_{2t} + \hat{\beta}_3 (A_{2t} \times T_i) + \hat{\beta}_4 A_{1t} + \hat{\beta}_5 (A_{1t} \times T_i) + \hat{\epsilon}_{it}$$

   The variable $A_{2t}$ is 1 at time $t = 2$ and 0 otherwise. The variable $A_{1t}$ is 1 at time $t = 1$ and 0 otherwise. How does the coefficient $\hat{\beta}_5$ help you test the identifying assumptions from (3)?