Question 1 Assume there is a new anti-flu pill on the market and you are interested in understanding if it improves upon flu cases. Suppose that you observe that families of nurses and doctors take the pill. Can you compare health outcomes of family member of nurses and doctors who take the pill with family members of nurses and doctors who don't take the pill? Do you expect this comparison to be a valid one?

Assume now that the pill is given to 500 patients and a placebo pill is given to other 500 patients, using random assignment of patients to treatment. How would you estimate the treatment effect of the pill? Suppose that you have data on weight, gender of each patient, could you use these data to improve your estimate? Suppose you also measured the cholesterol level of each patient before assigning them to treatment. Could you say whether these data improve your estimate? Explain.

Answer Comparing families of nurses and doctors who take the pill (treatment group) with family member of nurses and doctors who do not take the pill (control group) is not a valid comparison as individuals who take the pill might be different from those who do not. For example, it might be that families of doctors who take the anti-flu pill are more likely to get frequent flu cases hence, on average, less healthy than the comparison group.

Random assignment of patient to treatment solves the selection problem. Making the control group take a placebo ensures that treatment and control groups are as similar as possible. In this way we avoid that patients in the treatment group realize that they are part of an experiment and start to behave differently, i.e start to eat more healthy, exercise etc. (This is problematic because if patients start to behave differently then we do not know whether this is the effect of the pill or healthier lifestyle.) In case of random assignment we just need to compare means (which would be the equivalent of the regression of health outcomes on a dummy taking value one if the patient is in the treatment group).

Data on weight, gender of each patient should not affect the sign of the coefficient estimate, however, they can help us to estimate the causal effect of the pill on health outcomes with more precision, lower standard errors on the coefficient estimate. Same reasoning applies to the cholesterol level. In case of random assignment we expect mean differences on these covariates to be balanced among treatment and control group, no systematic differences between treatment and control.
For deeper understanding about this question look at AP Chapter 2 and lecture notes.

Question 2 Consider the following hypothetical example: some schools in California are forced to close for repairs due to a summer earthquake. District closer to the epicenter are most severely affected. A district with some closed schools need to "double up" their students temporarily increasing class size. This is an example of a quasi-experiment. How would you use this quasi-experiment to estimate the effect of class size on test scores?

Answer In this case we could use distance from the epicenter as an instrument for class size. Distance from the epicenter satisfies the condition for instrument relevance as it is correlated with class size. The idea is that the instrument affect test scores only through class size and should be unrelated to any other factors that might affect student performance. If this is the case, then the instrument is exogenous as it is uncorrelated with the error term. Thus the instrument could use to estimate the effect of class size on test scores.
See below for a more detailed answer on the conditions for instrument validity AP Chapter 4, Wooldridge Chapter 15 and Lecture Notes.

Question 3 In an article, Evans and Schwab (1995) studied the effects of attending a Catholic high school on the probability of attending college.

Answer For concreteness, let college be a binary variable equal to unity if a student attends college, and zero otherwise. Let $CathHS_i$ be a binary variable equal to one if the student attends a Catholic high school. The model to estimate is:

$$College_i = \alpha + \beta CathHS_i + X_i + u_i$$

where X_i includes gender, race, family income, and parental education.

(i) Why might $CathHS_i$ be correlated with u_i?

Answer Individuals who attend catholic high school may have different set of values, motivation and attitude towards studying (maybe intrinsic motivation or family expectations) which can affect both high school choice – i.e choosing catholic high school – and applying to college.

(ii) Evans and Schwab have data on a standardized test score taken when each student entered high school. What can be done with these variables to improve the ceteris paribus estimate of attending a Catholic high school?

Answer It may be used as a proxy for latent skills (ability), which may also be a cause of endogeneity, if the process of getting into Catholic high schools is more demanding. Correlation between the test score and ability is probably positive.

(iii) Let $CathRel_i$ be a binary variable equal to one if the student is Catholic. Discuss the two requirements needed for this to be a valid IV for $CathHS_i$ in the preceding equation. Which of these can be tested?

Answer The two conditions to meet are the relevance and exclusion restrictions. Relevance means a non zero correlation between religion and Catholic high school attendance. This condition can be tested.

The exclusion restriction is satisfied if being a Catholic would only affect college attendance only through attending a catholic high school. It could not have a direct effect on admission rates - e.g. due to college principals admitting more Catholic students independent of their performance. Nor could be there confounders related to both being a Catholic and getting admitted to college.

(iv) Not surprisingly, being Catholic has a significant effect on attending a Catholic high school. Do you think $CathRel_i$ is a convincing instrument for $CathHS_i$?

Answer When thinking of whether the instrument is a convincing one we should think of how the instrument might affect the outcome variable, $College_i$, directly or how it might be potentially correlated with the error term. (see in (iii) to see how exclusion restriction might be violated). If the exclusion restriction is violated, we say that the instrument is not valid.

Note that this is not the same as the case of weak instruments. We talk about weak instrument in cases when the correlation between the instrument and the endogenous variable is low. As a rule of thumb for the instrument to be considered relevant, the F-statistics of the first stage of the excluded instruments should be less than 10.

See Chapter 4 of AP, Wooldridge Chapter 15 and lecture notes.

Question 4 Consider the problem of estimating the effect of family income on college grade point average. It could be that, though family income is important for performance before college, it has no direct effect on college performance. To test

this, we might postulate the model:

$$colGPA_i = \beta_0 + \beta_1 faminc_i^* + \varepsilon_i$$

where $faminc^*$ is actual annual family income. Family income is self reported by students, hence could be easily measured with error.
Assume that we are in the case of classical measurement error. What are the consequences of the measurement error when estimating β_1?

Answer The consequences of the classical measurement error is attenuation bias in the parameter of interest. If we would observe the real family income, $faminc^*$, then we would just estimate β with OLS. However this is not the case as family income is self reported hence likely to be measured with error.

$$faminc = faminc_i^* + u_i$$

we observe $faminc$ which is a function of the real family $faminc^*$ and some measurement error. Given that we are in the classical measurement error case, then we know that:

$$u_i \sim iid(0, \sigma_u^2)$$
$$Cov(faminc^*, u_i) = 0$$

$$\begin{aligned} colGPA_i &= \beta_1 faminc_i^* + \varepsilon_i \\ &= \beta_1(faminc_i - u_i) + \varepsilon_i = \beta_1 faminc_i \underbrace{-\beta_1 u_i + \varepsilon_i}_{\tilde{\varepsilon}_i} \\ colGPA_i &= \beta_1 faminc_i + \tilde{\varepsilon}_i \qquad \text{estimated model} \end{aligned}$$

Estimating the model with OLS gives using the observed family income gives an underestimate of the true parameter β_1. To see this note that:

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{Cov\left(colGPA_i, faminc_i\right)}{Var\left(faminc_i\right)} \\
&= \frac{Cov\left(\beta_1 faminc_i + \tilde{\varepsilon}_i, faminc_i\right)}{Var\left(faminc_i\right)} \\
&= \frac{\beta_1 Cov\left(faminc_i, faminc_i\right)}{Var\left(faminc_i\right)} + \frac{Cov\left(\tilde{\varepsilon}_i, faminc_i\right)}{Var\left(faminc_i\right)} \\
&= \underbrace{\beta_1}_{A} + \underbrace{\frac{Cov\left(\tilde{\varepsilon}_i, faminc_i\right)}{Var\left(faminc_i\right)}}_{B}
\end{aligned}$$

The numerator in B can be written as:

$$\begin{aligned}
Cov\left(faminc_i, \tilde{\varepsilon}_i\right) &= Cov\left(faminc_i, -\beta_1 u_i + \varepsilon_i\right) = Cov\left(faminc_i, -\beta_1 u_i\right) + Cov\left(faminc_i, \varepsilon_i\right) \\
&= -\beta_1 Cov\left(faminc_i, u_i\right) = -\beta_1 Cov\left(faminc_i^* + u_i, u_i\right) \\
&= -\beta_1 \sigma_u^2
\end{aligned}$$

Hence, substituting we get that:

$$\begin{aligned}
&= \beta_1 - \beta_1 \frac{\sigma_u^2}{Var\left(faminc_i\right)} \\
&= \beta_1 \left[1 - \frac{\sigma_u^2}{\sigma_u^2 + Var\left(faminc_i^*\right)}\right]
\end{aligned}$$

The solutions to this example follows exactly as in the lecture notes and Wooldrigde Chapter 9.