1. Consider a single model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:

$$GPA = \beta_0 + \beta_1 PC + u$$

where PC is a binary variable indicating PC ownership.
(i) Why might PC ownership be correlated with u ?
(ii) Explain why PC is likely to be related to parent's annual income. Does this mean parental income is a good IV for PC? Why or why not?
(iii) Suppose that, four years ago, the university gave grants to buy computers to roughly one half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for PC.

(i) There are many potential factors that influence GPA. The effects of these potential regressors have to be captured wither by PC or u. If any of these is also correlated to PC (such as family income), the error term could also be correlated with PC.

Ai - you can always argue an omitted variable; give a story; e.g., PC could be correlated to family income; higher income makes it more likely to have private tutors; So maybe what's really going on is GPA being explained by tutors, not PC.

(ii) Parents who have higher income probably also have higher disposable income and can afford to buy a PC for their kids at college. For a good IV, we want something that is correlated with PC, but not correlated with u. In the case of income, there is still a chance that it is correlated with u. For example, income could be correlated with regional effects (better public school districts) which are captured by u.

Ai - income could be correlated to parent's education (omitted variable) or some other unobserved characteristic of the student (e.g., may be more motivated [or pressured] in school)

(iii) Satisfies 2 conditions: correlated to PC and not correlated to error term (since being selected for the grant was random, it shouldn't be related to any other parameters, so even if they are omitted and correlated to the error term, $Grant$ will not be)

Define $\mathbf{x}_i = \begin{bmatrix} 1 \\ PC_i \end{bmatrix}$, $\mathbf{z}_i = \begin{bmatrix} 1 \\ Grant_i \end{bmatrix}$

Three options:

1. **Directly** - $\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$

2. **2SLS** -

   a. Regress PC on $\mathbf{z}_i$

   b. Generate $\hat{PC} = \hat{\delta}_0 + \hat{\delta}_1 Grant$

   c. Regress GPA on $\hat{PC}$

3. **Stata** - `ivreg GPA (PC = GRANT)`

Ai - another potential instrument would be PC price (determined by market so it's probably not related to other factors that are student dependent)

2. In a recent article, Evans and Schwab (1995) studied the effects of attending a Catholic high school on the probability of attending college. For concreteness, let $college$ be a binary variable equal to unity if a student attends college, and zero otherwise. Let $CathHS$ be a binary variable equal to one if the student attends a Catholic high school. A linear probability model is

$$college = \beta_0 + \beta_1 CathHS + other\_factors + u$$

where the other factors include gender, race, family income, and parental education.
(i) Why might $CathHS$ be correlated with u?
(ii) Evans and Schwab have data on a standardized test score taken when each student was a sophomore. What can be done with this variable to improve the ceteris paribus estimate of attending a Catholic high school?
(iii) Let $CathRel$ be binary variable equal to one if the student is Catholic. Discuss the two requirements needed for this to be a valid IV for $CathHS$ in the preceding equation. Which of these can be tested?
(iv) Not surprisingly, being Catholic has a significant effect on attending a Catholic high school. Do you think $CathRel$ is a convincing instrument for $CathHS$?

(i) Reasons for regressor being correlated to error term: (a) simultaneous decision [LHS and RHS variables being jointly determined], (b) omitted variable, or (c) constraint relating LHS and RHS variables. Given that Catholic schools are private, a student who attends one probably have parents who are more concerned about their child's education and will push harder for them to attend college. In such a situation, one could argue that $college$ and $CathHS$ are jointly determined.

Ai - possible omitted variable for ability; "self-select"... better students go to private schools

(ii) If we assume students at Catholic high schools score better (or worse) on average than other students, we may be able to use the standardized test score as an instrumental variable for $CathHS$. The score is not jointly determined by the parents so it may solve the problem discussed in (i).

(iii) Two requirements is the IV being (highly) correlated to the regressor and being uncorrelated to the error term. The first one can be tested by regressing $CathHS$ on $CathRel$ and look for R^2 > 0.1 and significant coefficient on $CathRel$. Also want to check the impact (magnitude) of the coefficient on $CathRel$ (i.e., check size becase even if it's significant at 99.99%, a value of 0.1 doesn’t mean much)

Ai - Can't test the second one unless we have another instrument that we know is good; then model is over identified and we can use the Hausman test

(iv) No. We have to consider the direction of the relationship... there percentage of students who attend Catholic high school that are Catholic may be high, but the percentage of Catholics who attend Catholic high school may not be. (Kind of the smoking-lung cancer problem... % how have lung cancer that smoke is high, but not the other way around.)

Ai - $CathRel$ may be related to error term... didn't really cover why

3. For a large university, you are asked to estimate the demand for tickets to women's basketball games. You can collect time series data over 10 seasons, for a total of about 150 observations. One possible model is

$$\ln attend_t = \beta_0 + \beta_1 \ln price_t + \beta_2 winperc_t + \beta_3 rival_t + \beta_4 weekend + \beta_5 t + u_t$$

where *price* is the price of admission, probably measured in real terms, *winperc*, is the team's current winning percentage, *rival*, is a dummy variable indicating a game against a rival, and *weekend*, is a dummy variable indicating whether the game is on a weekend.
(i) Why is it a good idea to have a time trend in the equation?
(ii) The supply of tickets is fixed by the stadium capacity; assume this has not changed over the 10 years. This means that quantity supplied does not vary with price. Does this mean that price is necessarily exogenous in the demand equation?
(iii) Suppose that the nominal price of admission changes slowly. The athletic office chooses price based partly on last season's average attendance, as well as last season's team success. Under what assumptions is last season's winning percentage a valid instrumental variable for price?
(iv) Does it seem reasonable to include the (log of the) real price of men's basketball games in the equation? Can you think of another variable related to men's basketball that might belong in the women's attendance equation?
(v) If some games are sold out, what problems does this cause for estimating the demand function?

(i) Demand grows over time because of population growth. Since there is no variable for population in the model, including time may work (assuming steady, linear population growth).
Ai - *t* may capture macroeconomic events: population growth, income growth over time, bigger pool for alumni

(ii) Exogenous means $E[price \cdot u] = 0$ (i.e., uncorrelated to error term); since capacity is fixed, Q^D does not have to equal to Q^S, but school is still trying to maximize profit which is based on the capacity $\therefore$ price is not exogenous

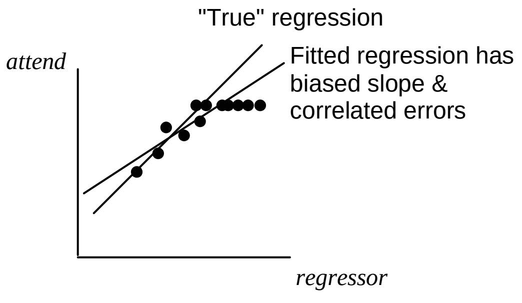
(iii) Want $winperc_{t\text{-}1}$ to be correlated to $price_t$, but not to error term u_t

$$\ln attend_t = \beta_0 + \beta_1(\delta_0 + \delta_1 attend_{t-1} + \delta_2 winperc_{t-1}) + \beta_2 winperc_t + \beta_3 rival_t + \beta_4 weekend + \beta_5 t + u_t$$

Ai - $price_t$ depends on $attend_{t\text{-}1}$ and $winperc_{t\text{-}1}$

(iv) The price of men's basketball games could make sense in the sense that men's games could be viewed as a substitute for the women's games. Unless the games are on the same night, however, the correlation may not be as strong. Another variable related to men's basketball that would be better is a binary variable: 1 if there is a men's game (home or away) at the same time as the women's game.
Ai - relative winning percentage of men vs. women (i.e., which team is doing better)

(v) Linear regression wouldn't work well because *attend* would not be linear... it will result in biased coefficient estimates and possibly correlated error terms



4. Discuss test and correction for heteroskedasticity and error term correlation in the 2SLS framework.

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$

Heteroskedasticity in 2SLS - i.e., $E(u_i^2 \mid \mathbf{z}_i) \neq \sigma^2$

Detecting -

(1) run 2SLS and get $\hat{\boldsymbol{\beta}}$

(2) compute consistent residuals: $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

(3) regress e_i^2 on $(1\ \mathbf{z}_i)$ (i.e., be sure to include a constant term if it's not already in $\mathbf{z}_i$)

(4) do overall test of significance (i.e., standard F-test to check if all parameters are simultaneously equal to zero)... if regression is significant, there's heteroskedasticity

Correcting -

(1) save fitted value of $\hat{e}_i^2$ (from regression in step (3) above)

(2) transform model: $\dfrac{y_i}{\hat{e}_i} = \dfrac{\mathbf{x}_i{}'}{\hat{e}_i}\boldsymbol{\beta} + \dfrac{u_i}{\hat{e}_i} = \beta_1 \dfrac{x_{1i}}{\hat{e}_i} + \hat{\beta}_2 \dfrac{x_{2i}}{\hat{e}_i} + \hat{\beta}_3 \dfrac{x_{3i}}{\hat{e}_i} + \dfrac{u_i}{\hat{e}_i}$

(3) do 2SLS on the transformed model; can use $\mathbf{z}_i = \begin{bmatrix} w_i \\ x_{2i} \\ x_{3i} \end{bmatrix}$ or $\mathbf{z}_i = \begin{bmatrix} w_i / \hat{e}_i \\ x_{2i} / \hat{e}_i \\ x_{3i} / \hat{e}_i \end{bmatrix}$... will give different results, but both have same statistical properties

Serial Correlation in 2SLS -

Detecting -

(1) same (1) and (2) as heteroskedasticity

(3) run $e_i = \rho e_{i-1} + \boldsymbol{\gamma}_i$ (or any other form); if $\hat{\rho}$ is significantly different than zero, there's serial correlation

Correcting -

(1) transform model:

$$(y_i - \hat{\rho} y_{i-1}) = \beta_1 (x_{1i} - \hat{\rho} x_{1i-1}) + \beta_2 (x_{2i} - \hat{\rho} x_{2i-1}) + \beta_3 (x_{3i} - \hat{\rho} x_{3i-1}) + (u_i - \hat{\rho} u_{i-1})$$

(2) do 2SLS on the transformed model; can use $\mathbf{z}_i - \hat{\rho}\mathbf{z}_{i-1}$, $\mathbf{z}_i$, or $\mathbf{z}_{i-1}$ (will have same statistical properties)