Econometrics Question 1

1.

If we estimate the regression $score = b_0 + b_1 private + u$ using OLS, the estimator of b_1 is very likely to be biased. We are omitting important variables such as family income ($faminc$) or ability.

For the omitted variable $faminc$, the bias is $b_{faminc} \cdot \frac{Cov(private, faminc)}{Var(private)}$.

$Cov(private, faminc) > 0$ because kids from rich families are more likely to go to private schools. $b_{faminc} > 0$ even when the type of school is kept constant. Maybe kids from poor families have family responsibilities and less time for homework. The omitted variable bias is positive.

For the omitted variable $ability$, the bias is $b_{ability} \cdot \frac{Cov(private, ability)}{Var(private)}$.

$b_{ability} > 0$ even when the type of school is kept constant. $Cov(private, ability) > 0$ for several reasons. Private schools give some financial aid to students of high ability. There could be a selection mechanism where parents of capable students make efforts to send them to private schools. And finally, if ability is genetic and parents of high ability are richer on average they will have smarter kids and will also afford private schools. The omitted variable bias is again positive.

The OLS estimate in the regression $score = b_0 + b_1 private + u$ is biased upwards.

2.

The omitted variable bias in the regression $score = b_0 + b_1 private + u$ can be eliminated by using $voucher$ as an instrumental variable. $voucher$ is a valid instrument because the two conditions apply. $Cov(voucher, private) > 0$ as a fraction of the students who receive vouchers go to private schools. Also, $Cov(voucher, u) = 0$ because the voucher assignments are completely random.

To construct an estimator for b_1, use the 2SLS procedure:

First stage: OLS regression $private = p_0 + p_1 voucher + v$ Fitted values $\widehat{private} = \hat{p}_0 + \hat{p}_1 voucher$.

Second stage: OLS regression $score = b_0 + b_1 \widehat{private} + u$

The OLS estimator of the second regression is the 2SLS estimator $\hat{b}_1^{2SLS}$, which is unbiased as long as the instrument is valid.

3.

The 2SLS estimator is biased if the instrumental variable is correlated with the error term. If the true regression is $score = b_0 + b_1 private + b_2 \, faminc + b_3 ability + v$ but we only include the $private$ variable in the regression, then the error term would be $u = b_2 \, faminc + b_3 ability + v$. In this case, $Cov(voucher, u) = Cov(voucher, b_2 \, faminc + b_3 ability + v) = b_2 Cov(voucher, \ faminc) + b_3 Cov(voucher, ability) + Cov(voucher, v)$. Since $Cov(voucher, \ faminc) < 0$ the exogeneity assumption is unlikely to hold, so $Cov(voucher, u) \neq 0$. Then the 2SLS estimator is biased.

$$\hat{b}_1^{2SLS} \approx \frac{Cov(voucher, score)}{Cov(voucher, \ private)} = \frac{Cov(voucher, b_0 + b_1 private + u)}{Cov(voucher, \ private)} = b_1 + \frac{Cov(voucher, u)}{Cov(voucher, \ private)}$$

Suppose we also have observations on $faminc$. Then we can just include it in the regression $score = b_0 + b_1 private + b_2 \, faminc + u$ and use $voucher$ as an instrumental variable for $private$. The 2SLS estimator is derived as follows:

First stage: OLS regression $private = p_0 + p_1 voucher + p_2 \, faminc + v$

Get the fitted values $\widehat{private} = \hat{p}_0 + \hat{p}_1 voucher + \hat{p}_2 \, faminc$.

Second stage: OLS regression $score = b_0 + b_1 \widehat{private} + b_2 \, faminc + u$

This new 2SLS estimator is unbiased as long as:

- Conditional on $faminc$, $voucher$ is randomly distributed, and
- $faminc$ is an exogenous variable (not correlated with $ability$.)

4.

(a) Only a small fraction of those who receive vouchers actually use them (5%), which explains why the effect of being given a voucher is small. Vouchers are good, but parents need to be persuaded to actually use them to send kids to private school. A politician needs to tell us how much money he is willing to spend for a given increase in test scores and then we will perform a cost/benefit analysis.

(b) Since only a small fraction of those who receive vouchers actually use them (5%), the effect of receiving a voucher is significant, but of small value. This does not contradict the fact that actually using the voucher to go to a private school has a large causal effect on scores.

Econometrics Question 2

1. Provide a derivation of the omitted variables bias formula

Imagine the *true* model is:
$Y = b_0 + b_1 X_1 + b_2 X_2 + u$
Assumptions: (1) $E(u_i \,|\, X) = 0$, (2) iid observations, (3) no perfect multicollinearity.

But we forgot about variable X_2, or we don't know how to measure it.
So we run the following regression instead:
$Y = g_0 + g_1 X_1 + v$

The OLS estimator of g_1 is $\hat{g}_1^{OLS} = \dfrac{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2} \approx \dfrac{Cov(X_1, Y)}{Var(X_1)}$.

The covariance term can be written as:

$$Cov(X_1, Y) = Cov(X_1, b_0 + b_1 X_1 + b_2 X_2 + u) =$$
$$= Cov(X_1, b_0) + Cov(X_1, b_1 X_1) + Cov(X_1, b_2 X_2) + Cov(X_1, u)$$

$Cov(X_1, b_0) = 0$ because b_0 is a constant.
$Cov(X_1, b_1 X_1) = b_1 Var(X_1)$
$Cov(X_1, b_2 X_2) = b_2 Cov(X_1, X_2)$
$Cov(X_1, u) = 0$ by assumption (1).

$$\boxed{\hat{g}_1^{OLS} \approx b_1 + \frac{b_2 Cov(X_1, X_2)}{Var(X_1)}}$$

If $b_2 \cdot Cov(X_1, X_2) > 0$, the omitted variable bias is *positive.*
If $b_2 \cdot Cov(X_1, X_2) < 0$, the omitted variable bias is *negative.*
If $b_2 = 0$ or $Cov(X_1, X_2) = 0$, there is *no omitted variable bias*.

2. Provide an example to explain how panel data can be used to eliminate certain kinds of omitted variable bias

One example of panel data is the wage regression. We want to estimate the effect of education on wages, but it is quite likely there are omitted variables. If we fail to include a variable that is correlated with education and affects wages even when wages are kept constant, then the OLS regression will be bias. Ability might be such an omitted variable. The problem with ability is that it's also difficult to measure. We could try using a proxy such as the IQ but we are not very sure this would work well.

The panel data approach is to include dummy variables for individual ability in order to separate the fixed effect of ability on wages from the effect of education on wages. A good approach to do this is to observe pairs of twins over time. Of all people, twins are the most likely to have similar abilities. So we can use a dummy variable to each pair of twins and run the regression this way. If we also include a constant, one of the dummies need to be dropped to avoid perfect multicollinearity.

Panel data allows us to eliminate the effects of unobserved variables, as long as they remain constant through time. However, if the unobserved variables change through time, panel data will not completely eliminate the bias.

3. Consider the case of longitudinal data (such as repeated observations on the same firms over time.) Discuss the problems that arise in obtaining appropriate standard errors and confidence intervals, and some solutions to these problems

Consider the following time series regression:

$$Y_{t+1} = b_0 + \left(b_1 Y_t + ... + b_p Y_{t-p+1}\right) + \sum_{j=1}^{k} \left(g_{j1} X_{jt} + ... + g_{jq_j} X_{j,t-q_j+1}\right) + u_t \quad \text{for } t = 1,2,...,T$$

OLS estimation to work well if the following assumptions hold:
(1) $E(u \mid X, Y \text{ lags}) = 0$ (orthogonality)
(2) The Y 's and the X 's have stationary distributions
(3) No perfect multicollinearity

The most problematic is assumption (2), stationarity. This is not necessarily the case in many situations. The distribution of our variables may change over time due to several reasons. Typical examples are *trends* (persistent long-term movements over time) and *breaks* (regression parameters change over time.) If the distribution is non-stationary, one cannot use OLS directly. If we do, we will encounter the following problems:

- OLS estimators are biased
- OLS standard errors do not have normal distribution, so we cannot use $\pm 1.96s$
- Spurious regression: two unrelated variables might nevertheless appear to explain each other if they have trends.

Checking the existence of a trend is done using the Dickey-Fuller(1979), which tests the existence of a unit root. If the time series has a unit root, we can try to get rid of it using differencing. If the transformation is stationary we can again use OLS. However, there is no guarantee that differencing results necessarily in a stationary distribution. We can also check the existence of a break using the Chow test.