1. **Stock and Watson, question 12.10.** You are told that you have an instrumental variable regression model with \( Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i \), where \( Z_i \) is an instrument. You are then told that \( W_i \) is unavailable, but you go ahead and estimate the model omitting \( W_i \).

   a. In the first part of the question, you are asked to suppose that \( Z_i \) and \( W_i \) are uncorrelated and asked to determine if the IV estimator is consistent. The key to answering this question is to note that

   \[
   \hat{\beta}_{T\!S\!L\!S} = \frac{s_{ZY}}{s_{ZX}} \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}
   \]

   \[
   = \frac{\text{Cov}(Z, \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i)}{\text{Cov}(Z, X_i)}
   \]

   \[
   = \frac{\beta_1 \text{Cov}(Z, X_i) + \beta_2 \text{Cov}(Z, W_i)}{\text{Cov}(Z, X_i)}
   \]

   \[
   = \beta_1 + \beta_2 \frac{\text{Cov}(Z, W_i)}{\text{Cov}(Z, X_i)}
   \]

   Hence, if \( Z_i \) and \( W_i \) are uncorrelated, the second term disappears and the estimator is consistent.

   b. Using the same result, \( Z_i \) and \( W_i \) are correlated, the second term does not disappear and the estimator is inconsistent.

2. In this second question, you are asked to consider a simple regression model

   \[
   Y_i = \beta_0 + \beta_1 X_i + u_i
   \]

   and let \( Z_i \) be an *binary* instrument for \( X_i \). You are then told to use equation (12.4) to show that the TSLS estimator of \( \beta_1 \) can be written as:

   \[
   \hat{\beta}_{T\!S\!L\!S} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}
   \]

   where \( \bar{Y}_1 \) and \( \bar{X}_1 \) denote the means of \( Y_i \) and \( X_i \) (respectively) over that part of the sample with \( Z_i = 1 \) and \( \bar{Y}_0 \) and \( \bar{X}_0 \) denote the means of \( Y_i \) and \( X_i \) (respectively) over that part of the sample with \( Z_i = 0 \).

   Here’s how the result can be shown. We have that:

   \[
   \hat{\beta}_{T\!S\!L\!S} = \frac{s_{ZY}}{s_{ZX}} \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}
   \]

   But, with \( Z_i \) being *binary* and letting \( n_1 \) and \( n_0 \) denote the number of 1’s and 0’s in the sample respectively,
note that $\bar{Z} = \frac{n_1 - n_0}{n_1 + n_0}$. We then have that the numerator in the previous equation becomes:

$$
\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z}) = \sum_{i=1}^{n} Y_i Z_i - \sum_{i=1}^{n} Y_i \bar{Z} - \sum_{i=1}^{n} \bar{Y} Z_i + \sum_{i=1}^{n} \bar{Y} \bar{Z}
$$

$$
= \sum_{Z_i = 1}^{n} Y_i - \bar{Z} \sum_{i=1}^{n} Y_i - \bar{Y} \sum_{i=1}^{n} Z_i + n \bar{Y} \bar{Z}
$$

$$
= n_1 \bar{Y}_1 - \bar{Z} (n_1 \bar{Y}_1 + n_0 \bar{Y}_0)
$$

$$
= n_1 (1 - \bar{Z}) \bar{Y}_1 - n_0 \bar{Z} \bar{Y}_0
$$

$$
= \frac{n_1 n_0}{n_1 + n_0} \bar{Y}_1 - \frac{n_1 n_0}{n_1 + n_0} \bar{Y}_0
$$

$$
= \frac{n_1 n_0}{n_1 + n_0} (\bar{Y}_1 - \bar{Y}_0).
$$

Using the same steps for the denominator yields

$$
\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) = \frac{n_1 n_0}{n_1 + n_0} (\bar{X}_1 - \bar{X}_0).
$$

Substituting this into our expression for the TSLS estimator yields:

$$
\hat{\beta}_{TSLS} = \frac{n_1 n_0}{n_1 + n_0} \frac{(\bar{Y}_1 - \bar{Y}_0)}{(\bar{X}_1 - \bar{X}_0)}
$$

$$
= \frac{(\bar{Y}_1 - \bar{Y}_0)}{(\bar{X}_1 - \bar{X}_0)}.
$$

3. In this question, you were asked to consider a simple model to estimate the effects of personal computer (PC) ownership on college grade point average for graduating seniors at a university:

$$
GPA_i = \beta_0 + \beta_1 PC_i + u_i \tag{3}
$$

where $PC_i$ is a binary variable indicating PC ownership.

a. The first part of the question asked why might $PC$ ownership be correlated with $u_i$? There are numerous omitted variables in the above model that might be correlated with $PC$ ownership. For example, one might expect one’s GPA to be correlated with income (since higher income households could afford to provide better early and extra educational opportunities and afford to buy a PC). The education level of your parents would likely be correlated with your GPA (given heredity) and more educated parents would more likely see the value in purchasing a PC.

b. In this next part the question, you were asked to explain why $PC$ is likely to be related to parent’s annual income (which I have already done). You were then asked whether or not parental income would be a good IV for PC. Clearly, parental income is likely to satisfy the instrument relevance criteria (i.e., it is likely correlated with $PC$ ownership). However, it is unlikely to satisfy the second requirement for a good instrument (i.e., instrument exogeneity).

c. In part c, you are told to suppose that, four years ago, the university gave grants to buy computers to half of the incoming students, and the students who received the grants were randomly chosen. The question is then how you would use this information to construct an instrumental variable for $PC$. In this case, you have a clear instrument for $PC$ purchases - the binary variable indicating whether or not the were awarded the grant. It satisfies both the instrument relevance criteria (i.e., since individuals are more likely to purchase a PC if they have a grant to do so) and it satisfies the second requirement for a good instrument (i.e., instrument exogeneity), since the grants were randomly assigned.
d. In this last question, you are given information about a sample of individuals, noting that

- that among those students who received the grants, 90% of them owned a PC (i.e., $X_1 = 0.90$) and
  the group had an average GPA of 3.05 (i.e., $\bar{Y}_1 = 3.05$) and
- that among those students who did not receive the grants, 75% of them owned a PC (i.e., $X_0 = 0.75$)
  and the group had an average GPA of 2.75 (i.e., $\bar{Y}_0 = 2.75$).

Using the results from question 2 above, our estimate of $\beta_{TSLS}^1$ would be:

$$
\hat{\beta}_{TSLS}^1 = \frac{(\bar{Y}_1 - \bar{Y}_0)}{(X_1 - X_0)} = \frac{(3.05 - 2.75)}{(0.90 - 0.75)} = 2.
$$


<table>
<thead>
<tr>
<th>Regressor</th>
<th>OLS</th>
<th>TSLS</th>
<th>TSLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MoreKids$</td>
<td>-5.387</td>
<td>-6.313</td>
<td>-5.821</td>
</tr>
<tr>
<td>Additional Regressors</td>
<td>Intercept</td>
<td>Intercept</td>
<td>Intercept, agem1, black, hispan, othrace</td>
</tr>
<tr>
<td>First Stage F-stat</td>
<td>1238.2</td>
<td></td>
<td>1280.9</td>
</tr>
</tbody>
</table>

The table shows the OLS and 2SLS estimates for the $MoreKids$ variable only. Values for the intercept and additional regressors are not shown.

a. The coefficient is -5.387, which indicates that women with more than 2 children work 5.387 fewer weeks per year than women with 2 or fewer children.

b. Both fertility and weeks worked are choice variables. A woman with a positive labor supply regression error (a woman who works more than average) may also be a woman who is less likely to have an additional child. This would imply that $MoreKids$ is positively correlated with the regression error, so that the OLS estimator of $\beta_{MoreKids}$ is positively biased.

c. The linear regression of $morekids$ on $samesex$ (a linear probability model) yields an estimated coefficient on $samesex$ of 0.066 so that couples with $samesex = 1$ are 6.6% more likely to have an additional child that couples with $samesex = 0$. The effect is highly significant (t-statistic=35.2).

d. $Samesex$ is random and is unrelated to any of the other variables in the model including the error term in the labor supply equation. Thus, the instrument is exogenous. From (c), the first stage F-statistic is large (F=1238) so the instrument is relevant. Together, these imply that $samesex$ is a valid instrument.

e. No, see the answer to (d).

f. See column (2) of the table. The estimated value of $Morekids = -6.313$.

g. See column (3) of the table. The results do not change in an important way. The reason is that $samesex$ is unrelated to $agem1$, black, hispan, othrace, so that there is no omitted variable bias in IV regression in (2).