Answers to Econometrics Questions

Question 7

1a)
Height is in inches, so 5 foot 9 inches=69 inches
The gender gap in earnings = -.820(coefficient on male)+.013 (coefficient on M*H)*69 = .077. Thus, log (male income)-log (female income)=log (male income/female income)=0.77. Then, male income/female income=e^(-.077)=1.08. Therefore, the percentage difference in earnings for a woman and a man is 8%.

2 points for correct formula
1 points for using correct coefficient values in formula (-.820 and .13 and 69 inches)
1 point for getting the difference=.077
1 point for converting that to percentage change (8 percent)
<u>5 points total</u>

1b)
No need to actually calculate the gender gap. From the table, the gender cap is the coefficient on gender in regression (2), .215, which is larger than the gender gap of .077 in regression (4).
The difference reason the estimates are different is that *Height* is correlated with *Male*. This can be understood as a standard OVB example: if *Height* is omitted, *Male* is correlated with the error term so *Male* picks up the effect of *Height* so the coefficient on *Male* is biased up when *Height* is excluded.

2 points for recognizing positive correlation between being male and office work or Height.
2 points for recognizing positive correlation between being wages and office work or Height
1 points for realizing both positive correlations lead to an upward bias on the coefficient on Gender in regression (2)
<u>5 points total</u>

2a)
This is testing whether the coefficients on *height* and *Male*height* are jointly significant in predicting education, controlling for the other covariates. The F-test output is already given in column 7 (130.9), so the statistic should be compared against the F distribution (Table 4) with m=2 degrees of freedom. Since 130.9> 3, which is the critical value at the 5% significance level, we can reject the null hypothesis and conclude that the effect of height is not zero.

2 points for correctly identifying that the null is that both coefficients equal zero
1 point for mentioning the word F-statistic

1 point for using the F-statistic from column 7 (130.9) and for correctly identifying that the degrees of freedom from the table; m=2
1 points for rejecting the null hypothesis
<u>5 points total</u>

2b)
For women, the effect of height on years of education is represented by the coefficient on *height.* The null hypothesis of zero effect can be tested by testing whether the coefficient is significantly different from zero. The t statistic =.093/.010 = 9.3 (this is significant even at the 1% level of significance with a one-sided null)

2 points for correctly identifying the *height* coefficient to test
2 points for the correct test statistic (9.3)
1points for rejecting the null hypothesis
<u>5 points total</u>

2c)
For male, the null hypothesis: the coefficient on *height(b1)* + the coefficient on *Male*height(b2)*=0. So, we are testing the null of b1+b2=0. If .093+.055=.148 is significantly different from zero, we reject the null. The t-statistics= b1+b2/[var(b1)+var(b2)+cov(b1,b2)]$^{1/2}$. We can't get the covariance of the two coefficients from the table, so the exact value of t-statistic cannot be computed.

2 points for correctly identifying the null hypothesis
2 points for mentioning t-statistics.
1 points for correctly writing t-statistics formula, including cov(b1, b2).
<u>5 points total</u>

3a)
We believe that intelligence is positively correlated with schooling and that intelligence has a positive effect on earnings. So if we omit "intelligence", we expect the coefficient on *Educ* to be upwardly biased.

2 points for recognizing positive correlation between being intelligence and education
2 points for recognizing positive correlation between being wages and intelligence
1 points for realizing both positive correlations lead to an upward bias
<u>5 points total</u>

3b)
For these variables to adequately control for intelligence, we would need the error term to be uncorrelated with *Educ*, once the control variables are included. A different way to say this is that, if *W* are the control variables (race, *Height*, etc) and *X* is *Educ*, then we need $E(u|X, W) = E(u|X)$, which is the conditional mean independence condition. These are two different ways to say the same idea. So, the question is, is it plausible that $E(u|X, W) = E(u|X)$? That is, among individuals with

the same value of *W*, are there remaining variations in intelligence that would affect outcomes? That is, among individuals of the same race, height, and gender, are there still variations in intelligence that affect earnings? Getting this far, with the correct question stated precisely, is worth all but one point. The remaining point would be for using common sense and saying that individuals of the same height/race/gender can have significant variations in intelligence that would affect earnings, so that the conditional mean independence assumption does not plausibly hold so that regression (5) does not adequately control for intelligence. It might be that a full credit answer reaches this conclusion without fully going through the error term argument but that will require careful reading.

2 points for mentioning "conditional mean independence"
2 points for discussing the logic of "conditional mean independence"
1 points for arguing that the regression (5) does not control for intelligence well. However, the full credit could be given if the answer explains the logic of conditional mean independence.
5 points total

3c)
i) The first-stage equation is regression (7). The null hypothesis: Both coefficients on *Height* and *Male×Height* are zero. The F-statistics is 130.9, which exceeds 10, so reject the null and conclude that the instruments are strong.

1 point for pointing out that the relevant regression is (7), the "first stage regression"
1 point for stating the relevance condition of corr(Z, X) being nonzero
1 points for mentioning the null hypothesis.
1 point for comparing F stat=130.9 to the critical value
1 point for mentioning the critical value is 10, and rejecting the null
5 points total

ii) The exogeneity condition is that cov(Z,u) = 0. Because intelligence is one of the main omitted variables of concern, and we don't want it to have a covariance with respect to *height* and *heightxgender*. Part b asserts a reason for height to be correlated with intelligence. If so, then height would be correlated with the error term, so *Height* would not be exogenous. If we have a sample in which early childhood nutrition is good throughout and if we postulate that the genetic aspect of height is uncorrelated with intelligence, then *Height* would be uncorrelated with intelligence and thus uncorrelated with that component of the error term (which is the question being asked – *Height* might be correlated with other items in the error term and it is fine if those are discussed however this discussion is not required and its absence should not be penalized).

However, a student who says that the instruments are valid without deeply addressing counterarguments should earn no more than 4 points total given the fact that these are unlikely to be good instruments in the study.

2 points for stating cov(Z,u) = 0
1 point for mentioning J-test
2 points for arguing that the instruments are not so exogenous.
5 points total

3d)
i) Entity fixed effects estimator can deal with the omitted variable of intelligence, given that intelligence is fixed over time

$Y_{it} = b_1X_{1it} + \ldots + b_kX_{kit} + a_i + u_{it}$
i=1,2,...,N
t=1,2,...T
X_{kit} means kth variable for individual i at time
a_i is entity fixed effects.

The fixed effects estimator can be estimated by including a dummy variable for each of n−1 individuals. The dummy variable captures the effect of variables that do not change over time, including intelligence.
$Y_{it} = b_1X_{1it} + \ldots + b_kX_{kit} + a_1D_1 + \ldots + a_{N-1}D_{N-1+}\ u_{it}$
t=1,2,...T

Thus the fixed effects regression controls for all such effects and therefore controls in particular for the effect of intelligence.

This point can be made a lot of different ways, including arguing from the deviations-from-means formulation, or the T = 2 intuition (its OK to do this even though T = 4).

3 points for mentioning entity FE.
2 points for explaining that how entity FE can get rid of OVB
5 points total

ii) Homoskedasticity-only SE, heteroskedasticity-robust SE, and clustered HAC SE. [Note someone could say Newey-West – so really there are four possibilities, cluster and NW being distinct]
The choice of standard error depends on the error structure. The homoskedasticity-only SE or heteroskedasticity-robust SE will underestimate standard errors if you have serial correlation in the error term. In this panel data, it is quite likely that we will have serial correlation since we have observations on earnings for four "consecutive" years. Suppose you have a positive error term in the first year. This means you have a better-paying job that year than would be predicted by the regression. Odds are that you will have the same job, or a similar job, the next year, so that your error term will be positive the next year too. That is, the error term will be positively serially correlated. Therefore we need to use a standard error that takes into account this possible serial correlation, as well as heteroskedasticity, that

is, we need a HAC standard error. This is best done using clustered standard errors.
1 point for identifying at least one standard error
2 points (each) for identifying 2 other different types of standard errors
1 points for mentioning serial correlation
1 points for suggesting clustered HAC SE to correct for serial correlation and heteroskedasticity
<u>5 points total</u>

iii) The big thing here is to think about the education variable in panel context. The data are for **working adults,** so for the vast majority of workers, *Educ* will not change. The two reasons it would change in the data are (a) some on-the-job training or schooling taken while working, or (b) errors in reporting for example recall errors or coding errors – so it will look like schooling has changed when it actually didn't. The latter is simply measurement error which will in general introduce bias (if classical measurement error, bias towards zero). The former would mean that any changes in earnings would reflect a very different type of education, e.g. a night course at a community college. Taking a night course at a community college is not exogenous, for example it might because getting some accreditation will give you a pay increase in your current position, or qualify you for a promotion. So whatever is estimated using this panel data regression, either from variation in (a) or (b), is unlikely to provide a good estimate of the effect of another year of schooling in the sense of going to college, finishing 11th grade, etc.

Even if we can get rid of OVB from intelligence by using entity fixed effect and corrects for serial correlation by using clustered HAC SE, we are still exposed to other OVB which are from the omitted variables that are constant across the entities but vary over time. To solve this problem, we can add <u>time fixed effects</u> along with the entity fixed effects. For example, the national level shock will affect everyone in the economy, and the effect of education on income each year could be changing, depending on the persistence of this shock.

2 points for mentioning the existence of little variation in using education in panel data .
3 points for mentioning time FE
<u>5 points total</u>