Exam

Please discuss each of the 3 problems on a separate sheet of paper, not just on a separate page!

Problem 1: (15 points)
Two researchers are investigating the effects of time spent studying on the examination marks earned by students on a certain course. For a sample of 100 students, they have the examination mark, M, total hours spent studying, H, hours on primary study, P, and hours spent on revision, R. By definition, $H = P + R$. The sample means of H, P, and R are 100 hours, 95 hours, and 5 hours, respectively. The sample correlation coefficients are 0.98 for H and P, 0.10 for H and R, and -0.11 for P and R. The standard deviations of the distributions of H, P, and R are 10.1, 10.1, and 2.1, respectively. Researcher A decides to regress M on P and R and fits the following regression (standard errors in parentheses; t statistics in square brackets):

$$\begin{array}{rcll} \hat{M} & = & 45.6 + 0.15P + 0.21R & R^2 = 0.243 \\ & & \begin{array}{lll}(2.8) & (0.03) & (0.14)\end{array} & \\ & & \begin{array}{lll}[16.30] & [5.49] & [1.51]\end{array} & \end{array} \qquad (1)$$

Researcher B decides to regress M on H, P, and R. However, the regression application refuses to fit the regression with all three explanatory variables. Instead, it drops R and the regression output is

$$\begin{array}{rcll} \hat{M} & = & 45.6 + 0.21H - 0.05P & R^2 = 0.243 \\ & & \begin{array}{lll}(2.8) & (0.14) & (0.14)\end{array} & \\ & & \begin{array}{lll}[16.30] & [1.51] & [-0.40]\end{array} & \end{array} \qquad (2)$$

Note: In answering the following questions, you should give your reasoning in general terms. Detailed mathematical analysis is not required and no credit will be given for it.

1. Researcher A says that her specification has better explanatory power than that of Researcher B because the coefficient of her main variable, P, has a high t statistic. Explain whether this assertion is correct.

2. She says that the insignificant coefficient of R in (1) is to be expected because the students, on average, spent much less time on revision than on primary study. Explain whether this assertion is correct.

3. Researcher B says that, assuming that his specification is in fact correct, not being able to include R in the regression has given rise to omitted variable bias, and this is responsible for the negative coefficient of P. Explain whether this assertion is correct.

Solution:

1. The assertion is incorrect. The specifications are equivalent and explanatory power, as measured by R^2, is the same.

2. The assertion is incorrect. The small mean of R has nothing to do with it. Since the estimated coefficient of R is actually greater than that of P, the relatively low t statistic is attributable to the much greater standard error. This, in turn, is attributable to the fact that the variance of R is smaller than that of P. (See the data on the sample standard deviations.)

3. This is nonsense since his specification cannot be correct. It involves exact multicollinearity.
   It is wrong to say that the negative coefficient is implausible. It is measuring the difference in the effects of P and R, not the absolute effect of P. However, it is true say that the estimator will have large variance since H and P are highly correlated.

Problem 2: (15 points)
Let $\{e_t : t = -1, 0, 1, ...\}$ be a sequence of independent, identically distributed random variables with mean zero and variance one. Define a stochastic process by

$$x_t = e_t - \frac{1}{2}e_{t-1} + \frac{1}{2}e_{t-2}, \qquad t = 1, 2, ...$$

1. Find $E[x_t]$ and $Var[x_t]$. Do either of these depend on t?
2. Compute $Corr(x_t, x_{t+1})$ and $Corr(x_t, x_{t+2})$.
3. What is $Corr(x_t, x_{t+h})$ for $h > 2$?

Solution:

1. Find $E[x_t]$ and $Var[x_t]$. Do either of these depend on t?

$$\begin{aligned} x_t &= e_t - \frac{1}{2}e_{t-1} + \frac{1}{2}e_{t-2} \\ E[x_t] &= E[e_t] - \frac{1}{2}E[e_{t-1}] + \frac{1}{2}E[e_{t-2}] \\ E[x_t] &= 0 \end{aligned}$$

$$\begin{aligned} Var[x_t] &= Var[e_t - \frac{1}{2}e_{t-1} + \frac{1}{2}e_{t-2}] \\ &= E[(e_t - \frac{1}{2}e_{t-1} + \frac{1}{2}e_{t-2})^2] \\ &= E[e_t^2] + \frac{1}{4}E[e_{t-1}^2] + \frac{1}{4}E[e_{t-2}^2] - 2\frac{1}{2}E[e_t \cdot e_{t-1}] \\ &\quad +2\frac{1}{2}E[e_t \cdot e_{t-2}] - 2\frac{1}{2}E[e_{t-1} \cdot e_{t-2}] \\ &= 1 + \frac{1}{4} + \frac{1}{4} = 1.5 \end{aligned}$$

   None of these expressions depend on t.

2. Compute $Corr(x_t, x_{t+1})$ and $Corr(x_t, x_{t+2})$.

$$Cov(x_t, x_{t+1}) = E\left[(e_t - \frac{1}{2}e_{t-1} + \frac{1}{2}e_{t-2})(e_{t+1} - \frac{1}{2}e_t + \frac{1}{2}e_{t-1})\right]$$

$$
\begin{aligned}
&= -E\left[e_t\frac{1}{2}e_t\right] - E\left[\frac{1}{2}e_{t-1}\frac{1}{2}e_{t-1}\right] \\
&= -\frac{1}{2} - \frac{1}{4} = -\frac{3}{4} \\
Corr(x_t, x_{t+1}) &= -\frac{1}{2}
\end{aligned}
$$

$$
\begin{aligned}
Cov(x_t, x_{t+2}) &= E\left[(e_t - \frac{1}{2}e_{t-1} + \frac{1}{2}e_{t-2})(e_{t+2} - \frac{1}{2}e_{t+1} + \frac{1}{2}e_t)\right] \\
&= E\left[e_t\frac{1}{2}e_t\right] = \frac{1}{2} \\
Corr(x_t, x_{t+2}) &= \frac{1}{3}
\end{aligned}
$$

3. What is $Corr(x_t, x_{t+h})$ for $h > 2$?
   $Corr(x_t, x_{t+h}) = 0$ for $h > 2$

Problem 3: (20 points)
The NLSY2000 data set contains the following data for a sample of 2,427 males and 2,392 females for the years $1980 - 2000$: weight in pounds, years of schooling, age, marital status in the form of a dummy variable MARRIED defined to be 1 if the respondent was married, 0 if single, and height in inches. Hypothesizing that weight is influenced by schooling, age, marital status, and height, the following regressions were performed for males and females separately:

(1) an ordinary least squares (OLS) regression pooling the observations

(2) a within-groups fixed effects regression

(3) a random effects regression

The results of these regressions are shown in the table. Standard errors are given in parentheses.

	Males			Females		
	OLS	FE	RE	OLS	FE	RE
Years of schooling	-0.98	-0.02	-0.45	-1.95	-0.60	-1.25
	(0.09)	(0.23)	(0.16)	(0.12)	(0.27)	(0.18)
Age	1.61	1.64	1.65	2.03	1.66	1.72
	(0.04)	(0.02)	(0.02)	(0.05)	(0.03)	(0.03)
Married	3.70	2.92	3.00	-8.27	3.08	1.98
	(0.48)	(0.33)	(0.32)	(0.59)	(0.46)	(0.44)
Height	5.07	dropped	4.95	3.48	dropped	3.38
	(0.08)		(0.18)	(0.10)		(0.21)
constant	-209.52	dropped	-209.81	-105.90	dropped	-107.61
	(5.39)		(12.88)	(6.62)		(13.43)
R^2	0.27	-	-	0.17	-	-
n	17,299	17,299	17,299	13,160	13,160	13,160
Hausman $\chi^2(3)$			7.22			92.94

1. Explain why height is excluded from the FE regression.
2. Evaluate, for males and females separately, whether the fixed effects or random effects model should be preferred.
3. For males and females separately, compare the estimates of the coefficients in the OLS and FE models and attempt to explain the differences.

4. Explain in principle how one might test whether individual-specific fixed effects jointly have significant explanatory power, if the number of individuals is small. Explain why the test is not practical in this case.

Solution:

1. Height is constant over observations. Hence, for each individual, $HEIGHT_{it} - \overline{HEIGHT}_i = 0$ for all t, where $\overline{HEIGHT}_i$ the mean height for individual i for the observations for that individual. Hence height has to be dropped from the regression model. The critical value of chi-squared, with three degrees of freedom, is 7.82 at the 5 percent level and 16.27 at the 0.1 percent level. Hence there is a possibility that the random effects model may be appropriate for males, but it is definitely not appropriate for females.

2. **Males**
   The OLS regression suggests that schooling has a small (one pound less per year of schooling) but highly significant negative effect on weight. The fixed effects regression eliminates the effect, indicating that an unobserved effect is responsible: males with unobserved qualities that have a positive effect on educational attainment, controlling for other measured variables, have lower weight as a consequence of the same unobserved qualities. We cannot compare estimates of the effect of height since it is dropped from the FE regression. The effect of age is the same in the two regressions. There is a small but highly significant positive effect of being married, the OLS estimate possibly being inflated by an unobserved effect.
   Females
   The main, and very striking, difference is in the marriage coefficient. The OLS regression suggests that marriage reduces weight by eight pounds, a remarkable amount. The FE regression suggests the opposite, that marriage leads to an increase in weight that is similar to that for males. The clear implication is that women who weigh less are relatively successful in the marriage market, but once they are married they put on weight . For schooling the story is much the same as for males, except that the OLS coefficient is much larger and the coeffi-

cient remains significant at the 5 percent level in the FE regression. The effect of age appears to be exaggerated in the OLS regression, for reasons that are not obvious.

3. One might test whether individual-specific fixed effects jointly have significant explanatory power by performing a LSDV regression, eliminating the intercept in the model and adding a dummy variable for each individual. One would compare RSS for this regression with that for the regression without the dummy variables, using a standard F test. In the present case it is not a practical proposition because there are more than 17,000 males and 13,000 females.