Exam

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

Problem 1: (15 points)
Suppose that in the model

$$Y = \beta_1 + \beta_2 X + u,$$

where the disturbance term u satisfies the regression model assumptions, the variable X is subject to measurement error, being underestimated by a fixed amount α in all observations.

1. Discuss whether it is true that the ordinary least squares estimator of β_2 will be downwards biased by an amount proportional to both α and β_2.
2. Discuss whether it is true that the fitted values of Y from the regression will be reduced by an amount $\alpha\beta_2$.
3. Discuss whether it is true that R^2 will be reduced by an amount proportional to α.

Solution:

1. Not true.
   Let the measured X be $\tilde{X}$, where $\tilde{X} = X - \alpha$. Then

$$\begin{aligned} b_2^{OLS} &= \frac{\sum_i (\tilde{X}_i - \bar{\tilde{X}})(Y_i - \bar{Y})}{\sum_i (\tilde{X}_i - \bar{\tilde{X}})^2} \\ &= \frac{\sum_i \left((X_i - \alpha) - (\bar{X} - \alpha)\right)(Y_i - \bar{Y})}{\sum_i \left((X_i - \alpha) - (\bar{X} - \alpha)\right)^2} \\ &= \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = b_2 \end{aligned}$$

   Thus this measurement error has no effect on the estimate of the slope coefficient.

2. The estimator of the intercept will be $b_1^{OLS} = \bar{Y} - b_2^{OLS}\bar{\tilde{X}} = \bar{Y} - b_2(\bar{X} - \alpha)$. Hence, the fitted value in observation i will be

$$\hat{Y}_i = \bar{Y} - b_2(\bar{X} - \alpha) + b_2\tilde{X}_i = \bar{Y} - b_2(\bar{X} - \alpha) + b_2(X_i - \alpha)$$

   which is what it would be in the absence of the measurement error.

3. Since R^2 is the variance of the fitted values of Y divided by the variance of the actual values, it will be unaffected.

Problem 2: (15 points)
Suppose that a time series process $\{y_t\}$ is generated by $y_t = z + e_t$, for all $t = 1, 2, ...$, where $\{e_t\}$ is an i.i.d. sequence with mean zero and variance σ_e^2. The random variable z does not change over time; it has mean zero and variance σ_z^2. Assume that each e_t is uncorrelated with z.

1. Find the expected value and variance of y_t.
2. Find $Cov(y_t, y_{t+h})$ for any t and h. Is $\{y_t\}$ covariance stationary?
3. Calculate $Corr(y_t, y_{t+h})$ for any t and h.
4. Covariance stationary sequences where $Corr(x_t, x_{t+h}) \to 0$ as $h \to \infty$ are said to be **asymptotically uncorrelated**. Is y_t an asymptotically uncorrelated process?

Solution:

1.

$$\begin{aligned} E[y_t] &= E[z + e_t] = E[z] + E[e_t] = 0 \\ Var[y_t] &= E[(y_t - E[y_t])^2] = E[(z + e_t)^2] = E[z^2] + 2E[z \cdot e_t] + E[e_t^2] \\ &= \sigma_z^2 + \sigma_e^2 \end{aligned}$$

2.

$$Cov[y_t, y_{t+h}] = E[y_t \cdot y_{t+h}] = E[(z + e_t)(z + e_{t+h})] = E[z^2] = \sigma_z^2$$

The process is covariance stationary.

3.

$$Corr[y_t, y_{t+h}] = \frac{\sigma_z^2}{\sigma_z^2 + \sigma_e^2}$$

4. No, the correlation does not die out.

Problem 3: (20 points)
A researcher has the following data for a sample of $1,498$ females drawn from the United States National Longitudinal Survey of Youth: Weight in kilos, Height in centimeters, 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 ethnicity in the form of a dummy variable Black defined to be 1 if the respondent was black, 0 otherwise.
These data were obtained for 1985 and 2000 for the same women. The respondents were aged 20-27 in 1985. Women who were divorced in either 1985 or 2000 were excluded from the sample. The researcher fits two regressions:
(1) an ordinary least squares (OLS) regression combining the observations for 1985 and the observations for 2000 with Weight as the dependent variable and Years of schooling, Married, Height, Age, and Black as explanatory variables,
(2) a first differences (FD) regression with the change in weight from 1985 to 2000 as the dependent variable and the change in Years of schooling, the change in Married, and the change in Age (15 years for all respondents) over the same period as explanatory variables. The FD regression was fitted without a constant.
The results of these regressions are shown in the following table. t-statistics are given in parentheses.

	OLS	FD
Years of schooling	−0.88	−0.06
	(−7.41)	(−0.25)
Married	−3.274	0.01
	(−5.28)	(0.02)
Height (cm)	0.37	-
	(11.51)	
Age	0.82	0.72
	(22.06)	(28.26)
Black	6.12	-
	(7.43)	
constant	−5.52	-
	(−1.03)	
R^2	0.20	0.49
n	2,996	1,498

1. Explain theoretically why OLS and FD regressions may yield different estimates of the parameters of the model.
2. The coefficients of Years of schooling and Married are negative and highly significant in the OLS regression, but near zero and not significant in the FD regression. Give an intuitive explanation of this.
3. Explain why Height and Black are excluded from the FD regression.
4. The change in age from 1985 to 2000 is the same for all respondents. Discuss the implications, if any, for the FD regression.
5. R^2 is much higher for the FD regression than for the OLS regression. Does this imply that the FD regression is a better specification?
6. Explain in principle how one might test whether individual-specific fixed effects jointly have significant explanatory power, if the number of individuals is small. (In this case, the number of individuals is large, and so the test is not practical.)

Solution:

1. Suppose that the model is written

   $$W_{it} = \beta_1 + \beta_2 S_{it} + \beta_3 MARRIED_{it} + \beta_4 H_i + \beta_5 A_{it} + \beta_6 BLACK_i + \alpha_i + u_{it},$$

   where W = weight, S = years of schooling, H = height, A = age, and α is a term capturing the unobserved characteristics of the individual. If α is correlated with any of the observed characteristics, OLS estimates will be subject to omitted variable bias (unobserved heterogeneity bias). However, in the FD regression

   $$\triangle W_{it} = \beta_2 \triangle S_{it} + \beta_3 \triangle MARRIED_{it} + \beta_5 \triangle A_{it} + \triangle u_{it}$$

   the unobserved heterogeneity disappears and one obtains unbiased estimates.
2. The differences suggest that both S and $MARRIED$ are correlated with the unobserved heterogeneity.

3. H and $BLACK$ will be the same in both years and hence their first differences are zero. It is not possible to include zero variables in a regression model. (The intercept drops out for the same reason.)

4. Usually a constant explanatory variable has to be dropped from a regression specification because its effect cannot be differentiated from the intercept. However this problem does not arise in the FD regression because there is no intercept.

5. The R^2 in the two regressions are not comparable because in the OLS regression R^2 is the proportion of the variance of weight explained by the regression while in the second it is the proportion of the variance of the change of weight. However, since it is generally much easier to explain variances of levels than variances of differences, the higher R^2 in the FD regression does suggest that it is a better specification. The reason is probably due to the fact that much of the variance in levels is attributable to variance in the unobserved fixed effects.

6. Fit the least squares dummy variable (LSDV) fixed effects model by replacing the intercept in the OLS model by a set of individual-specific dummy variables. Then evaluate the joint explanatory power of the dummy variables using the F statistic

$$F(1497, 1492) = \frac{(RSS_{OLS} - RSS_{LSDV})/1497}{RSS_{LSDV}/1492}$$

the null hypothesis being that the coefficients of the dummy variables are the same. Clearly in a sample such as the present one this test is impracticable because it requires inserting a very large number of individual-specific dummy variables.