CHAPTER 13

SOLUTIONS TO PROBLEMS

13.1 Without changes in the averages of *any* explanatory variables, the average fertility rate fell by .545 between 1972 and 1984; this is simply the coefficient on *y84*. To account for the increase in average education levels, we obtain an additional effect: $-.128(13.3 - 12.2) \approx -.141$. So the drop in average fertility if the average education level increased by 1.1 is .545 + .141 = .686, or roughly two-thirds of a child per woman.

13.3 We do not have repeated observations on the *same* cross-sectional units in each time period, and so it makes no sense to look for pairs to difference. For example, in Example 13.1, it is very unlikely that the same woman appears in more than one year, as new random samples are obtained in each year. In Example 13.3, some houses may appear in the sample for both 1978 and 1981, but the overlap is usually too small to do a true panel data analysis.

13.5 No, we cannot include age as an explanatory variable in the original model. Each person in the panel data set is exactly two years older on January 31, 1992 than on January 31, 1990. This means that $\Delta age_i = 2$ for all i. But the equation we would estimate is of the form

$$\Delta saving_i = \delta_0 + \beta_1 \Delta age_i + \ldots,$$

where δ_0 is the coefficient on the year dummy for 1992 in the original model. As we know, when we have an intercept in the model, we cannot include an explanatory variable that is constant across i; this violates Assumption MLR.3. Intuitively, since age changes by the same amount for everyone, we cannot distinguish the effect of age from the aggregate time effect.

13.7 (i) It is not surprising that the coefficient on the interaction term changes little when *afchnge* is dropped from the equation, because the coefficient on *afchnge* in (13.12) is only .0077 (and its *t* statistic is very small). The increase from .191 to .198 is easily explained by sampling error.

(ii) If *highearn* is dropped from the equation [so that $\beta_1 = 0$ in (13.10)], then we are assuming that, prior to the change in policy, there is no difference in average duration between high earners and low earners. But the very large (.256), highly statistically significant estimate on *highearn* in (13.12) shows this presumption to be false. Prior to the policy change, the high earning group spent about 29.2% [$\exp(.256) - 1 \approx .292$] longer on unemployment compensation than the low earning group. By dropping *highearn* from the regression, we attribute to the policy change the difference between the two groups that would be observed without any intervention.

SOLUTIONS TO COMPUTER EXERCISES

C13.1 (i) The *F* statistic (with 4 and 1,111 *df*) is about 1.16 and *p*-value $\approx$ .328, which shows that the living environment variables are jointly insignificant.

(ii) The F statistic (with 3 and 1,111 df) is about 3.01 and p-value $\approx$ .029, and so the region dummy variables are jointly significant at the 5% level.

(iii) After obtaining the OLS residuals, $\hat{u}$, from estimating the model in Table 13.1, we run the regression $\hat{u}^2$ on *y74*, *y76*, …, *y84* using all 1,129 observations. The null hypothesis of homoskedasticity is H_0: $\gamma_1 = 0, \gamma_2 = 0, \ldots, \gamma_6 = 0$. So we just use the usual F statistic for joint significance of the year dummies. The R-squared is about .0153 and $F \approx 2.90$; with 6 and 1,122 df, the p-value is about .0082. So there is evidence of heteroskedasticity that is a function of time at the 1% significance level. This suggests that, at a minimum, we should compute heteroskedasticity-robust standard errors, t statistics, and F statistics. We could also use weighted least squares (although the form of heteroskedasticity used here may not be sufficient; it does not depend on *educ*, *age*, and so on).

(iv) Adding *y74*·*educ*, …, *y84*·*educ* allows the relationship between fertility and education to be different in each year; remember, the coefficient on the interaction gets added to the coefficient on *educ* to get the slope for the appropriate year. When these interaction terms are added to the equation, $R^2 \approx .137$. The F statistic for joint significance (with 6 and 1,105 df) is about 1.48 with p-value $\approx$ .18. Thus, the interactions are not jointly significant at even the 10% level. This is a bit misleading, however. An abbreviated equation (which just shows the coefficients on the terms involving *educ*) is

$$\begin{aligned}\widehat{kids} = {} & -8.48 - .023\,educ + \ldots - .056\,y74\cdot educ - .092\,y76\cdot educ \\ & \quad (3.13) \quad (.054) \qquad\qquad (.073) \qquad\quad (.071) \\ & - .152\,y78\cdot educ - .098\,y80\cdot educ - .139\,y82\cdot educ - .176\,y84\cdot educ. \\ & \quad (.075) \qquad\quad (.070) \qquad\quad (.068) \qquad\quad (.070)\end{aligned}$$

Three of the interaction terms, *y78*·*educ*, *y82*·*educ*, and *y84*·*educ*, are statistically significant at the 5% level against a two-sided alternative, with the p-value on the latter being about .012. The coefficients are large in magnitude as well. The coefficient on *educ* – which is for the base year, 1972 – is small and insignificant, suggesting little if any relationship between fertility and education in the early seventies. The estimates above are consistent with fertility becoming more linked to education as the years pass. The F statistic is insignificant because we are testing some insignificant coefficients along with some significant ones.

C13.3 (i) Other things equal, homes farther from the incinerator should be worth more, so $\delta_1 > 0$. If $\beta_1 > 0$, then the incinerator was located farther away from more expensive homes.

(ii) The estimated equation is

$$\begin{aligned}\widehat{\log(price)} = {} & 8.06 - .011\,y81 + .317\log(dist) + .048\,y81\cdot\log(dist) \\ & \quad (0.51) \quad (.805) \qquad (.052) \qquad\quad (.082)\end{aligned}$$

$$n = 321, \quad R^2 = .396, \quad \bar{R}^2 = .390.$$

While $\hat{\delta}_1 = .048$ is the expected sign, it is not statistically significant (*t* statistic $\approx .59$).

(iii) When we add the list of housing characteristics to the regression, the coefficient on *y81* · log(*dist*) becomes .062 (se = .050). So the estimated effect is larger – the elasticity of *price* with respect to *dist* is .062 after the incinerator site was chosen – but its *t* statistic is only 1.24. The *p*-value for the one-sided alternative H_1: $\delta_1 > 0$ is about .108, which is close to being significant at the 10% level.

(iv) The fact that ldist has a much smaller coefficient and is insignificant in part (iii) indicates that the characteristics included in part (iii) largely capture the housing characteristics that are most important for determining housing prices.

C13.5 (i) Using pooled OLS we obtain

$$\widehat{\log(rent)} = \underset{(.535)}{-.569} + \underset{(.035)}{.262}\, d90 + \underset{(.023)}{.041}\log(pop) + \underset{(.053)}{.571}\log(avginc) + \underset{(.0010)}{.0050}\, pctstu$$

$n = 128$, $R^2 = .861$.

The positive and very significant coefficient on *d90* simply means that, other things in the equation fixed, nominal rents grew by over 26% over the 10 year period. The coefficient on *pctstu* means that a one percentage point increase in *pctstu* increases *rent* by half a percent (.5%). The *t* statistic of five shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

(ii) The standard errors from part (i) are not valid, unless we think a_i does not really appear in the equation. If a_i is in the error term, the errors across the two time periods for each city are positively correlated, and this invalidates the usual OLS standard errors and *t* statistics.

(iii) The equation estimated in differences is

$$\widehat{\Delta\log(rent)} = \underset{(.037)}{.386} + \underset{(.088)}{.072}\,\Delta\log(pop) + \underset{(.066)}{.310}\log(avginc) + \underset{(.0041)}{.0112}\,\Delta pctstu$$

$n = 64$, $R^2 = .322$.

Interestingly, the effect of *pctstu* is over twice as large as we estimated in the pooled OLS equation. Now, a one percentage point increase in *pctstu* is estimated to increase rental rates by about 1.1%. Not surprisingly, we obtain a much less precise estimate when we difference (although the OLS standard errors from part (i) are likely to be much too small because of the positive serial correlation in the errors within each city). While we have differenced away a_i, there may be other unobservables that change over time and are correlated with $\Delta pctstu$.

(iv) The heteroskedasticity-robust standard error on $\Delta pctstu$ is about .0028, which is actually much smaller than the usual OLS standard error. This only makes *pctstu* even more significant

(robust t statistic $\approx$ 4). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation.

C13.7 (i) Pooling across semesters and using OLS gives

$$\widehat{trmgpa} = \underset{(0.35)}{-1.75} \underset{(.048)}{-.058\, spring} + \underset{(.00015)}{.00170\, sat} \underset{(.0010)}{-.0087\, hsperc}$$
$$+ \underset{(.052)}{.350\, female} \underset{(.123)}{-.254\, black} \underset{(.117)}{-.023\, white} \underset{(.076)}{-.035\, frstsem}$$
$$\underset{(.00073)}{-.00034\, tothrs} + \underset{(0.104)}{1.048\, crsgpa} \underset{(.049)}{-.027\, season}$$
$$n = 732, \quad R^2 = .478, \quad \bar{R}^2 = .470.$$

The coefficient on *season* implies that, other things fixed, an athlete's term GPA is about .027 points lower when his/her sport is in season. On a four point scale, this a modest effect (although it accumulates over four years of athletic eligibility). However, the estimate is not statistically significant (t statistic $\approx$ –.55).

(ii) The quick answer is that if omitted ability is correlated with *season*, then, as we know from Chapters 3 and 5, OLS is biased and inconsistent. The fact that we are pooling across two semesters does not change that basic point.

If we think harder, the direction of the bias is not clear, and this is where pooling across semesters plays a role. First, suppose we used only the fall term, when football is in season. Then the error term and season would be negatively correlated, which produces a downward bias in the OLS estimator of β_{season}. Because β_{season} is hypothesized to be negative, an OLS regression using only the fall data produces a downward biased estimator. [When just the fall data are used, $\hat{\beta}_{season} = -.116$ (se = .084), which is in the direction of more bias.] However, if we use just the spring semester, the bias is in the opposite direction because ability and season would be positive correlated (more academically able athletes are in season in the spring). In fact, using just the spring semester gives $\hat{\beta}_{season} = .00089$ (se = .06480), which is practically and statistically equal to zero. When we pool the two semesters, we cannot, with a much more detailed analysis, determine which bias will dominate.

(iii) The variables *sat*, *hsperc*, *female*, *black*, and *white* all drop out because they do not vary by semester. The intercept in the first-differenced equation is the intercept for the spring. We have

$$\widehat{\Delta trmgpa} = \underset{(.206)}{-.237} + \underset{(.069)}{.019\, \Delta frstsem} + \underset{(.014)}{.012\, \Delta tothrs} + \underset{(0.119)}{1.136\, \Delta crsgpa} \underset{(.043)}{-.065\, \Delta season}$$
$$n = 366, \quad R^2 = .208, \quad \bar{R}^2 = .199.$$

Interestingly, the in-season effect is larger now. Term GPA is estimated to be about .065 points lower in a semester that the sport is in-season. The *t* statistic is about –1.51, which gives a one-sided *p*-value of about .065.

(iv) One possibility is a measure of course load. If some fraction of student-athletes take a lighter load during the season (for those sports that have a true season), then term GPAs may tend to be higher, other things equal. This would bias the results away from finding an effect of *season* on term GPA.

C13.9 (i) When we add the changes of the nine log wage variables to equation (13.33), we obtain

$$\widehat{\Delta\log(crmrte)} = \underset{(.021)}{.020} - \underset{(.027)}{.111}\,d83 - \underset{(.025)}{.037}\,d84 - \underset{(.0241)}{.0006}\,d85 + \underset{(.025)}{.031}\,d86 + \underset{(.025)}{.039}\,d87$$
$$- \underset{(.030)}{.323}\,\Delta\log(prbarr) - \underset{(.018)}{.240}\,\Delta\log(prbconv) - \underset{(.026)}{.169}\,\Delta\log(prbpris)$$
$$- \underset{(.022)}{.016}\,\Delta\log(avgsen) + \underset{(.027)}{.398}\,\Delta\log(polpc) - \underset{(.030)}{.044}\,\Delta\log(wcon)$$
$$+ \underset{(0.14)}{.025}\,\Delta\log(wtuc) - \underset{(.031)}{.029}\,\Delta\log(wtrd) + \underset{(.0212)}{.0091}\,\Delta\log(wfir)$$
$$+ \underset{(.014)}{.022}\,\Delta\log(wser) - \underset{(.102)}{.140}\,\Delta\log(wmfg) - \underset{(.172)}{.017}\,\Delta\log(wfed)$$
$$- \underset{(.096)}{.052}\,\Delta\log(wsta) - \underset{(.102)}{.031}\,\Delta\log(wloc)$$

$n = 540, \quad R^2 = .445, \quad \overline{R}^2 = .424.$

The coefficients on the criminal justice variables change very modestly and the statistical significance of each variable is also essentially unaffected.

(ii) Since some signs are positive and others are negative, they all cannot really have the expected sign. For example, why is the coefficient on the wage for transportation, utilities, and communications (*wtuc*) positive and marginally significant (*t* statistic $\approx$ 1.79)? Higher manufacturing wages lead to lower crime, as we might expect, but, while the estimated coefficient is by far the largest in magnitude, it is not statistically different from zero (*t* statistic $\approx$ –1.37). The *F* test for joint significance of the wage variables, with 9 and 529 *df*, yields $F \approx 1.25$ and *p*-value $\approx .26$.

C13.11. (i) Take changes as usual, holding the other variables fixed: $\Delta math4_{it} = \beta_1\Delta\log(rexpp_{it}) = (\beta_1/100)\cdot[\,100\cdot\Delta\log(rexpp_{it})] \approx (\beta_1/100)\cdot(\,\%\Delta rexpp_{it})$. So, if $\%\Delta rexpp_{it} = 10$, then $\Delta math4_{it} = (\beta_1/100)\cdot(10) = \beta_1/10$.

(ii) The equation, estimated by pooled OLS in first differences (except for the year dummies), is

$$\widehat{\Delta math4} = \underset{(.52)}{5.95} + \underset{(.73)}{.52\, y94} + \underset{(.78)}{6.81\, y95} - \underset{(.73)}{5.23\, y96} - \underset{(.72)}{8.49\, y97} + \underset{(.72)}{8.97\, y98}$$

$$- \underset{(2.76)}{3.45}\, \Delta\log(rexpp) + \underset{(1.029)}{.635}\, \Delta\log(enroll) + \underset{(.055)}{.025}\, \Delta lunch$$

$n = 3{,}300, \quad R^2 = .208.$

Taken literally, the spending coefficient implies that a 10% increase in real spending per pupil decreases the *math4* pass rate by about 3.45/10 ≈ .35 percentage points.

(iii) When we add the lagged spending change, and drop another year, we get

$$\widehat{\Delta math4} = \underset{(.55)}{6.16} + \underset{(.77)}{5.70\, y95} - \underset{(.79)}{6.80\, y96} - \underset{(.74)}{8.99\, y97} + \underset{(.74)}{8.45\, y98}$$

$$- \underset{(3.04)}{1.41}\, \Delta\log(rexpp) + \underset{(2.79)}{11.04}\, \Delta\log(rexpp_{-1}) + \underset{(1.18)}{2.14}\, \Delta\log(enroll)$$

$$+ \underset{(.061)}{.073}\, \Delta lunch$$

$n = 2{,}750, \quad R^2 = .238.$

The contemporaneous spending variable, while still having a negative coefficient, is not at all statistically significant. The coefficient on the lagged spending variable is very statistically significant and implies that a 10% increase in spending last year increases the *math4* pass rate by about 1.1 percentage points. Given the timing of the tests, a lagged effect is not surprising. In Michigan, the fourth grade math test is given in January, and so if preparation for the test begins a full year in advance, spending when the students are in third grade would at least partly matter.

(iv) The heteroskedasticity-robust standard error for $\hat{\beta}_{\Delta\log(rexpp)}$ is about 4.28, which reduces the significance of Δlog(*rexpp*) even further. The heteroskedasticity-robust standard error of $\hat{\beta}_{\Delta\log(rexpp_{-1})}$ is about 4.38, which substantially lowers the *t* statistic. Still, Δlog($rexpp_{-1}$) is statistically significant at just over the 1% significance level against a two-sided alternative.

(v) The fully robust standard error for $\hat{\beta}_{\Delta\log(rexpp)}$ is about 4.94, which even further reduces the *t* statistic for Δlog(*rexpp*). The fully robust standard error for $\hat{\beta}_{\Delta\log(rexpp_{-1})}$ is about 5.13, which gives Δlog($rexpp_{-1}$) a *t* statistic of about 2.15. The two-sided *p*-value is about .032.

(vi) We can use four years of data for this test. Doing a pooled OLS regression of $\hat{r}_{it}$ on $\hat{r}_{i,t-1}$, using years 1995, 1996, 1997, and 1998, gives $\hat{\rho} = -.423$ (se = .019), which is strong negative serial correlation.

(vii) The fully robust "*F*" test for $\Delta\log(enroll)$ and $\Delta lunch$, reported by Stata 7.0, is .93. With 2 and 549 *df*, this translates into *p*-value = .40. So we would be justified in dropping these variables, but they are not doing any harm.

C13.13 (i) We can estimate all parameters except β_0 and β_1. The intercept for the base year cannot be estimated, and neither can coefficients on the time-constant variable $educ_i$.

(ii) We want to test $H_0 : \gamma_1 = \gamma_2, ..., \gamma_7 = 0$, so there are seven restrictions to be tested. Using FD (which eliminates $educ_i$) and obtaining the *F* statistic gives *F* = .31 (*p*-value = .952). Therefore, there is no evidence that the return to education varied over this time period. (Also, each coefficient is individually statistically insignificant at the 25% level.)

(iii) The fully robust *F* statistic is about 1.00, with *p*-value = .432. So the conclusion really does not change. The γ_j are jointly insignificant.

(iv) The estimated union differential in 1980 is simply the coefficient on $\Delta union_{it}$, or about .106 (10.6%). For 1987, we add the coefficients on $\Delta union_t$ and $\Delta d87_t \cdot union_{it}$, or $-.041$ (–4.1%). The difference, –14.7%, is statistically significant (*t* = –2.15, whether we use the usual pooled OLS standard error or the fully robust one).

(v) The usual *F* statistic is 1.03 (*p*-value = .405) and the statistic robust to heteroskedasticity and serial correlation is 1.15 (*p*-value = .331). Therefore, when we test all interaction terms as a group (seven of them), we fail to reject the null that the union differential was constant over this period. Most of the interactions are individually insignificant; in fact, only those for 1986 and 1987 are close. We can get joint insignificance by lumping several statistically insignificant variables in with one or two statistically significant ones. But it is hard to ignore the practically large change from 1980 to 1987. (There might be a problem in this example with the strict exogeneity assumption: perhaps union membership next year depends on unexpected wage changes this year.)

C13.15 (i) Tabulating the observations by year, the year with the most number of observations is the most recent, 2006 (2,986 people). The year with the fewest observation is 2004 with only 1,337 people. Out of 17,137 total people in the sample, 5,260, or about 30.69%, report being "very happy."

(ii) Regressing *vhappy* on the year dummies *y*96, *y*98, *y*00, *y*02, *y*04, and *y*06 gives R^2 = .0005, but we cannot use the *R*-squared form of the *F* statistic if we want robustness to heteroskedasticity. Instead, use the robust option in Stata followed by the test command. The *p*-value for the heteroskedasticity-robust test is about .199. Therefore, unless we choose a

significance level of 20%, we fail to reject the null. There is no strong evidence that the proportion of "very happy" people has changed over time.

(iii) The coefficient on *occattend* is about .0043 (robust se = .0080) and that on *regattend* is about .112 (robust se = .011). This implies that the probability that someone who regularly attends a religious service is .112, or 11.2 percentage points, higher than someone who never attends a religious service. The coefficient on *occattend* is also relative to the base group: those who never attend, but its affect is very small and statistically significant. Thus, it appears that "regular" attendance at religious services has a large, statistically significant effect on being very happy; occasional attendance has no effect.

(iv) One must be careful to define *highinc* so that it is missing whenever *income* is missing, as it is for 2,092 of the observations. Including it along with the other controls drops the coefficient on *regattend* somewhat: to about .096 (robust se = .015). Its *t* statistic is still well above six, so it is still very significant (but less so without the controls, and on the larger sample).

(v) The coefficients (robust ses) of the four new controls in part (iv) are

highinc: .1011 (.0100)
*unem*10: –.0881 (.0095)
educ: .0039 (.0017)
teens: –.0168 (.0093)

Higher income makes people happier and the coefficient is very statistically significant. Being unemployed recently makes people less happy. An unemployment spell in the last 10 years lowers the probability of being very happy by about .088. Education positively effects happiness although the partial effect is not large. The difference between a high school and four-year college education is only .0156. Having teenagers reduces the probability of being happy, but the effect is only marginally significant: $t = -1.81$.

(vi) In this data set, there is a gender variable and an indicator for whether the respondent is black. First just include *female* and *black* as explanatory variables. The coefficient on *female* is small, .0022 (robust se = .0094) and statistically insignificant. By contrast, the coefficient on *black* is –.0502, which implies that, controlling for other factors, blacks are five percentage points less likely to be "very happy." The robust *t* statistic of about –3.84 shows that it is statistically significant, too.

We can interact *female* and *black* and add the interaction to the regression. Now the three gender/race variables are individually insignificant (but they are jointly very significant, with robust *p*-value = .0011). The coefficient on the interaction, –.0202, suggests that black females are less happy than black males (who are both less happy than non-blacks), but the estimate is not statistically significant.