Επιλεγμένες Ασκήσεις

Κεφάλαιο 8

8.1 Parts (ii) and (iii). The homoskedasticity assumption played no role in Chapter 5 in showing that OLS is consistent. But we know that heteroskedasticity causes statistical inference based on the usual *t* and *F* statistics to be invalid, even in large samples. As heteroskedasticity is a violation of the Gauss-Markov assumptions, OLS is no longer BLUE.

8.3 False. The unbiasedness of WLS and OLS hinges crucially on Assumption MLR.4, and, as we know from Chapter 4, this assumption is often violated when an important variable is omitted. When MLR.4 does not hold, both WLS and OLS are biased. Without specific information on how the omitted variable is correlated with the included explanatory variables, it is not possible to determine which estimator has a small bias. It is possible that WLS would have more bias than OLS or less bias. Because we cannot know, we should not claim to use WLS in order to solve "biases" associated with OLS.

8.5 (i) No. For each coefficient, the usual standard errors and the heteroskedasticity-robust ones are practically very similar.

(ii) The effect is –.029(4) = –.116, so the probability of smoking falls by about .116.

(iii) As usual, we compute the turning point in the quadratic: .020/[2(.00026)] $\approx$ 38.46, so about 38 and one-half years.

(iv) Holding other factors in the equation fixed, a person in a state with restaurant smoking restrictions has a .101 lower chance of smoking. This is similar to the effect of having four more years of education.

(v) We just plug the values of the independent variables into the OLS regression line:

$$\widehat{smokes} = .656 - .069 \cdot \log(67.44) + .012 \cdot \log(6,500) - .029(16) + .020(77) - .00026(77^2) \approx .0052.$$

Thus, the estimated probability of smoking for this person is close to zero. (In fact, this person is not a smoker, so the equation predicts well for this particular observation.)

SOLUTIONS TO COMPUTER EXERCISES

8.6 (i) Given the equation

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u,$$

the assumption that the variance of u given all explanatory variables depends only on gender is

$$Var(u \mid totwrk, educ, age, yngkid, male) = Var(u \mid male) = \delta_0 + \delta_1 male$$

Then the variance for women is simply δ_0 and that for men is $\delta_0 + \delta_1$; the difference in variances is δ_1.

(ii) After estimating the above equation by OLS, we regress $\hat{u}_i^2$ on $male_i$, i = 1,2, …,706 (including, of course, an intercept). We can write the results as

$$\begin{array}{lllll} \hat{u}^2 = & 189{,}359.2 & - & 28{,}849.6\ male & + \\ residual & & & & \\ & (20{,}546.4) & & (27{,}296.5) & \end{array}$$

$$n = 706, \quad R^2 = .0016.$$

Because the coefficient on *male* is negative, the estimated variance is higher for women.

(iii) No. The t statistic on *male* is only about –1.06, which is not significant at even the 20% level against a two-sided alternative.

8.8 After estimating equation (8.18), we obtain the squared OLS residuals $\hat{u}^2$. The full-blown White test is based on the R-squared from the auxiliary regression (with an intercept),

$$\hat{u}^2 \text{ on } llotsize,\ lsqrft,\ bdrms,\ llotsize^2,\ lsqrft^2,\ bdrms^2,$$
$$llotsize \cdot lsqrft,\ llotsize \cdot bdrms, \text{ and } lsqrft \cdot bdrms,$$

where "*l* " in front of *lotsize* and *sqrft* denotes the natural log. [See equation (8.19).] With 88 observations the n-R-squared version of the White statistic is $88(.109) \approx 9.59$, and this is the outcome of an (approximately) χ_9^2 random variable. The p-value is about .385, which provides little evidence against the homoskedasticity assumption.

8.10 (i) By regressing *sprdcvr* on an intercept only we obtain $\hat{\mu} \approx .515$ se $\approx .021$). The asymptotic t statistic for H_0: $\mu = .5$ is $(.515 - .5)/.021 \approx .71$, which is not significant at the 10% level, or even the 20% level.

(ii) 35 games were played on a neutral court.

(iii) The estimated LPM is

$\widehat{sprdcvr}$ =.490 +.035 *favhome* +.118 *neutral* –.023 *fav25* +.018 *und25*

(.045) (.050) (.095) (.050) (.092)

$n = 553$, $R^2 = .0034$.

The variable *neutral* has by far the largest effect – if the game is played on a neutral court, the probability that the spread is covered is estimated to be about .12 higher – and, except for the intercept, its *t* statistic is the only *t* statistic greater than one in absolute value (about 1.24).

(iv) Under H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, the response probability does not depend on any explanatory variables, which means neither the mean nor the variance depends on the explanatory variables. [See equation (8.38).]

(v) The *F* statistic for joint significance, with 4 and 548 *df*, is about .47 with *p*-value $\approx$ .76. There is essentially no evidence against H_0.

(vi) Based on these variables, it is not possible to predict whether the spread will be covered. The explanatory power is very low, and the explanatory variables are jointly very insignificant. The coefficient on *neutral* may indicate something is going on with games played on a neutral court, but we would not want to bet money on it unless it could be confirmed with a separate, larger sample.

8.12 (i) The heteroskedasticity-robust standard error for $\hat{\beta}_{white} \approx .129$ is about .026, which is notably higher than the nonrobust standard error (about .020). The heteroskedasticity-robust 95% confidence interval is about .078 to .179, while the nonrobust CI is, of course, narrower, about .090 to .168. The robust CI still excludes the value zero by some margin.

(ii) There are no fitted values less than zero, but there are 231 greater than one. Unless we do something to those fitted values, we cannot directly apply WLS, as $\hat{h}_i$ will be negative in 231 cases.

8.14 (i) I now get $R^2 = .0527$, but the other estimates seem okay.

(ii) One way to ensure that the unweighted residuals are being provided is to compare them with the OLS residuals. They will not be the same, of course, but they should not be wildly different.

(iii) The *R*-squared from the regression $\breve{u}_i^2$ on $\breve{y}_i$, $\breve{y}_i^2$, $i = 1,...,807$ is about .027. We use this as $R^2_{\hat{u}^2}$ in equation (8.15) but with $k = 2$. This gives $F = 11.15$, and so the *p*-value is essentially zero.

(iv) The substantial heteroskedasticity found in part (iii) shows that the feasible GLS procedure described on page 279 does not, in fact, eliminate the heteroskedasticity. Therefore, the usual standard errors, *t* statistics, and *F* statistics reported with weighted least squares are not valid, even asymptotically.

(v) Weighted least squares estimation with robust standard errors gives

$$\widehat{cigs} = \underset{(37.31)}{5.64} + \underset{(.54)}{1.30}\log(income) - \underset{(8.97)}{2.94}\log(cigpric) - \underset{(.149)}{.463}\,educ + \underset{(.115)}{.482}\,age - \underset{(.0012)}{.0056}\,age^2 - \underset{(.72)}{3.46}\,restaurn$$

$n = 807,\ R^2 = .1134$

The substantial differences in standard errors compared with equation (8.36) further indicate that our proposed correction for heteroskedasticity did not fully solve the heteroskedasticity problem. With the exception of *restaurn*, all standard errors got notably bigger; for example, the standard error for log(*cigpric*) doubled. All variables that were statistically significant with the nonrobust standard errors remain significant, but the confidence intervals are much wider in several cases.

8.15 (i) The usual OLS standard errors are in (·), the heteroskedasticity-robust standard errors are in [·]:

$$\widehat{nettfa} = \underset{\substack{(15.31)\\[19.09]}}{1.50} + \underset{\substack{(.062)\\[.102]}}{.774}\,inc - \underset{\substack{(0.77)\\[1.08]}}{1.60}\,age + \underset{\substack{(.009)\\[.014]}}{.029}\,age^2 + \underset{\substack{(2.05)\\[2.06]}}{2.47}\,male + \underset{\substack{(2.13)\\[2.19]}}{6.98}\,e401k$$

$n = 2{,}017,\ \ R^2 = .128.$

(ii) The smallest $\hat{h}_i$ is about 12.83 and the largest is about 58,059.74. Thus, there is wide variation in the estimated conditional variances.

(iii) The usual WLS standard errors are in (·), the standard errors robust to misspecified variance are in [·]:

$$\widehat{nettfa} = \underset{\substack{(9.94)\\[8.19]}}{-2.58} + \underset{\substack{(.058)\\[.062]}}{.456}\,inc - \underset{\substack{(.541)\\[.408]}}{.613}\,age + \underset{\substack{(.007)\\[.005]}}{.013}\,age^2 + \underset{\substack{(1.03)\\[0.82]}}{1.42}\,male + \underset{\substack{(1.23)\\[1.14]}}{4.26}\,e401k$$

$n = 2{,}017,\ \ R^2 = .062.$

Interestingly, except for the income coefficient, the robust standard errors are actually smaller than the usual standard error. This could just be sampling variation, or it could be that the variance function is misspecified in such a way that, when it is used in WLS, the usual standard errors overestimate the actual sampling variation.

(iv) The robust standard error for the *e401k* coefficient is 2.19, while that for WLS is 1.14. Thus, the WLS standard error is just over half as large as the OLS standard error. Assuming that the zero conditional mean assumption actually holds – something that is not clear given some nontrivial changes in the WLS estimates as compared with OLS – the smaller robust standard error for WLS suggests it is the more efficient procedure, whether or not we have properly specified the "skedastic" function.

Κεφάλαιο 9

9.1 There is functional form misspecification if $\beta_6 \neq 0$ or $\beta_7 \neq 0$, where these are the population parameters on $ceoten^2$ and $comten^2$, respectively. Therefore, we test the joint significance of these variables using the *R*-squared form of the *F* test: $F = [(.375 - .353)/(1 - .375)][(177 - 8)/2] \approx 2.97$. With 2 and ∞ *df*, the 10% critical value is 2.30 awhile the 5% critical value is 3.00. Thus, the *p*-value is slightly above .05, which is reasonable evidence of functional form misspecification. (Of course, whether this has a practical impact on the estimated partial effects for various levels of the explanatory variables is a different matter.)

9.3 (i) Eligibility for the federally funded school lunch program is very tightly linked to being economically disadvantaged. Therefore, the percentage of students eligible for the lunch program is very similar to the percentage of students living in poverty.

(ii) We can use our usual reasoning on omitting important variables from a regression equation. The variables log(*expend*) and *lnchprg* are negatively correlated: school districts with poorer children spend, on average, less on schools. Further, $\beta_3 < 0$. From Table 3.2, omitting *lnchprg* (the proxy for *poverty*) from the regression produces an upward biased estimator of β_1 [ignoring the presence of log(*enroll*) in the model]. So when we control for the poverty rate, the effect of spending falls.

(iii) Once we control for *lnchprg*, the coefficient on log(*enroll*) becomes negative and has a *t* of about –2.17, which is significant at the 5% level against a two-sided alternative. The coefficient implies that $\Delta\widehat{math10} \approx -(1.26/100)(\%\Delta enroll) = -.0126(\%\Delta enroll)$. Therefore, a 10% increase in enrollment leads to a drop in *math10* of .126 percentage points.

(iv) Both *math10* and *lnchprg* are percentages. Therefore, a ten percentage point increase in *lnchprg* leads to about a 3.23 percentage point fall in *math10*, a sizeable effect.

(v) In column (1) we are explaining very little of the variation in pass rates on the MEAP math test: less than 3%. In column (2), we are explaining almost 19% (which still leaves much variation unexplained). Clearly most of the variation in *math10* is explained by variation in *lnchprg*. This is a common finding in studies of school performance: family income (or related factors, such as living in poverty) are much more important in explaining student performance than are spending per student or other school characteristics.

9.5 The sample selection in this case is arguably endogenous. Because prospective students may look at campus crime as one factor in deciding where to attend college, colleges with high crime rates have an incentive not to report crime statistics. If this is the case, then the chance of appearing in the sample is negatively related to *u* in the crime equation. (For a given school size, higher *u* means more crime, and therefore a smaller probability that the school reports its crime figures.)

9.6 (i) To obtain the RESET F statistic, we estimate the model in Computer Exercise 7.5 and obtain the fitted values, say $\widehat{lsalary}_i$. To use the version of RESET in (9.3), we add $(\widehat{lsalary}_i)^2$ and $(\widehat{lsalary}_i)^3$ and obtain the F test for joint significance of these variables. With 2 and 203 *df*, the F statistic is about 1.33 and p-value $\approx$ .27, which means that there is not much concern about functional form misspecification.

(ii) Interestingly, the heteroskedasticity-robust F-type statistic is about 2.24 with p-value $\approx$ .11, so there is stronger evidence of some functional form misspecification with the robust test. But it is probably not strong enough to worry about.

9.8 (i) If the grants were awarded to firms based on firm or worker characteristics, *grant* could easily be correlated with such factors that affect productivity. In the simple regression model, these are contained in u.

(ii) The simple regression estimates using the 1988 data are

$$\begin{aligned} \widehat{\log(scrap)} &= .409 \quad + .057\ grant \\ &\quad\ (.241) \quad\ (.406) \end{aligned}$$

$$n = 54, \quad R^2 = .0004.$$

The coefficient on *grant* is actually positive, but not statistically different from zero.

(iii) When we add $\log(scrap_{87})$ to the equation, we obtain

$$\begin{aligned} \widehat{\log(scrap_{88})} &= .021 \quad - .254\ grant_{88} \quad + .831\ \log(scrap_{87}) \\ &\quad\ (.089) \quad\ (.147) \quad\ (.044) \end{aligned}$$

$$n = 54, \quad R^2 = .873,$$

where the year subscripts are for clarity. The t statistic for H_0: $\beta_{grant} = 0$ is $-.254/.147 \approx -1.73$. We use the 5% critical value for 40 *df* in Table G.2: -1.68. Because $t = -1.73 < -1.68$, we reject H_0 in favor of H_1: $\beta_{grant} < 0$ at the 5% level.

(iv) The t statistic is $(.831 - 1)/.044 \approx -3.84$, which is a strong rejection of H_0.

(v) With the heteroskedasticity-robust standard error, the t statistic for $grant_{88}$ is $-.254/.142 \approx -1.79$, so the coefficient is even more significantly less than zero when we use the heteroskedasticity-robust standard error. The t statistic for H_0: $\beta_{\log(scrap_{87})} = 1$ is $(.831 - 1)/.071 \approx -2.38$, which is notably smaller than before, but it is still pretty significant.

9.10 With *sales* defined to be in billions of dollars, we obtain the following estimated equation using all companies in the sample:

$$\widehat{rdintens} = 2.06 + .317\,sales - .0074\,sales^2 + .053\,profmarg$$
$$(0.63)\quad (.139)\quad (.0037)\quad (.044)$$

$$n = 32,\ R^2 = .191,\ \bar{R}^2 = .104.$$

When we drop the largest company (with sales of roughly \$39.7 billion), we obtain

$$\widehat{rdintens} = 1.98 + .361\,sales - .0103\,sales^2 + .055\,profmarg$$
$$(0.72)\quad (.239)\quad (.0131)\quad (.046)$$

$$n = 31,\ R^2 = .191,\ \bar{R}^2 = .101.$$

When the largest company is left in the sample, the quadratic term is statistically significant, even though the coefficient on the quadratic is less in absolute value than when we drop the largest firm. What is happening is that by leaving in the large sales figure, we greatly increase the variation in both *sales* and $sales^2$; as we know, this reduces the variances of the OLS estimators (see Section 3.4). The *t* statistic on $sales^2$ in the first regression is about –2, which makes it almost significant at the 5% level against a two-sided alternative. If we look at Figure 9.1, it is not surprising that a quadratic is significant when the large firm is included in the regression: *rdintens* is relatively small for this firm even though its sales are very large compared with the other firms. Without the largest firm, a linear relationship between *rdintens* and *sales* seems to suffice.

9.12 (i) 205 observations out of the 1,989 records in the sample have *obrate* > 40. (Data are missing for some variables, so not all of the 1,989 observations are used in the regressions.)

(ii) When observations with *obrat* > 40 are excluded from the regression in part (iii) of Problem 7.16, we are left with 1,768 observations. The coefficient on *white* is about .129 (se ≈ .020). To three decimal places, these are the same estimates we got when using the entire sample (see Computer Exercise C7.8). Perhaps this is not very surprising since we only lost 203 out of 1,971 observations. However, regression results can be very sensitive when we drop over 10% of the observations, as we have here.

(iii) The estimates from part (ii) show that $\hat{\beta}_{white}$ does not seem very sensitive to the sample used, although we have tried only one way of reducing the sample.

9.14 (i) The equation estimated by OLS is

$$\widehat{nettfa} = 21.198 - .270\,inc + .0102\,inc^2 - 1.940\,age + .0346\,age^2$$
$$(9.992)\quad (.075)\quad (.0006)\quad (.483)\quad (.0055)$$

$$+ 3.369\,male + 9.713\,e401k$$
$$(1.486)\quad (1.277)$$

$$n = 9{,}275,\ R^2 = .202$$

The coefficient on *e401k* means that, holding other things in the equation fixed, the average level of net financial assets is about $9,713 higher for a family eligible for a 401(k) than for a family not eligible.

(ii) The OLS regression of $\hat{u}_i^2$ on inc_i, inc_i^2, age_i, age_i^2, $male_i$, and $e401k_i$ gives $R^2_{\hat{u}^2} = .0374$, which translates into $F = 59.97$. The associated p-value, with 6 and 9,268 *df*, is essentially zero. Consequently, there is strong evidence of heteroskedasticity, which means that u and the explanatory variables cannot be independent [even though $E(u|x_1,x_2,\ldots,x_k) = 0$ is possible].

(iii) The equation estimated by LAD is

$$\widehat{nettfa} = \underset{(1.382)}{12.491} - \underset{(.010)}{.262}\, inc + \underset{(.00008)}{.00709}\, inc^2 - \underset{(.067)}{.723}\, age + \underset{(.0008)}{.0111}\, age^2$$

$$+ \underset{(.205)}{1.018}\, male + \underset{(.177)}{3.737}\, e401k$$

$n = 9{,}275$, Psuedo $R^2 = .109$

Now, the coefficient on *e401k* means that, at given income, age, and gender, the median difference in net financial assets between families with and without 401(k) eligibility is about $3,737.

(iv) The findings from parts (i) and (iii) are not in conflict. We are finding that 401(k) eligibility has a larger effect on mean wealth than on median wealth. Finding different mean and median effects for a variable such as *nettfa*, which has a highly skewed distribution, is not surprising. Apparently, 401(k) eligibility has some large effects at the upper end of the wealth distribution, and these are reflected in the mean. The median is much less sensitive to effects at the upper end of the distribution.

Κεφάλαιο 10

10.1 (i) Disagree. Most time series processes are correlated over time, and many of them strongly correlated. This means they cannot be independent across observations, which simply represent different time periods. Even series that do appear to be roughly uncorrelated – such as stock returns – do not appear to be independently distributed, as you will see in Chapter 12 under dynamic forms of heteroskedasticity.

(ii) Agree. This follows immediately from Theorem 10.1. In particular, we do not need the homoskedasticity and no serial correlation assumptions.

(iii) Disagree. Trending variables are used all the time as dependent variables in a regression model. We do need to be careful in interpreting the results because we may simply find a spurious association between y_t and trending explanatory variables. Including a trend in the regression is a good idea with trending dependent or independent variables. As discussed in Section 10.5, the usual R-squared can be misleading when the dependent variable is trending.

(iv) Agree. With annual data, each time period represents a year and is not associated with any season.

10.3 Write

$$y^* = \alpha_0 + (\delta_0 + \delta_1 + \delta_2)z^* = \alpha_0 + LRP \cdot z^*,$$

and take the change: $\Delta y^* = LRP \cdot \Delta z^*$.

10.5 The functional form was not specified, but a reasonable one is

$$\log(hsestrts_t) = \alpha_0 + \alpha_1 t + \delta_1 Q2_t + \delta_2 Q3_t + \delta_3 Q3_t + \beta_1 int_t + \beta_2 \log(pcinc_t) + u_t,$$

Where $Q2_t$, $Q3_t$, and $Q4_t$ are quarterly dummy variables (the omitted quarter is the first) and the other variables are self-explanatory. This inclusion of the linear time trend allows the dependent variable and $\log(pcinc_t)$ to trend over time (int_t probably does not contain a trend), and the quarterly dummies allow all variables to display seasonality. The parameter β_2 is an elasticity and $100 \cdot \beta_1$ is a semi-elasticity.

10.7 Let *post79* be a dummy variable equal to one for years after 1979, and zero otherwise. Adding *post79* to equation 10.15) gives

$$\begin{array}{lllll} \widehat{i3}_t & = 1.30 & + .608\ inf_t & + .363\ def_t & + 1.56\ post79_t \\ & (0.43) & (.076) & (.120) & (0.51) \end{array}$$

$$n = 56, \quad R^2 = .664, \quad \overline{R}^2 = .644.$$

The coefficient on *post79* is statistically significant (*t* statistic ≈ 3.06) and economically large: accounting for inflation and deficits, *i3* was about 1.56 points higher on average in years after 1979. The coefficient on *def* falls once *post79* is included in the regression.

10.9 Adding log(*prgnp*) to equation (10.38) gives

$$\widehat{\log(prepop_t)} = \underset{(1.26)}{-6.66} \underset{(.040)}{-.212 \log(mincov_t)} \underset{(.222)}{+.486 \log(usgnp_t)} \underset{(.080)}{+.285 \log(prgnp_t)} \underset{(.005)}{-.027\, t}$$

$$n = 38, \quad R^2 = .889, \quad \bar{R}^2 = .876.$$

The coefficient on log($prgnp_t$) is very statistically significant (*t* statistic ≈ 3.56). Because the dependent and independent variable are in logs, the estimated elasticity of *prepop* with respect to *prgnp* is .285. Including log(*prgnp*) actually increases the size of the minimum wage effect: the estimated elasticity of *prepop* with respect to *mincov* is now –.212, as compared with –.169 in equation (10.38).

10.11 (i) The coefficient on the time trend in the regression of log(*uclms*) on a linear time trend and 11 monthly dummy variables is about –.0139 (se ≈ .0012), which implies that monthly unemployment claims fell by about 1.4% per month on average. The trend is very significant. There is also very strong seasonality in unemployment claims, with 6 of the 11 monthly dummy variables having absolute *t* statistics above 2. The *F* statistic for joint significance of the 11 monthly dummies yields *p*-value ≈ .0009.

(ii) When *ez* is added to the regression, its coefficient is about –.508 (se ≈ .146). Because this estimate is so large in magnitude, we use equation (7.10): unemployment claims are estimated to fall 100[1 – exp(–.508)] ≈ 39.8% after enterprise zone designation.

(iii) We must assume that around the time of *EZ* designation there were not other external factors that caused a shift down in the trend of log(*uclms*). We have controlled for a time trend and seasonality, but this may not be enough.

10.13 (i) The estimated equation is

$$\widehat{gc}_t = \underset{(.0019)}{.0081} \underset{(.067)}{+.571\, gy_t}$$

$$n = 36, \quad R^2 = .679.$$

This equation implies that if income growth increases by one percentage point, consumption growth increases by .571 percentage points. The coefficient on gy_t is very statistically significant (*t* statistic ≈ 8.5).

(ii) Adding gy_{t-1} to the equation gives

$$\widehat{gc}_t = .0064 + .552\, gy_t + .096\, gy_{t-1}$$
$$\quad\quad (.0023) \quad\quad (.070) \quad\quad (.069)$$
$$n = 35, \quad R^2 = .695.$$

The t statistic on gy_{t-1} is only about 1.39, so it is not significant at the usual significance levels. (It is significant at the 20% level against a two-sided alternative.) In addition, the coefficient is not especially large. At best there is weak evidence of adjustment lags in consumption.

(iii) If we add $r3_t$ to the model estimated in part (i) we obtain

$$\widehat{gc}_t = .0082 + .578\, gy_t + .00021\, r3_t$$
$$\quad\quad (.0020) \quad\quad (.072) \quad\quad (.00063)$$
$$n = 36, \quad R^2 = .680.$$

The t statistic on $r3_t$ is very small. The estimated coefficient is also practically small: a one-point increase in $r3_t$ reduces consumption growth by about .021 percentage points.

10.15 (i) The sign of β_2 is fairly clear-cut: as interest rates rise, stock returns fall, so $\beta_2 < 0$. Higher interest rates imply that T-bill and bond investments are more attractive, and also signal a future slowdown in economic activity. The sign of β_1 is less clear. While economic growth can be a good thing for the stock market, it can also signal inflation, which tends to depress stock prices.

(ii) The estimated equation is

$$\widehat{rsp500}_t = 18.84 + .036\, pcip_t - 1.36\, i3_t$$
$$\quad\quad (3.27) \quad\quad (.129) \quad\quad (0.54)$$
$$n = 557, \quad R^2 = .012.$$

A one percentage point increase in industrial production growth is predicted to increase the stock market return by .036 percentage points (a very small effect). On the other hand, a one percentage point increase in interest rates decreases the stock market return by an estimated 1.36 percentage points.

(iii) Only *i3* is statistically significant with t statistic ≈ -2.52.

(iv) The regression in part (i) has nothing directly to say about predicting stock returns because the explanatory variables are dated contemporaneously with *rsp500*. In other words, we do not know $i3_t$ before we know $rsp500_t$. What the regression in part (i) says is that a change in *i3* is associated with a contemporaneous change in *rsp500*.

10.17 (i) The variable *beltlaw* becomes one at $t = 61$, which corresponds to January, 1986. The variable *spdlaw* goes from zero to one at $t = 77$, which corresponds to May, 1987.

(ii) The OLS regression gives

$$\widehat{\log(totacc)} = 10.469 + .00275\, t - .0427\, feb + .0798\, mar + .0185\, apr$$
$$(.019) \quad (.00016) \quad (.0244) \quad (.0244) \quad (.0245)$$

$$+ .0321\, may + .0202\, jun + .0376\, jul + .0540\, aug$$
$$(.0245) \quad (.0245) \quad (.0245) \quad (.0245)$$

$$+ .0424\, sep + .0821\, oct + .0713\, nov + .0962\, dec$$
$$(.0245) \quad (.0245) \quad (.0245) \quad (.0245)$$

$n = 108$, $R^2 = .797$

When multiplied by 100, the coefficient on *t* gives roughly the average monthly percentage growth in *totacc*, ignoring seasonal factors. In other words, once seasonality is eliminated, *totacc* grew by about .275% per month over this period, or, 12(.275) = 3.3% at an annual rate.

There is pretty clear evidence of seasonality. Only February has a lower number of total accidents than the base month, January. The peak is in December: roughly, there are 9.6% accidents more in December over January in the average year. The *F* statistic for joint significance of the monthly dummies is $F = 5.15$. With 11 and 95 *df*, this give a *p*-value essentially equal to zero.

(iii) I will report only the coefficients on the new variables:

$$\widehat{\log(totacc)} = 10.640 + \ldots + .00333\, wkends - .0212\, unem$$
$$(.063) \quad (.00378) \quad (.0034)$$

$$- .0538\, spdlaw + .0954\, beltlaw$$
$$(.0126) \quad (.0142)$$

$n = 108$, $R^2 = .910$

The negative coefficient on *unem* makes sense if we view *unem* as a measure of economic activity. As economic activity increases – *unem* decreases – we expect more driving, and therefore more accidents. The estimate that a one percentage point increase in the unemployment rate reduces total accidents by about 2.1%. A better economy does have costs in terms of traffic accidents.

(iv) At least initially, the coefficients on *spdlaw* and *beltlaw* are not what we might expect. The coefficient on *spdlaw* implies that accidents dropped by about

5.4% *after* the highway speed limit was increased from 55 to 65 miles per hour. There are at least a couple of possible explanations. One is that people because safer drivers after the increased speed limiting, recognizing that the must be more cautious. It could also be that some other change – other than the increased speed limit or the relatively new seat belt law – caused lower total number of accidents, and we have not properly accounted for this change.

The coefficient on *beltlaw* also seems counterintuitive at first. But, perhaps people became less cautious once they were forced to wear seatbelts.

(v) The average of *prcfat* is about .886, which means, on average, slightly less than one percent of all accidents result in a fatality. The highest value of *prcfat* is 1.217, which means there was one month where 1.2% of all accidents resulting in a fatality.

(vi) As in part (iii), I do not report the coefficients on the time trend and seasonal dummy variables:

$$\widehat{prcfat} = \underset{(.103)}{1.030} + \ldots + \underset{(.0055)}{.00063\ wkends} - \underset{}{.0154\ unem}$$

$$+ \underset{(.0206)}{.0671\ spdlaw} - \underset{(.0232)}{.0295\ beltlaw}$$

$n = 108,\ R^2 = .717$

Higher speed limits are estimated to increase the percent of fatal accidents, by .067 percentage points. This is a statistically significant effect. The new seat belt law is estimated to decrease the percent of fatal accidents by about .03, but the two-sided *p*-value is about .21.

Interestingly, increased economic activity also increases the percent of fatal accidents. This may be because more commercial trucks are on the roads, and these probably increase the chance that an accident results in a fatality.

Κεφάλαιο 11

11.1 Because of covariance stationarity, $\gamma_0 = \text{Var}(x_t)$ does not depend on t, so $\text{sd}(x_{t+h}) = \sqrt{\gamma_0}$ for any $h \geq 0$. By definition, $\text{Corr}(x_t,x_{t+h}) = \text{Cov}(x_t,x_{t+h})/[\text{sd}(x_t) \cdot \text{sd}(x_{t+h})] = \gamma_h /(\sqrt{\gamma_0} \cdot \sqrt{\gamma_0}) = \gamma_h / \gamma_0$.

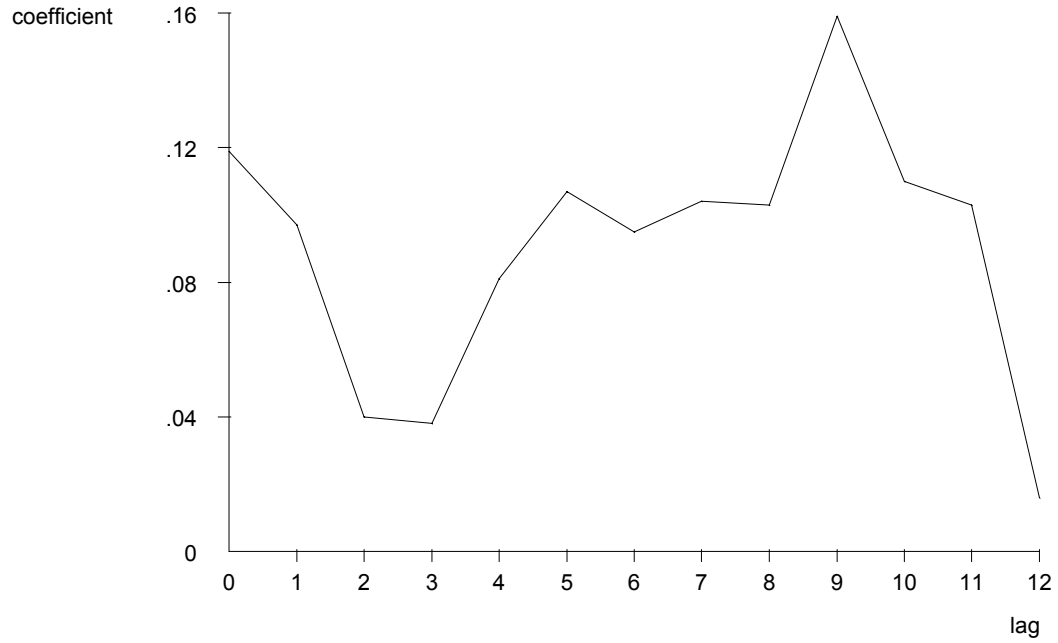
11.3 (i) $\text{E}(y_t) = \text{E}(z + e_t) = \text{E}(z) + \text{E}(e_t) = 0$. $\text{Var}(y_t) = \text{Var}(z + e_t) = \text{Var}(z) + \text{Var}(e_t) + 2\text{Cov}(z,e_t) = \sigma_z^2 + \sigma_e^2 + 2 \cdot 0 = \sigma_z^2 + \sigma_e^2$. Neither of these depends on t.

(ii) We assume $h > 0$; when $h = 0$ we obtain $\text{Var}(y_t)$. Then $\text{Cov}(y_t,y_{t+h}) = \text{E}(y_t y_{t+h}) = \text{E}[(z + e_t)(z + e_{t+h})] = \text{E}(z^2) + \text{E}(ze_{t+h}) + \text{E}(e_t z) + \text{E}(e_t e_{t+h}) = \text{E}(z^2) = \sigma_z^2$ because $\{e_t\}$ is an uncorrelated sequence (it is an independent sequence and z is uncorrelated with e_t for all t. From part (i) we know that $\text{E}(y_t)$ and $\text{Var}(y_t)$ do not depend on t and we have shown that $\text{Cov}(y_t,y_{t+h})$ depends on neither t nor h. Therefore, $\{y_t\}$ is covariance stationary.

(iii) From Problem 11.1 and parts (i) and (ii), $\text{Corr}(y_t,y_{t+h}) = \text{Cov}(y_t,y_{t+h})/\text{Var}(y_t) = \sigma_z^2 /(\sigma_z^2 + \sigma_e^2) > 0$.

(iv) No. The correlation between y_t and y_{t+h} is the same positive value obtained in part (iii) now matter how large is h. In other words, no matter how far apart y_t and y_{t+h} are, their correlation is always the same. Of course, the persistent correlation across time is due to the presence of the time-constant variable, z.

11.5 (i) The following graph gives the estimated lag distribution:



By some margin, the largest effect is at the ninth lag, which says that a temporary increase in wage inflation has its largest effect on price inflation nine months later. The smallest effect is at the twelfth lag, which hopefully indicates (but does not guarantee) that we have accounted for enough lags of *gwage* in the FLD model.

(ii) Lags two, three, and twelve have *t* statistics less than two. The other lags are statistically significant at the 5% level against a two-sided alternative. (Assuming either that the CLM assumptions hold for exact tests or Assumptions TS.1′ through TS.5′ hold for asymptotic tests.)

(iii) The estimated LRP is just the sum of the lag coefficients from zero through twelve: 1.172. While this is greater than one, it is not much greater, and the difference from unity could be due to sampling error.

(iv) The model underlying and the estimated equation can be written with intercept α_0 and lag coefficients $\delta_0, \delta_1, \ldots, \delta_{12}$. Denote the LRP by $\theta_0 = \delta_0 + \delta_1 + \ldots + \delta_{12}$. Now, we can write $\delta_0 = \theta_0 - \delta_1 - \delta_2 - \ldots - \delta_{12}$. If we plug this into the FDL model we obtain (with $y_t = gprice_t$ and $z_t = gwage_t$)

$$\begin{aligned} y_t &= \alpha_0 + (\theta_0 - \delta_1 - \delta_2 - \ldots - \delta_{12})z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \ldots + \delta_{12} z_{t-12} + u_t \\ &= \alpha_0 + \theta_0 z_t + \delta_1(z_{t-1} - z_t) + \delta_2(z_{t-2} - z_t) + \ldots + \delta_{12}(z_{t-12} - z_t) + u_t. \end{aligned}$$

Therefore, we regress y_t on z_t, $(z_{t-1} - z_t)$, $(z_{t-2} - z_t)$, $\ldots$, $(z_{t-12} - z_t)$ and obtain the coefficient and standard error on z_t as the estimated LRP and its standard error.

(v) We would add lags 13 through 18 of $gwage_t$ to the equation, which leaves 273 – 6 = 267 observations. Now, we are estimating 20 parameters, so the *df* in the unrestricted model is $df_{ur} = 267$. Let R^2_{ur} be the *R*-squared from this regression. To obtain the restricted *R*-squared, R^2_r, we need to reestimate the model reported in the problem but with the same 267 observations used to estimate the unrestricted model. Then $F = [(R^2_{ur} - R^2_r)/(1 - R^2_{ur})](247/6)$. We would find the critical value from the $F_{6,247}$ distribution.

11.7 (i) We plug the first equation into the second to get

$$y_t - y_{t-1} = \lambda(\gamma_0 + \gamma_1 x_t + e_t - y_{t-1}) + a_t,$$

and, rearranging,

$$y_t = \lambda\gamma_0 + (1 - \lambda)y_{t-1} + \lambda\gamma_1 x_t + a_t + \lambda e_t,$$

$$\equiv \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t,$$

where $\beta_0 \equiv \lambda\gamma_0$, $\beta_1 \equiv (1 - \lambda)$, $\beta_2 \equiv \lambda\gamma_1$, and $u_t \equiv a_t + \lambda e_t$.

(ii) An OLS regression of y_t on y_{t-1} and x_t produces consistent, asymptotically normal estimators of the β_j. Under $E(e_t|x_t,y_{t-1},x_{t-1}, \ldots) = E(a_t|x_t,y_{t-1},x_{t-1}, \ldots) = 0$ it follows that $E(u_t|x_t,y_{t-1},x_{t-1}, \ldots) = 0$, which means that the model is dynamically complete [see equation (11.37)]. Therefore, the errors are serially uncorrelated. If the homoskedasticity assumption $Var(u_t|x_t,y_{t-1}) = \sigma^2$ holds, then the usual standard errors, *t* statistics and *F* statistics are asymptotically valid.

(iii) Because $\beta_1 = (1 - \lambda)$, if $\hat{\beta}_1 = .7$ then $\hat{\lambda} = .3$. Further, $\hat{\beta}_2 = \hat{\lambda}\hat{\gamma}_1$, or $\hat{\gamma}_1 = \hat{\beta}_2 / \hat{\lambda} = .2/.3 \approx .67$.

11.8 (i) The first order autocorrelation for log(*invpc*) is about .639. If we first detrend log(*invpc*) by regressing on a linear time trend, $\hat{\rho}_1 \approx .485$. Especially after detrending there is little evidence of a unit root in log(*invpc*). For log(*price*), the first order autocorrelation is about .949, which is very high. After detrending, the first order autocorrelation drops to .822, but this is still pretty large. We cannot confidently rule out a unit root in log(*price*).

(ii) The estimated equation is

$$\widehat{\log(invpc_t)} = \underset{(.040)}{-.853} + \underset{(0.96)}{3.88}\,\Delta\log(price_t) + \underset{(.0016)}{.0080}\,t$$

$$n = 41, \quad R^2 = .501.$$

The coefficient on $\Delta\log(price_t)$ implies that a one percentage point increase in the growth in price leads to a 3.88 percent increase in housing investment above its trend. [If $\Delta\log(price_t) = .01$ then $\Delta\widehat{\log(invpc_t)} = .0388$; we multiply both by 100 to convert the proportionate changes to percentage changes.]

(iii) If we first linearly detrend $\log(invpc_t)$ before regressing it on $\Delta\log(price_t)$ and the time trend, then $R^2 = .303$, which is substantially lower than that when we do not detrend. Thus, $\Delta\log(price_t)$ explains only about 30% of the variation in $\log(invpc_t)$ about its trend.

(iv) The estimated equation is

$$\begin{aligned} \widehat{\Delta\log(invpc_t)} &= .006 + 1.57\,\Delta\log(price_t) + .00004t \\ &\quad (.048) \quad (1.14) \quad (.00190) \end{aligned}$$

$$n = 41, \quad R^2 = .048.$$

The coefficient on $\Delta\log(price_t)$ has fallen substantially and is no longer significant at the 5% level against a positive one-sided alternative. The *R*-squared is much smaller; $\Delta\log(price_t)$ explains very little variation in $\Delta\log(invpc_t)$. Because differencing eliminates linear time trends, it is not surprising that the estimate on the trend is very small and very statistically insignificant.

11.10 (i) The estimated equation is

$$\begin{aligned} \widehat{return}_t &= .226 + .049\ return_{t-1} - .0097\ return_{t-1}^2 \\ &\quad (.087) \quad (.039) \quad (.0070) \end{aligned}$$

$$n = 689, \quad R^2 = .0063.$$

(ii) The null hypothesis is H_0: $\beta_1 = \beta_2 = 0$. Only if both parameters are zero does $E(return_t|return_{t-1})$ not depend on $return_{t-1}$. The *F* statistic is about 2.16 with *p*-value $\approx$ .116. Therefore, we cannot reject H_0 at the 10% level.

(iii) When we put $return_{t-1} \cdot return_{t-2}$ in place of $return_{t-1}^2$ the null can still be stated as in part (ii): no past values of *return*, or any functions of them, should help us predict $return_t$. The *R*-squared is about .0052 and $F \approx 1.80$ with *p*-value $\approx$ .166. Here, we do not reject H_0 at even the 15% level.

(iv) Predicting $return_t$ based on past returns does not appear promising. Even though the *F* statistic from part (ii) is almost significant at the 10% level, we have many observations. We cannot even explain 1% of the variation in $return_t$.

11.12 (i) The estimated equation is

$$\widehat{\Delta gfr} = \underset{(1.05)}{-1.27} \; \underset{(.027)}{-.035\,\Delta pe} \; \underset{(.028)}{-.013\,\Delta pe_{-1}} \; \underset{(.027)}{-.111\,\Delta pe_{-2}} \; \underset{(.0242)}{+.0079\,t}$$

$$n = 69, \quad R^2 = .234, \quad \overline{R}^2 = .186.$$

The time trend coefficient is very insignificant, so it is not needed in the equation.

(iii) The estimated equation is

$$\widehat{\Delta gfr} = \underset{(.582)}{-.650} \; \underset{(.032)}{-.075\,\Delta pe} \; \underset{(.033)}{-.051\,\Delta pe_{-1}} \; \underset{(.028)}{+.088\,\Delta pe_{-2}} \; \underset{(2.83)}{+4.84\,ww2} \; \underset{(1.00)}{-1.68\,pill}$$

$$n = 69, \quad R^2 = .296, \quad \overline{R}^2 = .240.$$

The F statistic for joint significance is $F = 2.82$ with p-value $\approx .067$. So *ww2* and *pill* are not jointly significant at the 5% level, but they are at the 10% level.

(iii) By regressing Δgfr on Δpe, $(\Delta pe_{-1} - \Delta pe)$. $(\Delta pe_{-2} - \Delta pe)$, *ww2*, and *pill*, we obtain the LRP and its standard error as the coefficient on Δpe: $-.075$, se = .032. So the estimated LRP is now negative and significant, which is very different from the equation in levels, (10.19) (the estimated LRP was .101 with a t statistic of about 3.37). This is a good example of how differencing variables before including them in a regression can lead to very different conclusions than a regression in levels.

11.14 (i) If $E(gc_t|I_{t-1}) = E(gc_t)$ – that is, $E(gc_t|I_{t-1})$ = does not depend on gc_{t-1}, then $\beta_1 = 0$ in $gc_t = \beta_0 + \beta_1 gc_{t-1} + u_t$. So the null hypothesis is H_0: $\beta_1 = 0$ and the alternative is H_1: $\beta_1 \neq 0$. Estimating the simple regression using the data in CONSUMP.RAW gives

$$\widehat{gc_t} = \underset{(.004)}{.011} \; \underset{(.156)}{+.446\,gc_{t-1}}$$

$$n = 35, \quad R^2 = .199.$$

The t statistic for $\hat{\beta}_1$ is about 2.86, and so we strongly reject the PIH. The coefficient on gc_{t-1} is also practically large, showing significant autocorrelation in consumption growth.

(ii) When gy_{t-1} and $i3_{t-1}$ are added to the regression, the R-squared becomes about .288. The F statistic for joint significance of gy_{t-1} and $i3_{t-1}$, obtained using the Stata "test" command, is 1.95, with p-value $\approx .16$. Therefore, gy_{t-1} and $i3_{t-1}$ are not jointly significant at even the 15% level.

11.16 (i) The first order autocorrelation for *prcfat* is .709, which is high but not necessarily a cause for concern. For *unem*, $\hat{\rho}_1 = .950$, which is cause for concern in using *unem* as an explanatory variable in a regression.

(ii) If we use the first differences of *prcfat* and *unem*, but leave all other variables in their original form, we get the following:

$$\widehat{\Delta prcfat} = \underset{(.105)}{-.127} + \ldots + \underset{(.0072)}{.0068\, wkends} + \underset{(.0161)}{.0125\, \Delta unem}$$

$$- \underset{(.0238)}{.0072\, spdlaw} + \underset{(.0265)}{.0008\, bltlaw}$$

$n = 107,\ R^2 = .344,$

where I have again suppressed the coefficients on the time trend and seasonal dummies. This regression basically shows that the change in *prcfat* cannot be explained by the change in *unem* or any of the policy variables. It does have some seasonality, which is why the *R*-squared is .344.

(iii) This is an example about how estimation in first differences loses the interesting implications of the model estimated in levels. Of course, this is not to say the levels regression is valid. But, as it turns out, we can reject a unit root in *prcfat*, and so we can at least justify using it in level form; see Computer Exercise 18.13. Generally, the issue of whether to take first differences is very difficult, even for professional time series econometricians.

Κεφάλαιο 12

12.1 We can reason this from equation (12.4) because the usual OLS standard error is an estimate of $\sigma / \sqrt{SST_x}$. When the dependent and independent variables are in level (or log) form, the AR(1) parameter, ρ, tends to be positive in time series regression models. Further, the independent variables tend to be positive correlated, so $(x_t - \bar{x})(x_{t+j} - \bar{x})$ – which is what generally appears in (12.4) when the $\{x_t\}$ do not have zero sample average – tends to be positive for most t and j. With multiple explanatory variables the formulas are more complicated but have similar features.

If $\rho < 0$, or if the $\{x_t\}$ is negatively autocorrelated, the second term in the last line of (12.4) could be negative, in which case the true standard deviation of $\hat{\beta}_1$ is actually less than $\sigma / \sqrt{SST_x}$.

12.3 (i) Because U.S. presidential elections occur only every four years, it seems reasonable to think the unobserved shocks – that is, elements in u_t – in one election have pretty much dissipated four years later. This would imply that $\{u_t\}$ is roughly serially uncorrelated.

(ii) The t statistic for H_0: $\rho = 0$ is $-.068/.240 \approx -.28$, which is very small. Further, the estimate $\hat{\rho} = -.068$ is small in a practical sense, too. There is no reason to worry about serial correlation in this example.

(iii) Because the test based on $t_{\hat{\rho}}$ is only justified asymptotically, we would generally be concerned about using the usual critical values with $n = 20$ in the original regression. But any kind of adjustment, either to obtain valid standard errors for OLS as in Section 12.5 or a feasible GLS procedure as in Section 12.3, relies on large sample sizes, too. (Remember, FGLS is not even unbiased, whereas OLS is under TS.1 through TS.3.) Most importantly, the estimate of ρ is *practically* small, too. With $\hat{\rho}$ so close to zero, FGLS or adjusting the standard errors would yield similar results to OLS with the usual standard errors.

12.5 (i) There is substantial serial correlation in the errors of the equation, and the OLS standard errors almost certainly underestimate the true standard deviation in $\hat{\beta}_{EZ}$. This makes the usual confidence interval for β_{EZ} and t statistics invalid.

(ii) We can use the method in Section 12.5 to obtain an approximately valid standard error. [See equation (12.43).] While we might use $g = 2$ in equation (12.42), with monthly data we might want to try a somewhat longer lag, maybe even up to $g = 12$.

12.7 Regressing $\hat{u}_t$ on $\hat{u}_{t-1}$, using the 69 available observations, gives $\hat{\rho} \approx .292$ and se($\hat{\rho}$) $\approx .118$. The *t* statistic is about 2.47, and so there is significant evidence of positive AR(1) serial correlation in the errors (even though the variables have been differenced). This means we should view the standard errors reported in equation (11.27) with some suspicion.

12.9 (i) The test for AR(1) serial correlation gives (with 35 observations) $\hat{\rho} \approx -.110$, se($\hat{\rho}$) $\approx .175$. The *t* statistic is well below one in absolute value, so there is no evidence of serial correlation in the accelerator model. If we view the test of serial correlation as a test of dynamic misspecification, it reveals no dynamic misspecification in the accelerator model.

(ii) It is worth emphasizing that, if there is little evidence of AR(1) serial correlation, there is no need to use feasible GLS (Cochrane-Orcutt or Prais-Winsten).

12.11 (i) Using the data only through 1992 gives

$$\widehat{demwins} = \underset{(.107)}{.441} - \underset{(.354)}{.473}\, partyWH + \underset{(.205)}{.479}\, incum + \underset{(.036)}{.059}\, partyWH \cdot gnews - \underset{(.028)}{.024}\, partyWH \cdot inf$$

$$n = 20, \quad R^2 = .437, \quad \bar{R}^2 = .287.$$

The largest *t* statistic is on *incum*, which is estimated to have a large effect on the probability of winning. But we must be careful here. *incum* is equal to 1 if a Democratic incumbent is running and –1 if a Republican incumbent is running. Similarly, *partyWH* is equal to 1 if a Democrat is currently in the White House and –1 if a Republican is currently in the White House. So, for an incumbent Democrat running, we must add the coefficients on *partyWH* and *incum* together, and this nets out to about zero.

The economic variables are less statistically significant than in equation (10.23). The *gnews* interaction has a *t* statistic of about 1.64, which is significant at the 10% level against a one-sided alternative. (Since the dependent variable is binary, this is a case where we must appeal to asymptotics. Unfortunately, we have only 20 observations.) The inflation variable has the expected sign but is not statistically significant.

(ii) There are two fitted values less than zero, and two fitted values greater than one.

(iii) Out of the 10 elections with *demwins* = 1, 8 of these are correctly predicted. Out of the 10 elections with *demwins* = 0, 7 are correctly predicted. So 15 out of 20 elections through 1992 are correctly predicted. (But, remember, we used data from these years to obtain the estimated equation.)

(iv) The explanatory variables are *partyWH* = 1, *incum* = 1, *gnews* = 3, and *inf* = 3.019. Therefore, for 1996,

$$\widehat{demwins} = .441 - .473 + .479 + .059(3) - .024(3.019) \approx .552.$$

Because this is above .5, we would have predicted that Clinton would win the 1996 election, as he did.

(v) The regression of $\hat{u}_t$ on $\hat{u}_{t-1}$ produces $\hat{\rho} \approx$ -.164 with heteroskedasticity-robust standard error of about .195. (Because the LPM contains heteroskedasticity, testing for AR(1) serial correlation in an LPM generally requires a heteroskedasticity-robust test.) Therefore, there is little evidence of serial correlation in the errors. (And, if anything, it is negative.)

(vi) The heteroskedasticity-robust standard errors are given in [·] below the usual standard errors:

$$\begin{aligned}\widehat{demwins} = &.441 - .473\, partyWH + .479\, incum + .059\, partyWH\cdot gnews \\ &(.107) \quad (.354) \quad (.205) \quad (.036) \\ &[.086] \quad [.301] \quad [.185] \quad [.030] \\ &-.024\, partyWH\cdot inf \\ &(.028) \\ &[.019]\end{aligned}$$

$$n = 20, \quad R^2 = .437, \quad \overline{R}^2 = .287.$$

In fact, all heteroskedasticity-robust standard errors are less than the usual OLS standard errors, making each variable more significant. For example, the *t* statistic on *partyWH*·*gnews* becomes about 1.97, which is notably above 1.64. But we must remember that the standard errors in the LPM have only asymptotic justification. With only 20 observations it is not clear we should prefer the heteroskedasticity-robust standard errors to the usual ones.

12.13 (i) The iterated Prais-Winsten estimates are given below. The estimate of ρ is, to three decimal places, .293, which is the same as the estimate used in the final iteration of Cochrane-Orcutt:

$$\begin{aligned}\widehat{\log(chnimp)} = &-37.08 + 2.94 \log(chempi) + 1.05 \log(gas) \\ &(22.78) \quad (.63) \quad (.98) \\ &+ 1.13 \log(rtwex) \\ &(.51) \\ &- .016\, befile6 - .033\, affile6 - .577\, afdec6 \\ &(.319) \quad (.322) \quad (.342)\end{aligned}$$

$$n = 131, \quad R^2 = .202$$

(ii) Not surprisingly, the C-O and P-W estimates are quite similar. To three decimal places, they use the same value of $\hat{\rho}$ (to four decimal places it is .2934 for C-O and .2932 for P-W). The only practical difference is that P-W uses the equation for $t = 1$. With $n = 131$, we hope this makes little difference.

12.15 (i) Here are the OLS regression results:

$$\widehat{\log(avgprc)} = -.073 - .0040\, t - .0101\, mon - .0088\, tues + .0376\, wed + .0906\, thurs$$
$$(.115) \quad (.0014) \quad (.1294) \quad (.1273) \quad (.1257) \quad (.1257)$$

$n = 97,\ R^2 = .086$

The test for joint significance of the day-of-the-week dummies is $F = .23$, which gives p-value = .92. So there is no evidence that the average price of fish varies systematically within a week.

(ii) The equation is

$$\widehat{\log(avgprc)} = -.920 - .0012\, t - .0182\, mon - .0085\, tues + .0500\, wed + .1225\, thurs$$
$$(.190) \quad (.0014) \quad (.1141) \quad (.1121) \quad (.1117) \quad (.1110)$$
$$+ .0909\, wave2 + .0474\, wave3$$
$$(.0218) \quad (.0208)$$

$n = 97,\ R^2 = .310$

Each of the wave variables is statistically significant, with *wave2* being the most important. Rough seas (as measured by high waves) would reduce the supply of fish (shift the supply curve back), and this would result in a price increase. One might argue that bad weather reduces the demand for fish at a market, too, but that would reduce price. If there are demand effects captured by the wave variables, they are being swamped by the supply effects.

(iii) The time trend coefficient becomes much smaller and statistically insignificant. We can use the omitted variable bias table from Chapter 3, Table 3.2 to determine what is probably going on. Without *wave2* and *wave3*, the coefficient on *t* seems to have a downward bias. Since we know the coefficients on *wave2* and *wave3* are positive, this means the wave variables are negatively correlated with *t*. In other words, the seas were rougher, on average, at the beginning of the sample period. (You can confirm this by regressing *wave2* on *t* and *wave3* on *t*.)

(iv) The time trend and daily dummies are clearly strictly exogenous, as they are just functions of time and the calendar. Further, the height of the waves is not influenced by past unexpected changes in log(*avgprc*).

(v) We simply regress the OLS residuals on one lag, getting $\hat{\rho} = .618, \text{se}(\hat{\rho}) = .081, t_{\hat{\rho}} = 7.63$. Therefore, there is strong evidence of positive serial correlation.

(vi) The Newey-West standard errors are $\text{se}(\hat{\beta}_{wave2}) = .0234$ and $\text{se}(\hat{\beta}_{wave3}) = .0195$. Given the significant amount of AR(1) serial correlation in part (v), it is somewhat surprising that these standard errors are not much larger compared with the usual, incorrect standard errors. In fact, the Newey-West standard error for $\hat{\beta}_{wave3}$ is actually smaller than the OLS standard error.

(vii) The Prais-Winsten estimates are

$$\widehat{\log(avgprc)} = -.658 - .0007\, t + .0099\, mon + .0025\, tues + .0624\, wed + .1174\, thurs$$
$$(.239) \quad (.0029) \quad (.0652) \quad (.0744) \quad (.0746) \quad (.0621)$$
$$+ .0497\, wave2 + .0323\, wave3$$
$$(.0174) \quad (.0174)$$

$n = 97,\ R^2 = .135$

The coefficient on *wave2* drops by a nontrivial amount, but it still has a *t* statistic of almost 3. The coefficient on *wave3* drops by a relatively smaller amount, but its *t* statistic (1.86) is borderline significant. The final estimate of ρ is about .687.

Κεφάλαιο 13

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 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) 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 $y74 \cdot educ, \ldots, y84 \cdot educ$ 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

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

Three of the interaction terms, $y78 \cdot educ$, $y82 \cdot educ$, and $y84 \cdot educ$ 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.

13.9 (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

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

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

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 $y81 \cdot \log(dist)$ 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.

13.11 (i) Using pooled OLS we obtain

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

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

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 thing 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

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

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

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 ≈ 4). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation.

13.13 (i) Pooling across semesters and using OLS gives

$$\widehat{trmgpa} = -1.75 \underset{(0.35)}{-.058\, spring} \underset{(.00015)}{+.00170\, sat} \underset{(.0010)}{-.0087\, hsperc}$$

$$\underset{(.052)}{+.350\, female} \underset{(.123)}{-.254\, black} \underset{(.117)}{-.023\, white} \underset{(.076)}{-.035\, frstsem}$$

$$\underset{(.00073)}{-.00034\, tothrs} \underset{(0.104)}{+1.048\, crsgpa} \underset{(.049)}{-.027\, season}$$

$$n = 732, \quad R^2 = .478, \quad \overline{R}^2 = .470.$$

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 form 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

$$\widehat{\Delta trmgpa} = -.237 \underset{(.206)}{+.019\, \Delta frstsem} \underset{(.069)}{+.012\, \Delta tothrs} \underset{(.014)}{+1.136\, \Delta crsgpa} \underset{(0.119)}{-.065\, season} \underset{(.043)}{}$$

$$n = 366, \quad R^2 = .208, \quad \overline{R}^2 = .199.$$

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.

13.15 (i) When we add the changes of the nine log wage variables to equation (13.33) we obtain

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

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

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 cannot all 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 ≈ 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 ≈ -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 $\approx .26$.

13.17. (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.

Κεφάλαιο 14

14.1 First, for each $t > 1$, $\text{Var}(\Delta u_{it}) = \text{Var}(u_{it} - u_{i,t-1}) = \text{Var}(u_{it}) + \text{Var}(u_{i,t-1}) = 2\sigma_u^2$, where we use the assumptions of no serial correlation in $\{u_t\}$ and constant variance. Next, we find the covariance between Δu_{it} and $\Delta u_{i,t+1}$. Because these each have a zero mean, the covariance is $\text{E}(\Delta u_{it} \cdot \Delta u_{i,t+1}) = \text{E}[(u_{it} - u_{i,t-1})(u_{i,t+1} - u_{it})] = \text{E}(u_{it}u_{i,t+1}) - \text{E}(u_{it}^2) - \text{E}(u_{i,t-1}u_{i,t+1}) + \text{E}(u_{i,t-1}u_{it}) = -\text{E}(u_{it}^2) = -\sigma_u^2$ because of the no serial correlation assumption. Because the variance is constant across t, by Problem 11.1, $\text{Corr}(\Delta u_{it}, \Delta u_{i,t+1}) = \text{Cov}(\Delta u_{it}, \Delta u_{i,t+1})/\text{Var}(\Delta u_{it}) = -\sigma_u^2/(2\sigma_u^2) = -.5$.

14.3 (i) $\text{E}(e_{it}) = \text{E}(v_{it} - \lambda\bar{v}_i) = \text{E}(v_{it}) - \lambda\text{E}(\bar{v}_i) = 0$ because $\text{E}(v_{it}) = 0$ for all t.

(ii) $\text{Var}(v_{it} - \lambda\bar{v}_i) = \text{Var}(v_{it}) + \lambda^2\text{Var}(\bar{v}_i) - 2\lambda \cdot \text{Cov}(v_{it}, \bar{v}_i) = \sigma_v^2 + \lambda^2\,\text{E}(\bar{v}_i^2) - 2\lambda \cdot \text{E}(v_{it}\bar{v}_i)$. Now, $\sigma_v^2 = \text{E}(v_{it}^2) = \sigma_a^2 + \sigma_u^2$ and $\text{E}(v_{it}\bar{v}_i) = T^{-1}\sum_{s=1}^{T} E(v_{it}v_{is}) = T^{-1}[\sigma_a^2 + \sigma_a^2 + \ldots + (\sigma_a^2 + \sigma_u^2) + \ldots + \sigma_a^2] = \sigma_a^2 + \sigma_u^2/T$. Therefore, $\text{E}(\bar{v}_i^2) = T^{-1}\sum_{t=1}^{T} E(v_{it}\bar{v}_i) = \sigma_a^2 + \sigma_u^2/T$. Now, we can collect terms:

$$\text{Var}(v_{it} - \lambda\bar{v}_i) = (\sigma_a^2 + \sigma_u^2) + \lambda^2(\sigma_a^2 + \sigma_u^2/T) - 2\lambda(\sigma_a^2 + \sigma_u^2/T).$$

Now, it is convenient to write $\lambda = 1 - \sqrt{\eta}/\sqrt{\gamma}$, where $\eta \equiv \sigma_u^2/T$ and $\gamma \equiv \sigma_a^2 + \sigma_u^2/T$. Then

$$\begin{aligned}
\text{Var}(v_{it} - \lambda\bar{v}_i) &= (\sigma_a^2 + \sigma_u^2) - 2\lambda(\sigma_a^2 + \sigma_u^2/T) + \lambda^2(\sigma_a^2 + \sigma_u^2/T) \\
&= (\sigma_a^2 + \sigma_u^2) - 2(1 - \sqrt{\eta}/\sqrt{\gamma})\gamma + (1 - \sqrt{\eta}/\sqrt{\gamma})^2\gamma \\
&= (\sigma_a^2 + \sigma_u^2) - 2\gamma + 2\sqrt{\eta}\cdot\sqrt{\gamma} + (1 - 2\sqrt{\eta}/\sqrt{\gamma} + \eta/\gamma)\gamma \\
&= (\sigma_a^2 + \sigma_u^2) - 2\gamma + 2\sqrt{\eta}\cdot\sqrt{\gamma} + (1 - 2\sqrt{\eta}/\sqrt{\gamma} + \eta/\gamma)\gamma \\
&= (\sigma_a^2 + \sigma_u^2) - 2\gamma + 2\sqrt{\eta}\cdot\sqrt{\gamma} + \gamma - 2\sqrt{\eta}\cdot\sqrt{\gamma} + \eta \\
&= (\sigma_a^2 + \sigma_u^2) + \eta - \gamma = \sigma_u^2.
\end{aligned}$$

This is what we wanted to show.

(iii) We must show that $\text{E}(e_{it}e_{is}) = 0$ for $t \neq s$. Now $\text{E}(e_{it}e_{is}) = \text{E}[(v_{it} - \lambda\bar{v}_i)(v_{is} - \lambda\bar{v}_i)] = \text{E}(v_{it}v_{is}) - \lambda\text{E}(\bar{v}_i v_{is}) - \lambda\text{E}(v_{it}\bar{v}_i) + \lambda^2\text{E}(\bar{v}_i^2) = \sigma_a^2 - 2\lambda(\sigma_a^2 + \sigma_u^2/T) + \lambda^2\text{E}(\bar{v}_i^2) = \sigma_a^2 - 2\lambda(\sigma_a^2 + \sigma_u^2/T) + \lambda^2(\sigma_a^2 + \sigma_u^2/T)$. The rest of the proof is very similar to part (ii):

$$
\begin{aligned}
\mathrm{E}(e_{it}e_{is}) &= \sigma_a^2 - 2\lambda(\sigma_a^2 + \sigma_u^2/T) + \lambda^2(\sigma_a^2 + \sigma_u^2/T) \\
&= \sigma_a^2 - 2(1 - \sqrt{\eta}/\sqrt{\gamma})\gamma + (1 - \sqrt{\eta}/\sqrt{\gamma})^2\gamma \\
&= \sigma_a^2 - 2\gamma + 2\sqrt{\eta}\cdot\sqrt{\gamma} + (1 - 2\sqrt{\eta}/\sqrt{\gamma} + \eta/\gamma)\gamma \\
&= \sigma_a^2 - 2\gamma + 2\sqrt{\eta}\cdot\sqrt{\gamma} + (1 - 2\sqrt{\eta}/\sqrt{\gamma} + \eta/\gamma)\gamma \\
&= \sigma_a^2 - 2\gamma + 2\sqrt{\eta}\cdot\sqrt{\gamma} + \gamma - 2\sqrt{\eta}\cdot\sqrt{\gamma} + \eta \\
&= \sigma_a^2 + \eta - \gamma = 0.
\end{aligned}
$$

14.5 (i) For each student we have several measures of performance, typically three or four, the number of classes taken by a student that have final exams. When we specify an equation for each standardized final exam score, the errors in the different equations for the same student are certain to be correlated: students who have more (unobserved) ability tend to do better on all tests.

(ii) An unobserved effects model is

$$score_{sc} = \theta_c + \beta_1 atndrte_{sc} + \beta_2 major_{sc} + \beta_3 SAT_s + \beta_4 cumGPA_s + a_s + u_{sc},$$

where a_s is the unobserved student effect. Because SAT score and cumulative GPA depend only on the student, and not on the particular class he/she is taking, these do not have a *c* subscript. The attendance rates do generally vary across class, as does the indicator for whether a class is in the student's major. The term θ_c denotes different intercepts for different classes. Unlike with a panel data set, where time is the natural ordering of the data within each cross-sectional unit, and the aggregate time effects apply to all units, intercepts for the different classes may not be needed. If all students took the same set of classes then this is similar to a panel data set, and we would want to put in different class intercepts. But with students taking different courses, the class we label as "1" for student A need have nothing to do with class "1" for student B. Thus, the different class intercepts based on arbitrarily ordering the classes for each student probably are not needed. We can replace θ_c with β_0, an intercept constant across classes.

(iii) Maintaining the assumption that the idiosyncratic error, u_{sc}, is uncorrelated with all explanatory variables, we need the unobserved student heterogeneity, a_s, to be uncorrelated with $atndrte_{sc}$. The inclusion of SAT score and cumulative GPA should help in this regard, as a_s, is the part of ability that is not captured by SAT_s and $cumGPA_s$. In other words, controlling for SAT_s and $cumGPA_s$ could be enough to obtain the ceteris paribus effect of class attendance.

(iv) If SAT_s and $cumGPA_s$ are not sufficient controls for student ability and motivation, a_s is correlated with $atndrte_{sc}$, and this would cause pooled OLS to be biased and inconsistent. We could use fixed effects instead. Within each student we compute the demeaned data, where, for each student, the means are computed across classes. The variables SAT_s and $cumGPA_s$ drop out of the analysis.

14.6 (i) This is done in Computer Exercise 13.5(i).

(ii) See Computer Exercise 13.5(ii).

(iii) See Computer Exercise 13.5(iii).

(iv) This is the only new part. The fixed effects estimates, reported in equation form, are

$$\widehat{\log(rent_{it})} = .386\, y90_t + .072 \log(pop_{it}) + .310 \log(avginc_{it}) + .0112\, pctstu_{it},$$
$$\qquad\qquad (.037) \qquad\quad (.088) \qquad\quad (.066) \qquad\quad (.0041)$$

$N = 64, \quad T = 2.$

(There are $N = 64$ cities and $T = 2$ years.) We do not report an intercept because it gets removed by the time demeaning. The coefficient on $y90_t$ is identical to the intercept from the first difference estimation, and the slope coefficients and standard errors are identical to first differencing. We do not report an R-squared because none is comparable to the R-squared obtained from first differencing.

14.8 (i) 135 firms are used in the FE estimation. Because there are three years, we would have a total of 405 observations if each firm had data on all variables for all three years. Instead, due to missing data, we can use only 390 observations in the FE estimation. The fixed effects estimates are

$$\widehat{hrsemp_{it}} = -1.10\, d88_t + 4.09\, d89_t + 34.23\, grant_{it}$$
$$\qquad\qquad\quad (1.98) \qquad\quad (2.48) \qquad\quad (2.86)$$

$$+.504\, grant_{i,t-1} - .176 \log(employ_{it})$$
$$(4.127) \qquad\quad (4.288)$$

$n = 390, \quad N = 135, \quad T = 3.$

(ii) The coefficient on *grant* means that if a firm received a grant for the current year, it trained each worker an average of 34.2 hours more than it would have otherwise. This is a practically large effect, and the t statistic is very large.

(iii) Since a grant last year was used to pay for training last year, it is perhaps not surprising that the grants does not carry over into more training this year. It would if inertia played a role in training workers.

(iv) The coefficient on the employees variable is very small: a 10% increase in *employ* increases predicted hours per employee by only about .018. [Recall: $\Delta\widehat{hrsemp} \approx (.176/100)\,(\%\Delta employ)$.] This is very small, and the t statistic is practically zero.

14.10 (i) Different occupations are unionized at different rates, and wages also differ by occupation. Therefore, if we omit binary indicators for occupation, the union wage differential may simply be picking up wage differences across occupations. Because

some people change occupation over the period, we should include these in our analysis.

(ii) Because the nine occupational categories (*occ1* through *occ9*) are exhaustive, we must choose one as the base group. Of course the group we choose does not affect the estimated union wage differential. The fixed effect estimate on *union*, to four decimal places, is .0804 with standard error = .0194. There is practically no difference between this estimate and standard error and the estimate and standard error without the occupational controls ($\hat{\beta}_{union}$ = .0800, se = .0193).

14.12 (i) If there is a deterrent effect then $\beta_1 < 0$. The sign of β_2 is not entirely obvious, although one possibility is that a better economy means less crime in general, including violent crime (such as drug dealing) that would lead to fewer murders. This would imply $\beta_2 > 0$.

(ii) The pooled OLS estimates using 1990 and 1993 are

$$\widehat{mrdrte}_{it} = \underset{}{-5.28} - \underset{(4.43)}{2.07\, d93_t} + \underset{(2.14)}{.128\, exec_{it}} + \underset{(.263)}{2.53\, unem_{it}} \quad \underset{(0.78)}{}$$

$$N = 51,\ T = 2,\ R^2 = .102$$

There is no evidence of a deterrent effect, as the coefficient on *exec* is actually positive (though not statistically significant).

(iii) The first-differenced equation is

$$\widehat{\Delta mrdrte}_i = \underset{}{.413} - \underset{(.209)}{.104\, \Delta exec_i} - \underset{(.043)\quad(.159)}{.067\, \Delta une}$$

$$n = 51,\ R^2 = .110$$

Now, there is a statistically significant deterrent effect: 10 more executions is estimated to reduce the murder rate by 1.04, or one murder per 100,000 people. Is this a large effect? Executions are relatively rare in most states, but murder rates are relatively low on average, too. In 1993, the average murder rate was about 8.7; a reduction of one would be nontrivial. For the (unknown) people whose lives might be saved via a deterrent effect, it would seem important.

(iv) The heteroskedasticity-robust standard error for $\Delta exec_i$ is .017. Somewhat surprisingly, this is well below the nonrobust standard error. If we use the robust standard error, the statistical evidence for the deterrent effect is quite strong ($t \approx -6.1$). See also Computer Exercise 13.12.

(v) Texas had by far the largest value of *exec*, 34. The next highest state was Virginia, with 11. These are three-year totals.

(vi) Without Texas in the estimation, we get the following, with heteroskedasticity-robust standard errors in [·]:

$$\widehat{\Delta mrdrte}_i = \underset{\substack{(.211)\\[.200]}}{.413} - \underset{\substack{(.105)\\[.079]}}{.067}\,\Delta exec_i - \underset{\substack{(.160)\\[.146]}}{.070}\,\Delta une$$

$n = 50,\ R^2 = .013$

Now the estimated deterrent effect is smaller. Perhaps more importantly, the standard error on $\Delta exec_i$ has increased by a substantial amount. This happens because when we drop Texas, we lose much of the variation in the key explanatory variable, $\Delta exec_i$.

(vii) When we apply fixed effects using all three years of data and all states we get

$$\widehat{mrdrte}_{it} = \underset{(.75)}{1.73}\,d90_t + \underset{(.71)}{1.70}\,d93_t - \underset{(.160)}{.054}\,exec_{it} + \underset{(.285)}{.395}\,unem_{it}$$

$N = 51,\ T = 3,\ R^2 = .068$

The size of the deterrent effect is only about half as big as when 1987 is not used. Plus, the t statistic, about $-.34$, is very small. The earlier finding of a deterrent effect is not robust to the time period used. Oddly, adding another year of data causes the standard error on the *exec* coefficient to markedly increase.

14.14 (i) The OLS estimates are

$$\begin{aligned}\widehat{pctstck} = {} & 128.54 + \underset{(55.17)}{11.74}\,choice + \underset{(6.23)}{14.34}\,prftshr + \underset{(7.23)}{1.45}\,female - \underset{(6.77)}{1.50}\,age \\ & \underset{(.78)}{} \\ & + .70\,educ - \underset{(16.02)}{15.29}\,finc25 + \underset{(15.72)}{.19}\,finc35 - \underset{(17.80)}{3.86}\,finc50 \\ & - \underset{(16.02)}{13.75}\,finc75 - \underset{(15.72)}{2.69}\,finc100 - \underset{(17.80)}{25.05}\,finc101 - \underset{(.0128)}{.0026}\,wealth89 \\ & + \underset{(6.68)}{6.67}\,stckin89 - \underset{(6.38)}{7.50}\,irain89\end{aligned}$$

$n = 194,\ R^2 = .108$

Investment choice is associated with about 11.7 percentage points more in stocks. The t statistic is 1.88, and so it is marginal significant.

(ii) These variables are not very important. The F test for joint significant is 1.03. With 9 and 179 *df*, this gives p-value = .42. Plus, when these variables are dropped from the regression, the coefficient on *choice* only falls to 11.15.

(iii) There are 171 different families in the sample.

(iv) I will only report the cluster-robust standard error for *choice*: 6.20. Therefore, it is essentially the same as the usual OLS standard error. This is not very surprising because at least 171 of the 194 observations can be assumed independent of one another. The explanatory variables may adequately capture the within-family correlation.

(v) There are only 23 families with spouses in the data set. Differencing within these families gives

$$\Delta\widehat{pctstck} = \underset{(10.94)}{15.93} + \underset{(15.00)}{2.28\,\Delta choice} - \underset{(16.92)}{9.27\,\Delta prftshr} + \underset{(21.49)}{21.55\,\Delta female} - \underset{(9.00)}{3.57\,\Delta age} - \underset{(3.43)}{1.22\,\Delta educ}$$

$$n = 23,\ R^2 = .206,\ \bar{R}^2 = -.028$$

All of the income and wealth variables, and the stock and IRA indicators, drop out, as these are defined at the family level (and therefore are the same for the husband and wife).

(vi) None of the explanatory variables is significant in part (v), and this is not too surprising. We have only 23 observations, and we are removing much of the variation in the explanatory variables (except the gender variable) by using within-family differences.