Solutions to Problem Set 5 (Due April 27)

EC 2228 03, Spring 2015 Prof. Baum, Mr. Zhang

Maximum number of points for Problem set 5 is: 90

Problem 9.3

(i) (2 pts) 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) (2 pts) 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, $_3$¡ 0. From Table 3.2, omitting *lnchprg* (the proxy for *poverty*) from the regression produces an upward biased estimator of $_1$ [ignoring the presence of log(*enroll*) in the model]. So when we control for the poverty rate, the effect of spending falls.

(iii) (2 pts) 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 $\Delta ma\hat{t}h10(1.26/100)(\%\ Deltaenroll) = .0126(\%\Delta enroll)$. Therefore, a 10% increase in enrollment leads to a drop in $math10$ of .126 percentage points.

(iv) (2 pts) 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) (2 pts) 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.

Problem 9.C3

(i) (2 pt) 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) (2 pts) The simple regression estimates using the 1988 data are

$$\begin{array}{llll} \log(scrap) = & .409 & +.057 & grant \\ & (.241) & (.406) & \end{array}$$

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

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

(iii) (3 pts) When we add $\log(scrap_{87})$ to the equation, we obtain

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

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

where the year subscripts are for clarity. The t statistic for $H_0 : \beta_{grant} = 0$ is $-.254/.147 = -1.73$. 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) (2 pts) The t statistic is $(.831 - 1)/.044 = -3.84$, which is a strong rejection of H_0.

(v) (2 pts) With the heteroskedasticity-robust standard error, the t statistic for $grant_{88}$ is $-.254/.146 = -1.74$, 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 $(.831-1)/.074 = -2.28$, which is notably smaller than before, but it is still pretty significant.

Problem 9.C8

(i) (2 pt) The mean of *stotal* is .047, its standard deviation is .854, the minimum value is -3.32, and the maximum value is 2.24.

(ii) (2 pts) In the regression of *jc* on *stotal*, the slope coefficient is .011 (se = .011). Therefore, while the estimated relationship is positive, the t statistic is only one: the correlation between *jc* and *stotal* is weak at best. In the regression of *univ* on *stotal*, the slope coefficient is 1.170 (se = .029), for a t statistic of 39.7. Therefore, *univ* and *stotal* are positively correlated.

(iii) (3 pts) When we add *stotal* to (4.17) and estimate the resulting equation by OLS, we get

$$\begin{array}{lccccc} \log(wage) = & 1.495 & +.0631 \; jc & +.0686 \; univ & +.00488 \; exper & +.0494 \; stotal \\ & (.021) & (.0068) & (.0026) & (.00016) & (.0068) \end{array}$$

$$n = 6,763, R^2 = .228.$$

Let $\theta = \beta_{jc} - \beta_{univ}$. Then, we can test $\beta_{jc} = \beta_{univ}$ by testing $H_0 : \theta = 0$ and $H_1 : \theta < 0$. In Stata, *lincom* $jc - univ$ offers you t statistic of -.81. Hence, at the 95% level, we cannot reject H_0. Compared with what we found without *stotal*, the evidence is even weaker against H_1. The t statistic from equation (4.27) is about -1.48, while here we have obtained only -.81.

(iv) (1 pt) When $stotal2$ is added to the equation, its coefficient is .0019 (t statistic = .40). Therefore, there is no reason to add the quadratic term.

(v) (1 pt) The F statistic for the significance of the interaction terms $stotal \cdot jc$ and $stotal \cdot univ$ is about 1.96; with 2 and 6,756, this gives p-value = .141. So, even at the 10% level, the interaction terms are jointly insignificant. It is probably not worth complicating the basic model estimated in part (iii).

(vi) (1 pt) I would just use the model from part (iii), where *stotal* appears only in level form. The other embellishments were not statistically significant at small enough significance levels to warrant the additional complications.

Problem 9.C11

(i) (2 pts) The coefficient and t statistic of *exec* is respectively .085 and .30. As the coefficient of *exec* is not statistically significant, this regression doesn't give any evidence for a deterrent effect of capital punishment.

(ii) (2 pts) 34 executions were reported for Texas during 1993. It is much larger than any other states. The second largest number of executions is 11 in Virginia. In addition, there was no excecution in more than 30 states. However, after adding dummy for Texas, its coefficient isn't statistically significant (t statistic = -.32). From this, Texas doesn't appear to be an "outlier".

(iii) (2 pts) After employing the lagged murder rate, the coefficient decreases to negative (.295 $\Longrightarrow$ -.071). But t statistics in absolute value increases (.41 $\Longrightarrow$ 2.34) so that it becomes statistically significant.

(iv) (2 pts) Dropping Texas, both β_{exec} (-.071 $\Longrightarrow$ -.045) and its t statistic (-2.34 $\Longrightarrow$ -.60) decrease in absolute value so that it is not statistically different from zero. From this, Texas is not an outlier.

Problem 10.2

(4 pts) We follow the hint and write

$$gGDP_{t-1} = \alpha_0 + \delta_0 int_{t-1} + \delta_1 int_{t-2} + u_{t-1},$$

and plug this into the right-hand-side of the int_t equation:

$$\begin{aligned} int_t &= \gamma_0 + \gamma_1(\alpha_0 + \delta_0 int_{t-1} + \delta_1 int_{t-2} + u_{t-1} - 3) + v_t \\ &= (\gamma_0 + \gamma_1\alpha_0 - 3\gamma_1) + \gamma_1\delta_0 int_{t-1} + \gamma_1\delta_1 int_{t-2} + \gamma_1 u_{t-1} + v_t. \end{aligned}$$

Now by assumption, u_{t-1} has zero mean and is uncorrelated with all right-hand-side variables in the previous equation, except itself of course. So

$$\mathrm{Cov}(int, u_{t-1}) = \mathrm{E}(int_t u_{t-1}) = \gamma_1 \mathrm{E}(u_{t-1}^2) > 0$$

because $\gamma_1 > 0$. If $\sigma_u^2 = \mathrm{E}(u_t^2)$ for all t then $\mathrm{Cov}(int, u_{t-1}) = \gamma_1\sigma_u^2$. This violates the strict exogeneity assumption, TS.2. While u_t is uncorrelated with int_t, int_{t-1}, and so on, u_t is correlated with int_{t+1}.

Problem 10.C1

(4 pts.) Define $post79$ as a dummy variable equal to one for years after 1979, and zero otherwise. Adding it to equation 10.15 gives

$$\begin{array}{rlllllll} \widehat{i3_t} = & 1.30 & +0.608 & inf_t & +0.363 & def_t & +1.56 & post79_t \\ & (0.43) & (0.076) & & (0.120) & & (0.51) & \end{array}$$

$$n = 56, R^2 = 0.664, \bar{R}^2 = 0.644.$$

The coefficient on $post79$ is statistically significant (t-statistic$\approx$ 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.

Problem 10.C7

(i) (2 pts.) The estimated equation is

$$\begin{array}{rclll} \widehat{gc_t} & = & 0.0081 & + & 0.571 \quad gy_t \\ & & (0.0019) & & (0.067) \end{array}$$

$$n = 36, R^2 = 0.679.$$

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.

(ii) (2 pts.) Adding gy_{t-1} to the equation gives

$$\begin{array}{rlllll} \widehat{gc_t} = & 0.0064 & +0.552 & gy_t & +0.096 & gy_{t-1} \\ & (0.0023) & (0.070) & & (0.069) & \end{array}$$

$$n = 35, R^2 = 0.695.$$

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) (2 pts.) If we add $r3_t$ to the model in part (i) we have

$$\begin{array}{rlllll} \widehat{gc_t} = & 0.0082 & +0.578 & gy_t & -0.00021 & r3_t \\ & (0.0020) & (0.072) & & (0.00063) & \end{array}$$

$$n = 36, R^2 = 0.680.$$

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.

Problem 10.C9

(i) (2 pts.) The sigh of β_2 should be negative: as interest rates rise, stock returns fall. Higher interest rates imply that T-bill or bond investments are more attractive, and also signal a future slowdown in the economic actively. The sigh 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) (2 pts.) The estimated equation is

$$\begin{array}{llllll} \widehat{rsp500}_t = & 18.84 & +0.036 & prip_t & -1.36 & i3_t \\ & (3.27) & (0.129) & & (0.054) & \end{array}$$

$$n = 557, R^2 = 0.012.$$

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) (2 pts.) Only $i3$ is statistically significant with t statistic $\approx$-2.52.

(iv) (2 pts.) 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 $i3t$ before we know $rsp500t$. What the regression in part (i) says is that a change in $i3$ is associated with a contemporaneous change in $rsp500$.

Problem 10.C11

(i) (2 pts) 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) (3 pts) The OLS regression gives

$$\begin{aligned} \widehat{\log(totacc)} = {} & \underset{(.019)}{10.469} + \underset{(.00016)}{.00275}\, t - \underset{(.0244)}{.0427}\, feb + \underset{(.0244)}{.0798}\, mar + \underset{(.0245)}{.0185}\, apr \\ & + \underset{(.0245)}{.0321}\, may + \underset{(.0245)}{.0202}\, jun + \underset{(.0245)}{.0376}\, jul + \underset{(.0245)}{.0540}\, aug \\ & + \underset{(.0245)}{.0424}\, sep + \underset{(.0245)}{.0821}\, oct + \underset{(.0245)}{.0713}\, nov + \underset{(.0245)}{.0962}\, dec \end{aligned}$$

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

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) (3 pts)

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

$$- \underset{(.0126)}{.0538}\, spdlaw + \underset{(.0142)}{.0954}\, beltlaw$$

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

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) (2 pts) 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) (2 pts) 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) (3 pts)

$$\widehat{prcfat} = \underset{(.103)}{1.030} + \ldots + \underset{(.00616)}{.00063}\, wkends - \underset{(.0055)}{.0154}\, unem$$

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

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

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.

Problem 12.1

(2 pts) 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 $\sigma/\sqrt{SST_x}$.

Problem 12.C9

(i) (2 pts) Here are the OLS regression results:

$$\begin{array}{lllllllllll}
\log(avgprc) = & -.073 & -.004 & t & -.010 & mon & -.009 & tues & +.038 & wed & +.091 & thurs \\
 & (.115) & (.001) & & (.129) & & (.127) & & (.126) & & (.126) &
\end{array}$$

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

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) (2 pts) The equation is

$$\begin{array}{lllllllllll}
\log(avgprc) = & -.920 & -.001 & t & -.018 & mon & -.009 & tues & +.050 & wed & +.123 & thurs \\
 & (.190) & (.001) & & (.114) & & (.112) & & (.112) & & (.111) & \\
 & +.091 & wave2 & +.047 & wave3 & & & & & & & \\
 & (.022) & & (.021) & & & & & & & &
\end{array}$$

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

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) (2 pts) The time trend coefficient becomes much smaller and statistically insignificant. We can use the omitted variable bias table from Chapter 3, Table 3.2 (page 92) 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)(2 pts) 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) (2 pts) We simply regress the OLS residuals on one lag, getting $\rho = .618$, $\text{se}(\rho) = .081$, $t_\rho = 7.62$. Therefore, there is strong evidence of positive serial correlation.

(vi) (1 pt) The Newey-West standard errors are $\text{se}(\beta_{wave2}) = .023$ and $\text{se}(\beta_{wave3}) = .019$. 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 β_{wave3} is actually smaller than the OLS standard error.

(vii) (1 pt) The Prais-Winsten estimates are

$$\begin{array}{lllllllll}
\log(avgprc) = & -.658 & -.001\ t & -.010\ mon & -.003\ tues & +.062\ wed & +.117\ thurs \\
 & (.239) & (.003) & (.065) & (.074) & (.075) & (.062) \\
 & +.050\ wave2 & +.032\ wave3 & & & & \\
 & (.017) & (.017) & & & &
\end{array}$$

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

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.