The Islamic University of Gaza
Faculty of Commerce
Economics Department
Econometrics & Quantitative Analysis – Dr. Samir Safi
8/12/2012

Midterm Examination #2 - SOLUTION

Question #1

Do problem #9 in chapter 8 on page 275-276

9. A full-scale regression model for the total annual gross sales in thousands of dollars of J. C. Quarter's durable goods for the last 26 years produces the following result (all measurements are in real dollars—or billions of real dollars). Standard errors are in parentheses:

$$\widehat{SQ}_t = -7.2 + \underset{(250.1)}{200.3PC_t} - \underset{(125.6)}{150.6PQ_t} + \underset{(40.1)}{20.6Y_t} - \underset{(10.6)}{15.8C_t} + \underset{(103.8)}{201.1N_t}$$

where:
- SQ_t = sales of durable goods at J. C. Quarter's in year t
- PC_t = average price of durables in year t at J. C. Quarter's main competition
- PQ_t = the average price of durables at J. C. Quarter's in year t
- Y_t = U.S. gross domestic product in year t
- C_t = U.S. aggregate consumption in year t
- N_t = the number of J. C. Quarter's stores open in year t

a. Hypothesize signs, calculate t-scores, and test hypotheses for this result (5 percent level).
b. What problems (out of omitted variables, irrelevant variables, and multicollinearity) appear to exist in this equation? Explain.
c. Suppose you were now told that the $\bar{R}^2$ was .821, that $r_{Y,C}$ was .993, and that $r_{PC,PQ}$ was .813. Would this change your answer to the above question? How?
d. What recommendation would you make for a rerun of this equation with different explanatory variables? Why?

Solution #1

(a) Coefficient: β β β β β

Hypothesized sign: + – + + +

t-value: 0.801 –1.199 0.514 –1.491 1.937

$t_C = 1.725$ at the 5% level, so only β is significantly different from zero in the expected direction.

(b) The obviously low *t*-scores could be caused by irrelevant variables, by omitted variables biasing the estimated coefficients toward zero, or by severe imperfect multicollinearity.

(c) The high simple correlation coefficient between Y and C indicates that the two are virtually identical (redundant), which makes sense theoretically. The simple correlation coefficient between the two price variables is not as high, but mild multicollinearity exists nonetheless.

(d) Y_t and C_t both serve as measures of the aggregate buying power of the economy, so they are redundant, and one should be dropped. It doesn't matter statistically which one is dropped, but Y_t seems analytically more valid than C_t, so we'd drop C. Dropping one of the price variables would be a mistake, since they have opposite expected signs. While forming a relative price variable is an option, the low level of multicollinearity, the reasonable coefficients, and the possibility that C_t is also multicollinear with prices (so dropping it will improve things) all argue for making just one change.

Question #2

Do problem #13 in chapter 9 on page 340-341

13. You're hired by Farmer Vin, a famous producer of bacon and ham, to test the possibility that feeding pigs at night allows them to grow faster than feeding them during the day. You take 200 pigs (from newborn piglets to extremely old porkers) and randomly assign them to feeding only during the day or feeding only at night and, after six months, end up with the following (admittedly very hypothetical) equation:

$$\hat{W}_i = 12 + 3.5G_i + 7.0D_i - 0.25F_i$$
$$(1.0) \quad (1.0) \quad (0.10)$$
$$t = 3.5 \quad 7.0 \quad -2.5$$
$$\bar{R}^2 = .70 \quad n = 200 \quad DW = 0.50$$

where: W_i = the percentage weight gain of the *i*th pig
G_i = a dummy variable equal to 1 if the *i*th pig is a male, 0 otherwise
D_i = a dummy variable equal to 1 if the *i*th pig was fed only at night, 0 if only during the day
F_i = the amount of food (pounds) eaten per day by the *i*th pig

a. Test for serial correlation at the 5 percent level in this equation.
b. What econometric problems appear to exist in this equation? (*Hint:* Be sure to make and test appropriate hypotheses about the slope coefficients.)
c. The goal of your experiment is to determine whether feeding at night represents a significant improvement over feeding during the day. What can you conclude?
d. The observations are ordered from the youngest pig to the oldest pig. Does this information change any of your answers to the previous parts of this question? Is this ordering a mistake? Explain your answer.

Solution #2

(a) This is a cross-sectional dataset and we normally wouldn't expect autocorrelation, but we'll test anyway since that's what the question calls for. DL for a 5% one-sided, K = 3, test is approximately 1.61, substantially higher than the DW of 0.50. (Sample sizes in Table B-4 only go up to 100, but the critical values at those sample sizes turn out to be reasonable estimates of those at 200.) As a result, we can reject the null hypothesis of no positive serial correlation, which in this case seems like evidence of impure serial correlation caused by an omitted variable or an incorrect functional form.

(b)

Coefficient:	β	β	β
Hypothesized sign:	+	+	–?
t-value:	3.5	7.0	–2.5
$t_C = 1.645$ (5% one-sided with infinite d.f.)	reject	reject	reject

We certainly have impure serial correlation. In addition, some students will conclude that F has a coefficient that is significant in the unexpected direction. (As it turns out, the negative coefficient could have been anticipated because the dependent variable is in percentage terms but F is in aggregate terms. We'd guess that the more food a pig eats, the bigger it is, meaning that its chances of growing at a high *rate* are low, thus the negative sign.)

(c) The coefficient of D is significant in the expected direction, but given the problems with this equation, we'd be hesitant to conclude much of anything just yet.

(d) In this case, the accidental ordering was a lucky stroke (not a mistake), because it allowed us to realize that younger pigs will gain weight at a higher rate than their older counterparts. If the data are ordered by age, positive residuals will be clustered at one end of the dataset, while negative ones will be clustered at the other end, giving the appearance of serial correlation.

Question #3

Do problem #9 in chapter 10 on page 379-380

9. What makes a college library great? While quality of holdings and ease of use are important, the simplest measure of the quality of a library is the number of books it holds. Suppose you've been hired by the Annual American Research on the Number of Books project (AARON) to build a model of the number of books in a cross section of 60 U.S. university and college libraries. After researching the literature, you estimate (standard errors in parentheses):

$$\widehat{VOL_i} = -1842 + 0.038STU_i + 1.73FAC_i + 1.83SAT_i \qquad (10.32)$$
$$\qquad\qquad (0.026) \quad (0.44) \quad (0.82)$$
$$t = 1.45 \quad 3.91 \quad 2.23$$
$$\bar{R}^2 = .81 \quad n = 60$$

where: VOL_i = thousands of books in the *i*th school's library
STU_i = the number of students in the *i*th school
FAC_i = the number of faculty in the *i*th school
SAT_i = the average SATs of students in the *i*th school

a. The simple correlation coefficient between STU and FAC is 0.93, a Park test on the residuals of Equation 10.32 produces a t-score of 3.50, and the Durbin–Watson d for the equation is 1.91. Given this information, what econometric problems appear to exist in this

equation? Explain. Which problem do you think you should attempt to correct first? Why?

b. You decide to create a linear combination of students and faculty ($TOT_i = 10FAC_i + STU_i$) and to rerun Equation 10.32, obtaining Equation 10.33 below. Which equation do you prefer? Explain your reasoning.

$$\widehat{VOL}_i = -1704 + 0.087TOT_i + 1.69SAT_i \qquad (10.33)$$
$$(0.007) \quad (0.84)$$
$$t = 12.87 \quad 2.02$$
$$\bar{R}^2 = .80 \quad n = 60 \quad DW = 1.85$$

c. Use the data in Table 10.2 (filename BOOKS10) to test for heteroskedasticity in the residuals of Equation 10.33 using the Park test and/or the White test (depending on the tests typically used in your class).

d. If you run Weighted Least Squares (WLS) (using TOT as your proportionality factor) on Equation 10.33, you get Equation 10.34, but take a look at it! Why don't we provide a t-score for the coefficient of the inverse of TOT? Why didn't WLS work very well? What alternative remedy would you suggest?

$$\widehat{VOL/TOT}_i = 0.089 - 63.6\,(1/TOT_t) + 0.079SAT_i/TOT_i$$
$$(0.010) \quad (0.057)$$
$$t = 8.57 \quad 1.39$$
$$\bar{R}^2 = .02 \quad n = 60 \quad DW = 1.89 \qquad (10.34)$$

e. Start from the equation you prefer (between 10.32 and 10.33) and use the data in Table 10.2 to correct for heteroskedasticity using some method other than WLS (*Hint:* You might reformulate the equation or calculate HC $SE(\hat{\beta})$s.)

TABLE 10.2 DATA ON COLEGE AND UNIVERSITY LIBRARY HOLDINGS

Observation	VOL	FAC	STU	SAT
1	11.5	5	58	850
2	200.0	138	2454	954
3	70.0	44	573	874
4	100.0	98	2172	941
5	7000.0	1651	31123	1185
6	70.0	26	295	874
7	125.0	35	1131	902
8	2200.0	1278	22571	1048
9	400.0	365	6554	960
10	110.0	47	793	930
11	6000.0	1650	36330	1142
12	58.4	39	522	800
13	212.0	69	1041	1060
14	400.0	57	1059	1000
15	1888.0	896	16411	1150
16	486.0	125	1678	1170
17	439.0	135	2529	1100
18	1900.0	653	19082	1080
19	155.0	114	3523	1026

TABLE 10.2 (continued)

Observation	VOL	FAC	STU	SAT
20	6.9	11	207	873
21	509.0	346	6781	1097
22	180.0	25	147	990
23	53.0	21	214	920
24	100.0	44	764	900
25	30.0	18	176	1176
26	157.0	99	3682	930
27	475.0	223	5665	1037
28	613.6	384	4411	960
29	483.6	141	3341	860
30	2500.0	2021	41528	1086
31	142.0	66	1251	1030
32	210.0	73	1036	1000
33	20.0	10	120	1070
34	150.0	94	2344	858
35	300.0	195	2400	1180
36	233.5	70	1416	910
37	235.0	165	4148	1001
38	460.7	316	9738	980
39	1632.0	355	5578	1060
40	93.0	30	505	930
41	263.0	185	3724	1124
42	144.5	101	2387	945
43	770.0	148	1900	1190
44	1700.0	960	16750	1057
45	1100.0	284	2833	1310
46	1900.0	905	15762	1090
47	60.0	55	875	848
48	1200.0	445	6603	1060
49	1600.0	623	14727	1120
50	1289.0	412	11179	1230
51	1666.0	1607	9251	883
52	15.0	26	608	800
53	160.0	48	656	1010
54	200.0	281	3892	980
55	263.0	195	2987	1070
56	487.0	275	5148	1060
57	3300.0	867	11240	1260
58	145.0	37	569	843
59	205.0	28	628	980
60	7377.0	2606	34055	1160

Solution #3

(a) Multicollinearity and heteroskedasticity (but not positive serial correlation) appear to exist. We'd tackle the multicollinearity first. Since the heteroskedasticity could be impure, you should get the best specification you can before worrying about correcting for heteroskedasticity.

(b) For all intents and purposes, the two equations are identical. Given that, and given the reasonably strong *t*-score of STU, we'd stick with Equation 10.22. Note that the ratio of the FAC/STU coefficients is far more than 10/1 in Equation 10.22. This means that Equation 10.22 overemphasizes the importance of faculty compared to Equation 10.23. (On second thought, what's wrong with overemphasizing the importance of faculty?)

(c) Both methods show evidence of heteroskedasticity. For instance, if TOT = Z, the Park test

t = 4.85 > 2.67, the critical *t*-value for a two-sided, 1% test with 57 degrees of freedom (interpolating).

(d) There are many possible answers to this question, including HC standard errors, but the interesting possibility might be to reformulate the equation, using SAT and STU/FAC (the student/faculty ratio) as proxies for quality:

$$\widehat{\quad} = \quad + \quad -$$

(0.00007) (0.0013)

t = 1.59 −3.44

$\bar{}$ = N .19 = DW 60 = 2.11