Solutions to Problem Set 3 (Due March 11)

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

Maximum number of points for Problem set 3 is: 76

Problem 4.1

(i) (2 pts.) Heteroskedasticity generally causes the t statistics not to have a t distribution under H_0. Homoskedasticity is one of the CLM assumptions.

(ii) (2 pts.) The CLM assumptions contain no mention of the sample correlations among independent variables, except to rule out the case where the correlation is one. If two independent variables are perfectly correlated, then the X matrix is not of full rank and we have a problem. Otherwise, partial correlations are acceptable (and likely).

(iii) (2 pts.) An important omitted variable violates Assumption MLR.4 (zero conditional mean), so then the t statistics dont have a t distribution under H_0. For example, suppose we are trying to predict consumption of cigarettes. On the right hand side, we include income but we do not include education. Since income and education are almost surely positively correlated, then the errors would not have zero conditional mean. This would lead to biased estimates of β .

Problem 4.3

(i) (4 pts.) Holding $profmarg$ fixed,

$$\Delta \widehat{rdintents} = .321 \Delta log(sales) = (.321/100)[100 \Delta log(sales)] \approx .00321(\%\Delta sales)$$

Therefore, if $\%\Delta sales = 10, \Delta \widehat{rdintens} \approx .032$, or only about 3/100 of a percentage point. For such a large percentage increase in sales, this seems like a very small effect.

(ii) (4 pts.) $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 > 0$, where β_1 is the population slope on $log(sales)$. The t statistic is $.321/.216 \approx 1.486$. The 5% critical value for a one-tailed test, with df = 32-3 = 29, is obtained from Table G.2 as 1.699; so we cannot reject H_0 at the 5% level. But the 10% critical value is 1.311; since the t statistic is above this value, we reject H_0 in favor of H_1 at the 10% level.

(iii) (2 pts.) With an increase of profit margin by 1 percentage point, expenditures on R&D rise by 0.05 percentage points. Economically that is quite significant, as given a 10% increase in profit margin then they will increase expenditures on R& D by 0.5 percentage point.

(iv) (2 pts.) Not really. Its t statistic is only 0.05/0.046=1.087, so we are not able to reject at even the 10% level.

Problem 4.5

(i) (2 pts.) $.412 \pm 1.96(.094)$, or about [.228 , .596].

(ii) (2 pts.) No, because the value .4 is well inside the 95% CI.

(iii)(2 pts.) Yes, because 1 is well outside the 95% CI.

Problem 4.12

(i) (4 pts.) The coefficient on *lexpend* means that the pass rate ($math10$) increases by 11.16%p, as the expenditure increases by 100% (1.00). Hence, if *expend* increases by 10% (0.1), then the estimated percentage point change in $math10$ is $1.116 = 11.16 \times 0.1$. And the negative intercept means that the estimation is not precise, since a sufficiently low *lexpend* predicts a negative pass rate, even though the pass rate cannot be negative in reality. It seems because of low variation of explanatory variable (*lexpend*). As noted, the minimum value of *lexpend* is 8.11, which is very close to the mean value (8.37). And notice that we cannot evaluate the estimation at the zero of *expend*, since we dont know the value of $log(0)$.

(ii) (4 pts.) Yes. Schools in wealthy towns probably tend to spend more money than in poor district. And the test pass rate can be higher in a wealthy town, as parents can put more effort or out-of-school resources on their children. It implies the common variation of *lexpend* and $math10$ is actually partly accounted by other factors. Hence, if expenditure were randomly assigned to schools to estimate the effect of the expenditure per se, then R-squared would be less.

(iii)(2 pts.) The coefficient on *lexpend* is lower. But it is still statistically significant at the 1% level, as its t-statistic (7.75/3.04=2.55) is still higher than the critical value (2.33).

(iv)(2 pts.) The R-squared is much higher. Hence, we can think the additional two variables improve the models prediction. Beyond those variables, the number of teachers or the number of students per a teacher can be a good candidate as an explanatory variable.

Problem C 4.1

(i) (2 pts.) Holding other factors fixed,

$$\Delta voteA = \beta_1 \Delta log(expendA) = (\beta_1/100)[100\Delta log(expendA)] \approx (\beta_1/100)(\%\Delta expendA)$$

So a .01 increase in expenditure will result in a $(\beta_1/100) * (100 * .01) = .01\beta_1$ change in the vote for A.

(ii) (2 pts.) The null hypothesis is $H_0 : \beta_2 = -\beta_1$, which means a $z\%$ increase in expenditure by A and a $z\%$ increase in expenditure by B leaves voteA unchanged. We can equivalently write $H_0 : \beta_1 + \beta_2 = 0$.

(iii) (4 pts.) The estimated equation (with standard errors in parentheses below estimates) is

$$\widehat{voteA} = 45.08(3.93) + 6.08(0.38)log(expendA) - 6.62(0.39)log(expendB) + .15(0.06)prtystrA$$

$$n = 173, R^2 = .793$$

The coefficient on $log(expendA)$ is very significant (t statistic ≈ 15.92), as is the coefficient on $log(expendB)$ (t statistic ≈ -17.45). The estimates imply that a 10%, ceteris paribus, increase in spending by candidate A increases the predicted share of the vote going to A by about .61 percentage points. [Recall that, holding other factors fixed, $\Delta\widehat{voteA} \approx (6.083/100)\%\Delta log(expendA)$] Similarly, a 10% ceteris paribus increase in spending by B reduces As vote by about .66 percentage points. These effects certainly cannot be ignored.

```
. reg  voteA lexpendA lexpendB prtystrA

      Source |       SS       df       MS              Number of obs =     173
-------------+------------------------------           F(  3,   169) =  215.23
       Model |  38405.1089     3   12801.703           Prob > F      =  0.0000
    Residual |  10052.1396   169  59.4801161           R-squared     =  0.7926
-------------+------------------------------           Adj R-squared =  0.7889
       Total |  48457.2486   172  281.728189           Root MSE      =  7.7123

------------------------------------------------------------------------------
       voteA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lexpendA |   6.083316     .38215    15.92   0.000     5.328914    6.837719
    lexpendB |  -6.615417   .3788203   -17.46   0.000    -7.363247   -5.867588
    prtystrA |   .1519574   .0620181     2.45   0.015     .0295274    .2743873
       _cons |   45.07893   3.926305    11.48   0.000     37.32801    52.82985
------------------------------------------------------------------------------
```

```
. test lexpendA=-lexpendB

 ( 1)  lexpendA + lexpendB = 0

       F(  1,   169) =    1.00
            Prob > F =    0.3196
```

While the coefficients on $log(expendA)$ and $log(expendB)$ are of similar magnitudes (and opposite in sign, as we expect), we do not have the standard error of $\hat{\beta}_1 + \hat{\beta}_2$, which is what we would need to test the hypothesis from part (ii).

(iv) (2 pts.) We fail to reject $\beta_1 + \beta_2 = 0$.

Problem C4.3

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

```
. regress lprice sqrft bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   60.73
       Model |  4.71671468     2  2.35835734           Prob > F      =  0.0000
    Residual |  3.30088884    85  .038833986           R-squared     =  0.5883
-------------+------------------------------           Adj R-squared =  0.5786
       Total |  8.01760352    87  .092156362           Root MSE      =  .19706

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sqrft |   .0003794   .0000432     8.78   0.000     .0002935    .0004654
       bdrms |   .0288844   .0296433     0.97   0.333    -.0300543    .0878232
       _cons |   4.766027   .0970445    49.11   0.000     4.573077    4.958978
------------------------------------------------------------------------------
```

$$\widehat{log(price)} = 4.766(0.10) + .000379(.000043)sqrft + .0289(.0296)bdrms$$

$$n = 88, R^2 = .588$$

Therefore, $\hat{\theta}_1 = 150(.000379) + .0289 = .858$, which means that an additional 150 square foot bedroom increases the predicted price by about 8.6%.

(ii) (2 pts.) $\beta_2 = \theta_1 - 150\beta_1$, and so $log(price) = \beta_0 + \beta_1 sqrft + (\theta_1 - 150\beta_1)bdrms + u = \beta_0 + \beta_1(sqrft - 150bdrms) + \theta_1 bdrms + u$.

(iii) (2 pts.) From part (ii) we run the regression

```
. gen sqrft150=sqrft-150*bdrms

. regress  lprice sqrft150 bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   60.73
       Model |  4.71671468     2  2.35835734           Prob > F      =  0.0000
    Residual |  3.30088884    85  .038833986           R-squared     =  0.5883
-------------+------------------------------           Adj R-squared =  0.5786
       Total |  8.01760352    87  .092156362           Root MSE      =  .19706

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sqrft150 |   .0003794   .0000432     8.78   0.000      .0002935    .0004654
       bdrms |   .0858013   .0267675     3.21   0.002      .0325804    .1390223
```

Really, $\hat{\theta}_1 = .0858$; note we also get $se(\hat{\theta}_1) = .0268$.The 95% confidence interval is .0326 to .1390 (or about 3.3% to 13.9%).

Problem C4.5

(i) (4 points) If we drop $rbisyr$ the estimated equation becomes

$$\begin{array}{lllllll} \widehat{log(salary)} = & 11.02 & + & .0677 & years+ & .0158 & gamesyr \\ & (0.27) & & (.0121) & & (.0016) & \\ & & + & .0014 & bavg+ & .0359 & hrunsyr \\ & & & (.0011) & & (.0072) & \end{array}$$

$$n = 353, R^2 = .625.$$

Now $hrunsyr$ is very statistically significant (t-statistic ≈ 4.99), and its coefficient has increased by about two and one-half times.

(ii) (4 points) The equation with $runsyr$, $fldperc$, and $sbasesyr$ added is

$$\begin{array}{lllllllll} \widehat{log(salary)} = & 10.41 & + & .0700 & years+ & .0079 & gamesyr & & \\ & (0.20) & & (.0120) & & (.0027) & & & \\ & & + & .00053 & bavg+ & .0232 & hrunsyr & & \\ & & & (.00110) & & (.0086) & & & \\ & & + & .0174 & runsyr+ & .0010 & fldperc\text{ -} & .0064 & sbasesyr \\ & & & (.0051) & & (.0020) & & (.0052) & \end{array}$$

$$n = 353, R^2 = .639.$$

Of the three additional independent variables, only *runsyr* is statistically significant (t-statistic = .0174/.0051 ≈ 3.41). The estimate implies that one more run per year, other factors fixed, increases predicted salary by about 1.74%, a substantial increase. The stolen bases variable even has the wrong sign with a t-statistic of about -1.23, while $fldperc$ has a t-statistic of only .5. Most major league baseball players are pretty good fielders; in fact, the smallest $fldperc$ is 800 (which means .800). With relatively little variation in $fldperc$, it is perhaps not surprising that its effect is hard to estimate.

(iii) (4 points) From their t-statistics, *bavg*, $fldperc$, and *sbasesyr* are individually insignificant. The F-statistic for their joint significance (with 3 and 345 df) is about .69 with p-value ≈ .56. Therefore, these variables are jointly very insignificant.

Problem C4.9

(i) (2 points) The results from the OLS regression, with standard errors in parentheses, are

$$\begin{array}{llllllll} \widehat{log(psoda)} = & \underset{(0.29)}{-1.46} & + & \underset{(.031)}{.073} & prpblck + & \underset{(.027)}{.137} & log(income) & \underset{(.133)}{+.380} \; prppov \end{array}$$

$$n = 401 R^2 = .087.$$

The p-value for testing $H_0 : \beta_1 = 0$ against the two-sided alternative is about .018, so that we reject H_0 at the 5% level but not at the 1% level.

(ii) (2 points) The correlation is about -.84, indicating a strong degree of multicollinearity. Yet each coefficient is very statistically significant: the t statistic for $\hat{\beta} log(income)$ is about 5.1 and that for $\hat{\beta} prppov$ is about 2.86 (two-sided p-value = .004).

(iii) (2 points) The OLS regression results when $log(hseval)$ is added are

$$\begin{array}{lllllll} \widehat{log(psoda)} = & \underset{(0.29)}{-.84} & + & \underset{(.029)}{.098} & prpblck\ - & \underset{(.038)}{.053} & log(income) \\ & & + & \underset{(.134)}{.052} & prppov\ + & \underset{(.018)}{.121} & log(hseval) \end{array}$$

$$n = 401 R^2 = .184.$$

The coefficient on $log(hseval)$ is an elasticity: a one percent increase in housing value, holding the other variables fixed, increases the predicted price by about .12 percent. The two-sided p-value is zero to three decimal places.

(iv) (4 points) Adding $log(hseval)$ makes $log(income)$ and *prppov* individually insignificant (at even the 15% significance level against a two-sided alternative for $log(income)$, and *prppov* is does not have a t statistic even close to one in absolute value). Nevertheless, they are jointly significant at the 5% level because the outcome of the $F_{2,396}$ statistic is about 3.52 with p-value = .030. All of the control variables - $log(income)$, *prppov*, and $log(hseval)$ - are highly correlated, so it is not surprising that some are individually insignificant.

(v) (2 points) Because the regression in (iii) contains the most controls, $log(hseval)$ is individually significant, and $log(income)$ and $prppov$ are jointly significant, (iii) seems the most reliable. It holds fixed three measures of income and affluence. Therefore, a reasonable estimate is that if the proportion of blacks increases by .10, $psoda$ is estimated to increase by 1%, other factors held fixed.