Solutions to Problem Set 2 (Due October 6)

EC 228 02, Fall 2010 Prof. Baum, Ms Hristakeva

Maximum number of points for Problem set 2 is: 100

Problem C.1.2

- (i) (2 pts.) You can see the number of women total as well as the number of women who smoked by using the command **tabulate cigs**. We can see there are 694 women (from total observations) in the sample. We can also see that 574 women reported smoking no cigarettes, so 120 women, or about 17 percent, smoke.
- (ii) (2 pts.) Use the command **summarize cigs**. We can see the average number of cigarettes smoked per day by women is 2.41. This is not a good measure of the typical woman since 574/694 or about 83 percent of women were in fact non-smokers.
- (iii) (2 pts.) Use the command **summarize cigs if cigs>0** which gives the average daily consumption for smokers of 13.95 cigarettes. Clearly this estimate is much higher than our estimate in part (ii) due to the fact that we are only looking now at the 120 women that actually smoke.
- (iv) (1 pt.) Use the command **summarize fatheduc** and we can see there are only 589 observations instead of 694. There are missing values meaning some people did not answer the question about father's education level.
- (v) (1 pt.) Use the command **summarize faminc** and see that the mean family income is 29036 dollars with a standard deviation of 18534 dollars.

Problem C.1.3

- (i) (1 pt.) Use the command **summarize math4**. Then you can see the highest value is 100 and the lowest value is zero. This means the best school had a pass rate of 100 percent and the worst school had a pass rate of zero. This makes intuitive sense.
- (ii) (2 pts.) Use the command **tabulate math4 if math4==100**. You will see that only 19 schools had perfect pass rates, or dividing by 729 (the total number of schools), then 2.6 percent of schools had perfect pass rates.
- (iii) (2 pts.) Use the command **tabulate math4 if math4==50**. So 6 schools have pass rates of exactly 50 percent.

- (iv) (2 pts.) Use the command **summarize math4 read4**. We can see that the pass rate for math is higher (72.89 percent relative to 60.81 percent) yet we can't be sure of the significance since the standard deviations are 19.76, 19.16 respectively, meaning the means lie within one standard deviation of each other.

- (v) (2 pts.) Use the command **correlate math4 read4** and we see there is a positive correlation of 86.86 percent, meaning there is a strong positive relationship between schools which do well on the reading test and schools which do well on the math test.

- (vi) (2 pts.) Use the command **summarize exppp** to find that mean spending is 5168 dollars with a standard deviation of 1057 dollars. Dividing 1057/5168 tells us that the standard deviation is about 20 percent of the mean. Therefore, since we know approximately 95 percent of schools spend within 2 deviations of the mean, we can compute an estimate that 95 percent of schools spend between $[5168 - 2 * 1057, 5168 + 2 * 1057] = [3054, 7282]$ which doesn't seem like much variation.

- (vii) (1 pt.) A's spending exceeds B's by $(6000 - 5500)/5500 = 9.09$ percent. We get a lower percentage when taking logs: $100 * [log(6000) - log(5500)] = 8.70$ percent.

Problem C.1.4

- (i) (2 pts.) Use the command **tabulate train** to see that $185/445 = 41.6$ percent of men receive job training.

- (ii) (2 pts.) Use the command **bysort train: summarize re78** to see that average earnings from men not receiving job training is 4555 dollars and that average earnings for men who receive job training is 6349 dollars. This is a difference of $6349-4555/4555$ or 39.36 percent.

- (iii) (2 pts.) Use the command **tabulate unem78 train, column**. You can see that unemployment for those who did not receive job training is 35.38 percent while it is only 24.32 for those who received job training. This seems very economically significant.

- (iv) (2 pts.) We cannot tell the effectiveness of the job training program from these results only. Perhaps **unobserved characteristics** are at work; for example, only the more gifted workers were offered job training. Since we cannot observe these differences between workers, if we want to measure the effectiveness of the job training program we should look at differences in wages earned for the same worker before training and after training.

Problem 2.4

- (i) (2 pts.) When cigs = 0, predicted birth weight is 119.77 ounces. When cigs = 20, $\widehat{bwght} = 109.49$. This is about an 8.6% drop.

- (ii) (2 pts.) Not necessarily. There are many other factors that can affect birth weight, particularly overall health of the mother and quality of prenatal care. These could be correlated with cigarette smoking during birth. Also, something such as caffeine consumption can affect birth weight, and might also be correlated with cigarette smoking.

- (iii) (4 pts.) If we want a predicted *bwght* of 125, then $cigs = (119.77 - 125)/(524) \approx -10.18$, or about negative 10 cigarettes! This is nonsense, of course, and it shows what happens when we are trying to predict something as complicated as birth weight with only a single explanatory variable.

- (iv) (2 pts.) Yes. Since about 80 percent of women did not smoke, but we only have one birthweight estimate when $cigs = 0$ (since we are only using cigs to explain birth weight) then the predicted birthweight for $cigs = 0$ is in the middle of the entire distribution of birth weights when $cigs = 0$. If we believe that non-smokers have heavier babies than smokers, then we would under-predict high birth weights.

Problem 2.5

- (i) (4 pts.) The intercept implies that when income is 0, consumption is predicted to be negative 124.84 dollars. This, of course, cannot be true, and is reflective of this consumption function being a poor predictor of consumption at very low income levels. On the other hand, relative to annual income, 124.84 dollars is not so far from zero.

- (ii) (2 pts.) Just plug 30,000 into the equation: $\widehat{cons} = -124.84 + .853(30,000) = 25,465.16$ dollars.

- (iii) (2 pts.) The MPC is a straight line at $\beta_1 = .853$. While the APC has a negative intercept, for incomes of very low range, ie $inc = 1000$, the APC is greater than zero. As *inc* goes to infinity, APC approaches MPC from below.

Problem 2.6

- (i) (2 pts.) The coefficient is .312 meaning there is a positive relationship between housing prices and distance from a garbage incinerator. This is what we would expect. Increasing a house's distance from a garbage incinerator, holding everything else constant, should result in an increased home price.

- (ii) (2 pts.) If the city chose to locate the incinerator in an area away from more expensive neighborhoods, then $log(dist)$ is positively correlated with housing quality. This would make OLS estimation biased.

- (iii) (2 pts.) Size of the house, number of bathrooms, size of the lot, age of the home, proximity to parks, and quality of the neighborhood (including school quality) are just a handful of factors that could influence price. As mentioned in part (ii), these could

certainly be correlated with dist [and $log(dist)$]. For example, it might be likely that a city planner would want to place the garbage incinerator away from the park and closer to a manufacturing area.

Problem C.2.1

- (i) (2 pts.) Use the command **summarize prate** to find the mean participation rate is 86.88 percent. Use the command "summarize mrate" to find that companies, on average, match 75 cents per dollar of worker's contribution.
- (ii) (4 pts.) Use the command **regress prate mrate**. So $\widehat{prate} = 82.0138 + 6.482331 * mrate$. Sample size is 767, and R^2 is 0.0895.
- (iii) (4 pts.) If $mrate = 0$, the predicted participation rate is 82.0138 percent, meaning 82 percent of people would participate in a 401k plan even if their company matched nothing. Coefficient in $mrate$ implies that a one dollar increase in the match rate is estimated to increase participation by 6.482331 percentage points. This assumes, of course, that this change in prate is possible (ie prate cannot be more than 100 percent.)
- (iv) (2 pts.) If we plug 3.5 in the equation, we get $prate = 82.0138 + 3.5 * 6.482331 = 104.702$. This is impossible, as we can have at most a 100 percent participation rate. This illustrates that, especially when dependent variables are bounded (as in the case of percentages), a simple regression model can give strange predictions for extreme values of the independent variable. (In the sample of 765 firms, only 15 have $mrate > 3.5$.)
- (v) (2 pts.) From our estimate of R^2, $mrate$ explains 8.95 percent of the variation. This is not much, and many other factors may affect participation rate.

Problem C.2.4

- (i) (2 pts.) Use the command **summarize IQ wage**. So the average salary is about 957.95 dollars and the average IQ is about 101.28. The sample standard deviation of IQ scores is 15.05 which is fairly close to the population standard deviation of 15.
- (ii) (4 pts.) Use the command **regress wage IQ**. We get that $\widehat{wage} = 116.99 + 8.30IQ$, n=935, $R^2 = 0.096$. An increase in IQ by 15 points would raise predicted monthly salary by $8.30 * 15 = 124.50$. IQ only explains 9.6 percent of variation in wage which isn't very much.
- (iii) (4 pts.) Use the command **regress lwage IQ** where lwage is the log(wage). So $\widehat{log(wage)} = 5.89 + 0.0088IQ$, n=935, $R^2 = .099$. If $\triangle IQ = 15$ then $\triangle \widehat{log(wage)} = 0.0088(15) = 0.132$, which is (approximate) proportinate change in predicted wage. The percentage increase is therefore approximately 13.2.

Problem C.3.2

(i) (2 pts.)

```
    . regress price sqrft bdrms
          Source |       SS       df       MS          Number of obs =      88
-------------+------------------------------          F(  2,    85) =   72.96
       Model |  580009.152     2  290004.576          Prob > F      =  0.0000
    Residual |  337845.354    85  3974.65122          R-squared     =  0.6319
-------------+------------------------------          Adj R-squared =  0.6233
       Total |  917854.506    87  10550.0518          Root MSE      =  63.045

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sqrft |   .1284362   .0138245     9.29   0.000     .1009495    .1559229
       bdrms |   15.19819   9.483517     1.60   0.113    -3.657582    34.05396
       _cons |    -19.315   31.04662    -0.62   0.536    -81.04399      42.414
------------------------------------------------------------------------------
```

The estimated equation is

$$\widehat{price} = -19.32 + .128sqrft + 15.20bdrms$$

$$n = 88,\ R^2 = .632$$

(ii) (2 pts.) Holding square footage constant,$\widehat{\triangle price} = 15.20\triangle bdrms$, and so $\widehat{price}$ increases by 15.20, which means \$15,200.

(iii) (2 pts.) Now $\widehat{\triangle price} = .128\triangle sqrft + 15.20\triangle bdrms = .128(140) + 15.20 = 33.12$,or \$33,120. Because the size of the house is increasing, this is a much larger effect than in(ii).

(iv) (2 pts.) About 63.2% from R^2.

(v) (2 pts.) The predicted price is $-19.32 + .128(2,438) + 15.20(4) = 353.544$, or \$353,544.

(vi) (2 pts.) From part (v), the estimated value of the home based only on square footage and number of bedrooms is \$353,544. The actual selling price was \$300,000, which suggests the buyer underpaid by some margin. But, of course, there are many other features of a house (some that we cannot even measure) that affect price, and we have not controlled for these.

Problem C.3.4

(i) (2 pts.)The minimum, maximum, and average values for these three variables are given in the table below. Use the command "summarize atndrte priGPA ACT".

Variable	Average	Minimum	Maximum
atndrte	81.71	6.25	100
priGPA	2.59	0.86	3.93
ACT	22.51	13	32

(ii) (4 pts.)

```
. regress atndrte priGPA ACT
      Source |       SS       df       MS              Number of obs =     680
-------------+------------------------------           F(  2,   677) =  138.65
       Model |  57336.7612     2  28668.3806            Prob > F      =  0.0000
    Residual |  139980.564   677  206.765974            R-squared     =  0.2906
-------------+------------------------------           Adj R-squared =  0.2885
       Total |  197317.325   679   290.59989            Root MSE      =  14.379

------------------------------------------------------------------------------
     atndrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   17.26059   1.083103    15.94   0.000     15.13395    19.38724
         ACT |  -1.716553    .169012   -10.16   0.000    -2.048404   -1.384702
       _cons |    75.7004   3.884108    19.49   0.000     68.07406    83.32675
------------------------------------------------------------------------------
```

The estimated equation is: $\widehat{atndrte} = 75.70 + 17.26priGPA - 1.72ACT$

$$n = 680, R^2 = 0.291$$

The intercept means that, for a student whose prior GPA is zero and ACT score is zero, the predicted attendance rate is 75.7%. But this is clearly not an interesting segment of the population. (In fact, there are no students in the college population with $priGPA = 0$ and $ACT = 0$, or with values even close to zero.)

(iii) (2 pts)The coefficient on $priGPA$ means that, if a students prior GPA is one point higher (say, from 2.0 to 3.0), the attendance rate is about 17.3 percentage points higher. This holds ACT fixed. The negative coefficient on ACT is, perhaps initially a bit surprising. Five more points on the ACT is predicted to lower attendance by 8.6 percentage points at a given level of $priGPA$. As $priGPA$ measures performance in college (and, at least partially, could reflect, past attendance rates), while ACT is a measure of potential in college, it appears that students that had more promise (which could mean more innate ability) think they can get by with missing lectures.

(iv) (2 pts)We have $\widehat{atndrte} = 75.70 + 17.267(3.65) - 1.72(20) \approx 104.3$. Of course, a student cannot have higher than a 100% attendance rate. Getting predictions like this is always possible when using regression methods for dependent variables with natural upper or lower bounds. In practice, we would predict a 100% attendance rate for this student. (In fact, this student had an actual attendance rate of 87.5%.)

(v) (2 pts)The difference in predicted attendance rates for A and B is $17.26(3.1 - 2.1) - (21 - 26) = 25.86$.