Problem Set 4 Solution

Chapter 18
1. 20 and 25
5. (i) histogram for the sum. It is becoming a normal curve.
   (ii) histogram for the product.
   (iii) histogram for numbers to be drawn.

Chapter 20.
5. average weight for a guest : 150 lbs.
   4 tons = 8000 lbs.
   50 x 150 = 7500 lbs.
   SE = 35 x $\sqrt{50}$ = 247.5
   $\therefore$ 7500 ± 2 SE = 7500 ± 2(247.5) = 7995 and 7005.
   And the range of 7005 lbs. and 7995 lbs. covers more than 95.45 percentage of selected 50 people's sum of weights. Therefore, the percentage of the group's being 8000 lbs. is far right side of a curve, which is about **2.275**. (100 – 95.45 = 4.55, 4.55/2 = 2.275)

6. (ii) The sample size here is 0.1 percentage of the total population in each state. For California, the sample size is 30,000 and the sample size of Nevada is 1,000. With a larger sample size, the accuracy is expected to be higher in California than in Nevada.

8. Total population : 30,000 Total Democrats : 12,000
   Pr(Democrats) = 12,000/30,000 = .4
   Having 50-50 chance implies the symmetry of the theoretical sampling distribution. Since the theoretical sampling distribution is symmetric around the estimated mean,

   $\therefore$ E(Democrats in sample) = .4 x 1,000 [Pr(Dem) x sample size] = **400**.

Chapter 21.
1. 15.8 percentage of the total American household is expected to have computer. Therefore,
   E(HH with computer in the town with 25,000 population) = 25,000 x .158 = 3,950.

a. In order to calculate the mean and the SE of the sample,
   79/500 = .158 (15.8 %). SE = [$\sqrt{}$ (.158)x(1-.158)] / $\sqrt{}$ (500) = .365 / $\sqrt{}$ (500) = .0163 (1.63 %).
   $\therefore$ The percentage of households in the town with computers is estimated as **15.8 %** : this estimate is likely to be off by **1.63 %** or so.

b. CI = .158 ± 2 x .0163 = .1906 and .1254. Therefore, the confidence intervals are 12.54 % and 19.06%.

2. Pr(HH with refrigerator of the sample) = 498/500 = .996 (99.6).
   SE = [$\sqrt{}$ (.996)x(1-.996)] / $\sqrt{}$ (500) = .00282 (.282%).

a. The percentage of households in the town with refrigerators is estimated as **99.6 %**; this estimate is likely to be off by **.282 %**.

b. CI = .996 ± 2 x .00282 = .1.00164 and .99036. The upper bound of confidence interval is greater than 100 %. We cannot create the upper CI in this case, but the lower bound of the confidence interval is 99.036 %.

12. (i) irrelevant
(ii) a histogram for the numbers drawn.
(iii) a probability histogram for the sum.

14. sample size = 1,500.
Pr(renters of the town from the sample) = 1035/1500 = .69 (69 %).
E(renters of the sample) = .69.
SE(renters of the sample) = [√ (.69)x(1-.69)] / √ (1500) = .012 (1.2%).

a. The expected value for the percentage of sample persons who rent is **exactly equal to** 69 %.

*note: the question is asking the expected value and SE of the sample not the population that we can estimate from the sample. Therefore, the values are all exactly equal to the calculated numbers from the sample.

b. The SE for the percentage of sample persons who rent is **estimated from the data** 1.2 %.

Chapter 23.

10. population size = 80,000 SD = 1.75.
sample size = 625 average no. of persons in a household = 2.30.

a. True.
SE = 1.75 / √625 = .07

b. False.
There is no point to calculate the CI for the sample. We calculate the CI to check out whether our estimates safely fall in the range of the population.

c. True.
2.30 ± 2 x .07 = 2.44 and 2.16.

d. False.
This is simply a misinterpretation of a confidence interval.

e. False.
The Central Limit Theorem is the claim that if you repeat the drawing of the samples from the population, the shape of the sample averages becomes a normal curve.

f. True.
Explained above.

12. 400 is the size of a population not a sample. A confidence interval is used to confirm the accuracy of the estimates obtained from a sample. Thus, the confidence interval, in this case, is meaningless.

Chapter 26.

2. Pr(red numbers) = 18/38 = .474
sample size = 3800 red numbers in the sample = 1890.
Pr(red numbers in the sample) = 1890/3800 = .497

a. H_0 : Pr(red numbers) = .474
* interpretation : the difference between .474(population) and .497(sample) is due to a chance error. OR .479 is obtained due to a chance error.
H_1 : Pr(red numbers) > .474

* interpretation : the difference between .474(population) and .497(sample) is not due to a chance error but to a systematic effect.

b. $Z = (.497 - .474) / SE$
$SE = SD / \sqrt{3800} = [\sqrt{(.474)x(1-.474)}] / \sqrt{(3800)} = .0081$
$\therefore Z = (.497368 - .473684) / .0081 = 2.924$
p-value = 1 - .99825 = .00175. (less than .05, 5 % of significance level)

c. Both of Z score and p-value indicate there are too many reds and it is not by chance error.

4. population = 900 students ; final average = 63 & SD = 20
a section = 30 students ; final average = 55
H_0 : the mean of final = 63
H_1 : the mean of final ≠ 63
$SE = 20 / \sqrt{30} = 3.651$
$Z = (55 - 63) / 3.651 = - 2.19$
p-value = .0139
$\therefore$Both of Z score and p-value show that the difference between the population average and the sample average is not caused by a chance error. The section of this TA did poorer job than the average.

6. venire = 350 ; women = 102. Pr(women in the venire) = 102/350 = .2914.
juror group = 100 ; women in juror group = 9. Pr(women juror) = 9/100 = .09.
However, a majority of the eligible jurors in the district were female; namely, more than half of the eligible jurors in the district were women. Is that a good selection?

a. mean = .2914 ; and let's assume that (at least) 50 percent of the population is women. $SE = [\sqrt{(.5)x(1-.5)}] / \sqrt{(350)} = .0267$.
$Z = (.2914 - .5) / .0267 = -7.6142$
p-value = .0000…1
Therefore, the under-representation of women in the venire selection is not due to a chance error. Something's wrong!

b. E(women juror) = .2914 x 100 = 29.14 Since there are 102 women out of 350 people in the venire, we expect to see 29 women jurors. Actual number of women juror = 9 (.09)
$SE = [\sqrt{(.2914)x(1-.2914)}] / \sqrt{(100)} = .0454$
$Z = (.09 - .2914) / .0454 = - 4.4361$
p-value = .001
Again, the under-representation of women jurors is statistically significant.

c. Therefore, there's something wrong. It's very unlikely for this kind of juror selection to happen by chance.

7. total patients in a month = 1022
odd days : 580 even days : 442
it should be evenly divided and showing 50-50 entrance rate if there is no error whatsoever.
Pr(odd days in the sample) = 580/1022 = .5675
Expected Pr(odd days) = .5
$SE = [\sqrt{(.5)x(1-.5)}] / \sqrt{(1022)} = .0156$
$Z = (.5675 - .5) / .0156 = 4.32$
p-value = .0008

From the Z score and p-value, we can see that more people came to the hospital on odd days. We must therefore disagree with the observer's treatment of this like a coin toss.

Chapter 29.

1. (a) True. Even though the difference is highly significant (say, p = .01), there is still the possibility that the cause of the difference is chance error (very unlikely, though.). This is exactly what p-value means.
   (b) False. A statistically significant number is not only dependent of the actual number, but also the size of a sample.
   (c) It could be true and false. P-value of .047 and .052 are just about the same magnitude, but can be treated differently. For instance, when a researcher set the critical value as .05 (as in most cases), the estimate with .052 p-value is not significant and the null hypothesis should fail to be rejected, whereas the one with .047 is treated as statistically significant and the null should be rejected.

2. (i) Is the difference due to chance?
   The whole idea of hypothesis testing is to see whether the difference between expected values and observed values are caused by chance. Thus, Z scores are (intuitively) normalized differences and p-values represent the probability that the normalized Z-score can emerge by chance. Apparently, the smaller a p-value, the lower the probability that the difference is due to a chance error.

3. average of box = 50
   X_1 : sample size = 100, SE = SD / $\sqrt{(100)}$ = 10 / 10 = 1
   X_2 : sample size = 300, SE = SD / $\sqrt{(900)}$ = 10 / 30 = .3333
   The statement is FALSE. Z-scores and p-values are not only dependent on average differences, but also of standard errors. Here, the investigator 2 has a larger sample size, and it results in different SE's for the two investigators. Therefore, the investigator whose **z-score** (not average) is further from 0 will get the smaller p-value, which might be the case for the investigator 2.

6. β = .07 ; SE = .05
   Z = .07 / .05 = 1.4
   Even though we did not set the critical value, conventional wisdom provides us with Z = 1.96 and p-value ≤ .05 as cut-off values for statistical significance. Here, Z score is not statistically significant according to the p-value = .05, which confirms that there is "no impact." However, if we set the cut-off value higher than .05, namely, p = .1, the conclusion is completely different: the impact is statistically significant. Therefore, to be accurate, we can conclude that it is more likely there is a positive relationship between inflation and voting behavior, but the actual magnitude of the influence is not precisely estimated

8. female employment in the United States = 50.4 % in 1985.
   female employment in the United States = 54.1 % in 1993.

a. The question asks whether the change in women's employment is statistically significant between 1985 and 1993. Even though it is based on population survey, if female employment in 1985 and 1993 are considered as realizations of an economic theory of the United States, comparing the difference makes sense for hypothesis testing.

b. However, we cannot perform the test because it is a cluster sample and doesn't have sufficient information. All the numbers given are from the population not from a sample. Even though we can calculate the Z score, it is meaningless.

c. H_0 : female employment rate in 1985 = female employment rate in 1993.
H_1 : female employment rate in 1985 ≠ female employment rate in 1993.
$SE_{1985} = \sqrt{} (.504)x(1-.504) / \sqrt{}50,000 = .002236$
$SE_{1993} = \sqrt{} (.541)x(1-.541) / \sqrt{}50,000 = .002229$
$SE = \sqrt{} (.5225)x(1-.5225) / \sqrt{}50,000 = .00223$
$Z = (54.1 – 50.4) / \sqrt{} (.00223) = 16.6$: p-value = .000….1
Thus, we can conclude that the change is highly significant.

11. sample size = 250 TV = 38 % ; Radio = 30 %
Statistically, the question makes sense, therefore, you can answer it. Assume that TV viewing rates and Radio listening rates are the same and set the Radio listening rate as a mean.
$SE = \sqrt{} (.34)x(1-.34) / \sqrt{}250 = .03$
$Z = (.38 - .30) / .03 = 2.676$: p-value = 1 – .9907 = .0093.
Thus, we can conclude that the respondents spend more time watching TV than listening to the radio. The problem here is how accurate the responses were. That is, even though it proved that people spend more time on TV than on radio according to the test result, it may be difficult to state so unless you know how reliable people's memories were when they answered the question.

PART II.

1. $Z = (X – 0)/1 = X$
   a. $Pr(X \geq 0) = .5$
   b. $Pr(X \geq .84) = .2005$
   c. $Pr(X \geq 1.96) = .025$
   d. $Pr(-1.96 \leq X \leq 1.96) = .05$

2. a. $Pr(X< Z) = .975 \Rightarrow 1.96$
   b. $Pr(X< Z) = .95 \Rightarrow 1.645$
   c. $Pr(-Z \leq X \leq Z) = .975 \Rightarrow 2.24$
   d. $Pr(-Z \leq X \leq Z) = .95 \Rightarrow 1.96$

3. a. $X \sim N(4, 9)$
   $Z = (X – 4) / 3 = (6.5 – 4) / 3 = 2.5/3 = .8333…$
   p-value = 1 - .7995* = .2005.
   * note: you can find this value from the table at the end of any statistics book.
   b. $X \sim N(-3, 4)$
   $Z = (X + 3) / 2 = (6.5 + 3) / 2 = 9.5/2 = 4.75$
   p-value = very close to zero. (.00…1)

4. $X \sim T(0,1)$ d.f. = 20

a. $t = (X – 0) / 1 = 2.09 = .025$.
b. .05
c. $t = 2.09$
d. $t = 2.85$.

5. $X \sim T(3, 2.25)$
$t = (X – 3) / \sqrt{2.25}$ d.f. = 20
a. $Pr(X > 1.155) = (1.155 – 3) / \sqrt{2.25} = -1.845 / 1.5 = -1.23$.
According to the t-table, the area covered above –1.23 with d.f. of 20 is around 85 %.
b. $(X – 3) / 1.5$ with d.f. of 20 to cover 99 %, ± t should be 2.85.

Note: When you calculate z-score or t-score, the equation is :

$$\frac{X - mean}{SE}$$

Usually, (X-mean) is calculated in absolute term, and the order does not matter in 2-tailed test. But, if you are doing 1-tailed test, be careful about the order not to be (mean – X). If you have a correct intuition about this, it won't be a big problem (since you can convert it in the context of a normal distribution), but it could be confusing.

Part III.

Question A

1. The first part of the problem asks you to run the multiple regression to predict room choice.

```
. reg firstchoice yearbuilt roomsize

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =    2.12
       Model |  3963.17801     2  1981.58901           Prob > F      =  0.1901
    Residual |  6530.42199     7  932.917427           R-squared     =  0.3777
-------------+------------------------------           Adj R-squared =  0.1999
       Total |    10493.60     9  1165.95556           Root MSE      =  30.544

------------------------------------------------------------------------------
 firstchoice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   yearbuilt |   .9285734   .8766258     1.06   0.325    -1.144317    3.001464
    roomsize |  -.1171777    .688022    -0.17   0.870    -1.744091    1.509736
       _cons |  -1717.581   1616.999    -1.06   0.323    -5541.176    2106.013
------------------------------------------------------------------------------
```

```
2. The second part of the problem asks you to run the two bivariate components
of part (1).

. reg firstchoice  yearbuilt

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    4.80
       Model |  3936.11796     1  3936.11796           Prob > F      =  0.0598
    Residual |  6557.48204     8  819.685255           R-squared     =  0.3751
-------------+------------------------------           Adj R-squared =  0.2970
       Total |    10493.60     9  1165.95556           Root MSE      =   28.63

------------------------------------------------------------------------------
 firstchoice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   yearbuilt |   .7945975   .3626074     2.19   0.060    -.0415767    1.630772
       _cons |  -1473.848   705.5833    -2.09   0.070    -3100.926    153.2298
------------------------------------------------------------------------------

. reg firstchoice roomsize

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    3.08
       Model |  2916.41791     1  2916.41791           Prob > F      =  0.1174
    Residual |  7577.18209     8  947.147761           R-squared     =  0.2779
-------------+------------------------------           Adj R-squared =  0.1877
       Total |    10493.60     9  1165.95556           Root MSE      =  30.776

------------------------------------------------------------------------------
 firstchoice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    roomsize |   .5368167   .3059215     1.75   0.117    -.1686396    1.242273
       _cons |  -5.423699   45.29416    -0.12   0.908    -109.8722    99.02482
------------------------------------------------------------------------------
```

Comparing the coefficients between the multivariate and bivariate cases shows that something may be a bit amiss in the multivariate case. The standard errors are really big in the multivariate regression compared to the bivariate regressions and the coefficients have changed a lot. The variable roomsize is even a different sign! This sounds a lot like multicollinearity – the use of two independent variables measuring the same underlying factor. In this case, it is possible that the newer the building is, the larger the rooms in response to students' expressed preferences over the years. To see if this is the case, consider the following regression:

```
. reg roomsize yearbuilt

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =   33.08
       Model |  8149.61788     1  8149.61788           Prob > F      =  0.0004
    Residual |  1970.78212     8  246.347765           R-squared     =  0.8053
-------------+------------------------------           Adj R-squared =  0.7809
       Total |    10120.40     9  1124.48889           Root MSE      =  15.695

------------------------------------------------------------------------------
    roomsize |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   yearbuilt |   1.143357   .1987867     5.75   0.000     .6849536     1.60176
       _cons |  -2080.029   386.8112    -5.38   0.001    -2972.017   -1188.041
------------------------------------------------------------------------------
```

Clearly the older the building is, the larger the rooms. More than 80% of the variance in roomsize is explained by the year of construction. This is definitely a multicollinearity problem.

3. So which model is best? The multivariate case is clearly the wrong one to use as described in part (2). Since there is reason to believe that roomsize is a function of how recently the dorm was built, and that there are also advantages to newer buildings generally, the best model is probably the bivariate case using yearbuilt as the independent variable.

Beyond the theoretical reasons to use the bivariate case with yearbuilt, this model is also the only one with a statistically significant coefficient, and the highest R^2 (0.37) which further establishes our confidence in this conclusion.

It is a shame, however, to simply throw away the information that's found in the size-of-rooms variable. If this regression was part of a larger study in which it was important to control for the physical characteristics of a building, but only as a control to eliminate omitted variables bias, then a common solution would be to create a scale that would combine the yearsbuilt and roomsize variables. You could do this by subtracting each variable from its mean and dividing by its standard deviation and then adding together the z-scores. You would then have a unitless measure of "building quality" which might predict firstchoice better than either variable would alone. Turns out in this case that there is no real improvement by building such a combined variable (see below) – it really appears to be just a set of collinear variables.

```
summ  yearbuilt roomsize
```

```
    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
   yearbuilt |      10      1945.7   26.31877       1910       1981
    roomsize |      10       144.6    33.5334         97        200

. gen zyearbuilt=(1945.7-yearbuilt)/26.31877

. gen zroomsize=(144.6- roomsize)/33.5334

. gen quality= zyearbuilt+ zroomsize

. reg firstchoice quality

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    4.16
       Model |   3591.4989     1   3591.4989           Prob > F      =  0.0756
    Residual |   6902.1011     8  862.762637           R-squared     =  0.3423
-------------+------------------------------           Adj R-squared =  0.2600
       Total |    10493.60     9  1165.95556           Root MSE      =  29.373

------------------------------------------------------------------------------
 firstchoice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quality |  -10.25477   5.026131    -2.04   0.076    -21.84505    1.335507
       _cons |       72.2   9.288502     7.77   0.000     50.78068    93.61932
------------------------------------------------------------------------------
```

Once again, I am inclined to trust the bivariate regression with yearbuilt. From this combined variable regression, you can see that the t-statistic and R^2 have become smaller relative to the yearbuilt.

Question B

```
. reg cvote82 pvote80 newtown;

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  2,    61) =  111.19
       Model |  .429321715     2  .214660857           Prob > F      =  0.0000
    Residual |  .117767969    61  .001930622           R-squared     =  0.7847
-------------+------------------------------           Adj R-squared =  0.7777
       Total |  .547089684    63  .008683963           Root MSE      =  .04394

------------------------------------------------------------------------------
     cvote82 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     pvote80 |   1.088667     .08096    13.45   0.000     .9267773    1.250556
     newtown |  -.0674145   .0110094    -6.12   0.000    -.0894291      -.0454
       _cons |  -.0025771    .057215    -0.05   0.964    -.1169856    .1118314
------------------------------------------------------------------------------
```

Courter did more poorly in newtowns by 0.067 percentage of vote. That is, he lost 0.067 percentage of votes in new towns.

```
. reg cvote82 pvote80;

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  1,    62) =  116.38
       Model |   .35693151     1   .35693151           Prob > F      =  0.0000
    Residual |  .190158174    62  .003067067           R-squared     =  0.6524
-------------+------------------------------           Adj R-squared =  0.6468
       Total |  .547089684    63  .008683963           Root MSE      =  .05538

------------------------------------------------------------------------------
     cvote82 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     pvote80 |   1.100501    .102014    10.79   0.000      .896578    1.304424
       _cons |  -.0466514   .0715417    -0.65   0.517    -.1896611    .0963583
------------------------------------------------------------------------------
. bys newtown : reg cvote82 pvote80;

_______________________________________________________________________________
-> newtown = 0

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   24.45
       Model |  .028806365     1  .028806365           Prob > F      =  0.0000
    Residual |  .032991833    28   .00117828           R-squared     =  0.4661
-------------+------------------------------           Adj R-squared =  0.4471
       Total |  .061798198    29  .002130972           Root MSE      =  .03433

------------------------------------------------------------------------------
     cvote82 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     pvote80 |   .7077853   .1431468     4.94   0.000     .4145625    1.001008
       _cons |   .2639357   .1003594     2.63   0.014     .0583587    .4695127
------------------------------------------------------------------------------
```

```
-> newtown = 1

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  1,    32) =  142.20
       Model |  .33065637     1  .33065637             Prob > F      =  0.0000
    Residual | .074410661    32  .002325333            R-squared     =  0.8163
-------------+------------------------------           Adj R-squared =  0.8106
       Total | .405067031    33  .012274759            Root MSE      =  .04822

------------------------------------------------------------------------------
     cvote82 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     pvote80 |   1.181061   .0990436    11.92   0.000     .9793155    1.382806
       _cons |  -.1343421   .0694759    -1.93   0.062    -.2758598    .0071755
------------------------------------------------------------------------------
```

Partisanship plays an important role overall if you look at the first regression. It shows that Courter received 1.1 % additional votes, if the percentage of vote for Reagan of a town increases by 1 %. However, the deviation gets smaller if it is a new district, while it does larger in an old district. In a new district, Courter received an additional 1.18 % of the vote as the people of the district vote for Reagan increased 1 %. In old districts, the partisan effect attenuates to 0.708 %. Therefore, partisanship has more effect in a new town. The votes that Courter got in an old town is due to another reason, namely the name recognition effect driven by the incumbency advantage.

```
. gen newt_p = newtown*pvote80;


. reg cvote82 pvote80 newtown newt_p;

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,    60) =   81.88
       Model |  .43968719     3  .146562397            Prob > F      =  0.0000
    Residual | .107402494    60  .001790042            R-squared     =  0.8037
-------------+------------------------------           Adj R-squared =  0.7939
       Total | .547089684    63  .008683963            Root MSE      =  .04231

------------------------------------------------------------------------------
     cvote82 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     pvote80 |   .7077853   .1764367     4.01   0.000     .3548594    1.060711
     newtown |  -.3982779   .1379026    -2.89   0.005    -.6741242   -.1224315
      newt_p |   .4732753   .1966758     2.41   0.019     .0798653    .8666854
       _cons |   .2639357   .1236988     2.13   0.037     .0165013    .5113702
------------------------------------------------------------------------------
```

cvote 82 = $\beta_0 + \beta_1$ pvote80 + β_2 newtown + β_3 pvote80 * newtown + ε
when it is a new town: slope = $\beta_1 + \beta_3$ & intercept = $\beta_0 + \beta_2$
when it is an old town: slope = β_1 & intercept = β_0

Including interaction term considers the different effect of partisanship playing in new towns and old towns. This provides the same result as if you had run two separate regressions of new towns and old towns. The result gives you the same coefficients as in the previous two regressions, confirming the bigger role of partisanship in new towns and incumbency effect in old towns.

Question C

```
. reg rate93q totfac totstu;

      Source |       SS       df       MS              Number of obs =     109
-------------+------------------------------           F(  2,   106) =   45.02
       Model |  34.7971203     2  17.3985601           Prob > F      =  0.0000
    Residual |  40.9693812   106  .386503596           R-squared     =  0.4593
-------------+------------------------------           Adj R-squared =  0.4491
       Total |  75.7665015   108   .70154168           Root MSE      =  .62169

------------------------------------------------------------------------------
     rate93q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totfac |   .0198392   .0054385     3.65   0.000     .0090569    .0306215
      totstu |   .0054181    .001323     4.10   0.000     .0027951    .0080412
       _cons |   1.964446   .1103758    17.80   0.000     1.745615    2.183276
------------------------------------------------------------------------------

. corr rate93q totfac totstu, cov;
(obs=109)

             |  rate93q   totfac   totstu
-------------+---------------------------
     rate93q |  .701542
      totfac |  7.55772  217.865
      totstu |  31.7926  597.155  3681.26
```

$b_1 = Cov(X_1\ Y) / Var(X_1) – b_2\ \ Cov(X_1\ X_2) / Var(X_1)$
rearrange this into :

$Cov(X_1\ Y) / Var(X_1) = b_1 + b_2\ \ Cov(X_1\ X_2) / Var(X_1)$
Here, b_1 is a direct effect and $b_2\ \ Cov(X_1\ X_2) / Var(X_1)$ is an indirect effect. (same logic for b_2).
Plug the numbers obtained variance-covariance table, we can get the following answers:

.034689 = .0198392 + .0054181 x (597.155/217.865)
.008636 = .0054181 + .0198392 x (597.155/3681.26)

	Gross effect	Direct effect	Indirect effect
totfac	.034689	.0198392	.01485
totstu	.008636	.0054184	.0032182

Part C.2.

I used the following variables:

pub_fac : The ratio of the total number of program publications in the period 1988-1992 to the number of program faculty. My assumption is that if the program is effective, the ratio of the publication to the number of faculty will be high.

myd : Median time lapse from entering graduate school to receipt of Ph.D. in years. This is a distributed median with multiple degrees awarded in the median year proportioned over the year. (it is important to me!) The program should let Ph.D. students graduate sooner (lets the program save money and be productive.) if it is effective enough.

suppfac : Percentage of program faculty with research support in the period 1988 to 1992. The quality and effectiveness of the program depends on the institutional and external research support.

fac_stu : And lastly, I created the variable of faculty-student ratio (**fac_stu**) using the total number of faculty divide by the total number of students. gen fac_stu = totstu/totfac

```
. reg rate93e pub_fac myd fac_stu suppfac;

      Source |       SS       df       MS              Number of obs =     109
-------------+------------------------------           F(  4,   104) =   28.09
       Model |  36.9698165     4  9.24245412           Prob > F      =  0.0000
    Residual |  34.2244696   104  .329081438           R-squared     =  0.5193
-------------+------------------------------           Adj R-squared =  0.5008
       Total |   71.194286   108  .659206352           Root MSE      =  .57366

------------------------------------------------------------------------------
     rate93e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     pub_fac |   .0546024   .0173104     3.15   0.002      .0202753    .0889295
         myd |  -.1197654   .0336794    -3.56   0.001     -.1865528   -.0529779
     fac_stu |   .0671608   .0320218     2.10   0.038      .0036604    .1306611
     suppfac |   .0189985    .003511     5.41   0.000      .0120362    .0259609
       _cons |   2.413536   .3219996     7.50   0.000      1.774999    3.052073
------------------------------------------------------------------------------
```

Of course, you should always look at the bivariate graphs. These are attached as follows:

File Name	Graph of:
PSet4-C3-pub_fac	rate93e and pub_fac
PSet4-C3-myd	rate93e and myd
PSet4-C3-fac_stu	rate93e and fac_stu
PSet4-C3-suppfac	rate93e and suppfac

pub_fac and myd each have one massive outlier, probably due to input error. Once omitted, you get the following graphs and regression:

File Name	Graph of:
PSet4-C3-NEWpub_fac	rate93e and pub_fac w/ outlier omitted
PSet4-C3-NEWmyd	rate93e and myd w/ outlier omitted

. reg rate93e pub_fac myd fac_stu suppfac

Source	SS	df	MS
Model	37.0885633	4	9.27214082
Residual	34.0812883	103	.330886295
Total	71.1698516	107	.6651388

Number of obs	=	108
F(4, 103)	=	28.02
Prob > F	=	0.0000
R-squared	=	0.5211
Adj R-squared	=	0.5025
Root MSE	=	.57523

rate93e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pub_fac	**.0706044**	.0298839	2.36	**0.020**	.0113367	.1298721
myd	**-.0974805**	.047835	-2.04	**0.044**	-.1923499	-.0026111
fac_stu	.059939	.0339344	1.77	0.080	-.0073618	.1272399
suppfac	**.0180133**	.0038259	4.71	**0.000**	.0104256	.0256011
_cons	2.2304	.4263319	5.23	0.000	1.384872	3.075929

As the ratio of publication to the number of faculty and the percentage of program faculty with research support grows, the effectiveness of the program increases at statistically significant levels. In addition, as the median time spent by Ph.D. student in the program increases, the effectiveness of the program declines, which implies that a more effective program lets students graduate sooner. You notice that the faculty student ration is no longer significant at the .05 level once the outliers are omitted. In addition, the outliers caused over estimation in myd and underestimation in pub_fac which would be important from a policy perspective.

Looking at the graph with rate93e and fac_stu, you notice an outlier which is probably not an input error (at least it is not obviously an error). Taking the natural log to create lnfac_stu yields a much more linear looking relationship. There is a non-linearity in the graph of rate93e and pub_fac also which is linearized nicely with the natural log (see graphs). Regressing with lnfac_stu and lnpub_fac you get:

File Name	Graph of:
PSet4-C3-lnpub_fac	rate93e and log of pub_fac
PSet4-C3-lnfac_stu	rate93e and log of fac_stu

```
. reg rate93e lnpub_fac myd lnfac_stu suppfac

      Source |       SS       df       MS              Number of obs =     106
-------------+------------------------------           F(  4,   101) =   31.01
       Model |  33.8241391     4  8.45603476           Prob > F      =  0.0000
    Residual |  27.5428644   101  .272701628           R-squared     =  0.5512
-------------+------------------------------           Adj R-squared =  0.5334
       Total |  61.3670034   105  .584447652           Root MSE      =  .52221

------------------------------------------------------------------------------
     rate93e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lnpub_fac |   .3852297   .1155941     3.33   0.001     .1559222    .6145372
         myd |  -.1151531   .0455439    -2.53   0.013    -.2054999   -.0248063
   lnfac_stu |   .2108884   .0871956     2.42   0.017     .0379158    .3838609
     suppfac |   .0116026   .0037141     3.12   0.002     .0042348    .0189704
       _cons |   2.499508   .4151781     6.02   0.000     1.675907     3.32311
------------------------------------------------------------------------------
```

This nets us an additional 3% of explanatory power in our R^2 and we now also have all of our variables with t-scores well above 2. The faculty-student ratio which we expected to be important now shows that it is.

Part C.3.

```
. reg rate93e lnpub_fac myd lnfac_stu suppfac, beta

      Source |       SS       df       MS              Number of obs =     106
-------------+------------------------------           F(  4,   101) =   31.01
       Model |  33.8241391     4  8.45603476           Prob > F      =  0.0000
    Residual |  27.5428644   101  .272701628           R-squared     =  0.5512
-------------+------------------------------           Adj R-squared =  0.5334
       Total |  61.3670034   105  .584447652           Root MSE      =  .52221

------------------------------------------------------------------------------
     rate93e |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
   lnpub_fac |   .3852297   .1155941     3.33   0.001                 .3409101
         myd |  -.1151531   .0455439    -2.53   0.013                -.1817654
   lnfac_stu |   .2108884   .0871956     2.42   0.017                 .1829051
     suppfac |   .0116026   .0037141     3.12   0.002                 .2903878
       _cons |   2.499508   .4151781     6.02   0.000                        .
------------------------------------------------------------------------------
```

Using the beta command, I created standardized coefficients. Standardized coefficients simply rescale the variables into standard deviations from the mean, which results in unitless coefficients. This enables us to compare the variables by their relative effects. In this case, the faculty publication record is the most significant on the effectiveness of the program, followed closely by faculty with research support.