Problem Set 3 Solutions

Chapter 9.

4. **answer: just about the same**. Think about two scatterplots that have been superimposed and shifted a bit. Move female scatterplot toward southwest (lower height and weight) to reflect the differences in average weights and heights. It turns out the shift is roughly along the "standard deviation line." Therefore, the resulting new scatterplot should has roughly the same correlation as the two earlier ones.

7. Again, this is an ecological correlation based on percentage, which usually generates an overestimated correlation coefficient. In addition, the correlation coefficient does not necessarily means causation. Certainly, there is a somewhat high association between the percentage population of native-born and votes received by Johnson. However, there might be some other factors which links the native-born and Johnson's votes, such as socioeconomic status. In addition, there is a possible aggregation fallacy. If the fraction of voters in a county is a very small and particular group of the whole population, then the voters are different people compared to the population.

Chapter 10.

1. A – (i)
   B – (iii)
   C – (ii)

3. average height of husbands = 68 SD of husbands = 2.7
   averageheight of wives = 63 SD of wives = 2.5 r = .25

$$r = \frac{\text{Cov(weight, height)}}{SD_w \text{ x } SD_h} = .25$$

$$\exists_w = r \text{ x } \frac{SD_w}{SD_h} = \frac{.25\text{x}2.5}{2.7}$$

(a) $\frac{.25\text{x}2.5}{2.7}$ x 4 ≈ 1, predicted height of a wife = 64

(b) $\frac{.25\text{x}2.5}{2.7}$ x -4 ≈ -1, predicted height of a wife = 62

(c) $\frac{.25\text{x}2.5}{2.7}$ x 0 = 0, predicted height of a wife = 63

(d) we do not have any information about husband's height, therefore, have to predict the wife's height is an average, which is 63.

7. Both doctors are wrong. This question is about the regression fallacy and regression effect. When the first measurent is too high or too low, the second measurement tends to regress toward mean.

Chapter 11.

1. (v) $\sqrt{(1-r^2)}$ x SD of y

2. Yes, something is wrong. $\sqrt{(1-r^2)}$ = 3.12, and 2xr.m.s. = 6.24 which covers 95% of the data. Even if the average is 0, ± 6.24 is way to above and below. Since GPA scale is usually 4.0, it does not make sense the highest value of the data can be 6.24.

6. NO. A correlation is not causation. We can conclude that a student who does homework tends to have better GPA, because he/she probably is studious. However, we cannot assert that doing homework makes the student's GPA higher.

9. NO. This is a regression effect again. Rookie of the year is the outstanding player of the year, which implies that he is a high outlier case. In the second year, he supposedly regresses toward average level.

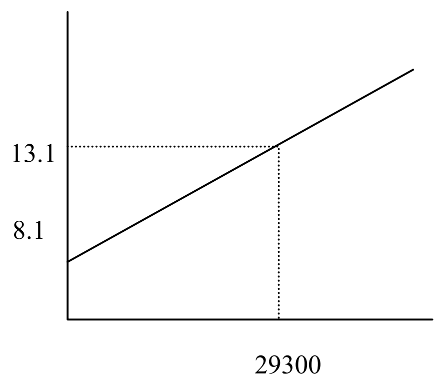
Chapter 12.

4. (a) **about 1.** The line represents the average of the data. All the data points are located between zero and four, thus we can guess the SD of y is 1. (remember, 2x SD usually covers 95 percentage of data.) And SD of y is the r.m.s. for predicting y by its average. Therefore, it should be around 1.
   (b) NO. The regression line seems to be left-downward.

11. $\text{Slope} = r \times \dfrac{SD_y}{SD_x} = .0000617$

$\therefore\ .37 \times \dfrac{SD_y}{SD_x} = .0000617$

When x = 0, y = 8.1 years.
8.1 = 13.1 - y, y = 5.
5 = 29300 x yr/$
yr/$ = 5/29300 = .00017065, which is different from the slope coefficient in the equation. (.0000617)
However, we cannot figure it out until we know SDs.



Part II.

A.1.

```
. reg  abortion  attend

      Source |       SS       df       MS              Number of obs =    1748
-------------+------------------------------           F(  1,  1746) =  301.17
       Model |  309.274994     1  309.274994           Prob > F      =  0.0000
    Residual |  1793.01334  1746  1.02692631           R-squared     =  0.1471
-------------+------------------------------           Adj R-squared =  0.1466
       Total |  2102.28833  1747  1.20337054           Root MSE      =  1.0134

------------------------------------------------------------------------------
    abortion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |   .2362405   .0136129    17.35   0.000     .2095411    .2629398
       _cons |   2.166483   .0476642    45.45   0.000     2.072998    2.259968
------------------------------------------------------------------------------

. predict yhat1
```

This regression shows a positive and statistically significant effect of church attendance on abortion law liberalism. Note the reverse coding in both variables. The regression explains only about 14% of the variance here, so there are a lot of other factors going on or a transformation is needed.

A.2.

```
. gen attend2=attend^2
(18 missing values generated)

. reg abortion  attend attend2

      Source |       SS       df       MS              Number of obs =    1748
-------------+------------------------------           F(  2,  1745) =  158.69
       Model |  323.516155     2  161.758077           Prob > F      =  0.0000
    Residual |  1778.77217  1745  1.01935368           R-squared     =  0.1539
-------------+------------------------------           Adj R-squared =  0.1529
       Total |  2102.28833  1747  1.20337054           Root MSE      =  1.0096

------------------------------------------------------------------------------
    abortion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |   .4361105   .0551665     7.91   0.000     .3279111      .54431
     attend2 |  -.0369555   .0098871    -3.74   0.000    -.0563473   -.0175637
       _cons |   2.016965   .0620911    32.48   0.000     1.895184    2.138745
------------------------------------------------------------------------------

. predict yhat2
```

The data looked a little curved to me, so I threw in a polynomial transformation. These are a little harder to interpret, which is why the graph in A.3. is important. The regression is again highly significant and positive in the universe of possible values for attend (see A.3.), but the transformed model only explains a little bit more of the variance.

A.3.

```
. graph  abortion yhat1 yhat2 attend, connect (.ss) symbol (Oii) jitter(3)
see attached graph PSet3GraphA3
```

As you can see from the graph, the linear model underpredicts the effect for most of the range of attendance and overpredicts at the extremes.

B.1.

```
. gen lnbooks=ln(books)

. reg reading  lnbooks

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    8.41
       Model |  639.043888     1  639.043888           Prob > F      =  0.0062
    Residual |  2886.05611    38   75.948845           R-squared     =  0.1813
-------------+------------------------------           Adj R-squared =  0.1597
       Total |     3525.10    39  90.3871795           Root MSE      =  8.7149

------------------------------------------------------------------------------
     reading |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnbooks |   9.226154   3.180651     2.90   0.006     2.787264    15.66504
       _cons |   139.6288   24.59164     5.68   0.000     89.84563     189.412
------------------------------------------------------------------------------
```

I chose to take the log of books in order to mute the effect of some outliers like Iowa and Kansas. You can see the difference by looking at attached graphs PSet3GraphB1 and PSet3GraphB1ln.

B.2.

```
. predict yhat

. gen resid= reading-yhat
```

see attached graph PSet3GraphB2

Also try:
. rvfplot, s([state]) yline(0)

The residual plot should look like noise around a straight line, but you can see some curve to it between 7.8 and 8. There could be some heteroskedasticity here (correlation between the independent variable and the residuals), but more important is the presence of notable outliers, especially Washington DC which is very poorly predicted. Also Hawaii and Connecticut raise some concern.

B.3.

The Regression Equation above is:
reading = ln(books)*9.226 + 139.629

Since a regression line always runs through the point (X_{avg}, Y_{avg}) the STATA calculated value of X_{avg} (the average of lnbooks) can be used rather than solving for lnbooks. It is 7.780 which translates to 2392.275 books.

To calculate the number of books which generates an increase of 5 points in reading we can solve by subtracting from the number of books needed to create a reading score of 215.85 from the mean number of books.

Solving for books:
215.85 = ln(books)*9.226 + 139.629
76.221 = ln(books)*9.226
76.221/9.226 = ln(books)
8.262 = ln(books)
$e^{8.262}$ = books
3873.834 = books

So to increase reading scores by 5, we need an additional (3873.834 – 2392.275) 1481.559 books per hundred students.

C.1.

```
. gen lnchaspend = ln(chaspend)
(7 missing values generated)

. reg  incvote lnchaspend

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  1,    24) =    6.13
       Model |  4.0910e+12     1  4.0910e+12           Prob > F      =  0.0207
    Residual |  1.6010e+13    24  6.6708e+11           R-squared     =  0.2035
-------------+------------------------------           Adj R-squared =  0.1703
       Total |  2.0101e+13    25  8.0403e+11           Root MSE      =  8.2e+05

------------------------------------------------------------------------------
     incvote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lnchaspend |   183283.2   74010.97     2.48   0.021     30532.07    336034.3
       _cons |   -1501158    1006837    -1.49   0.149     -3579167    576850.1
------------------------------------------------------------------------------
```

Again, I chose to use a natural log of the independent variable to linearize it. When using income or spending, the natural log is a commonly used transformation. In this case there was not really a big advantage gained by doing so (see graphs PSet3GraphC1 and PSet3GraphC1ln). The resulting regression seems to indicate the counter intuitive result that challenger spending has a significant and POSITIVE effect on incumbent votes. What is working against us here is probably large states. Bigger states require candidates to spend more, so the increase in overall spending is being related to the number of possible votes more than the effect we are trying to measure. Also bear in mind that incumbent spending is not being controlled for.

C.2.

```
. gen twoptyvote=incvote/(incvote+chavote)
(4 missing values generated)

. reg  twoptyvote lnchaspend

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  1,    24) =   45.84
       Model |   .13145517     1   .13145517           Prob > F      =  0.0000
    Residual |   .06882913    24   .00286788           R-squared     =  0.6563
-------------+------------------------------           Adj R-squared =  0.6420
       Total |    .2002843    25  .008011372           Root MSE      =  .05355

------------------------------------------------------------------------------
  twoptyvote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lnchaspend |  -.0328547   .0048528    -6.77   0.000    -.0428703   -.0228391
       _cons |    1.05211   .0660163    15.94   0.000      .915859    1.188361
------------------------------------------------------------------------------
```

This looks much better. Not only is the coefficient significant, but it’s sign is now in the expected direction. Using a percentage vote to mute the effect of state size as a confounding variable, we are also modeling the data better. This second model explains about 66% of the variance compared to 20% in the first.

C.3.

Clearly, the correct model to use is the second which implies that the more money a challenger spends, the lower an incumbent's vote share and therefore the incumbent’s probability of victory.