Introductory Stats

Prof. Green

Solution Set: Supplementary Homework 8

1. Produce five number summaries of the four variables: gnp92, gnp73, im92, im73. Comment on the apparent trends in infant mortality and GNP.

Variable	N	N*	Mean	Median	TrMean	StDev
gnp92	152	55	5177	1430	4120	8006
im92	194	13	47.06	32.00	44.03	41.11
gnp73	140	67	1371	530	1095	1988
im73	134	73	90.85	85.50	88.81	61.28

Variable	SE Mean	Minimum	Maximum	Q1	Q3
gnp92	649	60	36080	548	5208
im92	2.95	4.50	162.10	14.00	68.00
gnp73	168	60	12050	235	1605
im73	5.29	10.00	229.00	30.75	142.75

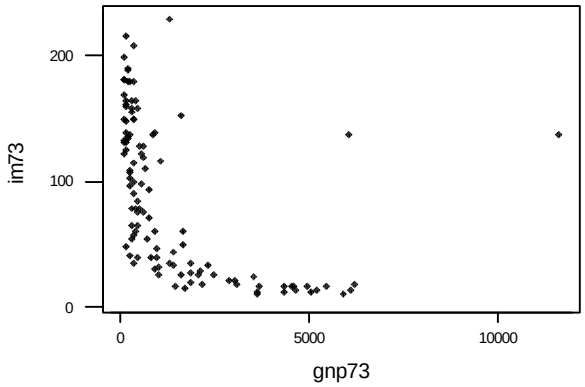
2. Notice that the summaries produced in problem #1 are based on different set of countries, because not all of the countries supply data for one or more variable. To get around this problem, click on Manip > Subset Worksheet. Click on 'Condition' and specify that you want only those rows for which 'gnp92 > 0 and gnp73 > 0 and im92 > 0 and im73 > 0.' This will eliminate so-called 'missing data.' Now print out five number summaries and interpret.

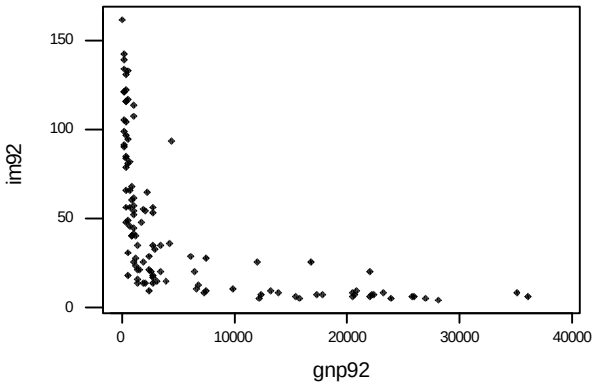
Variable	N	Mean	Median	TrMean	StDev	SE Mean
gnp92	112	6253	1840	5212	8825	834
im92	112	46.56	32.00	43.39	41.14	3.89
gnp73	112	1446	555	1201	1917	181
im73	112	87.93	78.00	85.47	61.42	5.80

Variable	Minimum	Maximum	Q1	Q3
gnp92	60	36080	550	7495
im92	4.50	162.10	11.45	67.50
gnp73	70	11630	250	1845
im73	10.00	229.00	27.25	138.00

3. Examine the degree to which infant mortality varies as a function of per capita GNP. Begin by plotting gnp92 against im92 (make sure you draw the graph so that im92 is the dependent variable). Repeat this exercise for these variables as measured in 1973. Do the two graphs look similar? Is there an apparent relationship between GNP and infant mortality (IM)? What are the correlations between these two pairs of variables? Is the GNP-IM relationship a linear relationship?

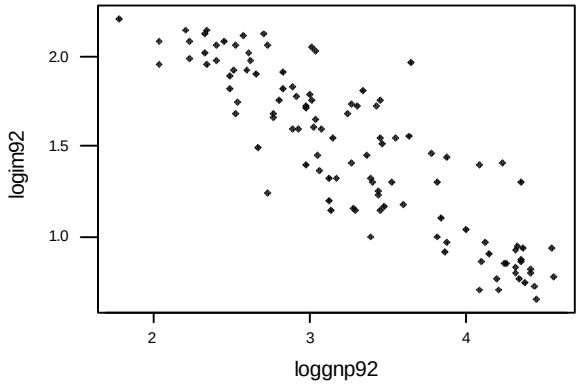
	gnp92	im92	gnp73
im92	**-0.571**		
gnp73	0.865	-0.528	
im73	-0.592	0.854	**-0.491**





The relationship is consistent across years (save for some OPEC outliers in 1973). The relationship is not linear.

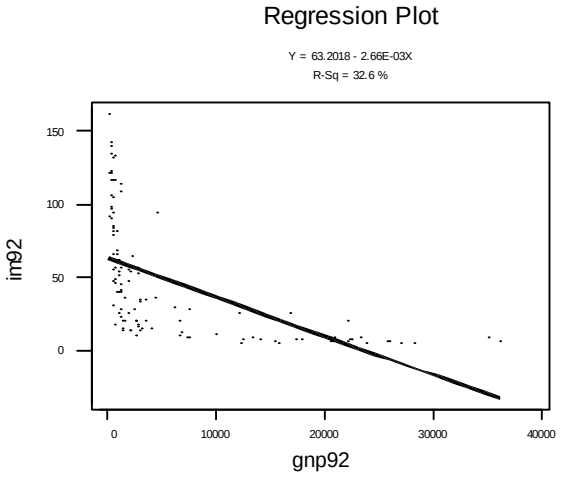
4. One trick to straighten out nonlinear relationships between variables is to use logarithms. (Note: ordinarily one would use natural logs, but to maintain comparability with some of the built-in commands below, use base-10 logs.) Use the Calc > Calculator menu to generate the base-10 log of GNP and IM in each year. Redo the plots in problem #3. Are the transformed variables linearly related? What are the correlations between these transformed variables?

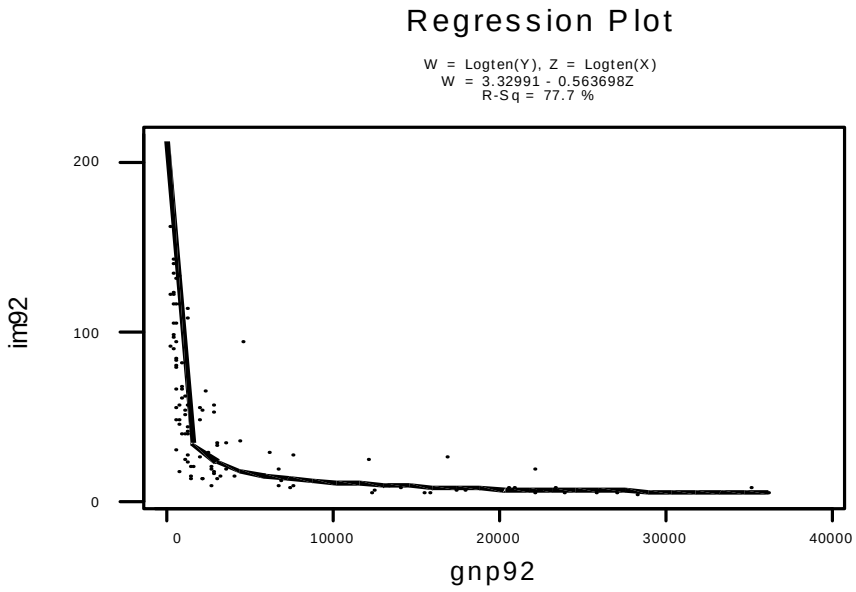


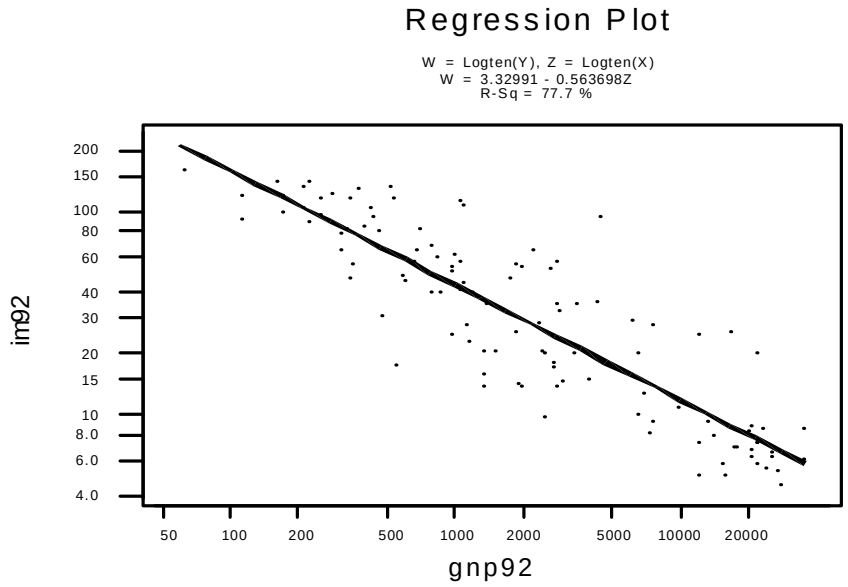

```
Correlation of loggnp92 and logim92 = -0.881, P-Value = 0.000
```

This correlation is much higher than when the variables were in raw form.

5. Click on Regression > Fitted Line Plot and regress IM92 (the dependent variable) on GNP92. Now repeat this exercise by selecting Options and clicking LOGTEN transformations of X and Y. Try this with and without the scaling options to see how the resulting graphs differ. How do the R-squared's change as a result of the log-transformation? What does the R-squared mean?







The R-squared improves dramatically, to 77.7%. This means that 77.7% of the variance in Y is predicted by the independent variable, log-gnp92.

6. Having determined that the log transformations work well, go back to the Regression menu and select Regression. Regress LOGIM92 on LOGGNP92. Interpret the following output: slope, constant, s, and standard error of the slope (under the heading labeled standard deviation). How is the t-statistic of the slope calculated?

```
The regression equation is logim92 = 3.33 - 0.564 loggnp92

Predictor        Coef       StDev          T        P
Constant      3.32991     0.09731      34.22    0.000
loggnp92     -0.56370     0.02879     -19.58    0.000

S = 0.2139      R-Sq = 77.7%     R-Sq(adj) = 77.5%
```

Slope: the effect of a one unit change in the independent variable (log of GNP92) on the expected value of the dependent variable (Log of IM92). Here, a one unit increase in LOGGNP92 decreases LOGIM by .56 units.

Constant: the expected value of LOGIM92 when LOGGNP=0. LOGGNP = 0 when GNP is $1 per capita. Note that the antilog of 3.32 is greater than 1000: when a country is that poor, no babies survive (i.e., the infant mortality rate is over 1000 per 1000 births).

S: The estimated standard deviation of the disturbance distribution. Note that this statistic is sometimes called the 'standard error of the regression' or the 'standard error of estimate.'

Standard error of slope: the standard deviation of the sampling distribution of the slope over hypothetical replications of the analysis.

T: ratio of the coefficient to its standard error.

7. Using the menu Basic Stats > Covariance, generate the variances and covariance for LOGIM92 and LOGGNP92. Show that the slope estimate from your regression equals the covariance divided by the variance of the independent variable (in this case LOGGNP92). Show that the estimated intercept can by calculated by subtracting the slope times the mean of X (LOGGNP92) from the mean of Y (LOGIM92). Show that the R-square is the covariance squared divided by the product of the two variances.

```
Variable             N       Mean     Median     TrMean      StDev    SE Mean
loggnp92           112     3.3056     3.2648     3.3089     0.7052     0.0666
logim92            112     1.4666     1.5049     1.4709     0.4510     0.0426

          loggnp92   logim92
loggnp92  0.497272
logim92  -0.280311  0.203364
```

-.56 = -.28/.497.
3.33= 1.47 - -.56*3.31
.777=.28*.28/(.50*.20)

8. What countries are 'outliers' or unusual observations?

```
Obs   loggnp92    logim92        Fit   StDev Fit    Residual    St Resid
 34 gabon 3.65     1.9731     1.2733      0.0225      0.6998        3.29R
 84 qatar 4.22     1.4150     0.9488      0.0333      0.4661        2.21R
 94 sri lk2.73     1.2455     1.7897      0.0261     -0.5441       -2.56R
```

9. Recall that several countries do not report their infant mortality rates in 1973. Pick 3 such countries that do report GNP73 and predict their 1973 infant mortality rates. For the sake of practice, do this by hand based on a regression of IM73 on GNP73. To check your work, click on Regression > Regression > Options and insert a value of LOGGNP73 on which to base your predictions. Just for fun, request a "prediction interval" (PI) which gives a 95% range into which that observation is likely to fall.

```
Predictor        Coef       StDev          T        P
Constant       3.3530      0.1213      27.64    0.000
loggnp73     -0.55225     0.04222     -13.08    0.000

S = 0.2497      R-Sq = 60.9%     R-Sq(adj) = 60.5%


     Fit  StDev Fit         95.0% CI             95.0% PI
  1.8398     0.0238   (  1.7926,  1.8871)  (  1.3427,  2.3370) mongolia
  1.6521     0.0260   (  1.6005,  1.7037)  (  1.1545,  2.1496) reunion
  2.1381     0.0352   (  2.0683,  2.2078)  (  1.6383,  2.6378) vietnam

Exponentiating these results give IM rates of 68, 45, 135.
```

10. Find a country that reported IM for 1992 but not 1973. Can a 'backwards' regression of LOGIM73 on LOGIM92 be used to estimate what that country's infant mortality rate was in 1973? Try it.

```
The regression equation is logim73 = 0.615 + 0.806 logim92

Predictor        Coef       StDev          T        P
Constant      0.61492     0.05222      11.78    0.000
logim92       0.80562     0.03405      23.66    0.000

S = 0.1618      R-Sq = 83.6%     R-Sq(adj) = 83.4%

Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
  2.1134     0.0203   (  2.0731,  2.1537)  (  1.7903,  2.4365)
```

The predicted value for Myanmar (IM92=72) is 10 to the 2.11, or 129. This model, although nonsensical from the standpoint of causality, provides a high degree of predictive accuracy.