NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

IOMS Department

Statistics and Data Analysis

Professor William Greene
Phone: 212.998.0876
Office: KMC 7-78
Home page: www.stern.nyu.edu/~wgreene
Email: wgreene@stern.nyu.edu
Course web page: www.stern.nyu.edu/~wgreene/Statistics/Outline.htm

Assignment 4
Linear Regression Model

Part I. Law of Large Numbers.

1. In Notes (slides) 10, we looked at the idea that in estimating a mean of a population, a larger sample is better than a small one. “Better” is quantified in the idea of the “standard error of the mean,” which is computed as $\sigma/\sqrt{n}$ for a sample of n observations. A useful question to consider is “how much better.” Suppose I have drawn a sample of 10,000 observations on the number of minutes that arriving flights are late at airports around the world. I find that the sample mean is 24.75 and the sample standard deviation is 9.32. What is the estimate of the standard error of the mean? Now, the question. How much better would you say a sample of 100,000 observations would be?

The estimated standard error of the mean is $9.32/\sqrt{10{,}000} = 9.32/100 = 0.0932$. If the sample were 100,000 instead, then the standard error of the mean would be $9.32/\sqrt{100{,}000} = 9.32/316.228 = 0.02947$. So, for a sample 10 times as large and 10 times as expensive, the improvement is only about a factor of 3.2. This questions the virtue of having a huge sample when you already have a precise estimate of a parameter, such as the mean.

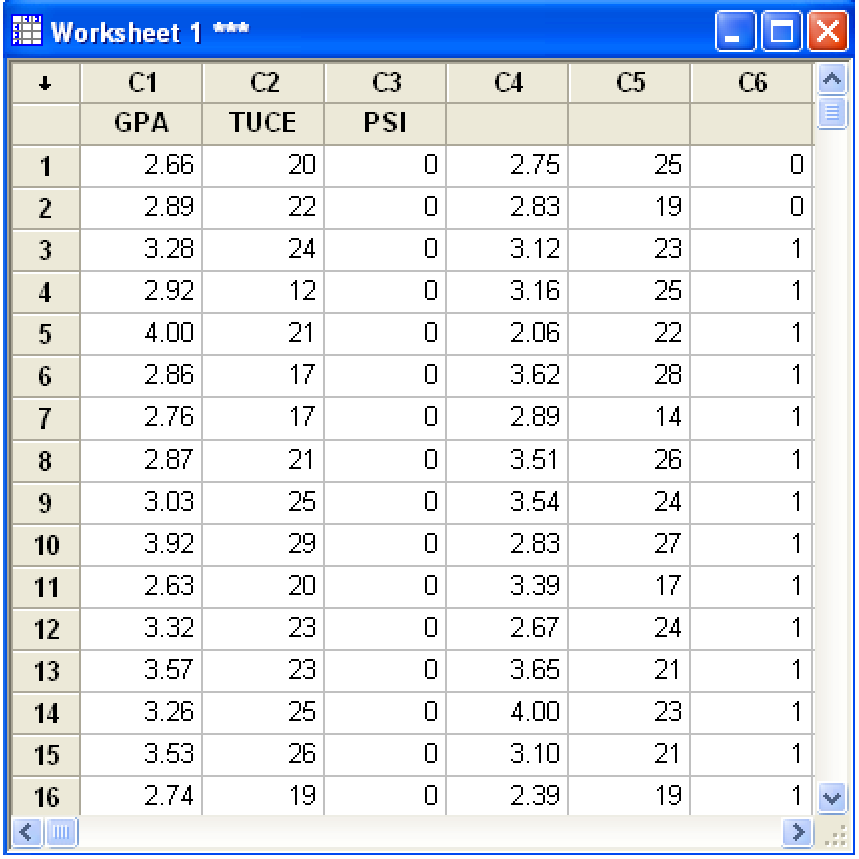
2. We have found (and will continue to find) many uses for the empirical rule: 95% of almost any distribution will lie within two standard deviations of the mean. One of the ways we use this result is to form a range of estimates around an estimate of the mean of a population that we can feel accounts for the uncertainty (sampling variability) of that estimator. For the results in problem 1 above, what would you report as your range of estimates for the average number of minutes late for flights assuming that the sample used is 10,000 flights?

The mean plus and minus two standard errors of the mean is 24.75 plus and minus 2*0.0932 which is 24.5636 to 24.9364.

Part II. Linear Regression Analysis

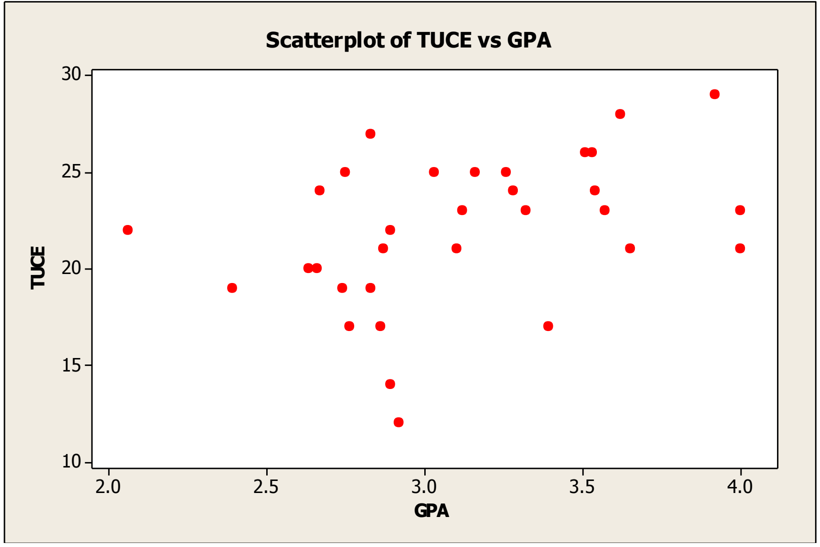
3. The data contained in the data file EconGrades.mpj (it is on the course website), is a sample of 32 observations on high school students. (The data were examined in a study in the Journal of Economic Education.) The variables in the data set are

GPA = the student’s grade point average
TUCE = the student’s grade on a test of economic literacy
PSI = 1 if the student participated in a special economics course, 0 if not.



↓	C1	C2	C3	C4	C5	C6
	GPA	TUCE	PSI			
1	2.66	20	0	2.75	25	0
2	2.89	22	0	2.83	19	0
3	3.28	24	0	3.12	23	1
4	2.92	12	0	3.16	25	1
5	4.00	21	0	2.06	22	1
6	2.86	17	0	3.62	28	1
7	2.76	17	0	2.89	14	1
8	2.87	21	0	3.51	26	1
9	3.03	25	0	3.54	24	1
10	3.92	29	0	2.83	27	1
11	2.63	20	0	3.39	17	1
12	3.32	23	0	2.67	24	1
13	3.57	23	0	3.65	21	1
14	3.26	25	0	4.00	23	1
15	3.53	26	0	3.10	21	1
16	2.74	19	0	2.39	19	1

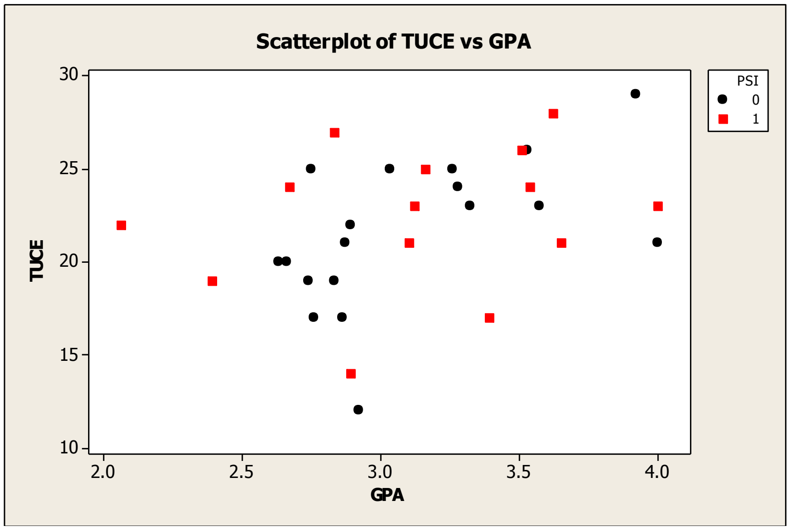
a. Construct a scatter plot of **TUCE** (on the vertical axis) against **GPA** on the horizontal axis.



b. Does the scatter plot suggest that there is a relationship between **GPA** and **TUCE**? Describe it if your answer is yes.

The graph doesn't suggest much. If there is any relationship, is is slightly positive.

c. Now, produce a scatter plot that separates the two groups. [Graph → Scatter Plot → With Groups: Then specify the variables and **PSI** as the categorical variable for the grouping.] Does the plot suggest that there is a different relationship for the two groups? (We will pursue this below.)



There appears to be a more discernible positive relationship, more so for those students who did not have the PSI.

4. For the n=32 observations in problem 3, the basic statistics are:

$$\begin{aligned}
\overline{\text{GPA}} &= \tfrac{1}{n}\Sigma_{i=1}^{n}\text{GPA}_i &&= 3.1172\\
\overline{\text{TUCE}} &= \tfrac{1}{n}\Sigma_{i=1}^{n}\text{TUCE}_i &&= 21.938\\
S^2_{\text{GPA}} &= \tfrac{1}{n-1}\Sigma_{i=1}^{n}(\text{GPA}_i - \overline{\text{GPA}})^2 &&= 0.217821\\
S^2_{\text{TUCE}} &= \tfrac{1}{n-1}\Sigma_{i=1}^{n}(\text{TUCE}_i - \overline{\text{TUCE}})^2 &&= 15.221774\\
S_{\text{GPA,TUCE}} &= \tfrac{1}{n-1}\Sigma_{i=1}^{n}(\text{GPA}_i - \overline{\text{GPA}})(\text{TUCE}_i - \overline{\text{TUCE}}) &&= 0.704657
\end{aligned}$$

a. Compute the correlation coefficient between GPA and TUCE. Interpret the result.

The correlation between GPA and TUCE is

$$r_{\text{GPA,TUCE}} = \frac{S_{\text{GPA,TUCE}}}{\sqrt{S^2_{\text{GPA}}S^2_{\text{TUCE}}}} = \frac{0.704657}{\sqrt{0.217821 \times 15.221774}} = 0.386986$$

b. Compute the constant term, a, and the slope, b, in the linear least squares regression of TUCE (the dependent variable) on GPA (the independent variable). The result should help to confirm the conclusion you drew in part 3.b. above.

$$b = \frac{S_{\text{GPA,TUCE}}}{S^2_{\text{GPA}}} = \frac{0.704657}{0.217821} = 3.235027$$

$$a = \overline{\text{TUCE}} - b\overline{\text{GPA}} = 21.938 - 3.235027(3.1172) = 11.8538$$

c. Use the Stat → Regress feature in Minitab to confirm the computations in part b.

Regression Analysis: TUCE versus GPA

```
The regression equation is
TUCE = 11.9 + 3.24 GPA


Predictor     Coef  SE Coef      T      P
Constant    11.853    4.434   2.67  0.012
GPA          3.235    1.407   2.30  0.029
S = 3.65699   R-Sq = 15.0%   R-Sq(adj) = 12.1%
```

d. Compute the residual standard deviation, S_e using the statistics given above.

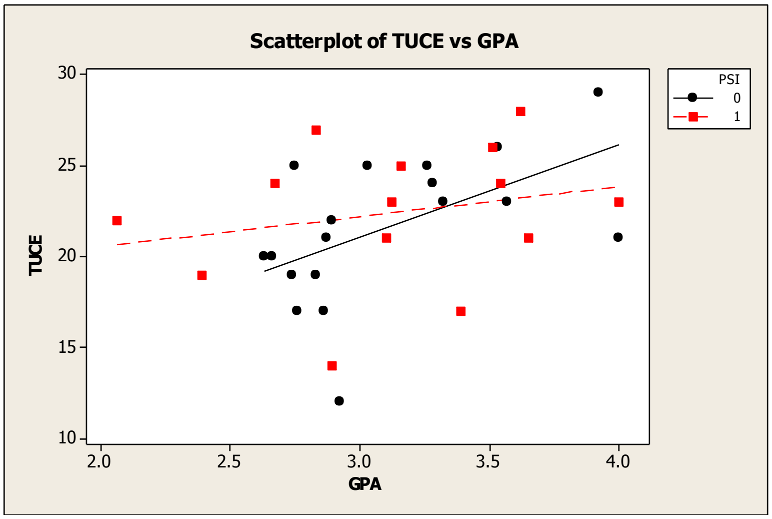
The residual standard deviation is computed as

$$s_e = \sqrt{\frac{(n-1)[S^2_{\text{TUCE}} - b^2 S^2_{\text{GPA}}]}{(n-2)}} = \sqrt{\frac{31[15.221774 - 3.235027^2(0.217821)]}{30}} = 3.656993$$

Note that this appears in the regression results that Minitab gave above.

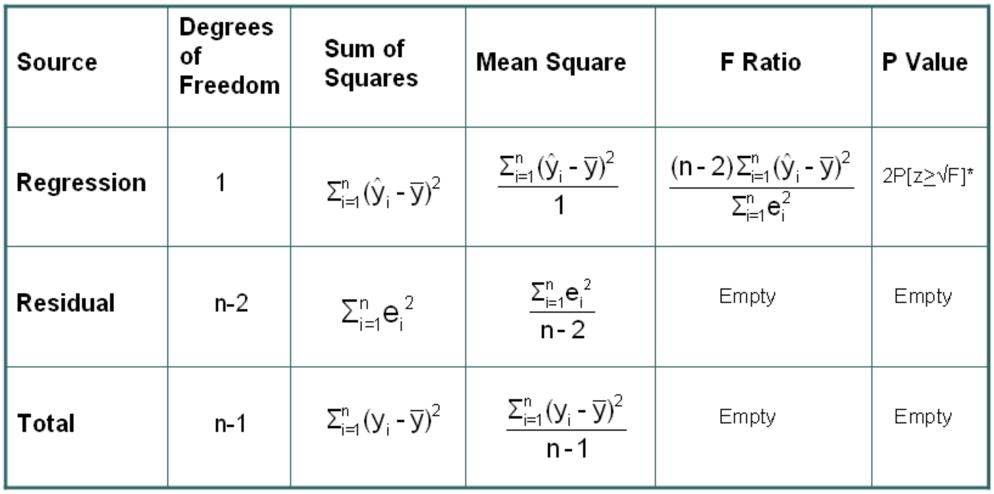
5. To pursue the question raised in part 3.c. earlier, produce a scatter plot that contains the two regressions in it for the two groups of students. Does this enable you to reach a conclusion about the difference in the relationship between GPA and TUCE for the two groups of students?

Using Graph→Scatter Plor → With Regression and Groups, then TUCE on GPA with groups defined by PSI, I get:



This does agree with the suggestion made earlier.

6. A set of basic statistics for the two variables GPA and TUCE in a sample of n = 32 students is computed in problem 4. One of the sets of results that would be produced from the regression would be an analysis of variance table, containing the results shown below. (This is from class notes 14).



Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	P Value
Regression	1	$\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{1}$	$\frac{(n-2)\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\Sigma_{i=1}^{n}e_i^2}$	2P[z≥√F]*
Residual	n-2	$\Sigma_{i=1}^{n}e_i^{\,2}$	$\frac{\Sigma_{i=1}^{n}e_i^{\,2}}{n-2}$	Empty	Empty
Total	n-1	$\Sigma_{i=1}^{n}(y_i - \bar{y})^2$	$\frac{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$	Empty	Empty

Using the basic statistics given in problem 4, fill in the numbers for this analysis of variance table. (Hint: There is another very useful slide in your class notes.) Now that you have filled in the table, what is the R^2 in this regression?

$$\begin{aligned}
\overline{GPA} &= \tfrac{1}{n}\Sigma_{i=1}^{n}GPA_i &&= 3.1172\\
\overline{TUCE} &= \tfrac{1}{n}\Sigma_{i=1}^{n}TUCE_i &&= 21.938\\
S_{GPA}^2 &= \tfrac{1}{n-1}\Sigma_{i=1}^{n}(GPA_i - \overline{GPA})^2 &&= 0.217821\\
S_{TUCE}^2 &= \tfrac{1}{n-1}\Sigma_{i=1}^{n}(TUCE_i - \overline{TUCE})^2 &&= 15.221774\\
S_{GPA,TUCE} &= \tfrac{1}{n-1}\Sigma_{i=1}^{n}(GPA_i - \overline{GPA})(TUCE_i - \overline{TUCE}) &&= 0.704657
\end{aligned}$$

The values of the statistics in the table are

$$n-2=30$$

$$n-1=31$$

$$\Sigma_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)S_{TUCE}^2 = 31(15.221774) = 471.874994$$

$$\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (n-1)b^2S_{GPA}^2 = 31[3.235027^2(0.217821)] = 70.667099$$

$$\Sigma_{i=1}^{n}e_i^{\,2} = (n-1)[S_{TUCE}^2 - b^2S_{GPA}^2] = 31[15.221774 - 3.235027^2(0.217821)] = 401.207895$$

$$\text{where } b = \frac{S_{GPA,TUCE}}{S_{GPA}^2} = \frac{0.704657}{0.217821} = 3.235027$$

$$\frac{\Sigma_{i=1}^{n}e_i^{\,2}}{n-2} = \frac{401.207895}{30} = 13.373597$$

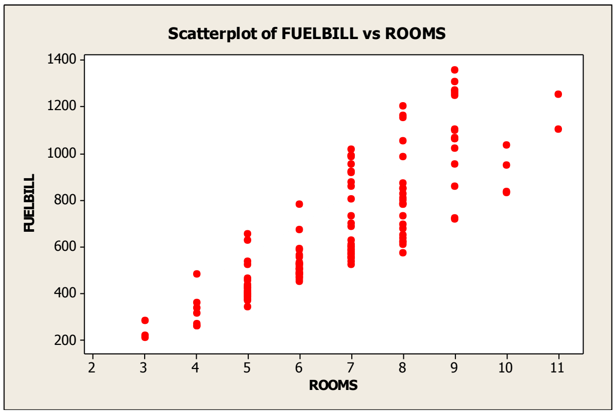
$$\frac{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = S_{TUCE}^2 = 15.221774$$

$$F = \frac{(n-2)\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\Sigma_{i=1}^{n}e_i^{\,2}} = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\left(\Sigma_{i=1}^{n}e_i^{\,2}\right)/(n-2)} = \frac{70.667099}{13.373597} = 5.284076$$

$$R^2 = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} = \frac{70.667099}{471.874994} = 0.149758$$

7. The file heating.mtp deals with the heating bill for dwelling units of various numbers of rooms.
(a) What are the names of the variables in this data set?
(b) Obtain a scatterplot of the two variables. Which variable should be on the horizontal axis?
(c) Find the linear regression equation resulting from the regression of FUELBILL on ROOMS.
(d) Examine the residual-fitted-plot from this regression. Is there a tendency for larger dwelling units to have more variable heating bills?

a. The file contains 147 observations on two variables, ROOMS and FUELBILL.
b. The scatter plot appears below. Logically, we would be predicting the fuel bill by the number of rooms, so the fuel bill should appear on the vertical axis.

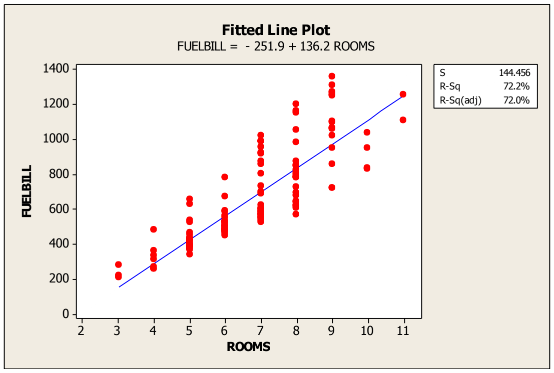


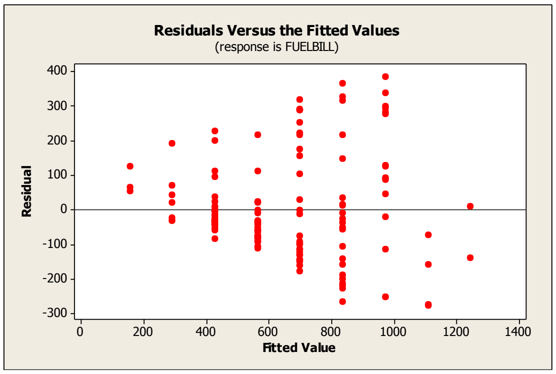
c. Letting Minitab do the work, we find

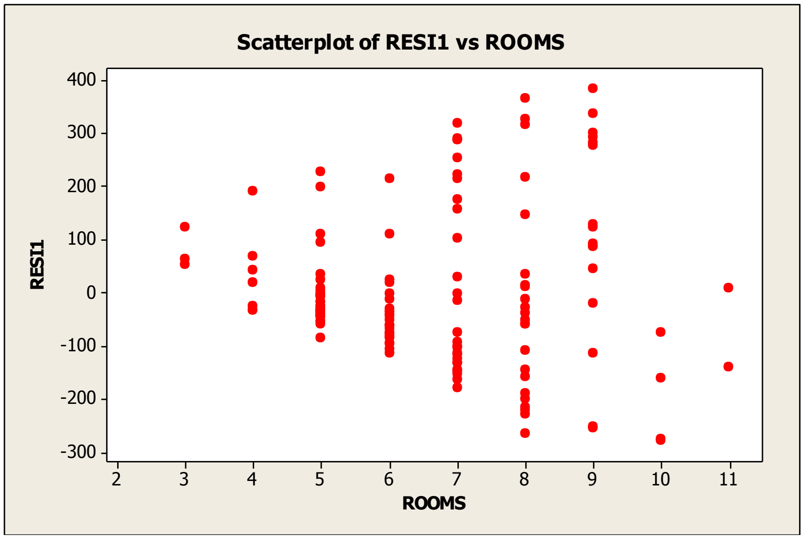
Regression Analysis: FUELBILL versus ROOMS

```
The regression equation is
FUELBILL = - 252 + 136 ROOMS
Predictor      Coef  SE Coef      T      P
Constant    -251.89    48.44  -5.20  0.000
ROOMS       136.169    7.098  19.18  0.000
S = 144.456   R-Sq = 72.2%   R-Sq(adj) = 72.0%
```

d. The regression and the residual plots appear below. It certainly does appear that the variation of the residuals increases as the number of rooms increases. (Note that because we are using only ROOMS as the predictor, the plot on the right, of the residuals against the fitted value, is a plot of the residuals against a + b*Rooms. This is essentially the same as the plot of the residuals against ROOMS, as the additional figure shows. Note, you can save the residuals by using Stat→Regression→Regression, then set the variables, and choose Storage, then check the Residuals choice under Diagnostic Measures. Then, make the scatter plot.

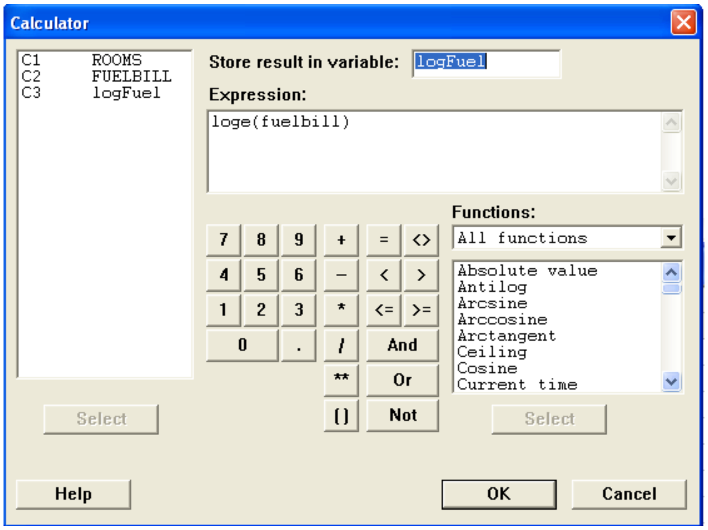






8. The increasing variation of the residuals as the number of rooms increases that you see in the residual-versus-fitted plot of the previous problem suggest that the constant variance assumption of the regression model is wrong for these data. A common corrective action is to replace the dependent variable, here FUELBILL, with its logarithm. Take this action and repeat the regression. Note the revised fitted model equation and give the new residual-versusfitted plot. Finally, can you compare the value of R^2 for the two versions?

Minitab HINT: Use **Calc → Calculator** and then set up the resulting panel as follows:



```
C1    ROOMS
C2    FUELBILL
C3    logFuel
```

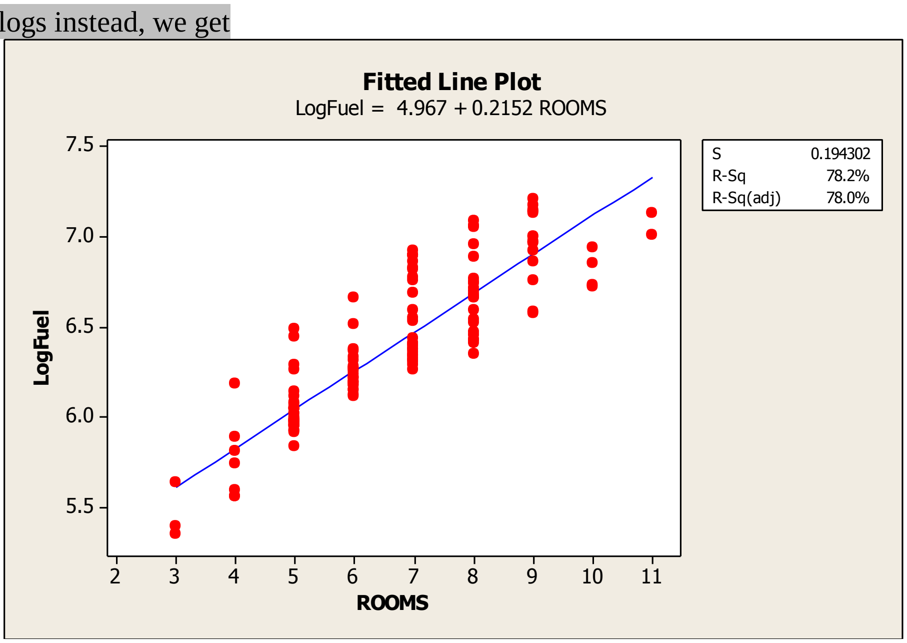
Store result in variable: logFuel

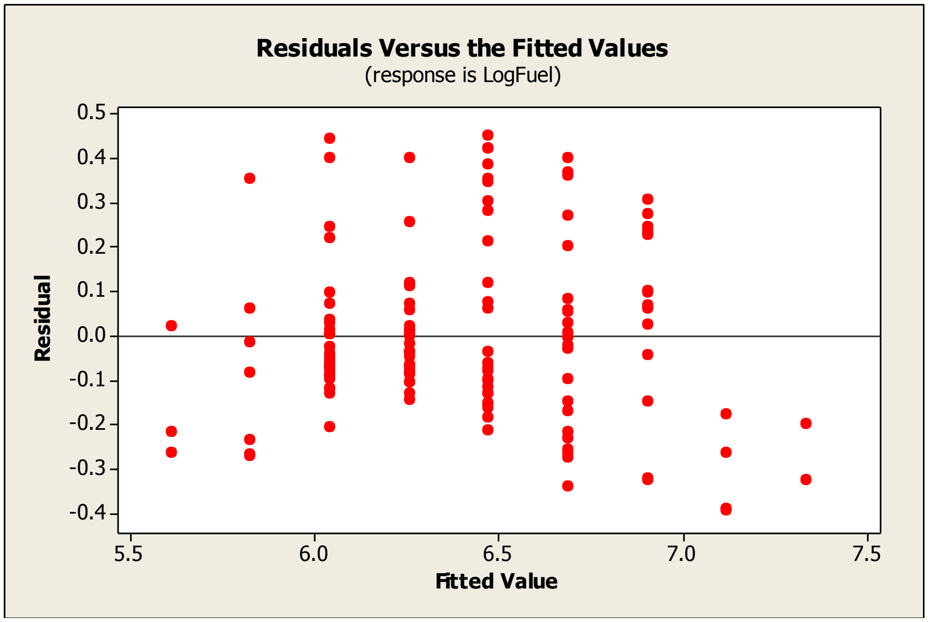
Expression:

```
loge(fuelbill)
```

In Minitab, the function name for natural logs is LOGE. In the **Functions** list, this is identified as "Natural log." (You can just type the "loge," you need not use the menu.) You get to select the name for **Store result in variable**. Here logFUEL seems a reasonable choice. The new variable will be placed in the next open column. Instead of giving a name to the new variable, you can select a column.

Using the logs instead, we get

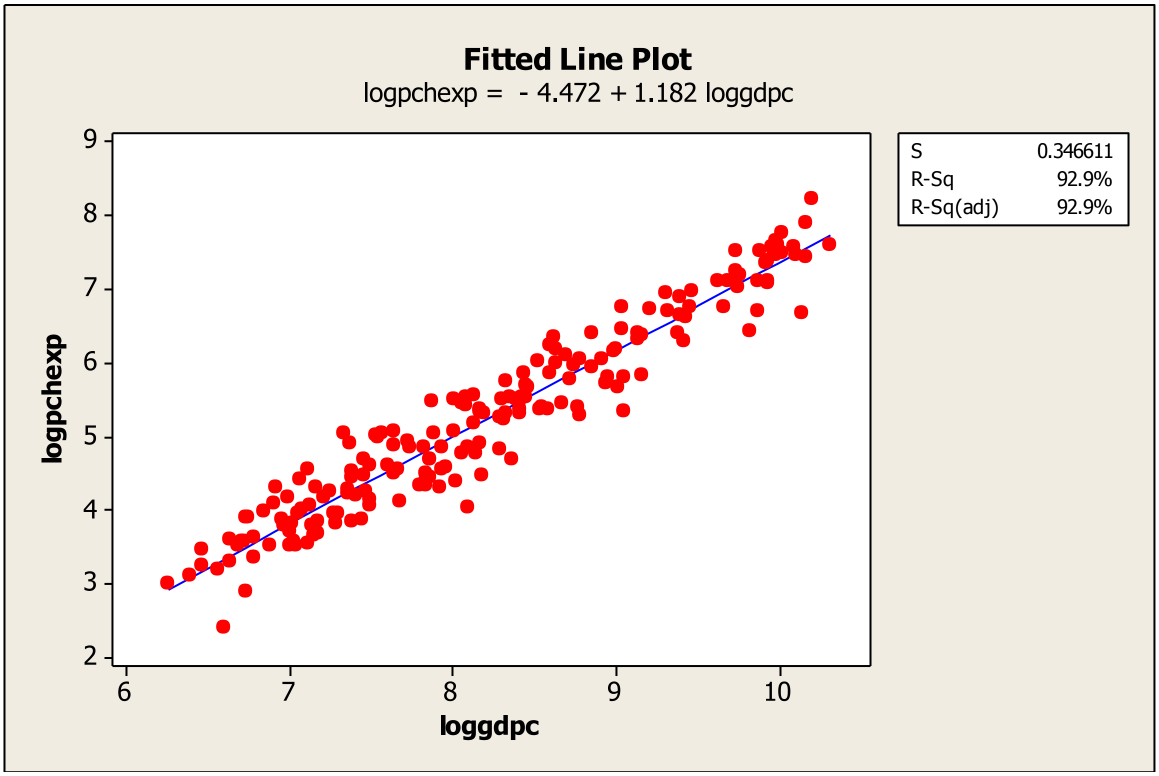




which is more in line with expectations. The R^2 seems to go up from 72% to 78% when we use logs. However, it is not appropriate to compare these two values, since they describe two different variables and two different regressions.

9. A number that many researchers have analyzed in the study of health care is the elasticity of health care expenditure with respect to per capita income. You can obtain an estimate of this important parameter by using a log-on-log regression of the log of health care expenditure (PCHEXP) on the log of per capita income (GDPC) using the WHO health care data, WHO-HealthStudy.mpj. (Note, you will have to obtain the logs of the two variables first. Use the hint in Problem 8.) Carry out the regression. What is your estimate of the elasticity? Is the provision of health care expenditure *elastic* (elasticity > 1) or *inelastic* (elasticity < 1)?

First use Calc to obtain the logs of the two variables. Then the regression, with the plot of actual and fitted values is as follows. The estimate of the elasticity is 1.182 which suggests that per capital health expenditure is elastic with respect to per capita GDP. Many other researchers have found this same result.


Fitted Line Plot
logpchexp = - 4.472 + 1.182 loggdpc
S 0.346611
R-Sq 92.9%
R-Sq(adj) 92.9%
logpchexp
loggdpc
9
8
7
6
5
4
3
2
6
7
8
9
10

===
===
===
===

13. (Continuing our art theme from Notes 15.) Suppose that the relationship that we discovered for Monet's paintings in Notes 15 (and the discussion of prediction using the model in Notes 16) also applied to Salvador Dali's paintings. The Hallucinogenic Torreador is one of Dali's most famous, really huge surrealistic paintings. (If you've never seen it, it is definitely worth the effort. It lives in the Dali museum in St. Petersburg, Florida. Some discussion is at http://www.cdc.gov/ncidod/EID/vol6no5/cover.htm.) The exact dimensions of the painting are 398.8 cm by 299.7 cm, which at 2.54 cm/inch is 157" by 118" (or 13.08 feet by 9.83 feet). Assuming these dimensions, use the model that we developed for Monet to compute a prediction of the sale price of this Dali painting. Also form a prediction interval using the results and formulas shown in class (and in notes 15 and 16). (Note, the surface area used in the model is in square inches, which is 157x118 = 18,526.) Do note, when you reach your answer for the predicted sale price, the number you get illustrates the danger of extrapolating a regression line far (in this case, extremely far) outside the range of experience. More specifically, Monet never in his entire career, painted a single canvas even remotely the size of this one.)

SOLUTION

The estimated regression equation was ln($price) = 2.825 + 1.725 ln(Surface Area). For a painting that is 18,526 square inches, the prediction of the log of the price would be
2.825 + 1.725 log(18,526) =19.776455. Using the simple formula given in the notes for the prediction of the price, exp(2.825 + 1.725 log(18,256) gives $387,976,369.13
To form a prediction interval for the log of the price, we use

$$\text{Estimated log price} \pm 1.96 s_e \sqrt{1+\frac{1}{n}+\frac{(\text{logArea}^* - \overline{\text{logArea}})}{\Sigma_{i=1}^{n}(\text{logArea}_i - \overline{\text{logArea}})^2}}$$

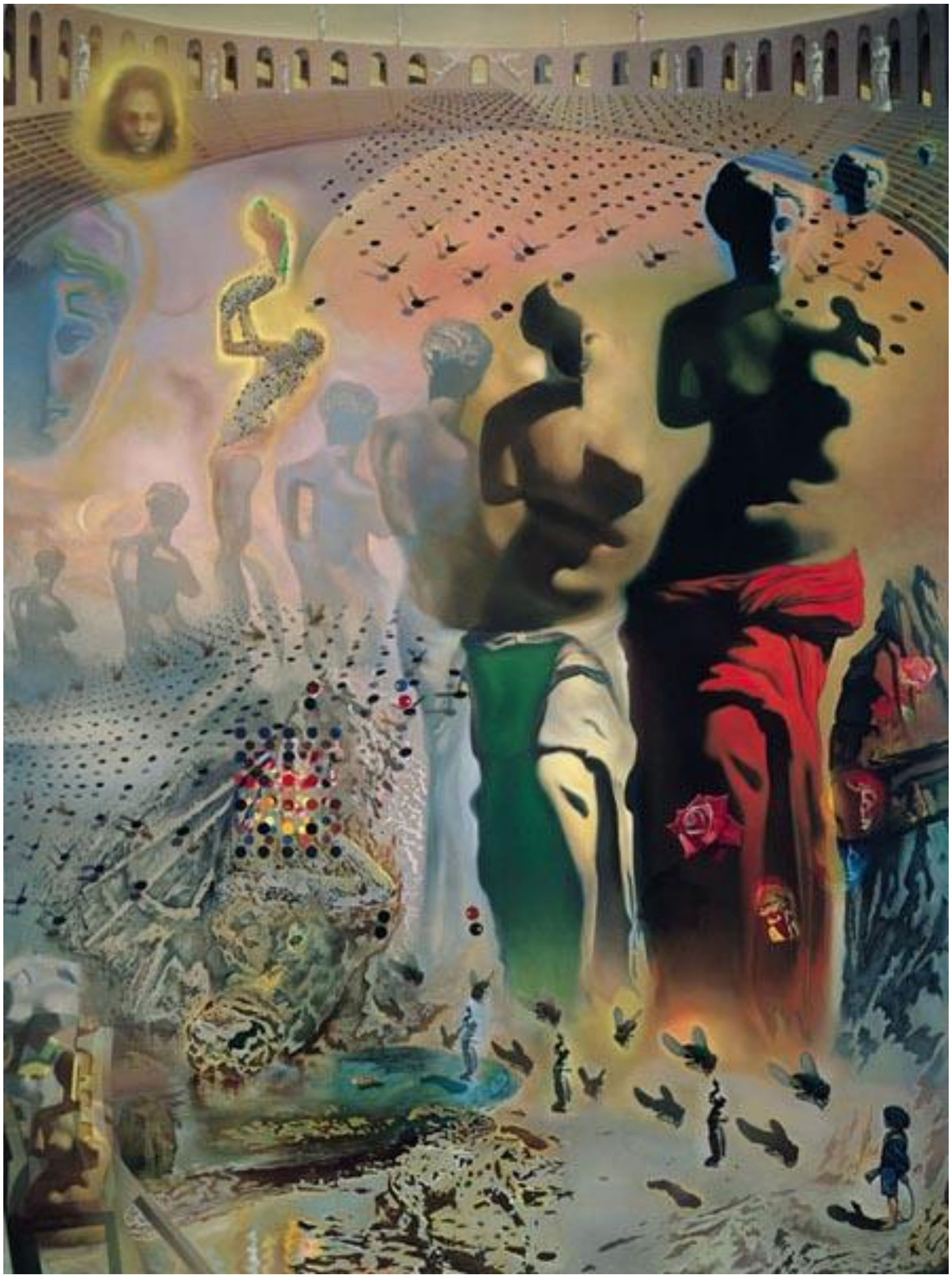
For the sample used in class, n = 328. We derived the denominator of the term in the square root in class as 27.82452. So, the prediction interval is

$$\begin{aligned}&19.776455 \pm 1.96(1.00645)\sqrt{1+\frac{1}{328}+\frac{(9.826930-6.72918)^2}{27.82452}}\\&=19.776455 \pm 1.96(1.00645)(1.161002)\\&=17.46822 \text{ to } 22.06670\end{aligned}$$

Taking the exponents of the lower and upper bounds gives

Lower	=	$ 39,579,052.28
Expected	=	$387,976,372.50
Upper	=	$3,832,181,269.

Of course, noone is likely to pay $383 million for a painting. (But, hey, you never know. Ron Lauder paid $150 million for a painting by Klimt last year. And, Steve Wynn punched his elbow through a painting that he had just sold for $100M.) Of course, the Dali painting is not for sale. The problem here is that we are using the model for prices of Monet's paintings to try to value a painting by a different artist, and in a completely range of sizes. The Hallucinogic Toreador is about 10 times as the largest Monet in the sample, so it is far outside the range of experience represented by that sample. The model could not possibly represent the price for a painting this large.



Hallucinogenic Torreador, Salvador Dali, 1969. Best viewed in color.

14. The file KansasCtyPopn.mtp gives the populations of the 105 counties of Kansas in the years 1980 and 2000. The objective is the relationship between the 2000 population and the 1980 population.

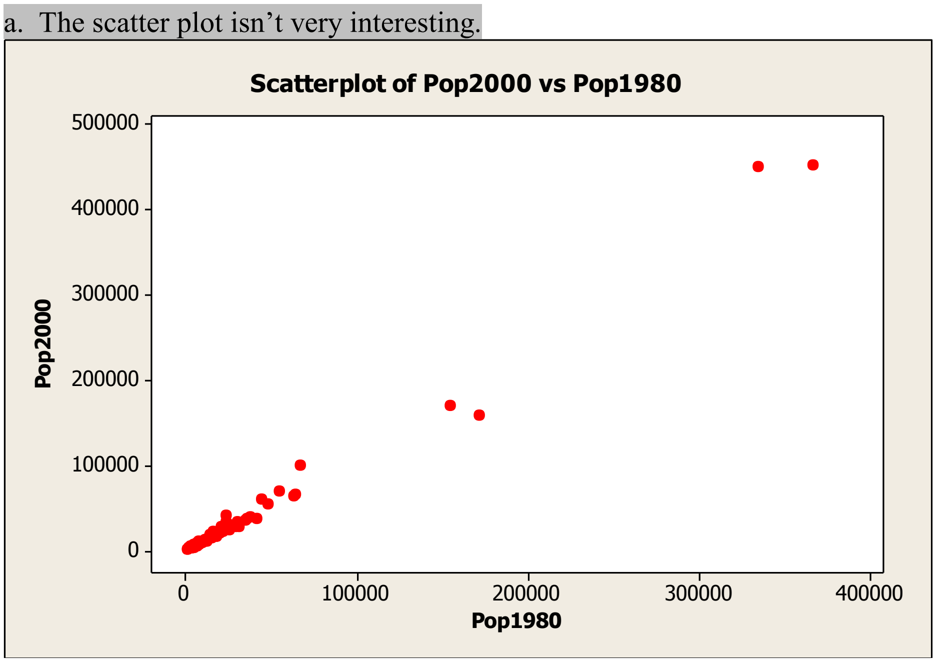
(a) Prepare a scatterplot of the data. The figure should suggest that both the population figures should be replaced by their logarithms.

(b) Replace both Popn1980 and Popn2000 by their base-*e* logarithms. Use Minitab command **Calc → Calculator** for this. The names LP1980 and LP2000 would be reasonable for the transformed variables. Then find the regression of LP2000 on LP1980. [At this step you should also ask for the residual versus fitted plot, as it will be needed in part (d).] Be sure to note the estimated slope.

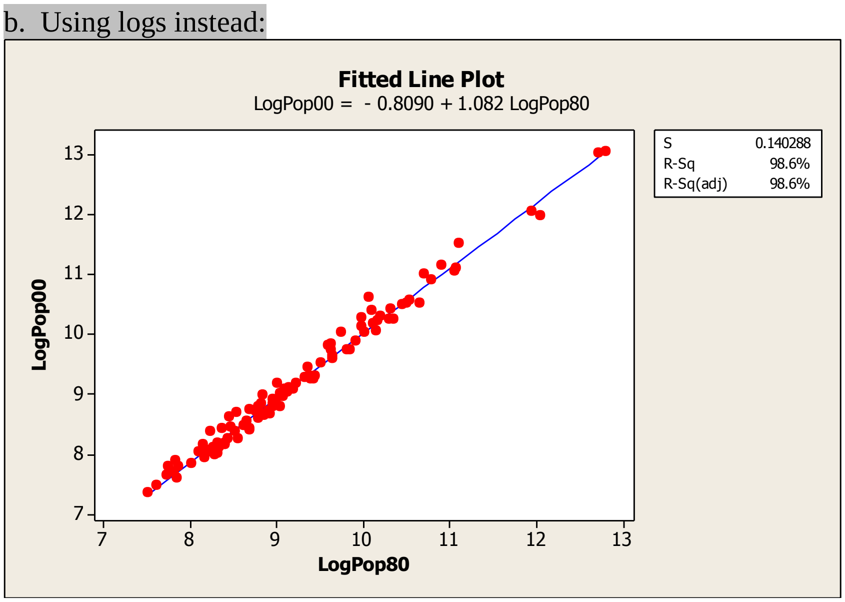
(c) Since the estimated slope exceeds 1, the suggestion is that the 2000 numbers have a greater internal variability than the 1980 numbers. Confirm this by getting the coefficient of variation for Popn1980 and Popn2000. Use **Stat → Basic Statistics → Display Descriptive Statistics → Statistics** and check off the coefficient of variation. Verify that Popn2000 has the larger coefficient of variation. NOTE: The coefficient of variation is defined as $s/\overline{x} = \text{standard deviation/mean}$. Minitab reports this as a percent so that a coefficient of variation of 2.00 would be reported as 200 (for 200%).

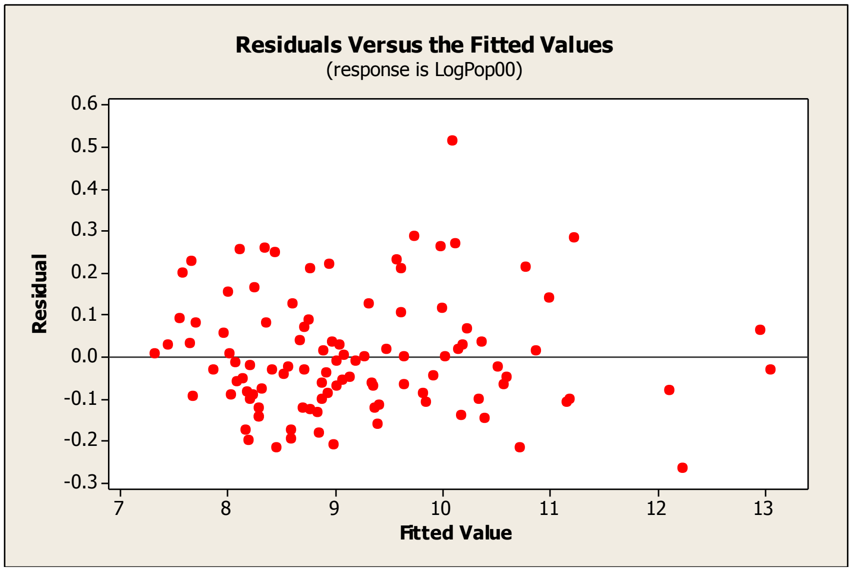
(d) Examine the residual versus fitted plot for this regression. Does it suggest that there are problems? There will be one very large residual. Identify the county and give the corresponding data numbers. (This point can be found on the list of unusual observations.)

a. The scatter plot isn't very interesting.



b. Using logs instead:





c. Descriptive Statistics: Pop1980, Pop2000

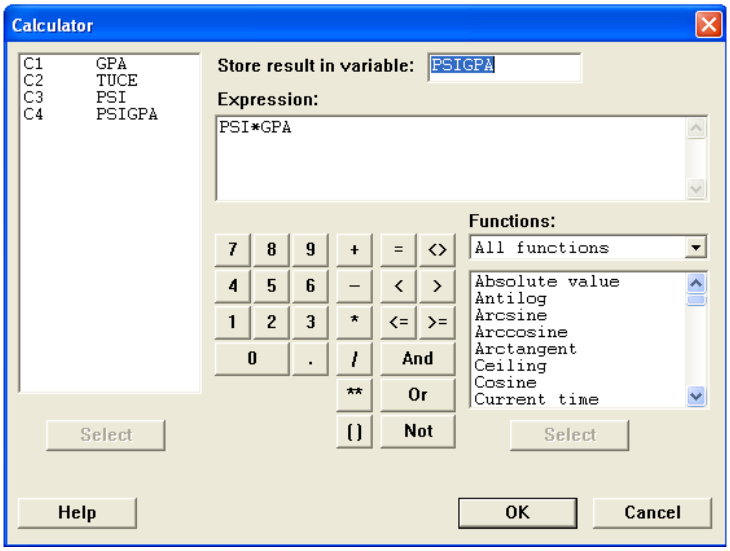
```
Variable     N  N*   Mean  SE Mean  StDev  CoefVar  Minimum     Q1  Median     Q3
Pop1980    105   0  23125     5106  52318   226.24     1845   4513    8234  20897
Pop2000    105   0  25601     6376  65333   255.20     1534   3790    7673  22558

Variable  Maximum
Pop1980    367088
Pop2000    452869
```

d. There is one huge residual, Finney county which grew by over 50% over the period, while the other counties grew at more moderate paces or, in most cases, actually shrunk. The regression results are strange. Most counties shrunk, but the regression suggests an average growth of about 8%. The simple average of the growth rates is -5%, not +8%.

6. This question will look ahead a bit to later in the course, and use a trick to compute two separate regressions with one set of computations. We are interested to know, does it appear that different regression equations apply to students who took PSI and those who did not. In principle, you could find out by computing the regression using the 18 observations on students with PSI = 0 then repeat it with the 14 students with PSI = 1. Here is how you can do both at the same time:

(1) Use CALC to compute a new variable, PSI * GPA.



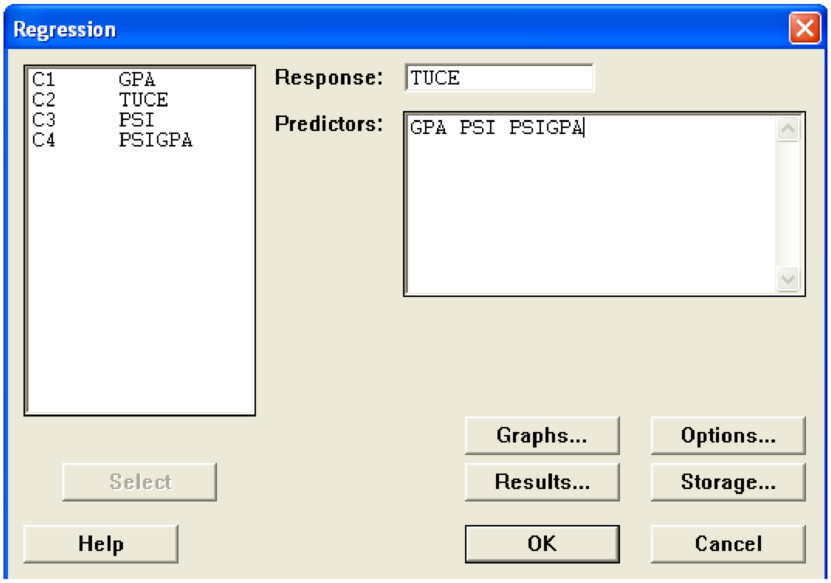
```
C1      GPA
C2      TUCE
C3      PSI
C4      PSIGPA
```

Store result in variable: PSIGPA

Expression: PSI*GPA

Functions: All functions

(2) Now, use Stat → Regression → Regression to compute a regression, but instead of specifying just one "Predictor," specify 3, TUCE, PSI and PSIGPA (one at a time, unfortunately).



```
C1        GPA
C2        TUCE
C3        PSI
C4        PSIGPA
```

Response: TUCE

Predictors: GPA PSI PSIGPA

(3) Now, push the OK button. The results will look as follows:

Regression Analysis: TUCE versus GPA, PSI, PSIGPA

```
The regression equation is
TUCE = 5.77 + 5.09 GPA + 11.5 PSI - 3.44 PSIGPA
Predictor     Coef   SE Coef       T       P
Constant    ******     6.582    0.88   0.388
GPA         ******     2.104    2.42   0.022
PSI         ******     8.950    1.28   0.210
PSIGPA      ******     2.840   -1.21   0.236
```

Where I have covered the numbers you need. The a and b for the PSI = 0 group are the coefficients on CONSTANT and GPA in the results. The a in the PSI=1 group will be the CONSTANT + Coefficient on PSI. The b in the PSI=1 group will be the Coefficient on GPA + Coefficient on PSIGPA. (These are exactly the numbers you would get if you split the data set into the two groups and did your computations separately.) So, to come to the point, what did you find? Is the relationship for one group stronger than for the other? What is your conclusion? (Note for the inquiring minds… You can deduce the values of the coefficients from the numbers that I have not covered in the table above. Do you see how? You will have to search a bit in your text for this. We have not covered it in class yet.)

The results appear below. The coefficients can actually be deduced as the product of T times SE Coef. (Try it.) The reason is that (apparently obviously) T = b / SE.

Regression Analysis: TUCE versus GPA, PSI, PSIGPA

```
The regression equation is
TUCE = 5.77 + 5.09 GPA + 11.5 PSI - 3.44 PSIGPA
Predictor       Coef   SE Coef       T       P
Constant       5.774     6.582    0.88   0.388
GPA            5.089     2.104    2.42   0.022
PSI           11.471     8.950    1.28   0.210
PSIGPA        -3.437     2.840   -1.21   0.236
S = 3.66938    R-Sq = 20.1%    R-Sq(adj) = 11.5%
```

Note that compared to the regression in Part 4.c., R^2 has increased but R^2(adj) has decreased..

7. Consider a set of (x, Y) data in which $n = 25$, $\Sigma_i x_i = 400$, $\Sigma_i x_i^2 = 9{,}800$, $\Sigma_i y_i = 146{,}000$, $\Sigma_i y_i^2 = 944{,}000{,}000$, and $\Sigma_i x_i y_i = 2{,}120{,}000$. Find the fitted equation resulting from the regression of Y on x. Then obtain the estimate of s_ε, the estimate of the noise standard deviation.

We first need to find the means, variances and covariance.

$$\bar{x} = \frac{400}{25} = 16,\ \ \bar{y} = \frac{146{,}000}{25} = 5{,}840$$

$$S_x^2 = \frac{\left(\Sigma_{i=1}^n x_i^2\right) - n\bar{x}^2}{n-1} = \frac{9{,}800 - 25(16^2)}{24} = 141.667$$

$$S_y^2 = \frac{\left(\Sigma_{i=1}^n y_i^2\right) - n\bar{y}^2}{n-1} = \frac{944{,}000{,}000 - 25(5{,}840^2)}{24} = 3{,}806{,}667$$

$$S_{xy} = \frac{\left(\Sigma_{i=1}^n x_i y_i\right) - n\bar{x}\,\bar{y}}{n-1} = \frac{2{,}120{,}000 - 25(16)(5840)}{24} = -9000$$

Now, use the usual formulas for a, b, and s_e^2.

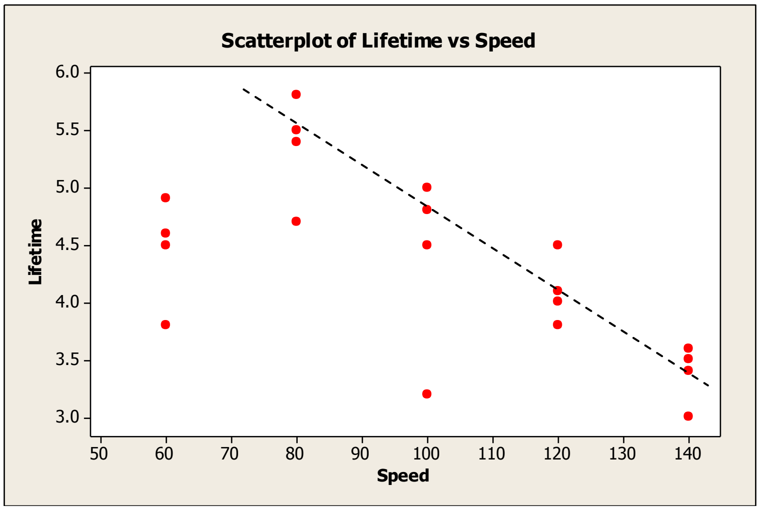
$$b = \frac{S_{xy}}{S_x^2} = \frac{-9{,}000}{141.667} = -63.5294,$$

$$a = \bar{y} - b\bar{x} = 5{,}840 - (-63.5294)16 = 6{,}856.47$$

$$s_e = \sqrt{\frac{(n-1)(S_y^2 - b^2 S_x^2)}{n-2}} = \sqrt{\frac{24(3{,}806{,}667 - 63.5294^2(141.667))}{23}} = 1837.27$$

10. HOG, problems 11.7 and 11.8, page 518-519.

Here is the scatter plot of the data. There does appear to be a relation, however, not a linear one. On the other hand, if one considers only the observations with speed greater than 60, then there does appear to be a linear relationship at work. With the first group of observations, the relationship appears to be quadratic. There is an outlier in the data set, the low point in the Speed=100 group. This is unlikely to be a very influential observation, however, because it is right in the middle of the data set. In fact, the speed = 100 of this data point exactly equals the sample mean of speed, so the influence measure for this point will be zero.

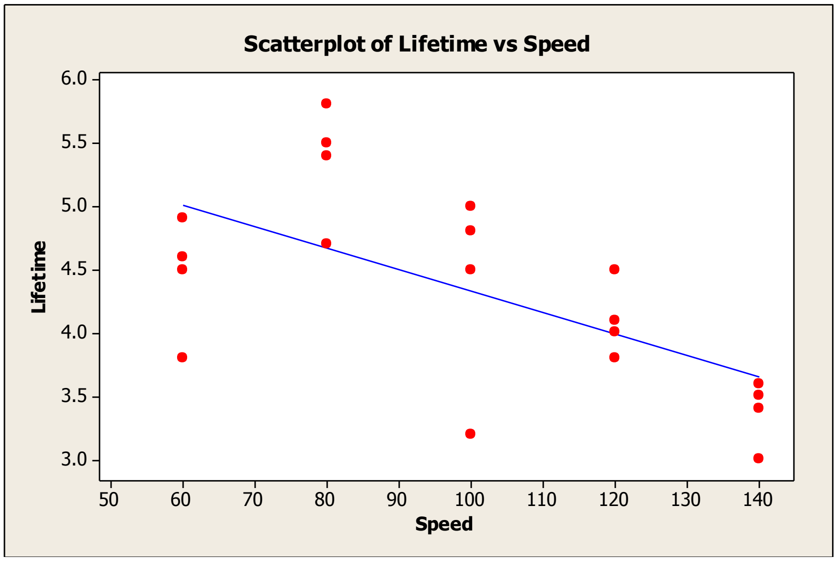


For HOG, Problem 8:

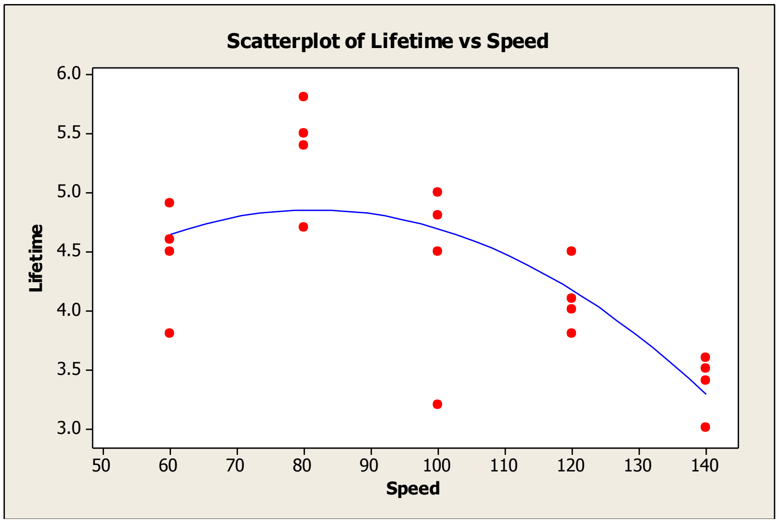
a. The intercept and slope appear at the end of the results; a = 6.03 and b = -0.017.

b. The negative slope indicates that higher speeds are associated with lower lifetimes. Friction being what it is, this seems natural.

c. The residual standard deviation is given at the top of the results, $s_e = 0.6324$. It is an estimate of the standard deviation of the disturbances around the regression at specific values of Speed. I have placed a dashed line in the figure above. It is not the regression line, since it ignores the leftmost 4 points. However, essentially, the residual standard deviation is a measure of the dispersion of the individual groups around the line shown. Note that it will be considerably distorted by the odd four points. They will serve considerably to increase the computes s_e. The figure below, which includes all the points in the scatter makes the problem clear.



Minitab's quadratic regression shown in the next figure shows how the "model" improves when more detailed account is taken of the configuration of the data.


Scatterplot of Lifetime vs Speed
Lifetime
Speed