4.6

(f) $\hat{y} = 10^{(-1094.51 + 0.558(year))}$

(g) $\hat{y} = 10^{(-1094.51 + 0.558(1982))} = 12178673.85$ acres

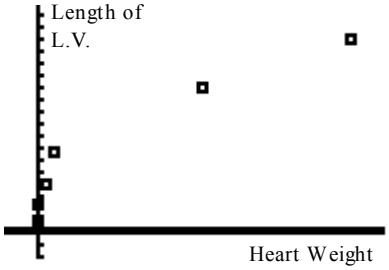
4.11

(b) The transformed data looks like two lines; a line with one slope for the years up to 1880, and a line with a lesser slope for the years after 1880. The slope corresponds to the rate of increase of the exponential function, so we would expect the years after 1880 to show slower rate of increase.

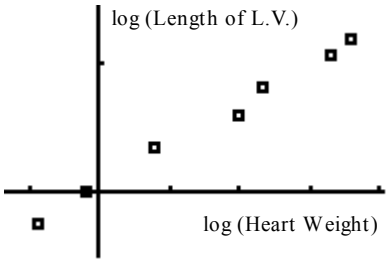
(c) Answers will vary. If students used all of the data, their prediction will be too high. It should improve if they edit their data.

4.14

The scatterplot of the Length of the cavity of the left ventricle (y) vs. the Heart weight (x) shows an extremely strong, but clearly non-linear relationship.

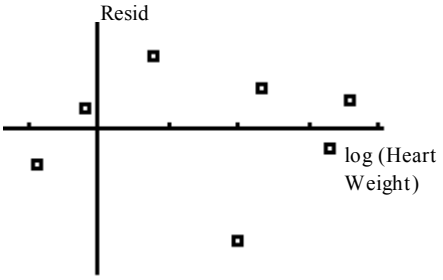


Applying our methods of transforming data we plot $\log(y)$ vs. $\log(x)$, and observe that there is a very strong, positive, linear relationship between $\log(y)$ and $\log(x)$. We therefore perform a linear regression of $\log(y)$ onto $\log(x)$.



$\log \hat{y} = 0.0467 + 0.316478 \cdot \log(x)$, where $r = 0.996$, indicating a very strong, positive, linear relationship.

We also check the residual plot, and notice that there is no discernable pattern, also indicating that the linear model relating $\log(y)$ and $\log(x)$ is good.



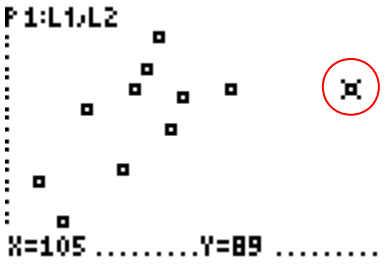
Based on our work above, we can conclude that the Length of the cavity of the left ventricle in various mammals is related to their Heart weight by the power model

$$\text{Lenght of L.V.} = 10^{0.0467}(\text{Heart Weight})^{0.316478}$$

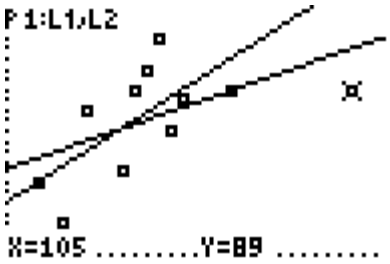
4.30
If x is the explanatory variable (# of beds) and y is the response (median number of days of stay), there may be a lurking variable z representing, for example, the types of procedures done in small and large hospitals that may explain why people remain longer in larger hospitals. For example, maybe more serious procedures, requiring more serious and lengthy care are performed at larger hospitals.

4.31
We would expect the correlation between the individual stocks' change in January vs. how much it changed throughout the year to be lower. This is because the S&P 500 index is an average, and averaged data reduces the variation that we would see from stock to stock.

4.32
The graph with the potential outlier is shown to the right.

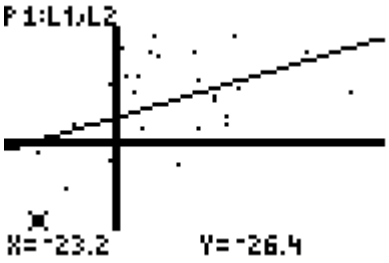


And here is the graph with the two regression lines. The first line, $\hat{y} = 20.49 + 0.754x$ omits the outlier. We can see this because the second line is "pulled towards the influential point (with smaller slope), and does not go through the middle of most of the data.

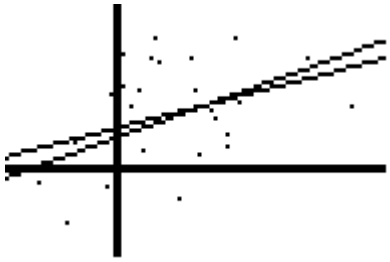


4.40
Possible lurking variables may include: level of academic aptitude, size of school, location of school, college admission requirements. Since there may be other variables influencing success in college, we should not say that taking algebra and geometry causes success. We only observe the association, which may or may not have a causal link.

4.48
(a) The year with the largest residual is (-23.2, -26.4), 1974



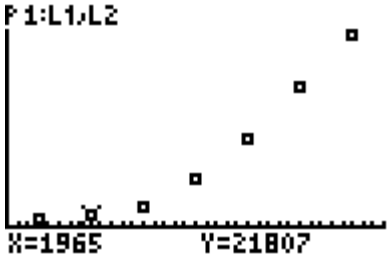
The plot without this residual is shown here, with the new LSRL.



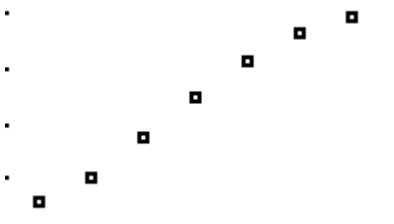
We can see that the line did not change much so we say that the point is not influential.

There do not appear to be any distinguishable patterns in the time vs. residual plot.

4.76

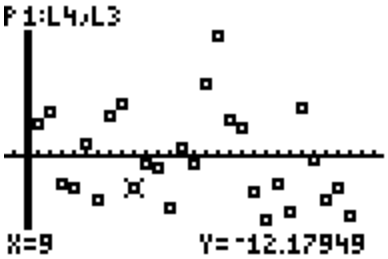


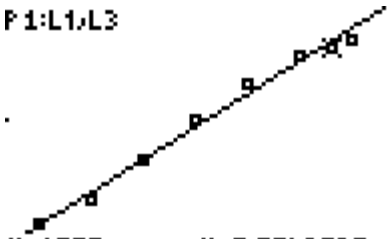
This plot looks much more exponential than linear.



The transformed data seems much more linear, and so the exponential model may be a good fit.

$\log(\text{social ins expenditures}) = -98.63833 + 0.05244(1988) = 5.61239$,
so $(\text{expenditures}) = 10^{5.61239} = 4092628.34$ million dollars.



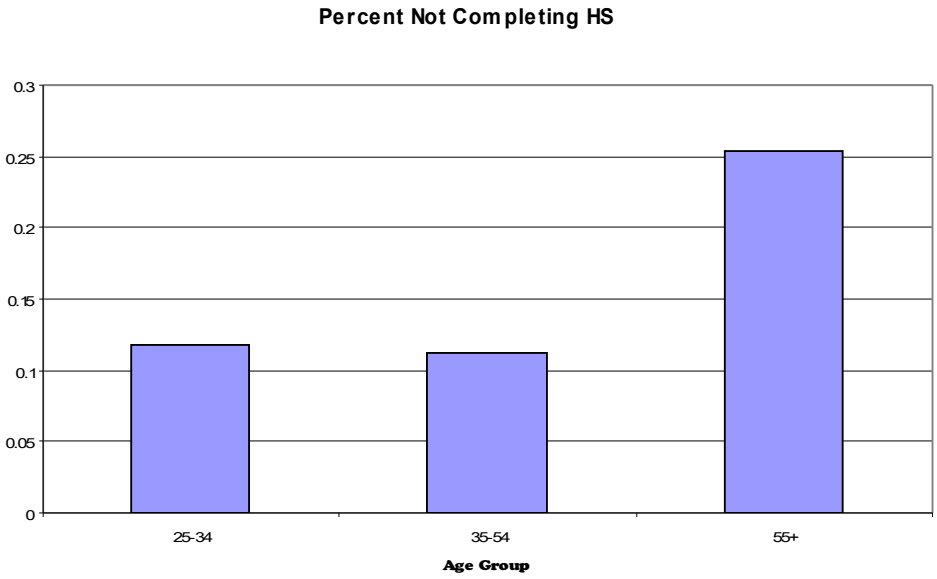


. The plot with the added data is shown. The point for 1988 falls not too far from the line, just slightly below. There is not evidence that the trend of growth changed in a major way.

4.51
The marginal distribution of "Age" is

Age	
25-34	0.215637
35-54	0.464732
55+	0.319626

4.52



The bar graph to the left shows the percentage in each age group that did not complete high school. We note that while the two younger age groups are close to the same, the 55+ group is notably higher. That is more 55+ year olds did not complete high school than in the other two groups.

4.55
6014 men took part in the study. 1.26% of them died in the first 5 years
The explanatory variable is blood pressure.

The proportions are presented in the table below:

	Died
Low	0.008
High	0.016

From the table we see that twice as many with high blood pressure died as with low blood pressure, so one might say there is an association.

4.60

(a)

	Admit	Deny
Male	490	210
Female	280	220

(b)

	Admit
Male	0.7
Female	0.56

(c)

	Business Admit	Law Admit
Male	0.8	0.1
Female	0.9	0.33

(d) We could note that the Law School has many more female applicants than male applicants, and the Business School has many more male applicants than female applicants. Overall they have many more Male applicants. So it makes sense that they admit more Males overall, but the difference when we look at the schools separately shows that of the females that apply to those schools more get admitted than of the males that apply.

4.70

(a)

Player	Hits	At Bats	Avg
Joe	120	500	0.24
Moe	130	500	0.26

(b) Moe has a slightly higher batting average

(c)

Right Pitchers

Player	Hits	At Bats	Avg
Joe	40	100	0.4
Moe	120	400	0.3

Left Pitchers

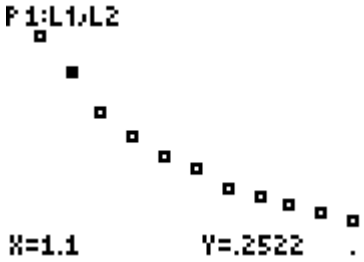
Player	Hits	At Bats	Avg
Joe	80	400	0.2
Moe	10	100	0.1

Joe has a higher average against both types of pitcher.

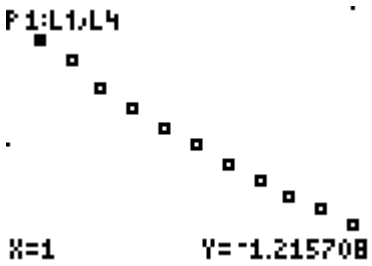
(d) The varible (L- or R-handed Pitcher) influences batting average. If we ignore the effect of this "lurking variable" in our analysis, we may not get a clear picture and be mislead into thinking that only the hitter influences batting average.

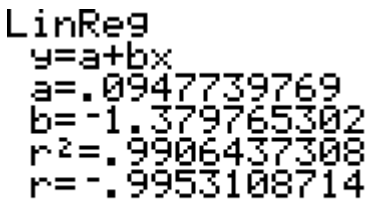
4.72

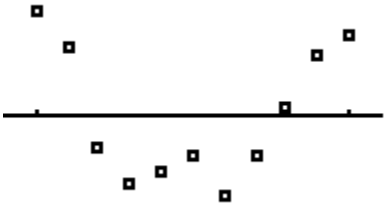
The scatterplot shows a very strong, non-linear decreasing relationship between light intensity and distance.



Applying an exponential transformation we obtain:

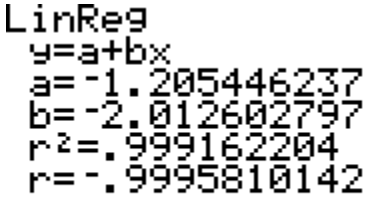


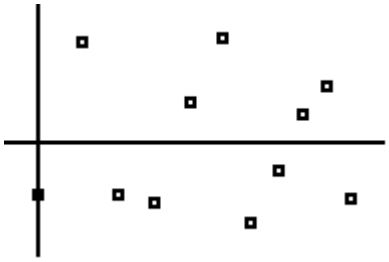




ln y vs x Resid vs. x

The patterned resudual plot suggests that our data may not be exponentially related... so we try the power transformation:



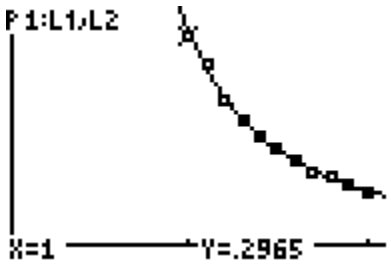


ln y vs ln x Resid vs ln y.

The power model gives much more satisfactory results.

Thus, the power model is (for x = distance from source in m, and y = Light Intensity (cd))

$\hat{y} = e^{-1.2054} x^{-2.0126}$ This equation is plotted with the original data below:



The intensity of the light varies inversely with the square of the distance.

4.78

There is a strong, non-linear relationship between the year and the number of EFT's. We also note tha the pattern of growth seems to be different for the last three years.

We observe the relationship log(EFT) vs. years for the years 1985 thru 1996. There is a strong, positive linear relationship ($r = 0.997$). The residual plot shows a fairly even scatter BUT, there may be some indication of cyclical data in the residual plot. This is beyond the scope of our class, but could be an interesting investigation.

For the years 1997 thru 1999 we perform a similar analysis and obtain relativiely satisfactory results to confirm an exponential growth model (although it is difficult to tell for sure with just three data points.)

Here are our results:

Years 1985 thru 1996	Years 1997 thru 1999
$\log(\hat{EFT}) = 3.289 + 0.0483(\text{Years after 1980})$ $\hat{EFT} = 10^{3.289} \cdot 10^{0.0483(\text{Years after 1980})}$	$\log(\hat{EFT}) = 3.892 + 0.0123(\text{Years after 1980})$ $\hat{EFT} = 10^{3.892} \cdot 10^{0.0123(\text{Years after 1980})}$

4.80

The timeplot seems to indicate a negative association with year and the percentage of voters. That is, over time, it appears that the percentage of voters has decreased.

We note the significant drop in percentage beginning with the 1972 election. A partial explanation using the fact that the voting age changed in 1970 might be that a lot of new 18-yr old eligible voters did not exercise their right to vote... and have not much since then.

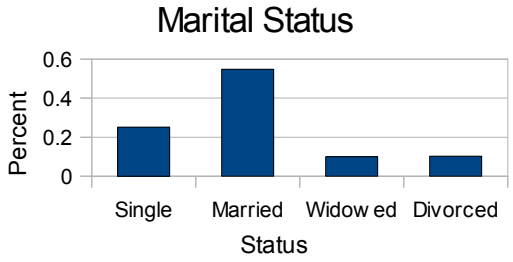
4.81

(a) Here is my completed table. The total entry differs from that given due to roundoff error.

Age	Single	Married	Widowed	Divorced	Total
15-24	16121	2694	21	203	19039
25-39	7409	19925	212	2965	30511
40-64	3553	29687	2338	6797	42375
65 plus	680	8223	8490	1344	18737
	27763	60529	11061	11309	110662

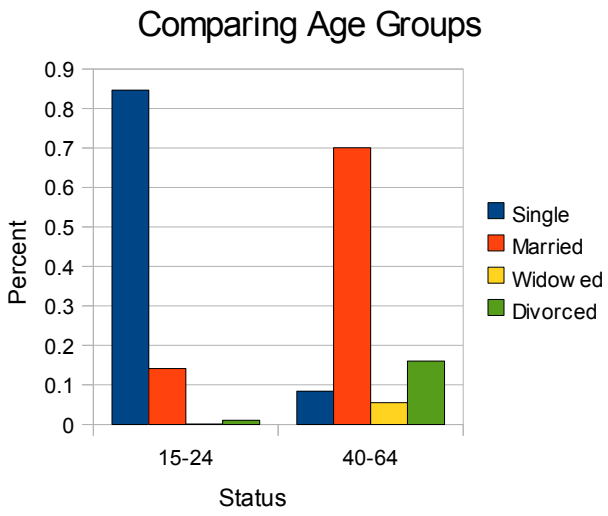
(b) The marginal distribution is:

Single	Married	Widowed	Divorced
0.25	0.55	0.1	0.1



(c) Here are the conditional distributions for the two age groups.

Age	Single	Married	Widowed	Divorced
15-24	0.85	0.14	0	0.01
40-64	0.08	0.7	0.06	0.16



We can see that most of the younger women are single (85%) while most of the older women are married (70%).

(d) Here is the requested distribution:

Age	Single
15-24	0.58
25-39	0.27
40-64	0.13
65 plus	0.02

4.82

This is another example of Simpson's paradox. When we consider the "lurking" variable of "field of study", we realize that this variable affects "salary". So, clearly, when we consider its influence, we see "more" than we otherwise would.