Unit 2 Answer Key

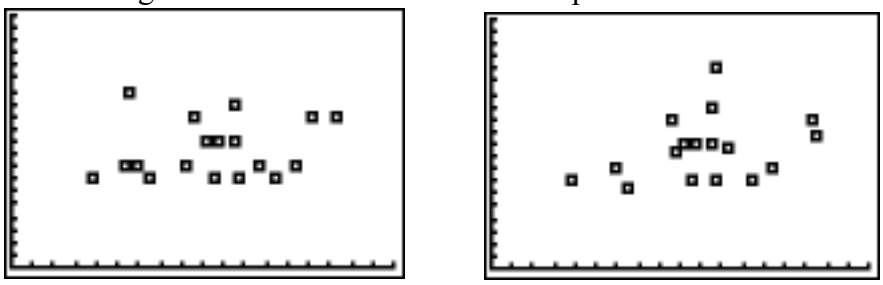
2.1 Homework Answer Key

2.2 Homework Answer Key

1.
   a. There appears to be a relationship between all 4 plots. Plots 1, 2, and 4 look like a linear model could fit, and plot 3 a linear relationship looks less likely.
   b. Scatterplot 1 is positive, Scatterplot 2 is negative, Scatterplot 4 is negative.

2.
   a. Positive correlation, as temperature increases, the energy costs to cool a home would likely increase as well.
   b. Close to 0, as it is unlikely that height and IQ are related in any way.
   c. Positive correlation, as it is likely that foot size increases (generally) as height increases.
   d. Positive correlation, if both tests are valid assessments of mathematical knowledge/ability, as SAT score increases, ACT score should increase as well.

3.
   a. $r = 0.204$, this indicates a weak, positive, linear trend
   b. $r = 0.241$, r increased in value because as we set the price to a fixed amount, any difference in price that was due to a difference in the amount of cereal is limited by doing this.
   c. Per Serving Data: Per Cup Data:



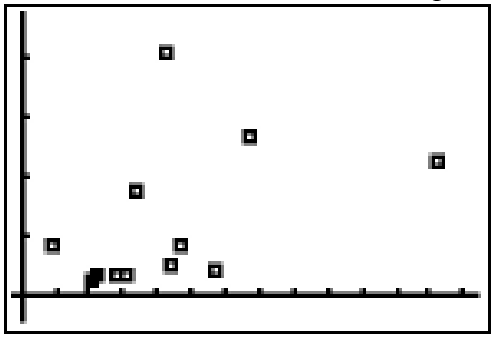
   The plots seem very similar, both indicating a weak positive linear trend, though the per cup data is slightly more clustered and appears to have an outlier at (56, 28), while the per serving data has what appears to be the corresponding point as an outlier at (28,14).

4.

a. The correlation between Average Hopping Power and Arch Height is a *weak, negative, linear correlation*. The fact that it is negative means that increases in arch height generally result in decreases in hopping power (thought the trend is very weak).
b. The correlation coefficients do support this conclusion – "flat-footedness" does not appear to be a disadvantage in this age group. This is because the correlations are all very close to zero (the greatest distance from zero being 0.10), so the data between each motor ability test does not correlate well (linearly) with arch height.

5.

a. $r = 0.431$. This indicates a *weak, positive, linear* correlation.



b. Given that they indicate the game is "open-world" (involving exploration) but you only earn experience for defeating "monsters", it could be that in some of the sessions the player explored extensively without fighting and didn't earn experience during long periods of playing the game. In addition, many games have an experience progression that is more exponential than linear, so later portions of the game may have much higher experience rewards than a linear trend would indicate.

2.3 Homework Answer Key

1.

a. $r = 0.56$, $\hat{y} = 51.222 + 0.166x$. $r^2 = 0.315$, so 31.5% of the variation in the percentage of alumni who strongly agree their education was worth it is explained by the college ranking.
b. When $x = 50$, $\hat{y} = 59.508$, so we would predict about 50.5% of the students would agree if their college was ranked 50.
c. It would not be appropriate to extrapolate to a value of $x = 10$ because our data set is from $x = 28$ to $x = 98$. 10 is relatively far out from the data set, and the linear trend may not continue.

2.

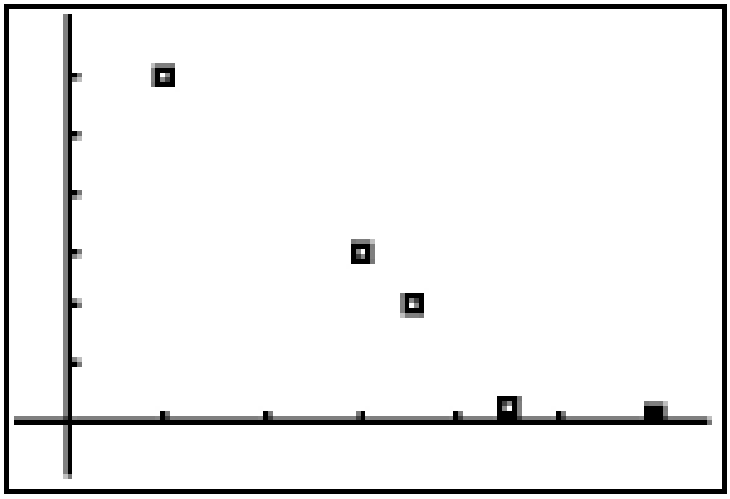
a. $r = 0.942$, $\hat{y} = 10.144 + 0.974x$. The value of r indicates a *strong, positive, linear* correlation. This tends to indicate that the linear function would be good to predict values for the Wright meter.
b. Since $r^2 = 0.888$, that means 88.8% of the variation in the Wright meter is explained by the Mini-Wright meter.
c. If $x = 525$, the predicted value for the Wright meter would be 521.3. If $x = 400$, the predicted value for the Wright meter would be 399.6.

3.

a. The independent (explanatory) variable is x = hours of television watched and the dependent (response) variable is y = the number of fruit and vegetable servings.
b. It would have a negative slope because there is an average decrease in fruit and vegetable servings for each hour of TV viewed.
c. Since $r = 0.64$, $r^2 = (0.64)^2 = 0.4096$, so 40.96% of the variation in fruit and vegetable servings is explained by the hours of TV viewed.
d. Since (1,4.45) is $(\overline{x}, \overline{y})$, and the slope is –0.14, the point-slope form from algebra would give us $\hat{y} - 4.45 = -0.14(x - 1)$, and solving for y would give us $\hat{y} = 4.59 - 0.14x$

4.

a. The relationship looks like it is *negative*, with a strong correlation. The data does not appear to be necessarily linear (though it may be). It seems to have an exponential trend rather than linear.



b. $\hat{y} = 101.328 - 9.2956x$, $r = -0.96$, $r^2 = 0.922$
c. 92.2% of the survival rate is attributable to the mean time to defibrillator use.
d. The predicted average survival rate for a mean time to defibrillator use of 10 minute is 8.37 percent.

e. The predicted average survival rate for a mean time to defibrillator use of 0.10 minute is 100.4 percent, and the predicted average survival rate for a mean time to defibrillator use of 11 minutes is –0.923 percent. Extrapolating outside the data set on the low end gives an unrealistic survival rate (above 100%). While 11 minutes is not outside the data set, using the linear model for this value gives a negative survival rate, which is not possible. These facts tend to indicate that a linear model may not be appropriate for this data, or it may only be appropriate for a limited range of x–values.

2.4 Homework Answer Key

1.

a. The slope would be –4000. Because the home prices decrease as you move toward the central valley, the slope would be negative.

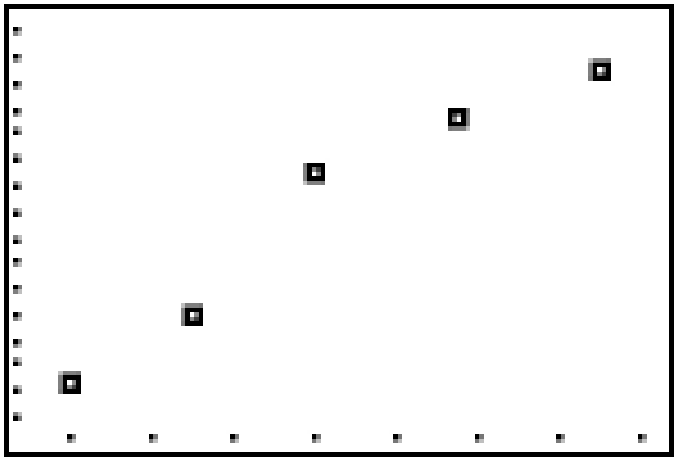
b. $\hat{y} = 450,000 - 4,000x$

c. When x = 30 miles, the predicted average home price is $300,000. If the actual value used for this distance was $315,500, the residual value is 15,500.

d. Since there was no pattern to the residual plot, this tends to indicate that the linear trend is valid, and that the linear model may be a reasonable one.

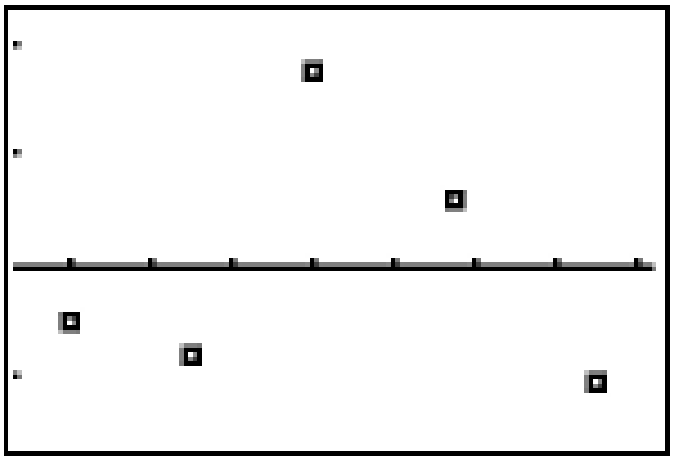
2.

a. The pattern appears to possibly be linear, though it appears that it may have a curve to it.



b. $\hat{y} = 492.729 + 14.764x$, r = 0.974, and 94.8% of the relationship between walk distance and age is explained by the representative age.

c. The residual plot may have a pattern (it looks possibly cubic or perhaps sinusoidal), though with so few points, it is difficult to tell – you may assert that it is scattered as well:

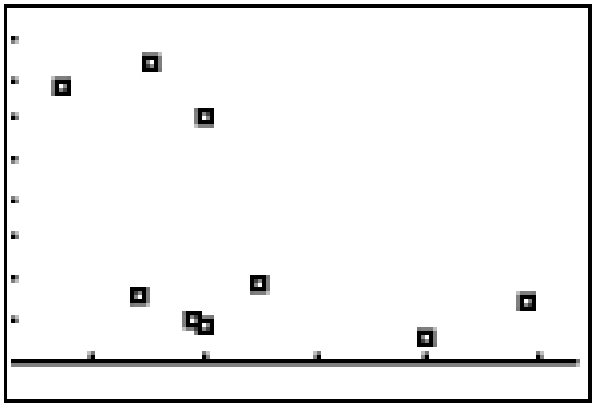


d. The predicted average y-value when $x = 0$ is 492.729 yards. This extrapolation does not make sense, because it would be the distance a new-born infant could walk if the model continued in the linear pattern outside of the range of x–values in our data set, and since new-born infants cannot walk, this is not a reasonable assumption.

e. It is not likely that we could extrapolate to the age of 45 – this is very far outside our data set, and it is not reasonable to assume that this trend would continue to be linear outside of the range of x–values. We would not be able to calculate a residual for this value; to calculate a residual, we need an actual data point to compare our predicted value to, and we do not have any data for $x = 45$.

3.

a. $\hat{y} = 232.2575 - 2.9255x$



b. $r^2 = 0.300$, so 30.0% of the variation in algal colony density is explained by the rock surface area.

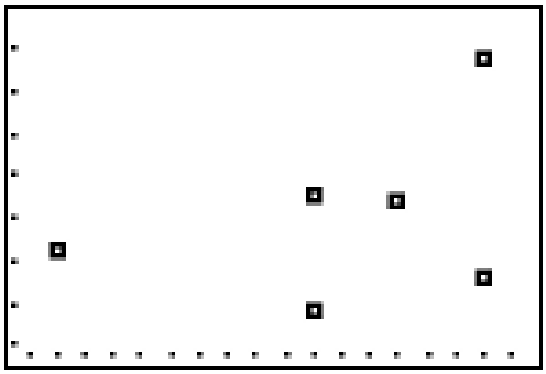
c. The predicted average algal colony density is 85.982 for a rock surface area of 50.

d. The residual value for (50, 152) is 66.018, and the residual value for (50, 22) is –63.892

e. The relationship appears to be *moderate*, *negative*, and *linear* because $r = -0.5475$.
f. Potentially, (79, 35) could be an influential point, and it appears to be a high leverage point.
g. $\hat{y} = 302.8628 - 4.422x$. The *y*–intercept is over 70 units higher without the influential point, and the slope value changes from about –2.9 to –4.4, roughly a 150% of the value with the influential point.

4.

a. $\hat{y} = 86.923 + 0.359x$

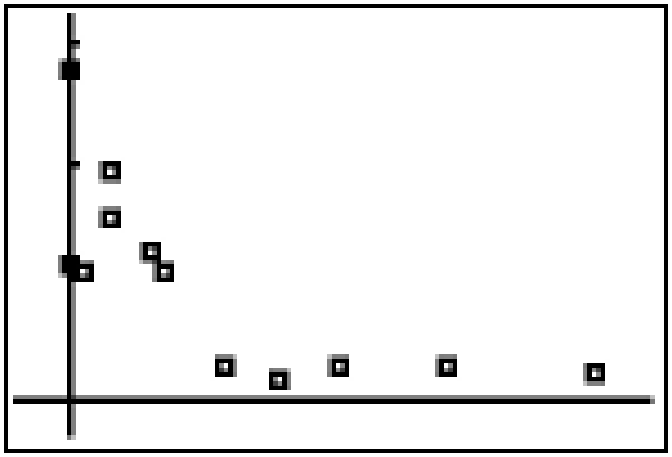


b. $r = 0.379$, $r^2 = 0.144$. This relationship appears to be *weak*, *positive*, and *linear*.
c. $\hat{y} = 173.08$ when $x = 240$, so the predicted average *y*-value when $x = 240$ is 173.08 micrograms per kilogram of French fries. The residual for (240, 120) is –53.08, and the residual for (240, 190) is 16.92.
d. The residual plot appears to be scattered, so a linear model may be reasonable.

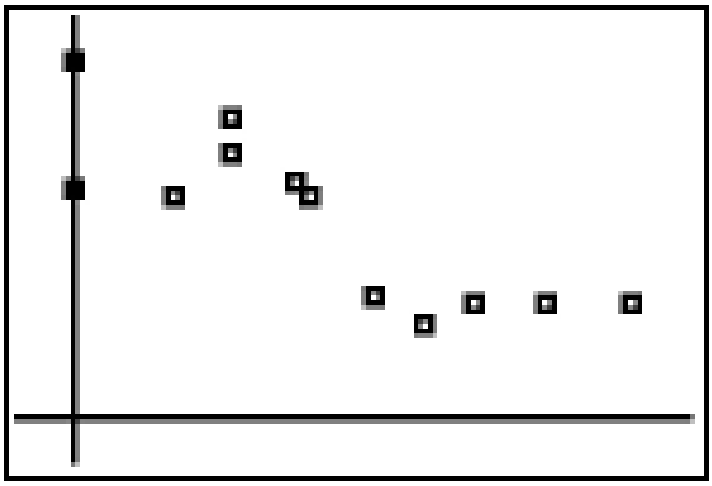
2.5 Homework Answer Key

1.

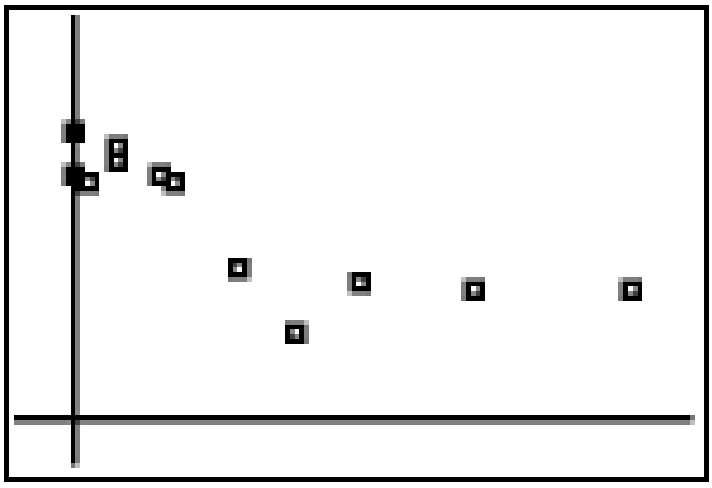
a. The pattern seems significantly curved – it does not appear to be linear at all.



b. Based on this scatterplot, the plot of $\sqrt{x}$ against $\sqrt{y}$ looks much more reasonable, as the transformed data looks more linear. However, it still looks quite curved, so this is still not likely the best transformation.

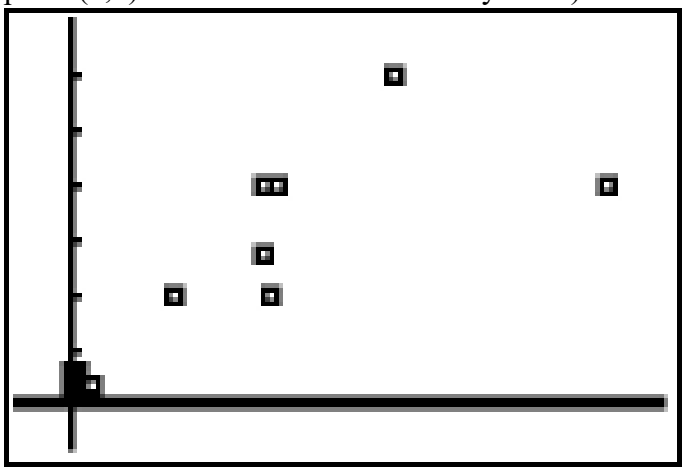


c. Below is the plot of x against $\ln\left(\ln\left(y\right)\right)$. It does appear to be more linear (I chose to use the natural log twice, because when I did one transformation by using the natural log, the data still looked very curved).

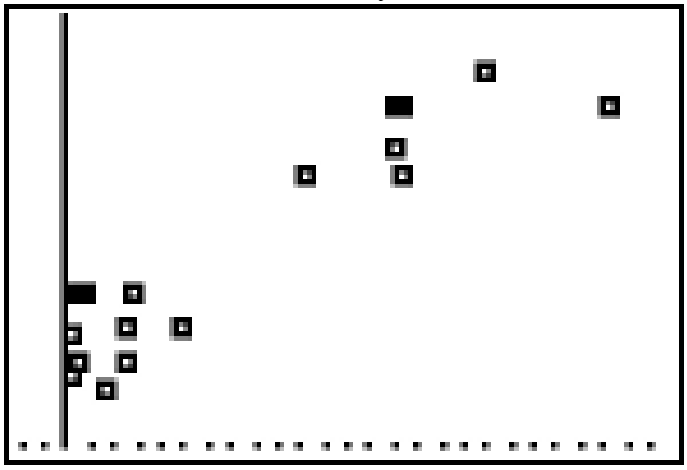


2.

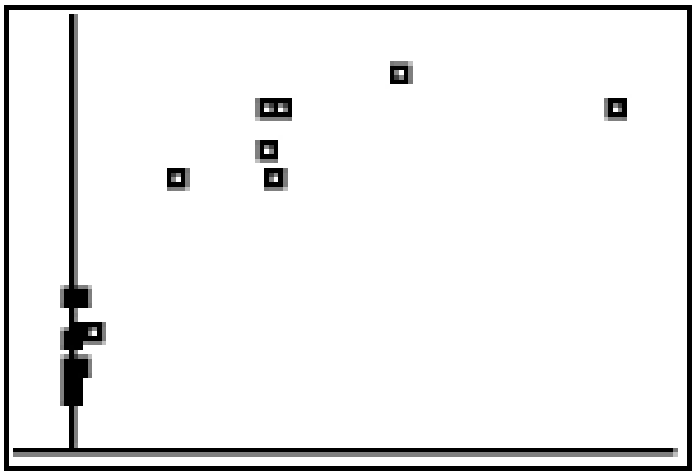
a. This transformation looks like it is possibly linear, thought there could be some curve to it as well. (But there are a lot of the points clustered close to each other near the point (0,0) that it is difficult to tell any trend).



b. This plot looks more linear than the plot in part a. It still could have a slight curvature to it and not be truly linear, however.

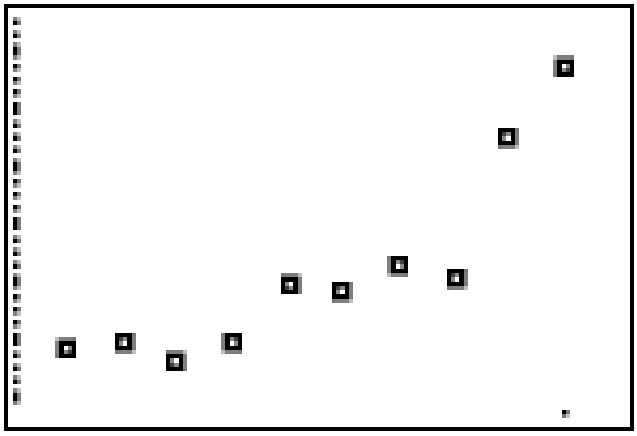


c. No, this seems to exaggerate the curve of the plot. The plot in part b appears to be the most linear.



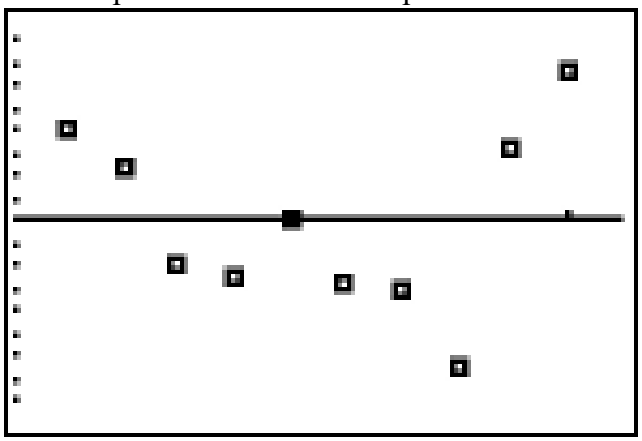
3.

a. The number of heart transplants over the years 2006 to 2015 appears to be increasing in a non-linear pattern. It appears to be increasing at an increasing rate. (Note that the y-axis is extremely distorted, it starts at $y = 2050$ and goes to $y = 2920$ – this makes the non-linear nature of the trend more obvious).

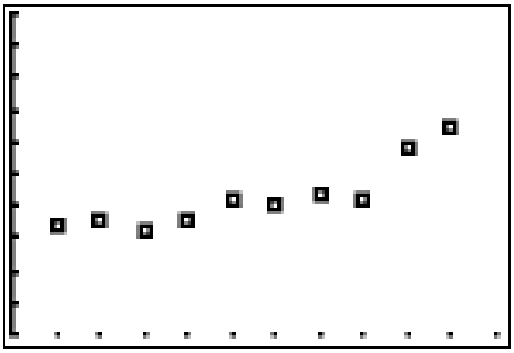


b. $\hat{y} = 2026.733 + 60.92x$. While $r = 0.877$, the nature of the pattern in the scatterplot indicates that a linear model may not be the best description of the data.

c. The residual plot makes it look like transformation could be appropriate, because the residual plot looks like it has a pattern to it.



d. Since I assumed a quadratic relationship, I am going to try plotting x against $\sqrt{y}$. This looks a bit more linear, but still not completely linear.

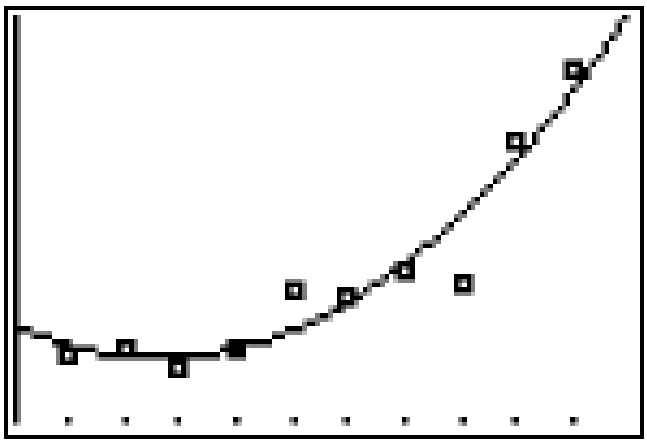


e. $\widehat{\sqrt{y}} = 45.163 + 0.617x$. For this regression, $r = 0.882$, so it is slightly better than the linear trend's r value. The predicted average number of heart surgeries in 2016 would be 2699.

f. I must be willing to assume that the trend over the past ten years continues for at least one more year. This assumption may not be valid for 2036 because projecting a trend from only 10 years of data to 20 years past the data set does not seem reasonable. There could be such a variety of reasons for the increase in surgeries over a short period that may not be valid over the longer period of time.

(Note: the best regression for this would likely be quadratic, and not transformed – below is the quadratic regression equation and graph. This is not covered on the AP test, but it is useful to see that there are a lot of varieties of regression beyond linear ones).

$$\hat{y} = 10.655x^2 - 56.287x + 2261.15$$



For this regression, $r^2 = 0.919$ and $r = 0.959$, which indicates that this is likely a very good fit.