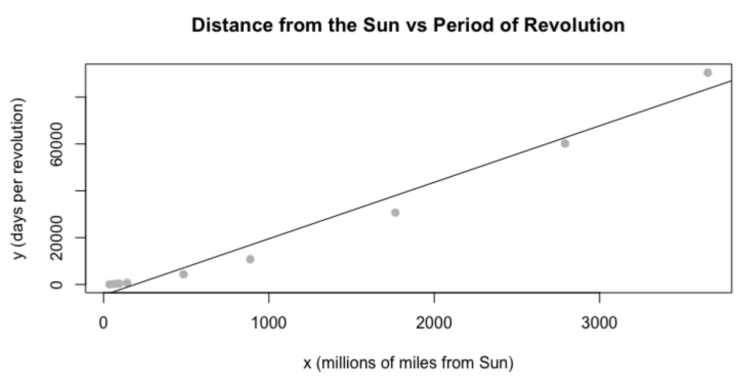
MATH 140 **Name** ________________________ **Date:**
Worksheet: Regression II

The Scene: The following table lists the average distance from the sun (in millions of miles), and period of revolution (Period) around the sun (in Earth days) of the nine planets in the solar system (including Pluto!). Below the table is a scatterplot of the data, produced in RStudio, which includes the least-squares regression line.

	planet	distance (x)	revolution (y)	Predicted revolution ($\hat{y}$)	residual ($y - \hat{y}$)
1	Mercury	36	88		
2	Venus	67	225		
3	Earth	93	365		
4	Mars	142	687		
5	Jupiter	484	4332		
6	Saturn	887	10760		
7	Uranus	1765	30684		
8	Neptune	2791	60188		
9	Pluto	3654	90467		

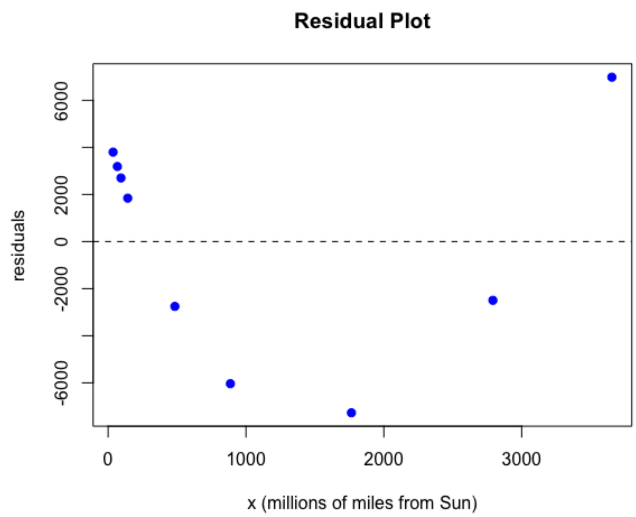


1. In a sentence or two describe the relationship between a planet's distance from the Sun and its period of revolution.

2. Using RStudio we find that r^2 for these data is 0.978. Quite high! Does this value seem to indicate a strong linear relationship between distance and period of revolution? Explain in a sentence or two.

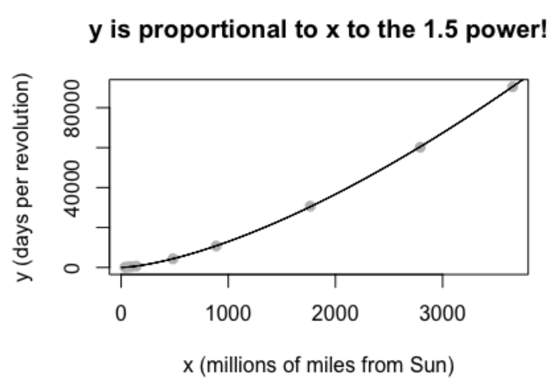
3. Compute the predicted revolution periods for each of the 9 planets using the best fit line, which is
$$\hat{y} = 24.1x - 4578.8.$$
Then compute the residuals for each of the planets. Record your results in the last 2 columns of the table at the start of this worksheet.

4. Below is a scatterplot of residuals vs distance, this plot is called a residual plot. Does the residual plot reveal any pattern? What does this tell us about the goodness of a linear model for the relationship between period and distance? Explain.



When a straight line is a reasonable model, the residual plot should reveal a seemingly random scattering of points. When the residual plot reveals a pattern of some kind, as is the case here, a non-linear model would fit the data better.

5. It turns out that a planet's period (y) is not a *linear* function of its distance from the Sun (x), but rather y is proportional to $x^{1.5}$! This power, 1.5, reveals itself when one finds the least-squares regression line of the log of the data. The plot below shows the original data with the curve $\hat{y} = 0.41x^{1.5}$.



For Mars and Neptune, predict y from x using the equation $\hat{y} = 0.41x^{1.5}$.

(a) Period of revolution for Mars, as predicted by the polynomial: ________________

(b) Period of revolution for Neptune, as predicted by the polynomial: ________________

Discussion Question: Are we overfitting the data? With such a small sample, perhaps this curve overfits the data, and wouldn't continue to be a good model for the relationship between a planet's distance from its Sun, and how long its year is. It turns out that the 1.5 power relationship is correct, and this relationship, observed by Copernicus and Kepler, was, in fact, mathematically derived by Sir Isaac Newton using the methods of Calculus he established!

	planet	distance	revolution	pred_y	residual	powerpred	res2
1	Mercury	36	88	-3711.20	3799.20	87.48	-0.52
2	Venus	67	225	-2964.10	3189.10	222.11	-2.89
3	Earth	93	365	-2337.50	2702.50	363.23	-1.77
4	Mars	142	687	-1156.60	1843.60	685.31	-1.69
5	Jupiter	484	4332	7085.60	-2753.60	4312.44	-19.56
6	Saturn	887	10760	16797.90	-6037.90	10698.93	-61.07
7	Uranus	1765	30684	37957.70	-7273.70	30031.16	-652.84
8	Neptune	2791	60188	62684.30	-2496.30	59716.56	-471.44
9	Pluto	3654	90467	83482.60	6984.40	89455.66	-1011.34