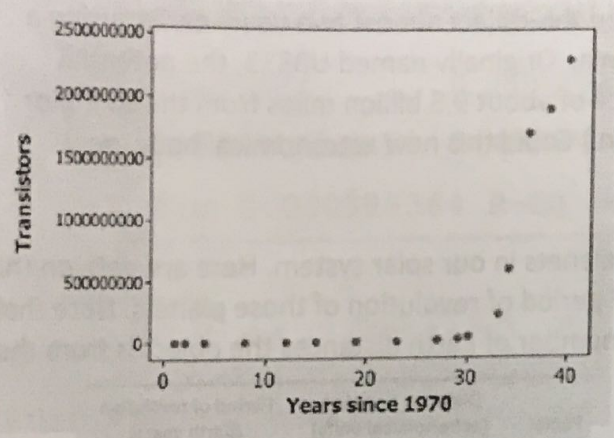
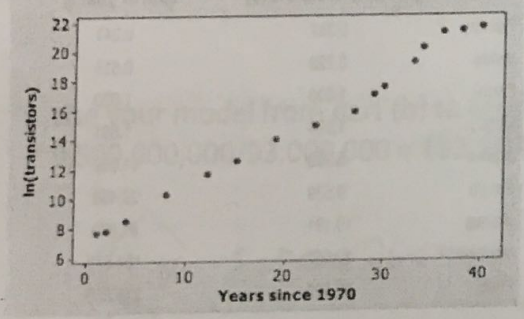
Example: Moore's Law & Computer Chips

Gordon Moore, one of the founders of Intel Corporation, predicted in 1965 that the number of transistors on an integrated circuit chip would double every 18 months. This is Moore's law, one way to measure the revolution in computing. Here are data on the dates and number of transistors for Intel microprocessors:





Processor	Date	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386	1985	275,000
486 DX	1989	1,180,000
Pentium	1993	3,100,000
Pentium II	1997	7,500,000
Pentium III	1999	24,000,000
Pentium 4	2000	42,000,000
Itanium 2	2003	220,000,000
Itanium 2 w/9MB cache	2004	592,000,000
Dual-core Itanium 2	2006	1,700,000,000
Six-core Xeon 7400	2008	1,900,000,000
8-core Xeon Nehalem-EX	2010	2,300,000,000

(a) A scatterplot of the natural logarithm (log base *e* or ln) of the number of transistors on a computer chip versus years since 1970 is shown. Based on this graph, explain why it would be reasonable to use an exponential model to describe the relationship between number of transistors and years since 1970.

It is not linear, when doing ln of y vs x data becomes more linear. also if you look at the residual plot.

(b) Minitab output from a linear regression analysis on the transformed data is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	7.0647	0.2672	26.44	0.000
Years since 1970	0.36583	0.01048	34.91	0.000

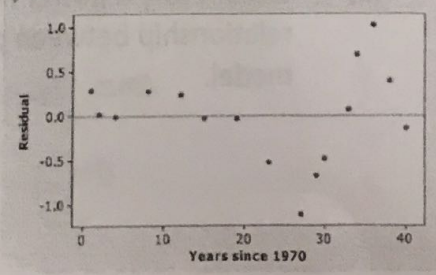
S = 0.544467 R-Sq = 98.9% R-Sq(adj) = 98.8%

$\ln \hat{y} = 7.0647 + .36583x$

(c) Use your model from part (b) to predict the number of transistors on an Intel computer chip in 2020. Show your work.

$\ln \hat{y} = 25.3562$ 102,815,299,900

(d) A residual plot for the linear regression in part (b) is shown below. Discuss what this graph tells you about the appropriateness of the model.



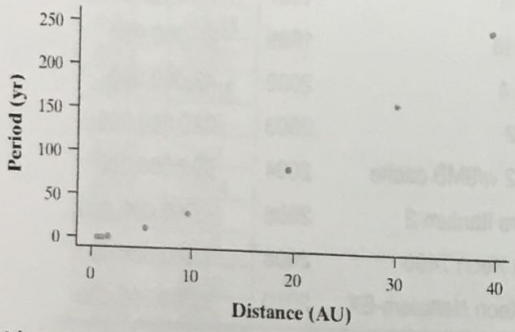
no pattern
exp is good as
data is linear now

Transforming to Achieve Linearity HW

What's a Planet, Anyway?

On July 31, 2005, a team of astronomers announced that they had discovered what appeared to be a new planet in our solar system. They had first observed this object almost two years earlier using a telescope at Caltech's Palomar Observatory in California. Originally named UB313, the potential planet is bigger than Pluto and has an average distance of about 9.5 billion miles from the sun. (For reference, Earth is about 93 million miles from the sun.) Could this new astronomical body, now called Eris, be a new planet?

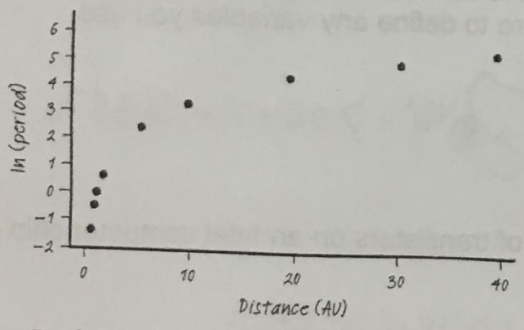
At the time of the discovery, there were nine known planets in our solar system. Here are data on the distance from the sun (in astronomical units, AU) and period of revolution of those planets. Note that distance is measured in astronomical units (AU), the number of Earth distances the object is from the sun.

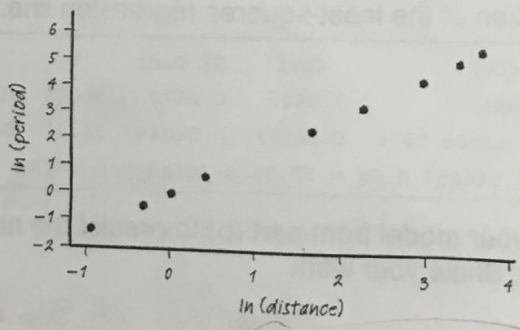


Planet	Distance from sun (astronomical units)	Period of revolution (Earth years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.456
Uranus	19.191	84.070
Neptune	30.061	164.810
Pluto	39.529	248.530

Above is a scatterplot of the planetary data. There appears to be a strong curved relationship between distance from the sun and period of revolution.

PROBLEM: The graphs below show the results of two different transformations of the data. The first graph plots the natural logarithm of period against distance from the sun for all nine planets. The second graph plots the natural logarithm of period against the natural logarithm of distance from the sun for the nine planets.





(a) Explain why a power model would provide a more appropriate description of the relationship between period of revolution and distance from the sun than an exponential model. Once transformed the data is more linear.

(b) Minitab output from a linear regression analysis on the transformed data in the second graph is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	0.0002544	0.0001759	1.45	0.191
ln(distance)	1.49986	0.00008	18598.27	0.000

S = 0.000393364 R-Sq = 100.0% R-Sq(adj) = 100.0%

$\widehat{\ln y} = .0002544 + 1.49986 \ln(x)$

(c) Use your model from part (b) to predict the period of revolution for Eris, which is 9,500,000,000/93,000,000 = 102.15 AU from the sun. Show your work.

$\widehat{\ln y} = .0002544 + 1.49986 \ln(102.15)$

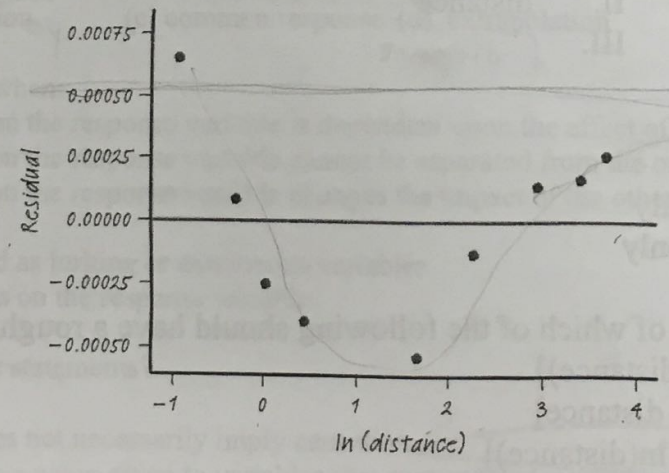
$\widehat{\ln y} = 6.9334$

4.626

1032.02 earth years

(d) A residual plot for the linear regression in part (b) is shown below. Do you expect your prediction in part (c) to be too high, too low, or about right? Justify your answer.

too high



Multiple Choice: Select the best answer for Exercises 1 to 4

exp

1. Suppose that the relationship between a response variable y and an explanatory variable x is modeled by $y = 2.7(0.316)^x$. Which of the following scatterplots would approximately follow a straight line?
   (a) A plot of y against x
   (b) A plot of y against $\log x$
   (c) A plot of $\log y$ against x
   (d) A plot of $\log y$ against $\log x$
   (e) None of (a) through (d)

Exercises 2 to 4 refer to the following setting. Some high school physics students dropped a ball and measured the distance fallen (in centimeters) at various times (in seconds) after its release. If you have studied physics, then you probably know that the theoretical relationship between the variables is distance = $490(\text{time})^2$. A scatterplot of the students' data showed a clear curved pattern.

$\hat{y} = 490(x)^2$ 226.576 Theo

2. At 0.68 seconds after release, the ball had fallen 220.4 centimeters. How much more or less did the ball fall than the theoretical model predicts? actual
   (a) More by 226.576 centimeters
   (b) More by 6.176 centimeters
   (c) No more and no less
   (d) Less by 226.576 centimeters
   (e) Less by 6.176 centimeters

3. Which of the following single transformations should linearize the relationship?
   I. time^2
   II. distance^2
   III. $\sqrt{\text{distance}}$

power

   (a) I only
   (b) II only
   (c) III only
   (d) I and II only
   (e) I and III only

2. A scatterplot of which of the following should have a roughly linear pattern?
   (a) [time, ln(distance)]
   (b) [ln(time), distance]
   (c) [ln(time), ln(distance)]
   (d) [ln(distance), time]
   (e) [distance, ln(time)]

Practice Multiple Choice

1. Suppose that the scatterplot of (log x, log y) shows a strong positive correlation. Which of the following must be true?

I. The variables x and y also have a correlation close to 1.
II. A scatterplot of (x, y) shows a strong nonlinear pattern.
III. The residual plot of the variables x and y shows a random pattern.

(a) I only (b) II only (c) III only (d) I and II (e) I, II, and III

2. What is the purpose of residual plots?
(a) To determine causation.
(b) To assess the type of relationship that exists between x and y.
(c) To check the appropriateness and fit of the regression equation for the data.
(d) To measure the variability in the residuals.
(e) To provide predictions for the response variable.

3. Fourth grade children were asked what emotion they associated with the color red. The responses for emotion and gender of the children are summarized in the following two-way table.

	Anger	Pain	Happiness	Love
Male	35	27	12	38
Female	27	17	19	39

What proportion of the males associate the color red with love?
(a) 0.5234 (b) 0.3598 (c) 0.3393 (d) 0.1822 (e) 0.1775

4. A strong negative linear relationship between Average State SAT scores and Percentage of students taking the SAT in each of those states reflects which underlying relationship?
(a) Causation (b) correlation (c) common response (d) extrapolation (e) confounding

5. Two variables are confounded when:
(a) The effect of one variable on the response variable is dependent upon the effect of the other variable.
(b) The effect of one variable on the response variable cannot be separated from the other variable.
(c) The effect of one variable on the response variable changes the impact of the other variable on the response variable.
(d) Both variables are classified as lurking or extraneous variables.
(e) They interact in their effects on the response variable.

6. Which of the following are true statements?

I. High correlation does not necessarily imply causation.
II. A lurking variable is a name given to variables that cannot be identified or explained.
III. Successful prediction requires a cause and effect relationship.

(a) I only (b) II only (c) III only (d) I and III only (e) I and II only

7. If the model for the relationship between the score on and AP Statistics Test (y) and the number of hours spent preparing for the test (x) was , determine the residual if a student studied 9 hours and earned an 85.

(a) 6.53 (b) 3.14 (c) 15.23 (d) 0 (e) –4.86

8. Which of the statements is true?

I. Two variables are confounded if their effects on a response variable cannot be distinguished from each other.

II. A lurking variable has an effect on the relationship among variables in the study but is not included among the variables studied.

III. Observational studies of the effect of one variable on another variable can fail if the explanatory variable is confounded with a lurking variable.

(a) I only (b) II only (c) III only (d) I and II only (e) I, II, and III

9. A study was conducted to determine the effectiveness of varying amounts of vitamin C in reducing the number of common colds. A survey of 450 people provided the following information:

	Daily amount of Vitamin C taken		
	None	500 mg	1000 mg
No colds	57	26	17
At least one cold	223	84	43

450

What conclusion can be made?

(a) The data proves that vitamin C reduces the number of common colds.

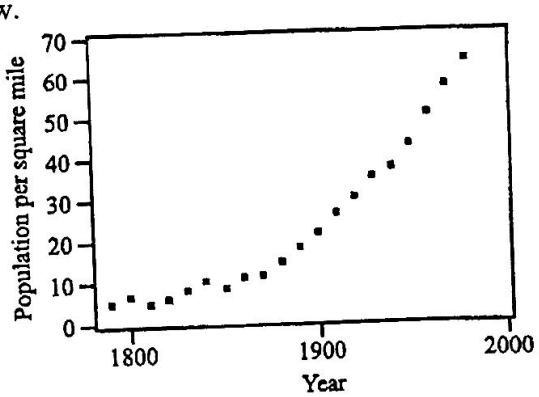
(b) The data proves that vitamin C has no effect on the number of common colds.

(c) There appears to be a strong association between consumption of vitamin C and the occurrence of common colds.

(d) There appears to be little association between consumption of vitamin C and the occurrence of colds.

(e) Since common colds are caused by viruses, there is no reason to conclude that vitamin C could have any effect.

10.) For the past two-hundred years population per square feet in a northwest suburb can be modeled using the exponential equation . The scatter plot of the data is shown below.



Which of the following statements is true?

(a) If an attempt is made at fitting a straight line to the original data, the corresponding residual plot would be approximately linear.

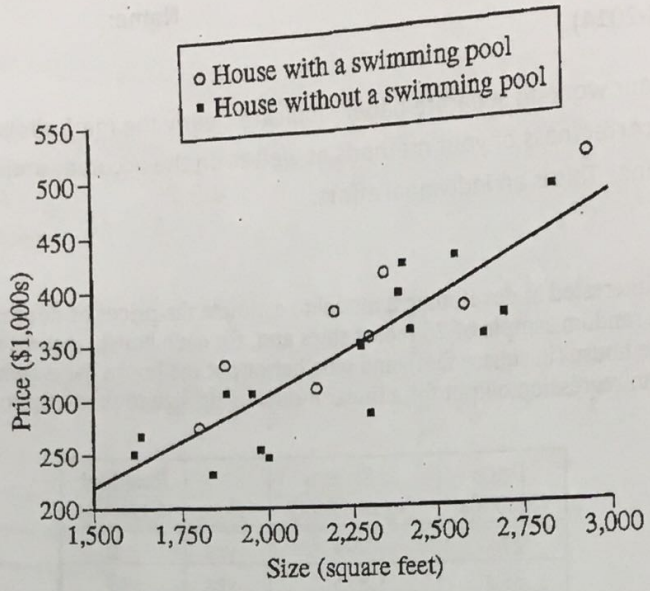
(b) If an attempt is made at fitting a straight line to the original data, the corresponding residual plot would be scattered and show no pattern.

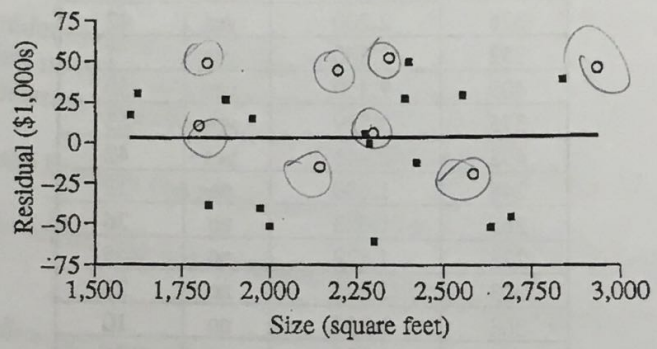
(c) If an attempt is made at fitting a straight line to the original data, the corresponding residual plot would be a straight line.

(d) Plotting the logarithm of population per square mile against year should be approximately linear.

(e) Plotting the logarithm of population per square mile against the logarithm of year should be approximately linear.

(a) 6.53





Linear Fit

Price = –28.144 + 0.165 Size

Summary of Fit

RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

(a) Interpret the slope of the least squares regression line in the context of the study.

As the size (in square feet) goes up by 1 the cost goes up by .165 (in thousands of dollars

obs - exp

(b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study.
was 49 (in thousands) dollars more than exp

The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool.

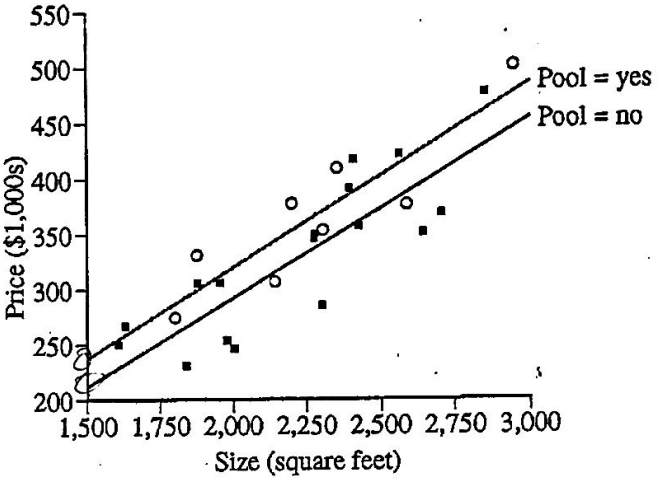
27.505 (in thousands more

To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.

Linear Fit (Pool = yes)
Price = −11.602 + 0.166 size

Linear Fit (Pool = no)
Price = −27.382 + 0.160 size

o House with a swimming pool
▪ House without a swimming pool



(d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is (−0.099, 0.110). Based on this interval, is there a significant difference in the two slopes? Explain your answer. no as 0 is contained in the interval to show no difference

(e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c)?

about 16 (thousands dollars) more per square foot