Straightening Data in a Scatterplot
Selecting a Good Re-Expression Model

What Is All This Stuff?

Here's what is included:

- Page 3: Graphs of the three main patterns of data points that the student is likely to encounter in scatter plots, along with suggestions on which re-expression models to try.
- Page 4: Graphs of residual plots that the student is likely to encounter, along with suggestions on which re-expression models to try.
- Pages 5 to 10: Scatter plots developed using various model types, their associated regression lines, re-expressions used to make the models linear, and some ideas on when to use the re-expressions.

Explanation

In an effort to assist the student in selecting a model to use in "straightening" data in a scatterplot, I developed the series of models described on pages 5 to 10. Using the patterns in the scatterplots and residual plots generated by these models, I created the summaries on pages 3 and 4.

It is not possible to identify and chart every pattern that may arise in a scatterplot. However, I am hopeful that reviewing the information in this package will help the student identify patterns and quickly determine which re-expressions make the most sense when data are not nearly linear.

If the student runs into a pattern different from those included in this packet, they should try to identify a pattern in the summaries that has similar characteristics. Sometimes it may be necessary for the student to draw on their knowledge of Algebra and functions to select an appropriate model. In any case, after re-expressing data, remember to check the residual plot for randomness and lack of a pattern to determine if the model you selected is a good one.

Why Is it Difficult to Pick a Model?

Rarely will it be obvious that a single model is much better than every other model. There will often be more than one model which makes a decent choice in "straightening" data. Why is this? The graphs on the next page may help.

Figure 1, below, summarizes a number of models that work with scatterplots where the data are simultaneously rising and flattening. Notice that all of these models have the same general pattern. And … there are many more models that have similar patterns.

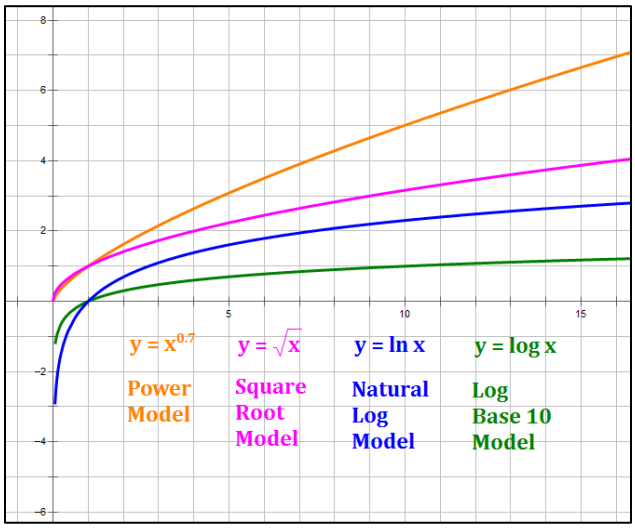


Figure 1

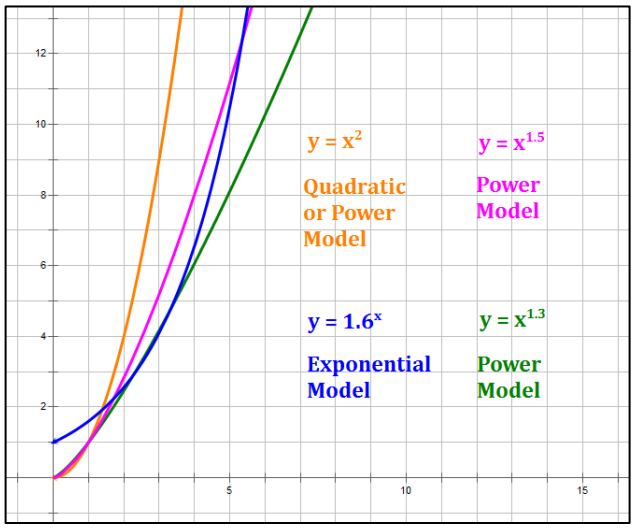


Figure 2

Figure 2, above, summarizes a number of models that work with scatterplots where the data are simultaneously rising and concave up. Notice that all of these models have the same general pattern. Again … there are many more models that have similar patterns.

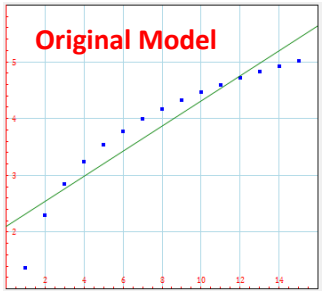
Finding the right model is difficult because there are so many possibilities. Additionally, the best model is not always the model with the best fit (i.e., the highest R^2). Selecting the right model often involves understanding the nature of the data, including what causes it to have its general shape. In hard science (e.g., chemistry, physics, biology, astronomy), fitting data with the proper model depends on the underlying scientific principles involved. In the social sciences (i.e., sciences relating to society and human behavior), however, it is often more of a challenge to select the right model for a given set of data; and, even if you find a good model, it may need to change over time.

The Log Models – A Safe Haven

It is noteworthy that three models involving logarithms (i.e., the logarithmic model, the power model and the exponential model) are very common. Notice in the summary on page 3 that at least one of these models fits each of the patterns shown. The student should become proficient at using these models and knowing when each should be used. When in doubt, try the Power model; it is very, uh … Power-ful. Best wishes!

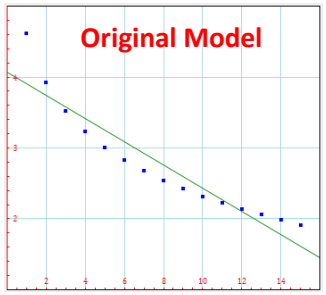
Earl

Re-Expression Models to Consider Based on the Pattern of Points in a Scatterplot



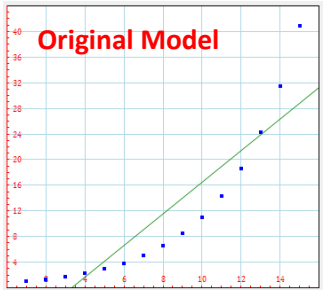
Models to consider:

- Logarithmic
  $\hat{y} = a + b \log x$
- Square Root
  $\hat{y}^2 = a + bx$
- Power
  $\log \hat{y} = a + b \log x$



Models to consider:

- Logarithmic
  $\hat{y} = a + b \log x$
- Exponential
  $\log \hat{y} = a + bx$
- Reciprocal
  $\frac{1}{\hat{y}} = a + bx$



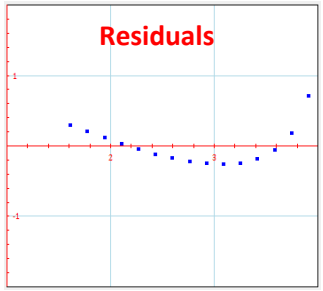
Models to consider:

- Quadratic (for counts)
  $\sqrt{\hat{y}} = a + bx$
- Exponential
  $\log \hat{y} = a + bx$
- Power
  $\log \hat{y} = a + b \log x$

Summary of When to Use Specific Models

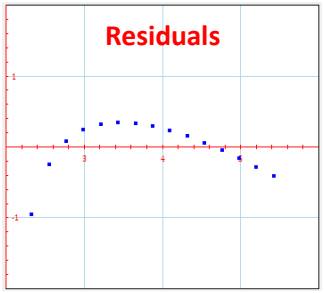
- **Logarithmic Model:** $\hat{y} = a + b \log x$ Use when values flatten out on the right.
- **Exponential Model:** $\log \hat{y} = a + bx$ Use when values increase or decrease at a constant percentage rate.
- **Power Model:** $\log \hat{y} = a + b \log x$ Similar to the Square Root and Quadratic Models, but allows powers other than ½ or 2. A very nice feature is that the regression determines the appropriate power (i.e., the value of b) to be used in the model.
- **Square Root Model:** $\hat{y}^2 = a + bx$ For values that look like a square root function.
- **Quadratic Model:** $\sqrt{\hat{y}} = a + bx$ Start here when counts are involved.
- **Reciprocal Model:** $\frac{1}{\hat{y}} = a + bx$ Use when values are ratios, like "miles per hour." Alternatively, invert the ratio and try a linear model.

Re-Expression Models to Consider Based on a Graph of the Residuals



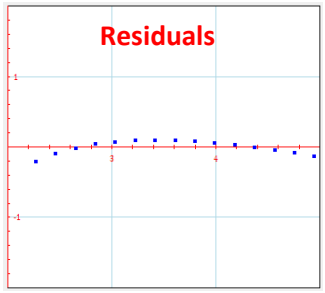
Models to consider:

- Logarithmic
  $\hat{y} = a + b \log x$



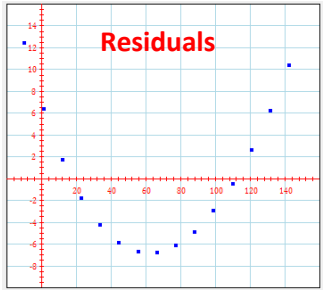
Models to consider:

- Logarithmic
  $\hat{y} = a + b \log x$



Models to consider:

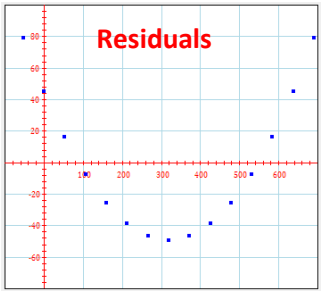
- Square Root
  $\hat{y}^2 = a + bx$
- Power
  $\log \hat{y} = a + b \log x$



Bottom
Tilts Left

Models to consider:

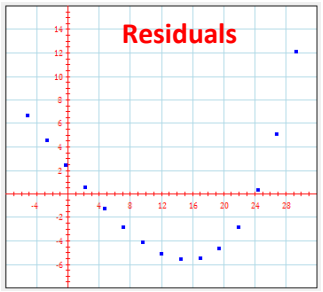
- Power
  $\log \hat{y} = a + b \log x$



Bottom
Centered

Models to consider:

- Quadratic
  $\sqrt{\hat{y}} = a + bx$
- Power
  $\log \hat{y} = a + b \log x$



Bottom
Tilts Right

Models to consider:

- Exponential
  $\log \hat{y} = a + bx$
- Reciprocal
  $\frac{1}{\hat{y}} = .024 + .06x$

Straightening Data in a Scatterplot
Description of Individual Models

The following pages provide the background material used to create the summaries on pages 3 and 4. It is not necessary to study the detail for each of these models, but some familiarity with situations in which they should be used (green boxes) is recommended.

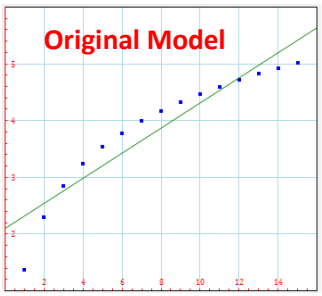
I attempted to choose models that the student may encounter on the AP Exam. Other models are certainly possible, so this list should not be considered exhaustive. Two models are shown on each page. Here is an explanation of what is shown for each model.

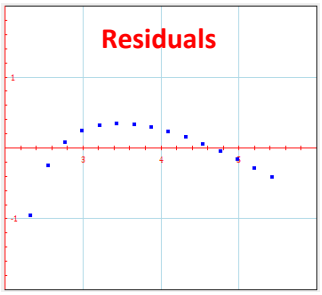
- **Top Line:** Name of the model and the function I used to generate the data points. The specific function I used is not of major importance to the student; rather, the pattern of the points in the scatterplot and in the residual plot should be noted.
- **Top Two Graphs:** These are the scatterplot and residual plot generated by the function chosen. <u>These are very important</u> and represent the kind of patterns the student should look for when choosing a re-expression model. Note that, for consistency, I used x-values from 1 to 15 in every model. The regression equation associated with the scatterplot is also shown.

  Note that the residual plots shown in this packet are based on the definition of residual plots in your textbook. A point is plotted for each point in the original scatterplot. The abscissa (i.e., horizontal coordinate) of each point is the predicted value, $\hat{y}$, associated with the corresponding point in the original scatterplot, and the ordinate (i.e., vertical coordinate) of each point is the residual, $y - \hat{y}$, associated with that point. So, you can think of each ordered pair in the residual plot as having the coordinates: $(\hat{y},\ y - \hat{y})$.
- **Action Line:** A description of the action to be taken by the student to produce a data re-expression that will attempt to "straighten" the data.
- **Bottom Graph:** A scatterplot of the re-expressed data. Note that the re-expressed data for each model in this packet all lie on straight lines; this results directly from the manner in which I created the original scatterplots. *In your work, you will want the re-expression to produce 1) a scatterplot with points that are close to a straight line, and 2) a residual plot with randomly distributed points that have no apparent pattern.*

  I show the general description of the re-expression model being used and the regression equation for the re-expressed data to the right of the graph. Note that this regression equation is equivalent to the sample function shown in the top line; this also results directly from the manner in which I created the original scatterplots.
- **Green Box:** Comments on when the model should be used and anything else I found particularly interesting in developing this packet.

Logarithmic Model sample function: $y = 3.1\log(x) + 1.37$

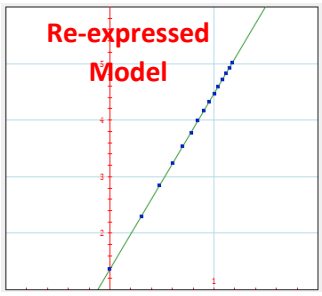




Regression Equation:

$\hat{y} = 2.10 + 0.22x$

Action: Change x-axis values from x to $\log(x)$



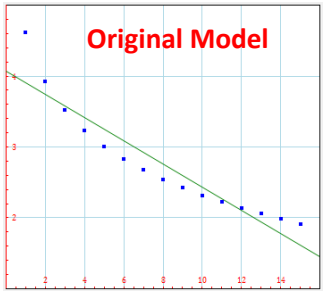
Model: $\hat{y} = a + b\log x$

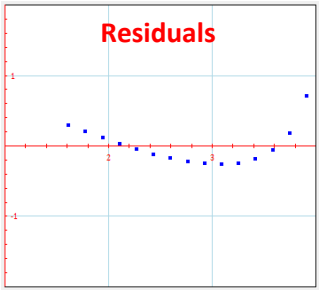
Regression Equation:

$\hat{y} = 1.37 + 3.10 \cdot \log x$

Use when values flatten out on the right.

Logarithmic Model sample function: $y = -2.3\log(x) + 4.62]$

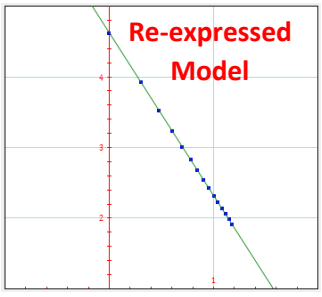




Regression Equation:

$\hat{y} = 4.07 - 0.16x$

Action: Change x-axis values from x to $\log(x)$



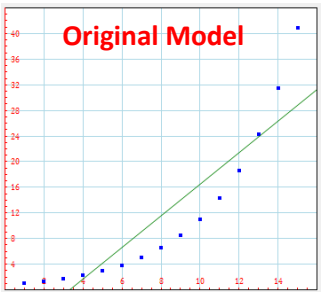
Model: $\hat{y} = a + b\log x$

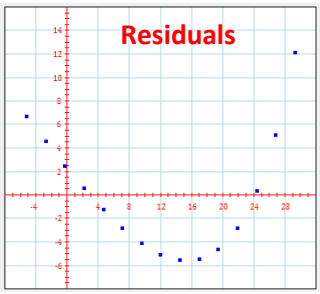
Regression Equation:

$\hat{y} = 4.62 - 2.30 \cdot \log x$

Use when values flatten out on the right.

Exponential Model sample function: $y = 0.8 \cdot (1.3)^x$

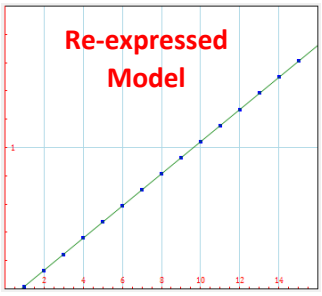




Regression Equation:

$\hat{y} = -8.11 + 2.46x$

Action: Change y-axis values from y to $\log(y)$



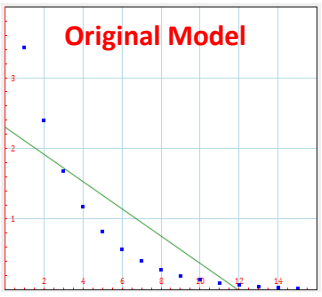
Model: $\log \hat{y} = a + bx$

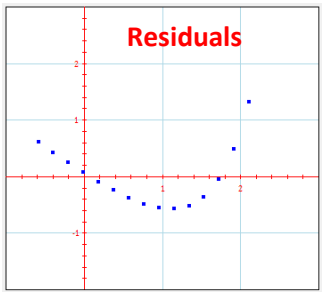
Regression Equation:

$\log \hat{y} = -0.1 + 0.11x$

Use when values increase or decrease at a constant percentage rate.

Exponential Model sample function: $y = 4.9 \cdot (0.7)^x$

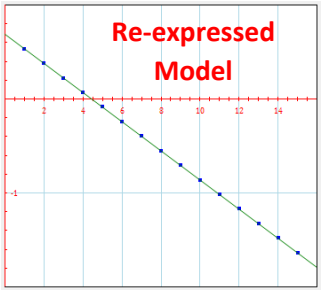




Regression Equation:

$\hat{y} = 2.30 - 0.19x$

Action: Change y-axis values from y to $\log(y)$



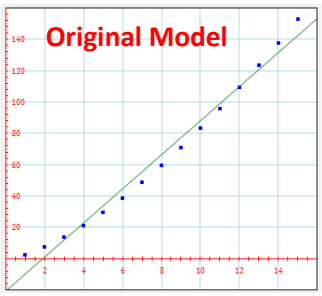
Model: $\log \hat{y} = a + bx$

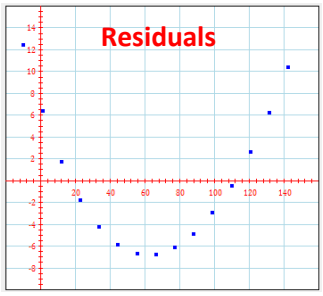
Regression Equation:

$\log \hat{y} = 0.69 - 0.15x$

Use when values increase or decrease at a constant percentage rate.

Power Model sample function: $y = 2.63 \cdot x^{1.5}$

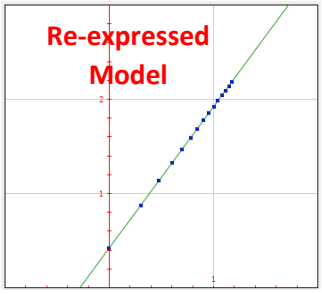




Regression Equation:

$\hat{y} = -20.72 + 10.88x$

Action: Change x-axis values from x to $\log(x)$ AND y-axis values from y to $\log(y)$



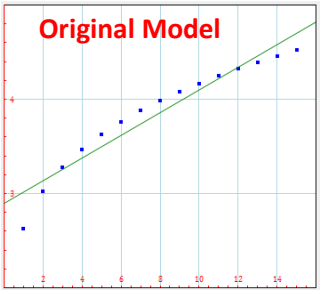
Model: $\log \hat{y} = a + b \log x$

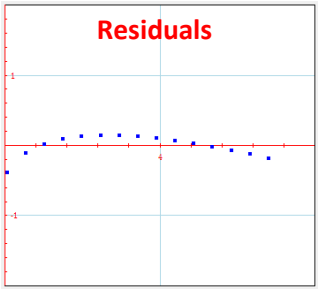
Regression Equation:

$\log \hat{y} = 0.42 + 1.50 \cdot \log x$

Similar to the Square Root and Quadratic Models, but allows powers other than ½ or 2. A very nice feature is that the regression determines the appropriate power (i.e., value of b) to be used in the model.

Power Model sample function: $y = 2.63 \cdot x^{0.2}$

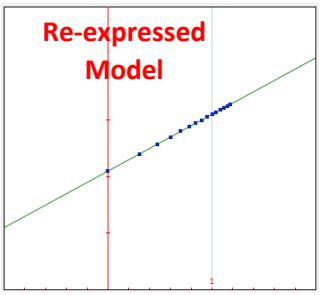




Regression Equation:

$\hat{y} = 2.89 + 0.12x$

Action: Change x-axis values from x to $\log(x)$ AND y-axis values from y to $\log(y)$



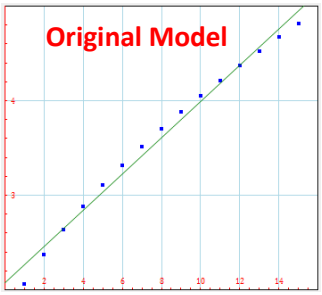
Model: $\log \hat{y} = a + b \log x$

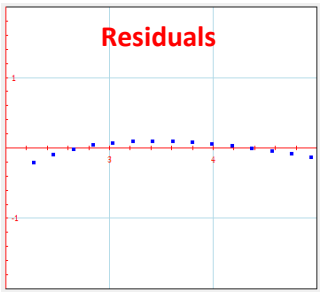
Regression Equation:

$\log \hat{y} = 0.42 + 0.2 \cdot \log x$

Similar to the Square Root and Quadratic Models, but allows powers other than ½ or 2. A very nice feature is that the regression determines the appropriate power (i.e., value of b) to be used in the model.

Square Root Model sample function: $y = \sqrt{1.35x + 2.91}$

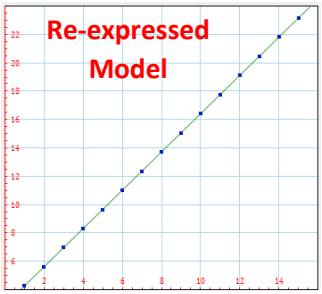




Regression Equation:

$\hat{y} = 2.08 + 0.19x$

Action: Change y-axis values from y to y^2



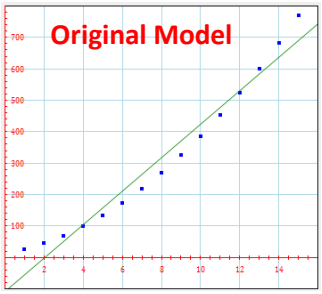
Model: $\hat{y}^2 = a + bx$

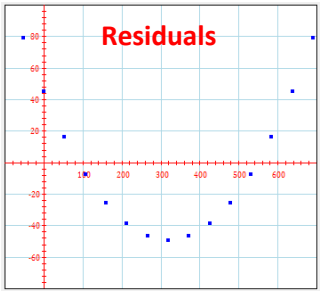
Regression Equation:

$\hat{y}^2 = 2.91 + 1.35x$

For values that look like a square root function.

Quadratic Model sample function: $y = (1.62x + 3.44)^2$

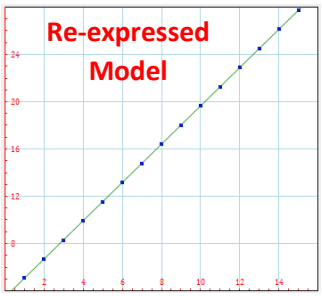




Regression Equation:

$\hat{y} = -107.14 + 53.14x$

Action: Change y-axis values from y to $\sqrt{y}$



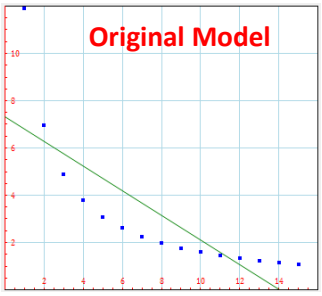
Model: $\sqrt{\hat{y}} = a + bx$

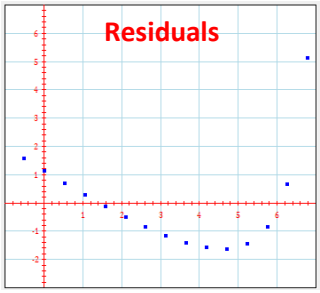
Regression Equation:

$\sqrt{\hat{y}} = 3.44 + 1.62x$

Start here when counts are involved.

Reciprocal Model sample function: $y = \frac{50}{3x+1.2}$





Regression Equation:

$\hat{y} = 7.31 - 0.52x$

Action: Change x-axis values from y to $\frac{1}{y}$

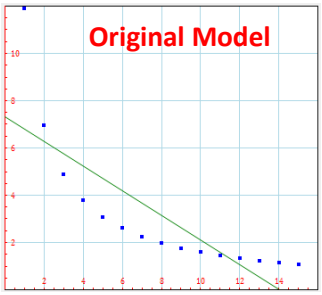
Model: $\frac{1}{\hat{y}} = a + bx$

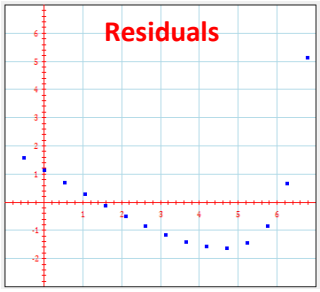
Regression Equation:

$\frac{1}{\hat{y}} = .024 + .06x$

Use when values are ratios, like "miles per hour." Alternatively, invert the ratio and trv a linear model.

Negative Reciprocal Model sample function: $y = \frac{50}{3x+1.2}$

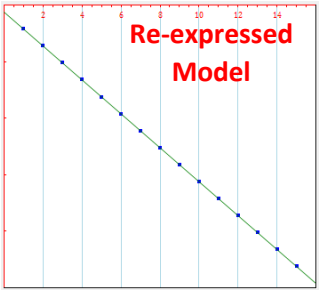




Regression Equation:

$\hat{y} = 7.31 - 0.52x$

Action: Change x-axis values from y to $\frac{-1}{y}$



Model: $\frac{-1}{\hat{y}} = a + bx$

Regression Equation:

$\frac{-1}{\hat{y}} = -.024 - .06x$

Like the Reciprocal Model, but preserves the direction of the original curve.