

Re-expression and Linear Regression

In this session, we will explore the important connection between re-expression or linearization of a function and functional models based on regression techniques. I assume that the techniques of regression itself are well understood, and that the computations and graphical presentation of the residuals is also understood. We will begin from that basis. Statisticians consider all polynomial functions as linear models, since you solve linear equations to find the coefficients. So $y = a + bx$ and $y = a + bx + cx^2$ are both linear models. In this presentation, I will use the more familiar terms of linear functions for $y = a + bx$ and non-linear function for $y = a + bx + cx^2$.

Mathematical Principles of Re-expression

Both data sets presented below were collected in the kitchen one summer afternoon and represent two phenomena with which students should be familiar. The first is the relationship between the depth of water in an urn and time, as the urn is being emptied. The second describes the relationship between the temperature of a cup of coffee and the time that the cup has been sitting on the table cooling. The temperature of the room at the time these measurements were taken was approximately 81 °F.

Time (secs)	0	30	60	100	140	180	210	265	330	390
Depth (in)	12.5	11.25	10.5	9.25	8.0	7.0	6.0	4.75	3.5	2.5

Table 1: Time and Depth of water in an urn

Time (min)	15	22	24	27	32	37	40	50	64
Temp (°F)	130	122	119	116	112	108	106	100	94

Table 2: Time and Temperature of coffee

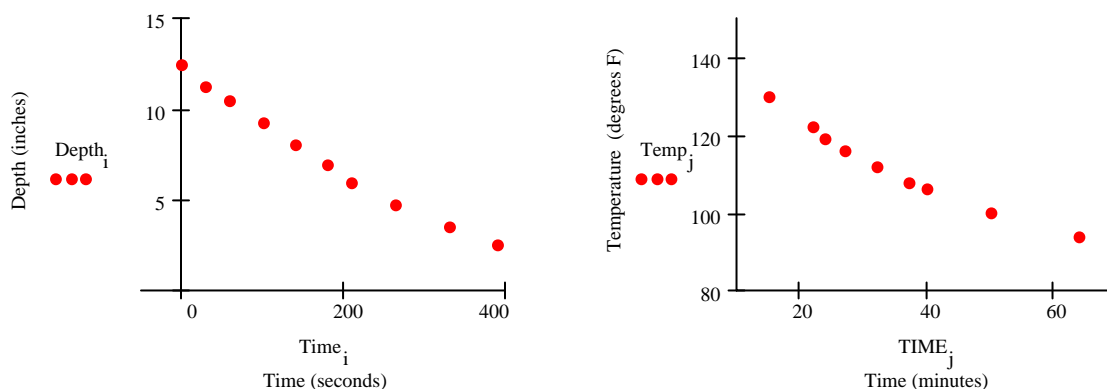


Figure 1: Plots of Depth vs. Time and Temperature vs. Time

The first question we ask is, “Is the relationship between the variables linear?” Students can fit lines to data and partition the phenomena they represent into linear and non-linear categories. If the data is linear, they can interpret the equation of the line in the context of the phenomena. For the two phenomena being investigated here, students should be able to argue

from their personal experience that the relationship between the variables is non-linear. At some time in their lives, most students have gotten a drink from an urn with a stop-cock drain at the bottom. They recognize that a cup fills quickly if the urn is nearly full and very slowly if the urn is nearly empty. If the relationship is linear, then the change in the depth of the liquid in the urn over any fixed interval of time will be the same (to within measurement accuracy). Similarly, the change in temperature of a cup of coffee over any fixed interval of time will be constant. However, students know that a cup of coffee will lose more heat in the first 10 minutes than it will in the interval between 20 and 30 minutes.

If students can judge whether the urn is nearly full or nearly empty by how quickly the cup fills, then the relationship cannot be linear. If the change in the dependent variable allows students to predict the value of the independent variable, then the relationship cannot be linear. As one student told me, “Being linear means that you don’t know where you are.” This is an important way to think about linear functions.

Data analysis offers support to these experiential arguments. If we fit a line to the data, the curved shape of the data becomes apparent as the characteristic U-shaped residual plot indicates curvature in the data.

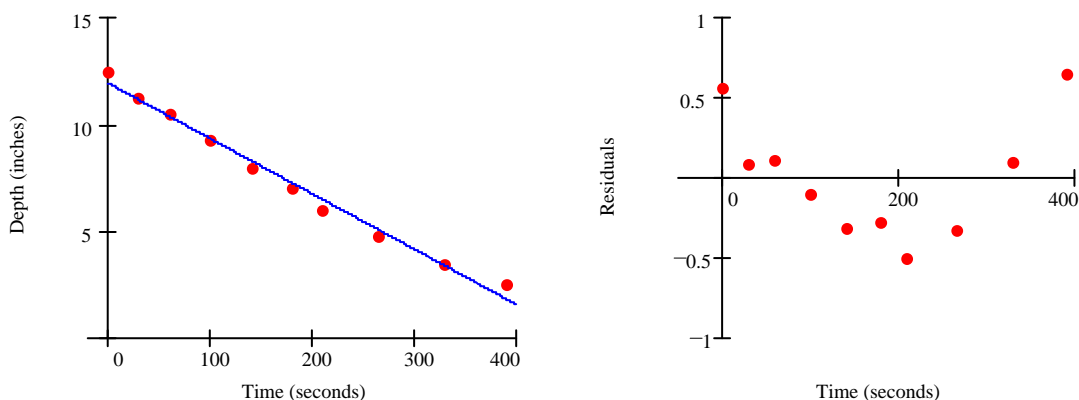


Figure 2: Linear Model of Depth and Time with Residuals

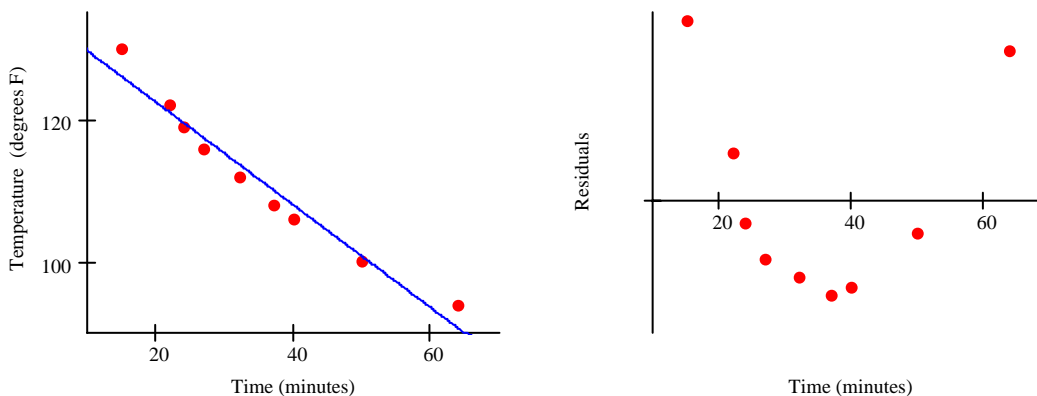


Figure 3: Linear model of Temperature and Time with Residuals

Clearly, both data sets represent non-linear phenomena.

Automatic Graphing

Why not use the automatic graphing capabilities of your calculator and just fit quadratic functions? The least squares quadratic fit for the relationship between time and the depth of

lemonade in the urn is $D = 0.00002542t^2 - 0.03556t + 12.46$, and for the relationship between time and the temperature of a cup of coffee $T = 0.009735t^2 - 1.490t + 149.8$. Notice both models do a good job of describing the data (see Figure 7).

In second year algebra, students study two basic forms for quadratic functions, $y = ax^2 + bx + c$ and $y = a(x-h)^2 + k$. In the first form, the initial condition is easy to see. At $t = 0$, the depth of lemonade is estimated at 12.5 inches and the temperature of the coffee around 150 degrees. This is already cool for coffee, so, if our model is correct, the coffee had already been cooling for a while before our data gathering began.

In the second form, we can read off the location of the vertex. If we rewrite these two quadratics, we find that for the lemonade data the quadratic is very close to the perfect square, $D = 0.00002542(t - 699.5)^2$. The vertex is $(699.5, 0)$, and we interpret this to mean that after approximately 700 seconds, the urn will be empty. For the coffee problem, completing square produces the quadratic $T = 0.009735(t - 76.53)^2 + 92.79$. The vertex is at $(76.53, 92.79)$, which we interpret to mean that the terminal temperature is around 93 degrees and is reached in a little over 75 minutes. Students might be worried about this second equation based on the knowledge that the room temperature at the time the data was taken was around 81 degrees.

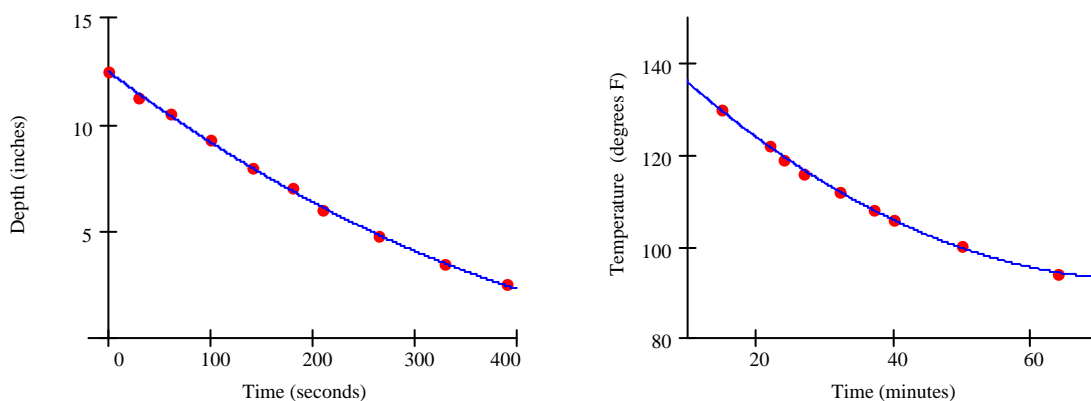


Figure 4: Automatic Least Squares Quadratic Fits

Be cautious about quadratic fits. Most simple functions can be approximated quite well over a limited domain with a quadratic. Never have too much confidence in automatic quadratic fits like these.

Re-expressing to Linearize Data

A more refined and informative modeling technique than just fitting quadratics is to re-express and linearize the data. To re-express the data, students must understand clearly the effects of transformations on the shapes of the basic functions. The lemonade data cannot represent values from the quadratic functions $y = ax^2$ or $y = ax^2 + b$, but could be a portion of a quadratic with a horizontal shift, $y = (ax - b)^2$, since we expect the vertex to lie on the x -axis (time at which the urn is empty).

If $y = (ax - b)^2$, then $\sqrt{y} = \pm(ax - b)$, that is, the data set (x, \sqrt{y}) is linear with a slope of $\pm a$ and intercept of $\mp b$.

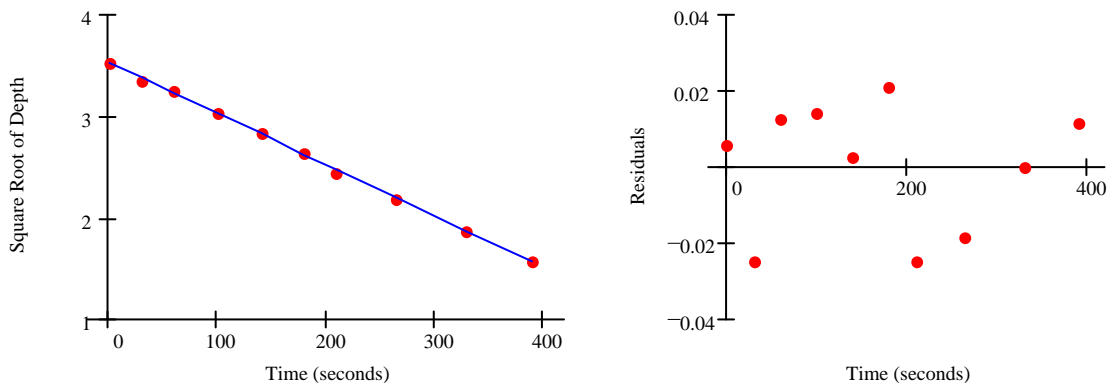


Figure 5: $Y = -0.005026X + 3.530$, with where $Y = \sqrt{D}$ and $X = t$

The least squares line for the data set (t, \sqrt{D}) is $Y = -0.005026X + 3.530$, where $Y = \sqrt{D}$ and $X = t$. This re-expression generates the model $\sqrt{D} = -0.005026t + 3.530$, or $D = (-0.005026t + 3.530)^2$. Compare this to the previous quadratic model.

For the coffee temperature problem, we noted earlier that the quadratic gave us an unsatisfactory final temperature of 92 degrees. Our experience tells us that the final temperature will approach the temperature of the room, or around 81 degrees, so our quadratic model is questionable. Instead of quadratic, if the data represent an exponential function, then the form should be $T = ae^{-bt} + 81$.

To linearize data that follows this functional form, we can start by taking logarithms of both sides of the equation. So, if $T = ae^{-bt} + 81$, then $\ln(T) = \ln(ae^{-bt} + 81)$. Unfortunately, logarithms don't simplify sums, so we can't linearize the data this way. The vertical shift has created a problem that logs can't solve. So, we first must remove the vertical shift.

If if $T = ae^{-bt} + 81$, then $T - 81 = ae^{-bt}$. Now, $\ln(T - 81) = \ln(ae^{-bt})$ and using the rules of logs on the right-hand side, gives us $\ln(T - 81) = \ln(a) - bt$. This means the ordered pairs $(t, \ln(T - 81))$ will be linear. Our semi-log re-expression should linearize the data.

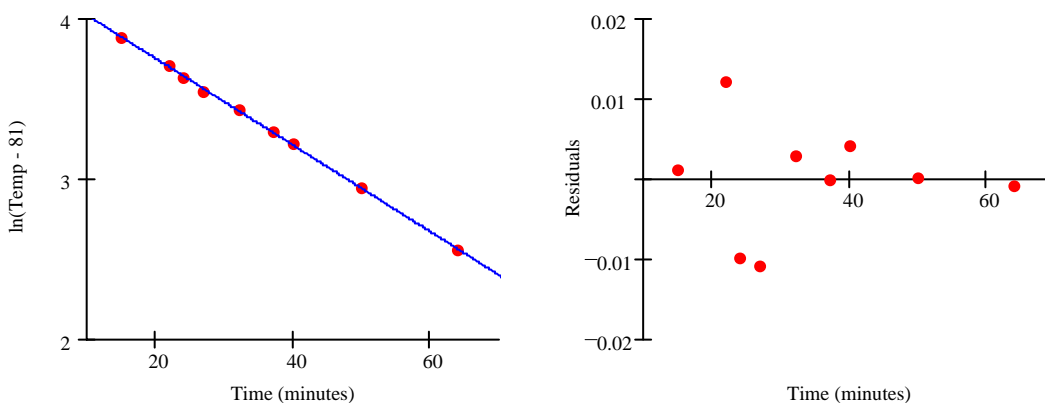


Figure6: Semi-log re-expression with $Y = -0.02704X + 4.296$

If $\ln(T - 81) = -0.02704t + 4.296$, then $T = 73.341e^{-0.02704t} + 81$.

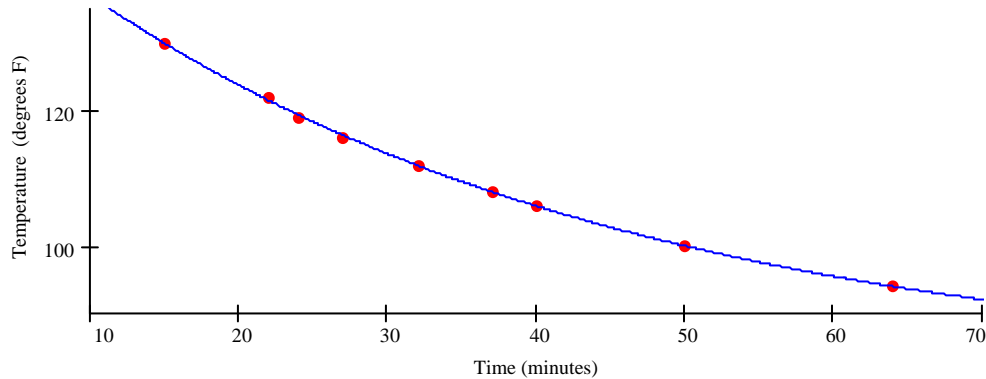


Figure 7: $T = 73.341e^{-0.02704t} + 81$ plotted against the original data

Does this exponential model fit the data better than the quadratic model we used before? No. In fact, it may not fit as well, since we have only two parameters to work with rather than three in the quadratic model. However, we believe this is an example of exponential behavior *because the semi-log re-expression linearized the data*. That is our evidence for exponential decay.

The Mathematical Principle

High school students are taught that $f^{-1}(f(x)) = x$, over the domain of f . When we compose a function with its inverse, we obtain the linear function $y = x$. In the case of re-expression to linearize data, we don't require the specific line $y = x$, we only require a linear function. So we don't need to find *the* inverse of f , only a kind of "semi-inverse", a function g where $g(f(x)) = mx + b$. Since we don't know f (that's what we're trying to figure out), but we do know g , we can solve for f as $f(x) = g^{-1}(mx + b)$. Linearization is just an application, but a very powerful application, of composition of functions and inverse functions. If you look back at the two examples given, we have simply composed the unknown function f with what we expect to be its inverse g , so $g(f(x)) = mx + b$. For quadratic behavior, we used square roots. For exponential behavior, we used logs. The square root re-expression won't linearize the Coffee Cooling data and the logarithmic re-expression won't linearize the Depth in the Urn data.

Predicting the 2nd Half of the Data from the 1st Half

As one final argument for not using the automatic fitting features of the calculator when investigating unknown data, consider the following example. Table 3 repeats the Coffee Cooling data.

Time (min)	15	22	24	27	32	37	40	50	64
Temp (°F)	130	122	119	116	112	108	106	100	94

Table 3: Time and Temperature of coffee

Now, suppose we use only the first 4 data points to fit a model. If our model truly captures the underlying functional relationship between the two variables, then it should do a reasonable job

of fitting the rest of the data too (since they all live on the same function). If, however, our model just does a good job of fitting the data, but doesn't capture the underlying functional relationship, then it should not be expected to fit the rest of the data.

Using just the four points

Time (min)	15	22	24	27					
Temp (°F)	130	122	119	116					

to fit a quadratic function using the automatic function fitting capabilities of the calculator, we find $Q(x) = 0.00189x^2 - 1.26x + 148.5$. Repeating the semi-log re-expression (first subtracting 81), we find $E(x) = 75e^{-0.0281x} + 81$. Both functions fit these 4 points well.

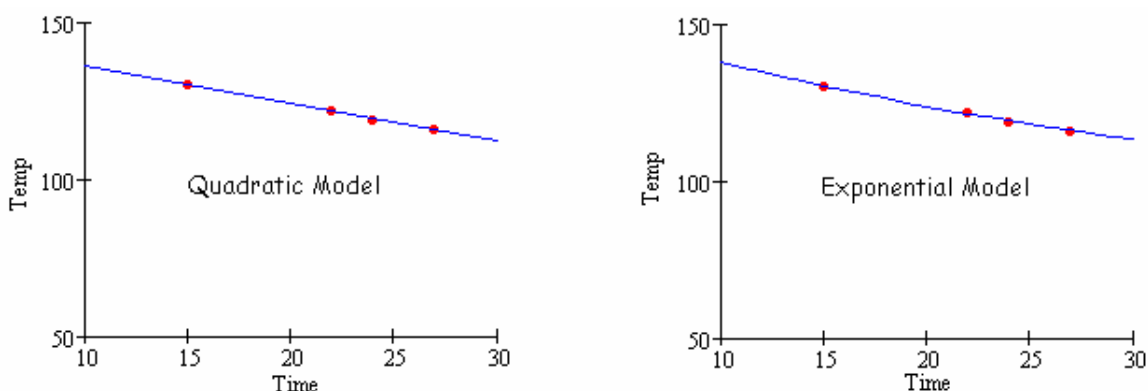


Figure 8: Both Models Through 1st Four Data Points

But, how do they fit the rest of the data set?

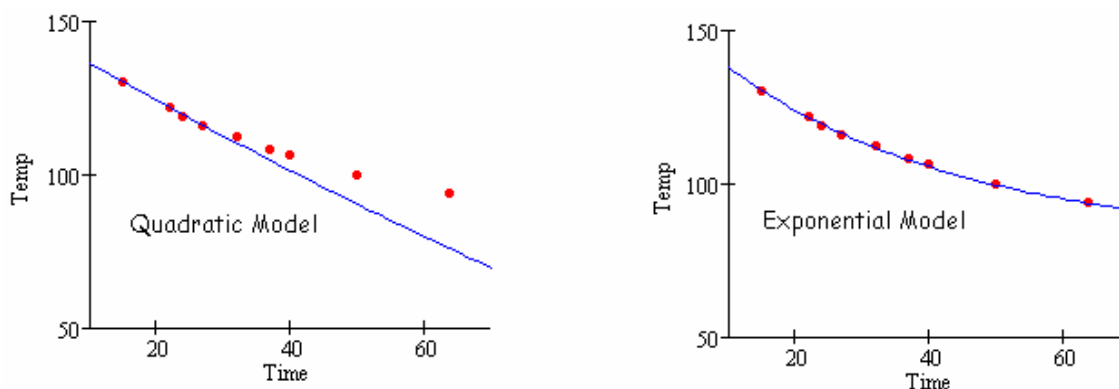


Figure 9: Both Models Through All Data Points

We can see from the graphs above, that the quadratic model did not capture the underlying functional form. It just fit the given data well. The exponential model is correct. It captures the underlying functional form of the data. We knew it would, since we were able to use function composition to linearize the data.

It is the linearization of the data, not the apparent quality of the fit, that is our evidence that the exponential model is correct.

Power Functions

If the data describes a power relationship, $y = ax^n$, then we can re-express the data using the n th root. The ordered pair $(x, \sqrt[n]{y})$ should be linear. However, there are many different kinds of power functions, from simple quadratic $y = ax^2$, to reciprocal $y = \frac{a}{x}$, square root $y = a\sqrt{x}$, inverse square laws $y = \frac{a}{x^2}$. All of these functions can be re-expressed using the log-log transformation. If $y = ax^n$, then $\ln y = \ln(ax^n)$. Simplifying we find that $y = \ln(a) + n \ln(x)$ is a linear function. If we graph $(\ln x, \ln y)$, we will have a slope of n and an intercept of $\ln a$.

On the side of a large bag of Alpo dog food, the following table is shown relating the weight of the dog to be fed and the number of cups of dog food required. What is the relationship being described here?

Weight	7	20	36	57	79	102	130	159	190
Number of Cups	1	2	3	4	5	6	7	8	9

Table 4: Weight of Dog and Cups of Food

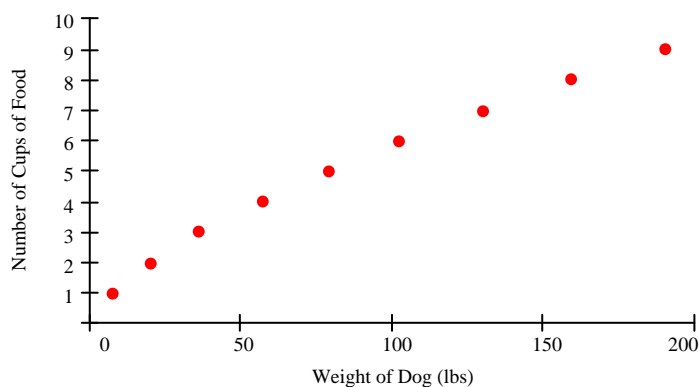


Figure 10: Scatterplot of Weight of Dog and Number of Cups of Alpo

The data appear to be some sort of root function, being increasing. We will try a log-log re-expression and see what happens. The graph of the ordered pair $(\ln W, \ln C)$ is shown below, with its residual plot.

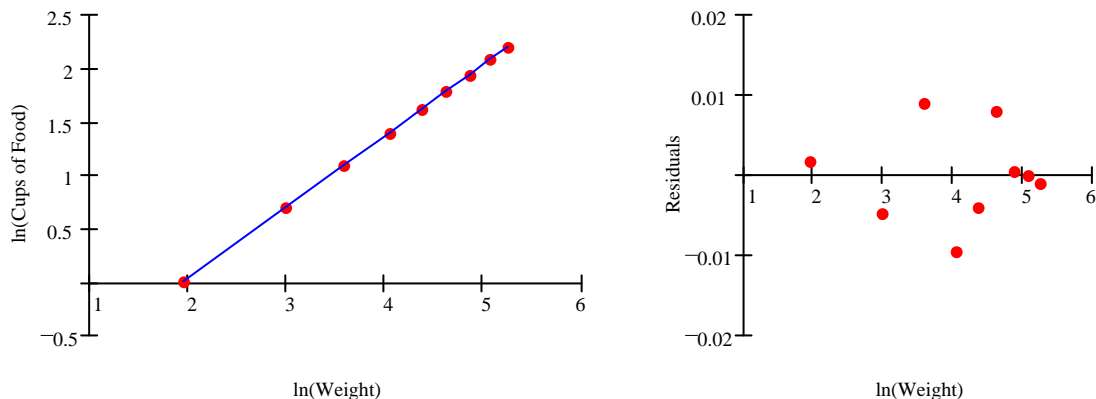


Figure 11: Log-log plot and Residuals

The linear model for this re-expression is $\ln C = -1.298 + 0.666 \ln W$, so $C = 0.273x^{0.666}$. According to the folks at Alpo, the number of cans of food a dog should have each day is apparently proportional to the 2/3 power of its weight.

Managing Vertical Shifts

The CO₂ concentration in the atmosphere has been measured at the Observatory on Mauna Loa since 1959. It has been increasing from 316.1 ppm in 1959 to 369.4 in 2000. During the period from 1959 until 1985, the growth in CO₂ concentration appears to be exponential. What function models the average yearly concentration of CO₂ during this time period?

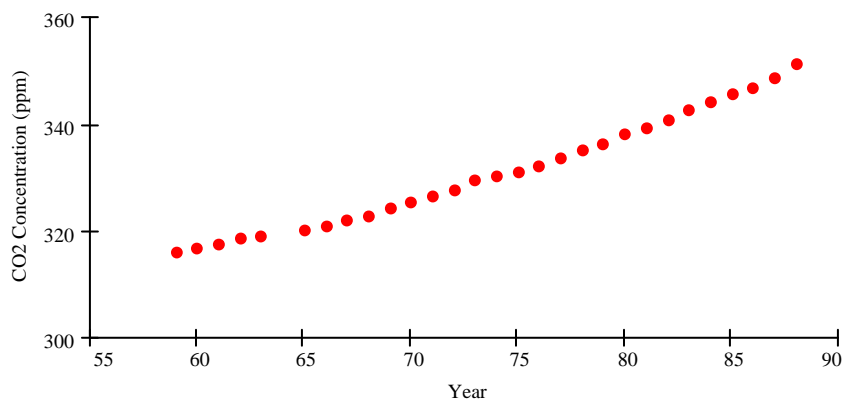


Figure 12: CO₂ Concentration at Mauna Loa Observatory

Year	59	60	61	62	63	65	66	67	68	69
CO ₂	316.1	317.0	317.7	318.6	319.1	320.4	321.1	322.0	322.8	324.2

Year	70	71	72	73	74	75	76	77	78	79
CO ₂	325.5	326.5	327.6	329.8	330.4	331.0	332.1	333.6	335.2	336.5

Year	80	81	82	83	84	85	86	87	88
CO ₂	338.4	339.5	340.8	342.8	344.3	345.7	346.9	348.6	351.2

Table 5: CO₂ Data from Mauna Loa Observatory

If this is an example of exponential growth, there must be a vertical shift involved in the model. The function $C(t) = ae^{kt}$ has a horizontal asymptote at $C = 0$, which is unrealistic. We are seeking a model of the form $C(t) = ae^{kt} + b$, where b represents the asymptotic level of CO₂. We can approximate this value by considering the scatterplot from a longer view. The scatterplot below compares the data (with the origin showing) to the possible asymptote at 250.

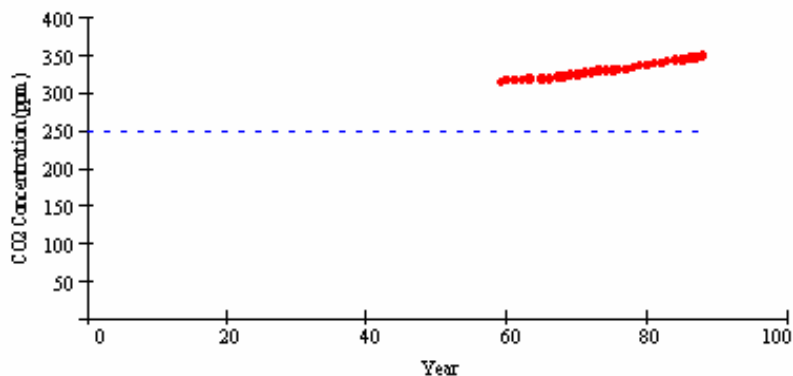


Figure 13: CO₂ Concentration at Mauna Loa Observatory and Asymptote

Look at the two re-expressions $(\text{Year}, \ln(\text{CO}_2 - 250))$, and $(\text{Year}, \ln(\text{CO}_2 - 310))$. The residual plots are also shown below.

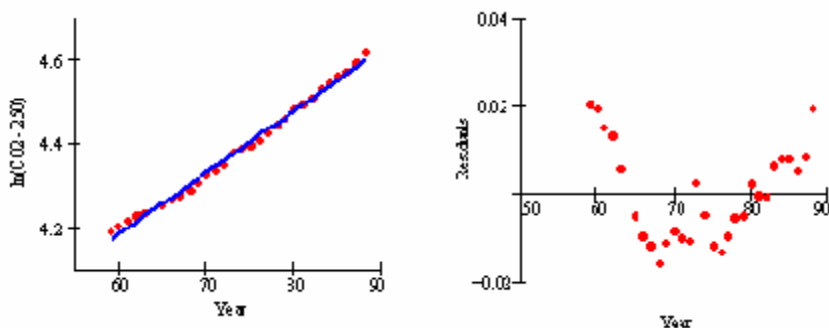


Figure 14: Scatterplot of $(\text{Year}, \ln(\text{CO}_2 - 250))$ and Residual from Linear Fit

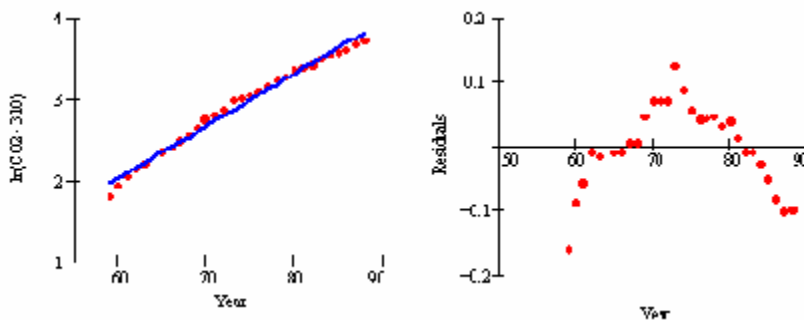


Figure 15: Scatterplot of $(\text{Year}, \ln(\text{CO}_2 - 310))$ and Residual from Linear Fit

Notice the pattern in the residuals for $(\text{Year}, \ln(\text{CO}_2 - 250))$ and compare them to the pattern for $(\text{Year}, \ln(\text{CO}_2 - 310))$. The concavity of the residuals has changed. This is extremely important. When assuming an asymptote of 250, we have a residual pattern that is concave up. When assuming an asymptote of 310, we have a residual pattern that is concave down. Since straight is between concave up and concave down, we expect there will be a value between 250 and 310 that will linearize the data. After fiddling around, using the concavity of the residuals as a guide, we consider a proposed asymptote of 290. The ordered pair $(\text{Year}, \ln(\text{CO}_2 - 290))$ appears linear with a nicely scattered residual plot.

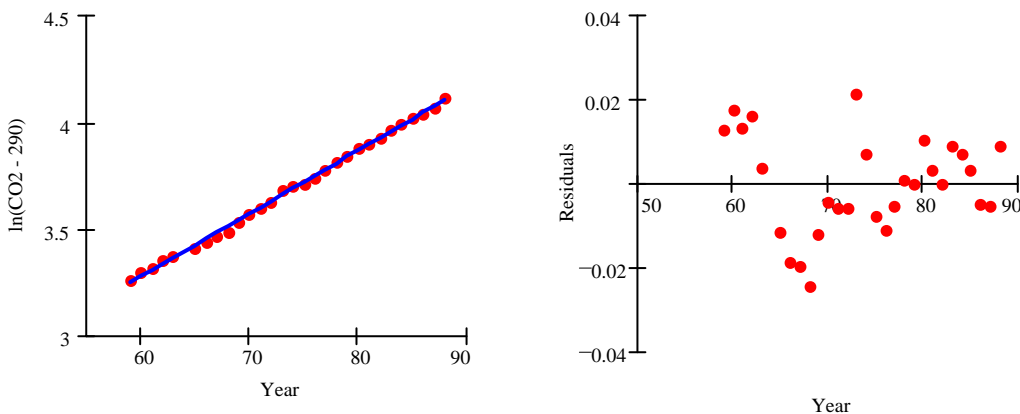


Figure 16: Linear model $\ln(\text{CO}_2 - 290) = 1.506 + 0.02953 \cdot \text{Year}$ and Residual Plot

The regression line is $\ln(\text{CO}_2 - 290) = 1.506 + 0.02953 \cdot \text{Year}$, so our model is $\text{CO}_2 = 4.51e^{0.02953 \cdot \text{Year}} + 290$.

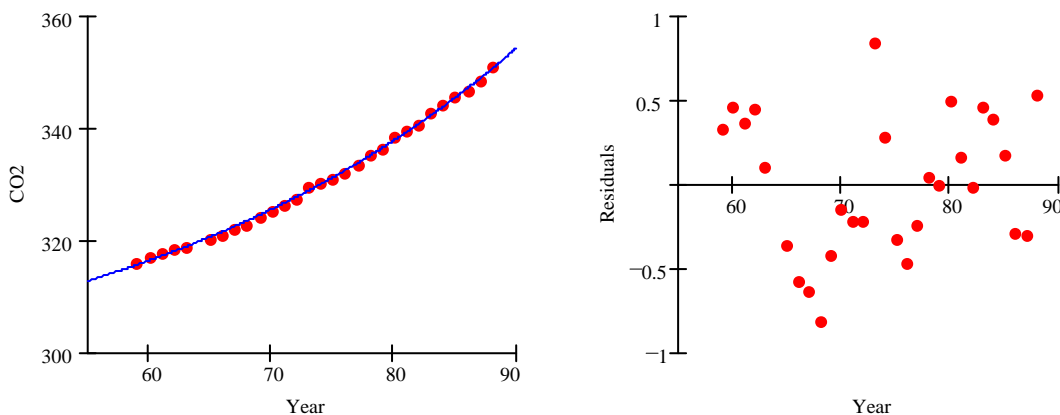


Figure 17: The model $\text{CO}_2 = 4.51e^{0.02953 \cdot \text{Year}} + 290$ and Residual Plot

Concluding Comments

For the AP Statistics program, the log-log power re-expression and the semi-log exponential re-expression are critical tools, as is the analysis of the residuals to assess the quality of the linear fit. These two principles of re-expression are the most commonly used tools and should be a part of every student's toolkit. These principles of re-expression can be extended to

much more sophisticated modeling, well beyond the scope of the AP curriculum. We can fit logistic models. We can tailor our model to our understanding of the phenomenon, fitting, for example, quadratic fits of $y = ax + bx^2$ and $y = a + bx^2$, and perform other wondrous investigations. Moreover, all of our techniques for finding prediction intervals or other inference for regression techniques assumes a linear model. We can assess the quality of the linear model in much more sophisticated ways than we can non-linear models at the introductory level.

Advanced Topics Beyond the AP Course

Logistic Growth

Logistic (or constrained) growth models play an important role in modeling the biological growth patterns. Below is the population of the US based on the US Census from 1790 to 1940. The year 1790 is coded as 0 and 1800 as 1, so each unit of time is 10 years.

$D_i :=$ $p_i :=$

0	3.9
1	5.3
2	7.2
3	9.6
4	12.9
5	17.1
6	23.2
7	31.4
8	38.6
9	50.2
10	62.9
11	76
12	92
13	105.7
14	122.8
15	131.4

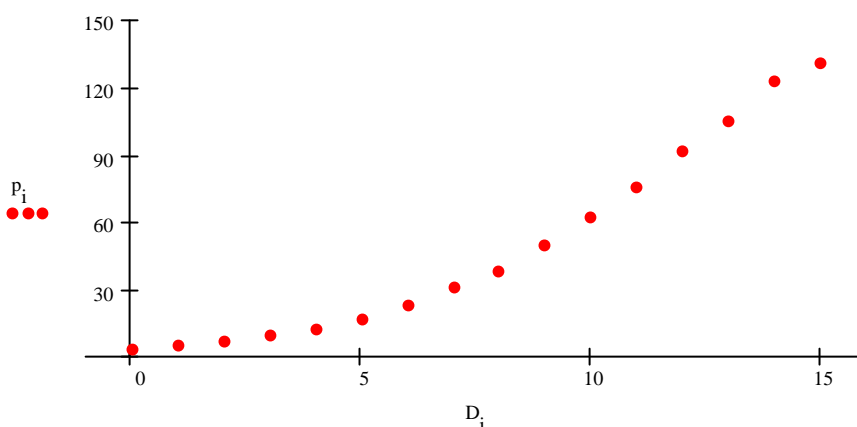


Figure 18: US Census 1790-1940

We can use our knowledge of re-expression and linearization to find a model for this growth pattern. The logistic growth model is $P(t) = \frac{a}{1 + be^{ct}}$. This is a function that models population

growth that has an upper bound. Since the model is $P = \frac{a}{1 + be^{ct}}$, we can linearize it with the

following re-expressions. First, we note that $\frac{1}{P} = \frac{1}{a} + \frac{b}{a}e^{ct}$. The reciprocal of the dependent

variable is an exponential function. Subtracting $\frac{1}{a}$ gives is a simple exponential model that a

semi-log graph should linearize $\frac{1}{P} - \frac{1}{a} = \frac{b}{a}e^{ct}$. So the graph of $\left(t, \ln\left(\frac{1}{P} - \frac{1}{a}\right) \right)$ should be linear

since $\ln\left(\frac{1}{P} - \frac{1}{a}\right) = \ln\left(\frac{b}{a}\right) + ct$. The slope is c and the intercept is $\ln\left(\frac{b}{a}\right)$. Of course, we don't know a , so we have to fiddle around a little to get it. I started out with $a = 160$ (this was just a guess) and looked at the residuals from the regression. They were concave down.

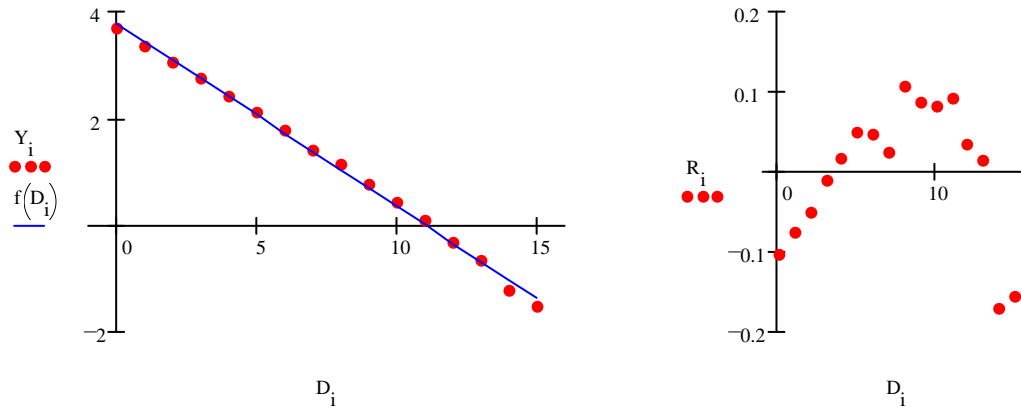


Figure 19: Re-expression with $a = 160$

I then tried $a = 220$ (another guess). The residuals were concave up, and much better.

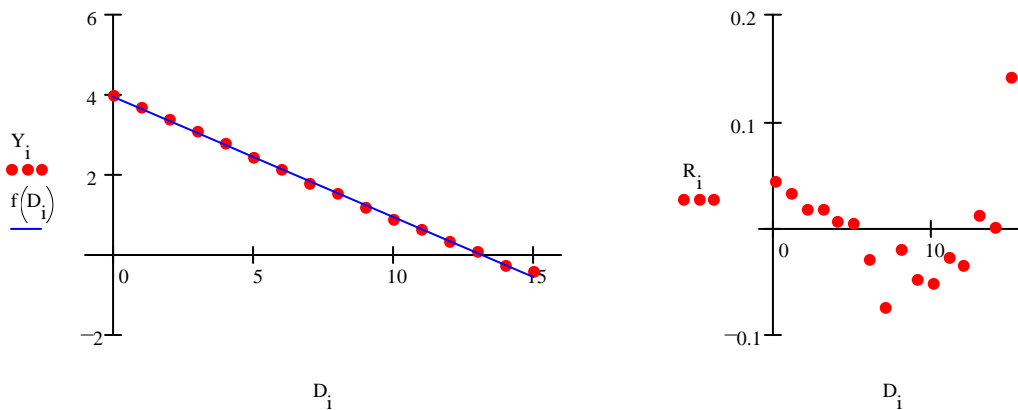


Figure 19: Re-expression with $a = 220$

Since straight is between concave up and concave down, I believe there is a value of a between 160 and 220 that will linearize the data. I fiddled a little and settled on $a = 190$.

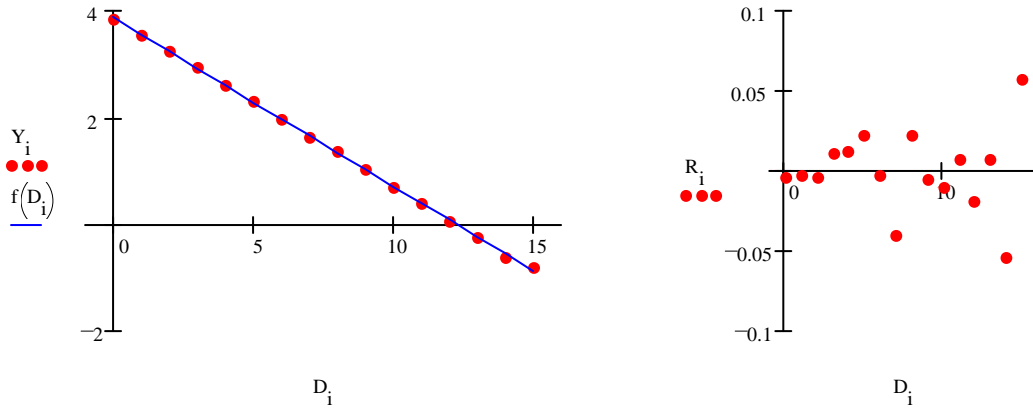


Figure 20: Re-expression with $a = 190$

The key to being successful in this investigation is to find one value of a that gives residuals that are concave down and another that gives residuals that are concave up. Bisection will lead you quickly to an appropriate value. This is difficult on the calculator. Software like MathCad makes this kind of investigation much easier. Anyway, the final model had was $\ln\left(\frac{1}{P} - \frac{1}{190}\right) = -1.377 - 0.316t$, so we have $a = 190$, $\ln\left(\frac{b}{190}\right) = -1.377$, so $b = 47.94$, and $c = -0.316$. The final function is $P(t) = \frac{190}{1 + 47.94e^{-0.316t}}$. The graph of the function against the data looks fine.

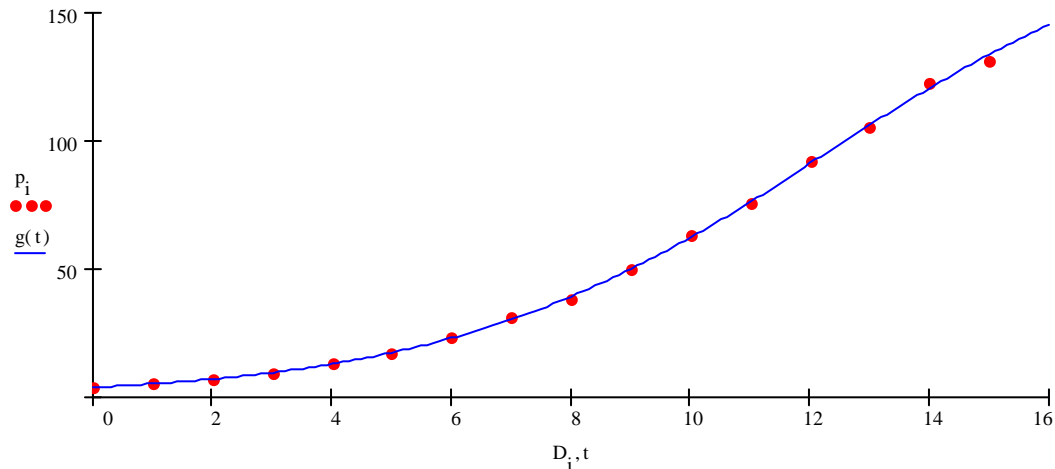


Figure 21: Logistic Fit

Wind Chill Table Investigation

The standard wind-chill table from an almanac provides a challenging investigation in re-expression and curve fitting. The table is presented below:

Wind (mph)	Temperature in degrees Fahrenheit (Winds greater than 45 mph have little additional cooling effect)															
	35	30	25	20	15	10	5	0	-5	-10	-15	-20	-25	-30	-35	

5	33	27	21	16	12	7	0	-5	-10	-15	-21	-26	-31	-36	-42
10	22	16	10	3	-3	-9	-15	-22	-27	-34	-40	-46	-52	-58	-64
15	16	9	2	-5	-11	-18	-25	-31	-38	-45	-51	-58	-65	-72	-78
20	12	4	-3	-10	-17	-24	-31	-39	-46	-53	-60	-67	-74	-81	-88
25	8	1	-7	-15	-22	-29	-36	-44	-51	-59	-66	-74	-81	-88	-96
30	6	-2	-10	-18	-25	-33	-41	-49	-56	-64	-71	-79	-86	-93	-101
35	4	-4	-12	-20	-27	-35	-43	-52	-58	-67	-74	-82	-89	-97	-105
40	3	-5	-13	-21	-29	-37	-45	-53	-60	-69	-76	-84	-92	-100	-107
45	2	-6	-14	-22	-30	-38	-46	-54	-62	-70	-78	-85	-93	-102	-109

Consider each row in the table. Obviously, the values are rounded to the nearest degree. The wind-chill temperatures for a wind of 5 mph gives the following set of data.

Actual Temp	35	30	25	20	15	10	5	0	-5	-10	-15	-20	-25	-30	-35
Wind-Chill Temp	33	27	21	16	12	7	0	-5	-10	-15	-21	-26	-31	-36	-42

Students should recognize that as the Actual Temperature is decreased by 5 degrees, the Wind-Chill Temperature decreases by either 5 or 6 degrees. Thus, a linear model will represent this data well.

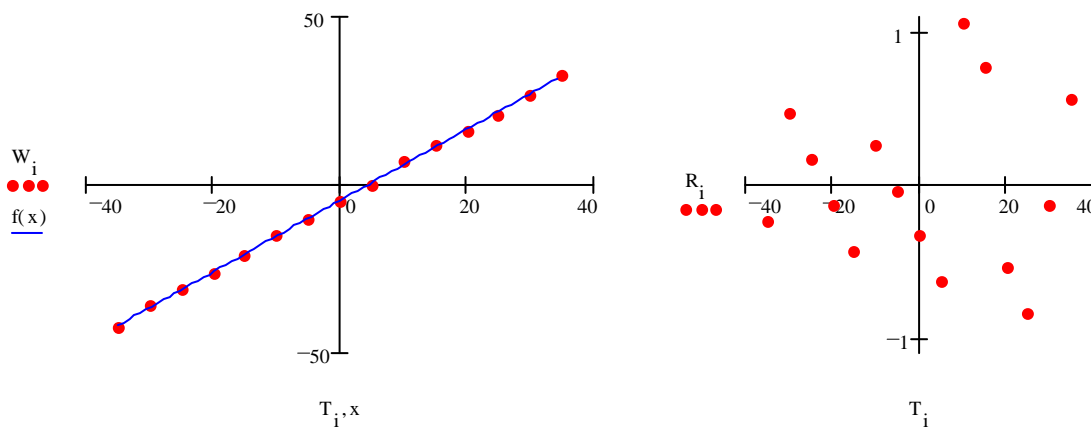


Figure 22: Linear Model for 5 mph Wind

Students may note a sawtooth aspect to the residual plot. This is characteristic of data that has been rounded. Not only is this set of data linear, but the corresponding data for each wind velocity is linear. This is a good practice set for entering data and fitting linear models.

For each wind velocity, we can find the linear model. Here is the list:

Wind Velocity	Linear Least-Squares Model
5 mph	$C_5 = 1.058T - 4.650$
10 mph	$C_{10} = 1.24T - 21.30$
15 mph	$C_{15} = 1.343T - 31.34$
20 mph	$C_{20} = 1.425T - 38.46$
25 mph	$C_{25} = 1.479T - 43.90$
30 mph	$C_{30} = 1.521T - 48.10$
35 mph	$C_{35} = 1.551T - 50.71$

40 mph	$C_{40} = 1.573T - 52.49$
45 mph	$C_{45} = 1.586T - 53.76$

As the wind velocity changes, the slopes and intercepts of the regression lines change. This means the slopes and intercepts are a function of the wind velocity. It appears that the wind-chill temperature can be written as $C = f(w)T + g(w)$ for some functions f and g . The data defining functions f and g are given below:

Wind Velocity	5	10	15	20	25	30	35	40	45
Slope	1.058	1.235	1.43	1.25	1.479	1.521	1.551	1.573	1.586
Intercept	-4.650	-21.30	-31.34	-38.46	-43.90	-48.10	-50.71	-52.49	-53.76

The scatterplots for the slopes and intercepts are given below:

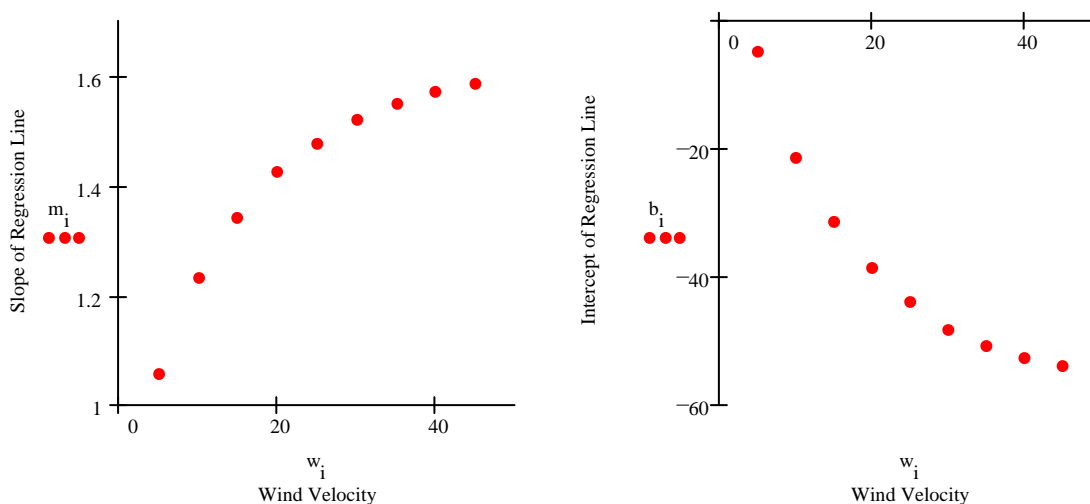


Figure 23: Slope and Intercept Functions

What functions model this behavior?

A statement on the table that wind velocities greater than 45 mph have little additional cooling effect suggests asymptotic behavior. We can try exponential functions since they also illustrate this asymptotic behavior. If the relationship between Wind Velocity and Slope is exponential, what kind of exponential function do we have? We can argue that an exponential function of the form $m = -Ae^{-kw} + B$ might fit. From our graph, we can consider an initial guess of $B = 1.7$. If this is correct, then the re-expression $(w, \ln(1.7 - m))$ should be linear. The residual plot indicates this is not linear. The re-expression $(w, \ln(1.6 - m))$ is then tried and we see that this residual plot has the opposite curvature. There is a value between 1.6 and 1.7 that should work.

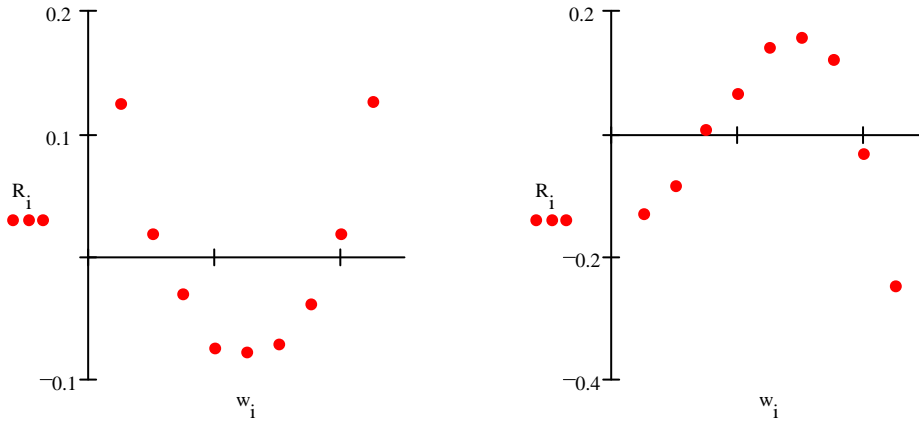


Figure 24: Residual Plots Indicating Change in Concavity

By trial and error, we find 1.623 is a good value.

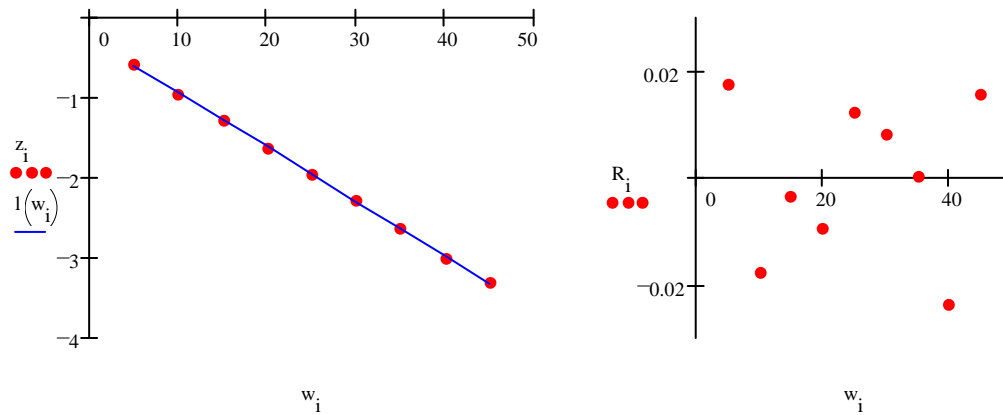


Figure 25: Linearized Data

The model, then is, $f(w) = 1.623 - 0.78e^{-0.068w}$.

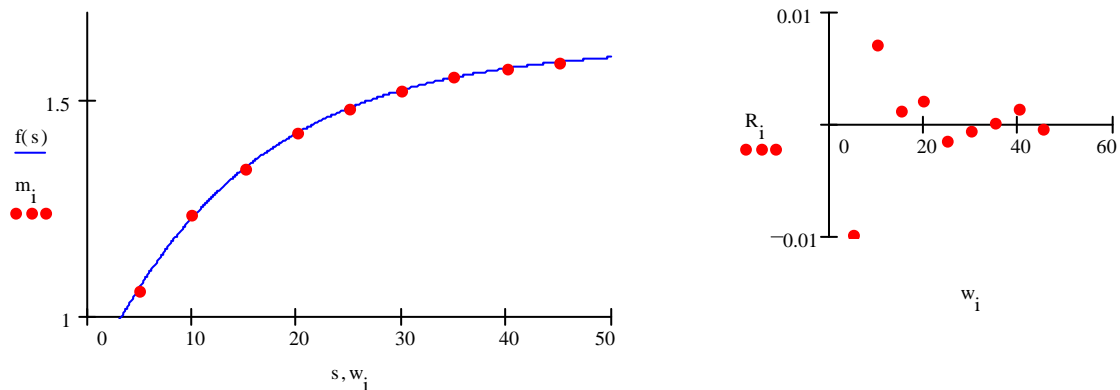


Figure 26: Exponential Model with Residuals

With similar techniques, we can show that the relationship between the wind velocity and the intercept of the regression line is $g(w) = 70.316e^{-0.66w} - 57.5$. So the wind-chill function can be defined as $C(w, T) = (1.623 - 0.78e^{-0.068w}) \cdot T + (70.316e^{-0.66w} - 57.5)$. The numbers in the wind-chill table are the values of this function rounded to the nearest integer.