1.7 Linear and Exponential Data Fitting

1.7.1 The Nature of the Data

When information or data is collected, it frequently contains some amount of error. Therefore, data with an inherently linear relationship may not appear to fit this relationship perfectly. This can also be said of data which truly have attributes lending them to the use of an exponential model. In other words, although we have found the correct underlying relationship, sometimes our models may not appear to fit the data perfectly. Other times, we may select a model which appears to fit the data well but it isn't the best choice for prediction. This lesson will challenge you to think about the data so we can begin choosing the best type of model, based upon the nature of the data.

We've discussed the *nature of the data* previously in class. Our intuition regarding the *nature of the data* is that it would be what the data would look like if there was no randomness associated with the collection, as if we lived in a "perfect world." However, we know that there is some degree of randomness in almost everything that happens in the world. For example, a person may think that firing a bullet from a rifle exactly the same way will result in hitting the same spot on the target. But, how many times do two bullets go through the same hole? Not often. Even if we take the human completely out of the loop and fire the rifle from mounts on the ground, the bullets won't go through the same hole. Why? The answer is in the randomness that exists in the world, possibly: wind resistance, tiny abnormalities in the bullet affecting trajectory, wear on the rifle barrel, percentage of powder igniting, and many more.

The result is that the data we see most likely cannot be *perfectly* modeled with any reasonable model. But, we know that we can model the trajectory of the bullet, determining where it will strike, with a parabolic function (thanks to Sir Isaac Newton), because we understand that the nature of the data is a parabolic trend.

> The **nature of the data** consists of the underlying attributes of the data which describe the pattern it will take both within and outside of the collected region.

In addition to examining our knowledge of the nature of the data, we can apply quantitative tests to determine what we may expect the nature of the data to be.

- Is the rate of change of the data constant (or near constant)?
  - If yes, try a linear model.
- If not, then try an exponential model, unless the data follow a cycle, which we will address in later lessons.

1.7.2 Data Fitting

In section 1.7.1, we discussed using the nature of the data to determine what type of model we will use to model data we are given. It is now time to fully develop the model that we selected. Finally, we will plot our models against the data and apply the "eyeball," SSE, and r^2 tests to determine if a particular model is good or not.

Example 1 *In doing some research on Mexico, we come across the data in Table 1.11 regarding Mexico's population. We are interested in modeling the population as a function of time but need to determine which function will best fit the data.*[1]

Year	Population (millions)
2000	100
2001	102
2002	104.04
2003	106.12
2004	108.24
2005	110.41

Table 1.11: Population of Mexico

Step 1: *Transform the problem.* We are **given** a table of data, we must examine the table.

- Define the variables:
  - Independent variable (input): year
  - Dependent variable (output): population

- Nature of the data: Apply the quantitative measures of the nature of the data to gain an idea of which type of function may be most appropriate. See Table 1.12.

Mexico Population		
Year	**Population (Millions)**	**Rate of Change**
2000	100	
2001	102	2
2002	104.04	2.04
2003	106.12	2.08
2004	108.24	2.12
2005	110.41	2.17

Table 1.12: Quantitative Measure of the Nature of the Data

[1] Problem adapted from *Functions Modeling Change: a Preparation for Calculus*, 3rd Edition. Connally, Hughes-Hallett, Gleason, et al., 2007. p. 23.

Notice that the data are increasing, as is the rate; therefore, we need a function that is increasing at an increasing rate, modeled by an exponential function. Let's graph the data to verify our conjecture. See Figure 1.34.

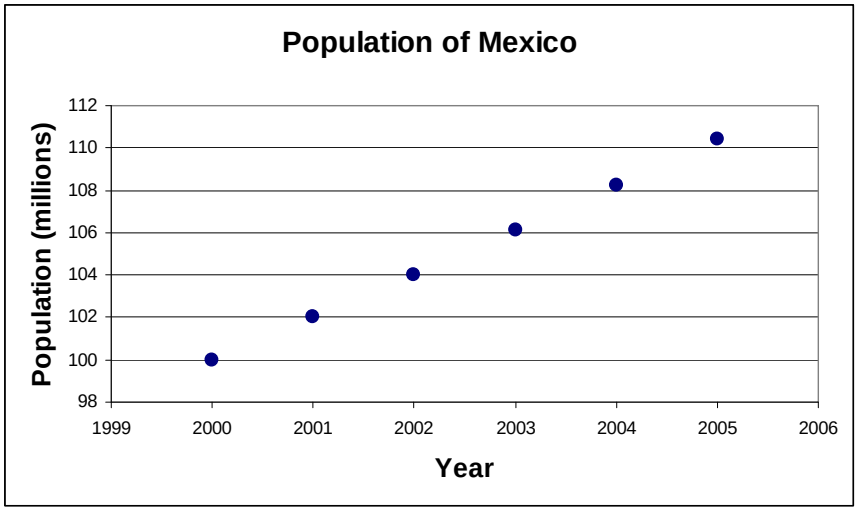

Figure 1.34: Scatterplot of Table 1.11 Data.

The data in Figure 1.34 seems to exhibit a *linear* pattern. But our quantitative measure of the nature of the data and our visual measure seem to conflict. We know we will have to **find** the model that fits through the data best.

- Assumptions
  - Based on the scatterplot in Figure 1.34, we will assume the plot of the data appears to increase at a constant rate. We'll assume a linear model might be a good choice.
  - Since the data do not increase in a perfectly linear manner (or, the rate of change would be constant), we'll also assume a generalized exponential model might model the data. We select this because we can develop a model that incorporates a bend in the data and a vertical shift.
  - The data accurately represent what will happen to the population in the coming years.

Our **plan** is to use the model development and evaluation techniques that we have used to this point in the course to find the best model to determine Mexican population in the future.

Step 2. *Solve the problem using appropriate solution techniques.* To **solve**, we will develop a linear and an exponential model, then evaluate the two to decide which is best.

- <u>Linear model development</u>. First, we'll try the linear model. In general, our linear model will be:

$$Population(year) = a(year) + d$$

- Linear Model Parameter Estimation: Using modeling skills developed to date, we need to estimate the parameters (a and d) for a linear model. To estimate two parameters, we need two data points that represent the data well.

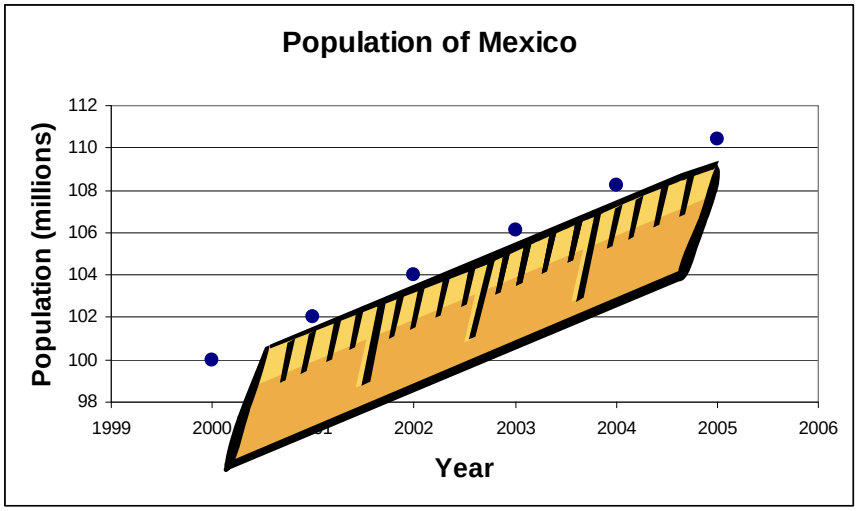

Figure 1.35: Determining Representative Data Points for Parameter Estimation

- The apparent linear trend of the data, seen in Figure 1.35, indicates that any of the data will represent it well; therefore, we'll select the first and last data points to estimate our parameters.

- Now, as we have many times in the past, we'll create two linear equations using the data points to estimate our parameter values. Our two equations are:

$$100 = a(2000) + d$$
$$110.41 = a(2005) + d$$

```
Solve[{100 == a * 2000 + d, 110.41 == a * 2005 + d}, {a, d}]

{{a → 2.082, d → -4064.}}
```

Figure 1.36: Using *Mathematica* to Solve Two Simultaneous Linear Equations

Using the solution from Figure 1.36 yields the final model:

$$Population(year) = 2.082(year) - 4064$$
$$Domain: \{year \mid 2000 \le year \le 2005\}$$
$$Range: \{population \mid 100 \le population \le 110.41$$

The model is plotted in Figure 1.37; by visual inspection, it fits the data well.

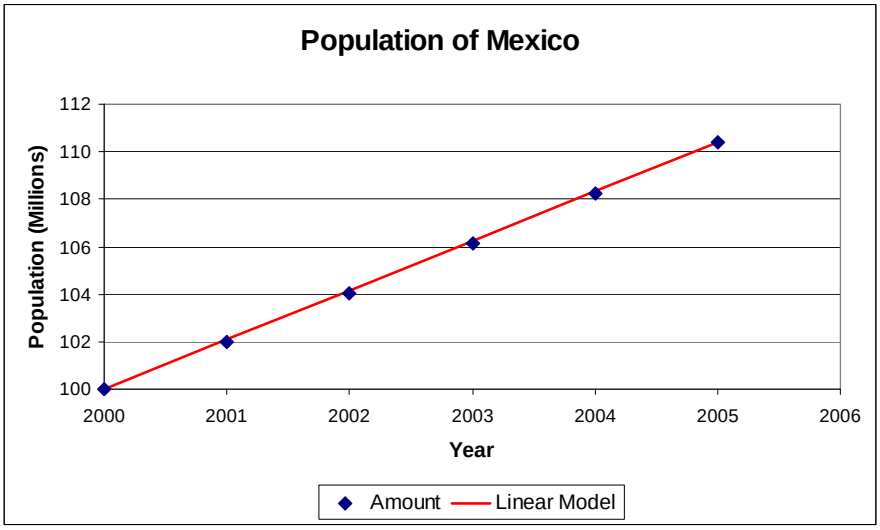

Figure 1.37: Mexican Population Data Overlaid with Linear Model

- <u>Exponential model development</u>. Now that we've developed a linear model, it is time to develop our generalized exponential model, of the form:

$$Population(year) = a(b)^{year} + d$$

- As we saw earlier in the example, each of the points seems representative. It make sense to solve for all three parameters if we can, so we'll use the second, fourth, and sixth points to form three equations in three unknowns and let *Mathematica* do the work for us. See Figure 1.38 for the results.

- **NOTE** The independent variable in the equation (year) is scaled by subtracting 2000 from each actual year value. It is common practice to scale the independent variable when working with exponential functions because raising a number to the 2000th power is very often problematic.

```
In[1]:= Solve[{102 == a * b^1 + d, 106.12 == a * b^3 + d, 110.41 == a * b^5 + d}, {a, b, d}]

Out[1]= {{d → 2.15059, a → -97.851, b → -1.02042}, {d → 2.15059, a → 97.851, b → 1.02042}}
```

Figure 1.38: Using *Mathematica* to Solve for Three Parameters

- Since the b parameter needs to be greater than 1 (to ensure an increasing exponential function), we choose the second solution returned by *Mathematica*: $a = 97.851$, $b = 1.02042$, and $d = 2.15059$.

Using the second solution from Figure 1.38, yields the final model:

$$Population(year) = 97.851(1.02042)^{year} + 2.15059$$
$$Domain: \{year \mid 0 \le year \le 5\}$$
$$Range: \{population \mid 100 \le population \le 110.41$$

The exponential model is plotted in Figure 1.39, to see how well it appears to fit the data.

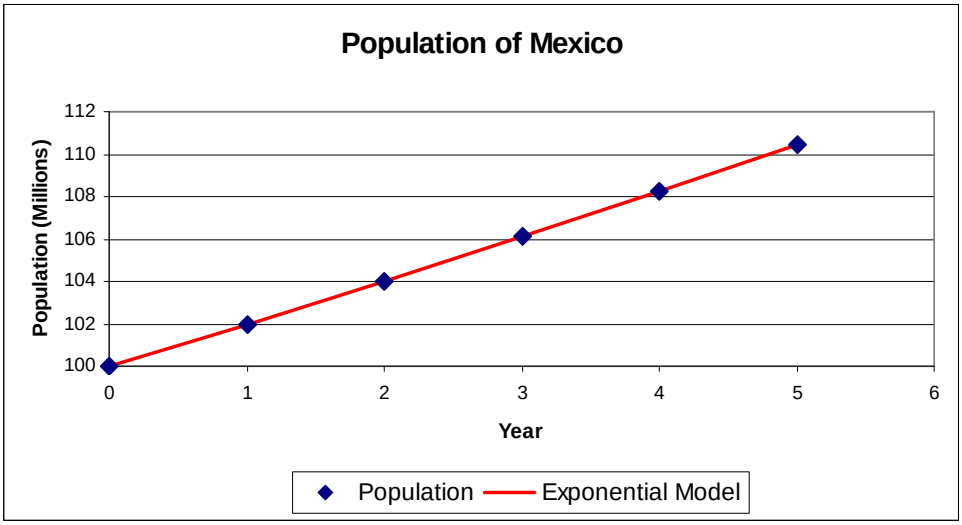

Figure 1.39: Mexican Population Data Overlaid with Exponential Model

Step 3. *Interpret the results of the solution.* The estimated linear model certainly seems to fit the data as well as the estimated exponential model. Overall, the linear model seems to be our best model – it fits almost as well as the exponential and has one fewer parameter (it's simpler). Though it may not seem necessary to compute sum of squared error and r^2 in this case, we do it anyway for two reasons.

- Quantitative, objective justifications are useful to back-up qualitative, subjective observations.
- Valuable practice in the computation of SSE is useful for cases that are not so clear cut.

Recall from Section 1.5 that we compute SSE by using the following formula:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^{\,2}\,.$$

Derive the error at each point by subtracting the predicted value from your model from the actual value of the data. Square each of these values, summing these squared provides SSE. See the results in Figure 1.40.

Year	Population	Linear Model	Squared Mean Error	SST	Model Error	Squared Model Error	SSE			
2000	100	100	26.368225	75.83	0.00000000	0.00000000	0.04572000		a	2.082
2001	102	102.082	9.828225		0.08200000	0.00672400			b	
2002	104.04	104.164	1.199025		0.12400000	0.01537600	**R-Sq**		c	
2003	106.12	106.246	0.970225		0.12600000	0.01587600	0.99939709		d	-4064
2004	108.24	108.328	9.641025		0.08800000	0.00774400				
2005	110.41	110.41	27.825625		0.00000000	0.00000000				
Mean:	105.14									
Year	**Population**	**Exponential Model**	**Squared Mean Error**	**SST**	**Model Error**	**Squared Model Error**	**SSE**			
0	100	100.00159	26.368225	75.83	0.00159000	0.00000253	0.00001203		a	97.851
1	102	101.9997074	9.828225		-0.00029258	0.00000009			b	1.02042
2	104.04	104.0386264	1.199025		-0.00137360	0.00000189	**R-Sq**		c	
3	106.12	106.1191801	0.970225		-0.00081990	0.00000067	0.99999984		d	2.15059
4	108.24	108.2422187	9.641025		0.00221871	0.00000492				
5	110.41	110.4086098	27.825625		-0.00139023	0.00000193				
Mean:	105.14									

Figure 1.40: Computation of Linear and Power Sum of Squared Error

As anticipated through visual inspection, the linear and exponential models have similar SSE and r^2. In fact, the linear model accounts for nearly 99.94% of the variation in the data whereas the exponential model accounts for greater than 99.99%.

The nature of the data would appear to support an exponential model, but when we use the other tests at our disposal; we realize that we can use a simpler, two-parameter linear model instead of the more complicated three-parameter exponential model. So, it seems both of our models are very good. So, which do we use? Good question! Let's see if we can determine what to do as we **reflect**.

We have determined the following: The linear model would probably prove useful in extrapolating for a short time period either prior to 2000 or after 2005 but wouldn't be useful for the long-term, because it does not fit the nature of the data. The problem is identifying an appropriate domain for our model. How far out can we consider the linear model useful?

This question is an important question that highlights the fact that modeling has components that are "science," as in the development and evaluation of the models, and "art," as in the interpretation of the models and their relevance over certain domains. In fact, our final model may be a piecewise function that enables us to choose different functions over different ranges of the domain.

Question 1 *In Example 1, over what domain would a linear function be most appropriate? An exponential function?*

Question 2 *The following table contains data on the population of two countries in millions. One of the countries experiences relatively constant growth between 1950 and 2000 while the other does not.*[2]

a. *Develop two models (one linear and one generalized exponential model) to predict each country's population over time. Outline your problem solving process ensuring you include your assumptions and parameter estimation process. Discuss the fit of your models.*

b. *Which of the two countries exhibits non-constant growth?*

c. *Using the model you believe to be best for each country, estimate the populations in 1993.*

Year	1950	1960	1970	1980	1990	2000
Country A	8.2	9.8	12.4	15.1	14.7	23.9
Country B	7.5	9.9	12.5	14.9	17.2	19.2

Population Data for Countries A and B

Question 3 *Go back to one of the models that you have worked with in class, the one you found most interesting. Estimate the parameters of the model to achieve a better fit. Communicate your results in terms of the steps you took to develop the model, evaluate the model, and conduct a sensitivity analysis.*

[2] Problem adapted from *Functions Modeling Change: a Preparation for Calculus*, 3rd Edition. Connally, Hughes-Hallett, Gleason, et al., 2007, p. 25.

Challenge Question.

In Section 1.4, we considered data pertaining to the launch of the Space Shuttle Challenger. This data, on the percent expansion of an o-ring, appeared to indicate a linear relationship between the surrounding temperature and the percent expansion. Unfortunately, our attempts to model this data failed to identify that the space shuttle should not have launched.

Now, let's consider the *nature of this data*. O-rings are made of rubber (picture an o-ring as a rubber band). When heated, rubber expands allowing it to stretch beyond the expansion it would be capable of at room temperature but with a limitation - eventually, the rubber will break. When cooled, rubber becomes brittle and is capable of very little expansion before breaking. Therefore, if we were to describe an o-ring's expansion based on the temperature of its surroundings, we would realize that at low temperatures, the o-ring would not be capable of much expansion. As the temperature increased, the o-ring would expand quickly but, over the long term, the o-ring would reach its maximum expansion and would eventually break. Graphically, o-ring expansion would resemble something like the graph shown in Figure 1.41.

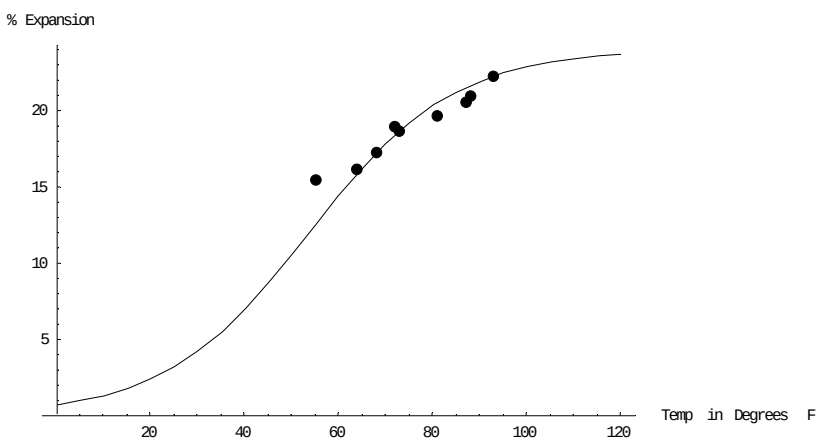

Figure 1.41: Nature of the Data: O-ring Expansion

Develop, fully evaluate, and communicate a model that would effectively represent the nature of the data illustrated in Figure 1.41 and predict the explosion of the Challenger.