FITTING LINEAR MODELS TO DATA

Supplement to Unit 9B

MATH 1001

In the handout we will learn how to find a linear model for data that is given and use it to make predictions. We will also learn how to measure how closely the model "fits" the given data. We will learn how to find the linear model that best fits a set of given data.

Finding a Linear Model for Data and Making Predictions

First, we consider the following table that gives the population for Spalding County, Georgia from 1960 to 2000. (Source: U.S. Census Bureau)

Year	Pop. (thousand)	Change
1960	35.4	
1970	39.5	4.1
1980	47.9	8.4
1990	54.5	6.6
2000	58.4	3.9

The third column of this table shows (for each decade year) the change in population during the preceding decade. We see the population of Spalding County increased by about 4 thousand people in the 1950s and 1990s. In the 1970s and 1980s the population increased by roughly 7 to 8 thousand people. We might wonder whether this qualifies as almost linear population growth. So, we will plot the data and look. We plot the year on the *x*-axis and the population on the *y*-axis.

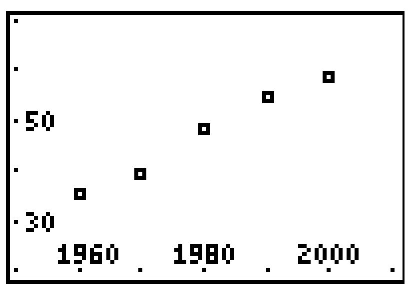

Figure 1

It appears from Figure 1 that these points do appear to lie on or near some straight line. But, how can we find a straight line that passes through or near each data point? One way is to simply pass a straight line through the first and the last data points. To make this easier, we will let *t* be the number of years after 1960. Thus, our data will look like:

t (years since 1960)	P (in thousands)
0	35.4
10	39.5
20	47.9
30	54.5
40	58.4

To find the slope between the first data point (0, 35.4) and the last data point (40, 58.4), we use the formula for the slope:

$$\begin{aligned} \text{slope} = m &= \frac{\text{change in } P}{\text{change in } t} \\ &= \frac{58.4 - 35.4}{40 - 0} \\ &= \frac{23}{40} \\ &= 0.575 \end{aligned}$$

Note that the P-intercept is (0, 35.4). Thus, a linear model for the population of Spalding County is

$$P(t) = 0.575t + 35.4$$

The graph of this line and the data points is shown in Figure 2 below. You can see that this line passes through or near each data point.

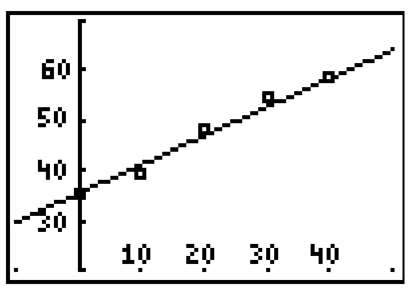

Figure 2

One of the reasons to find a model for real-world is to use the model to make predictions. These predictions fall into two categories: (1) making predictions within the scope of the data and (2) make predictions beyond the scope of the data.

As an example of the first type of prediction use the model we found for Spalding County population to predict the population for the year 1995. Notice that 1995 is 35 years after 1960. Thus, we substitute t = 35 into our equation:

$$P(35) = 0.575 \times 35 + 35.4 \approx 55.5$$

Thus, the population in Spalding County in 1995 was approximately 55.5 thousand people.

As an example of the second type of prediction use the model we found for Spalding County population to predict the population for the year 2010. Notice that 2010 is 50 years after 1960. Thus, we substitute $t = 50$ into our equation:

$$P(50) = 0.575 \times 50 + 35.4 \approx 64.2$$

Thus, the population in Spalding County in 2010 will be approximately 64.2 thousand people.

NOTE: The farther predictions get from the ends of the data, the less reliable they become. For example, using our model to prediction the population of Spalding County in 2040 would not give accurate results.

Measuring How Closely the Model Fits the Data

To measure how closely the model above fits the data, we begin by comparing the actual population values with the ones predicted by the model we found. The **error** is the difference in the actual value and the predicted value; that is,

$$\text{error} = (\text{actual value}) - (\text{predicted value})$$

To get the predicted value, we substitute the values for t into our equation $P(t) = 0.575t + 35.4$

$t = 0$ $\quad P(0) = 0.575 \times 0 + 35.4 = 35.4$
$t = 10$ $\quad P(10) = 0.575 \times 10 + 35.4 = 41.15$
$t = 20$ $\quad P(20) = 0.575 \times 20 + 35.4 = 46.9$
$t = 30$ $\quad P(30) = 0.575 \times 30 + 35.4 = 52.65$
$t = 40$ $\quad P(40) = 0.575 \times 40 + 35.4 = 58.4$

Year	t	P (Actual)	$P(t)$ (Predicted)	Error, E_i $P - P(t)$
1960	0	35.4	35.4	0
1970	10	39.5	41.15	−1.65
1980	20	47.9	46.9	1
1990	30	54.5	52.65	1.85
2000	40	58.4	58.4	0

The more useful way to measure how closely the model fits the data is by calculating the sum of the squares of errors and the average error.

Definition: The phrase "**Sum of Squares of Errors**" is so common in data modeling that it is abbreviated **SSE**. Thus, the SSE associated with data modeling based on *n* data points is defined by

$$\text{SSE} = E_1^2 + E_2^2 + E_3^2 + \cdots E_n^2$$

To find the SSE we first begin by finding the squares of the errors.

t	P (Actual)	$P(t)$ (Predicted)	Error, E_i $P - P(t)$	E_i^2
0	35.4	35.4	0	0
10	39.5	41.15	−1.65	2.7225
20	47.9	46.9	1	1
30	54.5	52.65	1.85	3.4225
40	58.4	58.4	0	0

Thus, the SSE is

$$\begin{aligned}\text{SSE} &= 0 + 2.7225 + 1 + 3.4225 + 0 \\ &= 7.145\end{aligned}$$

The smaller the SSE is the better the model fits the data. This allows you to compare two or more different models to determine which one is the best.

Another way to compare models is by finding the average error.

Definition: The **average error** in a linear model fitting *n* given data points is defined by

$$\text{average error} = \sqrt{\frac{\text{SSE}}{n}}$$

The average error for our model is

$$\begin{aligned}\text{average error} &= \sqrt{\frac{7.145}{5}} \\ &= \sqrt{1.429} \\ &\approx 1.195\end{aligned}$$

Fitting Linear Models to Data

Example:

(a) Find a linear model for the population of Spalding County using the first and fourth data points; that is, (0, 35.4) and (30, 54,5).

(b) Use your model to predict the Spalding County population in the years 1995 and 2010.

(c) Find the SSE and average error. Use these to determine whether the model found in this example or the previous model is a better fit for the data.

(a)

(b)

(c)

t	P (Actual)	$P(t)$ (Predicted)	Error, E_i $P - P(t)$	E_i^2
0	35.4			
10	39.5			
20	47.9			
30	54.5			
40	58.4			

SSE =

average error =

Finding the Best-Fit Linear Model for Given Data

The big question is: how do we find the "best-fit" linear model for the data. In short, we find it by finding a value for the slope of the line and the y-intercept that makes both the SSE and average error as small as possible. Our calculators will find the best-fit linear model for us. The steps are outlined below.

1. Select **STAT, 1:Edit…**.
2. Enter the x-values for the data in **L1** and the y-values in **L2**.
3. Press **STAT** and arrow over to **CALC**.
4. Select **4:LinReg(ax+b)**.
5. Then enter L1 and L2 (or which ever lists you have your x- and y-values stored in).
6. Press **L1,L2**.
7. Press **ENTER**.

Fitting Linear Models to Data

Example:

(a) Find the best-fit linear model for the population data for Spalding County.
(b) Use your model to predict the population in 1995 and 2010.
(c) Find the SSE and the average error for the model.

(a) $P(t) = 0.61t + 34.94$
(b) $P(35) = 0.61 \times 35 + 34.94 \approx 56.3$ thousand
$P(50) = 0.61 \times 50 + 34.94 \approx 65.4$ thousand
The population of Spalding County was about 56.3 thousand in 1995 and will be about 65.4 thousand in 2010.
(c)

t	P (Actual)	$P(t)$ (Predicted)	Error, E_i $P - P(t)$	E_i^2
0	35.4	34.94	0.46	0.2116
10	39.5	41.04	−1.54	2.3716
20	47.9	47.14	0.76	0.5776
30	54.5	53.24	1.26	1.5876
40	58.4	59.34	−0.94	0.8836

SSE = 0.2116 + 2.3716 + 0.5776 + 105876 + 0.8836 = 5.632

average error = $\sqrt{\dfrac{5.632}{5}} \approx 1.061$

<u>Exercises</u>:

In each of problems 1 and 2 the population census data for a U.S. city is given.

(a) Find a linear model for the data using the first and last data points. Let $t = 0$ in the year 1950. Use it to predict the population in 2000. Calculate the average error of the model.
(b) Find the linear model that best fits this census data. Let t be 0 in the year 1950. Use it to predict the population in 2000. Calculate the average error of the model.

1. San Diego, California:

Year	1950	1960	1970	1980	1990
Pop. (thous)	334	573	697	876	1111

2. Riverside, California:

Year	1950	1960	1970	1980	1990
Pop. (thous)	47	84	140	171	227

In each of problems 3 and 4 the population census data for a U.S. city is given.

(a) Find a linear model for the data using the second and fourth data points. Let $t = 0$ in the year 1950. Use it to predict the population in 2000. Calculate the average error of the model.

(b) Find the linear model that best fits this census data. Let t be 0 in the year 1950. Use it to predict the population in 2000. Calculate the average error of the model.

3. Garland, Texas:

Year	1950	1960	1970	1980	1990
Pop. (thous)	11	39	81	139	181

4. Santa Anna, California:

Year	1950	1960	1970	1980	1990
Pop. (thous)	46	100	156	204	294

5. The following table gives the number of compact discs (in millions) sold in the United States for the even-numbered years 1988 through 1996.

Year	1988	1990	1992	1994	1996
Sales, S, (millions)	149.7	286.5	407.5	662.1	778.9

Source: *The World Almanac and Book of Facts 1998.*

(a) Find the linear model $S(t) = mt + b$ that best fits this data. Let $t = 0$ in 1988.

(b) Compare the model's prediction for the year 1995 with the actual 1995 CD sales of 722.9 million.

(c) Use the model to predict the CD sales for the year 2002.

(d) Which prediction, the one for 1995 or the one from 2002, is likely to be closer to actual sales? Why?

6. The table below lists the number of passenger cars (in millions) in the United States for the years 1940 through 1990.

Year	1940	1950	1960	1970	1980	1990
Number of Cars, N, (millions)	27.5	40.3	61.7	89.3	121.6	133.7

Source: *Statistical Abstracts of the United States.*

(a) Find the best-fit linear model for the data. Let $t = 0$ in 1940.

(b) Use your model to predict the number of passenger cars in the year 2000 and in the year 2010.

Fitting Linear Models to Data

Thus far we have constructed linear models for data that represent a function of some independent variable. Frequently in the real world, we are confronted with data that does not actually describe a function, but that suggests a correlation that might be modeled by a liner function. (For more information on correlation, review Unit 5E.) Exercises 7 and 8 are examples of such data.

7. In a 1977 study of 21 of the best American female runners, researchers measured the average stride rate, *S*, at different speeds, *v*. The data are given in the table below.

Speed in ft/sec, *v*	15.86	16.88	17.50	18.62	19.97	21.06	22.11
Stride rate in steps/sec, *S*	3.05	3.12	3.17	3.25	3.36	3.46	3.55

Source: R.C. Nelson, C.M. Brooks, and N.L. Pike, “Biomechanical Comparison of Male and Female Distance Runners.” *The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies*, ed. P. Milvy, pp. 793-807, (New Your: New York Academy of Sciences, 1977).

(a) Find the best-fit linear model $S(v) = mv + b$ for the data.
(b) Use your model to predict the stride rate when the speed is 18 ft/sec and when the speed is 10 ft/sec.
(c) Which of predictions in part (b) do you have more confidence in? Why?

8. The table below gives the height, *h*, (in inches) and the weight, *W*, (in pounds) of seven infielders on the Los Angeles Dodgers roster on July 12, 1997.

Height in inches, *h*	70	74	71	71	76	70	73
Weight in pounds, *W*	163	170	180	145	222	185	200

(a) Make a scatter plot and observe that the data appears to have a liner relationship. (See the PowerPoints for Unit 5E for how to draw a scatter plot on your calculator.)
(b) Find the linear model $W(h) = mh + b$ that best fits this data.
(c) Use the linear model to predict the weight of a major league infielder who is 6 feet tall. (Hint: First, convert 6 feet to inches.)
(d) Should you use this model to predict the weight of any American male who is 6 feet tall? Why or why not?

<u>Answers</u>:

1. (a) $P(t) = 19.425t + 334$; $P(50) = 1305$ thousand; average error ≈ 29.37
   (b) $P(t) = 18.57t + 346.8$; $P(50) \approx 1275$ thousand; average error ≈ 26.42

2. (a) $P(t) = 4.5t + 47$; $P(50) = 272$ thousand; average error ≈ 6.29
   (b) $P(t) = 4.47t + 44.4$; $P(50) \approx 268$ thousand; average error ≈ 5.33

3. (a) $P(t) = 5t - 11$; $P(50) = 239$ thousand; average error ≈ 11.06
   (b) $P(t) = 4.4t + 2.2$; $P(50) \approx 222$ thousand; average error ≈ 7.00

4. (a) $P(t) = 5.2t + 48$; $P(50) = 308$ thousand; average error ≈ 17.11
   (b) $P(t) = 6t + 40$; $P(50) = 340$ thousand; average error ≈ 10.04

5. (a) $S(t) = 81.7t + 130.14$
   (b) $S(7) = 702.04$ million. This prediction is below what the actual sales were in 1995.
   (c) $S(14) = 1273.9$ million
   (d) The prediction for 1995 is more accurate because it is between two known data points. The year 2002 is six years after the last data point.

6. (a) $N(t) = 2.29t + 21.70$
   (b) N(60) ≈ 159.3 million; $N(70) \approx 182.2$ million

7. (a) $S(v) = 0.08v + 1.77$
   (b) $S(18) \approx 3.21$ steps/sec; $S(10) \approx 2.57$ steps/sec
   (c) We have more confidence in the prediction for the speed of 18 ft/sec since it is within the scope of the given data. The speed of 10 ft/sec is much lower than the lowest speed given.

8. (b) $W(h) = 7.33h - 348.33$
   (c) $W(72) = 179.33$ pounds
   (d) No, a professional baseball player is not a typical American male.