1. The least-squares regression line

Often a finite set of data points \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \) exhibits a linear pattern (more or less). One would like to have an agreed-upon criterion to decide if one line fits the data better than another, with an eye to finding the line that “best fits” the points with respect to the criterion. The agreed-upon way to measure goodness of fit of the points to a line with equation \( y = mx + b \) is to look at the sum of the squares of the vertical distances from each point to the line, that is, \( \sum_{i=1}^{n} (y_i - (mx_i + b))^2 \). Consider the data set \( \{(0, 5), (3, 4), (5, -1), (6, 1), (7, -2)\} \) with respect to the line \( y = 6 - x \) as shown in Figure 1 below. The line looks like a reasonable one in that some of the points are above the line and some are below.

More precisely, with respect to our sum of the squares criterion, the goodness of fit is

\[
(5 - 6)^2 + (4 - 3)^2 + (-1 - 1)^2 + (1 - 0)^2 + (-2 - (-1))^2 = 1 + 1 + 4 + 1 + 1 = 8.
\]

It turns out that using techniques of calculus (which are beyond the scope of this course), one can find the slope and \( y \)-intercept of the line of best fit using some simple arithmetic. The line is also unique. These are reasons why this particular line is used. We call it the least-squares regression line. Another reason for its use is because it has this very nice property: the line always passes through \((\bar{x}, \bar{y})\), where \( \bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \) and \( \bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \). If we looked instead for a line that minimized the sum of the vertical distances, it would not necessarily pass through this point - or even be unique. Furthermore, finding such a line presents serious computational challenges.

Writing \( y = mx + b \) for the least-squares regression line, it can be shown that its slope is given by

\[
m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},
\]

and once that is calculated we get \( b = \bar{y} - m\bar{x} \) because \((\bar{x}, \bar{y})\) is a point on the line.

We return to the data set initially under discussion.

**Example 1.1.** Find the equation of the least squares regression line to the data set \( \{(0, 5), (3, 4), (5, -1), (6, 1), (7, -2)\} \).

**Solution:** We can easily verify that \( \bar{x} = 4.2 \) and \( \bar{y} = 1.4 \). We make a table whose ultimate purpose is to compute \( \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \) and \( \sum_{i=1}^{n}(x_i - \bar{x})^2 \).
The entries in the last row are obtained by summing the entries in the columns directly above. From these, we get \( m = -30.4 + 30.8 = -0.8 \) and as \( 1.4 = -0.987(4.2) + b = -4.15 + b \), we get \( b = 5.55 \). The line of best fit is \( y = -0.987x + 5.55 \), so that our rough guess \( y = 6 - x \) was not so far off after all!

It may be the case that our data set does not exhibit a linear pattern but some other pattern instead. In two important cases we can use our techniques to fit the data to a line by making a change of variables. Suppose that our data seems well-suited to an exponential model, i.e., a function of the form \( y = ar^x \). Applying the natural logarithm to both sides, we obtain

\[
\ln y = \ln a + x \cdot \ln r.
\]

Thus, putting \( Y = \ln y \), we see that \( Y \) is a linear function of \( x \). So if we fit the transformed data set \( \{(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)\} \) by a least-squares regression line, this leads to an exponential curve that fits the original data. Note that this curve will generally not minimize the sum of squared errors of the original data. Instead, it will minimize the sum of squared errors for the transformed data. Nevertheless, the resulting exponential curve is a good fit to the original data.

**Example 1.2.** Fit an exponential function to the data set \( \{(0,4), (1,6), (3,28), (4,49)\} \).

**Solution:** We work with the transformed data set

\[
\{(0, \ln 4), (1, \ln 6), (3, \ln 28), (4, \ln 49)\} = \{(0, 1.38), (1, 1.79), (3, 3.33), (4, 3.89)\},
\]

from which we get \( \bar{x} = \frac{1}{4}(0 + 1 + 3 + 4) = 2 \) and \( \bar{Y} = \frac{1}{4}(1.38 + 1.79 + 3.33 + 3.89) = 2.60 \). We now form a table analogous to the previous one where we compute \( \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) \) and \( \sum_{i=1}^{n}(x_i - \bar{x})^2 \).

<table>
<thead>
<tr>
<th>( i )</th>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( Y_i )</th>
<th>( x_i - \bar{x} )</th>
<th>( Y_i - \bar{Y} )</th>
<th>( (x_i - \bar{x})(Y_i - \bar{Y}) )</th>
<th>( (x_i - \bar{x})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1.38</td>
<td>-2</td>
<td>-1.22</td>
<td>2.44</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6</td>
<td>1.79</td>
<td>-1</td>
<td>-0.81</td>
<td>0.81</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>28</td>
<td>3.33</td>
<td>1</td>
<td>0.73</td>
<td>0.73</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>49</td>
<td>3.89</td>
<td>2</td>
<td>1.29</td>
<td>2.58</td>
<td>4</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td></td>
<td></td>
<td></td>
<td>-3.0</td>
<td>-1.28</td>
<td>6.56</td>
<td>10</td>
</tr>
</tbody>
</table>

Using this table to produce the best fit line \( Y = mx + b \), we get \( m = \frac{5.56}{10} = 0.556 \). Since \( (2, 2.60) \) is on the line, we can substitute the values and solve for \( b \) to obtain \( Y = 0.656x + 1.29 \). Now we need to transform this equation back to get the best fitting exponential curve \( y = ar^x \). We use that

\[
\ln y = Y = 0.656x + 1.29 = (\ln r)x + \ln a,
\]

so matching up the respective terms we see that \( \ln a = 1.29 \) and \( \ln r = 0.656 \). This yields \( a = e^{1.29} = 3.63 \) and \( r = e^{0.656} = 1.92 \). In summary, the desired exponential function is \( y = 3.63(1.92)^x \).

Least-squares lines can also be indirectly used to model data using power functions of the form \( y = ax^r \). (Note that for power functions the variable \( x \) is the base, while for exponential functions the variable \( x \) occurs in the exponent.) Taking natural logarithms of both sides, we get \( \ln y = \ln a + r \ln x \), that is, \( Y = \ln y \).
is a linear function of $X = \ln x$. Once more we can fit a linear function to the transformed data, and then solve for the values of $a$ and $r$. As with the exponential model, the resulting power function model will not minimize the sum of squared errors for the original data, but will be a good fit for the original data. We illustrate this with an example of data on cumulative AIDS cases.

**Example 1.3.** (Cumulative AIDS cases) In the early years of the disease, the cumulative number of AIDS cases in the United States could be modeled by an allometric (or power) model of the form $y = ax^r$, where $x$ is the number of years after 1980 and $y$ is the number of cumulative AIDS cases. The data on the number of AIDS cases is given in the table below (from page 21, E. K. Yeargers, R. W. Shonkwiler, and J. V. Herod, 1996, An Introduction to the Mathematics of Biology: with Computer Algebra Models, Birkhuser, Boston). Find the best-fitting power model for this data, and predict the cumulative number of AIDS cases in the year 1995 under this model.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$x$</th>
<th>$y$</th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97</td>
<td>4</td>
<td>6,928</td>
<td>7</td>
<td>52,902</td>
</tr>
<tr>
<td>2</td>
<td>709</td>
<td>5</td>
<td>15,242</td>
<td>8</td>
<td>83,903</td>
</tr>
<tr>
<td>3</td>
<td>2,698</td>
<td>6</td>
<td>29,944</td>
<td>9</td>
<td>120,612</td>
</tr>
<tr>
<td>10</td>
<td>161,711</td>
<td>11</td>
<td>206,247</td>
<td>12</td>
<td>257,085</td>
</tr>
</tbody>
</table>

Solution: We first graph the data to see whether the power model is appropriate. If it is, then the graph of the transformed data set consisting of data points $(X_i, Y_i) = (\ln x_i, \ln y_i)$ should be close to a straight line. Figure 2 shows the original and transformed data. Indeed, the graph of the transformed data is close to a straight line, so the power model is appropriate.

![Figure 2. On the left, cumulative AIDS cases in years after 1980. On the right, transformed data where years after 1980 and cumulative AIDS cases are replaced by their natural logarithms.](image)

We now proceed to find the best-fit regression line to the transformed data $(X_i, Y_i) = (\ln x_i, \ln y_i)$. The first two transformed data values (to 3 decimal places) are $(\ln 1, \ln 97) = (0, 4.575)$ and $(\ln 2, \ln 709) = (0.693, 6.564)$. We compute $\bar{X} = 1.666$ and $\bar{Y} = 9.869$. 


From the values in the last row, we compute \( m = 20.57 \div 6.28 = 3.2755 \). Substituting \( \bar{X} = 1.666 \) and \( \bar{Y} = 9.869 \) into the equation \( Y = mX + b \) and solving for \( b \) gives \( b = 4.412 \), so

\[
\ln y = Y = 3.2755X + 4.412 = r \ln x + \ln a,
\]

so \( r = 3.2755 \) and \( \ln a = 4.412 \). Solving the latter equation for \( a \) we obtain \( a = e^{4.412} = 82.4342 \), so the best power model to the data is given by \( y = ax^r = 82.4342x^{3.2755} \). Figure 3 shows the best-fit regression line to the transformed data on the left and the power model resulting from it with the original data on the right. Note that more advanced methods can produce an even better fitting power model, but those are beyond the scope of these notes.

![Figure 3](image-url)

**Figure 3.** On the left, transformed data with least-squares regression line. On the right, cumulative AIDS cases in years after 1980 with the model derived from the linear regression line of the transformed data.

The obtain the predicted cumulative number of AIDS cases for the year 1995, we substitute \( x = 1995 - 1980 = 15 \) into the model equation: \( y = 82.4342 \cdot 15^{3.2755} = 586,671 \).

**Exercises on least-squares regression lines**

1. Find the equation of the least-squares regression line for these data sets. Graph the data together with their least-squares regression line.
   
   (a) \( \{(3,5), (5,5), (10,6)\} \)  
   (b) \( \{(3,7), (-1,-2), (4,10), (7,16)\} \)  
   (c) \( \{(0,0), (6,3), (-4,-1), (8,3)\} \).

2. Fit an exponential function to the following data sets. Graph the data together with the function obtained from the least-squares regression line for the transformed data.
   
   (a)\( \{(0,1.5), (1,5), (2,30), (3,120)\} \)  
   (b)\( \{(-1,6), (0,3), (2,.2), (3,.08)\} \)
3. In 1967, bald eagles were listed under the Endangered Species Preservation Act, and as a result, the number of breeding pairs was monitored closely. Conservation methods and the ban of DDT have helped the growth of the eagle population, and in 2007, bald eagles were removed from the list of endangered species. The table below gives the number of bald eagle breeding pairs in the lower 48 states for selected years. (Source: U.S. Fish & Wildlife Service). Fit an exponential function to the data and use it to predict the number of breeding pairs for 2016.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Pairs</td>
<td>487</td>
<td>791</td>
<td>1188</td>
<td>3035</td>
<td>6471</td>
<td>9789</td>
</tr>
</tbody>
</table>

4. Fit a power function to the following data sets. Graph the data together with the function obtained from the least-squares regression line for the transformed data.

(a)\{(1, 1), (2, 7), (3, 40)\}  
(b)\{(5, 2), (9, 3), (27, 5), (45, 7)\}.

5. Below is a table that shows measurements of the length [in cm], mass [in grams], and surface area [in cm²] of several dogs.

<table>
<thead>
<tr>
<th>Length</th>
<th>Mass</th>
<th>Surface Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>3460</td>
<td>2245</td>
</tr>
<tr>
<td>63</td>
<td>5150</td>
<td>3250</td>
</tr>
<tr>
<td>72</td>
<td>5350</td>
<td>3865</td>
</tr>
<tr>
<td>76</td>
<td>9980</td>
<td>5000</td>
</tr>
<tr>
<td>97</td>
<td>17220</td>
<td>7900</td>
</tr>
<tr>
<td>101</td>
<td>26070</td>
<td>8995</td>
</tr>
<tr>
<td>105</td>
<td>33260</td>
<td>10535</td>
</tr>
</tbody>
</table>

(a) Make a graph with length as the independent variable and mass as the dependent variable. What type of model seems to fit this data?
(b) Make a graph of \((x, Y) = (x, \ln y)\) (exponential model) and a graph of \((X, Y) = (\ln x, \ln y)\) (power model). Which of these graphs is closer to a straight line?
(c) For the model for which the transformed data is closer to a straight line, perform the computations necessary to find the least-squares regression line.
(d) From the least-squares regression line, find the appropriate constants for the function that fits the original data.
(e) Predict the mass of a dog that is 120 centimeters long.
(f) Repeat (a) - (d) using length as the independent variable and surface area as the dependent variable.
(g) Predict the surface area of a dog that is 120 centimeters long.
(h) The graph below shows surface area versus mass of the dog, and the data is reasonably close to a straight line. Find the least-squares regression line for these pairs of data. Draw it into the graph of the data. What does the linear model predict as the surface area of a dog with a mass of 15,000 grams?

![Graph showing surface area versus mass of a dog. The data points are scattered, and the regression line is a straight line.]
6. (a) Show that \( y = \frac{1}{3} \) is the least-squares line for the data set \( \{(0, 0), (1, 1), (2, 0)\} \).
(b) Show that the line \( y = 0 \) is a line that better fits the data set with respect to minimizing the sum of vertical distances from the points to the line.

2. Expected Value

We often deal with a random quantity, such as the outcome of a die roll, that we cannot predict in advance even though we know what the possible results are and how likely each result is. If each result amounts to a numeric quantity (such as the value of a die roll), we are often interested in what the average result is - that is, the result that would be obtained on average if the random event were repeated many times over. This is called the expected value of that quantity. For example, life insurance companies take great interest in the expected value of how long people will live, depending on their age, sex, health, and so on. Let us review how an average is computed and motivate the definition of the expected value by looking at the average grade of students in Calculus at Cal State LA.

Example 2.1. The table below gives the grade distribution of the 3,111 students who completed a calculus course from Fall 2014 through Fall 2015 at Cal State LA (Source: Cal State LA Institutional Research.)

What is the average grade in a calculus course?

<table>
<thead>
<tr>
<th>Letter grade</th>
<th>A</th>
<th>A-</th>
<th>B+</th>
<th>B-</th>
<th>C+</th>
<th>C-</th>
<th>D+</th>
<th>D-</th>
<th>F or WU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade point</td>
<td>4.0</td>
<td>3.7</td>
<td>3.3</td>
<td>3.0</td>
<td>2.7</td>
<td>2.0</td>
<td>1.7</td>
<td>1.3</td>
<td>1.0</td>
</tr>
<tr>
<td>Number of students</td>
<td>431</td>
<td>112</td>
<td>123</td>
<td>426</td>
<td>134</td>
<td>168</td>
<td>657</td>
<td>91</td>
<td>67</td>
</tr>
</tbody>
</table>

Solution: The average grade, \( AG \), is computed as the sum of the grade points, divided by the number of students:

\[
AG = \frac{431 \times 4.0 + 112 \times 3.7 + 123 \times 3.3 + 426 \times 3.0 + \cdots + 353 \times 1.0 + 47 \times 0.7 + 502 \times 0.0}{3111} = 2.09.
\]

That is, on average, a student taking a calculus course would receive a grade just above a C.

Note that we can rewrite equation above by breaking up the fraction according to the values of the grade points, so that each grade point value is multiplied with its respective relative frequency, such as \( \frac{431}{3111} \):

\[
AG = \frac{431 \cdot 4.0 + 112 \cdot 3.7 + 123 \cdot 3.3 + 426 \cdot 3.0 + \cdots + 353 \cdot 1.0 + 47 \cdot 0.7 + 502 \cdot 0.0}{3111} = 2.09.
\]

The relative frequencies can be used as an estimate of the probability that the corresponding grade point value will be awarded to a student taking a calculus course. This motivates the definition of the expected value of a random quantity as the sum of the possible values, each multiplied by the probability that it will occur.

In the general setting, when we conduct an experiment where each outcome can be associated with a numerical value, we group the outcomes into subsets, called events, such that each outcome in an event has the same numerical value. In the example above, the outcomes are the letter grades, and the associated numerical values are the grade points. Suppose these events are \( A_1, A_2, \ldots, A_k \) and the value assigned to each outcome of \( A_i \) is \( v_i \). Then the expected value or expectation \( E \) of these values is defined by the formula

\[
E = v_1 P(A_1) + v_2 P(A_2) + \cdots + v_k P(A_k) = \sum_{i=1}^{k} v_i P(A_i).
\]

The expected value is also called the mean. Let’s look at a few examples.

Example 2.2. Three fair coins are tossed. What is the expected number of heads obtained in three coin tosses?
Solution: Here each outcome has probability $\frac{1}{8}$. The number of heads that are possible are 0, 1, 2, or 3. The associated events are \{TTT\}, \{TTH, THT, HTT\}, \{HHT, THH, HTH\}, and \{HHH\}. The expected value is thus

$$E = 0 \cdot P(\{TTT\}) + 1 \cdot P(\{TTH, THT, HTT\}) + 2 \cdot P(\{HHT, THH, HTH\}) + 3 \cdot P(\{HHH\})$$

$$= 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2}$$

This makes intuitive sense in that on the average, each flip of a fair coin should result in $\frac{1}{2}$ head, and so three flips on the average should result in $\frac{3}{2}$ heads.

**Note:** The average or expected value does not have to be one of the possible values that can occur. However, the average value is always between the smallest and the largest possible value. In the example above, the expected value has to be between 0 and 3.

**Example 2.3.** In a city, a survey was taken of women in their forties as to how many husbands they had had. The results: 41 percent had never been married, 33 percent had had one husband, 17 percent had had two husbands, 6 percent had had three husbands, and 3 percent had had 4 husbands. Find the expected number of husbands of a woman chosen at random from this survey group.

Solution: We have to convert percents to decimals; for example, the probability of having had 3 husbands is .06. With this in mind,

$$E = 0(.41) + 1(.33) + 2(.17) + 3(.06) + 4(.03) = .33 + .34 + .18 + .12 = .97$$ husbands.

That is, on average, a randomly selected woman from the group has had approximately one husband.

**Example 2.4.** Two fair dice are rolled. Compute the expectation for the difference for their upturned faces.

Solution: We represent outcomes as ordered pairs of upturned faces. Thus (1,5) means that the first die shows 1 and the second shows 5. In particular, (1,5) is a different outcome from (5,1). The possible values for the difference are 0, 1, 2, 3, 4, and 5. These are associated with the events $A_0, A_1, \ldots, A_5$ described below:

$$A_0 = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$$
$$A_1 = \{(1,2), (2,1), (2,3), (3,2), (3,4), (4,3), (4,5), (5,4), (5,6), (6,5)\}$$
$$A_2 = \{(1,3), (3,1), (2,4), (4,2), (3,5), (5,3), (4,6), (6,4)\}$$
$$A_3 = \{(1,4), (4,1), (2,5), (5,2), (3,6), (6,3)\}$$
$$A_4 = \{(1,5), (5,1), (2,6), (6,2)\}$$
$$A_5 = \{(1,6), (6,1)\}.$$

Since there are 36 outcomes each having probability $\frac{1}{36}$, we compute

$$E = 0 \cdot \frac{6}{36} + 1 \cdot \frac{10}{36} + 2 \cdot \frac{8}{36} + 3 \cdot \frac{6}{36} + 4 \cdot \frac{4}{36} + 5 \cdot \frac{2}{36}$$

$$= \frac{10 + 16 + 16 + 10 + 8 + 10}{36} = \frac{70}{36} = 1.94.$$
earns 80 singles, 28 doubles, 4 triples and 16 home runs. What is the expected number of bases the batter achieves per time at bat?

Solution: The numerical values that can be achieved here are 0, 1, 2, 3, or 4. Clearly, the batter earns no bases in $400 - (80 + 28 + 4 + 16) = 272$ official times at bat. We compute

$$E = 0 \cdot \frac{272}{400} + 1 \cdot \frac{80}{400} + 2 \cdot \frac{28}{400} + 3 \cdot \frac{4}{400} + 4 \cdot \frac{16}{400} = .20 + .14 + .03 + .16 = .53 \text{ bases.}$$

The baseball jargon for the expected number of bases earned is *slugging percentage*.

**Exercises on expected value**

1. Four fair coins are tossed with 16 possible outcomes.
   (a) Find the probability of getting exactly two heads.
   (b) Find the expected value of the number of heads.

2. One fair die is rolled. Find the expected value of the upturned number.

3. Two fair dice are rolled. Find the expected value for the sum of their upturned faces.

4. One card is drawn from a deck (see image below). Assigning a value 1 to an ace, a value 10 to a face card, and the numerical value to any number card,
   (a) Find the probability that a value of 10 is achieved.
   (b) Find the expected value of the number obtained.

![Figure 4. A deck of playing card consists of 4 suits, with 13 cards each.](image)

5. A player draws one card from a deck. If it is an ace, the player wins 5 dollars. If it is a face card, the player wins 2 dollars. Otherwise, the player loses 1 dollar. Thus the values that can be obtained in this game are 5, 2, and -1.

   (a) What is the expected value of this game to the player?
   (b) Adjust the value earned if an ace is drawn so that the expected value is zero (the game is then called fair).

6. In 500 official times at bat, a baseball player has 90 singles, 25 doubles, 5 triples, and 10 home runs.
(a) What is the probability that the player does not make an out? Note: this is called his batting average.
(b) Find the player’s slugging percentage.

7. Consider the spinner shown on page 784 of Blitzer. Suppose that if the spinner lands on the yellow region you win 8 dollars, if it lands on one of the green regions you win 5 dollars, and otherwise you lose 3 dollars. Find the expectation of this game.

8. Consider the square target shown on page 785 of Blitzer. If the dart hits the yellow area you win 5 dollars and if it hits the white area you lose 2 dollars. Find the expectation of this game.

9. Consider the game of roulette as described by Blitzer on page 782.

(a) If a player bets 1 dollar on red and the ball lands on red, the player wins a dollar. Otherwise the bet is lost. Find the expected value of this game.
(b) If a player bets 1 dollar on the two green numbers, and the ball lands on either 0 or 00, then the player wins 17 dollars. Otherwise the bet is lost. Find the expected value of this game.
(c) If a player bets 1 dollar on the numbers \{1, 2, 3, \ldots, 12\} as a group, and one of these numbers is obtained, the player wins 2 dollars. Otherwise the bet is lost. Find the expected value of this game.

What do you guess happens for all bets one might place in roulette?

10. 100 married couples in Atlanta were surveyed as to the number of children they had. The results: 35 had no children, 32 had one child, 20 had two children, 10 had three children, and 3 had four children. Find the expected number of children of a couple selected randomly from the survey group.