Mathematical Models & Linear Statistical Models

Basic Concepts & Computations

Nicholas Bennett
nick@nickbenn.com

October 11, 2018

Copyright and License Information

Things Change

(Adapted with permission from "Things Change" by the Maryland Collaborative for Teacher Preparation [1].)

Introduction

The only thing certain in life is change. From birth we each grow taller, heavier (and sometimes lighter), older, wiser, richer (and sometimes poorer). We live in human communities with populations that are changing minute by minute through births and deaths and immigration, and at any instant of time many of those people are in motion on foot, bicycles, cars, buses, trains, and airplanes. The physical features of the world around us are in constant motion—pushed by forces of wind and water and gravity—and our planet Earth is racing around the sun at nearly 1,700,000 miles per day. In our economic lives the prices we pay for food, clothes, shelter, transportation, education, and entertainment go up and down in response to consumer demand and producer supply.

Many of the most important problems in mathematics beyond arithmetics require description and prediction of changes in related quantitative variables—in other words, construction and use of models of change. In some cases those problems involve analysis of changes in variables as time passes; in other cases the problem is to understand the ways that changes in some variables cause changes in other variables. Algebra and calculus are at the heart of this study of change.

Patterns of Change

When we have information about a relationship between two or more variables, one of the best ways to present that information is with coordinate graphs of (x, y) data pairs. Such graphs may display specific pairs of related numerical values, a line or curve representing the general relationship between the x and y values, or both.

The following statements describe nine different situations in which two variables are (or at least seem to be) related to each other. Match each situation to the graph that you believe is most likely to represent the relation between those variables. Then explain as carefully as you can what the shape of the graph tells about the ways the variables change in relation to each other.

1. When a tennis player hits a high lob shot, its height changes as time passes. What pattern seems likely to relate time and height?
2. The senior class officers at Lincoln High School decided to order and sell souvenir baseball caps with the school insignia, name, and "Class of '95" on them. One supplier said it would charge $100 to make the design and then an additional $4 for each cap made. How would the total cost of the order be related to the number of caps in the order?
3. The population of the world has been increasing for as long as data or estimates have been available. What pattern of population growth has occurred over that time?

4. In planning a bus trip to Florida for spring break, a travel agent worked on the assumption that each bus would hold at most 40 students. How would the number of buses be related to the number of student customers?

5. The depth of water under the U. S. Constellation in Baltimore Harbor changes due to tides as time passes in a day. What pattern would that (time, depth) data fit?

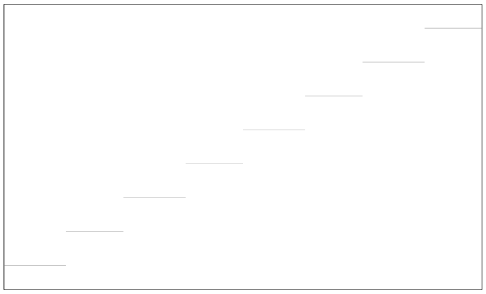
6. When the Lincoln High School class officers decided to order and sell t-shirts with names of everyone in the Class of `95, they checked with a sample of students to see how many would buy at various proposed prices. How would sales be related to price charged?

7. How does the height of a bungee jumper vary as time passes in the jump?

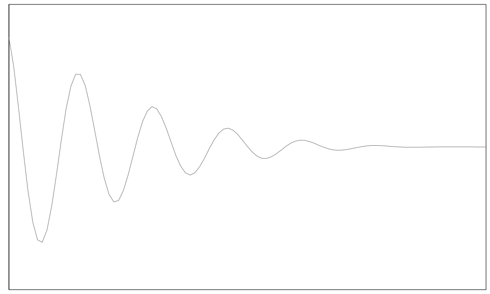
8. In a wildlife experiment, all fish were removed from a lake and the lake was restocked with 1000 new fish. The population of fish then increased over the years as time passed. What pattern would likely describe change in fish population over time?

9. According to *The Old Farmer's Almanac*, if you live near crickets, you can estimate the nighttime outdoor temperature in degrees Fahrenheit by counting the number of cricket chirps in 14 seconds and adding 40 to that number. If you tested that by gathering data and plotting chirps vs. temperature, and the data seemed to support the rule of thumb, what might the graph of observations look like?

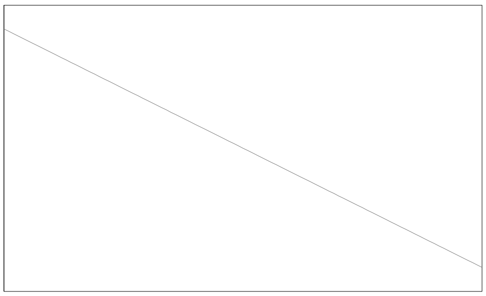
For each of the scenarios above, select the graph from the following page that you think most closely matches the pattern or relationship described. Try not to focus on the units or scales used on the axes of each graph (in fact, none of the graphs have such notations), but on the general nature of the relationships between the quantities described.



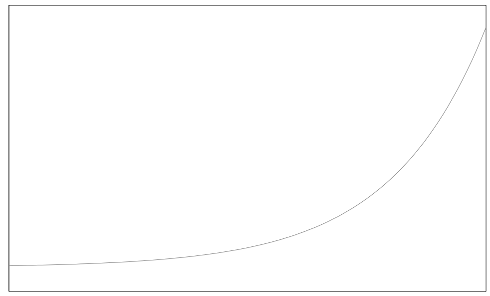
A



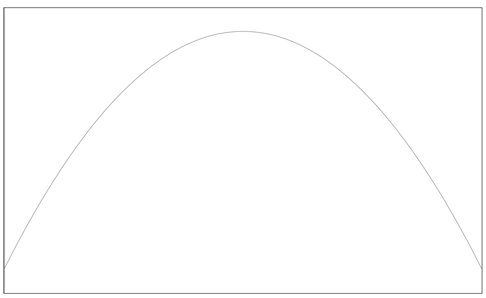
B



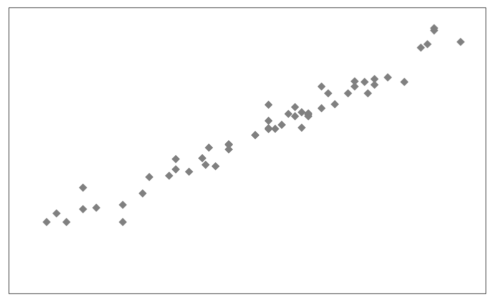
C



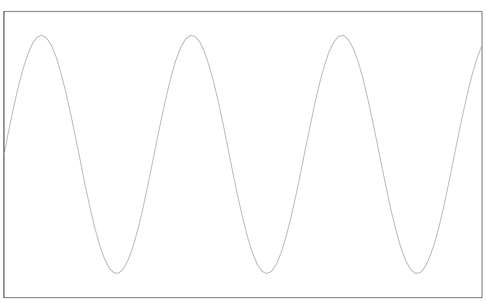
D



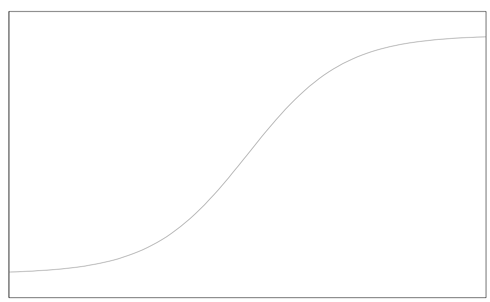
E



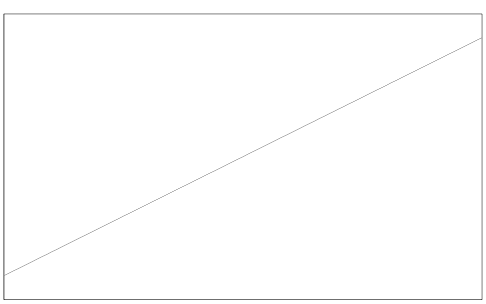
F



G



H



I

Mathematical Models

Introduction

When the relationships between the quantities of a real-world problem can be expressed in mathematical formulas, we refer to those formulas—along with the “key” that maps the variables and constants in the formulas to real-world quantities and units of measure—as a *mathematical model* of the problem.

Example

While exiting the Lunar Module, Apollo 14 mission commander Alan Shepard rolls a golf ball off the platform at the top of the ladder.[1] The golf ball leaves the platform moving horizontally and falls to the Moon's surface from a height of 3 meters. The Moon's gravitational acceleration close to the surface is 1.63 meters per second per second. (In other words, after 1 second, a body will be falling at 1.63 meters per second; after 2 seconds, it will be falling at 3.26 meters per second; and so on.) Since the Moon's atmosphere is so thin, we can ignore atmospheric drag.

If we want to know how the height above the surface of the golf ball at some time *t* after it leaves the platform, up to the moment it hits the ground, we can start with the general formulas for a body falling in a vacuum, and adapt them to our problem.

Given

$$\begin{aligned} g &= \text{gravitational acceleration} \\ t &= \text{time} \\ v_0 &= \text{initial vertical speed} \\ h_0 &= \text{initial height} \\ v_t &= \text{vertical speed at time } t \\ h_t &= \text{height at time } t \end{aligned} \tag{1}$$

Then

$$\begin{aligned} v_t &= v_0 + g\,t \\ h_t &= h_0 + v_0 t + \frac{g}{2} t^2 \end{aligned} \tag{2}$$

For our scenario,

$$\begin{aligned} g &= -1.63 \text{ meters/sec}^2 \\ v_0 &= \quad 0 \text{ meters/sec} \\ h_0 &= \quad 3 \text{ meters} \end{aligned} \tag{3}$$

1 In reality, while Alan Shepard famously used a makeshift golf club to hit two golf balls on the Moon, he didn't roll any golf balls off the LM platform, as far as we know.

Taken together, (1), (2), and (3) make up a mathematical model that can be used to answer a number of questions about our hypothetical golf ball. We can even set $h_t = 0$ and use the quadratic formula to solve the second equation of (2) for *t*, to find out how long it will take the ball to reach the ground.

Mathematical Formulas as Patterns of Change

By itself, (2) might not be a very useful model, since without (1) it might not be clear what the different variables refer to, and since it doesn't include the information in (3) that's specific to our scenario. But general formulas like those in (2) are the essential core of a mathematical model: just as a graph visually conveys the relationship between variables, a formula does the same thing symbolically.

See if you can match the equations shown below to the corresponding scenarios or graphs in "Patterns of Change". While none of the formulas is a complete model, some are tied very specifically to the corresponding scenarios. Others fit the scenarios to some degree (at least in the general shape), but might not express the underlying mathematics accurately. Some are written in a general form, where c_0, c_1, etc. represent constant values in the equations.

This is intended to be a challenging exercise, especially when it comes to the last few formulas. Don't worry if you don't understand all of the symbols and notation used[2]; see if you can guess their meaning from their appearance and use. If you can't match all of the formulas to scenarios or graphs, keep in mind that it's possible that some of the formulas don't correspond to any of the scenarios or graphs, and vice versa. On the other hand, it's also possible that one (or more) of the general formulas matches more than one scenario, or that more than one formula can be matched with a single scenario.

i. $y = 4x + 100$

ii. $y = -16(x - 2)^2 + 70$

iii. $y = \left\lceil \frac{x}{40} \right\rceil$

iv. $y = x - 40 + \varepsilon$

v. $y = c_0 + c_1 x$

vi. $y = c_0 + c_1 \sin x$

vii. $y = c_0 + c_1 \frac{\cos x}{x + 1}$

viii. $y = \frac{c_0}{1 + e^{c_1 - x}}$

ix. $y = c_0 e^{c_1 (x + c_2)}$

2 A review of some of the symbols and their usage is found in Appendix B.

Statistical Models

Introduction

The final “Patterns of Change” scenario (page 4) mentions a relationship claimed to exist between temperature and the rate of cricket chirps. Of course, even if such a relationship exists, there could be many other factors besides temperature that affect the rate of crickets' chirps; because of this, creating a formal model for this scenario presents a challenge: Even if we analyze the data very carefully, and quantity all the factors we can, we won't be able to predict the outcome entirely; some portion will remain unpredictable. This unpredictable part might be due to known factors that are beyond our practical ability to measure or control them, or it might be due to factors we're not even aware of. Whatever the reason, when we build mathematical models that acknowledge or include some randomness, uncertainty, or unpredictability in the *dependent* (output) variables, given known *independent* (input) variable values, we call them *statistical models*.

Statistical Inference

Statistical models are usually constructed by *inference*. There's not a formal definition of the statistical inference process, but we can think of it generally as:

1. Collect, explore, and analyze data.
2. Quantify relationships in the data.
3. Test the model elements—i.e. the relationships quantified in step 2.
4. If the relationships aren't sufficiently supported by the test results, use those results to revise assumptions and search for other relationships.
5. Repeat steps 1–4 as appropriate.

This might sound a bit like what you've learned as the scientific method. In fact, statistical inference is often an important part of scientific inquiry.

Value of Statistical Models

Can a statistical model still be useful, or does the inherent uncertainty make such a model too inaccurate? That's not an easy question to answer. In part, it depends on how well the model fits the data. So we not only need to quantify the relationships—we also need to quantify how well we can quantify those relationships!

Ultimately, however, the answer boils down to our needs. A model that leaves a majority of the variation in the dependent variable unexplained may be sufficient for some purposes, while other uses may require a model that explains almost all of that variation.

Types of Statistical Models

Just as the set of even the most commonly used types of mathematical models is far too broad to show in "Patterns of Change", there's a wide variety of commonly used statistical model types, and we can only hope to touch on a few here. For example, in a *polynomial* model, the dependent variable is expressed as a polynomial function of one or more independent variables (a simple form of the polynomial model is the *linear* model, where there are no terms with degree higher than 1); in a *cyclical* time series model, the dependent variable cycles in value over time (the independent variable), either in a relatively smooth sinusoidal (sine wave-shaped) fashion, or as the sum of multiple sinusoidal cycles of different lengths; in an *autoregressive* time series model, the value of the dependent variable at one moment in time is used as an input value at a later moment, following a lag period. (Of course, there are also hybrid models that use combinations of these and other types.)

Discussion

1. Besides the cricket chirps scenario, which of the "Patterns of Change" scenarios (if any) could be described by statistical models?
2. Assume that we're developing statistical models for each of the following groups of variables. What are the independent and dependent variables in each? What type (or combination of types) of statistical model, from those mentioned briefly above, might best fit real data in each case?
   a) Daily low temperature in Socorro, New Mexico vs. day of the year.
   b) Populations of rabbits in the wild in New Mexico, measured or estimated monthly, based on the previous month's rabbit and coyote populations.
   c) Total monthly precipitation, measured at Denver International Airport, vs. month of the year.
   d) Annual number of hurricanes in the Atlantic ocean over the past 75 years.
   e) End-of-day value of the Dow Jones Industrial Average (a set of representative stocks listed on the New York Stock Exchange) over the past 50 years, taking into account long-term trends, seasonal variations throughout the year, and the fact that the ending value on one day strongly influences the early prices for those stocks in the following day.

Linear Models

Definition

One of the simplest statistical models is applicable to a wide range of problems. In the *linear model*, a dependent variable is expressed as a linear combination of independent variables and an error term:[3]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon, \tag{4}$$

where

X_i are the independent variables;

Y is the dependent variable;

$\beta_i \in \mathbb{R},\ i=0,1,2,\ldots$ (the coefficients are real numbers);

ε is the error term, a quantity not explained by the model.

Formally, the independent variables are assumed to be continuous over real value ranges; in practice, this condition is often relaxed to allow for integral or other discrete numeric values.

Simple Linear Models

In a simple linear model, there's only one independent variable and one dependent variable, so (4) becomes

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{5}$$

This is often written as

$$Y = \alpha + \beta X + \varepsilon. \tag{6}$$

As you've probably figured out already, a simple linear model is easy to show graphically, along with the actual data. It's essentially a straight line through the data points, fitting them as closely as possible—though we haven't yet said what "as closely as possible" really means.

Finding the Best Fit

How do we find the straight line that best fits the data? To illustrate the problem, let's look at actual data for cricket chirps and temperature, collected by Dr. Peggy LeMone [3]. We'll begin by expanding graph F (page 5), adding some details on the scale and units of measure (Figure 1, data from Appendix A).

3 In this usage, *linear model* is synonymous with *linear regression model*. In other contexts, the same term can refer to other model types—e.g. the *general linear model*, which allows for multiple dependent variables, linear combinations of functions (possibly non-linear) of the independent variables, and categorical values.

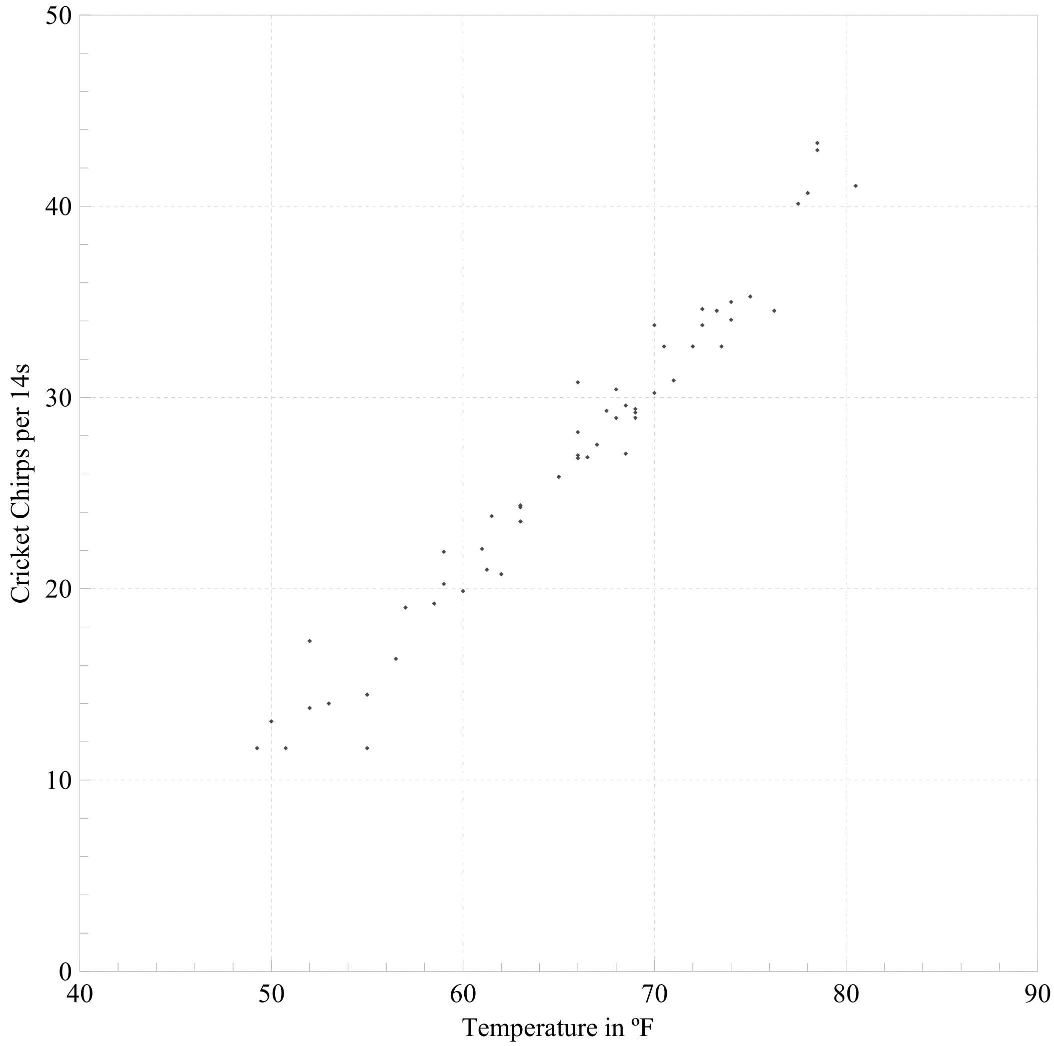


Figure 1: Cricket Chirps vs. Temperature

Arguably, the most important question to ask at this point is whether a linear model makes sense for the relationship between temperature and cricket chirp rate. Even if the answer is yes, we shouldn't assume that any apparent relationship will hold under conditions far outside the bounds of the observed data. For instance, it seems quite likely that very high or very low temperatures would have a disastrous effect on the crickets themselves—and consequently on their chirps.

A linear model seems appropriate in this case, for the range of temperatures observed. But two different people, fitting a straight line to the data by sight, will probably draw slightly different lines. How can we measure how well each line fits, so that we can select the best fitting one?

It might seem reasonable to evaluate the fit by adding up the differences (called *residuals* or *errors*) between the Y values on the line (denoted by $\hat{y}_i$) and the actual Y values for each line under consideration, and use the line with the sum closest to zero. However, for any non-vertical straight line that passes through $(\bar{x}, \bar{y})$, the positive and negative residuals cancel each other out, and the sum is zero. (For a proof of this, see Appendix D.)

Linear Least-Squares Regression

If we want to use the residuals to assess the quality of fit, we need to sum them in such a way that the negative values don't cancel out the positive values. One general approach is to treat all of the residuals as positive values in some way; for example, we could use the absolute values or the squared values of the residuals. Both approaches are used, but taking the squared values makes an analytical solution easier. The method of finding the best fitting line by minimizing the sum of the squared residuals is called *linear* (or *ordinary*) *least-squares regression*.[4]

If we denote our model estimates for α and β by *a* and *b*, respectively,[5] and the fitted line by

$$\hat{Y} = a + bX, \tag{7}$$

then the sum of the squared residuals (*sum of squared errors*, or SSE) for the model is given by

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2. \tag{8}$$

We could calculate SSE for multiple lines, and choose the line with the lowest value as the best fit among them. However, instead of comparing specific alternatives lines, we can use calculus to find the values of *a* and *b* that give the smallest possible SSE. Even better, we can derive general formulas for *a* and *b* that can be used with any appropriate data set. This is the aim of least-squares regression. The steps of the analysis can be found in Appendix D, but we'll summarize the results here, so that we can make use of them:

The values of *a* and *b* in (7) that give the minimum value of SSE in (8) can be computed as

$$\begin{aligned} a &= \frac{\sum x^2 \sum y - \sum x \sum x\,y}{n \sum x^2 - \left(\sum x\right)^2} \\ b &= \frac{n \sum x\,y - \sum x \sum y}{n \sum x^2 - \left(\sum x\right)^2} \end{aligned} \tag{9}$$

4 The linear least-squares method is most useful when the residuals aren't correlated with each other or with the independent variable, and when they follow a normal distribution. However, even when these conditions are not strictly satisfied, this method is still often used, at least as an exploratory tool.

5 The fact that *b* is typically used for the estimated slope in the linear regression model can cause some confusion, since the standard slope-intercept form of the equation of a line is $y = mx + b$, where *b* is the intercept.

A Simple Example

Let's use (9) to find a and b for a simple data set. Given the x and y values provided in Table 1, we begin by calculating and filling in the values in the xy and x^2 columns. Then, we sum the values in each column, and write the totals in the bottom row.

x	y	xy	x^2
0	-0.5	0	0
1	0	0	1
2	1.5	3	4
3	2	6	9
$\sum x = 6$	$\sum y = 3$	$\sum xy = 9$	$\sum x^2 = 14$

Table 1: Data and calculations for simple regression example

Finally, we can substitute the values from the bottom row of Table 1, along with the number of data points ($n = 4$), in place of the corresponding sums in (9).

$$a = \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - \left(\sum x\right)^2}$$

$$= \frac{14 \cdot 3 - 6 \cdot 9}{4 \cdot 14 - 6^2} = -\frac{12}{20} = -\frac{3}{5}$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \left(\sum x\right)^2}$$

$$= \frac{4 \cdot 9 - 6 \cdot 3}{4 \cdot 14 - 6^2} = \frac{18}{20} = \frac{9}{10}$$

We now have our fitted line, found by linear least-squares regression.

$$\hat{Y} = -\frac{3}{5} + \frac{9}{10} X \qquad (10)$$

To visualize the fit, we can plot the data points along with the fitted regression line (Figure 2). It's easy to see that the fit is pretty good—and because we used least-squares regression, we know that this fitted line minimizes SSE. But we're still not sure how to quantify *how good* the fit is.

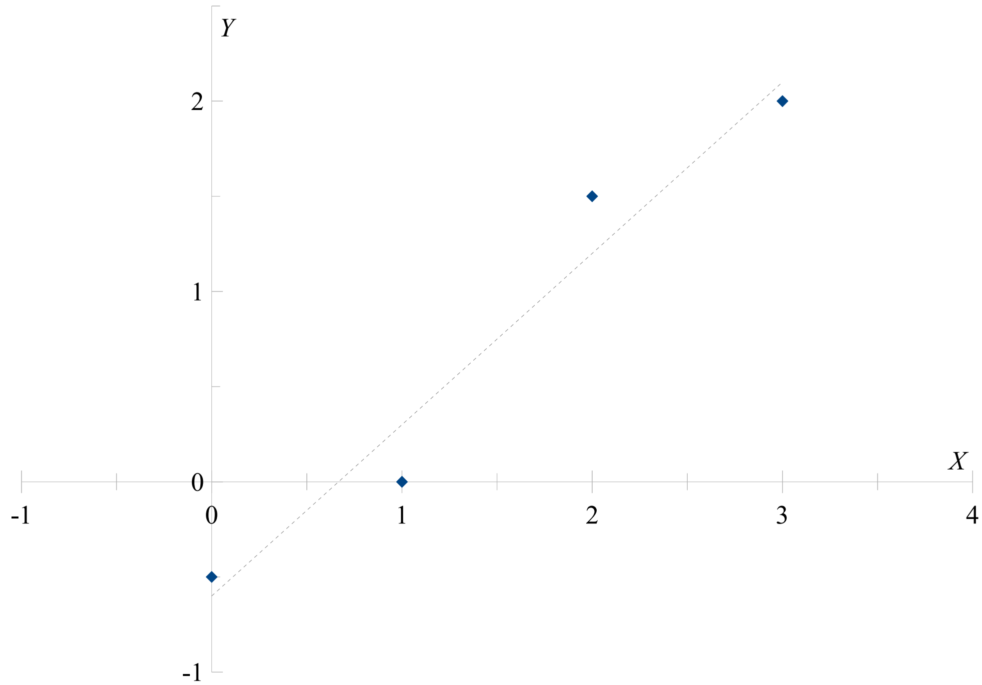


Figure 2: Simple example data points and fitted regression line

The Coefficient of Determination

To assess the quality of a model, or to compare alternative models, we usually need some way to quantify how well a model fits the data—to measure its *goodness-of-fit*. (This isn't the only type of measure of interest when evaluating a statistical model, but it's one of the most important.)

One way to measure goodness-of-fit is to measure how much of the change in the dependent variable is accounted for by the model. First, we need to quantify the total variation in the dependent variable; we can do this by summing the squared deviations of its observed values from its sample mean. We call this the *total sum of squares*, or SST.

$$\begin{aligned} \text{SST} &= \sum (y - \bar{y})^2 \\ &= \sum y^2 - \frac{\left(\sum y\right)^2}{n} \end{aligned} \tag{11}$$

SST can also be expressed as the sum of SSE and the *sum of squares of regression* (SSR).[6]

$$\text{SST} = \text{SSE} + \text{SSR}, \tag{12}$$

where SSE is given by (8), and

6 Unfortunately, while these abbreviations are used in many textbooks, some others define SSE and SSR with meanings opposite to these. In those texts, SSR is the *sum of squared residuals* (which we're calling the sum of squared errors), while SSE is the *sum of squares explained* (i.e. the regression sum of squares). Adding more confusion, still other texts use ESS for *explained sum of squares*, RSS for *residual sum of squares*, and TSS for *total sum of squares*. The best way to avoid confusion is to define these terms explicitly when using them.

$$SSR = \sum \left(\hat{y} - \bar{y} \right)^2. \tag{13}$$

The larger that SSR is in relation to SST, the more that the change in the dependent variable is explained by the model. This leads us to a useful goodness-of-fit measure: the *coefficient of determination*, or R^2.

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{14}$$

We can interpret R^2 as the fraction of the variation in the dependent variable that's determined or explained by the model.[7]

Simple Example Revisited: Computing the Coefficient of Determination

Let's calculate R^2 for the simple data set in Table 1. To begin, let's add a few more columns to the table, for y^2, $\hat{y}$, and $\left(y - \hat{y} \right)^2$. Fill in those columns using the original data and the fitted line equation given by (10); then, compute the sums for y^2 and $\left(y - \hat{y} \right)^2$ (Table 2).

x	y	xy	x^2	y^2	$\hat{y}$	$\left(y - \hat{y} \right)^2$
0	-0.5	0	0	0.25	-0.6	0.01
1	0	0	1	0	0.3	0.09
2	1.5	3	4	2.25	1.2	0.09
3	2	6	9	4	2.1	0.01
$\sum x = 6$	$\sum y = 3$	$\sum xy = 9$	$\sum x^2 = 14$	$\sum y^2 = 6.5$		$\sum \left(y - \hat{y} \right)^2 = 0.2$

Table 2: Calculations for coefficient of determination in simple regression example

Finally, we can use (8), (11), (12), and (14) to find R^2.

$$\begin{aligned} \text{SSE} &= \sum \left(y - \hat{y} \right)^2 \\ &= 0.2 \\ \text{SST} &= \sum y^2 - \frac{\left(\sum y \right)^2}{n} \\ &= 6.5 - \frac{3^2}{4} = 4.25 \\ \text{SSR} &= \text{SST} - \text{SSE} \\ &= 4.25 - 0.2 = 4.05 \end{aligned}$$

7 While R^2 is useful, it's often misused. It's easy to fall into the trap of adding more independent variables or polynomial terms to a model to increase R^2, at the expense of a loss of general predictive power. For this reason, an *adjusted* R^2, which discounts the increase from additional terms, is often used when building multivariate or polynomial models.

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$
$$= \frac{4.05}{4.25} \approx 0.95$$

From this, we conclude that approximately 95% of the total change in the dependent variable is determined or explained by the independent variable.

Discussion

1. Are there any other curves besides a straight line that fit the data in Table 1?
2. Should we consider an R^2 value of 0.95 to be high enough for a usable model?
3. Do you think a non-linear model could have a higher R^2 value in this case?
4. What other factors, besides a difference in R^2, might lead us to prefer a linear model over a non-linear model, or vice versa?

What about the crickets?

Even though the cricket data set isn't very large, we probably wouldn't want to do the chirps vs. temperature linear regression by hand. However, these calculations can be performed easily in a spreadsheet program; in fact, virtually all of the most widely used spreadsheet programs—including Microsoft Excel, Google Sheets, LibreOffice Calc, OpenOffice Calc, and Apple Numbers—include functions specifically for computing a, b, and $\hat{Y}$ in (7), from X and Y input values. If you have access to one of these programs, we suggest you attempt this analysis, using the data in Appendix A.

In the accompanying document, "Linear Statistical Models: Basic Computation with Python and SciPy", we'll use the Python programming language to analyze the cricket chirps data—first writing code to compute a, b, $\hat{Y}$, and R^2 according to formulas (9) and (14), then using the SciPy library to perform the equivalent analysis.

References

[1] Maryland Collaborative for Teacher Preparation. “Unit III: Things Change (Part A)”. Retrieved Sep. 21 2011 from http://www.towson.edu/csme/mctp/Courses/Mathematics/UnitIIIA.html

[2] “Cricket Chirps: Nature's Thermometer”. Retrieved Oct. 7, 2018 from http://www.almanac.com/cricket-chirps-temperature-thermometer

[3] M. A. LeMone. (Oct. 5, 2007). “Measuring temperature using crickets”. Retrieved Oct. 7, 2018 from https://www.globe.gov/explore-science/scientists-blog/archived-posts/sciblog/index.html_p=45.html

Appendix A: Cricket Chirps vs. Temperature

The following observations were recorded by Dr. Margaret LeMone in Boulder, Colorado, over a 30 day period in August and September, 2007 [3]. According to Dr. LeMone, the measurements were originally in chirps per 30 seconds (averaged over multiple successive observations, and halved for chirps per 15 seconds) and degrees Fahrenheit (taking the average reading from multiple thermometers); the column for chirps per 14 seconds was derived from the original data.

Date	Time	Chirps/15s	Chirps/14s	Temp (ºF)
21 Aug	2030	44	41.067	80.5
21 Aug	2100	46.4	43.307	78.5
21 Aug	2200	43.6	40.693	78
24 Aug	1945	35	32.667	73.5
24 Aug	2015	35	32.667	70.5
24 Aug	2100	32.6	30.427	68
24 Aug	2200	28.9	26.973	66
24 Aug	2230	27.7	25.853	65
25 Aug	0030	25.5	23.8	61.5
25 Aug	0330	20.375	19.017	57
25 Aug	0500	12.5	11.667	55
25 Aug	2000	37	34.533	76.25
25 Aug	2030	37.5	35.0	74
25 Aug	2100	36.5	34.067	74
25 Aug	2200	36.2	33.787	72.5
26 Aug	0530	33	30.8	66
26 Aug	2030	43	40.133	77.5
26 Aug	2200	46	42.933	78.5
27 Aug	2000	29	27.067	68.5
27 Aug	2030	31.7	29.587	68.5
27 Aug	2100	31	28.933	68
27 Aug	2200	28.75	26.833	66
28 Aug	0240	23.5	21.933	59
28 Aug	2010	32.4	30.24	70
28 Aug	2050	31	28.933	69
28 Aug	2200	29.5	27.533	67
29 Aug	0240	22.5	21.0	61.25
29 Aug	0440	20.6	19.227	58.5
29 Aug	2000	35	32.667	72
29 Aug	2050	33.1	30.893	71
29 Aug	2200	31.5	29.4	69
29 Aug	2330	28.8	26.88	66.5
30 Aug	0330	21.3	19.88	60
30 Aug	2000	37.8	35.28	75

Date	Time	Chirps/15s	Chirps/14s	Temp (ºF)
30 Aug	2055	37	34.533	73.25
30 Aug	2200	37.1	34.627	72.5
1 Sep	2200	36.2	33.787	70
2 Sep	0330	31.4	29.307	67.5
2 Sep	0600	30.2	28.187	66
4 Sep	0240	31.3	29.213	69
4 Sep	0505	26.1	24.36	63
5 Sep	0500	25.2	23.52	63
6 Sep	0600	23.66	22.083	61
7 Sep	0215	22.25	20.767	62
7 Sep	0525	17.5	16.333	56.5
9 Sep	2010	15.5	14.467	55
9 Sep	2110	14.75	13.767	52
10 Sep	2115	15	14.0	53
10 Sep	2210	14	13.067	50
11 Sep	0315	18.5	17.267	52
16 Sep	2100	27.7	25.853	65
17 Sep	2200	26	24.267	63
18 Sep	0130	21.7	20.253	59
19 Sep	0415	12.5	11.667	50.75
19 Sep	0435	12.5	11.667	49.25

Appendix B: Mathematical symbols and usage

Most of the mathematical symbols (beyond $+$, $-$, $\cdot$, $/$, $=$, etc.) used in the body of this document are defined informally below—along with some others that are closely related to them.

Concept	Symbol	Definition	Examples
Floor	$\lfloor\ldots\rfloor$	Rounding down (towards $-\infty$) of a non-integral real number, to the next integer value.	$\lfloor 1.75\rfloor = 1$ $\lfloor -1.75\rfloor = -2$ $\lfloor 1\rfloor = 1$
Ceiling	$\lceil\ldots\rceil$	Rounding up (towards ∞) of a non-integral real number, to the next integer value.	$\lceil 1.75\rceil = 2$ $\lceil -1.75\rceil = -1$ $\lceil 1\rceil = 1$
Exponent	b^n (superscript)	Number of times (not necessary integral) a base b is multiplied by itself in a product.	$x^2 = x\cdot x$ $3^4 = 3\cdot 3\cdot 3\cdot 3 = 81$
Enumeration	s_i (subscript)	Numbered terms of an ordered sequence.	$\boldsymbol{S} = \{s_1, s_2, s_3, \ldots\}$ $\boldsymbol{F} = \{1, 1, 2, 3, 5, \ldots\}$ ($\boldsymbol{F}$ is Fibonacci sequence.)
Sum	$\sum$	Sum of terms in a sequence. $\sum_{i=m}^{n} s_i = s_m + s_{m+1} + \ldots + s_n$ (If the bounds m and n are well understood, they are often omitted from the $\sum$ operator notation.)	$\sum_{i=1}^{4} f_i = f_1 + f_2 + f_3 + f_4$ $= 1 + 1 + 2 + 3$ (Sum of 1^{st} 4 terms of Fibonacci sequence.)
Product	$\prod$	Product of terms in a sequence. $\prod_{i=m}^{n} s_i = s_m \cdot s_{m+1} \cdot \ldots \cdot s_n$	$\prod_{i=3}^{5} \frac{i}{i+1} = \frac{3}{4}\cdot\frac{4}{5}\cdot\frac{5}{6}$ $= \frac{1}{2}$
Factorial	$n!$	$n! = \prod_{i=1}^{n} i$ $= 1\cdot 2\cdot\ldots\cdot n$ $0! = 1$	$5! = 1\cdot 2\cdot 3\cdot 4\cdot 5$ $= 120$
Euler's number	e	Base of natural logarithms. $e = \sum_{i=0}^{\infty} \frac{1}{n!}$ $= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots$ $\approx 2.71828\ldots$	

Appendix C: Proof that the Sum of Residuals is Zero

For any straight line that passes through the point $(\bar{x}, \bar{y})$, the sum of errors (residuals) is zero. We can prove this using $\hat{y} = \bar{y} + m(x - \bar{x})$ as a general equation for any such line.

Given

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\hat{y} = \bar{y} + m(x - \bar{x})$$

Then

$$\begin{aligned}
\text{SE} &= \text{ sum of errors (residuals)} \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} \left[y_i - \bar{y} - m(x_i - \bar{x}) \right] \\
&= \sum_{i=1}^{n} (y_i - \bar{y}) - m \sum_{i=1}^{n} (x_i - \bar{x}) \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} - m \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \right) \\
&= n\bar{y} - n\bar{y} - m \left(n\bar{x} - n\bar{x} \right) \\
&= 0
\end{aligned}$$

Appendix D: Minimizing the sum of squared errors (SSE)

(Note: The analysis below uses basic calculus.)

We know that if a differentiable function of a single variable (where the domain of that variable is unbounded) has a finite minimum or maximum value, it occurs at a *stationary point*—i.e. a point where the first derivative of the function is equal to zero. For example, the minimum y value of an upward-opening parabola occurs at its vertex, where the slope is zero.

The same rule applies to differentiable functions of N variables, as well: if a finite minimum or maximum value of the function exists, it must be located at a point where all N of the *partial derivatives* of the function are equal to zero.[8] Therefore, to find a and b that give the minimum value of SSE, we need to find the combination where the partial derivatives of SSE with respect to a and b are equal to zero.[9] The first step is thus to compute these partial derivatives.

$$\begin{aligned}
\frac{\partial(\text{SSE})}{\partial a} &= \sum_{i=1}^{n}\left[-2\left(y_i - a - b\,x_i\right)\right] \\
&= -2\left(\sum_{i=1}^{n} y_i - a\,n - b\sum_{i=1}^{n} x_i\right) \\
\frac{\partial(\text{SSE})}{\partial b} &= \sum_{i=1}^{n}\left[-2\,x_i\left(y_i - a - b\,x_i\right)\right] \\
&= -2\left(\sum_{i=1}^{n} x_i\,y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2\right)
\end{aligned} \qquad (15)$$

By setting the value of both partial derivatives in (15) to 0, we get 2 equations in 2 unknowns. (The x and y values aren't unknowns in this case; they're assumed to come from actual data.) Since all of our summations are over the same range, we'll leave out the limits and indices from here on.

$$\begin{aligned}
0 &= -2\left(\sum y - a\,n - b\sum x\right) \\
0 &= -2\left(\sum x\,y - a\sum x - b\sum x^2\right)
\end{aligned}$$

$$\begin{aligned}
a\,n \quad + b\sum x &= \sum y \\
a\sum x + b\sum x^2 &= \sum x\,y
\end{aligned} \qquad (16)$$

8 In a function of two or more variables, the partial derivative of the function with respect to one of the variables is obtained by taking the derivative with respect to that variable, while treating all other variables as constants.

9 Just as we can use the second derivative of a function of a single variable to determine whether a stationary point is a minimum, maximum, or inflection point, we can use the matrix of second partial derivatives (called the *Hessian* matrix) to do this with a function of multiple variables. Though it's outside the scope of this document, it can be shown from its Hessian that the sole stationary point of SSE is in fact a minimum, for any data set with at least two distinct values of the independent variable.

We can now solve the equations in (16) simultaneously to find *a* and *b*.

$$\begin{aligned}
a\,n\sum x^2 + \quad & b\sum x\sum x^2 = \sum x^2\sum y \\
a\left(\sum x\right)^2 + \quad & b\sum x\sum x^2 = \sum x\sum x\,y \\
\hline
a\left[n\sum x^2 - \left(\sum x\right)^2\right] \quad & \qquad = \sum x^2 y - \sum x\sum x\,y
\end{aligned}$$

$$\begin{aligned}
a\,n\sum x + b\left(\sum x\right)^2 \quad & = \sum x\sum y \\
a\,n\sum x + b\,n\sum x^2 \quad & = n\sum x\,y \\
\hline
b\left[n\sum x^2 - \left(\sum x\right)^2\right] & = n\sum x\,y - \sum x\sum y
\end{aligned}$$

$$a = \frac{\sum x^2\sum y - \sum x\sum x\,y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$b = \frac{n\sum x\,y - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad (17)$$

Acknowledgements

Original text for “Things Change” is from course materials for “Introductory Mathematics at College Park” (Math 110), developed by the Maryland Collaborative for Teacher Preparation, funded by the National Science Foundation, Division of Undergraduate Education, Collaboratives for Excellence in Teacher Preparation (NSF Cooperative Agreement No. DUE 9814650) [1]. Used with permission.

These materials were written with the invaluable editorial and creative assistance of L.D. Landis, Drew Einhorn, Joel Castellanos, Elizabeth Kallman, and David Janecky.