Linear Regression

We have demonstrated in this section that constant, linear, and piecewise combinations of these two function types can be used to model a variety of phenomena inspired by real-world situations. What happens, as if often the case in real-world situations, when we are given data sets that are not precisely linear, but still have a definite linear trend? An example of this is Skippy's time and temperature data from Example **??** in Section **??**.

In that example, t represented the time (number of hours after 6 a.m.) and T represented the outdoor temperature in degrees Fahrenheit. The data Skippy collected along with a plot of the function $T = f(t)$ are given below. Even though the data points as t varies from $t = 0$ to $t = 8$ do not all lie on the same line - a fact we could prove analytically by checking slopes - there does appear to be a linear **trend** evident. The same can be said for the data as t varies from $t = 8$ to $t = 12$. As we'll see, there are statistical methods which can produce linear functions that are in some sense 'closest' to all of the data, and they are represented below by the dashed lines below on the right.

t: hours after 6 a.m.	T: temperature $^\circ$F
0	64
2	67
4	75
6	80
8	83
10	83
12	82

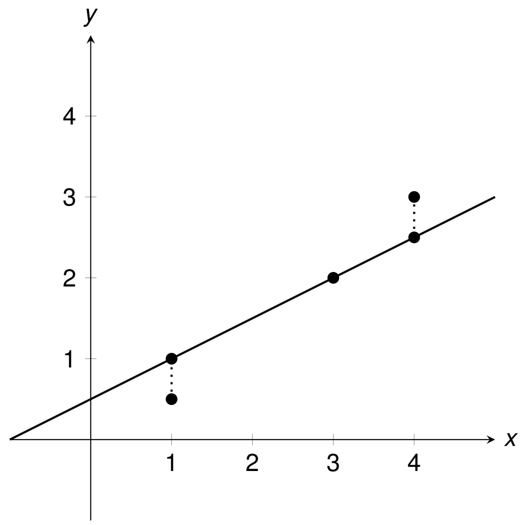
Desmos link: `https://www.desmos.com/calculator/aqedfuqmuw`

How do we measure how 'close' a set of points is to a given line? Let's leave Skippy's data for the moment and focus on a smaller data set. Suppose we collected three data points: $\{(1, 0.5), (3, 2), (4, 3)\}$. At the top of the next page (on the left) we plot these points along with the line $y = 0.5x + 0.5$. The way we measure how close the line is to these points is by computing the **total squared (vertical) error** between the data points and the line as follows. For each of our data points, we find the vertical distance between the point and the line. To accomplish this, we need to find a point on the line directly above or below each data point. In other words, we need a point on the line with the same x-coordinate as our data point.

For example, to find the point on the line directly above $(1, 0.5)$, we plug $x = 1$ into $y = 0.5x + 0.5$ and we get the point $(1, 1)$. Similarly, we find $(3, 2)$ is on the line already and $(4, 2.5)$ is the point on the line directly beneath $(4, 3)$. We find the total squared error E by taking the sum of the squares of the differences of the y-coordinates of each data point and its corresponding point on the line. For the data and line in this discussion $E = (0.5 - 1)^2 + (2 - 2)^2 + (3 - 2.5)^2 = 0.5$.

Using advanced mathematical machinery,[1] it is possible to find the line which results in the lowest value of E. This line is called the **least squares regression line**, or sometimes the 'line of best fit'. The formula for the line of best fit requires notation we won't present until Chapter **??**, so we will revisit it then. Most graphing utilities have a built-in regression feature, so at this point we turn the computations over to the technology. A screenshot from desmos is given on the right at the top of the next page.

[1] like Calculus or Linear Algebra ...



Desmos link: https://www.desmos.com/calculator/ofrcaiemar

Our graphing utility produces the model[2] $y = mx + b$ where the slope is $m \approx 0.821$ and the y-coordinate of the y-intercept is $b \approx -0.357$. The value r is the **correlation coefficient** and is a measure of how close the data is to being on the same line. The closer $|r|$ is to 1, the better the linear fit.[3] Having $r \approx 0.997$ tells us that the points have a strong, positive correlation - that is, they are very close to being on a line with a positive slope, namely $y = 0.821x - 0.357$. Indeed, the total squared error between our data set and this line is $E \approx 0.018$. The mathematics tells us that this is the smallest we can get E by modifying the parameters m and b, even though none of the data points actually lie on the line.

Now that we have this new mathematical machinery, let's revisit Skippy's time and temperature data.

Example 1.

1. Use a graphing utility to find best fit linear models for each of the data sets below. Comment on the fit and interpret the slope of each.

t: hours after 6 a.m.	T: temperature °F
0	64
2	67
4	75
6	80
8	83

t: hours after 6 a.m.	T: temperature °F
8	83
10	83
12	82

2. Use your models to predict the temperature at 7 a.m. and 3 p.m., rounded to one decimal place.

Explanation. 1. For our first set of data, we get the line $T = F(t) = 2.55t + 63.6$. The value $r = 0.987$ tells us that it is a fairly good fit and we see this graphically, too.[4] Thus we can be confident in using this model to predict the temperature during between the hours of 6 a.m. and 2 p.m. with reasonable accuracy.

To interpret the slope, we recognize t as the independent variable (input) and T as the dependent variable (output), so the slope $m = \dfrac{\Delta T}{\Delta t}$ is the rate of change of temperature with respect to time. In this

[2] We chose to use three decimal places for the approximations in this demonstration. How many you get to use in reality varies from one application to another.

[3] The value r^2 is called the **coefficient of determination** and is also a measure of the goodness of fit. We refer the interested reader to a course in Statistics to explore the significance of r and r^2.

[4] We use F as the name of the function here to distinguish it from f - the function determined solely by the given set of data.

case, $m = 2.55$ means that the temperature is increasing (getting warmer) at a rate of 2.55°F per hour. A screenshot from Desmos is given below.

Desmos link: `https://www.desmos.com/calculator/jhkc4szlj7`

For the second set of data, we get $T = G(t) = -0.25t + 85.167$ and we have $r = -0.866$. Here, the negative sign on r indicates a negative correlation which means our line has a negative slope.[5] While the fit looks OK, it certainly isn't as strong as with the first data set, so using this model to predict the temperature between 2 p.m. and 6 p.m. (let alone beyond) is a bit risky.

The slope in this case is $m = -0.25$ which corresponds to the temperature decreasing (getting cooler) at a rate of 0.25°F per hour. That's what a negative correlation means - an increase in input (more time passes) yields a decrease in output (cooler temperatures). As screenshot from Desmos is given at the top of the next page.

Desmos link: `https://www.desmos.com/calculator/rqhhlc0enf`

2. The time 7 a.m. corresponds to $t = 1$. This falls between $t = 0$ and $t = 8$ so we use our first model. Substituting $t = 1$ gives $T = F(1) = 2.55(1)+63.6 \approx 66.2$. Therefore, the model predicts the temperature to be 66.2°F at 7 a.m.. Likewise, 3 p.m. corresponds to $t = 9$. This is greater than 8, so we use the second model: $T = G(9) = -0.25(9) + 85.167 \approx 82.9$. The model predicts the temperature at 3 p.m. to be 82.9°F. Based on the goodness of fit of each model, we have more confidence in the former prediction than in the latter.

Examples **??**, **??** and **??** (among others) represent three different levels of mathematical modeling. In Example **??**, the mathematical model (the cost function) was provided and our task was to use the model to **interpret** the mathematics in that context. In Example **??**, we were given a minimal amount of information, namely, two data points, and then asked to **construct** a model which fit those data exactly. Lastly, in Example **??**, we were given several data points and we used statistical methods to construct a **best fit** model to the data.

The validity of the models rests on the validity of the underlying assumptions used to create the models. For instance, is there any reason to assume a price-demand function would be linear? Is it reasonable to assume that the temperature changes at a constant rate? These are questions for economists and scientists. Mathematicians often take on a role of equal parts translator and prophet: they codify ideas into formulas and then use them to make predictions about yet-to-be observed phenomena.

[5] We use G as the function name here to distinguish it from the given function f and the regression for the first data set, F.