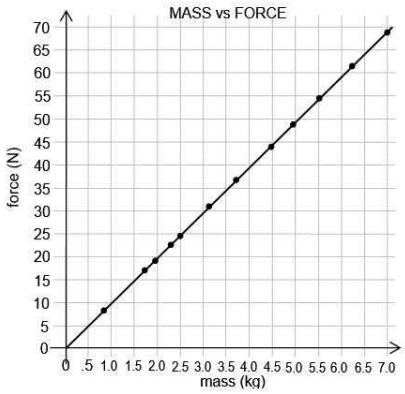
Correlation & Regression Lines

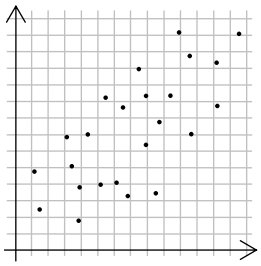
If you've taken a science class and had to do a lab report, then you're familiar with the idea of graphing a batch of data to make a scatterplot. Often, you would be asked to draw a line of best fit through the data. In a high school science class, you had to guess where this line was, but different students with the same data might put the line in different places. In statistics, we have a method for finding the equation of the best line.

Consider a physics experiment where we put various objects on a scale to determine their mass in kilograms, and then we hang them from a force meter to determine the amount of gravitational force acting on them. According to physics, you should be able to calculate the force exactly if you knew the mass: you'd multiply the mass by 9.8, which is the constant acceleration due to gravity.

If a student were to do this experiment perfectly, and create a scatterplot of her data, plotting mass vs. force, it should look like the graph to the right: all the points lie exactly on a perfectly straight line, as they should for force and mass. The relationship is 100% predictable.



If, on the other hand, the student made any tiny mistakes during the experiment or something went wrong — maybe the spring in the force meter was rusty or old and didn't work well, maybe the student graphed a data point slightly wrong, or maybe she didn't zero the scale before using it — then the points are going to vary slightly around that same line. The data will still suggest a line, but you can't put a ruler on the paper and draw a line through all of them.

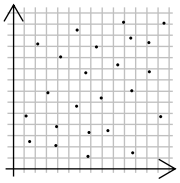


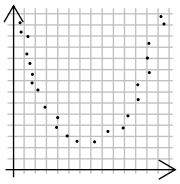
In real life, issues in business, health and public opinion aren't as predictable as a physics experiment. You might get a scatterplot that looks like the one on the left. Does this data still suggest a line? How do we differentiate between real associations in data vs. randomness? Can we make any predictions based on this data?

When we measure data two ways, such as taking a sample of children and finding both their age and height, one measurement goes on the horizontal axis and one goes on the vertical axis. The measurement on the horizontal axis is the **independent variable** or the **explanatory variable**. It should be one that we cannot control for. The passage of time is often the explanatory variable. The remaining measurement is the **dependent variable** or the **response variable**. Its value is dependent on the other measurement.

Authored by Darren Rigby

We can analyse data to determine its degree of **correlation**: how linear a set of data is, and how easy or useful it is to make predictions about the population by using a line of best fit as a model. Correlation has the variable **r** and can range from −1 to +1, inclusive. A correlation of +1 shows a very strong **positive relationship** or **direct relationship** between the variables: as one variable increases, so does the other variable. A correlation of −1 shows a very strong **negative relationship** or **inverse relationship** between the variables: as one variable increases, the other one decreases. A correlation of 0 means that a linear model doesn't fit the data well. *This does not mean that there is no relationship between the variables!* Both of the graphs at the right would have correlations close to 0, but there is clearly a relationship between the variables in the second graph. Correlation only looks for linear relationships.





To calculate r, you need to know the mean and standard deviation for both variables ($\bar{x}$, $\bar{y}$, s_x, s_y). Then:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{1}{(n-1)s_x s_y}\sum(x_i-\bar{x})(y_i-\bar{y})$$

Put into words, for each data point subtract the mean for x and y from the x-value and y-value for the point. Multiply them and divide by the product of the two standard deviations. Add up the results for each data point. Divide the total by one less than the number of data points you have. This is, frankly, a long and tedious calculation for any data set of more than about 5 points. (It's slightly easier if you use the second, simplified form of the formula.) If you have Excel or a Texas Instruments graphing calculator (or any other calculator that does statistical calculations), we advise you to let the computer do the arithmetic for you. (And after a week or two, you won't have to perform the calculation again for any of your assignments.)

In a Texas Instruments calculator, the setting that displays r is turned *off* as a default. To turn it on, start on a clear line in the calculation window and press [2nd][CATALOG 0]. This will open a list of all functions and commands programmed into the calculator. Scroll down to the line "DiagnosticOn". Press [Enter]. This will return you to the calculation window. Press [Enter] again to execute the command. You should only need to do this once (unless you reset the calculator).

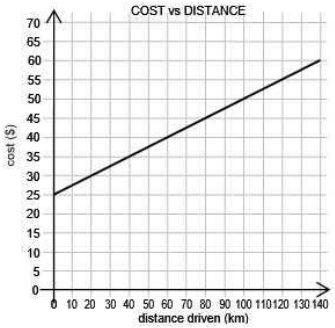
Once we know the correlation, we can find the formula for the line of best fit through a data set, which is known in statistics as a **regression line**. Specifically, we'll look at the **least-squares regression line**. This is one way to find a line of best fit, but it's not the only one (though it's the only one you'll encounter in this course). To define a line we need its **slope** and its y-**intercept**. We can calculate them from r, the means and the standard deviations:

$$\hat{y} = a + bx, \text{ where } b = r\cdot\frac{s_y}{s_x} \text{ and } a = \bar{y} - b\cdot\bar{x}$$

The slope is the number multiplied by x and the intercept is the other number.
It is worth noting that the TI calculators *reverse* the meanings of a and b from what it says in the Moore textbook. In the formula on the previous page, a is the intercept and b is the slope; in the calculator a is the slope and b is the intercept. To help reduce confusion, in this worksheet, we'll use the terms rather than variable names.

Statisticians use marks to distinguish between forms of a variable. If y is a measured, **observed** value for a statistical variable Y, then $\bar{y}$ ("y-bar") is the mean value of all Y, and $\hat{y}$ ("y-hat") is a **predicted** value for y for a given value of x.

To understand what the slope and intercept mean for a line, consider a pricing scheme for renting a car. A rental company will often charge you a flat fee for the rental plus another amount for every kilometre driven. A graph of distance driven vs. cost would be a straight line with a meaningful slope and intercept. The intercept of the line is the y-value that results when x = 0. The amount you would pay for the rental if you didn't drive the car at all would be just the rental fee, which is $25 according to this graph. From there, we can see how much the overall cost of the rental increases per kilometre driven. This is what slope measures: the rate of change in the response variable when we increase the explanatory variable. In the graph at right, every additional kilometre driven adds $0.25 to the cost. In a graph, a line that has a positive slope goes up to the right, and it occurs when r > 0; a line that has negative slope goes down to the right, and such a line occurs when r < 0.



We can use a regression line for **interpolation** of data, to predict a typical value for y given a value for x. Usually this is only done for values of x between the highest and lowest x-values encountered in a study. Beyond those values we cannot be sure how the graph behaves, since we're only making a prediction. For the car rental price, the equation is $\hat{y} = 25 + 0.25x$. To find the value for $\hat{y}$ when we've driven our rental car 200 km, we simply plug 200 in for x in the equation and solve:

$$\begin{aligned}\hat{y} &= 25 + 0.25(200)\\ &= 25 + 50\\ &= 75\end{aligned}$$

Example 1: What would $\hat{y}$ be for an x value of 36, if $\bar{x} = 29$, $s_x = 12.4$, $\bar{y} = 109.4$, $s_y = 3.7$ and $r = -0.68$?

Solution:

$$\text{slope} = r \cdot \frac{s_y}{s_x} = -0.68 \cdot \frac{3.7}{12.4} = -0.20290\ldots$$

$$\text{intercept} = \bar{y} - [\text{slope}] \cdot \bar{x} = 109.4 - (-0.20290)(29) = 115.28419\ldots$$

$\therefore \hat{y} \approx 115.28 - 0.203x$

$y(36) = 115.28 - 0.203(36)$

$= 107.972$

EXERCISES

A. 1. Explain why distance was on the horizontal axis in the car rental graph on page 3 of this worksheet.

2. In each pair of variables, determine which should be the explanatory variable (x) and which one should be the response variable (y).

a) plant growth, exposure to sunlight

b) amount of alcohol consumed, reaction time

c) amount of money invested, salary

d) body mass index, happiness calculated from various factors besides weight

B. A researcher wants to determine whether there is a linear relationship between boys' ages and their heights during their adolescent and pre-adolescent years. Below are the ages of 8 boys and their heights in cm. Also included are the means and standard deviations of each set of data points.

Subject	A	B	C	D	E	F	G	H	mean	s.d.
Age (x)	10	12	13	15	15	16	16	17	14.25	2.375
Height (y)	102	108	117	122	134	146	178	162	133.625	26.715

Calculate r.

C. Determine an equation for $\hat{y}$ from the data above and your answer to B.

D. 1. Use the equation from C to determine predicted values for the following ages to the nearest cm.

a) 11 years old

b) 15 years old

c) 65 years old

d) 2 years old

2. Explain why your answer to 1b) does not match either of the data points from the table in B.

3. Explain why the extrapolated values for $\hat{y}$ from 1c) and 1d) are not reasonable heights.

4. For what values of x does $\hat{y}$ make sense?

SOLUTIONS

A. (1) Distance is the explanatory variable; we would use the distance travelled to predict what price we will be charged when we return the vehicle.

(2)a) x: exposure, y: growth (b) x: alcohol, y: reaction (c) x: salary, y: investments

(d) both are possible, depending on which you see as cause and which as effect

B. r = 0.859 C. $\hat{y} = -4.059 + 9.662x$ D. (1)a) 102 (b) 141 (c) 624 (d) 15

(2) They don't match because 141 cm is a predicted value based on the data as a whole. The model says a 15-year-old should be close to 141 cm.

(3) The experiment examined heights of teenagers. Beyond the given points, people's growth rates do not stay the same.

(4) x = [10, 17]. To include 18 and 19, a 19-year-old subject should be added to the study.