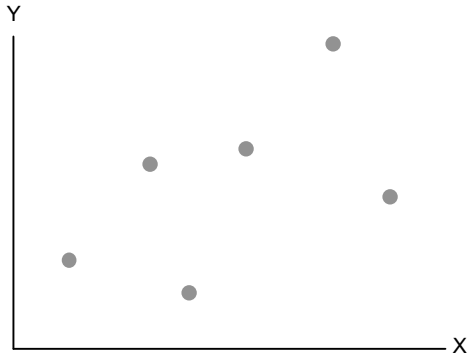
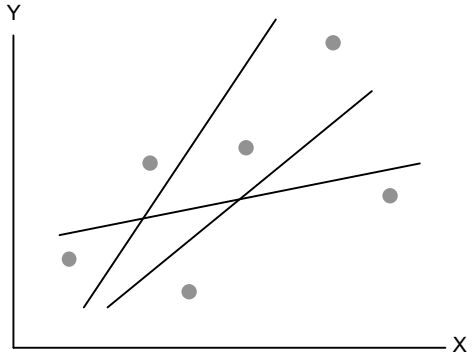
Why we use “least squares” regression instead of “least absolute deviations” regression.

We have a scatterplot showing the relationship between variables x and y.



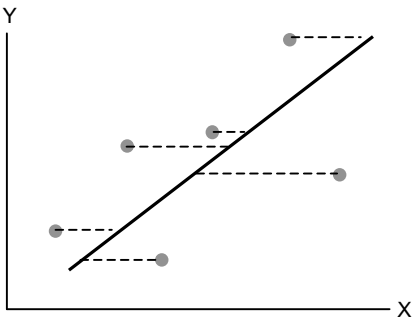
We would like to find a line that best describes the relationship between the variables. How do we determine which line is best?



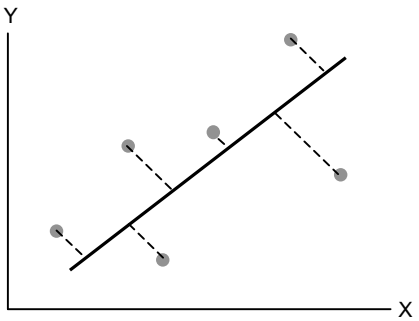
Which line best describes the relationship between x and y?

The best line will be one that is “closest” to the points on the scatterplot. In other words, the best line is one that minimizes the total distance between itself and all the observed data points.

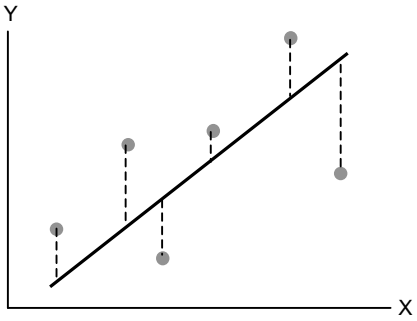
We can measure this distance in a variety of ways. We could measure distance from the points to the line horizontally, perpendicularly, or vertically. Since we oftentimes use regression to predict values of Y from observed values of X, we choose to measure the distance vertically. This distance represents the error we would have if we used our regression line to predict values of Y from observed values of X.



Horizontal Distance

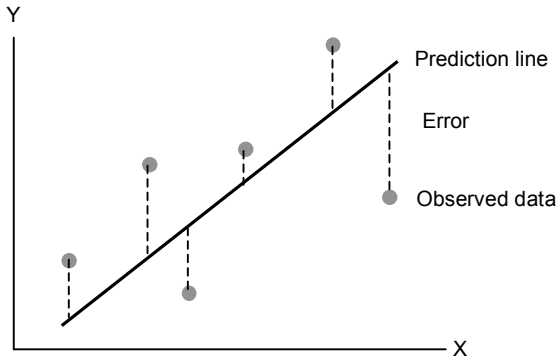


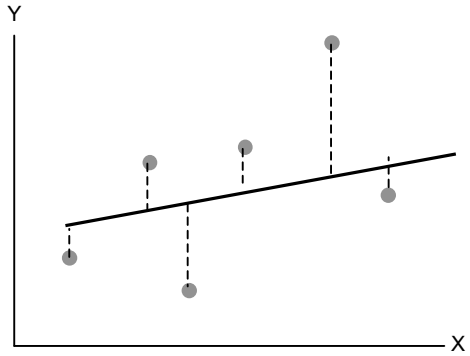
Perpendicular Distance



** Vertical Distance **

We want to find the line that minimizes the vertical distance between itself and the observed points on the scatterplot. So here we have 2 different lines that may describe the relationship between X and Y. To determine which one is best, we can find the vertical distances from each point to the line...





We want to minimize the total distance from the points to the line, so we want to minimize the sum of the lengths of all the dotted lines. From the two lines drawn above, the sum of the lengths of the dotted lines is:

So based on this, the line on the right is better than the line on the left in describing the relationship between X and Y. The line on the right has less total error. Unfortunately, this would be an extremely tedious method to use to find the *best* line. After all, we could draw an infinite number of lines on the scatterplot. We need to find another (faster) way to minimize the sum of the distances from the regression line to the observed data points.

Let's first label our scatterplot:

The observed data are points with coordinates (X_i, Y_i).

From algebra, we know the formula for a straight line is: y = mx + b. Just to make things more difficult for students, we typically rewrite this line as $\hat{Y} = b_0 + b_1 X_i$ (there really are other reasons for rewriting it this way, but it's not important).

We can then define the error to be the difference between the coordinates and the prediction line.

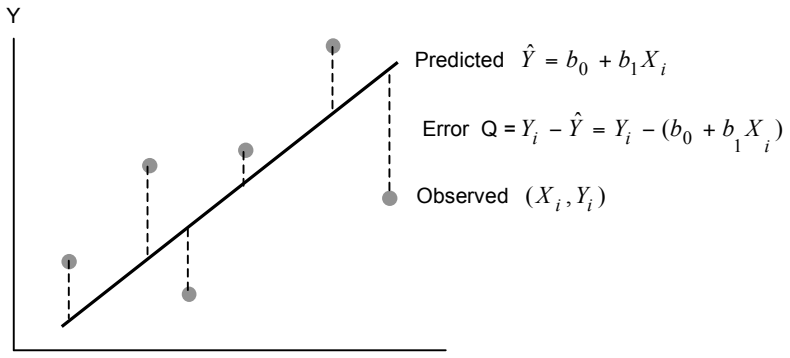
The coordinate of one point: (X_i, Y_i)

The prediction line: $Y = b_0 + b_1 X_i$

So error = distance from one point to the line = Coordinate – Prediction = $Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$

We can now take the sum of these distances for all the points for our n observations: $\sum_{i=1}^{n} Y_i - b_0 - b_1 X_i$

Therefore, the best-fitting line will be the one that minimizes: $\sum_{i=1}^{n} Y_i - b_0 - b_1 X_i$.



We still have one major problem. If we use the above formula (or if we look at the dotted error lines above) we notice that some of the errors will be positive and some will be negative. The problem is that when we add positive and negative values, they tend to cancel each other out.

When we found the total length of the lines on the previous page, we made the assumption that all the errors were positive. Whether the observed data point was above or below the line did not matter – we always assumed that the error (vertical distance) was positive.

We have not yet made this assumption in our formula. In order to make all the errors positive, we could decide to use absolute values. Absolute values would force all the errors to become positive (this is because absolute values typically represent distances). When all the errors are positive, we know we can add them without worrying about some errors canceling out other errors.

So, using absolute values, our formula becomes: $\sum_{i=1}^{n} \left| Y_i - b_0 - b_1 X_i \right|$. We want to find the line that minimizes this formula.

Whenever we have a formula that we want to maximize or minimize, we can use Calculus to find that maximum or minimum. We use a process called *differentiation*.

So now we simply need to use some Calculus on our formula and we will be finished...

Not so fast!

Ask any Calculus professor (or student who has taken Calculus) how they minimize a formula using Calculus. They should quickly be able to tell you something along the lines of, "Oh, you just take the first derivative and set it equal to zero."

Now show them the formula we want to minimize. In fact, point out the absolute value bars on the formula and watch as the look of fear or dread overcomes them. That's because **absolute values are difficult to work with in mathematics (especially Calculus).**

In fact, for what we want to do, we cannot use our formula with absolute values. In mathematical terms, we would say this is because the use of absolute values results in discontinuous derivatives that cannot be treated analytically.

In plain English, we can describe the problem this way. We already know that we could draw a bunch of dotted lines on a scatterplot and measure the total distance of all those lines. We could then try different regression lines and repeat the process. We would prefer to find a **quick and easy** way to find the line that minimizes the sum of the errors. Using absolute values does not make the process any easier for us. In fact, using absolute values requires us to basically go through the entire process of drawing regression lines, measuring the error distances, and summing those errors. We could use a computer to do this (using something called Wesolowsky's direct descent method), but we still wouldn't understand what's going on.

Let me stress that we would really like to use those absolute value bars in our formula. Absolute values are the easiest way to make all values positive and absolute values make sense. **We simply choose not to use absolute values because of the difficulties we have in working with them mathematically**.

So we need to use a different method to make all the errors positive. To see what method we'll use, let's examine the concept of standard deviation that we learned earlier. Recall that the standard deviation tells us, on average, the distance from each observation to the overall mean. For example, if we had the following data set:

Observations: 10 10 10 30 30 30 (the mean is 20)

Since each score is 10 units away from the mean, the standard deviation (average distance from the mean) would equal 10. We also know we could use the formula for standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(X - \overline{X}\right)^2}{n}} = \sqrt{\frac{(10-20)^2 + (10-20)^2 + (10-20)^2 + (30-20)^2 + (30-20)^2 + (30-20)^2}{6}} = \sqrt{\frac{600}{6}} = 10$$

The values in the numerator represent the squared distances from each observation to the mean. Do you remember why we need to square everything in the numerator of the formula? Let's see what would happen if we forget to square everything...

$$\sqrt{\frac{(10-20)^2 + (10-20)^2 + (10-20)^2 + (30-20)^2 + (30-20)^2 + (30-20)^2}{6}} = \sqrt{\frac{-10-10-10+10+10+10}{6}} = 0$$

When we do not square all the values in the numerator, the positive and negative values cancel each other out and we're always left with a value of 0.

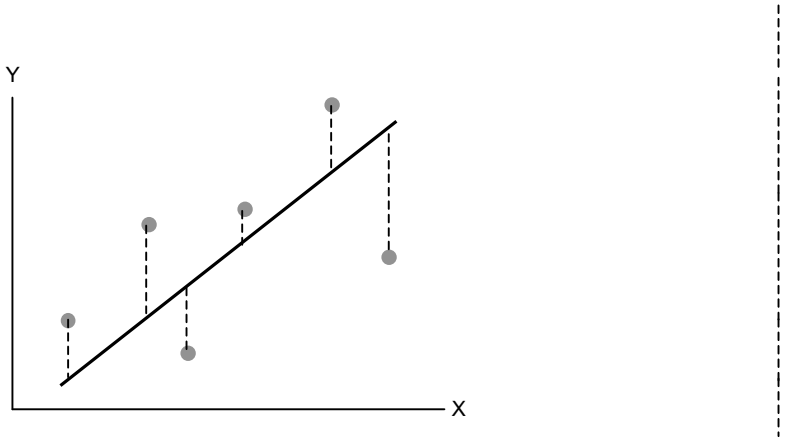
We could have solved this problem by using absolute values in the numerator (this is called the mean absolute deviation), but we instead chose to square all the values (and then take a square root at the end).

We will use this same process with our regression problem. Instead of using absolute values, let's square all the values. If we need to, we can always take a square root at the end.

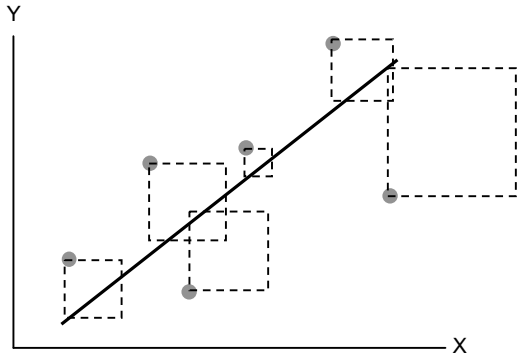
When we use this method, the formula that we want to minimize changes from $\sum_{i=1}^{n}\left|Y_i - b_0 - b_1 X_i\right|$ to $\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X_i\right)^2$.

We can now use (relatively) straightforward Calculus methods to find the regression line that will minimize this formula.

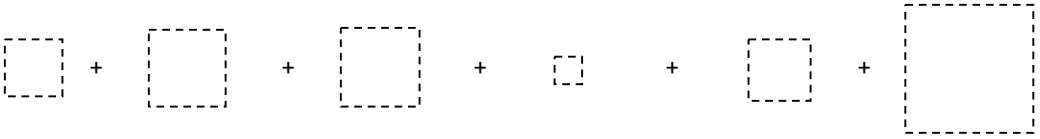
Before we do that, we better understand what's happening when we square all the values. Recall that when we used absolute values, we were simply finding the total distance of all our error lines (the vertical dotted lines).



When we square those error lines, we are literally making squares from those lines. We can visualize this as...



So we want to find the regression line that minimizes the sum of the areas of these error squares. For this regression line, the sum of the areas of the squares would look like this...



If we could find a line that better fits the data, the total area of all the squares would be smaller.

So how did this make anything easier?

Honestly, if you're not interested in the mathematics of regression, this didn't make anything easier for you. It's much more intuitive to use absolute values than squares. But remember that if you use absolute values, you will be forced to draw many different regression lines and measure the error distances from each observation to each regression line. Even if you do this, you'll never be completely certain that you've found the best-fitting regression line.

If you want to use simple formulas to find the best-fitting regression line, we'll have to use the formula with squared values. The last two pages of this handout detail the Calculus we can use to minimize the sum of squared deviations (to find the best-fitting regression line).

If you don't want to navigate through the Calculus, the important results are:

1. We define the best-fitting regression line (the line that is closest to the data) as the line that **minimizes the sum of squared errors**. We call this the **ordinary least squares regression line**.

2. We know the formula for a line is $Y = b_0 + b_1 X_i$. Using Calculus, we can prove that the **ordinary least squares regression line** is:

$$Y = b_0 + b_1 X_i \quad \text{where} \quad b_0 = \overline{Y} - b_1 \overline{X} \quad \text{and} \quad b_1 = r \frac{S_y}{S_x}$$

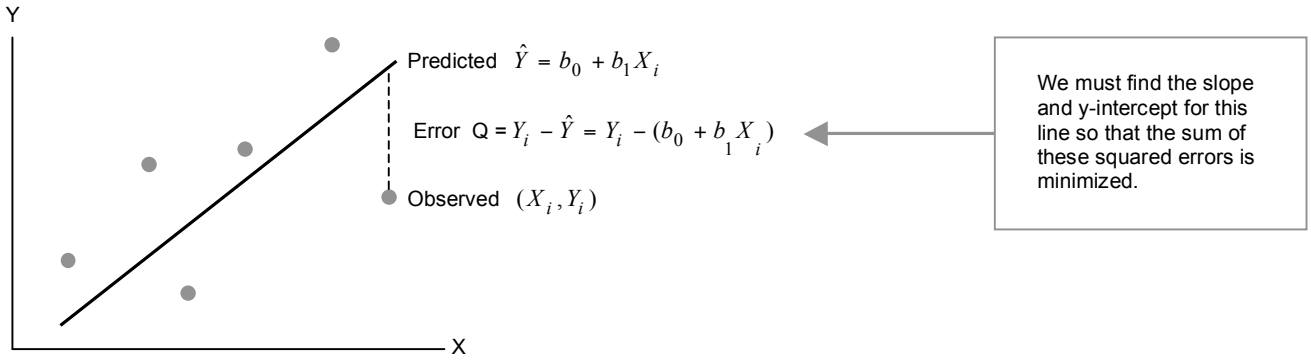
and where: r is the correlation coefficient between X and Y
S_y is the standard deviation of Y
S_x is the standard deviation of X

This **ordinary least squares regression line** is not necessarily the best method to use. In fact, using absolute values in our formula would yield a regression line that is more robust than what we get from our least squares method. The advantages of our least squares method are:

1. It is easy to find the best-fitting regression line (using the above formulas)
2. We are sure to find only **one** best-fitting line. (If we used absolute values, we would find more than one line is "best")

We simply choose to use the least squares method because it is much easier to work with mathematically.

Derivation of the Parameters of the Least Squares Regression Line



Let Q represent the sum of squared errors: $Q = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$

We need to find values for b_0 and b_1 that will minimize Q. We know that to minimize a function, we must set its first derivative equal to zero and solve. Because we have two variables in this function, we'll need to take partial derivatives of Q with respect to b_0 and b_1.

Partial derivative of Q with respect to b_0: (we treat b_0 as a variable and all other terms as constants)

(Chain Rule)

$$\frac{\partial Q}{\partial b_0} = \frac{\partial \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2}{\partial b_0} = 2\sum (Y_i - b_0 - b_1 X_i) \frac{\partial (Y_i - b_0 - b_1 X_i)}{\partial b_0} = -2\sum (Y_i - b_0 - b_1 X_i)$$

We set this partial derivative equal to zero: $-2\sum (Y_i - b_0 - b_1 X_i) = 0$ $\quad \sum (Y_i - b_0 - b_1 X_i) = 0$

$$\sum Y_i = n b_0 + b_1 \sum X_i$$

Partial derivative of Q with respect to b_1: (we treat b_1 as a variable and all other terms as constants)

(Chain Rule)

$$\frac{\partial Q}{\partial b_1} = \frac{\partial \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2}{\partial b_1} = 2\sum (Y_i - b_0 - b_1 X_i) \frac{\partial (Y_i - b_0 - b_1 X_i)}{\partial b_1} = -2\sum X_i (Y_i - b_0 - b_1 X_i)$$

Set the partial derivative equal to zero: $-2\sum X_i (Y_i - b_0 - b_1 X_i) = 0$ $\quad \sum X_i (Y_i - b_0 - b_1 X_i) = 0$

$$\sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

Now we must solve this system of two *normal* equations…

System of normal equations:

$$\sum Y_i = nb_0 + b_1 \sum X_i$$
$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

This system can be solved to get: $b_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$

and

$$b_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \overline{Y} - b_1 \overline{X}$$

We can rewrite b_1 given the following information: $S_{xy} = \sum (x_i - \overline{X})(y_i - \overline{Y}) = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i$

$$S_{xx} = \sum (x_i - \overline{X})^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$$

$$S_{yy} = \sum (y_i - \overline{Y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$

Therefore, $b_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \dfrac{S_{xy}}{S_{xx}} = r\dfrac{S_y}{S_x}$

So, the line that minimizes the sum of squared errors has the following slope and y-intercept parameters:

$$b_0 = \overline{Y} - b_1 \overline{X} \quad \text{and} \quad b_1 = r\frac{S_y}{S_x}$$