Regression I - the least squares line

The difference between correlation and regression.

> Correlation describes the relationship between two variables, where neither variable is "independent" or used to "predict".
>
> In correlation it would be irrelevant if we changed the axes on our graph.
>
> This is NOT true for regression.

In regression, we often have a variable that is *independent* and another that *depends* on this variable.

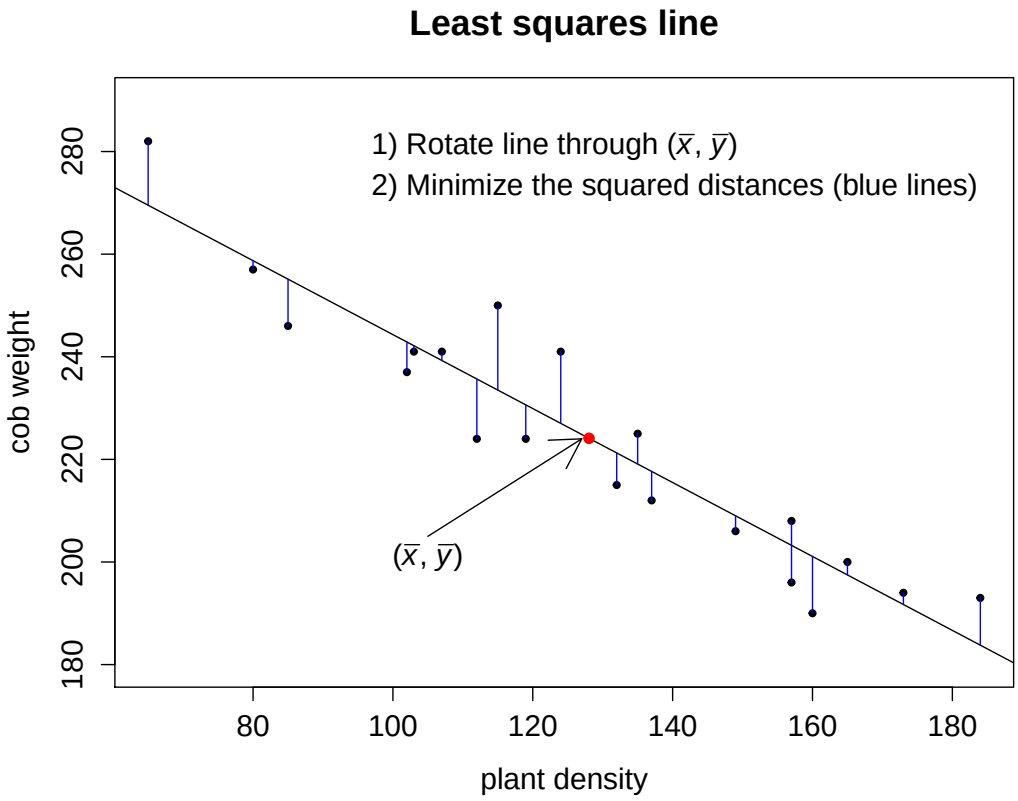
> For example,temperature (independent) and activity level (dependent) in cold blooded animals.
>
> > Activity level obviously depends on temperature (and NOT vice-versa).
>
> Often we also use this "independent" variable to *predict* the "dependent" variable.
>
> We're also interested in testing for significance, though in this case we look to see if a line going through the data points is significant.

The basic idea in regression is to model the relationship between our variables.

> Let's assume there is a relationship between our variables. This relationship is then modeled using an equation.
>
> > In the simplest case (as in our class), this is the equation for a line.
>
> This raises the question: is the line significant? This is one place where statistics come in.
>
> > There are actually many different ways of estimating the line, but we'll only learn the most common method.

Fitting the line.

> We'll use what is called the *least squares line*.



We draw a line through the point $(\bar{x}, \bar{y})$, and then rotate it until the vertical distances (which we'll actually call *residuals* or *errors* instead of distances) are at a minimum:

$$\sum_{i=1}^{n} d_i = minimum$$

Except that (for mathematical reasons we don't want to go into) we actually use the sum of the *squared* distances:

$$\sum_{i=1}^{n} d_i^2 = minimum$$

> To do this, we use calculus (e.g., differentiate, set to 0, make sure we have a minimum).

Fortunately, we don't actually need to do the calculus (it's been done for us). All we really need to know are the answers, which consist of the slope and intercept for our least squares line:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Where b_1 is the slope and b_0 the intercept. Notice that we could rewrite the equation for the slope in terms of the Sum of Cross Products (SS_{cp}) and the Sum of Squares for x (SS_x).

$$b_1 = \frac{SS_{cp}}{SS_x}$$

So we finally get the equation for our least squares line:

$$\hat{Y} = b_0 + b_1 X$$

Where $\hat{Y}$ is the predicted value of Y for a particular X.

A comment on a different approach (that is actually the same).

Recently, for some reason, many textbooks have presented a different approach to calculating b_1:

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

Where r is the correlation coefficient. The two approaches do give the same answer for b_1:

$$b_1 = r\left(\frac{s_y}{s_x}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \times \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We will stick with the original approach. In those instances where you don't have to calculate everything yourself, you'll be given SS_{cp}, SS_x and SS_y to help you (instead of r, s_x, and s_y).

Let's summarize. We calculate b_0 and b_1 and then arrange everything into the equation for a line.

Notice also that (as usual) we calculated the *estimates*:

$$b_1 \textit{ estimates } \beta_1$$

$$b_0 \textit{ estimates } \beta_0$$

As usual, β_1 and β_0 are unknown, and we estimate them with b_1 and b_0.

So we know how to calculate the equation of a line. There is, however, a similar equation that is also used a lot in regression:

$$y_i = b_0 + b_1 x_i + e_i$$

What's the difference? Notice that in this case the equation does not give us $\hat{Y}$. Instead, it gives us the actual values for Y (the actual y_i's). It does this because the last term (e_i) is an *error* term. It adds in the *difference* between the actual y_i and the estimated value (i.e., $\hat{y}_i$).

It is used quite a bit; often we want to calculate the value of the e_i's, and we can use this equation to do that.

Notice (as mentioned) that we use the term *error* here. The e_i's are also known as the *residuals* ($e_i = r_i$), which abbreviation we use depends a little on what we're doing, but for use they're mostly interchangeable.

Let's illustrate things before we get lost with an example based on Wikipedia that discusses soil respiration. Soil respiration is the amount of CO_2 that is released from the soil by microorganisms and other soil inhabitants. We want to know what the effect of temperature is on soil respiration. CO_2 levels are measured as CO_2 exchange rate (NCER), measured in µmol/m^2.

First, you should figure out what you expect to happen - if temperature increases, what should happen to the CO_2 levels being released from the soil (more on formulating hypotheses in the next set of notes)?

We have the following results, loosely based on a real study:

Temperature	17	18	19	20	21	22	23	24
NCER	0.31	0.27	0.45	0.43	0.51	0.55	0.65	0.63

We know how to calculate the following:

$$SS_x = 42 \quad SS_y = 0.1334 \quad SS_{cp} = 2.26 \quad \bar{x} = 20.5 \quad \bar{y} = 0.475$$

But to make sure, let's do the calculation for SS_{cp} (our **S**um of **C**ross **P**roducts) one more time:

$$\begin{array}{lll}
\text{for:} & i=1:(17-20.5)(.02-.31) & =0.5775 \\
& i=2:(18-20.5)(.25-.27) & =0.5125 \\
& i=3:(19-20.5)(.54-.45) & =0.0375 \\
& i=4:(20-20.5)(.69-.43) & =0.0225 \\
& i=5:(21-20.5)(1.07-.51) & =0.0175 \\
& i=6:(22-20.5)(1.50-.55) & =0.1125 \\
& i=7:(23-20.7)(1.74-.65) & =0.4375 \\
& i=8:(24-20.5)(1.74-.63) & =\underline{0.5425} \\
& \textbf{Sum} & =\mathbf{2.26}
\end{array}$$

So now that we know how to get SS_{cp} we can proceed to calculate our slope:

$$b_1 = \frac{2.26}{42} = 0.05381$$

And now we use this to calculate our intercept:

$$b_0 = 0.475 - 0.05381 \times 20.5 = -0.62809$$

From the slope and intercept we get the equation for our least squares line:

$$\hat{Y} = -0.62809 + 0.05381X$$

And finally we can see what it all looks like: