Linear Regression

Motivation

One of the most common techniques for data analysis in a broad range of disciplines is linear regression. It's easy. It looks simple but scientific. It's ubiquitous. But what does it mean in any particular application? What can be inferred from it?

Bare bones

Suppose you have two matching sequences of numbers; they might be measurements of two quantities in the same location at different times, for example, but the possibilities are endless. Stripping everything down to the fundamental technique, we are just starting with the two sequences, X_i and Y_i, where i ranges from 1 to n, or from 0 to $n-1$.

In the following we will keep the notation as minimal as possible by omitting the limits from summations—they should be obvious from the context—and by using Einstein's summation convention whenever possible to avoid even the summation symbol ($\sum$). In this convention, the appearance of a repeated index in a product implies summation over all values of that index, so $\sum X_i X_i$ is abbreviated as simply $X_i X_i$.

Now we want to find a straight line that fits the scatter plot. But what is the criterion for a good fit? We will use the sum of the squares of the deviations of the Y variable from the line. This is standard linear least squares. Notice that it is breaking the initial symmetry–it is treating the X and Y dimensions differently.

Here we will take a shortcut to make the notation simpler. Define the mean of a variable via

$$\bar{X} = \frac{1}{n}\sum X_i \tag{1}$$

and use lower case for a sequence with its mean subtracted out:

$$x_i = X_i - \bar{X}. \tag{2}$$

The straight line will be

$$\hat{y}_i = a + bx_i. \tag{3}$$

We need to find the values of a and b that minimize the sum of squared deviations of y_i from $\hat{y}_i$; we can call this the cost function, C:

$$\begin{aligned} C(a,b) &= \sum (y_i - a - bx_i)^2 &(4)\\ &= \sum [y_i^2 + a^2 + (x_i b)^2 &(5)\\ &\qquad + 2(-y_i a - y_i x_i b + a x_i b)]. \end{aligned}$$

(The equation above is using an explicit summation everywhere because some of the terms do not involve repeated indices.)

To minimize C we set its partial derivatives to zero:

$$\frac{\partial C}{\partial a} = 2\sum (a - y_i + x_i b) = 0 \tag{6}$$

The solution is $a = 0$ because $\sum y_i = n\bar{y}_i = 0$ and similarly for the last term. The solution for b is more interesting:

$$\frac{\partial C}{\partial b} = 2\sum (x_i^2 b - y_i x_i - a x_i) = 0. \tag{7}$$

The last term is zero because x is zero-mean, leaving (dropping the summation sign and using the Einstein convention)

$$b = \frac{y_i x_i}{x_i x_i}. \tag{8}$$

Now that we have found the best fit in the least-squares sense based on this choice of cost function, what does it mean? Under what circumstances can we infer something more interesting from it?

Suppose X and Y result from an experiment or observational campaign that can be repeated many times; or from a single campaign that can be extended effectively to infinite length. Further, suppose that when the experiment is repeated or the time series lengthened, there is a particular combination of regularity and randomness such that

$$y_i = \beta x_i + \epsilon_i, \tag{9}$$

where β is constant, and ϵ is a zero-mean zero-mean residual, possibly but not necessarily random. We don't need to specify any more about it at this point. Notice that we are talking about regularity and randomness, but *not* about causality or dynamics of any kind.

Under these conditions, let's see whether b from (8) is an unbiased estimator of β. The question

is whether $E[b] = \beta$. To get the expected value, we will take the limit as the number of points increases; this could be viewed as extending a time series, or stacking a set of experiments. First, we need to substitute (9) in (8):

$$b = \frac{\beta x_i x_i + \epsilon_i x_i}{x_i x_i} \quad (10)$$

$$= \beta + \frac{\epsilon_i x_i}{x_i x_i} \quad (11)$$

Then the bias,

$$E[\beta - b] = \lim_{n\to\infty} \frac{\epsilon_i x_i}{x_i x_i} \quad (12)$$

$$= 0 \text{ iff } \epsilon_i x_i \to 0. \quad (13)$$

Therefore b is an unbiased estimator if ϵ and x are uncorrelated. Of course for any finite n, the *sample* correlation of ϵ with x will not be zero, so b could be higher or lower than β—or even of a different sign.

Suppose the condition for zero bias is met. Then if we calculate a b from our x and y data set, and subsequently have another sample of x from a process or system that we believe is identical, we can use it in (3) to predict the corresponding value of y, knowing that the prediction will not be perfect; b is not exactly β, and we have no way of knowing the corresponding ϵ. Under the stated conditions, however, the prediction will be the best bet we can make with no additional information, under the *assumption* that (9) is a good description of the *statistical* relationship.

Reversing roles

What happens if we fit a line the other way around by swapping the roles of X and Y? We might expect to get the reciprocal of b, but in general we don't. For

$$\hat{x_i} = c + dy_i. \quad (14)$$

we still get $c = 0$, but

$$d = \frac{y_i x_i}{y_i y_i}. \quad (15)$$

With the squared correlation coefficient

$$r^2 = \frac{(y_i x_i)^2}{(y_i y_i)(x_i x_i)}, \quad (16)$$

we find

$$b = r^2 \frac{1}{d}. \quad (17)$$

Only if x and y are perfectly correlated or anti-correlated will b and d be reciprocals of each other. Otherwise, the slope calculated by minimizing the squared deviations of y will always be smaller than the slope calculated by minimizing the squared deviations of x.

This is easy to understand if you think about an extreme case: suppose there is no correlation between x and y. Then the scatter plot will be just an elliptical cloud with its major axis on one of the axes; and if you scale x and y by their standard deviations, the ellipse will collapse to a circle. Then you simply cannot do better in predicting y based on x than to say "zero"; the straight line fit will be a horizontal line through the origin. But if you are asked to predict x based on y, you still can't do better than to say "y is most likely to be closer to zero than to anything else, regardless of x", so the straight line fit will be a vertical line.

An application: lack of fit

Perhaps you have a time series of temperature measurements, and you want to know whether they are trending up, so you assign your de-meaned temperature series to y and your de-meaned time variable to x. Then you calculate b using (8). Simple. What could go wrong?

Suppose your record is 2.5 years long. What do you think the residuals from your fit will look like? How might b change if the start and end of your record were shifted by 3 months?

The problem here is *lack of fit*; there is much more going on than just a trend—there is a prominent annual cycle, which, over the 2.5 year period of data, is correlated with time. In other words, your ϵ in (9) is very badly behaved. It actually dominates over the linear term, and to make matters worse, $\epsilon_i x_i$ is large. Any inference you make about the real long-term trend will almost certainly be in error, and it will certainly be unjustified. Of course, the problem is greatly exacerbated by the shortness of the time series.

In this case the lack of fit part of the problem can be addressed by using multiple linear regression,

which just means extending the model to include an annual cycle. In the simplest case one includes just the annual harmonic, and the model looks like this:

$$y_i = a + bt_i + c\cos(2\pi t_i/T) + s\sin(2\pi t_i/T) + \epsilon_i, \tag{18}$$

where T is one year. The same optimization procedure that was used to estimate b is used for c and s. Notice that because the cosine and sine parts are not zero-mean in general, a will not be zero. Multiple least squares problems like this have a neat linear algebra solution which we will not demonstrate here. The analytic solution is not good for computation, however; instead, use functions that are designed for this purpose.

Main points so far

- Regression is about predictions on the basis of statistical relationships, not about dynamics or causality.
- Linear regression fits are not symmetric with respect to the dependent variable (Y) and the independent variable (X). This is an inherent property, not a defect.
- Linear regression provides coefficients for an assumed model; it is up to you choose a model that fits well.
- If the model doesn't fit well, then ϵ and x may be correlated, in which case the linear regression estimate of β , b, will be a biased estimator.
- Even with a good model, no amount of statistical magic can conjure up information that is not present in the data set being analyzed.

Noise in X: a minor diversion

Returning to the single line fit, what happens if our X values are noisy?

Again we will go straight to the de-meaned variables and try to keep everything as simple as possible. Define x_i as the "true" value, and $x'_i = x_i + \delta_i$ as the measured value. Then the model (9) becomes, in terms of x',

$$y_i = \beta x'_i + \epsilon_i - \beta\delta_i. \tag{19}$$

Our least-squares fit regression coefficient looks the same as before, but with x' substituted for x:

$$b' = \frac{y_i x'_i}{x'_i x'_i}. \tag{20}$$

The analog of (10) is

$$b' = \frac{\sum(\beta x_i + \epsilon_i)(x_i + \delta_i)}{\sum(x_i + \delta_i)^2} \tag{21}$$

$$= \frac{\beta x_i x_i + b x_i\delta_i + \epsilon_i x_i + \epsilon_i\delta_i}{x_i x_i + 2x_i\delta_i + \delta_i\delta_i}. \tag{22}$$

If x, ϵ, and δ are all mutually uncorrelated, then upon taking the expected value this collapses down to

$$E[b'] = E\left[\frac{\beta x_i x_i}{x_i x_i + \delta_i\delta_i}\right] \tag{23}$$

$$= \frac{\beta}{1+\lambda}, \tag{24}$$

$$\lambda = \frac{E[\delta_i\delta_i]}{E[x_i x_i]}. \tag{25}$$

In this case the noise in x' biases b' lower in absolute value.

There is another possibility, though, which is that the errors in x, δ, are *uncorrelated* with x' and therefore *correlated* with x. This is a subtle distinction, but it leads to the result that b' becomes an *unbiased* estimator of β provided ϵ is still uncorrelated with x:

$$b' = \frac{\beta x'_i x'_i + \epsilon_i x'_i - \beta\delta_i x'_i}{x'_i x'_i} \tag{26}$$

$$= \beta + \frac{\epsilon_i x'_i}{x'_i x'_i} - \frac{\beta\delta_i x'_i}{x'_i x'_i}, \tag{27}$$

leaving only β after taking the expectation.

I labeled this "a minor diversion" because I think that in most cases the question of whether measurement errors are correlated with the measured values, or with the "true" values is of minor importance compared to the problems in the formulation of a model and in the adequacy of the sampling. Notice also that even in the case where b' is a biased estimator of β, it is providing the best coefficient to use for predicting any additional values of y based on new values of x', given the chosen model.