

Lines, models, and errors: Regression in the field

The fitting of a line to data is one of the commonest statistical processes in modern marine ecology. However, surprisingly few workers think closely about what they are actually doing. Fitting a line and its associated equation to data implies that there is some (possibly idealized) set of variable values that lie on a line with formula $Y = \beta_0 + \beta_1 X$, and usually, that there is scientifically useful information in the slope (β_1) and intercept (β_0) of this relationship. The practical problem faced by most scientists, with the possible exception of some laboratory-based physicists and engineers, is that the observed y and x values do not lie on the line. The aim of regression is to recover the parameters β_0 and β_1 as precisely as possible given only the y and x values. The problem is complicated by the nature of the relationship between Y and X , whether it is symmetric or not, and by the reasons the pairs (y, x) of values do not lie on the line (i.e., the error structure).

Sources of error

In this paper, I consider two main types of error: measurement error and equation error (Fuller 1987). Measurement error (δ), as its name implies, is the result of recording values that are randomly different from the actual values the scientist is attempting to measure. The measuring process itself introduces variation. The observed values (y and x) cannot therefore lie on the line, even if the actual values do. Unfortunately there is no reason to assume that even if I could measure the actual values precisely that they would lie on the line either. These deviations, the equation errors (ϵ), are commonly produced by intrinsic variation between the sample units because of the effects of randomly varying unmeasured factors or natural heterogeneity of the sampling units (genetic variation, etc.). These are what are traditionally called error.

Asymmetric relationships

Asymmetry is the traditional linear regression situation. It is simply defined as the case when there is no equation error added to X (although, as I show later, there may be measurement error). There is a dependent and an independent variable: the value of Y is assumed to depend on the value of X . In this case, the scientist is interested in the value of Y for a given value of X . The pairs (y, X) do not lie on the line. Variation of different kinds forces y to deviate from its expected value (Y) on the line. I therefore model $Y = E(y)$, the expected value of y , for given values of X . This is usually relevant when it can be assumed that the value of Y is in some sense a response of the system to, or is conditional on, the value of X .

In the simplest regression case, the value of X is assumed to be measured without error. In this case, it does not matter whether the error in y is due to measurement or equation

error. The procedure for estimating β_1 is the same—ordinary least squares regression (OLS).

The OLS estimate of the slope is given by $\hat{\beta}_{OLS}$.

$$\hat{\beta}_{OLS} = \frac{\text{cov}(Xy)}{\text{var}(X)}$$

If measurement error is added to X to give x , then clearly $\text{var}(x) > \text{var}(X)$. Less obviously, perhaps, $\text{cov}(xy) < \text{cov}(Xy)$. As a result, the OLS estimate of the slope calculated by $[\text{cov}(xy)]/[\text{var}(x)]$ will be biased and closer to zero than the OLS when there is no measurement error on X . This effect—making the slope closer to zero—is called attenuation. Clearly, the degree of attenuation depends on the magnitude of the measurement error ($\text{var}[\delta_x]$). I therefore cannot correct for it properly without knowing $\text{var}(\delta_x)$. If I do know it, then it is simple to estimate the corrected slope, sometimes called the method of moments estimator (Carroll and Ruppert 1996).

$$\hat{\beta}_{MM} = \frac{\text{var}(x)}{\text{var}(x) - \text{var}(\delta_x)} \hat{\beta}_{OLS} \quad (1)$$

Significance tests for this estimator (when $\text{var}[\delta_x]$ is estimated from direct measurements) are given in Fuller (1987).

There is an exception to the above analysis. When the values of x were chosen, and fixed, by the experimenter, then, as Berkson showed in 1950, the OLS solution is still unbiased.

I can produce reliable estimates of β_1 if more than two variables have been measured. By the method of instrumental variables, it is possible to estimate the error rate, provided that the third variable is correlated with X (Fuller 1987).

Symmetric relationships

In this situation, there is no clear dependent and independent variables; no X and Y , just Y_1 and Y_2 related by the following equations.

$$Y_1 = \beta_0 + \beta_1 Y_2 \quad \text{and} \quad Y_2 = \frac{Y_1}{\beta_1} - \frac{\beta_0}{\beta_1}$$

These equations define a single line that relates the two variables. The problem is that the actual pairs of values (y_1, y_2), even if observed without measurement error, will not lie on the line. Clearly, there is no reason to assume that these deviations originate in just one of the variables. Both are likely to be at the mercy of unmeasured factors; both could have intrinsic variability. Equally clearly, there is no reason to assume that the magnitude of these deviations will be the same in both variables ($\text{var}[\epsilon_{y1}] \neq \text{var}[\epsilon_{y2}]$). Equally, the measurement error will, if present, often be different in both ($\text{var}[\delta_{y1}] \neq \text{var}[\delta_{y2}]$). Statistically, I am arguing that a particular pair of values (y_1, y_2) will be a random sample from a bivariate distribution whose centroid (Y_1, Y_2) lies on the

line. The variances of this distribution will be $(\text{var}[\delta_{y_1}] + \text{var}[\varepsilon_{y_1}])$ and $(\text{var}[\delta_{y_2}] + \text{var}[\varepsilon_{y_2}])$, and there could easily be a covariance (correlation) between the sources of the deviations.

How should you handle this situation? The starting point is to calculate the two extremes. Assume that there is no error (of either kind) on the Y_2 . This is the observed value (i.e., $y_2 = Y_2$). The appropriate regression is now simple OLS: y_1 on y_2 . I estimate the parameters of this line and draw it. Next, I assume that there is no error of either kind on the Y_1 . All the error is on the Y_2 . The appropriate regression is simple OLS: y_2 on y_1 . If I estimate the parameters of this relationship and reorganize the equation to give Y_1 in terms of Y_2 , then I can make a statement that is undeniably true: my best estimate of the true line lies between these two extremes. In some situations, this will be enough to make a definitive answer to the question of interest; in most cases, it will not, but it will let me know the approximate region of the desired line.

If I have been able in some way to estimate the measurement error, perhaps by repeatedly measuring the same sampling unit, then these extremes can be narrowed by using the two corresponding methods of moments lines.

This process is unnecessary if I already have the error variances for the two variables. In fact, if the deviations are independent, only their ratio (λ) is needed. The well-known maximum likelihood estimate of the slope of the structural or functional equation (Kuhry and Marcus 1977; Laws and Archie 1981; Fuller 1987; McArdle 1988) gives an unbiased estimate of the slope. This is also known as orthogonal regression (Carroll and Ruppert 1996). As Carroll and Ruppert point out (and as noted by McArdle), although this method is normally spoken of with reference to measurement error (δ), equation error (ε) is often at least as important.

$$\beta_{\text{ML}} = \{\text{var}(y_1) - \lambda \text{var}(y_2) + ([\text{var}(y_1) - \lambda \text{var}(y_2)]^2 + 4\lambda \text{cov}^2(y_1, y_2))^{1/2}\} / [2 \text{cov}(y_1, y_2)] \quad (2)$$

Once again, the major problem is that I generally do not even have the ratio of the error variances. What course is left open to me? There are two main strategies here. The first is to again identify the plausible extreme cases. For example, it might be acceptable to assert that the variance of one variable will always be greater than that of the other—for example, $\text{var}(\delta_{y_1}) + \text{var}(\varepsilon_{y_1}) > \text{var}(\delta_{y_2}) + \text{var}(\varepsilon_{y_2})$ —so at minimum, in this case, $\lambda = 1$. At the other extreme, it might be possible to argue that the ratio could not be greater than, say, 3. These limits will depend on the particular situation, and care should be taken to err on the conservative side (to avoid confrontations with referees, etc.). These two extreme lines will be far closer than the OLS extremes considered above and might be sufficient for the purpose for which the line is needed. Standard errors and confidence intervals for the parameters are given in Fuller (1987).

The alternative strategy, considerably more widespread in biology and related disciplines, is to simply guess a value for λ . The two most common methods for fitting lines to data where there is error on both variables are the major axis (MA) and the reduced major axis (RMA) methods. The ma-

major axis method simply assumes that $\lambda = 1$: the error on Y_1 is equal to the error on Y_2 —that is, $\text{var}(\delta_{y_1}) + \text{var}(\varepsilon_{y_1}) = \text{var}(\delta_{y_2}) + \text{var}(\varepsilon_{y_2})$. The reduced major axis method assumes that $\lambda = \text{var}(y_1)/\text{var}(y_2)$ (i.e., that the error variances are proportional to the total variances). Both assumptions are virtually guaranteed to be wrong. The question is: To what extent are these methods robust to wrong values of λ ? Lakshminarayanan and Gunst (1984) suggest that in situations where there is appreciable error on both variables, β_{ML} is likely to be sensitive to wrong values of λ . McArdle (1988) showed that the RMA method was more robust than the MA method but that when the assumed λ value is badly wrong, the bias in both could be quite large, especially when the relationship is not strong (a small correlation coefficient). In fact, the largest bias (as a proportion of the RMA slope) possible is $[(1/|r|) - 1]$, the difference between the β_{OLS} for x on y and the β_{RMA} .

If the errors in one variable are not independent of errors in the other, then none of these methods are appropriate. Fuller (1987) and Reilman et al. (1985) offer a method for this situation. However, the sheer unwieldiness of the formula makes it unlikely that it will be used often (even if the correlation between the deviations were known).

Least squares regression

The relationships between these methods can be clarified by recognizing that they are all in fact special cases of one least squares method. The best fitting line is the one that minimizes the squared distances from the observed points to the corresponding (X, Y) points on the line. What varies is how this distance is measured. In the error-in-variables situation, points can deviate from the line in both the vertical and horizontal directions, so any distance (for the i th data point) I calculate will depend on $\gamma_{xi} = \delta_{xi} + \varepsilon_{xi}$ and $\gamma_{yi} = \delta_{yi} + \varepsilon_{yi}$, the vertical and horizontal deviations. I want the statistical distance, so if the errors are independent, I use Eq. 3.

$$[i\text{th statistical distance}]^2 = \frac{\gamma_{xi}^2}{\sigma_{\gamma x}^2} + \frac{\gamma_{yi}^2}{\sigma_{\gamma y}^2} \quad (3)$$

If they are not independent, I have to use the more complex Mahalanobis distance, which has been scaled by $\Sigma_{\gamma\gamma}^{-1}$, the inverse of the error variance covariance matrix (Fuller 1987).

The commonly used methods I described above are special cases of the independent errors model. For example, if I assume there is no error in X ($\gamma_x = 0$), then when I minimize the sum of the squared distances in Eq. 3, I get OLS y on X . If I assume instead that there is no error on Y ($\gamma_y = 0$), then the least squares solution yields OLS x on Y , and so on. One insight that this approach gives me is that if I use principal component analysis (Seber 1984) to produce a trend line, I am minimizing the sum of squared orthogonal distances. That is, I am implicitly assuming that the observed points deviate from the points on the line (X, Y) by random error with equal variances. This might or might not be plausible in any given situation. It cannot be recommended as a general strategy.

Model I and Model II regression

Biologists and scientists weaned on the text *Biometry* by Sokal and Rohlf (1969, 1981) use a different terminology. Based on an analogy with Model I and Model II analysis of variance, the asymmetric case with no measurement error on X is referred to as Model I. The original definition of Model I regression was based on four assumptions, the first of which was:

1. The independent variable X is measured without error. We therefore say that the X 's are "fixed." We mean by this that only Y , the dependent variable, is a random variable; X does not vary at random but is under the control of the investigator (Sokal and Rohlf 1969, p. 408–409).

Clearly, this is a definition of classical OLS regression y on X .

Model II regression, on the other hand, is basically everything else.

In *Model II regression*, the independent variable is also measured with error. We do not consider X to be fixed and at the control of the investigator (Sokal and Rohlf 1969, p. 410).

This dichotomy, Model I and Model II regression, does not seem to me very useful. The distinction between dependent and independent variables (the asymmetry of the relationship) is split between both. The corrected OLS estimator (the method of moments estimator, Eq. 1) is a Model II method, although it is appropriate for an asymmetric situation with error on X . Basically, Model I is just an extreme Model II situation—zero error on X . The OLS method is simply a special form of the more general least squares approach. There seems to be little point in separating them. The main difference is in the error variances, and that will depend on whether there is equation error in the X as well as the Y (i.e., whether the relationship is symmetric or not).

Practical considerations

The first step is to accept that there will nearly always be some error in x , particularly in the field sciences. The following series of questions should help decide what to do.

1. What is the relation between X and Y , the unobservable points on the line, that you want to know about? Is the relationship asymmetric with the equation error in y , the dependent variable?
2. If so, is the measurement error in x enough to worry about? Basically express your worst-case guess of $\text{var}(\delta_x)$ as a proportion (p) of the total $\text{var}(x)$. Now calculate $p/(1-p)$. This is the worst-case amount of attenuation as a proportion of the OLS estimate. If you think this would not worry even a hostile reader (I said this was a worst case), then go ahead and use OLS, but state your argument as to why you feel that level of attenuation can be ignored. Your readers have to be convinced too.
3. If the relationship between X and Y is likely to be symmetric—equation error on both variables—then you have to consider the ratio (λ) of total error variances for both variables. Consider first the measurement error: Which variable will have more? Then consider the

other forcing variables that perturb the Y_1 and Y_2 away from the line—the equation error. Sometimes it is possible to get some idea of the relative range of these values. Put the two sources of error together and make a guess as to the extremes that λ might take. Plug them into Eq. 3 and estimate the range of slopes that might be possible. You now have to add sampling error to these estimates, so putting confidence intervals (Fuller 1987) on these estimates would now be essential. Alternatively some workers will simply try fitting major or reduced major axis regression models (both are principal components solutions, McArdle 1988). However, they would have to be aware that the assumed λ will be wrong and the slope estimate biased, so they should explain why this does not affect the conclusions they want to draw. It will be up to the reader to decide whether to believe them or not.

Conclusions

In real situations, any line fitted through data will have parameter estimates that are biased. The important step is to show that it does not matter to the conclusions that are being drawn. If there is no interest in the parameters of the equation (apart from demonstrating that $\beta \neq 0$) and the main aim is simply a cosmetic line through the data, then it does not matter much which method is used. The test of $\rho = 0$ will suffice for inference. However, if the slope, the intercept, or both parameters of the line are important, then care must be taken that the scientific conclusions follow from the data. If the two variables are highly correlated, then there is unlikely to be a problem—the OLS y on X is r^2 times the OLS x on Y . All available solutions lie between these two extremes. If, however, the range of possible slopes is too large for the purposes of the study, then more information will be needed before it can be improved. In the absence of such information, a rigorous scientist will accept that there is no way a model sensibly can be fitted to the data, so there is no way of making inferences about the model's parameters.

Calbet's problem with error

The specific problem addressed in Calbet (2001) and identified correctly by Laws (this issue) seems relatively straightforward. The conclusions in the original paper—that the slope of the line was not 1—clearly depended on the measurement error in x , the productivity measurement. No estimate of this error was given in the paper. Laws justifiably shows that the conclusion was sensitive to the magnitude of the measurement error. Calbet and Prairie (this issue) present evidence that the measurement error is so small that the conclusions remain unaffected.

Basically the whole exchange could be rewritten as:

Laws: You didn't give an error for X —your conclusion depends on that.

Calbet: Oops, sorry. Look, it's small; my conclusion stands.

Provided readers accept the estimate of measurement er-

ror, the matter is settled. However, because there will seldom be agreement on the size of measurement error, it seems to me that a slightly more sophisticated approach is possible.

The problem faced by Calbet (2001) can plausibly be seen as a problem of the first kind: an asymmetric relationship where Y is dependent on X . The grazing can be seen as a response to the primary productivity. The interesting question was: Given this amount of primary productivity, what amount of grazing could I expect? The primary productivity to which the grazing responds is not measured precisely. Measurement error from a number of sources creeps in. The real problem is estimating the measurement error. In this particular example, there seems little chance of getting any direct estimate. Certainly, guessing the percentage of the total $\text{var}(x)$ that is measurement error, perhaps from other studies, and using that in the method of moments estimator could only improve on the OLS estimate. However no significance test would be possible. Perhaps a more sensible approach would be to work backwards. The null hypothesis that a slope estimate equals 1 can be tested by seeing whether 1 lies inside the confidence interval for the slope. One simple heuristic approach would be to take, in this case, the upper limit of the OLS confidence interval (which, Calbet showed, does not include 1) and see how large the measurement error could be before the interval includes 1. So long as this level of error was implausibly large, then readers might be convinced that the true slope was indeed <1 —without, sadly, being able to say what the slope was precisely. This leaves the essential decision to the reader and summarizes all the available information to help them make it. Informed readers will have their own opinions of what a plausible level of measurement error is.

Brian H. McArdle

Department of Statistics
University of Auckland
Private Bag 92019
Auckland, New Zealand

References

- BERKSON, J. 1950. Are there two regressions? *J. Am. Stat. Assoc.* **45**: 164–180.
- CALBET, A. 2001. Mesoplankton grazing effect on primary production: A global comparative analysis in marine ecosystems. *Limnol. Oceanogr.* **46**: 1824–1830.
- CARROLL, R., AND D. RUPPERT. 1996. The use and misuse of orthogonal regression in measurement error models. *Am. Statistician* **50**: 1–6.
- FULLER, W. A. 1987. *Measurement error models*. Wiley.
- KUHRY, B., AND L. F. MARCUS. 1977. Bivariate linear models in biometry. *Syst. Zool.* **26**: 201–209.
- LAKSHMINARAYANAN, M. Y., AND R. F. GUNST. 1984. Estimation of parameters in linear structural relationships: Sensitivity to the choice of the ratio of error variances. *Biometrika* **71**: 569–573.
- LAWS, E. A., AND J. W. ARCHIE. 1981. Appropriate use of regression analysis in marine biology. *Mar. Biol.* **65**: 13–16.
- MCARDLE, B. H. 1988. The structural relationship: Regression in biology. *Can. J. Zool.* **66**: 2329–2339.
- REILMAN, M. A., R. F. GUNST, AND M. LAKSHMINARAYANAN. 1985. Structural model estimation with correlated measurement errors. *Biometrika* **72**: 669–672.
- SEBER, G. A. F. 1984. *Multivariate observations*. Wiley.
- SOKAL, R. R., AND F. J. ROHLF. 1969. *Biometry*. W.H. Freeman.
- , AND ———. 1981. *Biometry*, 2nd Edition. W.H. Freeman.

Received: 2 October 2002

Accepted: 14 December 2002

Amended: 18 December 2002