# Inverse Regression from Longitudinal Data

Geoffrey Jones[1]
[1] Massey University, Private Bag 11222, Palmerston North, New Zealand

## Abstract
Inverse regression, or statistical calibration, uses the estimated relationship between a response *Y* and a covariate *x* to infer the values of unknown *x*'s from their observed *Y*'s. Typically *x* is univariate but *Y* may be multivariate. A brief review of the basic theory will be given, followed by consideration of the problems involved in extending these approaches to longitudinal data, i.e. where the training data consists of groups of observations on distinct individuals. A Bayesian analysis using MCMC is shown to give a flexible framework for solving these problems. An example concerning the age determination of tern chicks from their wingspan and weight measurements will be used for illustration.

**Key Words:** Calibration, Hierarchical model, Multilevel model, MCMC

## 1. A Brief Review

Inverse regression, or statistical calibration, arises in a variety of contexts, most notably in laboratory analyses where known amounts or concentrations *x* of a target chemical are treated to produce responses *Y*; the resulting pairs (*x*, *Y*) are used as training data to establish a calibration curve. Further samples with unknown concentrations can then be treated in the same way and their responses used to infer the unknown concentrations, along with their precision. If the training data comprise repeated measures on distinct groups or individuals with significantly different characteristics, then the methodology must be adapted to account for this structure. One place where this occurs is in ecological settings where body measurements are used to determine the age of an animal, based on repeated measures on a sample of animals of the same species. Such a situation was encountered by Keedwell (2002), who hoped to use measurements on black-fronted tern chicks in New Zealand to estimate the age of chicks in an ongoing conservation effort.

We begin with a brief review of the methods used when the training data are assumed to be independent. Many different approaches have been proposed in the literature, governed in part by the response *Y* is univariate or multivariate, and whether the relationship is linear or nonlinear. For a more thorough review, see Osborne (1991) and the references contained therein.

### 1.1 Univariate
Figure 1 plots the age (days) and wing length (mm) of an individual tern chick measured over a number of days. The growth can be seen to be approximately linear over this range. A linear model $Y = \alpha + \beta x + \varepsilon$ has been fitted to the data, assuming i.i.d. errors. The fitted line and 95% prediction limits are shown on the graph. Given another wing length measurement $Y_0 = 90$ mm, the classical estimator

$$\hat{x}_0 = (Y_0 - \hat{\alpha}) / \hat{\beta},$$

found by inverting the estimated regression line, gives the *x*-value of the point on the calibration curve at which *y* is $Y_0$. The prediction limits may be similarly inverted to give an exact confidence interval for the true value $x_0$. To see that this is also the maximum likelihood estimator, we can consider adding the point $(x_0, Y_0)$ to the training data and refitting the model: the error sum of squares is clearly minimized when $(x_0, Y_0)$ is on the line fitted to the training data alone since this line minimizes the sum over the training data and the new point will contribute zero. Note too that the addition of the new point does not change the estimated calibration curve (in contrast to multivariate calibration considered in section 1.2.).

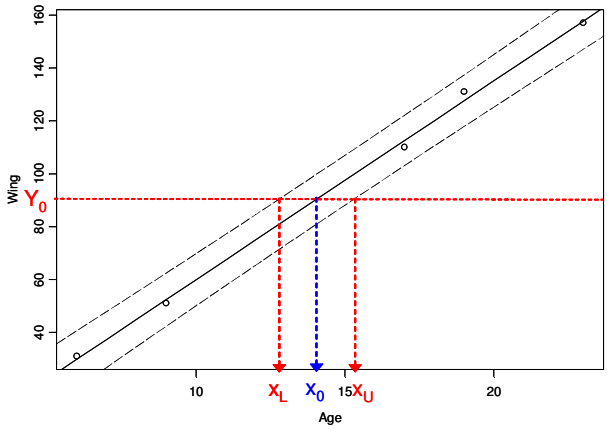


**Figure 1:** The classical estimator $X_0$ and a confidence interval $(X_l, X_U)$ given response $Y_0$ are obtained by inverting the regression line and prediction interval fitted to the training data.

Unfortunately the classical estimator has infinite variance, so some prefer an alternative estimator $\tilde{x}_0$ derived from regressing $x$ on $Y$. This alternative may not be very sensible, however, when the relationship is nonlinear. Other alternatives include Bayesian approaches and nonparametric local smoothing.

## 1.2 Multivariate

We focus here on situations where the dimension $q$ of the response $Y$ is greater than the dimension $p$ of the covariate $x$. In some cases $q$ is much greater than $p$, as in mass spectrometry where $Y$ is a spectrum. In our tern chick example we have measurements of wing span (mm) and weight (g), so $q = 2$. A major consequence of having $q > p$ is that we cannot now, in general, find an $x_0$ for which the fitted $\hat{Y}_0$ from the calibration model is equal to the observed $Y_0$. The situation is illustrated in Figure 2. Appropriate functions are fitted to each component of $Y$ using the training data, in a multivariate framework to allow for correlation between the error terms in the component models. We can then plot the fitted curve traced out in $R^q$ as $x$ varies. The predicted value of $x_0$ given a new response $Y_0$ should correspond to a point on the fitted curve that is in some sense as close as possible to $Y_0$. The metric for this distance should take into account the correlation between the component errors, as estimated by the model. This gives essentially the estimator proposed by Brown (1982) for the linear and Clarke (1992) the nonlinear case. It can be represented graphically by adding an ellipse based on the estimated error covariance.

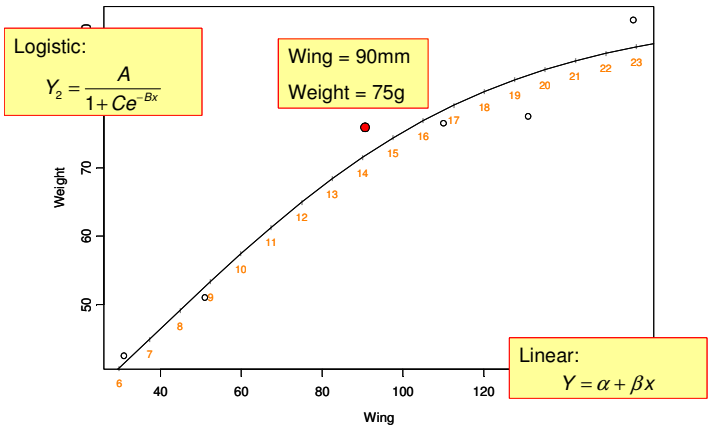


**Figure 2:** Fitted model for wing length and weight of a tern chick from training data, with new point $Y_0$ = (90, 75) added. The fitted curve is drawn for a range of age values that are shown along the curve.

Brown and Sundberg (1989) point out that there may be useful information in the distance of the new point from the fitted line: if this distance is unusually large, it suggests that the new point does not come from the same population as the training data. They also propose an alternative, maximum likelihood, approach. Note that we can obtain the maximum likelihood estimator by adding the new point $(x_0, Y_0)$ to the training data and refitting the model for a range of values of $x_0$, then using the resulting *profile likelihood* to obtain a point estimate and confidence limits – this is a very general approach that can be used in a wide variety of calibration settings. Note too that because $Y_0$ is not on the fitted curve, the addition of the new point will change the parameter estimates of the calibration model, so that the classical estimator given by Brown (1982) is not now the same as the maximum likelihood estimator.

# 2. Longitudinal

## 2.1 Univariate Longitudinal

We now suppose that the training data consist of repeated measures on $I$ individuals: $(x_{ij}, Y_{ij})$, for $j = 1,...,n_i$, $i = 1,...I$, and that the parameters describing the relationship between $Y$ and $x$ vary significantly between individuals. An example is shown in Figure 3, which plots data on a number of different tern chicks. The calibration model estimated from the training data now consists of a family of growth curve $Y = \alpha_i + \beta_i x$ in conjunction with a model to describe how the random vector $(\alpha_i, \beta_i)'$ varies between individuals; a common approach would be to assume that they are normally distributed with, say, mean $(\alpha, \beta)'$ and precision matrix $\Gamma$.

Given a single measurement $Y_0$ on a new individual, there are various ways in which a predicted value and confidence interval for $x_0$ could be derived. Following the approach of the classical estimator, we could derive a point estimate and confidence interval by considering the mean and variance of

$$Y_0 - \hat{\alpha} - \hat{\beta} x_0 = (Y_0 - \alpha_0 - \beta_0 x_0) + (\alpha_0 + \beta_0 x_0 - \alpha - \beta x_0) - (\hat{\alpha} + \hat{\beta} x_0 - \alpha - \beta x_0)$$

where $(\hat{\alpha}, \hat{\beta})'$ is the estimate from the training data and $(\alpha_0, \beta_0)'$ is the parameter vector for the new individual. This suggests a point estimate of $\hat{x}_0 = (Y_0 - \hat{\alpha}) / \hat{\beta}$. The confidence interval needs a Satterthwaite approximation to the degrees of freedom of the estimated variance of the above expression. See Wood et al. (1981).

We again have the simple option of adding the new point to the training data and fitting the model for a range of values of $x_0$ to derive the profile likelihood function. The maximizing $x_0$ is then the MLE, and an approximate confidence interval can be found by using a chi-squared approximation for the profile likelihood ratio statistic.

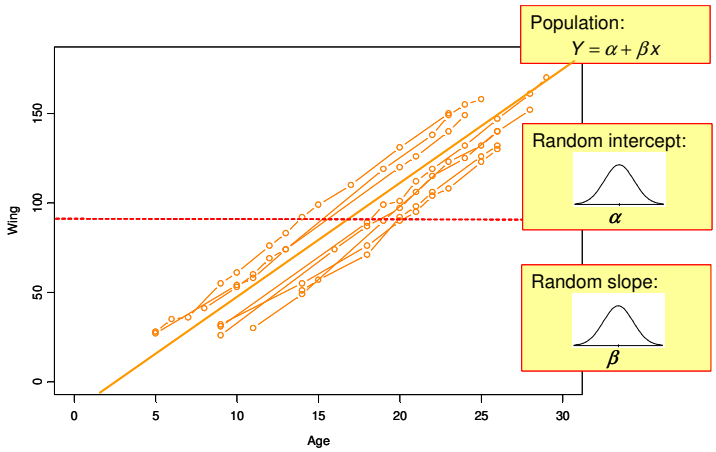


**Figure 3:** Longitudinal data on wing length for a sample of tern chicks. Data for the same chick are joined with lines. A population average curve (solid line) has been estimated from the training data and added to the plot, and a horizontal line at 90mm representing the wing length of a new individual of unknown age.

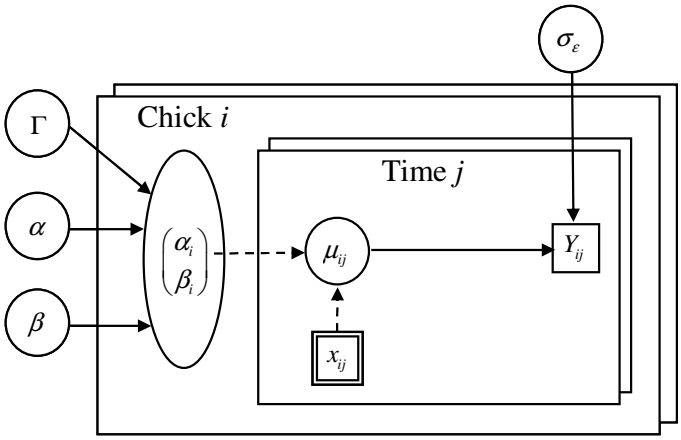


**Figure 4:** Graphical model for Tern wing length data. See Gilks et al. (1995).

A third alternative is to use a Bayesian framework and Markov chain Monte Carlo (MCMC) estimation. Uninformative priors can be specified for the model parameters $\alpha$, $\beta$, $\Gamma$ and $\sigma_\varepsilon$, and the unknown age $x_0$ estimated simply by making it a stochastic node in the MCMC program. This method produces an approximation to the exact posterior distribution of $x_0$ given the data.

All three of the above approaches can be adapted to the situation where repeated measures are available on the new individual, and to nonlinear specification of the model relating $Y$ to $x$. The availability of repeated measures can potentially lead to much better inference, as there should now be information available on the growth parameters of the new individual. That this does not happen to any significant extent in the tern wing span example can be explained by reference to Figure 3: the slope parameters $\beta_i$ vary little between birds, with most of the variability in the population coming from the intercepts $\alpha_i$. Figure 3 shows that the age of a new chick cannot be determined accurately from its wing length alone. We now consider the utility of adding other body measurements.

## 2.1 Multivariate Longitudinal

As in the non-hierarchical case of section 1.2, we cannot now choose $x_0$ so as to place the data for a new individual on the population growth curve. This suggests that there is now information, even from a single observed $Y_0$, about the growth parameters of the new individual. To examine this issue we first consider a linear example in which the population model parameters are assumed known. Suppose

$$Y_0 \sim N(X_0\beta_0, \Omega) \quad \text{where} \quad \beta_0 \sim N(\beta, \Gamma)$$

where, for example with $q = 2$,

$$X_0 = \begin{pmatrix} 1 & x_0 & 0 & 0 \\ 0 & 0 & 1 & x_0 \end{pmatrix} \quad \text{and} \quad \beta = (\alpha_1, \beta_1, \alpha_2, \beta_2)'$$

and let

$$\hat{\beta}_0 = (X_0^t \Omega X_0)^- X_0^t \Omega Y_0 \quad \text{and} \quad \Xi_0 = X_0^t \Omega X_0 .$$

If we regard $x_0$ and $\beta_0$ as fixed but unknown parameters, then a little algebra gives

$$\begin{aligned} -2 \log L(X_0, \beta_0) = K &+ (Y_0 - X_0\hat{\beta}_0)^t \Omega (Y_0 - X_0\hat{\beta}_0) + (\hat{\beta}_0 - \beta)^t \Xi_0 (\Xi_0 + \Gamma)^{-1} \Gamma (\hat{\beta}_0 - \beta) \\ &+ [\beta_0 - (\Xi_0 + \Gamma)^{-1}(\Xi_0\hat{\beta}_0 + \Gamma\beta)]^t (\Xi_0 + \Gamma)[\beta_0 - (\Xi_0 + \Gamma)^{-1}(\Xi_0\hat{\beta}_0 + \Gamma\beta)] \end{aligned} \tag{1}$$

where $K$ is a constant. For any fixed value of $x_0$ the likelihood will be maximized by setting

$$\beta_0 = (\Xi_0 + \Gamma)^{-1}(\Xi_0\hat{\beta}_0 + \Gamma\beta) \tag{2}$$

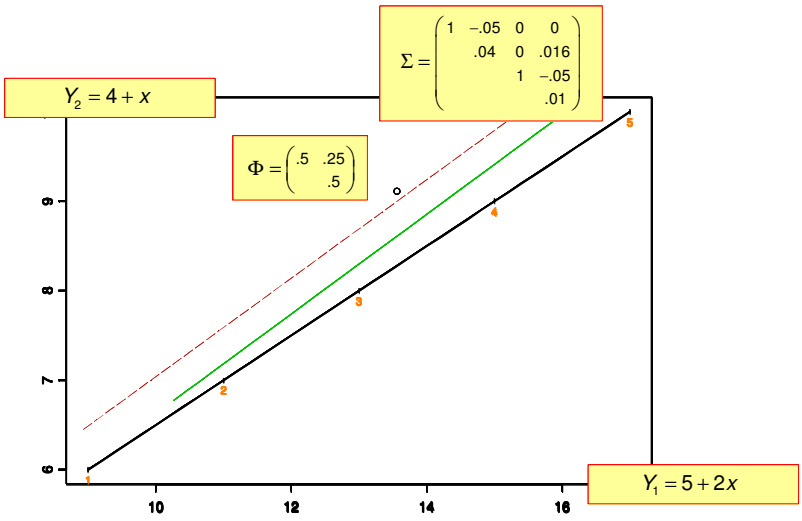


**Figure 5:** Simulated bilinear hierarchical model with parameters as shown. The covariance matrices shown, Σ and Φ, are the inverses of the precision matrices Γ and Ω. The solid black line shows the population response path with $x$ values alongside. The solid green line is the simulated response path for a new individual, the plotted point a simulated response $Y_0$ for this individual at $x_0 = 4$, and the red broken line the estimated response path for this individual.

and $Y_0 - X_0\hat{\beta}_0 = 0$ so the profile likelihood for $x_0$, to within a constant, is

$$-2\log L_P(X_o) = (\hat{\beta}_0 - \beta)^t \Xi_0(\Xi_0 + \Gamma)^{-1}\Gamma(\hat{\beta}_0 - \beta).$$

Alternatively we can regard $\beta_0$ as a random effect and integrate it out of (1) to give the marginal likelihood for $x_0$

$$-2\log L_M(X_o) = -2\log L_P(X_o) + \log\left|\Xi_0 + \Gamma\right|.$$

The point estimates and confidence intervals derived from these two expressions tend to be very similar in practice.

Figure 5 illustrates this situation for a bilinear hierarchical model, i.e. for which $q = 2$ and both components of $Y$ are linearly related to $x$. The population response curve is a straight line in $R^2$ traced out by varying the value of $x$. Individual response paths cluster around this population curve. Given a response $Y_0$ for a new individual, an estimate of $x_0$ can be obtained by maximizing the profile or marginal likelihood as given above. The estimated response path for the new individual, derived from (2), is an average of the population response path and the path implied by $\hat{\beta}$ that passes through the observed $Y_0$; this average is weighted according to the precisions Γ and Ω in the two levels of the hierarchical model. Thus the information about $\beta_0$ derivable from the response $Y_0$ will depend on the relative sizes of the two components of variability.

In practice the population model parameters will not be known and will have to be estimated from the training data. An analytical analysis now becomes intractable even for the linear case. Fitting the population model to the training data is not straightforward; standard software has routines for fitting univariate linear and nonlinear multilevel models, and these can be adapted to fit multivariate versions by introducing indictors for the different components of the response, although there is insufficient flexibility in specifying the full covariance structure. See the discussion by Davidian and Giltinan (1995).

The Bayesian analysis of the previous section is however easily extendible to the multivariate case, even when there are nonlinear component models. It can even cope with additional complexities such as a mixture population model, as in the case of the terns where some chicks were identified visually as “slow-growing”. Figure 6 shows some examples of posterior distributions from the tern chick data. Age determination can be seen to be considerable more precise, to within one day, when using both the wing length and weight measurements.

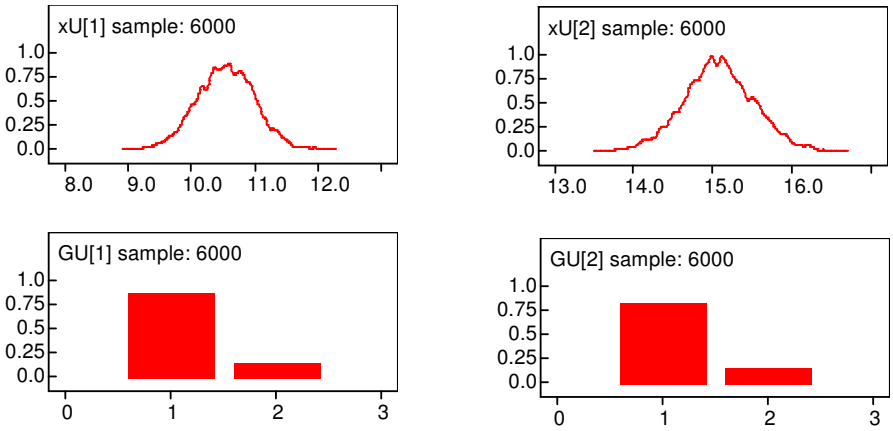


**Figure 6:** Posterior distributions for age (xU) and group membership (GU = 1 "normal" or 2 "slow") of two tern chicks with known wing length and weight..

## Acknowledgements

The author wishes to thank Jane Keedwell for providing the tern chick data.

## References

Brown, P.J (1987) Multivariate calibration (with discussion). *J. R. Statist. Soc.* B, 44, 287-321.

Brown, P.J and Sundberg, R. (1989) Prediction giagnostics and updating in multivariate calibration. *Biometrika*, 76, 349-361.

Clarke, G.P.Y. (1992) Inverse estimates from a multiresponse model. *Biometrics*, 48, 1081-1094.

Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall.

Gilks, W.R, Richardson, S., and Spiegelhalter, D.J. (1995) Introducing Markov chain Monte Carlo. In: *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics* (eds W.R. Gilks, S. Richardson, and D.J. Spiegelhalter), pp. 1--19. London: Chapman \& Hall.

Keedwell, J.R. (20022) Blach-fronted terns and banded dotterels: causes of mortality and comparisons of survival. Unpublished PhD thesis, Department of Ecology, Massey University NZ.

Osborne, C. (1991) Statistical calibration: a review. *International Statistical Review*, 59, 309-336.

Wood, J.T., Carpenter, S.M, and Poole, W.E. (1981) Confidence intervals for ages of marsupials determined from body measurements. *Australian Wildlife Research*, 8, 269-274.