

Stat 544 Spring 2008
Mini-Project #2

The web page http://www.ncsu.edu/chemistry/resource/NeXS/linear_calibration.html has some data from what purports to be a real "linear calibration" problem from chemistry. (I'm actually a bit skeptical as to whether the data are real, given the fact that the ranges in y for fixed x are almost too uniform to believe, but we'll ignore this worry.) Below are the data from the page. x is (supposedly "known") concentration of Riboflavin (in mcg/mL) of "standard" liquid samples run through some kind of instrument that reads out an "Intensity" y in units peculiar to the instrument.

x	y
0.00	8
0.00	5
0.10	17
0.10	21
0.30	47
0.30	44
0.50	70
0.50	73
0.70	95
0.70	98

The ultimate goal in a calibration problem like this is to be able to take a new observed intensity y_{new} and estimate the corresponding concentration of Riboflavin that produced it, x_{new} .

Begin analysis of this situation using the usual SLR model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for the ϵ_i iid $N(0, \sigma^2)$. This is $y_i \sim \text{ind } N(\beta_0 + \beta_1 x_i, \sigma^2)$ and for $i = 1, 2, \dots, 10$ the 3 model parameters are β_0, β_1 , and σ^2 . If one adds to the modeling an 11th data pair $(x_{\text{new}}, y_{\text{new}})$, the resulting model has 4 unknown parameters, $\beta_0, \beta_1, \sigma^2$, and x_{new} . Use what you convince yourself are fairly noninformative prior assumptions and do a Bayes analysis here for cases where $y_{\text{new}} = 10, 20, 30, 40, 50, 60, 70, 80, 90$. (You should be able to accomplish this without making a separate run for each value of y_{new} .) (What are credible intervals for all the model parameters?)

Investigate whether taking the fairly obviously discrete/rounded nature of the measured intensities into account materially changes your initial conclusions.

Limit what you type up to turn in to a cover page (**use one!**) plus at most 6 typewritten pages (including whatever figures you want to include). Use at least 11 point fonts, 1.5 line spacing, and 1 inch left and right margins. Also include an Appendix with "commented" WinBUGS and/or R code that you have used. (This Appendix does not count in the above "6 typewritten pages" limit.)

A Note on the Berkson Model, Identifiability, and Bayes Analysis for This Problem

The "standard" solutions fed into the instrument during this calibration study were made up with the intention of producing Riboflavin concentration x as in the table above. It is, of course, inconceivable that the *actual* concentrations were *exactly* as in the table, and presumably the instrument reacts to what is actually fed, not what the experimenter intends! Berkson's famous analysis of the situation is then as follows. Suppose that the actual concentration produced in standard solution i is

$$x'_i = x_i + \eta_i$$

where the η_i are iid $N(0, \sigma_\eta^2)$ independent of the ϵ_i (that remain as above), and that the observed intensity is then describable as

$$y_i = \beta_0 + \beta_1 x'_i + \epsilon_i = \beta_0 + \beta_1 x_i + \beta_1 \eta_i + \epsilon_i$$

This model for the 10 pairs (x_i, y_i) has 4 model parameters $\beta_0, \beta_1, \sigma_\eta^2$, and σ^2 , and if an 11th data pair $(x_{\text{new}}, y_{\text{new}})$ is included, there are 5 parameters. Investigate what is possible for a Bayes analysis here. (Note the attached discussion regarding Berkson's model.) Note that

in the Berkson model

$$y_i = \beta_0 + \beta_1 x_i + \beta_1 \eta_i + \epsilon_i$$

if we let

$$\beta_1 \eta_i + \epsilon_i = \gamma_i$$

then with

$$\sigma_\gamma^2 = \beta_1^2 \sigma_\eta^2 + \sigma^2$$

we have nothing but the usual SLR model with error variance σ_γ^2 . In particular, the Berkson model with parameters $\beta_0, \beta_1, \sigma^2$, and σ_η^2 is not identifiable. (See Section 3.3.1 of the course outline. For a given set of SLR parameters β_0, β_1 , and σ_γ^2 , there are many σ_η^2 and σ^2 pairs that produce the same σ_γ^2 and therefore same distribution of the observables. So, for example, if b_0, b_1 , and SSE/n are the usual Stat 511 least squares MLEs of the SLR parameters, any $(\sigma_\eta^2, \sigma^2)$ pair with $SSE/n = b_1^2 \sigma_\eta^2 + \sigma^2$ will maximize the Berkson likelihood. It is therefore really not possible to estimate all of the Berkson parameters.)

If one tries to do Bayes estimation of all 4 Berkson parameters, what one gets will be very prior-dependent. For example, one seemingly sensible way to proceed is to place priors on β_0, β_1 , and σ_γ^2 as if one had the SLR model (which one does) and then make some prior assumption about

$$k = \frac{\sigma_\eta^2}{\sigma^2}$$

For example, independent flat priors on β_0, β_1 , and $\ln \sigma_\gamma$ produce inferences like those from standard linear models theory for $\beta_0, \beta_1, y_{\text{new}}$, and x_{new} . Then an independent prior on k leads to (completely prior-dependent) inferences for σ_η^2 and σ^2 based on the identities

$$\sigma^2 = \frac{\sigma_\gamma^2}{k\beta_1^2 + 1} \text{ and } \sigma_\eta^2 = k\sigma^2$$

But one should not fool oneself ... in terms of separating σ_η^2 and σ^2 , all one is looking at in the posterior is what one has put in in terms of prior assumptions about k .

Some sets of Bayes assumptions here may lead to seemingly bizarre behavior of MCMC simulations. Sometimes, what that kind of thing is telling you is that you've got a likelihood/posterior that has a "ridge" in it where the sampler wanders around with wildly different values of the parameter vector giving very similar posterior densities. That is, lack of identifiability can lead to poorly behaved MCMC.

A Note on Bayes Simple Linear Calibration

What at first seems like a fairly innocuous/simple problem of Bayes inference using a simple linear regression model turns out to be more subtle than one might think. That is, under a model for n observables

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the ϵ_i are iid $N(0, \sigma^2)$ we consider the Bayes estimation of x_{new} that produces an $(n + 1)$ st observation y_{new} .

Probably the most obvious analysis of this situation treats $\beta_0, \beta_1, \sigma^2$, and x_{new} as parameters and (for $f(\cdot|\mu, \sigma^2)$ the $N(\mu, \sigma^2)$ pdf) uses likelihood function

$$L(\beta_0, \beta_1, \sigma^2, x_{\text{new}}) = \left(\prod_{i=1}^n f(y_i | \beta_0 + \beta_1 x_i, \sigma^2) \right) f(y_{\text{new}} | \beta_0 + \beta_1 x_{\text{new}}, \sigma^2)$$

When one then supplies a prior for $(\beta_0, \beta_1, \sigma^2, x_{\text{new}})$, say a prior of independence of $(\beta_0, \beta_1), \sigma^2$, and x_{new} ,

$$g_1(\beta_0, \beta_1) g_2(\sigma^2) g_3(x_{\text{new}})$$

one then has a joint distribution of $\mathbf{y} = (y_1, \dots, y_n, y_{\text{new}})$ and $(\beta_0, \beta_1, \sigma^2, x_{\text{new}})$ specified by the density

$$\left(\prod_{i=1}^n f(y_i | \beta_0 + \beta_1 x_i, \sigma^2) \right) f(y_{\text{new}} | \beta_0 + \beta_1 x_{\text{new}}, \sigma^2) g_1(\beta_0, \beta_1) g_2(\sigma^2) g_3(x_{\text{new}}) \quad (1)$$

This joint distribution has the corresponding DAG in Figure 1.

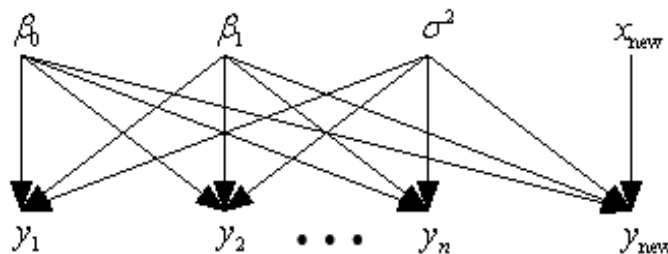


Figure 1: DAG for a First Analysis of the Calibration Problem

Notice that since x_i are not random, they do not appear in this DAG, while x_{new} does appear because of the prior that it is given.

Analysis with the joint distribution (and thus the conditional/posterior $(\beta_0, \beta_1, \sigma^2, x_{\text{new}})$ given \mathbf{y}) has a number of potentially initially counter-intuitive features, at least for many choices of prior. The DAG in Figure 1 and form (1) imply that the potentially completely hypothetical y_{new} will in general contribute to the posterior for $(\beta_0, \beta_1, \sigma^2)$. Its use in this present form tends to attenuate the slope β_1 (reduce the size of the slope) and increase the error variance σ^2 (say relative to least squares estimates of these). Further, this effect is exacerbated if more than one hypothetical future value of y is employed and included in (1). And what is more, in general if multiple future (even completely hypothetical) future values of y are employed, what is obtained for an inference for one of the corresponding x 's depends upon what other y 's are included in the analysis!!! This is hardly an appealing circumstance.

A second line of thinking about this problem is the following. Simple algebra says that if $y = \beta_0 + \beta_1 x + \epsilon$, then

$$x = \frac{y - \beta_0 - \epsilon}{\beta_1}$$

and this *perhaps* motivates the following thinking. If I think of y_{new} as fixed, maybe a sensible ?prior? distribution for x_{new} conditioned on β_0, β_1 and σ^2 is

$$x_{\text{new}} \sim N\left(\frac{y_{\text{new}} - \beta_0}{\beta_1}, \frac{\sigma^2}{\beta_1^2}\right) \quad (2)$$

So, letting y_{new} disappear from a DAG representation of a second model for this situation (since it is now treated as a fixed value), one has the representation in Figure 2.

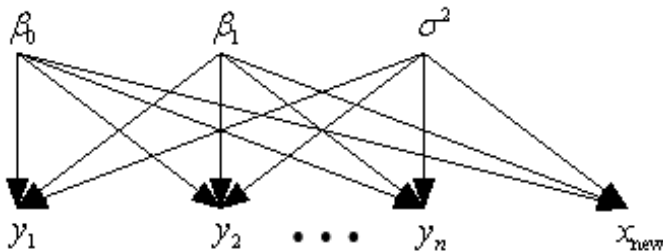


Figure 2: DAG for a Second Analysis of the Calibration Problem

With the assumption (2) and priors for β_0, β_1 , and σ^2 as before, corresponding to Figure 2 is a joint distribution specified by

$$\left(\prod_{i=1}^n f(y_i | \beta_0 + \beta_1 x_i, \sigma^2) \right) f\left(x_{\text{new}} | \frac{y_{\text{new}} - \beta_0}{\beta_1}, \frac{\sigma^2}{\beta_1^2}\right) g_1(\beta_0, \beta_1) g_2(\sigma^2) \quad (3)$$

As x_{new} is not observed, it does not contribute to inferences about $(\beta_0, \beta_1, \sigma^2)$ and when there are multiple new y 's under discussion, predictive posteriors of new x 's are not affected by the "other" new values of y included in the analysis. This structure seems to behave

more "sensibly" in the calibration context than the first one. The question of how much sense the (?prior?) assumption (2) makes seems to me to be up for debate.

What is perhaps a bit comforting is that one can realize the structure (3) as a special instance of structure (1) (that, of course, has properties that structure (1) does not possess in general). To this end, notice that

$$f\left(x_{\text{new}} \mid \frac{y_{\text{new}} - \beta_0}{\beta_1}, \frac{\sigma^2}{\beta_1^2}\right) = |\beta_1| f(y_{\text{new}} \mid \beta_0 + \beta_1 x_{\text{new}}, \sigma^2)$$

so that the density (3) can be rewritten as

$$\left(\prod_{i=1}^n f(y_i \mid \beta_0 + \beta_1 x_i, \sigma^2)\right) f(y_{\text{new}} \mid \beta_0 + \beta_1 x_{\text{new}}, \sigma^2) (|\beta_1| \cdot g_1(\beta_0, \beta_1)) g_2(\sigma^2) \quad (4)$$

Then form (4) is formally a version of form (1) where the (improper) prior assumption on x_{new} is that it is uniform on \mathfrak{R} and the prior on (β_0, β_1) is

$$|\beta_1| \cdot g_1(\beta_0, \beta_1)$$

Notice that relative to the prior $g_1(\beta_0, \beta_1)$, this choice makes values of β_1 with small magnitude much less likely (presumably then combatting the attenuation effects referred to above).