Week 5: Modelling Data

by Michael Escobar

February 8, 2016

1 Introduction

Here we quickly review some of the basic models used in statistics. First, there is a discussion of the basic linear model which includes simple linear regression and the analysis of variance model. The next part discusses some of the specific computing and modelling issues in the bayesian setting. The basic linear model is then greatly extended to models of binary and count outcomes and other models with the general linear models theory (GLM). So, the parts in this hand are the following:

Section 2: The Analysis of Variance and Linear Regression Model.

Section 3: Some quick comments on priors for mean model.

Section 4: Priors for precision parameters for ANOVA models.

Section 5: Some Computing and Modelling Issues for the Bayesian Linear Model.

Section 6: The General Linear Model.

2 Analysis of Variance and Linear Regression Models

2.1 Introduction

The analysis of variance and linear regression model is one of the center pieces of applied statistics. So, first, we look at these models and see how flexible they are.

2.2 Basic regression

The simplest regression is where a variable Y is predicted by a variable X with a simple line. Let y_i be the i-th observation of Y and let x_i be the i-th observation of X. Let the formula for the predicting line of Y given X be:

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

In this equation, we need to find the unknown parameters (β_0, β_1). The error is expressed in the term ϵ. Often this is consider to be a normal distribution with mean zero and unknown, constant variance σ^2. Also, conditioning on the terms β_0, β_1, and x_i, the different observations are considered to be independent and the X's independent of ϵ.

Alternatively, one can say that the expected $Y_i|X_i, \beta_0, \beta_1, \sigma^2$ have independent normal distributions with means $\beta_0 + \beta_1 X_i$ and common variance σ^2.

When there are more than one "X" variable, one can expand the above model with the following equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon.$$

Note matrix notation: $y = X\beta + \epsilon$, where X is a matrix with column vectors: $[1, x_1, \ldots, x_p]$, where the column vector x_j is $(x_{1j}, x_{2j}, \ldots, x_{nj})'$, and where β is the column vector with $\beta' = [\beta_0 \beta_1 \ldots \beta_p]$. Sometimes the matrix X is called the design matrix.

2.3 Categorical variables (Analysis of Variance)

- Basically, one has a factor that is, a categorical variable which has several levels. For example, this could be fixed dose levels of a drug or it could be different types of drugs or treatments. The simplest analysis of variance model, the one-way ANOVA, has only one factor. Say that the factor is A and it has L. The usual assumption is the response variable, $Y_i|\alpha_1, \ldots, \alpha_L$ has expected α_l when the i-th observation has factor level l.

- Note: by using dummy variables, then one can turn this into a linear equation with $L - 1$ dummy variables. In this way, one can convert the simple one-way ANOVA model into a linear regression. To see this, let there be 9 observations total with the first three from the first level, the next three observations from the second level, and the last three from the third level. So, might record the variable A as:

  $$A = (1, 1, 1, 2, 2, 2, 3, 3, 3)'$$

  and so we model Y_i as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_1 \\ \alpha_1 \\ \alpha_2 \\ \alpha_2 \\ \alpha_2 \\ \alpha_3 \\ \alpha_3 \\ \alpha_3 \end{bmatrix} + \epsilon = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \epsilon$$

So, when

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \text{ and } \beta = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix},$$

then the one-way ANOVA is just another example of linear regression. Also, note that the "variable" A got turned into three x's in this example. This is a trick that is repeated below. It should be also noted that the above matrix X is a called the design matrix.

- For 2-way and higher ANOVA, one can use similar trick. However, note that one needs to worry about how to parameterizes the model. This leads to questions of how to write the "design matrix". There are a number of different ways to parameterize this model. If factor A has a levels and factor B has b levels, then their will be a total of $1 + (a - 1) + (b - 1)$ independent columns in the design matrix Xif there are no interaction terms.

- An interaction term can be interpreted two ways. One way to think of it is that a change between two specific levels of factor A are different for different levels of B. Sometimes this is referred to as a multiplicative effect because one way to incorporate these into a design matrix is to "multiple" the dummy variables that are used for the two factors. A problem with this is that the number of columns of X grow at a multiplicative way when adding new factors. That is, when a model has factor A with a levels and factor B with b levels, then the model with interaction terms will have $a * b$ columns in the design matrix.

- In a "designed experiment", then one might use "contrast" to test for differences between different "levels" of the "factors". Often, one would use "orthogonal" contrast to divide up the factors in the model.

- If the levels of the factor really are increments of some continuous scale, then one can use orthogonal polynomials to see what degree of a polynomial should be used to model the effect. Also, from this, one can show that a continuous fit is a model which is just a restriction of the categorical model (where one models each level of the factor separately).

- Note: one can mix categorical and continuous variables. These models are sometimes called analysis of covariance.

- More use of "designed experiments" where one chooses the X values to tease out different factors, etc.

- Note: in statistical research and in statistical courses, the study of more of these tricks is covered in the area of "Design of Experiments".

2.4 More flexible regression

- Sometimes, the effects of the "x" is not linear on the outcome "y". In such cases, it is useful to perhaps add a x^2 term to the model. This adds a "bend" to the fit of the functional relationship between x and y. That is, the model becomes:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

- Now, we can extend the above by noting that this is just the first part of a taylor series and we can (often) approximate a function with a polynomial of high enough order. Therefore, we can approximate the functional relationship between x and y by:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots \beta_p x_i^p + \epsilon_i.$$

  Note, in the above equation, the design matrix, X, has columns which correspond to the the powers of x. So, although our fit in one sense is not longer linear in x, we still have a linear model and this is just another case where the model is: $Y = X\beta + \epsilon$.

- Please note that when using the different powers of x as variables in the model, these variables are often highly correlated. For example if $x_i = i$ (that is, the value of x for the i-th observation is i), then x and x^2 are highly correlated. Similarly for the higher powers. Therefore, one might want to use "orthogonal points". See the experimental design literature.

- Instead of using polynomials (or orthogonal polynomials), one might use a set of different bases functions. For example, "fourier analysis" is one such method where one uses a series of sines and cosines to model the data. This type of model is sometimes used for data which is cyclical.

- Another very popular technique which can be put in this form are spline models. These are "non-parametric" models which fit the data in a more local sense. For an excellent paper which describes how to do this with WinBugs and R, see the paper by Crainiceanu CM, Ruppert D, Wand MP, "Bayesian Analysis for Penalized Spline Regression Using WinBUGS," *J of Statistical Software*, *14*, Issue 14, Oct 2005. In that paper they fit the following model:

$$y_i = m(x_i) + \epsilon_i,$$

where $m(x, \theta)$ is a smooth function of x_i and the function has the parameters θ. The basic idea is to define some "knots" which are places you would like to allow "bends" in your function. Let there be a total of K knots at the points $\kappa_1 < \kappa_2 < \ldots < \kappa_K$. The function that they fit to $m(\cdot)$ is what is called a "thin-plate" spline and has the following form:

$$m(x, \theta) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k |x - \kappa_k|^3,$$

where the parameter vector $\theta = (\beta_0, \beta_1, u_1, \ldots, u_k)'$. So, we see that once again, we are in the linear regression form.

- Comment on the spline model: yes, there are lots of knots. So, in the usual frequentist literature, one then uses a penalty function to control the large number of regression terms. In the above paper, they show how one can changed the penalty function that one might use into a reparameterization of the model.

2.5 Functional predictors

Using "functional predictors": Say that the X values are a series of measurements over time. Then, one could use all the X's in the regression, or one can make summary scores of these values.

- use a simple measure like the mean values.
- use some other measure like the max, mean, or "area under the curve"
- first remake the score by doing something like a factor analysis.
- "functional predictors": fit the series of X's by a polynomial. Then use the coefficients of these polynomials as "predictors". a) use a simple polynomial of some order. b) fit X with an orthogonal polynomial, c) fit with some other basis function such as a spline model. d) note: could simple use these coefficients and treat like data or one can use the two series of regressors in the Bayesian model. Note 1: this is related to growth curve models. Note 2: I have seen these types of analysis come up more in recent conferences, etc.
- More fancy stuff: With the increase in statistical power over the last 10-15 years, there has been more and interesting ideas. Other ideas that are somewhat related include general additive models (GAM). These are somewhat related to neural nets. The computer science/machine learning community has lots of rather clever ideas which are making their way into the statistical community. One idea is called kernelization. Basically, takes several variables and looks at various transforms growing the number of variables in the linear model. Then, they look at the $X'X$ matrix and perform some computer science magic. {I'm still learning the details...} This trick is part of the underlying structure of support vector machines and its relatives.

2.6 Multilevel model

Basically, one wants to break larger models into smaller parts by considering it as a conditionally independent hierarchical model. Then, one can model each model as described above. {possible more on this} {Also, related concepts are longitudinal models and spatial data analysis.}

2.7 Other issues: missing values and variable selection

The issue of variable selection is covered in an upcoming lecture. I don't think I will have time to discuss missing values in this course. Check out some of the examples in the WinBugs manual for some information on how to do this.However, this is something that people will want to consider doing.

2.8 A Final Note On Linear Regression

One can mix up all the above ideas as needed....

3 Some ideas on the Prior structure for mean parameters.

- If the ϵ random variable in the above model is normal, then that will mean that the likelihood function is based on the normal distribution. Therefore, for the basic prior for the β's, one might first consider the normal distribution, which is the conjugate prior. People have considered other choices. When one uses a "double exponential" distribution, it will tend to shrink the value of β closer to zero. (Need to check: but this is similar to the lasso regression...). Another alternative is to use a type of t-distribution for the error distribution.... this will make the regression more resistant to outlying values.

- When one has a series of β's for the values of a categorical variable, then perhaps one might let this set of β's come from a common distribution. For example, it one is testing a series of similar drugs. One might have a set of β's which each represent the effect of each of the individual drugs. Then perhaps one would consider having this set of β's come from a common normal distribution with an unknown mean and precision. With several drugs, then one is in a position to estimate the unknow mean and precision paramters.

- It one has a β which is a "slope" parameter for a continuous measure, then one might use a prior which has a fixed mean and precision. With only one β being sampled from the prior distribution, one does not have multiple values to estimate the mean and precision of the prior distribution for that β parameter.

- if one use the normal distribution, there are still questions as to what distribution to put on the hyperparameters, especially σ^2 which is the variance of the error term ϵ. The choice that is probably the most common, because it is the conjugate prior is to use an inverse gamma. However there are other suggestions such as maybe using a flat prior on σ, σ^2, or perhaps $1/\sigma^2$. This is discussed in the next section.

4 Priors for the Precision Parameter

This document looks at priors for the spread parameter of the normal. The conjugate prior uses a gamma distribution on the precision parameter. As we saw in the second week, one might consider a "low information" prior to be a gamma distribution with a low value for the alpha and perhaps also beta parameter. (That is the, the shape and scale parameter.) In early Winbugs programs, it was common to see one use a Gamma(.001,.001) prior for the precision parameters of normally distributed parameters. However, it is possible that this choice of a seemingly low informative prior may actually have unintended consequences.

First, the mathematics of the problem are explained in detail. Then, some alternative priors are discussed. Then, 2 datasets are randomly generated and fit with the different priors to demonstrate the problem.

4.1 The Basic Problem

Basically, the problem is that if one use a gamma prior for the precision parameter, then even if one sets the scale parameter, alpha, to be zero, the posterior distribution is a gamma distribution with scale parameter $n/2$. However, look at what this implies for the simple 1-way ANOVA model. The 1-way ANOVA model has the following form:

$$\begin{aligned} X_{ij}|\mu_j,\tau &\sim N(\mu_j,\tau) \quad i=1,\ldots,I; j=1,\ldots,J \\ \mu_j|\mu_0,\tau_0 &\sim N(\mu_0,\tau_0) \end{aligned}$$

For priors for μ_0, τ, and τ_0, one might use the following conjugate priors:

$$\begin{aligned} \mu_0|\mu_{00},\tau_{00} &\sim N(\mu_{00},\tau_{00}) \\ \tau|\alpha,\beta &\sim \text{Gamma}(\alpha,\beta) \\ \tau_0|\alpha_0,\beta_0 &\sim \text{Gamma}(\alpha_0,\beta_0) \end{aligned}$$

Now, for the 1-way ANOVA, an important issue is whether all the μ_j's are equal. This is what is often of interest in this model. Let σ_0 equal $1/\sqrt{\tau_0}$. When all the μ_j's are equal, this is equivalent to σ_0 and σ_0^2 equal to zero. This is therefore equivalent to τ_0 to be infinity. Basically, the gamma posterior distribution does not give weight to τ_0 being infinity.

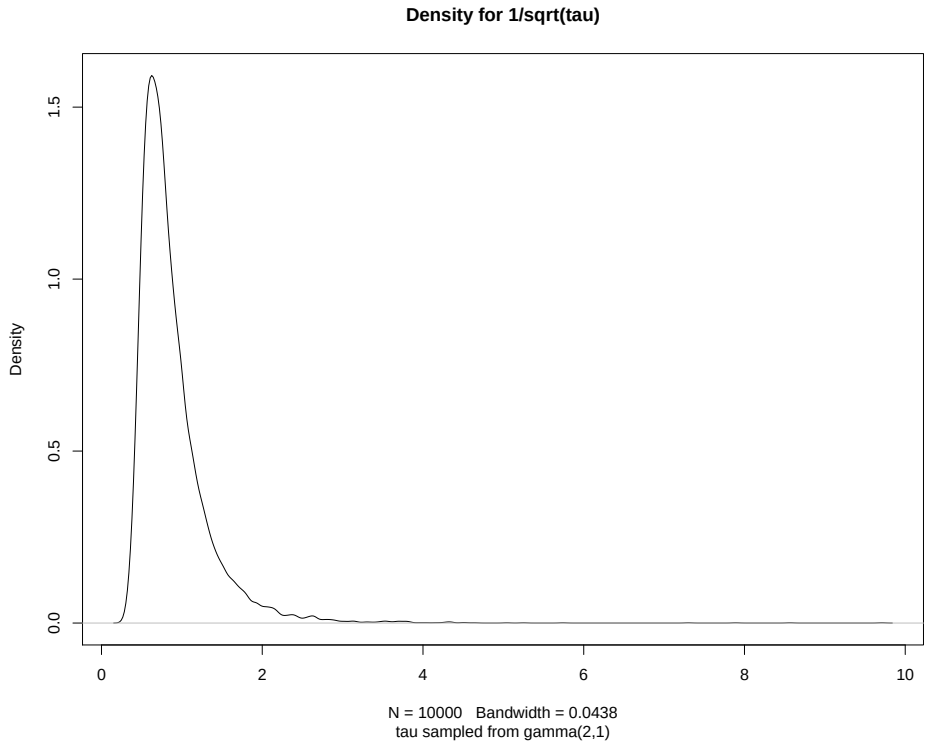

Figure 1: The above plot is generated by first generating 10,000 τ_0 sample from a gamma(2,1). These are converted into samples of σ_0 by setting $\sigma_0 = 1/\sqrt{(}\tau_0)$. The sampled values of σ_0 are then plotted with a kernel density estimator.

To see this, note that when J is 4, then the posterior for τ_0 has a shape parameter, $\alpha_0 + n/2$, of at least 2. Figure 1 gives an estimate of the shape of the posterior distribution for σ_0 when τ_0 is from a gamma with shape parameter 2 and scale 1. Note that there does not appear to be any mass when σ_0 is near zero. Therefore, when a gamma prior is used, it appears to have no mass in the small neighborhood around zero. (Aside: this might be somewhat mitigated when a prior like gamma(.001,.001) is used. Then the β parameter may slide the mass near zero, but the left tail will still decrease to zero right next to zero.)

4.2 Alternative Priors

Since it important to put prior weight on very small values of σ_0 or σ_0^2, one might consider priors where the prior density does not go to zero as σ_0 goes to zero.

First, there are priors one might put directly on σ_0 or σ_0^2. Some popular choices are:

- A uniform prior. Place the lower end of the prior on 0 and the upper end at some place where one is fairly confident (without being too extreme) that the σ_0 or σ_0^2 is smaller.

- A "half normal" prior. To define a half normal, first start with a normal centered at 0. Then, restrict the random variable to be non-negative. So, the normal is truncated to be greater than or equal to zero.

The density of both of these priors do not go to zero as the value of the random variable goes to zero. Therefore, both priors have mass in a neighborhood around zero.

Below, these priors are used in a simple, generated data set and compared to the use of either a gamma prior or a uniform prior on τ_0.

4.3 Generated Data

To compare the effect of these prior on estimates of the between group standard deviation two datasets are generated and then fit with the different priors. In both datasets, there are 5 groups and 10 observations per group. Also, the within group standard deviation, σ, is set to 2 in both datasets. For the first dataset, the group means are all equal and therefore the between group standard deviation,σ_0, is zere. In the second data set, the between group standard deviation is set to 1. If the first data set, the F-statistic which test for equal means is 0.735 (pval=0.57 on 4,45 dfs) and for the second dataset, the F-statistic which test for equal means is 3.0153 (pvalue=0.02753 on 4, 45 df). The boxplots for the two data sets are in figure 2.

Table 1 shows the results from fitting each dataset with the different priors. In this table, the credible regions are given for different priors. The column of numbers under the heading "Data 1" are the credible regions for the data when the means are actually all equal and therefore when σ_0 is equal to 0. Here we see that the 2.5% is equal to zero (within roundoff error) when uniform or half normal prior is used for σ_0. When the priors are placed on the precision parameter τ_0 or the variance parameter σ_0^2, then the credible region appears to be bounded off of the actual value zero. For data 2, where the parameter σ_0 is truly not zero, then all the priors appear to produce credible regions that contain the true value which is that σ_0 equals 1.

4.4 Some Closing Thoughts

- So, from the above, it would appear that putting a uniform or a half normal prior on σ_0 has some advantages. When all the means are equal, it appears to "allow" the posterior to find this. Although, when the means are not all equal, it appears from this data set, that the models where the priors are placed on the τ_0 parameter result in narrower credible regions. (Then again, perhaps they are narrower since they don't have weight on the possibility that τ_0 is infinite.)

- This is only one simulation for one combinination of sampled mean, numbers of samples within group, and for possible values for σ_0. So, this simulation does give some insight as to what is going on, but this is not a full blown "simulation study". Usually, a simulation study would not just simulate one data set (per condition)

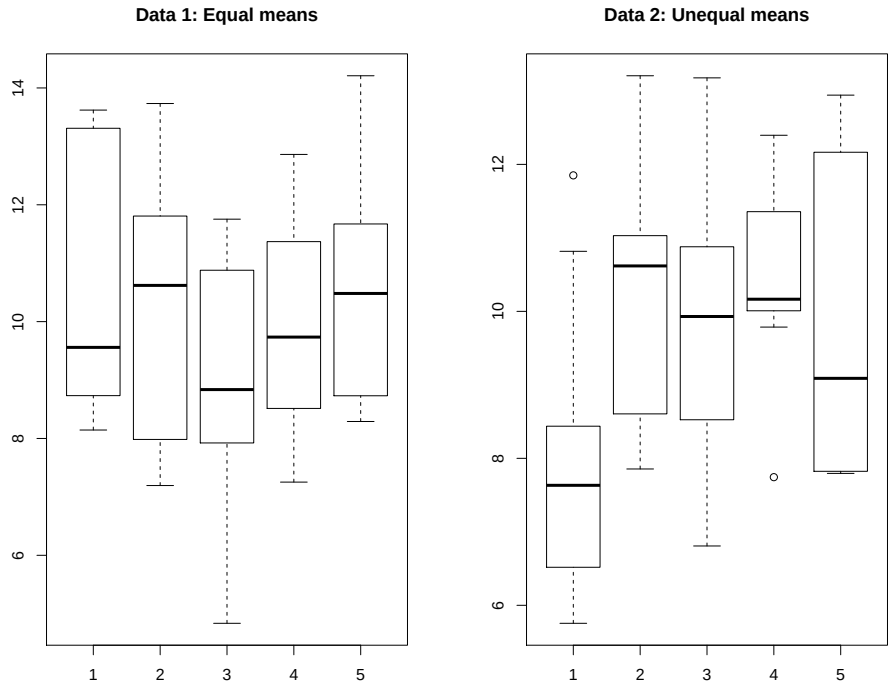

Figure 2: The data set on the left is generated with equal group mean values. In this data, $\sigma_0 = 0$. The data set on the right is generated with different group means. In this data, $\sigma_0 = 1$.

Table 1: Posterior credible regions resulting from different priors for ANOVA

	DATA 1 $\sigma_0 = 0$		Data 2 $\sigma_0 = 1$	
	2.5%	97.5%	2.5%	97.5%
$\tau_0 \sim$				
gamma(.01, .01)	0.1	1.2	0.1	2.4
gamma(1, 1)	0.4	1.8	0.5	2.1
unif(0, 10)	0.3	0.8	0.3	1.6
unif(0, 100)	0.3	0.8	0.3	1.6
$\sigma_0 \sim$				
unif(0, 4)	0.0	2.2	0.2	3.1
unif(0, 10)	0.0	2.5	0.3	4.0
1/2Norm(0, .1)	0.0	2.1	0.2	2.9
1/2Norm(0, .01)	0.0	2.2	0.2	3.7
$\sigma_0^2 \sim$				
unif(0, 16)	0.2	2.5	0.5	3.6
unif(0, 100)	0.2	5.5	0.5	5.8
1/2Norm(0, .01)	0.2	3.0	0.4	3.5
1/2Norm(0, .0001)	1.7	2.6	0.5	6.6

but would simulate many. If one wants to get accuracy of a credible region value to 2 decimal places, one should simulate perhaps 10,000 datasets for a given choice of a) number of groups, b) number of observations within groups, and c) size of σ_0. Also, one should look at different combinations of the those three factors. So, this little study provides some insight, more could be done.

- One might question whether Bayesian should care if the credible region "covers" the "true value". That is, if one has stated one's true prior and one collects data, then the posterior is one's "true" posterior belief. Well, that may be true. However, there is something to be said for having Bayesian procedures which have good "frequentist" properties. (See Rubin (1984)[1] for some discussion of this.)

5 Computing for ANOVA models

5.1 Introduction

The main problem with modelling a 3-way ANOVA (or anything more than a 1-way ANOVA) is that fact that there is an over specification of the parameters in the model. For example, let us consider the typical two way ANOVA with factors A and B which have I and J factor levels respectively. Then, if for the i-th level of A and the j-th level of B, we can model the expected value E_{ij} as:

$$E_{ij} = m_0 + \alpha_i + \beta_j.$$

The problem, of course, with the above parametrization is that we can add some constant, say c, to m_0 and subtract c from each α_i and we get the same model. So, the parameters are unidentified. Because of this, when one runs the MCMC, then the chains for m_0, the α_i's, and the β_j's can wander around aimlessly. So, it is necessary to do something in order to make the model identifiable.

Note, that another way to write the equation is with matrix notation:

$$E = X\gamma,$$

where E is a column vector and it contains the expected values of the k-th observation. The parameter vector γ is defined by:

$$\gamma' = [m_0|\alpha_1 \dots \alpha_I|\beta_1 \dots \beta_J].$$

The matrix X is called the design matrix and it controls the restrictions on the parametrization. Also, one can consider X as a block matrix of the form, $X = [X_m|X_\alpha|X_\beta]$ where X_m is usually a column of ones when the parameter m_0 is in the model and a column of

[1] Rubin,DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics, 12(4), 1151-1172.

zeroes when one wants to remove the parameter m_0. The matrices X_α and X_β specify how one parameterizes the α_i's and β_j's. When constructing the matrices X_α and X_β from the template matrices Z_α and Z_β. Given Z_α, define the k-th row of X_α as equal to the i'-th row of Z_α when the k-th observation is from the i'-th level of factor A. Similarly, define the rows of X_β from the rows of Z_β.

In the different models that I have included, I have considered three different ways to deal with the unidentifiability.

1. The first way was to set the last factor level equal to the negative of the sum of the other terms in the model and to then put a prior distribution on m_0 and the other levels of each factor. This is equivalent to having a design matrix for the factors where the last column of Z_α and Z_β are columns of -1's. The file which contains the WinBugs code for this model is "dimondBMod.txt".

2. Another way that I used to model a restriction on the parameters is to use the design matrix based on an orthonormal matrix. Basically, for Z_α one uses an orthonormal matrix which is orthogonal to the column vector of ones to be the first $I-1$ columns of Z_α (and sets the last column to zeroes). Similarly one defines Z_β. One can get these matrices by using, for example, the R command: `contr.poly( )`. This creates the matrix of polynomial contrasts. If the data is balanced, then this will produce parameters which are uncorrelated. In this example, the factor effects are no the values α_i or β_j since these parameters are present with different weights in all the factor levels, so one needs to do a conversion to get the factor levels. You can see the WinBugs code in the file: "diamondMod.txt".

3. Another way of doing this is to use a trick of data-augmentation. This trick I first learned from the new book by Gelman and Hill (Reference: Gelman and Hill, 2007, *Data Analysis using Regression and Multilevel/Hierarchical Models*,(New York: Cambridge University Press), pg 419-434). There are two conceptual parts of the data-augmentation trick. In the first part, one simply samples the parameters m_0, the α_i's, and the β_j's with out any restriction at all. So, yes, these parameters are unidentifiable and if one looks at various convergence diagnostics one finds that these parameters will wander all over the place. For the second part of this trick, one defines identifiable parameters from the unidentified one. Define:

$$\begin{aligned} m_0^* &= m_0 + \frac{1}{I}\sum_{i=1}^{I}\alpha_i + \frac{1}{J}\sum_{j=1}^{J}\beta_j, \\ \alpha_{i'}^* &= \alpha_{i'} - \frac{1}{I}\sum_{i=1}^{I}\alpha_i \quad \text{where } i' \in \{1,\ldots,I\}, \text{ and} \\ \beta_{j'}^* &= \beta_{j'} - \frac{1}{J}\sum_{j=1}^{J}\beta_j \quad \text{where } j' \in \{1,\ldots,J\}. \end{aligned}$$

From the above, note that m_0^*, the α_i^*'s, and the β_j^*'s are identifiably. We looking at convergence criteria for the chain, one ignores the unidentifiable parameters and just looks at the identifiable parameters. You can see the WinBugs code in the file: "diamondBRedMod.txt".

When I had worked with smaller datasets (and with a larger PC), I usually used either parameterizations using methods 1 or 2 above. In those cases, I found usually found that using the method with an orthonormal matrix (technique 2) resulting in an algorithm which converged in very few steps and was fairly fast. However, when I set up both of these techniques for a very large data set, I found that WinBugs took a painfully long time to perform each iteration of the MCMC. (Someone suggested that possible my machine did not have enough memory so I was potentially having problems with memory swapping.) So, I looked around for another way to do things and I came across the data-augmentation method. This algorithm iterated much faster than the other two methods and it seemed to converge in very few steps also.

6 General Linear Models

6.1 Overview of GLM

The General Linear Model (GLM) extends the usual linear regression/analysis of variance model. In these models, there is a measure of some outcome values and some explanitory variables which are measured, which are thought to affect the outcome values. So, these models have the basic look of a regression model. It is useful for a wide variety of models such as models where the outcomes are binary values or counts. The likelihood side of this model has been well studied in the frequentist literature. So, first, the likelihood portion of this model is discussed.

There are three basic parts of the (likelihood portion of this) model:

- Random/Stochastic component.
- Systematic component
- Link function.

6.1.1 Random Component

For a sample of size N with outcome $Y_i (i = 1, \ldots, N)$ and assume that conditioning on a parameter θ, then $[Y_i|\theta_i]$ has some distributional form with parameter θ_i. For example, if Y_i is equal to one or zero (a success or a failure), we might consider $Y_i|\theta_i$ to be a sample from a bernoulli distribution with parameter θ_i. As another example, suppose that the value Y_i is a count of events, then we might consider $Y_i|\theta_i$ to be from a poisson distribution with θ_i as the parameter for the mean. Usually, the class of distribution considered for the random component belong to the one parameter exponential family.

6.1.2 Systematic Component

This component describe the relationship of the explanatory variable to the rest of the model. Suppose that for each observation i, we have k variables, $X_{i1}, \ldots, X_{ik}$. For the general linear model, we assume that the effect of the explanitory variables can be described as a linear function. That is, the effect of the X's on the outcome Y_i is of the form:

$$\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}.$$

Please note, that this part of the model is the same as for the linear regression/anova type of model. One can use the same types of structures one used there for these models.

6.1.3 Link Function

The purpose of the link function is to link the random component to the systematic component. This is done by modelling a function of the mean, $E(Y_i|\theta_i)$, by the systematic component. That is,

- Let $\mu_i = E(Y_i|\theta_i)$.
- Let $g(\cdot)$ be the link function.
- Model μ_i by the systematic component by the the equation:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}.$$

6.1.4 Basic message

The main difference between these models and the linear regression/anova models is with the link function and the random component. However, the systematic component is the same as the linear regression/anova models. Frequentist theory also tells us that the distribution of the β parameter is approximately the same as for the linear regression/anova models. There are some caveats. One, the random component must be from the exponential family. As it turns out, members of this family include the binomial, poisson, and multinomial family is what these models are usually used for. Also, there are some minor details like overdispersion and offsets that you might need to worry about.

6.2 Some Details

6.2.1 Some Common Link Functions and Random Components

The following are some common link functions used with some of the common random components:

- When Y_i takes on values of zero or one, than one might consider the binomial distribution or the bernoulli distribution for the random component. Two common link functions used when the bernoulli/binomial random component is used are the logit and the probit link. That is:

$$\begin{aligned} \text{logit:} \quad & \log\left(\frac{\mu}{1-u}\right) \\ \text{probit:} \quad & \Phi^{-1}(\mu), \end{aligned}$$

  where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the normal distribution. When the logit function is used, this analysis is sometimes called a logistic regression, and when the probit function is used, this analysis is sometimes called a probit regression.

- When the outcome variable Y_i is number of counts of some events, then the it is common to use a Poisson distribution as the random component. For this model, the log function is often used as the link function. Sometimes this analysis is referred to as a Poisson regression. Also, when this model is used to analyse data from a contingency table, the model is sometimes referred to as log-linear model.

  Also, note that sometimes the identity link function is with the Poisson random component. The identity link function is simply the function that maps a value to itself. That is, $g(\mu) = \mu$.

- When Y_i takes on continuous values, then one might consider using the normal distribution for the random component. Note that one might want to transform the Y_i value with some function, such as a power function or the log function, so that the residual look normal. The identity link is commonly used with the normal random component.

- Please note, that one needs to check to see if the random component, the systematic component, and the link function result in a model which is indeed modelling the data. The above suggested forms for the random component and link functions are often used for the data which is described, but they need to be checked. Please see reference on categorical data analysis or general linear linear models for more details.

6.2.2 Frequentist Inference

Using maximum likelihood theory, under some basic regularity conditions, it can be shown that the joint distribution of the maximum likelihood estimators is asymptotically normal. Therefore, one can do approximate frequentist inference

6.2.3 Common Bayesian model for GLM

The random component defines the likelihood part of the Bayesian model. The link function is used to link the mean of the predictor variables in a linear function. Therefore, one needs to put priors on the β's. Since the MLE is approximately normal, it seems reasonable to use normal priors for the β's. Please note that this is not a conjugate prior set up. Also, getting the conditional distribution and getting the algorithm to sample the different parameters is somewhat complicated. However, we let WinBugs worry about the algorithm for now.

6.2.4 Overdispersion and the Bayes Approach

Note, that for the the binomial and Poisson models, there is only one parameter, so this one parameter specifies both the mean and variance of the distribution. However, if one models the mean parameter, sometimes the variance of data is does not have the correct variance. For example, for the Poisson model, the mean and the variance are equal. However, the data might not have the mean equal to the variance. This is often referred to as being overdisperse (if the variance is too large) or under disperse (if the variance is too small). When there is overdispersion in the model, one can model this in the Bayesian model, then one can include the a "random effect" term.

6.2.5 offset

This is just used in some models. This is discussed for the poisson model.