`breakdown` — An Example of a Normal Hierarchical Model

This example describes the use of a normal hierarchical model to combine information in a set of "breakdown" baseball statistics (Albert, 1994). Consider the group of 154 regular major league baseball players during the 1992 season. For each player, the number of hitting opportunities (at-bats) and number of hits is recorded for various situations. For example, the datafile `ahead.dat` records the hitting performance of each player during the pitch counts "ahead in count" and "two strikes". The first row of this file (shown below) contains the values 134, 49, 248, 62. This particular ballplayer (Roberto Alomar) had 49 hits out of 134 at-bats when he was ahead in the count (pitch counts of 1-0, 2-0, 3-0, 2-1, and 3-1), and had 62 hits in 248 chances when there were two strikes in the count (pitch counts of 0-2, 1-2, 2-2, and 3-2).

```
134    49   248    62
158    58   290    65
 81    24   222    35
164    61   264    68
158    53   264    61
123    36   171    33
110    36   175    41
152    52   276    53
132    39   305    66
124    40   252    48
. . .
```

The objective is to simultaneously estimate the hitting probabilities of the 154 players under the two situations. For the pitch count situation, let p_{i1} denote the hitting probability of player i in the "ahead in count" situation and p_{i2} denote the hitting probability of this player when the pitch count is "two strikes". We observe x_{ij} hits in n_{ij} at-bats of the ith player in the jth situation. We will focus on learning about the difference in situation probabilities $p_{i2} - p_{i1}$ for all players.

We first apply a reexpression to make the observed data approximately normally distributed. Define the observed logit $y_{ij} = \log[x_{ij}/(n_{ij} - x_{ij})]$ which is approximately normally distributed with mean $\mu_{ij} = \log[p_{ij}/(1 - p_{ij})]$ and variance $\sigma_{ij}^2 = 1/[n_{ij}p_{ij}(1 - p_{ij})]$. We can represent the data for the $N = 154$ players under the two situations as the matrix

	Player			
	1	2	. . .	N
ahead in count	y_{11}	y_{21}	. . .	y_{N1}
two strikes	y_{12}	y_{22}	. . .	y_{N2}

We estimate the variances $\{\sigma_{ij}^2\}$ by substituting the sample proportions for the probabilities in the formula. Since the sample sizes $\{n_{ij}\}$ are large, we regard these sampling variances as known.

The mean logits for the ith player are modeled as $\mu_{i1} = \mu_i + \alpha_i$ and $\mu_{i2} = \mu_i - \alpha_i$, where μ_i measures the hitting ability of player i, and α_i represents the situational effect that measures the change in the player's hitting ability due to the pitch count. The hitting abilities $\{\mu_i\}$ are assigned a flat prior. To reflect the belief that the situational effects across players are exchangeable, we assign the $\{\alpha_i\}$ the two-stage normal hierarchical prior:

1. α_i distributed $N(\mu_\alpha, \sigma_\alpha^2)$, $i = 1, ..., 154$

2. μ_α distributed uniform, σ_α^2 distributed Inverse Gamma(a, b).

In this prior, the only hyperparameter values to specify are the Inverse Gamma parameters a and b.

This model can be fit to the pitch count situational data using two Matlab commands. The data is stored in the text file `ahead.dat`. The command

```
[p1,p2,y,s2]=setup('ahead');
```

will compute the observed proportions for each situation and place them in the columns `p1` and `p2`. In addition, this function will compute the observed logits $\{y_{ij}\}$ and the estimated variances $\{\sigma_{ij}^2\}$ and store them in the columns `y` and `s2`. The command

```
[Mp1,Mp2]=gibbs(y,s2,hpar,m);
```

has four inputs: the observed logits and associated sampling variances `y`, `s2`, the hyperparameters of the prior variance, a, b, stored as a two element vector `hpar`, and the number of iterations of the simulation algorithm `m`. A Gibbs sampling algorithm is used to sample from the joint posterior distribution of $(\{\mu_i\}, \{\alpha_i\}, \mu_\alpha, \sigma_\alpha^2)$. Simulated values of the hitting probabilities $\{p_{i1}, p_{i2}\}$ are generated from this algorithm and stored in the Matlab columns `Mp1` and `Mp2`.

In this example, the hyperparameters were set to the relatively uninformative values $a = b = .5$ and $m = 10000$ iterations were used. The Matlab plot command

```
plot(p2-p1,mean(Mp2-Mp1))
```

produces a scatterplot (shown below) of the observed difference in hitting proportions $\{x_{i2}/n_i - x_{i1}/n_i\}$ against the posterior means of $\{p_{i2} - p_{i1}\}$.

In the figure, the dotted line $y = x$ and the dashed horizontal line $y = -.123$ have been drawn. The diagonal line represents estimates that are identical to the observed proportions and the horizontal line represents estimates that are shrunk entirely toward the average effect across all players. This graph demonstrates that the observed differences in hitting proportions between the "two strikes" and "ahead in count" situations are shrunk approximately one half of the way towards the average difference of .123 of all players.

The table below presents the observed season batting averages and the posterior means of the respective probabilities for four players. Jose Canseco appeared to have a large situational effect for the 1992 season. He batted for an average of .358 when ahead in the count and only .133 when there were two strikes, for a season difference of $-.225$. Using the exchangeable model, the posterior mean estimates of the underlying batting probabilities are .324 and .151 and the difference in these estimates is $-.173$. Similarly, at the other extreme, Tony Gwynn actually batted better in 1992 under two strikes than ahead in the count, for an observed difference of .0390. The exchangeable model estimate at the difference in hitting probabilities is $-.0370$, indicating that Gwynn actually was a better hitter when ahead in the count.

PLAYER	SEASON			POSTERIOR		
	Ahead in Count	Two Strikes	Diff.	Ahead in Count	Two Strikes	Diff.
Buhner	.317	.148	-.169	.299	.157	-.142
Canseco	.358	.133	-.225	.3240	.151	-.173
Gwynn	.252	.291	.039	.298	.262	-.037
Hayes	.337	.168	-.169	.321	.176	-.146

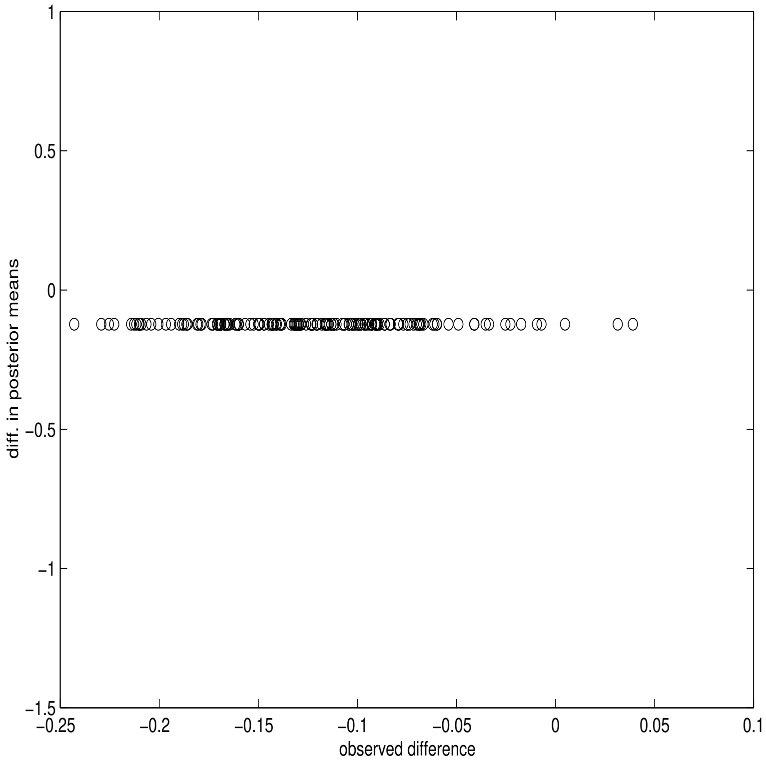

Figure 1: Plot of posterior mean of the difference in hitting probabilities against the observed batting average difference for all players.

In the subdirectory `hierarch`, there are datasets corresponding to other situational variables described in Albert (1994). These variables including batting against pitchers of the opposite side versus the same side, batting against groundball versus flyball pitchers, batting home versus away, day versus night, on grass versus artificial turf, batting when the runners are in scoring position versus none on/out, and batting before and after the All Star game.