Final Exam Takehome Portion / Last HW Assignment
SOLUTIONS

Background (from the assignment sheet)

Lillard & Panis (2000) discuss data on the decisions of 501 mothers to deliver babies in a hospital vs. at home or elsewhere. The mothers have varying numbers of children, ranging from 1 to 10, and make separate decisions about where to deliver each child. There are a total of 1060 births in the data set. The available variables are

```
'data.frame':   1060 obs. of  6 variables:
 $ hospital: int  0 0 1 0...   1 = hospital birth, 0 = birth elsewhere
 $ loginc  : num  4.33 5.62... logarithm of family income (log dollars)
 $ distance: num  1.7 7.9...   distance (miles) from nearest hospital
 $ dropout : int  0 0 0 0 0... 1 = mother did not complete hs, 0 = completed hs
 $ college : int  1 0 0 0 0... 1 = mother attended college, 0 = did not
 $ mom     : int  1 2 2 2 2... unique identifier for each mother
```

Note that family income varies from the birth of one child to the next, hence family income is recorded for each child, rather than once only for each mother. Note also that if both `dropout` and `college` are zero, then the mother completed high school but did not go on to college. The mother's `group` identifier appears once for each of her children; thus, the number of children per mother could be obtained as `n.kids <- table(mom)`.

The data are available in the file `hosp.txt` under a link for the takehome portion of the final, at `http://www.stat.cmu.edu/~brian/463`.

Exercises

1. A simple model (perhaps good, perhaps bad) would be to allow the probability of birth to depend on family income at the time of birth, with a random intercept for each mother. As a multilevel model, this is

   Level 1:

   $$\text{logit}(P[y_i = 1]) = \alpha_{0j[i]} + \beta_1 x_i, \quad i = 1, \ldots, 1060$$

   Level 2:

   $$\alpha_{0j} = \beta_0 + \eta_j, \quad \eta_j \sim N(0, \tau^2), \quad j = 1, \ldots, 501$$

   where $y_i = 1$ if the i^{th} child was born in a hospital, $j[i]$ is the mom on the i^{th} child, and x_i is the log(income) for the family at the time of that child's birth. Equivalently in `lmer()`'s modeling language

   ```
   lmer.inc <- lmer(hospital ~ loginc + (1|mom), data=hosp, family=binomial)
   ```

(a) Write this as a hierarchical Bayes model, adding prior distributions wherever needed.

Here is one acceptable way to do it:

Likelihood:

$$\begin{aligned} y_i &\sim Bernoulli(p_i) \\ p_i &= \exp(\alpha_{0j[i]} + \beta_1 x_i)/(1 + \exp(\alpha_{0j[i]} + \beta_1 x_i)), \quad i = 1, \ldots, 1060 \end{aligned}$$

Priors:

$$\begin{aligned} \alpha_{0j} &\sim N(\beta_0, \tau^2), \quad j = 1, \ldots, 501 \\ \beta_0 &\sim N(0, 1000) \\ \beta_1 &\sim N(0, 1000) \\ \tau^2 &\sim Gamma(0.01, 0.01) \end{aligned}$$

Notes:

- *I will write Normal distributions parametrized by mean and variance unless otherwise noted.*
- *Any relatively flat prior for β_0 and β_1 is fine, since they are fixed effects (fixed effects get flat priors in Bayesian models; random effects get priors with a variance that you estimate as part of the model.*
- *Many different prior densitites are possible for τ^2. A Gamma with small α and β (as I have written above) is common. Another common possibility is a Uniform distribution on a large positive interval (e.g. Unif(0,100) or Unif(0,10), etc.).*

(b) Write this as s variance components model.

Substituting the level-2 model into the level-1 model we get

$$\begin{aligned} y_i &\sim Ber(p_i), \ with \\ logit(p_i) &= \beta_0 + \beta_1 x_i + \eta_{j[i]} \end{aligned}$$

where $i = 1, \ldots, 1060$, $j[i]$ is the mother ($j = 1, \ldots, 501$) of the i^{th} child, and $\eta_j \sim N(0, \tau^2)$. Note that since this is not a Bayesian model we don't have to specify prior distributions.

2. In the file `2012-final-takehome-rcode.r` in the takehome final area on `http://www.stat.cmu.edu/~brian`, there is R code to fit two different WinBUGS models to the hosp.txt data:

   - model.00 and rube.00
   - model.01 and rube.01

(a) Fit each model and inspect the output and diagnostic graphs. Is there anything remarkable to point out, or is the MCMC algorithm working well all parameters in each model? Write a few sentences about your findings, illustrated with appropriate graphs or numerical output.

Here is the result of fitting model.00:

```
> rube.00
Rube Results:
Run at 2012-12-03 17:29 and taking 16.19 secs
            mean      sd MCMCerr    2.5%     25%     50%     75%   97.5% Rhat n.eff
b0[1]    -25.579 18.3045  0.9967 -68.270 -36.172 -21.470 -11.418  -2.182 1.00  1000
b0[2]     -3.112  1.1913  0.0778  -5.451  -3.896  -3.119  -2.304  -0.905 1.00   750
b0[3]     24.137 19.4425  0.8678  -1.568   8.775  20.245  35.425  70.037 1.00  1000
b0[4]     -2.332  1.2916  0.0780  -4.783  -3.137  -2.375  -1.524   0.245 1.01   460
b0[5]     24.893 18.6502  0.8167  -0.206  10.395  21.525  35.940  68.060 1.00  1000
b0[6]    -26.558 19.7207  1.0950 -74.437 -38.112 -22.330 -11.387  -2.255 1.00  1000
b0[7]    -26.507 17.5909  0.7746 -69.023 -36.922 -21.610 -13.145  -4.952 1.00  1000
b0[8]    -28.430 18.3735  0.8906 -70.039 -38.690 -24.155 -13.887  -5.183 1.00   410
b0[9]     24.285 19.3886  0.8304  -2.221   8.990  20.705  35.795  69.804 1.00  1000
b0[10]    22.941 20.2336  0.8697  -2.797   7.249  18.590  33.135  71.885 1.00  1000
b1         0.484  0.0871  0.0106   0.304   0.433   0.486   0.545   0.640 1.07    54
deviance 579.534 19.6944  0.8968 544.300 566.325 578.100 591.900 622.185 1.01   140

DIC = 730.135
```

This model has a different fixed-effect intercept (flat normal prior!) for each mom, and a single fixed-effect slope on `loginc` *(again, a flat normal prior). There are 501 fixed-effect intercepts, of which only 10 are printed here.*

We see that although all of the parameters in this report seem to have converged Markov chains ($\hat{R} \leq 1.1$), very few of the visible fixed effect intercepts are significantly different from zero. We can explore this furtheer with the p3() command.

Exploration with the `p3()` *command does not suggest any serious problems with estimating any of the parameters. One thing we do notice is that, although most of the intercepts are not significantly different from zero, their CI's lie almost entirely on one side of zero or the other. These are log-odds, so a value less than zero corresponds to a baseline probability less than 1/2 to use a hospital for birth, and a value greater than zero corresponds to a baseline probability greater than 1/2. This suggests that there may be another factor (in this data, or possibly not collected by the experimenters) that would help us to predict when the baseline*[1] *probability of using a hospital for birth is less than or greater than 1/2.*

Here is the result of fitting model.01:

[1] Note that this is a baseline for loginc=0, which corresponds to an income of $1. The intercept might be much more interpretable if we had first centered log-income.

```
> rube.01
Rube Results:
Run at 2012-12-03 17:49 and taking 15.54 secs
             mean      sd MCMCerr     2.5%     25%      50%      75%    97.5% Rhat n.eff
b0[1]      -0.775  1.4223 0.05225  -3.6984  -1.652   -0.734    0.236    1.817 1.01   330
b0[2]       0.183  0.8714 0.03576  -1.6428  -0.388    0.180    0.757    1.824 1.00  1000
b0[3]       1.137  1.2605 0.05776  -1.3202   0.319    1.120    1.927    3.724 1.01   150
b0[4]       0.374  0.9460 0.04161  -1.5740  -0.221    0.412    0.990    2.142 1.00   760
b0[5]       2.081  1.1203 0.06041   0.0492   1.304    2.039    2.782    4.317 1.03    76
b0[6]      -0.670  1.3277 0.05161  -3.3443  -1.572   -0.600    0.201    1.786 1.01   270
b0[7]      -1.025  1.3188 0.06584  -3.7811  -1.861   -1.045   -0.175    1.450 1.01   150
b0[8]      -1.181  1.1918 0.05070  -3.7607  -1.927   -1.073   -0.367    0.894 1.01   280
b0[9]       1.173  1.3913 0.07488  -1.4248   0.257    1.085    1.970    4.385 1.00  1000
b0[10]      1.078  1.3480 0.05411  -1.4186   0.175    1.063    1.921    3.765 1.01   170
b1         -0.140  0.0185 0.00124  -0.1763  -0.152   -0.139   -0.127   -0.103 1.02   130
tau2        2.554  0.6417 3.62400   1.4060   2.116    2.516    2.916    3.983 1.18    16
deviance 1011.806 38.4398 0.07188 936.4100 986.225 1011.000 1036.000 1088.000 1.08    32

DIC = 1223.61
```

This model replaces the fixed intercepts for each mom with a random intercept (the `b0`*'s now have a normal prior with a variance* τ^2 *that we are estimating from the data, rather than the flat priors in* `model.00`*), and sa single fixed-effect slope on* `loginc`*. Again all of the intercepts seem to be converged, and so does the slope. The magnitudes of the intercepts are much smaller than in* `rube.00` *and the CI's are more nearly centered on zero; this may be in part because the prior density is not very flat and is centered at zero, so the model is going to prefer intercepts closer to zero overall.*

Exploration with the `p3()` *command does not suggest any serious problems with estimating any of the parameters. It is notable that* τ^2 *is not completely converged (*$\hat{R} \approx 1.18$*).*

A couple of troubling things to note: the slope on income is significantly different from zero in both models, but of opposite sign: positive in `rube.00` *and negative in* `rube.01`*; and as indicated above the magnitudes of the random intercepts in* `rube.01` *are much smaller than for the fixed intercepts in* `model.00`*. It's hard to know what to make of this without a lot more exploration of the data and model fits.*

(b) Write model.01 as a hierarchical Bayes model. Find a `glm()` or `lmer()` model that is more or less equivalent to model.01, and fit it. Submit the model you found, the output of `summary()` or `display()` on the fitted model, and a sentence or two comparing parameters estimates between your models and rube.01.

`model.01` *is almost the same as the random-intercepts model considered in problem 1, except, that no fixed effect intercept is estimated. So the model would be*

```
>  lmer.inc.0 <- lmer(hospital ~ loginc + (1|mom) -1, data=hosp, family=binomial)
```

```
> display(lmer.inc.0)
glmer(formula = hospital ~ loginc + (1 | mom) - 1, data = hosp,
    family = binomial)
coef.est  coef.se
   -0.14     0.02

Error terms:
 Groups   Name        Std.Dev.
 mom      (Intercept) 1.23
 Residual             1.00
---
number of obs: 1060, groups: mom, 501
AIC = 1294.5, DIC = 1290.5
deviance = 1290.5
```

The lmer estimate of τ^2 is $1.23^2 = 1.51$, a little smaller than the rube estimate. The slope on loginc is -0.14, same as the rube estimate.

You were not required to do this for the problem, but if we had instead fitted exactly the random-intercepts model considered in problem 1, we would get

```
> lmer.inc <- lmer(hospital ~ loginc + (1|mom), data=hosp, family=binomial)
> display(lmer.inc)
glmer(formula = hospital ~ loginc + (1 | mom), data = hosp, family = binomial)
            coef.est coef.se
(Intercept) -4.44     0.42
loginc       0.53     0.06

Error terms:
 Groups   Name        Std.Dev.
 mom      (Intercept) 1.46
 Residual             1.00
---
number of obs: 1060, groups: mom, 501
AIC = 1174, DIC = 1168
deviance = 1168.0
```

This makes a little more sense and is a little more consistent with the results of rube.00. The two lmer fits together suggest that model.01 should have estimated a fixed effect intercept as well as the random intercepts.

(c) Write model.00 as a hierarchical Bayes model. Find a `glm()` or `lmer()` model that is more or less equivalent to model.00, and fit it. Submit the model you found, the output of `summary()` or `display()` on the fitted model, and a sentence or two comparing parameters estimates between your models and rube.00.

model.00 is an all-fixed effects model, that can be fitted using glm:

```
> fmom <- as.factor(hosp$mom)
```

```
> glm.inc <- glm(hospital ~ fmom - 1 + loginc, data=hosp,family=binomial)
> display(glm.inc)
glm(formula = hospital ~ fmom - 1 + loginc, family = binomial,
    data = hosp)
        coef.est coef.se
fmom1     -23.53 17730.37
fmom2      -4.11     1.14
fmom3      16.95 17730.37
fmom4      -3.20     1.13
fmom5      16.00 10010.33
fmom6     -23.04 17730.37
fmom7     -25.50 12435.47
fmom8     -25.31 10086.76
fmom9      16.68 17730.37
fmom10     16.37 17730.37
fmom11    -23.75 12486.19
fmom12    -23.17 17730.37
  .          .        .
  .          .        .
  .          .        .
fmom497   -25.34 17730.37
fmom498   -23.89 12536.83
fmom499    -4.08     1.55
fmom500    15.85 17730.37
fmom501    15.31 17730.37
loginc      0.69     0.11
---
  n = 1060, k = 502
  residual deviance = 420.6, null deviance = 1469.5 (difference = 1048.8)
```

The estimates of the fixed mom intercepts are comparable in sign and magnitude to the estimates from rube.00. The estimate of the slope on loginc is also similar, although a bit larger for glm than rube.

One difference is that the standard errors tend to be much larger in glm() than rube. This is in part because prior distributions contribute some information about where we think the parameter values are, and hence reduce standard errors.

(d) Using only the fitted model objects rube.00 and rube.01, and any numerical or graphical output you can derive from them, which model fits the data better? Why? [Advice: Don't kill yourself doing lots of different things to answer this question!]

The overwhelming evidence from DIC is to prefer model.00, the fixed-intercepts model.

However, as alluded to above, the problem may just be that we should have put a fixed intercept (as well as random intercepts) in model.01!

3. Assume the children are listed in birth order for each mother, in the data set. We can obtain the birth order as follows:

   ```
   b.ord <- unlist(lapply(split(mom,mom),function(x){1:length(x)}))
   ```

   Do mothers tend to be more likely to have hospital births with each successive child?

   (a) Answer this question using `glm()` models that relate the probability of hospital birth to the `b.ord` variable. Feel free to add other covariates from `hosp.txt` if they make for a better model.

   The coefficient of `b.ord` *in the simplest model*

   ```
   > display(
   +   glm.00 <- glm(hospital ~ b.ord, data=hosp, family=binomial)
   + )
   glm(formula = hospital ~ b.ord, family = binomial, data = hosp)
               coef.est coef.se
   (Intercept) -0.76     0.12
   b.ord       -0.05     0.05
   ---
     n = 1060, k = 2
     residual deviance = 1288.8, null deviance = 1289.9 (difference = 1.1)
   ```

   is not significantly different from zero (so birth order does not affect hospital usage). A little crude variable selection

   ```
   > glm.03 <- glm(hospital ~ (b.ord + . - mom)^2, data=hosp, family=binomial)
   > display(
   +   glm.AIC.2 <- stepAIC(glm.00,scope=list(lower=formula(glm.00),
   +                                 upper=formula(glm.03)))
   + )
   glm(formula = hospital ~ b.ord + dropout + loginc + college +
       distance + dropout:distance + college:distance, family = binomial,
       data = hosp)
                    coef.est coef.se
   (Intercept)      -2.60     0.40
   b.ord            -0.02     0.05
   dropout          -1.94     0.27
   loginc            0.46     0.06
   college           1.39     0.47
   distance         -0.10     0.04
   dropout:distance  0.10     0.06
   ```

```
college:distance -0.14     0.09
---
  n = 1060, k = 8
  residual deviance = 1067.2, null deviance = 1289.9 (difference = 222.7)
```

provides a model that fits the data much better, but also does not show a coefficient of `b.ord` *that is significantly different form zero.*

(b) Does your model fit better if you allow a random intercept or slope, or both? Try to assess this, using `lmer()` and related tools (not `WinBUGS`).

Here are all three models (rand int, rand slope, and rand both), and a comparison with the best glm model:

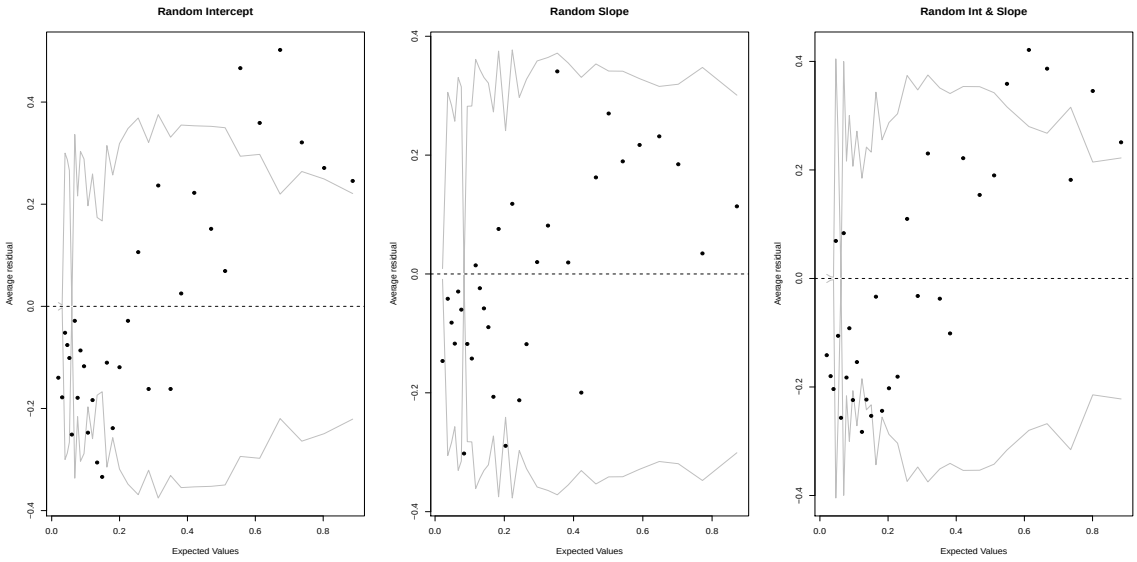
```
> formula(glm.AIC.2)
hospital ~ b.ord + dropout + loginc + college + distance + dropout:distance +
    college:distance
> lmer.inter <- lmer(hospital ~ b.ord + dropout + loginc + college +
+                    distance + dropout:distance + college:distance +
+                    (1|mom),
+                    data=hosp, family=binomial)
>
> lmer.slope <- lmer(hospital ~ b.ord + dropout + loginc + college +
+                    distance + dropout:distance + college:distance +
+                    (b.ord - 1|mom),
+                    data=hosp, family=binomial)
>
> lmer.both <- lmer(hospital ~ b.ord + dropout + loginc + college +
+                    distance + dropout:distance + college:distance +
+                    (b.ord|mom),
+                    data=hosp, family=binomial)
>
> anova(lmer.inter,lmer.slope,lmer.both)
Data: hosp
Models:
lmer.inter: hospital ~ b.ord + dropout + loginc + college + distance + dropout:distanc
lmer.inter:     college:distance + (1 | mom)
lmer.slope: hospital ~ b.ord + dropout + loginc + college + distance + dropout:distanc
lmer.slope:     college:distance + (b.ord - 1 | mom)
lmer.both: hospital ~ b.ord + dropout + loginc + college + distance + dropout:distance
lmer.both:     college:distance + (b.ord | mom)
           Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
lmer.inter  9 1059.7 1104.4 -520.84
lmer.slope  9 1066.8 1111.5 -524.43 0.0000      0    1.00000
```

```
lmer.both  11 1063.3 1118.0 -520.68 7.5021      2    0.02349 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> AIC(glm.AIC.2)
[1] 1083.159
>
> -deviance(glm.AIC.2)/2
[1] -533.5795
```

Clearly, all three lmer models fit better than the corresponding glm model, and the random intercept model fits best, in terms of AIC.

Checking residual plots for the conditional residuals suggests a somewhat different answer:

```
> par(mfrow=c(1,3))
> binnedplot(fitted(lmer.inter),residuals(lmer.inter),main="Random Intercept")
> binnedplot(fitted(lmer.slope),residuals(lmer.slope),main="Random Slope")
> binnedplot(fitted(lmer.both),residuals(lmer.both),main="Random Int & Slope")
```



The conditional residuals show a modest preference for the random slopes only model, even though this model is the least good random effects model according to AIC.

(c) Does the answer to the question change if you use an `lmer()` model from part (b) rather than a `glm()` model from part (a)? Explain.

Let's look at the fixed effect estimates for all three models:

```
> round(summary(lmer.inter)@coefs,2)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.19       0.47   -6.82     0.00
b.ord              -0.04       0.06   -0.63     0.53
dropout            -2.30       0.33   -6.88     0.00
loginc              0.56       0.07    8.36     0.00
college             1.75       0.56    3.11     0.00
distance           -0.10       0.04   -2.15     0.03
dropout:distance    0.09       0.07    1.46     0.15
college:distance   -0.18       0.11   -1.71     0.09
> round(summary(lmer.slope)@coefs,2)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.67       0.44   -6.07     0.00
b.ord              -0.17       0.08   -2.06     0.04
dropout            -2.05       0.30   -6.77     0.00
loginc              0.51       0.06    8.20     0.00
college             1.49       0.51    2.94     0.00
distance           -0.10       0.04   -2.32     0.02
dropout:distance    0.10       0.06    1.58     0.11
college:distance   -0.15       0.10   -1.53     0.13
> round(summary(lmer.both)@coefs,2)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.11       0.46   -6.72     0.00
b.ord              -0.06       0.07   -0.95     0.34
dropout            -2.26       0.33   -6.86     0.00
loginc              0.55       0.07    8.35     0.00
college             1.71       0.56    3.08     0.00
distance           -0.10       0.04   -2.16     0.03
dropout:distance    0.09       0.06    1.46     0.14
college:distance   -0.18       0.11   -1.68     0.09
```

In all three models, the estimated slope on `b.ord` *is negative, but it is only significantly different from zero in the random slopes model, which did not have a good AIC value but had good conditional residuals.*

At this point, I would conclude that birth order does not have a significant effect on whether a hospital is preferred for birth. If there were an effect, I would expect it to be negative (the later the childbirth, the less likely to use a hospital). It would be worthwhile to explore the models and data further, to see if the random slopes model really is better at capturing important variation in the data. If that were the case, then I would want to say that increasing birth order has a negative effect on the likelihood of using a hospital for childbirth.