Hierarchical Linear Models

Statistics 220

Spring 2005

Hierarchical Linear Models

The linear regression model

$$\begin{aligned} y &\sim N(X\beta, \Sigma_y) \\ \beta|\sigma^2 &\sim p(\beta|\sigma^2) \\ \sigma^2 &\sim p(\sigma^2) \end{aligned}$$

can be extended to more complex situations. We can put more complex structures on the βs to better the describe the structure in the data.

In addition to allowing for more structure on the βs, it can also used to model the measure error structure Σ_y.

For example, consider the one-way random effects model discussed earlier

$$\begin{aligned} y_{ij}|\theta,\sigma^2 &\overset{ind}{\sim} N(\theta_j,\sigma^2) \\ \theta_j|\mu,\tau^2 &\overset{iid}{\sim} N(\mu,\tau^2) \end{aligned}$$

This is an equivalent model to (after integrating out the θs)

$$y|\mu,\Sigma_y \sim N(\mu,\Sigma_y)$$

where

$$\begin{aligned} \mathrm{Var}(y_i) &= \sigma^2+\tau^2=\eta^2 \\ \mathrm{Cov}(y_{i_1},y_{i_2}) &= \begin{cases} \rho\eta^2 & \text{if } i_1 \text{ and } i_2 \text{ in group } j \\ 0 & \text{if } i_1 \text{ and } i_2 \text{ in different groups} \end{cases} \end{aligned}$$

and

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

In this framework, ρ is often referred to as the interclass correlation.

Note that this correspondence with the usual ANOVA formulation of the model. See the text for the regression formulation of the equivalence.

This approach can be used the model the equal correlation structure

$$\Sigma_y = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

discussed last time as long as $\rho \geq 0$ (each observation is in its own group). (Note that in general that ρ can be positive. However this hierarchical model can not be used to deal with this case.)

General Hierarchical Linear Model

$$\begin{aligned} y|X, \beta, \Sigma_y &\sim N(X\beta, \Sigma_y) \\ \beta|X_\beta, \alpha, \Sigma_\beta &\sim N(X_\beta\alpha, \Sigma_\beta) \\ \alpha|\alpha_0, \Sigma_\alpha) &\sim N(\alpha_0, \Sigma_\alpha) \end{aligned}$$

The first term is the likelihood, the second term is 'population distribution' (process), and the third term is the 'hyperprior distribution'.

The X is the set of covariates for the responses y and X_β is the set of the covariates for the βs.

Often $\Sigma_y = \sigma^2 I$, $X_\beta = 1$ and $\Sigma_\beta = \sigma_\beta^2 I$.

Usually the hyprerprior parameters α_0 and Σ_α are treated as fixed. Often the noninformative prior $p(\alpha) \propto 1$ is used.

Note that this can be treated as a single linear regression with the structure

$$y_*|X_*, \gamma, \Sigma_* \sim N(X_*\gamma, \Sigma_*)$$

with $\gamma = (\beta\ \alpha)^T$ and

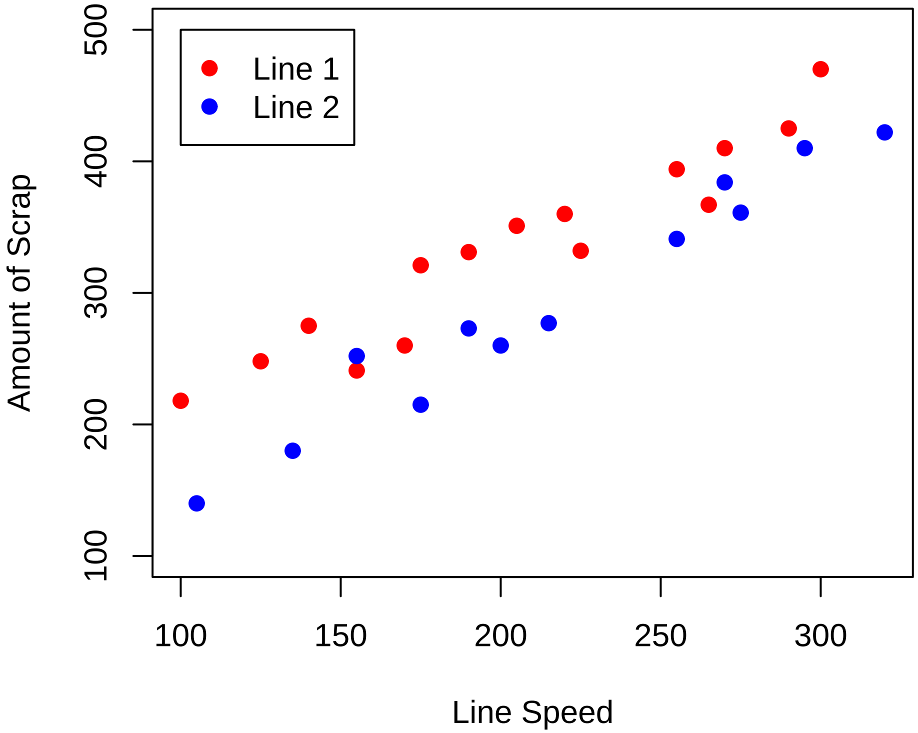
$$y_* = \begin{bmatrix} y \\ 0 \\ \alpha_0 \end{bmatrix} \quad X_* = \begin{bmatrix} X & 0 \\ I_J & -X_\beta \\ 0 & I_K \end{bmatrix} \quad \Sigma_* = \begin{bmatrix} \Sigma_y & 0 & 0 \\ 0 & \Sigma_\beta & 0 \\ 0 & 0 & \Sigma_\alpha \end{bmatrix}$$

While this is sometimes useful for computation, as many conditional distributions just fall out, it is less useful in terms of interpretation.

Regression Example

Soap Production Waste

- y: Amount of scrap
- x_1: Line speed
- x_2: Production line (1 or 2)



There are $n_1 = 15$ observations on Line 1 and $n_2 = 12$ observations on Line 2.

Want to fit a model allowing different slopes and intercepts for each production line (i.e. an interaction model).

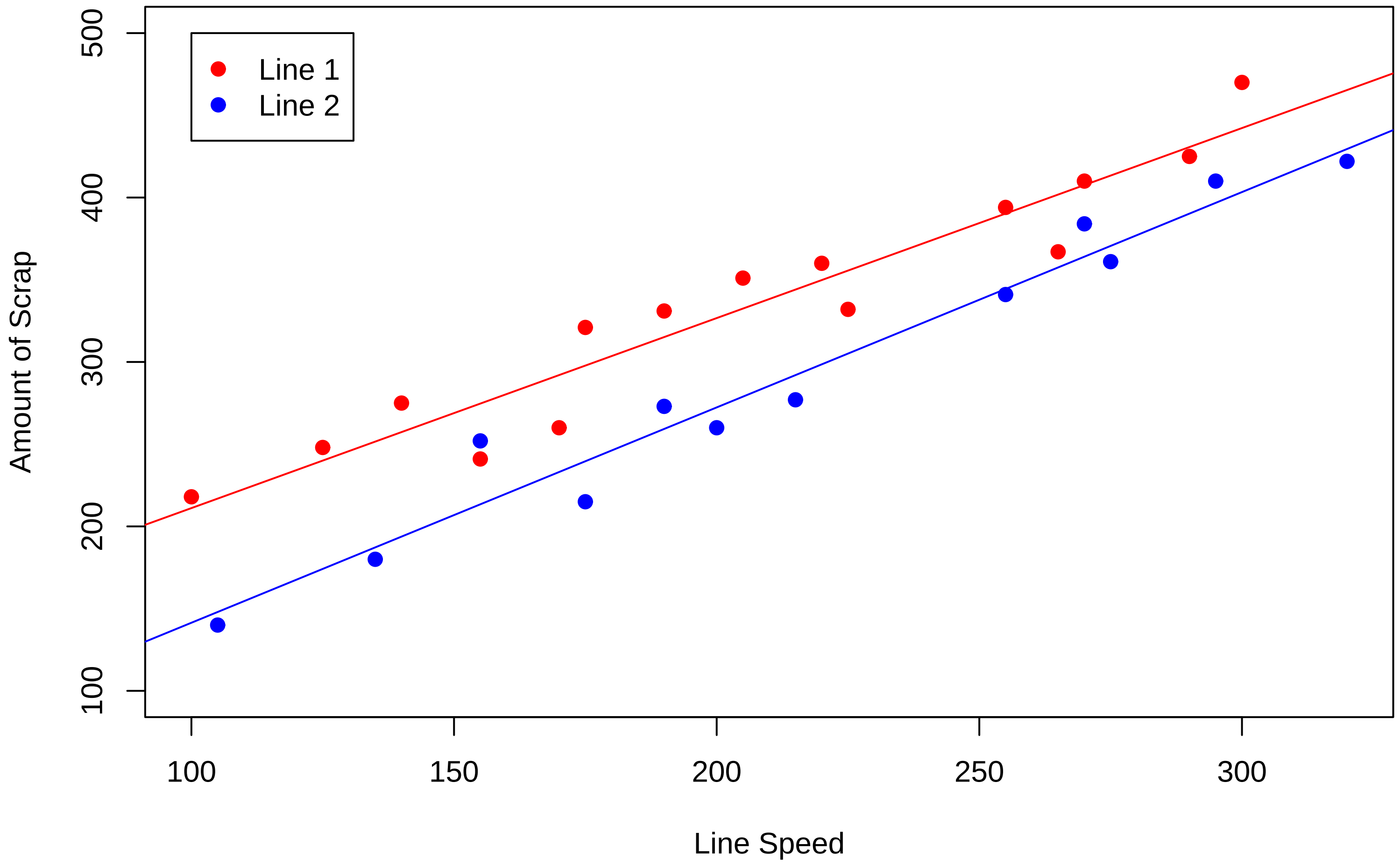
We can use the following model

$$\begin{aligned} y_{ij}|\beta_j,\sigma_j^2 &\overset{ind}{\sim} N(\beta_{0j}+\beta_{1j}x_{1ij},\sigma_j^2); \qquad i=1,\ldots,n_j, j=1,2 \\ \beta_{0j}|\alpha_0 &\overset{iid}{\sim} N(\alpha_0,100) \\ \beta_{1j}|\alpha_1 &\overset{iid}{\sim} N(\alpha_1,1) \\ \alpha_0 &\sim N(0,10^6) \\ \alpha_1 &\sim N(0,10^6) \end{aligned}$$

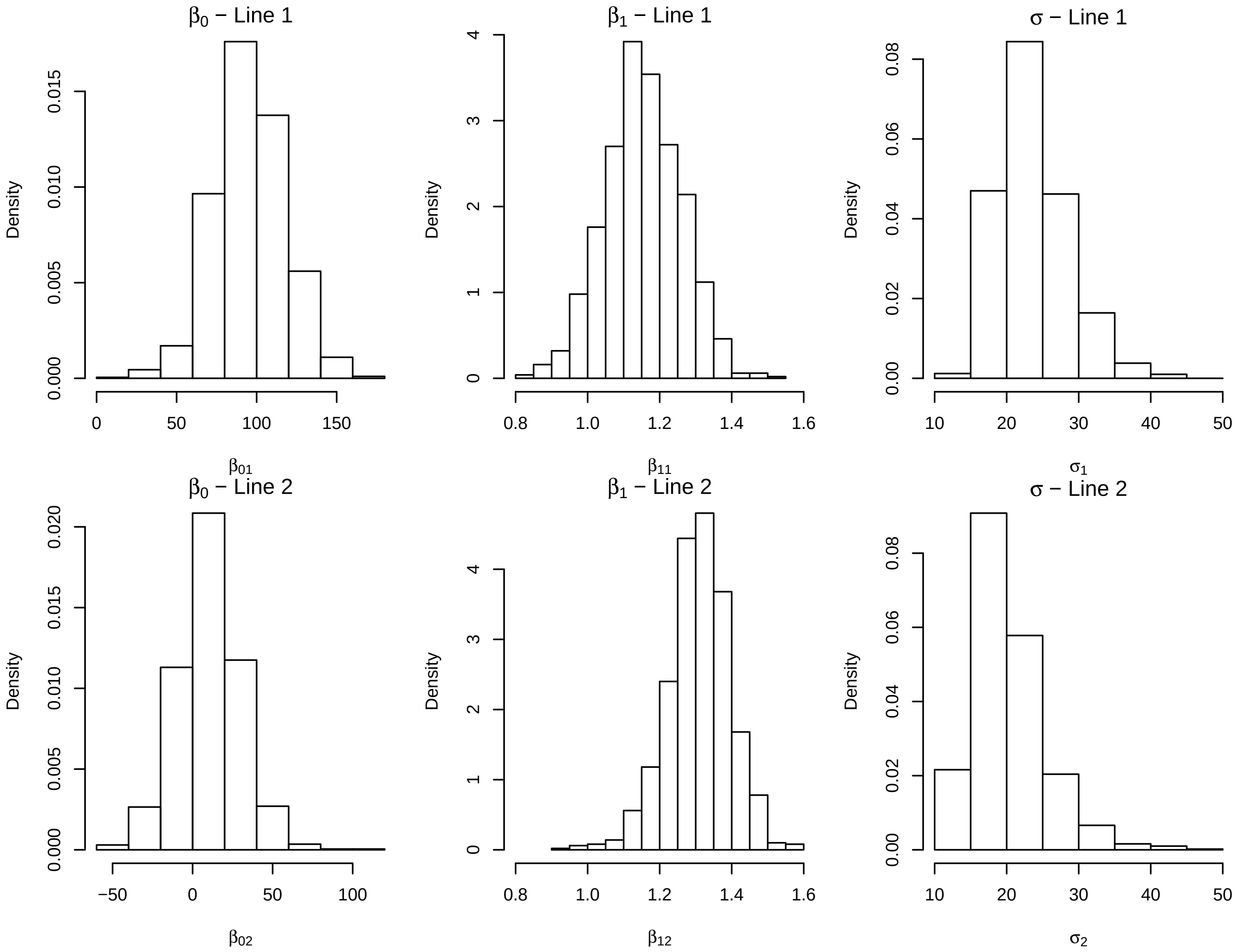
This model forces the two regression lines to be somewhat similar, though the prior form for the lines is vague.

Note that this does fit into the framework mentioned earlier with $\beta = (\beta_{01}\ \beta_{11}\ \beta_{02}\ \beta_{12})^T$, $\alpha = (\alpha_0\ \alpha_1)^T$ and X and X_β have the forms

$$X = \begin{bmatrix} 1 & x_{111} & 0 & 0 \\ & & \vdots & \\ 1 & x_{1n_1 1} & 0 & 0 \\ 0 & 0 & 1 & x_{112} \\ & & \vdots & \\ 0 & 0 & 1 & x_{1n_2 2} \end{bmatrix} \qquad X_\beta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$



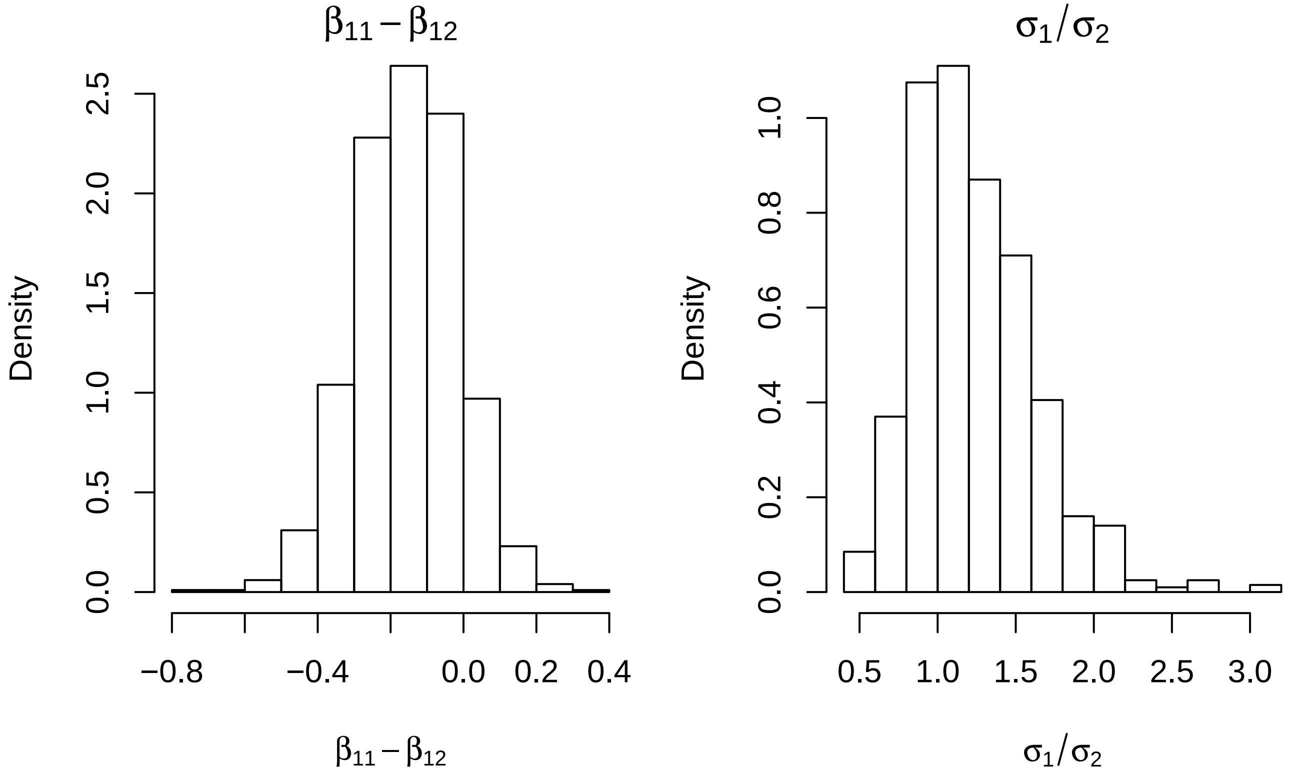
The posterior mean lines suggest that the intercepts are quite different but the slopes of the lines are similar, though the slope for line 1 appears to be a bit flatter.


β0 − Line 1
Density
0.000
0.005
0.010
0.015
0
50
100
150
β01
β1 − Line 1
Density
0
1
2
3
4
0.8
1.0
1.2
1.4
1.6
β11
σ − Line 1
Density
0.00
0.02
0.04
0.06
0.08
10
20
30
40
50
σ1
β0 − Line 2
Density
0.000
0.005
0.010
0.015
0.020
−50
0
50
100
β02
β1 − Line 2
Density
0
1
2
3
4
0.8
1.0
1.2
1.4
1.6
β12
σ − Line 2
Density
0.00
0.02
0.04
0.06
0.08
10
20
30
40
50
σ2

The similarity of the slopes is also suggested by the previous histograms and the following posterior summary statistics. It also appears that variation around the regression lines are similar for the two lines, though it appears that the standard deviation is larger for line 1.

Parameter	Mean	SD
β_{01}	95.57	22.56
β_{02}	10.44	19.60
β_{01}	1.156	0.107
β_{02}	1.310	0.088
σ_1	23.68	4.82
σ_2	20.24	5.03

We can examine whether there is a difference between the slopes by examining the distribution of $\beta_{11} - \beta_{12}$ and a difference in variance about the regression line by looking at the distribution of $\frac{\sigma_1}{\sigma_2}$.



The is marginal evidence for a difference in slopes as

$$
\begin{aligned}
P[\beta_{11} > \beta_{12}|y] &= 0.125 \\
E[\beta_{11} > \beta_{12}|y] &= -0.154 \\
\mathrm{Med}(\beta_{11} > \beta_{12}|y) &= -0.150
\end{aligned}
$$

There is less evidence for a difference in σs as

$$\begin{aligned} P[\sigma_1 > \sigma_2|y] &= 0.694 \\ E[\sigma_1 > \sigma_2|y] &= 1.24 \\ \mathrm{Med}(\sigma_1 > \sigma_2|y) &= 1.17 \end{aligned}$$

This is also supported by comparing this model with the model where $\sigma_1^2 = \sigma_2^2$.

Model	DIC	p_D
Common σ^2	247.2	5.6
Different σ^2	249.2	7

In this case the smaller model with $\sigma_1^2 = \sigma_2^2$ appears to be giving the better fit, though it is not a big difference.

Fitting Hierarchical Linear Models

Not surprisingly, exact distributional results for these hierarchical models do not exist and Monte Carlo methods are required.

The usual approach is some form of Gibbs sampler. There are a wide range of approaches that can be used for sampling the regression parameters

- All-at-once Gibbs

  All regression parameters $\gamma = (\beta \ \alpha)^T$ are drawn jointly given y and the variance parameters. While this is simple in theory, for some problems the dimensionality can be huge and this can be inefficient.

- Scalar Gibbs

  Draw each parameter separately. This can be much faster as the dimension of each draw is small. Unfortunately, the chain may mix slowly in some cases

- Blocking Gibbs

  Sample the regression parameters in blocks. This helps with the dimensionality problems and will tend to mix faster than Scalar Gibbs.

- Scalar Gibbs with a linear transformation

  By rotating the parameter space the Markov Chain will tend to mix quickly. Thus working with

$$\xi = A^{-1}(\gamma - \gamma_0)$$

  where $A = V_\gamma^{1/2}$ will mix much better. After this transformation, sample the component of ξ one by one and then transform back at the end of each scan to give γ.

  This approach can also be used with the Blocking Gibbs form.

ANOVA

Many Hierarchical Linear Models are examples of ANOVA models. This should not be surprising as any ANOVA model can be written as an regression model where all predictor variables are indicator variables.

In this situation, the βs will fall in blocks, corresponding to the different factors in the studies. For example consider a two way design with the interaction terms

$$y_{ijk} \overset{ind}{\sim} N(\mu + \phi_i + \theta_j + (\phi\theta)_{ij}, \sigma^2)$$

In this case there are three blocks, the main effects ϕ_i and θ_j and the interactions $(\phi\theta)_{ij}$.

A common approach is to put a separate prior structure on each block. For this example, put the prior on the treatment effects

$$\begin{aligned}\phi_i &\overset{iid}{\sim} N(0, \sigma^2_\phi) \\ \theta_j &\overset{iid}{\sim} N(0, \sigma^2_\theta) \\ (\phi\theta)_{ij} &\overset{iid}{\sim} N(0, \sigma^2_{\phi\theta})\end{aligned}$$

For the variance parameters, the conjugate hyperprior

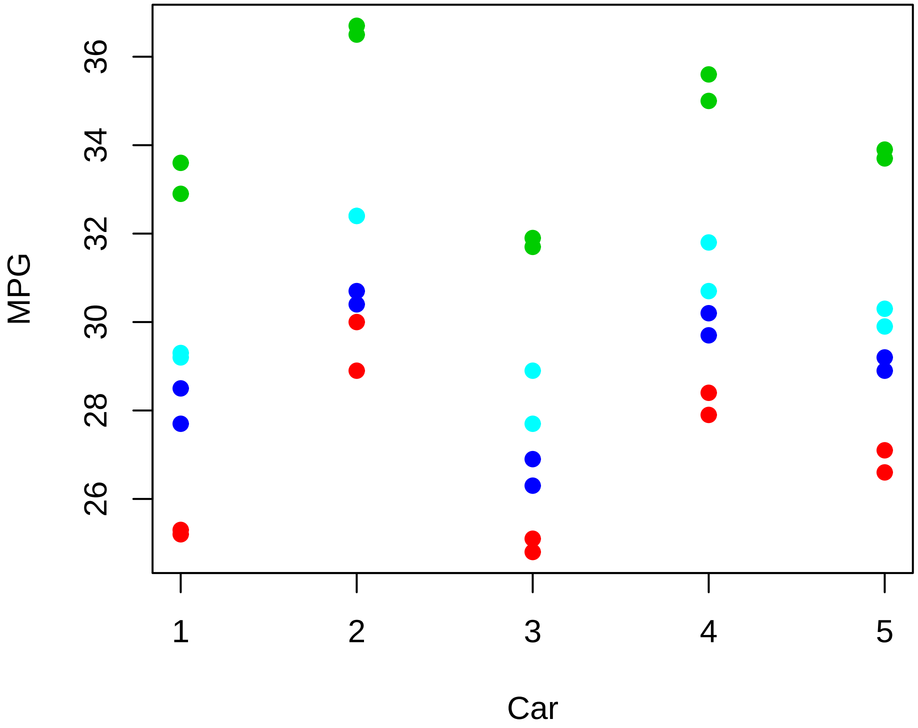
$$\begin{aligned}\sigma^2_\phi &\sim \text{Inv}-\chi^2(\nu_\phi, \sigma^2_{0\phi}) \\ \sigma^2_\theta &\sim \text{Inv}-\chi^2(\nu_\theta, \sigma^2_{0\theta}) \\ \sigma^2_{\phi\theta} &\sim \text{Inv}-\chi^2(\nu_{\phi\theta}, \sigma^2_{0\phi\theta})\end{aligned}$$

Note that as in standard ANOVA analyzes, the interaction terms are only included if all of the lower order effects included in the interaction are in the model. For example, for a three way ANOVA, the three-way interaction will only be included if all the main effects and two-way interactions are included in the model.

ANOVA Example

MPG: The effect of driver (4 levels) and car (5 levels) were examined. Each driver drove each car over a 40 mile test course twice.

From the plot of the data, it appears that both driver and car have an effect on gas mileage. As the pattern of MPG for each driver seems to be the same for each car (points are roughly shifted up or down as the car level changes) is appears that the interaction effects are small.



The standard ANOVA analysis agrees with this hypothesis as

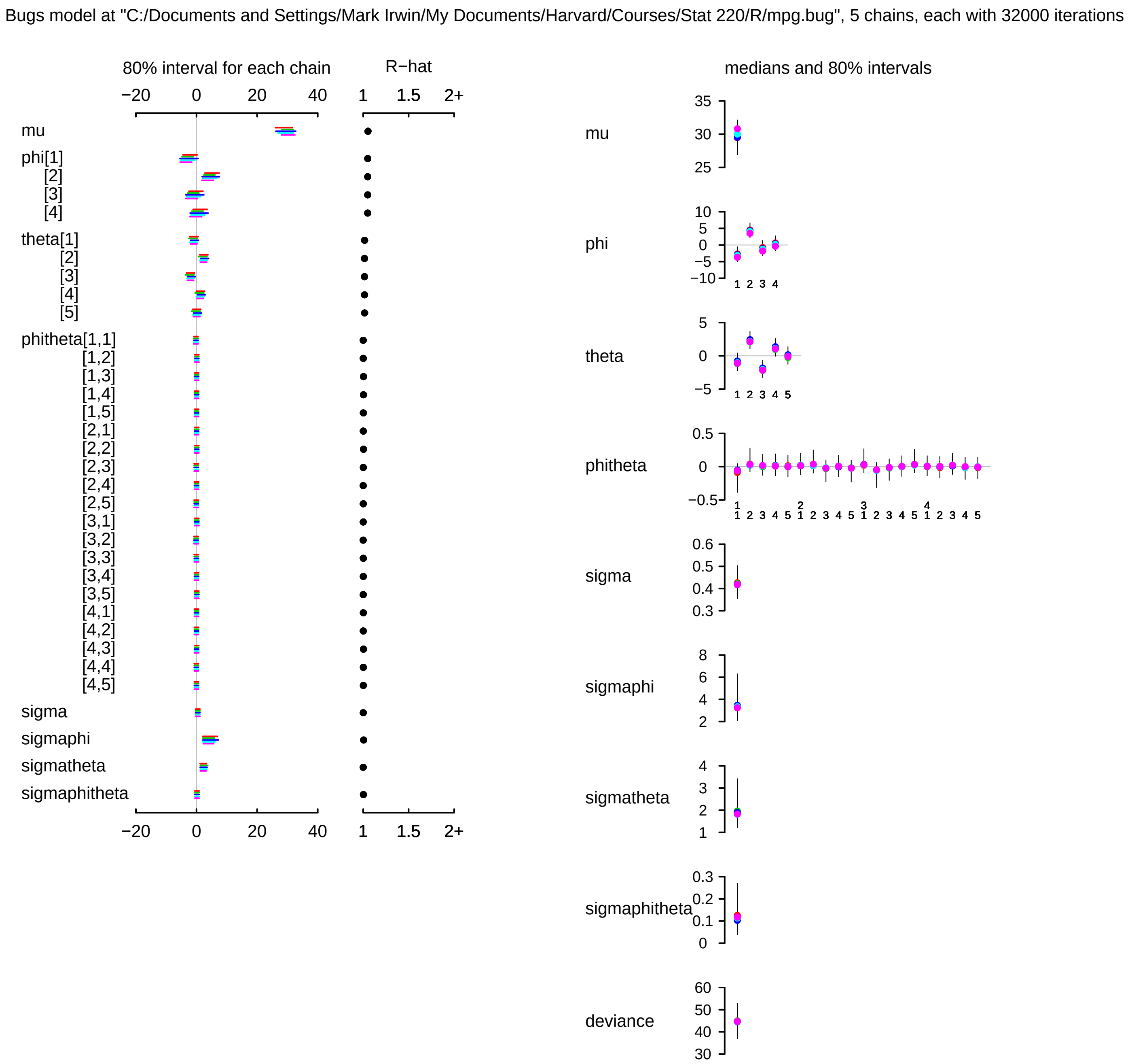
```
Analysis of Variance Table

Response: MPG
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Car        4  94.713  23.678  134.73 3.664e-14 ***
Driver     3 280.285  93.428  531.60 < 2.2e-16 ***
Car:Drive 12   2.446   0.204    1.16    0.3715    
Residuals 20   3.515   0.176                      
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Inference for Bugs model at "mpg.bug"
 5 chains, each with 32000 iterations (first 16000 discarded),
 n.thin = 40, n.sims = 2000 iterations saved
Time difference of 23 secs
              mean  sd 2.5%  25%  50%  75% 97.5% Rhat n.eff
mu            29.8 2.1 25.4 28.5 30.0 31.2  33.4  1.1    84
phi[1]        -3.0 1.9 -6.3 -4.1 -3.1 -1.9   1.5  1.0    99
phi[2]         4.2 1.9  1.0  3.1  4.1  5.3   8.8  1.0    98
phi[3]        -1.1 1.9 -4.4 -2.2 -1.2  0.0   3.5  1.1    98
phi[4]         0.4 1.9 -3.0 -0.8  0.2  1.4   4.9  1.0    99
theta[1]      -0.9 1.1 -3.0 -1.6 -1.0 -0.4   1.5  1.0   230
theta[2]       2.3 1.1  0.2  1.7  2.2  2.8   4.9  1.0   250
theta[3]      -2.0 1.1 -4.1 -2.6 -2.0 -1.4   0.5  1.0   250
theta[4]       1.2 1.1 -0.9  0.6  1.1  1.8   3.7  1.0   220
theta[5]       0.0 1.1 -2.1 -0.6  0.0  0.6   2.5  1.0   220
```

```
              mean  sd 2.5%  25%  50%  75% 97.5% Rhat n.eff
phitheta[1,1] -0.1 0.2 -0.6 -0.2 -0.1  0.0   0.1  1.0  2000
phitheta[1,2]  0.1 0.2 -0.2  0.0  0.0  0.1   0.5  1.0  2000
phitheta[1,3]  0.0 0.1 -0.3  0.0  0.0  0.1   0.4  1.0  2000
phitheta[1,4]  0.0 0.1 -0.3  0.0  0.0  0.1   0.4  1.0  2000
phitheta[1,5]  0.0 0.2 -0.3 -0.1  0.0  0.1   0.3  1.0  1300
phitheta[2,1]  0.0 0.1 -0.3  0.0  0.0  0.1   0.4  1.0  2000
phitheta[2,2]  0.1 0.2 -0.2  0.0  0.0  0.1   0.4  1.0  1100
phitheta[2,3]  0.0 0.1 -0.4 -0.1  0.0  0.0   0.2  1.0  1600
phitheta[2,4]  0.0 0.1 -0.3 -0.1  0.0  0.1   0.3  1.0  2000
phitheta[2,5]  0.0 0.2 -0.4 -0.1  0.0  0.0   0.2  1.0  2000
phitheta[3,1]  0.1 0.2 -0.2  0.0  0.0  0.1   0.5  1.0  2000
phitheta[3,2] -0.1 0.2 -0.5 -0.2 -0.1  0.0   0.2  1.0  2000
phitheta[3,3]  0.0 0.1 -0.4 -0.1  0.0  0.0   0.3  1.0  2000
phitheta[3,4]  0.0 0.2 -0.3 -0.1  0.0  0.1   0.3  1.0  2000
phitheta[3,5]  0.1 0.2 -0.2  0.0  0.0  0.1   0.5  1.0  2000
```

```
                mean  sd 2.5%  25%  50%  75% 97.5% Rhat n.eff
phitheta[4,1]    0.0 0.1 -0.3 -0.1  0.0  0.1   0.3  1.0  1900
phitheta[4,2]    0.0 0.1 -0.3 -0.1  0.0  0.1   0.3  1.0  2000
phitheta[4,3]    0.0 0.1 -0.2  0.0  0.0  0.1   0.4  1.0  2000
phitheta[4,4]    0.0 0.1 -0.3 -0.1  0.0  0.0   0.3  1.0  2000
phitheta[4,5]    0.0 0.1 -0.3 -0.1  0.0  0.1   0.3  1.0  2000
sigma            0.4 0.1  0.3  0.4  0.4  0.5   0.6  1.0  2000
sigmaphi         4.0 2.2  1.7  2.5  3.4  4.7  10.2  1.0   890
sigmatheta       2.2 1.2  1.0  1.5  1.9  2.6   5.0  1.0  2000
sigmaphitheta    0.1 0.1  0.0  0.1  0.1  0.2   0.4  1.0  1000
deviance        44.9 6.3 32.4 40.7 44.7 48.9  57.7  1.0  2000
 pD = 20.1 and DIC = 65 (using the rule, pD = var(deviance)/2)
```


Bugs model at "C:/Documents and Settings/Mark Irwin/My Documents/Harvard/Courses/Stat 220/R/mpg.bug", 5 chains, each with 32000 iterations
80% interval for each chain
R−hat
−20
0
20
40
1
1.5
2+
mu
phi[1]
[2]
[3]
[4]
theta[1]
[2]
[3]
[4]
[5]
phitheta[1,1]
[1,2]
[1,3]
[1,4]
[1,5]
[2,1]
[2,2]
[2,3]
[2,4]
[2,5]
[3,1]
[3,2]
[3,3]
[3,4]
[3,5]
[4,1]
[4,2]
[4,3]
[4,4]
[4,5]
sigma
sigmaphi
sigmatheta
sigmaphitheta
medians and 80% intervals
mu
35
30
25
phi
10
5
0
−5
−10
1 2 3 4
theta
5
0
−5
1 2 3 4 5
phitheta
0.5
0
−0.5
1 2 3 4
1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
sigma
0.6
0.5
0.4
0.3
sigmaphi
8
6
4
2
sigmatheta
4
3
2
1
sigmaphitheta
0.3
0.2
0.1
0
deviance
60
50
40
30