Hierarchical Model Specification in Quantitative Research.

Chapters 14-15, Multilevel Logistic Regression and More

JEFF GILL
Distinguished Professor, Department of Government
Professor, Department of Mathematics & Statistics
Member, Center for Neuroscience
American University

Dichotomous Outcomes Overview

- We will create a regression model for dichotomous outcome variables: vote/not-vote, war/no-war, pass/fail, etc.
- Note that this is different than having dichotomous explanatory variables.
- Remember that regression is really conditional average, $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$, which does not have the same implications for $0/1$ outcomes on the LHS.
- Consider the probability that a single case has a 0 or a 1 as the outcome:

$$\pi_i = p(Y_i) = p(Y = 1|\mathbf{X} = \mathbf{x}_i), \quad \text{where} \quad \pi \in [0{:}1].$$

- So:

$$\mathbb{E}(Y_i|\mathbf{x}_i) = (\pi_1)(1) + (1 - \pi_i)(0) = \pi_i.$$

  (recall that for discrete RV $\mathbb{E}(A) = \sum_{\text{over events}} P(\text{A}) \times A$)
- This means that we are *estimating* an underlying probability value for given levels of a vector of explanatory variable values.

Conceptual Model

- Start with the linear predictor $\boldsymbol{\eta} = \alpha + \beta\mathbf{x}$.
- Now let's specify a link function that relates the linear additive RHS component to the expected value of the nonlinear LHS component:

$$\pi_i = g^{-1}(\eta_i) = p(\alpha_i + \beta_i x) \;\Rightarrow\; g(\pi_i) = \eta_i = \alpha_i + \beta_i x.$$

- Objectives for $g^{-1}()$:
  - ▷ smooth on [0:1]
  - ▷ For a positive effect of $\mathbf{x}_i$ on π_i:
    - $g^{-1} \rightarrow 0$ as $x_i \rightarrow, -\infty$
    - $g^{-1} \rightarrow 1$ as $x_i \rightarrow, +\infty$.
  - ▷ For a negative effect of $\mathbf{x}_i$ on π_i:
    - $g^{-1} \rightarrow 1$ as $x_i \rightarrow, -\infty$
    - $g^{-1} \rightarrow 0$ as $x_i \rightarrow, +\infty$.

New Conceptual Model

- There are two common solutions for $g^{-1}()$.
- Logit:

$$\Lambda(\eta_i) = [1 + \exp(-\eta_i)]^{-1}$$

- Probit:

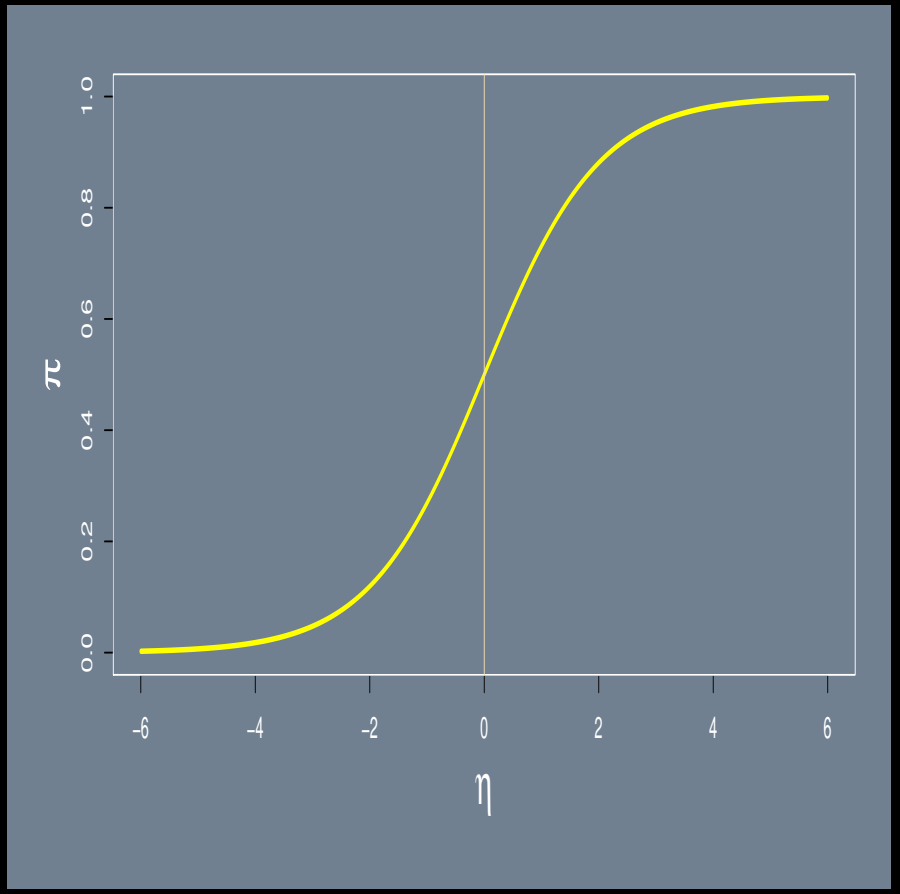
$$\Phi(\eta_i) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\eta_i} \exp[-\frac{1}{2}\eta_i^2] d\eta_i$$

- These are sometimes given in $g()$ form: $\Phi^{-1}(\pi_i)$ and $\Lambda^{-1}(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{p_i}{1-p_i}\right)$.
- Less common is the cloglog function:

$$g(\mu) = -\log\left(-\log(1-\mu)\right) \qquad g^{-1}(\eta) = 1 - \exp\left(-\exp(\eta)\right)$$

Latent Variable Justification

- Humans make dichotomous decisions from smooth preference structures, but we only see discrete choices in the data.
- The Index Function (Utility) model states that if *benefits* - *costs* = U is greater than zero then the choice should be a one, and vice-versa.



Latent Variable Justification

- Utility model states: $U_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ (subsume the constant into the vector), and $p(U_i > 0) = p(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i > 0) = p(\boldsymbol{\epsilon}_i > -\mathbf{x}_i\boldsymbol{\beta})$.
- Political Example:
  - U^R, the utility of voting for the Republican candidate
  - U^D, the utility of voting for the Democratic candidate
  - direction is arbitrary, so pick $Y = 1$ the decision to vote for the Republican candidate
  - Define the two utility functions in regression terms:

$$U_i^R = \mathbf{x}_i\boldsymbol{\beta}_R + \boldsymbol{\epsilon}_{iR} \qquad\qquad U_i^D = \mathbf{x}_i\boldsymbol{\beta}_D + \boldsymbol{\epsilon}_{iD}$$

  - So now:

$$\begin{aligned} p(Y_i = 1|\mathbf{x}_i) &= p(U_i^R > U_i^D) \\ &= p(\mathbf{x}_i\boldsymbol{\beta}_R + \boldsymbol{\epsilon}_{iR} > \mathbf{x}_i\boldsymbol{\beta}_D + \boldsymbol{\epsilon}_{iD}|\mathbf{x}_i) \\ &= p(\mathbf{x}_i[\boldsymbol{\beta}_R - \boldsymbol{\beta}_D] + \boldsymbol{\epsilon}_{iR} - \boldsymbol{\epsilon}_{iD} > 0) \\ &= p(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon} > 0) \end{aligned}$$

    which is just 1-CDF.

Binomial Regression Model

- If Y_i for $i = 1, \ldots, n$ is iid binomial $B(n_i, p_i)$, then:

$$p(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

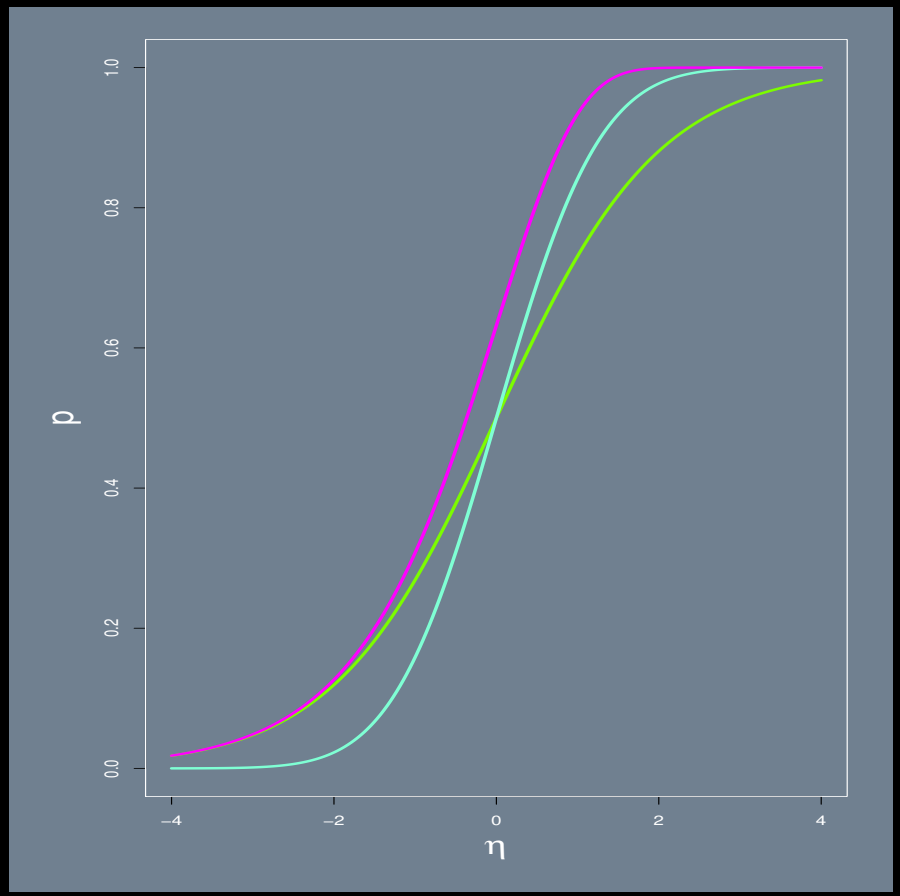
- Further suppose that these are affected by the same q predictors (covariates, explanatory variables), $x_{i1}, \ldots, x_{iq}$.
- The tool that connects these predictors to p is the linear predictor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}.$$

- We still need a link function, $\eta_i = g(p_i)$, that is not an identity ($\eta_i = p_i$) since we need $0 \leq p_i \leq 1$.

Binomial Link Functions

- Logit (logistic): $\eta = \log\left(\frac{p}{1-p}\right)$, $p = \frac{\exp(\eta)}{1+\exp(\eta)}\,[1+\exp(-\eta)]$.
- Probit: $\eta = \Phi^{-1}(p)$, $p = \Phi(\eta)$.
- Complementary log-log: $\eta = \log(-\log(1-p))$, $p = 1-\exp(-\exp(\eta))$.




```
ruler <- seq(-4,4,length=200)
postscript("Class.MLE/faraway.ch2.fig3.ps")
par(col.axis="white",col.lab="white",col.sub="white",
    col="white", bg="slategray",cex.lab=2,mar=c(6,6,2,2))
plot(ruler,exp(ruler)/(1+exp(ruler)),type="l",lwd=3,
     col="lawngreen",ylim=c(0,1),
     xlab=expression(eta),ylab="p")
lines(ruler,pnorm(ruler),lwd=3,col="aquamarine")
lines(ruler,1-exp(-exp(ruler)),lwd=3,col="magenta")
dev.off()
```

Binomial Model Estimation

- Define a likelihood function for observed iid y_i, where $i = 1, \ldots, n$ from $f(y|p)$.
- Then the *joint distribution* of these observed data is:

$$p(y_1, y_2, \ldots, y_n) = p(y_1|\boldsymbol{\beta}, \mathbf{x}_1) f(y_2|\boldsymbol{\beta}, \mathbf{x}_2) \cdots f(y_n|\boldsymbol{\beta}, \mathbf{x}_n) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}, \mathbf{x}_i).$$

- If we consider that p is really the unknown and the y_i are known, then it makes sense to think of this joint function as a function that reveals something about $\boldsymbol{\beta}$.
- Denote it $L(\boldsymbol{\beta}|\mathbf{x}, \mathbf{y})$, which is called a likelihood function.

Binomial Model Estimation

- More precisely, we can incorporate the information that Y can only be 0 or 1:

$$L(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) = \prod_{y_i=0}[1-F(\mathbf{X}_i\boldsymbol{\beta})]\prod_{y_i=1}[F(\mathbf{X}_i\boldsymbol{\beta})]$$

$$= \prod_{i=1}^{n}[1-F(\mathbf{X}_i\boldsymbol{\beta})]^{1-y_i}\,[F(\mathbf{X}_i\boldsymbol{\beta})]^{y_i}$$

$$\ell(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n}[(1-y_i)\log(1-F(\mathbf{X}_i\boldsymbol{\beta})) + y_i\log(F(\mathbf{X}_i\boldsymbol{\beta}))]$$

- The log-likelihood is concave to the x-axis for common choices of $F()$, and produces coefficient estimates that are distributed student's-t.
- Generally with the binomial setup it is easier to think in terms of the CDF, $F()$, rather than the PDF, $f()$, since the former directly describes the S-curve of theoretical interest.

Binomial Model MLE

- The gradient is given by:

$$G = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \left[\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f + i}{1 - F_i} \right] \mathbf{x}_i$$

- The Hessian is given by:

$$H = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \ell(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \frac{f_i^2}{F_i(1 - F_i)} \mathbf{x}_i \mathbf{x}_i'$$

- The Variance-Covariance Matrix is calculated as:

$$VC_{\boldsymbol{\beta}} = E\left[-H^{-1}\right]$$

Common Forms

- Probit, where $\phi_i = \phi_i(\mathbf{x}_i\boldsymbol{\beta})$ and $\Phi_i = \Phi_i(\mathbf{x}_i\boldsymbol{\beta})$:

$$G = \sum_{y=0} \frac{-\phi_i}{1-\Phi_i}\boldsymbol{\beta}\mathbf{x}_i + \sum_{y_i=1} \frac{\phi_i}{\Phi_i}\boldsymbol{\beta}\mathbf{x}_i$$

$$H = \left\{ \sum_{i=0} \left[-\frac{-\phi_i^2}{(1-\Phi_i)^2} + \frac{\mathbf{x}_i\boldsymbol{\beta}\phi_i}{1-\Phi_i} \right] + \sum_{i=1} \left[-\frac{\mathbf{x}_i\boldsymbol{\beta}\phi_i}{\Phi_i} - \phi_i^2 \right] \right\} \mathbf{x}_i\mathbf{x}_i'$$

$$VC_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\phi_i^2}{\Phi_i(1-\Phi_1)}\mathbf{x}_i\mathbf{x}_i'$$

- Logit, where $\Lambda_i = 1/[1+\exp(\mathbf{X}_i\boldsymbol{\beta})]$:

$$G = \sum_{i=1}^{n} (y_i - \Lambda_i)\mathbf{x}_i \qquad H = \sum_{i=1}^{n} \{-\Lambda_i(1-\Lambda_i)\}\,\mathbf{x}_i\mathbf{x}_i'$$

$$VC_{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \{\Lambda_i(1-\Lambda_i)\}\,\mathbf{x}_i\mathbf{x}_i' \right]^{-1}$$

Interpretation of Individual Binomial $\boldsymbol{\beta}$ Results

- sign of the parameter estimate
- predicted/fitted values
- marginal effects, including first differences
- derivative methods
- note: $\text{logit}(\boldsymbol{\beta}) \approx \frac{\pi}{\sqrt{3}}\text{probit}(\boldsymbol{\beta})$
- Wald (t-tests) for significance:

$$W = (R\hat{\boldsymbol{\beta}} - q)\left[R(VC_{\hat{\boldsymbol{\beta}}})R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - q)$$

for $H_0 : R\hat{\boldsymbol{\beta}} = q$ (commonly $R = 1, q = 0$, so that $W \sim F_{df=J,n-K}$. (where J is the number of restrictions stipulated in R). For individual coefficients, this reduces to:

$$W_k = (\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k / VC_{\hat{\boldsymbol{\beta}}}[k,k])^{\frac{1}{2}} \sim t_{df=n-k}$$

(where $n \times k$ is the dimension of the $\mathbf{X}$ matrix).

- We know that the F-test is more robust than the t-test (Hauck-Donner effect, JASA 1977).

Dichotomous Example

- Consider depression era economic and electoral data, to calculate a dichotomous regression model for using each of these link functions according to the following model specification:

$$FDR \sim f[(POST.DEP - PRE.DEP) + FARM]$$

where: FDR indicates whether or not Roosevelt carried that state in the 1932 presidential elections, $PRE.DEP$ is the mean per-state income before the onset of the Great Depression (1929) in dollars, $POST.DEP$ is the mean per-state income after the onset of the great depression (1932) in dollars, and $FARM$ is the total farm wage and salary disbursements in thousands of dollars per state in 1932.

- The first three cases look like this...

State	FDR	PRE.DEP	POST.DEP	FARM
Alabama	1	323	162	4067
Arizona	1	600	321	6100
Arkansas	1	310	157	8134

Dichotomous Example

```
fdr.out <- glm(FDR~I(POST2.DEP-PRE.DEP)+FARM, family=binomial(link="probit"),
    data=fdr)
summary.glm(fdr.out)

 Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -0.9304  -0.5412  -0.3447  -0.1672   2.5301

 Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
 (Intercept)          -2.752e+00  9.701e-01  -2.837  0.00455
 I(POST2.DEP - PRE.DEP) -8.295e-03  3.662e-03  -2.265  0.02351
 FARM                 -6.566e-05  3.812e-05  -1.723  0.08496
 ---
 (Dispersion parameter for binomial family taken to be 1)

        Null deviance: 36.434  on 48  degrees of freedom
        Residual deviance: 29.093  on 46  degrees of freedom
        AIC: 35.093
```

Percent Predicted Correctly

- Compares actual against predicted in a 2-by-2 table:

		Prediction 0	1
Data	0	correct	incorrect
	1	incorrect	correct

- But wait! These models do not produce predicted 0/1 values, for instance:

```
round(logitmod2$fitted.values,3)
    1     2     3     4     5     6     7     8     9    10    11    12    13
0.939 0.859 0.829 0.603 0.430 0.375 0.375 0.375 0.322 0.274 0.230 0.230 0.230
   14    15    16    17    18    19    20    21    22    23
0.230 0.158 0.130 0.086 0.086 0.069 0.069 0.045 0.036 0.023
```

from the Bernoulli treatment.

Percent Predicted Correctly

- The naïve criteria:

$$p_i = 1 \text{ if, } F(\mathbf{x}_i\boldsymbol{\beta}) > 0.5 \qquad p_i = 0 \text{ if, } F(\mathbf{x}_i\boldsymbol{\beta}) < 0.5$$

- Create the table:

```
ppc <- cbind(orings2$damage, round(logitmod2$fitted.values,3))
( naive <- matrix(c(
    nrow(ppc[(ppc[,1] == 0) & (ppc[,2] < 0.5),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 0) & (ppc[,2] > 0.5),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 1) & (ppc[,2] < 0.5),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 1) & (ppc[,2] > 0.5),])/nrow(ppc)),
    byrow=TRUE,ncol=2) )
```

```
        [,1]    [,2]
[1,] 0.69565 0.00000
[2,] 0.13043 0.17391
```

- Better criteria: mean of $\hat{y}_i$, substantive/theoretical point.

Binomial Model Comparison

- Compare two models, one with ℓ parameters and one with s parameters such that $\ell > s$ and every parameter in the s set is also in the ℓ set: nesting.
- Denote the first as $L(p|\mathbf{y}, \mathbf{X}_L) = L_L$ and the second as $L(p|\mathbf{y}, \mathbf{X}_S) = L_S$.
- A tool for comparing these models is the likelihood ratio statistic:

$$LRT = 2\log\frac{L_L}{L_S} = 2(\log(L_L) - \log(L_S)) = -2\log\frac{L_S}{L_L} = -2(\log(L_S) - \log(L_L)).$$

- This is distributed asymptotically χ^2 with degrees of freedom the difference between the number of parameters in the two models.
- Tail values support the nesting values, meaning that the restricted values are not supported.

Binomial Model Comparison

- The most extreme case of L_L fits a "covariate" to every datapoint as an indicator function, and is thus a regression model where every datapoint is a separate inference.
- This is called the saturated model and provides no data-reduction and no modeling value, but serves as a reference point.
- For the binomial model, the saturated model can be described by $\hat{p}_i = y_i/n_i$, which is the number of success over the number of trials for the i th case (frequently $n_i = 1$).
- Another reference point is a model that uses β_0 only and is called a *mean model*.
- Thus any model we specify "lives" between these two extremes of model fit.
- Residuals in the nonlinear regression sense are called deviances to distinguish them from the assumptions in linear models.

Binomial Model Comparison

- So it should be clear that:

$$\sum D_{\text{saturated model}} < \sum D_{\text{our specified model}} < \sum D_{\text{mean model}}$$

- For the binomial model, the LRT reduces to a ratio of the saturated model to the specified model, given by:

$$D = 2\sum_{i=1}^{n}\{y_i \log(y_i/\hat{y}_i) + (n_i - y_i)\log((n_i - y_i)/(n_i - \hat{y}_i))\},$$

  where $\hat{y}_i$ are the fitted values from the smaller (specified) model.

- The mean model provides a large value of D called the *null deviance*.

- D for assessing a model with p covariates is asymptotically distributed χ^2_{n-p}, where $n - p$ is the degrees of freedom.

- Here is a contrived example ($n = 23$):

```
summary(logitmod)
    :
    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
```

Binomial Model Comparison

- Formal tests:
  - Specified model versus saturated model:

    ```
    pchisq(deviance(logitmod),df.residual(logitmod),lower=FALSE)
    0.71641
    ```

    which is not in the χ^2_{21} tail, so it is statistically "close" to the saturated model and therefore a good fit.
  - Mean model versus saturated model:

    ```
    pchisq(38.9,22,lower=FALSE)
    0.014489
    ```

    which is in the χ^2_{22} tail, so it is statistically "far" from the saturated model and therefore not a good fit.
  - Specified model (with temperature) versus mean model ($D_S - D_L$):

    ```
    pchisq(38.9-16.9,1,lower=FALSE)
    2.7265e-06
    ```

    which is in the χ^2_{22} tail, so L_S is statistically "far" from L_L.

Binomial Model Comparison

- Cautions:
  - The approximation of D to a χ^2 distributed statistic is poor for small n_i and "lumpy" distribution of n_i as well.
  - Most texts recommend $n_i \geq 5$, $\forall i$, but this is just a rule-of-thumb.
  - We could also have done a Wald test on temperature:

```
            Estimate Std. Error z value Pr(>|z|)
temp         -0.2162     0.0532   -4.07  4.8e-05
```

    but differences of deviances are usually more accurate than tests on a single deviance.
  - When Wald provides significant results but a deviance comparison doesn't (the Hauck-Donner effect).

Binomial Model Comparison

- Confidence interval for the j th coefficient: $\hat{\beta}_j \pm z^{\alpha/2} se(\hat{\beta}_j)$.
- Low-tech method:

```
c(-0.2162-1.96*0.0532,-0.2162+1.96*0.0532)
-0.32047 -0.11193
```

- Hi-tech method:

```
summary(logitmod)$coefficients[,1]
        + qnorm(0.975) * t(c(-1,1) %o% summary(logitmod)$coefficients[,2])
(Intercept)  5.20243 18.12355
temp        -0.32046 -0.11201
```

- Profile likelihood version (accounts for covariance):

```
library(MASS)
confint(logitmod)
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept)  5.57520 18.73760
temp        -0.33266 -0.12018
```

Real Example: Model of Vote Choice 1994 American National Election Study

	Parameter Estimate	Standard Error	z-statistic	p-value
Choice Parameters				
Intercept	-1.116	0.387	-2.882	0.004
Democratic Support for Clinton	-0.015	0.008	-1.943	0.052
Republican Support for Clinton	0.030	0.011	2.701	0.007
Democratic Crime Concern	0.044	0.009	4.960	0.000
Republican Crime Concern	0.007	0.009	0.699	0.485
Democratic Gvt. Help Disadv.	0.029	0.011	2.698	0.007
Republican Gvt. Help Disadv.	-0.006	0.013	-0.438	0.661
Democratic Gvt. Spending	0.114	0.025	4.633	0.000
Republican Gvt. Spending	-0.100	0.025	-4.030	0.000
Democratic Federal Healthcare	0.031	0.008	3.670	0.000
Republican Federal Healthcare	-0.017	0.010	-1.691	0.091
Democratic Ideology Entropy	0.104	0.131	0.794	0.427
Republican Ideology Entropy	0.303	0.068	4.437	0.000
Party Identification Scale	0.368	0.028	13.158	0.000

Goodness of Fit Test: $LRT = 359.3869, p < 0.0001$ for $\chi^2_{df=19}$
Percent Correctly Classified: 78.66% (using the "naive criteria")

New Example in Chapter 14

- Estimating state-level opinions from national polls correcting for non-response at the group (state) level.
- Data come from a 1988 CBS News poll with random digit dialing (RDD) across 51 groups.
- Two steps: fit the multilevel model for all groups, then fit group-level predictions: MRP = Multilevel Regression + Poststratification (Gelman and Little 1997, Gelman, Park, Bafumi 2004, 2006).
- Data Details:
  - Outcome is the binary choice: $y = 1$ for Republican vote, $y = 0$ for Democratic vote.
  - It is assumed that there is no binomial overdispersion.
  - θ_ℓ = average response for each cross-classification of state and categorical demographics, sex (male, female), race (African American, other), age (4 categories), education (4 categories).
  - Therefore $\ell = 1, \ldots, L = 2 \times 2 \times 4 \times 4 \times 51 = 3264$, but we can reduce this by keeping states separated.
  - Cross-classification example: male, African American, age group 2, education group 4, in New York.

Primary Quantity of Interest

- Define N_ℓ as the number of survey respondents in category ℓ, from national demographics (census data).
- Note that this *not* the number of respondents in each category from the survey, which we could label n_ℓ.
- The estimated population average of y in state j is:

$$\hat{\theta}_j = \frac{\sum_{\ell \in j} N_\ell \theta_\ell}{\sum_{\ell \in j} N_\ell}$$

  where the summation is over the $\ell = 2 \times 2 \times 4 \times 4 = 64$ demographic categories in state j.
- This weighting by population average is currently called *poststratification*.
- We do this because some categories in some states will be small or empty and we want to weight accordingly.
- Another cross-classification example: male, African American, age group 2, education group 4, in Wyoming.

Primary Quantity of Interest

- A non-multilevel model cannot handle this issue directly, because the state-level effects are not treated at a different level of hierarchy.
- The key quantity of interest is the average value of y within each of the ℓ poststratification categories, which should be labeled y_ℓ.
- MRP cannot fix a bad model.
- If you randomly choose a survey question from a survey and randomly choose any state-level predictor, there is a good possibility that your predictions will be wrong: you have to have a *reason* for picking that predictor.
- Note also that large sample sizes, typical of national surveys, are necessary to be unbiased.
- The standard alternative is disaggregating national survey data to the state level and computing means within.

The Simplest Model

- Specify:

$$p(y_i = 1) = \text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}$$

- Which for two covariates means:

$$p(y_i = 1) = \text{logit}^{-1}(\alpha_{j[i]} + \beta^{\text{female}} \cdot \text{female} + \beta^{\text{black}} \cdot \text{black}), \qquad i = 1, \ldots, n$$

$$\alpha_j \sim N(\mu_\alpha, \sigma^2_{\text{state}}), \qquad j = 1, \ldots, 51.$$

- Data setup (Gelman):

```
lapply(c("lme4","arm"),library, character.only=TRUE)
data (state)                    # "state" IS AN R DATA FILE
state.abbr <- c (state.abb[1:8], "DC", state.abb[9:50])
dc <- 9
not.dc <- c(1:8,10:51)
region <- c(3,4,4,3,4,4,1,1,5,3,3,4,4,2,2,2,2,3,3,1,1,1,2,2,3,2,4,2,4,
            1,1,4,1,3,2,2,3,4,1,1,3,2,3,3,4,1,3,4,1,2,4)
```

More Data Setup

- Get the file from: http://pages.wustl.edu/files/pages/imce/jgill/polls.dta.

```
library(foreign)
polls <- read.dta("Class.Multilevel/Archive/examples/election88/polls.dta")
attach(polls)

# SELECT JUST THE DATA FROM THE LAST SURVEY (#9158)
table (survey)                  # look at the survey id's

survey
 9152  9153  9154  9155 9156a 9156b  9157  9158
 1611  1653  1833  1943   684  1478  2149  2193

ok <- survey==9158              # define the condition
polls.subset <- polls[ok,]      # select the subset of interest
detach(polls)
```

More Data Setup

```
print (polls.subset[1:5,])
         org year survey bush state edu age female black weight
11352 cbsnyt    7   9158   NA     7   3   1      1     0    923
11353 cbsnyt    7   9158    1    39   4   2      1     0    558
11354 cbsnyt    7   9158    0    31   2   4      1     0    448
11355 cbsnyt    7   9158    0     7   3   1      1     0    923
11356 cbsnyt    7   9158    1    33   2   2      1     0    403

# CREATE CASEWISE DELETED DATASET(!)
polls.subset.delete <- NULL
for (i in 1:nrow(polls.subset))
    if ( sum(is.na(polls.subset[i,])) == 0 )
        polls.subset.delete <- rbind(polls.subset.delete,polls.subset[i,])
y <- polls.subset.delete$bush
```

The Model

```
M1 <- glmer(y ~ black + female + (1 | state), family=binomial(link="logit"),
      data=polls.subset.delete)
summary(M1)

Formula: y ~ black + female + (1 | state)
   Data: polls.subset
  AIC  BIC logLik deviance
 2667 2689  -1329     2659
Random effects:
 Groups Name        Variance Std.Dev.
 state  (Intercept) 0.169    0.411
Number of obs: 2015, groups: state, 49

Fixed effects:                                    Correlation of Fixed Effects:
            Estimate Std. Error z value Pr(>|z|)        (Intr) black
(Intercept)   0.4452     0.1009    4.41  1.0e-05     black  -0.114
black        -1.7416     0.2080   -8.37  < 2e-16     female -0.551 -0.006
female       -0.0970     0.0946   -1.03     0.31
```

Notes On the Shorter Model

- We do not get σ_y^2 because it is fixed in the logit model at 1.6 (roughly $\pi/\sqrt{3}$) to identify the scale, whereas probit is fixed at 1.0 (see `http://andrewgelman.com/2006/06/take_logit_coef/`).
- We can explicitly create an analagous variance component once we're modeling in `bugs` /`jags` .
- The 51 coefficients estimate vectors by state are given by:

```
coef(M1)
$state
   (Intercept)   black    female    (Intercept)   black    female    (Intercept)   black    female
1    0.9905732 -1.7416 -0.097046 3    0.6861961 -1.7416 -0.097046 4    0.3149191 -1.7416 -0.097046
5    0.3064689 -1.7416 -0.097046 6    0.4050408 -1.7416 -0.097046 7    0.5254054 -1.7416 -0.097046
8    0.2079747 -1.7416 -0.097046 9    0.3516474 -1.7416 -0.097046 10   0.5550147 -1.7416 -0.097046
11   0.6803185 -1.7416 -0.097046 13   0.2466995 -1.7416 -0.097046 14   0.1273424 -1.7416 -0.097046
15   0.6035602 -1.7416 -0.097046 16  -0.0026701 -1.7416 -0.097046 17   0.7726262 -1.7416 -0.097046
18   0.5872641 -1.7416 -0.097046 19   0.5910221 -1.7416 -0.097046 20   0.2515000 -1.7416 -0.097046
21  -0.1121011 -1.7416 -0.097046 22  -0.0427777 -1.7416 -0.097046 23   0.2749340 -1.7416 -0.097046
24  -0.0275582 -1.7416 -0.097046 25   0.8466771 -1.7416 -0.097046 26   0.3433893 -1.7416 -0.097046
27   0.3080935 -1.7416 -0.097046 28   0.4347462 -1.7416 -0.097046 29   0.5339232 -1.7416 -0.097046
30   0.4746324 -1.7416 -0.097046 31   0.4785591 -1.7416 -0.097046 32   0.2224769 -1.7416 -0.097046
33  -0.0155143 -1.7416 -0.097046 34   0.8088724 -1.7416 -0.097046 35   0.4100830 -1.7416 -0.097046
36   0.7300007 -1.7416 -0.097046 37   0.5490738 -1.7416 -0.097046 38  -0.0097895 -1.7416 -0.097046
39   0.2640775 -1.7416 -0.097046 40   0.1995630 -1.7416 -0.097046 41   0.8028206 -1.7416 -0.097046
42   0.3687074 -1.7416 -0.097046 43   1.0934871 -1.7416 -0.097046 44   0.5629053 -1.7416 -0.097046
45   0.9725737 -1.7416 -0.097046 46   0.5676579 -1.7416 -0.097046 47   0.9125061 -1.7416 -0.097046
48   0.4002012 -1.7416 -0.097046 49   0.4215526 -1.7416 -0.097046 50   0.1547878 -1.7416 -0.097046
51   0.5676579 -1.7416 -0.097046
```

Notes On the Shorter Model

- The output from `display` and `summary` gives *model*, but not *null*, deviance:

```
AIC = 2666.7, DIC = 2658.7
deviance = 2658.7
```

```
     AIC  BIC logLik deviance
    2667 2689  -1329     2659
```

- This is rectified by running a null model:

```
M0 <- glmer(y ~ (1 | state), family=binomial(link="logit"),
            data=polls.subset.delete)
summary(M0)
  AIC  BIC logLik deviance
 2749 2760  -1373     2745
Random effects:
 Groups Name        Variance Std.Dev.
 state  (Intercept) 0.13599  0.36877
Number of obs: 2015, groups: state, 49

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26229    0.07746   3.386 0.000709
```

More On the Shorter Model

- So the deviance comparison comes from comparing the model:

```
 AIC  BIC logLik deviance
2667 2689  -1329     2659
```

to the null model:

```
 AIC  BIC logLik deviance
2749 2760  -1373     2745
```

- This is mechanically done by:

```
pchisq(2745-2659, df=2, lower.tail=FALSE)
[1] 2.1151e-19
```

- Note also that the output of `coef()` is a list so the matrix needs to be pulled out with:

```
dim(coef(M1)[[1]])
[1] 49  3
```

(we lost Alaska and Idaho in the casewise deleting process).

A Fuller Model with Non-Nested Factors

- Now include hierarchies for four factors and some interactions:

$$p(y_i = 1) = \text{logit}^{-1}\Big(\beta^0 + \beta^{\text{female}} \cdot \text{female}_i + \beta^{\text{black}} \cdot \text{black}_i + \beta^{\text{female.black}} \cdot \text{female}_i \cdot \text{black}_i + \alpha^{\text{age}}_{k[i]} + \alpha^{\text{edu}}_{\ell[i]} + \alpha^{\text{age.edu}}_{k[i],\ell[i]} + \alpha^{\text{state}}_{j[i]}\Big)$$

$$\alpha^{\text{state}}_j \sim N(\alpha^{\text{region}}_{m[j]} + \beta^{\text{v.prev}} \cdot \text{v.prev}_j, \sigma^2_{\text{state}})$$

- Notice that these are 4 *non-nested* hierarchies.
- The remaining coefficients are modeled as:

$$\alpha^{\text{age}}_k \sim N(0, \sigma^2_{\text{age}}), \qquad \text{for } k = 1, \ldots, 4$$

$$\alpha^{\text{edu}}_\ell \sim N(0, \sigma^2_{\text{edu}}), \qquad \text{for } \ell = 1, \ldots, 4$$

$$\alpha^{\text{age.edu}}_{k,\ell} \sim N(0, \sigma^2_{\text{age.edu}}), \qquad \text{for } k = 1, \ldots, 4, \ell = 1, \ldots, 4$$

$$\alpha^{\text{region}}_m \sim N(0, \sigma^2_{\text{region}}), \qquad \text{for } m = 1, \ldots, 5$$

A Fuller Model with Non-Nested Factors, Data Setup

```
v.prev=c(0.57,0.63,0.61,0.54,0.54,0.58,0.57,0.53,0.15,0.61,0.52,
         0.52,0.66,0.54,0.58,0.52,0.60,0.53,0.58,0.55,0.50,0.49,
         0.55,0.51,0.58,0.53,0.57,0.67,0.60,0.63,0.56,0.55,0.52,
         0.57,0.61,0.55,0.60,0.54,0.53,0.48,0.57,0.60,0.55,0.59,
         0.69,0.56,0.61,0.55,0.49,0.53,0.63)

attach(polls.subset.delete)
n.edu <- 4                          # 4 CATEGORIES OF EDUCATION
age.edu <- n.edu*(age-1) + edu      # 4 times (age-1) + edu
region.full <- region[state]        # 2193 LONG DATA-LEVEL VECTOR FOR REGIONS
v.prev.full <- v.prev[state]        # 2193 LONG DATA-LEVEL VECTOR FOR PREVIOUS VOTE
black.female <- black*female       # CREATES AN INTERACTION VARIABLE
detach(polls.subset.delete)
```

A Fuller Model with Non-Nested Factors, Estimation

```
M2 <- glmer(y ~ black + female + black.female + v.prev.full + (1 | age) + (1 | edu)
              + (1 | age.edu) + (1 | state) + (1 | region.full),
          data=polls.subset.delete,family=binomial(link="logit"))
summary(M2)
```

```
Formula: y ~ black + female + black.female + v.prev.full + (1 | age) +
             (1 | edu) + (1 | age.edu) + (1 | state) + (1 | region.full)
   Data: polls.subset.delete
  AIC  BIC logLik deviance
 2650 2706  -1315     2630
Random effects:
 Groups      Name        Variance   Std.Dev.
 state       (Intercept) 3.9330e-02 1.9832e-01
 age.edu     (Intercept) 2.2414e-02 1.4971e-01
 region.full (Intercept) 3.1180e-02 1.7658e-01
 edu         (Intercept) 1.1169e-02 1.0568e-01
 age         (Intercept) 1.0243e-09 3.2004e-05
Number of obs: 2015, groups: state, 49; age.edu, 16; region.full, 5; edu, 4; age, 4
```

A Fuller Model with Non-Nested Factors, Estimation

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.45146    0.98249  -3.513 0.000443
black        -1.63303    0.32445  -5.033 4.82e-07
female       -0.09002    0.09784  -0.920 0.357503
black.female -0.17916    0.41956  -0.427 0.669369
v.prev.full   6.96836    1.75620   3.968 7.25e-05

Correlation of Fixed Effects:
            (Intr) black  female blck.f
black       -0.026
female      -0.053  0.181
black.femal  0.023 -0.764 -0.233
v.prev.full -0.990  0.009 -0.006 -0.009
```

These Results Can Be Hard To Summarize In Detail

```
coef(M2)
$state
   (Intercept)  black    female black.female v.prev.full
1      -3.2988 -1.633 -0.089996     -0.17918      6.9682
3      -3.4214 -1.633 -0.089996     -0.17918      6.9682
4      -3.5024 -1.633 -0.089996     -0.17918      6.9682
5      -3.4129 -1.633 -0.089996     -0.17918      6.9682
6      -3.4898 -1.633 -0.089996     -0.17918      6.9682
7      -3.4195 -1.633 -0.089996     -0.17918      6.9682
8      -3.5007 -1.633 -0.089996     -0.17918      6.9682
9      -3.4531 -1.633 -0.089996     -0.17918      6.9682
10     -3.6953 -1.633 -0.089996     -0.17918      6.9682
11     -3.3341 -1.633 -0.089996     -0.17918      6.9682
13     -3.5220 -1.633 -0.089996     -0.17918      6.9682
14     -3.5321 -1.633 -0.089996     -0.17918      6.9682
15     -3.3935 -1.633 -0.089996     -0.17918      6.9682
16     -3.5560 -1.633 -0.089996     -0.17918      6.9682
17     -3.3592 -1.633 -0.089996     -0.17918      6.9682
18     -3.4007 -1.633 -0.089996     -0.17918      6.9682
```

19	-3.4759	-1.633	-0.089996	-0.17918	6.9682
20	-3.4970	-1.633	-0.089996	-0.17918	6.9682
21	-3.5688	-1.633	-0.089996	-0.17918	6.9682
22	-3.4978	-1.633	-0.089996	-0.17918	6.9682
23	-3.4909	-1.633	-0.089996	-0.17918	6.9682
24	-3.5443	-1.633	-0.089996	-0.17918	6.9682
25	-3.3692	-1.633	-0.089996	-0.17918	6.9682
26	-3.4208	-1.633	-0.089996	-0.17918	6.9682
27	-3.4853	-1.633	-0.089996	-0.17918	6.9682
28	-3.5423	-1.633	-0.089996	-0.17918	6.9682
29	-3.4384	-1.633	-0.089996	-0.17918	6.9682
30	-3.4526	-1.633	-0.089996	-0.17918	6.9682
31	-3.4040	-1.633	-0.089996	-0.17918	6.9682
32	-3.5080	-1.633	-0.089996	-0.17918	6.9682
33	-3.5556	-1.633	-0.089996	-0.17918	6.9682
34	-3.3694	-1.633	-0.089996	-0.17918	6.9682
35	-3.4761	-1.633	-0.089996	-0.17918	6.9682
36	-3.2177	-1.633	-0.089996	-0.17918	6.9682
37	-3.4895	-1.633	-0.089996	-0.17918	6.9682
38	-3.5627	-1.633	-0.089996	-0.17918	6.9682
39	-3.4199	-1.633	-0.089996	-0.17918	6.9682

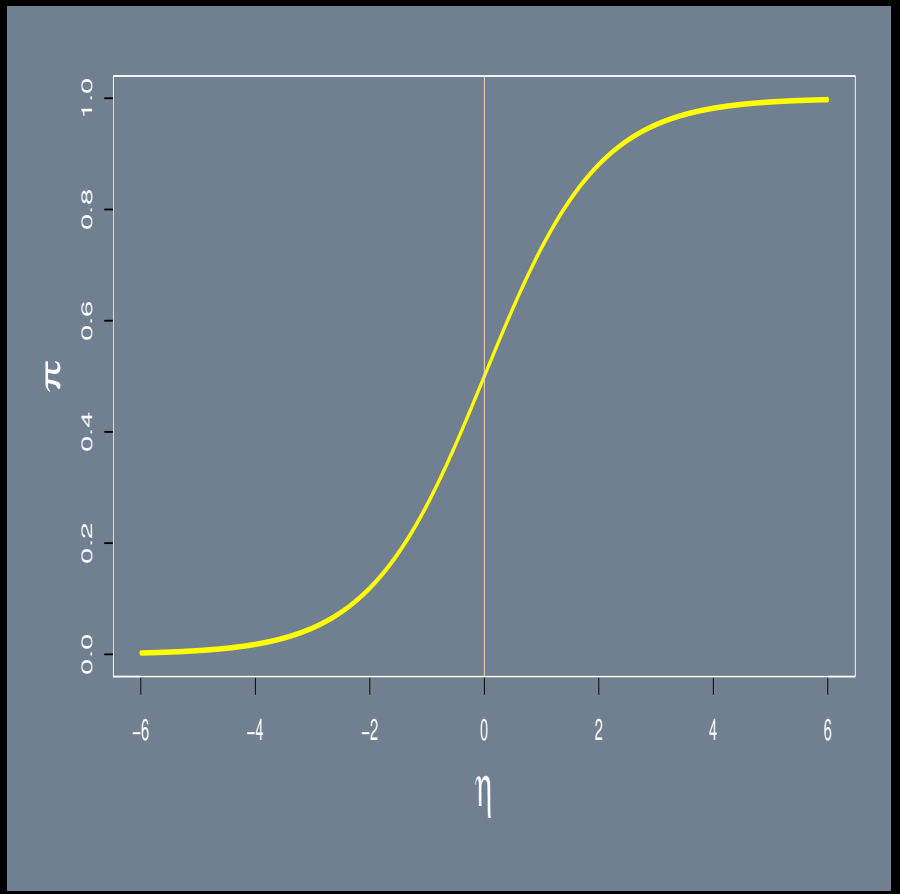
```
40     -3.4622 -1.633 -0.089996     -0.17918      6.9682
41     -3.3786 -1.633 -0.089996     -0.17918      6.9682
42     -3.4786 -1.633 -0.089996     -0.17918      6.9682
43     -3.2179 -1.633 -0.089996     -0.17918      6.9682
44     -3.6143 -1.633 -0.089996     -0.17918      6.9682
45     -3.3620 -1.633 -0.089996     -0.17918      6.9682
46     -3.4178 -1.633 -0.089996     -0.17918      6.9682
47     -3.4213 -1.633 -0.089996     -0.17918      6.9682
48     -3.4367 -1.633 -0.089996     -0.17918      6.9682
49     -3.3320 -1.633 -0.089996     -0.17918      6.9682
50     -3.5036 -1.633 -0.089996     -0.17918      6.9682
51     -3.4297 -1.633 -0.089996     -0.17918      6.9682

$age.edu
   (Intercept)  black    female black.female v.prev.full
1      -3.4929 -1.633 -0.089996     -0.17918      6.9682
2      -3.3408 -1.633 -0.089996     -0.17918      6.9682
3      -3.3491 -1.633 -0.089996     -0.17918      6.9682
4      -3.4286 -1.633 -0.089996     -0.17918      6.9682
5      -3.3884 -1.633 -0.089996     -0.17918      6.9682
6      -3.6042 -1.633 -0.089996     -0.17918      6.9682
```

```
7      -3.4658 -1.633 -0.089996     -0.17918      6.9682
8      -3.5248 -1.633 -0.089996     -0.17918      6.9682
9      -3.4256 -1.633 -0.089996     -0.17918      6.9682
10     -3.4184 -1.633 -0.089996     -0.17918      6.9682
11     -3.4118 -1.633 -0.089996     -0.17918      6.9682
12     -3.4401 -1.633 -0.089996     -0.17918      6.9682
13     -3.6582 -1.633 -0.089996     -0.17918      6.9682
14     -3.4803 -1.633 -0.089996     -0.17918      6.9682
15     -3.3920 -1.633 -0.089996     -0.17918      6.9682
16     -3.4076 -1.633 -0.089996     -0.17918      6.9682

$region.full
  (Intercept)  black    female black.female v.prev.full
1     -3.5383 -1.633 -0.089996     -0.17918      6.9682
2     -3.5283 -1.633 -0.089996     -0.17918      6.9682
3     -3.2117 -1.633 -0.089996     -0.17918      6.9682
4     -3.5324 -1.633 -0.089996     -0.17918      6.9682
5     -3.4528 -1.633 -0.089996     -0.17918      6.9682

$edu
  (Intercept)  black    female black.female v.prev.full
```

```
1     -3.5308 -1.633 -0.089996     -0.17918      6.9682
2     -3.4703 -1.633 -0.089996     -0.17918      6.9682
3     -3.3581 -1.633 -0.089996     -0.17918      6.9682
4     -3.4490 -1.633 -0.089996     -0.17918      6.9682

$age
  (Intercept)  black    female black.female v.prev.full
1     -3.4515 -1.633 -0.089996     -0.17918      6.9682
2     -3.4515 -1.633 -0.089996     -0.17918      6.9682
3     -3.4515 -1.633 -0.089996     -0.17918      6.9682
4     -3.4515 -1.633 -0.089996     -0.17918      6.9682
```

The Divide-By-4 Rule

- The logistic curve is steepest at its center, where $\alpha + \beta x = 0$, and $\text{logit}^{-1}(\alpha + \beta x) = 0.5$ (to generalize, see Nagler's Scobit model).



The Divide-By-4 Rule

- The slope at this inflection point is the biggest of anywhere on the curve, solving:

$$\begin{aligned}\frac{d}{dx}\left[1+\exp(x\beta)\right]^{-1} &= \left[1+\exp(x\beta)\right]^{-2}(-1)\exp(x\beta)\beta \\ &= \left.\frac{-\exp(x\beta)\beta}{(1+\exp(x\beta))^2}\right|_{x=0} \\ &= -\frac{\beta}{4}.\end{aligned}$$

- Therefore 4 is the maximum change in $p(Y=1)$ for a one-unit change in x, since this change cannot exceed 1 in absolute value (the sign isn't important here).
- So we can take logistic regression coefficients (other than the constant term) and divide them by 4 to get an upper bound of the predictive difference corresponding to a unit difference in x.
- Assumptions: we are near the middle on the x-axis, and there are no large covariances between model coefficients.

Gelman & Hill Observations

- The intercept,

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.45146    0.98249  -3.513 0.000443
```

  is not easily interpretable since it corresponds to a case in which black, female, and v.prev are all 0, but `v.prev` typically takes on values near 0.5 and is never 0.

- The coefficient for black is:

```
             Estimate Std. Error z value Pr(>|z|)
black        -1.63303    0.32445  -5.033 4.82e-07
```

  Dividing by 4 (see page 82) yields a rough estimate that African-American men were 40% less likely than other men to support Bush, after controlling for age, education, and state.

- The coefficient for female is:

```
             Estimate Std. Error z value Pr(>|z|)
female       -0.09002    0.09784  -0.920 0.357503
```

  Dividing by 4 shows that non-African-American women were very slightly less likely than non-African-American men to support Bush, after controlling for age, education, and state, although the standard error does not let us make this claim.

Gelman & Hill Observations

- The large standard error on the coefficient for black:female,

```
             Estimate Std. Error z value Pr(>|z|)
black.female -0.17916    0.41956  -0.427 0.669369
```

  indicates that the sample size is too small to estimate this interaction precisely.

- The coefficient for v.prev.full is:

```
             Estimate Std. Error z value Pr(>|z|)
v.prev.full   6.96836    1.75620   3.968 7.25e-05
```

  which, when divided by 4, is 1.7, suggesting that a 1% increase in a state's support for Republican candidates in the previous election mapped to a predicted 1.7% difference in support for Bush in 1988.

Gelman & Hill Observations

- The state-level errors have an estimated standard deviation of roughly 0.2 on the logit scale:

```
Groups       Name        Variance   Std.Dev.
state        (Intercept) 3.9330e-02 1.9832e-01
```

  Dividing by 4 tells us that the states differed by approximately ±5% on the probability scale (over and above the differences explained by demographic factors).

- The differences among age-education groups and regions are also approximately ±5% on the probability scale:

```
Groups       Name        Variance   Std.Dev.
age.edu      (Intercept) 2.2414e-02 1.4971e-01
region.full (Intercept) 3.1180e-02 1.7658e-01
```

- Very little variation is found among age groups or education groups after controlling for the other predictors in the model.

```
Groups       Name        Variance   Std.Dev.
edu          (Intercept) 1.1169e-02 1.0568e-01
age          (Intercept) 1.0243e-09 3.2004e-05
```

Using the Model Inferences to Estimate Average Opinion For Each State

- The model gives the probability that any adult will prefer Bush, given the ethnicity, age, education level, and state of the person.
- We can now compute weighted averages of these probabilities to represent the proportion of Bush supporters in any specified subset of the population: *poststratification*.
- We first extract from the U.S. Census the counts N_ℓ in each of the 3264 crossclassification cells and create a 3264×6 data frame, census, indicating the sex ethnicity, age, education, state, and number of people.
- This corresponds each of corresponding to each cell according to our data:

	org	year	survey	bush	state	edu	age	female	black	weight
11353	cbsnyt	7	9158	1	39	4	2	1	0	558
11354	cbsnyt	7	9158	0	31	2	4	1	0	448
11355	cbsnyt	7	9158	0	7	3	1	1	0	923
11356	cbsnyt	7	9158	1	33	2	2	1	0	403
11357	cbsnyt	7	9158	1	33	4	4	1	0	317
11358	cbsnyt	7	9158	1	39	2	2	0	0	1532
11359	cbsnyt	7	9158	1	20	2	4	1	0	896
:										

Using the Model Inferences to Estimate Average Opinion For Each State

- The estimated population average of y in state j is:

$$\hat{\theta}_j = \frac{\sum_{\ell \in j} N_\ell \theta_\ell}{\sum_{\ell \in j} N_\ell}$$

  where the summation is over the $\ell = 2 \times 2 \times 4 \times 4 = 64$ demographic categories in state

- Then compute y^{pred} for each category.

- Our output will be $p(y_i) = 1$, meaning the probability that one of these distinct cases votes for Bush.

Using the Model Inferences to Estimate Average Opinion For Each State

```
# CREATE A GRID OF ALL POSSIBLE CROSSES
polls.X <- expand.grid(state=1:51,edu=1:4,age=1:4,female=0:1,black=0:1)
rbind(polls.X[1:7,],polls.X[(nrow(polls.X)-6):nrow(polls.X),]) #ILLUSTRATION
     state edu age female black
1        1   1   1      0     0
2        2   1   1      0     0
3        3   1   1      0     0
4        4   1   1      0     0
5        5   1   1      0     0
6        6   1   1      0     0
7        7   1   1      0     0
3258    45   4   4      1     1
3259    46   4   4      1     1
3260    47   4   4      1     1
3261    48   4   4      1     1
3262    49   4   4      1     1
3263    50   4   4      1     1
3264    51   4   4      1     1
```

Using the Model Inferences to Estimate Average Opinion For Each State

```
polls.X[1,]           state edu age female black
                  1     1   1   1      0     0

# LOOK AT ALABAMA
coef(M2)$state[1,]
  (Intercept)  black    female black.female v.prev.full
1     -3.2988 -1.633 -0.089996     -0.17918      6.9682

# THIS IS THE OFFSET FROM THE US MEAN, BUT WE'LL DIRECTLY USE THE INTERCEPT
ranef(M1)$state[1,]          [1] 0.54534

# MRP PREDICTION OF THE CASE DESCRIBED BY polls.X[1,]
( y1.pred <- inv.logit( coef(M2)$state[1,1]
                + polls.X[1,5]*coef(M2)$state[1,]$black
                + polls.X[1,4]*coef(M2)$state[1,]$female
                + polls.X[1,4]*polls.X[1,5]*coef(M2)$state[1,]$black.female
                + v.prev[1]*coef(M2)$state[1,]$v.prev.full ) )
[1] 0.72225
```

Non-Nested Overdispersed Model for Death Sentence Reversals

- We can modify models in this chapter to account for outcomes that are proportions.
- So y_i is the probability of a success, not the observation of a success.
- Or y_i is the number of successes out of n_i attempts.
- Section 6.3 describes data where the outcome variable is *the number of death penalty sentences per state that are reversed by a higher court.*
- G&H create a non-nested model for state, $j = 1, \ldots, J$, and year, $t = 1, \ldots, T$, coefficients.

Non-Nested Overdispersed Model for Death Sentence Reversals

- Explanatory variables: frequency that the death sentence was imposed, the backlog of capital cases in the appeals courts, the level of political pressure on judges, indicators for the years from 1973 to 1995 for the 34 states (all of those in this time span that had death penalty laws).
- The regression model with all these predictors is:

$$y_i \sim \text{Bin}(n_i, p_i)$$
$$p_i = \text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta} + \alpha_{j[i]} + \gamma_{t[i]})$$

  where j indexes states and t indexes years.

- The necessary distributions for the state and year coefficients are:

$$\alpha_j \sim N(0, \sigma_\alpha^2)$$
$$\gamma_t \sim N(a + bt, \sigma_\gamma^2)$$

- The coefficients for year are include as a linear time trend to capture the overall increase in reversal rates over the time of the study.

Non-Nested Overdispersed Model for Death Sentence Reversals

- The model for the γ_t hierarchy includes an intercept, and so we do not need to include a constant term in the hierarchy for α_j or at the indvidual level.
- The multilevel structure here is just to take care of data heterogeneity and get better estimates for the $\boldsymbol{\beta}$ terms.
- There is an overdispersion problem for these data, with this model:
  - ▷ Create the standardized residuals:

$$z_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)/n_i}}$$

    where $\hat{p}_i = \text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta} + \alpha_{j[i]} + \gamma_{t[i]})$.

  - ▷ Under the binomial assumptions $z_i \sim N(0,1)$, but a better test comes from:

$$\sum_{i=1}^{n} z_i^2 \sim \chi^2_{df=(T\times J)-k}$$

    where $T \times J = 520$, and k is the number of explanatory variables including the constant ($T \times J$ is less than $23 \times 34 = 782$ because not all states had death penalty laws in all years).

Non-Nested Overdispersed Model for Death Sentence Reversals

- Fix #1: use the beta-binomial distribution instead of the binomial distribution:

$$y_i \sim \text{beta-binomial}(n_i, p_i, \omega),$$

  where $\omega \geq 1$ is the overdispersion parameter and the model with $\omega = 1$ reduces to the binomial.

- In `R`, use the `glm` function with `quasibinomial(link="logit")`

- Fix #2: use the binomial-normal model instead of the binomial distribution by adding normally distributed errors on the logistic scale:

$$p_i = \text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta} + \alpha_{j[i]} + \gamma_{t[i]} + \xi_i)$$
$$\xi_i = N(0, \sigma_\xi^2)$$

  where the model reduces to the binomial when $\sigma_\xi^2 = 0$.

- Generally this has to be done with `bugs` /`jags` .

Non-Nested Overdispersed Model for Death Sentence Reversals

- With moderate sample sizes, it is typically difficult to distinguish between the beta-binomial and binomial-normal models.
- The beta-binomial model adds only one new parameter and so it can be easier to fit
- The binomial-normal model has the advantage that the new error term, ξ_i, is on the same scale as the group-level predictors, $\alpha_{j[i]}$ and $\gamma_{t[i]}$ which can make the fitted model easier to explain.

Overdispersed Poisson Regression

- Data that are fit with a GLM that have greater variance than assumed by the model used are called overdispersed.
- The binomial and Poisson regression models lack a natural parameter to deal with overdispersion so one must be added.
- Chapter 15 focuses on the *police stops data* from chapter 6, except that the data are not made available to us.
- So instead consider the data from: Koch, M. T. Cranmer, S. (2007). "Testing the 'Dick Cheney' Hypothesis: Do Governments of the Left Attract More than Goverments of the Right?" *Conflict Management and Peace Science* **24**, 311-326.

Data Description

- In this example we look at terrorist activity in 22 Asian democracies over 8 years (1990-1997) with data subsetted from Koch and Cranmer (2007), $n = 150$.
- The outcome of interest a count of violent terrorists attacks in a country/year pair, `ATT`.
- `DEM` for these countries is the Polity IV 21-point democracy scale ranging from -10 indicating a hereditary monarchy to +10 indicating a fully consolidated democracy.
- The variable `FED` is assigned zero if sub-national governments do not have substantial taxing, spending, and regulatory authority, and one otherwise.
- Governmental systems, `SYS`, coded as: (0) for direct presidential elections, (1) for strong president elected by assembly, and (2) dominant parliamentary government.
- `AUT` is a dichotomous variable indicating whether or not there are autonomous regions not directly controlled by central government.
- `LEF` is 1 if the government is coded left-of-center, 0 otherwise.
- `CID` is the country-identifying code.
- `SUM` is the total number of attacks by country over the period of study.

Poisson Regression with an Offset

- Generalize from the standard Poisson GLM:

$$y_i \sim \text{Poisson}(\theta_i), \qquad \theta_i = \exp(\mathbf{X}_i\boldsymbol{\beta}),$$

  to:

$$y_i \sim \text{Poisson}(u_i\theta_i), \qquad \theta_i = \exp(\mathbf{X}_i\boldsymbol{\beta}).$$

- Here u_i is the exposure, and $\log(u_i)$ is the offset.
- The idea is to scale the effect of the rate.
- Typical strategy: put the log of the exposure into the model as an offset which forces its regression coefficient to be 1.
- Here we will scale the count per case by the sum total.

Using an Offset

- We just modeled these as counts independent of the amount of exposure.
- But the counts are actually out of a number of cases exposed.
- This is called a rate model in the count literature: events per unit of exposed.
- Thus we want to put exposure on the RHS of the model, being careful about logs:

$$\log\left(\frac{E[Y|\boldsymbol{\beta},\mathbf{X}]}{\text{exposure}}\right) = \mathbf{X}\boldsymbol{\beta}$$

$$\log(E[Y|\boldsymbol{\beta},\mathbf{X}]) - \log(\text{exposure}) = \mathbf{X}\boldsymbol{\beta}$$

$$\log(E[Y|\boldsymbol{\beta},\mathbf{X}]) = \mathbf{X}\boldsymbol{\beta} + \log(\text{exposure})$$

$$E[Y|\boldsymbol{\beta},\mathbf{X}] = \exp\left[\mathbf{X}\boldsymbol{\beta} + \log(\text{exposure})\right]$$

  which justifies putting a log-constant on the RHS to reflect the number exposed in each case.

- In `R` this is done with the `offset()` specification.

Terrorism Count Data

```
library(foreign)
data_one <- read.dta("CheneyData.dta", convert.factors = TRUE)
# GET THE CONDENSED DATA SET (1990 TO 1997, ASIA ONLY, CASEWISE DELETED):
asia.sub.df <-
    read.table("https://pages.wustl.edu/files/pages/imce/jgill/cheney.asia_.sub_.txt",
    header=FALSE)
names(asia.sub.df) <- c("ATT","DEM","FED","SYS","AUT","LEF","CID","SUM")

table(asia.sub.df$ATT)
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 16 38
83 14  8 12 10  4  3  4  3  3  1  1  1  1  1  1

table(asia.sub.df$LEF)
 0  1
99 51
```

Terrorism Count Data, Simple Poisson GLM

```
asia.1 <- glm(ATT ~  DEM + FED + SYS + AUT + LEF, family=poisson, data=asia.sub.df)
summary(asia.1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7523     0.1126    6.68  2.4e-11
DEM           0.0600     0.0116    5.16  2.5e-07
FED          -0.5194     0.1944   -2.67   0.0075
SYS          -0.3213     0.0717   -4.48  7.4e-06
AUT           0.0027     0.3046    0.01   0.9929
LEF           0.6795     0.1399    4.86  1.2e-06
Null deviance: 763.32  on 149  degrees of freedom
Residual deviance: 707.82  on 144  degrees of freedom
AIC: 928.6

# ESTIMATED OVERDISPERSION
z <- (asia.sub.df$ATT - asia.1$fitted.values)/sd(asia.1$fitted.values)
pchisq(sum(z^2), df=asia.1$df.residual, lower.tail=FALSE)
[1] 0
```

Terrorism Count Data, Model With Log-Exposure

```
asia.2 <- glm(ATT ~  DEM + FED + SYS + AUT + LEF, offset=log1p(SUM), family=poisson,
              data=asia.sub.df)
summary(asia.2)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.024354   0.106984  -18.92  < 2e-16
DEM          0.000402   0.011437    0.04    0.972
FED         -0.436426   0.197849   -2.21    0.027
SYS         -0.100306   0.067730   -1.48    0.139
AUT         -0.572688   0.314415   -1.82    0.069
LEF          0.685259   0.135114    5.07  3.9e-07
Null deviance: 359.64  on 149  degrees of freedom
Residual deviance: 334.22  on 144  degrees of freedom
AIC: 555.1

# ESTIMATED OVERDISPERSION
z <- (asia.sub.df$ATT - asia.2$fitted.values)/sd(asia.2$fitted.values)
pchisq(sum(z^2), df=asia.2$df.residual, lower.tail=FALSE)
[1] 2.0001e-05
```

Terrorism Count Data, Modeling With Overdispersion

```
asia.3 <- glm(ATT ~  DEM + FED + SYS + AUT + LEF, offset=log1p(SUM),
              family=quasipoisson, data=asia.sub.df)
summary(asia.3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.024354   0.169716  -11.93   <2e-16
DEM          0.000402   0.018143    0.02   0.9823
FED         -0.436426   0.313862   -1.39   0.1665
SYS         -0.100306   0.107444   -0.93   0.3521
AUT         -0.572688   0.498779   -1.15   0.2528
LEF          0.685259   0.214340    3.20   0.0017

(Dispersion parameter for quasipoisson family taken to be 2.5166)

Null deviance: 359.64  on 149  degrees of freedom
Residual deviance: 334.22  on 144  degrees of freedom
```

A Multilevel Poisson Regression Model

- Add a hyperparameter σ_ϵ that measures the amount of overdispersion accounting for country as a group:

$$y_i \sim \text{Poisson}(u_i\theta_i), \qquad \theta_i = \exp(\mathbf{X}_i\boldsymbol{\beta} + \epsilon_j[i]), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

- We can use `CID` to create a country-level in the model

```
asia.4 <- glmer(ATT ~  DEM + FED + SYS + AUT + LEF + (1|CID), offset=log1p(SUM),
                family=poisson, data=asia.sub.df)
summary(asia.4)
```

- Model Summaries:

```
 AIC BIC logLik deviance
 345 366   -165      331
```

- Random Effects:

```
 Groups Name        Variance Std.Dev.
 CID    (Intercept) 0.0434   0.208
Number of obs: 150, groups: CID, 21
```

A Multilevel Poisson Regression Model

- Fixed Effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0933	0.1498	-13.97	< 2e-16
DEM	0.0160	0.0152	1.05	0.292
FED	-0.3973	0.2786	-1.43	0.154
SYS	-0.1755	0.0979	-1.79	0.073
AUT	-0.3795	0.4026	-0.94	0.346
LEF	0.7318	0.1500	4.88	1.1e-06

- Correlation of Fixed Effects:

	(Intr)	DEM	FED	SYS	AUT
DEM	-0.343				
FED	0.237	-0.210			
SYS	-0.620	-0.103	-0.395		
AUT	-0.170	0.402	0.102	-0.094	
LEF	-0.382	0.038	-0.401	0.147	-0.256

Observations

- This is a good example of why we need to move to `bugs` /`jags` .
- `lmer` gives (decent) approximations for hierarchical parameters.
- It also assumes that σ_y is constant in groups.
- We also do not get uncertainty estimates for terms like σ_α.
- `lmer` is a great learning tool and a good place to get rough estimates for more complicated models.
- Moving to `bugs` /`jags` also gives us more flexibility in defining hierarchies: third levels and beyond, non-normal distributional assumptions, and more.

Getting Started with the `bugs` Langauge

- Models in the `bugs` language are specified more theoretically than in other packages.
- We will follow G&H and use vague priors everywhere, but this is an important issue in some instances.
- Example from the assigned reading for next week: "Bayesian Analytical Methods: A Methodological Prescription for Public Administration."
- Data from the 1998 and 2004 rounds of the American State Administrator's Project (ASAP) survey.
- Our outcome variable of interest `grp.influence` is an index of the respondents' (senior executives') perceptions of the influence that clientele groups have on the total agency budget, specialized program budgets, and agency policies. Each of these questions is a seven-point scale and they have been summed to create a single outcome variable (ranging from 3 to 21).

Getting Started with the bugs Langauge

- Explanatory variables: contracting, gov.influence, leg.influence, elect.board, years.tenure, education, party.ID, category2, category3, category4, category5, category6, category7, category8, category9, category10, category11, category12, med.time, medt.contr, gov.ideology, lobbyists, nonprofits.
- Data for jags needs to be in list form:

```
asap.jags.list <- list(STATES <- 50, SUBJECTS <- 713,
    state.id <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...),
    contracting <- c(6, 2, 0, 0, 0, 1, 0, 3, 3, 6, 0, 1, 5, 1, 1, 1, 0,...),
    :
    nonprofits <- c(1.9783, 0.509, 2.0701, 1.3639, 15.6682, 2.7968,...)
)
```

- Or read this in from a file that you've constructed.

Getting Started with the `bugs` Langauge

- Markov chains require starting points, so create the following to modify:

```
"tau" <- 1
"tau.alpha" <- 1
"alpha" <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
             0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
             0,0,0,0,0,0,0,0,0,0,0,0)
"beta" <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
"gamma" <- c(0,0,0)
```

- Then make them different.

The First Part of the jags Code

```
model  {
  for (i in 1:SUBJECTS) {
    mu[i] <- alpha[state.id[i]]
      + beta[1]*contracting[i] + beta[2]*gov.influence[i] + beta[3]*leg.influence[i]
      + beta[4]*elect.board[i] + beta[5]*years.tenure[i]  + beta[6]*education[i]
      + beta[7]*party.ID[i]    + beta[8]*category2[i]     + beta[9]*category3[i]
      + beta[10]*category4[i]  + beta[11]*category5[i]    + beta[12]*category6[i]
      + beta[13]*category7[i]  + beta[14]*category8[i]    + beta[15]*category9[i]
      + beta[16]*category10[i] + beta[17]*category11[i]   + beta[18]*category12[i]
      + beta[19]*med.time[i]   + beta[20]*medt.contr[i]
    grp.influence[i] ~ dnorm(mu[i],tau)
  }
  for (j in 1:STATES) {
    eta[j]  <- gamma[1]*gov.ideology[j] + gamma[2]*lobbyists[j]
             + gamma[3]*nonprofits[j]
    alpha[j] ~ dnorm(eta[j],tau.alpha)
  }
```

The Second Part of the jags Code

```
  beta[1]    ~ dnorm(0.070,1)     # PRIOR MEANS FROM KELLEHER AND YACKEE 2009, MODEL 3
  beta[2]    ~ dnorm(-0.054,1)
  beta[3]    ~ dnorm(0.139,1)
  beta[4]    ~ dnorm(0.051,1)
  beta[5]    ~ dnorm(0.017,1)
  beta[6]    ~ dnorm(0.056,1)
  beta[7]    ~ dnorm(0.039,1)
  beta[8]    ~ dnorm(0.0,1)       # DIFFUSE PRIORS
  :
  beta[18]   ~ dnorm(0.0,1)
  beta[19]   ~ dnorm(0.184,1)     # PRIOR MEANS FROM KELLEHER AND YACKEE 2009, MODEL 3
  beta[20]   ~ dnorm(0.156,1)
  gamma[1]   ~ dnorm(0.0,1)       # DIFFUSE PRIORS
  gamma[2]   ~ dnorm(0.0,1)
  gamma[3]   ~ dnorm(0.0,1)
  tau        ~ dgamma(1.0,1)
  tau.alpha  ~ dgamma(1.0,1)
}
```

Running jags From R

```
# LOAD LIBRARY AND SOURCE FILES
library(rjags); library(arm); library(coda); library(superdiag)

# DEFINE THE MODEL
asap.model2.rjags  <- function()  {
  for (i in 1:SUBJECTS) {
  :
}

# SAVE MODEL TO A FILE
write.model(asap.model2.rjags, "Article.JPART/asap.model2.rjags")
```

Running jags From R

```
# RUN THE SAMPLER AND COLLECT coda SAMPLES
asap2.model <- jags.model(file="Article.JPART/asap.model2.rjags",
    inits=asap.inits, data=asap.jags.list, n.chains=3,  n.adapt=5000)
update(asap2.model, n.iter=2500)
asap2.mcmc <- coda.samples(model=asap2.model, variable.names=names(asap.jags.list),
    n.iter=2500)
summary(asap2.mcmc)

# CHECK CONVERGENCE
superdiag(as.mcmc.list(asap2.mcmc), burnin=0)

# GET THE DEVIANCE AND THE DIC
asap2.dic <- dic.samples(asap2.model, n.iter=25000, type="pD")
```

Exercises for Next Week

14.5. Multilevel logistic regression with non-nested groupings: the folder speed.dating contains data from an experiment on a few hundred students that randomly asssigned each participant to 10 short dates with participants of the opposite sex (Fisman et al., 2006). For each date, each person recorded several subjective numerical ratings of the other person (attractiveness, compatibility, and some other characteristics) and also wrote down whether he or she would like to meet the other person again. Label

$$y_{ij} = \begin{cases} 1 \text{ if person } i \text{ is interested in seeing person } j \text{ again} \\ 0 \text{ otherwise} \end{cases}$$

and $r_{ij1}, \ldots r_{ij6}$ as person i's numerical ratings of person j on the dimensions of attractiveness, compatibility, and so forth.

(a) Fit a classical logistic regression predicting $Pr(y_{ij} = 1)$ given person i's 6 ratings of person j. Discuss the importance of attractiveness, compatibility, and so forth in this predictive model.

(b) Expand this model to allow varying intercepts for the persons making the evaluation; that is, some people are more likely than others to want to meet someone again. Discuss the fitted model.

(c) Expand further to allow varying intercepts for the persons being rated. Discuss the fitted model.

Exercises for Next Week

14.6. Varying-intercept, varying-slope logistic regression: continuing with the speed- dating example from the previous exercise, you will now fit some models that allow the coefficients for attractiveness, compatibility, and the other attributes to vary by person.

(a) Fit a no-pooling model: for each person i, fit a logistic regression to the data y_ij for the 10 persons j whom he or she rated, using as predictors the 6 ratings $r_{ij1}, \ldots r_{ij6}$. (Hint: with 10 data points and 6 predictors, this model is difficult to fit. You will need to simplify it in some way to get reasonable fits.)

(b) Fit a multilevel model, allowing the intercept and the coefficients for the 6 ratings to vary by the rater i.

(c) Compare the inferences from the multilevel model in (b) to the no-pooling model in (a) and the complete-pooling model from part (a) of the previous exercise.

Exercises for Next Week

15.1. Multilevel ordered logit: using the National Election Study data from the year 2000 (data available in the folder nes), set up an ordered logistic regression predicting the response to the question on vote intention (0 = Gore, 1 = no opinion or other, 2 = Bush), given the predictors shown in Figure 5.4 on page 84, and with varying intercepts for states. (You will fit the model using Bugs in Exercise 17.10.)

15.2. Using the same data as the previous exercise:

(a) Formulate a model to predict party identification (which is on a five-point scale) using ideology and demographics with a multilevel ordered categorical model allowing both the intercept and the coefficient on ideology to vary over state.

(b) Fit the model using lmer() and discuss your results.