Stat 414 - Day 15
Multilevel Logistic Regression (Ch. 17)

Last Time: Logistic Regression
With a binary response variable, we can predict the probability of success using the logistic "link function" to create a linear relationship with the log-odds.

$$ln\big(\pi/(1-\pi)\big) = \beta_0 + \beta_1 x$$

where we assume a Binomial distribution with $E(Y) = n\pi$ and $Var(Y) = n\pi(1-\pi)$. Note that we have one parameter here π and not separate parameters for the mean and variance. Also note that the logistic models works just as well for binomial observations (response = number or proportions of successes) or Bernoulli observations (response = success or failure).

Example 1: Data were collected on 5,366 women who recently gave birth in Bangledesh. One question we can ask is whether mother's age ($mage$) predicts whether or not the mother receives prenatal care during pregnancy ($antemed$).
(a) Explain what props$angemed represents.

(b) What appears to be the association between mom's age and probability of prenatal care?

(c) Would a linear model appear appropriate? How are you deciding?

These observations were taken across 361 communities. Are there substantial community to community differences in the likelihood of receiving prenatal care?
(d) Is the association between "whether or not prenatal care" and "community" statistically significant?

Fit a model with a different intercept for each community
(e) Does the "effect" of mom's age change much when we added community to the model? What does this tell you?

(f) What is the predicted antemed when momage = 33 for the average community?

(g) What is the average predicted antemed when momage = 33 across these communities?

```
new_data <- data.frame(mage = 33, comm = levels(bang$comm))
predicted_values <- predict(m2, newdata = new_data)
mean(predicted_values)
```

Now let's fit a logistic model instead

```
model.glm = glm(antemed ~ mage, data = bang, family=binomial)
summary(model.glm)
```

Note, these models can get complicated to run and you will start to notice they take a few minutes. "Fisher scoring iterations" is one approach.

(h) Interpret the slope coefficient in context.

Now add the community variable to the model

```
model.glm2 =  glm(antemed ~ mage + comm, data = bang, family = binomial, contrasts
= list(comm = contr.sum))
```

(i) Interpret the slope of momage in context.

(j) What is the predicted probability of prenatal care for 33-year-old moms in the average community?

(k) What is the average (across the communities) predicted probability for 33-year-old moms? (Verify the predicted probability for a mom from community 1)

Definition: The *conditional effect* is for a particular group (e.g., the community) and the *marginal effect* is averaging across the groups (e.g., communities on average). For linear models, these are the same, but in nonlinear models, they differ and which one you want to examine may depend on the research question.

We do see the significant differences among communities, but again, the individual communities aren't my primary interest and that's a lot of coefficients to print out! Once again, we have the option of treating the community id as a random effect rather than a fixed effect. Here is the random intercepts model

$$ln\left(\pi_j/(1-\pi_j)\right) = \beta_0 + u_{0j} \text{ where } u_{0j} \sim N(0, \tau_0^2)$$

(l) Explain what u_{0j} and τ_0^2 represent in this context.

We will use the "glmer" function (in lme4 package) to fit multilevel logistic regression models.

```
#The random intercepts model
model0 = glmer(antemed~ 1 + (1 | comm), family=binomial, data = bang)
summary(model0); confint(model0)
#can use fitted.values to see the (back-transformed) predicted probabilities. Note
how we assume the same probability for every woman in the same community
head(fitted.values(model0), 20)
```

(m) Interpret the intercept in context.

Notice that we are only given an estimate for τ_0 but not σ. That's because there is no separate "within community variation" parameter in logistic regression. (We are assuming the same odds for each woman within the same community.) This has led to different suggestions for calculating the **intraclass correlation coefficient**. I'm partial to

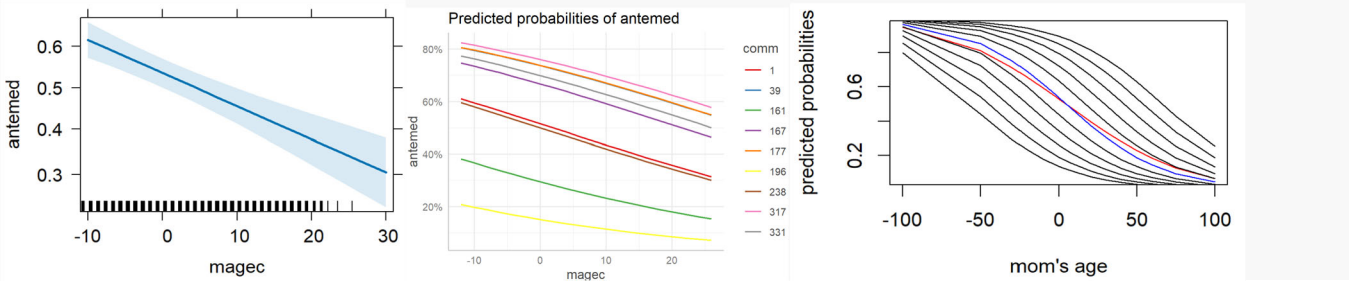
$$ICC = \tau_0^2/(\tau_0^2 + \pi^2/3)$$

where $\pi^2/3$ comes from the variance of the logistic distribution.

(n) Use the suggested formula to calculate an ICC.

Fit the random intercepts model to predict the probability of receiving prenatal care from the mother's age when the child was born (grand mean centered), while allowing for the odds to vary among the communities.

```
model1.mlm = glmer(antemed~ 1 + magec +  (1 | comm), family=binomial, data = bang)
summary(model1.mlm, corr=FALSE)
```



(o) Write out the estimated model equation.

In the special case of random intercepts, you can convert the "subject specific" regression coefficient into the "marginal" (think population average) coefficient with the equation

$$\beta_{marginal} = \hat{\beta}_1^{SS} / \sqrt{1 + .356\,\hat{\tau}_0^2}$$

(p) Calculate and interpret the marginal coefficient.

In most cases the conditional effect will be larger (in abs value) from the marginal effect. The difference between these values increases as the cluster curves are more spread out (larger variance of the intercepts).

Let's continue to explore our model.

(q) Does random slopes significantly improve the model?

(r) Interpret the slope/intercept covariance in context.

(s) Should urban be added to this model? Is the coefficient positive or negative?

(t) Describe the interaction between mother's age and type of community.

(t) Would it make sense to consider adding random slopes for urban?

Computer Problem 15 (due Friday 9am): The data in shot.attempts.csv represent shot attempts in an imaginary hockey league.
Fit a random intercepts model to predict the probability of scoring a goal from the distance and angle of the shot, while allowing for the odds to vary among the players.

Turns out it is trivial to incorporate a second set of random intercepts, often referred to as a "crossed-model" or "cross-classified" or "imperfect hierarchy." (Ch. 13)

(a) Include random effects for goalie as well. How many parameters does this add to the model? How do you interpret the parameter(s)? How do the variance components compare? What does this tell you? What assumption are we making in this model that you might want to question?

Notes:

- The penalized quasi-likelihood algorithm is an alternative to the Fisher Scoring Iterations.
  ```
  #library(MASS)
  model0b <- glmmPQL(antemed ~ 1, random = ~1|comm, family = binomial, data = bang)
  ```
  The results do vary a bit. One advantage is you can control the number of iterations and the "number of quadrature points", i.e., granularity of the search space. This can be useful if your model is having trouble converging.
- The "slope" of the marginal relationship applies to a "randomly selected person" (population average). The slope of the conditional relationship (what you get in the model summary) applies to a particular community.