Stat 414 – Day 33
Logistic regression (Ch. 17)

Ordinary 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.

$$log\ odds = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$$

In running the generalized linear model, we also specify the distribution of the errors. For the logistic model, the default is the binomial distribution. For a binomial random variable, recall that E(Y) = $n\pi$ and Var(Y) = $n\pi(1-\pi)$. Note that we have one parameter here, π, rather than separate parameters for the mean and variance. Also note that the logistic model 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) Fit a logistic regression model. Summarize the "effect" of mom's age on the response variable.

These observations were taken across 361 communities. Are there substantial community to community differences in the likelihood of receiving prenatal care?
(b) Why did I use the "mean" function in the aggregate command?

(c) What is the average proportion?

(d) Is the association between "whether or not prenatal care" and "community" statistically significant?

So let's add the community variable to the model. Keep in mind that the individual communities aren't the primary interest, so instead of 360 dummy variables, we will treat their effects as a random sample from a (normally distributed) population with variance τ^2. Here is the random intercepts (null) model:

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

(e) Explain what u_{0j} and τ_0^2 represent in this context. What's "missing" in this model and why?

We will use the "glmer" function (in lme4 package) to fit multilevel logistic regression models.
(f) Fit the random intercepts model and interpret the intercept in context.

Because we don't have a "σ" parameter, this has led to different suggestions for calculating the intraclass correlation coefficient. I'm partial to

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

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

(g) Use the suggested formula to calculate an ICC. Note, this agrees with the performance package.

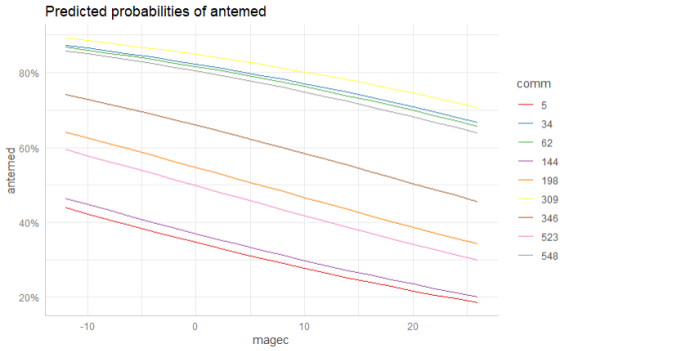
(h) Fit the random intercepts model to predict the probability of receiving prenatal care from the mother's age when the child was born (grand mean centered), while allowing for the odds to vary among the communities. Write out the estimated model equation.

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

```
Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.462    1.209
Number of obs: 5366, groups:  comm, 361

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.144604   0.071781   2.015    0.044 *
magec       -0.032357   0.005235  -6.181 6.37e-10 ***
```

```
plot(ggeffects::ggpredict(model1.mlm, terms=c("magec", "comm [sample=9]"), type="re"), ci = F)
```



Example 2: (a) Consider the hockey goal data from Sam Ventura.
(a) 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."

```
model3.mlm = glmer(goal ~ 1 + angle + dist + (1 | shooter_id) + (1 | goalie_ie), family=binomial, …
```

(b) 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?

Next questions: Does random slopes significantly improve the model? Interpret the slope/intercept covariance. Should urban be added to the model? With random slopes? interaction?