Stat 414 - Day 13
Random Slopes/Logistic Regression

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: Bangledesh prenatal care

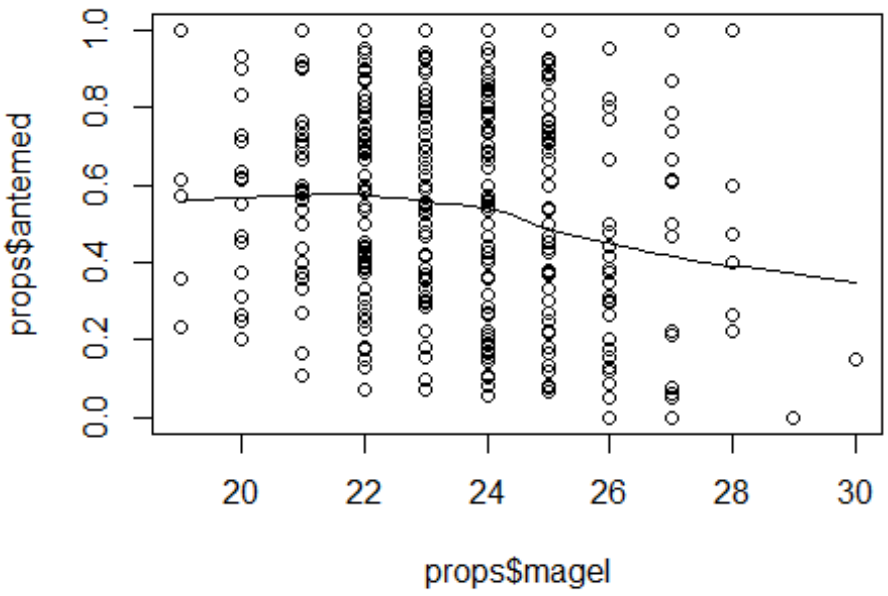
Data were collected on 5,366 women who recently gave birth in Bangledesh. One question we can ask is whether mother's age ($mage$) predicts whether or not the mother receives prenatal care during pregnancy ($antemed$).

```
bang = read.delim("https://www.rossmanchance.com/stat414/data/Bangladesh.txt", head
er=TRUE, "\t")
head(bang$antemed)  #note, the response variable is in "ungrouped" (Bernoulli) form
at
[1] 0 1 1 0 0 1
```

Data Exploration

Aggregate the data to the community level and simplify the age variable for now.

```
props = aggregate(bang, by = list(bang$comm), FUN = mean)
props$mageI <- round(props$mage)
scatter.smooth(props$antemed ~ props$mageI)
```



(a) Explain what props$angemed represents.

The mean of a 0/1 variable is the proportion of 1s and since 1 = yes, this is the proportion of moms in each community who received prenatal care.

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

Older moms tend to be less likely to get prenatal care than younger moms

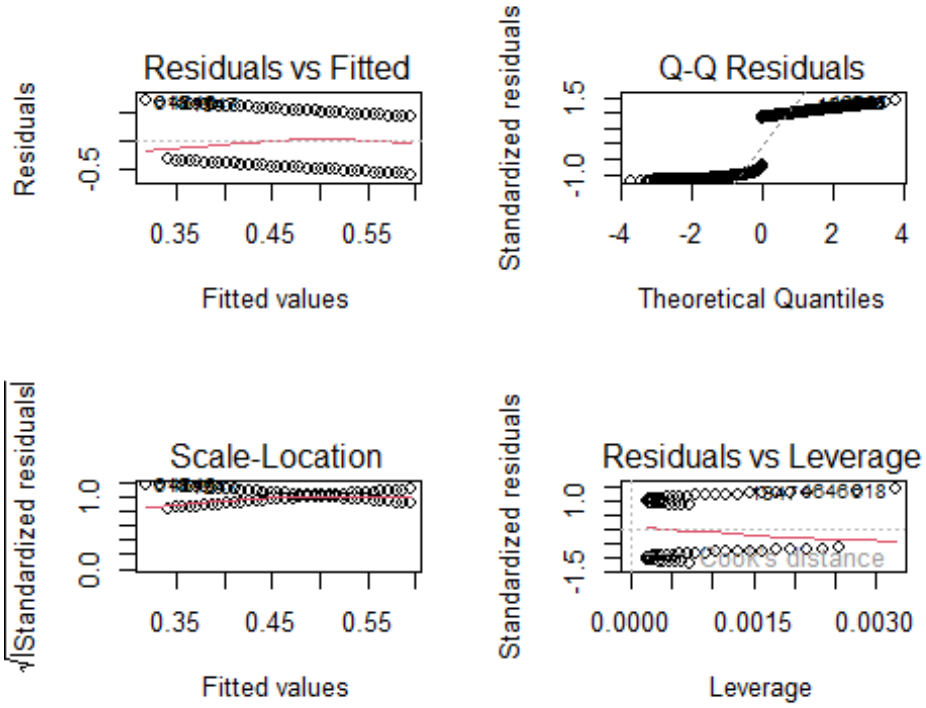
(c) Would a linear model appear appropriate? How are you deciding?

By default could say no because the reponse variable is binary, but the association between the proportion of yeses at each age and age is not terribly nonlinear

Model 1

```
m1 <- lm(antemed ~ mage ,  contrasts = list(comm = contr.sum), data = bang)
summary(m1)

Call:
lm(formula = antemed ~ mage, data = bang, contrasts = list(comm = contr.sum))

Residuals:
   Min     1Q Median     3Q    Max
-0.595 -0.518  0.421  0.474  0.682

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.69479    0.02650   26.21  < 2e-16 ***
mage        -0.00769    0.00108   -7.09  1.5e-12 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.498 on 5364 degrees of freedom
Multiple R-squared:  0.0093,    Adjusted R-squared:  0.00911
F-statistic: 50.3 on 1 and 5364 DF,  p-value: 1.47e-12
par(mfrow=c(2,2))
plot(m1)
```



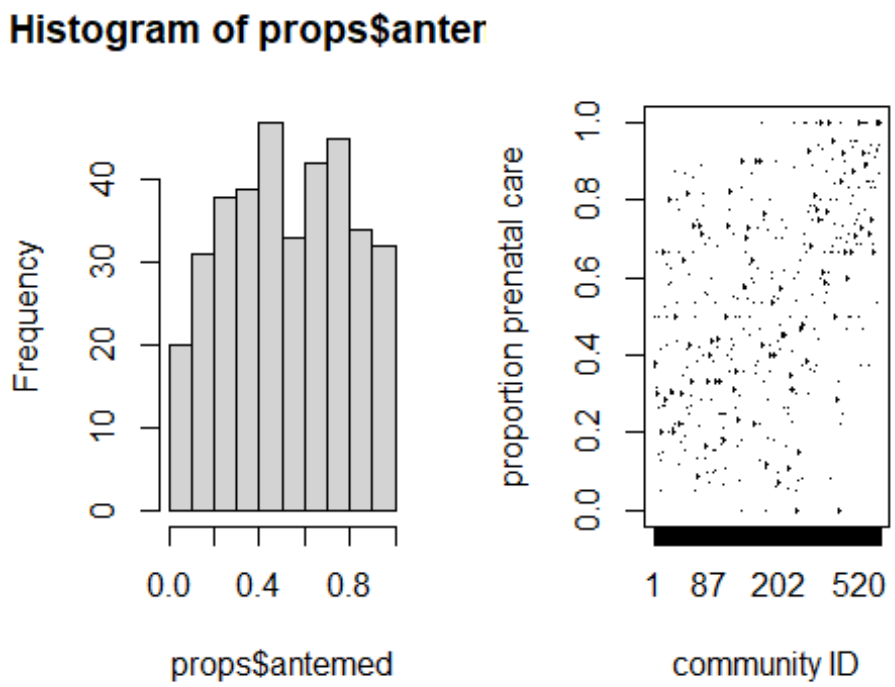

```
par(mfrow=c(1,1))
```

Note, the residual plots aren't very helpful with binary responses

Adding communities

These observations were taken across 361 communities. Are there substantial community to community differences in the likelihood of receiving prenatal care?

```
par(mfrow=c(1,2))
hist(props$antemed)
plot(props$antemed ~ as.factor(props$comm), ylab= "proportion prenatal care", xlab
= "community ID")
```




```
summary(props$antemed); sd(props$antemed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.312   0.533   0.532   0.750   1.000
[1] 0.2686
counts = aggregate(bang, by = list(bang$comm), FUN = sum)
chisq.test(counts$antemed)

    Chi-squared test for given probabilities

data:  counts$antemed
X-squared = 815, df = 360, p-value <2e-16
```

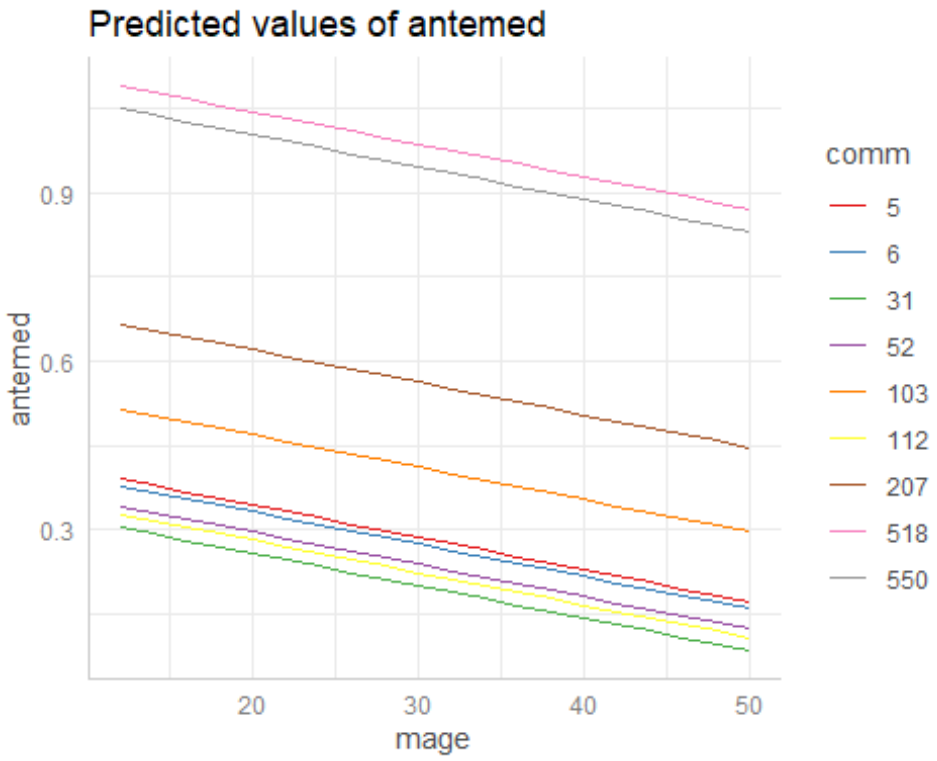
(d) Is the association between "whether or not prenatal care" and "community" statistically significant?

We are essentially testing $H_0: \pi_{c1} = \ldots = \pi_{c361}$ against the alternative that at least one community has a different underlying probability of prenatal care. The chi-square statistic is quite large and the p-value small, so we will reject the null hypothesis and conclude that there are some differences among communities for the likelihood of prenatal care.

Fit a linear model with a different intercept for each community (fixed effects)

```
bang$comm = factor(bang$comm)
m2 <- lm(antemed ~ mage + comm,  contrasts = list(comm = contr.sum), data = bang)
#Don't print out everything!
summary(m2)$coefficients[1:5,1:4]
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  0.668360  0.0243322 27.4681 7.788e-155
```

```
mage         -0.005778  0.0009991 -5.7832  7.773e-09
comm1        -0.023908  0.1162421 -0.2057  8.371e-01
comm2        -0.169172  0.0998290 -1.6946  9.021e-02
comm3        -0.142130  0.0949851 -1.4963  1.346e-01
anova(m2)
Analysis of Variance Table

Response: antemed
            Df Sum Sq Mean Sq F value  Pr(>F)
mage         1     12   12.46   65.70 6.5e-16 ***
comm       360    379    1.05    5.55 < 2e-16 ***
Residuals 5004    949    0.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
plot(ggeffects::ggpredict(m2, terms=c("mage", "comm [sample = 9]")), show_ci = FALS
E)
```



(e) Does the "effect" of mom's age change much when we added community to the model? What does this tell you?

Without community in the model, the slope of age was -.0077 which changed to -.0058, which is not a huge change but does indicate a little bit of association between *age* and *community* - that is some communities tend to be older than olders, and the within-community association is a bit more telling (quicker decrease in probabilitiy of prenatal care witha age.

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

We used 'effect coding' so for the average community we assume zero for all the communitites and just use 0.668 - .0058*33 = 0.48. Nothing fancy to do here than to realize the regression predicts the mean response so the probabiity of prenatal care.

(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)
[1] 0.4777
```

It's the same (up to rounding). This is why in linear models we ignored the distinction of whether the line was predicting one person or the average.

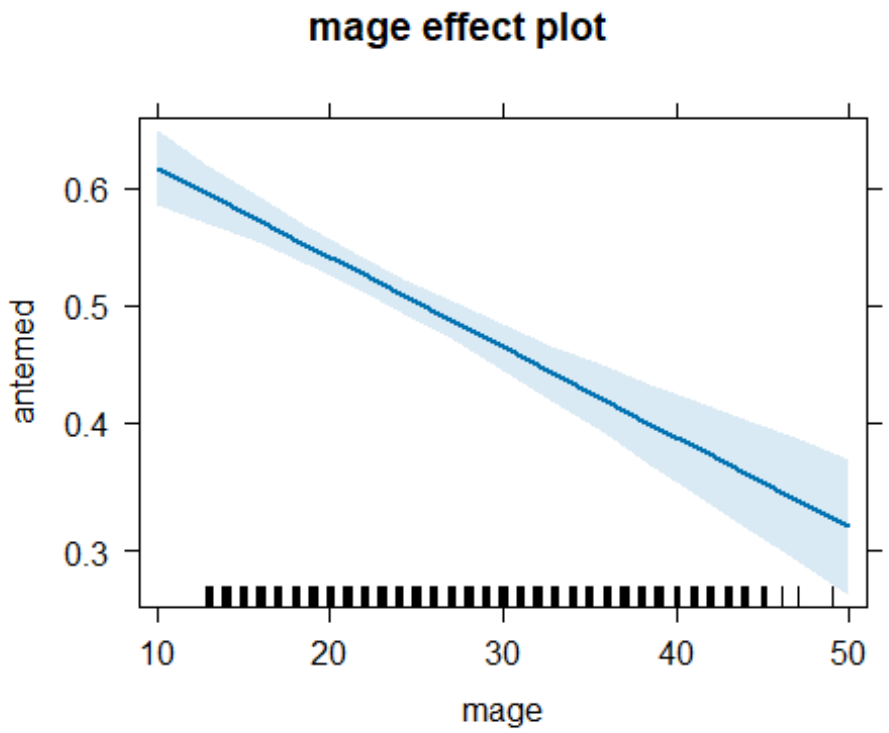
Logistic model

Now let's fit a logistic model instead

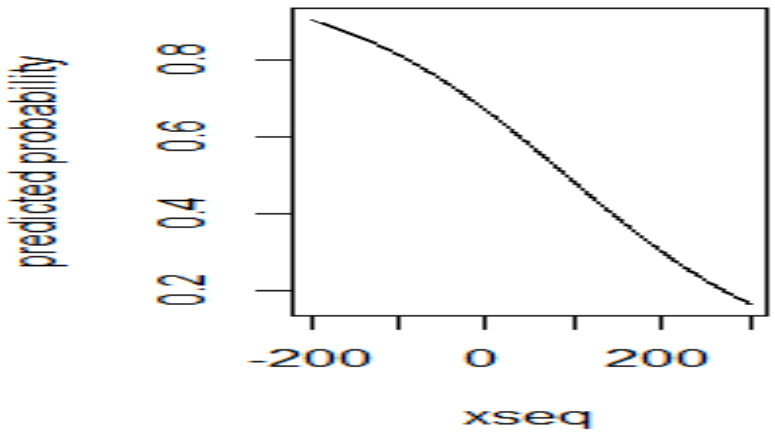
```
model.glm = glm(antemed ~ mage, data = bang, family=binomial)
summary(model.glm)

Call:
glm(formula = antemed ~ mage, family = binomial, data = bang)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.78558    0.10785    7.28  3.2e-13 ***
mage        -0.03103    0.00442   -7.03  2.1e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7435.2  on 5365  degrees of freedom
Residual deviance: 7385.1  on 5364  degrees of freedom
AIC: 7389

Number of Fisher Scoring iterations: 4
```

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.

```
par(mfrow=c(1,2))
#library(effects)
plot(effects::allEffects(model.glm))
```




```
#expanding the x-values to ridiculous numbers to see the "S-shaped" curve
xseq = seq(-200, 300)
preds = exp(.694793 - .007690*xseq)/(1 + exp(.694793 - .007690*xseq))
plot(preds~xseq, type="l", ylab = "predicted probability")
```



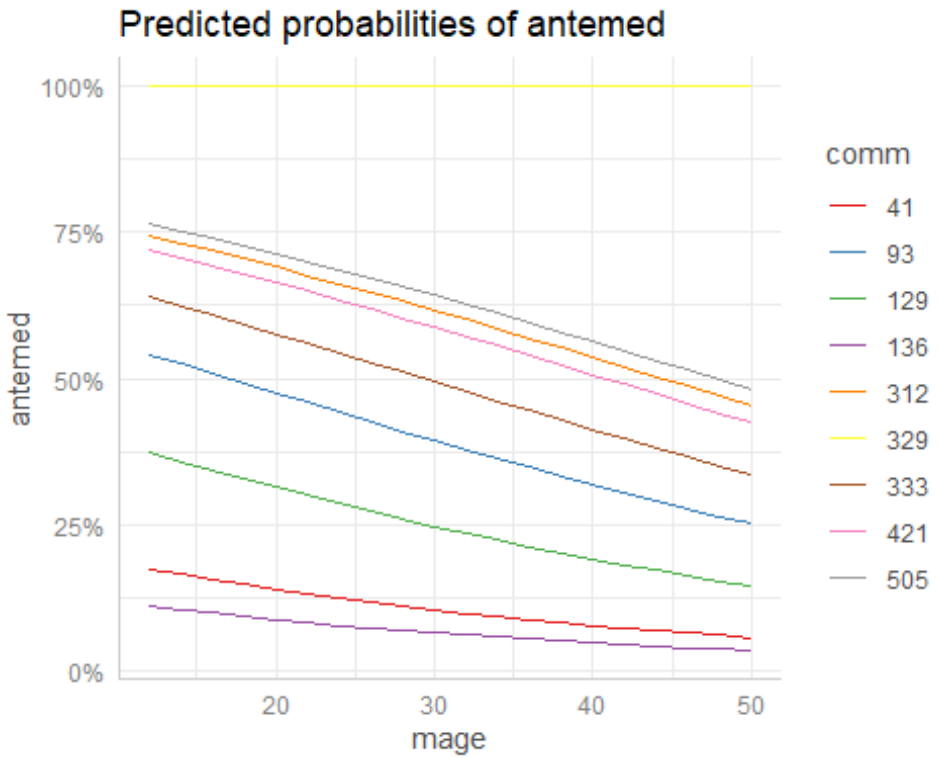
(h) Interpret the slope coefficient in context.

exp(-.031) = 0.969, the predicted odds of prenatal care decreases by the factor 0.97 with each additional year in age

Please interpret the intercept in terms of the estimated proportion, so exp(.785578)/(1 + exp(.785578)) = 0.687, the predicted probability a mom of 0 years of age will get prenatal care!

Now add the community variable to the model (fixed effects)

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

plot(ggeffects::ggpredict(model.glm2, terms=c("mage", "comm [sample=9]")), show_ci
= F)
```




```
summary(model.glm2)$coefficients[1:5,1:3]
            Estimate Std. Error  z value
(Intercept)  1.37010   9.749665  0.14053
mage        -0.03299   0.005541 -5.95289
comm1       -0.54559   9.763491 -0.05588
comm2       -1.17173   9.760474 -0.12005
comm3       -1.03160   9.759187 -0.10571
```

(i) Interpret the slope of momage in context.

exp(-.033) = 0.967, the predicted odds of prenatal care are 0.967 times smaller with each additional year in age, after adjusting for the community the mom is in, meaning for a mom in a particular community, say the typical community.

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

predicted log odds = 1.3701 - .03299*33 = 0.28, so predicted probability = exp(.28)/(1+exp(.28)) = 0.5695

(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)

```
#community 1
lo1 <- 1.3701 - .03299*33 - 0.5456
lo1
[1] -0.2642
p1 <- exp(lo1)/(1+exp(lo1))
new_data <- data.frame(mage = 33, comm = levels(bang$comm))
predicted_values <- predict(model.glm2, newdata = new_data)  #Log odds
predicted_probs <- exp(predicted_values)/(1 + exp(predicted_values))
head(predicted_probs)
     1      2      3      4      5      6
0.4344 0.2911 0.3208 0.6057 0.2586 0.2494
mean(predicted_probs)
[1] 0.4763
```

Noticeably lower at 0.4763

Key Idea

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.

Ranom effects

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

$$ln\left(\pi_j/(1-\pi_j)\right) = \beta_0 + u_{0j}$$

where

$$u_{0j} \sim N(0, \tau_0^2)$$

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

random effect for community j; variance of the community effects

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

```
#library(lme4)
#The random intercepts model
model0.mlm = glmer(antemed~ 1 + (1 | comm), family=binomial, data = bang)
summary(model0.mlm)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
```

```
 Family: binomial  ( logit )
Formula: antemed ~ 1 + (1 | comm)
   Data: bang

      AIC       BIC    logLik -2*log(L)  df.resid
     6640      6653     -3318      6636      5364

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.778 -0.746  0.342  0.712  2.678

Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.46     1.21
Number of obs: 5366, groups:  comm, 361

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1481     0.0718    2.06    0.039 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
confint(model0.mlm)
               2.5 % 97.5 %
.sig01      1.091360 1.3428
(Intercept) 0.007511 0.2899
#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.mlm), 20)
     1      2      3      4      5      6      7      8      9     10     11
0.5060 0.5060 0.5060 0.5060 0.5060 0.5060 0.5060 0.5060 0.5060 0.5060 0.5060
    12     13     14     15     16     17     18     19     20
0.5060 0.5060 0.5060 0.3898 0.3898 0.3898 0.3898 0.3898 0.3898
```

(m) Interpret the intercept in context.

exp(.14808)/(1 + exp(.14808)) = 0.5369 predicted probability of prental care for moms age = 0 in the average community (not necessarily the same as the average probability)

Notice that in the multilevel model output 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.

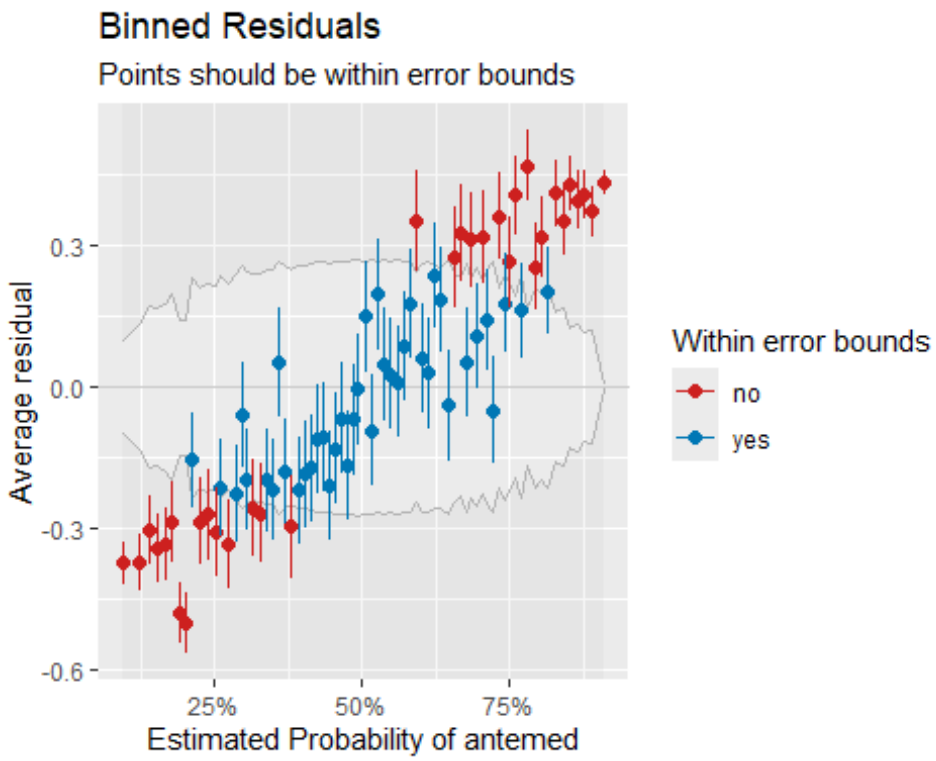
1.464/ (1.464 + 3.1415^2/3) = .308

Note, this agrees with the performance package.

```
performance::icc(model0.mlm)
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.308
  Unadjusted ICC: 0.308
```

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)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: antemed ~ 1 + magec + (1 | comm)
   Data: bang

      AIC       BIC    logLik -2*log(L)  df.resid
     6603      6623     -3299      6597      5363

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.976 -0.743  0.336  0.719  3.236

Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.46     1.21
Number of obs: 5366, groups:  comm, 361

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14460    0.07178    2.01    0.044 *
magec       -0.03236    0.00523   -6.18  6.4e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#Do the residuals look like they're supposed to if the model is well specified?
#library(tidyverse)
performance::binned_residuals(model1.mlm) %>% plot()
```

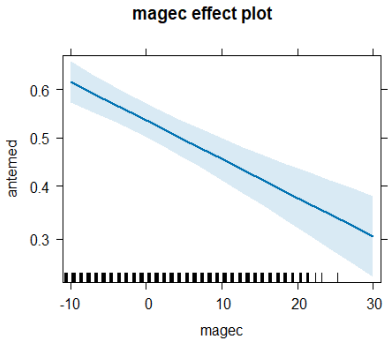


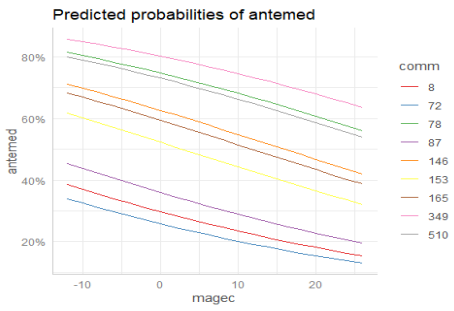

```
performance::performance_hosmer(model1.mlm)
# Hosmer-Lemeshow Goodness-of-Fit Test

  Chi-squared: 77.271
           df:  8
      p-value:  0.000
```

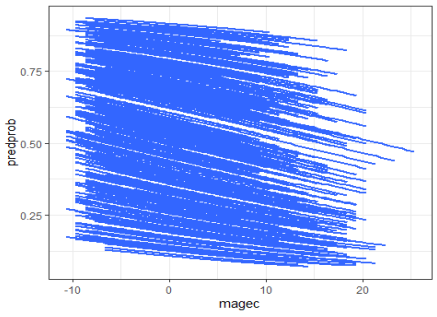
(o) Write out the estimated model equation.

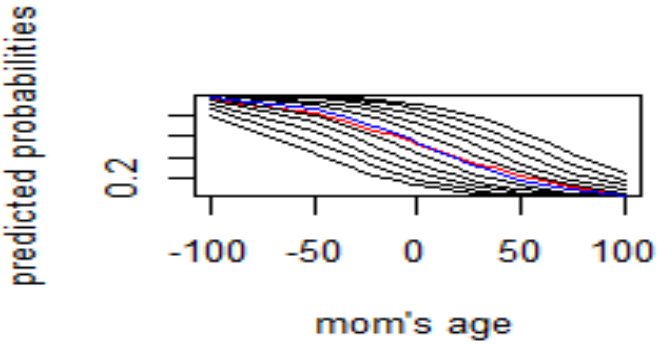
Predicted logs odds of prenatal care = $0.145 - .0324momage + \hat{u}_{0j}$

But the back-transformed probability functions will vary by community:









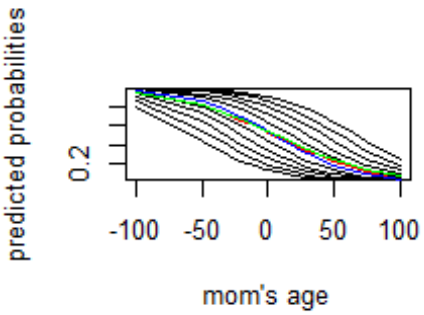
Turns out, you can find the "marginal coefficient" from the "conditional coefficient" using $\hat{\beta}/(\sqrt{1 + .365 \times \widehat{\tau_0^2}}$

(p) Calculate and interpret the marginal coefficients (intercept and slope).

Intercept: .144/sqrt(1 + .356*1.462) = 0.1168

Slopes: -.03236/sqrt(1 + .356*1.462) = -.0262

```
[1] -0.02624
```



-.033 / sqrt(1 + .356 * 1.461) = -.0268

Note the correspondence between the red and green curves, and how the overall association is "flatter" for this marginal association.

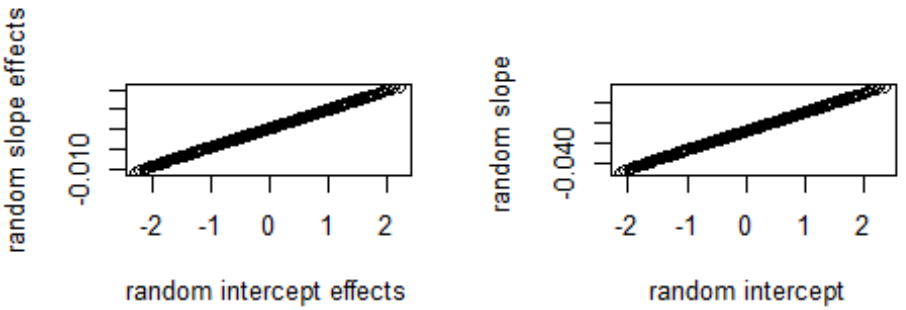
We would interpret the (back transformed) marginal intercept as the predicted probability of prenatal care for mom's of average age averaged across the communities (i.e., population-averaged) and the marginal slope as the effect of mother's age averaged over the communities. In most cases the conditional effect (within community) will be larger (in abs value) than the marginal effect (the fixed effects are shruken towards zero when convert conditional to marginal).

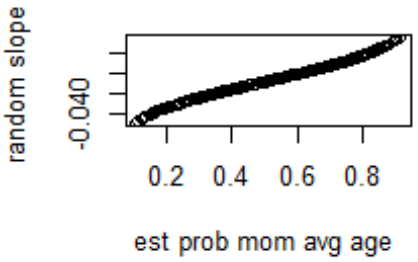
Computer Problem 14 - due 7am Wednesday

Let's continue to explore our model.

(q) The likelihood of prenatal care changes with mother's age at birth. Does this rate of change seem to vary across the communities? (Be clear how you are deciding.)

```
model2.rs = glmer(antemed~ 1 + magec +  (magec | comm), family=binomial, data = bang)

summary(model2.rs)

p1 <- plot(effects::allEffects(model2.rs))

p2 <- plot(ggeffects::ggpredict(model2.rs, terms=c("magec", "comm [sample=9]"), type="random"), show_ci = F)

predprob2=fitted(model2.rs)

p3 <- ggplot(data = bang, aes(y=predprob2, x=magec, group=comm)) + geom_smooth(method="loess", se=F) +  theme_bw()

p1 + p2 + p3

anova(model1.mlm, model2.rs)
```





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

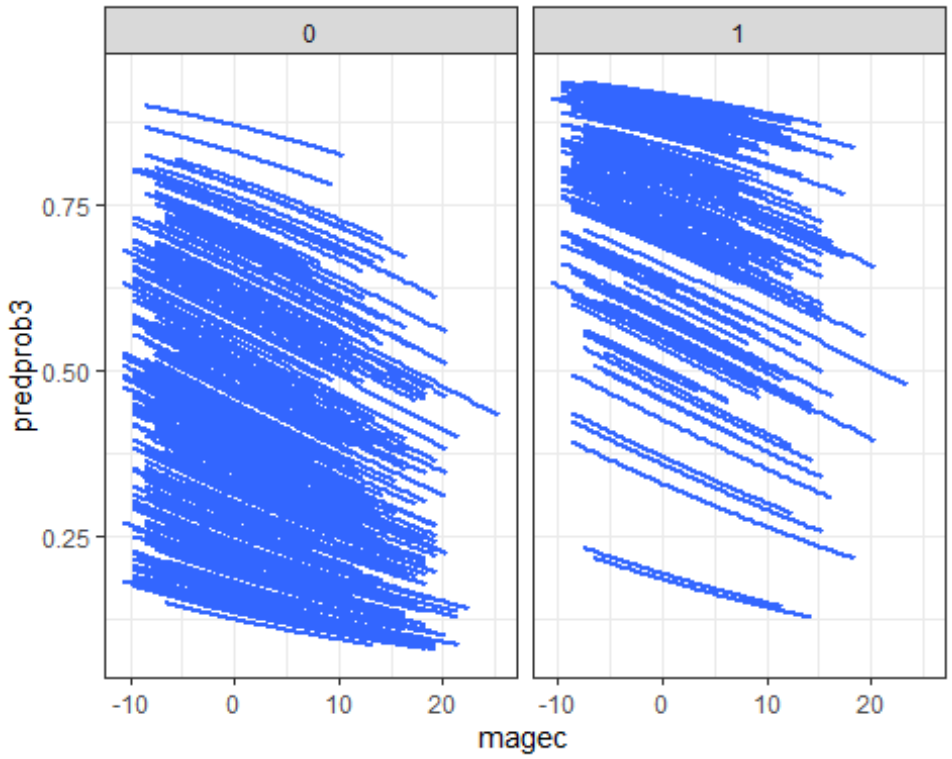
Adding a Level 2 variable

Communities have been designated as urban (urban = 1) or rural (urban = 0).

(s) Does "urban" explain significant variation in the response (in the intercepts at the community level)? Is the coefficient positive or negative? How do you interpret that? (Keep in mind this is a categorical variable, so just saying 'positive association' is not very clear.)

```
model3.rs = glmer(antemed~ 1 + magec +  urban + (1 + magec | comm),
family=binomial, data = bang)
summary(model3.rs, corr=FALSE)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: antemed ~ 1 + magec + urban + (1 + magec | comm)
   Data: bang

      AIC       BIC    logLik -2*log(L)  df.resid
     6496      6536     -3242      6484      5360

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.155 -0.734  0.322  0.721  3.163

Random effects:
 Groups Name        Variance   Std.Dev. Corr
 comm   (Intercept) 0.97008107 0.984927
        magec       0.00000035 0.000592 -1.00
Number of obs: 5366, groups:  comm, 361

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34642    0.07429   -4.66  3.1e-06 ***
magec       -0.03252    0.00524   -6.20  5.5e-10 ***
urban        1.49530    0.13328   11.22  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
optimizer (Nelder_Mead) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
predprob3=fitted(model3.rs)
ggplot(data = bang, aes(y=predprob3, x=magec, group=comm)) +
  facet_wrap(~urban) +
  geom_smooth(method="loess", se=F) +
  theme_bw()
```

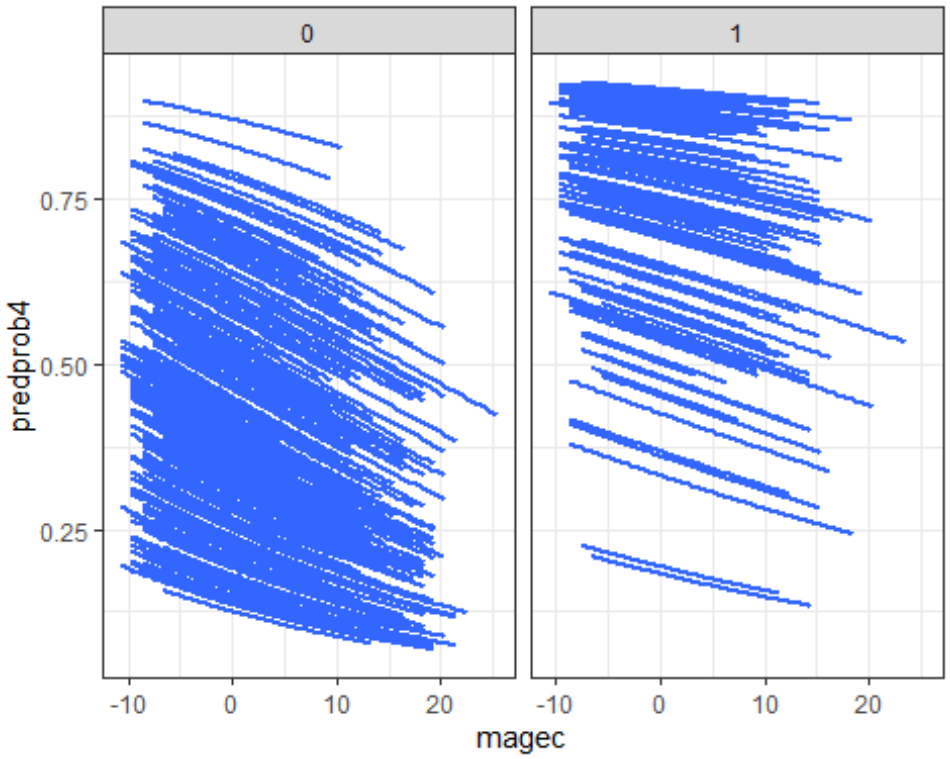



```
anova(model2.rs, model3.rs)
Data: bang
Models:
model2.rs: antemed ~ 1 + magec + (magec | comm)
model3.rs: antemed ~ 1 + magec + urban + (1 + magec | comm)
           npar  AIC  BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model2.rs     5 6607 6640  -3298      6597
model3.rs     6 6496 6536  -3242      6484   113  1     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(t) Include and describe the interaction between mother's age and type of community (rural vs. urban) (*Hint*: Do better than 'higher slopes for urban communities')

```
model4.rs = glmer(antemed~ 1 + magec +  urban + magec*urban + (magec | comm), famil
y=binomial, data = bang)
summary(model4.rs)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: antemed ~ 1 + magec + urban + magec * urban + (magec | comm)
   Data: bang

      AIC       BIC    logLik -2*log(L)  df.resid
     6496      6542     -3241      6482      5359

Scaled residuals:
   Min     1Q Median     3Q    Max
```

```
-3.066 -0.738  0.330  0.717  3.375

Random effects:
 Groups Name        Variance  Std.Dev. Corr
 comm   (Intercept) 0.9689861 0.98437
        magec       0.0000041 0.00203  1.00
Number of obs: 5366, groups:  comm, 361

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34786    0.07428   -4.68  2.8e-06 ***
magec       -0.03681    0.00615   -5.99  2.1e-09 ***
urban        1.49465    0.13280   11.25  < 2e-16 ***
magec:urban  0.01683    0.01228    1.37     0.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) magec  urban
magec        0.046
urban       -0.561 -0.027
magec:urban -0.028 -0.515  0.026
optimizer (Nelder_Mead) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
predprob4=fitted(model4.rs)
ggplot(data = bang, aes(y=predprob4, x=magec, group=comm)) +
  facet_wrap(~urban) +
  geom_smooth(method="loess", se=F) +
  theme_bw()
```




```
anova(model3.rs, model4.rs)
Data: bang
Models:
model3.rs: antemed ~ 1 + magec + urban + (1 + magec | comm)
model4.rs: antemed ~ 1 + magec + urban + magec * urban + (magec | comm)
          npar  AIC  BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model3.rs    6 6496 6536  -3242      6484
model4.rs    7 6496 6542  -3241      6482  1.84  1       0.17
```