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.

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

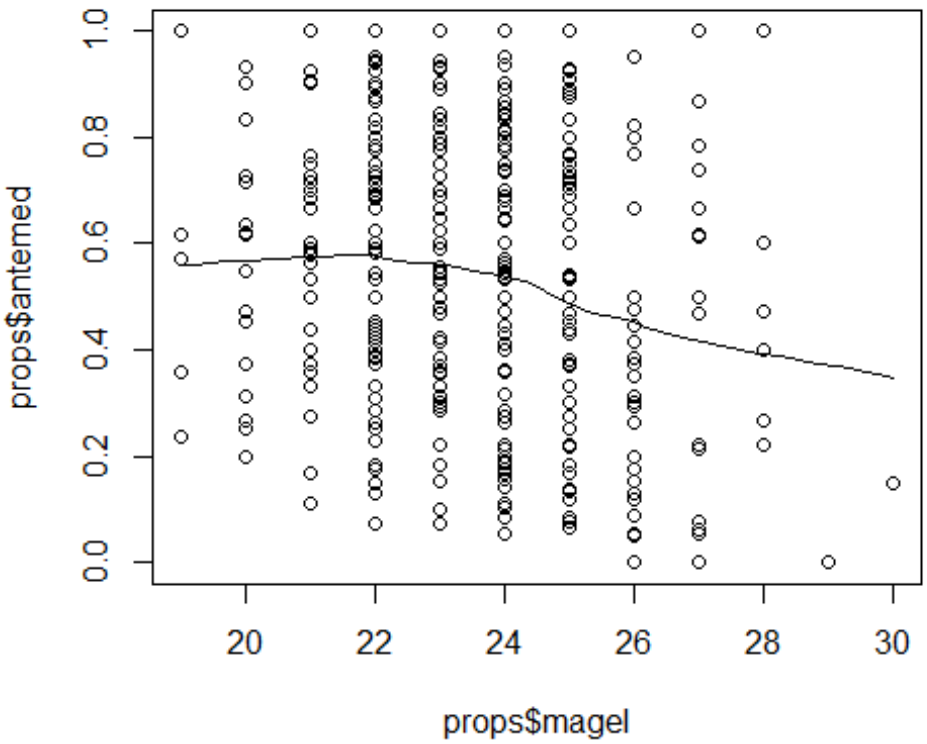
where we assume a Binomial distribution with $E(Y) = n\pi$ and $Var(Y) = n\pi(1-\pi)$. Note that we have one parameter here π and not separate parameters for the mean and variance. Also note that the logistic models works just as well for binomial observations (response = number or proportions of successes)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

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

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$antemed represents.

proportion of mom's in each community who received prenatal care (1 = yes, 0 = no)

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

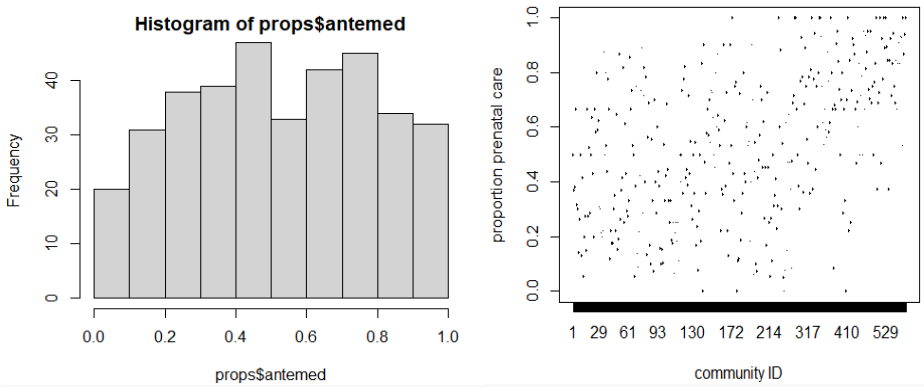
the likelihood of prenatal care decreases with age

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

probably not the best idea but more than some we will see even though the respoinse is proportions and need to be limited to be between 0 and 1

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

```
hist(props$antemed)
```




```
summary(props$antemed); sd(props$antemed)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3125  0.5333  0.5323  0.7500  1.0000
## [1] 0.2686424
```

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

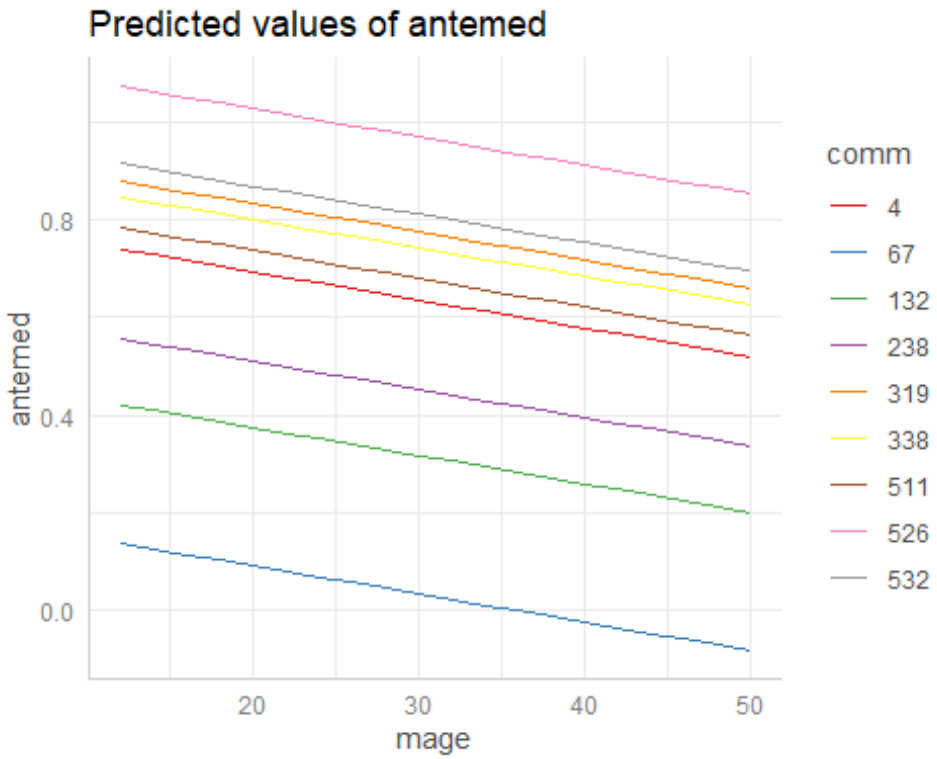
```
counts = aggregate(bang, by = list(bang$comm), FUN = sum)
chisq.test(counts$antemed)
##
##  Chi-squared test for given probabilities
##
## data:  counts$antemed
## X-squared = 815.43, df = 360, p-value < 2.2e-16
```

Fit a model with a different intercept for each community

First added the model without community.

```
bang$comm = factor(bang$comm)
m1 <- lm(antemed ~ mage, data = bang); summary(m1)
##
## Call:
## lm(formula = antemed ~ mage, data = bang)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5948 -0.5179  0.4206  0.4744  0.6820
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.694793   0.026505  26.214  < 2e-16 ***
## mage        -0.007690   0.001084  -7.094 1.47e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4976 on 5364 degrees of freedom
```

```
## Multiple R-squared:  0.009295,   Adjusted R-squared:  0.00911
## F-statistic: 50.33 on 1 and 5364 DF,  p-value: 1.471e-12

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)  0.668359745 0.0243322088 27.468108 7.787976e-155
## mage        -0.005778066 0.0009991081 -5.783225  7.773035e-09
## comm1       -0.023908084 0.1162420739 -0.205675  8.370532e-01
## comm2       -0.169171926 0.0998289629 -1.694618  9.021019e-02
## comm3       -0.142130265 0.0949851010 -1.496343  1.346274e-01
#library(ggeffects)
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?

there is some relationship between community and mom's ages if you feel not much changed, then is not much of a relationship between the ages across the communities

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

.668 - .00578*33 = 0.477

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

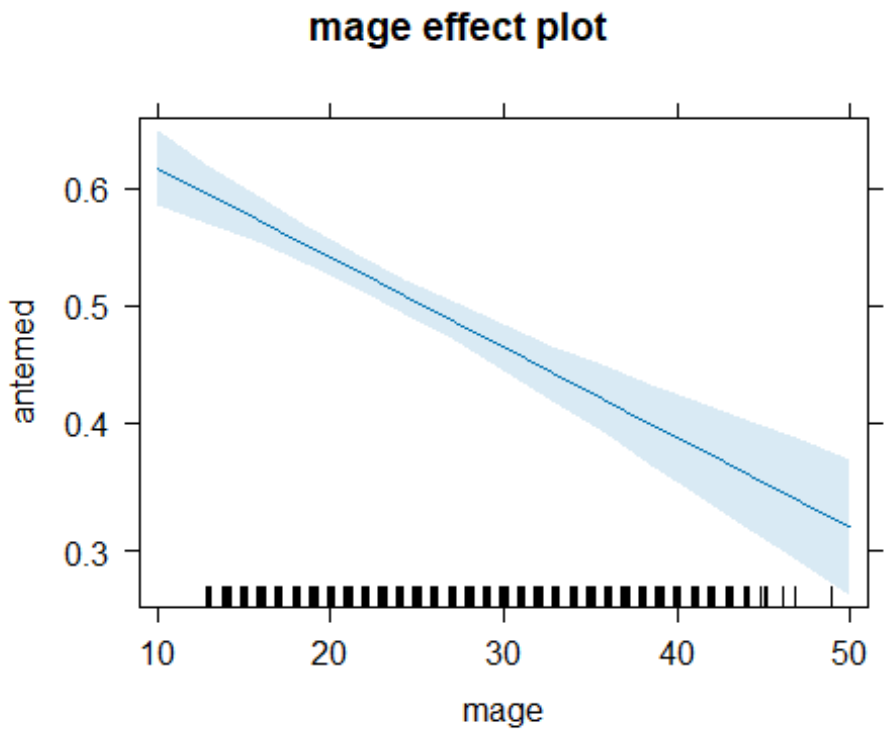
```
new_data <- data.frame(mage = 33, comm = levels(bang$comm))

predicted_values <- predict(m2, newdata = new_data)
mean(predicted_values)
## [1] 0.4776836
```

Now let's fit a logistic model instead

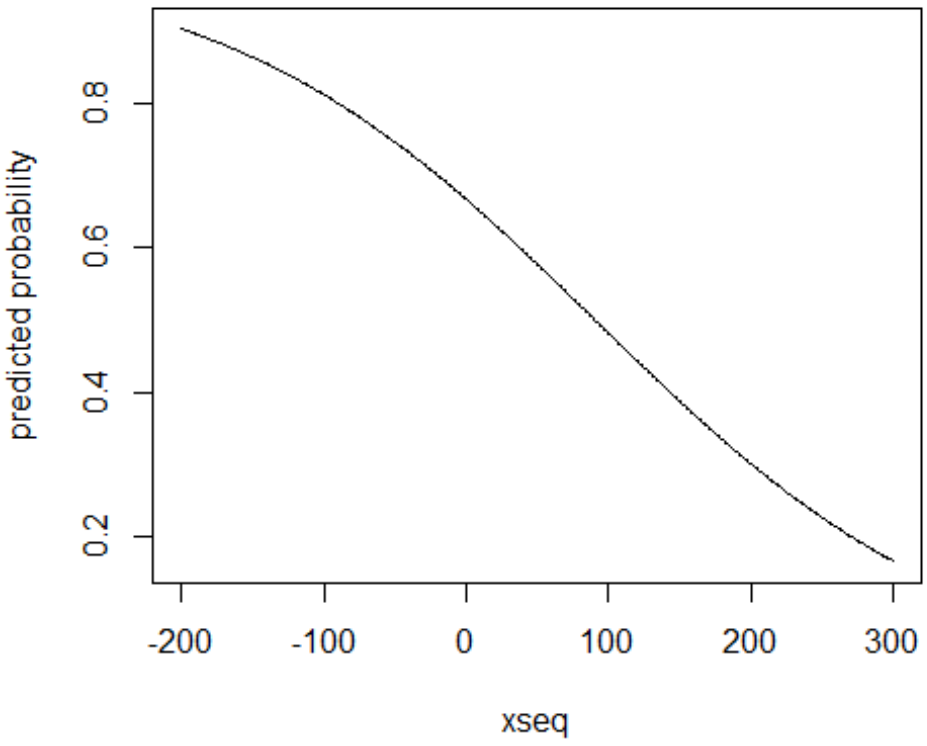
```
model.glm = glm(antemed ~ mage, data = bang, family=binomial)
summary(model.glm)
##
## Call:
## glm(formula = antemed ~ mage, family = binomial, data = bang)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.785578   0.107847   7.284 3.24e-13 ***
## mage        -0.031029   0.004415  -7.028 2.10e-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: 7389.1
##
## 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.

```
#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")
```

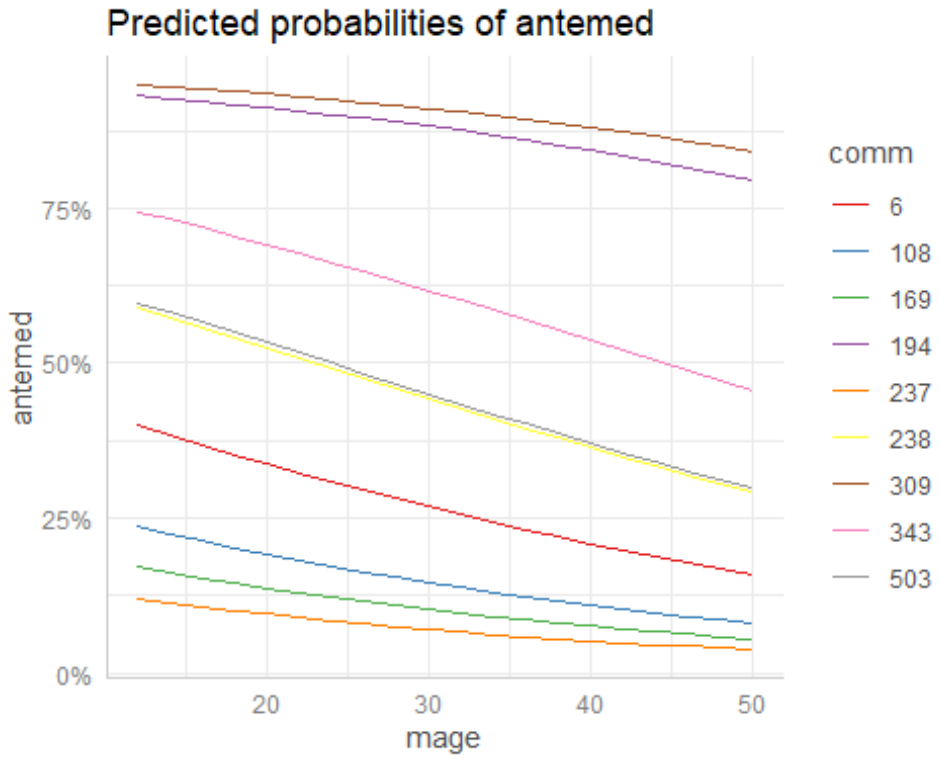


(h) Interpret the slope coefficient in context.

Each one year increase in mom age is associated with a exp(-.007690) = 0.992 times smaller odds of prenatal care.

Now add the community variable to the model

```
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)
## Data were 'prettified'. Consider using `terms="mage [all]"` to get
##   smooth plots.
```




```
summary(model.glm2)$coefficients[1:5,1:3]
##                Estimate  Std. Error     z value
## (Intercept)  1.37009837 9.749664504  0.14052774
## mage        -0.03298677 0.005541306 -5.95288701
## comm1       -0.54559224 9.763490704 -0.05588086
## comm2       -1.17173005 9.760474257 -0.12004848
## comm3       -1.03159693 9.759186907 -0.10570521
```

(i) Interpret the slope of momage in context.

exp(-.033) is the multiplicative decrease in predicted odds of prenatal care per year in a particular community, in a typical community

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

1.37 - .033*33 = .281 #predicted log odds

exp(.281) = 1.32 #predicted odds

1.32/(1 + 1.32) = .569 # predicted probability

(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)

```
lo1 <- 1.3701 - .03299*33 - 0.5456
lo1
## [1] -0.26417
p1 <- exp(lo1)/(1+exp(lo1))
new_data <- data.frame(mage = 33, comm = levels(bang$comm))
predicted_values <- predict(model.glm2, newdata = new_data)
predicted_probs <- exp(predicted_values)/(1 + exp(predicted_values))
head(predicted_probs)
##         1         2         3         4         5         6
## 0.4343666 0.2910696 0.3208078 0.6056791 0.2585588 0.2493996
mean(predicted_probs)
## [1] 0.47628
```

Not the same! (can only assume the same with linear relationships)

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

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

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 = glmer(antemed~ 1 + (1 | comm), family=binomial, data = bang)
summary(model0)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: antemed ~ 1 + (1 | comm)
##    Data: bang
##
##      AIC      BIC   logLik deviance df.resid
##   6639.5   6652.7  -3317.8   6635.5     5364
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7779 -0.7458  0.3423  0.7118  2.6784
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  comm   (Intercept) 1.464    1.21
## Number of obs: 5366, groups:  comm, 361
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.14809    0.07178   2.063   0.0391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
confint(model0)
## Computing profile confidence intervals ...
##                   2.5 %    97.5 %
## .sig01      1.091360254 1.3428293
## (Intercept) 0.007511176 0.2898802
#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)
##         1         2         3         4         5         6         7         8
## 0.5060448 0.5060448 0.5060448 0.5060448 0.5060448 0.5060448 0.5060448 0.5060448
##         9        10        11        12        13        14        15        16
## 0.5060448 0.5060448 0.5060448 0.5060448 0.5060448 0.5060448 0.3898461 0.3898461
##        17        18        19        20
## 0.3898461 0.3898461 0.3898461 0.3898461
```

(m) Interpret the intercept in context.

exp(intercept) predicted probability of prental care for moms age = 0 in the average community

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.

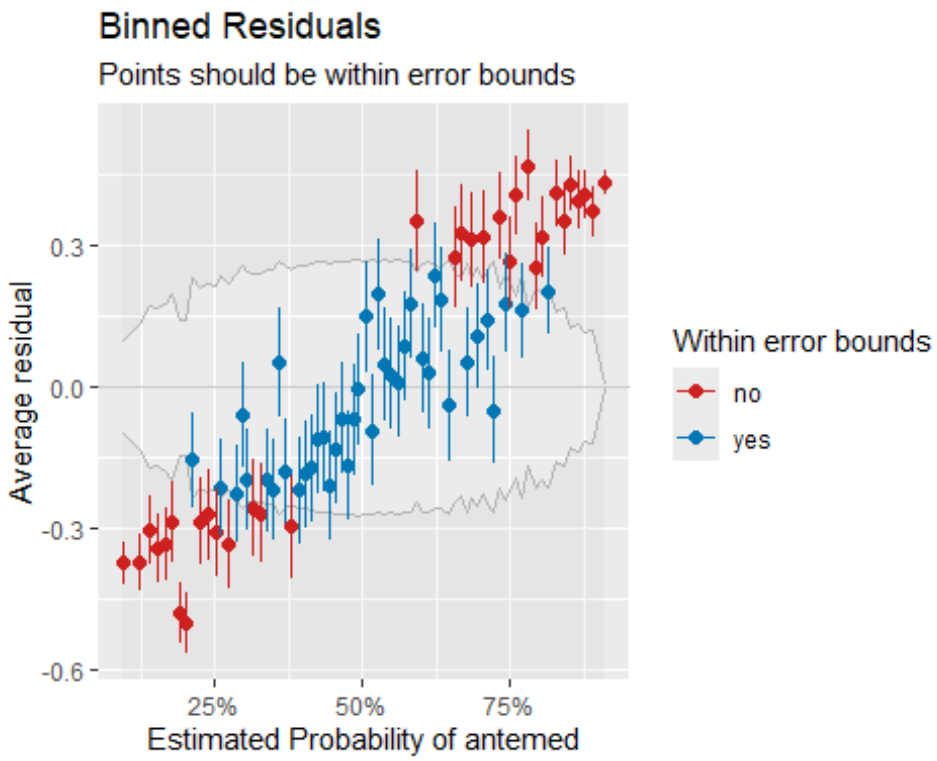
1.464/ (1.464 + 3.1415^2/3) = .308

Note, this agrees with the performance package.

```
performance::icc(model0)
## # 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.

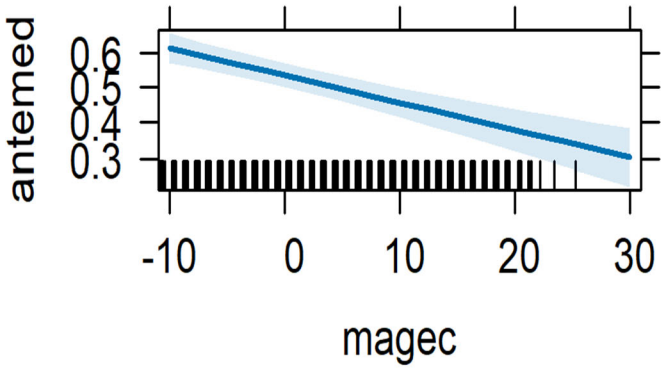
```
model1.mlm = glmer(antemed~ 1 + magec +  (1 | comm), family=binomial, data = bang)
summary(model1.mlm, corr=FALSE)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: antemed ~ 1 + magec + (1 | comm)
##    Data: bang
##
##      AIC      BIC   logLik deviance df.resid
##   6603.4   6623.2  -3298.7   6597.4     5363
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9757 -0.7431  0.3357  0.7190  3.2357
##
## 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 ***
## ---
## 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()
```

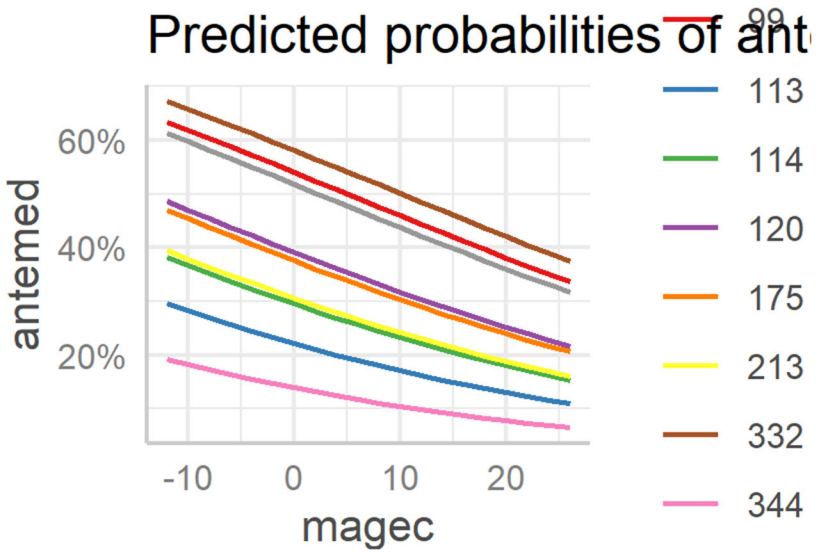


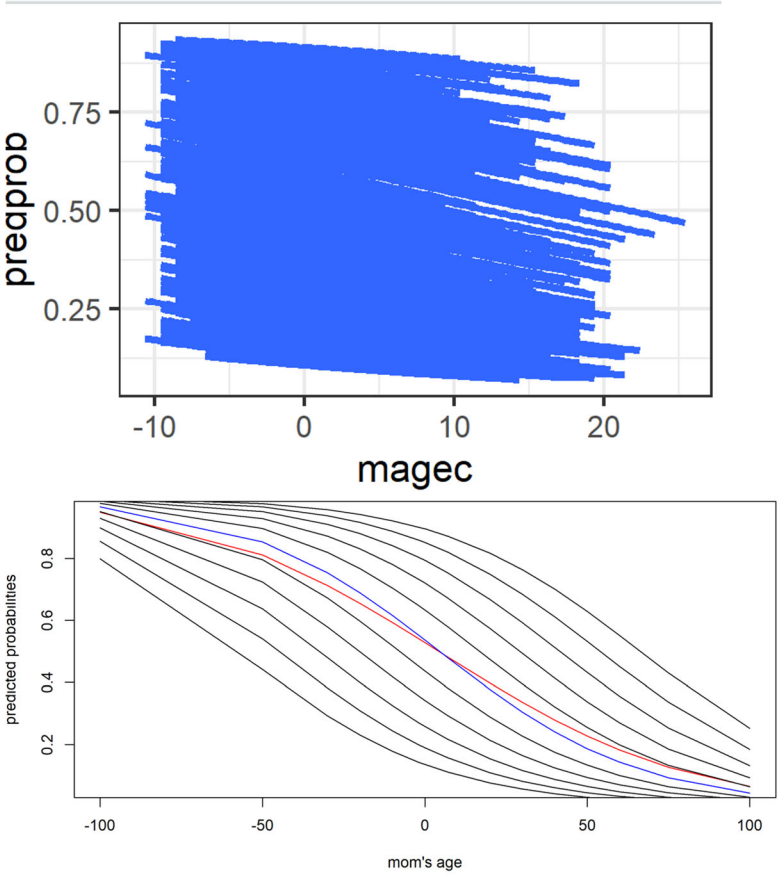

```
performance::performance_hosmer(model1.mlm)
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 77.271
##            df:  8
##       p-value:  0.000
## Summary: model does not fit well.
```

(o) Write out the estimated model equation.

Predicted logs odds of prenatal care = 0.145 - .0324 mom age + $\hat{u}_j$


antemed
0.6
0.5
0.4
0.3
-10
0
10
20
30
magec


Predicted probabilities of ant
antemed
60%
40%
20%
-10
0
10
20
magec
99
113
114
120
175
213
332
344



(p) Calculate and interpret the marginal coefficient.

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

[1] -0.02676551

"effect" of mom's age for a randomly selected mom /averaging across all of the communities / population effect

rather than within a specific community

See Quiz 15