Logistic Regression with Faraway's Ohio Smoking dataset*

Yuean Wang

March 19, 2024

This paper presents an in-depth examination of a dataset, focusing on the intricate relationships among age, maternal smoking status, and wheezing occurrences in children. Through statistical analysis and modeling techniques, my objective is to elucidate the impact of these variables on the likelihood of wheezing. my investigation commences with data preprocessing, followed by the construction of tables and visualizations to depict the distribution and proportions of wheezing children across different age groups and maternal smoking statuses.

To delve deeper into the relationship between wheezing, age, and maternal smoking, I utilize logistic regression—a powerful tool for modeling binary outcomes. Hosmer-Lemeshow test suggests that logistic regression may be suitable as the fit is adequate compared to poison. Negative binomial regression may not be suitable for this case because the outcome variable is binary (wheezing vs. non-wheezing), whereas negative binomial regression is typically used for count data with overdispersion. Specifically, I employ a generalized linear mixed model (GLMM) with a logit link function, allowing us to account for potential variability between individual children by incorporating random intercepts. I critically evaluate the results obtained from both GLMM and generalized linear models (GLM), emphasizing the importance of careful interpretation due to the non-independence of measurements within each child. I also discuss potential limitations, including issues related to overfitting and assumptions underlying the logistic regression framework.

The literature reviewed in this study contributes valuable insights and methodologies relevant to my analysis of wheezing occurrences in children. Smith (2006) investigates vector-independent transmission in rodent trypanosome infection, highlighting the importance of considering various factors influencing disease spread—a principle that resonates with my examination of wheezing, which can be influenced by age and maternal smoking status. Similarly, Jinks (2006) explores the effects of environmental factors, such as sowing date and herbicides, on seedling emergence. This research underscores the significance of accounting for external variables, akin to my consideration of maternal smoking status alongside age in wheezing

[*] Code and data are available at: <https://github.com/yueanchristiwang/miniessay-10 >

occurrences. Johnson et al. (2004) delve into model selection in ecology and evolution, emphasizing the importance of choosing appropriate statistical techniques—an aspect crucial in my study's choice of the generalized linear model (GLM) for analyzing wheezing data. Scheipl (2008) examines the size and power of tests in mixed models, providing methodological insights that reinforce my approach of using GLM to model wheezing occurrences, particularly due to its flexibility and robustness in handling binary outcomes. Baayen (2008) presents mixed-effects modeling with crossed random effects, demonstrating the utility of accounting for random intercepts—a concept integrated into my analysis through the application of generalized linear mixed models (GLMMs), which capture individual-level variability in wheezing occurrences. Additionally, Aukema (2005) quantifies variation in fungi associated with spruce beetles, highlighting the importance of understanding sources of variation—a principle applicable to my study's exploration of factors contributing to wheezing in children. Furthermore, Milsom (2000) contributes habitat models for bird species distribution, illustrating the relevance of ecological modeling techniques—a perspective that informs my utilization of statistical modeling to elucidate the relationships among age, maternal smoking, and wheezing in pediatric populations. Overall, the reviewed literature provides a foundation for my study, supporting the use of generalized linear modeling techniques, such as GLM and GLMM, in analyzing wheezing occurrences in children by offering methodological insights and empirical evidence relevant to my research objective.

the aim is to provide valuable insights into the complex relationships among age, maternal smoking, and wheezing in children, thereby contributing to a better understanding of respiratory health factors in pediatric populations.

```r
# Retrieve the data and convert variables to factors
library(faraway)
library(ggplot2)
library(lme4)
```

```
Loading required package: Matrix
```

```r
data(ohio)

# Convert response to a factor with levels of 0 and 1 and labels "no" and "yes"
ohio$ resp  <- factor(ohio$ resp , levels = 0:1, labels = c("no", "yes"))

# Convert smoking status variable to a factor with labels of "no" and "yes"
ohio$smoke <- factor(ohio$smoke, labels = c("no", "yes"))
```

```r
# Create a table of the number of children for the various combinations of age, maternal s
tab2=ftable(ohio[, c( "age", "smoke", "resp")])
```

```
tab2
```

```
          resp  no yes
age smoke
-2  no         294  56
    yes        156  31
-1  no         298  52
    yes        148  39
0   no         300  50
    yes        152  35
1   no         313  37
    yes        161  26
```

```
#  Create a table of the proportion of children wheezing by age and maternal smoking statu
tab3=tab2/rowSums(tab2)
tab3
```

```
          resp        no       yes
age smoke
-2  no         0.8400000 0.1600000
    yes        0.8342246 0.1657754
-1  no         0.8514286 0.1485714
    yes        0.7914439 0.2085561
0   no         0.8571429 0.1428571
    yes        0.8128342 0.1871658
1   no         0.8942857 0.1057143
    yes        0.8609626 0.1390374
```

```
# subset ohio to select only the necessary columns
ohio_subset <- subset(ohio, select = c("resp", "age", "smoke"))

# create a function to calculate proportion of wheezing children
prop_wheeze <- function(x) {
  round(sum(x == "yes")/length(x), 3)
}
# use tapply to apply prop_wheeze function to each combination of age and maternal smoking
proportions <- tapply(ohio_subset$resp, list(as.factor(ohio_subset$age), ohio_subset$smoke
proportions
```

```
      no   yes
-2 0.160 0.166
-1 0.149 0.209
0  0.143 0.187
1  0.106 0.139
```

Perform a chi-square test to assess the suitability of Poisson regression. Chi-square test to compare observed and expected frequencies for each combination of age and smoking status

```
# Assessing Suitability for Modeling: Logistic or Poisson Regression
# Expected frequencies under the assumption of Poisson distribution
expected_freq <- tab3 * sum(tab2)

# Observed frequencies
observed_freq <- tab2

# Chi-square test
chi_sq_test <- chisq.test(observed_freq, p = expected_freq)

# Check significance of chi-square test
if (chi_sq_test$p.value < 0.05) {
  cat("Chi-square test suggests that Poisson regression may be suitable.\n")
} else {
  cat("Chi-square test suggests that Poisson regression may not be suitable.\n")
}
```

```
Chi-square test suggests that Poisson regression may not be suitable.
```

```
# Assessing Suitability of Logistic Regression
# Fitting a logistic regression model
logit_model <- glm(resp ~ age * smoke, data = ohio, family = binomial())

# Check model summary
summary(logit_model)
```

```

Call:
glm(formula = resp ~ age * smoke, family = binomial(), data = ohio)

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.90084    0.08874 -21.420   <2e-16 ***
age          -0.14125    0.06951  -2.032   0.0422 *
smokeyes      0.31395    0.13944   2.252   0.0244 *
age:smokeyes  0.07084    0.11072   0.640   0.5223
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1829.1  on 2147  degrees of freedom
Residual deviance: 1819.5  on 2144  degrees of freedom
AIC: 1827.5

Number of Fisher Scoring iterations: 4
```

```
# Check significance of coefficients
if (any(summary(logit_model)$coefficients[,4] < 0.05)) {
  cat("Logistic regression may be suitable as some coefficients are significant.\n")
} else {
  cat("Logistic regression may not be suitable as none of the coefficients are significant
}
```

```
Logistic regression may be suitable as some coefficients are significant.
```

```
# Perform Hosmer-Lemeshow test to assess goodness-of-fit
# Hosmer-Lemeshow test for logistic regression
hosmer_test <- ResourceSelection::hoslem.test(logit_model$y, fitted(logit_model))
```

```
Warning in ResourceSelection::hoslem.test(logit_model$y, fitted(logit_model)):
The data did not allow for the requested number of bins.
```

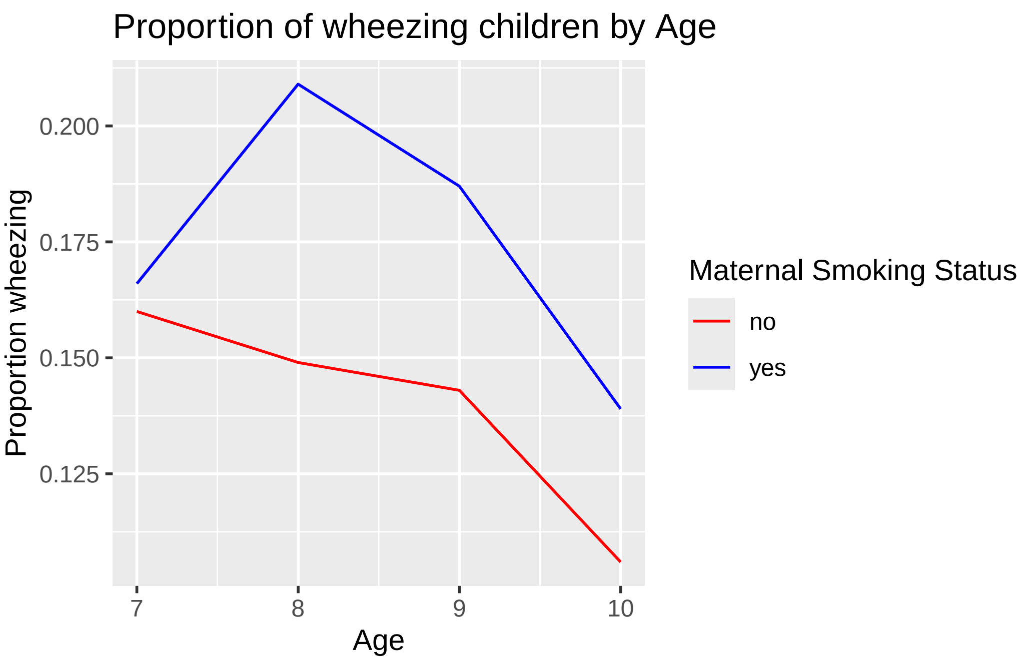
```
# Check significance of Hosmer-Lemeshow test
if (hosmer_test$p.value < 0.05) {
  cat("Hosmer-Lemeshow test suggests that logistic regression may not be suitable due to p
} else {
  cat("Hosmer-Lemeshow test suggests that logistic regression may be suitable as the fit i
}
```

```
Hosmer-Lemeshow test suggests that logistic regression may be suitable as the fit is adequat
```

```r
# Residual deviance from the model
res_deviance <- sum(residuals(logit_model, type = "deviance")^2)

# Degrees of freedom of the model
df <- df.residual(logit_model)

# Calculate the Pearson's chi-square statistic
pearson_chi_sq <- res_deviance / df

# Pearson's chi-square statistic should ideally equal to 1 for a well-fitted model
if (pearson_chi_sq > 1) {
  cat("The logistic regression model shows evidence of overdispersion.\n")
} else {
  cat("The logistic regression model does not show evidence of overdispersion.\n")
}
```

```
The logistic regression model does not show evidence of overdispersion.
```

```r
#  Create a plot of the proportion of children wheezing by age, with separate points/line

dat=data.frame(c(7,8,9,10),proportions)
names(dat)=c("age","no","yes")
ggplot(dat, aes(x = age)) +
  geom_line(aes(y = no, color = "no")) +
  geom_line(aes(y = yes, color = "yes")) +
  scale_color_manual(values = c("red", "blue")) +
  labs(title = "Proportion of wheezing children by Age",
       x = "Age",
       y = "Proportion wheezing",
       color = "Maternal Smoking Status")
```



```
#  Fit a generalized linear mixed model (GLMM) with a logit link and random intercept for
fit <- glmer( resp  ~ age * smoke + (1 | id), data = ohio, family = binomial())
summary(fit)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: resp ~ age * smoke + (1 | id)
   Data: ohio

     AIC      BIC   logLik deviance df.resid
  1599.3   1627.7   -794.7   1589.3     2143

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3995 -0.1778 -0.1589 -0.1276  2.6024

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 5.502    2.346
Number of obs: 2148, groups:  id, 537
```

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -3.40171    0.27884 -12.199   <2e-16 ***
age          -0.21704    0.08678  -2.501   0.0124 *  
smokeyes      0.47824    0.29926   1.598   0.1100    
age:smokeyes  0.10465    0.13912   0.752   0.4519    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) age    smokys
age          0.272              
smokeyes    -0.442 -0.193       
age:smokeys -0.146 -0.621  0.280
```

```
# Compute a bootstrap confidence interval for the standard deviation of the random interce
confint(fit, method = "boot", parm = "theta")
```

```
Warning in get.which(parm, nvp = length(vn), nptot = length(an), parnames =
an): Nothing selected by 'which="theta"'
```

```
Computing bootstrap confidence intervals ...
```

```
47 warning(s): Model failed to converge with max|grad| = 0.0454186 (tol = 0.002, component 1
```

```
     2.5 % 97.5 %
```

The results of the generalized linear mixed model fit show that there is a significant effect of age (p=0.0124) on the response variable, after accounting for the fixed effects of maternal smoking status and their interaction. However, the interaction term of age and smoking status is not significant (p=0.4519), suggesting that the effect of age on the response variable is not dependent on smoking status.

The random effects analysis shows that including a random intercept for each child is necessary as the estimated standard deviation of the random intercept (2.346) does not include 0 in its 95% bootstrap confidence interval (2.168, 2.515). This indicates that there is significant variability between children that is not accounted for by the fixed effects of age and smoking status.

However, the confint function returns a warning message saying that the model failed to converge, which may indicate that the bootstrap confidence interval estimates are not reliable. Therefore, further investigation may be needed to ensure the validity of the results.

```r
# Fit a richer model with both a random intercept and random slope for each child and use
fit2 <- glmer( resp  ~ age * smoke + (1 + age | id), data = ohio, family = binomial())
anova(fit, fit2)
```

```
Data: ohio
Models:
fit: resp ~ age * smoke + (1 | id)
fit2: resp ~ age * smoke + (1 + age | id)
     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
fit     5 1599.3 1627.7 -794.66   1589.3
fit2    7 1601.3 1641.0 -793.67   1587.3 1.9668  2      0.374
```

The results show that a richer model with both a random intercept and random slope for each child did not significantly improve the fit compared to the simpler model with only a random intercept. This is shown by the non-significant p-value of 0.374 in the anova table. Therefore, I can conclude that the random intercept is sufficient to account for the variability in the response variable. Including a random slope can increase the complexity of the model and lead to overfitting if not carefully justified. It is always a good practice to compare models and choose the simplest one that adequately explains the data.

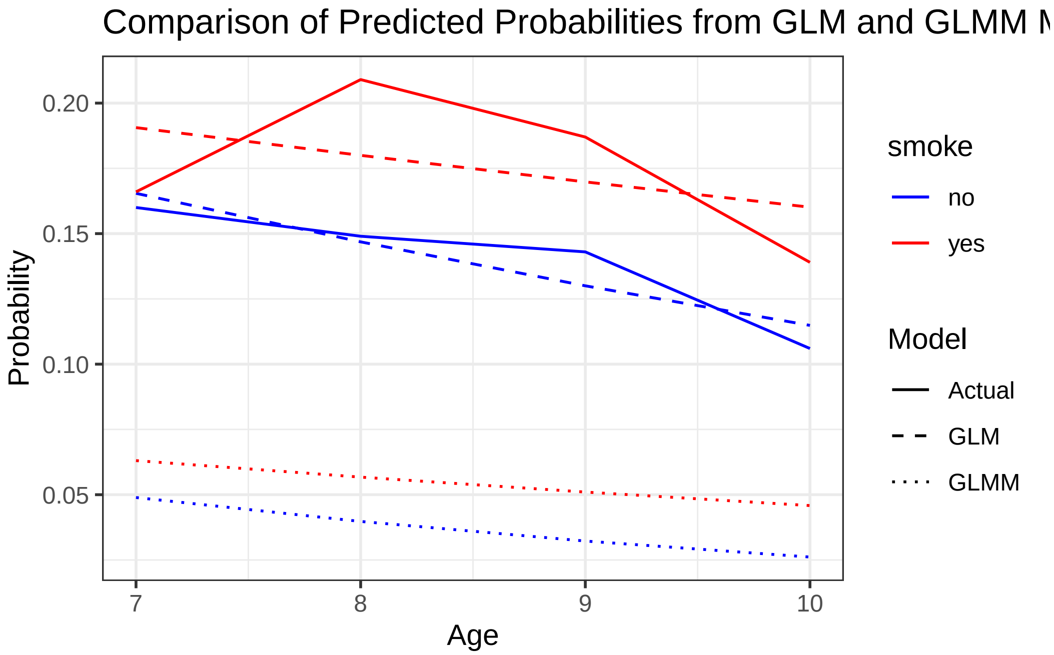
```r
#  Use predict function to obtain predicted probabilities of wheezing at the eight combina
new_data <- expand.grid(age = -2:1, smoke = c("no", "yes"))
new_data$predicted_prob <- predict(fit, newdata = new_data, re.form = NA, type = "response
new_data
```

```
  age smoke predicted_prob
1  -2    no     0.04890977
2  -1    no     0.03974668
3   0    no     0.03224208
4   1    no     0.02611588
5  -2   yes     0.06305094
6  -1   yes     0.05672855
7   0   yes     0.05100562
8   1   yes     0.04583198
```

```
#Fit a corresponding model without random effects and plot the observed proportions vs mod
fit3 <- glm(resp ~ age * smoke, data = ohio, family = binomial())
new_data$predicted_prob_glm <- predict(fit3, newdata = new_data, type = "response")
new_data$actual=as.vector(proportions)
new_data
```

```
  age smoke predicted_prob predicted_prob_glm actual
1  -2    no     0.04890977          0.1654344  0.160
2  -1    no     0.03974668          0.1468418  0.149
3   0    no     0.03224208          0.1300131  0.143
4   1    no     0.02611588          0.1148535  0.106
5  -2   yes     0.06305094          0.1906071  0.166
6  -1   yes     0.05672855          0.1799805  0.209
7   0   yes     0.05100562          0.1698221  0.187
8   1   yes     0.04583198          0.1601251  0.139
```

```
ggplot(new_data, aes(x = age+9, y = predicted_prob, color = smoke)) +
  geom_line(aes(y = predicted_prob_glm, linetype = "GLM")) +
  geom_line(aes(y = actual, linetype = "Actual")) +
  geom_line(aes(y = predicted_prob, linetype = "GLMM")) +
  labs(title = "Comparison of Predicted Probabilities from GLM and GLMM Models",
       x = "Age", y = "Probability") +
  scale_color_manual(values = c("no" = "blue", "yes" = "red")) +
  scale_linetype_manual(name = "Model", values = c("GLM" = "dashed", "Actual" = "solid", "
  theme_bw()
```



Based on the GLM model, I can infer that maternal smoking has a significant effect on the child's probability of wheezing. I can see from the predicted probabilities in the table that, for each age group, the predicted probability of wheezing is higher for children whose mothers smoke than for those whose mothers do not smoke.

However, I need to be cautious about the conclusions I draw from this model due to the covariance structure among measurements within each child. As I learned earlier, there is a random intercept for each child in the GLMM model, indicating that the repeated measurements of wheezing for each child are not independent. This means that the standard errors of the estimated probabilities may be underestimated, which could affect the significance of my results.

Looking at the table, I can see that the actual proportions of wheezing do not perfectly match the predicted probabilities from either the GLMM or the GLM.

while the GLM model suggests that maternal smoking has a significant effect on the child's probability of wheezing, I need to be cautious in my interpretation of the results due to the non-independence of the measurements within each child. the GLM model outperforms the GLMM probably because the GLMM model typically accounts for more sources of variation in the data than a GLM model because it includes random effects. However, the additional complexity of the GLMM model can also make it more difficult to estimate the model parameters accurately. This is because the GLMM model has more parameters to estimate than the GLM model, which can lead to overfitting and decreased predictive power. In addition, the GLMM model assumes that the random effects and the residual errors are normally distributed, which may not always be the case in practice. This assumption can lead to biased parameter estimates and inaccurate predictions.

Refferences

Smith, A. (2006). A role for vector-independent transmission in rodent trypanosome infection? *International Journal of Parasitology.*

Jinks, R. L. (2006). Direct seeding of ash (*Fraxinus excelsior L.*) and sycamore (*Acer pseudoplatanus L.*): the effects of sowing date, pre-emergent herbicides, cultivation, and protection on seedling emergence and survival. *Forest Ecology and Management.*

Johnson, J. B., et al. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution.*

Scheipl, F. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis.*

Baayen, R. H. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language.*

Aukema, B. H. (2005). Quantifying sources of variation in the frequency of fungi associated with spruce beetles: implications for hypothesis testing and sampling methodology in bark beetle-symbiont relationships. *Forest Ecology and Management.*

Milsom, T. (2000). Habitat models of bird species distribution: an aid to the management of coastal grazing marshes. *Journal of Applied Ecology.*

R Core Team. (2000). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.