

STAT 426 Final Project

Jefferson Mathews & Sam DerMatoian

Introduction

- Dr A. Leach tested the effects of a drug on 50 children with a history of otitis media in the Northern Territory of Australia. The children were randomized to the drug or the a placebo, and also to receive active encouragement to comply with taking the drug-. (R Documentation)
- According to the Mayo Clinic, Otitis Media is an infection of the middle ear, which is the air-filled space behind the eardrum. It also notes that these infections are more common in children than adults. Additionally, it notes that infections can go away without treatment over time.

References: <https://www.mayoclinic.org/diseases-conditions/ear-infections/symptoms-causes/syc-20351616>
<https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/bacteria.html>

Introduction (cont.)

- We seek to answer three questions from this analysis:
- Is the treatment effective?
- How do they behave over time?
- Any compliance effect?

Initial Data Analysis

| | Bacteria Absent | Bacteria Present |
|---------|-----------------|------------------|
| Active | 12 | 84 |
| Placebo | 31 | 93 |

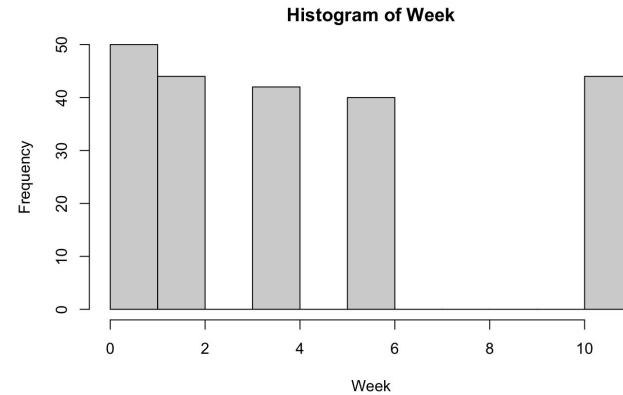
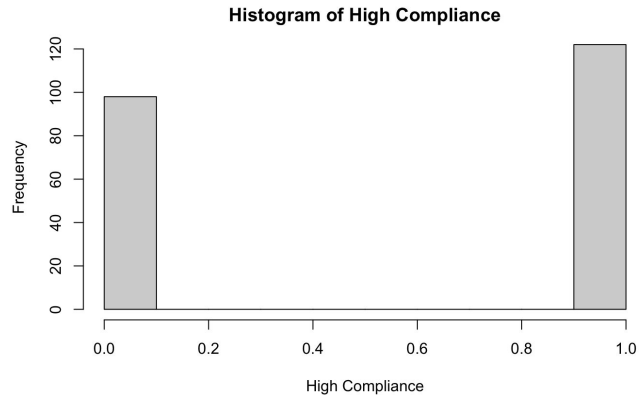
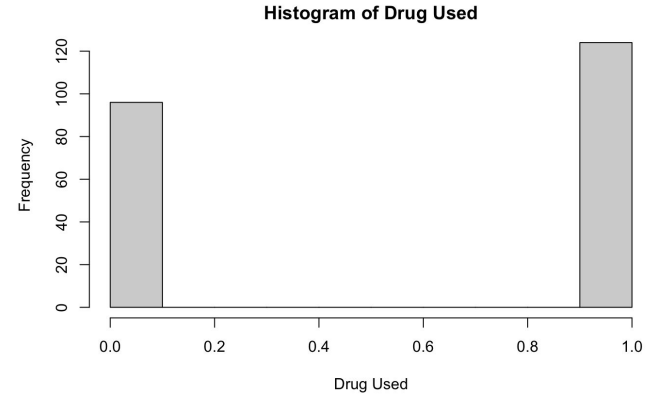
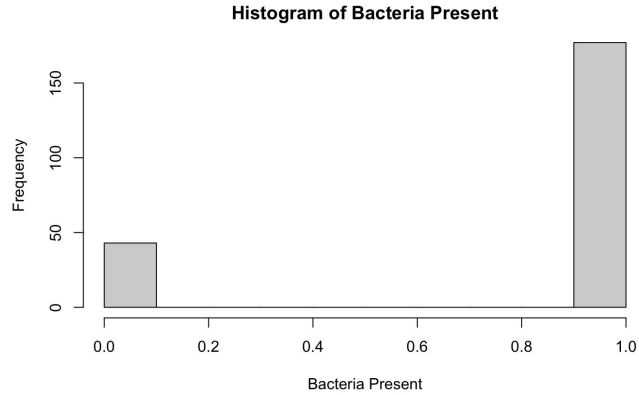
| | Bacteria Absent | Bacteria Present |
|---------|-----------------|------------------|
| Week 0 | 5 | 45 |
| Week 2 | 4 | 40 |
| Week 4 | 11 | 31 |
| Week 6 | 11 | 29 |
| Week 11 | 12 | 32 |

| | Bacteria Absent | Bacteria Present |
|-----------------|-----------------|------------------|
| High Compliance | 24 | 74 |
| Low Compliance | 19 | 103 |

| | Bacteria Absent | Bacteria Present |
|---------|-----------------|------------------|
| Placebo | 12 | 84 |
| Drug | 18 | 44 |
| Drug+ | 13 | 49 |

Initial Data Analysis (cont.)

Note: When checking for NA values, there were none



Initial Data Analysis (cont.)

Number of observations for
each patient

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| X01 | X02 | X03 | X04 | X05 | X06 | X07 | X08 | X09 | X10 |
| 4 | 4 | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 2 |
| X11 | X12 | X13 | X14 | X15 | X16 | X17 | X18 | X19 | X20 |
| 5 | 5 | 5 | 3 | 5 | 5 | 5 | 5 | 5 | 4 |
| X21 | Y01 | Y02 | Y03 | Y04 | Y05 | Y06 | Y07 | Y08 | Y09 |
| 5 | 5 | 5 | 5 | 4 | 3 | 4 | 5 | 4 | 4 |
| Y10 | Y11 | Y12 | Y13 | Y14 | Z01 | Z02 | Z03 | Z05 | Z06 |
| 5 | 5 | 2 | 4 | 3 | 3 | 5 | 5 | 5 | 2 |
| Z07 | Z09 | Z10 | Z11 | Z14 | Z15 | Z19 | Z20 | Z24 | Z26 |
| 5 | 4 | 5 | 5 | 5 | 3 | 5 | 4 | 5 | 5 |

Methods

The variables in the original dataset were changed in name and format for easier modeling. The changes were as follows:

| Old Variable Name | New Variable Name |
|-------------------|-------------------|
| y | Bacteria_Present |
| ap | Drug_Used |
| hilo | High_Compliance |

| Old Format | New Format |
|------------|------------|
| y/n | 0/1 |
| a/p | 0/1 |
| hi/lo | 0/1 |

The Week variable was treated as a continuous variable due to the sporadic time intervals the dataset was in. The other variables were stored as binary variables.

The trt variable, which combines High_Compliance and Drug_Used wasn't changed, but was treated as a factor variable with levels 0, 1, 2

Explanation of Model Methods

The Generalized Linear Model helps to answer the research questions by modeling the presence of bacteria as a function of time(Week), the treatment, and compliance. Due to the trt predictor's collinearity, it was left out and ID was also left out due to its irrelevance in the GLM model. The model was built as an additive model and the compliance predictor was dropped from the final model because it wasn't significant

The Generalized Linear Mixed Model was built in a similar way to the GLM, but it used the ID predictor as a random effect variable. This gave a random intercept to each ID instead of one intercept for the whole model. The analysis of the data showed this may be required for an accurate model, so it was taken into account.

Interaction

Interaction in the model was checked by multiplying trt and week. Compliance and Drug Use were checked as well against each other and week. None of these interactions provided any significant predictors and all models with interaction terms had poor AIC values, deviance values, and predictive strength.

Therefore interaction was not required to answer any of the research questions

Results

GLM model with the lowest AIC (210.95) used just week and Drug_Used.

GLMM model with lowest AIC(206.4) used Week and Drug_Used with a random effect of Patient ID

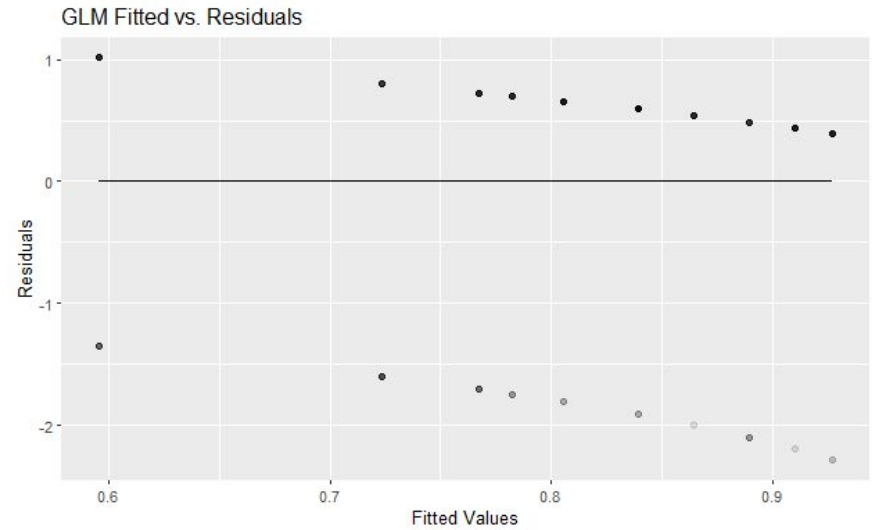
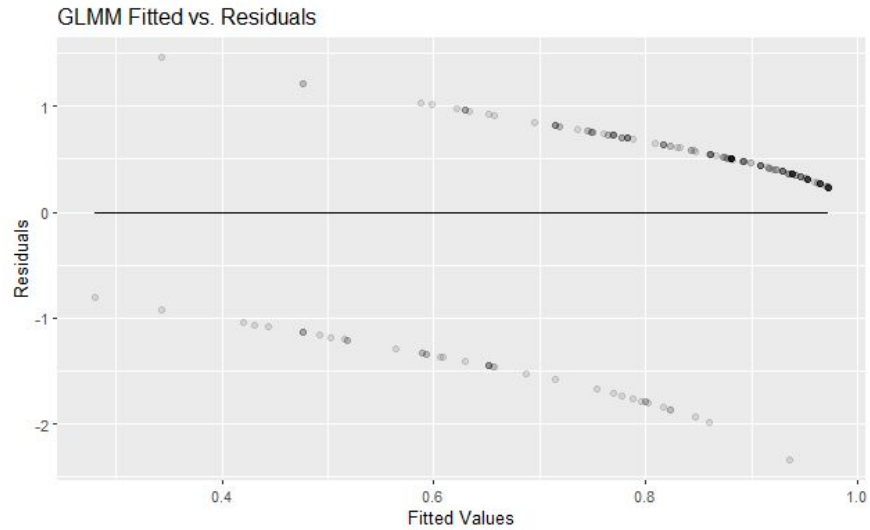
GLM Model

| Variable | Coefficient |
|-----------|-------------|
| Week | -0.11479 |
| Drug Used | -0.89034 |

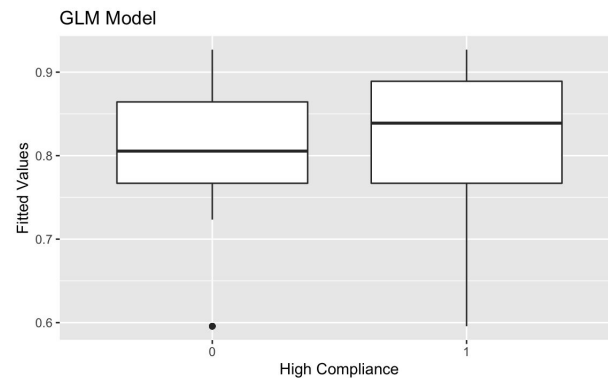
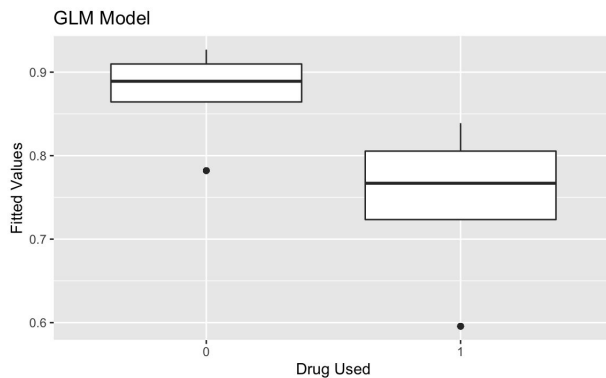
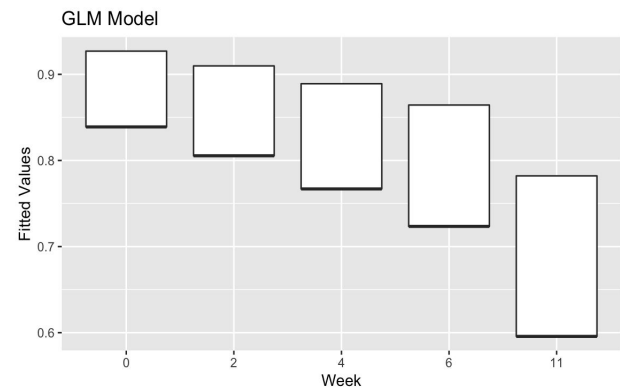
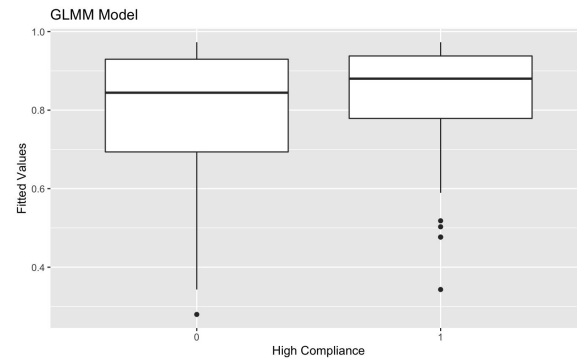
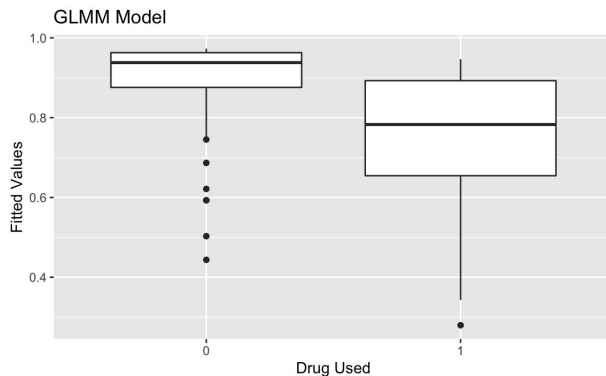
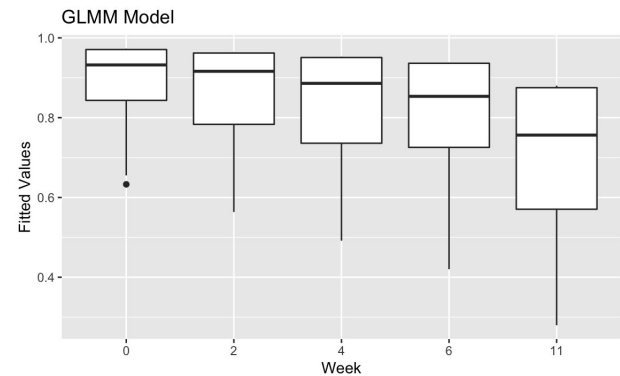
GLMM Model

| Variable | Coefficient | Variance |
|----------------------------------|-------------|----------|
| Week | -0.14439 | |
| Drug Used | -1.07565 | |
| Patient ID (Random effect) | | 1.401 |

Residual Analysis



Fitted Values vs Predictor Boxplots



Summary

The GLMM model is preferred based the results of the random effects and the comparison of the AIC values of the two models.

Time did have an impact on the presence, the probability of the bacteria being there decreased with time. This corresponds to the knowledge discovered in the beginning of presentation that these infections go away with time.

The drug did have an impact according to both models, so taking it may have helped the infection go away

Compliance had no impact either was, so there was most likely no compliance effect

R Code

```
library(MASS)
library(ggplot2)
bacteria1 = bacteria
```

****Initial Data Analysis****

```
bacteria1$y<-ifelse(bacteria1$y=="y",1,0)
bacteria1$ap<-ifelse(bacteria1$ap=="a",1,0)
bacteria1$hilo<-ifelse(bacteria1$hilo=="hi",1,0)
colnames(bacteria1) = c("Bacteria_Present", "Drug_Used", "High_Compliance", "Week", "ID", "trt")
bacteria1
```

```
summary(bacteria1)
```

```
hist(bacteria1$Bacteria_Present, main = "Histogram of Bacteria Present", xlab = "Bacteria Present")
hist(bacteria1$Drug_Used, main = "Histogram of Drug Used", xlab = "Drug Used")
hist(bacteria1$High_Compliance, main = "Histogram of High Compliance", xlab = "High Compliance")
hist(bacteria1$Week, main = "Histogram of Week", xlab = "Week")
```

```
summary(bacteria1$ID)
```

```
bact_drug_tab = table(bacteria1$Drug_Used, bacteria1$Bacteria)
colnames(bact_drug_tab) = c("No Bacteria", "Bacteria")
rownames(bact_drug_tab) = c("Active", "Placebo")
bact_drug_tab
```

```
bact_comp_tab = table(bacteria1$High_Compliance, bacteria1$Bacteria)
colnames(bact_comp_tab) = c("No Bacteria", "Bacteria")
rownames(bact_comp_tab) = c("High Compliance", "Low Compliance")
bact_comp_tab
```

```
bact_week_tab = table(bacteria1$Week, bacteria1$Bacteria)
colnames(bact_week_tab) = c("No Bacteria", "Bacteria")
rownames(bact_week_tab) = c("Week 0", "Week 2", "Week 4", "Week 6", "Week 11")
bact_week_tab
```

```
bact_trt_tab = table(bacteria1$trt, bacteria1$Bacteria)
colnames(bact_trt_tab) = c("No Bacteria", "Bacteria")
bact_trt_tab
```

```
which(is.na(bacteria1))
```

****Modeling****

```
glm_mod = glm(Bacteria_Present ~ Week + High_Compliance + Drug_Used,
family = binomial, data = bacteria1)
summary(glm_mod)
```

```
glm_mod1 = glm(Bacteria_Present ~ Week + Drug_Used, family = binomial, data
= bacteria1)
summary(glm_mod1) ##Best GLM model
```

```
glm_mod2 = glm(Bacteria_Present ~ factor(Week) + High_Compliance +
Drug_Used, family = binomial, data = bacteria1)
summary(glm_mod2)
```

```
glm_mod3 = glm(Bacteria_Present ~ factor(Week) * trt, family = binomial, data =
bacteria1)
summary(glm_mod3)
```

```
library(lme4)
glmm_mod = glmer(Bacteria_Present ~ Week + High_Compliance + Drug_Used
+(1 | ID) , data = bacteria1, family = binomial)
summary(glmm_mod)
```

```
glmm_mod2 = glmer(Bacteria_Present ~ factor(Week)+ High_Compliance +
Drug_Used +(1 | ID) , data = bacteria1, family = binomial)
summary(glmm_mod2)
```

```
glmm_mod3 = glmer(Bacteria_Present ~ factor(Week) + Drug_Used +(1 | ID) ,
data = bacteria1, family = binomial)
summary(glmm_mod3)
```

```
glmm_mod4 = glmer(Bacteria_Present ~ Week + Drug_Used + (1 | ID) , data =
bacteria1, family = binomial)
summary(glmm_mod4) ##Selected model
```

R Code (cont.)

****Model Analysis****

```
ggplot(bacteria1, aes(x = factor(bacteria1$Week), y = glm_mod1$fitted.values)) +  
  geom_boxplot() +  
  labs(title = "GLM Model", x = "Week", y = "Fitted Values")
```

```
ggplot(bacteria1, aes(x = factor(bacteria1$Drug_Used), y =  
  glm_mod1$fitted.values)) +  
  geom_boxplot() +  
  labs(title = "GLM Model", x = "Drug Used", y = "Fitted Values")
```

```
ggplot(bacteria1, aes(x = factor(bacteria1$High_Compliance), y =  
  glm_mod1$fitted.values)) +  
  geom_boxplot() +  
  labs(title = "GLM Model", x = "High Compliance", y = "Fitted Values")
```

```
ggplot(bacteria1, aes(x = factor(bacteria1$Week), y = fitted.values(glm_mod4))) +  
  geom_boxplot() +  
  labs(title = "GLMM Model", x = "Week", y = "Fitted Values")
```

```
ggplot(bacteria1, aes(x = factor(bacteria1$Drug_Used), y =  
  fitted.values(glm_mod4))) +  
  geom_boxplot() +  
  labs(title = "GLMM Model", x = "Drug Used", y = "Fitted Values")
```

```
ggplot(bacteria1, aes(x = factor(bacteria1$High_Compliance), y =  
  fitted.values(glm_mod4))) +  
  geom_boxplot() +  
  labs(title = "GLMM Model", x = "High Compliance", y = "Fitted Values")
```

```
ggplot(mapping = aes(x = fitted.values(glm_mod4), y = resid(glm_mod4))) +  
  geom_point(alpha = 0.1) +  
  geom_line(y = 0) +  
  labs(x = "Fitted Values", y = "Residuals", title = "GLMM Fitted vs. Residuals")
```

```
ggplot(mapping = aes(x = fitted.values(glm_mod1), y = resid(glm_mod1))) +  
  geom_point(alpha = 0.1) +  
  geom_line(y = 0) +  
  labs(x = "Fitted Values", y = "Residuals", title = "GLM Fitted vs. Residuals")
```