Data Analysis Exam

Xinyi Lin

11/29/2019

## -- Attaching packages --------------------------------------------------------------------------- tidyverse 1.2.1 --

## <U+2713> ggplot2 3.2.1 <U+2713> purrr 0.3.3  
## <U+2713> tibble 2.1.3 <U+2713> dplyr 0.8.3  
## <U+2713> tidyr 1.0.0 <U+2713> stringr 1.4.0  
## <U+2713> readr 1.3.1 <U+2713> forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

##   
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':  
##   
## collapse

## sid Treatment Gender ND0 ND30   
## Min. : 1.00 1:106 Female:143 Min. : 44.0 Min. : 16.00   
## 1st Qu.: 79.25 2:107 Male :171 1st Qu.: 68.0 1st Qu.: 41.25   
## Median :157.50 3:101 Median :154.0 Median : 73.00   
## Mean :157.50 Mean :130.3 Mean : 78.24   
## 3rd Qu.:235.75 3rd Qu.:184.0 3rd Qu.:112.75   
## Max. :314.00 Max. :252.0 Max. :167.00   
## ND60 Relapse  
## Min. : 17.00 0:164   
## 1st Qu.: 34.00 1:150   
## Median : 68.00   
## Mean : 67.52   
## 3rd Qu.: 91.00   
## Max. :175.00

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | p | test |
| n | 106 | 107 | 101 |  |  |
| Gender = Male (%) | 52 (49.1) | 62 (57.9) | 57 (56.4) | 0.381 |  |
| ND0 (mean (SD)) | 125.71 (62.70) | 134.08 (59.11) | 131.02 (59.62) | 0.594 |  |
| ND30 (mean (SD)) | 84.78 (42.16) | 89.27 (41.83) | 59.69 (27.63) | <0.001 |  |
| ND60 (mean (SD)) | 83.13 (42.26) | 60.26 (28.39) | 58.81 (28.46) | <0.001 |  |
| Relapse = 1 (%) | 87 (82.1) | 49 (45.8) | 14 (13.9) | <0.001 |  |

## Question 1

In the 30 days following randomization:

, where follows Poisson distribution.

##   
## Call:  
## glm(formula = ND30 ~ Treatment + Gender + ND0, family = poisson(),   
## data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3075 -0.7638 -0.1089 0.8062 3.2370   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.5442082 0.0266509 132.986 <2e-16 \*\*\*  
## Treatment2 -0.0156420 0.0147692 -1.059 0.29   
## Treatment3 -0.3999763 0.0167293 -23.909 <2e-16 \*\*\*  
## GenderMale 0.5592504 0.0395059 14.156 <2e-16 \*\*\*  
## ND0 0.0039075 0.0003108 12.572 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 6504.4 on 313 degrees of freedom  
## Residual deviance: 399.5 on 309 degrees of freedom  
## AIC: 2310.7  
##   
## Number of Fisher Scoring iterations: 4

## Waiting for profiling to be done...

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Exp.Est | Std.Error | CIL | CIU |
| (Intercept) | 3.5442 | 34.6123 | 0.0267 | 3.4919 | 3.5964 |
| Treatment2 | -0.0156 | 0.9845 | 0.0148 | -0.0446 | 0.0133 |
| Treatment3 | -0.4000 | 0.6703 | 0.0167 | -0.4328 | -0.3672 |
| GenderMale | 0.5593 | 1.7494 | 0.0395 | 0.4818 | 0.6367 |
| ND0 | 0.0039 | 1.0039 | 0.0003 | 0.0033 | 0.0045 |

According to results above, we can find that the total number of drinks ratio between people who get Treatment 2 and people who get Treatment 1 is 0.9844797 and the total number of drinks ratio between people who get Treatment 3 and people who get Treatment 1 is 0.6703359. This means for the 30 days following randomization, the effect of Treatment 3 is better than that of Treatment 2 and the effect of Treatment 2 is better than that of Treatment 1.

In the 30 days between the and day follow-up:

, where follows Poisson distribution.

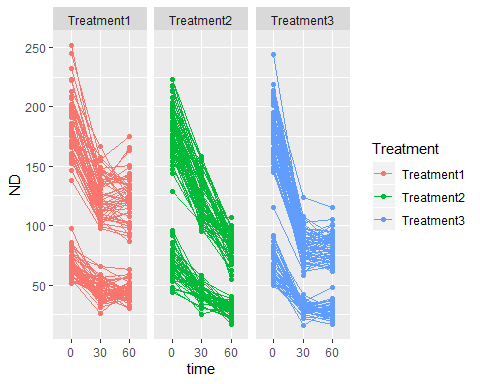
##   
## Call:  
## glm(formula = ND60 ~ Treatment + Gender + ND0 + ND30, family = poisson(),   
## data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.11224 -0.71270 -0.05254 0.64616 2.91976   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.4741130 0.0304486 114.098 <2e-16 \*\*\*  
## Treatment2 -0.3899279 0.0164667 -23.680 <2e-16 \*\*\*  
## Treatment3 -0.3531096 0.0251297 -14.051 <2e-16 \*\*\*  
## GenderMale 0.5567705 0.0438727 12.691 <2e-16 \*\*\*  
## ND0 0.0034088 0.0004005 8.512 <2e-16 \*\*\*  
## ND30 0.0012420 0.0005328 2.331 0.0198 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 5846.8 on 313 degrees of freedom  
## Residual deviance: 337.1 on 308 degrees of freedom  
## AIC: 2202.1  
##   
## Number of Fisher Scoring iterations: 4

## Waiting for profiling to be done...

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Exp.Est | Std.Error | CIL | CIU |
| (Intercept) | 3.4741 | 32.2692 | 0.0304 | 3.4144 | 3.5337 |
| Treatment2 | -0.3899 | 0.6771 | 0.0165 | -0.4222 | -0.3577 |
| Treatment3 | -0.3531 | 0.7025 | 0.0251 | -0.4024 | -0.3039 |
| GenderMale | 0.5568 | 1.7450 | 0.0439 | 0.4708 | 0.6427 |
| ND0 | 0.0034 | 1.0034 | 0.0004 | 0.0026 | 0.0042 |
| ND30 | 0.0012 | 1.0012 | 0.0005 | 0.0002 | 0.0023 |

According to results above, we can find that the total number of drinks ratio between people who get Treatment 2 and people who get Treatment 1 is 0.6771056 and the total number of drinks ratio between people who get Treatment 3 and people who get Treatment 1 is 0.7024999. This means for the 30 days following randomization, the effect of Treatment 2 is slightly better than that of Treatment 3 and the effect of Treatment 3 is better than that of Treatment 1.

## Question 2



Let be the numbers of drinks consumed in the 30 days for subject at visit . Assume follows poisson distribution

Model: .

##   
## Call:  
## geeglm(formula = ND ~ Treatment + time + Treatment \* time + ND0,   
## family = poisson(), data = spr\_data, id = sid, corstr = "ar1")  
##   
## Coefficients:  
## Estimate Std.err Wald Pr(>|W|)   
## (Intercept) 3.3011838 0.0226743 21196.951 <2e-16 \*\*\*  
## Treatment2 0.0049309 0.0228209 0.047 0.829   
## Treatment3 -0.3768966 0.0249864 227.529 <2e-16 \*\*\*  
## time60 -0.0196647 0.0233668 0.708 0.400   
## ND0 0.0080519 0.0001221 4345.400 <2e-16 \*\*\*  
## Treatment2:time60 -0.3733159 0.0319773 136.292 <2e-16 \*\*\*  
## Treatment3:time60 0.0047926 0.0340681 0.020 0.888   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Estimated Scale Parameters:  
## Estimate Std.err  
## (Intercept) 1.793 0.1015  
##   
## Correlation: Structure = ar1 Link = identity   
##   
## Estimated Correlation Parameters:  
## Estimate Std.err  
## alpha 0 0  
## Number of clusters: 628 Maximum cluster size: 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Exp.Est | Std.Error | CIL | CIU |
| (Intercept) | 3.3012 | 27.1448 | 0.0227 | 3.2567 | 3.3456 |
| Treatment2 | 0.0049 | 1.0049 | 0.0228 | -0.0398 | 0.0497 |
| Treatment3 | -0.3769 | 0.6860 | 0.0250 | -0.4259 | -0.3279 |
| time60 | -0.0197 | 0.9805 | 0.0234 | -0.0655 | 0.0261 |
| ND0 | 0.0081 | 1.0081 | 0.0001 | 0.0078 | 0.0083 |
| Treatment2:time60 | -0.3733 | 0.6884 | 0.0320 | -0.4360 | -0.3106 |
| Treatment3:time60 | 0.0048 | 1.0048 | 0.0341 | -0.0620 | 0.0716 |

According to model results, we can find that the ratio of drinks consumed changes between ND30 and ND60 for group in Treatment 2 versus Treatment 1 is significant with value equals to 0.6884 and the ratio of drinks consumed changes between ND30 and ND60 for group in Treatment 3 versus Treatment 1 is not significant with value equals to 1.0048. So the pattern of change in the number of drinks consumed between the Treatment 1 group and Treatment 2 group is difference and the same between the Treatment 1 group and Treatment 3 group.

## Question 3

Model: , where follows Poisson distribution.

##   
## Call:  
## glm(formula = ND0 ~ Gender, family = poisson(), data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.387 -1.112 -0.094 1.032 4.836   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.2110 0.0102 413.5 <2e-16 \*\*\*  
## GenderMale 0.9975 0.0116 85.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 9235.21 on 313 degrees of freedom  
## Residual deviance: 744.34 on 312 degrees of freedom  
## AIC: 2816  
##   
## Number of Fisher Scoring iterations: 4

## Waiting for profiling to be done...

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Exp.Est | Std.Error | CIL | CIU |
| (Intercept) | 4.2110 | 67.427 | 0.0102 | 4.1910 | 4.231 |
| GenderMale | 0.9975 | 2.712 | 0.0116 | 0.9747 | 1.020 |

According to above results, we can find that confidence interval of Gender across 0, which means coefficient of Gender is significant with value equals to 0.9975. This means on average, total numder drinks male consumed in baseline 30-day period is 2.7115 times compared to that of female.

## Question 4

Let be the numbers of drinks consumed in the 30 days for subject at visit . Assume follows poisson distribution

Model: .

##   
## Call:  
## geeglm(formula = ND ~ Treatment + time + Treatment \* time + ND0 +   
## Gender \* Treatment, family = poisson(), data = spr\_data,   
## id = sid, corstr = "ar1")  
##   
## Coefficients:  
## Estimate Std.err Wald Pr(>|W|)   
## (Intercept) 3.551946 0.024646 20770.60 <2e-16 \*\*\*  
## Treatment2 -0.065641 0.026406 6.18 0.013 \*   
## Treatment3 -0.424745 0.028744 218.35 <2e-16 \*\*\*  
## time60 -0.019665 0.017760 1.23 0.268   
## ND0 0.003956 0.000235 282.45 <2e-16 \*\*\*  
## GenderMale 0.538337 0.034562 242.62 <2e-16 \*\*\*  
## Treatment2:time60 -0.373316 0.024639 229.56 <2e-16 \*\*\*  
## Treatment3:time60 0.004793 0.026260 0.03 0.855   
## Treatment2:GenderMale 0.064446 0.027364 5.55 0.019 \*   
## Treatment3:GenderMale 0.033038 0.029957 1.22 0.270   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Estimated Scale Parameters:  
## Estimate Std.err  
## (Intercept) 1.17 0.0664  
##   
## Correlation: Structure = ar1 Link = identity   
##   
## Estimated Correlation Parameters:  
## Estimate Std.err  
## alpha 0 0  
## Number of clusters: 628 Maximum cluster size: 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Exp.Est | Std.Error | CIL | CIU |
| (Intercept) | 3.552 | 34.881 | 0.025 | 3.504 | 3.600 |
| Treatment2 | -0.066 | 0.936 | 0.026 | -0.117 | -0.014 |
| Treatment3 | -0.425 | 0.654 | 0.029 | -0.481 | -0.368 |
| time60 | -0.020 | 0.980 | 0.018 | -0.054 | 0.015 |
| ND0 | 0.004 | 1.004 | 0.000 | 0.004 | 0.004 |
| GenderMale | 0.538 | 1.713 | 0.035 | 0.471 | 0.606 |
| Treatment2:time60 | -0.373 | 0.688 | 0.025 | -0.422 | -0.325 |
| Treatment3:time60 | 0.005 | 1.005 | 0.026 | -0.047 | 0.056 |
| Treatment2:GenderMale | 0.064 | 1.067 | 0.027 | 0.011 | 0.118 |
| Treatment3:GenderMale | 0.033 | 1.034 | 0.030 | -0.026 | 0.092 |

According to results above, we can find that for treatment 1, the log count difference between

## Question 5

Model

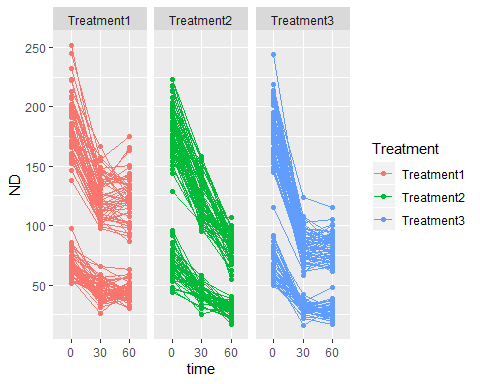
##   
## Call:  
## glm(formula = Relapse ~ Treatment + Gender + ND0, family = binomial(link = "logit"),   
## data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.928 -0.586 -0.518 0.644 2.049   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.81883 0.59800 3.04 0.0024 \*\*   
## Treatment2 -1.70055 0.32082 -5.30 1.2e-07 \*\*\*  
## Treatment3 -3.36665 0.38619 -8.72 < 2e-16 \*\*\*  
## GenderMale 0.49704 0.90733 0.55 0.5838   
## ND0 -0.00429 0.00749 -0.57 0.5670   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 434.67 on 313 degrees of freedom  
## Residual deviance: 328.23 on 309 degrees of freedom  
## AIC: 338.2  
##   
## Number of Fisher Scoring iterations: 4

## Waiting for profiling to be done...

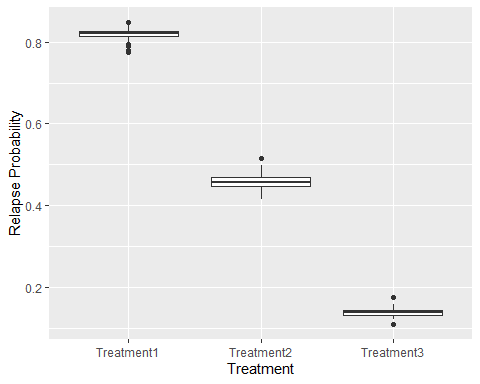
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Exp.Est | Std.Error | CIL | CIU |
| (Intercept) | 1.819 | 6.165 | 0.598 | 0.666 | 3.02 |
| Treatment2 | -1.701 | 0.183 | 0.321 | -2.349 | -1.09 |
| Treatment3 | -3.367 | 0.034 | 0.386 | -4.160 | -2.64 |
| GenderMale | 0.497 | 1.644 | 0.907 | -1.290 | 2.28 |
| ND0 | -0.004 | 0.996 | 0.008 | -0.019 | 0.01 |

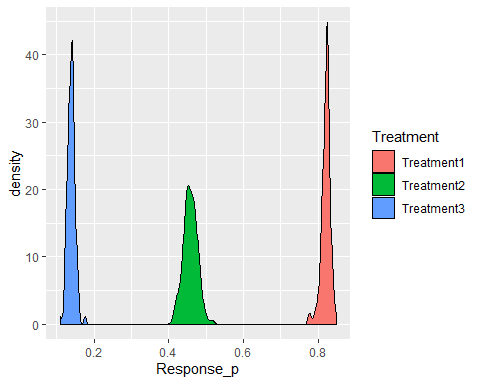
As confidence interval of coefficients for treatment 2 and treatment 3 do not across 0, we can conclude that treatment effects difference between treatment 2, treatment 3 versus treatment 1 are significant. The odd ratio of relapse between people in treatment 2 versus people in treatment 1 is 0.183 and the odd ratio of relapse between people in treatment 2 versus people in treatment 1 is 0.035. In general, treatment 3 have best treatment effect on preventing relapse and treatment effect of treatment 2 is better than treatment 1.

## Question 6



According to the above plot, we can find that for people in Treatment group 1 and 3, the treatment effects decrease in the 30 days between the and day follow-up. Total numbers of drinks consumed in the 30 days following randomization are similar to total numbers of drinks consumed in the 30 days between the and day follow-up. While total numbers of drinks consumed in 30 days keep decrease for people in Treatment group2. According to these trends, treatment 2 might be beneficial once the treatment has stopped.





## Others

Let be the average numbers of drinks consumed in the 30 days for treatemnt groups at visit .

There is no difference among groups.

### Question 2

Patterns of change among different groups are the same

## Loading required package: RColorBrewer

## Loading required package: reshape

##   
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':  
##   
## rename

## The following objects are masked from 'package:tidyr':  
##   
## expand, smiths

## Loading required package: lavaan

## This is lavaan 0.6-5

## lavaan is BETA software! Please report any bugs.

## Call:  
## pbg(data = data[, 4:6], group = data$Treatment)  
##   
## Hypothesis Tests:  
## $`Ho: Profiles are parallel`  
## Multivariate.Test Statistic Approx.F num.df den.df p.value  
## 1 Wilks 0.454 75.1 4 620 6.33e-52  
## 2 Pillai 0.619 69.7 4 622 8.75e-49  
## 3 Hotelling-Lawley 1.044 80.6 4 618 4.79e-55  
## 4 Roy 0.856 133.1 2 311 1.75e-42  
##   
## $`Ho: Profiles have equal levels`  
## Df Sum Sq Mean Sq F value Pr(>F)   
## group 2 12213 6107 3.27 0.039 \*  
## Residuals 311 580275 1866   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $`Ho: Profiles are flat`  
## F df1 df2 p-value  
## 1 656 2 310 4.03e-112

## Question 5

Let be whether relapse after 6 months for subject . follows poisson distribution

Model: where .

knitr::opts\_chunk$set(echo = FALSE)  
library(gee)  
library(tidyverse)  
library(geepack)  
library(nlme)  
library(tableone)  
data = read.csv("./ALCDEP.txt", sep = "\t") %>%   
 mutate(Gender = ifelse(Gender==0, "Male", "Female"),  
 Gender = as.factor(Gender),  
 Treatment = as.factor(Treatment),  
 Relapse = as.factor(Relapse))  
summary(data)  
# Table1  
factorVars <- c("Treatment","Gender","Relapse")  
vars <- c("Gender","ND0", "ND30", "ND60", "Relapse")  
tableOne <- CreateTableOne(vars = vars, strata = "Treatment", data = data, factorVars = factorVars)  
tableOne %>% kableone()  
# Table2  
glm.model1 = glm(ND30~Treatment+Gender+ND0, family = poisson(), data = data)  
summary(glm.model1)  
confint\_glm <- function(object, parm, level = 0.95, ...) {  
 coef = coef(summary(object)) %>% as.data.frame()  
 coef\_CI = object %>% confint() %>% as.data.frame()  
 table = cbind(coef, coef\_CI) %>%   
 mutate(Exp.Est = round(exp(Estimate),4),  
 CIL = round(`2.5 %`,4),  
 CIU = round(`97.5 %`,4),  
 Std.Error = round(`Std. Error`,4),  
 Estimate = round(Estimate,4)) %>%  
 select(Estimate, Exp.Est, Std.Error, CIL, CIU)  
 rownames(table) <- rownames(coef)  
 return(table)  
}  
confint\_glm(glm.model1) %>% knitr::kable()  
# Table 3  
glm.model2 = glm(ND60~Treatment+Gender+ND0 + ND30, family = poisson(), data = data)  
summary(glm.model2)  
confint\_glm(glm.model2) %>% knitr::kable()  
# Figure1  
seq\_data = data %>%   
 gather(key = "time", value = "ND", ND0:ND60) %>%   
 mutate(time = str\_replace(time, "ND", "")) %>%   
 mutate(Gender = as.factor(Gender),  
 Treatment = str\_c("Treatment",Treatment))  
seq\_data %>%   
 ggplot(aes(x = time, y = ND, group = sid, color = Treatment)) +  
 geom\_point()+  
 geom\_line()+  
 facet\_grid(. ~ Treatment)  
# Table4  
spr\_data = data %>%   
 gather(key = "time", value = "ND", ND30:ND60) %>%   
 mutate(time = str\_replace(time, "ND", "")) %>%   
 mutate(Gender = as.factor(Gender),  
 Treatment = as.factor(Treatment))  
gee.model = geeglm(ND ~ Treatment + time + Treatment\*time + ND0, id = sid, data = spr\_data, family = poisson(), corstr="ar1")  
summary(gee.model)  
# CI  
confint\_geeglm <- function(object, parm, level = 0.95, ...) {  
 cc <- coef(summary(object))  
 mult <- qnorm((1+level)/2)  
 citab <- with(as.data.frame(cc),  
 cbind(Estimate = round(Estimate,4),  
 Exp.Est = round(exp(Estimate),4),  
 Std.Error = round(Std.err,4),  
 CIL=round(Estimate-mult\*Std.err,4),  
 CIU=round(Estimate+mult\*Std.err,4)))  
 rownames(citab) <- rownames(cc)  
 citab[parm,]  
}  
  
confint\_geeglm(gee.model) %>% knitr::kable()  
# Table5  
glm.model3 = glm(ND0 ~ Gender, data = data, family = poisson())  
summary(glm.model3)  
confint\_glm(glm.model3) %>% knitr::kable()  
# Table6  
gee.model2 = geeglm(ND ~ Treatment + time + Treatment\*time + ND0 + Gender\*Treatment, id = sid, data = spr\_data, family = poisson(), corstr="ar1")  
summary(gee.model2)  
confint\_geeglm(gee.model2)%>% knitr::kable()  
# Table7  
glm.model3 = glm(Relapse ~ Treatment + Gender + ND0, data = data, family = binomial(link = "logit"))  
summary(glm.model3)  
confint\_glm(glm.model3) %>% knitr::kable()  
# delete  
seq\_data = data %>%   
 gather(key = "time", value = "ND", ND0:ND60) %>%   
 mutate(time = str\_replace(time, "ND", "")) %>%   
 mutate(Gender = as.factor(Gender),  
 Treatment = str\_c("Treatment",Treatment))  
seq\_data %>%   
 ggplot(aes(x = time, y = ND, group = sid, color = Treatment)) +  
 geom\_point()+  
 geom\_line()+  
 facet\_grid(. ~ Treatment)  
# Figure2  
data %>%   
 mutate(Response\_p = predict(glm.model3,type="response"),  
 Treatment = str\_c("Treatment", Treatment)) %>%   
 ggplot(aes(y = Response\_p, x = Treatment)) +  
 geom\_boxplot() +  
 scale\_y\_continuous(name="Relapse Probability")  
# delete  
predic\_data = data %>%   
 mutate(Response\_p = predict(glm.model3,type="response"),  
 Treatment = str\_c("Treatment", Treatment))  
predic\_data %>%   
 ggplot(aes(x = Response\_p, fill = Treatment)) +  
 geom\_density(stat = "density")   
library(profileR)  
model = pbg(data[,4:6], data$Treatment)  
summary(model)