Final

Sijia Yue

12/9/2019

Abstract

Introudction

Alcohol has been part of human culture for a long time and the history of alcohol could be traced back to 10,000 year ago. There's also an association between public health status and alcohol consumption. Alcohol is viewed as a risk factor of various disease, especially for alcohol-use disovers.

Based on these facts, a group of researchers designed a randomized control trail to study the longitudinal effect among 3 types of interventions. The aim of this study is to see if the interventions were help in reducing alcohol dependence. There were 314 subjects (n=314) that suffer from alcohol disorder were recruited in this study and they were randomized into 3 treatment type (1, 2, 3). Then the total number of drinks consumed in the past 30 days were recorded at day 0 (before randomization), day 30 and day 60. The subjects were also followed up 6 months after the end of treatment. Based on an a priori set criteria, the researchers classified the subjects as having relapsed into alcohol dependence or not.

In this report, the data analysis and model building are based on the results from this research. Exploratory data analysis are provided to have a deeper understand of this dataset. Several generalized linear model and mixed effect model are build to make inferences.

Details of data

There are 314 observations in this dataset(n=314).

There are 7 columns, including oberservation ID(sid), treatment, gender, number of consumption count at day 0, day 30 and day 60(ND0, ND30, ND60) and relapse status after the trail.

- Observation ID is the unique identifier for each patient, it is an integer.
- Treatment is a 3 level character, indicating 3 treatment arms (1, 2, 3).
- Gender is a 2 level character (0=Male, 1=Female).
- Number of consumption counts are count numbers.
- Relapse is a 2 level character, indicating the obervation having relapsed into alcohol dependence or not (0=No, 1=Yes).

There is no missing data, so it is a balanced design.

Objectives

There are 6 main questions this paper would focus on:

- Is there evidence to suggest that the treatments differ in their effects on alcohol dependence, as reflected by the number of drinks consumed in a given 30 day period?
- Is there a difference in the pattern of change in the number of drinks consumed between the various treatment groups over the duration of the study?

- Alcohol-use disorders are among the most disabling disease categories for the global burden of disease especially for men. Is there evidence to suggest that males tend to have a higher alcohol dependence than females?
- Do men and women respond differently to treatment?
- Is there any evidence to suggest that the treatments differ in their effects on subjects with regard to relapsing into alcohol dependence ?
- Even in the case that the treatments might differ in their pattern of change or on how subjects relapse into alcohol dependence, is there any evidence to suggest that any of the treatments might be beneficial once the treatment has stopped.

Data Analysis

Explortory Data Analysis

To have a thorough and vivid understanding about this dataset, this paper would provide several plots.

- Explore the change of consumption according to time stratifying treatment.

`Figure 1` is a spagatti plot illustrating the trend of change in differnt treatment group. Overall, there's a decrease in alcohol consumption according to time and there's an obvious drop from baseline to day 30 in all treatment groups. From day 30 to day 60, the pattern of change differs according to treatment. In treatment 2, it still shows a decrease in 30 days. However, in treatment 1 and 3, the plot keeps flat and some observations even have an increasing trend of alcohol consumption.

These trends also reflected by `Figure 2`. The boxplot demonstated the distribution of alcohol consumption along time in different treatment group.

So, we could have an idea that the effect of different treatment groups are quiet different. Also, we could notice that there's an gap between observations in `Figure 1`, it probobaly shows the gender difference. So, let's plot another one stratifying the gender.

- Explore the change of consumption according to time stratifying gender.

`Figure 3` is a spagatti plot illustrating the trend of change in differnt gender. The change of consumptions are pretty clear – male consumes more alcohol than female. Also, female are more likely to control their consumption since most of the lines keep flat after day 30. However, the trend varies a lot in men. Some lines from male show a decreasing trend, but most of the lines keep flat or even increases.

`Figure 4` shows the overall trend of change in consumption statifying two genders. The number of consumptions are decreasing along time in both genders, but male has much higher consumptions than female. This results demonstrate that the gender difference is pretty high, so it should be considered.

- Check the normality of data

Before we jump to model fitting, the last thing we need to do in EDA is to check the data normality. This is really important because each model has different assumptions, so we need to check the data first.

`Figure 5` shows the distribution of consumption stratifying gender and timepoints. The density plots are not strictly normal distribution. Since the number of consumption is count data, so we assume the consumption follows Poisson distribution.

Model Fitting

After data exploraty, we would fit multiple statistical models to discuss about the objectives.

- Is there evidence to suggest that the treatments differ in their effects on alcohol dependence, as reflected by the number of drinks consumed in a given 30 day period?

To find the relationship between alcohol consumption in a given 30 day period in different treatment group, generilized linear model (GLM) with Poisson distribution is introducted. GLM is an extension of linear model with 3 key components:

- Random component - probability ditribution of the outcome variable
- Systematic conmponent - combination of linear predictors
- Link function - the link between random component and systematic conmponent

Since we want to find the relationship in a given 30 day period, two models should be built at day 30 and day 60. According to the EDA, gender should clearly different trend of change in

Model I (At day 30, controling for gender and baseline consumption):

$$log(Y_{ij}) = \beta_0 + \beta_1 * I(treatment = 2) + \beta_2 * I(treatment = 3) + \beta_3 * I(gender = female) + \beta_4 * ND0$$

Model II (At day 30, controling for gender, day 30 consumption and baseline consumption):

$$log(Y_{ij}) = \beta_0 + \beta_1 * I(treatment = 2) + \beta_2 * I(treatment = 3) + \beta_3 * I(gender = female) + \beta_4 * ND30 + \beta_5 * ND0$$

If the estimates on treatments differ in two models, then we could conclude treatment affected differently on alcohol dependence in a given 30 day period.

- Is there a difference in the pattern of change in the number of drinks consumed between the various treatment groups over the duration of the study?

Based on the results from `Figure 3` and `Figure 4`, there should be a difference in the pattern of change in the number of drinks condumed between various groups over time. To test this hypothesis, mixed effect Model with random intercept is introducted. Mixed effect model is an extension of GLM, which models the trend of outcome on subject level, taking account the different among each individual by adding random intercept or random slope.

To know the difference in pattern of change, we need to include interaction term (time:treatment) into the model.

Model III (with random intercept):

$$log(Y_{ij}) = \beta_0 + \beta_1 * I(treatment = 1) * I(time = 0) + \beta_2 * I(treatment = 2) * I(time = 0) + \beta_3 * I(treatment = 3) * I(time = 0) + \beta_4$$

If the estimates for all the fixed effects are different in each treatment group, then we could conclude that there's difference in pattern of change on outcome in various treatments.

Conclusion

- Is there evidence to suggest that the treatments differ in their effects on alcohol dependence, as reflected by the number of drinks consumed in a given 30 day period?

The results of Model I is showed in `Table 1`.

In a given 30 day period from baseline to day 30, the effect of alcohol dependence reacts differently on treatment groups.

To be specific, the log rate ratio of an average individual in treatment 2 comparing to treatment 1 is -0.02 controling with gender and baseline consumption, and the difference is not statistical significant. So, treatment 2 has no significant effect on alcohol consumption comparing to treatment 1.

Similarly, the log rate ratio of an average individual in treatment 3 comparing to treatment 1 is -0.40 controling with gender and baseline consumption, also the difference is statistical significant. So, treatment 3 has a significant decrease effect on alcohol consumption comparing to treatment 1.

The reults of Model II is showed in `Table 2`.

In a given 30 day period from day 30 to day 60, the effect of alcohol dependence reacts differently on treatment groups.

To be specific, the log rate ratio of an average individual in treatment 2 comparing to treatment 1 is -0.39 controling with gender, baseline consumption and day 30 consumption, also the difference is statistical significant. So, treatment 2 has a significant decrease effect on alcohol consumption comparing to treatment 1.

Similarly, the log rate ratio of an average individual in treatment 3 comparing to treatment 1 is -0.35 controling with gender, baseline consumption and day 30 consumption, also the difference is statistical significant. So, treatment 3 has a significant decrease effect on alcohol consumption comparing to treatment 1.

- Is there a difference in the pattern of change in the number of drinks consumed between the various treatment groups over the duration of the study?

Discussion

Reference

Appendix

Read Data

```
# Read the txt file into r, the data is wide format
df = read.table("ALCDEP.txt") %>%
  janitor::clean_names() %>%
  mutate(treatment = as.factor(treatment),
         gender = as.factor(gender))

# Reshape the data into long format
df_long = df %>%
  gather(key = "time", value = "consumption", nd0:nd60) %>%
  mutate(time = substr(time, 3, nchar(time)),
         time = as.numeric(time))
```

EDA

```
# Figure 1
# spagatti plot for each individual along time indicating treatment group
p1 =
  ggplot(df_long,aes(x = time, y = consumption, color = treatment, group = sid)) +
  geom_line() +
  ggthemes::theme_few() + ggthemes::scale_color_few() +
  labs(x = "Time points",
```

```
       y = "Count of alcohol consumptions")
grid.arrange(p1, bottom="Figure 1: Spagatti plot among time stratifying treatment group")
```

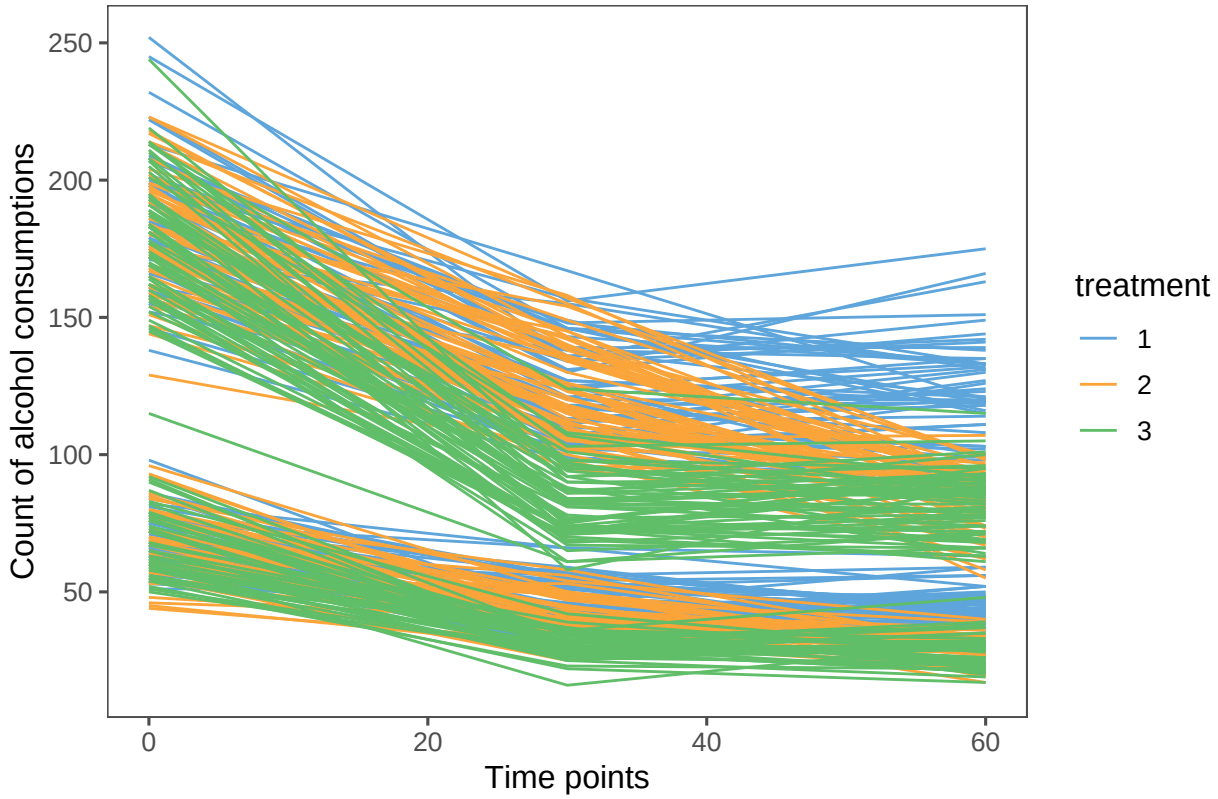

Figure 1: Spagatti plot among time stratifying treatment group

```
# Figure 2
# Distribution between treatment and alcohol consumption indicating time points
p2 =
  ggplot(df_long, aes(x = time, y = consumption, fill = treatment)) +
  geom_boxplot() +
  ggthemes::theme_few() + ggthemes::scale_fill_few()
grid.arrange(p1, bottom="Figure 2: Spagatti plot among time stratifying treatment group")
```

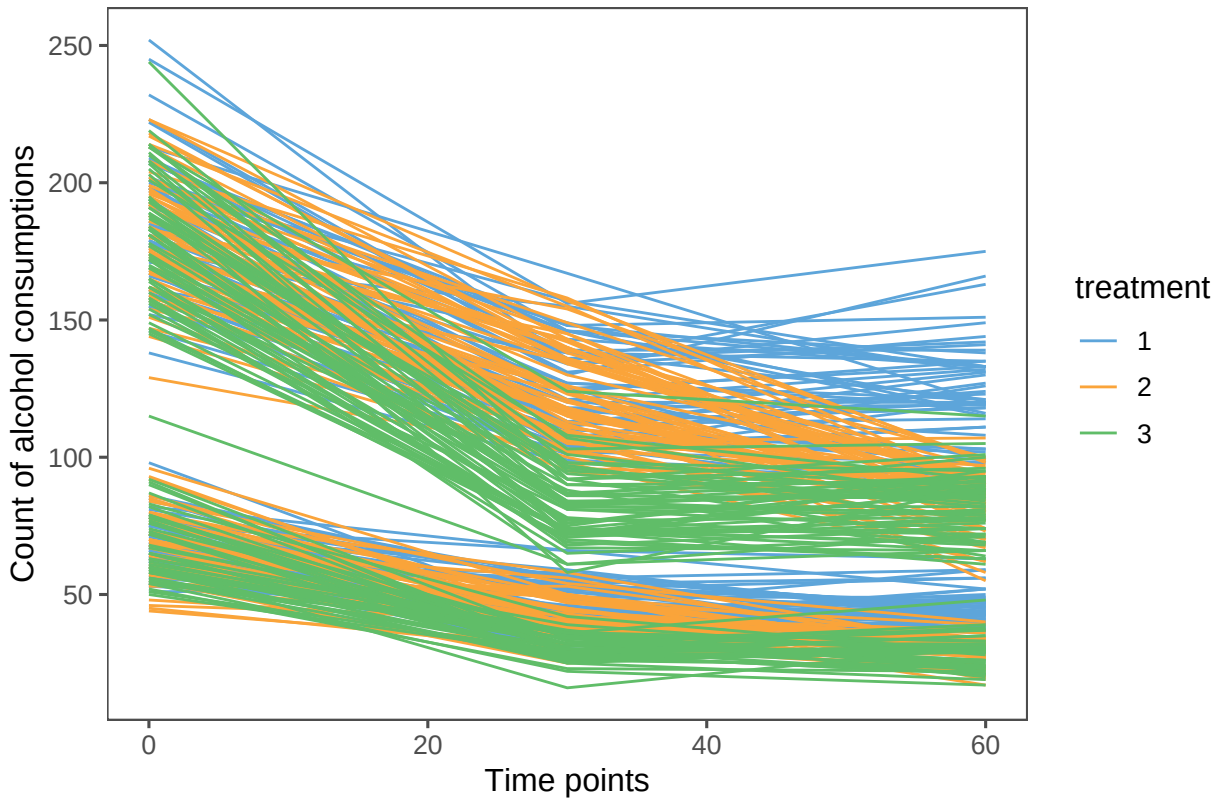

Figure 2: Spagatti plot among time stratifying treatment group

```
# Figure 3
# spagatti plot for each individual along time indicating gender
p3 =
  ggplot(df_long, aes(x = time, y = consumption, color = gender, group = sid)) +
  geom_line() +
  ggthemes::theme_few() + ggthemes::scale_color_few() +
  labs(x = "Time points",
       y = "Count of alcohol consumptions"
       )
grid.arrange(p3, bottom="Figure 3: Spagatti plot among time stratifying treatment group")
```

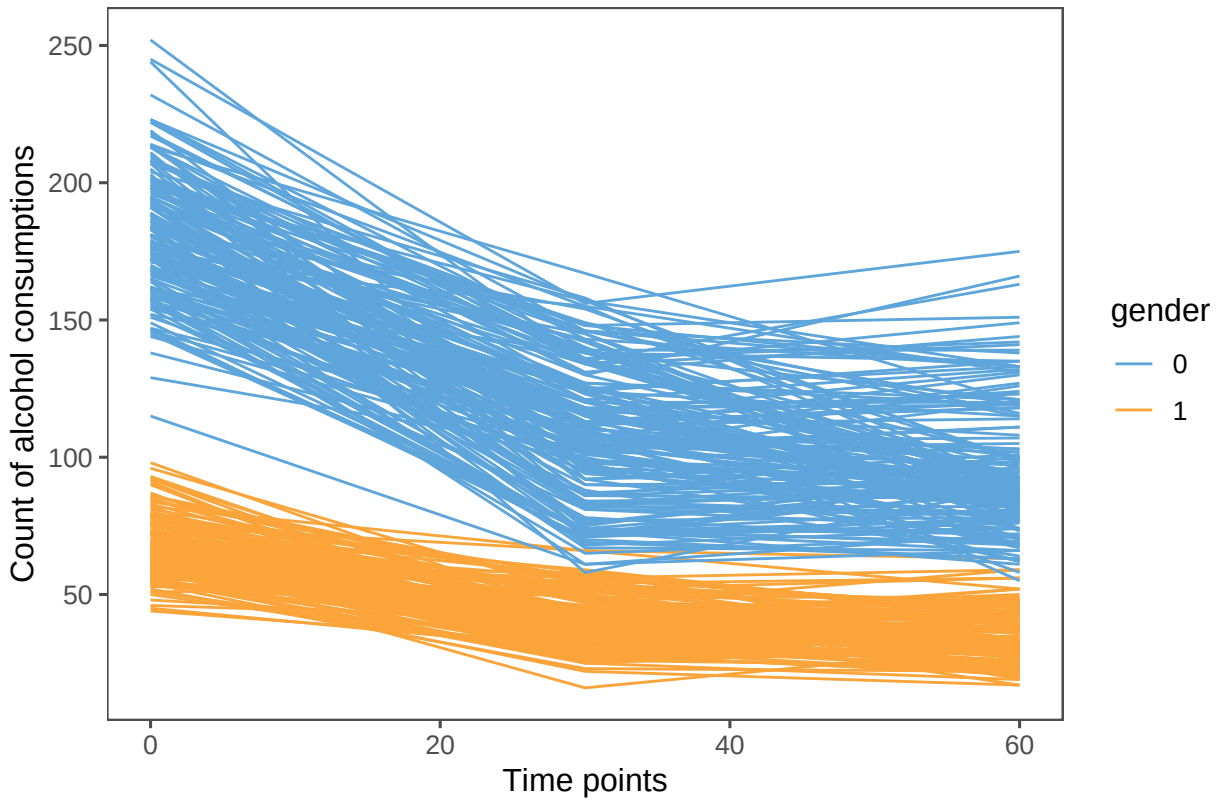

Figure 3: Spagatti plot among time stratifying treatment group

```
# Figure 4
# Distribution between time and alcohol consumption indicating gender
p4 =
  ggplot(df_long, aes(x = time, y = consumption, fill = gender)) +
  geom_boxplot() +
  ggthemes::theme_few() + ggthemes::scale_fill_few()
grid.arrange(p4, bottom="Figure 4: Spagatti plot among time stratifying treatment group")
```

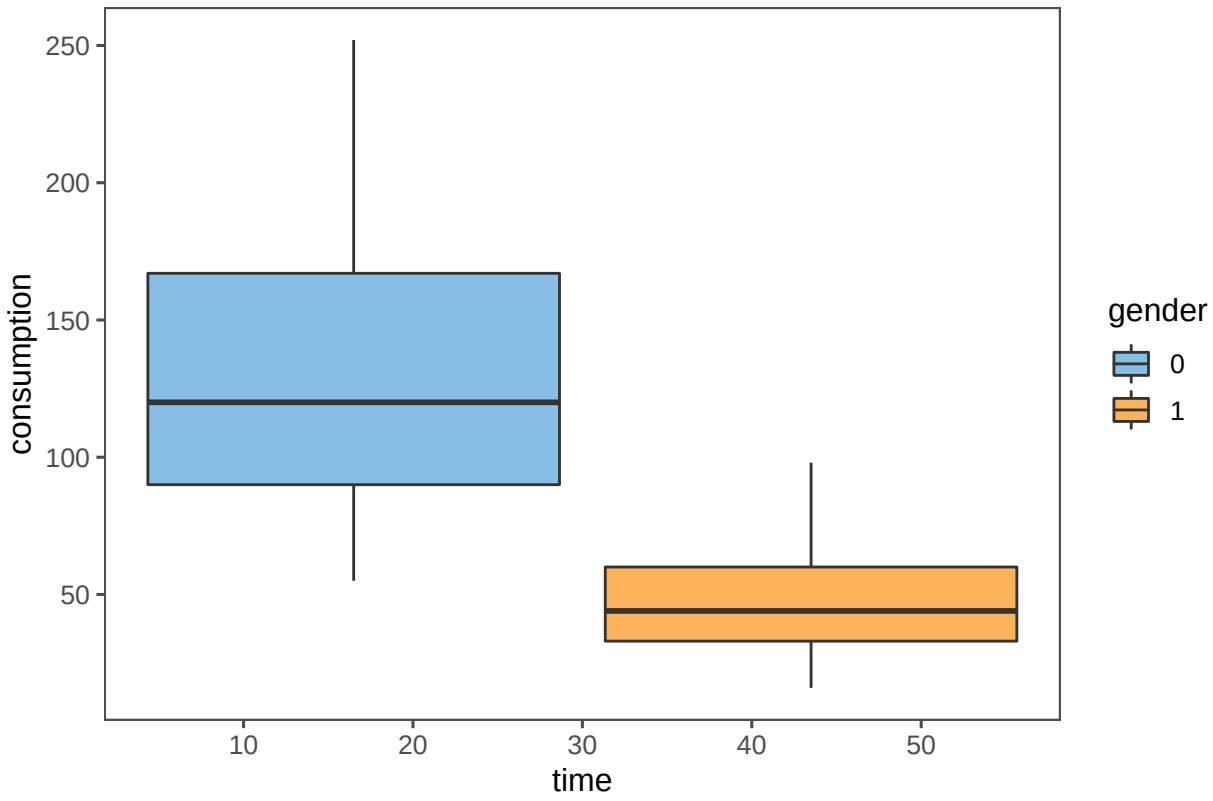

Figure 4: Spagatti plot among time stratifying treatment group

```
# Figure 5
# Ditribution of alcohol consumption
p5 =
  ggplot(df_long, aes(x = consumption, fill = gender)) +
  geom_density()+
  facet_wrap(~gender+~time) +
  ggthemes::theme_few() + ggthemes::scale_fill_few()
grid.arrange(p5, bottom="Figure 5: Spagatti plot among time stratifying treatment group")
```

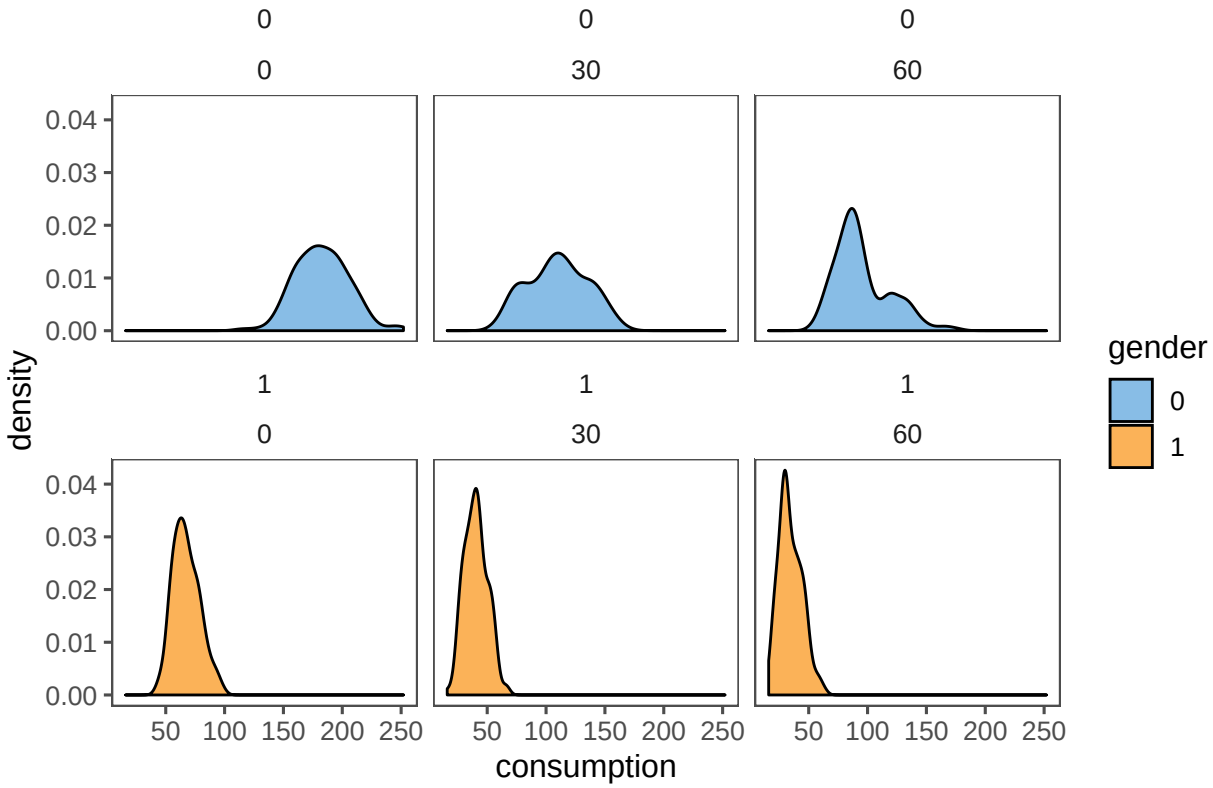

Figure 5: Spagatti plot among time stratifying treatment group

Question 1

Is there evidence to suggest that the treatments differ in their effects on alcohol dependence, as reflected by the number of drinks consumed in a given 30 day period?

```
fit1_1 = glm(nd30 ~ treatment + gender + nd0, family = "poisson", data = df)
fit1_2 = glm(nd60 ~ treatment + gender + nd30 + nd0, family = "poisson", data = df)

summary(fit1_1)$coef %>%
  as_tibble(rownames = 'terms') %>%
  janitor::clean_names() %>%
  rename(p_value = pr_z) %>%
  mutate(sign = if_else(p_value <= 0.05 ,'*',' '),
         ci_low = estimate - 1.96*std_error,
         ci_high = estimate + 1.96*std_error) %>%
  select(terms,estimate, ci_low, ci_high, p_value, sign) %>%
  knitr::kable(digits = 2, caption="table")
```

Table 1: table

terms	estimate	ci_low	ci_high	p_value	sign
(Intercept)	4.10	3.99	4.22	0.00	*
treatment2	-0.02	-0.04	0.01	0.29	
treatment3	-0.40	-0.43	-0.37	0.00	*
gender1	-0.56	-0.64	-0.48	0.00	*
nd0	0.00	0.00	0.00	0.00	*

```r
summary(fit1_2)$coef %>%
    as_tibble(rownames = 'terms') %>%
  janitor::clean_names() %>%
  rename(p_value = pr_z) %>%
  mutate(sign = if_else(p_value <= 0.05 ,'*',' '),
         ci_low = estimate - 1.96*std_error,
         ci_high = estimate + 1.96*std_error) %>%
  select(terms, estimate, ci_low, ci_high, p_value, sign) %>%
  knitr::kable(digits = 2, caption = "table")
```

Table 2: table

terms	estimate	ci_low	ci_high	p_value	sign
(Intercept)	4.03	3.90	4.16	0.00	*
treatment2	-0.39	-0.42	-0.36	0.00	*
treatment3	-0.35	-0.40	-0.30	0.00	*
gender1	-0.56	-0.64	-0.47	0.00	*
nd30	0.00	0.00	0.00	0.02	*
nd0	0.00	0.00	0.00	0.00	*

Question 2

Is there a difference in the pattern of change in the number of drinks consumed between the various treatment groups over the duration of the study?

```r
fit2 = glmer(consumption ~ treatment:time + (1 | sid), family = poisson,
             data = df_long)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00200619
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```r
summary(fit2)$coef %>%
  as_tibble(rownames = 'terms') %>%
  janitor::clean_names() %>%
  rename(p_value = pr_z) %>%
  mutate(sign = if_else(p_value <= 0.05 ,'*',' '),
         ci_low = estimate - 1.96*std_error,
         ci_high = estimate + 1.96*std_error) %>%
  select(terms,estimate, ci_low, ci_high, p_value, sign) %>%
  knitr::kable(digits = 2)
```

terms	estimate	ci_low	ci_high	p_value	sign
(Intercept)	4.71	4.65	4.76	0	*
treatment1:time	-0.01	-0.01	-0.01	0	*
treatment2:time	-0.01	-0.01	-0.01	0	*
treatment3:time	-0.01	-0.02	-0.01	0	*

Question 3

Alcohol-use disorders are among the most disabling disease categories for the global burden of disease especially for men. Is there evidence to suggest that males tend to have a higher alcohol dependence than females?

```
glmer(consumption ~ gender + time + (1|sid), df_long, family = poisson)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00675845
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: consumption ~ gender + time + (1 | sid)
##    Data: df_long
##       AIC       BIC    logLik  deviance  df.resid
##  8746.675  8766.067 -4369.337  8738.675       938
## Random effects:
##  Groups Name        Std.Dev.
##  sid    (Intercept) 0.1348
## Number of obs: 942, groups:  sid, 314
## Fixed Effects:
## (Intercept)      gender1         time
##      5.1606      -1.0025      -0.0116
## convergence code 0; 2 optimizer warnings; 0 lme4 warnings
```

Question 4

Do men and women respond differently to treatment?

```
glmer(consumption ~ gender:treatment + time + (1|sid), df_long, family = poisson)
```

```
## fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.0102199
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: consumption ~ gender:treatment + time + (1 | sid)
##    Data: df_long
##       AIC       BIC    logLik  deviance  df.resid
##  8606.333  8645.117 -4295.166  8590.333       934
## Random effects:
##  Groups Name        Std.Dev.
##  sid    (Intercept) 0.09886
## Number of obs: 942, groups:  sid, 314
```

```
## Fixed Effects:
##        (Intercept)                time  gender0:treatment1
##             4.0428             -0.0116              1.2364
## gender1:treatment1  gender0:treatment2  gender1:treatment2
##             0.2206              1.1234              0.1015
## gender0:treatment3
##             1.0038
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
## convergence code 0; 2 optimizer warnings; 0 lme4 warnings
```

Question 5

Is there any evidence to suggest that the treatments differ in their effects on subjects with regard to relapsing into alcohol dependence?

```
glmer(relapse ~ treatment + gender + time + consumption + (1|sid),
      df_long, family = poisson)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00303067
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: relapse ~ treatment + gender + time + consumption + (1 | sid)
##    Data: df_long
##       AIC       BIC    logLik  deviance  df.resid
## 1410.2551 1444.1911 -698.1275 1396.2551       935
## Random effects:
##  Groups Name        Std.Dev.
##  sid    (Intercept) 0.3821
## Number of obs: 942, groups:  sid, 314
## Fixed Effects:
## (Intercept)   treatment2   treatment3      gender1         time
##  -0.2320330   -0.6003733   -1.7988653   -0.0092026   -0.0001595
## consumption
##  -0.0001726
## convergence code 0; 2 optimizer warnings; 0 lme4 warnings
```

Question 6

Even in the case that the treatments might differ in their pattern of change or on how subjects relapse into alcohol dependence, is there any evidence to suggest that any of the treatments might be beneficial once the treatment has stopped.