STAT 601 Final Project: Schizophrenia Data Set

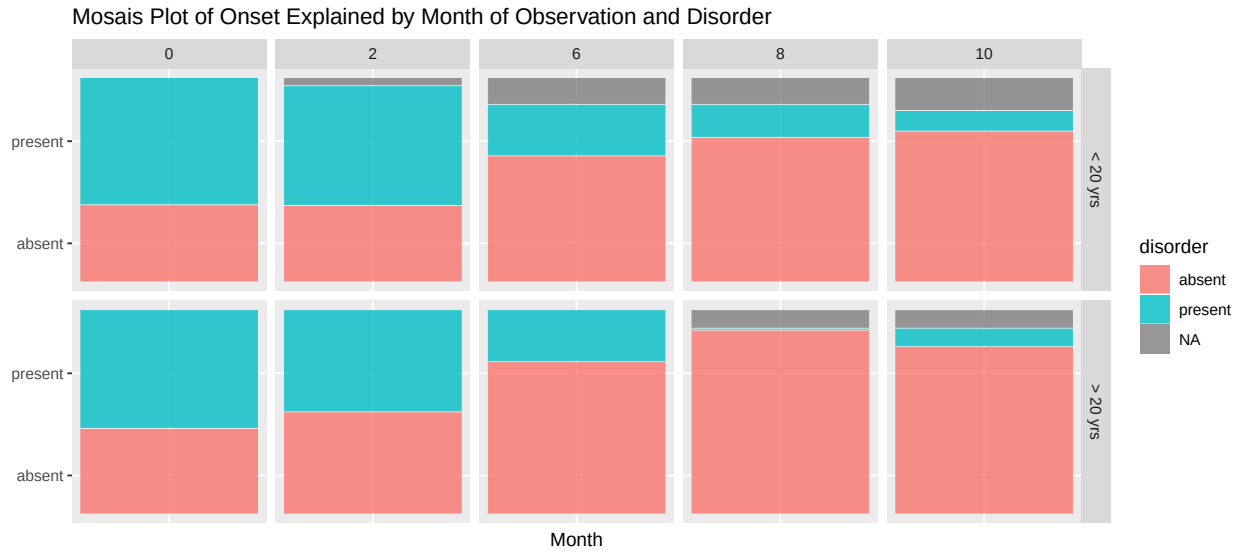
Joseph McDonald

Introduction

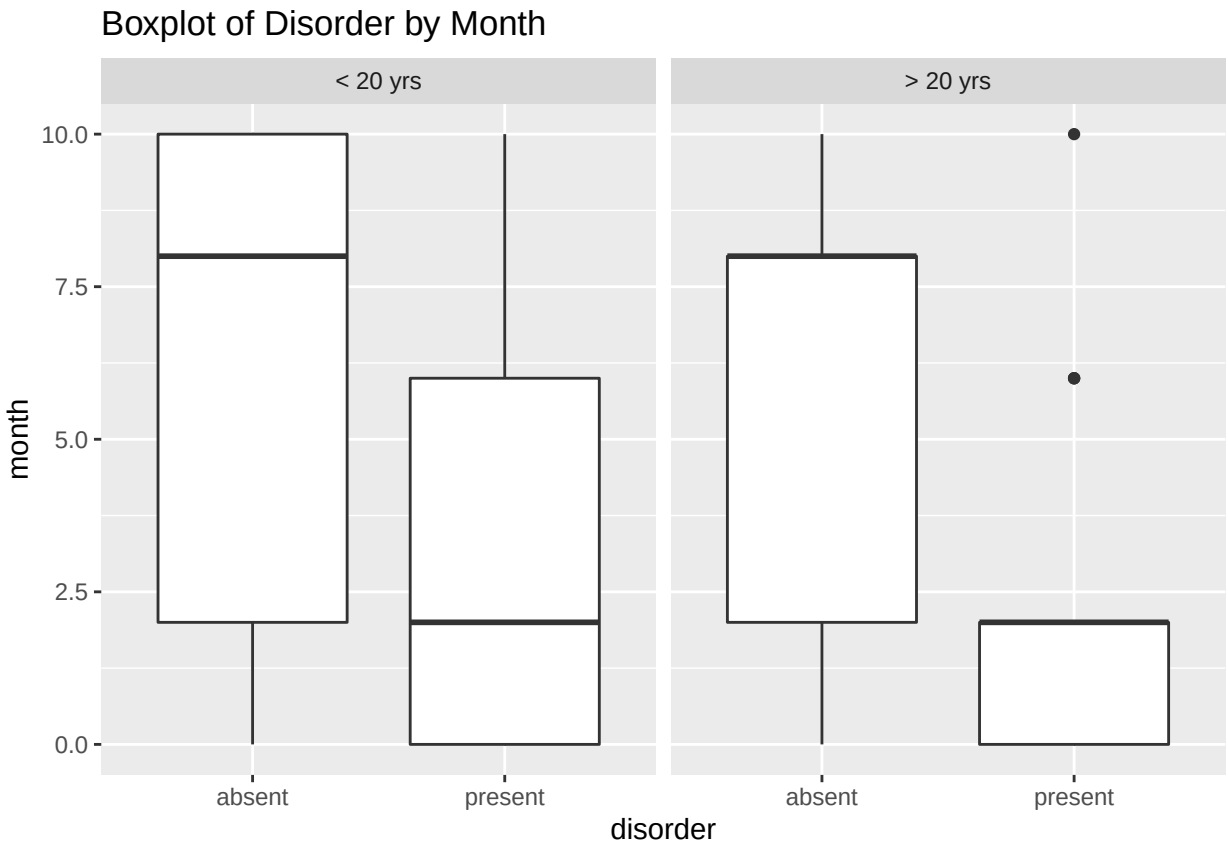
In this problem, we are trying to determine whether the course of schizophrenia differs over time depending on whether a person has early or late onset schizophrenia. We will attempt to do this by determining whether the binary onset variable, less than 20 years or greater than 20 years, is alone a useful predictor variable for whether the disorder will be present or absent at given intervals following hospitalization. If it is a good predictor, than the implication would be that there are two distinct patterns of the illness differing based upon the onset.

Exploration

A brief examination of a summary of the data reveals that there are a few missing values of the disorder variable. This is likely due to people who dropped out of the study over time. For now, we will leave these observations in the dataset and continue our explorations. The mosaic plot shows the relationship between observations of the disorder and the month in which they were observed, considering early and late onset patients seperately. We see that with both groups there is a distinct trend of declining presence of thought disorder as the month of observation increases but perhaps with differing rate. We also note, as should be expected, that NA values increase as time progresses. Intuitively we assume that this is due to the fact that people are more likely to drop out of the study as time goes on for one reason or another. Because there are relatively few dropouts, for the purpose of this analysis we will remove the observations with NA values.



A box plot comparison of the onset variable gives a further sense of the distribution of the data. We notice right away that the median months for presence or absense for both early and late onset are the same. Additionally, the overall distribution of both appear to be fairly similar even though there appears to be some distinction between the levels of overlapping quartiles. This is would lead us to believe from a brief overview that there is not a substantial difference in the course of the disease based upon whether it is early or late onset, but we will need further analysis to make a final conclusion.



Model Training and Selection

Because of the repeated measures in this data set, we assume that there is some type of correlation between the measurements over time. Once a model of best fit is found, we can determine whether the onset of schizophrenia is useful in predicting the course of the disease over time. This would cause problems for generlized linear models because it violates the assumption of the independence of individual measurements. We will train several generalized estimating equations (gee) and linear mixed models and determine the model of best fit for this data since in various ways these models account for the correlation structure between the repeated measures and link the explanatory variables to the repeated measures, similary to multiple regression.

We start by training three different gee models with indepedent, exchangeable, and, unstructured correlation structures respectivley. We find that the three models are very similary in their fit. QIC indidcates that the unstructured model is the best fit of the three, but the difference between the three is small enough that we will examine the coefficient structures of each of the models further.

Table 1: Comparison of GEE Models

Correlation.Model	QIC
Independent	269.6359
Exchangeable	269.5094
Unstructured	269.4647

Examining the coeffiecient summaries of the three models, we note that the onset variable has z scores that will not produce significant p-values for all three models. However, we also observe that there are relatively

large differences between the naive and robust z scores. This is concerning because it could indicate that our correlation structures are not actually representative of the true correlation structure of the data, one of the key assumptions of the gee model. Because of these differencesit is difficult to consider any of these gee models a good fit for this data.

We will also consider two linear mixed effect models, one with random intercept and one with random intercept and random slope. In these two models, we are able to make an assumption of independence of the repeated measures conditional on the random effects chosen. We train both models and a quick AIC comparison indicates that the random intercept random slope model is a much better fit of the data than the random intercept model. We confirm that there is a significant difference between the two models with an ANOVA test which returns a p-value of 1.362e-09. Therefore we conclude that the random intercept random slope model is the better fit between the two lmer models.

Since none of the gee models are a good fit, we examine the the better of the two lmer models a bit more closely to make sure that we do not have a case where none of the models are a good fit. The confusion matrix reveals decent results and so we conclude that the lmer model with random slope and random intercept is our model of best fit.

Table 2: Confusion Matrix for random intercept and slope model

	absent	present
absent	129	15
present	3	57

Conclusion and Discussion

With our model of best fit we conduct a coefficient test and find that the onset variable is, in fact, not significant as a predictor of the behavior of thought disorder over time with a p-value of .289.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lmer(formula = nstat ~ onset + (month | subject), data = schizo,
##     REML = FALSE, na.action = na.omit)
##
## Linear Hypotheses:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept) == 0    0.33187    0.04742   6.998  2.6e-12 ***
## onset> 20 yrs == 0 -0.09404    0.08865  -1.061    0.289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Univariate p values reported)
```

It is also interesting to note that the onset variable was not a significant predictor in any of the models which we discarded in model selection either. Because it has been shown that the onset variable is not a significant predictor in the model of best fit, or any model tested for that matter, we are unable to conclude that the course of schizophrenia differs between patients with early or late onset. One assumption that could have impacted this conclusion include the assumption that the missing values from drop out do not provide relevant information. We cannot conclude definitively that there is no difference, but only that we were not able to find one with the given data set and assumptions.