Homework #10

Justin Robinette

No collaborators for any problem

Problem #1, Part A: Consider the **respiratory** data from the **HSAUR3** package. Investigate the use of other correlation structures than the independence and exchangeable structures used in the text for the respiratory data.

Results: For this exercise, I investigated the use of the *'AR-M'*, and *'unstructured'* correlation structures within the gee model. I chose these two models because, in the next exercise, I will explore QIC through the use of **geeglm** which works with these two correlation structures.

Figure 1.1a and **Figure 1.1b** show the coefficient values and P-Values of the *Auto Regressive* correlation structure. **Figure 1.1b** shows that *treatment*, *centre*, and *baseline* are statistically significant predictor variables. **Figure 1.1a** shows an increase in SE values when going from naive to robust.

Figure 1.2a and **Figure 1.2b** are constructed the same as the prior set, this time examining the *Unstructured* correlation structure. Here, each pair of observations is allowed to have a different correlation. **Figure 1.2a** shows a better (smaller) change when going from glm to robust. Another major change in this structure is that *centre* is no longer a statistically significant predictor variable at alpha = 0.05.

Figure 1.3 and **Figure 1.4** summarize the p-value tables, for both *naive (glm)* and *robust sandwich*, of each predictor based on the correlation structure used in the model. An alpha value of 0.05 is indicated on each plot for convenience. Bar plots below the line indicate the variable is a significant predictor of *respiratory status* in the model. As we can see in **Figure 1.4**, in the *Unstructured* model *centre*'s p-value is greater than the alpha of 0.05 indicating that it is not a statistically significant predictor variable at this level.

Figure 1.1a: Auto Regressive Structure Coefficients & P-Values

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	**-0.9629448**	**0.4455161**	**-2.1614142**	**0.4611607**	**-2.0880894**
centre2	**0.7427015**	**0.3146448**	**2.3604438**	**0.3562300**	**2.0848932**
trttrt	**1.2472824**	**0.3112665**	**4.0071210**	**0.3518974**	**3.5444492**
gendermale	**0.1132323**	**0.3883671**	**0.2915599**	**0.4494506**	**0.2519348**
baselinegood	**1.9113953**	**0.3167921**	**6.0335952**	**0.3501873**	**5.4582088**
age	**-0.0169164**	**0.0116652**	**-1.4501525**	**0.0129273**	**-1.3085816**

Figure 1.1b: Auto Regressive Structure P-Values

rn	Naive P-Value	Robust P-Value
(Intercept)	**0.031**	**0.037**
centre2	**0.018**	**0.037**
trttrt	**0.000**	**0.000**
gendermale	**0.771**	**0.801**
baselinegood	**0.000**	**0.000**
age	**0.147**	**0.191**

Figure 1.2a: Unstructured Coefficients

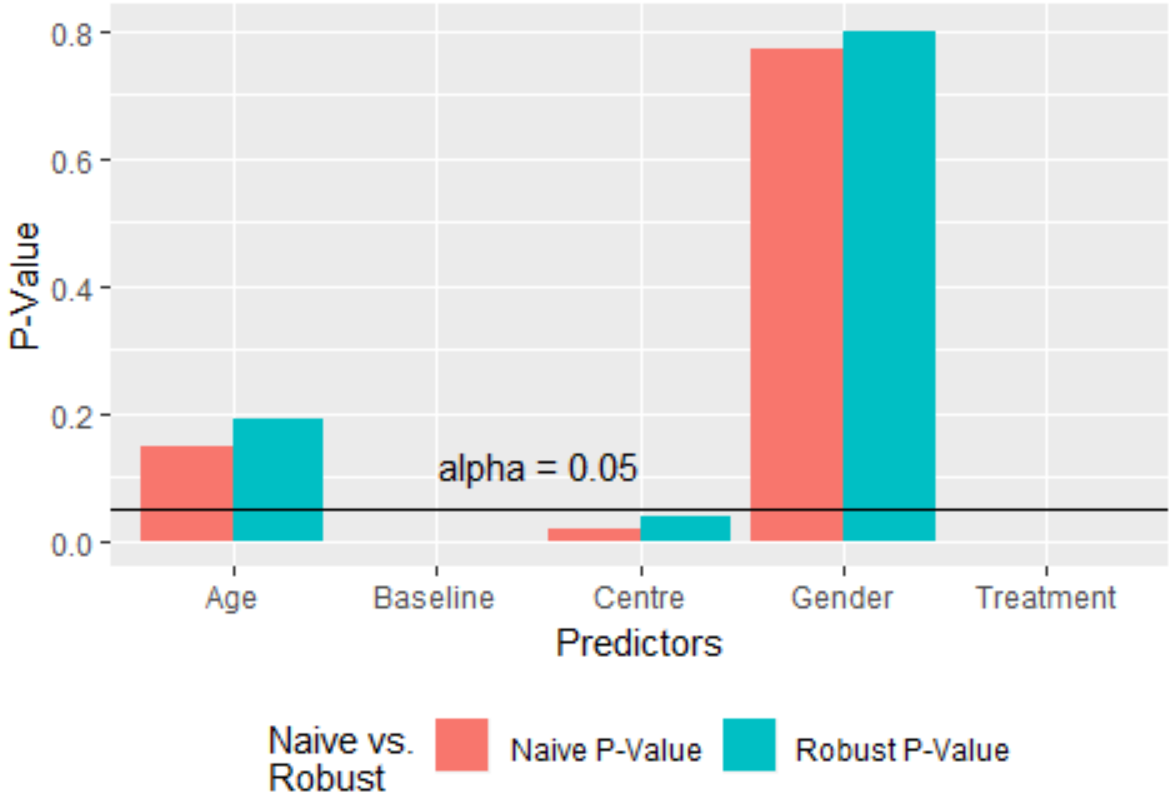
	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	**-0.9312798**	**0.4791852**	**-1.9434655**	**0.4612499**	**-2.0190352**
centre2	**0.6727947**	**0.3390779**	**1.9841895**	**0.3548202**	**1.8961568**
trttrt	**1.2789154**	**0.3354409**	**3.8126404**	**0.3494500**	**3.6597956**
gendermale	**0.0946735**	**0.4172964**	**0.2268736**	**0.4436295**	**0.2134068**
baselinegood	**1.9346252**	**0.3428184**	**5.6432949**	**0.3480468**	**5.5585200**
age	**-0.0168892**	**0.0125574**	**-1.3449620**	**0.0129054**	**-1.3086948**

Figure 1.2b: Unstructured P-Values

rn	Naive P-Value	Robust P-Value
(Intercept)	**0.052**	**0.043**
centre2	**0.047**	**0.058**
trttrt	**0.000**	**0.000**
gendermale	**0.821**	**0.831**
baselinegood	**0.000**	**0.000**
age	**0.179**	**0.191**

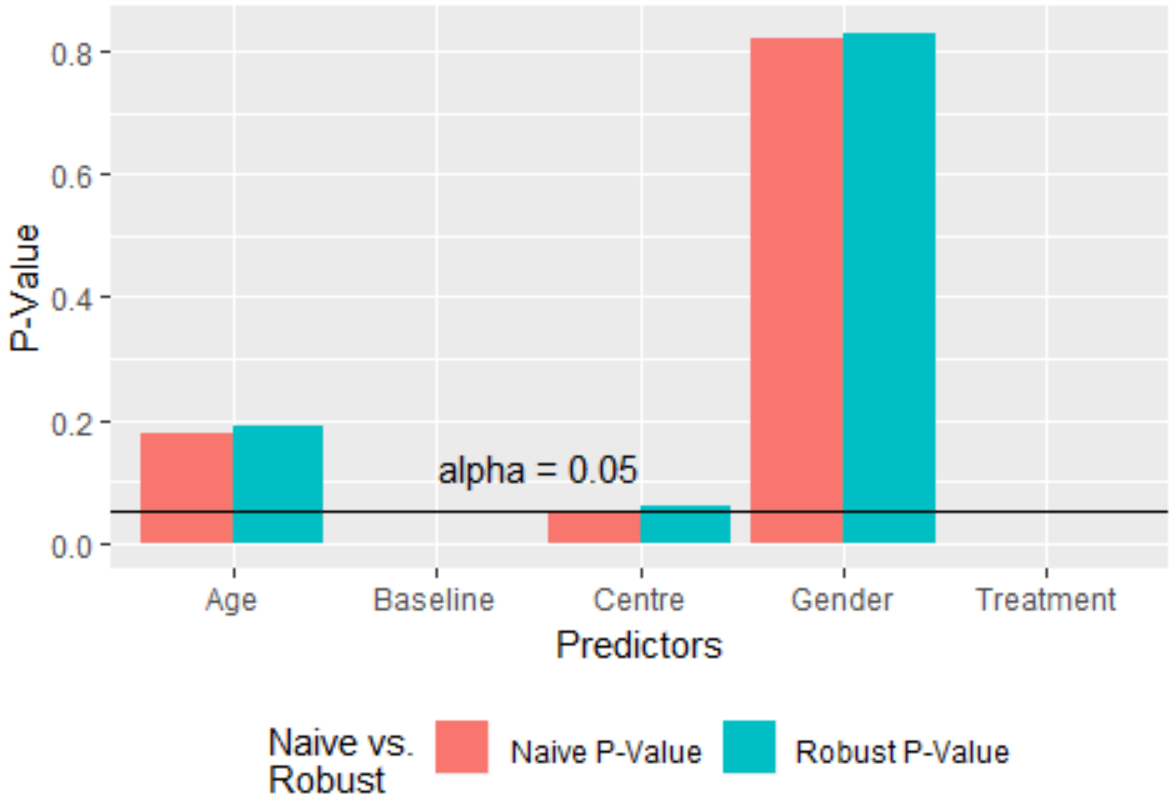
Autoregressive P-Values

Figure 1.3



Unstructured P-Values

Figure 1.4



Problem #1, Part B: Which model is the best? Compare the following models: - independent - exchangable - AR-m - unstructured

Justify your answer. Use QIC (in **MESS**), misclassification rate, comparison of naive vs. robust Z-score.

Results: For this exercise, I compared the QIC, misclassification rate, naive vs. robust Z-score by variable by model and naive vs. robust Z-score total by model.

First, I fit geeglm models that match the gee models from the prior exercise for *independent*, *exchangeable*, *autoregressive* and *unstructured* correlation structured models. **Figure 1.5** shows the QIC scores of each model. The QIC score (Quasilikelihood under the Independence model Criterion) is similar to the AIC score we've used many times this semester. The QIC can be used to compare models with the lower score being the superior model. As we see, the the *unstructured* model is the best, according to QIC. The *independent* model has the worst QIC score among the 4 models. Because 3 of the scores are quite close, I decided to further compare the 4 models to find the best model.

Figure 1.6 compares the misclassification rate which, oddly enough, is identical among the 4 models.

Figure 1.7 shows the difference between the naive and robust z-scores for each variable by model. After comparing these scores, we can see that the *Unstructured* and *Exchangeable* models again appear to be somewhat close using this metric.

Lastly, I used **Figure 1.8** to plot the total difference between the naive and robust Z-scores by model. Here we can see that the *Unstructured* model has a slightly lower total than the *Exchangeable* model. Therefore, I've determined that the best *Generalized Estimation Equation* model is the one that uses *Unstructured* as it's correlation structure.

Figure 1.5: QIC Comparison of Models

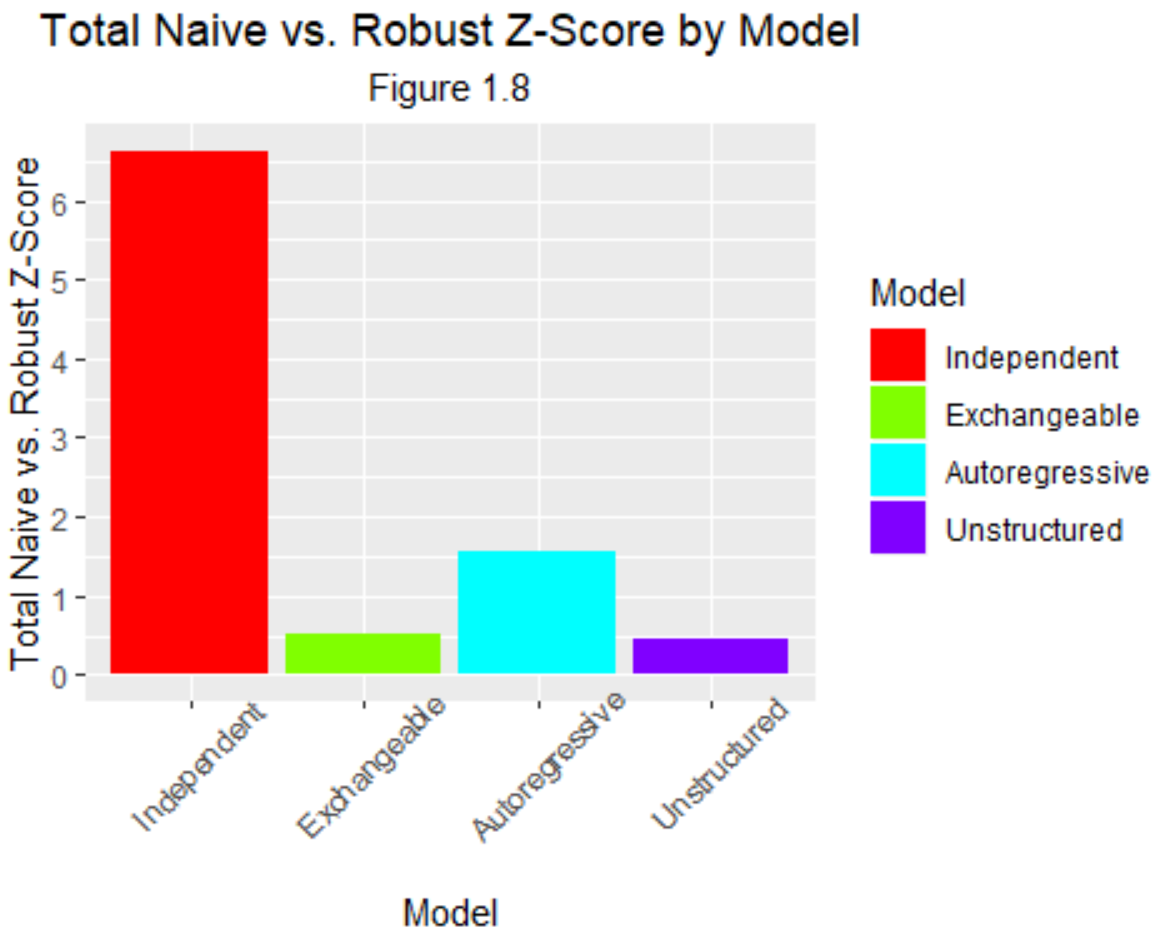
	QIC
IndependentModel_QIC	**508.5300**
ExchangeableModel_QIC	**495.7950**
AutoregressiveModel_QIC	**496.8349**
UnstructuredModel_QIC	**495.7418**

Figure 1.6: Misclassification Rate by Model

Independent Model	Exchangeable Model	Autoregressive Model	Unstructured Model
0.259009	**0.259009**	**0.259009**	**0.259009**

Figure 1.7: Naive vs. Robust Z-Score Difference

	Independent Model	Exchangeable Model	Autoregressive Model	Unstructured Model
(Intercept)	**0.7104613**	**0.0741245**	**0.0733248**	**0.0755697**
centre2	**0.9212111**	**0.0961787**	**0.2755506**	**0.0880327**
trttrt	**1.7817907**	**0.1673937**	**0.4626718**	**0.1528448**
gendermale	**0.1356169**	**0.0165246**	**0.0396250**	**0.0134669**
baselinegood	**2.4234212**	**0.1279470**	**0.5753865**	**0.0847749**
age	**0.6523892**	**0.0492845**	**0.1415709**	**0.0362672**



Problem #2, Part A: The data set **schizophrenia2** from **HSAUR3** package was collected in a follow-up study of women patients with schizophrenia (Davis, 2002). The binary response recorded at 0, 2, 6, 8, and 10 months after hospitalization was "thought disorder" (absent or present). The single covariate is the factor indicating whether a patient had suffered early or late onset of her condition (age of onset less than 20 years of age or age of onset 20 years or above). The question of interest is whether the course of the illness differs between patients with early and late onset schizophrenia. Investigate the question using plots and summary statistics.

Results: Figure 2.1 and **Figure 2.2** show the frequency of 'Thought Disorder' classification by month for Early Onset patients (less than 20 years old at diagnosis) and Late Onset patients (over 20 years old at diagnosis). I've added a category to these plots to show patients who dropped out of the study.

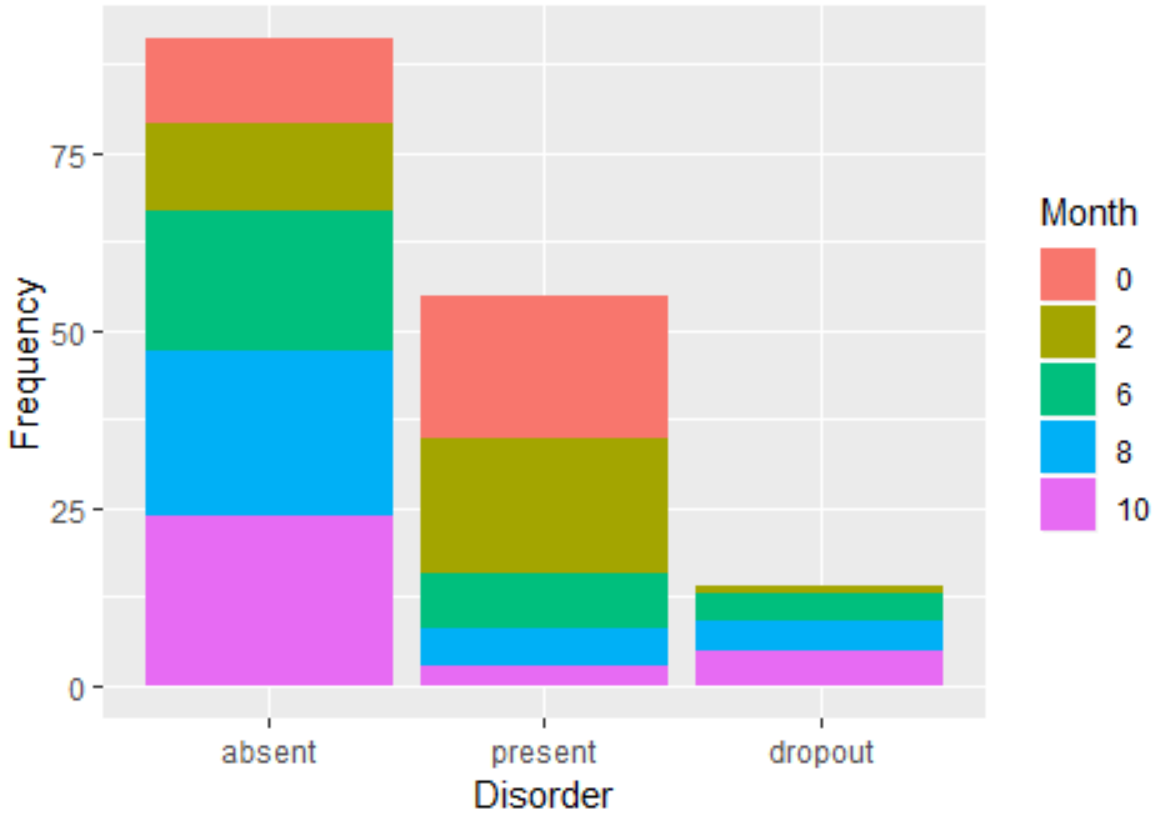
The question of interest is whether the course of the illness was different depending on early or late onset of schizophrenia. These two plots, in my opinion, do not show a distinct difference. We see that as more time passes since hospitalization, the 'present' classification gets smaller and the 'absent' classification gets larger. One noticeable difference is that, with early onset patients, there were dropouts earlier in the study. With late onset patients, there weren't NA values until month 8.

Figure 2.3 shows the frequency of classification by onset status. Again, when looking at the entirety of the data, we don't see a markedly different pattern between early and late onset patients. **Figure 2.4** looks for a difference in patient behavior, based on onset, by examining proportions. Here we see a bigger gap in the proportion of present and dropout classifications among late onset patients than early onset patients. **Figure 2.5** summarizes these plots.

Comparable base R plots are shown.

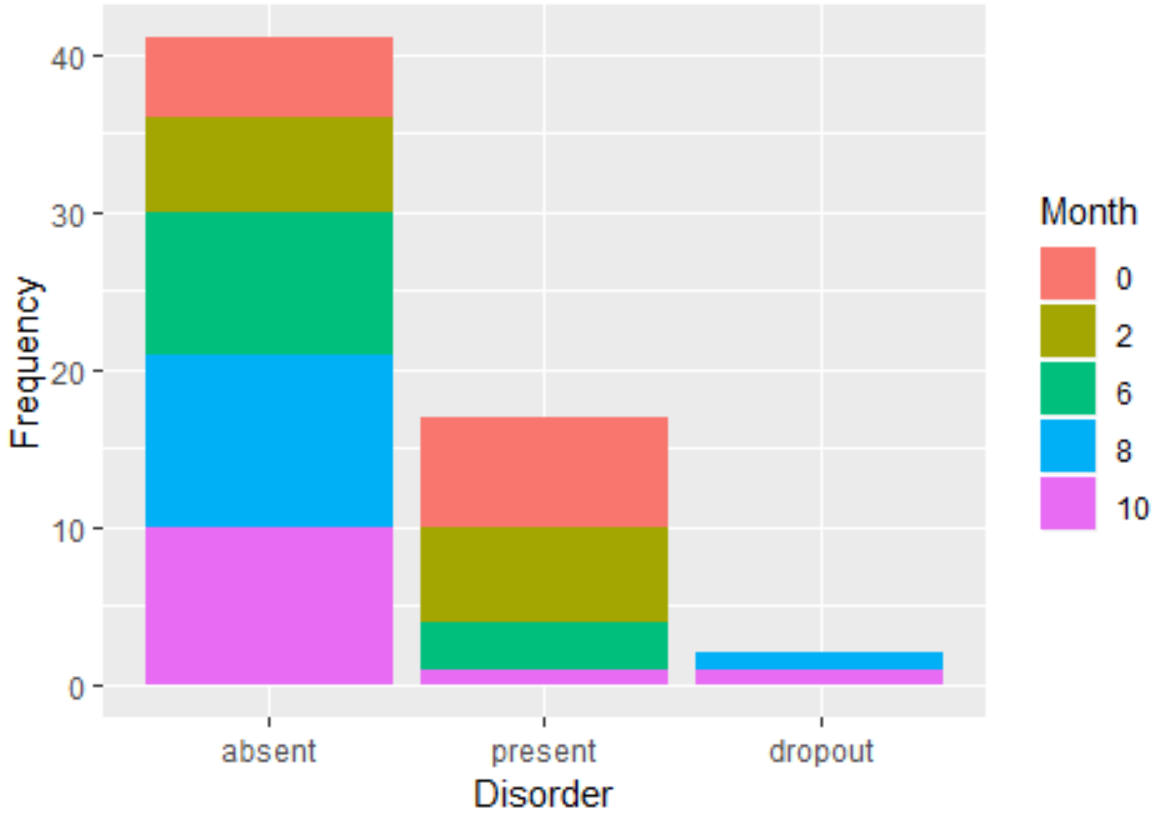
Early Onset Disorder Count by Month

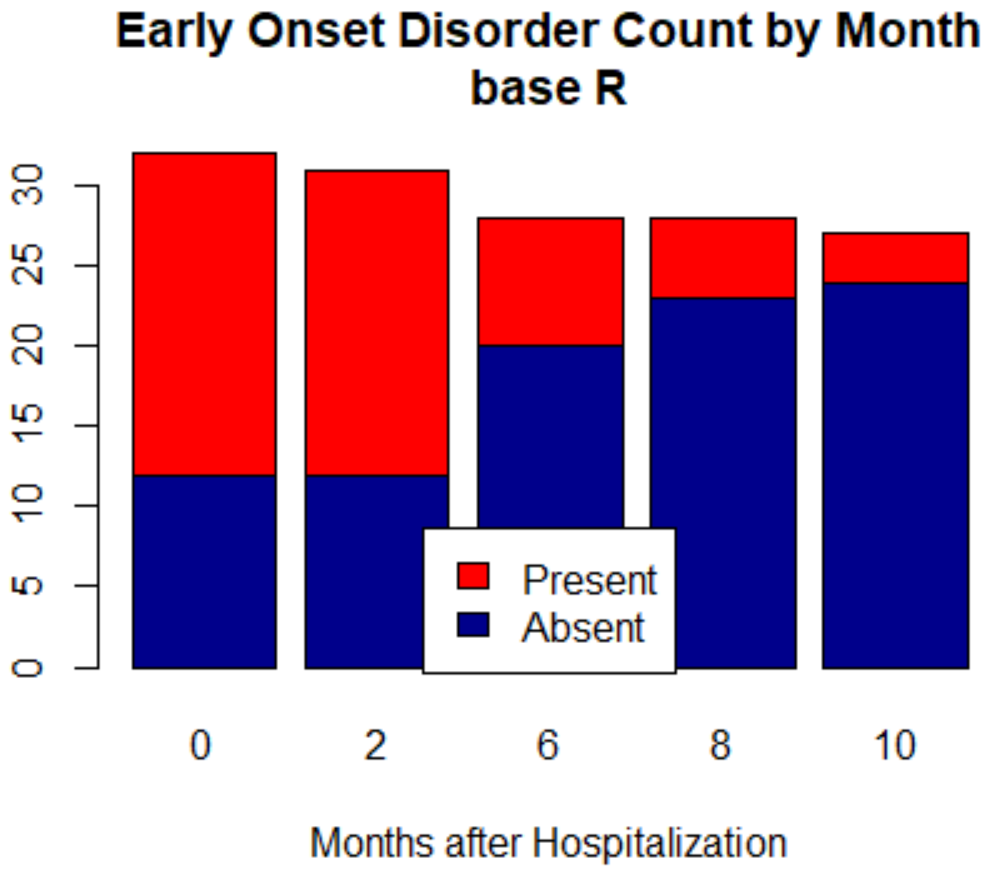
Figure 2.1

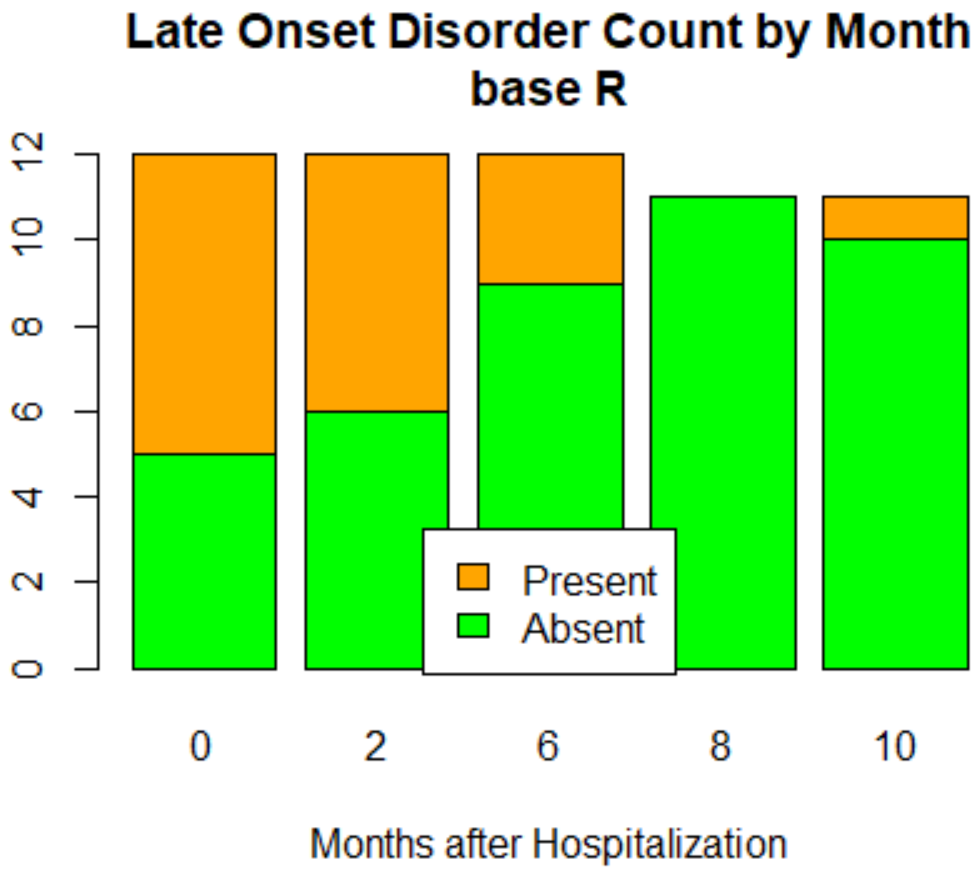


Late Onset Disorder Count by Month

Figure 2.2

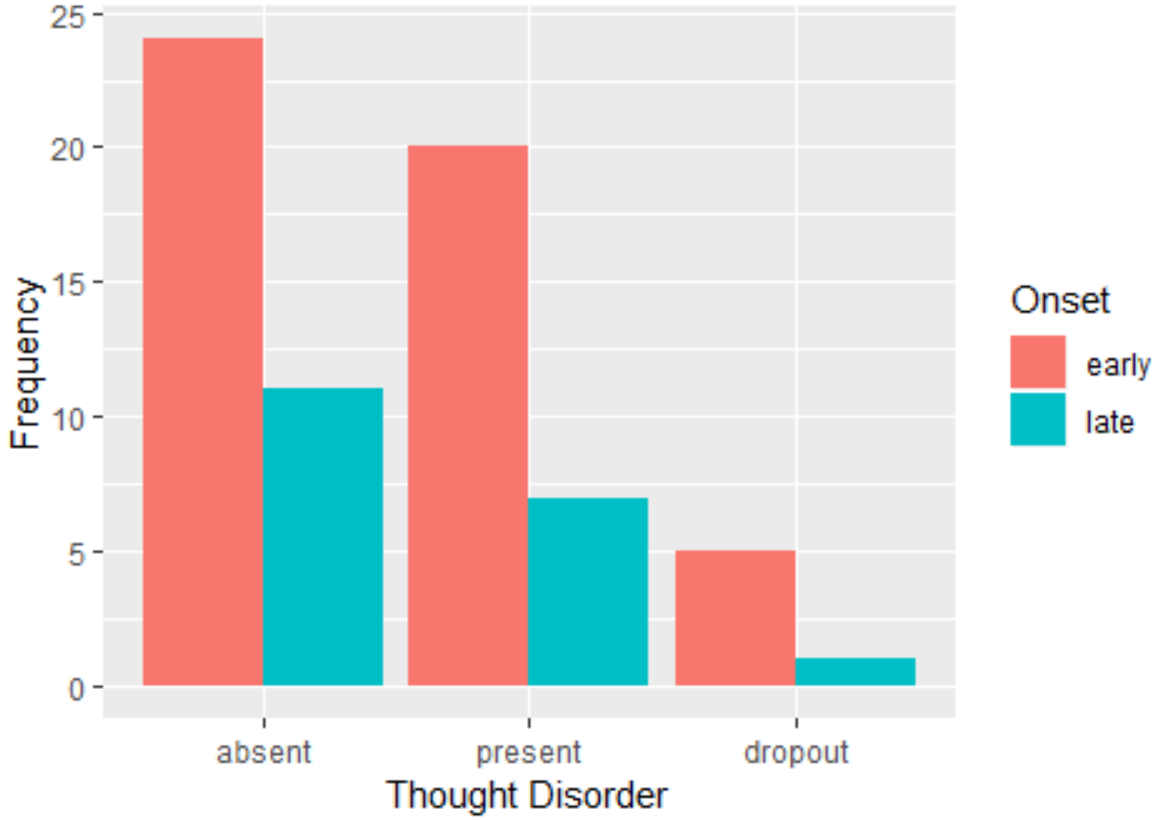



Early Onset Disorder Count by Month
base R
0
5
10
15
20
25
30
Present
Absent
0
2
6
8
10
Months after Hospitalization


Late Onset Disorder Count by Month
base R
0
2
4
6
8
10
12
Present
Absent
0
2
6
8
10
Months after Hospitalization

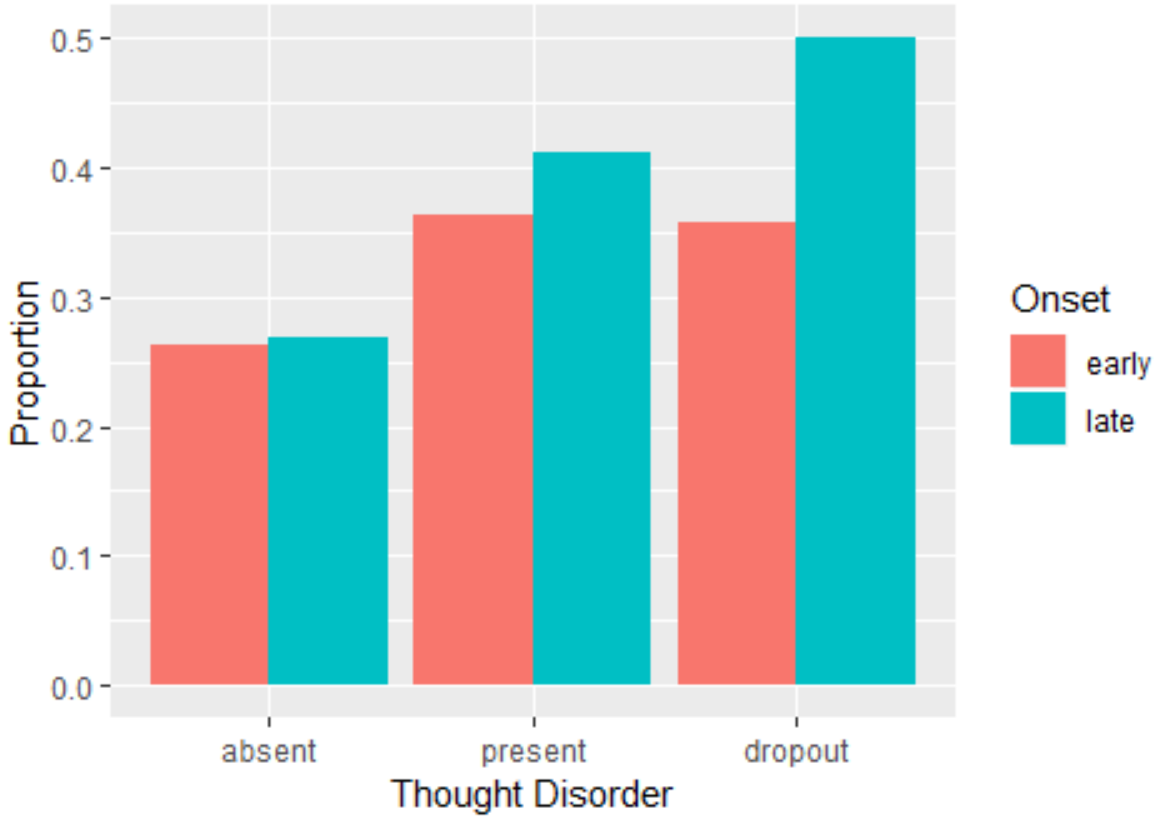
Thought Disorder Count by Onset

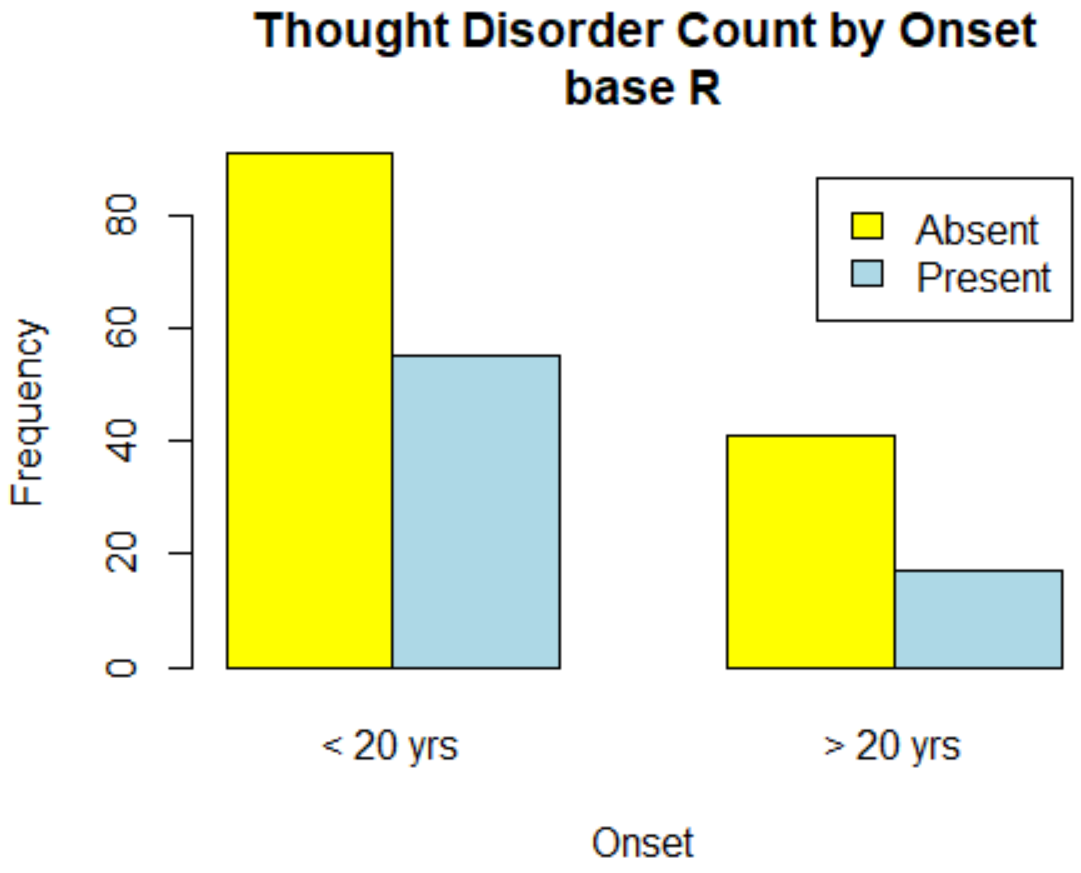
Figure 2.3



Thought Disorder Proportion by Onset

Figure 2.4




Thought Disorder Count by Onset
base R
Frequency
0
20
40
60
80
Absent
Present
< 20 yrs
> 20 yrs
Onset

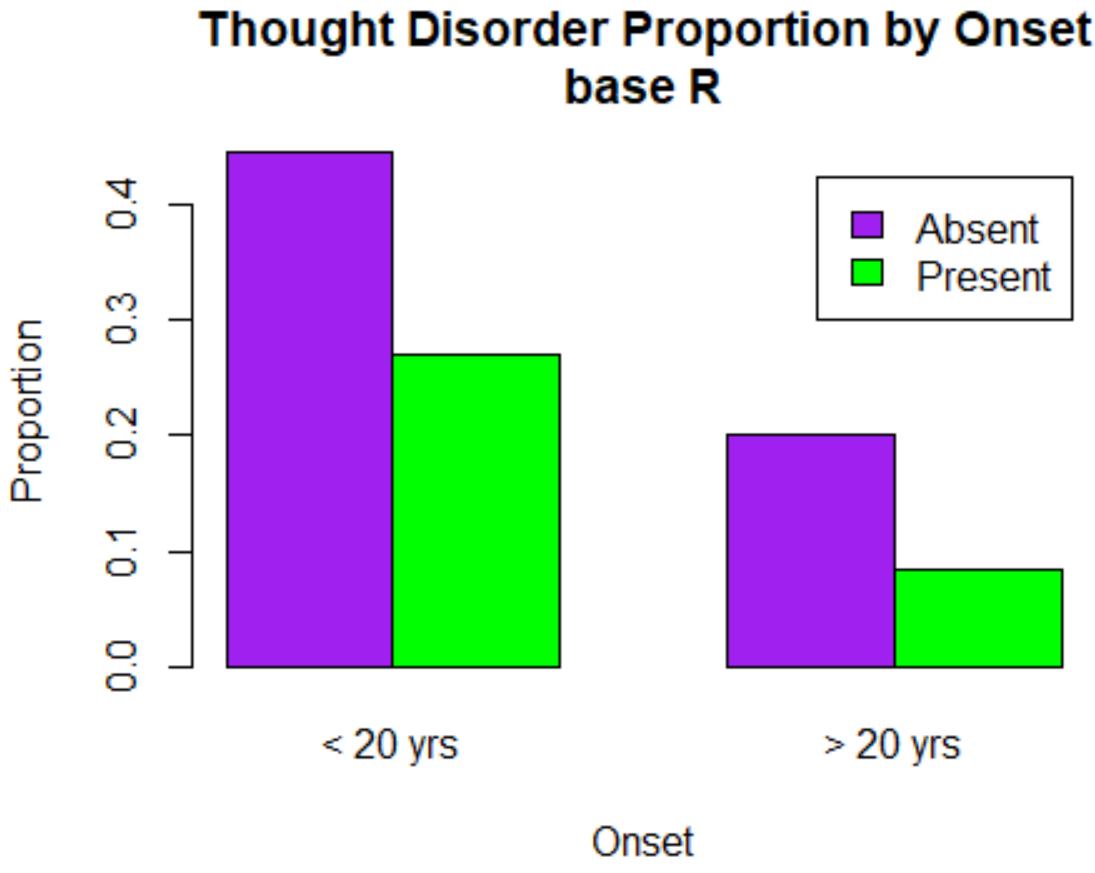

Thought Disorder Proportion by Onset
base R
Proportion
0.0
0.1
0.2
0.3
0.4
Absent
Present
< 20 yrs
> 20 yrs
Onset

Figure 2.5: Frequency and Percentage of Obs by Disorder & Month

Disorder	Month of Eval	Number of Observations	% of Obs by Disorder
absent	0	12	13.186813
absent	2	12	13.186813
absent	6	20	21.978022
absent	8	23	25.274725
absent	10	24	26.373626
present	0	20	36.363636
present	2	19	34.545454
present	6	8	14.545454
present	8	5	9.090909
present	10	3	5.454546
dropout	2	1	7.142857
dropout	6	4	28.571429
dropout	8	4	28.571429
dropout	10	5	35.714286

Problem #2, Part B: Investigate the question using the GEE approach.

Results: I used the GEE approach with correlation structures of 'independence', 'exchangeable', 'unstructured', and 'ar1'. I fit the models using the *geeglm* function. **Figure 2.6** shows the QIC scores by 'corstr'. Here we don't see a big different between models so all will be used going forward.

Because the question of interest is whether the disorder progression is affected by onset, we examine the p-values for onset as a predictor of the disorder from each model in **Figure 2.7**. We see that onset is not statistically significant in any of the models.

Figure 2.6: QIC Score for GEE by Correlation Structure

	Independent	Exchangeable	Unstructured	Autoregressive
QIC	269.6359	267.5885	267.7774	267.2283

Figure 2.7: Onset P-Values by Model

	Independent P-Vals	Exchangeable P-Vals	Unstructured P-Vals	Autoregressive P-Vals
(Intercept)	0.0146440	0.0146659	0.0261017	0.0176799
onset> 20 yrs	0.3640173	0.3547650	0.2993175	0.3741108

Problem #2, Part C: Investigate the question using mixed effects model (lmer) from the previous chapter.

Results: Here I used the *lmer* function to fit a linear mixed-effects model. **Figure 2.8** shows the coefficients within that model. Most notably, we see again that onset is not statistically significant as a predictor of the disorder classification.

Figure 2.8: Model Coefficients in Linear Mixed-Effect Model

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Linear Hypotheses:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept) == 0    -0.5665     0.2314  -2.448   0.0144 *
## onset> 20 yrs == 0  -0.4336     0.4443  -0.976   0.3290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Univariate p values reported)
```

Problem #2, Part D: Is there a difference? Which model(s) work(s) best? Describe your results.

Results: Since, up to this point, I have not seen a difference in onset's effectiveness as a predictor of the disorder progression, we can take a look at the classification error rate from each of the above referenced models.

Figure 2.9 summarizes these error rates. As we can see, among the GEE models, there is no difference in the error rate. The LMER model was slightly better at predicting the disorder classification, but it still has a very high error rate at over 30%.

Ultimately, it does not appear that onset is a good predictor for this dataset.

Figure 2.9: Misclassification Rate by Model

Independent Error	Exchangeable Error	Unstructured Error	Autoregressive Error	Linear Mixed-Effects Error
0.3529412	**0.3529412**	**0.3529412**	**0.3529412**	**0.3039216**