Homework 10

Alex Soupir

November 23, 2019

Packages: HSAUR3, gee, lme4, Matrix, multcomp, ggplot2, plyr, tidyr, MESS, MuMIn

Collaborators:

Chapter 14

Please do the following problems from the text book R Handbook and stated.

1. Consider the **respiratory** data from the **HSAUR3** package.

   a. Investigate the use of other correlational structures than the independence and exchangeable structures used in the text for the respiratory data.

```
##  (Intercept)      centre2       trttrt   gendermale baselinegood          age
##  -0.90017133   0.67160098   1.29921589   0.11924365   1.88202860  -0.01816588

##  (Intercept)      centre2       trttrt   gendermale baselinegood          age
##  -0.90017133   0.67160098   1.29921589   0.11924365   1.88202860  -0.01816588

##  (Intercept)      centre2       trttrt   gendermale baselinegood          age
##  -0.90017133   0.67160098   1.29921589   0.11924365   1.88202860  -0.01816588

##  (Intercept)      centre2       trttrt   gendermale baselinegood          age
##  -0.90017133   0.67160098   1.29921589   0.11924365   1.88202860  -0.01816588

##
## Coefficients for correlation structure 'stat_M_dep':

## $coefficients
##                Estimate Naive S.E.    Naive z Robust S.E.    Robust z
## (Intercept)  -0.96294479 0.42071989 -2.2888026  0.46116071 -2.0880894
## centre2       0.74270146 0.29713257  2.4995626  0.35622997  2.0848932
## trttrt        1.24728240 0.29394223  4.2432909  0.35189738  3.5444492
## gendermale    0.11323226 0.36675166  0.3087437  0.44945059  0.2519348
## baselinegood  1.91139525 0.29916032  6.3892005  0.35018727  5.4582088
## age          -0.01691639 0.01101599 -1.5356209  0.01292727 -1.3085816

##
## Coefficients for correlation structure 'non_stat_M_dep':

## $coefficients
##                Estimate Naive S.E.    Naive z Robust S.E.    Robust z
## (Intercept)  -0.98263481 0.41928627 -2.3435893   0.4633336 -2.1207932
## centre2       0.74848974 0.29571189  2.5311452   0.3573102  2.0947897
## trttrt        1.24711331 0.29268905  4.2608814   0.3520803  3.5421273
## gendermale    0.11671467 0.36524841  0.3195488   0.4510766  0.2587469
## baselinegood  1.90029171 0.29758567  6.3856963   0.3505178  5.4213839
## age          -0.01646838 0.01096429 -1.5020019   0.0129291 -1.2737454

##
## Coefficients for correlation structure 'AR-M':

## $coefficients
##                Estimate Naive S.E.    Naive z Robust S.E.    Robust z
```

```
## (Intercept)   -0.96294479 0.44551609 -2.1614142  0.46116071 -2.0880894
## centre2        0.74270146 0.31464483  2.3604438  0.35622997  2.0848932
## trttrt         1.24728240 0.31126647  4.0071210  0.35189738  3.5444492
## gendermale     0.11323226 0.38836711  0.2915599  0.44945059  0.2519348
## baselinegood   1.91139525 0.31679209  6.0335952  0.35018727  5.4582088
## age           -0.01691639 0.01166525 -1.4501525  0.01292727 -1.3085816
##
## Coefficients for correlation structure 'unstructured':

## $coefficients
##               Estimate Naive S.E.    Naive z Robust S.E.    Robust z
## (Intercept) -0.93127982 0.47918516 -1.9434655  0.46124991 -2.0190352
## centre2      0.67279473 0.33907786  1.9841895  0.35482020  1.8961568
## trttrt       1.27891545 0.33544088  3.8126404  0.34944996  3.6597956
## gendermale   0.09467354 0.41729637  0.2268736  0.44362952  0.2134068
## baselinegood 1.93462516 0.34281837  5.6432949  0.34804681  5.5585200
## age         -0.01688917 0.01255736 -1.3449620  0.01290535 -1.3086948

##  (Intercept)      centre2        trttrt   gendermale baselinegood          age
## 4.348356e-02 5.793932e-02 2.524165e-04 8.310097e-01 2.720717e-08 1.906378e-01
```

Looking at the other correlation structures, minus the "fixed" correlation structure because the command never seemed to finish for some reason, is shown above with the coefficients for each covariate. The starting point for each variable is the same between models, as mentioned in the lecture video. However, the differences between the naive standard error and the robust standard error does differ between the correlation structures stat_M_dep, non_stat_M_dep, AR_M, and unstructured. The correlation structure that resulted in the greatest differences overall between the the naive s.e. and the robust s.e. was the non_stat_M_dep, but it wasn't great as the difference seen in the lecture model with the correlation structure of independence. The least difference between naive s.e. and robust s.e. was with the model using unstructured correlation structure.

Looking closer at the model model with the most similar naive s.e./robust s.e. we can see which of those variables are significant. Treatment is significant and so is the baseline, but gender and age do no show significance.

b. Which model is the best? Compare the following models: independent, exchangable, and what ever models you tried in part (a). Justify your answer. (Hint: use QIC (in **MESS**), MSE, misclassification rate, comparison of naive vs robust Z-score, or another method, be sure to state your method)

```
##  (Intercept)      centre2        trttrt   gendermale baselinegood          age
## -0.90017133   0.67160098   1.29921589   0.11924365   1.88202860  -0.01816588

##  (Intercept)      centre2        trttrt   gendermale baselinegood          age
## -0.90017133   0.67160098   1.29921589   0.11924365   1.88202860  -0.01816588

##
## Coefficients for correlation structure 'independence':

## $coefficients
##               Estimate  Naive S.E.   Naive z Robust S.E.    Robust z
## (Intercept) -0.90017133 0.337653052 -2.665965  0.46032700 -1.9555041
## centre2      0.67160098 0.239566599  2.803400  0.35681913  1.8821889
## trttrt       1.29921589 0.236841017  5.485603  0.35077797  3.7038127
## gendermale   0.11924365 0.294671045  0.404667  0.44320235  0.2690501
## baselinegood 1.88202860 0.241290221  7.799854  0.35005152  5.3764332
## age         -0.01816588 0.008864403 -2.049306  0.01300426 -1.3969169
```

```
##
## Coefficients for correlation structure 'exchangeable':

## $coefficients
##                 Estimate Naive S.E.    Naive z Robust S.E.    Robust z
## (Intercept)  -0.90017133  0.4784634 -1.8813796  0.46032700 -1.9555041
## centre2       0.67160098  0.3394723  1.9783676  0.35681913  1.8821889
## trttrt        1.29921589  0.3356101  3.8712064  0.35077797  3.7038127
## gendermale    0.11924365  0.4175568  0.2855747  0.44320235  0.2690501
## baselinegood  1.88202860  0.3419147  5.5043802  0.35005152  5.3764332
## age          -0.01816588  0.0125611 -1.4462014  0.01300426 -1.3969169

##
## MSE of state_M_dep model:

## IF/ELSE 1 or 0 for the model fitted values from lecture

## [1] 0.259009

## Raw value MSE

## [1] 0.1826503

##
## MSE of non_stat_M_dep model:

## IF/ELSE 1 or 0 for the model fitted values from lecture

## [1] 0.259009

## Raw value MSE

## [1] 0.1826893

##
## MSE of AR-M model:

## IF/ELSE 1 or 0 for the model fitted values from lecture

## [1] 0.259009

## Raw value MSE

## [1] 0.1826503

##
## MSE of unstructured model:

## IF/ELSE 1 or 0 for the model fitted values from lecture

## [1] 0.259009

## Raw value MSE

## [1] 0.18265

##
## MSE of independence model:

## IF/ELSE 1 or 0 for the model fitted values from lecture

## [1] 0.259009

## Raw value MSE

## [1] 0.1825935
```

```
## 
## MSE of exchangeable model:

## IF/ELSE 1 or 0 for the model fitted values from lecture

## [1] 0.259009

## Raw value MSE

## [1] 0.1825935
```

Adding to the coefficient naive and robust s.e. from Part A we can see that the independence correlation structure has a large difference. This means that we cannot trust the naive z values and want to use the robust, which show less significance due to them being more conservative (2.8 naive v 1.9 robust z). The MSE for the models using the code from lecture code results in the same MSE for all fo the models correcting to either 1 or 0 where 1 is good status and 0 is poor status, depending if the fitted value is >0.5 (1) or <0.5 (0). Not correcting to either 1 or 0 of the nstat and just keeping the raw value for MSE calculations resulted in some divergence but not a great deal (0.1825935 to 0.1826893). Because of this, sticking with the summary results comparing the naive s.e. to the robust s.e. makes more sense where the model with exchangeable, or even unstructured, correlation structures provided the least difference between them and allowing to trust the context of the model. If we were to go further, we coulld check for interaction terms between the insignificant or less significant variables as suggested in lecture.

2. The data set **schizophrenia2** from **HSAUR3** package were 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 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. (https://www.rdocumentation.org/packages/HSAUR3/versions/1.0-9/topics/schizophrenia2) Investigate this question using

   a. plots and summary statistics

```
## 
## Raw means for the onset age of the disease

##   < 20 yrs.0   > 20 yrs.0   < 20 yrs.2   > 20 yrs.2   < 20 yrs.6   > 20 yrs.6
##   0.62500000   0.58333333   0.61290323   0.50000000   0.28571429   0.25000000
##   < 20 yrs.8   > 20 yrs.8  < 20 yrs.10  > 20 yrs.10
##   0.17857143   0.00000000   0.11111111   0.09090909

## 
## Raw variance for the onset age of the disease

##   < 20 yrs.0   > 20 yrs.0   < 20 yrs.2   > 20 yrs.2   < 20 yrs.6   > 20 yrs.6
##   0.24193548   0.26515152   0.24516129   0.27272727   0.21164021   0.20454545
##   < 20 yrs.8   > 20 yrs.8  < 20 yrs.10  > 20 yrs.10
##   0.15211640   0.00000000   0.10256410   0.09090909
```
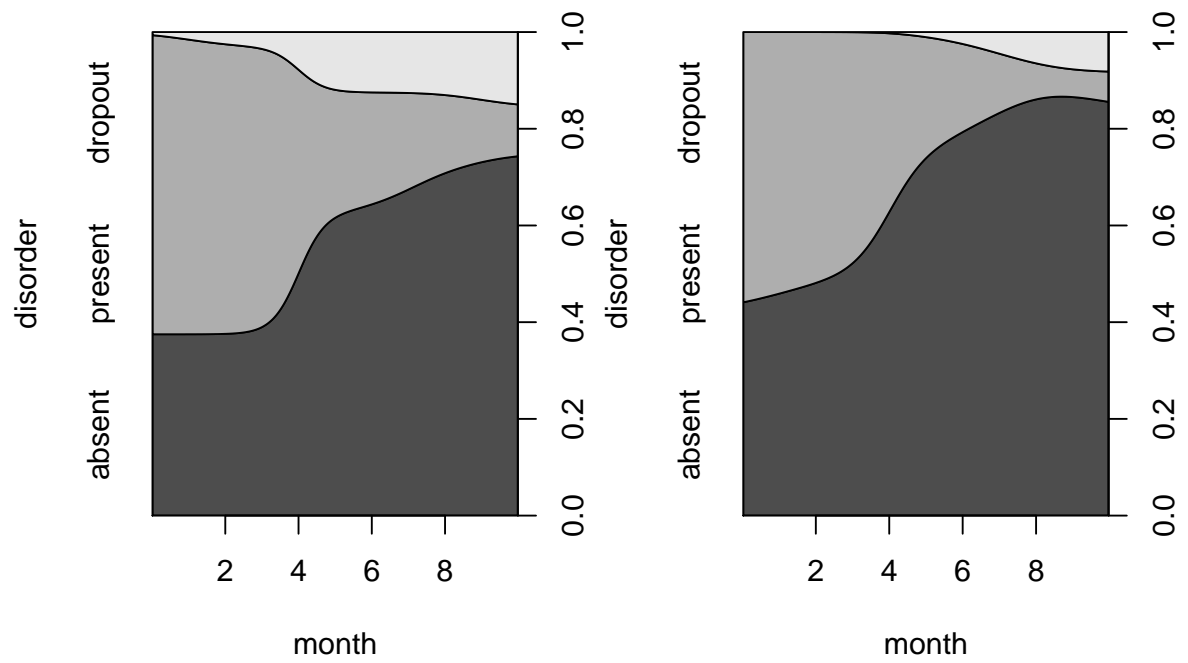
xtabs(~onset + month + disorder, data = sc)
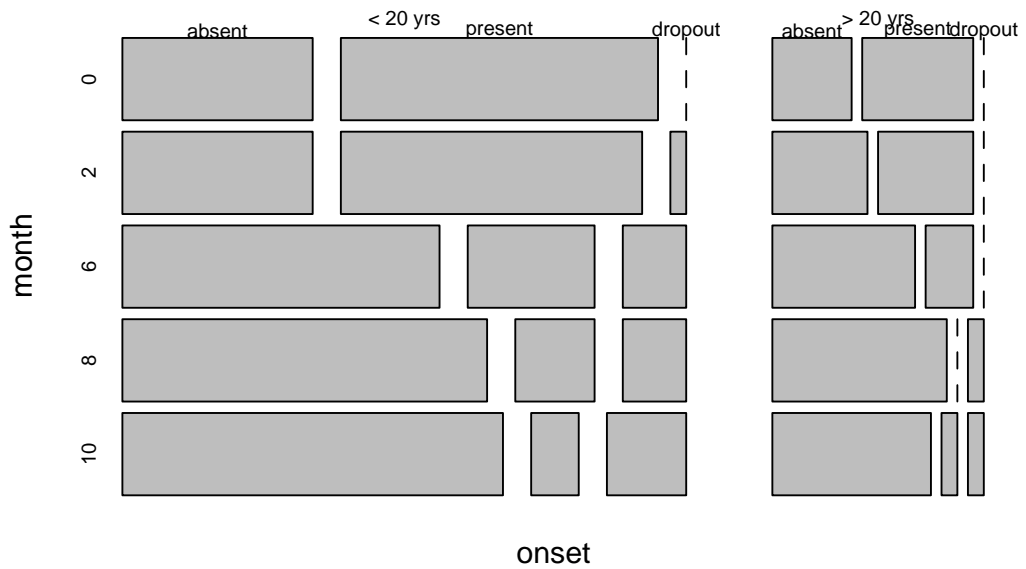

```
## 
## Summary statistics of onset before 20 years old:
```

```
##      subject              onset            disorder        month
## 1         :   5     < 20 yrs:160    absent :91     Min.    : 0.0
## 3         :   5     > 20 yrs:  0    present:55     1st Qu.: 2.0
## 4         :   5                     dropout:14     Median : 6.0
## 5         :   5                                    Mean    : 5.2
## 6         :   5                                    3rd Qu.: 8.0
## 7         :   5                                    Max.    :10.0
## (Other):130
##
## Summary statistics of onset after 20 years old:
##      subject            onset           disorder        month
## 2         : 5     < 20 yrs: 0     absent :41     Min.    : 0.0
## 10        : 5     > 20 yrs:60     present:17     1st Qu.: 2.0
## 11        : 5                     dropout: 2     Median : 6.0
## 13        : 5                                    Mean    : 5.2
## 15        : 5                                    3rd Qu.: 8.0
## 22        : 5                                    Max.    :10.0
## (Other):30
```

16 data points were removed because they didn't have full records like last week for use calculating the mean and variance like lecture, and the response (being categorical) was converted to a 1 (praesent) or 0 (absent). Plots and summary statsitics were done using the raw schizophrenia2 data.

The means and variance of each at time point don't appear to be too different from each other besides at time 8 where those patients with onset after 20 years of age was all absent thought disorder. In order to plot the data in a way that was appropriate for binary/factor response some kind of density plotting had to be used. I used a conditional density plot and a mosaic plot to display the data between the different onset times for the length after hospitalization. I kept the NA values in at each time point. As one would expect, as the time goes on the number of dropouts from the study increases. The number of absent disorders also increases, and this could be a reason that some of the patients dropped out, and the number of present disorders decreased. The mosaic plot displays this as well but shows the zero mean and varaince for the 8 month time point for those with the onset before 20 years old.

The main data frame was split by the onset variable and then *summary()* was run on both new data frames to show some summary statistics. The number of records for the less than 20 onset is 160 vs 60 records for the onset after than 20 years old. 9% of the time points from the under 20 onset are missing and 3% of the time points from over 20 onset is missing. The ratios can be more easily visualized than described through numbers.

   b. the GEE approach

```
##      (Intercept)    onset> 20 yrs baselinepresent
##        1.4651447        0.4007224      -0.9289455

##      (Intercept)    onset> 20 yrs baselinepresent
##        1.4651447        0.4007224      -0.9289455

##      (Intercept)    onset> 20 yrs baselinepresent
##        1.4651447        0.4007224      -0.9289455

##      (Intercept)    onset> 20 yrs baselinepresent
##        1.4651447        0.4007224      -0.9289455

##
```

```
## Coefficients for correlation structure 'independence':

## $coefficients
##                  Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)     1.4651447  0.3671049  3.9910791   0.4543513  3.2246954
## onset> 20 yrs   0.4007224  0.4184393  0.9576597   0.4665384  0.8589269
## baselinepresent -0.9289455  0.4066898 -2.2841622   0.4901649 -1.8951693

## geeglm QIC, not gee QIC:

##      QIC
## 191.1318

##
## Coefficients for correlation structure 'non_stat_M_dep':

## $coefficients
##                  Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)     1.3544067  0.4048959  3.3450738   0.4316356  3.1378479
## onset> 20 yrs   0.3390988  0.4622082  0.7336495   0.4711819  0.7196771
## baselinepresent -0.9644094  0.4502463 -2.1419595   0.4715176 -2.0453309

## geeglm QIC, not gee QIC:

##      QIC
## 191.1022

##
## Coefficients for correlation structure 'exchangeable':

## $coefficients
##                  Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)     1.3974811  0.4139557  3.3759198   0.4500977  3.1048393
## onset> 20 yrs   0.4356897  0.4829817  0.9020833   0.4685839  0.9298009
## baselinepresent -0.8597299  0.4622047 -1.8600629   0.4862539 -1.7680680

## geeglm QIC, not gee QIC:

##      QIC
## 191.1022

##
## Coefficients for correlation structure 'unstructured':

## $coefficients
##                  Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)     1.426753  0.3888937  3.6687482   0.4283397  3.3308913
## onset> 20 yrs   0.379727  0.4325486  0.8778828   0.4281747  0.8868505
## baselinepresent -1.198680  0.4292882 -2.7922490   0.4634501 -2.5864263

## geeglm QIC, not gee QIC:

##      QIC
## 190.1386

##
##
## gee model QIC using MuMIn::QIC
```

PACKAGES INTERFERE WITH EACH OTHER. The MESS package QIC function does not work on gee models but the MuMIn QIC fuction does work on them. The

MuMIn package was pulled from with MuMIn::QIC when calculating the QIC from the gee models. The following was written before finding this out.

Model selection table (Intrc) basln onset corstr qLik QIC delta weight schiz_gee6 1.427 + + u -92.704 190.1 0.00 0.342 schiz_gee3 1.354 + + n_s_M_d -91.928 190.8 0.69 0.242 schiz_gee4 1.397 + + e -91.567 191.1 0.98 0.209 schiz_gee1 1.465 + + i -91.549 191.1 1.01 0.206 Abbreviations: corstr: e = 'exchangeable', i = 'independence', n_s_M_d = 'non_stat_M_dep', u = 'unstructured' Models ranked by QIC(x)

The correlation structures that didn't work were the *stat_M_dep* and the *AR-M* structures. The error output says that the clustsize=1 and through searching the internet it says that the ID column (subject) should be sorted and have more than 1 in a row, which they were. I even went through and sorted again through the order() command and the error persists so I removed them from running in the code. Along with the *onset* parameter that is mentioned as being the only covariate, baseline was also added like with the respiratory data from lecture.

Additionally, The QIC command from the MESS package worked once, but not again after (*now noting that it was probably because the MuMIn package was loaded after or something along those lines*), on the models produced using gee(). Because of this, I copied the output from the time that it had worked into the text above. I have also tried getting the models to work using the geeglm function that QIC is supposed to work with but that gives an error for scale.value even with scale.fix is set to false (supposedly not using the scale.value in this case). Removing the scale.value parameter does give the models that work and the standard error is similar to that of the robust s.e., but uses a w (wald) value rather than z value. However, these models also throw an error using the recommended model.sel() because the rank function (QIC) returns a vector greater than 1 in length.. Furthermore, the results from the QIC on each model constructed using the geeglm() function from MESS package is different from the QIC of the table produced with QIC on the models from gee() function.

The coefficients from the gee() models suggest the best model is the "non_stat_M_dep" model with the lowest differences between naive and robust, but exchangeable is extremely close as well. This is supported by the QIC of the geeglm() models where the exchangeable and non_stat_M_dep have the same QIC.

AFTER FINDING OUT MESS QIC DOES NOT WORK ON GEE MODELS AND USING MuMIn QIC

Basically same story as above, but now able to repeatedly print out the model.sel QIC table and view the QIC of the gee models. The QIC values are different between the QIC from MESS and MuMIn (most likely due to the difference in models; gee vs geeglm). Because of this, the QIC to compare should probably be from the MuMIn package because it uses the gee models that are being used to compare the naive and robust s.e. values. According to the QIC of the gee models, the unstructured correlation structure produced the lowest QIC for the data. This is interesting because depending on the covariate that is being looked at, the non_stat_M_dep and exhangeable has slightly lower differences. The baseline of the unstructured correlation structure is the most significant of all models, and the onset factor is not significant for any model.

c. mixed effects model (lmer) from previous chapter

AIC of random intercept LME:

[1] 165.5155

AIC of random slop LME:

[1] 168.0985

d. Is there a difference? What model(s) work best? Describe your results.

QIC is quasi-AIC, so I would image they would be comparable? If so, the random intercept model produces the lowest AIC value at 166 and the random slope produces an AIC of 168. These values are a fair amount lower than those QIC/AIC values from the gerneralized estimation equations where the values were between 190.1 and 191.1. This suggests that the linear mixed effect models are a better fit for the data than are the generalized estimation equation models.

References

- journal.r-project.org
- cran.r-project.org (HSAUR3 documentation for the schizophrenia2 data)