Discovering Associations - Model 2

Binomial data of rose vase life days.

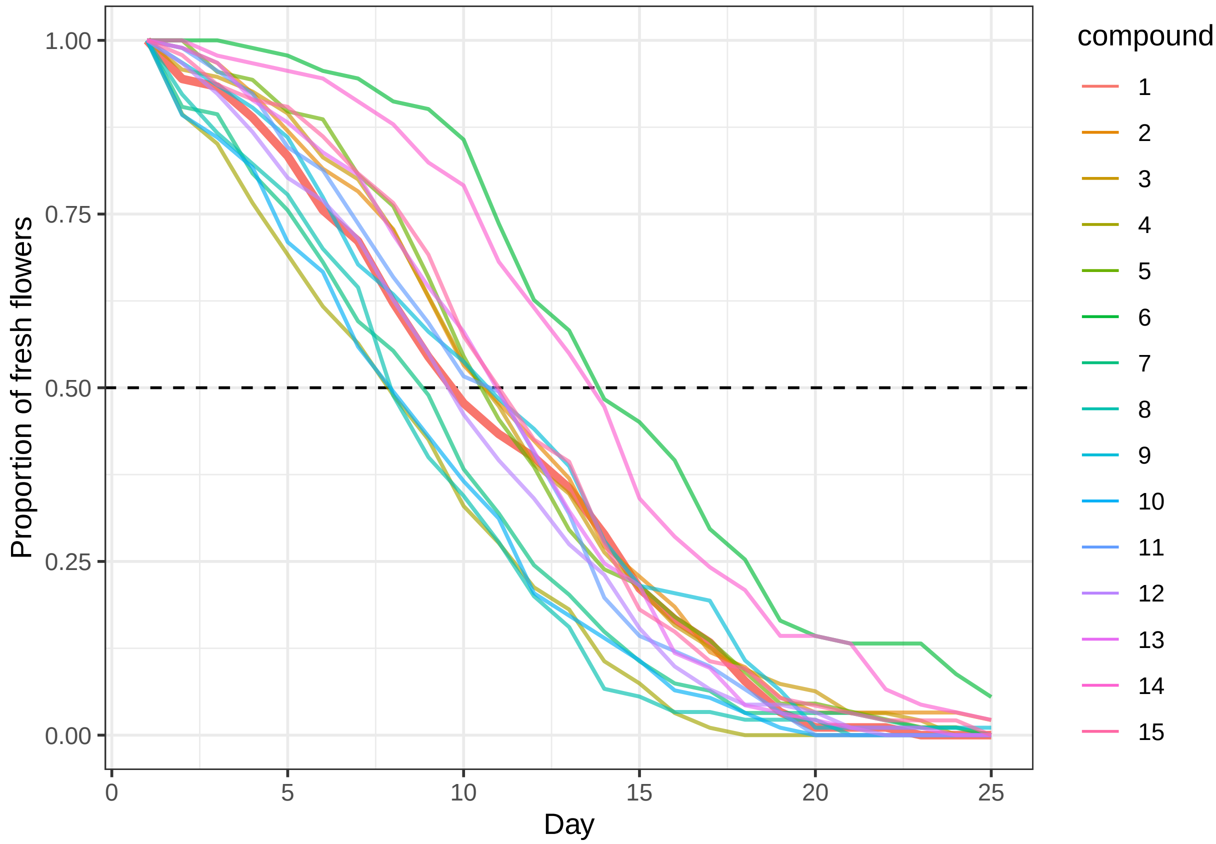
Out of 1440 datapoints, we have 60 missing outcomes (total vase days), which is 4.1% of the data. More data description, can copy some from simulation description.

Fit a longitudinal binary data predicting vase life. First need to transform the data into a binary outcome per day.

```
outmat<-matrix(nrow = nrow(d), ncol=max(d$tot.vase.days))

outmat[is.na(outmat)]<-1
for (i in 1:nrow(outmat)){
  outmat[i,c(d[i,tot.vase.days]:25)]<-0
  outmat[i,d[i,tot.vase.days]]<-1

}

outdf<-as.data.frame(outmat)
names(outdf)<-paste0("newVar_",names(outdf))

d_full<-d %>%
  bind_cols(outdf %>% as.data.frame()) %>%
  pivot_longer(contains("newVar"), names_to="day", values_to = "fresh") %>%
  mutate(day=as.numeric(gsub("newVar_V","",day)))

data_full_cc <- aggregate(fresh ~ compound + day, data = d_full, FUN = mean) %>%
  mutate(water=ifelse(compound==1,T,F))

ggplot(data = data_full_cc)+
  geom_hline(yintercept=0.5, linetype="dashed")+
  geom_line(aes(x = day, y = fresh, color = compound, size=water, alpha=water)) +
  scale_size_discrete(range=c(0.7,1.5),guide="none")+
  scale_alpha_discrete(range=c(0.65,1), guide="none")+
  theme_bw()+
  ylab("Proportion of fresh flowers")+
  xlab("Day")
```

```
## Warning: Using size for a discrete variable is not advised.
```

```
## Warning: Using alpha for a discrete variable is not advised.
```




```
#included this just to see if all datapoints are there..
#ggplot(data = data_full_cc)+geom_line(aes(x = day, y = fresh, color = compound))+facet_wrap(~compound)
```

We test Generalized Estimating Equations (GEEGLM function).

We Consider various specifications for the 'working' correlation structure Model1 = Independence Model2 = Exchangeable Model3 = Auto-regressive

```
gee_out1 <- geeglm(fresh ~ day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="independence")
summary(gee_out1)
```

```
##
## Call:
## geeglm(formula = fresh ~ day + day:compound, family = "binomial",
##     data = d_full, id = flowerID, corstr = "independence")
##
##  Coefficients:
##                  Estimate   Std.err     Wald Pr(>|W|)
## (Intercept)      3.673283  0.085808 1832.528  < 2e-16 ***
## day             -0.353073  0.017006  431.052  < 2e-16 ***
## day:compound2    0.021756  0.022610    0.926 0.335946
## day:compound3    0.017360  0.022097    0.617 0.432096
## day:compound4   -0.093738  0.026864   12.175 0.000484 ***
## day:compound5    0.015731  0.022176    0.503 0.478105
```

```
## day:compound6   0.098110  0.019487   25.347 4.79e-07 ***
## day:compound7  -0.047708  0.027396    3.033 0.081606 .
## day:compound8  -0.082888  0.027166    9.310 0.002279 **
## day:compound9   0.015367  0.022547    0.464 0.495529
## day:compound10 -0.072791  0.027067    7.232 0.007161 **
## day:compound11 -0.006274  0.021952    0.082 0.775011
## day:compound12 -0.018950  0.023574    0.646 0.421491
## day:compound13  0.006972  0.021335    0.107 0.743825
## day:compound14  0.078748  0.019713   15.958 6.48e-05 ***
## day:compound15  0.021031  0.021554    0.952 0.329181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.9163  0.3788
## Number of clusters:   1380  Maximum cluster size: 25
```

For model2 the interaction effects do not show the compound effects observed in the graphs.

```
gee_out2 <- geeglm(fresh ~ day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="exchangeable")
summary(gee_out2)
```

```
##
## Call:
## geeglm(formula = fresh ~ day + day:compound, family = "binomial",
##     data = d_full, id = flowerID, corstr = "exchangeable")
##
##  Coefficients:
##                 Estimate  Std.err    Wald Pr(>|W|)
## (Intercept)      3.56553  0.09528 1400.45   <2e-16 ***
## day             -0.33227  0.01791  344.29   <2e-16 ***
## day:compound2   -0.01015  0.02444    0.17   0.6779
## day:compound3   -0.00473  0.02495    0.04   0.8497
## day:compound4   -0.07091  0.02622    7.31   0.0068 **
## day:compound5   -0.01622  0.02598    0.39   0.5322
## day:compound6    0.05274  0.02312    5.20   0.0225 *
## day:compound7   -0.06140  0.02634    5.43   0.0198 *
## day:compound8   -0.08872  0.02733   10.53   0.0012 **
## day:compound9    0.00895  0.02367    0.14   0.7054
## day:compound10 -0.05717  0.02680    4.55   0.0329 *
## day:compound11 -0.02125  0.02510    0.72   0.3974
## day:compound12 -0.02558  0.02540    1.01   0.3139
## day:compound13 -0.01478  0.02434    0.37   0.5437
## day:compound14  0.04042  0.02318    3.04   0.0813 .
## day:compound15 -0.00432  0.02400    0.03   0.8570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
```

```
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.931   0.359
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.321   0.117
## Number of clusters:   1380  Maximum cluster size: 25
```

```
gee_out3 <- geeglm(fresh ~ day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="ar1")
summary(gee_out3)
```

```
##
## Call:
## geeglm(formula = fresh ~ day + day:compound, family = "binomial",
##     data = d_full, id = flowerID, corstr = "ar1")
##
##  Coefficients:
##                 Estimate  Std.err    Wald Pr(>|W|)
## (Intercept)      3.62152  0.08188 1956.13  < 2e-16 ***
## day             -0.34566  0.01656  435.48  < 2e-16 ***
## day:compound2    0.01515  0.02212    0.47  0.49331
## day:compound3    0.00965  0.02175    0.20  0.65739
## day:compound4   -0.08346  0.02522   10.95  0.00093 ***
## day:compound5    0.00539  0.02224    0.06  0.80839
## day:compound6    0.08768  0.01945   20.31  6.6e-06 ***
## day:compound7   -0.05232  0.02469    4.49  0.03406 *
## day:compound8   -0.08605  0.02561   11.29  0.00078 ***
## day:compound9    0.01472  0.02182    0.45  0.49998
## day:compound10 -0.07062  0.02538    7.74  0.00539 **
## day:compound11 -0.00889  0.02188    0.17  0.68457
## day:compound12 -0.02110  0.02297    0.84  0.35833
## day:compound13  0.00206  0.02120    0.01  0.92249
## day:compound14  0.06891  0.01959   12.38  0.00043 ***
## day:compound15  0.01389  0.02109    0.43  0.51009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.909   0.327
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.832   0.064
## Number of clusters:   1380  Maximum cluster size: 25
```

We compare the models using QIC: the auto-regressive working correlation seems to work better.

```r
QIC <- MuMIn::QIC
model.sel(gee_out1, gee_out2, gee_out3, rank = QIC)
```

```
## Model selection table
##          (Int)    day cmp:day corstr   qLik   QIC delta weight
## gee_out3  3.62 -0.346       +    ar1 -12111 24488   0.0  0.999
## gee_out1  3.67 -0.353       + indpnd -12104 24503  15.1  0.001
## gee_out2  3.57 -0.332       + exchng -12203 24587  99.7  0.000
## Abbreviations:
##  corstr: exchng = 'exchangeable', indpnd = 'independence'
## Models ranked by QIC(x)
```

We test more complex models.

Model4: adding species.

```r
gee_out4 <- geeglm(fresh ~ species + day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="ar1")
```

Model5: adding garden.

```r
gee_out5 <- geeglm(fresh ~ garden + day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="ar1")
```

Model6: adding rater.

```r
gee_out6 <- geeglm(fresh ~ rater + day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="ar1")
summary(gee_out6)
```

```
##
## Call:
## geeglm(formula = fresh ~ rater + day + day:compound, family = "binomial",
##     data = d_full, id = flowerID, corstr = "ar1")
##
##  Coefficients:
##                Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)     4.02824  0.14147 810.81  < 2e-16 ***
## rater2          0.02784  0.16167   0.03   0.8633
## rater3          2.16279  0.16814 165.47  < 2e-16 ***
## rater4          2.12926  0.17471 148.54  < 2e-16 ***
## rater5          1.01089  0.16644  36.89  1.2e-09 ***
## rater6         -1.65509  0.15827 109.36  < 2e-16 ***
## day            -0.43822  0.01593 757.16  < 2e-16 ***
## day:compound2   0.01529  0.02120   0.52   0.4709
## day:compound3   0.01289  0.02082   0.38   0.5360
## day:compound4  -0.11720  0.02608  20.20  7.0e-06 ***
## day:compound5   0.00728  0.02144   0.12   0.7342
## day:compound6   0.10995  0.01864  34.78  3.7e-09 ***
## day:compound7  -0.07929  0.02491  10.14   0.0015 **
## day:compound8  -0.12515  0.02636  22.54  2.1e-06 ***
## day:compound9   0.01132  0.02124   0.28   0.5939
```

```
## day:compound10 -0.09051  0.02558  12.51   0.0004 ***
## day:compound11 -0.01446  0.02156   0.45   0.5023
## day:compound12 -0.03201  0.02271   1.99   0.1587
## day:compound13  0.00176  0.02027   0.01   0.9306
## day:compound14  0.09084  0.01873  23.51  1.2e-06 ***
## day:compound15  0.01477  0.02023   0.53   0.4652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.846   0.583
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.788   0.142
## Number of clusters:   1380  Maximum cluster size: 25
```

Model7: adding rater and garden.

```r
gee_out7 <- geeglm(fresh ~ rater + garden + day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="ar1")
summary(gee_out7)
```

```
##
## Call:
## geeglm(formula = fresh ~ rater + garden + day + day:compound,
##     family = "binomial", data = d_full, id = flowerID, corstr = "ar1")
##
##  Coefficients:
##                Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)     3.71685  0.14454 661.22  < 2e-16 ***
## rater2          0.05522  0.15847   0.12  0.72749
## rater3          2.25502  0.17095 174.01  < 2e-16 ***
## rater4          2.19912  0.17552 156.98  < 2e-16 ***
## rater5          1.04050  0.16491  39.81  2.8e-10 ***
## rater6         -1.68434  0.16043 110.23  < 2e-16 ***
## garden2         0.77607  0.09702  63.99  1.2e-15 ***
## day            -0.45205  0.01653 747.70  < 2e-16 ***
## day:compound2   0.01675  0.02162   0.60  0.43853
## day:compound3   0.01733  0.02101   0.68  0.40937
## day:compound4  -0.11851  0.02626  20.36  6.4e-06 ***
## day:compound5   0.00983  0.02188   0.20  0.65313
## day:compound6   0.11598  0.01892  37.58  8.8e-10 ***
## day:compound7  -0.07961  0.02501  10.13  0.00146 **
## day:compound8  -0.12624  0.02696  21.92  2.8e-06 ***
## day:compound9   0.01546  0.02157   0.51  0.47345
## day:compound10 -0.09130  0.02598  12.35  0.00044 ***
## day:compound11 -0.01413  0.02136   0.44  0.50828
## day:compound12 -0.03177  0.02297   1.91  0.16655
```

```
## day:compound13  0.00443  0.02058   0.05  0.82942
## day:compound14  0.09565  0.01896  25.46  4.5e-07 ***
## day:compound15  0.01859  0.02060   0.82  0.36664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.905    1.05
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.786   0.238
## Number of clusters:   1380  Maximum cluster size: 25
```

```
model.sel(gee_out1, gee_out2, gee_out3, gee_out4, gee_out5, gee_out6,
          gee_out7, rank = QIC)
```

```
## Model selection table
##          (Int)    day cmp:day spc grd rtr corstr   qLik   QIC delta weight
## gee_out7  3.72 -0.452       +       +   +    ar1  -9428 19176     0      1
## gee_out6  4.03 -0.438       +           +    ar1  -9626 19578   402      0
## gee_out5  3.36 -0.352       +       +        ar1 -11946 24166  4990      0
## gee_out3  3.62 -0.346       +                ar1 -12111 24488  5312      0
## gee_out4  3.70 -0.346       +   +            ar1 -12110 24499  5324      0
## gee_out1  3.67 -0.353       +             indpnd -12104 24503  5327      0
## gee_out2  3.57 -0.332       +             exchng -12203 24587  5412      0
## Abbreviations:
##  corstr: exchng = 'exchangeable', indpnd = 'independence'
## Models ranked by QIC(x)
```

We compare with a similar model corresponding to Model 7, but using an independence working correlation.

```
gee_out8 <- geeglm(fresh ~ rater + day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="independence")
summary(gee_out8)
```

```
##
## Call:
## geeglm(formula = fresh ~ rater + day + day:compound, family = "binomial",
##     data = d_full, id = flowerID, corstr = "independence")
##
##  Coefficients:
##                Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)     4.08875  0.14375 809.08  < 2e-16 ***
## rater2          0.04228  0.16040   0.07  0.79208
## rater3          2.19993  0.16583 175.99  < 2e-16 ***
## rater4          2.19563  0.17866 151.03  < 2e-16 ***
## rater5          1.00518  0.16523  37.01  1.2e-09 ***
```

```
## rater6         -1.68894  0.15797 114.31  < 2e-16 ***
## day            -0.45022  0.01638 755.03  < 2e-16 ***
## day:compound2   0.02408  0.02185   1.21  0.27055
## day:compound3   0.02018  0.02120   0.91  0.34109
## day:compound4  -0.12011  0.02754  19.02  1.3e-05 ***
## day:compound5   0.01685  0.02154   0.61  0.43418
## day:compound6   0.12117  0.01897  40.79  1.7e-10 ***
## day:compound7  -0.06294  0.02731   5.31  0.02119 *
## day:compound8  -0.11458  0.02889  15.73  7.3e-05 ***
## day:compound9   0.01840  0.02167   0.72  0.39576
## day:compound10 -0.09298  0.02799  11.04  0.00089 ***
## day:compound11 -0.00687  0.02123   0.10  0.74619
## day:compound12 -0.02545  0.02369   1.15  0.28270
## day:compound13  0.00698  0.02110   0.11  0.74063
## day:compound14  0.10022  0.01897  27.91  1.3e-07 ***
## day:compound15  0.02422  0.02149   1.27  0.25968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.845   0.555
## Number of clusters:   1380  Maximum cluster size: 25
```

Adding compounds to the model provides a lower QIC but the effect is not anymore present in the interaction term.

```
gee_out9 <- geeglm(fresh ~ compound + rater + garden + day + day:compound,
id=flowerID, data=d_full, family="binomial", corstr="ar1")
summary(gee_out9)
```

```
##
## Call:
## geeglm(formula = fresh ~ compound + rater + garden + day + day:compound,
##     family = "binomial", data = d_full, id = flowerID, corstr = "ar1")
##
##  Coefficients:
##                Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)     3.56588  0.34519 106.71  < 2e-16 ***
## compound2       0.33553  0.50429   0.44   0.5058
## compound3       0.39623  0.48357   0.67   0.4126
## compound4      -0.65581  0.41387   2.51   0.1131
## compound5       0.63419  0.50132   1.60   0.2059
## compound6       1.50878  0.55067   7.51   0.0061 **
## compound7      -0.47156  0.45178   1.09   0.2966
## compound8      -0.41656  0.44867   0.86   0.3532
## compound9      -0.08784  0.43337   0.04   0.8394
## compound10     -0.70343  0.42252   2.77   0.0959 .
## compound11      0.56514  0.47422   1.42   0.2334
## compound12     -0.00679  0.45304   0.00   0.9880
## compound13      0.77219  0.50434   2.34   0.1258
```

```
## compound14      1.32316  0.52604   6.33   0.0119 *
## compound15      0.61900  0.51780   1.43   0.2319
## rater2          0.04778  0.15912   0.09   0.7640
## rater3          2.29281  0.17422 173.21  < 2e-16 ***
## rater4          2.24668  0.17867 158.12  < 2e-16 ***
## rater5          1.03818  0.16735  38.48  5.5e-10 ***
## rater6         -1.72236  0.15928 116.93  < 2e-16 ***
## garden2         0.79515  0.09794  65.92  4.4e-16 ***
## day            -0.44312  0.02600 290.54  < 2e-16 ***
## compound2:day  -0.00784  0.04320   0.03   0.8559
## compound3:day  -0.01126  0.04016   0.08   0.7792
## compound4:day  -0.05714  0.03952   2.09   0.1482
## compound5:day  -0.03789  0.04329   0.77   0.3814
## compound6:day   0.02307  0.04094   0.32   0.5731
## compound7:day  -0.03511  0.04392   0.64   0.4239
## compound8:day  -0.08439  0.04413   3.66   0.0558 .
## compound9:day   0.02199  0.03618   0.37   0.5433
## compound10:day -0.02817  0.03970   0.50   0.4780
## compound11:day -0.05711  0.04019   2.02   0.1553
## compound12:day -0.03014  0.03964   0.58   0.4469
## compound13:day -0.05224  0.04037   1.67   0.1956
## compound14:day  0.01227  0.03890   0.10   0.7524
## compound15:day -0.02632  0.04171   0.40   0.5281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.906   0.923
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.786    0.21
## Number of clusters:   1380  Maximum cluster size: 25
```

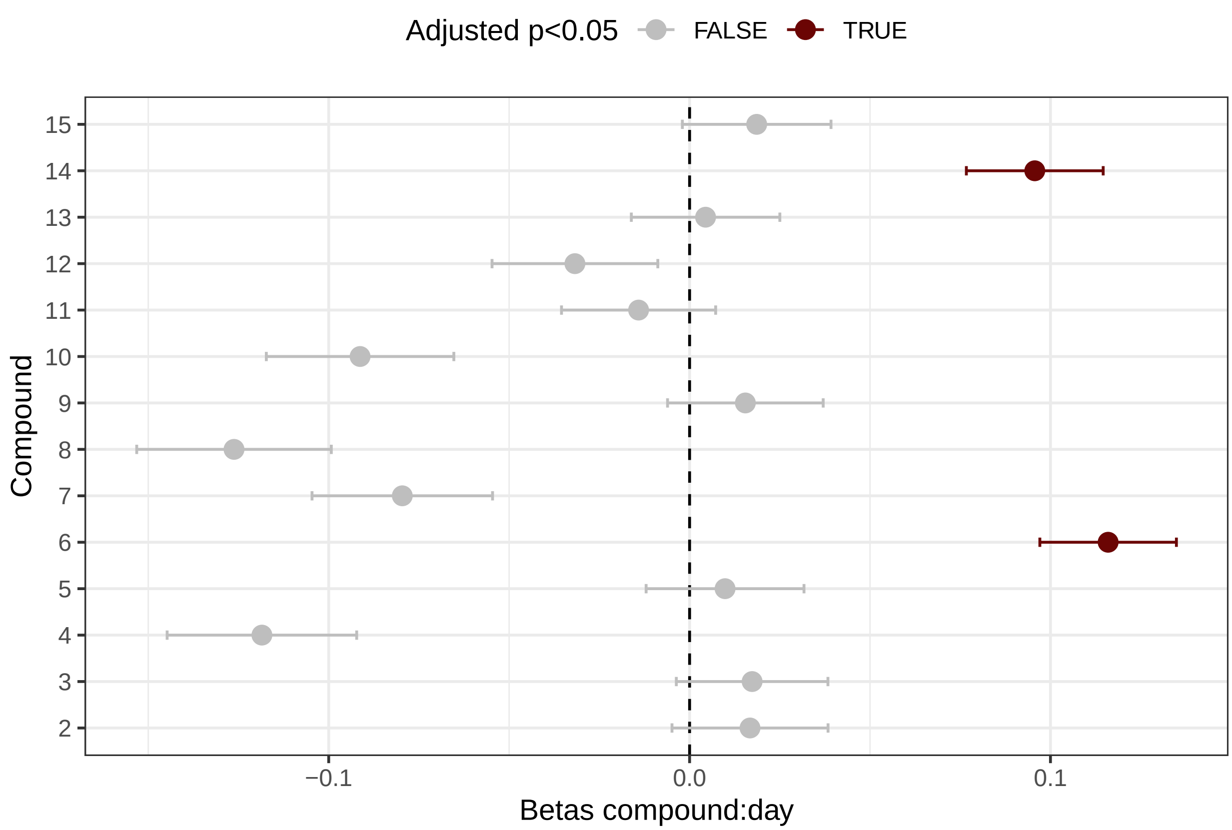
```
model.sel(gee_out1, gee_out2, gee_out3, gee_out4, gee_out5, gee_out6,
          gee_out7, gee_out8,gee_out9, rank = QIC)
```

```
## Model selection table
##          (Int)    day cmp:day spc grd rtr cmp corstr   qLik   QIC delta weight
## gee_out9  3.57 -0.443       +       +   +   +    ar1  -9312 19065     0      1
## gee_out7  3.72 -0.452       +       +   +        ar1  -9428 19176   111      0
## gee_out6  4.03 -0.438       +           +        ar1  -9626 19578   513      0
## gee_out8  4.09 -0.450       +           +     indpnd  -9621 19602   538      0
## gee_out5  3.36 -0.352       +       +            ar1 -11946 24166  5102      0
## gee_out3  3.62 -0.346       +                    ar1 -12111 24488  5423      0
## gee_out4  3.70 -0.346       +   +                ar1 -12110 24499  5435      0
## gee_out1  3.67 -0.353       +                 indpnd -12104 24503  5438      0
## gee_out2  3.57 -0.332       +                 exchng -12203 24587  5523      0
## Abbreviations:
##  corstr: exchng = 'exchangeable', indpnd = 'independence'
```

```
## Models ranked by QIC(x)
```

Results with model 7

```
gee_coefficients<-as.data.frame(summary(gee_out7)$coefficients) %>%
  rownames_to_column("predictor") %>%
  filter(grepl("day:compound",predictor)) %>%
  dplyr::rename(pval=`Pr(>|W|)`) %>%
  dplyr::mutate(one_sided_pval=ifelse(Estimate>0, pval/2, (1-pval/2)),
                p_adjusted=p.adjust(one_sided_pval, method="holm"),
         significant_higher=ifelse(p_adjusted<0.05, T, F))

ggplot(gee_coefficients %>%
         mutate(compound=factor(gsub("day:compound","",predictor), levels=2:15)),
       aes(x=compound, y=Estimate, color=p_adjusted<0.05))+
  geom_hline(yintercept=0, linetype="dashed")+
  geom_errorbar(aes(ymin=Estimate - Std.err, ymax=Estimate +Std.err), width=0.2)+geom_point(size=3)+ther
  scale_color_manual(values=c("grey","#6B0504"), name="Adjusted p<0.05")+coord_flip()+
  ylab("Betas compound:day")+
  xlab("Compound")+
  theme(legend.position = "top")
```



```
#scale_()+
#theme(axis.title.x = element_text(size=15))
```