Stat 414 - Day 26
Model Diagnostics: Influential Observations (Ch. 10)

Last Time: Model Diagnostics: Residuals

- Conditional (within group) vs. Marginal (overall "average") residuals
  - residuals(model1) reports the conditional residuals
  - Examine residuals for gross violations of normality, unusual observations
  - patterns may also suggest potential model improvements (including unused variables)

There are several ways to measure the amount of influence of observation(s) to a model, e.g.,

- DFBetas are measured for each observation in the dataset and indicate how much the slope coefficient of a variable changes if that observation is removed (will consider large if > $\frac{2}{\sqrt{n}}$).
- Cook's Distance is measured for each observation and provides an overall measure of how much all of the model predictions change when that observation is removed (large if > $4/(n-2)$). Found by considering each observation's leverage and residual.

The influence.ME package appears to be a good one for detecting influence of Level 2 groups.
https://journal.r-project.org/archive/2012-2/RJournal_2012-2_Nieuwenhuis~et~al.pdf
The influence function looks at

- DFBetas (can specify which variables/parameters you want to focus on)
- Cook's distance (normally across all variables/parameters)
- Sigtest (you can supply cut-off value for significance and then see whether judgement of significance changes)

If the group sizes are not too large, you can use obs=TRUE to also (separately) analyze the influence of Level 1 observations.

Example 1: Cigarette study

```
#install.packages("influence.ME")
library(influence.ME)
modelinfstats <- influence(model1, "subjectID")
plot(modelinfstats,
     which="dfbetas",
     xlab="DFbetaS",
     ylab="subject ID")

dfbetas(modelinfstats)
```

(a) Are any of the dfbeta variable values larger than $2/\sqrt{(n)}$? Show your work (n = number of subjects, we are focusing on level 2 here). Which subject(s)? What does this tell you?

```
cigstudy2 = cigstudy[which(cigstudy$subjectID!=7),]
model1b = lmer(Cigs ~ Time + BDI + Time:BDI + FTND + (1 |subjectID),  …
```

Now examine the Cook's Distances

```
plot(modelinfstats, which="cook",
     cutoff=.5, sort=TRUE,
     xlab="Cook´s Distance",
     ylab="Subject ID")
```

(b) Where did the "cut-off" value come from? What do you learn?

(c) Does removing any of the subjects change the significance of Time variable?

```
sigtest(modelinfstats, test=-1.96)$Time
summary(model1b)
```

(d) Does removing any of the subjects change the significance of FTND variable?

```
sigtest(modelinfstats, test=-1.96)$FTND
```

(e) What would it mean to have an influential observation at Level 1?

```
modelinfstats.obs <- influence(model1, obs=TRUE)
cooks.d <- cooks.distance(modelinfstats.obs)
plot(cooks.d)
which(cooks.d > 4/8)
sigtest(modelinfstats.obs, test=-1.96)$Time
sigtest(modelinfstats.obs, test=-1.96)$FTND
```

(f) Summarize what you learn from the above output

See for yourself if removing that one observation changes the model much:

```
model2 <- exclude.influence(model1, obs=28)
summary(model2)

#to exclude subject 7
model3 <- exclude.influence(model1, grouping = "subjectID", level = "7")
summary(model3)
```

If you want just a couple of graphs, remember you can use plot(model1) and plot(ranef(model1)) to focus on the level 1 and level 2 residuals.

The above graphs are a bit old school, what do you find with the following commands?

```
#install.packages("performance")
performance::check_model(model1)
```

(g) What new information is presented in this output?