Zygosity and Response Rate

My strategy for longitudinal genetic models in CoTwins has been to fit a multilevel growth model to the phenotype, use the conditional means of the random effects and the estimated fixed effects to get growth model parameters for each twin, and use those estimated parameters as the phenotypes in a standard OpenMx multivariate twin model. Multilevel models implement partial pooling, where random effect estimates for a twin or twin pair will be pulled toward the sample mean in proportion to the amount of information available for that unit.

If monozygotic twins are significantly more correlated on their response rates than dizygotic twins, this might inflate our heritability estimates. I'm not yet convinced that this would happen because twin pairs with high sample size and little pooling seem like they could cancel out the effect of twin pairs with low sample size and a lot of pooling. In other words, is a systematic skew actually introduced?

```r
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)

sub_use <- read_rds("../data/processed/remote_substance_use.rds")
id_mapping <- read_csv("../data/processed/id_mapping.csv", col_types = "ccccc")

sub_use <- sub_use %>% group_by(user_id) %>% summarize(N = n())

id_mapping <- select(id_mapping, T1_id = T1_alternate_id, T2_id = T2_alternate_id, bestzygos)

id_mapping <- left_join(id_mapping, sub_use, by = c("T1_id" = "user_id"))
id_mapping <- select(id_mapping, T1_id:bestzygos, T1_N = N)

id_mapping <- left_join(id_mapping, sub_use, by = c("T2_id" = "user_id"))
id_mapping <- select(id_mapping, T1_id:T1_N, T2_N = N)
```

Now that we have the number of weekly substance use surveys completed for twins 1 and 2 in each twin pair, we can split by zygosity and calculate the correlations.

```r
mz_data <- filter(id_mapping, bestzygos == "MZ") %>% na.omit()
dz_data <- filter(id_mapping, bestzygos %in% c("DZ", "OS")) %>% na.omit()
```

The dizygotic correlation is 0.3963885 and the monozygotic correlation is 0.5592712 so monozygotic twins are more correlated. I'll note that this difference isn't significant when tested but I don't think that matters too much either way:

```r
psych::r.test(
  n = nrow(dz_data),
  n2 = nrow(mz_data),
```

```
  r12 = cor(dz_data$T1_N, dz_data$T2_N),
  r34 = cor(mz_data$T1_N, mz_data$T2_N))
```

```
## Correlation tests
## Call:psych::r.test(n = nrow(dz_data), r12 = cor(dz_data$T1_N, dz_data$T2_N),
##     r34 = cor(mz_data$T1_N, mz_data$T2_N), n2 = nrow(mz_data))
## Test of difference between two independent correlations
##  z value 1.79    with probability  0.07
```

Let's plot what this looks like.

```
id_mapping$bestzygos[id_mapping$bestzygos == "OS"] <- "DZ"
ggplot(id_mapping, aes(T1_N, T2_N, color = bestzygos)) +
  geom_point() +
  geom_smooth(method = "lm")
```