Exploratory Data Analysis: Axon Regeneration

```
df <- read_table("../Data/untreated.sni.untreated.ddi.txt",
                 show_col_types = F)
```

Couple notes

Kruskal-Walis is the right idea, but it's non-parametric. Generally, if parametric assumptions seems reasonably, that's preferred because it makes computing confidence intervals and p-values easier. There's exceptions, but you'll also get more power

Quick and Dirty Look at Replication etc.

- Two "treatments": time (hours) and `Group`
- Measurements were *not* repeated –> independence between time points

```
df %>%
  group_by(Group, Hour) %>%
  summarise(count = n())
```

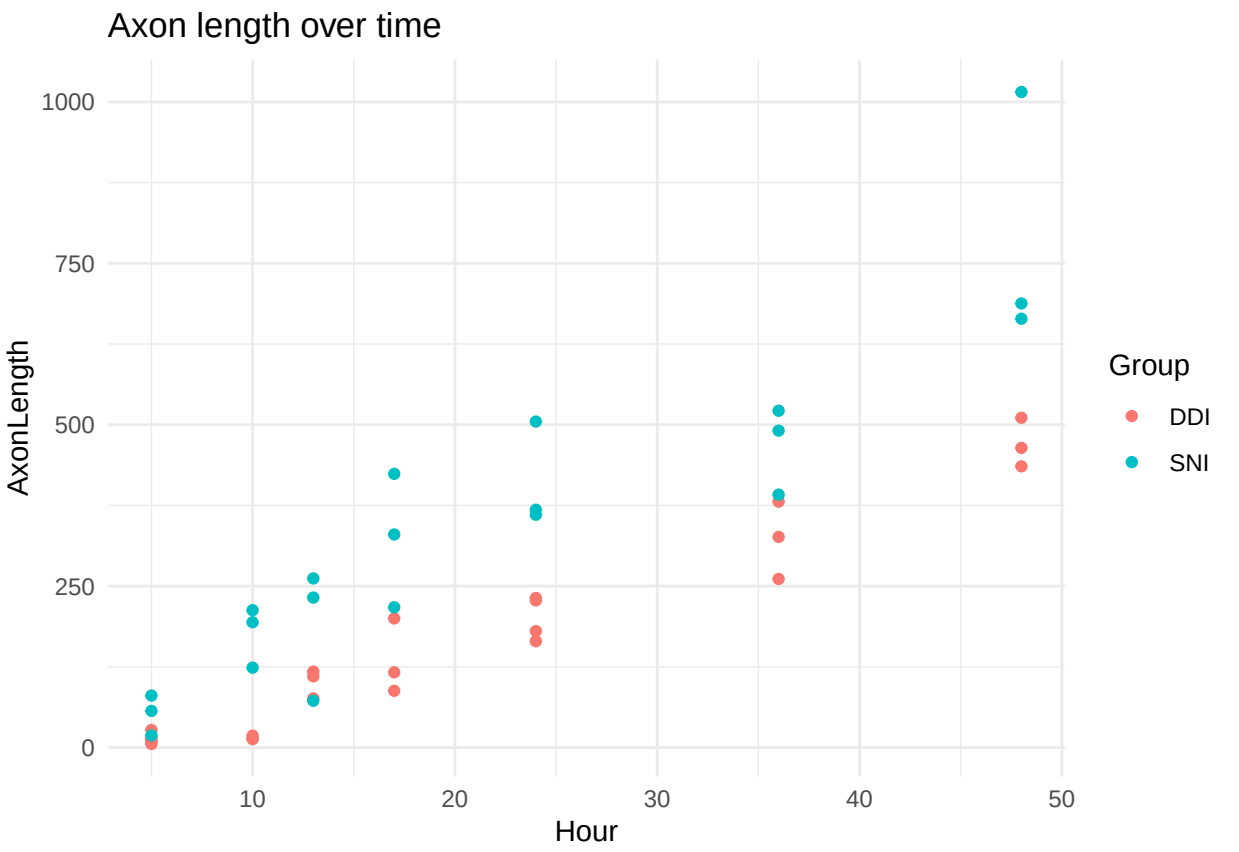
```
## `summarise()` has grouped output by 'Group'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 14 x 3
## # Groups:   Group [2]
##    Group  Hour count
##    <chr> <dbl> <int>
##  1 DDI       5     5
##  2 DDI      10     3
##  3 DDI      13     3
##  4 DDI      17     3
##  5 DDI      24     4
##  6 DDI      36     3
##  7 DDI      48     3
##  8 SNI       5     3
##  9 SNI      10     3
## 10 SNI      13     3
## 11 SNI      17     3
## 12 SNI      24     3
## 13 SNI      36     3
## 14 SNI      48     3
```

Vast majority of treatment levels have three replicates, a couple have four or five.

EDA Plot: What do the measurements look like over time?

```
df %>%
  ggplot(aes(x = Hour, y = AxonLength, color = Group)) +
  geom_point() +
  theme_minimal() +
  ggtitle("Axon length over time")
```



Slopes are well separated. The `SNI` sample in the top right gives me pause–it may be high leverage and disproportionately drive the slope estimate. Slopes are reasonably linear. Looks like we can force the intercept to zero. There's some funneling of variance–data get a bit more variable as time passes, but nothing so extreme as to justify a transformation in my opinion. Most of that increase in variance is just from that outlier. In other words, no need to log or square root `AxonLength`.

Model 1: No Intercept

Let Y_i be the random variable representing axon length for subject i. Let $X_{i,1}$ be the time (in hours) that the measurement was taken, and let $X_{i,2}$ be the indicator random variable for group (i.e. $X_{1,2} = 1$ if subject 1 is `DDI`). As usual, let σ^2 be an unknown but fixed variance parameter to a centered Gaussian. The error term is iid $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

We do not include an intercept because axon length at time zero is reasonably assumed to be zero (or close enough).

Let $\beta_1, \beta_2, \beta_3$ be the associated regression coefficients for time (`Hours`), group (`isDDI`), and an interaction term. Our model is then

$Y_i \sim \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i$

The model is fit below, with light pre-processing.

```
data <- df %>%
  dplyr::mutate(isDDI = ifelse(Group == "DDI", 1, 0))

model <- lm("AxonLength ~ Hour + isDDI + Hour:isDDI - 1", data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: AxonLength
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Hour        1 4373024 4373024 814.071 < 2.2e-16 ***
## isDDI       1  263111  263111  48.980  1.46e-08 ***
## Hour:isDDI  1   87628   87628  16.313 0.0002234 ***
## Residuals  42  225615    5372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Before accounting for multiple testing, etc. we are asking the following questions, with some abuse of language to speed things up.

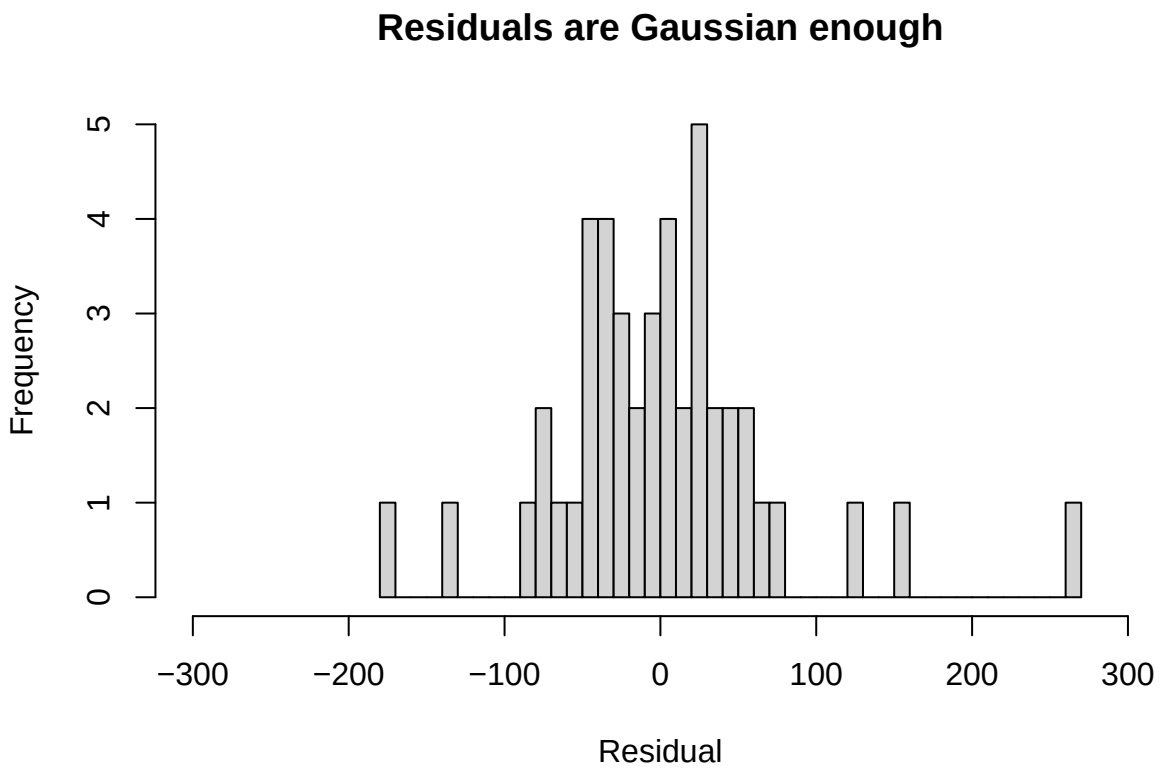
1. Is there an association between `Hour` and `AxonLength` after adjusting for group and interaction? Yes
2. Is there a difference in adjusted means between `DDI` and `SNI` (not quite the same as a t-test result but pretty close)? Yes.
3. Finally, is there a time by group interaction? Yes.

Number 3 is the real question of interest. You can already tell–and I'm sure you've already modeled this–that the interaction between `Hour` and `Group` is significant. We really should do this marginally–what do we get

Model diagnostics

Residuals grossly non-normal? No.

```
hist(model$residuals, breaks = 40,
     main = "Residuals are Gaussian enough",
     xlab = "Residual",
     xlim = c(-300, 300))
```



How do the predictions look? Pretty damn good.

```
y <- data$AxonLength
y.hat <- predict(model, data)

data.frame(y=y, y=y.hat) %>%
  ggplot(aes(x = y, y = y.hat)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
  xlab("y (ground truth)") +
  ylab("y.hat (predicted)") +
  labs(caption = "Red line indicates perfect prediction") +
  ggtitle("Predictions from model are reasonable")
```

Question of Interest: At what timepoint is there a significant difference between groups?

This is really getting at "what's the earliest time point that has a difference...

If you wanted to be really precise and ask all sorts of interesting questions of the data, you could fit a Bayesian model. This gets you posterior distributions that you can play with. For example, at time (plug in some number between 0 and 50 hours), what's the probability that the axon length for `DDI` and `SNI` is more than 50 microns different? This is why I was curious about practical significance.

I'd be happy to set this up, but it will take more time and I would not be able to get to it until after the semester wraps up.

In frequentist spaces like traditional linear regression / ANOVA above, we can't quite ask those questions because we're dealing with coverage probabilities (some bullshit). We've established the linear model fits well above.

```
all_hours <- sort(unique(data$Hour))

# Pack p-values into data frame
tests <- data.frame(Hr = all_hours,
                    p = rep(NA, length(all_hours)))

for (ii in 1:nrow(tests)){
  hh <- tests[ii, "Hr"]
```

```
  aa <- data %>% dplyr::filter(Hour == hh, Group == "DDI") %>% pull(AxonLength)
  bb <- data %>% dplyr::filter(Hour == hh, Group == "SNI") %>% pull(AxonLength)
  tests[ii, "p"] <- t.test(aa, bb)$p.val
}

tests
```

```
##   Hr          p
## 1  5 0.16156040
## 2 10 0.02659029
## 3 13 0.27268360
## 4 17 0.06664284
## 5 24 0.03405613
## 6 36 0.05074635
## 7 48 0.10148051
```

Hour 13 is why the formulation is problematic. How do you handle a significant difference (hour 10), followed by an insignificant one?

I'll think about it some more, this is all I had time for at the moment. I think Bayes is the way to go... But I'm sure someone has come across this before. Especially in clinical trial literature. How do you decide when curves/functions are "far enough" away from each other.