BIOS 6612 HW 5: Linear Mixed Models

In a study of exercise therapies, 37 patients were assigned to one of two weightlifting programs. In the first program (treatment 1), the number of repetitions was increased as subjects became stronger. In the second program (treatment 2), the number of repetitions was fixed but the amount of weight was increased as a subject became stronger. Measures of strength were taken at baseline (day 0), and on days 2,4,6,8,10, and 12.

The raw data are stored in an external file: `exercise_therapy.dat`

Each row of the data set contains the following variables:

- `id`: the subject identifier
- `time`: time in days
- `trt`: a categorical variable coded 1 = Program 1 (increased number of repetitions), 2 = Program 2 (increased amount of weight)
- `y`: the outcome, a measure of strength

Exercise 1: EDA (20 points)

Perform exploratory data analysis to better understand the impact of the two weigthlifting programs on strength over time.

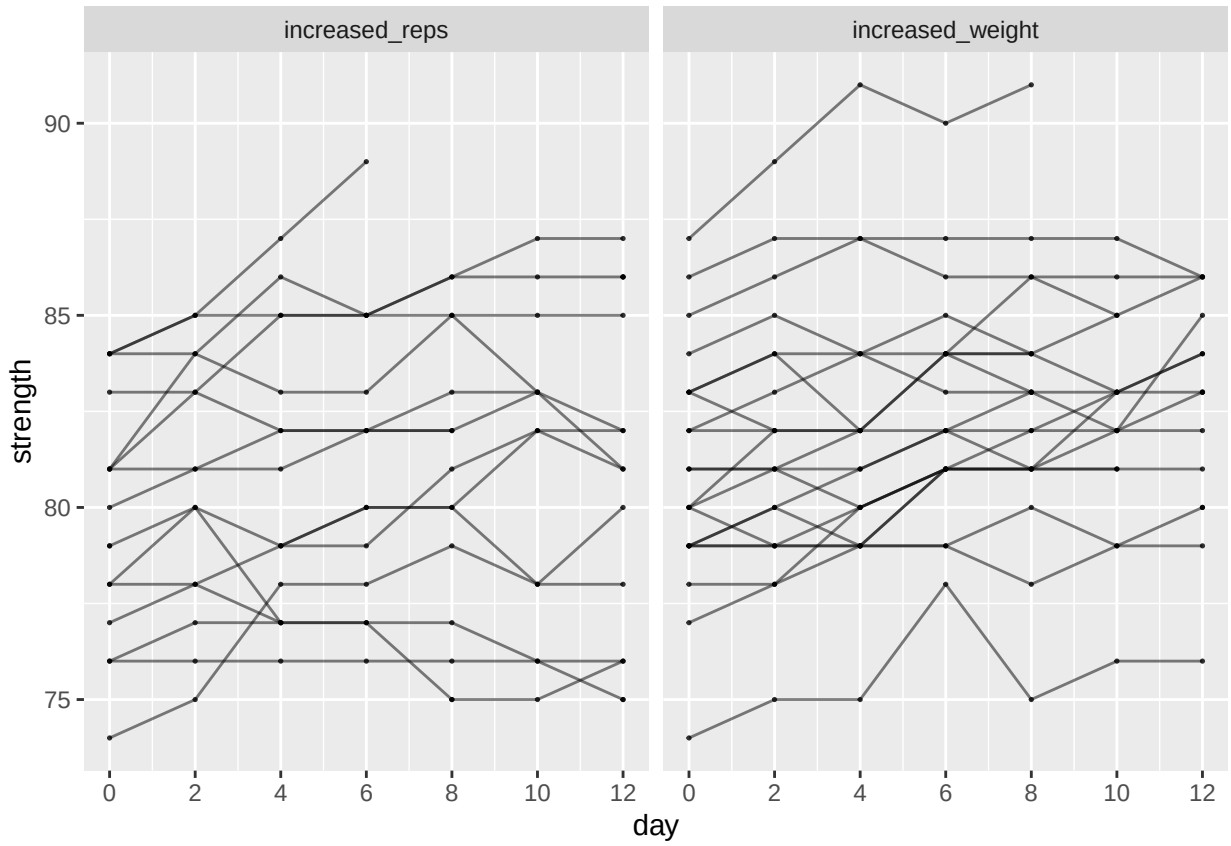
First I load the data. I'm also converting the treatment variable from numeric to factor.

```
strength = read_dta(here::here("hw", "homework_5", "exercise_therapy.dta"))

# convert treatment variable from numeric to factor
strength = strength %>%
  mutate(trt = factor(trt, levels = 1:2, labels = c("increased_reps", "increased_weight")))
```

1a. (10 pts) Construct spaghetti plots that display the strength versus time (in days) for each of the two treatment groups. Interpret trends that you see.

```
strength %>%
  ggplot(aes(time, y)) +
  geom_point(size = 0.25, alpha = 0.8) +
  geom_line(aes(group = id), alpha = 0.5) +
  scale_x_continuous(breaks=c(0,2, 4, 6, 8, 10, 12)) + # labels axes at study days
  labs(x = "day", y = "strength") +
  facet_wrap(~trt)
```



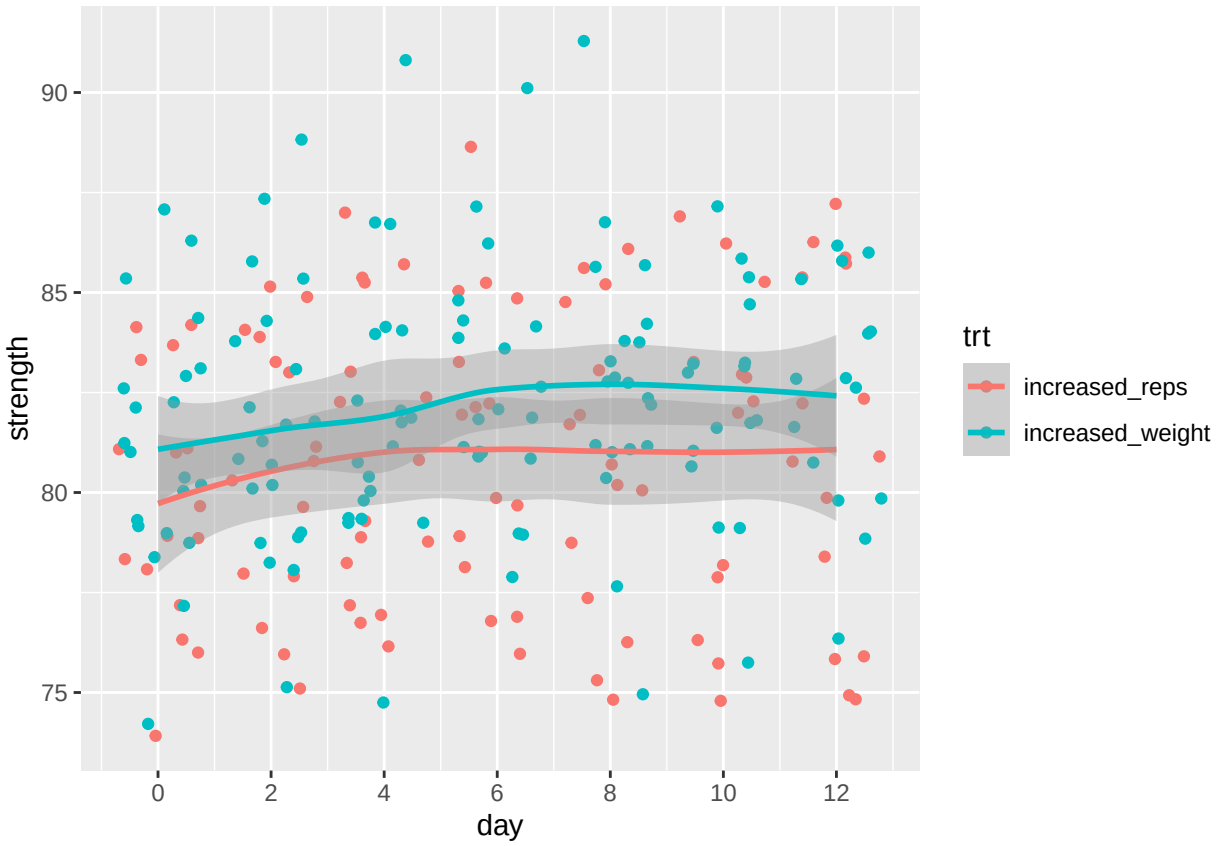
Some observations about the data:

- All subjects are measured at the same set of time points. A few subjects are missing observations, but overall the data is fairly balanced.
- There is quite a bit of variability across subjects at baseline in both treatment groups
- High variability within subjects as well
- Does strength increase over time in either group?
  - Lot's of heterogeneity across subjects in both groups
  - Some subjects increasing, others remaining relatively stable
  - Slight increase in the population overall

1b. (5 pts) Construct a plot of the smoothed means for the two groups (on one plot). Interpret trends that you see.

```
strength %>%
  ggplot(aes(time, y, group = trt, color = trt)) +
  geom_jitter() +
  geom_smooth() +
  scale_x_continuous(breaks=c(0,2, 4, 6, 8, 10, 12)) + # labels axes at study days
  labs(x = "day", y = "strength")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The mean in the increased weight group is slightly higher than the mean in the increased reps group at all time points. Both groups follow a similar trend, with a relatively modest increase in strength that seems to level off at day 6 or 8.

1c. (5 pts) From the plots do you see evidence that a random intercept should be included in a model? A random slope?

The plots indicates that random intercepts should be included in the model, as we can clearly see that the subjects are shifted up and downward by contant values from the mean strength values in each of the groups. However, there is not obvious evidence from the plots to support a random slope in the model, as the relationship between strength and time appears to be the same across subjects and we do not see a difference in variance at the different time points.

Exercise 2: Random Intercept Model (60 points)

2a. (10 pts) Write out the notation and assumptions for a model with a randomly varying intercept for subject. Include in your model (1) treatment, (2) a linear time trend, and (3) a treatment by linear time trend interaction.

The model is:

$$Y_{ij} = \beta_0 + \beta_1 trt_{ij} + \beta_2 time_{ij} + \beta_3 trt_{ij} \times time_{ij} + b_i + \epsilon_{ij}$$

And our assumptions are:

- $b_i \sim N(0, \sigma_b^2)$
- $\epsilon_{ij} \sim N(0, \sigma^2)$ are *iid* subject-specific measurement errors
- b_i and ϵ_{ij} are assumed to be independent of each other

2b.(5 pts) In this model, which are the fixed effects and which are the random effects?

The fixed effects in this model are for time, treatment, and the interaction between treatment and time. This model has one random effect (the random intercept that varies by subject).

2c. (10 pts) Fit this model in R. As output, provide table of fixed effects with estimate, Std. Error, and t value, and table of random effects variance estimates.

```
mod_randint = lmer(y ~ (1 | id) + trt + time + trt*time, data = strength)
```

The table of fixed effects is below:

```
as_tibble(summary(mod_randint)$coefficients, rownames = "term") %>%
  kable(digits = 2)
```

term	Estimate	Std. Error	t value
(Intercept)	80.11	0.84	95.50
trtincreased_weight	1.22	1.11	1.09
time	0.12	0.03	4.50
trtincreased_weight:time	0.03	0.04	0.93

The table of random effects variance estimates is below:

```
re_variance = as_tibble(summary(mod_randint)$varcor) %>%
  rename("variance" = vcov,
         "sd" = sdcor,
         group = grp,
         Name = var1)  %>%
  select(-var2)

re_variance %>%
  kable(digits = 3)
```

group	Name	variance	sd
id	(Intercept)	10.678	3.268
Residual	NA	1.212	1.101

2d. (10 pts) What is the estimated variance of the random intercepts? Calculate and interpret the ICC for this model.

- The variance of the random intercepts is $\sigma_b^2 = 10.68$.

We calculate the ICC by

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} =$$

0.898.

This is a large ICC value (close to 1).

- 89.8% of the variability in strength is attributable to differences among subjects relative to variation within subjects
- High ICC indicates it's important to account for repeated measurements within a subject!

2e. (5 pts) Test the significance of the fixed effects in the model.

A table of fixed effects with p-values is given below.

```
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
mod_randint = lmer(y ~ (1 | id) + trt + time + trt*time, data = strength)

as_tibble(summary(mod_randint)$coefficients, rownames = "term") %>%
  rename(p_value = "Pr(>|t|)") %>%
  mutate(p_value = format.pval(p_value)) %>%
  kable(digits = 2)
```

term	Estimate	Std. Error	df	t value	p_value
(Intercept)	80.11	0.84	37.51	95.50	< 2.22e-16
trtincreased_weight	1.22	1.11	37.52	1.09	0.28161
time	0.12	0.03	199.06	4.50	1.1305e-05
trtincreased_weight:time	0.03	0.04	199.16	0.93	0.35534

The fixed effect for time is statistically significant, but treatment and the interaction between time and treatment are not significant.

2f. (5 pts) Test the significance of the random intercept term.

```
## model without the random intercept
mod_lm = lm(y ~ time + trt + time*trt, data = strength)

## test for the significance of the random intercept
exactLRT(mod_randint, mod_lm)
```

```
## No restrictions on fixed effects. REML-based inference preferable.
```

```
## Using likelihood evaluated at REML estimators.
```

```
## Please refit model with method="ML" for exact results.
```

```
##
##  simulated finite sample distribution of LRT. (p-value based on 10000
##  simulated values)
##
## data:
## LRT = 381.39, p-value < 2.2e-16
```

- $H_0 : \sigma_b^2 = 0$
- $H_1 : \sigma_b^2 > 0$

We reject the null hypothesis that the variance of the random effect is 0, which indicates that the random intercept is statistically significant.

2g. (5 pts) Based on the analysis up to this point, interpret the effect of treatment on changes in strength. Does your analysis suggest a difference between the two groups?

The analysis suggests that there is no difference between the treatment groups and changes in strength. The coefficient for increasing weight versus increasing repitions was not found to be statistically significant. In addition the coefficient for the interaction between time and the increased weight program was also not found to be statistically significant.

Refitting a model without the interaction (see output below), we still do not see a significant association between strength and treatment.

```
library(lmerTest)
mod_randint = lmer(y ~ (1 | id) + trt + time, data = strength)

as_tibble(summary(mod_randint)$coefficients, rownames = "term") %>%
  rename(p_value = "Pr(>|t|)") %>%
  mutate(p_value = format.pval(p_value)) %>%
  kable(digits = 2)
```

term	Estimate	Std. Error	df	t value	p_value
(Intercept)	80.01	0.83	36.11	96.29	< 2.22e-16
trtincreased_weight	1.41	1.09	34.97	1.29	0.20629
time	0.14	0.02	200.17	7.65	8.2229e-13

Exercise 3: Random Slope Model (30 points)

3a. (15 pts) Refit a model that includes a random slope and main effects for time and treatment but no interaction term. Provide a table of fixed effects with p-values. Interpret the fixed effects.

```
mod_randslope = lmer(y ~ (1 + time | id) + trt + time, data = strength)

as_tibble(summary(mod_randslope)$coefficients, rownames = "term") %>%
  rename(p_value = "Pr(>|t|)") %>%
  mutate(p_value = format.pval(p_value)) %>%
  kable(digits = 2)
```

term	Estimate	Std. Error	df	t value	p_value
(Intercept)	80.09	0.80	35.11	100.06	< 2.22e-16
trtincreased_weight	1.21	1.06	35.01	1.14	0.26128
time	0.15	0.03	34.80	4.46	8.1737e-05

β_0: the average population-level strength at baseline for those in the increased reps treatment group is 80.1 β_1: (the treatment variable) the average difference in strength comparing those in the increased weight treatment group to those in the increased reps group is 1.211, controlling for day β_2: (the time variable) the average difference in strength for each one day difference in time, controlling for treatment

3b. (10 pts) What is the estimated variance of the random intercept in this model? What is the estimated variance of the random slope in this model? What is the correlation between the random intercept and slope in this model?

```
summary(mod_randslope)
```

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
```

```
 id       (Intercept) 9.9760   3.1585
          time        0.0334   0.1827   -0.03
 Residual             0.6517   0.8073
Number of obs: 238, groups:  id, 37
```

- The estimated variance of the random intercept is 9.976
- The estimated variance of the random slope is 0.0334
- The correlation between the random intercept and slope is -0.03

3c. (5 pts) Plot the random effects for both the random intercept and random slope models. Interpret.

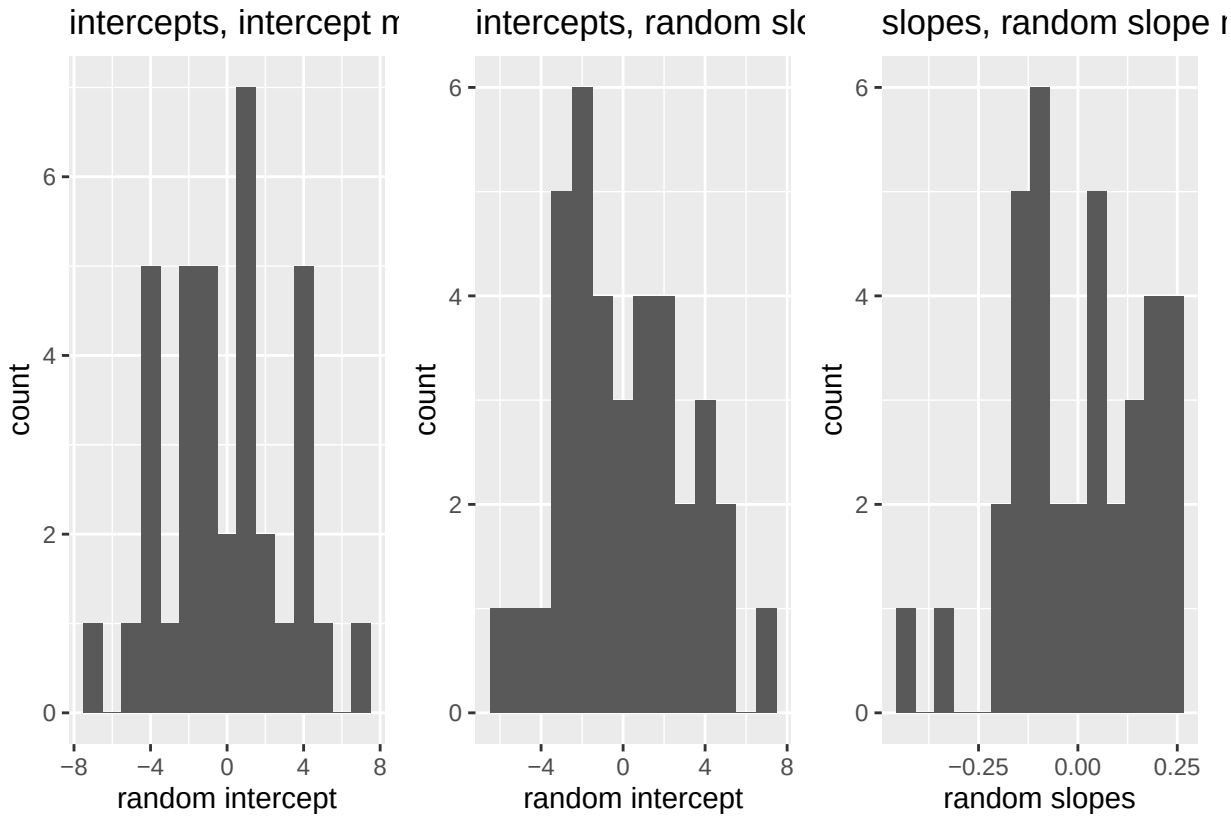
Below are histograms of the random intercepts and random slopes

```
p1 = tibble(intercept = ranef(mod_randint)$id[["(Intercept)"]]) %>%
ggplot(aes(x = intercept)) +
  geom_histogram(binwidth = 1) +
  labs(x = "random intercept", title = "intercepts, intercept model")

p2 = tibble(intercept = ranef(mod_randslope)$id[["(Intercept)"]]) %>%
ggplot(aes(x = intercept)) +
  geom_histogram(binwidth = 1) +
  labs(x = "random intercept", title = "intercepts, random slope model")

p3 = tibble(intercept = ranef(mod_randslope)$id[["time"]]) %>%
ggplot(aes(x = intercept)) +
  geom_histogram(bins = 15) +
  labs(x = "random slopes", title = "slopes, random slope model")

p1 + p2 + p3
```



Based on these plots, the random effects are distributed around zero and mostly normally distributed (though the random slopes appear to be slightly skewed). The values for the random slopes are quite small.

Below I also look at the fitted values from both models.

```
## random intercept model
int_mod = strength %>% filter(!is.na(y)) %>%
  mutate(fitted_values = fitted(mod_randint)) %>%
ggplot(aes(x = time, y = y, group = id)) +
  geom_point() + geom_path(alpha = .4) +
  labs(x = "Day", y = "Strength", title = "random intercept model fitted values") +
  stat_smooth(method = "lm", aes(group = NULL), se = FALSE, color = "red", size = 1.1) +
  geom_line(aes(y = fitted_values), color = "blue", alpha = .3)

## random slope model
slope_mod = strength %>% filter(!is.na(y)) %>%
  mutate(fitted_values = fitted(mod_randslope)) %>%
ggplot(aes(x = time, y = y, group = id)) +
  geom_point() + geom_path(alpha = .4) +
  labs(x = "Day", y = "Strength", title = "random slope model fitted values") +
  stat_smooth(method = "lm", aes(group = NULL), se = FALSE, color = "red", size = 1.1) +
  geom_line(aes(y = fitted_values), color = "blue", alpha = .3)

int_mod + slope_mod
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```