STAA 554 Homework 5

Contents

Libraries you may want:

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(lme4)
library(pbkrtest)
library(RLRsim)
library(lmerTest)
library(nlme)
```

Q1 Earthquakes

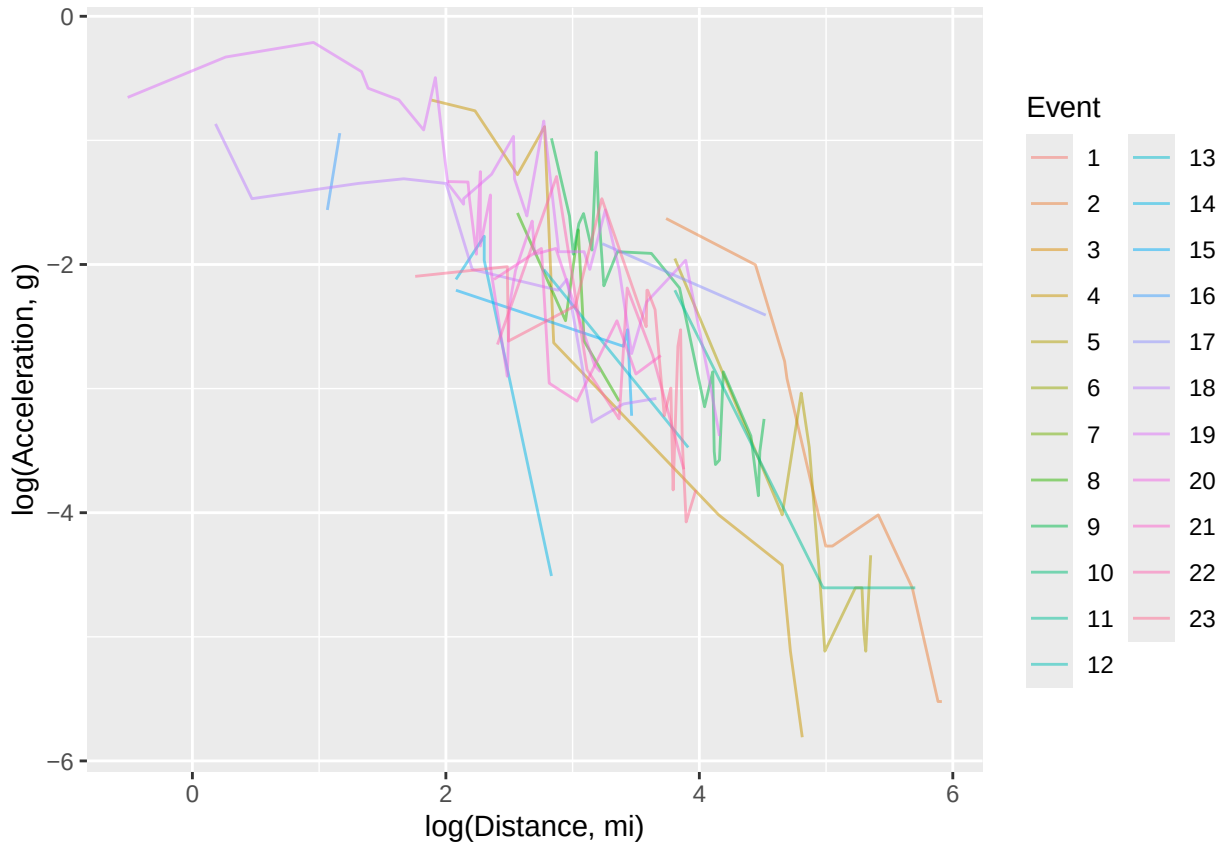
The attenu.csv data gives peak accelerations measured at various observation stations for 23 earthquakes in California. The data has been used by various workers to estimate the attenuating effect of distance on ground acceleration.

(a) 1pt Plot

Plot lines showing how the acceleration changes with distance for each quake. Make transformations of both axes so that the relationship is easier to see and replot.

```
attenu  <- read.csv("Data/attenu.csv")
attenu <- na.omit(attenu)
attenu2 <- attenu %>%
  mutate(log_accel = log(accel),  # natural log
         log_dist  = log(dist))

ggplot(attenu2, aes(x = log_dist, y = log_accel,
                    group = event, colour = factor(event))) +
  geom_line(alpha = .5) +
  labs(x = "log(Distance, mi)", y = "log(Acceleration, g)",
       colour = "Event")
```



(b) 2pt Mixed Model

Fit a mixed effects model with the transformed variables which takes account of both events and stations as random effects; include magnitude as a fixed effect. Express the effect of magnitude on the acceleration.

```
m1 <- lmer(log_accel ~ mag + log_dist + (1 | event) + (1 | station),
           data = attenu2, REML = FALSE)
summary(m1)$coef              # fixed effects table
```

```
            Estimate Std. Error      df  t value   Pr(>|t|)
(Intercept) -2.05872    0.60735  14.135  -3.3897 4.3497e-03
mag          0.43817    0.10802  17.244   4.0565 7.9956e-04
log_dist    -0.98109    0.05558 112.095 -17.6518 3.7727e-34
```

Each 1 point increase in magnitude increases log_acceleration by .438.

(c) 2pt Quadratic Term

Does adding a quadratic term in distance improve the model?

```
m2 <- update(m1, . ~ . + I(log_dist^2))
anova(m1, m2)   # LRT because both ML-fitted
```

```
Data: attenu2
Models:
m1: log_accel ~ mag + log_dist + (1 | event) + (1 | station)
m2: log_accel ~ mag + log_dist + (1 | event) + (1 | station) + I(log_dist^2)
   npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
m1    6 337 355   -162      325
m2    7 277 299   -131      263  61.7  1    3.9e-15
```

Yes, adding the quadratic form does improve our model significantly. this makes sense as we see a curved relationship between distance and acceleration in the plot on part a. To model this non-linearity, we do need a quadratic term.

(d) 2pt Station variance component

Can we remove the station variation term? *(hint: attenu2 <-na.omit(attenu) may be useful if you use REML methods for comparisons, recall methods based on likelihoods require the same data for all models)*

```
m_full <- lmer(log_accel ~ mag + log_dist + I(log_dist^2) +
                 (1 | event) + (1 | station),
               data = attenu2,
               REML = TRUE)

m_nostat <- lmer(log_accel ~ mag + log_dist + I(log_dist^2) +
                   (1 | event),
                 data = attenu2,
```

```
                 REML = TRUE)


anova(m_nostat, m_full)
```

```
Data: attenu2
Models:
m_nostat: log_accel ~ mag + log_dist + I(log_dist^2) + (1 | event)
m_full: log_accel ~ mag + log_dist + I(log_dist^2) + (1 | event) + (1 | station)
         npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
m_nostat    6 277 296   -133      265
m_full      7 277 299   -131      263  2.47  1       0.12
```

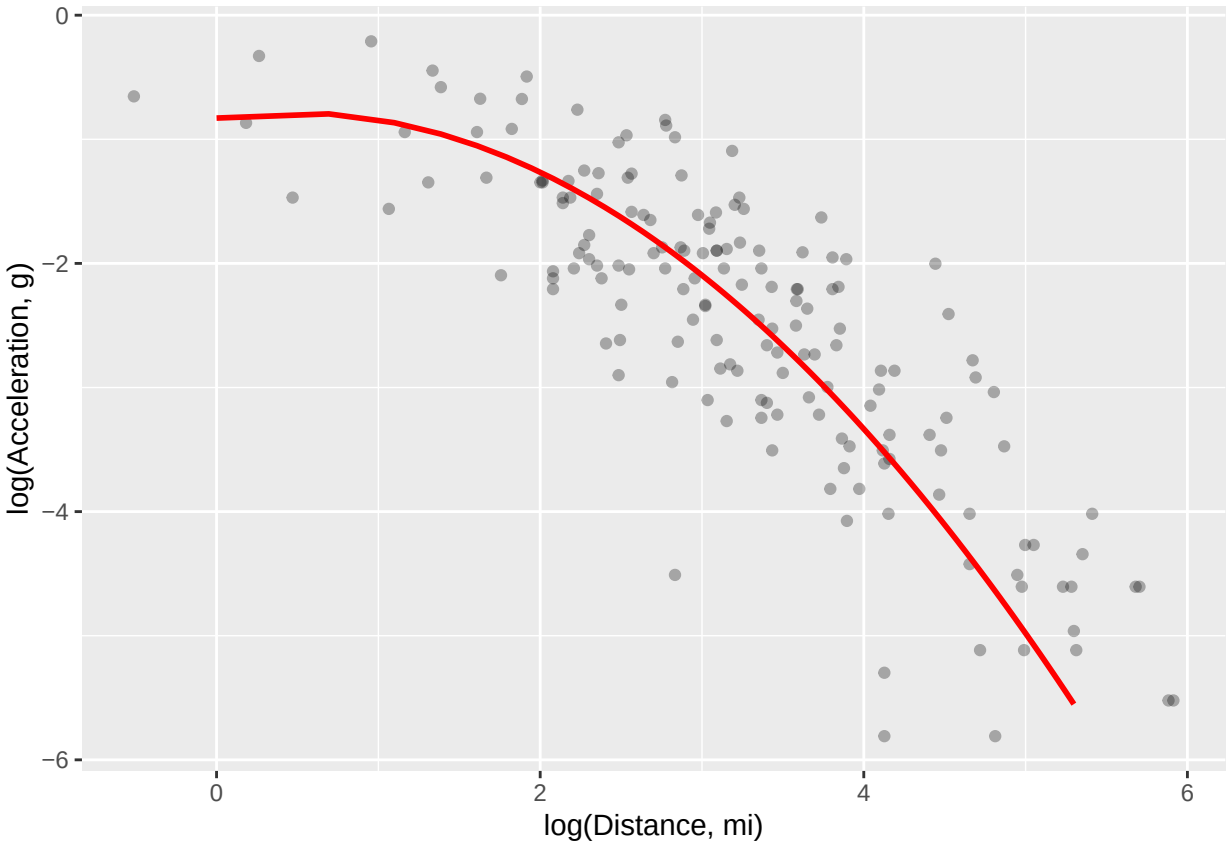
We can remove the station variance term because the full model is not signifficantly better (p = .12) than the reduced model.

(e) 2pts Prediction

Using the model with a quadratic term and random effects for both station and event, as in part c: For a new magnitude 6 quake, predict the acceleration for up to a distance of 200 miles. Make a plot of the data and show your predicted curve on top of the data in a different color.

```
newdata <- data.frame(
  mag       = 6,
  log_dist  = log(seq(1, 200, by = 1))
)
newdata$pred <- predict(m2, newdata, re.form = NA)   # fixed-effects only

ggplot(attenu2, aes(log_dist, log_accel)) +
  geom_point(alpha = .3) +
  geom_line(data = newdata, aes(log_dist, pred), colour = "red", size = 1) +
  labs(y = "log(Acceleration, g)", x = "log(Distance, mi)")
```

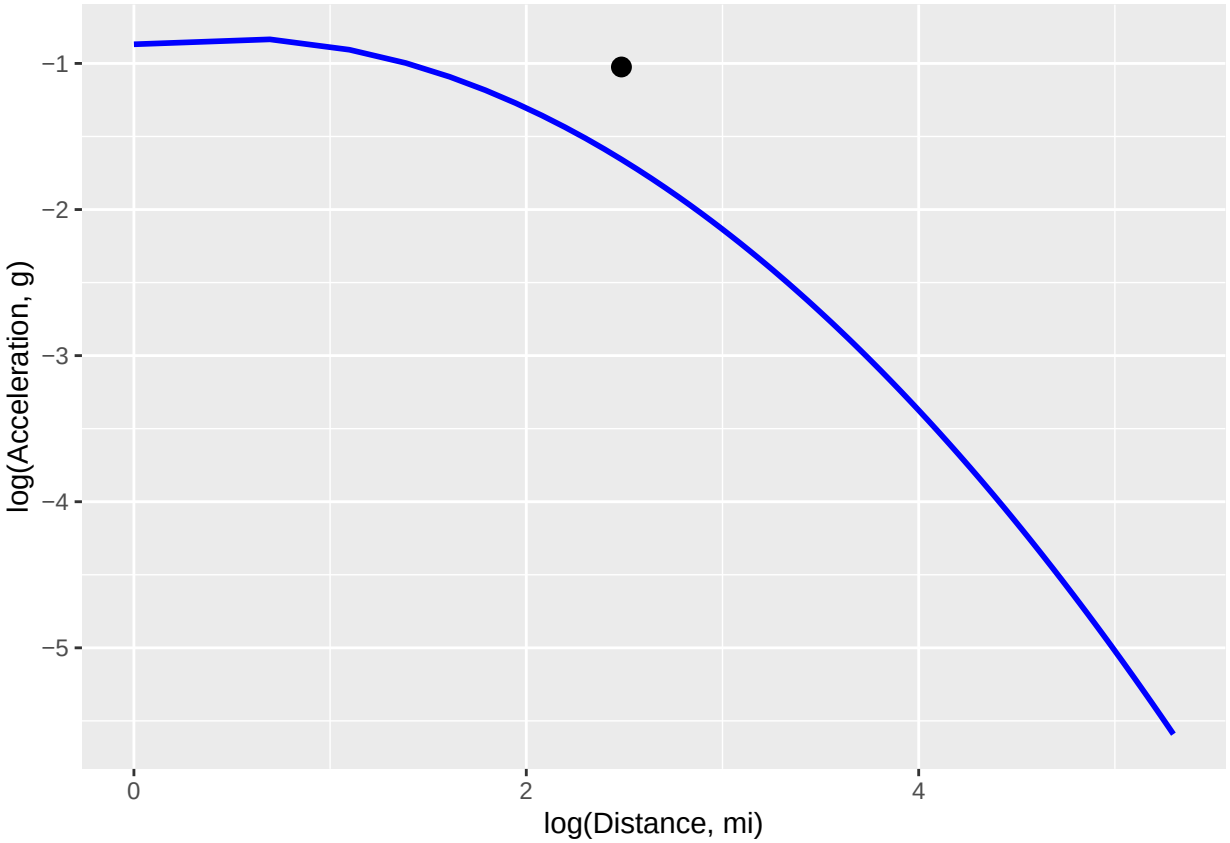


(f) 3pts Prediction for observation 1

Predict how the acceleration varied for the first event where only one observation was available. Show the predicted acceleration up to 200 miles in a plot. Add the actual observation to the plot. (just point predictions, no intervals needed)

```
event1_re   <- ranef(m2)$event[1, ]
station_re0 <- 0
newdata$pred_evt1 <- with(newdata,
  pred + event1_re + station_re0)

ggplot() +
  geom_line(data = newdata, aes(log_dist, pred_evt1),
            colour = "blue", size = 1) +
  geom_point(data = attenu2 %>% filter(event == unique(event)[1]),
             aes(log_dist, log_accel), colour = "black", size = 3) +
  labs(y = "log(Acceleration, g)", x = "log(Distance, mi)")
```



Q2 Rat Drink Data

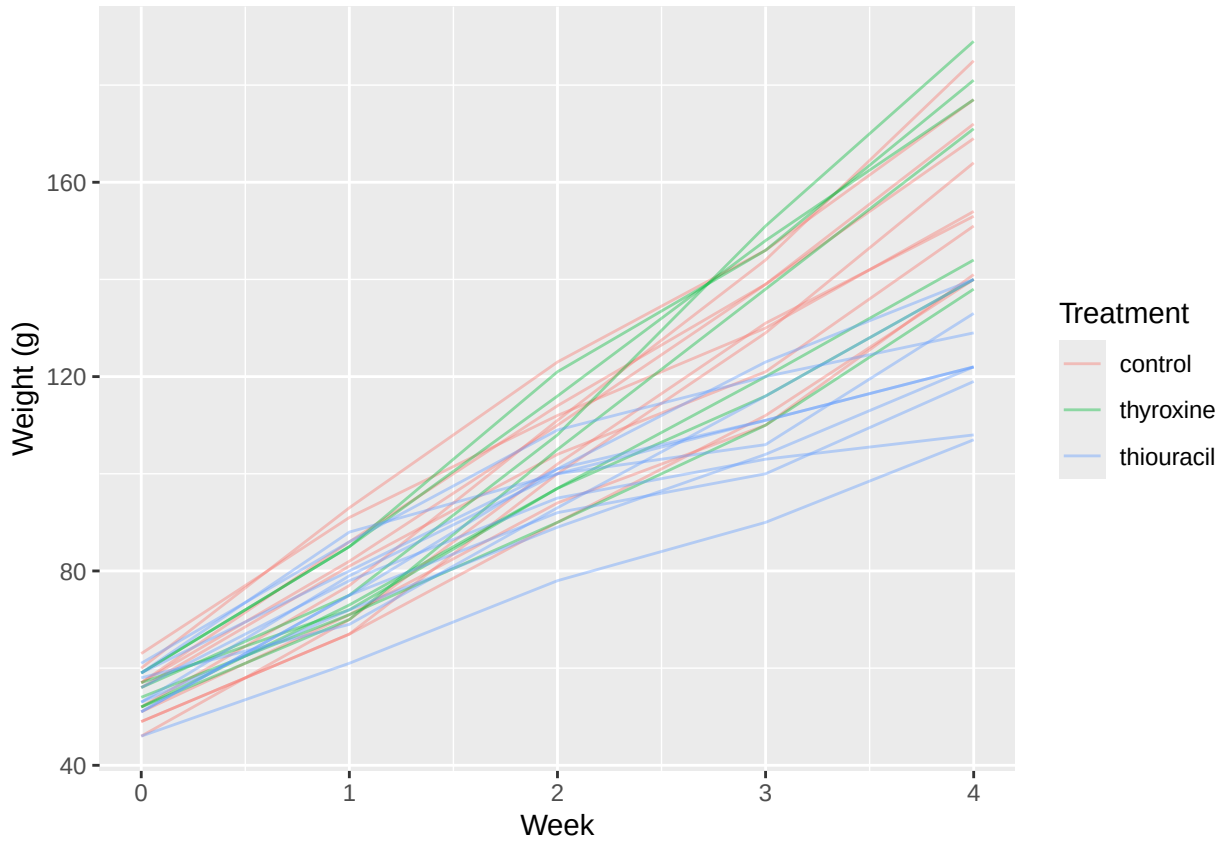
The ratdrink.csv data consist of five weekly measurements of body weight for 27 rats. The first 10 rats are on a control treatment while 7 rats have thyroxine added to their drinking water. Ten rats have thiouracil added to their water.

```
rats <- read.csv("Data/ratdrink.csv") %>%      # wt, weeks, subject, treat
  mutate(
    subject = factor(subject),
    treat   = factor(treat,        # put control first for clear contrasts
                     levels = c("control","thyroxine","thiouracil"))
  )
```

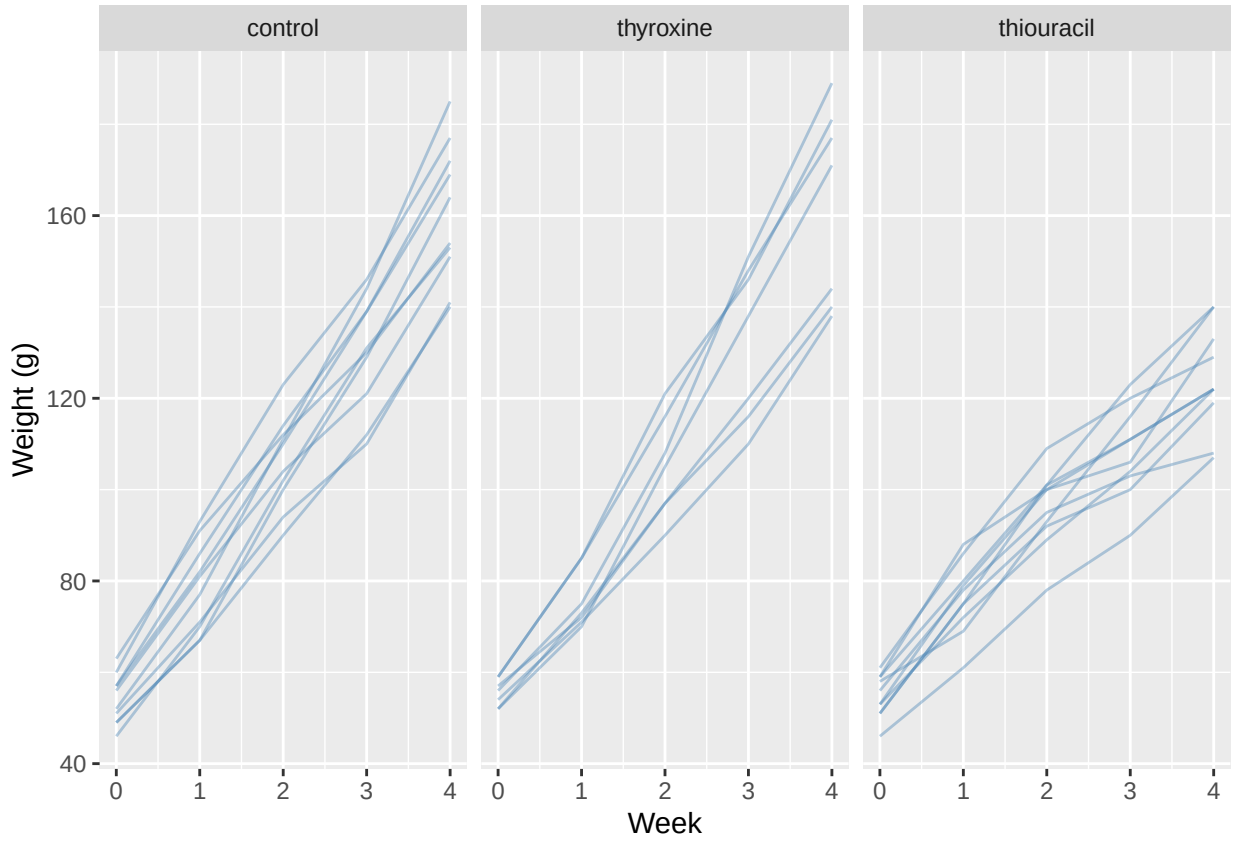
(a) 1pt Plots

Plot the data showing how weight increases with age on a single panel, taking care to distinguish the three treatment groups. Now create a three-panel plot, one for each group. Discuss what can be seen.

```
ggplot(rats, aes(weeks, wt, colour = treat, group = subject)) +
  geom_line(alpha = 0.4) +
  labs(x = "Week", y = "Weight (g)", colour = "Treatment")
```




```
ggplot(rats, aes(weeks, wt, group = subject)) +
  geom_line(alpha = 0.4, colour = "steelblue") +
  facet_wrap(~ treat, nrow = 1) +
  labs(x = "Week", y = "Weight (g)")
```



All three treatment groups display roughly linear weight gain. Control and Thyroxine groups rise in parallel; Thiouracil rats gain weight more slowly which may indicate a treatment, week interaction.

(b) 3pt Mixed Model Interpretation

Fit a linear longitudinal model that allows for a random slope and intercept for each rat. Each group should have a different mean line. Give interpretation for the following estimates:

i. The fixed effect intercept term.
ii. The interaction between thiouracil and week.
iii. The intercept random effect SD.

```
m1 <- lmer(wt ~ treat*weeks + (weeks | subject), data = rats, REML = FALSE)
summary(m1)$coef[ ,1:2]    # estimate & SE for brevity
```

```
                    Estimate Std. Error
(Intercept)          52.88000     1.9740
treatthyroxine       -0.79429     3.0762
treatthiouracil       4.78000     2.7916
weeks                26.48000     1.1937
treatthyroxine:weeks  0.66286     1.8602
treatthiouracil:weeks -9.37000     1.6881
```

i

Intercept = 52.9 g = mean week-0 weight for a control rat. ### ii Thiouracil × week = −9.37 g/week thiouracil suppresses weekly gain by =9 g relative to control ### iii Random-intercept SD - 5.8 g indicates baseline weights vary ±5.8 g across rats.

(c) 2pt Test

Check whether there is a statistically significant treatment effect.

```
m0 <- update(m1, . ~ weeks + (weeks | subject))
anova(m0, m1)
```

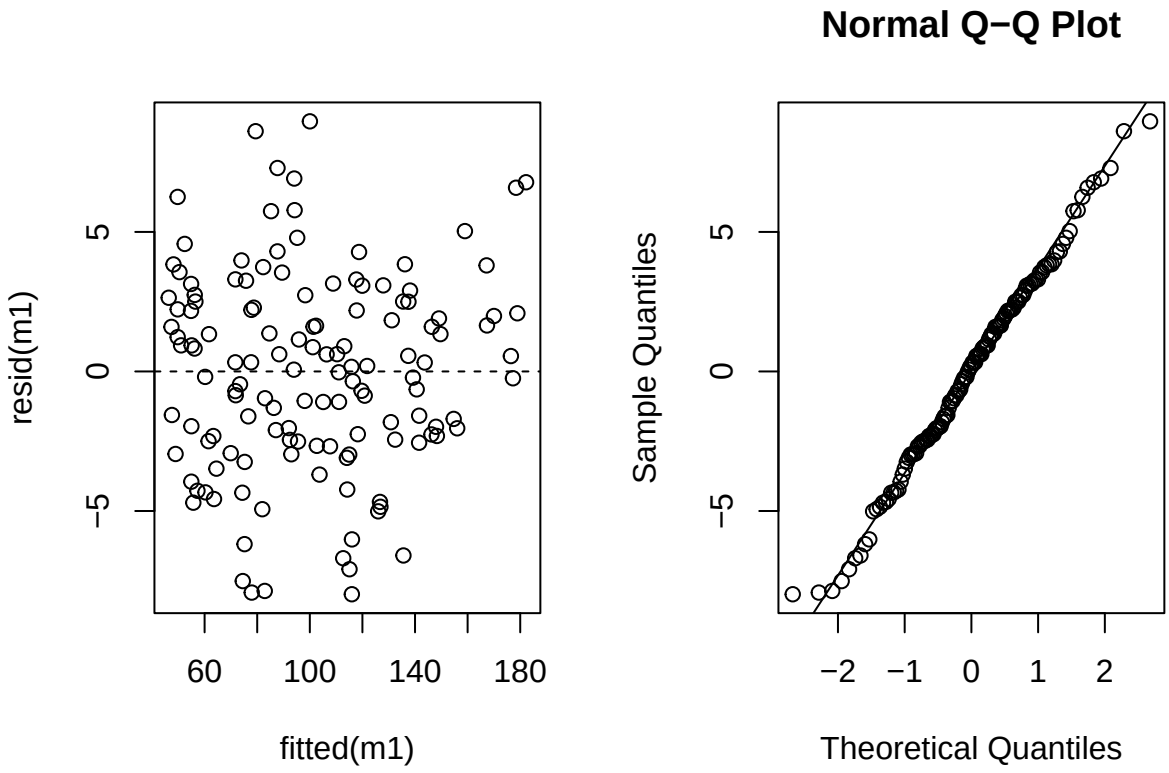
```
Data: rats
Models:
m0: wt ~ weeks + (weeks | subject)
m1: wt ~ treat * weeks + (weeks | subject)
   npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
m0    6 933 950   -460      921
m1   10 916 945   -448      896  25.1  4    4.7e-05
```

The new model is signifficantly better which indicates that growth grajectories do differ by treatment. This aligns with what we saw in the plots for part a.

(d) 2pt Diagnostic

Construct diagnostic plots showing the residuals against the fitted values and a QQ plot of the residuals. Interpret.

```
par(mfrow = c(1,2))
plot(fitted(m1), resid(m1));  abline(h = 0, lty = 2)
qqnorm(resid(m1)); qqline(resid(m1))
```



Residual vs fitted shows no fanning so we dont have concerns about inequal varaince. The QQ plot also resembles a straight line indicating that our Normality of residuals assumption remains strong.

(e) 2pt Confidence Intervals

Construct confidence intervals for the parameters of the model.

i. Which random effect terms may not be significant?
ii. Is the thyroxine group significantly different from the control group?

```
confint(m1, oldNames = FALSE)
```

```
                                2.5 %   97.5 %
sd_(Intercept)|subject          3.45063  7.66548
cor_weeks.(Intercept)|subject  -0.52610  0.37941
sd_weeks|subject                2.60641  4.86878
sigma                           3.75556  5.11423
(Intercept)                    48.86929 56.89071
treatthyroxine                 -7.04453  5.45596
treatthiouracil                -0.89201 10.45201
weeks                          24.05474 28.90526
treatthyroxine:weeks           -3.11663  4.44235
treatthiouracil:weeks         -12.79983 -5.94017
```

i

The slope–intercept correlation CI covers 0 meaning there is little evidence of a systematic relationship between baseline weight and growth rate ### ii Thyroxine main and interaction CIs both include 0, so Thyroxine does not differ from Control.

(f) 2pt Covariance structure

Fit this same model from (b) using lme() and extract the marginal covariance matrix for observations from a particular rat. Describe the observed structure and comment on if it makes sense in this context.

```
m1_lme <- lme(wt ~ treat*weeks, random = ~ weeks | subject, data = rats)
getVarCov(m1_lme, individual = "1")
```

```
Random effects variance covariance matrix
            (Intercept)   weeks
(Intercept)     32.4950 -2.8447
weeks           -2.8447 14.1390
  Standard Deviations: 5.7004 3.7602
```

Positive variances on the diagonal and a small negative covariance between intercept and slope imply slightly slower growth in heavier-starting rats—biologically plausible “catch-up” growth.

(g) 3pt Covariance Structure

Compare the covariance matrix from (f) to the covariance matrix in each of the following models:

i. fit the same model in (b), but without the random slope and assume a compound symmetric matrix. Does the compound symmetric assumption make sense in this context?

ii. fit the same model in (b), but without the random slope and assume an unstructured covariance matrix. Describe any general trends in the structure.

iii. fit the same model in (b), but without the random slope and assume an autoregressive 1 structure in the covariance matrix. Why might we consider this structure in this context?

```
# have tio fix issue of week 0 for corSymm to work
rats <- read.csv("Data/ratdrink.csv") %>%
  arrange(subject, weeks) %>%
  group_by(subject) %>%
  mutate(weekID = row_number()) %>%
  ungroup()
```

i.

```
m_cs <- lme(wt ~ treat*weeks, random = ~ 1 | subject,
            correlation = corCompSymm(form = ~ weeks | subject),
            data = rats)

m_cs
```

```
Linear mixed-effects model fit by REML
  Data: rats
  Log-restricted-likelihood: -474.22
  Fixed: wt ~ treat * weeks
         (Intercept)       treatthiouracil        treatthyroxine
            52.88000               4.78000              -0.79429
               weeks treatthiouracil:weeks  treatthyroxine:weeks
            26.48000              -9.37000               0.66286

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:      8.4384    7.157

Correlation Structure: Compound symmetry
 Formula: ~weeks | subject
 Parameter estimate(s):
Rho
  0
Number of Observations: 135
Number of Groups: 27
```

The Compound symetric forces equal correlations at all lags meaning this structure is too rigid for data where week-1 vs week-2 measurements correlate more than week-1 vs week-5

ii.

```
# Had to increase max iterations so the model could converge
ctrl <- lmeControl(msMaxIter   = 200,
                   maxIter     = 200,
                   pnlsMaxIter = 50)

m_un <- lme(wt ~ treat*weeks,
            random      = ~ 1 | subject,
            correlation = corSymm(form = ~ weekID | subject),
            weights     = varIdent(form = ~ 1 | weekID),
            data        = rats,
            method      = "REML",
            control     = ctrl)

m_un
```

```
Linear mixed-effects model fit by REML
  Data: rats
  Log-restricted-likelihood: -406.06
  Fixed: wt ~ treat * weeks
         (Intercept)       treatthiouracil        treatthyroxine
             55.5034              -1.8577                2.9438
               weeks treatthiouracil:weeks  treatthyroxine:weeks
             26.1390              -7.4680               -1.4272
```

```
Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:      3.9406   2.9917

Correlation Structure: General
 Formula: ~weekID | subject
 Parameter estimate(s):
 Correlation:
  1     2     3     4
2 0.976
3 0.724 0.823
4 0.385 0.447 0.792
5 0.063 0.168 0.603 0.899
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | weekID
 Parameter estimates:
     1      2      3      4      5
1.0000 2.8543 3.1589 4.1645 5.2618
Number of Observations: 135
Number of Groups: 27
```

Unstructured allows each variance and covariance to differ; captures rising variance and declining correlations but at the cost of 22 parameters. we see that correlations are stronger betwen closer weeks and weaker between weaks that are farther apart.

iii.

```
m_ar1 <- lme(wt ~ treat*weeks, random = ~ 1 | subject,
             correlation = corAR1(form = ~ weeks | subject),
             data = rats)

m_ar1
```

```
Linear mixed-effects model fit by REML
  Data: rats
  Log-restricted-likelihood: -441.12
  Fixed: wt ~ treat * weeks
         (Intercept)       treatthiouracil        treatthyroxine
            53.56221               1.92169               1.09097
               weeks treatthiouracil:weeks  treatthyroxine:weeks
            26.64143              -9.27979               0.19621

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:   0.0016792   11.498

Correlation Structure: AR(1)
 Formula: ~weeks | subject
```

```
 Parameter estimate(s):
   Phi
0.8467
Number of Observations: 135
Number of Groups: 27
```

AR1 does a good job in matching the correlations into lags for equally spaced weeks.

(h) 2pt Compare using information criteria

Compare the 4 models from f and g using AIC and BIC. Which model appears best?

```
AIC(m1_lme, m_cs, m_un, m_ar1)
```

```
       df    AIC
m1_lme 10 898.68
m_cs    9 966.45
m_un   22 856.11
m_ar1   9 900.24
```

```
BIC(m1_lme, m_cs, m_un, m_ar1)
```

```
       df    BIC
m1_lme 10 927.27
m_cs    9 992.19
m_un   22 919.03
m_ar1   9 925.97
```

In this case, the flexibility of the unstructured approach outweighs its complexity cost as seeing with the lowest AIC so we will proceed with that.

Q3

The National Youth Survey collected a sample of 11–17 year olds, 117 boys and 120 girls, asking questions about marijuana usage. The data is presented in potuse.csv.

Potuse levels: 1: “non-user” 2: “light” 3: “Heavy”

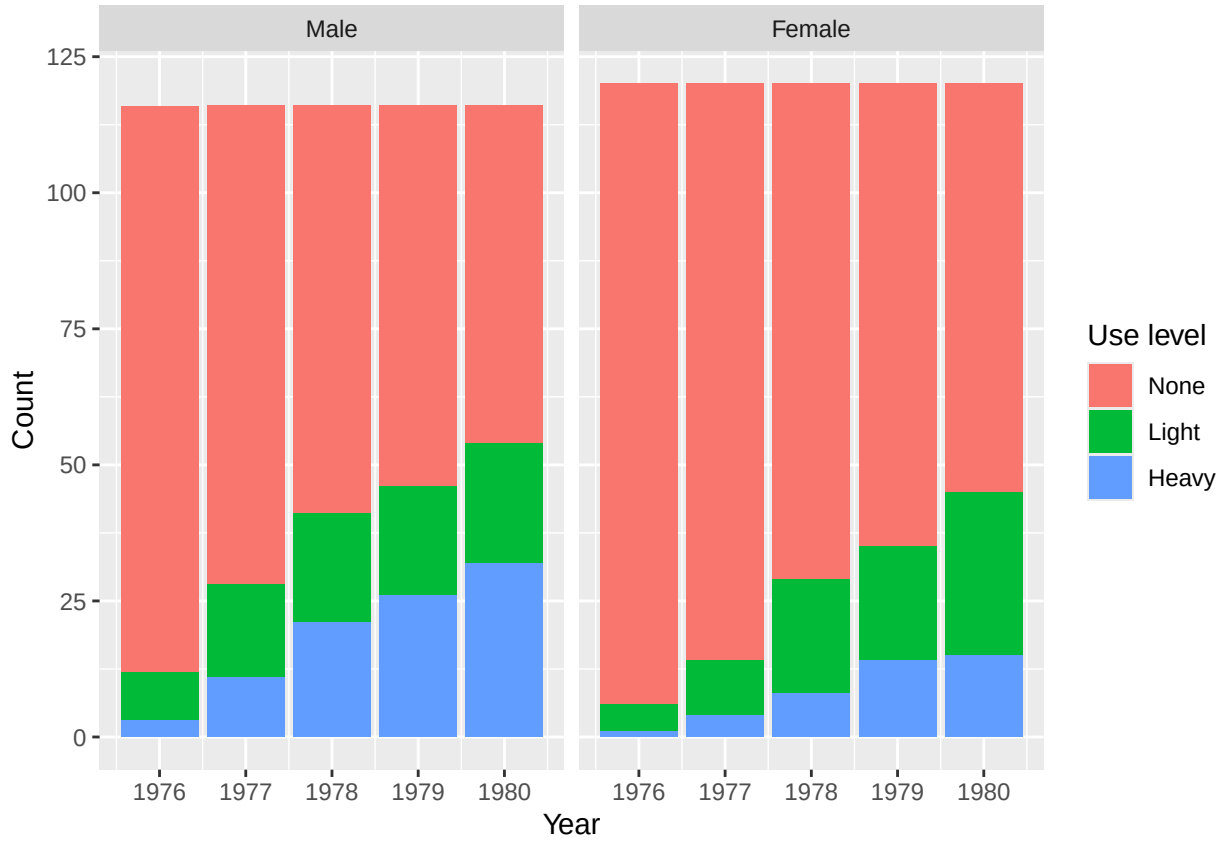
Sex: 1: male. 2: female

```
pot <- read.csv("Data/potuse.csv") %>%
  pivot_longer(cols   = starts_with("year."),
               names_to  = "year",
               values_to = "use_lvl") %>%
  mutate(
    year_num = as.integer(sub("year\\.", "19", year)),  # 1976-1980
    sex      = factor(sex, labels = c("Male","Female")),
    use_lvl  = factor(use_lvl, levels = 1:3,
                      labels = c("None","Light","Heavy"))
  )
```

(a) 1pt Plot

Plot the total number of people falling into each usage category as it varies over time separately for each sex.

```
ggplot(pot, aes(year_num, count, fill = use_lvl)) +
  geom_col(position = "stack") +
  facet_wrap(~ sex) +
  labs(x = "Year", y = "Count", fill = "Use level")
```



(b) 2pt Format and fit model

Condense the levels of the response into whether the person did or did not use marijuana that year. Turn the year into a numerical variable. Fit a GLMM for the now binary response with an interaction between sex and year as a predictor using Gauss-Hermite quadrature. Comment on the effect of sex.

Hint: The idea of this problem is to fit a GLMM modeling a binary response (i.e. , will a person with these characteristics be likely to use pot or not) to the data. The tricky part is that the potuse.csv data comes in the form of count data, not individual data, so you have to 'wrangle' it until it represents individuals. Consider the following functions:

```
tidyr::gather() to make the data long instead of wide (i.e., to start to turn the years into a numeric v
tidyr::pivot_longer(), newer than gather(), possibly more intuitive?
dplyr::uncount() function to turn the count data into individual data... very cool little function.
```

```r
pot_wide <- read.csv("Data/potuse.csv") %>%
  mutate(patternID = row_number())

pot_expanded <- pot_wide %>%
  uncount(count) %>%
  mutate(id = row_number())

pot_long <- pot_expanded %>%
  pivot_longer(
    cols      = starts_with("year."),
    names_to  = "year",
    values_to = "use_lvl"
  ) %>%
  mutate(
    year_num = as.integer(sub("year\\.", "19", year)),
    sex      = factor(sex, labels = c("Male", "Female")),
    use_bin  = (use_lvl != 1)
  )

m_full <- glmer(
  use_bin ~ sex * scale(year_num) + (1 | id),
  data   = pot_long,
  family = binomial,
  nAGQ   = 10
)

summary(m_full)$coef
```

```
                         Estimate Std. Error  z value   Pr(>|z|)
(Intercept)              -1.79898    0.33290 -5.40406 6.5150e-08
sexFemale                -1.12529    0.45988 -2.44692 1.4408e-02
scale(year_num)           1.22269    0.15936  7.67259 1.6856e-14
sexFemale:scale(year_num) 0.22034    0.22939  0.96053 3.3679e-01
```

(c) 2pt Test

Fit a reduced model without sex and use it to test for the significance of sex in the larger model.

```r
m_full <- glmer(
  use_bin ~ sex * scale(year_num) + (1 | id),
  data   = pot_long,
  family = binomial,
  nAGQ   = 10
)

m_nosex <- update(m_full, . ~ scale(year_num) + (1 | id))

lrt_c <- anova(m_nosex, m_full)
print(lrt_c)
```

```
Data: pot_long
```

```
Models:
m_nosex: use_bin ~ scale(year_num) + (1 | id)
m_full: use_bin ~ sex * scale(year_num) + (1 | id)
        npar  AIC  BIC logLik deviance Chisq Df Pr(>Chisq)
m_nosex    3 1007 1022   -500     1001
m_full     5 1004 1030   -497      994  6.36  2      0.042
```

The model with sex is significantly better than the model without it (p = .042). This is clear as the deviance increases when we remove sex.

(d) 3pt Linearity

Fit a model with year as a factor. (For simplicity, No sex term in the model.)

i. Should this model be preferred to the model with year as just a linear term?

ii. Interpret the estimated effects of the year in the factor version of the model.

```r
m_yearFac <- glmer(
  use_bin ~ factor(year_num) + (1 | id),
  data   = pot_long,
  family = binomial,
  nAGQ   = 10
)

lrt_d   <- anova(m_nosex, m_yearFac)
aic_d   <- AIC(m_nosex, m_yearFac)
print(lrt_d)
```

```
Data: pot_long
Models:
m_nosex: use_bin ~ scale(year_num) + (1 | id)
m_yearFac: use_bin ~ factor(year_num) + (1 | id)
          npar  AIC  BIC logLik deviance Chisq Df Pr(>Chisq)
m_nosex      3 1007 1022   -500     1001
m_yearFac    6 1001 1031   -494      989  11.9  3     0.0076
```

```r
print(aic_d)
```

```
          df    AIC
m_nosex    3 1006.8
m_yearFac  6 1000.9
```

i

The linear trend is inadequate to capture year to year differences.; retain the factor model to capture a year-linear pattern.

ii

Use climbed steeply through 1978, plateaued in 1979, then fell in 1980—mirroring national survey data of the late-1970s “peak” in teen marijuana use. Sex differences remain constant across years because the interaction is non-significant.