Problem Set 8

matt krasnow (the goat)

Due Friday, December 6, 2024 at 11:59pm

Problem set policies. *Please provide concise, clear answers for each question while making sure to fully explain your reasoning. For problems asking for calculations, show your work in addition to clearly indicating the final answer. For problems involving* `R`*, be sure to include the code and output in your solution.*

Please submit the PDF of your knit solutions to Gradescope and be sure to assign which pages of your solution correspond to each problem. Make sure that the PDF is fully readable to the graders; e.g., make sure that lines don't run off the page margin.

We encourage you to discuss problems with other students (and, of course, with the teaching team), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution. Be aware that simply copying answers found online, whether human-generated or machine-generated, is a violation of the Honor Code.

Question 1

Consider the following model:

$$Y_{ij} = \beta_0 + \alpha_j + \beta_1 x_{ij} + \varepsilon_{ij}$$
$$i = 1, \ldots, n_j$$
$$j = 1, \ldots, J$$
$$\alpha_j \sim N(0, \sigma_\alpha^2) \perp\!\!\!\perp \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

a) Compute $\text{Var}(Y_{ij})$.

We know that β_0 and x_{ij} are deterministic and thus constant => 0 variance

$\text{Var}(Y_{ij}) = \text{Var}(\alpha_j) + \text{Var}(\varepsilon_{ij})$

Thus

$$Var(Y_{ij}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$$

This is possible because of liinearity of variance. in this case, there is no covariance between alpha and epsilon because they are independent normals.

b) Compute $\text{Cov}(Y_{ij}, Y_{i'j'})$ where $i' \neq i$ and $j' \neq j$.

$\alpha_j and \alpha_{j'}$ are independent for $j! = j'$

$$Y_{ij} = \beta_0 + \alpha_j + \beta_1 x_{ij} + \varepsilon_{ij}$$
$$Y_{i'j'} = \beta_0 + \alpha_{j'} + \beta_1 x_{i'j'} + \varepsilon_{i'j'}$$

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{i'j'}) &= \text{Cov}(\beta_0 + \alpha_j + \beta_1 x_{ij} + \varepsilon_{ij}, \beta_0 + \alpha_{j'} + \beta_1 x_{i'j'} + \varepsilon_{i'j'}) \\ &= \text{Cov}(\alpha_j, \alpha_{j'}) + \text{Cov}(\alpha_j, \varepsilon_{i'j'}) + \text{Cov}(\varepsilon_{ij}, \alpha_{j'}) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) \\ &= 0 + 0 + 0 + 0\end{aligned}$$

$$\boxed{\text{Cov}(Y_{ij}, Y_{i'j'}) = 0}$$

c) Compute $\text{Cov}(Y_{ij}, Y_{i'j})$ where $i' \neq i$.

$$Y_{ij} = \beta_0 + \alpha_j + \beta_1 x_{ij} + \varepsilon_{ij}$$
$$Y_{i'j} = \beta_0 + \alpha_j + \beta_1 x_{i'j} + \varepsilon_{i'j}$$

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{i'j}) &= \text{Cov}(\beta_0 + \alpha_j + \beta_1 x_{ij} + \varepsilon_{ij}, \beta_0 + \alpha_j + \beta_1 x_{i'j} + \varepsilon_{i'j}) \\ &= \text{Cov}(\alpha_j, \alpha_j) + \text{Cov}(\alpha_j, \varepsilon_{i'j}) + \text{Cov}(\varepsilon_{ij}, \alpha_j) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j})\end{aligned}$$

- $\text{Cov}(\alpha_j, \alpha_j) = Var(\alpha_j) = \sigma_\alpha^2$

- $\text{Cov}(\alpha_j, \varepsilon_{i'j}) = 0$ (by independence assumption)
- $\text{Cov}(\varepsilon_{ij}, \alpha_j) = 0$ (by independence assumption)
- $\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j}) = 0$ (different observations within group)

Thus:

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{i'j}) &= \sigma_\alpha^2 + 0 + 0 + 0 \\ &= \sigma_\alpha^2\end{aligned}$$

$$\boxed{\text{Cov}(Y_{ij}, Y_{i'j}) = \sigma_\alpha^2}$$

observations within the same group have covariance equal to the variance of the random effect (σ_α^2). This correlation is because observations within the same group share the same random effect α_j, while all other components remain independent.

d) The ratio of the answer to c) and the answer to a) is called the **intraclass correlation coefficient**. Explain in a sentence or two what this measures.

The intraclass correlation coefficient (ICC) is:

$$\begin{aligned}\text{ICC} &= \frac{\text{Cov}(Y_{ij}, Y_{i'j})}{\text{Var}(Y_{ij})} \\ &= \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}\end{aligned}$$

The ICC measures the proportion of total variance in the response that is attributed to *between-group variation*. It is the correlation between any two observations within the same group, indicating how much of the total variation in the data is explained by the grouping structure.

Question 2

The dataset `growth.csv` contains a certain type of growth data on a sample of 11 girls and 16 boys. Specifically, in this study, the distance between the center of the pituitary to the pteryomaxillary fissure (a teardrop-shaped opening located in the human skull) was measured at four occasions: at 8, 10, 12 and 14 years. The variables in the dataset are:

`subjid`: a unique subject identifier

`sex`: the sex of each child

`distance_8`: the distance (in mm) at age 8

`distance_10`: the distance (in mm) at age 10

`distance_12`: the distance (in mm) at age 12

`distance_14`: the distance (in mm) at age 14

a) The data are currently in "wide" format. Wrangle the dataset into "long" format that is suitable for analysis with, for example, the `lmer()` function. That is, there should be a column for age, and another for `distance`.

```
library(tidyr)
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
```

```
library(readr)

library(ggplot2)

library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x nlme::collapse() masks dplyr::collapse()
## x dplyr::filter()  masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Read the data
growth_data <- read_csv("data/growth.csv")
```

```
## Rows: 27 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): subjid, sex
## dbl (4): distance_8, distance_10, distance_12, distance_14
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

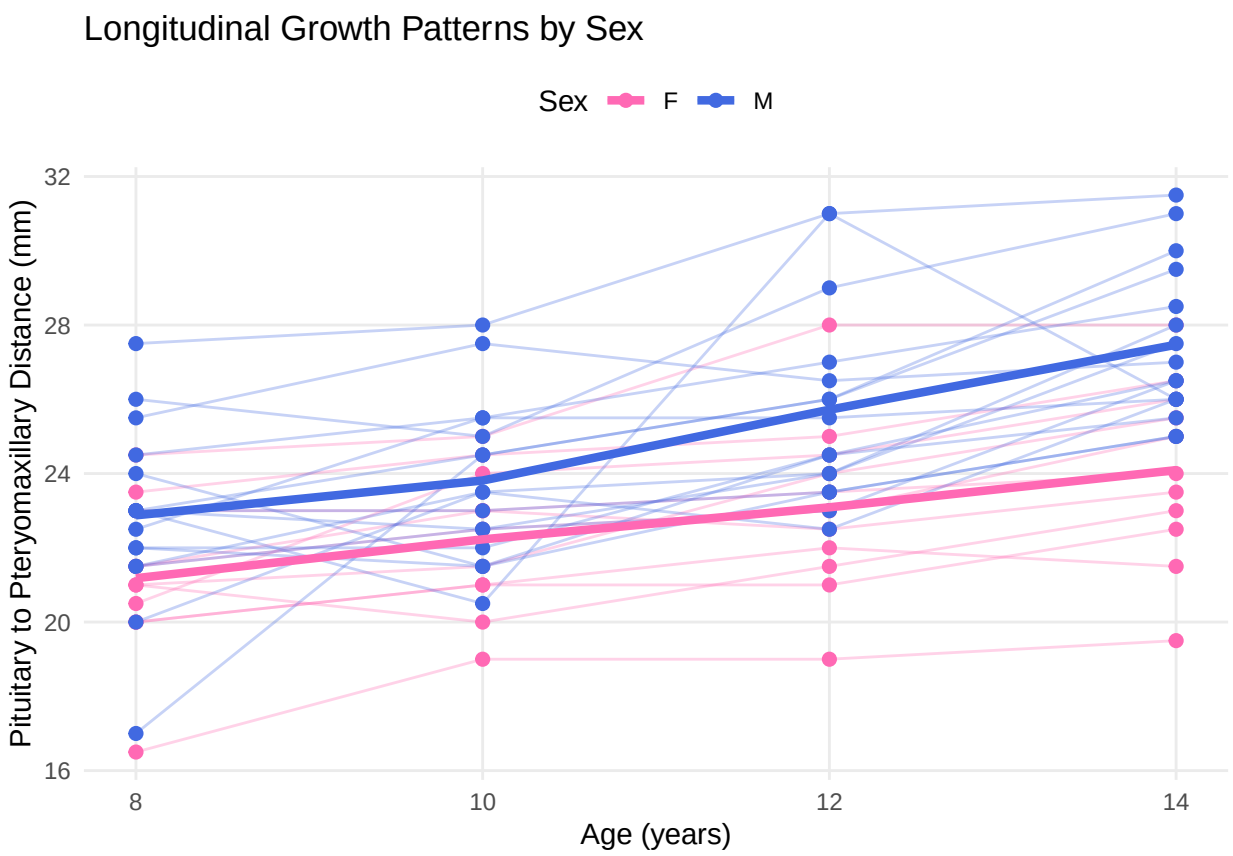
```
#  wide to long
growth_long <- growth_data %>%
  pivot_longer(
    cols = starts_with("distance_"),
    names_to = "age",
    values_to = "distance",
    # get numeric age from column names
    names_prefix = "distance_"
  ) %>%
  # change age to numeric
  mutate(age = as.numeric(age)) %>%
  # Arrange by subject ID and age
  arrange(subjid, age)


head(growth_long)
```

```
## # A tibble: 6 x 4
##   subjid sex     age distance
##   <chr>  <chr> <dbl>    <dbl>
## 1 F001   F         8     21
## 2 F001   F        10     20
## 3 F001   F        12     21.5
## 4 F001   F        14     23
## 5 F002   F         8     21
## 6 F002   F        10     21.5
```

b) Plot these data longitudinally, using different colored series for males and females. Comment on any strange observations you observe.

```
# GORGEOUS!!@!@@!


ggplot(growth_long, aes(x = age, y = distance, group = subjid)) +
  #map color here
  geom_line(aes(color = sex), alpha = 0.3) +
  geom_point(aes(color = sex), size = 2) +
  # mean traj
  stat_summary(aes(group = sex, color = sex),
               fun = mean,
               geom = "line",
               linewidth = 1.5) +
```

```
  scale_color_manual(values = c("F" = "#FF69B4", "M" = "#4169E1")) +
  #x
  scale_x_continuous(breaks = c(8, 10, 12, 14)) +
  labs(
    title = "Longitudinal Growth Patterns by Sex",
    x = "Age (years)",
    y = "Pituitary to Pteryomaxillary Distance (mm)",
    color = "Sex"
  ) +
  theme_minimal() +
  theme(
    legend.position = "top",
    panel.grid.minor = element_blank()
  )
```



Observations: - there is a catch up period for male distance around age 10 where the slope chanegs - similarly, but less drastically, there is a slope change for females around age 10 - there are strange behaviors where thesome mailes experience greatsdden and sharp deceases in distance. non monotonic patterns

c) Fit three mixed models that control for age, sex, and their interaction. The first model should incorporate a random intercept for age, the second model should have independent random intercepts and slopes for age, and the third model should have a random intercept and slope for age and allow for correlation between them. Pick which you think is the most appropriate model for these data using AIC or BIC.

```
growth_long <- growth_long %>%
  mutate(age_centered = age - mean(age))

# Model 1: Random Intercept only
```

```r
m1 <- lme(distance ~ age_centered * sex,
          random = ~ 1 | subjid,
          data = growth_long,
          method = "ML")  # Using ML for AIC/BIC comparison

# Model 2: Independent Random Intercept and Slope
m2 <- lme(distance ~ age_centered * sex,
          random = list(subjid = pdDiag(~ 1 + age_centered)),
          data = growth_long,
          method = "ML")

# Model 3: Correlated Random Intercept and Slope
m3 <- lme(distance ~ age_centered * sex,
          random = ~ age_centered | subjid,
          data = growth_long,
          method = "ML")

# Compare models using AIC and BIC
models_comparison <- data.frame(
  Model = c("Random Intercept",
            "Independent Random Intercept & Slope",
            "Correlated Random Intercept & Slope"),
  AIC = c(AIC(m1), AIC(m2), AIC(m3)),
  BIC = c(BIC(m1), BIC(m2), BIC(m3))
)

# View comparison
print(models_comparison)
```

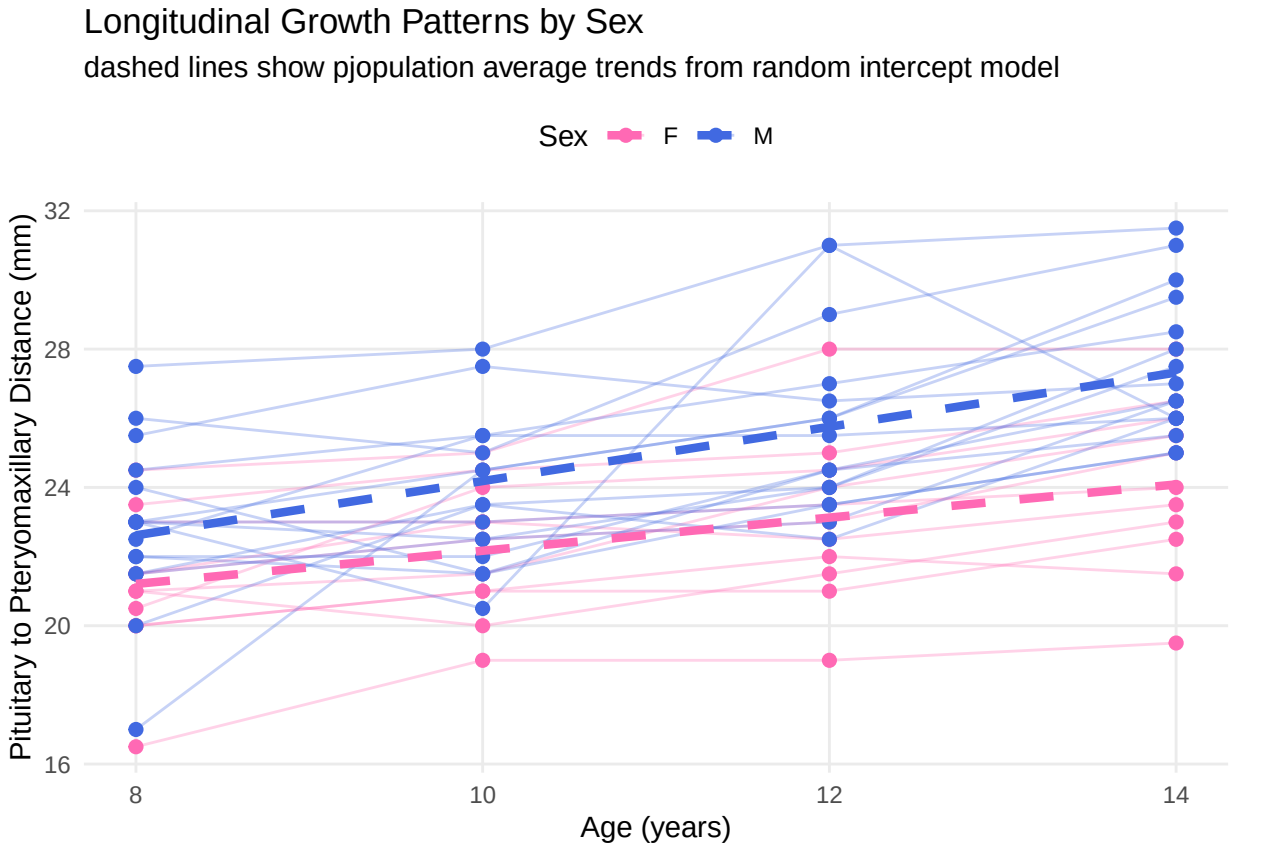
```
##                                   Model      AIC      BIC
## 1                     Random Intercept 440.6391 456.7318
## 2 Independent Random Intercept & Slope 442.0878 460.8627
## 3  Correlated Random Intercept & Slope 443.8060 465.2630
```

INterpretation: - Random intercept has the lowest AIC (best, most parsimonous) - increased complexity from 2 and 3 does not give enough ofa areason to use them - The correlation between random effects (Model 3) adds complexity without substantial benefit - Model 1 (Random Intercept only) provides the best balance between model fit and complexity

d) Use the model you selected from part b) to add the population average trends to your plot in part b) for both boys and girls separately.

```r
mean_age <- mean(growth_long$age)

new_data <- expand.grid(
  age_centered = seq(min(growth_long$age_centered),
                     max(growth_long$age_centered),
                     length.out = 100),
  sex = c("F", "M")
) %>%
  mutate(age = age_centered + mean_age)

new_data$predicted <- predict(m1, newdata = new_data, level = 0)

ggplot() +
```

```
  geom_line(data = growth_long,
            aes(x = age, y = distance, group = subjid, color = sex),
            alpha = 0.3) +
  geom_point(data = growth_long,
            aes(x = age, y = distance, color = sex),
            size = 2) +
  geom_line(data = new_data,
            aes(x = age, y = predicted, color = sex),
            linewidth = 1.5,
            linetype = "dashed") +
  scale_color_manual(values = c("F" = "#FF69B4", "M" = "#4169E1")) +
  scale_x_continuous(breaks = c(8, 10, 12, 14)) +
  labs(
    title = "Longitudinal Growth Patterns by Sex",
    subtitle = "dashed lines show pjopulation average trends from random intercept model",
    x = "Age (years)",
    y = "Pituitary to Pteryomaxillary Distance (mm)",
    color = "Sex"
  ) +
  theme_minimal() +
  theme(
    legend.position = "top",
    panel.grid.minor = element_blank()
  )
```



e) Formally test whether the distance between the pituitary to the pteryomaxillary fissure differs between boys and girls at birth (hint: you might want to use the `lmerTest` package).

```r
growth_long <- growth_long %>%
  mutate(age_birth = age - 0)  # age centered at birth

# Fit model with age centered at birth using nlme
m1_birth <- lme(distance ~ age_birth * sex,
                random = ~ 1 | subjid,
                data = growth_long,
                method = "REML")  # Use REML for inference

# Extract summary with test statistics
summary(m1_birth)
```

```
## Linear mixed-effects model fit by REML
##   Data: growth_long
##        AIC      BIC    logLik
##   445.7572 461.6236 -216.8786
##
## Random effects:
##  Formula: ~1 | subjid
##         (Intercept) Residual
## StdDev:    1.816214 1.386382
##
## Fixed effects:  distance ~ age_birth * sex
##                    Value Std.Error DF   t-value p-value
## (Intercept)    17.372727 1.1835071 79 14.679023  0.0000
## age_birth       0.479545 0.0934698 79  5.130483  0.0000
## sexM           -1.032102 1.5374208 25 -0.671321  0.5082
## age_birth:sexM  0.304830 0.1214209 79  2.510520  0.0141
##  Correlation:
##                (Intr) ag_brt sexM
## age_birth      -0.869
## sexM           -0.770  0.669
## age_birth:sexM  0.669 -0.770 -0.869
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.59804400 -0.45461690  0.01578365  0.50244658  3.68620792
##
## Number of Observations: 108
## Number of Groups: 27
```

```r
# Get confidence intervals
intervals(m1_birth, which = "fixed")
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                     lower       est.      upper
## (Intercept)    15.0170153 17.3727273 19.7284392
## age_birth       0.2934984  0.4795455  0.6655925
## sexM           -4.1984797 -1.0321023  2.1342751
## age_birth:sexM  0.0631473  0.3048295  0.5465118
```

- The estimated difference between males and females at birth is -1.03mm
- This difference is not statistically significant ($p = 0.5082$)

- We cannot conclude there is a sex difference at birth

f) Formally test whether the change in distance over time is different for boys versus girls.

- Boys' distance increases about 0.30 mm/year faster than girls
- This difference is statistically significant ($p = 0.0141$)

There is strong evidence ($p = 0.0141$) that the change in pituitary to pteryomaxillary fissure distance over time differs between boys and girls, with boys showing a significantly faster rate of increase than girls.

Problem 3

You've been asked to determine how player performance, measured via batting average, progresses as major league baseball players age. The belief is that players peak at different ages, and that the mean peak is around 30 years of age.

The dataset `mlb_batting_data.csv` contains year-by-year batting records for all positional players with at least one at-bat (an opportunity to record a hit) from 1980 until 2021. A player's batting average (BA) is the proportion of at-bats in which he records a hit (BA = H/AB) will be the response variable, and a player's age in year (Age) will be the predictor. A mixed model to predict BA from AGE will be used to model the career arc of each player.

Let Y be batting average, and let X be a player's age. Then a reasonable mixed model for these data could be written as such:

$$\begin{aligned} Y_{i,j} &\sim N\left(\beta_{0,j} + \beta_{1,j}X_{i,j} + \beta_{2,j}X_{i,j}^2, \sigma_y^2/n_{i,j}\right) \\ \beta_{0,j} &\sim N\left(\mu_0, \sigma_0^2\right) \\ \beta_{1,j} &\sim N\left(\mu_1, \sigma_1^2\right) \\ \beta_{2,j} &\sim N\left(\mu_2, \sigma_2^2\right) \end{aligned}$$

for the i^{th} measurement for the j^{th} player. For example the Red Sox's David Ortiz was 39 years old in 2015, his 19^{th} year in the league, and had a batting average of 0.273 in 528 at-bats. Thus his measurements were $Y_{19,j'} = 0.273, X_{19,j'} = 39,$ and $n_{19,j'} = 528$ (Ortiz is the j'^{th} player in the database... the exact value for j' is not easy to determine and not really important).

(a) Determine how many unique players are in the data set. Create a histogram of the variable `Age`. Comment on what you notice.

```
mlb_data <- read_csv("data/mlbdata.csv")
```

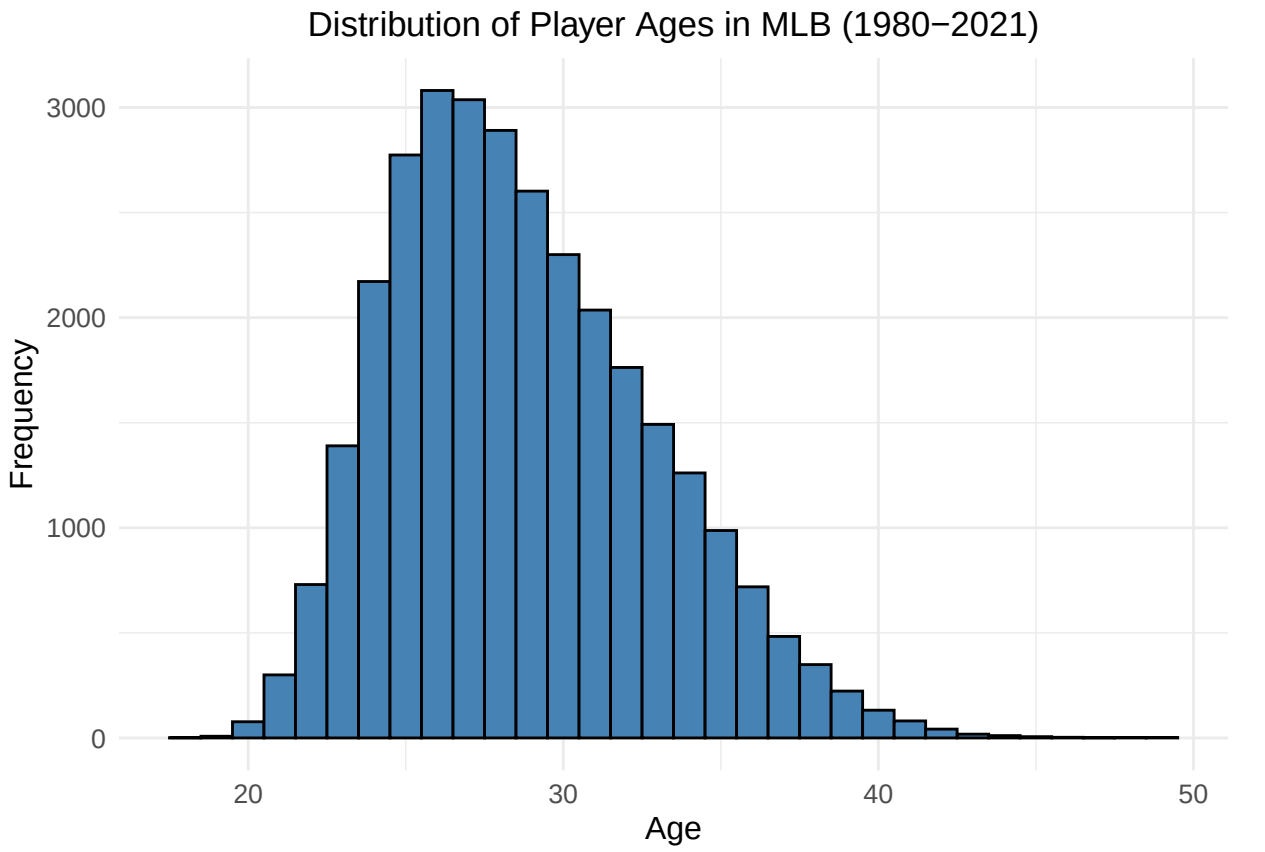
```
## Rows: 30974 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): playerID, nameFirst, nameLast, teamID, lgID
## dbl (5): yearID, Age, AB, H, BA
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# (a) Number of unique players
n_players <- length(unique(mlb_data$playerID))
cat("Number of unique players:", n_players, "\n")
```

```
## Number of unique players: 4858
```

```
# Create histogram of Age
ggplot(mlb_data, aes(x = Age)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black") +
  labs(
    title = "Distribution of Player Ages in MLB (1980-2021)",
    x = "Age",
    y = "Frequency"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_text(size = 10),
```

```
    axis.title = element_text(size = 12)
  )
```



there are 4858 players in the league

Observations: - strong right skew - age centered around mid to late twenties - non normal distribution

(b) Create two plots: (i) the histogram of years played by individuals, and (ii) the scatterplot of career batting average vs. years played for each individual. Briefly comment on what you notice.

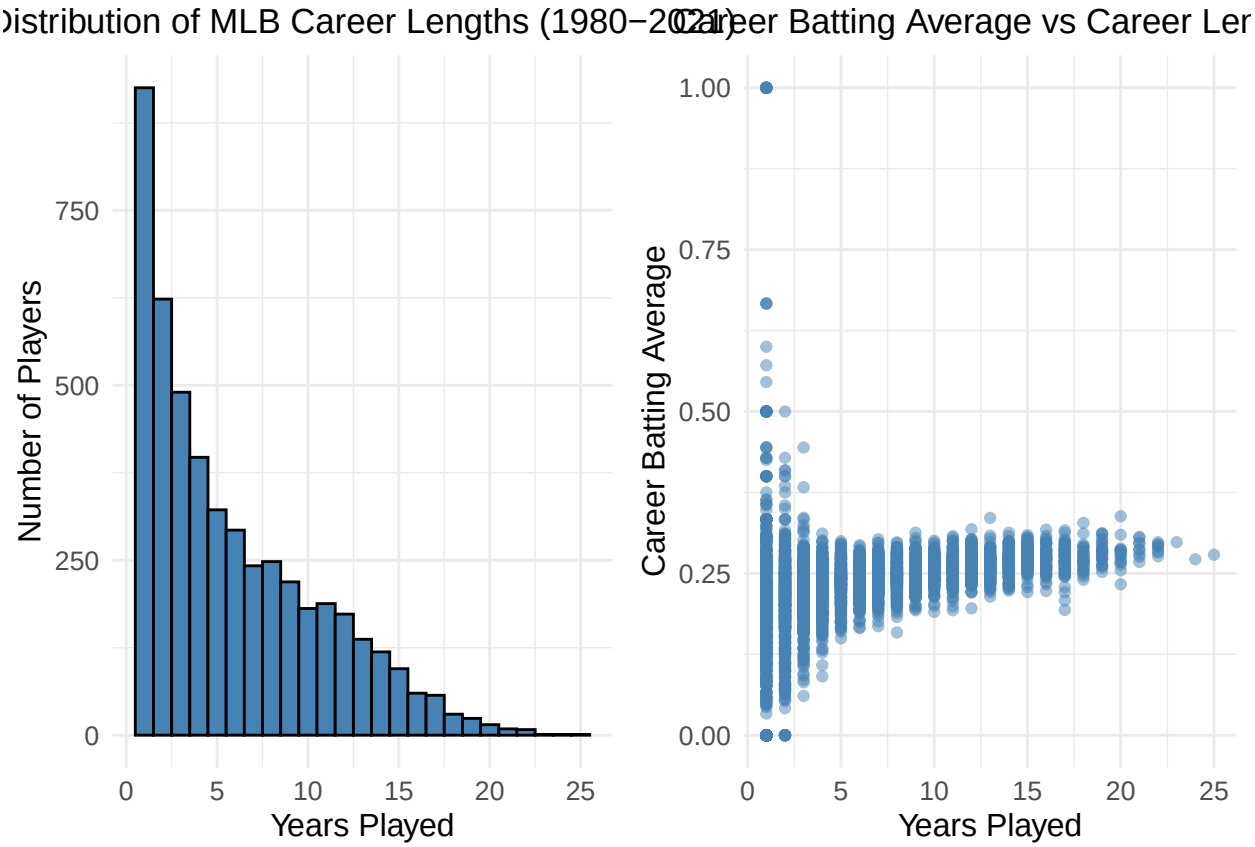
```
career_stats <- mlb_data %>%
  group_by(playerID) %>%
  summarise(
    years_played = n_distinct(yearID),
    career_ba = sum(H) / sum(AB)
  )

# years played
p1 <- ggplot(career_stats, aes(x = years_played)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black") +
  labs(
    title = "Distribution of MLB Career Lengths (1980-2021)",
    x = "Years Played",
    y = "Number of Players"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_text(size = 10),
```

```
    axis.title = element_text(size = 12)
  )

#batting average vs years played
p2 <- ggplot(career_stats, aes(x = years_played, y = career_ba)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  labs(
    title = "Career Batting Average vs Career Length",
    x = "Years Played",
    y = "Career Batting Average"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12)
  )

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(p1, p2, ncol = 2)
```



- Most players have very short careers (1-3 years)
- There's a rapid decline in frequency as career length increases
- Very few players make it past 15-20 years in MLB
- There appears to be a positive relationship between career length and batting average
- Players with longer careers (>10 years) tend to maintain batting averages above ~.225
- The spread of batting averages narrows as career length increases
- potential "survivor bias" - players with better batting averages tend to have longer careers
- outliers appear at very high batting averages (>.500), possible erorr

(c) Fit a quadratic regression model using the `lm` function in R to predict `BA` from `Age` and `Age`^{2}, using the argument `weight=AB` to account for the fact that there is more information/certainty in estimating the true batting average for players when they have more at-bats (which also mimics the stated variance σ_y^2 in the probabilistic model statement above).

```
mlb_data$BA <- mlb_data$H / mlb_data$AB


mlb_data$Age_squared <- mlb_data$Age^2


quad_model <- lm(BA ~ Age + Age_squared,
                 data = mlb_data,
                 weights = AB)
```

```r
summary(quad_model)
```

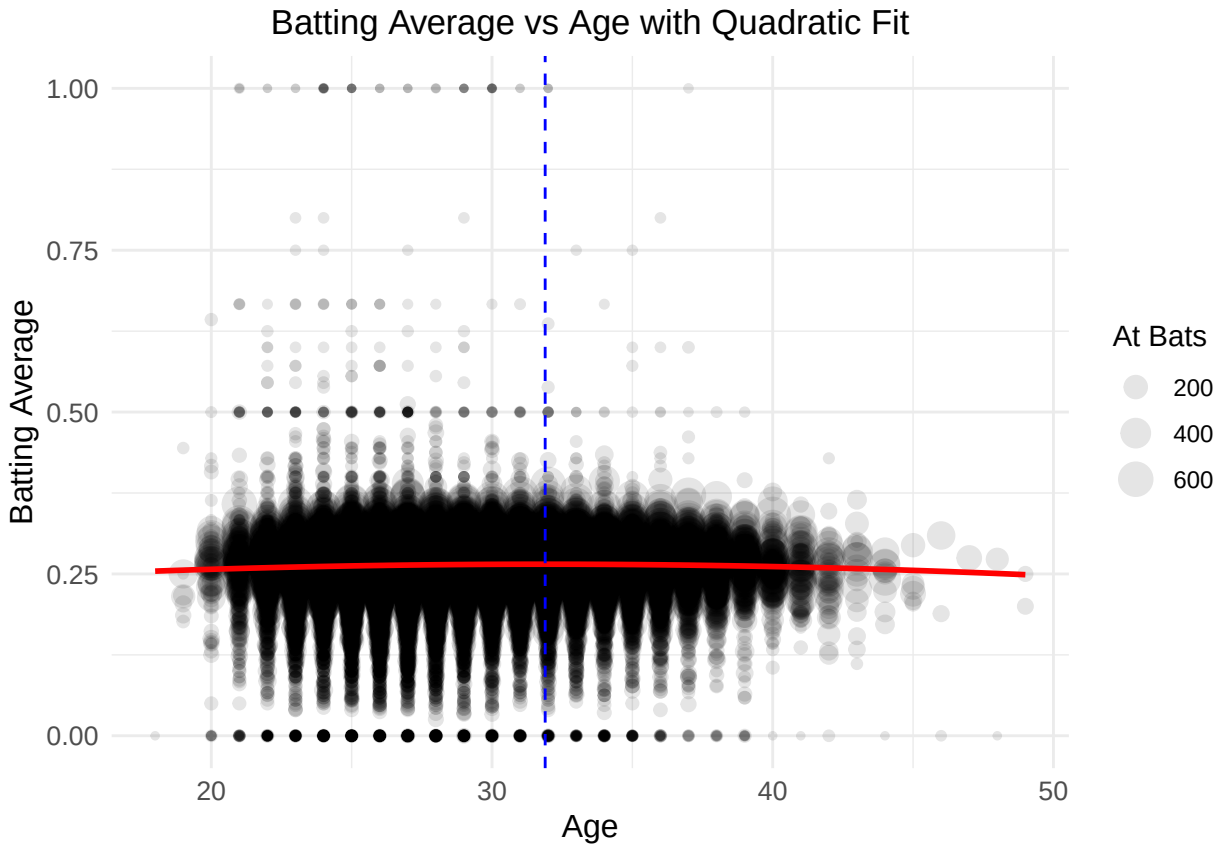
```
##
## Call:
## lm(formula = BA ~ Age + Age_squared, data = mlb_data, weights = AB)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1588 -0.4960 -0.1744  0.2133  3.0522
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.088e-01  9.127e-03  22.873  < 2e-16 ***
## Age          3.527e-03  6.127e-04   5.756 8.67e-09 ***
## Age_squared -5.529e-05  1.016e-05  -5.442 5.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5699 on 30971 degrees of freedom
## Multiple R-squared:  0.001437,   Adjusted R-squared:  0.001372
## F-statistic: 22.28 on 2 and 30971 DF,  p-value: 2.138e-10
```

```r
beta_age <- coef(quad_model)[2]
beta_age2 <- coef(quad_model)[3]
peak_age <- -beta_age / (2 * beta_age2)

cat("\nPeak batting average occurs at age:", round(peak_age, 2))
```

```
##
## Peak batting average occurs at age: 31.89
```

```r
age_range <- seq(min(mlb_data$Age), max(mlb_data$Age), by = 0.1)
predicted_ba <- predict(quad_model,
                        newdata = data.frame(Age = age_range,
                                             Age_squared = age_range^2))


ggplot(mlb_data, aes(x = Age, y = BA)) +
  geom_point(alpha = 0.1, aes(size = AB)) +  # Raw data points, sized by AB
  geom_line(data = data.frame(Age = age_range, BA = predicted_ba),
            color = "red", size = 1) +  # Fitted curve
  geom_vline(xintercept = peak_age, linetype = "dashed", color = "blue") +  # Peak age
  labs(
    title = "Batting Average vs Age with Quadratic Fit",
    x = "Age",
    y = "Batting Average",
    size = "At Bats"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12)
  )
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



(d) Use the `lmer` function in the `lme4` package for R to fit the mixed model suggested in this study (this could be called a 'random intercept, slope, and quadratic term' model), and be sure to use the argument `weight=AB` here too. Note: you can ignore the warnings, or you can attempt to fix them using more iterations in the optimization function.

```
library(Matrix)
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(lme4)
```

```
##
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:nlme':
##
##     lmList
```

```r
mlb_data$BA <- mlb_data$H / mlb_data$AB
mlb_data$Age_squared <- mlb_data$Age^2


mixed_model <- lmer(
  BA ~ Age + Age_squared + (1 + Age + Age_squared | playerID),
  data = mlb_data,
  weights = AB,
  control = lmerControl(optCtrl = list(maxfun = 100000))
)
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
summary(mixed_model)
```

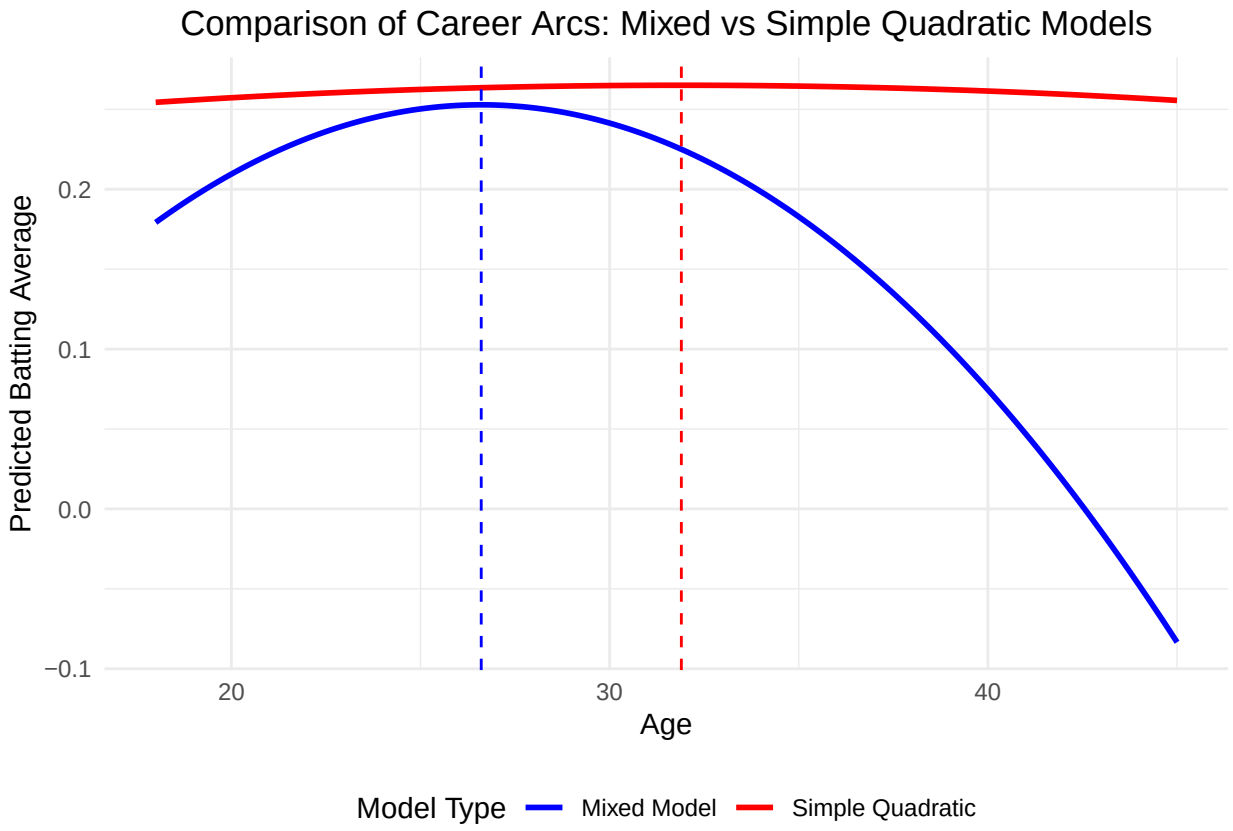
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: BA ~ Age + Age_squared + (1 + Age + Age_squared | playerID)
##    Data: mlb_data
## Weights: AB
## Control: lmerControl(optCtrl = list(maxfun = 1e+05))
##
## REML criterion at convergence: -83943.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9077 -0.7371 -0.1075  0.5092  4.1087
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  playerID (Intercept) 1.807e-01 0.425073
##           Age         1.192e-02 0.109171 -0.10
##           Age_squared 1.672e-05 0.004089 -0.03 -0.99
##  Residual             1.697e-01 0.411995
## Number of obs: 30974, groups:  playerID, 4858
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) -4.505e-01  1.638e-02  -27.50
## Age          5.287e-02  1.991e-03   26.55
## Age_squared -9.936e-04  6.533e-05  -15.21
##
## Correlation of Fixed Effects:
##             (Intr) Age
## Age         -0.544
## Age_squared  0.275 -0.956
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

there is a clear population-level pattern in how batting average changes with age, there is substantial player-to-player variation in career trajectories

q: why are the t values negative?

(e) Use your estimates for the two models to plot the average 'career arc' for each model on one set of axes (line plots make the most sense). Determine the estimated peaks of each arc.

```r
# Simple quadratic model coefficients (from part c)
simple_coef <- c(0.2088, 0.003527, -0.00005529)

# Mixed effects model coefficients (from part d)
mixed_coef <- c(-0.4505, 0.05287, -0.0009936)


age_seq <- seq(18, 45, by = 0.1)


predict_ba <- function(age, coef) {
    coef[1] + coef[2] * age + coef[3] * age^2
}


mixed_pred <- predict_ba(age_seq, mixed_coef)
simple_pred <- predict_ba(age_seq, simple_coef)


calc_peak <- function(coef) {
    -coef[2] / (2 * coef[3])
}

mixed_peak <- calc_peak(mixed_coef)
simple_peak <- calc_peak(simple_coef)


ggplot() +
    geom_line(aes(x = age_seq, y = mixed_pred, color = "Mixed Model"), size = 1) +
    geom_line(aes(x = age_seq, y = simple_pred, color = "Simple Quadratic"), size = 1) +
    geom_vline(xintercept = mixed_peak, linetype = "dashed", color = "blue") +
    geom_vline(xintercept = simple_peak, linetype = "dashed", color = "red") +
    labs(
        title = "Comparison of Career Arcs: Mixed vs Simple Quadratic Models",
        x = "Age",
        y = "Predicted Batting Average",
        color = "Model Type"
    ) +
    theme_minimal() +
    theme(
        plot.title = element_text(hjust = 0.5),
        legend.position = "bottom"
    ) +
    scale_color_manual(values = c("Mixed Model" = "blue", "Simple Quadratic" = "red"))
```




```
# Print peak ages
cat("Peak age (Mixed Model):", round(mixed_peak, 2), "\n")
```

```
## Peak age (Mixed Model): 26.61
```

```
cat("Peak age (Simple Quadratic):", round(simple_peak, 2), "\n")
```

```
## Peak age (Simple Quadratic): 31.9
```

(f) Compare the results of the two models in the previous parts. What is the interpretation of each? Why do they differ in value?

- 26.61 is likely more accurate for individual player development, while the simple model's later peak 31.90 might better represent the careers of successful long-term players.
- The difference highlights the importance of choosing appropriate statistical methods that match the data structure
- The mixed model challengse the idea that players peak around age 30, => pure batting skill might peak earlier. later peaks might be more a result of survivor bias than true skill development.