Project Part 3

```r
set.seed(7000)

library(readr)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v purrr     1.0.2
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
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 becom
```

```r
library(lme4)
```

```
Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```r
library(performance)
library(ggplot2)
library(ICC)
```

```r
#df_baseball <- read_csv("https://www.dropbox.com/scl/fi/2bcvc8eabdinum2e3r9oj/statcast_pi

load("df_baseball_clean.RData")
```

```r
set.seed(7000)
baseball_sample <- df_baseball_clean[sample(nrow(df_baseball_clean), 1000), ] |> arrange(d
head(baseball_sample)
```

```
# A tibble: 6 x 15
  type  hit_distance_sc pitch_type dist_from_cen launch_angle launch_speed
  <fct>           <dbl> <fct>              <dbl>        <dbl>        <dbl>
1 X                 397 SI                  2.47           29        103
2 X                 199 FF                  3.31           30         63.9
3 X                  78 SI                  2.64            4         98.7
4 X                 213 FF                  3.06           21         72.6
5 X                   2 SI                  2.42          -53         65.3
6 X                 336 CU                  2.49           40         94.1
# i 9 more variables: game_type <fct>, batter <fct>, stand <fct>,
#   plate_x <dbl>, plate_z <dbl>, dist_from_cen_gmc <dbl>,
#   hit_distance_gmc <dbl>, launch_angle_gmc <dbl>, launch_speed_gmc <dbl>
```

Include the following modeling steps. This may not find the best model, but will be an opportunity for you to build a multilevel model in a coherent fashion. You should be using your cleaned data set with quantitative variables grand-mean centered.

1. Include a graph exploring the variability in the response variable across the Level-2 units. Fit an ANOVA using OLS for your response variable and the Level 2 grouping variable (the Level 2 units). Does the variation in the response across the Level 2 units appear to be statistically significant?

```r
balanced_sample <- df_baseball_clean |>
  group_by(batter) |>
  mutate(total_hits = n()) |>
  filter(total_hits >= 100)
```

```r
set.seed(7000)
hits_over_100 <- df_baseball_clean |>
  group_by(batter) |>
  mutate(total_hits = n()) |>
  filter(total_hits >= 100) |>
```

```r
    ungroup()

  random_batters <- hits_over_100 |>
    distinct(batter) |>
    slice_sample(n = 100) |>
    pull(batter)

  final_data <- hits_over_100 |>
    filter(batter %in% random_batters) |>
    group_by(batter) |>
    slice_head(n = 100) |>
    ungroup()


  final_data |> distinct(batter) |> nrow()  # Should return 100
```

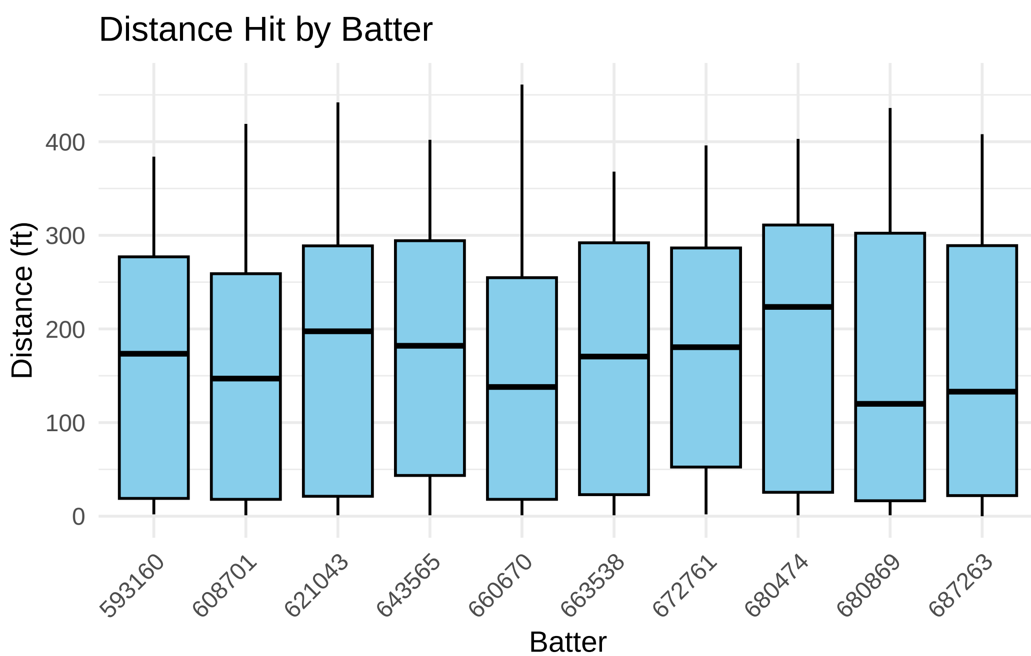
```
[1] 100
```

```r
final_data |> group_by(batter) |> summarise(total_hits = n())
```

```
# A tibble: 100 x 2
   batter total_hits
   <fct>       <int>
 1 457759        100
 2 516782        100
 3 518595        100
 4 521692        100
 5 543257        100
 6 543760        100
 7 543877        100
 8 545341        100
 9 571745        100
10 572138        100
# i 90 more rows
```

```r
final_data <- final_data |>
  mutate(is_fastball = as.integer(pitch_type %in% c("SI", "FF", "CU", "FA")))
```

```r
top_batters <- final_data |>
  distinct(batter) |>
  slice_sample(n = 10)

filtered_data <- final_data |>
  filter(batter %in% top_batters$batter)

ggplot(filtered_data, aes(x = batter, y = hit_distance_sc)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Distance Hit by Batter",
       x = "Batter",
       y = "Distance (ft)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
model0 <- lm(hit_distance_sc ~ batter, data= final_data)
anova(model0)
```

```
Analysis of Variance Table

Response: hit_distance_sc
```

```
             Df    Sum Sq Mean Sq F value    Pr(>F)    
batter      99   4768384   48165  2.6984 < 2.2e-16 ***
Residuals 9900 176713984   17850                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the results of an ANOVA fitting hit distance by batter, we have strong evidence that batter explains a significant amount of variation in hit distance ($F = 2.6984$, $p < .0001$). We will proceed with caution because we have a small F-value but a significant p-value.

2. Fit the "random intercepts only" (null) model. Interpret each of the estimated parameters in context. Interpret the intraclass correlation coefficient in context. Does the value of the ICC seem "substantial" to you? Report the likelihood, deviance, and AIC values for later comparison.

```r
nullmodel <- lmer(hit_distance_sc ~ 1 + (1 | batter), data=final_data)
summary(nullmodel)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: hit_distance_sc ~ 1 + (1 | batter)
   Data: final_data

REML criterion at convergence: 126371.2

Scaled residuals: 
     Min       1Q   Median       3Q      Max 
-1.48876 -1.03538 -0.01705  0.88674  2.27881 

Random effects:
 Groups   Name        Variance Std.Dev.
 batter   (Intercept)   303.2   17.41  
 Residual             17849.9  133.60  
Number of obs: 10000, groups:  batter, 100

Fixed effects:
            Estimate Std. Error t value
(Intercept)  167.743      2.195   76.43
```

```r
ICC::ICCbare(y=final_data$hit_distance_sc, x = final_data$batter)
```

```
Warning in ICC::ICCbare(y = final_data$hit_distance_sc, x = final_data$batter):
Missing levels of 'x' have been removed
```

```
[1] 0.0167
```

```r
logLik(nullmodel)
```

```
'log Lik.' -63185.58 (df=3)
```

```r
performance(nullmodel)
```

```
# Indices of model performance

AIC       |      AICc |       BIC | R2 (cond.) | R2 (marg.) |   ICC |    RMSE |   Sigma
---------------------------------------------------------------------------------------
1.264e+05 | 1.264e+05 | 1.264e+05 |      0.017 |      0.000 | 0.017 | 133.180 | 133.604
```

$$\tau_0^2:$$

The batter to batter variance in average hit distance is 303.2.

$$\sigma^2:$$

The variance in average hit distance for each batter is 17849.9.

$$\beta_0:$$

The average hit distance across all batters is 167.743.

$$ICC: \frac{303.2}{303.2 + 17849.9} = 0.01670238$$

The correlation between two hits by the same batter is .015. 1.5% of the variation is explained by within batter variation in hit distance rather than between batters. This is not substantial. The log likelihood of the null model is -63185.58. The deviance is 133.604 feet. The AIC is 126400.

3. Add 1-3 Level 1 variables. Carry out a likelihood ratio test to compare this model to the model in step 2 (using ML, clearly explain how you find the chi-square value and df). Include details. Also report/compare the AIC values to the intercepts only model. Calculate a "proportion of variation explained" for this set of variables and interpret the results in context (be clear variation in what). Did the Level 2 variance decrease? What does the tell you? Remove (one at a time) any insignificant variables.

```r
# dist by launch angle, pitch type, launch speed, random batter intercepts
model1 <- lmer(hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball + (1 |
summary(model1)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball +
    (1 | batter)
   Data: final_data

     AIC      BIC   logLik deviance df.resid
119930.5 119973.8 -59959.3 119918.5     9994

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3948 -0.7942  0.0329  0.8371  2.7293

Random effects:
 Groups   Name        Variance Std.Dev.
 batter   (Intercept)   50.42   7.10
 Residual             9411.47  97.01
Number of obs: 10000, groups:  batter, 100

Fixed effects:
                  Estimate Std. Error t value
(Intercept)      168.23067    1.60621 104.738
launch_angle_gmc   2.76021    0.03474  79.457
launch_speed_gmc   2.72632    0.06698  40.706
is_fastball       -3.78555    1.96226  -1.929

Correlation of Fixed Effects:
            (Intr) lnch_n_ lnch_s_
lnch_ngl_gm -0.009
lnch_spd_gm  0.046 -0.145
is_fastball -0.663  0.009  -0.089
```

```r
anova(nullmodel, model1)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: final_data
Models:
nullmodel: hit_distance_sc ~ 1 + (1 | batter)
model1: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball + (1 | batter)
          npar    AIC    BIC logLik deviance Chisq Df Pr(>Chisq)
nullmodel    3 126381 126402 -63187   126375
model1       6 119931 119974 -59959   119919  6456  3  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
performance(model1)
```

```
Model was not fitted with REML, however, `estimator = "REML"`. Set
  `estimator = "ML"` to obtain identical results as from `AIC()`.
```

```
# Indices of model performance

AIC       |      AICc |       BIC | R2 (cond.) | R2 (marg.) |   ICC |   RMSE |  Sigma
-------------------------------------------------------------------------------------
1.199e+05 | 1.199e+05 | 1.200e+05 |      0.481 |      0.478 | 0.005 | 96.843 | 97.013
```

```r
# dist by launch angle, launch speed, random batter intercepts
model2 <- lmer(hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + (1 | batter), data=

# dist by launch angle, random batter intercepts
model3 <- lmer(hit_distance_sc ~ launch_angle_gmc + (1 | batter), data=final_data, REML =
anova(nullmodel, model3, model2, model1)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: final_data
Models:
nullmodel: hit_distance_sc ~ 1 + (1 | batter)
model3: hit_distance_sc ~ launch_angle_gmc + (1 | batter)
```

```
model2: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + (1 | batter)
model1: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball + (1 | batter)
          npar    AIC    BIC logLik deviance     Chisq Df Pr(>Chisq)
nullmodel    3 126381 126402 -63187   126375
model3       4 121460 121489 -60726   121452 4922.3127  1    < 2e-16 ***
model2       5 119932 119968 -59961   119922 1530.0115  1    < 2e-16 ***
model1       6 119931 119974 -59959   119919    3.7194  1    0.05378 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# dist by launch angle, random pitch type intercepts, launch speed, random batter intercep
model4 <- lmer(hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + (1 | is_fastball) +
performance(model4)
```

```
Model was not fitted with REML, however, `estimator = "REML"`. Set
  `estimator = "ML"` to obtain identical results as from `AIC()`.
```

```
# Indices of model performance

AIC       |      AICc |       BIC | R2 (cond.) | R2 (marg.) |   ICC |   RMSE |  Sigma
-------------------------------------------------------------------------------------
1.199e+05 | 1.199e+05 | 1.200e+05 |      0.481 |      0.478 | 0.006 | 96.847 | 97.021
```

In model 1, we added our three level 1 variables of interest (launch speed, launch angles, and if its a fastball) and used a likelihood ratio test to compare. Using an anova to compare the null model and model 1, the likelihood test gave us a large chi-square test statistic of 6456 and a p-value less than 0.001. The likelihood test had 3 degrees of freedom, which is the difference in the number of parameters between the null model and model 1. The AIC of model 1 is 119931 which is less than the null model AIC of 126381. Model 1 also has a BIC of 119974 which is less than the null models BIC of 126402. Therefore there is strong evidence to conclude that model 1 with all the level 1 variables is a better fit for the data than the random intercepts null model.

The level 1 variables is_fastball, launch angle, and launch speed explain (R^2 = 0.481) 48.1% the variation in hit distance. Yes, the level 2 variance decreased from 303.2 (in null model) to 50.42 (in model 1).

We then created model 2 that includes launch speed, launch angle, and batter random intercepts. We also created model 3 which only includes launch angle and batter random intercepts. We then ran an anova of models 1, 2, and 3. We found that model 2, is a better fit of the data than model 3 and launch speed is

a statistically significant predictor of hit distance (chi-square = 1530, p-value < 0.001). However, we found that model 1 was not significantly better fit of the data than model 2 and the variable is_fastball is not a statistically significant predictor of hit distance (chi-square = 3.7194, p-value = 0.05378).

4. Add 1-3 Level 2 variables. Carry out a likelihood ratio test to compare the models (using ML). Include details. Also report/compare the AIC values. Calculate a "proportion of variation explained" for each level and interpret the results in context. Remove (one at a time) any insignificant variables.

```
# dist by launch angle, pitch type, launch speed, stand, random batter intercepts
model5 <- lmer(hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball + stand

summary(model5)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball +
    stand + (1 | batter)
   Data: final_data

     AIC      BIC   logLik deviance df.resid
119932.2 119982.7 -59959.1 119918.2     9993

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4011 -0.7943  0.0342  0.8361  2.7325

Random effects:
 Groups   Name        Variance Std.Dev.
 batter   (Intercept)   49.75   7.053
 Residual             9411.66  97.014
Number of obs: 10000, groups:  batter, 100

Fixed effects:
                  Estimate Std. Error t value
(Intercept)      169.08631    2.30277  73.427
launch_angle_gmc   2.76003    0.03474  79.447
launch_speed_gmc   2.72657    0.06698  40.710
is_fastball       -3.77499    1.96230  -1.924
standR            -1.29265    2.49371  -0.518

Correlation of Fixed Effects:
```

```
            (Intr) lnch_n_ lnch_s_ is_fst
lnch_ngl_gm -0.016
lnch_spd_gm  0.037 -0.145
is_fastball -0.456  0.009  -0.089
standR      -0.717  0.013  -0.007  -0.009
```

```r
anova(model1, model5)
```

```
Data: final_data
Models:
model1: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball + (1 | batter)
model5: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball + stand + (1 | ba
       npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
model1    6 119931 119974 -59959   119919
model5    7 119932 119983 -59959   119918 0.2676  1     0.6049
```

```r
performance(model5)
```

```
Model was not fitted with REML, however, `estimator = "REML"`. Set
  `estimator = "ML"` to obtain identical results as from `AIC()`.

# Indices of model performance

AIC       |      AICc |       BIC | R2 (cond.) | R2 (marg.) |   ICC |   RMSE |  Sigma
-------------------------------------------------------------------------------------
1.199e+05 | 1.199e+05 | 1.200e+05 |      0.481 |      0.478 | 0.005 | 96.846 | 97.014
```

Model 5 includes three level 1 variables and where the batter stands as level 2 variable. Model 5 has an AIC of 119932 which is barely larger than the AIC for model 1 which is 119931.

$$\frac{49.75}{49.75 + 9411.66} = 0.005258$$

Level 2 batter variance explains 0.5% of the total variance in distance hit.

$$\frac{9411.66}{49.75 + 9411.66} = 0.9947418$$

Level 1 hit variance explains 99.5% of the total variation in distance hit.

Since we only have one level 2 variable, the likelihood test for model 1 and model 5 produces a p-value of 0.6049 and a chi-square statistic of 0.2676 (df=1). This tells us that stand is not a statistically significant predictor of distance hit.

5. Consider random slopes for one Level 1 variable. (This could involve putting back in one of the variables that was removed earlier…) Include a graph illustrating variability in the estimated random slopes and discuss what you learn in context. Interpret the amount of group-to-group variation in these slopes in context. Once you have a model with at least one set of random slopes, compare this model to the model in step 4, is adding random slopes a significant improvement (REML, be clear how you are determining degrees of freedom)?

```r
model6 <- lmer(hit_distance_sc ~ is_fastball + launch_angle_gmc + launch_speed_gmc + stand
```

```
Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
Model failed to converge with max|grad| = 1.40873 (tol = 0.002, component 1)
```

```
Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearl
 - Rescale variables?
```

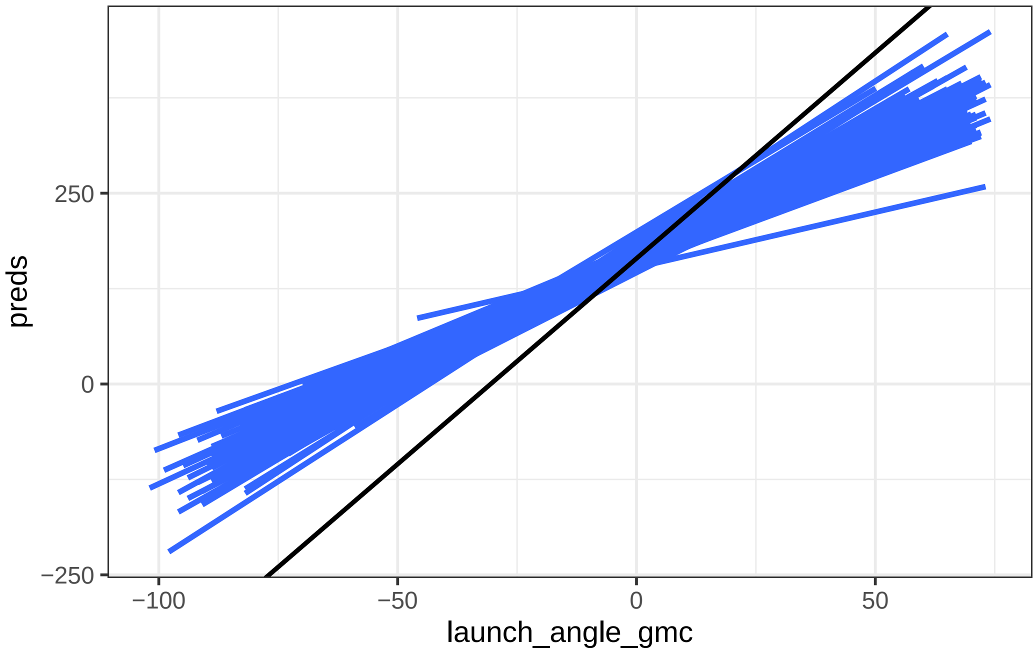
```r
summary(model6)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: hit_distance_sc ~ is_fastball + launch_angle_gmc + launch_speed_gmc +
    stand + (1 + launch_angle_gmc | batter)
   Data: final_data

REML criterion at convergence: 119863.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6783 -0.7941  0.0190  0.8242  2.6793

Random effects:
 Groups   Name             Variance  Std.Dev. Corr
 batter   (Intercept)        66.8888  8.1786
          launch_angle_gmc    0.1644  0.4054  0.67
 Residual                  9284.0320 96.3537
Number of obs: 10000, groups:  batter, 100
```

```
Fixed effects:
                 Estimate Std. Error t value
(Intercept)     168.82279    2.34374  72.031
is_fastball      -3.74606    1.95575  -1.915
launch_angle_gmc   2.81219    0.05342  52.642
launch_speed_gmc   2.68660    0.06702  40.084
standR           -0.23179    2.49797  -0.093

Correlation of Fixed Effects:
            (Intr) is_fst lnch_n_ lnch_s_
is_fastball -0.446
lnch_ngl_gm  0.165  0.006
lnch_spd_gm  0.032 -0.089 -0.097
standR      -0.708 -0.009  0.014  -0.005
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 1.40873 (tol = 0.002, component 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
```

```r
preds = predict(model6, newdata = final_data)
ggplot(final_data, aes(x = launch_angle_gmc , y = preds , group = batter)) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_abline(intercept = 165.80, slope = 2.86 + 2.53) +
geom_abline(intercept = 165.80 - 2.74, slope = 2.86 + 2.53) +
geom_abline(intercept = 165.80 + .23, slope = 2.86 + 2.53) +
geom_abline(intercept = 165.80 + .23 - 2.74, slope = 2.86 + 2.53) +
  theme_bw()
```

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



The least amount of batter-to-batter variation in estimated launch angle slopes occurs at about -20 degrees.

```r
anova(model5, model6)
```

```
refitting model(s) with ML (instead of REML)

Data: final_data
Models:
model5: hit_distance_sc ~ launch_angle_gmc + launch_speed_gmc + is_fastball + stand + (1 | ba
model6: hit_distance_sc ~ is_fastball + launch_angle_gmc + launch_speed_gmc + stand + (1 + la
       npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
model5    7 119932 119983 -59959   119918
model6    9 119881 119946 -59932   119863 54.821  2  1.247e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compared to model5 which we fit in part 4, based on the results of adding random slopes for grand mean centered launch angle this seemed to improve the model, as shown by Chisq = 97.79 and p < .0001. We have 2 df for this test because adding random slopes introduced a variance component for the random launch angle slopes, as well as the covariance between the random intercepts and the random launch angle slopes.

6. Add and interpret a cross-level interaction (you may have to use insignificant variables, focus on interpreting the interaction). Are you able to explain much of the slope variation you found in step 5? Is this a significantly better model?

```
model7 <- lmer(hit_distance_sc ~ is_fastball + launch_angle_gmc + launch_speed_gmc + stand
```

```
Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
Model failed to converge with max|grad| = 1.93446 (tol = 0.002, component 1)
```

```
Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is near
 - Rescale variables?
```

```
summary(model7)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: hit_distance_sc ~ is_fastball + launch_angle_gmc + launch_speed_gmc +
    stand + is_fastball * stand + (1 + launch_angle_gmc | batter)
   Data: final_data

REML criterion at convergence: 119853.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6964 -0.7920  0.0238  0.8266  2.6696

Random effects:
 Groups   Name             Variance  Std.Dev. Corr
 batter   (Intercept)        63.5412  7.971
          launch_angle_gmc    0.1681  0.410   0.67
 Residual                  9280.2805 96.334
Number of obs: 10000, groups:  batter, 100

Fixed effects:
                    Estimate Std. Error t value
(Intercept)        165.54499    2.75041  60.189
is_fastball          2.45953    3.36245   0.731
launch_angle_gmc     2.81254    0.05377  52.312
launch_speed_gmc     2.68838    0.06701  40.120
standR               4.74902    3.32993   1.426
is_fastball:standR  -9.33901    4.12268  -2.265
```

```
Correlation of Fixed Effects:
            (Intr) is_fst lnch_n_ lnch_s_ standR
is_fastball -0.655
lnch_ngl_gm  0.137  0.002
lnch_spd_gm  0.022 -0.043 -0.097
standR      -0.802  0.539  0.009   0.004
is_fstbll:R  0.533 -0.814  0.001  -0.011  -0.668
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 1.93446 (tol = 0.002, component 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
```

```
anova(model6,model7)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: final_data
Models:
model6: hit_distance_sc ~ is_fastball + launch_angle_gmc + launch_speed_gmc + stand + (1 + la
model7: hit_distance_sc ~ is_fastball + launch_angle_gmc + launch_speed_gmc + stand + is_fast
       npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
model6    9 119881 119946 -59932   119863
model7   10 119878 119950 -59929   119858 5.2221  1     0.0223 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We fit a model with an interaction between stand and fastball. This is a significantly better model because when doing an ANOVA comparing model 6 and our new model, we got a p-value of 0.022. This means we have evidence that the model with an interaction term is significantly better, and therefore explains more variation in the centered hit distance, than model 6. This new model explains a little bit more of the slope variation in the random centered launch angle slope. It used to have a variation of 0.161 and now, with the interaction, it has a variation of 0.1681.

Keep in mind: Doing what I tell you to do is ~ B work. Doing more or less will move your grade up or down. Possible Extras: Enhanced graphs; More than 2 levels; Compare model in step 3 to a random effects ANCOVA model (using OLS); Testing additional random slopes; Cross validation (or at least consider possible multiple comparison issues); Including and interpreting confidence intervals