Comparing Hitting Talent Accross Different Baseball Leages

Lee Przybylski

November 21, 2020

1 Comparing Leagues

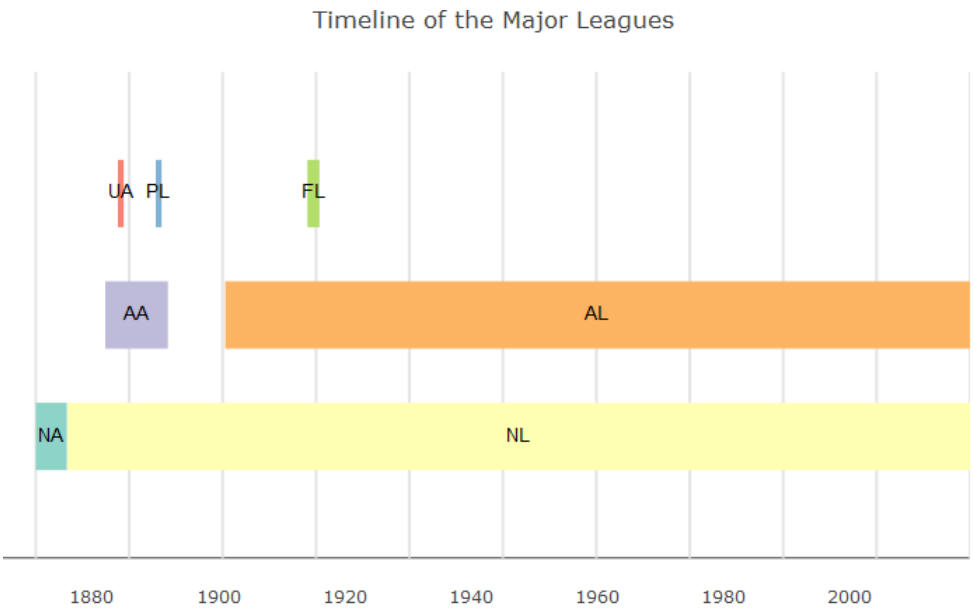

Figure 1: A rough timeline for the seven leagues considered as the Major Leagues.

We explore the prospect of comparing hitting talent accross diffent leagues using the Lahman database, available in the `Lahman` package in `R`. We will first try to recreate the findings of Richard Cramer first published in the 1980 *Baseball Research Journal* where he showed that average batting skill improved over the years by comparing players who played in the same league accross consecutive years. This article can be found at

`https://ourgame.mlblogs.com/average-batting-skill-through-major-league-history-landmarks-of-sabermetrics-part-i-bb5849adae0b`

We will treat the American and National Leagues (AL and NL) as two different leagues and we can compare the two leagues to each other. We can also use the same method to compare different seasons of the same league. When we model OPS for a particular player in a particular league, our model will only use data from a span of 2 years, because we want to limit ourselves to the moderate assumption that a hitter's ability does not change significantly between two consecutive seasons. This will allow us to compare consecutive seasons directly. Indirect comparisons will be based on different models.

In each model, we want to determine two test statistics, one to represent the differences in two different league effecs on OPS and one to represent the differences in average hitter talent. For the purposes of

discussion, we denote our players by $i = 1, 2, ..., n$ and leagues by $j = 1, 2, ..., m$. Each league has its own roster of players. We denote the fact that player i is played in league j by $i \in L_j$. We let n_j denote the number of players in league j.

2 The OPS Distribution

OPS stands for on base plus slugging percentage. It is defined as

$$OPS := \frac{H + BB + HBP}{AB + BB + HBP + SF} + \frac{TB}{AB}.$$

Just as its name suggests, this is on base percentage (OBP) plus slugging percentage ($SLUG$). We refer to any plate appearance counted as an at-bat, base on balls, hit by pitch, or sacrifice fly as a batter's attempt. Sacrifice bunts do not count as batter's attempts. Let n equal the number of attempts, so $n = AB + BB + HBP + SF$. Given a batter playing a season in a given league, we define the following parameters: p_w equals the probability that a single batter's attempt results in a walk or a hit by pitch, p_j $(j = 1, 2, 3, 4)$ is the probability that a single batter's attempt results in a hit that advances the batter to the jth base, (p_1 is the probability the batter hits a single etc.), and $p_B = p_1 + p_2 + p_3 + p_4 + p_w$. We let p_0 denote the probabilty that a player's plate appearance results in an at bat but not a hit. This would include strikeouts, ground outs, and reaching base by a fielding error. Finally, we let p_s denote the probability that a player's plate appearnce results in a sacrifice fly or sacrifice hit. These instances do not count as at-bats. We have

$$p_s + p_w + p_0 + p_1 + p_2 + p_3 + p_4 = 1$$

so our model has six parameters, $\boldsymbol{\theta} = (p_s, p_w, p_1, , p_2, p_3, p_4)'$. These are readilly estimated using the standard statistics for a hitter in a given season. For example, $\hat{p}_w = \frac{BB+HBP}{n}$, $\hat{p}_s = \frac{SF+SH}{n}$, etc. According the model,

$$\mathbb{E}[OBP] = \mathbb{E}\left[\frac{H + BB + HBP}{n}\right] = p_B$$

so $OBP = \hat{p}_B$ is an estimator of of p_B. We also have

$$\text{Var}[OBP] = \frac{p_B(1 - p_B)}{n}.$$

We can model TB as a multinomial distribution, meaning a sum of n i.i.d. trials of A_j such that $P[A = j] = p_j$ for $j = 0, 1, 2, 3, 4$ as defined above. Since

$$\mathbb{E}[A_j] = p_1 + 2p_2 + 3p_3 + 4p_4.$$

we get $TB = n\hat{p}_1 + 2n\hat{p}_2 + 3np_3 + 4n\hat{p}_4$ and

$$\mathbb{E}[TB] = np_1 + 2np_2 + 3np_3 + 4np_4.$$

Since $AB = n(1 - \hat{p}_w - \hat{p}_s)$, we find that

$$SLUG = \frac{n(\hat{p}_1 + 2\hat{p}_2 + 3\hat{p}_3 + 4\hat{p}_4)}{n(1 - \hat{p}_w)} = \frac{\hat{p}_1 + 2\hat{p}_2 + 3\hat{p}_3 + 4\hat{p}_4}{1 - \hat{p}_w - \hat{p}_s}.$$

Putting everything together,

$$OPS = \hat{p}_w + \hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \frac{\hat{p}_1 + 2\hat{p}_2 + \hat{p}_3 + \hat{p}_4}{1 - \hat{p}_w - \hat{p}_s}.$$

Using Taylor's theorem, we can show that

$$\mathbb{E}[OPS] \approx p_w + p_1 + p_2 + p_3 + p_4 + \frac{p_1 + 2p_2 + 3p_3 + p_4}{1 - p_w - p_s} \tag{1}$$

and

$$\text{Var}[OPS] \approx D'\text{Var}[\hat{\boldsymbol{\theta}}]D, \tag{2}$$

where

$$D = \begin{bmatrix} \frac{\partial OPS}{\partial p_s} \\ \frac{\partial OPS}{\partial p_w} \\ \frac{\partial OPS}{\partial p_1} \\ \frac{\partial OPS}{\partial p_2} \\ \frac{\partial OPS}{\partial p_3} \\ \frac{\partial OPS}{\partial p_4} \end{bmatrix} = \begin{bmatrix} \frac{p_1+2p_2+3p_3+4p_4}{(1-p_w-p_s)^2} \\ 1+\frac{p_1+2p_2+3p_3+4p_4}{(1-p_w-p_s)^2} \\ 1+\frac{1}{1-p_w-p_s} \\ 1+\frac{2}{1-p_w-p_s} \\ 1+\frac{3}{1-p_w-p_s} \\ 1+\frac{4}{1-p_w-p_s} \end{bmatrix}.$$

To understand $\text{Var}[\hat{\boldsymbol{\theta}}]$, first observe that for one of our parameters p_i,

$$\mathbb{E}[\hat{p}_i] = p_i, \quad \text{Var}[\hat{p}_i] = \frac{p_i(1-p_i)}{n}.$$

For two different parameters, p_i and p_k, we find their covariance by first defining B_{ij} as 1 if the event corresponding to i occurred at the jth at bat, and 0 otherwise. We define B_{kj} in basically the same way. Thus

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n} B_{ij}, \;\; \hat{p}_k = \frac{1}{n}\sum_{j=1}^{n} B_{kj}.$$

Thus,

$$\begin{aligned} \text{Cov}[\hat{p}_i, \hat{p}_k] &= \mathbb{E}[\hat{p}_i\hat{p}_k] - \mathbb{E}[\hat{p}_i]\mathbb{E}[\hat{p}_k] = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{l=1}^{n}\mathbb{E}[B_{ij}B_{kl}] - p_ip_k = \frac{1}{n^2}\sum_{1\le l\ne j\le n}\mathbb{E}[B_{ij}]\mathbb{E}[B_{kl}] - p_ip_k \\ &= \frac{1}{n^2}\sum_{1\le l\ne j\le n} p_ip_k - p_ip_k = \frac{n^2-n}{n^2}p_ip_k - p_ip_k = -\frac{1}{n}p_ip_k \end{aligned}.$$

We note that the third equality above is due to the fact that the events used to define p_i and p_k cannot happen simultaneously, and events corresponding to different plate appearances are assumed to be independent. Hence

$$\text{Var}[\hat{\boldsymbol{\theta}}] = n^{-1}\begin{bmatrix} p_s(1-p_s) & -p_sp_w & -p_sp_1 & -p_sp_2 & -p_sp_3 & -p_sp_4 \\ & p_w(1-p_w) & -p_wp_1 & -p_wp_2 & -p_wp_3 & -p_wp_4 \\ & & p_1(1-p_1) & -p_1p_2 & -p_1p_3 & -p_1p_4 \\ & & & p_2(1-p_2) & -p_2p_3 & -p_2p_4 \\ & & & & p_3(1-p_3) & -p_3p_4 \\ & & & & & p_4(1-p_4) \end{bmatrix}$$

From (2), this suggests that $\text{Var}[OPS] = \frac{\sigma^2}{n}$ for some $\sigma^2 > 0$.

2.1 Residuals in the Data

To test the theoretical result above, we took $n = 72087$ observations of hitters from the lahman data set in R. We fit a mixed effects model using maximum likelihood and plot the squared residuals against the number of plate appearances. The model is

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma_k + p_k + t_{js} + \varepsilon_{ijkl}, \;\; p_k \sim N(0,\sigma_p^2), t_{js} \sim N(\sigma_t^2), \varepsilon_{ijkl} \sim N(0,\sigma_{ijkl}^2), \tag{3}$$

where the intercepts λ_{ij} and γ_k represent the effect on OPS playing in the ith league during the jth season and the effect on OPS of being born the same year as the kth hitter. The random effects p_k and t_{js} are

related to the kth player and the team s during season j. In figure 2 we plot the squared residuals and two regression function based on PA^{-1} and $PA^{-1/2}$. It is troubling that the decay in the sqared residuals seems to have a better linear fit with respect to $PA^{-1/2}$. More evidence of this phenomenon appears when we do a

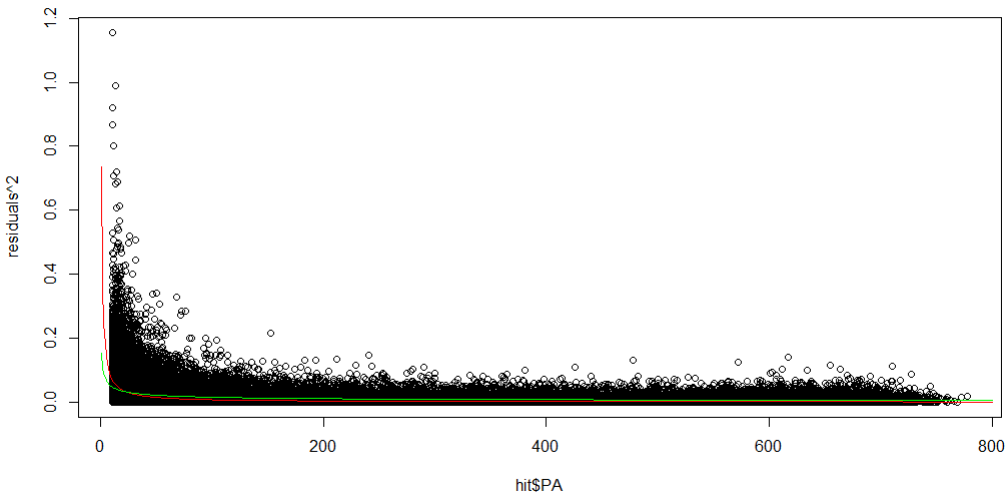

Figure 2: A plot of the squared residuals based on the fit of the mixed model in 3. The red curve represents $e = 0.7375PA^{-1}$. The green curve represents $e = 0.1532PA^{-1/2}$.

linear regression on the log of the squared residuals on the log of the number of plate appearnaces. Since the expected value of our squared residuals is equal to the variance of ε_{ijkl} according to our model. We expect from the theoretical argument above that $\sigma^2_{ijkl} = \sigma^2/PA_{ijkl}$. If we talke the logarithm of both sides, this is equivalent to

$$\log \sigma^2_{ijkl} = -1 \cdot \log PA_{ijkl} + \log \sigma^2.$$

Unfortunately, when we take a regression of the log of the squared residuals on the log of the number of plate apperances, our regression coefficient is much closer to $-1/2$ rather than -1.

```
> residuals <- residuals(model)
> sq_res <- residuals^2
> sum(sq_res == 0)
[1] 0
> lm(log(sq_res) ~ log(hit$PA))

Call:
lm(formula = log(sq_res) ~ log(hit$PA))

Coefficients:
(Intercept)  log(hit$PA)
    -3.2442      -0.5102
```

3 Linear Model Selection

In this section, we consider a series of linear models to estimate the true mean of OPS for a given player in a given league. Our models assume there is a portion of the mean that is attributed to the player and another portion of the mean that results from the players environment. Estimates of the term that is unique for the

player give us an estimate of the player's talent. Due to the theoretical result in the previous section, we suspect that the variance is inversely related to the number of plate appearances. We find in practice that this is not exactly the case. To illustrate this, we fit an unweighted model to the Lahman data from 1899 to 1972. This model is given by

$$OPS_{ijk} = \mu + \rho_i + \lambda_{jk} + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{4}$$

where OPS_{ij} OPS for player i playing in league j during season k. In our code, we avoid singularities by creating a new factor, `lg.yr` by pasting the name of the league to the season for each observation. This makes sense since not every league plays in every season. We then obtain the residuals of model (4). We expect that a weighted model will have standard deviation proportional to a power of the number of plate appearances, that is,

$$\text{Var}[OPS_{ijk}] = PA_{ijk}^{\theta}\sigma^2$$

for some $\theta < 0$. Equivalently,

$$\log(\text{Var}[OPS_{ijk}]) = \theta \log(PA_{ijk}) + 2\log(\sigma).$$

Thus we can estimate θ by performing a linear regression of the log of the squared residuals of the unweighted model with respect to $\log(PA)$. When doing this, we notice that a collection of the squared residuals are extremely small. These are for the instances where we only observe a player once. With one observation, it is impossible to get a reasonable estimate of the player's effect on OPS. This will also be common in the college baseball data.

A major problem with (4) is that for players that have only one observation, the residual is very small. This is because there many observations of a particular league-year, but only one observation for the player.

Because of how common it is for us to have few observations of a single player, we would like to consider a model with random player effects. Consider

$$OPS_{ijk} = \mu + \lambda_{jk} + p_i + \varepsilon_{ijk}PA_{ijk}^{-1/2},\ p_i \sim N(0, \sigma_p^2),\ \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{5}$$

3.1 League-Year Additive Model 1

Our first model is simple because it does not estimate player effects at all. We only consder an intercept from the league and a slope associated with the year.

$$OPS_{ij} = \lambda_i + \gamma x_j + \varepsilon_{ijk}PA_{ijk}^{-1/2}, \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{6}$$

The intercept for the ith league is represented by λ_i, and the slope for representing the change in OPS over accross different seasons is represented by γ. We fit (6) using the lahman data in `FitMisedEffModel-MLB.R` as `lm0`. The computer takes less that 1 second to fit the model. In the R output, we provide our estimates for the coefficients in 6:

```
> summary(lm0)

Call:
lm(formula = OPS ~ LG + YR, data = hit, weights = PAsc)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.24640 -0.06887 -0.03122  0.01558  0.62099

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.3770611  0.0268284 -14.055  < 2e-16 ***
LGAL         0.0344442  0.0037442   9.199  < 2e-16 ***
```

```
LGFL         0.0120713  0.0079737   1.514    0.130
LGNA        -0.0079360  0.0079991  -0.992    0.321
LGNL         0.0254440  0.0036957   6.885 5.83e-12 ***
LGPL         0.0828031  0.0108619   7.623 2.50e-14 ***
LGUA        -0.0506231  0.0127539  -3.969 7.22e-05 ***
YR           0.0005421  0.0000141  38.451  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07669 on 72079 degrees of freedom
Multiple R-squared:  0.03149,Adjusted R-squared:  0.03139
F-statistic: 334.8 on 7 and 72079 DF,  p-value: < 2.2e-16
```

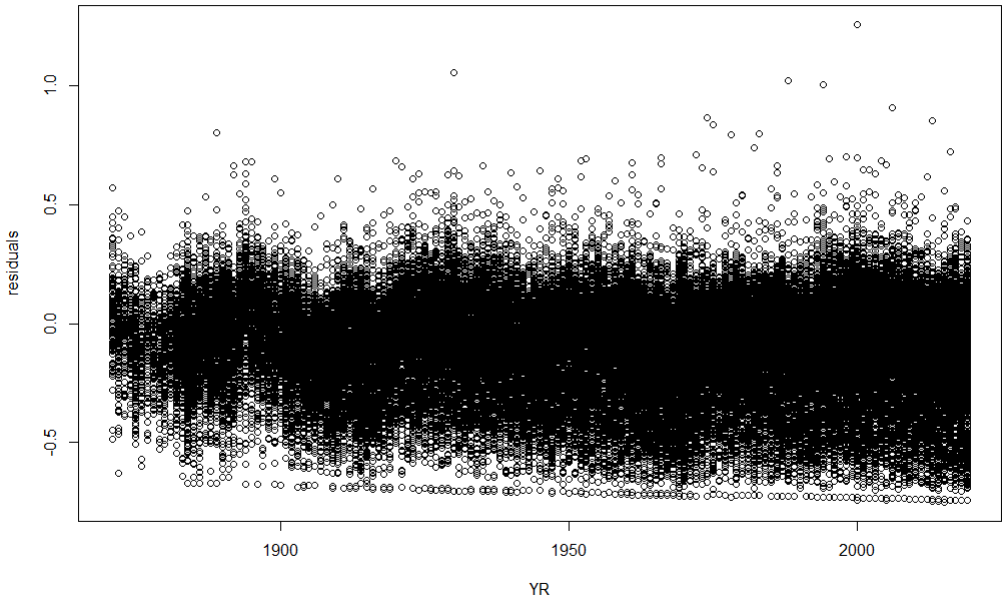

Figure 3: A residual plot based on the fit of model (6)

In Figure 3 we plot the residuals resulting in the fit of the model with respect to year. Notice that there is some sort of non-linear trend in the residual plot. The magnitude of the residuals also seems to increase sightly in more recent seasons.

3.2 League-Year-Factor Additive Model

Our second model assigns an intercept for each year.

$$OPS_{ij} = \lambda_i + \gamma_j + \varepsilon_{ijk} PA_{ijk}^{-1/2}, \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{7}$$

The intercept for the ith league is represented by λ_i, and the intercept for season j is represented by γ_j. We fit (7) using the lahman data in `FitMisedEffModel-MLB.R` as `lm1`. The computer takes about 2 seconds to fit the model. Assigning a unique intercept for each season takes care of the non-linear trend in the residuals as we see in figure 3. There is still an issue with the magnitude increase in the residuals for more recent seasons. The residuals also seemed to skew left.

The improved fit between (6) to (7) is significant. We see this in the following anova table:

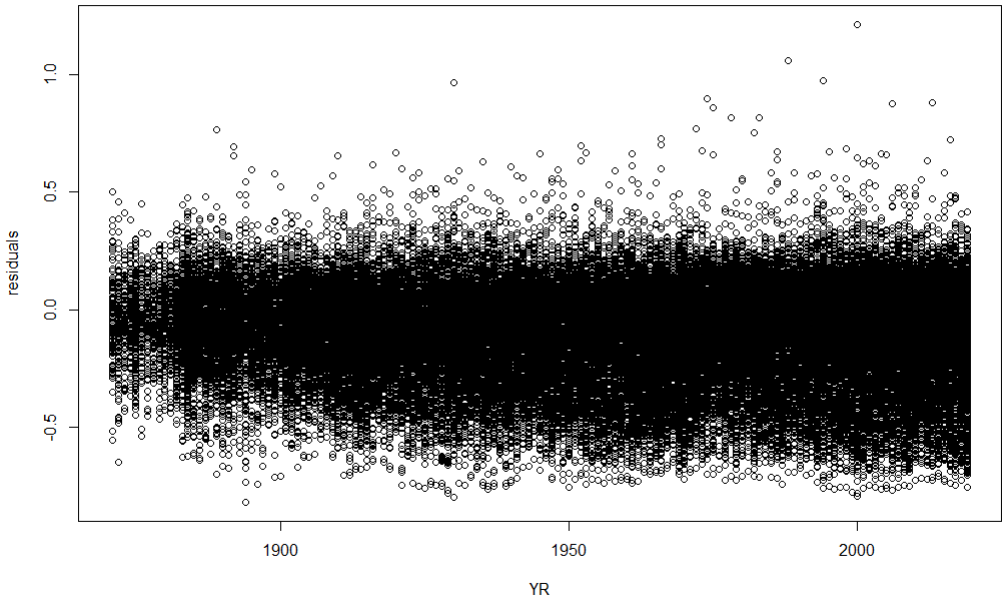

Figure 4: A residual plot based on the fit of model (7)

```
> anova(lm0,lm1)
Analysis of Variance Table

Model 1: OPS ~ LG + YR
Model 2: OPS ~ LG + YRf
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1  72079 423.89
2  71933 397.96 146    25.939 32.114 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3.3 League-Year-Interaction Model

The next model allows for interactions between the league and year effects. This makes sense since each league may not follow the same shape of nonlinear path. To avoid missing interactions, since not every league plays in every season, we create a new factor called `lg.yr` by pasting together the predictors `LG` and `YR`.

$$OPS_{ij} = \lambda_{ij} + \varepsilon_{ijk} PA_{ijk}^{-1/2}, \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{8}$$

The intercept for the ith league in season j is represented by λ_{ij}. We fit (8) using the lahman data in `FitMisedEffModel-MLB.R` as `lm2`. The computer takes about 5 seconds to fit the model. We check the residual plot in figure 5. The residuals are still skewed left, and slightly larger for more recent seasons.

We check the quality of fit with anova again. The improved fit is significant.

```
Analysis of Variance Table

Model 1: OPS ~ LG + YRf
Model 2: OPS ~ lg.yr
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1  71933 397.96
```

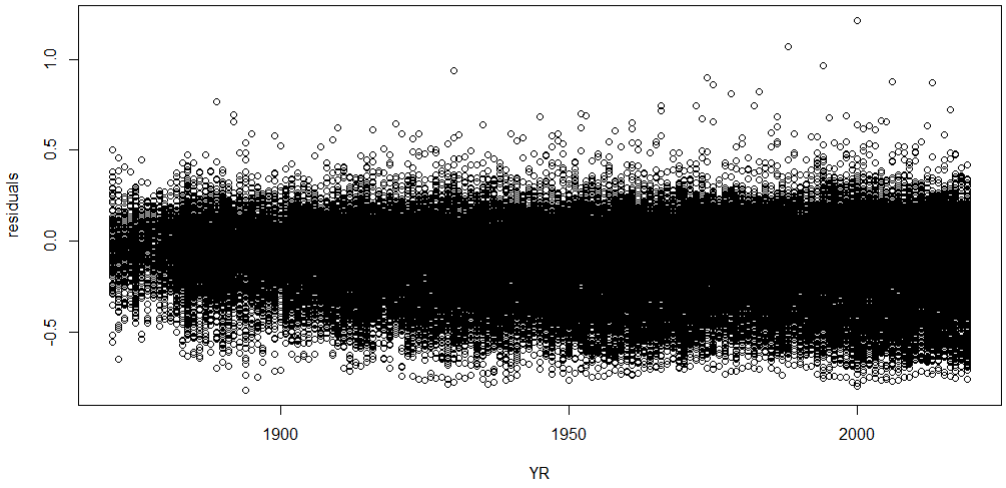

Figure 5: A residual plot based on the fit of model (8)

```
2  71805 395.64 128    2.3106 3.2762 < 2.2e-16 ***
```

3.4 League-Year-Interaction Mixed Model

Our first mixed model is similar to (8), but we add random intercepts for each different player.

$$OPS_{ij} = \lambda_{ij} + p_k + \varepsilon_{ijk} PA_{ijk}^{-1/2},\ p_k \sim N(0, \sigma_p),\ \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{9}$$

The intercept for the ith league in season j is represented by λ_{ij}, and p_k are the normally distributed player effects. We fit (9) using the lahman data in `FitMixedEffModel-MLB.R` as `mixed1`. The computer takes about 1.5 minutes to fit the model. We check the residual plot in figure 6. The residuals are still skewed left, but it seems the magnitude of the residuals is fairly consistent accross seasons.

We check the quality of fit using a likelihood ratio test compared to the fit in model (8). The improved fit from the addition of random player effects is significant.

```
Data: hit
Models:
lm2: OPS ~ lg.yr
mixed1: OPS ~ lg.yr + (1 | Plyr)
        Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
lm2    283 -39532 -36932  20049   -40098
mixed1 284 -80831 -78222  40699   -81399 41301      1  < 2.2e-16 ***
---
```

Another important observation regarding this model is the estimated lg.yr effects that result from from its fit. Observe in figure 7 how the estimated effects oscillate over time. A little research could help line up the peaks and valleys with the eras of high offense and defense in the Major leagues. For example, the peak in the last half of the 1990's could be due to the steroid era. The blue vertical line is for 1973, the first season of the designated hitter rule, which seems to correspond to a rise in offense. It is likely that each oscillation could be explained by some significant rule change or technological development.

3.5 Environment-Birthyr Mixed Model 1

Our second mixed model maintains the random player effects and adds a fixed effect in the form of a slope for birth.yr. The rationale is that the talent of hitters improves from generation to generation due to many

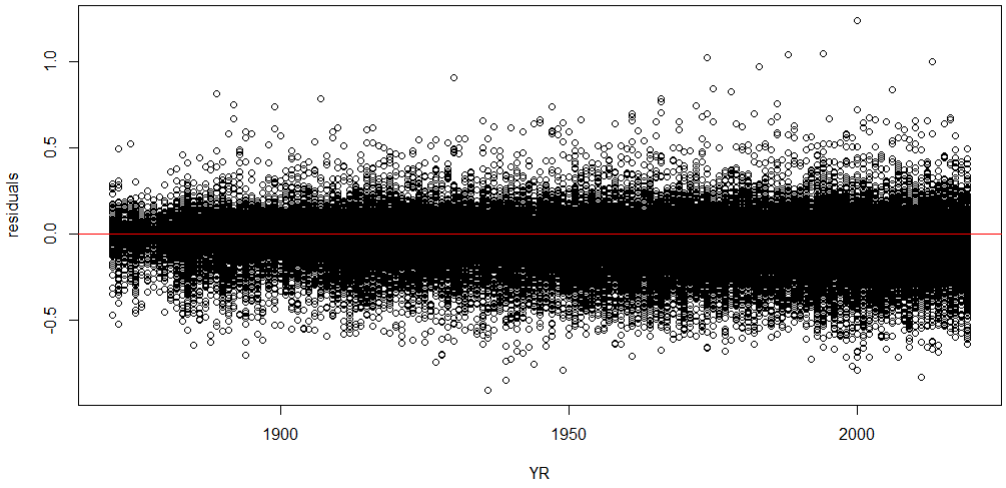

Figure 6: A residual plot based on the fit of model (9)

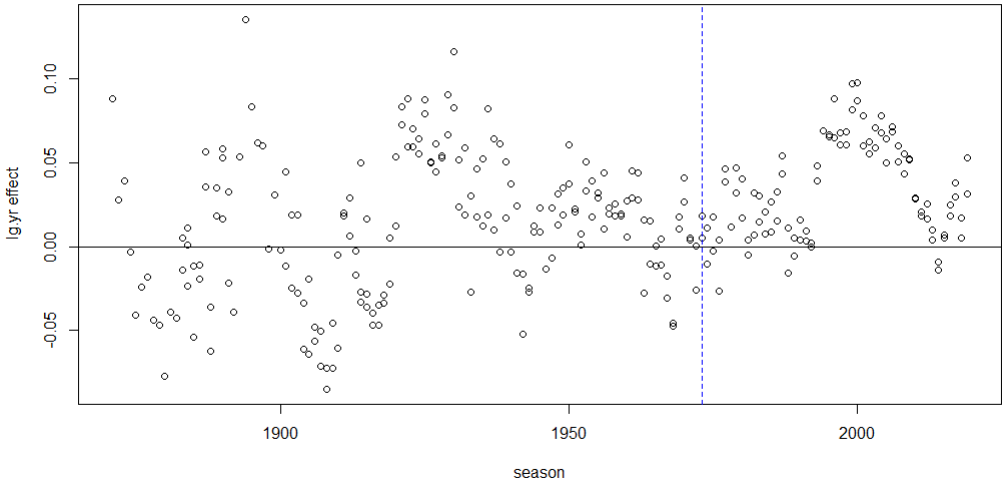

Figure 7: A plot of the lg.yr effects estamated by fitting model (9).

factors including a growing talent pool and better health and training.

$$OPS_{ij} = \lambda_{ij} + \gamma x_k + p_k + \varepsilon_{ijk} PA_{ijk}^{-1/2},\ p_k \sim N(0, \sigma_p),\ \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{10}$$

The intercept for the ith league in season j is represented by λ_{ij}. γ is the slope associated with the improvement of hitting ability for players over the years. x_k is the year player k was born. As before, p_k are the normally distributed player effects. We fit (10) using the lahman data in `FitMixedEffModel-MLB.R` as `mixed2`. The computer takes about 1.5 minutes to fit the model. We check the residual plot in figure 8. The residuals are similar to what we saw in the fit of model (9).

We check the quality of fit using a likelihood ratio test compared to the fit in model (9). The improved fit from the addition of birth year is significant.

```
Data: hit
```

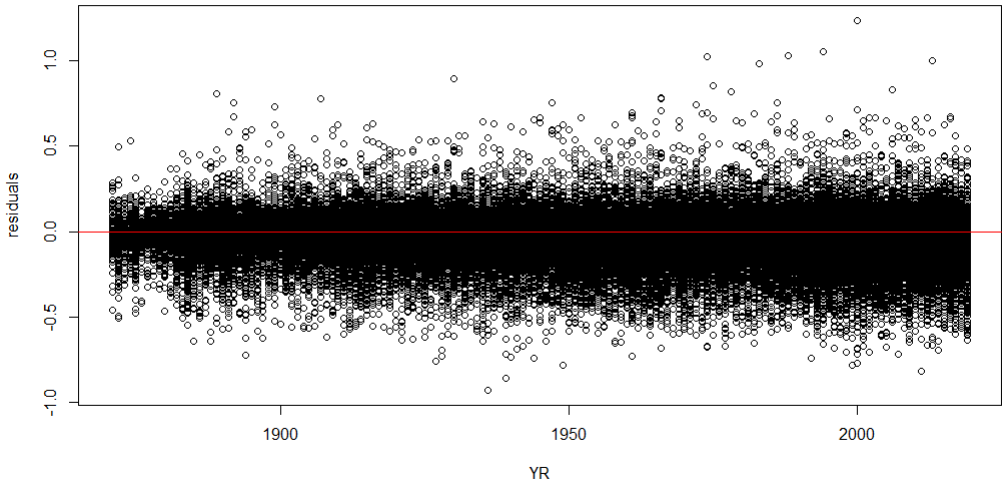

Figure 8: A residual plot based on the fit of model (10).

```
Models:
mixed1: OPS ~ lg.yr + (1 | Plyr)
mixed2: OPS ~ lg.yr + birth.yr + (1 | Plyr)
        Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
mixed1 284 -80831 -78222  40699   -81399
mixed2 285 -81300 -78682  40935   -81870 471.43      1  < 2.2e-16 ***
```

Our fit gives an estimate of $\hat{\gamma} = 0.0024$. This suggest that on average, hitting talent is improving over the years. As in the previous subsection, we plot the lg.yr effects for this model. See figure 9. We continue to see oscillations, but this time there is a general trend downward. Thus according to this model, players are getting progressively better at hitting over the years, and it is getting harder to maintain a high OPS in major league baseball over time.

3.6 Environment-Birthyr Mixed Model 2

Our third mixed model builds on the previous mixed model by assigning a fixed intercept for each birth.yr. This is intended to accomodate the fact that the improvement of hitters being born from year to year may not be at a constant rate.

$$OPS_{ij} = \lambda_{ij} + \gamma_k + p_k + \varepsilon_{ijkl} PA_{ijkl}^{-1/2},\ p_k \sim N(0, \sigma_p),\ \varepsilon_{ijk} \sim N(0, \sigma^2) \tag{11}$$

The intercept for the ith league in season j is represented by λ_{ij}. The intercept associated with the hitting ability for players born the same year as player k is represented by γ_k. As before, p_k are the normally distributed player effects. We fit (11) using the lahman data in `FitMixedEffModel-MLB.R` as `mixed3`. The computer takes about 3 minutes to fit the model. We check the residual plot in figure 13. The residuals are similar to what we saw in the fit of model (10).

We check the quality of fit using a likelihood ratio test compared to the fit in model (10). The improved fit obtained by treating birth.yr as a factor is significant.

```
Data: hit
Models:
mixed2: OPS ~ lg.yr + birth.yr + (1 | Plyr)
mixed3: OPS ~ lg.yr + birth.yrf + (1 | Plyr)
        Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
```

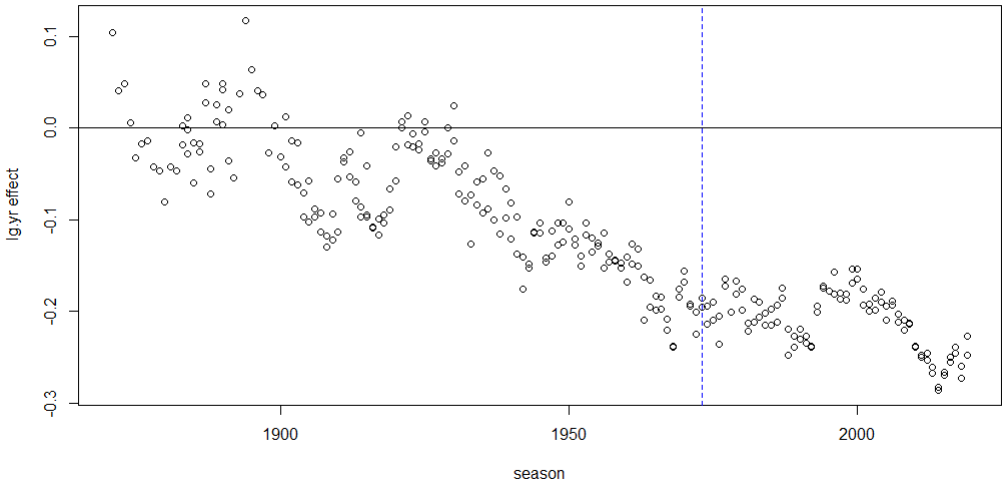

Figure 9: A plot of the lg.yr effects estamated by fitting model (10).

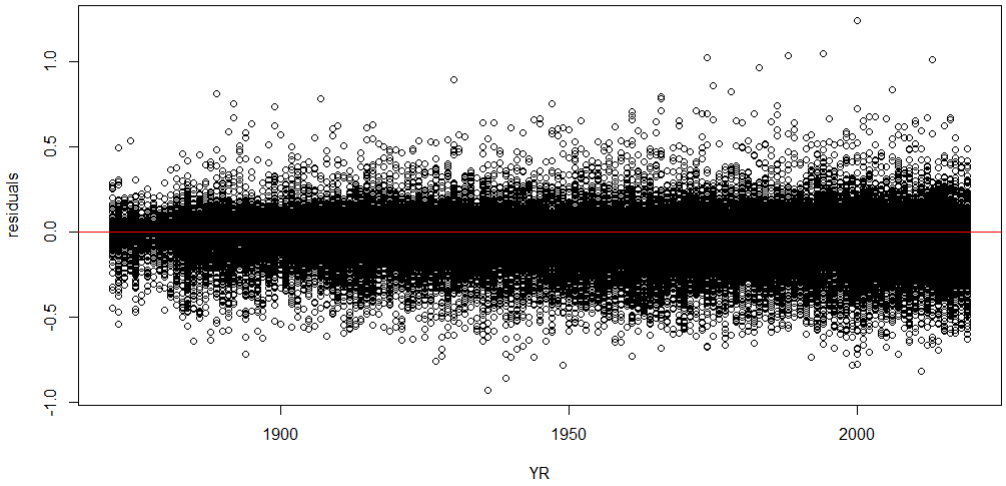

Figure 10: A residual plot based on the fit of model (10).

```
mixed2 285 -81300 -78682  40935   -81870
mixed3 446 -81194 -77097  41043   -82086 215.5     161   0.002672 **
```

For future reference, we save the information on the variance associated with the random effects of this model in the output below:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Plyr     (Intercept) 0.016152 0.12709
 Residual             0.002271 0.04765
Number of obs: 72087, groups:  Plyr, 13175
```

As in the previous subsection, we plot the lg.yr effects for this model. See figure 11. We see basically the

same oscillations and downward trend as in the previous model. Hence this model agrees wtih the previous mixed model that it is getting harder to maintain a high OPS in major league baseball over the years. We also present a scatter plot of the estimated birth year effects on OPS in figure 12. Notice that there is a general trend of improvement in hitting talent over the years. The plot also suggests that players born in the years 1998-1999 are amazing hitters. These estimates are inflated since those young players are only getting chances to play in the MLB because they are so talented. As the rest of the players born the same year break into the majors, the estimate of talent born that year will regress back to the normal trend.

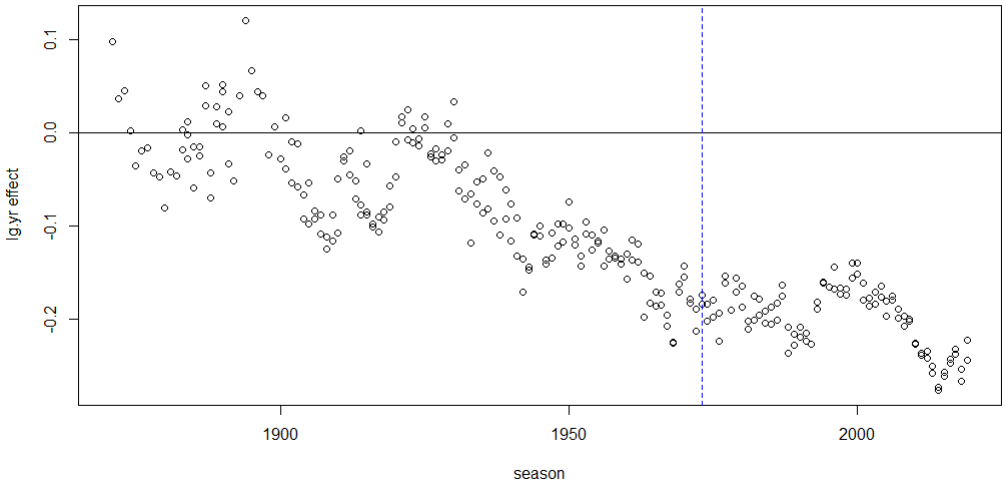

Figure 11: A plot of the lg.yr effects estimated by fitting model (11).

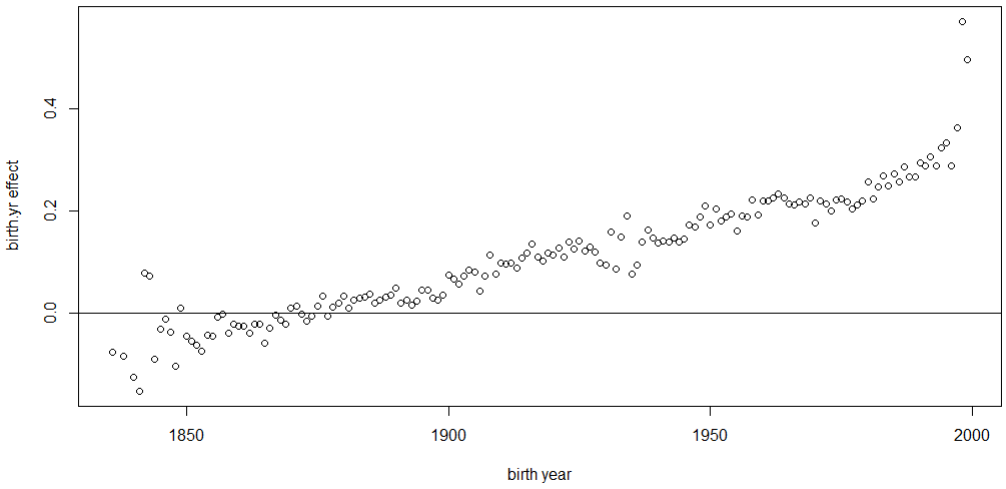

Figure 12: A plot of the birth.yr effects estimated by fitting model (11).

3.7 Environment-Birthyr Mixed Model 3

Our fourth mixed model builds on the previous mixed model by adding random effects for tm.yr. This is a factor assigned by pasting together the factors `TM`, and `YR`. The goal is to account for variance that occurs from players that play on the same team during a given season. Those observations should be correlated somewhat because they face the same opponents on the same days.

$$OPS_{ij} = \lambda_{ij} + \gamma_k + p_k + t_{sj} + \varepsilon_{ijkl} PA_{ijkl}^{-1/2},\ p_k \sim N(0, \sigma_p),\ t_{sj} \sim N(0, \sigma_t),\ \varepsilon_{ijk} \sim N(0, \sigma^2) \qquad (12)$$

The intercept for the ith league in season j is represented by λ_{ij}. The intercept associated with the hitting ability for players born the same year as player k is represented by γ_k. As before, p_k are the normally distributed player effects. The t_{sj} represent the random effects for playing on a team s during the jth season. We fit (12) using the lahman data in `FitMixedEffModel-MLB.R` as `mixed4`. The computer takes almost 6 minutes to fit the model. We check the residual plot in figure **??**. The residuals are similar to what we saw in the fit of model (12), although the magnitude is uniformly smaller.

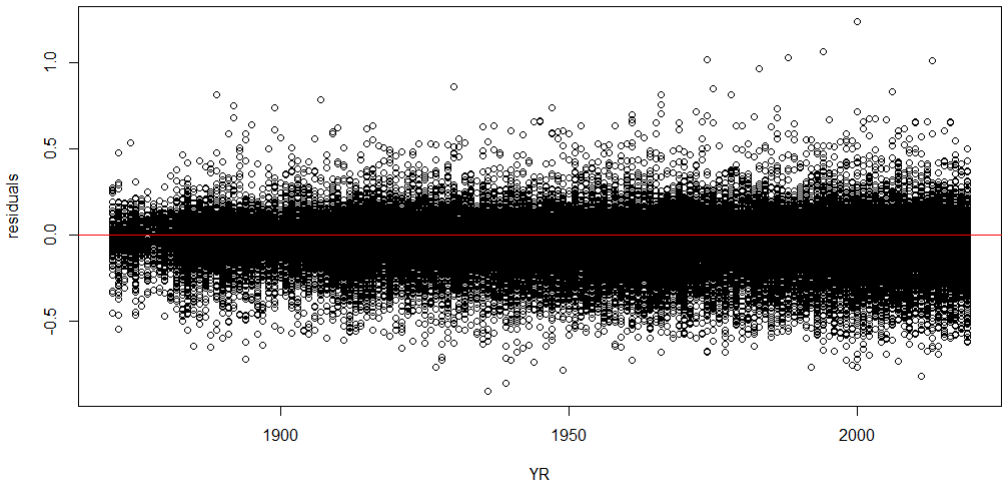

Figure 13: A residual plot based on the fit of model (12).

We check the quality of fit using a likelihood ratio test compared to the fit in model (11). The improved fit obtained by adding the tm.yr random effects is significant.

```
Data: hit
Models:
mixed3: OPS ~ lg.yr + birth.yrf + (1 | Plyr)
mixed4: OPS ~ lg.yr + birth.yrf + (1 | Plyr) + (1 | tm.yr)
        Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
mixed3 446 -81194 -77097  41043   -82086
mixed4 447 -81602 -77496  41248   -82496 410.46      1  < 2.2e-16 ***
---
```

As in the previous subsection, we plot the lg.yr effects for this model. See figure 14. We see basically the same oscillations and downward trend as in the previous model. Hence this model agrees wtih the previous mixed model that it is getting harder to maintain a high OPS in major league baseball over the years. We also present a scatter plot of the estimated birth year effects on OPS in figure 15. Again we see that only the best are representing the hitters born during 1998-1999.

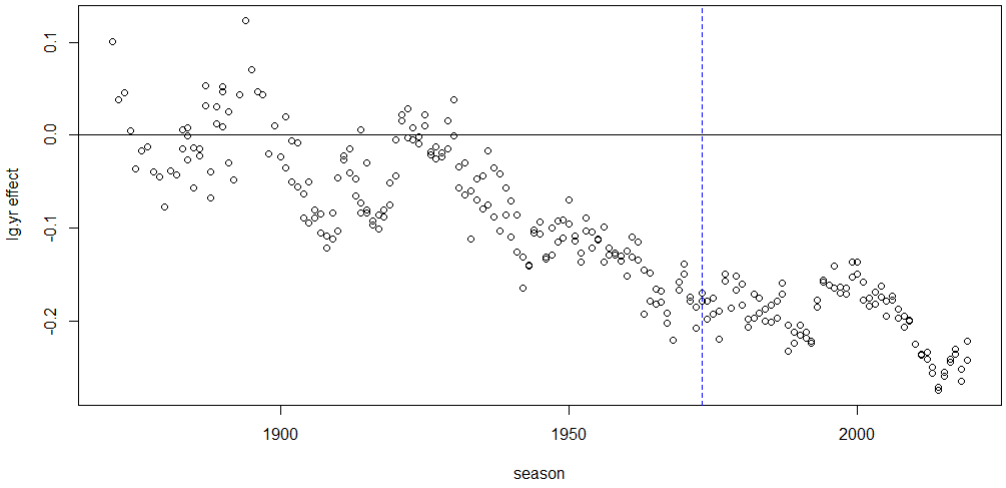

Figure 14: A plot of the lg.yr effects estimated by fitting model (12).

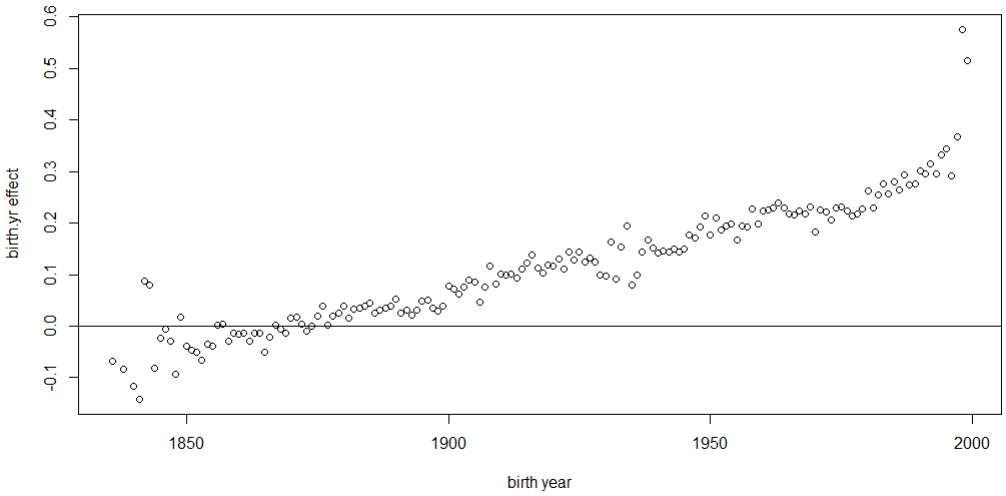

Figure 15: A plot of the birth.yr effects estimated by fitting model (12).

4 A Player and League Effects Model

Looking at the theoretical result in the previous section provides some insight about the variance for a linear model predicting OPS. The amount of variance for OPS under a given observation should be inversely related to the number of plate appearances. Thus we consider the model

$$OPS_{ij} = \mu + \rho_i + \lambda_{jk} + \varepsilon_{ijk} PA_{ij}^{-1/2}, \ \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{13}$$

Here, λ_{jk} is the league effect on OPS for league j in year k. The ρ_i denotes the effect of player i on OPS. One concern is that the player effect might also change over time, but for now we ignore this.

The additive model in (13) give us a natural way to quantify the the differences in league effects on OPS and the differences in average hitting ability between players of different leagues. First, suppose we would

like to know if the the league effect on OPS in the jth league has a significant change after the kth season. The null hypothesis associated with this question is

$$H_0 : \lambda_{jk} = \lambda_{j(k+1)}$$

We test this with a t-statistic. We could also compare how easy or difficult it is for hitters in different leagues accross the same season. The null hypothesis related to this question is

$$H_0 : \lambda_{j_1 k} = \lambda_{j_2 k} = ... = \lambda_{j_n k}.$$

This can be tested using an F-statistic. To do this we construct a matrix C, with $n-1$ rows such that

$$C\hat{\boldsymbol{\beta}} = (\lambda_{j_1 k} - \lambda_{j_2 k}, \lambda_{j_1 k} - \lambda_{j_3 k}, ..., \lambda_{j_1 k} - \lambda_{j_n k})'$$

where $\hat{\boldsymbol{\beta}}$ is the vector of parameter estimated from the model. The hypothesis is equivalent to

$$H_0 : C\boldsymbol{\beta} = \mathbf{0}.$$

To address questions about how talented the hitters are in a given league, we define the average hitting talent in league j during season k as

$$\Theta_{jk} = \frac{1}{n} \sum_{i \in L_{jk}} p_i.$$

This is just an average of the player effects for those players belonging to league j in the kth season. If we want to know if hitting talent in league j has changed between two consecutive seasons, we can test the null hypothesis

$$H_0 : \Theta_{jk} = \Theta_{j,k+1}$$

using a t-statistic. The final question we might ask is if the average hitting talent was different between the different Major Leagues in a given season. The null hypothesis associated with this question is

$$H_0 : \Theta_{j_1 k} = \Theta_{j_2 k} = ... = \Theta_{j_n k}.$$

Just as we did with league effects, this is easily tested with an F-statistic.

In the plot of hitting talent, during the seasons where there is a significant difference in average hitting talent among the leagues in a given season, the seasons are plotted with a "$*$".

If the null hypothesis is rejected, we connect the seasons with a solid line instead of a dashed line in the plot. This means the change in how difficult it was to hit in the league was significant between these two seasons. As with the question about league effects, we connect the seasons on the plot with a solid line instead of a dashed line. In the plot of league effects on OPS, during the seasons where there is a significant difference between the different leagues in a season, those seasons are plotted with a "$*$".

We can perform season to season comparisons by fitting model (13) to two seasons of data at a time. By repeating this for all seasons from 1871 to 2018, we can compare consecutive seasons.Below we plot the year to year league effects and comparisons of average hitting ability for the Major Leagues under this model. In the league effects plot, we add vertical lines for the following years that had significant rule changes that we might expect to have effected OPS uniformly for all players in a league:

- 1884 - Pitchers were allowed to deliver the ball overhand.
- 1889- The modern rule of 4 balls or 3 strikes is first used. In 1887 for example, players were allowed 4 strikes.
- 1893 - The pitcher is moved back from a pitchers box that is 50 ft from home plate to a rubber that is 60 ft 6 in from home.
- 1901 - Foul balls begin to count as strikes. This is first implemented in the NL and in 1903 it is also implemented in the AL.

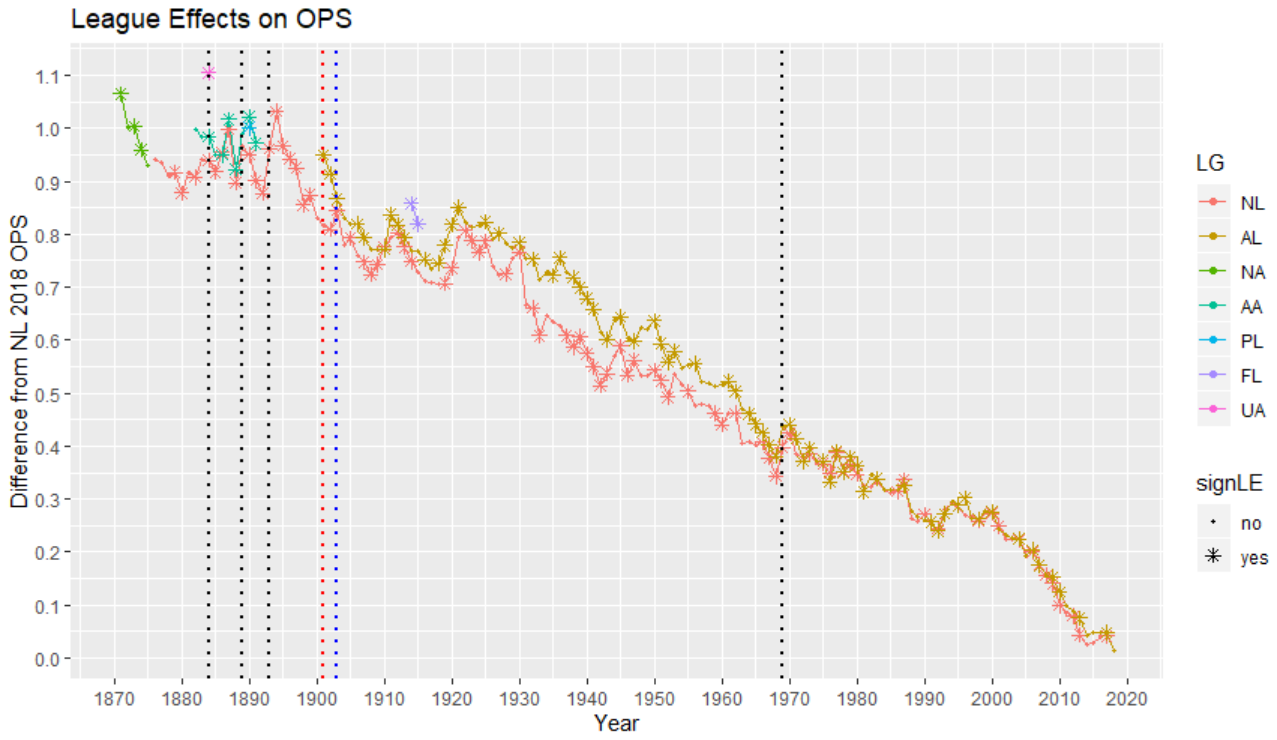

Figure 16: Year to year comparisons of league effects on OPS for the seven historical major leagues using model (13).

- 1969 - After an offensively frustrating season where Carl Yastrzemski lead the AL with a .301 BA, the mound was lowered from 15 to 10 inches high and the top of the strike zone is lowered from the top of the hitter's shoulders to his armpits.

To better understand how to interpret these statistics, look at the following output:

```
> d.comps[d.comps$start.yr == 1972,c(3:5,7:8,10:11)]
         LG1     LG2        LG.ef     LE.pval     Tal.diff      TD.pval comn.plyrs
1081 NL.1972 NL.1973 -0.009365313 0.214909632 -0.010504143 1.382065e-01        218
1082 NL.1972 AL.1973 -0.017349391 0.345697916 -0.072911588 5.635844e-05         30
1083 NL.1972 AL.1972  0.006797458 0.719799481  0.007182892 7.049179e-01          9
1084 NL.1973 NL.1972  0.009365313 0.214909632  0.010504143 1.382065e-01        218
1085 NL.1973 AL.1973 -0.007984077 0.672203915 -0.062407445 7.069133e-04         12
1086 NL.1973 AL.1972  0.016162772 0.402741540  0.017687035 3.564599e-01         20
1087 AL.1973 NL.1972  0.017349391 0.345697916  0.072911588 5.635844e-05         30
1088 AL.1973 NL.1973  0.007984077 0.672203915  0.062407445 7.069133e-04         12
1089 AL.1973 AL.1972  0.024146849 0.002500686  0.080094480 0.000000e+00        165
1090 AL.1972 NL.1972 -0.006797458 0.719799481 -0.007182892 7.049179e-01          9
1091 AL.1972 NL.1973 -0.016162772 0.402741540 -0.017687035 3.564599e-01         20
1092 AL.1972 AL.1973 -0.024146849 0.002500686 -0.080094480 0.000000e+00        165
```

The columns `LG1` and `LG2` indicate the two leagues being compared. The column `LG.ef` indicates the estimated difference in the league effects on OPS and `LE.pval` indicates the p-value associated with this estimate. The column `Tal.dff` indicates the estimated difference in the average hitter's effect on OPS and `TD.pval` is the associated p-value. The first row tells us that the 1972 NL was more difficult to hit in league wide by about 0.009 OPS although this difference is not statistically significant. Similarly the average hitter's effect was smaller by 0.011 OPS and this difference was not statistically significant. If anything

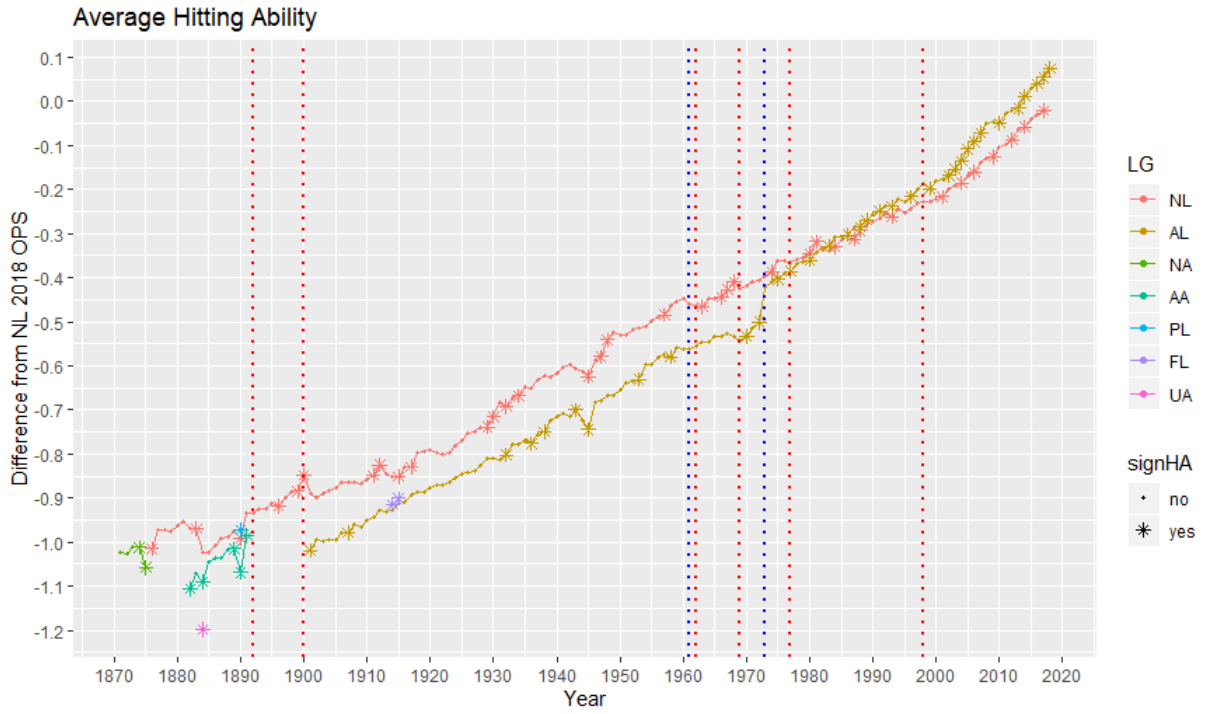

Figure 17: Year to year comparisons of average hitting ability based on OPS for the seven historical major leagues using model (13). Red vertical lines indicate a change in the number of NL teams. Blue vertical lines indicate a change in the number of AL teams.

changed in the NL between these two years, the evidence says that the averge hitter got slightly better in 1973. The last row of the output suggests that things also got easier for hitters in the AL between 1972 and 1973 by about 0.024 OPS which was a significant difference. On the other hand, the ability of the average hitter is estimated to have significantly improved by 0.080, which is probably due to the designated hitter rule.

Comparing these results to the results in Cramer's study, we might be a bit uneasy. The steps that we obtain using our model are larger. There are some key differences that add some confusion. Cramer bases his comparisons on batter win average, which is not defined in his article, and when he makes a season to season comparison, he only uses players who play in both leagues. In order to better compare our method with Cramer's we perform the same comparisons as in model (13) but with BA as the output instead. I have also changed the column `rel.HA` to represent the average hitting talent in the league compared to the NL in 1976, just as in the table provided by Cramer. Notice our difference in batting average for the NL in 1960 is -0.0396 where Cramer estimated it to be -0.004, and for the AL it was -0.0943 where Cramer's estimate was -0.017. If we go all the way back to the first NL season, 1876, we get a difference of -0.2951, where Cramer came up with -0.123. Our number suggests that the average player from the first season of the NL would almost never get a hit in the 1976 season. This might be possible considering the 1876 rules still required the pitcher to deliver the ball underhanded and players did not wear gloves in the field, but it still seems like a stretch.

```
LG   YR       rel.HA     HA.pval LG   YR       rel.HA      HA.pval
NL 1979  0.007498388 0.214128013 AL 1979 -0.001443420 3.805310e-01
NL 1978  0.005015233 0.281706810 AL 1978 -0.003056244 4.201078e-01
NL 1977  0.002595796 0.397728412 AL 1977 -0.010337636 1.522183e-03
NL 1976  0.000000000 0.327069418 AL 1976 -0.015861695 2.211858e-02
NL 1975 -0.001332416 0.582842733 AL 1975 -0.020554222 7.395716e-02
NL 1974 -0.009789343 0.002410404 AL 1974 -0.024711244 6.975659e-02
```

```
NL 1973 -0.013974676 0.146479945 AL 1973 -0.030179950 3.652036e-02
NL 1972 -0.020028541 0.018183812 AL 1972 -0.056184474 1.110223e-15
NL 1971 -0.024206956 0.094175827 AL 1971 -0.064344813 3.612706e-03
NL 1970 -0.026907561 0.306337503 AL 1970 -0.071208198 1.490615e-02
NL 1969 -0.029372613 0.381022898 AL 1969 -0.077498498 1.931780e-02
NL 1968 -0.023853880 0.028040788 AL 1968 -0.076551146 7.146779e-01
NL 1967 -0.032935080 0.003328016 AL 1967 -0.074191418 4.035065e-01
NL 1966 -0.038655497 0.048057140 AL 1966 -0.078228096 1.072795e-01
NL 1965 -0.038101694 0.849250471 AL 1965 -0.080221265 5.160448e-01
NL 1964 -0.037562652 0.849086529 AL 1964 -0.085551282 8.699337e-02
NL 1963 -0.044375507 0.025692892 AL 1963 -0.087909936 4.275607e-01
NL 1962 -0.043614973 0.782429847 AL 1962 -0.090259058 3.891868e-01
NL 1961 -0.041620493 0.530123887 AL 1961 -0.091816095 6.255709e-01
NL 1960 -0.039595488 0.543317632 AL 1960 -0.094259275 4.390831e-01
```

It might be that there are not enough observations in two seasons to accurately assess the effect of each hitter. We relax the assumptions that player effects change every season and fit the model using every season in the Lahman data. Each player is assigned a parameter ρ_i. The league effects λ_j are indexed by both the league and the year. This is done by creating a new factor, `lg.yr`, by pasting the name of the league and the year for each observation. This fit took just under 50 minutes. The results seem a little more reasonable since the range of league effects and player effects is smaller, close to 0.6. In figure 18 we plot the league effects relative to the 1976 season of the NL, and in figure 19 we plot the estimated hitting ability relative to the 1976 season of the NL. It is noteworthy that according to this new fit, the AL had more hitting talent than the NL even in its first decade. The jump in 1973 due to the DH rule is also more distinct, but it is really about the same size as in the previous fit.

An interesting observation from figure 19 is that it is more common for the difference between leagues

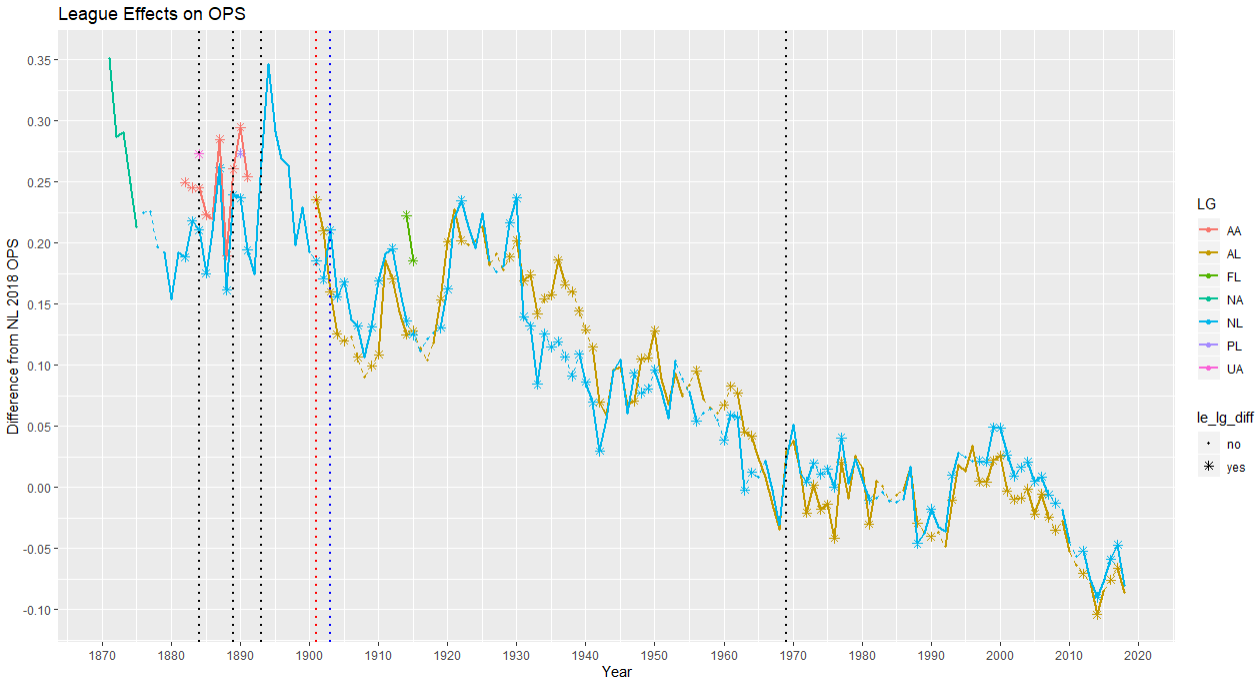

Figure 18: Year to year comparisons of league effects on OPS for the seven historical major leagues using model (13) fit with all seasons of Lahman data at once.

to be significant during a given season than for the average talent to change significantly between two sea-

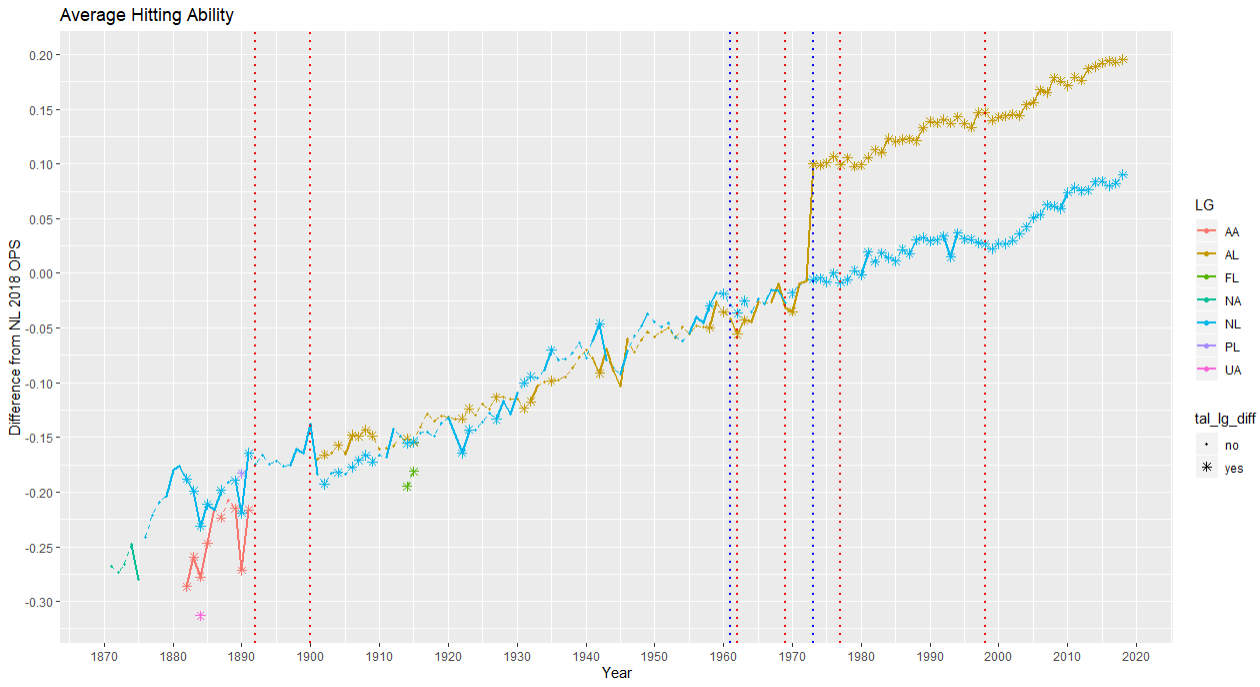

Figure 19: Year to year comparisons of average hitting ability based on OPS for the seven historical major leagues using model (13). This fit uses all seasons of Lahman data at once.

sons. Perhaps this makes sense since great hitters typically have longer careers than great pitchers. Great pitchers are not always great for their entire career either. They might get hurt or the league just figures out how to handle them. Pitching ability is probably a significant factor in determining the league effects on OPS, so this quantity should be less stable season after season. You can see how the DH rule produced a significant change in average hitting talent for the AL. According to the model, if we estimate the average hitting talent in AL 1973 minus the average hitting talent in AL 1972 we get a 95% confidence interval of (0.0953, 0.1189).

5 Modeling OPS with College Hitters

We now look to design a linear mixed effects model to predict OPS for the college baseball data provided by the Baseball Cube. This model should be similar to the models we fit in section 3 using MLB data from the Lahman data set. Here we first take the time to describe the college baseball data. This exploratory work is saved in the script `ExpHittingData.R`. Our data for college hitters is contained in two spread sheets stored in the folder `ComparisonsWithDI`.

The first spreadsheet contains data for NCAA Division I hitters. The NCAA data contains $n_1 = 57859$ observations of 45 variables. We provide a summary of this data frame in table 2. The variables included in the NCCA data are:

```
> names(ncaa0)
 [1] "playerid"     "year"         "teamName"     "LeagueAbbr"   "leagueName"
 [6] "Level"        "lastName"     "firstName"    "G"            "AB"
[11] "R"            "H"            "Dbl"          "Tpl"          "HR"
[16] "RBI"          "SB"           "CS"           "BB"           "IBB"
[21] "SO"           "SH"           "SF"           "HBP"          "GDP"
[26] "Bavg"         "Slg"          "obp"          "OPS"          "Age"
[31] "HT"           "WT"           "Bats"         "Throws"       "posit"
[36] "borndate"     "Place"        "hsname"       "hsplace"      "mlbid"
```

```
[41] "draft_year"    "draft_Round"   "draft_overall" "Draft_Team"    "X"
```

For nearly all the observations in NCAA, 57723 out of 57859, the players are given a non-zero player-id number. The observations with player-id of 0 are useless because the do not include player names. They could help compute league averages, but do not work for fitting our mixed effects models. We remove observations with no assigned non-zero player-id. We use ddply to collect biographical information on the remaining players and store this in the data frame `players_ncaa`. We can observe that of the 26632 NCAA hitters we have, only 1035 of them play for multiple schools during their career. We save players data as `D1_hitter_bios.csv`. We also save the cleaned up hitting data as `D1Batting_2010-2020.csv`.

In the summer data set, we have 43534 observations, but we immediately remove thos from before 2005 because it is unlikely that there will be any common hitters from the NCAA data which does not begin until 2010. This leaves us with $n_2 = 42648$ observations of 44 variables. The variables included in the summer data are:

```
> names(Summer0)
 [1] "playerid"       "year"           "teamName"       "LeagueName"     "lastname"
 [6] "firstname"      "G"              "AB"             "R"              "H"
[11] "Dbl"            "Tpl"            "HR"             "RBI"            "SB"
[16] "CS"             "BB"             "IBB"            "SO"             "SH"
[21] "SF"             "HBP"            "GDP"            "Bavg"           "Slg"
[26] "obp"            "OPS"            "Age"            "HT"             "WT"
[31] "Bats"           "Throws"         "posit"          "borndate"       "place"
[36] "colleges"       "draft_Year"     "draft_Round"    "draft_overall"  "draft_teamabbr"
[41] "status"         "teamname"       "CurrentTeam"    "X"
```

Once the earlier observations are removed we find that there 27856 of these observations include nonzero player id numbers. 22443 of these ids match ids in our collection of NCAA players. We narrow our focus to only seasons with these players. To identify these seasons, consider table 1. Based on this table, we elect to consider only the 2008 and later seasons of the summer data. It is possible that some of the unidentified hitters in the summer data could be tied to known players.

Season	# DI players
2006	1
2007	22
2008	65
2009	162
2010	2007
2011	1944
2012	1920
2013	1920
2014	2262
2015	2263
2016	2637
2017	2492
2018	2380
2019	2368

Table 1: The number of of players with ids in the NCAA data set that apper during each season of the summer league data.

We deal with the unidentified summer hitters by first trying to connect observations from the unidentified summer hitters. We collect all the unique first and last name combinations of the unidentified players. For each name, we look at the numer of observations

Looking at the sets of unique players in the NCAA and summer data sets, we found 13341 common players.

NCAA					Summer
#Conferences	#Teams	Seasons	#Leagues	#Teams	Seasons
35	310	2010 - 2020	28	409	1996-2019

Table 2: A summary of the two data frames in the college baseball data.

Year	Team	League	last name	first name
2015	Geneva Red Wings	New York Collegiate	Rodriguez	Alex
2017	Valley Blue Sox	New England Collegiate League	Rodriguez	Alex
2017	Texarkana Twins	Texas Collegiate League	Rodriguez	Alex
2017	Concord Athletics	Southern Collegiate League	Rodriguez	Alex
2017	Riverside Bulldogs	Southern California League	Rodriguez	Alex
2018	North Adams SteepleCats	New England Collegiate League	Rodriguez	Alex
2018	Savannah Bananas	Coastal Plain League	Rodriguez	Alex
2018	Academy Barons	California Collegiate League	Rodriguez	Alex

Table 3: This is an example of multiple observations of the same player's name appearing multiple times in one season.