Case Study #1

Il Shan Ng

Professor Laura Chihara
Math 315: Advanced Statistical Modeling

May 1 2016

1 Introduction

The Colorado Rockies, a Major League Baseball team, started an experiment on June 20th 2012. Under the belief that restricting the number of pitches would train its pitchers to perform better, the team decided to impose a limit of 75 pitches per starting pitcher until the end of the year. In this case study, I construct a hierarchical model for pitcher performance in order to assess whether the pitch count limit had any effect on performance.

2 Exploratory Data Analysis

Here, I use 2012 data on Rockies pitchers collected by a team of students (Lampert, Friedrich, Sturz) from FanGraphs. The data set consists of 118 observations, each representing one game, made of 7 different pitchers. To measure pitcher performance during a game, I have chosen average fastball velocity (vFA) as the response variable. Because the pitch count limit was intended to conserve arm strength and produce better pitches, vFA should give a good indication of whether or not the limit had any effect. The main explanatory variable of interest is whether or not the pitch count limit was in effect when the game was played (PCL). Other explanatory variables that should be controlled for are whether or not the game was played at home on Coors field (Coors), and the age of the pitcher (Age).

Table 1 gives summary statistics on the numeric variables vFA and Age. To avoid model fitting problems, I centered the Age variable by subtracting the mean from individual values. The binary variables PCL and Coors have sufficient counts in each category (PCL: 75 yes, 43 no; Coors: 57 yes, 61 no) to allow for accurate modeling.

Table 1: Summary statistics for numeric variables

Variable	Min	Max	Mean	SD
vFA	83.6	94.4	90.2	2.92
Age	24	33	26.9	3.58

Figures 1 and 2 show the distribution of the response grouped by PCL and Coors. Unfortunately for the Rockies, Figure 1 suggests that average fastball velocity tended to be lower when the pitch count limit was in effect. The response also shows a lot more variability under the pitch count limit. Figure 2 suggests that the distribution of average fastball velocity is not affected by whether or not the game was played on Coors field, and so the Coors variable may end up being insignificant.

Figure 1: Boxplot of vFA grouped by PCL

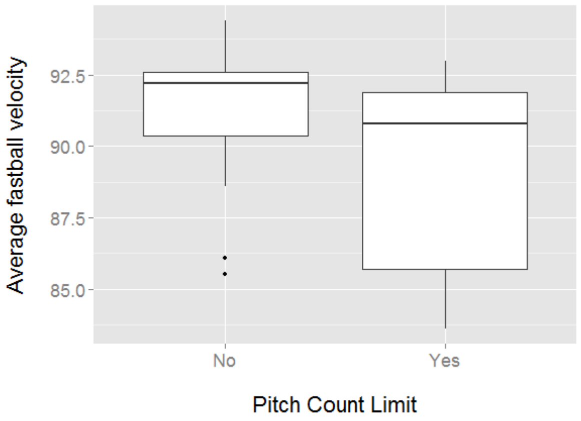


Figure 2: Boxplot of vFA grouped by Coors

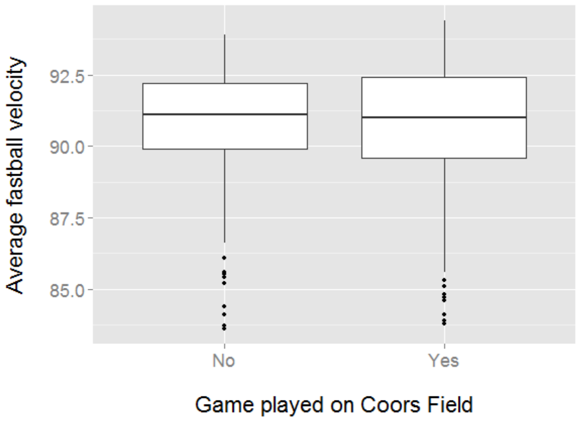
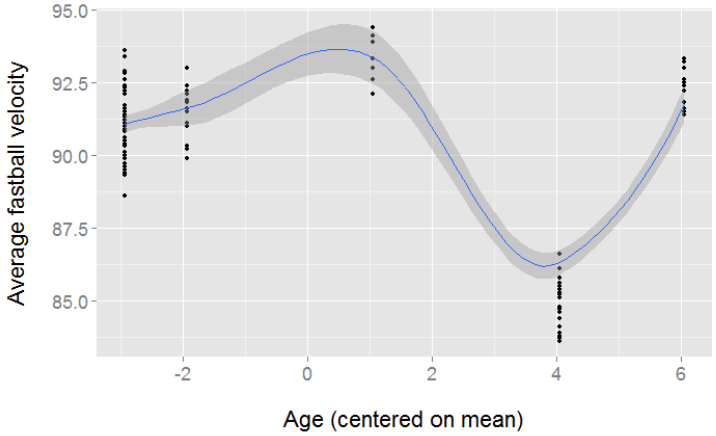


Figure 3 shows the relationship between the response and Age. To take into account the obvious curvature, I will include both a quadratic and cubic term in Age.

Figure 3: Scatter plot of vFA against Age



3 Results

To take into account the correlation among average fastball velocity measurements taken from the same pitcher, I created a two level model, shown below. As expected, the Coors variable was insignificant and so could be left out. I also found all possible interaction terms to be unnecessary.

Level one (pitcher i, game j):

$$\text{vFA}_{i,j} = a_i + b_i \text{PCL}_{i,j} + \epsilon_{i,j}$$

Level two (pitcher i):

$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 \text{Age}_i + \alpha_2 \text{Age}_i^2 + \alpha_3 \text{Age}_i^3 + u_i \\ b_i &= \beta + v_i \end{aligned}$$

where $\epsilon_{i,j} \sim N(0, \sigma^2)$ and $\left[\begin{array}{c} u_i \\ v_i \end{array}\right] \sim \text{MVNorm}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \left[\begin{array}{c} 0 \\ 0 \end{array}\right]$ and $\boldsymbol{\Sigma} = \left[\begin{array}{cc} \sigma_u^2 & \sigma_{u,v} \\ \sigma_{u,v} & \sigma_v^2 \end{array}\right]$.

Tables 2 and 3 give the estimates for the fixed effect parameters and variance components. The variable PCL, with a small t-value of -0.49, is not significant, but had to be left in the model because the random slope for PCL was needed.

Table 2: Estimates for fixed effect parameters

Parameter	Notation	Estimate	Standard error	t-value
Intercept	α_0	93.51	1.73	54.02
Age	α_1	-1.787	0.536	-3.34
Age^2	α_2	-0.528	0.209	-2.53
Age^3	α_3	0.128	0.041	3.09
PCL	β	-0.251	0.511	-0.49

Table 3: Estimates for variance components

Component	Notation	Estimate
SD of random intercept	σ_u	2.21
SD of random slope	σ_v	1.27
Correlation	$\rho_{u,v}$	-0.653

Diagnostic procedures flagged no outliers and revealed no major problems with the model. The estimates above are therefore reliable.

4 Discussion

As mentioned earlier, the fixed slope for PCL ended up being insignificant. Although the EDA showed an obvious decrease in the mean response under the pitch count limit, the sheer increase in variability was captured by the model as a highly significant random slope for PCL. Thus, controlling for age and the correlation between measurements taken from the same pitcher, my model shows that the pitch count limit had no significant impact on pitcher performance.

One may contend that average fastball velocity is not the best measure of pitcher performance during a game, and that other variables, such as earned runs per nine innings (ERA), would make a better response variable. I have tried to model ERA, but encountered too many scaling and model fitting issues. Hence, given the limitations of the data, I chose the response variable that was least likely to incur problems. In addition, the cubic term in Age was included in the model only to control for the effects of pitcher age. Its interpretation would make little sense, since it is unlikely for average fastball velocity to vary in a cubic fashion within a population of pitchers.