

The affect of body mass index (BMI) on baseball performance

October 15, 2008

Abstract

We examine the impact of body mass index (BMI) on performance in the sport of baseball. We show that high bmi is associated with performance gains in hitting, but is also associated with increased strikeouts and performance losses in base running. We also look for evidence that a shift in bmi has occurred over time.

↳ and what do you find?

nice abstract

1 Introduction

This paper examines whether the physical attributes of baseball players affects the way the sport is played. Specifically, we explore whether a player's body mass index (BMI) can help explain his success or failure on the field. We also explore whether trends in BMI over time can help illustrate changes in the way the sport has been played.

We will first look at the effect of physical attributes on batting performance. We want to see whether larger athletes perform better at the plate. Next we will examine the relationship between bmi and strikeouts. Then we will explore how weight and BMI affects a players performance on the bases. Do heavier players attempt to steal bases less often? Do they get caught more frequently when they do?

The final part of the paper will analyze the average BMI of baseball players' over time. We will try to find evidence that changes in BMI over time have paralleled known changes in the way the sport has been played.

The data we analyze has been collected by the Baseball Databank, a volunteer organization devoted to accumulating and disseminating baseball data. We combined the publicly available datasets `batting.csv` and `players.csv` to obtain height and weight information on 64715 professional baseball players. Our observations begin in 1924 and are compiled through 2007.

To complete this analysis, we created a number of important variables from the available data.

The first of these, body mass index (bmi) is a commonly used statistic to measure an individual's mass. It is calculated with the following formula:

$$\text{BMI} = \frac{\text{weight (lbs)} \times 703}{\text{height (in)}^2} \quad (1)$$

BMI provides a better description of an individual's physique than either height or weight alone. The higher an individual's BMI, the more massive he is compared to people of similar height. Large BMI's indicate above average amounts of fat or muscle. It is rare for a person to have a BMI ≥ 30 due to muscle alone. In the medical field, an individual with a BMI this large is considered to be clinically obese.

The other variables we create are familiar baseball statistics. The formulas we used to calculate them appear below.

↳ Table 1.

2 Body size and Hitting

A player's physique could affect his performance at the plate in many ways. It could help him:

when were height & weight measured?

I think you could combine these paragraphs into fewer bigger ones.

Variable	Formula	Description
slg	$slg = \text{total bases} \div \text{at bats}$	Slugging percentage
ba	$ba = \text{hits} \div \text{at bats}$	Batting average
sba	$sba = sb + cs$	Stolen base attempts
ob	$ob = h + bb - hr$	Number of times a player gets on base

using / is more common

Table is missing a label & caption

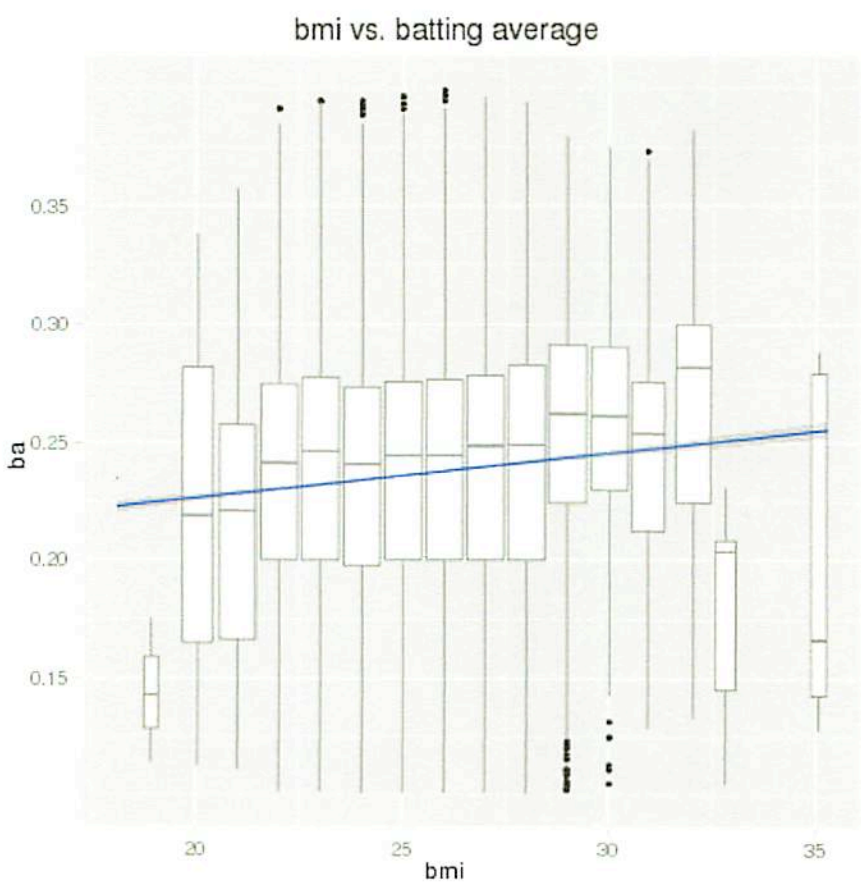
- successfully hit the ball
- propel the ball further when he does hit it

2.1 Hitting the ball

Prowess at hitting the ball is most commonly measured by a player's batting average (ba). Better batters get more hits, and hence have higher batting averages.

From the graph below, we see that bmi is positively associated with batting average (ba). Due to the large amount of data, we use boxplots to avoid overplotting. We also ignore outlying values of ba that result from players who had very few appearances at bat.

need more details



This looks like a png graphic - pdf would've been fine here.

More massive players appear to hit the ball successfully at a higher rate than smaller players. One reason

for this could be that massive players hit the ball further. To get a hit, a batter must not only strike the ball, but also arrive safely at first base before the ball is thrown there. Hitting the ball further gives the batter a larger head start and could inflate his number of hits.

This graph also illustrates a phenomenon that occurs across all of our findings. After a certain point, increases in BMI are no longer associated with performance gains. In fact, players with BMI's above 32 have much lower median batting averages than the other players. We will return to this topic at the end of the section.

2.2 Propelling the ball

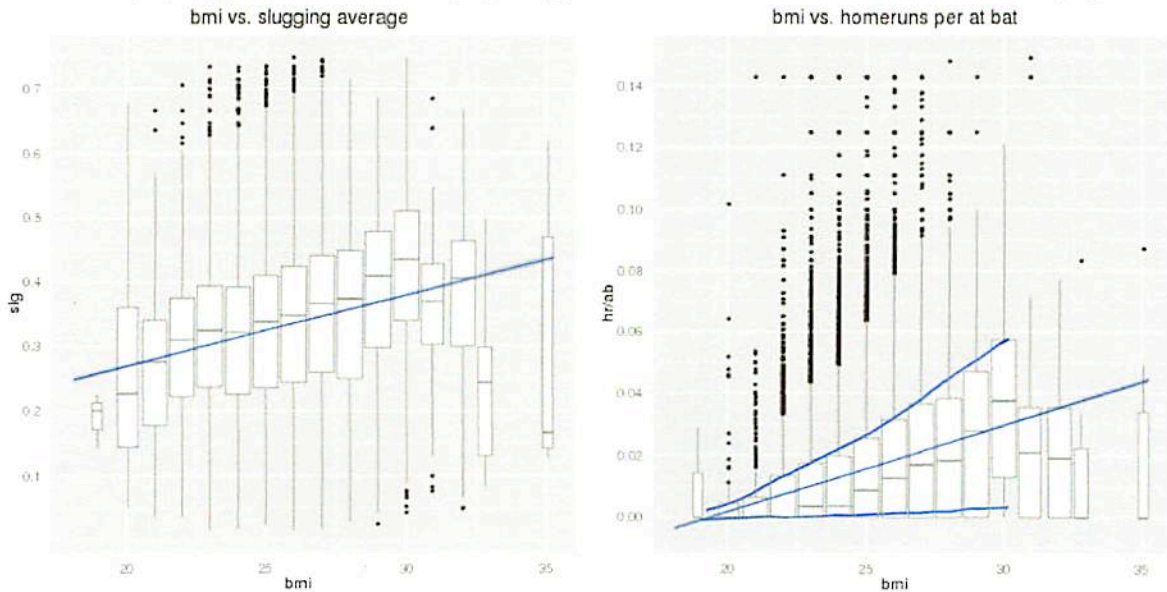
A batter's slugging percentage (slg) quantifies how far he hits the ball when he does hit it. It is calculated as follows:

$$\begin{aligned} \text{slg} &= \frac{\text{total bases}}{\text{at bats}} \\ &= \frac{1 \times \text{singles} + 2 \times \text{doubles} + 3 \times \text{triples} + 4 \times \text{homeruns}}{\text{at bats}} \end{aligned}$$

The number of homeruns a player hits is also an important indicator of how far he can hit the ball. A homerun is usually the most valued outcome of an at bat, and only players who can hit the ball out of the park will be able to hit them. To compare players' ability to hit homeruns, we must standardize the number of homeruns they hit (hr) by the number of times they appear at bat (ab).

As illustrated below, a strong positive relationship exists between bmi and both slugging percentage slg and homeruns (i.e., hr / ab). More massive players appear to hit the ball further than less massive players.

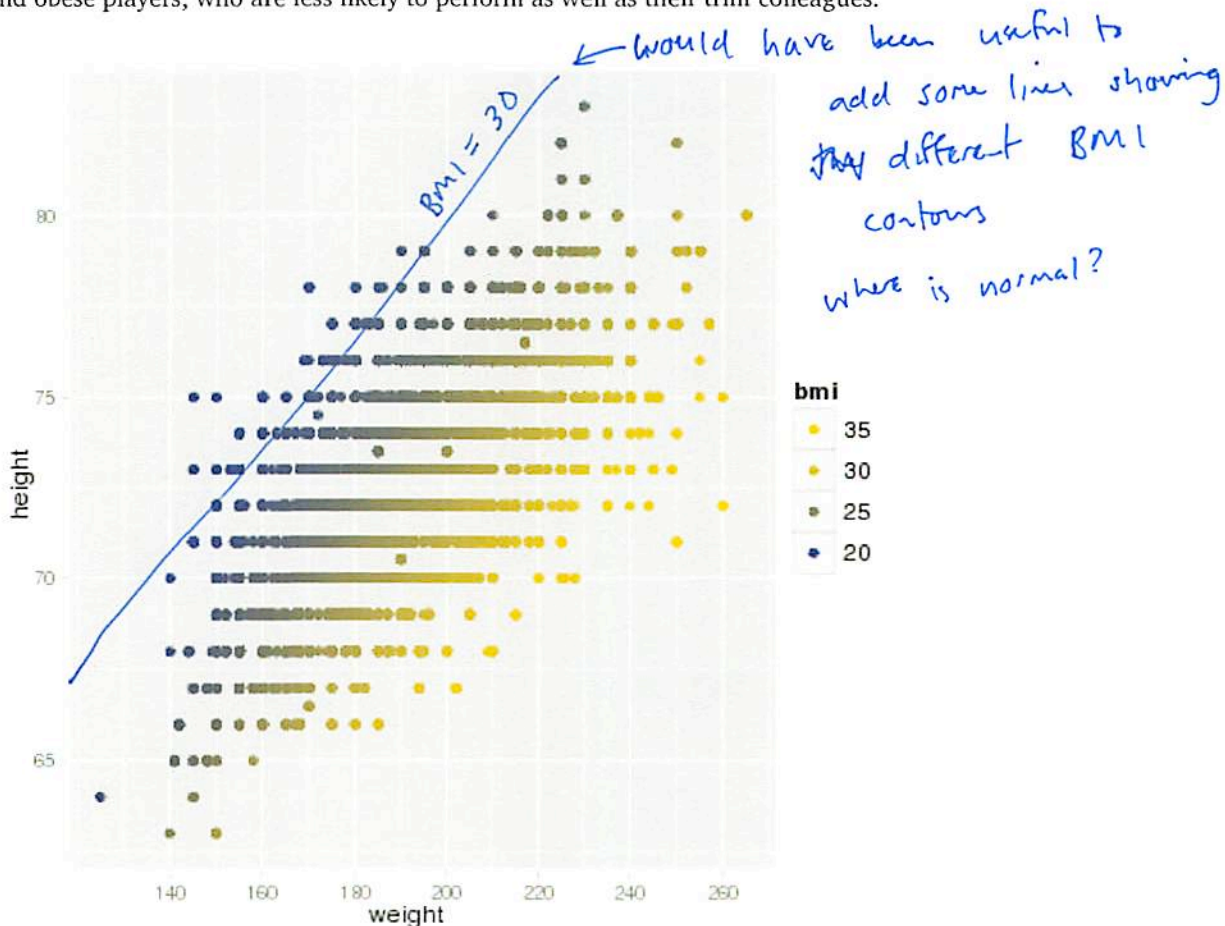
* variation increases!



Again we see a decrease in median performance for the highest BMI's. For BMI's under 30, an increase in BMI is associated with an increase in performance. However, for BMI's > 30, an increase in BMI is associated with a decrease in performance. This suggests that the relationship between BMI and performance is not strictly linear. — but how many players have bmi > 30?

What might be causing the trend to break down at these high values? Human weights vary more than human heights. Since the BMI is a ratio of weight vs. height, the largest BMI scores indicate individuals

who are much heavier than other players of their height (see graph below). These large values are capturing overweight and obese players, who are less likely to perform as well as their trim colleagues.



The increased negative performance associated with extreme BMI's will also be noticeable in regards to strikeouts and stealing bases.

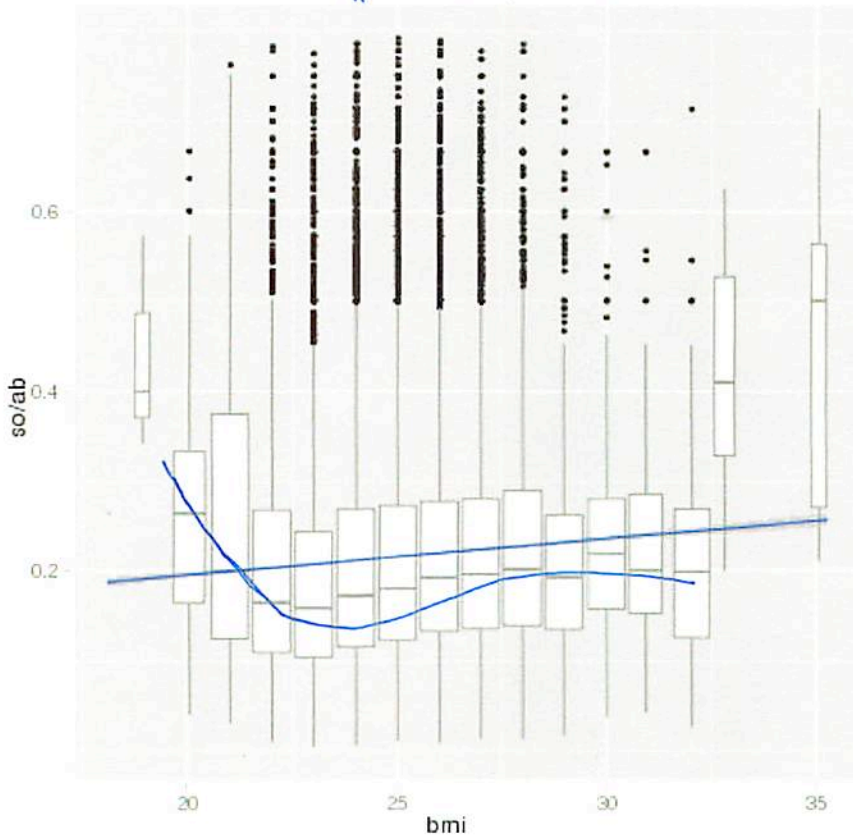
3 BMI and strike outs

Although a high bmi is associated with longer and more frequent hits, it is also associated with increased instances of striking out, as the graph below demonstrates. To compare the rates at which players struck out, we standardized their number of strikeouts (*so*) by the number of times they appeared at bat (*ab*). As with the other graphs, we first removed the players who only appeared at bat once or twice.

the graph shows something more complex!

(why not say) < 20 times?

average
bmi vs. strike outs per at bat

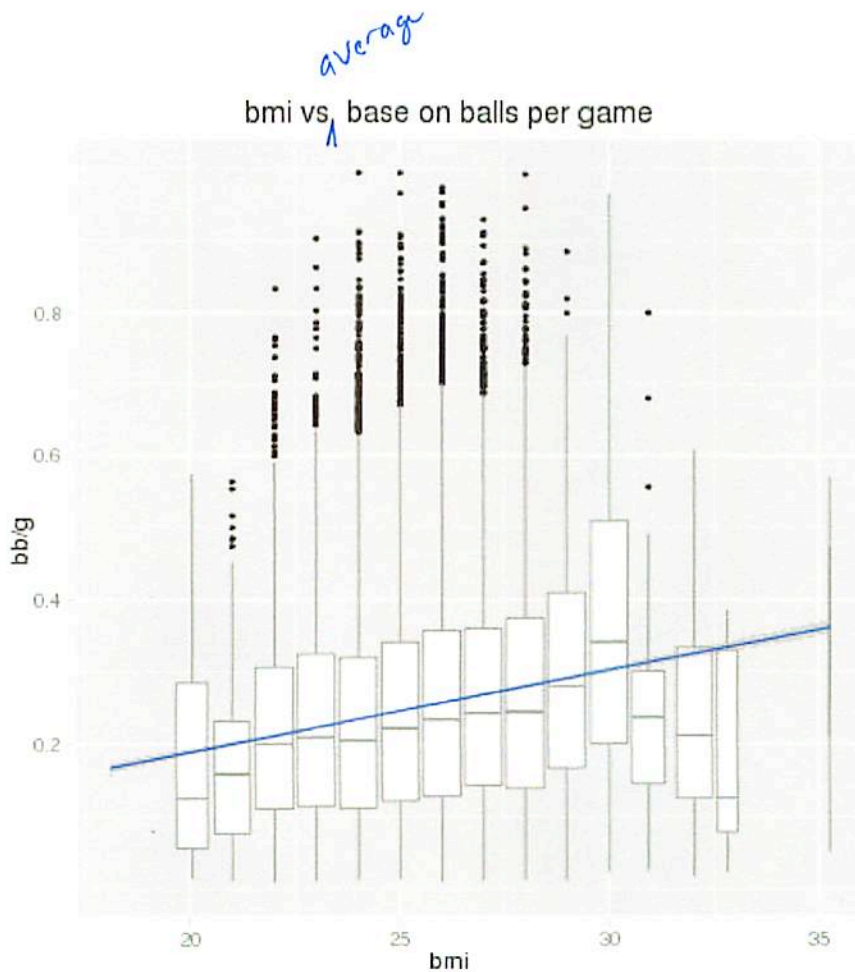


It may seem odd that massive players could have both higher batting averages and more strikeouts. However, strikeouts are not directly related to a player's number of hits. There are many other ways to receive an out. A player may ground out, be thrown out at first, hit a pop fly, etc. All of these cases prevent the player from being awarded a hit. Our findings seem to suggest that massive players minimize these types of outs. *—why?*

One reason massive players may frequently strike out is because they are more picky about which pitches to swing at. The previous sections suggest that a large player's strength is his ability to hit the ball far and score homeruns. If this is his goal, it makes sense for him to wait for a good pitch (instead of swinging at an unsure one just to get on base). This is especially true if a large player's size makes them weak at base running, which is the topic of the next section. *nice hypothesis*

A finding that supports this interpretation is the positive relationship between bmi and bases on balls bb (see below). A player is granted a base on balls if the pitcher throws four pitches that are outside the strike zone. If a player swings at one of these pitches and misses, the pitch is considered a strike and does not count towards the base on balls count. Hence, the association suggests that massive players have increased discipline when it comes to not swinging at unsure pitches. *interesting idea*

for all of these you should be referring to figures by number using {ref{—}}



4 Body size and base running

One way to quantify a player's confidence on base is to examine how often he tries to steal bases. A fast player who believes he can outrun the ball is likely to steal base more often than one who feels less sure of himself. Since a team's third base coach often decides who will steal and when, this metric also reveals how confident a manager is in a player's on base abilities.

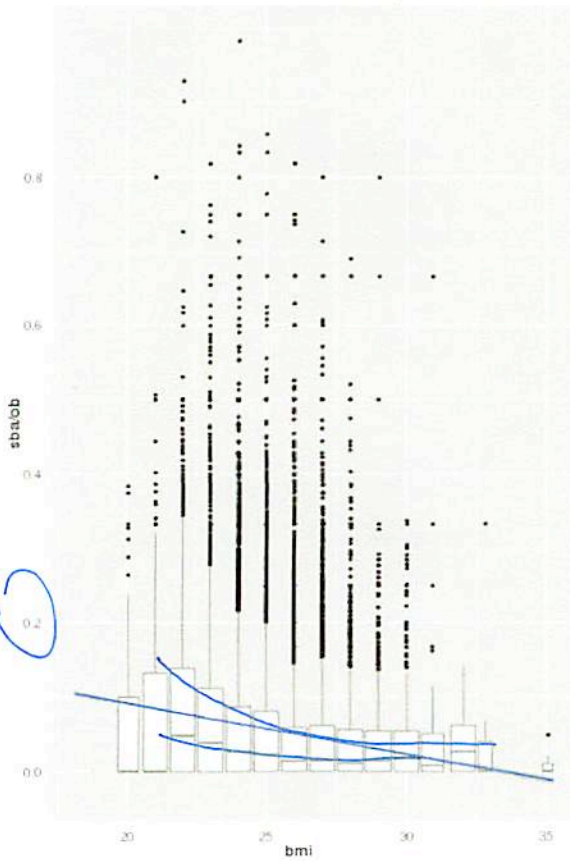
To compare how often each player attempts to steal, we standardize the number of times they attempt to steal by the number of times they have an opportunity to steal (i.e. the number of times they are left on base).

$$\text{Attempt ratio} = \frac{\text{stolen base attempts (sba)}}{\text{number of times on base (ob)}} \quad (2)$$

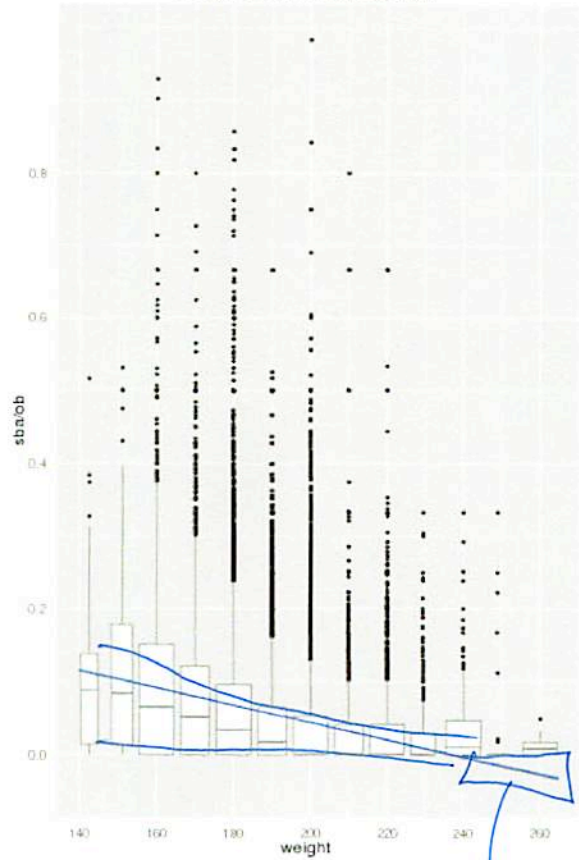
Players who have no hits are left out of this analysis for obvious reasons.

Our findings show that massive players attempt to steal base less often than lighter players. This affect is magnified if you just look at a player's weight. It seems that heavier and more massive players decide to steal base less frequently than lighter players.

bmi vs stolen base attempts



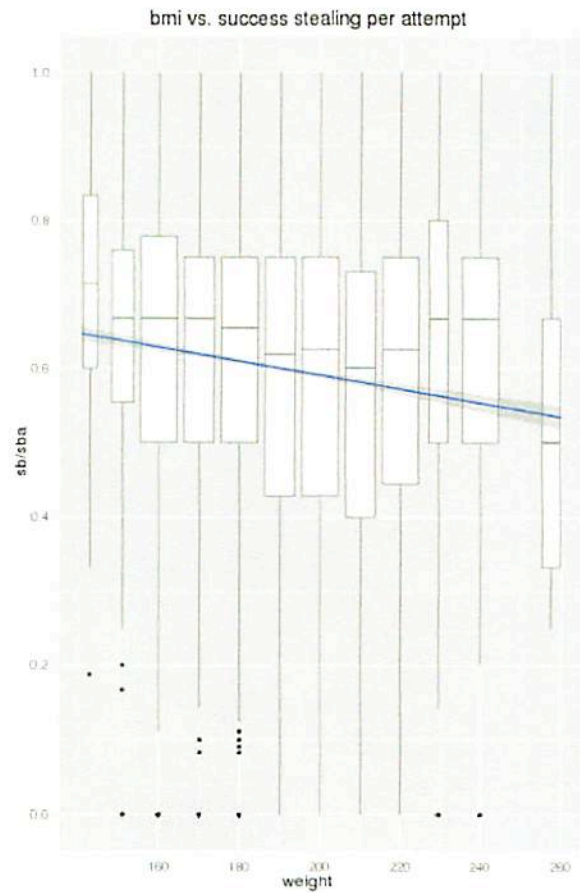
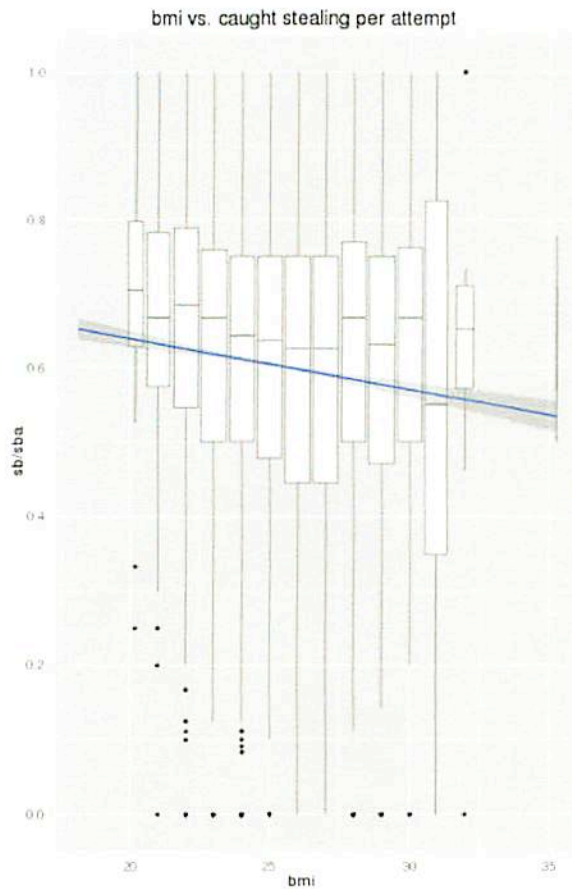
weight vs stolen base attempts



If we look at the success rate of players who attempt to steal bases, we see that these decisions are well founded. Heavier and more massive players succeed less often than lighter players when they attempt to steal base.

these numbers seem very high! many players are attempting to steal >40% of the time?!

a linear fit isn't very useful here!



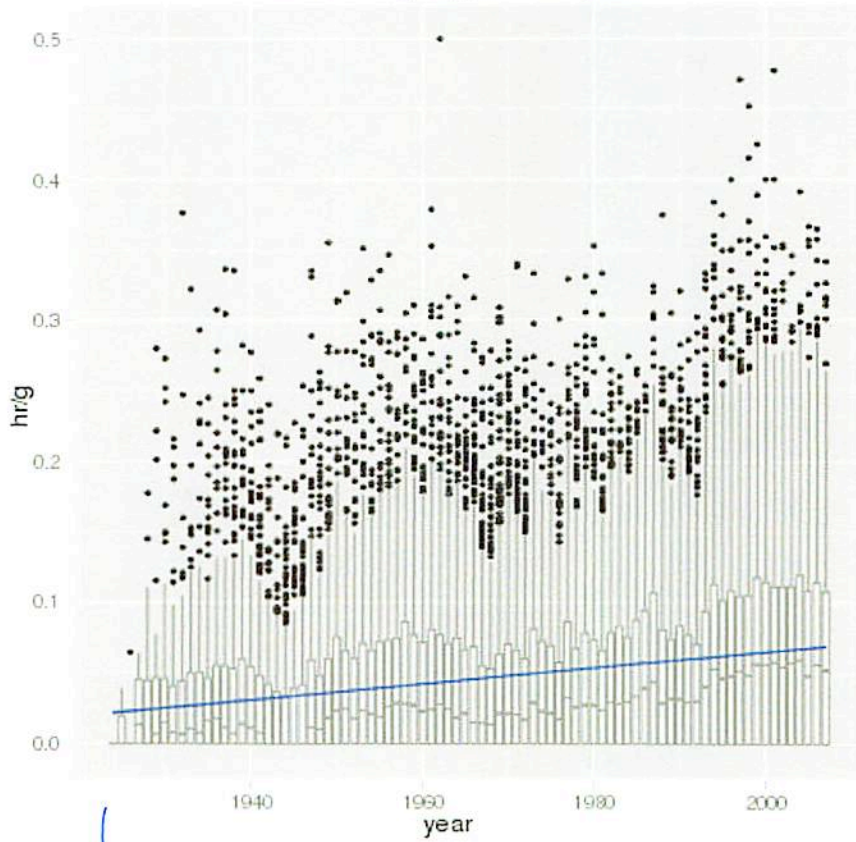
5 BMI over time

In the history of baseball, the years before 1920 are known as the "dead ball" years. Strategy during this period focused on base running abilities and power hitting was deemphasized. Power hitting gradually gained importance after 1920 and continues to grow as a primary strategy among baseball teams.

A plot of homeruns per game against year illustrates this trend towards powerhitting. The mean number of homeruns per game steadily increases over time.

how can you tell that from the plot?

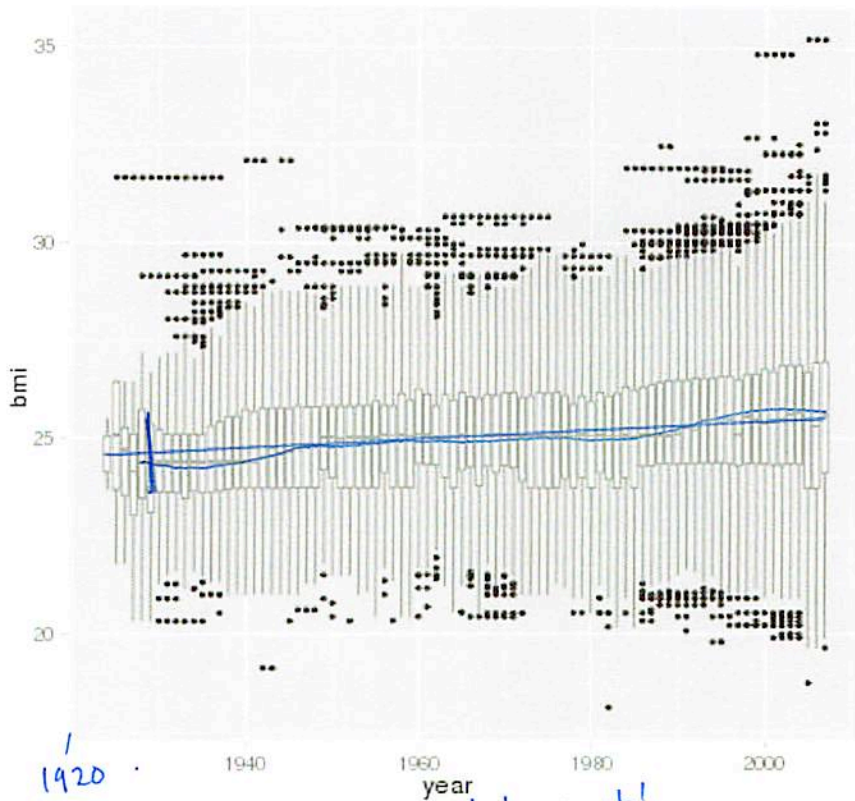
Homeruns per game vs. year



Since massive players appear better at power hitting and weaker at base running, it seems likely that baseball teams would have begun recruiting more massive players as the power hitting trend grew. Can we find evidence of this type of phenomenon? Yes.

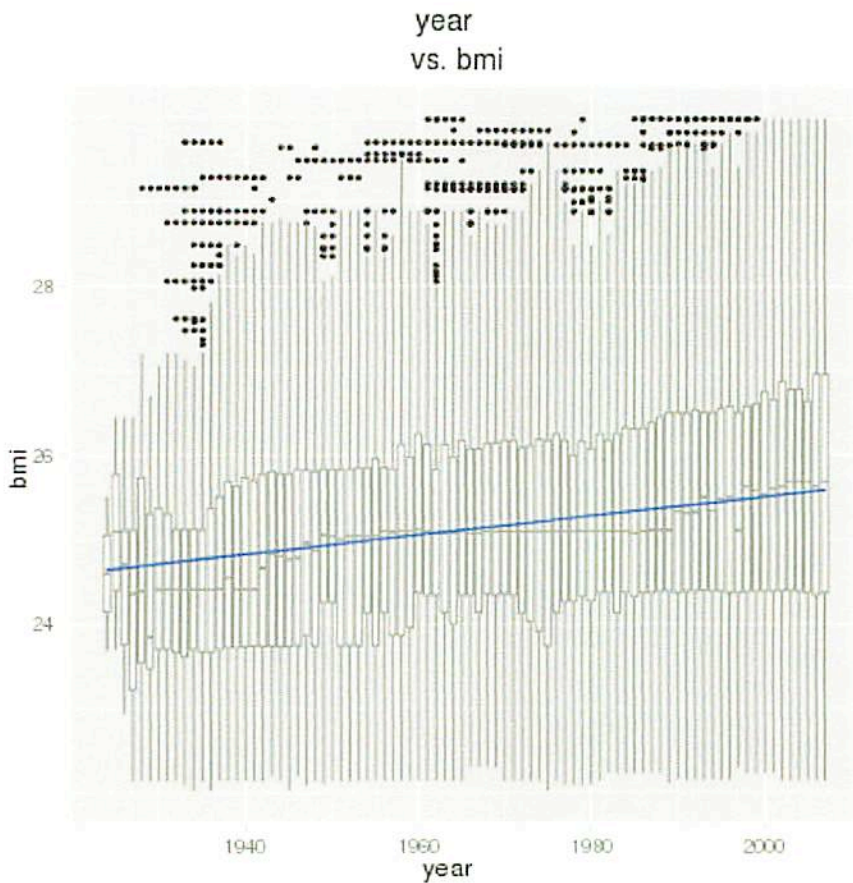
why does year
data

year vs.
bmi



The mean bmi of players increases over time. This suggests that teams have obtained more massive and powerful players. However, we have seen that the advantages of BMI decrease for BMI's ≥ 30 . It seems likely that baseball teams would increase players with high BMI's, but not with extremely high BMI's. If remove players with these extreme BMI's, we see an even more interesting phenomenon. Median BMI scores for baseball players remain largely stable, except for two dramatic and permanent increases in BMI. The first of these upward shifts in BMI occurred during the late 1940's. The second occurred during the early 1990's.

an alternative cause
might be the discreteness
of the data



6 Conclusion

The findings of this paper reveal that more massive players have a higher batting average than their lighter teammates. Moreover, massive players also appear to hit the ball farther when they do hit it than their lighter teammates do. These players have higher slugging percentages and hit more homeruns.

These offensive talents, however, come with a tradeoff. Massive players receive more strikeouts and seem to be slower and less skilled on base than their teammates. They attempt to steal bases ~~much~~ less often and fail more frequently than lighter players.

It would be interesting in future research to examine the relationship between BMI and defensive abilities. Do players with a high bmi have more errors? Are they confined to playing certain positions? And, do they contribute to getting as many outs as their teammates do?

Our research also shows that mean **BMI** has increased steadily over time, which reflects a well known change in the way baseball has been played. But our findings also suggest a more interesting and complicated progression.

The median bmi of baseball players has remained relatively constant except for an abrupt and permanent jump that occurred around 1945 and another that occurred in the early 1990s. These two jumps suggest that a linear model may not be an appropriate way to describe the behavior of bmi over time.


It also focuses scrutiny on the years that BMI changed. What could have occurred to effect such a shift? The late 1940's were a tumultuous time for baseball. Ballplayers who had joined the military during WWII

again you paragraphs are too short

were slowly finishing their service and returning to the sport. In 1948, Jackie Robinson became the first of many african american ball players admitted to the major leagues. Baseball attendance began to soar above any previous levels. And in the 1950's, televised games ruined many local leagues and sent their best players to the majors.

The early 1990s seem less remarkable, but perhaps they corresponded with advances in exercise science or steroid use.

Surprisingly, though, the rule changes that ended the "dead ball" era of baseball occurred 29 years before either of these two shifts. It seems unlikely that the rule changes could've lead to such delayed, and yet abrupt changes in bmi. A third shift may have happened around 1929 but the dearth of height and weight information for baseball players of this time makes it impossible to research it. Further research should seek to identify the hidden conditions that created the two observable changes in the size and shape of baseball players.



A Code

```
# Load ggplot2
library(ggplot2)

# Create merged data set with all the relevant variables
options(stringsAsFactors = FALSE)
b <- read.csv("batting.csv")
p <- read.csv("players.csv")
bp <- merge(b, p, by.x = "id", by.y = "id" by = "id")

bp$bmi <- bp$weight * 703 / bp$height ^ 2
bp$slg <- with(bp, ((h - hr - X3b - X2b) + 2*X2b + 3* X3b + 4*hr) / ab)
bp$ba <- with(bp, h / ab)
bp$sba <- with(bp, sb + cs)
bp$ob <- with(bp, h - hr + bb )

# Remove outliers in bmi
qplot(bmi, data = bp, geom = "histogram", binwidth = 1)
qplot( height, weight, data = bp)
bp <- bp[bp$height > 50 & bp$weight < 275,]

# What does the distribution of ab look like?

qplot(ab, data = bp, geom = "histogram", binwidth = 5)

# Are these zeroes for ab incorrect?
noab <- subset(bp, ab == 0)
dim(noab)
sum(noab$h + noab$X2b + noab$X3b + noab$hr)

# Remove players who batted less than 5 times per year
bp <- subset(bp, ab > 4)

# How does bmi affect batting average?

# Find and remove outliers from ba
qplot(ba, data = bp, geom = "histogram")
baclean = subset(bp, bp$ba < 0.4 & bp$ba > 0.1)
l = geom_smooth(aes(group = 1), method="lm")
qplot(bmi, ba, data = baclean, geom = "boxplot", group = round_any(bmi, 1), main
      = "bmi vs. batting average") + l
ggsave(file = "bmibattingaverage.png", width = 6, height = 6)

# How does bmi affect slugging percentage?
```

could you have
used a single
within call here?

use with?

```

# Find and remove outliers from slg
qplot(slg, data = bp, geom = "histogram")
slgclean = subset(bp, bp$slg < 0.75 & bp$slg != 0)
qplot(bmi, slg, data = slgclean, geom = "boxplot", group = round_any(bmi, 1),
      main = "bmi vs. slugging average") + 1
ggsave(file = "bmiandslugging.png", width = 6, height = 6)

# How does bmi affect homerun rate?

# Standardize hr and look for outliers
qplot(hr / ab, data = bp, geom = "histogram", binwidth = 0.01)
summary(bp$hr / bp$ab)
hrabclean = subset(bp, bp$hr / bp$ab < 0.15)
qplot(bmi, hr / ab, data = hrabclean, geom = "boxplot", group = round_any(bmi,
  1), main = "bmi vs. homeruns per at bat") + 1
ggsave(file = "bmiandhrbyab.png", width = 6, height = 6)

# How does bmi affect strikeouts per at bat?

# Standardize strikeouts and remove outliers
qplot(so / ab, data = bp, geom = "histogram")
soabclean = subset(bp, bp$so / bp$ab < 0.8 & bp$so / bp$ab != 0)
qplot(bmi, so / ab, data = soabclean, geom = "boxplot", group = round_any(bmi,
  1), main = "bmi vs. strike outs per at bat") + 1
ggsave(file = "bmistrikeout.png", width = 6, height = 6)
qplot(height, so / ab, data = soabclean, geom = "boxplot", group = round_any
  (height, 1), main = "height vs. strike out per game") + 1

# BMI and bb
qplot(bb / g, data = bp, geom = "histogram")
bbgclean <- subset(bp, bb / g != 0 & bb / g < 1)
qplot(bmi, bb / g, data = bbgclean, geom = "boxplot", group = round_any(bmi,
  1), main = "bmi vs. base on balls per game") + 1
ggsave(file = "bmibbg.png", width = 6, height = 6)

# BMI and base running

# How does bmi impact whether a player attempts to steal bases?
sbaclean <- subset(bp, bp$ob > 0)
sbaclean <- subset(sbaclean, sbaclean$sba / sbaclean$ob < 1)
qplot(sba / ob, data = sbaclean, geom = "histogram", binwidth = 0.01)
length(with(sbaclean, sba / ob > 0 & sba / ob < 1))
qplot(height, sba / ob, data = sbaclean, geom = "boxplot", group = round_any
  (height, 1), main = "height vs stolen base attempts") + 1

```



```

qplot(weight, sba / ob, data = sbaclean, geom = "boxplot", group = round_any
  (weight, 10), main = "weight vs stolen base attempts") + 1
ggsave(file = "weightandattemptsstealing.png", width = 6, height = 9)
qplot(bmi, sba / ob, data = sbaclean, geom = "boxplot", group = round_any(bmi,
  1), main = "bmi vs stolen base attempts") + 1
ggsave(file = "bmiandattemptsstealing.png", width = 6, height = 9)

# How does bmi affect a player's success at stealing bases?
qplot(sba, data = bp, geom = "histogram")

# remove players who did not attempt to steal base enough to yield useful data
sbclean <- subset(bp, bp$sba > 2 & bp$sba < 100)
qplot(sb / sba, data = sbclean, geom = "histogram")
qplot(bmi, sb / sba, data = sbclean, geom = "boxplot", group = round_any(bmi,
  1), main = "bmi vs. caught stealing per attempt") + 1
ggsave(file = "bmiandsuccessstealing.png", width = 6, height = 9)
qplot(weight, sb / sba, data = sbclean, geom = "boxplot", group =
  round_any(weight, 10), main = "bmi vs. success stealing per attempt") + 1
ggsave(file = "weightandsuccessstealing.png", width = 6, height = 9)

# How has bmi changed over time?

# First, what has been the trend over time for power hitting?
# Standardize hr on g
qplot(hr / g, data = bp, geom = "histogram")
hpgclean <- subset(bp, g > 10)
qplot(hr / g, data = hpgclean, geom = "histogram")
qplot(year, hr / g, data = hpgclean, geom = "boxplot", group = year,
  main = "Homeruns per game vs. year") + 1
ggsave(file = "10cleanhpgbyyearline.png", width = 6, height = 6)
qplot(year, slg, data = slgclean, geom = "boxplot", group = year,
  main = "Slugging percentage vs. year") + 1

# Plot bmi over time
bmiclean = subset(bp, bp$bmi > 22 & bp$bmi < 30)
qplot(year, bmi, data = bp, geom = "boxplot", group = year, main = "year vs.
  bmi") + 1
ggsave(file = "bmibyyear.png", width = 6, height = 6)
qplot(year, bmi, data = bmiclean, geom = "boxplot", group = year, main = "year
  vs. bmi") + 1
ggsave(file = "healthybmiyear.png", width = 6, height = 6)

# Examine bmi over 30
muted <- identity

```

```
qplot(weight, height, data = bp, colour = bmi)
ggsave(file = "heightweightbmi.png", width = 6, height = 6)
qplot(bmi, g, data = bp, geom = "boxplot", group = round_any(bmi, 1),
      main = "bmi vs. games") + 1
ggsave(file = "30bmigames.png", width = 9, height = 6)
qplot(bmi, ab, data = bp, geom = "boxplot", group = round_any(bmi, 1),
      main = "bmi vs. at bats") + 1
ggsave(file = "30bmiatbats.png", width = 9, height = 6)
```