Project 2

Luke Terry

29 November, 2021

Figure 1: Luke Terry

My Video

```
# <video width="320" height="240" controls>
#   <source src="usingvideoinrmd.mp4" type="video/mp4">
# Your browser does not support the video tag.
# </video>
```

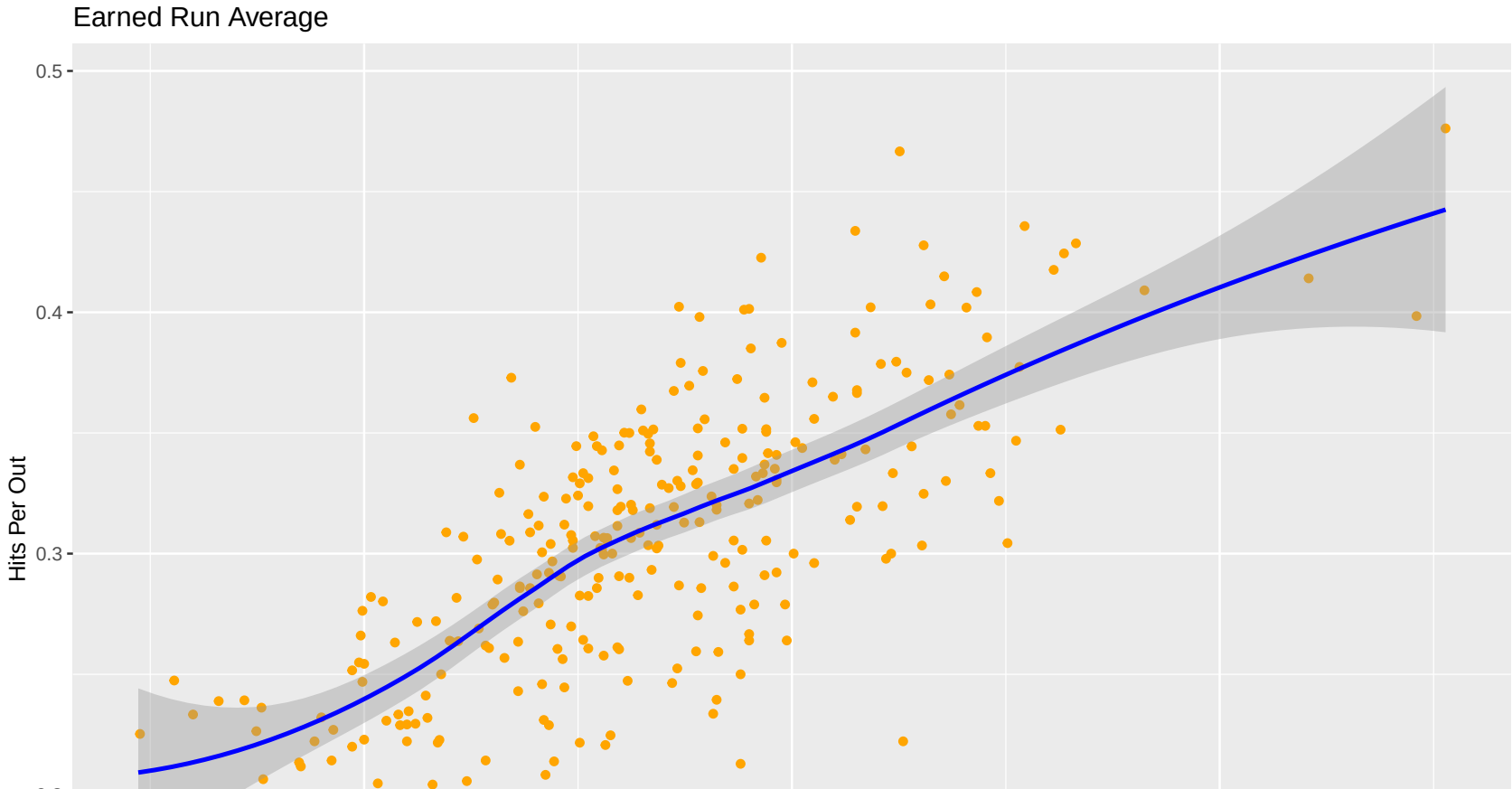
Introduction

Baseball is a timeless sport which was popularized in the United States in the mid 1850s (Wikipedia (2021)). Like many other sports in the United States, watching baseball, talking about baseball, and making predictions about baseball is a popular pastime among many people.

Compared to other sports, baseball and statistics have a close relationship. Perhaps it's something about the leisurely pace of a baseball game, or the easily measurable individual actions in each baseball play. Many people believe that the use of statistics in baseball was introduced by Billy Beane, the General Manager of the Oakland A's, since this narrative was popularized in Michael Lewis's 2003 book *Moneyball* (Lewis (2003)), and 2011 movie of the same name. Statistics in baseball, however, have been a consideration within coaching staff far before Billy Beane. There exist evidence of statisticians assisting baseball managers as early as 1940, and many general managers used statistical analysis to inform their decisions prior to Beane, such as Danny Evans of the Dodgers and Doug Melvin of the Brewers. Beane's fascination

Premliminary Plots and Interpretation of the data

In order to decide which variables to analyze using linear regression, it's useful to make preliminary plots and analyze them subjectively for seemingly linear patterns. Below are a few which I've constructed using some traditional pitching statistics. All graphs are a plot of a traditional pitching statistic versus a pitcher's Hits Per Out (HPO).



Theory needed to carry out SLR

It is my belief that a pitcher's HPO tends to increase as their ERA does. Thus, when choosing between pitchers, a pitcher with a higher ERA will tend to give up more hits. I would like to create a model which relates the rate at which a pitcher gives up hits (HPO) to the pitchers, ERA, so I can use the more ubiquitous ERA to predict which pitchers are likely to give up more hits.

As with many sources of real data, however, an unknown number of factors and randomness affects a pitcher's ability to throw pitches, even if we assume that the skill of the pitcher does not change (or changes in a predictable career-aging curve ((n.d.))). In other words, in the preliminary plots, there seemed to be a vaguely linear association, but all of the points did *not* line up in a straight line. Because of this, a deterministic model cannot be used, and instead I will need to construct a probabilistic model.

One type of probabilistic model is simple linear regression (SLR). SLR assumes that there is a linear association between the variables, and uses data driven methods to construct a line which

Estimating Parameters

In order to estimate the parameters $/beta_0$ and β_1, I will now use the Method of Least Squares. This involves finding the line in which the deviations from the expected value of the model (the regression line) are minimized. If all of the assumptions about ϵ which I outlined previously are true, the Method of Least Squares produces a line which is identical to the Method of Maximum Likelihood (Mendenhall (2016)). Since the population least squares line is represented by $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$, my predicted line base on the 2019 MLB season (my sample) will be $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i$. For each X_i, I am able to calculate a residual $\hat{\epsilon}$, which is the deviation of the data point from my least squares regression line.

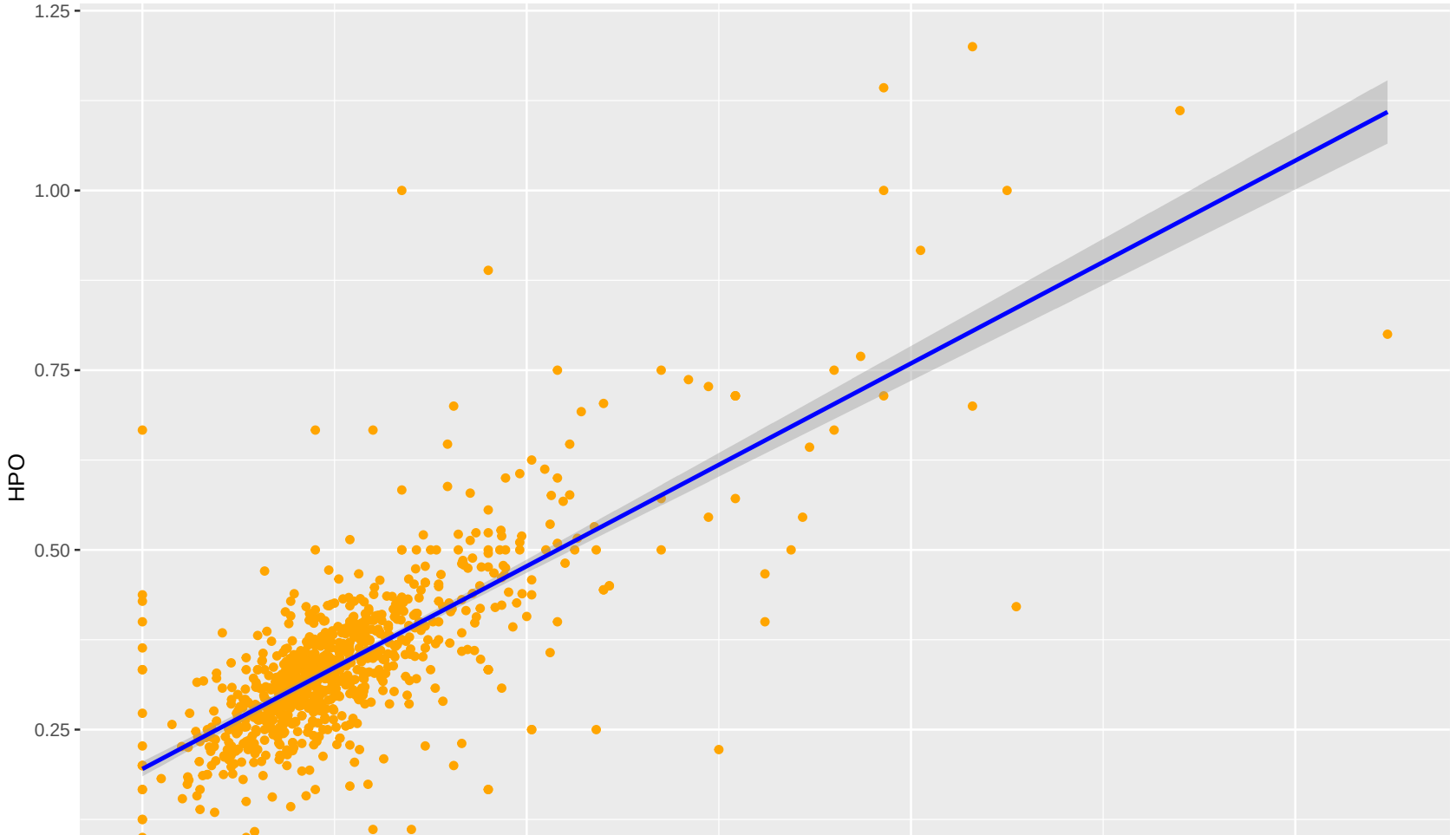
$$\hat{\epsilon} = (y_i - \hat{y}_i) = (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Therefore, the least squares regression line will be the one which minimizes the sum of the squares for error (SSE). The SSE is given by

$$\text{SSE} = \sum^{n} \hat{\epsilon}^2 = \sum^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

Check for outliers

```
g <- ggplot(data = pt, aes(x = ERA, y = HPO))
g = g + geom_point(col = "orange")
g = g + stat_smooth(method = "lm", col = "blue")
g
```



Development of model without Outliers

```
#remove data with an excessively high Cook's Distance
baddata <- subset(cooksdf, cd >= (1/873), select = obs)$obs
ptf <- pt[-c(baddata),]

ptf <- subset(ptf, is.finite(HPO))

hpof.lm <- with(ptf, lm(HPO ~ ERA))
smf <- summary(hpof.lm)
betas <- smf$coefficients
rsq <- smf$r.squared
rsqa <- smf$adj.r.squared
```

In the above block of code, I have produced another linear model, this time with the potential outliers removed. The values predicted by this linear model are $\hat{\beta}_0 = 0.1844$ and $\hat{\beta}_1 = 0.0302$. This model also has an r^2 value of 0.6901. It has an adjusted r^2 value of '$rround(rsqa, 4)$, however. Adjusted r^2 is a similar metric to r^2, in that it is a measure of how well the model fits the data. Adjusted

Validity with mathematical expressions

Now that I have an estimate for the values in my linear model, it's important to analyze whether the assumptions that this was built on are true. This section will focus on the theory behind verifying assumptions; the next section will focus on actually verifying the assumptions in this data set.

If you recall an earlier section, there are four assumptions on which my model is built. These assumptions involve the distribution of the error, specifically that it is normally distributed, with a mean of zero, and a constant variance. Additionally, all errors should be independent of all other errors. To verify these assumptions, I will first use the residuals to find the Residual Sum of Squares (RSS). It is calculated using the following equation, where $\hat{y}$ is a value predicted by our regression line, and each y_i is a data point.

$$\text{RSS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

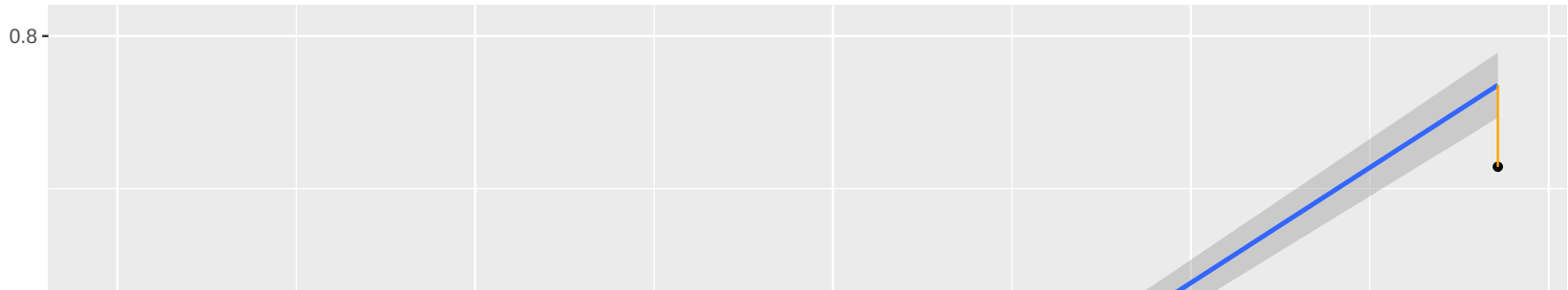
After this, I will find the Model Sum of Squares (MSS), which is

Actual tests for Validity

In this section, I will put the theory introduced in the last section to work on our data set. The scatterplot at the end of the previous section seems to illustrate that a simple linear model is a good fit for the data, but many statistics exist to test whether this model is the best one, and to test whether the assumptions that I made when creating the model and removing its outliers are true.

Residual Plot

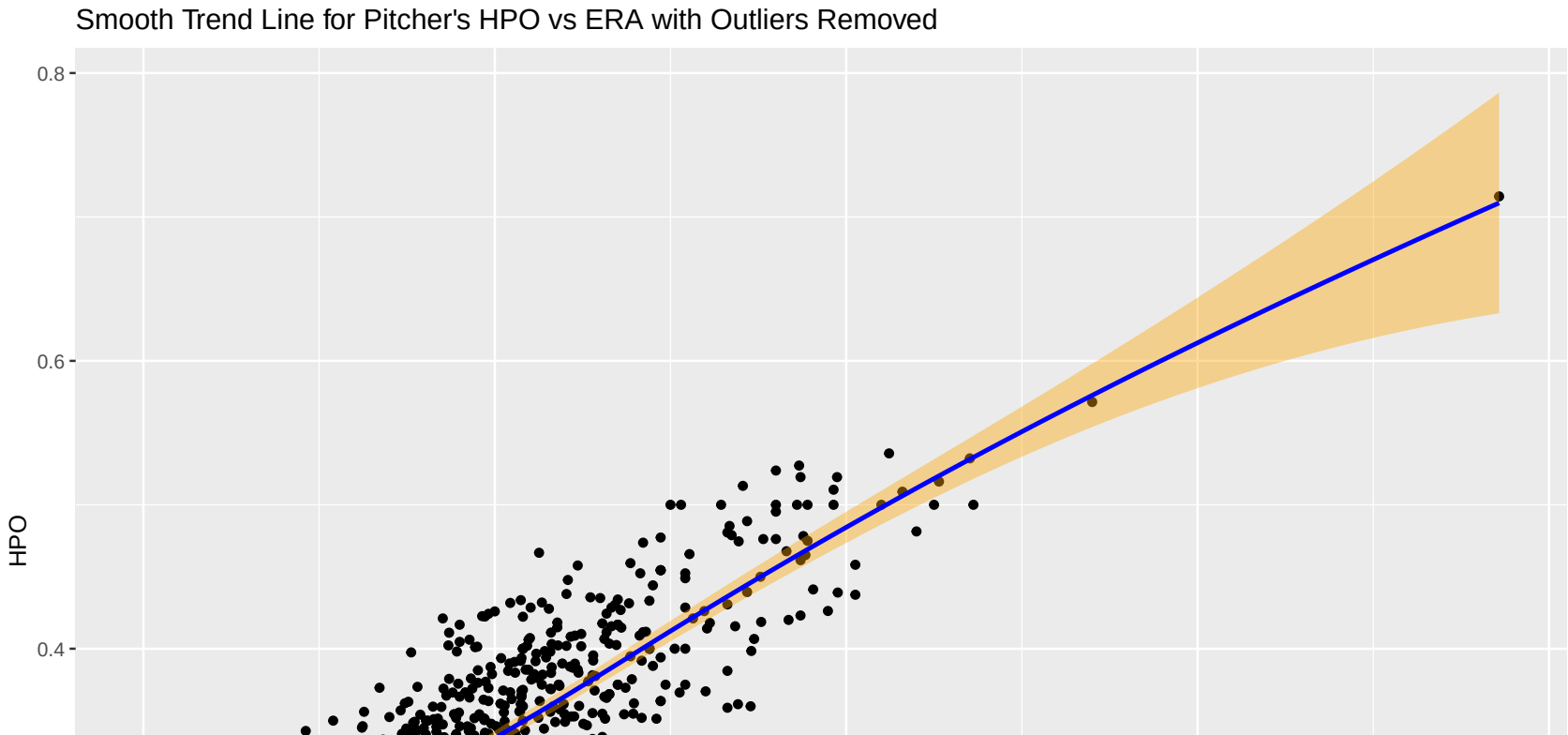
```
e <- ggplot(ptf, aes(x = ERA, y = HPO))
e = e + geom_point()
e = e + geom_smooth(method = 'lm')
e = e + geom_segment(aes(x=ERA, xend=ERA, y=HPO, yend=hpof
e
```



Assesing the Use of a Linear Model

Loess Smooth Plot for the data

```
f <- ggplot(data = ptf, aes(x = ERA, y = HPO))
f = f + geom_point()
f = f + geom_smooth(method = "loess", col = "blue", fill =
f = f + ggtitle("Smooth Trend Line for Pitcher's HPO vs ERA
f
```



Analysis of Linear Model

Point Estimates and Model Summary

```
summary(hpof.lm)
```

```
##
## Call:
## lm(formula = HPO ~ ERA)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.108280 -0.028982  0.000372  0.028104  0.111980
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.184399   0.003908   47.19   <2e-16 ***
## ERA         0.030246   0.000745   40.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Predictive Use of the Model

Now that the preceding analysis has been completed, it's time to apply the model on a practical level. If I choose a pitcher, I should be able to use this regression line to predict the rate at which they will give up hits, and therefore compare them to other pitchers and increase my liklihood of keeping my streak.

Since it's currently the baseball off season, I will look retroactively at a call that had to be made without the help of this model. Take for instance August 1st, 2021. On this day, the Milwaukee Brewers were scheduled to play the Atlanta Braves, who were starting the pitcher Charlie Morton, who at the time had an ERA of 3.69. The lead off hitter for the Brewers, Tyrone Taylor, would likely face Morton a few times, so knowing the HPO of the pitcher is very useful.

```
predict(hpof.lm, data.frame(ERA=c(3.69)))
```

```
##         1
## 0.2960087
```

Conclusion

Research Question and Implications

I set out to find a relationship between a pitcher's ERA, and the rate at which the same pitcher gives up hits. Based on the SLR which I have performed, a linear model is appropriate and allows for meaningful predictions to be made. Although there are many other factors which may also be factored into a batter's ability to get a hit, the information from this model will actually help me with maintaining my streak, since it allows for an analysis of a pitcher's average propensity to give up hits given their ERA.

Suggest ways to improve model or experiment

At the end of the day, the practical use for this entire analysis is to be able to predict the rate of hits that a pitcher gives up, given their ERA, which is a standard and ubiquitous statistic. For my purposes (coming closer to beating the streak), further analysis will still have to be done on different sabermetric data to predict whether a batter will face a pitcher, and what other factors may be affecting the pitcher or the batter.

When I constructed the preliminary plots section, I was surprised to

References

n.d. https://www.billjamesonline.com/aging_patterns/.

Baccellieri, Emma. 2021. "Five Takeaways One Month into Sticky-Stuff Enforcement." *Sports Illustrated*. Sports Illustrated. https://www.si.com/mlb/2021/07/21/sticky-stuff-crackdown-one-month-the-opener.

Baseball Almanac, Inc. n.d. "Baseball Almanac 2021: Baseball History, Baseball Records and Baseball Research." *Baseball Almanac*. https://www.baseball-almanac.com/.

"Cook's Distance." 2021. *Wikipedia*. Wikimedia Foundation. https://en.wikipedia.org/wiki/Cook%27s_distance.

Douglas, Joe. 2017. "Ottoneu 101: Plate Appearances by Lineup Spot." *RotoGraphs Fantasy Baseball*. https://fantasy.fangraphs.com/buying-generic-plate-appearances-by-lineup-spot/.

Lewis, Michael (Michael M.). 2003. *Moneyball : The Art of Winning an Unfair Game / Michael Lewis.* 1st ed.. New York: