A Study of Baseball Attendance: the Effects of Winning and Top-End Player Acquisitions

John Pette

Introduction

This is a statistical examination of one of the factors most critical to measuring a major league team's success: attendance, the most direct driver of a team's revenue. In my analysis, I sought to answer several questions. The first two were fairly obvious:

1. What is the relationship between team wins and attendance?
2. How does winning a league pennant or World Series relate to attendance?

The deeper, more intriguing questions I wanted to address were related to payroll and high-end player acquisitions:

3. Is there a relationship between payroll and attendance?
4. How does a top-end player acquisition relate to attendance?

I chose to look at the years 2000-2017, as that 18-year period would allow for an adequate number of data points (540), while covering the period containing all modern mega-contracts. I should note that this is an observational study, so it cannot make any causal conclusions.

Methodology and Data

First, let's examine attendance itself. Attendance figures are far from perfect. If a game sells out, there is an artificial cap on the attendance figure for that game, and it ceases to be a good measure of demand. However, the number of sellouts in baseball is fairly negligible when compared to the overall quantity of games played, so the straight attendance figures are likely the best measure. They certainly represent the most practical measure - I looked at using attendance as a percentage of overall capacity, but reading my interpretations of relationships between anything and attendance percentages would have driven most of my readers to violence, so I stayed with the straight attendance figures.

My starting point for data was the Teams table from the 2017 version of the Lahman database (available at seanlahman.com, copyright 1996-2018 by Sean Lahman). This is a very clean data set, which cut down tremendously on data clean up. However, while the Lahman database is a phenomenal resource, there were additional variables I wanted to assess, so I had to compile data from other sources.

Payroll: To examine team payroll, I pulled the Opening Day payroll figures from Cot's Baseball Contracts, which, conveniently enough, covered the precise years I was examining, 2000-2017. In order to be comparable, though, the payroll figures needed to be in constant dollars, so I adjusted them all to 2017 equivalent amounts using a calculator based on the consumer price indices for the years in question.

The Lahman database provides indicator variables showing whether a team won the World Series, League Championship Series, or Wild Card game in a given year. I thought these would be interesting, but attendance effects during those years would likely be captured by the wins variable. In lieu of these, I created new lag variables to show whether a team won the World Series or league pennant in the previous year, making the assumption that we could see more of an effect of success after the fact.I also added an indicator variable to show whether a team had won the league pennant last year, but lost the World Series. This made the two variables mutually exclusive and eliminated the need for interaction terms between them.

I created an additional indicator variable to show whether a team opened a new ballpark in a given year. We can generally assume that we would see an attendance spike in those years, but I wanted to build it into the model.

Finally, I add two more indicator variables to track top-end acquisitions: one to show whether a team made a top-end player acquisition, by signing a free agent or through a trade, in the off-season before a given year. The other shows whether a team made a mid-season trade to acquire a top-end player. This is the only subjective field in my analysis, and I will discuss it in more depth below.

Now, let's load the data and look at the top few rows.

```
# Load the Teams table from the Lahman Database, 2017 version, copyright 1996-2018 by Sean Lahman.
Teams <- read.csv("baseballdatabank-master_2018-03-28/baseballdatabank-master/core/Teams.csv")
```

```
# Create 2000-2017 subset of data.
tm <- Teams[which(Teams$yearID >= '2000' & Teams$yearID <= '2017'),]
head(tm)
```

```
##    yearID lgID teamID franchID divID Rank   G Ghome  W  L DivWin WCWin
## 41   2000   AL    ANA      ANA     W    3 162    81 82 80      0     0
## 42   2001   AL    ANA      ANA     W    3 162    81 75 87      0     0
## 43   2002   AL    ANA      ANA     W    2 162    81 99 63      0     1
## 44   2003   AL    ANA      ANA     W    3 162    82 77 85      0     0
## 45   2004   AL    ANA      ANA     W    1 162    81 92 70      1     0
## 46   2005   AL    LAA      ANA     W    1 162    81 95 67      1     0
##    LgWin LgWinLastYr WSWin WSWinLastYr LgWinWSLossLastYr   R   AB    H X2B
## 41     0           0     0           0                 0 864 5628 1574 309
## 42     0           0     0           0                 0 691 5551 1447 275
## 43     1           0     1           0                 0 851 5678 1603 333
## 44     0           1     0           1                 0 736 5487 1473 276
## 45     0           0     0           0                 0 836 5675 1603 272
## 46     0           0     0           0                 0 761 5624 1520 278
##    X3B  HR  BB   SO  SB CS HBP SF  RA  ER  ERA CG SHO SV IPouts   HA HRA
## 41  34 236 608 1024  93 52  47 43 869 805 5.00  5   3 46   4344 1534 228
## 42  26 158 494 1001 116 52  77 53 730 671 4.20  6   1 43   4313 1452 168
## 43  32 152 462  805 117 51  74 64 644 595 3.69  7  14 54   4357 1345 169
## 44  33 150 476  838 129 61  56 50 743 680 4.28  5   9 39   4294 1444 190
## 45  37 162 450  942 143 46  73 41 734 692 4.28  2  11 50   4363 1476 170
## 46  30 147 447  848 161 57  29 39 643 598 3.68  7  11 54   4393 1419 158
##    BBA  SOA   E  DP    FP                          name          park
## 41 662  846 134 182 0.978                Anaheim Angels Angel Stadium
## 42 525  947 103 142 0.983                Anaheim Angels Angel Stadium
## 43 509  999  87 151 0.986                Anaheim Angels Angel Stadium
## 44 486  980 105 138 0.982                Anaheim Angels Angel Stadium
## 45 502 1164  90 126 0.985                Anaheim Angels Angel Stadium
## 46 443 1126  87 139 0.986 Los Angeles Angels of Anaheim Angel Stadium
##    attendance MaxCapacity attendancepct TopAcq MidYrAcq NewPark AprPayroll
## 41    2066982     3649050     0.5664439      0        0       0   79403400
## 42    2000919     3649050     0.5483397      0        0       0   66113206
## 43    2305547     3649050     0.6318212      0        0       0   84126632
## 44    3061094     3649050     0.8388742      0        0       0  105270180
## 45    3375677     3649050     0.9250838      1        0       0  130493998
## 46    3404686     3649050     0.9330335      0        0       0  122645279
##    BPF PPF teamIDBR teamIDlahman45 teamIDretro
## 41 102 103      ANA            ANA         ANA
## 42 101 101      ANA            ANA         ANA
## 43 100  99      ANA            ANA         ANA
## 44  98  97      ANA            ANA         ANA
## 45  97  97      ANA            ANA         ANA
## 46  98  97      LAA            ANA         ANA
```

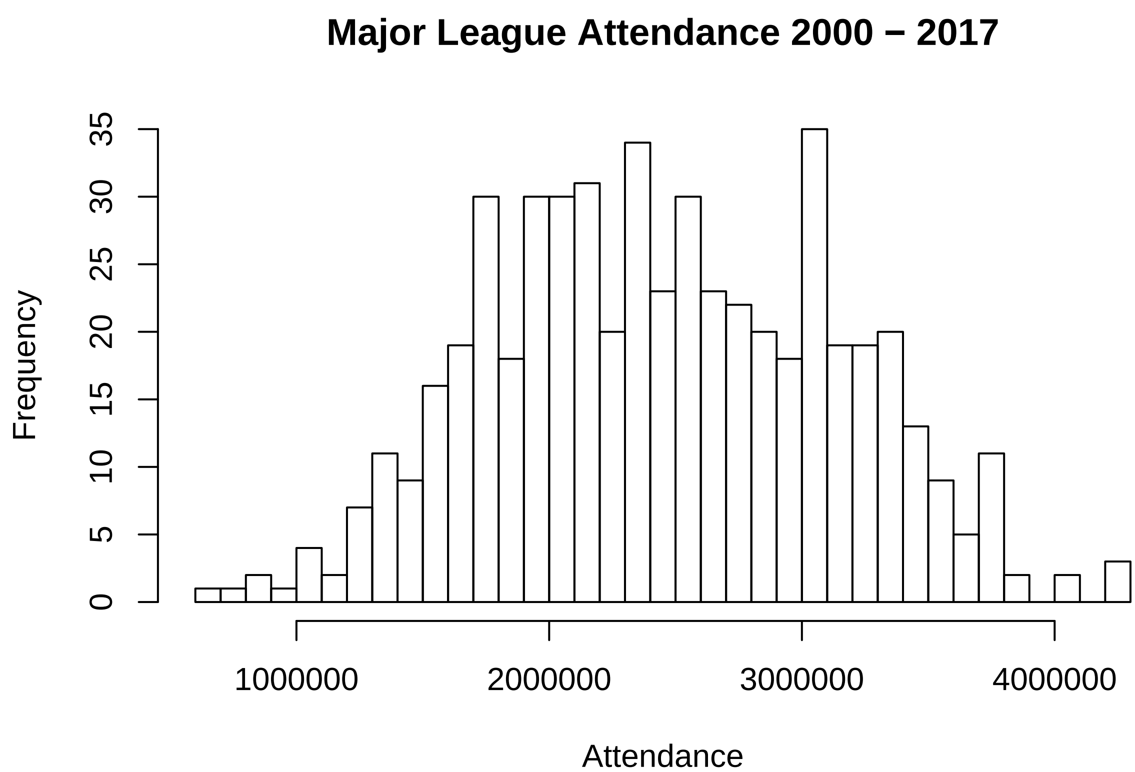
```r
#Count rows of data.
cat("Total data points:", nrow(tm))
```

```
## Total data points: 540
```

Attendance

Below is a histogram of the attendance figures. It shows a more-or-less normal distribution, with a spike at the 3 million mark.

```r
# Print attendance histogram.
hist(tm$attendance, breaks = 50, xlab="Attendance", main="Major League Attendance 2000 - 2017")
```



Initial Models: Attendance, Wins, and Payroll

For our first models, I will look at two classic relationships: wins vs. attendance and payroll vs. wins. Throughout this analysis, I will use robust standard errors with my linear models, as it is just good standard practice.

```r
# Run linear model,
lm1 <- lm(attendance ~ W, data = tm)

# Print model output summary.
summary(lm1)
```

```
##
## Call:
## lm(formula = attendance ~ W, data = tm)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1697340  -410993    -5366   429309  1617808
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   137489     185694    0.74               0.459
## W              28577       2271   12.58 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 599200 on 538 degrees of freedom
## Multiple R-squared:  0.2274, Adjusted R-squared:  0.2259
## F-statistic: 158.3 on 1 and 538 DF,  p-value: < 0.00000000000000022
```

```
# Print coefficients report with robust standard errors.
coeftest(lm1, vcovHC(lm1))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 137488.6   175594.7   0.783               0.434
## W            28577.1     2174.5  13.142 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Adjust standard errors.
cov1         <- vcovHC(lm1)
robust_se    <- sqrt(diag(cov1))

# Adjust F-statistic .
wald_results <- waldtest(lm1, vcov = cov1)

# Print stargazer table of linear model output with robust standard errors.
stargazer(lm1, type="latex",
          se = list(NULL, robust_se),
          header = FALSE,
          title = "Model 1 - Attendance and Wins",
          model.numbers=FALSE,
          no.space = TRUE,
          omit.stat = c("rsq", "f"))
```

```
lm2 <- lm(W ~ AprPayroll, data = tm)
coeftest(lm2, vcovHC(lm2))
```

```
##
## t test of coefficients:
##
##                     Estimate       Std. Error t value
## (Intercept) 71.2844009652774  1.1523145409259 61.8619
## AprPayroll   0.0000000926849  0.0000000096653  9.5894
##                          Pr(>|t|)
## (Intercept) < 0.00000000000000022 ***
## AprPayroll  < 0.00000000000000022 ***
## ---
```

Table 1: Model 1 - Attendance and Wins

	Dependent variable:
	attendance
W	28,577.060***
	(2,271.243)
Constant	137,488.600
	(185,694.000)
Observations	540
Adjusted R^2	0.226
Residual Std. Error	599,201.100 (df = 538)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cov2         <- vcovHC(lm2)
robust_se    <- sqrt(diag(cov2))
wald_results <- waldtest(lm2, vcov = cov2)

stargazer(lm2, type="latex",
          se = list(NULL, robust_se),
          header = FALSE,
          title = "Model 2 - Wins and Payroll",
          model.numbers=FALSE,
          no.space = TRUE,
          omit.stat = c("rsq", "f"))
```

Table 2: Model 2 - Wins and Payroll

	Dependent variable:
	W
AprPayroll	0.00000***
	(0.000)
Constant	71.284***
	(1.186)
Observations	540
Adjusted R^2	0.125
Residual Std. Error	10.628 (df = 538)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
cat("Additional attendance associated with each win above the baseline:", lm1$coefficients[2],"\nAdditi
```

```
## Additional attendance associated with each win above the baseline: 28577.06
## Additional payroll associated with each additional win over baseline: 10789248
```

This is interesting. I had always assumed that there was an inconsistent relationship between wins and payroll, having watched many high-dollar teams implode. The model here suggests a highly statistically significant relationship between wins and payroll, with a p-value of effectively zero. It is just not a particularly useful one. Each additional win is associated with a $10.8 million payroll increase. The relationship has statistical significance, but no practical significance.

On the other hand, each win is associated with an attendance increase of 28,577. This is also a highly statistically significant result, and this is the relationship I will explore further. The model has an adjusted R-squared value of 0.2259, meaning it only explains 22.6 percent of the variation in attendance. This tells us that there are many additional omitted variables, or that there are many other factors related to attendance than just wins, which is a reasonable assumption anyway.

World Series and League Pennant Wins

Now, let's take a look at the same model, accounting for whether a team won the World Series in the preceding year. For posterity, I will first test a model accounting for whether a team wins the World Series in a given year.

```
lm3 <- lm(attendance ~ W + WSWin, data = tm)
coeftest(lm3, vcovHC(lm3))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 154029.3   180693.4  0.8524              0.3944
## W            28345.4     2254.6 12.5724 <0.0000000000000002 ***
## WSWin        66460.2   152853.6  0.4348              0.6639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cov3         <- vcovHC(lm3)
robust_se    <- sqrt(diag(cov3))
wald_results <- waldtest(lm3, vcov = cov3)

stargazer(lm3, type="latex",
          se = list(NULL, robust_se),
          header = FALSE,
          title = "Model 3 - Attendance and Concurrent Year World Series Win",
          model.numbers=FALSE,
          no.space = TRUE,
          omit.stat = c("rsq", "f"))
```

Table 3: Model 3 - Attendance and Concurrent Year World Series Win

	Dependent variable:
	attendance
W	28,345.410***
	(2,330.254)
WSWin	66,460.180
	(147,379.700)
Constant	154,029.300
	(189,417.000)
Observations	540
Adjusted R^2	0.225
Residual Std. Error	599,645.200 (df = 537)
Note:	*p<0.1; **p<0.05; ***p<0.01

As I suspected at the outset, we do not see a statistically significant effect. Any effect is likely accounted for in the wins variable. Now, we will examine a model showing whether a team won the World Series the previous year.

```
lm4 <- lm(attendance ~ W + WSWinLastYr, data = tm)
coeftest(lm4, vcovHC(lm4))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value              Pr(>|t|)
## (Intercept) 184685.0   175638.3  1.0515                0.2935
## W            27779.9     2184.6 12.7161 < 0.00000000000000022 ***
## WSWinLastYr 520509.2   116525.2  4.4669            0.00000968 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cov4         <- vcovHC(lm4)
robust_se    <- sqrt(diag(cov4))
wald_results <- waldtest(lm4, vcov = cov4)

stargazer(lm4, type="latex",
          se = list(NULL, robust_se),
          header = FALSE,
          title = "Model 4 - Attendance, Wins, and Preceding Year World Series Win",
          model.numbers=FALSE,
          no.space = TRUE,
          omit.stat = c("rsq", "f"))
```

Table 4: Model 4 - Attendance, Wins, and Preceding Year World Series Win

	Dependent variable:
	attendance
W	27,779.860***
	(2,256.319)
WSWinLastYr	520,509.200***
	(142,703.600)
Constant	184,685.000
	(184,061.700)
Observations	540
Adjusted R^2	0.243
Residual Std. Error	592,464.500 (df = 537)
Note:	*p<0.1; **p<0.05; ***p<0.01

This is a different story. We get a highly statistically significant result showing an 520,000-person increase in attendance related to putting a World Series champion on the field. What if we add in the preceding year's World Series losers?

```
lm5 <- lm(attendance ~ W + WSWinLastYr + LgWinWSLossLastYr, data = tm)
coeftest(lm5, vcovHC(lm5))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        250505.6   177609.1  1.4104             0.158992
```

```
## W                   26830.7     2222.4 12.0731 < 0.00000000000000022 ***
## WSWinLastYr       537158.9   116406.8  4.6145          0.000004936 ***
## LgWinWSLossLastYr 314208.6   121220.5  2.5920             0.009801 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cov5         <- vcovHC(lm5)
robust_se    <- sqrt(diag(cov5))
wald_results <- waldtest(lm5, vcov = cov5)

stargazer(lm5, type="latex",
          se = list(NULL, robust_se),
          header = FALSE,
          title = "Model 5 - Attendance, Wins, and Preceding Year World Series or Pennant Win",
          model.numbers=FALSE,
          no.space = TRUE,
          omit.stat = c("rsq", "f"))
```

Table 5: Model 5 - Attendance, Wins, and Preceding Year World Series or Pennant Win

	Dependent variable:
	attendance
W	26,830.720***
	(2,290.333)
WSWinLastYr	537,158.900***
	(142,414.000)
LgWinWSLossLastYr	314,208.600**
	(144,259.400)
Constant	250,505.600
	(185,896.100)
Observations	540
Adjusted R^2	0.248
Residual Std. Error	590,409.900 (df = 536)
Note:	*p<0.1; **p<0.05; ***p<0.01

There appears to be a relationship between attendance and winning the pennant, but losing the World Series. It is a somewhat smaller, but still substantial, effect: an attendance bump of 314,000. None of the above is a particularly shocking outcome.

Top-End Player Acquisition

Now, let's see what happens when we account for top-end player acquisitions. There will be some gray area in this part, as what qualifies as "top-end" will differ from person to person. My definition of a top-end player acquisition is one involving a superstar or one with a very high dollar value (naturally, these often coincide). I am not looking at how these acquisitions played out, just how they would have been viewed at the time. Essentially, I am trying to quantify the blockbuster. If it is a deal that made a team's fans say "wow" at the time (in a good way), I want it in here. If it made the rival team's fans say "uh-oh", I want it in here.

Closers are not on this list. Top closer acquisitions certainly excite people, but I am working under the assumption that no one goes to the ballpark to see the closer the way they would for a new clean-up hitter or ace. No one is saying, "Let's go to the game tonight to see Papelbon," unless it's, "Let's go to the game tonight to see if Papelbon chokes out Harper again."

Mid-season acquisitions present a quandary, as, by definition, we cannot look at those as affecting season-long attendance in the same way as off-season acquisitions. I have chosen to break out mid-season acquisitions as a separate indicator variable. In the event that those acquisitions were free agents-to-be, and the acquiring team re-signs that player in the subsequent off-season, I have treated those as new free agent signings. This is not perfect, but there were only a few of those cases, and in most (all?) of them, the fans never expected the team to re-sign the player, so they were met with the excitement of a big free agent signing. This would include Cespedes in 2015/16, Holliday in 2009/10, Manny in 2008/09, and Jason Schmidt in 2001/02 (hey, it was a big deal at the time). Here are the complete lists. This could be an endless debate, so we will proceed under the assumption that people will generally agree on most of these, and that they are sufficient for analysis.

Top-end Offseason Acquisitions Team Year Player Mode CIN 2000 Ken Griffey Jr. (Trade) DET 2000 Juan Gonzalez (Trade) LAD 2000 Shawn Green (Trade) NYM 2000 Mike Hampton (Trade) BOS 2001 Manny Ramirez (FA) CLE 2001 Juan Gonzalez (FA) COL 2001 Mike Hampton (FA) NYY 2001 Mike Mussina (FA) SEA 2001 Ichiro Suzuki (FA) TEX 2001 Alex Rodriguez (FA) NYY 2002 Jason Giambi (FA) SFG 2002 Jason Schmidt (Trade/FA) NYY 2003 Hideki Matsui (FA) NYY 2003 Jose Contreras (FA) PHI 2003 Jim Thome (FA) PHI 2003 Kevin Millwood (FA) LAA 2004 Vladimir Guerrero (FA) BAL 2004 Miguel Tejada (FA) BOS 2004 Curt Schilling (Trade) HOU 2004 Roger Clemens (FA) NYY 2004 Alex Rodriguez (Trade) ARI 2005 Shawn Green (Trade) DET 2005 Magglio Ordonez (FA) MIL 2005 Carlos Lee (Trade) NYM 2005 Pedro Martinez (FA) NYM 2005 Carlos Betran (FA) NYY 2005 Randy Johnson (Trade) BOS 2006 Josh Beckett (Trade) CHW 2006 Jim Thome (Trade) NYM 2006 Carlos Delgado (Trade) WAS 2006 Alfonso Soriano (Trade) BOS 2007 Daisuke Matsuzaka (FA) BOS 2007 J.D. Drew (FA) CHC 2007 Alfonso Soriano (FA) HOU 2007 Carlos Lee (FA) SFG 2007 Barry Zito (FA) LAA 2008 Torii Hunter (FA) ARI 2008 Dan Haren (Trade) DET 2008 Miguel Cabrera (Trade) NYM 2008 Johan Santana (Trade) LAD 2009 Manny Ramirez (Trade/FA) NYY 2009 Mark Teixeira (FA) NYY 2009 A.J. Burnett (FA) NYY 2009 C.C. Sabathia (FA) PHI 2010 Roy Halladay (Trade) SEA 2010 Cliff Lee (Trade) STL 2010 Matt Holliday (Trade/FA) LAA 2011 Vernon Wells (Trade) BOS 2011 Adrian Gonzalez (FA) BOS 2011 Carl Crawford (FA) MIL 2011 Zack Greinke (Trade) PHI 2011 Cliff Lee (FA) TEX 2011 Adrian Beltre (FA) WAS 2011 Jayson Werth (FA) LAA 2012 Albert Pujols (FA) DET 2012 Prince Fielder (FA) MIA 2012 Jose Reyes (FA) TEX 2012 Yu Darvish (FA) WAS 2012 Gio Gonzalez (Trade) LAA 2013 Josh Hamilton (FA) ATL 2013 Justin Upton (Trade) KAN 2013 James Shields (Trade) NYY 2014 Masahiro Tanaka (FA) NYY 2014 Jacoby Ellsbury (FA) SEA 2014 Robinson Cano (FA) TEX 2014 Prince Fielder (Trade) BOS 2015 Hanley Ramirez (FA) BOS 2015 Pablo Sandoval (FA) CHC 2015 Jon Lester (FA) MIA 2015 Dee Gordon (Trade) SDP 2015 Matt Kemp (Trade) TOR 2015 Josh Donaldson (Trade) WAS 2015 Max Scherzer (FA) ARI 2016 Zack Greinke (FA) BOS 2016 David Price (FA) CHC 2016 Jason Heyward (FA) DET 2016 Justin Upton (FA) DET 2016 Jordan Zimmermann (FA) NYM 2016 Yoenis Cespedes (Trade/FA) CLE 2017 Edwin Encarnacion (FA)

Top-end Mid-season Acquisitions Team Year Player Mode ARI 2000 Curt Schilling (Trade) SFG 2001 Jason Schmidt (Trade/FA) HOU 2004 Carlos Beltran (Trade) STL 2004 Larry Walker (Trade) NYY 2006 Bobby Abreu (Trade) ATL 2007 Mark Teixeira (Trade) LAA 2008 Mark Teixeira (Trade) LAD 2008 Manny Ramirez (Trade/FA) MIL 2008 C.C. Sabathia (Trade) CHW 2009 Jake Peavy (Trade) PHI 2009 Cliff Lee (Trade) STL 2009 Matt Holliday (Trade/FA) PHI 2010 Roy Oswalt (Trade) TEX 2010 Cliff Lee (Trade) ANA 2012 Zack Greinke (Trade) LAD 2012 Adrian Gonzalez (Trade) DET 2014 David Price (Trade) NYM 2015 Yoenis Cespedes (Trade/FA) TOR 2015 David Price (Trade) LAD 2017 Yu Darvish (Trade)

Now, what happens when we incorporate these acquisitions into the model?

```
lm6 <- lm(attendance ~ W + WSWinLastYr + LgWinWSLossLastYr + TopAcq + MidYrAcq, data = tm)
coeftest(lm6, vcovHC(lm6))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        446375.1   177611.7  2.5132              0.01226 *
## W                   23737.9     2254.4 10.5295 < 0.00000000000000022 ***
## WSWinLastYr        551862.9   118036.0  4.6754          0.000003721 ***
```

```
## LgWinWSLossLastYr 260358.7   123949.3  2.1005             0.03615 *
## TopAcq            307413.2    65941.5  4.6619         0.000003962 ***
## MidYrAcq          432019.3   108584.8  3.9786         0.000078882 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cov6         <- vcovHC(lm6)
robust_se    <- sqrt(diag(cov6))
wald_results <- waldtest(lm6, vcov = cov6)

stargazer(lm6, type="latex",
          se = list(NULL, robust_se),
          header = FALSE,
          title = "Model 6 - Attendance, Wins, World Series/Pennant Wins, and Top-End Acquisitions",
          model.numbers=FALSE,
          no.space = TRUE,
          omit.stat = c("rsq", "f"))
```

Table 6: Model 6 - Attendance, Wins, World Series/Pennant Wins, and Top-End Acquisitions

	Dependent variable:
	attendance
W	23,737.900***
	(2,313.620)
WSWinLastYr	551,862.900***
	(139,165.600)
LgWinWSLossLastYr	260,358.700*
	(141,758.400)
TopAcq	307,413.200***
	(76,126.160)
MidYrAcq	432,019.300***
	(133,730.600)
Constant	446,375.100**
	(185,451.900)
Observations	540
Adjusted R^2	0.283
Residual Std. Error	576,747.700 (df = 534)
Note:	*p<0.1; **p<0.05; ***p<0.01

This is very interesting. Both variables produce highly statistically significant results. A top off-season acquisition is associated with a 307,000-person increase in attendance, while a mid-season acquisition is associated with an even larger attendance spike of 432,000.

I tested this model with an interaction term between wins and mid-season acquisitions, as logically, a mid-season top-end acquisition would only occur in conjunction with larger win totals. The resulting term was statistically insignificant, so I dropped it from the regression.

New Ballpark

For my next model, I will add in the opening of a new ballpark, which I would expect would be associated with an attendance bump.

```
lm7 <- lm(attendance ~ W + WSWinLastYr + LgWinWSLossLastYr + TopAcq + MidYrAcq + NewPark, data = tm)
summary(lm7)
```

```
##
## Call:
## lm(formula = attendance ~ W + WSWinLastYr + LgWinWSLossLastYr +
##     TopAcq + MidYrAcq + NewPark, data = tm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1590390  -388278   -43118   386036  1751844
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)         405224     183720   2.206             0.027834 *
## W                    24063       2289  10.510 < 0.0000000000000002 ***
## WSWinLastYr         565534     137655   4.108            0.0000461 ***
## LgWinWSLossLastYr   275984     140234   1.968             0.049582 *
## TopAcq              294011      75362   3.901             0.000108 ***
## MidYrAcq            445346     132281   3.367             0.000816 ***
## NewPark             544100     149812   3.632             0.000309 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570300 on 533 degrees of freedom
## Multiple R-squared:  0.3067, Adjusted R-squared:  0.2989
## F-statistic: 39.29 on 6 and 533 DF,  p-value: < 0.00000000000000022
```

```
coeftest(lm7, vcovHC(lm7))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value              Pr(>|t|)
## (Intercept)       405224.4   175182.1  2.3132               0.02109 *
## W                  24062.8     2224.6 10.8166 < 0.00000000000000022 ***
## WSWinLastYr       565534.4   117946.4  4.7948         0.00000211519 ***
## LgWinWSLossLastYr 275984.2   123538.2  2.2340               0.02590 *
## TopAcq            294011.3    67453.0  4.3588         0.00001570031 ***
## MidYrAcq          445345.9   108139.0  4.1183         0.00004423767 ***
## NewPark           544099.7    99477.0  5.4696         0.00000006945 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cov7         <- vcovHC(lm7)
robust_se    <- sqrt(diag(cov7))
wald_results <- waldtest(lm7, vcov = cov7)

stargazer(lm7, type="latex",
          se = list(NULL, robust_se),
          header = FALSE,
          title = "Model 7 - Attendance, Wins, World Series/Pennant Wins, Top-End Acquisitions, and New
          model.numbers=FALSE,
          no.space = TRUE,
          omit.stat = c("rsq", "f"))
```

As expected, we see a sizable attendance increase associated with a new ballpark: 544,000. Despite its small sample size, it is a highly statistically significant effect. I included the full model summary in this case to see

Table 7: Model 7 - Attendance, Wins, World Series/Pennant Wins, Top-End Acquisitions, and New Ballpark

	Dependent variable:
	attendance
W	24,062.770***
	(2,289.404)
WSWinLastYr	565,534.400***
	(137,655.300)
LgWinWSLossLastYr	275,984.200**
	(140,233.500)
TopAcq	294,011.300***
	(75,362.220)
MidYrAcq	445,345.900***
	(132,280.700)
NewPark	544,099.700***
	(149,812.000)
Constant	405,224.400**
	(183,720.400)
Observations	540
Adjusted R^2	0.299
Residual Std. Error	570,275.100 (df = 533)
Note:	*p<0.1; **p<0.05; ***p<0.01

whether we had seen any improvement in the Adjusted R-squared value. The improvement is modest (0.299), especially considering this value increases automatically upon the addition of variables. There are clearly many more factors that explain the variance in major league attendance, but these are still have interesting results.

Conclusions

As this is an observational study, we cannot draw any causal conclusions from these models. However, there are still some strong relationships here. If we assume, for the sake of argument, that there were causal relationships here, how would a team best proceed in an effort to bolster attendance next year? Well, first off, win the World Series. There we go. Easy. Of course, this is very difficult to do, as is winning the pennant. A team is only going to build a new ballpark once in a generation (for billions of dollars, no less), so that is not what we would call "good business strategy" for increasing attendance alone.

A top-end acquisition, though, is something within a team's power. Ignoring the literal interpretation of the WAR statistic, let's assume that a top-end player acquisition adds five wins to a team's total. Just estimating that effect using the coefficients in our model gives us the following:

```
cat("Effect of acquiring a top-end player on attendance:", (5*lm7$coefficients[2]) + lm7$coefficients[5
```

```
## Effect of acquiring a top-end player on attendance: 414325.2
```

```
cat("Associated ticket revenue: $", 32*((5*lm7$coefficients[2]) + lm7$coefficients[5]))
```

```
## Associated ticket revenue: $ 13258405
```

The result is a bump of over 400,000 to attendance figures and over $13 million in ticket revenue in the first year (using a major league average ticket price of $32). This is before accounting for concessions, merchandise, and, with any luck, playoff appearances.