Does Left Handed Pitching Matter

Carl Ganz

September 26, 2015

I want to explore whether more left handed pitching is better. Let's manipulate `Lahman's` baseball data before we do any modeling.

```
#load packages
library(Lahman)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(ggplot2)
#load datasets as datatables
pitching<-tbl_df(Pitching)
master<-tbl_df(Master)
teams<-tbl_df(Teams)
#select lefthandedness from master table
master<-select(master,playerID,throws)
#filter data for post war era
pitching<-filter(pitching,yearID>1945)
teams<-filter(teams,yearID>1945)
#merge in handedness with pitching data
pitching<-merge(pitching,master,by="playerID")
#generate summary statistics for teams: number of innings pitched by lefties
#and number of games started by lefties for each team each year
leftyip<-pitching %>% group_by(yearID,teamID) %>%
  filter(throws=="L") %>% summarise(leftyips=sum(IPouts),leftystarts=sum(GS))
#merge summary statistics in with team statistics
teams<-merge(teams,leftyip,by=c("yearID","teamID"))
#calculate percentage of innings pitched by lefties
teams$leftypercent<-100*teams$leftyips/teams$IPouts
```

Now lets use linear regressions to model how lefthandness correlates with winning.

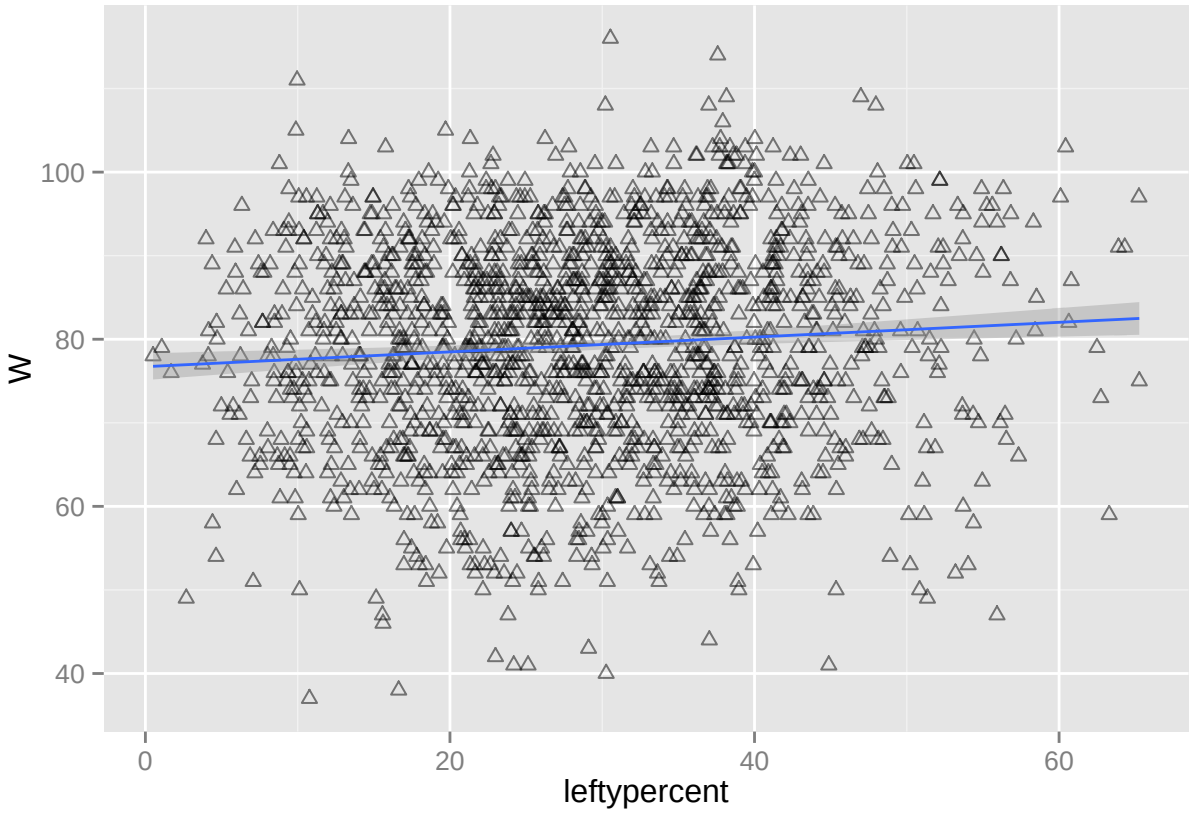
```
lm1<-lm(W~leftypercent,teams)
summary(lm1)
```

```
##
## Call:
## lm(formula = W ~ leftypercent, data = teams)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.670  -8.904   0.653   9.454  36.583
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  76.71696    0.81597  94.019  < 2e-16 ***
## leftypercent  0.08843    0.02610   3.388  0.00072 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.61 on 1653 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.006897,   Adjusted R-squared:  0.006296
## F-statistic: 11.48 on 1 and 1653 DF,  p-value: 0.0007199
```

```
ggplot(teams,aes(x=leftypercent,y=W)) + geom_point(shape=2,alpha=1/2) + geom_smooth(method=lm)
```

```
## Warning: Removed 1 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



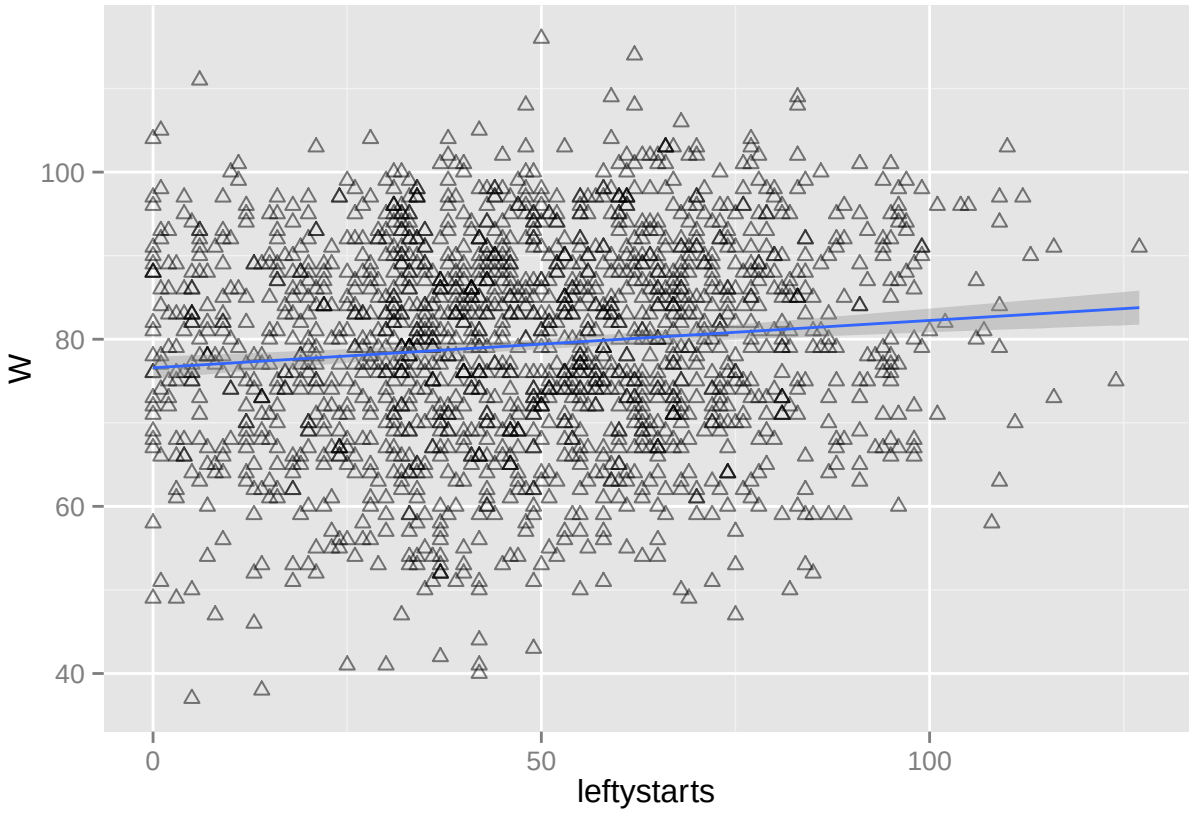
We get a statistically significant positive correlation (each additional percent of innings thrown by left hander is associated with an extra 0.09 wins per season), but as the graph and the R^2 both indicate the proportion of innings pitched by lefties isn't a good predictor of team performance.

Let's look at the number of games started by lefties instead.

```
lm2<-lm(W~leftystarts,teams)
summary(lm2)
```

```
##
## Call:
## lm(formula = W ~ leftystarts, data = teams)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.867  -8.724   0.745   9.364  36.584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 76.58340    0.67154 114.042  < 2e-16 ***
## leftystarts  0.05665    0.01250   4.533 6.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.58 on 1654 degrees of freedom
## Multiple R-squared:  0.01227,    Adjusted R-squared:  0.01167
## F-statistic: 20.55 on 1 and 1654 DF,  p-value: 6.23e-06
```

```
ggplot(teams,aes(x=leftystarts,y=W)) + geom_point(shape=2,alpha=1/2) + geom_smooth(method=lm)
```



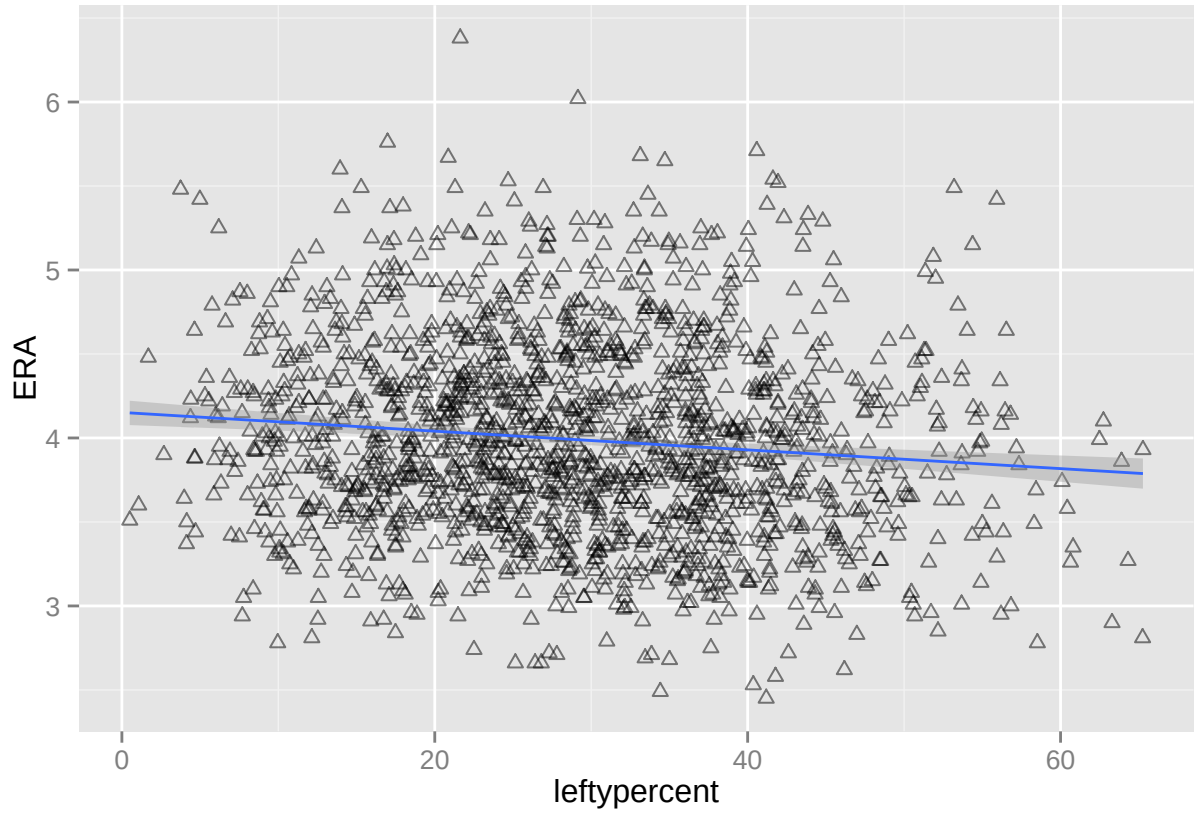
We get similar results. Each additional start by a lefty is associated with an additional 0.06 wins per season.

Let's instead model ERA instead of wins

```
lm3<-lm(ERA~leftypercent,teams)
summary(lm3)
```

```
##
## Call:
## lm(formula = ERA ~ leftypercent, data = teams)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4724 -0.4116 -0.0485  0.3631  2.3486
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.151970   0.037486 110.762  < 2e-16 ***
## leftypercent -0.005575   0.001199  -4.649  3.6e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5794 on 1653 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.01291,    Adjusted R-squared:  0.01231
## F-statistic: 21.62 on 1 and 1653 DF,  p-value: 3.596e-06
```

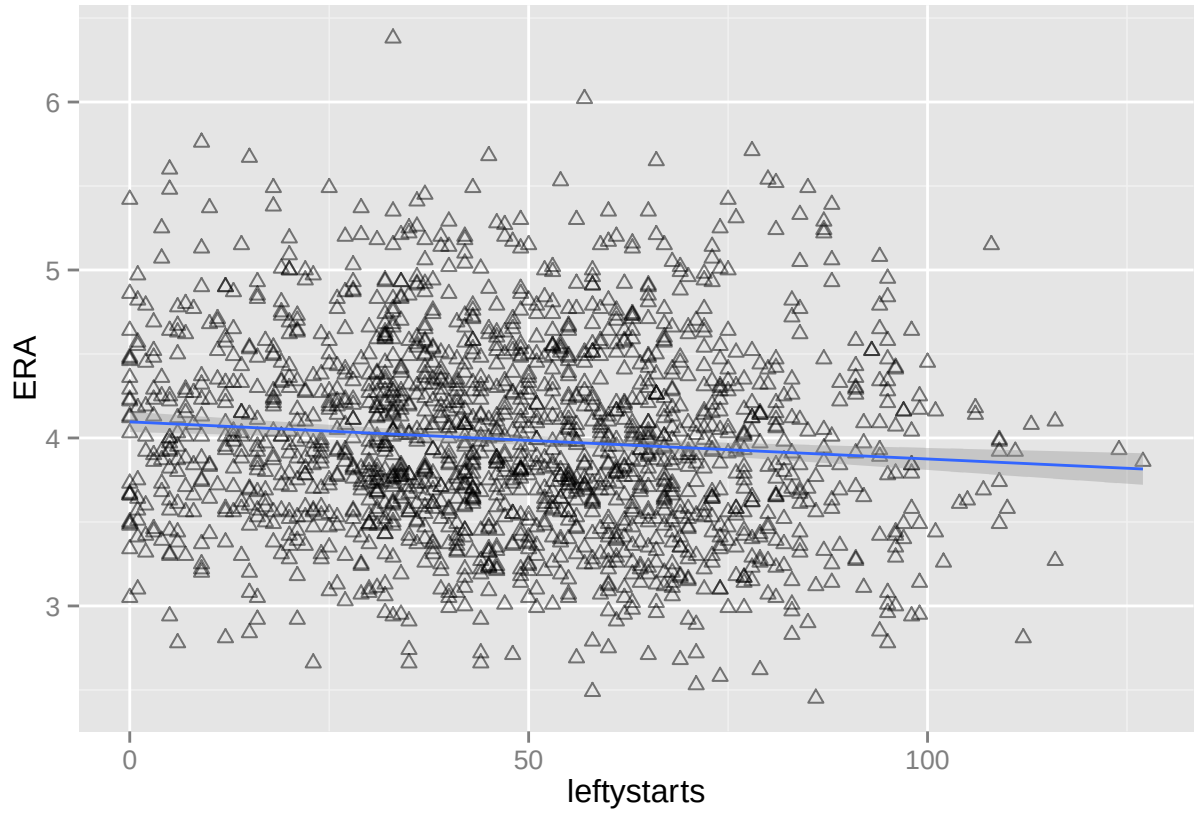
```
ggplot(teams,aes(x=leftypercent,y=ERA)) + geom_point(shape=2,alpha=1/2) + geom_smooth(method=lm)
```




```
lm4<-lm(ERA~leftystarts,teams)
summary(lm4)
```

```
##
## Call:
## lm(formula = ERA ~ leftystarts, data = teams)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47775 -0.41556 -0.04806  0.36801  2.35700
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0959320  0.0309827 132.200  < 2e-16 ***
## leftystarts -0.0022100  0.0005766  -3.833 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5806 on 1654 degrees of freedom
## Multiple R-squared:  0.008805,   Adjusted R-squared:  0.008205
## F-statistic: 14.69 on 1 and 1654 DF,  p-value: 0.0001313
```

```
ggplot(teams,aes(x=leftystarts,y=ERA)) + geom_point(shape=2,alpha=1/2) + geom_smooth(method=lm)
```



Each additional percent of innings pitched by a lefty is associated with 0.006 decline in ERA over the season. Each additional start by a lefty is associated with 0.002 decline in ERA.