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
```

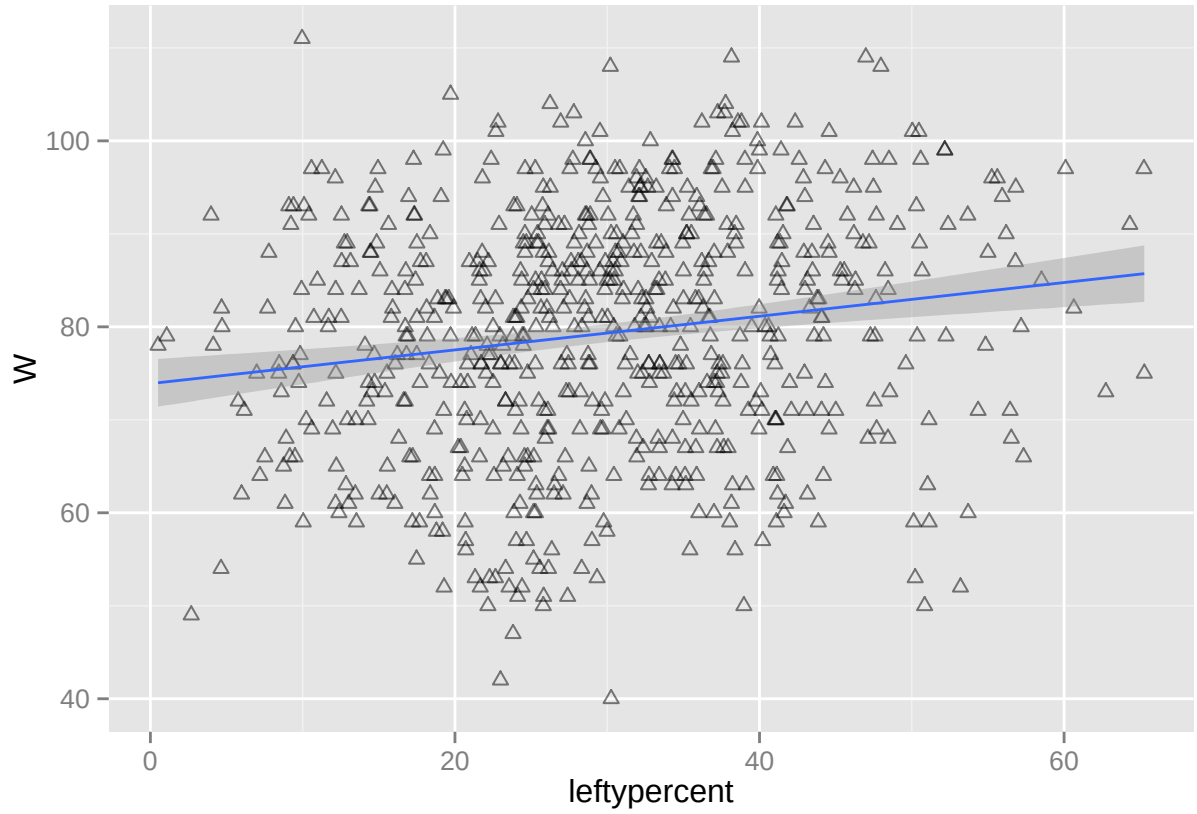
```
library(magrittr)
library(ggplot2)
#load datasets as datatables
pitching<-tbl_df(Pitching)
master<-tbl_df(Master)
teams<-tbl_df(Teams)
#select lefthandedness from master table
master<-select(master,playerID,throws)
#filter data for post war era
pitching<-filter(pitching,yearID>1945 & yearID<1980)
teams<-filter(teams,yearID>1945 & yearID<1980)
#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
## -39.375  -9.570   1.207   9.390  35.302
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.89402    1.32373  55.823  < 2e-16 ***
## leftypercent  0.18115    0.04128   4.389 1.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.83 on 666 degrees of freedom
## Multiple R-squared:  0.02811,    Adjusted R-squared:  0.02665
## F-statistic: 19.26 on 1 and 666 DF,  p-value: 1.325e-05
```

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



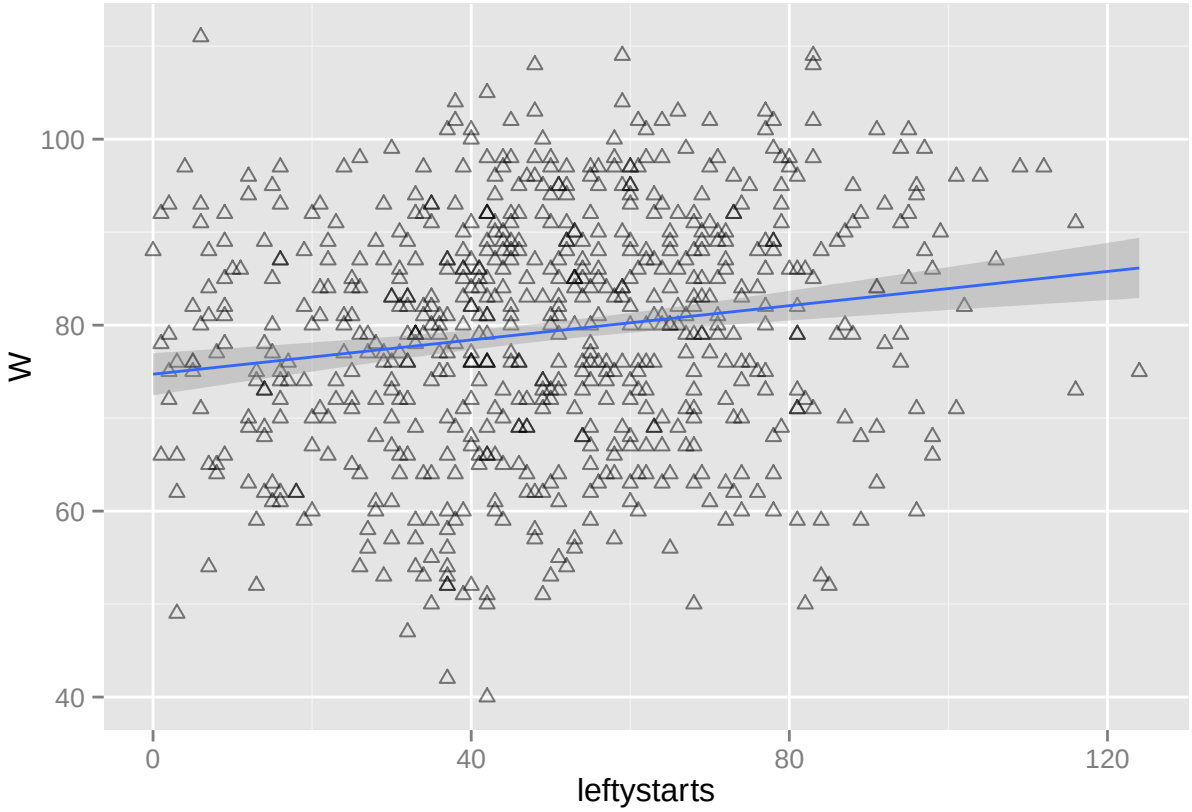
We get a statistically significant positive correlation (each additional percent of innings thrown by left hander is associated with an extra 0.09 wins per season), but as the graph and the R^2 both indicate the proportion of innings pitched by lefties isn't a good predictor of team performance.

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
## -38.594  -9.376   1.215   9.327  35.721
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.72624    1.15338  64.789  < 2e-16 ***
## leftystarts  0.09209    0.02105   4.374 1.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.83 on 666 degrees of freedom
## Multiple R-squared:  0.02792,    Adjusted R-squared:  0.02646
## F-statistic: 19.13 on 1 and 666 DF,  p-value: 1.417e-05
```

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



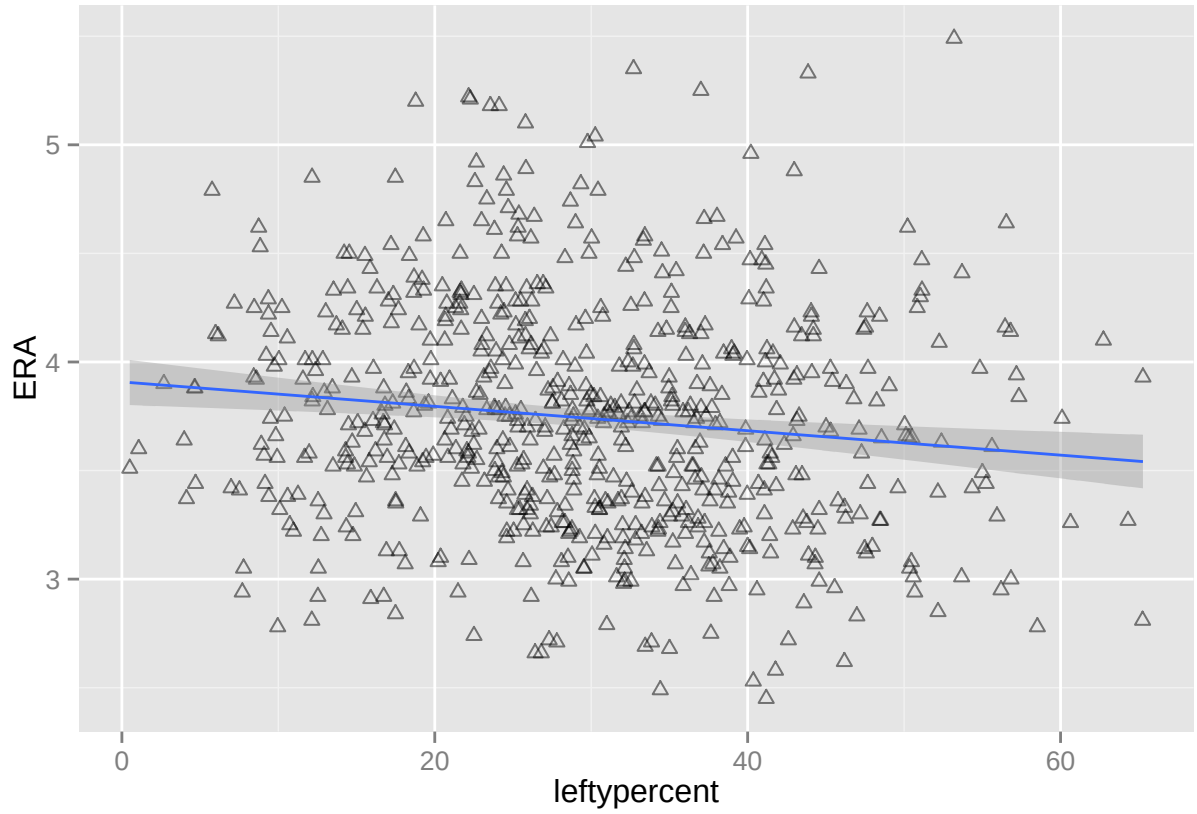
We get similar results. Each additional start by a lefty is associated with an additional 0.06 wins per season.

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
## -1.22669 -0.37185 -0.01032  0.34172  1.88080
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.908151   0.053603  72.909  < 2e-16 ***
## leftypercent -0.005620   0.001671  -3.362 0.000817 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5194 on 666 degrees of freedom
## Multiple R-squared:  0.01669,    Adjusted R-squared:  0.01522
## F-statistic: 11.31 on 1 and 666 DF,  p-value: 0.0008171
```

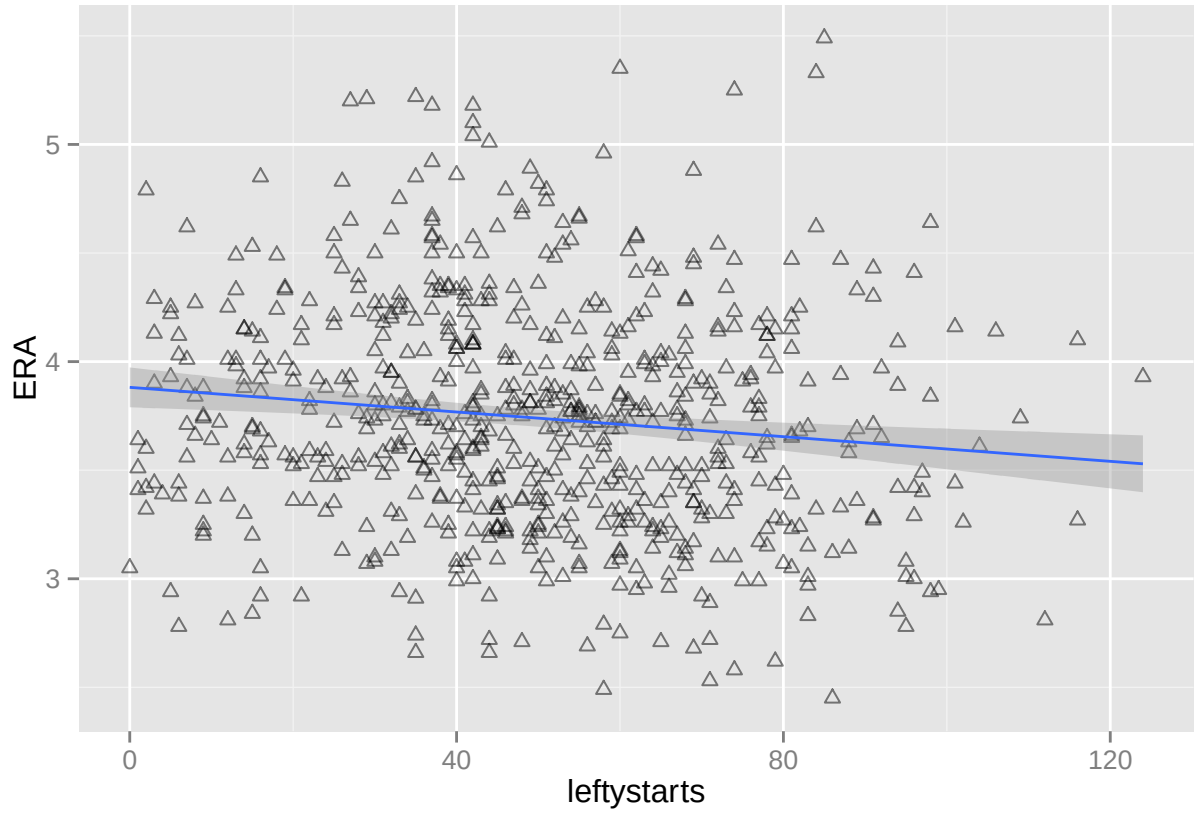
```
ggplot(teams,aes(x=leftypercent,y=ERA)) + geom_point(shape=2,alpha=1/2) + geom_smooth(method=lm)
```



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

```
##
## Call:
## lm(formula = ERA ~ leftystarts, data = teams)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22678 -0.37444 -0.00969  0.32890  1.84988
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.8814655  0.0467079   83.10  < 2e-16 ***
## leftystarts -0.0028393  0.0008526   -3.33 0.000916 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5195 on 666 degrees of freedom
## Multiple R-squared:  0.01638,    Adjusted R-squared:  0.0149
## F-statistic: 11.09 on 1 and 666 DF,  p-value: 0.0009163
```

```
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.