

Source: <http://www.baseball-reference.com>

Project member :

- 1) Tae Kyun Kim: SAS data output analyst
- 2) Hyun Jung Kim: project editing
- 3) Hwan Suk Lee: project design & mentor

Project Design:

Based on the record of 2008 Major league season, we assumed that good batters are batters who has AVG (batting average) higher than league average. Also, we assumed that good pitchers are pitchers who has ERA (earned runs average) lower than league average. From the record, we set AVG and ERA as our two variables. From these two variables, we compare each team and counts hitters who are above the entire season of AVG, and pitchers who are below the entire season of ERA. Then, we find out the relationship between wins and both AVG and ERA. We only considered batters who attended more than 100 times at bat. Also, we only considered pitchers who attended equal or more than 40 innings for same reason. Batters and hitters who have not been satisfied above these conditions are excluded due to the possibility to be outliers.

We analyze 2 correlations: one is batting AVG and WINS, and the other is between ERA and WINS. Figure out each correlation to know which factor (ERA, AVG) affect more on the numbers of WINS.

- i) Present scatter plot
- ii) Present correlation coefficient

Prediction!

In the baseball game, at least the team is going to win at the game when the pitcher played well in the game. However, no matter how well hitters hit the ball; there are still possibilities that the team might lose the game. For those reasons, some people are saying that pitcher has more influence on the number of wins and teams reputation than the hitters.

In order to prove this hypothesis, we did our project following this prediction:

Higher correlation coefficient is more effective to number of WINS. According to our research, we got too low correlations that between number of wins and either our factors AVG or ERA have too weak relationships. So we infer there must be some other lurking variables that lower our results. To get more clear correlation, we added one more variable to each pitcher and hitter variable, and find out another correlation coefficient by putting in these two variables in our study: WHIP (Walks and Hits per Innings Pitched) that is related to pitchers' ability, and OPS (On-Base Percentage + Slugging Percentage) that is related to hitters' ability.

Pitcher who has lower WHIP tends not to load bases, so make fielders less pressure of defense. Thus, we assumed that the pitcher who has lower WHIP contribute team's win as same as

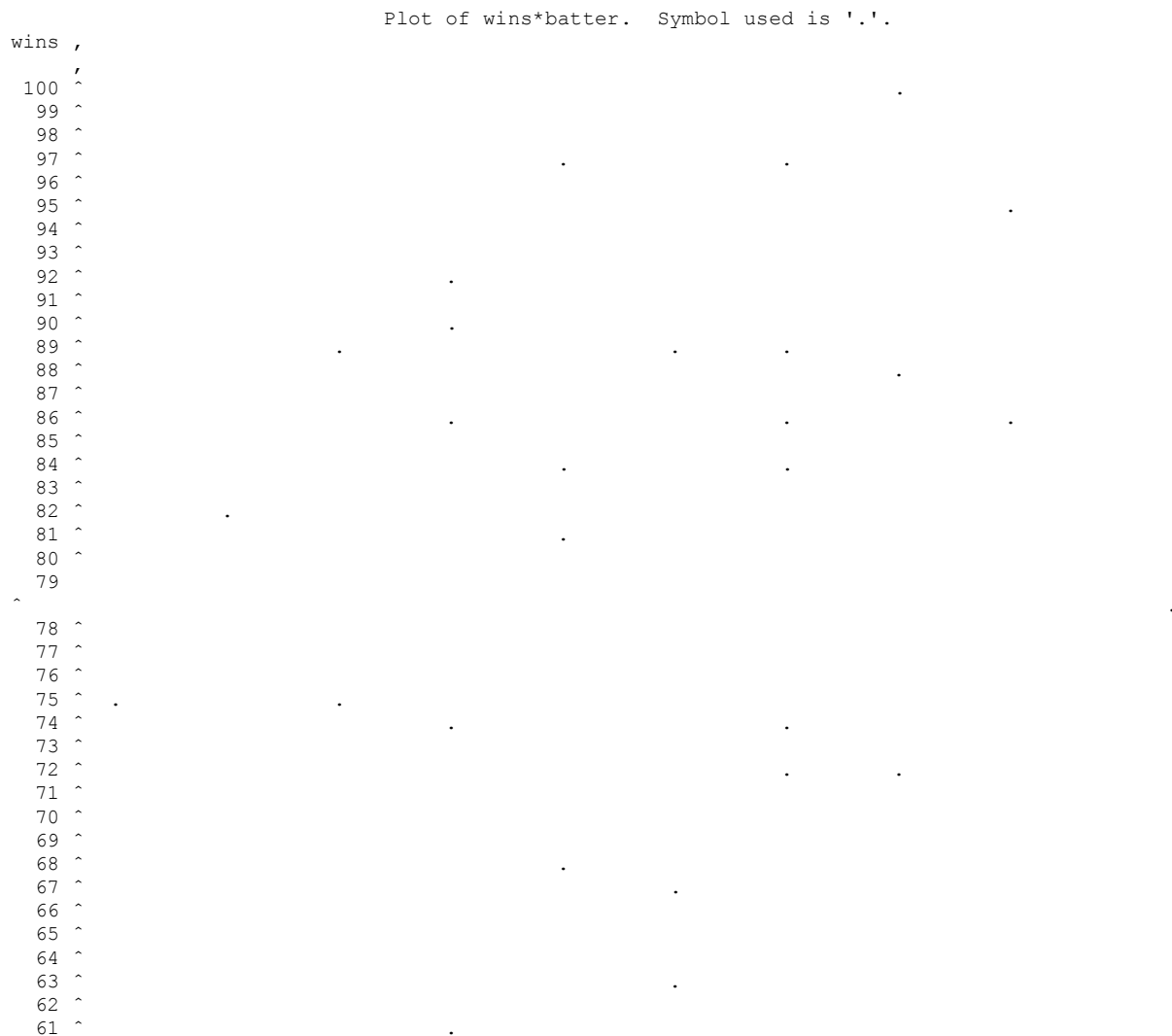
the pitcher who has lower ERA. On the other hand, OPS is generally used for defining power hitter. Of course, power hitter tends to make more extra base hits than just contact hitter and extra base hit has more possibility that runner makes score than just single base hit. Hence, power hitters have high slugging percentage as well. Also, for preventing extra base hit and power hitters usually hit so hard and lose their contact, pitchers threw more balls to outside of strike zone against power hitter than contact hitter, power hitters usually have more four-ball than contact hitters. Thus, we assumed that OPS, which is the sum of hitter's On Base Percentage and Slugging Percentage contributes for team's win. We only included WHIP that is above the season average and OPS that is below the season average.

First, we found out correlation between Average at bat (AVG) and number of Wins and also about Earned Runs Average (ERA) and Wins

SAS output indicate that two variables (AVG and ERA) do not show strong relationship with number of Wins.

The SAS System

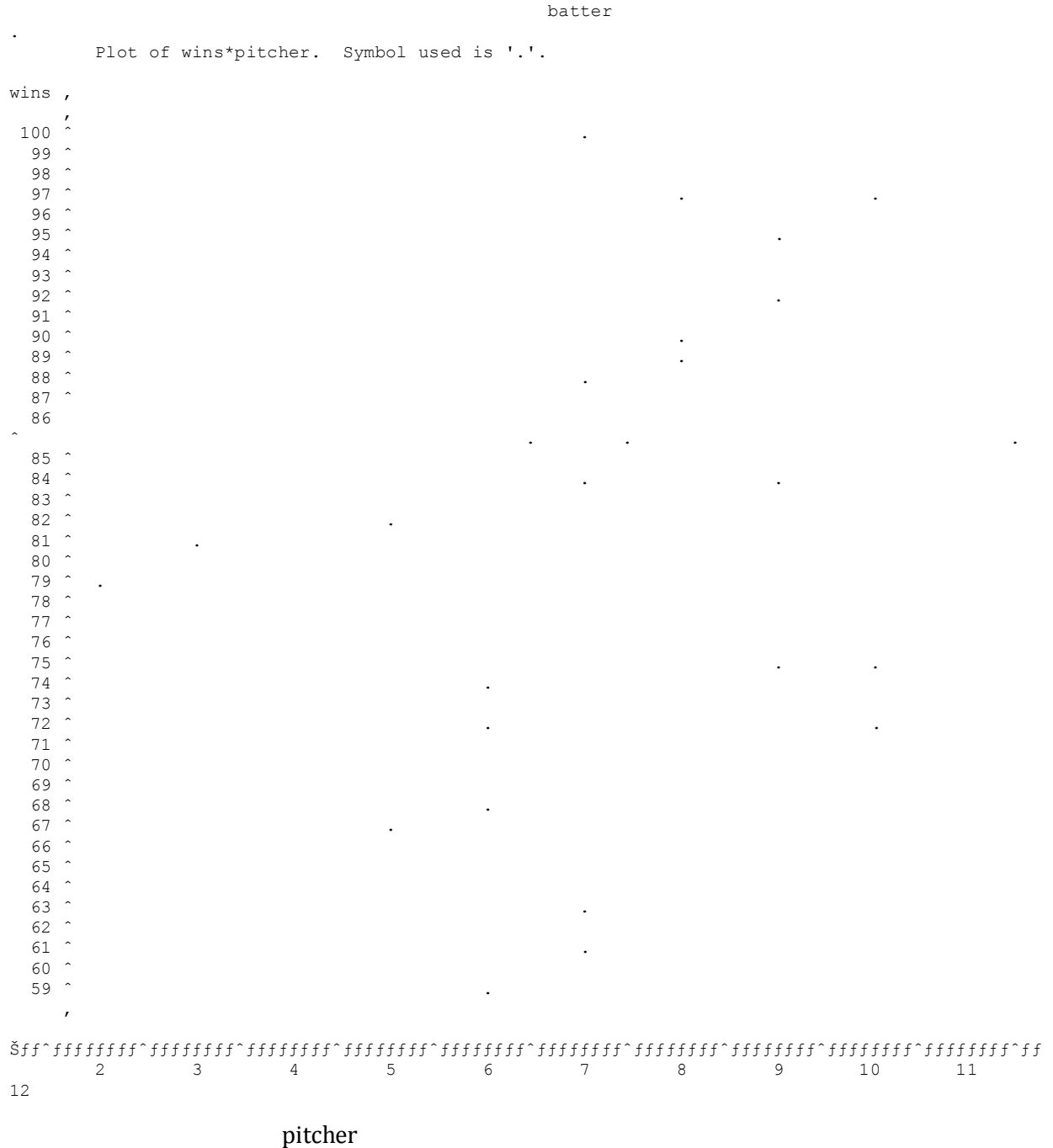
08:54 Monday, April 27, 2009 1



```

60 ^
59 ^
,
Šff^ffffffff^ffffffff^ffffffff^ffffffff^ffffffff^ffffffff^ffffffff^ffffffff^ffffffff^ff
3 4 5 6 7 8 9 10 11 12
13

```



Batters and wins: point estimate r is 0.18041

Pitchers and wins: point estimate r is 0.33924

Patterns of scatter plots do not show the linear pattern which means these two variables do not show consistency.

Through the result, pitchers' absolute value of r is higher than that of batters.

However, correlations are too low to evaluate precise effect of between two variables. Therefore we can assume that there are other lurking variables that are affect the number of wins. So we are including other variables such as WHIP and OPS to see the relationship and correlation with the number of wins.

So after considering these additional values, we get the results;

Correlation

BOPS (batter + OPS) and wins: point estimate r is 0.25455; r -square is 0.0638,

95% Confidence interval is 54.54347 to 87.6493

PWHIP (pitcher + WHIP) and wins: point estimate r is 0.50532; r -square is 0.3106,

95% Confidence interval is 48.69236 to 76.25275

It is clear that the correlation of BOPS and wins is higher than that of batters and wins. Also, the correlation of PWHIP and wins is higher than that of pitchers and wins.

What if we contain more variables in our experiments above? Are we going to get better results?

Yes, we will get more trusted results. Therefore, we concluded that:

It is a fact that good pitchers and batters affect the number of wins. However, we cannot judge the number of wins by these two or more variables. For example, teams' concentration, players' patience or endurance, the weather of the day, etc. are influencing variables which can give us better results. However, we cannot express all variables to the number. If we can put all of the variables that can be expressed as number, we would get higher correlations.

SAS code:

```
DATA project ;
input name $ batter pitcher OPS WHIP BOPS PWHIP wins ;
datalines ;

CHC 9 10 9 8 19 18 97
NYM 9 8 8 7 17 15 89
PHI 6 9 8 8 14 17 92
STL 11 8 9 6 20 14 86
FLA 7 7 8 7 15 14 84
ATL 10 10 10 8 20 18 72
MIL 6 8 8 6 14 14 90
COL 9 6 10 5 19 11 74
PIT 8 5 6 5 14 10 67
ARI 4 5 8 9 12 14 82
HOU 9 7 5 9 14 16 86
CIN 6 6 8 5 14 11 74
LAD 9 9 9 7 18 16 84
WSN 7 6 8 4 15 10 59
SFG 9 6 8 4 17 10 72
SDP 8 7 5 8 13 15 63
TEX 13 2 9 1 22 3 79
BOS 11 9 8 10 19 19 95
MIN 10 7 4 5 14 12 88
DET 9 6 10 3 19 9 74
CHW 5 8 6 8 11 16 89
CLE 7 3 5 4 12 7 81
NYY 8 8 8 6 16 14 89
BAL 7 6 6 3 13 9 68
TBR 7 8 8 8 15 16 97
LAA 10 7 8 8 18 15 100
TOR 6 12 5 7 11 19 86
KCR 5 9 3 5 8 14 75
SEA 6 7 3 4 9 11 61
OAK 3 10 1 6 4 16 75
;
run ;

proc plot data = project ;
plot wins * batter = '.' ;
run ;
proc plot data = project ;
plot wins * pitcher = '.' ;
run ;
proc corr data = project ;
var batter wins ;
run ;
proc corr data = project ;
var pitcher wins ;
run ;
proc corr data = project ;
var BOPS wins ;
run ;
proc corr data = project ;
var PWHIP wins ;
```

```
run ;
```

```
proc reg ;
```

```
model wins = batter ops / clb ;
```

```
run ;
```

```
proc reg ;
```

```
model wins = pitcher whip / clb ;
```

```
run ;
```

```
proc plot data = project ;
```

```
plot wins * bops = '.' ;
```

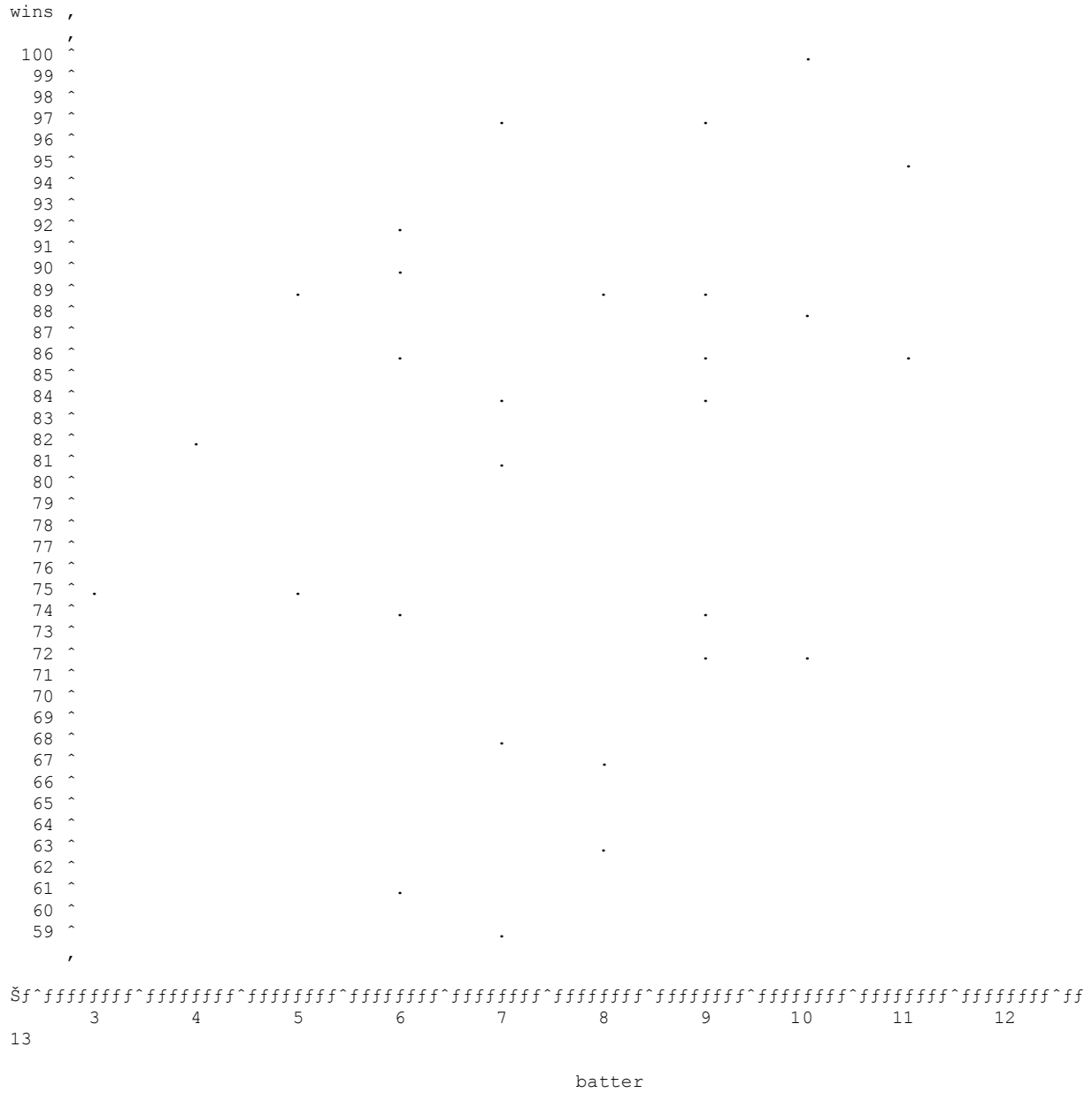
```
run ;
```

```
proc plot data = project ;
```

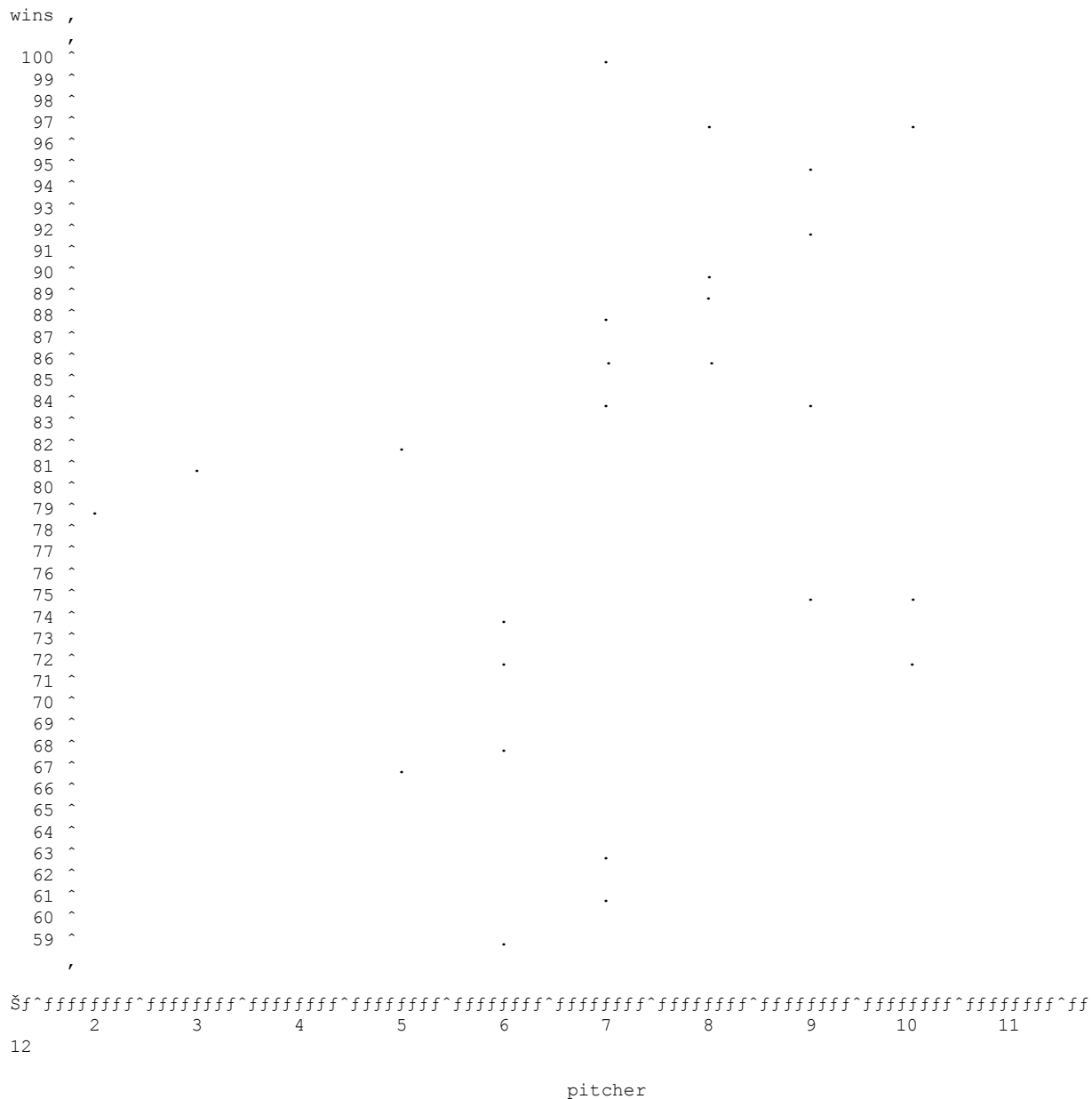
```
plot wins * pwhip = '.' ;
```

```
run ;
```

Plot of wins*batter. Symbol used is '.'.



Plot of wins*pitcher. Symbol used is '.'.



NOTE: 4 obs hidden.

The CORR Procedure

2 Variables: batter wins

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
batter	30	7.80000	2.23453	234.00000	3.00000	13.00000
wins	30	80.93333	11.08566	2428	59.00000	100.00000

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	batter	wins
batter	1.00000	0.18041 0.3401
wins	0.18041 0.3401	1.00000

The CORR Procedure

2 Variables: pitcher wins

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pitcher	30	7.30000	2.08690	219.00000	2.00000	12.00000
wins	30	80.93333	11.08566	2428	59.00000	100.00000

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	pitcher	wins
pitcher	1.00000	0.33924 0.0667
wins	0.33924 0.0667	1.00000

The CORR Procedure

2 Variables: BOPS wins

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
BOPS	30	14.86667	3.98041	446.00000	4.00000	22.00000
wins	30	80.93333	11.08566	2428	59.00000	100.00000

Pearson Correlation Coefficients, N = 30

Prob > |r| under H0: Rho=0

	BOPS	wins
BOPS	1.00000	0.25455 0.1746
wins	0.25455 0.1746	1.00000

The CORR Procedure

2 Variables: PWHIP wins

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PWHIP	30	13.43333	3.72950	403.00000	3.00000	19.00000
wins	30	80.93333	11.08566	2428	59.00000	100.00000

Pearson Correlation Coefficients, N = 30
 Prob > |r| under H0: Rho=0

	PWHIP	wins
PWHIP	1.00000	0.50532 0.0044
wins	0.50532 0.0044	1.00000

The REG Procedure
 Model: MODEL1
 Dependent Variable: wins

Number of Observations Read 30
 Number of Observations Used 30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	227.42879	113.71439	0.92	0.4106
Error	27	3336.43788	123.57177		
Corrected Total	29	3563.86667			

Root MSE	11.11628	R-Square	0.0638
Dependent Mean	80.93333	Adj R-Sq	-0.0055
Coeff Var	13.73511		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits
Intercept	1	71.09642	8.06740	8.81	<.0001	54.54347
batter	1	0.34831	1.08851	0.32	0.7514	-1.88513
OPS	1	1.01233	1.06605	0.95	0.3507	-1.17501

The REG Procedure
 Model: MODEL1
 Dependent Variable: wins

Number of Observations Read 30
 Number of Observations Used 30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1106.89596	553.44798	6.08	0.0066
Error	27	2456.97071	90.99892		
Corrected Total	29	3563.86667			

Root MSE	9.53934	R-Square	0.3106
Dependent Mean	80.93333	Adj R-Sq	0.2595
Coeff Var	11.78666		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits
Intercept	1	62.47255	6.71605	9.30	<.0001	48.69236 76.25275
pitcher	1	0.14218	1.03939	0.14	0.8922	-1.99048 2.27484
WHIP	1	2.84068	1.02660	2.77	0.0101	0.73426 4.94710