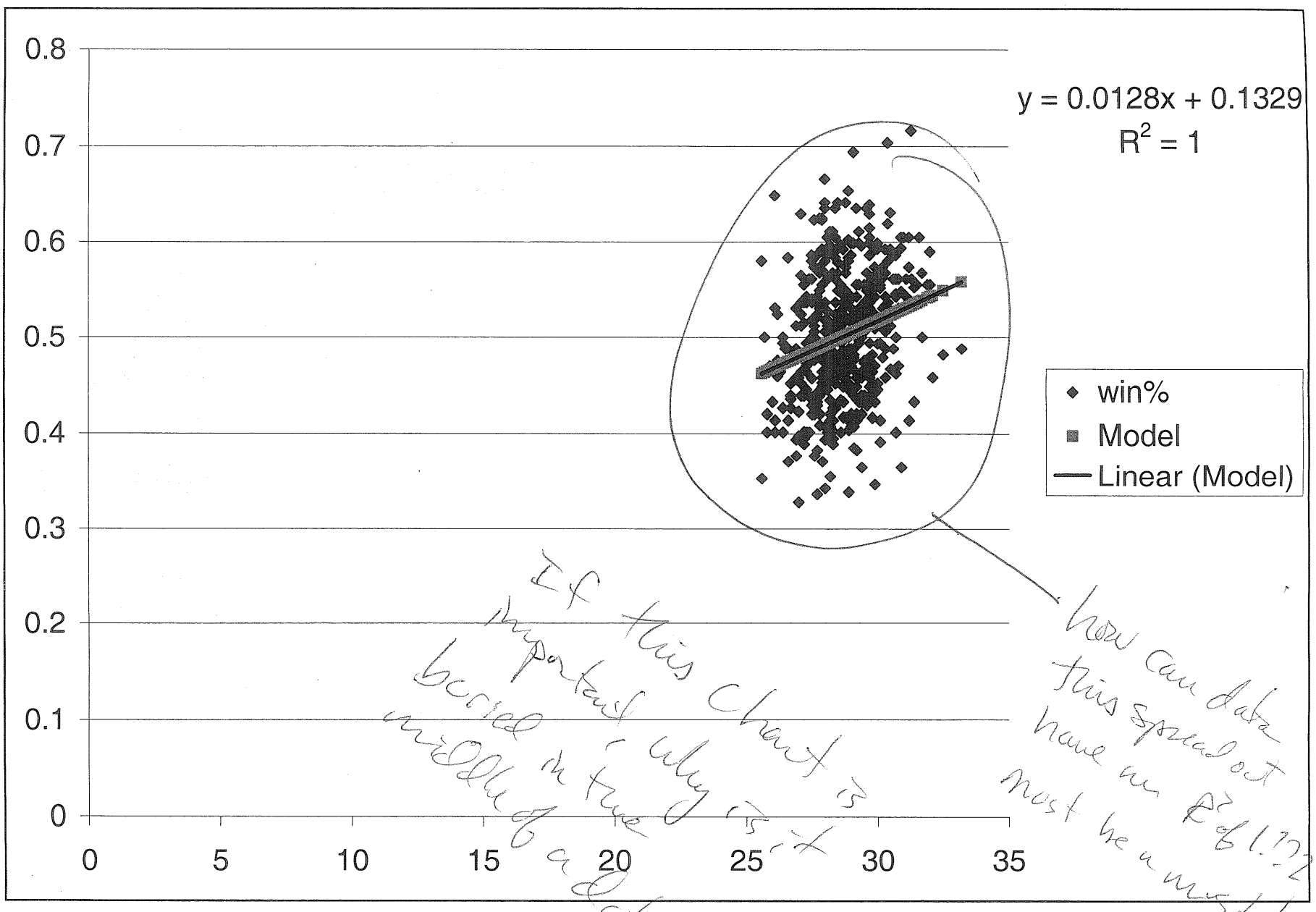


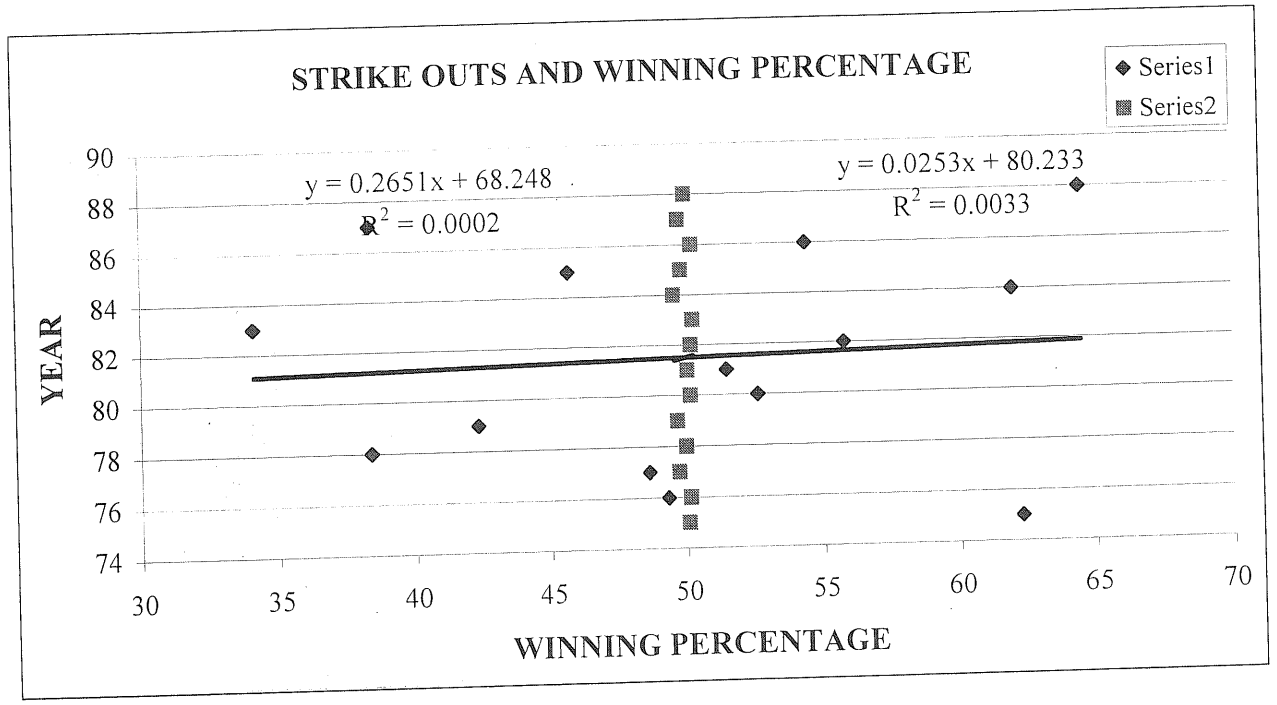

If this chart is important, why is it buried in the middle of a data set?

how can data this spread out have an $R^2$ of 1.??
must be a mistake

## STRIKE OUTS

| YEAR | STRIKE OUTS | WINNING PERCENT | WINNING PERCENT (EST) | ERROR | ERROR SQUARED |
|---|---|---|---|---|---|
| 75 | 773 | 62.3 | 50.06 | 12.24 | 149.82 |
| 76 | 752 | 49.3 | 50.14 | -0.84 | 0.70 |
| 77 | 862 | 48.6 | 49.72 | -1.12 | 1.26 |
| 78 | 791 | 38.4 | 49.99 | -11.59 | 134.38 |
| 79 | 875 | 42.3 | 49.67 | -7.37 | 54.39 |
| 80 | 743 | 52.6 | 50.17 | 2.43 | 5.89 |
| 81 | 773 | 51.5 | 50.06 | 1.44 | 2.07 |
| 82 | 736 | 55.8 | 50.20 | 5.60 | 31.36 |
| 83 | 714 | 34.1 | 50.28 | -16.18 | 261.88 |
| 84 | 898 | 62 | 49.59 | 12.41 | 154.05 |
| 85 | 821 | 45.7 | 49.88 | -4.18 | 17.46 |
| 86 | 715 | 54.4 | 50.28 | 4.12 | 16.98 |
| 87 | 842 | 38.4 | 49.80 | -11.40 | 129.95 |
| 88 | 776 | 64.5 | 50.05 | 14.45 | 208.84 |

| | |
|---|---|
| SSE | 1169.05 |
| SLOPE | 0.00 |
| Y INT | 52.98 |
| R^2 | 0.0002 |



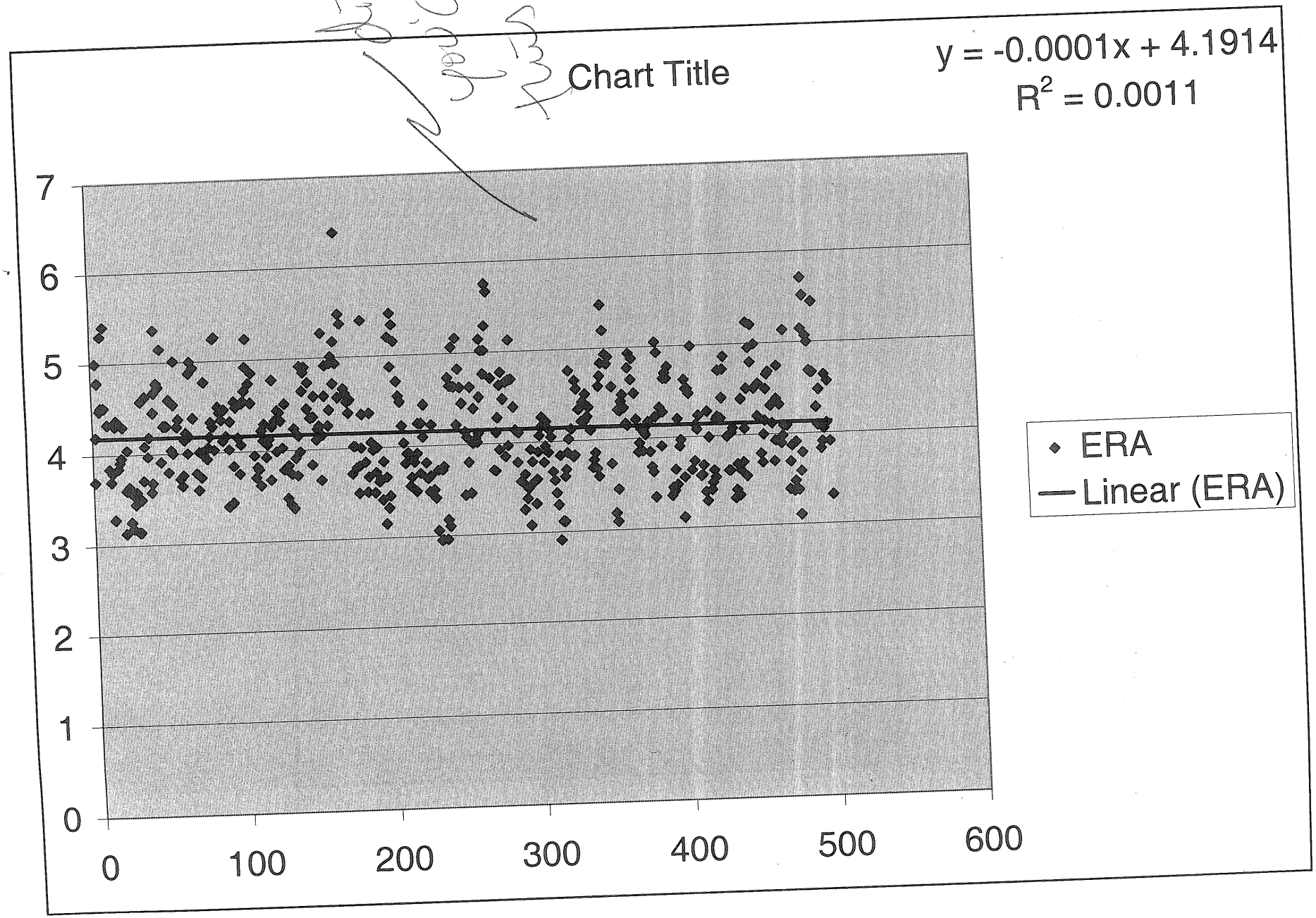
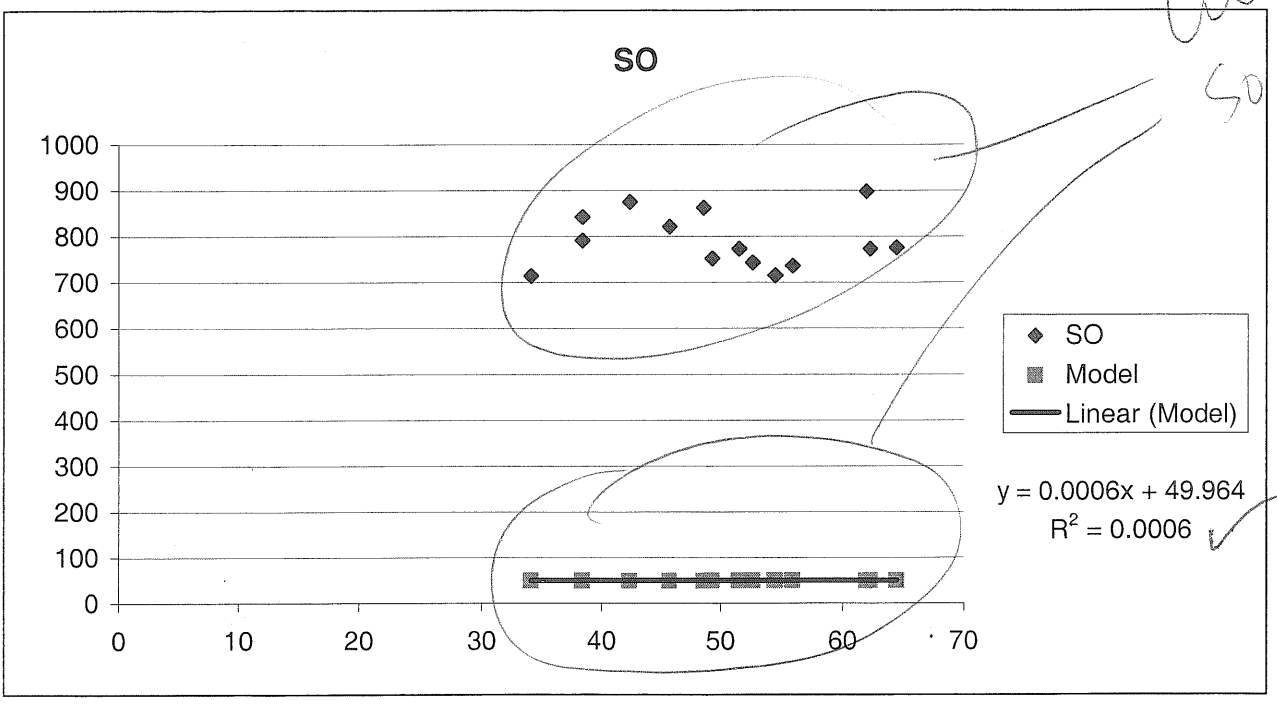


Chart Title
y = -0.0001x + 4.1914
R² = 0.0011
7
6
5
4
3
2
1
0
0
100
200
300
400
500
600
ERA
Linear (ERA)

| Win Pct | SO | Model | Error | Error^2 | SSE |
|---|---|---|---|---|---|
| 62.3 | 773 | 50.05989 | -12.2401 | 149.8202 | 1169.048 |
| 49.3 | 752 | 50.13916 | 0.839161 | 0.704191 | |
| 48.6 | 862 | 49.72395 | 1.123951 | 1.263267 | |
| 38.4 | 791 | 49.99195 | 11.59195 | 134.3733 | |
| 42.3 | 875 | 49.67488 | 7.374881 | 54.38887 | |
| 52.6 | 743 | 50.17313 | -2.42687 | 5.889684 | |
| 51.5 | 773 | 50.05989 | -1.44011 | 2.073906 | |
| 55.8 | 736 | 50.19956 | -5.60044 | 31.36498 | |
| 34.1 | 714 | 50.2826 | 16.1826 | 261.8765 | |
| 62 | 898 | 49.58806 | -12.4119 | 154.0561 | |
| 45.7 | 821 | 49.87871 | 4.178711 | 17.46163 | |
| 54.4 | 715 | 50.27882 | -4.12118 | 16.9841 | |
| 38.4 | 842 | 49.79944 | 11.39944 | 129.9473 | |
| 64.5 | 776 | 50.04857 | -14.4514 | 208.8438 | |

| | |
|---|---|
| **Slope** | -0.00377 |
| **Intercept** | 52.97769 |

why so different




In SLG versus SO, the SLG model is a more accurate representation of winning percentage because its SSE is smaller than SO's, respectively 877.6364 and 1169.048. In addition, the R^2 value for SLG (R^2=.2497) gives a better representation of actuality. The SO's R^2 value (R^2=.0006) is smaller than the SLG's, but like the linear model it does not accurately depict the best model in this case.

Discussion of minimizing SSE and best linear model of different predictor variables for win percentage of a hypothetical baseball team.

I chose the variables batting average and slugging average to model the relationship between these variables and the winning percentage of a baseball team. Comparing these two resulting models of minimized error of fitting a line to the respective data sets has convinced me that batting average is a better predictor of the team's winning percentage. There are many pieces of information and data analysis that support this assertion. First, the lower value for the Sum of Squared Error of fitting a line of best fit to the data for winning percentages at certain values for batting average show less variation between the data points and the line of best fit that minimizes variability between them than does the minimized value for fitting a line to winning percentages at certain slugging percentage values.

The Sum of Squared Error for batting average in relation to win percentage describes the amount of difference between each of the data points for win percentage at varying batting average values and the linear line of best fit that minimizes this variation, or error, in the model. The resulting linear model is given by the equation:
$Y = 6.2527\,x - 115.44$, where y is the winning percentage at different values of x, batting average. Batting average is thus the predictor variable in this model. The value for SSE is 784.4458 for this data set, lower than the value for relating slugging percentage and win percentage.

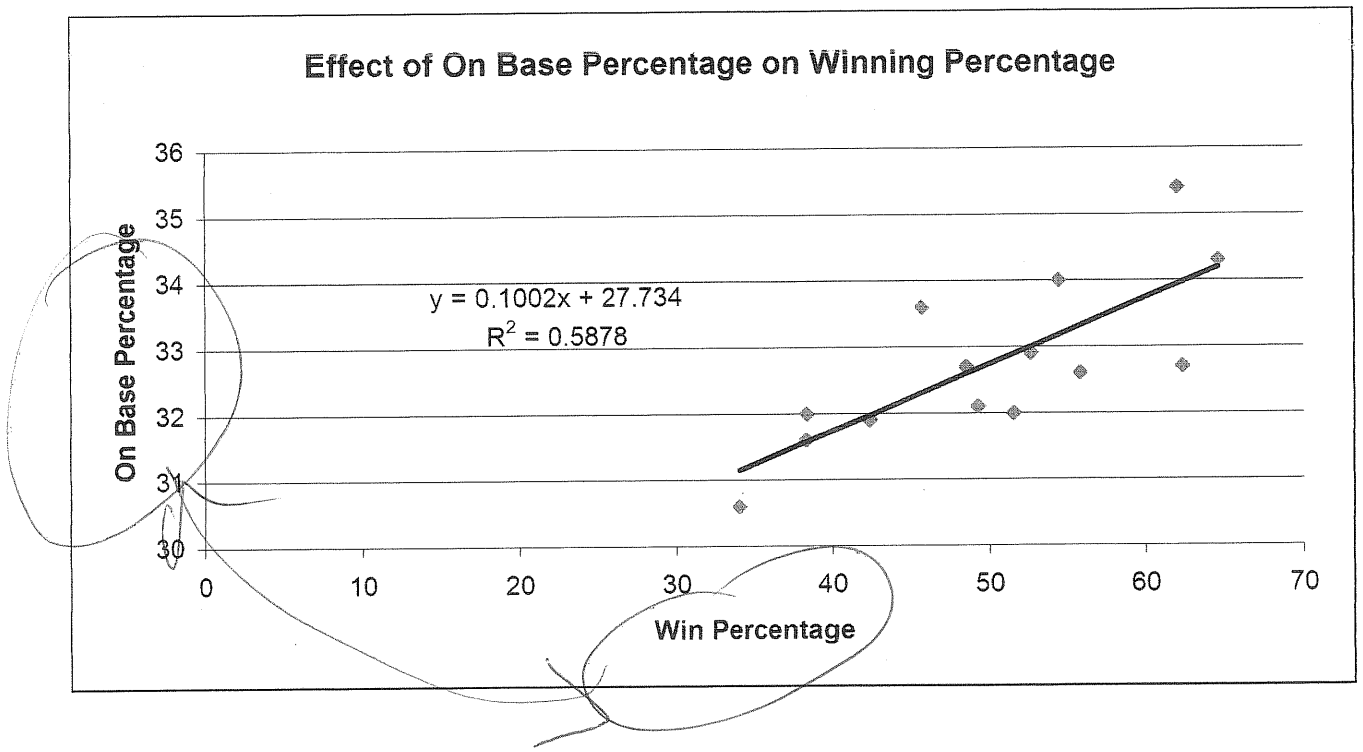
The minimized SSE for relating slugging percentage to win percentage is given by the linear model:
$Y = 2.3881\,x - 45.649$, where y is the winning percentage at different values of x, slugging average. The value for SSE for this data set is 877.6364, showing the model to have more error and thus be less efficient at predicting the win percentage of the team.

The second related piece of data to support batting average as the superior predictor of win percentage is the proportion of variability in the win percentage that is explained by the value of batting average. This is explained by the $R^2$ value. The higher $R^2$ value of .3294 for batting average shows that the variation of the win percentage is explained better by the batting average than that of the slugging percentage, .2497. The lower value for slugging percentage shows that there is a higher amount of variation in the data fitting that is not explained by the relation of slugging percentage to win percentage. The higher value for batting average of shows that there is less variation in the data fitting that is not explained by the relation of batting average to win percentage.

The results of such an analysis might be useful when teams are deciding where to allocate resources to improve areas of the team that might be more important to winning games. This data analysis shows that the team's batting average is a better predictor of overall win percentage than is its slugging percentage. This might convince the team to spend more money on reliable, on-base hitters, as opposed to the type of hitter who has more doubles and triples than lends to a good batting average. Such a model could be used to make goals for the season, by setting goals for batting average, the team can have an idea of a realistic expectation for win percentage. It makes itself a powerful tool for evaluating possible ways to allocate resources to maximize results.

Good

| Win Percentage | OBA | Linear Model | Error | e Squared | SSE |
|---|---|---|---|---|---|
| 62.3 | 32.7 | 31.01054 | -1.68946 | 2.854275092 | 57.97686341 |
| 49.3 | 32.1 | 30.95042 | -1.14958 | 1.321534176 | |
| 48.6 | 32.7 | 31.01054 | -1.68946 | 2.854275092 | *Variance* |
| 38.4 | 31.6 | 30.90032 | -0.69968 | 0.489552102 | 4.141204529 |
| 42.3 | 31.9 | 30.93038 | -0.96962 | 0.940162944 | |
| 52.6 | 32.9 | 31.03058 | -1.86942 | 3.494731136 | |
| 51.5 | 32 | 30.9404 | -1.0596 | 1.12275216 | |
| 55.8 | 32.6 | 31.00052 | -1.59948 | 2.55833627 | |
| 34.1 | 30.6 | 30.80012 | 0.20012 | 0.040048014 | *Slope* |
| 62 | 35.4 | 31.28108 | -4.11892 | 16.96550197 | 0.1002 |
| 45.7 | 33.6 | 31.10072 | -2.49928 | 6.246400518 | |
| 54.4 | 34 | 31.1408 | -2.8592 | 8.17502464 | *Intercept* |
| 38.4 | 32 | 30.9404 | -1.0596 | 1.12275216 | 27.734 |
| 64.5 | 34.3 | 31.17086 | -3.12914 | 9.79151714 | |



Good summary

For my evaluation I chose to analyze the impact of various baseball teams' batting averages and errors on their winning percentages. I began by plotting winning percentage as a function of batting average. These data have an SSE of 784.45 and an $R^2$ of .3294, meaning that 32.94% of the winning percentages can be explained by their corresponding batting average. I then did the same analysis, this time plotting winning percentage as a function of errors. The SSE for this comparison is 942.49 and the $R^2$ is .1934. The winning percentage as a function of batting average has a lower SSE and a higher $R^2$, and therefore this model is a better predictor than the error-win percentage one, though it is not a great predictor itself. We could use such a model and look at the correlation and the correlation coefficient to visualize how much a team's winning percentage can be steeped in it's players' batting capability.

Any reason why BA is better than E?

Fig. 1: Batting Avg. vs. Win Pct. (with SSE, $R^2$)

| | | | |
|---|---|---|---|
| Slope | 6.252681 | SSE Line | 784.4458 |
| Intercept | -115.435 | SSE y | 1169.709 |
| | | my R^2 | 0.329367 |
| | | comp. R^2 | 0.3294 |

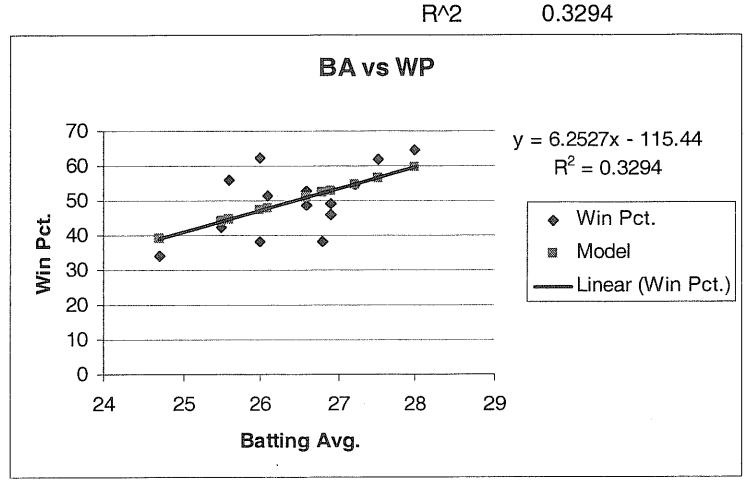
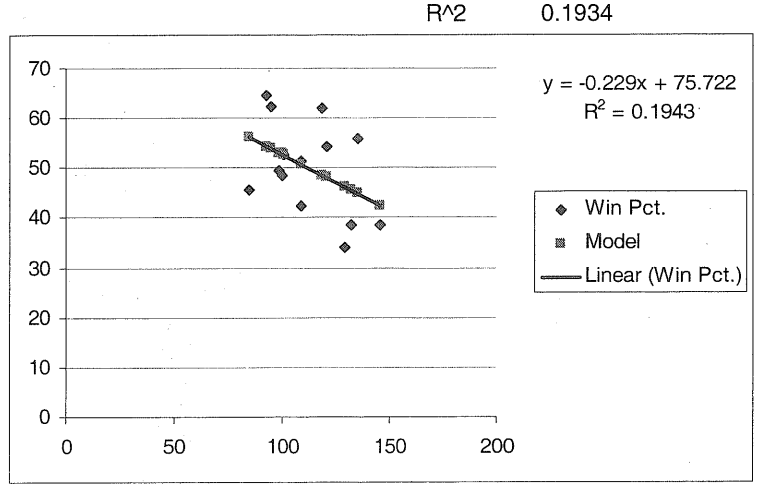



Fig. 2: Errors vs. Win Pct. (with SSE, $R^2$)

| | | | |
|---|---|---|---|
| Slope | -0.22899 | SSE Line | 942.4903 |
| Intercept | 75.72154 | SSE y | 1169.709 |
| | | my R^2 | 0.194253 |
| | | comp. R^2 | 0.1934 |

Nice embedding