

Logistic Regression James Harden

Jason Spector, Jericho Lawson, and Dr. Ekstrom

Purpose: Each activity illustrates a little sports analytics, a little statistics and a little *R*.

Sports Analytics: Examine the relationship between the number of points James Harden scored and the likelihood of the Houston Rockets winning.

Statistics: logistic regression, sensitivity, specificity

R: Continue our development and reinforcement of R skills, including: reading a csv, logistic modeling, and plotting.

Introduction

Having a game-changing player in the NBA can affect how successful his team is throughout the season. James Harden is no exception. While he was with the Rockets, he has made a name for himself through lots and lots of scoring (even if plenty of those points came from the free throw line). With that said, how well does James Harden need to do to ensure that the Rockets get a win? We will use logistic regression to answer this question!

What is Logistic Regression?

The term logistic curve came into being in the 1840's as mathematicians tried to model population growth in an environment with limited resources. Essentially, this meant that such curves had two limiting values, zero, and what they called carrying capacity.

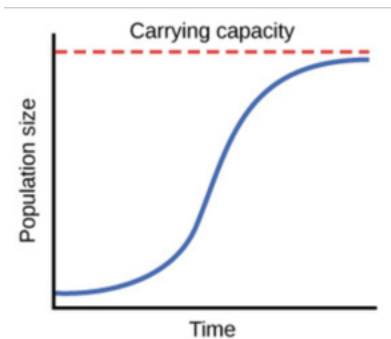


Figure 1: Typical Logistic Curve

It didn't take long before mathematicians and statisticians used such curves to model similar situations, in particular, situation where there are two distinct levels of output (binary variables). In sports, examples of such variables include game results (win or lose), shots on goal (score or save), field goals attempted (made or miss), etc... In logistic regression we are going to try to predict the output of such a binary variable using other variables. Such as in soccer, use the distance of the shot on goal attempted to predict whether it results in a goal or a save. Or in baseball, use velocity of a pitch in the strike zone to predict whether it results in contact or not.

Demonstration - Points by James Harden and Wins by his Rockets

In this demonstration, we are going to examine the relationship between the number of points scored by James Harden and whether his Rockets would win or lose. The data comes from 545 games played by James Harden and the Houston Rockets from October 31, 2012 to April 9, 2019.

Reading and Preparing the Data

```
harden = read.csv("harden_revised.csv", header = TRUE)
```

As you look at the data, you will see the last variable (Dummy) is the binary variable (with 1 for a win and 0 for a loss). We will also want to add color based on the variable Result. So the next will make sure that R reads Result as a categorical variable (with two categories, W and L).

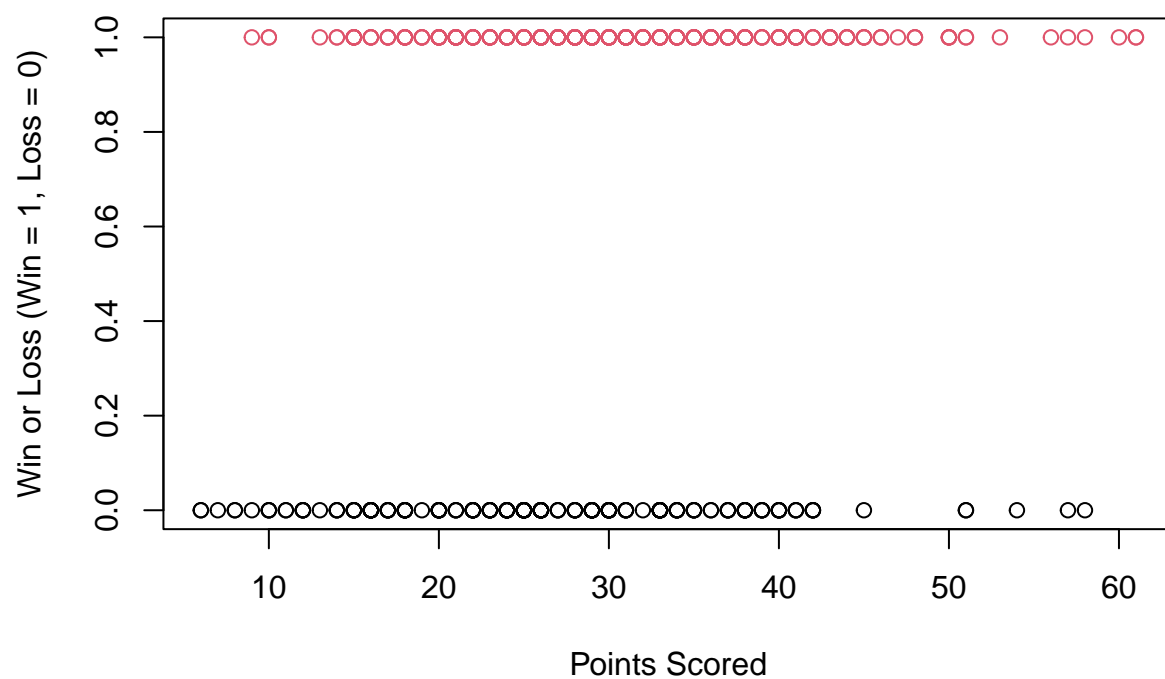
```
harden$Result = as.factor(harden$Result)
```

Plotting the Data

To analyze how Harden's points affects whether the Rockets get wins, we begin with a simple plot.

```
plot(harden$Dummy~harden$PTS, col = harden$Result,  
     xlab = "Points Scored", ylab = "Win or Loss (Win = 1, Loss = 0)",  
     main = "Rockets Results based on Points Scored by James Harden")
```

Rockets Results based on Points Scored by James Harden



Creating A Logistic Model

We will now use logistic regression. To do that, we use the “glm” function.

```
harden_log = glm(Result~PTS, data = harden, family = "binomial")
```

Note the format of the glm function.

Make sure you ALWAYS have the family = "binomial" parameter in the glm function. This tells the function that we want to use logistic regression.

Predicting From Our Logistic Model

Now we will use the predict function to determine the likelihood the Rockets get a win or loss based off each of Harden’s performances.

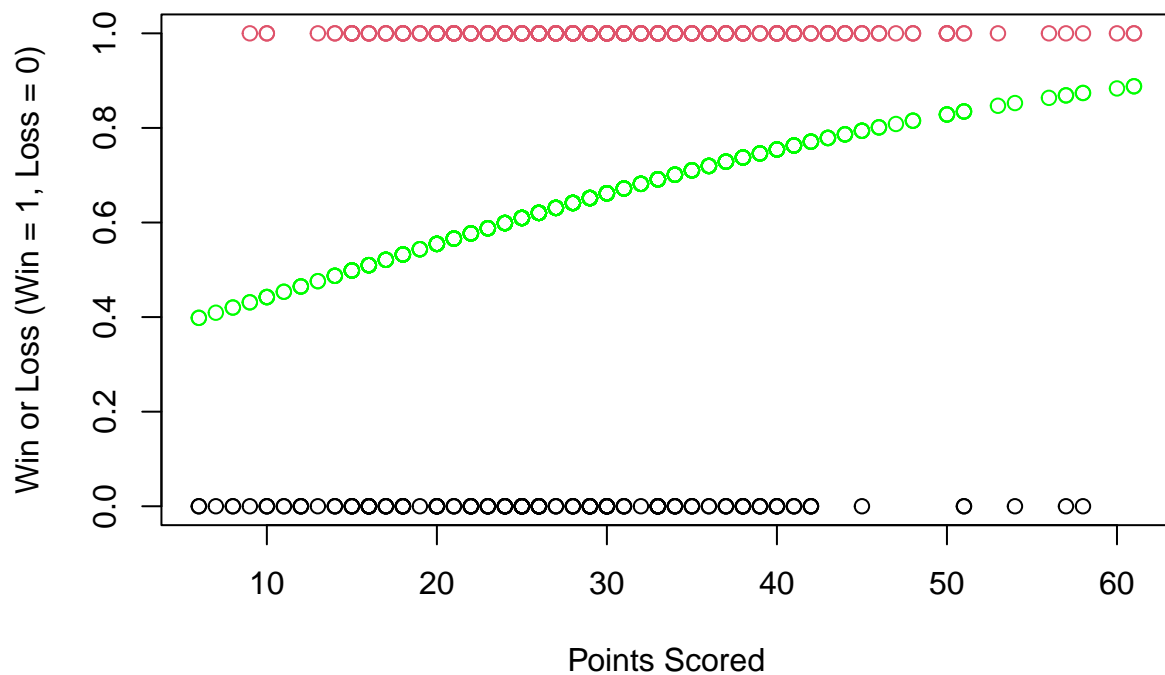
```
harden_probs = predict(harden_log, type = "response")
head(harden_probs)
```

```
##           1           2           3           4           5           6
## 0.7286239 0.7939519 0.5988325 0.4985460 0.5323689 0.5547724
```

We see that according to our model, the Rockets had a 72.9 percent chance to win game 1 (Harden scored 37), a 79.4 percent chance to win game 2 (Harden scored 45), a 59.9 percent chance to win game 3 (Harden scored 24), a 49.9 percent chance to win game 4 (Harden scored 15), etc... Let’s add this logistic curve (giving the probability of Rocket’s win based on Harden’s points) to our plot.

```
plot(harden$Dummy~harden$PTS, col = harden$Result,
     xlab = "Points Scored", ylab = "Win or Loss (Win = 1, Loss = 0)",
     main = "Rockets Wins based on Points Scored by James Harden")
points(harden$PTS,harden_probs,col="green")
```

Rockets Wins based on Points Scored by James Harden



Poll Questions:

Examine the following bits of R-code. What do you think they represent? Do the values seem to correlate with what you see in the plot above?

```
mean(harden_probs[harden$PTS < 20])
```

```
## [1] 0.4973586
```

```
mean(harden_probs[harden$PTS < 30])
```

```
## [1] 0.5805304
```

```
mean(harden_probs[harden$PTS >= 40])
```

```
## [1] 0.7925473
```

```
mean(harden_probs[harden$PTS >= 50])
```

```
## [1] 0.8534024
```

Comparing Predicted vs Actual Results

We can now compare our predictions from the model to the actual results. First we will create a vector with just 545 L's in it.

```
harden_log_pred = rep("L", 545)
```

If the our model gives a probability above 0.5, then it is predicting a win, so we turn those entries into a W in our vector.

```
harden_log_pred[harden_probs > .5] = "W"
```

It may be interesting to note how many wins our model predicts:

```
sum(harden_probs > .5)
```

```
## [1] 508
```

Out of 545 games, it predicts 508 wins. Wow. Note that

```
sum(harden$Dummy)
```

```
## [1] 352
```

Of 545 games, the Rockets actually won 352 of them.

Next we construct a table that gives us a breakdown of how accurate the model is when in predicts Wins and Losses.

```
harden_table = table(harden_log_pred, harden$Result)
harden_table
```

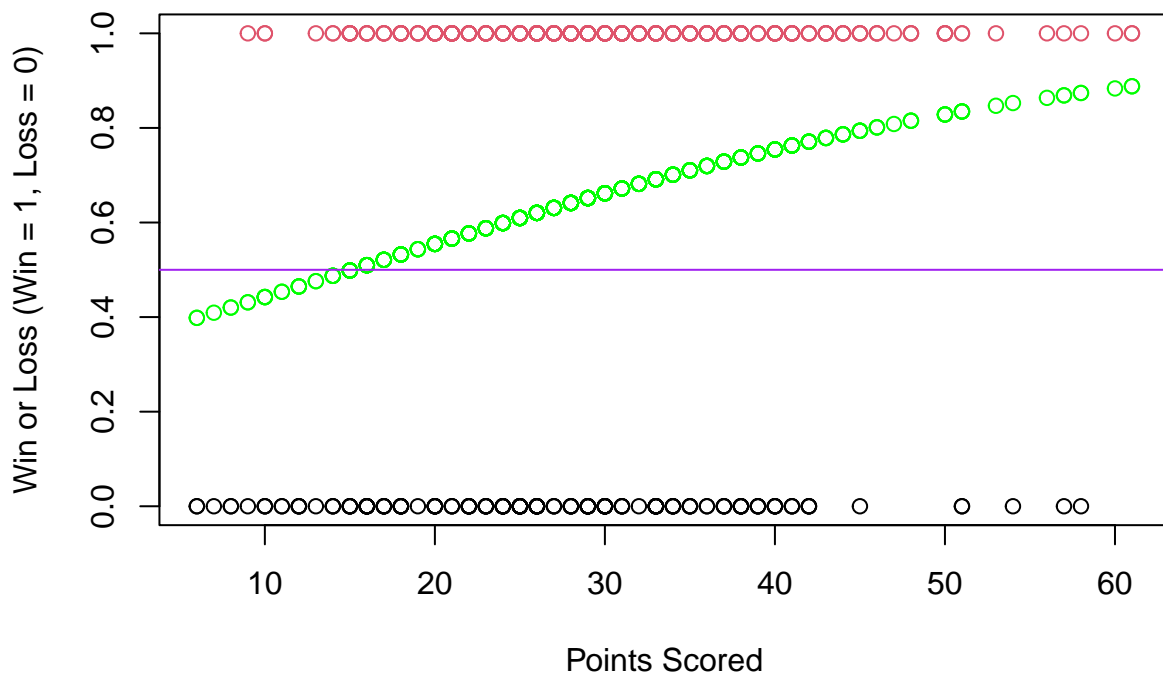
```
##
## harden_log_pred  L   W
##                L  24  13
##                W 169 339
```

Note how we read this table: the row L 24 13 says that when the model predicted a Loss, the Rockets lost 24 times and won 13 times, and the row W 169 339 says that when the model predicted a Win, the Rockets lost 169 times and won 339 times.

We can visualize this on our plot by adding a line at 0.5

```
plot(harden$Dummy~harden$PTS, col = harden$Result,
     xlab = "Points Scored", ylab = "Win or Loss (Win = 1, Loss = 0)",
     main = "Rockets Wins based on Points Scored by James Harden")
points(harden$PTS,harden_probs,col="green")
abline(a=0.5,b=0, col="purple")
```

Rockets Wins based on Points Scored by James Harden



Class Activity

1. Choose a sport you wish to investigate.
2. Determine variables you wish to perform a logistic regression and analysis on. Remember that the response variable must be binary.
3. Get the appropriate data by downloading it. (You may want to revisit your decisions for steps 1 and 2 if this proves difficult...)
4. Perform the logistic regression.
5. Use your resulting model to predict the results from your data as we did above. Create a table that compares predicted results to actual results. Create a plot that includes points for each actual result, points for each predicted result, and a line at the cutoff (presumed to be 0.5).

Turn in your resulting RMD file and your CSV data file into the appropriate Gradescope Class Activity assignment. It is due at the end of class.