Data Analytics 2 - Coding 1 term project: Analysing BoxRec.com

Oszkar Egervari

22 December 2021

Introduction

This is an analysis of active professional boxers' performance. The data for the analysis is scraped from BoxRec.com.

In the following pages I examine the connection between the current points held by each athlete and other variables listed by BoxRec.

My main goal for this project is to show whether these relationships exist (if they are statistically significant from zero) and how well they explain the the variations in the dependent variable, which is the current points of boxers.

Data

The data I used for the analysis is available on my github and the raw data can be found here.

Before deciding on the right hand side variables, let's take a glance at the descriptive statistics of the major variables of the dataset in table 1:

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Rank on BoxRec.com	4153.28	3691.00	2888.10	1.00	9997.00	274.70	9276.60
Points on BoxRec.com	4.11	0.36	20.69	0.00	691.00	0.01	14.97
Age	29.27	29.00	5.64	15.00	52.00	21.00	39.00
Number of professional bouts	18.64	15.00	15.07	1.00	248.00	3.00	47.00
Years spent as a professional boxer	6.70	6.00	4.87	0.00	28.00	0.00	16.00
Age of becoming a professional boxer	22.56	22.00	3.98	12.00	44.00	17.00	30.00
Height to average height per division ratio (%)	100.00	100.00	2.79	86.00	114.00	95.00	104.00

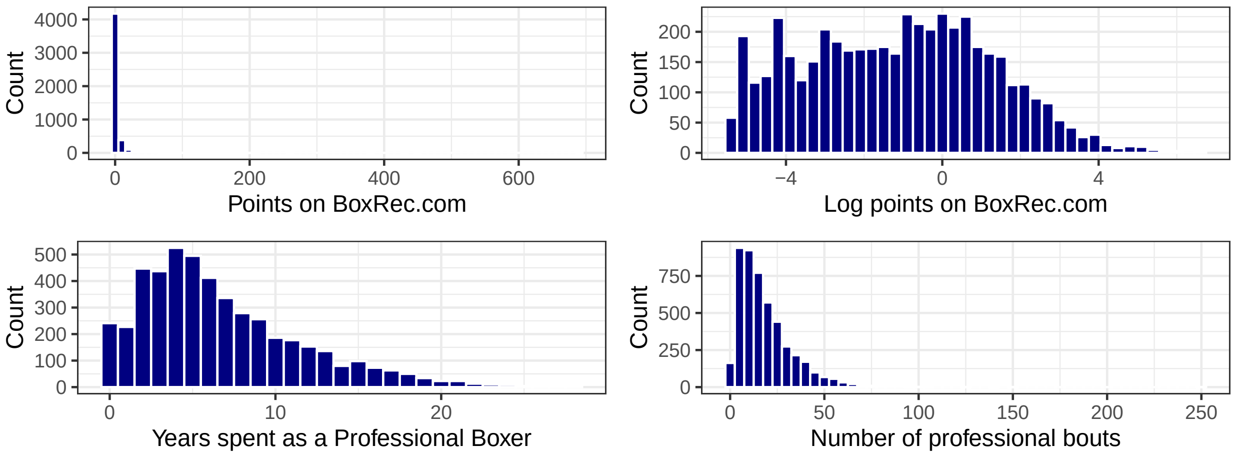
The number of observations is 4795 for all of our key variables.

We can see, that the ranks are fairly normally distributed, they are a little skewed to the right, which is understandable, because we can expect more athletes ranking the same as we are going down on the leaderboard.

However the points have a massive long right tail, the highest being 691, meanwhile the median is merely 0.356. Based on this, it's safe to say that it's probably better to use the log of the points, instead of their normal value, but let's take a look at some plots in the next section.

I calculated a new variable, the registered height and the mean height per division. Barring the heavy weight division, where there is no weight limit, so being taller seems almost always more advantageous, I'm not sure if height is an advantage or not. Because shorter athletes can put on more weight, given that everyone in a division has to fit in the same weight limit, meanwhile taller athletes have (in most cases) the benefit of extra reach. Thus I thought this is an interesting question to figure out.

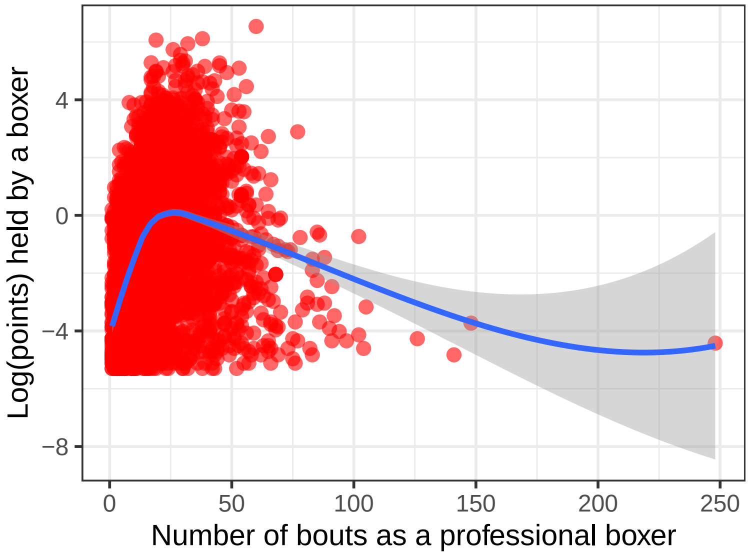
As the focus is the points achieved, the next Figure shows the histogram for this variable.



The plots confirm my theory, that the points have a log distribution, as we can see on the log points graph. The ‘years spent as a professional boxer’ and the number of professional bouts graphs look pretty similar, which is not surprising. At the end I decided to choose the number of bouts as the explanatory variable, because it’s coefficients were significant on all levels with one piecewise linear spline knot, meanwhile the years spent variable needed two knots in my opinion, and was not significant (on a 95% significance level) for all splines (details can be found in appendix).

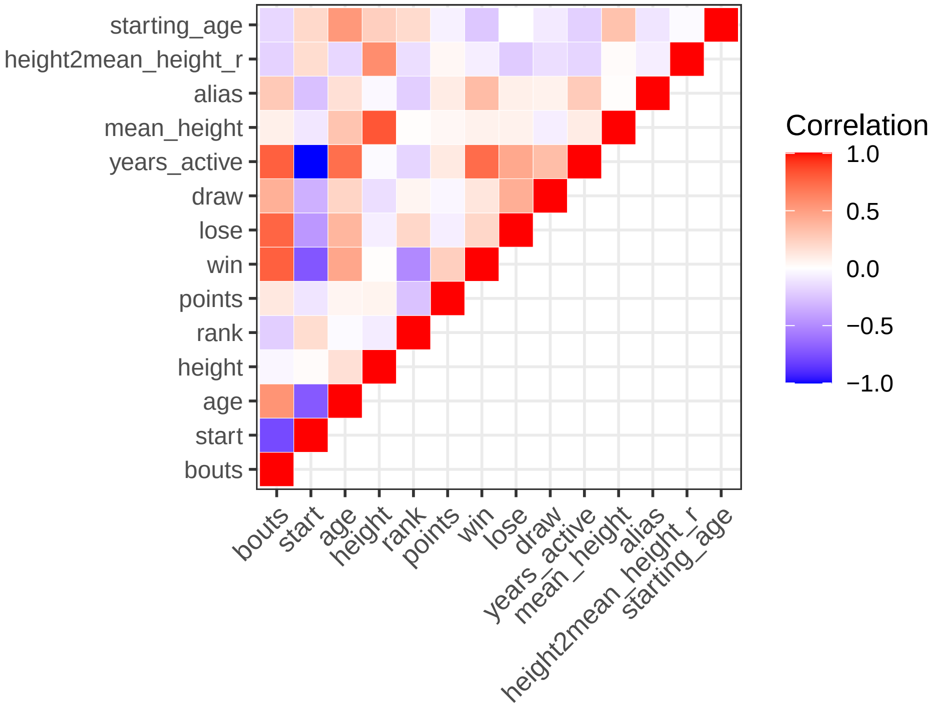
Also based on the graph, it would make sense trying to use the log of the number of professional bouts. I ended up not going this route, because the log-log regression of points-bouts provided an inferior model to the log-level regression with splines (for details please see appendix).

The key pattern of association is:



This is a scatterplot displaying the log(points) per number of bouts. The lowess non-parametric regression shows us where to have our knot, if we want to use piecewise linear splines.

Heatmap



The heatmap displays the correlation between all numeric variables in the dataset. For example we can see, that the 'start' and 'years active' variables are strongly correlated negatively, which means complete sense, since the earlier (the smaller) the start of a current professional career, the more the active years. We can also confirm, that the number of bouts and active years are positively correlated, the exact value is 0.78.

Based on the correlations of the numeric variables, I will use alias, lose, height2mean_height_r, age, starting_age and years_active as control variables. Alias is a binary variable and it takes a value of 1 when an athlete has an alias listed on BoxRec (2158 has an alias out of 4795). Lose is the number of losses, I have already talked about the height2mean_height_r, starting_age is the age of athletes turning professional, while years_active is the years of being a professional boxer.

Model

My preferred model is:

$\log(\text{points}) = -14.32 + 0.23\ (bouts < 25) + 0.14\ (bouts \geq 25) + \delta Z$

where Z is standing for the controls, which includes controlling for height to mean height per division, number of losses, current age, starting age, whether the boxer has an alias on BoxRec.com and active years. From this model we can infer:

- In case of log-level regression, the intercept is practically meaningless (in this case it means the average log points in case of a fighter having 0 bouts)
- when the number of bouts is one unit larger, but below the value of 25, we expect boxers to have 23 % more points on average
- when the number of bouts is one unit larger, with the value above or equal to 25, we expect boxers to have 14 % more points on average.

Based on the heteroskedastic robust standard errors, these results are statistically different from zero. To show that, I have run a two-sided hypothesis test:

$$H_0 := \beta_1 = 0$$

$$H_A := \beta_1 \neq 0$$

I have the t-statistic as 31.4 and the p-value is basically 0 for when the number of bouts is less than 25, and the t-statistic as 14.22 and the p-value is again basically 0 for when the number of bouts is more than 25, which confirms my conclusion.

We compare multiple models to learn about the stability of the parameters.

Table 2: Models to uncover relation between log points and the number of professional bouts

	(1)	(2)	(3)	(4)	(5)
Intercept	-1.798***	-3.266***	-3.325***	-14.35***	-14.32***
	(0.0727)	(0.0629)	(0.0639)	(1.146)	(1.080)
bouts	0.0347***				
	(0.0038)				
bouts (<25)		0.1523***	0.1431***	0.1811***	0.2280***
		(0.0047)	(0.0048)	(0.0084)	(0.0073)
bouts (>=25)		-0.0542***	-0.0546***	0.0812***	0.1387***
		(0.0053)	(0.0054)	(0.0079)	(0.0098)
alias_dummy			0.4417***	0.1040	0.0613
			(0.0662)	(0.0566)	(0.0530)
Losses				-0.1922***	-0.2199***
				(0.0130)	(0.0093)
height to division mean height ratio				0.1110***	0.0917***
				(0.0116)	(0.0100)
Age	No	No	No	No	Yes
Observations	4,795	4,795	4,795	4,795	4,795
R2	0.04785	0.20522	0.21272	0.42041	0.49011

The first table shows us, that both the intercept and the β are significant on a high level (significance code means p-value is between 0 and 0.001).

The second table, I added a spline (at 25 bouts). The coefficients are significant on the same level as with the first table.

On the third table I added the alias binary variable as a control. The coefficient of the alias variable, just as the other two are significant again on the same level. On a side note, it looks like a boxer having an alias, we can expect his points to be 44% higher on average. We'll see if that holds up with more variables in the mix.

On the fourth table I added another two control variables, the number of losses and the height to mean height per division. Both of them are significant with p values between 0 and 0.001. Interpreting the coefficients, with every other variable unchanged, we can expect a boxer with 1 additional loss to have on average 19% lower score. Finally we can see if there is indeed a relationship between the height to mean height per division relationship and points. The coefficient is 0.11, which means, that if the explanatory variable takes on a higher value by one unit, the expected value of points is 11% higher. In this case, the height ratio is a percentage (height/meanheight * 100), so if this ratio is 1% higher, we can expect the points to be 11% higher. Also the alias binary variable is not significant in this table, so most likely the two new variables explain the relationship of alias and log(points).

On the final table I added the age variable with two splines, the starting year and the years active with two splines. Even though these new variables seem like they are highly correlated, they are all significant with a p-value between 0 and 0.001 (the second spline of years_active was omitted because of collinearity). The summary table of this table can be found in the appendix.

Conclusion

Having done the analysis, we can conclude, that other than the alias binary variable, whose coefficient turned out to be not significant after involving other variables, every variable improved the previous models. We can see that by checking the R^2 value, which finally ended up being 0.49.

The analysis could be strengthened if the data was better. Unfortunately there are a lot of missing values for all types of variables throughout the dataset. I really wanted to use the reach (wingspan) variable, but there were just too many of it missing (only 1575 out of almost 10000 observations). I ended up not using the nationality as an explanatory variable, because on top of having 114 unique values, a lot of the country coefficients are not significant, which is expected, because there are 24 countries with only one observation.

Another thing, that would help the analysis are variables, that are more correlated to professional ranking. I'm not familiar with the advanced boxing statistics, but I'm sure, there are measures that would help, like area covered in a bout, hits received, hits given or even amateur results.

Appendix

Table 3: Comparing log(points) - number of bouts and log(points) - active years regressions

	Years spent in pro boxing	Number of bouts
(Intercept)	-3.165*** (0.0769)	-3.266*** (0.0629)
lspline(years_active,c(5))1	0.5196*** (0.0212)	
lspline(years_active,c(5))2	-0.0217* (0.0094)	
lspline(bouts,25)1		0.1523*** (0.0047)
lspline(bouts,25)2		-0.0542*** (0.0053)
Observations	4,795	4,795
R2	0.09951	0.20522

Table 4: Comparing log(points) - log number of bouts and log(points) - number of bouts regressions

	Log number of boutse	Number of bouts
(Intercept)	-3.978*** (0.0968)	-3.266*** (0.0629)
log(bouts)	1.082*** (0.0387)	
lspline(bouts,25)1		0.1523*** (0.0047)
lspline(bouts,25)2		-0.0542*** (0.0053)
Observations	4,795	4,795
R2	0.14667	0.20522

Table 5: Regression table 5

	(5)
(Intercept)	-14.32*** (1.080)
lspline(bouts,25)1	0.2280*** (0.0073)
lspline(bouts,25)2	0.1387*** (0.0098)
height2mean_height_r	0.0917*** (0.0100)
lose	-0.2199*** (0.0093)
lspline(age,c(27,31))1	-0.1093*** (0.0241)
lspline(age,c(27,31))2	-0.2235*** (0.0243)
lspline(age,c(27,31))3	-0.2600*** (0.0177)
starting_age	0.1764*** (0.0166)
alias	0.0613 (0.0530)
lspline(years_active,5)1	0.2061*** (0.0338)
Observations	4,795
R2	0.49011