MKT 500T Customer Analytics Using Probability Models

Final Project

Meng Guo, Student ID: 465982, Email: mengguo@wustl.edu

1. Background

It's believed that sports teams have better performance at their home, especially for soccer teams. That might be the result of familiar surroundings, the backing of a vociferous home crowd and the lack of traveling involved. How many goals they can get at home becoming the critical component determines the result of the game and have a great impact on the end ranking of the league for the season.

Spanish La Liga Football League is the men's top professional football division of the Spanish football league system. It's one of the most popular professional sports leagues in the world, with an average attendance of 26,983 for league matches in the 2017–18 season. The project, using *Spanish La Liga Football League* as a studying example, tried to find the pattern of the number of home-team goals for the season of 2017-2018, hoping to generate valuable insights for better understanding of the *La Liga Football League* and soccer teams' performance.

2. About the Data

The dataset is from https://datahub.io/sports-data/spanish-la-liga. It contains data for last 10 seasons of *Spanish La Liga Football League*. In this project, data of 2017-2018 season was used for analysis. The project counted the number of "FTHG" which stands for "Full-Time Home Goal" for each game, and generated the aggregated table as the following:

Table 1. Model Data

Home Goal	Count
0	95
1	114
2	94
3	45
4	16
5	10
6	5
7	1
Total	380

The first column recorded the number of full-time goals scored by the home team and the second column counted the number of home teams that scored the specific number of full-time goals throughout the whole season. As a total of 380 games played in the 2017-2018 season, the total of home games was also 380. A clear pattern can be seen in this table: most of the home teams got 0 to 2 goals per game while the number of goals vary from 0 to 7. The mean of goals was 1.5473, the frequency of 0 goals was 0.25 and the variance was 1.9.

3. Model Fitting

On the individual level, the project tried to measure how many goals a home team got per game, which fell into the category of counting problems. Poisson distribution was the first distribution used to fit, which expressed the probability of a given number of events occurring in a fixed interval of time. The first estimation approach tried was Maximum Likelihood Estimation. Then, the project considered bringing heterogeneity to the model because of the possibility that every team had its own chance of scoring. So, Negative Binomial Distribution based MLE model was the second model to fit. Another step the project took was assuming that there was a spike at zero that some teams just wouldn't score when they were at home, and then the project brought this assumption into model fitting. Next, instead of assuming every team's chance of scoring was totally different, the project also considered the possibility that infinite number of segments existed in the *La Liga Football League*. Since the actual number of segments was unknown, the project tried testing the number range from 2 to 4 to fit the data in order to find the more reasonable segregation. Other estimation approaches including Means & Zeros and Methods of Movement based on NBD were also implemented.

4. Results and Comparisons

Table 2. Summary of Model Fit

Model	*LL*	# Parameters	BIC	Chi-square p-value
NBD	-613.21985	2	1238.32004	0.583480137
ZNBD	-613.19332	3	1244.20715	0.431853616
Poisson	-617.23546	1	1240.41109	0.012052302
2 Seg Poisson	-613.12456	3	1244.06963	0.484273278
3 Seg Poisson	-612.94706	5	1255.59497	0.222018056
4 Seg Poisson	-612.94704	7	1267.47529	Not Valid
Means & Zeros	-613.234	2	1238.34835	0.597506399
MoM	-613.22031	2	1238.32096	0.586623552

- **General Results Discussion**

Firstly, let's take a look at the results of Log Likelihood. Generally speaking, when we include more parameters into the model, the log likelihood will always increase. However, it is the key to find balance between the flexibility and parsimony. It's not always worthwhile to include more parameters into the model for better fitness. For example, we can use the Likelihood Ratio Test to check whether an actual improvement will be brought when adding a spike at zero to the existing NBD model. Note that the frequency of spike at zero was 0.01452.

Table 3. Likelihood Ratio Test

χ^2	0.05305996
df	1
p-value	0.81782178

Since the p-value was very large, we cannot reject the null hypothesis that NBD and NBD with a spike at zero didn't have significant difference. So, the one more parameter added to the model was not necessary.

- **Model Comparisons**

From Table 2 above, we can make different model comparisons and evaluations based on different model measurement criteria:

(1) In-sample Fit

χ^2 p-value measures how the model fits on in-sample data. From the Summary of Model Fit table, we can infer that the in-sample fitness of Means & Zeros based on Negative Binomial Distribution generated the highest χ^2 p-value. In fact, the p-values from the models were all very large expect the 1-Segment Poisson Model and the 4-Segment Poisson Model, meaning using these models' results we can't reject the null hypothesis that the difference of the true number of goals and the estimated number of goals was significant from zero. In other words, these models performed quite well on in-sample data. At least we can say that they can give us reasonable estimates on in-sample data. MoM and NBD (using MLE) generated the second and third highest χ^2 p-value. However, the p-value of 1-Segment Poisson Model was less than 0.05 so that we can reject the null hypothesis on the 95% confidence level, suggesting the 1-Segment Poisson model performed poorly on predicting the number of goals. For 4-Segment Poisson Model, the χ^2 p-value was not valid because the degree of freedom was not positive anymore since we added too much parameters in the model using limited data.

From this comparison we can infer that Negative Binomial Distribution based models performed better than Poisson Distribution based models on in-sample fit. We can also take a look the parameters estimated by Means & Zeros, MoM and NBD (using MLE):

Table 4. Parameters of NBD

	Means & Zeros	MoM	MLE
r	6.419280753	6.78753933	6.87298595
alpha	4.148514772	4.38650501	4.44169371

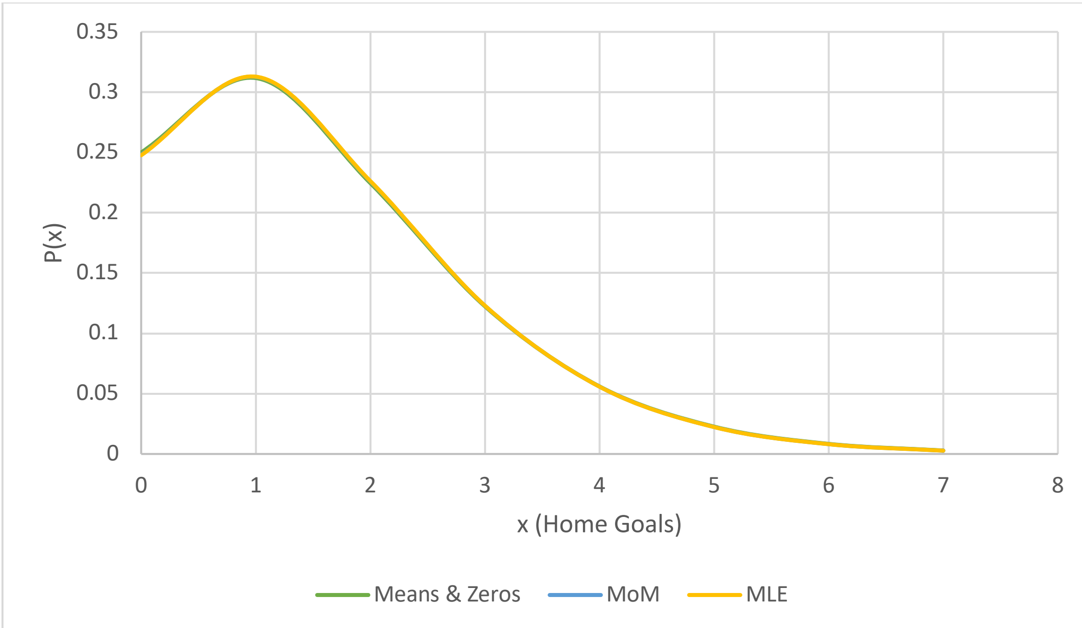


Figure 1. NBD Models Fit

From the table and figure above, we can see that different estimating methods generated similar parameter results, suggesting mean which was calculated by r/alpha ≈ 1.5. With r around 6 and alpha around 4 we can infer that the heterogeneity was not very disperse – there was a spike where most of the teams get an average of around 1 goal per game.

(2) Out of Sample Fit

As Bayesian Information Criterion (BIC) can be used as a proxy of out of sample measurement of fitness, we can make the following comparison:

From Table 2 we can conclude that NBD using MLE generated the lowest BIC value of 1238.32004, followed by MoM and Means & Zeros. Surprisingly, 1-Segment Poisson Model actually generated fair out of sample fit. The result of in-sample fit and out of sample fit was similar - NBD based models performed relatively better than Poisson Distribution based models.

(3) Parsimony

Just like what we've discussed in the Log Likelihood section, the more parameters are not always what we want in a model. 1-Segment Poisson only needed one parameter which was the fewest. NBD based models including those using MLE, Means & Zeros and MoM needed two parameters while NBD with a spike at zero needed three parameters. More than three parameters were required for greater than two segments' Poisson models. The number of parameters can be calculated by the formula of 2 * number of segments - 1.

(4) Story

The story behind our models is what we value greatly. Let's go through what the models were telling us:

- ***NBD Based Models***

NBD based models told us the story that each team in the *Spanish La Liga Football League* was different. They had different propensities of scoring in each home game. Given the r was around 6 - a greater than 1 value - we can infer that most of teams had a great propensity at getting around 1 goal per game. This is relatively harder to interpret and quite different from our common sense. In reality, we commonly assume that there are some teams that are more competitive in the league, and they will have greater propensity of scoring more goals than regular teams do. While there are also some teams that perform relatively worse than regular teams with the propensity of scoring fewer goals per game. According to this logic, we can check the result from finite segment models.

- ***Finite Segment Models***

Table 5. Parameters of Finite Segments

	Seg 1	Seg 2	Seg 3	Seg 4	LL
mean	1.54736842				-617.23546
mean	1.21027525	2.600217569			-613.12456
class size	0.75747771	0.242522291			
mean	2.86893859	1.3646139	0.00005		-612.94706
class size	0.15148561	0.815442164	0.03307222		
mean	1.3646756	2.86909881	1E-05	2.869118396	-612.94704
class size	0.81547035	0.124136936	0.03308353	0.027309183	

From the finite segment models, we can tell a totally different story. The project had tried the number of segments from 2 to 4, with the results listed in the above Table 5. So, how many segments were there? From Table 5 we can argue that 4 segments were relatively less reliable in this case because there were two groups of teams that had the same mean goals of 2.869 per home game. 2 Segments and 3 Segments were both acceptable while 2 Segments Model had better in-sample and out of sample fit.

The story behind 2 Segment Model was that the *La Liga League* had two different level of teams: There was a group of more competitive teams that generally would score 2 times than the other group of teams on average. We can relate this to reality that there were indeed some teams in the league that keep absolute extraordinary record of scoring and winning like Barcelona and Real Madrid.

The story behind 3 Segment Model also made sense. We can say there could be an additional group of teams that didn't perform as well as other teams. Those were the ones would not compete in the same level of league the next season.

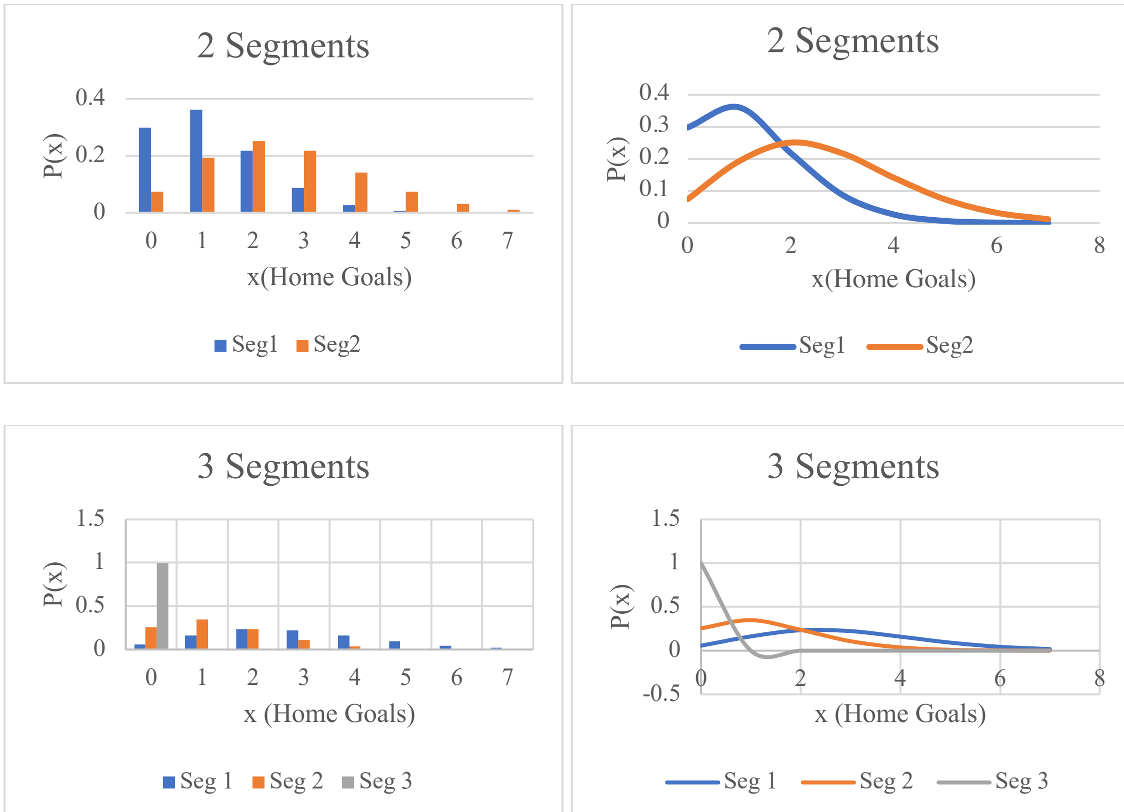


Figure 2. Models with 2 and 3 Segments

5. Conclusion

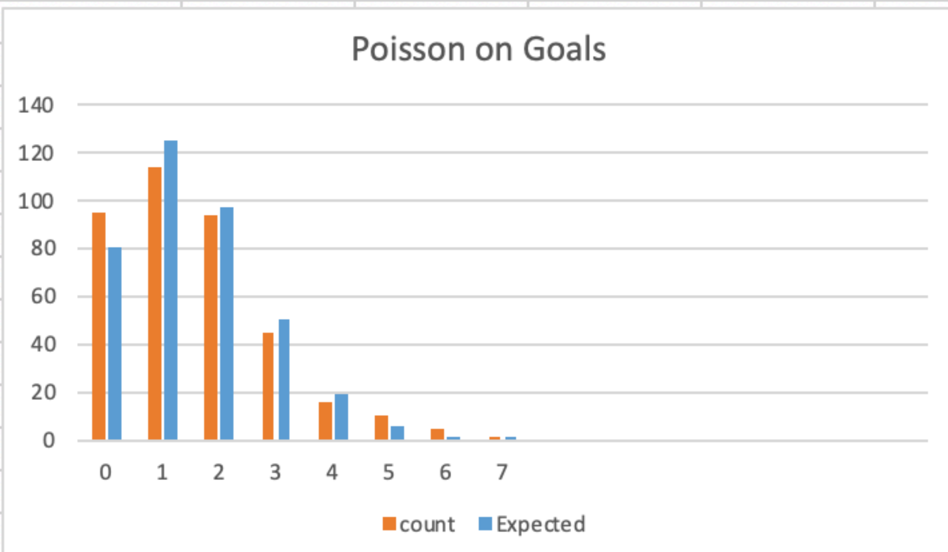
Analysis above provided us clearer understanding of *Spanish La Liga Football League*. The generally better fitness performance of NBD based models told us that it was highly possible that heterogeneity existed among the chance of scoring in the home game across teams. Since the NBD with spike at zero didn't pass the Likelihood Ration Test, we can confirm that the possibility of the existence of never-scoring team was very low. It was also reasonable in reality: even though the performance of different team can vary a lot, each team should always have a chance to shoot goals in a home game.

Based on personal opinion, finite segment models seemed to be more plausible because they told a better story than NBD based models, in spite of better model fitness of the latter set of models. From the model fitting result, we can conclude that there were at least two and probably less than four segments in the *La Liga League*. If only one model can be chosen, the 2 Segment Model would be the winner model since the fitness is good enough and the story behind was is also solid. In the 3 Segment Model, the size of the “third segment” was 0.033, which was much smaller than the other two groups. Even though based on common knowledge, we would naturally assume a three-group division in a sports league, however, this particular result that 2 Segment Model fitted a lot better did make more sense if we consider real-life situation. *La Liga League* is one of the best soccer leagues around the whole world. The teams that can compete in this league should be generally more competitive on average. Since the number of goals of a team can somehow reflect its ability to score, weaker teams with lower ability to score, to be more specific, with fewer goals, will leave *La Liga League* at the end of every season. After years of competition, the remaining teams should be even stronger on average. The number of teams that belong to the “third segment” was so low that a 2 Segment Model fits better on this data.

To conclude, the project found that there were two group of teams in *Spanish La Liga League*: one with higher average home-game goal of 2.6, and the other with relatively lower average home-game goal of 1.2.

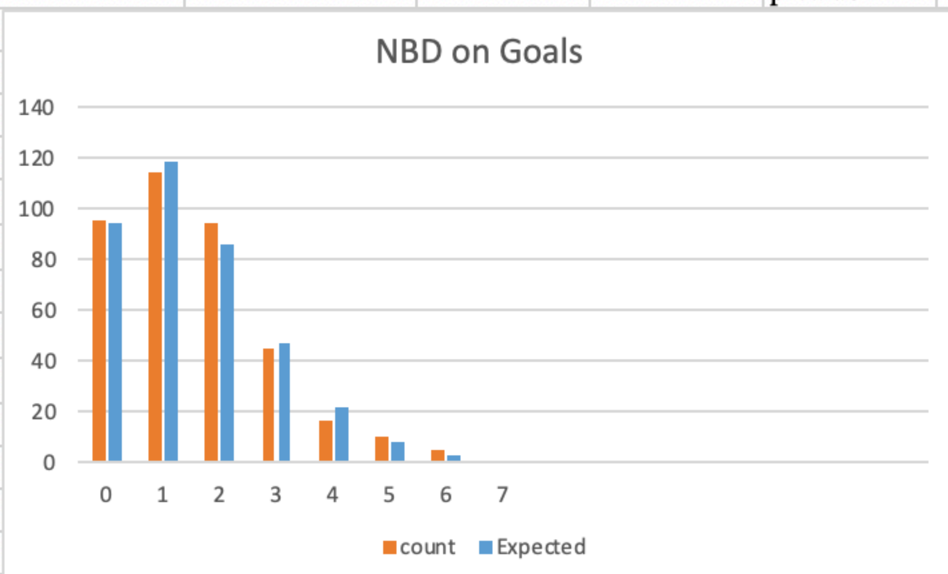
Appendix 1. Poisson

lambda	1.54736842		LLSum	-617.23546	
BIC	1240.41109				
HomeGoal	count	P(X=x)	LL	Expected	Chiq
0	95	0.21280726	-147	80.8667575	2.47009461
1	114	0.32929123	-126.63265	125.130667	0.99009899
2	94	0.25476742	-128.536	96.8116213	0.08165563
3	45	0.13140636	-91.325736	49.9344152	0.48760866
4	16	0.05083351	-47.667192	19.3167343	0.56949204
5	10	0.01573163	-41.520817	5.97802093	2.70596504
6	5	0.00405711	-27.536427	1.54170013	7.75756432
7	1	0.00089683	-7.0166399	0.34079687	1.27509611
total	380				16.3375754
				df	6
				p value	0.0120523



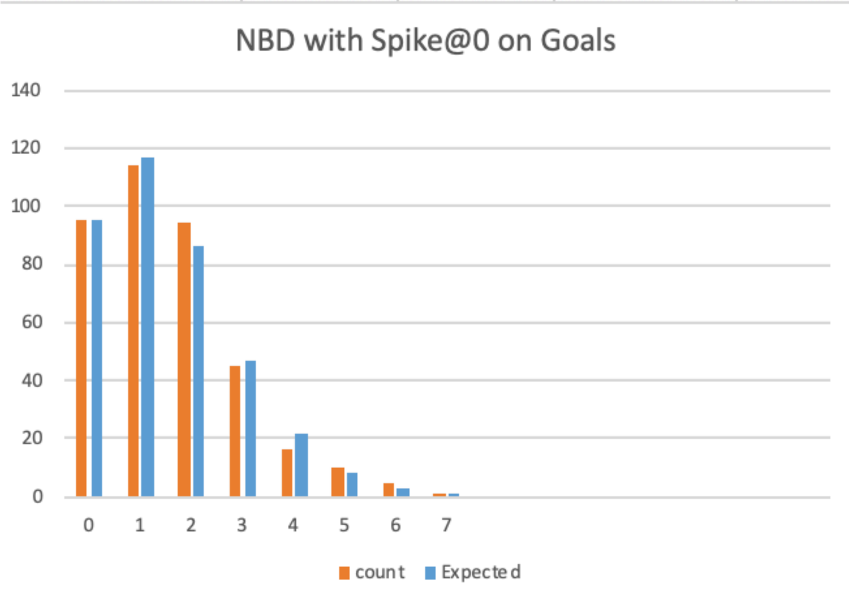
Appendix 2. NBD

r	6.872985948		LLSum	-613.21985	
alpha	4.441693713				
BIC	1238.320042				
HomeGoal	count	P(X=x)	LL	Expected	Chiq
0	95	0.24768653	-132.58118	94.1208821	0.00821123
1	114	0.31283386	-132.47746	118.876867	0.20007115
2	94	0.22630239	-139.67302	85.9949083	0.74517776
3	45	0.12299957	-94.30085	46.7398359	0.06476336
4	16	0.05579021	-46.178509	21.200281	1.27559263
5	10	0.02229476	-38.034037	8.47200777	0.27558524
6	5	0.00810732	-24.074939	3.08078191	1.19560495
7	1	0.00273984	-5.8998573	1.04113756	0.00162543
total	380				3.76663175
				df	5
				p value	0.58348014



Appendix 3. NBD with spike at zero

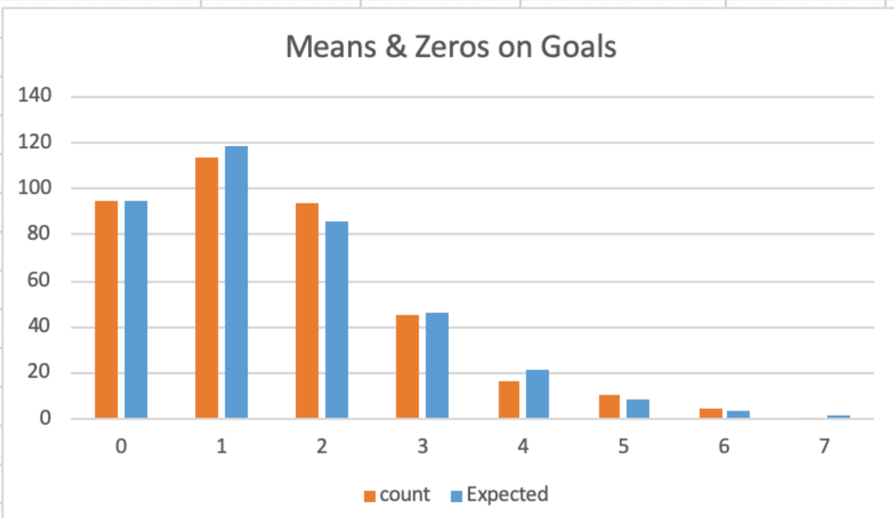
r	7.8590379			LLSum	-613.1933	
alpha	5.0052134					
spike at 0	0.0145206					
BIC	1244.2072					
HomeGoal	count	P(X=x)	P(with spike)	LL	Expected	Chiq
0	95	0.2389484	0.24999936	-131.6982	94.999756	6.26725E-10
1	114	0.3127124	0.30817162	-134.1892	117.10521	0.082339257
2	94	0.2306605	0.22731116	-139.2549	86.37824	0.672521571
3	45	0.1262287	0.12439576	-93.79292	47.27039	0.109046478
4	16	0.0570638	0.05623523	-46.05139	21.369388	1.34914164
5	10	0.0225378	0.02221056	-38.07187	8.4400126	0.288336127
6	5	0.0080434	0.00792662	-24.18764	3.0121173	1.311926938
7	1	0.0026518	0.00261333	-5.947129	0.9930668	4.84047E-05
total	380					3.813360417
					df	4
					pvalue	0.431853616



Appendix 4. Means & Zeros

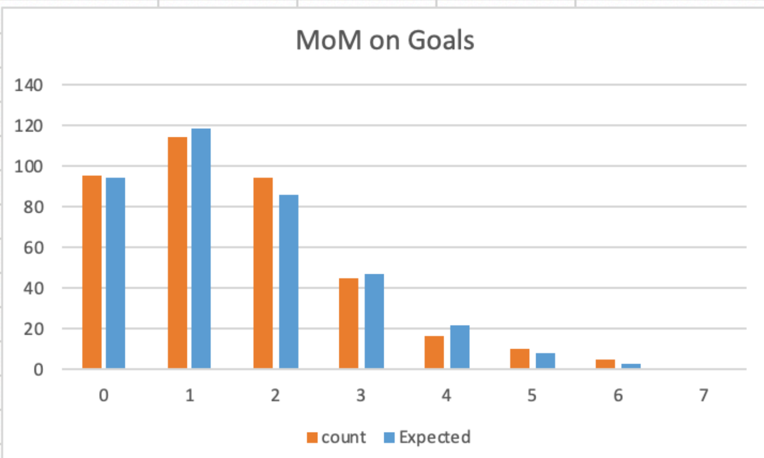
r	6.4192808		Mean	1.5473684		
alpha	4.1485148		Zero	0.25		
BIC	1238.3483		Estimated Zero	0.25	differencesq	3.4199E-24
LL	-613.234					

HomeGoal	count	P(X=x)	Expected	Chiq	LL
0	95	0.25	95	5.198E-21	-131.698
1	114	0.3117055	118.4480765	0.1670385	-132.8894
2	94	0.224592	85.34495606	0.8777295	-140.3862
3	45	0.1224239	46.52106337	0.049733	-94.51197
4	16	0.055994	21.27773621	1.3090913	-46.12016
5	10	0.0226635	8.612142226	0.2236551	-37.86998
6	5	0.0083779	3.183587058	1.0363643	-23.91081
7	1	0.002887	1.097067052	0.0085884	-5.847531
total	380			3.6722	
			df	5	
			p value	0.5975064	



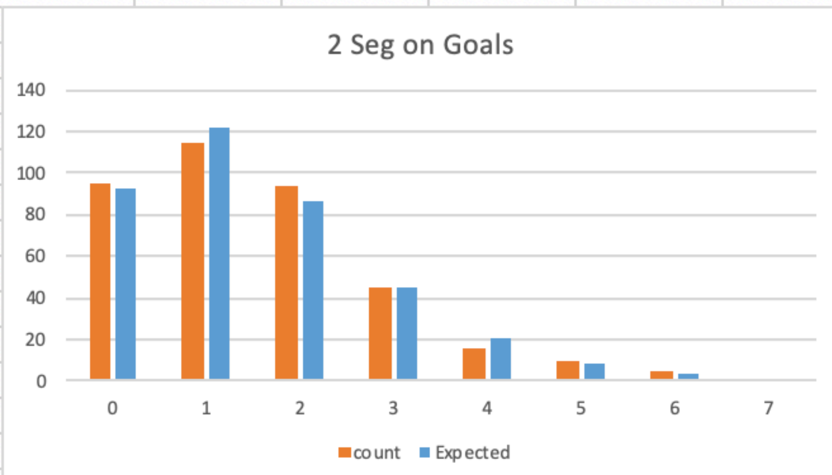
Appendix 5. MoM

r	6.78753933			Mean	1.54736842		
alpha	4.38650501			Variance	1.90012498	0.01492543	
BIC	1238.32096			Estimated Variance	1.90012498	differencesq	3.4705E-19
LL	-613.22031						
HomeGoal	count	variance	P(X=x)	Expected	Chiq	LL	
0	95	227.463158	0.24810183	94.27869515	0.00551854	-132.42202	
1	114	34.1557895	0.31263332	118.8006602	0.19399167	-132.55057	
2	94	19.2582825	0.2259948	85.87802408	0.76814172	-139.80087	
3	45	94.9562327	0.1228959	46.70044323	0.06191605	-94.338792	
4	16	96.2464266	0.05582695	21.21423929	1.28160577	-46.167978	
5	10	119.206648	0.0223609	8.497140181	0.26580562	-38.004416	
6	5	99.1296399	0.00815556	3.099114482	1.16593491	-24.045274	
7	1	29.7311911	0.0027659	1.051040604	0.00247863	-5.8903905	
total	380				3.74539291		
				df	5		
				p value	0.58662355		



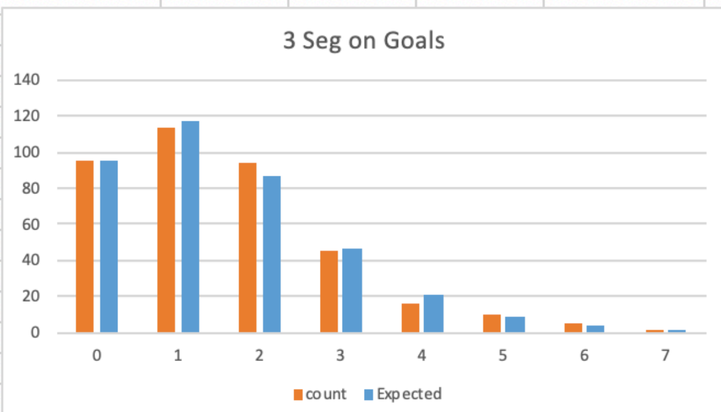
Appendix 6. 2 Segments

lambda_1	1.210275						
lambda_2	2.600218						
theta_1	1.1389	3.123332					
theta_2	0	1					
LL	-613.1246						
BIC	1244.07						
		0.757478	0.242522				
HomeGoal	count	Seg1	Seg2	P(X=x)	LL	Expected	Chiq
0	95	0.298115	0.074257	0.243825	-134.074	92.65339	0.05943207
1	114	0.360801	0.193085	0.320127	-129.8504	121.6481	0.48084188
2	94	0.218335	0.251032	0.226264	-139.6888	85.98048	0.74799165
3	45	0.088082	0.217579	0.119488	-95.60438	45.40533	0.00361837
4	16	0.026651	0.141438	0.054489	-46.55601	20.70594	1.06954254
5	10	0.006451	0.073554	0.022725	-37.84291	8.635487	0.21560975
6	5	0.001301	0.031876	0.008716	-23.71279	3.312205	0.86004712
7	1	0.000225	0.011841	0.003042	-5.795224	1.155979	0.02104652
total	380						3.45812989
						df	4
						p value	0.48427328



Appendix 7. 3 Segments

lambda_1	2.868939							
lambda_2	1.364614							
lambda_3	0.00005							
theta_1	1.521797	4.580448						
theta_2	3.205037	24.6564						
theta_3	0	1						
LL	-612.9471							
BIC	1255.595							
		0.151486	0.815442	0.033072225				
HomeGoal	count	Seg1	Seg2	Seg3	P(X=x)	LL	Expected	Chiq
0	95	0.056759	0.255479	0.999950001	0.249997	-131.699	94.99899	1.0632E-08
1	114	0.162838	0.348631	4.99975E-05	0.308957	-133.8989	117.4038	0.0986852
2	94	0.233587	0.237873	1.24994E-09	0.229357	-138.4128	87.15558	0.53750009
3	45	0.223382	0.108202	2.08323E-14	0.122071	-94.64173	46.38711	0.04147875
4	16	0.160217	0.036913	2.60404E-19	0.054371	-46.59069	20.66111	1.05153768
5	10	0.091931	0.010074	2.60404E-24	0.022141	-38.10308	8.413716	0.29907098
6	5	0.043957	0.002291	2.17003E-29	0.008527	-23.8224	3.240381	0.95552325
7	1	0.018016	0.000447	1.55002E-34	0.003093	-5.778492	1.175483	0.02619719
total	380							3.00999314
							df	2
							p value	0.22201806



Appendix 8. 4 Segments

lambda_1	1.3646756								
lambda_2	2.8690988								
lambda_3	1E-05								
lamba_4	2.8691184								
theta_1	3.3965421	29.860665							
theta_2	1.5141623	4.5456116							
theta_3	0.1918127	1.2114436							
theta_4	0	1							
LL	-612.947								
BIC	1267.4753								
		0.8154703	0.1241369	0.033083534	0.0273092				
HomeGoal	count	Seg1	Seg2	Seg3	Seg4	P(X=x)	LL	Expected	Chiq
0	95	0.2554635	0.05675	0.99999	0.0567489	0.2500007	-131.6977	95.00026023	7.128E-10
1	114	0.3486249	0.1628215	9.9999E-06	0.1628194	0.3089522	-133.9008	117.4018308	0.0985713
2	94	0.2378799	0.2335755	4.99995E-11	0.2335741	0.2293581	-138.4123	87.15606948	0.5374197
3	45	0.1082096	0.2233837	1.66665E-16	0.2233839	0.1220724	-94.64136	46.38749365	0.0415012
4	16	0.0369178	0.1602275	4.16663E-22	0.1602287	0.0543712	-46.59073	20.66105785	1.0515173
5	10	0.0100762	0.0919417	8.33325E-28	0.091943	0.0221411	-38.10322	8.413600575	0.2991184
6	5	0.0022918	0.043965	1.38888E-33	0.0439659	0.0085272	-23.82245	3.240346717	0.9555705
7	1	0.0004468	0.01802	1.98411E-39	0.0180205	0.0030934	-5.77848	1.175497378	0.0262011
total	380								3.0098996
								df	0
								p value