Amanda Kuznecov
anr431

Homework 6

DS-GA 3001.015
March 31, 2021

Question 2: Pitching Statistics

Part a: Probabilities described below were computed on pitcher data from 2012 to 2017.

Part a.i: Probability of a plate appearance ending with a walk: 0.0749

Part a.ii:. Probability of a plate appearance ending with a strikeout: 0.2064

Part a.iii: Conditional probability of a plate appearance ending with a strikeout given a walk did not occur: 0.2231

Part a.iv: Average number of home runs per plate appearance: 0.0276

Part a.v: Conditional probability of a plate appearance ending with a home run given neither a walk nor a strikeout occurred: 0.0384

Part a.vi: Conditional probability of a plate appearance ending with a non-HR hit given that the plate appearance didn't end with a walk, strikeout, or homerun: 0.2915

Part b: For each season and pitcher, *k-rate*, the average number of strikeouts per plate appearance that did not end in a walk (the second McCracken component) was computed. Pitchers with less than 500 batters faced were not included in this analysis. The top 10 pitchers in 2016 according to this statistic is provided below.

Pitcher	k-rate	Batters Faced
Fernandez_Jose_605228	0.370968	731
Scherzer_Max_453286	0.335697	900
Strasburg_Stephen_544931	0.330325	597
Kershaw_Clayton_477132	0.322702	543
Syndergaard_Noah_592789	0.310984	742
Ray_Robbie_592662	0.309220	772
Salazar_Danny_517593	0.309021	581
Velasquez_Vincent_592826	0.300395	550
Verlander_Justin_434378	0.300236	902
Archer_Chris_502042	0.297573	850

The top 10 pitchers in 2017 according to this statistic is provided below.

Pitcher	k-rate	Batters Faced
Sale_Chris_519242	0.381188	851
Scherzer_Max_453286	0.369655	778
Ray_Robbie_592662	0.367003	662
Kluber_Corey_446372	0.357625	775
Hill_Rich_448179	0.330020	551
Peacock_Brad_502748	0.329243	546
Archer_Chris_502042	0.314394	852
Severino_Luis_622663	0.314208	783
Strasburg_Stephen_544931	0.311927	696
Kershaw_Clayton_477132	0.311248	679

Part c: For each season and pitcher, *h-rate*, he average number of non- HR hits that did not end in a walk, strikeout, or home run (the fourth McCracken component) was computed. Pitchers with less than 500 batters faced were not included in this analysis. The top 10 pitchers in 2016 according to this statistic is provided below.

Pitcher	h-rate	Batters Faced
Ray_Robbie_592662	0.347732	772
Paxton_James_572020	0.346260	508
Pelfrey_Mike_460059	0.341981	541
Cole_Gerrit_543037	0.339726	503
Perdomo_Luis_606131	0.336066	655
McHugh_Collin_543521	0.335185	795
Bradley_Archie_605151	0.334951	630
Duffey_Tyler_608648	0.334118	593
Wacha_Michael_608379	0.333333	600
Syndergaard_Noah_592789	0.332627	742

The top 10 pitchers in 2017 according to this statistic is provided below.

Pitcher	h-rate	Batters Faced
Montero_Rafael_606160	0.361345	545
Taillon_Jameson_592791	0.348148	584
Richard_Clayton_453385	0.346154	852
Nelson_Jimmy_519076	0.333333	727
Bauer_Trevor_545333	0.333333	749
Gausman_Kevin_592332	0.331471	816
Miley_Wade_489119	0.329060	727
Pivetta_Nick_601713	0.328729	584
Colon_Bartolo_112526	0.328629	648
Boyd_Matt_571510	0.327830	602

Question 3: Predicting McCracken Components

Using data from the preceding season, we forecast each McCracken component. We restrict this analysis to rows with at least 200 batters faced.

Part a: bb-rate

(i) Intercept: 0.0335, Coefficient: 0.5474

(ii.A) New features were created for cumulative average of lineouts rate, popouts rate, flyouts rate, and groundouts rate for each pitcher over all previous seasons they had pitched. The general methodology for selecting features to improve the model was to perform a backward selection; all features were added to the model and then unimportant features were incrementally removed. Features were removed if their confidence interval included 0, which meant that including the feature had little to no effect. Features with the smallest confidence interval including 0 were the first to be removed, as well as features whose confidence intervals centered on 0. Additionally, the R^2 value was observed for any changes while removing features.

With the basic model only using bbrate_prev, the R^2 value was 0.286. With the additional features, R^2 improved to 0.307. For predicting bb-rate, the following features provided the best model performance:

- bbrate_prev (coeff = 0.4817); most significant
- fo_rate_prev (coeff = -0.0634); not very significant
- go_rate_prev (coeff = -0.0847); not very significant
- po_rate_prev (coeff = -0.1549); significant

(ii.B) It is evident that bbrate_prev is the most important as it has a very significant coefficient. This makes sense because walk rate should be highly correlated with the previous season's walk rate, as we explored earlier. As well, po_rate_prev has a significant coefficient but in the opposite direction, informing us that popout rate is negatively correlated with walk rate. This makes sense because an out made on a popout does not lead to a walk. The same logic applies to the other two features which also have coefficients in the opposite direction and are very similar in magnitude with each other. There is a negative correlation between walk rate and flyout rate, as well as between walk rate and groundout rate, however not as significant compared to the negative correlation between walk rate and popout rate.

Part b: k-rate

(i) Intercept: 0.0583 Coefficient: 0.7474

(ii.A) With the basic model only using krate_prev, the R^2 value was 0.558. With the additional features, R^2 improved to 0.572. For predicting k-rate, the following features provided the best model performance:

- krate_prev (coeff = 0.6202); most significant
- fo_rate_prev (coeff = -0.3296); significant
- go_rate_prev (coeff = -0.2449); signifcant

(ii.B) It is evident that krate_prev is the most important as it has a highly significant coefficient. This makes sense because strikeout rate should be highly correlated with previous season's strikeout rate as we discovered from our previous analysis. The features fo_rate_prev and go_rate_prev are both negatively correlated with strikeout rate, however they are both still quite significant based on the magnitude of their coefficients. This informs us that fewer flyouts and groundouts lead to strikeouts. This may be due to the fact that when a hitter already has two strikes against them, they are more likely to take a less risky hit (rather than hit the ball very high where the chances of an outfielder catching it is also very high, or hitting it on the ground where it may easily be picked up.) It is likely that at this stage, batters attempt more conservative hits and this could be why flyouts and groundouts are negatively correlated with strikeouts.

Part c: hr-rate

(i) Intercept: 0.0273, Coefficient: 0.2954

(ii.A) With the basic model only using hrrate_prev, the R^2 value was 0.069. With the additional features, R^2 improved to 0.160. For predicting hr-rate, the following features provided the best model performance:

- hrrate_prev (coeff = 0.1716); significant
- lo_rate_prev (coeff = 0.2136); most significant
- go_rate_prev (coeff = -0.0367); not very significant
- po_rate_prev (coeff = 0.0693); not very significant
- fo_rate_prev (coeff = 0.0537); not ver significant

(ii.B) It is surprising in this model that hrrate_prev is not the most significant feature based on the magnitude of our coefficients, as we would expect to predict the new season's home run rate based on the previous season's home run rate. However, it appears that lineout rate is a better indicator of home run rate. One reason for this could be that batters who hit in such a way where the ball is hit hard and low make for difficult catches. These types of batters likely make similar hits from season to season, and these could result in home runs if not caught by the defense. One could also say that there is some luck involved in home runs especially when the ball is not hit into the stands and was just not caught by the defense, and for this reason, lineout rate could aid in our prediction of home run rate for the next season. The other three features, groundout rate, popout rate and flyout rate are not particulary significant especially in comparison with home run rate and lineout rate, however they do add some value to our prediction. Groundout rate has a negative correlation with home run rate, and both popout and flyout rates have a positive correlation with home run rate. This makes sense because batters who generally hit grounders do not have the capacity to hit home runs as often, and also batters who hit popouts and flyouts likely do have more capacity to hit home runs.

Part d: h-rate

(i) Intercept: 0.2404, Coefficient: 0.1720

(ii.A) With the basic model only using hrate_prev, the R^2 value was 0.029. With the additional features, R^2 improved to 0.073. For predicting h-rate, the following features provided the best model performance:

- hrate_prev (coeff = 0.1497); significant
- lo_rate_prev (coeff = 0.2740); very significant
- go_rate_prev (coeff = 0.0483); not very significant
- fo_rate_prev (coeff = 0.1341); significant
- po_rate_prev (coeff = -0.2988); most signifcant

(ii.B) Once again, it is suprising in this model that hrate_prev is not the most significant feature based on the magnitude of our coefficients, as we would expect to predict the new season's hit rate based on the previous sesaon's hit rate. However as we learned, the hit rate statistic is not very indicative of future hit rates and there is not a very high correlation between one season and the next. For this reason, we must rely on other features to help predict hit rate of the next season. We see that popout rate and lineout rate both provide high predictive power and in opposite directions! Popout rate has a negative correlation with hit rate and lineout rate has a positive correlation with hit rate. The other two features,

groundout rate and flyout rate are also positively correlated with hit rate, however groundout rate is pretty insignificant compared to the other predicting features.