Comparison of projected wins of three projection systems in Major League Baseball

1. Introduction

In Major League Baseball (MLB), there are many projection systems attempting to predict the numbers of wins achieved by teams in a season. These projections are usually made prior to the start of the season. Player Empirical Comparison and Optimization Test Algorithm (PECOTA), sZymborski Projection System (ZiPS), and Steamer are three well known projection systems in MLB. Initially, we wish to compare the predicted wins of these three projection systems for the seasons of 2008-2018. However, no data was available from Steamer for 2008-2012. But data from FanGraphs aggregating Steamer and ZiPS was available for 2013-2018. So FanGraphs will be used in place of Steamer in our data analysis for these six years. Thus the projected wins as well as the observed wins of MLB teams are compiled for this period to compare the effectiveness of the predictions of these three projection systems. Chu and Wang (2019) suggested that the preseason projected wins could be used to help assess whether a team's belief in analytics has a positive impact on the team.

We first consider the sum of squares of the difference between the projected wins and observed wins for each of these three systems. However, this squared mathematical distance does not take into account various random factors affecting the numbers of games won by teams in a season. Rather, the squared statistical distance or Mahalanobis distance between the projected wins and observed wins will be examined here. Three models are proposed based on which we assess the Mahalanobis distance of the three projection systems.

As a first approximation, we assume in Model 1 that the numbers of wins obtained by teams are independent. Each number of wins can be regarded as a binomial random variable. As there are 30 teams in MLB, the Mahalanobis distance follows approximately a chi-square distribution with 30 degrees of freedom. Under the null hypothesis that the projected wins are plausible values of the actual wins for all 30 MLB teams in a given year, we are able to calculate the observed Mahalanobis distance and the corresponding p-value. This p-value allows us to reach the conclusion whether all the differences between projected and actual

wins are statistically significant or not for a particular projection system.

Model 2 imposes the correlation structure for the numbers of wins obtained by teams. Three different kinds of correlation structures are considered: linear, squared, and logarithmic. However, only the squared and logarithmic structures are suitable for computing the Mahalanobis distance because the associated variance-covariance matrices are invertible.

The number of wins obtained by a team is further broken down into the sum of numbers of wins in the matchup games against each team. This approach in Model 3 gives us a more precise assessment for each number of wins in the matchup games. The covariance between the numbers of wins obtained by two different teams will be estimated. A new variance-covariance matrix can then be formulated to recalculate a more accurate Mahalanobis distance between the projected wins and observed wins. These three models are shown in Section 2.

Bonferroni confidence intervals, confidence ellipsoids in higher dimensional spaces, and Benjamini-Hochberg procedure for multiple hypothesis testing are used to compare the effectiveness of these three projection systems. These results are presented in Section 3.

In Section 4, simulations are implemented to generate 1,000 realizations to test whether the p-values of Mahalanobis distances are within the threshold of 5% level of significance. The 95% confidence intervals, Bonferroni confidence intervals, and Benjamini-Hochberg procedure for multiple hypothesis testing are employed again to compare these three projection systems. The checking of the validity of normality assumption will be discussed in Section 5. Finally, conclusion and comments are given in Section 6.

2. Modeling

Let $w_1, w_2, ..., w_{30}$ be the projected wins for the 30 MLB teams in a season. PECOTA, ZiPS, and FanGraphs will be one of the three projection systems that generates these projected wins. Suppose that $x_1, x_2, ..., x_{30}$ be the corresponding observed wins for these 30 MLB teams in that particular season. To measure the accuracy of the projected wins generated by a projection system, one may consider the squared mathematical distance

$$D_0 = (x_1 - w_1)^2 + (x_2 - w_2)^2 + ... + (x_{30} - w_{30})^2.$$

Then one may compare the three different values of D_0 produced by PECOTA, ZiPS, and FanGraphs. The smaller the value of D_0, the better the projection system is to generate the projected wins for MLB teams. However, where do we draw the line on the value of D_0 beyond which the projection system is deemed to be not effective in generating the projected wins for teams. We need to develop some models to help answer this question.

2.1. Model 1

As each MLB team usually plays $n = 162$ games in a season, the projected winning percentage for Team i is $w_i/n, i = 1, 2, ..., 30$. The observed wins x_i can be regarded as an observed value coming from a random variable X_i representing the number of games won by Team i in a season. For the time being, let's first assume the independence of the winning of games for each team, i.e., the winning of one game for a team does not affect its winning of another game. Hence the random variable X_i can be treated as a binomial random variable with parameters $n = 162$ and p_i, where n is the number of games played in a season and p_i is the probability of winning a game for Team i. (If the team plays less than 162 games in a season, then n will be changed accordingly.) Let μ_i and σ_i^2 be the mean and variance of X_i, respectively.

It is well known that X_i can be approximated by a normal distribution when n is large. In practice, this approximation is adequate whenever $n * p_i \geq 15$ and $n * (1 - p_i) \geq 15$. These conditions imply that the number of wins and number of losses are both at least 15 games in a season. Since 162 games are played by each team in a season, these two conditions are certainly satisfied by each team.

Suppose we further assume the independence of X_i and X_j for all i and j, $i \neq j$. We then consider the squared statistical distance or Mahalanobis distance of $X = [X_1, X_2, ..., X_{30}]'$ that is to measure the distance between X and its mean while taking into account the shape of the distribution. In this case, we have

$$D_1 = (\frac{X_1 - \mu_1}{\sigma_1})^2 + (\frac{X_2 - \mu_2}{\sigma_2})^2 + ... + (\frac{X_{30} - \mu_{30}}{\sigma_{30}})^2 \quad (1)$$

Each term $((X_i - \mu_i)/\sigma_i)^2$ in (1) has approximately a chi-square distribution with 1 degree of freedom. Based on the assumption of the independence of X_i, it seems that D_1 has

approximately a chi-square distribution with 30 degrees of freedom. Since there is rarely a tie in an MLB game, the sum of X_i can be treated as a constant that is the total number of games played in a season. Hence the number of degrees of freedom is adjusted to 29.

We wish to test H_0: Projected wins are plausible values of the actual wins for all 30 MLB teams in a given year versus H_a: Projected wins are not plausible values of the actual wins for all 30 MLB teams in that given year (i.e., at least one projected win is not plausible). We will consider three sets of hypothesis testing, one for each projection system: PECOTA, ZiPS, Fangraphs. Since the winning percentage is the number of wins/n, the above hypothesis testing is equivalent to testing the projected winning percentages ($w_i/n = \tilde{p}_i$) are plausible values of the actual winning percentages (p_i) for all MLB teams. Hence H_0 is changed to $p_i = \tilde{p}_i, i = 1, 2, ..., 30$. The mean and variance of X_i, under H_0, can be estimated by

$$\hat{\mu}_i = n\tilde{p}_i = w_i \tag{2}$$

$$\hat{\sigma}_i{}^2 = n\tilde{p}_i(1 - \tilde{p}_i) = w_i(1 - w_i/n) \tag{3}$$

Table 1 shows the distance D_1 between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, FanGraphs for 2008-2018. The corresponding p-values are also given in the parentheses. We find that all p-values, except the ones of ZiPS and FanGraphs in 2016, were less than 0.05. Other than these two instances, with 5% level of significance, there was sufficient evidence to show that the projected wins produced by each of these three projection systems were not plausible values of the actual wins for all 30 MLB teams for 2008-2018. It implies that at least one projected win was significantly different from the corresponding actual win. For ZiPS and FanGraphs in 2016, however, there was insufficient evidence to show any significant difference of at least one projected win and the corresponding actual win.

[Table 1 here.]

Comparing PECOTA and ZiPS during the period of 2008-2018, we see that PECOTA produced smaller values of D_1 (or larger p-values) than ZiPS for 6 out of 11 years. During the period of 2013-2018, FanGraphs had the smallest values of D_1 for 3 years, PECOTA for 2 years, and ZiPS for 1 year. A smaller value of D_1 implies a shorter statistical distance

between the projected wins and actual wins after variances are taken into account. So the smaller the value of D_1, the better the projection system performs.

2.2. Model 2

The assumption of independence of X_i and X_j, for $i \neq j$, may not hold true in Model 1. It is because the sum of all variables $X_1 + X_2 + ... + X_{30} = 162 * 30/2 = 2430$. When a team wins a baseball game, it means another team loses a game. This is due to the fact that it is a zero-sum game and there is (almost) no tie for a game. Hence some correlation may exist between X_i and X_j. Let the correlation between X_i and X_j, $i, j = 1, 2, ..., 30, i \neq j$, be

$$\rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j},$$

where $\sigma_{i,j}$ is the covariance between X_i and X_j, and σ_i is the standard deviation of X_i. Thus,

$$\sigma_{i,j} = \rho_{i,j} \sigma_i \sigma_j.$$

Negative (or non-positive) correlation is expected to exist between X_i and X_j because of the zero-sum games. Let $m_{i,j}$ be the number of baseball games played between Teams i and j, where $i \neq j$. Note that $m_{i,j} = m_{j,i}$ and $m_{i,i} = 0$ for all i. For the extreme cases: when $m_{i,j} = 0$, it implies that $\rho_{i,j} = 0$; when $m_{i,j} = 162$, it implies that $\rho_{i,j} = -1$. Furthermore, when $0 < m_{i,j} < 162$, we have $-1 < \rho_{i,j} < 0$. As $m_{i,j}$ increases from 0 to 162, the value of $\rho_{i,j}$ decreases from 0 to -1.

Here we consider three models for $\rho_{i,j}$, $i \neq j$, satisfying all conditions mentioned above.

Model 2a: $\rho_{i,j} = -m_{i,j}/162$ (a linear model);

Model 2b: $\rho_{i,j} = -(m_{i,j}/162)^2$ (a squared model);

Model 2c: $\rho_{i,j} = -\log_2((m_{i,j}/162) + 1)$ (a logarithmic model).

Fig. 1 displays the graphs of functions $\rho_{i,j}$'s in Models 2a-2c.

[Fig. 1 here.]

The correlation matrix is given by

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{1,2} & ... & \rho_{1,30} \\ \rho_{2,1} & 1 & ... & \rho_{2,30} \\ . & . & ... & . \\ . & . & ... & . \\ \rho_{30,1} & \rho_{30,2} & ... & 1 \end{bmatrix} \tag{4}$$

The variance-covariance matrix is given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & ... & \rho_{1,30}\sigma_1\sigma_{30} \\ \rho_{2,1}\sigma_2\sigma_1 & \sigma_2^2 & ... & \rho_{2,30}\sigma_2\sigma_{30} \\ . & . & ... & . \\ . & . & ... & . \\ \rho_{30,1}\sigma_{30}\sigma_1 & \rho_{30,2}\sigma_{30}\sigma_2 & ... & \sigma_{30}^2 \end{bmatrix} = \Lambda\boldsymbol{\rho}\Lambda \tag{5}$$

where Λ is a 30×30 diagonal matrix with diagonal elements $\sigma_1, \sigma_2, ..., \sigma_{30}$.

Suppose that $X = [X_1, X_2, ..., X_{30}]'$ follows a multivariate normal distribution $N_{30}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean vector is $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_{30}]'$ and variance-covariance matrix $\boldsymbol{\Sigma}$ is given by (5) with $\rho_{i,j}$ being one of the values shown in Model 2a, 2b or 2c, and σ_i^2 estimated by (3) under H_0. Let

$$D_2 = (X - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}) \tag{6}$$

be the Mahalanobis distance between X and $\boldsymbol{\mu}$ while taking into account the covariances among those variables in X. Since the sum of X_i can be treated as a constant, Johnson and Wichern (2019) shows that D_2 has a chi-square distribution with 29 degrees of freedom.

When $\rho_{i,j} = 0, i \neq j$, D_2 is reduced to D_1 that involves no correlations among the variables. Hence D_2 is a generalization of D_1 when correlations are taken into account. In our situation, there are 30 variables involved as 30 MLB teams play in a season. So each correlation between any two variables might not be too large. Would D_2 generate a significantly different value from that of D_1 when all correlations are involved in calculating D_2? Or would D_1 be able to provide a good approximation for D_2?

Let's first consider Model 2a. For fixed j, $\Sigma_{i=1,i\neq j}^{30}\rho_{i,j} = -\Sigma_{i=1,i\neq j}^{30} m_{i,j}/162 = -1$. It implies that $\Sigma_{i=1}^{30}\rho_{i,j} = 0$ as $\rho_{i,i} = 1$ for all i. In this case, the determinant of $\boldsymbol{\rho}$ in (4) is zero. Consequently, the inverse of $\boldsymbol{\rho}$ does not exist. It implies that the inverse of the corresponding variance-covariance matrix $\boldsymbol{\Sigma}$ in (5) does not exist either. As a result, we will not be able to compute D_2 in (6) when $\rho_{i,j} = -m_{i,j}/162$ as shown in Model 2a.

However, the correlation matrix (4) with $\rho_{i,j} = -(m_{i,j}/162)^2$ given in Model 2b will not generate zero determinant. Therefore, the inverse of the correlation matrix exists and so does the corresponding variance-covariance matrix $\boldsymbol{\Sigma}$ in (5). Likewise, with the condition $\rho_{i,j} = -\log_2((m_{i,j}/162) + 1)$ given in Model 2c, both the inverse of correlation matrix and the inverse of variance-covariance matrix exist. Therefore, we will compute D_2 only for the values of $\rho_{i,j}$ given in Models 2b and 2c.

Table 2 shows the distance D_2 between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, FanGraphs for 2008-2018, using Model 2b. All p-values, except the ones of ZiPS and FanGraphs in 2016, were less than 0.05. Other than these two instances, with 5% level of significance, there was sufficient evidence to show that the projected wins produced by these three projection systems were not plausible values of the actual wins for all 30 MLB teams for 2008-2018.

For the period of 2013-2018, FanGraphs had the smallest value of D_2 for 3 years, PECOTA for 2 years, and ZiPS for 1 year. For the longer period of 2008-2018 and without considering FanGraphs, PECOTA had the smaller value of D_2 for 6 out of 11 years.

[Table 2 here.]

Table 3 shows the distance D_2 between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, FanGraphs for 2008-2018, using Model 2c. The p-values of PECOTA (2014), ZiPS (2016, 2014), and FanGraphs (2016, 2014, 2013) were greater than 0.05. With 5% level of significance, there was insufficient evidence to reject the null hypothesis that the projected wins of the 30 MLB teams were plausible values of their actual wins for those years. Besides these six instances, there was sufficient evidence to show that at least one projected win was not a plausible value of the actual win of these 30 MLB teams.

PECOTA, ZiPS, and FanGraphs each had the smallest value of D_2 for two years during

the period of 2013-2018. However, PECOTA had the smaller value of D_2 than that of ZiPS for 7 out of 11 years during the period of 2008-2018.

[Table 3 here.]

2.3. Model 3

To improve the covariance structure of $\mathbf{\Sigma}$, we decompose each X_i (the total number of games won by Team i in a season) into $X_{i,j}$ (the number of games won by Team i over Team j in that season). More specifically,

$$X_1 = X_{1,1} + X_{1,2} + ... + X_{1,30} \quad (7)$$

$$X_2 = X_{2,1} + X_{2,2} + ... + X_{2,30} \quad (8)$$

$$\vdots \quad \vdots$$

$$X_{30} = X_{30,1} + X_{30,2} + ... + X_{30,30} \quad (9)$$

Note that $X_{i,i} = 0, i = 1, 2, ..., 30$. Because of zero-sum games, we have $X_{i,j} + X_{j,i} = m_{i,j}$ (or $m_{j,i}$) that is the number of games played between Teams i and j, where $i, j = 1, 2, ..., 30$ and $i \neq j$. For example, Arizona and Colorado played 19 games against each other in 2018; however, Arizona and Baltimore did not play any game against each other in that year. The numbers of matchup games between teams in 2012 were changed to new numbers of matchup games in 2013 and onwards, e.g., Arizona and Colorado played only 18 games against each other in 2012.

Let us first consider the covariance between X_1 and X_2. Due to the independence of games Teams 1 and 2 played against Teams $3, 4, ..., 30$ (i.e., $Cov(X_{1,2}, X_{2,j}) = 0, j = 3, 4, ..., 30$ and $Cov(X_{1,i}, X_{2,j}) = 0, i, j = 3, 4, ..., 30$), the covariance between X_1 and X_2 becomes

$$\begin{aligned} Cov(X_1, X_2) &= Cov(X_{1,1} + X_{1,2} + ... + X_{1,30}, X_{2,1} + X_{2,2} + ... + X_{2,30}) \\ &= Cov(X_{1,2}, X_{2,1}) \\ &= Cov(X_{1,2}, m_{1,2} - X_{1,2}) \\ &= -Cov(X_{1,2}, X_{1,2}) \end{aligned} \quad (10)$$

$Cov(X_{1,2}, m_{1,2})$ is zero as $m_{1,2}$ is a fixed number. The covariance term $Cov(X_{1,2}, X_{1,2})$ is simply the variance term $Var(X_{1,2})$. $X_{1,2}$ can be regarded as a binomial random variable with

parameters $m_{1,2}$ and $p_{1,2}$, where $m_{1,2}$ is the number of games played between Teams 1 and 2, and $p_{1,2}$ is the probability that Team 1 will win over Team 2 in a game. Thus the variance of $X_{1,2}$ is $m_{1,2} * p_{1,2} * (1 - p_{1,2})$. Similarly, the variance of $X_{2,1}$ is $m_{2,1} * p_{2,1} * (1 - p_{2,1})$, where $m_{2,1} = m_{1,2}$ and $p_{2,1}$ is the probability that Team 2 will win over Team 1 in a game. Note that $p_{1,2} + p_{2,1} = 1$. Since $Cov(X_1, X_2) = Cov(X_2, X_1)$, it implies that $Var(X_{1,2}) = Var(X_{2,1})$. Thus $p_{1,2} * (1 - p_{1,2}) = p_{2,1} * (1 - p_{2,1})$, and this equation is always true since $p_{1,2} + p_{2,1} = 1$.

Under the equivalent percentage version of H_0, i.e., $p_i = w_i/n, i = 1, 2, ..., 30$, $p_{1,2}$ can be estimated by $p_1/(p_1 + p_2) = (w_1/n)/((w_1/n) + (w_2/n)) = w_1/(w_1 + w_2)$. Similarly, $p_{2,1}$ can be estimated by $p_2/(p_1 + p_2) = (w_2/n)/((w_1/n) + (w_2/n)) = w_2/(w_1 + w_2)$. With these estimations, $p_{1,2} + p_{2,1}$ is always 1. Hence, $Cov(X_1, X_2)$ in (10) can be estimated by $-m_{1,2} * w_1/(w_1 + w_2) * w_2/(w_1 + w_2)$ under H_0. Similarly, $Cov(X_2, X_1)$ can be estimated by $-m_{2,1} * p_{2,1} * (1 - p_{2,1}) = -m_{1,2} * w_2/(w_1 + w_2) * w_1/(w_1 + w_2)$, which is the estimated value of $Cov(X_1, X_2)$. By following the above procedure for the general terms $i \neq j$, $Cov(X_i, X_j) = Cov(X_j, X_i)$ can be estimated by

$$-m_{i,j} * w_i/(w_i + w_j) * w_j/(w_i + w_j) \text{ under } H_0. \quad (11)$$

The mean and variance of $X_i, i = 1, 2, ..., 30$, are estimated by (2) and (3), respectively, under H_0. Therefore, the variance-covariance terms $Cov(X_i, X_j), i, j = 1, 2, ..., 30$, shown in (3) and (11) can be computed directly to form the entries of $\mathbf{\Sigma}$ in (6). With this matrix $\mathbf{\Sigma}$, we are able to compute its inverse. The value of D_2 in (6) can then be calculated to compare the distance between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, FanGraphs for 2008-2018 under H_0. The corresponding p-values can also be evaluated using the chi-square distribution with 29 degrees of freedom. The results are given in Table 4.

The p-values of PECOTA (2016), ZiPS (2016), and FanGraphs (2016, 2014) were greater than 0.05. With 5% level of significance, there was insufficient evidence to reject the null hypothesis that the projected wins of the 30 MLB teams were plausible values of their actual wins for those years. Besides these four instances, there was sufficient evidence to show that at least one projected win was not a plausible value of the actual win of these 30 MLB teams.

FanGraphs had the smallest value of D_2 for 4 years, PECOTA 2 years, and ZiPS none during the period of 2013-2018. For the longer period of 2008-2018, ZiPS had the smaller value of D_2 than that of PECOTA for 6 out of 11 years.
[Table 4 here.]

3. Confidence regions

3.1. Bonferroni confidence intervals

From Model 1, X_i can be treated as a binomial random variable with parameters $n = 162$ and p_i. The unknown p_i can be estimated by $\hat{p}_i = x_i/n$, where x_i is the number of observed wins for Team i in the season. Hence an approximately 95% confidence interval for $p_i, i = 1, 2, ..., 30$, is

$$\hat{p}_i \pm z_{0.025}\sqrt{\hat{p}_i(1-\hat{p}_i)/n}, \tag{12}$$

where $z_{0.025} = 1.960$ is the upper 2.5% critical value of the standard normal distribution.

Consider the projected winning percentages $\tilde{p}_i = w_i/n, i = 1, 2, ..., 30$, and see how many of them fall in the corresponding confidence interval for p_i shown in (12). If any one of the $\tilde{p}_i$'s does not fall in the corresponding confidence interval for p_i, then we can say that at least one of the projected wins is different from one of the actual wins with probability approximately $1 - (0.95)^{30} \approx 0.7854$. Therefore, the chances that all 30 $\tilde{p}_i$'s fall in the corresponding confidence interval simultaneously are approximately 0.2146.

In order to adjust the overall confidence level from 21.46% to 95%, we use the Bonferroni confidence interval as follows:

$$\hat{p}_i \pm z_{0.025/30}\sqrt{\hat{p}_i(1-\hat{p}_i)/n}, \tag{13}$$

where $i = 1, 2, ..., 30$ and $z_{0.025/30} = 3.144$. When $z_{0.025/30}$ is used in (13) instead of $z_{0.025}$, there are 95% chances that all 30 projected winning percentages $\tilde{p}_i$'s fall in the corresponding Bonferroni confidence interval simultaneously. Table 5 shows the number of projected winning percentages for PECOTA, ZiPS, and FanGraphs, falling in the corresponding 95% Bonferroni confidence interval for 2008-2018.

[Table 5 here.]

It is desirable to see all 30 projected winning percentages simultaneously fall in the corresponding 95% Bonferroni confidence interval. However, only PECOTA (2013, 2008), ZiPS (2016, 2015, 2013), and FanGraphs (2016, 2015, 2013) had achieved this. We see that PECOTA and ZiPS produced the same mean (28.45) and comparable values (1.13, 1.29) for standard deviation. It seems that FanGraphs produced more accurate and precise results with higher mean (29.5) and smaller standard deviation (0.55). Nevertheless, the one-way analysis of variance shows that there is insufficient evidence at 5% level of significance to support that the average numbers of projected wins simultaneously falling in the corresponding 95% Bonferroni confidence interval are not the same for these three projection systems.

3.2. Confidence ellipsoids

Under Models 2b-2c and 3, $X = [X_1, X_2, ..., X_{30}]'$ follows a multivariate normal distribution $N_{30}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Recall that D_2 in (6) has a chi-square distribution with 29 degrees of freedom. Hence a 95% confidence ellipsoid for $\boldsymbol{\mu}$ is

$$D_2 = (X - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}) \leq \chi^2_{29}(0.05) \tag{14}$$

where $\chi^2_{29}(0.05) = 42.56$ is the upper 5% critical value of the chi-square distribution with 29 degrees of freedom. Under H_0, $\boldsymbol{\mu}$ can be estimated by $[w_1, w_2, ..., w_{30}]'$ shown in (2). With confidence 95%, the vector of observed wins $x = [x_1, x_2, ..., x_{30}]'$ should fall in the above ellipsoid given in (14).

When D_2 is less than or equal to 42.56, this is equivalent to the associated p-value greater than 0.05 as seen in Tables 2-4. Consequently, we obtain the same results and conclusions as presented in Sections 2.2 and 2.3.

3.3. Multiple hypothesis testing

The usual naive method of statistical testing on a single hypothesis may not be suitable for testing multiple hypotheses with the same significance level. The Bonferroni method, however, is usually more conservative and results in more acceptance of the status quo

H_0. The Benjamini-Hochberg procedure for multiple hypothesis testing can be used to test the significance of multiple statements. This procedure tends to balance the effect of the previous two approaches and is helpful in reducing false positives (type I error). For more details about the Benjamini-Hochberg procedure, see their paper (1995) or Tan et. al (2019).

Suppose we wish to test $H_0 : p_1 = w_1/162$ versus $H_a : p_1 \neq w_1/162$. Under H_0, the test statistic is

$$Z \approx \frac{x_1/162 - p_1}{\sqrt{p_1 * (1 - p_1)/162}} = \frac{x_1/162 - w_1/162}{\sqrt{w_1/162 * (1 - w_1/162)/162}} = \frac{x_1 - w_1}{\sqrt{w_1 * (162 - w_1)/162}}. \quad (15)$$

We calculate the observed test statistic z and then find the corresponding p-value called PV_1. Repeat the same process for the other 29 teams to obtain $PV_2, PV_3, ..., PV_{30}$. Rearrange these p-values in descending order to obtain

$$PV_{(30)} \geq PV_{(29)} \geq ... \geq PV_{(1)}. \quad (16)$$

Compare these ordered p-values term-wise with significance levels

$$\alpha > \frac{29}{30}\alpha > ... > \frac{1}{30}\alpha, \quad (17)$$

i.e., compare $PV_{(i)}$ with $(i/30)\alpha, i = 1, 2, ..., 30$. Choose the largest K such that $PV_{(K)} \leq (K/30)\alpha$ to declare that K of the projected winning percentages are statistically different from the actual winning percentages with significance level approximately α, say 5%. The smaller the value of K, the fewer the projected winning percentages are different from the actual ones and hence the better the projection system is. It is desirable to have $K = 0$, indicating that all 30 pairs of projected and actual winning percentages are not statistically significantly different.

Table 6 displays the multiple hypothesis testing for PECOTA, ZiPS, and FanGraphs, showing K distinct projected winning percentages from the actual winning percentages. PECOTA (2015, 2013, 2008), ZiPS (2016, 2015), and FanGraphs (2015, 2013) have $K = 0$, indicating that there was no significant difference between the projected and actual winning percentages for these years. By comparing the mean and standard deviation of K, it seems that FanGraphs is preferable. Nevertheless, the one-way analysis of variance shows that

there is insufficient evidence at 5% level of significance to support that the true average numbers of projected winning percentages distinct from the actual winning percentages are not the same for these three projection systems.

[Table 6 here.]

4. Simulations

Instead of having only one instance/realization of the observed wins for each team to be compared with the projected wins, simulations are implemented to generate 1,000 realizations of the observed wins. This allows us to have more extensive comparison of the observed wins with the projected wins. By doing so, we may be able to achieve more reliable result on the effectiveness of the projection systems in forecasting the actual wins of MLB teams.

From Model 3, $X_{i,j}$ (the number of games won by Team i over Team j in a season), $i \neq j$, can be regarded as a binomial random variable with parameters $m_{i,j}$ and $p_{i,j}$, where $m_{i,j}$ is the number of games played between Teams i and j, and $p_{i,j}$ is the probability of winning for Team i over Team j. The parameter $p_{i,j}$ can be estimated by $\hat{p_{i,j}} = x_{i,j}/m_{i,j}$, where $x_{i,j}$ is the observed wins for Team i over Team j in $m_{i,j}$ games. $\hat{p_{i,j}}$ will be modified to $(x_{i,j}+1)/(m_{i,j}+2)$ if $x_{i,j} = 0$ or $m_{i,j}$ so that $\hat{p_{i,j}}$ will not be 0 or 1 in the simulated binomial distribution. Searching from the MLB record books, we are able to find all the observed wins $x_{i,j}$ for Team i over Team j in each season of 2008-2018.

Now we run 1,000 simulations using R to generate the simulated values for the binomial distribution of $X_{i,j}$ with parameters $m_{i,j}$ and $\hat{p_{i,j}}$. We only need to generate the simulated values of $X_{i,j}$ for $1 \leq i < j \leq 30$. It is not necessary to generate $X_{j,i}$ because $X_{j,i} = m_{i,j} - X_{i,j}$. Note that $X_{i,i} = 0, i = 1, 2, ..., 30$. Using equations (7)-(9) and the simulated values $X^*_{i,j}$, we obtain the simulated value X^*_i for $X_i, i = 1, 2, ..., 30$. Let $X^* = [X^*_1, X^*_2, ..., X^*_{30}]'$ be the vector of simulated values for X. Then the simulated Mahalanobis distance is

$$D_3 = (X^* - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X^* - \boldsymbol{\mu}) \tag{18}$$

that follows approximately a chi-square distribution with 29 degrees of freedom.

Under the equivalent percentage version of H_0, $\boldsymbol{\mu}$ in D_3 can be replaced by $[w_1, w_2, ..., w_{30}]'$

shown in (2). Applying the estimates of $\text{Var}(X_i)$ in (3) and $Cov(X_i, X_j), i \neq j$, in (11), we are able to compute the inverse of $\mathbf{\Sigma}$ in (6) and hence the value of D_3 in (18) using the simulated value of X^*.

We obtain 1,000 observed values of D_3. First, we compare them with $\chi^2_{29}(0.05)$ to see how many of them falling outside the 95% confidence interval. Second, we compare them with $\chi^2_{29}(0.05/1000)$ to see how many of them falling outside the 95% Bonferroni confidence interval. Third, we do a multiple hypothesis testing using Benjamini-Hochberg procedure to test the significance of multiple statements based on these 1,000 simulated Mahalanobis distances.

In Benjamini-Hochberg procedure, we first obtain the corresponding p-values of these 1,000 simulated observed values of D_3 and arrange them in descending order, i.e.,

$$PV^*_{(1000)} \geq PV^*_{(999)} \geq ... \geq PV^*_{(1)}. \tag{19}$$

Then we compare them term-wise with

$$\alpha > \frac{999}{1000}\alpha > ... > \frac{1}{1000}\alpha. \tag{20}$$

Choose the largest K such that $PV^*_{(K)} \leq (K/1000)\alpha$ to declare that the projected wins are not plausible values of the actual wins for K simulated instances. The results are given in Table 7.

[Table 7 here.]

Table 7 shows that, except for 2016, at least 98.9% of these 1,000 simulated observed wins have rejected the null hypothesis that the PECOTA projected wins were plausible values of the actual wins at $\alpha = 0.05$ level of significance. Year 2016, however, had a distinct pattern from other years with the rejection rate of 72.5% at $\alpha = 0.05$, 10.7% at $\alpha =$ 5E-5, and 67.8% using the Benjamini-Hochberg procedure. Similar results happened to ZiPS and FanGraphs. The Benjamini-Hochberg procedure, a balance between the liberal naive approach and conservative approach, still shows the rejection rate of at least 98.9% for the three projection systems in all those years except 2016. For 2016, the rejection rates were 67.8%, 61.4%, and 30.5% for PECOTA, ZiPS, and FanGraphs, respectively. It looks like FanGraphs might be preferable among these three projection systems on some occasions.

5. Checking the validity of normality assumption

The results obtained above is based on the multivariate normal assumption for the vector of actual wins $X = [X_1, X_2, ..., X_{30}]'$. Due to the limitation of the visual effect on higher dimensions, we have used the 95% probability plot to check the univariate normality for each component $X_i, i = 1, 2, ..., 30$. As well, we have used the 95% confidence contour plot to check the bivariate normality for each pair $(X_i, X_j), 1 \leq i \neq j \leq 30$. Note that there are totally ${}^{30}C_2 = 435$ such pairs. There are no significant violations of the univariate normality for any component $X_i, i = 1, 2, ..., 30$ nor the bivariate normality for any pair $(X_i, X_j), 1 \leq i \neq j \leq 30$, for the years of 2008-2018. To save space, we will not display the probability plots or confidence contour plots here.

6. Conclusion and comments

With Models 1 and 2b, Tables 1-2 show that 26 out of all 28 projected wins were not plausible values of the actual wins of MLB teams at the 5% level of significance, except for ZiPS (2016) and FanGraphs (2016). Table 3 using Model 2c shows that 22 out of 28 projected wins were not plausible values of the actual wins of MLB teams at $\alpha = 5\%$ except for FanGraphs (2016, 2014, 2013), ZiPS (2016, 2014), and PECOTA (2014). Since the assumption of independence of the numbers of wins by teams is violated, the number of wins by a team is further decomposed into the sum of numbers of wins in the matchup games against each team. This approach in Model 3 gives a more precise assessment for each number of wins in the matchup games and hence provides more accurate results. Table 4 using Model 3 reveals that 24 out of 28 projected wins were not plausible values of the actual wins of MLB teams at $\alpha = 5\%$ except for FanGraphs (2016, 2014), ZiPS (2016), and PECOTA (2016). These results are based on the multivariate normal assumption for the vector of numbers of wins of the 30 MLB teams. The checking of the validity of normality doesn't show any significant violation of the assumption.

It is extremely difficult to accurately predict the outcomes of the numbers of wins achieved by all MLB teams in a season of 162 games. There are so many unpredictable variables evolving in teams during the season. Some of these variables could be injuries of key players,

adaptation of new players, errors made by players in games, etc. It seems that these three projection systems were not effective in predicting the numbers of wins achieved by MLB teams, although FanGraphs looked more promising than the other two systems. These projected wins, however, could serve as the expectation of the performance of each team prior to the start of a new season. As the season progresses, the update of the projected wins (as some projection systems are doing) is necessary to provide more accurate predicted numbers of wins for all MLB teams.

References

Baseballprospectus.com, 2008-2018. 'PECOTA projection system wins'. URLs:
https://web.archive.org/web/20180322120134/https://legacy.baseballprospectus.com/fantasy/dc/
https://web.archive.org/web/20170402051004/http://www.baseballprospectus.com/fantasy/dc/
http://web.archive.org/web/20160404005757/http://www.baseballprospectus.com/fantasy/dc/
https://web.archive.org/web/20150330052122/http://www.baseballprospectus.com/fantasy/dc
https://web.archive.org/web/20140320120227/http://www.baseballprospectus.com/fantasy/dc/
https://web.archive.org/web/20130217222041/http://www.baseballprospectus.com/fantasy/dc
https://web.archive.org/web/20120315124341/http://www.baseballprospectus.com/fantasy/dc/
https://web.archive.org/web/20110227231147/http://www.baseballprospectus.com/fantasy/dc
https://web.archive.org/web/20100225083122/http://www.baseballprospectus.com/fantasy/dc
https://web.archive.org/web/20090225104533/http://baseballprospectus.com/fantasy/dc
https://web.archive.org/web/20080330230800/http://baseballprospectus.com/fantasy/dc/

Baseball-reference.com, 2008-2018. 'MLB Standings'.
https://www.baseball-reference.com/leagues/MLB/2018-standings.shtml
(Substitute 2008-2017 for 2018 to get the corresponding URL.)

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological), 289-300.*

Chu, D., Wang, C., 2019. Empirical study on relationship between sports analytics and success in regular season and postseason in Major League Baseball. *Journal of Sports*

Analytics, (5) 205-222.

'FanGraphs projection system wins'. 2013-2018.
https://www.fangraphs.com/depthcharts.aspx?position=Standings

Johnson, R., Wichern, D., 2019. Applied Multivariate Statistical Analysis, 6th ed. Pearson.

'Matchup games'. 2008-2018.
https://www.flashscore.com/baseball/usa/mlb-2018/results/

Tan, P., Steinbach, M., Karpatne, A., Kumar, V., 2019. Introduction to Data Mining, 2nd ed. Pearson, pp. 772-775.

'ZiPS projection system wins'. 2008-2018.
https://mail.google.com/mail/u/0/#search/DSzymborski%40gmail.com/FFNDWMklqMBbXnPWVkjBVvKvdJcmCwpq?projector=1&messagePartId=0.1