WARM-UP
Choose a position player who is a free agent in the upcoming off-season. Using information you can find on Baseball Savant and/or Fangraphs, what do you think will be this player's wOBA in the next season? Write down your guess, and explain your reasoning.

7 PROJECTIONS

7.1 MARCEL THE MONKEY

Perhaps the most famous (and certainly the most simple) projection model is the Marcel the Monkey[1] Forecasting System, developed by sabermetrics pioneer Tom Tango in the early 2000s.[2] Suppose we have a metric we want to project for a player in year Y. The population mean for this metric is μ_0, and we've observed the following performance for the player in the three years leading up to year Y:

- In year $Y-3$, the player's average performance was $\bar{x}_{-3}$ in n_{-3} plate appearances.
- In year $Y-2$, the player's average performance was $\bar{x}_{-2}$ in n_{-2} plate appearances.
- In year $Y-1$, the player's average performance was $\bar{x}_{-1}$ in n_{-1} plate appearances.

Ignoring aging effects momentarily, the core of Marcel is to project the following performance for year Y:

$$\frac{2 \cdot 600 \cdot \mu_0 + 3 \cdot n_{-3} \cdot \bar{x}_{-3} + 4 \cdot n_{-2} \cdot \bar{x}_{-2} + 5 \cdot n_{-1} \cdot \bar{x}_{-1}}{2 \cdot 600 + 3 \cdot n_{-3} + 4 \cdot n_{-2} + 5 \cdot n_{-1}}. \tag{1}$$

Marcel is not intended to be a competitive forecasting system. As Tango put it, "These forecasts are the minimum level of competence that you should expect from any forecaster." In other words, Marcel presents a baseline against which to compare any attempt at building a projection system. Nevertheless, Marcel turns out to be surprisingly difficult to beat, relative to its simplicity.

There are two aspects of the Marcel formula to note. First, the projection depends not only on the player's performance but also on the number of plate appearances. This is a critical point to understand about projections—if a model ignores the sample size of the performance, it is missing the fundamental job of a projection: regression to the mean. Second, the weight decreases for performances further back in time. This is another fundamental aspect of projections: that more recent performance gets more weight.

7.2 GRID SEARCH

We can write can approximation of the Marcel formula in a more general form by replacing the hard-coded coefficients with variables c and w.

$$\frac{c \cdot \mu_0 + w^2 \cdot n_{-3} \cdot \bar{x}_{-3} + w \cdot n_{-2} \cdot \bar{x}_{-2} + n_{-1} \cdot \bar{x}_{-1}}{c + w^2 \cdot n_{-3} + w \cdot n_{-2} + n_{-1}}. \tag{2}$$

In this formula, c represents the amount of mean regression applied to the metric, and $w \in (0, 1)$ represents a year-to-year multiplicative dropoff in the weight on observed performance. As we saw in the chapter on BABIP, FIP and DIPS, the ideal amount of mean regression is different depending on the metric being forecasted, but Marcel does not allow for this possibility. By allowing c and w to vary by metric, this more general formulation has the potential to improve upon Marcel.

The annual dropoff w in weight depends on two things: the noise in the data and the extent to which player talent tends to change from year to year. If player talent changes very little from year to year, then the optimal w will be closer to one, putting relatively more weight on older performance. If player talent changes a lot from year to year, then the optimal w will be closer to zero, putting less weight on older performance. If the noise is very low, then again the optimal w will be closer to one because less historical

[1] This is a reference to a character on the 90s sitcom *Friends*. This projection system "uses as little intelligence as possible."
[2] https://www.tangotiger.net/archives/stud0346.shtml

data is required to supplement the most recent data. The higher the noise, the closer this moves the optimal w to zero as more sharing across seasons is necessary to stabilize the prediction.

We cannot use linear regression to determine the optimal c and w in equation (2) because these coefficients appear in the denominator as well as in the numerator. Instead, we employ grid search, similar to how we found the optimal exponent α in the Pythagorean formula. Grid search is very simple: We specify a grid of candidate values for (c, w) and try all of them to find out which one yields the predictions with the lowest error. Once we have the optimal (c^*, w^*), we plug those numbers into equation (2) to get the projections.

7.3 Projections as Multivariable Regression to the Mean

An alternative to grid search for determining the optimal amount of weight to put on each season of historical performance is to recognize the similarity between equation (2) and regression to the mean. In the single-variable case, regression to the mean took the following form:

$$\frac{n \cdot \bar{x} + c \cdot \mu_0}{n + c},$$

where the optimal c turns out to be σ_X^2/σ_μ^2 under the assumption of the probability model

$$\begin{aligned} X &\sim \text{Normal}(\mu, \sigma_X^2/n) \\ \mu &\sim \text{Normal}(\mu_0, \sigma_\mu^2). \end{aligned}$$

When dealing with multiple years of data, we can use the vector $\vec{X} = (\bar{x}_{-3}, \bar{x}_{-2}, \bar{x}_{-1}, \bar{x}_0)^T$ to represent a player's performance over their career. The natural extension of the model above is to replace the single-variable normal distributions with multivariate normal distributions:

$$\begin{aligned} \vec{X} &\sim \text{MVN}(\vec{\mu}, \Sigma_X) \\ \vec{\mu} &\sim \text{MVN}(\vec{\mu}_0, \Sigma_\mu). \end{aligned}$$

In this model, $\vec{\mu}$ represents the player's true talent over time, and $\vec{\mu}_0$ represents the league mean, which may be constant over time or may vary. The matrix Σ_X is the noise covariance matrix for $\bar{X}$, and it is natural to assume that this is a diagonal matrix with entries σ_X^2/n_i. Lastly, Σ_μ is the prior covariance matrix on true talent, reflecting the amount of spread between players and the extent to which their true talent varies from season to season.

Given the observed performance $\vec{X}$, we can write the posterior mean of $\vec{\mu}$ as

$$\mathbb{E}\left[\vec{\mu} \,\middle|\, \vec{X} = \vec{x}\right] = \left(\Sigma_X^{-1} + \Sigma_\mu^{-1}\right)^{-1} \left(\Sigma_X^{-1}\vec{x} + \Sigma_\mu^{-1}\mu_0\right).$$

For a given entry in $\vec{\mu}$, this turns out to be a linear combination of the entries of the observed $\vec{x}$ and $\vec{\mu}_0$, with the weights summing to one. In other words, we get a projection that looks like

$$\frac{w_0 \cdot \mu_0 + w_{-3} \cdot \bar{x}_{-3} + w_{-2} \cdot \bar{x}_{-2} + w_{-1} \cdot \bar{x}_{-1}}{w_0 + w_{-3} + w_{-2} + w_{-1}},$$

mirroring equations (1) and (2). Note that the weights $w_\cdot$ depend on the sample sizes $n_\cdot$.

7.4 Aging Curves

An important component of projections we have disregarded to this point is the aging curve. The general model for the aging curve is that there is some function $f(\cdot)$ of the age which has an additive effect on player performance. If X_{it}, μ_{it} and a_{it} represent the performance, true talent and age of player i in season t, then

$$E[X_{it}] = \mu_{it} + f(a_{it}).$$

For the sake of the preceding sections, we can replace the performance X_{it} with the age-adjusted performance $X_{it} - f(a_{it})$ to project player performance relative to the aging curve. The challenge is estimating the aging curve $f(\cdot)$. There are two popular approaches, each with its drawbacks.

7.4.1 Regression

The regression-based method is to create a regression model for player performance, treating each player-season as an observation. In this setup, we can regress player performance on age, using a flexible method such as a Generalized Additive Model (GAM) to estimate the relationship between age and performance. This simple approach suffers from an obvious form of selection bias: Only the most talent players will be included in the sample at extremly young ages and extremely old ages, making the aging curve appear flatter than it really is.

Typically, the solution for this selection bias problem is to introduce player random effects to control for player talent. It is not clear how effect this approach is for mitigating the effects of selection bias, and some researchers have tried imputation to get around this problem.[3]

7.4.2 The Delta Method

The delta method is an alternative approach to aging curves, also championed by Tango.[4] Instead of considering player-seasons X_{it} as the observation unit, this method considers *changes* (i.e. the "delta") from season to season $X_{i(t+1)} - X_{it}$ as the observation unit. This has the nice feature that for calculating each step from one age to the next in the aging curve, the same players are included in the sample. By modeling $X_{i(t+1)} - X_{it}$ as a function of age, we get the derivative of the aging curve, which we integrate to get an estimate of the aging curve.

The downside to the delta method is that it suffers from a different kind of selection bias. Players with extremely low values of X_{it} (often due to luck) are more likely to drop from the sample, causing a downward bias on $X_{i(t+1)} - X_{it}$. In other words, the aging curve appears steeper than it really is. There may be a way to mitigate this selection bias by using propensity scores to adjust samples for player dropout probabilities although this has not been explored in public literature.

7.5 Level Translations

In baseball, a common problem is projecting the future Major League performance of minor league prospects. Of course it would be problematic to treat minor league performance the same as Major League performance when building projections, and that is where level translations come into play. The task of level translations is intimately connected to aging curves, with similar techniques to regression and the delta method (based on players moving between levels) available (with similar pitfalls). Especially for young minor leaguers, the question of aging curve and level translation are intertwined because prospects tend to move up through the minor league system (or else drop out) as they age, making it more difficult to separate the effects of level and the effects of aging.

7.6 Discussion

1. What do you view as the most substantial limitations of the Marcel the Monkey Forecasting System?
   (a) What incorrect assumption(s) does the projection system make, and what are the consequences of the assumption(s) being violated?
   (b) What are the most important factors that the projection system does not take into account, and how do you think you could modify the system to account for these factors?

[3] Schukers, Lopez and Macdonald (2021) "What does not get observed can be used to make age curves stronger: Estimating player age curves using regression and imputation" https://arxiv.org/abs/2110.14017

[4] https://www.tangotiger.net/aging.html