Final Project: Sports

Ferdie Taruc

Executive Summary:

In this analysis, we are determining the next 10 “Attack” or willingness to attack the basket metric of Lebron James following his 3/2/2019 career game. Assuming that Lebron’s body deteriorates like many other athletes, his willingness to attack the paint should follow a bell-shaped trend. With no clear frequency of seasonality, Lebron displays a consistent “Attack” metric from game to game throughout his career. Even under the assumption that seasonality exists following each NBA season, our parametric model with SMA(1)[32] noise predicts that Lebron’s next ten games will be both positive and consistently close.

1: Exploratory Data Analysis

From Figure 1 we do not see a clear pattern in the Lebron James’ willingness to attack metric (hereafter refer to simply as “attack”). We can see that there is a large peak in his 605th game on 2/25/11 against the Wizards, and had his smallest “attack” in his 1047th game against the Rockets on 3/12/17. If we zoom in his first 100 games, we see a slight increasing trend that continues until 2014. After his peak “attack”, there is a slight decreasing trend. It can be the case as Lebron’s body continues to deter, his willingness to attack declines. From online reports, Lebron’s body does deter as he started to miss games through injuries from 2017 onward (almost 14 years into his career). However, “attack” seems to be relatively consistent on average.

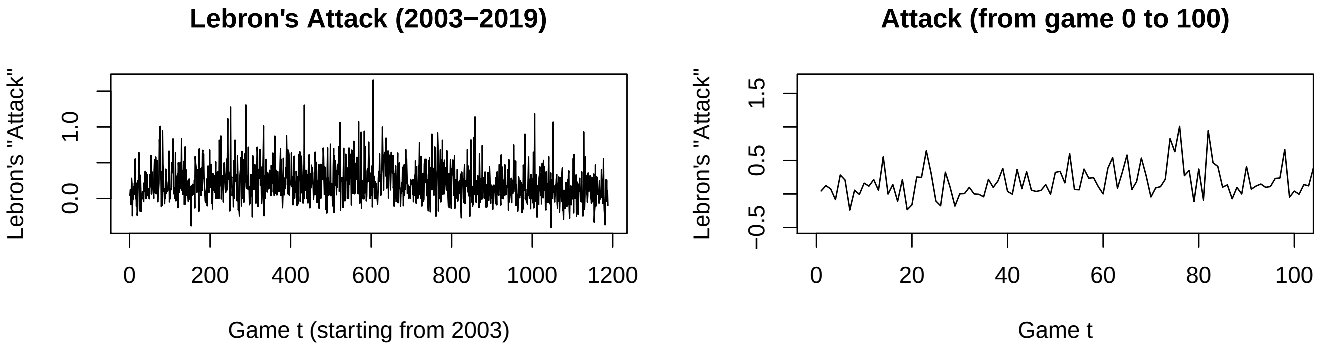


Figure 1: Lebron’s Attack throughout his 2003-2019 career. On the left panel is a small subsection that includes his career games 550 to 650.

Seasonal effects can’t be easily captured by certain periods (daily, monthly, etc) in basketball since games are spaced out abstractly by the NBA. Games are not systematically controlled; for example, a game at one certain day of a year may be under completely different conditions. However, Figure 2 shows that “attack” is consistent no matter which month or day it is in the season. Figure 3 also highlights that there are no common indicators or frequency that can model the seasonality that occurs within the data.

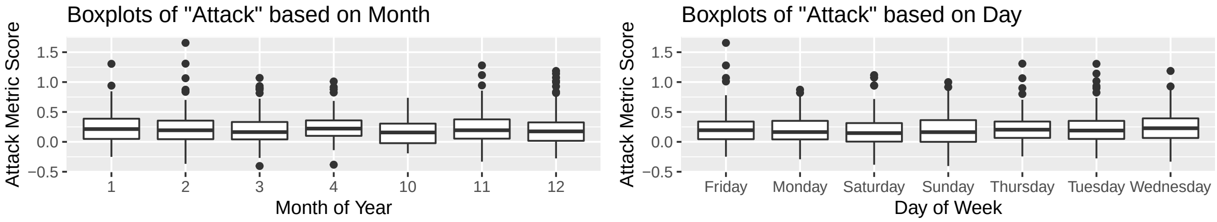


Figure 2: Boxplots of Lebron's Attack based on the day (left) and month (right) of the Year.

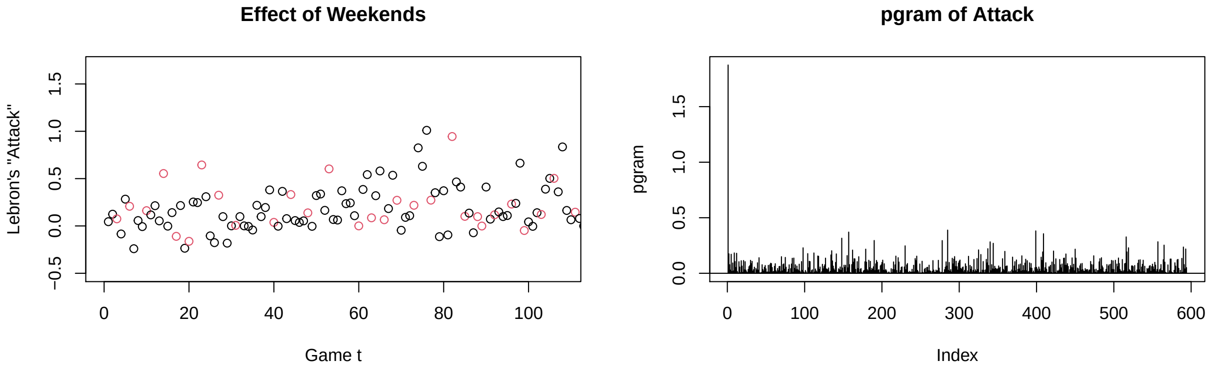


Figure 3: Analysis of seasonality of Lebron's Attack. In the left panel, red circles indicate Lebron's Attack for a game played on a weekend.

2: Models Considered

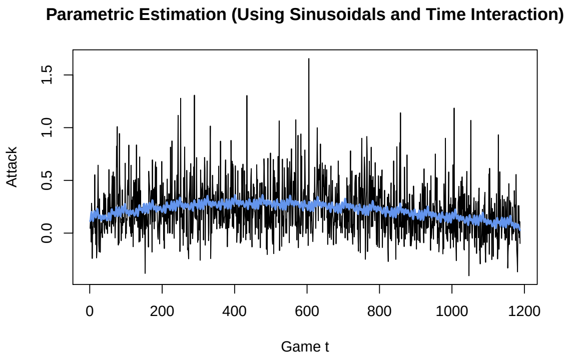
To model the natural signal in this data, both a parametric model and a seasonal differencing approach are used. Both of these models will be complimented with ARMA models for the remaining noise.

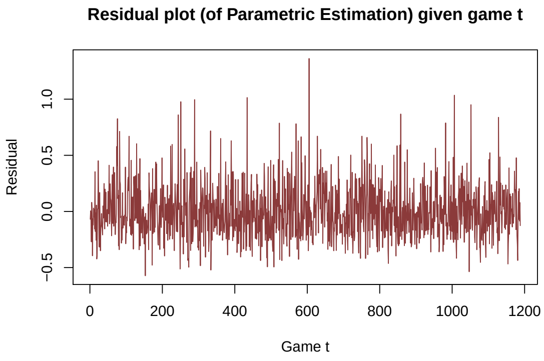
2.1: Parametric Signal Model

Let's first consider the following parametric model. We known seasonality exists (due to the up and down nature of "attack") and a bell-shaped trend also exists given how a player's deteriorating physical body and health affects how willing they are to attack the paint. We will capture this trend through a cubic function of game t, and given that the dataset includes only regular season games, we will assume that NBA is seasonal based off of 82 games, spanning usually over a six to seven month period. To account for Lebron's injuries, illnesses, or when he sits out games before playoffs, we will reduce the amount of games to 76, the average amount of games he plays for 16 seasons from 2003 to 2019 (including one outlier in 2019 when he played only 55 games due to an injury). As a result, we model the seasonality through a combination of sine and cosine functions with frequencies k/76, where k goes from 1 to 12. We choose K = 12 since there are is a period of 76/12, which is roughly six months for each season. (Stats taken from: https://www.basketball-reference.com/players/j/jamesle01.html)

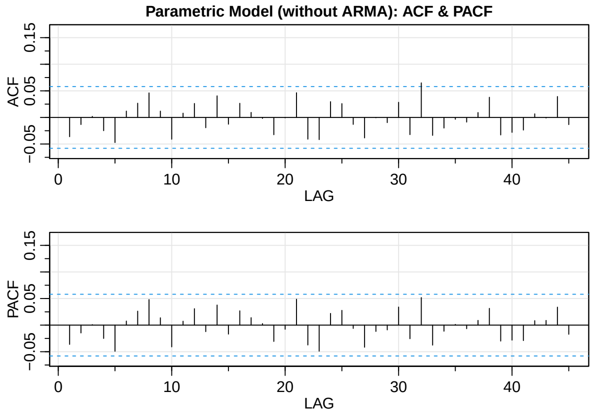
$$\begin{aligned}\text{Attack}_t &= \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{k=1}^{12}\left(\beta_{3+k} cos\left(\frac{2\pi kt}{76}\right) + \beta_{3+k} cos\left(\frac{2\pi kt}{76}\right)\right) \\ &+ \beta_{16} t cos\left(\frac{2\pi t}{76}\right) + \beta_{17} t^2 cos\left(\frac{2\pi t}{76}\right) + \beta_{18} t^3 cos\left(\frac{2\pi t}{76}\right) + X_t \qquad (1)\end{aligned}$$

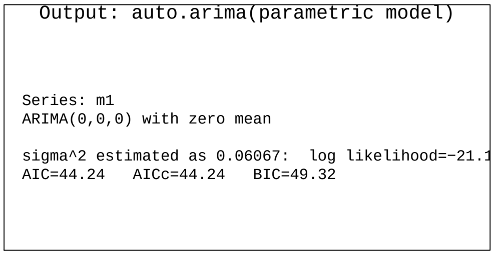
We will also include a game t interaction with our cubic, sinusoidal function as part of this model.





The residuals look fairly stationary! However, we did not exclude certain spikes or large dips (that may be outliers) using indicators. From this point, if we look at both the acf and pacf of the residuals from this model below, we see that the correlations look fairly like white noise (except for one correlation at lag = 32 that we will address in our first model). We can see that in the right panel below, auto.arima also suggests not to use any ARMA at all, which we will test if this holds true in the following models.





```
  Output: auto.arima(parametric model)



Series: m1
ARIMA(0,0,0) with zero mean

sigma^2 estimated as 0.06067:  log likelihood=-21.1
AIC=44.24   AICc=44.24   BIC=49.32
```

2.1.1: Parametric Signal Model with SMA(1)[32]

From the original ACF plot, we see that there is only one correlation with high magnitude at lag = 32. Under this assumption, we can then use a MA(Q=1) with lag 32, or SMA(1)[32] to fit this model. However, it's unclear what this seasonal lag of 32 represents under the context of Lebron's attack. If we look at the left panel of Figure 4, the SARIMA diagnostics, we see that the ACF and PACF look like white noise expect for one high correlation at lag 74 for both plots. Besides that, the high p-values for all lags from the Ljung-Box suggest that this model fits the data pretty well.

2.1.2: Parametric Model Signal with AR(3) x SMA(2)[37]

To address the additional high correlation at lag 74 as shown from the ACF from the previous model, let's try fitting a AR(3) x SMA(2)[37] model to see if we can have all correlations fit within the blue confidence intervals in both plots. Unfortunately, even after attempting to fit a really complex model to eliminate this large correlation, high correlation still exists as shown by the right panel of Figure 4.

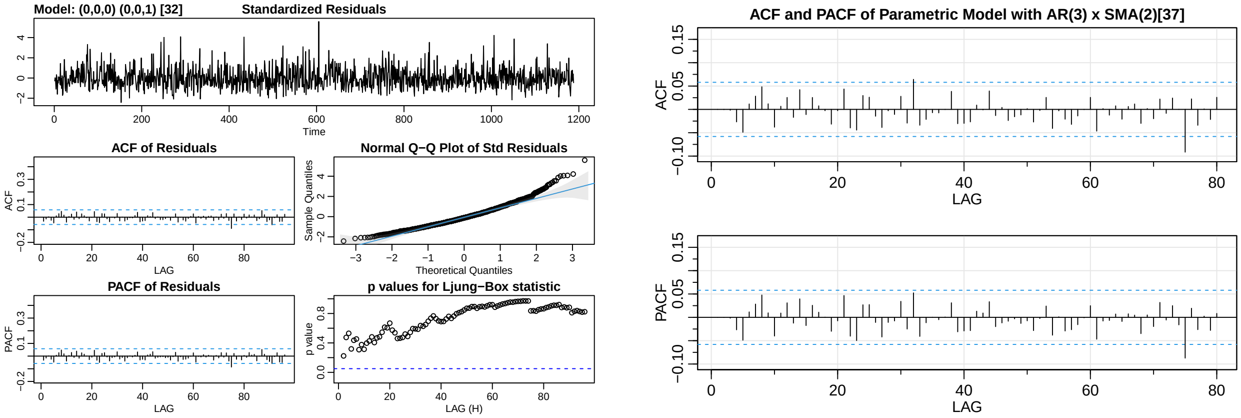
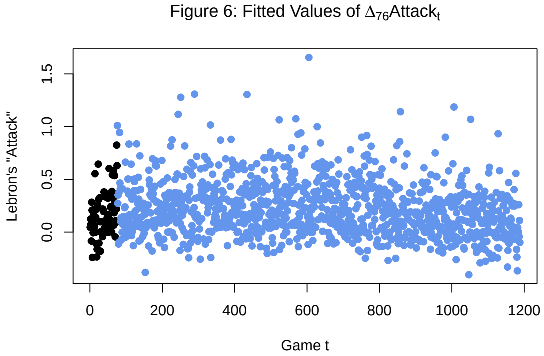
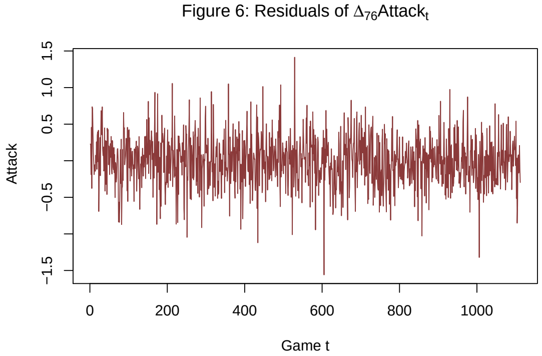


Figure 4: This panel provides performance analysis for the parametric model using SMA(1)[32] to eliminate noise (lags are provided by lags of games H). We see the model fits fairly well.

2.2: Seasonal Differencing (lag = 76)





We can also use seasonal differencing with lag = 76 (same logic from parametric model). The right panel of Figure 6 shows that the residuals of look fairly stationary (despite certain spikes that are hard to isolate).

If we look at ACF and PACF of the differencing model, it is unclear what type of ARMA model best fits the residuals. The lags with the largest values occurs at lags 21 and 23 for both the ACF and PACF. It's hard to determine whether there is remaining seasonality that can be eliminated or if it's due by chance. Since auto.arima suggests to not include an ARMA model, we will test the following models to see if the residuals are already white noise.

2.2.1: Seasonal Differencing with AR(P=1)[23]

We will first use AR(P=1)[23] to see if it eliminates the large correlations at lag 23 for both the ACF and PACF. As seen from Figure 7, the model fits fairly well: the Q-Q plot is normal throughout all quantiles, and the p-values are significant for almost all lags. The correlation values tend to stick within the blue-bands except for at lag 21, which was seen before.

2.2.2: Seasonal Differencing with ARMA(P=1, Q=1)[21]

To improve on the last model, let's now try eliminating the large correlation value at lag 21 instead of 23. By fitting an ARMA(1,3) we see that the ACF and PACF of the residuals look like fairly white noise. Like the last model, it fits extremely well, but still ignores the large correlation at lag 23. Still, the Ljung-Box statistics perform a tad worse than the last model, suggesting its complexity does not create extra benefits.

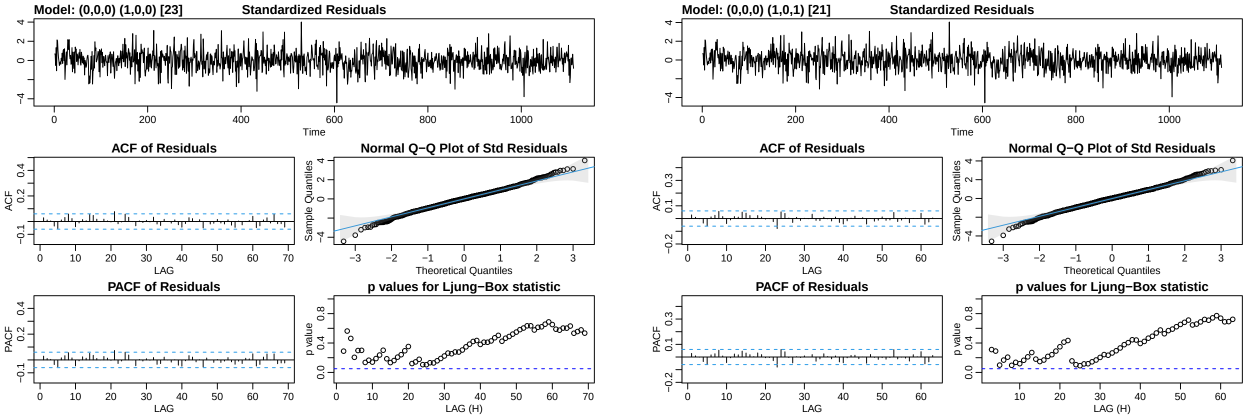


Figure 7: Left Panel is Performance Diagnostics for Seasonal Differencing with AR(P=1)[23] while the right panel eliminates the noise with ARMA(P=1, Q=1)[21].

3. Model Comparison and Selection

3.1: Model Comparison - Information Criterion

The parametric model outperforms the differencing model given the following Information Criterion:

Table 1: Information Criterion (AIC, AICc, BIC) for all 4 Models

Models	AIC	AICc	BIC
Parametric Signal Model with SMA(1)[32]	0.0363690	0.0363776	0.0491974
Parametric Model Signal with AR(3) x SMA(2)[37]	0.0449795	0.0450394	0.0749123
Seasonal Differencing (lag = 76) with AR(P=1)[23]	0.7153473	0.7153570	0.7288740
Seasonal Differencing (lag = 76) with ARMA(P=1, Q=1)[21]	0.7167969	0.7168163	0.7348325

3.2: Model Selection - Cross Validation These four models are no compared through time series cross validation. The testing sets for these four models rolls through the last 100 games of all games from 10-30-2003 to 11-25-2016, in 10 game segments. Thus, there will be 100 forecasting "attack" point for games across these ten windows. I have also used root-mean-square prediction error(RMSPE) as the metric to compare forecasting performances among these four models. The model with the lowest RMSPE will be chosen as the model to predict the ten next "attack" of Lebron James following his 3/2/2019 career game.

Tables 1 and 2 show that the parametric model with SMA(1)[32] does the best overall among the four models, both based by the comparison small RMSPE and small IC across the different models. Thus, we will use this model for future forecasting in Section 4, Results.

Table 2: Cross-validated out-of-sample root mean squared prediction error for the four models under consideration.

	RMSPE
Parametric Signal Model with SMA(1)[32]	0.1687400
Parametric Model Signal with AR(3) x SMA(2)[37]	0.1690846
Seasonal Differencing (lag = 76) with AR(P=1)[23]	0.3224692
Seasonal Differencing (lag = 76) with ARMA(P=1, Q=1)[21]	0.2436580

4. Results

To forecast the next ten “attack” of Lebron James following his 3/2/2019 career game, let “Attack” represent his willingness to attack the paint on game t with the additive noise term X_t, which is reinstated below as Equation 2. β coefficients can be found in Section 5.1 of the Appendix.

$$\begin{aligned}\text{Attack}_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{k=1}^{12} (\beta_{3+k} cos\left(\frac{2\pi kt}{76}\right) + \beta_{3+k} cos\left(\frac{2\pi kt}{76}\right)) \\ + \beta_{16} tcos\left(\frac{2\pi t}{76}\right) + \beta_{17} t^2 cos\left(\frac{2\pi t}{76}\right) + \beta_{18} t^3 cos\left(\frac{2\pi t}{76}\right) + X_t \end{aligned} \tag{2}$$

X_t is a stationary process defined by the equation below that models SMA(1)[32] or MA(Q=1)[32], where W_t is white noise with variance σ_W^2.

$$X_t = W_t + \theta_1 W_{t-32}$$

4.1: Results - Estimation of model parameters

Both estimates of the model parameters and noise estimates for our SMA[1][32] are given in Appendix 5.1. The intercept of the parametric model is initially at .148, which suggest that on average Lebron’s attack is fairly high.

4.2: Results - Prediction

Figure 8 shows the ten forecasted values of Lebron’s “attack” following 03-02-2019. We can see that the predictions are all positive, meaning that Lebron James is forecasted to “attack” the paint rather than shoot a three in the next ten games. However, it’s not extreme, as most of these values are less than .1. If we look at the light grey bands, Lebron can have his largest negative value of “attack”, meaning there is a possibility he will only shoot threes (however it ever so unlikely given this is only captured by the uncertainty of the SMA(1)[32] model).

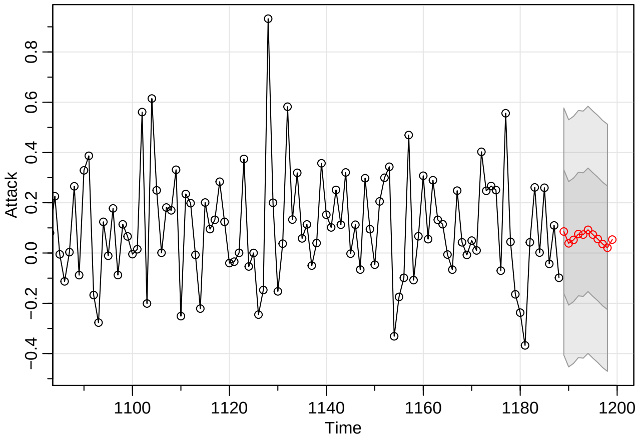


Figure 8: Forecasts of Lebron James Attack for his next 10 games following 11-25-2016. The black points are historical “attack” data of Lebron James throughout his 2003 to 2019 career. The red points are the forecasted “attack” points. The dark/light grey bands are standard error bands, representing 68% and 95% prediction intervals, respectively. The x-axis would ideally be dates of the games instead of t.

5.1: Appendix - Table of Parameter Estimates

Table 3: Estimates of the SMA(1)[32] model parameters in Equation 2, with their standard errors (SE).

	Estimate	SE
sma1	0.0651	0.0290
xmean	0.0000	0.0076

Table 4: Estimates of the forecasting model parameters in Equation 1, with their standard errors (SE).

Parameter	Estimate	SE	Coefficient Description [not required]
β_0	0.1421146	0.0291519	Intercept
β_1	0.0007143	0.0002125	$Time$ (Game t)
β_2	-0.000001	0.0000004	$Time^2$ (Game t)
β_3	0.000	0.000	$Time^3$ (Game t)
β_{4cos}	0.0206159	0.0411639	cos(k=1)
β_{4sin}	0.0142588	0.0102871	sin(k=1)
β_{5cos}	-0.008234	0.0102616	cos(k=2)
β_{5sin}	-0.0055267	0.0102556	sin(k=2)
β_{6cos}	-0.0024876	0.010257	cos(k=3)
β_{6sin}	-0.0072679	0.0102437	sin(k=3)
β_{7cos}	0.0075161	0.0102436	cos(k=4)
β_{7sin}	0.0051255	0.0102527	sin(k=4)
β_{8cos}	-0.0005145	0.0102449	cos(k=5)
β_{8sin}	0.0024589	0.0102491	sin(k=5)
β_{9cos}	0.0059104	0.0102522	cos(k=6)
β_{9sin}	-0.0029003	0.010241	sin(k=6)
β_{10cos}	0.0032655	0.0102492	cos(k=7)
β_{10sin}	-0.0077654	0.0102433	sin(k=7)
β_{11cos}	-0.0143242	0.0102443	cos(k=8)
β_{11sin}	0.0010516	0.0102478	sin(k=8)
β_{12cos}	-0.0104249	0.0102477	cos(k=9)
β_{12sin}	0.0036211	0.0102439	sin(k=9)
β_{13cos}	0.0009463	0.0102505	cos(k=10)
β_{13sin}	-0.0252824	0.0102408	sin(k=10)
β_{14cos}	0.0035723	0.0102464	cos(k=11)
β_{14sin}	-0.0102384	0.0102443	sin(k=11)
β_{15cos}	-0.0077414	0.0102431	cos(k=12)
β_{15sin}	-0.0053061	0.0102438	sin(k=12)
β_{16}	-0.0000604	0.0002985	Cosine × game t interaction
β_{17}	0.000	0.000001	Cosine × game t^2 interaction
β_{18}	0.000	0.000	Cosine × game t^3 interaction
σ_W^2	0.06041		Variance of White Noise