Contract Year Effect in the NBA*

Changhao Shi† Jonathan Liu‡ Terry II Culpepper§ Sean Choi¶

June 12, 2020

Abstract

We investigate whether being in a contract year has an effect on NBA player performance, with proxies such as win shares. Data is collected with custom scrapers on basketball websites and we evaluate the underlying theory of the data and how well they serve as proxies for performance. Existing heteroskedasticity, heterogeneity of players, and correlated observations drive our approach to tackle the issues individually with an array of methods. We use an IV regression controlling for possible endogenous effects. Subsequently, we employ a generalized estimating equation on a generalized weighted OLS to tackle the issue of correlated observations in addition to the issue of heteroskedasticity and heterogeneity. We also use a double LASSO inference method to tackle the curse of dimensionality problem that we encounter along an alternate approach to solving the correlated observations problem. The results we get show that the contract year has a mildly statistically significant effect on player performance, but not on most other measures of player performance. Furthermore, we compare the three approaches and demonstrate the usefulness of tackling underlying heteroskedastic, heterogeneous, and correlated observational trends in the data for inference and statistical power.

Keywords: Sports Economics, Athletes, Contracts, NBA

JEL Codes: Z20, Z22

*We would like to thank Professor Hortaçsu and Francisco del Villar Ortiz Mena for their invaluable insights.

†University of Chicago, The College, Economics
‡University of Chicago, The College, Economics
§University of Chicago, The College, Economics
¶University of Chicago, The College, Economics

Contents

1 Introduction

There has been numerous news articles highlighting the contract year phenomenon, a phenomenon where the athletes perform at a high level before their free agency than previous years, but once they sign we observe a drop in performance back to previous levels. Bismack Biyombo, Raef LaFrentz, and Luol Deng are just few of the many examples of this phenomenon. However, when watching the NBA, one finds that superstars such as Kobe Bryant, LeBron James, and Michael Jordan are not subject to the contract year phenomenon. This can be attributed to career concerns, as written by **?** and **?**. Such superstars play not simply to maximize their paychecks, but also to win championships, improve their legacies, and enter the conversation for greatest player of all time. However, many other players do not face the same career concerns, and hence may simply seek to maximize paychecks for minimal effort, forming the basis of the contract year phenomenon.

At its core, the contract year phenomenon is a principal-agent problem, complicated by incomplete information and time-variant incentives. In other words, a principal pays an agent, where the principal reaps utility as a function of the effort the agent puts in, while the agent reaps utility from the payment. The cost to the principal is the payment, while that to the agent is his/her effort level. Furthermore, principals are unable to observe the effort levels of agents directly, meaning that contracts offered by the principal may be incentive-incompatible. This can then lead to what **?** call "shirking" behavior, where an agent works with less effort than agreed upon in the contract. **?** further addresses the issue of moral hazard in a principal-agent interaction with imperfect information, and states that "[w]hen the same situation repeats itself over time, the effects of uncertainty tend to be reduced and dysfunctional behavior is more accurately revealed, thus alleviating the problem of moral hazard". However, a major issue with the contract year phenomenon is that principals, holding the expectation of high effort levels, offer long-term contracts, hence not allowing the "same situation" to play out numerous times, but only a few times before the player retires. This is intuitively suboptimal, and in fact, **?** finds that "contracts predicted by the

theory" have not been well-confirmed empirically. Thus, on the strategic level, agents on a contract year can engage in ex-ante opportunism, before undertaking ex post opportunism after signing the contract, terminology discussed in **?**. Namely, agents can first exert a high level of effort in one's contract year - ex ante opportunism - which enables them to sign a long-term contract. However, after signing, the players can now act with ex post opportunism, i.e. in this case putting in minimal effort. The degree to which this effort is minimal depends on an agent's dynamic problem. In other words, they can choose to exert no effort, hence guaranteeing that they never sign a new contract again. Alternatively, they could exert some positive level of effort, such that principals will be willing to give another contract once the agent reaches his next contract year. This is within the scope of the large body of IO literature on reputation.

That being said, sports-specific research has yielded conflicting results on opportunistic behavior. **?**, **?**, **?**, and **?** find evidence of the aforementioned opportunistic behavior. However, **?**, **?**, **?**, **?** attempt to explain the differing results by arguing that the test one uses to determine shirking affects whether or not one does find such behavior. In addition, **?** finds that multiple studies eliminates players who fail to acquire a new contract. This creates survivorship bias, as such players are likely to have systematically lower metrics of effort in contract years, and hence coefficients estimating the contract year phenomenon may be biased upwards.

Studies on the contract year phenomenon have used performance-based metrics, such as one's Player Efficiency Rating (PER, a statistic meant to measure a basketball player's per-minute performance while adjusting for the number of opportunities he has to perform), in order to estimate the impact of contract years on performance. We look at some such statistics, but instead of focusing on those, we would instead like to focus on other metrics that we believe are more closely related to effort level. This is because at the NBA level, simply exerting more effort may not be enough for many players who want to improve their performance: they face not only a talent barrier, but also a speed barrier during in-game

action, where the game moves quickly and is mostly continuous. This means that the effect of effort on performance could be completely outweighed by the need to think quickly and have extremely fast reactions. This is in contrast to a sport like baseball, where a player could exert more effort by practicing his batting more, and such effort can be demonstrated relatively clearly during in-game performance due to the slow, discrete nature of the game. We go more into detail of the exact statistics in the following section.

2 Literature Review

When watching the NBA, one finds that superstars such as Kobe Bryant, LeBron James, and Michael Jordan are not subject to the contract year phenomenon. This can be attributed to career concerns, as written by **?** and **?**. Such superstars play not simply to maximize their paychecks, but also to win championships, improve their legacies, and enter the conversation for greatest player of all time. However, many other players do not face the same career concerns, and hence may simply seek to maximize paychecks for minimal effort, forming the basis of the contract year phenomenon. At its core, the contract year phenomenon is a principal-agent problem, complicated by incomplete information and time-variant incentives. In other words, a principal pays an agent, where the principal reaps utility as a function of the effort the agent puts in, while the agent reaps utility from the payment. The cost to the principal is the payment, while that to the agent is his/her effort level. Furthermore, principals are unable to observe the effort levels of agents directly, meaning that contracts offered by the principal may be incentive-incompatible. This can then lead to what **?** call "shirking" behavior, where an agent works with less effort than agreed upon in the contract. **?** further addresses the issue of moral hazard in a principal-agent interaction with imperfect information, and states that "[w]hen the same situation repeats itself over time, the effects of uncertainty tend to be reduced and dysfunctional behavior is more accurately revealed, thus alleviating the problem of moral hazard". However, a major issue with the contract

year phenomenon is that principals, holding the expectation of high effort levels, offer long-term contracts, hence not allowing the "same situation" to play out numerous times, but only a few times before the player retires. This is intuitively suboptimal, and in fact, **?** finds that "contracts predicted by the theory" have not been well-confirmed empirically. Thus, on the strategic level, agents on a contract year can engage in ex-ante opportunism, before undertaking ex post opportunism after signing the contract, terminology discussed in **?**. Namely, agents can first exert a high level of effort in one's contract year - ex ante opportunism - which enables them to sign a long-term contract. However, after signing, the players can now act with ex post opportunism, i.e. in this case putting in minimal effort. The degree to which this effort is minimal depends on an agent's dynamic problem. In other words, they can choose to exert no effort, hence guaranteeing that they never sign a new contract again. Alternatively, they could exert some positive level of effort, such that principals will be willing to give another contract once the agent reaches his next contract year. This is within the scope of the large body of IO literature on reputation.

That being said, sports-specific research has yielded conflicting results on opportunistic behavior.

3 Data

3.1 Data collection

The NBA salaries of every player from the 2016-2017 season to the 2019-2020 season were collected from basketballreference.com. Note that basketballreference only displays data for the current season, so to get salary data for 2016-2017 to the 2018-2019 seasons, we utilized web.archive.org. Players were determined to be contract years if they did not have a salary for the following season. Furthermore, panel data of all advanced stats for each active NBA player was scraped from basketballreference.com for every regular season these players played in. Most of the advanced stats are measurements of a player's productivity on the

court. These include the age of the player, team the player played for, position the player played for, the number of games the player played for a particular season, average minutes played per game, player efficiency rating, true shooting percentage, three-point attempt rate, free-throw attempt rate, offensive rebound percentage, defensive rebound percentage, total rebound percentage, assist percentage, free throw attempt rate, offensive rebound percentage, defensive rebound percentage, assist percentage, steal percentage, block percentage, turnover percentage, usage percentage, offensive win shares, defensive win shares, win shares, win shares per 48 minutes, offensive box plus/minus, defensive box plus/minus, box plus/minus, and value over replacement player.

Another way to measure a player's productivity/effort on the court is with boxout data. In theory, the more a player boxes out opposing players in order to grab rebounds, the more productive that player is to the team. All boxout-related data was collected from stats.nba.com for every player from the 2016-2017 to the 2019-2020 regular season. These include the number of boxouts a player averages per game, number of boxouts on offense a player averages per game, number of boxouts on defense a player averages per game, average number of rebounds the team grabs as a result of a player boxing out, average number of rebounds the player grabs as a result of him boxing out, percentage of times a player boxes out on offensive, percentage of times a player boxes out on defense, the percentage of times the teams grabs a rebound when boxing out, and the percentage of times the player grabs a rebound when boxing out.

Another method of measuring a player's value on the court is by using data on how often the player touches the ball. Players who handle the ball more often are typically much more valuable to his team. Data for player touches was collected from stats.nba.com. These include the number of touches a player averages per game, the number of touches the player averages in the front court per game, the percentage of the time the player has the ball when he is on the court, average seconds the player has the ball when he touches it, the average number of dribbles the player takes whenever he touches the ball, the number of points the

player scores per touch, the average number of times a player touches the ball in the elbow part of the court, the average number of post-ups a player has per game, the average number of times a player touches the ball in the paint, points per elbow touch, points per post-up, and points per paint touch.

Note that when analyzing the data, the datasets for player salary, advanced player stats, boxouts, and touches were merged together by player name.

3.2 Dependent variables and proxies

To analyze the impact of a player playing during a contract year on his productivity on the court, we used the following functional form:

$$Production_{it} = a_1 * ContractYear_{it} + a_2 * Age_{it} + a_3 * Minutes_{it} + a_4 * Salary_{it}$$
$$+IndividualFixedEffects + SeasonalFixedEffects + PositionFixedEffects$$

The regressions are all weighted least squares by minutes played. This is because we do not want a superstar player who averages around 40 minutes per game to be weighted the same as another NBA player who barely gets any playing time. $ContractYear$ is a variable that takes on the value of 1 if the player is playing in a contract year; 0 otherwise. Age is the player's age on February 1st of the season. $Minutes$ is the average number of minutes a player averages per game. $Salary$ is the amount of money (in USD) a player is paid that season. Note that since only data from three seasons were analyzed, the effect of inflation will be marginal, at best. In addition to controlling for individual and season fixed effects, position fixed effects were also controlled for. This is because it is possible for some positions to be more valuable than others. For example, teams generally value big men/centers, so it is possible that these players can get longer contracts, so they will be less likely to be on a contract year. Furthermore, centers generally are more productive and/or valuable on the court (i.e. are responsible for grabbing many rebounds and scoring many points in the

paint).

We used 16 different variables as proxies of a player's productivity: usage rate, total win shares, win shares per 48 minutes, defensive win shares, offensive win shares, average total distance moved per game (in miles), average distance moved per game on defense (in miles), average distance moved per game on offense (in miles), average speed on the court (in miles per hour), average speed on defense (in miles per hour), average speed on offense (in miles per hour), average seconds per touch, average dribbles per touch, average box outs per game, average offensive box outs per game, and average defensive box outs per game.

A few of those proxies are rather weak. For example, every measurement involving a player's speed or distance moved on the court can be extremely noisy since a player is likely not hustling all of the time. For instance, on isolation plays where an offensive player seeks to attack a defensive player one-on-one, all other players on the court are likely standing around doing nothing. This is simply due to the nature of the play where involvement from the other players is not required, not because the other players are unwilling to put in any effort.

The strongest proxies are likely the win share variables: total win shares, win shares per 48 minutes, defensive win shares, and offensive win shares. These variables are commonly used in sports analytics to measure individual performance and/or how many wins can be attributed solely to a player (https://www.basketball-reference.com/about/ws.html). For example, a player can be on a terrible team (say only wins 20 out of 82 games), but can have 15 win shares. On the other hand, a player can be on a good team (say wins 60 out of 82 games) and also have 15 win shares. Both of those players are likely just as productive and/or valuable individually. However, in the former case, the player was unable to get much help from his teammates, but in the latter case, the player was surrounded by a strong supporting cast. Note that defensive win shares is a measure of the number of wins that can be attributed solely to a player's performance on defense, and offensive win shares is a measure of the number of wins that can be attributed solely to a player's performance on

offense.

4 Model and Analysis

4.1 Heteroskedasticity

In general, we cannot expect that the data for NBA players follows a homoskedastic trend. For example, the variance for win shares for players on a contract year can theoretically be different than the variance for offensive win shares for players not on a contract year. This is because a contract year can cause behavioral shifts that are reified in different ways: while a contract year could push one player to prioritize their win shares to impress future teams, another may feel ready to retire and gradually decrease the effort they put into each match, therefore affecting their performance. In a similar vein, we would also expect that, all else equal, players who have more time on the field to have a lower variance in performance than players who have less time on the field. An argument on the Law of Large Numbers applies here: all else equal (as in, players have the same ability and are given the same opportunities), players' performance converge to their expected performance as t, the time they spend on the field, goes to infinity. Heteroskedasticity in this sense may cause the OLS estimator to be unbiased but inconsistent, possibly leading to invalid inferences based on biased variances, so verifying whether the data contains underlying heteroskedasticity and tackling the problem will help with lowering the variance and therefore the precision of our results.

4.1.1 Breusch-Pagan test

We begin by looking at whether the data shows heteroskedastic trends. To do so, we use the Breusch-Pagan test from **?**. Our null hypothesis is

$$H_0 : \quad x_i \text{ exhibits homoskedastic trends}$$

for each regressor x_i we are using. To do so, we assume that the variance follows a functional form $h\left(z_i'\alpha\right)$. Then, our null hypothesis would be $\alpha = 0$, where 0 is the zero vector. Breusch and Pagan showed that this is equivalent to using the Lagrangian multiplier test statistic LM, where

$$\text{LM} = \hat{d}'\hat{\mathcal{P}}^{-1}\hat{d}$$

and

$$d = \frac{\partial l}{\partial \alpha}$$

and

$$\mathcal{P} = -\mathbb{E}\left(\frac{\partial^2 l}{\partial \alpha \partial \alpha'}\right)$$

and the hats are quantities that are evaluated with $\hat{\alpha}$ (**?**). In **R**, we use the Bresuch-Pagan test on our OLS regression of win shares on contract year and our controls, and obtain a p-value of 6.564×10^{-5}:

```
hetero.plot <- lm(formula = ws ~ contract_year + as.factor(name) + pos +
    as.factor(season) +salary_current, data = nba)


bptest(hetero.plot)
```

Table 1: Studentized Breusch-Pagan test

	Test Results		
Regressand	BP	df	p-value
Win Shares	501.44	386	6.564×10^{-5}
Usage Rate	580.35	386	5.266×10^{-10}
Boxouts	530.19	386	1.433×10^{-6}
Average Speed	548.27	386	9.901×10^{-8}
Average Dribbles per Touch	664.3	386	0

The Breusch-Pagan test allows us to reject the null hypothesis that the dataset exhibits homoskedastic trends very confidently. Due to this result, we would require a method to tackle the underlying heteroskedasticity of our data.

4.1.2 Weighted OLS

A weighted OLS gives us a partial solution to this issue. Because we have reason to believe that observations of players who are given little playtime are of a worse quality than observations of players who are given plenty of playtime, a possible weighting mechanism is to use the minutes they spend per match in order to prioritize the observations of players who spend the most amount of time on the field. There is, of course, an issue of whether players who are given little playtime are different from players who are given plenty of playtime. We will turn to tackle this issue in this next section.

We first note that, if our weights are are inversely proportional to the underlying variance of the data (in other words, W, our weight matrix, is the inverse of the variance-covariance matrix), then the weighted OLS estimator is the BLUE estimator (**?**). However, there is no reason to believe that the variance is the number of minutes in our data. Instead, it is reasonable to suppose that the variance correlates with the number of minutes. The potential disadvantage of a weighted least squares method arises when the weights are not precise relative to one another (**?**), but because minutes played is an external measurement that is precise and are not estimated from within our data, we can be confident that the issue of weights being imprecise relative to one another is limited. To formalize our argument, we state our assumption.

Assumption 1 *W, where W is the weight matrix consisting of the minutes played for a given player for each observation, approximates the inverse of the variance-covariance matrix of the regression.*

Our regression model is therefore

$$y_{it} = x'_{it}\beta + \epsilon_i$$

but we estimate β with the weighted least square estimator

$$\hat{\beta} = (X'WX)^{-1} X'WY$$

where X is the data matrix, Y is the vector of y_{it}'s, and W is the weight matrix. To be clear, W is the matrix that consists of the minutes played for a given player for each observation, such that $X'WX$ outputs the data matrix $\hat{X}'\hat{X}$ where $\hat{X}$ consists of, for each observation i and time period t, $w_{it}x_{it}$ This shows that we need to make the following assumption:

Assumption 2 *$X'WX$ is an invertible matrix.*

We would have to argue that we do not run into the problem of perfect multi-collinearity. This is a reasonable assumption - we do not believe, for example, that when we weight the players' age and income by the minutes played per game on the field, we obtain a linear relationship between the two. Concerns would arise if we have covariates such as total distance run and speed together. Note that these two covariates may not initially have a linear dependence, until, possibly, when we weight them by minutes played. However, we haven taken care to prevent possible multicollinearity by avoiding using covariates that have some relationship with minutes played, or may relate to another variable after being weighted by minutes played.

It is easy to conduct a weighted OLS on **R**. We simply call the function with an additional parameter, **weights**, using our choice of weight, minutes played:

```
reg.ws <- lm(formula = ws ~ contract_year + min + as.factor(name) + pos +
    as.factor(season)
+ salary_current, data = nba, weights = nba$min)
```

Finally, we use minutes played as a control once again, because the function of using minutes played as a weight is because we think it is a good proxy for the underlying heteroskedasticity of the data, but at the same time, minutes played as a covariate is also important as it will affect outcomes such as win shares. For instance, since win share is a measure of the amount of wins a player contributes to the team, it is natural to assume that the minutes played per match of a player will contribute to their win share. Therefore, we use minutes played as both a covariate and also the weight. We can show that the weighted OLS estimator remains unbiased if we assume that minutes played approximates the inverse of the variance-covariance matrix; however we cannot guarantee that the estimator still retains properties such as the BLUE property. We state this as a proposition:

Proposition 1 *The weighted OLS estimator for a regression with w_i as both the weight and a covariate is unbiased, but not necessarily BLUE.*

A trade-off exists here: it is clear that data from players who have low minutes played should not be weighted as heavily as data from players who have high minutes played. The solution of using a weighted OLS partially tackles the problem of the underlying heteroskedasticity, which prevents the OLS estimator from being BLUE. However, because of the necessity of using minutes played as a control as well, we are forced to use a weighted OLS estimator that is not BLUE. In essence, we are choosing between the lesser of two evils here. We believe that the weighted least squares approach is the superior one because the quality of the variance contributed by players with very low minutes played is unduly large, and the deviation from efficiency of the weighted OLS estimator with the weight as a control should be lower than the deviation from efficiency of the unweighted OLS estimator which puts undue weight on low quality data.

4.2 Heterogeneity

Player performance cannot be directly observed, and existing metrics such as distance covered on the field per minute may only be loosely correlated with player performance. For example, a player may specialize in 3-point throws and so roughly run the same amount of distance but attempt more 3-point shots per game. Another player may perform by running longer each game to gain tactical advantage on a field. It is also possible that the player's performance fails to translate into an observable metric; for example, he might put in effort into team-building exercises and coordinate much more on the field.

What is even worse is that the qualities of the players we have discussed that produces this heterogeneity cannot be observed. For example, their preferred playstyle, their cooperative attitude, and so on, are not something that we can observe in our data. These omitted variables, or rather, omitted fixed effects, that exhibit heterogeneity are therefore something we need to control for.

To control for fixed effects, we add indicator variables for players or teams as a covariate. This, along with the existing controls, gives us the regression model

$$y_{it} = \sum_{j=1}^{N} \alpha_j \mathbb{1}\left\{i = j\right\} + \mathbb{1}\left\{i_t \in \text{Contract}\right\}\beta + c_{it}'\gamma + u_{it}$$

where y_{it} is the outcome player statistics for player i in time period t, $\mathbb{1}\left\{i = j\right\}$ is the indicator variable for players, $\mathbb{1}\left\{i_t \in \text{Contract}\right\}$ is the indicator variable for whether a player is in a contract year in a given time period, and c_{it} are various controls, such as age and current salary. The identifying assumption is then that unobservable effects soaked up by the individual indicator variable that simultaneously affect the outcome player statistics and the explanatory variable and covariates are time-invariant. In other words,

$$\text{Cov}\left(x_{i1}, u_{it}\right) = \cdots = \text{Cov}\left(x_{iT}, u_{it}\right) = 0,$$

where x includes both the explanatory variable and the controls. This assumption is innocuous enough in this context. We don't expect unobserved characteristics (omitted variables) of an individual to change dramatically across years that also affect whether a given player is in a contract year. For example, a player may get married and be very happy that year. This would cause him to play better, therefore increasing his performance and we will see a change in his player statistics. However it is unlikely that it will affect whether he is in a given contract year, as that aspect largely depends on how many years ago the player signed the contract.

The implementation in **R** is straightforward. We simply add a dummy variable for the player's name.

```
reg.ws <- lm(formula = ws ~ contract_year + min + as.factor(name) + pos +
    as.factor(season)
+ salary_current, data = nba, weights = nba$min)
```

4.3 Correlated observations and within team (cluster) dependence

The performance put in by a particular player may also correlate with other players' performance. Since basketball is a team game, effort or performance put in by one player may synergize with effort or performance put in by other players. In general we expect observations to correlate strongly within-team.

Let us start by examining how effort could correlate between observations. We can look at this issue game-theoretically. Consider the standard public goods game, and let y_i be the energy of the player i devoted to the match, out of a total of 1 endowed unit of energy. The payoff function for player i can be thought of as

$$\Pi_i = (1 - y_i) + \alpha \sum_j y_j$$

where $0 < \alpha < 1$. The payoff here could represent the probability of winning a match. The sum of the players' effort correlates positively with the expected probability of winning a match. When $\alpha < 1$, the Nash equilibrium of this game is for all players to invest 0 energy into the match. However, when factoring in social norms, we expect players to fall into two categories. Based on empirical evidence, players will either put in more of their endowment when others put in more, or put in less when others put in more (**?**). This stylized game theoretical model and the empirical or experimental outcomes of human behavior shows that correlated observations with respect to effort is a significant issue that we need to address.

Furthermore, performance could also correlate between observations. For example, win share is a metric that estimates the contribution of a player to their team. While the win share of all players of a team does not exactly necessarily add up to the number of wins of a team (there is some more nuance than this), it does so approximately, which means that my increase in win share, all else equal, is your decrease in win share.

This makes estimating the contract year effect more difficult, as the outcome variable of an individual correlates with that of other individuals. Furthermore, this correlation is ex-ante unknown. While we may have some idea of the direction of the correlation, having some precise sense of its magnitude is difficult.

The first issue with correlated observations is that the OLS estimator is no longer efficient, and once again, loses the BLUE property. The second issue builds on the first, which is, once again, that an undue loss in statistical power can lead to bad inference results. The final issue is that we have very little intuition as to a possible functional form for the correlation between our outcomes.

4.3.1 Generalized estimating equation

? studied an extension of generalized linear models and introduced a class of estimating equations that give consistent estimates of regression parameters and their variance even under observations with unknown correlations. We use this method as one of our approaches

in tackling the issue of correlated observations. However, we will require an assumption.

Assumption 3 *Repeated observations for a subject are independent.*

The assumption is crucial for the generalized estimating equation method. We argue that this is a somewhat reasonable assumption. Since each season is independent of the previous season because the team performance in the previous season has no bearing on the initial conditions of the current season. For example, the winning team in 2017 does not get any innate advantage (in terms of competition) in 2018. While we obviously do not expect that player performance is completely independent of previous season performance (for example, the player could improve and learn from previous season mistakes), what is key is that the time dependence underlying our model is small enough and also not of key interest to our study of the contract year effect. Furthermore, we stress that, since we're attempting to correct for correlated observations that we believe is most prominent within-cluster, i.e., within-team. Hence, the subject in the assumption is emphatically the team, not the player. Hence, even if player performance is not completely independent, we have good reason to assume that team performance can be approximated as independent across observations. The estimating equation is then

$$U(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_{it}}{\partial \beta_k} V_i^{-1} (Y_i - \mu_i(\beta))$$

where μ_{it} is the model mean for team i at year t, V_i is the variance-covariance matrix, and β_k are parameters estimated by the GEE method (**?**). We then use the GEE method to re-estimate our weighted OLS model, this time using team as the **cluster** we're controlling for. The code implementation in **R** is straightforward:

```
nba <- nba %>%
arrange(team)
```

```
gee.ws <- geeglm(formula = ws ~ contract_year + min + as.factor(name) + pos +
    as.factor(season)
+ salary_current,
family = gaussian,
data = nba,
weights = nba$min,
id = team)
```

With this GEE method applied onto our weighted least squares model, we are finally able to correct for the three main challenges that we have discussed: heterogeneity, heteroskedasticity, and correlated observations.

4.4 IV Strategy

There is potential endogeneity bias between the outcome variable and the contract year dummy variable. Namely, not only could being on a contract year affect a player's outcome stats, but a player's outcome stats could affect whether a player is on a contract year. For example, a worse player may be given shorter-term contracts, since general managers would be less inclined to keep him on the roster in the long-run, meaning he is more likely to be in a contract year in general. As such, to solve this we would like an instrument that affects whether or not one is in a contract year in the current year, but has no direct effect on a player's outcome variables.

To do this, we consider the effect of LeBron James' free agency decisions. LeBron James is a superstar in contention for the greatest player of all time, and as such his free agency decisions can shape the landscape of the entire NBA. For example, between 2011-2018, LeBron made the NBA Finals every year, out of the Eastern Conference, meaning that other Eastern teams and players did not simply plan as they normally would have due to his dominance. While Western entities competed and planned almost as usual, those in the

East often chose to not compete, and instead "tank"; if a mid-tier team team decided that it could not compete against LeBron's team, it would often choose to "blow it up" and plan for a future without LeBron, as opposed to trying to upgrade and face LeBron in the playoffs (for an example of this, look at the 2015 Hawks, and how they traded/let their star players go in order to quickly rebuild).

As such, players and teams would plan contract years based on LeBron's contract years. LeBron telegraphs his contract years a few years in advance, in order to let his current team know that if they do not surround him with good players, he will leave and join another one. This telegraphing enables players around the league to pin their contract years to his. The significance is as follows: once LeBron finishes the last year of his contract, teams/players go into limbo in free agency, waiting for LeBron to announce which team he is joining. As soon as LeBron decides and the details of his contract are announced, other teams and players exit limbo and agree to contracts based on LeBron's choice. These contracts, though, are often similar to LeBron's, so that once LeBron enters a contract year again, the process can restart.

For a player's perspective, consider the following: a mid-tier player, currently in free agency, has only two offers: a low-paying, two-year offer from a contending team, and a high-paying, two-year offer from a bottom-tier team. He would like to win a championship; however, both teams are in the same conference as LeBron, so even if he joins the contending team, the chance is slim that they get to the NBA Finals. However, he knows that LeBron will enter a contract year in two years. As such, he chooses the high-paying offer on the bottom-tier team, then enters free agency at the same time as LeBron, where he can restart the process - and his quest for a championship - based on LeBron's free agency decision.

Thus, consider LeBron's latest free agency: the summer of 2018. It was known that LeBron was going through great turmoil and conflict on his team at the time, the Cleveland Cavaliers, and hence teams and players knew he was likely to leave. Based on the above argument, we will use an indicator for whether or not one's contract year is in the 2018

season as an instrumental variable.

Relevance: It clearly affects whether or not one is currently in a contract year.

Exclusion: It has no direct effect on a player's outcome stats, such as usage rate; it only affects outcomes through the contract year phenomenon.

Thus, this instrument provides an exogenous change in contract structure that affects contract year status.

4.5 Clustering data

4.5.1 Correlated observations

An alternate approach to solving the correlated observations problem we discussed earlier is to change the unit of analysis. Instead of approaching the question from the player level, we approach it from the team level. This is different from the approach used when we employed the method of generalized estimation equation. We reshape the entire dataset, such that each observation is a team, rather than each cluster being a team. This sidesteps the assumption that we require to argue that a team needs to have independent observations over time, but the reshaping of the dataset will reduce the number of observations from 720 with 71 variables to 90 observations with 65 variables. Hence, we run into the curse of dimensionality problem. The data reshaping code in **R** is as follows:

```
# drop duplicate columns
nba.LASSO <- nba %>%
select(-"player_y") %>%
select(-"season_1") %>%
select(-"age_2")

# filter teams that are TOT (meaning they were floating among teams)

nba.LASSO <- nba.LASSO %>%
```

```
filter(team != "TOT")

# produce weighted mean by team

nba.LASSO <- nba.LASSO %>%
group_by(team, season) %>%
mutate_all(funs(weighted.mean(.,min))) %>%
summarize_all(mean)

# drop columns that aren't meaningful when summed (produces NA)

nba.LASSO <- nba.LASSO %>%
select_if(~sum(!is.na(.)) > 0)
```

The curse of dimensionality problem, however, need not mean that we rapidly lose statistical power because our number of parameters in our regression model increases exponentially with the number of variables we have, since, for instance, we're not taking interaction terms with all variables. Our issue herein is a huge loss in statistical power and a corresponding huge increase in the variance of any regression we would run. That is to say, even a linear increase in our variance can be devastating when we only have 90 observations. Additionally, if we continue controlling for team fixed effects, this could get ugly very quickly.

4.5.2 Double LASSO inference selection

Our approach to tackling the curse of dimensionality problem is to employ the double LASSO-OLS and double LASSO-IV method described in **?** based on **?**. The double LASSO method serves as a tool for inference and inference selection by dropping spurious covariates that may not directly affect the data. However, in contrast with LASSO, we also attempt to ensure that we are not dropping covariates that have a small effect on the outcome variable

despite being correlated significantly with the remaining covariates. That is, we want to ensure that we do not inadvertently introduce an omitted variable bias when selecting inference variables. We note that other methods, such as relaxed LASSO proposed by **?** addresses some of the issues with only conducting a single LASSO (for example, a relaxed LASSO addresses the issue of computational complexity when the dimension of the dataset becomes very high, and also leads to a consistent variable selection under a prediction-optimal choice of the penalty parameters). However, the advantage of methods based on **?**'s is that it is heteroskedasticity robust, with possibly non-Gaussian disturbances (which we have implicitly assumed in our GEE model), and therefore addresses the underlying heteroskedasticity of our dataset. Furthermore, the coefficient of the final OLS after the double inference is also consistent. However, we need one assumption.

Assumption 4 *"Approximate sparsity". Only a relatively small number* $s = o(n)$ *of the regressors in the data are important for capturing accurately the main features of the regression function. In other words, the number of relevant regressors is much smaller than the sample size.*

The assumption follows from **?**, and is justified by carefully considering the covariates in our data. We don't have much reason to believe, for instance, that a player's dribbles per touch correlates so significantly with, say, the player's win share.

4.5.3 Pre-selection

Before we actually employ the double LASSO inference method, however, we need to take care not to use all our variables in the analysis, as several of them are correlated, or literally mean the same thing. For example, a player's offensive win shares is of course, highly correlated with their win shares. We have been hitherto circumspect and sidestepped the issue in our dataset by explicitly choosing our controls, but we now have to carefully look at all covariates in each regression and drop any covariates that serve as a proxy for effort or performance.

4.5.4 Double LASSO inference

We follow and outline the strategy in **?** for a double LASSO inference method we're going after. Our inference equation is

$$y_{it} = d_{it}\alpha + x'_{it}\beta + \epsilon_{it}$$

where d_{it} is the inference variable, that is, the percentage of players on a contract year, weighted by their minutes played on the field. x'_{it} are the usual covariates, and we are interested in the magnitude and direction of α.

5 Results

5.1 Generalized estimating equation method

5.1.1 Fixed effect weighted OLS

Before we discuss our results for the GEE method, let's first examine our results for the fixed effect weighted OLS. We run the fixed effect weighted OLS regression on 16 variables that we believe serve as a proxy for performance. While most of our results do not have statistical significance, such as in Table 2, we found that the contract year effect is statistically significant when we use win shares as the outcome variable, consistent with our discussion that win shares is the best metric for evaluating performance, in Table 3.

Table 2: Using Box Outs as the Dependent Variables

	Dependent variable:		
	Defensive Box Outs	Offensive Box Outs	Box Outs
	(1)	(2)	(3)
Contract Year	0.027	0.030	0.060
	(−0.107, 0.161)	(−0.036, 0.096)	(−0.113, 0.233)
	p = 0.697	p = 0.367	p = 0.500
Average Minutes Played	0.069	0.012	0.081
	(0.054, 0.084)	(0.005, 0.020)	(0.062, 0.101)
	p = 0.000***	p = 0.002***	p = 0.000***
Current Salary	0.000	0.000	0.000
	(−0.000, 0.00000)	(−0.000, 0.000)	(−0.000, 0.00000)
	p = 0.594	p = 0.083*	p = 0.312
Constant	0.602	−0.260	0.349
	(−0.245, 1.450)	(−0.677, 0.156)	(−0.743, 1.440)
	p = 0.165	p = 0.222	p = 0.532
Player Fixed Effects	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
Position Fixed Effects	Yes	Yes	Yes
Observations	820	820	820
R^2	0.913	0.838	0.912
Adjusted R^2	0.835	0.693	0.833
Residual Std. Error (df = 432)	2.840	1.400	3.660
F Statistic (df = 387; 432)	11.700***	5.770***	11.500***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: Using Win Shares as the Dependent Variables

	Dependent variable:			
	Win Shares	Offensive Win Shares	Defensive Win Shares	Win Shares per 48 Minutes
	(1)	(2)	(3)	(4)
Contract Year	0.401	0.282	0.102	0.003
	(0.083, 0.719)	(0.049, 0.515)	(−0.039, 0.243)	(−0.005, 0.011)
	p = 0.014**	p = 0.019**	p = 0.156	p = 0.445
Average Minutes Played	0.164	0.112	0.052	0.002
	(0.129, 0.200)	(0.086, 0.138)	(0.036, 0.068)	(0.001, 0.003)
	p = 0.000***	p = 0.000***	p = 0.000***	p = 0.0001***
Current Salary	−0.00000	−0.00000	−0.000	−0.000
	(−0.00000, 0.000)	(−0.00000, −0.000)	(−0.00000, 0.000)	(−0.000, 0.000)
	p = 0.052*	p = 0.035**	p = 0.331	p = 0.120
Constant	−0.098	−1.210	1.160	0.062
	(−2.110, 1.910)	(−2.690, 0.261)	(0.274, 2.050)	(0.012, 0.112)
	p = 0.924	p = 0.108	p = 0.011**	p = 0.016**
Player Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Position Fixed Effects	Yes	Yes	Yes	Yes
Observations	820	820	820	820
R^2	0.863	0.863	0.813	0.829
Adjusted R^2	0.741	0.740	0.646	0.676
Residual Std. Error (df = 432)	6.740	4.940	2.980	0.167
F Statistic (df = 387; 432)	7.060***	7.010***	4.860***	5.420***

Note: *p<0.1; **p<0.05; ***p<0.01

5.1.2 GEE Data

The GEE method gives us a smaller error by using the Wald statistic to compute the p-value and standard error. As we can see in Table 4, we are therefore able to bound the 95% confidence interval for our coefficients further. Hence, we can see that the GEE method improves statistical significance for datasets with underlying correlated observations and improves statistical power, serving as a method to limit the presence of correlated observations from exaggerating the lack of statistical power of our results. We can see these comparisons in Tables 5, 6, and 7.

Table 4: Using Win Shares as the Dependent Variables

	Dependent variable:		
	Win Shares	Offensive Win Shares	Defensive Win Shares
	(1)	(2)	(3)
Contract Year	0.401	0.282	0.102
	(0.126, 0.676)	(0.0578, 0.507)	(−0.0190, 0.223)
	p = 0.0043***	p = 0.0137**	p = 0.09829*
Average Minutes Played	0.164	0.112	0.0522
	(0.138, 0.191)	(0.0926, 0.131)	(0.0369, 0.0675)
	p = 0.000***	p = 0.000***	p = 0.000***
Current Salary	0.000	0.000	0.000
	(−0.000, 0.000)	(−0.000, 0.000)	(−0.000, 0.00000)
	p = 0.163	p = 0.0801*	p = 0.524
Constant	-0.0984	−1.21	1.16
	(−1.56, 1.36)	(−2.11, −0.316)	(0.344, 1.98)
	p = 0.895	p = 0.0081	p = 0.005***
Player Fixed Effects	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
Position Fixed Effects	Yes	Yes	Yes
Number of Clusters	31	31	31
Maximum Cluster Size	114	114	114
Degrees of Freedom	820	820	820
Residual Degrees of Freedom	432	432	432
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: Comparison of regression results with or without GEE for win shares

Coefficient/GEE	Relevant Statistics from weighted least squares		
	Coefficient	Confidence Interval (95)	p-value
Regular WOLS, Contract Year	0.401	(0.083, 0.719)	0.014**
WOLS with GEE, Contract Year	0.401	(0.126, 0.676)	0.0043***
Regular WOLS, Minutes Played	0.164	(0.129, 0.200)	0.000***
WOLS with GEE, Minutes Played	0.164	(0.138, 0.191)	0.000***

Table 6: Comparison of regression results with or without GEE for offensive win shares

Coefficient/GEE	Relevant Statistics from weighted least squares		
	Coefficient	Confidence Interval (95)	p-value
Regular WOLS, Contract Year	0.282	(0.049, 0.515)	0.019**
WOLS with GEE, Contract Year	0.282	(0.0578, 0.507)	0.0137**
Regular WOLS, Minutes Played	0.112	(0.086, 0.138)	0.000***
WOLS with GEE, Minutes Played	0.112	(0.0926, 0.131)	0.000***

Table 7: Comparison of regression results with or without GEE for defensive win shares

Coefficient/GEE	Relevant Statistics from weighted least squares		
	Coefficient	Confidence Interval (95)	p-value
Regular WOLS, Contract Year	0.102	(−0.039, 0.243)	0.156
WOLS with GEE, Contract Year	0.102	(−0.0190, 0.223)	0.09829*
Regular WOLS, Minutes Played	0.0522	(0.036, 0.068)	0.000***
WOLS with GEE, Minutes Played	0.0522	(0.0369, 0.0675)	0.000***

Additional tables and plots from the weighted OLS regression are presented in the appendix (A).

5.2 Instrumental variable method

5.3 Clustered data double LASSO inference method

The clustered data method ultimately did not lead to statistically significant results, despite ultimately only controlling on three covariates: contract year, age, and minutes played. That the number of observations was severely limited, being at 90, was a huge reason why we could not get statistically significant results.

Table 8: Double LASSO inference method with win share as the outcome variable

	Relevant Statistics from double LASSO		
Variable	Coefficient	Confidence Interval (95)	p-value
Contract Year	0.990	(−0.902, 2.88)	0.31
Note:		Selected covariates: Age, Minutes Played	

Table 9: Double LASSO inference method with offensive win share as the outcome variable

	Relevant Statistics from double LASSO		
Variable	Coefficient	Confidence Interval (95)	p-value
Contract Year	0.865	(−0.34, 2.07)	0.16
Note:		Selected covariates: Age, Minutes Played	

Table 10: Double LASSO inference method with defensive win share as the outcome variable

	Relevant Statistics from double LASSO		
Variable	Coefficient	Confidence Interval (95)	p-value
Contract Year	0.0227	(−0.956, 1)	0.96
Note:		Selected covariates: None	

6 Discussions

7 Conclusion

A Tables

We present all tables for our regression results in this section.

Table 11: Using Speed Metrics as the Dependent Variable

	Dependent variable:		
	Average Speed	Offensive Speed	Defensive Speed
	(1)	(2)	(3)
Contract Year	−0.006	−0.009	−0.009
	(−0.022, 0.011)	(−0.038, 0.021)	(−0.031, 0.014)
	p = 0.502	p = 0.555	p = 0.444
Average Minutes Played	−0.005	−0.005	−0.007
	(−0.007, −0.003)	(−0.009, −0.002)	(−0.010, −0.005)
	p = 0.00000***	p = 0.002***	p = 0.00000***
Current Salary	0.000	0.000	−0.000
	(−0.000, 0.000)	(−0.000, 0.000)	(−0.000, −0.000)
	p = 0.716	p = 0.506	p = 0.006***
Constant	4.590	5.070	4.220
	(4.490, 4.700)	(4.890, 5.260)	(4.080, 4.360)
	p = 0.000***	p = 0.000***	p = 0.000***
Player Fixed Effects	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
Position Fixed Effects	Yes	Yes	Yes
Observations	820	820	820
R^2	0.936	0.872	0.898
Adjusted R^2	0.880	0.758	0.806
Residual Std. Error (df = 432)	0.351	0.625	0.472
F Statistic (df = 387; 432)	16.500***	7.630***	9.780***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 12: Using Distance as the Dependent Variable

	Dependent variable:		
	Distance: Defensive	Distance: Offensive	Distance
	(1)	(2)	(3)
Contract Year	−0.004	−0.001	−24.800
	(−0.009, 0.002)	(−0.008, 0.005)	(−72.200, 22.500)
	p = 0.183	p = 0.722	p = 0.305
Average Minutes Played	0.033	0.039	379.000
	(0.032, 0.034)	(0.038, 0.039)	(373.000, 384.000)
	p = 0.000***	p = 0.000***	p = 0.000***
Current Salary	−0.000	0.000	0.00000
	(−0.000, 0.000)	(0.000, 0.000)	(−0.00000, 0.00001)
	p = 0.233	p = 0.002***	p = 0.095*
Constant	−0.0003	0.057	291.000
	(−0.036, 0.035)	(0.015, 0.099)	(−7.550, 590.000)
	p = 0.988	p = 0.009***	p = 0.057*
Player Fixed Effects	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
Position Fixed Effects	Yes	Yes	Yes
Observations	820	820	820
R^2	0.994	0.994	0.997
Adjusted R^2	0.989	0.989	0.994
Residual Std. Error (df = 432)	0.119	0.142	1,003.000
F Statistic (df = 387; 432)	194.000***	197.000***	366.000***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 13: Using Dribbles, Touches, and Usage Rate as the Dependent Variable

	Dependent variable:		
	Average Seconds per Dribble	Average Seconds per Touch	Usage Rate
	(1)	(2)	(3)
Contract Year	−0.025	−0.017	0.262
	(−0.100, 0.051)	(−0.086, 0.052)	(−0.229, 0.754)
	p = 0.528	p = 0.624	p = 0.297
Average Minutes Played	0.009	0.009	0.065
	(0.0004, 0.017)	(0.001, 0.017)	(0.010, 0.120)
	p = 0.041**	p = 0.025**	p = 0.022**
Current Salary	0.000	0.000	0.00000
	(0.000, 0.00000)	(−0.000, 0.000)	(−0.000, 0.00000)
	p = 0.049**	p = 0.078*	p = 0.062*
Constant	1.320	2.210	19.300
	(0.839, 1.800)	(1.780, 2.650)	(16.200, 22.400)
	p = 0.00000***	p = 0.000***	p = 0.000***
Player Fixed Effects	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
Position Fixed Effects	Yes	Yes	Yes
Observations	820	820	820
R^2	0.973	0.969	0.915
Adjusted R^2	0.949	0.941	0.839
Residual Std. Error (df = 432)	1.610	1.460	10.400
F Statistic (df = 387; 432)	40.100***	34.600***	12.000***
Note:			*p<0.1; **p<0.05; ***p<0.01

B Figures

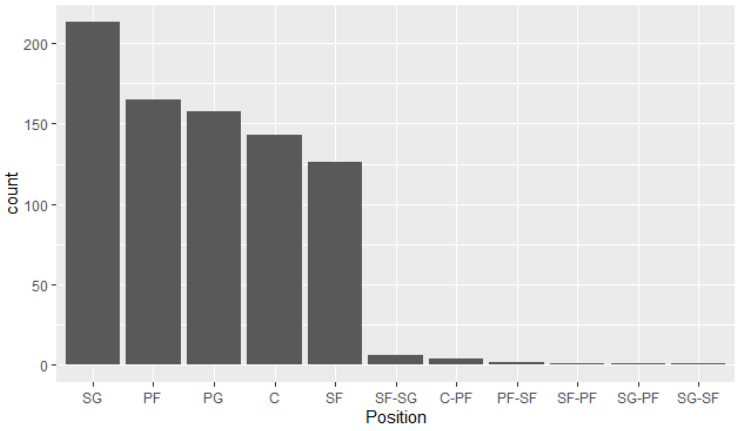

Figure 1: Position

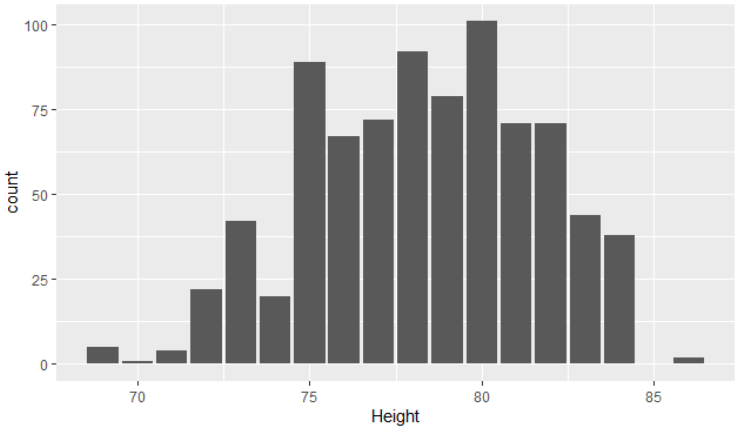

Figure 2: Height

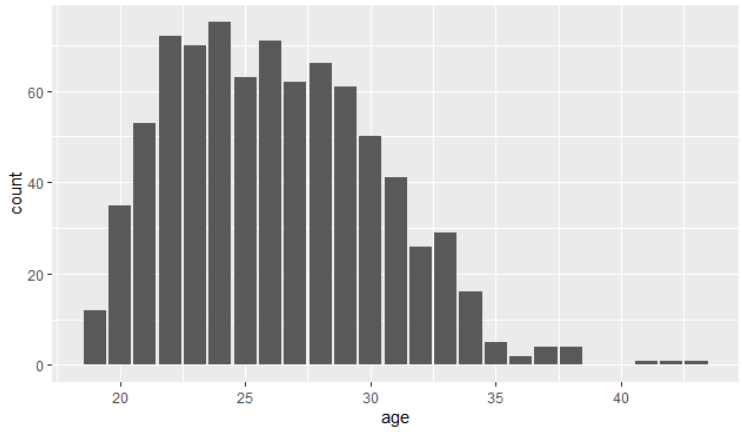

Figure 3: Age

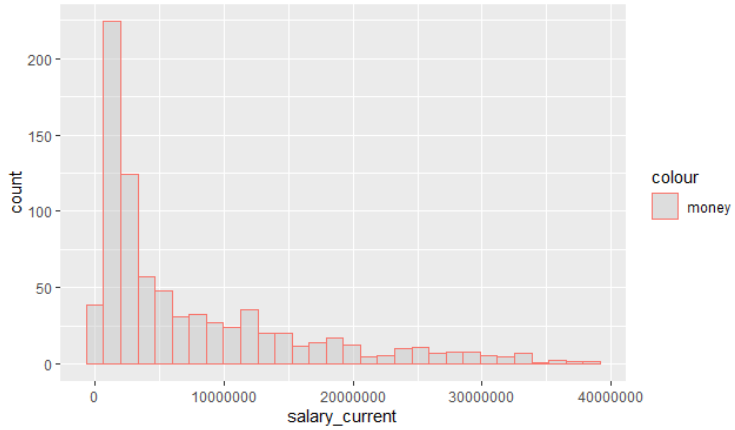

Figure 4: Salary

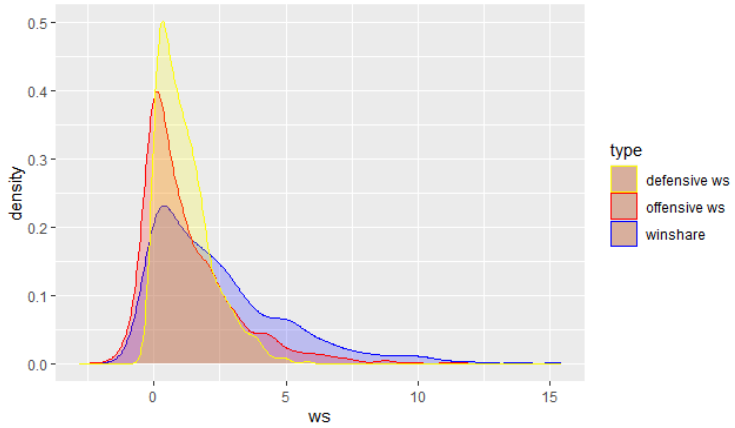

Figure 5: Win Shares

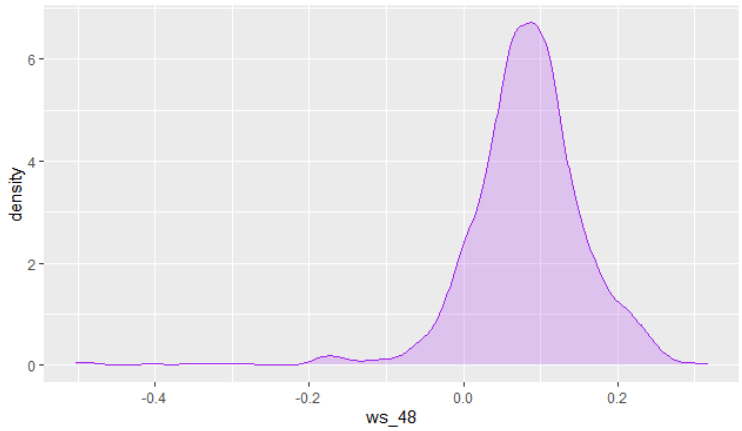

Figure 6: Win Shares at 48 min.

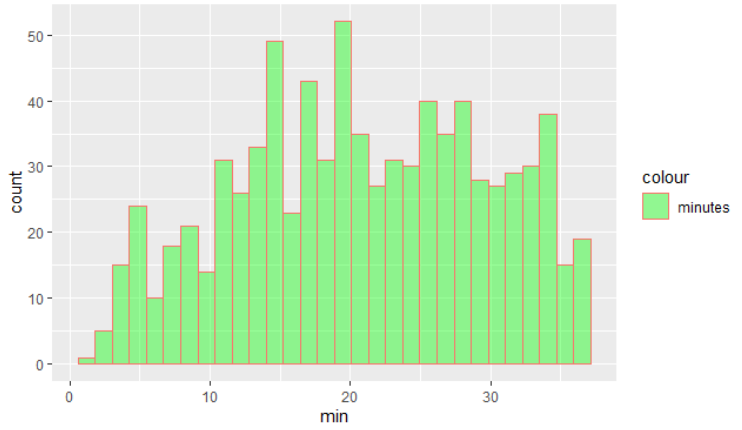

Figure 7: Minutes

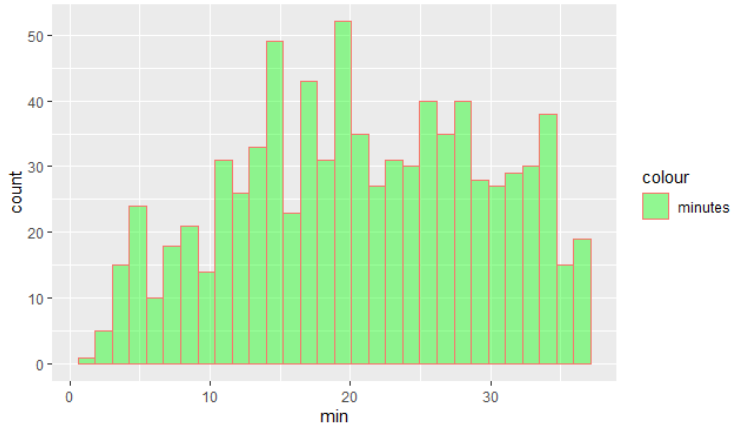

Figure 8: Minutes

C Code