The first step in compiling my dataset was accruing salary information for all arbitration eligible players and all free agents in my period of interest. For the arbitration data, my source was www.bizofbaseball.com, a site examining the financial aspects of baseball compiled by Maury Brown, the president of the Business of Sports Network as well as a writer for Forbes. On this site are charts with relevant information for arbitration-eligible players in each year dating back to the offseason that preceded the 2008 season. I created a single spreadsheet with all of this information, which gave me my initial arbitration data, which was then used to merge with the playing statistics. Variables that came from this source included first, last, team, position, servicetime, filed, winningbid, playerbid, clubbid as well as the prior and determined salary for the player as well as future years included if the player signed a multiple year extension to avoid arbitration, yearssal, total ($). To this data, I noted which players were the special case of players that became eligible for arbitration prior to three years of free agency (Super Twos); I also had Stata calculate the difference between the player and club bid as well as the midpoint, the percent change in a player's salary from the year prior, the difference from the midpoint to the ultimately determined salary, the dollar amount of change in salary, and the aav for the contract (only differed from the determined salary for extensions). For year, I made it the year for which the salary was determined (year after offseason of negotiation).

The basis for Free Agent data came from ESPN.com's Free Agent Tracker. For each year, the tracker provided a player's position, team and prior team, length of salary and total guaranteed value. In addition, it included whether the player was a type A, type B or qualifying offer free agent, which carries implications on compensation needed to give up that could affect a team's decision to sign a player. This was only a portion of the information I needed so I gathered the rest from the player Baseball-Reference.com and Baseballprospectus.com. This included the player's service time at signing as well as the yearly breakdowns (including previous salary) for multiyear contracts. Baseball Prospectus also indicated which players had not been tendered a contract despite being arbitration-eligible (these players were likely to see a salary decrease as their previous team did not see them as worthy of a likely raise through the arbitration process). I also made sure to note the players who came to MLB after playing professionally in international leagues as their service time numbers would be misleading. For example, Hiroki Kuroda was 33 during the first year he pitched in the Majors after pitching 11 years in Japan. Equating his first year of service time (and each year thereafter) with a 23-year-old rookie would clearly have deceiving results. For these as well, I had Stata calculate the percent change and absolute change in salary from the prior year. I set a fa variable equal to one to all players to distinguish them from the arbitration eligible players in the final dataset. Finally, I made sure the year corresponded to the year following the signed contract.

Finally, I needed to create spreadsheets with the playing statistics for each player. For this, I utilized the Baseball-Reference.com Play Index. This database allowed me to query each year of interest and get a spreadsheet of relevant statistics for each player in a given season. This also gave me the player's age during that season, which I could include in the dataset. Further, I could query aggregate statistics for multiple years to obtain three-year totals. Statistics that I used were bWAR (baseball reference wins above replacement), games, plate appearances, at bats, runs, hits, doubles, triples, home runs, runs batted in, walks, intentional walks, strikeouts, hit by pitches, bunts, sacrifice flies, times grounded into a double play, stolen bases, times caught stealing, batting average, on base percentage, slugging percentage, and On Base plus Slugging. I selected these statistics because they cover almost all of the traditional counting stats that a

player accrues throughout a season as well as batting average, the traditional “rate stat”, and more modern statistics like on base percentage and slugging percentage. Obviously, WAR is the most modern statistic and gives me a more or less comprehensive evaluation of a player’s worth in the interested years. Once I had these statistics, I simply merged the lists of free agents and arbitration eligible with the player’s numbers for the prior season from the Play Index spreadsheets and my data was complete