Using Regression Analysis to Predict Future Offensive Performance of Major League Baseball Players

By Kaven Kolb



Introduction:



Since the publication of the highly-influential book “Moneyball” in 2003, Major League Baseball decision-making has been progressing towards utilizing a much more advanced statistical-based analysis of player performance. Traditional statistics such as batting average, RBI’s, win/loss record, and earned run average have largely been replaced with newer statistics such as wOBA (weighted on-base average) and FIP (fielding-independent pitching). These stats place an emphasis on context-neutrality in an effort to give a player credit only for what he is directly responsible for.

In 2014, Major League Baseball began testing a new tracking system, called Statcast, to improve the collection and analysis of baseball statistics. Statcast is comprised of two tracking systems, a Trackman Doppler radar and high definition Chyron Hego cameras. Since its full implementation in 2015, Statcast has been used in tracking pitch information, batted-ball outcomes, as well as statistical quantification of basic player skills such as sprint speed, throwing-arm strength, etc. The mass collection of data such as exit-velocity and launch angle of batted balls has allowed for an additional level of statistical analysis. For example, on-base percentage has traditionally been calculated by dividing the number of times a player gets on base by the number of hitting opportunities he has had. Information collected by Statcast now allows us to predict the probability of a player getting on base based on how hard and how far he hits a ball. Essentially, this is an effort to remove any of the “luck” associated with batted-ball outcomes. By averaging these probabilities over many plate appearances, you can theoretically get a better idea of how good a player’s performance has actually been, as well as how good his performance will be in the future, than you can by simply looking at the outcomes of all of his plate appearances.

In this project, a dataset was generated of 176 players who received at least 400 plate appearances in both 2016 and 2017. A combination of statistics (taken from Fangraphs.com) and data collected by Statcast were used in this analysis, which used 2016 statistics to predict the total offensive value of a player in 2017. The goal was to use this dataset to reliably predict the wOBA of a player in 2017, based on statistics from his 2016 season, as well as creating a model that could be used for the same purpose in future seasons. The potential uses of this model are wide-ranging, including player evaluation and roster building by MLB teams, as well as potential legal uses in labor agreements between MLB teams and the player’s union.

Methods:

Data:

The full model had 28 predictor variables for the response (wOBA2)



|  |  |
| --- | --- |
| wOBA2 | Statistic designed to measure a player's overall offensive contribution. Weights each offensive outcome proportional to expected run value, then divides by plate appearances.  The number is then scaled to the league average on-base percentage (usually about .320, but varies from year-to-year. |
| Age2 | A player's age in years, at the beginning of 2017. |
| Speed1 | Maximum 2016 sprint speed in feet/s |
| GB1 | Number of batted balls classified as a ground ball divided by total number of batted balls |
| FB1 | Number of batted balls classified as a fly ball divided by total number of batted balls |
| GBFB1 | Ratio of ground balls to fly balls |
| IFFB1 | Percentage of flyballs that were hit to an infielder (ie: flyballs that weren't hit far) |
| Pull1 | Number of batted balls hit to a batter's pull side divided by total number of batted balls (Right-handed batter hits ball to Left Field, or vice-versa) |
| Oppo1 | Number of batted balls hit to the opposite of a batter's pull side divided by total number of batted balls (Right-handed batter hits ball to Right Field, or vice-versa) |
| Soft1 | Number of batted balls classified as "Soft Contact" divided by total numbers of batted balls (as defined by Baseball Info Solutions and supplied by Fangraphs.com) |
| Hard1 | Number of batted balls classified as "Hard Contact" divided by total numbers of batted balls (as defined by Baseball Info Solutions and supplied by Fangraphs.com) |
| OSwing1 | Percentage of pitches outside of the strike zone that a batter swings at |
| ZSwing1 | Percentage of pitches inside of the strike zone that a batter swings at |
| disc1 | Crude estimator of overall plate discipline, calculated as ZSwing1-OSwing1 |
| OContact1 | Percentage of swings a batter takes at pitches outside of the strike zone that result in the batter hitting the ball |
| ZContact1 | Percentage of swings a batter takes at pitches inside of the strike zone that result in the batter hitting the ball |
| FStrike1 | Percentage of plate appearances a batter has where the first pitch is ruled a strike |
| SwStr1 | Percentage of swings a batter takes that do not result in contact |
| VeloMax1 | Maximum exit velocity (in mph) of all batted balls a batter hit during the season |
| VeloAvg1 | Average exit velocity (in mph) of all batted balls a batter hit during the season |
| FBLDVelo1 | Average exit velocity (in mph) of all batted balls classified as a fly ball or a line drive that a batter hit during the season |
| GBVelo1 | Average exit velocity (in mph) of all batted balls classified as a ground ball that a batter hit during the season |
| DistMax1 | Maximum projected distance (in feet) traveled in the air of any ball a batter hit during the season |
| DistAvg1 | Average projected distance (in feet) traveled in the air of any ball a batter hit during the season |
| AvgHRDist1 | Average projected distance (in feet) traveled in the air of all home runs a batter hit during the season |
| ninemph1 | Percentage of batted balls a batter hit that had an exit velocity of at least 95 mph |
| BrlsBBE1 | Number of batted balls classified as a barrel divided by total number of batted balls |
| BrlsPA1 | Number of batted balls classified as a barrel divided by total number of plate appearances |
| langle1 | Average launch angle (in degrees) of all batted balls a player hit during a season (0 refers to ball hit perfectly horizontally) |

\*Barrels are defined by Statcast as any batted ball with a minimum expected batting average of .500 and a minimum expected slugging percentage of 1.500.

\*Any variable name followed by a 1 refers to data collected in 2016, 2 refers to data collected in 2017

\*Speed1, VeloMax1, VeloAvg1, FBLDVelo1, GBVelo1, DistMax1, DistAvg1, AvgHRDist1, ninemph1, BrlsBBE1, BrlsPA1, and langle1 were all recorded by Statcast and taken from baseballsavant.com, all other data is from Fangraphs.com

The total data set had 176 players, each with a minimum of 400 plate appearances in both years. This minimum requirement was set to make sure the data was taken only from seasons with large enough sample sizes to be statistically meaningful, but small enough where a model created from the dataset wouldn’t be too heavily influenced by extreme examples. All variable in this model are continuous.



Exploratory Analysis:



First off, relationships between the potential predictor variables and the response variable wOBA2 were examined. Scatterplots were analyzed to see if any of the relationships between the response and predictor variables were non-linear and thus needed to be transformed. None of the predictor variables had an obvious, non-linear relationship with wOBA2, so none were transformed. A histogram of each variable was also examined, all of which showed no obvious issues.

Nothing was done to address potential multicollinearity issues at this point, due to the large number of variables, but it can be inferred that certain predictor variable will have high correlation with each other. For example, VeloMax1 and VeloAvg1 are likely to have some significant overlap, thus it will probably not be necessary or useful to include both variables in the same model.

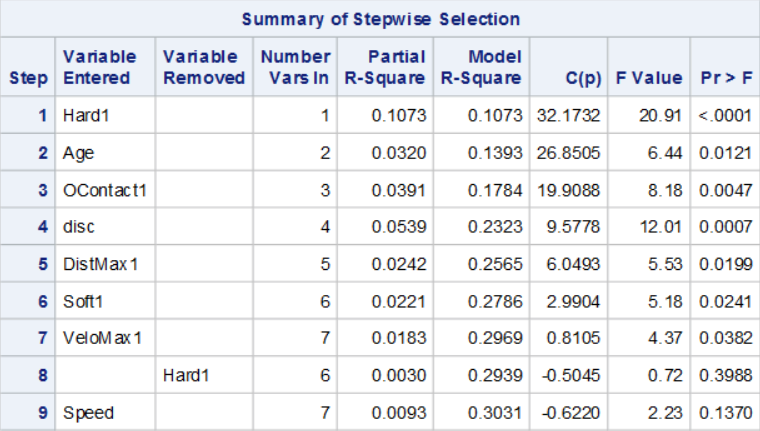


Model Building:



A Forward Stepwise Regression model building method was used to select the best model. This was chosen over a best or all subsets model primarily because of the data used. There were many predictor variables, none of which had an overwhelmingly strong correlation with the response variable. Due to the inherent statistical noise in many of these variables, a simpler model is preferable to avoid overfitting. Therefore, an all subsets model would have had many unnecessary predictors included. Selection using Mallow’s Cp Criterion or R-Squared/Adjusted R-Squared would have also produced complicated models that had several variables that did little to reliably improve the fit of the model. A stepwise model selection allowed for only variables that were significant at an alpha=0.15 level.





The stepwise selection method suggested a model with 7 variables. These variables were Age, OContact1, disc, DistMax1, Soft1, VeloMax1, and Speed. However, I eventually decided to exclude Speed from the model, since it was just barely significant at the chosen level, and it contributed minimally to the R-Square value of the total model. My concern was that Speed may no longer be significant if this analysis were repeated with a different dataset, and is possibly only in this model as an effect of overfitting.

Diagnostics:

Diagnostics for both the model with and without Speed were tested. Figures 1A-1C show diagnostic measures for the model without Speed, while 2A-2C show diagnostics for the model with speed. In both, the residual plot of the predicted values appears to be approximately normal with constant variance. The histogram shows a potential slight skew to the right. The leverage plot shows a few data points have high leverage or a large studentized residual, but neither have both. The qq plot also appears to mostly follow the straight line, showing the residuals are mostly normally distributed. Since the histogram is a bit uneven and also appeared to be a little skewed, I tested the residuals for normality. The results of this test for the model without Speed are given in figure 3. These tests test the null hypothesis that the data come from an independent, normally distributed population. They were tested against an alpha level of 0.05. In all four tests, there is not enough evidence to reject the null, confirming that the model meets the assumptions.

The last test I did was for multicollinearity. This is given in Figure 4. The table shows values for tolerance and variance inflation. None of the parameters included in this model show signs of multicollinearity.



Inferential Methods:



An ANOVA was used to determine if the fit for this model was appropriate and if the values for each parameter were significant. The F-test of the ANOVA was with alpha=0.05 and a null hypothesis of “there is no correlation between wOBA2 and the predictor variables. The parameters are tested against a null hypothesis of “the parameter is not statistically significant” and was tested against a maximum alpha of 0.15, though speed was excluded from the model despite being significant at that level.

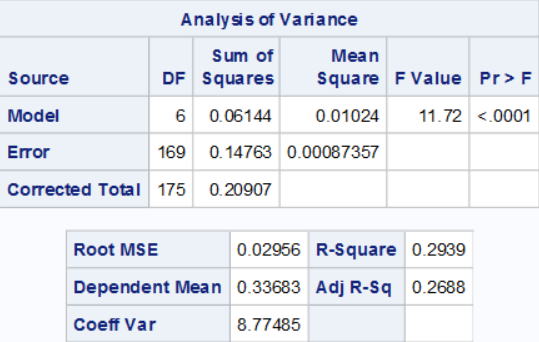


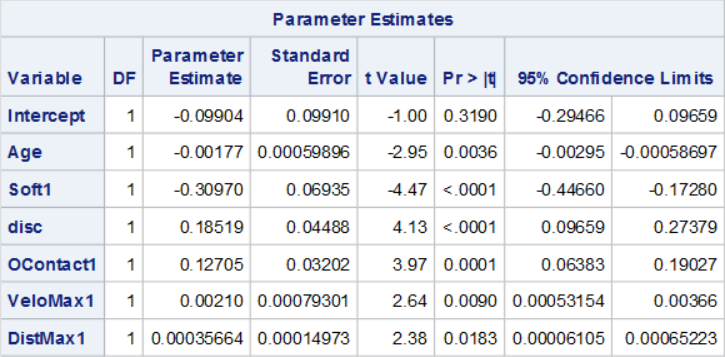
Results:



The ANOVA and Parameter Estimate table is given below for the full model of wOBA2 predicted by Age, Soft1, VeloMax1, DistMax1, OContact1, and Disc. The relationship is significant, and has an R-Squared value of 0.2939, an adjusted R-Squared value of 0.2688, a mean response of 0.33683, and standard deviation of 0.02956. All parameters are significant according to t-tests, and the 95% Confidence Limits for each parameter is also given in the table below. More importantly, it is logical that all parameters interact with the response variable in the way that they do, so the model passes that test.







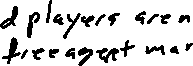
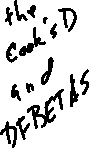
Discussion:

The goal of this project was to create a model that would adequately be able to predict batter performance in the upcoming year by using the previous year’s data. This goal was accomplished to a reasonable, with a R-Square value of .2939 and a standard deviation of 0.02956. Considering the league-average wOBA (approximately .320 each year), a player projected as average by this model would have a 68% (from the normal distribution) of finishing next year with a wOBA between .290 and .350. That’s approximately the range of league-average offense +/- 20%, or “bad, but playable” to “good, but not elite”. A better way of looking at this range might be noting that a player projected as average has an 84% of being at least 80% as good as league average, since missing high is a positive and desirable outcome. Perhaps the most useful way to use this model would be in team roster management. Considering there are 9 lineup spots in each MLB game, owners of a team could use this model with a Bonferroni correction of 0.05 = 0.45/9. That would greatly reduce the error in this model while giving a semi-reliable preseason prediction for how a team will perform, helping guide management in deciding when adding new players is necessary.



The uses of this model could potentially extend beyond that, however. One of the main topics of discussion over the 2017-2018 MLB offseason was the lack of activity on the free agent market. Many of the game’s best free agents went unsigned until very late in the offseason, and many ended up being forced to take contracts for much less than what they expected to earn. This is because most players don’t reach free agency until they’ve spent at least 6 years in the majors, meaning the average player reaches free agency around 30 years of age. Being hesitant to give large amounts of money to free agents is a decision backed by my model, since age has a significant negative correlation with offensive performance. This is important for teams to consider, but should also be addressed by the MLB players union. If recent evidence suggests that players are reaching decline before they become eligible for free agency, it should become a goal of the union to change the current pay structure (pre-free agency contracts are worth much less than those typically awarded in free agency).

To further assess the validity of my model, I compared it to other projection systems I could find. Information is somewhat limited, but I was able to find a couple comparisons. First off, I went with 2016 wOBA versus 2017 wOBA. This comparison gave a R-square value of .1839. This is essentially the baseline of my analysis. My model is more effective than simply looking at a player’s production from last year and assuming it will be the same next year. Next, I compared it to Statcast’s metric of xwOBA. This uses exit velocity and launch angle to give a player’s “deserved” wOBA. Comparing 2016’s xwOBA to 2017’s wOBA, gave an R-square value of .1737. Interestingly, xwOBA had even less predictive power than wOBA. Finally, I compared my model to Fangraph’s Steamer model. Comparing 2016 projections to 2016 results for players with at least 400 plate appearances gave a R-Square value of .2724. It’s important to note that while my value is technically higher, there are some flaws. First off, the data from 2016 is older than mine, and likely does not include Statcast info (since it was not released to the public until 2017) or any other improvements made to the model since then. Secondly, the model I created may produce a lower value if applied to a future sample. I can’t test my model on another sample since only a small portion of plays were tracked in 2015 and 2018 obviously hasn’t happened yet. It’s possible the difference (and maybe even more) between my model and Steamer is the result of overfitting, and applying this to a different sample won’t necessarily reproduce the same results. Additionally, my model had a number of points with very high leverage. While most of these points were centered around the mean prediction, it’s possible they are skewing my estimates. A larger sample could help fix that, but that isn’t possible now. Finally, the mean response of my model is .336, much higher than the actual league average of ~.320. This indicates significant survivorship bias in my model, since good players are much more likely to get 400+ plate appearances in consecutive seasons than are bad ones.



Overall, the model accomplished what it was supposed to do, reliably predicting future performance using stats that wouldn’t necessarily be the first thing you looked at in player evaluation. It is perhaps most interesting that a lot of the data collected by Statcast did not have a significant effect on the response variable. However, the parameters used are very logical, since there really are only three main components to hitting, all three of which are addressed by the model: power (MaxDist1, MaxVelo1, Soft1), plate discipline (disc), and contact ability (OContact1).

References:

1. *WOBA | FanGraphs Sabermetrics Library*, [www.fangraphs.com/library/offense/woba/](http://www.fangraphs.com/library/offense/woba/).
2. *Major League Leaderboards » 2017 » Batters » Dashboard | FanGraphs Baseball*, www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=400&type=8&season=2017&month=0&season1=2017&ind=0&team=&rost=&age=&filter=&players=.
3. “Statcast | Glossary.” *Major League Baseball*, m.mlb.com/glossary/statcast.”
4. “About Baseball Savant.” *Baseballsavant.com*, baseballsavant.mlb.com/about.

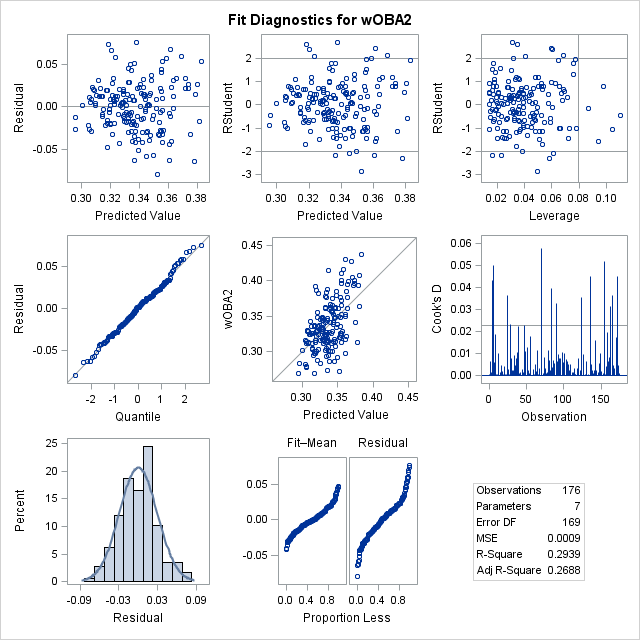


Figure 1A

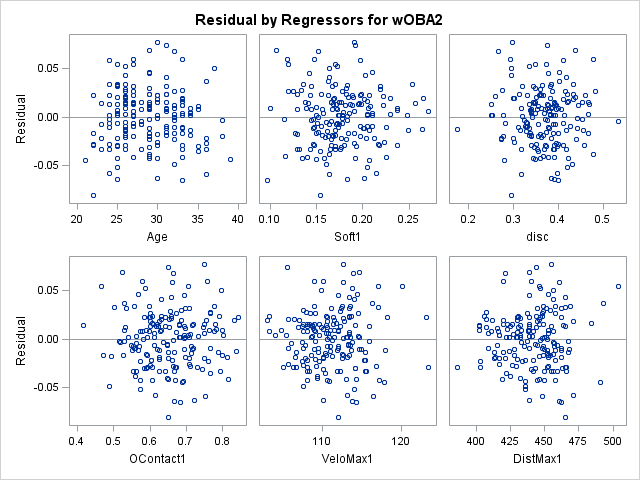


Figure 1B

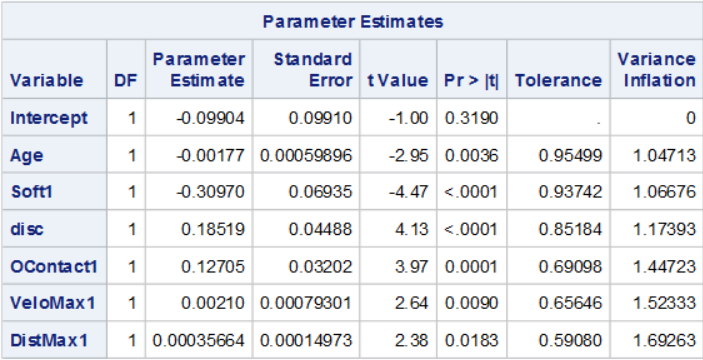


Figure 1C

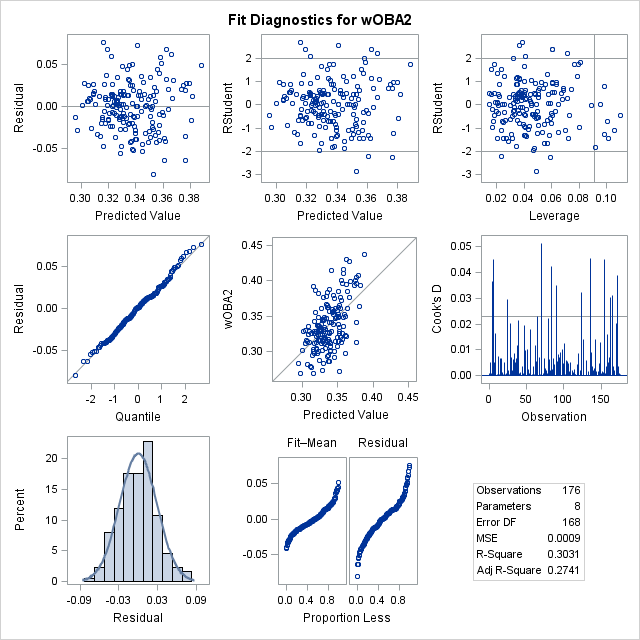


Figure 2A

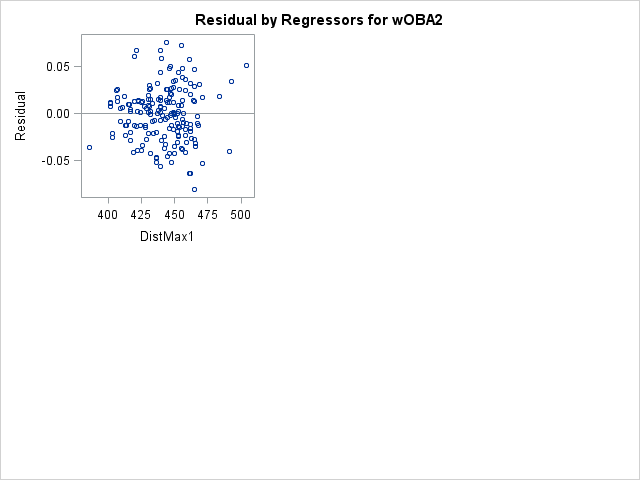
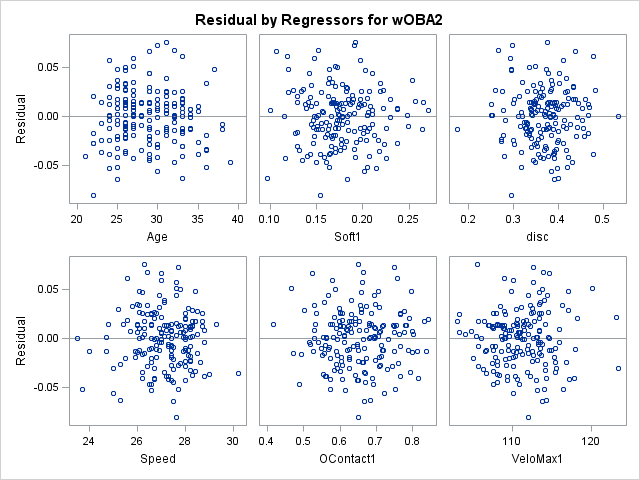


Figure 2B

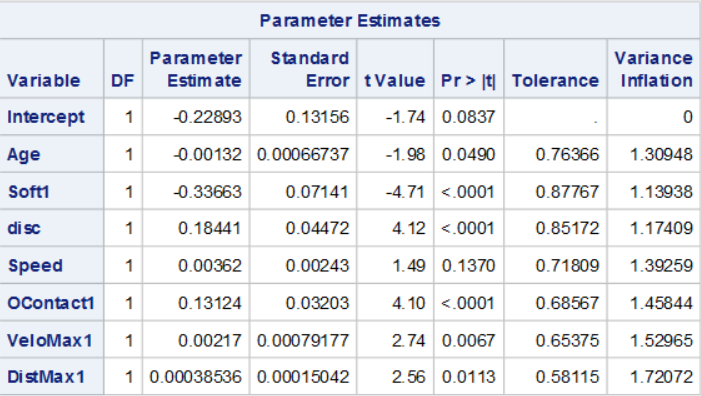


Figure 2C

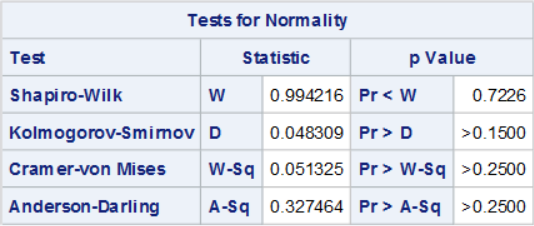


Figure 3

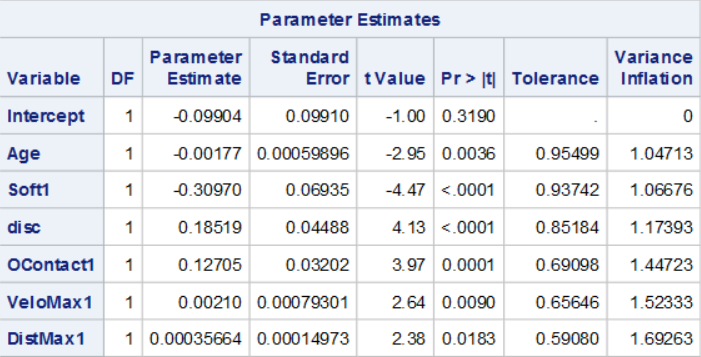


Figure 4