1 Introduction

College basketball has always been one of the most popular sports in the United States, with the NCAA Tournament being one of the most anticipated sporting events every year. During the NCAA Tournament or "March Madness", everyone tries to predict the winners of all of the games in order to get the elusive perfect bracket. However, no one ever talks about the score differential of those games. If there was a way to predict the score differential in a matchup, then maybe that helps sway your mind on a toss up game. If someone is trying to go back and forth on whether or not they want to pick an upset, but you find out it's most likely going to be a 20 point differential, then you are probably not going to pick that upset. But on the flip side, if that same model tells you it is going to be a two point game, you are probably going to take a chance and pick that upset.

A lot of studies have been done trying to predict the winner of athletic events, but rarely does anyone pay attention to the score and try to predict the differential of the game. The purpose of this study is to build a model that will accurately predict the differential of a basketball game by using box score statistics. A model that can accurately predict score differential could be used not only by fans, but also by coaches of basketball teams. If the model predicts accurately, then the factors put into the model obviously matter a lot to winning. A coach could use that information to focus on those particular variables or to recruit a certain way. As a fan, using it to predict March Madness games is an example, or using it to bet on a game if you are betting the spread. If the model says the point differential is going to be 2, but the spread is 11, then a fan can feel very confident in the bet that would be made. There is a place for this kind of prediction and it could be of use for a lot of people.

2 Literature Review

Perhaps the most important aspect to this research is figuring out exactly what statistics matter the most to the outcome of games. A study found "rebounding, assists, and 3-point shooting efficiency were determinant to win balanced and unbalanced games" (**?**). Rebounding, taking care of the ball, and shooting the three are all pretty standard as far as thinking of the most important statistics in basketball. But, there are several other statistics that are very important when looking at predicting the outcome of basketball games. All of these variables will be used in this study with offensive rebounds, assists, and three point field goals made being variables in some of the models. Another study by Daskalovski thought the most important variables were "minutes, field goals, field goals attempted, free throws, free throws attempted, assists, offensive rebounds, turnovers, and points" (**?**). Though minutes, point, and field goals are not used in this study, all of the other variables will be. The reasoning for not including field goals is that having field goals made, free throws made, and three point field goals made mess up the model. Having all three creates an unbalanced model, so not all of them can be used. The most important of these is three point field goals, so that will be included while normal field goals made and free throws made are not. Another common theme in basketball analytics is the four factors, these are "effective field goal percentage, free throw rate, turnovers per possession, and offensive rebound rate" (**?**). Every team does it differently, but one way or another this is shooting, free throws attempted, turnovers, and offensive rebounding. In this study three point field goals made, free throws attempted, turnovers, and offensive rebounds are all variables in the different models, meaning that the four factors are covered.

Another aspect that should be studied is actually predicting basketball games in general. West used "regression to model the number of NCAA Tournament games a team would win. His predictors were the teams proportion of

games won, cumulative points scored minus points allowed, Sangarin strength-of-schedule metric, and number of wins against teams with top-30 Sangarin rating" (**?**). There have been a lot of studies and models created to try and predict the winners of the game, but this study is trying to predict the score differential of the games as opposed to the winner of the game. However, this model will be similar and can provide direction on how a model should be formulated. Different metrics will be used as the outcome variable is different, but the formatting will be similar in a regression style.

Lastly, looking at how data is used overall in basketball is necessary to understand how the data works. Analytics and data are becoming extremely popular in basketball over the last decade. Demenius wrote about how data and analytics is changing basketball and said "advanced game statistics analysis gives a possibility to make better evaluations of everything that happens in a basketball game" (**?**). Everything in basketball is now analyzed more than it was before, from the amount of points on the scoreboard, to the distance players run, to how many times a guy high fives his teammates on the bench. Since everything is measured, it can be decided what the most valuable measurements are to actually winning basketball games. That is where this model is so important, finding out exactly how to accurately measure what matters in the final score differential.

3 Data

Data in this study was found online on Kaggle and downloaded so that it could be used. This data contained several different files that contained several years of college basketball information from regular season and postseason. The files that were needed for this study were regular season detailed game-by-game results, team names, and NCAA Tournament detailed results. The regular

season detailed results contained winning team ID, losing team ID, location, and box score stats for both teams and each row or observation was a game. The team names data matched the name of the school with the team ID that was used for that specific team. NCAA Tournament detailed data was the same as the regular season detailed results, but these were NCAA Tournament games instead of regular season games.

This data had games from several years, dating back to 2003, but for this study, the data was filtered to only data in the year 2019. This is because the 2021 data was not in the dataset and 2020 did not have an NCAA Tournament due to the Covid-19 pandemic. Having one season to look at made the data easier to work with and made it recent. Having the data be recent is important because basketball is played a lot different now than it was in 2003, so having the most recent season is important so that the playing style is the same. Data cleaning was done so that the name of the team was merged into the detailed results so that there was a team name attached to the result instead of just a team ID. Columns were also created to represent the differential between the winning team and losing team of games of important statistics. So, taking the number of steals by the winning team and subtracting the number of steals of the losing team gives you the differential column. For example, if the winning team had 11 steals and the losing team had 5 steals, then the differential column is 6 and that was done for each game.

This data is perfect for the problem that is trying to be solved, the data set provides everything that is needed. Using all three of field goals made, free throws made, or three point field goals made is difficult because it is too streamline and is confusing for the model created. But, the model is able to handle one and three point field goals made seems to be the most important of those variables, especially the way college basketball is played today. Having the data of other statistics such as assists, steals, blocks, etc also helps the model and helps use variables other than purely made shots to predict the outcome

of the basketball games at hand. Being able to predict score differential with statistics other than just purely shots made is important because it discovers the most important aspects of the game that are a little more hidden.

4 Methods

RStudio was used for the entirety of this study as this is the easiest way to upload, clean, and analyze the data. RStudio makes the whole process efficient and easy to keep track of, it is also easy to transfer the files from one device to another. The data was cleaned, the models were created, and the models were executed in RStudio using different programming practices in the software.

Several linear regression models were used to test the data and then create the prediction that predicted the score differential in a game based on box score stats. Statistics were chosen based on research and prior basketball knowledge and experience. Variables in the various models are free throws attempted (FTA), offensive rebounds (OR), assists (AST), turnovers (TO), steals (STL), blocks (BLK), and three point field goals made (FGM3). The variables in this model were the difference in these statistics between the winning team and the losing team. Estimating the model with the training data was the initial step to testing the model.

Average statistics for every team were then calculated for all of the variables, with the average covering the entire season leading up to the NCAA Tournament. To do this, the average had to be taken separately for wins and losses and then combined so that the entire season was calculated and could be fairly compared between the teams that were matched up. This is important because it is what is creating the model and prediction. The point of these models is to take what a team is averaging throughout the season in different

stats, compare it to their opponent, and predict what the differential of that game will be. The question is, if team a is averaging 5 more steals per game than team b, how much will that affect the expected differential of that game?

Next the model would take into account the differences in the matchup and see how much each individual variable would affect the outcome. Upon taking the differences of the teams into account, the model then made the predictions and showed exactly how much each variable in each model affected the outcome, which was the margin of victory or differential. The four linear models were structured as follows

$$margin = FTMdiff + ORdiff + ASTdiff + BLKdiff + STLdiff + TOdiff + FGM3diff \quad (1)$$

$$margin = ASTdiff * TOdiff + BLKdiff^2 + STLdiff^2 \quad (2)$$

$$margin = ASTdiff^2 + ORdiff^2 + TOdiff^2 + FTAdiff^2 + FGM3diff^2 \quad (3)$$

$$margin = STLdiff^2 + BLKdiff^2 + FTAdiff \quad (4)$$

$$margin = STLdiff + BLKdiff + FTAdiff + ORdiff + TOdiff \quad (5)$$

A linear model was the way to go for this study as it was the best way to see the effect of all of the variables on the outcome. Linear models also gave us a very easy way to compare the different models side by side, which was one of the goals of the study. Finding out which of the models was the best was key to the study and using the RMSE outputs of the linear models gave us a very easy way to do this. Several other models could have been used for this study and done completely fine, but a linear model seemed to be the correct model for the problem that was being solved and the goals the study wanted to accomplish.

5 Findings

Some of the general findings about the data included the fact that NCAA Tournament games were generally closer than regular season games in 2019. The average regular season game had a differential of 12.5 while the average NCAA Tournament game had a differential of 8.6. Regular season games also had a mean of 11 and NCAA Tournament games had a median of 8.6, obviously equal to the mean. This is not surprising as NCAA Tournament games are supposed to be between the best teams in the country, while there are some mismatches in the regular season that make the regular season mean higher.

The first linear regression model included the differentials of free throws attempted, offensive rebounds, assists, blocks, steals, and turnovers (see Appendix D). The differentials are the difference between the averages of the two teams throughout the regular season. This model included a testing RMSE of 9.8, meaning that it has a standard deviation of 9.8 from the actual values. As expected, some of the variables were better at predicting the outcome than other variables, which helps figure out what the most important variables are to prediction. Not surprisingly, the value that predicted the outcome the most was assists with a value of 0.820. This means that for every added assist, the differential goes up by 0.820. Assist differential and score differential have a very positive correlation (see Appendix A), which shows that the higher the assist differential is, the higher the differential will be. This makes sense because assists lead directly to scores, but it is good to see what was expected turn out to be correct. Second highest was actually blocked shots with blocks having a coefficient of 0.394. This is surprising from a basketball sense, as offensive rebounds and steals usually lead directly to points while blocks just stops the other team from scoring points. Something of note is that made three point field goals was not the highest valued variable when it came to prediction, it was second behind assists. A lot of coaches and fans would assume that the

team who makes the most threes has the best chance to win the game, and while that may be true, the model valued assists higher than made three point field goals when predicting the outcome. While made three point shots has a positive correlation to point differential (see Appendix B), it does not predict the outcome as much as assists. Interestingly, the only variable that had a negative effect was turnover differential. This makes sense when you look at the data as turnover differential and point differential have a negative relationship (see Appendix C). But, you can see in the graph that as the winning team has less turnovers, the score differential starts to get larger, which makes sense.

The RMSE value is what is the most important value, with this RMSE being 9.8, it means the study data is about 9.8 off of the actual values. This is a number that obviously would like to be as small as possible. Comparing this value to the other models will be what dictates the top models and separates the models from each other.

Model two was slightly more successful, this model emphasized fundamental basketball, which includes taking care of the ball and causing turnovers. The variables included in this model were assist differential multiplied by turnover differential, steal differential squared, and block differential squared (see Appendix E). It is not surprising that this particular model did well, there is a reason that coaches preach to take care of the ball and win the turnover battle. This model produces an RMSE value of 9.2, compared to the 9.8 of the first model. It should be noted that steals were more predictive than blocks in the model, with steals having one a half times the prediction value of blocks.

Model three was the worst of the models being compared, with an RMSE value of 10.1. This model focused on offense with the variables being assist differential squared, offensive rebounds differential squared, turnover differential squared, free throws attempted differential squared, and three point field goals made differential squared (see Appendix F). It is surprising that this model performed so poorly as it would stand to reason that offensive production

differential would predict games. Interestingly, when the variables were squared, the three point field goals made differential variable was more predictive than the assist variable. This was not the case in the first model where all of the variables were present.

Model four was a successful model, posting a RMSE value of 8.7. This model was the exact opposite of model three, this model focused on defense and its ability to predict the differential. This model included the variables of steal differential squared, block differential squared, and free throws attempted differential (see Appendix G). With only three variables it is surprising this model did well, but it shows that defense is more predictive than offense when looking at what variables predict the point differential in a college basketball game. It is not difficult to imagine that the team who plays better defense will win the basketball game. Free throws attempted differential was the highest value in this model with steal differential squared following closely behind.

The last model was the most successful model, this model was compiled of everything except for assists differential and three point field goals made differential. That means it was composed of steal differential, block differential, free throws attempted differential, offensive rebound differential, and turnover differential (see Appendix H). This model was testing to see how well statistics that did not lead directly to a score could predict score differential. This model had an RMSE value of 8.6, slightly better than model four. This is important to note because it shows that all of these variables that don't lead directly to scoring can factor into the prediction process.

6 Conclusion

There are several conclusions that can be made from this study with one of the main ones being that defensive variables used in this study are a bet-

ter predictor overall than the offensive variables used. All basketball fans have heard someone say that defense wins championships, this study would support that statement in a way but would revise it to say defense is a better predictor of score differential than offense. It also stands to reason that if a fan is trying to guess how close a game will be then they need to look at how well the teams compare defensively. If the projected score is not close, but the underdog gets more steals, more blocks, and shoots more free throws on average, then the score will probably be closer than the "experts" think. Also, if there is a coach looking at this data, the best way to close the gap with other teams, or extend your lead on other teams, is to get players who will get steals and block shots.

Overall, the models were extremely similar in how well they predicted the differentials of the NCAA Tournament testing data. All were within one and a half points of each other, but that one and a half points can be very important. As stated in the introduction, basketball games are notoriously hard to predict. Having a RMSE value of 8, 9, or even 10 is not something that is horrible, but it can definitely be improved upon. None of these models are perfect, but they provide a good baseline and can be built upon.

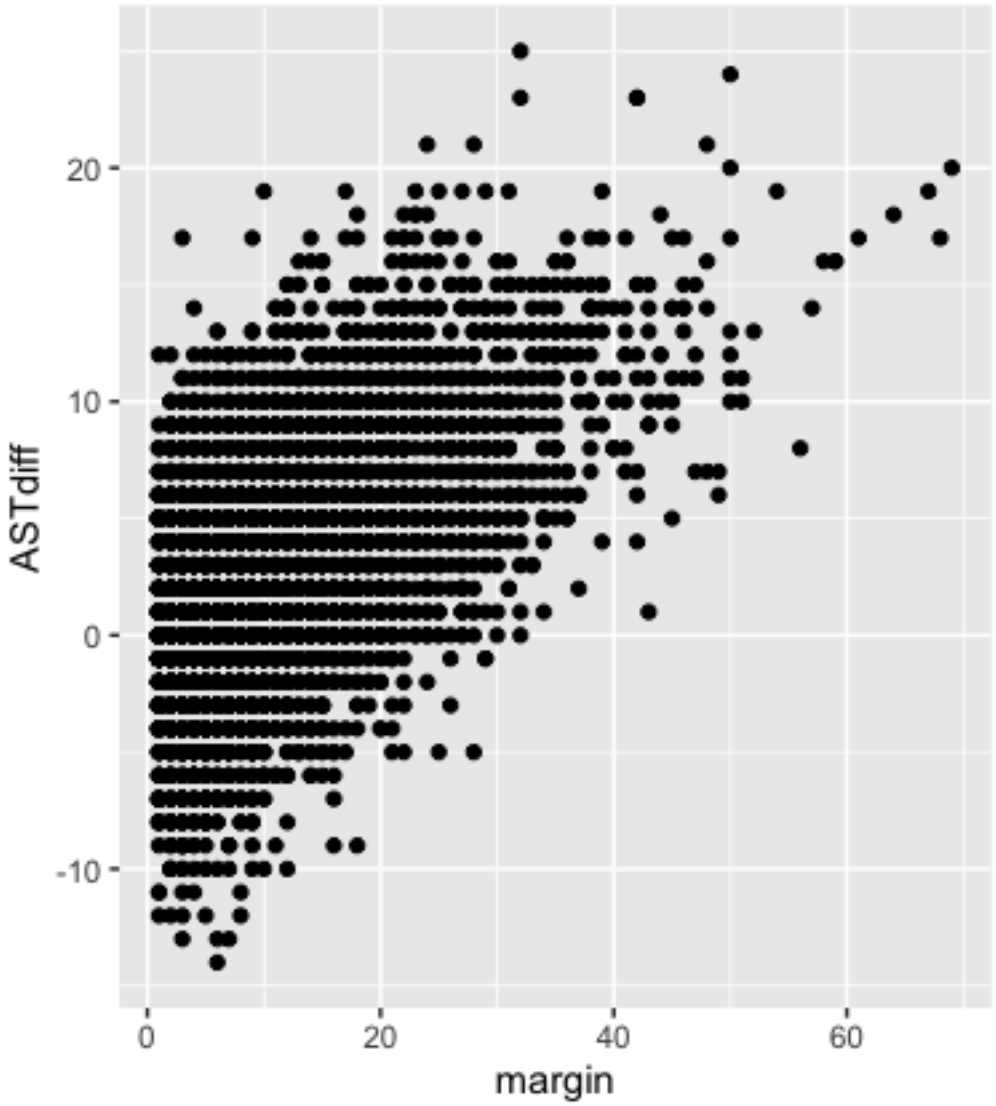
Another conclusion is that you have to build a well rounded team and have to build a well rounded model. The model that did the best was model five, which consisted of variables from all over, but none of them include points. If you get an assist or make a three point basket, that means points have gone up on the board. The statistics in model five do not mean that points have been put up on the board, just because you shoot a free throw does not mean it is going to go in. This shows that there is a lot that goes into the differential of a basketball game that is not just purely who made more shots.

This study was done to see if a model could be constructed to accurately predict the differential of NCAA Tournament basketball games. All of the models did well with some doing slightly better than others, but it was proven that box score statistics can be used to predict the score differential of games.

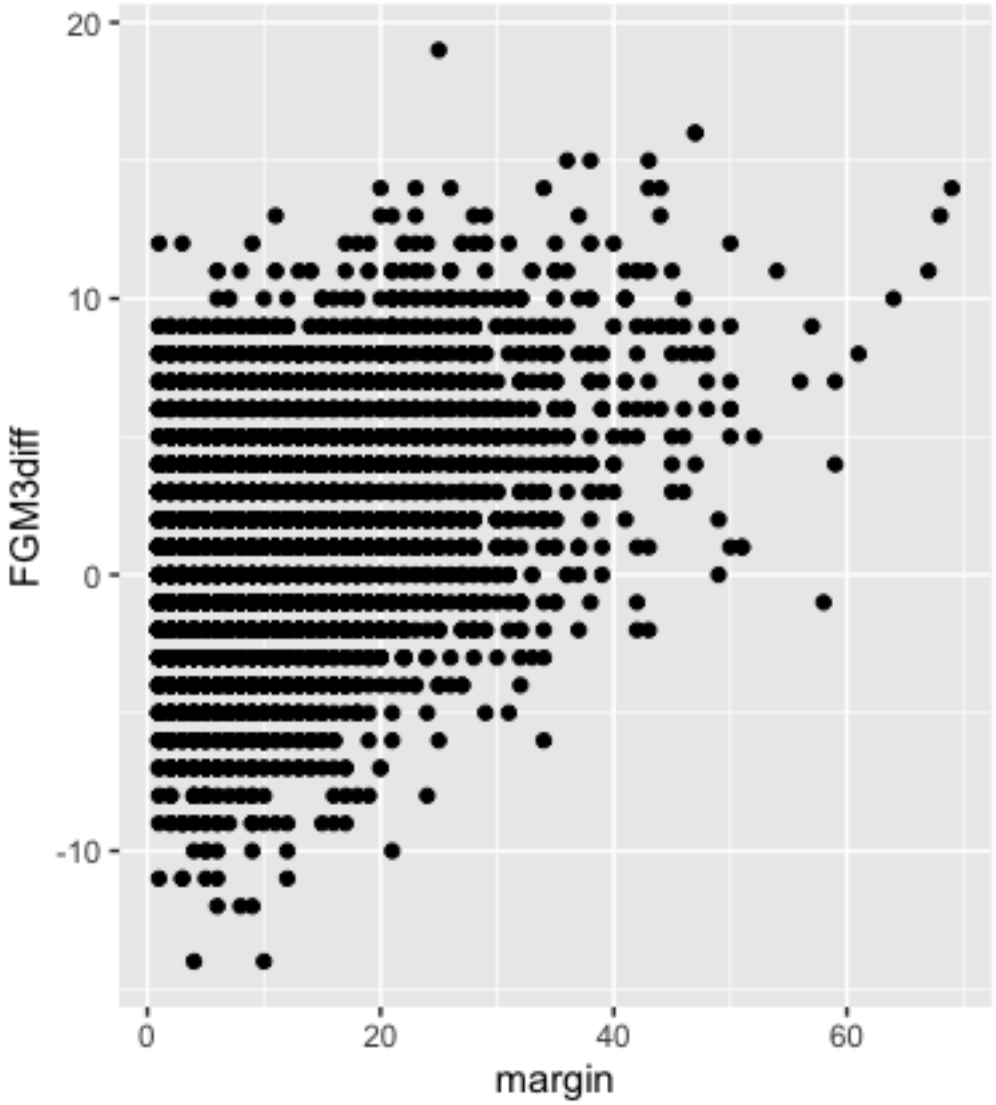
There are other statistics that could be used in the future, including shooting percentage instead of just shots made or how many shots were put up. This study did not explore percentage, but was using the raw box score statistics to try and predict the differential and it was relatively successful.

7 Appendix

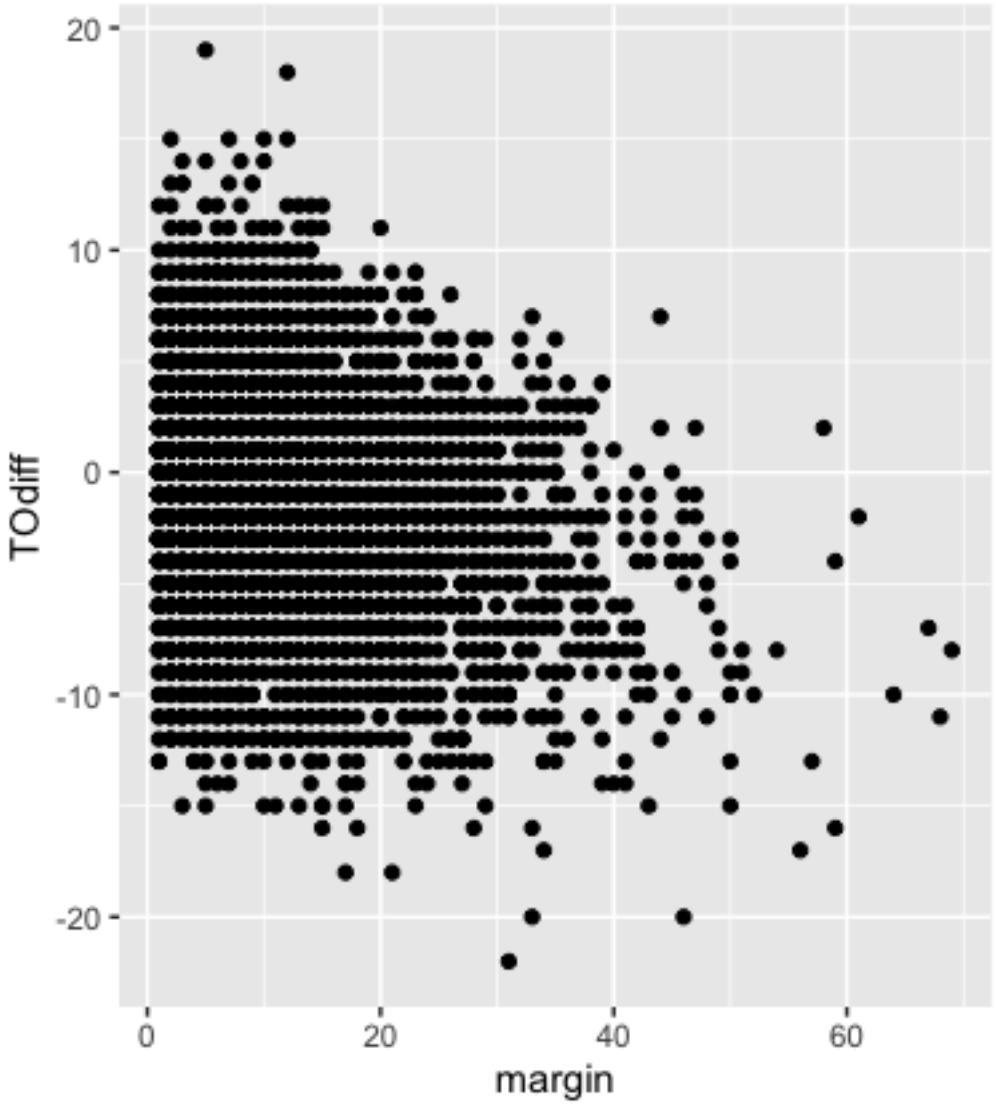
Appendix A



Appendix B



Appendix C



	Model 1
(Intercept)	7.075
	(0.132)
FTAdiff	0.181
	(0.012)
ORdiff	0.197
	(0.018)
ASTdiff	0.820
	(0.021)
BLKdiff	0.394
	(0.033)
STLdiff	0.233
	(0.036)
TOdiff	-0.243
	(0.027)
FGM3diff	0.562
	(0.028)
Num.Obs.	5463

Appendix D

	Model 2
(Intercept)	8.241
	(0.149)
ASTdiff	0.865
	(0.020)
TOdiff	-0.108
	(0.031)
poly(BLKdiff, 2, raw = TRUE)1	0.312
	(0.036)
poly(BLKdiff, 2, raw = TRUE)2	0.004
	(0.007)
poly(STLdiff, 2, raw = TRUE)1	0.204
	(0.038)
poly(STLdiff, 2, raw = TRUE)2	0.006
	(0.005)
ASTdiff × TOdiff	-0.041
	(0.004)
Num.Obs.	5463

Appendix E

(Intercept)	6.388
	(0.170)
poly(ASTdiff, 2, raw = TRUE)1	0.619
	(0.028)
poly(ASTdiff, 2, raw = TRUE)2	0.031
	(0.002)
poly(ORdiff, 2, raw = TRUE)1	0.150
	(0.017)
poly(ORdiff, 2, raw = TRUE)2	-0.001
	(0.002)
poly(TOdiff, 2, raw = TRUE)1	-0.306
	(0.021)
poly(TOdiff, 2, raw = TRUE)2	0.009
	(0.003)
poly(FTAdiff, 2, raw = TRUE)1	0.189
	(0.013)
poly(FTAdiff, 2, raw = TRUE)2	-0.001
	(0.001)
poly(FGM3diff, 2, raw = TRUE)1	0.422
poly(FGM3diff, 2, raw = TRUE)2	0.035

Appendix F

	Model 4
(Intercept)	10.973
	(0.172)
poly(STLdiff, 2, raw = TRUE)1	0.481
	(0.035)
poly(STLdiff, 2, raw = TRUE)2	0.019
	(0.006)
poly(BLKdiff, 2, raw = TRUE)1	0.514
	(0.043)
poly(BLKdiff, 2, raw = TRUE)2	0.010
	(0.008)
FTAdiff	-0.035
	(0.014)
Num.Obs.	5463

Appendix G

	Model 5
(Intercept)	11.134
	(0.138)
STLdiff	0.214
	(0.045)
BLKdiff	0.555
	(0.040)
FTAdiff	-0.040
	(0.014)
TOdiff	-0.335
	(0.034)
Num.Obs.	5463

Appendix H