Griffin Lyons

MET CS 555

Term Project

Data source: <https://www.kaggle.com/datasets/pablote/nba-enhanced-stats>

Data: Game data from NBA 2016-2017 season

Code: My code is attached separately

**Note:** In my submission there is a PDF from the data source explaining some of the statistics by their names in the dataset.

**1. Describe your research scenario and question(s).**

**Briefly describe your research scenario. Similar to our class examples, you should first describe the overall scenario and then specify a specific research question (or questions) based on it.**

My research scenario is this: I am interested in knowing how strongly the score (represented by difference in team scores) at the end of the first quarter of a basketball game predicts the outcome (W/L or difference in scores at the end of the game, with a positive difference representing a win and a negative difference representing a loss) in a game of NBA basketball.

**2. Describe the data set.**

**Briefly describe the data set. Describe each variable of the data set that you plan to use in your analysis. Describe any data cleaning you have performed. If possible, provide a link to the main data set source.**

My dataset is comprised of game information from the 2016-2017 NBA regular season (excluding any preseason exhibitions or the postseason (playoffs and the championship series/finals)). It is taken from this dataset on Kaggle: https://www.kaggle.com/datasets/pablote/nba-enhanced-stats

Game data includes team name abbreviations, each team’s score for each quarter (including overtime quarters), and various team statistics such as percentage of successful goals. There are two copies of each game, swapping the team described under ‘team’ and the team described under ‘opponent’. This also swaps the ‘result’ statistic, understandably, as the team statistic for the result can only be a win or a loss.

I plan to trim this dataset from the 2460 games by half to 1230 games to avoid this duplication’s potential impact on analysis. I will remove every second game, which will have the effect of making the ‘team’ always be the home team, and the ‘opponent’ always be the visiting game. For outcome, I will use two variables.

The first, for linear regression, is the point difference between teams after all of the quarters. I will make a column that adds each team’s points up across all quarters (if there is no overtime, the value of any quarter after the fourth is 0 in the dataset) and subtracts the opponent’s from the team’s. The second is the dichotomous ‘win’/‘loss’, which I will make an additional column transforming the “team results” column value into a 1 or a 0.

I will take a random sample of 120 games (to adhere to the “population should be at least 10 times larger than sample size” rule of thumb). I will create two more columns: point difference at the end of the first quarter (team points for the first quarter minus opponent points for the first quarter), and point difference at the end of the first half (team points for the first half minus opponent points for the first half). To make this latter column I will add the values of the first and second quarter for each team, and subtract the opponent’s value from the team’s value.

**3. Describe the statistical methods you plan to use.**

**Briefly describe the statistical methods you will be utilizing to investigate your research question(s).**

I will perform simple linear regression and multiple linear regression on the game-total point difference as the outcome variable/response variable (a positive point difference being a proxy for winning). I will also perform simple logistic regression and multiple logistic regression on the win/loss dichotomous variable as the outcome/response variable. First quarter point difference will be the explanatory variable in both cases.

I am using first quarter point difference for my explanatory variable rather than just first quarter score, and total point difference rather than just total score, because the score in the first quarter alone is not informative. Basically, the first quarter point difference is whether the team being focused upon is leading (positive difference) or trailing (negative difference) and whether that predicts a win (positive point difference when expressed numerically rather than categorically) or loss (negative point difference when expressed numerically).

Likewise, the total score on its own is meaningless, because of course, barring utter game-stopping disaster it will not remain equal to the number of points scored in the first quarter by the team, and barring utter game-stopping disaster and something like officials overturning points after the conclusion of the first quarter (overturning points well after they have been credited to a team is very rare, and in any case the official record after the conclusion of the game would not falsely record the wrong point total).

Other variables I will consider in the respective multiple regressions are the point difference between the team and opponent after the first two quarters (“at the half” as it is referred to), the team’s field goal percentage (how many of their shots were a success), and team turnovers (losing possession of the ball to the opponent before a shot has been taken at the opponent’s basket).

**4. Report your results.**

**Write up the results of your analysis. You should present tables and figures when relevant, and you should have a short write-up describing your results.**

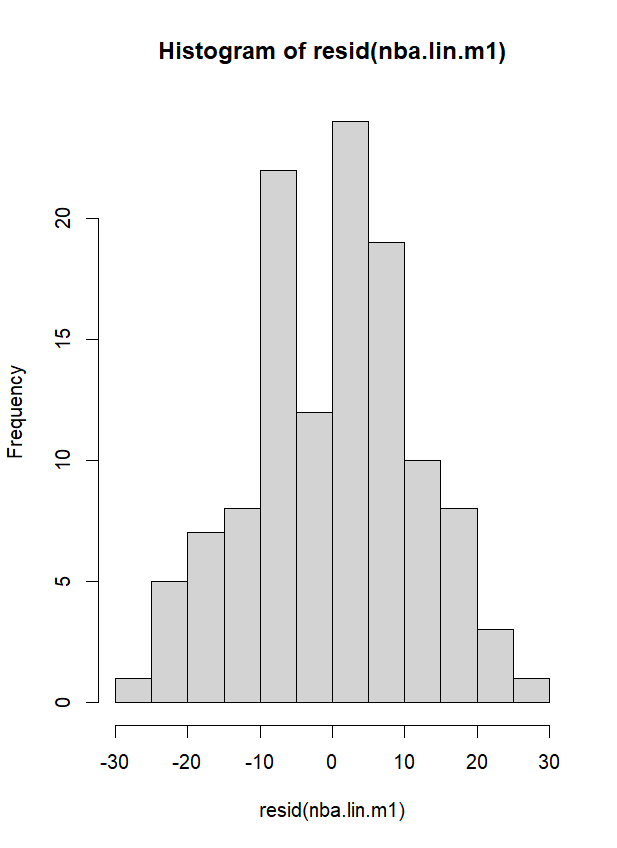
First, we perform a simple linear regression. We save our regression to a variable:

nba.lin.m1 <- lm(nba.samp$tpd ~ nba.samp$q1pd)

If we return that variable, we learn that βq1pd equals 0.6522. That is how much we expect the total point difference to increase for each unit increase in first quarter point difference.

**tpd** is our created column of values for total point difference between team and opponent, and **q1pd** is our created column of values for point difference at the end of the first quarter between team and opponent.

If we do a histogram of the residuals, it looks fairly normally distributed:



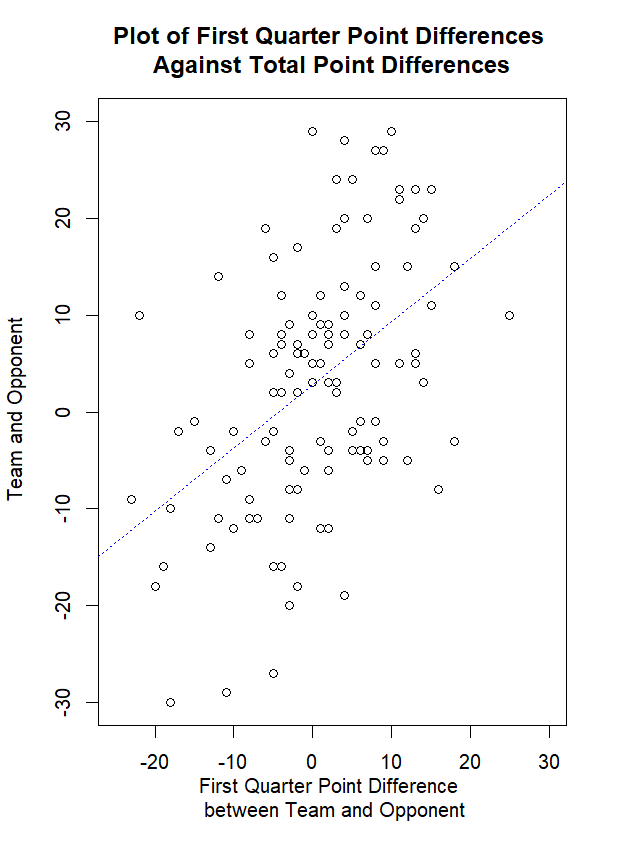
Let’s visualize this in a scatterplot with some code:plot(nba.samp$q1pd, nba.samp$tpd, xlim = c(-25,30), ylim = c(-30,30),

xlab = "First Quarter Point Difference \n between Team and Opponent",

ylab = "Total Point Difference between \n Team and Opponent",

main = "Plot of First Quarter Point Differences \nAgainst Total Point Differences")

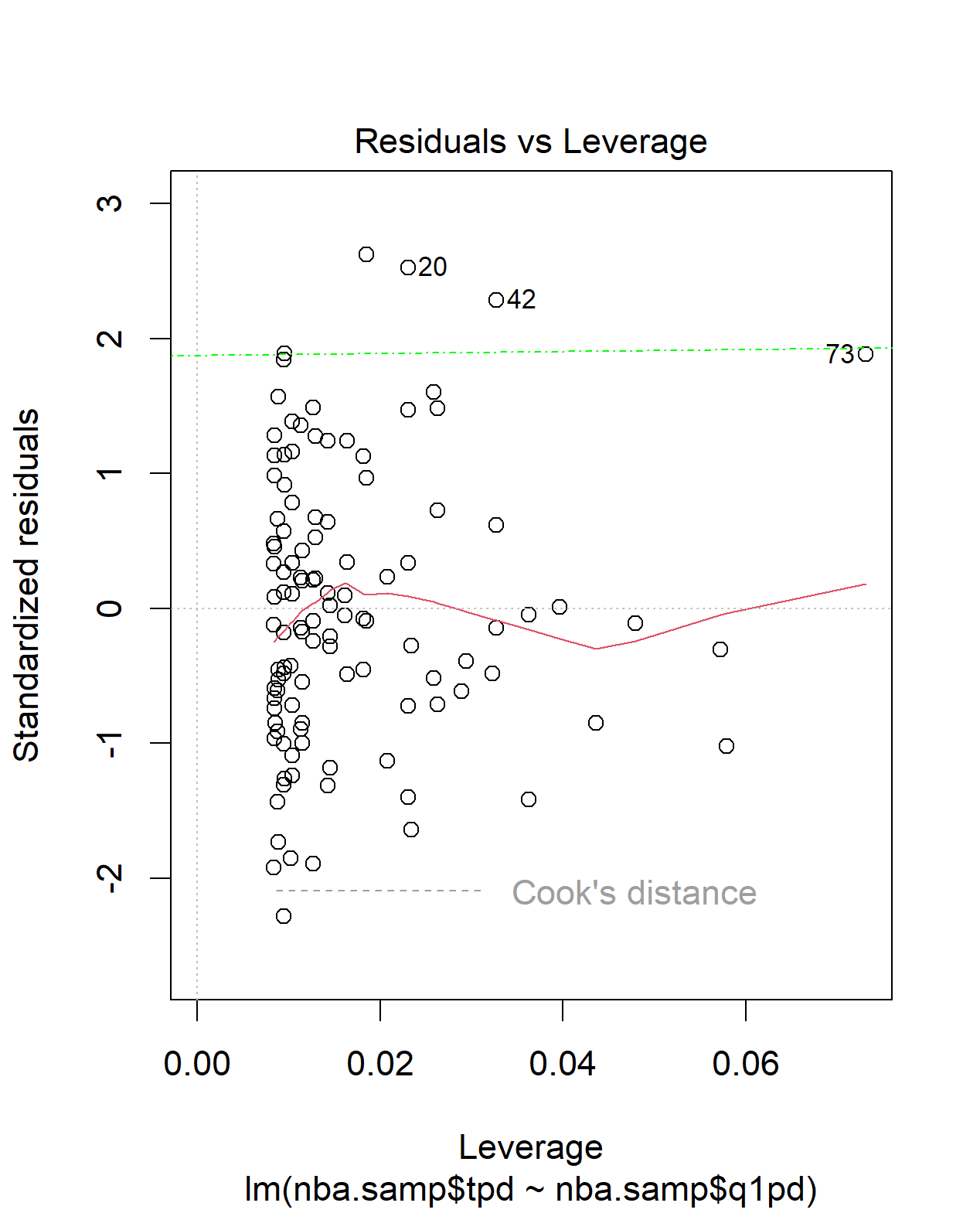
abline(nba.lin.m1, lty=3, col = "blue")



That looks like a positive association, and a strong one at that.

We can look for outliers by plotting the regression itself:  
  
plot(nba.lin.m1)

There appear to be three outliers, but they do not affect the regression line meaningfully. You can check with this code (all of the overlying lines overlay one another, which is why only the green line is visible):



outlier.1 <- lm(formula = nba.samp$tpd ~ nba.samp$q1pd,  
 data = nba.samp[ -20, ])  
abline(outlier.1, lty = 4, col = "red")  
  
outlier.2 <- lm(formula = nba.samp$tpd ~ nba.samp$q1pd,  
 data = nba.samp[ -42, ])  
abline(outlier.2, lty = 4, col = "yellow")  
  
outlier.3 <- lm(formula = nba.samp$tpd ~ nba.samp$q1pd,  
 data = nba.samp[ -73, ])  
abline(outlier.3, lty = 4, col = "green")

Now we’ll use the five steps to test the significance:

1. Set up hypotheses, select alpha
   1. H0: βq1pd = 0 (there is no linear association between point difference between the team and the opponent at the end of the first quarter and point difference between the team and the opponent at the end of the game)
   2. HA: βq1pd ≠ 0 (there **is** a linear association)
   3. We will set α at 0.05
2. We pick an appropriate test statistic. I will use the F-statistic, which is equal to the regression mean squares divided by the residual mean squares.
3. Decision Rule:
   1. We find the appropriate F-statistic with this R code:

qf(0.95, df1=1, df2=118)

The first number is 1-α, which in this case equals 0.05, and therefore is 0.95. df1 equals 1 because *k* equals 1 for simple linear regression, and df2 equals 118 because in a simple linear regression, the residual degrees of freedom equals n – k – 1, which is 120 – 1 – 1.

* 1. We find that the F value, F1,118,0.05, equals 3.921478.
  2. Decision Rule: Reject H0 if F ≥ 3.921478.
  3. Otherwise, do not reject H0.

1. We can use the following code to get an ANOVA table and find F (as well as *p*):

anova(nba.lin.m1)

(we can also use the summary function)  
  
F equals 32.774.

1. Conclusion:
   1. We reject H0 because 32.774 ≥ 3.921478. We have significant evidence at the α = 0.05 level that βq1pd does not equal 0. That is, there is evidence of a significant association between the point difference between team and opponent at the end of the first quarter and the total point difference between team and opponent at the end of the game.

We should note here, though, that R2 = 0.2174 (adding Reg SS and Residual SS from the ANOVA for the Total SS, and dividing Reg SS by Total SS; you can also use cor() on nba.samp$q1pd and nba.samp$tpd and square that to get the same.) The model only explains some of the variation.

Now we’ll try a multiple linear regression with our variables being first quarter point difference, first half point difference, field goal percentage, and team turnovers.

nba.lin.m2 <- lm(nba.samp$tpd ~

nba.samp$q1pd+nba.samp$fhpd+nba.samp$teamFG.+nba.samp$teamTO)

We perform the global test with the F statistic to determine whether these predictors are associated with total point difference, with 4 and 115 degrees of freedom, at an alpha level of 0.05. The numerator for F is because we have k = 4 predictors, and the denominator for F is because the denominator is equal to n minus k minus 1, which is 120 minus 4 minus 1, which is 115.

With this code, we find the F-statistic.

qf(0.95,4,115)

We get 2.450571. If F ≥ 2.4506, we have significant evidence that these predictors are associated with the total point difference. With the following code, we can find that out:

summary(nba.lin.m2)

We get an F-statistic of 37.52. That’s greater than 2.4506, so we do indeed have significant evidence.

Now we want to know if the individual coefficients are significant. We’ll use the t-statistic. We already have the values from the summary function conducted on nba.lin.m2, so we can compare them against the t-statistic for significance. We find it with this code, knowing our denominator (115) and alpha (0.05, divided in half because this is two-sided):

qt(0.025, 115)

We get -1.9808 as our t-statistic.  
  
The absolute values of the t-statistics determine whether each variable is significant.

First quarter point difference is 0.562, greater than -1.9808, so it is significant.

We can do a confidence interval test on our model:

confint(nba.lin.m2, level = 0.95)

We are 95% confident that our coefficient for first quarter point difference is between -0.1760 and 0.3156. The coefficient is 0.06976

We can repeat these for all of the other coefficients. First half point difference has a coefficient of 0.51673, team field goal percentage has a coefficient of 102.6648, and team turnovers has a coefficient of -0.6460. They all fall within their respective confidence intervals.

My interpretation is that all of these coefficients more strongly predict total point difference than first quarter point difference. First half point difference and team field goal percentage have positive relationships to total point difference, and turnovers has a negative relationship (which makes sense, as turning the ball over to the opponent, and possibly incurring penalties depending on the reason for the turnover, harms a team’s ability to score and increases the opportunities for the opponent to score).

Now, we’ll perform simple logistic regression and then multiple logistic regression, replacing total point difference with the categorical win/loss (nba.samp$winloss). The outcome ‘win’ is 1, and the outcome/non-event ‘loss’ is 0. We’ll save our model to a variable, and then perform a summary on it to help understand whether it is significant.

nba.log.m1 <- glm(nba.samp$winloss ~ nba.samp$q1pd, family = binomial)

summary(nba.log.m1)

I will skip the five steps here to summarize: if we use an alpha level of 0.05 (or even 0.001) we have significant evidence that there is an association between first quarter point difference and the outcome of a win occurring/a “success” occurring, because the p-value for first quarter point difference is 0.000331, or ≤ α (whether it’s 0.05 or 0.001). If we went to the trouble of finding an appropriate z-statistic (at α = 0.05 and divided by two, it would be 1.960), our summary’s z value of 3.590 would mean we could reject a null hypothesis that βq1pd = 0.

We want to test how well our model fits, so we use the pROC package to find the area under the curve.

library(pROC)

nba.samp$prob <- predict(nba.log.m1, type=c("response"))

nba.tg1 <- roc(nba.samp$winloss ~ nba.samp$prob)

print(nba.tg1)

We save the comparison of predictions and actual responses into a new column, “prob”, and then use the roc() function from the pROC package to find area under the curve. When we print it, we find that the area under the curve is equal to 0.6961. That’s better than chance, or an AUC of 0.50, but not particularly strong.

Now we want to disentangle that from the other variables we’ve already tested in the multiple linear regression setting, and compare that fit.

We can create a model and summarize it with this code:

nba.log.m2 <- glm(nba.samp$winloss ~ nba.samp$q1pd + nba.samp$fhpd +

nba.samp$teamFG. + nba.samp$teamTO, family = binomial)

summary(nba.log.m2)

The p-value for first quarter point difference (nba.samp$q1pd) is extremely unfavorable to its significance at 0.4599, far greater than α = 0.05 or even α = 0.1.

To determine whether all of these are significant, we will want to use the aod package, and use its significance test:

library(aod)

wald.test(b=coef(nba.log.m2), Sigma = vcov(nba.log.m2), Terms = 2:5)

We use 2 through 5 as the terms, because those correspond to the coefficients for first quarter point difference, first half point difference, team field goal percentage, and team turnovers. We get a p-value of 0.000022, meaning that we can reject the null hypothesis that all coefficients equal 0, because *p* is smaller than α = 0.05, and even smaller levels of alpha like 0.001.

To test the fit of this model, we can use the pROC package again with the same steps, creating a new variable and then using roc() to find the AUC:

nba.samp$prob2 <- predict(nba.log.m2, type=c("response"))

nba.tg2 <- roc(nba.samp$winloss ~ nba.samp$prob2)

print(nba.tg2)

By printing nba.tg2, we learn the area under the curve is 0.8243. That is better than the fit for first quarter point difference alone.  
  
We can visualize it with this code:

plot(1-nba.tg2$specificities, nba.tg2$sensitivities, type = "l",

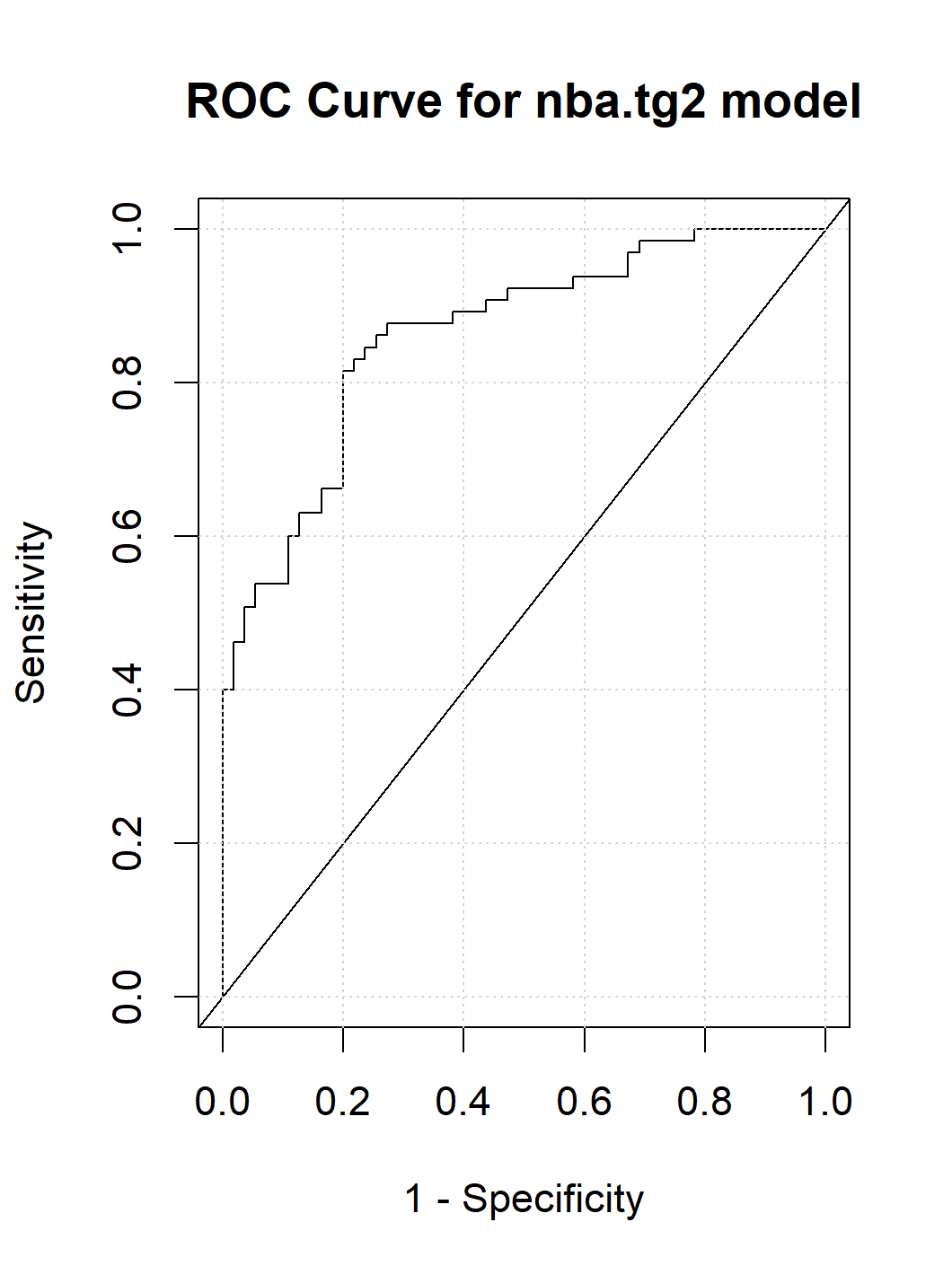
xlab = "1 - Specificity", ylab = "Sensitivity",

main = "ROC Curve for nba.tg2 model")

abline(a = 0, b = 1)

grid()

And get this:



**5. State your conclusions and discuss any limitations.**

**State the conclusion so that a none-statistician can understand. Discuss any potential limitations of your analysis. For example, are you suspicious that the assumptions of your test may not hold? Do you feel the analysis may have limitations for any other reasons?**

My conclusion is that there is a weak association between whether a team has a lead or is trailing at the end of the first quarter, and other factors like whether they are leading or trailing at the end of the first half, or their percentage of shots made for the whole game, have stronger associations with the outcome of the game (win or loss). There are a large number of reasons I think my analysis is limited. I am not very familiar with basketball statistics and analytics in general despite being a fan of the sport. I used a small sample size when I could have pulled from years of data and perhaps taken a much larger sample (though the disruptions of the pandemic to the 2019-2020 and 2020-2021 seasons, and to some extent 2021-2022 would make analyzing these seasons difficult if their data had been in the set I pulled this data from).

There are many real world variables that could explain performance – maybe some teams begin their games strong and take a lead after the first quarter, only to falter across the rest of the game when their opponents assert themselves, while a lead after the first half is more indicative of a more stable, consistently high-performing team. I could not find data for minutes played by or any absences of regular starting players (the five players on a team tasked with starting the game, who are commonly that team’s most important/best players), which might help account for some outcomes.

I knew first quarter point difference might have a weak association, but I lacked access to more detailed play-by-play data of individual games that might help me answer more detailed questions, and more importantly my lack of familiarity with basketball analytics and constraints on my time meant that it was probably a better idea to set my sights on something simple.

I found this project to be an enriching use of what I learned in this class, and it opened my eyes to the real-world applications of some techniques (and I am interested in learning more about basketball analytics as an application of this knowledge and for my own enjoyment of the sport), but relatively speaking I think my approach and methods were relatively crude for drawing conclusions from this data. That doesn’t mean I think my work was bad – it just means I can stand to learn more and how and what to apply, as well as what questions to ask about data.

I’ve attached my code separately in the submission.