KOBE’s Shot percentages

# Introduction

Many types of statistical modeling are used for different applications. For normally distributed data pulled from a single, or related variables there is the T-tests, linear regression, permutation test and more. When multiple groups are present we have ANOVA and MANOVA tests. There are ways to cut down on the number of variables to get an efficient model that does not “over fit” our data. But the limitations of these methods is the response variables has to be numeric, or analog. To describe a categorical variable, or digital, there is logistical regression. We will be running logistical regression on Kobe Bryant’s shots trying to predict if he will make or miss the shot. We will also be answering some questions about different scenarios and how they affect his shot percentage.

# Data Description

Using data provided by Kaggle for free at [**https://www.kaggle.com/c/kobe-bryant-shot-selection**](https://www.kaggle.com/c/kobe-bryant-shot-selection). The data consist of 29 years of Kobe Bryant’s shot history: 25 variables and roughly 30 thousand observations. The data set includes 5 thousand observations with an unknown value for the shot\_made\_flag response variable. To prevent leakage, predictive models must use only the observations that came before the missing value to predict the missing value.

# Exploratory Data Analysis (EDA)

Before starting on a project, we like to know the data and see what is useful and what we need to transform, because in most statistical models there are many assumptions that need to be meet. Logistical regression does not have these same limitations, the main assumptions focus on independence of variables, which we can assume, because Kobe cannot shoot the same shot twice. The purpose of the EDA is to visualize the different variables and their relationship to Kobe’s shot percentage. We used box plots to look at the different numeric variables and frequency counts on the different categorical variables. In fig.1 vs fig.2 we are displaying two different distributions.

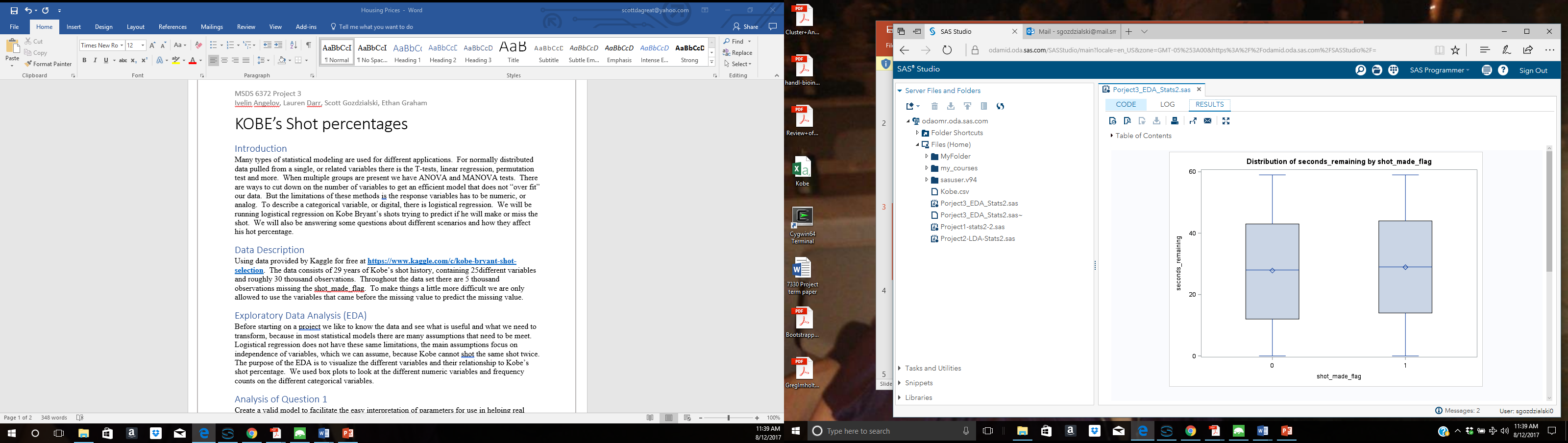


Fig. 1 (misses left, made right)

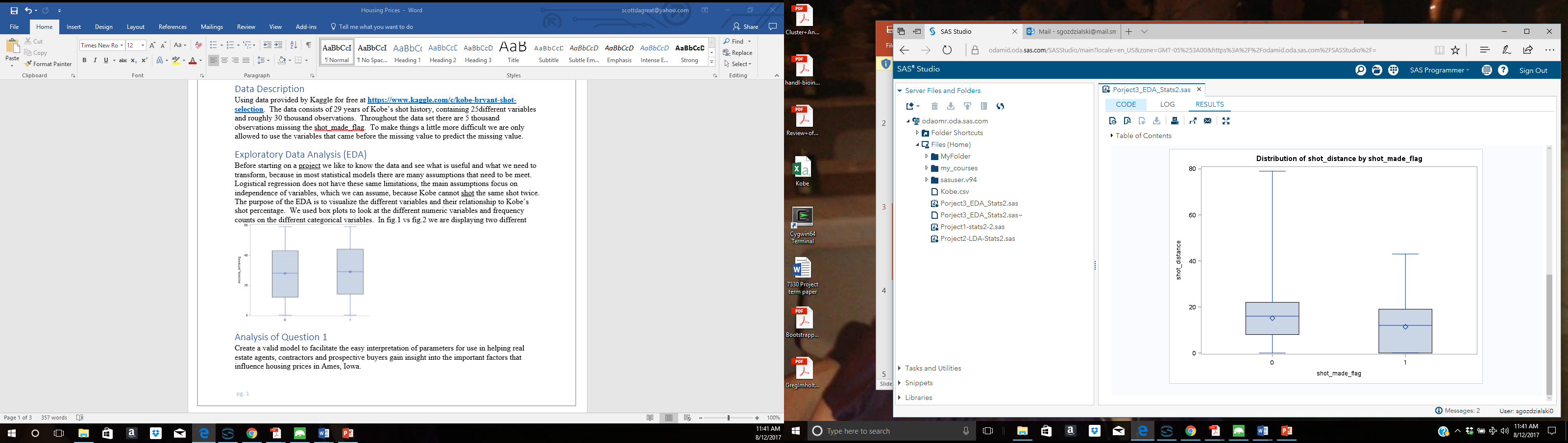
Seconds remaining

Fig. 2 (misses left, made right)

shot distance

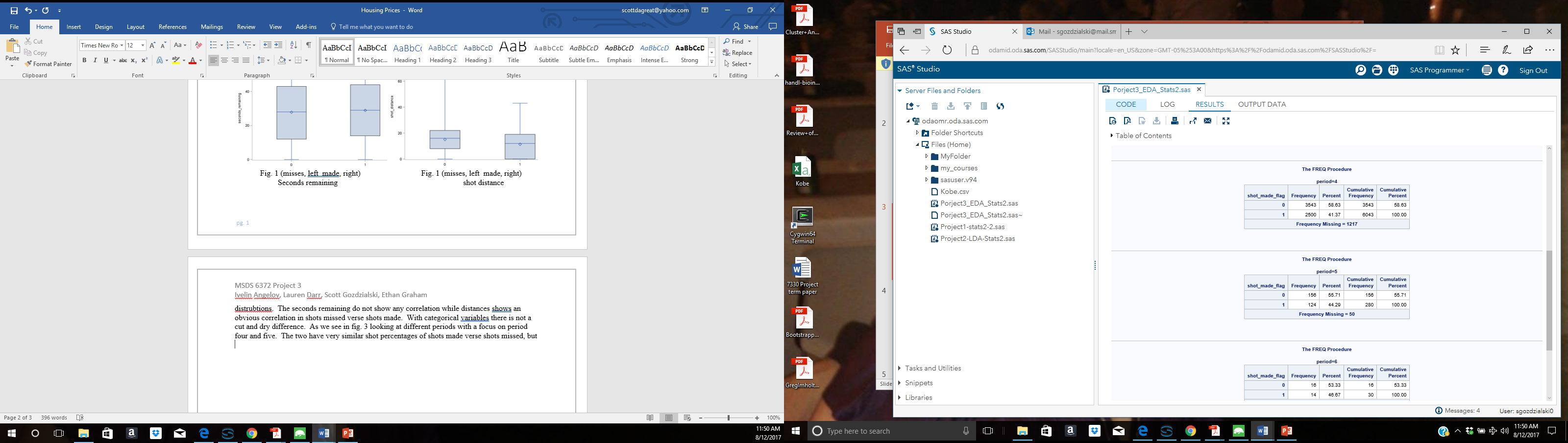
The seconds remaining do not show any correlation while shot distance shows an obvious correlation in shots missed verse shots made. With categorical variables there is not a cut and dry difference. As we see in fig. 3 looking at different periods with a focus on period four and five. The two have very similar shot percentages of shots made verse shots missed, but with different standard distributions we end up with completely different statistics. With the greater count of shots taken in period four than five the small percentage difference makes a greater difference than the difference in in period five.

Fig. 3

# Interpretation Models

# Model 1.

Is, Kobe’s shooting percentage is better at home than when he is away?To answer this question, we first check if the two teams in variable “matchup” are separated by “vs” meaning home game or “@” meaning away game. Than we create a variable “play\_home” which equals to 1 Kobe plays at home and 0 otherwise.

After creating the “play\_home” variable, we create a logistic regression model: *shot\_made\_flag(event ='1') = play\_home* and inspect the slope and intercept*.* After fitting the model, we get the flowing equation:

**shot\_made\_flag = -0.2556 + 0.0811 \* paly\_home**

From the equation we see that if Kobe plays at home will have around 8.1% greater chance of him making the shot.

# Model 2.

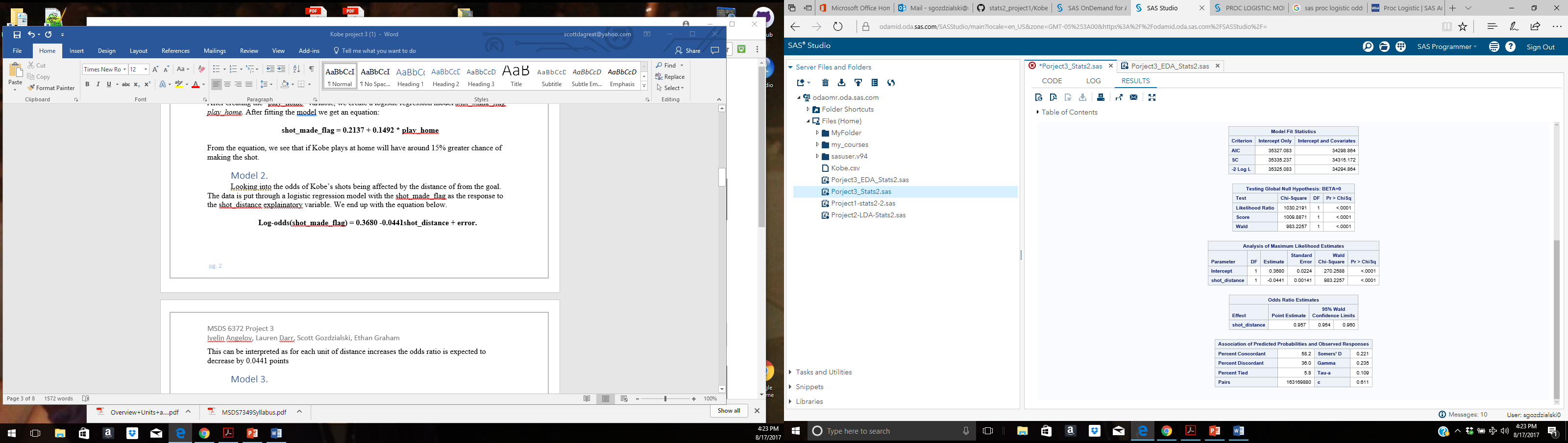
Looking into the odds of Kobe’s shots being affected by the distance of from the goal. The data is put through a logistic regression model with the shot\_made\_flag as the response to the shot\_distance explainatory variable. We end up with the equation below.

**Log-odds(shot\_made\_flag) = 0.3680 -0.0441shot\_distance + error.**

The odds ratio is calculated by taking the exp(Bo +B1 X), so the change in odds ratio, if we increase the distance by one unit, since there are no other variables is equal to the equation below.

**Odds-Ratio(shot\_made\_flag)= exp (-0.0441 (1-0) or 0.957**

SAS also calculates the 95% confidence ratio. So with a 95% confidence we can say the Odds-ratio of Kobe making the shot with each unit of distance measurement in between 0.954 and 0.960.

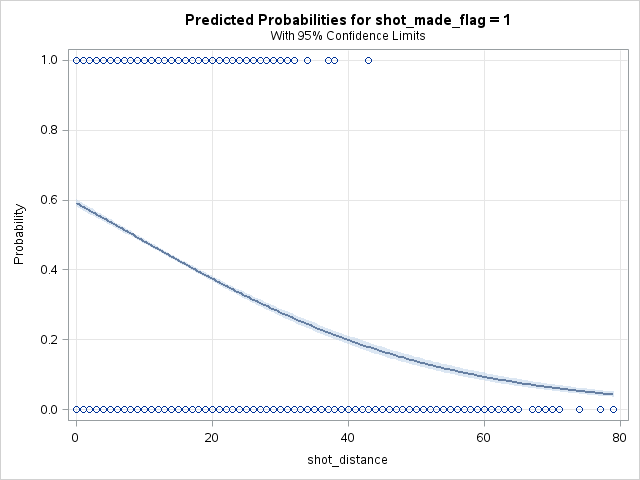


# Model 3.

Given the odds ratio, the probability of Kobe making a shot can be calculated as

This calculates the following probabilities for a few example distances,

|  |  |
| --- | --- |
| Distance | Probability |
| 5 | 0.5368 |
| 10 | 0.4817 |
| 20 | 0.3742 |
| 30 | 0.2778 |



# Model 4.

It is a practical assumption that the pressure of a playoff game could either bring out intense competitiveness and focus in a player, or result in failure. This kind of failure, especially when a player or team is favored to win, is colloquially referred to as “choking.” This model looks to establish or discredit a relationship between the distance Kobe was from the hoop, whether he was in a playoff game, and the probability of making a shot.

A preliminary exploration of the relationship between the distance Kobe is from the hoop, and the odds of him making the shot suggest that there is a threshold where the odds of Kobe making a basket should drop to zero (Figure 4.1). Specifically, this box plot shows that Kobe did not make any shots at distances greater than approximately 45units. Frequency tables of shots made and not made in playoff games and regular season games suggest that there is likely no difference in performance between playoff and regular season games, holding all other variables constant (Figure 4.2).

Logistic regression is appropriate to formally establish a model explaining the proportion of shots made by the distance to the goal and the game type condition. Specifically, binary logistic regression is appropriate when the response variable is one of two categories. The first model tested stated that the probability of making a shot is equal to the summary of shot distance, playoff status, and the interaction between these terms. SAS procedure Proc Logistic was used to first test the hypothesis that the probability of Kobe making a shot is independent of shot distance and playoff status. The test of this ‘global null hypothesis’ or goodness of fit test provided three criteria to choose from: likelihood ratio test, Score test, and Wald test (Figure 4.3). The likelihood ratio test uses the negative log likelihoods calculated for the full model and the reduced model (Figure 4.3). The reduced model reduces the explanatory variable terms to 0, leaving only the intercept. In this case, the negative log likelihood of the full model was smaller than the negative log likelihood of the reduced model by approximately 1,031. This is a large difference, and a chi-square test of independence results in a p-value <0.0001. Thus, the null hypothesis is rejected and it is concluded that the probability of Kobe making a shot is dependent upon either shot distance, playoff status, or both.

Now that it is established that one of the factors of interest are associated with the probability of Kobe making a shot, maximum likelihood estimators (MLE) can be used to build an appropriate model. MLEs use parameters in a model that maximize the probability of observing the outcome in the sample data that is used in the training set (Figure 4.4). In the full model, the coefficients for the interaction term (shot distance\*playoff status) and playoff status were not statistically significant (p-value>0.20). Only the coefficient for shot distance was statistically significant (p-value<0.0001). Before concluding that playoff status has no effect on the log-odds of Kobe making a shot, the interaction term was removed from the model leaving only shot distance and playoff status (Figure 4.5). The resulting MLE for the playoff term had an even p-value (0.6228) without the interaction term.

In conclusion, the log-odds of Kobe making a shot is different depending on the distance to the hoop, but it is not different depending on if he is in a playoff or regular season game. The only appropriate model only includes the shot distance explanatory variable:

**Log-odds(shot\_made\_flag) = 0.3680 -0.0441shot\_distance + error**

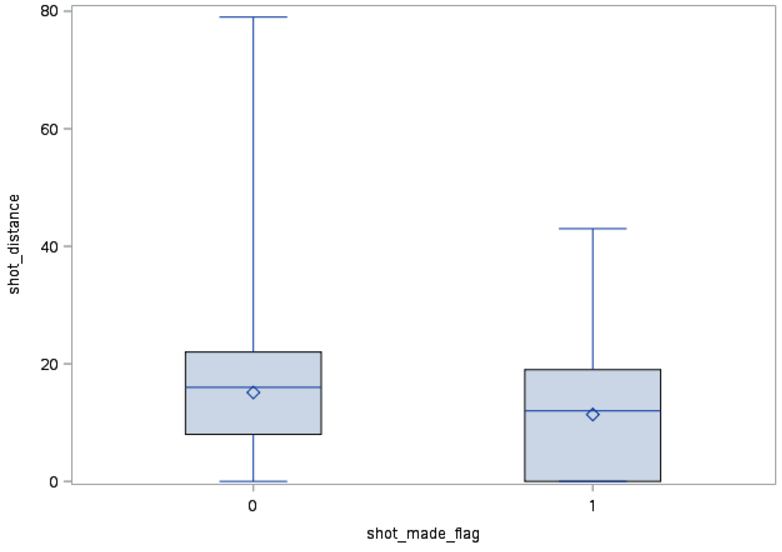
This model can be interpreted in terms of log-odds: As Kobe’s distance to the goal increases by one unit, the log-odds of him making the shot decreases by 0.0441. It may be more useful to understand the model in terms of probability with the use of a simple conversion. In this example, we will calculate the probability of Kobe making the shot at a distance of 10 units:

**Log odds(π)=0.3680 – 0.0441(10)**

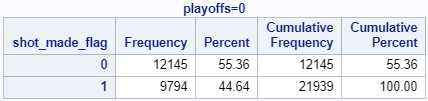
**π = e0.3680 – 0.0441(10)/ 1+ e0.3680 – 0.0441(10) = 0.482**

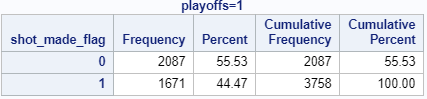
In a practical view, it appears that Kobe likely does not let the pressure to perform in a playoff game effect how he plays, but the distance to the goal will never be independent of the probability that he makes a shot.

**Figure 4.1: Distribution of distances for shots not made (0) and shots made (1)**

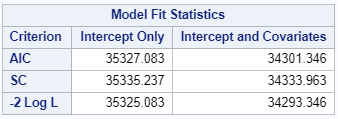


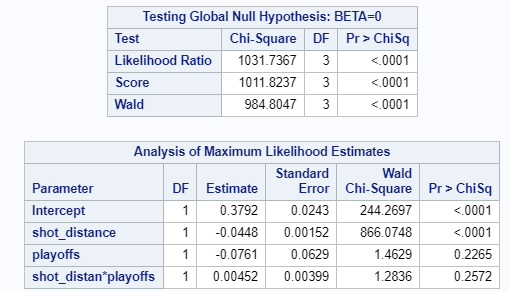
**Figure 4.2: Frequency of shots made and missed in regular season (playoffs=0) and during playoffs(playoffs=1)**



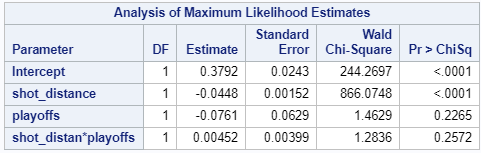


**Figure 4.3: Test statistics for goodness of fit**

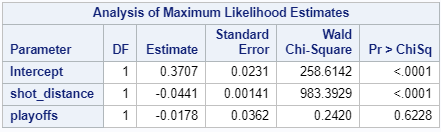




**Figure 4.4: MLE estimates for full model including interaction term**



**Figure 4.5: MLE estimates for model excluding interaction term**



# Predictive Model.

The data for the Kobe Bryant Kaggle completion is a time series having twenty years of historical shot records. It only makes sense to train the model on data recorded prior to the shot we are predicting and because of Kobe’s skills could be changing over time we use a rolling window of one year (365 days). This window is used for a traning dataset before every day we are making a prediction. We move the window every day we are making predictions.

To create the “rolling window” we use first extract all unique dates that have shot\_made\_flag equal to null (dates on which we have to predict) and we store them into a dataset called “dates”. Since the date type in SAS is just a number, we create the rolling window to be in between the current row from “dates” dataset and the same value minus 365. In this case we are training and predicting 1457 times in a loop.

We use logistic regression for making prediction because we are predicting a binary feature: “Shot Made” or “Shot Not Made” represented by 1 or 0. Because of the data in the rolling window is only fraction of the whole dataset, the categorical features will not include all of the categories and we found that using only the continuous features is giving us better score. Specifically, we are using minutes\_remaining, period, seconds\_remaining, and shot\_distance as features for the model.

The output from the logistic regression is in a form of logits: a probability in between 0 and 1 representing how likely is “Shot Made”. Normally, we would transform the output logits to be 0 if probability is less than .5 and 1 if the probability is greater than .5. However, the Kaggle competition requires a submission file with shot IDs and probability, so we don’t make any transformations of the logits. After each train/predict iteration, we collect the output of the logistic regression and finally we export the fields needed to a CSV file. We achieved a Kaggle score of 0.77881.

# Appendix

# Model 1

1. PROC IMPORT DATAFILE='/home/iangelov0/project3/data.csv' replace
2. DBMS=CSV
3. OUT=data;
4. GETNAMES=YES;
6. data data;
7. set data;
8. if substr(matchup, 5, 1) = '@' then play\_home = 0;
9. else play\_home = 1;
11. proc logistic datadata=data;
12. model shot\_made\_flag(event ='1') = play\_home;
13. run;

# Model 2

FILENAME REFFILE '/home/sgozdzialski0/Kobe.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=Kobe;

GETNAMES=YES;

RUN;

data Kobe;

set Kobe;

IF shot\_made\_flag = "." THEN DELETE;

run;

proc logistic data = Kobe;

"Kobe Shots": model shot\_made\_flag (event = '1') = shot\_distance;

run;

# Model 4.

\*Author: Lauren Darr

\*Import original data;

proc import

datafile='/home/ldarr1/Stats2\_Project3/kobedata.csv'

out=Kobe\_Full

dbms=CSV

replace;

getnames=yes;

datarow=2;

guessingrows=2000;

run;

\*Split data into train and test sets;

DATA Kobe\_Train Kobe\_Test;

SET Kobe\_Full;

IF shot\_made\_flag='0' or shot\_made\_flag='1' THEN OUTPUT Kobe\_Train;

ELSE OUTPUT Kobe\_Test;

RUN;

proc sort data = Kobe\_Train;

by shot\_made\_flag;

run;

\*Prelim look at distribution of shot\_distance v. shot\_made\_flag;

proc boxplot data = Kobe\_Train;

plot shot\_distance\*shot\_made\_flag;

run;

data playoff\_kobe;

set Kobe\_Train;

keep playoffs shot\_made\_flag;

run;

\*Prelim look at frequency of shots made v. playoff status;

Proc sort data=playoff\_kobe;

by playoffs;

run;

proc freq data =playoff\_kobe;

by playoffs;

run;

title 'Logistic Regression-Distance and Playoffs';

proc logistic data=Kobe\_Train;

model shot\_made\_flag(event='1')= shot\_distance playoffs shot\_distance\*playoffs/scale=none;

output out=No4LogRegOut predprobs=I p=probpreb;

run;

proc logistic data=kobe\_train;

model shot\_made\_flag(event='1')= shot\_distance playoffs/scale=none;

output out=No4LogRegOut2 predprobs=I p=probpreb;

run;

proc logistic data=kobe\_train;

model shot\_made\_flag(event='1') = shot\_distance/scale=none;

output out=No4LogRegOut3 predprobs=I p=probpreb;

run;

Predictive Model

1. PROC IMPORT DATAFILE='/home/iangelov0/project3/data.csv' replace
2. DBMS=CSV
3. OUT=data;
4. GETNAMES=YES;
6. proc sort data=data out=data;
7. by game\_date;
9. data data;
10. set data;
11. date=input(game\_date,yymmddd10.);
13. proc sql;
14. create table dates as
15. select distinct date
16. from data where data.shot\_made\_flag is null;
18. data result;
19. set data(obs=0);
20. logits = 0;
21. logits\_backup = 0;
23. %MACRO LOOP\_DATA;
24. /\* 1457 \*/
25. %DO i = 1 %TO 1457;
26. DATA current\_date;
27. SET dates(obs=&i firstobs=&i);
29. proc sql;
30. create table rolling\_data as
31. select \* from data a
32. inner join current\_date b
33. on  (a.date **>** b.date-365 and a.date **<** **b.date** and a.shot\_made\_flag is not null)
34. or (a.date = b.date);
36. proc logistic data=rolling\_data noprint;
37. model shot\_made\_flag = minutes\_remaining period seconds\_remaining shot\_distance;
38. output out=predicted\_backup(where=(shot\_made\_flag=.)) p=logits;
40. data predicted;
41. merge predicted predicted\_backup;
43. proc delete data=rolling\_data;
44. proc append base=result data=predicted force;
45. %END;
46. %MEND LOOP\_DATA;
47. %LOOP\_DATA;
49. data result;
50. set result;
51. shot\_made\_flag=logits;
52. keep shot\_id shot\_made\_flag;
54. proc export data=result dbms=csv
55. outfile='/home/iangelov0/project3/submition.csv'
56. replace;
57. run;