Assignment 3: Melvin's Performance Anaylsis

Jingyi(Abby) Liu

9/14/2018

First, clean the data by transform the column 'goal' into numerical. Replace 'Y' or 'y' by 1 and replace 'N' by 0.

```
#Import the data
df<-read.csv("kicksfootball.csv", header = TRUE, stringsAsFactors = FALSE)

#Clean the data
df[df == 'Y'|df == 'y']<-1
df[df == 'N']<-0
df$goal <- as.numeric(df$goal)
```

Part A

How would you describe Melvin's overall record? 1. In general, the probability that Melvin will hit the goal.

```
P1 = (nrow(subset(df, goal == 1)))/(nrow(df))
P1 = sprintf("%.4f", P1)
paste("Probability that Melvin will hit the goal",P1)
```

```
## [1] "Probability that Melvin will hit the goal 0.7973"
```

2. Will the probabilty be influenced by the nature of the attempts?(practice or match?)

```
#Probability that Melvin will hit the goal on a match
Pm = (nrow(subset(df, practiceormatch == 'M'& goal == 1)))/
  (nrow(subset(df, practiceormatch == 'M')))
Pm = sprintf("%.4f", Pm)
paste("Probability that Melvin will hit the goal on a match is",Pm)
```

```
## [1] "Probability that Melvin will hit the goal on a match is 0.7590"
```

```
#Probability that Melvin will hit the goal on a practice
Pp = (nrow(subset(df, practiceormatch == 'P'& goal == 1)))/
  (nrow(subset(df, practiceormatch == 'P')))
Pp = sprintf("%.4f", Pp)
paste("Probability that Melvin will hit the goal on a practice is",Pp)
```

```
## [1] "Probability that Melvin will hit the goal on a practice is 0.8009"
```

Melvin's rate of success on a practice is sligtly higher than that on a game.

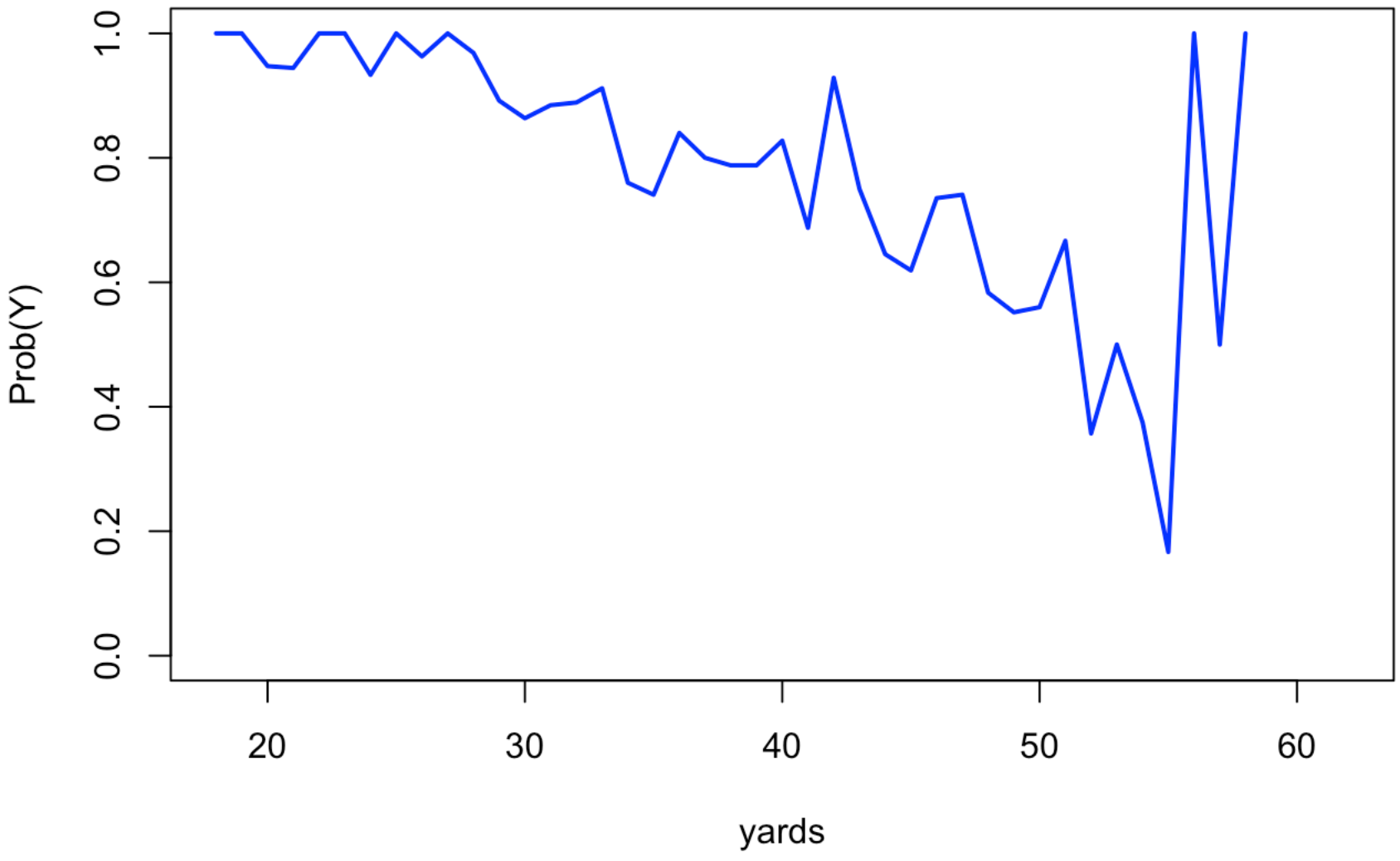
3. Explore how the probability is influenced by the distance.

```
Yards=seq(18,62,by=1)
Prab=rep(0,length(18:62))
#Output the probility of 'Y' under each 'yards'
for (i in 18:62) {
  a = subset(df, yards == i)
  Prab[i] = (nrow(subset(a, goal == 1)))/(nrow(a))
}
data.frame(Yards, Prab[18:62])
```

```
##    Yards Prab.18.62.
## 1     18   1.0000000
## 2     19   1.0000000
## 3     20   0.9473684
## 4     21   0.9444444
## 5     22   1.0000000
## 6     23   1.0000000
## 7     24   0.9333333
## 8     25   1.0000000
## 9     26   0.9629630
## 10    27   1.0000000
## 11    28   0.9687500
## 12    29   0.8918919
## 13    30   0.8636364
## 14    31   0.8846154
## 15    32   0.8888889
## 16    33   0.9117647
## 17    34   0.7600000
## 18    35   0.7407407
## 19    36   0.8400000
## 20    37   0.8000000
## 21    38   0.7878788
## 22    39   0.7878788
## 23    40   0.8275862
## 24    41   0.6875000
## 25    42   0.9285714
## 26    43   0.7500000
## 27    44   0.6451613
## 28    45   0.6190476
## 29    46   0.7352941
## 30    47   0.7407407
## 31    48   0.5833333
## 32    49   0.5517241
## 33    50   0.5600000
## 34    51   0.6666667
## 35    52   0.3571429
## 36    53   0.5000000
## 37    54   0.3750000
## 38    55   0.1666667
## 39    56   1.0000000
## 40    57   0.5000000
## 41    58   1.0000000
## 42    59          NaN
## 43    60   0.0000000
## 44    61          NaN
## 45    62   0.0000000
```

In general, with the increasement of ‘yards’, the probability that Melvin will hit the goal decreased.

```
# Visualize the relationship between yards and probability of scoring
plot(Yards, Prab[18:62], type="l",
     xlab="yards",
     ylab="Prob(Y)",
     col="blue",lwd=2)
```



Part B

.Write out the logistic function for: (1)Practices

The logistic function for practice = log (P(y=1)/(1−P(y=1))) =b0+b1x, For our equation b0 = 5.58180 and b1 = -0.10672.

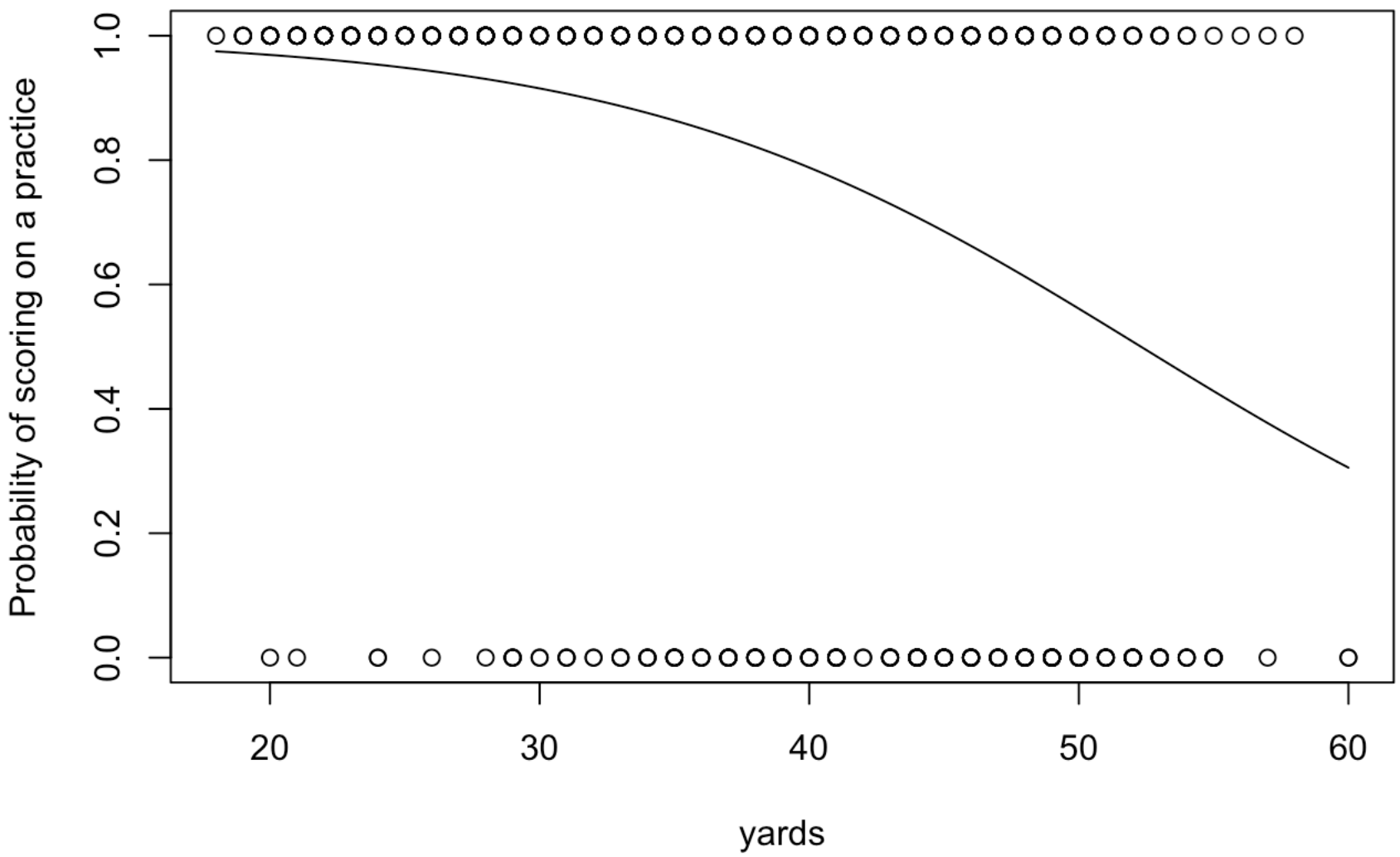
```r
#Select all the practice
df_p = subset(df, practiceormatch == 'P')

# fit a logistic regression model
practice.log = glm(goal ~ yards, data=df_p,
             family=binomial)

#See the results contained in goal.log
summary(practice.log)
```

```
##
## Call:
## glm(formula = goal ~ yards, family = binomial, data = df_p)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6377   0.2780   0.4207   0.6903   1.4441
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.58180    0.46705  11.951   <2e-16 ***
## yards       -0.10672    0.01099  -9.709   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 867.48  on 868  degrees of freedom
## Residual deviance: 748.93  on 867  degrees of freedom
## AIC: 752.93
##
## Number of Fisher Scoring iterations: 5
```

```r
# plot with yards on x-axis and sccoring or not (0 or 1) on y-axis
plot(goal~yards,data=df_p,
     xlab="yards", ylab="Probability of scoring on a practice")
curve(predict(practice.log, data.frame(yards=x),
              type="resp"),
      add=TRUE)
```



(2)Matches

The logistic function for practice = log (P(y=1)/(1−P(y=1))) =b0+b1x, For our equation b0 = 6.83393 and b1 = -0.14147 .

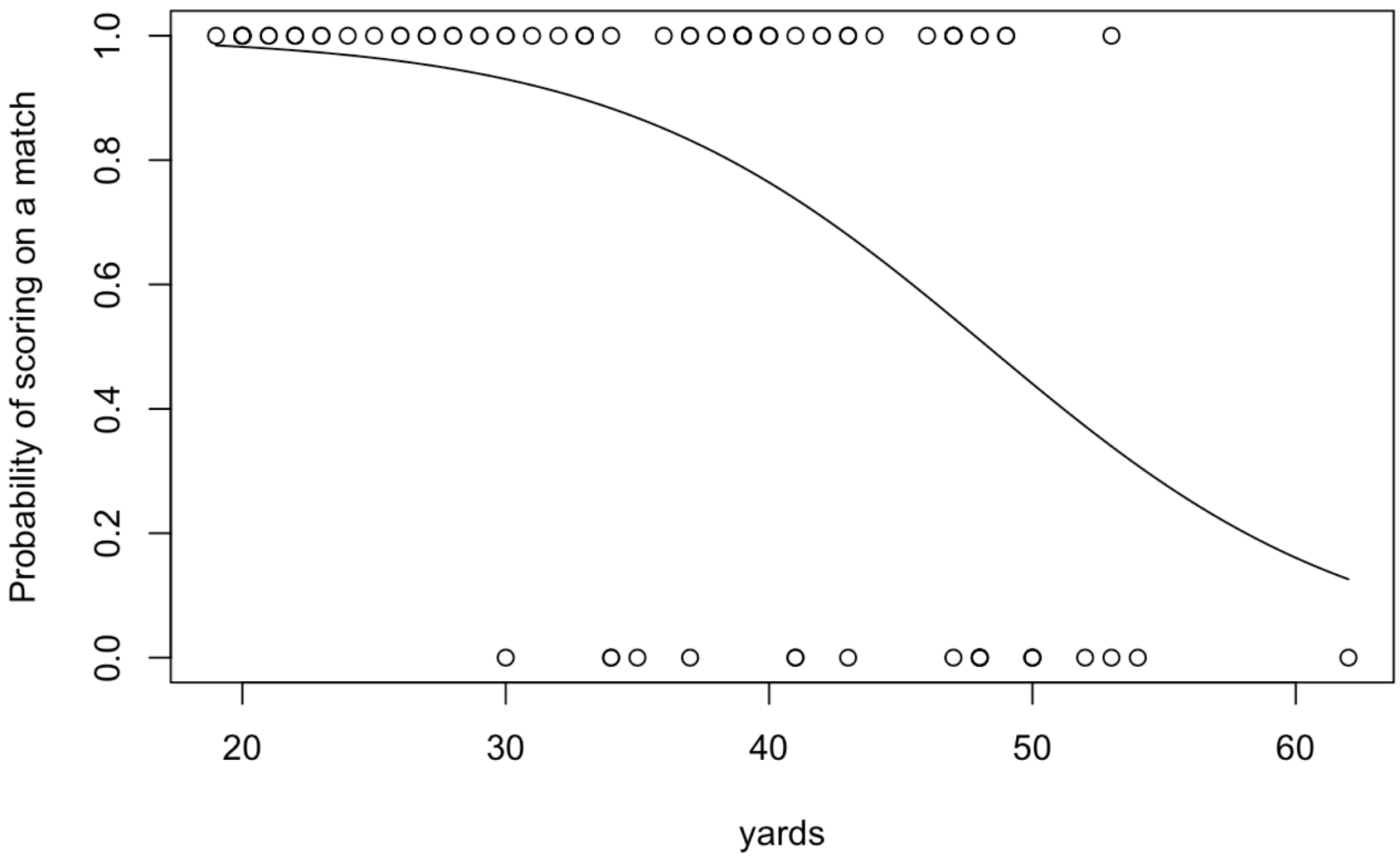
```
#Select all the match
df_m = subset(df, practiceormatch == 'M')

# fit a logistic regression model
match.log = glm(goal ~ yards, data=df_m,
              family=binomial)

#See the results contained in goal.log
summary(match.log)
```

```
##
## Call:
## glm(formula = goal ~ yards, family = binomial, data = df_m)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3074   0.1837   0.3553   0.6892   1.4692
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.83393    1.69900   4.022 5.76e-05 ***
## yards       -0.14147    0.03922  -3.607  0.00031 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 91.663  on 82  degrees of freedom
## Residual deviance: 71.523  on 81  degrees of freedom
## AIC: 75.523
##
## Number of Fisher Scoring iterations: 5
```

```
# plot with yards on x-axis and sccoring or not (0 or 1) on y-axis
plot(goal~yards,data=df_m,
     xlab="yards", ylab="Probability of scoring on a match")
curve(predict(match.log, data.frame(yards=x),
              type="resp"),
      add=TRUE)
```



b)

What is the probability of Melvin scoring a goal when he kicks from 20, 40 and 60 yards in practice?
Answer:The probabilities are 0.9692, 0.7880 and 0.3054

```
inp <- c(20,40,60)
newdata = data.frame(yards=inp)

predict(practice.log, newdata, type="response")
```

```
##         1         2         3
## 0.9691526 0.7880037 0.3054452
```

c. What is the probability of Melvin scoring a goal when he kicks from 20, 40 and 60 yards in matches?
   Answer:The probabilities are 0.9821, 0.7641 and 0.1605

```
inp <- c(20,40,60)
newdata = data.frame(yards=inp)

predict(match.log, newdata, type="response")
```

```
##         1         2         3
## 0.9820933 0.7640656 0.1605269
```

Part C

Plot the logistic models

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
yards_ <- c(18:62)
Probability = predict(practice.log, data.frame(yards=c(yards_)), type="response")
P_match = predict(match.log, data.frame(yards=c(yards_)), type="response")

df1 <- data.frame(yards_,Probability )
g <- ggplot(df1)
g <- g + geom_line(aes(x=yards_,y=Probability ,color='practice'))
g <- g + geom_line(aes(x=yards_,y=P_match,color='match'))
print(g)
```