به نام او
پروژهٔ تحلیل رگرسیون
دکتر صفدری
نیمسال دوم ۹۸ـ۹۹

محمدحسین کلهرجهاندوست
۹۵۱۰۹۹۶۱

---

* کتابخانه های مورد استفاده در طول تمرین:

```
> library(MASS)
> library(dplyr)
> library(readr)
> library(ggplot2)
> library(ISLR)
> library(class)
> library(gam)
> library(ggthemes)
> library (leaps)
> library(ROCR)
```

* seed را برابر ۱۰۰ قرار می دهیم:

```
> set.seed(100)
```

* توجه کنید که محیط RStudio از نیم فاصله پشتیبانی نمی کند. بنابراین در متن فارسی نیم فاصله رعایت نشده است. با توجه به این موضوع به جای «ی» بدل از کسره، در مواردی که به «هـ» پایانی ختم می شود از «همزه» استفاده شده است.

۱ طبقه بندی

۱.۱ خواندن و توضیح داده

ابتدا داده را می خوانیم و تغییرات لازم را انجام می دهیم.

```
> data = read_csv("epl.csv")
> epl = data %>% select(-X1)
> epl$FTR = ifelse(epl$FTR == 'H',0,1)
> epl$HTR = ifelse(epl$HTR == 'H',0,1)
```

برای طبقه بندی، دادهٔ لیگ برتر انگلیس از استفاده شده است. با توجه به این که هر فصل داده های خاص خودش را داشت، از ستون های مشترکی که در همهٔ فصل ها بود استفاده کردیم و یک دادهٔ جدید با ۷۲۶۰ مشاهده و ۲۲ ستون به نام epl ساختیم. جدول ۱ توضیح ستون های داده است.

نام ستون	توضیح
Date	تاریخ برگزاری بازی
HomeTeam	نام تیم میزبان
AwayTeam	نام تیم میهمان
FTHG	تعداد گل زدهٔ تیم میزبان در کل بازی
FTAG	تعداد گل زدهٔ تیم میهمان در کل بازی
FTR	نتیجهٔ نهایی بازی
HTHG	تعداد گل زدهٔ تیم میزبان در نیمهٔ اول
HTHG	تعداد گل زدهٔ تیم میهمان در نیمهٔ اول
HTR	نتیجهٔ بازی در نیمهٔ اول
Refree	نام داور مسابقه
HS	تعداد ضربه های تیم میزبان
AS	تعداد ضربه های تیم میهمان
HST	تعداد ضربه های درچارچوب تیم میزبان
AST	تعداد ضربه های درچارچوب تیم میهمان
HC	تعداد کرنرهای تیم میزبان
AC	تعداد کرنرهای تیم میهمان
HF	تعداد خطاهای تیم میزبان
AF	تعداد خطاهای تیم میهمان
HY	تعداد کارت های زرد تیم میزبان
AY	تعداد کارت های زرد تیم میهمان
HR	تعداد کارت های قرمز تیم میزبان
AR	تعداد کارت های قرمز تیم میهمان

جدول ۱: ستون های دادهٔ epl

۲.۱ توضیح مسئله

شاید بتوان گفت لیگ برتر انگلیس، جذاب ترین لیگ فوتبال است. مقایسهٔ برخی آمارها مثل تعداد دنبال کنندگان نیز این موضوع را نشان می دهد. این توصیف از لیگ برتر انگلیس بسیار متداول است که در یک بازی هیچ تیمی از پیش باخته نیست؛ هر چند این بازی بین تیم صدر و قعر جدول باشد.

یک سوال طبیعی در مواجهه با یک مسابقه که بین دو تیم برگزار می شود این است که در پایان مسابقه چه تیمی برنده خواهد شد. البته این سوال صورت بندی دقیقی ندارد. چون در مسئلهٔ طبقه بندی، متغیر هدف چند سطح مشخص دارد و مدلی که روی داده برازش می شود مشخص می کند که اگر متغیرهای پیش بینی کننده مقدار مشخصی بگیرند در این صورت متغیر هدف در کدام دسته قرار خواهد گرفت. بنابراین می توان صورت سوال رو به این شکل درآورد: برندهٔ بازی تیم مهمان است یا میزبان یا بازی بدون برنده (مساوی) است؟ تحت چه شرایطی هر کدام از حالت های گفته شده رخ می دهد؟ بنابراین در مسئلهٔ طبقه بندی متغیر هدف ما ستون FTR خواهد بود.

۳.۱ کاوش داده و انتخاب متغیرهای پیشگو

برای انتخاب پیشگوها از برخی از نمودارها و مدل ها کمک می گیریم. در ابتدا باید برخی از ستون ها که مشخصا به طور مستقیم روی ستون هدف اثر می گذارند را کنار بگذاریم. مثلا ستون هدف یعنی FTR دقیقا از تفاضل گل های زدهٔ تیم میزبان و میهمان در کل بازی، یعنی دو ستون FTHG و FTAG، مشخص می شود. چون تیم برنده تیمی است که گل زدهٔ بیشتری داشته باشد. بنابراین ستون های زیر را کنار می گذاریم:

```
> epldf = epl %>% select(-Div,-Date,-Referee,
+                        -FTHG,-FTAG,-HTHG,-HTAG
+                        ,-HomeTeam,-AwayTeam,-HTR)
>
```

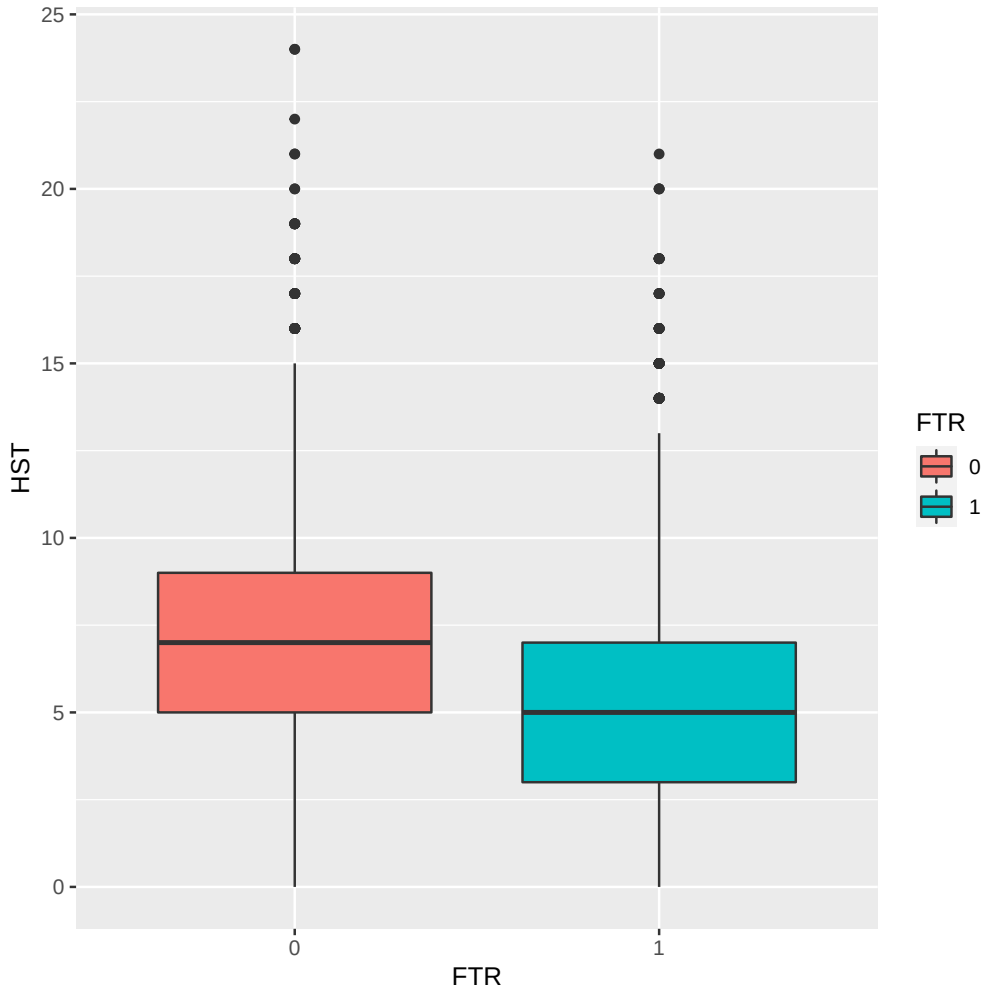
قبل از استفاده از ابزارهای آماری، سعی می کنیم به عنوان یک بینندهٔ فوتبال حدس بزنیم کدام یک از ستون های داده، با ستون هدف ارتباط دارد. برای این کار باید حدس بزنیم که هر کدام از ستون ها چه ارتباطی با افزایش گل (یا گل خوردن) دارد زیرا چیزی که در نهایت برنده را مشخص می کند تفاضل گل دو تیم است. مثلا ضربات در چارچوب رابطهٔ جدی با گل زدن دارد. به این معنی که هر چقدر ضربات در چارچوب بیشتر باشد احتمالا گل زده هم بیشتر می شود و بنابراین شانس برنده شدن بیشتر می شود. یا مثلا هر چقدر کرنر (و به طور عمومی تر ضربات شروع مجدد که اطلاعتی از آن ها در داده نیست) بیشتر باشد حدس می زنیم گل زده هم بیشتر شود زیرا کرنر جز موقعیت های مناسب گل زنی به شمار می رود. در مورد خطاها می توان گفت هر چقدر خطاها بیشتر باشد احتمال باختن بیشتر است. زیرا خطاها ممکن است در قسمت هایی از زمین انجام شود که موقعیت مناسبی برای گل کردن ضربات ایستگاهی باشد و به طور خاص مثلا پنالتی. در مورد تعداد کارت های زرد هم همین تحلیل وجود دارد. در مورد کارت قرمز علاوه بر تحلیل گفته شده حدس می زنیم تاثیر بیشتر باشد. زیرا با اخراج شدن یک بازیکن از زمین مسابقه فشار تیم مقابل بیشتر می شود و احتمال برنده شدن تیم مقابل بیشتر می شود. حالا با این توضیحات شاید مناسب باشد همبستگی ستون هدف با سایر ستون های داده را ببینیم:

```
> cor(epldf)[1,]
```

```
        FTR          HS          AS         HST         AST          HC
 000000000.1 196268376.0-  196560049.0 289861639.0-  240816800.0 019318310.0-
         AC          HF          AF          HY          AY          HR
 023875396.0  050114134.0  003470445.0 132118381.0  004202903.0 113683841.0
         AR
082430259.0-
```

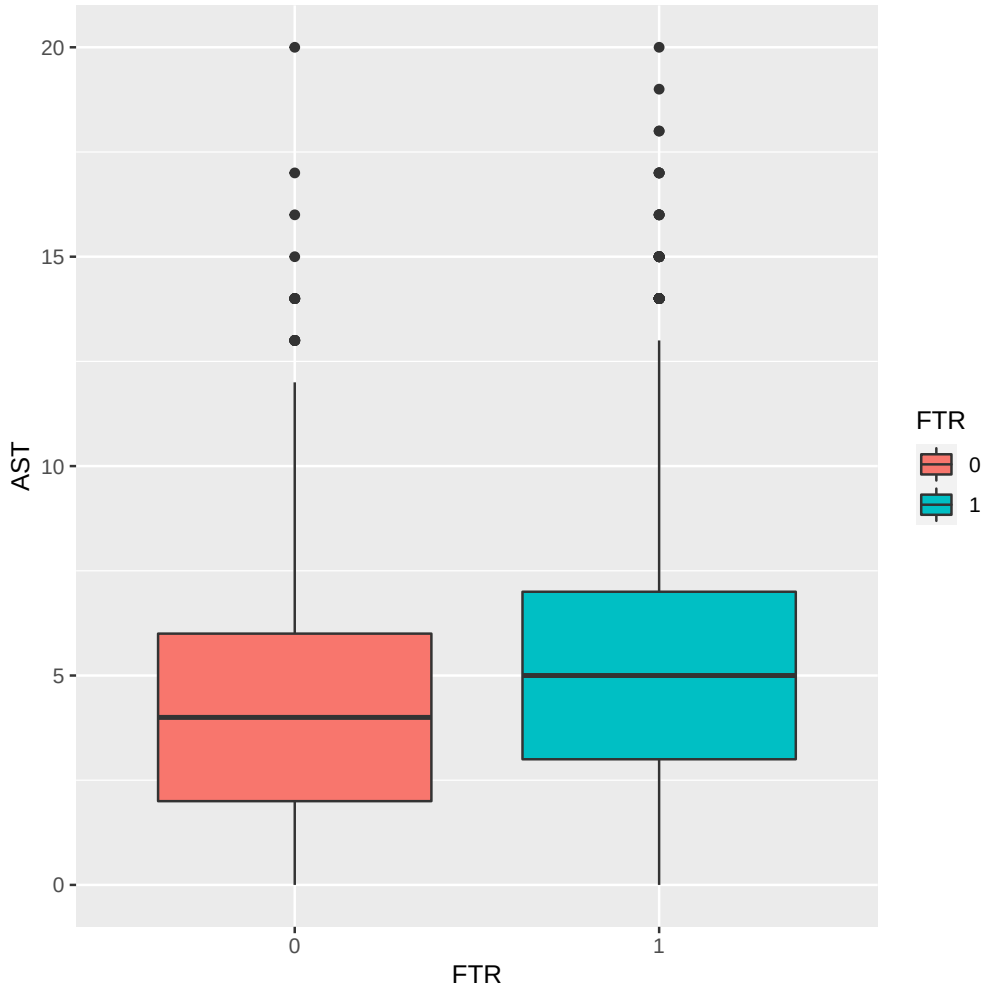
نتایج بالا نشان می دهد حدس ما دربارهٔ ضربات در چارچوب تا حدی درست بوده است. یعنی ضربات در چارچوب تیم میزبان هر چقدر بیشتر باشد احتمال برد تیم میهمان کمتر است و به طور مشابه ضربات در چارچوب تیم میهمان هر چقدر بیشتر باشد احتمال برد تیم میهمان بیشتر است. چند نمودار در این رابطه در ادامه آمده است. نمودار اول نمودار جعبه ای تعداد ضربات در چارچوب تیم میزبان بر حسب بردن یا نبردن تیم میهمان است:

```
> ggplot(data = epl,aes(x=as.factor(FTR),y=HST))+
+   geom_boxplot(aes(fill = factor(FTR)))+
+   labs(y="HST",x="FTR")+scale_fill_discrete(name = "FTR")
```



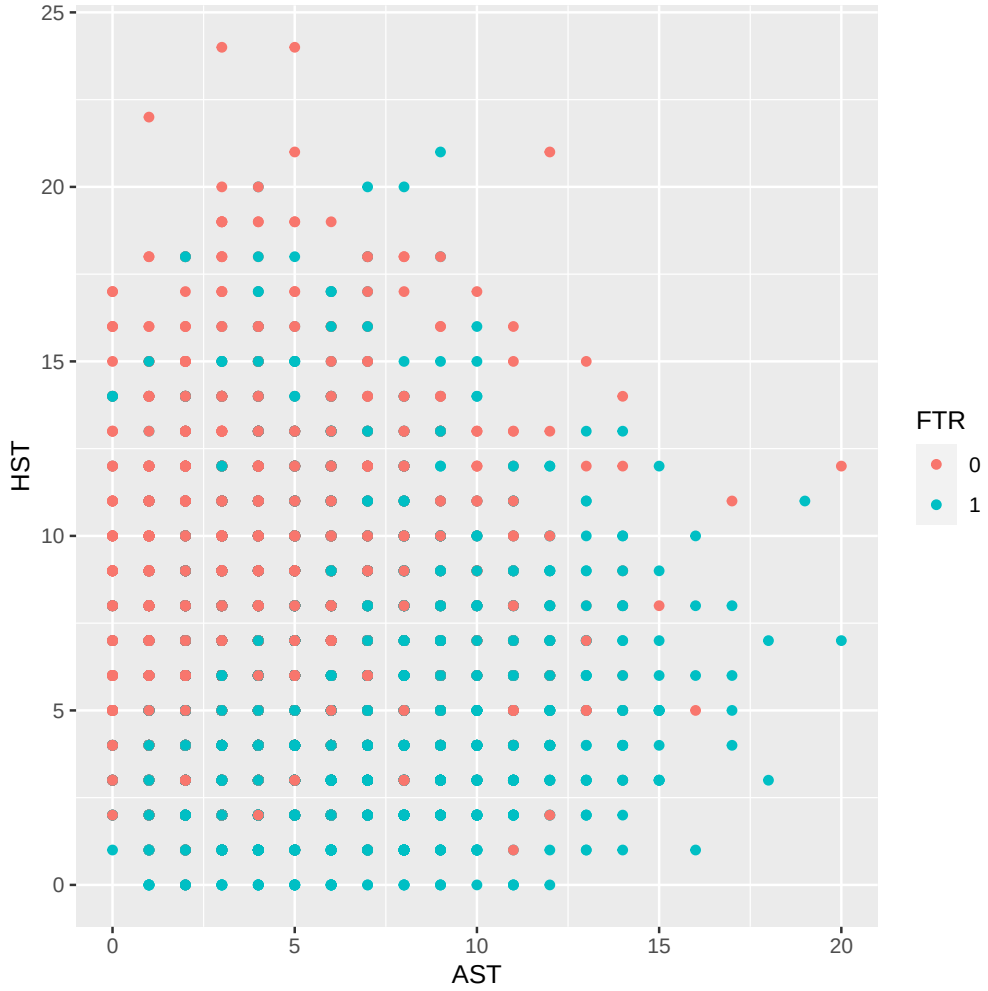
همانطور که قابل مشاهده است، میانگین ضربات در چارچوب تیم میزبان در هنگام برد میزبان بیشتر از میانگین ضربات در چارچوب تیم میزبان در زمان باخت تیم میزبان است که با شهود اولیهٔ ما سازگار است. نمودار دوم نیز نمودار جعبه ای تعداد ضربات در چارچوب تیم میهمان بر حسب بردن یا نبردن تیم میزبان است:

```
> ggplot(data = epl,aes(x=as.factor(FTR),y=AST))+
+   geom_boxplot(aes(fill = factor(FTR)))+
+   labs(y="AST",x="FTR")+
+   scale_fill_discrete(name = "FTR")
```



مانند نمودار قبل نیز این نمودار با حس اولیهٔ ما سازگار است. میانیگین ضربات در چارچوب تیم میهمان در زمان باخت تیم میزبان کمتر از این میانگین در زمان نبردن تیم میزبان است. نمودار بعدی نیز نمودا نقطه ای HST بر حسب AST است و رنگ نقاط نیز نشان دهندهٔ بردن یا نبردن میزبان (FTR) است:

```
> ggplot(data = epl,aes(x=AST,y=HST))+
+   geom_point(aes(col=as.factor(FTR)))+
+   labs(y="HST",x="AST")+
+   scale_color_discrete(name = "FTR")
```



این نمودار هم نشان می دهد که در زیر خط $x = y$ که ضربات در چارچوب مهمان بیشتر از میزبان است غالبا تیم میزبان برنده نبوده است. همچنین در بالای خط $x = y$ که ضربات در چارچوب میزبان بیشتر از میهمان بوده است غالبا میزبان برنده بوده است.

اما در مورد بقیهٔ ستون ها نتیجهٔ محاسبهٔ همبستگی ناامید کننده است. به این معنی که شاید بقیهٔ ستون ها توضیح دهندهٔ خیلی خوبی برای متغیر هدف نیستند. برای بهتر مشخص شدن این موضوع یک رگرسیون لجستیک روی داده برازش می کنیم تا ببینیم ضرایب کدام متغیرها از لحاظ آماری معنادار نیستند.

```
> model = glm(FTR~.,data = epldf,family = "binomial")
> summary(model)

Call:
glm(formula = FTR ~ ., family = "binomial", data = epldf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
7310.2-  0200.1-   4924.0   9717.0   4362.2
```

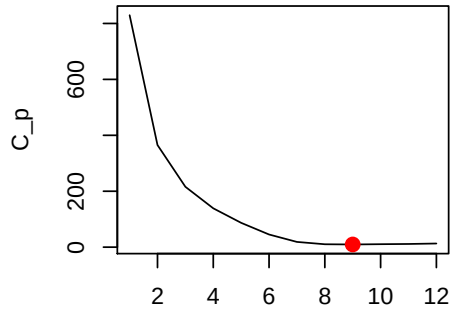
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 051075.0-   174997.0  292.0- 770392.0
HS           026140.0   007747.0   374.3 000741.0 ***
AS          014377.0-   009169.0  568.1- 116871.0
HST         262003.0-   012219.0 443.21-  < 2e16- ***
AST          262002.0   014534.0  027.18  < 2e16- ***
HC           090736.0   009848.0   214.9  < 2e16- ***
AC          088819.0-   011017.0  062.8- 51e.716- ***
HF          003920.0-   007513.0  522.0- 601861.0
AF           010247.0   007202.0   423.1 154837.0
HY           169157.0   024300.0   961.6 37e.312- ***
AY          027262.0-   022335.0  221.1- 222233.0
HR           855720.0   118186.0   240.7 47e.413- ***
AR          488926.0-   089394.0  469.5- 52e.408- ***
---
Signif. codes:  0 ‘***’ 001.0 ‘**’ 01.0 ‘*’ 05.0 ‘.’ 1.0 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.10024  on 7259  degrees of freedom
Residual deviance:  6.8579  on 7247  degrees of freedom
AIC: 6.8605

Number of Fisher Scoring iterations: 4
```

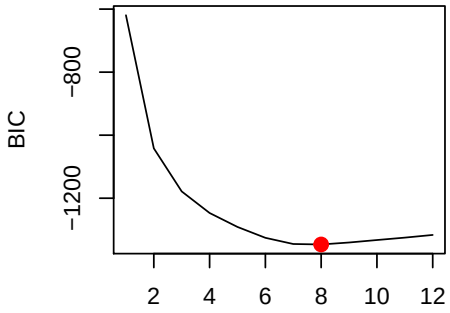
با توجه به نتیجهٔ مدل، ضرایب HS، HST، AST، HC، AC، AS، HY، HR و AR از لحاظ آماری معنادار هستتند. در واقع با این کار می توانیم بقیهٔ ضرایب را که از لحاظ آماری معنادار نیستند، کنار بگذاریم اما در مورد آن ها که معنی دار هستند نمی توانیم به طور قطعی نظری دهیم. اگر از روش forward selection استفاده کنیم به نتیجهٔ مشابه می رسیم:

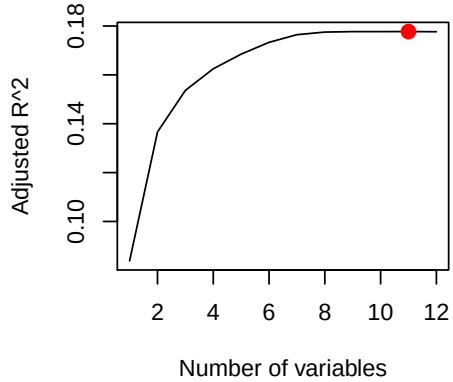
```
> regfit.fwd = regsubsets(FTR~., data= epldf , nvmax =19,
+                         method ="forward")
> summary(regfit.fwd)

Subset selection object
Call: regsubsets.formula(FTR ~ ., data = epldf, nvmax = 19, method = "forward")
12 Variables  (and intercept)
    Forced in Forced out
HS      FALSE      FALSE
AS      FALSE      FALSE
HST     FALSE      FALSE
AST     FALSE      FALSE
HC      FALSE      FALSE
AC      FALSE      FALSE
HF      FALSE      FALSE
AF      FALSE      FALSE
HY      FALSE      FALSE
AY      FALSE      FALSE
HR      FALSE      FALSE
AR      FALSE      FALSE
```

```
1 subsets of each size up to 12
Selection Algorithm: forward
          HS  AS  HST AST HC  AC  HF  AF  HY  AY  HR  AR
1  ( 1 )  " " " " "*" " " " " " " " " " " " " " " " " " "
2  ( 1 )  " " " " "*" "*" " " " " " " " " " " " " " " " "
3  ( 1 )  " " " " "*" "*" "*" " " " " " " " " " " " " " "
4  ( 1 )  " " " " "*" "*" "*" "*" " " " " " " " " " " " "
5  ( 1 )  " " " " "*" "*" "*" "*" " " " " "*" " " " " " "
6  ( 1 )  " " " " "*" "*" "*" "*" " " " " "*" " " "*" " "
7  ( 1 )  " " " " "*" "*" "*" "*" " " " " "*" " " "*" "*"
8  ( 1 )  "*" " " "*" "*" "*" "*" " " " " "*" " " "*" "*"
9  ( 1 )  "*" "*" "*" "*" "*" "*" " " " " "*" " " "*" "*"
10  ( 1 ) "*" "*" "*" "*" "*" "*" " " "*" "*" " " "*" "*"
11  ( 1 ) "*" "*" "*" "*" "*" "*" " " "*" "*" "*" "*" "*"
12  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

```
> mod.summary = summary(regfit.fwd)
> par(mfrow = c(2, 2))
> plot(mod.summary$cp, xlab = "Number of variables",
+      ylab = "C_p", type = "l")
> points(which.min(mod.summary$cp),
+        mod.summary$cp[which.min(mod.summary$cp)],
+        col = "red", cex = 2, pch = 20)
> plot(mod.summary$bic, xlab = "Number of variables",
+      ylab = "BIC", type = "l")
> points(which.min(mod.summary$bic),
+        mod.summary$bic[which.min(mod.summary$bic)],
+        col = "red", cex = 2, pch = 20)
> plot(mod.summary$adjr2, xlab = "Number of variables",
+      ylab = "Adjusted R^2",
+      type = "l")
> points(which.max(mod.summary$adjr2),
+        mod.summary$adjr2[which.max(mod.summary$adjr2)],
+        col = "red", cex = 2, pch = 20)
```







دقت کنید نتیجهٔ روش backward selection و best subset selection نیز مشابه همین بود. به نظر می رسد با توجه به نمودارهای بالا، انتخاب هشت پیشگو مناسب باشد. حتی به نظر می رسد می توان HS را هم کنار گذاشت. چون تقریبا با HST رابطه دارد و همانطور که از نتیجهٔ مدل لجستیک پیداست از بقیهٔ ضرایب معنادار کم اهمیت تر است و در روش forward selection نیز در مرحلهٔ هشتم اضافه شده است. پس دست آخر به هفت متغیر زیر به عنوان پیشگو می رسیم:

$$\mathrm{FTR} \sim \mathrm{HST} + \mathrm{AST} + \mathrm{HC} + \mathrm{AC} + \mathrm{HY} + \mathrm{HR} + \mathrm{AR}$$

```
> newModel = glm(FTR~HST+AST+HC+AC+HY+HR+AR,data = epldf,family = "binomial")
> summary(newModel)

Call:
glm(formula = FTR ~ HST + AST + HC + AC + HY + HR + AR, family = "binomial",
    data = epldf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
7546.2-  0233.1-   5017.0   9704.0   3595.2
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  098164.0   103423.0   949.0    343.0
HST         234268.0-   009388.0 955.24-  < 2e16- ***
AST          240599.0   011064.0  746.21  < 2e16- ***
HC           101472.0   009381.0  816.10  < 2e16- ***
AC          097163.0-   010535.0  223.9-  < 2e16- ***
HY           158952.0   022552.0   048.7 81e.112- ***
HR           825854.0   117322.0   039.7 93e.112- ***
AR          470520.0-   088668.0  307.5- 12e.107- ***
---
Signif. codes:  0 ‘***’ 001.0 ‘**’ 01.0 ‘*’ 05.0 ‘.’ 1.0 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.10024  on 7259  degrees of freedom
Residual deviance:  6.8595  on 7252  degrees of freedom
AIC: 6.8611

Number of Fisher Scoring iterations: 4
```

۴.۱ انتخاب مدل و محاسبهٔ دقت

برای انتخاب مدل تحلیل خاصی وجود ندارد که ما را برای استفاده از یک مدل خاص ترغیب کند یا بتوان با تقریب خوبی گفت که مدلی مناسب دادهٔ ما نیست. مثلا اگر رابطهٔ واقعی به خطی نزدیک باشد انتظار داریم روش های پارامتری مثل لجستیک و LDA بهتر از روش های غیرپارامتری مثل KNN عمل کنند. و برعکس اگر رابطهٔ واقعی خیلی غیرخطی باشد روش KNN عملکرد بهتری خواهد داشت. اما در حال حاضر حدس یا حسی نسبت به این «رابطهٔ واقعی» نداریم. در ادامه رگرسیون لجستیک، روش QDA و روش KNN را به کار می گیریم. چون روش LDA و روش لجستیک تقریبا مشابه

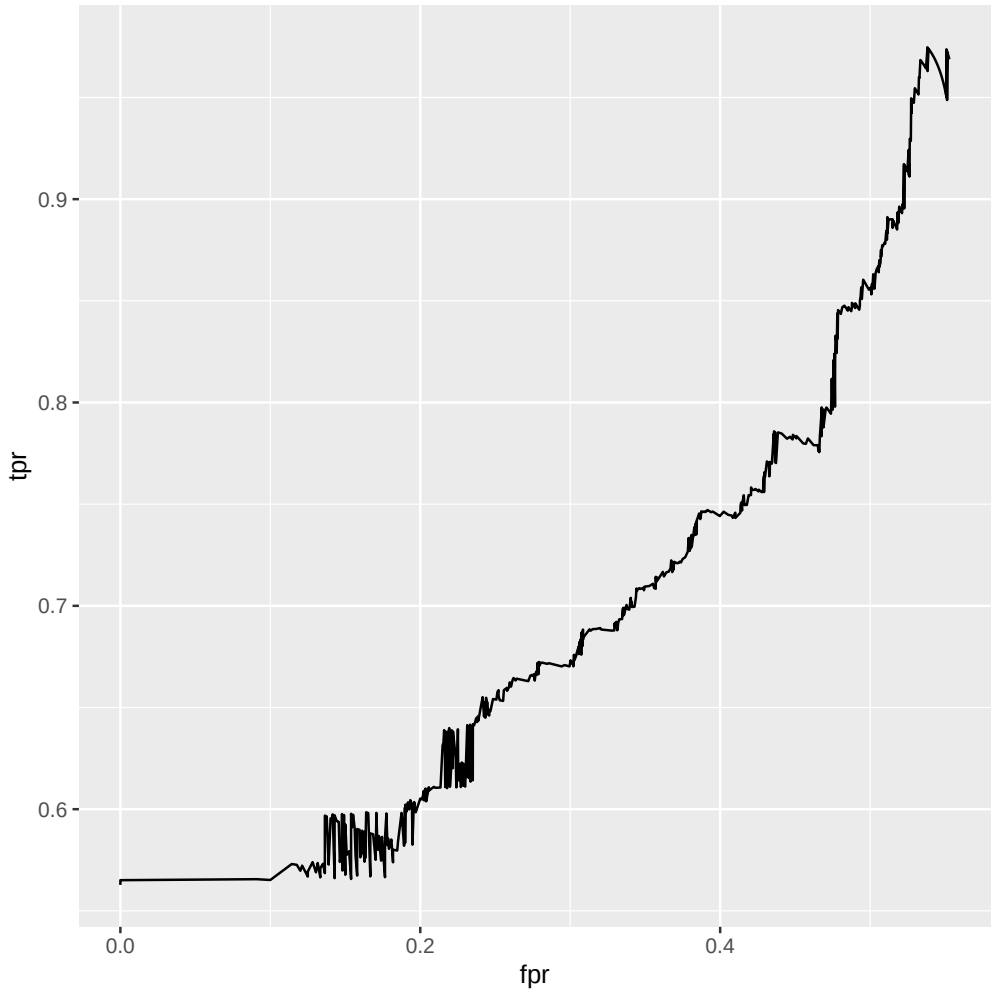
هستند از آن صرف نظر می کنیم. ابتدا داده را به دو بخش آموزش و آزمون تقسیم می کنیم و در هر سه مدل از همین داده ها استفاده می کنیم.

```
> epldf = epl %>% select(FTR,HST,AST,HC,AC,HY,HR,AR)
> n = nrow(epldf)
> train_samples = sample(n(n*8.0,
> trainEpl = epldf[train_samples,]
> testEpl = epldf[-train_samples,]
```

۱.۴.۱ رگرسیون لجستیک

```
> logModel = glm(FTR~HST+AST+HC+AC+HY+HR+AR,
+               data = trainEpl,family = "binomial")
> log.pred = ifelse(predict(logModel,testEpl %>% select(-FTR),
+                           type="response") > 29.0,1,0)
> logPred = predict(logModel,testEpl %>% select(-FTR),
+                           type="response")
> log.res = log.pred == testEpl$FTR
> table(log.pred,testEpl$FTR)

log.pred   0   1
```

```
       0 167  47
       1 469 769

> mean(log.res)

[1] 6446281.0

> tpr = c()
> fpr = c()
> #Long run time
> for (i in seq(from = 50,to = 900,by = 1)){
+   th=i/1000
+ logModel = glm(FTR~HST+AST+HC+AC+HY+HR+AR,
+                data = trainEpl,family = "binomial")
+ log.pred = ifelse(predict(logModel,testEpl %>% select(-FTR),
+                           type="response") > th,1,0)
+
+ fpr[itable(log.pred=[49-,testEpl$FTR)[1,2]/
+   (table(log.pred,testEpl$FTR)[1,2]+
+      table(log.pred,testEpl$FTR)[1,1])
+ tpr[itable(log.pred=[49-,testEpl$FTR)[2,2]/
+   (table(log.pred,testEpl$FTR)[2,2]+
+      table(log.pred,testEpl$FTR)[2,1])
+ }
> rocLog = cbind(tpr,fpr)
> rocLog = as.data.frame(rocLog)
> ggplot(data = rocLog,aes(x=fpr,y=tpr))+geom_line()
> logPred = predict(logModel,testEpl %>% select(-FTR),
+                           type="response")
> ROCRpred <- prediction(logPred, testEpl$FTR)
> ROCRperf <- performance(ROCRpred, 'tpr','fpr')
> plot(ROCRperf, colorize = TRUE, text.adj = c(2.0-((7.1,
```



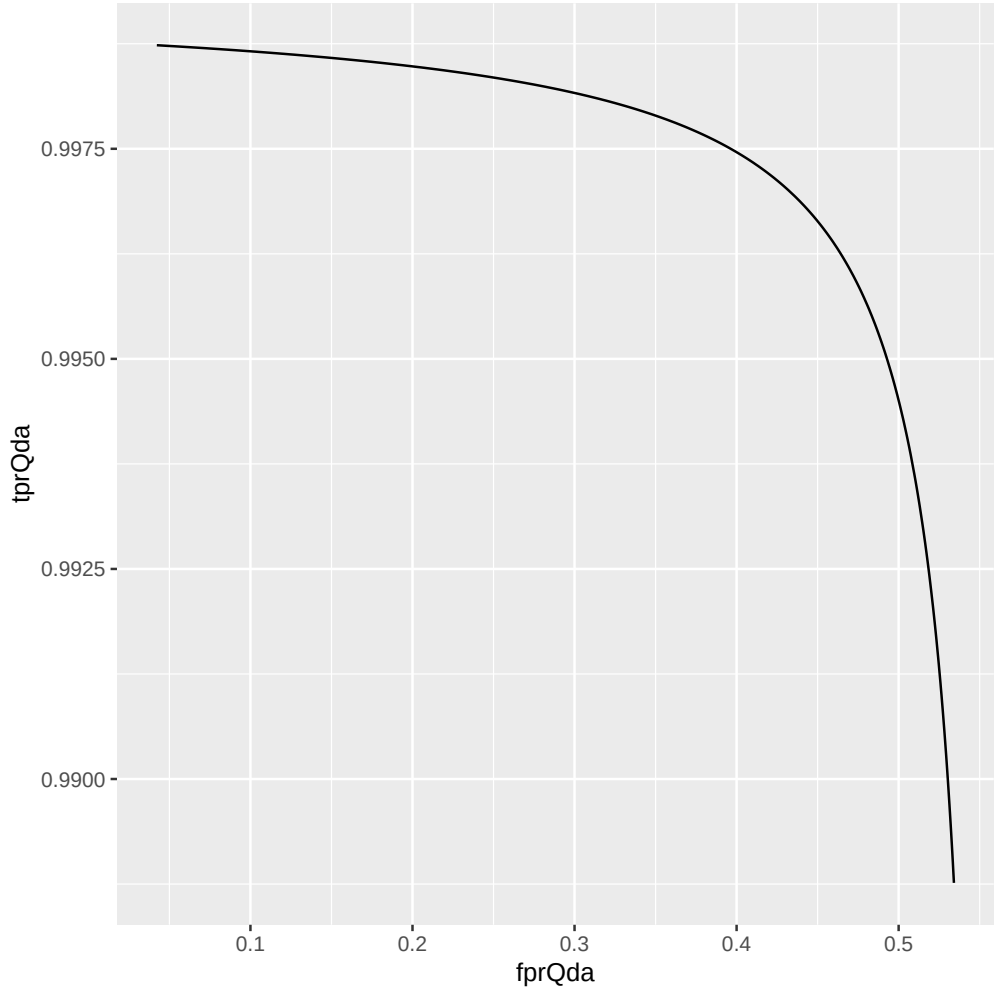
۲.۴.۱ روش QDA

```
> qdaModel = qda(FTR~HST+AST+HC+AC+HY+HR+AR,data = trainEpl)
> qda.pred = predict(qdaModel,testEpl %>% select(-FTR))
> predictions = qda.pred$class
> accuracy = sum(testEpl$FTR == predictions)/length(predictions)
> accuracy
```

```
[1] 6466942.0
```

```
> tprQda = c()
> fprQda = c()
> #Long run time
> for (i in seq(from = 50,to = 900,by = 1)){
+ th=i/1000
+ qdaModel = qda(FTR~HST+AST+HC+AC+HY+HR+AR,data = trainEpl)
+ qda.pred = predict(qdaModel,testEpl %>% select(-FTR))
+ qda.pred$class = ifelse(qda.pred$posterior[,2]>th,1,0)
```

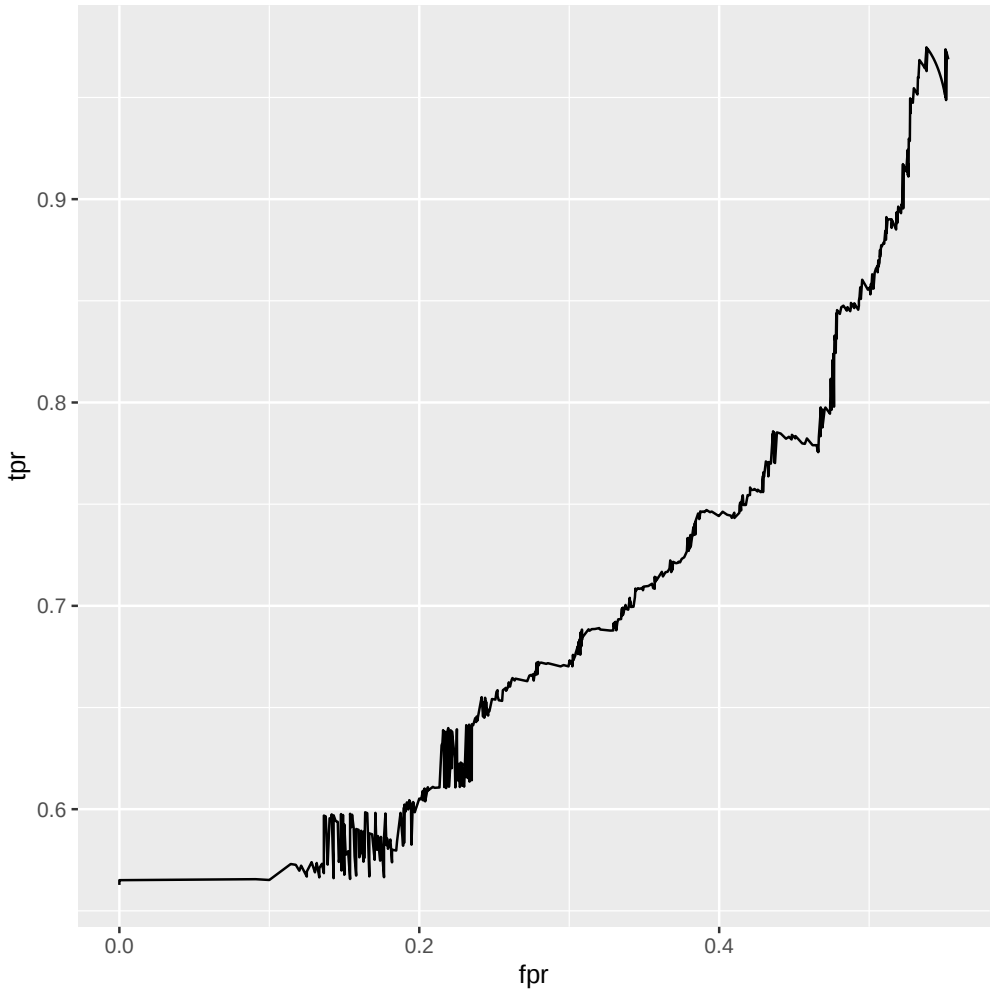
```
+ fprQda[itable(qda.pred$class=[49-,testEpl$FTR)[1,2]/
+   (table(qda.pred$class,testEpl$FTR)[1,2]+
+      table(log.pred,testEpl$FTR)[1,1])
+ tprQda[itable(qda.pred$class=[49-,testEpl$FTR)[2,2]/
+   (table(qda.pred$class,testEpl$FTR)[2,2]+
+      table(log.pred,testEpl$FTR)[2,1])
+ }
> rocQda = cbind(tprQda,fprQda)
> rocQda = as.data.frame(rocQda)
> ggplot(data = rocQda,aes(x=fprQda,y=tprQda)) + geom_line()
```



۳.۴.۱ روش KNN

```
> knn.pred = knn(trainEpl,testEpl,trainEpl$FTR,k=10,prob = TRUE)
> tabl = table(knn.pred,testEpl$FTR)
> acc = (tabl[1,1]+tabl[2,2])/1452
> acc
```

```
[1] 8422865.0

> probs = attr(knn.pred,"prob")
> tprKnn = c()
> fprKnn = c()
> #Long run time
> for (i in seq(from = 600,to = 950,by = 1)){
+   th=i/1000
+ knn.pred = ifelse(probs >= th,1,0)
+ fprKnn[itable(knn.pred=[599-,testEpl$FTR)[1,2]/
+   (table(knn.pred,testEpl$FTR)[1,2]+
+      table(log.pred,testEpl$FTR)[1,1])
+ tprKnn[itable(knn.pred=[599-,testEpl$FTR)[2,2]/
+   (table(knn.pred,testEpl$FTR)[2,2]+
+      table(log.pred,testEpl$FTR)[2,1])
+ }
> rocKnn = cbind(tprKnn,fprKnn)
> rocKnn = as.data.frame(rocLog)
> ggplot(data = rocLog,aes(y=tpr,x=fpr))+geom_line()
```



با توجه به نتایج بالا به نظر می رسد روش KNN-10 از همه بهتر باشد. دقت کنید برای $k = ۱, \ldots, ۱۰$ دقت محاسبه شد که در همین حدود بود و طبق گفتهٔ کتاب که به صورت تجربی $k = ۵$ و $k = ۱۰$ را پیشنهاد می دهد این مدل انتخاب شد. توجه کنید نمودار ROC مدل لجستیک با همین نمودار برای KNN تفاوت چندانی نمی کند اما دقت روش KNN بیشتر است.

با توجه به نتایج بالا طبق آنچه در درس آموختیم و با توجه به کتاب این حالتی که پیش آمده نشان از آن دارد که رابطهٔ واقعی جامعه غیرخطی است. (مطابق سناریو ۶ صفحهٔ ۱۵۲ کتاب)