predictFinish

David Wihl

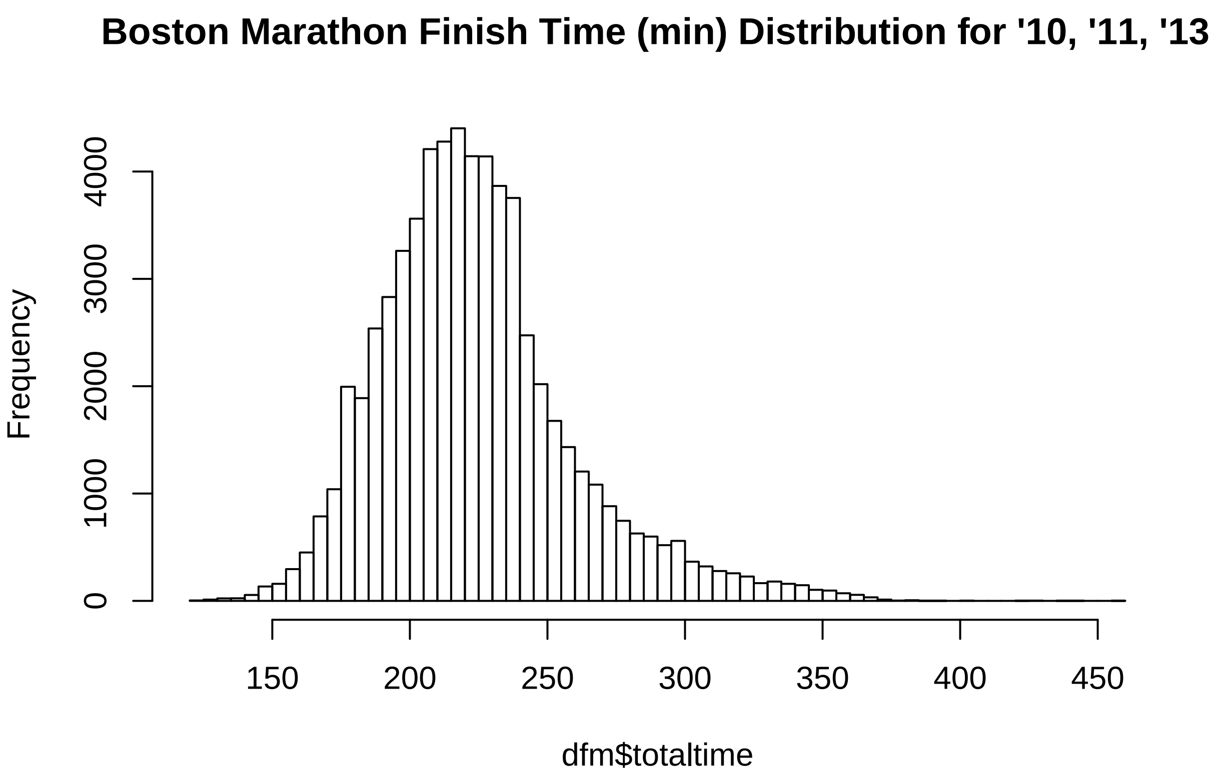
November 24, 2015

First let's read in the data and calculate finish times:

```
dfm <- read.csv("Previous Boston Marathon study/BAA data.txt",header=T,sep=" ")
times = as.matrix(dfm[,7:15], ncol=9)
dfm$totaltime = rowSums(times)
```

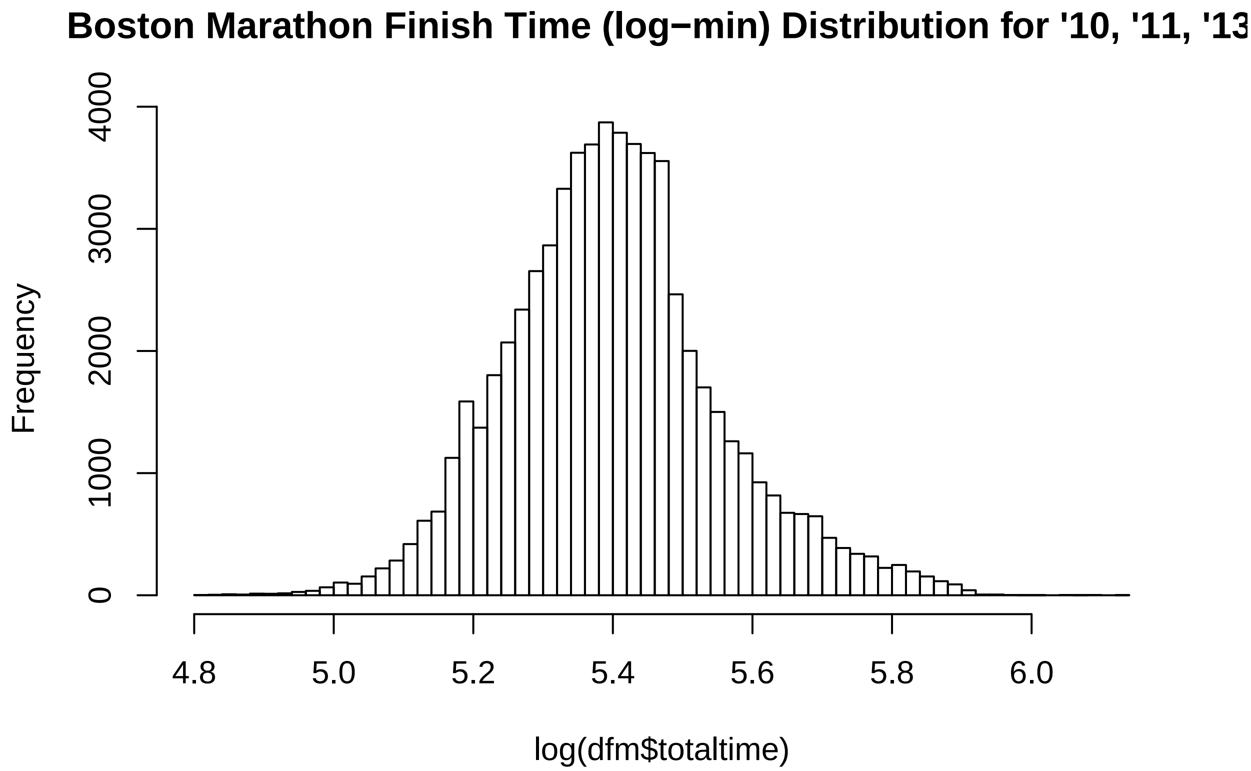
Some preliminary EDA of the data:

```
hist(dfm$totaltime,breaks=50, main="Boston Marathon Finish Time (min) Distribution for '10, '11, '13")
```



The data appears somewhat left skewed. Let's try a log transform:

```
hist(log(dfm$totaltime),breaks=50, main="Boston Marathon Finish Time (log-min) Distribution for '10, '1
```



That seems better.

Let's try a multiple regression on the non-log data:

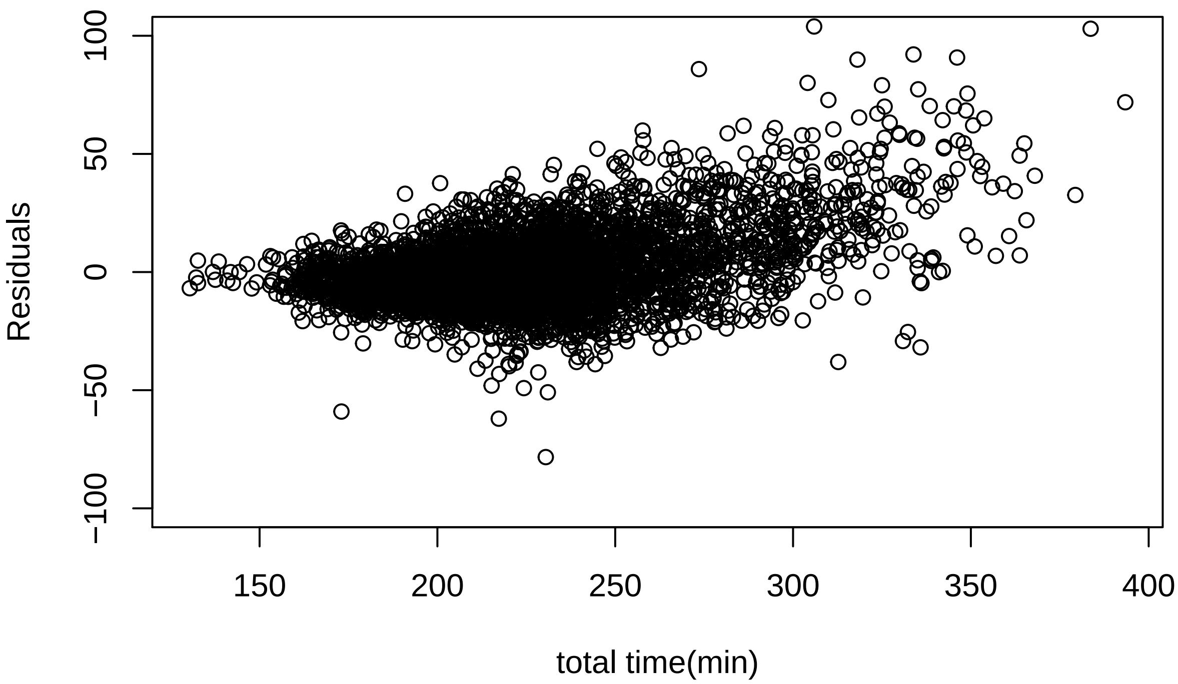
```
dfms = dfm[!is.na(dfm$totaltime), ]
dfms = dfms[order(dfms$totaltime),]
model = lm(totaltime~Age+Gender1F2M+K0.5, data=dfms)
summary(model)
```

```
##
## Call:
## lm(formula = totaltime ~ Age + Gender1F2M + K0.5, data = dfms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376.18   -9.26   -2.64    6.48  187.43
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.640157   0.595532  -41.38   <2e-16 ***
## Age           0.073888   0.006227   11.87   <2e-16 ***
## Gender1F2M    2.847267   0.139233   20.45   <2e-16 ***
## K0.5          9.703124   0.020687  469.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.43 on 64163 degrees of freedom
## Multiple R-squared:  0.8097, Adjusted R-squared:  0.8097
## F-statistic: 9.103e+04 on 3 and 64163 DF,  p-value: < 2.2e-16
```

```
model.resid = resid(model)
```

Let's plot the residuals

```
#plot(dfms$totaltime,model.resid,ylim=c(-100,100))
xydata = data.frame(x=dfms$totaltime, y=model.resid)
xydata = xydata[sample(1:nrow(xydata), 5000, replace=FALSE),]
plot(xydata$x,xydata$y,ylim=c(-100,100), xlab="total time(min)", ylab="Residuals")
```



This appears to be spreading out and trending up.

Let's try a polynomial transformation on the 5k time to see if that improves things:

```
bestresidsum = 9e9
bestpoly = 0
for (i in seq(from=1.0, to= 2.5, by=0.1)){
  dfms$k5xform = dfms$K0.5 ^ i
  model = lm(totaltime~Age+Gender1F2M+k5xform, data=dfms)
  model.residsum = sum(resid(model)^2)
  if (model.residsum < bestresidsum) {
    bestresid = resid(model)
    bestresidsum = model.residsum
    bestpoly = i
  }
}

cat("total error went from ",sum(model.resid^2), " (untransformed) to ", bestresidsum," (transformed)\n
```

```
## total error went from  15271875  (untransformed) to  15261576  (transformed)
```

```r
#plot(dfms$totaltime,bestresid,ylim=c(-100,100))
xydata = data.frame(x=dfms$totaltime, y=bestresid)
xydata = xydata[sample(1:nrow(xydata), 5000, replace=FALSE),]
plot(xydata$x,xydata$y,ylim=c(-100,100), xlab="total time(min)", ylab="Residuals")
```