Using Block Maxima and P.O.T. to model Slowest NFL Combine Athletes

Sri-Amirthan Theivendran

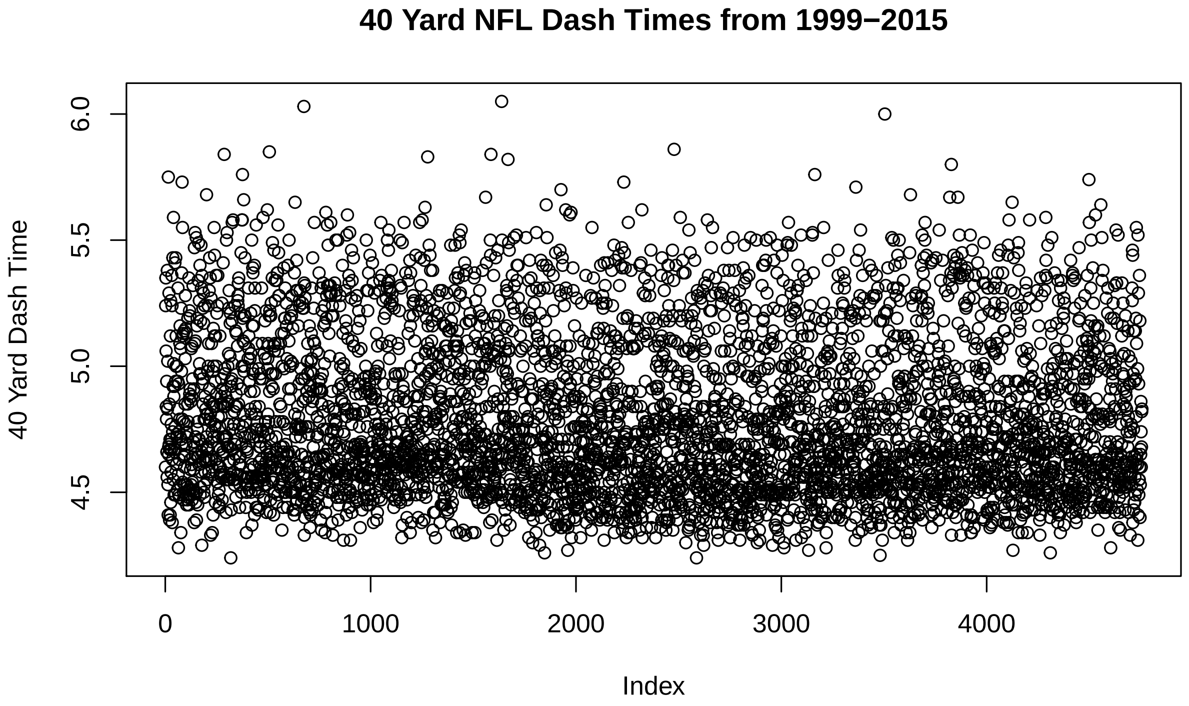
Every year roughly 300 NFL draft prospects attend the NFL combine and perform a myriad of physical challenges that test their strength and endurance. These tests are a means for teams to gage at a glance how ready a player is for the NFL. One of these tests is the forty yard dash. My aim is to see just how slow a player can be and still be in contention for the NFL draft.

To this end, I obtained yearly data for nfl combine stats from 1999-2015. Every year there are roughly 300 observations of 40 yard dash times. A snap shot of the data as well as a plot of the 40 yard dash times are given below.

```
head(nfldata)
```

```
##              name year fortyyd
## 1    Aaron Brooks 1999    4.60
## 2     Aaron Dalan 1999    5.24
## 3    Aaron Gibson 1999    5.35
## 4     Aaron Smith 1999    5.06
## 5   Aaron Stecker 1999    4.84
## 6 Aaron Williams 1999    4.79
```

```
plot(nflrundata, ylab="40 Yard Dash Time", main="40 Yard NFL Dash Times from 1999-2015")
```



```
summary(nflrundata)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.240   4.550   4.710   4.796   5.000   6.050
```

```
length(nflrundata)
```

```
## [1] 4756
```

We first analyze te data using the block maxima approach.

Block Maxima

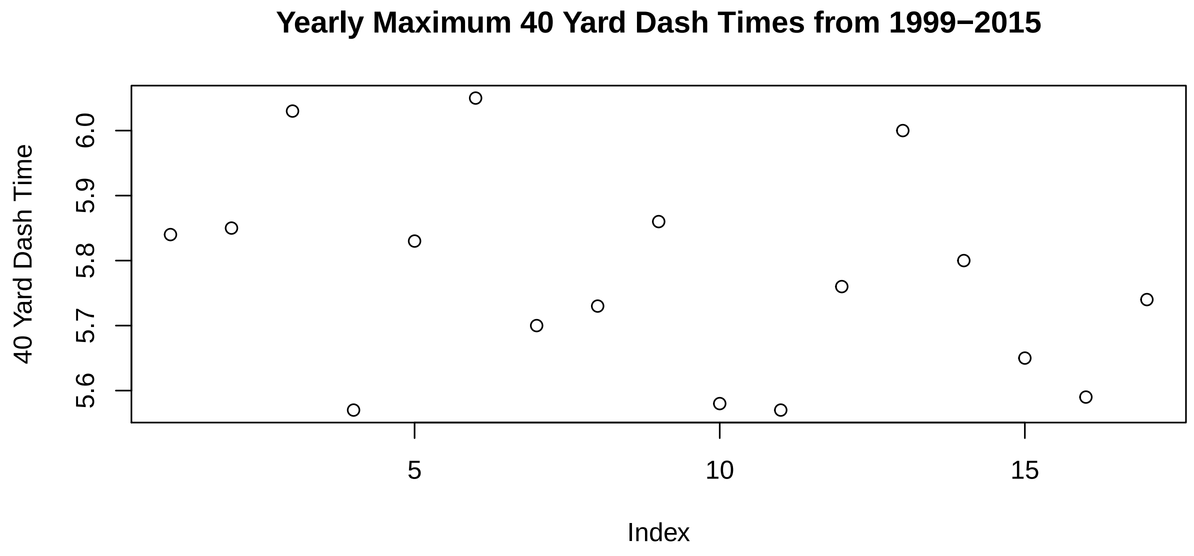
We first discuss assumptions of the model. The individual running times are quite independent of each other (each individual races alone) and identically distributed (as these running times came from one cohort of professional athletes, who share similar physical attributes).

We take our block size as one year, resulting in 17 observations. Our block size is largely based on how the data is collected, which results in this natural block size. For fitting GEV data to these block maxima, the yearly maximum are independent of each other for the reason stated before and it is reasonable to think that they are approximately GEV distributed as we are taking the maximum of 300 or so observations. A snaphshot as well as a plot of the data corresponding to the yearly maxima are given below.

```
head(nflmax)
```

```
##   Year  Max
## 1 1999 5.84
## 2 2000 5.85
## 3 2001 6.03
## 4 2002 5.57
## 5 2003 5.83
## 6 2004 6.05
```

```
plot(nflrunmaxdata, ylab="40 Yard Dash Time", main="Yearly Maximum 40 Yard Dash Times from 1999-2015")
```



```
length(nflrunmaxdata)
```

```
## [1] 17
```

We fit a GEV model to the yearly maximum data.

```
library(ismev)
```

```
## Loading required package: mgcv

## Loading required package: nlme

## This is mgcv 1.8-22. For overview type 'help("mgcv-package")'.
```

```
gfit<-gev.fit(nflrunmaxdata)
```

```
## $conv
## [1] 0
##
## $nllh
## [1] -8.291335
##
## $mle
## [1]  5.7117379  0.1375311 -0.1632304
##
## $se
## [1] 0.04021388 0.03058668 0.26882484
```

Approximate 95 percent confidence intervals for the mle estimates are given below.

```
LowerBounds <- gfit$mle - 1.96*gfit$se
UpperBounds <- gfit$mle + 1.96*gfit$se
LowerBounds
```

```
## [1]  5.63291873  0.07758125 -0.69012710
```

```
UpperBounds
```

```
## [1] 5.7905572 0.1974810 0.3636663
```

The estimated location parameter is $\hat{\mu} = 5.7117$, the estimated scale parameter is $\hat{\sigma} = 0.1375$ and the estimated shape parameter $\hat{\gamma} = -0.1632$. The estimated shape parameter is negative which implies that the distribution has an upper endpoint (although a 95 percent confidence interval for the parameter would include positive and negative values). The estimated upper endpoint would be

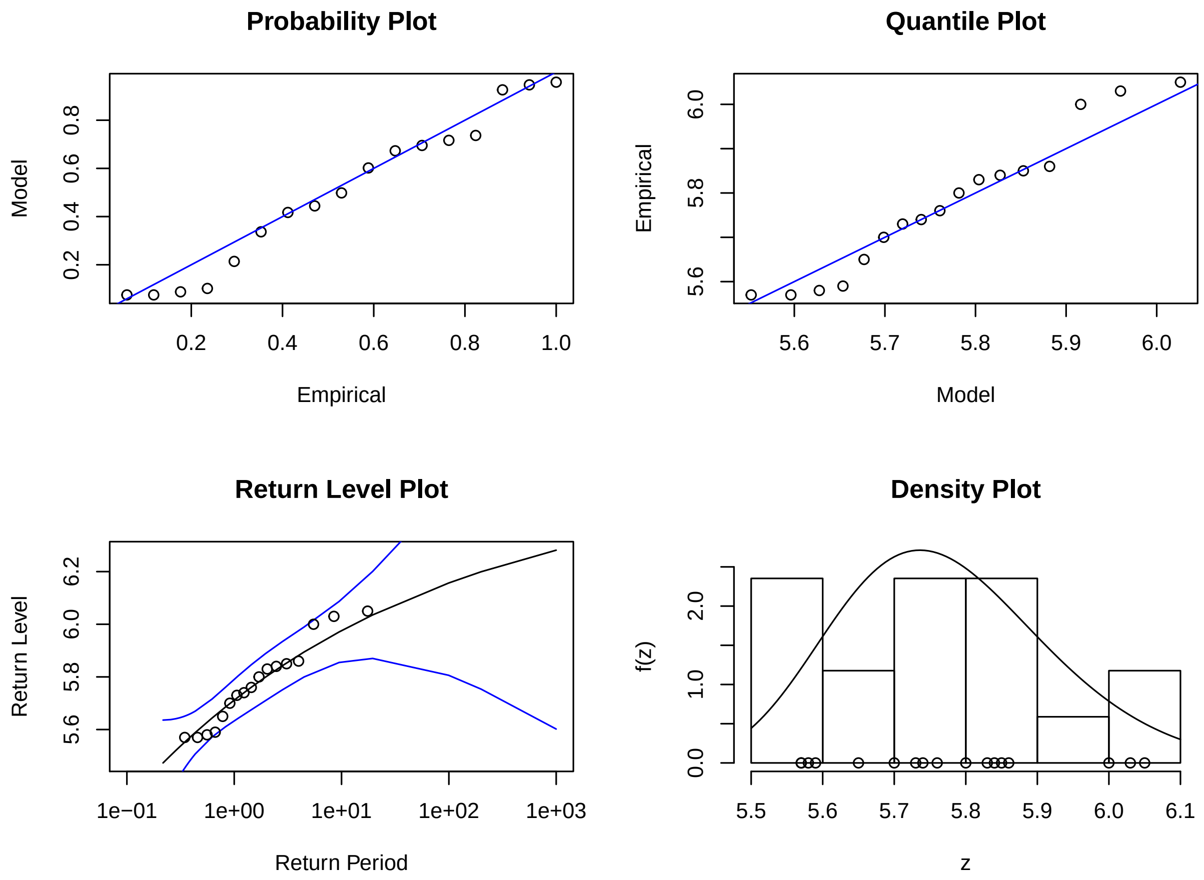
```
gfit$mle[1]-(gfit$mle[2]/gfit$mle[3])
```

```
## [1] 6.554296
```

which based on the data is not too unrealistic.

Next we verify that the estimated GEV distribution fits the data well. Below among other plots, a QQ-plot and probability plot are given. The distribution seems to fit the data well even into the tails, even though, given our block size there weren't many observations.

```
gev.diag(gfit)
```



Finally, we compute the return level with return period 1000 years.

```
library(evir)
qgev(0.999, xi=gfit$mle[3], mu=gfit$mle[1], sigma=gfit$mle[2])
```

```
## [1] 6.281434
```

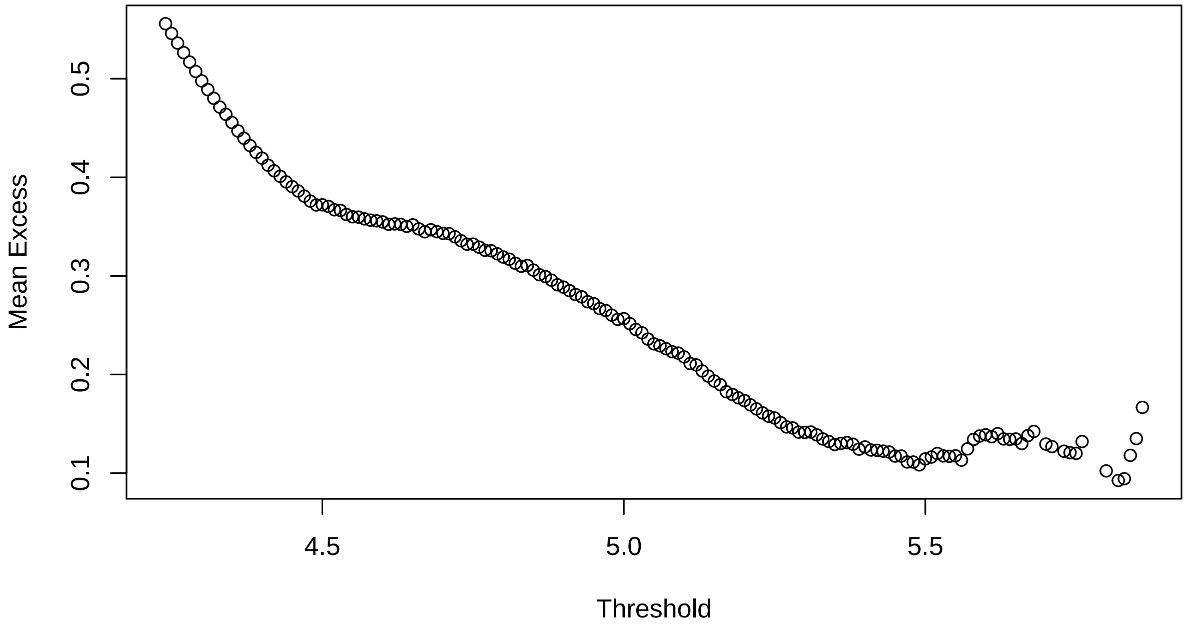
So on average we will see 1 player every 1000 years have a shot of making the NFL with a 40 yd dash time that exceeds 6.28 seconds at the combine. The estimate though will have an extremely large confidence interval because of the lack of observations due to a big block size. This phenomenon is illustrated in the return level plot given above.

Now we analyze the same data using the peaks-over-threshold method.

Peaks-Over-Thresholds

We first determine a suitable threshold by means of a mean-excess plot. A sample mean excess plot is given below

```
par(mfrow=c(1,1))
meplot(nflrundata)
```



The mean-excess plot is rougly linear from a threshold of 4.7 onwards. Hence after thresholding, we will still be using roughly 50 percent of the data. We fit a GPD to the excesses for this thresholded data.

```
fit<-gpd.fit(nflrundata, 4.7, npy=300)
```

```
## $threshold
## [1] 4.7
##
## $nexc
## [1] 2410
##
## $conv
## [1] 0
##
## $nllh
## [1] -320.5012
##
## $mle
## [1]  0.4451989 -0.3237586
##
## $rate
## [1] 0.5067283
##
## $se
## [1] 0.009328648 0.008362731
```

The estimated scale parameter is $\hat{\sigma} = 0.4452$ while the estimated shape parameter $\hat{\gamma} = -0.3237$. Approximate 95 percent confidence intervals for the mle estimates are given below. The peaks-over-threshhold method gives much tighter confidence intervals for the estimates because we are using a lot more data. The estimate for γ is negative like in the block maxima case but the confidence interval is now entirely negative.

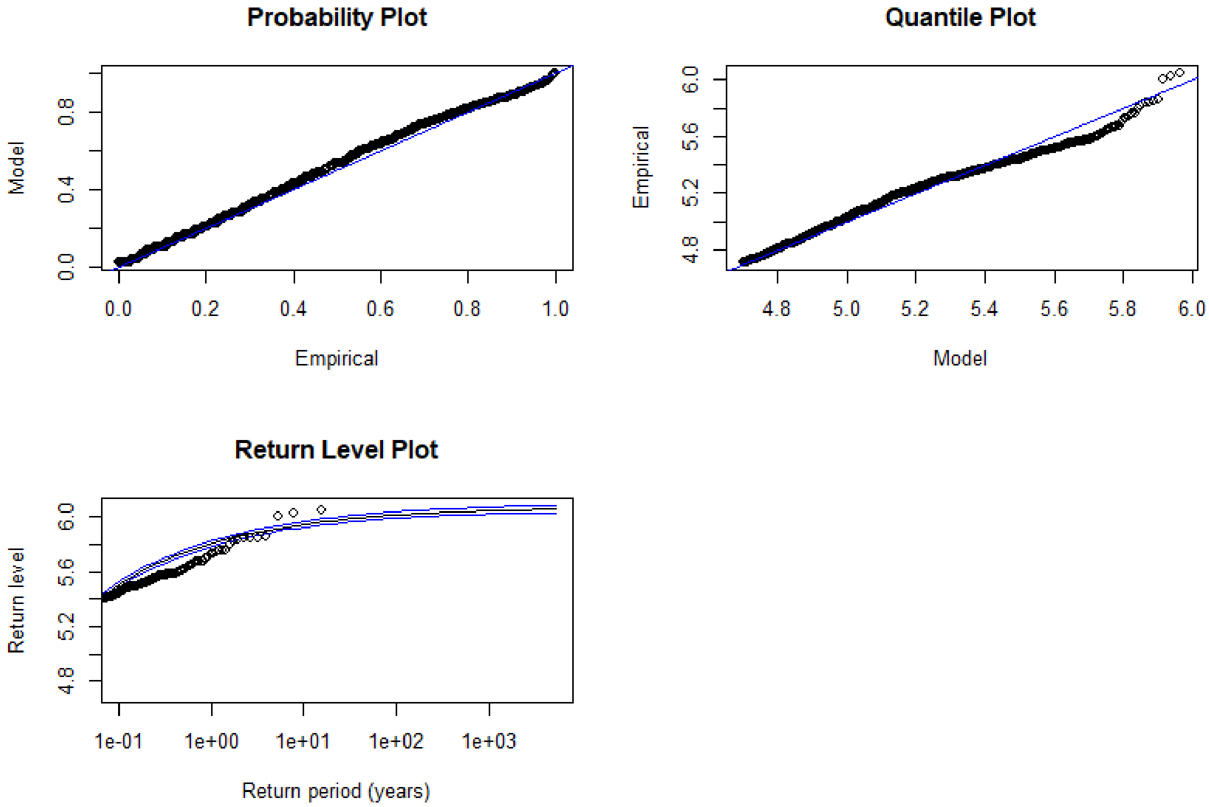


Figure 1: Diagonostic Plots for Estimated GPD distribution

```
lbounds<-fit$mle-1.96*fit$se
ubounds<-fit$mle+1.96*fit$se
lbounds
```

```
## [1]  0.4269148 -0.3401496
```

```
ubounds
```

```
## [1]  0.4634831 -0.3073677
```

Because the estimate for $\hat{\gamma}$ is negative the distribution has an upper endpoint. The estimated upper endpoint can be computed to be

```
4.7-fit$mle[1]/fit$mle[2]
```

```
## [1] 6.075095
```

As a comparison the block maxima gave an estimated upper endpoint of 6.55 seconds and the maximum time recorded is 6.05 seconds. Because of a larger amount of data, the peaks over threshold method gives an estimate that is nearer to the observed data.

Now we verify that the estimated GPD distribution fits the data well. Above a qq plot for the estimated pareto distribution among other diagnostic plots are given. The distribution seems to fit the data well even into the tails. The confidence intervals for the return levels are also a lot more narrow.

Now we compute the return level with return period 1000 years. The return level is estimated to be 6.06 seconds (using the formula from the slides, with $T = 1000 \times 280$ and the estimators of the paramters given above). In comparison the block maxima gave a return level of 6.28 seconds. The confidence interval associated with the estimate from the peaks over threshold method is much tighter as well.