```
download.file("https://github.com/nazareno/imdb-series/raw/master/data/series_from_imdb.
csv",destfile="series_from_imdb.csv")
base = read.csv("series_from_imdb.csv")
```

It is a large dataset, with more than 64,000 episodes of almost 890 TV series,

```
str(base)
'data.frame':   64018 obs. of  18 variables:
 $ series_name: Factor w/ 889 levels "'Allo 'Allo!",..: 137 137 137 137 137 137
137 137 137 137 ...
 $ episode    : Factor w/ 54090 levels "-30-","¡Viva los muertos!",..: 32314
7446 16 7176 17748 9562 1379 36218 17845 5553 ...
 $ series_ep  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ season     : int  1 1 1 1 1 1 1 2 2 2 ...
 $ season_ep  : int  1 2 3 4 5 6 7 1 2 3 ...
 $ user_rating: num  8.9 8.7 8.7 8.2 8.3 9.2 8.8 8.7 9.2 8.3 ...
```

Just pick a TV series, for instance Dan Harmon's Community,

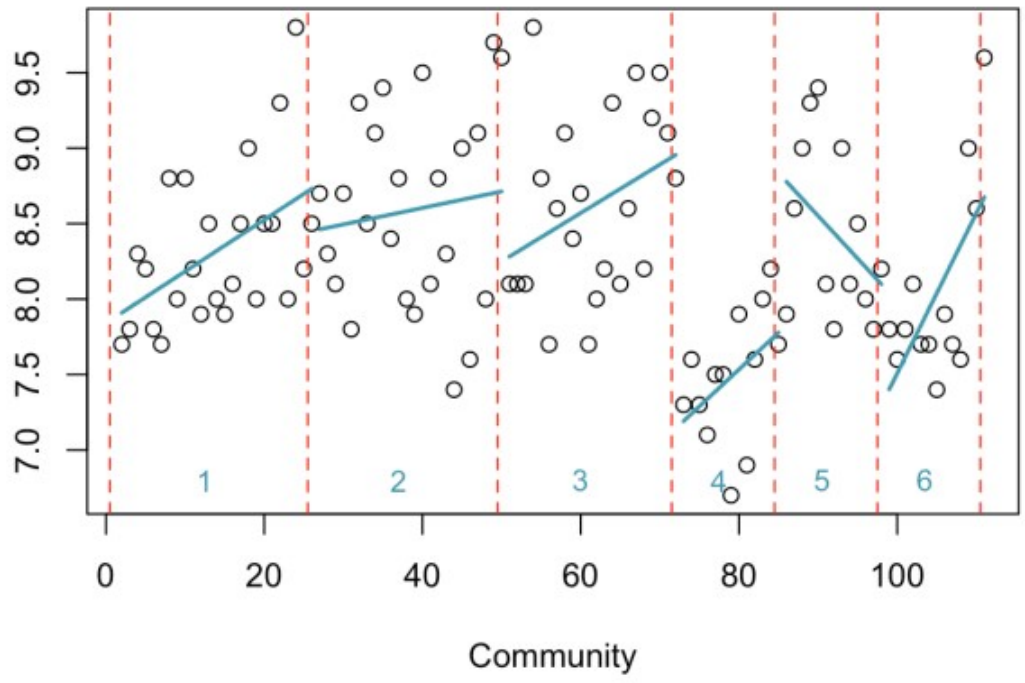
```
sbase = base[base$series_name=="Community",]
```

We can plot the evolution of the rating over the 110 episodes.

```
sbase=sbase[!duplicated(sbase[,c(1,2,4,5)]),]
sbase$series_ep=1:nrow(sbase)
```

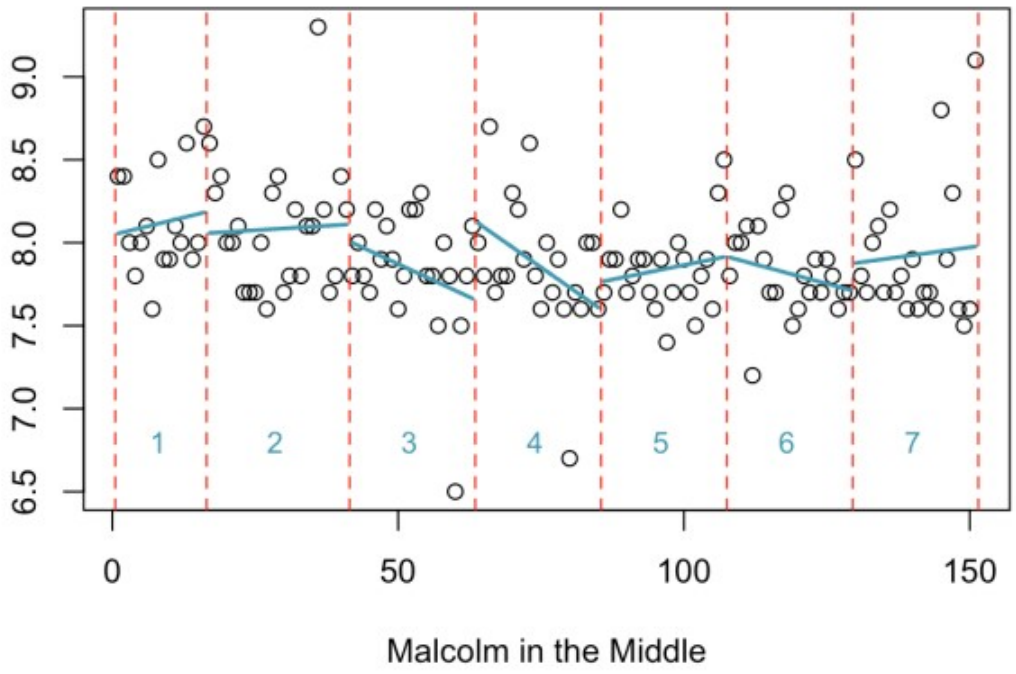
()since there could be some problem with the data (such as duplicates, let us clean it quickly)

```
plot(sbase$series_ep,sbase$UserRating,xlab=sbase$series_name[1])
idx=c(0,which(diff(sbase$season)!=0),nrow(sbase))
abline(v=idx+.5,lty=2,col=colr[2])
a = unique(sbase$season)
for(u in a){
  ssbase = sbase[sbase$season==u,]
  reg = lm(UserRating~series_ep,data=ssbase)
  lines(ssbase$series_ep,predict(reg),col=colr[3],lwd=2)
}
```



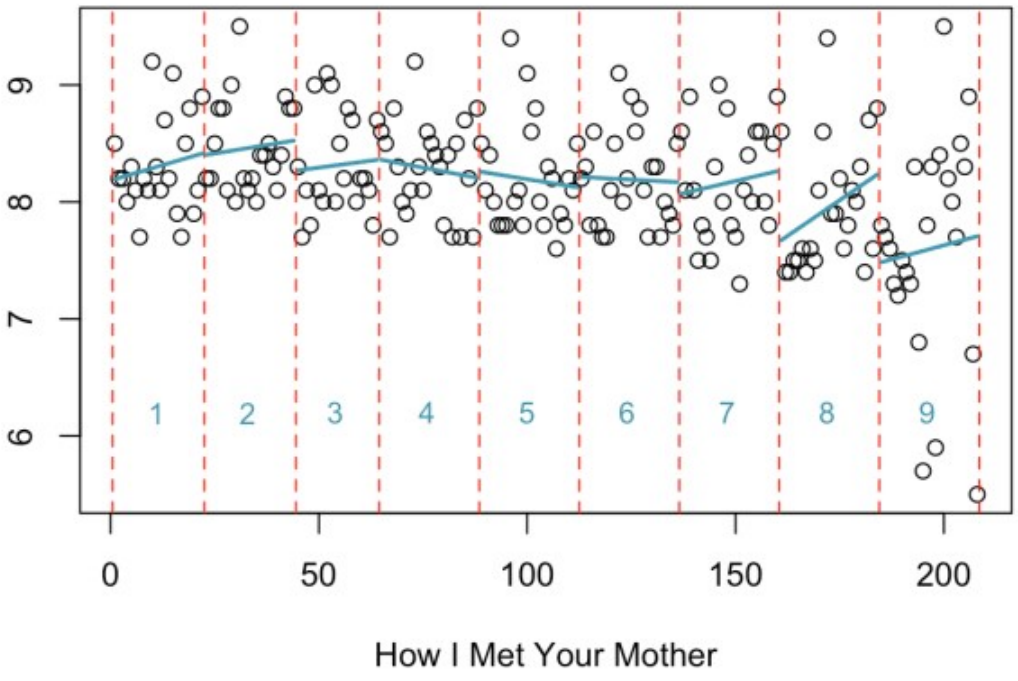
The vertical lines are here to visualize the seasons. On issue is that the lenght can vary with time. Consider Linwood Boomer's Malcom in The Middle,

```
sbase = base[base$series_name=="Malcolm in the Middle",]
```



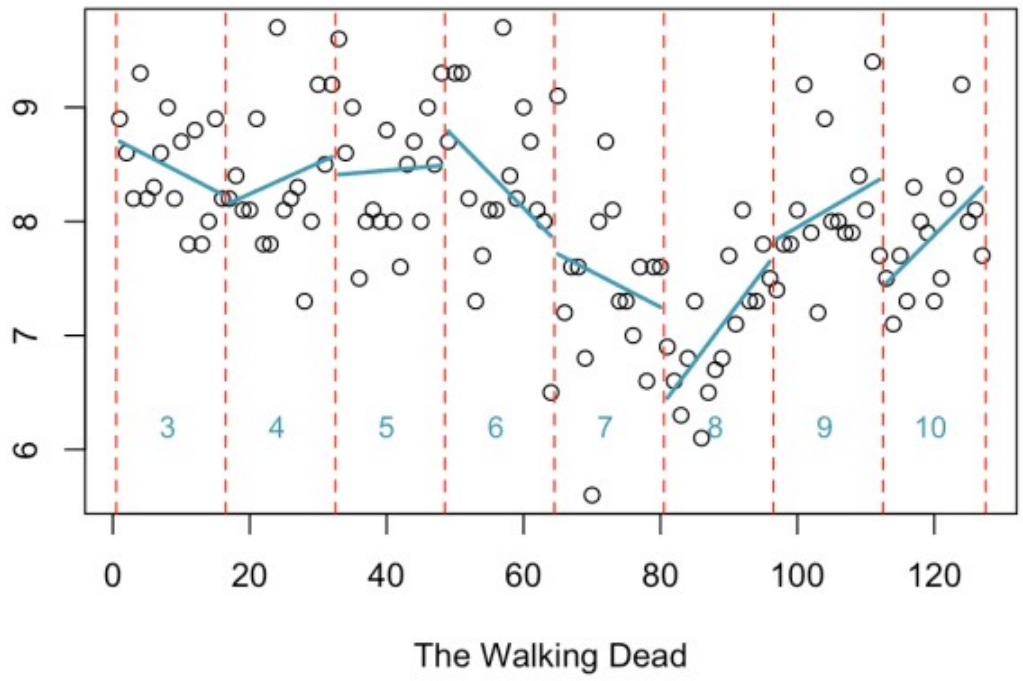
or Craig Thomas and Carter Bays's How I Met Your Mother,

```
sbase = base[base$series_name=="How I Met Your Mother",]
```



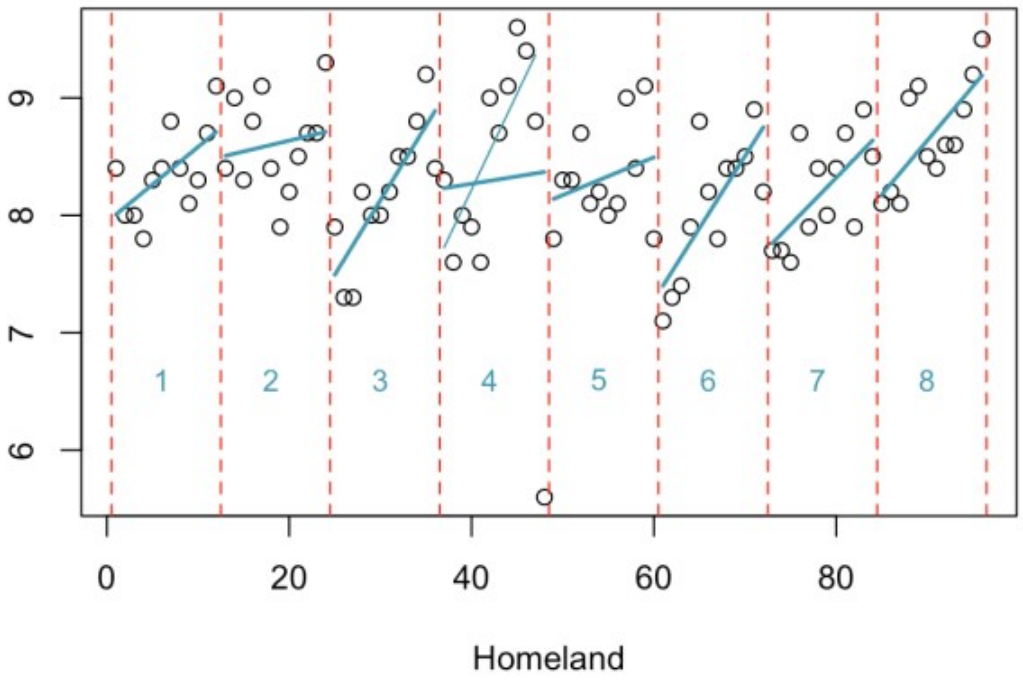
On those two, the evolution is rather stable. Look at AMC's The Walking Dead,

```
sbase = base[base$series_name=="The Walking Dead",]
```



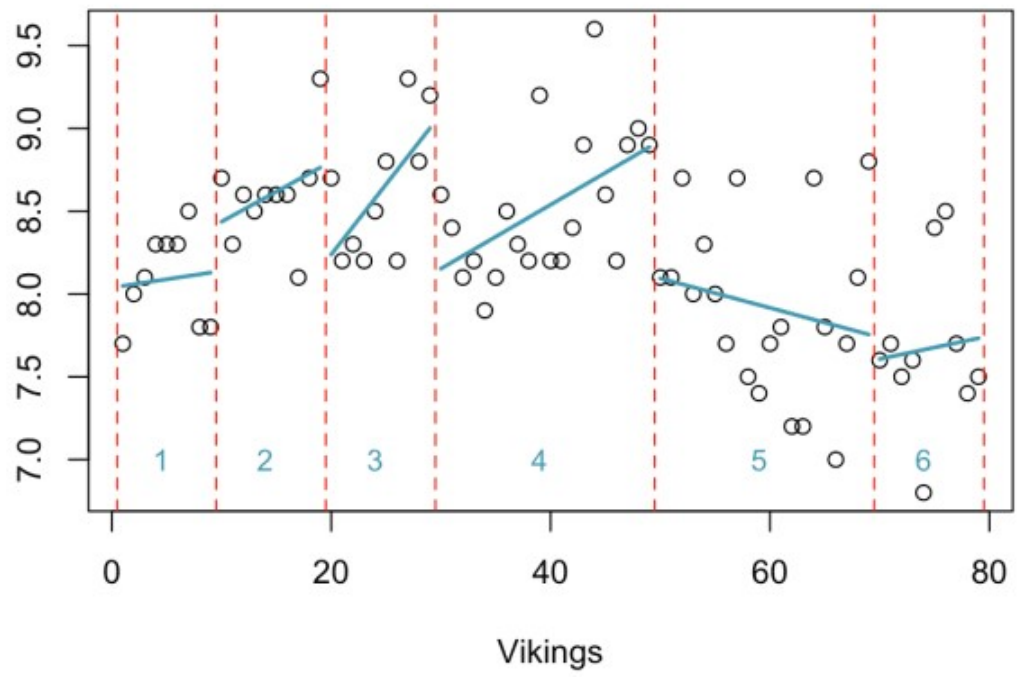
Now, look at Howard Gordon and Alex Gansa's Homeland,

```
sbase = base[base$series_name=="Homeland",]
```



There is an issue here with the last episode of season4, "Long Time Coming", that has a very poor rating. If we remove that point, we get the thin line. Note that the regression line is always increasing. For Michael Hirst's Vickings, we have

```
sbase = base[base$series_name=="Vicking",]
```



If we look more carefully on the previous graph, for five seasons (out of six), we have a positive slope. Well, to be honest, it is not *significantly* positive most of the time, but still. Out of 80 shows, and a total of 583 seasons, the slope is postive 75% of the time (433) and negative 25% of the time (150).

```
BASE = NULL
L80 = unique(base$series_name)
for(j in 1:length(L)){
sbase=base[base$series_name==L[j],]
sbase=sbase[!duplicated(sbase[,c(1,2,4,5)]),]
sbase=sbase[sbase$season>0,]
sbase$series_ep=1:nrow(sbase)
a=unique(sbase$season)
a=a[!is.na(a)]
for(u in a){
  ssbase=sbase[sbase$season==u,]
  reg=lm(UserRating~series_ep,data=ssbase)
  pente = NA
  if((!is.na(coefficients(reg)[2]))&(!is.na((summary(reg)$coefficients[2,4])))){
  if((summary(reg)$coefficients[2,4]<.05)&(coefficients(reg)[2]>0))
pente="positive"
  if((summary(reg)$coefficients[2,4]<.05)&(coefficients(reg)[2]<0))
pente="negative" sdf=data.frame(nom=sbase$series_name[1],season=u,slope=
coefficients(reg)[2],inf=confint(reg)[2,1],sup=confint(reg)[2,2],signe=pente)
BASE=rbind(BASE,sdf)} }} str(BASE) 'data.frame': 583 obs. of 6 variables: $ nom
: Factor w/ 80 levels "Friends","Game of Thrones",..: 1 1 1 1 1 1 1 1 1 1 ...

mean(BASE$slope>0)
[1] 0.7427101
table(BASE$signe)
negative positive
      15      144
```

x

```
sbase12 = sbase[sbase$season%in%c(a[ij],a[ij+1]),]
seuil = sbase12$series_ep[which(diff(sbase12$season)!=0)]+.5
s = function(x) (x-seuil)*(x>seuil)
```

```
reg = lm(UserRating~series_ep+s(series_ep)+I(series_ep>seuil),data=sbase12)
```

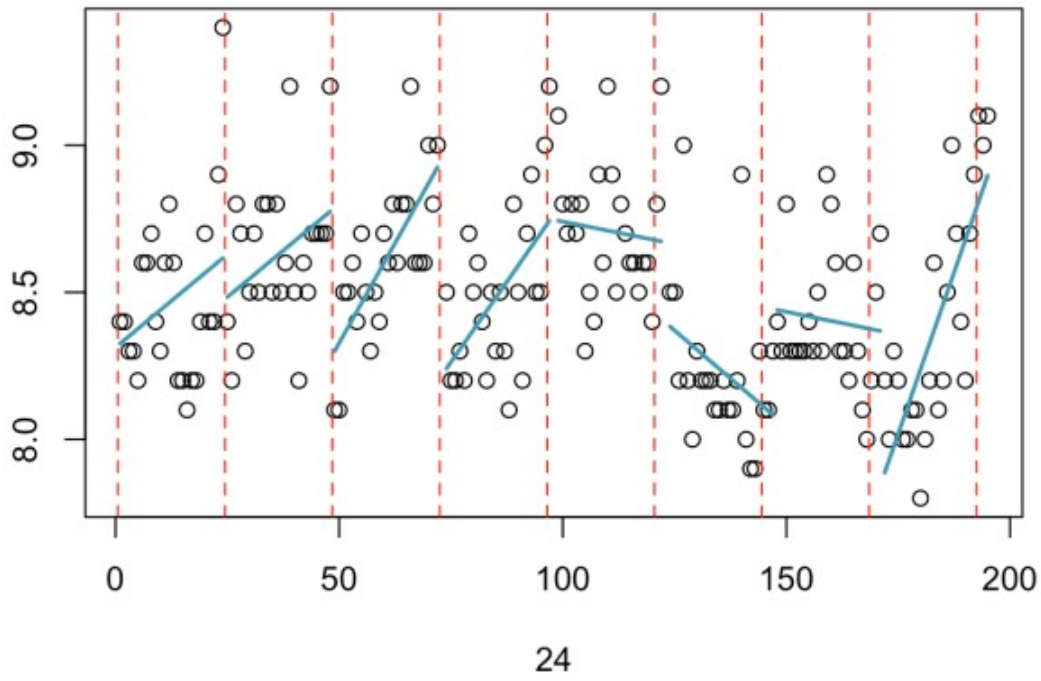
x

```
summary(reg)
Coefficients:
                        Estimate Std. Error t value  Pr(|t|)
(Intercept)              8.45000    0.16338  51.719    2e-16 ***
series_ep                0.10000    0.03235   3.091 0.008598 **
s(series_ep)             0.02000    0.04218   0.474 0.643291
I(series_ep)TRUE.       -1.01778    0.20486  -4.968 0.000257 ***
```

But again, most of the time, it is not significant. To be more specific, 72% of the time, the slope is not significant. But when it is, 90% of the time, it is positive (144 seasons).

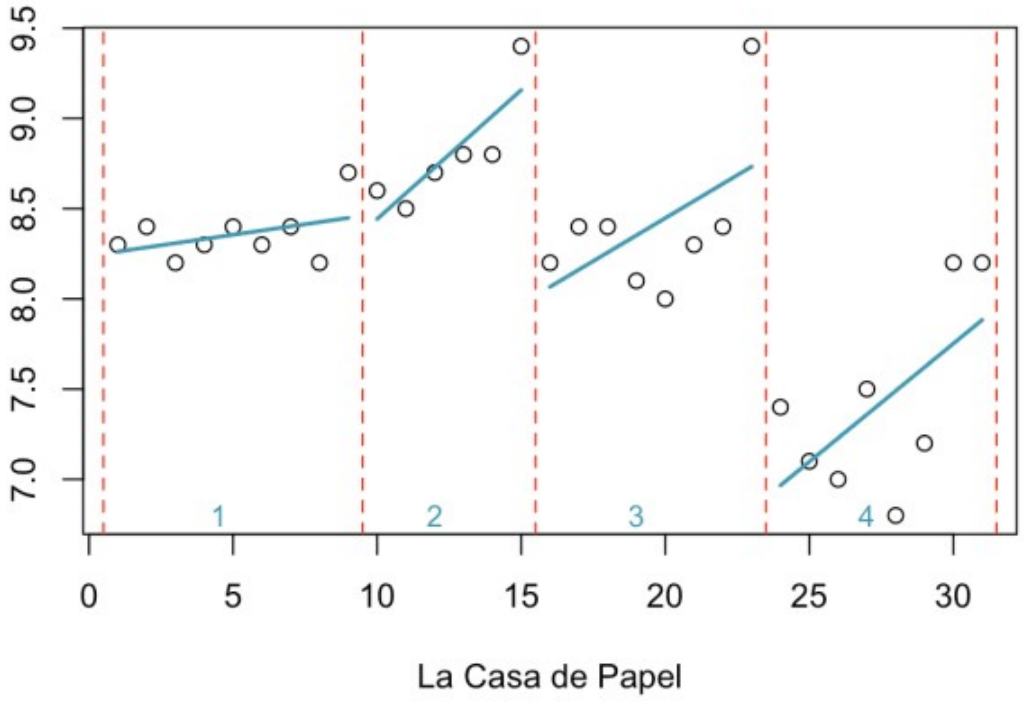
Joel Surnow and Robert Cochran's 24,

```
sbase = base[base$series_name=="Community",]
```

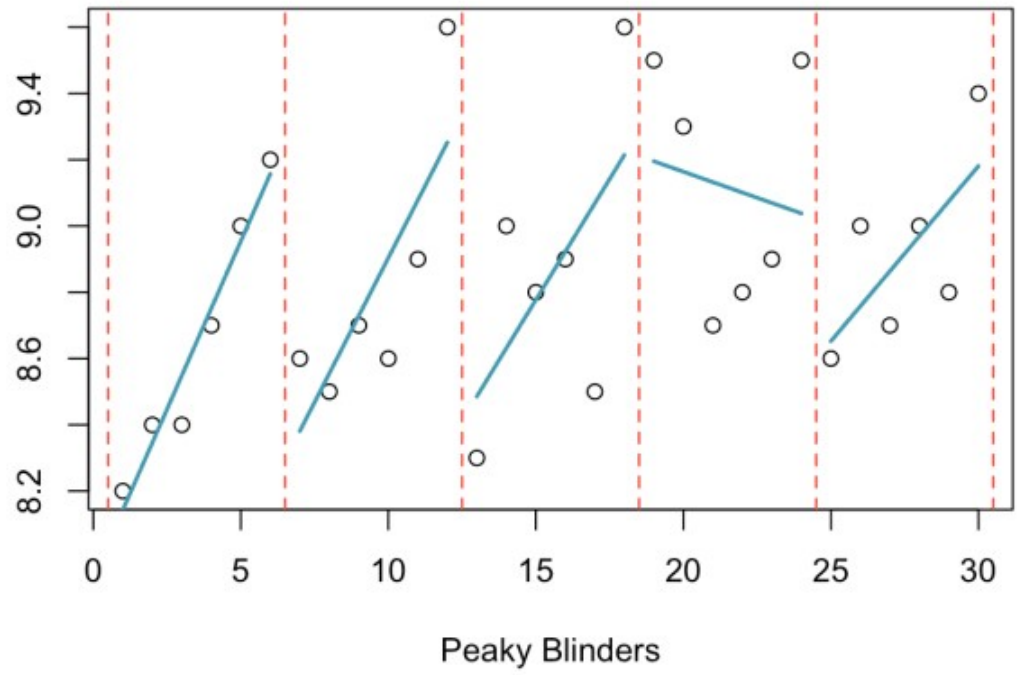


Álex Pina's La Casa de Papel,

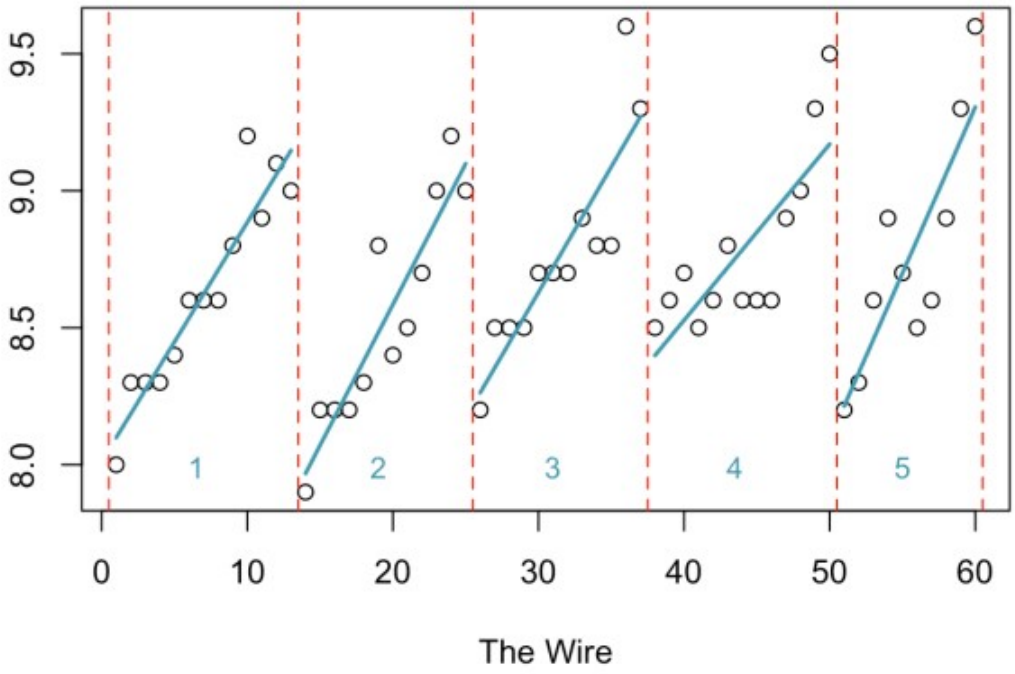
```
sbase = base[base$series_name=="Community",]
```



x Steven Knight's Peaky Blinders,

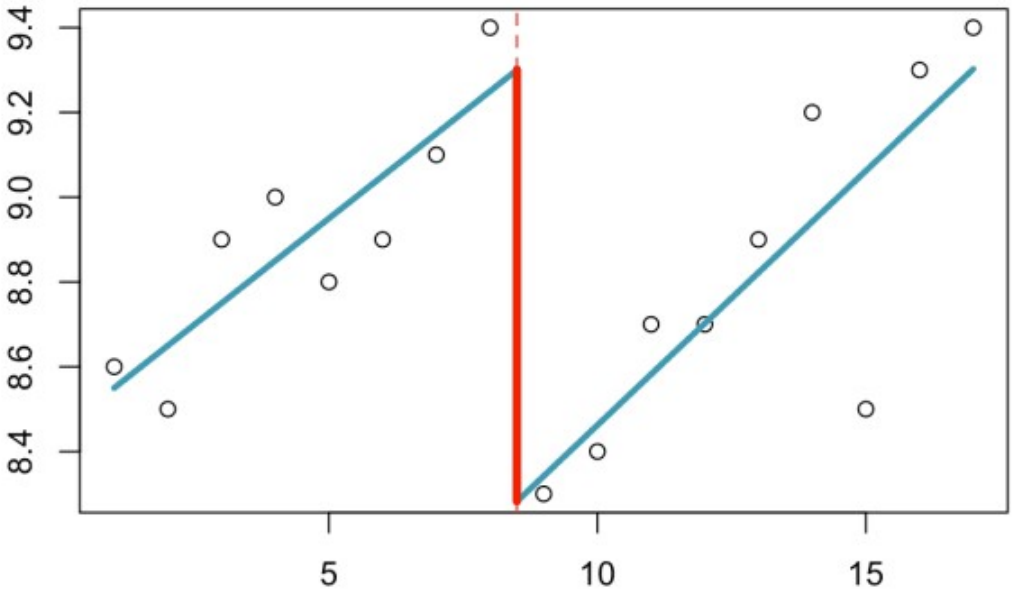


x David Simon's The Wire,

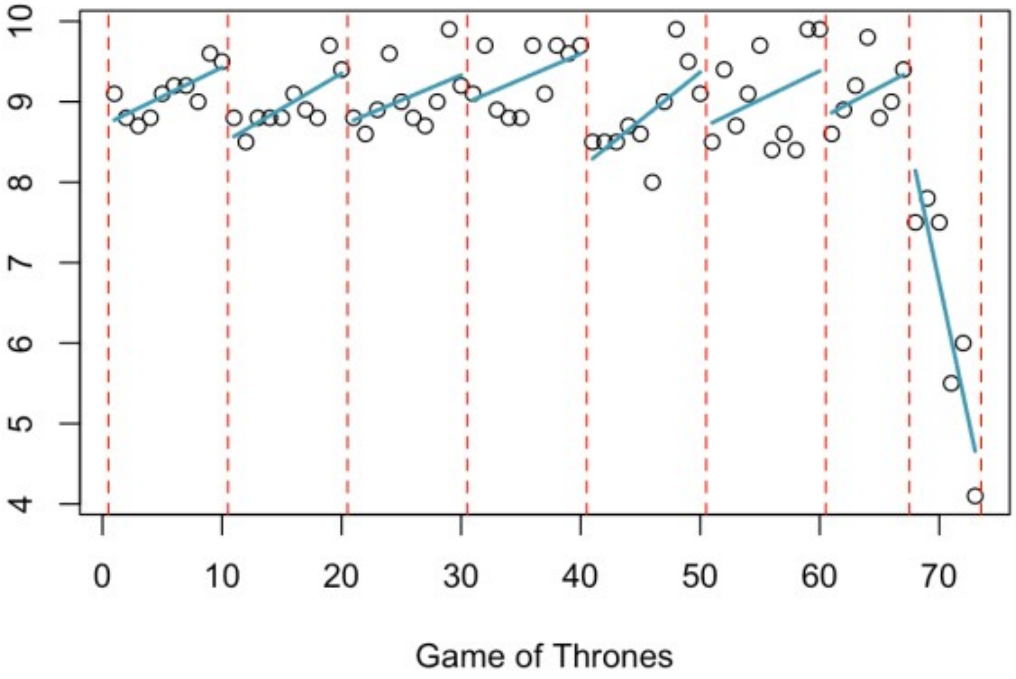


If we loop again over all our series, we have 485 pairs of consecutive seasons. As expected, in 75% of the casse, from season t-1 to season t, we observe a negative rupture. As previously, in 70% of the cases, it is not significat (with linear models before and after), and when it is significant, it is negative in 96% of the cases !

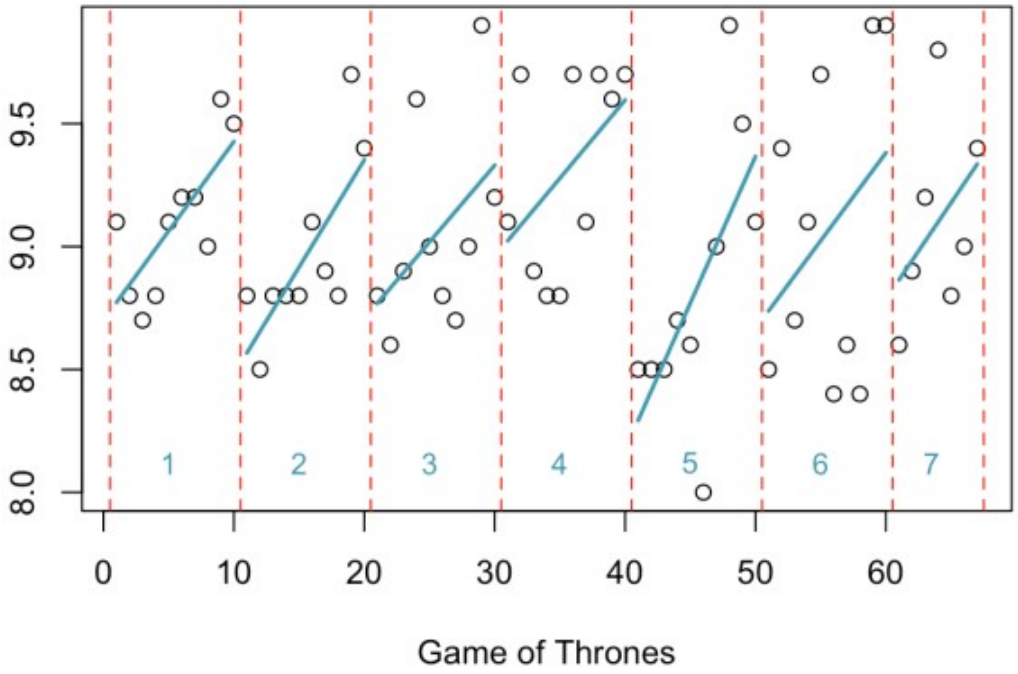
xxxxxxxx



x David Benioff and D. B. Weiss's Game of Thrones,


10
9
8
7
6
5
4
0
10
20
30
40
50
60
70
Game of Thrones

x


9.5
9.0
8.5
8.0
1
2
3
4
5
6
7
0
10
20
30
40
50
60
Game of Thrones