# Joinpoint regression with R

*Juan R Gonzalez*

## Contents



## 1 Introduction

**Objectives**

- Understand the concept of Jointpoint regression
- Learn how to perform Jointpoint regression with R
- Peform data analyses where the scientific question is to determine changes in temporal trends (incidence or mortality rates)

## 2 Package installation

Package can be installed from CRAN

```
install.packages("ljr")
```

After that it is loaded as usual

```
library("ljr")
```

It contains the following functions:

```
ls(2)
## [1] "ljr0"  "ljr01" "ljr1"  "ljr11" "ljrb"  "ljrf"  "ljrjk" "ljrk"  "ljrkk"
```

- *kcm*: Kentucky yearly cancer mortality from 1999-2005
- *ljr0*: MLE with 0 joinpoints
- *ljr01*: Perform test of 0 vs 1 joinpoints}

- *ljr1*: MLE with 1 joinpoint}
- *ljr11*: Test coefficients conditioned on K=1 joinpoint}
- *ljrb*: Perform backward joinpoint selection algorithm with upper bound K}
- *ljrf*: Perform forward joinpoint selection algorithm with unlimited upper bound
- *ljrjk*: Perform test of j vs k joinpoints
- *ljrk*: MLE with k joinpoints}
- *ljrkk*: Test coefficients conditioned on K=k joinpoint}

# 3 Data analysis

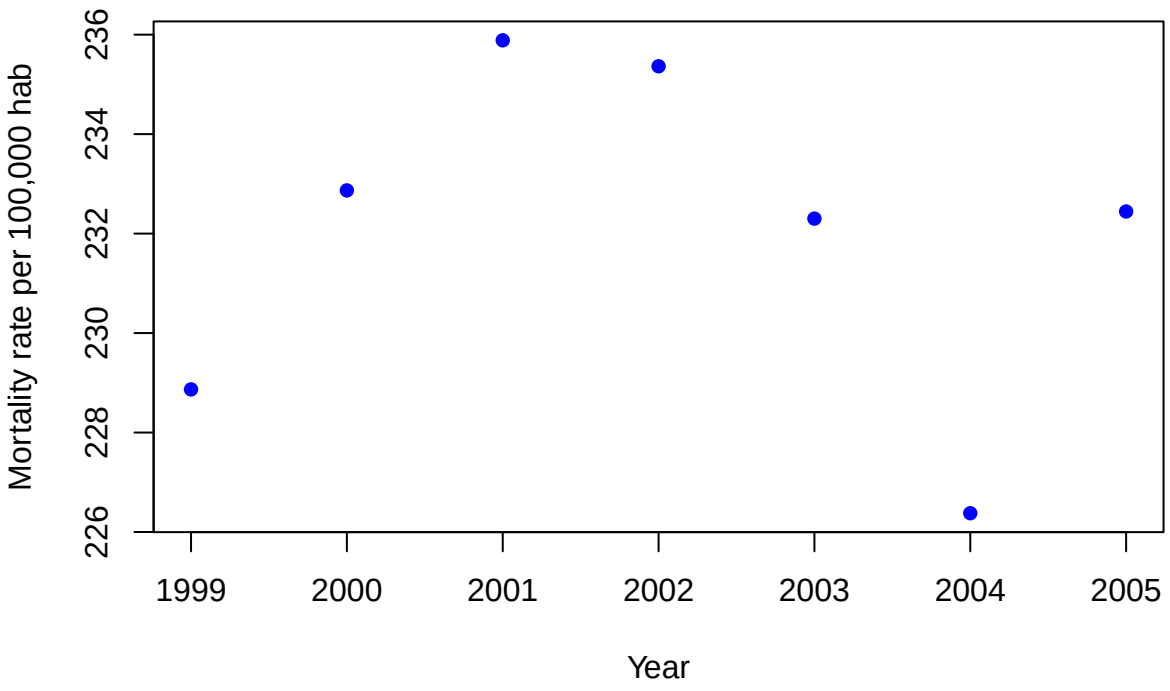
Let us use data available at the package

```
 data(kcm)
 head(kcm)
##   Year Count Population
## 1 1999  9196    4018053
## 2 2000  9412    4041769
## 3 2001  9595    4067643
## 4 2002  9624    4088977
## 5 2003  9558    4114489
## 6 2004  9373    4140427
```

## 3.1 Data visualization

As in any other statistical data analysis, first, let us have a look at the data

```
kcm$rate <- (kcm$Count/kcm$Population)*100000
plot(kcm$Year, kcm$rate, xlab="Year",
     ylab="Mortality rate per 100,000 hab",
     type="n")
points(kcm$Year, kcm$rate, pch=16, col="blue")
```




## 3.2 Simple trend analysis

One may be interested in estimating a Poisson model to determine whether there is any change in global tren. Here the null hypothesis is that slope is 0.

```
modPoisson <- glm(Count~Year+offset(log(Population)),
                  family=poisson, data=kcm)
modPoisson
##
## Call:  glm(formula = Count ~ Year + offset(log(Population)), family = poisson,
##     data = kcm)
##
## Coefficients:
## (Intercept)         Year
##  -4.1972066   -0.0009335
##
## Degrees of Freedom: 6 Total (i.e. Null);  5 Residual
## Null Deviance:       12.2
## Residual Deviance: 11.96     AIC: 92.94
```

Hence, the annual percentage change would be:

```
round((exp(modPoisson$coef[2])-1)*100, 2)
##  Year
## -0.09
```

with confidence interval

```
round((exp(confint(modPoisson)[2,])-1)*100, 2)
##  2.5 % 97.5 %
##  -0.47   0.29
```

In some ocassions there is overdispersion and negative binomial (NB) distribution must be used instead. Dispersion can be estimated by using residual deviance (null deviance / df residual). When this coefficient is $>1$ Poisson distribution is not adequate and NB has to be used. Here you can see an approximate test. Other better tests can be found in the library `pscl`.

$$H_0 : \text{There is no overdispersion}$$

The p-value corresponding to this test can be obtained by

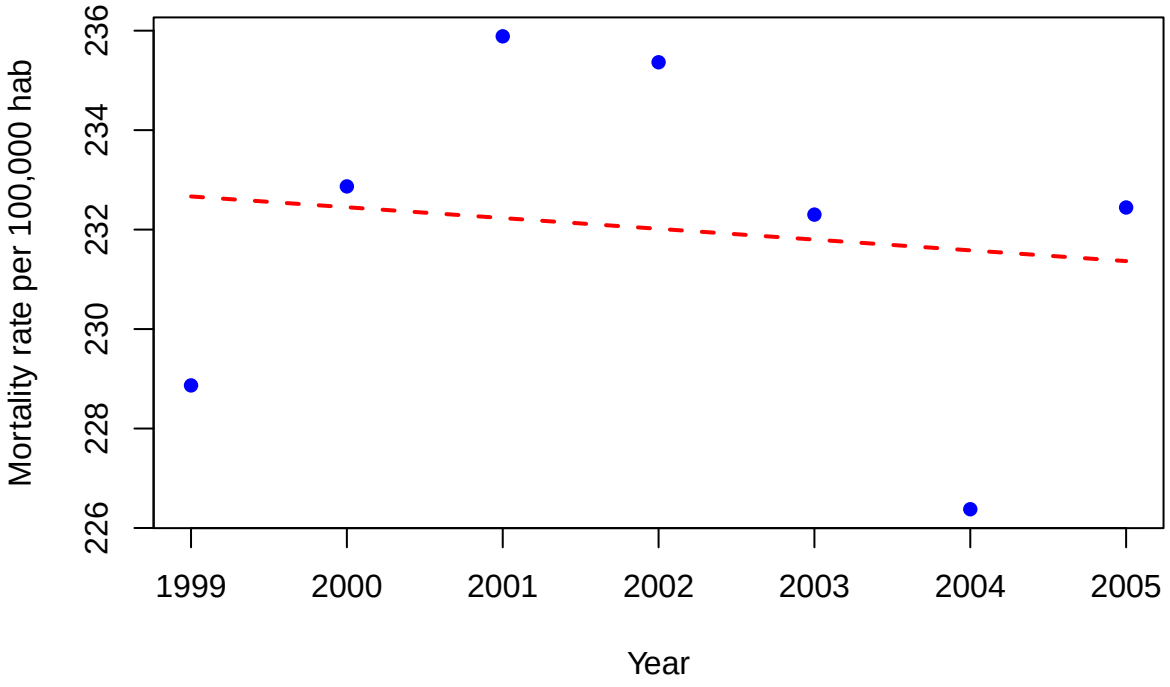
```
1 - pchisq(modPoisson$deviance, modPoisson$df.res)
## [1] 0.03527432
modPoisson
##
## Call:  glm(formula = Count ~ Year + offset(log(Population)), family = poisson,
##     data = kcm)
##
## Coefficients:
## (Intercept)         Year
##  -4.1972066   -0.0009335
##
## Degrees of Freedom: 6 Total (i.e. Null);  5 Residual
## Null Deviance:       12.2
## Residual Deviance: 11.96     AIC: 92.94
```

When rejecting the null hypothesis, negative binomial model can be fitted as

```
library(MASS)
modNB <- glm.nb(Count~Year+ offset(log(Population)),
                data=kcm)
modNB
##
## Call:  glm.nb(formula = Count ~ Year + offset(log(Population)), data = kcm,
##     init.theta = 13398.06067, link = log)
##
## Coefficients:
## (Intercept)         Year
##  -4.2218596   -0.0009212
##
## Degrees of Freedom: 6 Total (i.e. Null);  5 Residual
## Null Deviance:       7.141
## Residual Deviance: 7.01  AIC: 93.73
```

However, as illustrated in the next plot, linear trend is not valid

```
counts.pred <- predict(modPoisson, type="response")
rate.pred <- (counts.pred/kcm$Population)*100000
plot(kcm$Year, kcm$rate, xlab="Year",
     ylab="Mortality rate per 100,000 hab",
     type="n")
points(kcm$Year, kcm$rate, pch=16, col="blue")
lines(kcm$Year, rate.pred, lwd=2, lty=2, col="red")
```




# 4 Joinpoint analysis

## 4.1 Model for a given number of joinpoints

In general, one may be interested in estimating the best model for a given number of joinpoints (obtained by visualy inspecting the overal tren). For example, for 1 joinpont the model is estimated with the function `ljrk`

```
 ljrk(1, kcm$Count, kcm$Population, kcm$Year+.5)
## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t+g1*max(t-tau1,0)
##
##        Variables         Coef
## b0     Intercept -40.81272431
## g0             t   0.01737196
## g1 max(t-tau1,0)  -0.02418284
##
## Joinpoints:
##
## 1 tau1= 2001.273
## $Coef
##     Intercept             t max(t-tau1,0)
##  -40.81272431    0.01737196   -0.02418284
##
```

```
## $Joinpoints
##    tau1=
## 2001.273
##
## $wlik
## [1] -0.112523
```

Notice that we have used the Year variable +0.5 for interpreting purposes. The model for 2 joinpoints in estimated by

```
 ljrk(2, kcm$Count, kcm$Population, kcm$Year+.5)
## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t+g1*max(t-tau1,0)+g2*max(t-tau2,0)
##
##        Variables         Coef
## b0     Intercept -36.31521597
## g0             t   0.01512302
## g1 max(t-tau1,0)  -0.03460547
## g2 max(t-tau2,0)   0.04383173
##
## Joinpoints:
##
## 1 tau1= 2004.500
## 2 tau2= 2002.039
## $Coef
##     Intercept             t max(t-tau1,0) max(t-tau2,0)
##  -36.31521597    0.01512302   -0.03460547    0.04383173
##
## $Joinpoints
##    tau1=    tau2=
## 2004.500 2002.039
##
## $wlik
## [1] -0.1125223
```

## 4.2 Model selection using LRT

When two different models may fit the data, we can select the best one by using a likelihood ratio test (LRT) where the p-value is computed by using Montecarlo method. The function for that purpose is `ljrjk` and can be fitted in the case of comparing 1 an 2 joinpoints by executing

```
 ljrjk(1, 2, kcm$Count, kcm$Population, kcm$Year+.5,
             R=1000,alpha=.05)
## Testing H0:  1  joinpoint(s) vs. H1:  2  joinpoints
## p-value= 0.029
## Null hypothesis is rejected
##
## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t+g1*max(t-tau1,0)+g2*max(t-tau2,0)
##
##        Variables         Coef
## b0     Intercept -36.31521597
## g0             t   0.01512302
```

```
## g1 max(t-tau1,0)  -0.03460547
## g2 max(t-tau2,0)   0.04383173
##
## Joinpoints:
##
## 1 tau1= 2004.500
## 2 tau2= 2002.039
## $Coef
##     Intercept             t max(t-tau1,0) max(t-tau2,0)
##  -36.31521597    0.01512302   -0.03460547    0.04383173
##
## $Joinpoint
##    tau1=    tau2=
## 2004.500 2002.039
##
## $wlik
## [1] -0.1125223
##
## $pval
## [1] 0.029
```

Here we can conclude that

Once the model is selected, we can compute the anual percentage of change of each segment by

```
mod <- ljrk(2, kcm$Count, kcm$Population, kcm$Year+.5)
## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t+g1*max(t-tau1,0)+g2*max(t-tau2,0)
##
##        Variables         Coef
## b0     Intercept -36.31521597
## g0             t   0.01512302
## g1 max(t-tau1,0)  -0.03460547
## g2 max(t-tau2,0)   0.04383173
##
## Joinpoints:
##
## 1 tau1= 2004.500
## 2 tau2= 2002.039
cbind(year=c(1999, mod$Joinpoints),
      APC=round((exp(mod$Coef[-1])-1)*100,2))
##           year   APC
##       1999.000  1.52
## tau1= 2004.500 -3.40
## tau2= 2002.039  4.48
```

## 4.3 Automatic method of model selection

In some ocassion the user is interested in selecting the best modelby using an automatic method by using backward or forward approaches Si queremos escoger el modelo de forma automática, utilizamos un método de selección 'backward' o 'forward'. This can be performed by using the function `ljrb` or `ljrf`. Note that in 'barckward' method the argument `K` is required. This indicates the maximum number of joinpoints to be tested. The 'forward' method starts from the null model (e.g. linear trend or no joinpoints) and tests whether the inclusion of a new joinpoint is statistically significant or not.

```
ljrb(K=3, kcm$Count, kcm$Population, kcm$Year+.5)
## Backward algorithm for determining the number of joinpoints:
## Step  1 : Test H0: 0  joinpoint(s) vs H1: 3  joinpoint(s)
## p-value=  0.018
## Step  2 : Test H0: 0  joinpoint(s) vs H1: 2  joinpoint(s)
## p-value=  0.011
## Step  3 : Test H0: 1  joinpoint(s) vs H1: 2  joinpoint(s)
## p-value=  0.044
##
##  Estimated number of joinpoints=  1
##
## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=ofst+b0+g0*t+g1*max(t-tau1,0)
##
##        Variables         Coef
## b0     Intercept -40.81272431
## g0             t   0.01737196
## g1 max(t-tau1,0)  -0.02418284
##
## Joinpoints:
##
## 1 tau1= 2001.273
## $Coef
##     Intercept             t max(t-tau1,0)
##  -40.81272431    0.01737196   -0.02418284
##
## $Joinpoints
##    tau1=
## 2001.273
##
## $wlik
## [1] -0.112523
##
## $pvals
## [1] 0.018 0.011 0.044
```

```
ljrf(kcm$Count, kcm$Population, kcm$Year+.5)
## Forward algorithm for determining the number of joinpoints:
## Step  1 : Test H0: 0  joinpoint(s) vs H1: 1  joinpoint(s)
## p-value=  0.057
##
##  Estimated number of joinpoints=  0
##
## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t
##
##    Variables         Coef
## b0 Intercept -4.190069631
## g0         t -0.000935695
## $Coef
##    Intercept            t
## -4.190069631 -0.000935695
```

```
##
## $wlik
## [1] -0.1125237
##
## $pvals
## [1] 0.057 0.000 0.000 0.000 0.000
```

# 5 Exercise (to deliver)

---

**Exercise 1:** The file `mamaCat.txt` (available at https://github.com/isglobal-brge/TeachingMaterials/tree/master/Longitudinal_data_analysis/data) contains data about breast cancer mortality in women between 35 and 65 years old in Catalonia. Rates correspond to the period 1975-1997. Each column contains the next variables: year of mortality, number of deaths and at-risk population. Perform next tasks:

- Create a plot to visualize temporal tren
- Estimate a Poisson model (or another one that controls for overdispersion if exists) and estimate the percentage of annyal change in the mortality rates.
- Visualy detect how many changes of trens (e.g. joinpoints) whold fit the data
- Perform a test to determine whether the selected model is better than the one considering a single trend
- Use an automatic method to analize these data and determine how many joinpoints are neccesaries to model the temporal trend
- In case of having several trend changes, estimate the APC of each segment

**Exeercise 2** The file `pulmonCat.txt` (also available in the same repository) contains data about lung cancer mortality in males and females in Catalonia of the period 1975-1997. Each column contains the next variables: gender, year of mortality, number of deaths and at-risk population. Perform next tasks:

- Temporal trend analysis for males and females
- Which are your conclusions from the obtained results?

---

# 6 References


- The `pscl` package


# 7 Session information

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 14393)
##
## locale:
## [1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
## [3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Spain.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
## [1] MASS_7.3-45     ljr_1.4-0       knitr_1.15.1    BiocStyle_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] backports_1.0.5 magrittr_1.5    rprojroot_1.2   tools_3.3.2
##  [5] htmltools_0.3.5 yaml_2.1.14     Rcpp_0.12.9     stringi_1.1.2
##  [9] rmarkdown_1.3   stringr_1.1.0   digest_0.6.11   evaluate_0.10
```