Generalized linear models

Juan R Gonzalez

BRGE - Bioinformatics Research Group in Epidemiology
Barcelona Institute for Global Health (ISGlobal)
e-mail:juanr.gonzalez@isglobal.org
http://brge.isglobal.org
and Departament of Mathematics, UAB

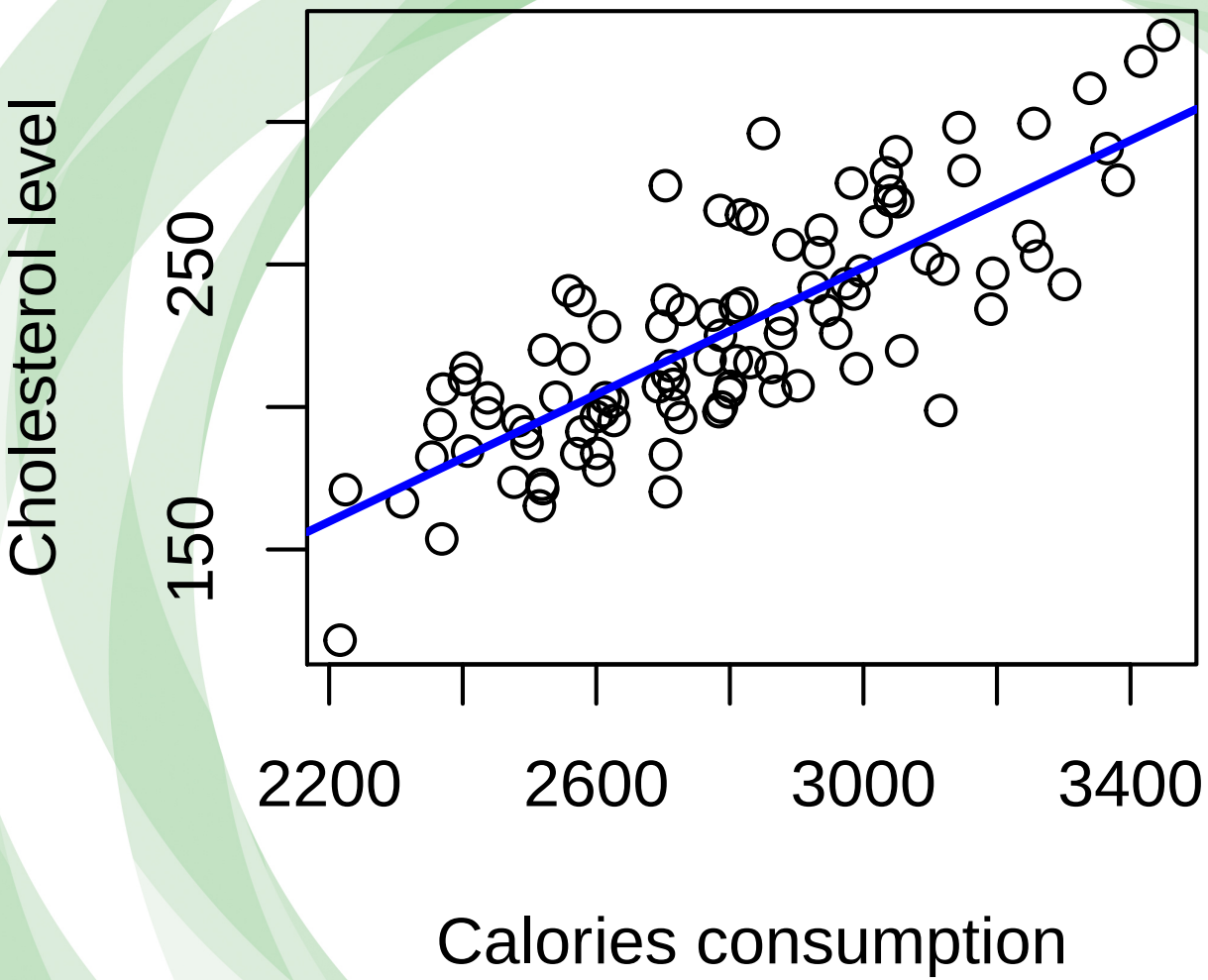
Outline

- Linear regression
- Logistic regression
- Poisson and negative binomial regression
- Jointpoint regression
- Survival analysis
- GEE and linear mixed models

Regression modeling

Outcome	Method	Example
Continuous	Linear regression	Factors that affects cholesterol levels
Binary	Logistic regression	Factors that affects developing cancer
Count	Poisson and NB regression	Incidence and mortality trends
Count	Joinpoint regression	Changes in longitudinal data
Time to event	Survival	Factors that affect time until developing cancer
All	Repeated/clustered measures	Factors that affect outcome complex data structure

Linear regression



Linear regression

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

- α correspond to the mean level of Y in the population
- β_j indicates the change in Y when X_j changes in 1 unit (after keeping the rest of X_k fixed)

Linear regression

Example: Researchers are interested in knowing the factors that better explain air Ozone levels (variable `Ozone` in data frame `airquality`). They measure solar radiation (`Solar.R`), average wind (`Wind`) and temperature (`Temp`) in different months ((`Months`) for 154 observations.

```
data(airquality)
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Linear regression

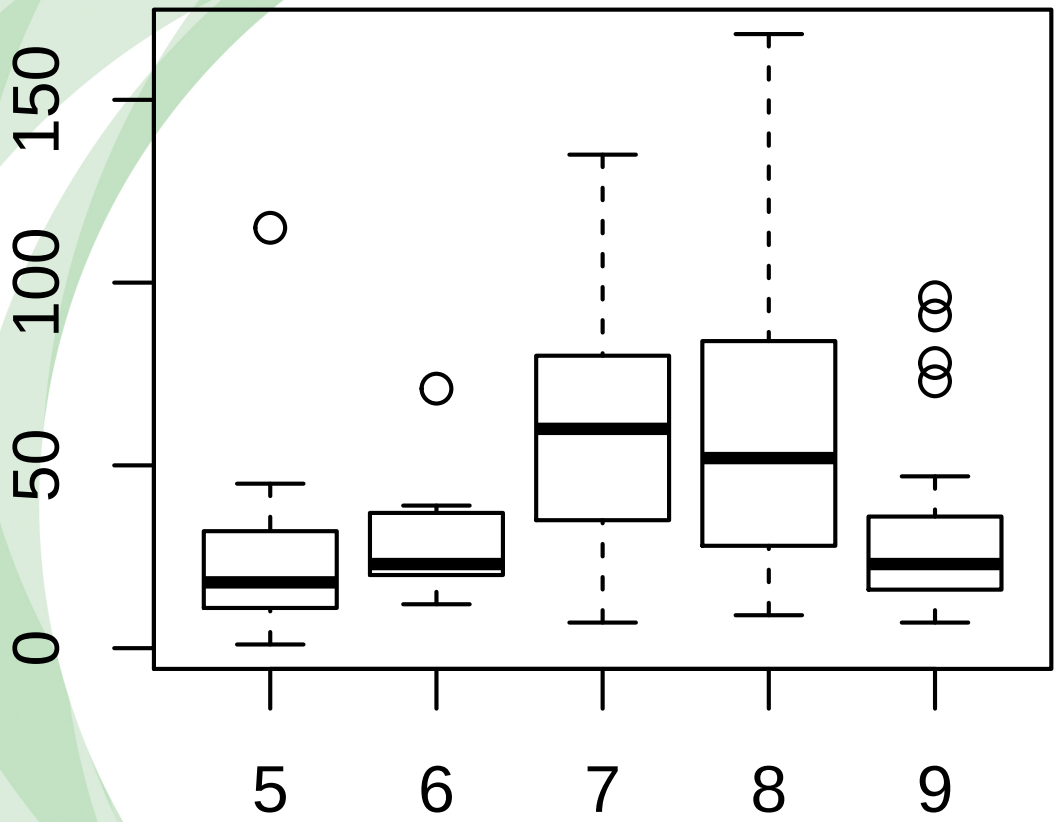
Simple linear regression

```
mod <- lm(Ozone ~ Temp, data=airquality)
summary(mod)

##
## Call:
## lm(formula = Ozone ~ Temp, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.729 -17.409  -0.587  11.306 118.271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -146.9955    18.2872  -8.038 9.37e-13 ***
## Temp           2.4287     0.2331  10.418  < 2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
## Residual standard error: 23.71 on 114 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.4877,Adjusted R-squared:  0.4832
## F-statistic: 108.5 on 1 and 114 DF,  p-value: < 2.2e-16
```

Multiple linear regression

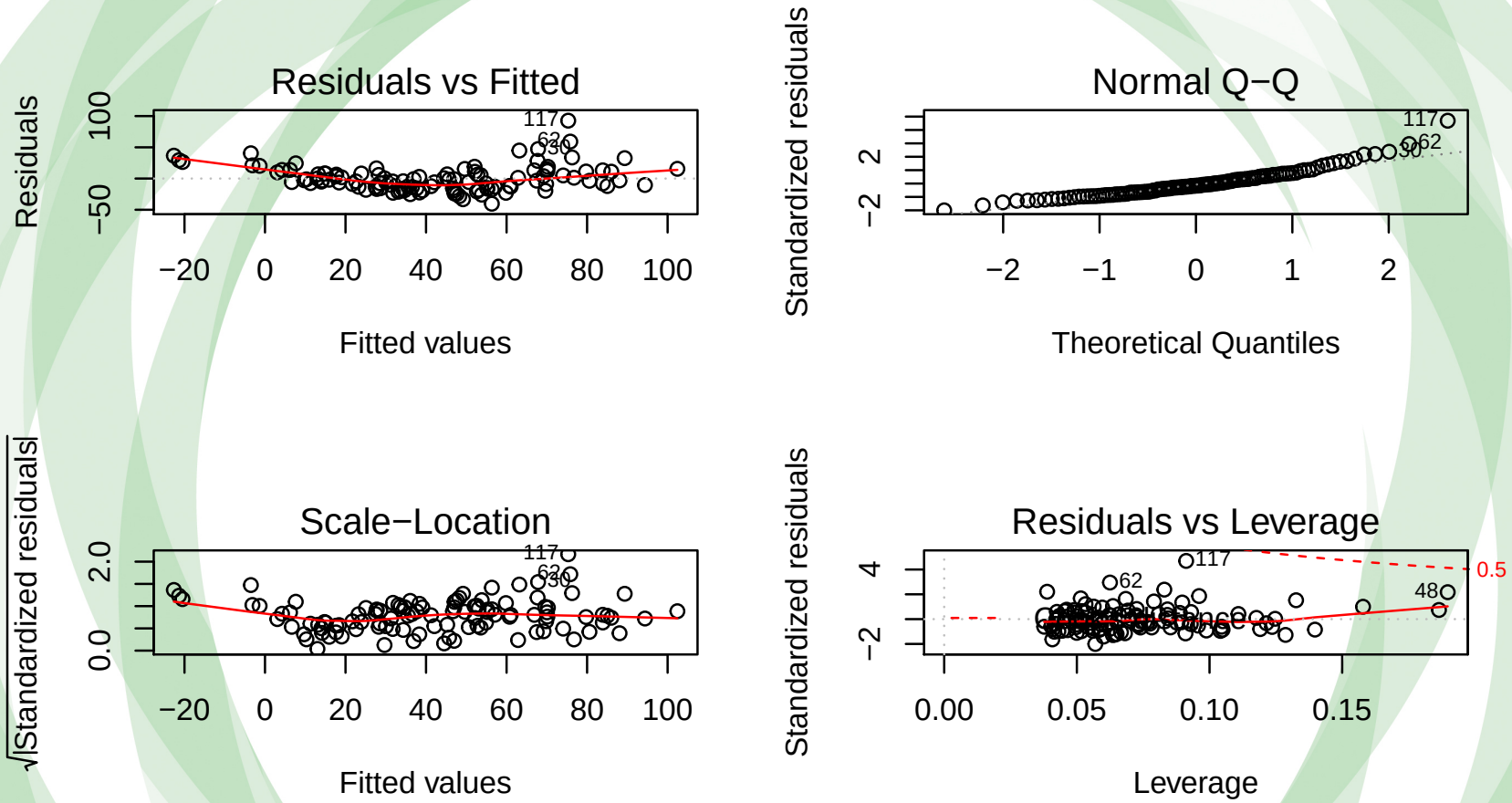
```
mod <- lm(Ozone ~ Solar.R + Wind + Temp + as.factor(Month),
          data=airquality)
summary(mod)

##
## Call:
## lm(formula = Ozone ~ Solar.R + Wind + Temp + as.factor(Month),
##     data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.344 -13.495  -3.165  10.399  92.689
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -74.23481   26.10184  -2.844  0.00537 **
## Solar.R             0.05222    0.02367   2.206  0.02957 *
## Wind               -3.10872    0.66009  -4.710 7.78e-06 ***
## Temp                1.87511    0.34073   5.503 2.74e-07 ***
## as.factor(Month)6 -14.75895    9.12269  -1.618  0.10876
## as.factor(Month)7  -8.74861    7.82906  -1.117  0.26640
## as.factor(Month)8  -4.19654    8.14693  -0.515  0.60758
## as.factor(Month)9 -15.96728    6.65561  -2.399  0.01823 *
##
```

Interpretation of categorical factors

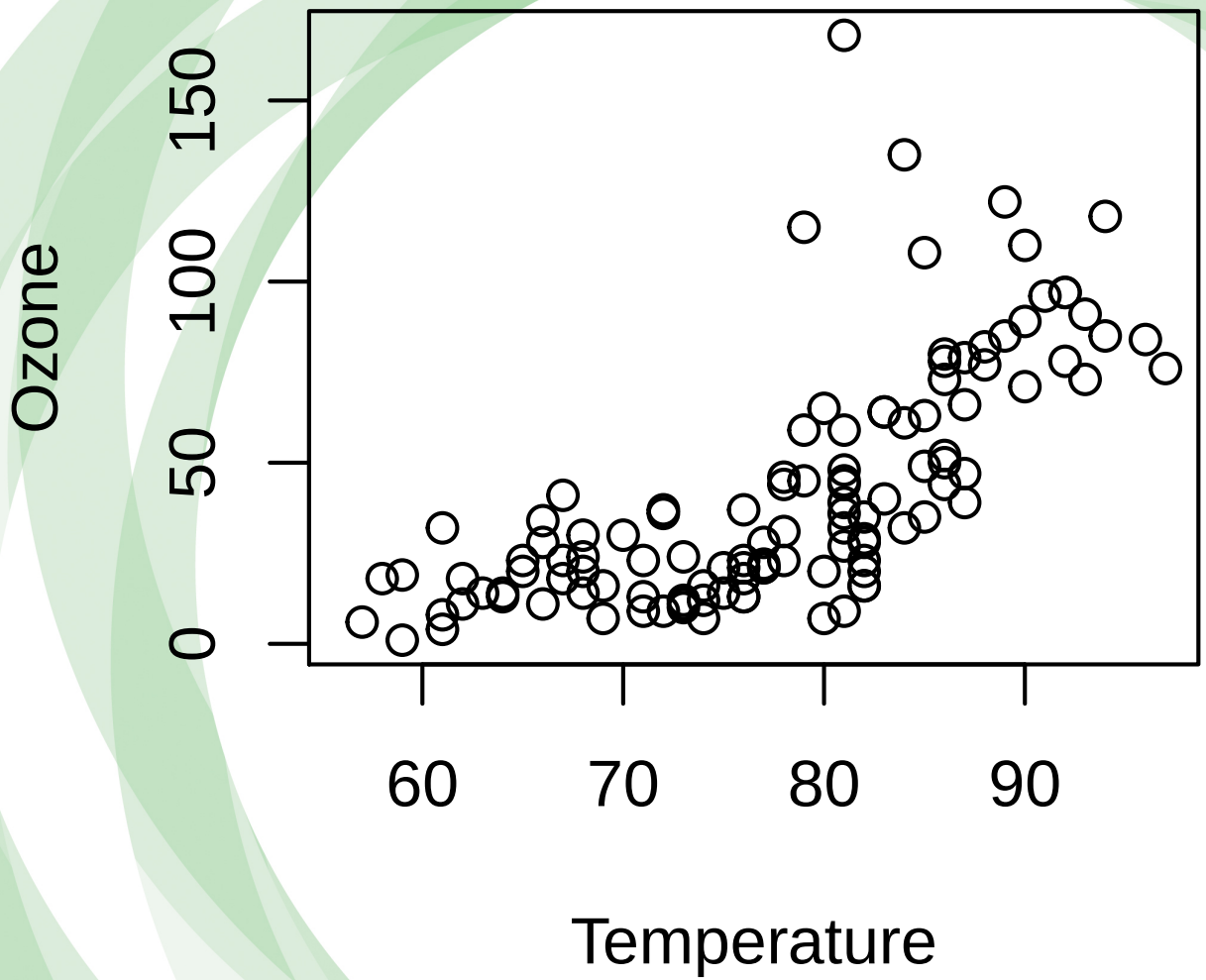


Model validation

```
par(mfrow=c(2,2))
plot(mod)
```



Linear regression



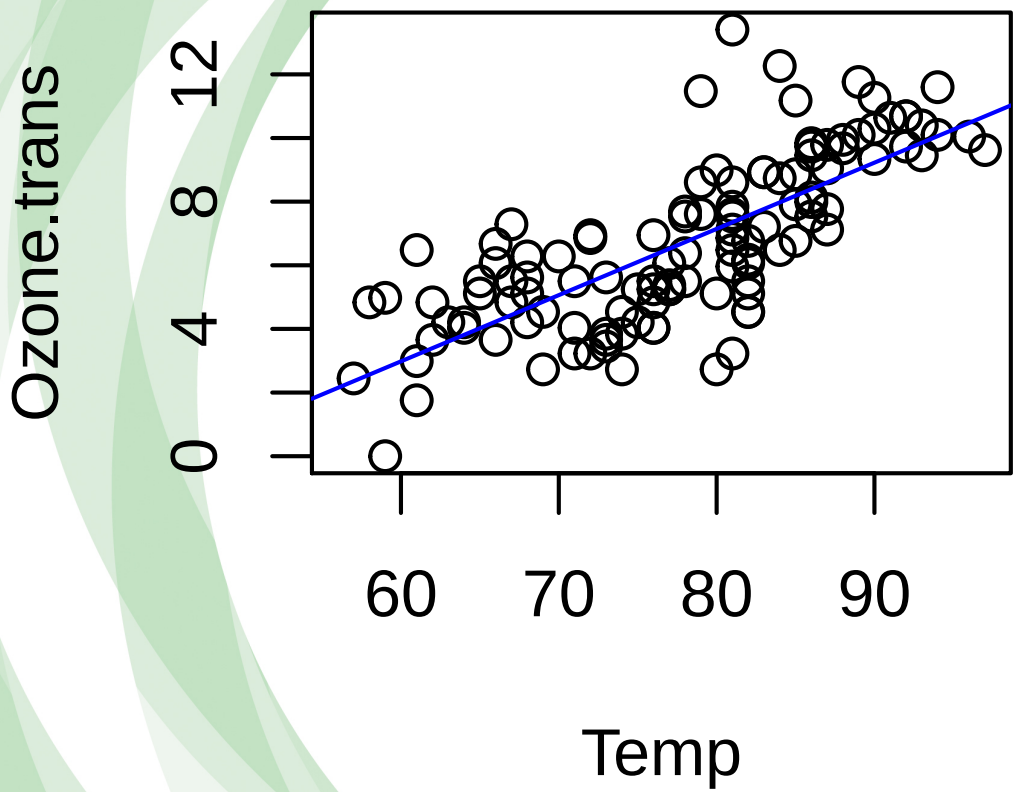
Linear regression

```
require(car)

## Loading required package:  car
## Loading required package:  carData

trans <- powerTransform(mod)
trans

## Estimated transformation parameter
##        Y1
## 0.2206725
```

```
Ozone.trans <- bcPower(airquality$Ozone,
                       coef(trans, round=TRUE))

mod.trans <- lm(Ozone.trans ~ Temp, data=airquality)
```

Linear regression

```
plot(Ozone.trans ~ Temp, data=airquality)
abline(mod.trans, col="blue")
```



Linear regression

Model validity can be measured by computing R^2

```
summary(mod)

##
## Call:
## lm(formula = Ozone ~ Temp, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.729 -17.409  -0.587  11.306 118.271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -146.9955    18.2872  -8.038 9.37e-13 ***
## Temp           2.4287     0.2331  10.418  < 2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
## Residual standard error: 23.71 on 114 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.4877,Adjusted R-squared:  0.4832
## F-statistic: 108.5 on 1 and 114 DF,  p-value: < 2.2e-16
```

Linear regression

```
summary(mod.trans)

##
## Call:
## lm(formula = Ozone.trans ~ Temp, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4144 -1.2733  0.0883  1.1028  6.0558
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.5085     1.3495  -7.046 1.49e-10 ***
## Temp          0.2082     0.0172  12.099  < 2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
## Residual standard error: 1.75 on 114 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.5622,Adjusted R-squared:  0.5584
## F-statistic: 146.4 on 1 and 114 DF,  p-value: < 2.2e-16
```

Logistic regression

Y variable is binary (case/control, relapse/non-relapse, mortality, ...).
In that case, the logit transformation guarantees linearity.

$$\log(p(Y=1)/(1-p(Y=1))) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k$$

$\exp(\beta_k)$ can be interpreted as the odds ratio (OR) of having/developing/being $Y = 1$

Logistic regression

Example: Reserchers are interested in determining whether a new treatment (varible `rx`) reduces mortality (variable `fustat`) in patients diagnosed with ovarian cancer. Data are available by typing:

```
data(ovarian, package="survival")
head(ovarian)
```

```
##   futime fustat     age resid.ds rx ecog.ps
## 1     59      1 72.3315        2  1       1
## 2    115      1 74.4932        2  1       1
## 3    156      1 66.4658        2  1       2
## 4    421      0 53.3644        2  2       1
## 5    431      1 50.3397        2  1       1
## 6    448      0 56.4301        1  1       2
```

Logistic regression

```
mod2 <- glm(fustat ~ rx, data=ovarian, family="binomial")
summary(mod2)

##
## Call:
## glm(formula = fustat ~ rx, family = "binomial", data = ovarian)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2435  -0.9854  -0.9854   1.1127   1.3824
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7783     1.2502   0.623    0.534
## rx           -0.6242     0.7966  -0.784    0.433
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35.890  on 25  degrees of freedom
## Residual deviance: 35.268  on 24  degrees of freedom
## AIC: 39.268
##
## Number of Fisher Scoring iterations: 4
```

Stepwise regression

```
modAll <- glm(fustat ~ ., data=ovarian, family="binomial")
modBest <- MASS::stepAIC(modAll)

## Start:  AIC=25.7
## fustat ~ futime + age + resid.ds + rx + ecog.ps
##
##             Df Deviance    AIC
## - rx         1   13.726 23.727
## - age        1   13.913 23.913
## - resid.ds   1   14.921 24.921
## <none>           13.700 25.701
## - ecog.ps    1   19.098 29.098
## - futime     1   25.179 35.179
##
## Step:  AIC=23.73
## fustat ~ futime + age + resid.ds + ecog.ps
##
##             Df Deviance    AIC
## - age        1   14.002 22.002
## - resid.ds   1   14.951 22.951
## <none>           13.726 23.727
## - ecog.ps    1   19.104 27.104
## - futime     1   26.115 34.115
##
```

Stepwise regression

```
summary(modBest)

##
## Call:
## glm(formula = fustat ~ futime + ecog.ps, family = "binomial",
##     data = ovarian)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.16625  -0.16213  -0.01266   0.43289   1.34125
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.121383   2.293838   0.925   0.3551
## futime      -0.012371   0.005812  -2.129   0.0333 *
## ecog.ps      2.833300   1.698989   1.668   0.0954 .
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35.890  on 25  degrees of freedom
## Residual deviance: 15.124  on 23  degrees of freedom
## AIC: 21.124
```

Poisson regression

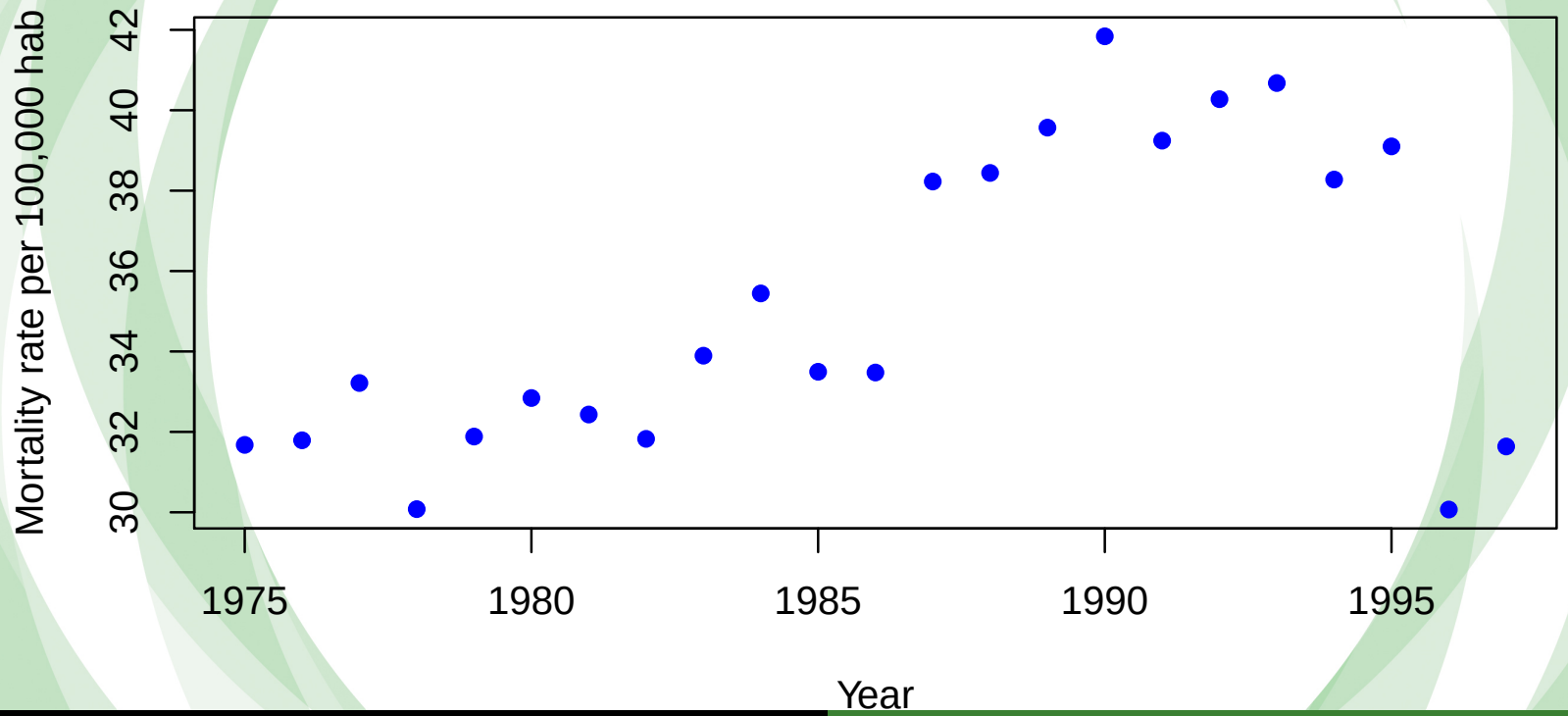
Let us analyze breast cancer mortality rates in Catalonia from 1943-1993. Our aim is to study the evolution of those rates.

```
breast <- read.delim("../../data/breastCat.txt")
head(breast)

##   year deaths population
## 1 1975     49     154692
## 2 1976     50     157279
## 3 1977     50     150531
## 4 1978     48     159583
## 5 1979     51     159954
## 6 1980     51     155291
```

Poisson regression

Let us visualize the evolution or mortality rates

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



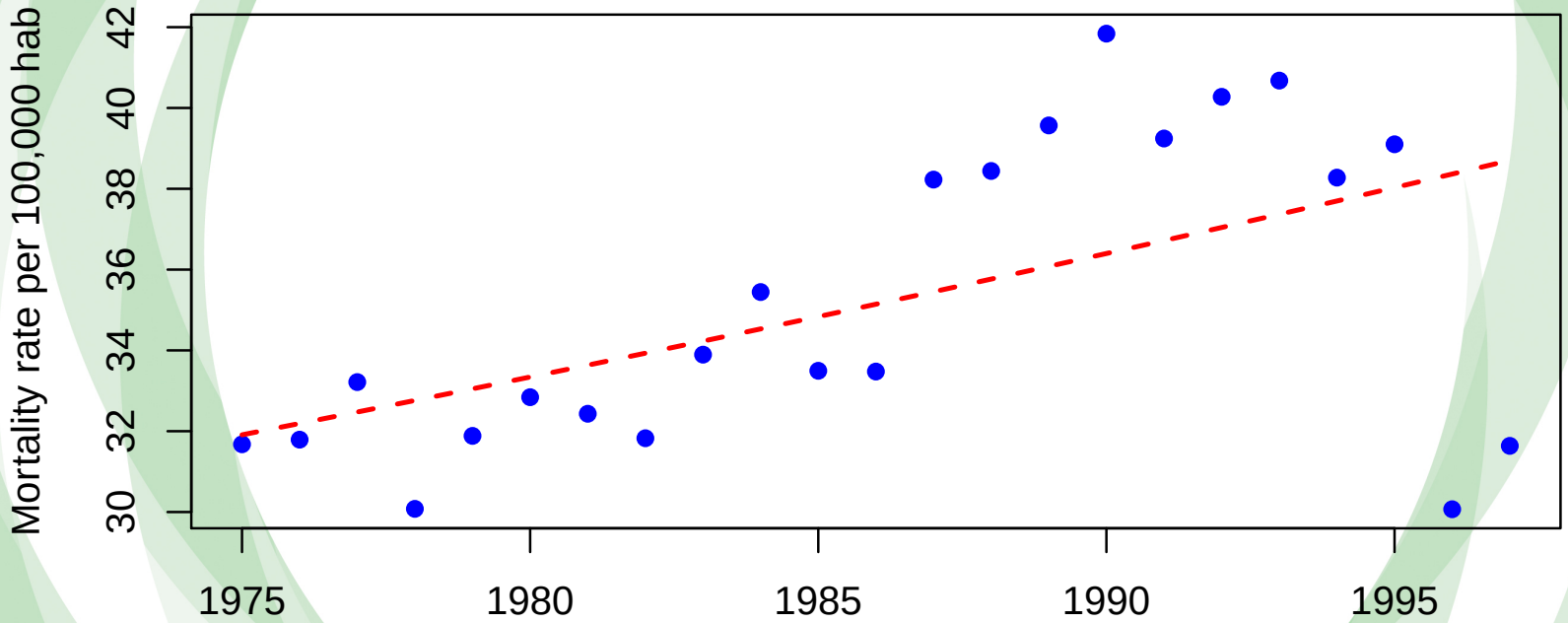
Poisson regression

This model evaluates the overal trend using a Poisson model:

```
modPoisson <- glm(deaths ~ year + offset(log(population)),
                  family=poisson, data=breast)
summary(modPoisson)

##
## Call:
## glm(formula = deaths ~ year + offset(log(population)), family = poi
##     data = breast)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7411  -0.2761  -0.0517   0.5354   1.1393
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.373696   8.431163  -3.010  0.00262 **
## year          0.008771   0.004244   2.067  0.03877 *
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 14.0971  on 22  degrees of freedom
```

Poisson regression

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



Poisson regression

Dispersion can be estimated using the residual deviance. A coefficient >1 indicates that overdispersion is present. The a negative binomial regression is required. An approximated test can be (see others in the library `pscl`)

$$H_0 : \text{There is no-overdispersion} \quad (1)$$

The associated p-value can be obtained by means of:

```
1 - pchisq(modPoisson$deviance, modPoisson$df.res)

## [1] 0.9811347

modPoisson

##
## Call:  glm(formula = deaths ~ year + offset(log(population)), famil
##     data = breast)
##
## Coefficients:
## (Intercept)         year
##  -25.373696     0.008771
##
## Degrees of Freedom: 22 Total (i.e. Null);  21 Residual
```

Negative Binomial regression

```
library(MASS)
modNB <- glm.nb(deaths ~ year + offset(log(population)),
                data=breast)

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit,
trace = control$trace > :  iteration limit reached
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit,
trace = control$trace > :  iteration limit reached

summary(modNB)

##
## Call:
## glm.nb(formula = deaths ~ year + offset(log(population)), data = br
##     init.theta = 2528200.758, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7411  -0.2761  -0.0517   0.5353   1.1393
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.373707   8.431254  -3.009  0.00262 **
## year          0.008772   0.004244   2.067  0.03877 *
##
```

Joinpoint Regression with R

Jointpoint regression aims to fit changes in rates over time. This estimate a regression with 1 change

```
library(ljr)

## ljr 1.4-0 loaded

ljrk(1, breast$deaths, breast$population, breast$year+.5)

## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t+g1*max(t-tau1,0)
##
##        Variables        Coef
## b0     Intercept -42.0406303
## g0             t   0.0171765
## g1 max(t-tau1,0)  -0.1009349
##
## Joinpoints:
##
## 1 tau1= 1993.905
## $Coef
##     Intercept             t max(t-tau1,0)
##   -42.0406303     0.0171765    -0.1009349
##
```

Joinpoint Regression with R

This tests whether a model with 1 joinpoint is statistically significant

```
ljrjk(0, 1, breast$deaths, breast$population, breast$year+.5,
      R = 1000)

## Testing H0:  0  joinpoint(s) vs. H1:  1  joinpoints
## p-value= 0.047
## Null hypothesis is rejected
##
## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t+g1*max(t-tau1,0)
##
##        Variables        Coef
## b0      Intercept -42.0406303
## g0              t   0.0171765
## g1 max(t-tau1,0)  -0.1009349
##
## Joinpoints:
##
## 1 tau1= 1993.905
## $Coef
##     Intercept             t max(t-tau1,0)
##   -42.0406303     0.0171765    -0.1009349
##
```

These are the changes of each segment

```
mod <- ljrk(1, breast$deaths, breast$population, breast$year+.5)

## Model:
## y~Binom(n,p) where p=invlogit(eta)
## eta=b0+g0*t+g1*max(t-tau1,0)
##
##        Variables        Coef
## b0     Intercept -42.0406303
## g0             t   0.0171765
## g1 max(t-tau1,0)  -0.1009349
##
## Joinpoints:
##
## 1 tau1= 1993.905

cbind(year=c(1975, mod$Joinpoints),
      APC=round((exp(mod$Coef[-1])-1)*100,2))

##           year   APC
##       1975.000  1.73
## tau1= 1993.905 -9.60
```