# NRES_798_12_201501

Generalized linear model application

Logistic regression

Poisson regression

Survival analysis

# Generalized linear model

- Linear predictor
- Link function

| Family (error) | Canonical link |
|---|---|
| Normal | Identity |
| Binomial | Logit |
| Poisson | Log |
| Gamma | reciprocal |

- Variance function (error structure)

# Logistic regression

- Simple logistic regression with one predictor variable

$$\beta_0 + age*\beta_1$$

- Example 1
  - Milicer, H. and Szczotka, F., 1966, Age at Menarche in Warsaw girls in 1965, Human Biology, 38, 199-203
  - three variables:
    - "Age" (average age of age homogeneous groups of girls)
    - "Total" (number of girls in each group)
    - "Menarche" (number of girls in the group who have reached menarche)

Type of distribution?

| | Age | Total | Menarche |
|---|---|---|---|
| 1 | 9.21 | 376 | 0 |
| 2 | 10.21 | 200 | 0 |
| 3 | 10.58 | 93 | 0 |
| 4 | 10.83 | 120 | 2 |
| 5 | 11.08 | 90 | 2 |
| 6 | 11.33 | 88 | 5 |
| 7 | 11.58 | 105 | 10 |
| 8 | 11.83 | 111 | 17 |
| 9 | 12.08 | 100 | 16 |
| 10 | 12.33 | 93 | 29 |
| 11 | 12.58 | 100 | 39 |
| 12 | 12.83 | 108 | 51 |
| 13 | 13.08 | 99 | 47 |
| 14 | 13.33 | 106 | 67 |
| 15 | 13.58 | 105 | 81 |
| 16 | 13.83 | 117 | 88 |
| 17 | 14.08 | 98 | 79 |
| 18 | 14.33 | 97 | 90 |
| 19 | 14.58 | 120 | 113 |
| 20 | 14.83 | 102 | 95 |
| 21 | 15.08 | 122 | 117 |
| 22 | 15.33 | 111 | 107 |
| 23 | 15.58 | 94 | 92 |
| 24 | 15.83 | 114 | 112 |
| 25 | 17.58 | 1049 | 1049 |

Data:
Average age of homogeneous group of girls

```
str(menarche)
'data.frame': 25 obs. of 3 variables:
$ Age : num 9.21 10.21 10.58 10.83 11.08
...
$ Total : num 376 200 93 120 90 88 105 111
100 93 ...
$ Menarche: num 0 0 0 2 2 5 10 17 16 29 ...
```

25 observations?

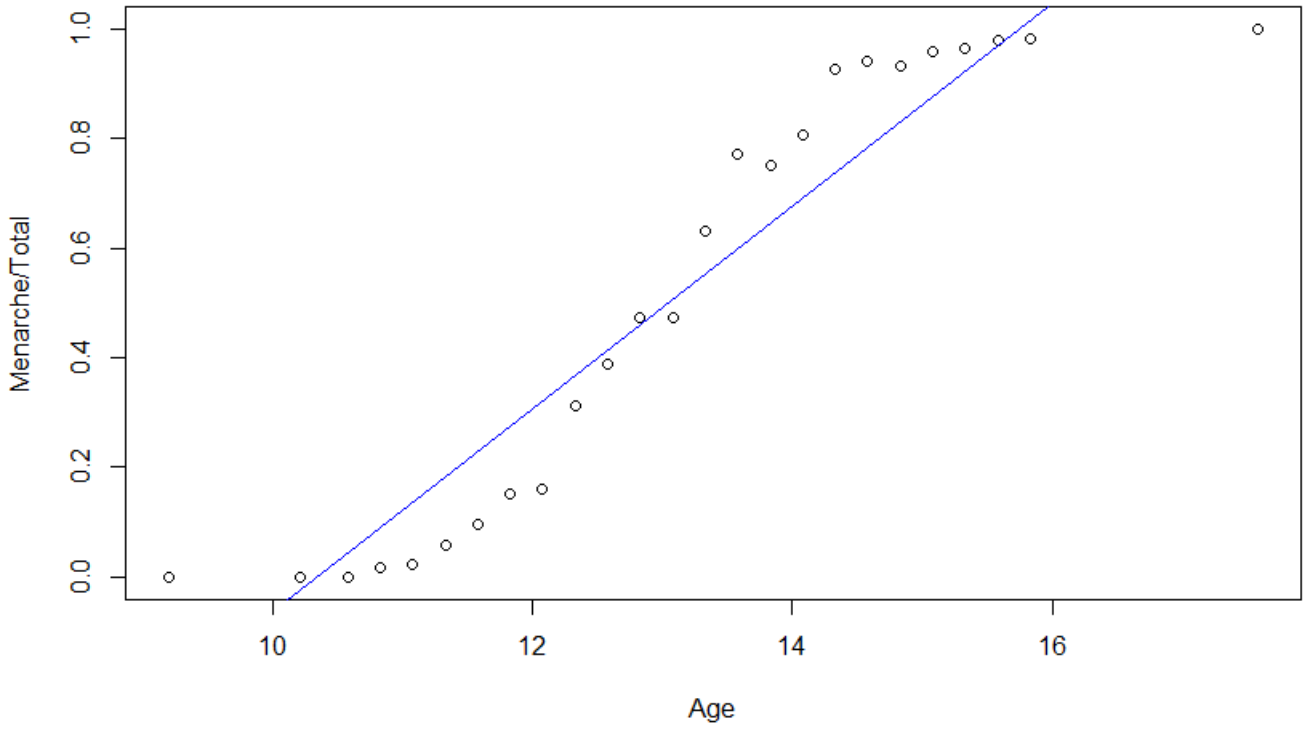
# Logit link function (Binomial family)

$$logit(p) = log\left(\frac{p}{1-p}\right) = \log\left(\frac{p}{q}\right)$$

$$\log\left(\frac{p}{q}\right) = \beta_0 + \beta_1 X + \epsilon_i$$

| | Age | Total | Menarche |
|---|---|---|---|
| 1 | 9.21 | 376 | 0 |
| 2 | 10.21 | 200 | 0 |
| 3 | 10.58 | 93 | 0 |
| 4 | 10.83 | 120 | 2 |
| 5 | 11.08 | 90 | 2 |
| 6 | 11.33 | 88 | 5 |
| 7 | 11.58 | 105 | 10 |
| 8 | 11.83 | 111 | 17 |
| 9 | 12.08 | 100 | 16 |
| 10 | 12.33 | 93 | 29 |
| 11 | 12.58 | 100 | 39 |
| 12 | 12.83 | 108 | 51 |
| 13 | 13.08 | 99 | 47 |
| 14 | 13.33 | 106 | 67 |
| 15 | 13.58 | 105 | 81 |
| 16 | 13.83 | 117 | 88 |
| 17 | 14.08 | 98 | 79 |
| 18 | 14.33 | 97 | 90 |
| 19 | 14.58 | 120 | 113 |
| 20 | 14.83 | 102 | 95 |
| 21 | 15.08 | 122 | 117 |
| 22 | 15.33 | 111 | 107 |
| 23 | 15.58 | 94 | 92 |
| 24 | 15.83 | 114 | 112 |
| 25 | 17.58 | 1049 | 1049 |

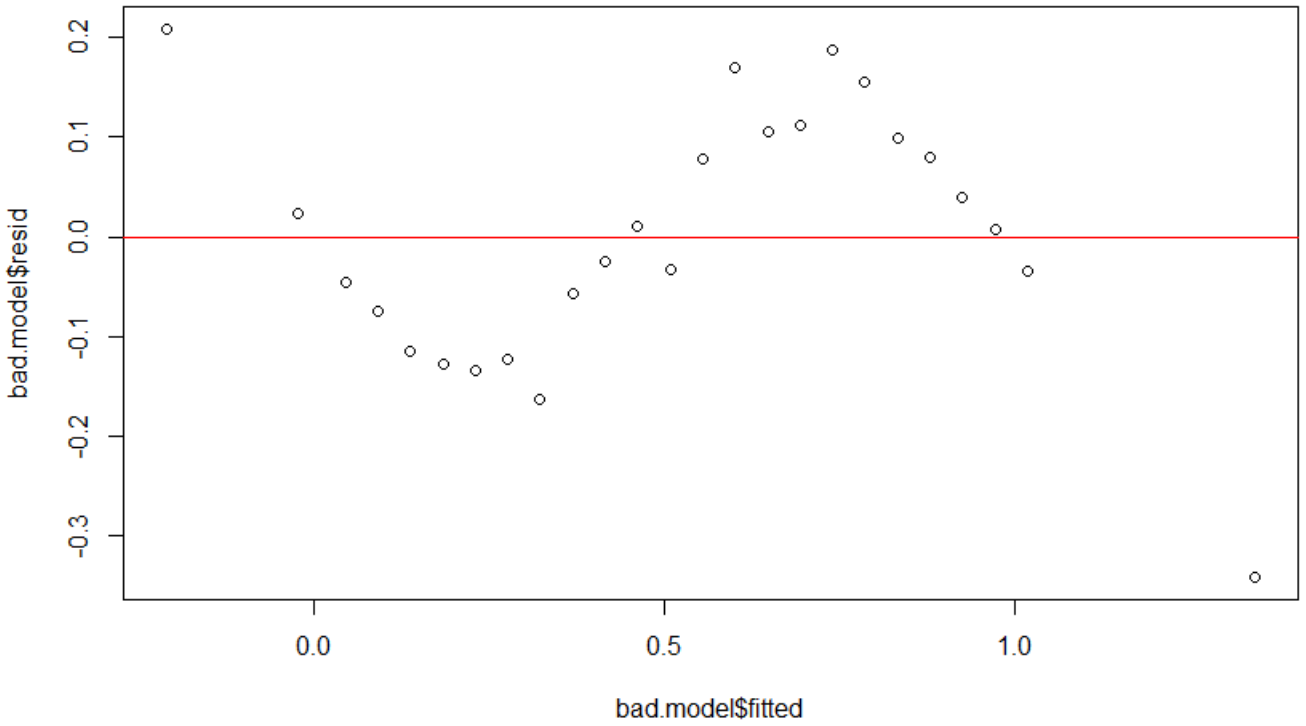
```
plot(Menarche/Total ~ Age, data=menarche)
abline(bad.model,col="blue")
```



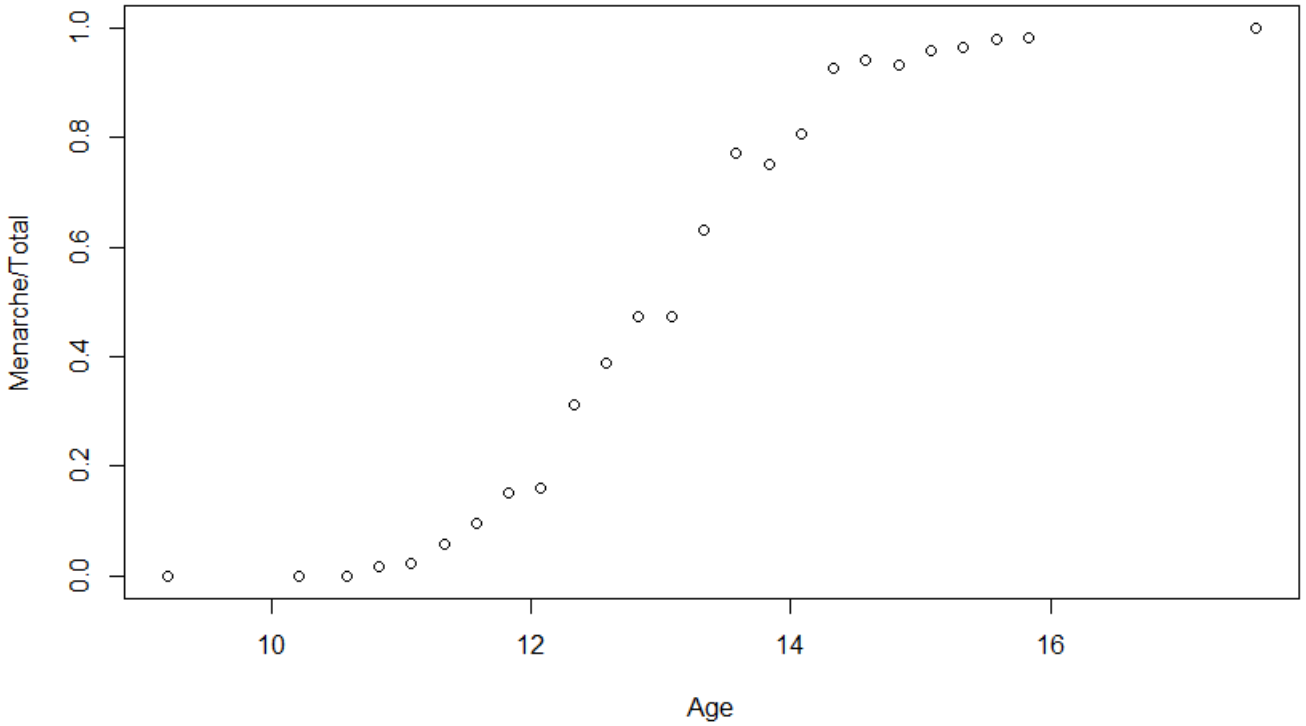
# Residuals

bad.model <- lm(Menarche/Total ~ Age, data=menarche)

plot(bad.model$fitted,bad.model$resid)



plot(Menarche/Total ~ Age, data=menarche)

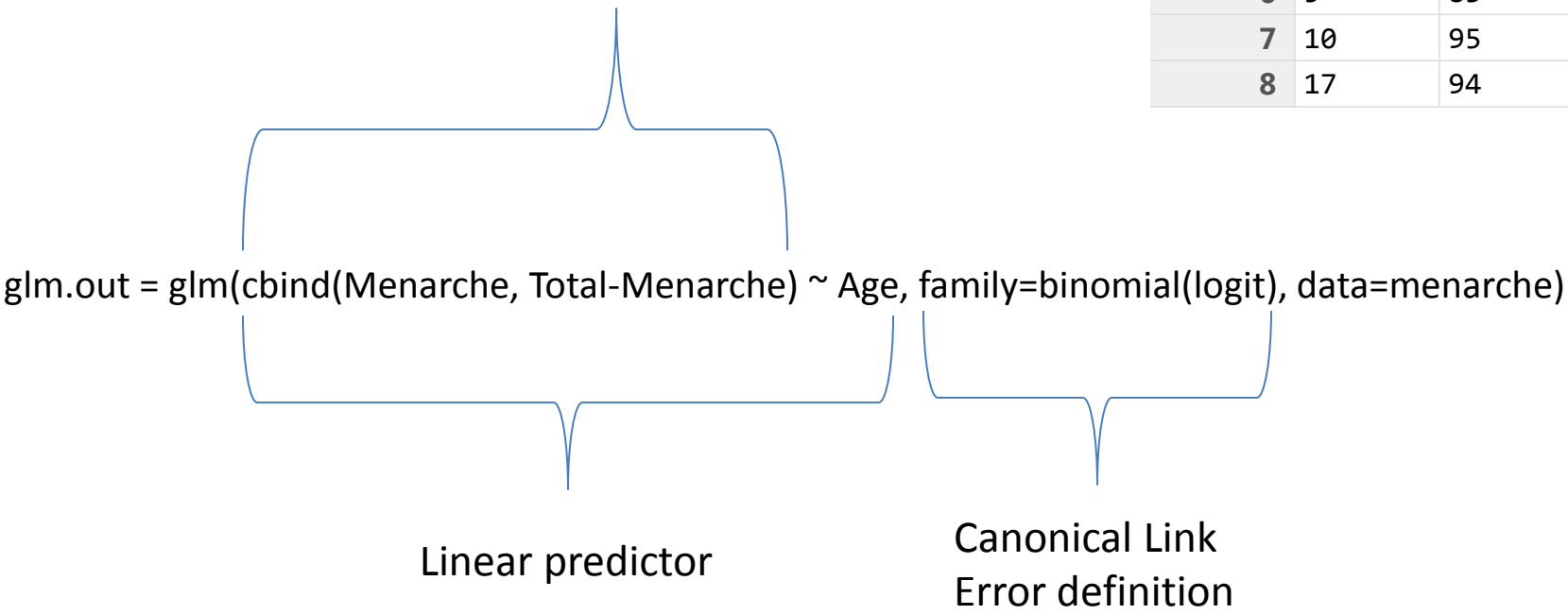



Correct model

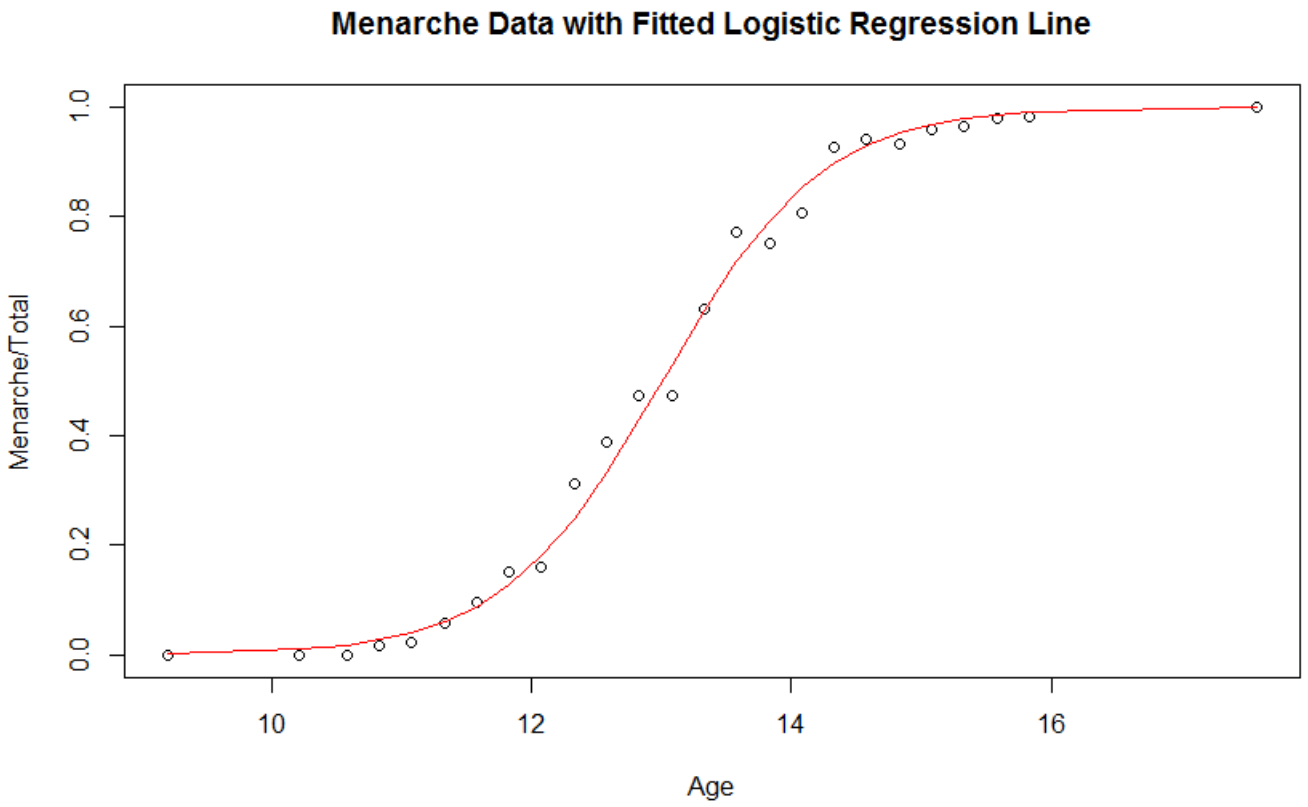
$$\log\left(\frac{p}{q}\right) = \beta_0 + \beta_1 X + \epsilon_i$$

| | p | q |
|---|---|---|
| 1 | 0 | 376 |
| 2 | 0 | 200 |
| 3 | 0 | 93 |
| 4 | 2 | 118 |
| 5 | 2 | 88 |
| 6 | 5 | 83 |
| 7 | 10 | 95 |
| 8 | 17 | 94 |

No binary response variable (0,1)
Send glm a matrix with p,q structure



```
plot(Menarche/Total ~ Age, data=menarche)

lines(menarche$Age, glm.out$fitted, type="l", col="red")

title(main="Menarche Data with Fitted Logistic Regression Line")
```



```
> summary(glm.out)

Call:
glm(formula = cbind(Menarche, Total - Menarche) ~ Age, family = binomial(logit),  data =
menarche)

Deviance Residuals:
   Min      1Q      Median      3Q     Max
-2.0363  -0.9953  -0.4900   0.7780   1.3675

Coefficients:
             Estimate    Std. Error    z value   Pr(>|z|)
(Intercept)  -21.22639    0.77068     -27.54     <2e-16 ***
Age           1.63197     0.05895      27.68     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

 Null deviance:     3693.884  on 24  degrees of freedom
Residual deviance:   26.703  on 23  degrees of freedom
AIC: 114.76

Number of Fisher Scoring iterations: 4
```

For every year the odds of having reached menarche increase by
exp(1.632) = 5.11 times

Deviance: how well the response is predicted.

Null deviance: how well a model with only an intercept (grand mean) predicts the data

Residual deviance: how well the model with just "age" predicts the data

```
> summary(glm.out)

Call:
glm(formula = cbind(Menarche, Total - Menarche) ~ Age, family = binomial(logit),  data =
menarche)

Deviance Residuals:
   Min      1Q      Median      3Q     Max
-2.0363  -0.9953  -0.4900   0.7780   1.3675

Coefficients:
             Estimate     Std. Error    z value   Pr(>|z|)
(Intercept)  -21.22639    0.77068     -27.54     <2e-16 ***
Age           1.63197     0.05895      27.68     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

 Null deviance:     3693.884  on 24  degrees of freedom
Residual deviance:  26.703    on 23  degrees of freedom
AIC: 114.76

Number of Fisher Scoring iterations: 4
```

Essentially a chi square distribution

i.e. 26.7 chi-square value on 23 d.f.

pchisq(26.7,23,lower.tail = FALSE)

[1] 0.2689471

H0: no difference between observed values and model

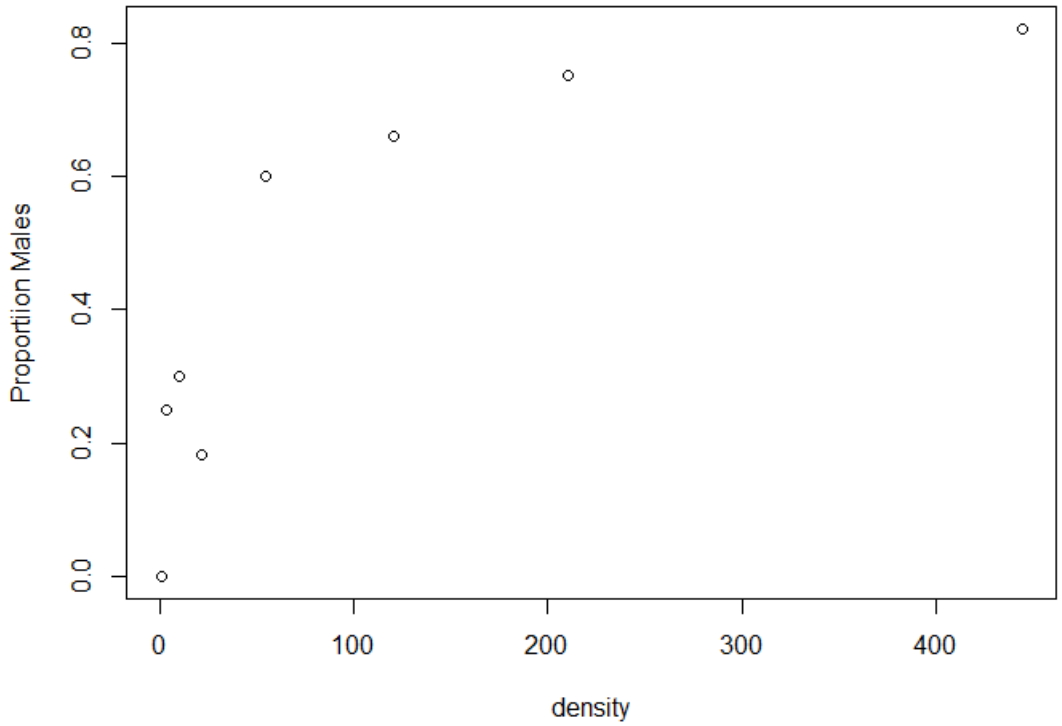
```
> anova(glm.out)

Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Menarche, Total - Menarche)

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev
NULL                    24     3693.9
Age   1   3667.2        23       26.7
```

# Logistic regression

- Example 2
  - Sex ration of insects is variable
  - Does population density influence the proportion of males?

| | density | females | males |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 2 | 4 | 3 | 1 |
| 3 | 10 | 7 | 3 |
| 4 | 22 | 18 | 4 |
| 5 | 55 | 22 | 33 |
| 6 | 121 | 41 | 80 |
| 7 | 210 | 52 | 158 |
| 8 | 444 | 79 | 365 |

```
datpath <- "C:/Users/Che/UNBC_work/Courses/NRES-798/NRES-798-Labs/therbook/"
numbers <- read.table(paste(datpath,"sexratio.txt",sep=""),header=TRUE)
attach(numbers)

p <- males/(males + females)

plot(density,p,ylab="Proportiion Males")
```



```
y <- cbind(males,females)
model <- glm(y~density,binomial)
summary(model)
```

```
Call:
glm(formula = y ~ density, family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.4619  -1.2760  -0.9911   0.5742   1.8795

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0807368  0.1550376   0.521    0.603
density     0.0035101  0.0005116   6.862 6.81e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  71.159  on 7  degrees of freedom
Residual deviance:  22.091  on 6  degrees of freedom
AIC: 54.618

Number of Fisher Scoring iterations: 4
```

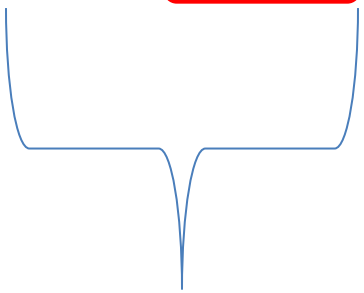
Exp(0.0035) = 1.0035

How to interpret these values?

```
> anova(model)
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

        Df Deviance  Resid. Df  Resid. Dev
NULL                        7      71.159
density  1   49.068         6      22.091
```

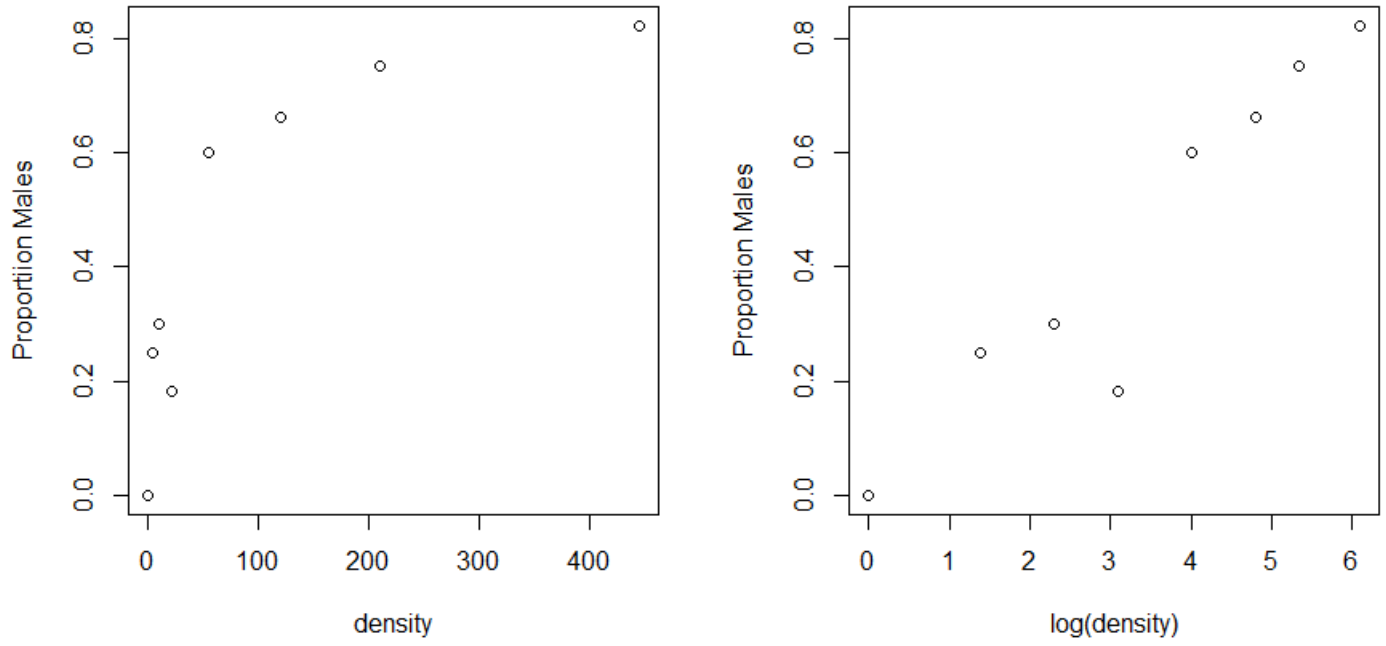
```
> pchisq(71.159,7,lower.tail = FALSE)
[1] 8.612937e-13
```

```
> pchisq(22.091,6,lower.tail = FALSE)
[1] 0.001165743
```



If Residual Df < Residual Deviance = Overdispersion
- There is extra, unexplained variation, in addition to the binomial variance assumed

```
par(mfrow=c(1,2))
plot(density, p, ylab="Proportiion Males")
plot(log(density), p, ylab="Proportion Males")
```




```
y <- cbind(males,females)
model1 <- glm(y~log(density),binomial)
```

```
y <- cbind(males,females)
model1 <- glm(y~log(density),binomial)

> summary(model1)

Call:
glm(formula = y ~ log(density), family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.9697  -0.3411   0.1499   0.4019   1.0372

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.65927    0.48758  -5.454 4.92e-08 ***
log(density)  0.69410    0.09056   7.665 1.80e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:      71.1593  on 7  degrees of freedom
Residual deviance:  5.6739  on 6  degrees of freedom
AIC: 38.201

Number of Fisher Scoring iterations: 4
```

No overdispersion
(previous ResDev 22.091)

```
> pchisq(5.6739,6,lower.tail = FALSE)
[1] 0.4606925
```

# Poisson regression

- Count data
- Variance ~ mean
- Canonical link = Log

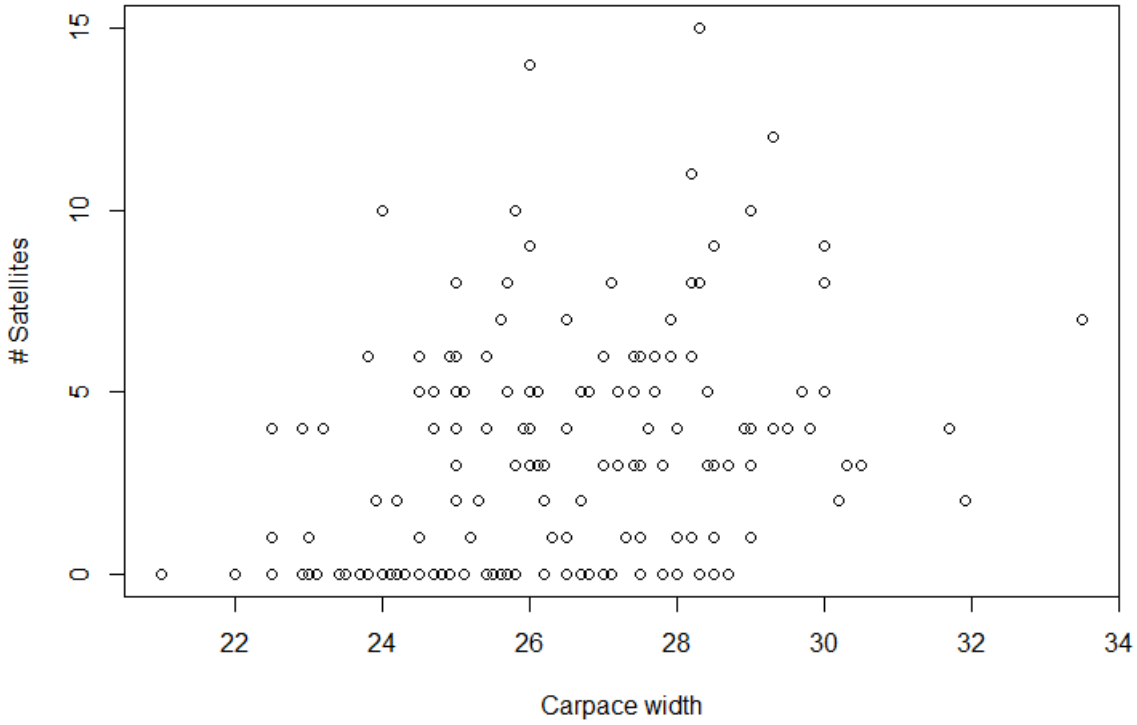
- Example 1
  - Female Horseshoe Crabs have 1 dominant male residing near their nests and potentially a number of “satellite” males.
  - Does the size and condition of a female crab influence how many satellite males reside near her?

# Poisson regression

| | C | S | W | Wt | Sa |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 28.3 | 3.05 | 8 |
| 2 | 3 | 3 | 26.0 | 2.60 | 4 |
| 3 | 3 | 3 | 25.6 | 2.15 | 0 |
| 4 | 4 | 2 | 21.0 | 1.85 | 0 |
| 5 | 2 | 3 | 29.0 | 3.00 | 1 |
| 6 | 1 | 2 | 25.0 | 2.30 | 3 |
| 7 | 4 | 3 | 26.2 | 1.30 | 0 |
| 8 | 2 | 3 | 24.9 | 2.10 | 0 |

Color, spine condition, weight, carpace width, # satellites

```
plot(crab$W,crab$Sa,xlab="Carpace width",ylab="# Satellites")
```




```
model=glm(crab$Sa~1+crab$W,family=poisson(link=log))
```

```
model=glm(crab$Sa~1, family=poisson(link=log))

Call:
glm(formula = crab$Sa ~ 1, family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4162  -2.4162  -0.5707   1.1045   4.9942

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0713     0.0445   24.07   <2e-16 ***
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 632.79  on 172  degrees of freedom
AIC: 990.09

Number of Fisher Scoring iterations: 5
```

```
model=glm(crab$Sa~1+crab$W,family=poisson(link=log))

Call:
glm(formula = crab$Sa ~ 1 + crab$W, family = poisson(link =
log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8526  -1.9884  -0.4933   1.0970   4.9221

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
crab$W       0.16405    0.01997   8.216  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: 927.18

Number of Fisher Scoring iterations: 6
```

What does this coefficient indicate?

> anova(model)
Analysis of Deviance Table

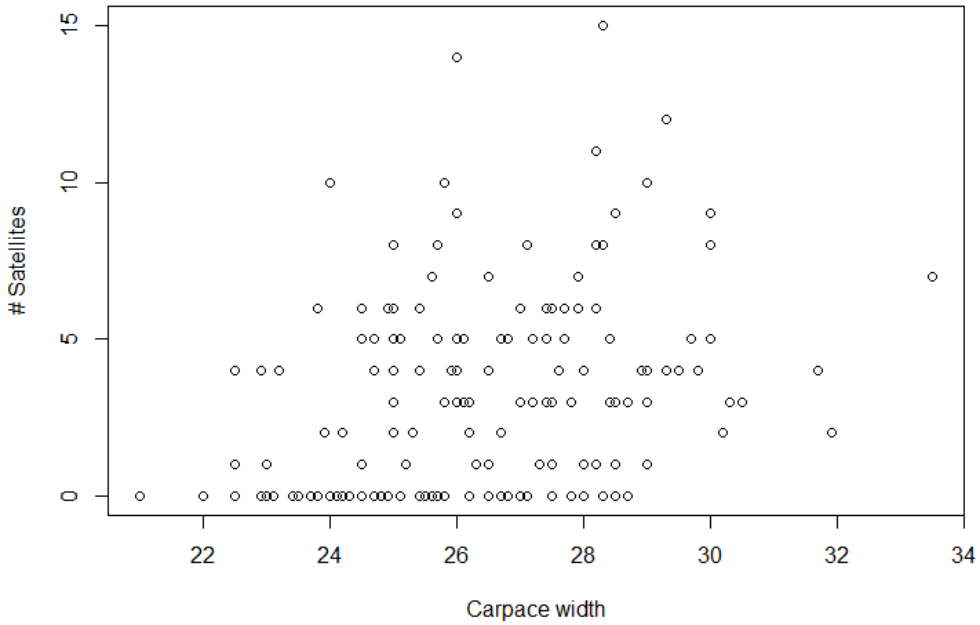
Model: poisson, link: log

Response: crab$Sa

Terms added sequentially (first to last)

| | Df | Deviance | Resid. Df | Resid. Dev |
|---|---|---|---|---|
| NULL | | | 172 | 632.79 |
| crab$W | 1 | 64.913 | 171 | 567.88 |

Good model?
Informative model?



# Poisson regression

- Example 2
- Number of stems broken as a function of tree type (Con, Dec), and stem tension

| | breaks | wool | tension |
|---|---|---|---|
| 1 | 26 | A | L |
| 2 | 30 | A | L |
| 3 | 54 | A | L |
| 4 | 25 | A | L |
| 5 | 70 | A | L |
| 6 | 52 | A | L |

Categorical

'data.frame': 54 obs. of 3 variables:
$ breaks : num 26 30 54 25 70 52 51 26 67 18 ...
$ type: Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
$ tension: Factor w/ 3 levels "L","M","H": 1 1 1 1 1 1 1 1 1 2 ...

# breaksmodel<-glm(breaks~wool*tension, warpbreaks, family=poisson)

Call:
glm(formula = breaks ~ wool * tension, family = poisson, data = warpbreaks)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -3.3383 | -1.4844 | -0.1291 | 1.1725 | 3.5153 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 3.79674 | 0.04994 | 76.030 | < 2e-16 *** |
| woolB | -0.45663 | 0.08019 | -5.694 | 1.24e-08 *** |
| tensionM | -0.61868 | 0.08440 | -7.330 | 2.30e-13 *** |
| tensionH | -0.59580 | 0.08378 | -7.112 | 1.15e-12 *** |
| woolB:tensionM | 0.63818 | 0.12215 | 5.224 | 1.75e-07 *** |
| woolB:tensionH | 0.18836 | 0.12990 | 1.450 | 0.147 |

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

What do these p values represent?

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 297.37 on 53 degrees of freedom
Residual deviance: 182.31 on 48 degrees of freedom
AIC: 468.97

Number of Fisher Scoring iterations: 4

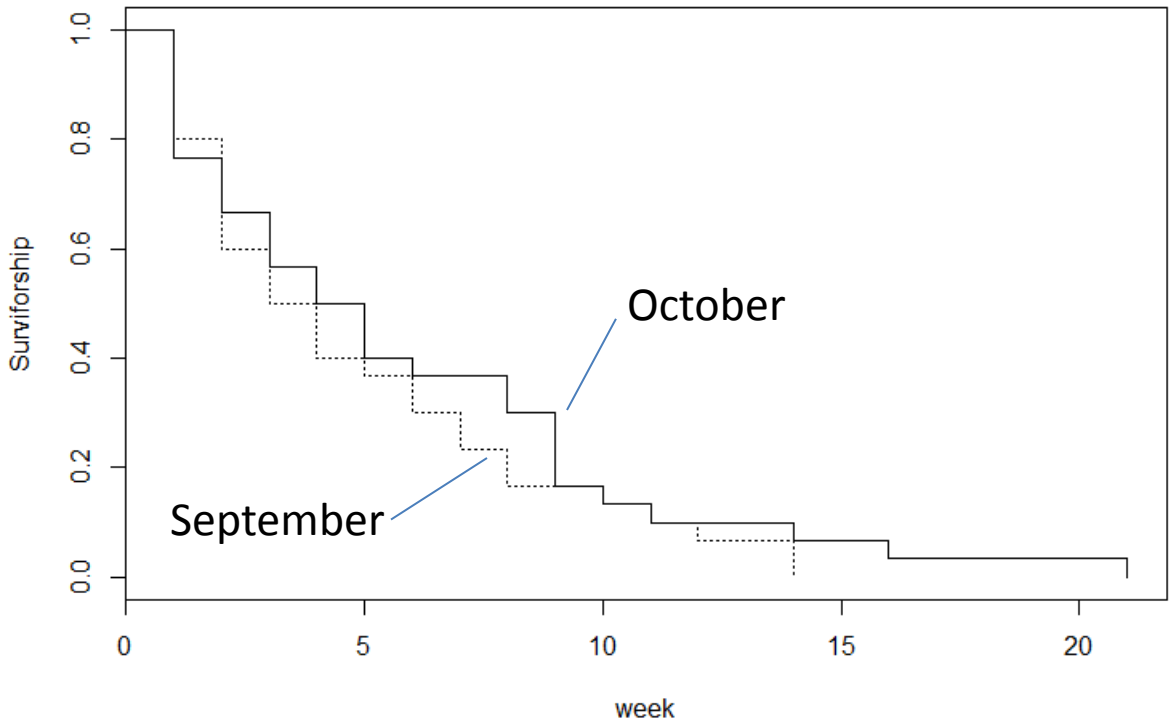
# Survival Analysis

- Often concerned with time to death
- Gamma error distribution

- Example
  - Survival of tree seedlings
  - Does size of canopy gap influence survival

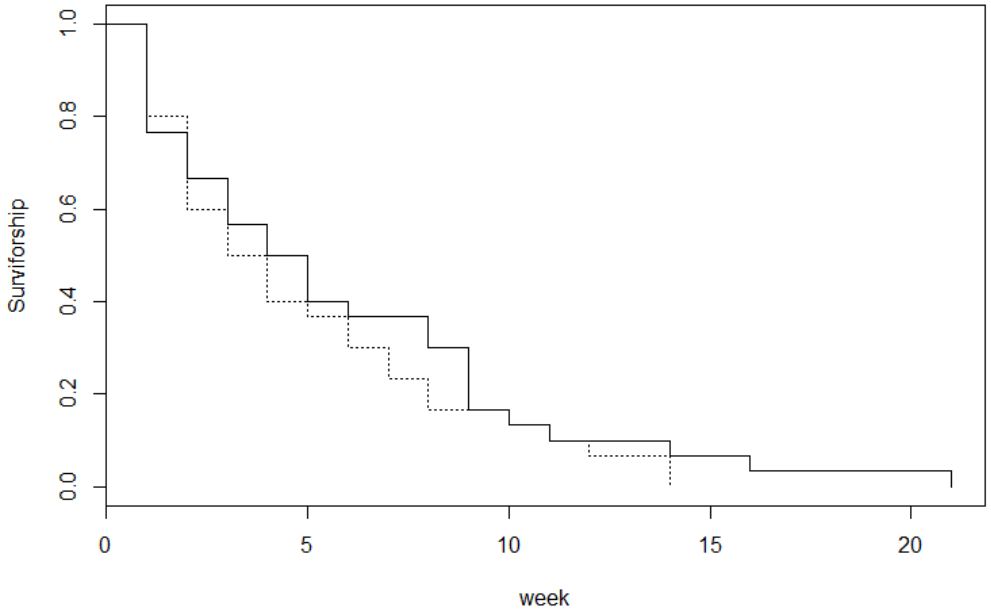
```
> head(seedlings)
    cohort      death    gapsize
1  September     7      0.5889
2  September     3      0.6869
3  September    12      0.9800
4  September     1      0.1921
5  September     4      0.2798
6  September     2      0.2607
```

# Survival analysis




Survival differences between cohorts?

model <- survfit(Surv(death,status)~cohort,data=seedlings)




model <- survfit(Surv(death,status)~cohort,data=seedlings)

Call: survfit(formula = Surv(death, status) ~ cohort, data = seedlings)

| | records | n.max | n.start | events | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|
| cohort=October | 30 | 30 | 30 | 30 | 4.5 | 3 | 9 |
| cohort=September | 30 | 30 | 30 | 30 | 3.5 | 2 | 7 |

Differences between cohorts?

# Survival analysis
# Cox's Proportional Hazard

model1 <- coxph(Surv(death,status)~strata(cohort)*gapsize)

```
Call:
coxph(formula = Surv(death, status) ~ strata(cohort) * gapsize)

  n= 60, number of events= 60

                                              coef     exp(coef) se(coef)     z       Pr(>|z|)
gapsize                                     -1.1863    0.3054   0.6210   -1.910   0.0561 .
strata(cohort)cohort=September:gapsize  0.5795    1.7852   0.8264    0.701   0.4831
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                            exp(coef) exp(-coef) lower .95 upper .95
gapsize                                       0.3054     3.2749   0.09042     1.031
strata(cohort)cohort=September:gapsize    1.7852     0.5602   0.35341     9.018

Concordance= 0.659  (se = 0.077 )
Rsquare= 0.076   (max possible= 0.993 )
Likelihood ratio test= 4.73  on 2 df,   p=0.09372
Wald test            = 4.89  on 2 df,   p=0.08682
Score (logrank) test = 5.04  on 2 df,   p=0.08046
```