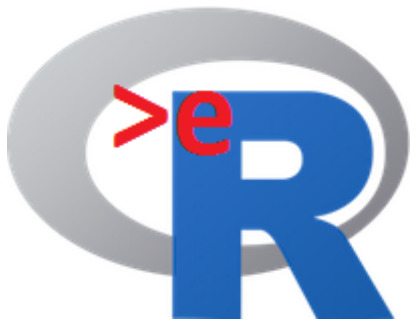This course was developed as a part of the VLIR-UOS Cross-Cutting project s:

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2020.

The >eR-Biostat initiative

Making R based education materials in statistics accessible for all

Applied Generalized Linear Models (GLM) using R (PART 2)

Developed by

Tadesse Awoke (Gondar University), Said Mussa (Mekele University) , Ziv Shkedy (Hasselt University), and Fetene Tekle (J & J)

LAST UPDATE: 07/09/2022

ER-BioStat

https://github.com/eR-Biostat

Email: erbiostat@gmail.com

@erbiostat

Reference list

- Main reference
  - Dobson (2002): An introduction to generalized linear models.
- Other references:
  - McCillagh and Nelder (1983): Generalized linear models (first edition).
  - Collet D(1994): Modeling Binary data.
  - Lindsey (1997): Applying generalized linear models.

Software

- Two main R functions:
  - Linear models in R: the lm() function.
  - Generalized linear models in R: the glm() function in R.

- All R programs for the examples presented in the slides are available online.

Datasets

- Data are given as a part of R programs for the course.

- External datasets (which are not given as a part of the R code) and used for illustration are available online.

Topics (part 2)

11. Poisson Regression

12. Beyond Poisson and binomial distributions: models with different link functions and/or distributions

13. Poisson regression and log linear models

14. Over dispersion

Chapter 11:
Poisson Regression

Donson: chapter 7.

Lindsey: Appendix B.

McCullagh & Nelder: chapter 2.

Count data

Count data:

- counts per unit of time/area/distance, etc
- contingency tables: counts cross-classified by categorical variables
- Covariates: categorical or continuous

Example 1: Stress data

- One randomly chosen member from each randomly chosen household in a sample from Oakaland, California, USA was interviewed. In a list of 41 events, respondents were asked to note which had occurred within the last 18 months. The result is given as:

Month	1	2	3	4	5	6	7	8	9
Respondents	15	11	14	17	5	11	10	4	8
Month	10	11	12	13	14	15	16	17	18
Respondents	10	7	9	11	3	6	1	1	14

Data in R

```
> stress <- read.table("C:...../stress.txt", sep=",",header=TRUE)
> attach(stress)
plot(respondents ~ month, xlab = "Months",ylab = "Subjects",
    xlim=c(0,20), ylim=c(0,20))
```
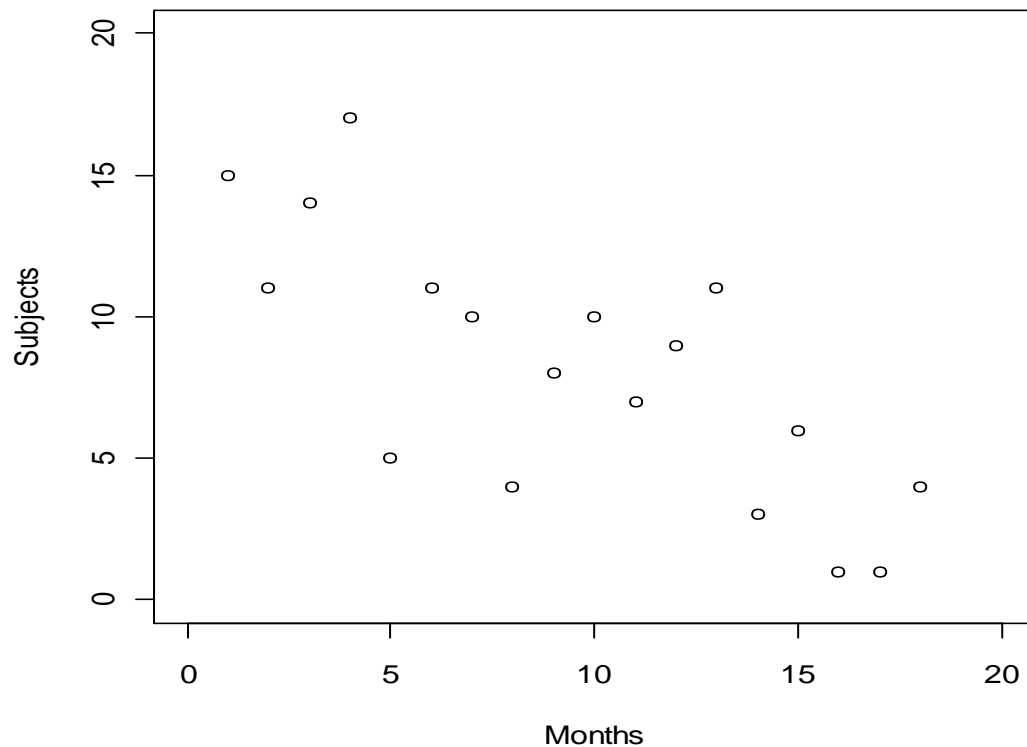
Model formulation

The distribution of the response variable

$$Y_t \sim Poisson(\mu_t)$$

$$f(Y_t, \mu) = \frac{\mu^{y_i} e^{-\mu}}{Y_t!}$$

$$E(Y_t) = \mu_t$$

The dependency on the predictor

$$\mu_t = f(month)$$

A proposal ?

$$\mu_t = \beta_0 + \beta_1 t$$

Model formulation

The distribution of the response variable

$$Y_t \sim Poisson(\mu_t)$$

$$E(Y_t) = \mu_t$$

The linear predictor

$$\mu_t = f(month)$$

$$\eta = \beta_0 + \beta_1 M_t$$

$$\mu_t = e^\eta = e^{\beta_0 + \beta_1 M_t}$$

$$g(E(Y_t)) = \log(\mu_t) = \eta$$

GLM for Poisson regression using glm()

$$\eta = \beta_0 + \beta_1 t$$

```
> respGLM <- glm(respondents ~ month,
        family=poisson, data=stress)
```

The relative risk

$$RR = \frac{E(Y_t \mid t+1)}{E(Y_t \mid t)} = \frac{e^{\beta_0 + \beta_1(t+1)}}{e^{\beta_0 + \beta_1 t}} = e^{\beta_1}$$

GLM for Poisson regression using glm()

> summary(respGLM)

Call:

glm(formula = respondents ~ month, family = poisson, data = stress)

Deviance Residuals:

   Min   1Q   Median   3Q   Max

-1.9886 -0.9631   0.1737   0.5131   2.0362

Coefficients:

        Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.80316   0.14816 18.920 < 2e-16 ***

month   -0.08377   0.01680 -4.986 6.15e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

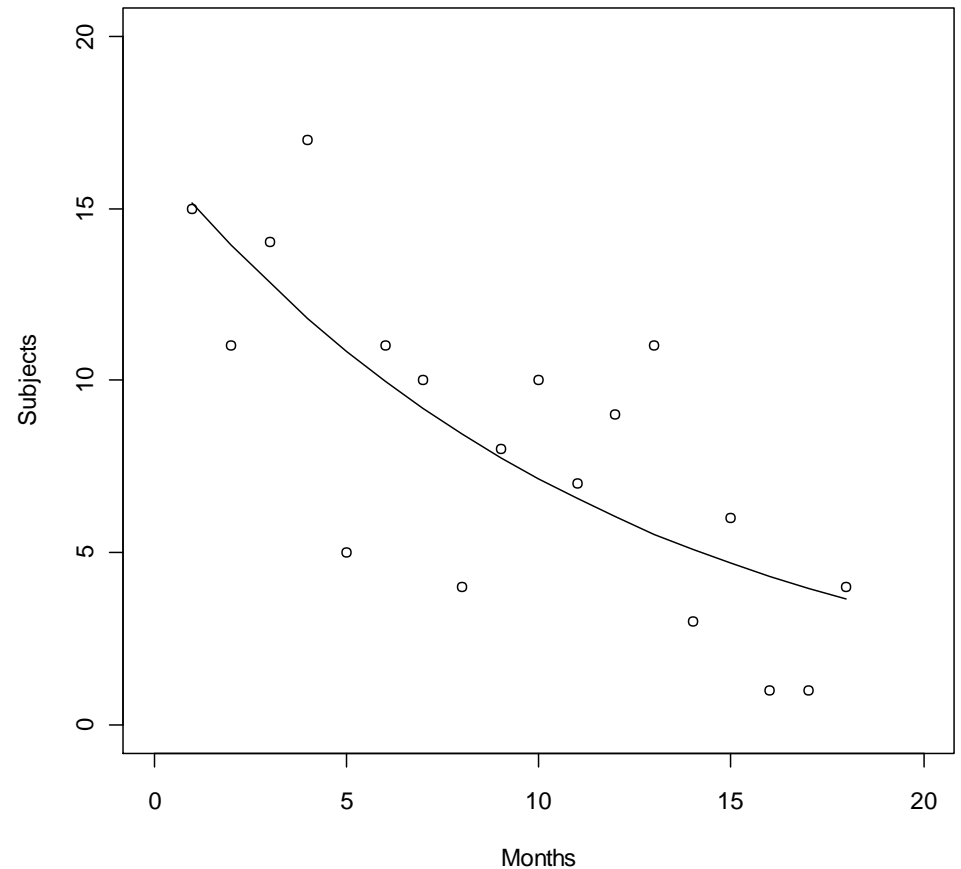   Null deviance: 50.843 on 17 degrees of freedom

Residual deviance: 24.570 on 16 degrees of freedom

AIC: 95.825

Number of Fisher Scoring iterations: 5

Data and fitted model

plot(respondents ~ month, xlab = "Months",
ylab = "Subjects", xlim=c(0,20), ylim=c(0,20))
lines(month,respGLM$fit)

Example 2: Ministerial resignation

- On October 18, 1995, `The Independent' reported on the numbers of ministerial resignations because of different reason. The years start in 1945-1951, with a Labour government, and 7 Resignations.

Term	45-51	51-57	55-57	57-63	63-64	64-70	70-74	74-76	76-79	79-90	90-95	97-05
Gov	Lab	con	con	con	con	lab	con	lab	lab	con	con	lab
Res	7	1	2	7	1	5	6	5	4	1	1	1
Year	6	4	2	6	1	6	4	2	3	1	5	8

- Main question: Is there any difference between Government (Labor and Conservative) in the rate of resignations?

Model formulation

The distribution of the response variable

$$Y_t \sim Poisson(\mu_t)$$

$$E(Y_t) = \mu_t$$

The linear predictor

$$\mu_t = f(gov : L/C)$$

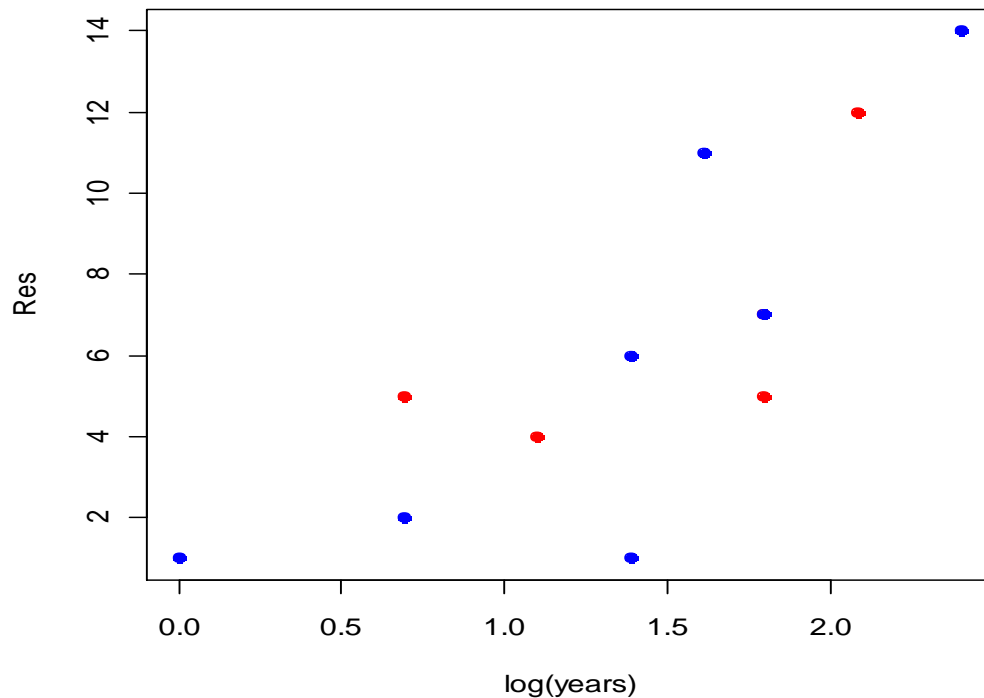$$G_t = \begin{cases} 1 & L \\ 0 & C \end{cases}$$

$$\eta = \beta_0 + \beta_1 G_t$$

$$\mu_t = e^\eta = e^{\beta_0 + \beta_1 G_t}$$

$$g(E(Y_t)) = \log(\mu_t) = \eta$$

The data in R

```
>Resignations <- read.table("C:....../resign.txt",header=T)
>attach(Resignations)
>plot(Res ~ log(years), pch=19, col=c(4,2)[Gov])
```

Model formulation (1)

> first.glm <- glm(Res ~ Gov , poisson)

$$\eta = \log(\mu_t) = \beta_0 + \beta_1 G_t$$

The relative risk

$$RR = \frac{E(Y_t \mid L)}{E(Y_t \mid C)} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Model 1 output in R

> first.glm <- glm(Res ~ Gov, poisson);

> summary(first.glm)

Call:

glm(formula = Res ~ Gov, family = poisson)

Deviance Residuals:

   Min   1Q Median   3Q   Max

-2.5331 -1.2942 -0.3255 0.7548 2.7793

Coefficients:

      Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.79176   0.15430   11.61 <2e-16 ***

Govlab   0.09531   0.23262   0.41   0.682

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

   Null deviance: 33.436 on 11 degrees of freedom

Residual deviance: 33.269 on 10 degrees of freedom

AIC: 78.459

Number of Fisher Scoring iterations: 5

Model formulation (2)

How can we take the number of government years into account ?

$$\mu_t = e^{\eta} = e^{\beta_0 + \beta_1 G_t + \beta_2 \log(years_t)}$$

$$\eta = \log(\mu_t) = \beta_0 + \beta_1 G_t + \beta_2 \log(years_t)$$

GLM with Poisson family

> first.glm <- glm(Res ~ Gov + log(years), poisson)

The same slope for log(year)

$$\eta = \log(\mu_t) = \beta_0 + \beta_1 G_t + \beta_2 \log(years_t)$$

Model 2: R output

glm(formula = Res ~ Gov + log(years), family = poisson)

Deviance Residuals:

Min   1Q   Median   3Q   Max

-2.2417 -0.3469 -0.1250   0.3917   1.6513

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.29973   0.41598   0.721   0.471

Govlab   0.03541   0.23271   0.152   0.879

log(years)   0.96636   0.22258   4.342 1.41e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 33.436 on 11 degrees of freedom

Residual deviance: 11.276 on 9 degrees of freedom

AIC: 58.466

Number of Fisher Scoring iterations: 4

Model 3: model formulation

Different intercepts and slopes

```
>  first.glm3 <- glm(Res ~ log(years)+Gov+Gov:log(years),
            poisson)
```

$$\eta = \beta_0 + \beta_1 G_t + \beta_2 \log(years_t) + \beta_3 G_t \log(year_t)$$

Model 3 output in R

\> summary(first.glm)

Call:

glm(formula = Res ~ log(years) +Gov+ Gov:log(years), family= poisson)

Estimate Std. Error z value Pr(>|z|)

(Intercept)   0.04443   0.51085   0.087   0.931

log(years)   1.11144   0.27306   4.070   4.69e-05 ***

Govlab   0.81973   0.82744   0.991   0.322

log(years):Govlab -0.46049   0.46880   -0.982   0.326

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

   Null deviance: 33.436 on 11 degrees of freedom

Residual deviance: 10.336 on 8 degrees of freedom

AIC: 59.526

Number of Fisher Scoring iterations: 4

Model 4: GLM with an offset variable: model formulation

Number of resignation per government year:

$$\frac{Y_t}{year_t} \sim Poisson(\mu_t)$$

$$Y_t \sim Poisson(years_t \times \mu_t)$$

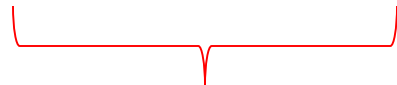# Model 4: GLM with an offset variable - model formulation

$$Y_t \sim Poisson(years_t \times \mu_t)$$

$$E(Y_t) = years_t \times \mu_t = years_t \times e^{\beta_0 + \beta_1 G_t}$$

$$g(E(Y_t)) = g(years_t \times \mu_t) = \log(years_t \times e^{\beta_0 + \beta_1 G_t})$$

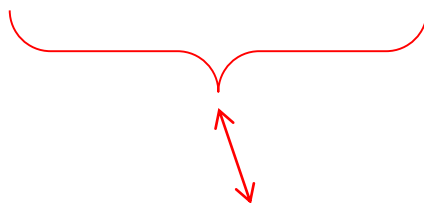$$g(E(Y_t)) = \underbrace{\log(years_t)}_{} + \beta_0 + \beta_1 G_t = \eta$$

offset variable

A model with offset in R

> next.glm<- glm(Res ~ Gov + offset(log(years)), poisson)

$$\beta_0 + \beta_1 G_t + \log(years_t)$$

Model 4 GLM with offset output in R

summary(first.glm4)

Call:

glm(formula = Res ~ Gov + offset(log(years)), family = poisson)

$\hat{\beta}_0$ Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.24116   0.15430   1.563   0.118

$\hat{\beta}_1$ Govlab   0.03647   0.23262 0.157   0.875

(Dispersion parameter for poisson family taken to be 1)

   Null deviance: 11.323 on 11 degrees of freedom

Residual deviance: 11.299 on 10 degrees of freedom

AIC: 56.488

Number of Fisher Scoring iterations: 4

Data and predicted models

```
 plot(Res ~ log(years), pch=19, col=c(4,2)[Gov]) # Use palette() to find out which colour corresponds
> legend(locator(1), legend= c("conservative", "labour"), col=c(4,2), pch=19)
> l <- (0:25)/10
> fv <- exp(0.3168 + 0.9654*l)# to plot fitted curve under last.glm
> lines(l,fv)
```

Ministerial Resignations against log(years)
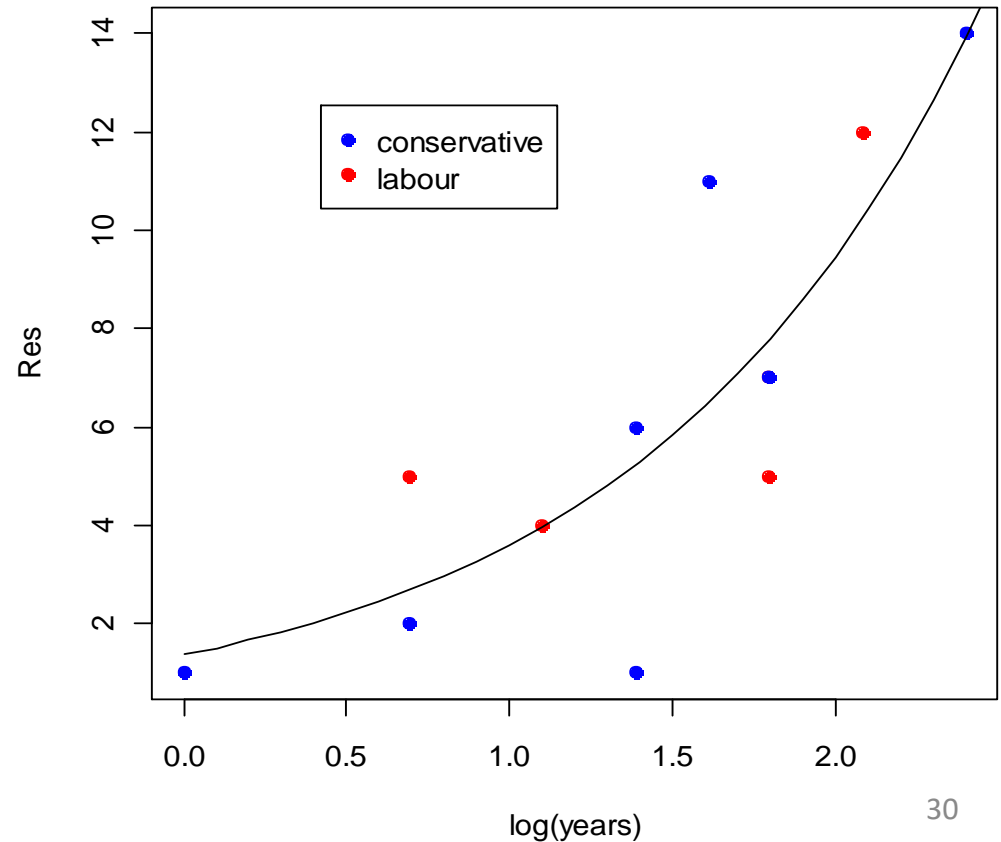
AIC for the different models

MODEL	No parameters	Deviance	AIC
1	2	33.269	78.45861
2	3	11.276	58.46574
3	4	10.336	59.52603
4	2	11.299	56.48846

$$g(E(Y_t)) = \log(years_t) + \beta_0 + \beta_1 G_t$$

Confidence interval for β_2 for model 2

first.glm2 <- glm(Res ~ log(years)+Gov, poisson); summary(first.glm)

> confint(first.glm2, level=0.95)
Waiting for profiling to be done...
              2.5 %   97.5 %
(Intercept) -0.5701863 1.0636716
log(years)   0.5447242 1.4185361
Govlab   -0.4268576 0.4893354

ANOVA for model 3 and model 2

> anova(first.glm3,first.glm2, test = "Chisq")
Analysis of Deviance Table

Model 1: Res ~ log(years) + Gov + Gov:log(years)
Model 2: Res ~ log(years) + Gov
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1   8   10.336
2   9   11.276 -1 -0.93971   0.3324

Model selection (II): the step() function in R

```
> step(first_glm, direction = "backward")
Start:  AIC=59.53
Res ~ log(years) + Gov + Gov:log(years)
                  Df Deviance    AIC
- log(years):Gov  1   11.276 58.466
<none>                10.336 59.526

Step:  AIC=58.47
Res ~ log(years) + Gov
             Df Deviance    AIC
- Gov         1   11.299 56.489
<none>            11.276 58.466
-    log(years)  1   33.269 78.459
```

```
Step:  AIC=56.49
Res ~ log(years)
             Df Deviance    AIC
<none>           11.299 56.489
-log(years)  1   33.436 76.626
-Call:  glm(formula = Res ~ log(years), family = poisson)
Coefficients:
(Intercept)   log(years)
    0.3168      0.9654
Degrees of Freedom: 11 Total (i.e. Null);  10 Residual
Null Deviance:      33.44
Residual Deviance: 11.3
AIC: 56.49
```

Model 4 with the offset variable is not included here.

Example 3: smoking and coronary death

```
> CHD
   age smoke death   pop
1   40    s    32 52407
2   50    s   104 43248
3   60    s   206 28612
4   70    s   186 12663
5   80    s   102  5317
6   40   ns     2 18790
7   50   ns    12 10673
8   60   ns    28  5710
9   70   ns    28  2585
10  80   ns    31  1462
```

A study about tobacco consumption and coronary heart disease among British doctors.

Smoking and coronary death: the data
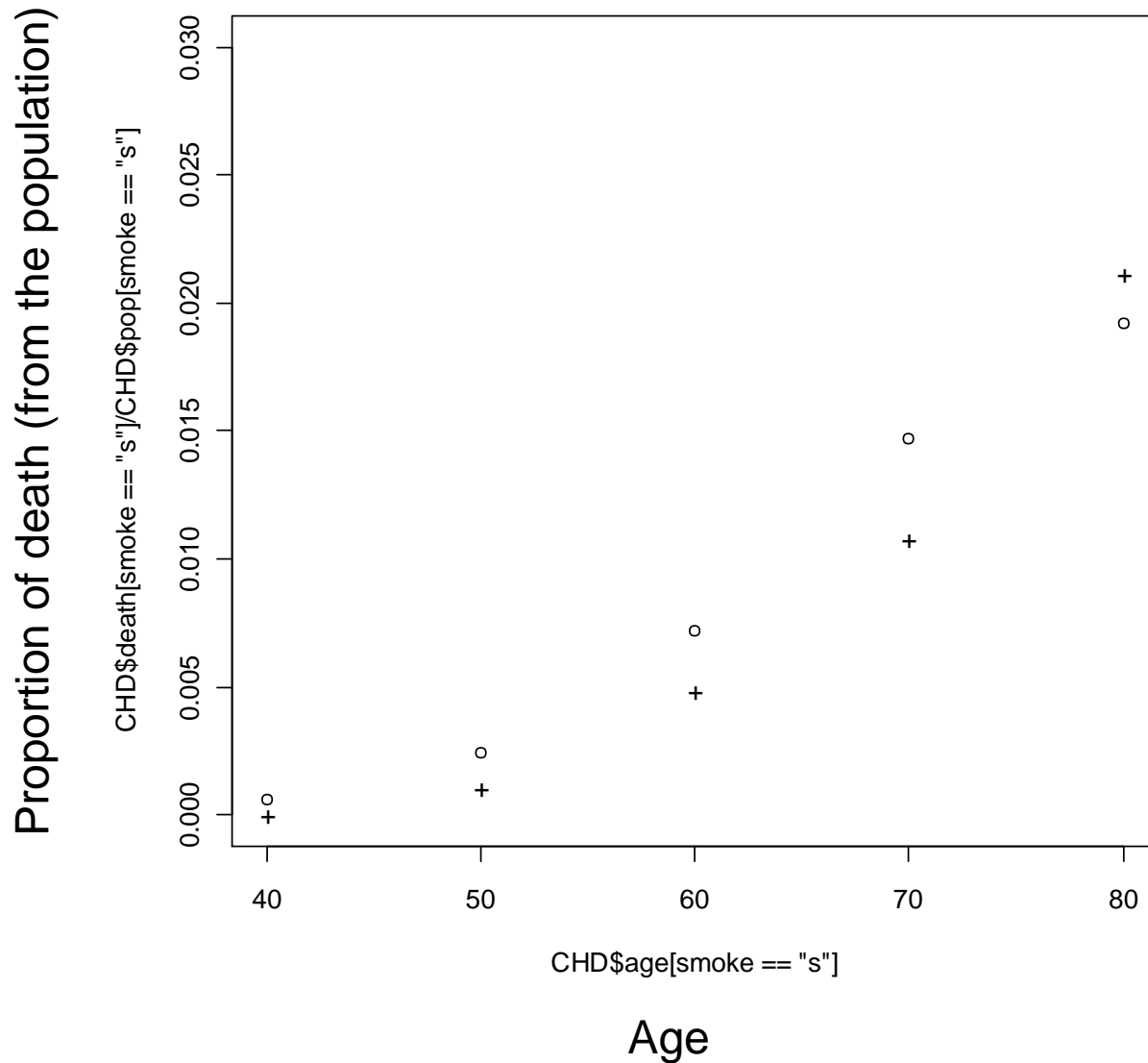
Proportion of death (from the population)

CHD$death[smoke == "s"]/CHD$pop[smoke == "s"]

CHD$age[smoke == "s"]

Age

36

GLM with an offset variable: model formulation

Number of deaths per population size:

$$\frac{Y_i}{n_i} \sim Poisson(\mu_i)$$

$$Y_i \sim Poisson(n_i \times \mu_i)$$

$$g(\mu_i) = X\beta$$

Smoking and coronary death

- Is the death rate higher for smokers than non smokers ?

- If so, by how much ?

- Is there differential effects of age ?

GLM with an offset variable: model formulation

Number of deaths per population size:

$$\frac{Y_i}{n_i} \sim Poisson(\mu_i)$$

$$Y_i \sim Poisson(n_i \times \mu_i)$$

$$g(\mu_i) = X\beta = \log(n_i) + \log(\mu_i)$$

Offset variable

$$\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...$$

Linear predictor:

$$\eta = f(age, smoking, pop.)$$

3 models

$$M_1 : \eta = \log(pop.) + \beta_1 age + \beta_2 smoke$$

$$M_2 : \eta = \log(pop.) + \beta_1 age + \beta_2 smoke + \beta_3 age \times smoke$$

$$M_3 : \eta = \log(pop.) + \beta_{11} age + \beta_{12} age^2 + \beta_2 smoke + \beta_3 age \times smoke$$

```
>fit.chd1<-glm(death ~ age + smoke+offset(log(pop)), poisson)
>fit.chd2<-glm(death ~ age + smoke+age:smoke+offset(log(pop)), poisson)
>age2<-age^2
>fit.chd3<-glm(death ~ age+age2+smoke+age:smoke+offset(log(pop)), poisson)
```

Model selection

```
> extractAIC(fit.chd1, k=2)
[1]   3.0000 130.2500
> extractAIC(fit.chd2, k=2)
[1]   4.0000 122.9614
> extractAIC(fit.chd3, k=2)
[1]   5.00000 66.70331
```

The model with quadratic age effect has the best goodness-of-fit.

R output (model 3)

> summary(fit.chd3)

Call:
glm(formula = death ~ age + age2 + smoke + age:smoke + offset(log(pop)),
  family = poisson)

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.970e+01 1.253e+00 -15.717 < 2e-16 ***
age   3.563e-01 3.632e-02   9.810 < 2e-16 ***
age2   -1.977e-03 2.737e-04 -7.223 5.08e-13 ***
smokes   2.364e+00 6.562e-01   3.602 0.000316 ***
age:smokes -3.075e-02 9.704e-03 -3.169 0.001528 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.0673 on 9 degrees of freedom
Residual deviance: 1.6354 on 5 degrees of freedom
AIC: 66.703

Number of Fisher Scoring iterations: 4

Data and predicted model

$$\hat{E}(Y_i) = n_i \times \exp(\mu_i) = n_i \times \exp(X\hat{\beta})$$
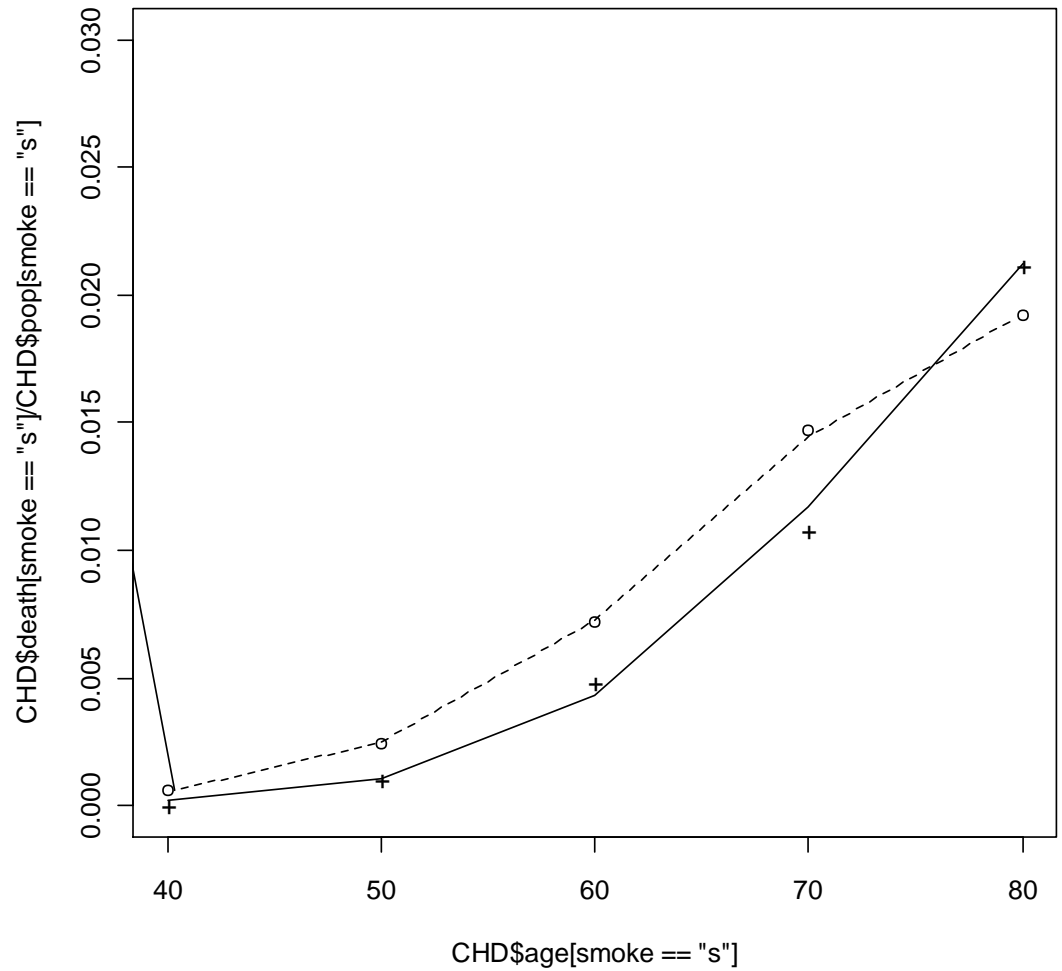
Chapter 12:
Beyond Poisson and binomial distributions: Models with different link functions and/or distributions

Lindsey: Chapter 4.

Example 1: Employment duration

- The employment duration of staff, age 25 to 44, recruited to the British post office in the first quarter of 1973 and classified in to two grades.

Month	Grade 1	Grade2	Month	Grade1	Grade2
1	22	30	13	0	1
2	18	28	14	0	0
3	19	31	15	0	0
4	13	14	16	1	1
5	5	10	17	1	1
6	6	6	18	1	0
7	3	5	19	3	2
8	2	2	20	1	0
9	2	3	21	1	3
10	1	0	22	0	1
11	0	0	23	0	1
12	1	1	24	0	0

The data in R

```
> employ <- read.table("C:.... /employ.txt",header=TRUE)
> attach(employ)
> employ
  month grade duration
```

	month	grade	duration
1	1	1	22
2	1	2	30
3	2	1	18
4	2	2	28
5	3	1	19
6	3	2	31
7	4	1	13
8	4	2	14
9	5	1	5
10	5	2	10

. . . .

Original sacle.
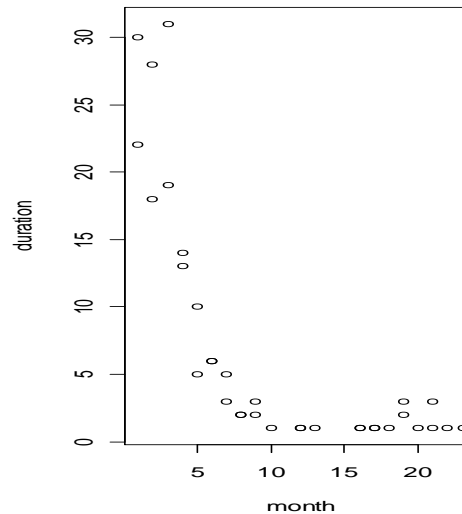
log transformation.
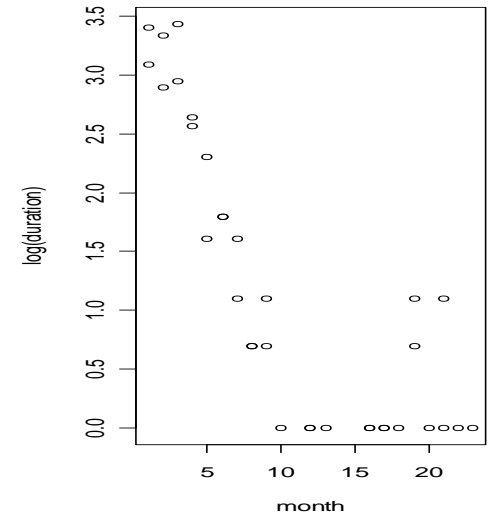
Model formulation

Models with normal error.

$$Y_t \sim N(\mu_t, \sigma^2)$$

$$\eta = \beta_0 + \beta_1 t + \beta_2 G$$

$$\mu = \beta_0 + \beta_1 t + \beta_2 G$$

$$\log(\mu) = \beta_0 + \beta_1 t + \beta_2 G$$

$$\frac{1}{\mu} = \beta_0 + \beta_1 t + \beta_2 G$$

$$\mu = e^{\beta_0 + \beta_1 t + \beta_2 G}$$

$$\mu = \frac{1}{\beta_0 + \beta_1 t + \beta_2 G}$$

identity link

Log link

invesre link

Models with normal error in R

m.normal.idt <- glm(duration ~ month + grade + month:grade,data = employ, family = gaussian(link = identity))

m.normal.inv <- glm(duration ~ month + grade + month:grade,data = employ, family = gaussian(link = inverse))

m.normal.log <- glm(duration ~ month + grade + month:grade,data = employ, family = gaussian(link = log))

m.normal.log1 <- glm(duration ~ month + grade ,data = employ,
family = gaussian(link = log))

Model selection

> extractAIC(m.normal.idt, k=2)
[1]   4.0000 228.5094
> extractAIC(m.normal.inv, k=2)
[1]   4.0000 198.3163
> extractAIC(m.normal.log, k=2)
[1]   4.0000 173.8542
> extractAIC(m.normal.log1, k=2)
[1]   3.0000 171.8545

The model with log link has the smallest AIC value.

GLM with normal error and log link in R

> m.normal.log <- glm(duration ~ month + grade + month:grade,data = employ1, family = gaussian(link = log))
> summary(m.normal.log)

Call:
glm(formula = duration ~ month + grade + month:grade, family = gaussian(link = log),
    data = employ1)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-4.6676 -1.5332 -0.2005   0.8915 10.8676

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4476971 0.1304802 26.423 < 2e-16 ***
month   -0.2709521 0.0444746 -6.092 1.08e-06 ***
grade2   0.3698490 0.1591805   2.323   0.0271 *
month:grade2 -0.0007872 0.0545143 -0.014   0.9886

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 8.219982)

    Null deviance: 2771.88 on 33 degrees of freedom
Residual deviance: 246.59 on 30 degrees of freedom
AIC: 173.85

Number of Fisher Scoring iterations: 7

Data and fitted model
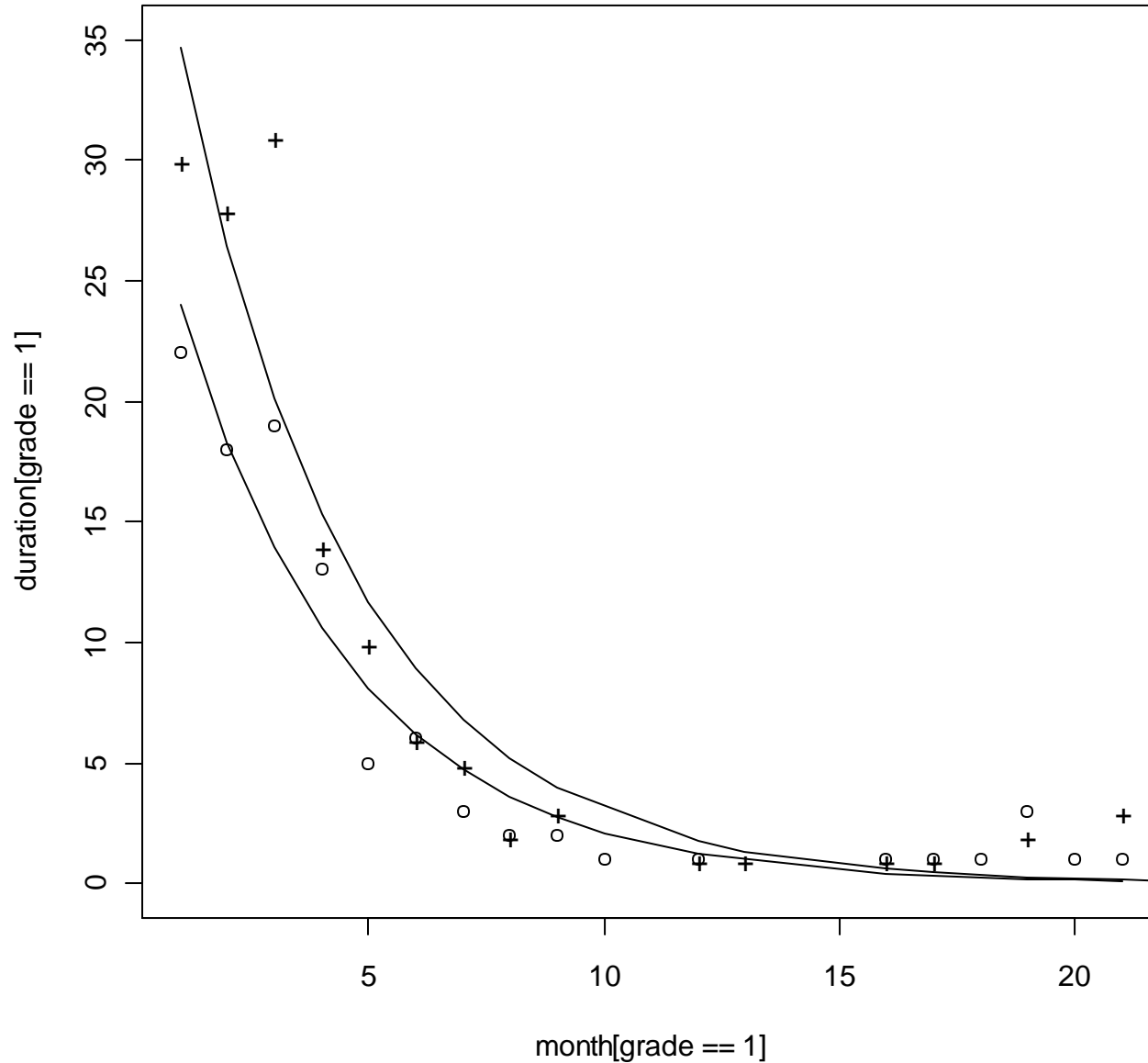
Other models

- Different distributions:
  - Normal.
  - Gamma.
  - …

$$Y_t \sim H(\mu_t)$$

$$g(\mu_t) = \eta = \beta_0 + \beta_1 t + \beta_2 G$$

$$\mu = \beta_0 + \beta_1 t + \beta_2 G$$

$$\log(\mu) = \beta_0 + \beta_1 t + \beta_2 G$$

$$\frac{1}{\mu} = \beta_0 + \beta_1 t + \beta_2 G$$

$$\mu = e^{\beta_0 + \beta_1 t + \beta_2 G}$$

$$\mu = \frac{1}{\beta_0 + \beta_1 t + \beta_2 G}$$

identity link

Log link

invesre link

Model formulation (model 1)

m.normal.log <- glm(duration ~ month + grade + month:grade, data = employ1, family = gaussian(link = log))

$$\eta = \log(\mu_i) = \beta_0 + \beta_2 t_i + \beta_1 G_i + \beta_3 t_i \times G_i$$

A model with normal error and log link:

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$g(\mu) = \eta, \quad \log(\mu) = \eta$$

R output : model 1

> m.normal.log <- glm(duration ~ month + grade + month:grade, data = employ, family = gaussian(link = log))
> summary(m.normal.log)

Call:
glm(formula = duration ~ month + grade + month:grade, family = gaussian(link = log),
    data = employ)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-4.6676 -1.5332 -0.2005   0.8915 10.8676

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.0778481 0.2764302 11.134 3.54e-12 ***
month   -0.2701650 0.0943705 -2.863 0.00759 **
grade   0.3698490 0.1591805   2.323 0.02712 *
month:grade -0.0007872 0.0545143 -0.014 0.98857

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 8.219982)

    Null deviance: 2771.88 on 33 degrees of freedom
Residual deviance: 246.59 on 30 degrees of freedom
AIC: 173.85

55

Model with gamma error and inverse link function: model formulation (model 2)

```
m.gamma.log <- glm(duration ~ month + grade + month:grade,
          data = employ1,
          family = Gamma(link = inverse))
```

$$\eta = 1/\mu_i = \beta_0 + \beta_2 t_i + \beta_1 G_i + \beta_3 t \times G_i$$

A model with gamma error and inverse:

$$Y_i \sim Gamma(\mu_i)$$

$$\frac{1}{\mu_i} = \eta_i$$

R output: model 2

> m.gamma.inv <- glm(duration ~ month + grade + month:grade, data = employ, family = Gamma(link = inverse))
> summary(m.gamma.inv)

Call:
glm(formula = duration ~ month + grade + month:grade, family = Gamma(link = inverse),
    data = employ)

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0122937 0.0396403 -0.310 0.75861
month   0.0492390 0.0154143   3.194 0.00329 **
grade   -0.0008403 0.0226236 -0.037 0.97062
month:grade -0.0089011 0.0090337 -0.985 0.33235

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.3661260)

    Null deviance: 51.752 on 33 degrees of freedom
Residual deviance: 9.887 on 30 degrees of freedom
AIC: 147.27

Model with gamma error and inverse link function: model formulation (model 3)

Grade is not included in the model:

$$\eta = 1/\mu_i = \beta_0 + \beta_2 t_i$$

A model with gamma error and inverse:

$$Y_i \sim Gamma(\mu_i)$$

$$\frac{1}{\mu_i} = \eta_i$$

R output: model 3

> m.gamma.inv1 <- glm(duration ~ month , data = employ, family = Gamma(link = inverse))
> summary(m.gamma.inv1)

Call:
glm(formula = duration ~ month, family = Gamma(link = inverse),
    data = employ)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
 -0.9163 -0.5284 -0.2795   0.2599   1.2685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.014033   0.010821 -1.297   0.204
month   0.035353   0.004445   7.954 4.45e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.3762360)

    Null deviance: 51.752 on 33 degrees of freedom
Residual deviance: 10.562 on 32 degrees of freedom
AIC: 145.63

Number of Fisher Scoring iterations: 6

Model selection

> AIC(m.normal.log)

[1] 173.8542

> AIC(m.gamma.inv)

[1] 147.2748

> AIC(m.gamma.inv1)

[1] 145.6314

> AIC(m.gamma.inv2)

[1] 145.3437

The model which give small AIC is the gamma model with inverse line and only duration in the model

Data and fitted models
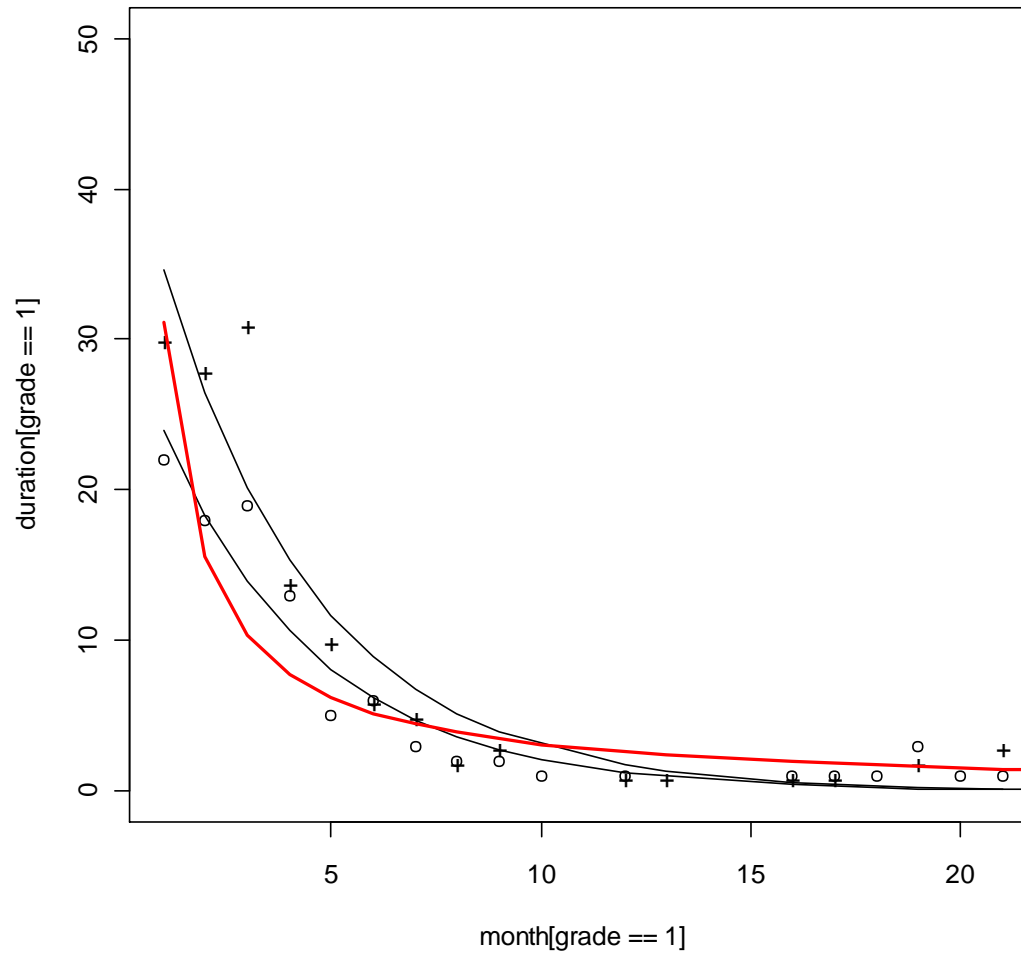
Chapter 13:
Poisson regression and log linear models

Based on Dobson: Chapter 9

Log linear models

The general frame work of Poisson regression is given by:

$$Y_i \sim Poisson(\mu_i)$$

$$g(\mu_i) = \log(\mu_i) = X\beta$$

Example 1: melanoma

Tumor Type	Site			
	Head & neck	Trunk	Extere mities	total
Hutchinson	22	2	10	34
Superficial melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

Cross sectional study of patients with form of skin cancer.

Different sites & different tumor types.

Example 1

Tumor Type	Site			
	Head & neck	Trunk	Exteremities	total
Hutchinson	22	2	10	34
Superficial melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

Y_{ij} Frequency of subjects with tumor type i and site j.

$$Y_{ij} \sim Poisson(\mu_{ij})$$

n Sample size.

Example 1: independence ?

	Site			
Tumor Type	Head & neck	Trunk	Extere mities	total
Hutchinson	22	2	10	34
Superficial melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

$$Y_{i.}$$

$$Y_{.j}$$

Chi-squared statistic for independence:

$$X^2 = \frac{\sum\limits_{ij}\left(Y_{ij} - e_{ij}\right)^2}{e_{ij}}$$

$$e_{ij} = \frac{Y_{i.} \times Y_{.j}}{n}$$

$$e_{11} = \frac{34 \times 68}{400}$$

Example 1: independence ?

$$X^2 = \frac{\sum_{ij}\left(Y_{ij} - e_{ij}\right)^2}{e_{ij}} = \frac{(22-5.78)^2}{5.78} + \ldots + \frac{(28-31.64)^2}{31.64} = 65.8$$

Model formulation

$$Y_{ij} \sim Poisson(\mu_{ij})$$

$$g(\mu_{ij}) = \mu \qquad g(\mu_{ij}) = \mu + \alpha_i + \beta_j \qquad g(\mu_{ij}) = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

The minimal
model

M1

Independence
model

M2

Second ordered
interaction model

M3

Data in R

> melanoma

	Tumor	Type	y
1	Hu	HN	22
2	Hu	Trunk	2
3	Hu	Extrem	10
4	Su	HN	16
5	Su	Trunk	45
6	Su	Extrem	115
7	Nod	HN	19
8	Nod	Trunk	33
9	Nod	Extrem	73
10	Ind	HN	11
11	Ind	Trunk	17
12	Ind	Extrem	28

Models in R

> M1<-glm(y~1, family=poisson, data=melanoma)

$$g(\mu_{ij}) = \mu$$

> M2<-glm(y~Tumor+Type, family=poisson, data=melanoma)

$$g(\mu_{ij}) = \mu + \alpha_i + \beta_j$$

> M3<-glm(y~Tumor+Type+Tumor:Type,family=poisson,
    data=melanoma)

$$g(\mu_{ij}) = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

> AIC(M1)
[1] 348.8361
> AIC(M2)
[1] 121.5482
> AIC(M3)
[1] 82.9297

Likelihood ratio test

> anova.glm(M2,M3,test="Chisq")
Analysis of Deviance Table

Model 1: y ~ Tumor + Type
Model 2: y ~ Tumor + Type + Tumor:Type
 Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1   6   50.618
2   0   0.000 6   50.618 3.533e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

Example 2: Death penalty

Death penalty verdicts for cases involving multiple murders in Florida between 1976 and 1987. This data is from an article that studied effects of racial characteristics on whether persons convicted of homicide received the death penalty. The 674 subjects classified is in to a 2x2x2 contingency table- two rows, two columns, and two layers.

		Victim's Race			
		White		Black	
		Defendant's Race		Defendant's Race	
		White	Black	White	Black
Death Penalty	Yes	53	11	0	4
	No	414	37	16	139
Percent Yes		11.3	22.9	0.0	2.8

Example 3: Death penalty

		Victim's Race			
		White		Black	
		Defendant's Race		Defendant's Race	
		White	Black	White	Black
Death Penalty	Yes	53	11	0	4
	No	414	37	16	139
Percent Yes		11.3	22.9	0.0	2.8

The response variable: death penalty verdicts

$$Y_{ijk} \sim Poisson(\mu_{ijk})$$

Example 3: Death penalty

		Victim's Race			
		White		Black	
		Defendant's Race		Defendant's Race	
		White	Black	White	Black
Death Penalty	Yes	53	11	0	4
	No	414	37	16	139

The mean structure

$$\log(\mu_{ijk}) = \mu + v_i + d_j + p_k + \text{interaction}$$

$$\log(\mu_{ijk}) = \mu + v_i + d_j + p_k + vd_{ij} + vp_{ik} + dp_{jk} + vdp_{ijk}$$

Exploring the Data in R

> deathpenalty <- data.frame(number = c(53,11,0,4,414,37,16,139),

+ victim = c("W","W","B","B","W","W","B","B"),

+ defendant = c("W","B","W","B","W","B","W","B"),

+ death = rep(c("yes","no"),rep(4,2)))

plot(xtabs(number~ victim + defendant+ death, deathpenalty))

xtabs(number ~ victim + defendant + death, deathpenalty)
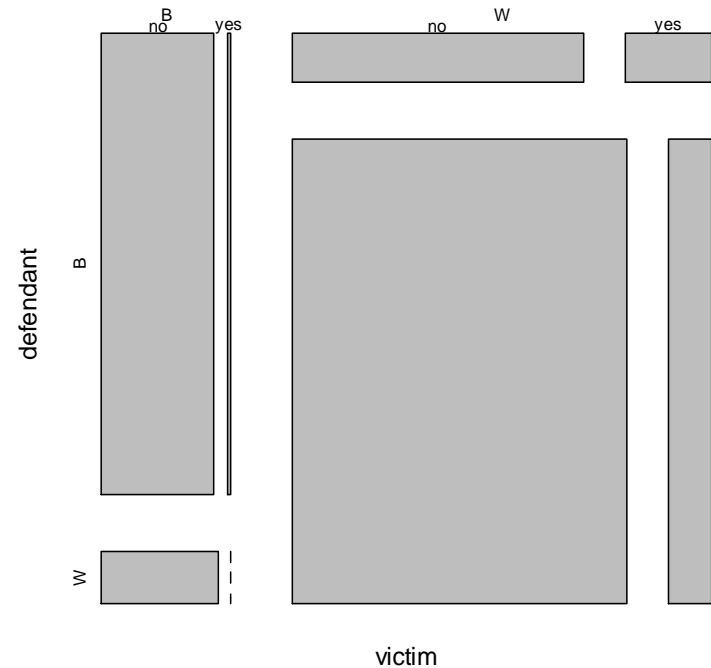
The model in R

$$\log(\mu_{ijk}) = \mu + v_i + d_j + p_k + vd_{ij} + vp_{ik} + dp_{jk} + vdp_{ijk}$$

M1<-glm(number~victim*defendant*death,
    family=poisson,
    data=deathpenalty)

Saturated model with Poisson family

> M1<-glm(number~victim*defendant*death, family=poisson, data=deathpenalty)

> summary(M1)
> Call:

Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.934e+00 8.482e-02 58.177 < 2e-16 ***
victimW   -1.324e+00 1.850e-01 -7.155 8.38e-13 ***
defendantW   -2.162e+00 2.640e-01 -8.189 2.63e-16 ***
deathyes   -3.548e+00 5.071e-01 -6.996 2.63e-12 ***
victimW:defendantW   4.577e+00 3.149e-01 14.536 < 2e-16 ***
victimW:deathyes   2.335e+00 6.125e-01   3.813 0.000137 ***
defendantW:deathyes   -2.153e+01 4.225e+04 -0.001 0.999593
victimW:defendantW:deathyes 2.068e+01 4.225e+04 0.00049 0.999609

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)

   Null deviance: 1.2251e+03 on 7 degrees of freedom
Residual deviance: 4.1224e-10 on 0 degrees of freedom
AIC: 54.04
Number of Fisher Scoring iterations: 20

Using the step function in R to identify the best model

step(the three way interaction model)

- A stepwise selection based on AIC.
- Starting point: the three way interaction model.

Step () function for model selection

> step(M1)

Start: AIC=54.04 ──────────────────────────────→

number ~ victim * defendant * death

               Df Deviance   AIC

- victim:defendant:death 1 0.37984 52.42

\<none>   0.00000 54.04

Step: AIC=52.42

number ~ victim + defendant + death + victim:defendant + victim:death + defendant:death

            Df Deviance   AIC

\<none>   0.38 52.42

- defendant:death   1   5.39 55.43

- victim:death   1   20.73 70.77

- victim:defendant 1   384.43 434.47

Three was interaction is excluded.

Two way interaction are excluded (one at the time).

The best model includes all two way interactions:

$$\log(\mu_{ijk}) = \mu + v_i + d_j + p_k + vd_{ij} + vp_{ik} + dp_{jk}$$

Output of the Step() function

Call: glm(formula = number ~ victim + defendant + death + victim:defendant +
            victim:death + defendant:death,
            family = poisson, data = deathpenalty)

Coefficients:
    (Intercept)   victimW   defendantW
      4.9358   -1.3298   -2.1746
      deathyes   victimW:defendantW   victimW:deathyes
      -3.5961   4.5950   2.4044
defendantW:deathyes
      -0.8678

Degrees of Freedom: 7 Total (i.e. Null); 1 Residual

Null Deviance: 1225

Residual Deviance: 0.3798   AIC: 52.42

The two way interaction model in R

$$\log(\mu_{ijk}) = \mu + v_i + d_j + p_k + vd_{ij} + vp_{ik} + dp_{jk}$$

M2<-glm(number~victim+defendant+death
  +victim:defendant
  +victim:death
  +defendant:death,
  family=poisson, data=deathpenalty)

Output in R

> summary(M2)

Call:
glm(formula = number ~ victim + defendant + death + victim:defendant +
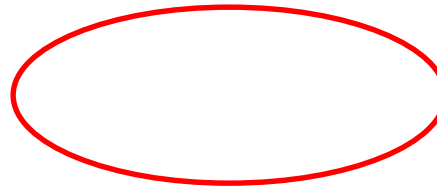   victim:death + defendant:death, family = poisson, data = deathpenalty)

Coefficients:
```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.93578    0.08471  58.265  < 2e-16 ***
victimW          -1.32980    0.18479  -7.196 6.19e-13 ***
defendantW       -2.17465    0.26377  -8.245  < 2e-16 ***
deathyes         -3.59610    0.50691  -7.094 1.30e-12 ***
victimW:defendantW  4.59497    0.31353  14.656  < 2e-16 ***
victimW:deathyes    2.40444    0.60061   4.003 6.25e-05 ***
defendantW:deathyes -0.86780    0.36707  -2.364   0.0181 *
---
```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1225.07955 on 7 degrees of freedom
Residual deviance: 0.37984 on 1 degrees of freedom
AIC: 52.42

Example 3: Antibiotic prescription

Diagnosis of Respiratory Tract Infections Hueston and Stott (2000) report a study of clinicians' diagnoses of respiratory tract infections over a 14-month period. The aim was to determine whether a reduction in prescription of antibiotics to acute bronchitis patients was due to clinicians assigning an alternative diagnosis.

	Time period				
Diagnosis	1-3/96	4-6/96	7-9/96	10-12/96	1-2/97
Acute bronchitis	113	58	40	108	100
Acute sinusitis	99	37	23	50	32
URI	410	228	125	366	304
Pneumonia	60	43	30	56	45
Total	682	366	218	580	481

The data in R

```
> diag <- rep(c("bron", "sinus", "URI", "pneu"), 5)
> time <- rep(c("win96", "spr96", "sum96", "aut96", "spr97"), rep(4, 5))
> rt <- data.frame(diag = factor(diag, unique(diag)),
+ time = factor(time, unique(time)), count = c(113, 99, 410, 60, 58, 37, 228, 43, 40, 23, 125, 30,
+ 108, 50, 366, 56, 100, 32, 304, 45))
```

Example 3: Antibiotic prescription

Diagnosis	Time period				
	1-3/96	4-6/96	7-9/96	10-12/96	1-2/97
Acute bronchitis	113	58	40	108	100
Acute sinusitis	99	37	23	50	32
URI	410	228	125	366	304
Pneumonia	60	43	30	56	45

Data in R

```
> rt
   diag  time count
1   bron win96  113
2  sinus win96   99
3    URI win96  410
4   pneu win96   60
5   bron spr96   58
6  sinus spr96   37
7    URI spr96  228
8   pneu spr96   43
9   bron sum96   40
10 sinus sum96   23
11   URI sum96  125
12  pneu sum96   30
13  bron aut96  108
14 sinus aut96   50
15   URI aut96  366
16  pneu aut96   56
17  bron spr97  100
18 sinus spr97   32
19   URI spr97  304
20  pneu spr97   45
```

Research question:
Diagnostic and time period are independent ?

Models formulation

$$Y_{ij} \sim Poisson(\mu_{ij})$$

$$g(\mu_{ij}) = \mu$$

Minimal model

$$g(\mu_{ij}) = \mu + D_i + T_j$$

Independence model

$$g(\mu_{ij}) = \mu + D_i + T_j + DT_{ij}$$

Second ordered interaction

Models in R

```
>
> M1 <- glm(count ~ 1, family=poisson, data=rt)
> M2 <- glm(count ~ diag+time, family=poisson, data=rt)
> M3 <- glm(count ~  diag+time+diag:time,family=poisson, data=rt)
>
>
> AIC(M1)
[1] 1915.304
> AIC(M2)
[1] 169.8704
> AIC(M3)
[1] 164.2791
```

Likelihood ratio test

> anova.glm(M2,M3,test="Chisq")
Analysis of Deviance Table

Model 1: count ~ diag + time
Model 2: count ~ diag + time + diag:time
 Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1   12   29.591
2   0   0.000 12   29.591 0.003216 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

Independence model is rejected.

Output model 3

> summary(M3)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.72739	0.09407	50.253	< 2e-16	***
diagsinus	-0.13227	0.13766	-0.961	0.33664	
diagURI	1.28877	0.10625	12.130	< 2e-16	***
diagpneu	-0.63304	0.15974	-3.963	7.40e-05	***
timespr96	-0.66694	0.16153	-4.129	3.64e-05	***
timesum96	-1.03851	0.18398	-5.645	1.66e-08	***
timeaut96	-0.04526	0.13457	-0.336	0.73664	
timespr97	-0.12222	0.13729	-0.890	0.37336	
diagsinus:timespr96	-0.31726	0.25143	-1.262	0.20702	
diagURI:timespr96	0.08013	0.18143	0.442	0.65872	
diagpneu:timespr96	0.33380	0.25693	1.299	0.19388	
diagsinus:timesum96	-0.42112	0.29568	-1.424	0.15438	
diagURI:timesum96	-0.14934	0.21045	-0.710	0.47795	
diagpneu:timesum96	0.34536	0.28957	1.193	0.23300	
diagsinus:timeaut96	-0.63784	0.21957	-2.905	0.00367	**
diagURI:timeaut96	-0.06827	0.15258	-0.447	0.65457	
diagpneu:timeaut96	-0.02374	0.22942	-0.103	0.91760	
diagsinus:timespr97	-1.00717	0.24536	-4.105	4.05e-05	***
diagURI:timespr97	-0.17691	0.15677	-1.128	0.25913	
diagpneu:timespr97	-0.16546	0.24029	-0.689	0.49107	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.7890e+03 on 19 degrees of freedom
Residual deviance: 1.5765e-14 on 0 degrees of freedom
AIC: 164.28

Number of Fisher Scoring iterations: 3

Chapter 14
Over dispersion

Lindsey: Chapter 3

Over dispersion parameter

- The general form of exponential family is defined as:

$$f(y) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

- Where $\emptyset$ is the dispersion parameter.
- a($\emptyset$): scale parameter.

Example: normal distribution

$$f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{\frac{-(y_i - \mu_i)^2}{2\sigma^2}}$$

$$= \exp\left\{ \left[y_i\mu_i - \frac{\mu_i^2}{2} \right] \frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right\}$$

The mean and the variance are separated parameters.

$$\theta_i = \mu_i,$$

$$b(\theta_i) = \theta_i^2 / 2$$

$$a_i(\phi) = \sigma^2$$

$$c(y_i, \, \phi) = -\left[y_i^2 / \phi + \log(2\pi\phi) \right] / 2.$$

Example: Binomial distribution

$$Z_i = \begin{cases} 1 \\ 0 \end{cases} \qquad Y_i = \sum_{i=1}^{n} Z_i \qquad Y_i \sim B(n, \pi_i)$$

$$p(y_i \mid \theta) = \binom{n_i}{y_i} \theta^{y_i} (1-\theta)^{n-y} =$$

$$\exp\left\{ y_i \log\left[\frac{\theta_i}{1-\theta_i}\right] + n_i \log(1-\theta_i) + \log\binom{n_i}{y_i} \right\}$$

The variance is a function of the mean.

$$a_i(\phi) = 1, \quad b(\theta_i) = \log(1 + \exp(\theta_i))$$

$$c(y) = \log\binom{n_i}{y_i}$$

$$E(y) = \mu = b'(\theta_i) = e^{\theta}(1 + \exp(\theta_i))^{-1}$$

$$\mathrm{var}(y) = n\mu(1-\mu)$$

Poisson distribution

$$Y_i \sim Poisson(\mu) \qquad\qquad f(y_i, \theta_i) \frac{\theta_i^{y_i} e^{-\theta}}{y_i!}$$

$$\boxed{a_i(\phi) = 1}$$

$$b(\theta) = \exp(\theta)$$

$$c(y) = -\log(y!)$$

E(y)=V(y).
$$E(y) = \mu = b'(\theta) = \exp(\theta)$$

$$\mathrm{var}(y) = \mu$$

Overdispersion

The binomial and Poisson distribution are a members of one parameter exponential family.

$$Y_t \sim Poisson(\mu_t) \qquad Y_i \sim B(n, \pi)$$

$$E(Y_t) = V(Y_t) = \mu_t \qquad E(Y_i) = n\pi$$

$$V(Y_i) = n\pi(1 - \pi)$$

$$\boxed{a_i(\phi) = 1}$$

Overdispersion

Often, we observed extra variability than expected:

$$V(Y_i) > E(Y_i)$$

$$V(Y_i) > n\pi(1-\pi)$$

$$V(Y_i) = \phi E(Y_i)$$

$$V(Y_i) = \phi n\pi(1-\pi)$$

$$\phi = 1 \implies \text{No problem with overdispersion}$$

Estimating over dispersion

Formula

$$\phi = \frac{\chi_p^2}{n - p}$$

where

$$\text{p} = \text{number of parametr in the model}$$
$$\text{n} = \text{number of observations}$$

Example 1: Germination of seeds from Orobanche

| o. aegyptiaco 75 | | | | o. aegyptiaco 73 | | | |
| bean | | cucumber | | bean | | cucumber | |
germ.	total	germ.	total	germ.	total	germ.	total
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	<u>81</u>	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	<u>4</u>	3	7
		10	13				

Data in R

> fac<-read.table("C:....../seed.txt", header = TRUE)

> fact <- transform(fac, prop = germ/total, extr.seed = interaction(extract,seed))

```
fact
   seed  extract germ total     prop      extr.seed
1  aeg75    bean   10    39 0.2564103    bean.aeg75
2  aeg75    bean   23    62 0.3709677    bean.aeg75
3  aeg75    bean   23    81 0.2839506    bean.aeg75
4  aeg75    bean   26    51 0.5098039    bean.aeg75
5  aeg75    bean   17    39 0.4358974    bean.aeg75
6  aeg75 cucumber   5     6 0.8333333 cucumber.aeg75
7  aeg75 cucumber  53    74 0.7162162 cucumber.aeg75
8  aeg75 cucumber  55    72 0.7638889 cucumber.aeg75
9  aeg75 cucumber  32    51 0.6274510 cucumber.aeg75
10 aeg75 cucumber  46    79 0.5822785 cucumber.aeg75
11 aeg75 cucumber  10    13 0.7692308 cucumber.aeg75
12 aeg73    bean    8    16 0.5000000    bean.aeg73
13 aeg73    bean   10    30 0.3333333    bean.aeg73
14 aeg73    bean    8    28 0.2857143    bean.aeg73
15 aeg73    bean   23    45 0.5111111    bean.aeg73
16 aeg73    bean    0     4 0.0000000    bean.aeg73
17 aeg73 cucumber   3    12 0.2500000 cucumber.aeg73
18 aeg73 cucumber  22    41 0.5365854 cucumber.aeg73
19 aeg73 cucumber  15    30 0.5000000 cucumber.aeg73
20 aeg73 cucumber  32    51 0.6274510 cucumber.aeg73
21 aeg73 cucumber   3     7 0.4285714 cucumber.aeg73
```

Exploring the data: Box-Plot

> plot(prop ~ extr.seed, data = fact, las = 1, ylab = "proportion")

> points(prop ~ extr.seed, data = fact, pch = 16)

> title("proportion of germinating seeds")

proportion of germinating seeds
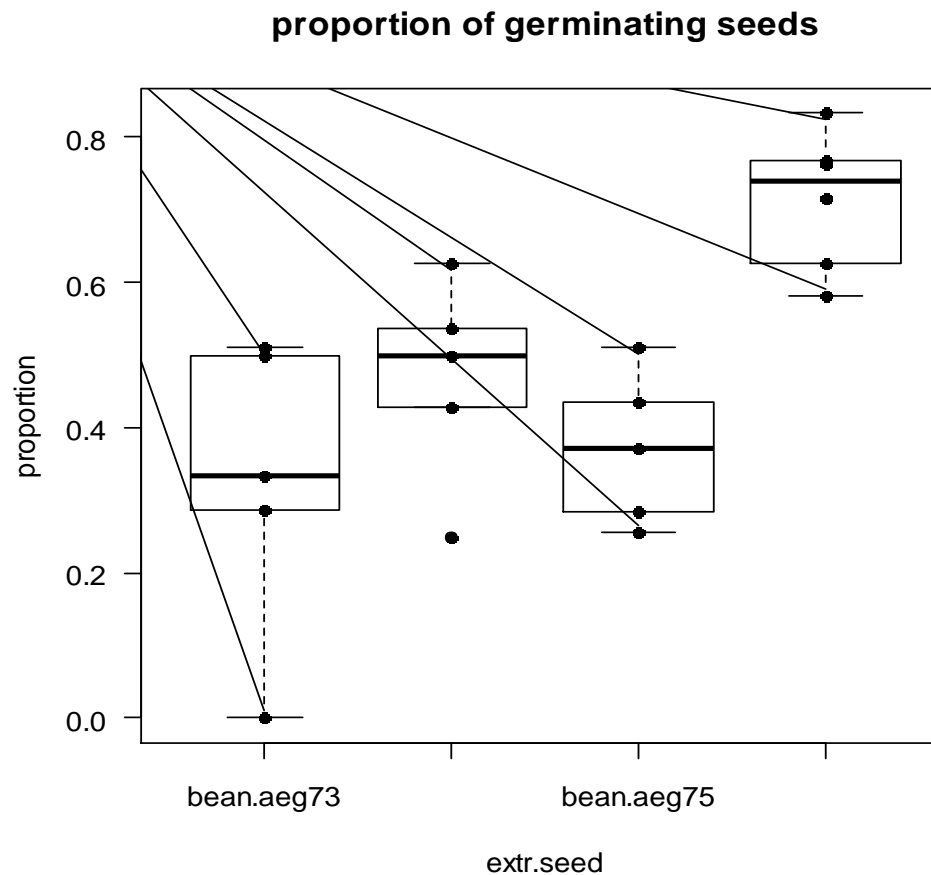
Model 1 formulation

- Binomial model

$$yi \sim Bin(n_i, \pi_i)$$

$$g(\pi_i) = \eta, \ \log it(\pi_i) = \eta$$

$$\log it(\pi_i) = \beta_0 + \beta_1 ext + \beta_2 seed + \beta_3 ext \times seed$$

Model 1: binomial family

```
> g <- glm(cbind(germ, total - germ) ~ extract + seed + extract:seed, family = binomial, data = fact)
> r.pears<-residuals(g, type="pearson")
> summary(g)
```

The assumption here is over dispersion parameter is $\Phi=1$

Model 1 output

Call:

glm(formula = cbind(germ, total - germ) ~ extract + seed + extract:seed,

   family = binomial, data = fact)

Deviance Residuals:

   Min   1Q   Median   3Q   Max

-2.01617 -1.24398   0.05995   0.84695   2.12123

Coefficients:

               Estimate Std. Error z value Pr(>|z|)

(Intercept)   -0.4122   0.1842 -2.238 0.0252 *

extractcucumber   0.5401   0.2498 2.162 0.0306 *

seedaeg75   -0.1459   0.2232 -0.654 0.5132

extractcucumber:seedaeg75 0.7781   0.3064 2.539 0.0111 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 98.719 on 20 degrees of freedom

Residual deviance: 33.278 on 17 degrees of freedom

AIC: 117.87

Number of Fisher Scoring iterations: 4

Estimating over dispersion in R

```
> X2 <- sum(residuals(g, type = "pearson")^2)
> X2
[1] 31.65114
> phi <- X2/g$df.residual
> phi
[1] 1.861832
> phi <- g$deviance/g$df.residual
> phi
[1] 1.957517
```

As we can see from the R output, the over dispersion parameter is greater than 1 and hence this is an indication of the presence of over dispersion

Confidence interval for model 1

> g <- glm(cbind(germ, total - germ) ~ extract + seed + extract:seed, family = binomial, data = fact)

> confint(g)

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	-0.77866159	-0.05469853
extractcucumber	0.05275108	1.03329024
seedaeg75	-0.58184899	0.29428969
extractcucumber:seedaeg75	0.17619697	1.37823747

Taking into account overdispersion in R

glm(model, family = quasibinomial,...)

Output quasi-binomial model 2

> g.over <- glm(cbind(germ, total - germ) ~ extract + seed + extract:seed, family = quasibinomial, data = fact)
> summary(g.over)

Call:
glm(formula = cbind(germ, total - germ) ~ extract + seed + extract:seed,
    family = quasibinomial, data = fact)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-2.01617 -1.24398 0.05995 0.84695 2.12123

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.4122   0.2513 -1.640 0.1193
extractcucumber   0.5401   0.3409   1.584 0.1315
seedaeg75   -0.1459   0.3045 -0.479 0.6379
extractcucumber:seedaeg75 0.7781   0.4181   1.861 0.0801 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.861832)

    Null deviance: 98.719 on 20 degrees of freedom
Residual deviance: 33.278 on 17 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

> summary(g.over)$dispersion
[1] 1.861832

Confidence Interval model2

```
> library(MASS)
> fact$prop <- with(fact, germ/total)
> g.over.alt <- glm(prop ~ extract + seed + extract:seed,
+ weights = total, family = quasibinomial, data = fact)
> confint(g.over.alt)
Waiting for profiling to be done...
                          2.5 %     97.5 %
(Intercept)            -0.91562380 0.07457178
extractcucumber        -0.12415989 1.21529201
seedaeg75              -0.74043242 0.45663323
extractcucumber:seedaeg75 -0.04413964 1.59702376
```

Estimates and CI for proportion with and without over dispersion

	$\hat{\pi}$	2.5%	97.5%
$\phi = 1$	0.36	0.31	0.42
$\hat{\phi} = 1.86$	0.36	0.28	0.45

To calculate this you need to install the R packages:

library(doBy)
library(R2HTML)

Example 2: Habitat preferences of lizards

- A study consists of two lizards type: Grohami and Opalinus.

- Response: number of sites (from the total) occupied by Grahami lizards.

- Covariates:

1. Height of the site (H).

2. Diameter (D).

3. Sun condition of the site (S, sun/ shade).

4. Time of the day (T).

Habitat preferences of lizards

```
> habitat
   G Total  S  D  H    T
1  20    22 S1 D1 H1 Early
2  8     9 S1 D1 H1  Mid
3  4     8 S1 D1 H1  Late
4  13    13 S1 D1 H2 Early
5  8     8 S1 D1 H2  Mid
6  12    12 S1 D1 H2  Late
7  8    11 S1 D2 H1 Early
8  4     5 S1 D2 H1  Mid
9  5     8 S1 D2 H1  Late
10 6     6 S1 D2 H2 Early
11 0     0 S1 D2 H2  Mid
12 1     2 S1 D2 H2  Late
13 34    45 S2 D1 H1 Early
14 69    89 S2 D1 H1  Mid
15 18    28 S2 D1 H1  Late
16 31    36 S2 D1 H2 Early
17 55    59 S2 D1 H2  Mid
18 13    16 S2 D1 H2  Late
19 17    32 S2 D2 H1 Early
20 60    92 S2 D2 H1  Mid
21 8    16 S2 D2 H1  Late
22 12    13 S2 D2 H2 Early
23 21    26 S2 D2 H2  Mid
24 4     8 S2 D2 H2  Late
```

S: sun conditions sun / shade).
D: diameter (<2 / > 2).
H: hight (< 5 / > 5).
T: time of day (early/ mid day/late).

Habitat preferences of lizards: model formulation

$$y_{ijkl} \sim B(n_{ijkl}, \pi_{ijkl})$$

Total ample size.

Number of sites occupied by Grahami lizards.

$$\pi_{ijkl} = \text{The probability that a site is occupied by Grahami lizards.}$$

$$g(\pi_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \dots$$

Habitat preferences of lizards: model formulation in R

Main effects model in R

$$g(\pi_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l$$

> f1<-glm((G/Total)~H+D+S+T,family="binomial",data=habitat)

R output

> summary(f1)

Call:
glm(formula = (G/Total) ~ H + D + S + T, family = "binomial",
    data = habitat)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-0.50878 -0.11019   0.02009   0.26466   0.52322

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0618   1.4060   1.466   0.143
HH2   1.0631   1.1222   0.947   0.343
DD2   -0.8798   1.0841 -0.812   0.417
SS2   -0.6415   1.0884 -0.589   0.556
TLate   -1.2054   1.2761 -0.945   0.345
TMid   0.0587   1.4590   0.040   0.968

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4.6730 on 22 degrees of freedom
Residual deviance: 1.5417 on 17 degrees of freedom
 (1 observation deleted due to missingness)
AIC: 28.658

Interpretation

```
Coefficients:
          Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0618     1.4060   1.466    0.143
HH2           1.0631     1.1222   0.947    0.343
DD2          -0.8798     1.0841  -0.812    0.417
SS2          -0.6415     1.0884  -0.589    0.556
TLate        -1.2054     1.2761  -0.945    0.345
TMid          0.0587     1.4590   0.040    0.968
```
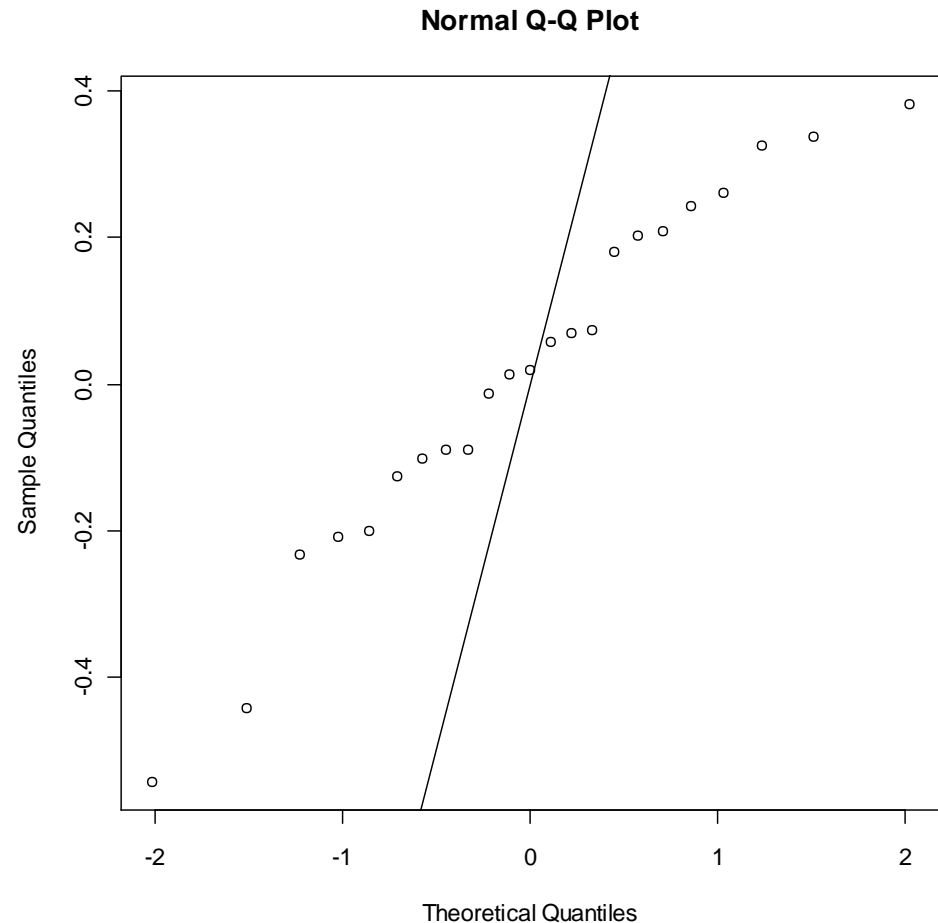
All the parameters estimates are not significant.

diagnostic

```
>r.pearson<-resid(f1, type="pearson")
> par(mfrow=c(1,1))
> qqnorm(r.pearson)
> abline(0,1)
```
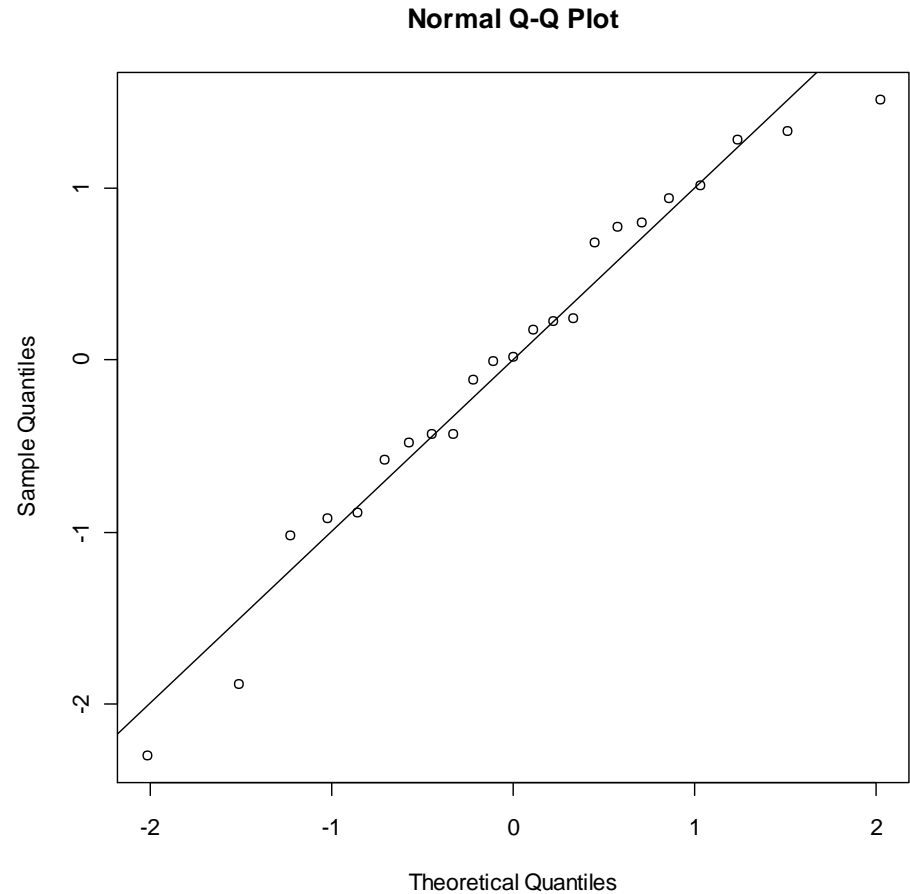
$$r_i^{\ P} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \sim N(0,1)$$

Normal Q-Q Plot

diagnostic

Normal Q-Q Plot

```
> mean(r.pearson)
[1] 0.01454735
> var(r.pearson)
[1] 0.05871372
```

**>r.s<-(r.pearson-mean(r.pearson))/
    sqrt((var(r.pearson)))**

```
> qqnorm(r.s)
> abline(0,1)
```

The variance of
pearson residual is
much smaller than 1

Taking into account overdispersion

glm((G/Total)~H+D+S+T,family="binomial",data=habitat)

glm((G/Total)~H+D+S+T,family=quasibinomial,data= habitat)

Taking into account overdispersion

```
> f1.over <- glm((G/Total)~H+D+S+T,family=quasibinomial,data= habitat)
> summary(f1.over)

Call:
glm(formula = (G/Total) ~ H + D + S + T, family = quasibinomial,
    data = habitat)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-0.50878 -0.11019  0.02009  0.26466  0.52322

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.0618    0.3883   5.310 5.76e-05 ***
HH2          1.0631    0.3099   3.430  0.00319 **
DD2         -0.8798    0.2994  -2.939  0.00918 **
SS2         -0.6415    0.3006  -2.134  0.04768 *
TLate       -1.2054    0.3524  -3.420  0.00326 **
TMid         0.0587    0.4029   0.146  0.88588
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for quasibinomial family taken to be 0.07626879)

```
    Null deviance: 4.6730  on 22  degrees of freedom
Residual deviance: 1.5417  on 17  degrees of freedom
  (1 observation deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 5
```
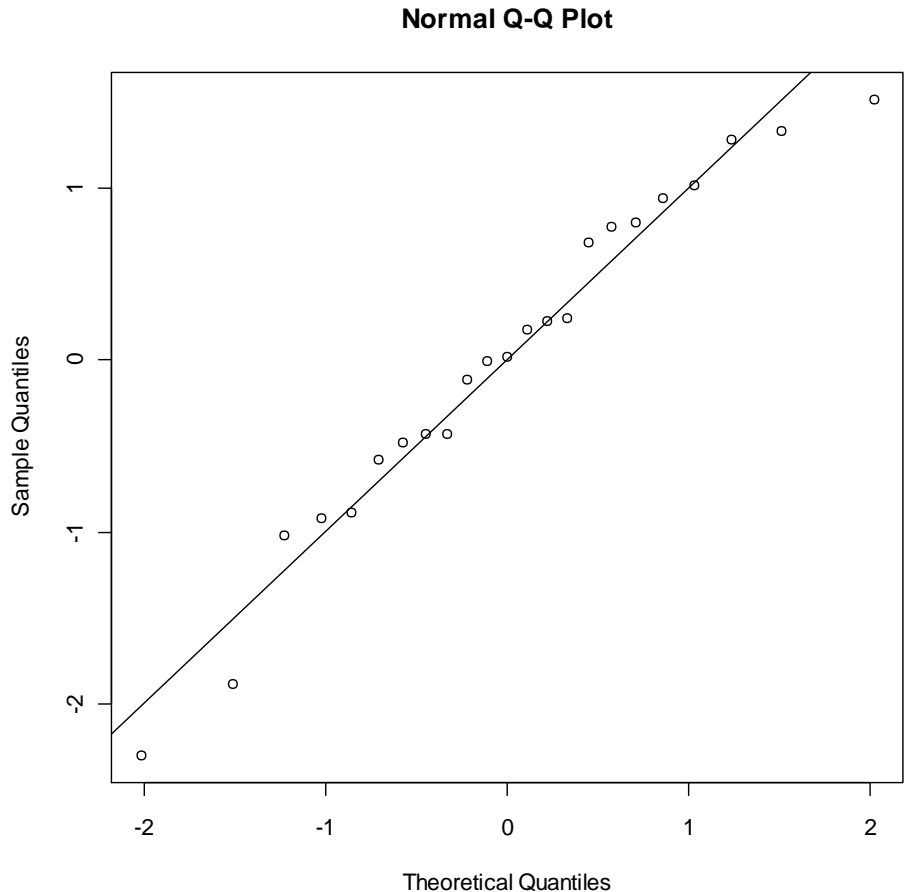
Diagnostic

Normal Q-Q Plot


```
> mean(r.pearson)
[1] 0.01454735
> var(r.pearson)
[1] 0.05871372
```

**>r.s<-(r.pearson-mean(r.pearson))/
sqrt((var(r.pearson)))**

```
> qqnorm(r.s)
> abline(0,1)
```

The variance of
pearson residual is
much smaller than 1

```
> summary(f1.over)$dispersion
[1] 0.07626879
```

Example 3: ship damage

- The ships data from the MASS package concern a type of damage caused by waves to the forward section of cargo-carrying vessels.

- The variables are
  - incidents number of damage incidents
  - service aggregate months of service
  - period period of operation : 1960-74, 75-79
  - year year of construction: 1960-64, 65-69, 70-74, 75-79
  - type type: '"A"' to '"E"'

- Here it makes sense to model the expected number of incidents per aggregate months of service.

Data in R

The data is available in the R the MASS library as:

> library(MASS)

> data(ships)

> ships2 <- subset(ships, service > 0)

> ships2$year <- as.factor(ships2$year)

➢ ships2$period <- as.factor(ships2$period)

> ships

	type	year	period	service	incidents
1	A	60	60	127	0
2	A	60	75	63	0
3	A	65	60	1095	3
4	A	65	75	1095	4
5	A	70	60	1512	6
.	.	.	.	.	.
37	E	70	60	1157	5
38	E	70	75	2161	12
39	E	75	60	0	0
40	E	75	75	542	1

Mean structure and model formulation in R

Model formulation:

$$Y_{ijk} \sim Poisson(\mu_{ijk})$$

$$g(\mu_{ijk}) = \mu + Tyep_i + Year_j + Period_k + \log(servise)$$

offset

Model formulation in R:

```
> glm1 <- glm(formula = incidents ~ type + year + period,
+         family = poisson(link = "log"), data = ships2,
+         offset = log(service))
```

$$\phi = 1$$

R output

> glm1 <- glm(formula = incidents ~ type + year + period,
+   family = poisson(link = "log"), data = ships2,
+   offset = log(service))
> summary(glm1)
Call:
glm(formula = incidents ~ type + year + period, family = poisson(link = "log"),
    data = ships2, offset = log(service))
Deviance Residuals:
   Min   1Q   Median   3Q   Max
-1.6768 -0.8293 -0.4370   0.5058   2.7912
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.40590   0.21744 -29.460 < 2e-16 ***
typeB   -0.54334   0.17759 -3.060 0.00222 **
typeC   -0.68740   0.32904 -2.089 0.03670 *
typeD   -0.07596   0.29058 -0.261 0.79377
typeE   0.32558   0.23588   1.380 0.16750
year65   0.69714   0.14964   4.659 3.18e-06 ***
year70   0.81843   0.16977   4.821 1.43e-06 ***
year75   0.45343   0.23317   1.945 0.05182 .
period75   0.38447   0.11827   3.251 0.00115 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 146.328 on 33 degrees of freedom
Residual deviance: 38.695 on 25 degrees of freedom
AIC: 154.56

Number of Fisher Scoring iterations: 5

Model 1: quasi-poisson log linear

> glm2 <- update(glm1, family = quasipoisson(link = "log"))

$$\phi \neq 1$$

> summary(glm2)
> Call:
glm(formula = incidents ~ type + year + period, family = quasipoisson(link = "log"),
   data = ships2, offset = log(service))

Coefficients:
```
        Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.40590   0.28276 -22.655  < 2e-16 ***
typeB       -0.54334   0.23094  -2.353  0.02681 *
typeC       -0.68740   0.42789  -1.607  0.12072
typeD       -0.07596   0.37787  -0.201  0.84230
typeE        0.32558   0.30674   1.061  0.29864
year65       0.69714   0.19459   3.583  0.00143 **
year70       0.81843   0.22077   3.707  0.00105 **
year75       0.45343   0.30321   1.495  0.14733
period75     0.38447   0.15380   2.500  0.01935 *
---
```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasipoisson family taken to be 1.691028)
   Null deviance: 146.328 on 33 degrees of freedom
Residual deviance: 38.695 on 25 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

Standard errors are changed since Φ>1.

Estimation of over dispersion parameter

> X2 <- sum(residuals(glm2, type = "pearson")^2)

> X2

[1] 42.27525

> phi <- X2/glm2$df.residual

> phi

[1] 1.69101

> phi <- g$deviance/glm2$df.residual

> phi

[1] 1.331111

CI for the parameters

> confint(glm2)

```
             2.5 %      97.5 %
(Intercept) -6.9789252 -5.86832189
typeB       -0.9793495 -0.07040017
typeC       -1.6043341  0.09972077
typeD       -0.8628299  0.63543426
typeE       -0.2880598  0.92322467
year65       0.3217115  1.08674462
year70       0.3882167  1.25564720
year75      -0.1562814  1.03712256
period75     0.0841713  0.68792187
```

Summary

- ANOVA table without and with assuming over dispersion

Effect	DF	$\phi = 1$	$\hat{\phi} = 1.67$
		P($>|\chi^2|$)	P($>$F)
Type	4	2.63E-11	2.29E-04
Period	1	1.1E-03	1.888E-02
Year	3	5.038e-09	5.777E-04