Applied Regression I

CMED6020 – Session 3

Eric Lau (ehylau@hku.hk)

School of Public Health
The University of Hong Kong

25 Jan 2021

Session 3 learning objectives

After this session, students should be able to

- Apply poisson and negative binomial regression models to count data
- Identify and apply suitable model to overdispersed data

Analysis of count data

Some characteristics of count data

- Nonnegative
- Always positively skewed
- Variance tends to increase with mean

→ Usually violates the assumptions of linear regression model

  - e.g. Homoscedasticity, Normality

- Linear regression model may not be appropriate to describe count data

- Examples of count data
  - daily number of ER visits, adverse drugs events, weekly number of influenza hospitalization

The Poisson distribution

- This statistical distribution can be very useful for modelling count data
- The Poisson distribution has only one parameter, μ
- The notation $Y \sim \text{Poisson}(\mu)$ indicates that a variable Y follows a Poisson distribution with parameter μ.
- Y can take any non-negative integer value, i.e. it could be 0, 1, 2, 3, …, 100, 101…

The probability mass function

- If $Y \sim \text{Poisson}(\mu)$, the probability of any particular value of Y is given by this probability mass function (pmf):

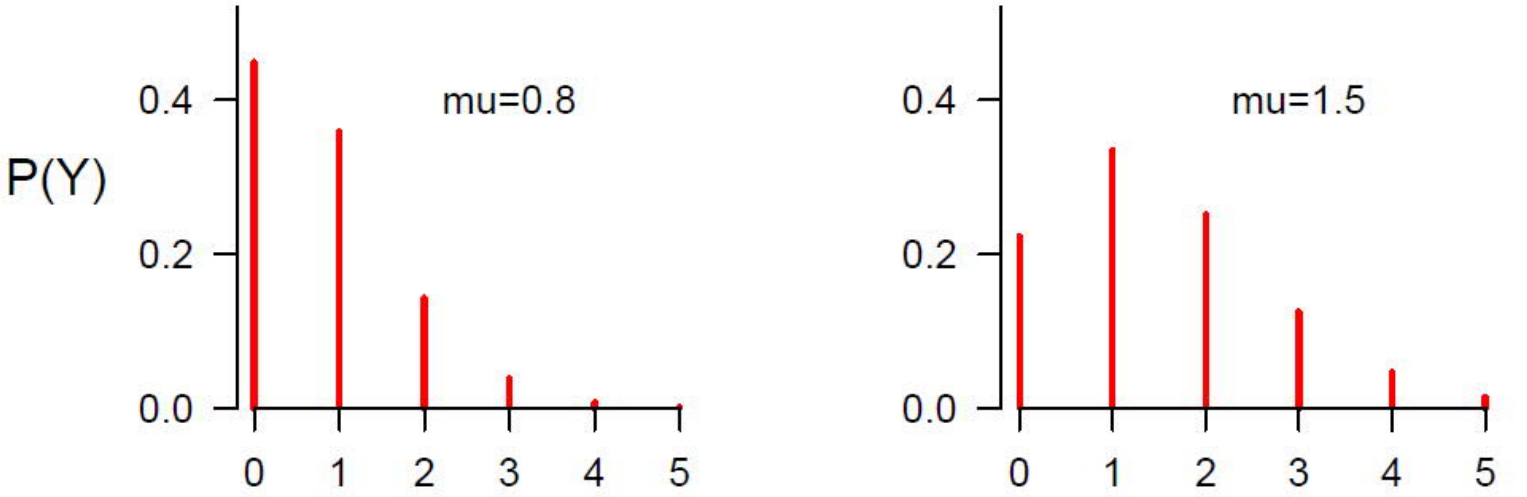
$$P(Y = y) = \frac{\mu^y \exp(-\mu)}{y!}$$

- Note that y! (y factorial) is the easy way to write the product

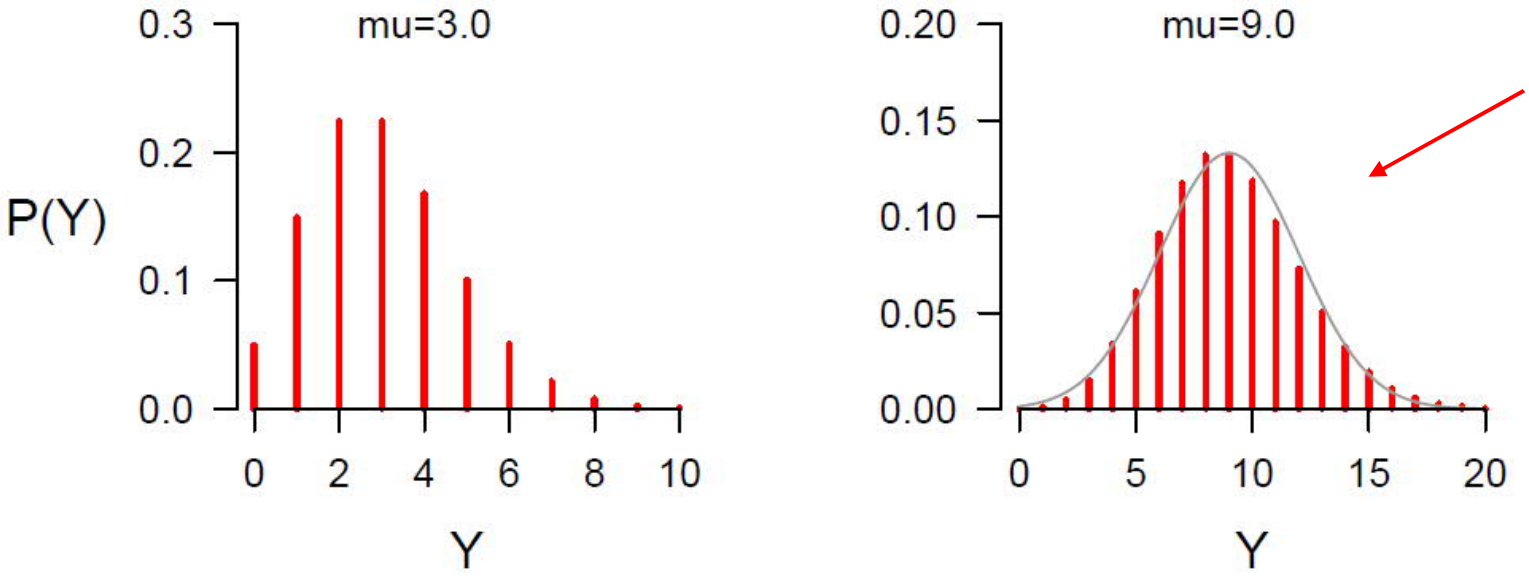
  y x (y – 1) x … x 1, e.g., 4! = 4 x 3 x 2 x 1 = 24

  0! = 1

Various Poisson distributions




```
plot(dpois(x=0:5, lambda=1.5),
type='h')
```

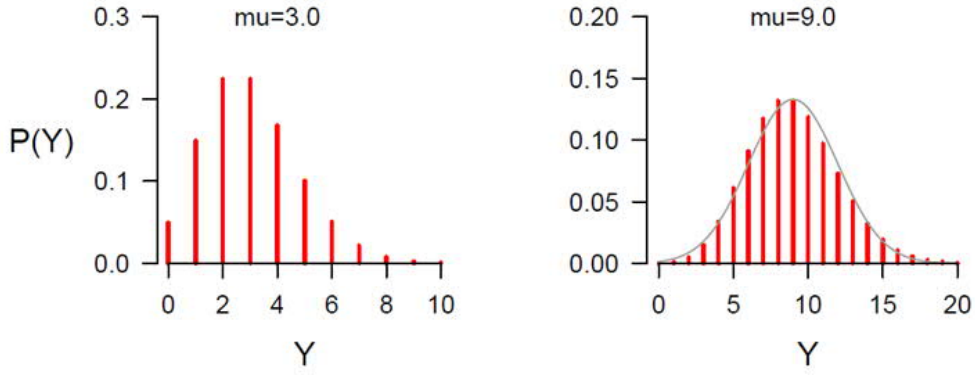


Here the distribution looks quite like a bell-shaped curve

Properties of the Poisson distribution

1. As μ increases, the 'mass' of the distribution shifts to the right. The mean value $E(Y) = \mu$
2. The variance equals the mean: $Var(Y) = E(Y) = \mu$
3. As μ increases, the probability of 0 decreases.
4. As μ increases (> 10), the Poisson distribution approximates a Normal distribution

- Count data often have properties which conflict with (2) and (3), and in those cases we might need to modify our models



Example dataset – accidents

- In the Prussian army, 1875-94, some unlucky soldiers died by being kicked by a horse (accidentally!)
- The full data are in “examplehorse.csv"
- The columns are ‘year’, ‘corps’ (army section), and number of deaths
- Can we predict the number of accidental deaths in 1895? Is there any difference over time, or between corps?

Dataset: examplehorse

	year	corps	deaths
1	1875	2	0
2	1876	2	0
3	1877	2	0
4	1878	2	2
5	1879	2	0
6	1880	2	2
7	1881	2	0
8	1882	2	0
9	1883	2	1
10	1884	2	1
11	1885	2	0
12	1886	2	0
13	1887	2	2
14	1888	2	1
15	1889	2	1
16	1890	2	0
17	1891	2	0
18	1892	2	2
19	1893	2	0
20	1894	2	0

Variables:

- ‘year’
- ‘corps’ – army section
- ‘deaths’ – number of deaths (outcome)

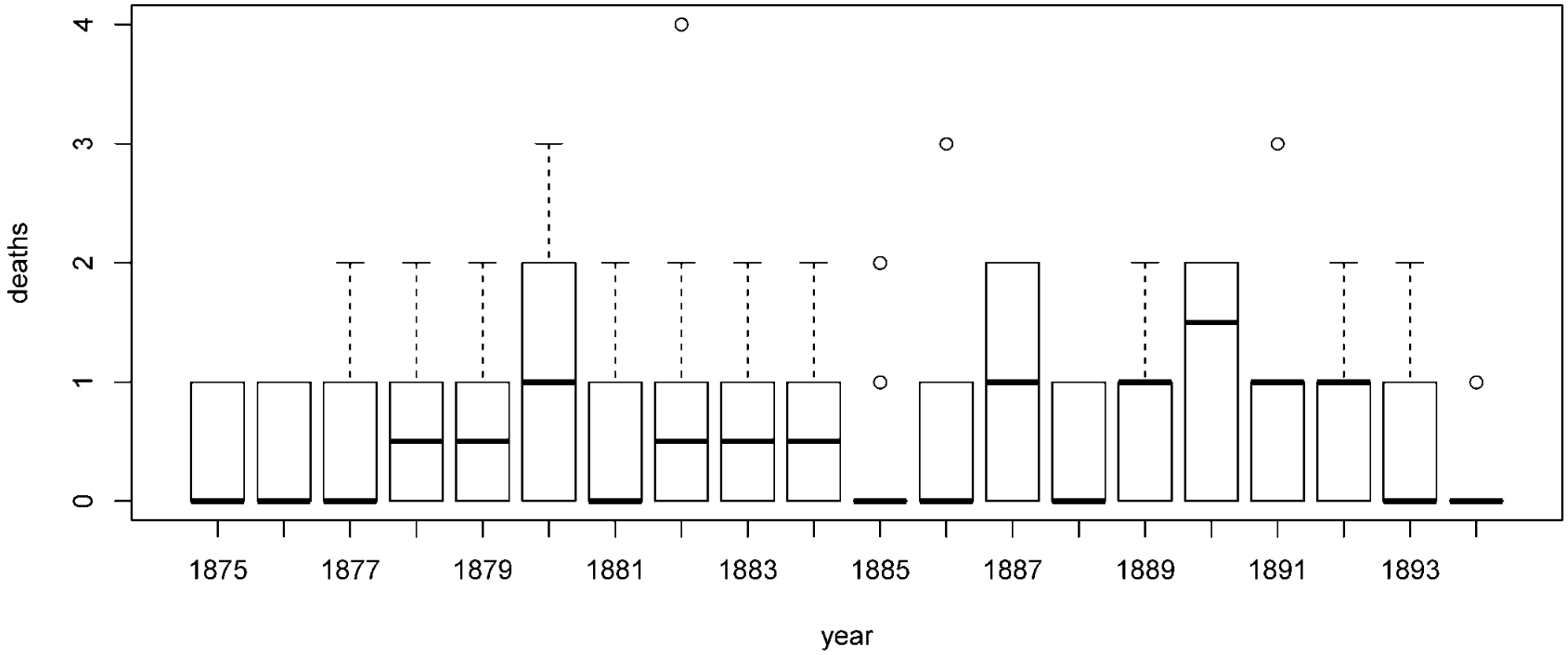
Your turn…
1. Plot the histogram for deaths
2. Calculate the mean and variance for deaths

Plot the histogram - output


```
> mean(horse$deaths)
[1] 0.61
> var(horse$deaths)
[1] 0.6109548
```

The mean is approximately equal to the variance

Boxplot - any obvious pattern over years?

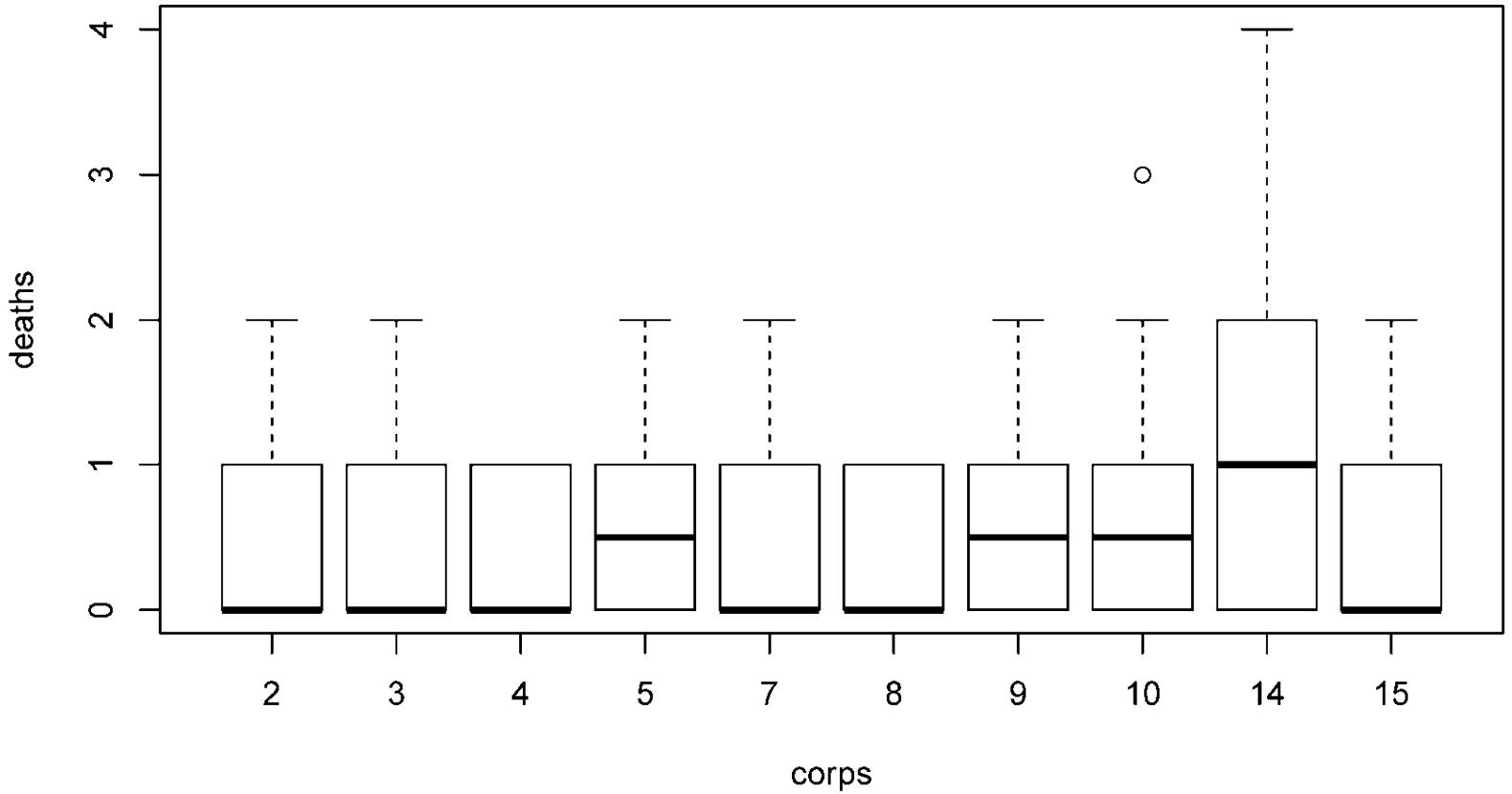



```
boxplot(deaths~year, data=horse, xlab='year', ylab='deaths')
```

Line plot - any obvious pattern over years?


```
horse.year <- aggregate(horse$deaths, by=list(horse$year), sum)
colnames(horse.year) <- c('year', 'deaths')
plot(horse.year$year, horse.year$deaths, type='l', xlab='year', ylab='deaths')
```

Boxplot - any obvious pattern between corps?




```
boxplot(deaths~corps, data=horse, xlab='corps', ylab='deaths')
```

Modelling the deaths

- The previous graphs suggested that there were no major differences between corps or between years
- Propose using the following simple model for the deaths:
  - Each year, and in each corps, deaths occurred randomly with mean μ (per year)
- To fit this model, we need to estimate μ
- What is the best way to estimate μ?

Fitting a Poisson model

- Our proposed model is a Poisson model, and the formal expression is:

$$Y_i \sim \text{Poisson}(\mu)$$

  where Y_i is the number of deaths in each corps in each year
- Using the method of *maximum likelihood*, the best estimate of μ is actually the sample mean, 0.61
- We can perform formal estimation in R…

Generalized Linear Model (GLM)

- Generalization of linear regression
- In multiple linear regression, the predicted value of the output Y, ie E(Y), equals $\beta_1 X_1 + \beta_2 X_2 + \ldots$
- In GLM, instead there is a function *l* such that

$$l(E(Y)) = \beta_1 X_1 + \beta_2 X_2 + \ldots$$

- *l* is called the link function
- Also, in GLMs errors don't have to follow the Normal distribution, but instead can follow Binomial, Poisson, Gamma distributions…

Families and link function

Regression model	Outcome variable	Distribution	Link function	Variance to mean relation
Ordinary linear	Continuous	Normal	Identity	Constant σ^2
Logistic	Binary	Bernoulli	Logit	$\mu(1 - \mu)$
Poisson	Count	Poisson	Log	μ
Negative binomial	Count	Negative binomial	Log	$\mu + \alpha\mu^2$

where logit(y) = log(y/(1-y))

GLM in R

glm(formula, data, subset, family, …)

- **formula** similar to lm, e.g. y~x+z
- ***data*** indicates the data frame to be used
- ***subset*** indicates that only some of the data should be used
- ***family*** is the distribution of the errors, for Poisson regression: family=poisson

A simple Poisson model in R - output

```
> summary(glm(deaths ~ 1, data=horse, family=poisson))

Call:
glm(formula = deaths ~ 1, family = poisson, data = horse)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1045  -1.1045  -1.1045   0.4567   2.8748

...
```

A simple Poisson model in R - output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.49430    0.09053   -5.46 4.77e-08 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 212.47  on 199  degrees of freedom
Residual deviance: 212.47  on 199  degrees of freedom
AIC: 414.21
```

Estimated mean from poisson model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.49430    0.09053   -5.46 4.77e-08 ***
```

```
> pois.horse <- glm(deaths ~ 1, data=horse,
   family=poisson)
> coef(pois.horse)
(Intercept)
 -0.4942963
> exp(coef(pois.horse))
(Intercept)
       0.61
```

Note that we are estimating a = log(μ), or equivalently μ = exp(a).

Why do we do this?

Why do we use a ‘log-link’?

- Estimating log(μ) instead of μ has a few useful consequences:

  1. we can use Normal distribution theory for the estimated parameters (e.g. to compose confidence intervals)
  2. The estimation procedure works better
  3. We can easily specify the effects of the explanatory variables to be proportional on the event rate (see later)

Interpreting the output

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.49430    0.09053   -5.46 4.77e-08 ***
> exp(coef(pois.horse))
(Intercept)
       0.61
> exp(confint(pois.horse))
2.5 %    97.5 %
0.5080600 0.7247332
```

Mean = exp(-0.494) = 0.61; 95% CI = (0.51, 0.73)

Under this model, the expected number of deaths (absolute risk) *in any corps in any year* is 0.61

Model checking

- We can compare the observed event counts to data that we might have expected, under a Poisson(0.61) model
- First, we can check the observed number of years/corps with 0 event, with 1 event etc. Then we work out how many years/corps we would expect to see with 0 event, 1 event, etc.

Observed data

```
> table(horse$deaths)

  0   1   2   3   4
109  65  22   3   1

> prop.table(table(horse$deaths))

    0     1     2     3     4
0.545 0.325 0.110 0.015 0.005
```

Expected data for 200 samples

- Recall the pdf of Poisson distribution:

$$P(Y = y) = \frac{\mu^y \exp(-\mu)}{y!}$$

- We could derive the expected frequencies for y = 0, 1, …, 5

```
> n <- nrow(horse)
> n*dpois(0:5, exp(coef(pois.horse)))
[1] 108.67017375  66.28880605  20.21808587   4.11101080
    0.62692915   0.07648536
```

Observed vs expected frequencies

Event count	Observed	Expected
0	109	108.67
1	65	66.29
2	22	20.22
3	3	4.11
4	1	0.63
5	0	0.08

Very good agreement!

Formal model goodness-of-fit

When we fit the Poisson model in R, at the bottom of the summary:

```
    Null deviance: 212.47  on 199  degrees of freedom
Residual deviance: 212.47  on 199  degrees of freedom
AIC: 414.21
```

This performs a formal goodness-of-fit test, comparing the observed data to the expected data (as in previous slide).

The null hypothesis is that the Poisson distribution fits the data well.

Under H_0, residual deviance follows a χ^2 distribution with 199 degrees of freedom

As a quick approximation, residual deviance/df should not be too much bigger than 1 (better if it is below 1).

Adjusting for independent (explanatory) variables

- We can generalise the Poisson distribution for a count outcome Y_i so it can depend on explanatory variables.
- By allowing the mean μ_i to vary, depending on some explanatory variables x_i through a log link. The pdf of Y_i is then given by:

$$P(Y_i = y_i) = \frac{\mu_i^{y_i} \exp(-\mu_i)}{y_i!}$$

where $\mu_i = \exp(\beta' x_i)$

The linear predictor $\beta' x_i$

- In regression models, we often use the linear predictor $\beta' x$. This is a shorthand form of an additive set of covariate effects, including an intercept term written as β_0:

$$\beta' x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

- The covariate effects are additive on this scale, but multiplicative on the log scale:

$$\mu = \exp(\beta' x) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots) =$$

$$\exp(\beta_0) \times \exp(\beta_1 x_1) \times \exp(\beta_2 x_2) \times \ldots$$

Allowing for differences between corps

- In the dataset, there are corps 2, 3, 4, 5, 7, 8, 9, 10, 14 and 15
- Assume corps 2 as the reference group, we can allow for possible differences between army corps by fitting the model

$$Y_i \sim \text{Poisson}(\mu_i), \text{ where}$$

$$\mu_i = \exp(\beta_0 + \beta_{c3}x_{c3i} + \beta_{c4}x_{c4i} + \ldots)$$

- Note if $\text{corps}_i = 3$, then $x_{c3i} = 1$ and all other $x_{c3j} = 0$ ($j = 4, 5, 7, \ldots, 15$)
- In this model, we have a baseline group (the 2nd corps), and each of the other corps can have a slightly different mean to the baseline group

A Poisson model with covariates in R

```
> class(horse$corps)
[1] "integer"
> horse$corps <- as.factor(horse$corps)
> summary(glm(deaths~corps, data=horse, family=poisson))

Call:
glm(formula = deaths ~ corps, family = poisson, data =
   horse)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5492  -1.0955  -0.8367   0.5438   2.0079
```

Continued...

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.108e-01  2.887e-01  -1.770   0.0768 .
corps3      -3.654e-09  4.082e-01   0.000   1.0000
corps4      -4.055e-01  4.564e-01  -0.888   0.3744
corps5      -8.701e-02  4.174e-01  -0.208   0.8349
corps7      -3.654e-09  4.082e-01   0.000   1.0000
corps8      -5.390e-01  4.756e-01  -1.133   0.2571
corps9       8.004e-02  4.003e-01   0.200   0.8415
corps10      2.231e-01  3.873e-01   0.576   0.5645
corps14      6.931e-01  3.535e-01   1.961   0.0499 *
corps15     -4.055e-01  4.564e-01  -0.888   0.3743
---
```

Continued...

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 212.47  on 199  degrees of freedom
Residual deviance: 196.89  on 190  degrees of freedom
AIC: 416.64

Number of Fisher Scoring iterations: 5
```

deviance/df is close to 1 (= 1.04)

Incidence rate ratios (IRR)

- Let's think about the predicted mean in army corps 14, compared to the rate in army corps 2. In our model the two predicted means are:

$$\mu_{14} = \exp(\beta_0 + \beta_{c14}) = \exp(\beta_0) \times \exp(\beta_{c14})$$
$$\mu_2 = \exp(\beta_0)$$

- So the ratio of the means is μ_{14} / μ_2, or simply $\exp(\beta_{c14})$. We can say that the mean in corps 14 is $\exp(\beta_{c14})$ times the mean in corps 2.

The IRR between corps - output

```
> pois.corps <- glm(deaths~corps, data=horse,
   family=poisson)

> round(exp(coef(pois.corps)),2)
(Intercept)      corps3      corps4      corps5      corps7
       0.60        1.00        0.67        0.92        1.00
corps8      corps9     corps10     corps14     corps15
  0.58        1.08        1.25        2.00        0.67
```

The IRR between corps - output

```
> round(exp(cbind(coef(pois.corps),
   confint(pois.corps))),2)
                  2.5 % 97.5 %
(Intercept) 0.60  0.32   1.01
corps3      1.00  0.44   2.25
corps4      0.67  0.26   1.61
corps5      0.92  0.40   2.09
corps7      1.00  0.44   2.25
corps8      0.58  0.22   1.45
corps9      1.08  0.49   2.41
corps10     1.25  0.59   2.72
corps14     2.00  1.02   4.14
corps15     0.67  0.26   1.61
```

Conclusions from example

- The overall mean number of deaths per corps per year from horse-kicks was 0.61
- A Poisson model fitted the data very well
- The corps were fairly similar, although the mean in corps 14 seemed to be a bit higher than the rest (IRR=2.0, 95% CI: 1.0-4.1)

IRR and relative risk

- It is very common to present the results of Poisson regression models in terms of the estimated IRRs, which can also be interpreted as relative risks
- If the event of interest is death (as in our horse-kicks example), then the IRR for a particular corps can be interpreted as the relative risk of death in that corps, compared to the baseline group (corps 2)

Counts and rates

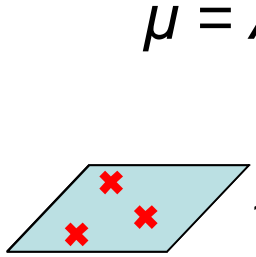
- Are the death rates the same in groups A and B?
  - group A: 5 deaths in 100 people
  - group B: 10 deaths in 200 people
- Are the death rates the same in groups C and D?
  - group C: 20 deaths in 1000 people in 1 year
  - group D: 20 deaths in 1000 people in 2 years

Poisson distribution is for counts

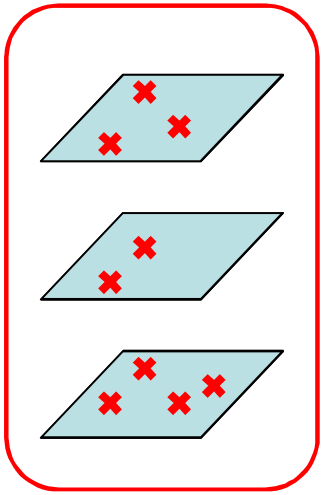
- In a Poisson model, the dependent variable must be an (integer) count
- The dependent variable cannot be a rate
- How can we model data with varying exposure periods (whether in terms of population size or time)?
- Is there a way to incorporate different lengths of exposure (different person-years) in the model?

Exposures and Rates

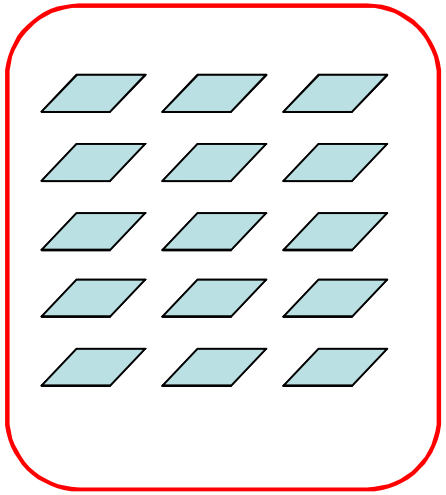
- A common way of using the Poisson distribution is for counting the occurrence of events where:
  - Each event happens with some rate λ per unit time;
  - Each group is exposed for a period A (often given in person-years);
  - The occurrence of one event doesn't have any effect on the chance of other events occurring
- Then the event count will be follow a Poisson distribution with mean $\mu = A\lambda$



~ Poisson(3)



~ Poisson(3×3)



~ Poisson(15×3)

Example 2 – Lung Cancer Rates

- The file ‘examplelung.csv’ contains data on the number of lung cancer cases (count) and population sizes (pop) by age groups in four Danish cities, in the year 1995
- We are interested in comparing the cancer incidence between cities. However the cities are not all the same size, so we wouldn’t expect the *counts* to be the same. What can we do?

Poisson regression with offsets

- We can model the lung cancer incidence Y_i (in a particular city and age group of size A_i) by a Poisson model:

$$Y_i \sim \text{Poisson}(A_i \lambda_i)$$

where $\lambda_i = \exp(\beta x_i)$, and x_i includes terms for city and age group

- Under the model, $\log(A_i)$ is the offset term:

$$\begin{aligned} E(Y_i) = \mu_i &= A_i \lambda_i \\ &= A_i \exp(\beta x_i) \\ &= \exp(\log(A_i) + \beta x_i) \end{aligned}$$

Prepare dataset in R

```
> # After reading and storing the dataset in "lung"
> summary(lung)
         city     age.gp      count              pop
 Fredericia:6   a50-54:4   Min.   : 2.000   Min.   : 509.0
 Horsens   :6   a55-59:4   1st Qu.: 7.000   1st Qu.: 628.0
 Kolding   :6   a60-64:4   Median :10.000   Median : 791.0
 Vejle     :6   a65-69:4   Mean   : 9.333   Mean   :1100.3
                a70-74:4   3rd Qu.:11.000   3rd Qu.: 954.8
                a75+  :4   Max.   :15.000   Max.   :3142.0
```

Specify offset in R

```
Call:
glm(formula = count ~ offset(log(pop)) + city + age.gp,
   family = poisson, data = lung)


Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.63573  -0.67296  -0.03436   0.37258   1.85267
```

Specify offset in R

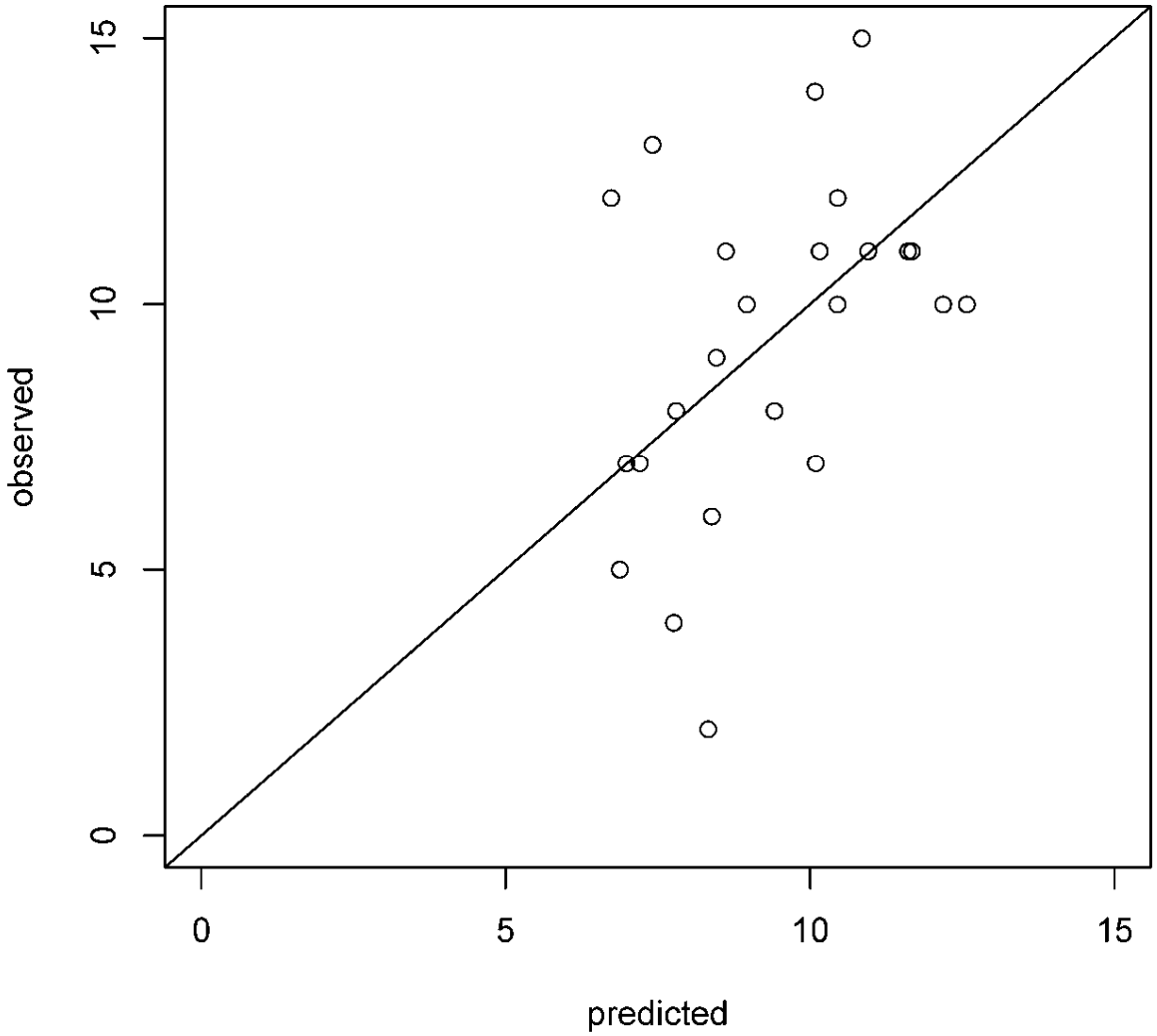
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.6321     0.2003 -28.125  < 2e-16 ***
cityHorsens   -0.3301     0.1815  -1.818   0.0690 .
cityKolding   -0.3715     0.1878  -1.978   0.0479 *
cityVejle     -0.2723     0.1879  -1.450   0.1472
age.gpa55-59   1.1010     0.2483   4.434 9.23e-06 ***
age.gpa60-64   1.5186     0.2316   6.556 5.53e-11 ***
age.gpa65-69   1.7677     0.2294   7.704 1.31e-14 ***
age.gpa70-74   1.8569     0.2353   7.891 3.00e-15 ***
age.gpa75+     1.4197     0.2503   5.672 1.41e-08 ***
...
```

R output

```
> pois.lung <- glm(count~offset(log(pop))+city+age.gp,
   data=lung, family=poisson)
> round(exp(cbind(coef(pois.lung), confint(pois.lung))),3)
                     2.5 % 97.5 %
(Intercept)  0.004 0.002  0.005
cityHorsens  0.719 0.503  1.026
cityKolding  0.690 0.476  0.995
cityVejle    0.762 0.525  1.099
age.gpa55-59 3.007 1.843  4.901
age.gpa60-64 4.566 2.907  7.236
age.gpa65-69 5.857 3.748  9.249
age.gpa70-74 6.404 4.043 10.212
age.gpa75+   4.136 2.523  6.762
```

We can see there is a statistically significant difference between Kolding (city 3) and Fredericia (city 1), since the IRR is 0.69 with 95% CI (0.476, 0.995), p-value = 0.048

Model checking




```
plot(predict(pois.lung, type='response'),
lung$count, xlim=c(0,15), ylim=c(0,15),
xlab='predicted', ylab='observed')
abline(a=0, b=1)
```

Predicted values

predict(object, newdata=NULL, type=c(“link”, “response”, “terms”), …)

- **object** is a fitted GLM model
- ***newdata*** (optional, data frame) to predict values for new data
- ***type:*** control the scale of the prediction: “response” will give the predicted values for the response variable E(Y); “link” will give the predicted values for the link function *l*(E(Y))

Model checking

Let's compare the observed and predicted incidence rates (per 1000 p-y) by age group, for Fredericia (city 1).

Recall $\mu_i = A_i\lambda_i = A_i\, exp(\beta x_i)$

Age	Observed I.R.	Predicted I.R.
50-54	3.60	3.58
55-59	13.75	10.77
60-64	15.49	16.35
65-69	17.21	20.98
70-74	21.61	22.93
75+	16.53	14.81

```
# observed incidences
with(lung,round(count[city=="Fredericia"]/
pop[city=="Fredericia"]*1000,2))

# predicted incidences
new <- data.frame(city="Fredericia",
age.gp=lung$age.gp[1:6],
pop=lung$pop[1:6])

round(predict(pois.lung, newdata=new,
type='response')/
lung$pop[lung$city=="Fredericia"]*1000,2)
```

Goodness-of-fit test

```
> summary(pois.lung)
...
    Null deviance: 129.908  on 23  degrees of freedom
Residual deviance:  23.447  on 15  degrees of freedom
AIC: 137.84

> deviance(pois.lung)/df.residual(pois.lung)
[1] 1.563165
```

Deviance/df = 1.56 is a bit larger than 1.

Conclusions from example

- We can present the estimated incidence rate ratios of lung cancer between the four cities, and in different age groups
- We can say that incidence increases with age
- We can say that the highest age-adjusted incidence is in Fredericia (city 1) and the lowest age-adjusted incidence is in Kolding (city 3)

		2.5 %	97.5 %
(Intercept)	0.004	0.002	0.005
cityHorsens	0.719	0.503	1.026
cityKolding	0.690	0.476	0.995
cityVejle	0.762	0.525	1.099
age.gpa55-59	3.007	1.843	4.901
age.gpa60-64	4.566	2.907	7.236
age.gpa65-69	5.857	3.748	9.249
age.gpa70-74	6.404	4.043	10.212
age.gpa75+	4.136	2.523	6.762

Mortality decrease according to socioeconomic groups during the economic crisis in Spain: a cohort study of 36 million people

Enrique Regidor, Fernando Vallejo, José A Tapia Granados, Francisco J Viciana-Fernández, Luis de la Fuente, Gregorio Barrio

	APR 2004–07 (before crisis, 95% CI)	APR 2008–11 (during crisis, 95% CI)	Effect size (95% CI)	p value
All causes				
Low (<72 m²)	1·7 (1·2 to 2·1)	3·0 (2·5 to 3·5)	1·4 (0·9 to 1·9)	<0·001
Medium (72–104 m²)	1·7 (1·3 to 2·1)	2·8 (2·5 to 3·2)	1·1 (0·8 to 1·5)	<0·001
High (>104 m²)	2·0 (1·4 to 2·5)	2·1 (1·6 to 2·7)	0·1 (–0·4 to 0·7)	0·610
Cancer (C00–C99)				
Low (<72 m²)	0·6 (–0·2 to 1·3)	0·6 (–0·1 to 1·4)	0·0 (–0·7 to 0·8)	0·996
Medium (72–104 m²)	1·0 (0·4 to 1·5)	0·9 (0·3 to 1·5)	–0·1 (–0·5 to 0·6)	0·842
High (>104 m²)	1·5 (0·7 to 2·3)	–0·3 (–1·1 to 0·6)	–1·8 (–2·6 to –0·9)	<0·001
Cardiovascular diseases (I00–I99)				
Low (<72 m²)	2·6 (1·6 to 3·6)	5·6 (4·5 to 6·6)	3·0 (2·0 to 4·0)	<0·001
Medium (72–104 m²)	3·1 (2·4 to 3·9)	5·2 (4·4 to 6·1)	2·1 (1·3 to 2·9)	<0·001
High (>104 m²)	2·6 (1·4 to 3·7)	5·4 (4·2 to 6·6)	2·9 (1·7 to 4·1)	<0·001
Respiratory diseases (J00–J99)				
Low (<72 m²)	0·6 (–1·1 to 2·4)	5·2 (3·4 to 7·1)	4·6 (2·8 to 6·4)	<0·001
Medium (72–104 m²)	–0·7 (–2·2 to 0·8)	5·9 (4·4 to 7·4)	6·6 (5·1 to 8·1)	<0·001
High (>104 m²)	1·6 (–0·7 to 3·9)	4·9 (2·5 to 7·2)	3·3 (0·9 to 5·6)	0·006
Gastrointestinal diseases (K00–K93)				
Low (<72 m²)	3·1 (1·3 to 4·9)	3·3 (1·4 to 5·3)	0·2 (–1·7 to 2·1)	0·818
Medium (72–104 m²)	1·7 (0·2 to 3·3)	3·0 (1·4 to 4·5)	1·2 (–0·3 to 2·8)	0·126
High (>104 m²)	0·1 (–2·4 to 2·5)	4·5 (2·1 to 6·9)	4·5 (2·0 to 6·9)	<0·001

Table 2: Comparison of time trends in premature mortality before and during the 2008 economic crisis, by cause of death (ICD-10 code*) and socioeconomic group according to household space in the 2001 Spanish Census Cohort of people aged 10–74 years

Next, we assessed how the trend in mortality risk changed in each socioeconomic group between pre-crisis and crisis periods. For that purpose, we modelled the age-adjusted mortality in each group and time period as a linear trend and computed the slope of that trend. We estimated the annual percentage change (APC) in mortality rate by fitting Poisson regression models in which the annual death count was the dependent variable and the total person-years observed was the offset of the regression. The APC was computed as $100 \times [\exp(\beta)-1]$, where β was the regression coefficient for the calendar-year. We computed 95% CIs of APC by

Regidor et al., Lancet, 2016

Overdispersion

Overdispersion

- In many situations, we find that count data have a particular property which makes the Poisson model unsuitable for use
- That is, in many applications we find that the variance of our count data is appreciably larger than the mean
  - (Poisson model assumes variance equals the mean: $Var(Y) = E(Y) = \mu$)
- This is known as ‘overdispersion’.
  - If we fit a Poisson model to overdispersed data, our results will not be reliable, so what can we do?

Generalising the Poisson model

- We can generalise the Poisson regression model to include a ‘random effect’ in the rate λ_i:

$$Y_i \sim \text{Poisson}(\mu_i), \text{ where } \mu_i = A_i\lambda_i \text{ and } \lambda_i = \exp(\beta' x_i + \varepsilon_i)$$

- Here ε_i is a ‘random effect’. We can think of ε_i as representing the effects of one or more unobserved explanatory variables

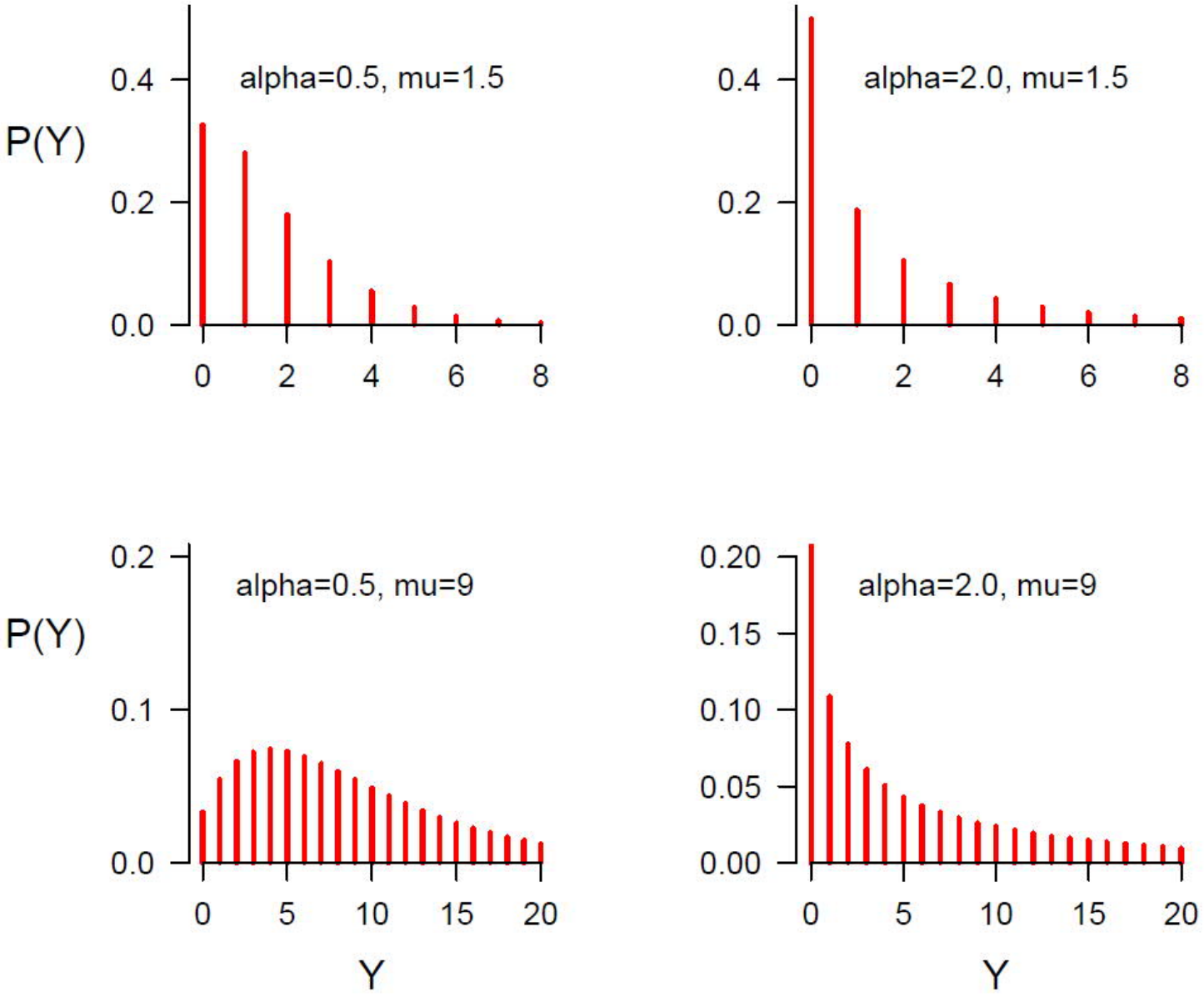
The Negative Binomial model

- In the special case where we allow $\exp(\varepsilon_i)$ to follow a gamma distribution with mean 1 and variance α, then Y_i follows a *negative binomial distribution* with
  - Mean = μ_i
  - Variance = $\mu_i + \alpha\mu_i^2$

- Check this link for pdf – http://en.wikipedia.org/wiki/Negative_binomial

```
#Example: lambda=5, gamma mean=1, variance=2
rpois(10000, 5*rgamma(10000,1/2,1/2))
```

Examples of NegBin distribution



Properties of the Negbin distribution

1. As μ increases, the 'mass' of the distribution shifts to the right. The mean value $E(Y) = \mu = A\lambda$
2. The variance is given by: $Var(Y) = \mu + \alpha\mu^2 > \mu$
3. As μ increases, the probability of 0 decreases.
4. As α decreases towards 0, the distribution becomes more like a Poisson

- Particularly note that α does not affect the mean
  - To check whether there is overdispersion in our data, we can fit the NegBin model and see how likely it is that $\alpha = 0$

Example 3 – Epileptic seizures

- The file ‘exampleepilepsy.csv’ contains data from a RCT which compared progabide (therapy=1) with a placebo (therapy=0)
- For each patient we know:
  - the number of seizures (*x*) experienced in 8 weeks prior to randomisation;
  - the number of seizures (*y*) experienced during the 8-week trial (let's set $A = 1$ since all patients had the same follow-up)
  - We also know each patient's age
- Was progabide more effective than placebo in reducing seizure frequency?

Check the overdispersion

```
> mean(epilepsy$y)
[1] 28.41379
> var(epilepsy$y)
[1] 823.4749
```

We can see that y seems to be overdispersed, since the variance (823.475) is much larger than the mean (28.41)

Practice

Fit a poisson regression model for the number of seizures with an intercept only.

Does your model fit the data satisfactorily?

Fit a NegBin regression model

```
> require(MASS)
> summary(glm.nb(y~1, data=epilepsy))
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.3469     0.1141   29.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.3898) family taken
   to be 1)
```

Hence the fitted mean (μ) is exp(3.347) = 28.4 (The observed mean was 28.4.)

Overdispersion

```
          Theta:  1.390
      Std. Err.:  0.253

 2 x log-likelihood:  -503.188
```

In glm.nb, θ = 1 / α, so the fitted variance ($\mu + \alpha\mu^2$) is $28.4 + 28.4^2 / 1.39 = 609.3$. (The observed variance was 823.5.)

Fit a NegBin regression model for the therapy effect

```
> nb.therapy1 <- glm.nb(y~therapy, data=epilepsy)
> round(exp(cbind(coef(nb.therapy1),
   confint(nb.therapy1))),3)
                    2.5 % 97.5 %
(Intercept) 34.393 25.522 47.760
therapy      0.664  0.428  1.028

> deviance(nb.therapy1)/df.residual(nb.therapy1)
[1] 1.14985
```

The estimated effect of therapy (unadjusted) is to reduce the rate of seizures by 34% (RR = 0.66, 95% CI = 0.43 -1.03)

The model fit the data satisfactorily.

Adjust for the original covariates (*x*, age)

```
> nb.therapy2 <- glm.nb(y~therapy+x+age, data=epilepsy)
> round(exp(cbind(coef(nb.therapy2),
   confint(nb.therapy2))),3)
                    2.5 % 97.5 %
(Intercept) 6.485 2.865 14.747
therapy     0.841 0.620  1.141
x           1.031 1.023  1.039
age         1.016 0.991  1.042
```

Adjust for log-transfomed *x* and age

```
> epilepsy$log10x <- log10(epilepsy$x)
> nb.therapy3 <- glm.nb(y~therapy+log10x+age,
   data=epilepsy)
> round(exp(cbind(coef(nb.therapy3),
   confint(nb.therapy3))),3)
                    2.5 % 97.5 %
(Intercept) 0.922 0.345  2.462
therapy     0.737 0.553  0.981
log10x      8.813 5.592 13.989
age         1.014 0.990  1.038
```

The estimated effect of therapy (adjusted) is to reduce the rate of seizures by 26% (RR = 0.74, 95% CI = 0.55-0.98)

Comparing models

- We can compare models based on the Akaike Information Criterion (AIC), which is defined as

$$AIC = 2k - 2ln(L)$$

  where k is the number of parameters in the model, L is the likelihood
- A lower AIC indicates a ‘better’ model
- As a rule of thumb, a difference in AIC of 2 or more can be regarded as significant

Model checking

Model	terms	log-likelihood	k	AIC
0	Intercept	-251.6	2	507.2
1	therapy	-249.9	3	505.9
2	therapy age x	-226.4	5	462.9
3	therapy age $\log_{10} x$	-222.8	5	455.7

Note that in addition to the explanatory variables, each model includes an intercept term β_0 and the 'scale' parameter α. From the above table we conclude that the AIC favors Model 3.

```
AIC(nb.epilepsy0, nb.therapy1, nb.therapy2, nb.therapy3)

logLik(…)
```

Association of gestational age and growth measures at birth with infection-related admissions to hospital throughout childhood: a population-based, data-linkage study from Western Australia

Jessica E Miller, Geoffrey C Hammond*, Tobias Strunk*, Hannah C Moore, Helen Leonard, Kim W Carter, Zulfiqar Bhutta, Fiona Stanley, Nicholas de Klerk, David P Burgner*

	Any infection	Invasive bacterial	Gastrointestinal	Lower respiratory tract
Gestational age (weeks)				
<28	2·91 (2·55–3·33)	2·63 (1·45–4·79)*	2·22 (1·60–3·09)	4·43 (3·66–5·36)
28–29	2·49 (2·24–2·78)	2·61 (1·65–4·13)	2·22 (1·76–2·80)	4·77 (4·07–5·59)
30–31	2·29 (2·07–2·53)	2·22 (1·49–3·33)	2·29 (1·90–2·75)	3·83 (3·32–4·43)
32–34	1·72 (1·64–1·81)	1·44 (1·15–1·81)*	1·64 (1·49–1·81)	2·48 (2·30–2·68)
35	1·57 (1·49–1·65)	1·72 (1·37–2·16)	1·64 (1·47–1·83)	1·91 (1·77–2·07)
36	1·44 (1·39–1·49)	1·28 (1·08–1·51)†	1·41 (1·32–1·51)	1·72 (1·61–1·84)
37	1·31 (1·28–1·34)	1·20 (1·07–1·35)	1·34 (1·28–1·40)	1·45 (1·38–1·52)
38	1·15 (1·13–1·17)	1·02 (0·94–1·12)§	1·13 (1·09–1·17)	1·22 (1·18–1·26)
39–40 (reference)	1·00	1·00	1·00	1·00
41	0·94 (0·92–0·96)	0·86 (0·78–0·95)†	0·92 (0·89–0·96)	0·96 (0·92–0·99)‡
≥42	0·99 (0·94–1·04)§	0·77 (0·60–1·00)‡	0·92 (0·83–1·01)§	1·08 (0·99–1·19)§

Table 2: Risk of childhood infection-related admissions to hospital

Statistical analysis

For each child included in the analysis, we calculated time at risk from birth-related hospital discharge to death, their 18th birthday, or end of the study (Dec 31, 2010), whichever occurred first. A χ^2 test of independence was done for children with and without recorded infection-related admissions to hospital with dichotomous and categorised measures. Likelihood ratio tests for linearity and departure from linearity were also done for ordered variables. The primary outcomes were the number and type of infection-related admissions to hospital. We calculated rate ratios (RR) for gestational age, birthweight, and birth length measures using a multilevel negative binomial regression framework (with births grouped by mother), and adjusted for maternal age at delivery (<20, 20–24, 25–29, 30–24, ≥35 years), birth year (2-year blocks), birth season, parity (previous

Miller et al., Lancet, 2016

Practice

Fit a poisson regression model for the number of seizures, to estimate the therapy effect adjusted for $\log_{10}x$ and age.

Does your model fit the data satisfactorily?

How does it compare with the negative binomial model?

Generalized linear model (GLM)

- Linear, logistic, poisson, negative binomial regression models all belong to GLM

Main components of GLM:

- Systematic component:
  - Link function *g* which links the outcome variable to the linear predictor
  - $g(E(Y)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$
- Random component:
  - Specifying variance to mean relation
  - e.g. poisson: $\sigma^2 = \mu$
  - negative binomial: $\sigma^2 = \mu + \alpha\mu^2$

Review

- We have seen how to describe, present and analyse count data with Poisson and negative binomial regression
- We have discussed the way to handle different exposures (offsets), and overdispersed data
- We have seen how to interpret and present the results of count regression models in terms of absolute and relative risks
- We have discussed model selection and model checking

Further reading

- Long, J. S. *Regression models for categorical and limited dependent variables*. Sage Publications, 1997
- Vittinghoff, E, et al. *Regression methods in biostatistics.* Springer, 2005