Poisson regression

Dae-Jin Lee
dlee@bcamath.org

Basque Center for Applied Mathematics

http://idaejin.github.io/bcam-courses/

Modeling count data

Introduction

- Response variable is a **count**
- **e.g.:** *number of cases of a disease, Number of infected plants, firms going bankrupt etc ...*
- When exploratory variables are categorical (i.e. you have a contingency table with counts in the cells), the models are usually called **'log-linear models'**
- If they are numerical/continuous, convention is to call them **poisson regression**
- The Poisson distribution describes the count of the number of random events within a fixed interval of time or space with a known average rate.

Modeling count data

Introduction

- Poisson regression may also be appropriate for rate data, where the rate is a count of events occurring to a particular unit of observation, divided by some measure of that unit's **exposure**.
- **e.g.:** *The number of deaths may be affected by the underlying population at risk, in demography death rates in geographic areas are modelled as the count of deaths divided by person-years, etc ...*
- In Poisson regression this is handled as an **offset**.
- Then, we include the exposure (exp) in the denominator and form a rate such as:

$$Y/exp = rate \Rightarrow Y = rate \times exp$$

Modeling count data

Introduction

- Linear regression methods (constant variance, normal errors) are not appropriate for count data for 5 main reasons:

Modeling count data

Introduction

- Linear regression methods (constant variance, normal errors) are not appropriate for count data for 5 main reasons:
  1. the linear model might lead to the prediction of negative counts
  2. the variance of the response variable is likely to increase with the mean
  3. the errors will not be normally distributed
  4. zeros are difficult to handle in transformations
  5. some distributions (e.g. log-normal or gamma) do not allow zeros

Poisson regression

Poisson distribution

- The Poisson distribution is widely used for the description of count data that refer to cases where we know how many times something happened, but we have no way of knowing how many times it did not happen.
- Recall that with the binomial distribution we know how many times something did not happen as well as how often it did happen.

Poisson regression

Poisson distribution

- The Poisson distribution is widely used for the description of count data that refer to cases where we know how many times something happened, but we have no way of knowing how many times it did not happen.
- Recall that with the binomial distribution we know how many times something did not happen as well as how often it did happen.

Definition

- a discrete random variable Y is ditributed as Poisson with parameter $\mu > 0$, if, for $k = 0, 1, 2, ...$, the probability mass function of Y is given by

$$\Pr(Y = k) = \frac{\mu^k e^{-\mu}}{k!}$$

- $\mathbb{E}(Y) = \mathbb{V}\text{ar}(Y) = \mu$

Poisson regression

Components of a GLM

- The components of a GLM for a count response are:
  1. **Random Component:** Poisson distribution and $\mathbb{E}(Y) = \mu$
  2. **Systematic component:** predictors/explanatory variables X
  3. **Link function:**
     - **Identity link:** which would give us $\mu = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ (but then we can get $\mu < 0$)
     - **Log link:** is the most common and canonical link

$$\log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

- Since the $\log(\mu)$ is a linear function of the X's and μ is a multiplicative function of X:

$$\begin{aligned} \mu &= e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k} \\ &= e^{\beta_0} e^{\beta_1 x_1} ... e^{\beta_p x_k} = \exp(\boldsymbol{X\beta}) \end{aligned}$$

Poisson regression

Components of a GLM

- The components of a GLM for a count response are:
  1. **Random Component:** Poisson distribution and $\mathbb{E}(Y) = \mu$
  2. **Systematic component:** predictors/explanatory variables X
  3. **Link function:**
     - **Identity link:** which would give us $\mu = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ (but then we can get $\mu < 0$)
     - **Log link:** is the most common and canonical link

$$\log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

- Since the $\log(\mu)$ is a linear function of the X's and μ is a multiplicative function of X:

$$\begin{aligned} \mu &= e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k} \\ &= e^{\beta_0} e^{\beta_1 x_1} ... e^{\beta_p x_k} = \exp(\boldsymbol{X\beta}) \end{aligned}$$

Poisson regression

log link and interpretation

$$\log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

- **What does it mean for μ? How do you interpret β?**

Poisson regression

log link and interpretation

$$\log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

- **What does it mean for μ? How do you interpret β?**
- The $\log$ is a transformation of the mean such that $\log(\mu)$ has a linear relationship with the predictors, e_i^{β} represents a multiplier effect of the i^{th} predictor on the mean, such that a $1-$unit increase on X_i has a multiplicative effect of e^{β_i} on μ.

Poisson regression

log link and interpretation

$$\log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

- **What does it mean for μ? How do you interpret β?**
- The $\log$ is a transformation of the mean such that $\log(\mu)$ has a linear relationship with the predictors, e_i^{β} represents a multiplier effect of the i^{th} predictor on the mean, such that a $1-$unit increase on X_i has a multiplicative effect of e^{β_i} on μ.
- Suppose of a single predictor x and 2 values of x (i.e. x_1 and x_2), with a difference of 1, i.e. $x_1 = 10$ and $x_1 = 11$, where $\mu_1 = \mathbb{E}(Y|x = 10)$ and $\mu_2 = \mathbb{E}(Y|x = 11)$
  - If $\beta = 0$, then $e^0 = 1$ and μ_2 is the same as μ_1. That is μ is not related to x.
  - If $\beta > 0$, $e^{\beta} > 1$ and μ_2 is e^{β} times **larger** than μ_1.
  - If $\beta < 0$, $e^{\beta} < 1$ and μ_2 is e^{β} times **smaller** than μ_1.

bcam basque center for applied mathematics

Poisson regression

Estimation, Deviances, Hypothesis testing, Goodness-of-Fit and residuals

- Once we are in a GLM setting with a $\log$ link, the parameters are estimated by Maximum Likelihood and Iterative Re-weighted Least Squares.
- The **Deviance** (discrepancy measure between observed and fitted values) for Poisson responses takes the form:

$$D = 2\sum\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\}$$

  for large samples $D \sim \chi^2_{n-p}$.
- **Goodness-of-Fit** measure is the Pearson's chi-squared statistic $\chi^2_p = \sum \frac{(y_i - \mu_i)^2}{\hat{\mu}_i}$
- **Likelihood ratio tests** can easily be constructed in terms of deviances, just as we did in logistic regression models.
- We will use **standardized residuals** $r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \sim \mathcal{N}(0, 1)$

bcam basque center for applied mathematics

Poisson regression

Case study: Agricultural experiment of species richness

- A longterm agricultural experiment had 90 grassland plots, each $25m \times 25m$, differing in `biomass`, soil `pH` and `species` richness (the count of species in the whole plot).

Poisson regression

Case study: Agricultural experiment of species richness

- A longterm agricultural experiment had 90 grassland plots, each $25m \times 25m$, differing in `biomass`, soil `pH` and `species` richness (the count of species in the whole plot).
- It is well known that species richness declines with increasing biomass, but the question addressed here is whether the slope of that relationship differs with soil pH. The plots were classified according to a 3-level factor as `high`, `medium` or `low` pH with 30 plots in each level.

Poisson regression

Case study: Agricultural experiment of species richness

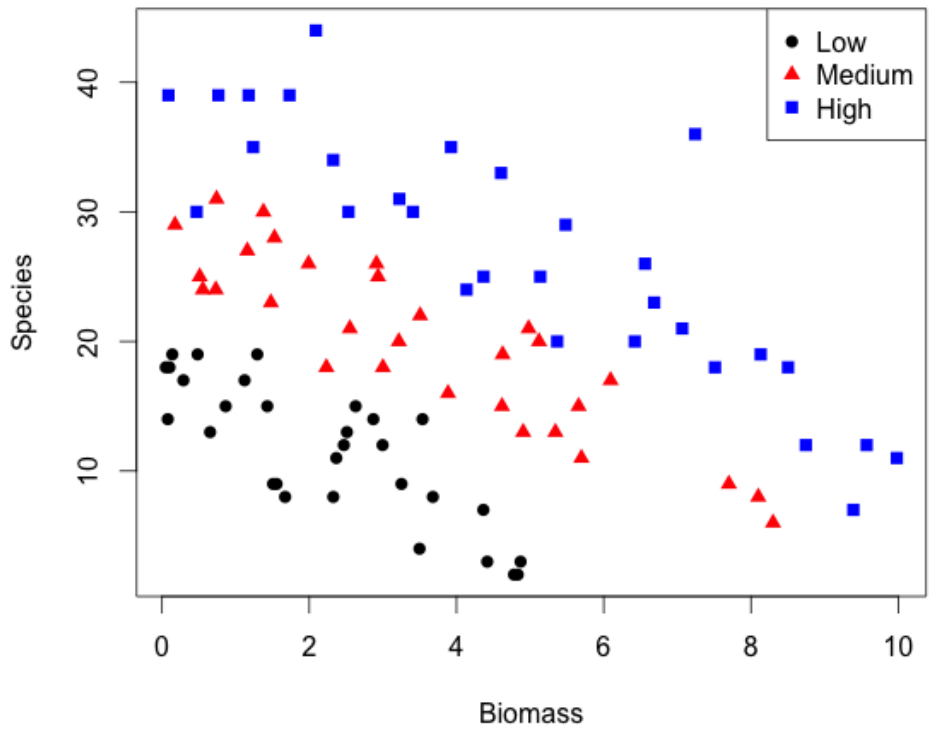
- A longterm agricultural experiment had 90 grassland plots, each $25m \times 25m$, differing in `biomass`, soil `pH` and `species` richness (the count of species in the whole plot).
- It is well known that species richness declines with increasing biomass, but the question addressed here is whether the slope of that relationship differs with soil pH. The plots were classified according to a 3-level factor as `high`, `medium` or `low` pH with 30 plots in each level.

See `Poisreg.R`

```
> species<-read.table("data/species.txt",header=TRUE)
> names(species)
```

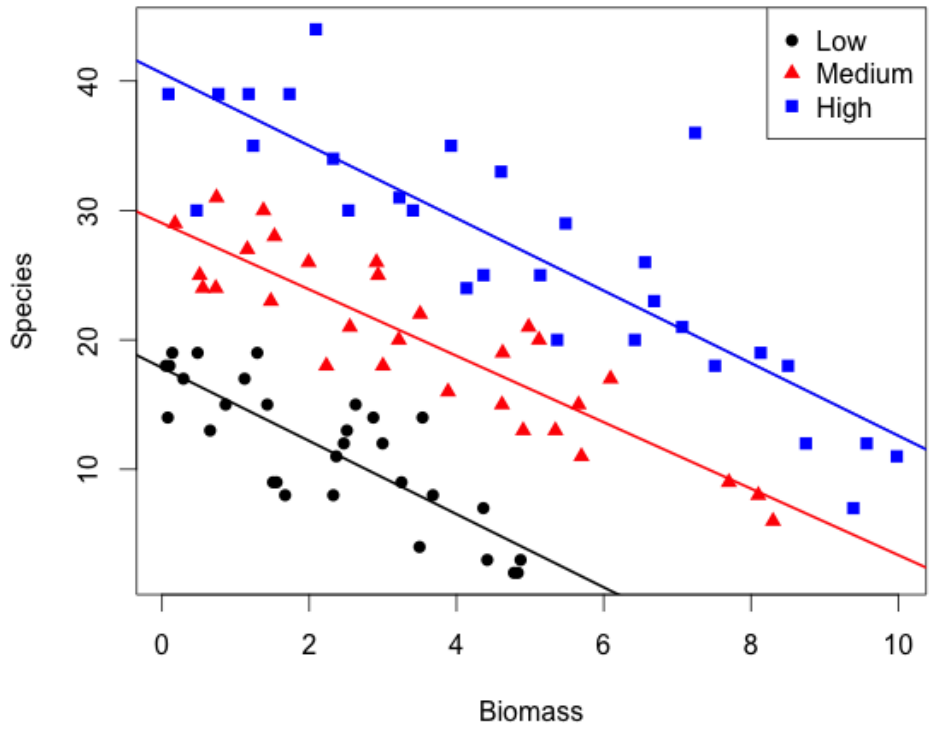
Poisson regression

We can plot the data for each level of soil `pH` and fit a `lm`



Poisson regression

We can plot the data for each level of soil `pH` and fit a `lm`



There is a clear difference in mean species richness with declining soil pH, but there is little evidence of any substantial difference in the slope of the relationship between species richness and biomass on soils of differing pH.

Poisson regression

Case study: Agricultural experiment of species richness

- We start fitting each predictor separately, are they significant?

Poisson regression

Case study: Agricultural experiment of species richness

- We start fitting each predictor separately, are they significant?

```
> m0=glm(Species~Biomass,family=poisson,data=species)
> anova(m0,test="Chisq")
> m1=glm(Species~pH,family=poisson,data=species)
> anova(m1,test="Chisq")
```

Poisson regression

Case study: Agricultural experiment of species richness

- We start fitting each predictor separately, are they significant?

```
> m0=glm(Species~Biomass,family=poisson,data=species)
> anova(m0,test="Chisq")
> m1=glm(Species~pH,family=poisson,data=species)
> anova(m1,test="Chisq")
```

- Include the interaction and test its significance

Poisson regression

Case study: Agricultural experiment of species richness

- We start fitting each predictor separately, are they significant?

```
> m0=glm(Species~Biomass,family=poisson,data=species)
> anova(m0,test="Chisq")
> m1=glm(Species~pH,family=poisson,data=species)
> anova(m1,test="Chisq")
```

- Include the interaction and test its significance *The effect of biomass on species richness depends on soil pH, and the strength of the relationship varies with soil pH*

Poisson regression

Case study: Agricultural experiment of species richness

- We start fitting each predictor separately, are they significant?

```
> m0=glm(Species~Biomass,family=poisson,data=species)
> anova(m0,test="Chisq")
> m1=glm(Species~pH,family=poisson,data=species)
> anova(m1,test="Chisq")
```

- Include the interaction and test its significance *The effect of biomass on species richness depends on soil pH, and the strength of the relationship varies with soil pH*

```
> m2 <- glm(Species~Biomass+pH,family=poisson,data=species)
> m3 <- glm(Species~Biomass*pH,family=poisson,data=species)
> anova(m2,m3,test="Chisq")
```

Hence assume that the slopes were the same is completely wrong: in fact, the slopes are very significantly different ($p < 0.00035$).

```
> summary(m3)



Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.95255    0.08240  35.833  < 2e-16 ***
Biomass         -0.26216    0.03803  -6.893 5.47e-12 ***
pHMedium         0.48411    0.10723   4.515 6.34e-06 ***
pHHigh           0.81557    0.10284   7.931 2.18e-15 ***
Biomass:pHMedium 0.12314    0.04270   2.884 0.003927 **
Biomass:pHHigh   0.15503    0.04003   3.873 0.000108 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: 514.39

Number of Fisher Scoring iterations: 4
```

Poisson regression

Case study: Agricultural experiment of species richness (cont.)

- The fitted model is

$$\log(\texttt{Species}) = 2{,}95 - 0{,}262 \times \texttt{Biomass} + 0{,}484\texttt{pH}_{medium} + 0{,}815 \times \texttt{pH}_{high} + 0{,}123 \times \texttt{pH}_{medium} \times \texttt{Biomass} + 0{,}155 \times \texttt{pH}_{high} \times \texttt{Biomass}$$

Poisson regression

Case study: Agricultural experiment of species richness (cont.)

- The fitted model is

$$\log(\texttt{Species}) = 2{,}95 - 0{,}262 \times \texttt{Biomass} + 0{,}484\texttt{pH}_{medium} + 0{,}815 \times \texttt{pH}_{high} + 0{,}123 \times \texttt{pH}_{medium} \times \texttt{Biomass} + 0{,}155 \times \texttt{pH}_{high} \times \texttt{Biomass}$$

- **Interpretation of the parameters:**
  - A 1-unit change in the Biomass implies a number of species of $\exp(-0{,}262) = 0{,}769$ times less in soil with **low** pH.

Poisson regression

Case study: Agricultural experiment of species richness (cont.)

- The fitted model is

$$\log(\texttt{Species}) = 2{,}95 - 0{,}262 \times \texttt{Biomass} + 0{,}484\texttt{pH}_{medium} + 0{,}815 \times \texttt{pH}_{high} + 0{,}123 \times \texttt{pH}_{medium} \times \texttt{Biomass} + 0{,}155 \times \texttt{pH}_{high} \times \texttt{Biomass}$$

- **Interpretation of the parameters:**
  - A 1-unit change in the Biomass implies a number of species of $\exp(-0{,}262) = 0{,}769$ times less in soil with **low** pH.
  - A 1-unit change in the Biomass implies a number of species of $\exp(-0{,}262 + 0{,}123) = 0{,}870$ times less in soil with **medium** pH.

Poisson regression

Case study: Agricultural experiment of species richness (cont.)

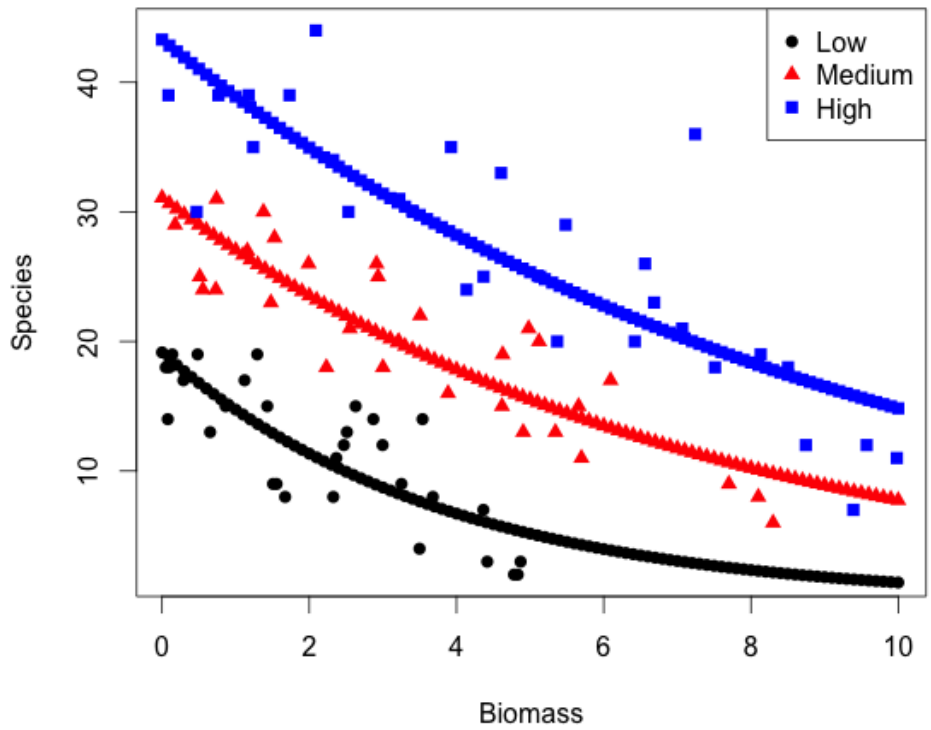
- The fitted model is

$$\log(\texttt{Species}) = 2{,}95 - 0{,}262 \times \texttt{Biomass} + 0{,}484\texttt{pH}_{medium} + 0{,}815 \times \texttt{pH}_{high} + 0{,}123 \times \texttt{pH}_{medium} \times \texttt{Biomass} + 0{,}155 \times \texttt{pH}_{high} \times \texttt{Biomass}$$

- **Interpretation of the parameters:**
  - A 1-unit change in the Biomass implies a number of species of $\exp(-0{,}262) = 0{,}769$ times less in soil with **low** pH.
  - A 1-unit change in the Biomass implies a number of species of $\exp(-0{,}262 + 0{,}123) = 0{,}870$ times less in soil with **medium** pH.
  - A 1-unit change in the Biomass implies a number of species of $\exp(-0{,}262 + 0{,}155) = 0{,}898$ times less in soil with **high** pH.

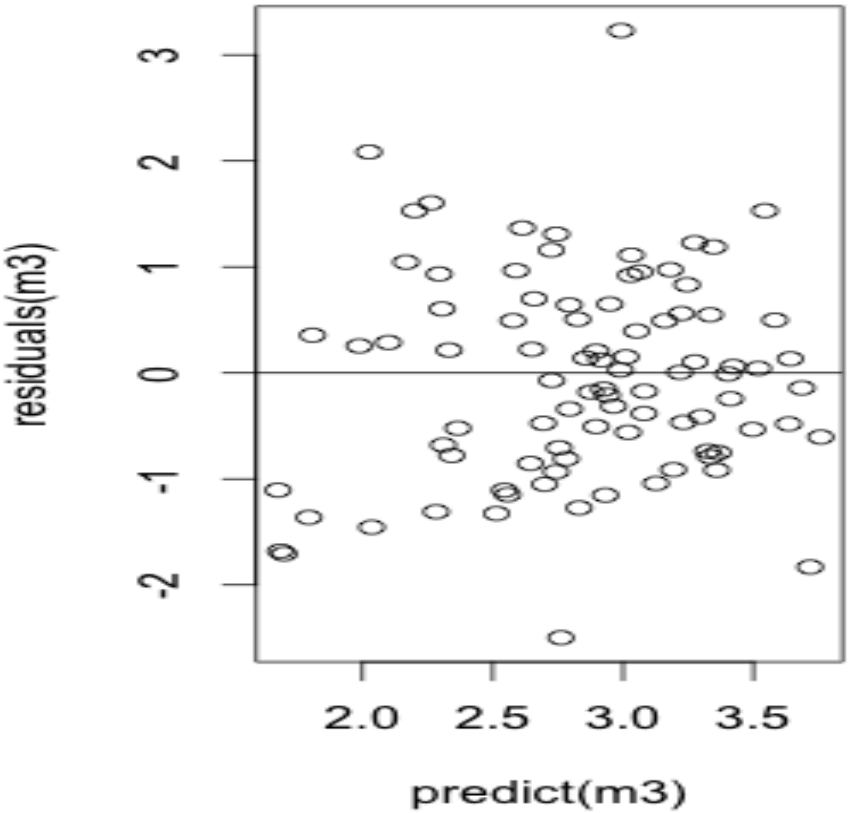
Poisson regression

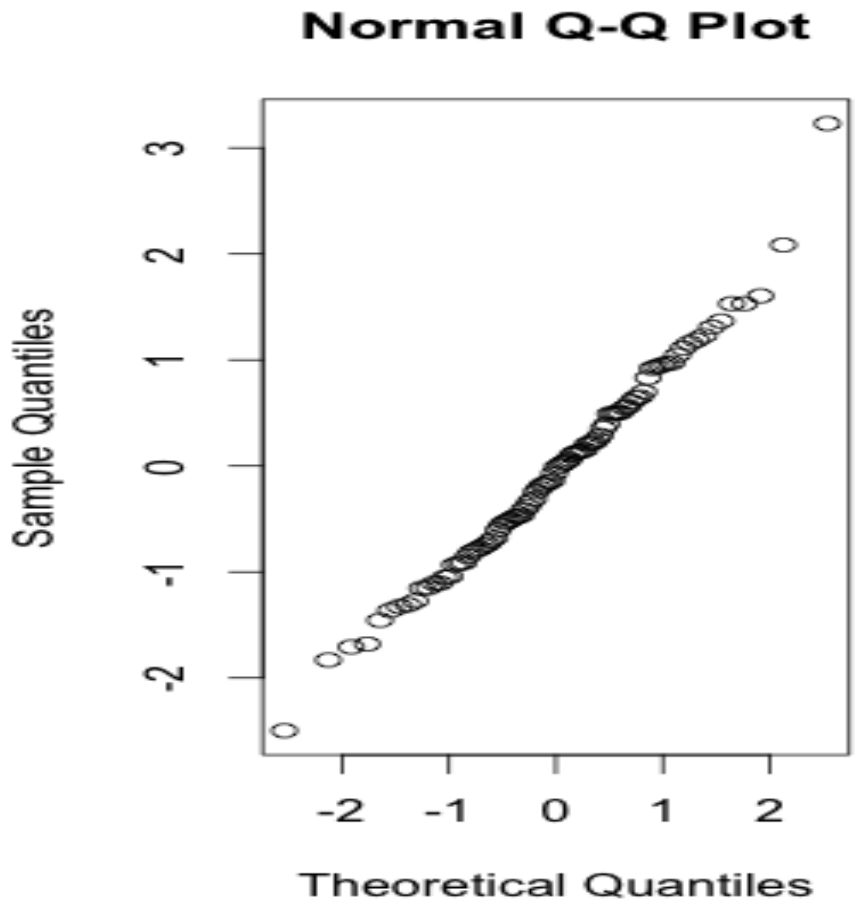
We can plot the data for each level of soil `pH`



Poisson regression

And check the fitted VS residuals plot and the QQ-plot of the residuals





Poisson regression

Overdispersion

- In Poisson regression, it is assumed that $\mathbb{V}\mathrm{ar}(Y) = \mu$
- Sometimes the data do not hold this assumption, this yields to wrong estimation of the standard errors in the parameters.
- **Overdispersion** is present if $\mathbb{V}\mathrm{ar}(Y) > \mu$
- A simple option to account for this phenomena is to include a **dispersion parameter** ϕ, such that $\mathbb{V}\mathrm{ar}(Y) = \phi\mu$, then when $\phi = 1$ we are in the Poisson case.
- The **dispersion parameter** ϕ can be estimated as

$$\hat{\phi} = \frac{\text{Deviance}}{n - p}$$

- In our example, we had that $\text{Deviance} = 83{,}2$ and $n - p = 84$, then $\phi \approx 1$.

Poisson regression

Overdispersion (cont.)

- In R, we can specify the `family=quasipoisson`

```
> m4 <- glm(Species~Biomass*pH,family=quasipoisson,data=species)
```

Poisson regression

Overdispersion (cont.)

- In R, we can specify the `family=quasipoisson`

```
> m4 <- glm(Species~Biomass*pH,family=quasipoisson,data=species)
```

- Or use the function `dispersiontest` in `library(AER)`

```
> library(AER)
> dispersiontest(m3)

Overdispersion test

data:  m3
z = -0.38285, p-value = 0.6491
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
 0.9369549
```

- The test is

$$H_0: \quad \phi^* = 1$$
$$H_1: \quad \phi^* \neq 1$$

Poisson regression

Other alternatives for overdispersion

- One common cause of overdispersion is excess zeros, which in turn are generated by an additional data generating process. In this situation, **zero-inflated Poisson** model should be considered.
- **Negative Binomial regression:** it can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the overdispersion. **E.g.:** `glm.nb` function in `library(MASS)`

Poisson regression

for incidence rates

- GLM's for rates

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- **Random component:** Response Y has a Poisson distribution, and t is index of the time or space; more specifically the expected value of rate Y/t, is $\mathbb{E}(Y/t) = \mu$ that is $\mathbb{E}(Y) = \mu t$.
- **Systematic component:** Any set of $X = (x_1, x_2, ..., x_k)$ can be explanatory variables. For now let's focus on a single variable x.
- **Link function:** Log of rate: $\log(Y/t)$

Poisson regression

for incidence rates (cont.)

- **Poisson loglinear regression model** for the expected rate of the occurrence of event is:

$$log(\mu/t) = \alpha + \beta x$$

This can be rearranged to:

$$\begin{aligned} \log(\mu) - \log(t) &= \alpha + \beta x \\ \log(\mu) &= \alpha + \beta x + \log(t) \end{aligned}$$

- The term $\log(t)$ is referred to as an **offset**. It is an adjustment term and a group of observations may have the same offset, or each individual may have a different value of t.
- $\log(t)$ is an observation and it will change the value of estimated counts:

$$\mu = \exp(\alpha + \beta x + \log(t)) = t \times \exp(\alpha)\exp(\beta x)$$

This means that mean count is proportional to t.

- Note that the interpretation of parameter estimates, α and β will stay the same as for the model of counts; you just need to multiply the expected counts by t.

Poisson regression

for incidence rates (cont.)

- In poisson regression we also have **relative risk (RR)** interpretations, i.e. the estimated effect of an explanatory variable is multiplicative on the rate, and thus leads to a risk ratio or relative risk.
- This is similar to logistic regression, but then the effect of an explanatory variable is multiplicative on the odds and thus leads to an odds ratio.

Poisson regression

for incidence rates (cont.)

- In poisson regression we also have **relative risk (RR)** interpretations, i.e. the estimated effect of an explanatory variable is multiplicative on the rate, and thus leads to a risk ratio or relative risk.
- This is similar to logistic regression, but then the effect of an explanatory variable is multiplicative on the odds and thus leads to an odds ratio.
- Let λ_0 be the incidence rate when x_0 and λ_1 when $X = x_0 + 1$, then

$$\begin{aligned} \log\left(\frac{\lambda_1}{\lambda_0}\right) &= \log(\lambda_1) - \log(\lambda_0) = \beta \\ \text{RR} &= \frac{\lambda_1}{\lambda_0} = \exp(\beta) \end{aligned}$$

i.e. $\exp(\beta)$ is the relative risk between the population with $X = x_0$ and $X = x_1$, and $\exp(\alpha)$ is the incidence rate when $X = 0$

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- This data set contains counts of incident lung cancer cases and population size in four neighbouring Danish cities by age group.
- The variables are:

`city`	a factor with levels `Fredericia`, `Horsens`, `Kolding`, and `Vejl`
`age`	a factor with levels `40-54`, `55-59`, `60-64`, `65-69`, `70-74`, and `75+`.
`pop`	the number of inhabitants.
`cases`	the number of lung cancer cases.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- This data set contains counts of incident lung cancer cases and population size in four neighbouring Danish cities by age group.
- The variables are:

`city`	a factor with levels `Fredericia`, `Horsens`, `Kolding`, and `Vejl`
`age`	a factor with levels `40-54`, `55-59`, `60-64`, `65-69`, `70-74`, and `75+`.
`pop`	the number of inhabitants.
`cases`	the number of lung cancer cases.

- **How does the expected number of lung cancer counts vary by age?**

See `Poisreg.r`

```
> lung <- read.table("data/lung.txt",header=TRUE)
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

See Poisreg.r

```
> head(lung)

        city   age  pop cases
1 Fredericia 40-54 3059    11
2    Horsens 40-54 2879    13
3    Kolding 40-54 3142     4
4      Vejle 40-54 2520     5
5 Fredericia 55-59  800    11
6    Horsens 55-59 1083     6
```

- The incidence rate is denoted by $\lambda = Y/E$, then for people aged 40-54 years, in `Fredericia`, $Y = 11$ and $E = 3059$, and then

$$\lambda = \frac{11}{3059} = 0{,}003595946$$

or $3{,}595946$ for each 1000 inhab.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

See Poisreg.r

```
> head(lung)

        city   age  pop cases
1 Fredericia 40-54 3059    11
2    Horsens 40-54 2879    13
3    Kolding 40-54 3142     4
4      Vejle 40-54 2520     5
5 Fredericia 55-59  800    11
6    Horsens 55-59 1083     6
```

- The incidence rate is denoted by $\lambda = Y/E$, then for people aged 40-54 years, in Fredericia, $Y = 11$ and $E = 3059$, and then

$$\lambda = \frac{11}{3059} = 0{,}003595946$$

or $3{,}595946$ for each 1000 inhab.

```
> boxplot(cases~age,data=lung,col="bisque")
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We start considering a model with age as covariate

$$\log(\lambda_i) = \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) + \\ + \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i)$$

  where $I(\cdot)$ is a indicator (1 if TRUE, 0 otherwise) for each range of age, with Age40-45 is used as baseline.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We start considering a model with age as covariate

$$\log(\lambda_i) = \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) + \\ + \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i)$$

  where $I(\cdot)$ is a indicator (1 if TRUE, 0 otherwise) for each range of age, with Age40-45 is used as baseline.

```
> lungmod1 <- glm(cases ~ age, family=poisson, data=lung)
```

```
> summary(lungmod1)


Call:
glm(formula = cases ~ age, family = poisson, data = lung)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4661  -0.4488  -0.1532   0.7773   1.5242

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.11021    0.17408  12.122   <2e-16 ***
age55-59    -0.03077    0.24810  -0.124    0.901
age60-64     0.26469    0.23143   1.144    0.253
age65-69     0.31015    0.22918   1.353    0.176
age70-74     0.19237    0.23516   0.818    0.413
age75+      -0.06252    0.25012  -0.250    0.803
---
Signif. codes:  0
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We start considering a model with age as covariate

$$\log(\lambda_i) = 2{,}11021 - 0{,}03077 \times I(\text{Age55-59}_i) + 0{,}26469 \times I(\text{Age60-64}_i) + \\ + 0{,}31015 \times I(\text{Age65-69}_i) + 0{,}19237(\text{Age70-74}_i) - 0{,}06252 \times I(\text{Age}{>}74_i)$$

- **Interpretation:**
  - $\exp(2{,}11021) = 8{,}24$ is the expected count of cancer cases among individuals aged $40 - 54$
  - $\exp(2{,}11021 - 0{,}03077) = 8{,}00$ is the expected count of cancer cases among individuals aged $55 - 59$
  - $\exp(-0{,}0377) = 0{,}97$ is the ratio of the expected counts comparing the $55 - 59$ aged group to the baseline group of age $40 - 54$. $\exp(\hat{\beta}_1)$ is also the relative rate.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We start considering a model with age as covariate

$$\log(\lambda_i) = 2{,}11021 - 0{,}03077 \times I(\text{Age55-59}_i) + 0{,}26469 \times I(\text{Age60-64}_i) + \\ + 0{,}31015 \times I(\text{Age65-69}_i) + 0{,}19237(\text{Age70-74}_i) - 0{,}06252 \times I(\text{Age}{>}74_i)$$

- **Interpretation:**
  - $\exp(2{,}11021) = 8{,}24$ is the expected count of cancer cases among individuals aged $40 - 54$
  - $\exp(2{,}11021 - 0{,}03077) = 8{,}00$ is the expected count of cancer cases among individuals aged $55 - 59$
  - $\exp(-0{,}0377) = 0{,}97$ is the ratio of the expected counts comparing the $55 - 59$ aged group to the baseline group of age $40 - 54$. $\exp(\hat{\beta}_1)$ is also the relative rate.

```
> lungmod1 <- glm(cases ~ age, family=poisson, data=lung)
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

If we calculate the **CI**'s for all ages we find that all contain 0, is there any association between cancer and age?

```
> confint(lungmod1)

                 2.5 %    97.5 %
(Intercept)  1.7484571 2.4330407
age55-59    -0.5200577 0.4572927
age60-64    -0.1863226 0.7247913
age65-69    -0.1357438 0.7664682
age70-74    -0.2671849 0.6587654
age75+      -0.5564824 0.4289194
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

We can perform a **Likelihood Ratio Test**:

```
> lungmod0 <- glm(cases ~ 1, family=poisson, data=lung)
> anova(lungmod0,lungmod1,test="Chisq")

Analysis of Deviance Table

Model 1: cases ~ 1
Model 2: cases ~ age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        23     27.704
2        18     22.756  5   4.9478   0.4223
```

and hence, we do not reject the hyphotesis of all $\beta_i = 0$, for $i = 1, ..., 5$.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We have considered the counts of lung cancer cases.
- Can we improve the analysis?

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We have considered the counts of lung cancer cases.
- Can we improve the analysis?
- Each city and age group has a different population size.
- So far, we have modeled expected counts for each population group, within the 4 year period of time, i.e., intrinsically rate $=$ counts/'4 years'

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We have considered the counts of lung cancer cases.
- Can we improve the analysis?
- Each city and age group has a different population size.
- So far, we have modeled expected counts for each population group, within the 4 year period of time, i.e., intrinsically rate $=$ counts/'4 years'
- It may be of more interest to know the rate per person, per 4 period of observation. We are interested in

$$r_i = \frac{\lambda_i}{\text{Pop}_i} = \frac{\mathbb{E}(\text{count}_i)}{\text{Pop}_i}$$

and model it by a log-linear model

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- We have considered the counts of lung cancer cases.
- Can we improve the analysis?
- Each city and age group has a different population size.
- So far, we have modeled expected counts for each population group, within the 4 year period of time, i.e., intrinsically rate $=$ counts/'4 years'
- It may be of more interest to know the rate per person, per 4 period of observation. We are interested in

$$r_i = \frac{\lambda_i}{\text{Pop}_i} = \frac{\mathbb{E}(\text{count}_i)}{\text{Pop}_i}$$

  and model it by a log-linear model
- Then, our model is

$$Y_i \sim \mathcal{P}\text{ois}(\lambda_i) = \mathcal{P}\text{ois}(r_i \times \text{Pop}_i)$$

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

- On a log-scale, our model is:

$$\log\left(\frac{\lambda_i}{\text{Pop}_i}\right) = \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) + \\ + \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age>74}_i)$$

  Exponentiating we have:

$$\left(\frac{\lambda_i}{\text{Pop}_i}\right) = \exp\left(\beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) + \right. \\ \left. + \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age>74}_i)\right)$$

- All counts are restricted to the same period of $1968 - 1971$, the values of λ_i are rates 'per 4-years'
- To obtain an easier interpretation of the rates we can divide by 4 to get rate per person-year and multiply by $10{,}000$ to get a rate per $10,000$ person-years, i.e. divide by 2500

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

This can be easily done by

```
> lungmod2 <- glm(cases ~ age + offset(log(pop/2500)),
+                 family=poisson, data=lung)
```

- `log(pop/2500)` is the offset
- The offset accounts for the population size, which could vary by age, region, etc ...
- It gives a convenient way to model rates per person-years, instead of modeling the raw counts.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

```
> summary(lungmod2)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9618     0.1741  11.270  < 2e-16 ***
age55-59      1.0823     0.2481   4.363 1.29e-05 ***
age60-64      1.5017     0.2314   6.489 8.66e-11 ***
age65-69      1.7503     0.2292   7.637 2.22e-14 ***
age70-74      1.8472     0.2352   7.855 4.00e-15 ***
age75+        1.4083     0.2501   5.630 1.80e-08 ***
---
    Null deviance: 129.908  on 23  degrees of freedom
Residual deviance:  28.307  on 18  degrees of freedom
AIC: 136.69
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971

Confidence intervals for the parameters can also be obtained as:

```
> confint(lungmod2)

                2.5 %   97.5 %
(Intercept) 1.6000371 2.284621
age55-59    0.5930558 1.570406
age60-64    1.0506603 1.961774
age65-69    1.3043879 2.206600
age70-74    1.3876652 2.313616
age75+      0.9143186 1.899720
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- The inclusion of the offset, implies that the interpretation of the coefficients should be done in terms of $\log(\lambda_i) - \text{offset}_i$
- In our example, λ_i is the expected number of cases observed in a particular age group, within a 4 year period of time.
- Hence in our case, with an offset of `log(pop/2500)`, we should think of the outcome as log rate per 10,000 person years.
- Then:
  - β_0 is the log rate of cancer cases per $10,000$ person years in the age group of $40 - 54$ (baseline)

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- The inclusion of the offset, implies that the interpretation of the coefficients should be done in terms of $\log(\lambda_i) - \text{offset}_i$
- In our example, λ_i is the expected number of cases observed in a particular age group, within a 4 year period of time.
- Hence in our case, with an offset of `log(pop/2500)`, we should think of the outcome as log rate per 10,000 person years.
- Then:
  - β_0 is the log rate of cancer cases per $10,000$ person years in the age group of $40-54$ (baseline)
  - β_1 is the log relative rate of cancer cases per $10,000$ person years comparing the age group of $50-59$ to the baseline age group $40-54$

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- The inclusion of the offset, implies that the interpretation of the coefficients should be done in terms of $\log(\lambda_i) - \text{offset}_i$
- In our example, λ_i is the expected number of cases observed in a particular age group, within a 4 year period of time.
- Hence in our case, with an offset of `log(pop/2500)`, we should think of the outcome as log rate per 10,000 person years.
- Then:
  - β_0 is the log rate of cancer cases per $10,000$ person years in the age group of $40-54$ (baseline)
  - β_1 is the log relative rate of cancer cases per $10,000$ person years comparing the age group of $50-59$ to the baseline age group $40-54$
  - β_2 is the log relative rate of cancer cases per $10,000$ person years comparing the age group of $60-64$ to the baseline age group $40-54$

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- Note that, including the offset, the regression coefficients are significant in the expected counts per year at age group compared to the baseline group.
- Now, with the offset, we are looking for differences in the expected counts per person-year, across age groups.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- Note that, including the offset, the regression coefficients are significant in the expected counts per year at age group compared to the baseline group.
- Now, with the offset, we are looking for differences in the expected counts per person-year, across age groups.
- **Likelihood Ratio Test** to test the effect of age

```
> lungmod3<-glm(cases ~ 1, family = poisson, data = lung,
+               offset = (log(pop/2500)))
> anova(lungmod3,lungmod2,test="Chisq")

Analysis of Deviance Table

Model 1: cases ~ 1
Model 2: cases ~ age + offset(log(pop/2500))
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        23    129.908
2        18     28.307  5    101.6 < 2.2e-16 ***
---
Signif. codes:  0
```

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- **What are the difference between model with and without the offset?**

Without offset	**With offset**
Do not reject H_0	Reject H_0
Expected cases across the age groups*	Expected cases per person year, across the age groups**

* There is no difference because the lower population `Age75+` might be counterbalanced by high rate of cases with increasing age

** The offset let us compare the rate of cancer among those who are alive, i.e. taking into account the number of people within cohort of interest.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- **Given the model including the offset what can we say about the rate of lung cancer in Denmark?**

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- **Given the model including the offset what can we say about the rate of lung cancer in Denmark?**
- Predicted **log expected count** of cancer from $1968-1971$ among $40-54$ year old group of age:

$$\log(\lambda_i) = \log\left(\frac{\text{Population Age 40-54}}{2500}\right) + \hat{\beta}_0$$
$$= \log\left(\frac{11600}{2500}\right) + 1{,}96 = 3{,}49$$

- where predicted rate of cancer per $10,000$ person years among $40-54$ year olds is $\log(\lambda_i) = \hat{\beta}_0 = 1{,}96$
- Taking exponentials we have the predicted expected count of cancer cases $1968-1971$ among age $40-54$

$$\lambda_i = \exp(3{,}49) = 32{,}9$$

per $10,000$ person years among $40-54$, is $\exp(\hat{\beta}_0) = 7{,}11$

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

per 10,000 person years among 40-54, is $\exp(\hat{\beta}_0) = 7{,}11$Then, based on this model, the prediction of cases of lung cancer per $10,0000$ people aged $40 - 54$ year old in Denmark, per year is $7,11$.

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

per 10,000 person years among 40-54, is $\exp(\hat{\beta}_0) = 7{,}11$Then, based on this model, the prediction of cases of lung cancer per $10,0000$ people aged $40 - 54$ year old in Denmark, per year is $7,11$.**And among $55 - 59$ age group?**

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

per 10,000 person years among 40-54, is $\exp(\hat{\beta}_0) = 7{,}11$Then, based on this model, the prediction of cases of lung cancer per $10,0000$ people aged $40 - 54$ year old in Denmark, per year is $7, 11$.**And among $55 - 59$ age group?**

- Log expected count is

$$\log(\lambda_i) = \log\left(\frac{\text{Population Age 55-59}}{2500}\right) + \hat{\beta}_0 + \hat{\beta}_1$$
$$= \log\left(\frac{3811}{2500}\right) + 1{,}96 + 1{,}08 = 3{,}46$$

and the predicted log rate of cancer per $10,000$ person years among $55 - 59$ years old $\hat{\beta}_0 + \hat{\beta}_1 = 1{,}96 + 1{,}08 = 3{,}04$

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- **What is exp($\hat{\beta}_1$)?**

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- **What is $\exp(\hat{\beta}_1)$?**
- $\exp(\hat{\beta}_1) = \exp(1,08) = 2,94$ is the relative rate of cancer cases per $10,000$ person years comparing $55-59$ and $40-54$ groups, i.e.

$$\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1)}{\exp(\hat{\beta}_0)} = \frac{20,9}{7,09} = 2,94$$

Poisson regression

Case study: Lung cancer incidence in Denmark 1968–1971 (cont.)

- **What is exp($\hat{\beta}_1$)?**
- $\exp(\hat{\beta}_1) = \exp(1,08) = 2,94$ is the relative rate of cancer cases per $10,000$ person years comparing $55-59$ and $40-54$ groups, i.e.

$$\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1)}{\exp(\hat{\beta}_0)} = \frac{20,9}{7,09} = 2,94$$

```
> exp(coef(lungmod2))

(Intercept)    age55-59    age60-64    age65-69    age70-74      age75+
   7.112069    2.951584    4.489204    5.756252    6.342177    4.088919
```

There is a increasing trend in relative rates compared to the baseline age group $40-54$ except for `Age75+`

Poisson regression

Some conclusions

- The offset allows for including the change the denominator or units of a rate
- When considering poisson regression on incidence rates, the model without an offset does not make much sense, and it fails to hold the Poisson assumptions
- We should try to use an offset if we suspect that the underlying population sizes differ for each of the observed counts