Section 3.3 Generalized Linear Models For Count Data

Outline

- Review of Poisson Distributions
- GLMs for Poisson Response Data
- Models for Rates
- Overdispersion and Negative Binomial Regression

Review of Poisson Distributions

A random variable Y has a Poisson distribution with parameter $\lambda > 0$ if

$$\mathrm{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

denoted as

$$Y \sim \text{Poisson}(\lambda).$$

One can show that

$$\mathbb{E}[Y] = \lambda, \quad \mathrm{Var}(Y) = \lambda \Rightarrow \text{SD}(Y) = \sqrt{\lambda}.$$

Poisson Approximation to Binomial

If $Y \sim$ binomial(n, p) with **huge** n and **tiny** p such that np moderate, then

$$Y \text{ approx.} \sim \text{Poisson}(np).$$

The following shows the values of $\mathrm{P}(Y = k)$, $k = 0, 1, 2, \ldots, 8$ for

$$Y \sim \text{Binomial}(n = 50,\ p = 0.03), \text{ and}$$
$$Y \sim \text{Poisson}(\lambda = 50 \times 0.03 = 1.5).$$

```
> dbinom(0:5, size=50, p=0.03)           # Binomial(n=50, p=0.03)
[1] 0.21806538 0.33721450 0.25551820 0.12644200 0.04594928 0.01307423

> dpois(0:5, lambda = 50*0.03)          # Poisson(lambda = 50*0.03)
[1] 0.22313016 0.33469524 0.25102143 0.12551072 0.04706652 0.01411996
```


Example (Fatalities From Horse Kicks)

The number of fatalities in a year that resulted from being kicked by a horse or mule was recorded for each of 10 corps of Prussian cavalry over a period of 20 years, giving 200 corps-years worth of data[1].

# of Deaths (in a corp in a year)	0	1	2	3	4	Total
Frequency	109	65	22	3	1	200

The count of deaths due to horse kicks in a corp in a given year may have a Poisson distribution because

- $p = P$(a soldier died from horsekicks in a given year) ≈ 0;
- $n = \#$ of soldiers in a corp was large (100's or 1000's);
- whether a soldier was kicked was (at least nearly) independent of whether others were kicked

[1] von Bortkiewicz (1898) *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.

Example (Fatalities From Horse Kicks — Cont'd)

- Suppose all 10 corps had the same n and p throughout the 20 year period. Then we may assume that the 200 counts all have the Poisson distn. with the same rate $\lambda = np$.
- How to estimate λ?
- MLE for the rate λ of a Poisson distribution is the **sample mean** $\overline{Y}$.
- So for the horsekick data:

# of Deaths (in a corp in a year)	0	1	2	3	4	Total
Frequency	109	65	22	3	1	200

the MLE for λ is

$$\widehat{\lambda} = \frac{0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{200} = 0.61$$

Example (Fatalities From Horse Kicks — Cont'd)

The fitted Poisson probability to have k deaths from horsekicks is

$$\mathrm{P}(Y = k) = e^{-\widehat{\lambda}}\widehat{\lambda}^k/k! = e^{-0.61}(0.61)^k/k!, \quad , k = 0, 1, 2, \ldots$$

k	Observed Frequency	Fitted Poisson Freq. $= 200 \times P(Y = k)$
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6
Total	200	199.9

```
> round(200*dpois(0:4, 0.61),1)
[1] 108.7  66.3  20.2   4.1   0.6
```


When Poisson Distributions Come Up

Variables that are generally Poisson:

- # of misprints on a page of a book
- # of calls coming into an exchange during a unit of time (if the exchange services a large number of customers who act more or less independently.)
- # of people in a community who survive to age 100
- # of customers entering a post office on a given day
- # of vehicles that pass a marker on a roadway during a unit of time (for light traffic only. In heavy traffic, however, one vehicle's movement may influence another)

GLMs for Poisson Response Data

Assume the response $Y \sim \text{Poisson}(\mu(x))$, where x is an explanatory variable.

Commonly used link functions for Poisson distributions are

- identity link: $\mu(x) = \alpha + \beta x$
  - sometimes problematic because $\mu(x)$ must be > 0, but $\alpha + \beta x$ may not
- log link: $\log(\mu(x)) = \alpha + \beta x \quad \Longleftrightarrow \quad \mu(x) = e^{\alpha + \beta x}$.
  - $\mu(x) > 0$ always
  - Whenever x increases by 1 unit, $\mu(x)$ is multiplied by e^{β}

Loglinear models use Poisson with log link

Inference of Parameters

- Wald, LR tests and CIs for β's work as in logistic models
- Goodness of fit:

$$\text{Deviance} = G^2 = 2\sum_i y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) = -2(L_M - L_S)$$

$$\text{Pearson's chi-squared} = X^2 = 2\sum_i \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

G^2 and X^2 are approx. $\sim \chi^2_{n-p}$, when all $\widehat{\mu}_i$'s are large (≥ 10),
where $n =$ num. of observations,
and $p =$ num. of parameters in the model.

Example (Mating and Age of Male Elephants)

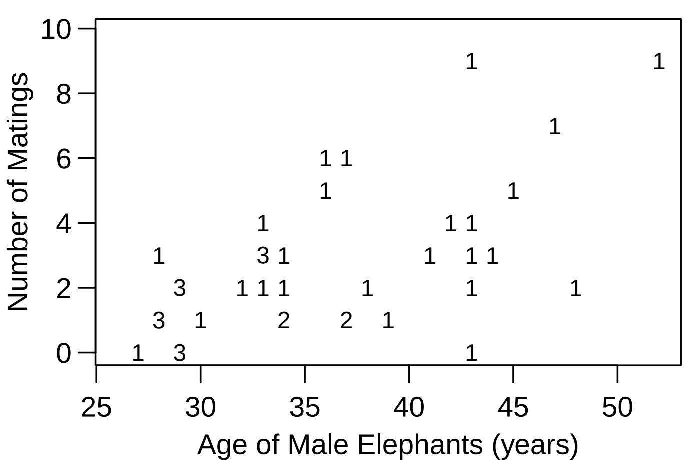
Joyce Poole studied a population of African elephants in Amboseli National Park, Kenya, for 8 years[2].

- Response: number of successful matings in the 8 years of 41 male elephants.
- Predictor: estimated ages of the male elephants at beginning of the study.

Age	Matings	Age	Matings	Age	Matings	Age	Matings
27	0	30	1	36	5	43	3
28	1	32	2	36	6	43	4
28	1	33	4	37	1	43	9
28	1	33	3	37	1	44	3
28	3	33	3	37	6	45	5
29	0	33	3	38	2	47	7
29	0	33	2	39	1	48	2
29	0	34	1	41	3	52	9
29	2	34	1	42	4		
29	2	34	2	43	0		
29	2	34	3	43	2		

[2]Data from J. H. Poole, "Mate Guarding, Reproductive Success and Female Choice in African Elephants", *Animal Behavior* **37** (1989): 842-499.

Example (Mating and Age of Male Elephants)



On the plot, '3' means there are 3 points at the same location.

Example (Elephant)

Let Y = number of successful matings $\sim$ Poisson(μ);

$$\text{Model 1}: \mu = \alpha + \beta\text{Age} \qquad \text{(identity link)}$$

```
> Age = c(27,28,28,28,28,29,29,29,29,29,29,30,32,33,33,33,33,33,34,34,
    34,34,36,36,37,37,37,38,39,41,42,43,43,43,43,43,44,45,47,48,52)
> Matings = c(0,1,1,1,3,0,0,0,2,2,2,1,2,4,3,3,3,2,1,1,2,3,
    5,6,1,1,6,2,1,3,4,0,2,3,4,9,3,5,7,2,9)
> eleph.id = glm(Matings ~ Age, family=poisson(link="identity"))
> summary(eleph.id)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.55205    1.33916  -3.399 0.000676 ***
Age          0.20179    0.04023   5.016 5.29e-07 ***
---
    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 50.058  on 39  degrees of freedom
AIC: 155.5
```

Fitted model 1: $\widehat{\mu} = \widehat{\alpha} + \widehat{\beta}\text{Age} = -4.55 + 0.20$ Age

- $\approx \widehat{\beta} = 0.20$ more matings if the elephant is 1 year older

Example (Elephant)

$$\text{Model 2: } \log(\mu) = \alpha + \beta\text{Age} \qquad \text{(log link)}$$

```
> eleph.log = glm(Matings ~ Age, family=poisson(link="log"))
> summary(eleph.log)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.58201    0.54462  -2.905  0.00368 **
Age          0.06869    0.01375   4.997 5.81e-07 ***
---
    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: 156.46
```

Fitted model 2: $\log(\widehat{\mu}) = -1.582 + 0.0687\text{Age}$

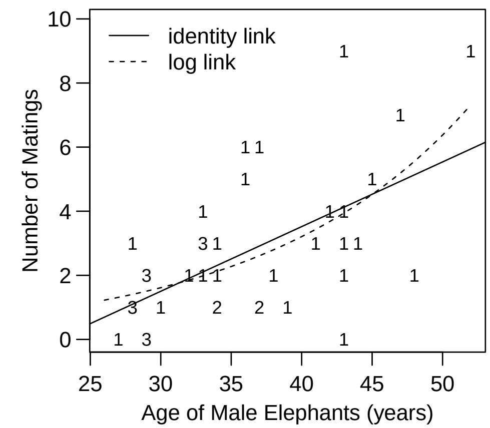
$$\widehat{\mu} = \exp(-1.582 + 0.0687\text{Age}) = 0.205(1.071)^{\text{Age}}$$

- expected number of matings increase by 7.1% for every extra year of age
- for a 40 year-old male, the expected number of matings is $\widehat{\mu} = \exp(-1.582 + 0.0687(40)) \approx 3.2$.

Which Model Better Fits the Data?

	AIC	Deviance	df
Model 1 (identity link)	155.50	50.058	39
Model 2 (log link)	156.46	51.012	39

- Based on AIC, Model 1 fits better
- Goodness of fit tests are not appropriate because ...
- Based on scatter plot...



Residuals

- Deviance residual:

$$d_i = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{2\left[y_i \log(y_i/\widehat{\mu}_i) - y_i + \widehat{\mu}_i\right]}$$

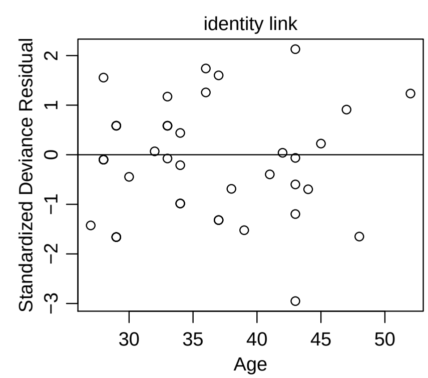
- Pearson's residual: $e_i = \dfrac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$
- Standardized Pearson's residual $= \dfrac{e_i}{\sqrt{1 - h_i}}$
- Standardized Deviance residual $= \dfrac{d_i}{\sqrt{1 - h_i}}$

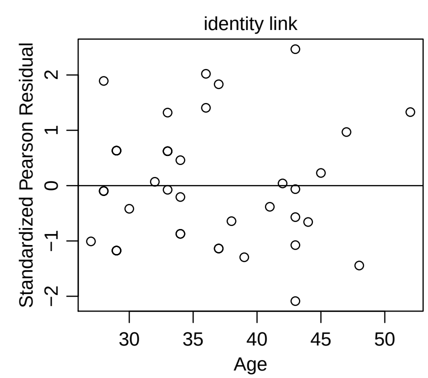
  where h_i = leverage of ith observation
- potential outlier if |standardized residual| > 2 or 3
- R function `residuals()` gives deviance residuals by default, and Pearson residuals with option `type="pearson"`.
- R function `rstandard()` gives standardized deviance residuals by default, and standardized Pearson residuals with option `type="pearson"`.

Residual Plots

```
plot(Age, rstandard(eleph.id),
     ylab="Standardized Deviance Residual", main="identity link")
abline(h=0)

plot(Age, rstandard(eleph.id, type="pearson"),
     ylab="Standardized Pearson Residual", main = "identity link")
abline(h=0)
```

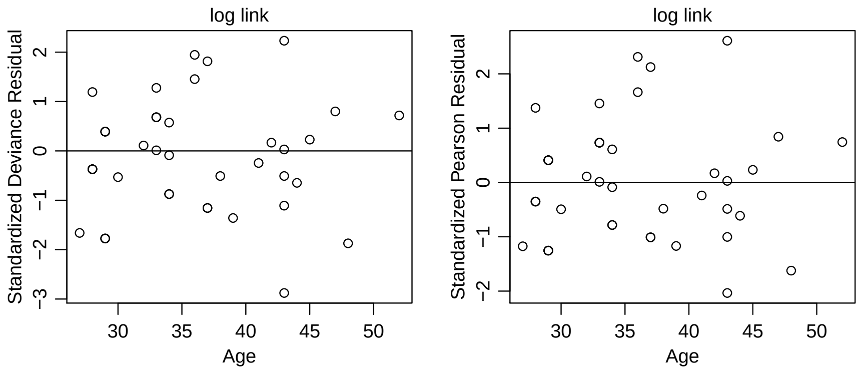




Residual Plots

```
plot(Age, rstandard(eleph.log),
     ylab="Standardized Deviance Residual", main="log link")
abline(h=0)

plot(Age, rstandard(eleph.log, type="pearson"),
     ylab="Standardized Pearson Residual", main = "log link")
abline(h=0)
```



Models for Rates

Sometimes y_i have different bases (e.g., number murders for cities with different pop. sizes)

Let y = count with base t. Assume $y \sim$ Poisson(μ), where

$$\mu = \lambda t$$

more relevant to model rate λ at which events occur.

Loglinear model:

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

i.e.,

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

$\log(t)$ is an offset.
See pp. 82-84 of text for discussion.

Example (British Train Accidents over Time)

Have collisions between trains and road vehicles become more prevalent over time?

- Total number of train-km (in millions) varies from year to year.
- Model annual rate of train-road collisions per million train-km with base t = annual number of train-km, and x = num. of years since 1975

```
> trains = read.table("traincollisions.dat", head=T)
> trains
   Year  KM Train TrRd
1  2003 518     0    3
2  2002 516     1    3
3  2001 508     0    4
4  2000 503     1    3
5  1999 505     1    2
...
27 1977 425     1    8
28 1976 426     2   12
29 1975 436     5    2
```



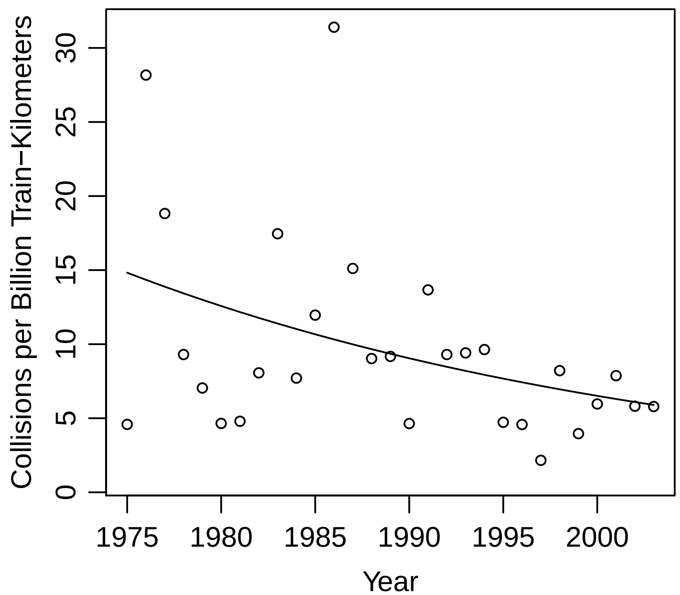
```
> trains1 = glm(TrRd ~ I(Year-1975), offset = log(KM),
                family=poisson, data=trains)
> summary(trains1)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.21142    0.15892  -26.50  < 2e-16 ***
I(Year - 1975) -0.03292    0.01076   -3.06  0.00222 **
---
    Null deviance: 47.376  on 28  degrees of freedom
Residual deviance: 37.853  on 27  degrees of freedom
AIC: 133.52
```

Fitted Model: $\log(\widehat{\lambda}) = \log(\widehat{\mu}/t) = -4.21 - 0.0329x$

$$\widehat{\lambda} = \frac{\widehat{\mu}}{t} = e^{-4.21-0.0329x} = e^{-4.21}(e^{-0.0329})^{x} = (0.0148)(0.968)^{x}$$

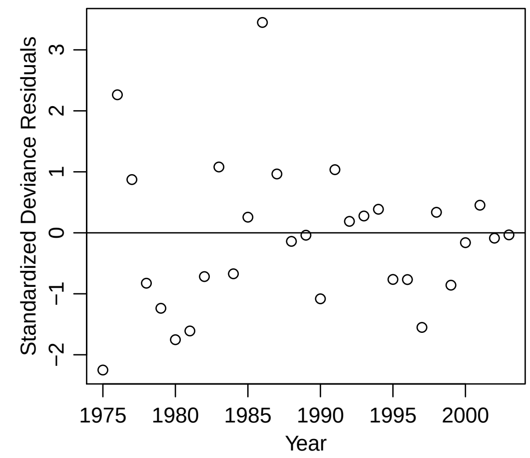
- Rate estimated to decrease by 3.2% per yr from 1975 to 2003.
- Est. rate for 1975 ($x = 0$) is 0.0148 per million km (15 per billion).
- Est. rate for 2003 ($x = 28$) is 0.0059 per million km (6 per billion).

```
plot(trains$Year, 1000*trains$TrRd/trains$KM,xlab="Year",
     ylab="Collisions per Billion Train-Kilometers",ylim=c(1,31.4))
curve(1000*exp(trains1$coef[1]+trains1$coef[2]*(x-1975)), add=T)
```



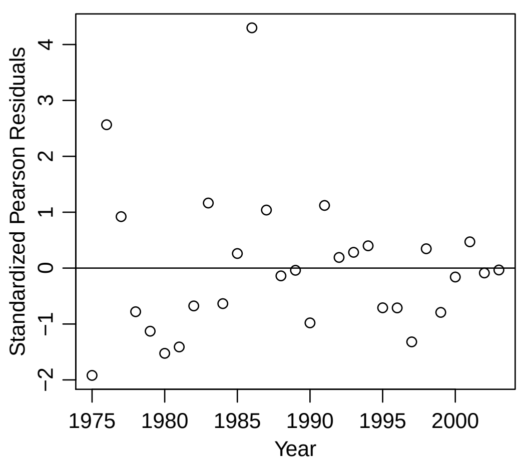
Train Data — Standardized Deviance Residuals

```
plot(trains$Year, rstandard(trains1),
     xlab="Year", ylab="Standardized Deviance Residuals")
abline(h=0)
```



Train Data — Standardized Pearson Residuals

```
plot(trains$Year, rstandard(trains1,type="pearson"),
     xlab="Year", ylab="Standardized Pearson Residuals")
abline(h=0)
```



There were 13 train-road collisions in 1986, a lot higher than the fitted mean 4.3 for that year.

Models for Rate Data With Identity Link

For $y \sim \text{Poisson}(\mu)$ with base t, where

$$\mu = \lambda t$$

the loglinear model

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

assumes the effect of the explanatory variable on the response to be multiplicative.
Alternatively, if we want the effect to be additive,

$$\begin{aligned} & \lambda = \mu/t = \alpha + \beta x \\ \Leftrightarrow \quad & \mu = \alpha t + \beta t x \end{aligned}$$

we may fit a GLM model with identity link, using t and tx as explanatory variables and with no intercept or offset terms.

Train Data — Identity Link

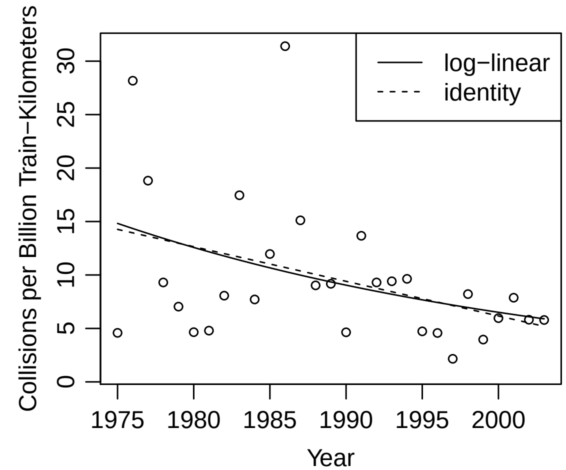
base t = annual num. of train-km, x = num. of years since 1975

```
> trains2 = glm(TrRd ~ -1 + KM + I(KM*(Year-1975)),
                family=poisson(link="identity"), data=trains)
> summary(trains2)
                        Estimate Std. Error z value Pr(>|z|)
KM                     1.426e-02  1.888e-03   7.557 4.14e-14 ***
I(KM * (Year - 1975)) -3.239e-04  9.924e-05  -3.264   0.0011 **
---
    Null deviance:    Inf  on 29  degrees of freedom
Residual deviance: 37.287  on 27  degrees of freedom
AIC: 132.95
```

Fitted Model: $\widehat{\lambda} = \widehat{\mu}/t = 0.0143 - 0.000324x$

- Estimated rate decreases by 0.00032 per million km (0.32 per billion km) per yr from 1975 to 2003.
- Est. rate for 1975 ($x = 0$) is 0.0143 per million km (14.3 per billion km).
- Est. rate for 2003 ($x = 28$) is 0.0052 per million km (5.2 per billion km).


```
plot(trains$Year, 1000*trains$TrRd/trains$KM,xlab="Year",
     ylab="Collisions per Billion Train-Kilometers",ylim=c(1,31.4))
curve(1000*exp(trains1$coef[1]+trains1$coef[2]*(x-1975)), add=T)
curve(1000*trains2$coef[1]+1000*trains2$coef[2]*(x-1975), add=T, lty=2)
legend("topright", c("log-linear","identity"), lty=1:2)
```



The loglinear fit and the linear fit (identity link) are nearly identital.

3.3.4 Overdispersion and Negative Binomial Regression

Overdispersion: Greater Variability than Expected

- One of the defining characteristics of Poisson regression is its lack of a parameter for variability:

$$\mathbb{E}(Y) = \mathrm{Var}(Y),$$

  and no parameter is available to adjust that relationship
- In practice, when working with Poisson regression, it is often the case that the variability of y_i about $\widehat{\lambda}_i$ is larger than what $\widehat{\lambda}_i$ predicts
- This implies that there is more variability around the model's fitted values than is consistent with the Poisson distribution
- This phenomenon is *overdispersion*.

Common Causes of Overdispersion

- Subject heterogeneity
  - subjects have different μ
    e.g., rates of infestation may differ from location to location on the same tree and may differ from tree to tree
  - there are important predictors not included in the model
- Observations are not independent – clustering

Negative Binomial Distribution

If Y has a negative binomial distribution, with mean μ and dispersion parameter $D = 1/\theta$, then

$$P(Y = k) = \frac{\Gamma(k+\theta)}{k!\Gamma(\theta)} \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{k}, \quad k = 0, 1, 2, \ldots$$

One can show that

$$\mathbb{E}[Y] = \mu, \quad \mathrm{Var}(Y) = \mu + \frac{\mu^2}{\theta} = \mu + D\mu^2.$$

- As $D = 1/\theta \downarrow 0$, negative binomial $\rightarrow$ Poisson.
- Negative binomial is a gamma mixture of Poissons, where the Poisson mean varies according to a gamma distribution.
- MLE for μ is the sample mean. MLE for θ has no close form formula.

Example (Known Victims of Homicide)

A recent General Social Survey asked subjects,

> 'Within the past 12 months, how many people have you known personally that were victims of homicide?"

Number of Victims	0	1	2	3	4	5	6	Total
Black Subjects	119	16	12	7	3	2	0	159
White Subjects	1070	60	14	4	0	0	1	1149

If fit a Poisson distribution to the data from blacks, MLE for λ is the sample mean

$$\widehat{\lambda} = \frac{0 \cdot 119 + 1 \cdot 16 + 2 \cdot 12 + \cdots + 6 \cdot 0}{159} = \frac{83}{159} \approx 0.522$$

Fitted $P(Y = k)$ is $e^{-\frac{83}{159}} \left(\frac{83}{159}\right)^k / k!$, $k = 0, 1, 2, \ldots$.

```
> round(dpois(0:6, lambda = 83/159),3)
[1] 0.593 0.310 0.081 0.014 0.002 0.000 0.000
> round(c(119,16,12,7,3,2,0)/159, 3)      # sample relative freq.
[1] 0.748 0.101 0.075 0.044 0.019 0.013 0.000
```


Example (Known Victims of Homicide)

Num. of Victims	0	1	2	3	4	5	6	Total	Mean	Variance
Black	119	16	12	7	3	2	0	159	0.522	1.150
White	1070	60	14	4	0	0	1	1149	0.092	0.155

Likewise, if we fit a Poisson distribution to the data from whites, MLE for λ is

$$\widehat{\lambda} = \frac{0 \cdot 1070 + 1 \cdot 60 + 2 \cdot 14 + \cdots + 6 \cdot 1}{1149} = \frac{106}{1149} \approx 0.092$$

Fitted $P(Y = k)$ is $e^{-\frac{106}{1149}} \left(\frac{106}{1149}\right)^k / k!$, $k = 0, 1, 2, \ldots$.

```
> round(dpois(0:6, lambda = 106/1149), 3) # fitted Poisson prob.
[1] 0.912 0.084 0.004 0.000 0.000 0.000 0.000
> round(c(1070,60,14,4,0,0,1)/1149, 3)    # sample relative freq.
[1] 0.931 0.052 0.012 0.003 0.000 0.000 0.001
```

- Too many 0's and too many large counts for both races than expected if the samples were drawn from Poisson distributions.
- It is not surprising that Poisson distributions do not fit the data because of the large discrepancies between sample mean and sample variance.

Example (Known Victims of Homicide)

Data:

$Y_{b,1}, Y_{b,2}, \ldots, Y_{b,159}$	answers from black subjects
$Y_{w,1}, Y_{w,2}, \ldots\ldots, Y_{w,1149}$	answers from white subjects

Poisson Model:

$$Y_{b,j} \sim \text{Poisson}(\mu_b), \quad Y_{w,j} \sim \text{Poisson}(\mu_w)$$

Neg. Bin. Model:

$$Y_{b,j} \sim \text{NB}(\mu_b, \theta), \quad Y_{w,j} \sim \text{NB}(\mu_w, \theta)$$

Goal: Test whether $\mu_b = \mu_w$.
Equivalent to test $\beta = 0$ in the log-linear model.

$$\log(\mu) = \alpha + \beta x, \quad x = \begin{cases} 1 \text{ if black} \\ 0 \text{ if white,} \end{cases}$$

Note $\mu_b = e^{\alpha+\beta}$, $\mu_w = e^{\alpha}$. So $e^{\beta} = \mu_b/\mu_w$.

Example (Known Victims of Homicide)

Negative binomial regression models can be fit using `glm.nb` function in the `MASS` package.

```
> nvics = c(0:6,0:6)
> race = c(rep("Black", 7),rep("White",7))
> freq = c(119,16,12,7,3,2,0,1070,60,14,4,0,0,1)
> data.frame(nvics,race,freq)
   nvics  race freq
1      0 Black  119
2      1 Black   16
3      2 Black   12
... (omit) ...
12     4 White    0
13     5 White    0
14     6 White    1
> race = factor(race, levels=c("White","Black"))
> hom.poi = glm(nvics ~ race, weights=freq, family=poisson)
> library(MASS)
> hom.nb = glm.nb(nvics ~ race, weights=freq)
```


Example (Known Victims of Homicide) — Poisson Fits

```
> summary(hom.poi)
Call:
glm(formula = nvics ~ race, family = poisson, data = homicide,
    weights = freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.051    0.000    5.257    6.216   13.306

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.38321    0.09713  -24.54   <2e-16 ***
raceBlack    1.73314    0.14657   11.82   <2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 962.80  on 10  degrees of freedom
Residual deviance: 844.71  on  9  degrees of freedom
AIC: 1122

Number of Fisher Scoring iterations: 6
```


Example (Known Victims of Homicide) — Neg. Binomial

```
> summary(hom.nb)
Call:
glm.nb(formula = nvics ~ race, weights = freq, init.theta = 0.2023119205,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-12.754    0.000    2.086    3.283    9.114

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6501     0.2077  -3.130  0.00175 **
raceWhite    -1.7331     0.2385  -7.268 3.66e-13 ***
---
(Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)

    Null deviance: 471.57  on 10  degrees of freedom
Residual deviance: 412.60  on  9  degrees of freedom
AIC: 1001.8

Number of Fisher Scoring iterations: 1
              Theta:  0.2023
          Std. Err.:  0.0409
 2 x log-likelihood:  -995.7980
```

```
> hom.nb$fit
         1          2          3          4          5          6          7
0.52201258 0.52201258 0.52201258 0.52201258 0.52201258 0.52201258 0.52201258
         8          9         10         11         12         13         14
0.09225413 0.09225413 0.09225413 0.09225413 0.09225413 0.09225413 0.09225413
> hom.nb$theta
[1] 0.2023119
```

- Fitted values given by the Neg. Bin model are simply the sample means — 0.522 ($=\frac{83}{159}$) for blacks and 0.0922 ($=\frac{106}{1149}$) for whites.
- Estimated common dispersion parameter is $\widehat{\theta} = 0.2023119$ with SE = 0.0409.
- Fitted $P(Y = k)$ is

$$\frac{\Gamma(k+\widehat{\theta})}{k!\Gamma(\widehat{\theta})}\left(\frac{\widehat{\theta}}{\widehat{\mu}+\widehat{\theta}}\right)^{\theta}\left(\frac{\widehat{\mu}}{\widehat{\mu}+\widehat{\theta}}\right)^{k}, \text{ where } \widehat{\mu} = \begin{cases}\frac{83}{159} & \text{for blacks}\\ \frac{106}{1149} & \text{for whites.}\end{cases}$$

- Textbook uses $D = 1/\theta$ as the dispersion parameter, estimated as $\widehat{D} = 1/\widehat{\theta} = 1/0.2023 \approx 4.94$.

Example (Known Victims of Homicide)

Black Subjects

Num. of Victims	0	1	2	3	4	5	6	Total
observed freq.	119	16	12	7	3	2	0	159
relative freq.	0.748	0.101	0.075	0.044	0.019	0.013	0	1
poisson fit	0.593	0.310	0.081	0.014	0.002	0.000	0.000	1
neg. bin.fit	0.773	0.113	0.049	0.026	0.015	0.009	0.006	0.991

White Subjects:

num. of victims	0	1	2	3	4	5	6	Total
observed freq.	1070	60	14	4	0	0	1	1149
relative freq.	0.931	0.052	0.012	0.003	0.000	0.000	0.001	0.999
poisson fit	0.912	0.084	0.004	0.000	0.000	0.000	0.000	1
neg. bin.fit	0.927	0.059	0.011	0.003	0.001	0.000	0.000	1.001

```
# neg. bin fit
> round(dnbinom(0:6, size = hom.nb$theta, mu = 83/159),3)  # black
[1] 0.773 0.113 0.049 0.026 0.015 0.009 0.006
> round(dnbinom(0:6,size = hom.nb$theta, mu=106/1149),3)   # white
[1] 0.927 0.059 0.011 0.003 0.001 0.000 0.000
```


Example (Known Victims of Homicide)

$$\text{Model:}\quad \log(\mu) = \alpha + \beta x, \quad x = \begin{cases}1 & \text{if black}\\ 0 & \text{if white,}\end{cases}$$

Model	$\widehat{\alpha}$	$\widehat{\beta}$	SE($\widehat{\beta}$)	Wald 95% CI for $e^{\beta} = \mu_B/\mu_A$
Poisson	−2.38	1.73	0.147	$\exp(1.73 \pm 1.96 \cdot 0.147) = (4.24, 7.54)$
Neg. Binom.	−2.38	1.73	0.238	$\exp(1.73 \pm 1.96 \cdot 0.238) = (3.54, 9.03)$

Poisson and negative binomial models give

- identical estimates for coefficients (this data set only, not always the case)
- but different SEs for $\widehat{\beta}$ (Neg. Binom. gives bigger SE)

To account for overdispersion, neg. binom. model gives wider Wald CIs (and also wider LR CIs).

Remark. Observe $e^{\widehat{\beta}} = e^{1.73} = 5.7$ is the ratio of the two sample means $\overline{y}_{\text{black}}/\overline{y}_{\text{white}} = 0.522/0.092$.

Wald CIs

```
> confint.default(hom.poi)
                2.5 %    97.5 %
(Intercept) -2.573577 -2.192840
raceBlack    1.445877  2.020412

> exp(confint.default(hom.poi))
                2.5 %    97.5 %
(Intercept) 0.0762623 0.1115994
raceBlack   4.2455738 7.5414329

> confint.default(hom.nb)
                2.5 %    97.5 %
(Intercept) -2.612916 -2.153500
raceBlack    1.265738  2.200551

> exp(confint.default(hom.nb))
                 2.5 %    97.5 %
(Intercept) 0.07332043 0.1160771
raceBlack   3.54571025 9.0299848
```

Likelihood Ratio CIs

```
> confint(hom.poi)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) -2.579819 -2.198699
raceBlack    1.443698  2.019231
> exp(confint(hom.poi))
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.0757877 0.1109474
raceBlack   4.2363330 7.5325339

> confint(hom.nb)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) -2.616478 -2.156532
raceBlack    1.274761  2.211746
> exp(confint(hom.nb))
Waiting for profiling to be done...
                 2.5 %    97.5 %
(Intercept) 0.07305976 0.1157258
raceBlack   3.57784560 9.1316443
```


If Not Taking Overdispersion Into Account ...

- SEs are underestimated
- CIs will be too narrow
- Significance of variables will be over stated (reported P values are lower than the actual ones)

How to Check for Overdispersion?

- Think about whether overdispersion is likely — e.g., important explanatory variables are not available, or dependence in observations.
- Compare the sample variances to the sample means computed for groups of responses with identical explanatory variable values.
- Large deviance relative to its deviance
- Examine residuals to see if a large deviance statistic may be due to outliers
- Large numbers of outliers are usually signs of overdispersion
- Check standardized residuals and plot them against them fitted values $\widehat{\mu}_i$.

Train Data Revisit

Recall Pearson's residual:

$$e_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$$

If no overdispersion, then

$$\mathrm{Var}(Y) \approx (y_i - \widehat{\mu}_i)^2 \approx \mathbb{E}(Y) \approx \widehat{\mu}_i$$

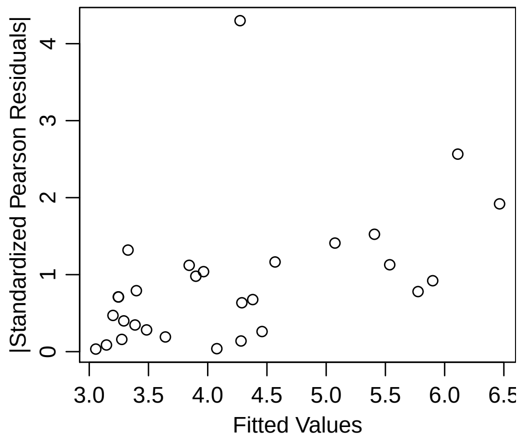
So the size of Pearson's residuals should be around 1.
With overdispersion,

$$\mathrm{Var}(Y) = \mu + D\mu^2$$

then the size of Pearson's residuals may increase with μ.
We may check the plot of the absolute value of (standardized) Pearson's residuals against fitted values $\widehat{\mu}_i$.

Train Data — Checking Overdispersion

```
plot(trains1$fit, abs(rstandard(trains1, type="pearson")),
     xlab="Fitted Values", ylab="|Standardized Pearson Residuals|")
```



The size of standardized Pearson's residuals tend to increase with fitted values. This is a sign of overdisperson.

Train Data — Neg. Bin. Model

```
> trains.nb = glm.nb(TrRd ~ I(Year-1975)+offset(log(KM)), data=trains)
> summary(trains.nb)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.19999    0.19584 -21.446  < 2e-16 ***
I(Year - 1975) -0.03367    0.01288  -2.615  0.00893 **
---
(Dispersion parameter for Negative Binomial(10.1183) family taken to be 1

    Null deviance: 32.045  on 28  degrees of freedom
Residual deviance: 25.264  on 27  degrees of freedom
AIC: 132.69

              Theta:  10.12
          Std. Err.:  8.00
 2 x log-likelihood:  -126.69
```

For year effect, the estimated coefficients are similar (0.0337 for neg. bin. model compared to 0.032 for Poisson model), but less significant (P-value $= 0.009$ in neg. bin. model compared to 0.002 in Poisson model)