Generalized Linear Models

based on Zuur et al.
Kuhnert and Venables
Normal Distribution example: weights of 1280 sparrows

Sparrows<-read.table("/Users/dbm/Documents/W2025/ZuurDataMixedModelling/Sparrows.txt",header=TRUE)
par(mfrow = c(2, 2))
hist(Sparrows$wt, nclass = 15, xlab = "Weight", main = "Observed data",xlim=c(10,30))
hist(Sparrows$wt, nclass = 15, xlab = "Weight", main = "Observed data", freq = FALSE,xlim=c(10,30))
Y <- rnorm(1281, mean = mean(Sparrows$wt), sd = sd(Sparrows$wt))
hist(Y, nclass = 15, main = "Simulated data",xlab = "Weight",xlim=c(10,30))

X <- seq(from = 0,to = 30,length = 200)
Y <- dnorm(X, mean = mean(Sparrows$wt), sd = sd(Sparrows$wt))
plot(X, Y, type = "l", xlab = "Weight", ylab = "Probablities", ylim = c(0, 0.25), xlim = c(10, 30),
main = "Normal density curve")
par(mfrow = c(1, 1))
\[ f(y_i; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} \]

\[ E(Y) = \mu \quad \text{and} \quad \text{var}(Y) = \sigma^2 \]
Poisson Distribution

\[ f(y; \mu) = \frac{\mu^y \times e^{-\mu}}{y!} \quad y \geq 0, \text{ } y \text{ integer} \]

\[ E(Y) = \mu \quad \text{and} \quad \text{var}(Y) = \mu \]

In practice the variance is often bigger than the mean -> overdispersion
Fig. 8.2  Poisson probabilities for $\mu = 3$ (A), $\mu = 5$ (B), $\mu = 10$ (C), and $\mu = 100$ (D). Equation (8.3) is used to calculate the probabilities for certain values. Because the outcome variable $y$ is a count, vertical lines are used instead of a line connecting all the points.
Gamma Distribution

\[ f(y; \mu, \nu) = \frac{1}{\Gamma(\nu)} \times \left( \frac{\nu}{\mu} \right)^\nu \times y^{\nu-1} \times e^{\frac{y\times\nu}{\mu}} \quad y > 0 \]

\[ E(Y) = \mu \quad \text{and} \quad \text{var}(Y) = \frac{\mu^2}{\nu} \]
**Fig. 8.4** Gamma distributions for different values of $\mu$ and $\nu$. The R function `dgamma` was applied, which uses a slightly different parameterisation: $E(Y) = a \times s$ and $\text{var}(Y) = a \times s^2$, where $a$ is called the shape and $s$ the scale. In our parameterisation, $\nu = a$ and $\mu = a \times s$. 
Binomial Distribution

\[ f(y; \pi) = \binom{N}{y} \times \pi^y \times (1 - \pi)^{N-y} \]

\[ E(Y) = N \times \pi \quad \text{var}(Y) = N \times \pi \times (1 - \pi) \]
Fig. 8.5  Binomial density curves $B(\pi, N)$ for various values of $\pi$ (namely 0.2, 0.5, and 0.7) and $N$ (namely 10, 20, and 100). R code to create this graph is on the book website
Natural Exponential Family

\[ f(y; \theta, \phi) = e^{\frac{y \times \theta - b(\theta)}{a(\phi)}} + c(y, \theta) \]

\[ E(Y) = b''(\theta) \]
\[ \text{var}(Y) = b'''(\theta) \times a(\phi) \]

to get, e.g., Poisson:

\[ \theta = \log(\mu), \phi = 1, a(\phi) = 1, b(\theta) = \exp(\theta), c(y, \phi) = -\log(y!) \]
Table 8.1  List of distributions for the response variable. Density means numbers per Area (or volume, range, etc), and in this case the offset option is needed in the Poisson or NB GLM

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Type of data</th>
<th>Mean – variance relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Continuous</td>
<td>Equation (8.2)</td>
</tr>
<tr>
<td>Poisson</td>
<td>Counts (integers) and density</td>
<td>Equation (8.4)</td>
</tr>
<tr>
<td>Negative binomial</td>
<td>Overdispersed counts and density</td>
<td>Equation (8.7)</td>
</tr>
<tr>
<td>Geometric</td>
<td>Overdispersed counts and density</td>
<td>Equation (8.8)</td>
</tr>
<tr>
<td>Gamma</td>
<td>Continuous</td>
<td>Equation (8.10)</td>
</tr>
<tr>
<td>Binomial</td>
<td>Proportional data</td>
<td>Equation (8.12)</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>Presence absence data</td>
<td>Equation (8.12) with $N = 1$</td>
</tr>
</tbody>
</table>
Generalized Linear Models

A GLM consists of three steps:

1. An assumption on the distribution of the response variable $Y_i$. This also defines the mean and variance of $Y_i$.
2. Specification of the systematic part. This is a function of the explanatory variables.
3. The relationship between the mean value of $Y_i$ and the systematic part. This is also called the link between the mean and the systematic part.
**Step 1:** In a Gaussian linear regression, we assume that the response variable \( Y_i \) is normally distributed with mean \( \mu_i \) and variance \( \sigma^2 \). The index \( i \) refers to a case or observation.

**Step 2:** In the second step, we specify the systematic part of the model. This means that we need to select the explanatory variables. Define the predictor function \( \eta(X_{i1}, \ldots, X_{iq}) \) by:

\[
\eta(X_{i1}, \ldots, X_{iq}) = \alpha + \beta_1 \times X_{i1} + \cdots + \beta_q \times X_{iq}
\]  

(9.1)

The systematic part is given by the predictor function \( \eta(X_{i1}, \ldots, X_{iq}) \).

**Step 3:** In the third step, we need to specify the link between the expected value of \( Y_i \) (which is \( \mu_i \)) and the predictor function \( \eta(X_{i1}, \ldots, X_{iq}) \). We use the identity link, which means that \( \mu_i = \eta(X_{i1}, \ldots, X_{iq}) \).

These three steps give the following GLM:

\[
Y_i \sim N(\mu_i, \sigma^2)
\]

\[
E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \sigma^2
\]

(9.2)

\[
\mu_i = \eta(X_{i1}, \ldots, X_{iq})
\]
Generalized Linear Models: Poisson regression

1. $Y_i$ is Poisson distributed with mean $\mu_i$. By definition of this distribution, the variance of $Y_i$ is also equal to $\mu_i$.
2. The systematic part is given by $\eta(X_{i1}, \ldots, X_{iq}) = \alpha + \beta_1 \times X_{i1} + \cdots + \beta_q \times X_{iq}$.
3. There is a logarithmic link between the mean of $Y_i$ and the predictor function $\eta(X_{i1}, \ldots, X_{iq})$. The logarithmic link (also called a log link) ensures that the fitted values are always non-negative.

$$Y_i \sim P(\mu_i)$$
$$E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \mu_i$$
$$\log(\mu_i) = \eta(X_{i1}, \cdots, X_{iq}) \quad \text{or} \quad \mu_i = e^{\eta(X_{i1}, \cdots, X_{iq})}$$
x <- 0:100
y <- rpois(101,exp(0.01+0.03*x))
MyData <- data.frame(x,y)
MyModel <- glm(y~x,data=MyData,family=poisson)
summary(MyModel)
coefs <- coef(MyModel)
plot(y~x,MyData,ylim=c(0,26))
par(new=TRUE)
plot(x,exp(0.01+0.03*x),ylim=c(0,26),type="l")
par(new=TRUE)
plot(x,exp(coefs[1]+coefs[2]*x),ylim=c(0,26),type="l",col="red")

Estimates parameters by maximizing the likelihood:

\[ L = \prod_{i} \frac{\mu_i \times e^{-\mu_i}}{y_i!}, \quad \mu_i = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d) \]
Fig. 9.3  Scatterplot of amphibian road kills versus distance (in metres) to a nearby Natural Park
1. $Y_i$, the number of killed animals at site $i$, is Poisson distributed with mean $\mu_i$.
2. The systematic part is given by $\eta(D.PARK_i) = \alpha + \beta \times D.PARK_i$.
3. There is a logarithm link between the mean of $Y_i$ and the predictor function $\eta(D.PARK_i)$.

As a result of these three steps, we have

\begin{align*}
Y_i &\sim p(\mu_i) \\
E(Y_i) &= \mu_i \quad \text{and} \quad \text{var}(Y_i) = \mu_i \\
\log(\mu_i) &= \alpha + \beta \times D.PARK_i \\ 
\text{or} \quad \mu_i &= e^{\alpha + \beta \times D.PARK_i}
\end{align*}

(9.6)
RoadKills <- read.table("/Users/dbm/Documents/W2025/ZuurDataMixedModelling/RoadKills.txt",header=TRUE)
RK <- RoadKills #Saves some space in the code
plot(RK$D.PARK, RK$TOT.N, xlab = "Distance to park", ylab = "Road kills")
M1 <- glm(TOT.N ~ D.PARK, family=poisson, data=RK)
summary(M1)

M2 <- glm(TOT.N ~ D.PARK, family=gaussian, data=RK)
summary(M2)
Call:
```
glm(formula = TOT.N ~ D.PARK, family = poisson, data = RK)
```

Deviance Residuals:
```
        Min      1Q  Median      3Q     Max
-8.1100 -1.6950 -0.4708  1.4206  7.3337
```

Coefficients:
```
                     Estimate  Std. Error   z value Pr(>|z|)
(Intercept)     4.316e+00  4.322e-02   99.87   <2e-16 ***
D.PARK           -1.059e-04  4.387e-06  -24.13   <2e-16 ***
```

---

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

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1071.4  on 51 degrees of freedom
Residual deviance: 390.9  on 50 degrees of freedom
AIC: 634.29

Number of Fisher Scoring iterations: 4
**Residual Deviance:**

Difference in $G^2 = -2 \log L$ between a maximal model where every observation has Poisson parameter equal to the observed value and your model.

**Null Deviance:**

Difference in $G^2 = -2 \log L$ between a maximal model where every observation has Poisson parameter equal to the observed value and the model with just an intercept.

Sometimes useful to compare residual deviances for two models. Differences are chi-square distributed with degrees of freedom equal to the change in the number of parameters.
We do not have an $R^2$ in GLM models, but the closest we can get is the explained deviance, which is calculated as

$$100 \times \frac{\text{null deviance} - \text{residual deviance}}{\text{null deviance}} = 100 \times \frac{1071.4 - 390.9}{1071.4} = 63.51\%$$

So the explanatory variable distance to the park explains 63.51% of the variation in road kills. Dobson (2002) called this proportional increase in explained deviance the pseudo $R^2$. 
MyData <- data.frame(D.PARK = seq(from = 0, to = 25000, by = 1000))
G <- predict(M1, newdata = MyData, type = "link", se = TRUE)
F <- exp(G$fit)
FSEUP <- exp(G$fit + 1.96 * G$se.fit)
FSELOW <- exp(G$fit - 1.96 * G$se.fit)
lines(MyData$D.PARK, F, lty = 1)
lines(MyData$D.PARK, FSEUP, lty = 2)
lines(MyData$D.PARK, FSELOW, lty = 2)
# Model Selection in Generalized Linear Models

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open lands (ha)</td>
<td>OPEN.L</td>
</tr>
<tr>
<td>Olive grooves (ha)</td>
<td>OLIVE</td>
</tr>
<tr>
<td>Montado with shrubs (ha)</td>
<td>MONT.S</td>
</tr>
<tr>
<td>Montado without shrubs (ha)</td>
<td>MONT</td>
</tr>
<tr>
<td>Policulture (ha)</td>
<td>POLIC</td>
</tr>
<tr>
<td>Shrubs (ha)</td>
<td>SHRUB</td>
</tr>
<tr>
<td>Urban (ha)</td>
<td>URBAN</td>
</tr>
<tr>
<td>Water reservoirs (ha)</td>
<td>WAT.RES</td>
</tr>
<tr>
<td>Length of water courses (km)</td>
<td>L.WAT.C</td>
</tr>
<tr>
<td>Dirty road length (m)</td>
<td>L.D.ROAD</td>
</tr>
<tr>
<td>Paved road length (km)</td>
<td>L.P.ROAD</td>
</tr>
<tr>
<td>Distance to water reservoirs</td>
<td>D.WAT.RES</td>
</tr>
<tr>
<td>Distance to water courses</td>
<td>D.WAT.COUR</td>
</tr>
<tr>
<td>Distance to Natural Park (m)</td>
<td>D.PARK</td>
</tr>
<tr>
<td>Number of habitat Patches</td>
<td>N.PATCH</td>
</tr>
<tr>
<td>Edges perimeter</td>
<td>P.EDGE</td>
</tr>
<tr>
<td>Landscape Shannon diversity index</td>
<td>L.SDI</td>
</tr>
</tbody>
</table>
library(AED)
library(lattice)
xyplot(Y ~ X, aspect = "iso", col = 1, pch = 16, data = RoadKills)
> RK$SQ.POLIC <- sqrt(RK$POLIC)
> RK$SQ.WATRES <- sqrt(RK$WAT.RES)
> RK$SQ.URBAN <- sqrt(RK$URBAN)
> RK$SQ.OLIVE <- sqrt(RK$OLIVE)
> RK$SQ.LPROAD <- sqrt(RK$L.P.ROAD)
> RK$SQ.SHRUB <- sqrt(RK$SHRUB)
> RK$SQ.DWATCOUR <- sqrt(RK$D.WAT.COUR)
> M2 <- glm(TOT.N ~ OPEN.L + MONT.S + SQ.POLIC +
          D.PARK + SQ.SHRUB + SQ.WATRES + L.WAT.C +
          SQ.LPROAD + SQ.DWATCOUR, family = poisson,
          data = RK)
> summary(M2)
> summary(M2)

Call:
  glm(formula = TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + D.PARK + SQ.SHARB +
      SQ.WATRES + L.WAT.C + SQ.LROAD + SQ.DWATCOUR, family = poisson,
      data = RK)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-6.8398  -1.3965  -0.1409   1.4641   4.3749

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.749e+00  1.567e-01 23.935  < 2e-16 ***
OPEN.L     -3.025e-03  1.580e-03 -1.915  0.055531 .
MONT.S      8.697e-02  1.359e-02  6.398  1.57e-10 ***
SQ.POLIC   -1.787e-01  4.676e-02 -3.822  0.000133 ***
D.PARK     -1.301e-04  5.936e-06 -21.923 < 2e-16 ***
SQ.SHARB   -6.112e-01  1.176e-01 -5.197  2.02e-07 ***
SQ.WATRES  2.243e-01  7.050e-02  3.181  0.001468 **
L.WAT.C    3.355e-01  4.127e-02  8.128  4.36e-16 ***
SQ.LROAD   4.517e-01  1.348e-01  3.351  0.000804 ***
SQ.DWATCOUR 7.355e-03  4.879e-03  1.508  0.131629

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1071.44  on 51  degrees of freedom
Residual deviance:  270.23  on 42  degrees of freedom
AIC: 529.62

Number of Fisher Scoring iterations: 5
How to select variables for inclusion in the model

• Hypothesis testing
  • Use the z-statistic produced by summary
    • `drop1(M2,test="Chi")`
    • `anova(M2,M3,test="Chi")`

• AIC, BIC, etc. - use `step(M2)` etc.
> drop1(M2, test="Chi")

Single term deletions

Model:

\[ \text{TOT.N} \sim \text{OPEN.L} + \text{MONT.S} + \text{SQ.POLIC} + \text{D.PARK} + \text{SQ.SHRUB} + \text{SQ.WATRES} + \text{L.WAT.C} + \text{SQ.LPROAD} + \text{SQ.DWATCOUR} \]

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>270.23</td>
<td>529.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPEN.L</td>
<td>1 273.93</td>
<td>531.32</td>
<td>3.69</td>
<td>0.0546474</td>
</tr>
<tr>
<td>MONT.S</td>
<td>1 306.89</td>
<td>564.28</td>
<td>36.66</td>
<td>1.410e-09 ***</td>
</tr>
<tr>
<td>SQ.POLIC</td>
<td>1 285.53</td>
<td>542.92</td>
<td>15.30</td>
<td>9.181e-05 ***</td>
</tr>
<tr>
<td>D.PARK</td>
<td>1 838.09</td>
<td>1095.48</td>
<td>567.85</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SQ.SHRUB</td>
<td>1 298.31</td>
<td>555.70</td>
<td>28.08</td>
<td>1.167e-07 ***</td>
</tr>
<tr>
<td>SQ.WATRES</td>
<td>1 280.02</td>
<td>537.41</td>
<td>9.79</td>
<td>0.0017539 **</td>
</tr>
<tr>
<td>L.WAT.C</td>
<td>1 335.47</td>
<td>592.86</td>
<td>65.23</td>
<td>6.648e-16 ***</td>
</tr>
<tr>
<td>SQ.LPROAD</td>
<td>1 281.25</td>
<td>538.64</td>
<td>11.02</td>
<td>0.0009009 ***</td>
</tr>
<tr>
<td>SQ.DWATCOUR</td>
<td>1 272.50</td>
<td>529.89</td>
<td>2.27</td>
<td>0.1319862</td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> step(M2)
Start:  AIC=529.62
TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + D.PARK + SQ.SHUB + SQ.WATRES +
     L.WAT.C + SQ.LPROAD + SQ.DWATCOUR

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>270.23</td>
<td>529.62</td>
</tr>
<tr>
<td>- SQ.DWATCOUR</td>
<td>272.50</td>
<td>529.89</td>
</tr>
<tr>
<td>- OPEN.L</td>
<td>273.93</td>
<td>531.32</td>
</tr>
<tr>
<td>- SQ.WATRES</td>
<td>280.02</td>
<td>537.41</td>
</tr>
<tr>
<td>- SQ.LPROAD</td>
<td>281.25</td>
<td>538.64</td>
</tr>
<tr>
<td>- SQ.POLIC</td>
<td>285.53</td>
<td>542.92</td>
</tr>
<tr>
<td>- SQ.SHUB</td>
<td>298.31</td>
<td>555.70</td>
</tr>
<tr>
<td>- MONT.S</td>
<td>306.89</td>
<td>564.28</td>
</tr>
<tr>
<td>- L.WAT.C</td>
<td>335.47</td>
<td>592.86</td>
</tr>
<tr>
<td>- D.PARK</td>
<td>838.09</td>
<td>1095.48</td>
</tr>
</tbody>
</table>

Call: glm(formula = TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + D.PARK + SQ.SHUB +
     SQ.WATRES + L.WAT.C + SQ.LPROAD + SQ.DWATCOUR, family = poisson,
     data = RK)

Coefficients:
(Intercept)        OPEN.L       MONT.S         SQ.POLIC       D.PARK       SQ.SHUB     SQ.WATRES       L.WAT.C       SQ.LPROAD       SQ.DWATCOUR
 3.7493885        -0.0030250       0.0869656       -0.1787178      -0.0001301      -0.6111864      0.2242561       0.3354676        0.4517172        0.0073554

Degrees of Freedom: 51 Total (i.e. Null); 42 Residual
Null Deviance:  1071
Residual Deviance:  270.2  AIC:  529.6
> step(M2,k=log(n))
Start: AIC=555.67
TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + D.PARK + SQ.SHRUB + SQ.WATRES +
    L.WAT.C + SQ.LPROAD + SQ.DWATCOUR

                  Df Deviance    AIC
- SQ.DWATCOUR  1     272.50   553.34
- OPEN.L       1     273.93   554.76
<none>                    270.23   555.67
- SQ.WATRES    1     280.02   560.86
- SQ.LPROAD    1     281.25   562.09
- SQ.POLIC     1     285.53   566.37
- SQ.SHRUB     1     298.31   579.14
- MONT.S       1     306.89   587.72
- L.WAT.C      1     335.47   616.30
- D.PARK       1     838.00  1118.92

Step: AIC=553.34
TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + D.PARK + SQ.SHRUB + SQ.WATRES +
    L.WAT.C + SQ.LPROAD

                  Df Deviance    AIC
<none>                    272.50   553.34
- OPEN.L       1     277.60   553.83
- SQ.WATRES    1     281.22   557.46
- SQ.POLIC     1     285.62   561.86
- SQ.LPROAD    1     286.31   562.54
- SQ.SHRUB     1     300.59   576.82
- MONT.S       1     311.91   588.14
- L.WAT.C      1     339.08   615.31
- D.PARK       1     843.43  1119.66

Call: glm(formula = TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + D.PARK + SQ.SHRUB +
    SQ.WATRES + L.WAT.C + SQ.LPROAD, family = poisson, data = RK)

Coefficients:
(Intercept)     OPEN.L       MONT.S     SQ.POLIC      D.PARK      SQ.SHRUB     SQ.WATRES     L.WAT.C     SQ.LPROAD
  3.8516007  -0.0034641  0.0892662  -0.1583426  -0.0001286  -0.6159894  0.2079514  0.3112948  0.4935665

Degrees of Freedom: 51 Total (i.e. Null); 43 Residual
Null Deviance: 1071
Residual Deviance: 272.5     AIC: 529.9
Overdispersion

Recall that in the Poisson model we assume the mean = variance

With no overdispersion, residual mean deviance should be close to 1:

\[
\frac{270.23}{42} = 6.43 \ldots \text{not close to 1}
\]

Or, consider mean squared residuals:

\[
\text{sum(resid(M2, type="pearson")^2) / 42 = 5.93}
\]

(suggests standard errors should be increased by \(\sqrt{5.93} = 2.44\))

Quasi-Poisson model accounts for overdispersion:
\[ M4 \leftarrow \text{glm}(\text{TOT.N} \sim \text{OPEN.L} + \text{MONT.S} + \text{SQ.POLIC} + \text{SQ.SHRUB} + \text{SQ.WATRES} + \text{L.WAT.C} + \text{SQ.LPROAD} + \text{SQ.DWATCOUR} + \text{D.PARK}, \]
\[ \text{family} = \text{quasipoisson}, \text{data} = \text{RK}) \]

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 3.749e+00 | 3.814e-01 | 9.830 | 1.86e-12 |
| OPEN.L    | -3.025e-03 | 3.847e-03 | -0.786 | 0.43604 |
| MONT.S    | 8.697e-02  | 3.309e-02 | 2.628 | 0.01194 |
| SQ.POLIC  | -1.787e-01 | 1.139e-01 | -1.570 | 0.12400 |
| SQ.SHRUB  | -6.112e-01 | 2.863e-01 | -2.135 | 0.03867 |
| SQ.WATRES | 2.243e-01  | 1.717e-01 | 1.306 | 0.19851 |
| L.WAT.C   | 3.355e-01  | 1.005e-01 | 3.338 | 0.00177 |
| SQ.LPROAD | 4.517e-01  | 3.282e-01 | 1.376 | 0.17597 |
| SQ.DWATCOUR | 7.355e-03 | 1.188e-02 | 0.619 | 0.53910 |
| D.PARK    | -1.301e-04 | 1.445e-05 | -9.004 | 2.33e-11 |

Dispersion parameter for quasipoisson family taken to be 5.928003

Null deviance: 1071.44 on 51 degrees of freedom
Residual deviance: 270.23 on 42 degrees of freedom
AIC: NA
Overdispersion

AIC not defined for quasi-Poisson model

Use `drop1(M4, test="F")` is approximately correct
or, better, do cross-validation

*Fig. 9.5* Fitted line of the optimal quasi-Poisson model using only `D.PARK` as the explanatory variables. R code to make this graph is given on the book’s website
Fig. 9.5  Fitted line of the optimal quasi-Poisson model using only D. PARK as the explanatory variables. R code to make this graph is given on the book’s website.
Residuals for Poisson Model

Because variance increases with the mean, standard residual doesn’t make much sense. Can use “Pearson residuals”

\[ \hat{e}_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(Y_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \]

For quasi-Poisson should also divide by the square root of the overdispersion parameter
Negative Binomial GLM

1. $Y_i$ is negative binomial distributed with mean $\mu_i$ and parameter $k$ (see also Chapter 8). By definition, the variance of $Y_i$ is also equal to $\mu_i$ and its variance is $\mu_i + \mu_i^2 / k$.

2. The systematic part is given by $\eta(X_{i1}, \ldots, X_{iq}) = \alpha + \beta_1 \times X_{i1} + \ldots + \beta_q \times X_{iq}$.

3. There is a logarithm link between the mean of $Y_i$ and the predictor function $\eta(X_{i1}, \ldots, X_{iq})$. The logarithmic link (also called log link) ensures that the fitted values are always non-negative.
\( Y_i \sim NB(\mu_i, k) \)

\( E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \mu_i + \frac{\mu_i^2}{k} \)

\( \log(\mu_i) = \eta(X_{i1}, \cdots, X_{iq}) \quad \text{or} \quad \mu_i = e^{\eta(X_{i1}, \cdots, X_{iq})} \)
> library(MASS)
> M6 <- glm.nb(TOT.N ~ OPEN.L + MONT.S + SQ.POLIC +
>             SQ.SHRUB + SQ.WATRES + L.WAT.C + SQ.LPROAD +
>             SQ.DWATCOUR + D.PARK, link = "log", data = RK)
| Estimate  | Std. Error | z value | Pr(>|z|) |
|-----------|------------|---------|----------|
| (Intercept) | 3.951e+00 | 4.145e-01 | 9.532 | <2e-16 |
| OPEN.L | -9.419e-03 | 3.245e-03 | -2.903 | 0.0037 |
| MONT.S | 5.846e-02 | 3.481e-02 | 1.679 | 0.0931 |
| SQ.POLIC | -4.618e-02 | 1.298e-01 | -0.356 | 0.7221 |
| SQ.SHRUB | -3.881e-01 | 2.883e-01 | -1.346 | 0.1784 |
| SQ.WATRES | 1.631e-01 | 1.675e-01 | 0.974 | 0.3301 |
| L.WAT.C | 2.076e-01 | 9.636e-02 | 2.154 | 0.0312 |
| SQ.LPROAD | 5.944e-01 | 3.214e-01 | 1.850 | 0.0644 |
| SQ.DWATCOUR | -1.489e-05 | 1.139e-02 | -0.001 | 0.9990 |

Dispersion parameter for Negative Binomial(5.5178) family taken to be 1

Null deviance: 213.674 on 51 degrees of freedom
Residual deviance: 51.803 on 42 degrees of freedom
AIC: 390.11
Theta: 5.52
Std. Err.: 1.41
2 x log-likelihood: -368.107
M7 <- glm.nb(formula = TOT.N ~ OPEN.L + L.WAT.C + SQ.LPROAD + D.PARK, data = RK)

plot(M7)
Contrast with quasi-Poisson model M4

Residuals vs Fitted

Normal Q-Q

Scale-Location

Residuals vs Leverage

Predicted values

Theoretical Quantiles

-8 -6 -4 -2 0 2 4

-2 -1 0 1 2

0.0 0.5 1.0 1.5

0.0 0.5 1.0 1.5

-3 -2 -1 0 1 2 3

-3 -2 -1 0 1 2 3

1.5 2.0 2.5 3.0 3.5 4.0 4.5

1.5 2.0 2.5 3.0 3.5 4.0 4.5

1.5 2.0 2.5 3.0 3.5 4.0 4.5

1.5 2.0 2.5 3.0 3.5 4.0 4.5

Contrast with quasi-Poisson model M4
What if zero is impossible?

Fig. 11.2  Frequency plot of the response variable \( N_{\text{days}} \), the number of days snake carcasses remain on the road. Note that a value of 0 cannot occur.
$$f(y_i; \mu_i | y_i \geq 0) = \frac{\mu^{y_i} \times e^{-\mu_i}}{y_i!}$$

$$f(0; \mu_i) = \frac{\mu^0 \times e^{-\mu_i}}{0!} = e^{-\mu_i}$$

$$f(y_i; \mu_i | y_i > 0) = \frac{\mu^{y_i} \times e^{-\mu_i}}{(1 - e^{-\mu_i}) \times y_i!}$$

$$L = \prod_i f(y_i; \mu_i | y_i > 0) = \prod_i \frac{\mu^{y_i} \times e^{-\mu_i}}{(1 - e^{-\mu_i}) \times y_i!}$$
Library(VGAM)
data(Snakes)
M3A <- vglm(N_days~PDayRain + Tot_Rain + Road_Loc +
PDayRain:Tot_Rain, family = posnegbinomial, data = Snakes)

M4A <- vglm(N_days~PDayRain + Tot_Rain + Road_Loc +
PDayRain:Tot_Rain, family = pospoisson, data = Snakes)
Too many zeroes much more common

Fig. 11.3 Intensity of parasites in cod. This is the same graph as Fig. 11.1, except that only frequencies between 0 and 10 are shown.

try plot(table(rpois(10000,lambda)))
<table>
<thead>
<tr>
<th>Model</th>
<th>Full name</th>
<th>Type of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZIP</td>
<td>Zero-inflated Poisson</td>
<td>Mixture</td>
</tr>
<tr>
<td>ZINB</td>
<td>Zero-inflated negative binomial</td>
<td>Mixture</td>
</tr>
<tr>
<td>ZAP</td>
<td>Zero-altered Poisson</td>
<td>Two-part</td>
</tr>
<tr>
<td>ZANB</td>
<td>Zero-altered negative binomial</td>
<td>Two-part</td>
</tr>
</tbody>
</table>
Zero-altered Models

I am not here, because the habitat is not good!

0 hippos

You didn’t see me! I was just under the water.

I am not here, but the habitat is good!

>0 hippos

You thought I was a crocodile.

Here we are!
Zero-altered Models

\[
f_{\text{ZAP}}(y; \beta, \gamma) = \begin{cases} 
  f_{\text{binomial}}(y = 0; \gamma) & y = 0 \\
  (1 - f_{\text{binomial}}(y = 0; \gamma)) \times \frac{f_{\text{Poisson}}(y; \beta)}{1 - f_{\text{Poisson}}(y = 0; \beta)} & y > 0 
\end{cases}
\]

Get the estimates of $\gamma$ and $\beta$ via maximum likelihood estimation

library(pscl)

f1 <- formula(Intensity ~ fArea * fYear + Length | fArea * fYear + Length)

H1A <- hurdle(f1, dist = "poisson", link = "logit", data = ParasiteCod2)
H1B <- hurdle(f1, dist = "negbin", link = "logit", data = ParasiteCod2)

AIC(H1A, H1B)

Can still have overdispersion in ZIP/ZAP models so check ZINB/ZANB
Zero-inflated Models

You thought I was a crocodile.

You didn't see me! I was just under the water.

I am not here, but the habitat is good!

0 hippos

I am not here, because the habitat is not good!

0 hippos

Here we are!

>0 hippos

Count process

Zero mass
\[ f(y_i = 0) = \pi_i + (1 - \pi_i) \times e^{-\mu_i} \]
\[ f(y_i | y_i > 0) = (1 - \pi_i) \times \frac{\mu^{y_i} \times e^{-\mu_i}}{y_i!} \]

\[ \mu_i = e^{\alpha + \beta_1 \times X_{i1} + \ldots + \beta_q \times X_{iq}} \]

\[ \pi_i = \frac{e^{\nu + \gamma_1 \times Z_{i1} + \ldots + \gamma_q \times Z_{iq}}}{1 + e^{\nu + \gamma_1 \times Z_{i1} + \ldots + \gamma_q \times Z_{iq}}} \]

\[ E(Y_i) = \mu_i \times (1 - \pi_i) \]
\[ \text{var}(Y_i) = (1 - \pi_i) \times (\mu_i + \pi_i \times \mu_i^2) \]
> library(pscl)
> f1 <- formula(Intensity ~ fArea*fYear +
>           Length | fArea * fYear + Length)
> Zipl <- zeroInfl(f1, dist = "poisson",
>               link = "logit", data = ParasiteCod2)