

# Chapter 10

## Poisson Regression

This chapter concerns poisson regression analysis for count data, or rates. The response variable in this situation is a quantitative variable, but has the property that it is discrete, taking on only integer values. The basic idea for this model is that the predictor information is related to the rate or susceptibility of the response to increase or decrease in counts.

### 10.1 Introduction to Poisson Distribution

A poisson random variable  $Y$  has probability density function,  $f(y) = P(Y = y)$  give as,

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!},$$

where the parameter  $\lambda$  is the mean value of the random variable  $Y$ ,  $E(Y) = \lambda$ . This random variable takes on values from zero to infinity, at integers. Larger values of the mean parameter  $\lambda$  will produce greater random variable values.

### 10.2 Poisson Regression Model

The basic model formulation is that the mean of the poisson random variable is a function of predictor information,

$$E(Y) = \mu = \lambda = \exp\{\alpha + \beta_1 x_1 + \dots + \beta_p x_p\},$$

because the log of this function produces a linear combination of the predictors, this model is said to have a log link function; the function that links the mean to the linearized predictor is the log function.

Consider a simple model with a single predictor  $x$ , we have,  $E(Y) = \lambda = \mu = \exp\{\alpha + \beta x\}$ ; and this function can be re-written as  $\exp\{\alpha\}(\exp\{\beta\})^x$ . When we consider a one unit increase in the predictor  $x$  we now have a mean function,

$$E(Y | X + 1) = \exp\{\alpha\}(\exp\{\beta\})^{x+1} = \exp\{\alpha\}(\exp\{\beta\})^x(\exp\{\beta\}) = E(Y | X)(\exp\{\beta\}),$$

so that the mean at  $X + 1$  is simply the mean at  $X$  multiplied by  $\exp\{\beta\}$  so the impact of a unit change is a multiple of the previous mean. This is important to remember, the poisson model has interpretation that is a multiple.

When a response count  $Y$  has an index  $t$  like population size or some other risk measure then,

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

by moving the  $\log(t)$  term to the right side of the equation, we get  $\log(\mu) = \log(t) + \alpha + \beta x$ , the  $\log(t)$  is called the offset. This tells us that the mean is proportional to the index  $t$ . Thus for a fixed  $x$ , doubling the population size would double the response  $Y$  of say the number of auto thefts.

### 10.2.1 Poisson Regression Model for Rates

A common situation for poisson regression is with a response that are counts or deaths in group  $j$  stratum  $i$ . We denote this response as  $y_{ij}$ , and the mortality or event rate is  $r_{ij} = y_{ij}/p_{ij}$ , where we denote person-years accumulated in group  $j$  stratum  $i$  as  $p_{ij}$ . We denote the poisson mean as  $E(Y_{ij}) = \lambda_{ij} = p_{ij} \exp(\beta'x)$ , where  $\beta'x$  defines the indicators for group and stratum.

#### Standardized Mortality Ratio

In this situation we often wish to determine the effects of groups or exposure levels, we do this through a ratio called the standardized mortality ratio (SMR). A SMR is defined as the ratio of two rates,

$$SMR = \frac{r_2}{r_1},$$

so that if a  $SMR = 2$  it would mean that numerator group 2 has twice the rate of denominator group 1. If we wish to compare two strata  $l$  and  $m$ , in an exposure group  $i$ , we compute,

$$SMR = \frac{r_{il}}{r_{im}} = \frac{e^{\alpha + \beta_{gr=i} + \beta_{str=l}}}{e^{\alpha + \beta_{gr=i} + \beta_{str=m}}} = e^{\beta_{str=l} - \beta_{str=m}},$$

To form confidence intervals for the true SMR we find the estimate and standard error for the linear combination of predictors and then exponentiate the confidence bounds to obtain bounds for SMR.

### 10.2.2 Hodgkins Rates in California

In this example we have data on Hodgkins rates for males and females in California in several age categories. The SAS program and output are given here. Note that it is necessary to specify link be log and dist to be poisson. A plot of the observed number of deaths versus the age group is given in Figure 10.1. This plot is not adjusted for the person-years of exposure, but shows clearly the higher risk of males vs females. A plot that does display correct, person-years-adjusted mortality rates, is displayed in Figure 10.2. This also shows clear evidence for higher mortality rates at every age for males than females.

```
* Hodgkins Disease mortality data ;
* California 1989 data, sex=1=males, sex=2=females;
* roughly 5 year age intervals starting at 30-34 to over 85 ;
* Page 470 Selvin ;
proc format ;
  value agecat 1='30-34'
              2='35-39'
              3='40-44'
```

```

4='45-49'
5='50-54'
6='55-59'
7='60-64'
8='65-69'
9='70-74'
10='75-79'
11='80-84'
12='85+' ;
value sexval 1='Males'
2='Females';

run;
data cal89;
input age sex peryr deaths;
lpyr=log(peryr);
format age agecat. sex sexval.;
datalines;
1 1 1299868 50
etc
;
proc genmod data=cal89;
class age sex;
model deaths = age sex / dist=poisson link=log offset=lpyr type3;
estimate 'sex smr' sex 1 -1 / exp;
output out=jon pred=predicted resdev=resid;
run;

```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	11	78.9574	7.1779
Scaled Deviance	11	78.9574	7.1779
Pearson Chi-Square	11	440.6367	40.0579
Scaled Pearson X2	11	440.6367	40.0579
Log Likelihood		978.7664	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-10.6274	0.3844	-11.3809	-9.8739	764.23	<.0001
age 30-34	1	0.5087	0.3941	-0.2637	1.2812	1.67	0.1968
age 35-39	1	0.4536	0.3958	-0.3222	1.2295	1.31	0.2518
age 40-44	1	0.3715	0.4003	-0.4130	1.1561	0.86	0.3533
age 45-49	1	0.6695	0.4171	-0.1481	1.4870	2.58	0.1085
age 50-54	1	-0.0117	0.4269	-0.8484	0.8250	0.00	0.9781
age 55-59	1	0.1616	0.4222	-0.6659	0.9890	0.15	0.7019
age 60-64	1	0.2485	0.4194	-0.5735	1.0705	0.35	0.5535
age 65-69	1	0.6032	0.4106	-0.2015	1.4080	2.16	0.1418

age	70-74	1	0.6288	0.4192	-0.1929	1.4505	2.25	0.1336
age	75-79	1	0.7216	0.4282	-0.1176	1.5608	2.84	0.0919
age	80-84	1	0.6958	0.4580	-0.2020	1.5935	2.31	0.1288
age	85+	0	0.0000	0.0000	0.0000	0.0000	.	.
sex	Females	1	-0.4159	0.0980	-0.6079	-0.2239	18.02	<.0001
sex	Males	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
age	11	18.68	0.0670
sex	1	18.28	<.0001

Label	Estimate	Standard Error	Confidence Limits	Chi-Square	Pr > ChiSq
sex smr	-0.4159	0.0980	-0.6079    -0.2239	18.02	<.0001
Exp(sex smr)	0.6598	0.0646	0.5445    0.7994		

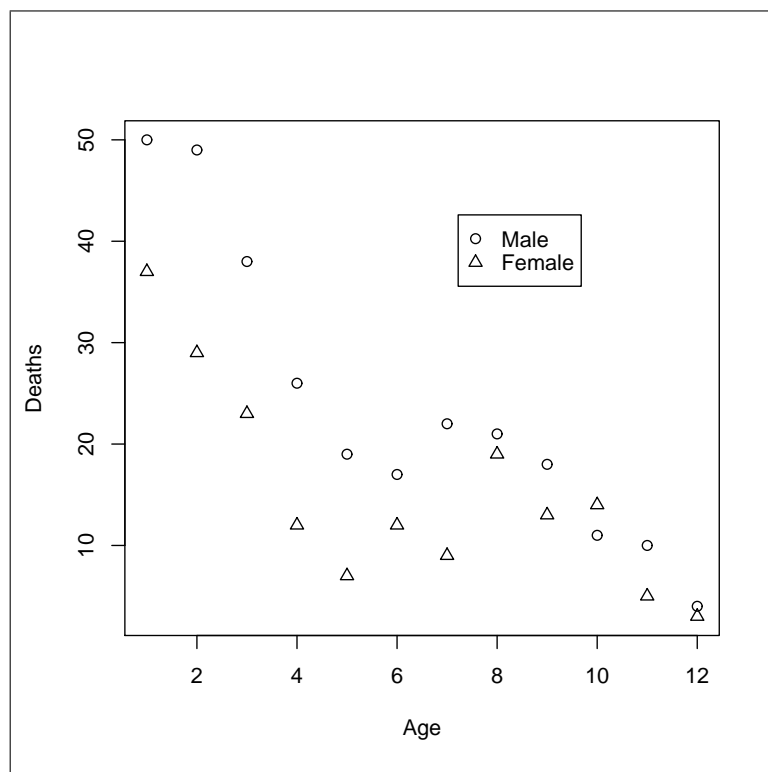


Figure 10.1: Observed Data Plot for Hodgkins Data (1=Male, 2=Female).

From the estimated output the mortality rate for females between 65 and 69 years is represented by,

$$\log(\hat{r}_{82}) = -10.6274 + .6032 + (-.4159) = -10.4401,$$

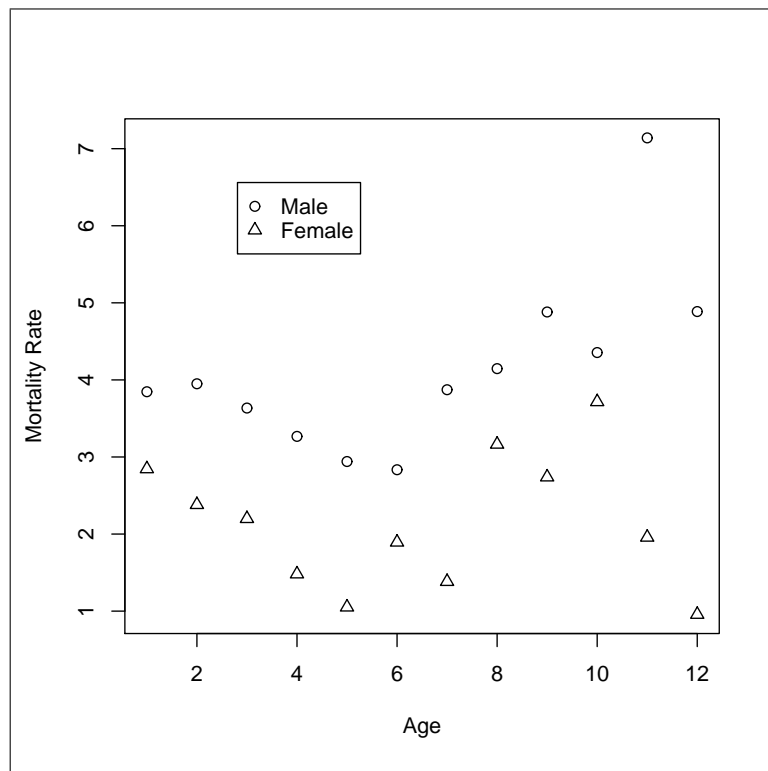


Figure 10.2: Observed Mortality Rates Plot for Hodgkins Data (1=Male, 2=Female).

and exponentiating this gives an estimate rate of .0000292 deaths per person-year. In more convenient units, this implies 2.92 deaths per 100 thousand person-years. The expected number of deaths or events for this observation is found by taking the estimated rate and multiply by the observed person-years  $p_{82} = 600455$  to obtain 17.555 deaths estimated for this observation, that had observed 19 events in this group. The estimated SMR for females relative to males is given by,  $\exp\{\hat{\beta}_{fem}\} = \exp(-.4159) = .6598$ , so that females are .6598 times as likely as males to have a Hodgkins death at any age. The confidence interval for this SMR is given in the output as (.5445, .7994), showing clear evidence that females have less risk of Hodgkins death than males at any age.

### 10.3 Model Inference

We have demonstrated how to perform inference for the SMR, and inference for individual model parameters is done as usual with confidence intervals or hypothesis tests provided in the output. Tests about sets of model terms are conducted using likelihood ratio tests. For example if we wished to test the null hypothesis that the sex of the subjects does not matter in Hodgkins mortality in CA after adjusting for age groups, we would form two models: one with just age terms, and one with both age and sex terms. The difference between the  $-2 \cdot \log$  likelihoods would form the test statistic. This kind of analysis is done automatically in SAS using the type3 option in the model statement. From the output given, we find a small p-value for this test for sex, suggesting strong evidence that the mortality rates for Hodgkins differ by sex, even after adjusting for age categories. The test for age is not as convincing. The degrees of freedom for these type3 tests is simply the difference in the number of

parameters estimated in the null and alternative hypothesis models.

Inference for arbitrary linear functions of model parameters is formed with the contrast statement in SAS. The terms involved are listed with appropriate multipliers to the parameter estimates. If it is needed to exponentiate the resulting functions, the `exp` option will present these results directly in the output along with confidence limits.

## 10.4 Model Goodness of Fit

Most goodness of fit methods for poisson regression models depends on the residual. The definition of a residual for observation  $i$  is,

$$resid_i = \frac{(y_i - \hat{y}_i)}{\sqrt{\hat{y}_i}},$$

where  $\hat{y}_i$  is the predicted count for observation  $i$  based on the poisson regression model. Large, positive values mean that the observed count was much larger than what was predicted by the model. A large, negative residual means the observed response was much less than that predicted by the model. A plot of observed versus predicted values will be helpful in identifying outliers and other odd observations. The plot in Figure 10.3 shows the deviance residuals for the Hodgkins data versus the predicted values, with different plot symbols for males and females. There is nothing particularly noteworthy about this plot. The model appears to model the data reasonably well, with no obvious outliers. The plot in Figure 10.4 shows that the observed number of deaths closely matches the predicted values from our model.

Global goodness of fit statistics of the null hypothesis, model fits; alternative hypothesis, doesn't fit can be found by using the Pearson chi-squared and deviance test statistics given in the SAS Proc GENMOD output. Large values of these statistics, and small p-values imply evidence that the model does not fit the observed data.

A common type of problem with poisson regression models is that the variation in the response is not equal to the mean, as the model assumes, but rather more variable than the mean. This potential violation can be observed through examining chi-squared test statistics divided by the degrees of freedom. If the model dispersion holds, and follows a poisson pattern this ratio should be approximately one, and larger than one for over-dispersed poisson counts. In our California Hodgkins data, we observe evidence for over-dispersed data with respect to the estimated main effect model.

## 10.5 Producing R Plots From SAS Output

The R commands used to produce the graphs in this section are given below. An output dataset was exported from SAS containing the residuals and predicted values.

```
> jon<-read.csv("jon.csv", header=T)
> attach(jon)
> plot(age, deaths, xlab="Age", ylab="Deaths", pch=as.numeric(sex))
> legend(locator(n=1), legend=c("Male","Female"),pch=1:2)
> plot(predicted,resid, xlab="Predicted Values", ylab="Dev Residual",
+ pch=as.numeric(sex))
> legend(locator(n=1), legend=c("Male","Female"),pch=1:2)
> plot(predicted,deaths, xlab="Predicted Values", ylab="Deaths",
+ pch=as.numeric(sex))
```

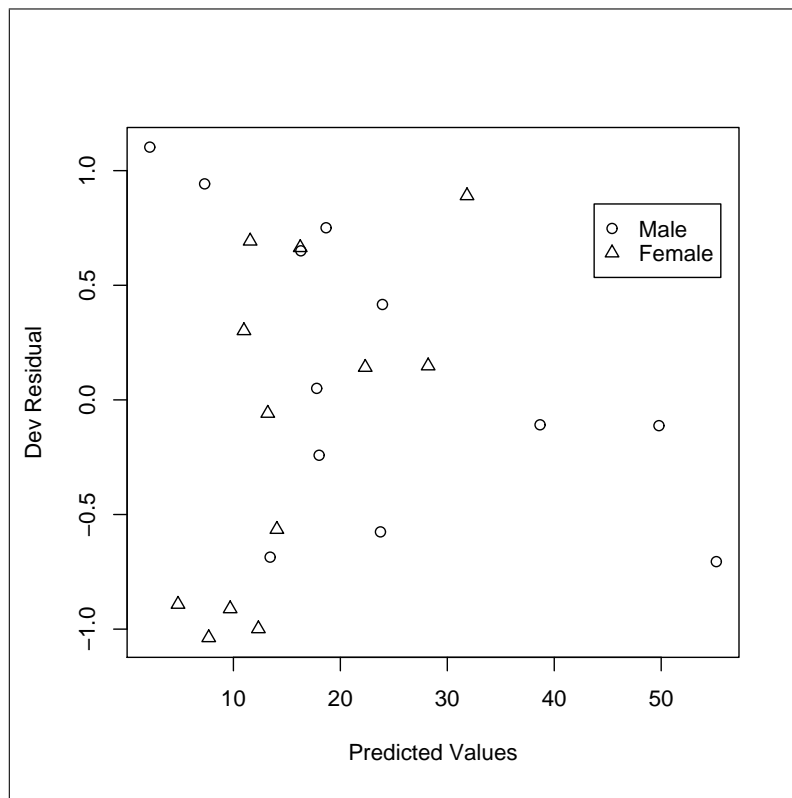


Figure 10.3: Residual Plot for Hodgkins Data.

```
> legend(locator(n=1), legend=c("Male","Female"),pch=1:2)
> mort_rate<-(deaths/peryr)*100000
> plot(age, mort_rate, xlab="Age", ylab="Mortality Rate", pch=as.numeric(sex))
> legend(locator(n=1), legend=c("Male","Female"),pch=1:2)
```

## 10.6 R Example

The R output given below shows the analysis commands for producing a poisson regression analysis of species count data modeled as a function of biomass and a grouping factor for soil ph. We produce a scatterplot with model fit, and a series of diagnostic plots from the glm function. There is no evidence of lack of fit of this estimated poisson regression model.

```
# Page 237 Crawley, number of different plant species on plots
# that have different biomass and different soil ph group levels
species<-read.table("species.txt",header=T)
attach(species)
names(species)
plot(Biomass,Species,type="n")

spp<-split(Species,pH)
```

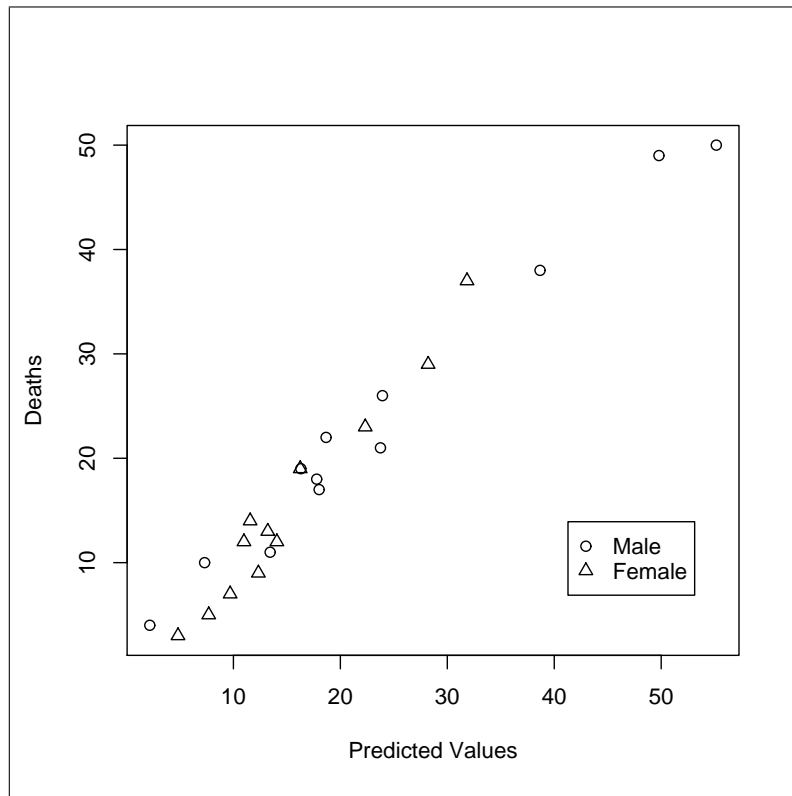


Figure 10.4: Observed vs Predicted Plot for Hodgkins Data.

```

bio<-split(Biomass,pH)

points(bio[[1]],spp[[1]],pch=16)
points(bio[[2]],spp[[2]],pch=17)
points(bio[[3]],spp[[3]])

legend(8.5,45, c("High","Mid","Low"), pch=c(16,1,17))

model1<-glm(Species~Biomass*pH,poisson)
summary(model1)
Call:
glm(formula = Species ~ Biomass * pH, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.49779  -0.74845  -0.04023   0.55745   3.22975

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.76812    0.06153  61.240 < 2e-16 ***

```

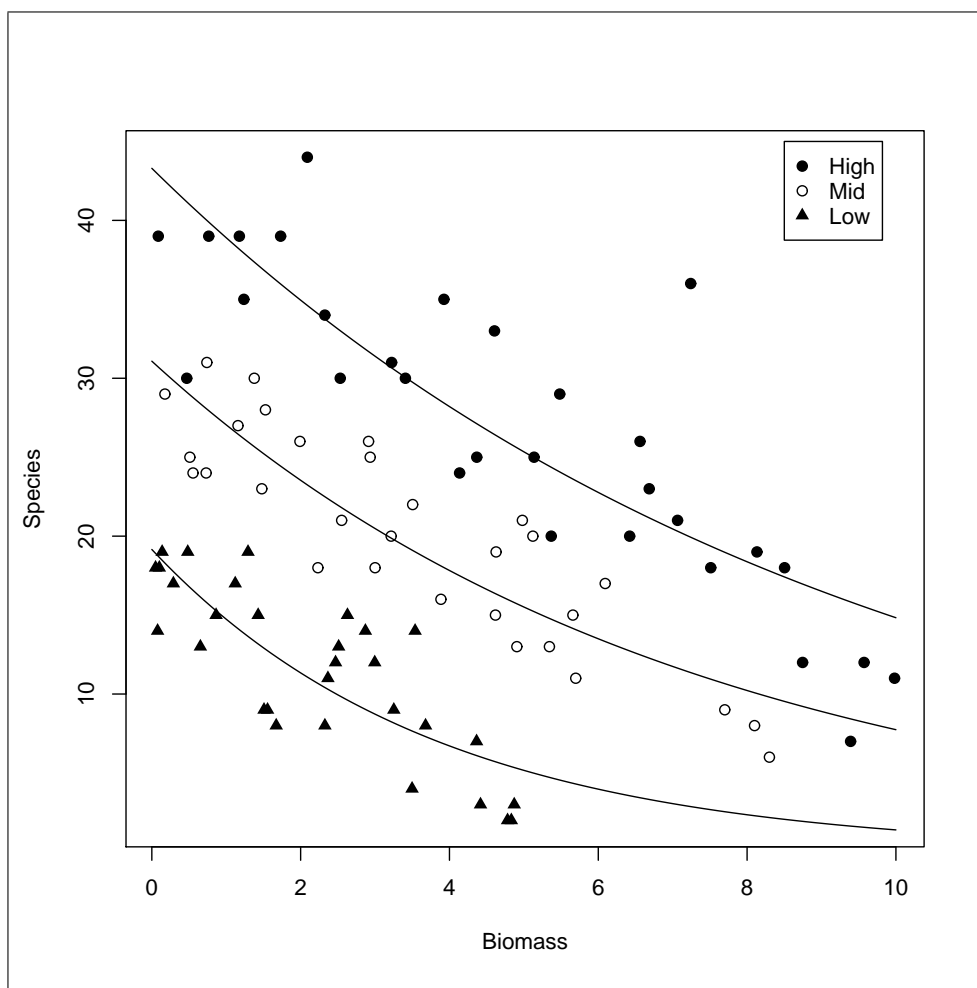


Figure 10.5: Data Plot of Species by Biomass with Predicted Values.

```

Biomass      -0.10713    0.01249   -8.577 < 2e-16 ***
pHlow        -0.81557    0.10284   -7.931 2.18e-15 ***
pHmid        -0.33146    0.09217   -3.596 0.000323 ***
Biomass:pHlow -0.15503    0.04003   -3.873 0.000108 ***
Biomass:pHmid -0.03189    0.02308   -1.382 0.166954

```

---

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

```
model2<-glm(Species~Biomass+pH,poisson)
```

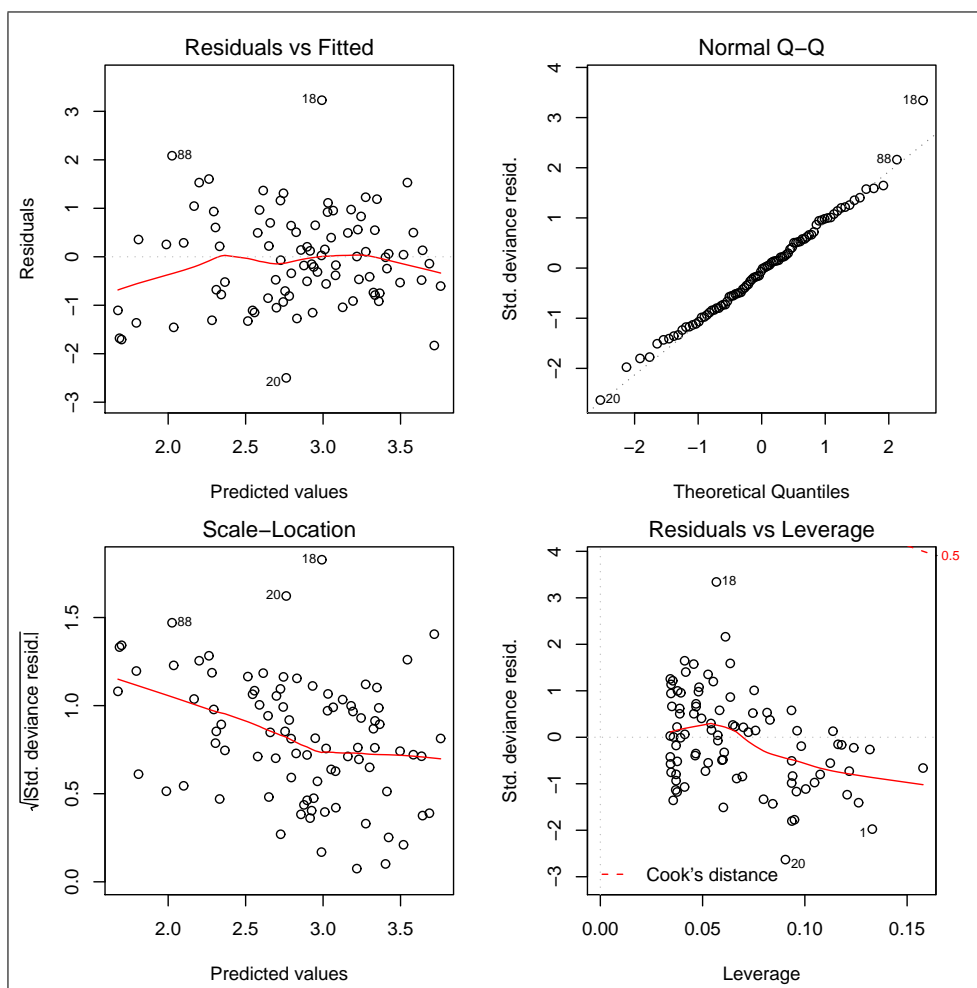


Figure 10.6: Fitted glm Model Object Plots.

```
summary(model2)
glm(formula = Species ~ Biomass + pH, family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.59586	-0.69887	-0.07373	0.66472	3.56040

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.84894	0.05281	72.885	< 2e-16 ***
Biomass	-0.12756	0.01014	-12.579	< 2e-16 ***
pHlow	-1.13639	0.06720	-16.910	< 2e-16 ***
pHmid	-0.44516	0.05486	-8.114	4.88e-16 ***

---

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: 99.242 on 86 degrees of freedom
AIC: 526.43
```

```
Number of Fisher Scoring iterations: 4
```

```
# The anova output given below shows evidence we need the interaction
anova(model2, model1, test="Chisq")
```

```
Analysis of Deviance Table
```

```
Model 1: Species ~ Biomass + pH
```

```
Model 2: Species ~ Biomass * pH
```

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi )
1	86	99.242			
2	84	83.201	2	16.040	0.0003288

```
xv<-seq(0,10,0.1)
```

```
levels(pH)
```

```
length(xv)
```

```
phv<-rep("high",101)
```

```
yv<-predict(model1,list(pH=factor(phv),Biomass=xv),type="response")
```

```
lines(xv,yv)
```

```
phv<-rep("mid",101)
```

```
yv<-predict(model1,list(pH=factor(phv),Biomass=xv),type="response")
```

```
lines(xv,yv)
```

```
phv<-rep("low",101)
```

```
yv<-predict(model1,list(pH=factor(phv),Biomass=xv),type="response")
```

```
lines(xv,yv)
```

```
# This code produces a 2 by 2 matrix of plots from the plot(model) structure:
```

```
par (mfrow=c(2,2), mar=c(4,4,2,2))
```

```
plot(model1)
```

```
# This code produces an assessment of dispersion:
```

```
model3<-glm(Species~Biomass*pH,quasipoisson)
```

```
summary(model3)
```