Chapter 14: Poisson modelling of count data: Two examples.

STATS 201/8

University of Auckland

Learning Outcomes

In this chapter you will learn about using a Poisson regression GLM to model:

- Earthquake frequencies using earthquake magnitude (numeric) and location (factor) as explanatory variables.
- Snapper counts using location (factor) and reservation status (factor) as explanatory variables.

Section 14.1
Example 1: Earthquake frequency

Earthquake magnitudes

The Gutenberg-Richter law

The Gutenberg-Richter law says that the expected number of earthquakes in a given period of time decreases multiplicatively with their magnitude. The formula is

$$\log_{10} N = a - bM$$

where N is the expected number of earthquakes of magnitude M or more on the Richter scale. Here, a and b are unknown parameters.

The Richter scale is logarithmic (base 10). So, for example, an earthquake that registers 5.0 on the Richter scale has a shaking amplitude 10 times that of an earthquake that registers 4.0. It can be shown that this corresponds to 30 times the release of energy.

FYI, the most powerful earthquake ever recorded was in Chile in 1960, measuring 9.5 on the Richter scale. The largest known seismic events on earth have been caused by asteroid impact, exceeding 13 on the Richter scale.

Earthquake magnitudes...

The Gutenberg-Richter law...

After applying a healthy dash of calculus, this formula can be re-expressed in a form that is more familiar to us

$$E[Y|x] = \exp(\beta_0 + \beta_1 x)$$

where Y is the number of earthquakes having magnitude between $x - \delta$ and $x + \delta$.[1]

E.g., if $x = 6$ and $\delta = 0.125$ then Y is the number of earthquakes between 5.875 and 6.125 in magnitude.

The above formula should look familiar. It is the one we use for a Poisson regression GLM when there is a single numeric explanatory variable x.

[1] In the above formula, β_0 and β_1 depend on a, b and δ in a complicated way that we are not going to concern ourselves with.

Sthn California and Washington earthquakes, 1987–2019

The research question is to quantify the rate of decrease in earthquake frequency (with increasing magnitude) in both Southern California (SC) and the State of Washington (WA), and to assess whether these rates are the same.

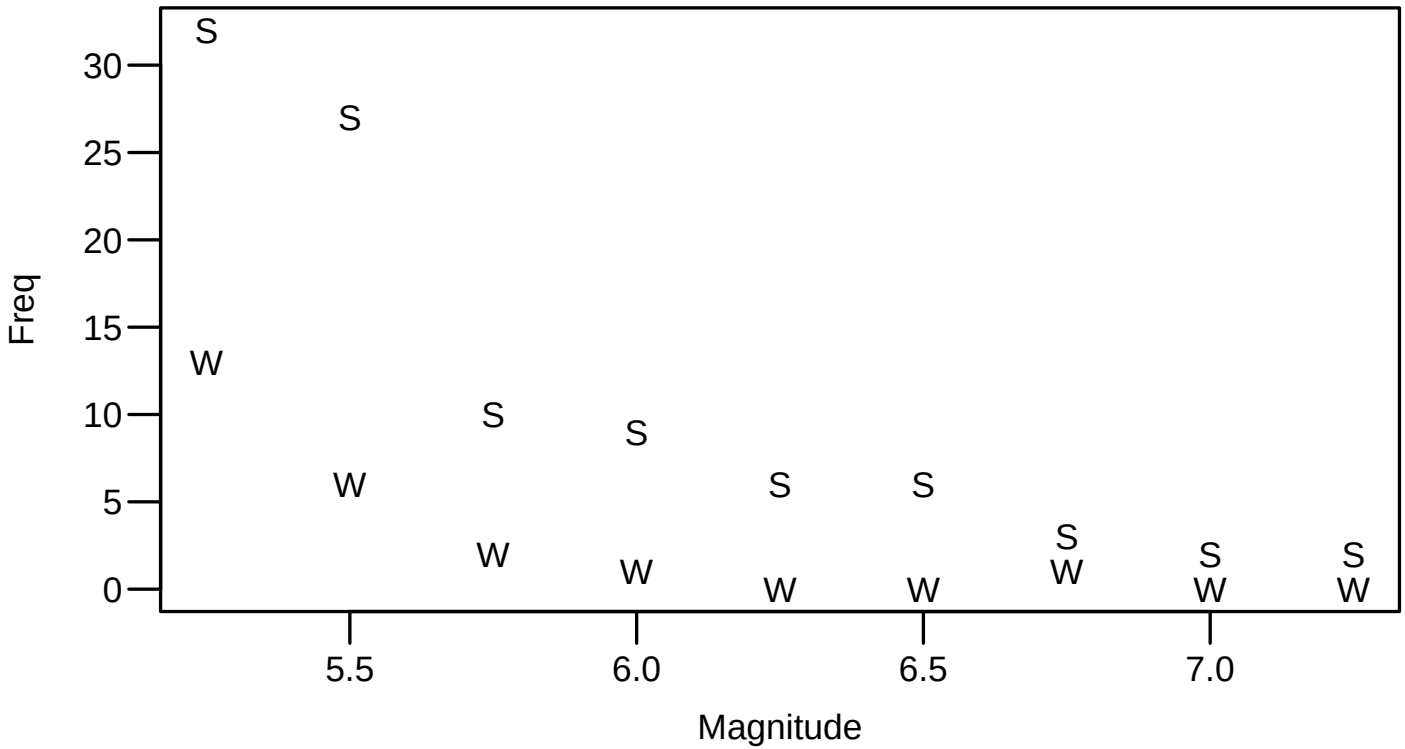
```
> Quakes.df=read.table("Data/EarthquakeMagnitudes.txt",header=TRUE)
> Quakes.df$Locn=as.factor(Quakes.df$Locn)
> subset(Quakes.df,subset=c(Locn=="SC"))[1:4,] #Print first 4 SC observations
  Locn Magnitude Freq
1   SC      5.25   32
2   SC      5.50   27
3   SC      5.75   10
4   SC      6.00    9
> subset(Quakes.df,subset=c(Locn=="WA"))[1:4,] #Print first 4 WA observations
   Locn Magnitude Freq
10   WA      5.25   13
11   WA      5.50    6
12   WA      5.75    2
13   WA      6.00    1
```

Here we have one explanatory variable that is a factor variable, and another that is numeric. We have seen this before in Chapter 8, and we handle it in much the same way as before, but using `glm` instead of `lm`.

Sthn CA and WA earthquakes, 1987–2019...

Plotting the data

```
> plot(Freq~Magnitude,data=Quakes.df,pch=substr(Locn,1,1))
```



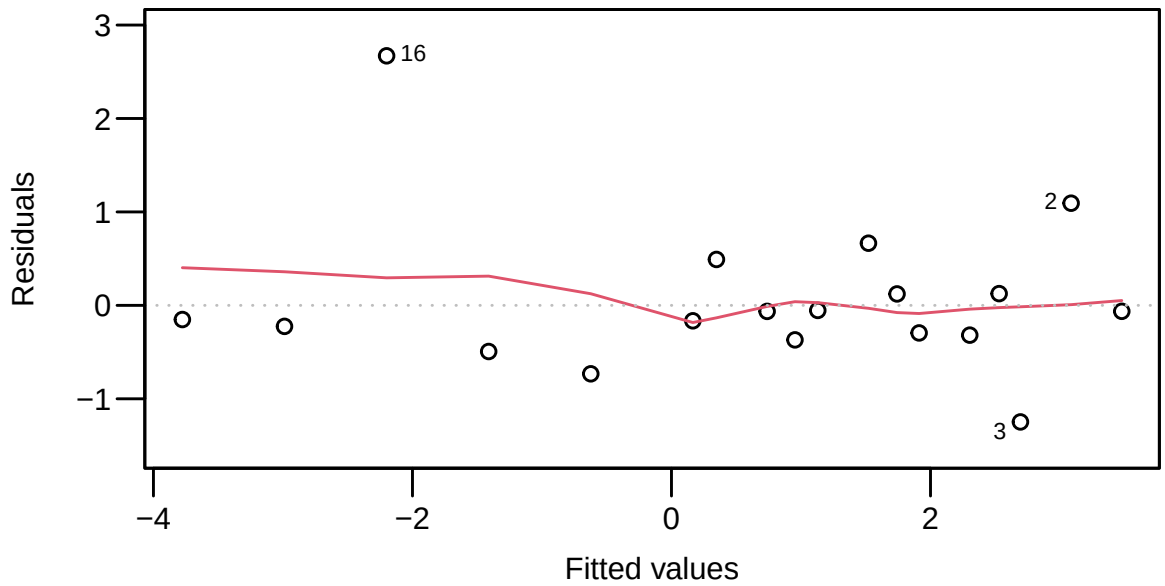
The data look consistent with an exponential decay. It is not clear if the rates of exponential decay are the same.

Sthn CA and WA earthquakes, 1987–2019...

Model fit and assumption checking

Recall from Chapter 8 that we fit the interaction model first, and then simplify if possible.

```
> Quake.gfit = glm(Freq ~ Locn * Magnitude, family = poisson,
+                  data = Quakes.df)
> plot(Quake.gfit, which = 1)
```

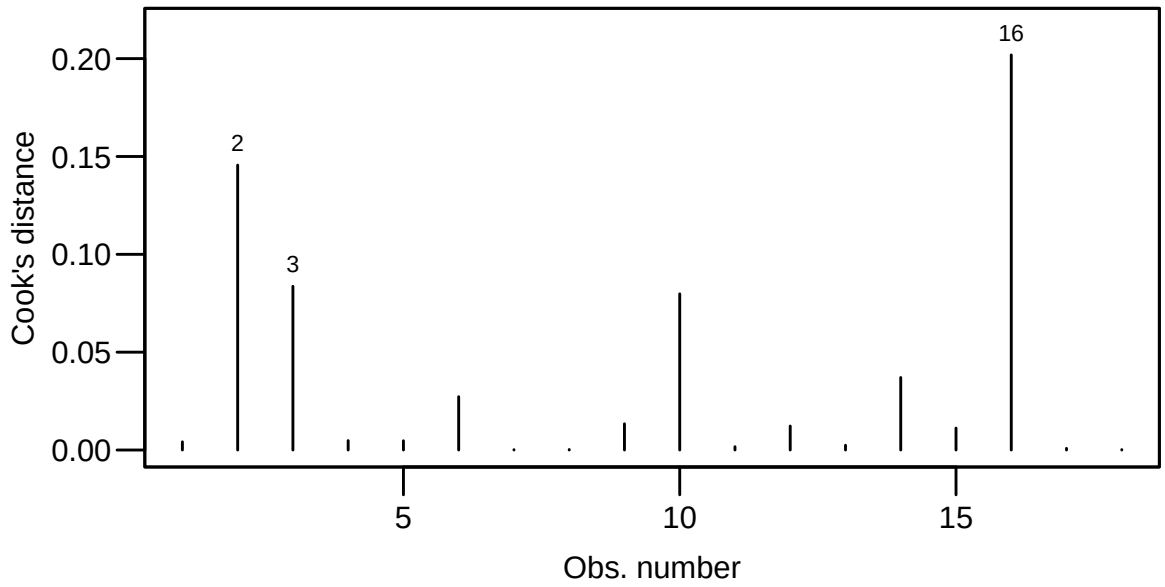


Looks OK, notwithstanding that observation 16 has a high residual. This observation has low expected value (approx $\exp(-2)$), so this residual is not reliable and no cause for concern.

Sthn CA and WA earthquakes, 1987–2019...

Checking influence

```
> plot(Quake.gfit, which = 4)
```



No influential points.

Sthn CA and WA earthquakes, 1987–2019...

Summary output

```
> summary(Quake.gfit)
```

```
Call:
glm(formula = Freq ~ Locn * Magnitude, family = poisson, data = Quakes.df)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      11.6923     1.1762   9.941  < 2e-16 ***
LocnWA            7.3923     3.9500   1.871   0.0613 .
Magnitude        -1.5648     0.2055  -7.616 2.61e-14 ***
LocnWA:Magnitude -1.5884     0.7199  -2.206   0.0274 *
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 176.1767  on 17  degrees of freedom
Residual deviance:   8.2295  on 14  degrees of freedom
```

The residual deviance of 8.23 is less than the 14 residual degrees of freedom, so there won't be any problems with the variance check.

```
> 1 - pchisq(8.23, 14)
[1] 0.8770025
```

Sthn CA and WA earthquakes, 1987–2019...

The interaction term P-value is 0.027, so we conclude that the effect of magnitude is different at the two locations.

Next, lets quantify these rates. First, for Southern California:

```
> Quake.cis = confint(Quake.gfit)
Waiting for profiling to be done...
> exp(Quake.cis[3,])
    2.5 %    97.5 %
0.1374743 0.3082437
> ## To interpret as percentage decreases
> 100*(1-exp(Quake.cis[3,]))
   2.5 %   97.5 %
86.25257 69.17563
```

Sthn CA and WA earthquakes, 1987–2019...

We change the baseline to get the rate for Washington:

```
> Quakes.df$Locn2=factor(Quakes.df$Locn,levels=c("WA","SC"))
>
> Quake2.gfit=glm(Freq~Locn2*Magnitude,family=poisson,data=Quakes.df)
> (Quake.WA.ci = exp(confint(Quake2.gfit)[3,]))
Waiting for profiling to be done...
      2.5 %      97.5 %
0.009077661 0.140175445
> ## To interpret as percentage decreases
> 100*(1 - Quake.WA.ci)
   2.5 %   97.5 %
99.09223 85.98246
```

Sthn CA and WA earthquakes, 1987–2019...

Executive Summary

Our Executive Summary would say that the rate of decline in the frequency of earthquakes (with increasing magnitude) is more rapid in WA than CA.

In WA, there is a 86 to 99% drop in the expected number of earthquakes for a one unit increase in their magnitude on the Richter scale. In CA, the decrease is between 69 to 86%.

Section 14.2
Example 2: Snapper counts in and around marine reserves

Snapper counts in and around marine reserves

Baited underwater video (BUV) is an established tool for counting fish such as snapper.

BUV was used at two coastal locations in New Zealand, Leigh and Hahei. Each location has a marine reserve. The BUV was deployed at sites inside the marine reserve, and at sites just outside the reserve. There was a total of 18 sites.

The three variables measured were

- Count of snapper
- Location (Leigh or Hahei)
- Reservation status (Yes or No)

It was of interest to explore the count of snapper with regard to location and reservation status.

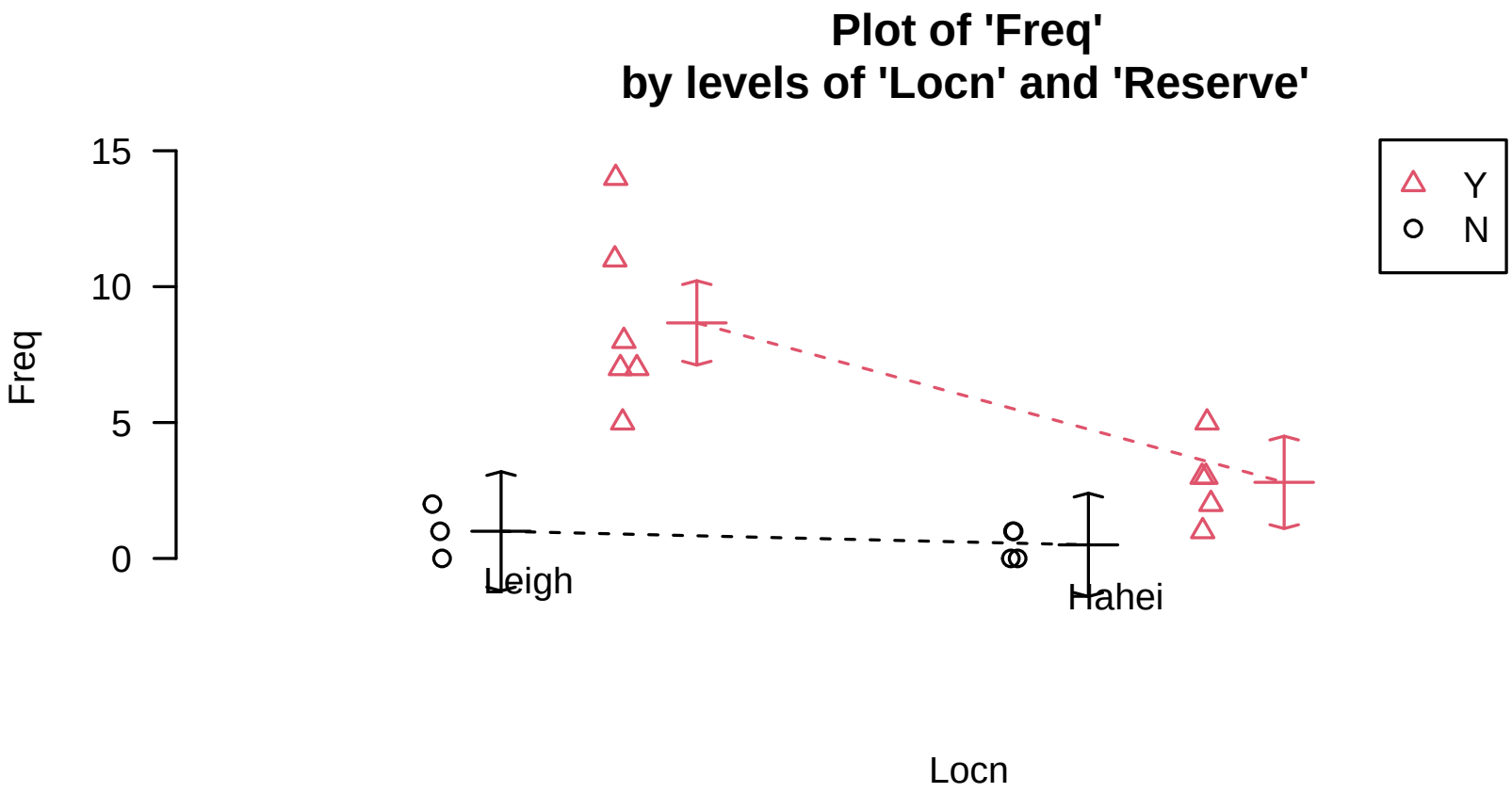
Snapper counts in and around marine reserves...

```
> Snap.df=read.table("Data/SnapperCROPvsHAHEI.txt",header=TRUE)
> with(Snap.df,{Locn=as.factor(Locn); Reserve=as.factor(Reserve)})
> Snap.df
    Locn Reserve Freq
1  Leigh       N    2
2  Leigh       N    1
3  Leigh       N    0
4  Leigh       Y    5
5  Leigh       Y   11
6  Leigh       Y    7
7  Leigh       Y    8
8  Leigh       Y    7
9  Leigh       Y   14
10 Hahei       N    1
11 Hahei       N    0
12 Hahei       N    1
13 Hahei       N    0
14 Hahei       Y    3
15 Hahei       Y    2
16 Hahei       Y    1
17 Hahei       Y    5
18 Hahei       Y    3
```

Snapper counts in and around marine reserves. . .

Plotting the data

```
> interactionPlots(Freq ~ Locn + Reserve, data = Snap.df)
```



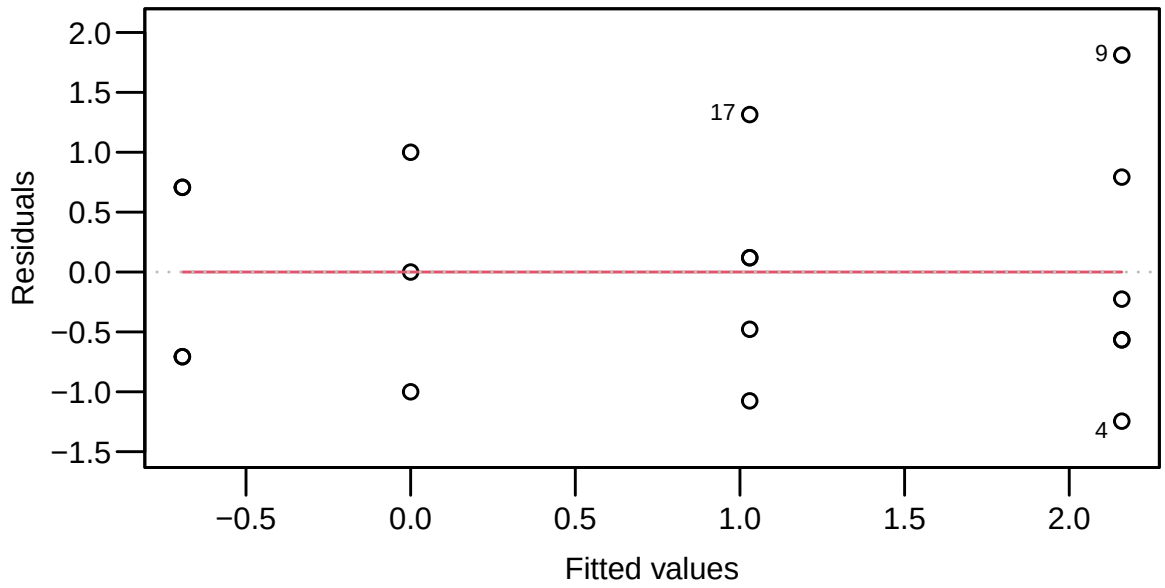
NOTE: Parallel lines no longer corresponds to lack of interaction. **Why?**

Snapper counts in and around marine reserves...

Model fit and assumption checking

There are two categorical explanatory variables, so we follow the steps from Chapter 12. First, we fit an interaction model:

```
> Snap.glm = glm(Freq ~ Locn*Reserve, family = poisson, data = Snap.df)
> plot(Snap.glm, which = 1)
```

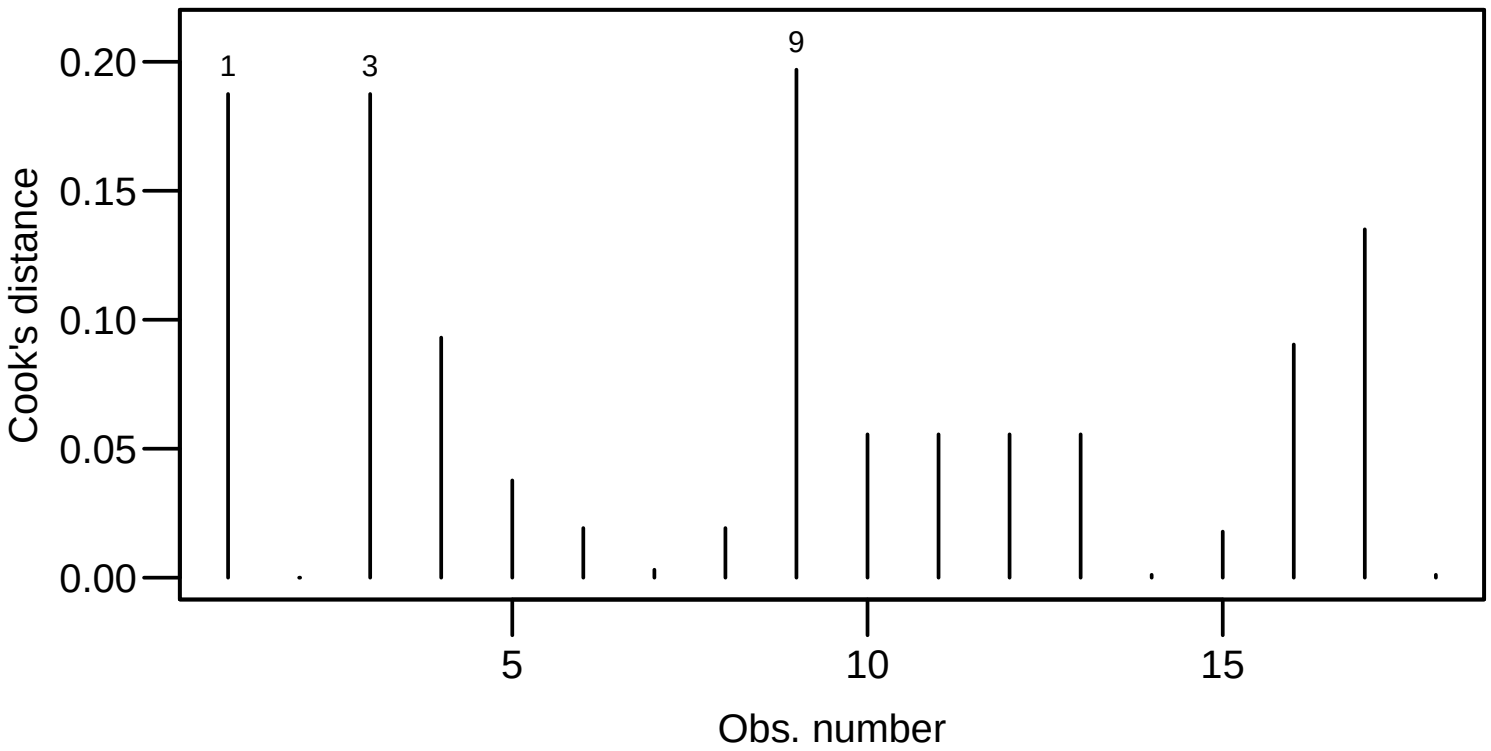


Looks fine.

Snapper counts in and around marine reserves. . .

Influence checking

```
> plot(Snap.glm, which = 4)
```



No overly influential points.

Snapper counts in and around marine reserves...

Assumption checking...

```
> summary(Snap.glm)
```

```
Call:
glm(formula = Freq ~ Locn * Reserve, family = poisson, data = Snap.df)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.6931     0.7071  -0.980   0.3270
LocnLeigh           0.6931     0.9129   0.759   0.4477
ReserveY            1.7228     0.7559   2.279   0.0227 *
LocnLeigh:ReserveY  0.4367     0.9612   0.454   0.6496
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70.453  on 17  degrees of freedom
Residual deviance: 14.678  on 14  degrees of freedom
```

The residual deviance is 14.678 on 14 dof. No problems there.

```
> 1 - pchisq(14.678, 14)
[1] 0.4005141
```

Snapper counts in and around marine reserves. . .

Apply Occam's razor

We see that the interaction between Location and Reserve is not required, so lets redo the `glm` without the interaction.

```
> Snap2.glm = glm(Freq ~ Locn + Reserve, family = poisson, data = Snap.df)
> summary(Snap2.glm)
```

```
Call:
glm(formula = Freq ~ Locn + Reserve, family = poisson, data = Snap.df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9491     0.4884  -1.943 0.051990 .
LocnLeigh     1.0894     0.2845   3.829 0.000128 ***
ReserveY      2.0105     0.4646   4.328 1.51e-05 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70.453  on 17  degrees of freedom
Residual deviance: 14.879  on 15  degrees of freedom
```

Snapper counts in and around marine reserves...

The residual deviance still indicates no evidence of a problem:

```
> 1 - pchisq(14.879, 15)
[1] 0.4601677
```

Lets calculate some confidence intervals, remembering to exponentiate them to get the multiplicative effects of location and reservation status.

```
> (Snap.cis <- exp(confint(Snap2.glm)))
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) 0.1298697  0.9105143
LocnLeigh   1.7443515  5.3626745
ReserveY    3.3224830 21.3481546
```

Snapper counts in and around marine reserves. . .

Executive Summary

We conclude that the expected count of snapper is between 3.3 and 21.3 times as high in marine reserves than in the area just outside of the reserve.

Moreover, the Leigh location has higher expected snapper counts than Hahei, - they are between 1.7 and 5.4 times as high at Leigh.

Closing remark – use of anova with a GLM

In situations where we need to test a joint hypothesis (see Chapter 9) we can continue to use the anova function.

However, be aware that anova for GLMs can use several different methods for calculating the appproximate P-value for the joint hypothesis. We recommend using test="Chisq".

By way of example:"

```
> Snap.anova=anova(Snap2.glm,test="Chisq")
> data.frame(Snap.anova) #Using data.frame removes extraneous output
        Df Deviance Resid..Df Resid..Dev     Pr..Chi.
NULL    NA       NA        17   70.45338           NA
Locn     1 22.65585        16   47.79753 1.937697e-06
Reserve  1 32.91888        15   14.87866 9.608571e-09
```

Note that the P-value for the reserve effect is different from that obtained from summary(Snap2.glm), but both P-values tell the same story – a very highly significant effect of reserve.