Poisson Models

Dr Josh Hodge

Introduction

The lectures introduced you to the concepts of generalised linear models and more specifically Poisson models and how count data is handled in linear models. Remember the three steps of generalised linear models:

1. Choosing a distribution for the response variable that makes assumptions about its error structure (here: Poisson)

2. We specify a linear function of covariates and/or fixed factors

3. Choosing a link function between the predictor function and the mean of the distribution (of the response variable) (here: log-linear)

In this handout, we are going to build on this conceptual knowledge and combine it with your programming skills in the R environment. We will consistently be going through the following steps:

1. Data exploration

2. Model building and fitting

3. Initial interpretations

4. Model validation

5. Model refitting (if necessary)

6. Model interpretation and plotting

Fisheries Data

```r
require(ggplot2)

## Loading required package: ggplot2

## Warning: replacing previous import 'vctrs::data_frame' by
'tibble::data_frame'
## when loading 'dplyr'

require(MASS)

## Loading required package: MASS

require(ggpubr)
```

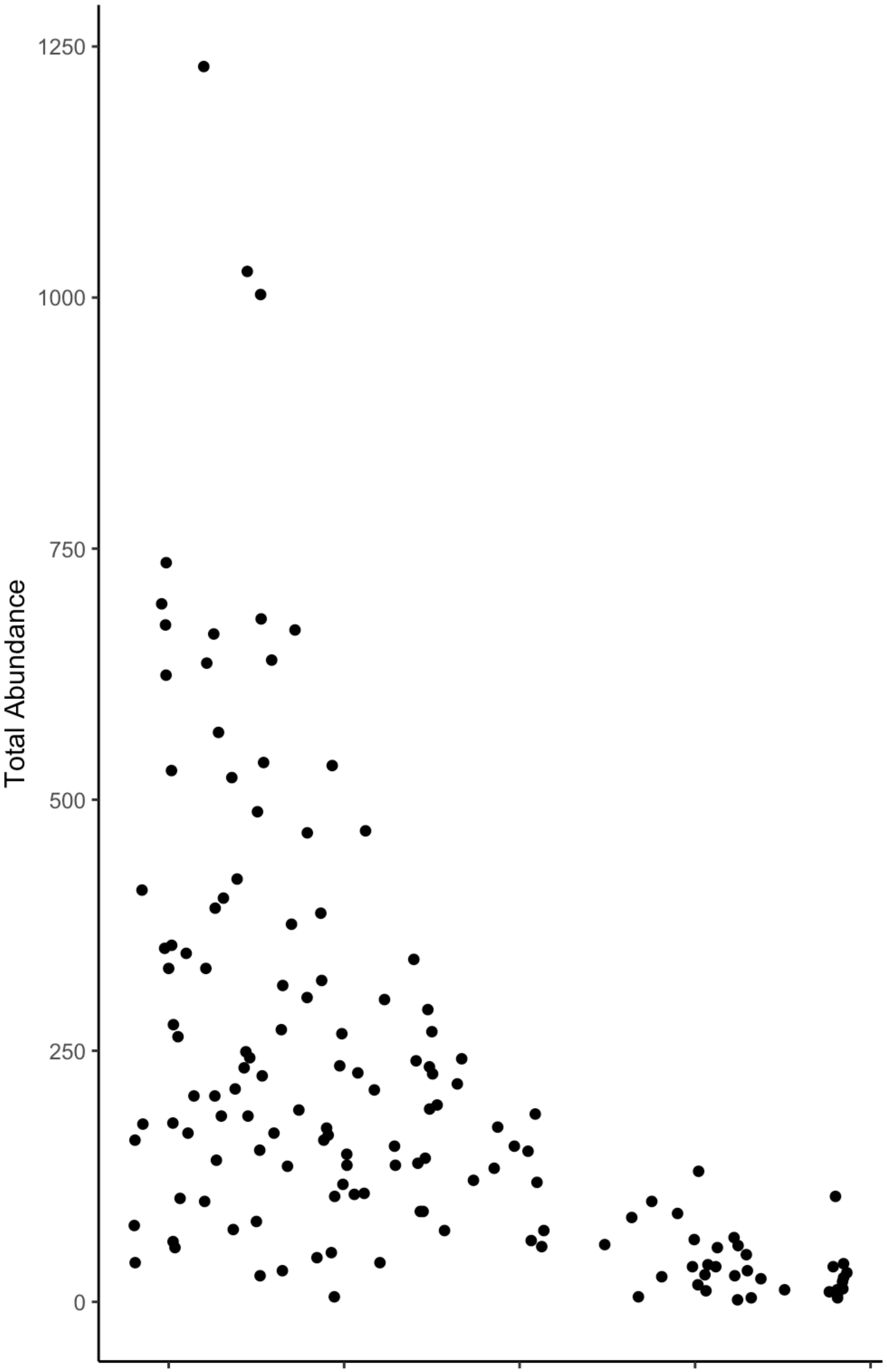
```
## Loading required package: ggpubr
```

```
fish<- read.csv("fisheries.csv", stringsAsFactors = T)
str(fish)
```

```
## 'data.frame':    146 obs. of  10 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Site     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ TotAbund : int  76 161 39 410 177 695 352 674 624 736 ...
##  $ Dens     : num  0.00207 0.00352 0.000981 0.008039 0.005933 ...
##  $ MeanDepth: num  0.804 0.808 0.809 0.848 0.853 0.96 0.977 0.982
0.985 0.986 ...
##  $ Year     : int  1978 2001 2001 1979 2002 1980 1981 1979 1982 1980
...
##  $ Period   : int  1 2 2 1 2 1 1 1 1 1 ...
##  $ Xkm      : num  98.8 76.8 103.8 91.5 107.1 ...
##  $ Ykm      : num  -57.5 178.6 -50.1 146.4 -37.1 ...
##  $ SweptArea: num  36710 45741 39775 51000 29831 ...
```

The dataset includes abundance data for a whole range of unique sites from 1977 to 2002. The other variables include density (“Dens”), mean depth in kilometres of the water column (“MeanDepth”), whether the catch was in period 1 (1979-1989) or period 2 (1997-2002) (“Period”), the lengths of the x and y areas sampled (“Xkm”, “Ykm”) and the total area sampled/swepted in at each site (“SweptArea”). For our initial analyses we are going to investigate whether total abundance changes with mean depth of the water column. See the initial scatterplot:

```
ggplot(fish, aes(x=MeanDepth, y=TotAbund))+
  geom_point()+
  labs(x= "Mean Depth (km)", y="Total Abundance")+
  theme_classic()
```


Total Abundance
1250
1000
750
500
250
0
Mean Depth (km)
1
2
3
4
5

Fitting the Model

Our basic model consists of total abundance as our response variable and mean depth as our explanatory variable. With the basic model equation:

$$\ln(TotAbund) = \beta_0 + \beta_1 * MeanDepth$$

```
M1<- glm(TotAbund~MeanDepth, data = fish, family = "poisson")
summary(M1)

## 
## Call:
## glm(formula = TotAbund ~ MeanDepth, family = "poisson", data = fish)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -25.544   -6.914   -3.046    3.901   35.744  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  6.64334    0.01273  521.70   <2e-16 ***
## MeanDepth   -0.62870    0.00670  -93.84   <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: 27779  on 145  degrees of freedom
## Residual deviance: 15770  on 144  degrees of freedom
## AIC: 16741
## 
## Number of Fisher Scoring iterations: 5
```

Initial Interpretation

The summary output is very similar to that we are used to for a regular linear model covered last term. We have estimated values for the intercept and slope parameters, standard errors, a *z*-value (synomynous with the *t*-value in the *t*-test) and a *p*-value. The null hypothesis of the *z*-value is that the estimate value is equal to zero with the associated *p*-value informing us the likelihood of this hypothesis. From this summary we can build our initial model equation and infer that as mean depth increases the total abundance of fish decreases. We can also calculate the Pseudo-R^2.

$$\ln(TotAbund) = 6.64 - 0.63 * MeanDepth$$

$$PseudoR^2 = 1 - (15770/27779) = 0.43$$

Model Validation

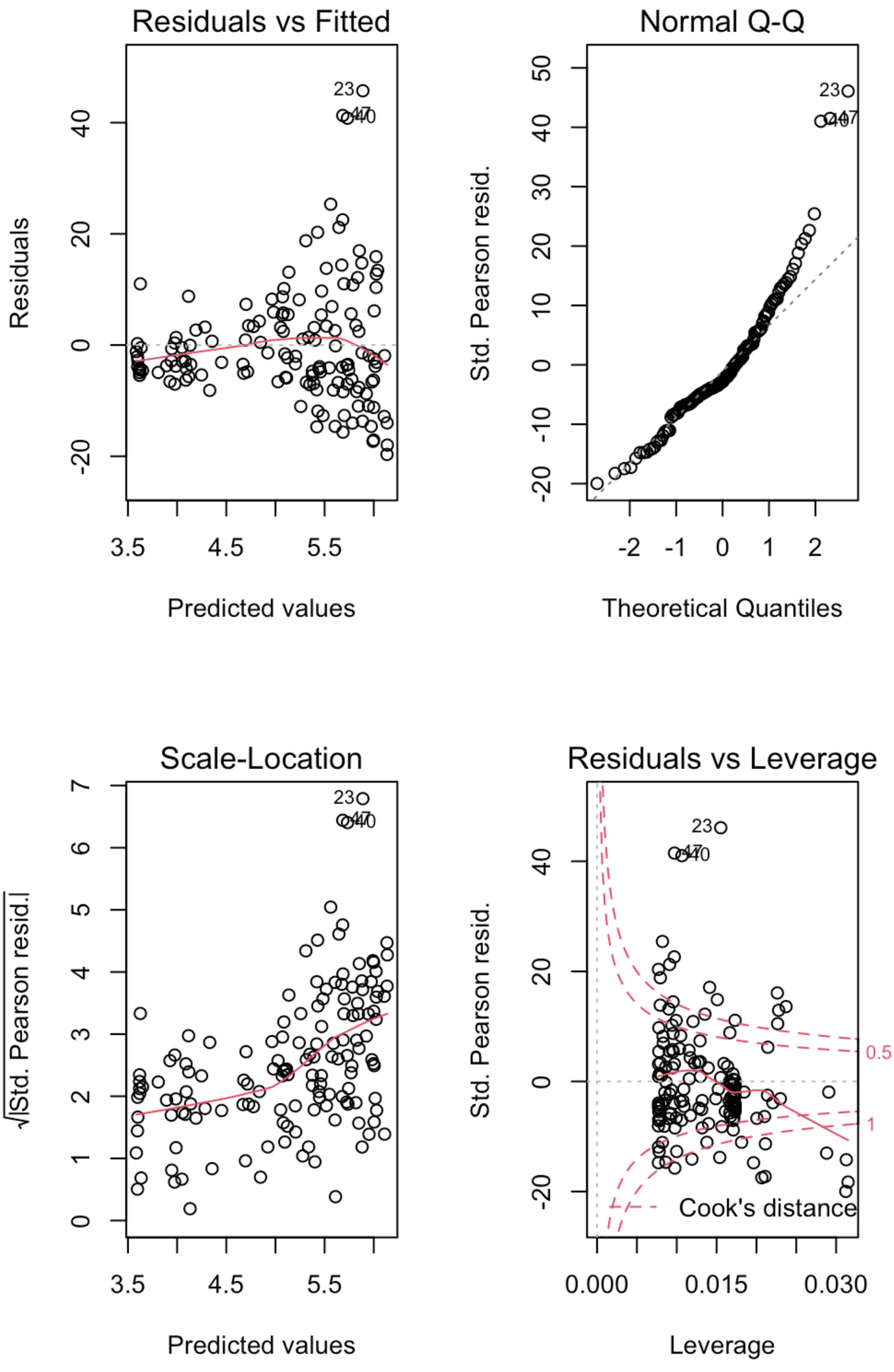
There are two steps in the model validation process:

1. Checking the diagnostics

2. Examining the dispersion parameter

So let’s plot the model diagnostics:

```
par(mfrow=c(2,2)) #partitioning the plot window into a 2X2
plot(M1)
```


Residuals vs Fitted
Residuals
Predicted values
-20
0
20
40
3.5
4.5
5.5
23
47
40
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
-20
-10
0
10
20
30
40
50
-2
-1
0
1
2
23
47
40
Scale-Location
√|Std. Pearson resid.|
Predicted values
0
1
2
3
4
5
6
7
3.5
4.5
5.5
23
47
40
Residuals vs Leverage
Std. Pearson resid.
Leverage
-20
0
20
40
0.000
0.015
0.030
0.5
1
Cook's distance
23
47
40

The “Std.Pearson resid. vs Leverage” plot highlights we potentially have a large number of outliers and can explore these in a little more detail. A Cook’s distance of more than 1 is generally considered to be an outlier so let’s see how many we have.

```
sum(cooks.distance(M1)>1)

## [1] 29
```

We have 29 outliers. If we had only one, we’d consider investigating it and potentially drop it but we do not want to drop over 20 observations so let’s just keep this in-mind for now and examine the dispersion parameter.

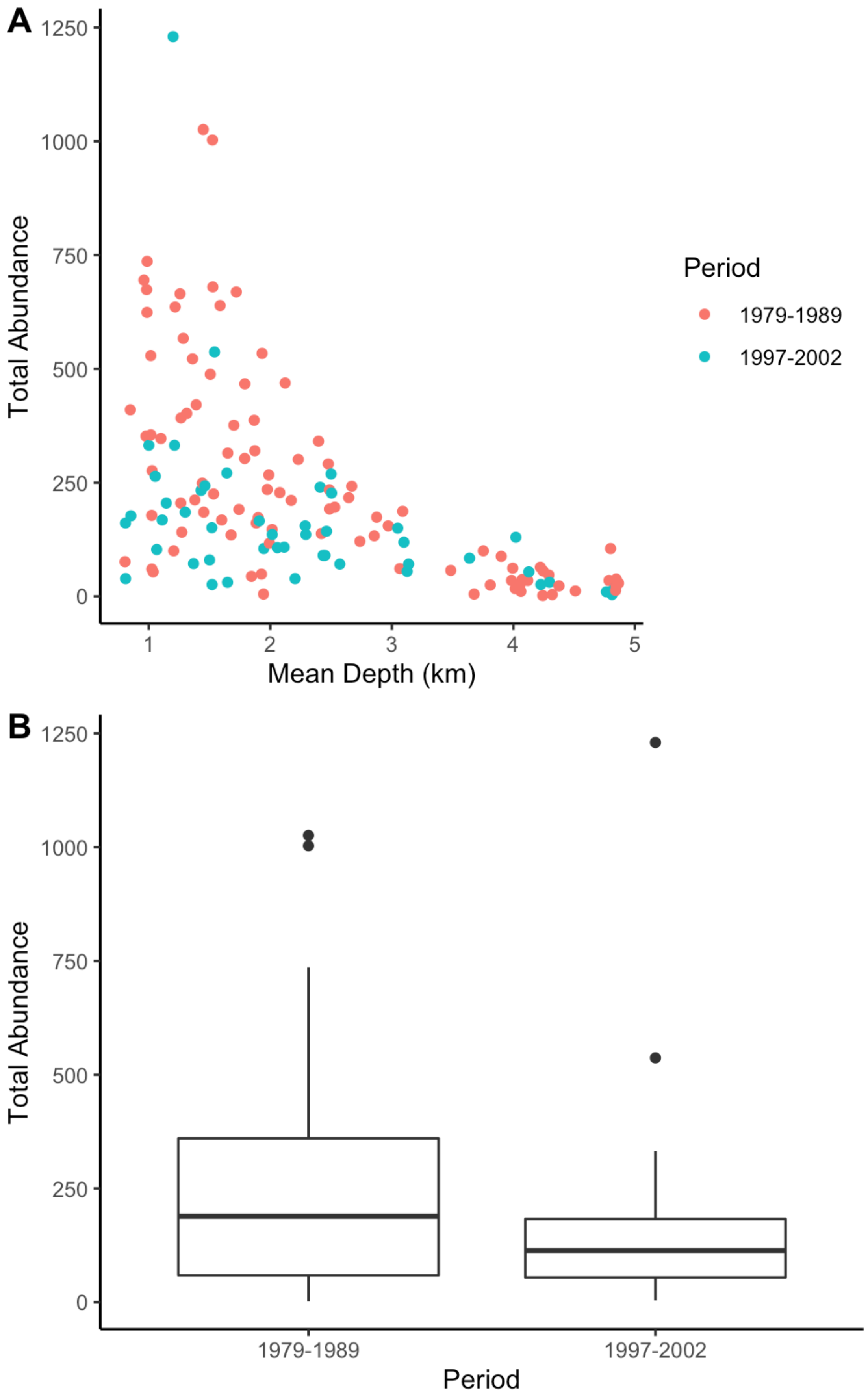
$$DispersionParameter = 15770/144 = 109.51$$

It is quite clear the model is overdispersed. In fact, the conditional variance is 109.51 times higher than the conditional mean and so our model is incredibly overdispersed. There are a number of reasons that could explain this overdispersion, one we have already identified in the outliers, but others might be:

- Transformation of covariates **not needed because we only have one continuous explanatory variable**
- Missing covariates/interactions **this may be a plausible avenue as we have collected other covariates and/or fixed factors**
- Zero-inflation **nope, because we have no zero’s**
- Inherent dependency **potentially, we could explore the random effect of year**

A possible avenue to explore from these options would be including additional covariates and/or fixed factors. We could fit “Year” as a fixed-factor but there are 13 levels and we would lose far too many degrees of freedom, thus statistical power, or we could fit “Period” as a fixed factor, but let’s explore this first through plotting.

```
scatterplot<-ggplot(fish, aes(x=MeanDepth, y=TotAbund,
color=factor(Period)))+
  geom_point()+
  labs(x= "Mean Depth (km)", y="Total Abundance")+
  theme_classic()+
  scale_color_discrete(name="Period", labels=c("1979-1989", "1997-
2002"))
boxplot<- ggplot(fish, aes(x=factor(Period, labels=c("1979-1989",
"1997-2002")), y=TotAbund))+
  geom_boxplot()+
  theme_classic()+
  labs(x="Period", y="Total Abundance")
ggarrange(scatterplot, boxplot, labels=c("A","B"), ncol=1, nrow=2)
```


A
1250
1000
750
500
250
0
Total Abundance
1
2
3
4
5
Mean Depth (km)
Period
● 1979-1989
● 1997-2002
B
1250
1000
750
500
250
0
Total Abundance
1979-1989
1997-2002
Period

From the plots, it looks like there could be a different relationship between MeanDepth and TotAbund (Plot A) and Period 2 has less total abundance than period 1 so let's include "Period" as a fixed factor and interact it with MeanDepth - based on our hypothesis that the affect of MeanDepth on TotAbund is different in both periods. We can examine the impact this has on the dispersion parameter.

Adding Period as a Fixed Factor

```
M2<- glm(TotAbund~MeanDepth*factor(Period), data = fish,
family="poisson")
summary(M2)
```

```
##
## Call:
## glm(formula = TotAbund ~ MeanDepth * factor(Period), family =
"poisson",
##     data = fish)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -25.298   -6.375   -1.721    3.323   44.621
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               6.832036   0.014837 460.473  < 2e-16 ***
## MeanDepth                -0.658858   0.007935 -83.031  < 2e-16 ***
## factor(Period)2          -0.674857   0.029189 -23.120  < 2e-16 ***
## MeanDepth:factor(Period)2  0.115712   0.014908   7.762 8.39e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 27779  on 145  degrees of freedom
## Residual deviance: 14293  on 142  degrees of freedom
## AIC: 15268
##
## Number of Fisher Scoring iterations: 5
```

```
anova(M2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: TotAbund
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       145      27779              
## MeanDepth                  1  12008.9       144      15770 < 2.2e-16
***
## factor(Period)             1   1417.9       143      14352 < 2.2e-16
***
## MeanDepth:factor(Period)  1     58.8       142      14293 1.713e-14
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary output and the analysis of deviance table, we can determine that Period does have a significant impact on the affect of MeanDepth on TotAbund and we can write two linear equations:

$$Period1{:}\ln(TotAbund) = 6.83 - 0.66 * MeanDepth$$

$$Period2{:}\ln(TotAbund) = (6.83 - 0.67)(-0.63 + 0.12) * MeanDepth$$
$$= 6.16 - 0.51 * MeanDepth$$

$$Period2{:}\ln(TotAbund) = 6.16 - 0.51 * MeanDepth$$

Now let's look at the dispersion parameter:

$$DispersionParameter = 14293/142 = 100.65$$

It is clear that we have reduced the dispersion (109.51 to 100.65) but our model is still overdispersed. Here, we two options: 1) fit a quasi-Poisson model or fit a negative binomial model. For this tutorial, we are going to do the latter but for completion you might also want to examine a quasi-Poisson approach.

Fitting a Negative Binomial

```
M3<- glm.nb(TotAbund~MeanDepth*factor(Period), data = fish)
summary(M3)

## 
## Call:
## glm.nb(formula = TotAbund ~ MeanDepth * factor(Period), data = fish, 
##     init.theta = 1.982326313, link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.3257  -0.8038  -0.1655   0.4164   2.7953  
## 
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                 7.02270    0.15928  44.091   <2e-16 ***
## MeanDepth                  -0.75841    0.05979 -12.686   <2e-16 ***
## factor(Period)2            -0.60372    0.27167  -2.222   0.0263 *  
## MeanDepth:factor(Period)2  0.08852    0.10120   0.875   0.3818    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for Negative Binomial(1.9823) family taken to
be 1)
##
##     Null deviance: 335.61  on 145  degrees of freedom
## Residual deviance: 158.27  on 142  degrees of freedom
## AIC: 1752.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.982
##           Std. Err.:  0.223
##
##  2 x log-likelihood:  -1742.201
```

```
anova(M3, test = "Chisq")
```

```
## Warning in anova.negbin(M3, test = "Chisq"): tests made without re-
estimating
## 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(1.9823), link: log
##
## Response: TotAbund
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       145     335.61
## MeanDepth                 1  167.393       144     168.21 < 2.2e-16
***
## factor(Period)            1    9.292       143     158.92  0.002302
**
## MeanDepth:factor(Period)  1    0.653       142     158.27  0.419108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that fitting a negative binomial has changed our model output significantly. The biggest difference is the change in the significance of the model term “MeanDepth:factor(Period)2” suggesting there is no significant difference between the slopes for each period, but there is a significant difference between the intercepts. At this point, we might want to run a reduced model (without the interaction) before moving onto to model validation.

```
M4<- glm.nb(TotAbund~MeanDepth+factor(Period), data = fish)
summary(M4)
```

```
##
## Call:
## glm.nb(formula = TotAbund ~ MeanDepth + factor(Period), data = fish,
##     init.theta = 1.97337518, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3119  -0.7930  -0.0909   0.4223   2.6109
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      6.95563    0.13596  51.159  < 2e-16 ***
## MeanDepth       -0.72986    0.04836 -15.091  < 2e-16 ***
## factor(Period)2 -0.39077    0.12553  -3.113  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.9734) family taken to
be 1)
##
##     Null deviance: 334.13  on 145  degrees of freedom
## Residual deviance: 158.23  on 143  degrees of freedom
## AIC: 1750.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.973
##           Std. Err.:  0.222
##
##  2 x log-likelihood:  -1742.852
```

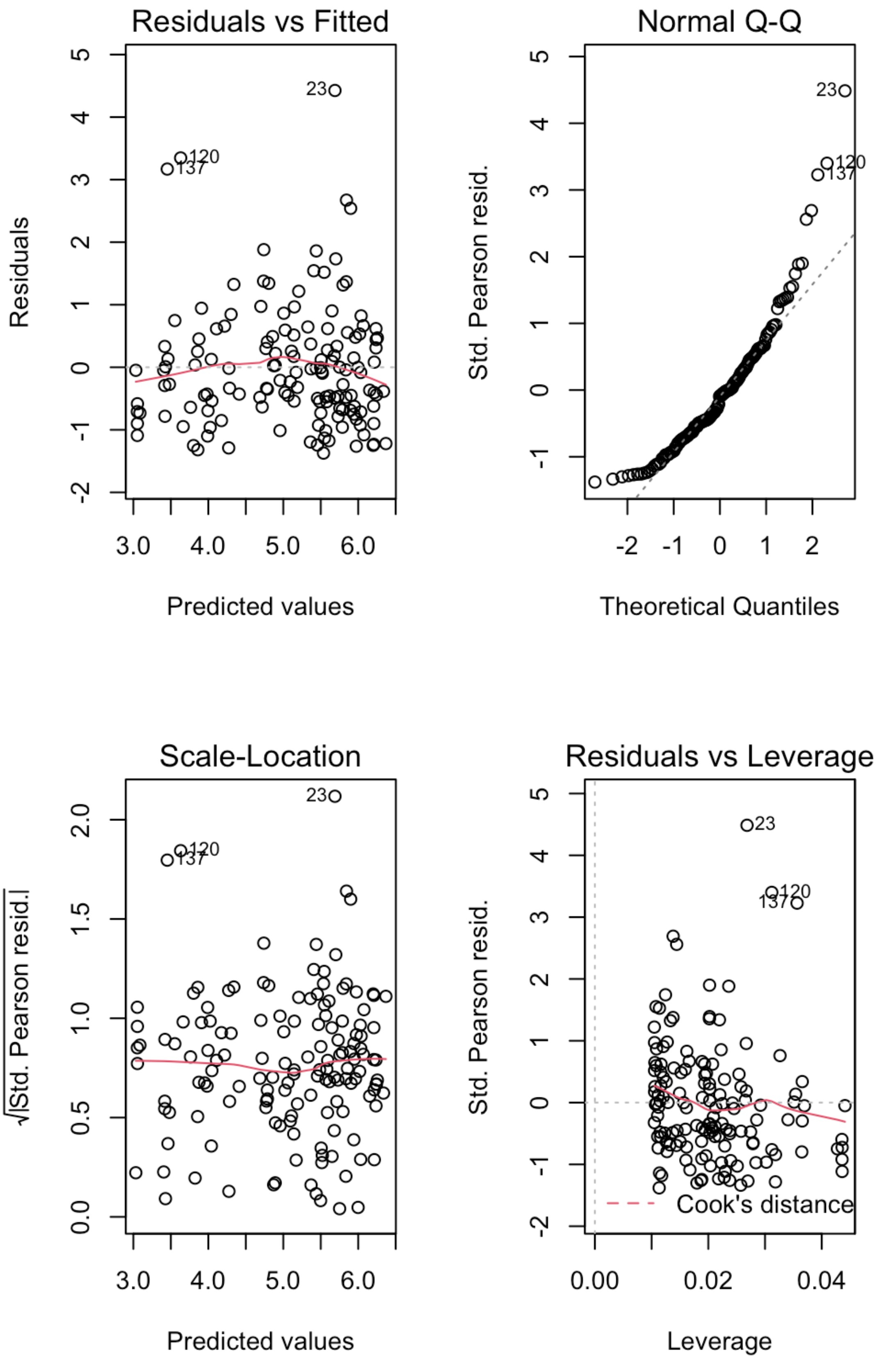
```
anova(M4, test = "Chisq")
```

```
## Warning in anova.negbin(M4, test = "Chisq"): tests made without re-
estimating
## 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(1.9734), link: log
##
## Response: TotAbund
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             145     334.13
## MeanDepth       1   166.65       144     167.48 < 2.2e-16 ***
## factor(Period)  1     9.25       143     158.23  0.002354 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now let's look at the model diagnostics for M4 and the dispersion parameter:

```
par(mfrow=c(2,2)) #partitioning the plot window into a 2X2
plot(M4)
```


Residuals vs Fitted
Residuals
Predicted values
3.0
4.0
5.0
6.0
-2
-1
0
1
2
3
4
5
23
120
137
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
-2
-1
0
1
2
Scale-Location
√|Std. Pearson resid.|
Predicted values
0.0
0.5
1.0
1.5
2.0
Residuals vs Leverage
Std. Pearson resid.
Leverage
0.00
0.02
0.04
Cook's distance

$$DispersionParameter = 158.23/143 = 1.11$$

An interesting point in the diagnostic plots and you can compare this to the original Poisson model, the assumed outliers are no more and this something to always bear in-mind. We never want to quickly remove outliers or potential outliers because they don't fit our first statistical analysis. With GLM's, there is a fair amount of tweaking you can do and find an approach that is suitable for your data. In this instance, we have reduced dispersion consecutively by adding a fixed factor and changing the family of our GLM. The negative binomial has a dispersion parameter of 1.11, which is a significant improvement from our first model. 1.11 isn't perfect but given we have only 146 data points we may conclude that this is acceptable and interpret the negative binomial model. (HINT: there are ways to improve this model with offsets but this beyond what we are examining in this module).

Negative Binomial Interpretation

So finally we have the model we can interpret that has two linear equations with different intercepts but the same slope of MeanDepth.

$$Period1{:}\ln(TotAbund) = 6.96 - 0.73 * MeanDepth$$

$$Period2{:}\ln(TotAbund) = (6.96 - 0.39) - 0.73 * MeanDepth$$
$$= 6.57 - 0.73 * MeanDepth$$

We can therefore interpret that "for every kilometer increase in mean depth total abundance decreased by a factor of *e^-0.73* or 0.48-fold". But what does this mean? Let's look at the exponentiated coefficients for Period 1.

$$For1km: TotAbund = 1053.63 * 0.48 = 505.74$$

$$For2km: 505.74 * 0.48 = 242.76$$

$$For3km: 242.76 * 0.48 = 116.52$$

I know the multiplicative nature of fold-change can be a difficult concept to grasp but hopefully those calculations make it a little clearer. The last concept to think about is the pseudo-R^2 for our negative binomial model, which is:

$$PseudoR^2 = 1 - (158.23/334.13) = 0.53$$

Finally, we can plot the model!

Plotting the Negative Binomial Model

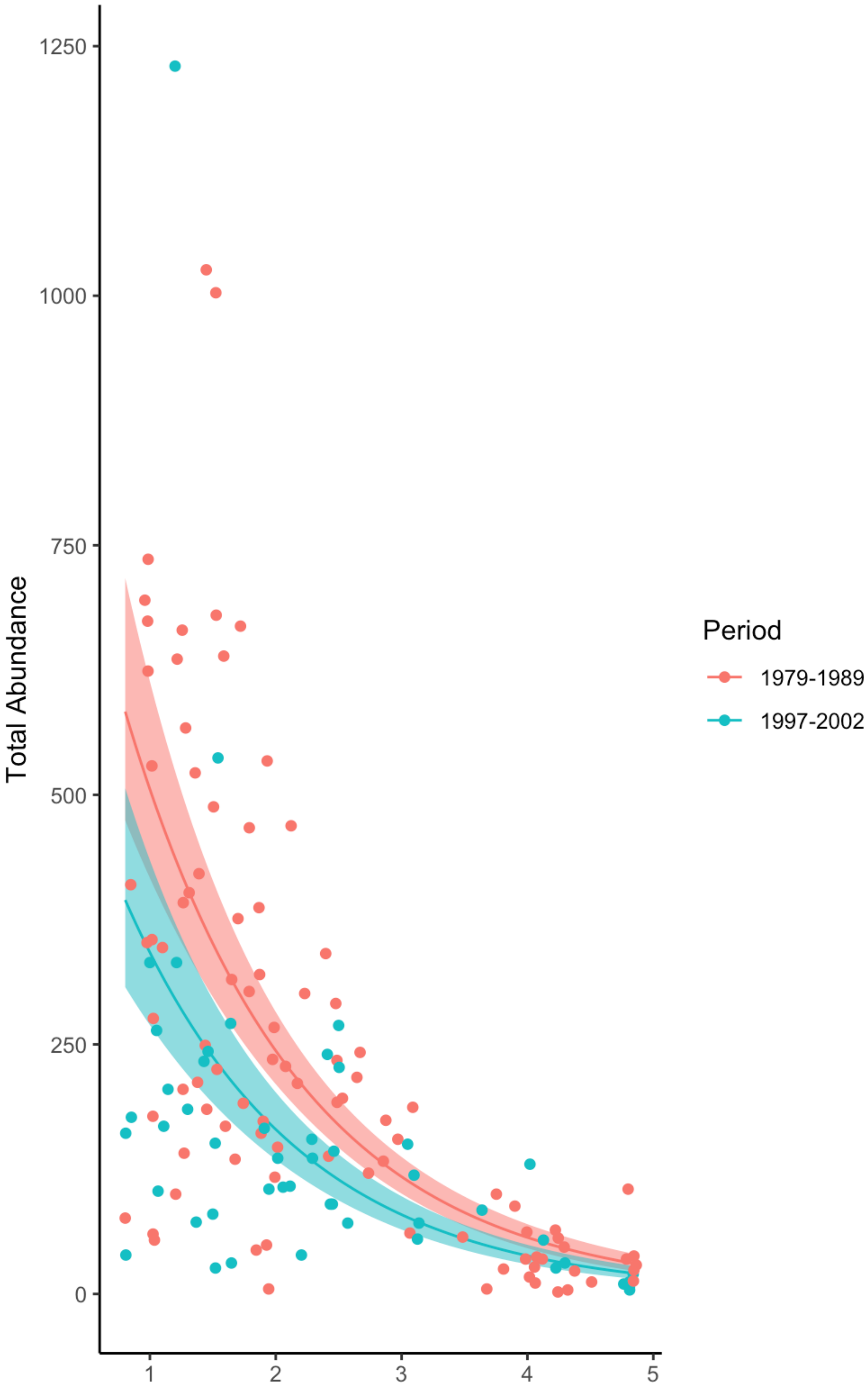
Plotting generalised linear models are not as easy as linear models. Instead, we need to use our model, make predictions and plot from these.

```
range(fish$MeanDepth) # Finding the range of MeanDepth

## [1] 0.804 4.865
```

```
period1 <- data.frame(MeanDepth=seq(from=0.804, to=4.865, length=100),
Period="1")
period2 <- data.frame(MeanDepth=seq(from=0.804, to=4.865, length=100),
Period="2")
period1_predictions<- predict(M4, newdata = period1, type = "link",
se.fit = TRUE) # the type="link" here predicted the fit and se on the
log-linear scale.
period2_predictions<- predict(M4, newdata = period2, type = "link",
se.fit = TRUE)
period1$pred<- period1_predictions$fit
period1$se<- period1_predictions$se.fit
period1$upperCI<- period1$pred+(period1$se*1.96)
period1$lowerCI<- period1$pred-(period1$se*1.96)
period2$pred<- period2_predictions$fit
period2$se<- period2_predictions$se.fit
period2$upperCI<- period2$pred+(period2$se*1.96)
period2$lowerCI<- period2$pred-(period2$se*1.96)
complete<- rbind(period1, period2)

# Making the Plot
ggplot(complete, aes(x=MeanDepth, y=exp(pred)))+
  geom_line(aes(color=factor(Period)))+
  geom_ribbon(aes(ymin=exp(lowerCI), ymax=exp(upperCI),
fill=factor(Period), alpha=0.3), show.legend = FALSE)+
  geom_point(fish, mapping = aes(x=MeanDepth, y=TotAbund,
color=factor(Period)))+
  labs(y="Total Abundance", x="Mean Depth (km)")+
  theme_classic()+
  scale_color_discrete(name="Period", labels=c("1979-1989", "1997-
2002"))
```


Total Abundance
1250
1000
750
500
250
0
Mean Depth (km)
1
2
3
4
5
Period
1979-1989
1997-2002

Bee Mites

The “bee_mites.csv” are the experimental results of the *in vitro* effect of four commercially available acaricides on mites (*Voroa* sp.). Each mite group was exposed to the tested pesticide and the number of dead mites counted for 24-hours. The underlying question was regardless of acaricide how does increasing the concentration impact the number of dead mites.

Fitting the Poisson Model

```
mites<- read.csv("bee_mites.csv")
mites_m1<- glm(Dead_mites~Concentration, data = mites, family =
"poisson")
summary(mites_m1)
```

```
##
## Call:
## glm(formula = Dead_mites ~ Concentration, family = "poisson",
##     data = mites)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84167  -0.61720  -0.00242   0.47048   1.82502
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.52820    0.09316   5.670 1.43e-08 ***
## Concentration  0.57181    0.08132   7.032 2.04e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 154.79  on 114  degrees of freedom
## Residual deviance: 109.25  on 113  degrees of freedom
## AIC: 398.71
##
## Number of Fisher Scoring iterations: 5
```

```
anova(mites_m1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Dead_mites
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            114     154.79
## Concentration  1   45.535       113     109.25 1.499e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Interpretation

Ok so let's interpret this model. Here are the two line equations:

$$\ln(NumberofDeadMites) = 0.53 + 0.57 * Concentration$$

$$NumberofDeadMites = e^{0.53+0.57*Concentration} = e^{0.53} * e^{0.57*Concentration}$$

We can interpret the increase in concentration as "for every gram per litre increase of acaricide concentration the number of dead mites increased by a factor of e^(0.57) or 1.77-fold".

$$PseudoR^2 = 1 - (109.25/154.79) = 0.29$$

Finally, we have a pseudo-R^2 of 0.29 meaning our model can explain 29% of variation in the number of dead mites.

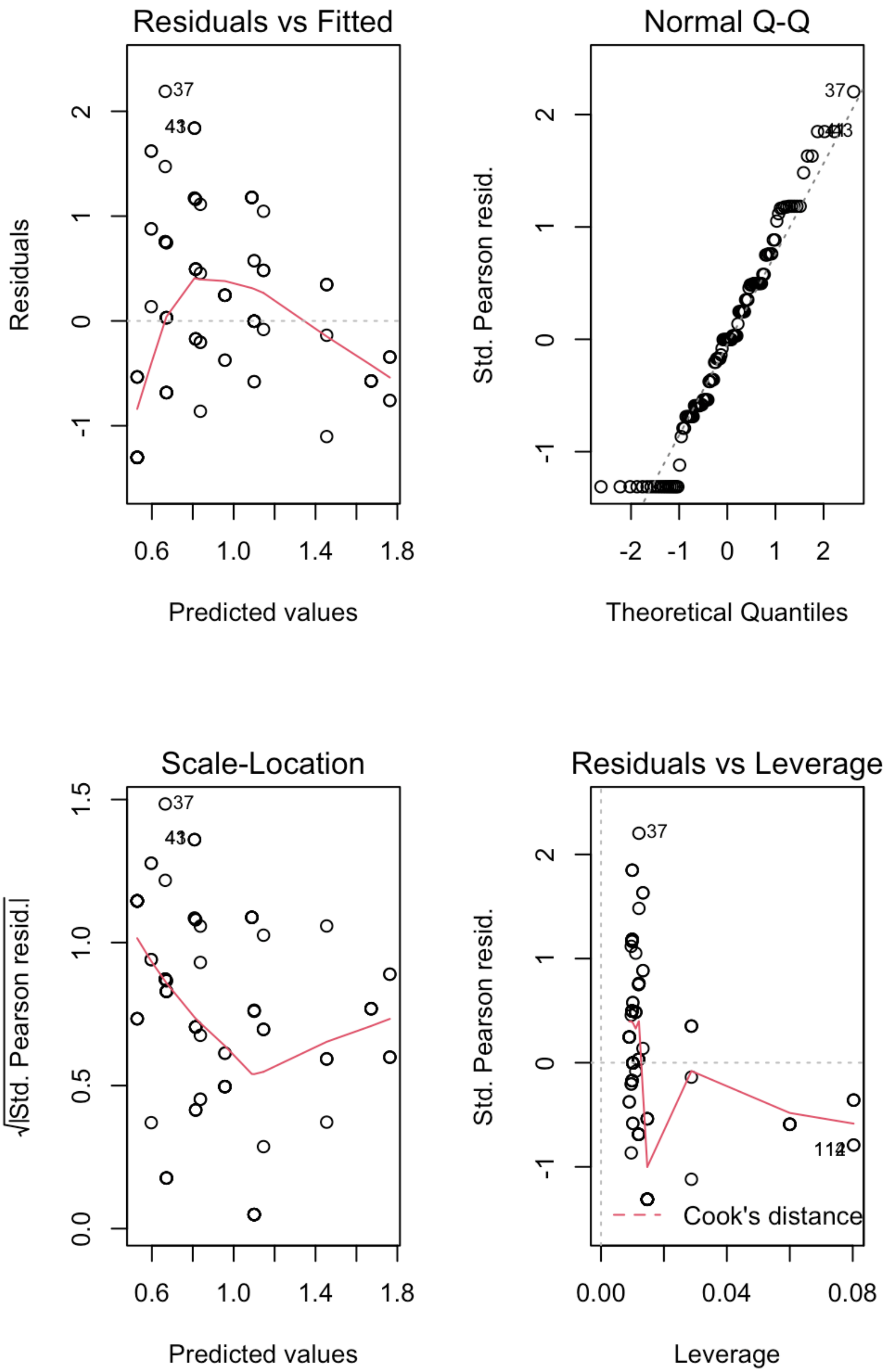
Model Validation

The final step (and hopefully last) is the model validation. Let's first look at the dispersion parameter:

$$DispersionParameter = 109.25/113 = 0.97$$

The dispersion parameter is very close of 1 and therefore we can conclude that dispersion is not an issue for our model! YAY! But let's check the model diagnostics.

```
par(mfrow=c(2,2)) #partitioning the plot window into a 2X2
plot(mites_m1)
```


Residuals vs Fitted
Residuals
Predicted values
2
1
0
-1
0.6
1.0
1.4
1.8
37
43
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
-2
-1
0
1
2
37
41 43
Scale-Location
√|Std. Pearson resid.|
Predicted values
1.5
1.0
0.5
0.0
0.6
1.0
1.4
1.8
37
43
Residuals vs Leverage
Std. Pearson resid.
Leverage
0.00
0.04
0.08
37
114
Cook's distance

So what is wrong with these diagnostics? What might be causing the issues?

You should have spotted that the “Residuals vs Fitted” and “Scale-Location” plots indicate that there are issues regarding the homogeneity of variances across the predicted values. This can be be seen as there is a distinct patterning in these plots and this can be seen by the curvature in the red spline. One reason for this might be because we have fitted an inappropriate model family and this is an important thing to consider. Here, we have fitted a Poisson family on the basis that we have the number of dead mites, however, if you examine the data frame closer you’ll notice that the total number of mites varied across the trials and therefore we strictly do not have Poisson data as it is constrained by the total number of dead mites. Tomorrow, we will look at reanalysing this data using a binomial family, but this is an important lesson that we can analyse the data using a Poisson family, interpret the model, validate the model with the dispersion parameter **BUT** we could have chosen the wrong model family in the first instance. This is an important lesson because R won’t tell you it’s wrong, we have to figure this out ourselves through our knowledge of the experimental design, the data collection and the statistical distributions and data types.

Regardless of this, let’s plot the Poisson model anyways and compare this with the results we get tomorrow.

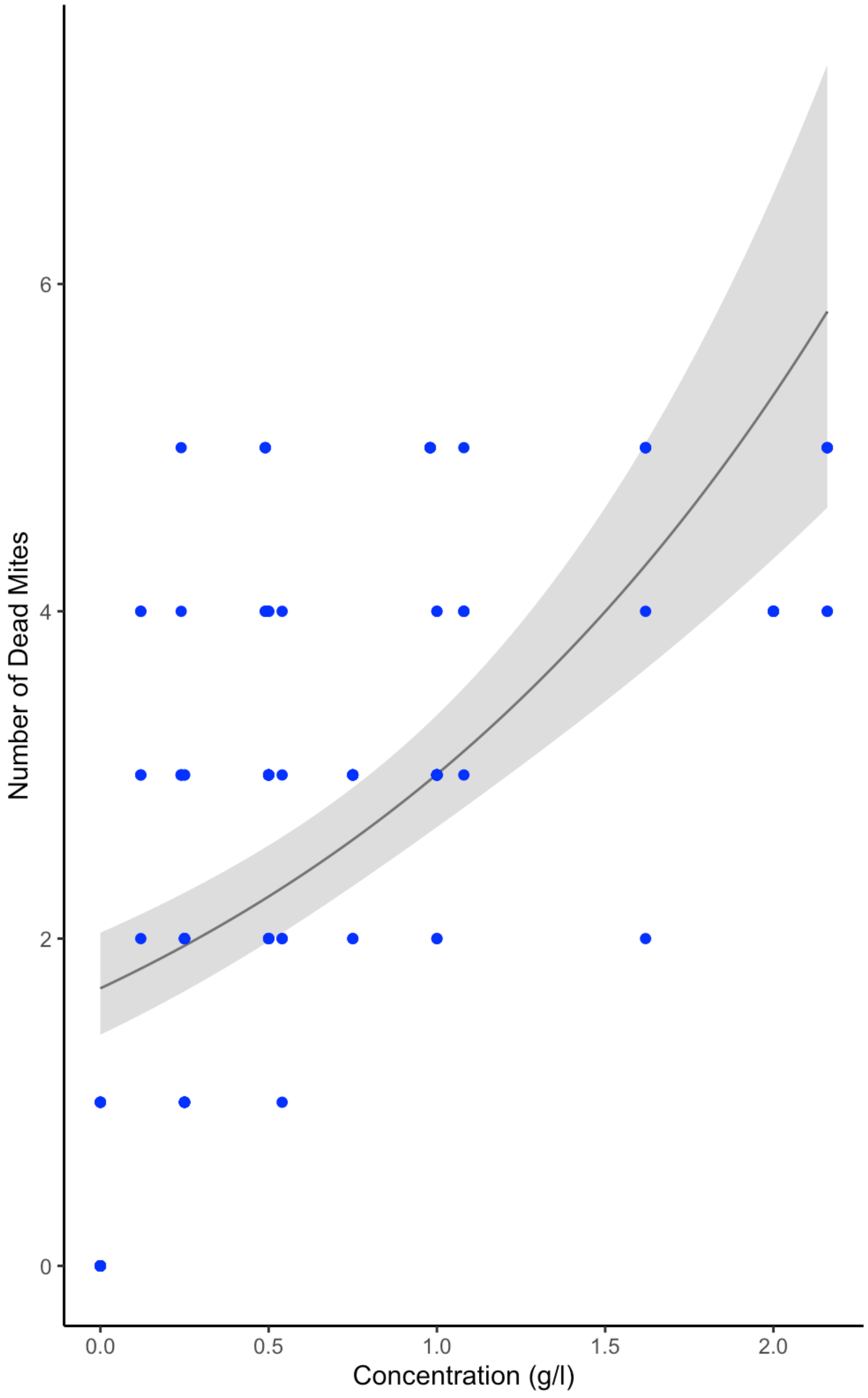
Plotting the Model

```
range(mites$Concentration) # Finding the range of concentration

## [1] 0.00 2.16

new_data <- data.frame(Concentration=seq(from=0, to=2.16, length=100))
predictions<- predict(mites_m1, newdata = new_data, type = "link",
se.fit = TRUE) # the type="link" here predicted the fit and se on the
log-linear scale.
new_data$pred<- predictions$fit
new_data$se<- predictions$se.fit
new_data$upperCI<- new_data$pred+(new_data$se*1.96)
new_data$lowerCI<- new_data$pred-(new_data$se*1.96)

# Making the Plot
ggplot(new_data, aes(x=Concentration, y=exp(pred)))+
  geom_line(col="black")+
  geom_ribbon(aes(ymin=exp(lowerCI), ymax=exp(upperCI), alpha=0.1),
show.legend = FALSE, fill="grey")+
  geom_point(mites, mapping = aes(x=Concentration, y=Dead_mites),
col="blue")+
  labs(y="Number of Dead Mites", x="Concentration (g/l)")+
  theme_classic()
```


Number of Dead Mites
0
2
4
6
0.0
0.5
1.0
1.5
2.0
Concentration (g/l)

Extra Tasks

I know this handout has been particularly long and thorough, but here are some data sets and research questions for you to practise with.

1. Species richness on the Galapagos islands (“gala.txt”):

- How does area of the island affect the number of plant species?

- The data set includes the “Species” (the number of species), “Endemics” (the number of endemic species), “Area” (area of the island in km^2), “Elevation” (highest elevation of the island metres), “Nearest” (distance from nearest island in km), “Scruz” (distance from Santa Cruz in km) and “Adjacent” (area of the adjacent island in square kilometres).

- HINT: you will need to log transform the variable “Area” as there is a lot of bunching - plot the relationship between Species~Area and Species~log(Area) to see what I mean.

2. Amphibian roadkills in Portugal (“RoadKills.txt”):

- How does the distance to the nearby park affect the number of road kills?

- The data set includes a whole lot of variable but I want you to focus on. “TOT.N” (the total number of roadkills) and “D.PARK” (the distance to nearest park in metres).