Data Analaysis and Modelling: Multiple Regression glm example

NES8010 Quantitative Research Skills

February 2021

# Introduction

This is an example of multiple linear regression, based on the basic straight line regression :

where is your response variable, your explanatory variable, is the intercept of your fitted regression line, and the gradient of your line. Note that sometimes your regression line goes straight through the origin of the graph and the intercept is not needed.

This also forms the basis for more complex regressions where curves can be fitted with different transformations of your predictor variables.

## Parental care in North American Passerines

A dataset compiled of 84 North American passerines to explore the relationship of parental care and nesting factors. Data are derived from The Birders Handbook, by Ehrlich, et al. (1988) compiled by Amy R. Moore, as a student at Grinnell College in 1999.

We are interested in the duration of parental care and its association with the nest type variables. Explore the dataset and determine feasible hypotheses to test and develop candidate models to test them.

### Getting started

First, read in the data. Ensure you are either using RProj or have the correct filepath to where your data are stored using the Session> set working >directory> choose working directory from the Rstudio menus.

The code in the notes uses the here package and the data are in the Data directory and checks that the packages we need for the exercise are loaded to our workspace.

wants <- c("ggfortify", "here")  
has <- wants %in% rownames(installed.packages())  
if(any(!has)) install.packages(wants[!has])  
  
  
library(here)

## here() starts at C:/Users/nacm4/OneDrive - Newcastle University/Documents/MSc/NES8010 Quantitative Skills/NES8010 Quantitative Skills

nest<-read.csv(here("Data", "BirdNest.csv"))  
head(nest)

## Species Common Page Length Nesttype  
## 1 Tyrannus tyrannus Eastern Kingbird 360 20.0 cup  
## 2 Myiodynastes luteiventris Sulphur-bellied Flycatcher 368 20.0 cavity  
## 3 Myiarchus cinerascens Ash-thoated Flycatcher 372 20.0 cavity  
## 4 Myiarchus tyrannulus Brown-crested Flycatcher 372 22.5 cavity  
## 5 Myarchus tuberculifer Dusky-capped Flycatcher 374 17.0 cavity  
## 6 Sayornis phoebe Eastern Phoebe 378 17.0 cup  
## Location No.eggs Color Incubate Nestling Totcare Closed.  
## 1 decid 3.5 1 17.0 17.0 34.0 0  
## 2 decid 3.5 1 15.5 17.0 32.5 1  
## 3 decid 4.5 1 15.0 15.0 30.0 1  
## 4 decid 4.5 1 14.0 16.5 30.5 1  
## 5 decid 4.5 1 14.0 14.0 28.0 1  
## 6 bridge 4.5 0 16.0 15.5 31.5 0

## Data Exploration

**Step 1:** Determine a biologically sensible hypothesis for the inclusion of each variable.

*e.g. in this example is there any reason to include the variable 'page'?*

**Step 2:** Data exploration. Plot the relationships between covariates. Are they measuring the same thing (we call this co-linearity, which can be observed in a scatterplot, and determined statistically with correlation)

summary(nest)

## Species Common Page Length   
## Length:84 Length:84 Min. :360.0 Min. : 9.00   
## Class :character Class :character 1st Qu.:430.0 1st Qu.:14.00   
## Mode :character Mode :character Median :469.0 Median :17.00   
## Mean :488.1 Mean :17.64   
## 3rd Qu.:552.5 3rd Qu.:20.00   
## Max. :646.0 Max. :31.50   
##   
## Nesttype Location No.eggs Color   
## Length:84 Length:84 Min. : 1.000 Min. :0.0000   
## Class :character Class :character 1st Qu.: 3.500 1st Qu.:1.0000   
## Mode :character Mode :character Median : 4.500 Median :1.0000   
## Mean : 4.583 Mean :0.8333   
## 3rd Qu.: 5.000 3rd Qu.:1.0000   
## Max. :12.500 Max. :1.0000   
##   
## Incubate Nestling Totcare Closed.   
## Min. :10.00 Min. : 8.00 Min. :19.00 Min. :0.0000   
## 1st Qu.:12.00 1st Qu.:11.38 1st Qu.:23.50 1st Qu.:0.0000   
## Median :13.00 Median :14.00 Median :27.50 Median :0.0000   
## Mean :13.35 Mean :14.44 Mean :27.73 Mean :0.3214   
## 3rd Qu.:14.00 3rd Qu.:17.00 3rd Qu.:31.00 3rd Qu.:1.0000   
## Max. :17.00 Max. :22.50 Max. :37.50 Max. :1.0000   
## NA's :1 NA's :1

There are several continuous variables and a few factors (*Nesttype*, *Location*).

Explore the data further using the table() and hist() commands, some suggestions are given below but this is not exhaustive, try some others of your own.

table (nest$Nesttype)

##   
## burrow cavity crevice cup pendant saucer spherical   
## 2 17 3 53 2 4 3

table (nest$Location)

##   
## bank bridge building cliff conif decid ground shrub   
## 3 1 2 1 14 24 19 17   
## snag   
## 3

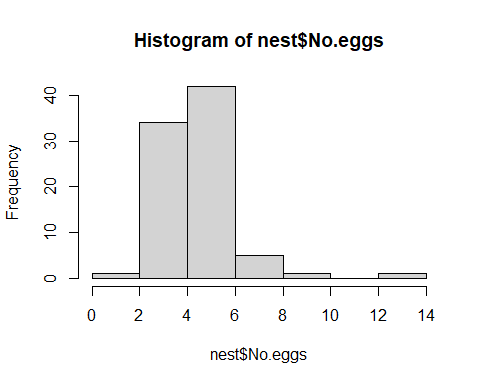
table (nest$Location, nest$Nesttype)

##   
## burrow cavity crevice cup pendant saucer spherical  
## bank 2 0 1 0 0 0 0  
## bridge 0 0 0 1 0 0 0  
## building 0 1 0 1 0 0 0  
## cliff 0 0 1 0 0 0 0  
## conif 0 2 0 9 2 1 0  
## decid 0 11 0 11 0 1 1  
## ground 0 0 1 16 0 2 0  
## shrub 0 0 0 15 0 0 2  
## snag 0 3 0 0 0 0 0

table (nest$Nesttype, nest$Closed.)

##   
## 0 1  
## burrow 0 2  
## cavity 0 17  
## crevice 0 3  
## cup 53 0  
## pendant 0 2  
## saucer 4 0  
## spherical 0 3

hist(nest$No.eggs)



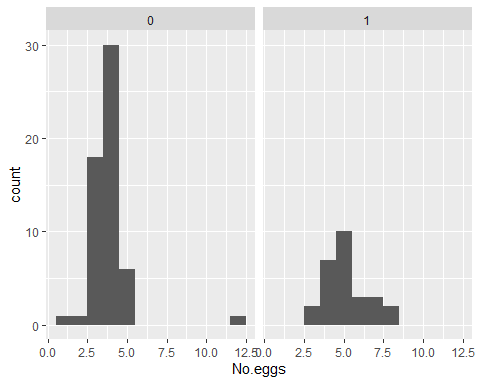
From these tables you begin to get an idea of the type of data we have. By tabulating two columns against each other (here *Nesttype* and *Closed* (which is a binary 0/1 variable to describe of the nest is closed or not) we clarify that only two nest types are open nests).

Try to visualise other combinations of the variables.

For example we could look to see if there is a difference in the number of eggs laid between open and closed nest types.

We can't do this easily with the hist() command so we will use the graphics package **ggplot2** to make these histograms quickly and side by side.

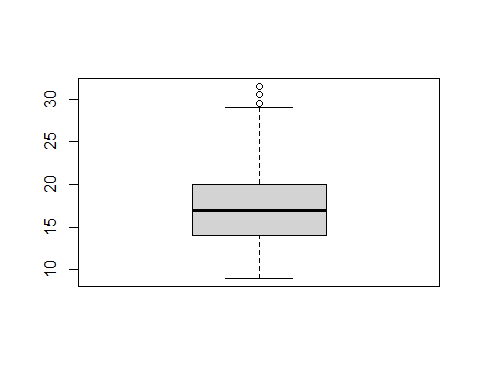
library(ggplot2)  
m<- ggplot(nest, aes(x=No.eggs))  
m <- m + geom\_histogram(binwidth = 1)  
m + facet\_grid( ~ Closed.)



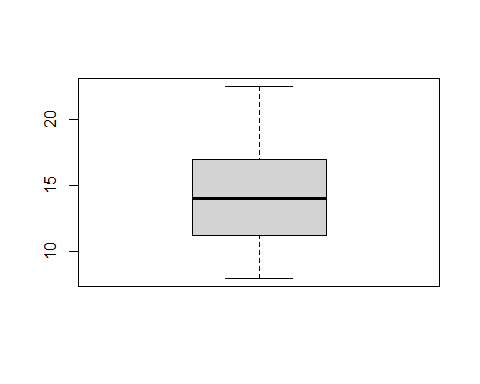
Remember there are lots of online resources and really useful help files <http://docs.ggplot2.org/current/> for every graph imaginable.

For the continuous variables you can plot boxplots to assess their distribution

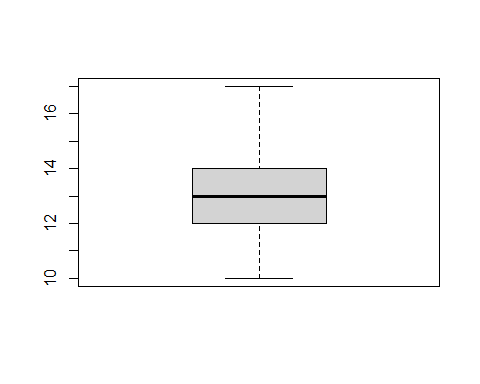
boxplot(nest$Length)



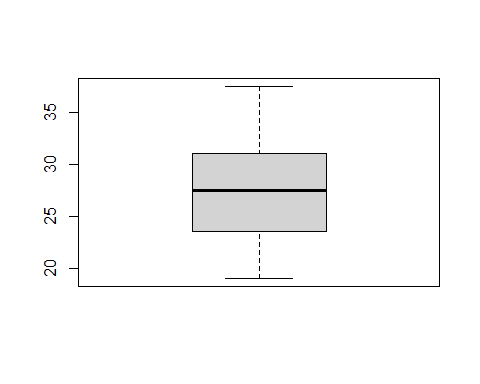
boxplot(nest$Nestling)



boxplot(nest$Incubate)

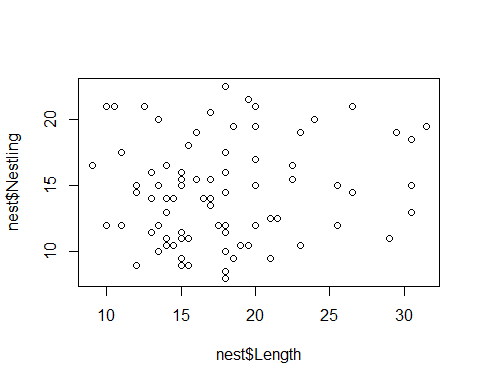


boxplot(nest$Totcare)

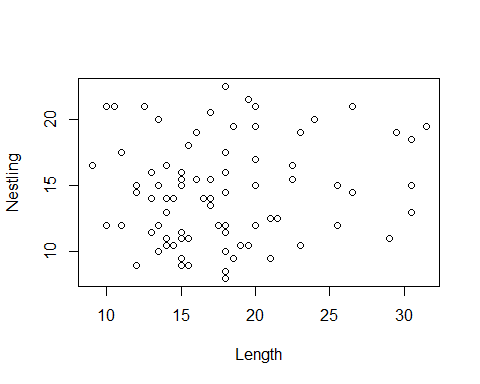


We can also plot some of these variables against each other in a scatterplot using base r commands

plot(nest$Length, nest$Nestling)



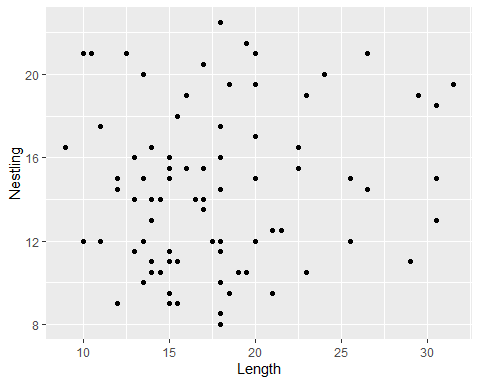
plot(Nestling~Length, nest)



Note the two different notations produce the same plot. If you are doing a lot of linear modelling you may find it easier to remember the formula method. It also mean you don't have to keep referring the dataframe.

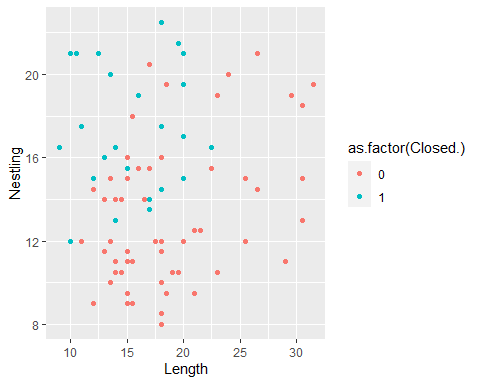
You can also produce the same plot in ggplot

ggplot(nest, aes(Length,Nestling))+geom\_point()



There doesn't look like there is a relationship here (no specific trend) but we could consider some of the other variables we have - let's add in number of eggs by using colour. [NB: we need to tell R that *Closed* is a factor not a continuous variable]

ggplot(nest, aes(Length,Nestling, colour=as.factor(Closed.)))+geom\_point()



# Model building

We have now had a look at the different variables in this data set. How might we come up with model?

**Step 3:** Create candidate models to compare:

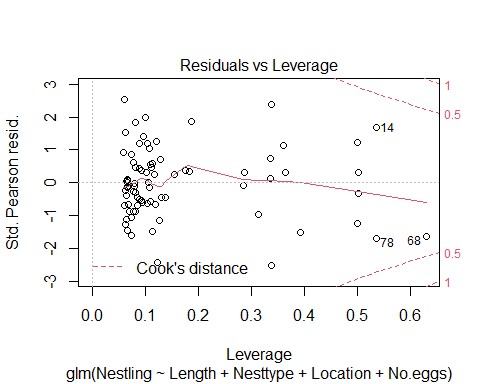
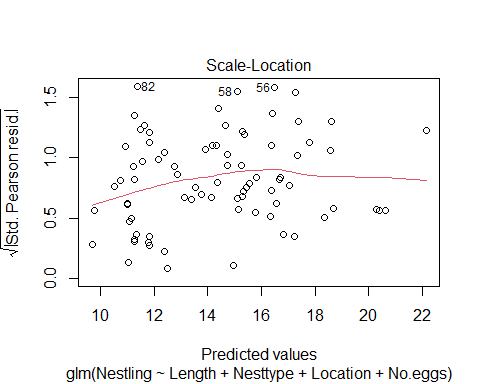
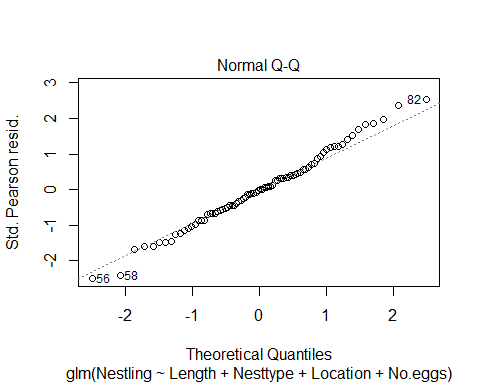
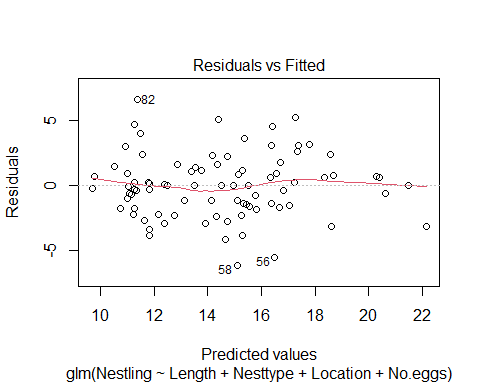
M1.nest<- glm(Nestling~Length+Nesttype + Location +No.eggs, data=nest)  
summary(M1.nest)

##   
## Call:  
## glm(formula = Nestling ~ Length + Nesttype + Location + No.eggs,   
## data = nest)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.101 -1.597 0.000 1.277 6.638   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.89176 2.58250 6.928 2.04e-09 \*\*\*  
## Length 0.18646 0.06793 2.745 0.00777 \*\*   
## Nesttypecavity -7.16826 4.44994 -1.611 0.11191   
## Nesttypecrevice -6.95071 3.34157 -2.080 0.04135 \*   
## Nesttypecup -8.77268 4.34282 -2.020 0.04738 \*   
## Nesttypependant -9.29550 4.90664 -1.894 0.06248 .   
## Nesttypesaucer -10.97270 4.52440 -2.425 0.01800 \*   
## Nesttypespherical -0.66340 4.63891 -0.143 0.88671   
## Locationbridge 3.04528 4.74053 0.642 0.52281   
## Locationbuilding 4.92032 4.36582 1.127 0.26376   
## Locationcliff 6.75718 3.84188 1.759 0.08317 .   
## Locationconif 3.57830 3.97127 0.901 0.37079   
## Locationdecid 1.75222 3.96624 0.442 0.66007   
## Locationground -0.81354 3.83313 -0.212 0.83256   
## Locationshrub -0.79336 3.96334 -0.200 0.84195   
## Locationsnag 2.98459 4.32052 0.691 0.49208   
## No.eggs 0.03686 0.26711 0.138 0.89066   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 7.344144)  
##   
## Null deviance: 1181.20 on 83 degrees of freedom  
## Residual deviance: 492.06 on 67 degrees of freedom  
## AIC: 422.88  
##   
## Number of Fisher Scoring iterations: 2

This model finds Length to be a significant predictor of Nestling duration, Nest type is important (duration varies and is longer in burrows that other nest types [BUT look at the number of burrowers]). Location is not significant but there may be something interesting with cliffs (low p value [- again check sample size]).

plot(M1.nest)

## Warning: not plotting observations with leverage one:  
## 6, 48, 49, 73



## Residual plots

When you plot the output object from a linear model you will obtain 4 plots:

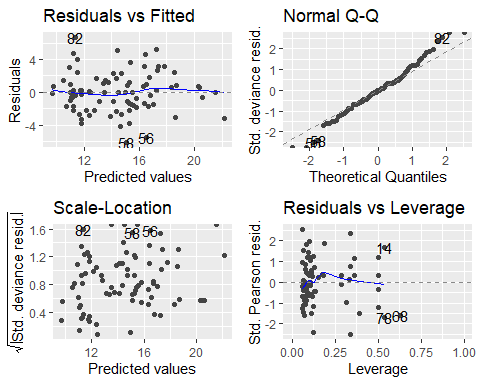
* Residuals vs Fitted
* Normal QQ plot
* Scale-Location
* Residuals vs Leverage

The first two plots are the most informative. You should not be able to see any particularly pattern in the *Residuals vs Fitted* plot, e.g. the residuals should not be negative at one set of fitted values, and positive at the other. It should look like a random scatter of points. The *QQ (quantile-quantile)* plot should ideally show you points on an approximate straight line, and gives an indication of the overall robustness of the linear model.

Any numbers on these residual plots label points that may be considered as outliers or influential points that may warrant further investigation. It is worth rechecking the raw data if any points stand out. If you think any data point may be unreliable you may consider removing it an re-running the model without this point to determine the innfluece it has on the model. Do not remove points without good reason.

The plot(model\_name) command allows us to check the assumptions of the model by assessing diagnostic plots. This base command is OK but we can also install the ggfortify package to produce ggplot like graphs.

library(ggfortify)  
autoplot(M1.nest)



If you only want to show one of the plots, e.g. the second QQ plot, use square brackets: autoplot(M1.nest)[2]

Now we need to consider that certain nest types are not possible in certain locations - are these variables in fact explaining much of the same thing? look back at the table of *Nesttype* and *Location*. Why might be more useful with this dataset is to group the open and closed nests instead of using *Nesttype*?

Try another candidate model:

M2.nest<- glm(Nestling~Length+Closed.+ Location +No.eggs, data=nest)  
summary(M2.nest)

##   
## Call:  
## glm(formula = Nestling ~ Length + Closed. + Location + No.eggs,   
## data = nest)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.5121 -1.8097 -0.3369 1.8036 7.0385   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.52928 2.73171 4.953 4.67e-06 \*\*\*  
## Length 0.18295 0.07324 2.498 0.01477 \*   
## Closed. 3.01780 1.01031 2.987 0.00385 \*\*   
## Locationbridge -0.50333 3.57565 -0.141 0.88845   
## Locationbuilding 0.70606 2.76974 0.255 0.79951   
## Locationcliff 2.02149 3.47606 0.582 0.56269   
## Locationconif -0.73296 2.02195 -0.362 0.71804   
## Locationdecid -2.29567 1.88768 -1.216 0.22791   
## Locationground -4.61402 2.06583 -2.233 0.02862 \*   
## Locationshrub -3.87041 2.02315 -1.913 0.05972 .   
## Locationsnag -1.88946 2.44869 -0.772 0.44286   
## No.eggs -0.14136 0.27650 -0.511 0.61073   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 8.96296)  
##   
## Null deviance: 1181.20 on 83 degrees of freedom  
## Residual deviance: 645.33 on 72 degrees of freedom  
## AIC: 435.65  
##   
## Number of Fisher Scoring iterations: 2

Now we can see that open/closed is significant and Location is now significant. *No.eggs* is still not important.

Let's try another few models to compare.

M3.nest<- glm(Nestling~Length+Closed.+ Location, data=nest)  
summary(M3.nest)

##   
## Call:  
## glm(formula = Nestling ~ Length + Closed. + Location, data = nest)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.4915 -1.8953 -0.2372 1.7664 7.1621   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.8693 2.3952 5.373 8.85e-07 \*\*\*  
## Length 0.1907 0.0713 2.674 0.00924 \*\*   
## Closed. 2.8342 0.9395 3.017 0.00352 \*\*   
## Locationbridge -0.6108 3.5514 -0.172 0.86393   
## Locationbuilding 0.6629 2.7544 0.241 0.81050   
## Locationcliff 2.0784 3.4566 0.601 0.54952   
## Locationconif -0.8695 1.9941 -0.436 0.66409   
## Locationdecid -2.3416 1.8760 -1.248 0.21594   
## Locationground -4.7396 2.0408 -2.322 0.02300 \*   
## Locationshrub -3.8676 2.0129 -1.921 0.05858 .   
## Locationsnag -1.8814 2.4362 -0.772 0.44247   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 8.872274)  
##   
## Null deviance: 1181.20 on 83 degrees of freedom  
## Residual deviance: 647.68 on 73 degrees of freedom  
## AIC: 433.96  
##   
## Number of Fisher Scoring iterations: 2

M4.nest<- glm(Nestling~Length+Closed.+Location +Color, data=nest)  
summary(M4.nest)

##   
## Call:  
## glm(formula = Nestling ~ Length + Closed. + Location + Color,   
## data = nest)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.1885 -1.7692 -0.1788 1.6101 7.3095   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.18282 2.36125 5.583 3.94e-07 \*\*\*  
## Length 0.19518 0.07015 2.782 0.00689 \*\*   
## Closed. 3.10140 0.93484 3.318 0.00143 \*\*   
## Locationbridge -1.00094 3.49838 -0.286 0.77561   
## Locationbuilding 2.08429 2.81311 0.741 0.46115   
## Locationcliff 1.40969 3.41777 0.412 0.68123   
## Locationconif 0.60421 2.11325 0.286 0.77576   
## Locationdecid -1.09386 1.96166 -0.558 0.57883   
## Locationground -3.50411 2.11270 -1.659 0.10155   
## Locationshrub -2.70242 2.07508 -1.302 0.19696   
## Locationsnag -1.89037 2.39560 -0.789 0.43264   
## Color -1.94076 1.03782 -1.870 0.06554 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 8.578826)  
##   
## Null deviance: 1181.20 on 83 degrees of freedom  
## Residual deviance: 617.68 on 72 degrees of freedom  
## AIC: 431.97  
##   
## Number of Fisher Scoring iterations: 2

M5.nest<- glm(Nestling~Length+Closed.+Color, data=nest)  
summary(M5.nest)

##   
## Call:  
## glm(formula = Nestling ~ Length + Closed. + Color, data = nest)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.0346 -2.5804 -0.7445 2.2572 7.8921   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.32612 1.58439 6.517 5.90e-09 \*\*\*  
## Length 0.23380 0.06965 3.357 0.00121 \*\*   
## Closed. 4.36030 0.77878 5.599 2.94e-07 \*\*\*  
## Color -1.69291 0.92594 -1.828 0.07123 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 9.990144)  
##   
## Null deviance: 1181.20 on 83 degrees of freedom  
## Residual deviance: 799.21 on 80 degrees of freedom  
## AIC: 437.62  
##   
## Number of Fisher Scoring iterations: 2

Now models M4 and M5 both look plausible - how do we choose between them?

We can compare the **AIC values** between the models. But remember we need to be comparing nested models - i.e. those that differ by only **ONE** variable.

AIC (M3.nest, M4.nest, M5.nest)

## df AIC  
## M3.nest 12 433.9579  
## M4.nest 13 431.9740  
## M5.nest 5 437.6176

So *Location* is important but what about different nest types in different locations - we can add in an interaction term to account for this. The \* denotes the interaction in the model formulae. The model will aslo include *Location* as a variable of its own (and associated degrees of freedom)

M6.nest<- glm(Nestling~Length+Closed.\*Location +Color, data=nest)  
summary(M6.nest)

##   
## Call:  
## glm(formula = Nestling ~ Length + Closed. \* Location + Color,   
## data = nest)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.183 -1.416 0.000 1.424 5.149   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.80979 2.84453 2.042 0.044988 \*   
## Length 0.20336 0.06607 3.078 0.003002 \*\*   
## Closed. 10.40447 2.04515 5.087 3.06e-06 \*\*\*  
## Locationbridge 6.23303 3.68700 1.691 0.095504 .   
## Locationbuilding 13.32381 3.77653 3.528 0.000755 \*\*\*  
## Locationcliff 1.32015 3.11699 0.424 0.673243   
## Locationconif 8.02389 2.79471 2.871 0.005449 \*\*   
## Locationdecid 6.87022 2.73184 2.515 0.014277 \*   
## Locationground 3.90187 2.70681 1.442 0.154032   
## Locationshrub 3.77523 2.51901 1.499 0.138581   
## Locationsnag -1.90673 2.18454 -0.873 0.385829   
## Color -2.09078 0.95033 -2.200 0.031204 \*   
## Closed.:Locationbridge NA NA NA NA   
## Closed.:Locationbuilding -14.99774 4.28705 -3.498 0.000830 \*\*\*  
## Closed.:Locationcliff NA NA NA NA   
## Closed.:Locationconif -7.43100 2.58555 -2.874 0.005404 \*\*   
## Closed.:Locationdecid -8.51517 2.28721 -3.723 0.000402 \*\*\*  
## Closed.:Locationground -7.98253 3.42568 -2.330 0.022771 \*   
## Closed.:Locationshrub NA NA NA NA   
## Closed.:Locationsnag NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 7.132147)  
##   
## Null deviance: 1181.20 on 83 degrees of freedom  
## Residual deviance: 484.99 on 68 degrees of freedom  
## AIC: 419.66  
##   
## Number of Fisher Scoring iterations: 2

There are NA's these occur where the model was not able to be fitted due to lack of data. Compare to the data in the table:

table(nest$Closed., nest$Location)

##   
## bank bridge building cliff conif decid ground shrub snag  
## 0 0 1 1 0 10 12 18 15 0  
## 1 3 0 1 1 4 12 1 2 3

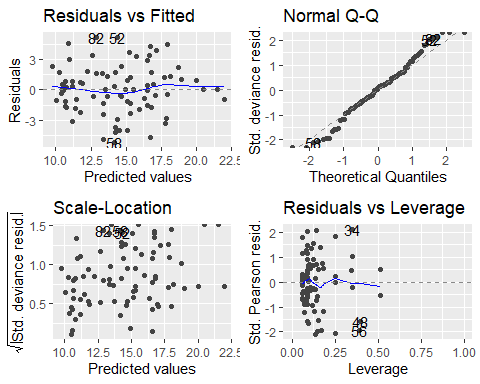
We can check the residuals of this model M6

autoplot(M6.nest)

## Warning: Removed 84 row(s) containing missing values (geom\_path).

## Warning: Removed 5 rows containing missing values (geom\_point).

## Warning: Removed 5 row(s) containing missing values (geom\_path).



There is little difference between M5 and M6. M6 is *expensive* in terms of parameters BUT is much harder to explain in terms of ecology/ biology, especially in light of the small sample sizes for some combinations of variables so proceed with caution and where ever possible opt for the simpler model.

You can repeat this analysis but explore how the **Incubation** or the **totalcare** data vary? Do you find the same variables are significant?

# Modelling protcol

1. DEFINE YOUR PERCEIVED RELATIONSHIP (A HYPOTHESIS)

* e.g. 'badger weight' is a function of 'length'

1. PLOT THE RESPONSE VARIABLE AGAINST THE PREDICTOR

* e.g. plot(badger$Length,badger$Weight)

1. WRITE THE MODEL INTO R

* test.glm<- glm(formula=response~predictor, family=errorfamily, dataframe name) e.g. test.glm <- glm(Weight~Length, gaussian, badger)

1. SUMMARY MODEL OUTPUT

e.g. summary(test.glm)

assess estimatse and coefficients (are they significant ?) assess amount of deviance (variation explained) (NULL-RESIDUAL)/NULL

1. ASSESS ERROR MODEL SUITABILITY

e.g. hist(residuals(test.glm)) ~ bell-shaped curved centred on zero formally: qqnorm(residuals(test.glm) ~ should be straight line