Binomial and Binary Models

Dr Josh Hodge

## Introduction

Remember the three steps of generalised linear models:

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

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

### Binary Models

#### *Varoa* spp in Honeycomb Cells

require(ggplot2)

## Loading required package: ggplot2

require(ggpubr)

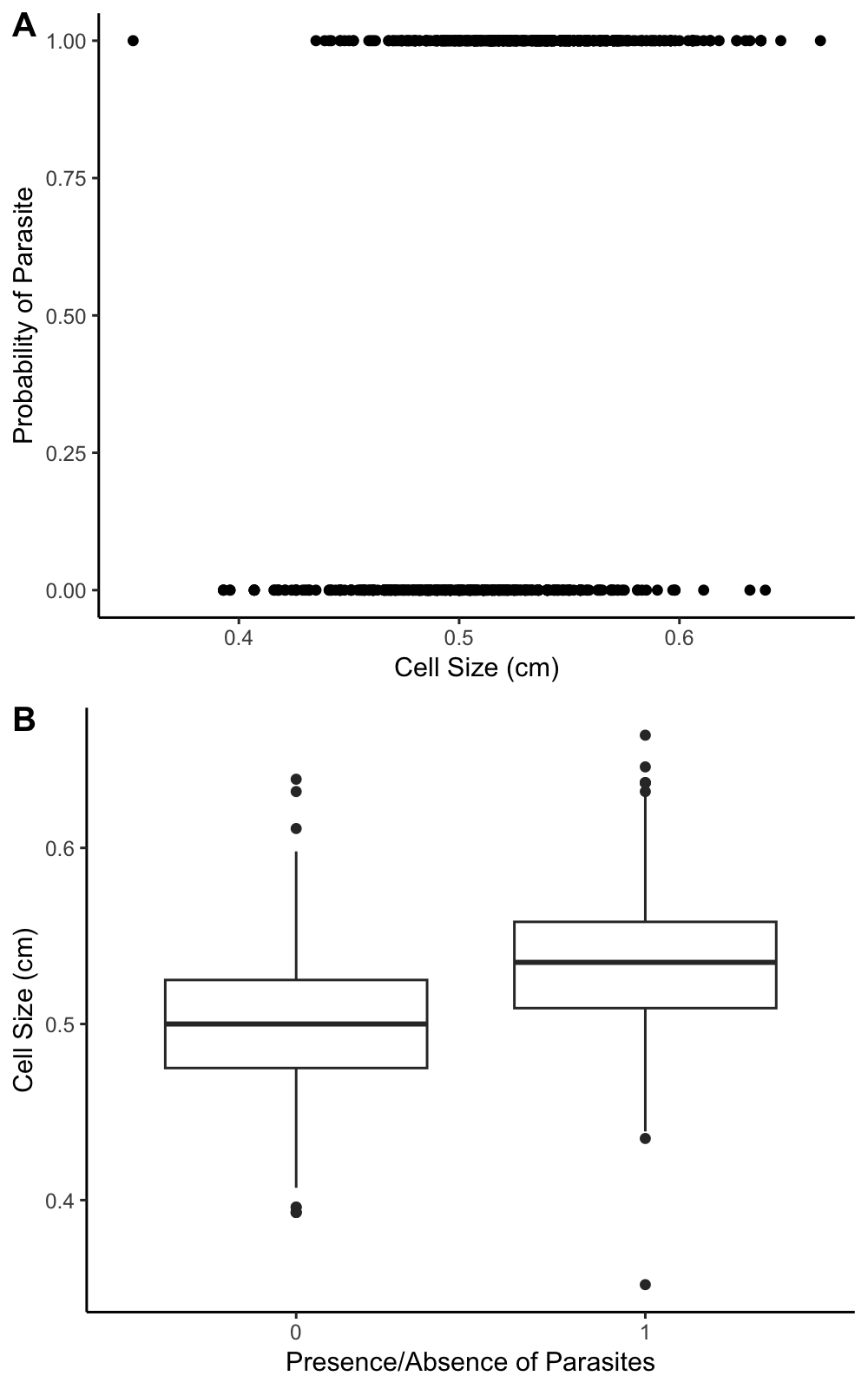
## Loading required package: ggpubr

worker<- read.csv("workerbees.csv", stringsAsFactors = T)  
str(worker)

## 'data.frame': 917 obs. of 2 variables:  
## $ Parasites: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CellSize : num 0.424 0.454 0.457 0.468 0.493 0.558 0.564 0.489 0.501 0.501 ...

We are going to analyse data collected on worker bee (*Apis mellifera*) brood honeycomb cell size and the prevalence of the parasitic mite (*Varoa destructor*). Let’s first investigate the data graphically:

scatterplot<-ggplot(worker, aes(x=CellSize, y=Parasites))+  
 geom\_point()+  
 labs(x= "Cell Size (cm)", y="Probability of Parasite")+  
 theme\_classic()  
boxplot<- ggplot(worker, aes(x=factor(Parasites), y=CellSize))+  
 geom\_boxplot()+  
 theme\_classic()+  
 labs(x="Presence/Absence of Parasites", y="Cell Size (cm)")  
ggarrange(scatterplot, boxplot, labels=c("A","B"), ncol=1, nrow=2)



#### Fitting the Model

M1<- glm(Parasites~CellSize, data = worker, family = "binomial")  
summary(M1)

##   
## Call:  
## glm(formula = Parasites ~ CellSize, family = "binomial", data = worker)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -11.245 1.052 -10.69 <2e-16 \*\*\*  
## CellSize 22.175 2.034 10.90 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1259.6 on 916 degrees of freedom  
## Residual deviance: 1104.9 on 915 degrees of freedom  
## AIC: 1108.9  
##   
## Number of Fisher Scoring iterations: 3

anova(M1, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Parasites  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 916 1259.6   
## CellSize 1 154.73 915 1104.9 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### Model 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 interpret the model and construct linear equation:

This line equation and the summary output tells us that increasing cell size of honeycomb increases the probability being infected by *Varoa destructor*. In the lecture, we examined this idea of identifying the value of *x* (here: Cell Size) where the probability flips from less likely to be infected or more likely to be infected. We can use the table presented in the lecture to do the following calculation:

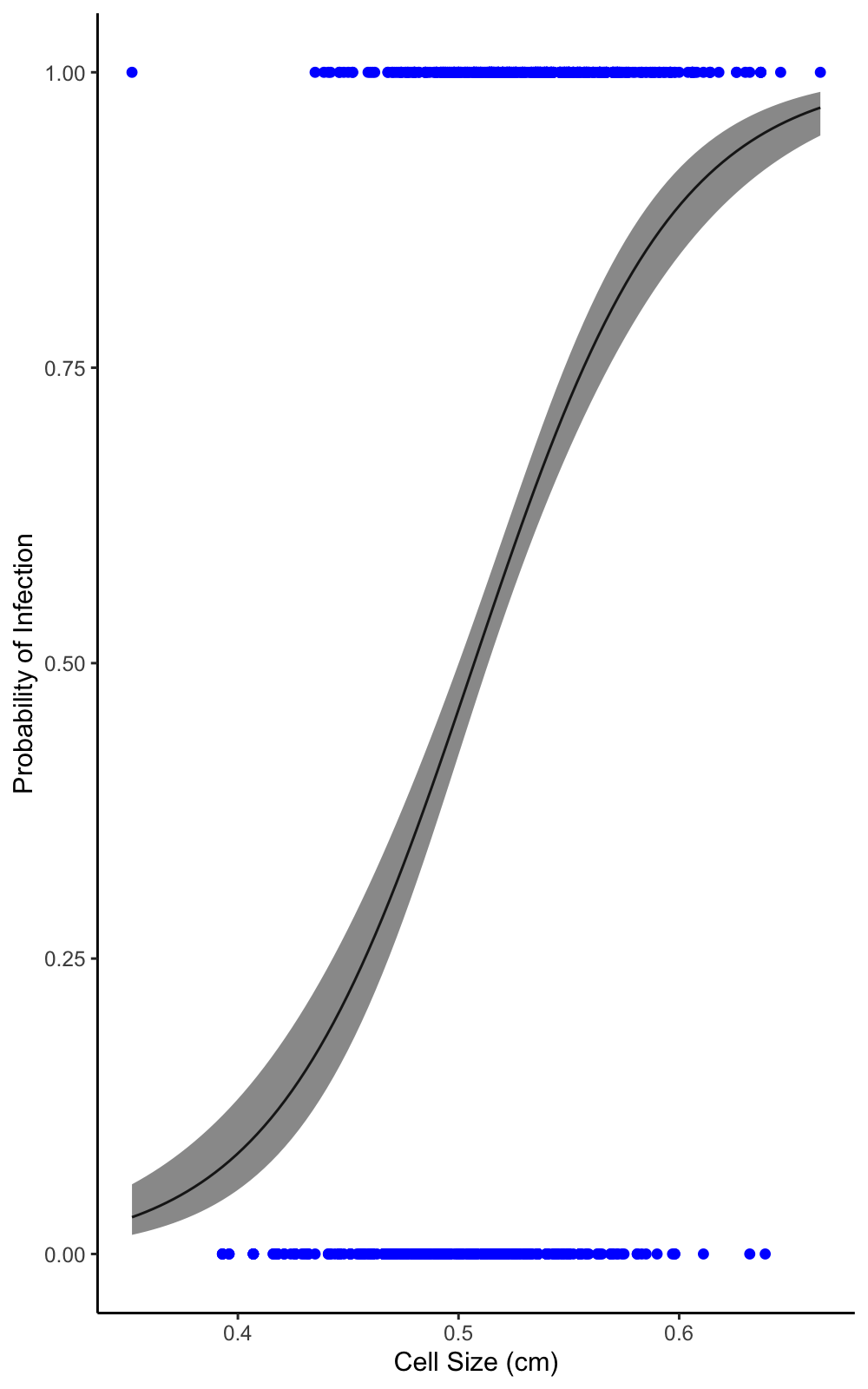
This suggests that a honeycomb cell size above 0.51cm is more likely to be infected by the *Varoa destructor* mite. Now we have this information, let’s see what this looks like graphically.

#### Plotting the Model

range(worker$CellSize) # Finding the range of Cell Size

## [1] 0.352 0.664

new\_data <- data.frame(CellSize=seq(from=0.352, to=0.664, length=100))  
predictions<- predict(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=CellSize, y=plogis(pred)))+   
 geom\_line(col="black")+  
 geom\_point(worker, mapping = aes(x=CellSize, y=Parasites), col="blue")+  
 geom\_ribbon(aes(ymin=plogis(lowerCI), ymax=plogis(upperCI), alpha=0.2), show.legend = FALSE)+   
 labs(y="Probability of Infection", x="Cell Size (cm)")+  
 theme\_classic()



Looking at the graph together with the coefficients and our flipping point, we can see that these inferences are graphically supported. The last thing aspect to examine is the pseudo-R^2, which tells us that this model was able to explain 12% of variation in the presence/absence of the *Varoa destructor* mite.

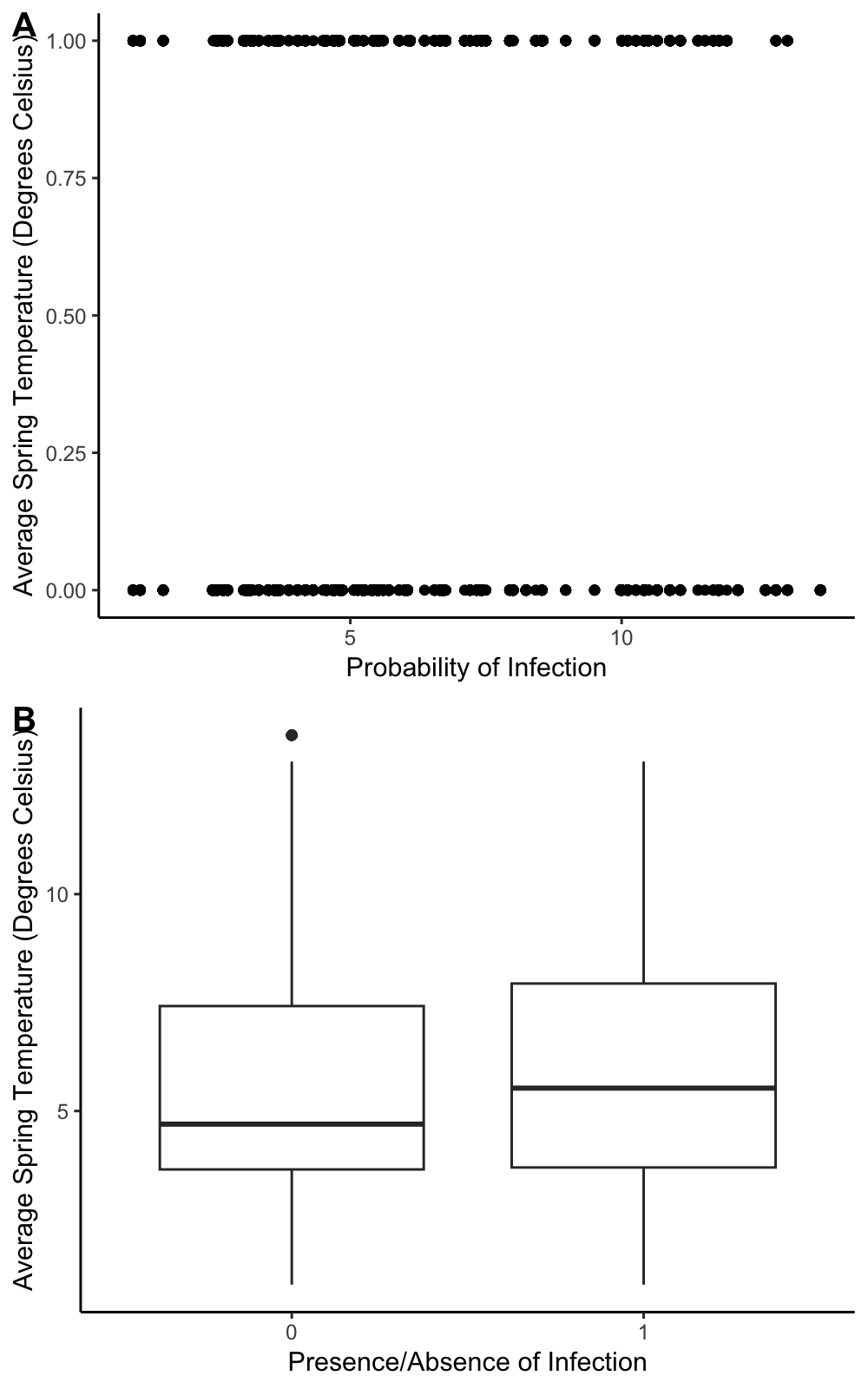
#### Chrytrid Infection Status in the Pyrenees

require(ggplot2)  
chytrid<- read.csv("chytrid.csv", stringsAsFactors = T)  
str(chytrid)

## 'data.frame': 6795 obs. of 10 variables:  
## $ Year : int 2003 2003 2003 2003 2003 2003 2003 2003 2007 2008 ...  
## $ Site : Factor w/ 10 levels "Etang d'Ayes",..: 3 3 3 3 3 3 3 3 5 1 ...  
## $ Habitat : Factor w/ 2 levels "lake","pond": 1 1 1 1 1 1 1 1 1 1 ...  
## $ LarvalStage : Factor w/ 4 levels "Adult","metamorph",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ InfectionStatus : int 1 1 1 0 1 1 1 1 1 0 ...  
## $ AnnualaverageRf : num 2.73 2.73 2.73 2.73 2.73 ...  
## $ AnnualaverageTmax: num 12.2 12.2 12.2 12.2 12.2 ...  
## $ AnnualaverageTmin: num 2.35 2.35 2.35 2.35 2.35 ...  
## $ AnnualaverageTavg: num 7.29 7.29 7.29 7.29 7.29 ...  
## $ Springavgtemp : num 11.5 11.5 11.5 11.5 11.5 ...

The dataset includes the “InfectionStatus” (1=positive, 0=negative) of amphibians sampled from a range of lakes and ponds in the Pyrenees from 2003 to 2018. The data also includes annual rainfall and temperature climate variables (AnnualaverageRf= rainfaill in mm; AnnualaverageTmax, AnnualaverageTmin, AnnualaverageTavg and Springavgtemp in degrees celsius). In the next analysis, we are going to examine the relationship between average spring temperature on chyrtid infection status. Now, we have significantly more data points than the previous example so let’s see whether the separation effect of spring temperature is clear in the infection status.

scatterplot<-ggplot(chytrid, aes(x=Springavgtemp, y=InfectionStatus))+  
 geom\_point()+  
 labs(x= "Probability of Infection", y="Average Spring Temperature (Degrees Celsius)")+  
 theme\_classic()  
boxplot<- ggplot(chytrid, aes(x=factor(InfectionStatus), y=Springavgtemp))+  
 geom\_boxplot()+  
 theme\_classic()+  
 labs(x="Presence/Absence of Infection", y="Average Spring Temperature (Degrees Celsius)")  
ggarrange(scatterplot, boxplot, labels=c("A","B"), ncol=1, nrow=2)



This degree of separation here is less apparent, but the ecological research has indicated that increasing the spring temperature increases the probability of chytrid fungus infection in amphibians. The scatterplot makes this inference difficult because there is a lot of overlap across the average spring temperature, but we can model this with a binary generalised linear model.

##### Fitting the Model

M2<- glm(InfectionStatus~Springavgtemp, data = chytrid, family = "binomial")  
summary(M2)

##   
## Call:  
## glm(formula = InfectionStatus ~ Springavgtemp, family = "binomial",   
## data = chytrid)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.057236 0.055585 -1.030 0.303   
## Springavgtemp 0.052629 0.008447 6.231 4.65e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9310.0 on 6794 degrees of freedom  
## Residual deviance: 9270.7 on 6793 degrees of freedom  
## AIC: 9274.7  
##   
## Number of Fisher Scoring iterations: 4

anova(M2, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: InfectionStatus  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 6794 9310.0   
## Springavgtemp 1 39.254 6793 9270.7 3.722e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### Model Interpretation

From the summary we can interpret the model and construct linear equation:

We can calculate the flipping point:

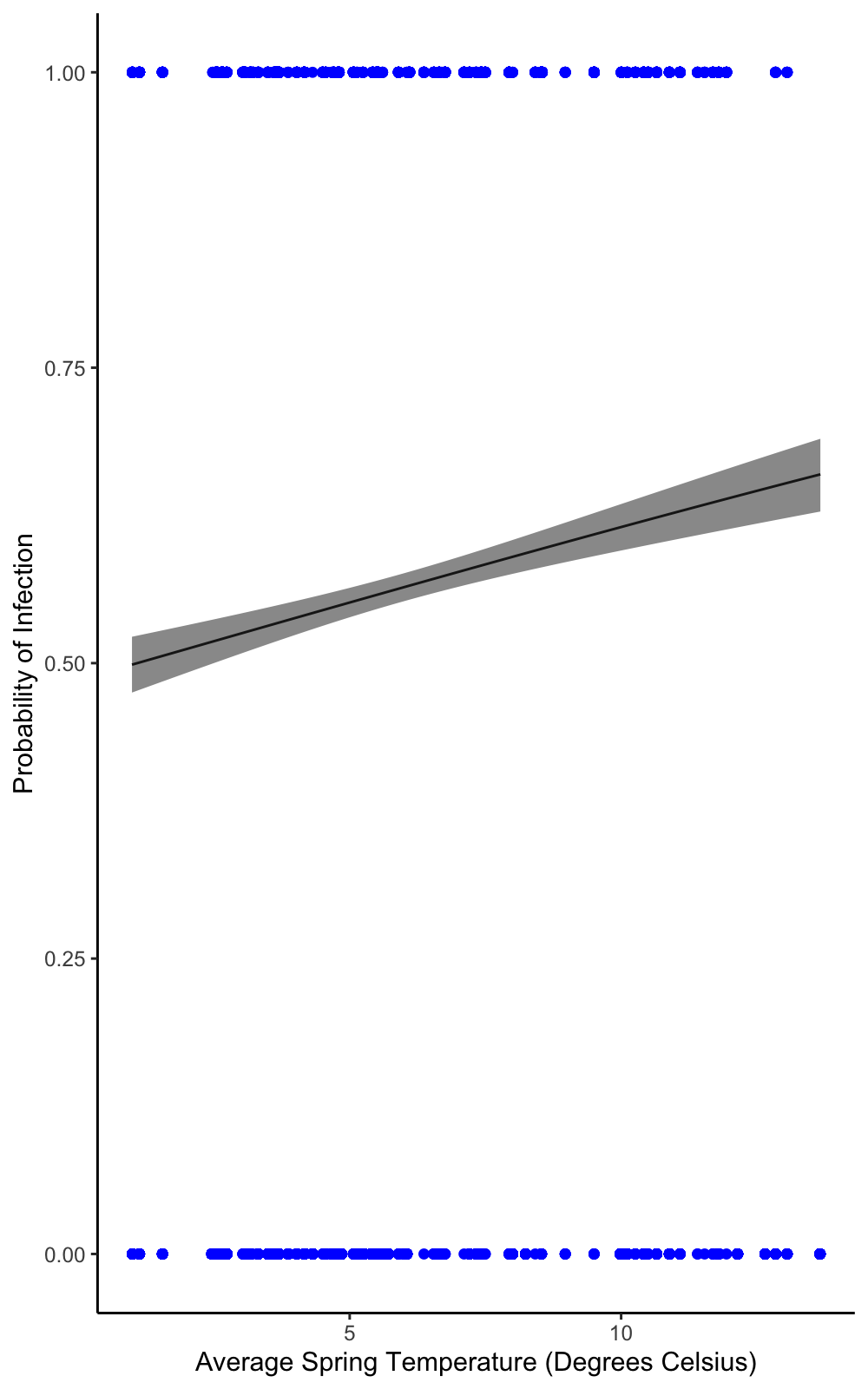
This allows us to infer that amphibians experiencing spring temperatures above 1.2 degrees celsius are more likely to be infected with chytrid. The last thing aspect to examine is the pseudo-R^2, which tells us that this model was able to explain 0.4% of variation in the presence/absence of the chytrid fungus.

##### Plotting the Model

range(chytrid$Springavgtemp) # Finding the range of Average Spring Temperature

## [1] 0.9968934 13.6638193

new\_data <- data.frame(Springavgtemp=seq(from=0.99, to=13.67, length=100))  
predictions<- predict(M2, 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=Springavgtemp, y=plogis(pred)))+   
 geom\_line(col="black")+  
 geom\_point(chytrid, mapping = aes(x=Springavgtemp, y=InfectionStatus), col="blue")+  
 geom\_ribbon(aes(ymin=plogis(lowerCI), ymax=plogis(upperCI), alpha=0.2), show.legend = FALSE)+   
 labs(y="Probability of Infection", x="Average Spring Temperature (Degrees Celsius)")+  
 theme\_classic()



This plot is informative and tells the reader about the effect of spring temperature on infection status and when combined with the linear equation and the summary output gives the reader all the information they need. However, it does appear the relationship is rather weak and chytrid fungus is very prevalent - our regression line is consistently above 0.5 regardless of spring averaget temperature.

Now in the next section, we’re going to analyse the same data but not as a binary outcome, but as a binomial outcome as:

### Binomial Models

chytrid\_binomial<- read.csv("chytrid\_binomial.csv", stringsAsFactors = T)  
str(chytrid\_binomial)

## 'data.frame': 175 obs. of 11 variables:  
## $ Year : int 2003 2007 2008 2008 2008 2008 2009 2009 2009 2009 ...  
## $ Site : Factor w/ 10 levels "Etang d'Ayes",..: 3 5 1 2 5 6 1 3 4 5 ...  
## $ Habitat : Factor w/ 2 levels "lake","pond": 1 1 1 1 1 1 1 1 1 1 ...  
## $ LarvalStage : Factor w/ 4 levels "Adult","metamorph",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Positives : int 7 129 0 43 23 0 0 157 192 298 ...  
## $ Total : int 8 141 41 51 34 61 35 251 485 531 ...  
## $ AverageRf : num 2.73 2.79 2.15 2.73 2.52 ...  
## $ AverageMaxTemp : num 12.17 11.86 12.64 7.03 11.03 ...  
## $ AverageMinTemp : num 2.348 2.339 2.582 -0.918 1.504 ...  
## $ AverageTemp : num 7.29 6.92 7.26 2.82 6.01 ...  
## $ AverageSpringTemp: num 11.54 7.94 4.69 5.42 6.55 ...

The dataset is a condensed version of the “chytrid.csv” dataset. The two new columns that are of relevance are “Positives” and “Total”. “Positives” are the number of positive samples per “Year”, “Site”, “Habitat” and “LarvalStage” andf the “Total” is the total number of samples. We can use these two values to formulate a binomial model to analyse whether average spring temperature affects the probability of chytrid infection. We have to feed the number of positives and the number of negatives into the *glm* function using cbind.

#### Fitting the Model

M3<- glm(cbind(Positives, Total-Positives)~AverageSpringTemp, data = chytrid\_binomial, family = "binomial")  
summary(M3)

##   
## Call:  
## glm(formula = cbind(Positives, Total - Positives) ~ AverageSpringTemp,   
## family = "binomial", data = chytrid\_binomial)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.403670 0.037311 -10.82 <2e-16 \*\*\*  
## AverageSpringTemp 0.088839 0.005572 15.94 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5055.4 on 174 degrees of freedom  
## Residual deviance: 4795.7 on 173 degrees of freedom  
## AIC: 5410.5  
##   
## Number of Fisher Scoring iterations: 4

anova(M3, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: cbind(Positives, Total - Positives)  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 174 5055.4   
## AverageSpringTemp 1 259.64 173 4795.7 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### Model Interpretation

From the summary we can interpret the model and construct linear equation:

The slope coefficient tells us that for every one degree celsius increase in average spring temperature the chytrid infection increases by log odds of 0.09 or an odds of 1.09. We can work out the flip point here too:

This suggests an average spring temperature of more than 4.44 degrees celsius is likely to lead to greater chytrid infections.

We can examine the pseudo-R^2, which tells us that this model was able to explain 5% of variation in the probability of chytrid fungus infection.

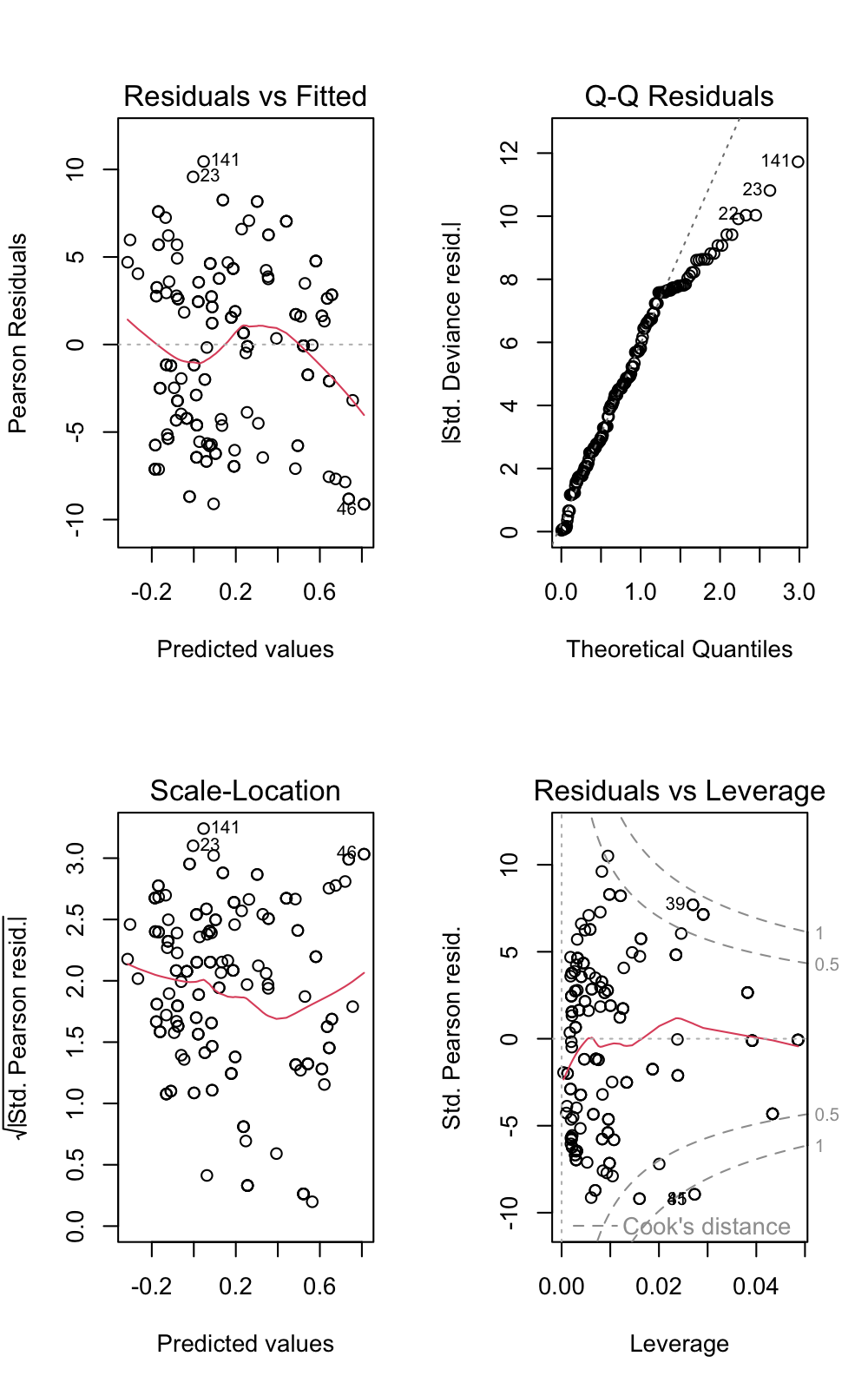
#### Model Validation

So let’s look at the dispersion parameter first:

The model is clearly overdispersed and there are numerous other reasons why this could be:

* Too simplistic: **there are a number of variables that could be random effects (“Year” and/or “Site”), fixed factors (“Habitat” and “LarvalStage”) and continuous covariates (“AverageRf”)**.
* One or zero outliers: **potentially (see diagnostic plots below)**

par(mfrow=c(2,2))  
plot(M3)



The “Residuals vs Leverage” plot suggests that the model may have a number outliers that are causing this overdispersion.

sum(cooks.distance(M3)>1)

## [1] 2

In total, 2 outliers have be identified as these have a Cook’s distance above 1. We could explore the options of adding covariates and/or random effects and you can explore these. For now, we are going to fit a quasi-binomial model.

#### Fitting a Quasi-Binomial Model

M4<- glm(cbind(Positives, Total-Positives)~AverageSpringTemp, data = chytrid\_binomial, family = "quasibinomial")  
summary(M4)

##   
## Call:  
## glm(formula = cbind(Positives, Total - Positives) ~ AverageSpringTemp,   
## family = "quasibinomial", data = chytrid\_binomial)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.40367 0.18066 -2.234 0.0267 \*   
## AverageSpringTemp 0.08884 0.02698 3.293 0.0012 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasibinomial family taken to be 23.44436)  
##   
## Null deviance: 5055.4 on 174 degrees of freedom  
## Residual deviance: 4795.7 on 173 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 4

anova(M4, test="F") # for quasi approaches we use the F test

## Analysis of Deviance Table  
##   
## Model: quasibinomial, link: logit  
##   
## Response: cbind(Positives, Total - Positives)  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev F Pr(>F)   
## NULL 174 5055.4   
## AverageSpringTemp 1 259.64 173 4795.7 11.075 0.001069 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

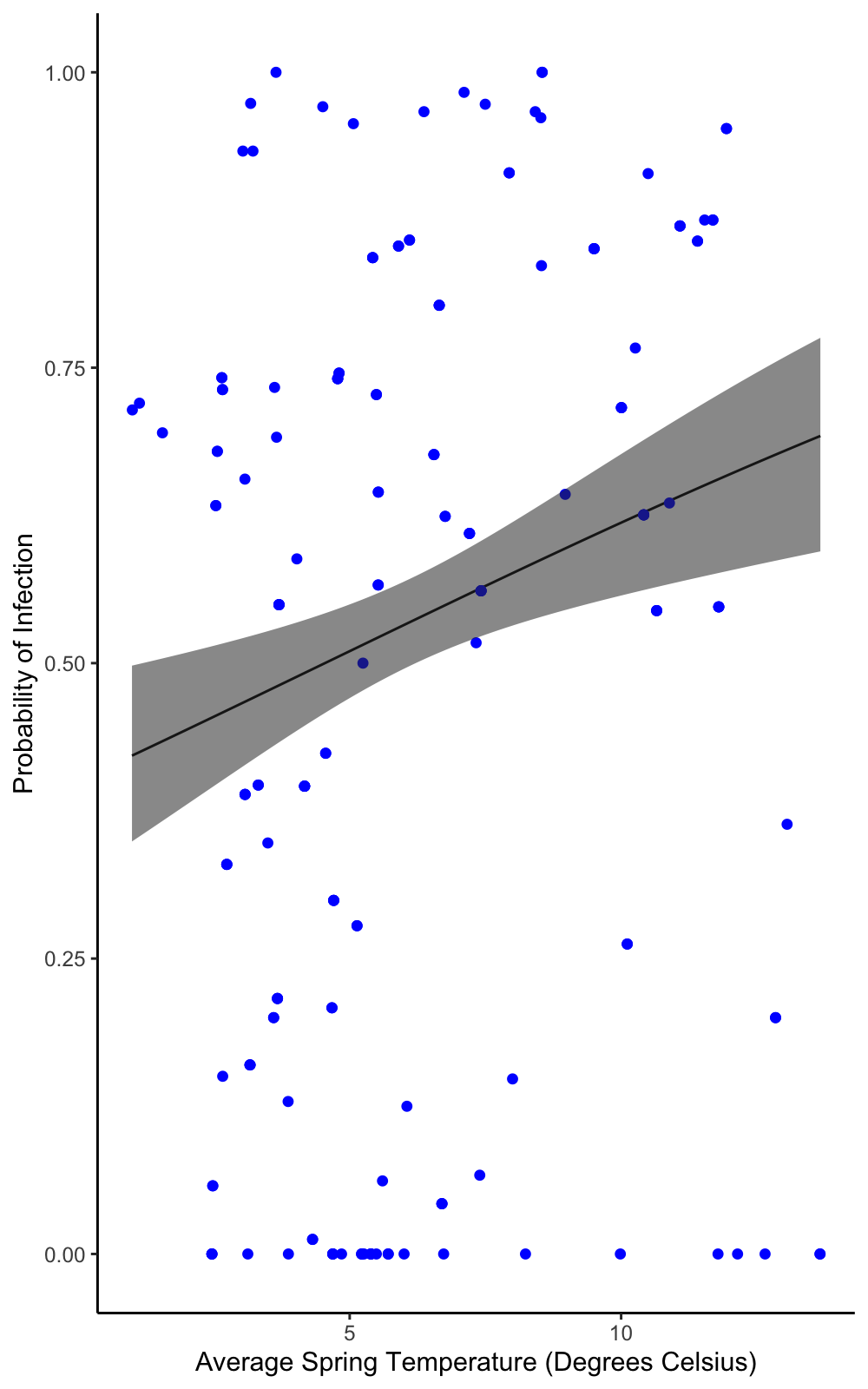
What you will notice from the output is that the estimate values do not change but the standard errors are inflated. This will be the model we will plot.

#### Plotting the Model

range(chytrid\_binomial$AverageSpringTemp) # Finding the range of Average Spring Temperature

## [1] 0.9968934 13.6638193

new\_data <- data.frame(AverageSpringTemp=seq(from=0.99, to=13.67, length=100))  
predictions<- predict(M4, 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=AverageSpringTemp, y=plogis(pred)))+   
 geom\_line(col="black")+  
 geom\_point(chytrid\_binomial, mapping = aes(x=AverageSpringTemp, y=(Positives/Total)), col="blue")+  
 geom\_ribbon(aes(ymin=plogis(lowerCI), ymax=plogis(upperCI), alpha=0.2), show.legend = FALSE)+   
 labs(y="Probability of Infection", x="Average Spring Temperature (Degrees Celsius)")+  
 theme\_classic()



### Revisiting the Bee Mites Data

Previously, we fitted a Poisson model to this data and concluded it might not have been the appropriate model family and a binomial model would be better.

#### Fitting the Model

mites<- read.csv("bee\_mites.csv")  
M5<- glm(cbind(Dead\_mites, Total-Dead\_mites)~Concentration, data = mites, family = "binomial")  
summary(M5)

##   
## Call:  
## glm(formula = cbind(Dead\_mites, Total - Dead\_mites) ~ Concentration,   
## family = "binomial", data = mites)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.8728 0.1670 -5.227 1.73e-07 \*\*\*  
## Concentration 2.9687 0.3275 9.065 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 347.77 on 114 degrees of freedom  
## Residual deviance: 194.82 on 113 degrees of freedom  
## AIC: 294.85  
##   
## Number of Fisher Scoring iterations: 5

anova(M5, test = "Chisq")

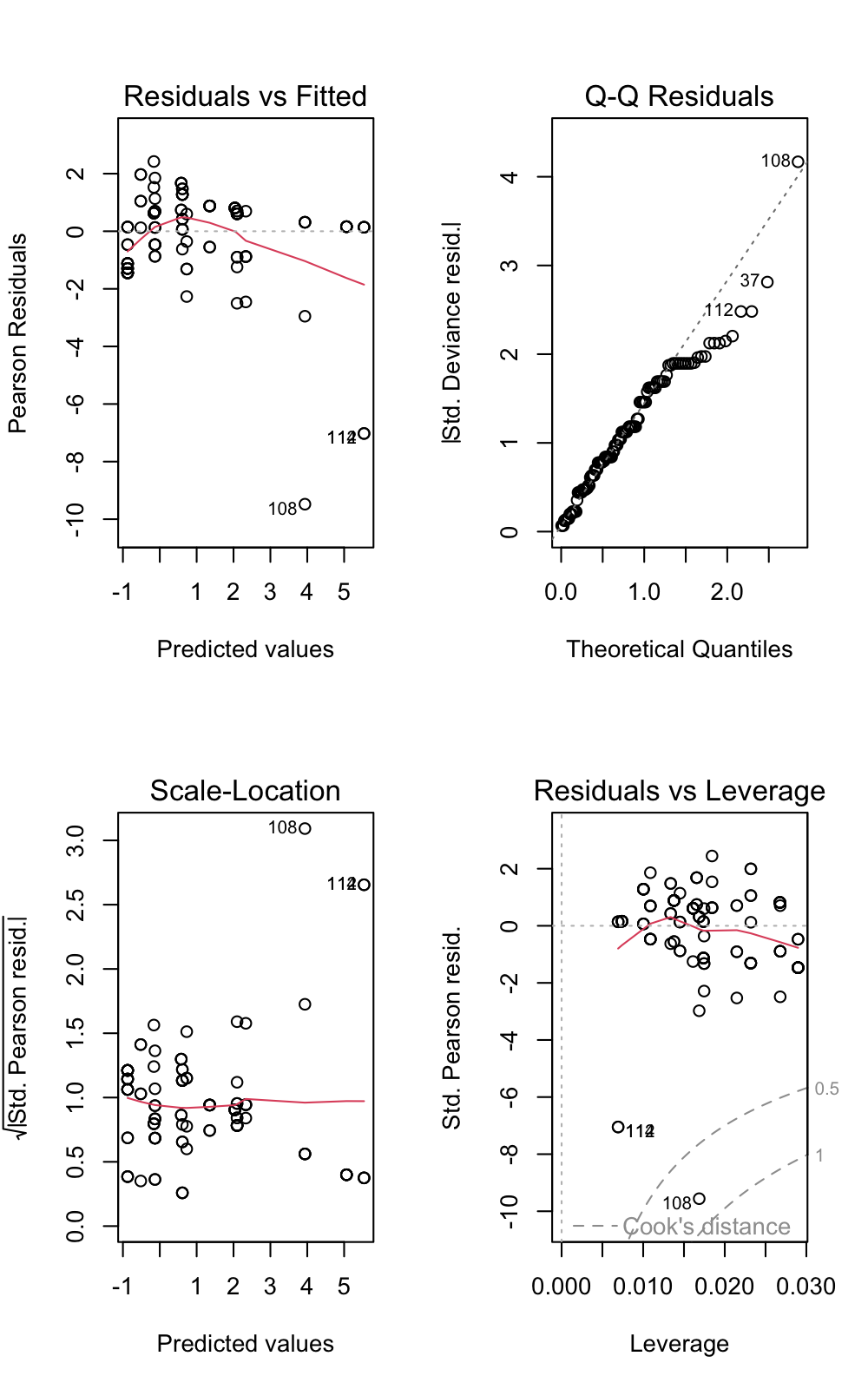
## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: cbind(Dead\_mites, Total - Dead\_mites)  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 114 347.77   
## Concentration 1 152.95 113 194.82 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### Model Validation

So let’s look at the dispersion parameter first:

The model is overdispersed and we know there are numerous other reasons why this could be. But let’s explore the model diagnostics first:

par(mfrow=c(2,2))  
plot(M5)



We can see from these plots that the previous criticism of unequal variances in the “Residuals vs Fitted” and the “Scale-Location” plots is not apparent and therefore changing the model family has corrected for this.

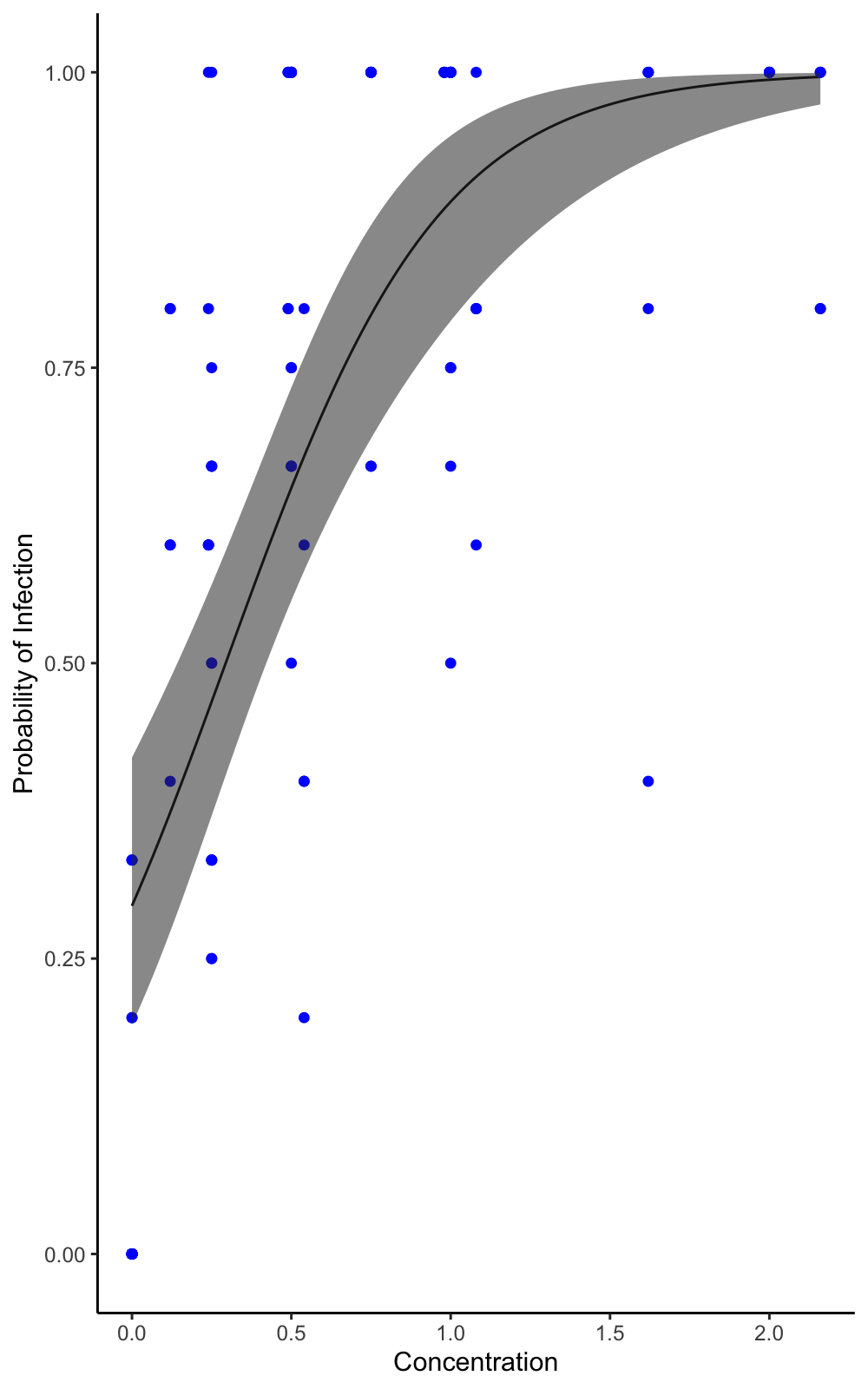
We could fit a quasi-binomial model to account for the overdispersion in the binomial model and let’s make a plot from this model.

### Refitting the Quasi-binomial Model and Plotting

M6<- glm(cbind(Dead\_mites, Total-Dead\_mites)~Concentration, data = mites, family = "quasibinomial")  
  
range(mites$Concentration)

## [1] 0.00 2.16

new\_data <- data.frame(Concentration=seq(from=0, to=2.16, length=100))  
predictions<- predict(M6, 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=plogis(pred)))+   
 geom\_line(col="black")+  
 geom\_point(mites, mapping = aes(x=Concentration, y=(Dead\_mites/Total)), col="blue")+  
 geom\_ribbon(aes(ymin=plogis(lowerCI), ymax=plogis(upperCI), alpha=0.2), show.legend = FALSE)+   
 labs(y="Probability of Infection", x="Concentration")+  
 theme\_classic()



## 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. Endemicity on the Galapagos islands (“gala.txt”):

* How does area of the island affect the endemicity (the proportion/probability of endemic species out of total 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 Endemicity~Area and Endemicity~log(Area) to see what I mean.
* You will have to use cbind to make the binomial odds ratio.