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: 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

Binary Models

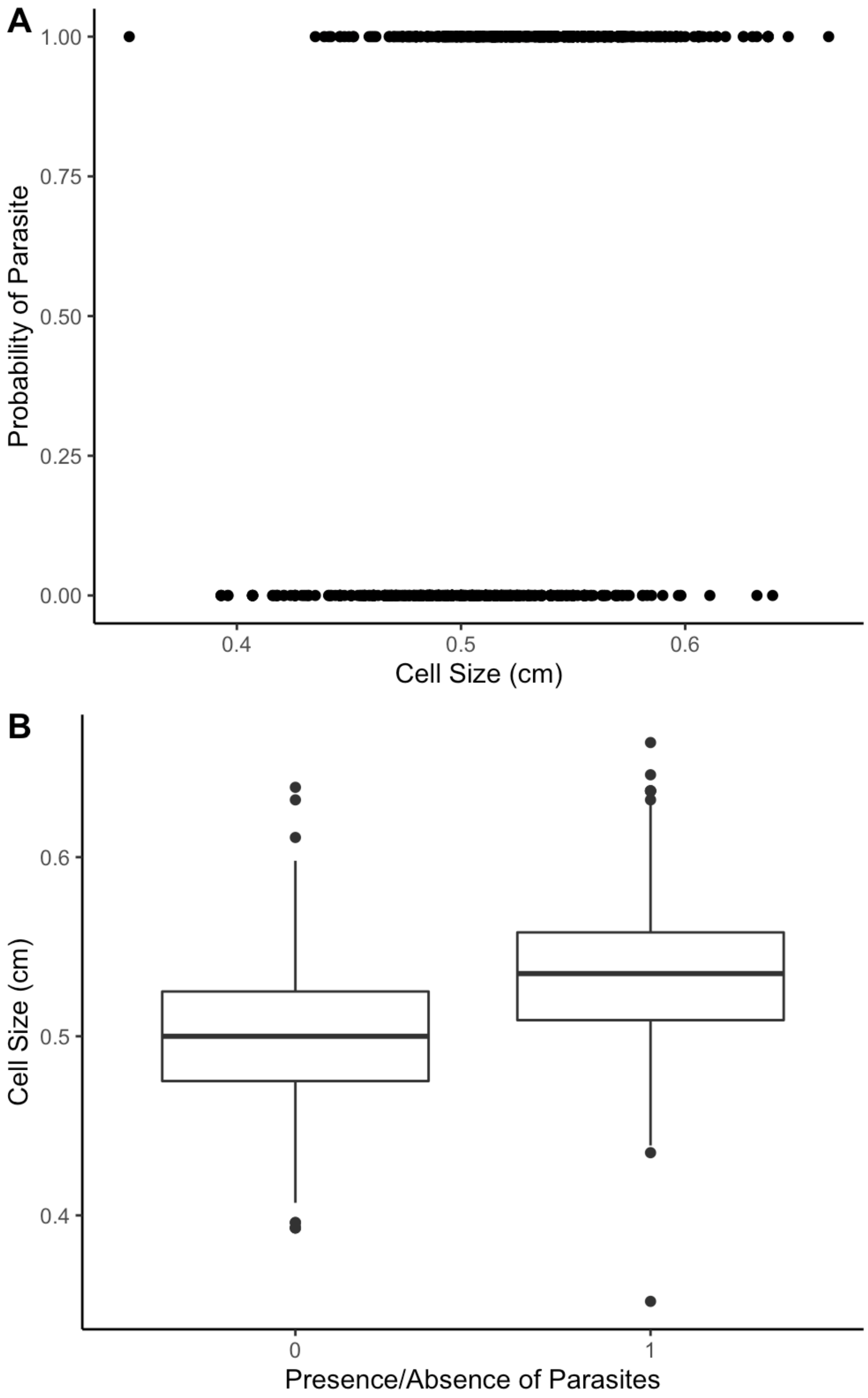
Varoa spp in Honeycomb Cells

```
require(ggplot2)

## Loading required package: ggplot2

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

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)
```


A
Probability of Parasite
1.00
0.75
0.50
0.25
0.00
0.4
0.5
0.6
Cell Size (cm)
B
Cell Size (cm)
0.6
0.5
0.4
0
1
Presence/Absence of Parasites

Fitting the Model

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

```
##
## Call:
## glm(formula = Parasites ~ CellSize, family = "binomial", data =
worker)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4403  -1.0570   0.5837   0.9878   2.6346
##
## 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:

$$logit(ProbabilityofParasites) = -11.25 + 22.18 * CellSize$$

This line equation and the summary output tells us that increasing cell size of honeycomb increases the probability being infected by *Varoa destructor*. A taking the inverse logit of 22.18 using the function *plogis* and we get the value of 1, meaning "for every centimeter increase in honeycomb cell size, the probability of being infected by the *Varoa destructor* mite increased by a factor of 1 or 100%".

```
plogis(coef(M1))

##  (Intercept)     CellSize
## 1.307548e-05 1.000000e+00
```

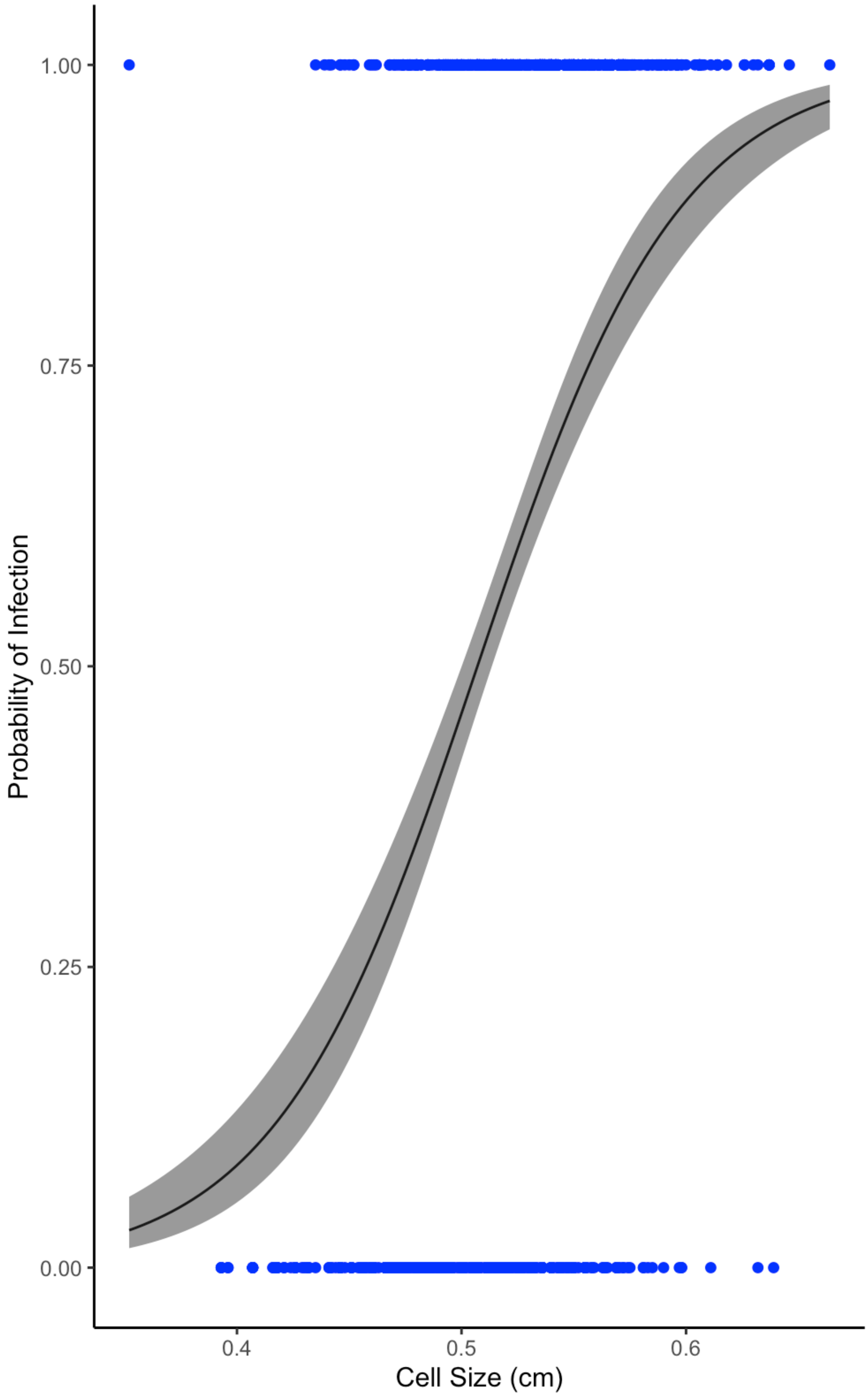
In the lecture, we examined this idea of identifying the value of *x* (here: Cell Size) where the probability flips from less likely to be infected or more likely to be infected. We can use the table presented in the lecture to do the following calculation:

$$\frac{\beta_0}{\beta_1} = \frac{11.25}{22.18} = 0.51cm$$

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()
```


1.00
0.75
0.50
0.25
0.00
Probability of Infection
0.4
0.5
0.6
Cell Size (cm)

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.

$$PseudoR^2 = 1 - (1104.9/1259.6) = 0.12$$

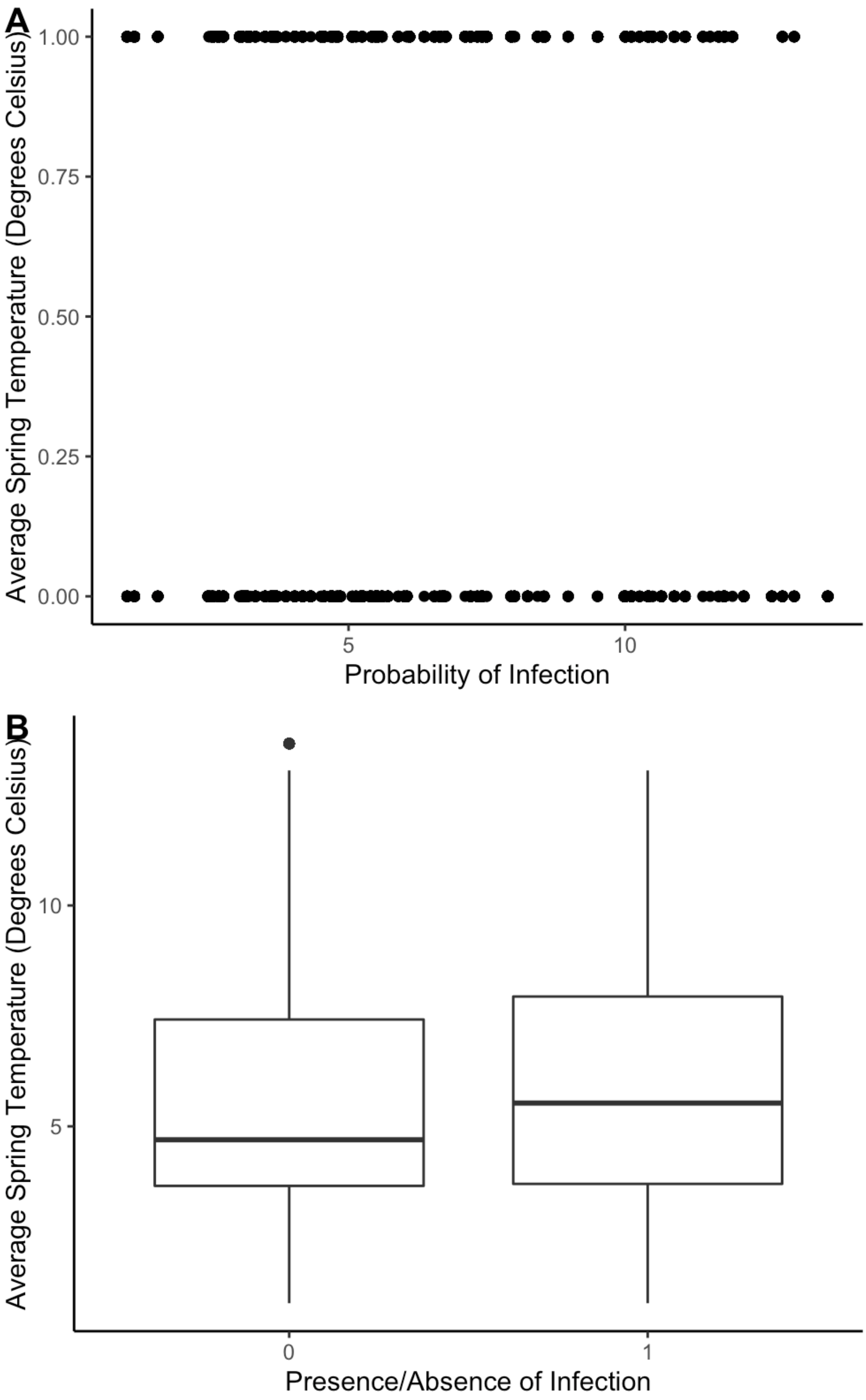
Chrytrid Infection Status in the Pyrenees

```
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)
```


A
Average Spring Temperature (Degrees Celsius)
1.00
0.75
0.50
0.25
0.00
5
10
Probability of Infection
B
Average Spring Temperature (Degrees Celsius)
10
5
0
1
Presence/Absence of Infection

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)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4683  -1.2470   0.9772   1.0860   1.1794
##
## 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:

$$logit(ProbabilityofInfection) = -0.06 + 0.05 * AverageSpringTemperature$$

We can calculate the flipping point:

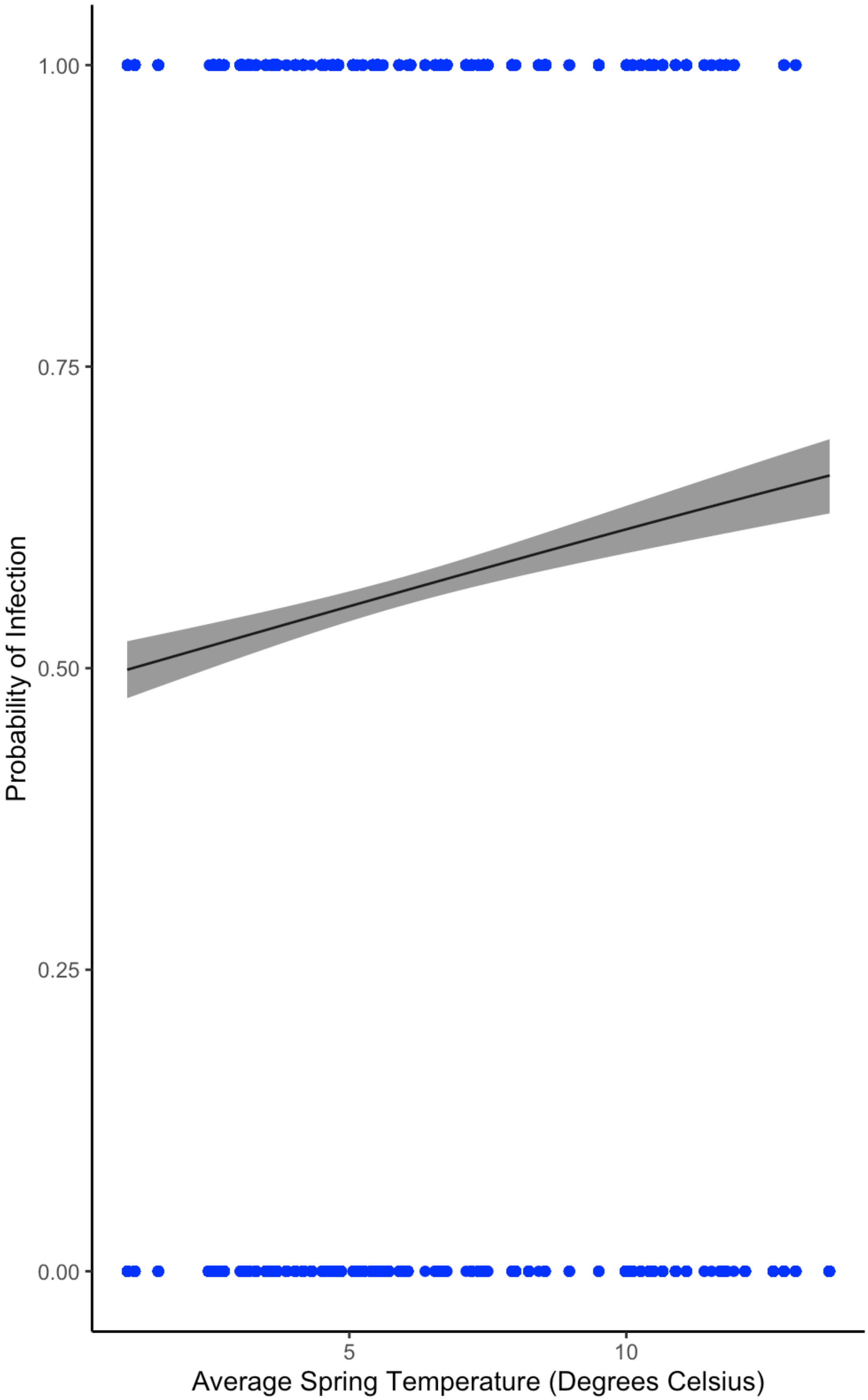
$$\frac{\beta_0}{\beta_1} = \frac{0.06}{0.05} = 1.2degreescelsius$$

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.

$$PseudoR^2 = 1 - (9270.7/9310.0) = 0.004$$

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()
```


1.00
0.75
0.50
0.25
0.00
Probability of Infection
5
10
Average Spring Temperature (Degrees Celsius)

This plot is really 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. 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:

$$\frac{NumberofPositives}{NumberofNegatives}$$

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)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.9963  -4.6700  -0.0673   3.2884  11.6684
##
## 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:

$$logit(Probabilityof Infection) = -0.4 + 0.09 * AverageSpringTemperature$$

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.

$$PseudoR^2 = 1 - (4795.7/5055.4) = 0.05$$

Model Validation

So let’s look at the dispersion parameter first:

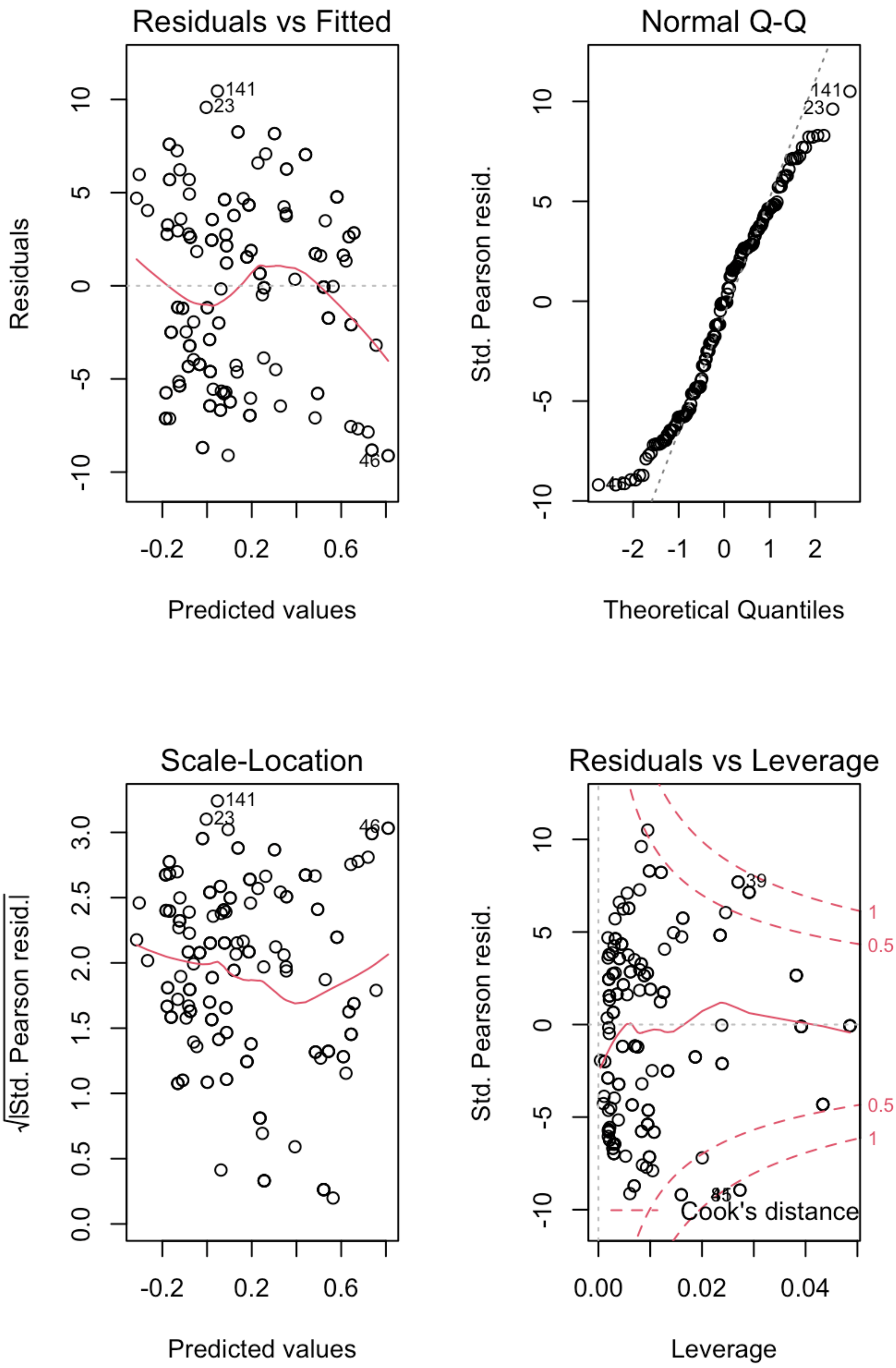
$$DispersionParameter = 4795.7/173 = 27.72$$

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)**

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


Residuals vs Fitted
Residuals
Predicted values
-0.2
0.2
0.6
-10
-5
0
5
10
141
23
46
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
-2
-1
0
1
2
-10
-5
0
5
10
141
23
46
Scale-Location
√|Std. Pearson resid.|
Predicted values
-0.2
0.2
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
141
23
46
Residuals vs Leverage
Std. Pearson resid.
Leverage
0.00
0.02
0.04
-10
-5
0
5
10
39
85
1
0.5
0.5
1
Cook's distance

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

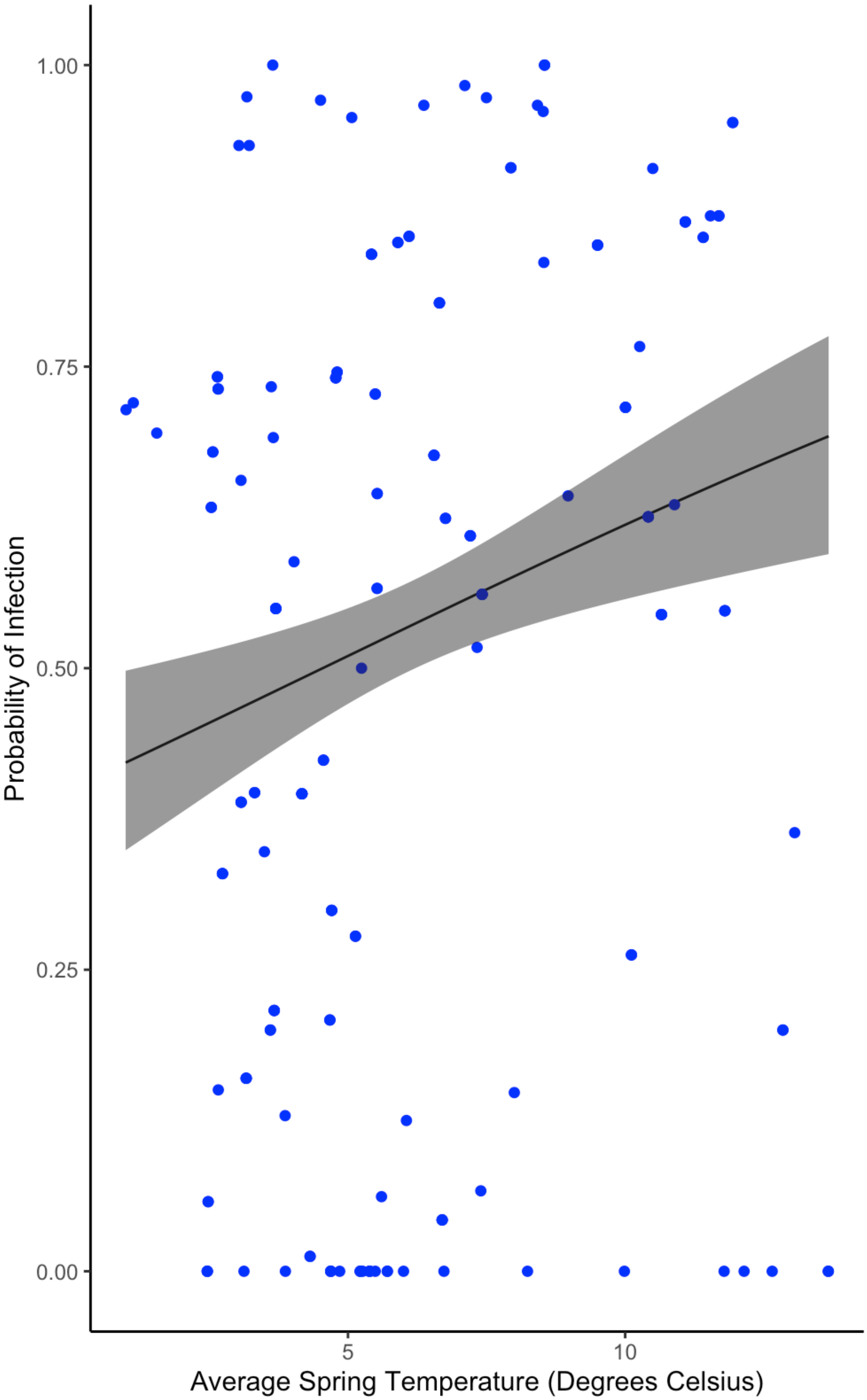
```
M4<- glm(cbind(Positives, Total-Positives)~AverageSpringTemp, data =
chytrid_binomial, family = "quasibinomial")
summary(M4)

##
## Call:
## glm(formula = cbind(Positives, Total - Positives) ~
AverageSpringTemp,
##     family = "quasibinomial", data = chytrid_binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.9963  -4.6700  -0.0673   3.2884  11.6684
##
## 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="Chisq")

## 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  Pr(>Chi)
## NULL                                174     5055.4
## AverageSpringTemp  1   259.64       173     4795.7 0.0008751 ***
## ---
## 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()
```


Probability of Infection
1.00
0.75
0.50
0.25
0.00
5
10
Average Spring Temperature (Degrees Celsius)

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)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1331  -0.8957   0.2244   0.9934   2.7866
##
## 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
```

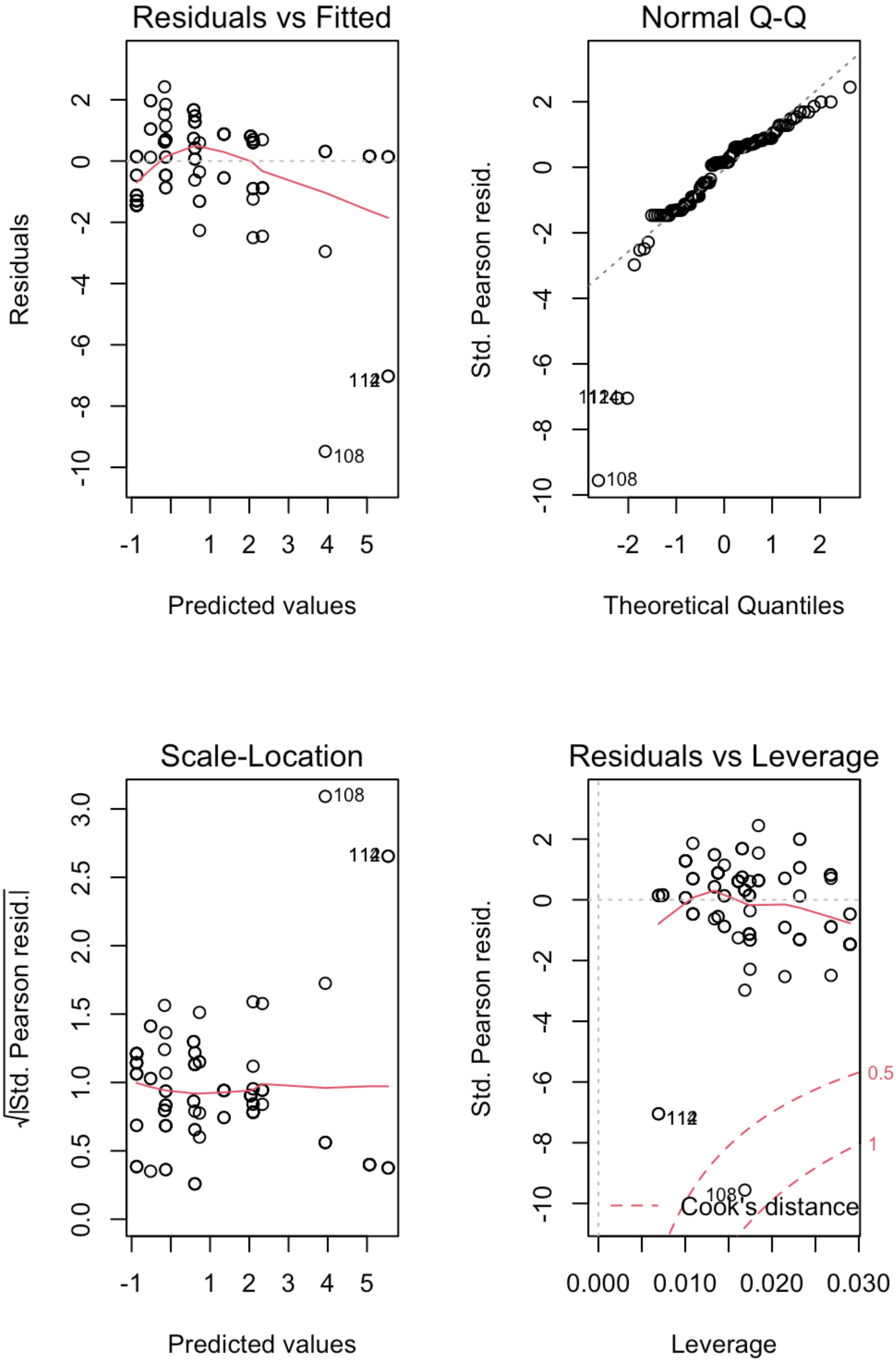
Model Validation

So let’s look at the dispersion parameter first:

$$DispersionParameter = 194.82/113 = 1.72$$

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)
```

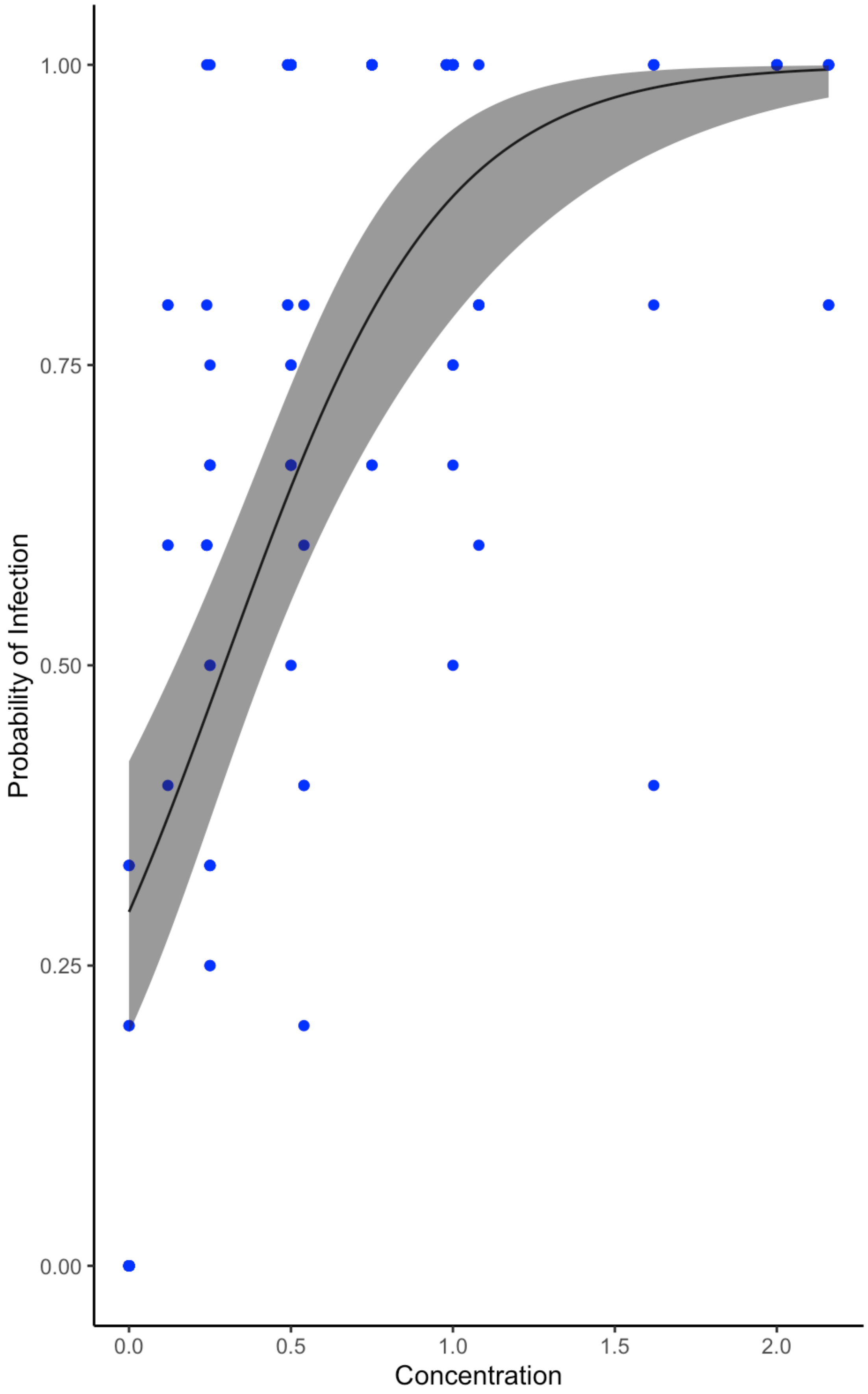

Residuals vs Fitted
Residuals
2
0
-2
-4
-6
-8
-10
Predicted values
-1
1
2
3
4
5
112
108
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
-2
-1
0
1
2
11240
108
Scale-Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Predicted values
108
112
Residuals vs Leverage
Std. Pearson resid.
Leverage
0.000
0.010
0.020
0.030
0.5
1
112
108
Cook's distance

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()
```


Probability of Infection
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
Concentration

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/probabilty 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.