Homework 2 - Probability models

Juliette J. Rubin

January 30, 2019

Question 1

Seedling survival in Thailand. We will first address questions 1a-b.

```
seedsurv <- read.csv("SEEDLING_SURVIVAL.csv")

#Tree height and seedling survival


#intercept: -0.063, slope: 0.14


#assessing strength of the relationship
yhat<--0.063+(0.14)*seedsurv$HEIGHT
y<- seedsurv$survival
m1<-glm(seedsurv$survival~seedsurv$HEIGHT, family="binomial")
coef(m1)
```

```
(Intercept) seedsurv$HEIGHT
-0.06271111      0.14071141
```

```
confint(m1)
```

```
                2.5 %    97.5 %
```

(Intercept) -0.5791061 0.4268167 seedsurv$HEIGHT 0.1038803 0.1815477

```
MAE<-function(yhat,y){
  return(mean(abs(y-yhat)))
}

MAE(yhat,y)
```

[1] 1.636803

```
#RMSE: 1.64

#Light level and seedling survival
m2 <- glm(seedsurv$survival~seedsurv$LIGHT, family="binomial")
coef(m2)
```

(Intercept) seedsurv$LIGHT 2.66194692 -0.06552684

```
confint(m2)
```

```
                2.5 %     97.5 %
```

(Intercept) 2.25136434 3.0876309 seedsurv$LIGHT -0.09841747 -0.0325795

```
#intercept: 2.66, slope: -0.066

#assessing the strength of the relationship
```

```
yhat<-2.66+(-0.066)*seedsurv$LIGHT
y<- seedsurv$survival

MAE<-function(yhat,y){
  return(mean(abs(y-yhat)))
}

MAE(yhat,y)
```

[1] 1.078313

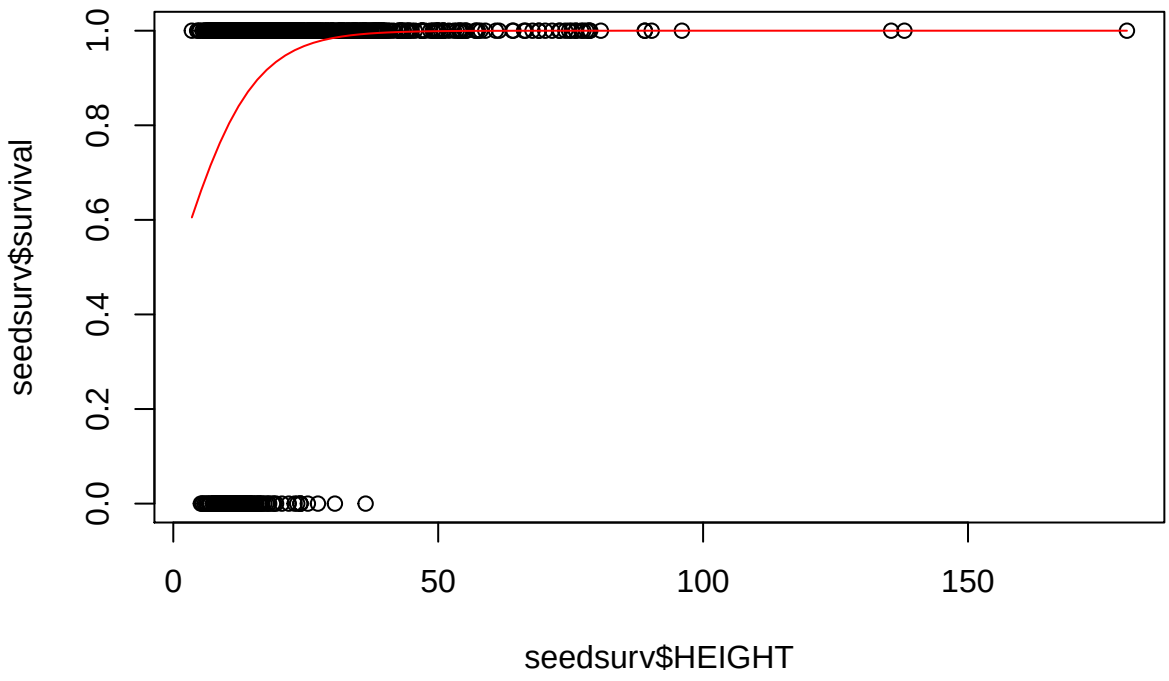
```
#RMSE: 1.08
```

Since the 95% confidence interval for height doesn't cross zero, it seems that height has a significant effect on seedling survival. The 95% confidence interval for light also does not cross zero and therefore has a significant effect. Interestingly, however, the MAE for light is lower than for height. Therefore it seems like light is the better predictor.

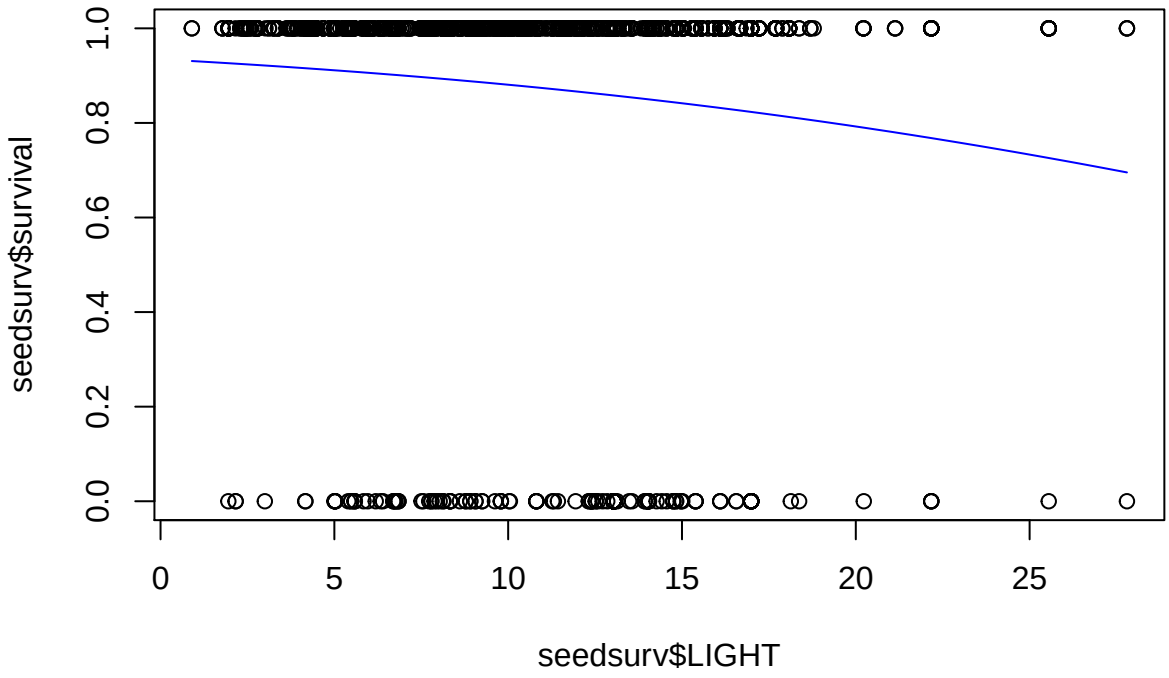
The baseline seedling survival rate is 14% in the tree height model and is 6.6% in the light model.

Below, we will visualize the data and the best-fit regression line.

```
plot(seedsurv$survival~seedsurv$HEIGHT)
curve(plogis(-0.063+0.14*x),add=T,col="red")
```




```
plot(seedsurv$survival~seedsurv$LIGHT)
curve(plogis(2.66+-0.066*x),add=T,col="blue")
```

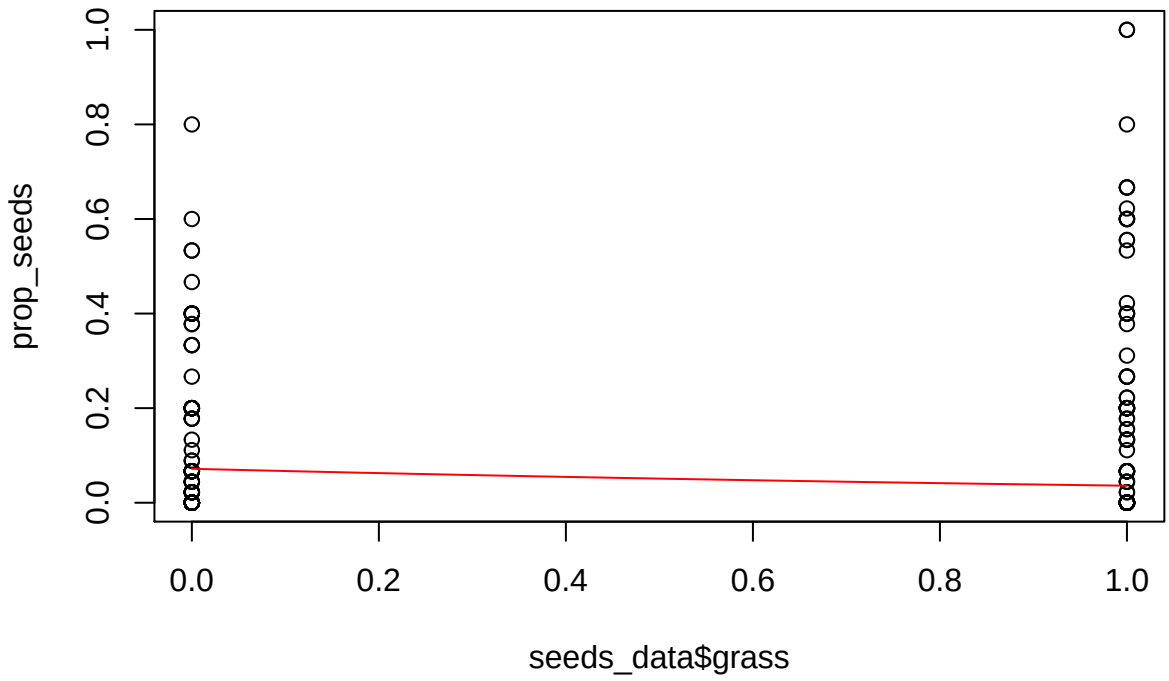


Question 2

We will now test the effect of grass in a plot on seedling survival success.

```
#read in data
seeds_data<-read.csv("seeds.csv")

#making data proportional by successes
prop_seeds<-seeds_data$recruits/seeds_data$seeds

#plotting the success of seedlings by grass in plot
plot(prop_seeds~seeds_data$grass)

#making a data matrix of the successes and duds
seed_success<-cbind(seeds_data$recruits,seeds_data$seeds-seeds_data$recruits)

#making a binomial model
seedsModel<-glm(seed_success~seeds_data$grass, family="binomial")

coef(seedsModel)
#Intercept: -2.56, slope: 0.728 -- when baseline grass is zero, 72.8% of seedlings survive
#0.728/4 = 0.182, thus the maximal effect of grass on seedling survival is 18.2%

plogis(-2.56)
#at baseline grass density, 7.3% of seedlings are recruited
```

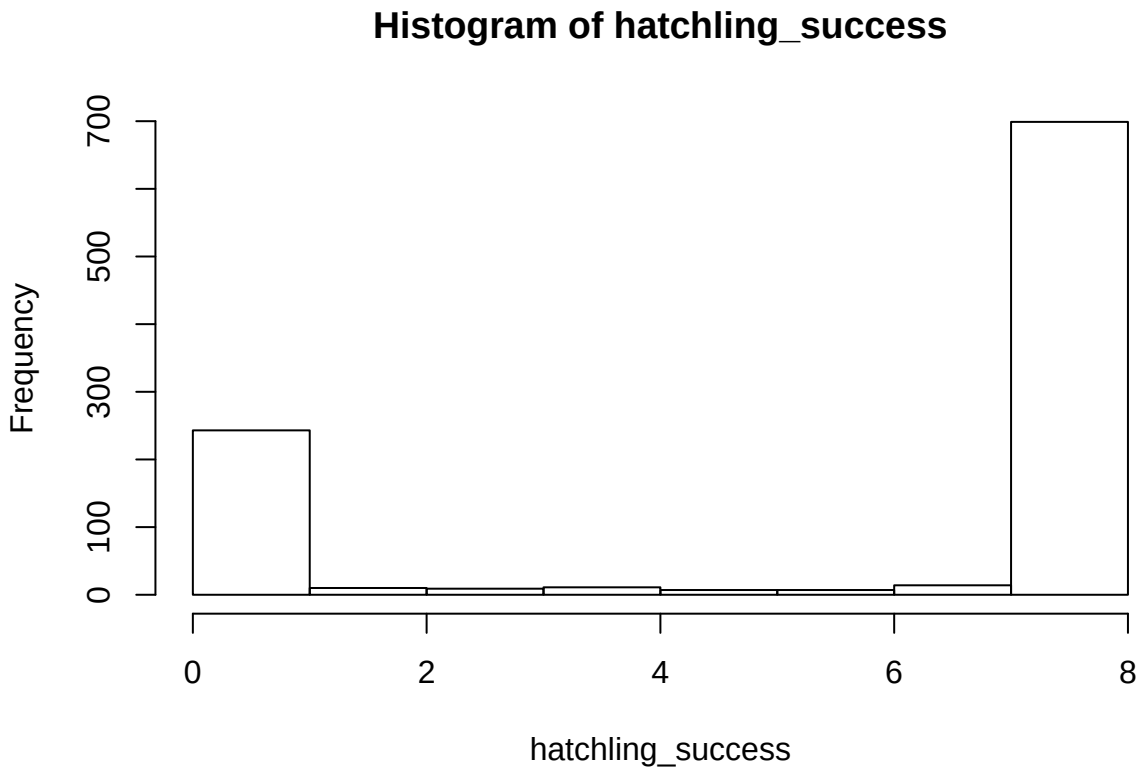
```r
confint(seedsModel)
#the slope does not cross zero
#0.558/4 = 0.14, 0.899/4 = 0.22 -- so the 95% CI for the maximum effect of grass on seedling recruitmen

plogis(confint(seedsModel))
#Baseline seedling success will range from 6.3%-8.07% (i.e., not huge variation)



#make plot
plot(prop_seeds~seeds_data$grass)
curve(plogis(-2.56+-.728*x),add=T,col="red")
```



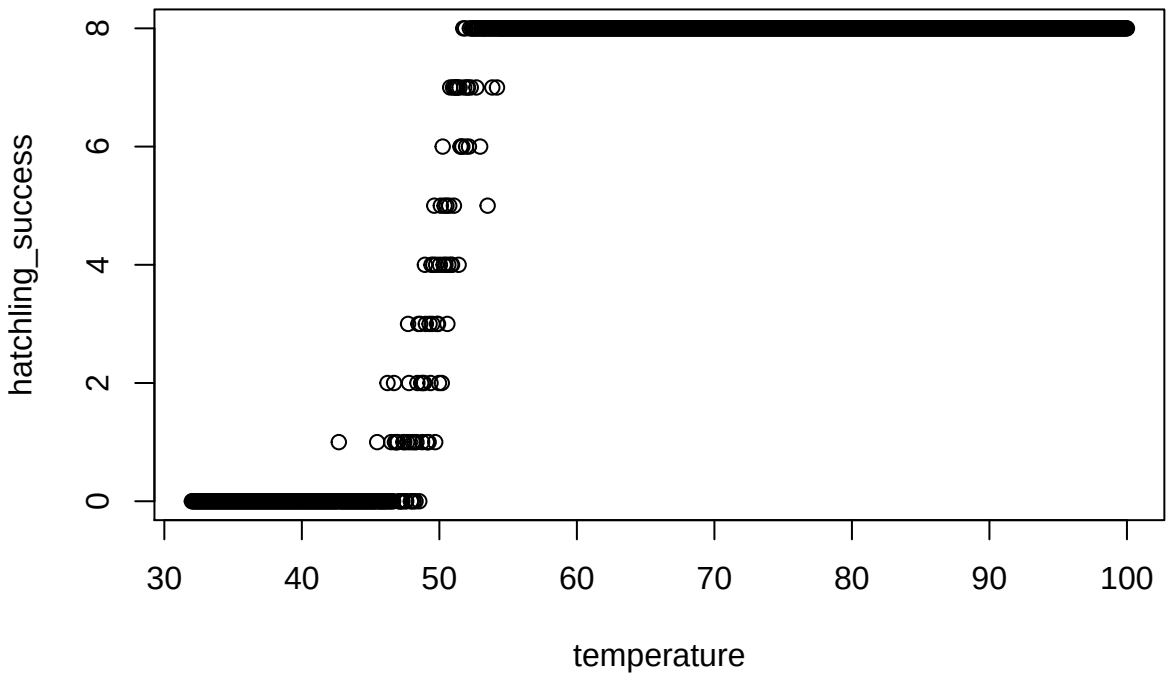
Becuase the maximal effect of grass density on seedling survival is 18.2% we can determine that grass has a fairly weak influence on seedling recruitment.

```r
to_be_filled=rep(NA,100)
to_be_filled[1]=1
for(i in 2:100) {to_be_filled[i]=to_be_filled[i-1]+1}

#length of predictor has to be the same as sample size
intercept=-50
slope=1
sample_size=1000
temperature=seq(from=32, to=100, length=1000)
hatchling_success=rbinom(n=sample_size, prob=plogis(intercept+slope*temperature), size=8)
```

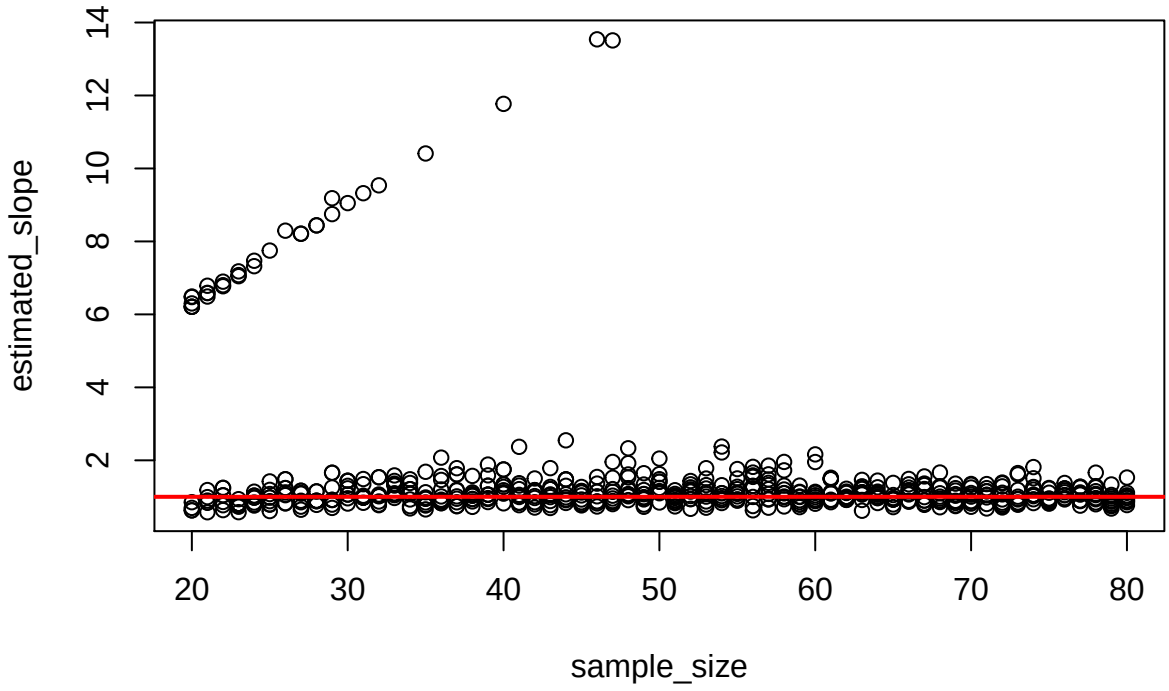
```r
hist(hatchling_success)
```




```r
plot(hatchling_success~temperature)
```




```
y=cbind(hatchling_success, 8-hatchling_success)
m1=glm(y~temperature, family="binomial")

coef(m1)

1.0108/4
#.2527 --> for every 1 degree rise in temperature the likelihood of nestling survival increases by 25%


#power analysis
sample_size = rep(seq(from=20, to=80), times=10)
estimated_slope=rep(NA,times=length(sample_size))
for(j in 1:length(sample_size)) {
  y=rbinom(n=sample_size[j],prob=plogis(intercept+slope*seq(from=32, to=100, length=sample_size[j])),si
  response=cbind(y,8-y)
  m1<-glm(response~seq(from=32, to=100, length=sample_size[j]),family="binomial")
  estimated_slope[j]=coef(m1)[2]
}

plot(estimated_slope~sample_size)
abline(h=1,col="red",lwd=2)
```



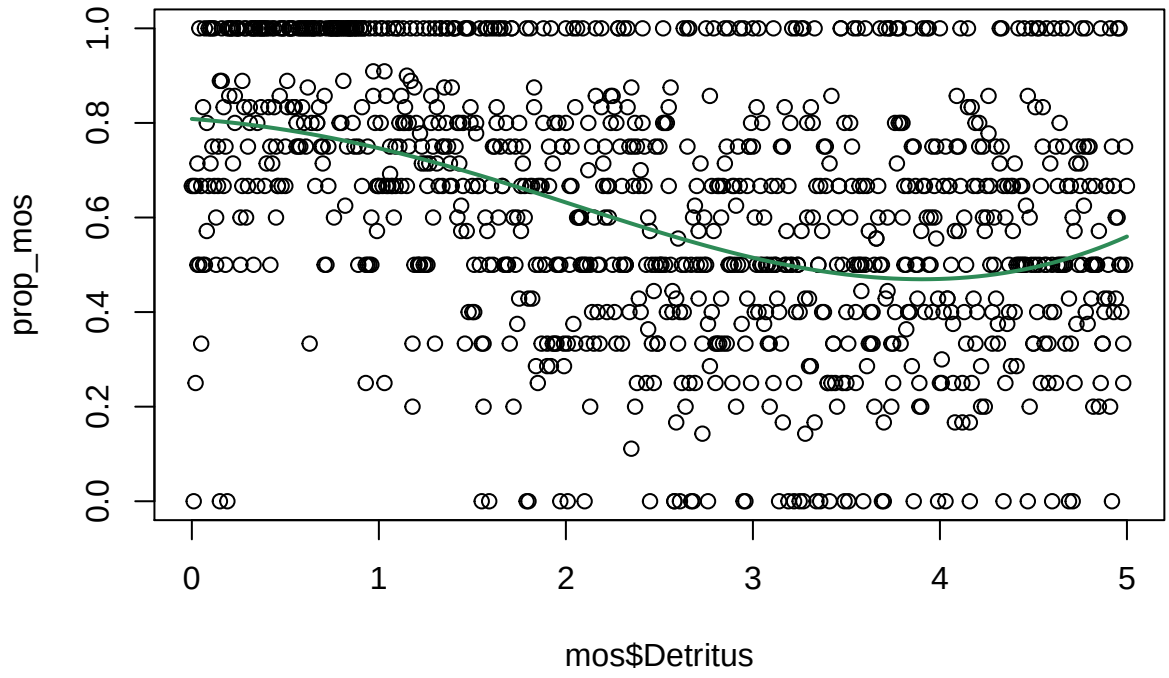

```
#running 60 different glms, one for each sample
```

coef(m1) (Intercept) temperature -50.520465 1.010844

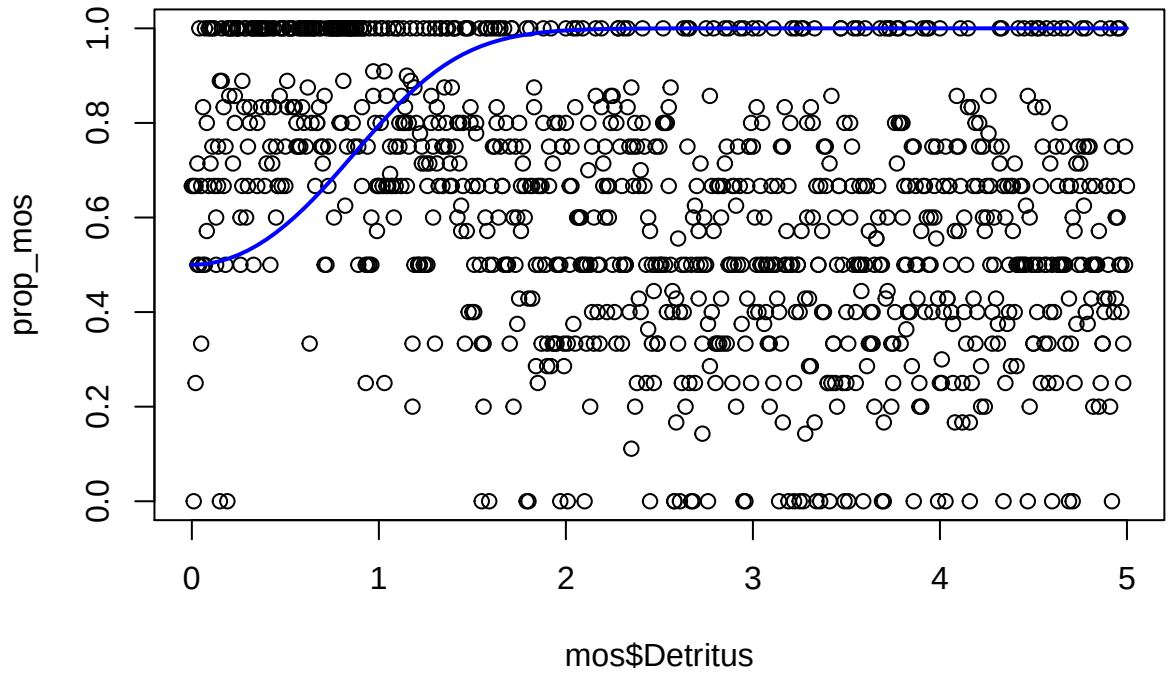
Question 3

We will now address the effect of detritus on mosquito population growth.

```
mos<-read.csv("mosquito_data.csv")
head(mos)

#turning the data into a proportion
prop_mos<-mos$Emergent_adults/mos$Egg_Count

#creating a plot for proportional mosquito data
plot(prop_mos~mos$Detritus)
#adding the polynomial curve
curve(plogis(1.44-(0.19*x)-(.21*(x^2))+(0.04*(x^3))),add=T,col="seagreen",lwd=2)
```




```
plot(prop_mos~mos$Detritus)
#Ricker curve
curve(plogis(10*x*((exp(1))^-2*x)),add=T,col="blue",lwd=2)
```



c) In the Ricker model we see a positive relationship between detritus and proportion of mosquitos that make it to adulthood. We also find a plateau in the proportion of mosquitos that make it to adulthood around 2units of detritus. In the polynomial model the effect of detritus has a much more varied effect, where it is negatively associated with proportion of adult mosquitos in low quantities, then this effect gets stronger between about 2 and 4 units of detritus and finally around 5 units the effect becomes positive.

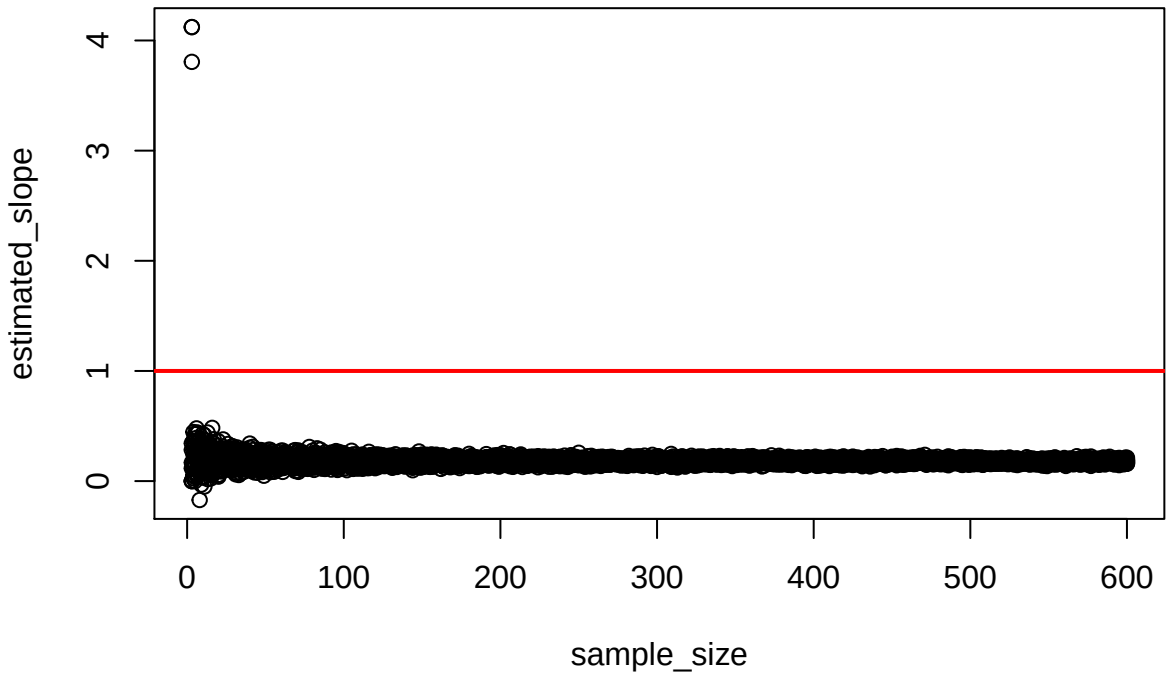
```
polynomial_likelihood<--sum(dbinom(x=mos$Emergent_adults,size=mos$Egg_Count,prob=plogis(1.44-(0.19*mos$
polynomial_likelihood
#1415.63


Ricker_likelihood<--sum(dbinom(x=mos$Emergent_adults,size=mos$Egg_Count,prob=plogis(10*mos$Detritus*exp
Ricker_likelihood
#1385.847
```

From our analysis of the different models using dbinom we find that the likelihood of the data given the hypothesis is higher under the Ricker model because that gives us the smalles output value.

Question 4

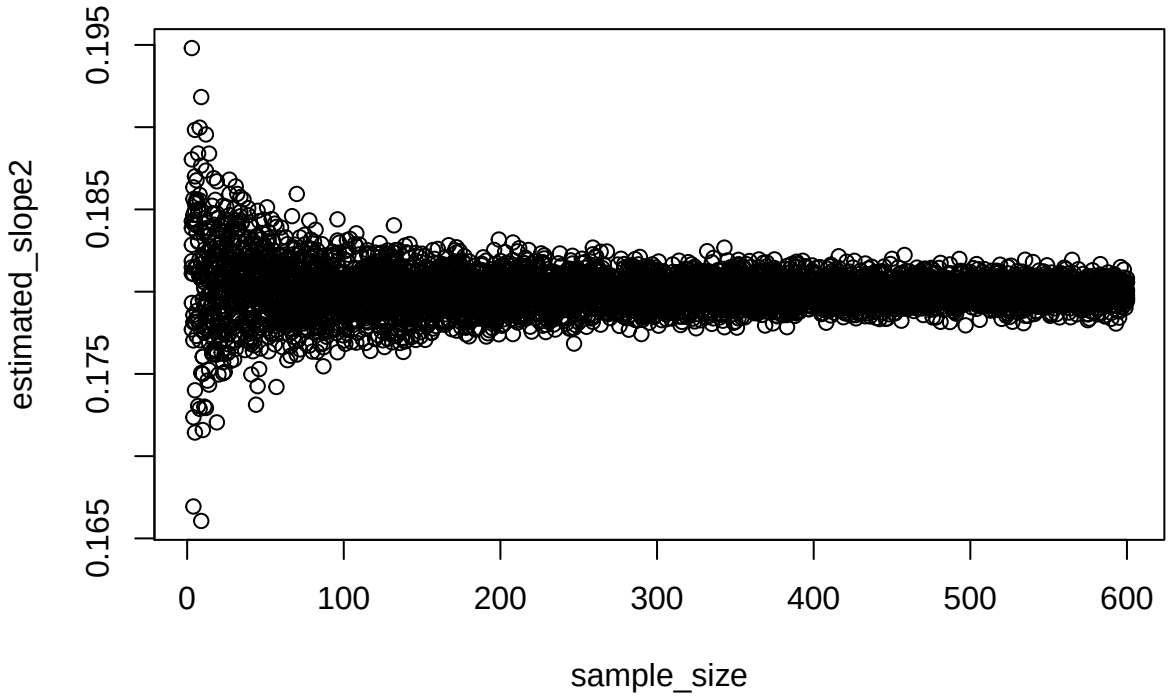
To determine how many moths we need to test in order to get an accurate measure of the effect of tail length on moth escape success (from bat predation), we will run a power analysis. Here, we draw information from a prior study that included 593 interactions between a bat and a saturniid moth of varying hindwing lengths. Each bat was allowed 4 saturniid interaction attempts per night. The slope of the line and approximate intercept are taken from the original study.

```r
#y=-0.71+0.18x (moth escape success by moth tail length)
#power analysis

sample_size = rep(seq(from=3, to=600), times=10)
estimated_slope=rep(NA,times=length(sample_size))
for(i in 1:length(sample_size)) {
  y=rbinom(n=sample_size[i],prob=plogis(-0.71+.18*seq(from=2, to=14, length=sample_size[i])),size=4)
  escape=cbind(y,4-y)
  m1<-glm(escape~seq(from=2, to=14, length=sample_size[i]),family="binomial")
  estimated_slope[i]=coef(m1)[2]
}

plot(estimated_slope~sample_size)
abline(h=1,col="red",lwd=2)
```




```r
confint(m1)
coef(m1)
#when the sample size was 3:150 my coef(m1) produced an intercept of -0.92 and slope of 0.20
#when the sample size was 3:600 my coef(m1) produced intercept of -0.69 and slope of 0.17


#linear regression
sample_size = rep(seq(from=3, to=600), times=10)
estimated_slope2=rep(NA,times=length(sample_size))
for(i in 1:length(sample_size)) {
  y=rnorm(n=sample_size[i],mean=(-0.71+.18*seq(from=2, to=14, length=sample_size[i])),sd=0.05)
```

```
  m2<-lm(y~seq(from=2, to=14, length=sample_size[i]))
  estimated_slope2[i]=coef(m2)[2]
}

plot(estimated_slope2~sample_size)
abline(h=1, col="red", lwd=2)
```




```
#somethign is wrong with the linear regression, clearly...
confint(m2)
#do not overlap zero
coef(m2)
#intercept: -0.71, slope:0.18

summary(m1)
#p-value:p<2e-16

summary(m2)
#p-value: p<2.2e-16
```

Interestingly, I am detecting no difference in accuracy between my binomial glm and linear regression models, in terms of p-value and difference from the known slope and intercept. However, the glm seems to indicate that I need a lower sample size than indicated by the linear regression (about 200 interactions vs 70 or so).

Discrete response variables are grouped into categories that they may or may not fit perfectly within. Continuous response variables are exact measurements of reality, leading to greater statistical and predictive power.