HW2_glm_binomial

Anna Roser

January 30, 2019

Question 1: Seedling Survival and Height

Height is a stronger predictor of seedling survival because the effect size of 1cm difference in height is greater than 1(unit) light access.

```
seeds<-read.csv("SEEDLING_SURVIVAL.csv")
head(seeds)
```

```
##   survival HEIGHT LIGHT
## 1        1   47.0  2.40
## 2        1   70.2 14.83
## 3        1   16.3  9.15
## 4        1   23.5  8.62
## 5        1   23.0  4.26
## 6        1   21.0  3.38
```

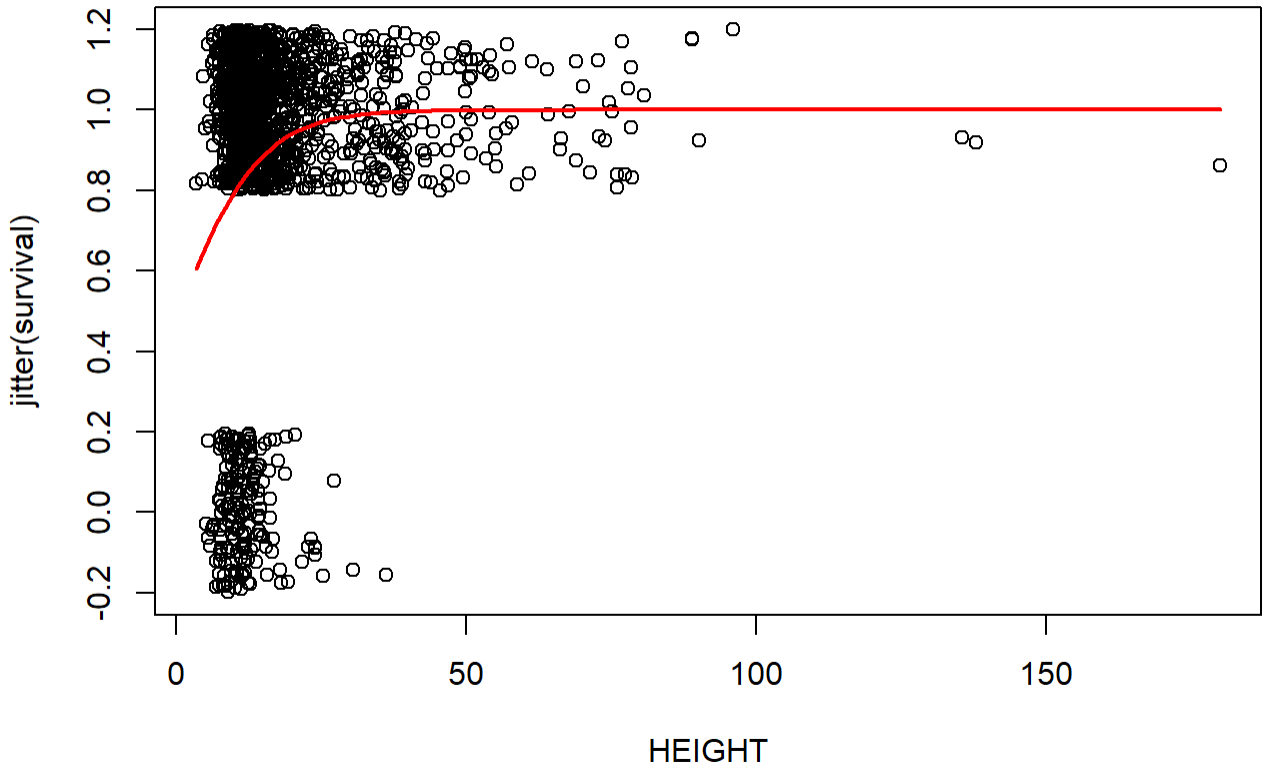
```
str(seeds)
```

```
## 'data.frame':    1435 obs. of  3 variables:
##  $ survival: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HEIGHT  : num  47 70.2 16.3 23.5 23 21 30 17.5 76 57.5 ...
##  $ LIGHT   : num  2.4 14.83 9.15 8.62 4.26 ...
```

A) Plot and Curve

A plot of raw data with the best-fit regression line (use curve) overlaid on the plot

```
seedmod1<-glm(survival~HEIGHT, data= seeds, family= "binomial")
plot(jitter(survival)~HEIGHT, data = seeds)
curve(plogis(-0.0627+0.141*x), add = T, col = "red", lwd = 2)
```



B) Point estimates for slope and intercept

As the height of the seedling increases by 1cm, the likelihood of survival increases by 4%. When the height of the seedling is zero, the probability of survival is 48%.

```
coef(seedmod1)
```

```
## (Intercept)      HEIGHT
## -0.06271111  0.14071141
```

```
coef(seedmod1)[2]/4     #need to divide slope by 4 to find steepest part of the curve--> max effect
```

```
##     HEIGHT
## 0.03517785
```

```
plogis(coef(seedmod1)[1])   #apply plogis to the intercept
```

```
## (Intercept)
##   0.4843274
```

C) Confidence intervals for slope and intercept

The 95% CI for height doesn’t cross zero, therefore there’s a significant positive relationship of hieght on seedling survival. When height of seedlings is zero, the probability of germination ranges from 35-60% (not very tight estimation)

```
confint(seedmod1)           #CI in data units--> interpret slope/predictor variable
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept) -0.5791061 0.4268167
## HEIGHT       0.1038803 0.1815477
```

```
plogis(confint(seedmod1))    #CI in proportion--> interpret intercept/response variable: at baseline of hgt
 at zero, there's a 35-60% chance of survival
```

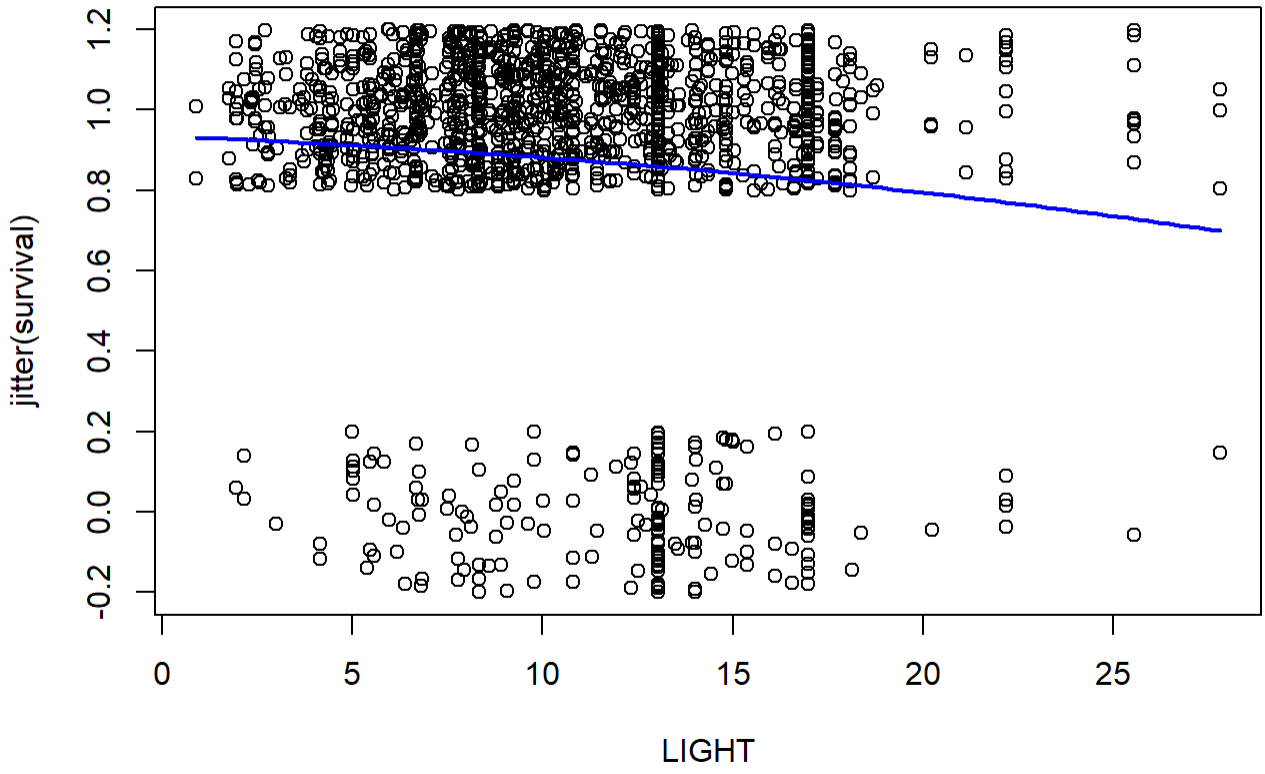
```
## Waiting for profiling to be done...
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.3591383 0.6051133
## HEIGHT      0.5259467 0.5452627
```

Question 1: Seedling Survival and Light

A) Plot and Curve

```
seedmod2<-glm(survival~LIGHT, data= seeds, family= "binomial")
plot(jitter(survival)~LIGHT, data = seeds)
curve(plogis(2.662+-0.0655*x), add = T, col = "blue", lwd = 2)
```



B) Point estimates for slope and intercept

When there's no light, the proportion of seedling survival is 93%. As the measurement of light reaching the seedlings increases by one, the probability of seedling survival drops by 2%.

```
coef(seedmod2)
```

```
## (Intercept)       LIGHT
##  2.66194692 -0.06552684
```

```
coef(seedmod2)[2]/4     #need to divide slope by 4 to find steepest part of the curve--> max effect
```

```
##       LIGHT
## -0.01638171
```

```
plogis(coef(seedmod2)[1])  #apply plogis to the intercept--> interpret seed survival as proportion
```

```
## (Intercept)
##   0.9347435
```

C) Confidence intervals for slope and intercept

The 95% CI of light doesn't cross zero, therefore there's a significant, negative effect on light on seedling survival success. When there's no light in a plot, there's a 90-96% chance of seedling survival (narrower estimate)

```
confint(seedmod2)          #CI in data units--> interpret slope/predictor variable
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept)  2.25136434  3.0876309
## LIGHT       -0.09841747 -0.0325795
```

```
plogis(confint(seedmod2))  #CI in proportion--> interpret intercept/response variable: at baseline of light
 at zero, there's a 90-96% chance of survival
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.9047682 0.9563796
## LIGHT       0.4754155 0.4918558
```

Question 2: Recruits and Light

Based on the evidence in A-C, I have concluded that light has a significant positive effect on seedling germination.

These data represent results from a seed addition experiment where 5, 15, or 45 seeds □□□germinating seedlings that emerged from the added seeds

```
seed<-read.csv("seeds.csv")
head(seed)
```

```
##   Site   Pile    DBH seedlings seeds recruits grass light
## 1   m1  m1.15 21.590         0    15        2     1  9.35
## 2   m1  m1.45  0.000         0    45        2     0 17.00
## 3   m1   m1.5 46.990         0     5        1     0  6.68
## 4  m10 m10.15  0.000         0    15        0     0  6.72
## 5  m10 m10.45 27.686         0    45        0     0  4.91
## 6  m10  m10.5  0.000         0     5        2     0  3.07
```

```
str(seed)
```

```
## 'data.frame':    281 obs. of  8 variables:
##  $ Site     : Factor w/ 94 levels "m1","m10","m11",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ Pile     : Factor w/ 281 levels "m1.15","m1.45",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ DBH      : num  21.6 0 47 0 27.7 ...
##  $ seedlings: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ seeds    : int  15 45 5 15 45 5 15 45 5 15 ...
##  $ recruits : int  2 2 1 0 0 2 6 0 1 1 ...
##  $ grass    : int  1 0 0 0 0 0 0 1 1 0 ...
##  $ light    : num  9.35 17 6.68 6.72 4.91 ...
```

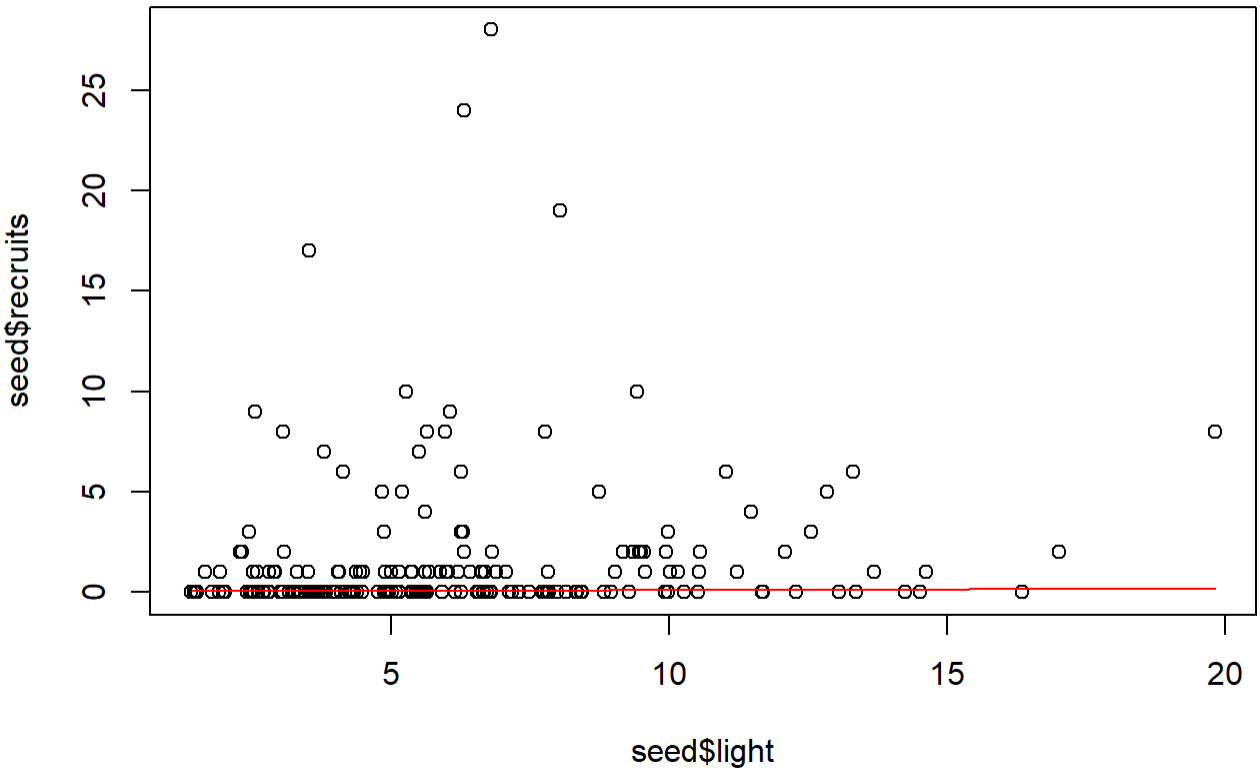
```
response<-cbind(seed$recruits, seed$seeds-seed$recruits)  #make proportional variable
head(response)
```

```
##      [,1] [,2]
## [1,]    2   13
## [2,]    2   43
## [3,]    1    4
## [4,]    0   15
## [5,]    0   45
## [6,]    2    3
```

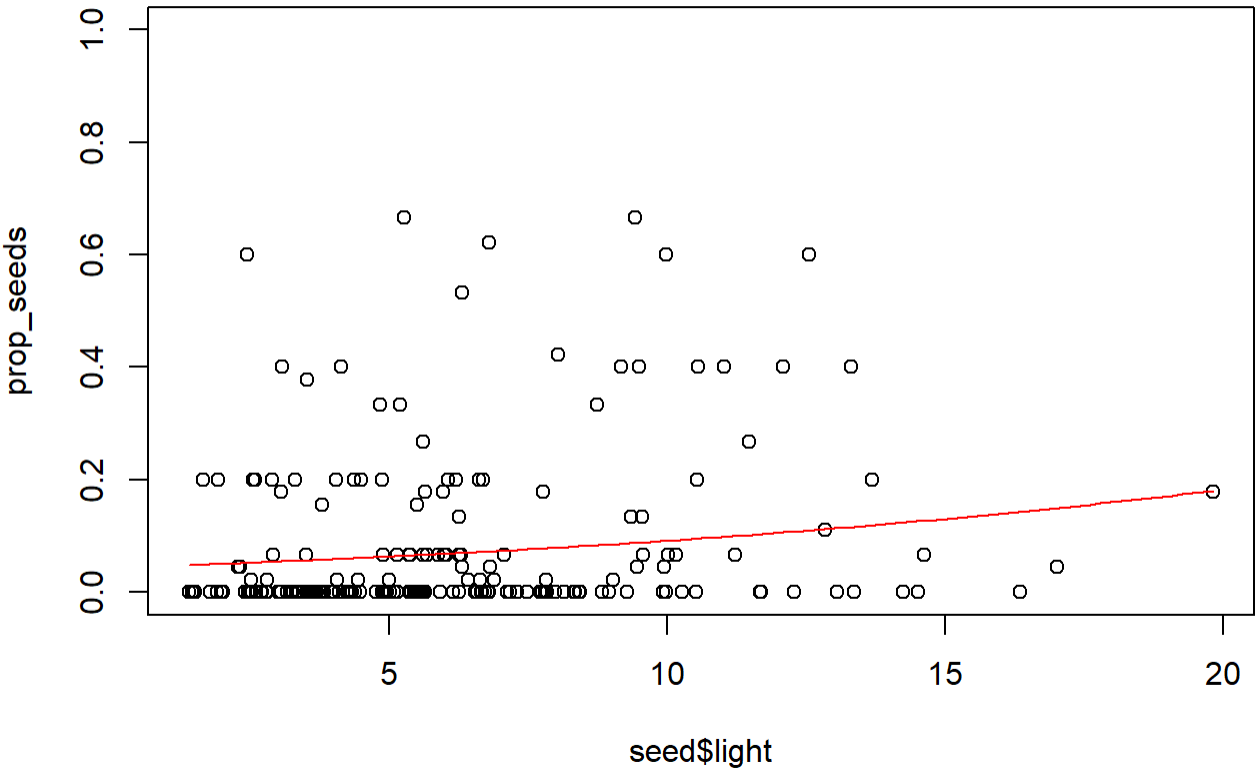
```
seedmod3<-glm(response~seed$light, family = "binomial")   #make the model
```

A) Plot and curve

```
plot(seed$recruits~seed$light)   #raw data w/ curve
curve(plogis(-3.093+0.0798*x), add = T, col = "red", lwd = 1.4)
```




```
prop_seeds<-seed$recruits/seed$seeds   #data as proportion w/ curve
plot(prop_seeds~seed$light)
curve(plogis(-3.093+0.0798*x), add = T, col = "red", lwd = 1.4)
```



B) Point Estimates

As the amount of light increases by one unit in the plot, the likelihood of recruit survival increases by 2%. When there's no light in the plot, the likelihood of recruit survival is 4%.

```
coef(seedmod3)
```

```
## (Intercept)  seed$light
##  -3.0936296   0.0798368
```

```
coef(seedmod3)[2]/4     #need to divide slope by 4 to find steepest part of the curve--> max effect
```

```
## seed$light
##  0.0199592
```

```
plogis(coef(seedmod3)[1])       #apply plogis to the intercept--> interpret recruit success as proportion
```

```
## (Intercept)
##  0.04337079
```

C) Confidence Intervals

The 95% CI for light as a predictor of recruit success does not cross zero, therefore light has a significant positive effect on recruit success. When there's no light in a plot, recruits have a 3-5% chance of success.

```
confint(seedmod3)     #CI in data values--real world, look at predictor variable
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept) -3.32799150 -2.8631191
## seed$light   0.04990323  0.1088193
```

```
plogis(confint(seedmod3))   #CI in proportion, look at response variable as percentage range
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %     97.5 %
## (Intercept) 0.0346233 0.05400712
## seed$light  0.5124732 0.52717801
```

Question 3: Mosquitos

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

```
##   Emergent_adults Egg_Count Detritus
## 1               2         3     0.00
## 2               0         2     0.01
## 3               2         3     0.01
## 4               4         6     0.02
## 5               1         4     0.02
## 6               5         7     0.03
```

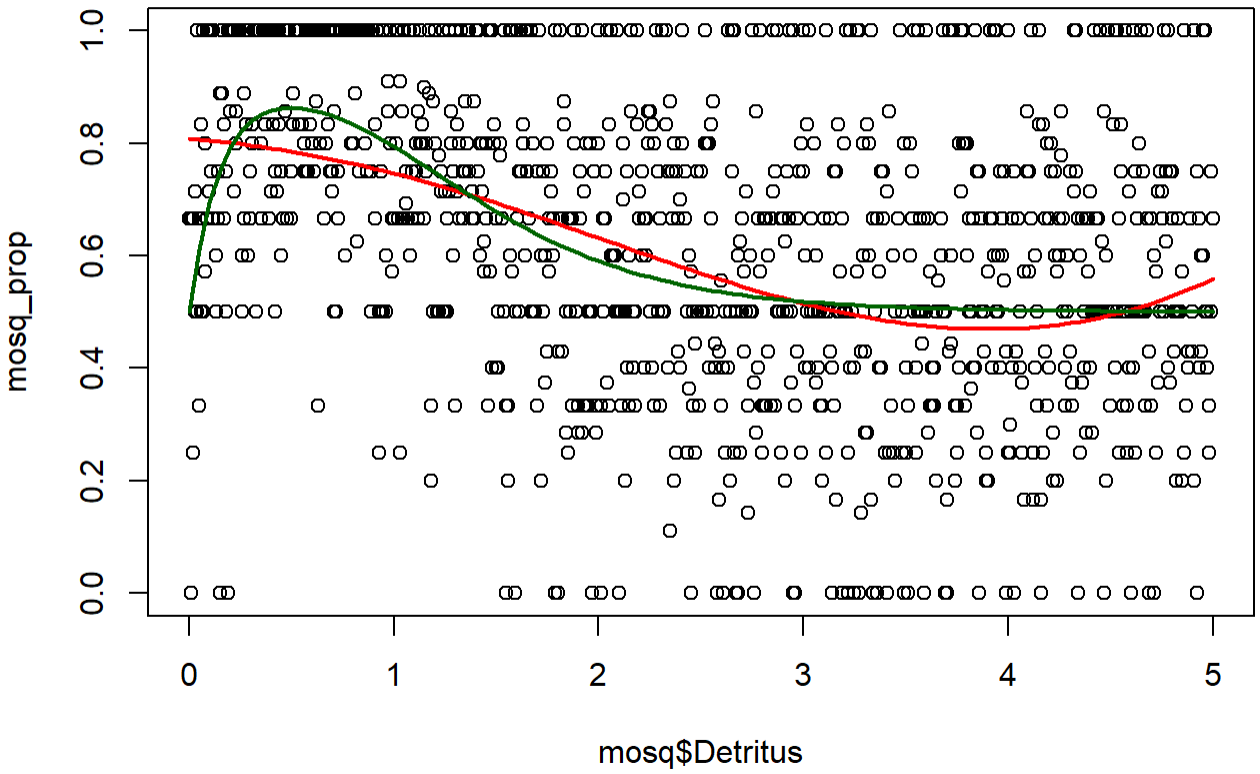
```
str(mosq)
```

```
## 'data.frame':    1000 obs. of  3 variables:
##  $ Emergent_adults: int  2 0 2 4 1 5 2 1 1 4 ...
##  $ Egg_Count      : int  3 2 3 6 4 7 4 1 2 6 ...
##  $ Detritus       : num  0 0.01 0.01 0.02 0.02 0.03 0.03 0.04 0.04 0.05 ...
```

A and B) Plot and Curves

Plot the data. (Again, note that plotting the proportion of eggs is necessary, since the number of eggs varies)

```
mosq_prop<-mosq$Emergent_adults/mosq$Egg_Count
plot(mosq_prop~mosq$Detritus)
curve(plogis(1.44-0.19*x-0.21*(x^2)+0.04*x^3), add = T, col = "red", lwd = 2)   #polynomial
curve(plogis(10*x*exp(-2*x)), add = T, col = "darkgreen", lwd = 2)              #Ricker
```



C) Polynomial vs. Ricker

The polynomial function shows a relationship where the success of mosquito hatching is highest at the lowest levels of detritus and the probability of hatching success drops below 50% as detritus increases. The Ricker model assumes the probability of success of mosquito hatching never drops below 50%, but also that hatching success attenuates as detritus increases.

D) dbinom comparison of MLE

you have to use sum because we're assuming that the samples are independent from each level of detritus

```
#binom function from book
binomMosq1 = function(p, k, N) {
  -sum(dbinom(k, prob = p, size = N, log = TRUE))
}


-sum(dbinom(x = mosq$Emergent_adults,      #success = x
            size = mosq$Egg_Count,         #size = trials
            prob = plogis(1.44-0.19*mosq$Detritus-0.21*(mosq$Detritus^2)+0.04*mosq$Detritus^3),  #prob = pr
edictor variable
            log = T))   #include log = T to switch from multiplying 1000 trials of prob to adding them, eas
ier for computing
```

```
## [1] 1415.63
```

```
-sum(dbinom(x = mosq$Emergent_adults,    #ricker
            size = mosq$Egg_Count,
            prob = plogis(10*mosq$Detritus*exp(-2*mosq$Detritus)),
            log = T))
```

```
## [1] 1385.847
```

E) According to dbinom, the likelihood of the data is higher for which model?

The ricker model is better for the mosquito data because the estimate of maximum likelihood is higher than the polynomial model (or the ricker negative log likelihood is lower than the polynomial).

Question 4: Power Analysis

Power analysis asks how much data is required to detect an effect. In this question, you will conduct two power analyses using stochastic simulation. The first power analysis will be for a linear regression (assume normally distributed data) and the second power analysis will be for binary data (assume logit-link function).

1. Decide true values

If we sample the same area of interest, where we've recored plant species evenness indicies in-situ, as we survey the same area (trials) at coarser spatial resolutions, we lose the ability to predict species richness from pixel heterogeneity.

Note to self: I've created these "true" values to say that as your sensor distance from the ground increases, the amount of species which you're able to detect decreases for by 6 per unit increase AGL. So then, I created a for loop to generate data to test the different amounts of trials/surveys to capture this relationship.

```
slope = -6      #as pixels become more coarse, ability to detect spp decreases by 6
intercept = 97   #max number spp in the plot if all can be detected
sigma = 6   #amount of variance in my sensor
```

2. Create Predictor Variable

?Why do we create a predictor variable if we just use seq in for loop????

```
resolution<-seq(from = 1, to = 100, length = 100)
```

3. Create a vector of sample size

```
sample_size<-c(3:100)
```

4. Create power vector normal dist

Create an empty vector to fill in with results from the power analysis. Note that the number of elements in the vector needs to be equal to the length of the sample size vector

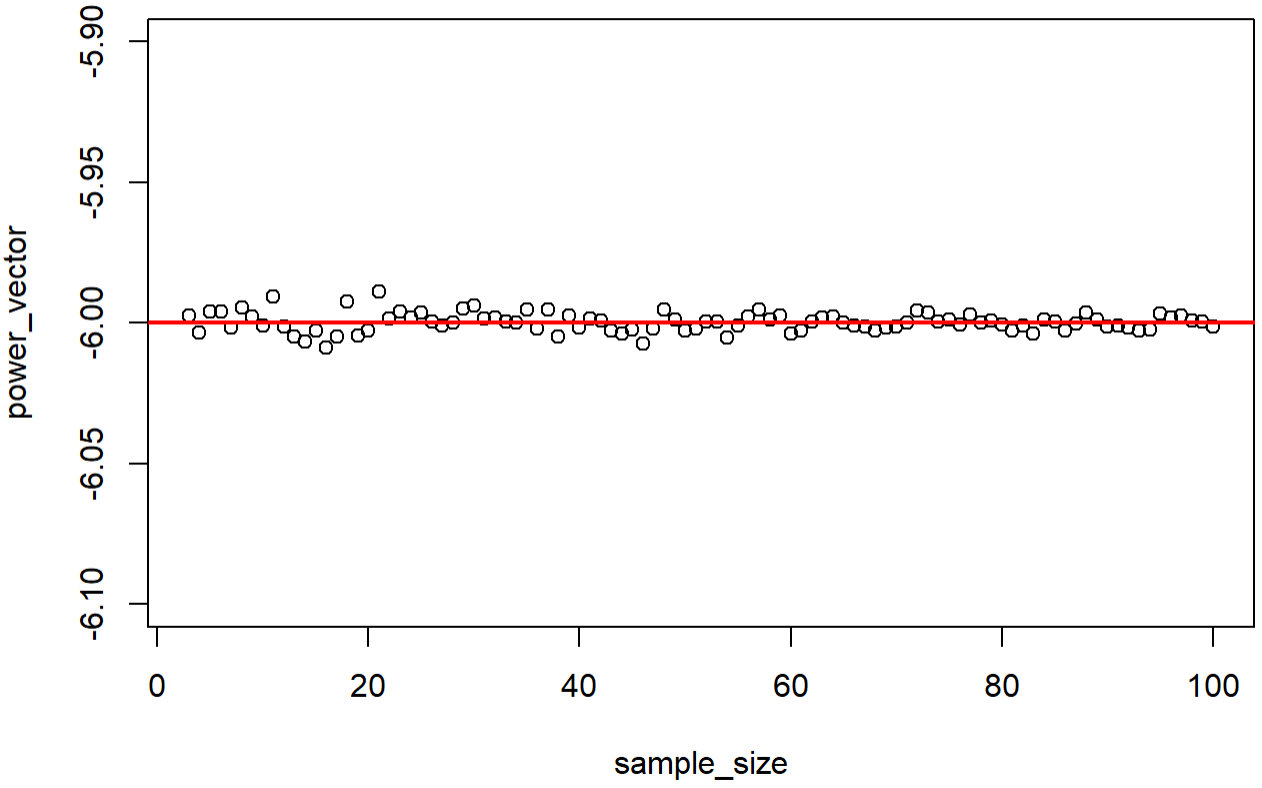
```
power_vector<-rep(NA, times = length(sample_size))
```

5. For loop: simulated analyses

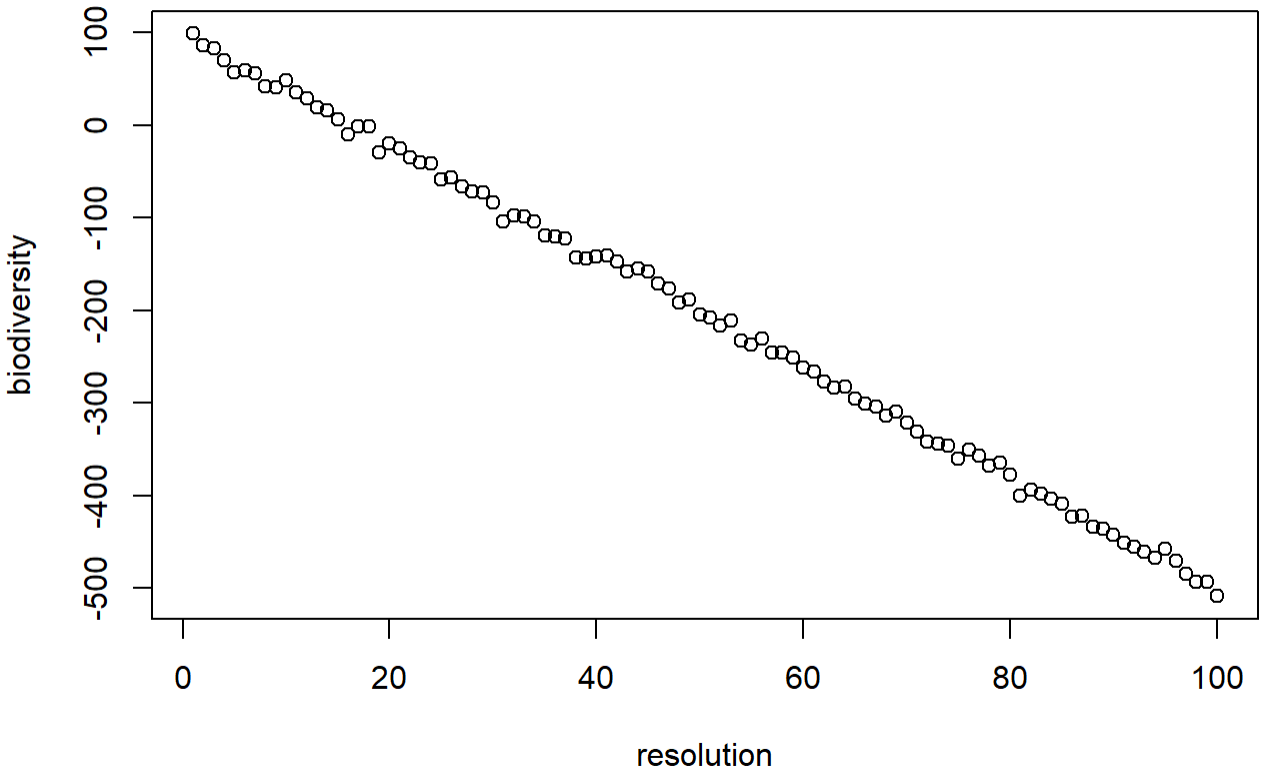
Write a for loop that fills in the empty vector with results from simulated analyses ### Linear model ###*?should I be using glm instead of lm in this loop?*

```
for(i in 1:length(sample_size)) {
  power_temp<-rnorm(n = sample_size[i], mean = intercept + slope*seq(from = 1, to = 1000, length = sample_s
ize[i]), sd = sigma)    #create temp vector with all rnorms
  res_mod<-lm(power_temp~seq(from = 1, to = 1000, length = sample_size[i]))  #run lm model for generated da
ta
  power_vector[i]<-coef(res_mod)[2]    #fill empty power vector with coef slope from lm model runs
}

#plot generated estimates of slope from trials as a function of the number of surverys completed
plot(power_vector~sample_size, ylim = c(-6.1,-5.9))
abline(h=-6, col= "red", lwd= 2)  #true population slope
```




```
####comparing generated biodiversity data to predictor variable, ground resolution####
biodiversity<-rnorm(n = 100, mean = intercept + slope*resolution, sd = sigma)
plot(biodiversity~resolution)
```



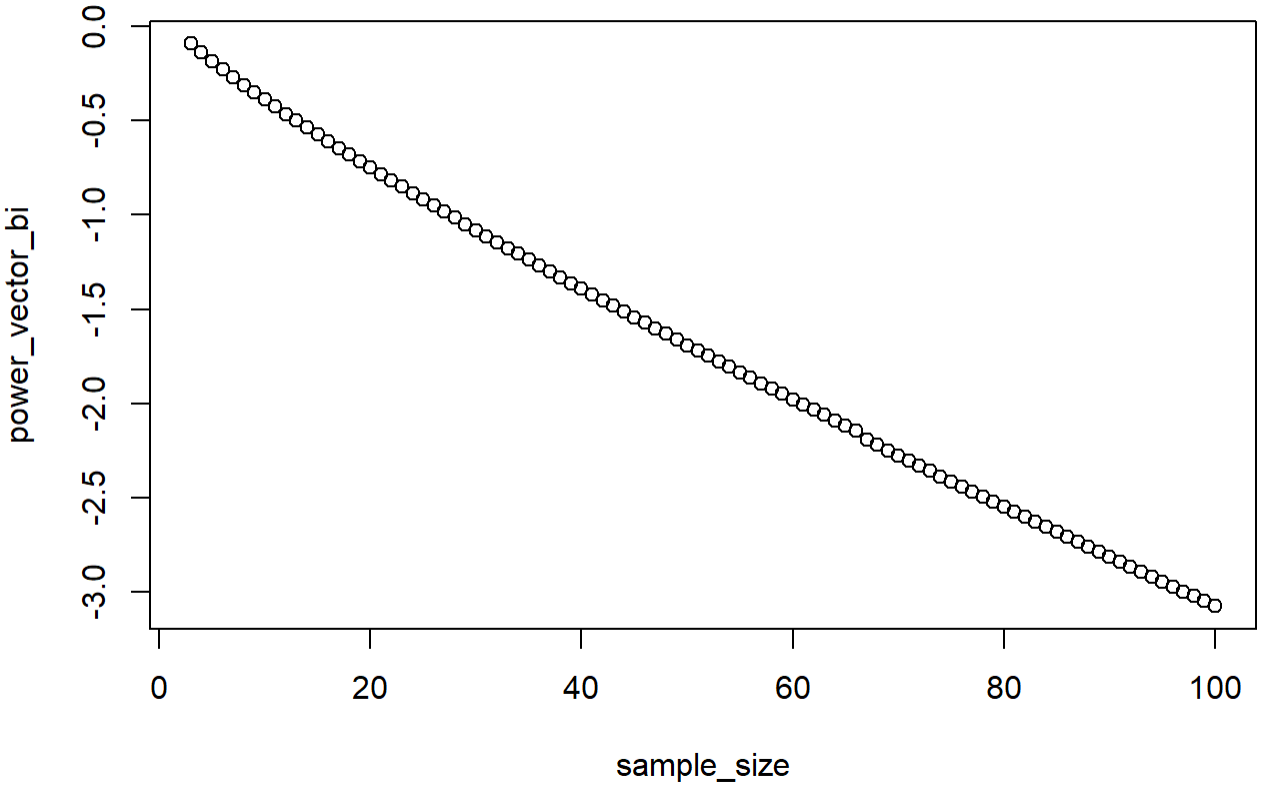
Binomial Model

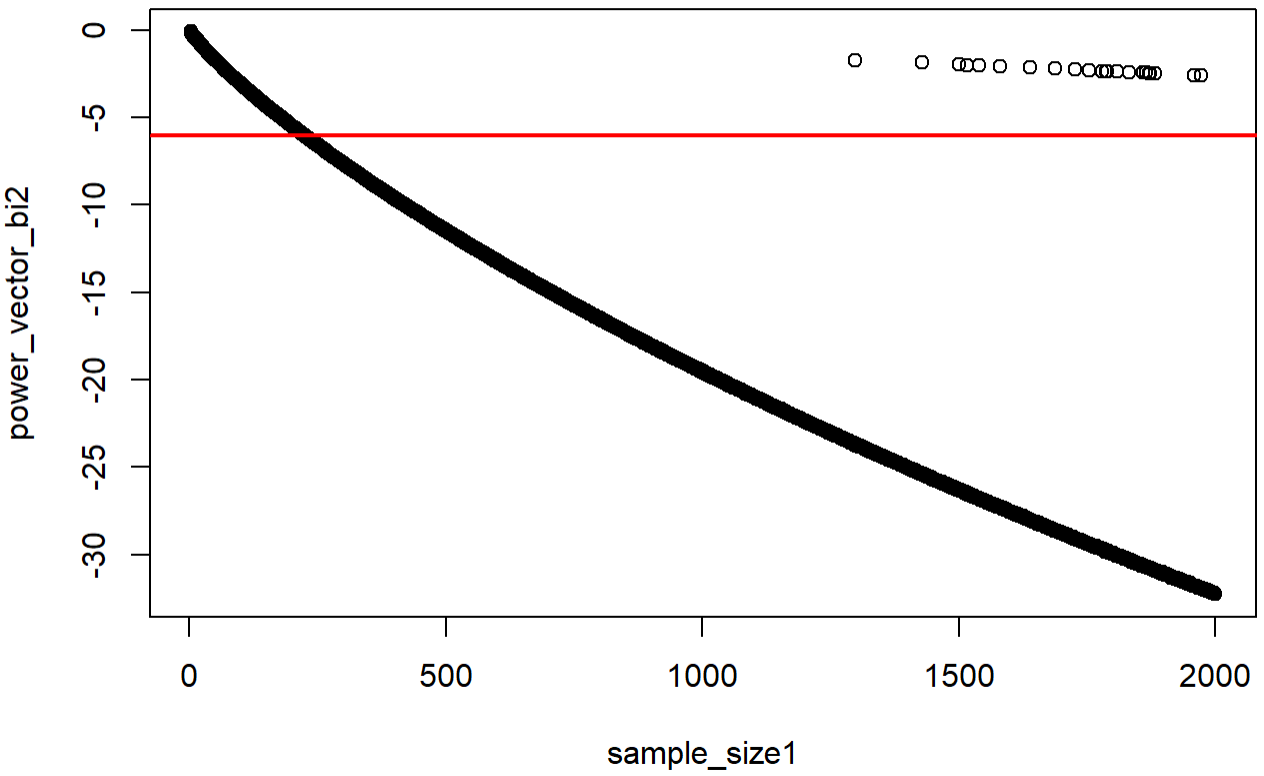
?Question–how to determine correct size for data?

```
#binomial power analysis with 100 trials
power_vector_bi<-rep(NA, times = length(sample_size))

for(i in 1:length(sample_size)) {
  power_temp<-rbinom(n = sample_size[i], size = 1, prob = plogis(intercept + slope*seq(from = 1, to = 1000,
 length = sample_size[i])))    #create temp vector with all rbinorms
  res_mod<-glm(power_temp~seq(from = 1, to = 1000, length = sample_size[i]), family = "binomial")  #run lm
 model for generated data
  power_vector_bi[i]<-coef(res_mod)[2]    #fill empty power vector with coef slope from lm model runs
}
```

```
plot(power_vector_bi~sample_size)   #100 trials is not enough
abline(h=-6, col= "red", lwd= 2)
```


power_vector_bi
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
0
20
40
60
80
100
sample_size


power_vector_bi2
0
-5
-10
-15
-20
-25
-30
0
500
1000
1500
2000
sample_size1

A) Number of samples needed?

How many samples do you need to accurately estimate the slope parameter in a binomial vs. linear regression? (remember: you know the true value of the slope parameter) Use MSE to calculate the accuracy and precision of your estimate vs. the real value:(slope-meanSlope)^2 For a very good parameter estimate of slope, I need to take 20 surveys based on the linear regression. For binomial, I need to take more than 200 surveys–realistically not feasible.

```
linear_MSE<-mean((-6-power_vector)^2)
linear_MSE
```

```
## [1] 1.136621e-05
```

```
binorm_MSE<-mean((-6-power_vector_bi)^2)
binorm_MSE
```

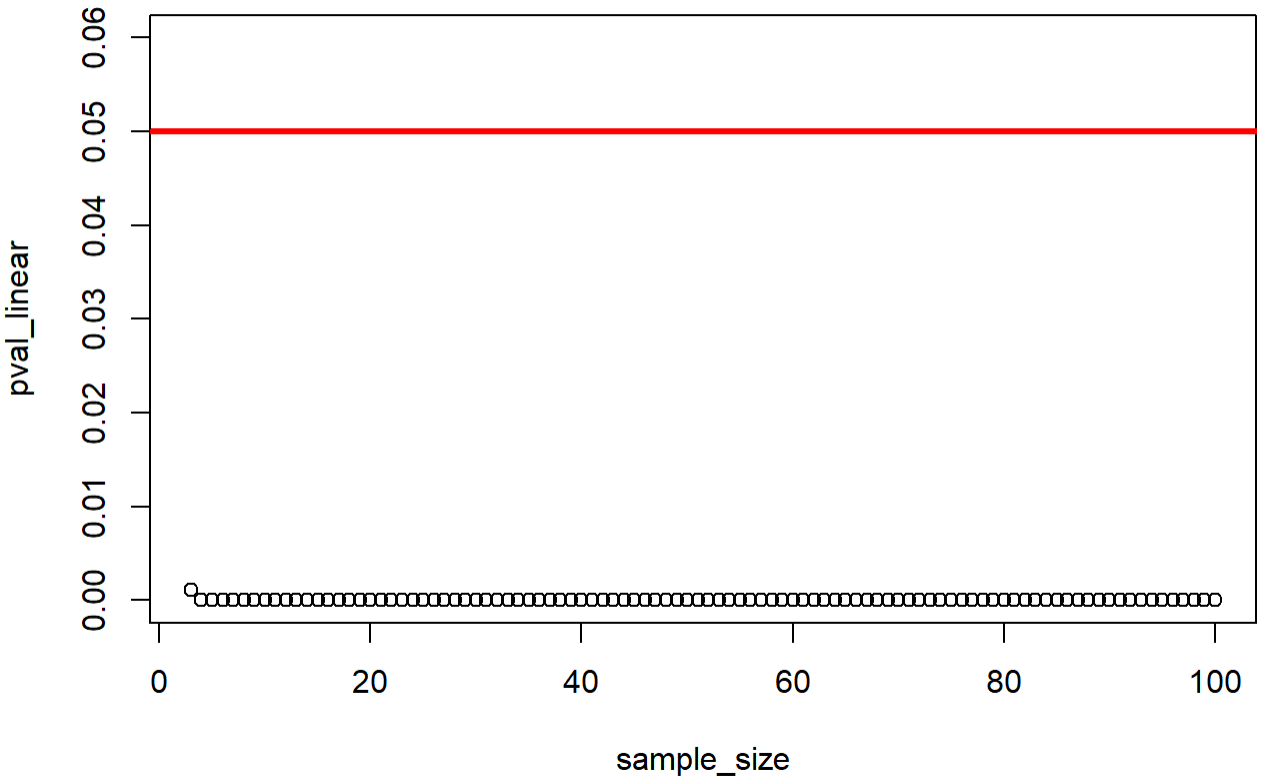
```
## [1] 19.26928
```

B) Evaluate p vals

To be safe I could do 3 surveys to ensure a p val of >0.05 with the linear regression. I would have to taken A LOT of surveys to get a significant p val with the binomial regression.

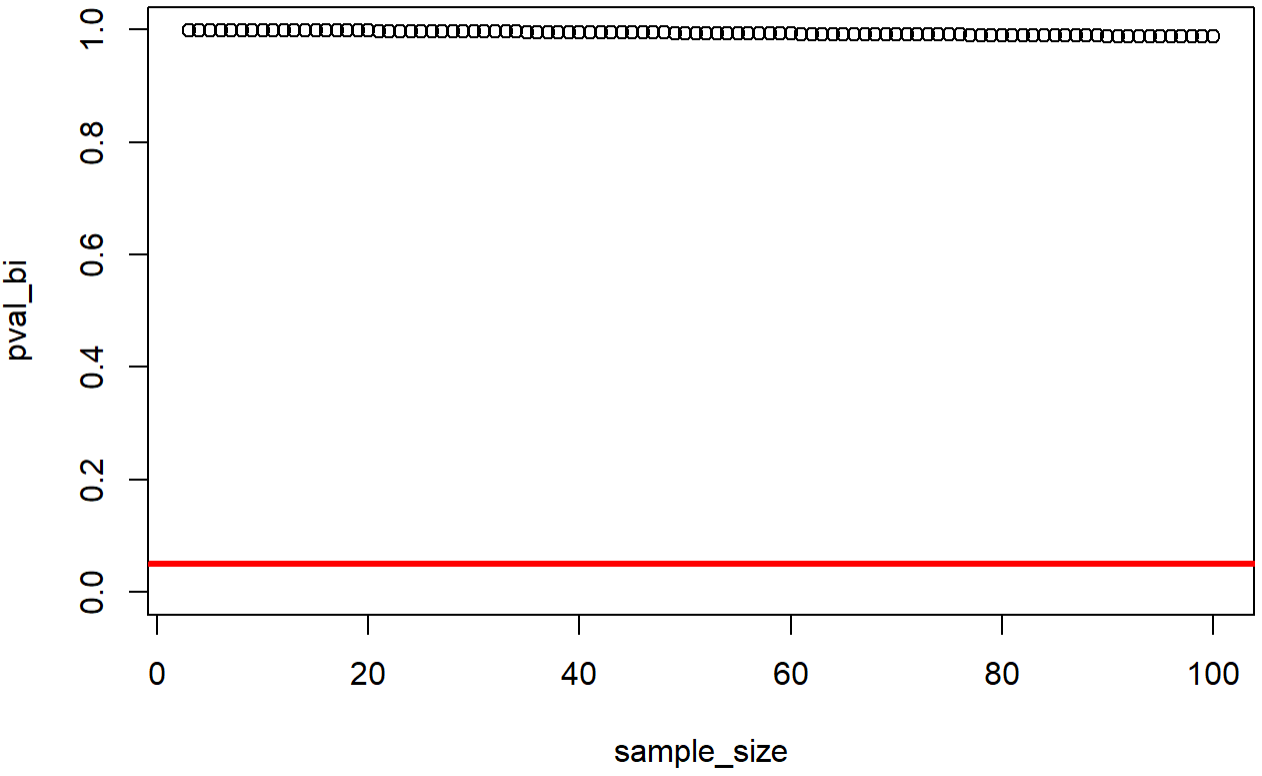
```
####Check p val####
pval_linear<-rep(NA, times = length(sample_size))
for(i in 1:length(sample_size)) {
  power_temp<-rnorm(n = sample_size[i], mean = intercept + slope*seq(from = 1, to = 1000, length = sample_s
ize[i]), sd = sigma)    #create temp vector with all rnorms
  res_mod<-lm(power_temp~seq(from = 1, to = 1000, length = sample_size[i]))  #run lm model for generated da
ta
  pval_linear[i]<-summary(res_mod)$coefficients[2,4]    #fill empty power vector with coef slope from lm mo
del runs
}

plot(pval_linear~sample_size, ylim = c(0,0.06))
abline(h=0.05, col = "red", lwd = 3)   #lm is very good
```




```
####Check p val####
pval_bi<-rep(NA, times = length(sample_size))
for(i in 1:length(sample_size)) {
  power_temp<-rbinom(n = sample_size[i], size = 1, prob = plogis(intercept + slope*seq(from = 1, to = 1000,
 length = sample_size[i])))    #create temp vector with all rbinorms
  res_mod<-glm(power_temp~seq(from = 1, to = 1000, length = sample_size[i]), family = "binomial")  #run lm
 model for generated data
  pval_bi[i]<-summary(res_mod)$coefficients[2,4]    #fill empty power vector with coef slope from lm model
 runs
}

plot(pval_bi~sample_size, ylim = c(0,1))
abline(h=0.05, col = "red", lwd = 3)   #as sample size increases, p val gets marginally better
```



C) Comparison of Statistical power: continuous vs discrete

Statistical power is generally higher for continuous than discrete response variables because continuous response variables include more variability of the data. I'm not sure–can we talk about this question in class?