Homework 2

Clara Buchholtz

1/31/2019

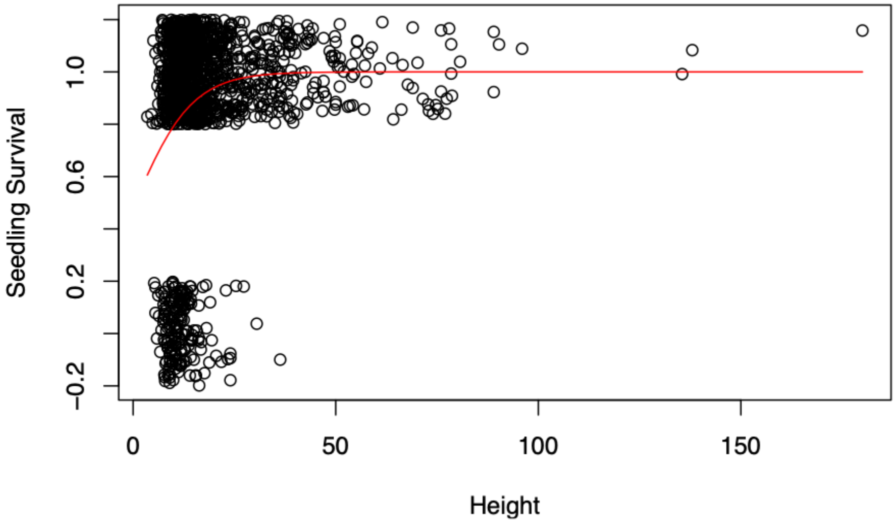
Question 1

Dataset: "SEEDLING_SURVIVAL.csv"

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

A) Effect of height on seedling survival

```
# a) Plot height as predictor for survival
plot(jitter(seedsurv$survival)~seedsurv$HEIGHT,xlab="Height", ylab="Seedling Survival")
# add line with parameters estimated from part b:
curve(plogis( -0.06271111 + 0.14071141*x), col="red", add=T)
```



```
# b) Estimate best fit ("most likely given the data") parameters from the glm
mHeight <- glm(seedsurv$survival~seedsurv$HEIGHT, family=binomial)
coef(mHeight)
```

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

```
plogis( -0.06271111)
```

```
## [1] 0.4843274
```

```
#Interpret the parameters
#Baseline seedling survival- convert to probability scale:
```

The seedlings have a baseline survival rate of roughly 48%.

```
#Effect of height
0.14071141/4
```

```
## [1] 0.03517785
```

There is a maximum height effect of a 3.5% increase in seedling survival

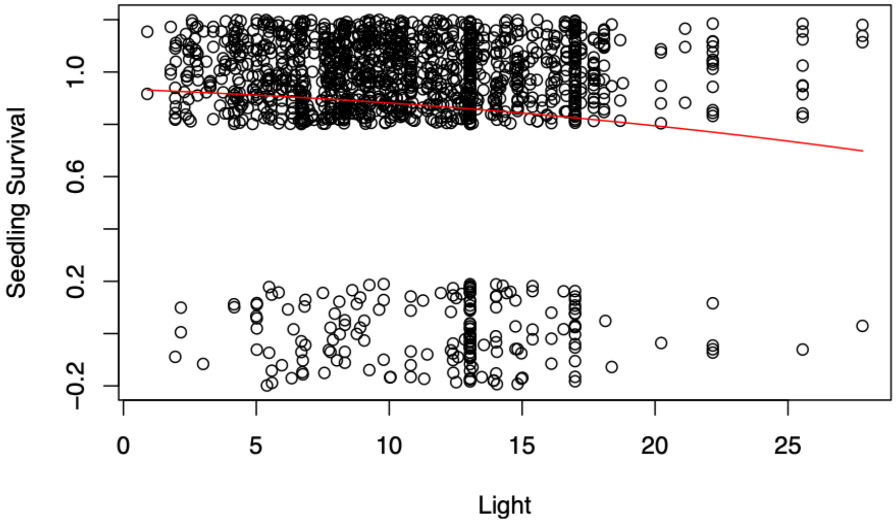
```
# c) Confidence intervals
confHeight <- confint(mHeight)
confHeight
```

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

From this output we can see that height has a significant and positive effect on seedling survival. The effect does not appear to be especially large, but the confidence interval does not cross zero.

B) Effect of light on seedling survival

```
# a) Plot light as predictor for survival
plot(jitter(seedsurv$survival)~seedsurv$LIGHT,xlab="Light", ylab="Seedling Survival")
# add line with parameters estimated from part b:
curve(plogis(2.66194692 -0.06552684 *x), col="red", add=T)
```



```
# b) Estimate best fit ("most likely given the data") parameters from the glm
mLight <- glm(seedsurv$survival~seedsurv$LIGHT, family=binomial)
coef(mLight)
```

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

```
#Interpret the parameters
#Baseline seedling survival- convert to probability scale:
plogis(2.66194692)
```

```
## [1] 0.9347435
```

The baseline survival for the seedlings is roughly 93%

```
#Maximum effect of light on seed survival
-0.06552684/4
```

```
## [1] -0.01638171
```

According to the model, light has a maximum effect of a -1.6% decrease in seedling survival

```
# c) Confidence Intervals
confint(mLight)
```

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

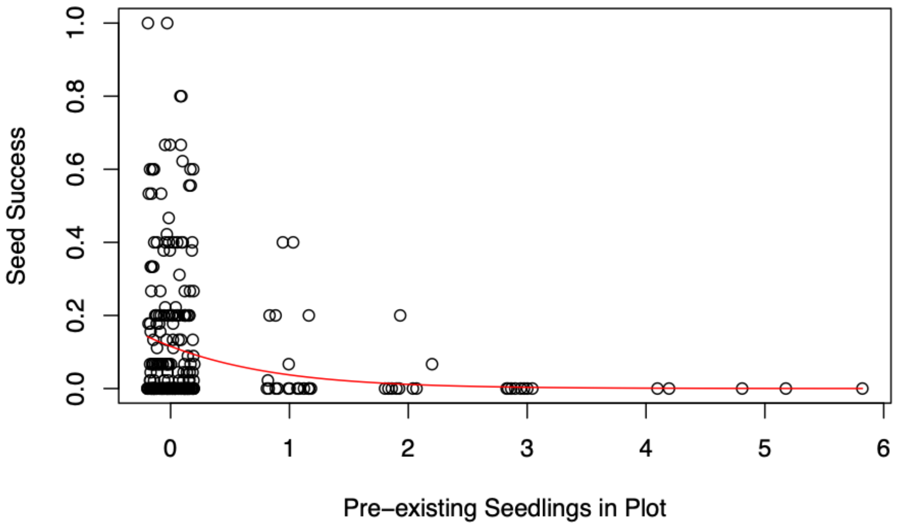
According to the model, there is a small, significant, negative effect of light on seedling survival.

Is height or light a stronger predictor of seedling survival?

Height seems to be a stronger predictor of seedling survival. Both predictors are significant, but the slope values for height are larger.

Question 2: Dataset: "seeds.csv"

```
Seeds <- read.csv("seeds.csv")
# a) a plot of raw data with the best-fit regression line
plot(jitter(Seeds$seedlings),Seeds$recruits/Seeds$seeds,
     xlab='Pre-existing Seedlings in Plot', ylab='Seed Success')
curve(plogis(-2.035570  -1.213717 *x), col='red', add=T)
```



```
# b) Point estimates for slope and intercept parameters,
#including a verbal description
#of the baseline and effect size for these parameters
response <- cbind(Seeds$recruits, Seeds$seeds-Seeds$recruits)
SeedParams <- glm(response~Seeds$seedlings, family=binomial)
coef(SeedParams)
```

```
##     (Intercept) Seeds$seedlings
##       -2.035570       -1.213717
```

```
#Baseline effect:
plogis(-2.035570)
```

```
## [1] 0.1155186
```

The baseline success of planted seeds in these plots when pre-existing seedlings = 0 is 11.55%

```
-1.213717/4
```

```
## [1] -0.3034292
```

The maximum effect that pre-existing seedlings in a plot are predicted to have on the planted seeds is a -30.34% decrease in success, suggesting the possibility that crowding or scarcity of some resource may be at play.

```
#c) Confidence intervals for slope and intercept
confint(SeedParams)
```

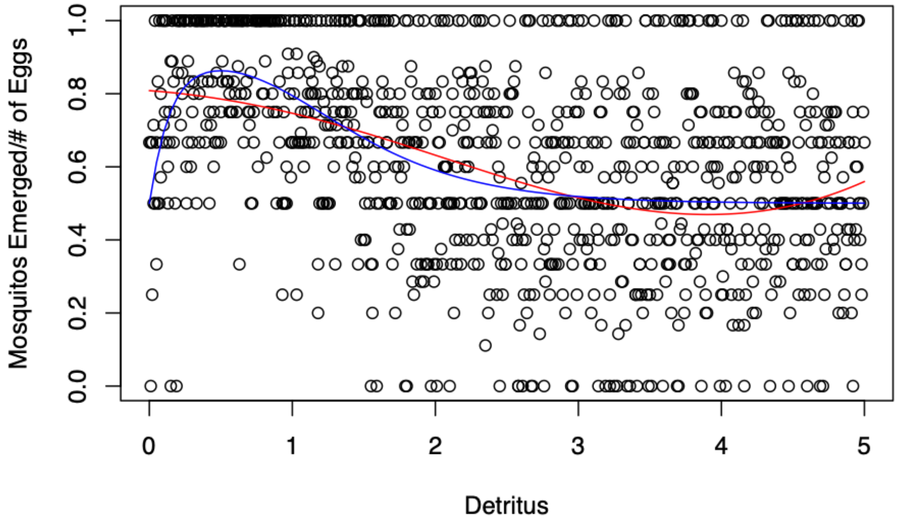
```
##                     2.5 %     97.5 %
## (Intercept)     -2.121808 -1.9511837
## Seeds$seedlings -1.592362 -0.8977661
```

Presence of seedlings already in the plot seems to have a significant (the interval does not cross zero), negative effect on seed success

Question 3 Dataset:“mosquito_data.csv”

```r
Mosquitos <-read.csv("mosquito_data.csv")
#a. Plot the data.
plot((Mosquitos$Emergent_adults/Mosquitos$Egg_Count)~Mosquitos$Detritus,xlim=c(0, 5),
     xlab="Detritus", ylab="Mosquitos Emerged/# of Eggs")
#b. Add curves
#Polynomial
curve(plogis(1.44-0.19*x-(0.21*x^2)+(0.04*x^3)), col='red', add=T)

#Ricker
curve(plogis(10*x*exp(-2*x)), col='blue', add=T)
```



c.

How are the biological implications of the polynomial model different from the Ricker model?

The Ricker model seems to predict a scenario where a having a little bit of detritus is beneficial to mosquito emergence, but after a certain point adding more detritus is actually not beneficial at all, and we go back to baseline mosquito emergence rates no matter how much more detritus we have.

The Polynomial model seems to suggest that increased detritus has a modest negative effect on mosquito emergence up until a point (detritus = ~3.9), when it starts to lose this effect and emergence rates start climbing again.

```r
#d. use dbinom to calculate the likelihood of both models
evalPoly<- -sum(dbinom(x=Mosquitos$Emergent_adults,
                      size=Mosquitos$Egg_Count,
                      prob= plogis(1.44-0.19*Mosquitos$Detritus-
                                  (0.21*Mosquitos$Detritus^2)+
                                  (0.04*Mosquitos$Detritus^3)),
                      log=TRUE))

evalPoly
```

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

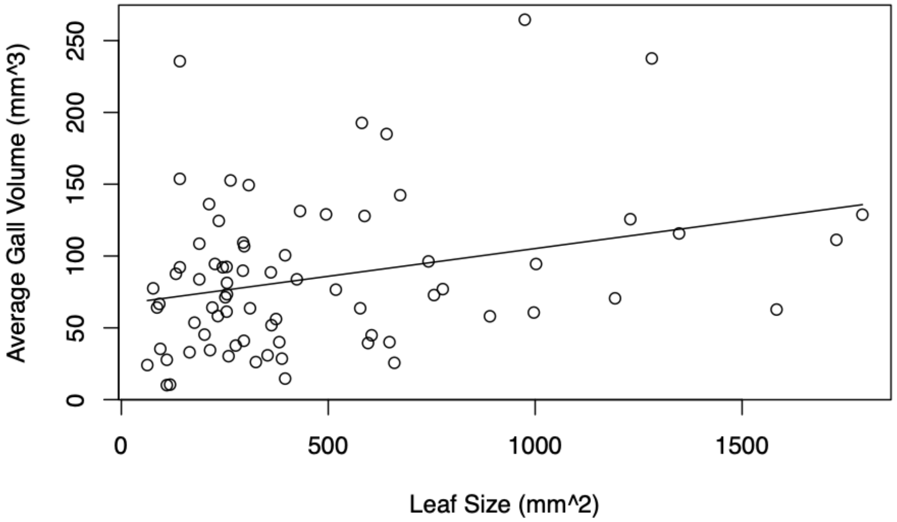
```
evalRicker <- -sum(dbinom(x=Mosquitos$Emergent_adults,
                          size=Mosquitos$Egg_Count,
                          prob= plogis(10*Mosquitos$Detritus*exp(-2*Mosquitos$Detritus)),
                          log=TRUE))
evalRicker
```

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

e. According to dbinom, the likelihood of the data is higher for which model? According to dbinom, the Ricker model has a higher likelihood, because the negative log likelihood is lower.

Question 4: Power analysis

```
#These data are taken from my thesis, and this analysis looks at the
#average gall size on a leaf, with
#leaf size being a predictor. I was hoping this relationship would be either
#very weak or non-significant, and I had collected 70+ leaves for my sample to verify that.
#This power analysis will help me evaluate whether my data can support my hypothesis,
#or if I would need to collect more data to be on solid footing making such a claim.
lfsize<-read.csv("LeafSize_12_21.csv")
plot(lfsize$AvgGallVol~lfsize$LeafArea,
     xlab="Leaf Size (mm^2)", ylab="Average Gall Volume (mm^3)")
curve(66.57264976+0.03863467*x, add=T)
```

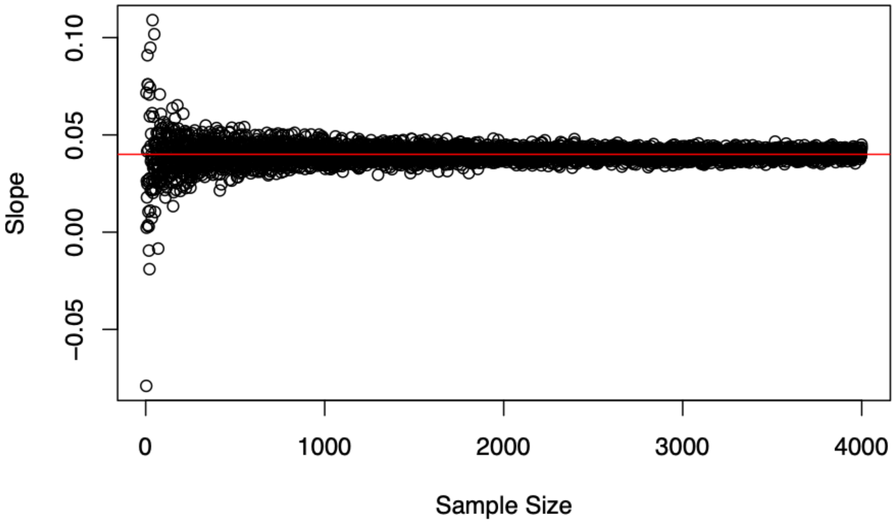


```
#Generating slope, intercept, and sd:
#mod_lf<-lm(lfsize$AvgGallVol~lfsize$LeafArea)
#coef(mod_lf)
# (Intercept) lfsize$LeafArea
# 66.57264976       0.03863467
#confint(mod_lf)
# confint(mod_lf)
# 2.5 %       97.5 %
#   (Intercept)     48.271916018 84.87338351
# lfsize$LeafArea  0.008829044  0.06844029
#sigma(mod_lf)
# [1] 50.8726
```

```
#NORMALLY DISTRIBUTED DATA
slope=0.04
intercept=48
standev=50

#Create a range of sample sizes to test:
sample_sizeL = c(3:4000)
#Create an empty vector to fill in with results from the power analysis.
power_vector<-rep(NA,times=length(sample_sizeL))

for(j in 1:length(sample_sizeL)){
  y=rnorm(n=sample_sizeL[j],
          mean=(intercept+slope*seq(from=50, to=1500,length=sample_sizeL[j])), sd=standev)
  m1<-lm(y~seq(from=50, to=1500, length=sample_sizeL[j]))
  power_vector[j]=coef(m1)[2]
  }
#plot
plot(power_vector~sample_sizeL, xlab="Sample Size", ylab="Slope")
abline(h=0.04, col="red", lwd=1)
```



```
#Around 3-4 thousand it levels out. But, probably due to the fact that my sd=50,
#it converges exactly on the line. This analysis does show considerable variability
#inside the range of the sample size I did collect, however, so I should probably up
#my sample size a little bit just to be sure!
```

```
#BINARY DATA

#This data is also from my thesis, and here involves the number of galls per
```

```
#leaf as a predictor, and the presence/absence of parasitism on that leaf as
#the response variable. My actual sample size was 1,126.
parasites <- read.csv('gallparasites.csv')
```

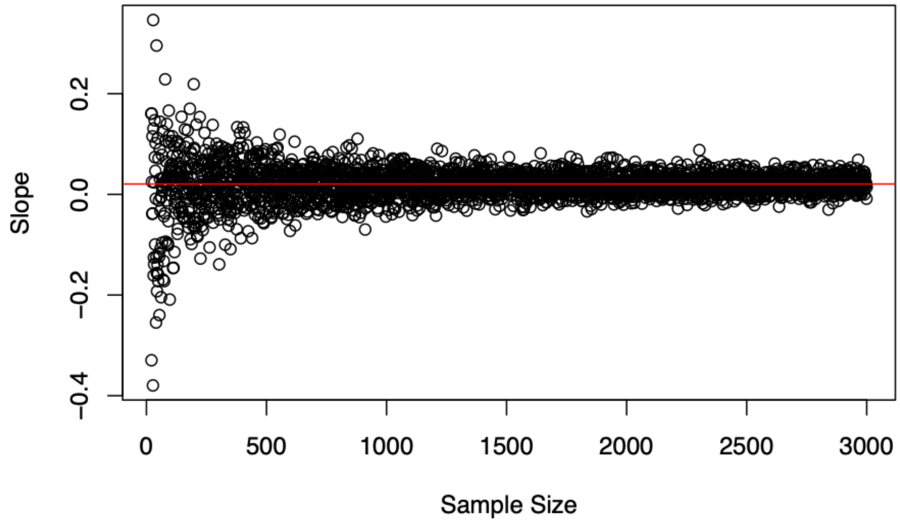
```
#Generating my parameters:
#getcoefs<-glm(parasites$Associates~parasites$Galls_On_Leaf, family = binomial)
#getcoefs
# (Intercept)  parasites$Galls_On_Leaf
# -1.72302                  0.02052
#sample size:
#dim(parasites)
# > dim(parasites)
# [1] 1126    3
```

```
slopeP =0.02052
interceptP = -1.72302

sample_size=seq(from=20, to=3000)
estimated_slope=rep(NA, times=length(sample_size))

for(j in 1:length(sample_size)){
  y=rbinom(n=sample_size[j],
           prob=plogis(interceptP+slopeP*seq(from=1, to=12,length=sample_size[j])),size=1)
  m1<-glm(y~seq(from=1, to=12, length=sample_size[j]), family="binomial")
  estimated_slope[j]=coef(m1)[2]
}

plot(estimated_slope~sample_size, xlab="Sample Size", ylab="Slope")
abline(h=0.02052 , col="red", lwd=1)
```



Given my actual sample size was 1126, it looks like I probably had enough samples to be reasonably sound.

a. How many samples do you need to accurately estimate the slope parameter in a binomial vs. linear regression? Use MSE to calculate the accuracy and precision of your estimate vs. the real value.

```
#Linear
MSEg = mean((0.04-power_vector)^2)
MSEg
```

```
## [1] 2.845793e-05
```

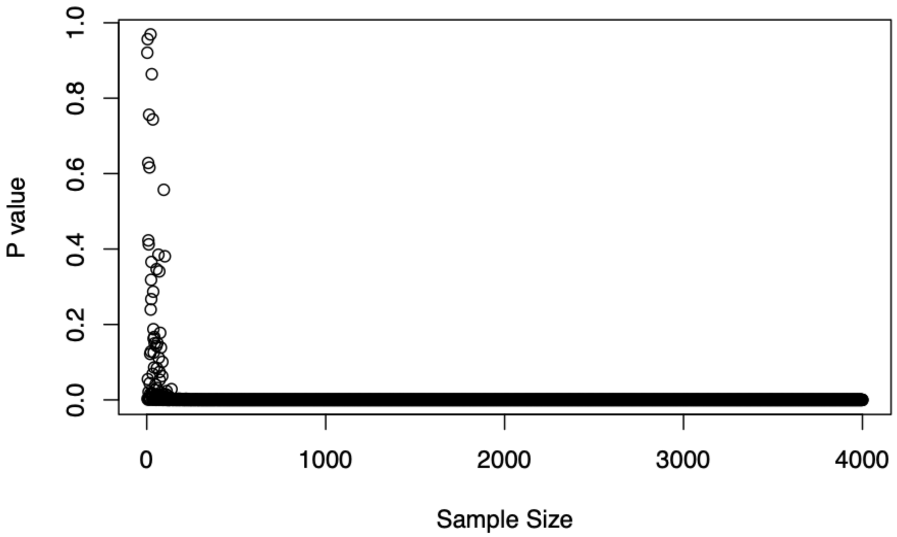
```
#Binomial
MSEp = mean((0.02052-estimated_slope)^2)
MSEp
```

```
## [1] 0.001294954
```

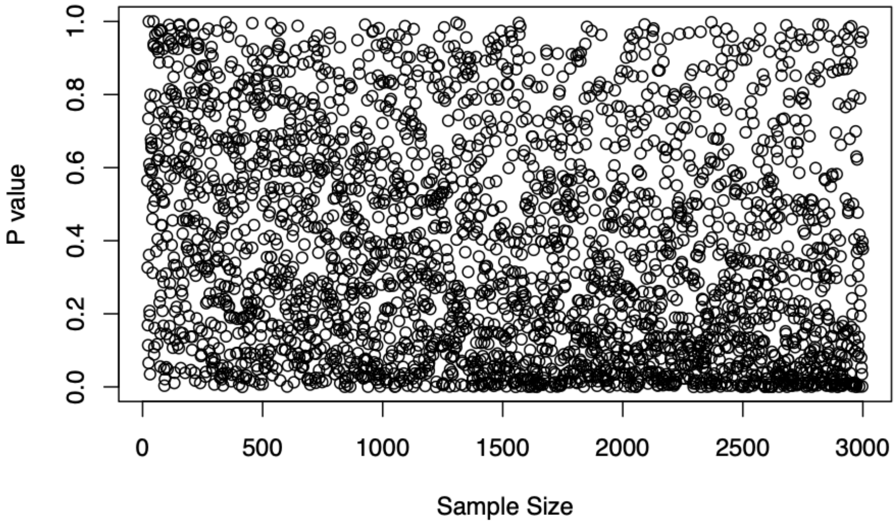
In both cases, I need quite a large sample size to make an accurate estimate- around 1000 in the linear regression, and well over 1000 (somewhere around 1200-1500 or so) in the binomial regression. My guess is that this sample size would be way lower in the linear regression if my sd weren't so incredibly high. In general, it seems a smaller sample size is needed for an accurate estimate using a linear regression.

b. How many samples do you need to ensure a p_value<0.05 for binomial vs. linear regression?

```
#Linear regression:
for(j in 1:length(sample_sizeL)){
  y=rnorm(n=sample_sizeL[j],
          mean=(intercept+slope*seq(from=50, to=1500,length=sample_sizeL[j])), sd=standev)
  m1<-lm(y~seq(from=50, to=1500, length=sample_sizeL[j]))
  power_vector[j]=summary(m1)$coefficients[2,4]
}
plot(power_vector~sample_sizeL, xlab='Sample Size', ylab='P value')
```



```
#Binary regression:
for(j in 1:length(sample_size)){
  y=rbinom(n=sample_size[j],
           prob=plogis(interceptP+slopeP*seq(from=1,to=12, length=sample_size[j])),size=1)
  m1<-glm(y~seq(from=1, to=12, length=sample_size[j]), family="binomial")
  estimated_slope[j]=summary(m1)$coefficients[2,4]
}
plot(estimated_slope~sample_size, xlab='Sample Size', ylab='P value')
```



For the linear regression, it looks like I need to collect somewhere in the low to mid 100s of samples to get a p value of .05 or less. In the binary regression, it looks like there is little hope of getting a certain p value in that range. I would guess this is because my predictor variable just doesn't seem to have much predictive power. It is interesting to see how much more likely significant p values are with larger samples sizes even under these circumstances.

c. In general, why is statistical power generally higher for continuous than discrete response variables?

I'm not entirely sure, but my guess is that continuous data carries more statistical information in a smaller range than discrete data. In continuous data, you can, for example, generate an infinite set of numerical possibilities even between just 2 numbers, and you can also measure an area under the curve for two numbers that are very close together. For discrete data, you need a vastly larger set of actual points in order to generate these kinds of comparisons.