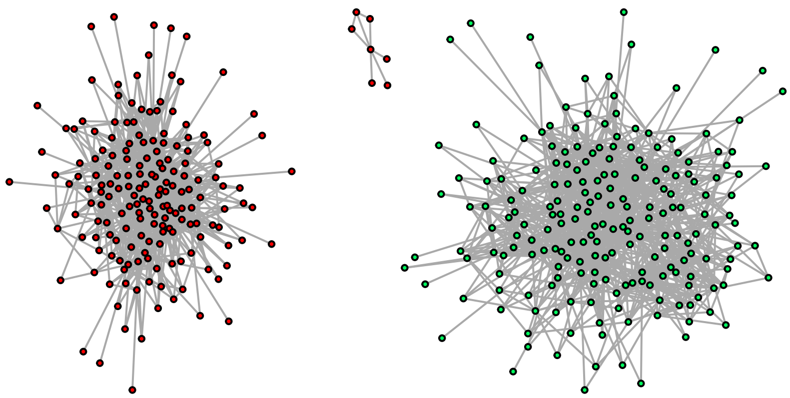
Homework 3 Assignment

Carrie Mecca, Charlie Marcou, Jessie Bustin and Jasmine Zhang

Each question is 10 points. Bonus is up to 5 points.

Starter code

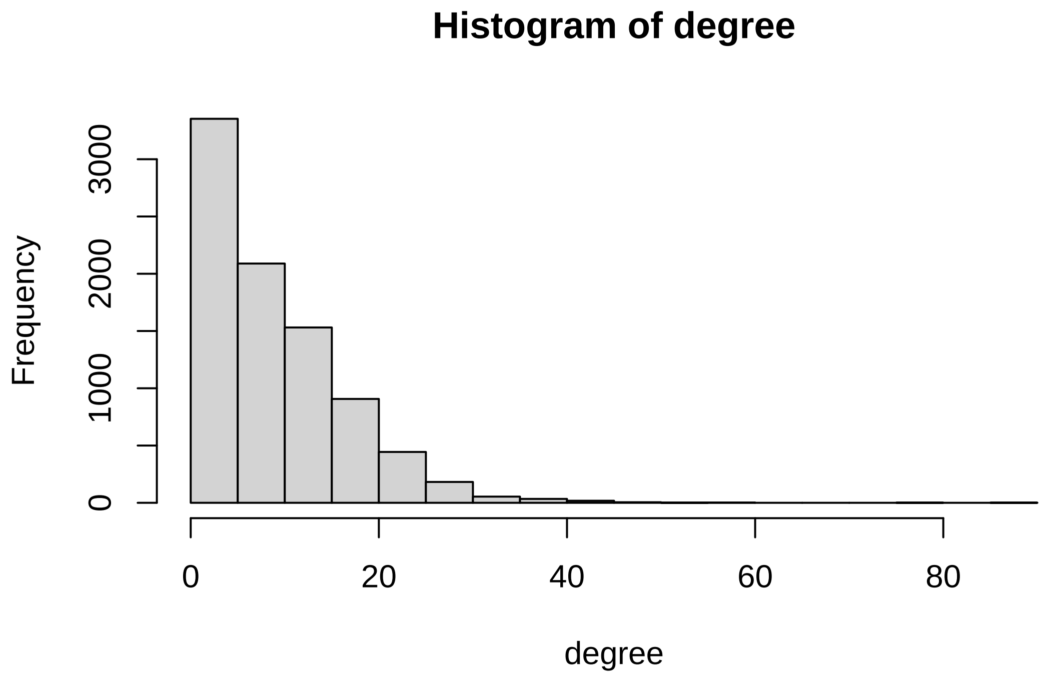


Question 1

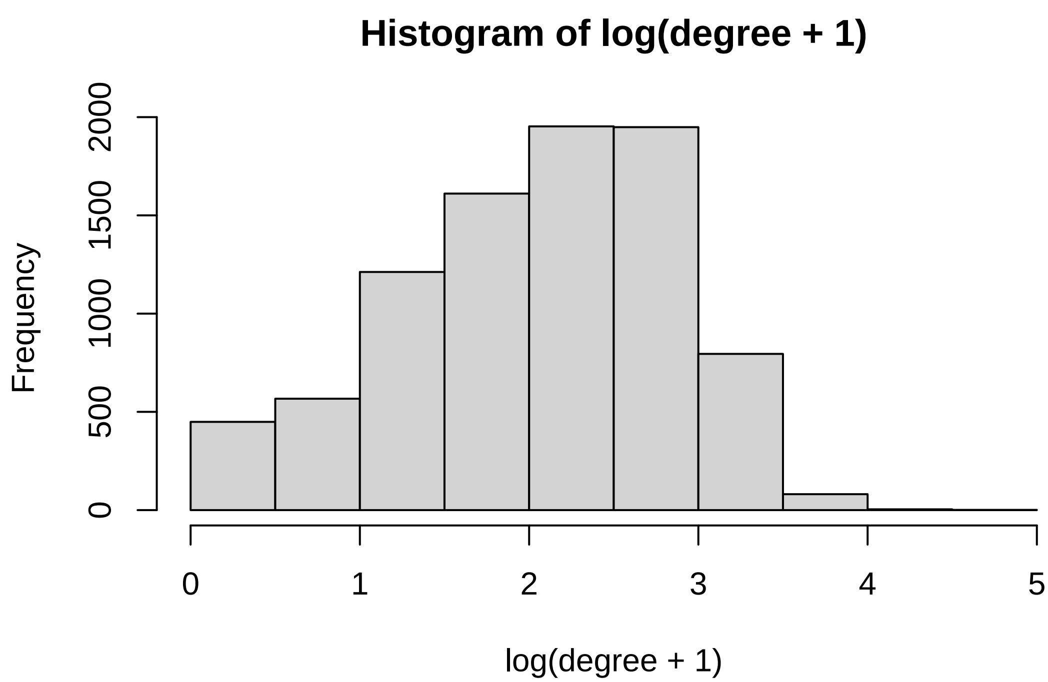
I'd transform degree to create our treatment variable *d*. What would you do and why?

We can see that degree is heavily left-skewed. A transformation to correct this might be a log or square root transformation. Because degree can take on the value of zero, we need to use a log(degree +1). When comparing histograms of the log and square root transformations, the log transformation seems to be more normally distributed so we will choose to use the log transformation.

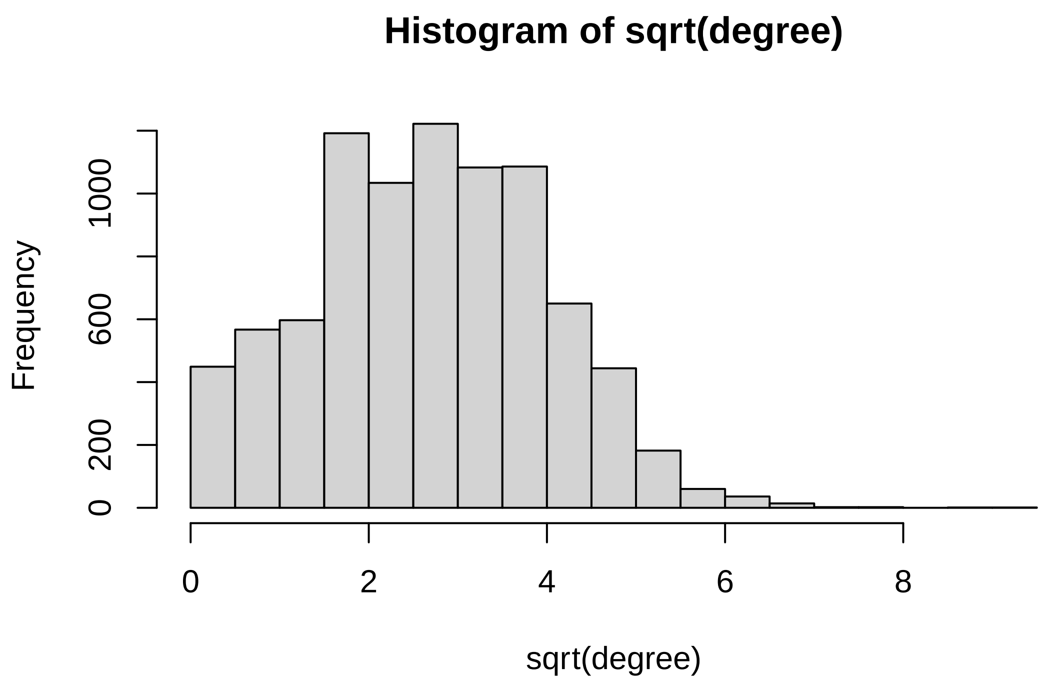
```
hist(degree)
```




```
hist(log(degree+1))
```




```
hist(sqrt(degree))
```




```
d=log(degree+1)
```

Question 2

Build a model to predict d from x, our controls. Comment on how tight the fit is, and what that implies for estimation of a treatment effect.

We can see that the R^2 is 0.06938527, which is not very large, which indicates the fit is not very good. This means that little of the treatment effect of degree on getting a loan is explained by the controls.

```
library(gamlr)

controls = hh[,c(4:5)]
v <- factor(hh$village)
v <- factor(v, levels=c(NA,levels(v)), exclude=NULL)

rl <- factor(hh$religion)
rl <- factor(rl, levels=c(NA,levels(rl)), exclude=NULL)

rf <- factor(hh$roof)
rf <- factor(rf, levels=c(NA,levels(rf)), exclude=NULL)

e <- factor(hh$electricity)
e <- factor(e, levels=c(NA,levels(e)), exclude=NULL)

o <- factor(hh$ownership)
o <- factor(o, levels=c(NA,levels(o)), exclude=NULL)

l <- factor(hh$leader)
l <- factor(l, levels=c(NA,levels(l)), exclude=NULL)

x = sparse.model.matrix(~ v + rl + rf + e + o + l, data=controls)[,-1]
dim(x)
```

```
# do LASSO of treatment on confounders
treat <- gamlr(x,d,lambda.min.ratio=1e-4)
##doing what we did in last hw to evaluate fit
treat.summary=summary(treat)
treat.summary[ (treat.summary$aicc ==min(treat.summary$aicc)),]
```

Question 3

Use predictions from Q2 in an estimator for effect of d on loan.

```
##Get predictions
y <- hh$loan
dhat <- predict(treat, x, type="response")
causal <- gamlr(cbind(d,dhat,x),y,free=2,lmr=1e-4)
coef(causal)["d",]
cor(drop(dhat),d)^2 #R^2
```

Question 4

Compare the results from Q3 to those from a straight (naive) lasso for loan on d and x. Explain why they are similar or different.

We can see that estimated treatments effects are very similar (.01616348 vs .01615451). This makes sense because we saw before that the R^2 between d and x is low. Since d seems to move very independently from x, it make sense that the naive lasso has a similar estimated treatment effect as the double lasso. The double lasso helps remove the influence of d_hat, but there wasn't much of an influence.

```
naive <- gamlr(cbind(d,x),y)
coef(naive)["d",]
```

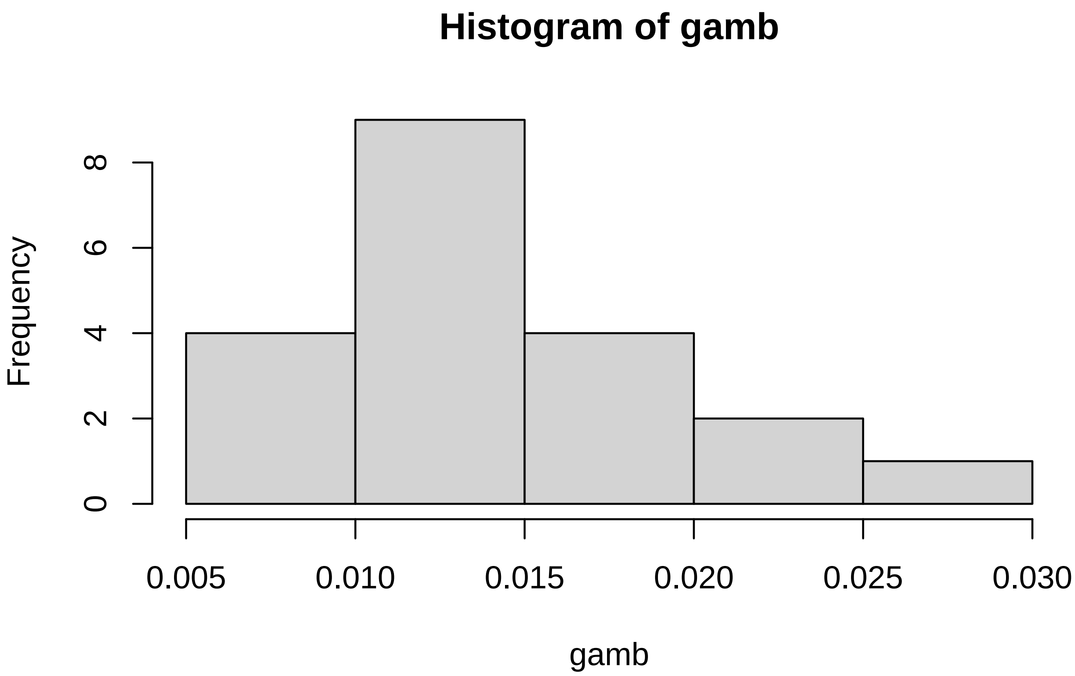
Question 5

Bootstrap your estimator from Q3 and describe the uncertainty.

The standard error for this bootstrapped estimator is only 0.005092135, with a confidence interval from 0.005349029 to 0.02571757. Because the distribution is not wide, and we have a small standard error, the uncertainty of the estimator is low.

```
## BOOTSTRAP
n <- nrow(x)

## Bootstrapping our lasso causal estimator is easy
gamb <- c() # empty gamma

for(b in 1:20){
    ## create a matrix of resampled indices
    ib <- sample(1:n, n, replace=TRUE)

    ## create the resampled data
    xb <- x[ib,]
    db <- d[ib]
    yb <- y[ib]
```

```r
    ## run the treatment regression
    treatb <- gamlr(xb,db,lambda.min.ratio=1e-3)
    dhatb <- predict(treatb, xb, type="response")
    fitb <- gamlr(cbind(db,dhatb,xb),yb,free=2)
    gamb <- c(gamb,coef(fitb)["db",])
}

summary(gamb)
hist(gamb)
```




```r
# get a standard error from Bootstrap
sd(gamb)
mean(gamb)+2*sd(gamb)
mean(gamb)-2*sd(gamb)
```

Bonus

Can you think of how you'd design an experiment to estimate the treatment effect of network degree?

You might randomly assign families to homes of varying degree but block by the control variables. Then we can observe what households seek loans. This randomizes treatment, so it is no longer just an observational study.