BUS 41201 Homework 4 Assignment

Group 11: YuTing Weng, Mengdi Hao, Elena Li, Minji park, Sarah Lee

2024-04-14

```
# Read in Data
setwd("C:/Users/user/Desktop/Big Data/HW/week4")
hh <- read.csv("microfi_households.csv", row.names="hh")
hh$village <- factor(hh$village)
```

```
zebra <- match(rownames(hh), V(hhnet)$name)
## number of commerce,friend,family connections
degree <- degree(hhnet)[zebra]
names(degree) <- rownames(hh)
degree[is.na(degree)] <- 0
```

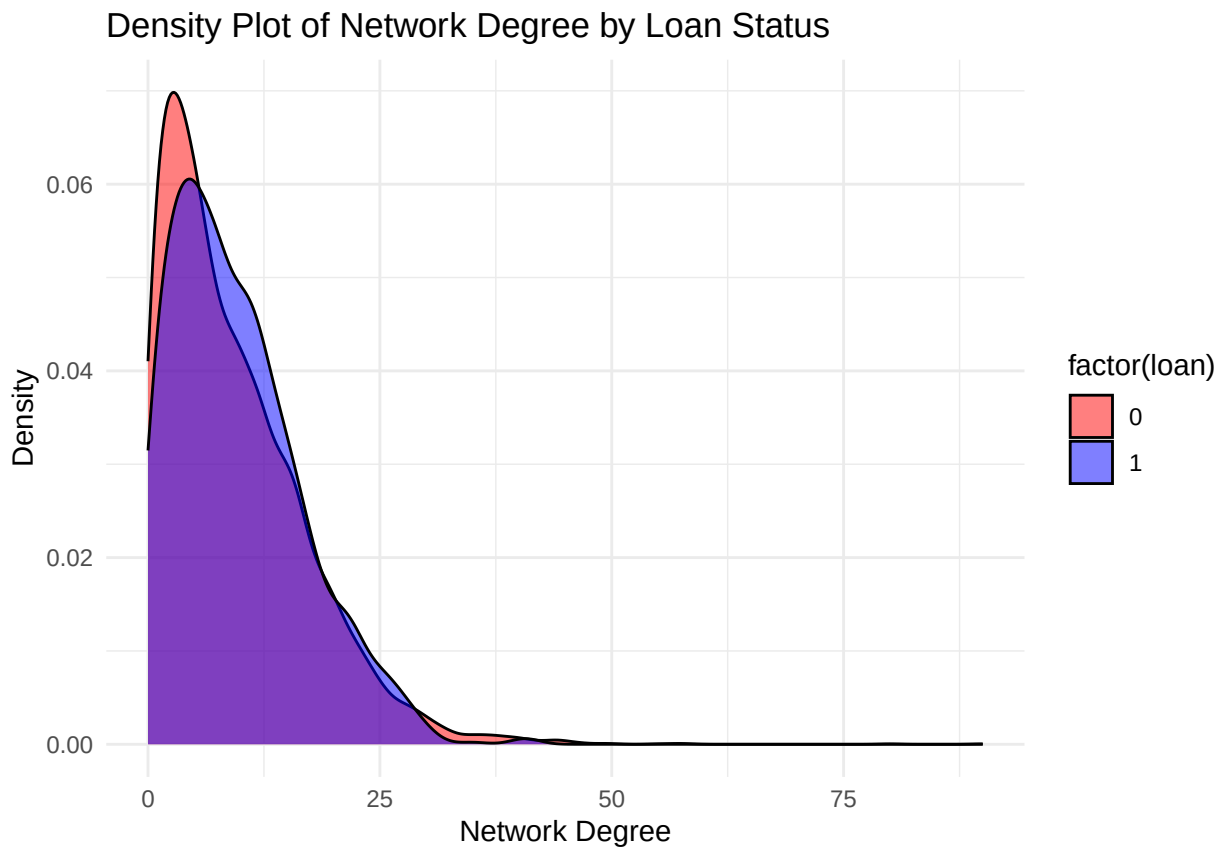
Question 1

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

```
library(ggplot2)

# Create a data frame with 'degree' and 'loan' variables
data <- data.frame(degree = degree, loan = hh$loan)

# Plot density plot
ggplot(data, aes(x = degree, fill =factor(loan))) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values=c('red','blue'))+
  labs(x = "Network Degree", y = "Density", title = "Density Plot of Network Degree by Loan Status") +
  theme_minimal()
```

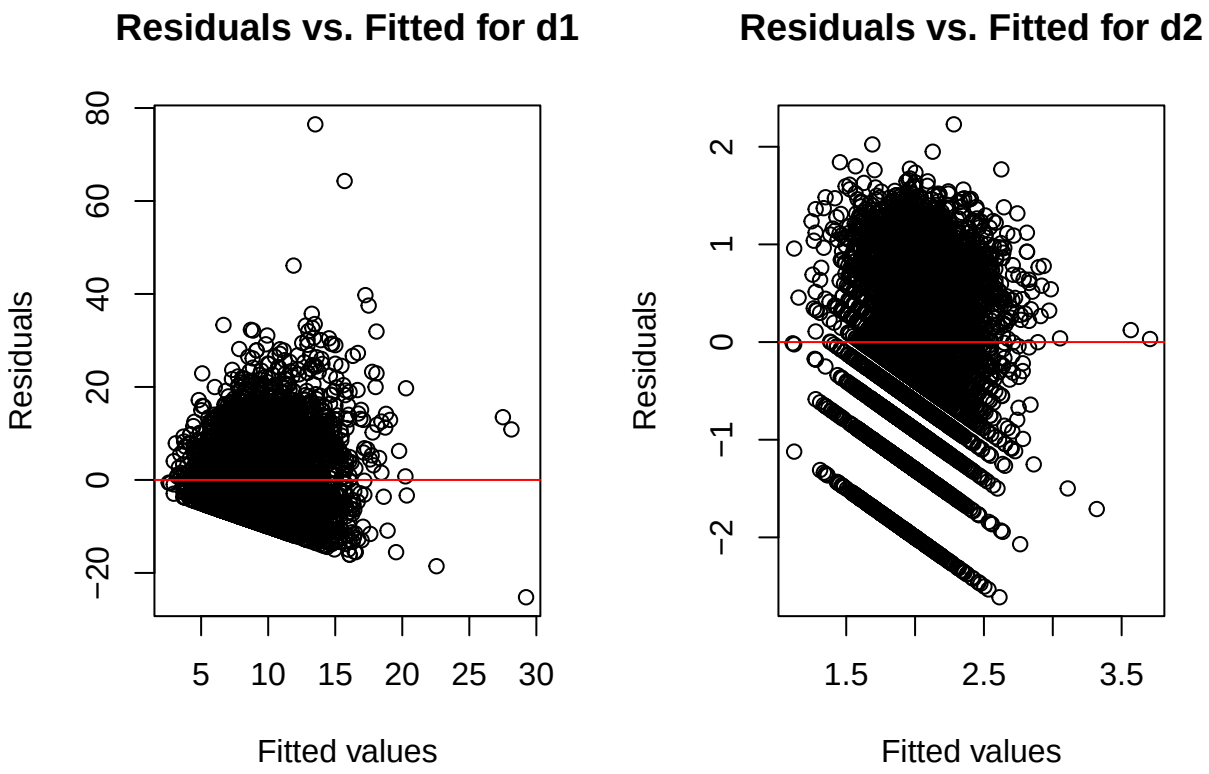


The log transformation is chosen for the degree variable due to its right-skewed distribution, typical in count data like network degrees. This transformation addresses skewness, aligning the data with linear regression assumptions, while also simplifying interpretation by linearizing relationships and enabling coefficients to represent percentage changes.

```
hh$degree <-degree
hh$degree_log <-log1p(degree)
```

```
x <- model.matrix(~ village + religion + roof + rooms + beds +
electricity + ownership + leader - 1, data=hh)
d1 <- hh$degree
d2 <- hh$degree_log
model1 <- gamlr(x, d1)
dhat1 <- predict(model1, x, type="response")
residuals1 <- d1 - dhat1
model2 <- gamlr(x, d2)

dhat2 <- predict(model2, x, type="response")
residuals2 <- d2 - dhat2
par(mfrow=c(1,2)) # Set up the graphics window to show two plots side by side
plot(dhat1, residuals1, xlab="Fitted values", ylab="Residuals", main="Residuals vs. Fitted for d1")
abline(h=0, col="red")
### Step 4: Plot Residuals vs. Fitted Values for d2
plot(dhat2, residuals2, xlab="Fitted values", ylab="Residuals", main="Residuals vs. Fitted for d2")
abline(h=0, col="red")
```



Based on these plots, the log transformation of the degree variable has improved the homogeneity of variance in the residuals, making the assumption of constant variance more tenable.

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.

```
hh <-naref(hh)

# Define the dependent variable
y <-hh$loan
# Define the treatment variable
d <-hh$degree_log

# Define control variables
controls <-data.frame( hh[, c("village","religion", "roof", "rooms", "beds",
                              "electricity", "ownership", "leader")])
x <- sparse.model.matrix(~.-1,  data=controls)
```

```
# First stage LASSO
treat <-gamlr(x,d)
# Predict treatment variable using control variable
dhat <- predict(treat,x, type='response')
```

```r
# Calculate in-sample R^2
cor(drop(dhat),d)^2
```

```
## [1] 0.08166425
```

The In-Sample R^2 is 0.0817 from the first stage of a two-stage LASSO model. This indicates that only about 8.17% of the variance in the treatment variable d is explained by the control variables x. This low R^2 suggests a weak predictive power of the controls over the treatment variable. It might indicate that the model has not captured all relevant confounding variables, potentially leading to omitted variable bias in estimating the treatment effect, risking biased results in the second stage.

Question 3

Use predictions from [2] in an estimator for effect of d on loan.

```r
# Second-stage LASSO
causal <- gamlr(cbind(d,dhat,x),y,free=2)

# Extract the treatment effect coefficient
coef(causal)['d',]
```

```
## [1] 0.01812596
```

```r
# Odd Multiplier
exp(coef(causal)["d",])
```

```
## [1] 1.018291
```

The coefficient for the treatment variable degree log is 0.0181, corresponding to an odds multiplier of exp(0.0181)=1.0183. This suggests a modest positive effect of the degree of connection on the likelihood of making loans. Specifically, a one percent increase in the degree of connection increases the odds of a household making a loan by about 1.83%.

Question 4

Compare the results from [3] to those from a straight (naive) lasso for loan on d and x. Explain why they are similar or different.

```r
# NAIVE LASSO: directly regress y on x and d
naive <- gamlr(cbind(d,x),y, family='binomial')

# Extract the treatment effect coefficient from th naive LASSO
coef(naive)["d",]
```

```
## [1] 0.1485588
```

```r
# Odd multiplier
exp(coef(naive)["d",])
```

```
## [1] 1.160161
```

In the naive LASSO model, the coefficient for degree log is 0.1486, translating to an odds multiplier of exp(0.1486)=1.1602. This indicates a stronger positive effect on the likelihood of making loans compared to the two-stage LASSO: a one percent increase in the degree of connection raises the odds by about 16.02%.

The naive LASSO model, by directly incorporating the treatment variable and control variables, may overlook confounding variables, potentially inflating the coefficient of the treatment variable due to confounding effects. Conversely, the two-stage LASSO model addresses this issue by including the predicted treatment values (dhat) alongside the control variables, effectively mitigating the impact of confounding variables and providing a more accurate estimation of the treatment effect.

Question 5

Bootstrap your estimator from [3] and describe the uncertainty.

```r
n <- nrow(x)

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

set.seed(123)

for(b in 1:100){
    ## 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]

    ## 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",])

    #print(b)
}
```

```r
#Summary Statistics of 100 estimators:
summary(gamb)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005384 0.014820 0.017946 0.017777 0.020996 0.030354
```

```r
cat("Standard Error of Treatment Effect: ", sd(gamb), "\n")
```

```
## Standard Error of Treatment Effect:  0.00488506
```

```r
cat("95% Confidence Interval: (", quantile(gamb, 0.025),",", quantile(gamb, 0.075), ")\n" )
```

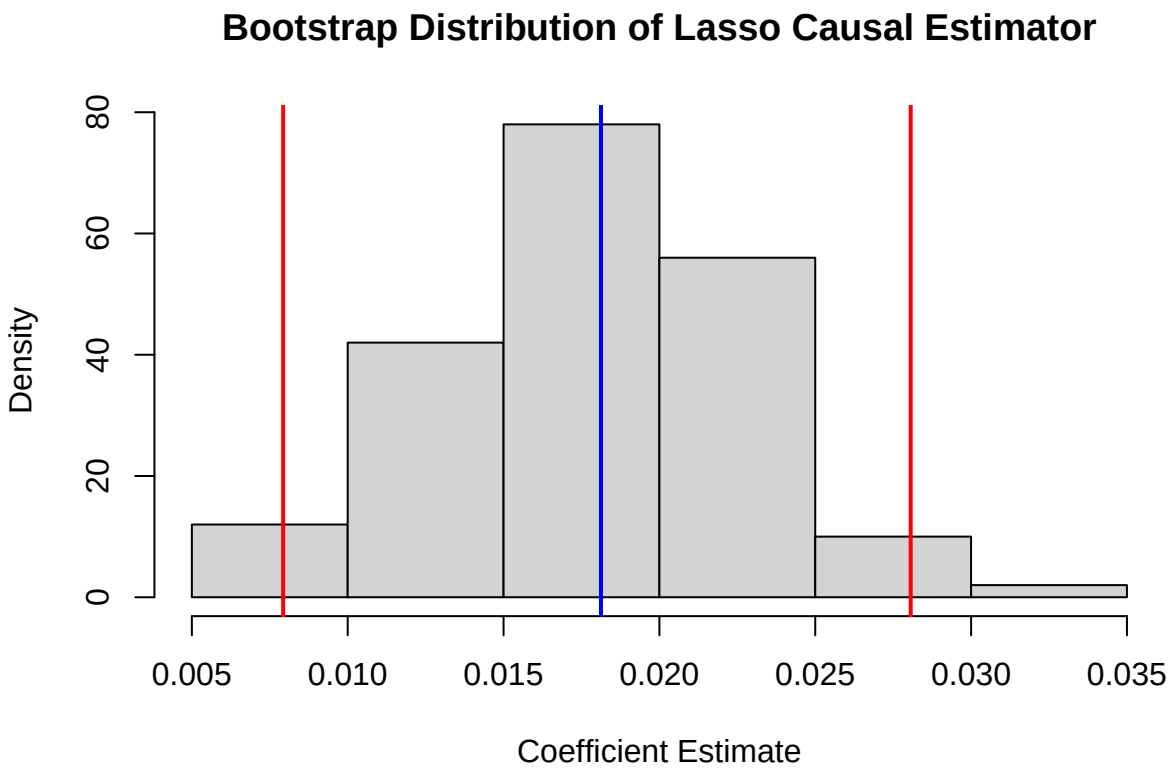
```
## 95% Confidence Interval: ( 0.00792944 , 0.01049061 )
```

```r
cat("Original Estimate: ", coef(causal)['d',], "\n")
```

```
## Original Estimate:  0.01812596
```

```r
# Plotting
hist(gamb, freq = FALSE, main = "Bootstrap Distribution of Lasso Causal Estimator",
     xlab = "Coefficient Estimate", ylab = "Density")
abline(v = coef(causal)["d",], col = "blue", lwd = 2)

# Confidence interval from bootstrap
abline(v = quantile(gamb, 0.025), col = "red", lwd = 2)
abline(v = quantile(gamb, 0.975), col = "red", lwd = 2)
```



In the bootstrapping procedure conducted 100 times, the resulting distribution of treatment effect coefficients showed a range from 0.0054 to 0.0304, with a mean of 0.0178. This suggests some minimal variability in the estimated treatment effect across the bootstrap samples. The standard error of the treatment effect was calculated to be 0.0049, indicating the precision of the estimate.

The original estimate from the model was 0.0181. Overall, the bootstrapping results suggest that while the original estimate is within the range of bootstrapped estimates, there is some uncertainty around its precise value.

Bonus Question

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

To assess the impact of network degree on the probability of obtaining a loan, a randomized controlled trial (RCT) could be designed. In this trial, participants would be randomly assigned to either a treatment group, where they receive targeted networking opportunities aimed at increasing their network degree, or a control group, where they do not receive any intervention. It's crucial to ensure that both groups are similar in all other relevant characteristics. By comparing the loan acquisition rates between the treatment and control groups and leveraging the increase in network degree induced by the intervention as an instrumental variable, we can estimate the causal effect of network degree on loan approval likelihood. This experimental design would provide valuable insights into whether broader networks directly impact the success rate of loan applications.