BUSN 41201 HW4

Group 11: Yu-Ting Weng, Mengdi Hao, Elena Li, Minji Park, Sarah Lee

```
hh <- read.csv("microfi_households.csv", row.names="hh")
hh$village <- factor(hh$village)
```

```
library(igraph)
edges <- read.table("microfi_edges.txt", colClasses="character")
## edges holds connections between the household ids
hhnet <- graph.edgelist(as.matrix(edges))
hhnet <- as.undirected(hhnet) # two-way connections.
```

```
## igraph is all about plotting.
V(hhnet) ## our 8000+ household vertices
## Each vertex (node) has some attributes, and we can add more.
V(hhnet)$village <- as.character(hh[V(hhnet),'village'])
```

```
## igraph is all about plotting.
V(hhnet) ## our 8000+ household vertices
## Each vertex (node) has some attributes, and we can add more.
V(hhnet)$village <- as.character(hh[V(hhnet),'village'])
## we'll color them by village membership
vilcol <- rainbow(nlevels(hh$village))
names(vilcol) <- levels(hh$village)
V(hhnet)$color = vilcol[V(hhnet)$village]
## drop HH labels from plot
V(hhnet)$label=NA
```

```
village1 <- induced.subgraph(hhnet, v=which(V(hhnet)$village=="1"))
village33 <- induced.subgraph(hhnet, v=which(V(hhnet)$village=="33"))
```

Question 1

Transform degree to create our treatment variable d. What would you do and why?

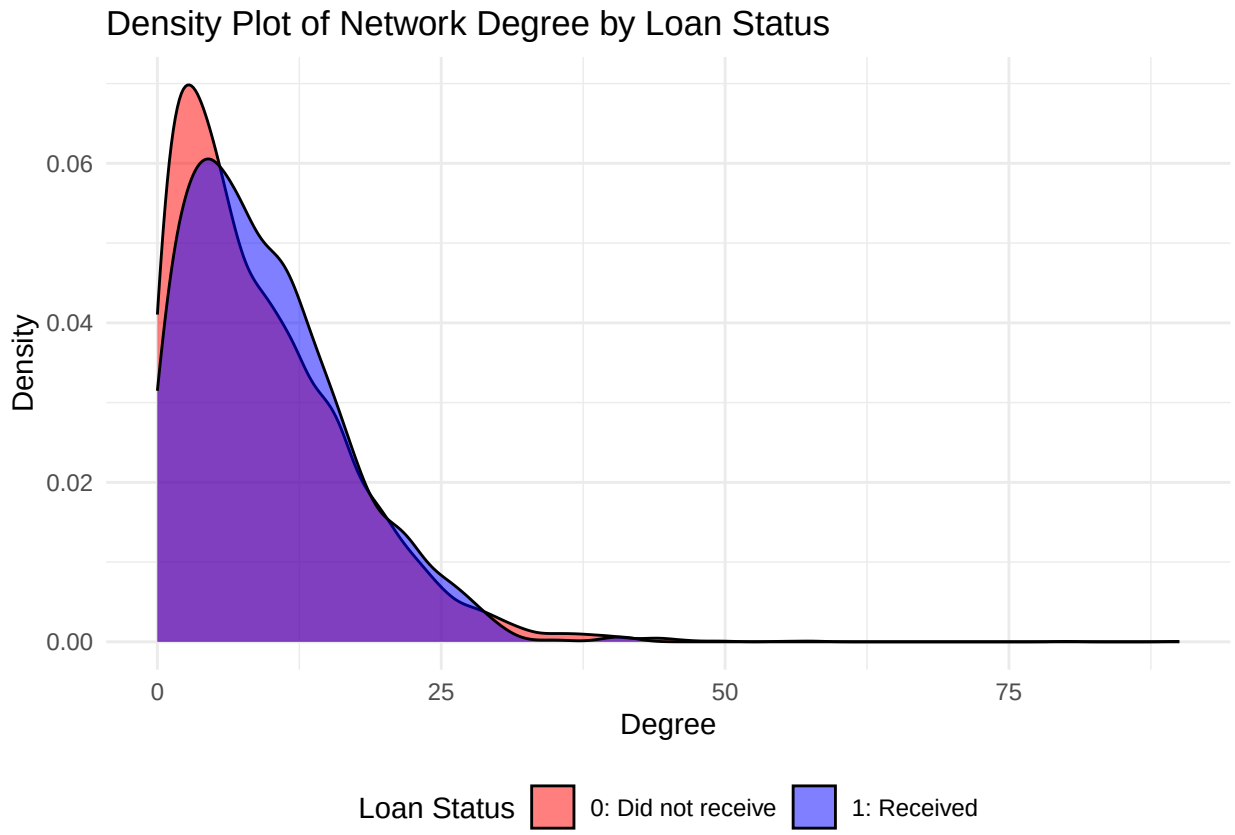
```
library(gamlr)
```

```
## Loading required package: Matrix
```

```
zebra <- match(rownames(hh), V(hhnet)$name)

degree <- degree(hhnet)[zebra]
names(degree) <- rownames(hh)
degree[is.na(degree)] <- 0
```

```
library(ggplot2)
ggplot(hh, aes(x=degree, fill=factor(loan))) +
  geom_density(alpha=0.5) +
  scale_fill_manual(values=c("red", "blue"),
```

```
                    name="Loan Status",
                    labels=c("0: Did not receive", "1: Received")) +
  labs(x="Degree", y="Density", title="Density Plot of Network Degree by Loan Status") +
  theme_minimal() +
  theme(legend.position="bottom")
```



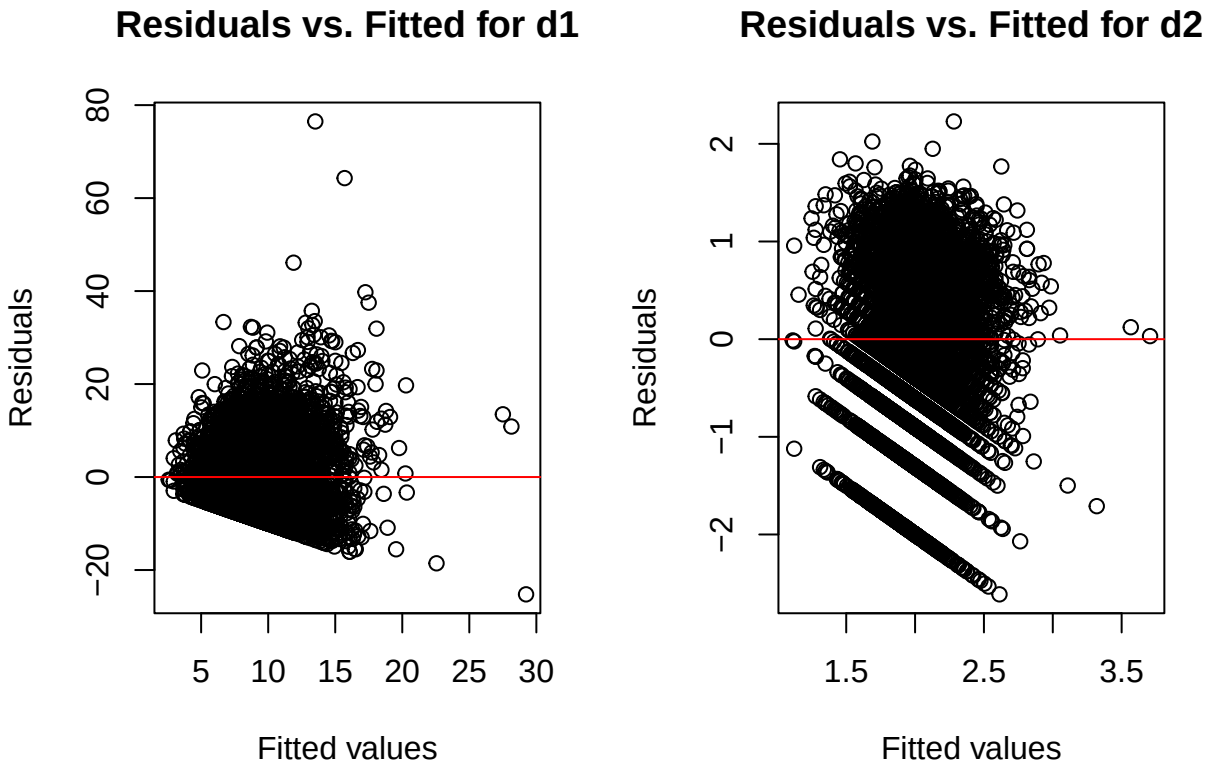

```
hh$degree <- degree
hh$degree_log <- log1p(degree)
```

From the plot, it appears that the degree variable is **right-skewed** for both groups, which is common for count data like network degrees—there are many households with a low degree and fewer with a high degree. In the context of a generalized linear model (GLM) with a continuous response, the **log transformation** was applied to the degree variable to help conform its distribution to the normality assumption required for the model's explanatory variables.

- Comparison of residuals before and after the transformation

```
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.

```
y <- hh$loan
d <- hh$degree_log
treat <- gamlr(x,d)

dhat <- predict(treat, x, type="response")
cor(drop(dhat),d)^2
```

```
## [1] 0.0818752
```

Considering that the R-squared value governs the amount of independent signal available for estimating, the value of **0.08187** indicates **a limited independent signal from the predictors for the variation in network degree.** This low level of explained variance suggests that the model may need to provide a stronger basis for estimating treatment effects reliably. Additional relevant predictors or a better-specified model may be necessary to capture the independent influence on the outcome.

Question 3

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

```
causal <- gamlr(cbind(d,dhat,x),y,family = "binomial")
```

```
## 'as(<dgeMatrix>, "dgCMatrix")' is deprecated.
## Use 'as(., "CsparseMatrix")' instead.
## See help("Deprecated") and help("Matrix-deprecated").
```

```
coef(causal)["d",]
```

```
## [1] 0.1562872
```

Question 4

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

```
library(glmnet)
```

```
## Loaded glmnet 4.1-8
```

```
treat <- gamlr(x,d,lambda.min.ratio = 1e-4)
dhat <- predict(treat, x, type = 'response')
causal <- gamlr(cbind(d,dhat,x),y,free=2)
coef(causal)["d",]
```

```
## [1] 0.0187176
```

The screenshots show two coefficient estimates for the variable d obtained from two gamlr models. In Question 3, the estimate is approximately 0.167, while in Question 4, which applies a stricter regularization, the estimate is around 0.018. The difference between these estimates suggests that the regularization strength greatly affects the coefficient size. The stricter regularization in Question 4 likely penalizes the coefficients more heavily, shrinking them towards zero, which might explain the reduced estimate for d. This can indicate either variable d having a less robust relationship with the outcome than initially indicated or the presence of overfitting in the less regularized model from Question 3.

Question 5

Bootstrap your estimator from 3 and describe the uncertainty.

```
n <- nrow(x)

## Bootstrapping our lasso causal estimator is easy

gamb <- c() # empty gamma

for(b in 1:30){
    ## create a matrix of resampled indices
```

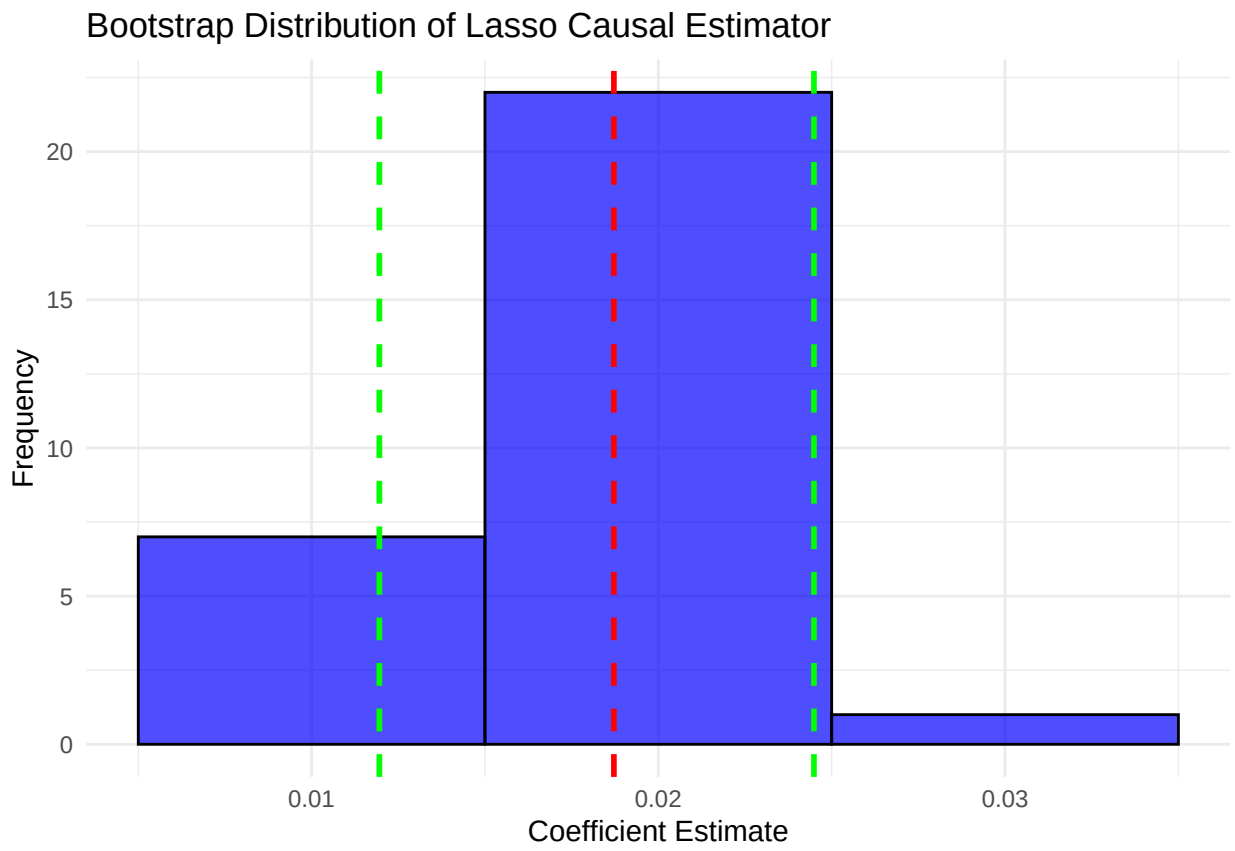
```
    ib <- sample(1:n, n, replace=TRUE)

    ## 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",])
}
summary(gamb)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.009643 0.015222 0.017337 0.018033 0.021308 0.025369
```

```
original_estimate <- coef(causal)["d",]

ci <- quantile(gamb, probs = c(0.025, 0.975))

bootstrap_df <- data.frame(Estimates = gamb)

p <- ggplot(bootstrap_df, aes(x = Estimates)) +
  geom_histogram(binwidth = 0.01, fill = "blue", color = "black", alpha = 0.7) +
  geom_vline(xintercept = original_estimate, color = "red", linetype = "dashed", size = 1) +
  geom_vline(xintercept = ci[1], color = "green", linetype = "dashed", size = 1) +
  geom_vline(xintercept = ci[2], color = "green", linetype = "dashed", size = 1) +
  labs(title = "Bootstrap Distribution of Lasso Causal Estimator",
       x = "Coefficient Estimate",
       y = "Frequency") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
print(p)
```




```
summary(gamb)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.009643 0.015222 0.017337 0.018033 0.021308 0.025369
```

```
# Print the original estimate and its CI
cat("Original Estimate: ", original_estimate, "\n")
```

```
## Original Estimate:  0.0187176
```

```
cat("95% CI: (", ci[1], ", ", ci[2], ")\n")
```

```
## 95% CI: ( 0.01195419 ,  0.02448857 )
```

The bootstrapped estimates exhibit wider variability than the initial model's uncertainty, indicating a higher level of uncertainty in the coefficient's estimation.

Additional Question

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

To determine the effect of network degree on the likelihood of obtaining a loan, I'd implement a randomized controlled trial providing some participants with targeted networking opportunities while others serve as a control group, ensuring both groups are similar on all other characteristics. By comparing the rates of loan acquisition between the groups and using the increase in network degree induced by the intervention as an instrumental variable, the causal impact of network degree on the probability of securing a loan could be estimated. This design would clarify whether broader networks directly influence the success rate of loan applications.