41201-01: Data Mining

Professor Taddy

Joanne Chen

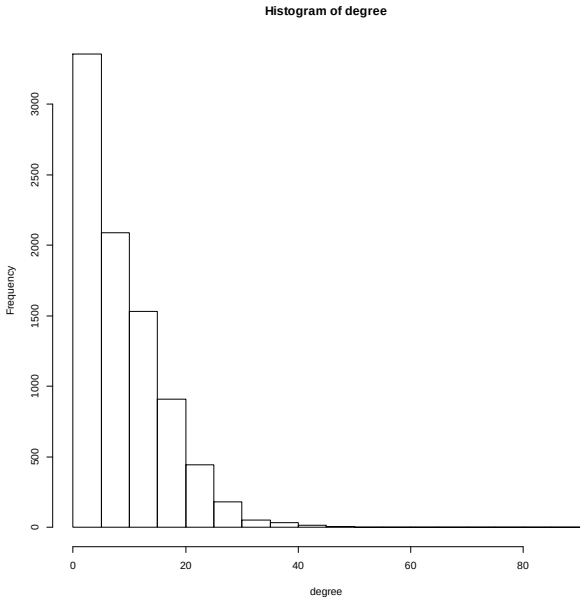
James Hardiman

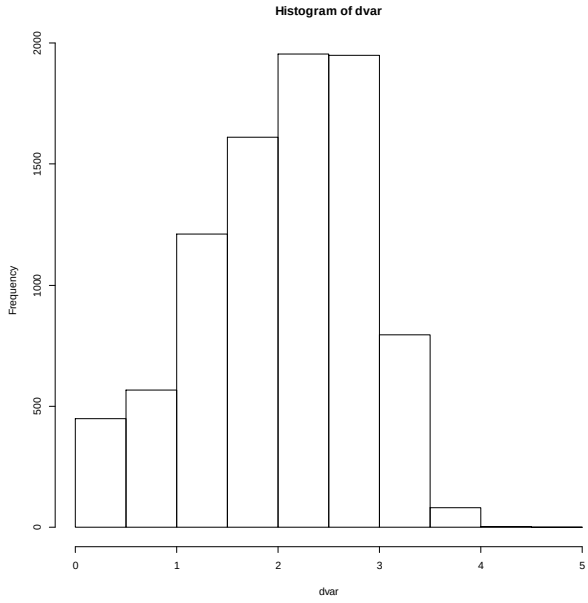
Evan Johnson

Honor code: We pledge our honor that we have not violated the Honor Code during the completion of this assignment.

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

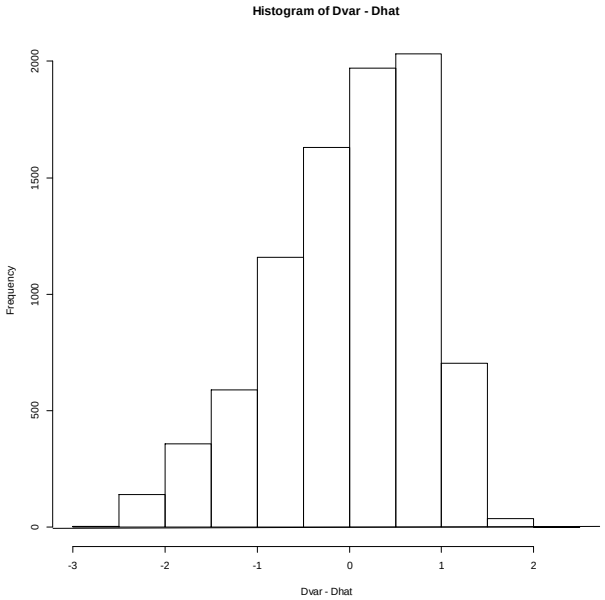
We chose to transform degree using a log transform. We mapped degree to a new vector called dvar = ln(degree + 1). This gives us a more normally distributed connection measure. It also has convenient properties. This includes mapping 0 degrees to 0 in dvar. It also makes linearizes percentage increases in degrees. This puts changes in degrees from 1 to 2 on the same scale as from 10 to 20, which makes intuitive sense as we would expect decreasing effect of marginal connections.

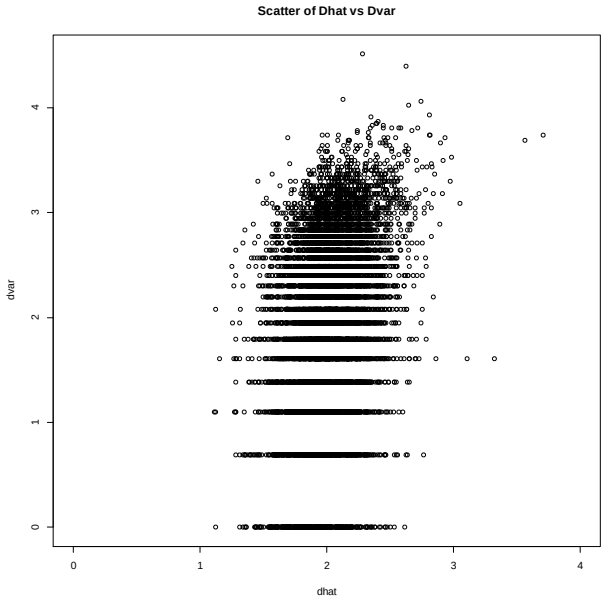




[2]. Build a model to predict d from x, our controls. Comment on how good you think the model is.

We built a model to predict dvar from our x's (prediction of dvar is dhat). This model performed mediocre and is able to predict only ~8% of the variance in dvar; however, as you can see from the histogram of the errors (mean = 0, s.d. = 0.8) and a plot of dvar vs. dhat (plotted with a basic linear regression between the two) there is a substantial portion of dvar that cannot be predicted by the x's. This gives us some confidence that we may be able to conduct an observational experiment with dvar.





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

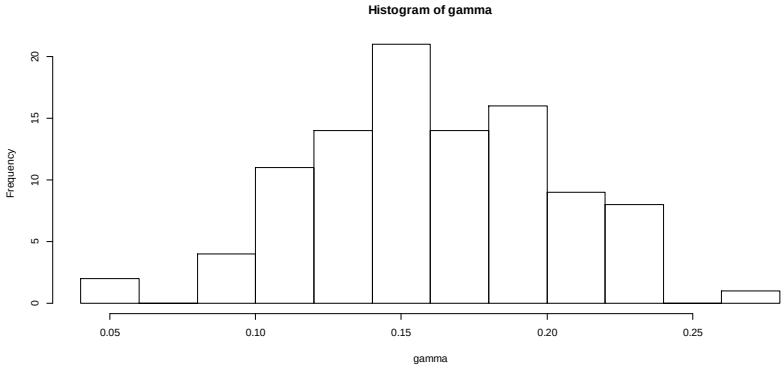
Including dhat in a gamlr regression of all covariates on loan (where we do not penalize dhat coefficients) we obtain a coefficient for dvar of 0.16089. This implies that for every unit increase in dvar (i.e., ln(degree + 1)) an individual is 16% more likely to obtain a loan.

[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.

If we do not include dhat in a gamlr regression of all covariates on loan, we obtain a coefficient for dvar of 0.1661. This is approximately -0.00521 difference between the co-efficient obtained including an estimator for dvar. This is not surprising considering how poorly x is at predicting dhat. Thus we would expect that including an unpenalized dhat as a function of x would not have a large influence on the estimated effect of dvar.

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

Using a bootstrapping 100 times to resample our data we can generate uncertainty around our estimate of the coefficient of dvar. Doing so we generated a distribution with mean 0.1613 and standard deviation of 0.041. A histogram of the 100 different dvar coefficients resulting from our bootstrapping process is shown below.



[+]. Fit the BCH algorithm and compare to results above. NB: loan is binary.