Big Data: Homework 4

Will Clark & Matthew DeLio
41201-01

April 30, 2015

1 Node Connectivity Transformation

Node connectivity (which we are calling `degree`) is measured by the number of edges for each node in a network. In this context, `degree` tells us the number of relationships that a household in our population has. We observe in Figure 1 that `degree` is distributed logarithmically.

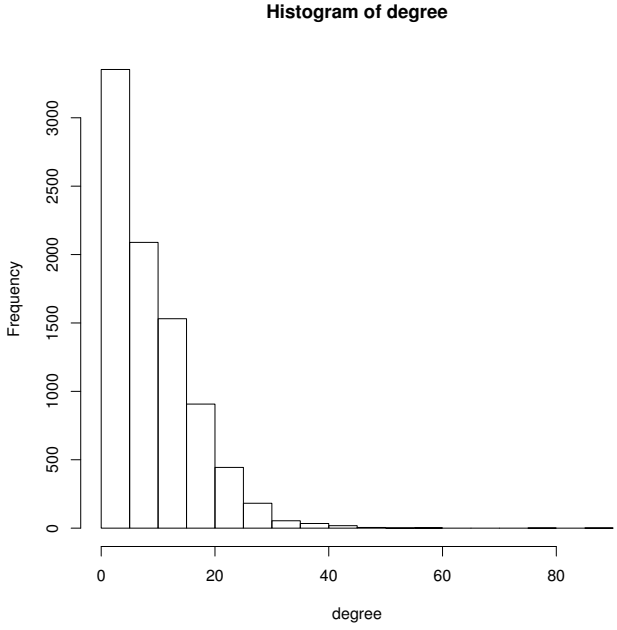

Figure 1: Distribution of `degree`

2 Predicting Node Connectivity from Controls

In this section, we build a model to predict a node's degree by using only our control variables. Our model is:

$$d = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$$

where d is a node's number of degrees, $\boldsymbol{X}$ is a vector of control variables including village, religion, type of roof on home, rooms and beds in home, a dummy variable for having electricity in home, whether a home is owned, and whether the person is a "leader" in the village.

We estimate this model using a Gamma-Lasso regression. In Figure 3 in the Appendix, we show the Gamma-Lasso path plots with five decision criteria marked: AIC, AICc, BIC, CV.Min, and CV.1se. The $\log(\lambda)$ selected by AICc and by CV.Min are reasonably close to each other (-4.60 and -4.46, respectively, shown in Table 1 in the Appendix), which provides us with a confirmation that our model is estimated reasonably well.

We then use the model selected by AICc to predict degree (which we will call $\hat{d}$). We plot d against $\hat{d}$ in Figure 2 and see that there is only a very rough correlation between the two ($\sigma_{d,\hat{d}} = 0.34$)

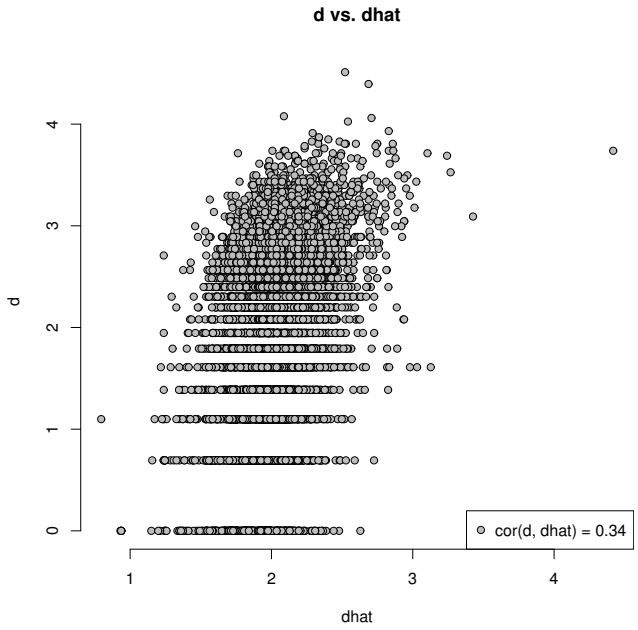

Figure 2: Degree and Predicted Degree

This is a positive result. It tells us that most of the variation in degree, which will be our treatment variable in Section 3, is exogenous and cannot be explained by the control variables that we observe. We can therefore measure the effect of degree as a treatment on the propensity to take out a loan and be reasonably sure that we are not simply measuring variation in other observed control variables.

3 Effect of Node Connectivity on Loan Propensity

4 Naive Estimation of Loan Propensity

5 Bootstrapping Uncertainty

6 Experimental Design

7 Appendix

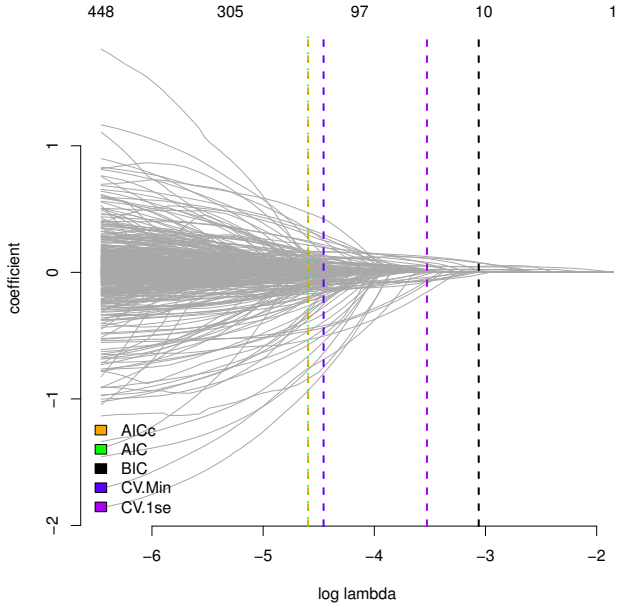

Figure 3: Gamma-Lasso Regression for Degree on Controls

	$\log(\lambda)$	Covariates Selected
AICc	-4.60	185
AIC	-4.60	185
BIC	-3.06	10
CV.Min	-4.46	161
CV.1se	-3.53	37

Table 1: Treatment IC Table