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.

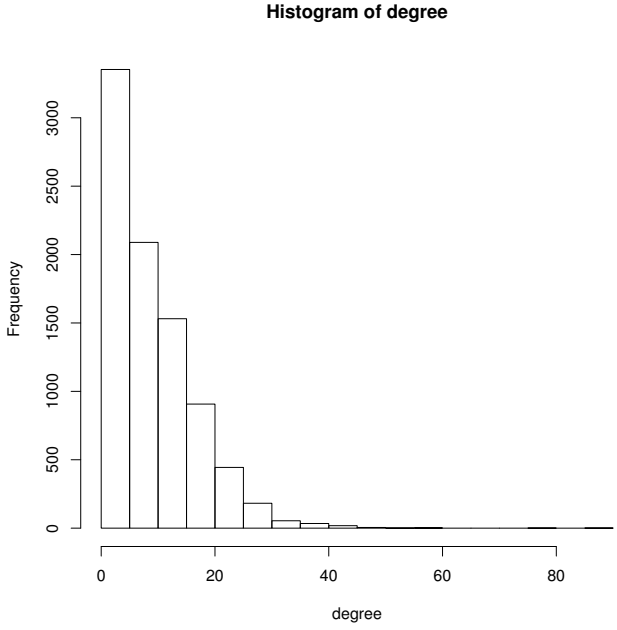

Figure 1: Distribution of `degree`

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:

$$d(x) = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \nu \tag{1}$$

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 3 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 ($\sigma_{d,\hat{d}} = 0.34$)

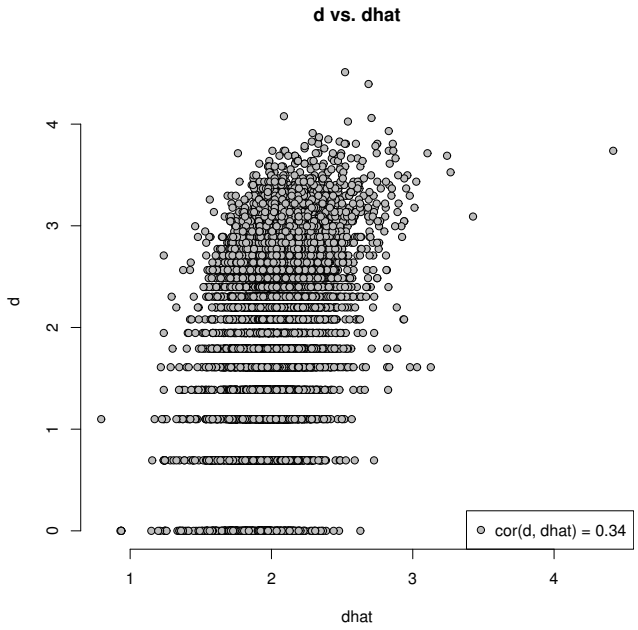

Figure 2: Degree and Predicted Degree

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

In this section, we use our estimate of predicted degree ($\hat{d}$) to build a model for loan propensity based on node connectivity. We include $\hat{d}$ in our regression model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + d\gamma + \hat{d}(x)\delta + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \tag{2}$$

where $p = \Pr(\text{loan} = 1|\boldsymbol{X})$. We can substitute in our definition of $\hat{d}(x)$ from equation 1:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + d\gamma + (\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \nu)\,\delta + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \tag{3}$$

which simplifies to

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \gamma\nu + \hat{d}(x)(\gamma + \delta) + \boldsymbol{X}\boldsymbol{\beta} \tag{4}$$

By estimating the model in equation 2, we can isolate the parameter γ, which is ultimately the effect of the independent variation in d on the propensity to take out a loan.

As in Section 2, we use a Gamma-Lasso regression and see that the $\log(\lambda)$ values for the AICc and CV.Min are roughly similar (-5.37 and -5.13, respectively, shown in Table 4 and plotted in Figure 4 in the Appendix). This tells us our model is reasonably estimated; we proceed using the model chosen by AICc.

The coefficent of interest in this regression is γ, the effect of exogenous changes in degree connectivity on propensity to take out a loan. In our model, γ is 0.0145, which means that having one additional connection (i.e. degree increased by one) increases the odds of taking out a microfinance loan by 1.46 percent[1]. This seems like a reasonable estimate —the coefficient is small but positive, indicating that knowing more people makes one very slightly more likely to take out a loan. We can imagine a story in which information travels more efficeintly to those who are more connected, but it is reasonable to assume that there are much more significant predictors of loan propensity than how many friends a person has.

In Table 1 and 2, we see the 5 most positive and negative predictors of loan propensity. At first glance, it appears that there are a few very positive factors that would increase propensity to take out a loan. For example, living in village 4 and having a thatch roof makes one 56.7 percent more likely to have a loan. Upon closer examination, though, there is only one person in our sample in village 4 with a thatch roof, so we cannot infer much from this coefficient. In fact, the top 5 coefficients have sample sizes of $n < 5$.

The coefficeints in Table 2 are more meaningful as the sub-sample sizes are sufficiently large to draw conclusions from. For example, Hindus in village 71 are 4.7 percent less likely to take out a loan. We are sufficiently confident in this coefficient because the sample size of Hindus in village 71 is 252.

This seems to be one drawback to using a matrix of control factors in which all possible terms interact with each other. In many cases, the Gamma-Lasso regression does not throw out coefficients that we might want to ignore anyway because of sample sizes that are insufficiently large.

[1] $100 \cdot (\exp(0.0145) - 1) = 1.46$

	x	β_j	Δ likelihood	n
1	village4:roofthatch	0.4483	56.5631	1
2	village21:ownershipRENTED	0.4229	52.6338	1
3	village3:ownershipSHARE OWNED	0.4122	51.0091	4
4	village65:ownershipLEASED	0.4118	50.9473	1
5	village1:roofthatch	0.4079	50.3722	1

Table 1: Most Significant Positive Predictors

	x	β_j	Δ likelihood	n
1	village50	-0.0417	-4.0834	244
2	village59:roofstone	-0.0441	-4.3095	268
3	village36	-0.0442	-4.3231	289
4	village71:religionhindu	-0.0483	-4.7140	252
5	village20:roofthatch	-0.0903	-8.6304	5

Table 2: Most Significant Negative Predictors

4 Naive Estimation of Loan Propensity

5 Bootstrapping Uncertainty

6 Experimental Design

7 Appendix

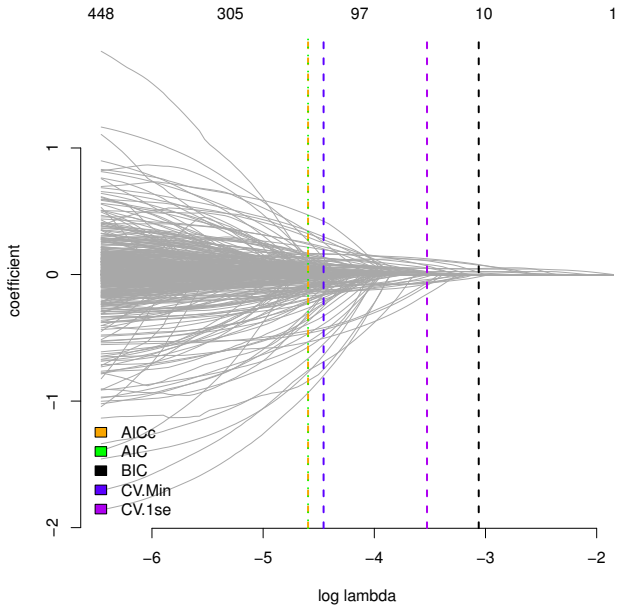

Figure 3: Gamma-Lasso Regression for Degree on Controls

	$\log(\lambda)$	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 3: Treatment IC Table

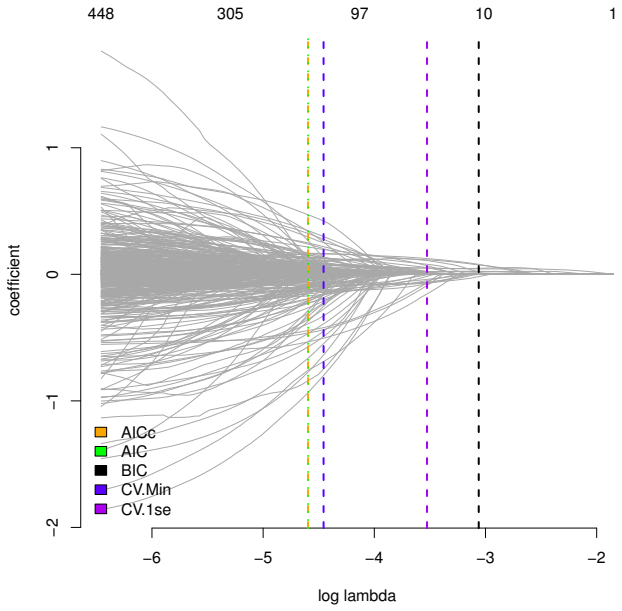

Figure 4: Gamma-Lasso Regression for Loan Propensity on Degree

	$\log(\lambda)$	Covariates Selected
AICc	-5.37	148
AIC	-5.37	148
BIC	-3.97	12
CV.Min	-5.13	116
CV.1se	-4.02	14

Table 4: Causal IC Table