Big Data: Homework 4

Will Clark & Matthew DeLio
41201-01

May 1, 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. Because one of the assumptions of linear regression requires all variables to be normally distributed, we will need to transform `degree` to make it more "normal".

One of the transformations we initially considered was to transform our metric from just `degree` to be the ratio of relationships a household has that also obtained loans. The hope here was that the fraction of one's immediate network of relationships that had loans would be roughly normally distributed. Unfortunately from Figure 2a we can see that this was not the case. While slightly less logarithmic, much of the distribution's mass was spread concentrated around 0, indicating that most of one's network had not obtained loans. An attempt to make it more normal by taking the log of this fraction is also shown Figure 2b. As we can clearly see from this figure, the histogram is no more normal than before. In the end this attempt was abandoned.

Another attempt was made to transform degree directly. This time, we focused on simply taking $\log(degree)$. However, since a few households appear to be isolated (i.e. not in the graph), taking the log directly would result in many NaN's. This is avoided by simply adding one to degree (since degree cannot be < 0 by definition) before taking the natural log: $\log(degree + 1)$. The resulting distribution is shown in Figure 3 and appears to be more normally distributed than the other normalization attempt. This transformation of `degree` will be used throughout these analyses as `d`

$$d = \log(degree + 1) \tag{1}$$

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:

$$d(x) = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \nu \tag{2}$$

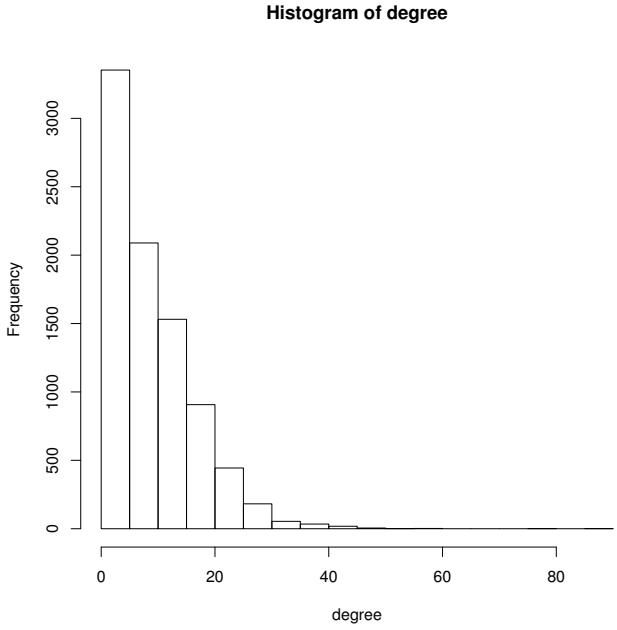


Figure 1: Distribution of `degree`

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 6 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.

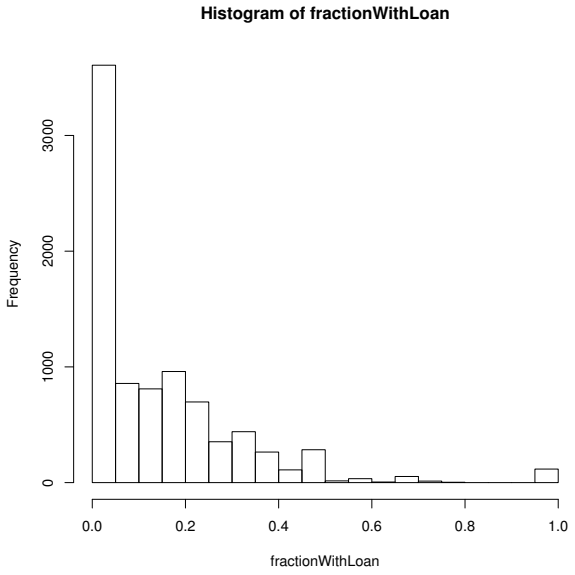
We then use the model selected by AICc to predict degree (which we will call $\hat{d}$). We plot d against $\hat{d}$ in Figure 4 and see that there is only a very rough correlation between the two ($\sigma_{d,\hat{d}} = 0.34$)

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.

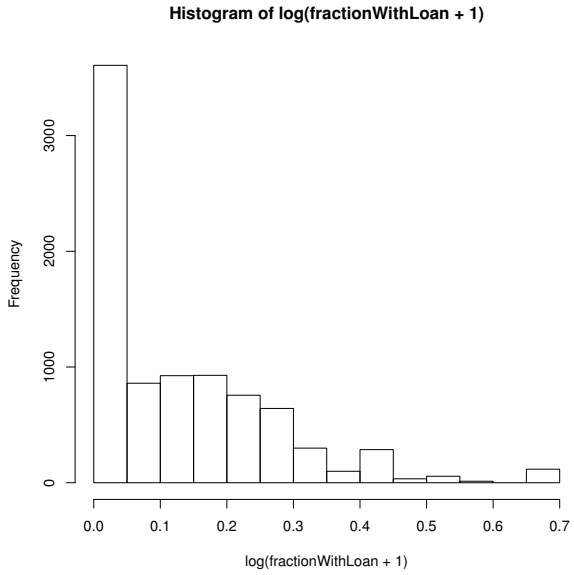
3 Effect of Node Connectivity on Loan Propensity

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:

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



(a) Distribution of Connected Households with Loans



(b) Log Distribution of Connected Network with Loans

Figure 2: Normalization of Degree: Fraction of Household's Network with Loans

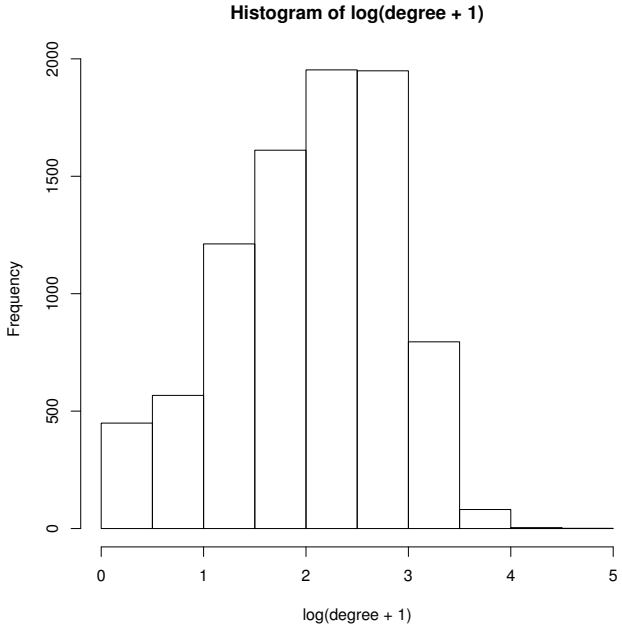


Figure 3: Distribution of $\log(\texttt{degree} + 1)$

where $p = \Pr(\text{loan} = 1|\boldsymbol{X})$. We can substitute in our definition of $\hat{d}(x)$ from equation 2:

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

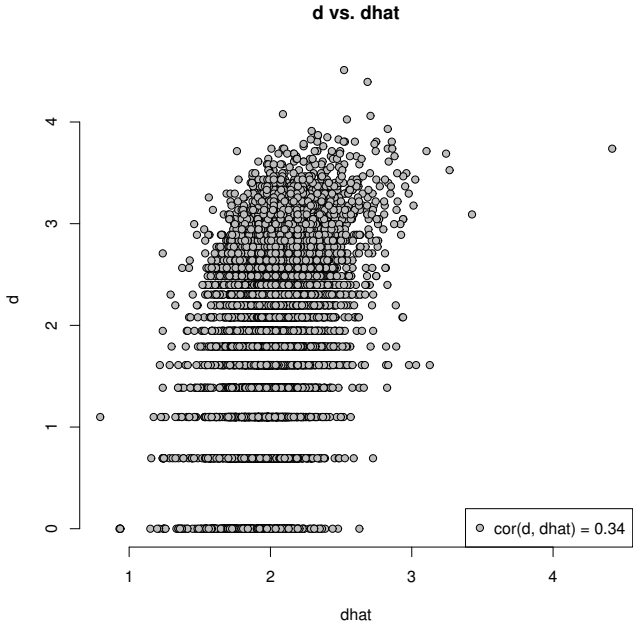


Figure 4: Degree and Predicted Degree

which simplifies to

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \gamma\nu + \hat{d}(x)(\gamma+\delta) + \boldsymbol{X}\boldsymbol{\beta} \tag{5}$$

By estimating the model in equation 3, 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 7 in the Appendix). This tells us our model is reasonably estimated; we proceed using the model chosen by AICc.

The coefficient 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 micro-finance 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 efficiently 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

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

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 coefficients 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.

	x	β_j	Δ odds (%)	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 odds (%)	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

In this section, run a simple regression of the $\log(odds)$ of obtaining a loan on `X` and `d` (see equation 6 for the model). Here we neglect to remove any of the influencers in `X` (i.e. the rest of the model) that are also correlated with `d`.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + d\gamma + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \tag{6}$$

As in the previous sections, we use a Gamma-Lasso regression. Here, again, we see that the AICc and the CV.Min are roughly similar (-5.39 and -5.07 respectively, shown in Table 5 and plotted in Figure 8 in the Appendix). We do note, perhaps with hindsight bias, that the there is a slightly larger divergence between the choices of $\log(\lambda)$ than in the previous sections. Using the Gamma-Lasso with the AICc selection criteria, we see that the coefficient (γ) associated with `d` is significant and equal to 0.0138.

In terms of odds, each additional connection increases the odds of taking out a loan by 1.39%[2]. This result is very similar from Section 3 where each connection increased the odds of taking out a loan by 1.46%. These vales are similar, but ultimately, differ since some of the effects of `d` can be predicted by `X` and have not been controlled for in our model. It seems that neglecting to control for these confounders may not lead to an outright incorrect answer, but rather to one that is subtly different. In this case the increase odds differ by each additional connection only differs by 5%.

5 Bootstrapping Uncertainty

In this section, we estimate our uncertainty in the γ found in Section 3 using the bootstrap method. To accomplish this, we successively take n random samples (with replacement from our data-set), and run the same two regressions discussed in Sections 2 & 3. That is, we run a Gamma-Lasso regression and predict each node's degree ($\hat{d}$) and then, using this prediction, run the Gamma-Lasso to determine the effect of (d) (γ) on the propensity to obtain a loan. These sample - regression - prediction - regression steps are performed 100 times; each time, the estimate of γ is saved.

From our new estimates of γ, uncertainty in our original γ is measured by finding the standard deviation of the samples. Figure 5 shows a histogram of these γ's. The distribution mean is 0.0168, which is notably different than the one found in Section 3 (0.0145). However, with a $\sigma = 0.00422$, the 95% confidence interval for the unconditional uncertainty in γ is $\mu \pm 1.96\sigma = [0.00849, 0.0250]$, which easily contains our previous estimate. In terms of log(odds) this confidence interval concludes that each additional degree of connection corresponds to a lift between $[0.852\%, 2.53\%]$ in the odds of obtaining a loan. This range quite a bit higher than we would have previously guessed and seems to suggest that the degree of household connectedness isn't the best predictor of loan propensity.

6 Experimental Design

Since degrees of connectedness amongst households are primarily observables, it is difficult to truly design an experiment to better estimate the treatment effect. However, one could, perhaps, setup an experiment whereby households in the treatment group are introduced by the experimenters to people they are currently unconnected to. Likewise households in the control group are left alone. If there is a noticeable lift in the propensity to obtain a loan in the treatment group, versus the control group, then one could possibly establish the effect of the degree of connectedness from the experiment.

7 Appendix

[2] $100 \cdot (\exp(0.0138) - 1) = 1.39$

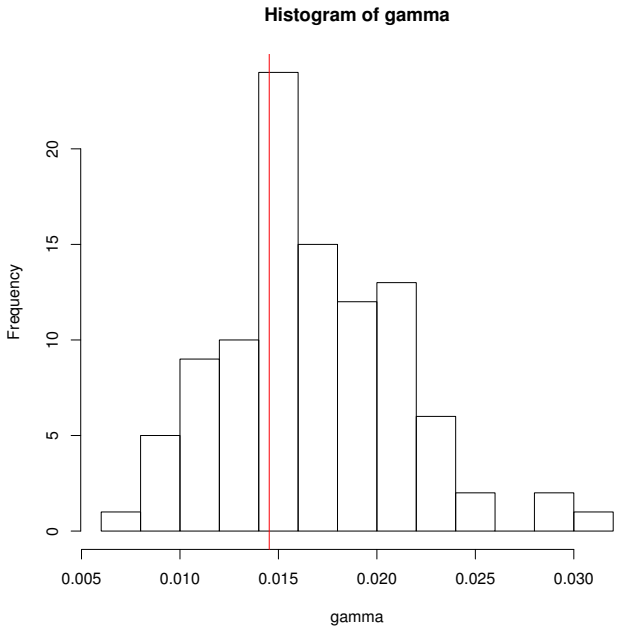


Figure 5: Distribution of γ via Bootstrap Method (k=100)

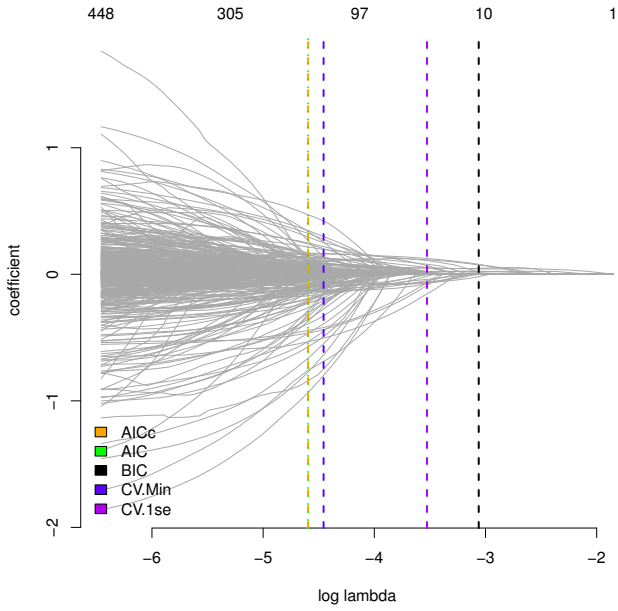


Figure 6: Gamma-Lasso Regression for Degree on Controls

	log(λ)	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

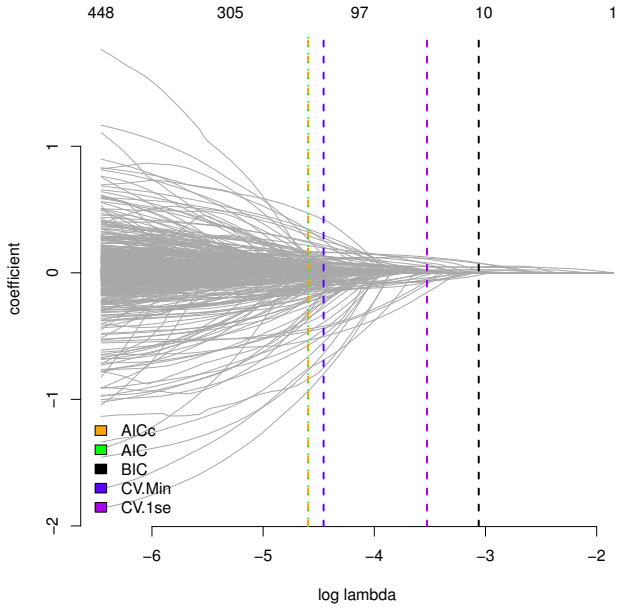


Figure 7: Gamma-Lasso Regression for Loan Propensity on Degree

	log(λ)	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

	log(λ)	Covariates Selected
AICc	-5.39	151
AIC	-5.44	159
BIC	-4.00	14
CV.Min	-5.07	105
CV.1se	-4.23	26

Table 5: Naive IC Table

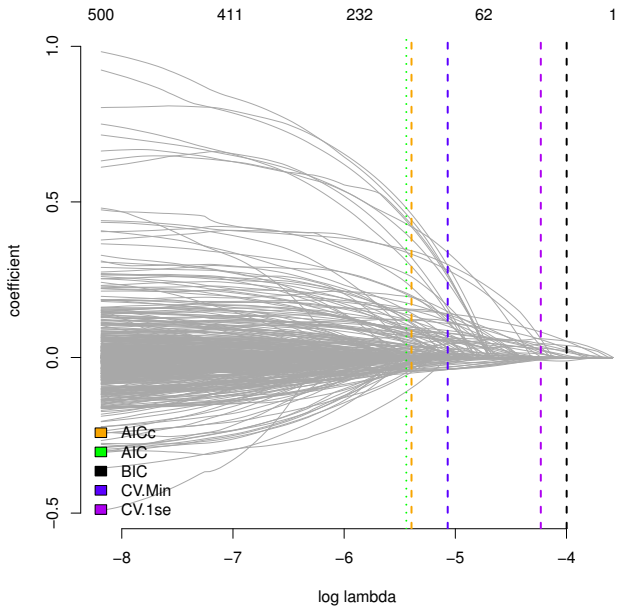


Figure 8: Gamma-Lasso Regression for Naive Loan Propensity