3-USDA Evaluation Probability of Loan

Robert Dinterman

2015-09-08

Broadband Loan Programs

In September 2005, the United States Department of Agriculture (USDA) Rural Utilities Service (RUS) released an Audit Report for their Broadband Grant and Loan Programs. These included two programs: Broadband Pilot Grant/Loan Program and 2002 Farm Bill Broadband Loan Program.

The Pilot Program formed from amendments to the Distance Learning and Telemedicine Program in fiscal year 2001 and 2002 in order to provide funds for rural, lower-income communities for the stated purpose "to encourage telecommunications carriers to provide broadband service to rural consumers where such service had not existed" and for "stimulating economic development and enhancing education and health care opportunities." The Loans were treasury rate loan funds totaling $100 million in 2001 and $80 million in 2002 while the grants totaled $20 million for 2002 and $10 million for 2003. The Pilot Program was meant to be small in scope so that RUS could determine the feasibility and effectiveness of providing loans and grants targeted for projects involving high-speed internet access and availability. The audit found that while the targeted areas were to be rural and low-income, a disproportionate amount of funds ended up in wealth suburban communities around the Houston area which triggered a reevaluation of the disbursement process. These would be corrected with the 2002 Farm Bill program.

The 2002 Farm Bill provided funding for RUS to establish the Rural Broadband Access Loan and Loan Guarantee Program and the Broadband Grant Program. These programs were established in the 2003 fiscal year but were provided funding through the Farm Bill from 2002 to 2007. The grant program expanded its stated directive, stating that funding will be provided for "broadband transmission service that fosters economic growth and delivers enhanced educational, health care, and public safety services." The Farm Bill provided for over $2.7 billion in Loans and Grants.

Determinants of Loan/Grant Receipt

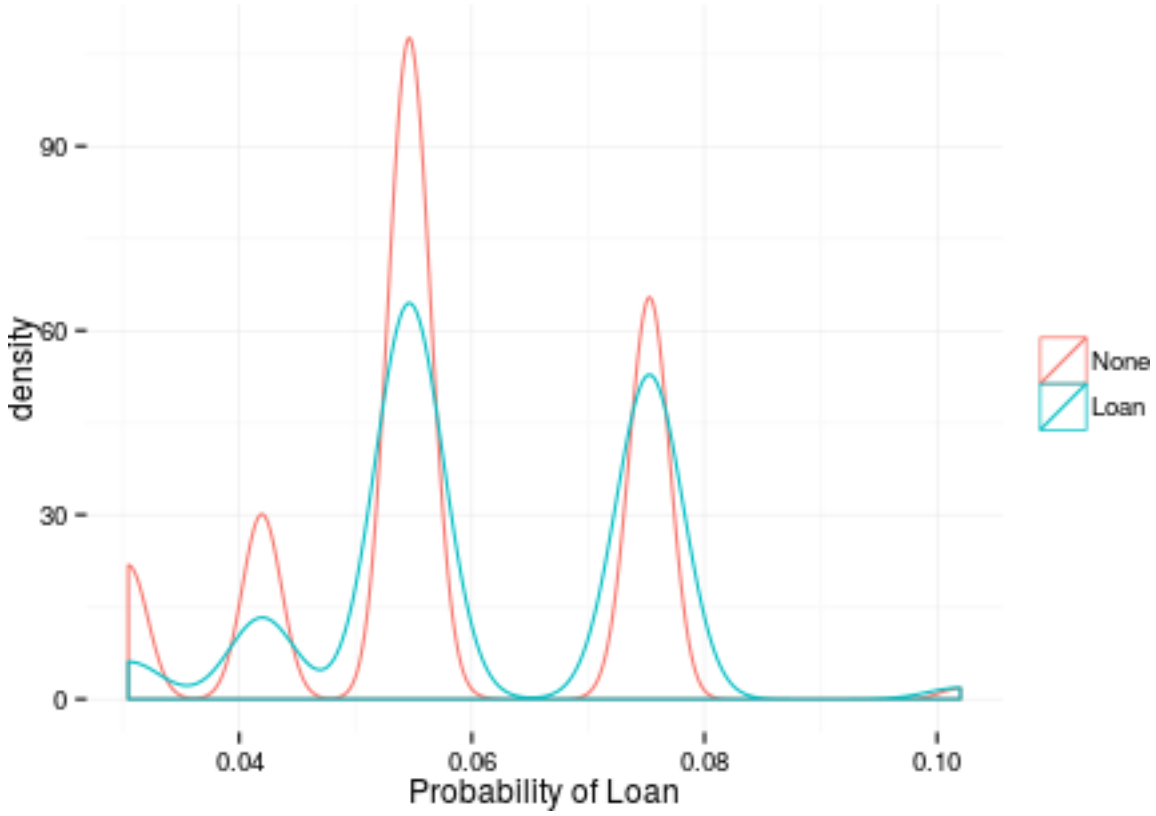
The stated objectives from RUS for disbursement of loans and grants were to target under-served rural communities with fewer than 20,000 inhabitants, as determined by the Census Bureau, and not located in a Standard Metropolitan Statistical Area. By using a logit specification, we can input the following covariates in attempt to predict whether a zip code will receive a loan/grant:

1. `I(Prov_alt < 2)TRUE` - an indicator variable for whether or not a zip code had fewer than 4 broadband providers as reported in December of 2000.
2. `I(SUMBLKPOP < 20000)TRUE` - an indicator variable for whether or not a zip code's population was fewer than 20,000 as of 2003 as reported by the Census Bureau. The year 2003 was chosen because that was the only available year I had for zip code level population statistics.
3. `rucadj` - an indicator variable for whether or not the zip code lies in a rural county that is adjacent to a metro county.
4. `rucnonadj` - an indicator variable for whether or not the zip code lies in a rural county that is not adjacent to a metro county.

The first set of results is via maximum likelihood methods of estimation and assumes homoskedasticity in the variance of the errors. The second set or results relaxes the homoskedasticity assumption to allow for heteroskedasticity in the errors.

Ignore the `Coefficients (binomial model with logit link):` and focus on `Latent scale model coefficients (with log link):`

```
##
## Call:
## glm(formula = iloans ~ I(Prov_alt < 2) + I(SUMBLKPOP < 20000) +
##     ruc, family = binomial(link = "logit"), data = data, subset = time ==
##     "2000-12-31")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4639  -0.3916  -0.3336  -0.2927   2.6425
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.12921    0.07234 -43.257  < 2e-16 ***
## I(Prov_alt < 2)TRUE        0.59939    0.08857   6.767 1.31e-11 ***
## I(SUMBLKPOP < 20000)TRUE -0.33130    0.08345  -3.970 7.19e-05 ***
## rucadj                     0.35480    0.06111   5.806 6.40e-09 ***
## rucnonadj                  0.02577    0.07238   0.356    0.722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13008  on 29637  degrees of freedom
## Residual deviance: 12892  on 29633  degrees of freedom
## AIC: 12902
##
## Number of Fisher Scoring iterations: 6
```




```
##
## Call:
## hetglm(formula = iloans ~ I(Prov_alt < 2) + I(SUMBLKPOP < 20000) +
##     ruc, data = data, subset = time == "2000-12-31", family = binomial(link = "logit"))
##
## Deviance residuals:
##     Min      1Q  Median      3Q     Max
## -0.6336 -0.3693 -0.3382 -0.2907  2.6579
##
## Coefficients (binomial model with logit link):
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.14303    0.08301 -37.863  < 2e-16 ***
## I(Prov_alt < 2)TRUE       -0.60210    0.50326  -1.196   0.2315
## I(SUMBLKPOP < 20000)TRUE -1.08076    1.20444  -0.897   0.3696
## rucadj                     1.18382    0.70976   1.668   0.0953 .
## rucnonadj                  2.39364    0.55257   4.332 1.48e-05 ***
##
## Latent scale model coefficients (with log link):
##                           Estimate Std. Error z value Pr(>|z|)
## I(Prov_alt < 2)TRUE        0.3455     0.1138   3.035  0.00241 **
## I(SUMBLKPOP < 20000)TRUE   0.1872     0.2710   0.691  0.48971
## rucadj                    -0.1612     0.2586  -0.624  0.53290
## rucnonadj                 -0.6965     0.4744  -1.468  0.14208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-likelihood: -6440 on 9 Df
## LR test for homoskedasticity: 11.94 on 4 Df, p-value: 0.01781
```

```
## Dispersion: 1
## Number of iterations in nlminb optimization: 21
```

Both results indicate that having fewer than 4 broadband providers increased the probability of being awarded a loan through either the Pilot or Farm Bill programs. We also see that rural counties adjacent to metro counties are more likely to receive loans/grants. The interesting, and noted problem in the Audit, is that population of zip code fewer than 20,000 reduces the probability of receiving a loan. This result is significant under the assumption of homoskedasticity, but when this is relaxed it is no longer significant.

Checking the fitted values, we see that there are limited predicted zip codes that have a probability of receiving a loan/grant in excess of 10%. This is likely because the model is too simple.

Although this particular model captures the *stated* criteria to obtain a loan/grant, it does a poor job capturing what the factors that actually affected disbursement.

More Accurate Model

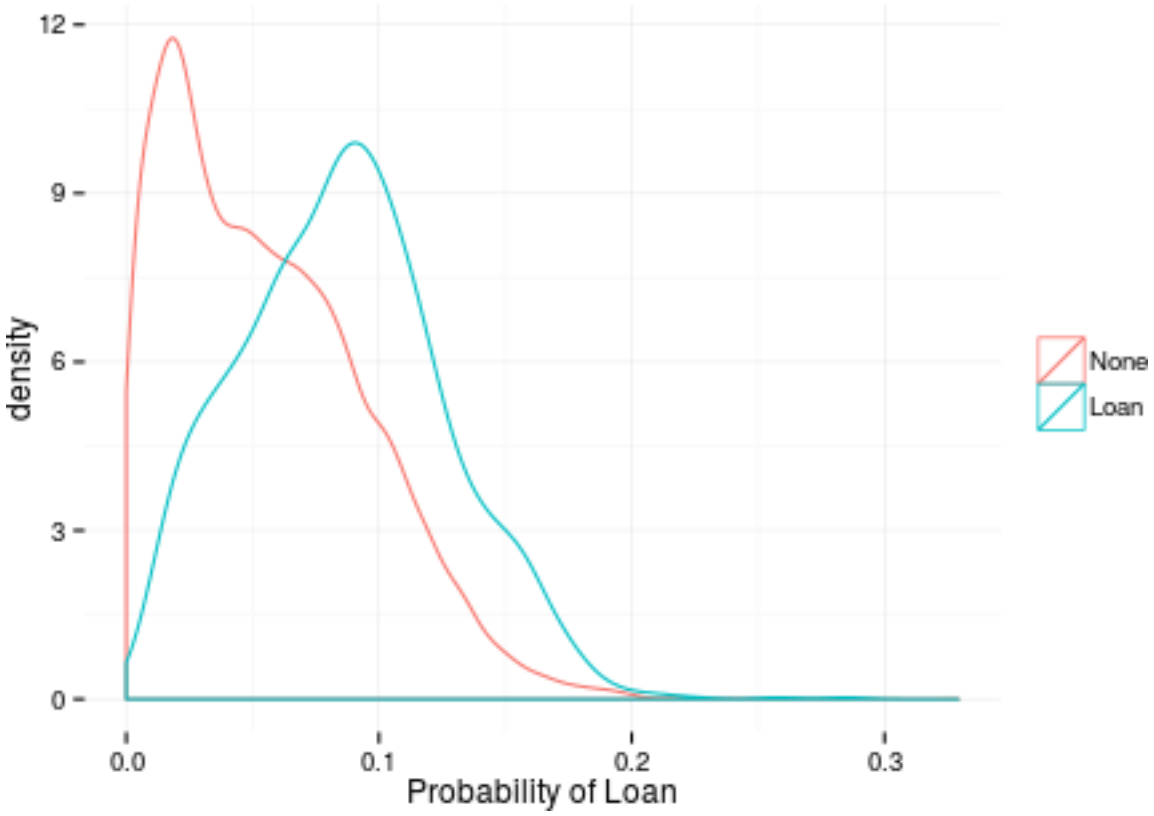
In order to better model the probability of receiving a loan, I include variables that have been associated with broadband availability:

1. `Prov_alt` - this is a count of number of broadband providers. This variable is slightly different than how I modeled it in the past, here I make the suppressed 1-3 count equal to 1 and subtract 2 from all values greater than 4.
2. `log(SUMBLKPOP + 1)` - this is a log of population in 2003. The one is because there are a few zip codes with no population and log(0) is undefined.
3. `log(est)` - log of the number of establishments in a zip code as a way to proxy for the demand of broadband in a zip code. More establishments should indicate a greater incentive for the community to apply for a loan.

4. `logINC` - log of mean income at the county level, data are from IRS.
5. `tri` - an index of topographical ruggedness that ranges from 0 (not rugged) to 100 (very rugged).

These variables add in more variation across zip codes as the previous model only had dummy variables. I still maintain the previous dummy variables as a way to verify whether or not the programs reached their intended goals. I eschew the homoskedastic model and only report for the relaxed heteroskedastic model:

```
##
## Call:
## hetglm(formula = iloans ~ I(Prov_alt < 2) + Prov_alt + log(SUMBLKPOP +
##     1) + I(SUMBLKPOP < 20000) + ruc + log(est) + logINC + tri, data = data,
##     subset = time == "2000-12-31", family = binomial(link = "logit"))
##
## Deviance residuals:
##     Min      1Q  Median      3Q     Max
## -0.8938 -0.4086 -0.3075 -0.1937  3.4869
##
## Coefficients (binomial model with logit link):
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -5.41259   40.22616  -0.135    0.893
## I(Prov_alt < 2)TRUE       -18.11980   19.71613  -0.919    0.358
## Prov_alt                    1.56292    1.76983   0.883    0.377
## log(SUMBLKPOP + 1)          3.10998    3.22983   0.963    0.336
## I(SUMBLKPOP < 20000)TRUE    4.76872    5.65981   0.843    0.399
## rucadj                     11.26398   11.41848   0.986    0.324
## rucnonadj                  18.42886   18.54027   0.994    0.320
## log(est)                   -0.06251    1.26062  -0.050    0.960
## logINC                     -4.72283    8.60484  -0.549    0.583
## tri                        -2.59688    2.69862  -0.962    0.336
##
## Latent scale model coefficients (with log link):
##                             Estimate Std. Error z value Pr(>|z|)
## I(Prov_alt < 2)TRUE        0.6202163  0.1140774   5.437 5.42e-08 ***
## Prov_alt                  -0.0934419  0.0226360  -4.128 3.66e-05 ***
## log(SUMBLKPOP + 1)        -0.0218333  0.0255821  -0.853  0.39340
## I(SUMBLKPOP < 20000)TRUE -0.0696923  0.0786805  -0.886  0.37574
## rucadj                     0.0279057  0.0638661   0.437  0.66215
## rucnonadj                 -0.2497792  0.0775545  -3.221  0.00128 **
## log(est)                   0.0409470  0.0260271   1.573  0.11566
## logINC                     0.2377665  0.0949211   2.505  0.01225 *
## tri                        0.0144322  0.0009375  15.394  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-likelihood: -6056 on 19 Df
## LR test for homoskedasticity: 234.6 on 9 Df, p-value: < 2.2e-16
## Dispersion: 1
## Number of iterations in nlminb optimization: 77
```



Ah, much better.

A likelihood ratio test of homoskedasticity strongly rejects the null, which is evidence in favor of the chosen model. Further, we see that number of broadband providers is a negatively related variable to the probability of receiving a loan/grant. This makes intuitive sense that areas with already high number of broadband providers are not likely to receive (or even apply for) these loans/grants as there is existing competition in that community.

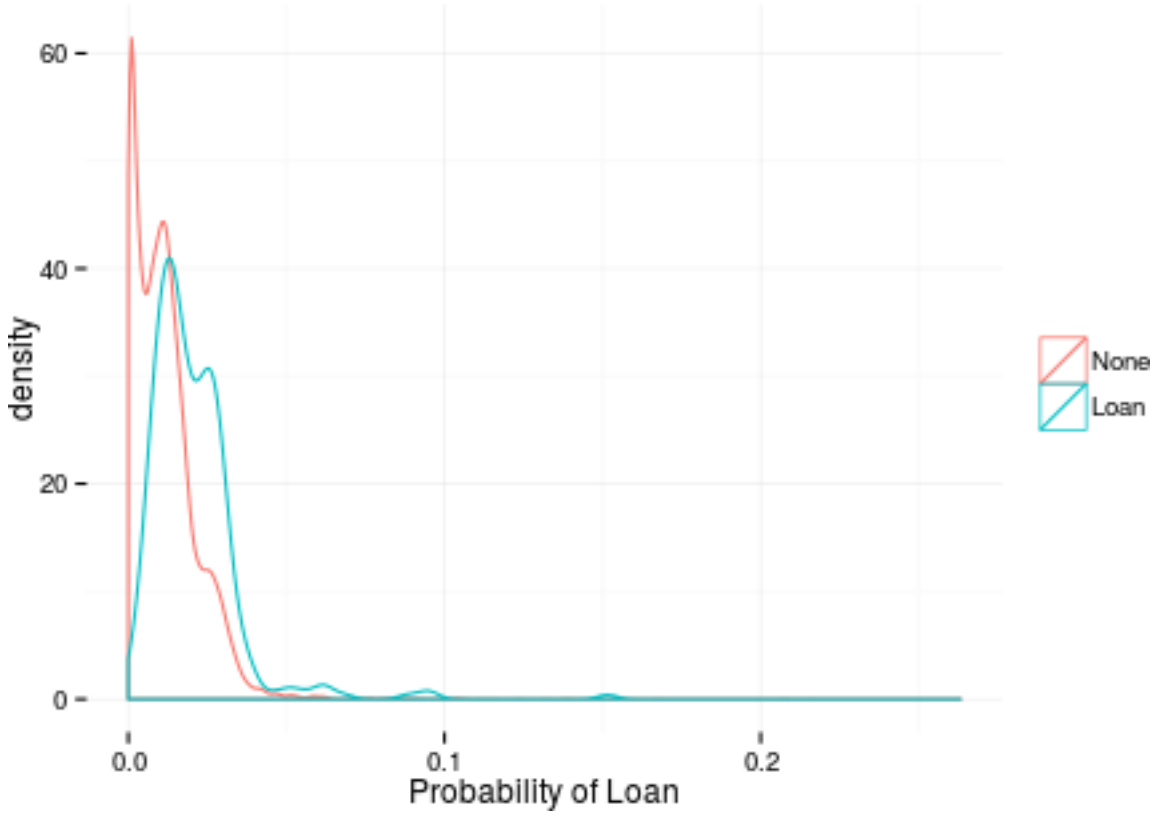
A takeaway from this regression is that it does appear that there is some evidence that the loan/grant programs targeted under-served areas and for smaller communities (although the coefficient associated with the threshold of 20,000 is the wrong sign, it is not statistically different from 0). One worrisome finding is that the rural non-adjacent areas appear to have a significantly negative association with receiving these loans/grants. Interpretation here is difficult because we only know communities which **received** loans/grants and not those which *applied for* loans/grants. If the rural non-adjacent areas are also associated with areas which were less likely to apply for grants, then the coefficient associated rural non-adjacent would be biased downward. Controlling for the probability of applying for a loan may render this finding insignificant or even change the sign on the coefficient.

Pilot Program

As a way to further disaggregate the data and focus on the specific programs, we can do the above analysis but only look at communities with Pilot loans/grants:

```
##
## Call:
## hetglm(formula = ipilot ~ I(Prov_alt < 2) + Prov_alt + log(SUMBLKPOP +
##     1) + I(SUMBLKPOP < 20000) + ruc + log(est) + logINC + tri, data = data,
##     subset = time == "2000-12-31", family = binomial(link = "logit"))
```

```
##
## Deviance residuals:
##     Min      1Q  Median      3Q     Max
## -0.7816 -0.1806 -0.1409 -0.0830  3.7915
##
## Coefficients (binomial model with logit link):
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              17839.83   32333.12   0.552    0.581
## I(Prov_alt < 2)TRUE       4970.87    7447.24   0.667    0.504
## Prov_alt                   617.38     928.88   0.665    0.506
## log(SUMBLKPOP + 1)          43.89     138.95   0.316    0.752
## I(SUMBLKPOP < 20000)TRUE  3029.35    5962.28   0.508    0.611
## rucadj                    -412.33     916.27  -0.450    0.653
## rucnonadj                  257.50     493.65   0.522    0.602
## log(est)                  -647.42     896.24  -0.722    0.470
## logINC                   -2700.57    4385.24  -0.616    0.538
## tri                        -52.90      77.54  -0.682    0.495
##
## Latent scale model coefficients (with log link):
##                           Estimate Std. Error z value Pr(>|z|)
## I(Prov_alt < 2)TRUE      -0.550120   0.458596  -1.200    0.230
## Prov_alt                 -0.186553   0.037840  -4.930 8.22e-07 ***
## log(SUMBLKPOP + 1)       -0.017666   0.037002  -0.477    0.633
## I(SUMBLKPOP < 20000)TRUE -0.324344   0.377390  -0.859    0.390
## rucadj                    0.198054   0.112765   1.756    0.079 .
## rucnonadj                 0.175625   0.119789   1.466    0.143
## log(est)                  0.173553   0.034648   5.009 5.47e-07 ***
## logINC                    0.701211   0.129295   5.423 5.85e-08 ***
## tri                      -0.001091   0.002965  -0.368    0.713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-likelihood: -1758 on 19 Df
## LR test for homoskedasticity: 54.13 on 9 Df, p-value: 1.784e-08
## Dispersion: 1
## Number of iterations in nlminb optimization: 139
```



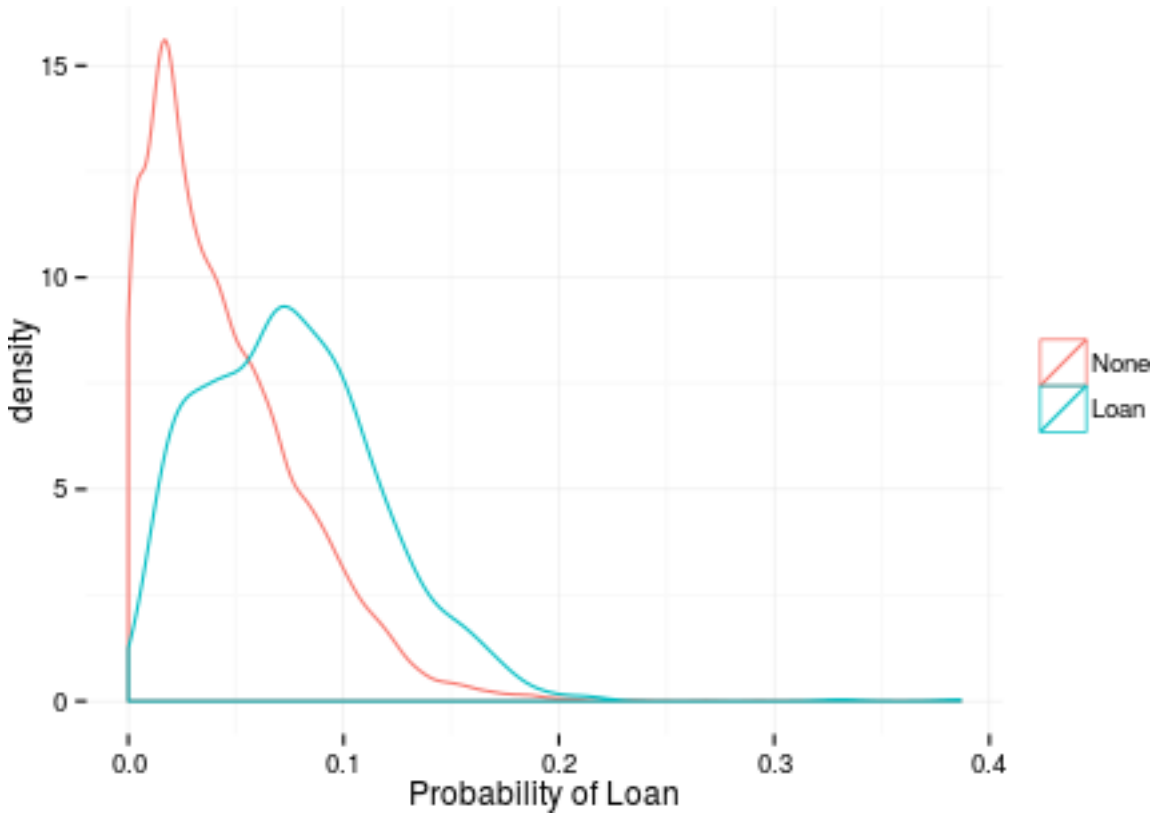
There appears to be about the same qualitative results as above, although the fit of the model indicates that very few zip codes had even a 10% chance at receiving these loans. This does jive with what one would expect. The Pilot Program was relatively small with approximately 30 loans/grants disbursed and the potential pool of communities that match the description is vast across the United States (~30,000). So the results make sense.

Farm Bill Program

We can perform the same analysis above, but for the Farm Bill loans/grants:

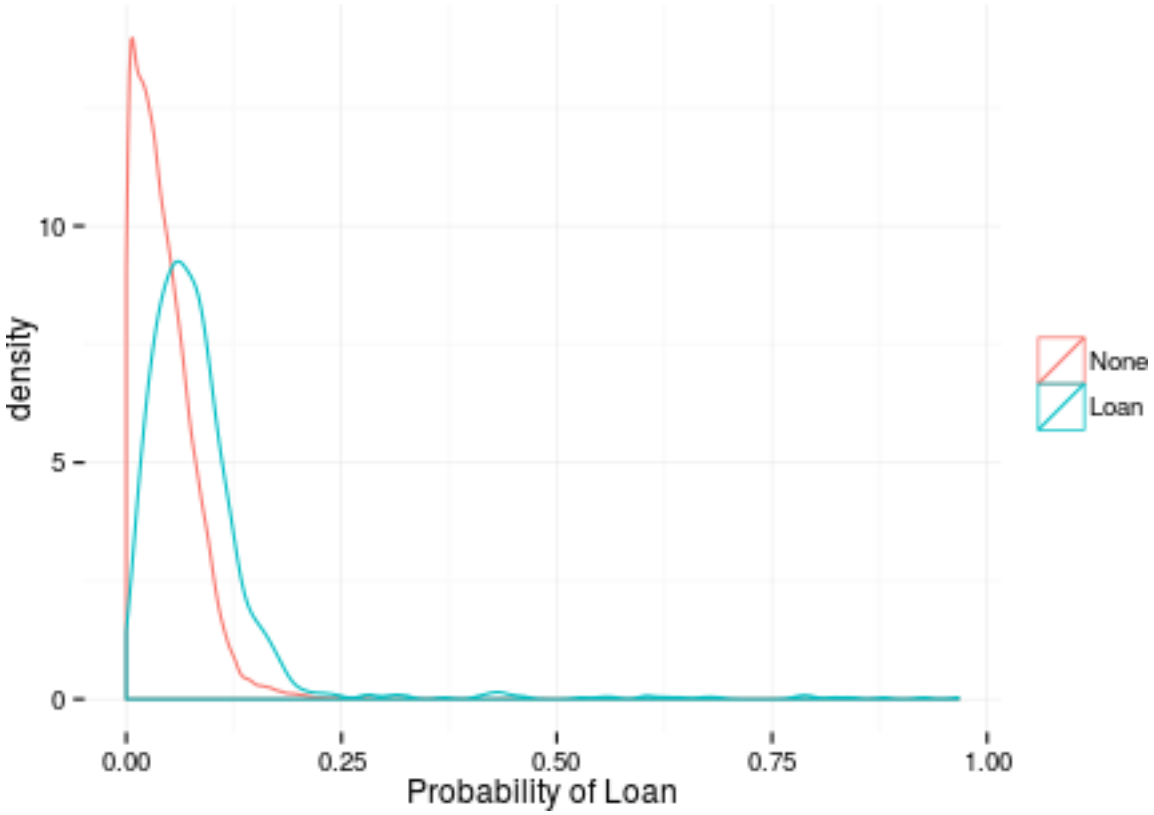
```
##
## Call:
## hetglm(formula = ibip1234 ~ I(Prov_alt < 2) + Prov_alt + log(SUMBLKPOP +
##     1) + I(SUMBLKPOP < 20000) + ruc + log(est) + logINC + tri, data = data,
##     subset = time == "2000-12-31", family = binomial(link = "logit"))
##
## Deviance residuals:
##     Min      1Q  Median      3Q     Max
## -0.9819 -0.3623 -0.2632 -0.1713  3.5905
##
## Coefficients (binomial model with logit link):
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -8.0435    35.3183  -0.228    0.820
## I(Prov_alt < 2)TRUE      -13.9205    15.4583  -0.901    0.368
## Prov_alt                   1.3599     1.5799   0.861    0.389
## log(SUMBLKPOP + 1)         4.4325     4.5892   0.966    0.334
## I(SUMBLKPOP < 20000)TRUE   8.3522     9.1301   0.915    0.360
```

```
## rucadj                      9.2980     9.5941   0.969    0.332
## rucnonadj                  20.1077    20.5910   0.977    0.329
## log(est)                    0.6657     1.4496   0.459    0.646
## logINC                     -6.0939     9.8663  -0.618    0.537
## tri                        -2.5436     2.6942  -0.944    0.345
##
## Latent scale model coefficients (with log link):
##                              Estimate Std. Error z value Pr(>|z|)
## I(Prov_alt < 2)TRUE        0.6018610  0.1074986   5.599 2.16e-08 ***
## Prov_alt                  -0.0726239  0.0227944  -3.186  0.00144 **
## log(SUMBLKPOP + 1)        -0.0427905  0.0248033  -1.725  0.08449 .
## I(SUMBLKPOP < 20000)TRUE -0.1445469  0.0784068  -1.844  0.06525 .
## rucadj                     0.0458948  0.0627591   0.731  0.46461
## rucnonadj                 -0.4562834  0.0773399  -5.900 3.64e-09 ***
## log(est)                   0.0270395  0.0255818   1.057  0.29052
## logINC                     0.2505371  0.0967572   2.589  0.00962 **
## tri                        0.0159330  0.0009365  17.012  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-likelihood: -5088 on 19 Df
## LR test for homoskedasticity: 279.7 on 9 Df, p-value: < 2.2e-16
## Dispersion: 1
## Number of iterations in nlminb optimization: 79
```



Because the Farm Bill did not come into effect until 2002, it may be more appropriate to use the values of number of broadband providers from 2002 in the regression instead:

```
##
## Call:
## hetglm(formula = ibip1234 ~ I(Prov_alt < 2) + Prov_alt + log(SUMBLKPOP +
##     1) + I(SUMBLKPOP < 20000) + ruc + log(est) + logINC + tri, data = data,
##     subset = time == "2002-12-31", family = binomial(link = "logit"))
##
## Deviance residuals:
##     Min      1Q  Median      3Q     Max
## -2.0304 -0.3569 -0.2651 -0.1674  3.6193
##
## Coefficients (binomial model with logit link):
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -0.4074    14.4538  -0.028    0.978
## I(Prov_alt < 2)TRUE       -7.1077     7.4428  -0.955    0.340
## Prov_alt                   1.3250     1.3454   0.985    0.325
## log(SUMBLKPOP + 1)         1.3284     1.3625   0.975    0.330
## I(SUMBLKPOP < 20000)TRUE   2.2515     2.3347   0.964    0.335
## rucadj                     4.1911     4.1101   1.020    0.308
## rucnonadj                  8.0697     7.8862   1.023    0.306
## log(est)                  -1.8558     1.9059  -0.974    0.330
## logINC                    -1.7280     2.9532  -0.585    0.558
## tri                       -0.9703     0.9803  -0.990    0.322
##
## Latent scale model coefficients (with log link):
##                            Estimate Std. Error z value Pr(>|z|)
## I(Prov_alt < 2)TRUE       0.3404213  0.0720828   4.723 2.33e-06 ***
## Prov_alt                 -0.1544668  0.0080536 -19.180  < 2e-16 ***
## log(SUMBLKPOP + 1)       -0.0396294  0.0210011  -1.887   0.0592 .
## I(SUMBLKPOP < 20000)TRUE -0.0868361  0.0545807  -1.591   0.1116
## rucadj                    0.0288573  0.0601810   0.480   0.6316
## rucnonadj                -0.4531036  0.0734290  -6.171 6.80e-10 ***
## log(est)                  0.1586255  0.0221377   7.165 7.76e-13 ***
## logINC                    0.1559924  0.0934106   1.670   0.0949 .
## tri                       0.0156121  0.0009296  16.794  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-likelihood: -5039 on 19 Df
## LR test for homoskedasticity: 430.1 on 9 Df, p-value: < 2.2e-16
## Dispersion: 1
## Number of iterations in nlminb optimization: 51
```



There is a marked improvement in the Farm Bill's disbursement of loans to smaller communities, but aside from that we see qualitatively the same results. It is still puzzling to see that rural non-adjacent communities are less likely to receive a loan. If only we had applicant data ...