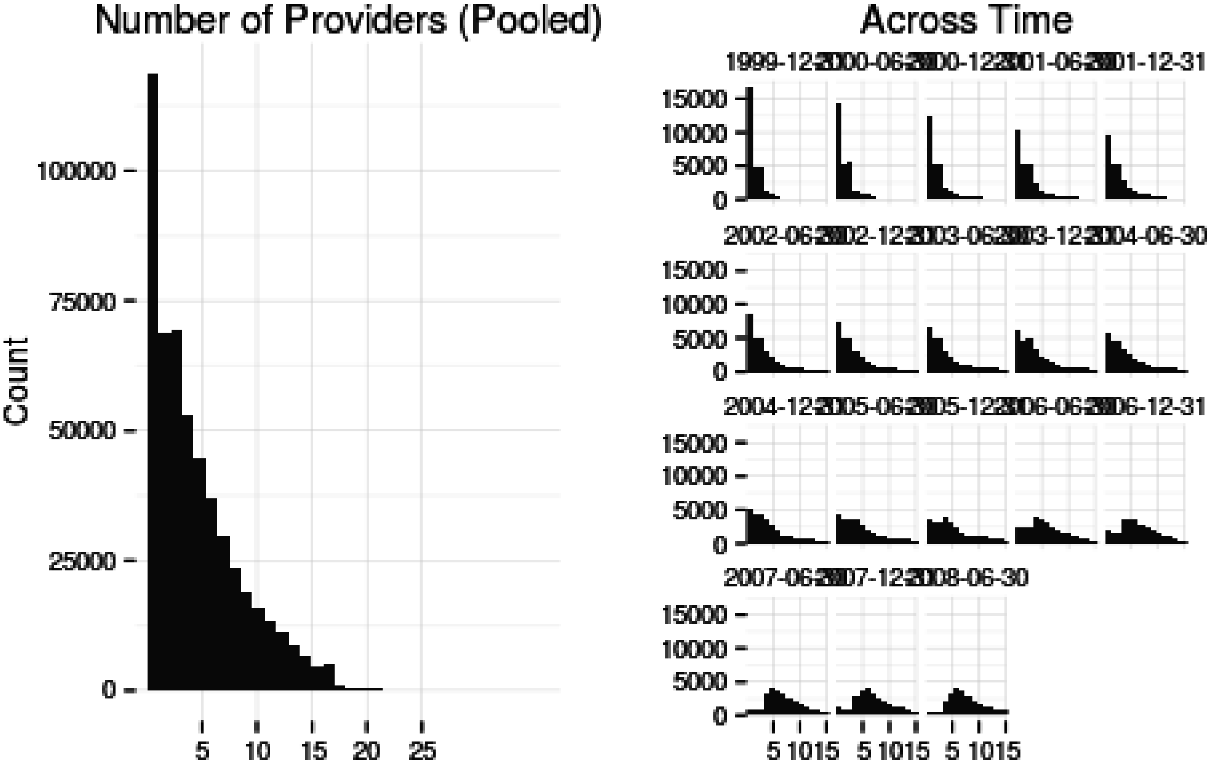
3-USDA Evaluation Basic

Robert Dinterman

2015-10-22

Basic Panel Regressions for Broadband Availability

The following regressions make use of the `plm` package in `R` in order to identify the relationship between broadband availability and the USDA Broadband Loan Program. Loosely, we wish to model the number of broadband providers by zip code across the years 1999 to 2008 and determine whether or not the USDA Broadband Loan Program had an impact on broadband availability as this was one of the intended benefits of the subsidized loans. As a quick reference, here is what the dependent variable looks like across time:



Data on broadband providers is measured twice a year (June 30 and December 31), ~~therefore the two values are averaged for the yearly value~~. Further, the variable takes on a count value of 0, 1-3*, 4, 5, 6, ... 31. The value 1-3* is a suppressed value of broadband providers for confidentiality purposes and has been coded as 2 to be consistent with the literature.

Other variables used include:

- `iloans` - an indicator variable for whether or not a zip code received a loan. No loans were given before 2002, so this variable does vary by time.
- `log(est)` - this is from zip code business pattern data and is the number of establishments in a particular zip code. I take the log of this variable because the distribution is right-skewed. It is possible to substitute this variable with number of employees, annual payroll, or first quarter payroll from the ZBP but I choose not to because those variables are **suppressed for approximately 9% of the zip codes**. Establishments is highly correlated with the other variables anyway, so I would rather use a less precise proxy than potentially bias the sample.

- `log(Pop_IRS)` - IRS has data on number of tax returns filed by county from 1989 until 2013. I use the number of exemptions per county as a way to proxy for the population of a county. This variable is also right-skewed and therefore the log of population is taken instead of population. The alternative for population would be to use US Census data which produce yearly estimates at the county level. These estimates are based off of the 2000 Census and use the demographic age distribution of a county in order to project forward the birth rate and death rate to determine what the population in a county should be. Since this is simply a function of initial conditions in 2000, I choose to use IRS data because there is more variation in the data and it reflects changes in economic conditions across counties that would drive migration (population change).
- `logINC` - this is tabulated from the same IRS data above using Adjusted Gross Income (AGI) at the county level for each year. This is divided by the number of households for a county and is therefore a proxy for mean income, as reported to the IRS, per year. Again, this is a right-skewed variable which is the justification for taking the logarithm of the variable.
- `tri` - stands for Terrain Ruggedness Index which uses elevation data for a given polygon to calculate the feature changes in a given area relative to the entire domain. This is at the ZCTA level across the United States and is thought of as a proxy for increased costs of broadband deployment due to rough terrain. This does not vary across years and so zip code fixed effects will take away this variable.
- `ruc` - the rural-urban continuum code, but for this study I simply use 3 classifications of a county: Metro, Rural but adjacent to a metro county, and Rural but non-adjacent to a metro county. Counties do change across time, but only in years that end in 3 (1993, 2003, . . .). I choose to use the values for 2003 as this would be a little bit before the halfway point in the analysis.
- `AREA_zcta` - through the zip code shapefile, the total square miles of the particular zip code is calculated through GIS software in R. Holding all else constant, it should be the case that larger zip codes have more broadband providers because they span a greater area which allows a different broadband provider to potentially serve the area. It is not the case that zip codes are randomly drawn, they are designed by the postal service and thus their goal is to efficiently serve the United States. This leads to zip code area being negatively correlated with population (value of -0.0709). Because of this, one needs to further control for density of people and firms.
- `I(Pop_IRS / AREA_cty)` - using the population data from the IRS, which is at the county level, I can proxy density at the county level by dividing by the square miles of the county. A denser county will attract more broadband providers as they have a larger customer base. The area is calculated through GIS software in R by using a county level shapefile.
- `I(est / AREA_zcta)` - to further proxy for density of firms, the yearly value of establishments per zip code is divided by the area of the zip code.

Static Models

I start with a naive relationship to model number of broadband providers in a Static Panel framework as follows:

$$Prov_{z,t} = \mu_z + \tau_t + \beta_1 Loan_{z,t} + \beta_2 X_{z,t} + \varepsilon_{z,t} + \alpha_z + \alpha_t$$

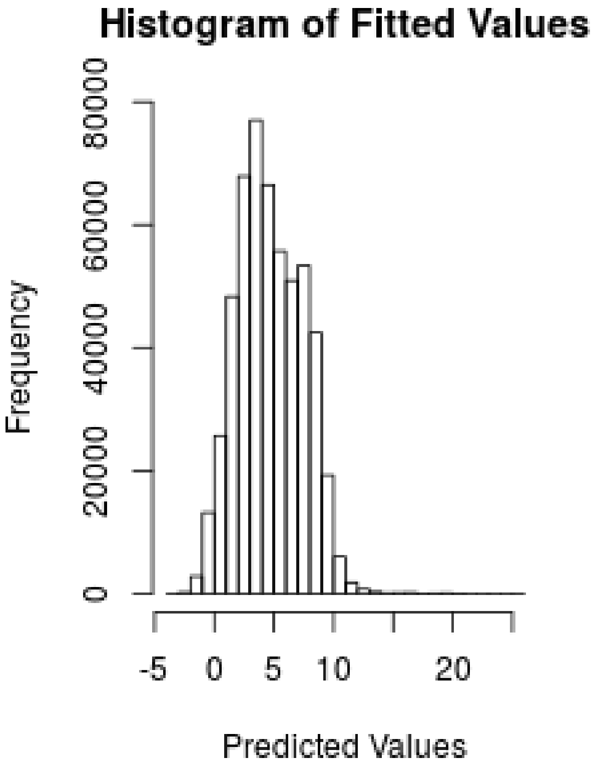
The variable $Prov_{z,t}$ is the number of providers in zip code z at time t. The μ_z parameter is a zip code level fixed effect that may or may not be present in the particular model; τ_t is a year fixed effect that may or may not be present; α_z is a zip code random effect that may or may not be present in the model; α_t is a year random effect that may or may not be present in the model; and the particular variable of interest is $Loan_{z,t}$ which is a dummy variable indicating if a zip code has been awarded (and assumed deployed) a subsidized loan from the USDA.

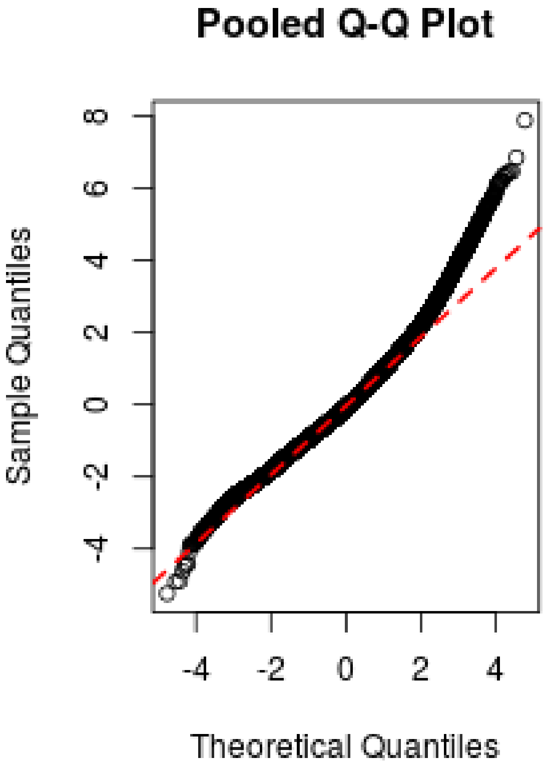
Pooled Regression

Assumes no fixed or random effects, effectively treating the regression as if the entire sample were taken from the same population:

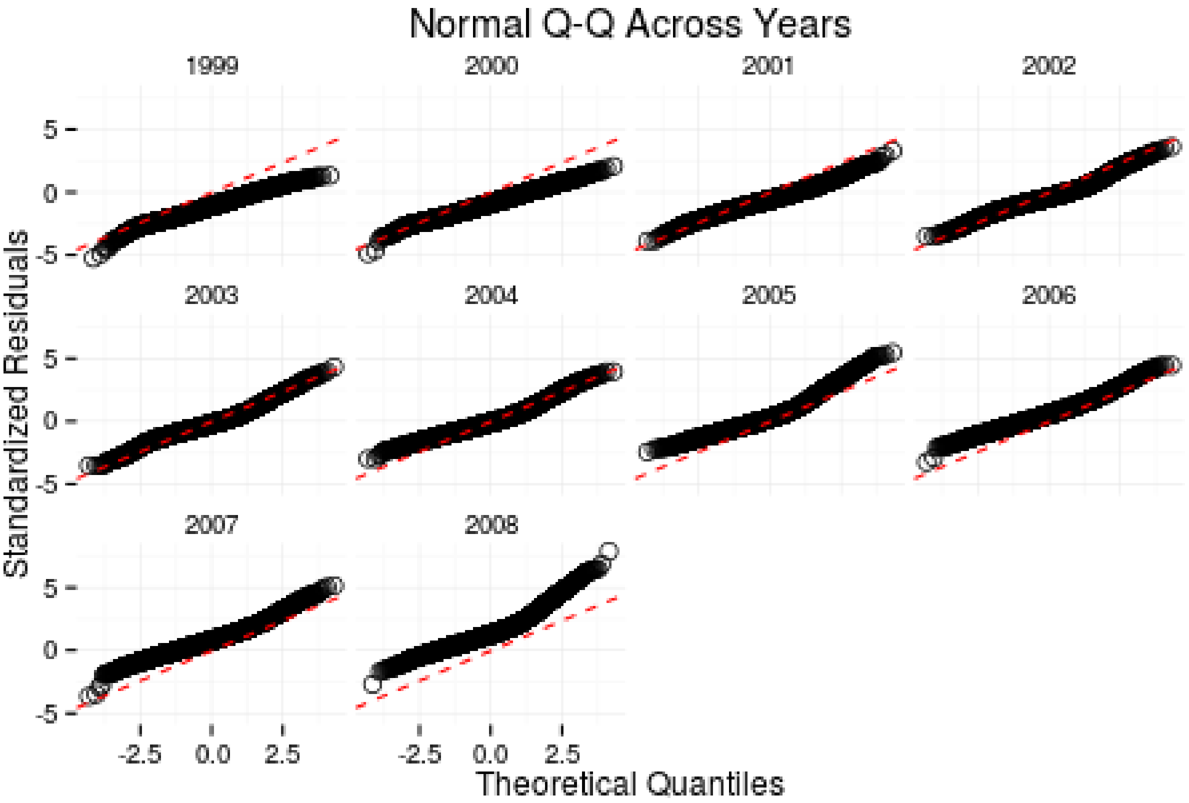
```
## Oneway (individual) effect Pooling Model
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta),
##     data = pdata, model = "pooling")
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -14.60   -1.88   -0.24    1.71   22.00
##
## Coefficients :
##                        Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)         -1.7526e+00  1.9393e-01  -9.0372 < 2.2e-16 ***
## iloans               1.7420e+00  2.4609e-02  70.7860 < 2.2e-16 ***
## log(est)             1.0477e+00  2.6355e-03 397.5350 < 2.2e-16 ***
## log(Pop_IRS)         5.5668e-01  3.7594e-03 148.0766 < 2.2e-16 ***
## logINC              -3.9560e-01  1.9200e-02 -20.6043 < 2.2e-16 ***
## tri                 -4.4120e-03  1.9532e-04 -22.5886 < 2.2e-16 ***
## rucadj              -8.2925e-02  1.1580e-02  -7.1614 7.998e-13 ***
## rucnonadj            2.4798e-01  1.3725e-02  18.0675 < 2.2e-16 ***
## poly(AREA_zcta, 2)1 -2.7003e+01  2.9681e+00  -9.0978 < 2.2e-16 ***
## poly(AREA_zcta, 2)2  1.9899e+01  2.8973e+00   6.8681 6.513e-12 ***
## I(Pop_IRS/AREA_cty)  9.7887e-05  1.9579e-06  49.9970 < 2.2e-16 ***
## I(est/AREA_zcta)     3.1664e-04  9.1809e-06  34.4892 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    8058700
## Residual Sum of Squares: 4127800
## R-Squared      :  0.48778
##       Adj. R-Squared :  0.48777
## F-statistic: 46105.6 on 11 and 532572 DF, p-value: < 2.22e-16
```

There are no controls for time or zip code effects and this gives the result that the Broadband Loan Programs were effective in adding 1.74 broadband providers per zip code across 1999 to 2008. However, if we turn to a few casual diagnostics we can see that a pooled model is not appropriate:





The fit of the model is poor as we can see that the distribution of fitted values appears to have a normal distribution to it even though the overall distribution of broadband providers is clearly right-skewed and contains no negative values. Also, there appears to be divergence from the normality assumption of the errors by referencing the Quantile-Quantile plot. If the distribution of residuals was normal, then the residuals should be neatly ordered around the red line. There is a fair disturbance form this, and we can stratify the residuals by year to see this better:



Fixed Effects Regressions

Three fixed effects models:

1. **Zip Code Fixed Effects:** this implies that `tri` and `ruc` cannot be estimated due to their time-invariant nature.
   - μ_z is present in the model.
2. **Year Fixed Effects:** this implies that there are separate time dummy variables for 2000 to 2008 (1999 is omitted). The effects of `tri` and `ruc` can be estimated.
   - τ_t is present in the model.
3. **Zip Code and Year Fixed Effects:** this is a combination of the two and therefore `tri` and `ruc` cannot be estimated.
   - μ_z and τ_t are present in the model.
4. **First Difference:** takes the difference between time t and $t-1$, which will not allow for `tri` or `ruc` to be estimated. An advantage to this model is that serial correlation should be reduced, although it is not guaranteed to be eliminated.

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     I(Pop_IRS/AREA_cty) + I(est/AREA_zcta), data = pdata, model = "within")
##
## Balanced Panel: n=29588, T=18, N=532584
```

```
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -59.300  -1.300  -0.127   1.230  22.800
##
## Coefficients :
##                        Estimate  Std. Error   t-value Pr(>|t|)
## iloans               2.3511e+00  2.7498e-02   85.5010  < 2e-16 ***
## log(est)             1.5430e+00  1.7427e-02   88.5420  < 2e-16 ***
## log(Pop_IRS)         1.3369e+01  5.2973e-02  252.3627  < 2e-16 ***
## logINC              -6.5964e+00  5.4367e-02 -121.3300  < 2e-16 ***
## I(Pop_IRS/AREA_cty)  1.3582e-03  3.3809e-05   40.1734  < 2e-16 ***
## I(est/AREA_zcta)     1.4898e-04  6.2159e-05    2.3968  0.01654 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    3203100
## Residual Sum of Squares: 2496700
## R-Squared      :  0.22052
##       Adj. R-Squared :  0.20826
## F-statistic: 23716.3 on 6 and 502990 DF, p-value: < 2.22e-16
```

```
## Oneway (time) effect Within Model
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta),
##     data = pdata, effect = "time", model = "within")
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -12.300  -1.330  -0.153   1.130  17.800
##
## Coefficients :
##                        Estimate  Std. Error  t-value  Pr(>|t|)
## iloans               2.8764e-02  1.8301e-02   1.5717     0.116
## log(est)             1.0283e+00  1.9390e-03 530.3211 < 2.2e-16 ***
## log(Pop_IRS)         4.7530e-01  2.7722e-03 171.4541 < 2.2e-16 ***
## logINC               3.9516e-01  1.4304e-02  27.6254 < 2.2e-16 ***
## tri                 -5.7013e-03  1.4368e-04 -39.6802 < 2.2e-16 ***
## rucadj              -4.8502e-02  8.5196e-03  -5.6930 1.249e-08 ***
## rucnonadj            2.4993e-01  1.0097e-02  24.7523 < 2.2e-16 ***
## poly(AREA_zcta, 2)1 -2.3237e+01  2.1832e+00 -10.6434 < 2.2e-16 ***
## poly(AREA_zcta, 2)2  1.4828e+01  2.1311e+00   6.9579 3.458e-12 ***
## I(Pop_IRS/AREA_cty)  9.7240e-05  1.4401e-06  67.5225 < 2.2e-16 ***
## I(est/AREA_zcta)     3.1288e-04  6.7531e-06  46.3322 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6029100
## Residual Sum of Squares: 2233200
## R-Squared      :  0.62959
##       Adj. R-Squared :  0.62956
## F-statistic: 82291 on 11 and 532555 DF, p-value: < 2.22e-16
```

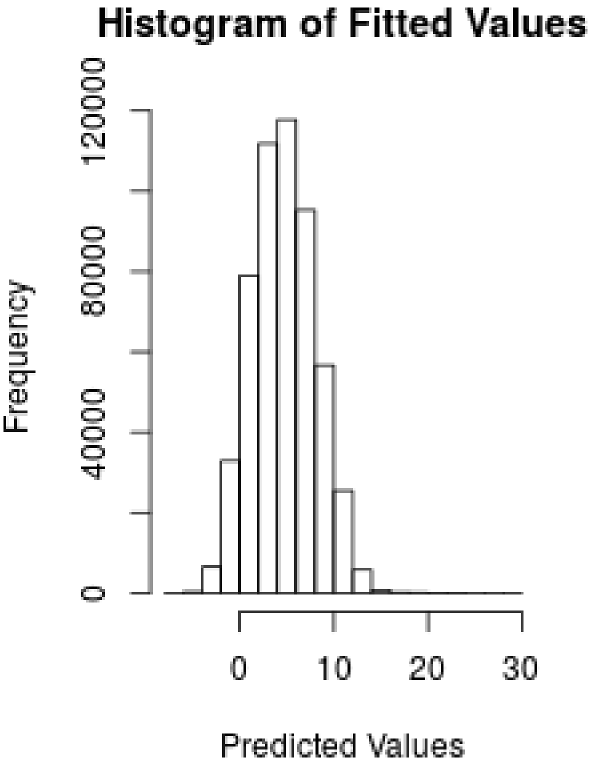
```
## Twoways effects Within Model
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     I(Pop_IRS/AREA_cty) + I(est/AREA_zcta), data = pdata, effect = "twoways",
##     model = "within")
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Residuals :
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -18.7000  -0.8800   0.0176   0.8900  12.7000
##
## Coefficients :
##                        Estimate  Std. Error  t-value  Pr(>|t|)
## iloans               9.6936e-02 1.8885e-02   5.1330 2.853e-07 ***
## log(est)             6.7343e-01 1.1891e-02  56.6353 < 2.2e-16 ***
## log(Pop_IRS)         2.9751e+00 4.1862e-02  71.0690 < 2.2e-16 ***
## logINC              -2.1312e+00 4.2181e-02 -50.5253 < 2.2e-16 ***
## I(Pop_IRS/AREA_cty)  2.8196e-04 2.2973e-05  12.2739 < 2.2e-16 ***
## I(est/AREA_zcta)     4.1301e-05 4.2094e-05   0.9812    0.3265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1173500
## Residual Sum of Squares: 1144900
## R-Squared      :  0.024301
##       Adj. R-Squared :  0.02295
## F-statistic: 2087.89 on 6 and 502973 DF, p-value: < 2.22e-16
```

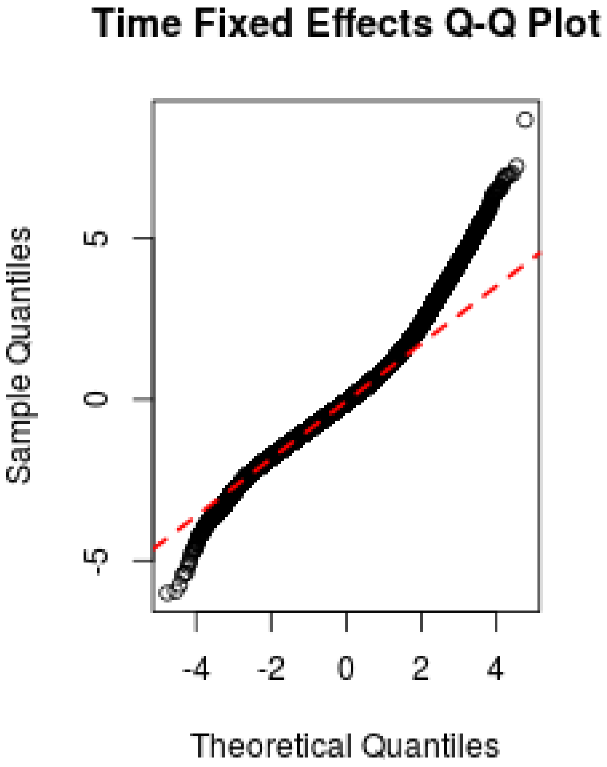
```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     I(Pop_IRS/AREA_cty) + I(est/AREA_zcta), data = pdata, model = "fd")
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##   -9.36   -0.43   -0.43    0.57   14.60
##
## Coefficients :
##                        Estimate  Std. Error  t-value  Pr(>|t|)
## (intercept)          4.2990e-01 1.7618e-03 244.0206 < 2.2e-16 ***
## iloans              -1.4948e-01 3.0126e-02  -4.9618 6.986e-07 ***
## log(est)             8.8593e-02 1.5163e-02   5.8425 5.145e-09 ***
## log(Pop_IRS)        -2.1550e+00 5.3852e-02 -40.0172 < 2.2e-16 ***
## logINC              -4.5493e-01 3.6473e-02 -12.4730 < 2.2e-16 ***
## I(Pop_IRS/AREA_cty) -4.4296e-04 3.0692e-05 -14.4325 < 2.2e-16 ***
## I(est/AREA_zcta)     3.9595e-05 5.4139e-05   0.7314    0.4646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    775730
## Residual Sum of Squares: 772420
## R-Squared      :  0.0042587
##       Adj. R-Squared :  0.0042586
## F-statistic: 358.538 on 6 and 502989 DF, p-value: < 2.22e-16
```

Standard errors are not robust in this setting, which likely implies that they are too small for proper testing inference. For the purposes of our interest, the significance of the loan β_1, is significantly positive for the first and third models but not significant for the second and fourth. The first, third, and fourth cannot identify the effects of zip code terrain on broadband deployment as well as the potential urban versus rural divide.

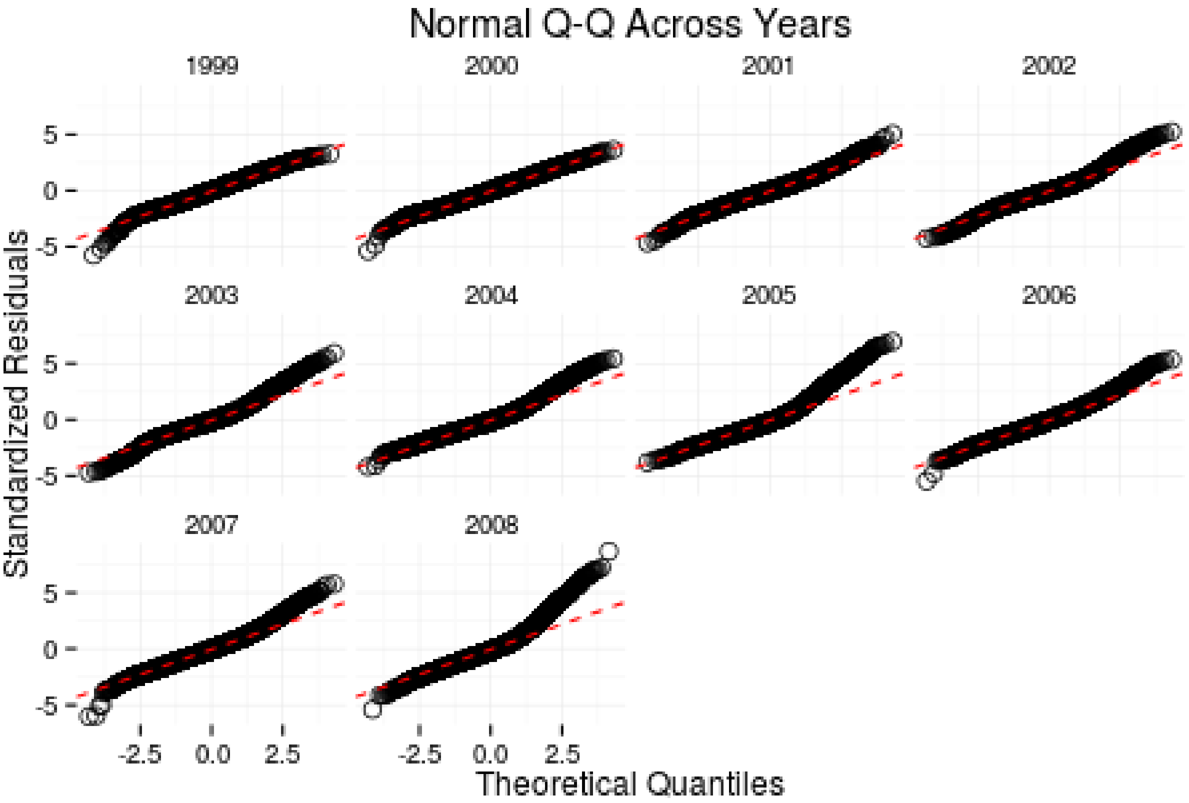
The first, third, and fourth models are likely incorrect for any inference that we would want to pursue. A big reason for this is because of the consistently negative associated coefficient with income. From an intuitive economic perspective, this does not make sense that areas with higher income would tend to have less broadband access. The logical conclusion of this would be that poorer areas have more access to broadband and common sense dictates this is not true (and the literature tends to find positive association with income and broadband access).

The second model is appealing because we see the expected signs across establishments, population, income, and terrain. It is slightly puzzling that rural non-adjacent counties tend to have more broadband access than urban and rural-adjacent counties. As a cross check, we can look at the time fixed effects model's fitted values and its Q-Q plot to verify whether or not the model has some bite to it:





We observe that the fitted values appear more normal than in the pooled ols model with the same problem of negative predicted values. Further, divergence from normality appears to still be a problem although the negative distribution of residuals looks to be better. To further inspect this model across time, here is a year-by-year look at the Q-Q plot:



It is enlightening that the divergence in the upper part of the distribution begins around 2002 and increases to 2008.

Random Effects Regressions

Three random effects models:

1. **Zip Code Random Effects:** implies an error component in the model which that is uncorrelated with the other regressors.
   - α_z is present.
2. **Year Random Effects:** implies
   - α_t is present.
3. **Zip Code and Year Random Effects:** implies
   - Both α_z and α_t are present.

All the models assume homoskedastic error variances, which may understate the true amount of variation present in the data. This will only matter in the event that we find a significant effect associated with the loan program. Failure to find a significant relationship with assumed homoskedastic errors should only be strengthened with robust standard errors.

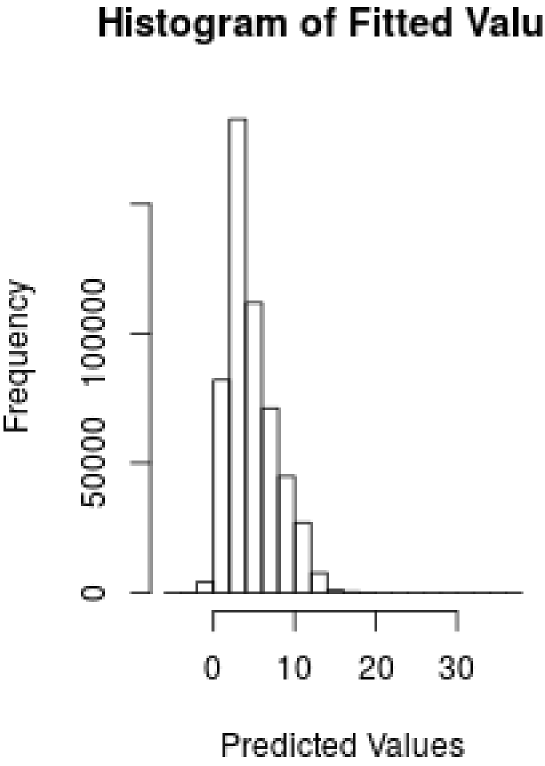
```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta),
##     data = pdata, model = "random")
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Effects:
##                 var std.dev share
## idiosyncratic 4.964   2.228 0.741
## individual    1.733   1.316 0.259
## theta:  0.6295
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -35.100  -1.520  -0.215   1.470  18.800
##
## Coefficients :
##                        Estimate  Std. Error   t-value  Pr(>|t|)
## (Intercept)          3.0020e+01  3.7106e-01   80.9033 < 2.2e-16 ***
## iloans               2.9309e+00  2.7849e-02  105.2441 < 2.2e-16 ***
## log(est)             1.1324e+00  5.8846e-03  192.4397 < 2.2e-16 ***
## log(Pop_IRS)         9.6661e-01  8.4980e-03  113.7456 < 2.2e-16 ***
## logINC              -3.8778e+00  3.6172e-02 -107.2026 < 2.2e-16 ***
## tri                 -2.2843e-03  4.5548e-04   -5.0151 5.303e-07 ***
## rucadj              -5.1480e-02  2.6860e-02   -1.9166 0.0552869 .
## rucnonadj            4.3262e-01  3.1895e-02   13.5637 < 2.2e-16 ***
## poly(AREA_zcta, 2)1 -2.3188e+01  6.9308e+00   -3.3456 0.0008211 ***
## poly(AREA_zcta, 2)2  2.0677e+01  6.7675e+00    3.0553 0.0022486 **
## I(Pop_IRS/AREA_cty)  9.9515e-05  4.5159e-06   22.0366 < 2.2e-16 ***
## I(est/AREA_zcta)     2.9307e-04  2.0567e-05   14.2499 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    3869700
## Residual Sum of Squares: 3099600
## R-Squared      :  0.19899
##       Adj. R-Squared :  0.19899
## F-statistic: 12027.9 on 11 and 532572 DF, p-value: < 2.22e-16
```

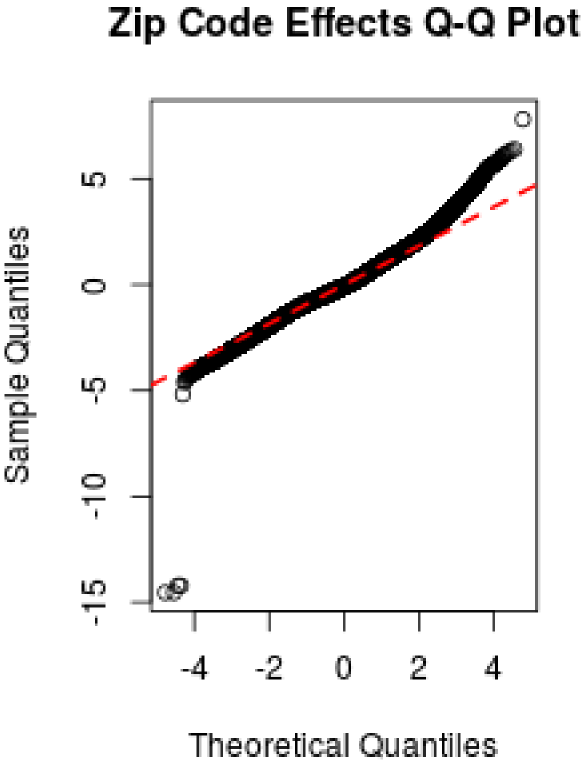
```
## Oneway (time) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta),
##     data = pdata, effect = "time", model = "random")
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Effects:
##                   var std.dev share
## idiosyncratic 4.19338 2.04777 0.985
## time          0.06195 0.24889 0.015
## theta:  0.9522
##
```

```
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -12.200  -1.330  -0.157   1.130  18.000
##
## Coefficients :
##                        Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)         -9.0515e+00  1.5592e-01 -58.0511 < 2.2e-16 ***
## iloans               3.2787e-02  1.8318e-02   1.7898   0.07348 .
## log(est)             1.0284e+00  1.9409e-03 529.8268 < 2.2e-16 ***
## log(Pop_IRS)         4.7549e-01  2.7749e-03 171.3573 < 2.2e-16 ***
## logINC               3.9328e-01  1.4318e-02  27.4676 < 2.2e-16 ***
## tri                 -5.6983e-03  1.4382e-04 -39.6203 < 2.2e-16 ***
## rucadj              -4.8585e-02  8.5279e-03  -5.6971 1.219e-08 ***
## rucnonadj            2.4992e-01  1.0107e-02  24.7275 < 2.2e-16 ***
## poly(AREA_zcta, 2)1 -2.3246e+01  2.1853e+00 -10.6371 < 2.2e-16 ***
## poly(AREA_zcta, 2)2  1.4840e+01  2.1332e+00   6.9567 3.487e-12 ***
## I(Pop_IRS/AREA_cty)  9.7242e-05  1.4415e-06  67.4575 < 2.2e-16 ***
## I(est/AREA_zcta)     3.1289e-04  6.7597e-06  46.2883 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6033700
## Residual Sum of Squares: 2237600
## R-Squared      :  0.62914
##       Adj. R-Squared :  0.62913
## F-statistic: 82134.3 on 11 and 532572 DF, p-value: < 2.22e-16
```

```
## Twoways effects Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta),
##     data = pdata, effect = "twoways", model = "random")
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Effects:
##                   var std.dev share
## idiosyncratic 2.27636 1.50876 0.539
## individual    1.88224 1.37195 0.446
## time          0.06201 0.24902 0.015
## theta  : 0.7491 (id) 0.9648 (time) 0.7488 (total)
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -9.3400 -0.9410 -0.0264  0.8890 14.2000
##
## Coefficients :
##                        Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)          2.6307e+00  3.1827e-01   8.2657 < 2.2e-16 ***
## iloans               1.2105e-01  1.8446e-02   6.5623 5.303e-11 ***
## log(est)             9.7940e-01  5.1849e-03 188.8943 < 2.2e-16 ***
## log(Pop_IRS)         6.4925e-01  7.7862e-03  83.3846 < 2.2e-16 ***
## logINC              -8.7445e-01  3.0400e-02 -28.7649 < 2.2e-16 ***
## tri                 -5.9009e-03  4.2279e-04 -13.9569 < 2.2e-16 ***
## rucadj              -1.0776e-01  2.4824e-02  -4.3409 1.420e-05 ***
## rucnonadj            2.2204e-01  2.9498e-02   7.5272 5.193e-14 ***
## poly(AREA_zcta, 2)1 -9.6867e+00  6.4324e+00  -1.5059    0.1321
## poly(AREA_zcta, 2)2  5.5794e+00  6.2834e+00   0.8880    0.3746
## I(Pop_IRS/AREA_cty)  1.0227e-04  4.1275e-06  24.7764 < 2.2e-16 ***
## I(est/AREA_zcta)     2.7276e-04  1.8124e-05  15.0500 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1481700
## Residual Sum of Squares: 1228200
## R-Squared      :  0.17106
##       Adj. R-Squared :  0.17106
## F-statistic: 9991.02 on 11 and 532572 DF, p-value: < 2.22e-16
```

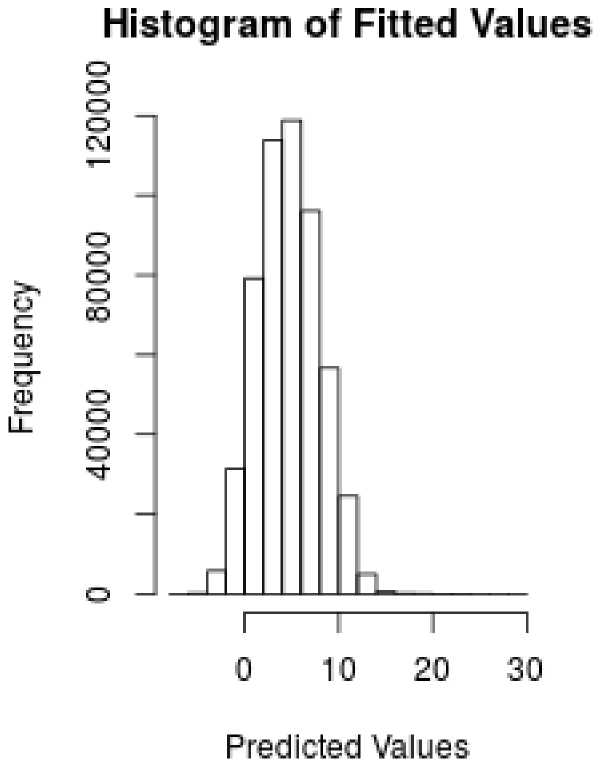
Allowing for random intercepts across zip codes results in a significant relationship for the loan program to the effect of 2.93 more broadband providers across 1999 to 2008. We also see the puzzling result that more income is associated with lower levels of broadband and rural non-adjacent counties have higher levels of broadband. These are at odds to what common sense dictates.

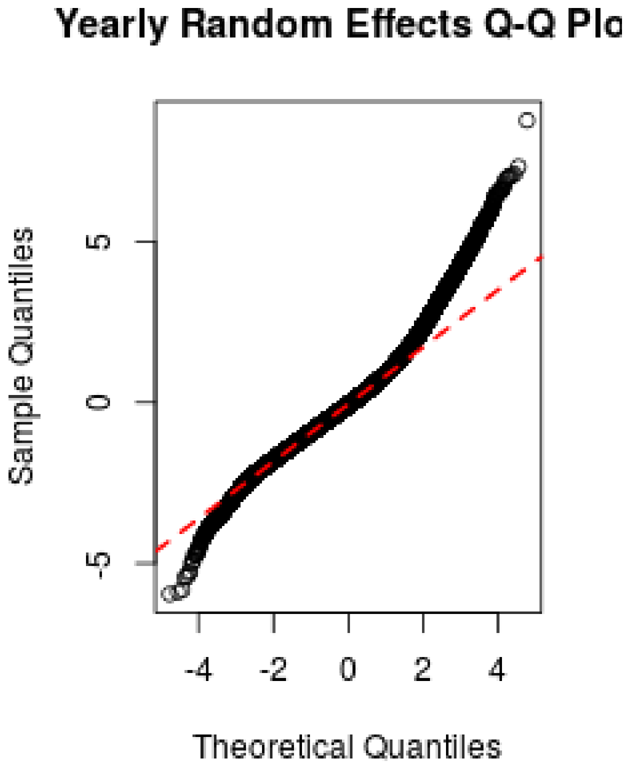




We do see that there is improvement of predicted values from the pooled and fixed effects models in that there are fewer negative values. Also, the distribution of fitted values is much improved as there is a slight right-skew to better match the observed values. However, the Q-Q plot points to divergence from normality of our residuals, thus making inference improper in this setting.

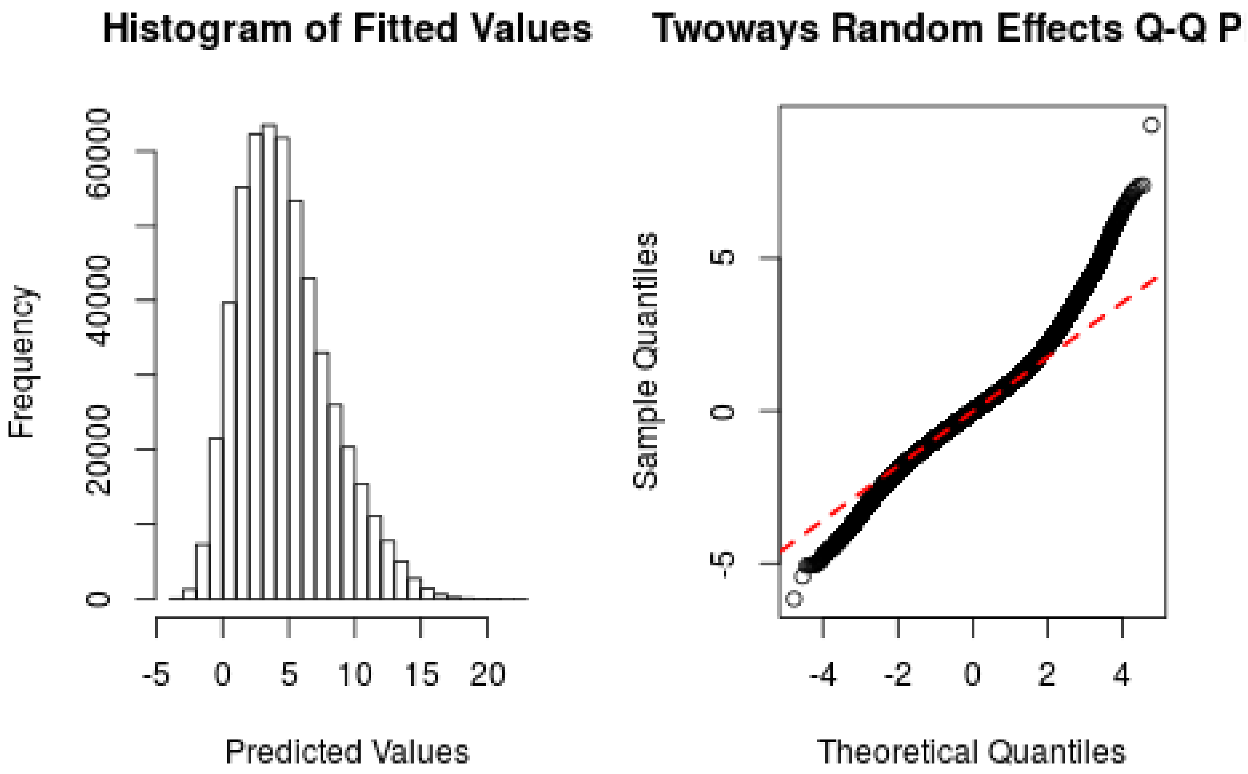
If we look at the random effects model across time, the coefficients make more sense in that income is positively related to broadband access. We still see that there is a negative association with rural-adjacent counties, which appears to be counter-intuitive. In this setting, the loans appear to not have an affect on broadband access. Further inspection of residuals and fitted values is needed:





We do see that there is improvement of predicted values from the pooled and fixed effects models in that there are fewer negative values. Also, the distribution of fitted values is still decidedly normal while our observed values are not. The Q-Q plot also points to divergence from normality of our residuals, thus making inference improper in this setting.

Finally, turning to a random effects model incorporating both the zip code and yearly random effects yields a mixture between the previous models. The counter-intuitive result that income is negatively related to broadband access still appears, which is concerning. Further inspection of the residuals is interesting:



The fitted values appear closest to the observed values, however the Q-Q plot of the residuals indicates a large divergence from the normality assumption. None of these models are appealing to consider inference upon the USDA Broadband Loan Program.

Model Testing and Diagnostics

The most obvious tests are to first check whether we have Fixed Effects or Random Effects through a Hausman test. Below are the Hausman tests matched up from the models above as 1. to 1.; 2. to 2.; and 3. to 3.

```
# Hausman
phtest(p1, r1)
```

```
##
##  Hausman Test
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + I(Pop_IRS/AREA_cty) +  ...
## chisq = 88552, df = 6, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

```
phtest(p1t, r1t)
```

```
##
##  Hausman Test
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + tri +  ...
## chisq = 35.377, df = 11, p-value = 0.0002147
## alternative hypothesis: one model is inconsistent
```

```
phtest(p12, r12)
```

```
##
##  Hausman Test
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + I(Pop_IRS/AREA_cty) +  ...
## chisq = 5976.9, df = 6, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Tests appear to indicate the model of choice would be the fixed effects estimator for 1. and 3. (zip code and two-way effects) of the competing models due to rejection of the null. It is interesting to note that for the 2. model (yearly effects) we see that the null is not rejected which would indicate that both estimators are consistent and the random effects estimator is more efficient.

OTHER DIAGNOSTIC CHECKS WERE DONE BUT I AM SUPPRESSING THEM BECAUSE IT SEEMS OBVIOUS TO ME THAT STATIC PANEL ESTIMATORS ARE NOT APPLICABLE TO BROADBAND DIFFUSION.

One implication for this would be to have a mixed model where there are fixed zip code effects but yearly random effects. But I take this as evidence that none of these models are appealing to use because there are better methods available for estimation.

For instance, knowing that the dependent variable is generally increasing over time, and that once a broadband provider forms in one zip code they are likely to exist for multiple periods further, we know that models which do not account for previous number of broadband providers will not accurately reflect reality. Because of this, we proceed onto a more complex model by lagging the dependent variable. As a quick check for whether or not this may be appropriate, I present tests of serial correlation for the six models that I have discussed in this section to motive that use of a Dynamic Panel Model:

```
# Serial Dependence?
pbgtest(p1)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + I(Pop_IRS/AREA_cty) +     I(est/AREA_z
## chisq = 300150, df = 18, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(p1t)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + tri +     ruc + poly(AREA_zcta, 2) + I
## chisq = 327310, df = 18, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(p12)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + I(Pop_IRS/AREA_cty) +     I(est/AREA_z
## chisq = 221110, df = 18, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```r
pbgtest(r1)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + tri +     ruc + poly(AREA_zcta, 2) + I
## chisq = 337020, df = 18, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```r
pbgtest(r1t)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + tri +     ruc + poly(AREA_zcta, 2) + I
## chisq = 327170, df = 18, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```r
pbgtest(r12)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + tri +     ruc + poly(AREA_zcta, 2) + I
## chisq = 228140, df = 18, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```r
#
# coeftest(r1, vcovHC)
# coeftest(r1t, vcovHC)
#
#
# # LaGrange Multiplier Tests
# plmtest(p1, type = "bp", effect = "time")
# plmtest(p1, type = "bp", effect = "individual")
# plmtest(p1, type = "bp", effect = "twoways")
#
#
# pFtest(p1, p1t) # F-Test of year effects
#
```

```
#
# pFtest(p1, pool1)
# pFtest(p1t, pool1)
# pFtest(p12, pool1)
#
# pFtest(p1b, p1)
# pFtest(p1b, p1t)
#
# #DO NO RUN: pcdtest(p1t, test = c("lm")) # Cross-sectional dependence?
#
# coeftest(p1,  vcov = function(x) vcovHC(x, cluster = "time", type = "HC1"))
# coeftest(p1t, vcov = function(x) vcovHC(x, cluster = "time", type = "HC1"))
```

And further, I have taken the year fixed effects model and calculated the correlation of residuals across time as another way to indicate serial correlation is present in the models:

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
1999	1.000	-0.002	0.038	0.041	-0.006	-0.017	-0.017	0.007	0.017	-0.007
2000	-0.002	1.000	0.019	0.040	-0.016	0.021	0.049	0.000	-0.056	-0.001
2001	0.038	0.019	1.000	0.047	-0.043	0.011	-0.010	0.009	0.000	0.010
2002	0.041	0.040	0.047	1.000	-0.047	-0.009	0.001	0.033	0.011	-0.019
2003	-0.006	-0.016	-0.043	-0.047	1.000	0.004	0.017	0.022	-0.013	0.016
2004	-0.017	0.021	0.011	-0.009	0.004	1.000	0.005	-0.026	-0.017	0.018
2005	-0.017	0.049	-0.010	0.001	0.017	0.005	1.000	0.030	-0.051	0.019
2006	0.007	0.000	0.009	0.033	0.022	-0.026	0.030	1.000	0.014	-0.043
2007	0.017	-0.056	0.000	0.011	-0.013	-0.017	-0.051	0.014	1.000	0.010
2008	-0.007	-0.001	0.010	-0.019	0.016	0.018	0.019	-0.043	0.010	1.000

Dynamic Models

The dynamic model implies a different form of broadband diffusion:

$$Prov_{z,t} = \mu_z + \rho Prov_{z,t-1} + \beta_1 Loan_{z,t} + \beta_2 X_{z,t} + \varepsilon_{z,t}$$

This is a more sophisticated model and requires different estimation technique to handle the endogenous $Prov_{z,t-1}$ on the right hand side. The obvious solution is to use a Generalized Method of Moments (GMM) estimator and have further lags of the dependent variable as instruments. In this setting, time-invariant effects cannot be estimated nor can year fixed/random effects.

The `plm` package allows for three estimators that can be useful here:

1. **Zip Code Fixed Effects OLS:** simply adds in a lagged value of number of broadband providers and treats it as exogenous. Obvious bias.
2. **GMM:** uses lagged values of broadband as an instrument, I also include `tri` as an instrument. No fixed effects are present, although implicitly the auto-regression of the dependent variable should be better capture variation over time.
3. **GMM with Zip Code Fixed Effects:** same instruments as above, however this includes a zip code fixed effect to capture unobserved heterogeneity across zip codes.

```
## Oneway (individual) effect Within Model
##
```

```
## Call:
## plm(formula = Prov_num ~ lag(Prov_num, 1) + iloans + log(est) +
##     log(Pop_IRS) + logINC + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta),
##     data = pdata, model = "within", index = c("zip", "time"))
##
## Unbalanced Panel: n=29588, T=17-18, N=532583
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -45.10000  -0.79700  -0.00102   0.98200  15.10000
##
## Coefficients :
##                        Estimate  Std. Error   t-value Pr(>|t|)
## lag(Prov_num, 1)     4.6147e-01  1.0821e-03  426.4428   <2e-16 ***
## iloans               1.5269e+00  2.3645e-02   64.5743   <2e-16 ***
## log(est)             1.1456e+00  1.4964e-02   76.5547   <2e-16 ***
## log(Pop_IRS)         8.6724e+00  4.6715e-02  185.6443   <2e-16 ***
## logINC              -5.0243e+00  4.6739e-02 -107.4990   <2e-16 ***
## I(Pop_IRS/AREA_cty)  1.2337e-03  2.8976e-05   42.5758   <2e-16 ***
## I(est/AREA_zcta)     8.7446e-05  5.3270e-05    1.6416   0.1007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    3203100
## Residual Sum of Squares: 1833700
## R-Squared      :  0.4275
##       Adj. R-Squared :  0.40375
## F-statistic: 53656.8 on 7 and 502988 DF, p-value: < 2.22e-16
```

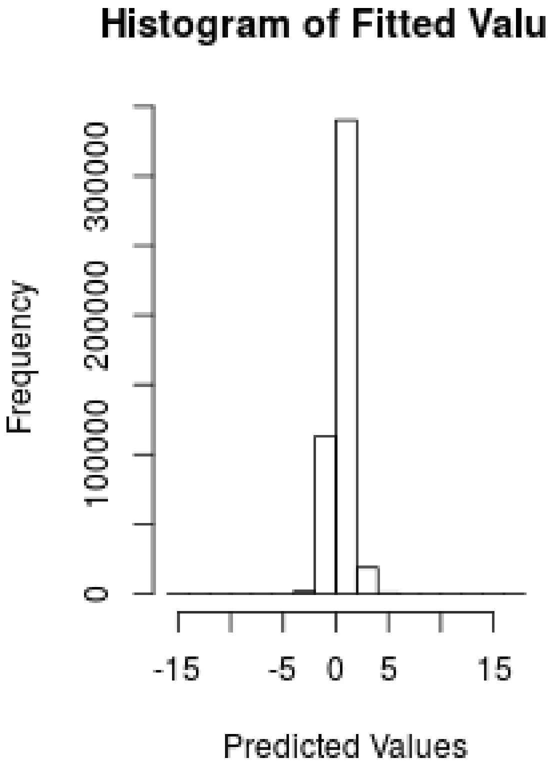
```
## Twoways effects One step model
##
## Call:
## pgmm(formula = Prov_num ~ lag(Prov_num, 1) + iloans + log(est) +
##     log(Pop_IRS) + logINC + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta) |
##     lag(Prov_num, 2:99) + tri, data = pdata, index = c("zip",
##     "time"))
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Number of Observations Used:  473408
##
## Residuals
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -16.98000  -0.64250  -0.09545   0.00000   0.61770  15.84000
##
## Coefficients
##                        Estimate  Std. Error  z-value  Pr(>|z|)
## lag(Prov_num, 1)     6.0374e-01  3.9783e-03 151.7587 < 2.2e-16 ***
## iloans               1.8899e-01  4.3005e-02   4.3947 1.109e-05 ***
## log(est)             1.4252e-01  2.2858e-02   6.2349 4.522e-10 ***
## log(Pop_IRS)         1.5378e+00  1.9831e-01   7.7546 8.860e-15 ***
## logINC              -1.5288e+00  5.0298e-01  -3.0394 0.0023703 **
## I(Pop_IRS/AREA_cty)  2.2035e-04  3.1800e-05   6.9293 4.228e-12 ***
## I(est/AREA_zcta)    -2.9923e-05  8.0306e-06  -3.7262 0.0001944 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sargan Test: chisq(151) = 11669.39 (p.value=< 2.22e-16)
## Autocorrelation test (1): normal = -112.8362 (p.value=< 2.22e-16)
## Autocorrelation test (2): normal = 14.50337 (p.value=< 2.22e-16)
## Wald test for coefficients: chisq(7) = 36553.3 (p.value=< 2.22e-16)
## Wald test for time dummies: chisq(16) = 35440.43 (p.value=< 2.22e-16)
```

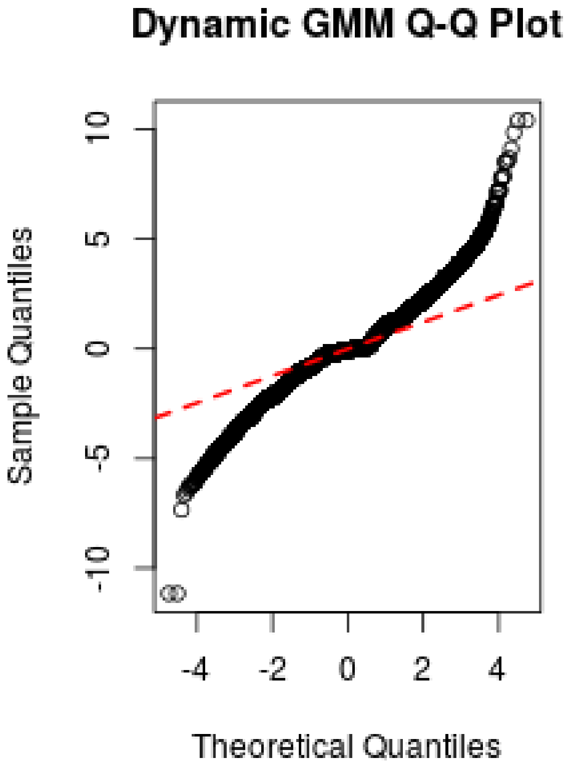
```
## Oneway (individual) effect One step model
##
## Call:
## pgmm(formula = Prov_num ~ lag(Prov_num, 1) + iloans + log(est) +
##     log(Pop_IRS) + logINC + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta) |
##     lag(Prov_num, 2:99) + tri, data = pdata, effect = "individual",
##     index = c("zip", "time"))
##
## Balanced Panel: n=29588, T=18, N=532584
##
## Number of Observations Used:  369198
##
## Residuals
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -16.23000  -0.86570   0.00000   0.07372   0.97060  14.97000
##
## Coefficients
##                       Estimate Std. Error  z-value Pr(>|z|)
## lag(Prov_num, 1)    9.4889e-01 2.1132e-03 449.0262  < 2e-16 ***
## iloans              5.5843e-01 4.9650e-02  11.2474  < 2e-16 ***
## log(est)            2.8154e-01 2.3402e-02  12.0305  < 2e-16 ***
## log(Pop_IRS)        1.5484e+00 8.4445e-02  18.3359  < 2e-16 ***
## logINC              7.9060e-02 4.4397e-02   1.7808  0.07495 .
## I(Pop_IRS/AREA_cty) 8.7354e-04 8.4066e-05  10.3911  < 2e-16 ***
## I(est/AREA_zcta)    1.3219e-04 9.8542e-05   1.3415  0.17977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sargan Test: chisq(151) = 15439.01 (p.value=< 2.22e-16)
## Autocorrelation test (1): normal = -126.1982 (p.value=< 2.22e-16)
## Autocorrelation test (2): normal = 15.11137 (p.value=< 2.22e-16)
## Wald test for coefficients: chisq(7) = 741290.5 (p.value=< 2.22e-16)
```

We can toss out the OLS estimator as the coefficients make no sense and we already know that the estimator is not consistent. It is useful to show because if this estimator was similar to GMM, then that would be indication that GMM is not necessary.

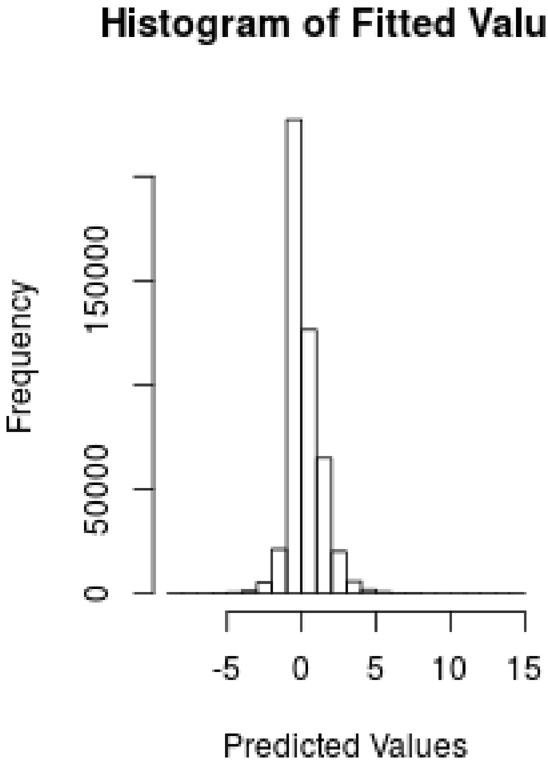
Both of the GMM estimators indicate the perplexing result that income is negatively associated with broadband access. This gives pause to use of these models and I have been attempting to figure out why this is the case. Possibly more puzzling is that when fixed effects are added, the auto-regressive parameter becomes larger than 1 which would imply an explosive process. I have to think this through a bit more, but it is likely the case that fitting a linear curve to a technology with rapid adoption would pose a problem.

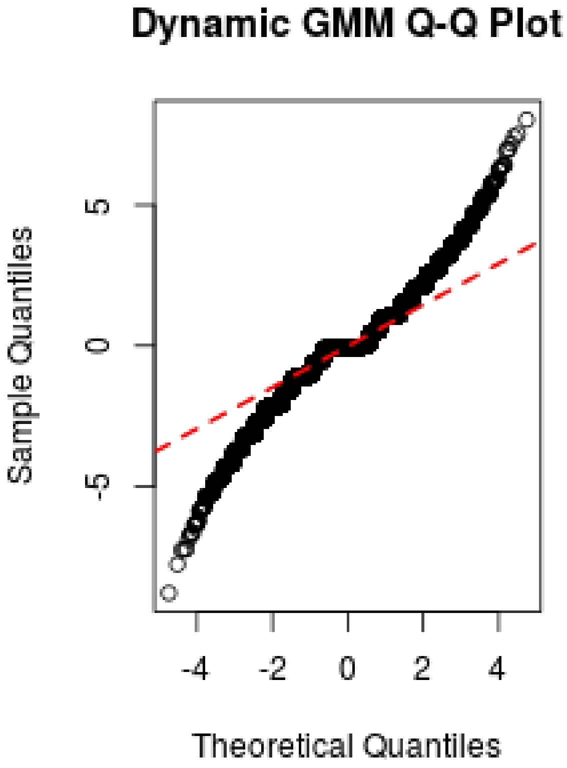
However, if we inspect the fitted values and Q-Q plots of the residuals we can see that a sophisticated model may not always be better:





The above is the second model. It is troublesome that there are fitted values that are drastically negative, but even more of a problem is that the Q-Q plot is nowhere close to normal. But adding in fixed effects is clearly not the solution:





While there is improvement of the predicted values, the Q-Q plot is even worse. A dynamic model is certainly not the solution here.

Next Step

We know that our dependent variable is a count variable and can only take on positive values. A model which predicts negative values for Broadband Providers indicates a problem. We know that this cannot happen and divergence from this too far will render this modeling exercise meaningless. This is why a Poisson Regression appears to be the best choice. I have results for this, but I need to format it so the results can be displayed. I am not sure how quickly I can do that.