SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

Predicting Highly Qualified Math Teachers in Secondary Schools in the United States

Bogdan Gadidov

Kennesaw State University

USERS PROGRAM

Predicting Highly Qualified Math Teachers in Secondary Schools in the United States

Bogdan Gadidov
Kennesaw State University
Advising Professor: Dr. Herman E. Ray

Abstract

Are secondary schools in the United States hiring enough qualified STEM (Science, Technology, Engineering, Mathematics) teachers? In which regions are there a disparity of qualified teachers? Data from an extensive survey conducted by the National Center for Education Statistics (NCES) is used for predicting qualified secondary school teachers across public schools in the U.S. The 3 criteria which are looked at in determining whether a teacher is qualified to teach a given subject are: whether the teacher has a degree in the subject they are teaching, teaching certification in the subject they are teaching, or 5 years of experience in the subject they are teaching. A qualified teacher is defined as one who has all 3 of the previous qualifications. The sample data includes socioeconomic data at the county level, which are used as predictors for hiring a qualified teacher. Data such as the number of students on free or reduced lunch at the school is used to assign schools as high needs or low needs schools. Other socioeconomic factors included are the income and education levels of working adults within a given school district. Some of the results show that schools with higher needs students (a school which has more than 40% of the students on some form of reduced lunch program) have less qualified teachers. The resultant model is used to score other regions and is presented on a heat map of the U.S. SAS' survey family of procedures such as Proc Surveyfreq and Proc Surveylogistic are used in the analyses since the data involves replicate weights.

Data Exploration

The focus for the model and graphics will be for math teachers from the survey. Math teachers were the majority of all survey responders. In total, there were 2,670 math teachers who responded in the survey. Data regarding the school district is used to assess the likelihood that a given school can hire a highly qualified teacher. The variables included in the final model are the region of the school, the percentage of students on a free or reduced lunch program (named NSLAPP_S), and the size of the school. Tables 1 and 2 below show some descriptive statistics for the NSLAPP_S variable, which is the percentage of students at the school who are on a free or reduced lunch program.

Table 1: Quantiles for Weighted Needs Variable using Proc Surveymeans

Quantiles						
Variable	Percentile		Estimate	Std Error	95% Confidence Limits	
NSLAPP_S	25%	Q1	20.17	1.88	16.43	23.91
	75%	Q3	62.04	1.78	58.52	65.57

Table 2: Descriptive Statistics for Unweighted Needs Variable using Proc Means

Analysis Variable : NSLAPP_S		
Lower Quartile	Mean	Upper Quartile
22.26	42.5	60.02

*The average for needs variable is 42.4% in the weighted sample.

Table 2 contains descriptive statistics of the original sample of 2,670 teachers while Table 1 has the weighted values corresponding to the needs variable. The purpose is to illustrate that the statistics are very similar in the unweighted or weighted case, but the weighted case is more relevant since the data comes from a survey. A similar Proc Surveyfreq can be used to measure a categorical variable, as shown in the next section.

Data Exploration (cont.)

The Proc Surveyfreq procedure is used to view the distribution of school sizes within the survey sample. The school size variable (named SCHSIZE) is created by binning the number of students in a school into one of 12 groupings. The creation of this variable is shown in Table 3. Table 4 shows the Proc Surveyfreq procedure and corresponding output. Note that both the unweighted and weighted frequencies are given by this procedure. The total weighted frequency of teachers is 142,832.

Table 3: Creation of School Size Variable

Number of Students	SCHSIZE
0-49	1
50-99	2
100-149	3
150-199	4
200-349	5
400-499	6
500-749	7
750-999	8
1000-1199	9
1200-1499	10
1500-1999	11
2000 or More	12

Table 4: Distribution of School Size Variable Using Proc Surveyfreq

SCHSIZE	Frequency	Weighted Frequency	Std Dev of Wgt Frequency	Percent	Std Err of Percent
1	37	1197	462	0.84	0.32
2	89	2465	506	1.73	0.35
3	104	2688	533	1.88	0.37
4	89	2735	743	1.91	0.52
5	279	10656	1317	7.46	0.94
6	230	9426	1130	6.60	0.80
7	328	16500	1482	11.55	1.04
8	263	15167	1767	10.62	1.16
9	177	11333	1423	7.93	0.90
10	230	14535	1499	10.18	0.96
11	385	29520	2441	20.67	1.50
12	313	26610	2784	18.63	1.71
Total	2524	142832	4892	100	

Model

The Proc Surveylogistic procedure is used to create the model. The resulting value from this procedure will give a probability of a math teacher at a given high school being highly qualified. The goal of the model will be to score a separate dataset, containing school and county level attributes. By scoring this dataset, the probability of hiring a highly qualified math teacher can be created for high schools across the U.S. Table 5 below contains the final variables selected in the model, and corresponding p-values. Table 6 contains the parameter estimates for the variables. The region variable contains 4 levels: Northeast, Midwest, South, and West (1,2,3,4, respectively), and the Western region is used as the baseline region for comparisons of the other regions. The c-stat for this model is 0.622.

Table 5: Variables in Final Model

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
REGION	3	10.1323	0.0175
NSLAPP_S	1	9.7111	0.0018
SCHSIZE	1	2.9672	0.0850
SCHSIZE*REGION	3	6.6769	0.0829

Table 6: Parameter Estimates of Final Model

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.9150	0.3840	5.6775	0.0172
REGION	1	1	-0.1489	0.8783	0.0288	0.8653
REGION	2	1	0.8926	0.4735	3.5544	0.0594
REGION	3	1	1.5443	0.5395	8.1933	0.0042
NSLAPP_S		1	-0.00874	0.00281	9.7111	0.0018
SCHSIZE		1	0.0671	0.0389	2.9672	0.0850
SCHSIZE*REGION	1	1	0.0767	0.1002	0.5858	0.4441
SCHSIZE*REGION	2	1	0.0103	0.0591	0.0306	0.8612
SCHSIZE*REGION	3	1	-0.1364	0.0624	4.7817	0.0288

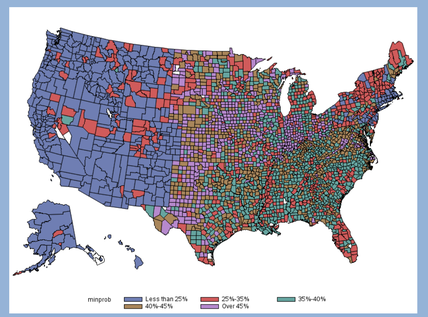
Predicting Highly Qualified Math Teachers in Secondary Schools in the United States

Bogdan Gadidov
Kennesaw State University
Advising Professor: Dr. Herman E. Ray

County Level Graphs of U.S. Secondary Schools

The resulting model is used to score high schools across the U.S., finding the probability that a math teacher at a given school is a highly qualified teacher. The probabilities are then aggregated at the county level. For instance, many counties have more than one high school represented in the scoring dataset, so there are several metrics used to summarize the probability of a highly qualified math teacher at the county level. The minimum, maximum, and average probabilities are used to create 3 county level graphs for the predicted probability of a highly qualified math teacher. For example, if a county has 8 high schools which are scored, each of those 8 schools has an individual probability that a math teacher at that school is highly qualified. The minimum of these 8 probabilities will be plotted in one graph, while the maximum of these 8 probabilities will be plotted in the next graph. Finally, the average of these probabilities will also be plotted. The goal is to have one probability represent the entire county.

Figure 1: County Level Graph Using Minimum of School Probabilities



County Level Graphs of U.S. Secondary Schools (cont.)

Figure 2: County Level Graph Using Maximum of School Probabilities

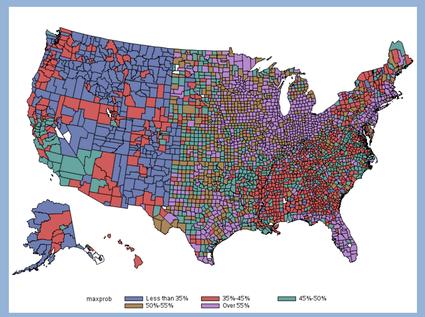
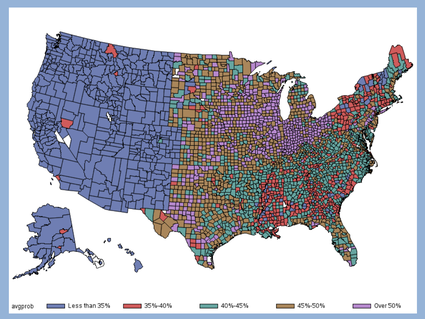


Figure 3: County Level Graph Using Average of School Probabilities



Predicting Highly Qualified Math Teachers in Secondary Schools in the United States

Bogdan Gadidov
Kennesaw State University
Advising Professor: Dr. Herman E. Ray

Conclusion

Using the minimum probability among schools within a county, it is apparent from the map that schools in the west seem to have the lowest probability of having a math teacher which is highly qualified. The midwest region has the highest probability, while counties in the south seem to consistently have probabilities in between the west and midwest regions.

It is interesting to see that when comparing the maximum probabilities among schools within a county, the south performs worse than it did using the minimum probability. The midwest region again performs the best, but many counties in the northeast perform very well with respect to the rest of the nation. Counties in the west which represent more urban areas perform better than when using the minimum probability, but overall the west again performs poorly when compared to the rest of the nation.

Using the average probability, the same sort of pattern is seen. The west performs the worst, while midwest and northeast perform better with respect to other counties in the nation.

Select SAS Code

```
proc sql;create table mins as select state, county, min(prob) as minprob from score
group by state, county;quit;

/* maximum of county probs */
proc sql;create table max as select state, county, max(prob) as maxprob from score
group by state, county;quit;

/* average of county probs */
proc sql;create table avg as select state, county, avg(prob) as avgprob from score
group by state, county;quit;



proc sort data=mins;by state county;run;
proc sort data=max;by state county;run;
proc sort data=avg;by state county;run;
data counties_min;merge score maps.uscounty;by state county;run;
data counties_max;merge score maps.uscounty;by state county;run;
data counties_avg;merge score maps.uscounty;by state county;run;


proc format;
  value mins low-0.25='Less than 25%'
             0.2500001-0.35='25%-35%'
             0.3500001-0.4='35%-40%'
             0.4000001-0.45='40%-45%'
             0.4500001-high='Over  45%';
run;

 proc gmap data = counties_min map =  maps.uscounty all;format minprob mins.;
   id state county;
   choro minprob/coutline=black discrete;
   run;quit;


   proc format;
  value maxs low-0.35='Less than 35%'
             0.35000001-0.45='35%-45%'
             0.45000001-0.5='45%-50%'
             0.50000001-0.55='50%-55%'
             0.55-high='Over 55%';
run;

   proc gmap data = counties_max map =  maps.uscounty all;format maxprob maxs.;
   id state county;
   choro maxprob/coutline=black discrete;
   run;quit;


    proc format;
  value avgs low-0.35='Less than 35%'
             0.35000001-0.40='35%-40%'
             0.40000001-0.45='40%-45%'
             0.45000001-0.50='45%-50%'
             0.50000001-high='Over  50%';
run;

   proc gmap data = counties_avg map =  maps.uscounty all;format avgprob avgs.;
   id state county;
   choro avgprob/coutline=black discrete;
   run;quit;
```

```
proc surveymeans data=test order=formatted varmethod=brr q1 mean q3;
var nslapp_s;
WEIGHT weight;
repweights trepwt1-trepwt88;where nslapp_s>=0;
RUN;

proc surveyfreq data=test order=formatted varmethod=brr;
tables schsize /row;WEIGHT weight;
repweights trepwt1-trepwt88;where nslapp_s>=0;
RUN;

proc surveymeans data=test order=formatted varmethod=brr q1 mean q3;
var nslapp_s;by region;
WEIGHT weight;
repweights trepwt1-trepwt88;where nslapp_s>=0;
RUN;


/* code for surveylogistic */
proc surveylogistic data=test varmethod=BRR;
class region (ref="4")
/param=ref;
model mathqual3 (desc)=region nslapp_s schsize schsize*region;
weight weight;repweight trepwt1-trepwt88;where nslapp_s>=0;run;
```

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL