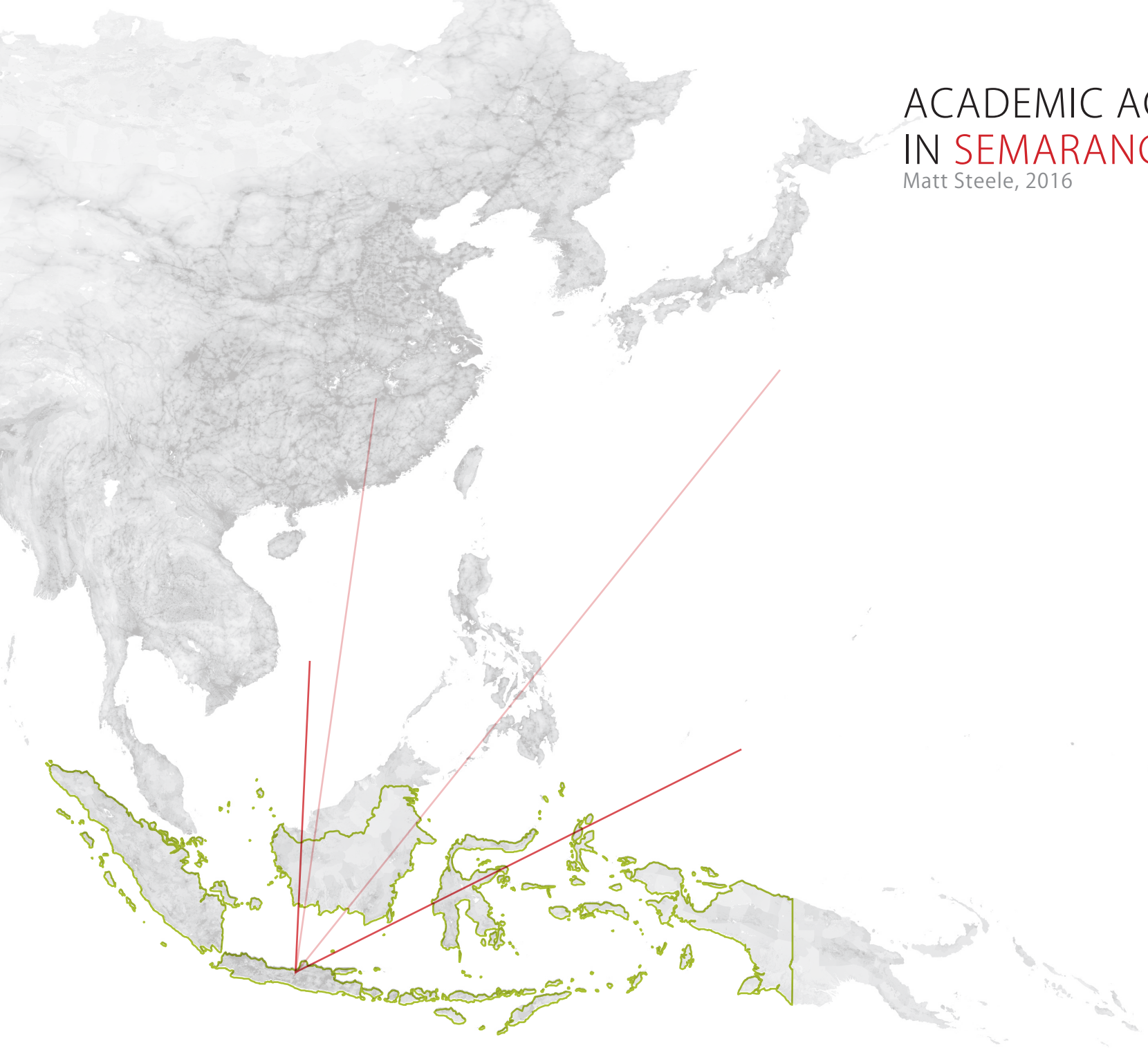
ACADEMIC ACHIEVEMENT IN SEMARANG INDONESIA

Matt Steele, 2016



PURPOSE

The intention of this study is to identify key factors associated with academic achievement in Indonesia. These factors are examined through a series of logistical models using the probability of school attendance or degree achievement as a function of individual, household and geographic characteristics. The chief interest of this study is to inform infrastructure and planning interventions that may improve educational achievement and access in the city of Semarang. To do so, four models were examined that explore different key indicators of achievement amidst the children population and the total population. Those models being:

4 Models Built

1,500,000+ observations
micro census dataset

children only
50% sample

1. Poor Attendance / Children only (ChldPoorAttMod)

The first model of this report provides an understanding of what factors might be associated with children not attending school. This model's dependent variable includes responses of both never attending and not currently attending for people under the age of 18, but over the age of 6.

2. Never Attended / Children Only (ChldNoAttMod)

The second model of this report examines factors that are associated with no attendance. There were far fewer observations of children under 18 never attending school, so only a few factors were significant.

whole dataset
25% sample

3. Never Attended / Total population (TotalPoorAttMod2)

The third model examines factors that might be associated with never attending school among the total population, looking at all age groups.

4. Educational Level achieved (TotEduMod2)

The last model attempts to understand factors that increase academic achievement.

SOURCE

The data for this analysis is built using anonymized individual-level data from the National Census, conducted by the National Statistics Agency (BPS). The data was acquired through the World Bank's micro-database. The last year the census data is available is 2010. While the Census is a comprehensive survey, the data across villages did not appear consistent with expectations. Most notably, the populations in each age-group varied; the younger the age group, the less observations. This suggests that younger children were less likely to be accounted for in the census. There may be some level of selection bias to this report because children not accounted for in the census may be more likely to also not attend school. Because school attendance is compulsory in Indonesia, it may be the fear of reprisal that is motivating parents to underreport children out of school.[1]

BACKGROUND

Indonesian Context

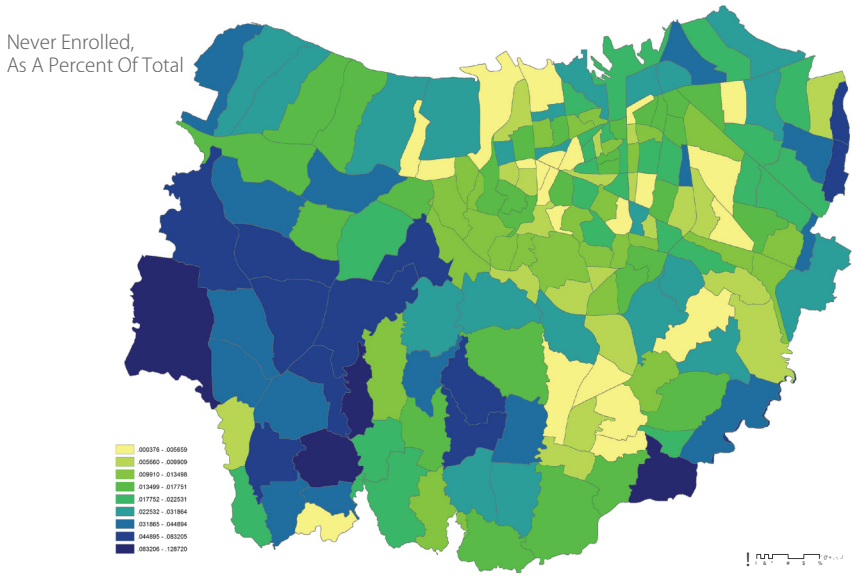
Indonesia is a major developing country, with the fourth largest population in the world. The country hosts some of the most rapidly urbanizing centers in the world. Indonesian rural migrants have moved to the country's urban centers at a staggering rate, which has put pressures on the little infrastructure that is in place.

Transportation and education are the two major issues that each of the major cities in Indonesia are chiefly concerned with. The former has been the subject of a number of technical assistance programs, including the Surabaya Urban Transit Corridor Study. The later has seen some focus nationally, but little focus has been placed on the dynamics at play within the major cities. In 2014, World Bank experts produced a report titled "Teacher Reform in Indonesia, The Role of Politics and Evidence in Policy Making". This analysis considered data provided by the national education department to understand the impacts of various policy reforms to the school systems. The major policy reforms have aimed to increase the quality of education in the schools, but they have mostly failed. Academic achievement in Indonesia trails nearly all other east Asian countries. The reasons for this are many, but a significant factor is a lack of teacher credentials. Teacher hiring and school budget allocations have been politicized and there is a lack of oversight. A similar report produced in 2010 by the Indonesia / Australia Basic Education Program examines various barriers from a qualitative perspective. They uncovered that issues surrounding poverty remain a major barrier for achieving universal access among Indonesia's youth.

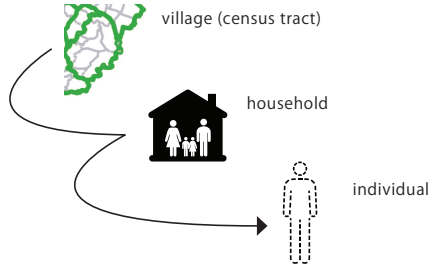
`

Semarang City Context

This study attempts to build a local understanding of the dynamics at play for educational achievement, using Semarang as a key case study. Semarang is a large coastal city in Indonesia. It is the 6th largest city in Indonesia. The city's southern wing hosts one of the leading universities in the country, the University of Diponogoro, otherwise known as UNDIP. As shown below, educational services appear to have less penetration into villages on the edges of the city.



DATA DEVELOPMENT



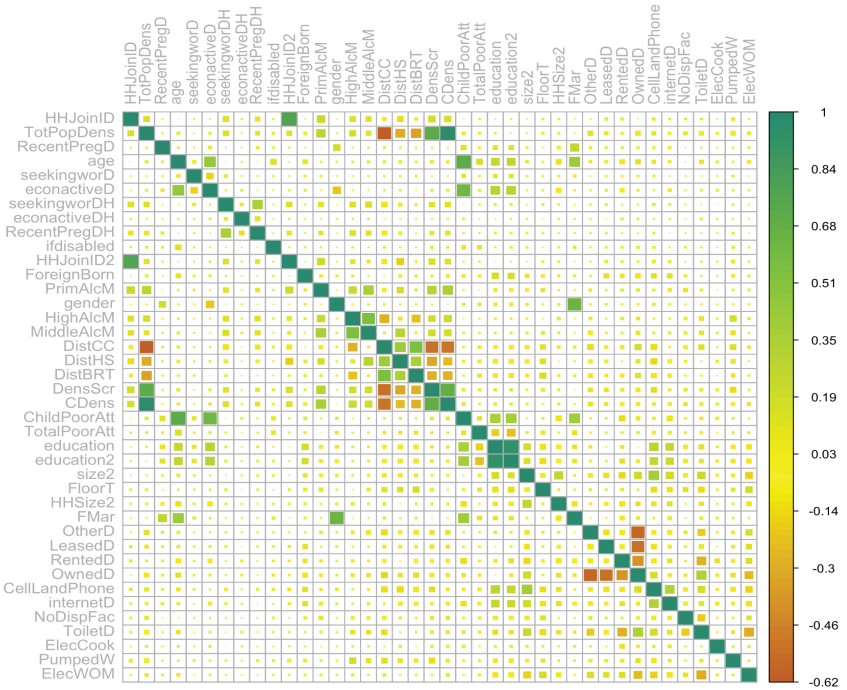
The total population in Semarang observed by the census was over 1.5 million people. The individual level data reflects an observation for each of these 1.5 million individuals. The census data available comes in a table for the household survey conducted as well as a table for the individual survey conducted. For this analysis, the household data and individual level data was merged by creating an ID based on the identified person number and household number and village number in the dataset.

The models were developed first using the full sample and then trained using random subsets of the sample. The models focused on the children only datasets were subsetted to fifty percent of the original dataset, consisting of 135,196 observations. The models focused on the total population were built using twenty-five percent of the dataset. The model making use of the smaller random samples produced roughly the same levels of significance. Variables whose significance appeared subject to changes from the sample group were removed from each model. This ensures that the model is generalizable, and not overfit.

The household data characteristics of a household member looking for work, economically active, or recently pregnant were generalized to all household members. Meaning, for example, the seekingwork dummy variable identified those whose households had a family member that was looking for work. Ultimately these variables were found to not be significant in the child specific models, and presented inherent collinearity with the more significant variables from which they were derived in the total population models.

In total there were 119 variables available from the various sources gathered. 34 education related variables were built for this analysis between the spatial data developed as well as the data from the household survey and the individual survey from the 2010 census. The raw data provided the answers to questions asked. Some of the responses were Boolean ("yes"/ "no") or categorical, and others were continuous. Categorical variables were transformed into dummy variables.

Figure 1: Correlation Plot of Variables



DEPENDENT VARIABLE

The census survey contains two questions relating to educational attainment: the current educational status of the respondent as well as the level of education they achieved. The educational status available responses were "currently enrolled", "was enrolled, but not currently", and "never enrolled". The number of affirmative responses for never enrolled were few in number across the whole dataset, but especially for children above age 6. The number of responses of those who were not currently enrolled, but previously enrolled, was greater among the child-aged population.

The educational level achieved was a numerical variable, one through nine. For purposes of this analysis vocational high schools were grouped in with regular high schools, and non-undergraduate post secondary school program were grouped in with undergraduate degrees. This ensured a continuum of educational attainment as a dependent variable, one through six.

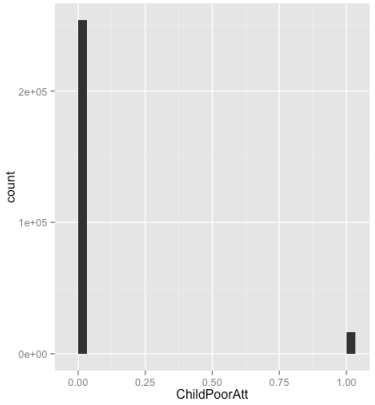


Figure 5: Proportion of Affirmative Responses to "Not Currently Attending" Within the Children Dataset.

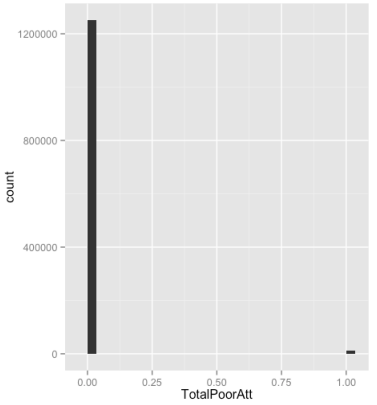


Figure 6: Proportion of Affirmative Responses to "Never have attended" Within the Children Dataset

Figure 4: Educational Attainment Histogram

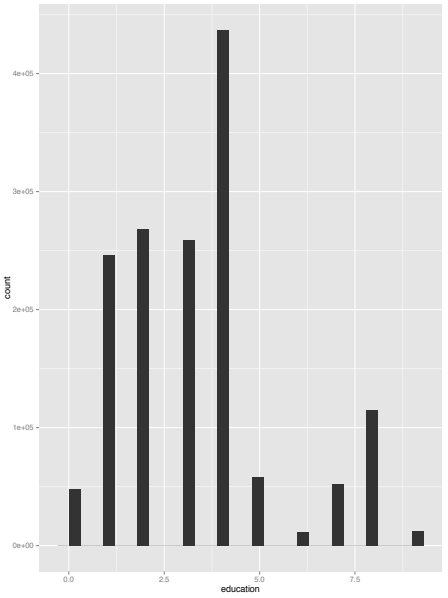


Figure 2: Educational Attainment by Age and by Gender

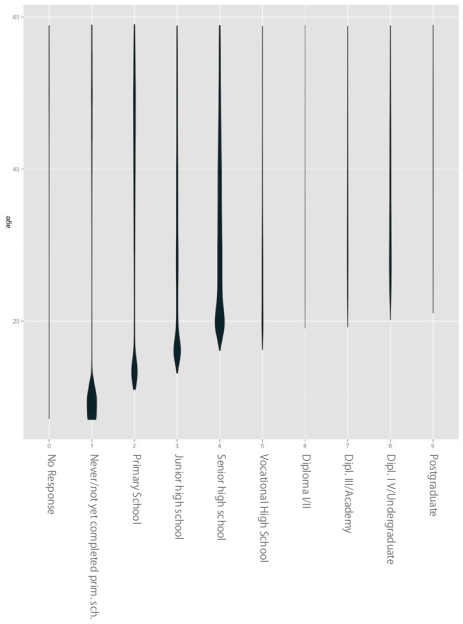
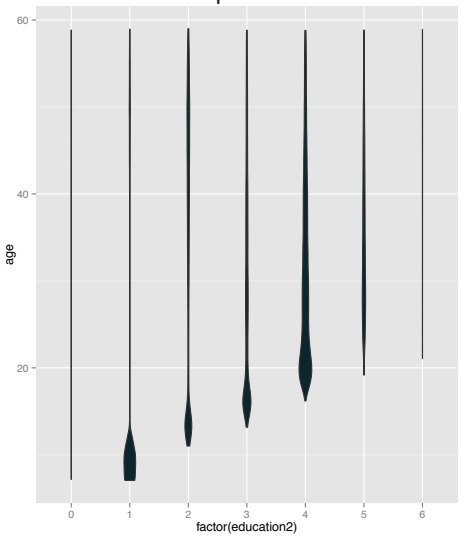


Figure 3: Educational Attainment by Age and by Gender, Merged into a Continuous Spectrum



SPATIAL VARIABLES

8 spatial variables were created for consideration in this analysis. Distance variables were computed using the zonal statistics tool in ArcGIS, considering the mean observation of the Euclidean distance to the points specified. The total and child population densities in each village was identified by using the calculate geometry tool in ArcGIS to identify the area of the villages in meters and then the total population associated with each village was divided by the area identified. A building density score was also included as a variable, which ranged from a scale of 1.0 to 30.0. It was derived as the weighted average score of both a kernel density overlay for building points and the average size of the building footprints in each village. For each village, three variables were created that reflected the spatially-weighted average of the populations in each of the areas surrounding middle, high and primary schools. Because catchment areas in Semarang are broadly defined, this metric shouldn't be directly interpreted as a proxy for school burden, but it does provide an interesting indicator of school service spatial efficiency. The school coverage zones were identified by first creating school zones making use of the cost allocation tool in ArcGIS. This takes a friction layer and a set of points as inputs. The friction layer indicates barriers and pathways for composing the zones. These zones were then tabulated based on the populations of the respective age groups for each school grade level. 6 to 11 for Primary school, 12 to 15 for middle school, and 16 to 18 for high schools. The larger the number, the more poorly served that area is by that particular set of schools.

The spatial data may be subject to some error because the administrative boundaries that the census uses differ from nationally ordained boundaries. The lack of consistent geographies at the sub-district level is a major barrier to doing sub-district statistical analysis in Indonesia. To avoid as much error as possible, the geographies used for this analysis was the 2011 PODES shapefile from BPN, which is the closest available geography to those the BPN used in 2010.

Figure 7: Distance to Center City

Figure 8: Number of Students Aged 6 to 8 near primary schools

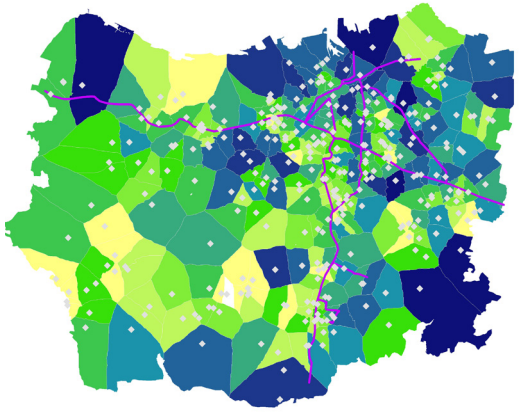


Figure 9: Distance to Bus Rapid Transit

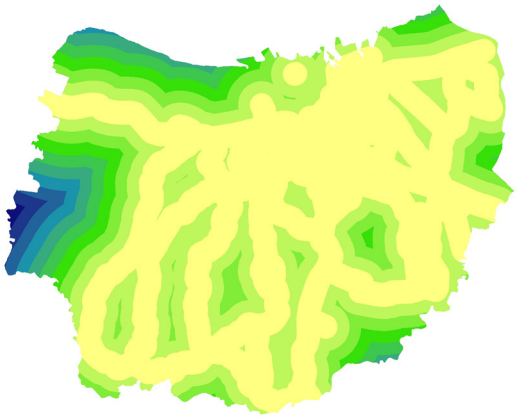
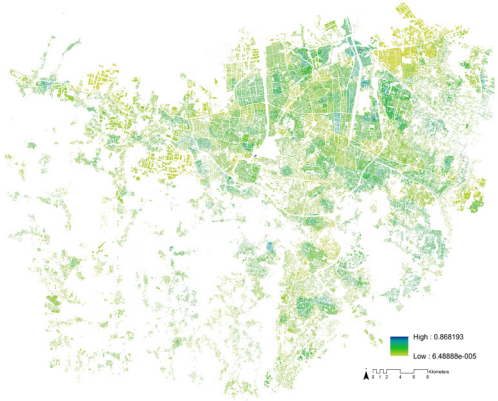


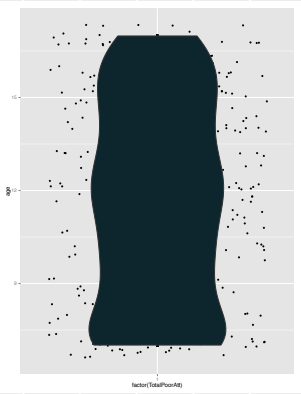
Figure 10: Building Density



The data was organized into two subsets: total population and children population. Both groups excluded observations of individuals under 6 years old, the age children are expected to attend school. The total dataset excluded observations from those above 60 because the proportion of those who had never attended school was disproportionately high since school was not compulsory in the 1960s, and the focus of this study is the general population. Within the child aged population the data was generally consistent as shown in the adjacent violin frequency plot figure. As discussed the last section of this report, given contradictory dynamics occuring across age categories, the more specific the age group of the analysis the more explanatory power the model will have.

The benefit of working with such a large, and presumably comprehensive dataset, is that the models have a high level of reliability, as the observations used for this analysis aren't built from a randomized sample, but is rather indicative of all phenomonon occuring. The implications of models with low R2, especially those build using multiple iterations of random samples, are still significant. Indeed, with such a large dataset, significance of variables is easier to find as well as nuances in their variations.

Figure 11: Age of Student vs poor attendance



Remove Outliers

The number of households was calculated by counting the number of observations in each household ID. Households with more than 50 members were excluded from the study. Such observations are likely of individuals residing in institutions, such as permanent residents of dormitories, hospitals or prisons. There appears to be a huge range of home size in terms of square meters. These studies excluded observations whose the size of the home was larger than 1,500 square meters.

Standardization

Interpreting the odds ratios of continuous variables is difficult when their terms change in small increments. For this reason the continuous variables were standardized to ease understanding. Households size was multiplied by the power of 10.

EDUCATIONAL ACCESS AND ACHIEVEMENT MODELS

Four models were built examining educational attainment in Semarang. Two made use of the children's subset of the data and examine factors that were associated with observations where the respondent had never attended school, or were not currently attending school. Two models were built making use of the total population dataset of all age groups. The first attempted to identify the probability of a respondent of any age having never attended school. The second was a normal regression, making use of educational level achieved as a continuous variable. The confusion tables were built considering observations identified with a probability higher than .3 as being likely. A Receiver Operating Characteristic (ROC) curve was built for each of the logistic models, it examines the specificity and the sensitivity rate of the model at predicting actual observations. Because the model was built with multiple random samples, a high level of predictive accuracy supports the notion that the model is robust.

Model 1 – Poor Attendance / Children Only

$$Ln\left(\frac{P(PoorAttendence\ Among\ Children)}{1-P(PoorAttendence\ Among\ Children)}\right) = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 internet + \beta_4 if Disabled + B_5 MiddleSchoolStudentBurden + B_6 DistanceToBRT + B_7 \#ofHouseHoldMembers + B_8 PopDen + B_9 CellPhoneandLandPhoneAccess + B_{10} EconomicallyActive + \epsilon$$

The first model had the greatest level of explanatory power, reflected in its high McFadden's R^2 of .55. The McFadden's R^2 metric suggests high explanatory power if it is above .2. Much of its explanatory power comes from the inclusion of the key variable, *economically active*. Those who are working are the ones least likely to also be in school. Though it doesn't explain all of the variation identified. The model run without the economically active factor is .33, which is still high for a McFadden's R^2.

Tests for multicollinearity and heteroscedasticity were conductd. The ANOVA test when applied to one model tests the probability that the average means of any of the variables are the same. Each value had a p-value of less than .05. The regressions for each of the following models was very close but consistently just below zero. Meaning that the key factors not included in this model would have suggested greater attendance among the observations. Each of hte models also had a greater positive range of residuals than negative, suggesting that the accuracy of the model had more variance in identifying poor attendence than not, which is likely a function of how few affirmative observations there are present in the dataset, though also suggests some level of heteroscedasticity.

The VIF function tests for multicollinearity within the model. The output is an absolute metric to compare the uniqueness of each variable's influence to the model. Values close to and above 4 present concerns. All values had low values, consistently around 1. The table in the following page provides a breakdown of the coefficient figures. Also featured is a chart depicting the odds ratio. To demonstrate high and low likelihood, the ratio was set to 0, with factors above indicating higher probability, and factors below indicating lower probability. Traditionally the ratio is considered with respect to 1.

Distance to Center City was surprisingly not a significant variable for this factor. This may be because there are other central nodes of significant development, notably the node surrounding UNDIP, though it did appear an important factor for the educational level achieved. Having a *household member seeking work* was also not a significant factor relating to educational achievement of children.

The older you are the more likely you are to not be in school, and if you are disabled there is an even greater chance that you are out of school. While *being Female* was a consistent factor suggesting less attendance,

Figure 12: Odds Ratio Comparison, Model 1

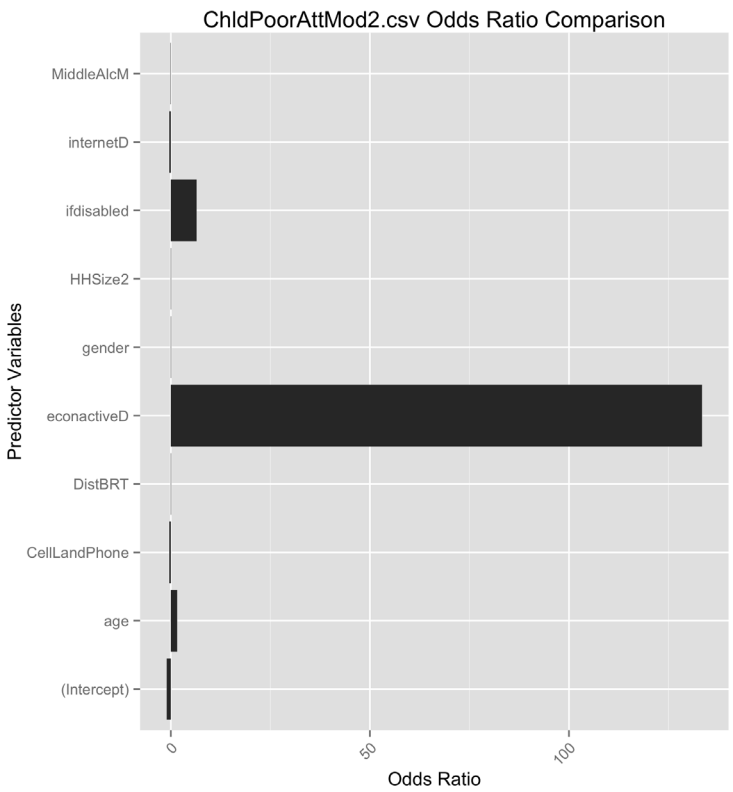


Table 1: Children, Poor Attendance, Coefficient Table, Model 1

		Coefficient	Significance	ANOVA	VIF	OR	OR Confidence Int	
	.rownames	estimate	p.value	p.value	VIF	0.0000	2.50%	96.50%
Model 1: ChildPoorAttMod	(Intercept)	-17.24	0.00	NA		0.00	0.00	0.00
	age	0.95	0.00	0.00	1.09	2.58	2.51	2.65
	gender	0.09	0.00	0.00	1.02	1.10	1.03	1.17
	internetD	-0.51	0.00	0.00	1.13	0.60	0.56	0.65
	ifdisabled	2.01	0.00	0.00	1.01	7.44	5.81	9.48
	MiddleAlcM	-0.07	0.01	0.00	1.01	0.93	0.90	0.96
	DistBRT	0.05	0.00	0.65	1.02	1.05	1.02	1.08
	HHSize2	0.10	0.00	0.00	1.00	1.10	1.08	1.13
	CellLandPhone	-0.50	0.00	0.04	1.16	0.61	0.56	0.66
	econactiveD	4.90	0.00	0.00	1.16	134.40	121.00	149.55

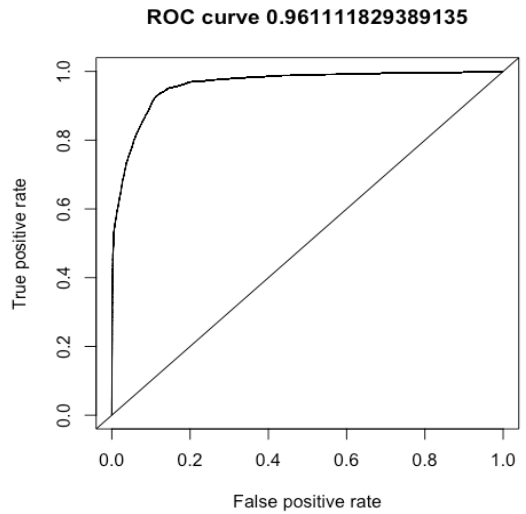
McF R2 =	0.57

	0	1	Accuracy	Sensitivity	Specificity
0	124158.00	3701.00	0.97	0.99	0.51
1	795.00	3924.00	0.97	0.99	0.51

Probability basis = 0.30

Deviance	**Residuals**	:		
Min	1Q	Median	3Q	Max
-2.80	-0.19	-0.05	-0.01	4.78

Figure 13: ROC Curve, Model 1



the effect was small. Relative to other developing countries, Indonesia has a high sense of educational equity between the sexes, at least among earlier ages, which plays out in interesting ways across these models. Interestingly, the gender trend reverses for the more absoltute measure of never attending school. This suggests that females persist to acheive some level of education at some point in their childhood, but they have more trouble remaining in the educational system than males.

Age and *household size* appear to have odds ratios of greater than one, meaning that an increase in household size or age is associated with an increased probability of stopping school. These factors compound in some instances, as seen in Figure 14. The probability of a 17 year old dropping out of school appears to grow the larger the household they

are in. This might be in part because large households reflect institutional housing of some kind, or that older children are entering group living situations like Kosts. This same increase is not seen among 6 year olds, who have a relatively low probability across household size.

The variable built using the *number of children per school* appeared significant. The proportion of *students near middle schools* provided significant influence for the model more than the other school service access variables considered. The implication is that the higher the number of students by schools, the less general infrastructure available, and that contributes to poor attendance. Though areas less served by schools might be economically disadvantaged in other ways.

The only particulary strange finding, was that *Internet access* consistently increased the odds of not attending school. This is perhaps the only factor in this analysis whose implications are unclear, and there may be some confusion around the variable's interpretation of a "yes" response.

Figure 15: Age vs Predicted Probabilities, Blue Reflects a Female Observation

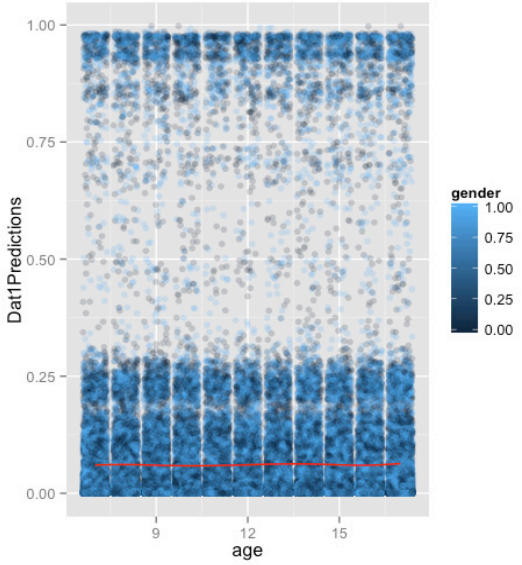


Figure 17: Profile of a 17 Year (red) and 6 year old (blue), Everything Else Held Constant Over Predicted Probabilities vs # of Children per Middle Schools

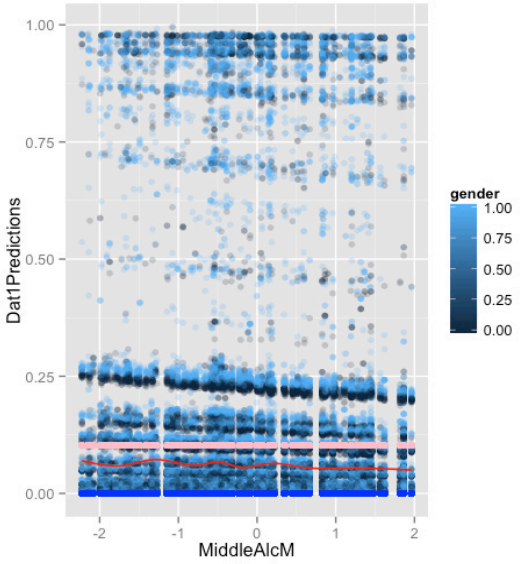


Figure 16: House-hold Size vs Predicted Probabilities of Having Poor Attendance

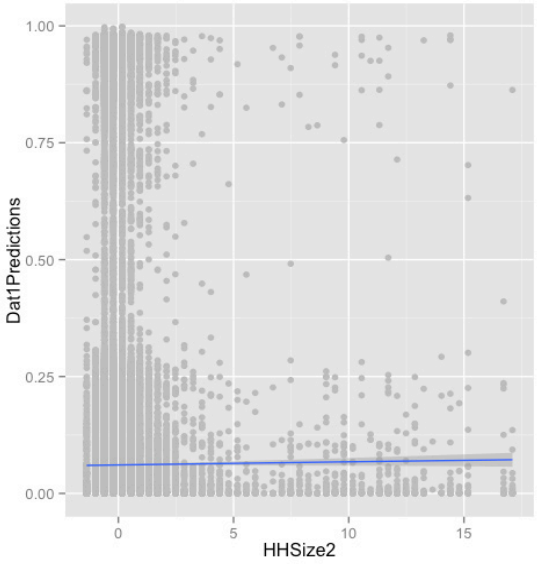
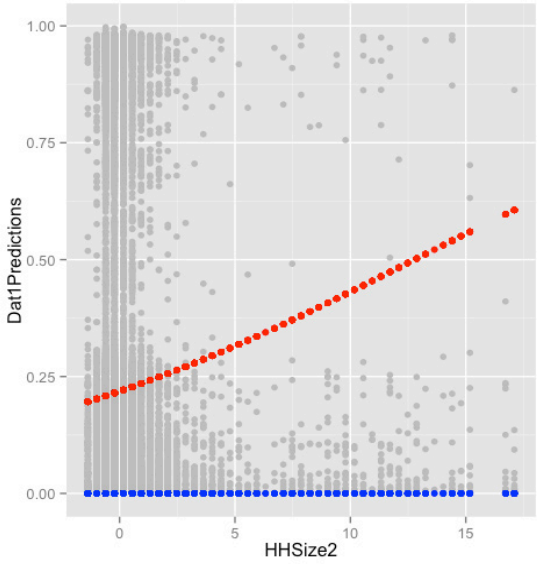


Figure 14: Profile Of A 17 Year (Red) And 6 Year Old (Blue), Everything Else Held Constant Over Predicted Probabilities Vs Household Size



Model 2 – Never Attended School / Children Only

$$Ln\left(\frac{P(ChildrenHavingNeverAttended\ School)}{1-P(ChildrenHavingNeverAttended\ )}\right) = \beta_0 + \beta_1 age + \beta_3 internet + \beta_4 if Disabled + \beta_5 \text{HouseholdSize} + \beta_6 \text{ifFemale} + \epsilon$$

The second model examining the incidence of children never attending school had the least number of significant variables. This is in part because the number of affirmative observations was so few. This made the model difficult to build, especially using random samples, which sometimes created different implications. Because younger children also appear under-reported in the census it is possible that there is omission bias in that those who would most likely not attend school are those that would also be unlikely to be reported at all.

The most important finding is the significance, explanatory power, and consistency of people being indicated as disabled. This variable was built aggregating responses among four survey responses relating to disability, such as difficulty hearing or seeing. Clearly being disabled is the main reason why students are kept from school, which suggests that infrastructure to accomodate for differently abled people is missing.

Because the affirmative responses were so small, as random sets were taken from the main dataset, this model was subject to the most inconsistencies. There were only a few hundred affirmative responses in any of the random sample-sets. Despite this, the few variables included were highly significant across all sampled subsets of the data. Given that there are more opportunities to enroll the older a person is, it makes sense that age would have an negative coefficient and consistent significance.

Strangely the use of Internet was a consistently significant factor across this model as well, always signaling a negative relationship with the dependent variable. This suggests that more access to Internet meant the student was more likely to have never attended school. This is an interesting factor, it might be due to a few things. My best guess is that it reflects some level of error bias in the model. Internet access reflects more education and status and those with out such access might be also the ones underreporting their children's presence. Relatedly, those who are disabled may also be in families who would be more open and understanding of their needs and more likely to identify them as disabled and not fear reprisal for keeping them from the government for keeping them out of school. Nevertheless the finding is strange and the justification unclear.

As shown in the adjacent chart, there is a general downward trend in people having never attended school, suggesting that people attend more as they get older, everything else held constant. Interestingly, there is a consistently higher probability of men never attending

Figure 18: Age vs Predicted Probabilities, Blue Reflects a Female Observation

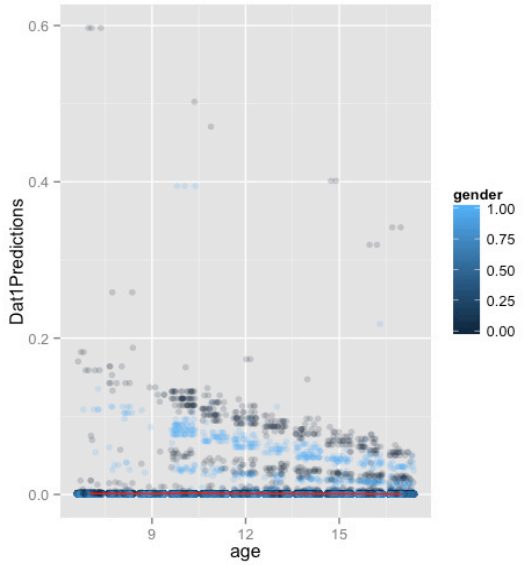


Figure 19: House-hold Size vs Predicted Probabilities, Blue Reflects a Female Observation

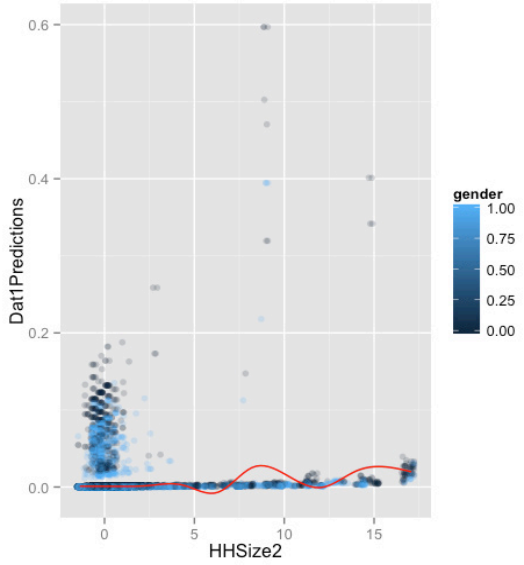


Figure 20: Odds Ratio Comparison

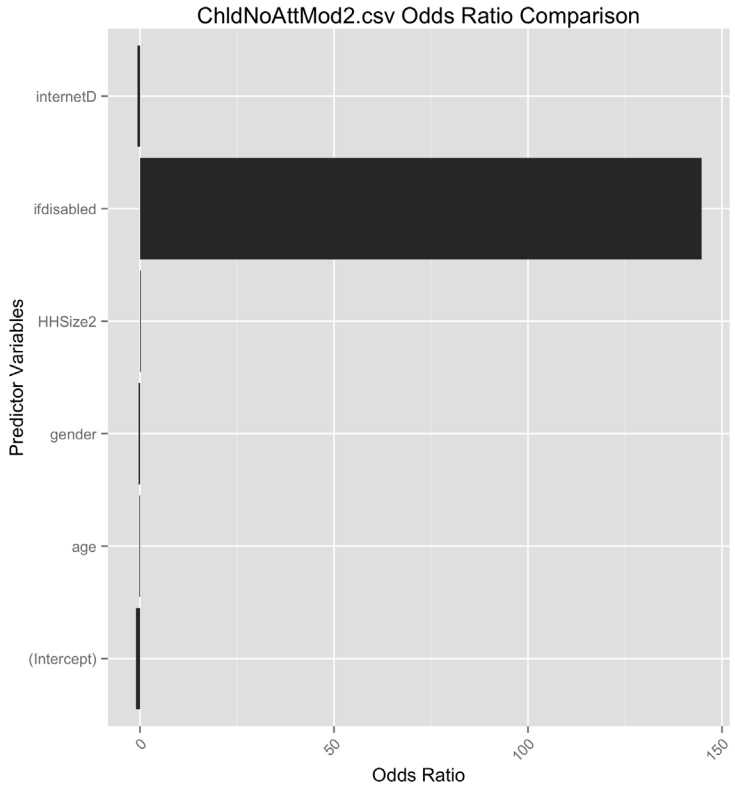


Table 2: Children, Never Attended, Coefficient Table, Model 2

		Coefficient	Significance	ANOVA	VIF	OR	OR Confidence Int	
	.rownames	estimate	p.value	p.value		0.0000	2.50%	96.50%
Model 2: ChldNoAtt Mod	(Intercept)	-5.72	0.00	0.00		0.00	0.00	0.01
	age	-0.12	0.00	0.00	1.05	0.88	0.83	0.93
	internetD	-0.92	0.00	0.00	1.01	0.37	0.22	0.58
	gender	-0.47	0.01	0.01	1.04	0.65	0.45	0.91
	ifdisabled	5.01	0.00	0.00	1.01	145.70	102.32	207.06
	HHSize2	0.06	0.00	0.00	1.03	1.24	1.16	1.31

McF R2 =	0.24	Deviance	Residuals:			
		Min	1Q	Median	3Q	Max
		-1.01	-0.04	-0.03	-0.02	4.27

Figure 22: IfDisabled vs Predicted Probabilities, Blue Reflects a Female Observation

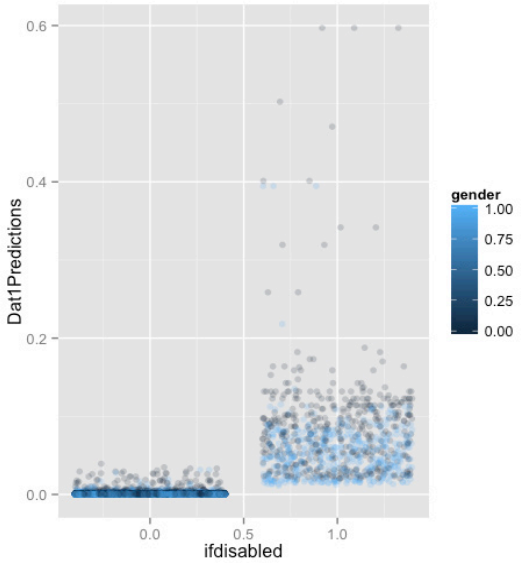
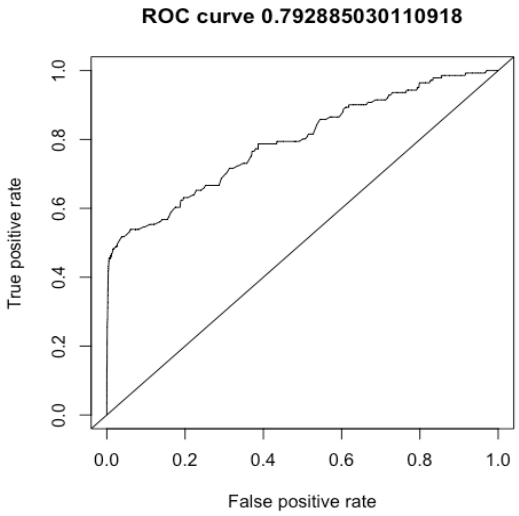


Figure 21: ROC Curve, Model 2



school than women, though this converges with age.
The output examining household size vs predictability is difficult to interpret. There is significant variation across sizes of households in terms of educational achievement, likley representing different types of households, including institutional or group living situations, such as Kosts, which abound in Semarang.

Reflecting on Figure 22, while being female appears to increase the overal probability of never attending school, it appears that males who are disabled are actually more inclined to have never attended school, and that rate declines over time, and in some small part the larger the household they are in. Disabled children though are the majority share of the individuals who never end up attending school.

The relatively lower ROC area under the curve is a reflection of how few affirmative observations there were.

Model 3 – Never Attended School / Total Population

$$Ln\left(\frac{P(ChildrenHavingNeverAttended\ School)}{1-P(ChildrenHavingNeverAttended\ )}\right)$$
$$= \beta_0 + \beta_1 internet + \beta_2 economicallyActive + \beta_3 ifDisabled + B_4 ForeignBorn + B_5 MiddleSchoolStudentBurden + B_6 \#ofHouseHoldMembers + B_7 DistHighSchools + B_8 DistanceToBRT + B_9 SqMtsofHome + B_{10} FemaleandMarried + B_{11} IfRented + B_{12} IfLeased + B_{13} CellPhoneandLandPhoneAccess + B_{14} PoorToiletFacilities + B_{15} \text{WaterSourcePumped} + B_{16} ElectricityUsedNonMetered + \epsilon$$

The third model had many variables but the least level of explanatory power. The dependent variable was again those *never having attended school*, but the dataset used reflected 25% of the whole population under 60. Because there are so many observations, the significance levels of the coefficients and the ANOVA test using the chi-square distribution was high, with all variables near-zero p-values. Despite the many variables, none of those in the final model exhibited multicollinearity issues, and no further variables were suggested to be removed by the stepwise regression method. That said, the model's fit, as a function of its McFadden's R^2 was only .13, though the area under the ROC curve covered 80%.

Again, *disabled individuals* had the highest probability of *never attending school*. Secondly the proportion of those who were *married and women* were of a much higher probability of having never attended school. *Gender* did not play as large a role in the child-subset data, but it appeared to play a much larger role in the total population data, at least in regards to the number of people who never attended school. Surprisingly, males were more likely to have *never attended school* than women. Interactions of *gender* with *being disabled* and being married appeared to greatly improve the model's explanatory power. The other factors considered reflected the socio-economic conditions of the respondent. Most notably the *housing floor type used as soil* variable was a significant indicator for people having never attended school. This, coupled with *no access to a flush toilet* or having to *pump your water from a ground well*, all reflected more informal settlement style living standards. Those who had *access to both a cellphone and a land line*, as well as the *Internet*, were likely to have attended school at some point in their life. The *access to internet* factor appeared to align with socio-economic factors among the larger population, but not in the children subset, strangely. Those who *rented* or *leased* were less likely to have never been to school. Housing law in Indonesia exists in a grey zone on a number of fronts. A large portion of slum dwellers have some claim to their land and consider themselves owners, though clearly many upper class members also own their own homes. The ROC curve for the model had a smoother area underneath it than Model 2 because there were more affirmative observations within the dataset.

Interestingly, as the floor area of a respondents home size decreases the likelihood of never attending school decreased, while household member increases was associated with an increase in having never attended school. This makes sense as large households are associated with wealth, and larger families are not. It might be worth taking a ratio of these two variables

Figure 23: Female & Married vs Predicted Probabilities,

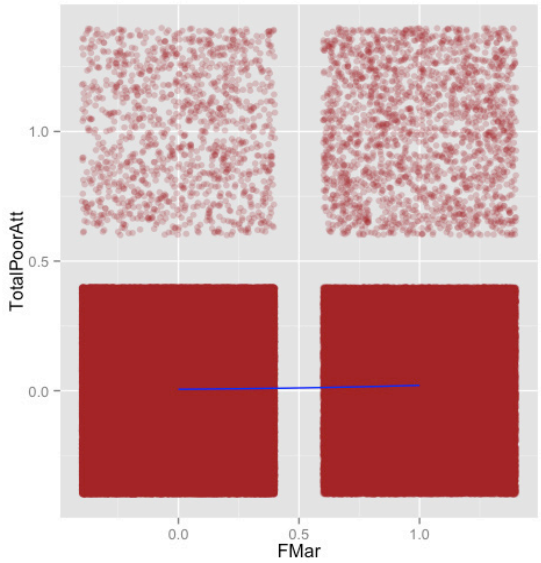


Figure 24: IfDisabled vs Predicted Probabilities

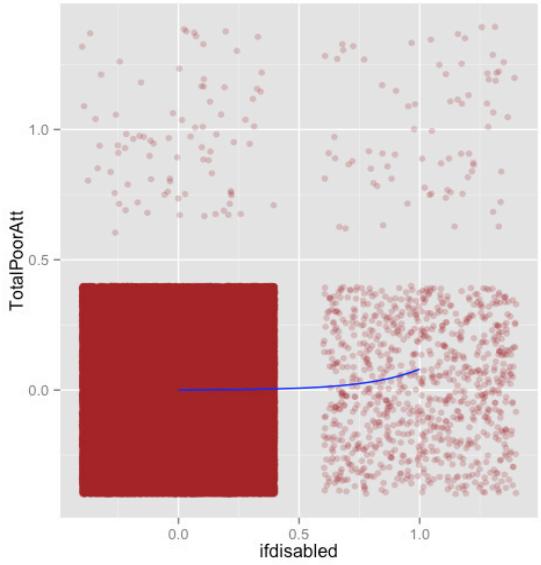
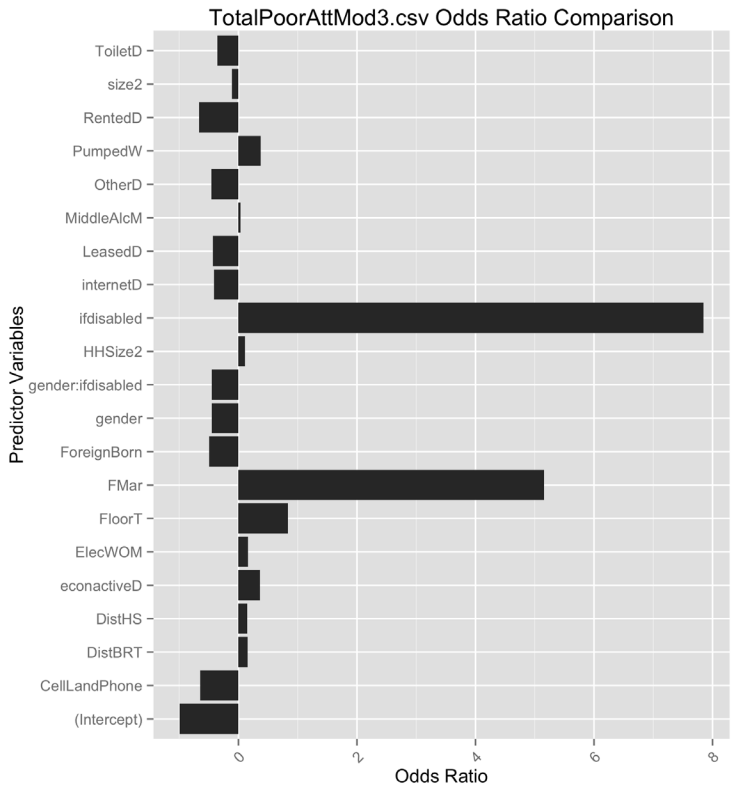


Figure 25: Odds Ratio Comparison



to create a new interaction variable for future models.

It would make sense that the distance to highschools and the distance to the BRT have similar associated probabilities since the high schools are considered pre-professional, have a regional draw and are located at central commercial points along the BRT.

It is important to remember that this model has a low R2, and as such implications are limited. That said, the variables were consistently significant across multiple random datasets drawn from a comprehensive dataset. While the variables included only explain a small proportion of the variation witnessed in the data, the model provides a small, but significant glimpse into the critical factors associated with a lack of educational attainment.

Table 3: Total Population, Never Attended, Coefficient Table,

		Coefficient	Significance	ANOVA	VIF	OR	OR Confidence Int	
	.rownames	estimate	p.value	p.value		0.0000	2.50%	96.50%
Model 3: TotalPoorAttMod	(Intercept)	-4.84	0.00	0.00		0.01	0.01	0.01
	internetD	-0.54	0.00	0	1.07	0.58	0.53	0.65
	econactiveD	0.38	0.00	0	1.01	1.46	1.36	1.57
	ifdisabled	1.83	0.00	0	1.01	6.23	5.49	7.04
	ForeignBorn	-0.67	0.00	0	1.02	0.51	0.42	0.62
	MiddleAlcM	0.04	0.04	0	1.24	1.04	1.00	1.08
	DistHS	0.14	0.00	0	1.43	1.15	1.10	1.19
	DistBRT	0.14	0.00	0	1.41	1.16	1.13	1.19
	size2	-0.12	0.03	0	1.24	0.89	0.83	0.94
	FloorT	0.61	0.00	0	1.19	1.83	1.63	2.05
	HHSize2	0.10	0.00	0	1.10	1.10	1.06	1.15
	FMar	1.28	0.00	0	1.02	3.61	3.36	3.89
	OtherD	-0.60	0.00	0	1.10	0.55	0.48	0.63
	LeasedD	-0.56	0.00	0	1.06	0.57	0.49	0.66
	RentedD	-1.14	0.00	0	1.06	0.32	0.24	0.42
	CellLandPhor	-1.04	0.00	0	1.15	0.35	0.31	0.41
	ToiletD	-0.44	0.00	0	1.27	0.64	0.59	0.71
	PumpedW	0.32	0.00	0	1.03	1.38	1.25	1.52
	ElecWOM	0.15	0.01	0	1.21	1.17	1.04	1.31

McF R2 =	0.13	Deviance	Residuals:			
		Min	1Q	Median	3Q	Max
		-1.18	-0.15	-0.10	-0.07	3.95

Figure 27: IfDisabled vs Predicted Probabilities, Blue Reflects a Female but Married Observation

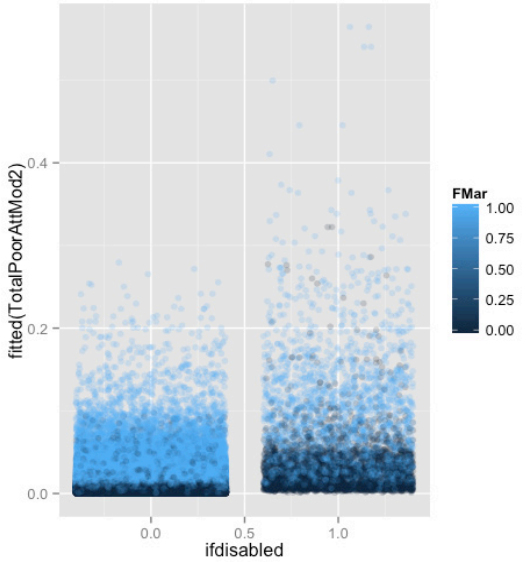
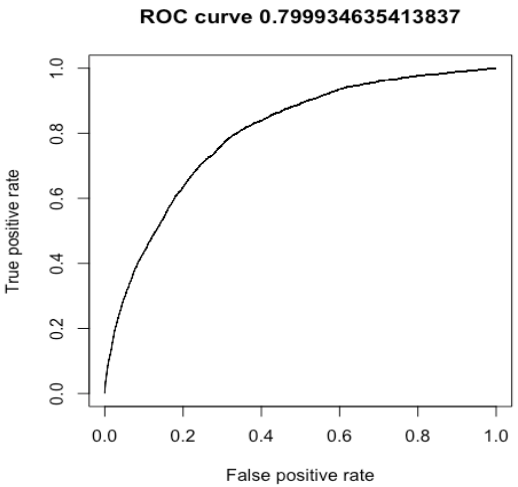


Figure 26: ROC Curve



Model 4 – Educational Achievement / Total Population

$$EducationLevelAChieved = \beta_0 + \beta_1 internet + \beta_2 economicallyActive + \beta_3 ifDisabled + B_4 ForeignBorn + B_5 a_{\epsilon} + B_6 \#ofHouseHoldMembers + B_7 SeekingWork + B_8 RecentlyPregnant + B_9 SqMtsofHome + B_{10} gender + B_{11} IfRented + B_{12} CellPhoneandLandPhoneAcce + B_{13} PoorToiletFacilities + B_{15} \text{WaterSourcePumped} + B_{16} NonMeteredElectricityA + B_{17} TotalPopDen + B_{18} FloorTypeSoil + B_{19} EconomicallyActive + B_{20} ElectricCool + \epsilon$$

The fourth model is the broadest in implications and is built using the linear OLS regression method. It is important to realize that the level of educational attainment variable created is not partitioned into an explicitly compatible continuum. Nevertheless, abstractly, this model helps provide an understanding of the factors that are associated with an increase in educational attainment and their general direction, clear direct shift implications are limited, though the charts of the data are illuminating. Since nearly all individuals completed at least primary school, and most completed highschool(4), coefficients below 1 suggest that the variable has relatively low predictive power.

The educational achiement dynamics at play is demonstrated by the plot in Figure 28, which shows *age* against *educational level achieved.* As expected, *never attending school,* level 0 in the chart, appeared to increase as you get older overall, and the strongest concentration of people with postgraduate degrees are between the ages of 35 and 50. After 50 overall attendance is low as well as post graduate attendance because available educational infrastructure was only modernized in the middle of the last century. Clearly, those who chose to not respond was greatest among the elderly and most likely reflects people who did not achieve any level of education. The gap in primary educational achievement reflects individuals who have moved on to a greater level of education, though older populations have the largest share of the population with only primary education. These contradictory trends in the data is why age as a factor alone is not a more powerful indicator of educational achievement. The model's R^2 jumped from .24 to .33 after accounting for the dummy variable, over 45, and the direction of all of the models shifted positively. This is because nearly everyone achieves at least the most basic level of education. While technically a better model the contradictions across age groups are still not entirely accounted for because older groups are both more likely to have the highest level of education, but older populations are also more likely to have never achieved any level of education.

As shown in the previous models, being male seemed to significantly increase the probability of the extreme case of never attending school. But across the total population, considering all age levels, *being female* decreases the overall acheivement level. Though there seems to be a slightly greater proportion of females attaining college in the 20 to 25 range, while there seems a disproportionate number of older women without education. After including a variable for the interaction of female and married though, this coefficient for *ifFemale* becomes positive, suggesting that those who are not married and female have an overall greater like-

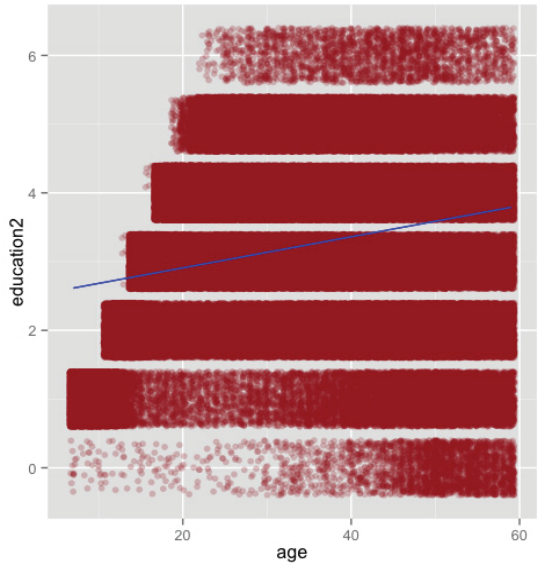


Figure 28: Age vs Predicted Probabilities

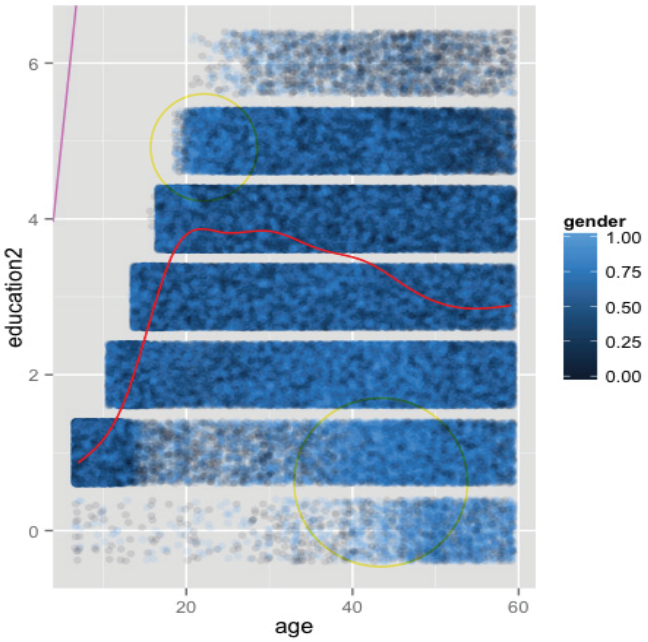


Figure 29: Age vs Predicted Probabilities, Blue Reflects a Female Observation

lihood of achieving a greater level of education than males. The interaction variable presents issues of multicollinearity though and was not included in the final model.

Confusingly, women who are *recently pregnant* have a greater liklihood of achieveing a greater level of education, while *being a woman* overall is less important. As shown in Figure 31, this is a reflection of the *age* dynamic at play, in that those who are more *recently pregnant* are more likely to be younger than older less educationed generations, but old enough to have achieved some level of education.

Similarly, those *looking for work* had a higher coefficient than those currently working, which may be due to a few factors, such as those looking for work are likely younger, and most younger people consistently attend school, additionally those who are still looking for work but not employed are those who have reason to believe they will find work, while those who aren't working but with less prospects have given up and would exist in the category of neither working nor looking for a job, which was not modeled.

Wealth indicators such as the *size of household*, *floor type, access to a cell phone,* and *if the household pumped water* were expectedly highly associated with educational achievement. Interestingly, being born in a different city suggests that the individual is more likely to achieve a greater level of education. This contradicts the general sentiment in Semarang that the incoming migrants are mostly rural poor migrants of lower socio-economic status.

The focus of this study was primarily school-aged children but suggested future studies relating to the whole population should subset the analysis by clear age groups to understand the dynamics occuring within subpopulations. These last two models were subsetted to only considering individuals under 60, but the dynamic of older folks being more likely to not have had any level of education persists across age groups above 25. Of most interest might be the age group directly after the college age, and future studies could focus on college achievement within this population.

Table 4: Total Population, Educational Achievement, Coefficient Table

		Coefficient	Significance	ANOVA	VIF
	.rownames	estimate	p.value	p.value	
	(Intercept)	-0.02	0.00	0.00	
	age	0.05	0.00	0.00	2.977878
	gender	-0.08	0.00	0.00	1.101742
	abv45	1.50	0.00	0.00	2.398055
	TotPopDens	0.00	0.00	0.00	1.025899
	internetD	0.39	0.00	0.00	1.141733
	RecentPregD	0.61	0.00	0.00	1.041186
	seekingworD	0.86	0.00	0.00	1.037975
	econactiveD	0.46	0.00	0.00	1.503391
	ifdisabled	-0.18	0.00	0.00	1.022838
Model	ForeignBorn	0.37	0.00	0.00	1.027199
TotEduMod	size2	0.07	0.00	0.00	1.284605
2	FloorT	-0.48	0.00	0.00	1.058987
	HHSize2	-0.06	0.00	0.00	1.105598
	RentedD	0.61	0.00	0.00	1.158622
	CellLandPhor	0.48	0.00	0.00	1.283729
	NoDispFac	-0.21	0.00	0.00	1.050867
	ToiletD	0.06	0.00	0.00	1.291998
	ElecCook	0.13	0.00	0.00	1.008496
	PumpedW	-0.17	0.00	0.00	1.021481
	ElecWOM	-0.19	0.00	0.00	1.157531

Adj	R2	33.00	=	Deviance	Residuals	:		
				Min	1Q	Median	3Q	Max
				-2.80	-0.19	-0.05	-0.01	4.78

Figure 31: IfDisabled vs Predicted Probabilities, Blue Reflects a Female Observation

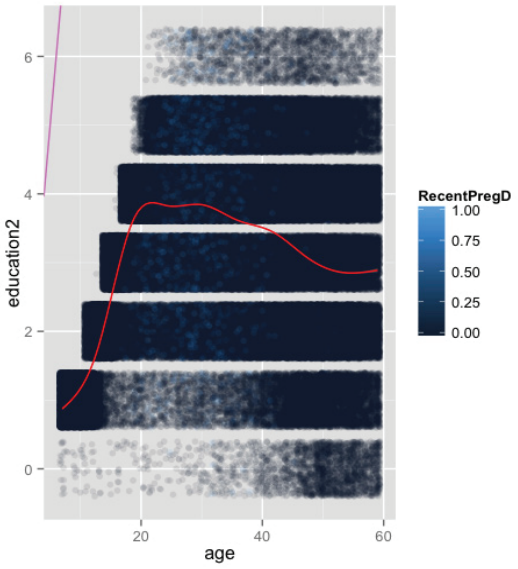


Figure 30: House-hold Size vs Predicted Probabilities