Are Students Using the Right Criteria When Choosing Schools?

An analysis of whether college matriculation rate is related to city crime rates

Sahil Patel[1] Robbie Ferrand[1] Ben Bowles[1]

[1]Department of Statistics, North Carolina State University

April 21, 2024

Abstract

Given the evidence that crime rates affect the learning ability of middle and high school students, this analysis focuses on trying to determine whether students across the US are using crime rates when picking a university or college to attend. Using a university's matriculation rate as a response and crime rates as covariates, we fit a beta regression, first without and then accounting for spatial autocorrelation. Given the evidence that the impact of crime rates may vary for universities with different matriculation rates, we lastly fit a quantile regression to show that only a small range of universities are marginally affected by the area's crime rate. This trend speaks to the apparent lack of student care in crime rates in the university's town or city, and given the importance of local crime in learning, students should be made more aware of the importance of crime rates in the colleges they plan on attending.

Keywords: Beta regression, spatial modelling, quantile regression, crime rate, university matriculation, college, pinball error, student, United States

1 Introduction

Many factors influence how students choose the university they attend. Whether it be the institution's cost or the campus's feel, most students have many criteria to evaluate where they plan on spending their first lengthy period away from home. A recent survey in 2022 collected from then-current college students asks them to analyze why they chose the school they did (Ezarik, 2022). Most students cited that they chose to attend their school because the program they were interested in was offered at the university. Second to this was the university's academic reputation, followed by other facts that include proximity to home and the feel of the campus. With the first two focused on the university's academics, it is easy to see that most students care about their education when picking a university. However, if students want to ensure they get a good education from their school, are there not factors outside of the classroom that would affect how well students learn through the year?

It has been shown in middle school students that the more unsafe they feel in and around their learning environment, the worse they perform academically (Lacoe, 2020). This trend was also noticed by Burdick-Will et al. (2011) when looking directly at crime rates. This has been further explored, and it has been noticed that crime rate even affects children's cognitive abilities (Sharkey, 2010). Consequently, it is natural to see that the safety of the environment a student is living in could have a major impact on a student's academic success. Crime naturally also links to stricter policing, but stricter policing has also been shown by Legewie and Fagan (2019) to have a negative impact on students' ability to learn. Moreover, they showed that this was disproportionately affecting minority children.

1.1 Research Question

As a consequence of the ever-present impact of crime on students' ability to learn and be successful in elementary, middle, and high school, we hoped to understand whether students are using the crime of the area where their university is located in their decision to attend said university. If students are not using crime rates in the city they hope to attend, then more work needs to be done to better inform students about the impact of non-academic factors outside of the school that affect their education. This should help both students succeed and the university administration to understand what external factors contribute to their varying matriculation rates.

2 Data

2.1 Covariates, Collecting, & Formatting

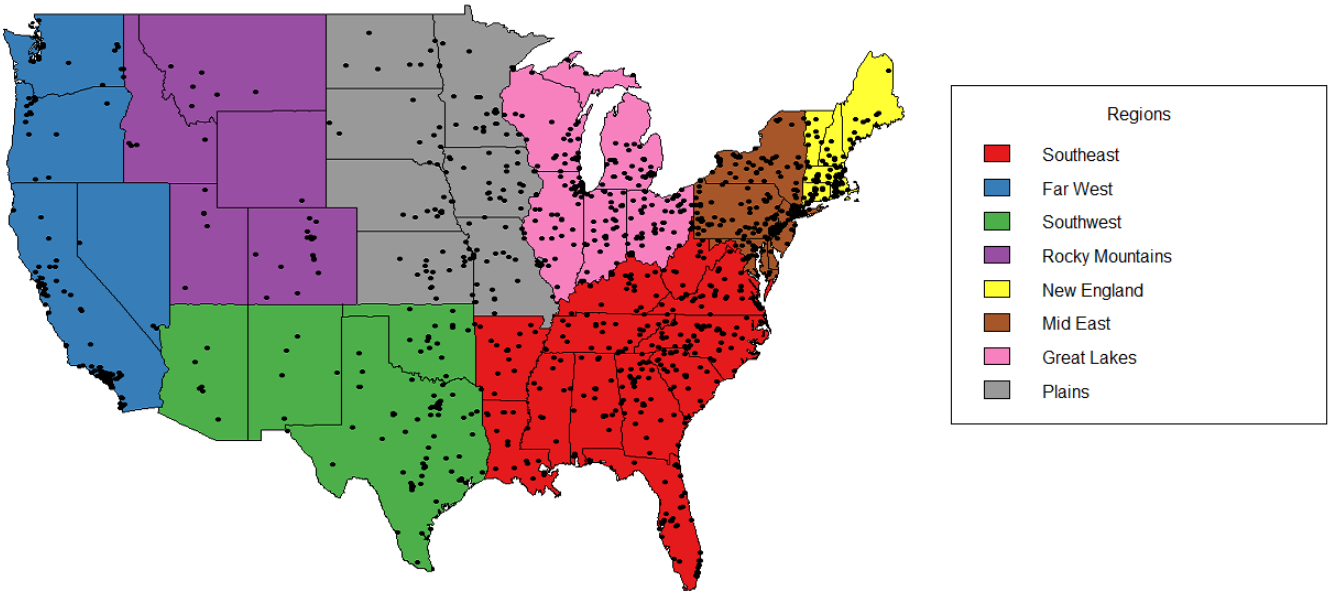

Figure 1: State map including data points for every college

To analyze whether a student is using crime as a factor in attending school, we compare the school's matriculation rate against the crime rates in the city where the school is located. The matriculation rate is a useful response in this circumstance, as it would give us information about students who were accepted into the school but eventually decided to attend (or not) based in part on, in this case, crime. Matriculation rate can be more explicitly defined as

$$\text{Matriculation} = \frac{\text{Total Students Enrolled}}{\text{Total Students Admitted}}$$

Naturally, there are a lot of confounders in this problem. It is reasonable to believe many factors go into a student's decision to attend school. We include acceptance rate as a predictor as it is highly correlated with the prestige of a university (Schmitz, 1993), and it has been shown (Ezarik, 2022) that the prestige of the university generally factors into how much someone will want to attend a university. Other factors like cost can influence a student's decision, but the prestige of the university would be one of the largest drivers of this. We then include city crime data and admission rates to try modeling matriculation rates.

To get the university's location and matriculation rates, data was collected from the Integrated Postsecondary Education Data System (IPEDS, 2021) from 2021. The data contains the number of students who were accepted and the number attended the university in 2021. Taking the ratio of these two quantities

gives us our matriculation rate. The latitudinal and longitudinal coordinates, city name, state name, and university admission rates are also included in this data set. We use the inverse of admission rates as a transformation. This is a consequence of admission rates being inversely associated with university prestige (Schmitz, 1993), and since admission rates are between 0 and 1, the inverse of admission rates would be positively associated to a university's prestige, and thus be positively associated with a school's matriculation rate.

The city and state names are used as a reference from the collegiate dataset and the FBI crime known to law enforcement (FBI, 2021) so that city crime rates can be adjoined to the college the city is located in. Number of cases of burglary, larceny, vehicle theft, murder, rape, robbery, and aggravated assaults are recorded in the FBI database. Number of cases was converted to rate per 100,000 by dividing the number of cases by the city population (included in the dataset) then multiplying by 100,000. We look for some type of relationship between these crimes and matriculation rates.

2.2 Exploratory Data Analysis

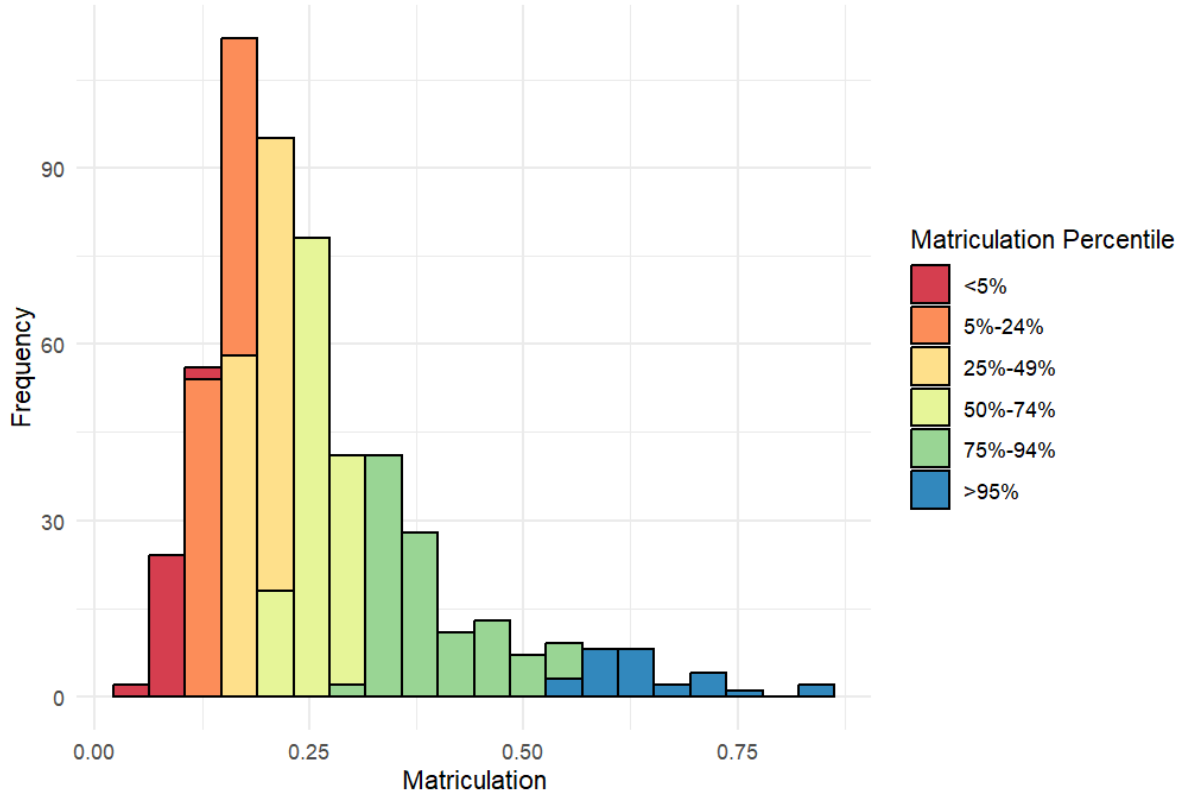

Figure 2: Distribution of Response Variable Matriculation Rate

Our variable of interest, matriculation rate, is a right-skewed distribution as per Figure 2. In particular, we notice that the top 5% of matriculation rate covers nearly 30% of values. The majority of the data is centered around $15\% - 30\%$, which appears to be the general IQR for US college matriculation rate. Note that the data has been divided into specific quantiles, which better display the spread of the data. The supplemental figures give distributions by percentile in Figure 5. Based on this data, we posed the research objective of dealing with the sharp right skew, whilst also keeping prediction intact for the rest (majority) of the data. As mentioned previously, we will be investigating how various crime rate variables and admission rate can explain and predict matriculation rate.

From Figure 3, we see that many of the crime variables are similarly distributed to the response variable: with a strong right skew. This implies that whilst the majority of the US is relatively tame with

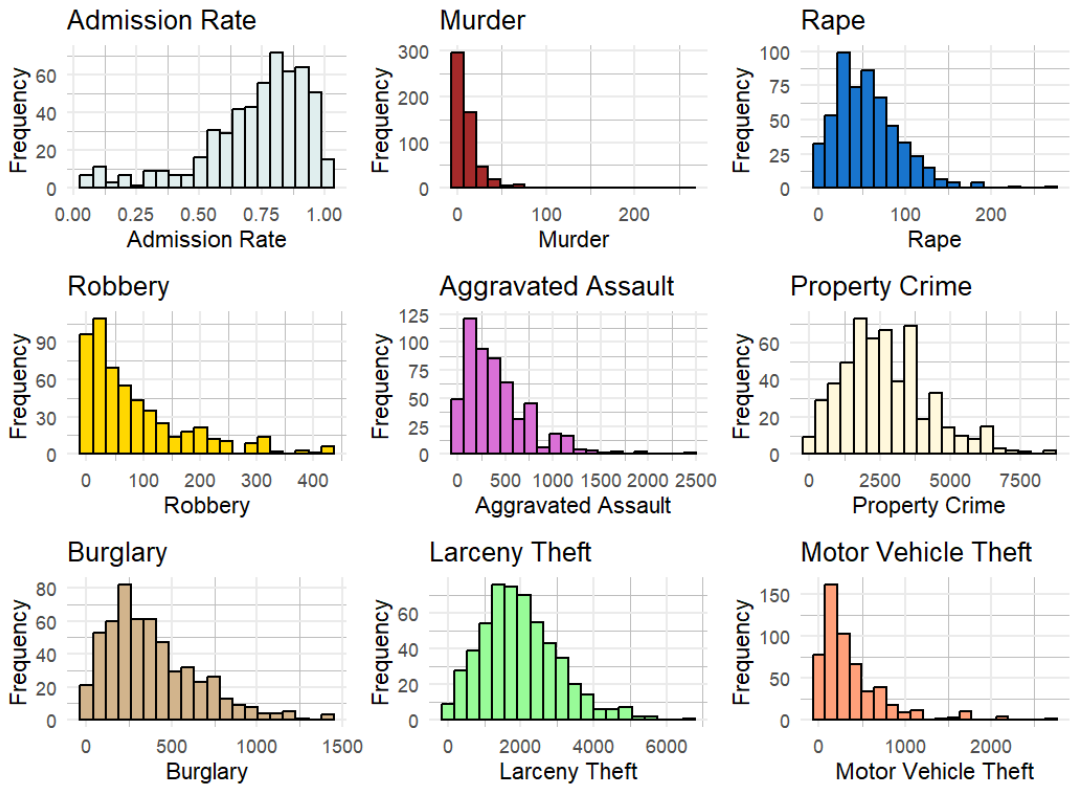

Figure 3: Distribution of Predictor Variables

regards to crime rate, there are a few areas that exhibit disproportionately high crime rates compared to the rest of the US. Otherwise, admission rates appear to be the opposite skew from the response. Most colleges appear to have an admission rate over 50%, whereas the left tail represents more selective or prestigious colleges. Our next investigation was into the correlations of these predictor variables, and to see if we are working with highly correlated data.

As far as variable transformations go, many of our predictor variables found no significant evidence requiring transformation. If we transformed one crime variable, we would likely need to transform all of the crime variables due to the similarity in distribution, as well as similarity to one another. We may consider transforming the response via logit, to account for the skew and to make it Gaussian. The admission rate variable will be transformed to the inverse in our beta and spatial models, as it was significant and will make the methods section easier to interpret.

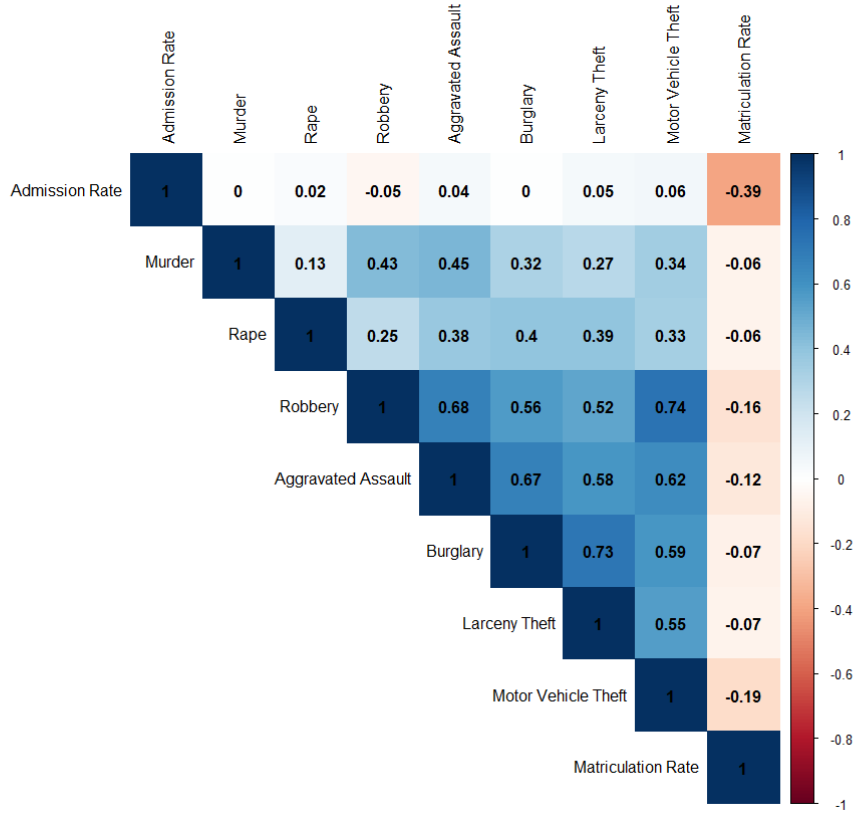

Figure 4: Correlation Plot of Predictor Variables

Figure 4 shows us that we have moderate to strong positive correlations in most of our crime data. That is, if there is a frequent occurrence of one crime, we generally would expect there to be more frequent occurrences of other crimes too - which extends to both violent and nonviolent crime. In our methods, we plan to analyze the possibility of spatial relationships of our crime rate variables because of this correlation.

3 Methods

3.1 Naive Approach

Our "naive" approach was to fit a beta regression model to the data. We were interested in this kind of model since matriculation rate is a continuous quantity between zero and one, which would not be suitable for models such as logistic regression that need a categorical response. Further, a Beta(α, β) distribution's density can take many different shapes for different values of the parameters α and β, which we believe will be useful for data that is right-skewed like ours, as seen in Figure 2. Recall that the probability density function of the beta distribution is often expressed as $f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, y \in (0,1)$, with $E(Y) = \frac{\alpha}{\alpha+\beta}$. We can slightly alter this density according to Ferrari and Cribari-Neto (2004)'s parameterization as follows: set $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha + \beta$, and then our density becomes $f(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, y \in (0,1)$ (Cribari-Neto and Zeileis, 2010). In this parameterization, $E(Y) = \mu$ and $\frac{1}{\phi}$ is a dispersion parameter. From here, much like other nonlinear regression methods, we set $x_i^T \beta = g(\mu_i)$, where g is the link function, the choice of which is left to the user. This gives us easy interpretations for the coefficients of β: e.g., under a logit link, a positive β

coefficient means that the expected value of our response increases as the value of the relevant predictor increases.

Beta regression begins with a few assumptions: First, our response variable follows a beta distribution. Second, we can model the mean response by independent linear predictors passed through a choice of link function (Cribari-Neto and Zeileis, 2010), and finally, no significant outliers in the data.

To test whether the data follows a beta distribution and the appropriateness of our link function, we will analyze diagnostic plots of the model. Plotting true versus predicted values, we should notice a roughly $y = x$ pattern among the points. Plotting linear predictors versus standardized residuals, we should see residuals that remain centered around zero and display minimal curvature around the $y = 0$ line. Finally, plotting generalized leverage versus predicted values, we should observe a roughly horizontal pattern. To check for the independence of our predictors, we first note the large amount of multicollinearity among the crime predictors in Figure 4, and we conclude that our predictor variables are not independent. To address this, we will instead predict using the principal components of the crime predictors to remove this multicollinearity. Next, we will deal with any outliers in our data as follows: if a data point has a standardized residual with a magnitude greater than 3 upon fitting the first beta model, we will remove the point from our training data and refit a new model. To choose the best number of principal components to work with, we will look at a plot and see if there are any large drops in training MSE. We will use the smallest number of principal components after which there is not a notably large drop in training MSE. To choose our link function, we will choose the link function that gives us the best training Pseudo-R^2 after removing outliers and choosing the best number of principal components.

3.2 Accounting for Spatial Autocorrelation

Considering that crime itself is reasonably suspect of spatial autocorrelation, it is reasonable to believe that a student might avoid attending school as a consequence of the crime in a nearby city. Consequently, the concept of beta regression can be extended into spatial regression. A spatial model takes the following form:

$$g(\boldsymbol{u}) = \tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

For our original beta regression model to be used, spatial covariance must be present. To verify this, a variogram is plotted on the original generalized linear model (the beta regression) residuals. A variogram plots semivariances over distance. This is done by collecting points within some distance of each other and calculating the variance. As distance increases, a non-spatial model would have a constant variance, but a spatial model would have some kind of pattern present in the variogram as distance changes.

A consequence of proceeding this route is the lack of support for different link functions in existing packages. Thus, we transform the variables using the link function determined in the original beta regression, denoted $\tilde{\boldsymbol{Y}}$, to fit what essentially is a normal linear regression onto $\tilde{\boldsymbol{Y}}$. Moreover, $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{0}, \sigma_s^2 \cdot \boldsymbol{R})$ represents a spatially co-varied random effect, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_e^2 \cdot \boldsymbol{I})$ represents random model error. In this procedure, we assume a common exponential decay of spatial covariance. Under this assumption, the following function denotes the spatial covariance structure.

$$\boldsymbol{R} = exp\left\{\frac{-\boldsymbol{M}}{\phi}\right\}$$

Where $\boldsymbol{M}$ is a matrix of the distance between all the data points, and ϕ is the range parameter (fit in the modelling phase) that controls how far university matriculation rates can be correlated with each other.

Given the multicollinearity found in the exploratory data analysis between crime rates, we used principal components for crime rates (the same from the fitted beta regression). We calculated the RMSE for a

regression fit at each number of principal components to determine the number of principal components. A jump in the RMSE will be used as a threshold to determine the number of components.

Once the number of components is determined, the model residuals and Cook's distances over the training data are inspected to verify the model's usability. Points with large residuals and/or Cook's distances are inspected further and removed if necessary. Root-mean-square errors over the training and testing data are compared to ensure the generalizability of our model; if similar, the model is re-fitted on all the data for inference. Then, AIC is compared between the nonspatial and spatial regression to ensure improvement in adding a spatial term.

Once refit on the full data, the fitted spatial covariance function is inspected to ensure a practical spatial covariance was fit over the data; a reasonable cutoff is verified in the spatial regression as we don't expect there to be autocorrelation between matriculation rates after a far distance.

3.3 Quantile Regression

Primarily motivated by analysis of the distribution of the response variable, we figured that Quantile Regression may be a good method of choice. Our intent of quantile regression is to estimate the conditional median matriculation rate, given specified quantiles in our data. In the literature, quantiles are often done correspondingly to trends in the data and are not necessarily required to be equidistant (Koenker and Hallock, 2001). Our quantiles of interest, τ are

$$\tau_i = \begin{cases} .05, & \text{if } i = 1 \\ .25, & \text{if } i = 2 \\ .50, & \text{if } i = 3 \\ .75, & \text{if } i = 4 \\ .95, & \text{if } i = 5 \end{cases}$$

which were chosen based on the distribution of the response variable. Most notably, the τ_5 quantile represents a spread of about 30% matriculation rate. Our mission primarily relies on determining if slopes are more or less significant at specified quantiles, especially if they are different from τ_3, the median quantile. No variable transformations were applied, and everything was taken as is. Interpretations will be made with regard to specific quantiles, which gives us a regression equation of

$$\boldsymbol{Q_\tau(Y|X)} = \boldsymbol{X\beta_\tau} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Q_\tau(Y|X)}$ represents a conditional regression equation of Y given X dependent on quantile τ.

In the Quantile regression section, we will fit both a standard multiple linear regression model, as well as a lasso model. We will complete these using qr.fit, qr.lasso, and qr.lasso.cv in the R package quantreg. (Koekner, 2023). After verifying the conditions, we will first apply the base fit of quantile regression with the 5 listed quantiles. Then, using the ANOVA function, we will conduct tests of pairwise slope differences between all possible groups. This will give us some indication of whether or not there is evidence, with all the predictors, of differences in model slopes at specific quantiles. The primary focus for doing this is to evaluate the appropriateness of quantile regression in the context of our dataset.

Next, we will begin to lasso the model with cross-validation to find the most useful coefficients at specific conditional quantiles within the data itself. We will evaluate which predictors are most valuable at specific quantiles, and see if any variables are viable all the way through. This is a natural second step to fitting the base quantile regression, as now we are interested in the best variables for prediction based on the model behavior from regression. To evaluate the fits, at every τ we will calculate the "pinball" loss function, as denoted in the literature. This takes the form of

$$L_\tau(y_i, \hat{y_i}) = \begin{cases} \tau|y_i - \hat{y_i}|, & y_i - \hat{y_i} \geq 0 \\ (1-\tau)|y_i - \hat{y_i}|, & y_i - \hat{y_i} < 0 \end{cases}$$

Where $L_\tau(y_i, \hat{y_i})$ represents the "pinball" loss at specified values of τ (Steinwart and Christmann, 2011). This loss function acts similar to an L1 regularization (or absolute error), except it introduces a penalty tied to the value of τ. For higher values of τ, we will penalize under-predictions (positive residuals) much more than over-predictions (negative residuals) and vice versa. Loss can be calculated component-wise for each value of τ and we will determine the total amount of loss and average loss per observation in the context of matriculation rate.

Quantile regression is similar to OLS, in that the model assumptions are relatively Gauss-Markovian. However, we assume a slightly lighter set of conditions: Design matrix is non-singular, matriculation rate is continuous, the relationship between predictors and matriculation rate is linear, and observations are independent and random with a large enough sample size. In most cases, quantile regression is used when the usual linear models assumptions are violated - especially in a case of heteroskedasticity or equivalent (Waldmann, 2018).

Most of these conditions follow from EDA since they are relatively light conditions. However, the dimensions of the design matrix were singular and aliased with too many predictors. To make quantile regression work, we removed the "Larceny" variable, since it was dwarfed in predictive power by the "Motor Vehicle Theft" variable. After making this correction, quantile regression assumptions are satisfied.

4 Results

4.1 Beta Regression Results

Before choosing the number of principal components (PCs) and link function, we removed four outlier points from our training data using a logit link starting model. We then chose the number of PCs to work with. As seen in Figure 6, we notice a large drop in MSE after using just the first four PCs, so we chose a beta regression model using the first four PCs. We next need to choose link function from among the choices of the logit, probit, clog-log, Cauchy, and log-log links. The Pseudo-R^2 values we got with different link functions are given in Table 6. With a training Pseudo R^2 of 0.301, we see that using a log-log link appears to be the best. The log-log link is $g(\mu_i) = log(-log(1-\mu_i))$. Cribari-Neto and Zeileis (2010)

Next, we predict on the testing data with the log-log link model with four PCs trained on the training data. The training MSE and Pseudo-R^2 are 0.0107 and 0.301, respectively, and 0.0117 and 0.362 for the testing data. We notice that the Pseudo-R^2 is higher on the testing set than on the training set. This suggests that the model is generalizing to new data well and is likely not overfitting. The training and testing MSE are similar, corroborating this; hence, the predictive power of this beta model seems reasonable.

For inference, we refit the model with the log-log link to the full dataset, save for five outlier points removed from the full dataset. The loadings for the PCs are given in Table 1. The corresponding coefficients in the fitted beta regression are given in Table 2.

Crime Type	PC1	PC2	PC3	PC4
Burglary	-0.420	0.189	0.160	-0.433
Larceny	-0.398	0.256	0.202	-0.506
Vehicle Theft	-0.413	0	0.265	0.476
Murder	-0.267	-0.627	-0.647	-0.243
Rape	-0.258	0.650	-0.630	0.319
Robbery	-0.414	-0.263	0.217	0.412
Agg. Assault	-0.431	0	0	0

Table 1: Principal component loadings for the principal components used in the beta regression.

Variable	Estimate	Std. Error	z-value	p-value
Intercept	-0.465	0.0170	-27.374	≈ 0
PC1	0.0183	0.0061	3.020	0.0025
PC2	0.0247	0.0125	1.971	0.0487
PC3	-0.0248	0.0147	-1.692	0.0906
PC4	-0.0719	0.0162	-4.431	9.39×10^{-6}
Inv. Admission Rate	0.0819	0.0068	12.088	≈ 0

Table 2: Beta regression coefficients for the final fitted model.

Notice that the coefficients for the first and fourth PCs are exceedingly significant in the full model, while the second and third PCs' coefficients are only marginally significant to non-significant.

For the diagnostic plots of this model: the linear predictors versus residuals (Figure 7), standardized residuals plot (Figure 8), and predicted values versus true values plots (Figure 9) look reasonable. Though, we believe the model is having difficulty predicting higher matriculation rates, as seen in Figure 9. Also, we notice a slight hyperbola-like shape on the generalized leverage versus predicted values plot (Figure 10), which suggests that the beta model may not be adequately predicting exceptionally high or low matriculation rates. This behavior leads us to believe that performing quantile regression may be useful for understanding the relationship between crime and matriculation rate in colleges with very high or low matriculation rates.

4.2 Adding Spatial Dependence

4.2.1 Model Motivation

Given the results of the previous section, the beta regression appears appropriate to model this problem, specifically with a log-log link function. To determine whether we need to account for spatial autocorrelation, the residuals from the beta regression fit earlier are used along with location data to create a variogram.

As seen in Figure 11, the increasing values up to a distance of 20 indicate the presence of spatial autocorrelation. Thus, fitting the model with a spatial autocorrelation is a reasonable strategy to model the crime's principal components and the reciprocal of admission rates against the matriculation rate at different principal components. We proceed to fit the model using the spmodel package (Dumelle et al., 2023) in R.

4.2.2 Spatial Model Results

Using the training data, 7 principal component regressions are fit, and training RMSE is calculated and plotted in Figure 12; we noticed a stark decline and little improvement after adding a 3rd principal component. Thus, that is the choice in the number of principal components. Fitting the model on the testing data yields a RMSE of 0.526, close to the training RMSE of 0.4699; validating the generalizability of the model, but interestingly larger than the RMSE of the original beta regression.

After re-fitting the model on the full data set and verifying the usability of the model through residual plots that show no pattern (Figure 13), we arrive that the first principal component and admission rates are the only significant non-intercept predictors, as seen in Table 3.

Table 3: Coefficients and their significance for the spatial beta regression

Component	Estimate	P-Value
Intercept	-1.49	$< 2 \cdot 10^{-16}$
PC1	0.0507	$6.29 \cdot 10^{-6}$
PC2	0.0093	0.686
PC3	-0.0165	0.537
Inv. Admission Rate	0.1093	$< 2 \cdot 10^{-16}$

Then, calculating AIC, we get 783.2 for the spatial model and 826.55 for the nonspatial model, providing evidence that a model accounting for spatial autocorrelation is better than one without. However, based on Figure 14, we see that after a distance of 20, the covariance of matriculation rates between schools does not exist (which is consistent with the fitted variogram). Thus, the use of spatial autocorrelation is validated in this model.

4.3 Quantile Investigation

After removing Larceny and fitting quantile regression using $\tau = .05, .25, .50, .75$ and $.95$, we noticed the differences between model slopes as seen in Table 7. Many of the quantile models turned out to be statistically different. Every quantile model is at least marginally ($\alpha = .1$) different from the base median quantile case at $\tau = .5$. We notice that we have insufficient evidence to believe slopes are significantly different between $\tau = .05$ and $\tau = .25$, as well as $\tau = .75$ and $\tau = .95$. Even without sufficient evidence to believe that some of the groups are different, we kept all of the initial groups in as metrics of comparison. Based on this, we fit the lasso and predicted on the testing set to determine what conclusions we can make about the corresponding fits.

Table 4: Loss Statistics for Quantile Regression

τ	Underpredictions	Overpredictions	% of Data Below	Total Pinball Loss
0.05	6	154	3.75%	1.331954
0.25	26	134	16.25%	4.55752
0.50	70	90	43.75%	6.757217
0.75	112	48	70%	6.07395
0.95	154	6	96.25%	2.409536

Table 4 contains metrics for the model, most notably the real quantile of the data that the predictions are capturing, and the total pinball loss. Since the loss is penalized, it should be viewed similarly to a penalized sum of absolute errors. We notice the prevalence of the right-skew in the quantiles, but the pinball loss is relatively small while accounting for the penalty. This implies that quantiles are relatively

predictable within the scope of our dataset, and especially that it is possible to predict the top and bottom 5% of the data relatively accurately. The loss function depicted in Figure 15 displays the nature of the prediction and corresponding loss per quantile. Our final coefficients per quantile are given in Table 5.

Table 5: Final Coefficients For Lasso Quantile Regression

	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
Intercept	1.79e-01	2.66e-01	3.86e-01	5.36e-01	7.83e-01
Murder	0	7.68e-05	0	5.10e-04	0
Rape	0	0	4.71e-05	0	0
Robbery	-1.19e-04	-1.51e-04	0	-1.57e-04	0
Aggravated Assault	0	3.29e-06	0	0	-5.97e-05
Property Crime	0	0	0	1.60e-05	0
Burglary	0	0	0	-2.94e-05	0
Admission Rate	-7.25e-02	-1.17e-01	-2.10e-01	-2.98e-01	-2.92e-01
Motor Vehicle Theft	-5.66e-06	-6.79e-06	-3.70e-05	-8.61e-05	-8.75e-05

5 Conclusion

From the results of our beta regression model, we first see the loadings for each type of crime in the first principal component are all negative, while the model's PC1 coefficient is positive and significant. This suggests that as overall crime rates decrease, we expect the matriculation rate of a given college to increase. The second principal component has a significant positive coefficient but has negative loadings for murder and robbery. Thus, we expect these violent crimes to have significantly lower expected matriculation rates than the first principal component originally suggested. The third principal component has only a marginally significant negative coefficient, so given the positive loadings on property crimes and robbery, we weakly expect these crimes will further lower matriculation rates. Finally, given the largely significant negative coefficient of the fourth principal component and the positive loadings of rape, robbery, and vehicle theft, we expect that increases in these crimes will further lower mean matriculation rates. We also see that, as expected, the inverse admission rate is strongly positively related to the matriculation rate, reinforcing the work of Schmitz (1993). While true for the original beta regression model, the results drastically change when accounting for spatial covariance. Instead of 3 principal components being significant when explaining matriculation rate, only the first principal component is significant, with a now stronger positive magnitude. This implies that overall crime is negatively correlated with matriculation rate. Moreover, with a stronger AIC than the model without spatial autocorrelation, we are left with the conclusion that there exists a trend between an overall decrease in crime rates and an increase in matriculation rates.

Quantile regression yielded us marginally significant results. With the justification step, we found extremely significant evidence that slopes differ between model quantiles. At different quantiles, we noticed that different crime rate variables were useful in predicting matriculation rate. Through every quantile, Admission Rate and Motor Vehicle Theft were important predictors, leading us to believe that we should do future investigation with these predictors. In the majority of cases, crime rate variables were negative, and if not, then they were close to zero (likely due to relationships with other covariates). A concatenated model using different coefficients at different slopes appears moderately effective. In discussing the quantile error, the model needs more significant predictor variables to capture the missing variability. However, due to the penalty term, the model does relatively well at the lower and upper echelons.

To synthesize, we notice a small negative correlation between students between crime rates and matriculation rates. Due to the number of confounders, it is hard to extract inferences or demonstrate any level

of "proof" that students are or are not using crime rates when deciding on a school. However, because our model includes only one confounder (admission rate), it provides the best chance for the crime rate to explain the matriculation rate. As a result of the spatial model and the quantile regression, we notice that there is some significance in predicting the matriculation rate with the crime rate, but the overall R^2 is quite low, demonstrating that crime rates may not adequately correlate with the matriculation rate. Consequently, we believe more work should be done to better inform students that crime rates around the city the school is located can drastically affect the student's education.

5.1 Limitations

As mentioned in the data section, there are a lot of confounders when trying to analyze this problem. Students have a plethora of varying criteria when choosing a school, not all of which can be captured in a data analysis. Without more dense data on these confounders, inferring whether students are using crime rates when choosing a college is considerably difficult. Regarding the modelling procedure, the data is not terribly dense with respect to midwestern schools, further throwing the model's reliability into question. Moreover, the residual plots are hard to analyze in beta regressions, so verifying the validity of the beta regression is hard (Ferrari and Cribari-Neto, 2004). Quantile regression proved to be useful in the context of our data (especially in dodging a few assumptions required for standard linear models). However, to get a more accurate read on patterns in the data, we most likely need a much bigger sample size (preferably above 1500, triple our current data size). As per Koenker and Hallock (2001), we note that in some cases data interpolation may be limited due to sample size, especially with a numerous amount of covariates. In our train and test splits, we still had enough colleges in each quantile for the Central Limit Theorem to hold, but we believe with more data points we would have stronger conclusions and better estimates.

5.2 Future Work

Given the spatial autocorrelation, further modelling can be done to enhance the regression performance. More specifically, spatial Gaussian processes are excellent (Gelfand and Schliep, 2016) non-linear models. A non-linear model might be able to separate out across different quantiles if a piecewise function is generated per quantile. Beyond different models that could be used, running a survey on whether students strongly factor in crime data would help make inferences on the strength of importance students place on the crime rates of their universities' home cities.

References

Burdick-Will, J., Ludwig, J., Raudenbush, S., Sampson, R., Sanbonmatsu, L., and Sharkey, P. (2011). *Converging evidence for neighborhood effects on children's test scores: An experimental, quasi-experimental, and observational comparison*, pages 255–276. Russell Sage Foundation.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in r. *Journal of Statistical Software*, 34(2):1–24.

Dumelle, M., Higham, M., and Ver Hoef, J. M. (2023). spmodel: Spatial statistical modeling and prediction in r. *PLOS ONE*, 18(3):1–32.

Ezarik, M. (2022). Students approach admissions strategically and practically. *Inside Higher Ed*.

FBI (2021). Bureau of investigation crime data explorer, crime rates by city known to law in 2021-2021.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Gelfand, A. E. and Schliep, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, 18:86–104.

IPEDS (2021). Integrated postsecondary education data system dataset on united states colleges.

Koekner, R. e. a. (2023). *Quantile Regression*. R package version 5.97.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.

Lacoe, J. (2020). Too scared to learn? the academic consequences of feeling unsafe in the classroom. *Urban Education*, 55(10):1385–1418.

Legewie, J. and Fagan, J. (2019). Aggressive policing and the educational performance of minority youth. *American Sociological Review*, 84(2):220–247.

Schmitz, C. C. (1993). Assessing the validity of higher education indicators. *The Journal of Higher Education*, 64(5):503–521.

Sharkey, P. (2010). The acute effect of local homicides on children's cognitive performance. *Proceedings of the National Academy of Sciences of the United States of America*, 107(26):11733–11738.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225.

Waldmann, E. (2018). Quantile regression: A short story on how and why. *Statistical Modelling*, 18:203–218.

6 Supplementals

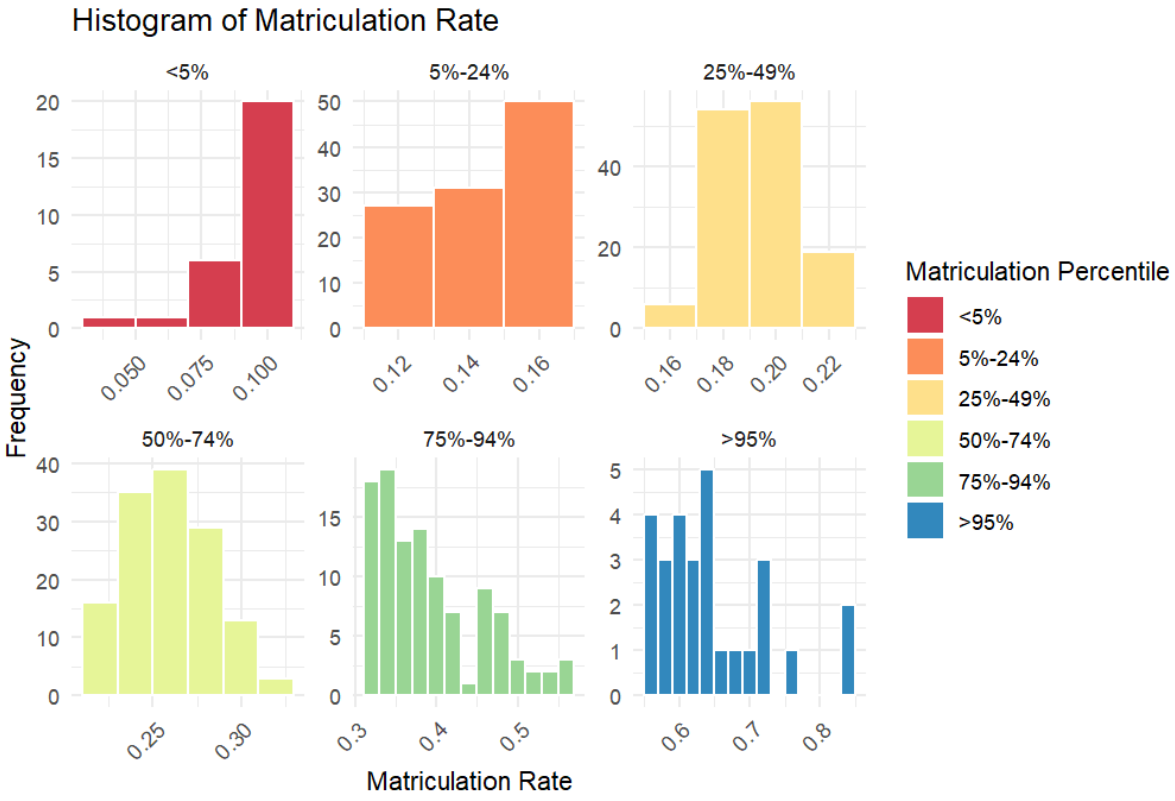

Figure 5: Distribution of Response Variable by Quantile

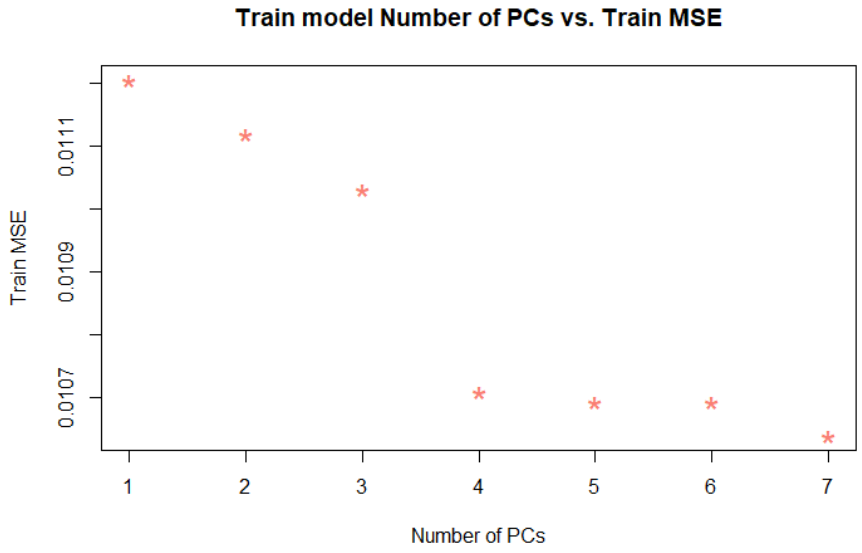

Figure 6: RMSE of the beta regression at a varying number of principals components used in the fit of the model

Link	Pseudo R^2
logit	0.252
probit	0.261
clog-log	0.223
Cauchy	0.179
log-log	0.301

Table 6: Comparing the training Pseudo-R^2 of the beta regression for different choices of link functions. The log-log link has the best RMSE and, consequently, is used in the beta regression modelling.

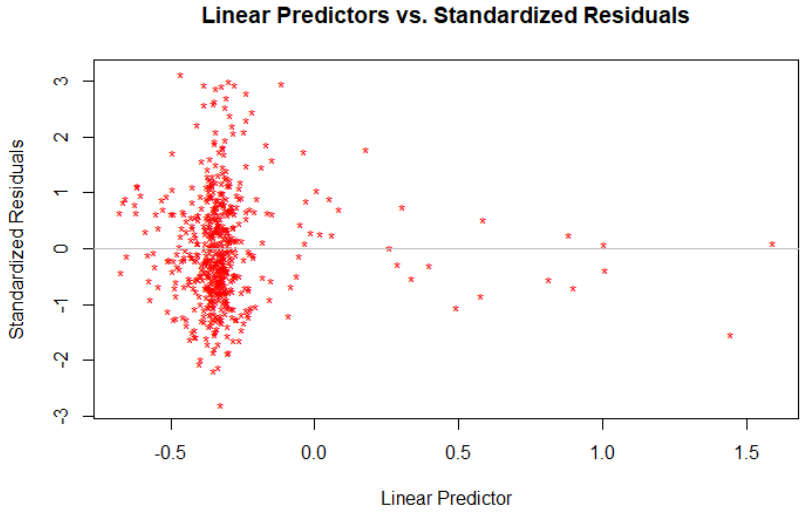

Figure 7: Model diagnostics for the beta regression. Linear predictors versus the residuals.

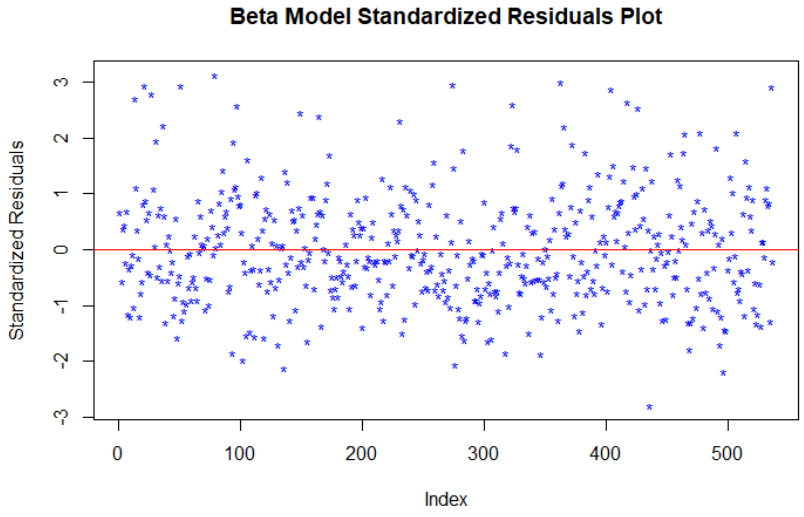

Figure 8: Standardized residual plot for the beta regression. A lack of pattern suggests a good fit.

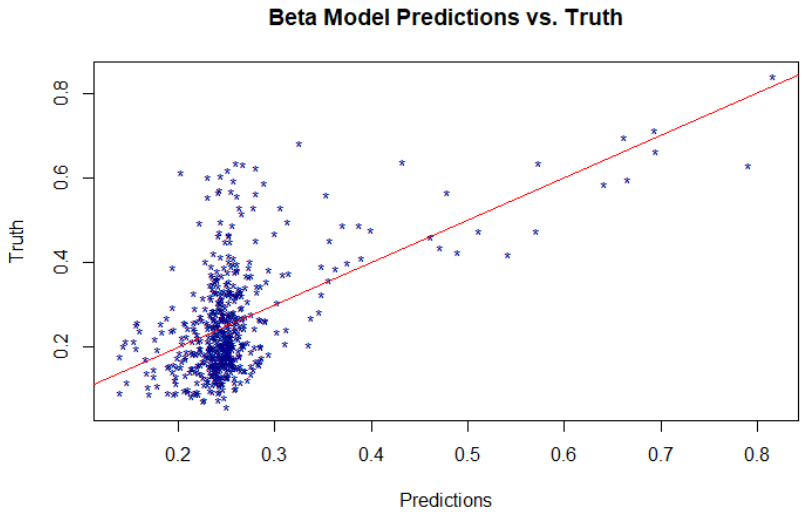

Figure 9: Model diagnostics for the beta regression; the predicted versus the true matriculation rate.

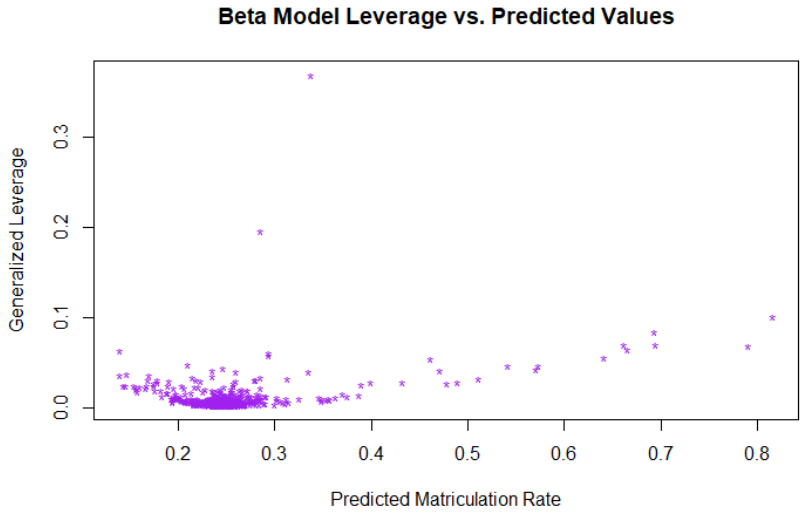

Figure 10: Model diagnostics for the beta regression. This leverage plot with a slight pattern suggests that the beta model might not be adequate at the extremes (very high/low matriculation rates).

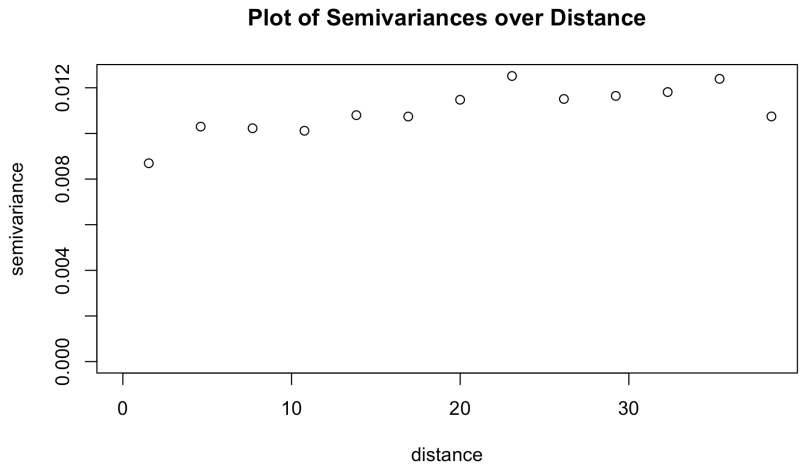

Figure 11: Variogram based on residuals from the original beta regression. The increasing trend indicates spatial autocorrelation.

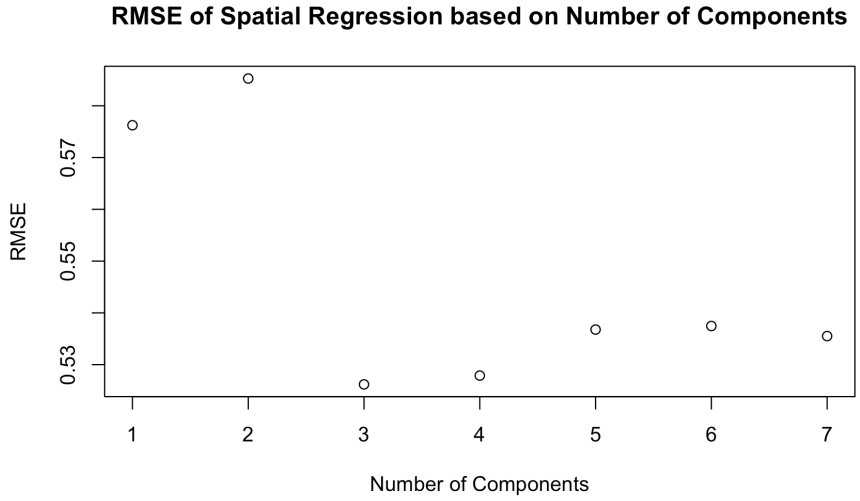

Figure 12: RMSE of the spatial regression fit with exponential decay of spatial covariance for each number of components used in the modelling.

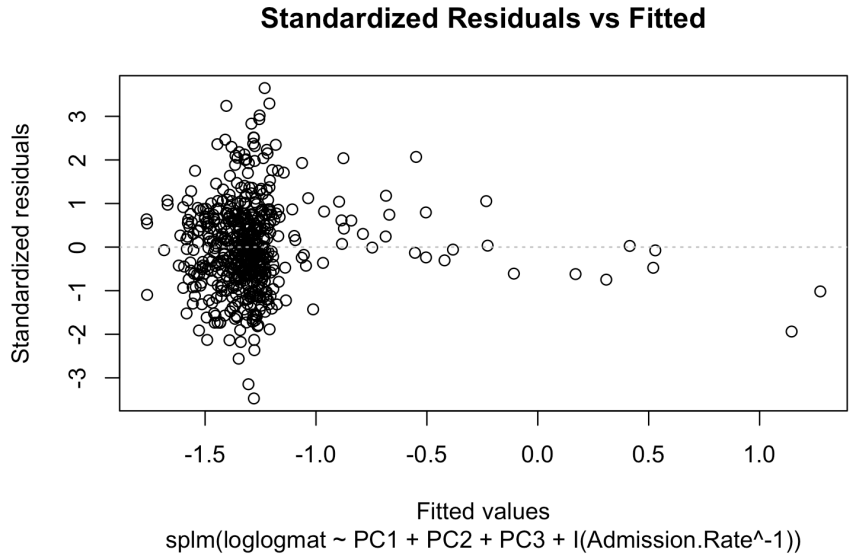

Figure 13: Residuals versus fitted values for the beta regression accounting for spatial autocorrelation. The lack of pattern suggests the model satisfies all necessary assumptions.

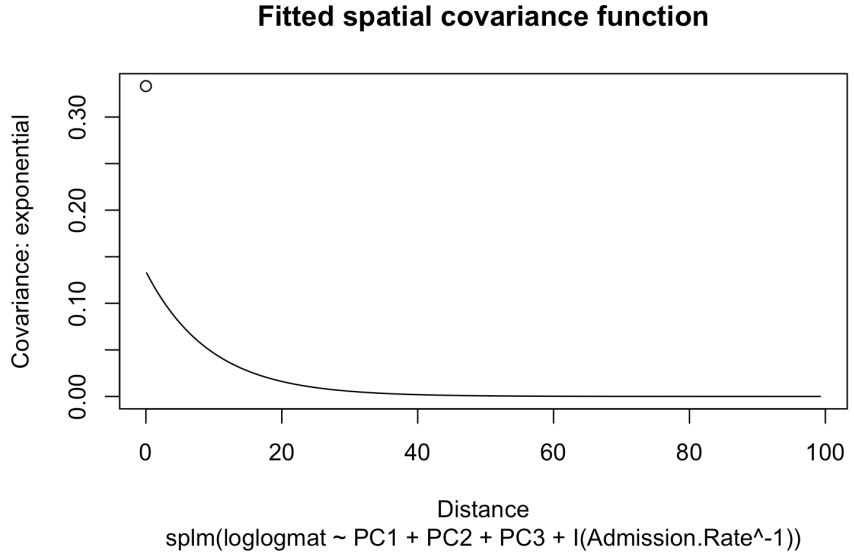

Figure 14: Fitted spatial covariance over distance. The taper towards 0 around 20 miles suggests that there is no correlation between matriculation rates of schools over 20 miles away from each other.

Table 7: ANOVA results for the quantile regression

τ	P-Value	Significance Code
$0.05, 0.25$	$0.221\,398$	
$0.05, 0.5$	3.2×10^{-5}	***
$0.05, 0.75$	2×10^{-6}	***
$0.05, 0.95$	$0.002\,914$	**
$0.25, 0.5$	$0.000\,178$	***
$0.25, 0.75$	8.2×10^{-5}	***
$0.25, 0.95$	$0.018\,465$	*
$0.5, 0.75$	$0.070\,384$	.
$0.5, 0.95$	$0.020\,653$	*
$0.75, 0.95$	$0.247\,265$	
$0.05, 0.25, 0.5$	0	***
$0.05, 0.25, 0.75$	8×10^{-6}	***
$0.05, 0.25, 0.95$	$0.005\,566$	**
$0.05, 0.5, 0.75$	4×10^{-6}	***
$0.05, 0.5, 0.95$	8×10^{-6}	***
$0.05, 0.75, 0.95$	2×10^{-6}	***
$0.25, 0.5, 0.75$	2×10^{-6}	***
$0.25, 0.5, 0.95$	9×10^{-5}	***
$0.25, 0.75, 0.95$	7.7×10^{-5}	***
$0.5, 0.75, 0.95$	$0.057\,269$	.
$0.05, 0.25, 0.5, 0.75$	0	***
$0.05, 0.25, 0.5, 0.95$	0	***
$0.05, 0.25, 0.75, 0.95$	5×10^{-6}	***
$0.05, 0.5, 0.75, 0.95$	5×10^{-6}	***
$0.25, 0.5, 0.75, 0.95$	6×10^{-6}	***
$0.05, 0.25, 0.5, 0.75, 0.95$	0	***

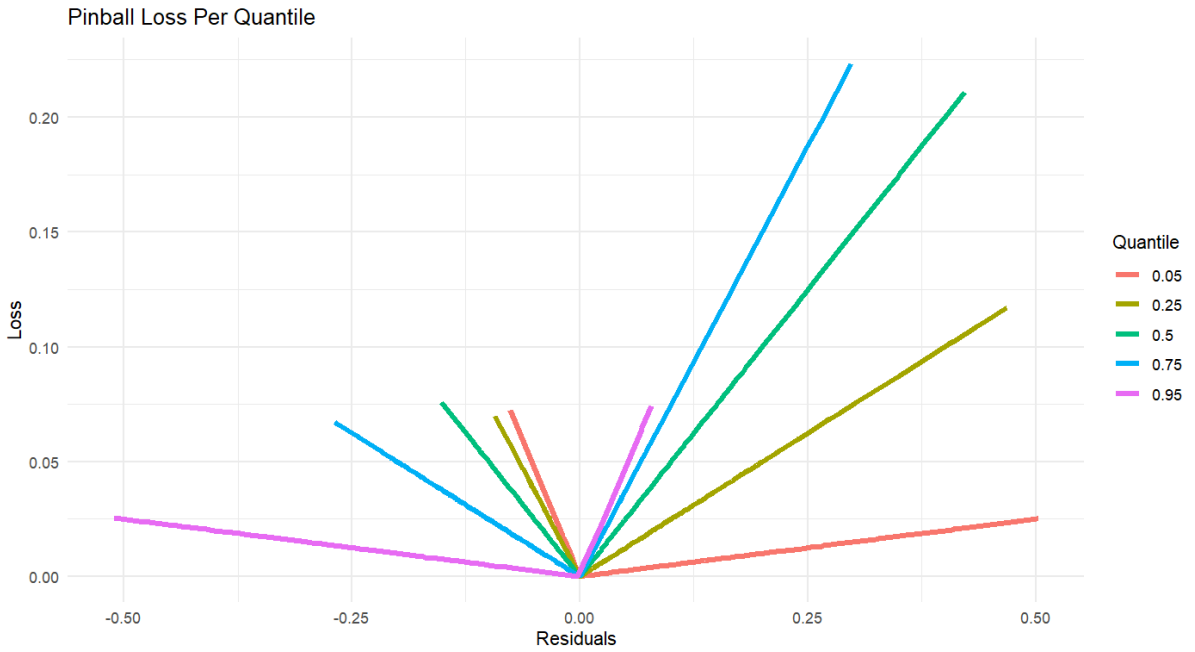

Figure 15: Loss Function at 5 Different τ Values