Important Predictors for U.S. Rent Prices

Daniela Shuman, Janet Li, Felicia Roman, Sam Wright

I. OVERVIEW & MOTIVATION

Rent prices have dramatically increased in cities across the globe in the past 2 years[1]. In the US particularly, this has put a burden on many American households. Furthermore, different cities can have large differences in baseline rent prices, as well as in the variations of baseline rents. This raises the question, what are some of the most influential factors that affect rent prices, across all cities in the U.S.?

II. HYPOTHESES

High Paying Jobs and Rent Price

1. *Hypothesis 1: We expect the number of high paying jobs to have a positive association with rent price*
   If there are more higher paying jobs in a region, we would expect individuals on average would be able to pay more for rent. The market then would account for this by raising rent prices. Further, more high paying jobs would suggest better amenities in the region, as in more expensive grocery stores or opera houses, which also would be associated with higher rent prices. It is possible that individuals who fill the higher paying jobs in the region wouldn't live in the area, and thus, the prices would be lower. However, we expect that companies would pay their employees more if the rent price is higher because cost of living is higher in the census tract. This would reflect in a positive association between rent price and high paying jobs.

2. *Hypothesis 2: We would expect the metropolitan region to have no effect on the high paying jobs association with rent price.*
   We don't expect there to be a difference in the rent price association with high paying jobs across regions. This phenomenon, we expect, is universal and thus, the association shouldn't change even if metropolitan regions change.

Mean Commute Time and Rent Price

1. *Hypothesis 1: We expect there is a negative association with mean commute time and rent price.*
   Generally, being closer to a job is preferable, so there would be higher demand for places to live with lower commute times, and thus higher rent price. Livability indices affect

[1] WSJ - Rent Prices Across the US

rent price in a region, and the closer a work location is to a home location, the higher the livability index. Individuals that are closer to their work location would likely live in more densely populated neighborhoods, which means higher demand and potentially lower supply of housing. There are a few different scenarios that could affect this association. If living in the center of the city is not desirable due to other factors, then the mean commute time would increase, but rent price would be lower. Additionally, it could be the case that a city is a suburb of a different, larger metropolitan area where there are many jobs. In this case, the rent would be lower here, but the mean commute time would be higher.

2. *Hypothesis 2: We expect that the metropolitan region should not affect whether or not there is a negative relationship between commute time and rent price.*
   There isn't an apparent reason as to why there would be a differential association between commute time and rent price depending on metropolitan area. However, we'd like to confirm this as there is a possibility that based on metropolitan area-specific organizational features that this effect could be different from city to city.

Race and Rent Price

1. *Hypothesis 1: We expect that neighborhoods with higher white population composition would have higher rent prices.*
   Due to historical segregation in the US, we would expect that rent prices are higher in white communities than in minority communities. Due to the historical wealth gap between races, we would expect that whiter communities have the capacity to pay more for rent than minority communities, raising rent prices.
2. *Hypothesis 2: We expect that the metropolitan region would affect the association of race with rent price.*
   Cities with more historical rates of segregation could have a stronger association of race with rent price than newer cities, or cities with less segregation. For example, we would expect to see Boston would have a stronger association between rent and racial composition because of a history of segregation. A city like Seattle, for example, with a shorter history of segregation would likely have a lesser association of rent and racial composition. Additionally, because much of the racial diversity in Seattle is directly from immigration, the relationship between race and rent price is a pretty different underlying relationship.
   Additionally, we would expect cities with higher population density to have

Foreign Share and Rent Price

1. *Hypothesis 1: We expect that neighborhoods with higher foreign population composition would have lower rent prices.*
   We would expect individuals born outside the US to pay less for rent than individuals born within the US. An individual born inside the US would likely have a better ability at

wealth building than an individual without roots in the country. Communities that have larger foreign populations, thus, would probably have lower rent prices.

2. *Hypothesis 2: We expect that the metropolitan region would affect the association of foreign share and rent price.*
   Cities with higher rates of educated foreign populations would likely have higher rent prices than cities with more working class foreign populations. Take Seattle, for example, a city with a very large educated foreign population, but a lower uneducated foreign population. Rent prices in Seattle are very high in comparison to other cities in Washington state with similar foreign population shares, but less educated foreign populations, and dramatically lower rent prices.

III. DATA DESCRIPTION & CLEANING

We have 56607 observations, with 22 predictors and 1 response variable. Our data is from the Opportunity Atlas from the Opportunity Insights Lab.
Our response is the average rent for a two-bedroom apartment in 2015 (*rent_twobed2015*).
The following are predictors in the data set:

- *czname:* Name of the commuting zone identifier
- *med_hhinc2016:* Median commute time of working adults in 2016
- *frac_coll_plus2010:* Fraction of Residents with a college degree or more in 2010
- *foreign_share2010*: Share of population Born Outside the US in 2010
- *popdensity2010*: population density in 2010
- *poor_share2010*: Share of poor individuals in 2010
- *share_*race*2010*: Share of individuals of each race (white, black, hispanic, and asian) in 2010
- *gsmn_math_g3_2013*: Average School District Level Standardized Test Scores in 3rd Grade in 2013
- *singleparent_share2010*: Share of Single-Headed Households with Children 2010
- *traveltime15_2010*: Share of Working Adults w/ Commute Time of 15 Minutes Or Less in 2010
- *kfr_pooled_pooled_p25*: Upward Mobility
  - The upward mobility metric is the mean percentile income rank in the national distribution at age 31-37 for children with parents at the 25th percentile of the national income distribution.
- *jail_pooled_pooled_p25*: Upward Mobility
  - The upward mobility metric is the mean percentile income rank in the national distribution at age 31-37 for children with parents at the 25th percentile of the national income distribution and who have been incarcerated.
- *emp2000*: Employment rates in 2000
- *ln_wage_growth_hs_grad*: Log wage growth for HS Grad., 2005-2014

- *ann_avg_job_growth_2004_2013*: Average Annual Job Growth Rate 2004-2013
- *mail_return_rate2010*: mail return rate in 2010
- *jobs_total_5mi_2015*: number of primary jobs within 5 miles in 2015
- *jobs_highpay_5mi_2015*: number of high paying jobs in 2015

Only *czname*, the commuting zone name, is a categorical variable. The distribution of the number of observations for each commuting zone is heavily right-skewed (see Figure 1 below), which makes sense because we expect to have more data for highly-populated commuting zones. For example, we have 3065 observations for Los Angeles, 2197 for New York, and 1545 for Chicago. Approximately 5% of the commuting zones (40 out of all 683) account for almost 50% of the observations. Furthermore, from Figure 2, it appears that these highly-populated commuting zones tend to have higher rent prices, and thus have the potential to heavily influence our results. We will need to account for this severe imbalance between the groups in later analyses.

Excluding the top 10 commuting zones in the number of census tracts, below is a distribution of the number of census tracts per commuting zone. As you can see from this right skewed distribution, the vast majority of cities have a small number of census tracts.

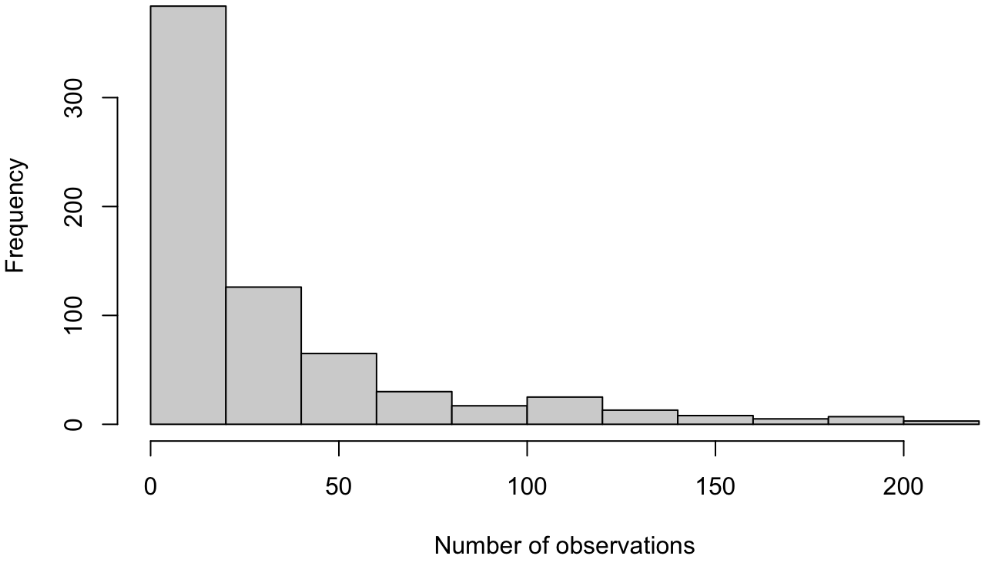


Figure 1. Distribution of the number of observations in each commuting zone, excluding the 68 largest commuting zones (~10%) to improve readability.

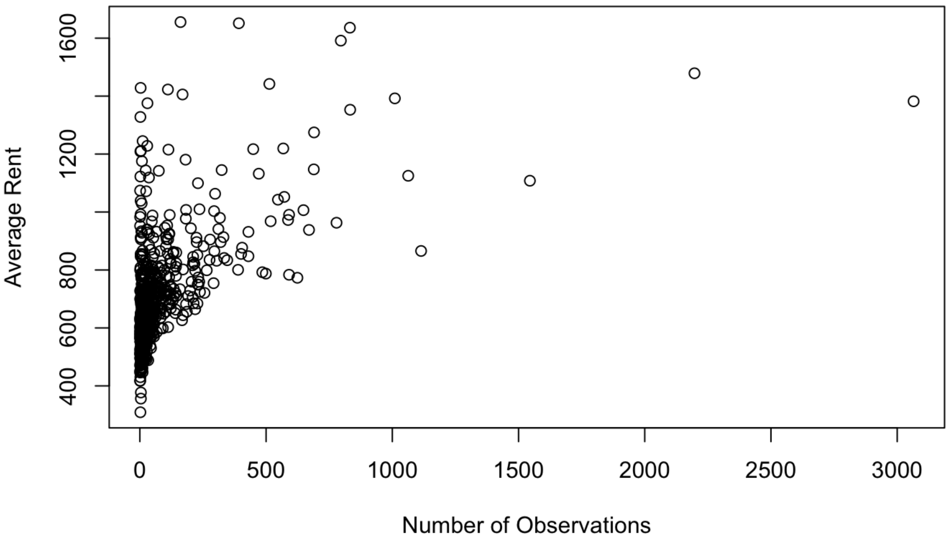


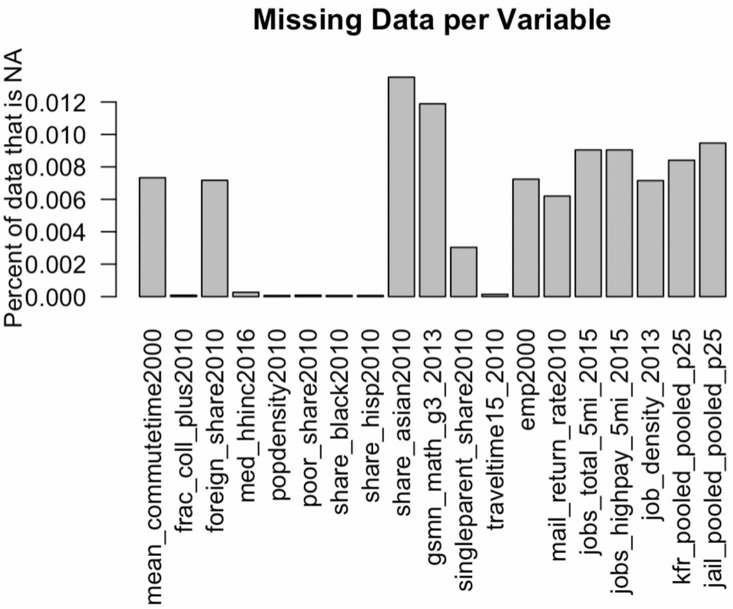
Figure 2. Average rent (response) vs. number of observations for each commuting zone.

The remaining 22 variables (including the response) are quantitative. From histograms of each quantitative variable (Figure 1 in the Appendix), we can see that many variables are heavily right-skewed, including rent. Even though our sample size is large enough that the Central Limit Theorem applies, it may still be helpful to log-transform the right-skewed predictors, to better satisfy the Normality assumptions of future models (e.g., baseline linear regression).

Handling Missing Data

We removed the few observations that had missing values for commuting zone name or average rent (the response variable). The reason we did this was to avoid making decisions off of an imputed response variable. Of the remaining rows, after removing the rows with rent price existing, only 0.7% had a missing community zone name, so we don’t expect removing these rows to significantly affect the results.

There are some predictors with more missing data than other predictors, but all the predictors are missing at most 0.012% of the data. This is a small amount, such that the imputation should not dramatically affect the results.

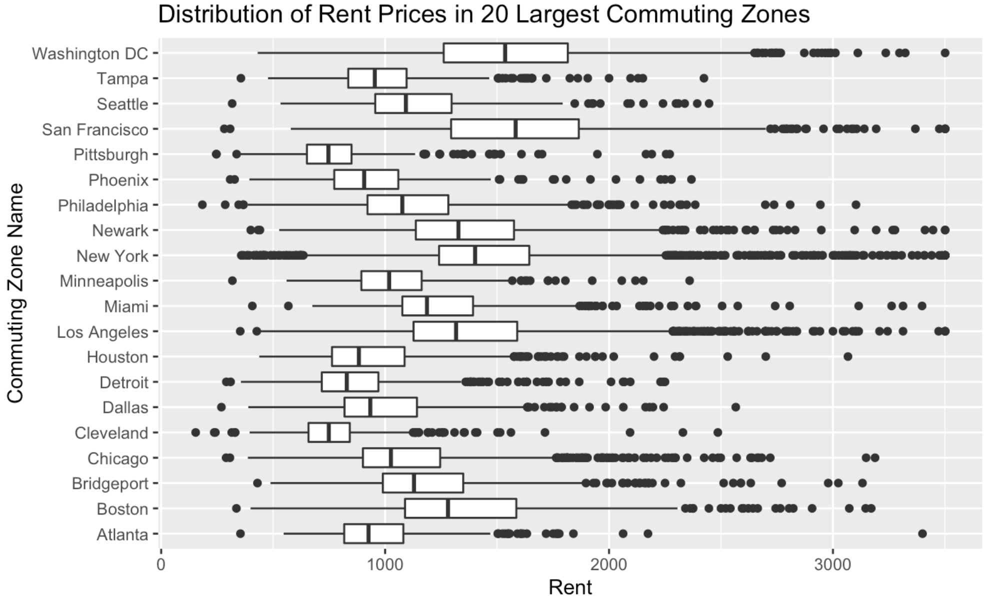


For the remaining 21 quantitative predictors, we imputed the missing data with the mean or median for that commuting zone. If the predictor was heavily skewed, we used the median; otherwise, we used the mean. If all values for that commuting zone were missing, we imputed the missing data with the mean or median across all observations in the dataset (the It is possible that our predictors vary per commuting zone. For example, the mean number of high paying jobs would likely be different in certain cities. Thus, it doesn't make sense to include an aggregator across all census tracts, but rather by bins. We decided to take the means of each of the following predictors per commuting zone because their distributions were not as strongly skewed by outliers and it is more interpretable: mean_commutetime2000, frac_coll_plus2010, gsmn_math_g3_2013, singleparent_share2010, traveltime15_2010, emp2000, ln_wage_growth_hs_grad, and ann_avg_job_growth_2004_2013. For the variables with more skewed distributions due to outliers, we took the median of the data per census tract to impute. This way, the aggregator is not as sensitive to the outliers.

Understanding the Response Variable

In order to build models to understand the influential factors in the rent price variable, we need to understand the distribution of our response. As expected, rent price is very right skewed, similar to other income related variables (as demonstrated by our plot in Figure 1 of the appendix). This means that there are a few very high rent census tracts but many lower rent census tracts, with a lower bound at 0, which is making the distribution right skewed. In all of our models, we considered a log transformation on the rent variable to account for this right skew.

Additionally, per commuting zone, the rent price varies wildly in its median and in its variance. This may be because some commuting zones have different numbers of observations, leading to different variances. In order to see what are the influencing factors of rent, we need to make sure we remove any baseline rent price from what we are predicting. Perhaps, just because the city is Detroit (there may be something special about Detroit), the median rent price is lower, than say San Jose (people generally like being near the beach) so the median rent price may be higher. We want to remove these effects so we can isolate the association between the rent price and our above predictors. This is why we choose to use fixed effects across commuting zones, which we will demonstrate below. Additionally, there is differing variance across commuting zones. Some commuting zones have a greater spread, say Washington DC, for example, in comparison to Pittsburgh. We need to account for this as well when we do fixed effects modelling, which is why we chose to also run the Random Intercept model with the fixed effects (explained below).



IV. MODELING APPROACH

In order to best understand what is the association between rent price and the above predictors, we employed several different methods, each addressing specific assumptions. We started with a baseline OLS regression model without fixed effects to observe how the predictors are associated with rent price. We then ran LASSO and a Random Forest to validate the variable importances to check if indeed our four hypothesis variables would be selected in a more generalizable setting. We chose to add a regularization term to our baseline model for variable selection, to remove the unimportant variables.

Baseline Model

Our base model is an OLS regression with rent price as the response and all the other variables as predictors. Almost all of the predictors were significantly associated with rent, which makes sense because our sample size is very large. The R^2 of the baseline model was 0.7589. The full linear regression output is in Figure 6 of the Appendix.

The linearity assumption seems satisfied in the baseline model. From the Residuals vs. Fitted plot you can see that there is almost no curvature on the plot and that for the most part the residuals are distributed on the residuals = 0 line.

The normality assumption is not satisfied in the baseline model. From the Normal QQ plot it is apparent that the standardized residuals do not fall on the standardized residuals = theoretical quantile, meaning that the residuals are not Normally distributed around the fitted line. This shouldn't be a huge problem since non-normally distributed residuals are typically not a concern for linear models unless we have high leverage points. There are a few leverage points as can be seen on the Residuals vs. Leverage plot so it's possible that we will need to remove these values to eliminate this concern.

The constant variance assumption is not satisfied well by the baseline model. The Scale-Location Plot shows that the square root of the standardized residuals increases as fitted values increase, so this points to variance of the observations around the fitted line is not constant. This is somewhat of a concern for a linear model (not as much as not satisfying the linearity assumption) because it doesn't bias the predictors but it does effect standard errors which does affect significance calculations, so it may be that we will need to consider other models for this data in the next stage of the project.

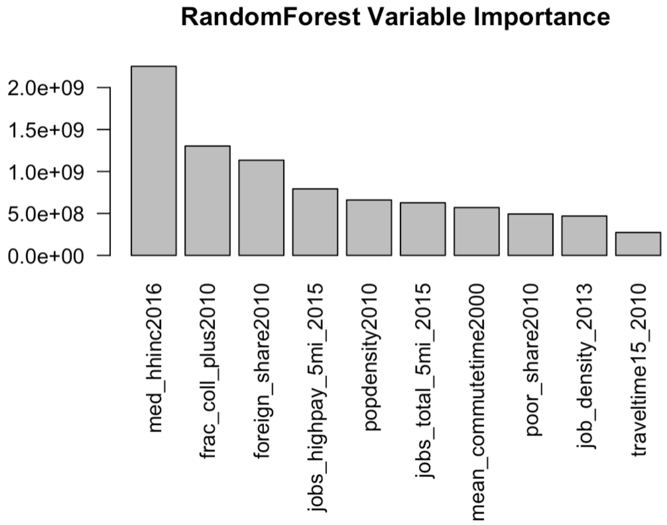
The independence assumption cannot be verified directly here. It is possible that this assumption can be violated because many of the predictors we used were from different points in time which may be correlated. Additionally, since the observations are from different geographical locations it's possible that some of them are correlated based on proximity to one another. Since this data type and these predictors are inherent to our project we may have to accept the possibility that this assumption may not be satisfied. Nevertheless, we may be able to minimize the disadvantages that come with this possibility by emphasizing model choice in the final phase of the project.

There is some amount of multicollinearity between a few of the predictors. Particularly, we see that the share of hispanic individuals is very correlated with the share of asian individuals in a census tract. We may consider removing a few of these variables or running regularization to deal with the multicollinearity. Employment is relatively correlated with the share of poor individuals, as expected. Check for the full variance covariance matrix in the appendix.

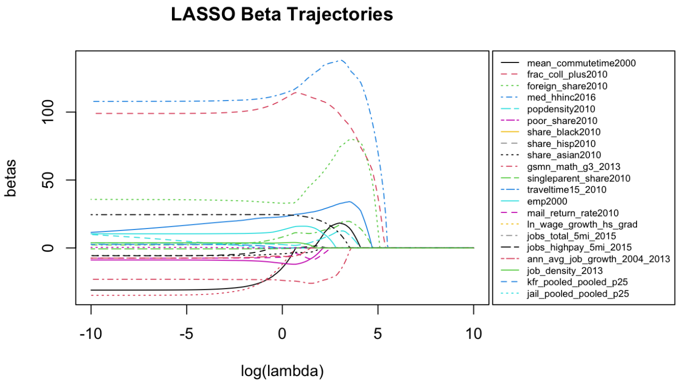
Variable Importance

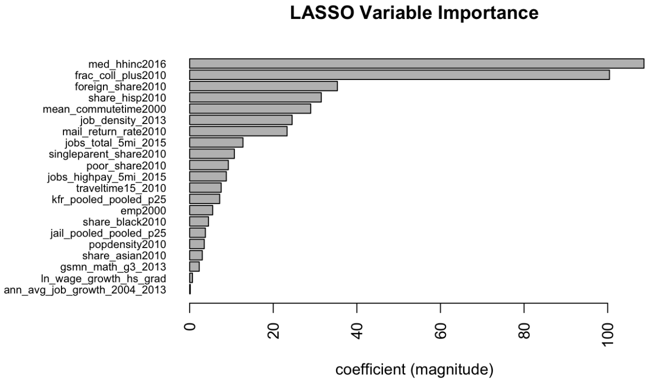
To determine variable importance, we ran two models: a random forest model and a LASSO model.

We chose to run a random forest in order to observe whether a non-parametric model would determine our four variables as important to determining rent price. Since we are not building a prediction model, we decided it would be useful to only run the random forest to determine variable importances. From our random forest model, it seems that the median household income, the fraction of individuals that have college degrees, and the foreign share of individuals per census tract are strong predictors of rent price. All three of our four hypothesis variables were included as important variables for the random forest. The reason the share of nonminority was not listed as an important predictor variable may be because the predictor is an aggregate variable of 3 different race variables, each currently listed as insignificant. We explore whether race is truly insignificant below.



We also fit a well-tuned LASSO regression model, based on the predictors used in the baseline linear regression model in order to observe which variables were chosen in variable selection. We chose a LASSO model over a Ridge regularized model in order to specifically run variable selection. From the plot of the hat(β) trajectory plots of the main effects, we can see that the hat(β)'s that shrink to 0 the latest correspond to the same 3 predictors identified in the random forest model: median household income, the fraction of individuals that have college degrees, and the foreign share of individuals per census tract. These 3 predictors also had the largest absolute magnitudes in our well-tuned LASSO regression model. Foreign share is clearly demonstrating some important association with rent price, but we need to observe this further below. The number of high paying jobs and mean commute time are also listed as somewhat important variables and are not dropped as frequently as other variables with LASSO regularization, however, we need to do further work in determining whether there is an association between these predictors and rent price. As hypothesized above, share of each racial variable is a less important predictor. This may be the case because the share of each individual minority doesn't seem to be as strongly associated with rent, but observe the aggregate, that the share of minority communities in general, we may observe a stronger association with rent price.





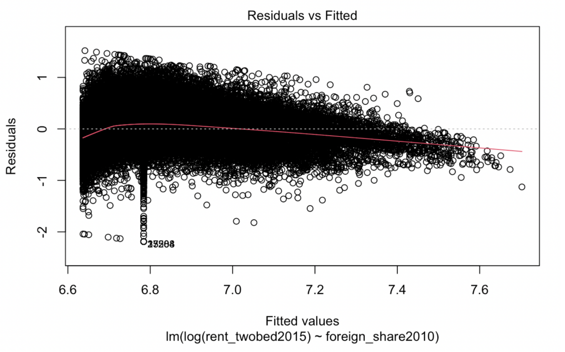
V. RESULTS & DISCUSSIONS

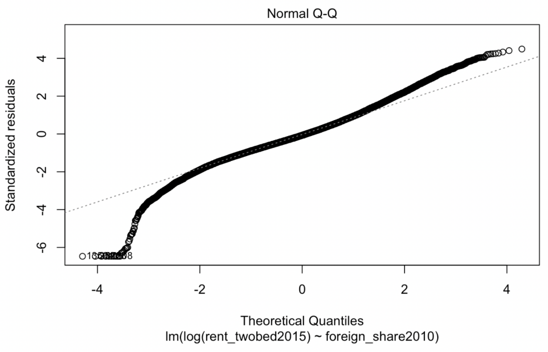
At this point, we can explore our hypotheses more directly, each with four different models that iteratively improve our model assumptions. For each of the four variables, we ran a basic OLS model, an OLS with fixed effects, a Random Intercept, and a Random Intercept with fixed effects. We ran an additional model for high paying jobs in order to explain the phenomenon we were seeing.

Foreign Share vs. Rent Price

OLS Model:

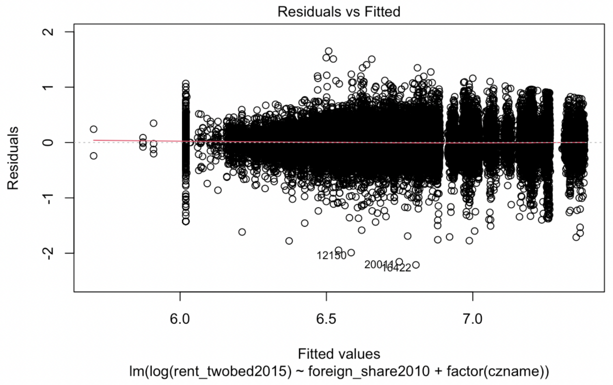
There is a statistically significant positive association between the share of foreign population and rent price per census tract. For each percentage point increase in the foreign share population, there is a 1.18 increase in the log of the rent price, which is a 1.18 multiplicative increase in the rent price. The p-value for this OLS model was $< 2 * 10^{-16}$, meaning the value was statistically significant. However, the data does not fit the assumptions of the OLS model. The residuals seem to be relatively normally distributed, except at the tail ends of the distribution. However, the residuals do not seem to have constant variance, which suggests there is some missing variable that is related to foreign share and rent price, that is skewing the residuals as foreign share increases. For this, we tried running an OLS regression with fixed effects across each city.

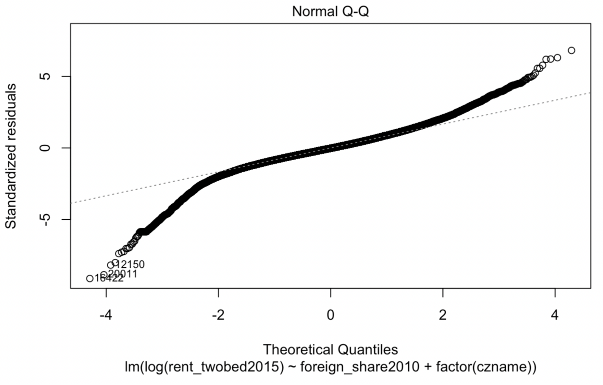




Fixed Effects Model:

When adding fixed effects per city, we observe that the relationship between foreign share and rent price is actually negative. So, after accounting for the differences in the baselines for each city, we still see an association between foreign share and rent price that is significant by a t-test (see model output in the appendix). For each 1% increase in foreign share, there is an approximate .035 decrease in the log of rent price, which means there is a multiplicative decrease of .035 in raw rent price. This suggests that the association, when observing each city, is actually negative between foreign share and rent price. This is more in line with our hypothesis, which was expecting an increase in foreign share would be associated with a decrease in rent price. The assumptions we were concerned with were addressed with the fixed effects model (see appendix for OLS Fixed Effects). The fixed effects OLS model had relatively constant variance in the residuals and relatively normally distributed residuals, in comparison to the regular OLS model.





Random Intercepts:

In running the random intercept model on foreign share to account for the differences in variance across the metropolitan regions, we found there was a 0.024 decrease in the log of rent price for each 1% increase in foreign share. This suggests that there is indeed a negative association even when accounting for the differences in variances across the model. However, there could still be at baseline some cities that have higher shares of foreign populations (take any large metropolis for example), and also, because of some other factor, higher rent prices. We want to take into account the differences in baseline rent prices, which is why we chose to run the fixed effects model.

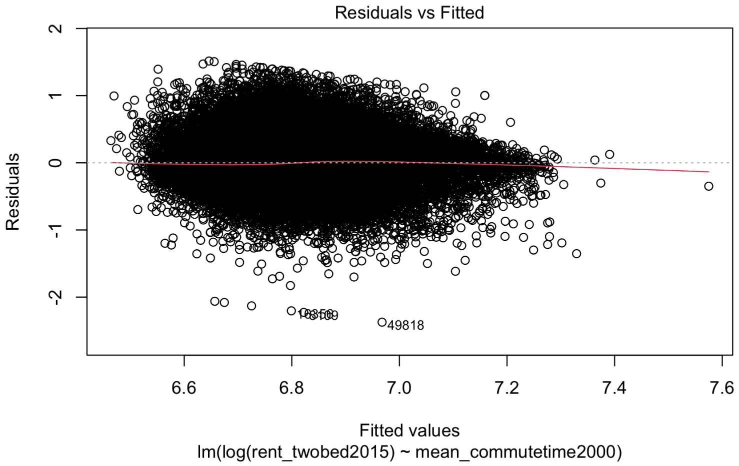
Random Intercepts + Fixed Effects:

In order to observe if this pattern is held across metropolitan districts, we ran a random intercepts model across the foreign share population. A random intercept model with fixed effects is a statistical model that is useful for analyzing data with repeated measures or hierarchical data structures. In the context of understanding the relationship between foreign share and rent price in a census tract, this type of model can account for the potential clustering of observations within each tract, allowing for a more accurate estimate of the relationship between the two variables. Additionally, fixed effects can be included in the model to control for any time-invariant factors that may confound the relationship of interest, such as the overall economic conditions in a given tract. This can help to isolate the effect of foreign share on rent prices, and provide a more precise estimate of the relationship between the two variables.We found that even when accounting for the variance differences across cities with the random intercepts model, the association between foreign share and rent price is still significantly negative. For every 1% increase in foreign share, there is a .035 decrease in the log of rent price, which means a 0.035 multiplicative decrease in the raw rent price (see appendix for specific results of the model).

Mean Commuting Time vs. Rent Price

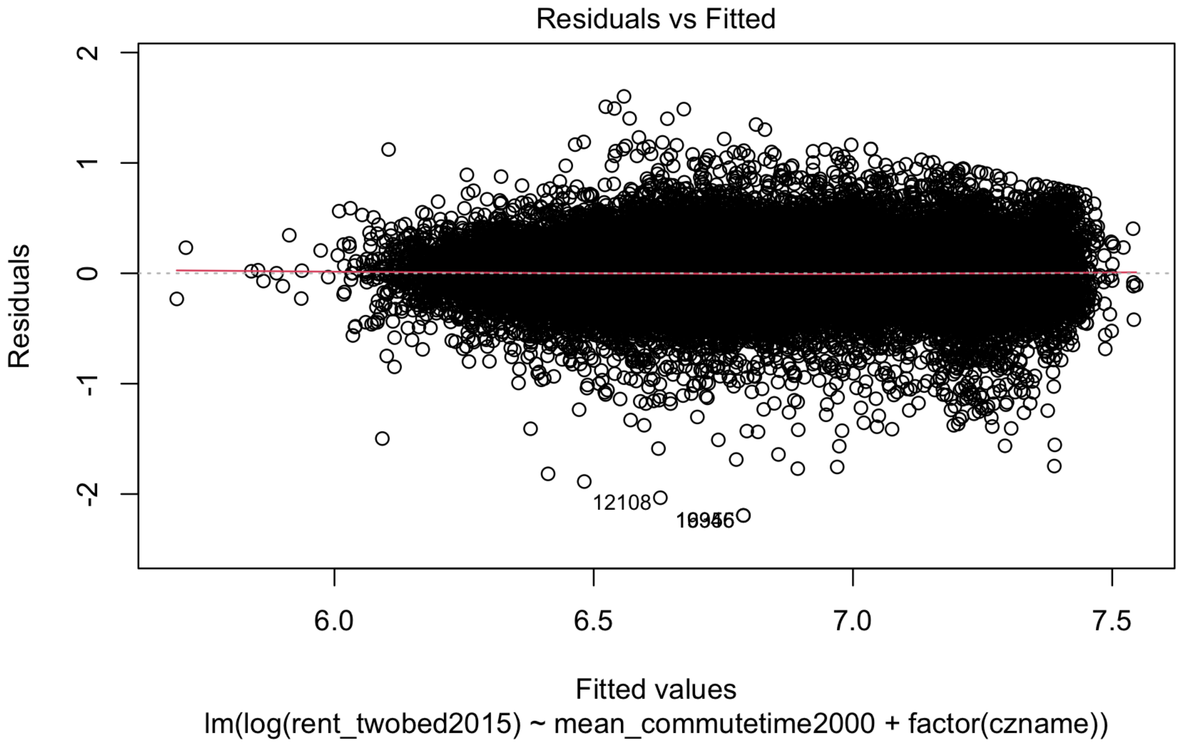
OLS Model:

An OLS model with mean commute time as the only predictor had an estimated slope of 0.016 suggesting that for every unit increase in mean commute time there is a 0.016 increase in the log of rent price (or a 0.016 multiplicative increase in raw rent price). The p-value was significant at < $2 * 10^{-16}$, but there was an issue with the heteroskedasticity of the model, likely owing to baseline differences in commute time between different commuting zones, so we decided to move forward with fixed effects and random intercepts model.



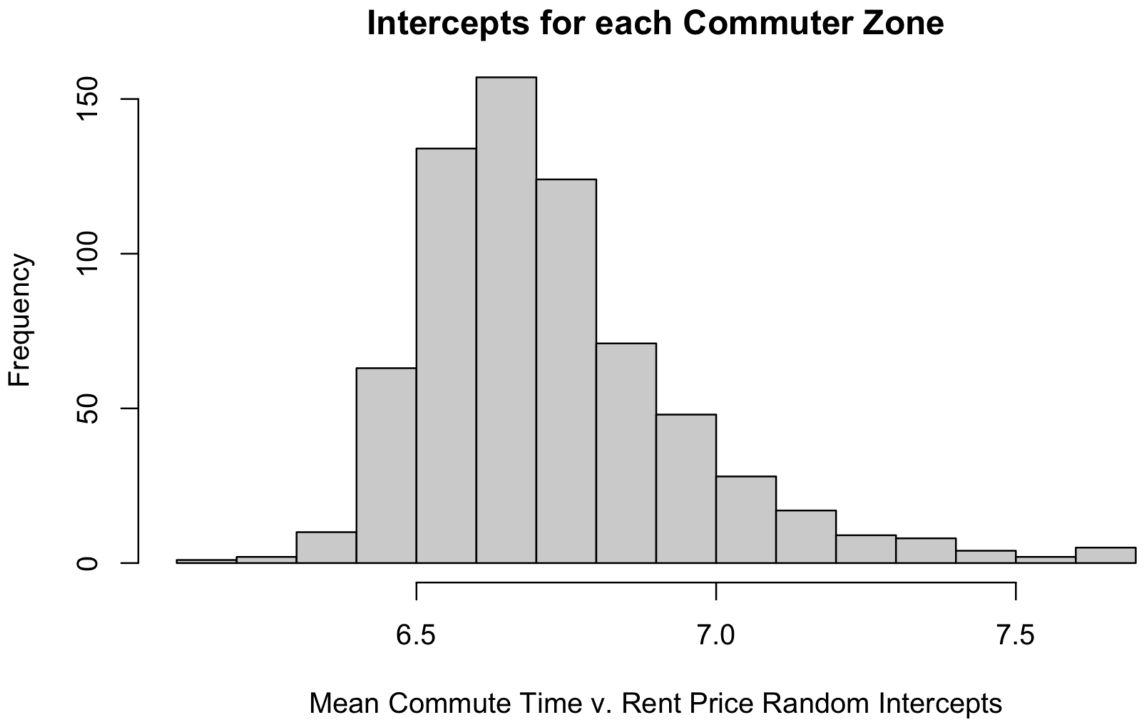
Fixed Effects Model:

The fixed effects model had a slope of -0.010 indicating that for every cubic increase in minority share there is a 0.010 decrease in the log of rent price (0.010 multiplicative decrease in raw rent price) when making commuter zone a fixed effect. From the residual plot it was clear that there was much less heteroskedasticity than before, which can give us more confidence in this model. The fact that the slopes have different signs in the OLS vs. the fixed effects model is probably of minor importance since they're already both so close to 0, and it seems most likely that there is just a weak relationship between commuter zone and rent in the first place.



Random Intercepts:

The results of the random intercept were generally in agreement with the fixed effects model. The slope that characterized the association between the log of rent price and the commuting time was the same as the fixed effects model indicating the same interpretations generally. The only difference is that the intercept is the overall model's intercept is the same, it's just that each group line has an intercept that is randomly generated (with a variance of 0.55 in this case), as opposed to the fixed effect where the intercepts of the groups are entirely separate and determined in a fixed way.

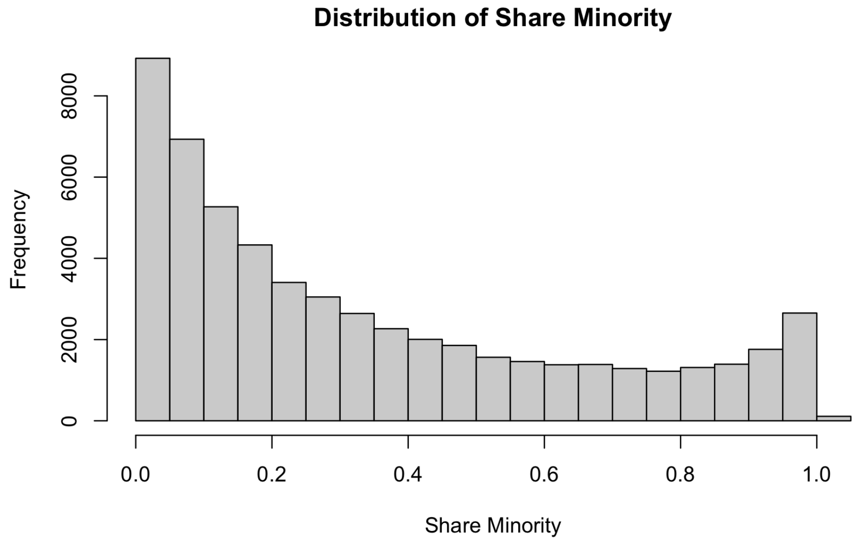


Fixed Effects + Random Intercepts:

This model had the similar results as both the fixed effects and random intercepts in that the differences from the intercepts and the slope were the same as in the fixed effects model. The variance in group intercept was lower, indicating that a lower amount of variance was needed to distinguish between the group-specific differences in baseline between the groups, likely because it is partly already accounted for by the fixed effects.

Minority Share vs. Rent Price

We originally intended to look for the association between the minority share of the population (percentage of the population that is Black, Asian or Hispanic since we only had this demographic information available to us). However, the distribution of the minority shares was right skewed and has many outliers high in the distribution. There were many outliers high in the distribution, and when plotting these values against the log of rent, it appeared as though the observations with high minority shares had a wide range of rent prices, so we chose to drop these points before fitting the model to improve equal variance for a linear model. Since the distribution was now right skewed after flipping, we needed to transform it and chose to cube root it as this seemed to make the distribution the most normal.

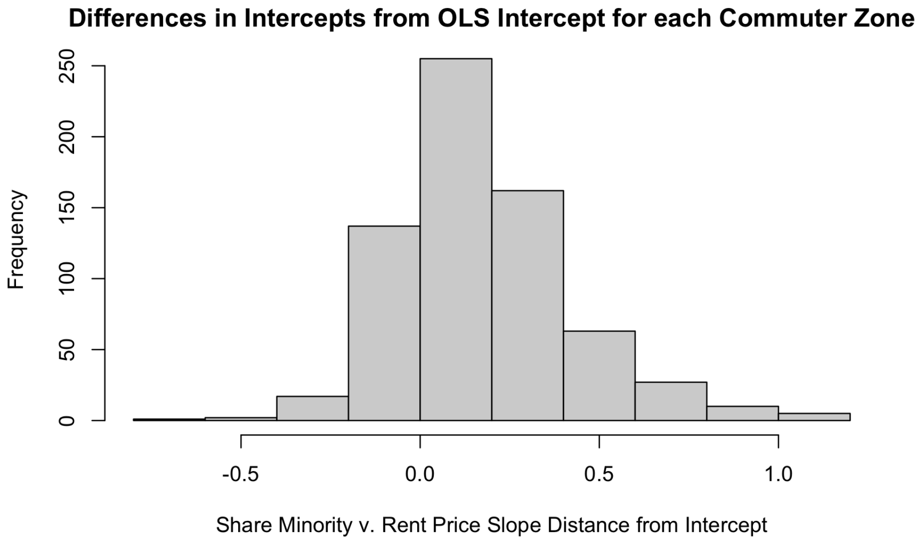


OLS Model:

An OLS model with the transformed share minority as the only predictor had an estimated slope of 0.504 suggesting that for every cubic increase in share minority there is a 0.504 increase in the log of rent price (or a 0.504 multiplicative increase in raw rent price). The p-value was significant at < $2 * 10^{-16}$, but it seemed unlikely that there would be a positive association between the share of minority and rent price, so we wondered if this owed to differences in commuter zone since more populous commuter zones are more likely to have both higher shares of minority and higher rents, so we decided to use fixed effects and random intercepts to see if distinguishing between commuter zone led to different results.

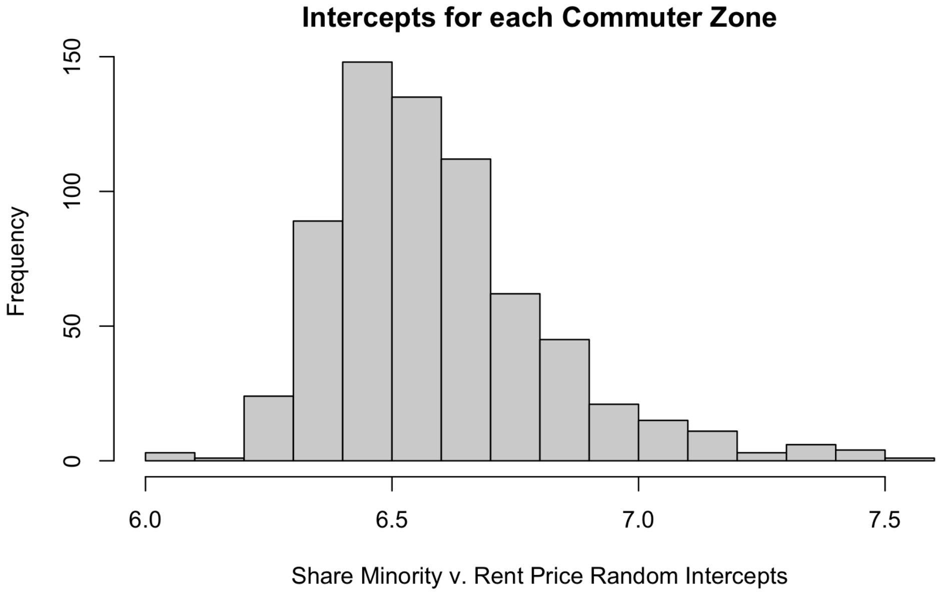
Fixed Effects Model:

The fixed effects model had a slope of -0.164 indicating that for every cubic increase in minority share there is a 0.164 decrease in the log of rent price (0.164 multiplicative decrease in raw rent price) when making commuter zone a fixed effect. This confirms that each commuter zone has a different baseline and that affects the overall prediction of rent price from share minority, as we see that when we distinguish between commuter zone the association is negative instead of positive (as the OLS model suggested).



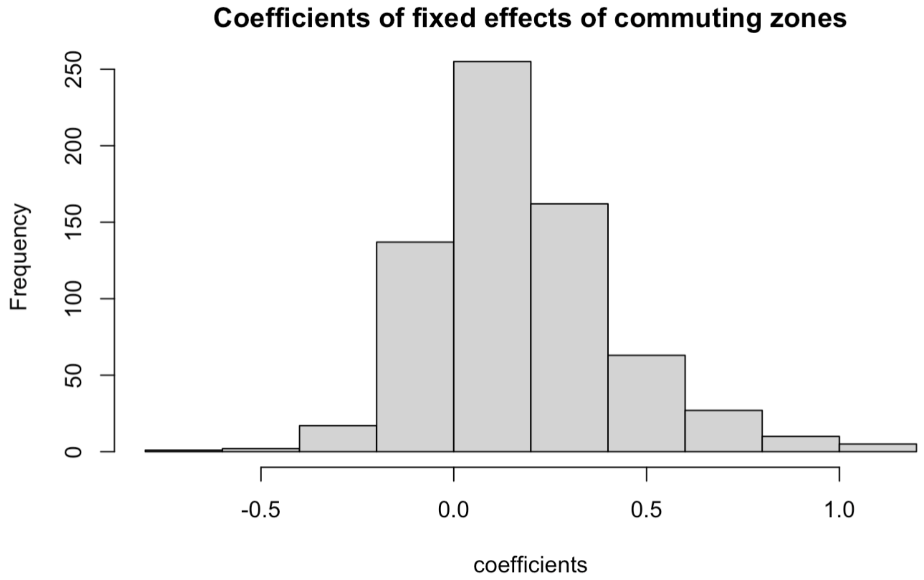
Random Intercepts:

The random effects model found an association between the log of rent price and the cube root of share minority of -0.154 which is decently close to that of the random effects model, with a variance around the overall intercept of 0.054. The only difference is that the intercept is the overall model's intercept is the same, it's just that each group line has an intercept that is randomly generated, as opposed to the fixed effect where the intercepts of the groups are entirely separate and determined in a fixed way.



Fixed Effects + Random Intercepts:

This model had the similar results as both the fixed effects and random intercepts in that the differences from the intercepts and the slope were the same as in the fixed effects model. The variance in group intercept was higher, indicating that a higher amount of variance was needed to distinguish between the group-specific differences in baseline between the groups, which is interesting because baseline differences between groups were already accounted for in part by the fixed effects that changed the intercepts. It could be that the extra amount of variance in the random effects layer was needed to explain further differences between groups that were not seen before.

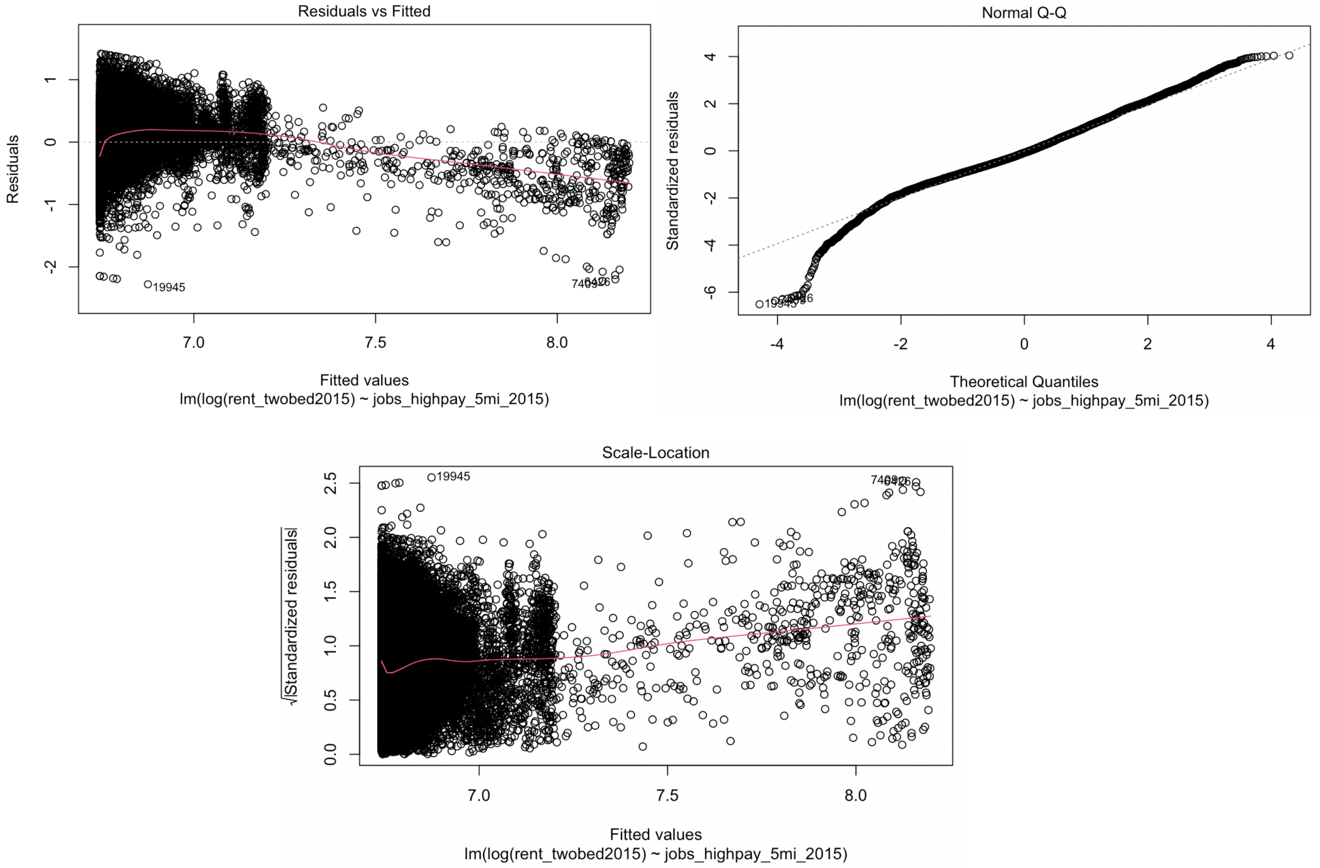


High Paying Jobs and Rent Price

OLS Model:

When fitting an OLS model to predict the logarithm of rent price from the number of high paying jobs in a 5 mile radius, we can see that there is a positive association between the amount of high paying jobs and rent price. The positive association is statistically significant since the p-value is less than < $2e^{-16}$. However, the slope is very small (8.117e-07), meaning that every one unit increase in high paying jobs is associated with a 8.117e-07 increase in the log of rent price (or, in other words, a 8.117e-07 multiplicative increase in rent price).

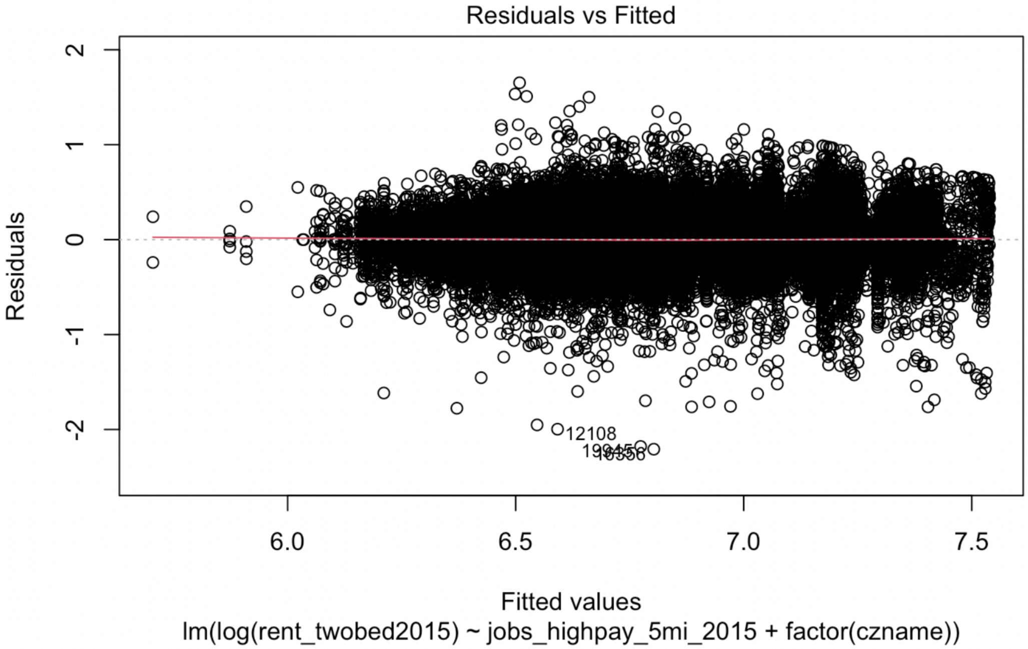
It appears that the assumptions of the linear model do not hold very well. From the residuals vs fitted plot we can see that the model tends to underestimate for low values and overestimate for higher values). We can also see that the residuals are not normally distributed and the variance is not constant across fitted values. This might be because of baseline differences in the number of high paying jobs between different commuting zones, so we decided to move forward with fixed effects and random intercepts models.



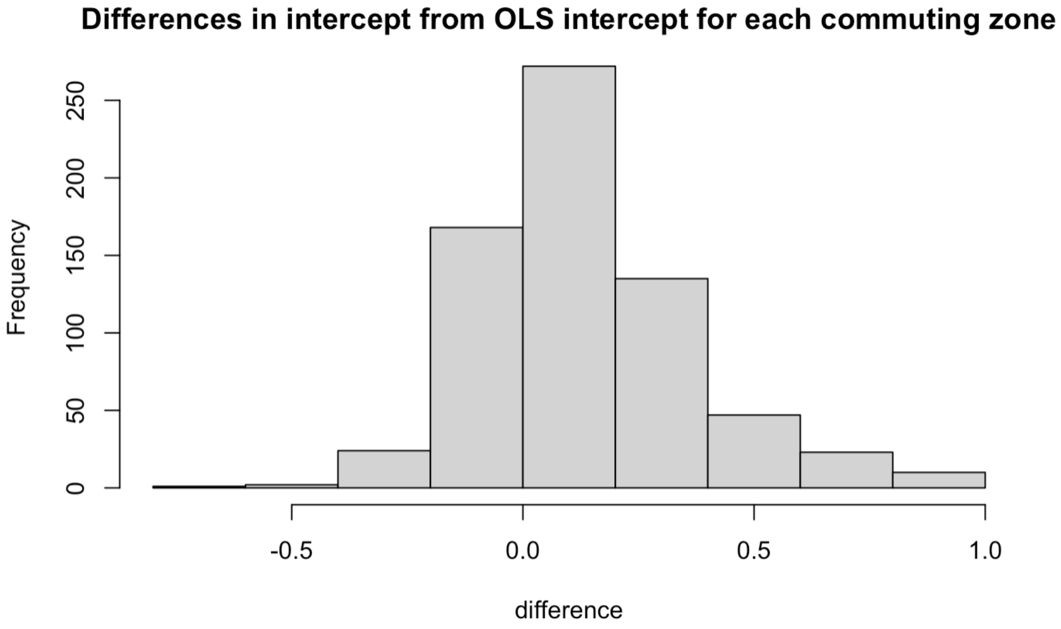
Given this very weak relationship between the number of high paying jobs and rent price , we decided to also look at the density of high paying jobs by creating a new predictor that is the ratio of the number of high paying jobs in a 5 mile radius and the total number of jobs in a 5 mile radius, since a percentage value is more informative than just looking at absolute numbers of jobs. Fitting a linear regression model on this new predictor (let's call it "high-pay job percentage"), gives us a much higher positive slope than the previous linear model. The slope in this case is 1.550931, meaning that every one unit increase in the density of high paying jobs is associated with a 1.550931 increase in the log of rent price (or, in other words, a 1.550931 multiplicative increase in rent price). We can see that there is a much stronger association between the density of high paying jobs and rent price, than between the number of high paying jobs and rent price. The p-value of the slope coefficient is less than < 2e-16, which means that the positive association is statistically significant.

Fixed Effects Model:

In the fixed effects model, jobs_highpay has a slope coefficient of 2.059e-07, which indicates that there is a positive relationship between the number of high paying jobs and rent price when we make commuting zone a fixed effect and that every one unit increase in high paying jobs is associated with a 2.059e-07 increase in the log of rent price (or, in other words, a 2.059e-07 multiplicative increase in rent price). The positive association is statistically significant since the p-value is less than < $2 \ * \ 10^{-16}$. From the residuals vs fitted plot we can see that the linearity and constant variance assumptions hold for the fixed effects model (much better than in the baseline OLS model).

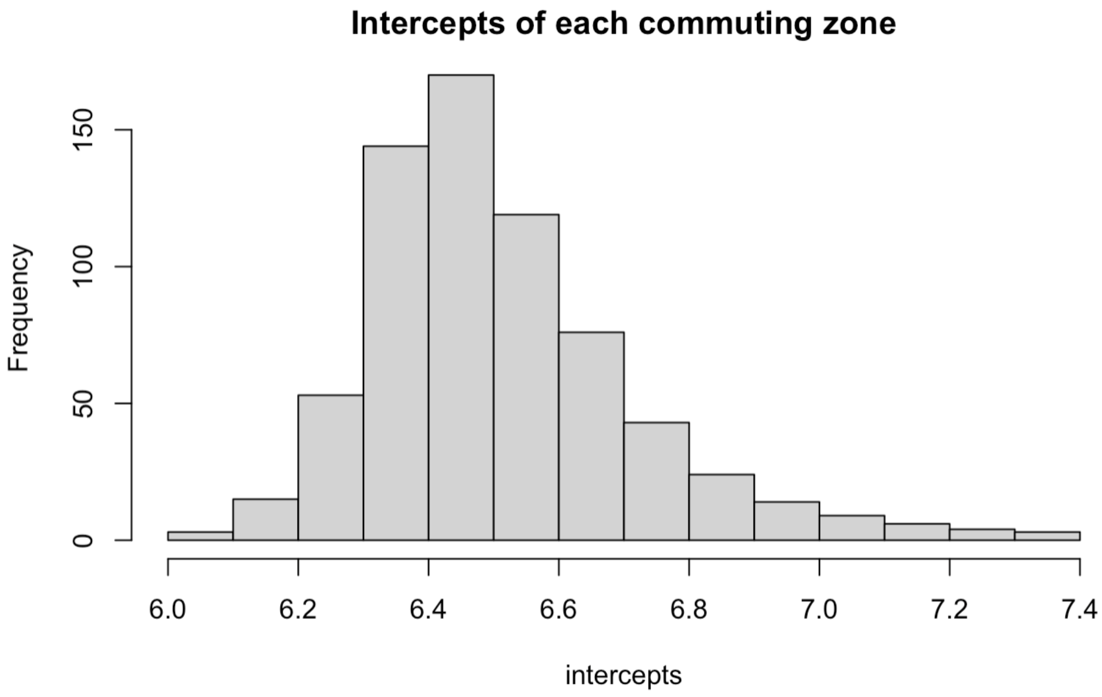


We can see that different commuting zones have different intercepts (some larger, some smaller than the OLS intercept). In general, the difference between each commuting zone intercept and the OLS intercept is normally distributed, centered around a positive value that is close to zero. This confirms that each commuting zone has a different baseline but we can see that each commuting zone has an effect on the overall prediction of rent price from jobs_highpay as for some of them the association is negative instead of positive.



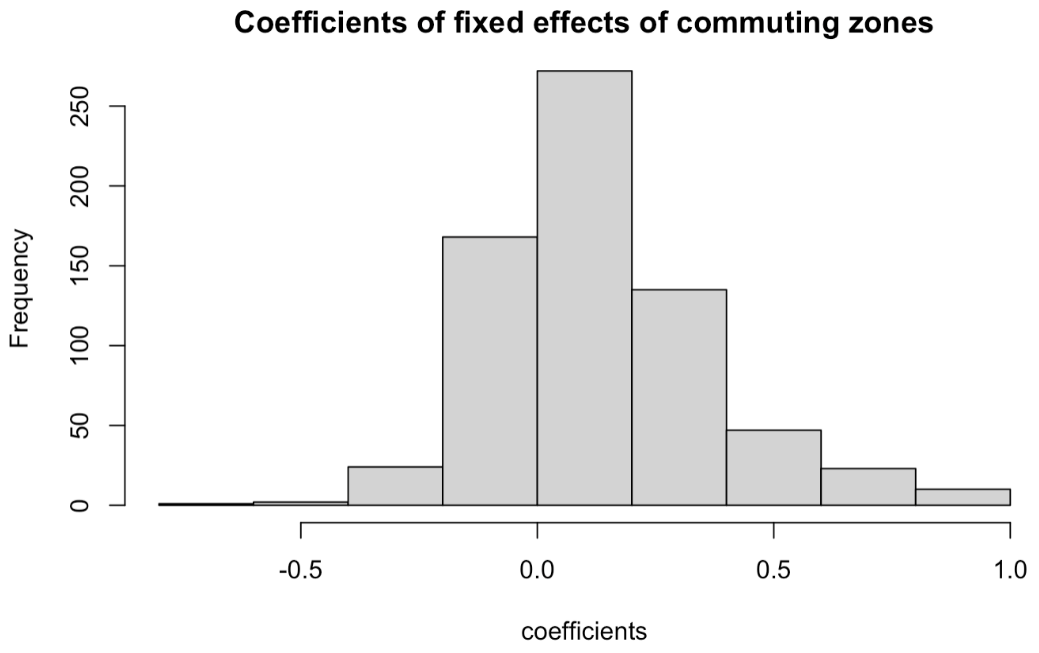
Random Intercepts:

In the random intercepts model, jobs_highpay has a slope coefficient of 2.083e-07 (very similar to the fixed effects model), indicating that there is a positive association between the number of high paying jobs and rent price, in particular that each one unit increase in high paying jobs is associated with a 2.083e-07 multiplicative increase in rent price. In this model the intercept for each commuting zone is randomly generated with an overall mean of 6.513 and variance of 0.04701.



Fixed Effects + Random Intercepts:

The relationship between jobs_highpay and rent price is very similar to the other two models above, with a slope coefficient of 2.059e-07. The variance in group intercept (0.005342) is lower than in the random intercept model above, indicating that less variance was needed to distinguish between the group-specific differences in baseline between the groups. This is because baseline differences between groups were already accounted for in part by the fixed effects that changed the intercepts. We can see that the coefficients of the fixed effects for each commuting zone are also very similar to the ones found in the fixed effects model when we compare the two distribution plots.



Interaction model:

We also fit an interaction model to explore if there is an interaction between the number of high paying jobs and mean commute time in the relationship with rent price. The slopes of the main effects of the number of high paying jobs and mean commute time are both positive and statistically significant (p-value is less than < 2e-16), meaning that more high paying jobs are associated with higher rent and that higher commute time has a weak positive association with rent prices. However, the interaction term between the two predictors is negative and statistically significant. This means that if there are a lot of high paying jobs and the mean commute time is also high then the rent prices get lower. This makes sense because this would mean that the people living in that area are not working in the high paying jobs but rather commuting elsewhere (perhaps to work in lower paying jobs) and thus rent prices are lower.

VI. CONCLUSION

Results:

Through the models we fit above, we were able to test our hypothesis and confirm them for the most part. In conclusion, we found that:

- There is a negative association between foreign share and rent price when we account for the fixed effects of each commuting zone.

- There is a weak negative association with mean commute time and rent price when we account for the fixed effects of each commuting zone.

- There is a negative association between the share of minority residents and rent prices. Also, each commuter zone has a different baseline and affects the overall prediction of rent price from the share of minority, with more populous commuter zones having both high share of minority and high rent prices.

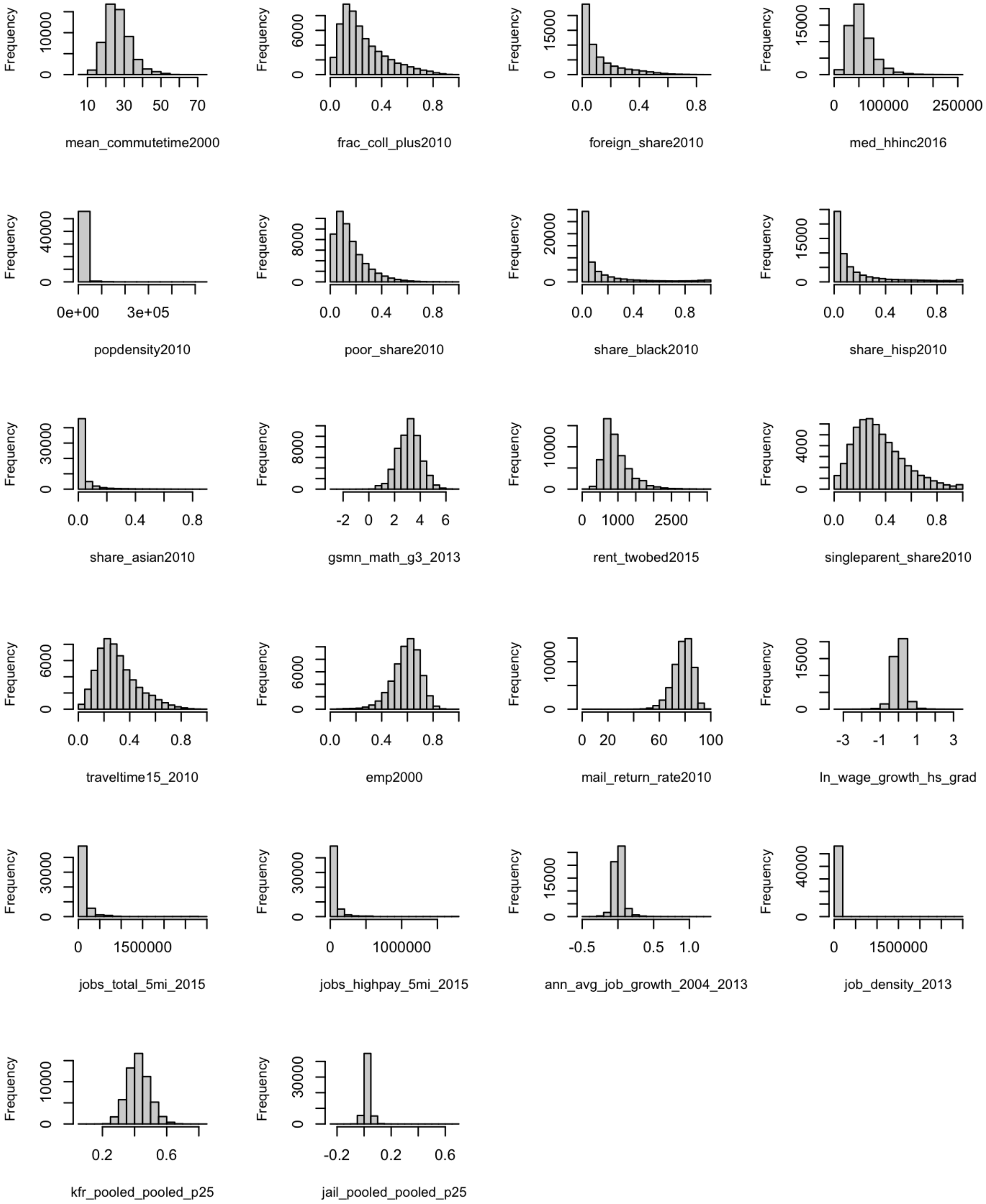
- There is a positive association between the number of high paying jobs and rent price and there is even a stronger positive association between the density of high paying jobs and rent price. This relationship shows some variation across different commuting zones and has an interactive effect with the mean commute time.

Future Work:

These models were helpful in understanding the relationship between our 4 predictors and rent price. We would likely want to choose the random intercept model with the fixed effect to determine any more specific associations between the predictor and the rent price. Random intercept model with fixed effects takes into account both the difference in baseline per city and the difference in variance per city, which allowed us to isolate the effect of the predictor on rent price. However, it would be useful to compare the OLS with Fixed Effects and the Random Intercept with Fixed Effects as a robustness check. This could be in the form of comparing the AIC or BIC outputs for the models.

In future work, we could also explore the predictive power of the models and evaluate them on a testing dataset by predicting rent prices from the various predictors and calculating the RMSE.

VII. APPENDIX

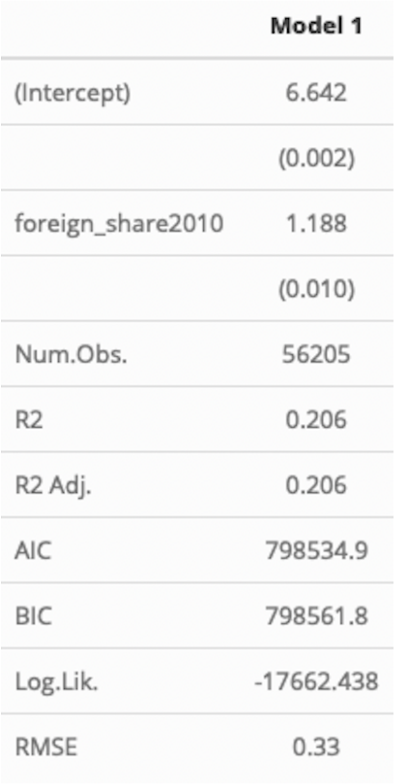


Appendix Figure 1. Distributions of all 44 quantitative variables, including the response (*rent_twobed_2015*).

Models output:

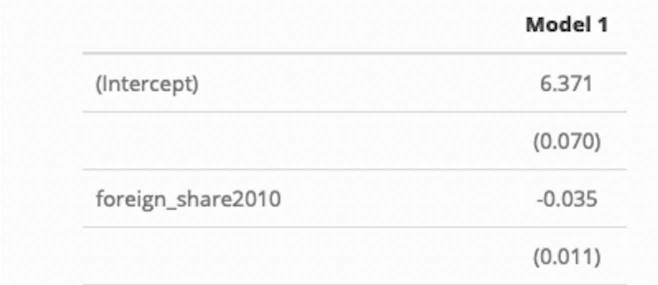
<u>Foreign Share vs. Rent Price</u>

- *OLS Model:*



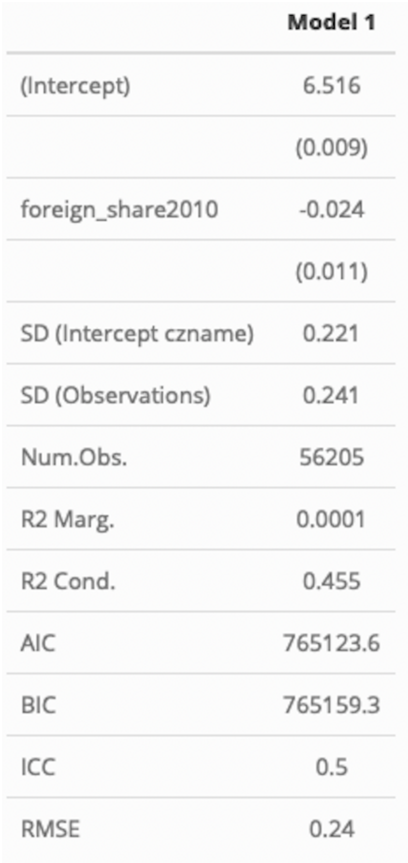
	Model 1
(Intercept)	6.642
	(0.002)
foreign_share2010	1.188
	(0.010)
Num.Obs.	56205
R2	0.206
R2 Adj.	0.206
AIC	798534.9
BIC	798561.8
Log.Lik.	-17662.438
RMSE	0.33

- *Fixed Effects Model:*



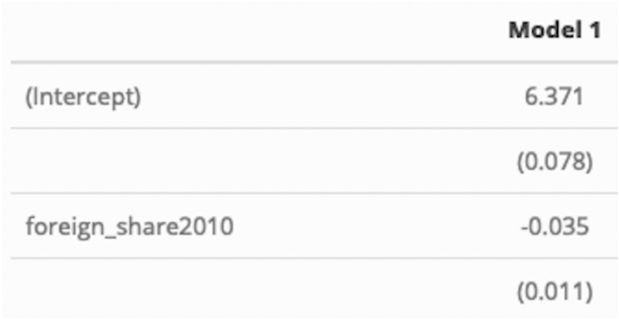
	Model 1
(Intercept)	6.371
	(0.070)
foreign_share2010	-0.035
	(0.011)

- *Random Intercepts:*



	Model 1
(Intercept)	6.516
	(0.009)
foreign_share2010	-0.024
	(0.011)
SD (Intercept czname)	0.221
SD (Observations)	0.241
Num.Obs.	56205
R2 Marg.	0.0001
R2 Cond.	0.455
AIC	765123.6
BIC	765159.3
ICC	0.5
RMSE	0.24

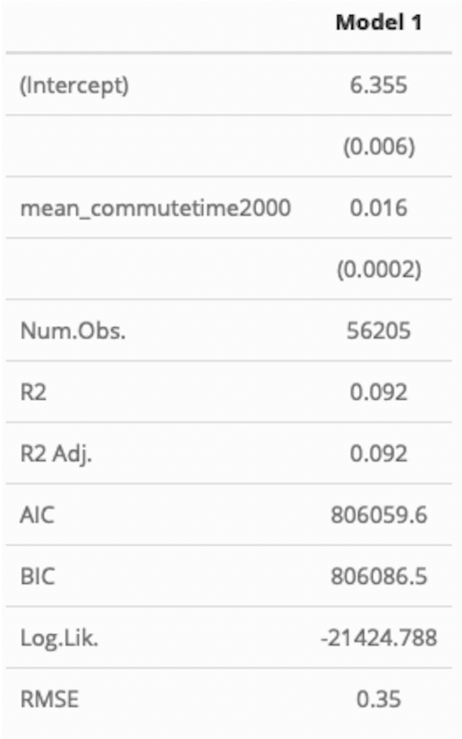
- *Fixed Effects + Random Intercepts:*



	Model 1
(Intercept)	6.371
	(0.078)
foreign_share2010	-0.035
	(0.011)

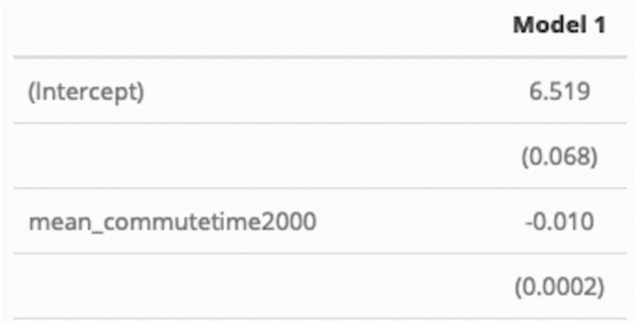
<u>Mean Commuting Time vs. Rent Price</u>

- *OLS Model:*



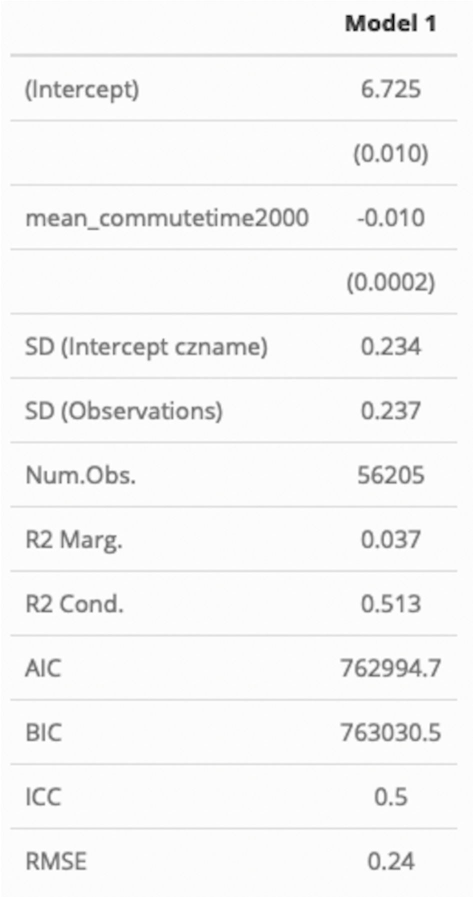
	Model 1
(Intercept)	6.355
	(0.006)
mean_commutetime2000	0.016
	(0.0002)
Num.Obs.	56205
R2	0.092
R2 Adj.	0.092
AIC	806059.6
BIC	806086.5
Log.Lik.	-21424.788
RMSE	0.35

- *Fixed Effects Model:*



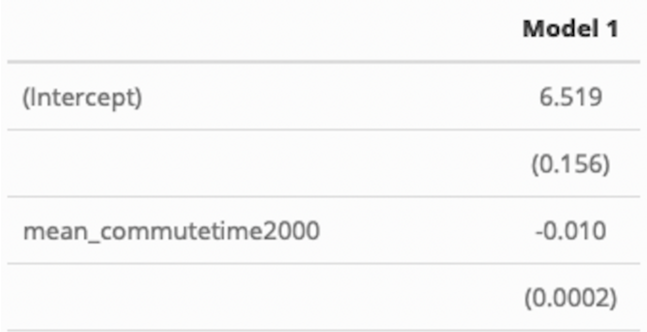
	Model 1
(Intercept)	6.519
	(0.068)
mean_commutetime2000	-0.010
	(0.0002)

- *Random Intercepts:*



	Model 1
(Intercept)	6.725
	(0.010)
mean_commutetime2000	-0.010
	(0.0002)
SD (Intercept czname)	0.234
SD (Observations)	0.237
Num.Obs.	56205
R2 Marg.	0.037
R2 Cond.	0.513
AIC	762994.7
BIC	763030.5
ICC	0.5
RMSE	0.24

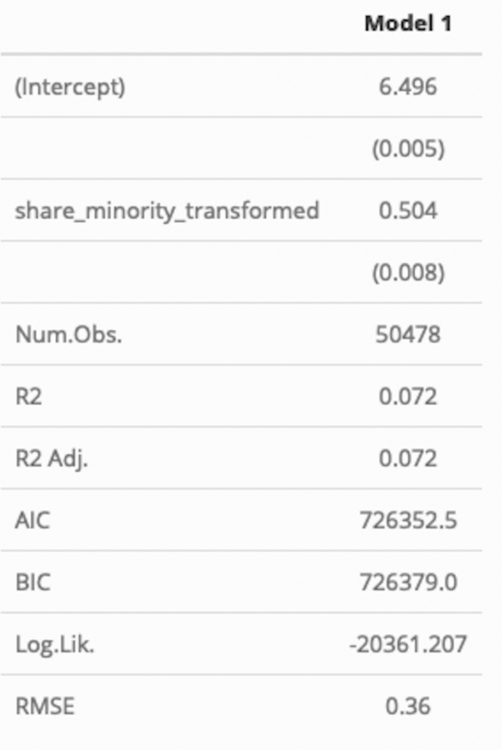
- *Fixed Effects + Random Intercepts:*



	Model 1
(Intercept)	6.519
	(0.156)
mean_commutetime2000	-0.010
	(0.0002)

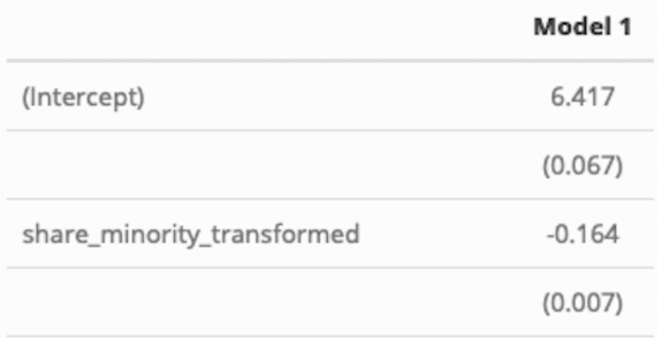
Minority Share vs. Rent Price

- *OLS Model:*



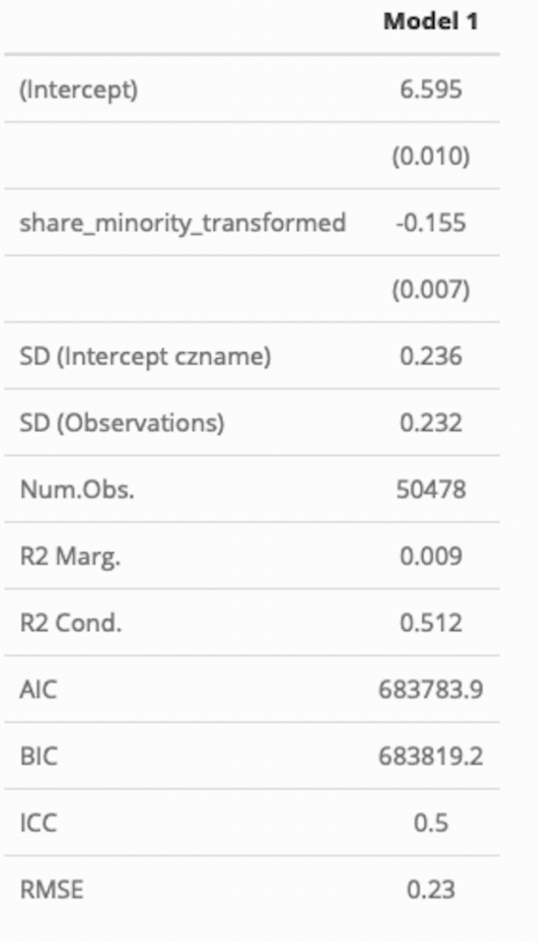
	Model 1
(Intercept)	6.496
	(0.005)
share_minority_transformed	0.504
	(0.008)
Num.Obs.	50478
R2	0.072
R2 Adj.	0.072
AIC	726352.5
BIC	726379.0
Log.Lik.	-20361.207
RMSE	0.36

- *Fixed Effects Model:*



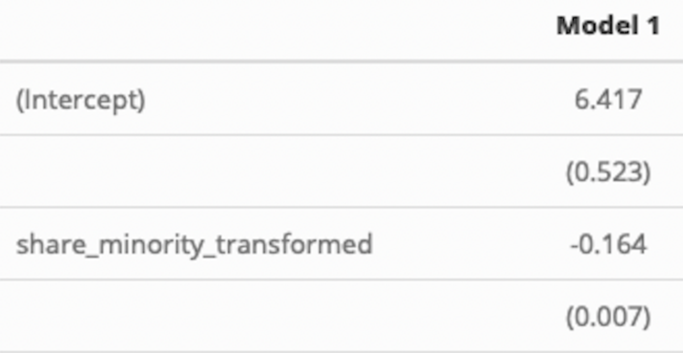
	Model 1
(Intercept)	6.417
	(0.067)
share_minority_transformed	-0.164
	(0.007)

- *Random Intercepts:*



	Model 1
(Intercept)	6.595
	(0.010)
share_minority_transformed	-0.155
	(0.007)
SD (Intercept czname)	0.236
SD (Observations)	0.232
Num.Obs.	50478
R2 Marg.	0.009
R2 Cond.	0.512
AIC	683783.9
BIC	683819.2
ICC	0.5
RMSE	0.23

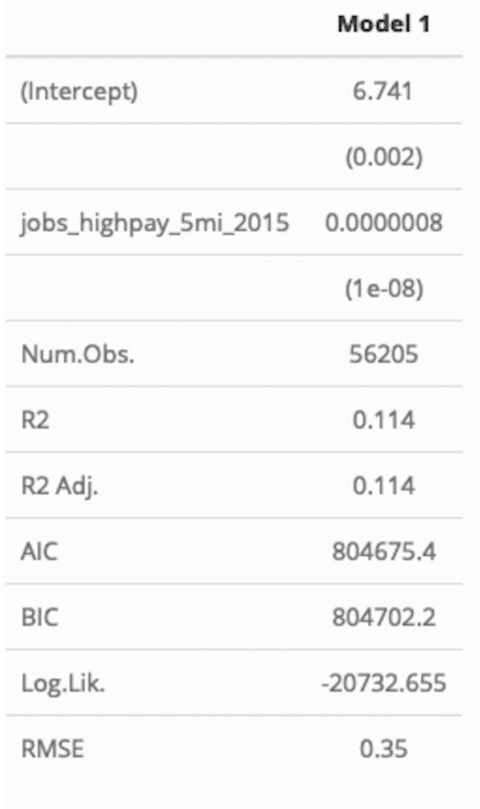
- *Fixed Effects + Random Intercepts:*



	Model 1
(Intercept)	6.417
	(0.523)
share_minority_transformed	-0.164
	(0.007)

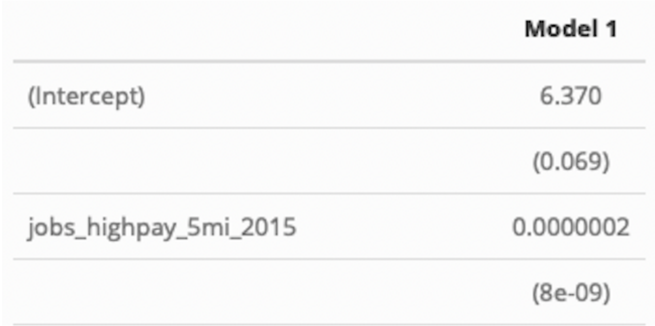
<u>High Paying Jobs vs. Rent Price</u>

- *OLS Model:*



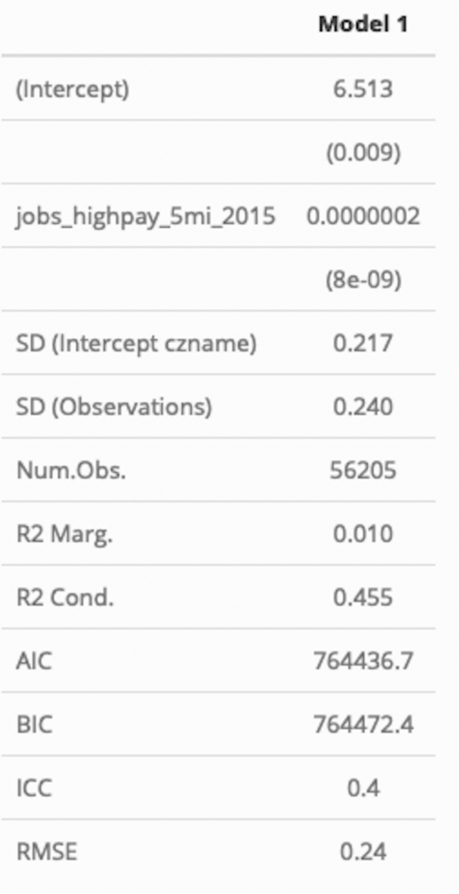
	Model 1
(Intercept)	6.741
	(0.002)
jobs_highpay_5mi_2015	0.0000008
	(1e-08)
Num.Obs.	56205
R2	0.114
R2 Adj.	0.114
AIC	804675.4
BIC	804702.2
Log.Lik.	-20732.655
RMSE	0.35

- *Fixed Effects Model:*



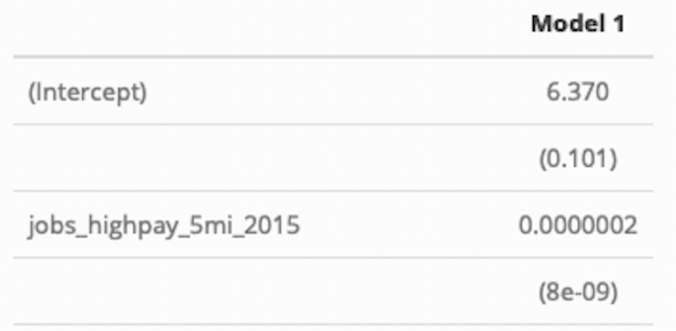
	Model 1
(Intercept)	6.370
	(0.069)
jobs_highpay_5mi_2015	0.0000002
	(8e-09)

- *Random Intercepts:*



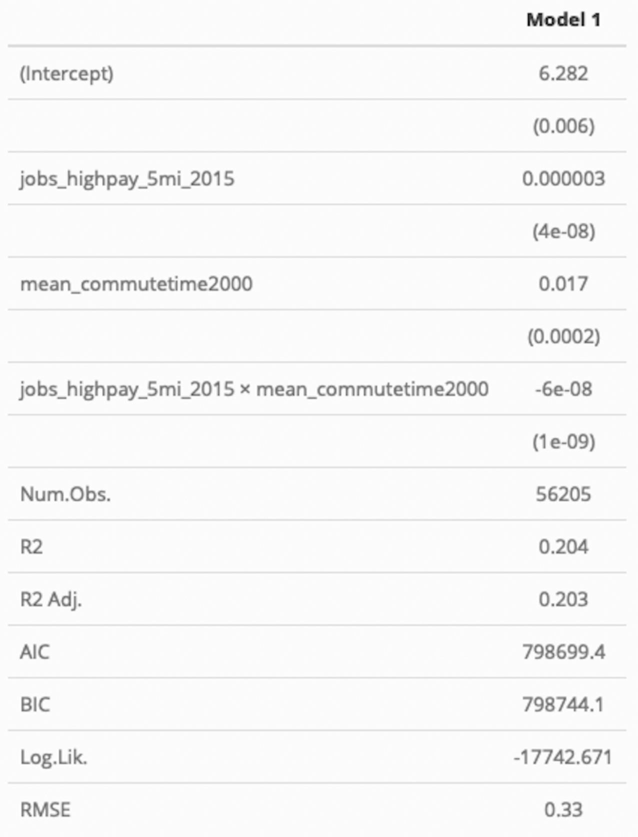
	Model 1
(Intercept)	6.513
	(0.009)
jobs_highpay_5mi_2015	0.0000002
	(8e-09)
SD (Intercept czname)	0.217
SD (Observations)	0.240
Num.Obs.	56205
R2 Marg.	0.010
R2 Cond.	0.455
AIC	764436.7
BIC	764472.4
ICC	0.4
RMSE	0.24

- *Fixed Effects + Random Intercepts:*



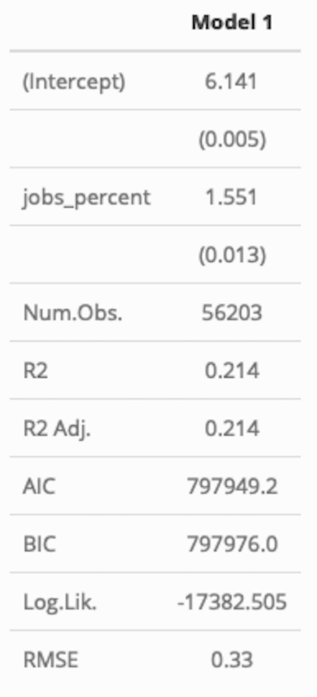
	Model 1
(Intercept)	6.370
	(0.101)
jobs_highpay_5mi_2015	0.0000002
	(8e-09)

- *Interaction model:*



	Model 1
(Intercept)	6.282
	(0.006)
jobs_highpay_5mi_2015	0.000003
	(4e-08)
mean_commutetime2000	0.017
	(0.0002)
jobs_highpay_5mi_2015 × mean_commutetime2000	-6e-08
	(1e-09)
Num.Obs.	56205
R2	0.204
R2 Adj.	0.203
AIC	798699.4
BIC	798744.1
Log.Lik.	-17742.671
RMSE	0.33

- *Job density OLS model:*



	Model 1
(Intercept)	6.141
	(0.005)
jobs_percent	1.551
	(0.013)
Num.Obs.	56203
R2	0.214
R2 Adj.	0.214
AIC	797949.2
BIC	797976.0
Log.Lik.	-17382.505
RMSE	0.33