SOCI: 20003 Final Project: Analysis

Contents

Analysis of demographic change in Berlin

Gentrification has emerged as a leading issue in urban studies, though attempts to quantitative evaluate it are highly varied in their theories, methods, and results. In this report, I attempt to gain insight into whether gentrification, indicated by changes in local price real estate valuations and demographics, is associated with a reduction in low-income residents, a process that can be described by displacement. This analysis is done using demographic data from Berlin's biannual Monitoring Soziale Stadtentwicklung report (monitoring social city development) and real estate valuation data from the Bodenrichtwert report, the city's yearly parcel value estimates.

My methods to prepare the data can be accessed at https://rpubs.com/cschwab98/cschwab_final_dataprep

1. Background and Literature Review

1.1 Background Gentrification is commonly considered to be established as a concept in England in the 1960s as working class neighborhoods became increasingly wealthy and poorer residents were forced to move out (Barton, 2014). It is usually defined as a confluence of real estate value increased caused by a rent gap (a gap between the offering prices of an area and the real value), the displacement of working class and/or minority residents, and cultural changes to suit to the wealthier incoming population.

1.2 Literature Review Despite the prevalence of gentrification in popular discourse, actually measuring the process is difficult given the confluence of variables, the lack of a robust definition, and the heterogeneity of areas said to be gentrified. In Preis et al (2020), the authors examine city government-lead attempts to measure gentrification from Seattle, Portland, Los Angeles, and Philadelphia, and apply each city's distinct method to Boston. The different models produce drastically different results in which census tracts they identify as at-risk for gentrifying and gentrified — the average percent of tracts recognized as gentrified in one method retained by other methods is 63%. In Berlin, there have been a few moderately successful attempts to quantify gentrification and displacement. Schulz (2017) examined a privately acquired data source on rental offerings on the post-code level and changes in poverty, and concluded that there were strong indications of displacement between 2007-2013 using a spatial autoregression model. Helweg (2018) compared demographic data to amenities data offered by Open Street Maps with numerous machine learning techniques but was unable to find any substantial relationship between amenity structure and demographic change.

2. Variables: Explaination and Descriptive statistics

In order to better understand the relationship between real estate price increase values and demographic change I compare the change in estimated parcel value with six demographic variables between 2003 and 2011: unemployment, long-time unemployment (greater than 1 year), Germans receiving welfare, seniors, foreigners, and EU citizens. The data is normalized to percent (the number of each group per 100 residents) and spatially aggregated to grid cells of 45,000 m^2 over the entirety of Berlin.

However, the normal causal relationship of the variables is here necessarily somewhat reversed — we can look at the negative correlation between

2.1 Variable descriptions: Each of these variables are measured continuously on an interval scale. The variable I chose to represent the displacement of low-income residents typically understood as displacement is those who are on welfare for longer than two years: `gWelfareC` = change in residents receiving welfare for longer than two years

Meanwhile, the predictor variables I chose to use are the six demographic variables and my real estate variable: `unempC` = change in unemployed working age residents `longunempC` = change in residents unemployed for >1 year `seniorsC` = change in seniors `youthC` = change in youth (under 18) `euC`= change in EU citizens `foreignC` = change in foreigners (those born outside Germany) `brwC` = change in local parcel price estimate

To provide further context for the `brwC` variable: each year the Berlin Gutachterausschuss provides a set of dummy parcels around the city and their estimated price per square meter. Similar to the Assessor's office, the methods for these valuations are kept close to the chest, but are said to be based off similar parcels of a similar condition and location (I'm currently working on obtaining the full list of actual property transactions from the Berlin Gutachterausschuss office, but post WWII data laws tend to make the Germans a bit stingy about providing open data). I calculated `brwC` by downloading these valuations for 2003 and 2011 alongside the years directly before and after (e.g. 2002, 2004), rasterizing each file, interpolating some of the missing values with no parcel intersections, and averaging the price; this is described in the `Data Prep` document.

It is also notable that the demographic variables are not reported for areas with insufficient population, as determined by the Berlin Senate's office based on undisclosed methods, and thus already excluded from this analysis. Additionally, the method of interpolation via rasterization eliminated a number of areas from the analysis — while robust interpretation of the potential adverse causes of this elimination is beyond the scope of this report, areas eliminated are likely sparsely populated.

2.2 Pre-regression assumptions and testing: Summary statistics for each variable is displayed below:

```
summary(bdataC[,2:8])
```

```
##        y              unempC          longunempC       gWelfareC
##  Min.   :3254104   Min.   :-14.600   Min.   :-7.200   Min.   :-2.200
##  1st Qu.:3264904   1st Qu.: -5.900   1st Qu.:-2.100   1st Qu.: 3.400
##  Median :3271384   Median : -4.500   Median :-1.400   Median : 6.700
##  Mean   :3271192   Mean   : -4.819   Mean   :-1.616   Mean   : 7.875
##  3rd Qu.:3276352   3rd Qu.: -3.500   3rd Qu.:-1.000   3rd Qu.:11.200
##  Max.   :3288448   Max.   :  9.500   Max.   : 4.100   Max.   :28.100
##     seniorsC          youthC          foreignC
##  Min.   :-13.500   Min.   :-14.20   Min.   :-66.900
##  1st Qu.:  2.300   1st Qu.: -3.50   1st Qu.: -0.600
##  Median :  5.600   Median : -1.30   Median :  0.200
##  Mean   :  5.099   Mean   : -1.76   Mean   :  0.207
##  3rd Qu.:  7.500   3rd Qu.: -0.10   3rd Qu.:  1.400
##  Max.   : 26.700   Max.   : 10.90   Max.   : 32.700
```

Before I perform any regression, there are certain assumptions that must be made about the relationship between the data I'm examining:

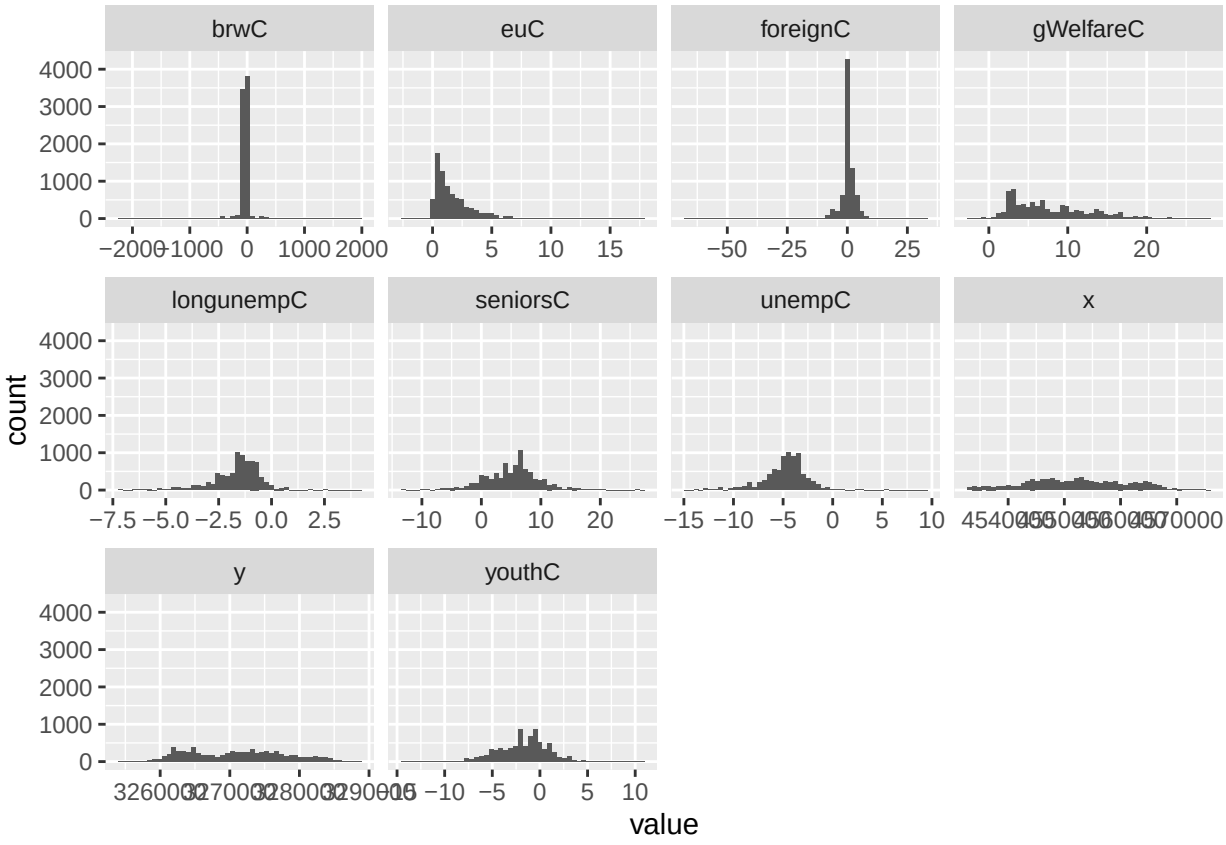
1. That the underlying data has been correctly processed and calculated
2. That the variables are linearly related
3. That the variables are normally distributed
4. That the variance across all levels of the predicted variables is constant (heterodasticity)

In addressing these assumptions, the first is something we can't really check without looking at the correct data, which I would be using if I had.

2.2.1 Normality

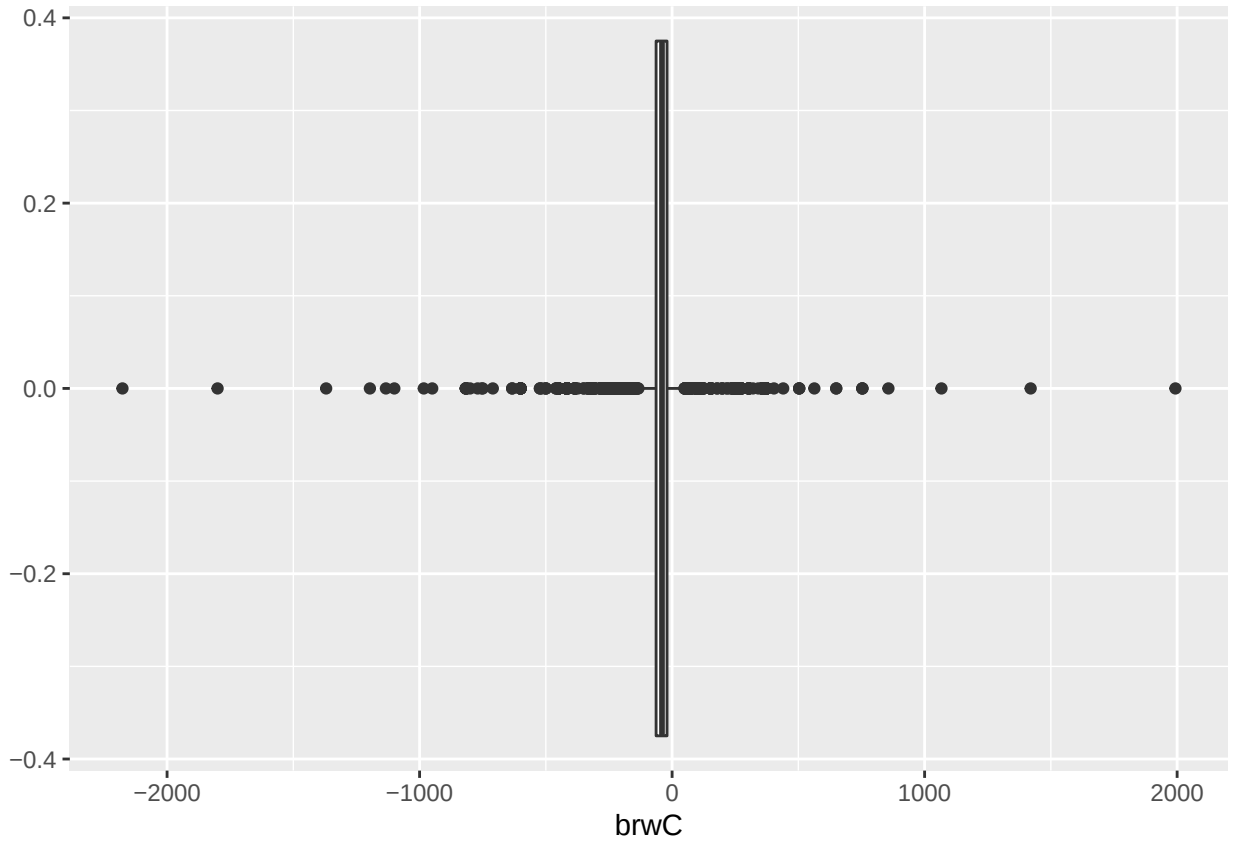
To address the second assumption, we examine the distribution of each variable:

```
ggplot(gather(bdataC), aes(value)) +
    geom_histogram(bins = 50) +
    facet_wrap(~key, scales = 'free_x')
```

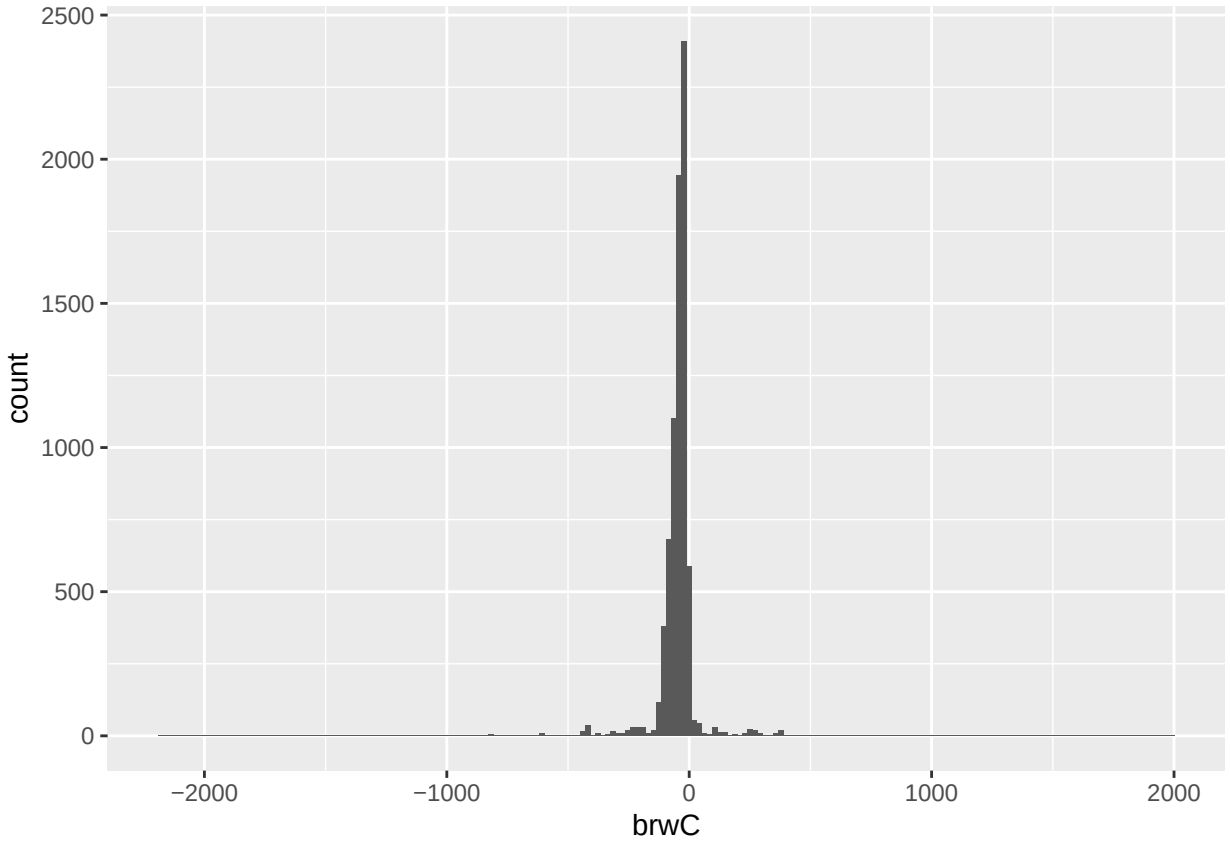


Based on these summary histograms, the variables are generally quite normally distributed, with some notable exceptions; `brwC` seems too strongly clustered around 0 to be meaningful here. Lets take a closer look at the distribution of the data with another histogram and box plot.

```
ggplot(bdataC, aes(brwC)) + geom_boxplot()
```




```
ggplot(bdataC, aes(brwC)) + geom_histogram(bins = 200)
```



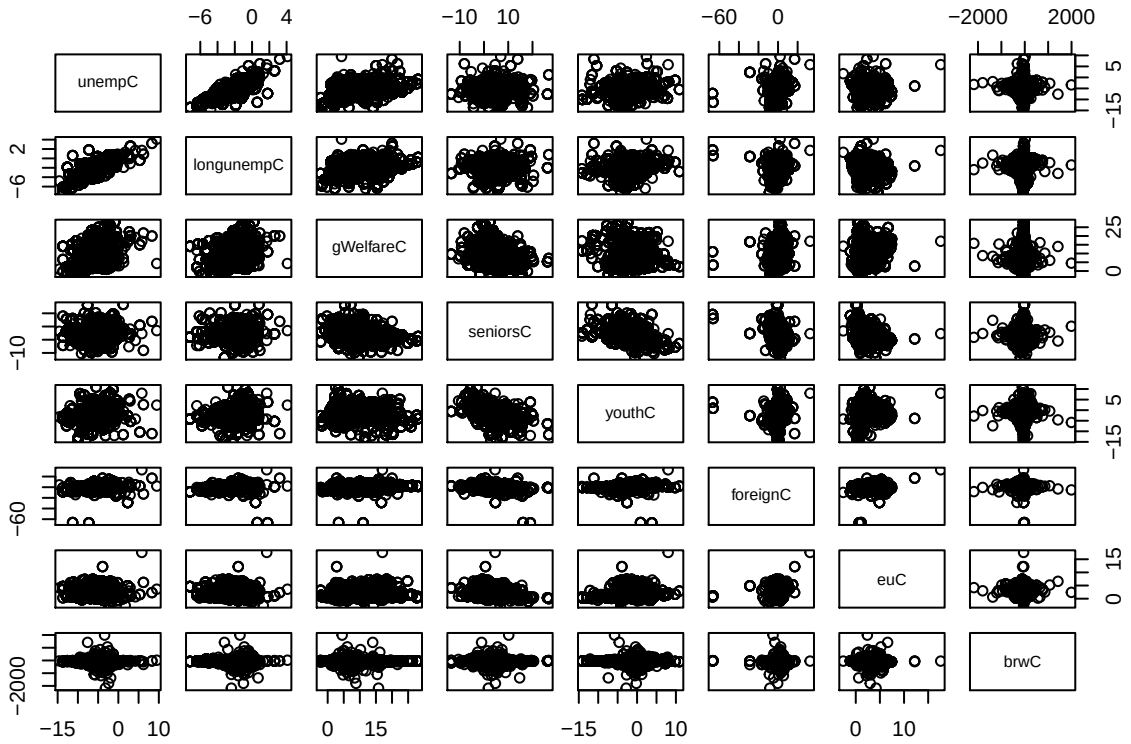
Both plots suggest that a substantial portion of the data is distributed around the mean except for a relatively small (but not necessarily trivial) set of values which lie substantially outside the mean. This could suggest that the data itself or my method of processing is flawed, or that the datasets I selected to measure real estate value change won't sufficiently capture value change over the eight years we're looking at. We'll take a closer look when we examine linearity in the next section

2.2.2 Linearity

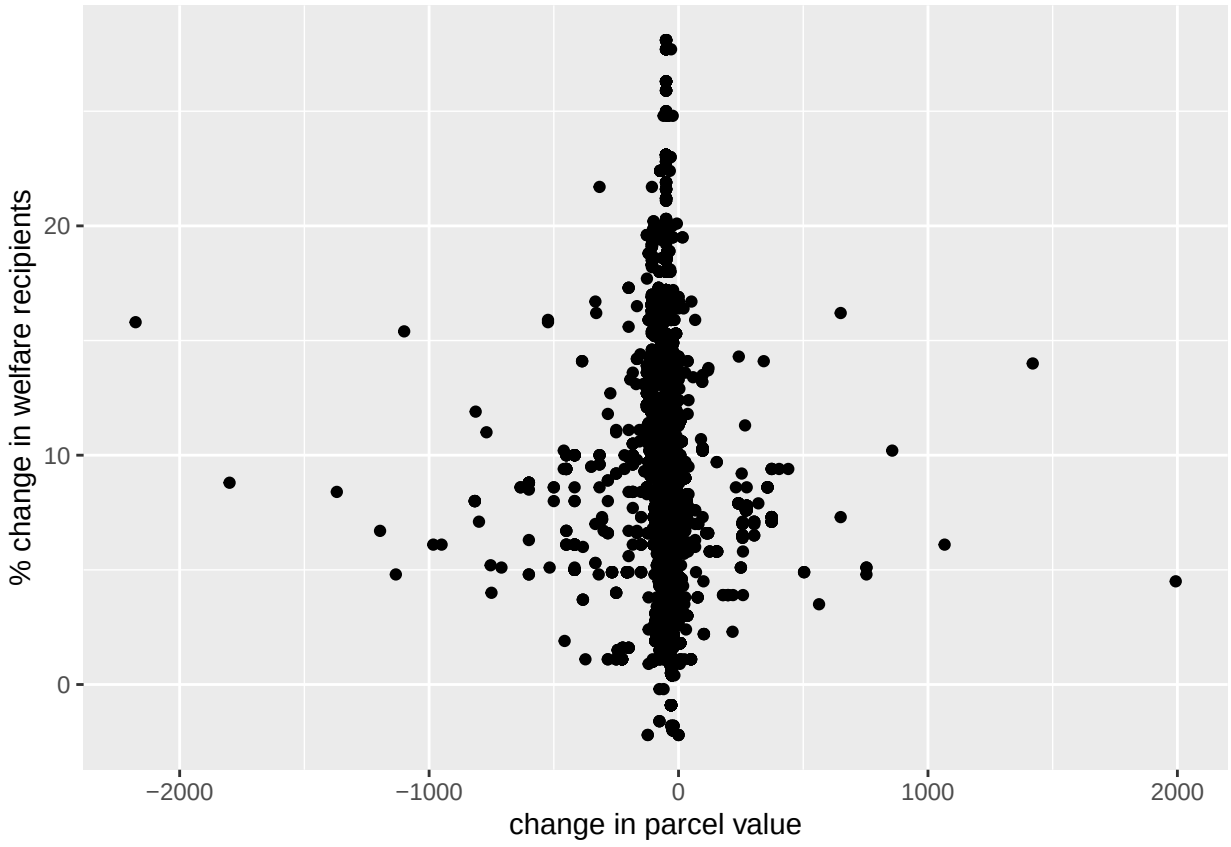
To assess the assumption of linearity, we plot all the indicator variables against each of the predictor variables.

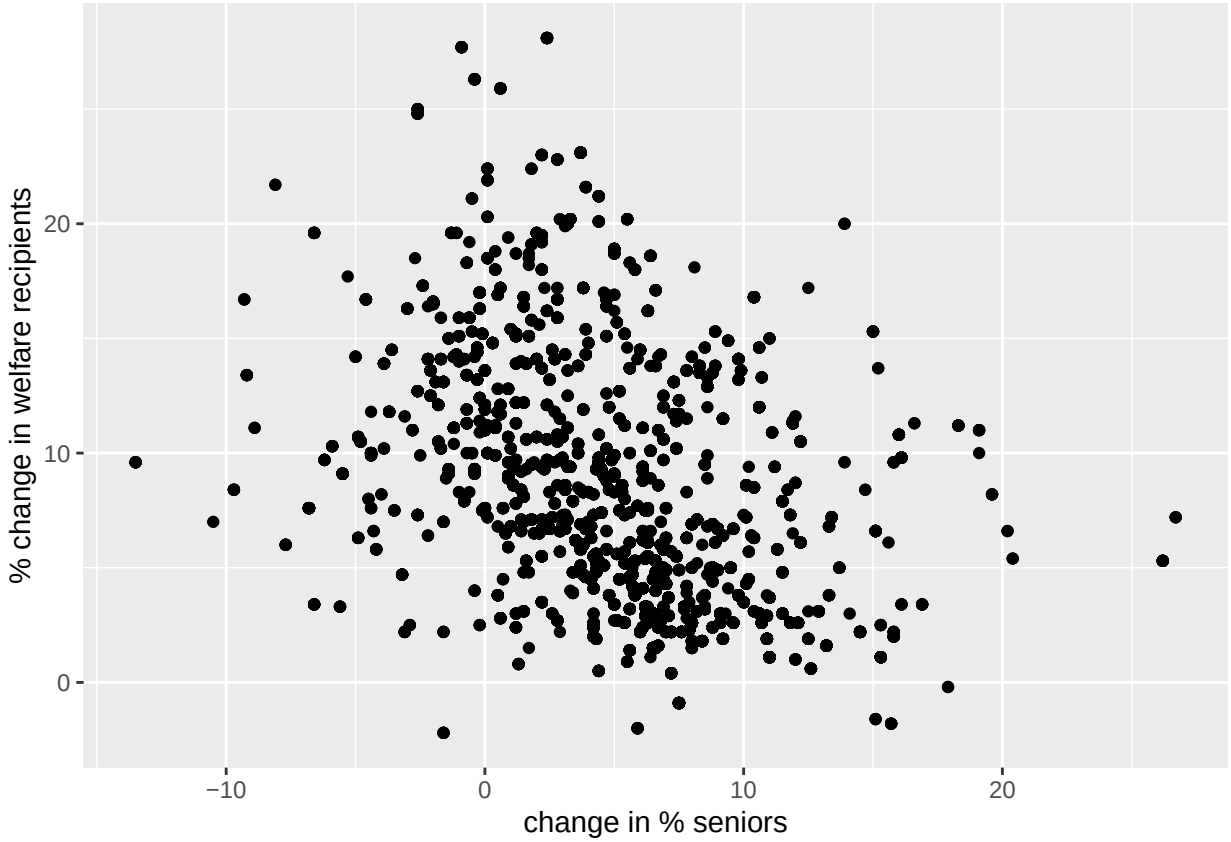
This code creates a scatterplot for each of the variables against each other:

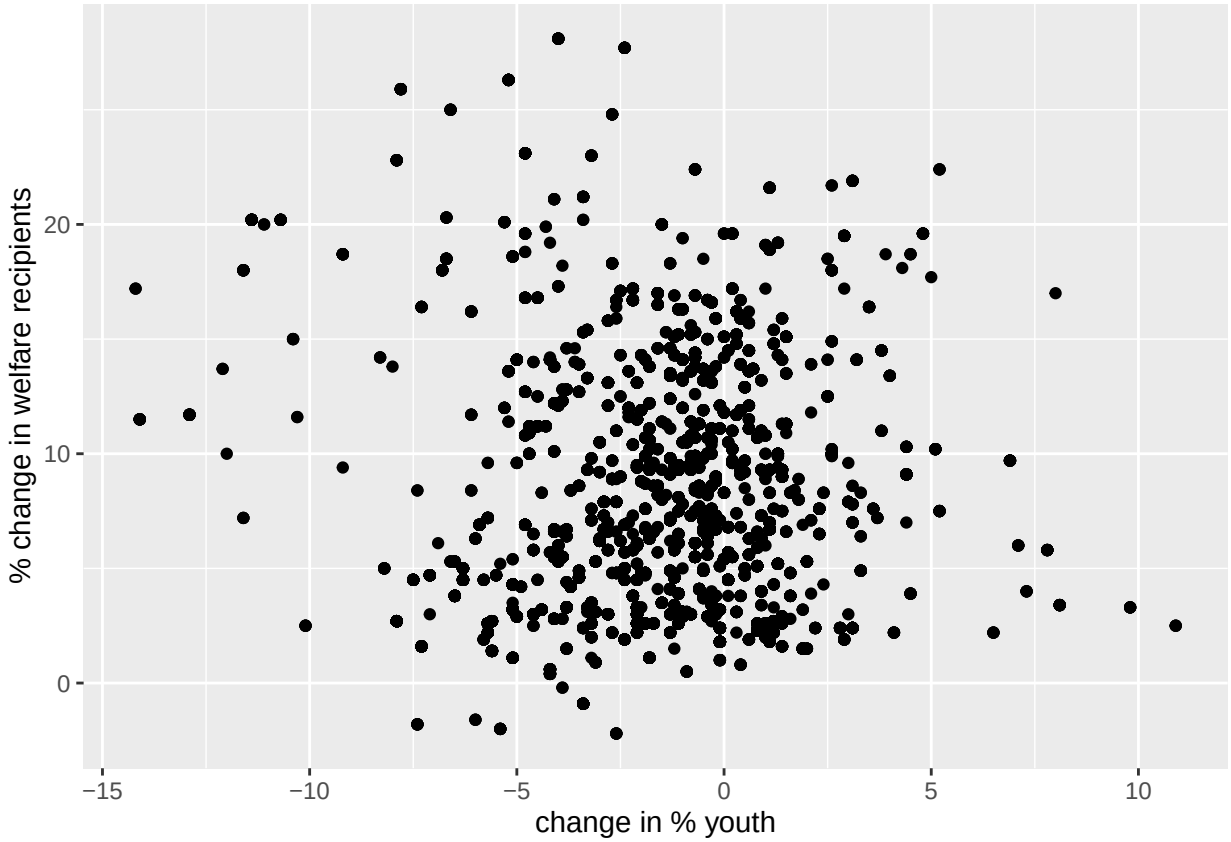
```
pwiseplot <- pairs(~unempC + longunempC + gWelfareC + seniorsC + youthC + foreignC + euC + brwC, data =
```

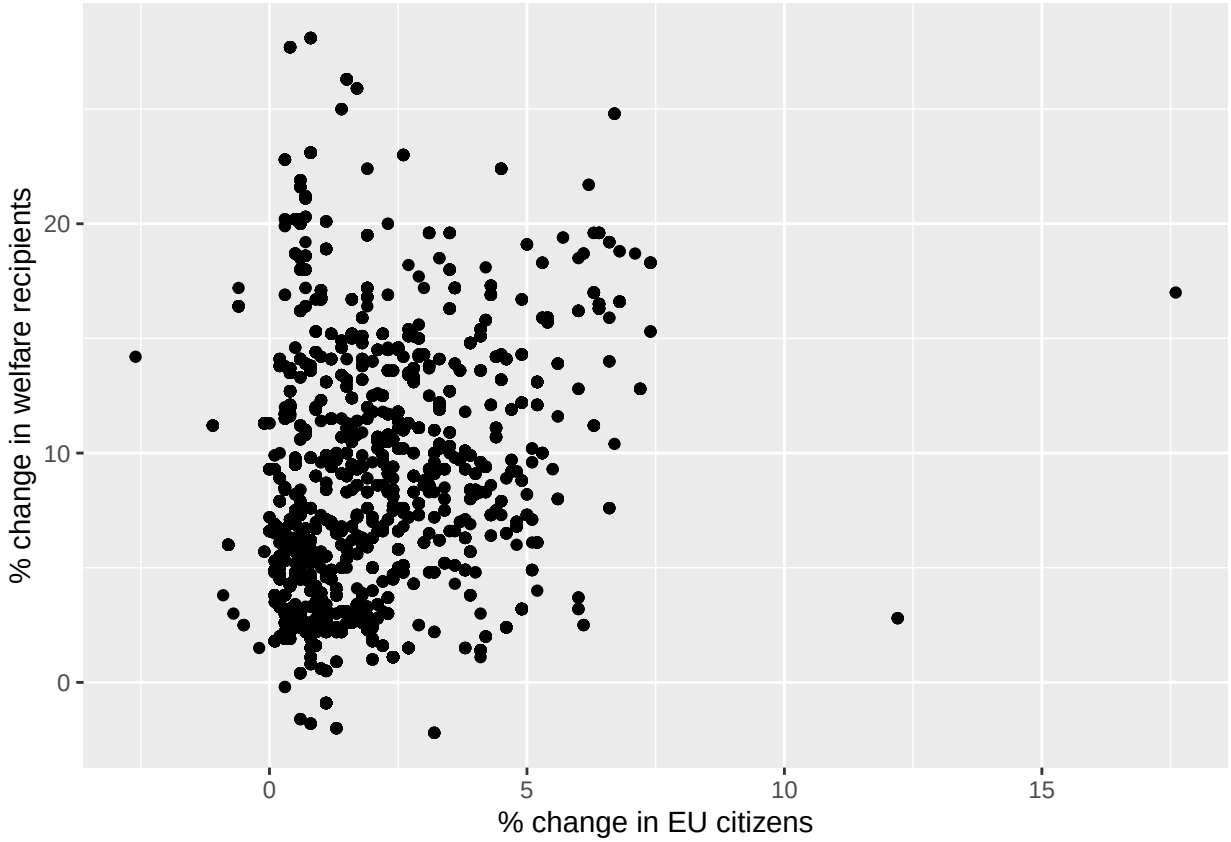


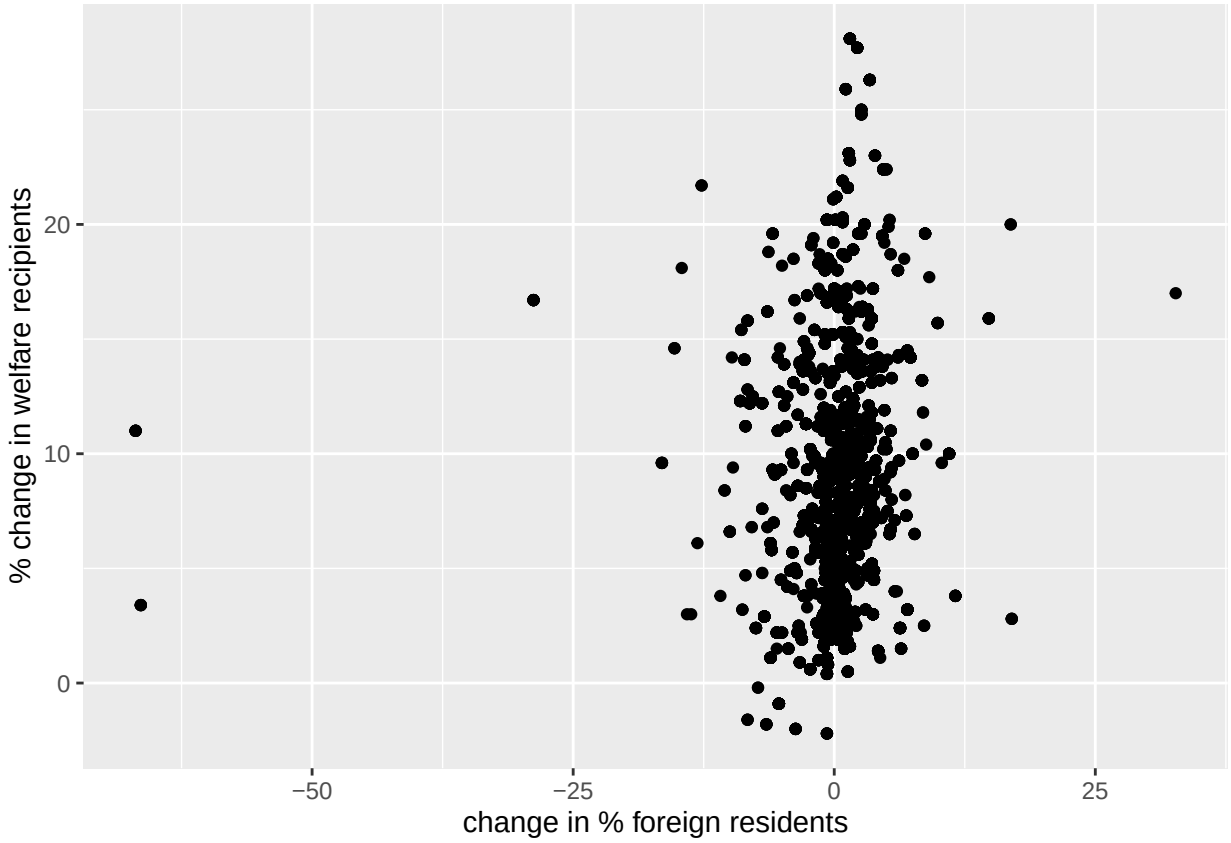
We can look more closely at the relationship between poverty and the predictor variables by creating scatter plots comparing each of the seven predictor variables:

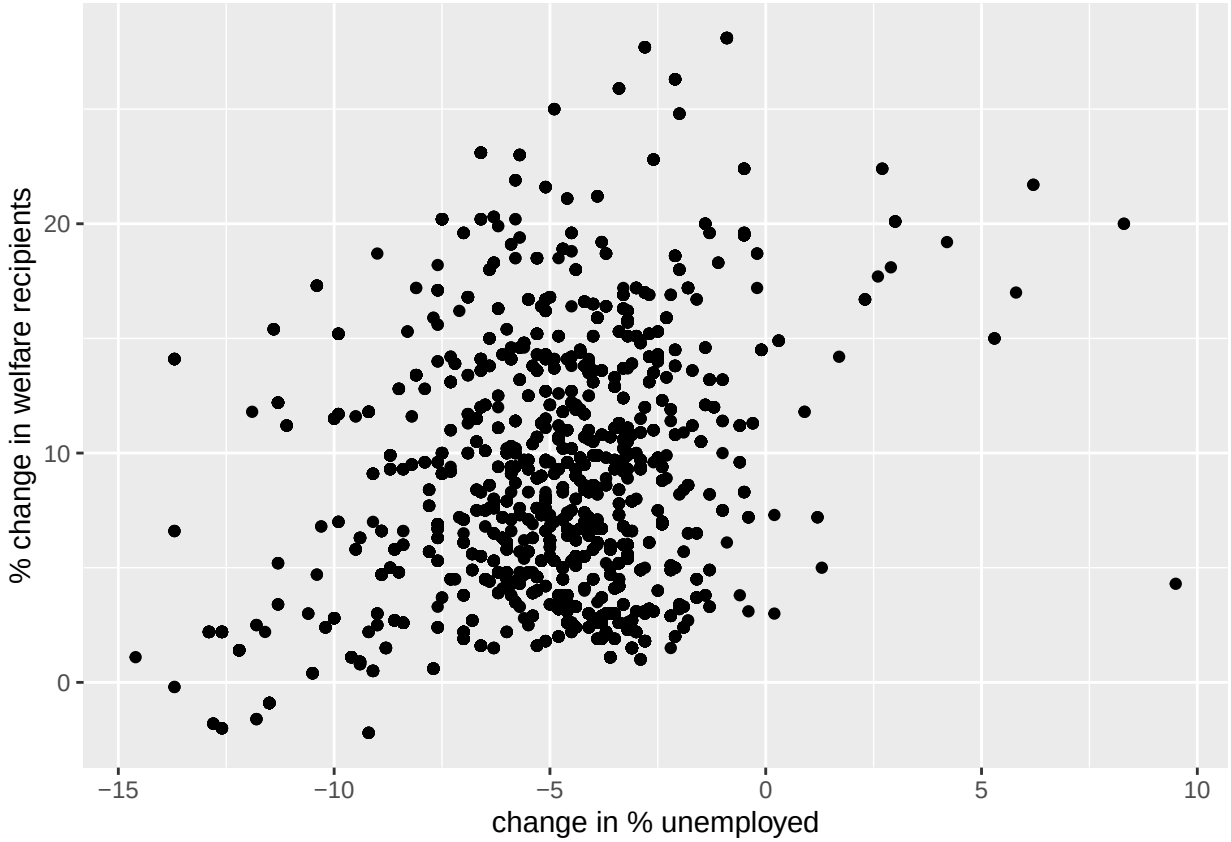


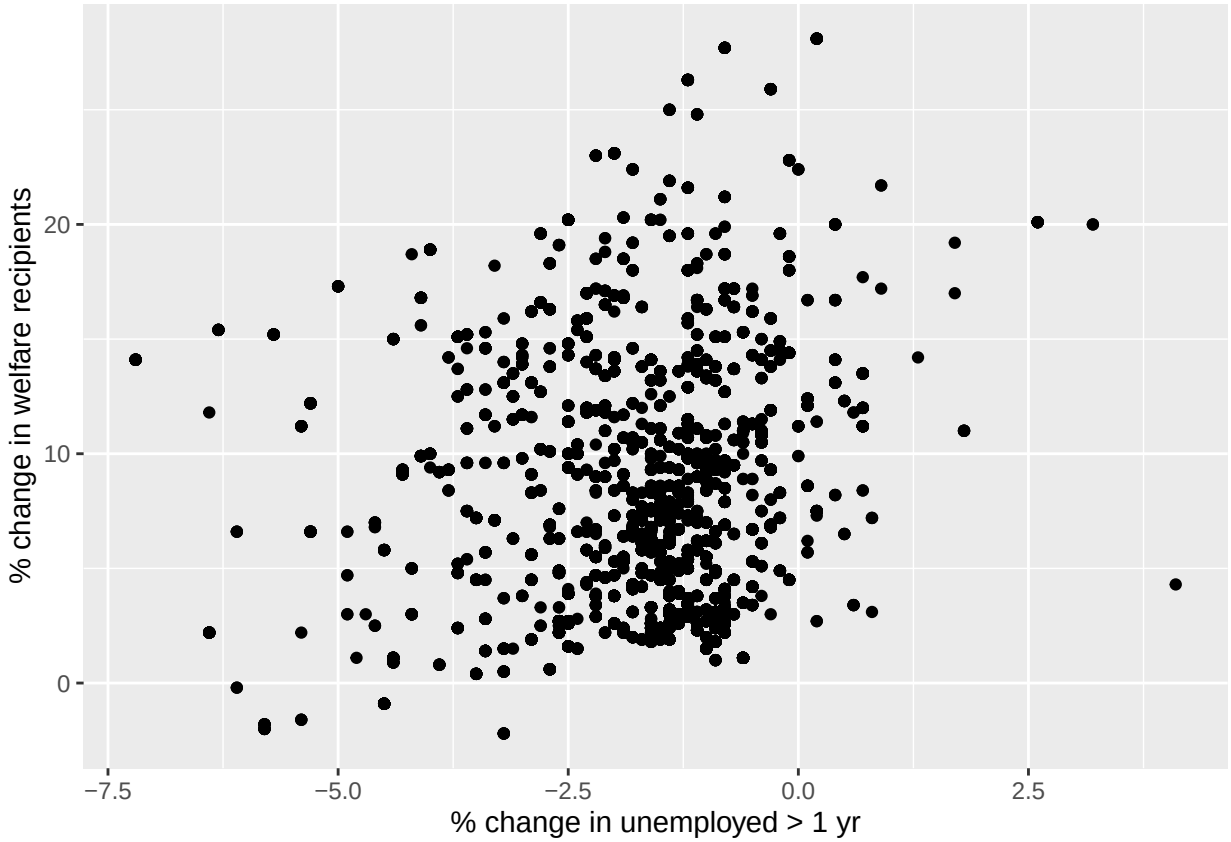

% change in welfare recipients
20
10
0
−10
0
10
20
change in % seniors


% change in welfare recipients
20
10
0
−15
−10
−5
0
5
10
change in % youth


% change in welfare recipients
20
10
0
0
5
10
15
% change in EU citizens

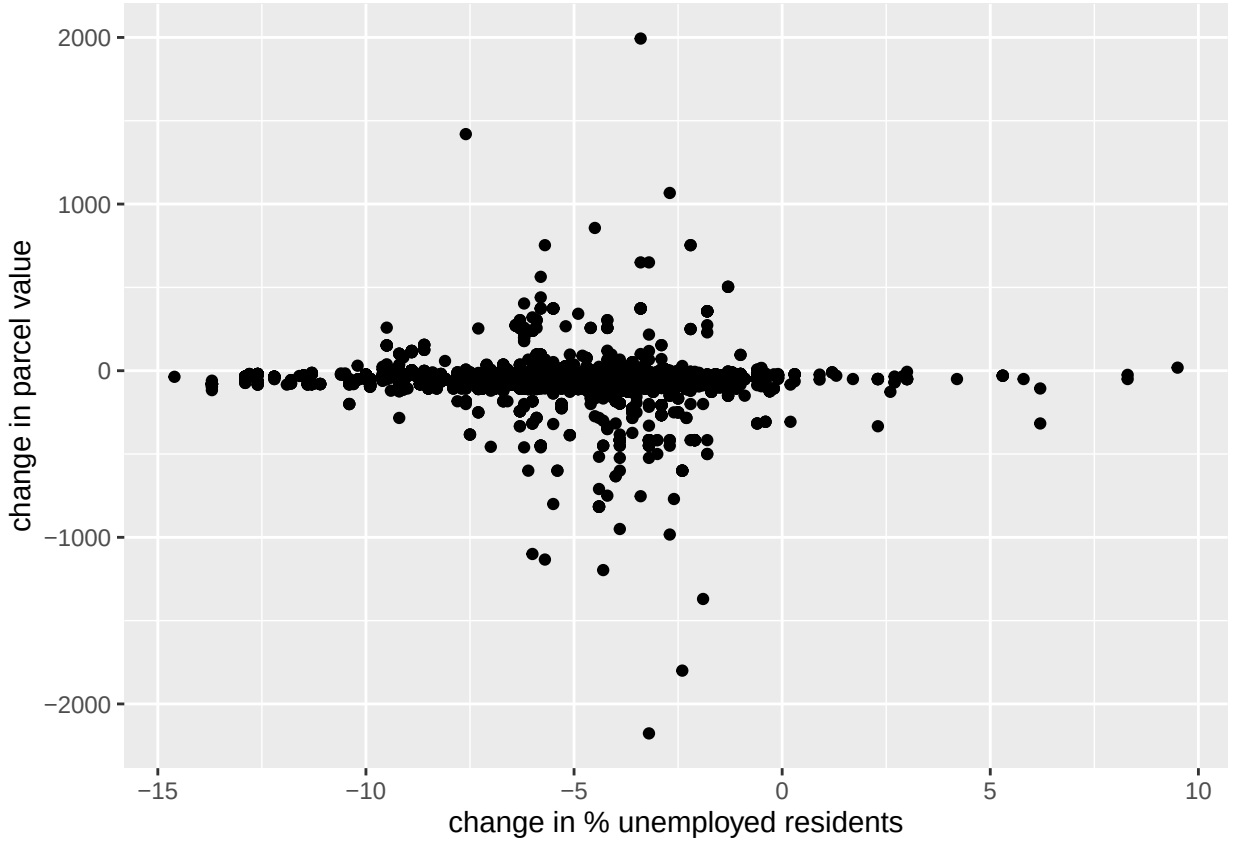

% change in welfare recipients
20
10
0
−50
−25
0
25
change in % foreign residents

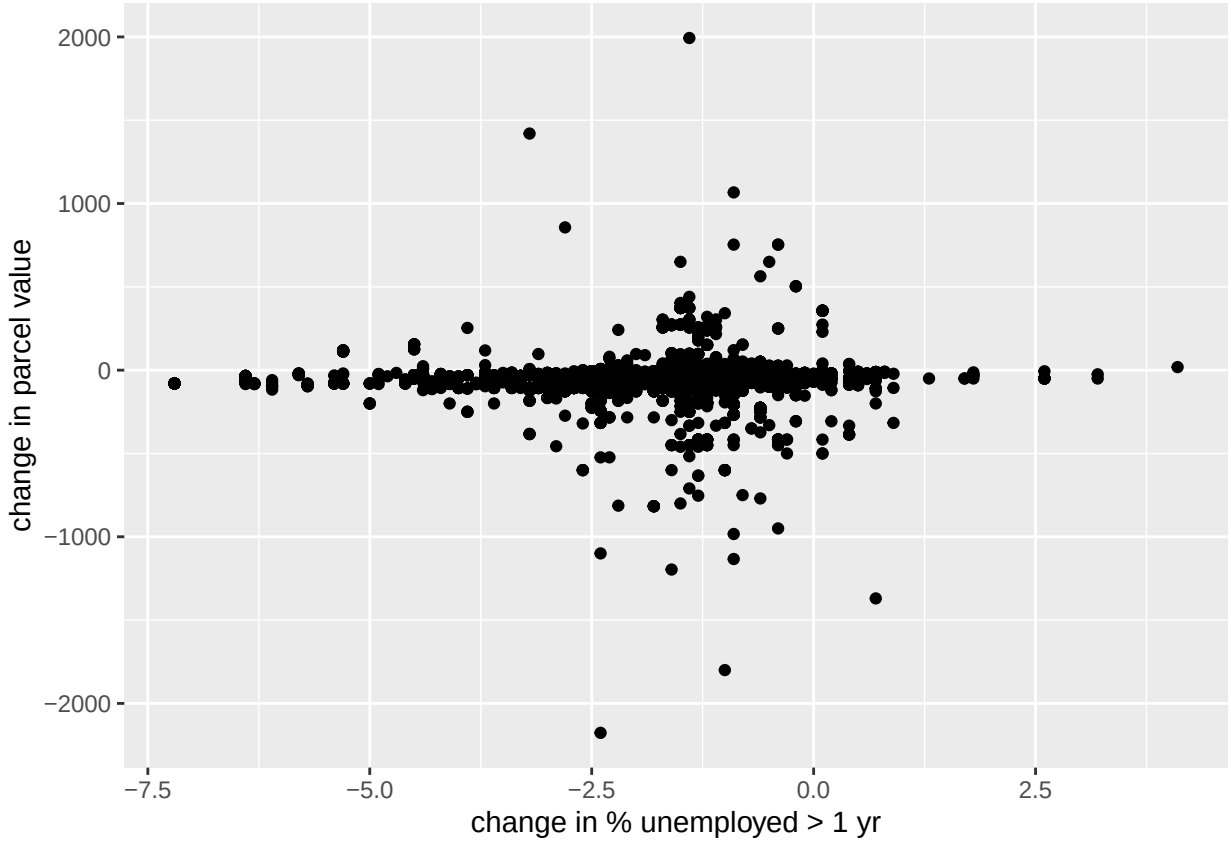

% change in welfare recipients
20
10
0
−15
−10
−5
0
5
10
change in % unemployed

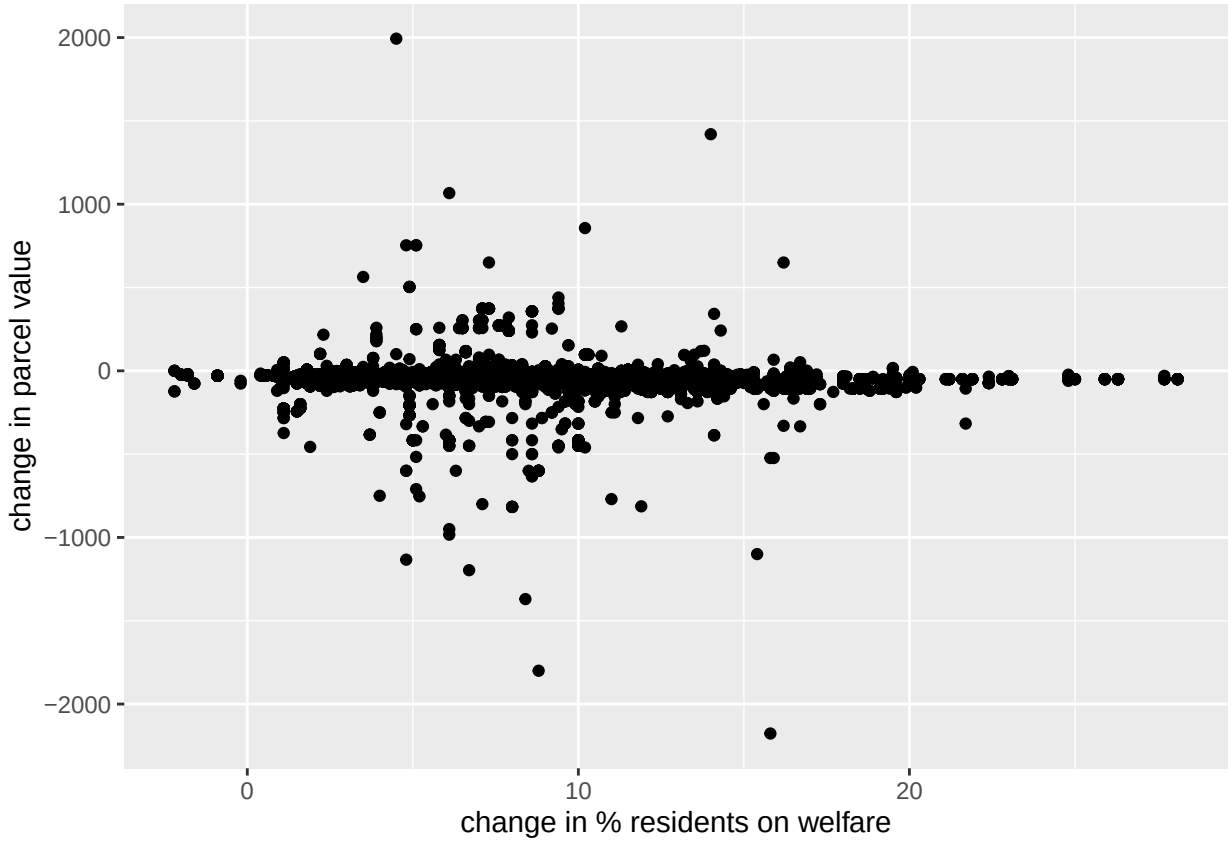

% change in welfare recipients
20
10
0
−7.5
−5.0
−2.5
0.0
2.5
% change in unemployed > 1 yr

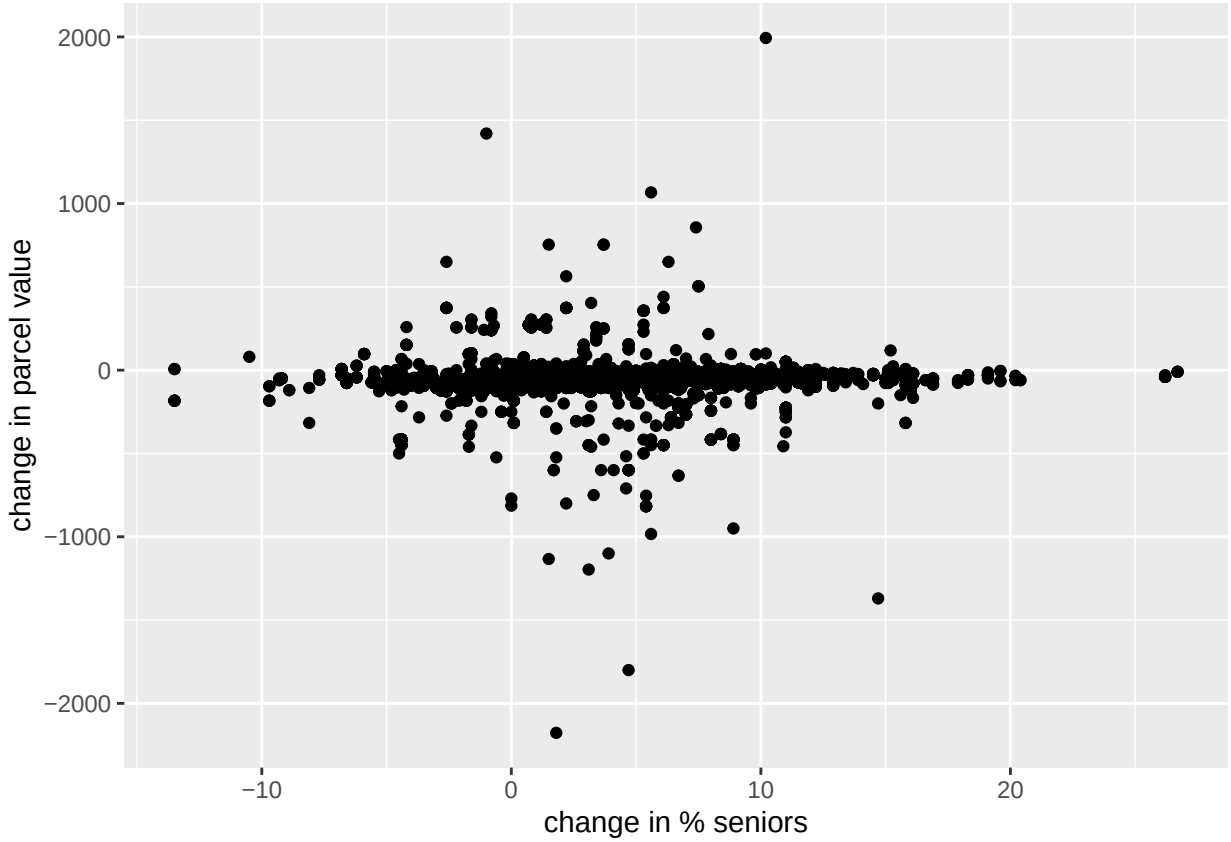
Overall, while none of the relationships stand out as particularly linearly related, `brwC`, `euC`, and `foreignC` seem particularly non-linear — I think this narrows down the model to four predictor variables: seniors, youth, and both unemployment variables.

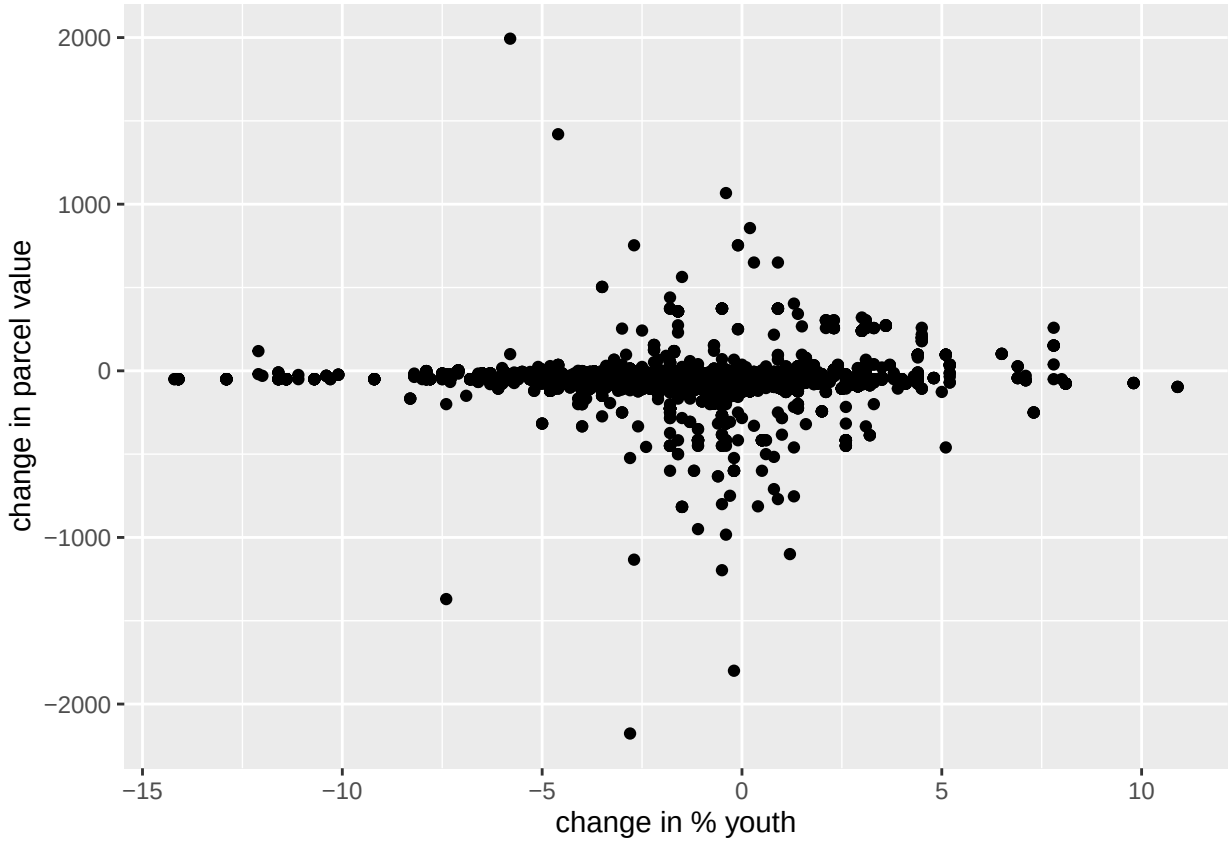
I'm also somewhat interested in assessing whether any of the demographic variables can predict price change in the parcels — similarly, before we perform any regressions or obtain any covariates, we need to examine our assumption that these variables are in fact linearly related. A scatterplot of real estate as the dependent variable compared to each demographic variable is shown below:

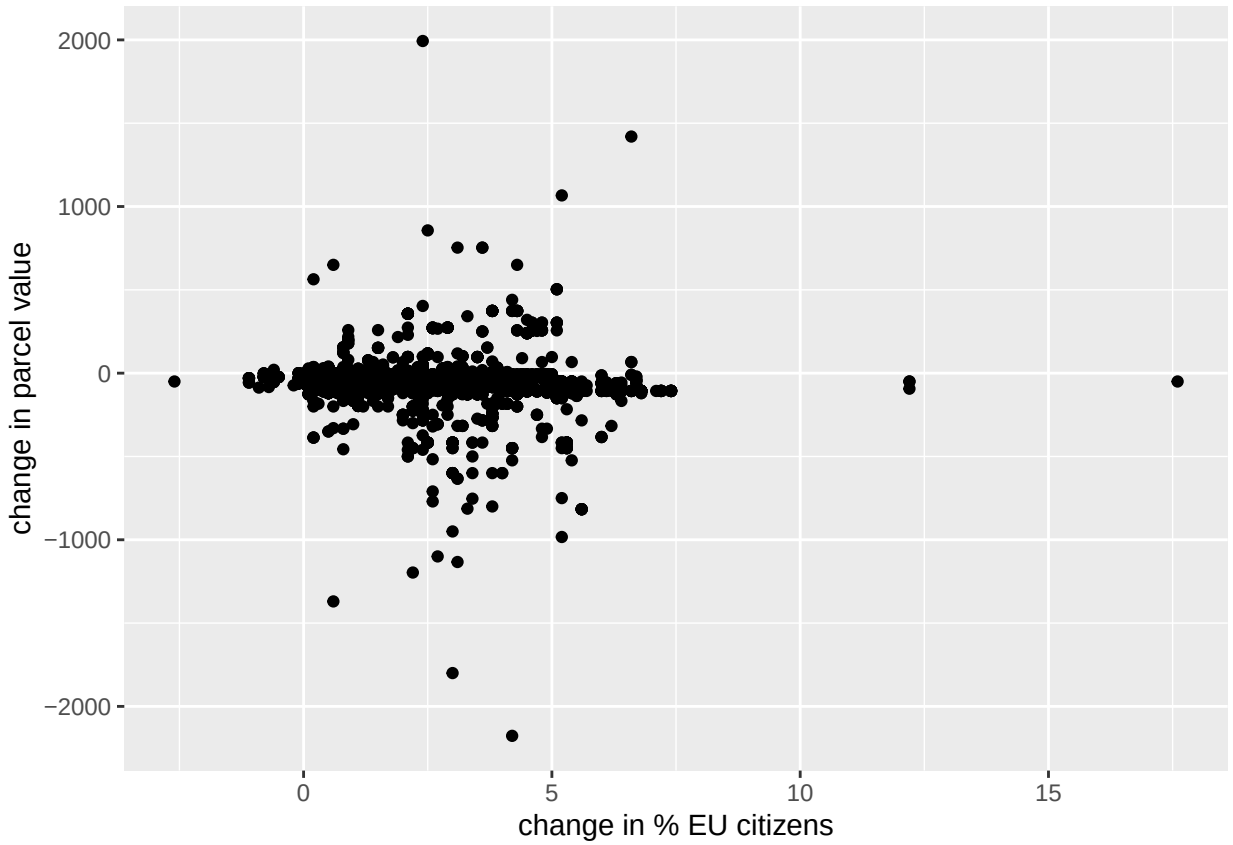



2000
1000
0
−1000
−2000
change in parcel value
−7.5
−5.0
−2.5
0.0
2.5
change in % unemployed > 1 yr


2000
1000
0
−1000
−2000
change in parcel value
0
10
20
change in % residents on welfare


2000
1000
0
−1000
−2000
change in parcel value
−10
0
10
20
change in % seniors


2000
1000
0
−1000
−2000
change in parcel value
−15
−10
−5
0
5
10
change in % youth



In my opinion, the total lack of linearity reflects that it is impossible to use model parcel change in any way, probably due to data quality issues in the underlying data or my processing. As such, I will not continue to analyze model parcel value until I can resolve the underlying data issues.

We will examine the assumption of homodasticity after modeling.

3 Modeling and Analysis

3.1 Model definition: predicting demographic change When attempting to measure displacement, one would look to see whether an area has lost a substantial amount of residents who are low-income, here represented by welfare recipients. In turn, one would look to which variables can best predict this displacement, such as valuation increases, increase in native/semi-native residents, and average age. This model represents the first method of integrating this exploration into the form of multiple regression: we ask how strong the association between the the four potential predictor variables are on any of the four indicator variables.

We represent this first model thus:

$$\text{gWelfareC} = \alpha + \beta_1(\text{seniorsC}) + \beta_2(\text{youthC}) + \beta_3(\text{unempC}) + \beta_4(\text{longunempC}) + \epsilon$$

In this model, β_1, β_2, β_4 and β_4 refer to their respective correlation coefficients. α represents the mean change in welfare of an area where there was no change in any of these variables, and ϵ refers to the random error associated with any one area.

As such, our null hypothesis is that β_1, β_2, β_4, β_4 are all equal to 0, meaning that they have no effect on change the change in number of people receiving welfare. We will see if we can invalidate this null hypothesis!

3.3 Model analysis Based on preliminary linearity analysis, we will evaluate the relationship between `gWelfareC` and the predictor variables.

```r
model <- lm(gWelfareC ~ seniorsC + youthC + unempC + longunempC, data = bdataC)
summary(model)
```

```
##
## Call:
## lm(formula = gWelfareC ~ seniorsC + youthC + unempC + longunempC,
##     data = bdataC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.717  -3.603  -1.346   2.568  18.235
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.29759    0.14057   94.60   <2e-16 ***
## seniorsC    -0.59896    0.01257  -47.63   <2e-16 ***
## youthC      -0.39047    0.02006  -19.47   <2e-16 ***
## unempC       0.86932    0.03517   24.72   <2e-16 ***
## longunempC  -0.70116    0.06989  -10.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.491 on 7763 degrees of freedom
## Multiple R-squared:  0.2613, Adjusted R-squared:  0.261
## F-statistic: 686.7 on 4 and 7763 DF,  p-value: < 2.2e-16
```

Based on this analysis, the p-value of each f-statistic is sufficiently small to reject the null hypothesis that there is no association between the predictors and the percent of people receiving welfare in an area. Based on the `Adjusted R-squared`, 26% of the variance of the dependent variable can be explained by the independent variables.

Additionally, we find out the alpha α, and each of the β values:

$$\widehat{\text{gWelfareC}} = 13.3 - 0.6(\text{seniorsC}) - 0.39(\text{youthC}) + 0.87(\text{unempC}) - 0.7(\text{longunempC})$$

We can examine specifically how significant each coefficient is in a coefficient table:

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 13.2975920 0.14057030  94.59745  0.000000e+00
## seniorsC    -0.5989637 0.01257490 -47.63170  0.000000e+00
## youthC      -0.3904696 0.02005884 -19.46621  1.917398e-82
## unempC       0.8693249 0.03516869  24.71872 6.415223e-130
## longunempC  -0.7011616 0.06989092 -10.03223  1.529880e-23
```

The most substantial effects are for seniors, with a t-value of -47, and the smallest is for long term unemployment. The p-value for each of these is sufficiently small that we can reject the null hypothesis and say that this model can predict, with some accuracy, the decline in residents who receive welfare.

In this case, a decreased number of both seniors and youth alongside increased short and longer term unemployment seems to be associated with a decrease of residents who receive welfare.

However, it is important to notice that the standard error ϵ is very high at 4.5 — this reflects that there is also a substantial amount of error in our model. We can find out the error rate by dividing the residual standard error by the mean of `gWelfareC`:

```r
sigma(model)/mean(bdataC$gWelfareC)
```

```
## [1] 0.5703154
```

This 57% residual standard error is quite high and means that the average deviation of the actual data from our model is substantial. I will expand on this in the results, but it isn't a good sign for the strenght of this model.

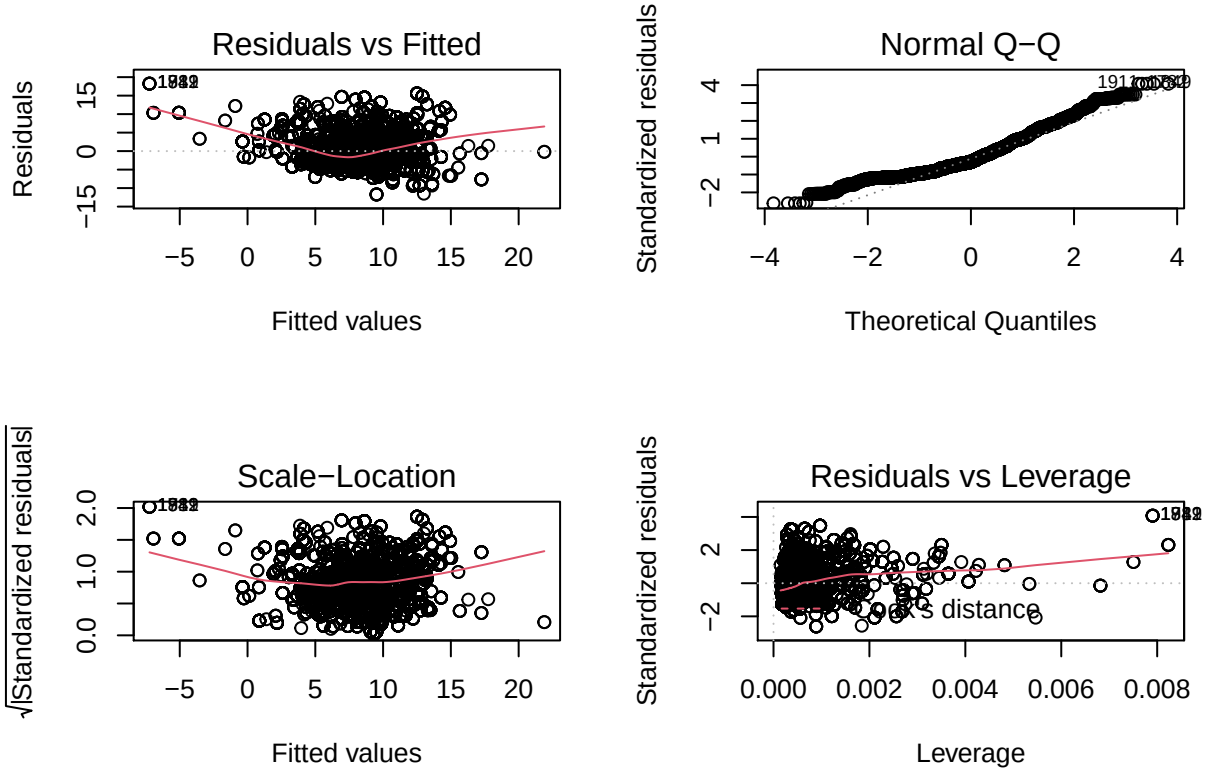
3.3 Post-regression assumption checks If we accept this model the association between displacement produced by a reduction in both youth and seniors and increased employment, we accept several assumptions along with it— in this section I check the assumptions that we haven't checked thus far (we looked at normality and linearity earlier briefly, but we'll return to them here too).

We assume that there aren't outliers producing the correlation coefficients in our model. While exactly defining an outlier is difficult, we can perform statistical tests to see how much of our data is far away from the median.

```
outlierTest(model)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 1782 4.080425         4.5404e-05       0.3527
```

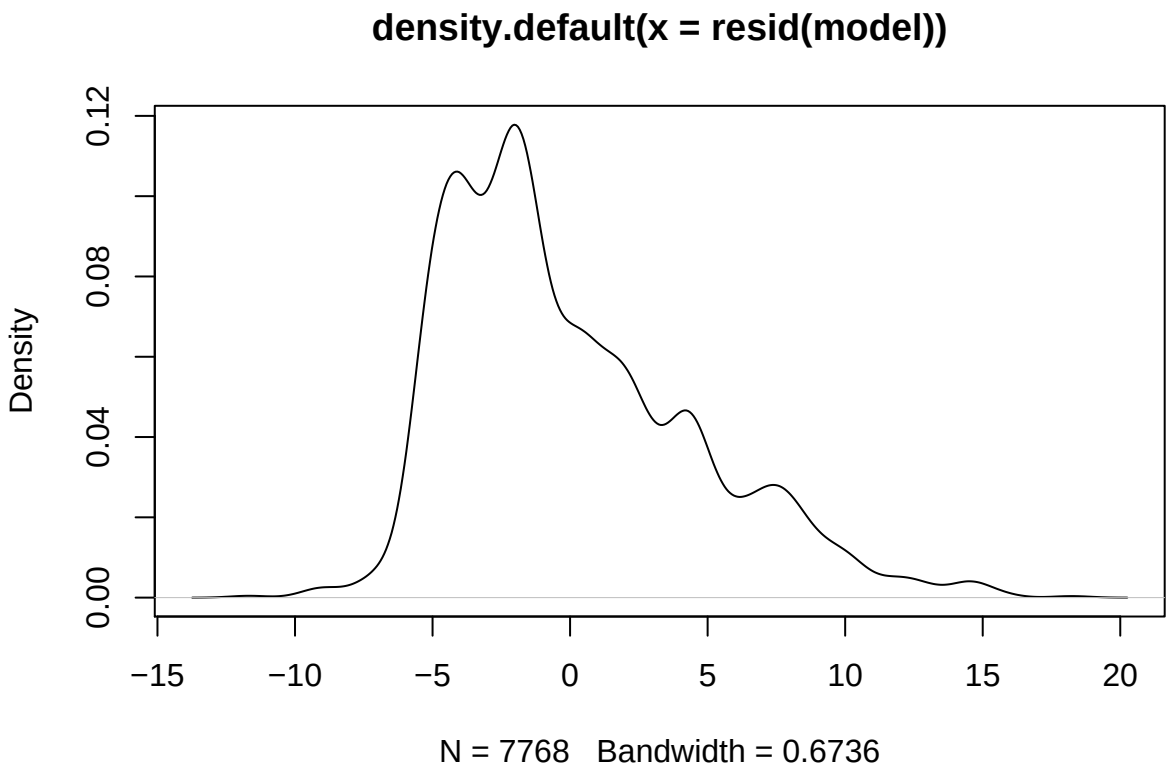
```
par(mfrow = c(2, 2))
plot(model)
```



This p-value of 0.35 indicates a very substantial chance that we might falsely reject the null hypothesis in this case — there seems to be a substantial chance that we would do so, as no residuals with a multiple-comparison test have a p-value of < 0.5.

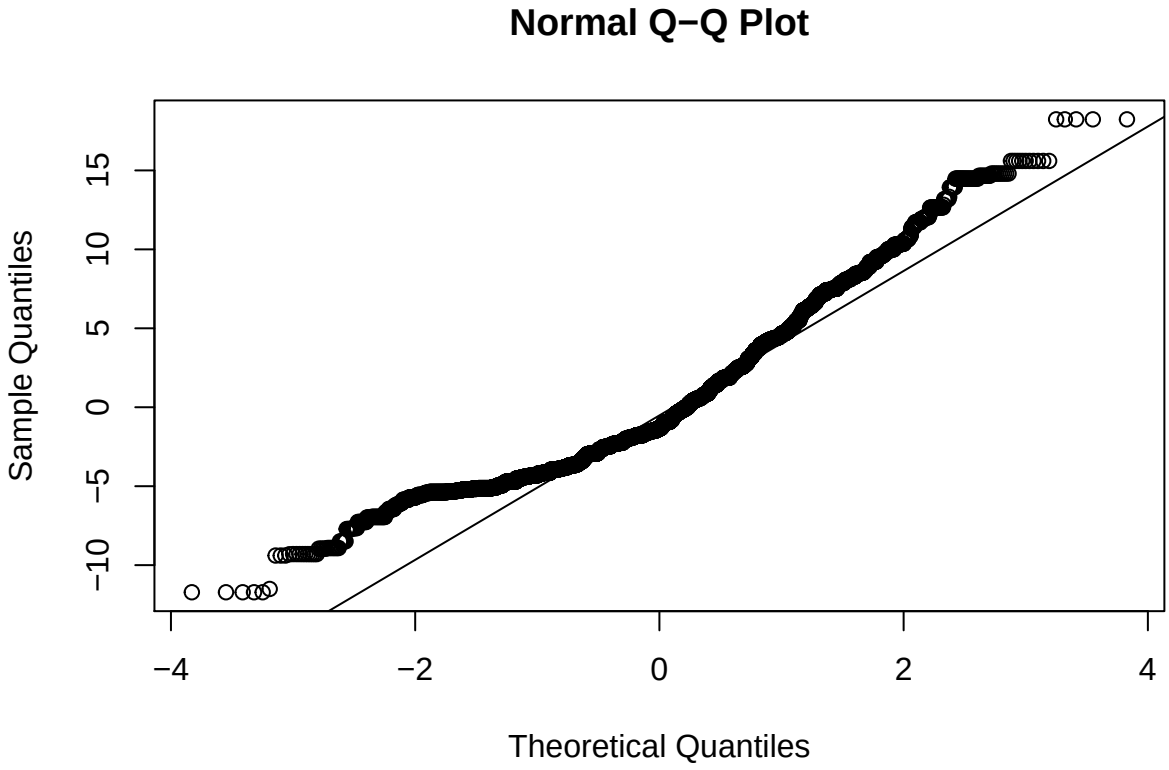
The plots also give us insight into the assumption of heterodasticity — the residuals are not distributed evenly across the x-axis, and the actual linearity of the residuals is small — this indicates that undermines the assumption of linearity which we explored, without statistical tests, earlier.

```
plot(density(resid(model)))
```



That the residuals are not centered on 0 and instead rightly skewed undermines the assumption of normality of residuals.

```
qqnorm(resid(model))
qqline(resid(model))
```



This Q-Q plot, which compares the data in a normal distribution to the distribution of my model's residuals, is curved slightly — this again reflects that we cannot assume the homodasticity of error.

Finally, we have assumed that the errors ϵ are independent and not being introduced by some other factor — because no time variables or spatial variables are present here, it is difficult to exactly check this, but understanding conceptually the model itself undermines this — for instance, lower unemployment being associated with reduced poverty could be due to people getting jobs, and thus no longer needing welfare, and older people might be generally more likely to be on welfare. I think that conceptually, while we can't verify this, we can safely reject this assumption.

4 Key findings and Limitations

I would say that based on my analysis, we cannot reject the null hypothesis that there is no association between reduction of poverty and changes in age, employment status, or foreign population. This analysis has not produced any significant results due either to the fact that there is no association between the predictor variables and the proportion of residents in an area receiving welfare, data quality issues arising from the original data or my methods of cleaning, or the influence of unknown variables affecting the number of people receiving welfare in an area. Moreover, I would say that a more robust methodolgy needs to be developed to analyze displacement than looking at the relationship between residents on welfare and any number of variables — the process is, in my view, too complex to be analyzed using multiple regression.

In returning to this assessment, it's a priority that more accurate data is found for real estate pricing, which is arguably the most substantial variable we hoped to analyze.