Heat Morality HW

Naomi Zubeldia and Sam Spackman

2023-04-13

```
library(tidyverse)
library(rgdal)
library(broom)
library(spdep)
library(nlme)
library(sf)
```

```
#reading in shape files
myShp <- st_read("./HoustonHeat.shp")
```

```
## Reading layer 'HoustonHeat' from data source
##   'C:\Users\naomi\OneDrive\Desktop\STAT469\HoustonHeat.shp' using driver 'ESRI Shapefile'
## Simple feature collection with 1487 features and 11 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -95.7332 ymin: 29.52989 xmax: -95.02184 ymax: 30.16357
## Geodetic CRS:  WGS 84
```

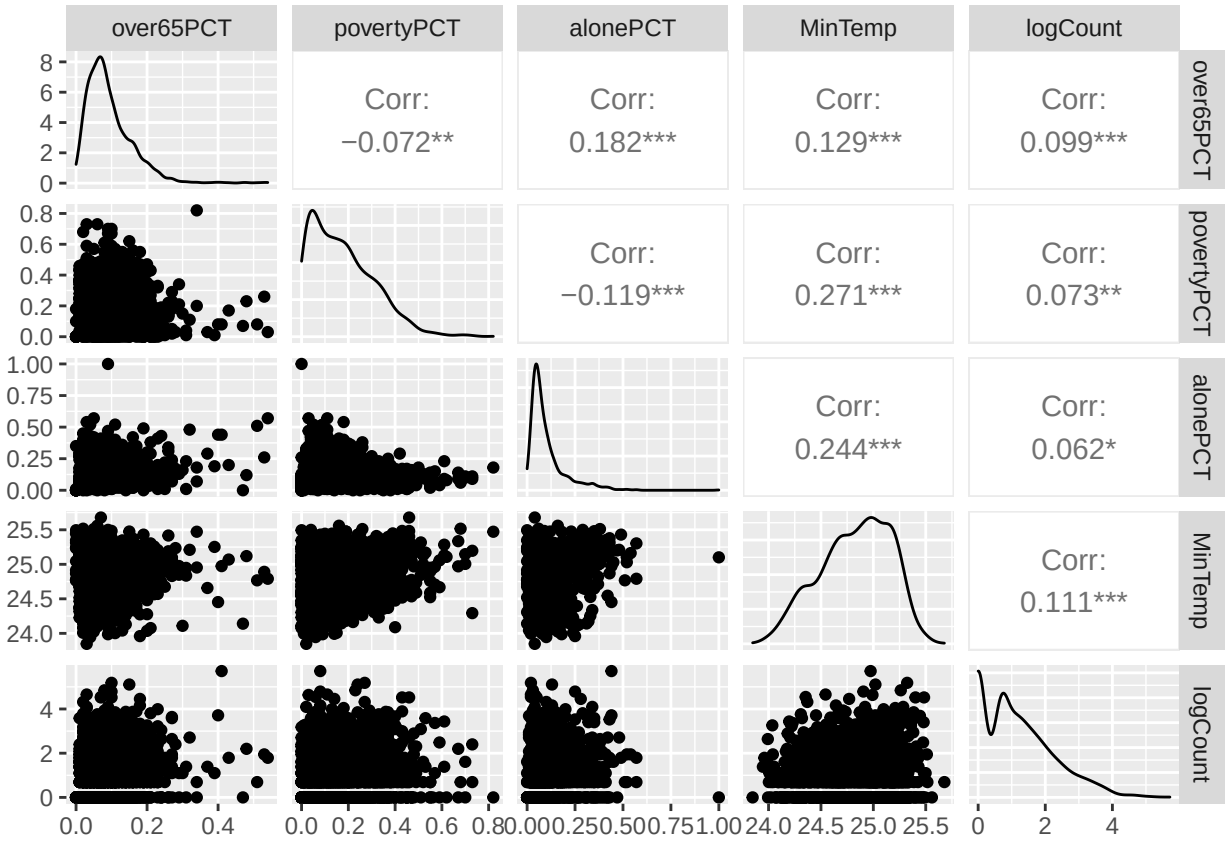
```
# #reformating the shapefile into a dataframe
# myShp@data$id <- rownames(myShp@data) #Assign ID to each polygon
# myShp.df <- tidy(myShp, region = "id") #Convert polygon info to data.frame()
# myShp.df <- merge(myShp.df, myShp@data, by = "id")

myShpDF <- data.frame(myShp) %>% dplyr::select(-geometry)
```

1.Exploratory Data Analysis

For our exploratory data analysis, we created a pairs plot to observe each of the variables interacting with the response variable of Log(Count+1). It looks like there is something going on with how the variables interact with Count since the graphs do not appear linear, which is something we will need to look into for our analysis.

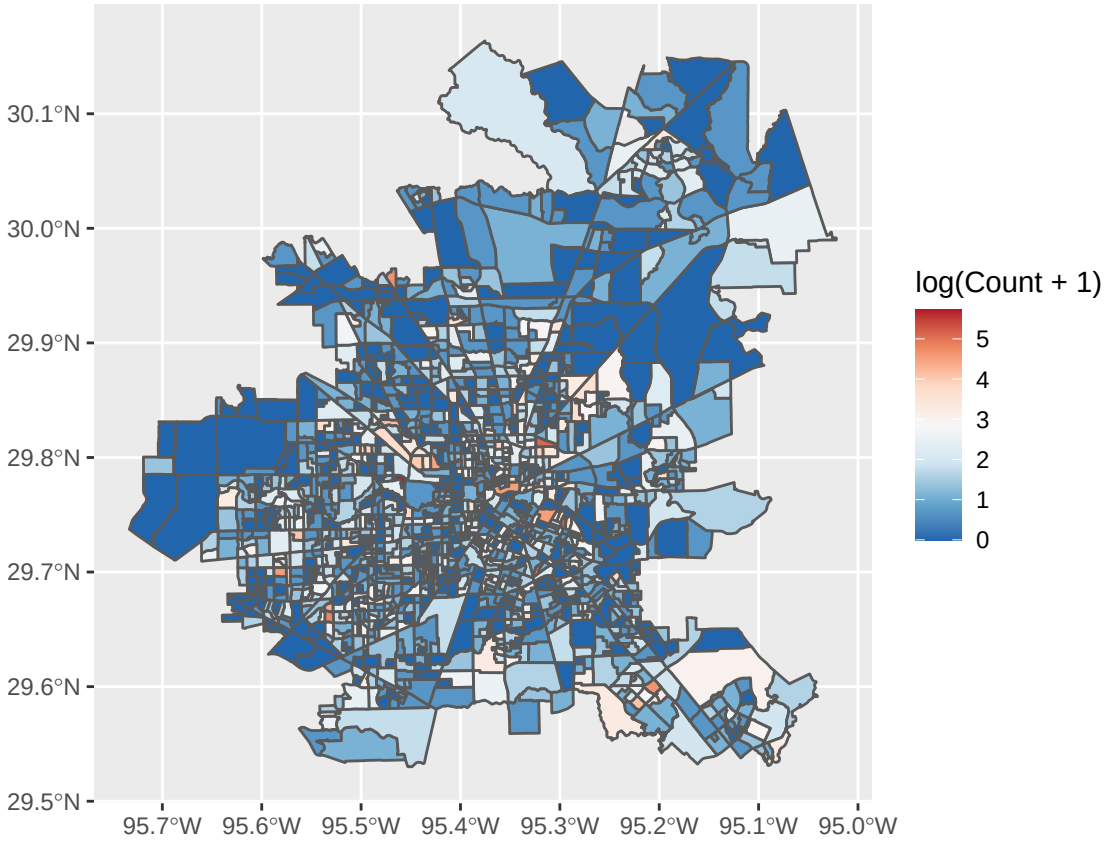
```
#observing the pairs plot of the data
myShpDF$logCount<-log(myShpDF$Count+1)
GGally::ggpairs(myShpDF[,8:12], progress=FALSE)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```



To observe if there is any correlation in the data since we have longitude and latitude measurements, we created the chloropeth map below.It appears that that there is some pattern in the map with there only being some spots of high morbility, which seems unusual since we found from a previous analysis that urban areas seem to be hotter and that would mean there would be more deaths than what this map shows.

```
#chloropeth map of count of data
ggplot(data=myShp) + geom_sf(mapping=aes(fill=log(Count+1))) + scale_fill_distiller(palette="RdBu")
```



2.Independent Linear Model

We created a basic linear model to see if it fits the data or if we need to look into a model that better fits for correlation.

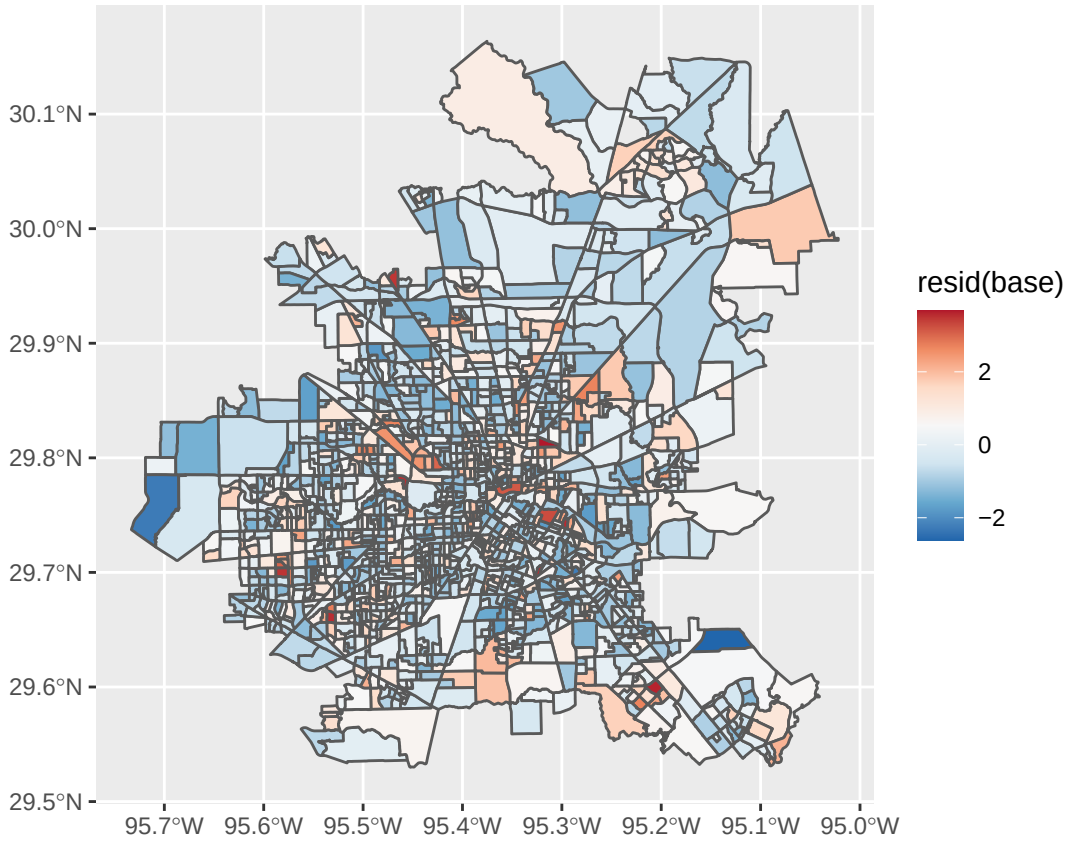
```
#creating a basic linear model
base<- lm(log(Count+1)~.-logCount,data=myShpDF)


#moran test of basic residuals
moran.test(x= resid(base), listw=nb2listw(poly2nb(st_make_valid(myShp))))
```

```
##
##  Moran I test under randomisation
##
## data:  resid(base)
## weights: nb2listw(poly2nb(st_make_valid(myShp)))
##
## Moran I statistic standard deviate = 15.736, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      0.2338486649     -0.0006729475      0.0002221256
```

To see if the model has any correlation, we conducted the moran test above on it. Because the pvalue is much smaller than the level of significance, we fail to reject the null hypothesis and find there to be correlation in the model.

We also mapped the residuals of the basic model in the cholorpeth map below. There seems to be weird patterns on it. Also, census areas next to each other have way different counts, which is unusal. There seems to be some areal correlation not being accounted for with this model.

```
#chloropeth map of the basic residuals
# data<-bind_cols(id=names(resid(base)),resid=resid(base))
# df<-left_join(myShp.df,data,by='id')
ggplot(data=myShp) + geom_sf(mapping=aes(fill=resid(base))) + scale_fill_distiller(palette="RdBu")
```



3. Areal Spatial Model

Because we found areal correlation in the basic model, we decided we need to use an areal spatial model that will account for the correlation. The model is explained below:

$$Y_i \overset{ind}{\sim} \text{Pois}(\lambda_i)$$

where

$$\log(\lambda_i) = \boldsymbol{x_i'\beta} + \boldsymbol{b_i'\theta}$$

$\boldsymbol{Y_i}$: A predicted count of number of heat-related morbidities for block group i.

λ_i: This is the mean count of the number of heat-related morbidities.

x_i': A vector containing all of the values for each of our explanatory variables as they relate to block group i. These are values such as the percent of homes without air conditioning, the median age of persons living in the block group, percent living alone in the block group, and others.

β: A vector of coefficients for each explanatory variable. In our case, these are the coefficients representing the intercept (the expected log mean count of number of heat-related morbidities if all of our explanatory values were 0) and the effect of the explanatory factors on our response (log mean count of number of heat-related morbidities). This is calculated using maximum likelihood estimation. These can be interpreted to mean that for every one unit increase in the value of x_j (median age, for instance), the log mean count of the number of heat-related morbidities increases by β_j, on average.

b_i': This vector represents the basis function values unique to block group i. In our case these basis functions are Moran basis functions based on eigenvectors of the adjacency matrix. Using only the eigenvectors from the adjacency matrix allows us to pull out only the important spatial effects from the data. Each observation (block group) will have a different value relating to each spatial basis function based on where it resides in space and this vector of spatial basis values will be multiplied against the θ vector to give us the full expected effect of spatial correlation related to that block group, similar to how the x matrix and β vector are multiplied against each other for predictions in normal linear regression.

θ: This represents a vector of the effects of the basis functions on our log response. This is calculated using maximum likelihood estimation.

4.

Below we will fit our new areal model we defined above.

```
#creating adjacency matrix
sf::sf_use_s2(FALSE) # Need to include this in latest R version
```

```
## Spherical geometry (s2) switched off
```

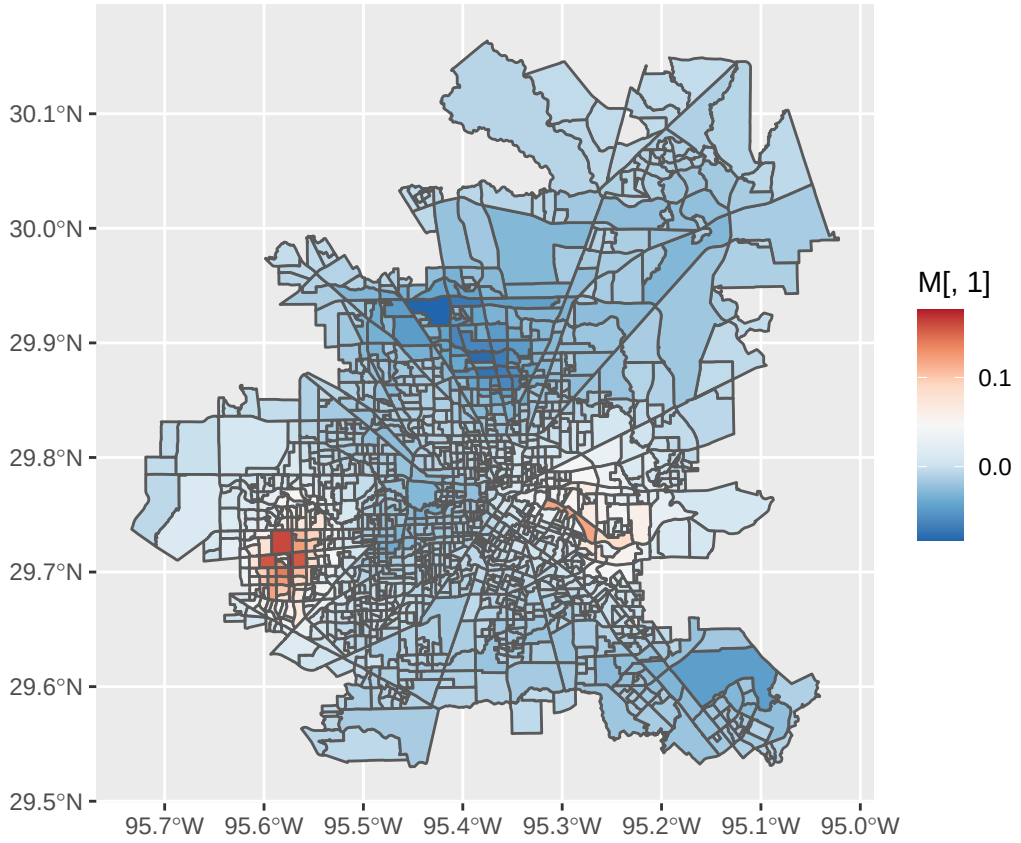
```
A <- nb2mat(poly2nb(myShp), style="B")
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they
## are planar
```

```
#creating moran spatial basis and chloropeth map
source("moranBasis.R")
```

```
## Warning: package 'RSpectra' was built under R version 4.2.3
```

```
# df2<- myShp.df[,-c(1:8)]%>% unique()
X<- model.matrix(base)
M <- moranBasis(X, A, tol=0.95)

#myShp.df$basis<-M[,1][match(myShp.df$id,myShp@data$id)]

ggplot(data=myShp) + geom_sf(mapping=aes(fill=M[,1])) + scale_fill_distiller(palette="RdBu")
```



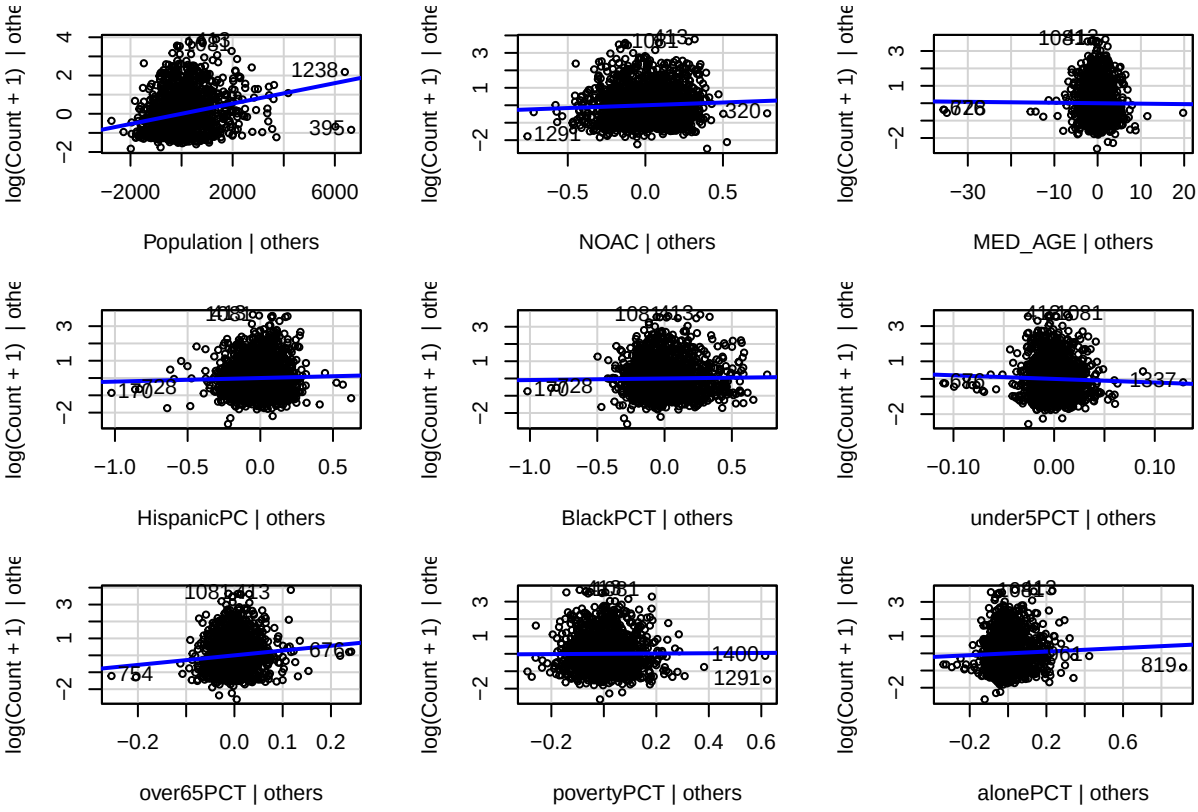

```
#merge bases on data set
myShpDF <- bind_cols(myShpDF,bases=M)
```

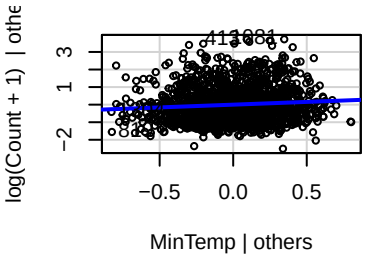
```
#fitting spatial glm model
myGLM <- glm(formula=Count~.-logCount, data=myShpDF, family=poisson)
```

Now that we have our newly-fitted areal model, we need to validate that it meets all of our necessary assumptions.

Below are added-variable plots to test whether our model meets the linearity assumption. No serious non-linear relationships are evident in these plots, so our model passes the linearity assumption. In this case we can say our model is log-linear.

```
#looking at avplots of model for linearity
library(car)
avPlots(base, ask=FALSE)
```


log(Count + 1) | othe
Population | others
NOAC | others
MED_AGE | others
HispanicPC | others
BlackPCT | others
under5PCT | others
over65PCT | others
povertyPCT | others
alonePCT | others



To check the independence assumption we check whether any correlation is left after subtracting out the effects of the basis functions. This is what we do below, then run a couple of tests designed to check the residuals for leftover correlation. Both tests return p-values near 1, which means our model passes the independence assumption.

```
#Checking independence assumption using moran and geary test
resid<-resid(myGLM,"pearson")
moran.test(x=resid , listw=nb2listw(poly2nb(myShp)))
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they
## are planar

##
##  Moran I test under randomisation
##
## data:  resid
## weights: nb2listw(poly2nb(myShp))
##
## Moran I statistic standard deviate = -5.1823, p-value = 1
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -0.0773988370     -0.0006729475      0.0002191977
```

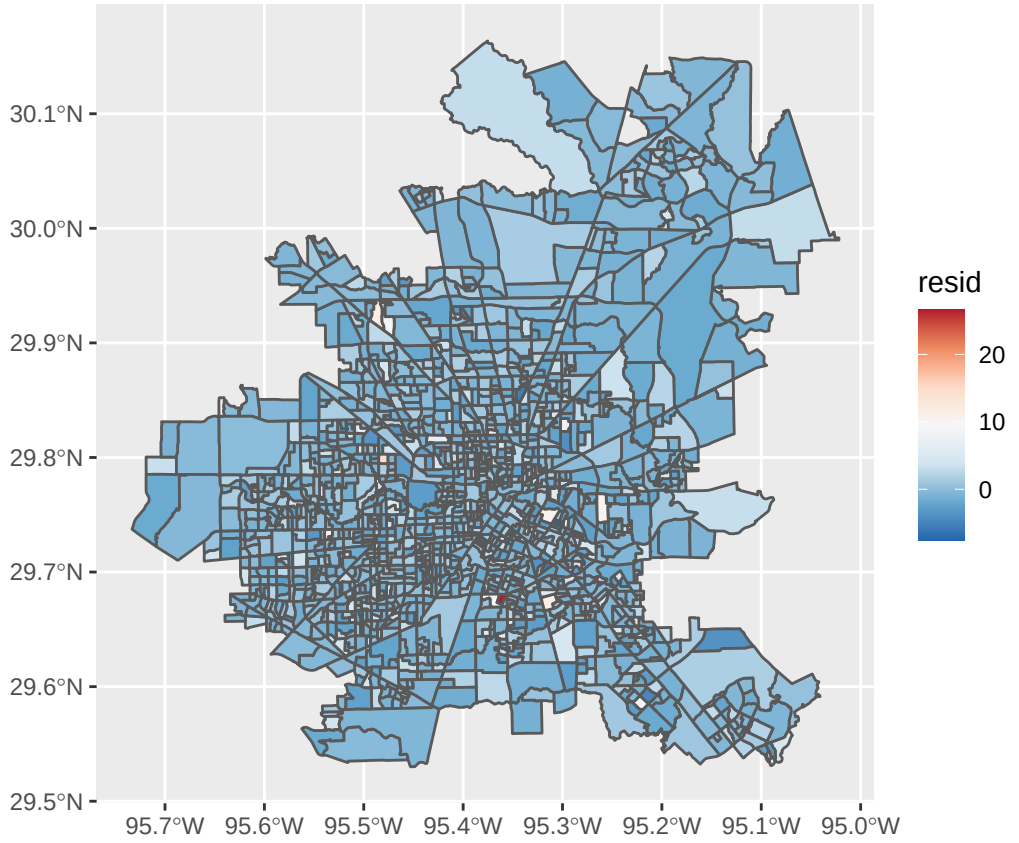
```
geary.test(x=resid , listw=nb2listw(poly2nb(myShp)))
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they
## are planar
```

```
##
##  Geary C test under randomisation
##
## data:  resid
## weights: nb2listw(poly2nb(myShp))
##
## Geary C statistic standard deviate = -3.2831, p-value = 0.9995
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##      1.0867596483      1.0000000000      0.0006983248
```

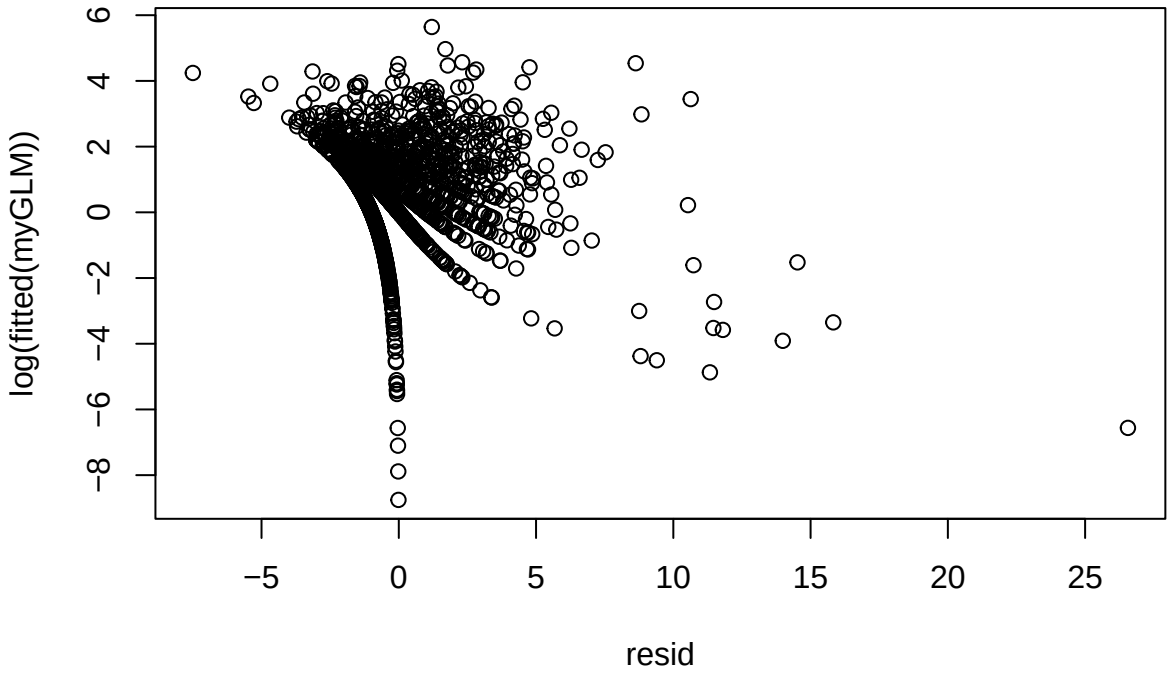
We can also plot another chloropleth map of the residuals to check if any areas still look like they are correlated with neighboring areas. In this case, the tests we ran above seem very correct; no area looks correlated with each other.

```
#chloropeth map of decorrelated residuals
# myShp.df$glmresid<- resid[match(myShp.df$id, myShp@data$id)]
ggplot(data=myShp) + geom_sf(mapping=aes(fill=resid)) + scale_fill_distiller(palette="RdBu")
```



To check for equal variance we can simply plot our standardized residuals against our fitted-values. This plot may look funny, but it doesn't show us any extra relationship that we have not yet already accounted for in the data so we can say that it passes the equal variance test.

```
#plot of fitted v residuals of glm model
plot(resid,log(fitted(myGLM)))
```



It is also worth noting that for our model we do not have to check any normality assumption because we are not assuming a normal relationship in our data. We are performing Poisson regression, instead.

5.

Below we have calculated confidence intervals for the effect of the explanatory variables on the count of number of heat-related morbidities. According to this table, block groups with a high percentage of Hispanics, people living alone, people over the age of 65, people living in poverty, and people with no air conditioning are the most at risk of higher counts of heat-related morbidities.

```
confint.default(myGLM)[2:11,]
```

```
##                    2.5 %        97.5 %
## Population  0.0003862391  0.0004530867
## NOAC        0.0371832952  0.4408534593
## MED_AGE    -0.0194028005  0.0071350667
## HispanicPC  0.3685940006  0.9375024855
## BlackPCT   -0.2282936852  0.1955099256
```

```
## under5PCT  -7.2285225025 -2.9043886141
## over65PCT   3.8809995138  6.0531089466
## povertyPCT  0.0673953490  0.8491026658
## alonePCT    0.2905381925  1.1384236106
## MinTemp     0.4766505657  0.8201270251
```

6.

Below is a graph that plots just the effects of the basis functions on the count of number of heat-related morbidities in a block group. This will show us which areas have the largest positive and negative effects on heat-related morbidities. In this case it looks like many inner-city areas are more at-risk to heat-related morbidities, while there are some pockets in the suburbs of the city that have strong effects away from heat-related morbidities.

```
#chloropeth map of the spatially correlated residuals
spatial <- M%*%matrix(coef(myGLM)[12:length(coef(myGLM))],ncol=1)
# myShp.df$allbasis<-spatial[match(myShp.df$id, myShp@data$id)]
ggplot(data=myShp) + geom_sf(mapping=aes(fill=spatial)) + scale_fill_distiller(palette="RdBu")
```