Drug Abuse by State

Jillian Warburton

2023-04-12

Reading in and Handling Shapefiles

1. Read in the drug abuse shapefile.

```
#load dataset
#myShp <- readOGR(dsn="~/R programming/STAT_469/Unit4",layer='DrugAbuse')
#library(broom) #contains tidy() function which converts polygons to data.frame
#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") #Merge data w/polygon data.frame
myShp <- st_read("~/R programming/STAT_469/Unit4/DrugAbuse.shp")
```

```
## Reading layer 'DrugAbuse' from data source
##   'C:\Users\jilli\OneDrive\Documents\R programming\STAT_469\Unit4\DrugAbuse.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 49 features and 6 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -124.7328 ymin: 24.95638 xmax: -66.96927 ymax: 49.37173
## Geodetic CRS:  NAD83
```

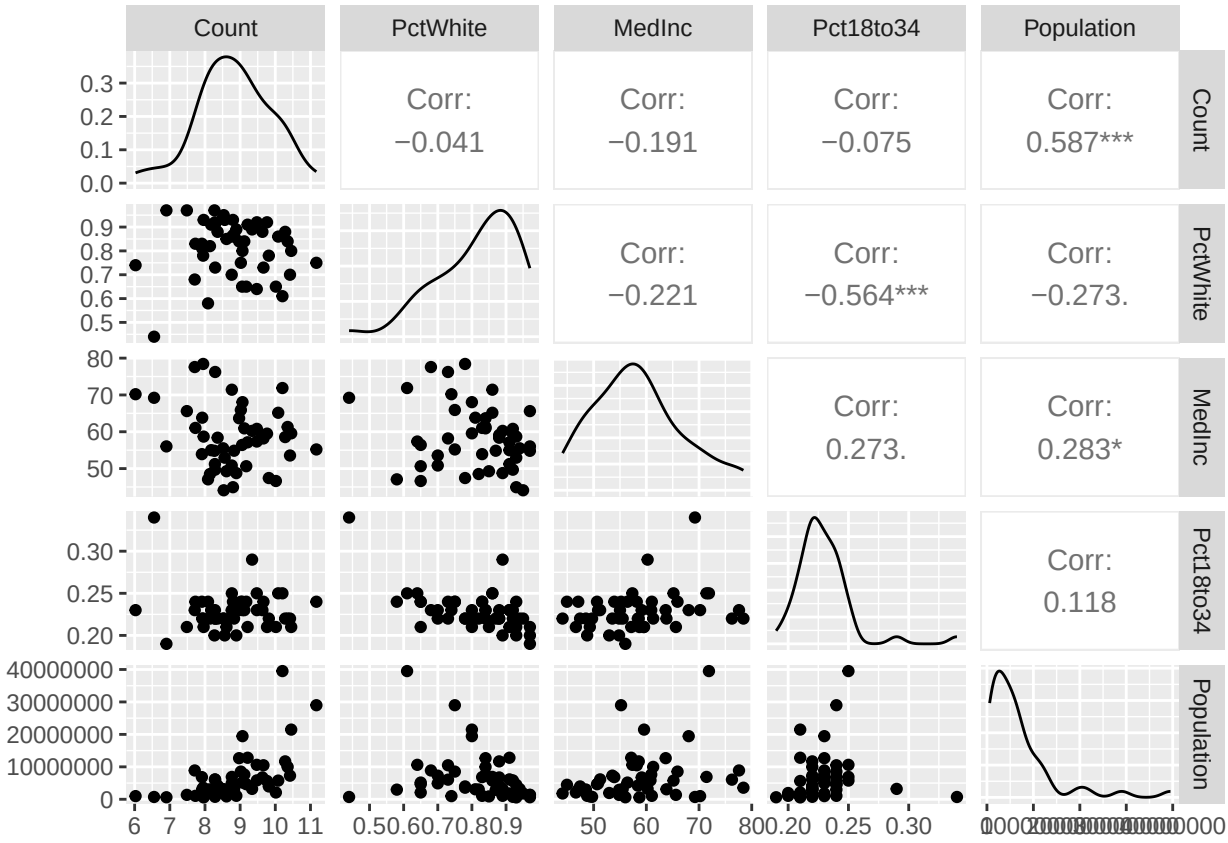
2. Reformat the shapefile data into a dataframe suitable for use with `ggplot()`.

```
data <- data.frame(myShp) %>% select(-geometry)
#data <- data.frame(myShp.df)
#data <- data %>% dplyr::select(-c(id, long, lat, order, hole, piece, group))
data_red <- data %>% dplyr::select(-c(State))
data_red <- data_red %>% mutate(Count = log(Count))
```

Exploratory Data Analysis

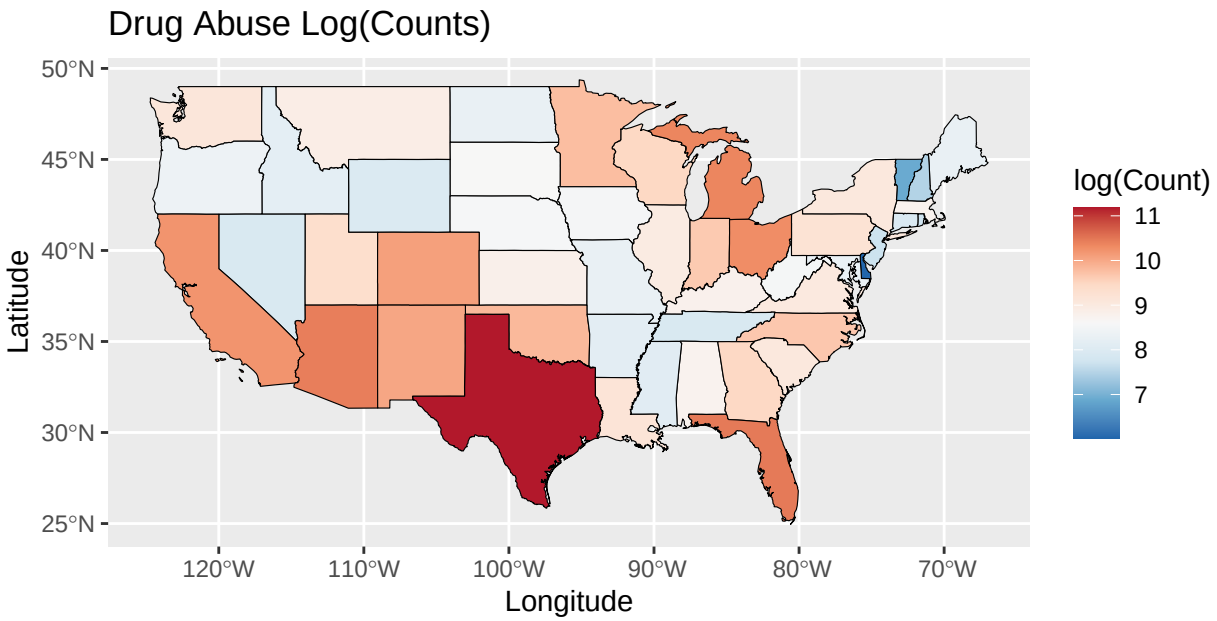
1. Create a pairs plot to assess the relationship between `log(Count)` and the explanatory variables (note we are using `log(Count)` here as the response because Poisson regression is log-linear).

```
ggpairs(data_red)
```



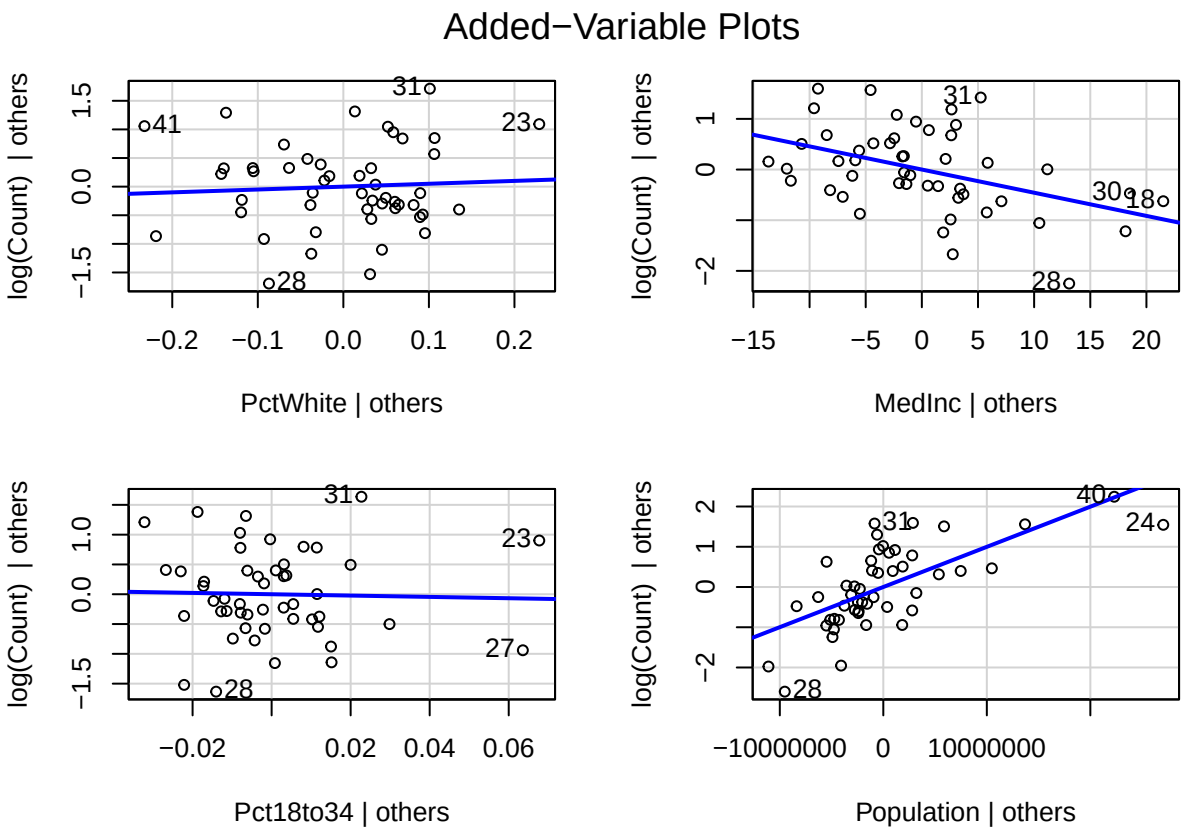
2. Create a choropleth map of `log(Count)`.

```
ggplot(data=myShp) +
  geom_sf(mapping=aes(fill=log(Count)), color="black") +
  scale_fill_distiller(palette="RdBu") +
  labs(title = "Drug Abuse Log(Counts)",
       x = "Longitude",
       y = "Latitude")
```



3. Fit a `lm()` of `log(Count)` using `Population`, `PctWhite`, `MedInc` and `Pct18to34` as explanatory variables. Perform a Moran's I test on the residuals to see if there is spatial correlation in the residuals.

```
data_lm <- lm(log(Count)~ PctWhite + MedInc + Pct18to34 + Population, data)
avPlots(data_lm, ask=FALSE)
```




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

```
##
##  Moran I test under randomisation
##
## data:  resid(data_lm)
## weights: nb2listw(poly2nb(myShp))
##
## Moran I statistic standard deviate = 2.5921, p-value = 0.004769
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.230238791      -0.020833333       0.009381812
```

4. Perform a Geary's C test on your residuals from #3 above to double check if there is spatial correlation in the residuals.

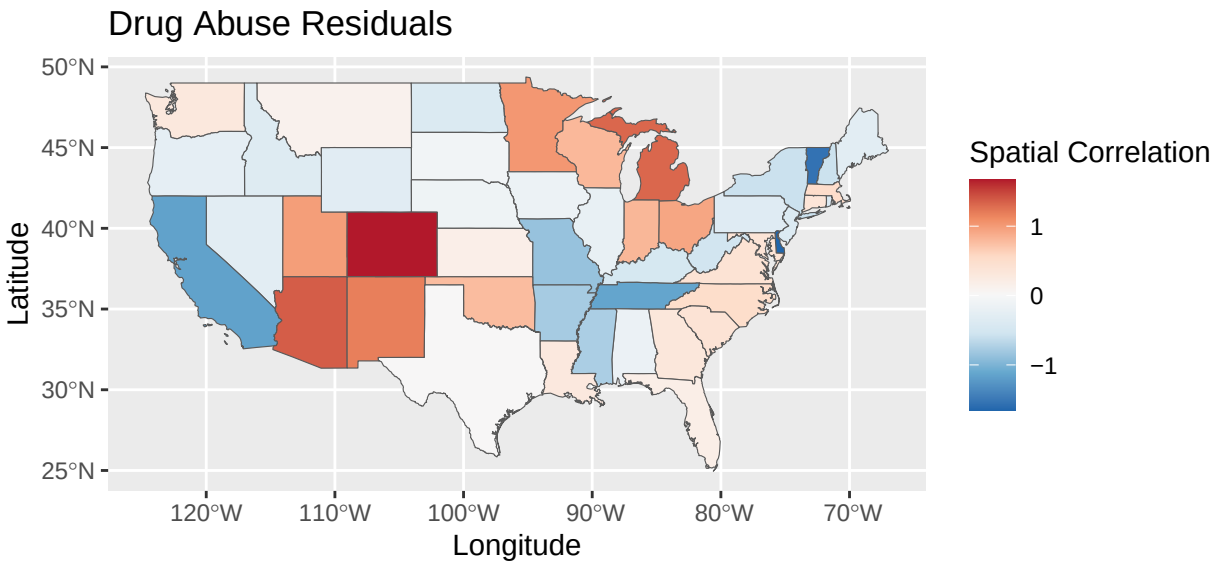
```
geary.test(x= resid(data_lm), listw=nb2listw(poly2nb(myShp)))
```

```
##
##  Geary C test under randomisation
##
## data:  resid(data_lm)
## weights: nb2listw(poly2nb(myShp))
```

```
##
## Geary C statistic standard deviate = 2.5474, p-value = 0.005427
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##        0.74378512        1.00000000        0.01011625
```

5. Map the residuals from the `lm()` fit to see if there is spatial correlation.

```
#plot residuals
ggplot(data=myShp) +
  geom_sf(mapping=aes(fill=data_lm$residuals)) +
  scale_fill_distiller(palette="RdBu", name="Spatial Correlation") +
  labs(title = "Drug Abuse Residuals",
       x = "Longitude",
       y = "Latitude")
```



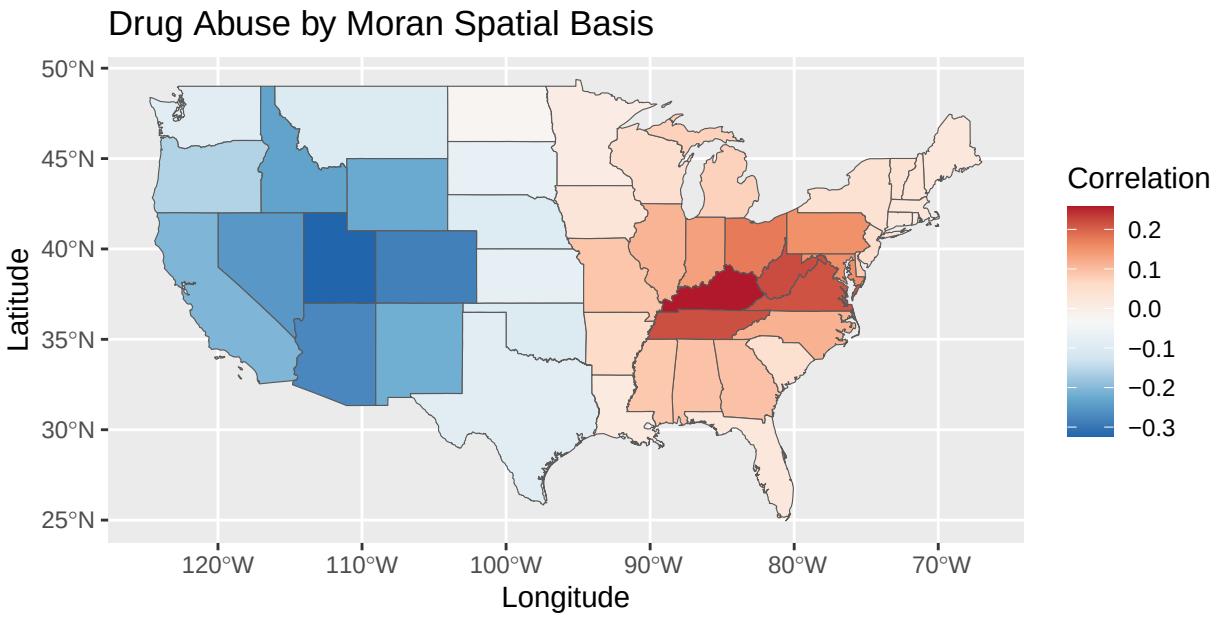
Defining Spatial Basis Functions

1. Create the adjacency matrix.

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

2. Create the Moran spatial basis and plot the first basis in a chloropleth map.

```
X <- model.matrix(data_lm)
M <- moranBasis(X, A, tol=0.95)
M <- data.frame(M)
ggplot(data=myShp) +
  geom_sf(mapping=aes(fill=M$B1)) +
  scale_fill_distiller(palette="RdBu", name="Correlation") +
  labs(title = "Drug Abuse by Moran Spatial Basis",
       x = "Longitude",
       y = "Latitude")
```



3. Merge the Moran spatial bases into your `myShpDF` data frame for use in fitting models later.

```
data <- bind_cols(data, M) #cbind() but efficient
```

Spatial GLM Model Fitting

1. Fit a spatial GLM model with `Count` as the response and using `PctWhite`, `MedInc`, `Population`, `Pct18to34` AND your spatial bases as explanatory variables. Print a `summary()` of the model to see your coefficient table.

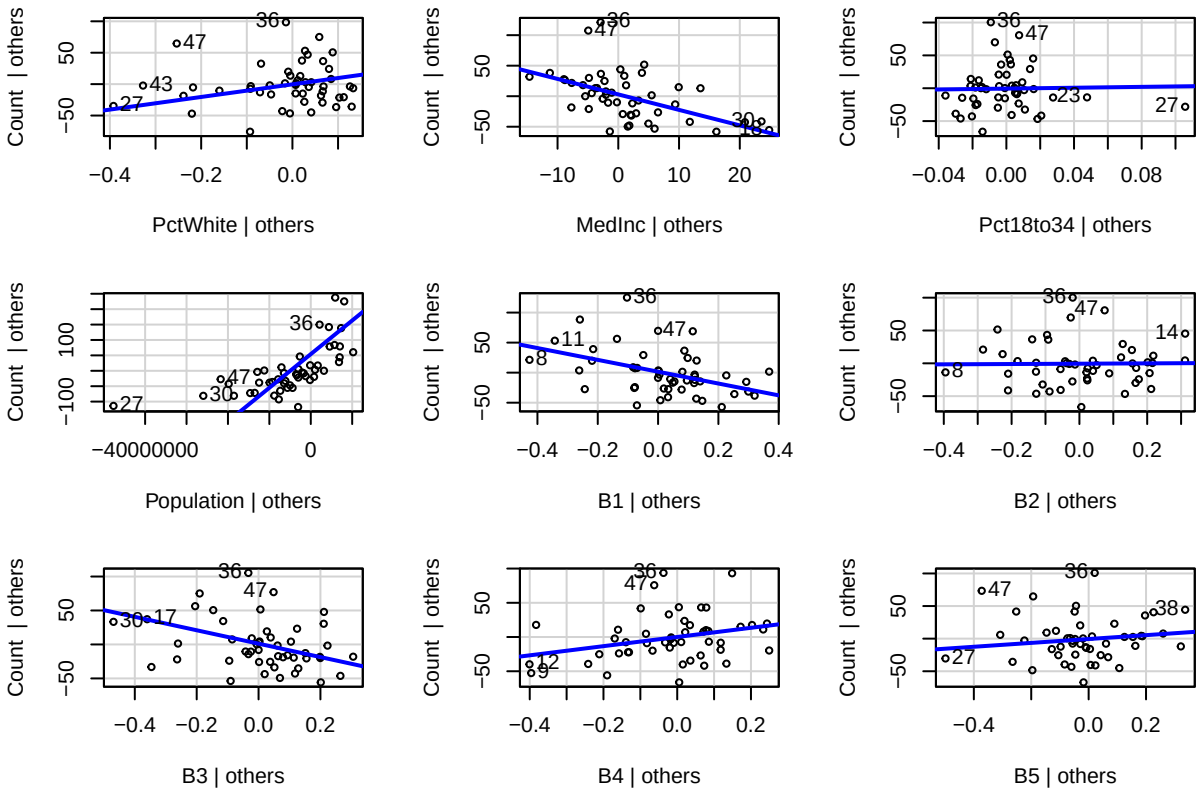
```
data_glm <- glm(formula=Count~.-State-Count, data=data, family=poisson)
summary(data_glm)
```

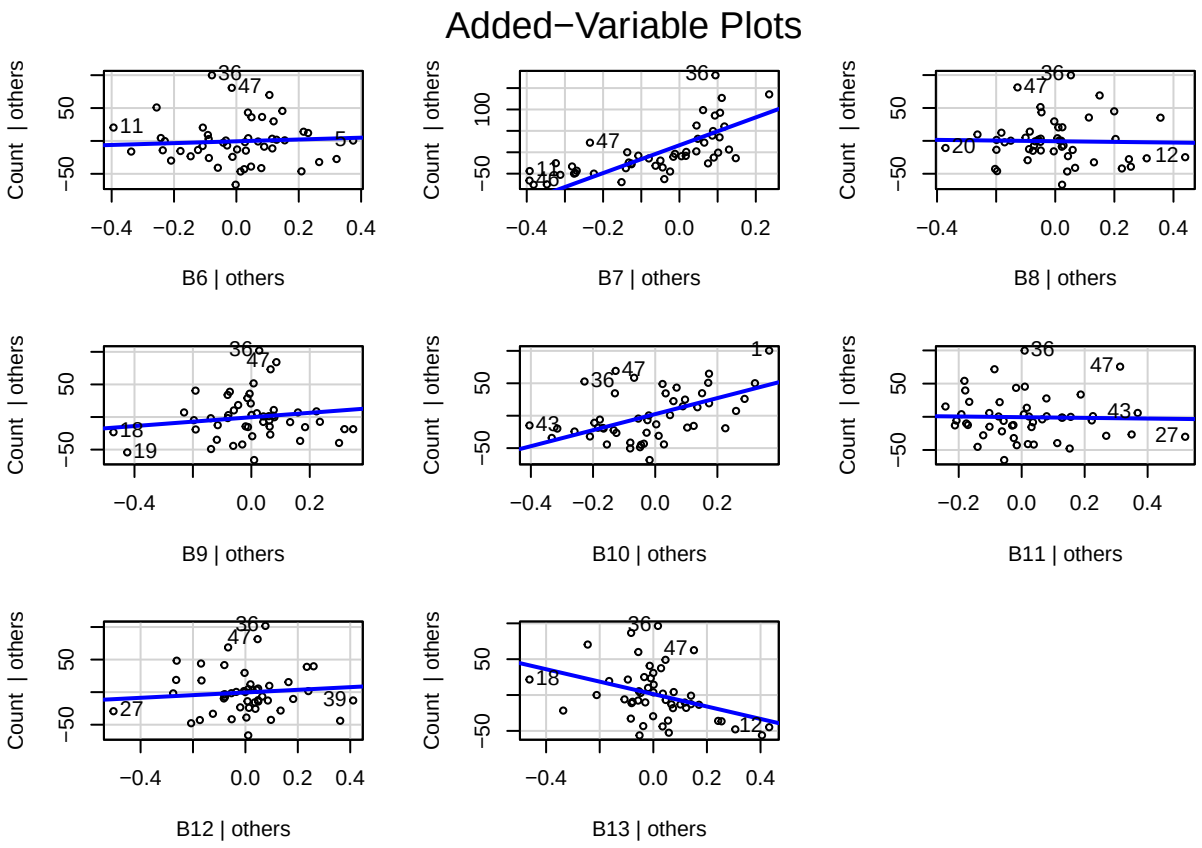
```
##
## Call:
## glm(formula = Count ~ . - State - Count, family = poisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -71.238  -23.869   -1.849   11.556   90.700
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  9.856e+00  2.699e-02  365.161  < 2e-16 ***
## PctWhite     6.351e-01  1.798e-02   35.327  < 2e-16 ***
## MedInc      -3.299e-02  2.491e-04 -132.425  < 2e-16 ***
## Pct18to34   -1.863e-01  9.051e-02   -2.059  0.03954 *
## Population   8.025e-08  1.877e-10  427.517  < 2e-16 ***
## B1          -1.202e+00  1.098e-02 -109.469  < 2e-16 ***
## B2          -1.085e-02  1.188e-02   -0.914  0.36089
## B3          -9.081e-01  1.017e-02  -89.267  < 2e-16 ***
## B4           9.765e-01  1.374e-02   71.047  < 2e-16 ***
## B5           1.921e-01  1.062e-02   18.084  < 2e-16 ***
## B6           1.199e-02  1.143e-02    1.049  0.29437
## B7           3.149e+00  1.123e-02  280.304  < 2e-16 ***
## B8          -3.388e-02  1.232e-02   -2.750  0.00595 **
## B9           3.500e-01  1.195e-02   29.282  < 2e-16 ***
## B10          1.164e+00  1.010e-02  115.231  < 2e-16 ***
## B11         -1.620e-01  1.207e-02  -13.423  < 2e-16 ***
## B12          1.128e-01  1.173e-02    9.619  < 2e-16 ***
## B13         -1.264e+00  1.258e-02 -100.454  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 496777  on 48  degrees of freedom
## Residual deviance:  53329  on 31  degrees of freedom
## AIC: 53887
##
## Number of Fisher Scoring iterations: 5
```

Validating Spatial MLR Model Assumptions and Predictions

1. Check the assumption of linearity using added-variable plots.

```
avPlots(data_glm, ask=FALSE)
```


Count | others
PctWhite | others
MedInc | others
Pct18to34 | others
Population | others
B1 | others
B2 | others
B3 | others
B4 | others
B5 | others



2. Check the assumption of independence by decorrelating residuals and performing Moran's I or Geary's C tests to make sure there is no more spatial correlation.

```
sres <- stdres.gls(data_glm)
moran.test(x= sres, listw=nb2listw(poly2nb(myShp)))
```

```
##
##  Moran I test under randomisation
##
## data:  sres
## weights: nb2listw(poly2nb(myShp))
##
## Moran I statistic standard deviate = -1.2132, p-value = 0.8875
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      -0.136699956      -0.020833333       0.009120747
```

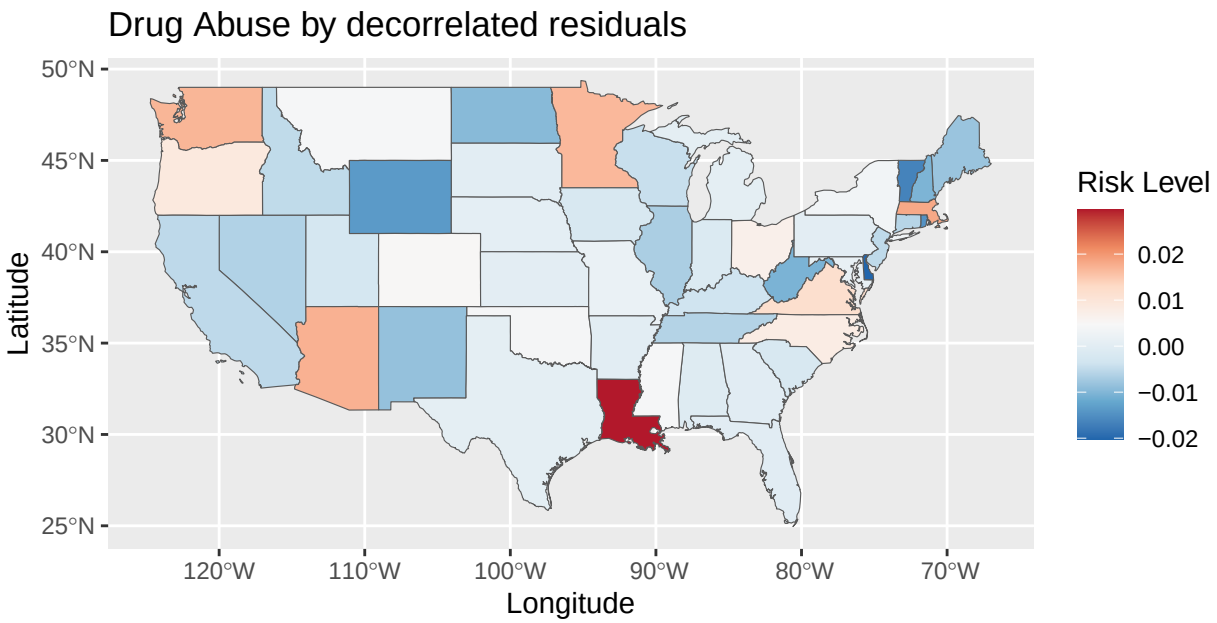
```
geary.test(x= sres, listw=nb2listw(poly2nb(myShp)))
```

```
##
##  Geary C test under randomisation
##
## data:  sres
```

```
## weights: nb2listw(poly2nb(myShp))
##
## Geary C statistic standard deviate = -0.50352, p-value = 0.6927
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##        1.05206138        1.00000000        0.01069049
```

3. Draw a choropleth map of the standardized and decorrelated residuals to visually verify that the residuals are no longer spatially correlated.

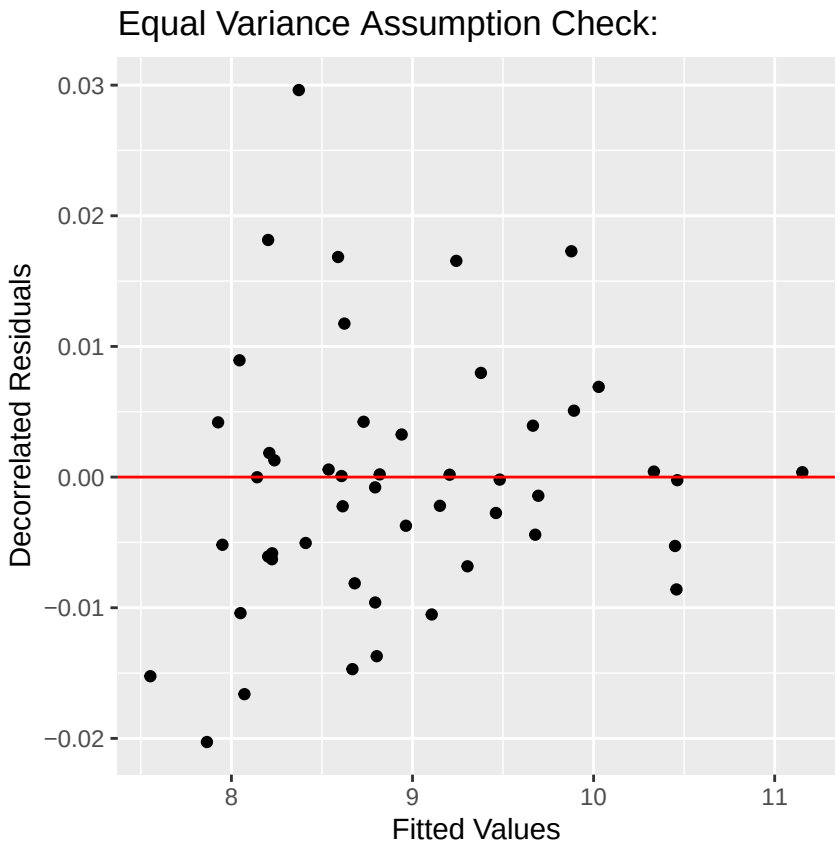
```
ggplot(data=myShp) +
  geom_sf(mapping=aes(fill=sres)) +
  scale_fill_distiller(palette="RdBu", name="Risk Level") +
  labs(title = "Drug Abuse by decorrelated residuals",
       x = "Longitude",
       y = "Latitude")
```



4. Check the assumption of equal variance by plotting the standardized and decorrelated residuals vs. the `log(fitted values)`.

```
ggplot(mapping = aes(x=log(fitted(data_glm)), y=sres)) +
  geom_point() +
  xlab('Fitted Values') +
```

```
  ylab('Decorrelated Residuals') +
  ggtitle('Equal Variance Assumption Check:') +
  geom_hline(yintercept = 0, col = "red") +
  theme(aspect.ratio = 1)
```



Statistical Inference

1. Print out the summary of the GLM model fit and identify the estimates and 95% confidence intervals of your explanatory variables.

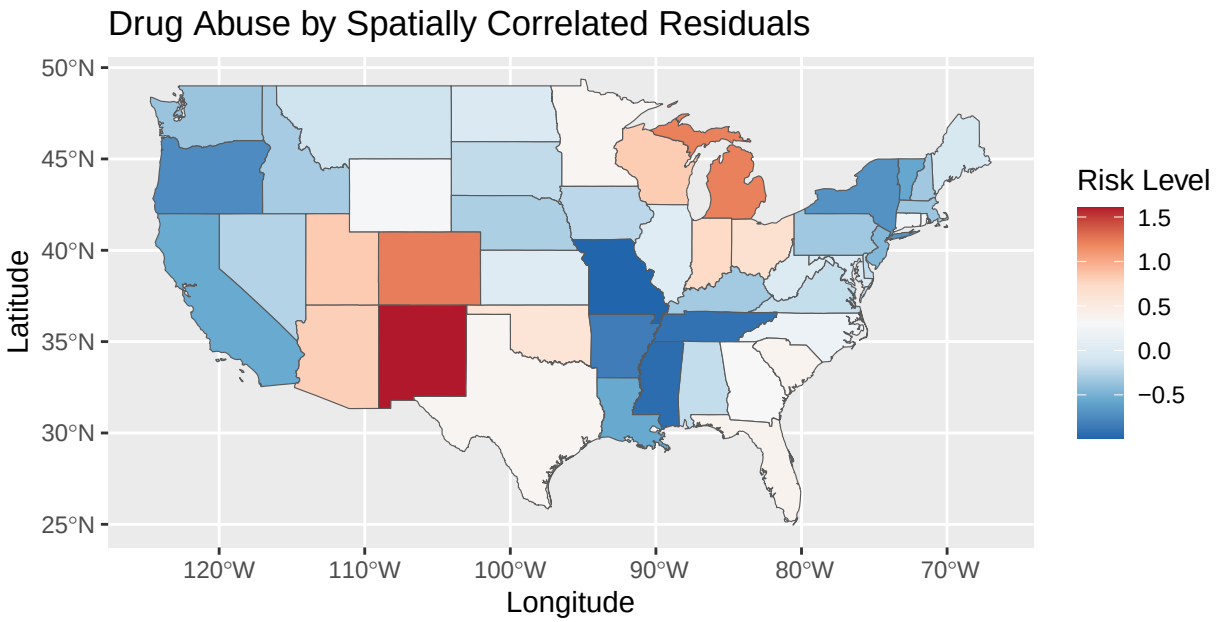
```
summary(data_glm)
```

```
##
## Call:
## glm(formula = Count ~ . - State - Count, family = poisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -71.238  -23.869   -1.849   11.556   90.700
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  9.856e+00  2.699e-02  365.161  < 2e-16 ***
```

```
## PctWhite     6.351e-01  1.798e-02   35.327  < 2e-16 ***
## MedInc      -3.299e-02  2.491e-04 -132.425  < 2e-16 ***
## Pct18to34   -1.863e-01  9.051e-02   -2.059  0.03954 *
## Population   8.025e-08  1.877e-10  427.517  < 2e-16 ***
## B1          -1.202e+00  1.098e-02 -109.469  < 2e-16 ***
## B2          -1.085e-02  1.188e-02   -0.914  0.36089
## B3          -9.081e-01  1.017e-02  -89.267  < 2e-16 ***
## B4           9.765e-01  1.374e-02   71.047  < 2e-16 ***
## B5           1.921e-01  1.062e-02   18.084  < 2e-16 ***
## B6           1.199e-02  1.143e-02    1.049  0.29437
## B7           3.149e+00  1.123e-02  280.304  < 2e-16 ***
## B8          -3.388e-02  1.232e-02   -2.750  0.00595 **
## B9           3.500e-01  1.195e-02   29.282  < 2e-16 ***
## B10          1.164e+00  1.010e-02  115.231  < 2e-16 ***
## B11         -1.620e-01  1.207e-02  -13.423  < 2e-16 ***
## B12          1.128e-01  1.173e-02    9.619  < 2e-16 ***
## B13         -1.264e+00  1.258e-02 -100.454  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 496777  on 48  degrees of freedom
## Residual deviance:  53329  on 31  degrees of freedom
## AIC: 53887
##
## Number of Fisher Scoring iterations: 5
```

2. Create a chloropleth map of the spatially correlated residuals (just the $\mathbf{b}_i'\hat{\theta}$ part) to identify states that, after accounting for the explanatory variables, have an elevated level of risk.

```
#X\beta +M\theta
M.theta = as.matrix(M) %*% coef(data_glm)[-(1:5)]
ggplot(data=myShp) +
  geom_sf(mapping=aes(fill=M.theta)) +
  scale_fill_distiller(palette="RdBu", name = "Risk Level") +
  labs(title = "Drug Abuse by Spatially Correlated Residuals",
       x = "Longitude",
       y = "Latitude")
```


Drug Abuse by Spatially Correlated Residuals
50°N
45°N
40°N
35°N
30°N
25°N
Latitude
120°W
110°W
100°W
90°W
80°W
70°W
Longitude
Risk Level
1.5
1.0
0.5
0.0
−0.5