Analysis before fitting the CAR model

Alvin Sheng

6/28/2021

```
library(here)
```

```
## here() starts at /Users/Alvin/Documents/NCSU_Fall_2021/NIH_SIP/flood-risk-health-effects
```

```
library(ape)
library(GGally)
```

```
## Loading required package: ggplot2
```

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

```
library(usdm)
```

```
## Loading required package: sp
```

```
## Loading required package: raster
```

```
##
## Attaching package: 'raster'
```

```
## The following objects are masked from 'package:ape':
##
##     rotate, zoom
```

```
fhs_model_df <- readRDS(here("intermediary_data/fhs_model_df_sw_states_census_tract.rds"))
```

Checking for multicollinearity among the covariates

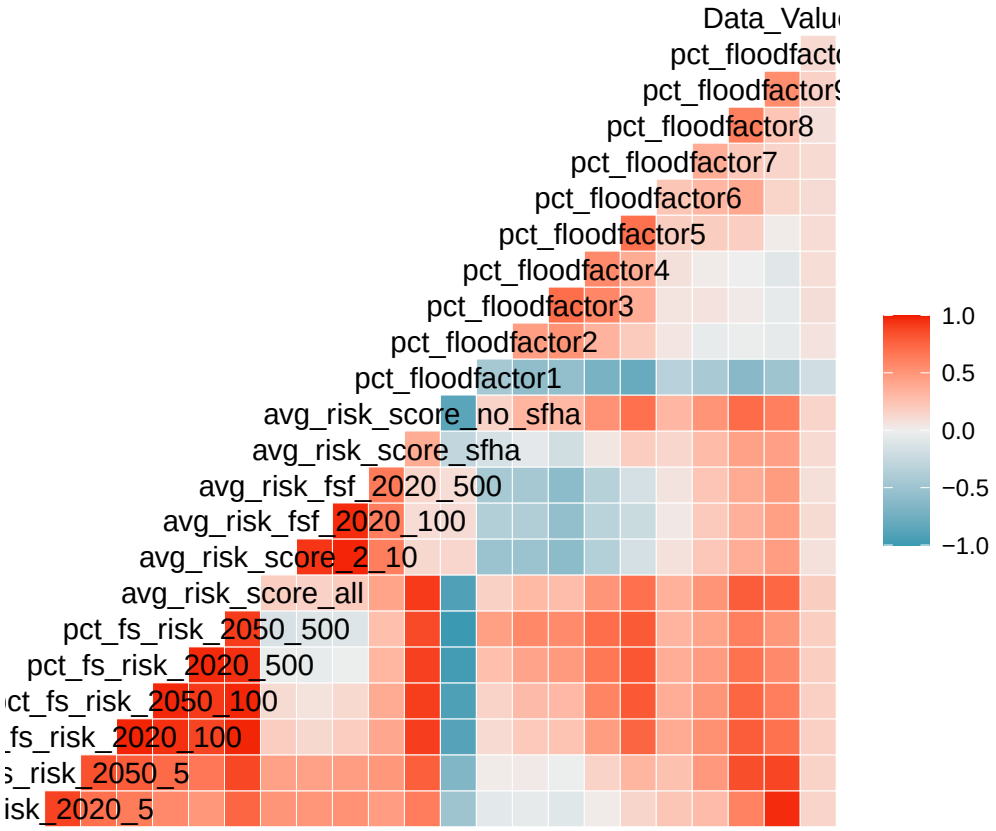
`S.CARleroux()` automatically puts a fixed ridge penalty on the beta coefficients. Therefore, the large number of covariates and multicollinearity would be accounted for.

Actually no, because the penalty is negligible.

Flood risk variables

```
ggcorr(data = fhs_model_df[, c(14:35, ncol(fhs_model_df))], progress = F)
```

```
## Warning: Ignoring unknown parameters: progress
```




```
flood_cor <- cor(fhs_model_df[complete.cases(fhs_model_df[, c(14:35, ncol(fhs_model_df))]), c(14:35, nc

flood_cor[nrow(flood_cor), ] # correlation with dependent variable
```

```
##     pct_fs_risk_2020_5     pct_fs_risk_2050_5   pct_fs_risk_2020_100
##              0.14582097             0.15653522             0.17415552
##   pct_fs_risk_2050_100   pct_fs_risk_2020_500   pct_fs_risk_2050_500
##              0.17112785             0.18469235             0.18376377
##     avg_risk_score_all    avg_risk_score_2_10  avg_risk_fsf_2020_100
##              0.18542687             0.07342158             0.11063517
##  avg_risk_fsf_2020_500    avg_risk_score_sfha avg_risk_score_no_sfha
##              0.08567016             0.11412867             0.15046242
##       pct_floodfactor1       pct_floodfactor2       pct_floodfactor3
##             -0.18386832             0.06727516             0.09533764
##       pct_floodfactor4       pct_floodfactor5       pct_floodfactor6
##              0.10025324             0.10659157             0.10935650
##       pct_floodfactor7       pct_floodfactor8       pct_floodfactor9
##              0.11302768             0.08609573             0.16379999
##      pct_floodfactor10         Data_Value_CHD
##              0.11571612             1.00000000
```

For each variable, I take the summary of its correlations with other variables, not including itself.

```
diag(flood_cor) <- NA

summary(flood_cor)
```

```
##  pct_fs_risk_2020_5 pct_fs_risk_2050_5 pct_fs_risk_2020_100
##  Min.   :-0.4867    Min.   :-0.6574    Min.   :-0.8862
```

```
##  1st Qu.: 0.1471    1st Qu.: 0.1940    1st Qu.: 0.1881
##  Median : 0.4913    Median : 0.4704    Median : 0.5027
##  Mean   : 0.3782    Mean   : 0.4491    Mean   : 0.4812
##  3rd Qu.: 0.6189    3rd Qu.: 0.7849    3rd Qu.: 0.8204
##  Max.   : 0.9625    Max.   : 0.8874    Max.   : 0.9836
##  NA's   :1          NA's   :1          NA's   :1
##  pct_fs_risk_2050_100 pct_fs_risk_2020_500 pct_fs_risk_2050_500
##  Min.   :-0.9264      Min.   :-0.9698      Min.   :-1.0000
##  1st Qu.: 0.2043      1st Qu.: 0.2854      1st Qu.: 0.2817
##  Median : 0.5584      Median : 0.5194      Median : 0.5276
##  Mean   : 0.4780      Mean   : 0.4637      Mean   : 0.4431
##  3rd Qu.: 0.8078      3rd Qu.: 0.7983      3rd Qu.: 0.7844
##  Max.   : 0.9836      Max.   : 0.9747      Max.   : 0.9698
##  NA's   :1            NA's   :1            NA's   :1
##  avg_risk_score_all avg_risk_score_2_10 avg_risk_fsf_2020_100
##  Min.   :-0.9180    Min.   :-0.6029     Min.   :-0.5591
##  1st Qu.: 0.2123    1st Qu.:-0.1232     1st Qu.:-0.1058
##  Median : 0.5188    Median : 0.1412     Median : 0.1122
##  Mean   : 0.4978    Mean   : 0.1368     Mean   : 0.1415
##  3rd Qu.: 0.8524    3rd Qu.: 0.4117     3rd Qu.: 0.4105
##  Max.   : 0.9787    Max.   : 0.9809     Max.   : 0.9629
##  NA's   :1          NA's   :1           NA's   :1
##  avg_risk_fsf_2020_500 avg_risk_score_sfha avg_risk_score_no_sfha
##  Min.   :-0.60025      Min.   :-0.2653     Min.   :-0.8639
##  1st Qu.:-0.09466      1st Qu.: 0.1190     1st Qu.: 0.1942
##  Median : 0.12881      Median : 0.3446     Median : 0.5169
##  Mean   : 0.15292      Mean   : 0.2724     Mean   : 0.4612
##  3rd Qu.: 0.43328      3rd Qu.: 0.4396     3rd Qu.: 0.7648
##  Max.   : 0.98093      Max.   : 0.6442     Max.   : 0.9209
##  NA's   :1             NA's   :1           NA's   :1
##  pct_floodfactor1  pct_floodfactor2   pct_floodfactor3   pct_floodfactor4
##  Min.   :-1.0000   Min.   :-0.51577   Min.   :-0.57945   Min.   :-0.60290
##  1st Qu.:-0.8497   1st Qu.:-0.06404   1st Qu.:-0.03908   1st Qu.:-0.07891
##  Median :-0.5740   Median : 0.05967   Median : 0.08198   Median : 0.08973
##  Mean   :-0.5341   Mean   : 0.04119   Mean   : 0.11502   Mean   : 0.08705
##  3rd Qu.:-0.3560   3rd Qu.: 0.19447   3rd Qu.: 0.36236   3rd Qu.: 0.36434
##  Max.   : 0.1530   Max.   : 0.51496   Max.   : 0.70795   Max.   : 0.70795
##  NA's   :1         NA's   :1          NA's   :1          NA's   :1
##  pct_floodfactor5   pct_floodfactor6  pct_floodfactor7   pct_floodfactor8
##  Min.   :-0.71639   Min.   :-0.8070   Min.   :-0.33125   Min.   :-0.4301
##  1st Qu.: 0.03627   1st Qu.: 0.1530   1st Qu.: 0.07139   1st Qu.: 0.1948
##  Median : 0.19093   Median : 0.3293   Median : 0.20357   Median : 0.2959
##  Mean   : 0.22319   Mean   : 0.3097   Mean   : 0.18374   Mean   : 0.2810
##  3rd Qu.: 0.55406   3rd Qu.: 0.6982   3rd Qu.: 0.32781   3rd Qu.: 0.4908
##  Max.   : 0.71642   Max.   : 0.8239   Max.   : 0.38890   Max.   : 0.6179
##  NA's   :1          NA's   :1         NA's   :1          NA's   :1
##  pct_floodfactor9  pct_floodfactor10 Data_Value_CHD
##  Min.   :-0.6149   Min.   :-0.4859   Min.   :-0.18387
##  1st Qu.: 0.1977   1st Qu.: 0.1226   1st Qu.: 0.09657
##  Median : 0.4225   Median : 0.4473   Median : 0.11358
##  Mean   : 0.4021   Mean   : 0.3597   Mean   : 0.11406
##  3rd Qu.: 0.6651   3rd Qu.: 0.6027   3rd Qu.: 0.16198
##  Max.   : 0.8488   Max.   : 0.9625   Max.   : 0.18543
##  NA's   :1         NA's   :1         NA's   :1
```

Many of the flood risk variables are very correlated.

Using VIF to exlude variables

```
fhs_model_df <- readRDS(here("intermediary_data/fhs_model_df_sw_states_census_tract.rds"))
```

```
X <- fhs_model_df[, 14:(ncol(fhs_model_df) - 1)]

X <- X[, names(X) != "pct_floodfactor1"]

X           <- scale(X) # Scale covariates

X <- as.data.frame(X)
```

```
vif(X)
```

```
##                 Variables          VIF
## 1      pct_fs_risk_2020_5 2.445053e+01
## 2      pct_fs_risk_2050_5 4.635759e+01
## 3    pct_fs_risk_2020_100 1.834865e+02
## 4    pct_fs_risk_2050_100 2.205339e+02
## 5    pct_fs_risk_2020_500 2.049111e+02
## 6    pct_fs_risk_2050_500 3.522854e+05
## 7      avg_risk_score_all 2.985917e+05
## 8     avg_risk_score_2_10 9.042961e+01
## 9   avg_risk_fsf_2020_100 3.008846e+01
## 10  avg_risk_fsf_2020_500 1.148603e+02
## 11    avg_risk_score_sfha 3.004347e+00
## 12 avg_risk_score_no_sfha 6.683746e+00
## 13       pct_floodfactor2 1.242851e+04
## 14       pct_floodfactor3 1.004510e+04
## 15       pct_floodfactor4 9.398613e+03
## 16       pct_floodfactor5 6.785731e+03
## 17       pct_floodfactor6 8.896006e+04
## 18       pct_floodfactor7 4.640655e+03
## 19       pct_floodfactor8 4.995250e+02
## 20       pct_floodfactor9 4.463901e+04
## 21      pct_floodfactor10 1.229045e+05
## 22                 EP_POV 3.788224e+00
## 23               EP_UNEMP 1.815064e+00
## 24                 EP_PCI 3.076932e+00
## 25              EP_NOHSDP 5.165273e+00
## 26               EP_AGE65 3.278709e+00
## 27               EP_AGE17 3.330624e+00
## 28              EP_DISABL 2.841119e+00
## 29              EP_SNGPNT 2.861687e+00
## 30              EP_MINRTY 3.720479e+00
## 31              EP_LIMENG 2.983930e+00
## 32               EP_MUNIT 2.023274e+00
## 33              EP_MOBILE 2.075563e+00
## 34               EP_CROWD 1.856420e+00
## 35               EP_NOVEH 2.668470e+00
## 36              EP_GROUPQ 1.517822e+00
```

```
## 37             EP_UNINSUR 2.612214e+00
## 38                     co 2.443587e+00
## 39                    no2 5.153443e+00
## 40                     o3 4.378701e+00
## 41                   pm10 2.057575e+00
## 42                   pm25 4.315523e+00
## 43                    so2 1.562599e+00
## 44    Data_Value_CSMOKING 7.702837e+00
```

```
vifstep(X)
```

```
## 9 variables from the 44 input variables have collinearity problem:
##
## pct_fs_risk_2050_500 avg_risk_score_all pct_fs_risk_2050_100 pct_fs_risk_2020_500 pct_fs_risk_2020_1
##
## After excluding the collinear variables, the linear correlation coefficients ranges between:
## min correlation ( EP_GROUPQ ~ pct_floodfactor7 ):  -2.996671e-05
## max correlation ( Data_Value_CSMOKING ~ EP_NOHSDP ):  0.7582366
##
## ---------- VIFs of the remained variables --------
##                 Variables      VIF
## 1   avg_risk_fsf_2020_100 4.825290
## 2     avg_risk_score_sfha 2.703222
## 3  avg_risk_score_no_sfha 8.878878
## 4        pct_floodfactor2 1.537687
## 5        pct_floodfactor3 2.386257
## 6        pct_floodfactor4 3.314557
## 7        pct_floodfactor5 3.058115
## 8        pct_floodfactor6 3.776082
## 9        pct_floodfactor7 1.343630
## 10       pct_floodfactor8 1.910825
## 11       pct_floodfactor9 3.536040
## 12      pct_floodfactor10 4.121883
## 13                 EP_POV 3.909199
## 14               EP_UNEMP 1.787206
## 15                 EP_PCI 3.110533
## 16              EP_NOHSDP 4.971912
## 17               EP_AGE65 3.257541
## 18               EP_AGE17 3.379153
## 19              EP_DISABL 2.894234
## 20              EP_SNGPNT 2.853221
## 21              EP_MINRTY 3.746649
## 22              EP_LIMENG 3.072245
## 23               EP_MUNIT 2.114477
## 24              EP_MOBILE 2.143679
## 25               EP_CROWD 1.907051
## 26               EP_NOVEH 2.673660
## 27              EP_GROUPQ 1.462990
## 28             EP_UNINSUR 2.591774
## 29                     co 2.385194
## 30                    no2 5.074562
## 31                     o3 4.149114
## 32                   pm10 2.013634
## 33                   pm25 4.237138
## 34                    so2 1.638362
```

```
## 35    Data_Value_CSMOKING 7.520142
```

This procedure detects that the following variables have collinearity problems. Let's exclude these variables and then rerun the analysis.

```
collin_var_names <- c("pct_fs_risk_2050_500", "avg_risk_score_all", "pct_fs_risk_2050_100",
                      "pct_fs_risk_2020_500", "pct_fs_risk_2020_100", "avg_risk_fsf_2020_500",
                      "pct_fs_risk_2050_5", "avg_risk_score_2_10", "pct_fs_risk_2020_5")
```

Non-spatial modeling

```
Y <- fhs_model_df$Data_Value_CHD

# extract the covariates matrix

X <- fhs_model_df[, 14:(ncol(fhs_model_df) - 1)]

X <- X[, names(X) != "pct_floodfactor1"]

# exclude some more variables selected by vifstep, to account for multicollinearity
# excluding all of the pct_fs_risk variables, as well as 3 of the avg_risk_score variables

collin_var_names <- c("pct_fs_risk_2050_500", "avg_risk_score_all", "pct_fs_risk_2050_100",
                      "pct_fs_risk_2020_500", "pct_fs_risk_2020_100", "avg_risk_fsf_2020_500",
                      "pct_fs_risk_2050_5", "avg_risk_score_2_10", "pct_fs_risk_2020_5")

X <- X[, !(names(X) %in% collin_var_names)]

X           <- scale(X) # Scale covariates
X[is.na(X)] <- 0        # Fill in missing values with the mean
```

```
fhs_lm <- lm(Y ~ X)
```

```
summary(fhs_lm)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3210 -0.5478 -0.0100  0.5264 10.9216
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.581987   0.008314 911.910  < 2e-16 ***
## Xavg_risk_fsf_2020_100   0.027546   0.018061   1.525 0.127251
## Xavg_risk_score_sfha     0.028514   0.013723   2.078 0.037742 *
## Xavg_risk_score_no_sfha -0.079681   0.020614  -3.865 0.000111 ***
## Xpct_floodfactor2       -0.021880   0.010166  -2.152 0.031393 *
## Xpct_floodfactor3       -0.010416   0.012883  -0.808 0.418837
## Xpct_floodfactor4        0.021847   0.014666   1.490 0.136363
## Xpct_floodfactor5       -0.014213   0.014087  -1.009 0.313044
```

```
## Xpct_floodfactor6        0.042080   0.015668   2.686 0.007248 **
## Xpct_floodfactor7       -0.010295   0.009418  -1.093 0.274338
## Xpct_floodfactor8       -0.004371   0.011634  -0.376 0.707133
## Xpct_floodfactor9       -0.028049   0.015427  -1.818 0.069060 .
## Xpct_floodfactor10       0.065390   0.014061   4.650 3.35e-06 ***
## XEP_POV                  0.367352   0.016076  22.851  < 2e-16 ***
## XEP_UNEMP                0.037585   0.011013   3.413 0.000645 ***
## XEP_PCI                 -0.174493   0.014368 -12.144  < 2e-16 ***
## XEP_NOHSDP               0.172869   0.018422   9.384  < 2e-16 ***
## XEP_AGE65                1.954234   0.015029 130.027  < 2e-16 ***
## XEP_AGE17                0.324119   0.015379  21.075  < 2e-16 ***
## XEP_DISABL               0.293560   0.013677  21.463  < 2e-16 ***
## XEP_SNGPNT              -0.117817   0.013948  -8.447  < 2e-16 ***
## XEP_MINRTY              -0.095578   0.015539  -6.151 7.93e-10 ***
## XEP_LIMENG               0.117819   0.014080   8.368  < 2e-16 ***
## XEP_MUNIT               -0.116269   0.011714  -9.926  < 2e-16 ***
## XEP_MOBILE               0.111058   0.011953   9.291  < 2e-16 ***
## XEP_CROWD               -0.037705   0.011393  -3.309 0.000937 ***
## XEP_NOVEH                0.136009   0.013333  10.201  < 2e-16 ***
## XEP_GROUPQ              -0.165014   0.010433 -15.817  < 2e-16 ***
## XEP_UNINSUR              0.012927   0.013099   0.987 0.323740
## Xco                      0.033148   0.012777   2.594 0.009489 **
## Xno2                     0.095318   0.018533   5.143 2.74e-07 ***
## Xo3                     -0.268664   0.016904 -15.893  < 2e-16 ***
## Xpm10                    0.007758   0.011701   0.663 0.507344
## Xpm25                    0.075061   0.016992   4.418 1.01e-05 ***
## Xso2                     0.060449   0.010378   5.825 5.85e-09 ***
## XData_Value_CSMOKING     0.778595   0.022539  34.545  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9367 on 12655 degrees of freedom
## Multiple R-squared:  0.8557, Adjusted R-squared:  0.8553
## F-statistic:  2144 on 35 and 12655 DF,  p-value: < 2.2e-16
```

Checking for spatial autocorrelation

```
W <- readRDS(here("intermediary_data", "census_tract_adj_reorganize_sw_states_census_tract.rds"))
```

Moran's I

```
(moran_results <- Moran.I(residuals(fhs_lm), W))
```

```
## $observed
## [1] 0.2207244
##
## $expected
## [1] -7.880221e-05
##
## $sd
## [1] 0.005162701
##
## $p.value
```

```
## [1] 0
```

The p-value is negligible, so we can reject the null hypothesis of zero spatial autocorrelation. Since the observed value of I is significantly greater then the expected value, the life expectancies are positively autocorrelated, in contrast to negatively autocorrelated. Thus, using a CAR model is justified.