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

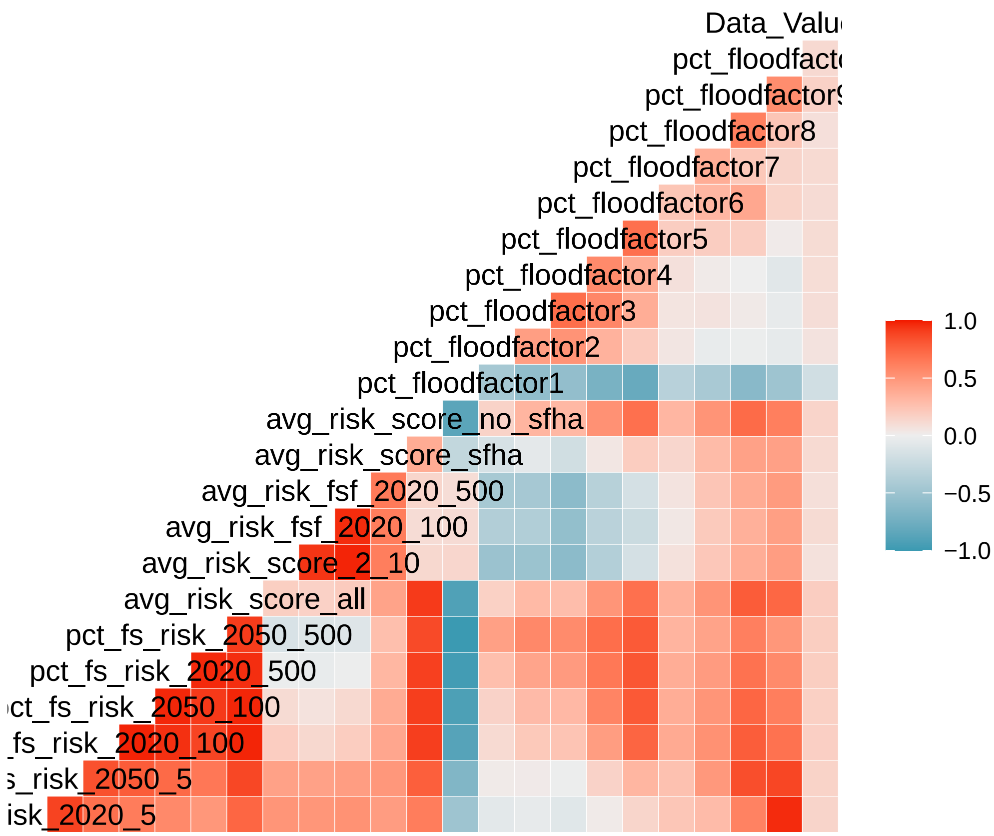
`S.CARleroux()` automatically puts a fixed ridge penalty on the beta coefficients. Therefore, the large number of covariates and multicollinearity would be accounted for.

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
```




```r
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.

```r
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.

Non-spatial modeling

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

```r
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"]

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

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

```r
summary(fhs_lm)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3892 -0.5446 -0.0078  0.5276 10.8411
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.581987   0.008303 913.182  < 2e-16 ***
## Xpct_fs_risk_2020_5      0.015678   0.038470   0.408 0.683626
## Xpct_fs_risk_2050_5      0.025576   0.056376   0.454 0.650079
## Xpct_fs_risk_2020_100   -0.170575   0.113321  -1.505 0.132288
## Xpct_fs_risk_2050_100    0.050143   0.121257   0.414 0.679228
## Xpct_fs_risk_2020_500    0.064952   0.118730   0.547 0.584352
## Xpct_fs_risk_2050_500  -20.801234   5.171286  -4.022 5.79e-05 ***
## Xavg_risk_score_all     13.893206   4.480820   3.101 0.001936 **
## Xavg_risk_score_2_10    -0.102668   0.076632  -1.340 0.180349
## Xavg_risk_fsf_2020_100  -0.121437   0.043862  -2.769 0.005638 **
## Xavg_risk_fsf_2020_500   0.265829   0.085708   3.102 0.001929 **
## Xavg_risk_score_sfha     0.020451   0.014050   1.456 0.145514
## Xavg_risk_score_no_sfha -0.083596   0.020838  -4.012 6.06e-05 ***
## Xpct_floodfactor2        3.511656   1.007135   3.487 0.000491 ***
## Xpct_floodfactor3        2.603303   0.887551   2.933 0.003362 **
## Xpct_floodfactor4        2.013200   0.835979   2.408 0.016046 *
## Xpct_floodfactor5        1.271077   0.713056   1.783 0.074679 .
## Xpct_floodfactor6        3.471896   2.567213   1.352 0.176272
## Xpct_floodfactor7        0.500765   0.571665   0.876 0.381059
## Xpct_floodfactor8        0.090543   0.195194   0.464 0.642757
## Xpct_floodfactor9        0.282674   1.875299   0.151 0.880187
## Xpct_floodfactor10      -0.296524   2.764210  -0.107 0.914574
```

```
## XEP_POV                    0.365663   0.016076  22.746  < 2e-16 ***
## XEP_UNEMP                  0.037642   0.011000   3.422 0.000623 ***
## XEP_PCI                   -0.171351   0.014420 -11.883  < 2e-16 ***
## XEP_NOHSDP                 0.166613   0.018571   8.972  < 2e-16 ***
## XEP_AGE65                  1.954923   0.015098 129.480  < 2e-16 ***
## XEP_AGE17                  0.322681   0.015366  20.999  < 2e-16 ***
## XEP_DISABL                 0.291074   0.013667  21.297  < 2e-16 ***
## XEP_SNGPNT                -0.118984   0.013939  -8.536  < 2e-16 ***
## XEP_MINRTY                -0.091515   0.015554  -5.884 4.12e-09 ***
## XEP_LIMENG                 0.117583   0.014114   8.331  < 2e-16 ***
## XEP_MUNIT                 -0.115448   0.011718  -9.852  < 2e-16 ***
## XEP_MOBILE                 0.116345   0.011993   9.701  < 2e-16 ***
## XEP_CROWD                 -0.037408   0.011390  -3.284 0.001025 **
## XEP_NOVEH                  0.139692   0.013399  10.426  < 2e-16 ***
## XEP_GROUPQ                -0.164887   0.010426 -15.815  < 2e-16 ***
## XEP_UNINSUR                0.012572   0.013187   0.953 0.340404
## Xco                        0.029732   0.012957   2.295 0.021770 *
## Xno2                       0.100091   0.018674   5.360 8.47e-08 ***
## Xo3                       -0.266974   0.017227 -15.498  < 2e-16 ***
## Xpm10                      0.017369   0.011969   1.451 0.146734
## Xpm25                      0.070791   0.017163   4.125 3.74e-05 ***
## Xso2                       0.060820   0.010457   5.816 6.17e-09 ***
## XData_Value_CSMOKING       0.786596   0.022682  34.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9353 on 12646 degrees of freedom
## Multiple R-squared:  0.8562, Adjusted R-squared:  0.8557
## F-statistic:  1711 on 44 and 12646 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] 0.2171751
##
## $expected
## [1] -7.880221e-05
##
## $sd
## [1] 0.005162714
##
## $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.

Using VIF to exlude variables

```
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 2.030010e+01
## 2      pct_fs_risk_2050_5 5.097545e+01
## 3    pct_fs_risk_2020_100 1.511176e+02
## 4    pct_fs_risk_2050_100 1.965348e+02
## 5    pct_fs_risk_2020_500 1.870803e+02
## 6    pct_fs_risk_2050_500 3.257059e+05
## 7      avg_risk_score_all 2.835344e+05
## 8     avg_risk_score_2_10 8.712119e+01
## 9   avg_risk_fsf_2020_100 2.918216e+01
## 10  avg_risk_fsf_2020_500 1.060027e+02
## 11    avg_risk_score_sfha 2.969112e+00
## 12 avg_risk_score_no_sfha 7.726870e+00
## 13       pct_floodfactor2 1.187250e+04
## 14       pct_floodfactor3 1.003070e+04
## 15       pct_floodfactor4 9.101649e+03
## 16       pct_floodfactor5 6.746566e+03
## 17       pct_floodfactor6 8.686556e+04
## 18       pct_floodfactor7 3.952087e+03
## 19       pct_floodfactor8 4.972143e+02
## 20       pct_floodfactor9 4.448970e+04
## 21      pct_floodfactor10 1.061683e+05
## 22                 EP_POV 3.765364e+00
## 23               EP_UNEMP 1.795772e+00
## 24                 EP_PCI 2.872660e+00
## 25              EP_NOHSDP 5.053680e+00
## 26               EP_AGE65 3.362087e+00
## 27               EP_AGE17 3.411777e+00
## 28              EP_DISABL 2.815037e+00
## 29              EP_SNGPNT 2.851369e+00
## 30              EP_MINRTY 3.583324e+00
## 31              EP_LIMENG 3.003150e+00
## 32               EP_MUNIT 1.999135e+00
## 33              EP_MOBILE 2.122437e+00
## 34               EP_CROWD 1.887161e+00
## 35               EP_NOVEH 2.540797e+00
## 36              EP_GROUPQ 1.547463e+00
## 37             EP_UNINSUR 2.518096e+00
## 38                     co 2.440484e+00
## 39                    no2 5.206342e+00
## 40                     o3 4.403776e+00
## 41                   pm10 2.119730e+00
```

```
## 42                   pm25 4.334853e+00
## 43                    so2 1.661277e+00
## 44    Data_Value_CSMOKING 7.442112e+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_2020_500 pct_fs_risk_2050_100 pct_fs_risk_2020_10
##
## After excluding the collinear variables, the linear correlation coefficients ranges between:
## min correlation ( o3 ~ EP_NOHSDP ):  0.0004107572
## max correlation ( Data_Value_CSMOKING ~ EP_NOHSDP ):  0.7700961
##
## ---------- VIFs of the remained variables --------
##                 Variables      VIF
## 1   avg_risk_fsf_2020_100 4.578968
## 2     avg_risk_score_sfha 2.643468
## 3  avg_risk_score_no_sfha 7.190601
## 4        pct_floodfactor2 1.549575
## 5        pct_floodfactor3 2.399075
## 6        pct_floodfactor4 3.215475
## 7        pct_floodfactor5 3.064580
## 8        pct_floodfactor6 3.843678
## 9        pct_floodfactor7 1.326846
## 10       pct_floodfactor8 2.023226
## 11       pct_floodfactor9 3.520129
## 12      pct_floodfactor10 2.964819
## 13                 EP_POV 3.974175
## 14               EP_UNEMP 1.812399
## 15                 EP_PCI 3.089545
## 16              EP_NOHSDP 4.880556
## 17               EP_AGE65 3.181606
## 18               EP_AGE17 3.367341
## 19              EP_DISABL 2.844562
## 20              EP_SNGPNT 2.803734
## 21              EP_MINRTY 3.518020
## 22              EP_LIMENG 2.884197
## 23               EP_MUNIT 2.023770
## 24              EP_MOBILE 2.063307
## 25               EP_CROWD 1.932280
## 26               EP_NOVEH 2.613338
## 27              EP_GROUPQ 1.516529
## 28             EP_UNINSUR 2.581742
## 29                     co 2.411719
## 30                    no2 5.058520
## 31                     o3 4.187400
## 32                   pm10 2.022643
## 33                   pm25 4.143421
## 34                    so2 1.589039
## 35    Data_Value_CSMOKING 7.493337
```

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")
```