Permeability Analysis

Jyotishka D.

December 6, 2018

Contents

1 Reading and cleaning data **1**

2 Correlation between selected variables **2**

3 Exploratory Analysis **2**
3.1 PCA 2
3.2 Principal Component with only a few selected variables 5
3.3 Principal Component with all variables 6

4 Factor Analysis **6**

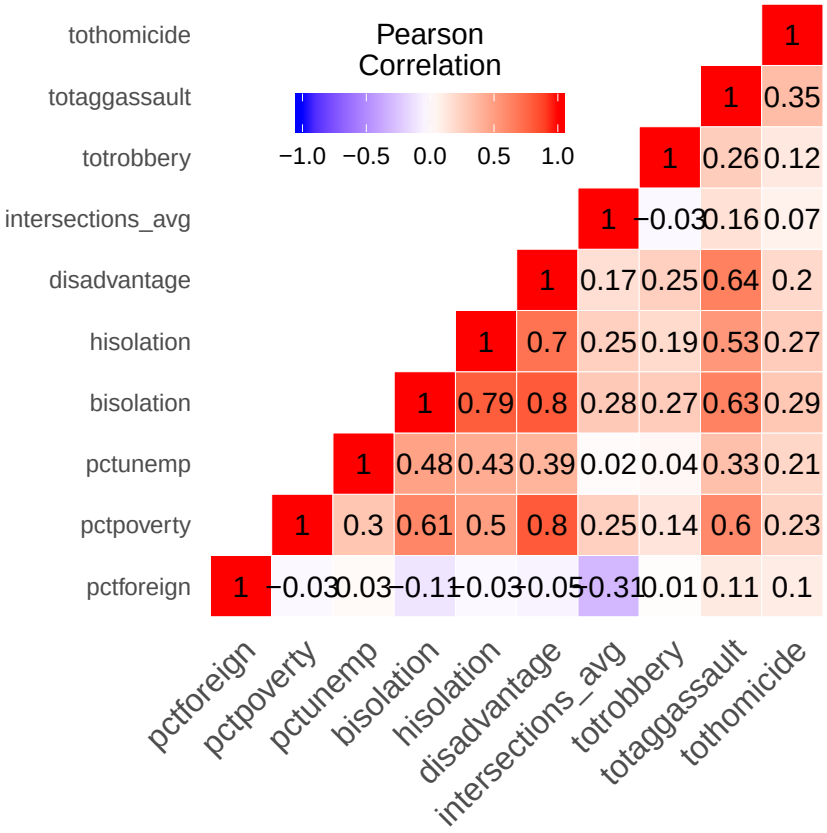
5 Model selection: Poisson regression with Elastic Net Penalty **7**

1 Reading and cleaning data

```
setwd("C:/Users/jd033/OneDrive/Documents/R/social")
library(readstata13)
dat <- read.dta13("Final Data - 155 Block Groups - 10-22-18.dta")
nm <- colnames(dat)
sel.dat <- dat[ ,!(nm %in% c("geoid1","geoid2","tract" ,"blockgroup","FID_1","geoid"))]
(colnames(sel.dat))
```

```
##  [1] "name"              "tpop"              "wpop"
##  [4] "bpop"              "apop"              "aipop"
##  [7] "hpop"              "opop"              "foreignpop"
## [10] "pctforeign"        "pctpoverty"        "pctunemp"
## [13] "pctnohsgrd"        "pctfhfamkids"      "pctmobile"
## [16] "wbdiss"            "whdiss"            "bhdiss"
## [19] "bisolation"        "hisolation"        "hisolation2"
## [22] "pctsnap"           "pctassist"         "pctlowskill"
## [25] "mhhincome"         "males"             "females"
## [28] "males_u18"         "females_u18"       "males_o18"
## [31] "females_o18"       "pctweakenglish"    "disadvantage"
## [34] "mhhincome2"        "drivedist"         "drivetime"
## [37] "walktime"          "transittime"       "intersections"
## [40] "intersections_avg" "walk_restrict"     "notransit"
## [43] "neighbor_count"    "totcrime"          "totrobbery"
## [46] "totaggassault"     "tothomicide"
```

2 Correlation between selected variables



3 Exploratory Analysis

3.1 PCA

Principal component with a few selected variables on the correlation matrix shown above.

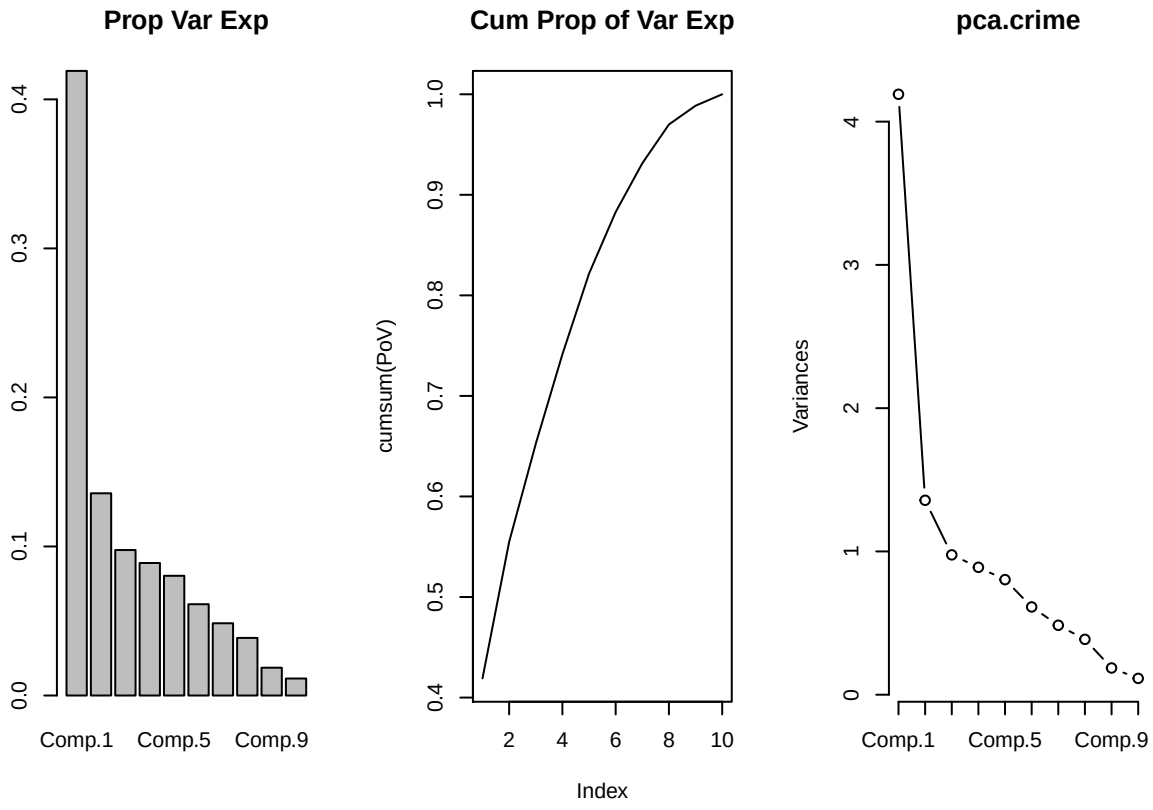
```
pca.crime <- princomp(scale(sel.dat2, scale = TRUE, center = TRUE), cor=TRUE)
summary(pca.crime) # print variance accounted for
```

```
## Importance of components:
##                            Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     2.0471789 1.1648431 0.98795676 0.94294880
## Proportion of Variance 0.4190941 0.1356859 0.09760586 0.08891524
## Cumulative Proportion  0.4190941 0.5547801 0.65238594 0.74130118
##                            Comp.5     Comp.6     Comp.7     Comp.8
## Standard deviation     0.89620789 0.78261699 0.69617282 0.62147671
## Proportion of Variance 0.08031886 0.06124893 0.04846566 0.03862333
## Cumulative Proportion  0.82162004 0.88286898 0.93133463 0.96995796
##                            Comp.9    Comp.10
## Standard deviation     0.43187908 0.33749195
## Proportion of Variance 0.01865195 0.01139008
## Cumulative Proportion  0.98860992 1.00000000
```

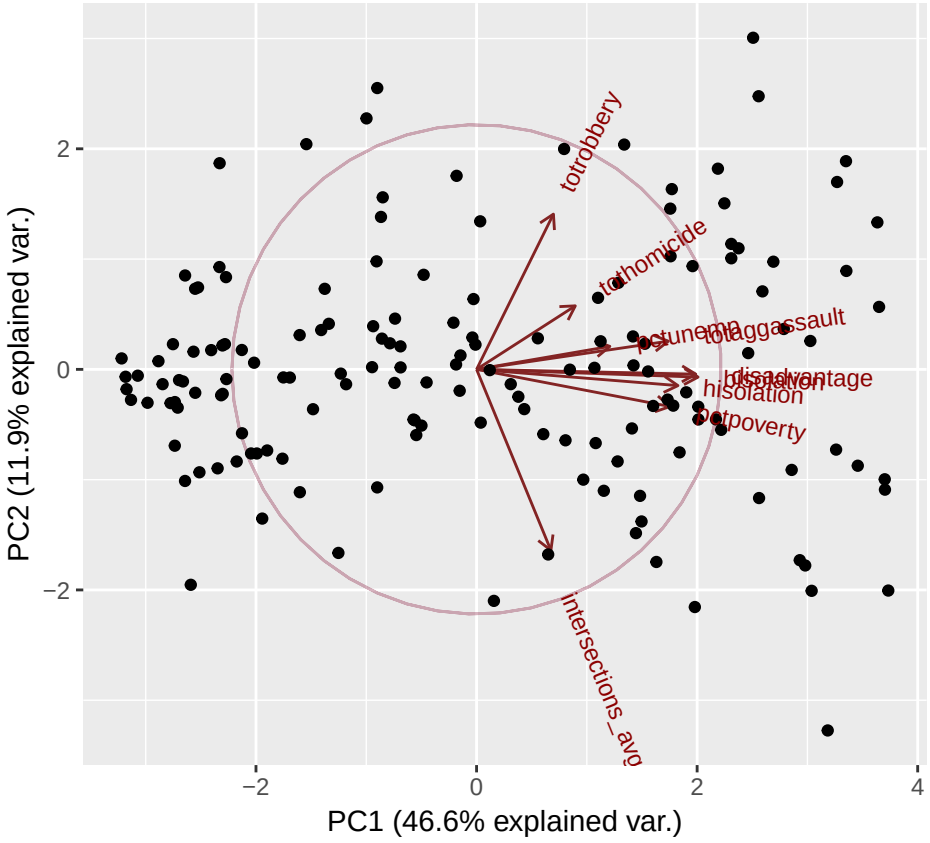
```
loadings(pca.crime, cutoff = 0)
```

```
##
## Loadings:
##                   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## pctforeign                0.689  0.167         0.425  0.521         0.184
## pctpoverty        -0.386                      0.407 -0.337  0.348  0.404
## pctunemp          -0.266  0.141  0.432 -0.221 -0.620  0.188  0.494
## bisolation        -0.443                     -0.117        -0.273
## hisolation        -0.402                     -0.122  0.253 -0.635
## disadvantage      -0.438               -0.238  0.173 -0.201         0.217
## intersections_avg -0.150 -0.608         0.373  0.206  0.587  0.239
## totrobbery        -0.153  0.197 -0.866        -0.272  0.208  0.197  0.149
## totaggassault     -0.382  0.158         0.105  0.237         0.226 -0.833
## tothomicide       -0.196  0.250  0.143  0.848 -0.195 -0.281         0.174
##                   Comp.9 Comp.10
## pctforeign         0.103
## pctpoverty        -0.281  0.441
## pctunemp
## bisolation         0.707  0.445
## hisolation        -0.585
## disadvantage       0.211 -0.764
## intersections_avg        -0.115
## totrobbery
## totaggassault
## tothomicide
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8
##                Comp.9 Comp.10
## SS loadings       1.0     1.0
## Proportion Var    0.1     0.1
## Cumulative Var    0.9     1.0
```

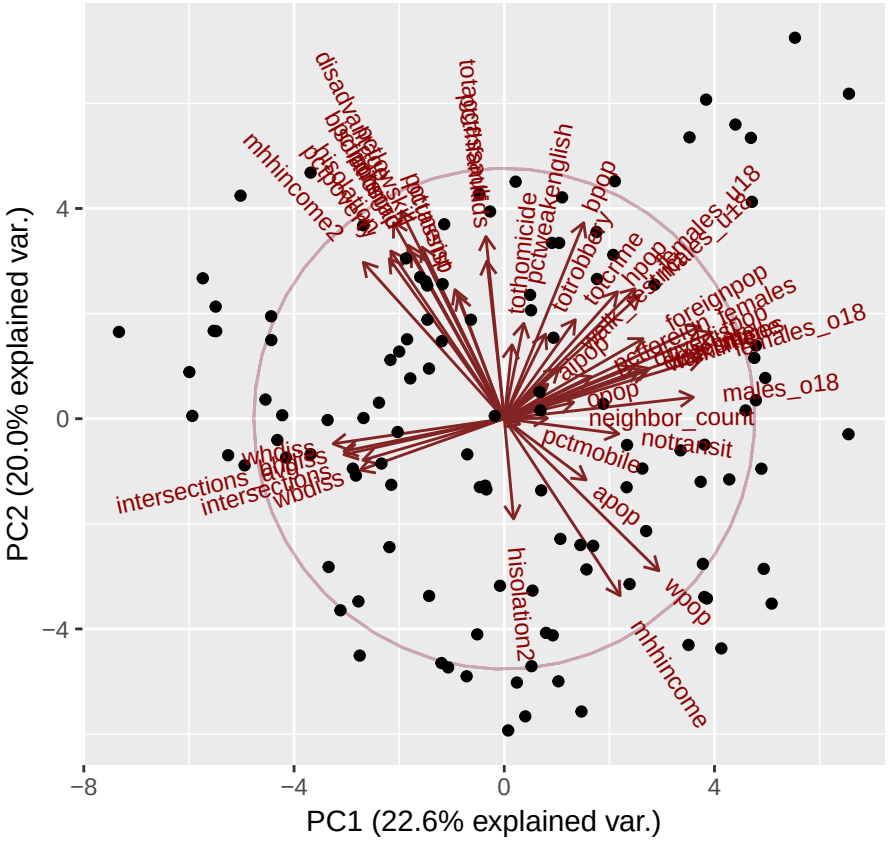
```
PoV <- pca.crime$sdev^2/sum(pca.crime$sdev^2)
par(mfrow=c(1,3))
barplot(PoV, main="Prop Var Exp")
plot(cumsum(PoV),type="l",main="Cum Prop of Var Exp")
screeplot(pca.crime, type="line")
```


Prop Var Exp
0.0
0.1
0.2
0.3
0.4
Comp.1
Comp.5
Comp.9
Cum Prop of Var Exp
cumsum(PoV)
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2
4
6
8
10
Index
pca.crime
Variances
0
1
2
3
4
Comp.1
Comp.5
Comp.9

3.2 Principal Component with only a few selected variables



3.3 Principal Component with all variables



4 Factor Analysis

Maximum Likelihood Factor Analysis entering raw data and extracting 3 factors, with varimax rotation.

```
##
## Call:
## factanal(x = sel.dat2, factors = 3, rotation = "varimax")
##
## Uniquenesses:
##        pctforeign        pctpoverty          pctunemp        bisolation
##              0.55              0.00              0.74              0.08
##        hisolation      disadvantage intersections_avg        totrobbery
##              0.32              0.18              0.74              0.90
##     totaggassault       tothomicide
##              0.46              0.87
##
## Loadings:
##                   Factor1 Factor2 Factor3
## bisolation         0.87
## hisolation         0.77
## disadvantage       0.66    0.60
## totaggassault      0.59    0.44
```

```
## pctpoverty          0.34    0.93
## pctforeign                          -0.66
## pctunemp            0.49
## intersections_avg                    0.48
## totrobbery          0.30
## tothomicide         0.31
##
##                Factor1 Factor2 Factor3
## SS loadings       2.70    1.61    0.85
## Proportion Var    0.27    0.16    0.09
## Cumulative Var    0.27    0.43    0.52
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 30.36 on 18 degrees of freedom.
## The p-value is 0.0341
```

5 Model selection: Poisson regression with Elastic Net Penalty

Poisson regression is used to model count data under the assumption of Poisson error, or otherwise nonnegative data where the mean and variance are proportional. We want to perform a variable selection, so we fit a penalized (or regularized) regression model with an Elastic Net ℓ_1 penalty. We optimize the penalized log-lielihood:

$$l(\beta \mid X, y) = \sum_{i=1}^{N} (y_i(\beta_0 + \beta' x_i) - e^{\beta_0 + \beta' x_i}) \min_{\beta_0, \beta} \frac{1}{N} l(\beta \mid X, y) + \lambda\{(1-\alpha) \sum_{i=1}^{N} \beta_i^2/2 + \alpha \sum_{i=1}^{N} |\beta_i|\}$$

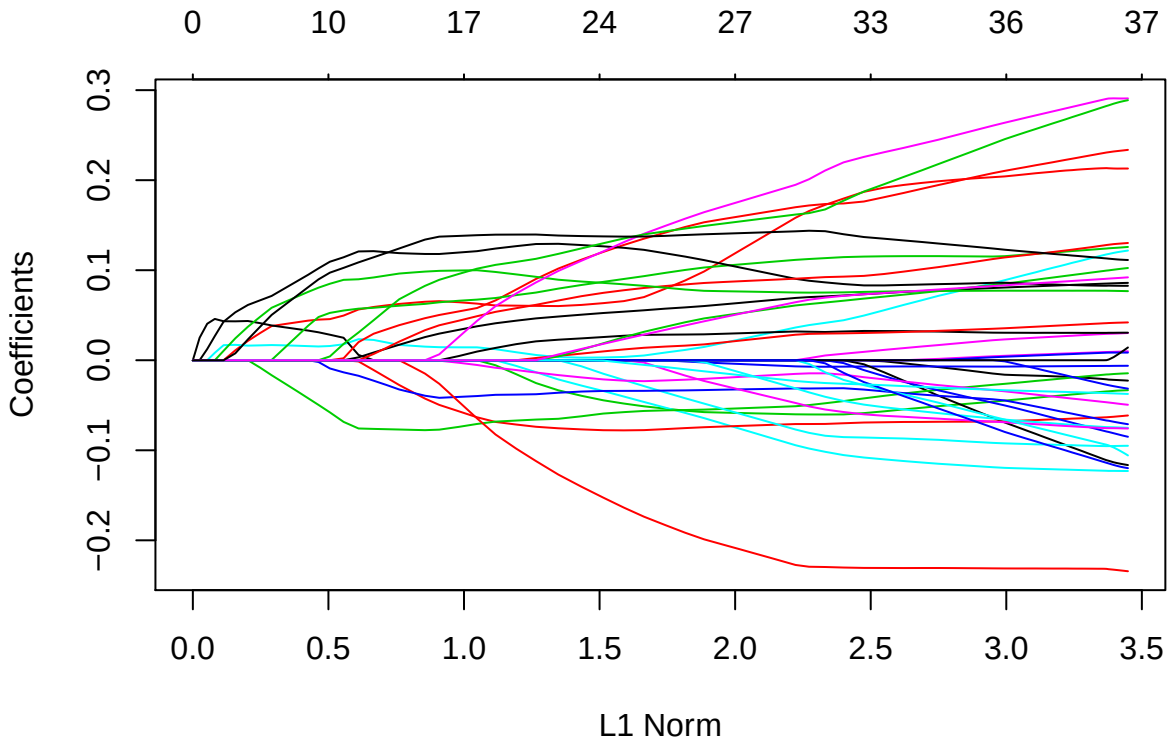
The idea behind penalized regression is that the extra penalty term will put a constraint on the parameter vector β, that will penalize for large coefficients, and the optimized estimate of β will be sparse, with many of the coefficients shrunk to zero.

This is helpful because it performs a model selection for us and makes the model more interpretable with fewer coefficients.

We fit the penalized Poisson regression with y being the `totcrime` and X being the set of all predictor variables, i.e. all columns minus `name` and the crime counts.

```
library(glmnet)
x = as.matrix(sel.dat[,c(2:43)])
x = scale(x, center = TRUE, scale = TRUE)
y = sel.dat[,44] # y is totcrime

fit = glmnet(x, y, family = "poisson")
plot(fit)
```




```
# coef(fit, s = 1)
cvfit = cv.glmnet(x, y, family = "poisson")
# opt.lam = c(cvfit$lambda.min, cvfit$lambda.1se)
(beta.sel <- coef(cvfit, s = cvfit$lambda.min))
```

```
## 43 x 1 sparse Matrix of class "dgCMatrix"
##                                1
## (Intercept)         4.819851140
## tpop                .
## wpop                .
## bpop                0.063806934
## apop                .
## aipop               .
## hpop                0.016025808
## opop                .
## foreignpop          .
## pctforeign          0.043013779
## pctpoverty          0.061662297
## pctunemp            .
## pctnohsgrd          .
## pctfhfamkids        .
## pctmobile           0.021644207
## wbdiss             -0.035341355
## whdiss              .
## bhdiss             -0.077312665
## bisolation          .
```

```
## hisolation         .
## hisolation2        .
## pctsnap            .
## pctassist          .
## pctlowskill        .
## mhhincome         -0.036064203
## males              .
## females            .
## males_u18          .
## females_u18        .
## males_o18          0.118464725
## females_o18        0.027291114
## pctweakenglish     0.097475996
## disadvantage       .
## mhhincome2         .
## drivedist          .
## drivetime          .
## walktime           .
## transittime       -0.007580568
## intersections      0.075590277
## intersections_avg  .
## walk_restrict      .
## notransit          .
## neighbor_count     0.129261896
```