Lab4_CaseyMicheline_MamrothAndrew_ArunimaKayath_Draf

Andrew Mamroth

August 13, 2017

Introduction

Exploratory Analysis

Building a Model

To build the model, we use a backwards approach. We first build a model that includes all the data we are given then remove the data with the least explanatory power. Following that, we then explain why the information for the variables removed is already incorporated into the model and thus why it is excluded from the final model.

```
summary(lm(crime$crmrte~crime$prbarr+crime$prbconv+crime$prbpris+crime$avgsen
           +crime$polpc+crime$density+crime$taxpc+crime$west+crime$central
           +crime$urban+crime$pctmin80+crime$wcon+crime$wtuc+crime$wtrd
           +crime$wfir+crime$wser+crime$wmfg+crime$wfed+crime$wsta+crime$wloc
           +crime$mix+crime$pctymle))
```

```
##
## Call:
## lm(formula = crime$crmrte ~ crime$prbarr + crime$prbconv + crime$prbpris +
##     crime$avgsen + crime$polpc + crime$density + crime$taxpc +
##     crime$west + crime$central + crime$urban + crime$pctmin80 +
##     crime$wcon + crime$wtuc + crime$wtrd + crime$wfir + crime$wser +
##     crime$wmfg + crime$wfed + crime$wsta + crime$wloc + crime$mix +
##     crime$pctymle)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0168836 -0.0039309 -0.0004161  0.0046227  0.0228050
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.333e-02  1.972e-02   0.676 0.501164
## crime$prbarr   -5.135e-02  9.919e-03  -5.177 2.24e-06 ***
## crime$prbconv  -1.854e-02  3.770e-03  -4.917 5.97e-06 ***
## crime$prbpris   4.159e-03  1.209e-02   0.344 0.731917
## crime$avgsen   -3.958e-04  4.241e-04  -0.933 0.354003
## crime$polpc     6.918e+00  1.546e+00   4.476 3.03e-05 ***
## crime$density   5.156e-03  1.400e-03   3.682 0.000464 ***
## crime$taxpc     1.676e-04  9.530e-05   1.759 0.083168 .
## crime$west     -2.416e-03  4.190e-03  -0.577 0.566193
## crime$central  -4.163e-03  2.869e-03  -1.451 0.151468
## crime$urban     5.814e-04  6.382e-03   0.091 0.927681
## crime$pctmin80  3.277e-04  9.886e-05   3.315 0.001484 **
## crime$wcon      2.406e-05  2.794e-05   0.861 0.392189
## crime$wtuc      5.257e-06  1.511e-05   0.348 0.729007
## crime$wtrd      2.896e-05  4.641e-05   0.624 0.534745
## crime$wfir     -3.482e-05  2.749e-05  -1.267 0.209657
## crime$wser     -1.887e-06  5.678e-06  -0.332 0.740741
## crime$wmfg     -8.792e-06  1.435e-05  -0.613 0.542111
## crime$wfed      2.981e-05  2.562e-05   1.164 0.248655
## crime$wsta     -2.326e-05  2.597e-05  -0.895 0.373764
## crime$wloc      1.337e-05  4.897e-05   0.273 0.785627
## crime$mix      -1.936e-02  1.472e-02  -1.315 0.192895
## crime$pctymle   1.035e-01  4.522e-02   2.288 0.025298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008317 on 67 degrees of freedom
## Multiple R-squared:  0.854,  Adjusted R-squared:  0.8061
## F-statistic: 17.81 on 22 and 67 DF,  p-value: < 2.2e-16
```

From here we trim the variables with the least explanatory power, but also it should be noted that some variables simply have very little correlation with our dependant variable crime rate, such as average sentence and probability of prison.

```
summary(lm(crime$crmrte~crime$density+crime$prbarr+crime$prbconv+
             crime$polpc+crime$pctmin80))$adj.r.squared
```

```
## [1] 0.7929876
```

```
dat_1<-data.frame(crime$crmrte,crime$avgsen,crime$prbpris)
cor(dat_1)
```

```
##               crime.crmrte crime.avgsen crime.prbpris
## crime.crmrte    1.00000000   0.01979653    0.04799540
## crime.avgsen    0.01979653   1.00000000   -0.09468083
## crime.prbpris   0.04799540  -0.09468083    1.00000000
```

From here, we see that these 5 variables contain almost all of the information of the other variables. Now that we have a model, we need to understand why these 5 variables cover all of the information we need for the model.

Density seems to be the strongest predictor of crime rate in the data. We include it first but it should be noted that if the urban flag is used in lew of density the model loses very little explanatory power because the two are highly correlated so little information is added by including it, and since density is more highly correlated with our dependant variable we choose to use it over the urban flag.

```
summary(lm(crime$crmrte~crime$urban+crime$prbarr+crime$prbconv+
             crime$polpc+crime$pctmin80))$adj.r.squared
```

```
## [1] 0.7302545
```

```
dat1<-data.frame(crime$crmrte,crime$density,crime$urban)
cor(dat1)
```

```
##               crime.crmrte crime.density crime.urban
## crime.crmrte     1.0000000     0.7277783   0.6150631
## crime.density    0.7277783     1.0000000   0.8206825
## crime.urban      0.6150631     0.8206825   1.0000000
```

Next we turn to wages. Even alone they seem to have little predictive power. It may be that case that what we really want to measure is not wages but unemployment as it may be that case that even if one doesn't have much money, they are at least employed and therefore will commit less crimes.

```r
summary(lm(crime$crmrte~crime$wcon+crime$wtuc+crime$wtrd+crime$wfir
           +crime$wser+crime$wmfg+crime$wfed+crime$wsta+crime$wloc))
```

```
##
## Call:
## lm(formula = crime$crmrte ~ crime$wcon + crime$wtuc + crime$wtrd +
##     crime$wfir + crime$wser + crime$wmfg + crime$wfed + crime$wsta +
##     crime$wloc)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.035348 -0.009720 -0.003703  0.006302  0.052214
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.898e-02  2.390e-02  -2.887  0.00501 **
## crime$wcon   6.737e-05  4.800e-05   1.404  0.16431
## crime$wtuc  -8.665e-07  2.747e-05  -0.032  0.97492
## crime$wtrd   1.245e-04  8.289e-05   1.501  0.13718
## crime$wfir  -6.460e-05  5.016e-05  -1.288  0.20150
## crime$wser  -5.261e-06  8.428e-06  -0.624  0.53424
## crime$wmfg   3.333e-05  2.573e-05   1.295  0.19889
## crime$wfed   7.975e-05  4.379e-05   1.821  0.07230 .
## crime$wsta   8.239e-05  4.497e-05   1.832  0.07062 .
## crime$wloc   1.162e-05  8.493e-05   0.137  0.89156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01636 on 80 degrees of freedom
## Multiple R-squared:  0.3253, Adjusted R-squared:  0.2494
## F-statistic: 4.285 on 9 and 80 DF,  p-value: 0.0001451
```

Verify Assumptions

Conclusions