Section1_Lab4_MichelleCutler_NishantVelagapudi_RangaMuva

Nishant Velagapudi, Ranga Muvararirwa, Michelle Cutler

December 8, 2017

Introduction

Our firm was hired to explore data from 1987 on crime rates from 90 counties across North Carolina and help a political campaign to understand the determinants of crime – as well as to generate policy suggestions that might be applicable to the local government.

Before proceeding, we had to decide whether to start with a parsimonious or a fully specified model. On the one hand, intuition suggested that we start by testing a series of pre-formulated hypotheses against the data. For example: evaluate strengths of relationships between economic variables and crime rates: using the findings to decide what should be included in our model. However, we were concerned that starting with a set of pre-formulated hypothesis could limit our model specification.Ultimately, we decided to start with a fully specified model and let the EDA guide which variables were irrelevant.

Exploratory Analysis

```
summary(Data)
```

```
##        X             county           year        crmrte
##  Min.   : 1.00   Min.   :  1.0   Min.   :87   Min.   :0.005533
##  1st Qu.:23.25   1st Qu.: 51.5   1st Qu.:87   1st Qu.:0.020604
##  Median :45.50   Median :103.0   Median :87   Median :0.030002
##  Mean   :45.50   Mean   :100.6   Mean   :87   Mean   :0.033510
##  3rd Qu.:67.75   3rd Qu.:150.5   3rd Qu.:87   3rd Qu.:0.040249
##  Max.   :90.00   Max.   :197.0   Max.   :87   Max.   :0.098966
##      prbarr           prbconv           prbpris           avgsen
##  Min.   :0.09277   Min.   :0.06838   Min.   :0.1500   Min.   : 5.380
##  1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.:0.3642   1st Qu.: 7.375
##  Median :0.27146   Median :0.45170   Median :0.4222   Median : 9.110
```

```
##  Mean   :0.29524   Mean   :0.55086   Mean   :0.4106   Mean   : 9.689
##  3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:0.4576   3rd Qu.:11.465
##  Max.   :1.09091   Max.   :2.12121   Max.   :0.6000   Max.   :20.700
##      polpc              density           taxpc             west
##  Min.   :0.0007459   Min.   :0.2034   Min.   : 25.69   Min.   :0.0000
##  1st Qu.:0.0012378   1st Qu.:0.5472   1st Qu.: 30.73   1st Qu.:0.0000
##  Median :0.0014897   Median :0.9792   Median : 34.92   Median :0.0000
##  Mean   :0.0017080   Mean   :1.4379   Mean   : 38.16   Mean   :0.2333
##  3rd Qu.:0.0018856   3rd Qu.:1.5693   3rd Qu.: 41.01   3rd Qu.:0.0000
##  Max.   :0.0090543   Max.   :8.8277   Max.   :119.76   Max.   :1.0000
##     central           urban            pctmin80           wcon
##  Min.   :0.0000   Min.   :0.00000   Min.   : 1.284   Min.   :193.6
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:10.024   1st Qu.:250.8
##  Median :0.0000   Median :0.00000   Median :24.852   Median :281.2
##  Mean   :0.3778   Mean   :0.08889   Mean   :25.713   Mean   :285.4
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0
##  Max.   :1.0000   Max.   :1.00000   Max.   :64.348   Max.   :436.8
##       wtuc            wtrd            wfir            wser
##  Min.   :187.6   Min.   :154.2   Min.   :170.9   Min.   : 133.0
##  1st Qu.:374.3   1st Qu.:190.7   1st Qu.:285.6   1st Qu.: 229.3
##  Median :404.8   Median :203.0   Median :317.1   Median : 253.1
##  Mean   :410.9   Mean   :210.9   Mean   :321.6   Mean   : 275.3
##  3rd Qu.:440.7   3rd Qu.:224.3   3rd Qu.:342.6   3rd Qu.: 277.6
##  Max.   :613.2   Max.   :354.7   Max.   :509.5   Max.   :2177.1
##       wmfg            wfed            wsta            wloc
##  Min.   :157.4   Min.   :326.1   Min.   :258.3   Min.   :239.2
##  1st Qu.:288.6   1st Qu.:398.8   1st Qu.:329.3   1st Qu.:297.2
##  Median :321.1   Median :448.9   Median :358.4   Median :307.6
##  Mean   :336.0   Mean   :442.6   Mean   :357.7   Mean   :312.3
##  3rd Qu.:359.9   3rd Qu.:478.3   3rd Qu.:383.2   3rd Qu.:328.8
```

```
##  Max.   :646.9   Max.   :598.0   Max.   :499.6   Max.   :388.1
##       mix             pctymle
##  Min.   :0.01961   Min.   :0.06216
##  1st Qu.:0.08060   1st Qu.:0.07437
##  Median :0.10095   Median :0.07770
##  Mean   :0.12905   Mean   :0.08403
##  3rd Qu.:0.15206   3rd Qu.:0.08352
##  Max.   :0.46512   Max.   :0.24871
```

It appears that all variables are ordinal with a few exceptions. First is "Year", which appears to be "87" in all cases.Second are "west", "central", and "urban". Each of these appear to be indicator variables. We will have to keep in mind sample problems: as our sample is fairly small.

It appears that probability of conviction is a conditional probability of the probability of arrest. Probability of prison time is subsequently conditional on the probability of conviction.

```
par(mar=c(1,1,1,1))
par(mfrow=c(4,5))


for (col in names(Data)){
 if(sapply(Data,class)[col] == "numeric"){
   hist(Data[!is.null(Data[col]), col], main=paste("",col), xlab=NULL, ylab=NULL,
        breaks=20, axes = FALSE, cex.lab=0.75, cex.axis=0.75, cex.main=0.75, cex.sub=0.75)
   axis(1, labels = FALSE)
   axis(2, labels = FALSE)
   print(paste(col, " has ", length(Data[is.na(Data[col]), col]), " null values"))
 } else {
   print(summary(Data[col]))
 }
}
```

```
##        X
```

```
##  Min.   : 1.00
##  1st Qu.:23.25
##  Median :45.50
##  Mean   :45.50
##  3rd Qu.:67.75
##  Max.   :90.00
##      county
##  Min.   :  1.0
##  1st Qu.: 51.5
##  Median :103.0
##  Mean   :100.6
##  3rd Qu.:150.5
##  Max.   :197.0
##       year
##  Min.   :87
##  1st Qu.:87
##  Median :87
##  Mean   :87
##  3rd Qu.:87
##  Max.   :87
## [1] "crmrte  has  0  null values"
## [1] "prbarr  has  0  null values"
## [1] "prbconv  has  0  null values"
## [1] "prbpris  has  0  null values"
## [1] "avgsen  has  0  null values"
## [1] "polpc  has  0  null values"
## [1] "density  has  0  null values"
```

```
## [1] "taxpc  has  0  null values"
##       west
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2333
##  3rd Qu.:0.0000
##  Max.   :1.0000
##     central
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3778
##  3rd Qu.:1.0000
##  Max.   :1.0000
##      urban
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.08889
##  3rd Qu.:0.00000
##  Max.   :1.00000
## [1] "pctmin80  has  0  null values"
## [1] "wcon  has  0  null values"
## [1] "wtuc  has  0  null values"
## [1] "wtrd  has  0  null values"
## [1] "wfir  has  0  null values"
## [1] "wser  has  0  null values"
```

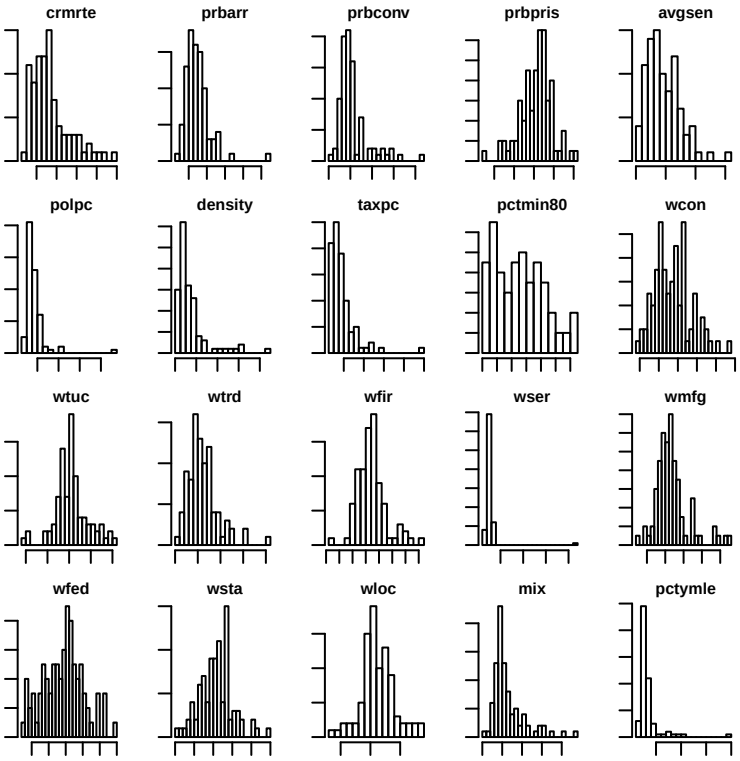
```
## [1] "wmfg  has  0  null values"
```

```
## [1] "wfed  has  0  null values"
```

```
## [1] "wsta  has  0  null values"
```

```
## [1] "wloc  has  0  null values"
```

```
## [1] "mix  has  0  null values"
```




```
## [1] "pctymle  has  0  null values"
```

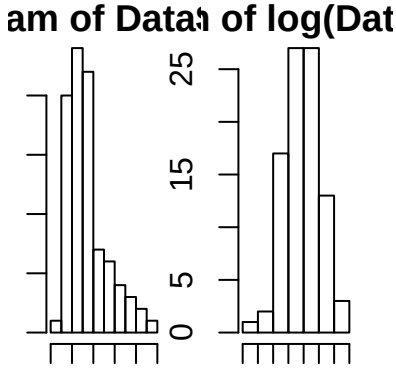
At a glance, variables pctmin80, crmrte, prbconv, mix, and pctymle appear non-normal. We can try transforms of each of these variables: though the near uniformity of the pctmin80 distribution suggests that normality may not be achievable. Of note is the fact that we are considering transforming our variable of interest (crmrte).

It is interesting to note that prbarr and prbconv are positively skewed, but after conviction, the distribution of prbpris appears to normalize.

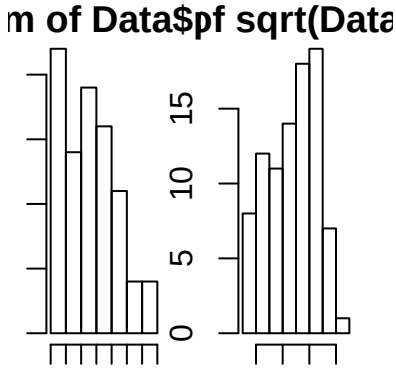
```
par(mar=c(1,1,1,1))
```

```
par(mfrow=c(1,2))
```

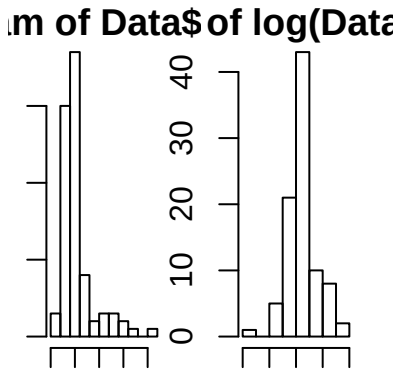
```
hist(Data$crmrte)


hist(log(Data$crmrte))
```



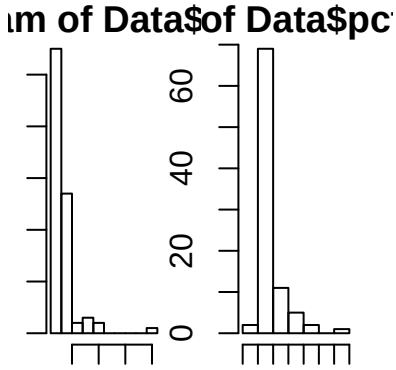

```
par(mfrow=c(1,2))

hist(Data$pctmin80)


hist(sqrt(Data$pctmin80))
```



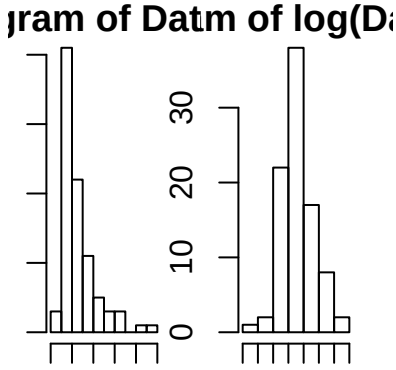

```
par(mfrow=c(1,2))

hist(Data$prbconv)


hist(log(Data$prbconv))
```




```
par(mfrow=c(1,2))

hist(Data$pctymle)


hist(Data$pctymle^(1/9))
```




```
par(mfrow=c(1,2))

hist(Data$mix)


hist(log(Data$mix))
```



Each transformation appears to generate more normal behavior with the exception of pctymle. Transforming

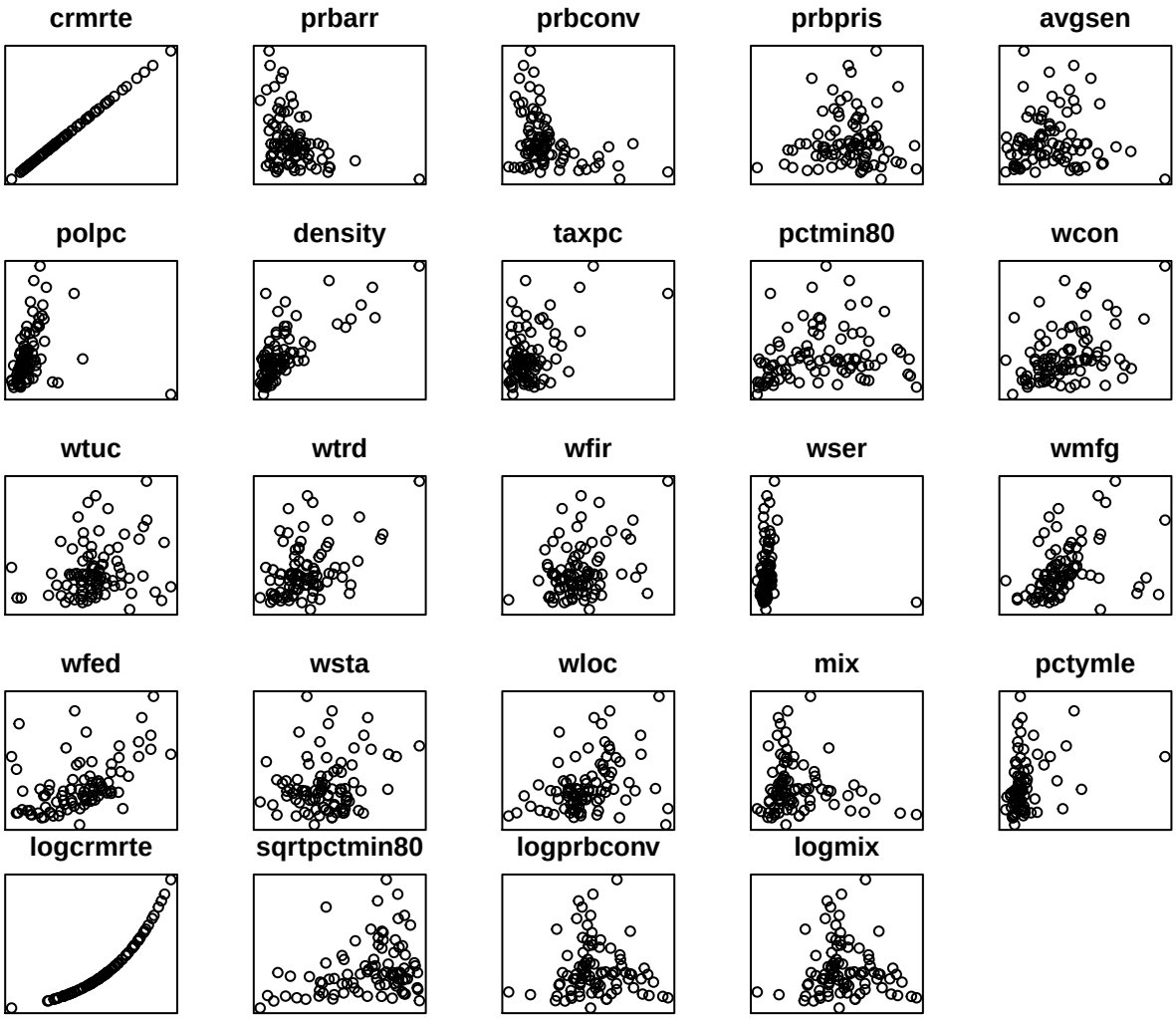
percentage male into a square root would also make for a difficult translation of effect: we elect to leave the variable as is.

Model Assumptions

Prior to even creating models, we will test the first two assumptions of linear regression:

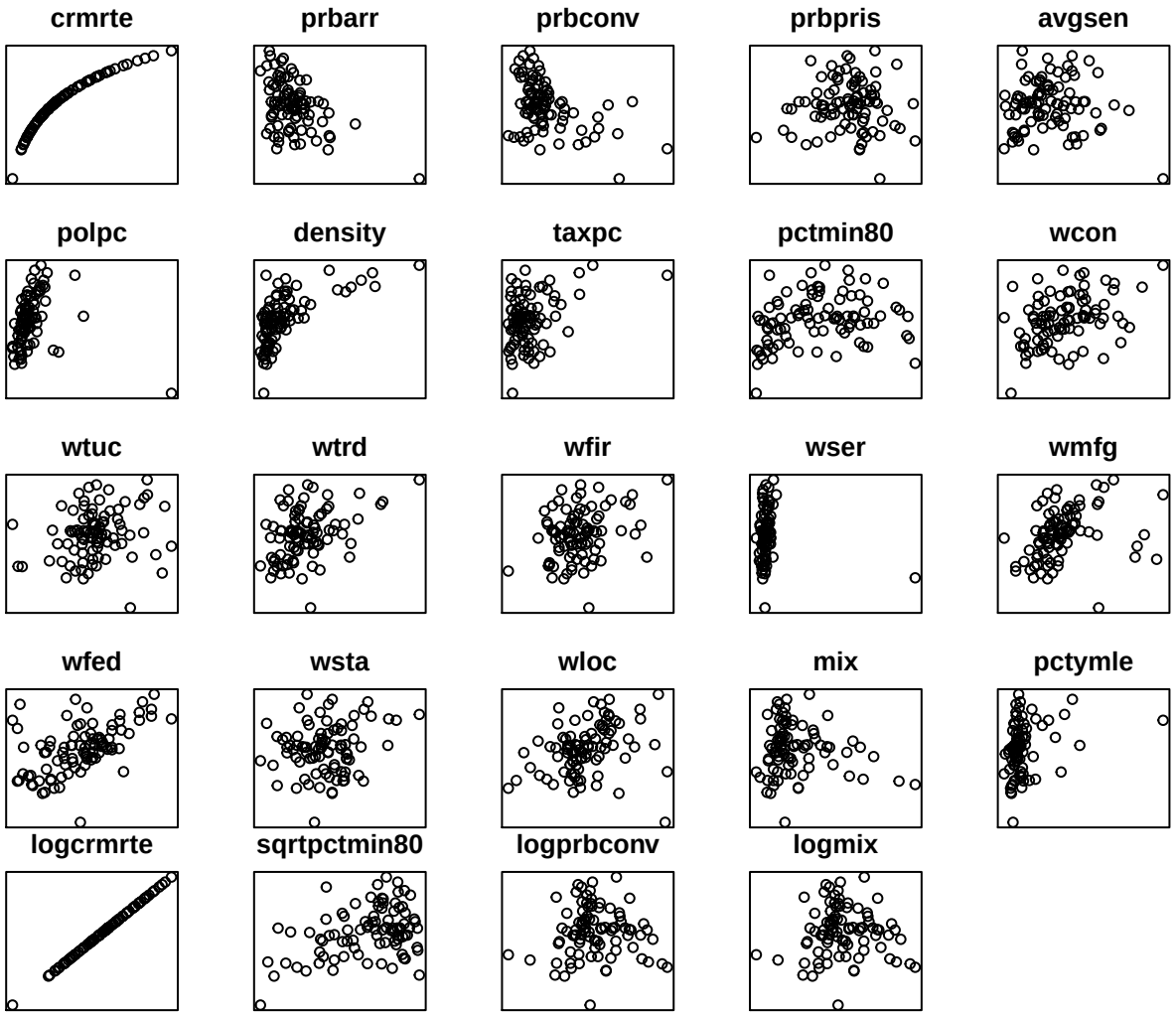
Assumption 1: Linear in Parameters

```
#transform variables
Data["logcrmrte"] <- log(Data$crmrte)


Data["sqrtpctmin80"] <- log(Data$pctmin80)
Data["logprbconv"] <- log(Data$mix)
Data["logmix"] <- log(Data$mix)


par(mfrow=c(4,5), mai=c(0.1,0.1,0.3,0.3))


for (col in names(Data[,-which(names(Data) %in% c("x","west","central","urban","X","county","year"))])){
 if(sapply(Data,class)[col] == "numeric")
 {
   plot(Data[!is.na(Data[col]) & !is.na(Data$crmrte), col], Data[!is.na(Data[col]) & !is.na(Data$crmrte)
 }
}
```



Here, we plot each of our variables against crmrte. We compare this to plots of each of our numeric variables against the logarithm of crmrte to continue to evaluate whether or not our output variable should be transformed.

```
par(mfrow=c(4,5), mai=c(0.1,0.1,0.3,0.3))


for (col in names(Data[,-which(names(Data) %in% c("x","west","central","urban","X","county","year"))])){
 if(sapply(Data,class)[col] == "numeric")
 {
   plot(Data[!is.na(Data[col]) & !is.na(Data$logcrmrte), col], Data[!is.na(Data[col]) & !is.na(Data$logc
 }
}
```

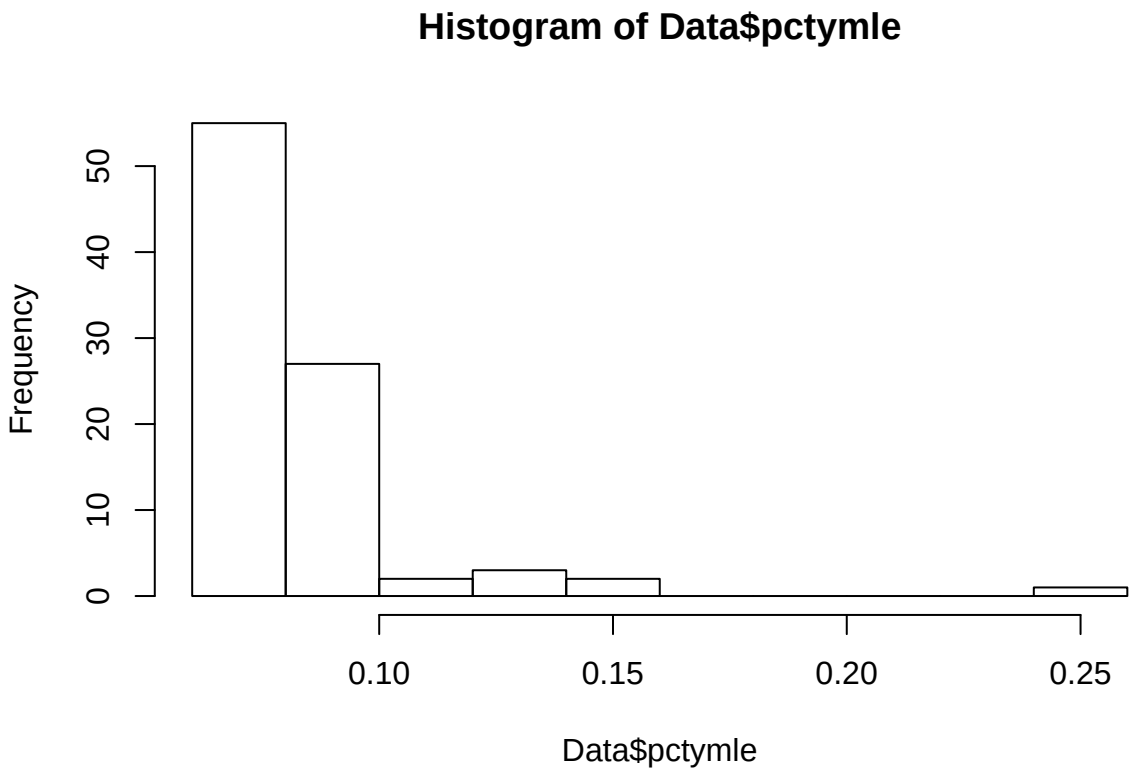


Assumption 2: Random Sampling

```
summary(Data[c("west","central","urban")])
```

```
##       west           central           urban
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.2333   Mean   :0.3778   Mean   :0.08889
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
```

```
hist(Data$pctymle)
```



We can see that ~37.8% of counties are "central", while 23.3% of counties are "west" and ~8.9% of counties are "urban". On the other hand, we can see that the distribution of percentage young male is highly positive skewed.

Assumption 3: No Perfect Multicollinearity

```
corMatrix <- cor(Data[,-which(names(Data) %in% c("x","west","central","urban","X","county","year"))])
VIFMatrix <- 1/(1-corMatrix^2)
```

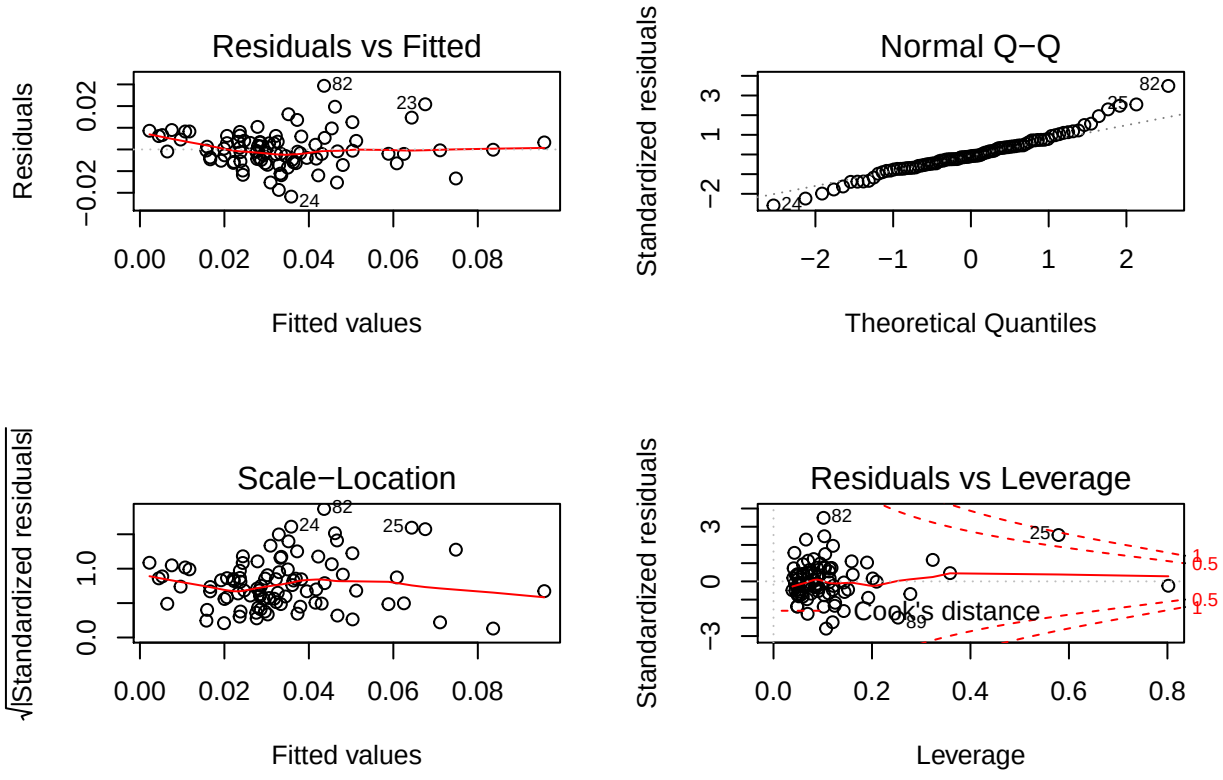
No VIF values are above 10, the most concerning correlation is between the logarithms of prbconv and mix. The prbconv variable is the probability of conviction, while mix is the mix of offenses. These two have a VIF of 8.8: since this is below 10, we will leave both variables in the model.

To test the rest of the assumptions, we will produce models and examine diagnostic plots/model properties.

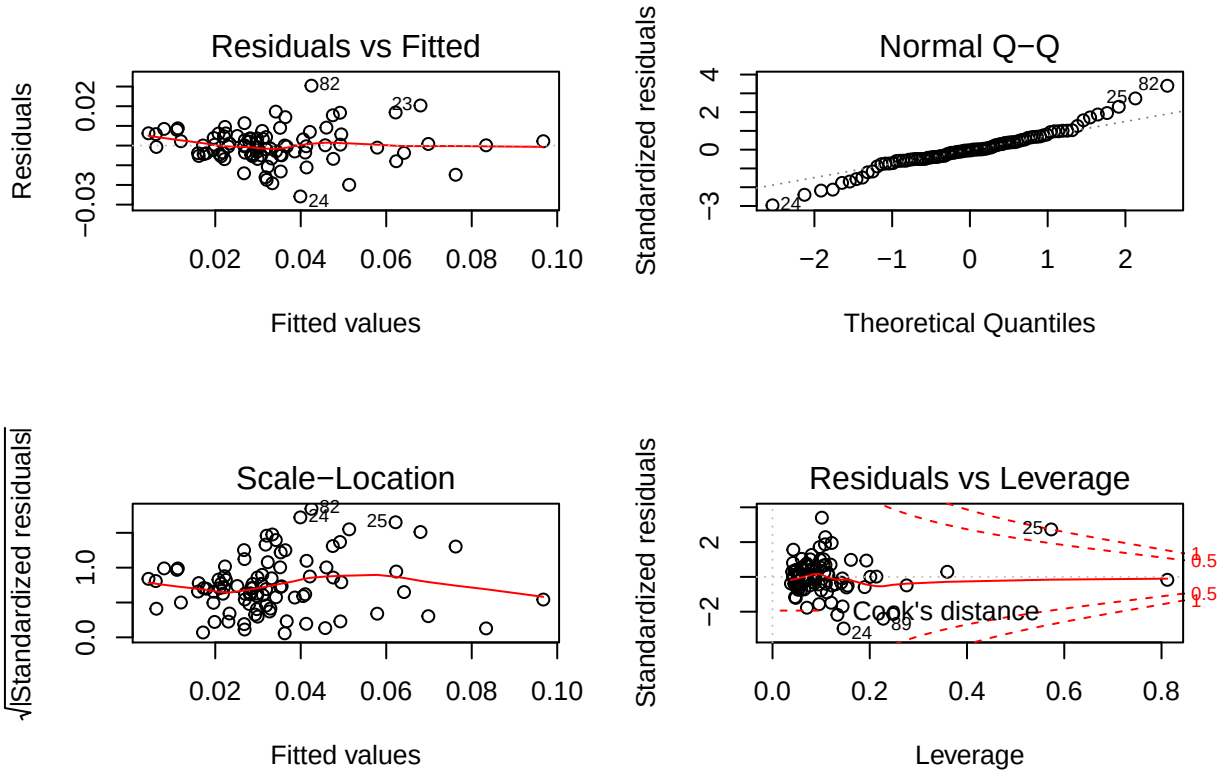
We will begin by creating three models, one with non-transformed variables and two with the variable transformed as shown above.

```r
#None transformed model
No_transform <- lm("crmrte ~ prbarr+prbconv+prbpris+avgsen+polpc+density+taxpc+west+central
                   +urban+pctmin80+wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc+mix+pctymle", data=Data)


#Applies transformations to crmrte, prbconv, pctmin80 and mix
transform <- lm("crmrte ~ prbarr+log(prbconv)+prbpris+avgsen+polpc+density+taxpc+west+central
                   +urban+sqrt(pctmin80)+wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc+log(mix)+pctymle",


#Removes the polpc
transform_no_polpc <- lm("crmrte ~ prbarr+log(prbconv)+prbpris+avgsen+density+taxpc+west+central
                   +urban+sqrt(pctmin80)+wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc+log(mix)+pctymle",
```

```r
par(mfrow=c(2,2))
plot(No_transform)
```




```r
par(mfrow=c(2,2))
plot(transform)
```




```
print(AIC(No_transform))
```

```
## [1] -583.5114
```

```
print(AIC(transform))
```

```
## [1] -572.5318
```

```
print(AIC(transform_no_polpc))
```

```
## [1] -557.7361
```

Examining the Akaike Information Criterion (AIC) of each model provides a metric that balances parsimony and fit by applying a penalty per included variable. We can see that the AIC of the regression on transformed variable is a lower absolute value: indicating a better fit.

Assumption 4: Zero Conditional Mean

Both models potentially violate the zero conditional mean assumption: though the model with the transformed variables appears to be a much more egregious violation (clear curvature in the residuals vs fitted plot).

Actions taken: We can see that the model with transformed variables is not appropriate due to this violation.

Although we appear to have violated zero-conditional-mean, we argue that given the fact that the sample size of we can apply OLS asymptotics. With the application of OLS asymptotics the critical argument becomes exogeneity – particularly if we want to explore a causal model.

Assumption 5: Homoskedasticity

Both models also appear to violate the assumption of homoskedasticity: the "width of band" in the residuals vs fitted plot and the non-level nature of the fitted line in the scale-location plot suggest that this assumption is violated in both cases, though the violation in the case of the transformed variable model seems worse as well.

Below we employ the Breusch-Pagan test to evaluate presence of of heteroskedasticity as follows:

```
#Test for presence of Heteroskedasticity
(bptest(transform))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  transform
## BP = 19.269, df = 9, p-value = 0.023
```

H_0 is that we have Homoskedasticity Given the large p-value, we fail to reject the null

Actions taken: We will use heteroskedasticity robust errors, as it is good practice.

Assumption 6: Normality of Errors

```
coeftest(transform, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
```

```
## (Intercept)   0.02811462  0.00959465  2.9302 0.0044127 **
## prbarr       -0.04850382  0.01312929 -3.6943 0.0004018 ***
## log(prbconv) -0.00904072  0.00347133 -2.6044 0.0109691 *
## prbpris       0.00833384  0.01364295  0.6109 0.5430281
## avgsen       -0.00062778  0.00039139 -1.6040 0.1126542
## polpc         5.90253517  1.99061282  2.9652 0.0039856 **
## density       0.00639248  0.00142500  4.4860 2.402e-05 ***
## taxpc         0.00008198  0.00023701  0.3459 0.7303369
## west         -0.01470509  0.00278537 -5.2794 1.087e-06 ***
## central      -0.00906657  0.00255571 -3.5476 0.0006537 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our coefficient tests, we can see that three variables are not statistically significant in the model: prbpris, avgsen, and taxpc. We can run an exclusion restriction to test that the coefficients on all three of these variables are not statistically significantly different than zero. We will do this test on the pre-transform test

Application of the Shapiro-Wilk test revealed to the pre-transform model resulted in a high p-value, suggesting that we fail to reject $H_0 : Normality of errors$ particularly given the sensitivity of this test to sample size.

```
#Appl the Shapiro-Wilk test of normality
shapiro.test(transform$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  transform$residuals
## W = 0.9697, p-value = 0.0337
```

```
restricted_transform <- lm("crmrte ~ prbarr+log(prbconv)+polpc+density+west+central+urban+sqrt(pctmin80)
                  +wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc+mix+pctymle", data=Data)
```

```
restricted_SSR <- sum((restricted_transform$residuals)^2)
noRestrict_SSR <- sum((No_transform$residuals)^2)

F_stat <- ((noRestrict_SSR - restricted_SSR) / noRestrict_SSR) * ((81-19-1)/3)
qf(.95,3,81-19-1)
```

```
## [1] 2.755481
```

```
coeftest(restricted_transform, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     0.0140976  0.0093503  1.5077 0.1355165
## prbarr         -0.0556610  0.0137768 -4.0402 0.0001208 ***
## log(prbconv)   -0.0106527  0.0032677 -3.2600 0.0016305 **
## polpc           6.2847051  2.2804024  2.7560 0.0072269 **
## density         0.0064695  0.0017275  3.7450 0.0003364 ***
## west           -0.0061384  0.0063999 -0.9591 0.3403411
## central        -0.0054406  0.0033860 -1.6068 0.1119871
## urban          -0.0013348  0.0078547 -0.1699 0.8654798
## sqrt(pctmin80)  0.0024032  0.0013284  1.8091 0.0741504 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have an F_stat of 1.43, while the critical value for α is 2.76. We fail to reject the null hypothesis that the coefficients of all of the restricted variables are non-zero. Therefore, we will conclude that these variables together have no relationship to crime rate in our regression. In the restricted model, we can see that variables prbarr, prbconv, polpc, density, and pctmin80 are statistically significantly different from zero even with Heteroskedasticity robust standard errors.

```r
stargazer(restricted_transform, transform, transform_no_polpc, type="text")
```

```
##
## ========================================================================================
##                                          Dependent variable:
##                        -----------------------------------------------------------------
##                                                 NA
##                              (1)                  (2)                   (3)
## ----------------------------------------------------------------------------------------
## prbarr                    -0.056***            -0.049***             -0.029***
##                            (0.009)              (0.009)               (0.009)
##
## log(prbconv)              -0.011***            -0.009***             -0.007***
##                            (0.002)              (0.002)               (0.002)
##
## prbpris                                          0.008                 0.013
##                                                 (0.013)               (0.014)
##
## avgsen                                          -0.001                0.0001
##                                                 (0.0004)              (0.0004)
##
## polpc                      6.285***             5.903***
##                            (1.230)              (1.457)
##
## density                    0.006***             0.006***              0.007***
##                            (0.001)              (0.001)               (0.001)
##
## taxpc                                            0.0001               0.0002**
##                                                 (0.0001)              (0.0001)
##
```

```
## west                          -0.006              -0.015***             -0.013***
##                              (0.004)              (0.003)               (0.003)
##
## central                       -0.005**            -0.009***             -0.008***
##                              (0.003)              (0.002)               (0.003)
##
## urban                         -0.001
##                              (0.006)
##
## sqrt(pctmin80)                0.002**
##                              (0.001)
##
## Constant                      0.014**              0.028***              0.017**
##                              (0.007)              (0.008)               (0.008)
##
## ------------------------------------------------------------------------------------
## Observations                    90                   90                    90
## R2                             0.784                0.776                 0.729
## Adjusted R2                    0.762                0.750                 0.703
## Residual Std. Error     0.009 (df = 81)      0.009 (df = 80)       0.010 (df = 81)
## F Statistic          36.659*** (df = 8; 81) 30.712*** (df = 9; 80) 27.304*** (df = 8; 81)
## ====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

Interpreting the Results (Associative Model)

Restricted Model (restricted_transform):

crmrte = -0.014 - 0.056prbarr - 0.011log(prbconv) + 6.3polpc + 0.006density - 0.006west -0.005central -0.001urban + 0.002sqrt(pctmin80)

Transformed Model (transform):

crmrte = -0.028 - 0.049prbarr + 0.009log(logprbconv) + 0.008prbpris - 0.001avgsen + 5.9polpc + 0.006density + 0.001taxpc - 0.015west -0.009*central

Transformed Model no PolPC (transform_no_polpc):

crmrte = -0.017 - 0.029prbarr - 0.007log(prbconv) + 0.013prbpris - 0.0001avgsen + 0.007density + 0.0002taxpc - 0.013west -0.008central

Drawing Inferences (Associative)

These inferences are based an associative model ("polpc" predictor included) We select the 'restricted_transform_model' on account of higher AIC. This model presents a number of critical relationships, which are worth noting:

CrimeRate vs PolicePerCapita

The model presents a single dominant predictor for crime-rate: police per capita. There is a clear and strong correlation between police per capita (polpc) and crime rates. This relationship aligns with the intuition that areas with higher crime rates tend to invest in more in their police force than the contrary. More specifically, controlling for all other factors, counties tend to invest in an incremental police officer per capita for each ~6.3% increase in crimes per period (assumed annual here).

Critique I: Levels of policing are typically consistent with ..and have been historically affected as a response to... levels of crime.

More explicitly, crime rates and police force levels are linearly and (typically) causally related in the reverse direction (higher crime rates lead to higher levels of policing). In effect, crime is arguably endogenous to per capita police levels. In other words, the correlation of per capita police levels with crime rate does not mean that there is a causal model.

CrimeRate vs Location

Location appears to have a very small/marginal impact on crime-rates, suggesting that levels of crime are similar/consistent across the 90 counties.

Model sufficiency

None of the remaining predictors are sufficiently influential to compare/compete with polpc. For example, the next strongest predictor (probability of arrest) is ~100x smaller than polpc.

Critique II: This suggests that the model is likely underspecified.

In other words, the predictor variables do not contain enough information to explain (or to account for) the rates of crime. For example, the wage data could have provided explanatory power if levels of employment by type, wage and location were also made available in the data.

Policy Prescriptions based on Associated Model (Includes "polpc" predictor)

This represents our associative model. The rest of our analysis and recommendations will be based on the restricted_no_tranform.

Policy I: Match increases in crime rates with proportionate increases in police force:

If the campaign is being conducted during a period of rising crime rates, then campaign would be advised to promote a platform that proposes to increase investment in police force levels by 1 police officer per capita for each 6.3% increases in crime rates.

Policy II: Match increases in crime rates with proportionate increases in enforcement.

There is a weaker and negative correlation between probability of arrest and conviction with levels of crime. Simultaneously increasing arrest and conviction rates by a factor of 100 will have a similar effect (100*(0.062 + 0.020) ~ 8%) on reducing crime rates as would be required to increase police personal to levels consistent with the new crime rates.

Policy III: Policy measures should be statewide.

Given negligible differences in levels of crime by location... as well as the lack of geo-specific data on employment patterns (wages, distribution of employment by type and wage etc), it is not possible to make policy recommendations that are location specific.

Policy IV: Increases in police levels are the most significant and practical responses to increases in crime rate.

Additional Exploration: Towards a more causal model

Motivation

Reduce model complexity by reducing dimensionality of predictor variables in an intuitive manner.

Conjecture

Neither of these reductions in dimension would result in information loss because: Location information is preserved in another factor column Wage data in its current form is not meaningful without knowing the relative proportions of employment by wage-category for each of the counties/regions. Average wage is therefore arguably more meaningful..

Approach

Restructure the Data set by: Combining "central", "western" and "urban" into a single factor column called "location" Combining "wcon", "wtuc","wtrd",. . . ."wloc" into a single variable called "w_avg" Explore relationships between reduced set of predictors and response (crmrte)

Explore relationships between reduced set of predictors and response (crmrte)

Attribution & References

Here we replicate an analysis by Alexandra Chouldehova Objective is to visually explore the impact of

Observations

```
Data2 <- read.csv("crime.csv",sep=",")


Data2$location[Data2$west == 1] <-"west"

Data2$location[Data2$urban == 1] <-"urban"

Data2$location[Data2$central == 1] <- "central"

Data2$location[is.na(Data2$location)] <- "other"


Data2$location <- factor(Data2$location)


Data2$w_avg <- rowMeans(Data2[,16:24])


transform2_no_polpc <- lm("crmrte ~ prbarr+prbconv+prbpris+avgsen+density+taxpc+pctmin80+w_avg+mix+pcty
```

```
# Calculate location-specific intercepts

intercepts <- c(transform2_no_polpc$coefficients["(Intercept)"],

                transform2_no_polpc$coefficients["(Intercept)"] + transform2_no_polpc$coefficients["prb

                transform2_no_polpc$coefficients["(Intercept)"] + transform2_no_polpc$coefficients["prb

                transform2_no_polpc$coefficients["(Intercept)"] + transform2_no_polpc$coefficients["prb
```

```
lines.df <- data.frame(intercepts = intercepts,

                       slopes = rep(transform2_no_polpc$coefficients["polpc"], 4),

                       location = levels(Data2$location))
```

Drawing Inferences (Causal)

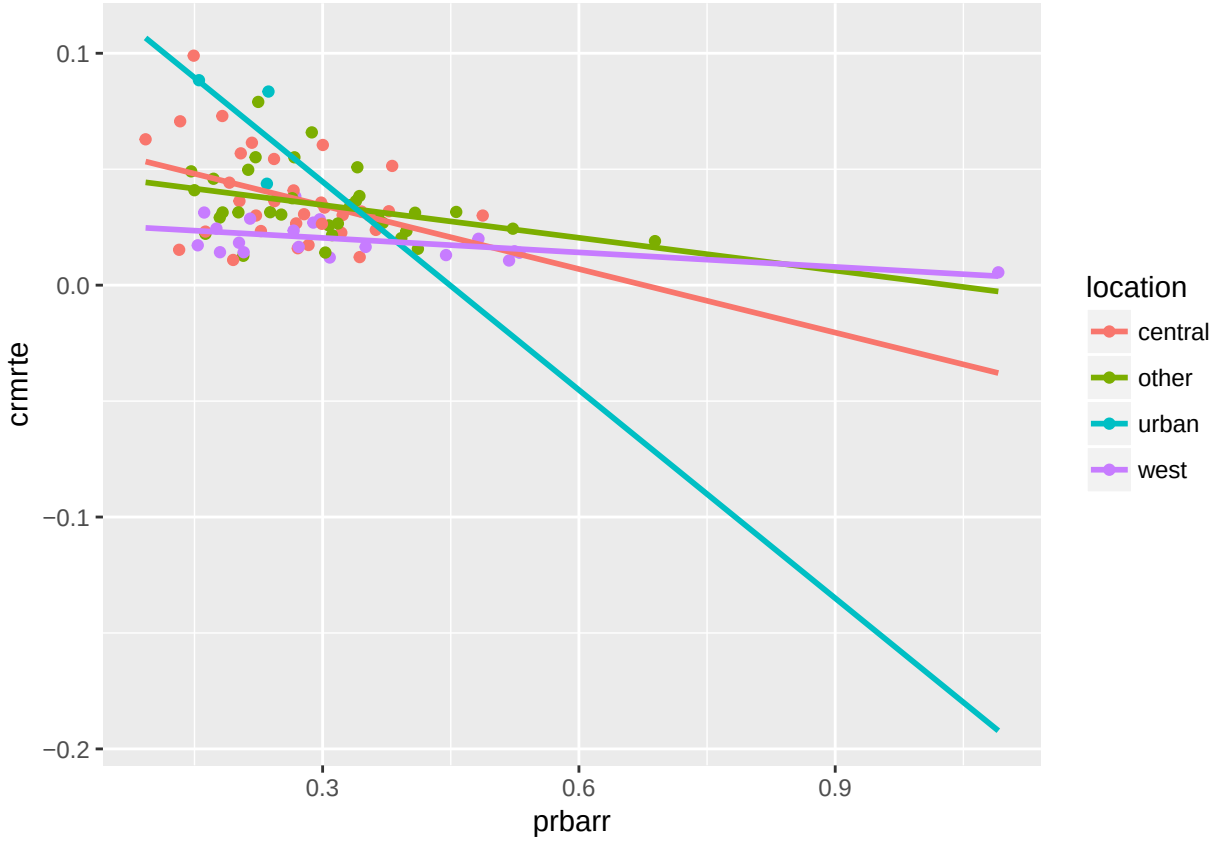
Here we argue that per capita police levels are a response to crime levels (refer discussion regarding associative model above.)

crmrte = -0.017 - 0.029prbarr - 0.007log(prbconv) + 0.013prbpris - 0.0001avgsen + 0.007density + 0.0002taxpc - 0.013west -0.008central

CrimeRate vs ProbabilityOfArrest

We observe a negative correlation between crime rates and probability of arrests. The model suggests that higher arrest rates results in decreases in crime rates (presumably because more would be criminals would be imprisoned)

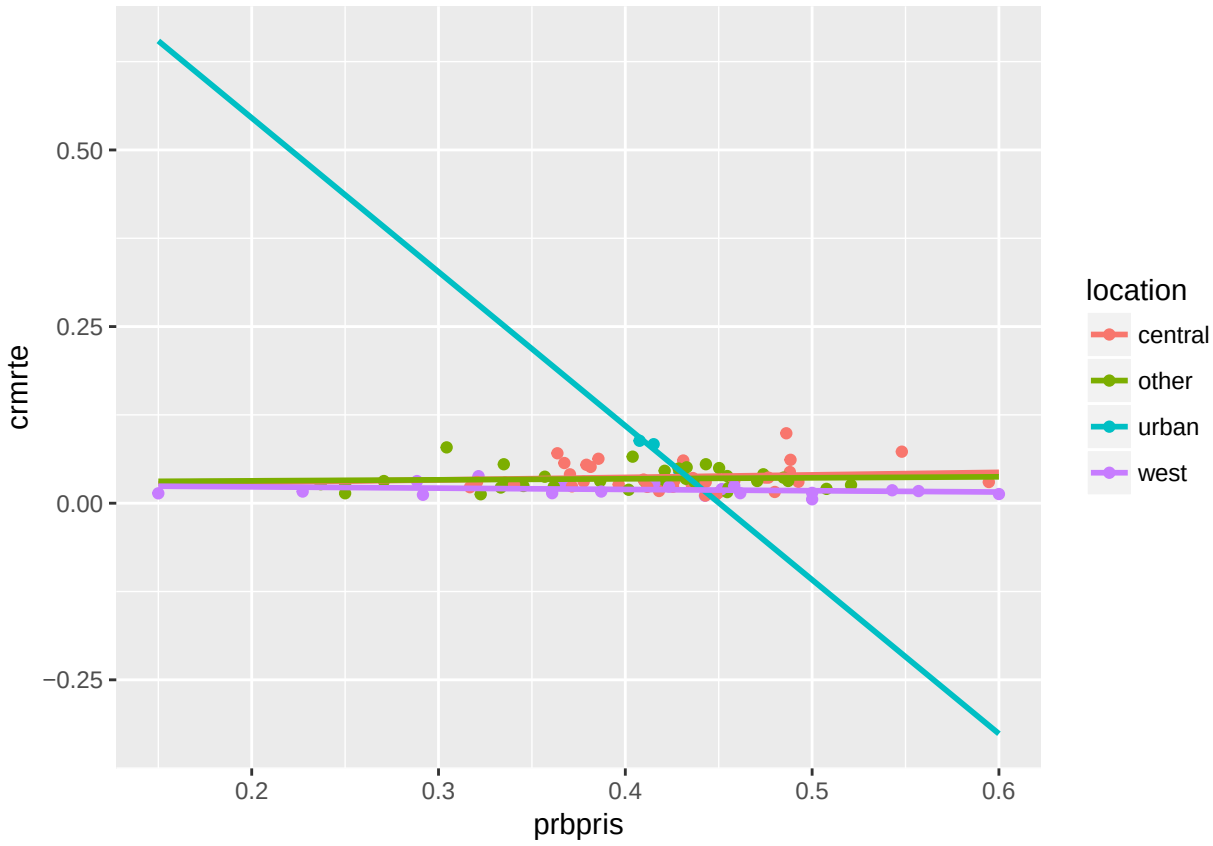
```
qplot(x = prbarr, y = crmrte, color = location, data = Data2) +
  stat_smooth(method = "lm", se = FALSE, fullrange = TRUE)
```



CrimeRate vs ProbabilityOfPrison

Crime rates appear to be positively correlated with probability of imprisonment. The model suggests a causal relationship between higher levels of arrest and crime. We speculate that this might be related to increases in the population of ex-convicts (arising from higher levels of imprisonment)

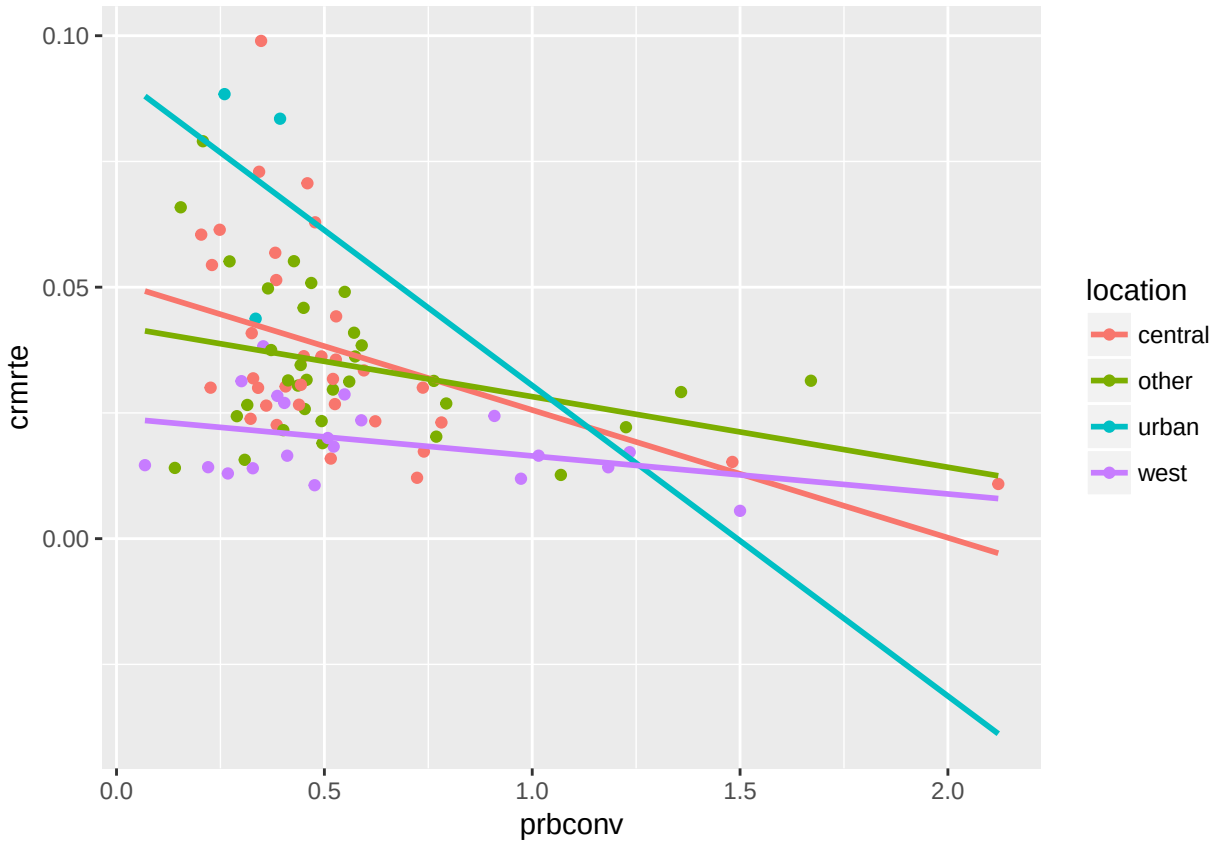
```
qplot(x = prbpris, y = crmrte, color = location, data = Data2) +
  stat_smooth(method = "lm", se = FALSE, fullrange = TRUE)
```



CrimeRates vs ProbabilityOfConviction

Roughly, a 1% gross change in probabilty of conviction is correlated with a ~1% decrease in crime rates

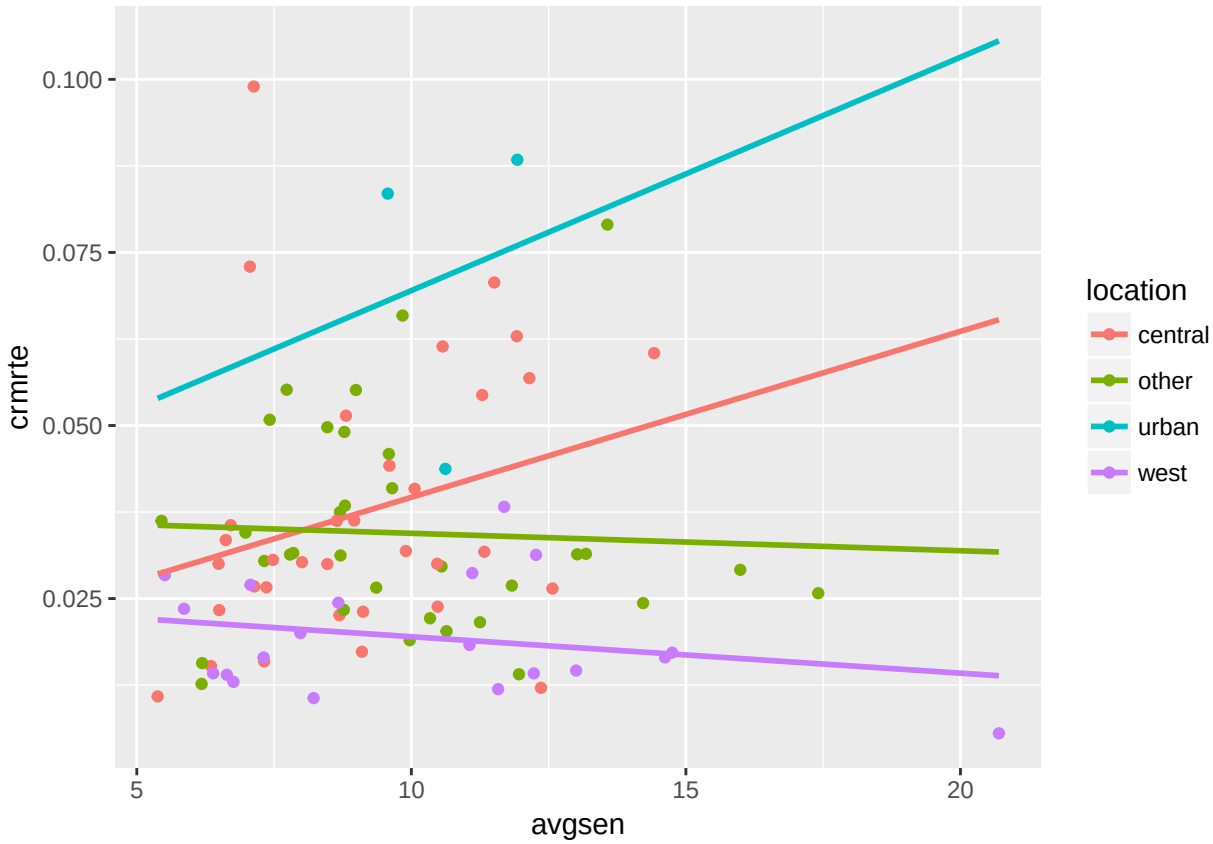
```
qplot(x = prbconv, y = crmrte, color = location, data = Data2) +

  stat_smooth(method = "lm", se = FALSE, fullrange = TRUE)
```



CrimeRate vs AverageSentences

Sentence lengths have minimal impact on levels of crime

```
qplot(x = avgsen, y = crmrte, color = location, data = Data2) +

  stat_smooth(method = "lm", se = FALSE, fullrange = TRUE)
```



CrimeRate vs Location

The model suggests that crime rates appear to vary by location with central and west having relatively lower levels of crime The preceding graphs present the same trends (parallel slopes)

Policy Prescriptions based on Causal Model (Excludes "polpc" predictor)

Policy I: Increase the probability of arrest:

Increase numbers of police per capita – the associative model above suggests ~1 additional police officer per each %1 increase in crime rate

Policy II: Reduce levels of recividism

In order to counter the negative impacts of imprisonment on reincoperation of ex-convicts into society, expand in-prison programs aimed at building skills of the inmate population

Policy III: Increase conviction rates

Ensure that the offices of the public prosecutor as well as the criminal investigative divisions are adequately funded Increase police officers training – particularly as it relates to evidence processing etc.

Policy IV: Moderate prison sentences

Longer stays in prison have no meaningful impact on crime rates

APPENDIX: Additional areas of study

CrimeRate vs AverageWages

```
library(gridExtra)

plot.complex <- qplot(x = w_avg, y = crmrte,

                      color = location, data = Data2) +

  geom_abline(aes(intercept = intercepts,

                  slope = slopes,

                  color = location), data = lines.df)


# Single intercept model

p <- ggplot(Data2, aes(x = w_avg, y = crmrte))

plot.simple <- p + geom_point(aes(colour = location)) + stat_smooth(method = "lm")

grid.arrange(plot.complex, plot.simple, ncol = 2)
```

```
## Warning: Removed 4 rows containing missing values (geom_abline).
```