Lab 4: Reducing Crime

Sullivan Swift, Jayanth Srinivasa

December 4, 2017

## Introduction

To address questions regarding the determinants of crime in North Carolina in 1987, we conducted an analysis of the state’s crime rate and possible related variables, including many of the following:

#description of the variables in the crime dataset  
desc

## variable label  
## 1 county county identifier  
## 2 year 1987  
## 3 crmrte crimes committed per person  
## 4 prbarr `probability` of arrest  
## 5 prbconv `probability` of conviction  
## 6 prbpris `probability` of prison sentence  
## 7 avgsen avg. sentence, days  
## 8 polpc police per capita   
## 9 density people per sq. mile  
## 10 taxpc tax revenue per capita  
## 11 west =1 if in western N.C.  
## 12 central =1 if in central N.C.  
## 13 urban =1 if in SMSA  
## 14 pctmin80 perc. minority, 1980  
## 15 wcon weekly wage, construction  
## 16 wtuc wkly wge, trns, util, commun  
## 17 wtrd wkly wge, whlesle, retail trade  
## 18 wfir wkly wge, fin, ins, real est  
## 19 wser wkly wge, service industry  
## 20 wmfg wkly wge, manufacturing  
## 21 wfed wkly wge, fed employees  
## 22 wsta wkly wge, state employees  
## 23 wloc wkly wge, local gov emps  
## 24 mix offense mix: face-to-face/other  
## 25 pctymle percent young male

#number of rows  
nrow(crime)

## [1] 90

summary(crime)

## 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

To address the question regarding the causes of crime, we examined variables crime rate (crmrte), density per square mile (density), tax revenue per capita (taxpc), percent minority in 1980 (pctmin80), percent young male (pctymle), and probability of arrest (prbarr) as our main variables of interest. We also wanted to examine wage data, so we calculated the median weekly wage (med\_wag) for each county. We believe these variables will give us a holistic view of each county in terms of population density, demographics, and wealth.

First, we created our new med\_wag variable, and performed a high level analysis to assess the quality of our data.

#Take the median for each county from all wage variables.  
crime$med\_wag <- apply(crime[,(16:24)], 1, median, na.rm=TRUE)  
  
#check for NAs among key variables  
C <- crime  
filter = !is.na(C$crmrte) | !is.na(C$density) | !is.na(C$taxpc) | !is.na(C$pctmin80) | !is.na(C$pctymle) | !is.na(C$med\_wag)  
C = C[filter,]  
summary(C)

## 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 med\_wag   
## Min. :0.01961 Min. :0.06216 Min. :231.7   
## 1st Qu.:0.08060 1st Qu.:0.07437 1st Qu.:294.3   
## Median :0.10095 Median :0.07770 Median :310.1   
## Mean :0.12905 Mean :0.08403 Mean :315.5   
## 3rd Qu.:0.15206 3rd Qu.:0.08352 3rd Qu.:331.1   
## Max. :0.46512 Max. :0.24871 Max. :436.8

nrow(C)

## [1] 90

#review the aggregate location data  
table(crime$west)

##   
## 0 1   
## 69 21

table(crime$central)

##   
## 0 1   
## 56 34

table(crime$urban)

##   
## 0 1   
## 82 8

summary(crime$county)

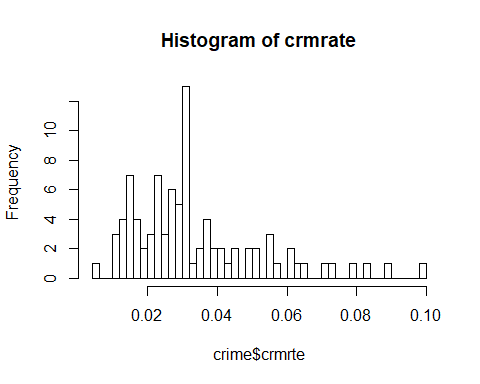
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 51.5 103.0 100.6 150.5 197.0

There are no NA values, so we were able to continue with the complete dataset. Above we briefly examined location data. We originally wanted to incorporate location into the analysis, but these three variables did not create groups of even and the county variable is too granular on its own, so we did not include location data in the analysis.

#examine key outcome variable, crmrte  
summary(crime$crmrte)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.005533 0.020604 0.030002 0.033510 0.040249 0.098966

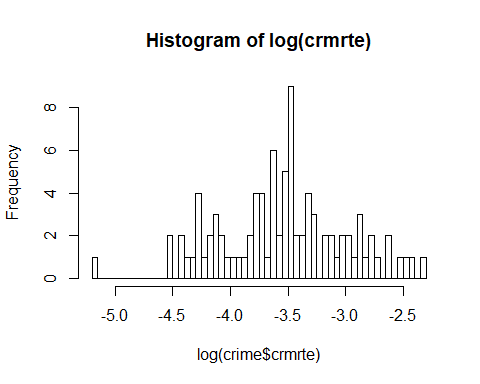
hist(crime$crmrte, breaks=50,  
 main="Histogram of crmrate")



#apply log transformation  
summary(log(crime$crmrte))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -5.197 -3.883 -3.506 -3.542 -3.213 -2.313

hist(log(crime$crmrte), breaks=50,  
 main="Histogram of log(crmrte)")



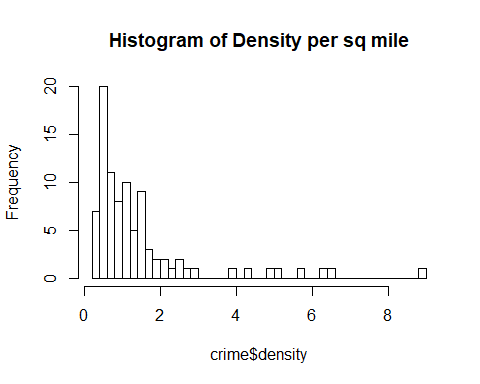
#store transformation  
crime$log\_crmrte <- log(crime$crmrte)

First we analyzed the outcome variable, crime rate (crmrte). In the original crmrte variable, the distribution is right tailed. We applied a log transformation to crmrte, and this variable had a more normal distribution. We chose to use the log of crmrte as our outcome variable in the models below.

#examine predictor variable, density  
summary(crime$density)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2034 0.5472 0.9792 1.4379 1.5693 8.8277

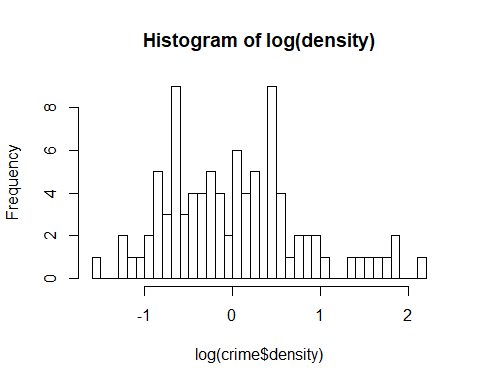
hist(crime$density, breaks=50,  
 main="Histogram of Density per sq mile")



#apply log transformation  
summary(log(crime$density))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.59247 -0.60298 -0.02112 0.01653 0.45060 2.17789

hist(log(crime$density), breaks=50,  
 main="Histogram of log(density)")



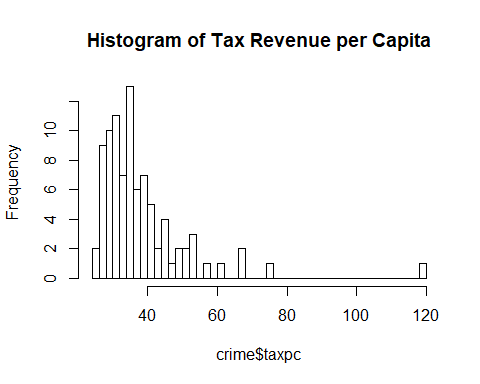
#store transformation  
crime$log\_density <- log(crime$density)

Next we examined density per square mile (density). There is a positive skew, so we again applied a log transformation. The log transformation of density is not quite normal, but since our , we can rely on the Central Limit Theorem.

#examine predictor variable, taxpc  
summary(crime$taxpc)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 25.69 30.73 34.92 38.16 41.01 119.76

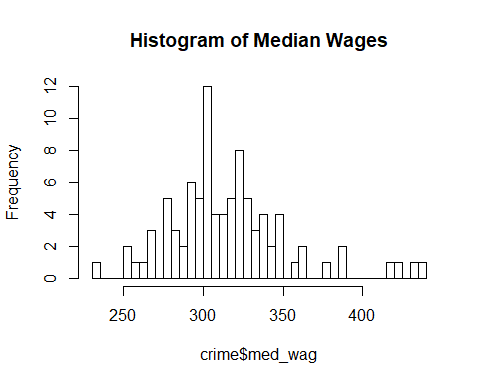
hist(crime$taxpc, breaks=50,  
 main="Histogram of Tax Revenue per Capita")



#examine predictor variable, med\_wag  
summary(crime$med\_wag)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 231.7 294.3 310.1 315.5 331.1 436.8

hist(crime$med\_wag, breaks=50,  
 main="Histogram of Median Wages")

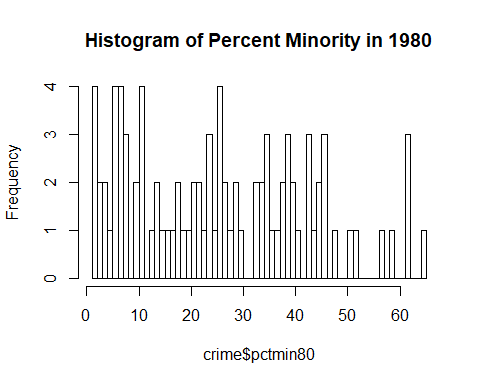


We then analyzed tax revenue per capita (taxpc) and median wage (med\_wag). We were concerned that these two variables have too much overlap in effect - wages are likely a very similar measure to the tax revenue per capita. After looking at the histograms of these two variables, we chose med\_wage over taxpc because it is closer to a normal distribution on its own and will be more clear to interpret in our models below.

#examine predictor variable, pctmin80  
summary(crime$pctmin80)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.284 10.024 24.852 25.713 38.183 64.348

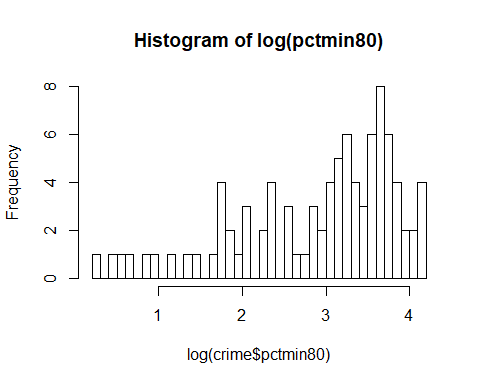
hist(crime$pctmin80, breaks=50,  
 main="Histogram of Percent Minority in 1980")



#apply log transformation  
summary(log(crime$pctmin80))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2497 2.3050 3.2127 2.9134 3.6424 4.1643

hist(log(crime$pctmin80), breaks=50,  
 main="Histogram of log(pctmin80)")

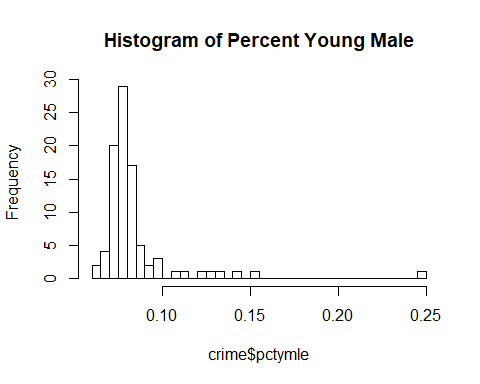


Next, we looked at percent minority in 1980 (pctmin80). This data’s distribution was not normal, so we applied a log transformation. This caused a left skew to the data, and we chose to continue with the untransformed variable, relying on the Central Limit Theorem for normality.

#examine predictor variable, pctymle  
summary(crime$pctymle)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.06216 0.07437 0.07770 0.08403 0.08352 0.24871

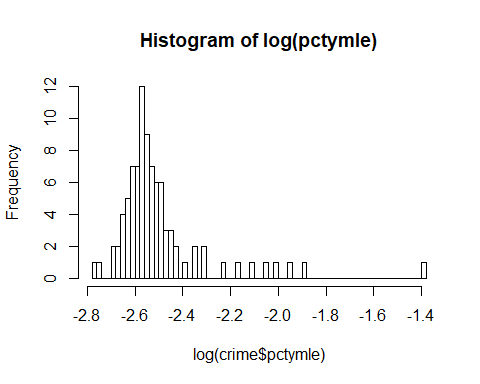
hist(crime$pctymle, breaks=50,  
 main="Histogram of Percent Young Male")



#apply log transformation  
summary(log(crime$pctymle))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.778 -2.599 -2.555 -2.501 -2.483 -1.391

hist(log(crime$pctymle), breaks=50,  
 main="Histogram of log(pctymle)")



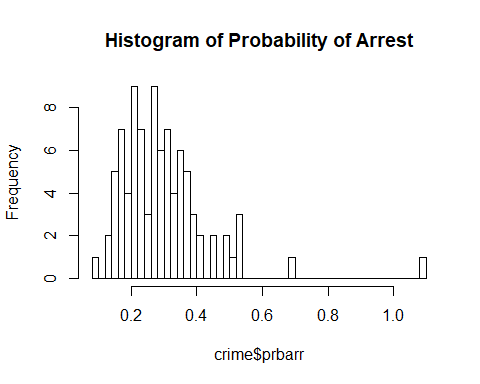
#store transformation  
crime$log\_pctymle <- log(crime$pctymle)

Then we examined the percentage of young males variable (pctymle). The values were all very small, between .06 and .25, and it appears to be in a decimal percentage format. pctymle also has a right skew. To address the skew, we used a log transformed pctymle. Using a log transformation we can also more easily interpret pctymle below in our models.

#examine predictor variable, prbarr  
summary(crime$prbarr)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.09277 0.20495 0.27146 0.29524 0.34487 1.09091

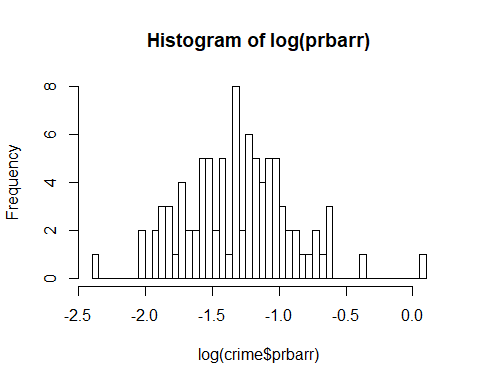
hist(crime$prbarr, breaks=50,  
 main="Histogram of Probability of Arrest")



#apply log transformation  
summary(log(crime$prbarr))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.37763 -1.58502 -1.30395 -1.30438 -1.06458 0.08701

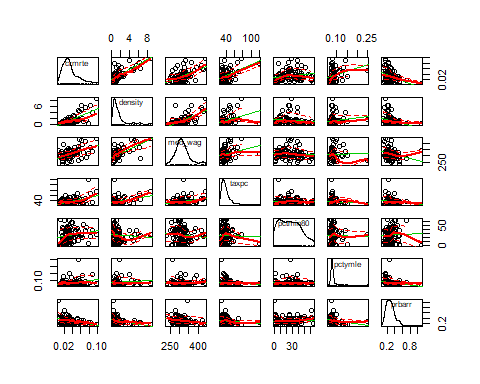
hist(log(crime$prbarr), breaks=50,  
 main="Histogram of log(prbarr)")



#store transformation  
crime$log\_prbarr <- log(crime$prbarr)

Lastly, we examined the probability of arrest (prbarr). While it is not as skewed as the rest, log(prbarr) is still slightly skewed to the right and has one fairly large outlier, so we decided to apply the log transform to this variable. The log transformation has a much more normal distribution.

#scatterplot matrix  
scatterplotMatrix(crime[,c('crmrte', 'density', 'med\_wag', 'taxpc', 'pctmin80', 'pctymle', 'prbarr')])



#correlation matrix  
(c <- with(crime, cor(cbind(crmrte, density, log\_density, med\_wag, taxpc, pctmin80, pctymle,log\_pctymle, prbarr))))

## crmrte density log\_density med\_wag taxpc  
## crmrte 1.0000000 0.72777835 0.7029129 0.510993818 0.44871512  
## density 0.7277783 1.00000000 0.8896747 0.635234332 0.32047367  
## log\_density 0.7029129 0.88967469 1.0000000 0.667310493 0.18412222  
## med\_wag 0.5109938 0.63523433 0.6673105 1.000000000 0.26431059  
## taxpc 0.4487151 0.32047367 0.1841222 0.264310586 1.00000000  
## pctmin80 0.1816506 -0.07470698 -0.1724073 -0.078267679 -0.02797739  
## pctymle 0.2903397 0.11478144 0.2051924 0.009117121 -0.09154375  
## log\_pctymle 0.3241625 0.14914089 0.2446371 0.070216272 -0.07649003  
## prbarr -0.3952830 -0.30053317 -0.3851474 -0.249962020 -0.13719105  
## pctmin80 pctymle log\_pctymle prbarr  
## crmrte 0.18165059 0.290339658 0.32416249 -0.39528302  
## density -0.07470698 0.114781444 0.14914089 -0.30053317  
## log\_density -0.17240732 0.205192423 0.24463711 -0.38514738  
## med\_wag -0.07826768 0.009117121 0.07021627 -0.24996202  
## taxpc -0.02797739 -0.091543750 -0.07649003 -0.13719105  
## pctmin80 1.00000000 -0.019256570 -0.01224664 0.04907002  
## pctymle -0.01925657 1.000000000 0.97454166 -0.18096201  
## log\_pctymle -0.01224664 0.974541661 1.00000000 -0.20725384  
## prbarr 0.04907002 -0.180962011 -0.20725384 1.00000000

Finally we examined a scatterplot matrix and correlation matrix to quickly assess the relationships between our variables. None of the variables we examined have a perfect correlation, though some have fairly strong relationships. Density has a strong correlation with the crmrte. We used density as a key variable in our models below for its intutive relationship to crmrte as well as the corrlational relationship. We chose to include med\_wag, but not taxpc, as med\_wag has a more normal distribution and has a more clear interpretation. We chose to include pctmin80, and the transformation log(pctymle) over pctymle. We can use this correlation matrix to confirm MLR3, no perfect multicollinearity, for our models below.

## Modeling Crime Rate and Addressing Assumptions

Addressing CLM1/MLR1, the linearity assumption, the three models we create below are linear in nature, so we meet this assumption.

Our sample appears to be nearly the entire population, and we ran into no NA values above. North Carolina has 100 counties, and our dataset contains 90. Enough of the population is included in the dataset for us to assume CLM2/MLR2, the assumption of random sampling is satisfied.

We addressed MLR3 above.

### Model 1

We defined our first model hoping to use only one or two key variables. We settled on log(crmrte) ~ log(density), as the more densely populated an area, the opportunity for crime to occur increases both from increased individuals and increased property in an area. Above we also see that density has a strong, positive correlation with crmrte.

#create the model  
(model1 <- lm(log\_crmrte ~ log\_density, data=crime))

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density, data = crime)  
##   
## Coefficients:  
## (Intercept) log\_density   
## -3.5498 0.4858

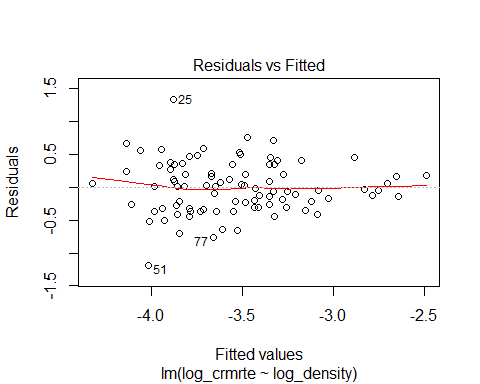
First, we create the model (model1) and examine its coefficients. For each 1% increase in density, there is a .49% increase in crmrte.

#examine the covariance between the predictor and the residuals  
cov(crime$log\_density, model1$residuals)

## [1] 1.476083e-18

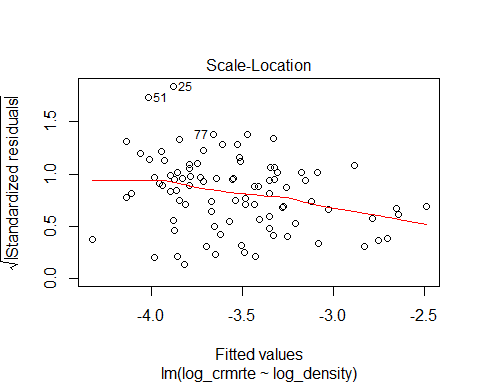
Examining the covariance, we see a very small relationship between log\_density and model1’s residuals. This supports exogeneity.

#resdiuals v fitted values plot  
plot(model1, which=1)



Next we examine the Residuals v Fitted plot to assess MRL4, the zero-conditional mean assumption. We see that the red spline line is very close to zero through the whole graph. The exception is the small uptick on the far left, but this is likely due to few data points.

#scale-location plot  
plot(model1, which=3)



#breusch-pagan test  
bptest(model1)

##   
## studentized Breusch-Pagan test  
##   
## data: model1  
## BP = 5.7062, df = 1, p-value = 0.01691

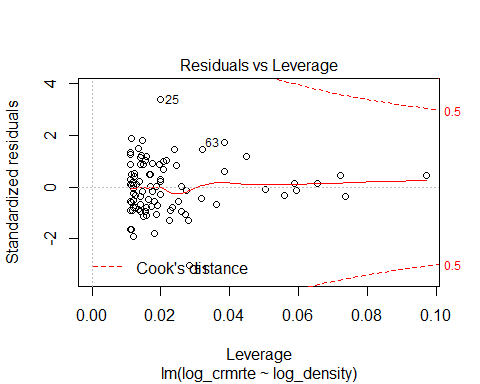
#score-test  
ncvTest(model1)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 7.951009 Df = 1 p = 0.004806057

#heteroskedatic robust standard errors  
se.model1 = sqrt(diag(vcovHC(model1)))

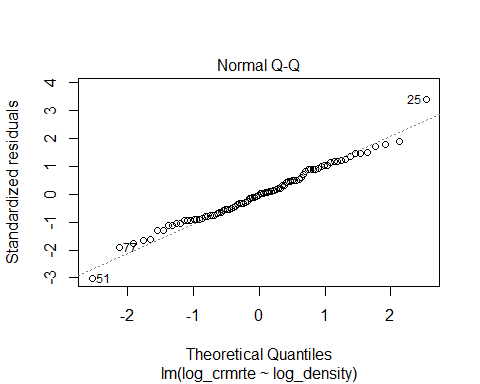
To assess MLR5, the homoskedasticity assumption, we look at the Residuals v Fitted plot above, as well as the Scale-Location plot. On both plots, the band of data points narrows as we move right on the graphs, providing evidence of a violation of homoskedasticity. Further supporting this violation, the Breusch-Pagan test and the Score-test both have significant p-values (bptest, ; score-test, ). Our sample is , however the sample is not extremely large. We will use the robust standard errors we have produced above to address the possibility of heteroskedasticity.

#residuals vs leverage plot  
plot(model1, which=5)

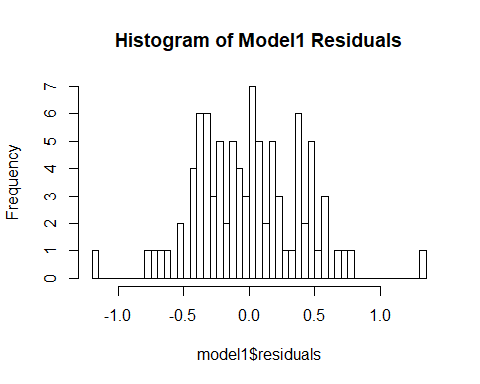


Throughout out analysis above, we noticed there was at least one quite large outlier. Here we looked at the Residuals v Leverage plot to determine if we need to remove this data point. The outlier fall inside of the Cook’s distance and does not appear to have any sizable bearing on model1, so we may leave it in our analysis.

#qqplot  
plot(model1, which=2)



#histogram of Model1 residuals  
hist(model1$residuals, breaks=50,  
 main="Histogram of Model1 Residuals")



When assessing MLR6, we examine the Q-Q plot of the model. Here, we can see that the data closely hugs the diagonal, indicating normal residuals. The histogram of the residuals also supports this conclusion. Both charts a few values on the extremes that vary, but on the whole support MLR6.

Next, we ran a summary on model1 and used our standard errors that are robust to heteroskedasticity.

(s1 <- summary(model1))

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density, data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.18454 -0.29595 -0.00466 0.26400 1.33734   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.54976 0.04201 -84.508 < 2e-16 \*\*\*  
## log\_density 0.48581 0.05403 8.991 4.28e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3984 on 88 degrees of freedom  
## Multiple R-squared: 0.4788, Adjusted R-squared: 0.4729   
## F-statistic: 80.84 on 1 and 88 DF, p-value: 4.28e-14

s1$coefficients[, 2] <- sqrt(diag(vcovHC(model1)))  
s1

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density, data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.18454 -0.29595 -0.00466 0.26400 1.33734   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.54976 0.04276 -84.508 < 2e-16 \*\*\*  
## log\_density 0.48581 0.04896 8.991 4.28e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3984 on 88 degrees of freedom  
## Multiple R-squared: 0.4788, Adjusted R-squared: 0.4729   
## F-statistic: 80.84 on 1 and 88 DF, p-value: 4.28e-14

We noticed that the R-squared value was s1$r.squared, which is a reasonable value.

coeftest(model1, vcov = vcovHC)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.549758 0.042762 -83.0129 < 2.2e-16 \*\*\*  
## log\_density 0.485809 0.048962 9.9222 5.193e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Model 2

Next, we wanted to expand our model with more independent variables, and we chose to add pctymle and pctmin80 incorporate some demographic information in our model. Since the pctymle distribution was skewed from the analysis of the histogram, we decided to use the log form of the variable in the analysis. The percent of young males (pctymle) and the percent minority in 1980 (pctmin80) are both reasonable variables to include in this model as they tell us about the demographic diversity of a county. Though the pctmin80 data is 7 years old, we chose to include it in our model so we could somewhat assess the impact of minorities on crime rate.

#create the model  
(model2 <- lm(log\_crmrte ~ log\_density + log\_pctymle + pctmin80, data=crime))

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density + log\_pctymle + pctmin80,   
## data = crime)  
##   
## Coefficients:  
## (Intercept) log\_density log\_pctymle pctmin80   
## -2.87900 0.50516 0.38753 0.01159

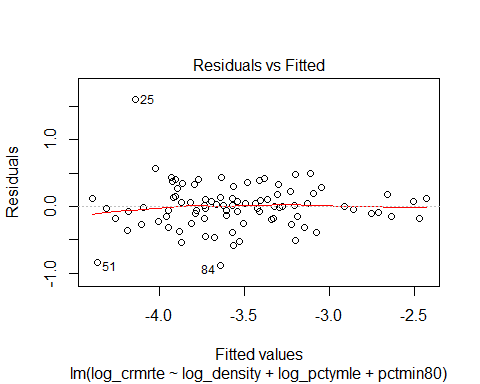
From examining the coefficients we see that, with everything else being held a constant an increase of 1% in density, we see an increase of 0.5% in the crime rate. Similarly, an increase in pctymle with everything else being constant, produces an increase of crime rate by 0.38% and finally any increase of minority population by 1 unit, results in a nearly 1% increase in the crime rate.

Now let us check the remaining assumptions. First, we will double check MLR3 with the variable’s Variance Inflation Factors. Then we will examine the zero conditional mean assumption by looking at the plot of residuals vs fitted values.

#Variance Inflation Factors  
vif(model2)

## log\_density log\_pctymle pctmin80   
## 1.097155 1.064703 1.031648

#residuals v fitted values plot  
plot(model2, which=1)



The VIF values are low, less than 10, supporting the MLR3 assumption. The red spline line is close to zero for all the fitted values, with only a small downturn on the left side. Compared to model1, the red spline line appears to fit zero better. This supports making the MLR4 assumption.

#examine covariance of predictor variables and model residuals  
cov(crime$log\_density, model2$residuals)

## [1] -1.580126e-18

cov(crime$log\_pctymle, model2$residuals)

## [1] -1.024933e-17

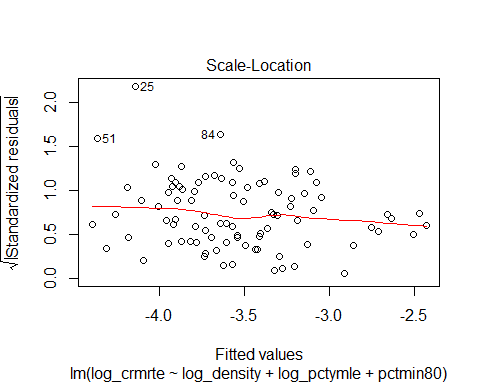
cov(crime$pctmin80, model2$residuals)

## [1] -1.138144e-16

The covariance between each of the independent variables and the model residuals are each very small, and it is clear that exogenity assumption holds.

To assess homoskedasticity, we looked at the graph of standardized residuals vs the fitted values below.

#scale-location plot  
plot(model2, which=3)



#breusch-pagan test  
bptest(model2)

##   
## studentized Breusch-Pagan test  
##   
## data: model2  
## BP = 4.5657, df = 3, p-value = 0.2065

#score test  
ncvTest(model2)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 14.46776 Df = 1 p = 0.0001425787

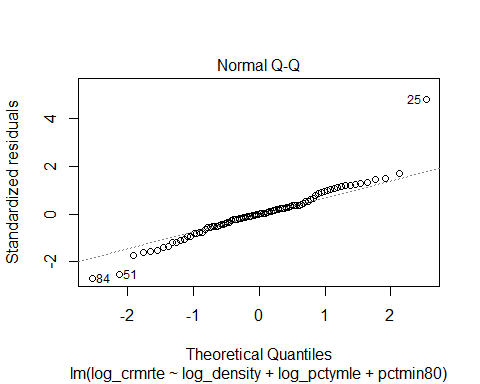
#heteroskedastic robust standard errors  
se.model2 = sqrt(diag(vcovHC(model2)))

The graph shows some outliers to the upper left edge and the band seems to narrow as we move to the right, indicating there may be heteroskedasticity.

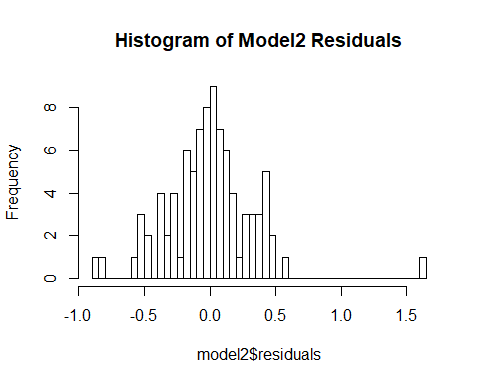
The Breusch-Pagan test did not yield significant results (), but the Non-constant Variance Score Test was significant with . We will be conservative and use the robust standard errors we have produced above to address any heteroskedacity in the data, just as in model1.

To test the normality of the errors, we examined the Q-Q plot of the residuals and the histogram of the residuals.

#qq plot  
plot(model2, which=2)



#histogram of model2 residuals  
hist(model2$residuals, breaks=50,  
 main="Histogram of Model2 Residuals")



The Q-Q plot has some variation around the ends, and both the histogram and the Q-Q plot contain an outlier to the far right. The histogram of the residuals is fairly normal, except for the outlier on the far right. model2 has more variation in the residuals than model1, but the evidence is strong enough to support making the MLR6 assumption.

Finally we can examine the summary command, using our heteroskedasticity robust standard errors from above.

(s2 <- summary(model2))

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density + log\_pctymle + pctmin80,   
## data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.88106 -0.17531 0.00089 0.14727 1.60223   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.879001 0.477722 -6.027 4.05e-08 \*\*\*  
## log\_density 0.505158 0.048554 10.404 < 2e-16 \*\*\*  
## log\_pctymle 0.387526 0.188409 2.057 0.0427 \*   
## pctmin80 0.011588 0.002167 5.349 7.22e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3418 on 86 degrees of freedom  
## Multiple R-squared: 0.6251, Adjusted R-squared: 0.612   
## F-statistic: 47.8 on 3 and 86 DF, p-value: < 2.2e-16

s2$coefficients[, 2] <- sqrt(diag(vcovHC(model2)))  
s2

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density + log\_pctymle + pctmin80,   
## data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.88106 -0.17531 0.00089 0.14727 1.60223   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.879001 0.265865 -6.027 4.05e-08 \*\*\*  
## log\_density 0.505158 0.045689 10.404 < 2e-16 \*\*\*  
## log\_pctymle 0.387526 0.101180 2.057 0.0427 \*   
## pctmin80 0.011588 0.002833 5.349 7.22e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3418 on 86 degrees of freedom  
## Multiple R-squared: 0.6251, Adjusted R-squared: 0.612   
## F-statistic: 47.8 on 3 and 86 DF, p-value: < 2.2e-16

That the rsquare value has improved to s2$r.squared and the adjusted R-square is s2$adj.r.squared in Model2. This indicates a better performance compared to Model1.

coeftest(model2, vcov = vcovHC)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.8790007 0.2658649 -10.8288 < 2.2e-16 \*\*\*  
## log\_density 0.5051581 0.0456889 11.0565 < 2.2e-16 \*\*\*  
## log\_pctymle 0.3875264 0.1011804 3.8301 0.0002428 \*\*\*  
## pctmin80 0.0115884 0.0028331 4.0903 9.67e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Model 3

We created one more model to include the previous covariates and along with the probability of arrests (prbarr) and the median wage (med\_wag) to assess the robustness of our first models.

#create the model  
(model3 <- lm(log\_crmrte ~ log\_density + log\_pctymle + pctmin80 + med\_wag + log\_prbarr, data=crime))

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density + log\_pctymle + pctmin80 +   
## med\_wag + log\_prbarr, data = crime)  
##   
## Coefficients:  
## (Intercept) log\_density log\_pctymle pctmin80 med\_wag   
## -3.3050608 0.4799760 0.2771288 0.0118819 -0.0005766   
## log\_prbarr   
## -0.2489919

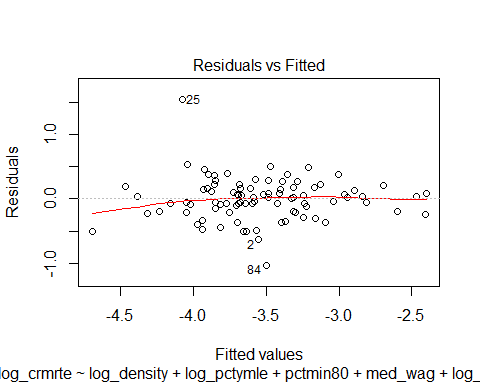
From the model3, we can say that as the population density increases by 1%, the crime rate goes up by 0.4799% with everything else being constant. Similarly, a percentage increase in the fraction of young males in the population by 1% increases the crime rate by 0.27% and an increase in minorities by 1% (compared to their population in 1980), causes an increase in crime-rate by 1.18%. Finally, the two new variables med\_wage and probability of arrests (prbarr), both have a negative effect on the crime rate. And as median wage increases by a unit of 1, the crime-rate decreases by 0.5% and as the probability of arrest goes up by 1%, the crime rate goes down by 0.24%.

As before, we confirm MLR3 with VIF and examine the plot of residuals vs fitted values to determine the zero conditional mean assumption.

#Variance Inflation Factors  
vif(model3)

## log\_density log\_pctymle pctmin80 med\_wag log\_prbarr   
## 2.073786 1.139526 1.037493 1.864206 1.267325

#residuals vs fitted values plot  
plot(model3, which=1)



Again, all VIF values are less than 10, supporting the MLR3 assumption. There is no evidence of a violation of MLR4. While the red spline line very close to zero, we also observe that to the left the line seems to move further below zero than in model2. This can likely be attributed to fewer data points on the left side of the fitted values.

#examine covariance of predictor variables and model residuals  
cov(crime$log\_density, model3$residuals)

## [1] -1.759015e-17

cov(crime$log\_pctymle, model3$residuals)

## [1] -9.811306e-18

cov(crime$pctmin80, model3$residuals)

## [1] 6.636779e-18

cov(crime$med\_wag, model3$residuals)

## [1] -2.881546e-15

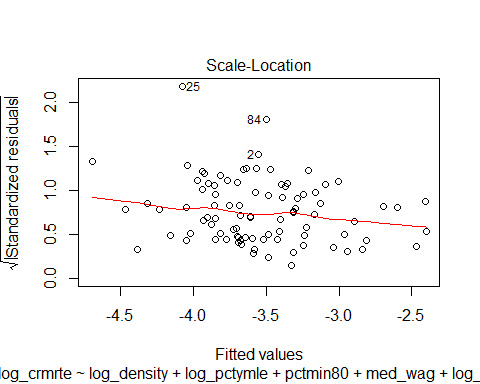
cov(crime$log\_prbarr, model3$residuals)

## [1] 5.153691e-18

Again, the covariance between each of the independent variables and the model residuals are very small, and we see exogeneity holds.

Next we look at model3’s scale-location plot, Breusch-Pagan test, and the Non-constant Variance Score Test to assess homoskedasticity.

#scale-location plot  
plot(model3, which=3)



#breusch-pagan test  
bptest(model3)

##   
## studentized Breusch-Pagan test  
##   
## data: model3  
## BP = 8.302, df = 5, p-value = 0.1404

#score test  
ncvTest(model3)

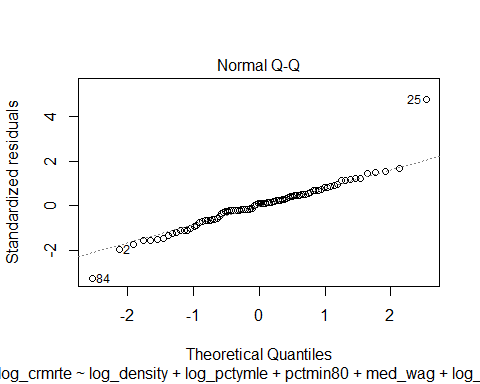
## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 8.565218 Df = 1 p = 0.00342646

#heteroskedastic robust standard error  
se.model3 = sqrt(diag(vcovHC(model3)))

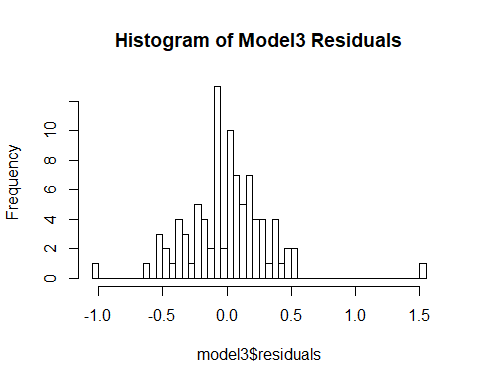
The graph clearly shows there are some outliers, showing the variables are heteroskedastic. Further supporting this violation, the Non-constant Variance Score Test has significant p-value (). The Breusch-Pagan test did not yield significant results () We will be conservative and use the robust standard errors we have produced above to address any heteroskedasticity in the data.

Again, we examine a Q-Q plot of the residuals and the histogram of the residuals to determine the MLR6 assumption.

#qq plot  
plot(model3, which=2)



#histogram of model3 residuals  
hist(model3$residuals, breaks=50,  
 main="Histogram of Model3 Residuals")



Both plots show a fairly normal distribution, with a smaller outlier to the left and a larger outlier to the right. model3’s residuals histogram looks marginally more normal than model2’s, and MLR6 still holds.

Finally we examine the summary of model3 with residuals robust to heteroskedasticity.

s3 <- summary(model3)   
s3$coefficients[, 2] <- sqrt(diag(vcovHC(model3)))  
s3

##   
## Call:  
## lm(formula = log\_crmrte ~ log\_density + log\_pctymle + pctmin80 +   
## med\_wag + log\_prbarr, data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.02477 -0.18983 0.02709 0.16743 1.53744   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.3050608 0.5116359 -5.462 4.71e-07 \*\*\*  
## log\_density 0.4799760 0.0826760 7.371 1.08e-10 \*\*\*  
## log\_pctymle 0.2771288 0.1306190 1.458 0.1487   
## pctmin80 0.0118819 0.0028858 5.606 2.58e-07 \*\*\*  
## med\_wag -0.0005766 0.0014211 -0.451 0.6529   
## log\_prbarr -0.2489919 0.1356380 -2.521 0.0136 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3334 on 84 degrees of freedom  
## Multiple R-squared: 0.6516, Adjusted R-squared: 0.6308   
## F-statistic: 31.42 on 5 and 84 DF, p-value: < 2.2e-16

Though the rsquare value has improved to s3$r.squared and the adjusted R-square is s3$adj.r.squared in model3, it hasn’t improved all that much in comparison to model2 and does not justify the addition of the extra variables.

coeftest(model3, vcov = vcovHC)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.30506084 0.51163591 -6.4598 6.484e-09 \*\*\*  
## log\_density 0.47997596 0.08267597 5.8055 1.109e-07 \*\*\*  
## log\_pctymle 0.27712877 0.13061897 2.1217 0.03681 \*   
## pctmin80 0.01188191 0.00288581 4.1174 8.923e-05 \*\*\*  
## med\_wag -0.00057661 0.00142110 -0.4057 0.68596   
## log\_prbarr -0.24899187 0.13563804 -1.8357 0.06994 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Comparison of Models

We used the Stargazer library function to compare our three models. Though we see little evidence of complete heteroskedasticity, we would like to be conservative and we are using the robust standard errors for all the three models.

#model comparison output with stargazer  
stargazer(model1, model2, model3, type = "text", omit.stat = "f",  
 se = list(se.model1, se.model2, se.model3),  
 star.cutoffs = c(0.05, 0.01, 0.001))

##   
## ===================================================================  
## Dependent variable:   
## -----------------------------------------------  
## log\_crmrte   
## (1) (2) (3)   
## -------------------------------------------------------------------  
## log\_density 0.486\*\*\* 0.505\*\*\* 0.480\*\*\*   
## (0.049) (0.046) (0.083)   
##   
## log\_pctymle 0.388\*\*\* 0.277\*   
## (0.101) (0.131)   
##   
## pctmin80 0.012\*\*\* 0.012\*\*\*   
## (0.003) (0.003)   
##   
## med\_wag -0.001   
## (0.001)   
##   
## log\_prbarr -0.249   
## (0.136)   
##   
## Constant -3.550\*\*\* -2.879\*\*\* -3.305\*\*\*   
## (0.043) (0.266) (0.512)   
##   
## -------------------------------------------------------------------  
## Observations 90 90 90   
## R2 0.479 0.625 0.652   
## Adjusted R2 0.473 0.612 0.631   
## Residual Std. Error 0.398 (df = 88) 0.342 (df = 86) 0.333 (df = 84)  
## ===================================================================  
## Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001

For model1, we see that the log\_density is statistically significant at the level. Also since we are examining the elastic relationship between the crime rate and the density of population, we see that for 1% increase in the density of population results in an increase in crmrte bye 0.486%. Since this is a pretty observable effect, the log\_density is practically significant too.

For model2, we see that log\_density, log\_pctymle, and pctmin80 are all statistically significant at the level. This clearly indicates statistical significance. Also, given everything else is constant, the crime rate increases by 0.5% with every 1% increase in the density of population, increases by 0.38% for every 1% increase in the fraction of the young male population and finally the crime rate will increase by 1.2% when the minority population increases by 1% over the 1980 value. The value of change signifies a practical significance too.

For model3, we see that the log\_density and pctmin80 are statistically significant at the p < 0.001 level, log\_pctymle is statistically significant at the 0.01 level, but med\_wag and log\_prbarr are not statistically significant. Practically, holding everything else constant, just increasing the density of the population by 1% increases the crime rate by 0.48%, only increasing the young male population by 1% causes an increase in crime rate by 0.27%, increasing the minority population by 1% increase crime rate by 1.2%. The median wage and the log\_prbarr (log of probability of arrest) have a negative effect on the crime rate. That is as the median wage increases by 1 unit, the crime decreases by 1% and while the probability of arrest increases by 1%, the crime rate decreases by 0.25%.

From the adjusted r-squared values we can clearly see an improvement from model1 to model2. However, while there is a modest increase in model3 when compared to model2, it does not justify the addition of two variables that are statistically insignificant.

## Omitted Variable Bias

Note that we did not analyze police per capita (polpc), probability of conviction (prbconv), probability of prison sentence (prbpris), average sentence (avgsen), or the offense mix (mix). There are several reasons we excluded polpc from our analysis. First, the number of police officers may be highly dependent on how much crime occurs in an area, and second, a county’s resources, or wealth, could impact the number of police. Since we already have measures for wealth, and due to this possible confounding relationship between police and crime, we excluded the polpc variable from analysis. avgsen and mix are more along the lines of a crime outcome, rather than a predictor of crime in an area. We chose between prbarr, prbcon, and prbpris as these variables measure similar events and including all three would be redundant, so we just include the transformed prbarr variable. We also briefly analyze aggregate location variables (west, central, urban), but these did not cut the data into groups of similar sizes.

The dataset also does not contain any variables that directly correspond to employability or employment availability in the counties. While, there are financial variables, it is reasonable to assume there is a relationship between crime and employment in an area. This creates a bias and since it is sensible to infer that it employability and employment availability would add to the bias of all the independent variables - log\_density, log\_pctymle, pctmin80, log\_prbarr and med\_wag.

Let us look at them one by one. Employability and employment availability in a county would likely increase the density of population and hence would lead to a positive bias and lead to overstating the effects of the density in the model. Same would be the case with the percent of young males, as it is realistic to assume that more young men would be a higher percentage of the population when there are more jobs. Given this to be true, our model will overstate the impact of density and the percentage of young males. The minority population is a similar case. The median wage would depend on the type and number jobs in an area, driving the median wage up or down. In general, it is reasonable to assume that the omitted variable of employment and employment availability would increase the impact of median wage. Finally, the probability of arrests can be reasonably expected to be negatively correlated with employment/employability and this would mean that we are understating the bias of the probability of arrest. Also, we would expect employment opportunities/employability to be negatively correlated with crime rate in general and this will lead to a negative bias on the constant.

## Conclusions and Implications

From our analysis, we can see that the variables log\_density, log\_pctymle and pctmin80 had the biggest impact on crmrte. Although these results are informative, no causal relationship can be drawn from our models. In general, regression itself does not imply causality; it’s basis is in correlation. In this instance, this general rule holds. The data we have is observational, and no experiment was performed. However, we can still see that population density, demographics, and crime rate have a relationship.

Also in model3, we observe that the median wage has a negative slope (while not significant statistically), we can hypothesize that increasing the median wage may decrease crime rate.

Now, since we cannot reduce the density of population in counties, we may need to look at the reasons for increased density, causing an increase in crime rate. It may be a good idea to ensure economic development to reduce the crime rate.

While there is a positive relation between the crime rate and demographic diversity of an area, this does not imply diversity has a negative impact on society. Instead, creating programs to aid underprivileged families and expanding the availability of education would not only help these individuals, but also improve society as a whole and potentially reduce crime.

In conclusion, the crime rate in North Carolina is a complex issue dependent on many variables - some of which cannot be controlled or even observed. However, there are measure we can take improve the social and economic status of an area to reduce the impact of factors that may contribute to an increased crime rate.