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.

```
model_1<-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)

summary(model_1)
```

```
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
```

If we trim the variables with the least explanatory power, we are left with only five variables. It should be noted, that many of the variables excluded, can be removed simply on the basis of have little to no correlation with the dependant variable, such as average sentence length and probability of prison.

```r
model_2<-lm(crime$crmrte~crime$density+crime$prbarr+crime$prbconv+
             crime$polpc+crime$pctmin80)
summary(model_2)$adj.r.square
```

```
[1] 0.7929876
```

```r
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
```

we see here that these 5 variables contain almost all of the predicative power of the other variables, as we only see our r squared drops by less than .01. signifi. 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. Also, to note, the central flag is more strongly correlated with density than with crimrte so it appears that once density is included in the model, most of the value of the central flag in terms of explanatory power is lost.

```r
model_3<-lm(crime$crmrte~crime$urban+crime$prbarr+crime$prbconv+
             crime$polpc+crime$pctmin80)
summary(model_3)$adj.r.squared
```

```
[1] 0.7302545
```

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

```
              crime.crmrte crime.density crime.urban crime.central
crime.crmrte     1.0000000     0.7277783   0.6150631     0.1658803
crime.density    0.7277783     1.0000000   0.8206825     0.3568285
crime.urban      0.6150631     0.8206825   1.0000000     0.1592702
crime.central    0.1658803     0.3568285   0.1592702     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
model_4<-lm(crime$crmrte~crime$wcon+crime$wtuc+crime$wtrd+crime$wfir
           +crime$wser+crime$wmfg+crime$wfed+crime$wsta+crime$wloc)
summary(model_4)
```

```
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
```

When looking at the flag for west, this variable is highly correlated with pctmin80 and is dropped from the model.

```r
cor(crime$west, crime$pctmin80)
```

```
[1] -0.6245144
```

For the remaining three variables, taxpc, mix, and pctymle that we did not include in the final model, they were dropped for the purpose of brevity. If we do include them, the predicative power gets only marginally better and the complexity of the model suffers.

```r
model_5<-lm(crime$crmrte~crime$prbarr+crime$prbconv+crime$polpc+
           crime$density+crime$taxpc+crime$pctmin80+crime$mix+crime$pctymle)
summary(model_5)
```

```
Call:
lm(formula = crime$crmrte ~ crime$prbarr + crime$prbconv + crime$polpc +
    crime$density + crime$taxpc + crime$pctmin80 + crime$mix +
    crime$pctymle)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0195160 -0.0054991 -0.0000579  0.0053447  0.0231196

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.019e-02  6.620e-03   3.050   0.0031 **
crime$prbarr   -5.074e-02  9.248e-03  -5.486 4.56e-07 ***
```

```
crime$prbconv  -2.052e-02  3.023e-03  -6.788 1.73e-09 ***
crime$polpc     6.468e+00  1.252e+00   5.167 1.67e-06 ***
crime$density   5.380e-03  6.871e-04   7.829 1.63e-11 ***
crime$taxpc     1.867e-04  7.765e-05   2.404   0.0185 *
crime$pctmin80  3.774e-04  5.431e-05   6.950 8.41e-10 ***
crime$mix      -2.324e-02  1.267e-02  -1.834   0.0703 .
crime$pctymle   8.322e-02  4.042e-02   2.059   0.0427 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008173 on 81 degrees of freedom
Multiple R-squared:  0.8296,    Adjusted R-squared:  0.8128
F-statistic: 49.29 on 8 and 81 DF,  p-value: < 2.2e-16
```

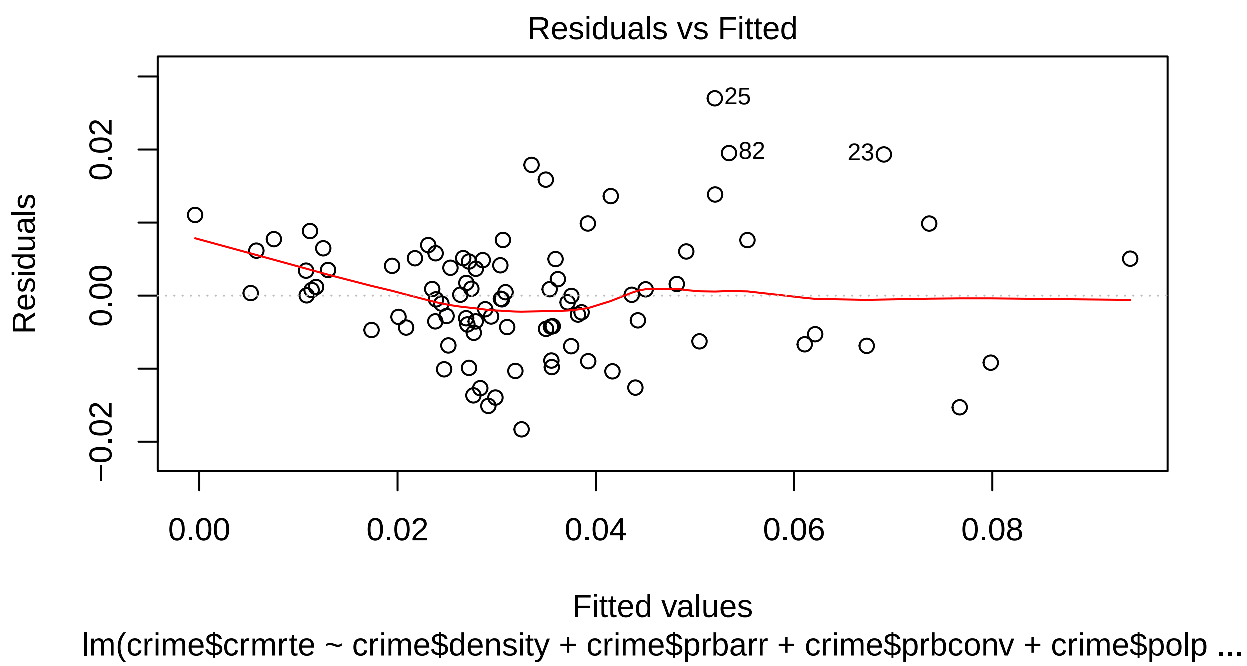
Verify Assumptions

Here we verify the the six assumptions of our model:

1) Linearity of the Parameters
2) Random Sampling
3) No Perfect Multicollinearity
4) Zero Conditional Mean
5) Homoskedasticity
6) Normality of Residuals

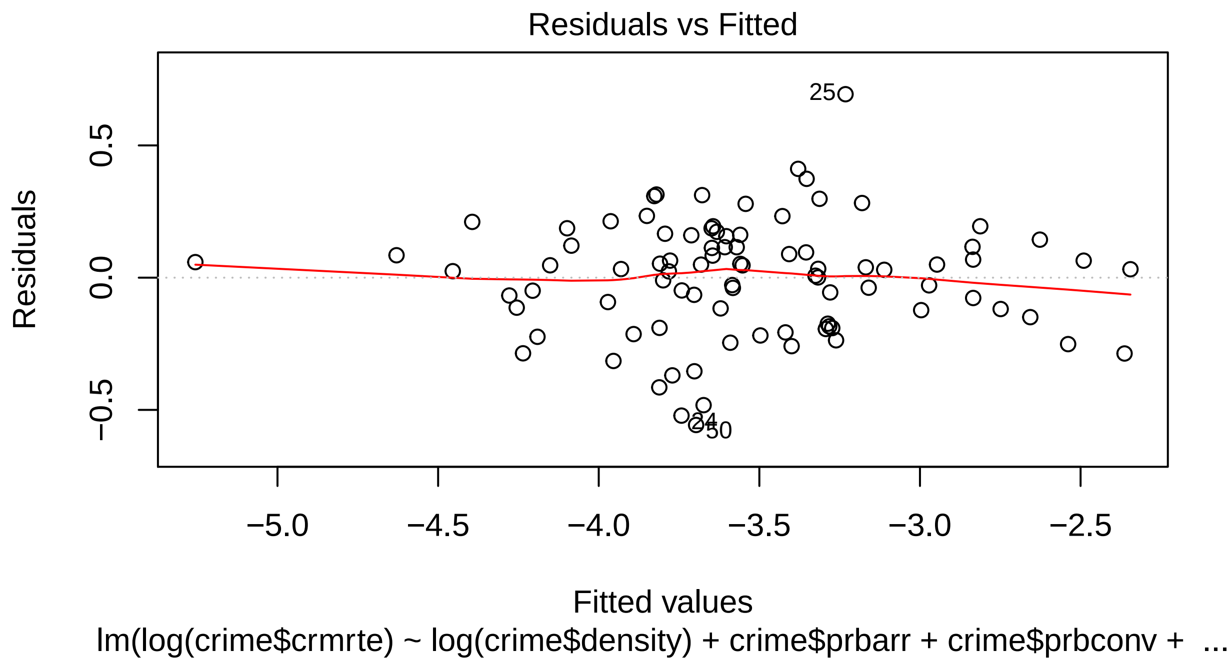
First we check for linearity by looking at the Residuals vs Fitted plot.

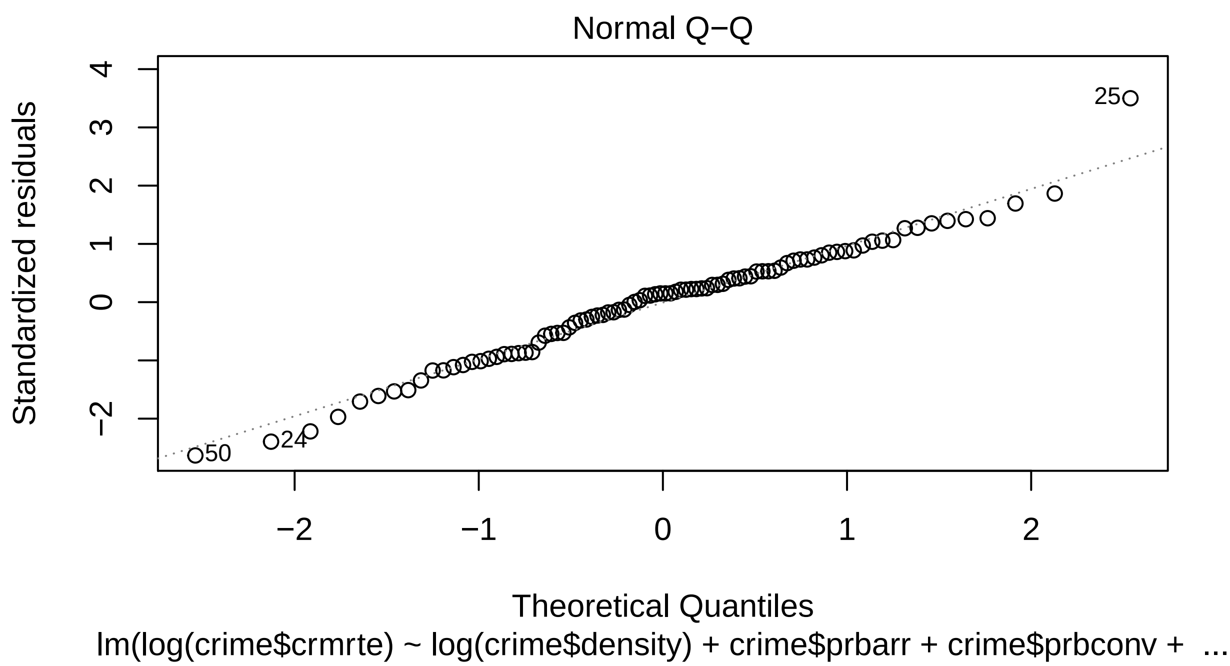
```
plot(model_2, which=1)
```

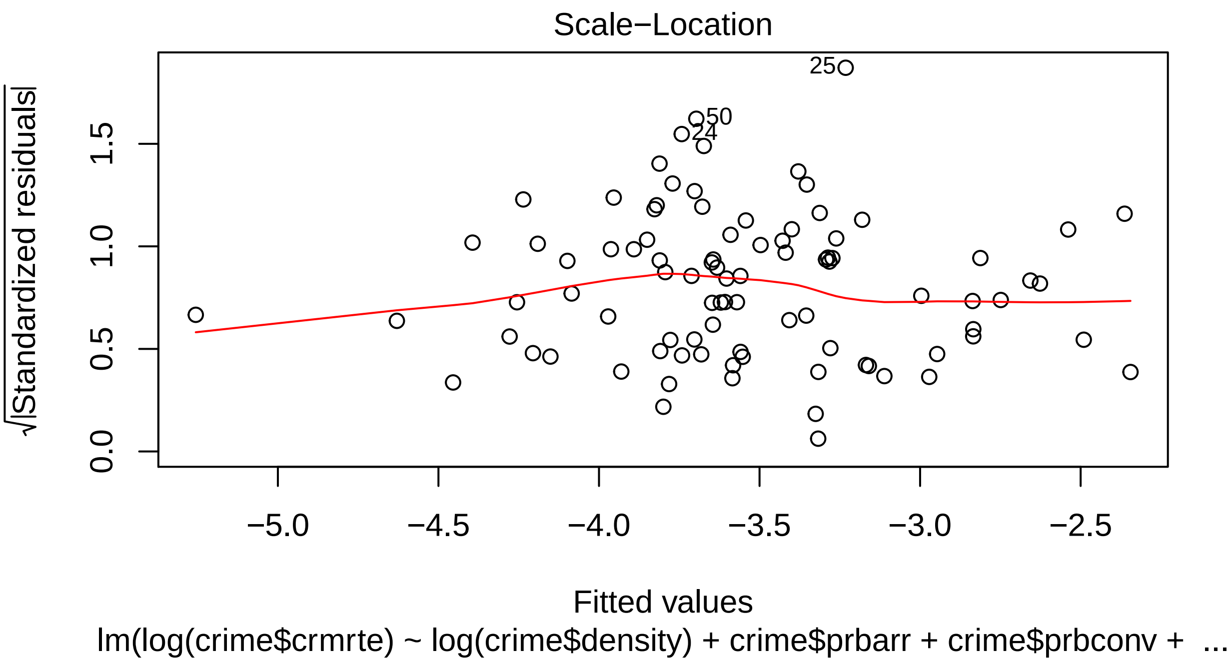


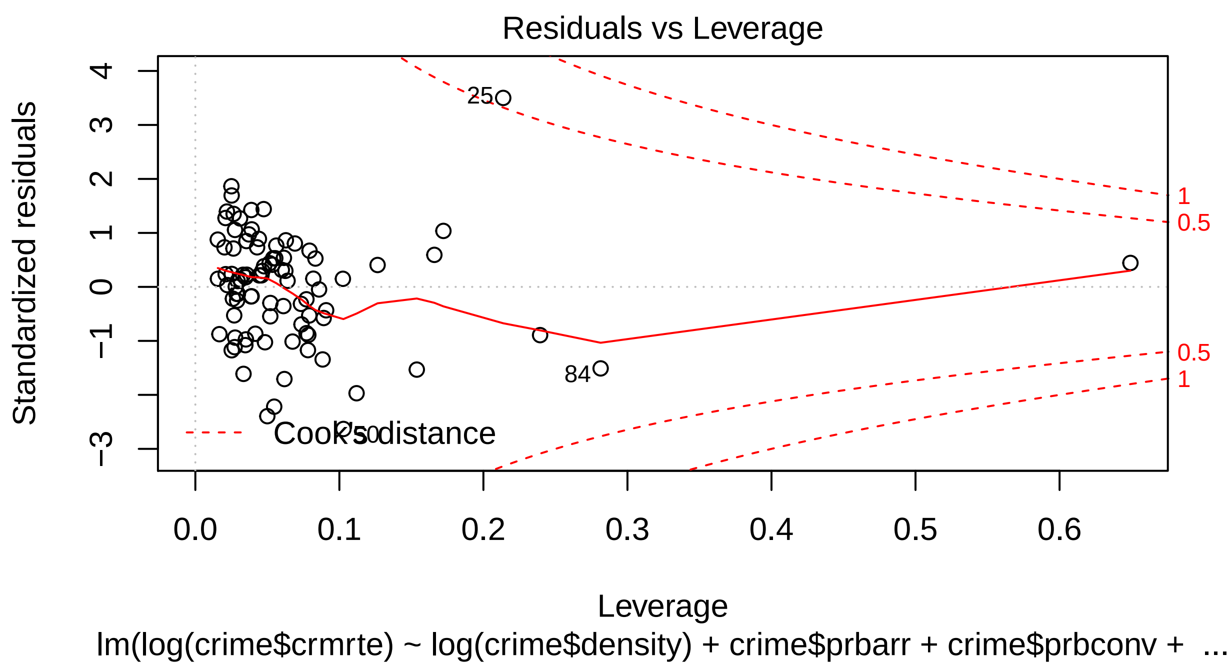
Here we see evidence of nonlinear relationship at the lower end of the range of our dependant variable. We address this by looking at the log-log relationship with respect to crimerte, density, and polpc, our variables that show the strongest evidence of skew.

```
model_f<-lm(log(crime$crmrte)~log(crime$density)+crime$prbarr+crime$prbconv+
             log(crime$polpc)+crime$pctmin80)
plot(model_f)
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
25
24
50
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5
Fitted values
lm(log(crime$crmrte) ~ log(crime$density) + crime$prbarr + crime$prbconv + ...


Normal Q−Q
Standardized residuals
4
3
2
1
0
−2
25
24
50
−2
−1
0
1
2
Theoretical Quantiles
lm(log(crime$crmrte) ~ log(crime$density) + crime$prbarr + crime$prbconv + ...





This plot is very strong evidence that the log-log transform takes care of the linearity assumption and the zero conditional mean assumption. Additionally, the normal QQ plot fits extremely well so we can safely assume we have normality of our residuals.

```r
dat_2<-data.frame(log(crime$density),log(crime$polpc),crime$prbarr,
                  crime$prbconv,crime$pctmin80)
vif(dat_2)
```

```
           Variables      VIF
1 log.crime.density. 1.578154
2   log.crime.polpc. 1.303623
3       crime.prbarr 1.420060
4      crime.prbconv 1.109106
5     crime.pctmin80 1.043042
```

To check for multicollinearity we use the measured variance inflation factors shown above. These values are sufficiently low for each of our independant variables so there is very little evidence of multicollinearity.

For the assumption of a random sample, we have to assume that the person gathering the data for the model took proper precautions to gather a truly random sample. We could gather a second sample and compare the distributions of the two samples and see how similar they are, but this would likely be costly and time consuming. For the purposes of this study, we assume that the person gathering the information used due diligence to gather a sample would truly is representative of the larger population of counties that the candidate intends to represent.

Referring back to the Residuals vs Fitted plot, we see strong evidence of f heteroskedasticity and will use robust standard errors when assessing our model from here.

```r
coeftest(model_f, vcov=vcovHC)
```

```

t test of coefficients:

                     Estimate Std. Error t value  Pr(>|t|)    
(Intercept)        -0.1056145  0.9616514 -0.1098 0.9128092    
log(crime$density)  0.2917154  0.0639652  4.5605 1.722e-05 ***
crime$prbarr       -1.6789191  0.2852187 -5.8864 7.848e-08 ***
crime$prbconv      -0.6122263  0.1029381 -5.9475 6.040e-08 ***
log(crime$polpc)    0.4546253  0.1297689  3.5033 0.0007391 ***
crime$pctmin80      0.0127636  0.0015183  8.4066 9.274e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
se.model = coeftest(model_f, vcov = vcovHC)[ , "Std. Error"]
```

Conclusions

```r
stargazer(model_f, type="text", se=list(se.model))
```

```

================================================
                         Dependent variable:    
                     ---------------------------
                               crmrte)          
------------------------------------------------
density)                      0.292***          
                               (0.064)          
                                                
prbarr                        -1.679***         
                               (0.285)          
                                                
prbconv                       -0.612***         
                               (0.103)          
                                                
polpc)                        0.455***          
                               (0.130)          
                                                
pctmin80                      0.013***          
                               (0.002)          
                                                
Constant                       -0.106           
                               (0.962)          
                                                
------------------------------------------------
Observations                     90             
R2                              0.843           
Adjusted R2                     0.834           
Residual Std. Error        0.223 (df = 84)      
F Statistic            90.516*** (df = 5; 84)   
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```