W203 Lab 3: Reducing Crime

Group 4: Chi Iong Ansjory, Tsung-Chin Han, Marcelo Queiroz

7/29/2018

Peer Feedback for Group 5

1. Introductions

The motivation of this report is clearly stated and easy to understand. However, they are found throughout the report. It would be interesting to have an introductory/context paragraph. The introduction of face-to-face crime rate comes with an interesting perspective through formula derivation from `mix` variable.

Points that were well-explained an could be part of the introduction:

- The two variables of interest and the assumption that made you transform `crmrte` and `mix` into a new variable (this was a really powerful insight, by the way).
- The focus on findings that can affect change without unintended consequences.
- The worry about not proposing experiements that are unethical.
- The point on adding the percentage of workers in each industry is really good, looking forward to see if this can be done.

2. The Initial EDA

The transformations of variables could be part of the initial or subsequent portion of the EDA analysis. Any particular reason to call the combined wage variable as `wave` instead of `wage`? Performed a sample mean without considering their weight composition may reduce influence of certain variables.

There are anomalous values identified but no further action is performed to have them adjusted or removed for EDA. Possible improvement for the last part: clearly state if the observations in section 3 resulted in data removing and why (or why not). Additionally, there is space for further analysis of each variable to see if further transformation is needed to expose linear relationships in scatterplots.

The scatterplots are a bit too small and busy to see the correlation. It is true that covariates may have casual effects, but they are more inspected for correlation rather than casual effect. Maybe a plot for each of the points raised in section 4 could help the reader to visualize your statements.

Correlation does not necessarily mean casuality. Perhaps address a bit more research question in terms of coming up a casual effect.

3. The Model Building Process

Although the six models came from reasonable assumptions and the transformations made a lot of sense, there is no explanation behind how the three models being come up with and related to the EDA. Some discussions of why the variables were chosen may help.

There are side-by-side comparisons on both outcome variables `crmrte` and `fcrmrte` with same three models for each. There is not enough explanation to understand the meaning through visualization. You can add this discussion either here or after the tables.

4. The Regression Table

More explanations are needed on how the model specifications are properly chosen to outline the boundary of reasonable choices. However, it is easy to find key coefficients in the regression table, but there is no further discussion of practical significance for key effects. Some discussion when comparing the three models for each variable may be needed to help ther reader on where to look. Also we would like to see comments on the AIC/R^2 relationship when we add more variables, since R-squared and AIC can help us understand the model predicting power. It would be helpful to see if adding variable would increase the predicting power.

5. The Omitted Variables Discussion

Did not see a section addressing this problem. We believe with the discussion around the regression tables and their similarities and differences can drive some insights on that.

6. Conclusion

The conclusion addresses the big-picture concerns about the political campaign in a qualitative manner. Interesting points are identified on police per capita against crime rate. A good point to add here is some discussion about the statistical versus the practical outcomes of our studies, since the coefficients are really low.

7. Throughout the Report

No significant errors, faulty logic, or unpersuasive writing that leads to less convincing conclusions.

Overall your report is clean and direct, with a lot of good insights. Most of our ideas in this peer review are related to add more narrative sections making the reading easier for people with no statistical background as well.

Following attaches report of Group 5 Lab 3 Part 1.

Lab 3 - Part 1

Tina, Debalina, Mark & Vivek

July 22, 2018

1. What do you want to measure?

This report discusses determinants of crime in counties across North Carolina. The data contains geographic, demographic, economic, and crime data. The data was collected by the campaign in 1987. The data is reported on a county level basis. We will explore the drivers of the crime rate in terms of the number of crimes committed per person.

We believe there are **2** variables of interest:

A. Overall crime rate as captured by the **crmrte** variable.
B. Face to face crime rate as captured by combination of **crmrte** and **mix** variables. We assume that the face to face variety of crime is more worrisome than the other variety. Ratio of face to face crime rate is defined as:

$$\text{Note: } mix = \frac{face}{other}$$

$$face = \frac{face}{total}$$

$$face = \frac{face}{face + other}$$

$$face = \frac{\frac{face}{other}}{\frac{face}{other} + \frac{other}{other}}$$

$$face = \frac{mix}{mix + 1}$$

$$face\ to\ face\ crime\ rate = crime\ rate * \frac{mix}{mix + 1}$$

2. What transformations should you apply to each variable?

We did the following transformations to the data:
- Removed the 'year' attribute
- Removed NAs
- Certain numeric attributes were imported as strings, so those were converted data to numeric
- Combined the wage variables as we noticed multicollinearity between them. We did a simple mean for now but a better solution could be weight them by sector employment. If time allows, we will research the sector employments in North Carolina to arrive at appropriate weights.
- Region was specified in 3 different columns. These were coalesced into a **region** factor variable.

R code for transformations:

```
df$year<-NULL
df<-df[!is.na(df$county),]
df<-data.frame(sapply(df, as.numeric))
df$wave<-(df$wcon+df$wtuc+df$wtrd+df$wfir+df$wser+df$wmfg+df$wfed+df$wsta+df$wloc)/9
df$region<-factor(ifelse(df$west==1, "west", ifelse(df$central ==1, "central", "other")))
```

```
df$urban<-factor(ifelse(df$urban == 1, "urban", "not urban"))
df$fcrmrte<-df$crmrte*(df$mix/(df$mix + 1))
```

3. Are your choices supported by EDA?

We have following observations from EDA:

1. Certain variables show some skew. In subsequent analysis, we plan to iteratively transform to see if the transformation improves goodness of fit. For example: prbarr, prbconv, polpc, taxpc, density, urban, pctymle
2. We see outliers in certain variables. These outliers need to be studied to establish if these could be typos or if they tell a different story. For example:

a. probability > 1 in prbconv and in prbarr

b. wser is approximately 10 times the average in one of the counties

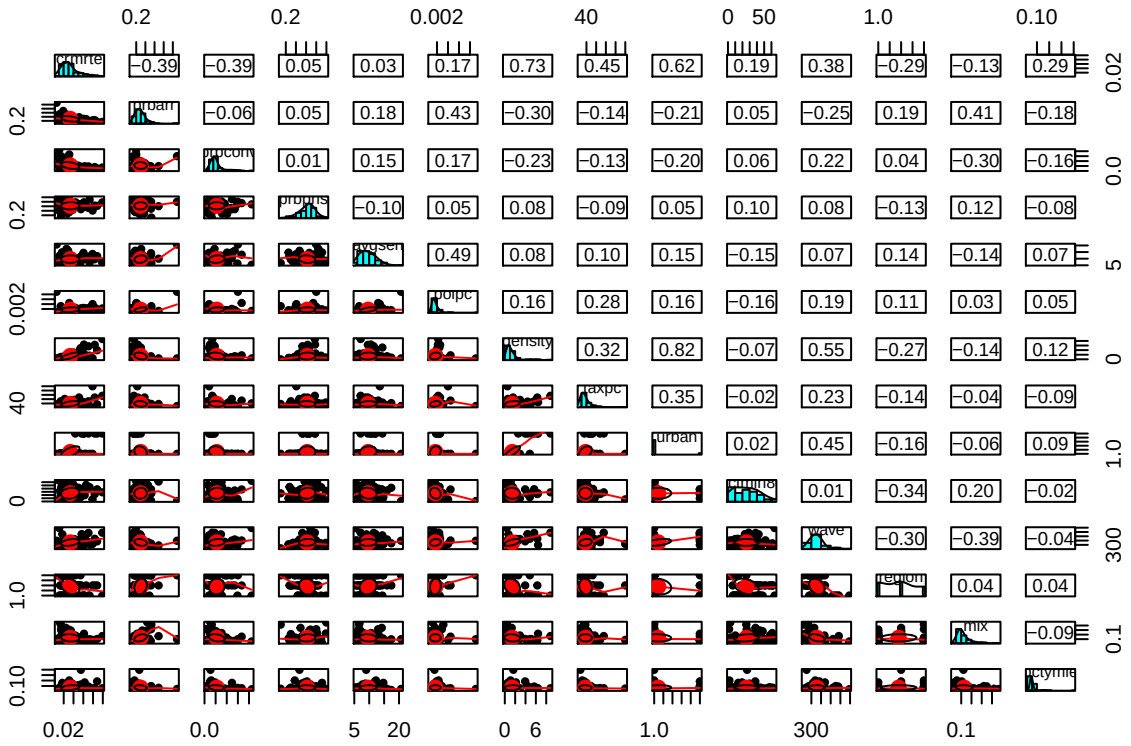
c. pctymle is exceptionally high for one of the counties.

d. crmrte is very low for a county where prbarr prbconv and prbpris are very high

4. What covariates help you identify a causal effect?

Checking pairwise correlation helps identify the covariates that that **may** have causal effect.

```
pairs.panels(df[,c("crmrte", "prbarr", "prbconv","prbpris","avgsen",
                   "polpc","density","taxpc","urban" ,"pctmin80","wave",
                   "region","mix","pctymle")],
             method="pearson",density = T, cex.cor = 1.5)
```



Looking at the correlations we pick the covariates, which show high correlation with crime rate leaving aside the covariate that seem to be collinear with other covariates. For example both density and urban have a high correlation with crime rate but they seem to be also highly correlated with themselves. Using this argument, we start our analysis with following covariates:

- density which seems to be strongly positively correlated with crime rate.
- prbarr which seems to be negatively correlated with crime rate.
- prbconv which seems to be positively correlated with crime rate.
- wages which is average wave and seems to be positively correlated with crime rate.

5. Linear Regression Models

We have done 6 models. 3 models with crime rate as the dependent variable and 3 more for face to face crime rate.

Model 1

Regression of Crime Rate vs population density, Probability of arrest, probability of conviction and average wage.

```
model1df<-df %>% select(crmrte,
                        density,
                        prbarr,
                        prbconv,
                        wave)
```

```
model1<-lm(formula = crmrte~density+prbarr+wave+prbconv, data = model1df)
```

Model 2

In addition to covariate in model1, we include Region, per capita tax revenue and percent young male.

```
model2df<-df %>% select(crmrte,
                        density,
                        prbarr,
                        prbconv,
                        wave,
                        region,
                        taxpc,
                        pctymle)
```

```
model2<-lm(formula = crmrte~density+prbarr+wave+prbconv+region+taxpc+pctymle, data = model2df)
```

Model 3

Regression of Crime rate with all variables.

```
model3df<-df %>% select(crmrte,
                        prbarr,
                        prbconv,
                        prbpris,
                        avgsen ,
                        polpc ,
                        density,
                        taxpc ,
                        urban ,
                        pctmin80,
                        wave,
                        region,
                        mix ,
                        pctymle)
```

```
model3<-lm(formula = crmrte~prbarr+prbconv+prbpris+avgsen+polpc+density+
             taxpc+urban+pctmin80+wave+region+mix+pctymle, data = model3df)
```

Comparison of the 3 regression models to predict crime rate

```
stargazer(model1, model2, model3, type = 'latex', report = "vc",
          add.lines=list(c("AIC", round(AIC(model1)), round(AIC(model2)), round(AIC(model3))),
                         c("BIC", round(BIC(model1)), round(BIC(model2)), round(BIC(model3)))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 22, 2018 - 6:22:21 PM

Table 1:

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
density	0.007	0.006	0.006
prbarr	−0.031	−0.022	−0.049
wave	0.00004	0.0001	0.00001
prbconv	−0.015	−0.014	−0.020
prbpris			0.006
avgsen			−0.0004
polpc			6.788
regionother		0.008	0.005
regionwest		−0.004	−0.0004
mix			−0.022
taxpc		0.0003	0.0002
urbanurban			−0.001
pctmin80			0.0003
pctymle		0.121	0.074
Constant	0.029	−0.004	0.018
AIC	-545	-578	-599
BIC	-530	-553	-559
Observations	91	91	91
R^2	0.632	0.766	0.838
Adjusted R^2	0.615	0.743	0.808
Residual Std. Error	0.012 (df = 86)	0.010 (df = 82)	0.008 (df = 76)
F Statistic	36.906*** (df = 4; 86)	33.576*** (df = 8; 82)	28.006*** (df = 14; 76)
Note:			*p<0.1; **p<0.05; ***p<0.01

Model 1.1

Regression of Face to face Crime Rate vs population density, Probability of arrest, probability of conviction and average wage.

```
model1.1df<-df %>% select(fcrmrte,
                      density,
                      prbarr,
                      prbconv,
                      wave)
```

```
model1.1<-lm(formula = fcrmrte~density+prbarr+wave+prbconv, data = model1.1df)
```

Model 2.1

In addition to covariate in model1.1, we include Region, per capita tax revenue and percent young male.

```
model2.1df<-df %>% select(fcrmrte,
                      density,
                      prbarr,
                      prbconv,
                      wave,
                      region,
                      taxpc,
                      pctymle)
```

```
model2.1<-lm(formula = fcrmrte~density+prbarr+wave+prbconv+region+taxpc+pctymle, data = model2.1df)
```

Model 3.1

Regression of Face to face Crime rate with all variables.

```
model3.1df<-df %>%
            select(fcrmrte,
                  prbarr,
                  prbconv,
                  prbpris,
                  avgsen ,
                  polpc ,
                  density,
                  taxpc ,
                  urban ,
                  pctmin80,
                  wave,
                  region,
                  pctymle)
```

```
model3.1<-lm(formula = fcrmrte~prbarr+prbconv+prbpris+avgsen+polpc+density+
               taxpc+urban+pctmin80+wave+region+pctymle, data = model3.1df)
```

Comparison of the 3 regression models to predict face to face crime rate

```
stargazer(model1.1, model2.1, model3.1, type = 'latex', report = "vc",
          add.lines=list(c("AIC", round(AIC(model1.1)), round(AIC(model2.1)), round(AIC(model3.1))),
                         c("BIC", round(BIC(model1.1)), round(BIC(model2.1)), round(BIC(model3.1)))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 22, 2018 - 6:22:21 PM

Table 2:

	Dependent variable:		
	fcrmrte		
	(1)	(2)	(3)
density	0.001	0.001	0.001
prbarr	0.0003	0.001	−0.002
wave	−0.00001	−0.00001	−0.00002
prbconv	−0.002	−0.002	−0.002
prbpris			0.003
avgsen			−0.0001
polpc			0.608
regionother		0.001	−0.0001
regionwest		−0.001	−0.00004
taxpc		0.00002	0.00002
urbanurban			−0.001
pctmin80			0.0001
pctymle		0.003	0.003
Constant	0.007	0.005	0.005
AIC	-889	-901	-921
BIC	-874	-876	-883
Observations	91	91	91
R^2	0.460	0.568	0.687
Adjusted R^2	0.435	0.526	0.635
Residual Std. Error	0.002 (df = 86)	0.002 (df = 82)	0.001 (df = 77)
F Statistic	18.291*** (df = 4; 86)	13.496*** (df = 8; 82)	13.023*** (df = 13; 77)
Note:			*p<0.1; **p<0.05; ***p<0.01

6. Conclusions

As a political organization we want to focus on variables that substantially impact our response variables, but also on which we can affect change. We also want to be mindful of unintended consequences of change. While reducing percent minority, for example, seems to reduce both crime rate and face-to-face crime rate, it will likely have unintended consequences toward diversity and culture. It is also quite hard to quantify what changes in each of these variables would cost, both monetarily and in tradeoffs to society. That being said, our conclusions will remain largely qualitative.

Probability of arrest, probability of conviction, and average sentence are all negatively correlated with our response variables. This would suggest that focusing on enforcement in our judicial system could promise lower crime rates. Even if one or all of these variables are confounded, improvements to the judicial system seem to be the most promising.

Interestingly, police per capita is highly correlated with crime rate and face-to-face crime rate. This is almost certainly not causal in nature, based on intuition. It is much more reasonable that high crime rates *require* more police force and the direction of causality goes in the other direction. It would be very interesting to run an experimental setup with police distribution, perhaps an AB test, to see its causal effect on our response variables. However, we should also keep in mind the ethics behind this type of experiment and not leave a community at risk as we try to determine optimal distribution of law enforcement.