Race, Experiences, and Criminal Sentencing

Julie Norman, Abraham Alhomadi, and Keaton Manwaring

12/6/2020

In this paper, we will undertake an econometric examination using the Crime1 dataset in the Wooldrige Package. The dataset contain a total of 2725 observations on 16 distinct variables.It was collected in 1986, in the State of California. Looking at the data, we will aspire to answer the following research question that is the focus of this paper: "How does an individual's income affect average sentencing outcome when they commit a crime? This question has been the subject of policy research for decades and, using the Crime1 dataset, we would like to contribute to the conversation on this issue of economics and criminal justice reform. Our belief as citizens is that we have a collective responsibility to promote better policies and the well-being of our nation through the efficient and just allocation of resources and ensuring that our criminal justice system is in fact "just" in its treatment of citizens who are convicted of a crime. In addition to our focus research question, we will investigate other equally critical questions, some of which include,"Does a person's employment status increase or decrease average sentencing?", " Do blacks get different average sentence lengths than nonblacks?", and "how does an individual's prior criminal history, if any, contribute to their average sentencing length?".

Now, to carry out our investigation, we will begin by focusing our annalysis on the the 134 individuals in our dataset who have been sentenced. We're filtering out the individuals who have not served prison time as our question is focused on what factors impact sentence length, not whether they've been found guilty at all. We'll regress our dependent variable "avgsen", which is the average sentence length of a convicted criminal, on variant explanatory variables: inc86 (legal income), nfarr86 (felony arrests), pcnv (proportion of prior convictions), durat (recent unemp duration), narr86(times arrested), Black, and Hispanic. Since our data treats black and hispanic as mutually exclusive, we're going to combine our race variable into a single dummy that illustrates whether an individual is black, hispanic, or neither. Our multiregression model would start with the following model and build on from there to test for interaction between variables and to reflect for possible increase or decrease in sentencing length in the short- and long-term. This is our original regression model:

The additional questions will help put our focus question in perspective, and allow us to identify alternative explanations besides an individual's income to why average sentencing length would differ from one person to another, given variables like employment status, criminal history, and race.

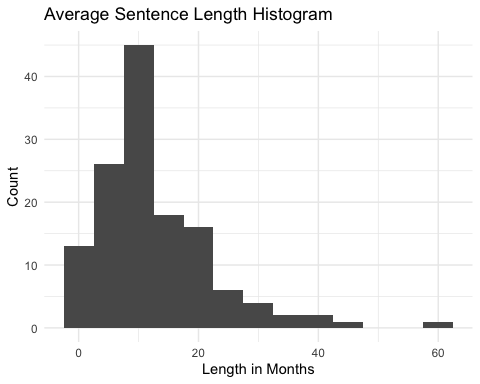
Lastly, for so long, policy makers have been trying to answer the above questions in order to improve economic and criminal justice policy making. They endeavor to identify our criminal justice system flaws and seek to address them, including issues like sentencing discrimination on the basis of income and race, and to improve the efficient and fair allocation of economic resources to all their citizens. This is an interesting economic inquiry, for it touches the heart of one of the challenges that has engulfed our nation. A stable and functioning economic system and a fair criminal justice system are important to maintaining a healthy life style, public safety to guard against social division because of discriminatory economic and criminal justice motives, and to promote a flourishing, economically strong country in the 21st century. For this reason, we are seeking to investigate whether average sentencing in, say, California or other states for that matter, are rendered based on effective, just, and fair laws, or are there discriminatory motives that impact our justice system rendering of sentencing. If discrimination does exist, then as citizens, we have a solemn obligation to put forth facts backed up by objective data and partake in making our country, as our founding fathers envisioned, a "more perfect union."

Our Crime1 dataset is included in the wooldridge package. We did not collect the data on our own nor do we know precisely how the data was collected. However, we know that the data it contains was collected in the State of California and the arrests data are for the year 1986. The observations reflect men who were born in 1960 or 1961, and the subject men in the dataset all had criminal history and a minimum of one arrest before 1986. So, our regression model, as written above, will provide a multidimensional analysis to explain whether our independent variables listed do or do not have a causal relationship with our dependent variable, Avgsen. We wanted to include gender as a dummy variable in our regression model, but the dataset only contains data on men. Therefore, instead of looking at the relationship between gender and average sentencing, we will dwell on race and average sentencing.

## Data Exploration

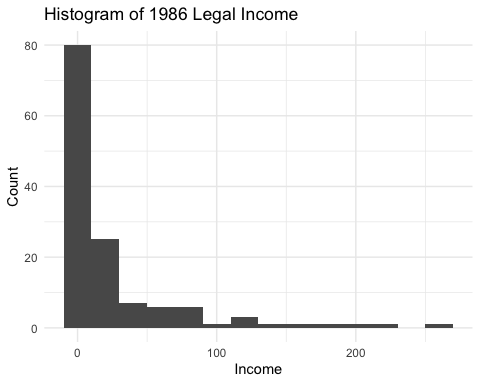
In order to better understand our data we'll look at each variable to check for any violations of the first Gauss Markov assumption: linear parameters. While there's logical expanations for relationships between our explanatory variable and our response variables, we also need to ensure that our data bares out that relationship.

### Average Sentence Length



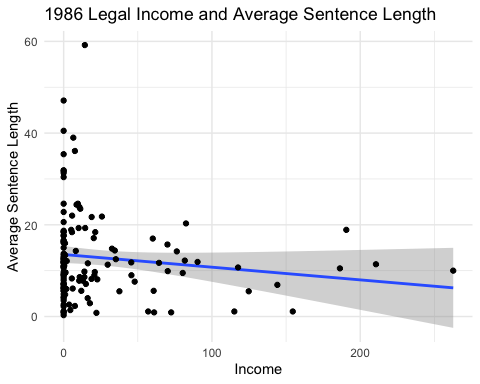
The histogram shows that that the distribution of sentence length is rightly skewed indicating that this model is proabably going to have issues with heteroskedasticity and that a log might be an appropriate adjustment for our model. However, since there are zero values in the variable a log transformation isn't appropriate. It also has a range of 58.9 which shows that this paramater has enough varition that we can use it in the model.

### Income from 1986



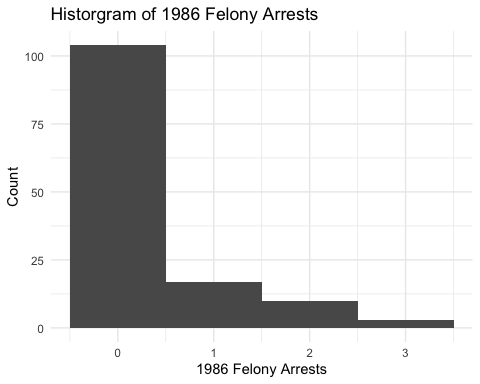
The histogram shows a strong right hand skew indicating that in order to meet the linear parameters assumption, we will need to take the log of the variable. However, since there are so many 0 values we won't be transforming the data Next we need to visualize the relationship between the income and average sentence length.

## `geom\_smooth()` using formula 'y ~ x'



The model indicates a week negative association between the income variable and the sentence variable. Due to this, we shouldn't expect a strong impact of this variable in our model.

### Felony arresting

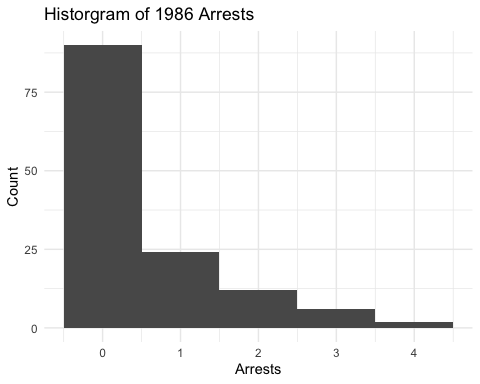


The histogram indicates a strong right hand skew in felony arrests. However, the ordinal nature of the data makes it unideal for a multiple regression model.

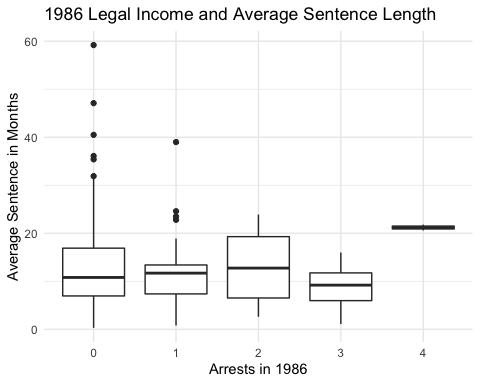


The side by side boxplots illustrates that amoung the average sentence length for each number of felony arrests has essentially no association on average. Due to this and the ordinal nature of the variable we should consider removing this variable from our model.

### Number of Arrests in 1986

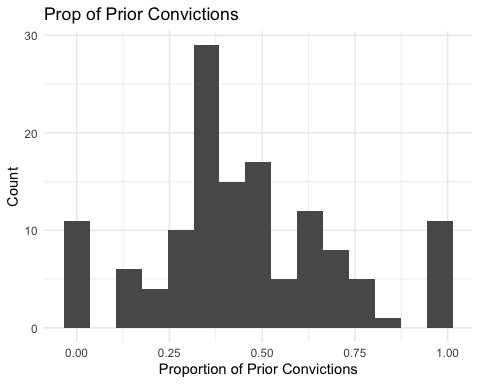


The histogram indicates a strong right hand skew in arrests. However, the ordinal nature of the data makes it again unideal for a multiple regression model.



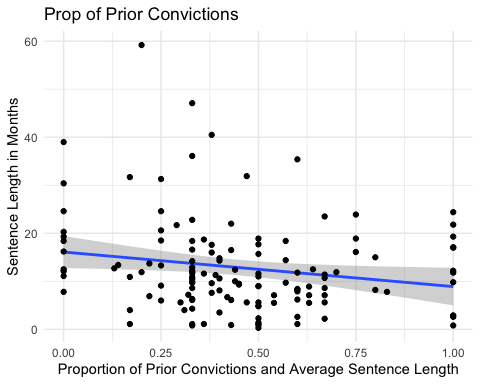
Again, he side by side boxplots indicates that amoung the average sentence length for each number of arrests there is essentially no association. While the average sentence for those with 4 felony arrests is higher than the rest, this is merely one data point and as it does not imply an overall trend in the data it's unlikely to be influencial in our model. Due to this and the ordinal nature of the variable we should consider removing this variable from our model.

### Proportion of prior convictions



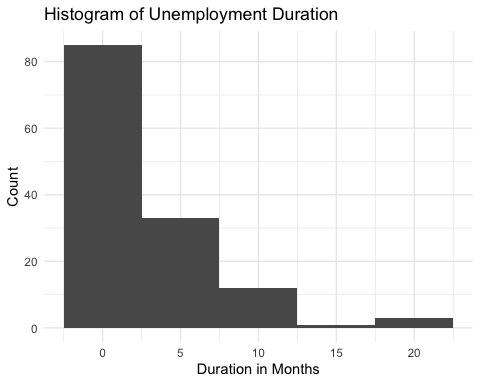
The histogram indicates that the distribution of th proportion of prior convictions is approximately normal, indicating that this variable is a good candidate for use in a multiple regression model.

## `geom\_smooth()` using formula 'y ~ x'



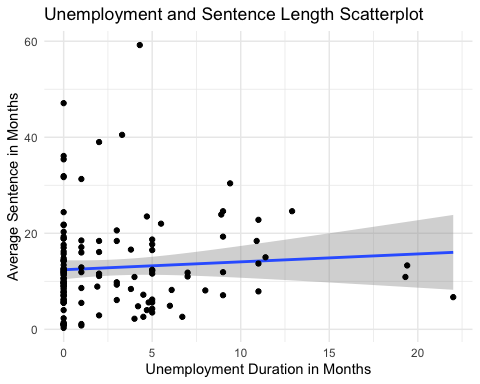
This scatterplot indicates a negative association between prior convictions and sentence length. This indicates that as convictions become more common for an individual, their sentence length decreases. This trend is particularly interesting as it does fit a typically assumption of repeat offenders having harsher punishments. While we can only speculate as to the exact nature of this relationship, its possible that other variables such as the types of crimes committed and/or race have a larger impact on average sentence then we wouldn've prevously thought. That makes this variable an ideal candidate for a multiple regression since we can see how this variable combined with other indicators appropriate predict average sentence length.

### Duration of Unemployment



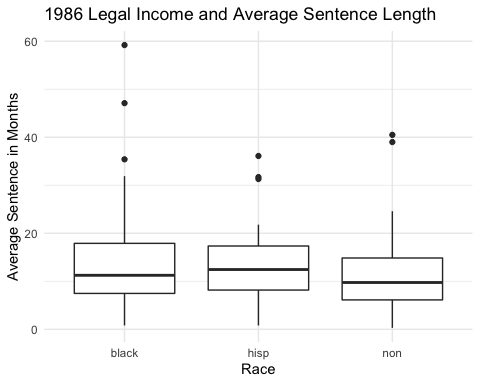
The histogram indicates a strong left hand skew in unemployment duration. This indicates that a logarithmic model might be more appropriate for this variable, however since we have zero values in our data a log isn't appropriate. The measurement unit for duration of unemployment is not given in our data set but since all the other time variables are measured in months we're assuming that here as well.

## `geom\_smooth()` using formula 'y ~ x'



This scatterplot illustrates a weak positive association between unemployment length and average sentence length. Due too the weakness of the association

### Race Comparisons



The boxplots of average sentence length show minmimal variation accross different racial categories. This indicates that even if there is a statistical significance between the regressions, it is likely of little practical significance. However, since race is a major point in our research question we'll keep this variable in our regression.

## Model Specification

Since we're considering eliminating two of the variables, it's important we include them in our models until we've determined the best form for the other variables in our model.

During our exploration of the variables, we found that due to skewness, a few of our variables could benefit from logs. However, since there are zero values in all of our variables this isn't an option. In order to be safe we'll compare a variatey of quadratic functions on the variables that weren't as skewed to see if it improves our model.

##   
## ==============================================================================================================================================  
## Dependent variable:   
## --------------------------------------------------------------------------------------------------------------------------  
## avgsen (avgsen) avgsen (avgsen) avgsen   
## (1) (2) (3) (4) (5) (6)   
## ----------------------------------------------------------------------------------------------------------------------------------------------  
## inc86 -0.019 -0.066 -0.020 -0.019 -0.016 -0.015   
## (0.019) (0.049) (0.019) (0.020) (0.019) (0.020)   
##   
## I(inc862) 0.0003   
## (0.0002)   
##   
## narr86 1.317 1.370 -2.097 1.275 0.899 1.282   
## (1.661) (1.661) (2.849) (1.673) (1.650) (1.664)   
##   
## I(narr862) 1.231   
## (0.837)   
##   
## nfarr86 -2.979 -3.095 -2.987 -1.959 -2.486 -2.830   
## (2.158) (2.160) (2.148) (4.183) (2.142) (2.170)   
##   
## I(nfarr862) -0.441   
## (1.546)   
##   
## pcnv -6.520\*\* -6.408\* -7.340\*\* -6.473\* -26.785\*\*\* -6.712\*\*   
## (3.290) (3.291) (3.322) (3.306) (10.099) (3.304)   
##   
## I(pcnv2) 20.145\*\*   
## (9.507)   
##   
## durat 0.143 0.104 0.186 0.128 0.163 0.482   
## (0.209) (0.212) (0.210) (0.216) (0.207) (0.480)   
##   
## I(durat2) -0.023   
## (0.029)   
##   
## racehisp -1.561 -1.347 -2.007 -1.575 -0.978 -1.528   
## (2.240) (2.249) (2.250) (2.249) (2.227) (2.244)   
##   
## racenon -3.119 -3.124 -3.358\* -3.078 -2.902 -3.136   
## (1.944) (1.944) (1.942) (1.957) (1.921) (1.948)   
##   
## Constant 17.777\*\*\* 18.231\*\*\* 18.684\*\*\* 17.735\*\*\* 21.255\*\*\* 17.387\*\*\*   
## (2.232) (2.274) (2.306) (2.245) (2.747) (2.290)   
##   
## ----------------------------------------------------------------------------------------------------------------------------------------------  
## Observations 134 134 134 134 134 134   
## R2 0.085 0.093 0.101 0.086 0.117 0.090   
## Adjusted R2 0.034 0.035 0.043 0.027 0.060 0.031   
## Residual Std. Error 9.511 (df = 126) 9.508 (df = 125) 9.467 (df = 125) 9.546 (df = 125) 9.382 (df = 125) 9.525 (df = 125)   
## F Statistic 1.673 (df = 7; 126) 1.600 (df = 8; 125) 1.748\* (df = 8; 125) 1.464 (df = 8; 125) 2.066\*\* (df = 8; 125) 1.537 (df = 8; 125)  
## ==============================================================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Since the regressions 2, 3 and 5 all saw an increase in and then we'll compare those and a mulitple combinations to the original next.

reg3A <- lm((avgsen) ~ inc86 + I(inc86^2) + narr86 + I(narr86^2) + nfarr86 + pcnv + durat + race, df)  
reg3B <- lm((avgsen) ~ inc86 + I(inc86^2) + narr86 + nfarr86 + pcnv + I(pcnv^2) + durat + race, df)  
reg3C <- lm(avgsen ~ (inc86) + narr86 + narr86sq + nfarr86 + pcnv + pcnvsq + durat + race, df)  
reg3D <- lm(avgsen ~ (inc86) + I(inc86^2) + narr86 + I(narr86^2) + nfarr86 + I(nfarr86^2) + pcnv + durat + race, df)  
  
stargazer(reg1, reg3A, reg3B, reg3C, reg3D, type = "text")

##   
## ============================================================================================================================  
## Dependent variable:   
## --------------------------------------------------------------------------------------------------------  
## avgsen (avgsen) avgsen   
## (1) (2) (3) (4) (5)   
## ----------------------------------------------------------------------------------------------------------------------------  
## inc86 -0.019 -0.065 -0.070 -0.017 -0.066   
## (0.019) (0.049) (0.048) (0.019) (0.049)   
##   
## I(inc862) 0.0002 0.0003 0.0003   
## (0.0002) (0.0002) (0.0002)   
##   
## narr86 1.317 -1.978 0.943 -2.480 -3.260   
## (1.661) (2.852) (1.647) (2.816) (3.114)   
##   
## I(narr862) 1.207 1.610\*   
## (0.838) (0.925)   
##   
## narr86sq 1.220   
## (0.826)   
##   
## nfarr86 -2.979 -3.097 -2.598 -2.495 0.916   
## (2.158) (2.151) (2.140) (2.132) (4.469)   
##   
## I(nfarr862) -1.736   
## (1.695)   
##   
## pcnv -6.520\*\* -7.216\*\* -27.445\*\*\* -27.502\*\*\* -7.297\*\*   
## (3.290) (3.324) (10.096) (10.063) (3.325)   
##   
## I(pcnv2) 20.929\*\*   
## (9.512)   
##   
## pcnvsq 20.051\*\*   
## (9.462)   
##   
## durat 0.143 0.149 0.119 0.206 0.105   
## (0.209) (0.214) (0.209) (0.208) (0.218)   
##   
## racehisp -1.561 -1.792 -0.710 -1.423 -1.992   
## (2.240) (2.261) (2.234) (2.237) (2.269)   
##   
## racenon -3.119 -3.359\* -2.900 -3.141 -3.279\*   
## (1.944) (1.942) (1.917) (1.919) (1.944)   
##   
## Constant 17.777\*\*\* 19.101\*\*\* 21.910\*\*\* 22.138\*\*\* 19.242\*\*\*   
## (2.232) (2.343) (2.795) (2.798) (2.347)   
##   
## ----------------------------------------------------------------------------------------------------------------------------  
## Observations 134 134 134 134 134   
## R2 0.085 0.108 0.127 0.132 0.115   
## Adjusted R2 0.034 0.043 0.064 0.069 0.043   
## Residual Std. Error 9.511 (df = 126) 9.467 (df = 124) 9.365 (df = 124) 9.338 (df = 124) 9.465 (df = 123)   
## F Statistic 1.673 (df = 7; 126) 1.665 (df = 9; 124) 2.004\*\* (df = 9; 124) 2.096\*\* (df = 9; 124) 1.604 (df = 10; 123)  
## ============================================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Since the is highest in regression 4 we'll use that model and double check the form later. For now we can consider the following regression our current model

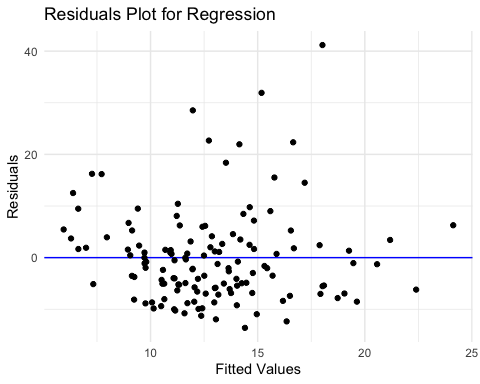
## Model Assumptions

We used an Ordinary Least Square (OLS) estimating method to compute the coefficients of our parameters. Our model complies with all five of OLS multiple regression assumptions. First, our regression model is linear in parameters, as illustrated by our model above. Additionally, the sampling of the observations was undertaken randomly so we've met that condition as well. The third condition is that there's no multicollinearity. Since the independent variables is 93.6621494 which is greater than zero and we have no variable repetition so we just need to check that we have no variables that are perfect multiples of each other

## $`Regress on inc86`  
## GVIF Df GVIF^(1/(2\*Df))  
## narr86 10.612666 1 3.257709  
## I(narr86^2) 8.029370 1 2.833614  
## nfarr86 3.553686 1 1.885122  
## pcnv 9.803458 1 3.131047  
## I(pcnv^2) 9.787451 1 3.128490  
## durat 1.065149 1 1.032060  
## race 1.147198 2 1.034927  
##   
## $`Regress on nfarr86`  
## GVIF Df GVIF^(1/(2\*Df))  
## inc86 1.178774 1 1.085713  
## narr86 8.115297 1 2.848736  
## I(narr86^2) 8.036588 1 2.834888  
## pcnv 9.725978 1 3.118650  
## I(pcnv^2) 9.713519 1 3.116652  
## durat 1.150120 1 1.072437  
## race 1.135812 2 1.032349  
##   
## $`Regress on narr86`  
## GVIF Df GVIF^(1/(2\*Df))  
## inc86 1.174464 1 1.083727  
## I(narr86^2) 2.733244 1 1.653252  
## nfarr86 2.707499 1 1.645448  
## pcnv 9.730195 1 3.119326  
## I(pcnv^2) 9.789315 1 3.128788  
## durat 1.152139 1 1.073377  
## race 1.177860 2 1.041774  
##   
## $`Regress on pcnv`  
## GVIF Df GVIF^(1/(2\*Df))  
## inc86 1.172175 1 1.082671  
## narr86 10.512828 1 3.242349  
## I(narr86^2) 8.017923 1 2.831594  
## nfarr86 3.505865 1 1.872396  
## I(pcnv^2) 1.042230 1 1.020897  
## durat 1.158303 1 1.076245  
## race 1.181577 2 1.042595  
##   
## $`Regress on durat`  
## GVIF Df GVIF^(1/(2\*Df))  
## inc86 1.082958 1 1.040653  
## narr86 10.585005 1 3.253460  
## I(narr86^2) 7.876173 1 2.806452  
## nfarr86 3.525278 1 1.877572  
## pcnv 9.849416 1 3.138378  
## I(pcnv^2) 9.810524 1 3.132176  
## race 1.165700 2 1.039074  
##   
## $`Regress on Black`  
## inc86 narr86 I(narr86^2) nfarr86 pcnv I(pcnv^2)   
## 1.134767 10.561517 7.979743 3.554117 9.833586 9.801718   
## durat hispan   
## 1.141349 1.111375   
##   
## $`Regress on Hispanic`  
## inc86 narr86 I(narr86^2) nfarr86 pcnv I(pcnv^2)   
## 1.174126 10.563169 8.004090 3.418493 9.835038 9.768656   
## durat black   
## 1.141275 1.077872

For our vif output, values between 1 and 5 are considered moderately correlated while values greater than 5 near perfect collinearity. From the output we can see there's strong correlation between inc86 and narr86, because of this we need to remove one of the variables entirely from the regression. Since we're focused oon jail time and non felony arrests can lead to prison time so we're going to eliminate nfarr86 and keep narr86 and its square.

Otherwise the highest variance inflation factor of all the parameter regressions is 3.555754 while most are between one and two. Since our strongest correlation is a moderate correlation we've shown there's no multicollinearity and we've met this assumption. The next assumption is the zero conditional mean assumption ie . We've already added logs to the model where appropriate and since the visualizations of our variables don't indicate a need for quadratics in our model, we can double check our model specificatioon by looking at the model residuals.



Since the residuals are more clustered below it indicates we may have violated our assumption. In order to double check we'll perform a RESET.

resettest(reg1, power = 2)

##   
## RESET test  
##   
## data: reg1  
## RESET = 1.7599, df1 = 1, df2 = 125, p-value = 0.187

resettest(reg4, power = 2)

##   
## RESET test  
##   
## data: reg4  
## RESET = 2.8921, df1 = 1, df2 = 124, p-value = 0.09152

The RESET test indicates that we can reject the null hypothesis at the 90% confidence level but not at the 95% confidence level. Normally we'd rework the model but since we already know that logs are likely needed for an ideal model and we know from the RESET test on reg1 that our model has improved so we'll keep reg4 as is.

The final assumption is homoskedasticity, in otherwords that . Our visualizations don't indicate any clear signs of homoskedasticity but to verify this we'll conduct a breusch-pagan test and a white test. The hypothesis test is as follows: $ H\_{0}: Var(u|inc, narr86, nfarr86, pcnv, durat, race) = Var(u|X) = ^2$ and the alternative where these values are not equal.

First we will use the BP test using the F statistic, then White test using F statistic, and both tests using LM ratio. Then, we will discuss the difference in the results they render. Keep in mind that a large F statistic or a large Lagrange multiplier statistic, (LM) indicate a rejection of the null hypothesis.

##   
## studentized Breusch-Pagan test  
##   
## data: reg4  
## BP = 6.2411, df = 8, p-value = 0.6202

Using the BP test, we see that our p-value is relatively large (p-value = 0.197) meaning that our F-statistic is relatively small. Consequently, we fail to reject the null hypothesis. This indicates that heteroskedasticity isn't a concern. However, in order to be thorough we'll also run a white test which is a less rigouros test for homoskedasticity.

##   
## Call:  
## lm(formula = u\_hat^2 ~ y\_hat + I(y\_hat^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -124.09 -73.81 -49.02 -1.16 1577.33   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -89.8107 188.3449 -0.477 0.634  
## y\_hat 19.0713 27.7237 0.688 0.493  
## I(y\_hat^2) -0.4178 0.9874 -0.423 0.673  
##   
## Residual standard error: 196.5 on 131 degrees of freedom  
## Multiple R-squared: 0.01815, Adjusted R-squared: 0.003164   
## F-statistic: 1.211 on 2 and 131 DF, p-value: 0.3012

We have an F statistic of 1.211 and a p value of 0.3012 >.05. Therefore we fail to reject the null hypothesis indicating that despite our earlier concerns, heteroskedasticity is not an issue. Due to the results of the bp and white test won't make any standard error adjustments to our model.

The regression as it stands now is

stargazer(reg4, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## avgsen   
## -----------------------------------------------  
## inc86 -0.017   
## (0.019)   
##   
## narr86 -4.089\*   
## (2.461)   
##   
## narr86sq 1.217   
## (0.827)   
##   
## pcnv -28.877\*\*\*   
## (10.009)   
##   
## pcnvsq 21.256\*\*   
## (9.420)   
##   
## durat 0.184   
## (0.207)   
##   
## racehisp -0.908   
## (2.196)   
##   
## racenon -3.127   
## (1.921)   
##   
## Constant 22.396\*\*\*   
## (2.794)   
##   
## -----------------------------------------------  
## Observations 134   
## R2 0.122   
## Adjusted R2 0.066   
## Residual Std. Error 9.352 (df = 125)   
## F Statistic 2.180\*\* (df = 8; 125)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

From the regression 4 output we can see our constant, pcnv and narr86 are all statistically significant. In order to determine if we should include the statistically insignificant variables we'll run a hypothesis test with the following null and alternative.

While is statistically insignificant we haven't included it in the our test given that is statistically significant and we already know that including the square is the correct form.

## Linear hypothesis test  
##   
## Hypothesis:  
## inc86 = 0  
## durat = 0  
## racehisp = 0  
## racenon = 0  
##   
## Model 1: restricted model  
## Model 2: avgsen ~ (inc86) + narr86 + narr86sq + pcnv + pcnvsq + durat +   
## race  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 129 11394   
## 2 125 10932 4 462.22 1.3214 0.2657

The results of the hyothesis test has a low F value of 1.3214 with a p value of 0.32657 which indicates that these values are not jointly significant so we'll remove those variables from our model, we can see the comparison with the removed variables here

reg5 <- lm(avgsen ~ narr86 + narr86sq + pcnv + pcnvsq, df)  
reg6 <- lm(avgsen ~ pcnv + pcnvsq, df)  
  
stargazer(reg4, reg5, reg6, type = "text")

##   
## ======================================================================================  
## Dependent variable:   
## ------------------------------------------------------------------  
## avgsen   
## (1) (2) (3)   
## --------------------------------------------------------------------------------------  
## inc86 -0.017   
## (0.019)   
##   
## narr86 -4.089\* -2.967   
## (2.461) (2.405)   
##   
## narr86sq 1.217 0.950   
## (0.827) (0.811)   
##   
## pcnv -28.877\*\*\* -30.336\*\*\* -29.332\*\*\*   
## (10.009) (9.989) (9.926)   
##   
## pcnvsq 21.256\*\* 22.325\*\* 22.040\*\*   
## (9.420) (9.380) (9.348)   
##   
## durat 0.184   
## (0.207)   
##   
## racehisp -0.908   
## (2.196)   
##   
## racenon -3.127   
## (1.921)   
##   
## Constant 22.396\*\*\* 21.091\*\*\* 20.197\*\*\*   
## (2.794) (2.555) (2.405)   
##   
## --------------------------------------------------------------------------------------  
## Observations 134 134 134   
## R2 0.122 0.085 0.075   
## Adjusted R2 0.066 0.057 0.060   
## Residual Std. Error 9.352 (df = 125) 9.398 (df = 129) 9.381 (df = 131)   
## F Statistic 2.180\*\* (df = 8; 125) 3.010\*\* (df = 4; 129) 5.275\*\*\* (df = 2; 131)  
## ======================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Since narr86 becomes statstically insignificant in when the other variables are removed, we can attribute the statistical significance to mere noise in our data. Indicating our best regression is the following.

Since this equation is univariate, I'll also include a graphic

x <- 0:1  
dat <- data.frame(x, y = 20.197 - 29.332\*x + 22.040\*x^2)  
f <- function(x) 20.197 - 29.332\*x + 22.040\*x^2  
ggplot(dat, aes(x,y)) +   
 ggtitle("Estminated Relationship from Regression Six") +  
 xlab("Propotion of Prior Convictions") +  
 ylab("Average Sentence Length") +  
 geom\_point() +  
 stat\_function(fun=f, colour="red")



## Model Interpretation and Implication

Since regression 4 and regression 6 are statistically significant, we can still offer interpretive results of each variable with the caveat that the values are not statistically significant. Regression 4 predicts that for every $100 increase in income theres a decrease in average prison sentence of 0.017 months. This is not only statistically insignificant but also practically insignificant as 0.017 months corresponds to less than a day. Since narr86 is a quadratic variable we need to solve the derivative of the quadratic is so and find the derivative which is positive. The solutions to our quadratic is 0 and 3.3599014, indicating that narr86 and avgsen has a relationship that is increasing at a decreasing rate when but it has decreasing at increasing rate when . These results indicat that the first 2 arrests increase average sentence length but arrests beyond the third decrease average sentence length which implies that individuals recieve lesser punishments for multiple arrests which is surprising given the expectations for harsher punishments for repeat offenders. However, since the result isn't statistically significant, we should let this result affect our judgement.

Model four also provides an estimation for durat, for every one month increase in unemployment duration theres a 0.184 month increase in average sentence length. 0.184 corresponds to less than a week which is still of little practical significance as well as the statistical insignificance. Regression 4 also contains variables for measuring race. We have hispanic and non black, non hispanic individuals which are interpreted relative to black individuals. Hispanic individuals have an estimated average sentence 0.908 months less than black individuals and non black and non hispanic individuals have an estimated averagee sentence length 3.127 months less than black individuals. While these variables aren't statistically significant relationship, the results correspond to what we'd expect through both typical descrimination practices and crime trends by race. In other words, there are multiple explanations for why black individals have the longest average sentence length while non black non hispanic individuals have the lowest length. Next, we'll focus on our final regression with the statistically significant results. Like narr86 before, pcnv (proportion of prior convictions) is quadratic so we will need to solve the quadratic and calculate the derivative. Since pcnv is the only variable in this regression we'll include the constant when solving our regression. The derivative of this regression is so $ pcnv = 0.665 $. For , pcnv and average sentence have a decreasing at an increasing rate but when $ pcnv > 0.665 $, pcnv and average sentence have a relationship that is increasing at an increasing rate. This implies that when conviction is less likely average sentence length decreases but when convictions are more probable, average sentence length is higher. While this is statistically significant, this fits as we'd expect that individuals who are arrested with higher liklihood of conviction are more likly to have convictions that lead to higher sentence length.

In regards to our focus question, we cannot say that income has an affect on average sentencing for two reasons. First, the income coefficient was statistically insignificant. Second, our graph shows that income and average sentencing have a very small correlation. This may have been unexpected, but it is worth mentioning that our surveyed males were young adults who are 25-years-old and with low income and a prior arrest history. This makes their employment prospects grim and explains their low income. We know that if we include older males of different ages and with different income levels and no criminal history, we can see that income would in fact have an impact on average sentencing.

In addition, looking at other key questions examined in this paper, we could say that the prior arrest ratio and recent unemployment variables were statistically insignificant, both individually and jointly. Our joint hypothesis proved that income, prior arrest ratio, and recent employment variables to be statistically insignificant, therefore their betas are unreliable. Consequently, we cannot say anything about whether recent unemployment and prior arrest ratio have an impact on average sentencing. Lastly, it was expected that an individual with prior arrest would have a higher average sentence. This coefficient is statistically significant. Conversely, we did not expect an individual with a felony conviction to have a lower average sentence. This coefficient is statistically significant at the 1% level. However, our explanation is that people who are convicted of a felony crime tend to put up a legal fight and, as a result, it leads them to getting a smaller average sentence.

Additionally, we ended up dropping income, unemployment status, and the arrest ratio variables from our model. This helped render a higher adjusted R squared and improved the overall relationship between our explanatory variables and the dependent variable. We kept the race, prior arrest, and felony convictions due to their statistical significance. What is interesting is the race question. It turns out that blacks fared worse than Hispanics and other races. Our regression model rendered a statistically significant coefficient that ties blacks’ disadvantage in average sentencing. In contrast, Hispanics tended to fare better than blacks. However, other races fared better than both of them.

Finally, our econometric examination throughout the paper imparts some lessons our policy makers and researchers could glean to improve criminal justice reform and policy making. We can say that blacks are disadvantaged more than Hispanics and other races. Given this point, they must use this finding to propose and implement better criminal justice policies, and help close the gap in discrimination toward blacks. For our nation to envision its prosperous future and for us to help make it “a more perfect union” as the founders admonished, it is critical that they address discrimination against blacks in the criminal justice system.